HIGH-DIMENSIONAL THEORY OF LORA FINE-TUNING IN A SOLVABLE ATTENTION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a high-dimensional statistical theory of low-rank adaptation (LoRA) in attention models, capturing the interplay between pre-training and fine-tuning. We introduce a solvable framework in which a single-head attention layer is first pre-trained on a data-abundant task and subsequently adapted via a rank-one LoRA update on limited data. In the high-dimensional limit, we obtain an asymptotic characterization of both stages in terms of a finite set of order parameters, yielding explicit predictions for test errors and representation alignment. Our analysis shows that the impact of pre-training on LoRA is summarized by an effective noise component, from which we derive prescriptions for the optimal pre-training procedure. We also demonstrate a regime with a mismatch between the value of the test error and representation quality, and propose an application of our theory to active fine-tuning.

1 INTRODUCTION

Large pre-trained attention models are rarely retrained from scratch for each new task. Instead, they are adapted to downstream data through parameter-efficient fine-tuning methods, which keep most pre-trained weights frozen and optimize only a small number of task-specific parameters. Among these methods, low-rank adaptation (LoRA) has become a standard approach: it injects trainable low-rank updates into the weight matrices of a frozen transformer, often matching or approaching full fine-tuning performance while substantially reducing memory, storage, and optimization costs Hu et al. (2022). This empirical success is part of a broader line of parameter-efficient adaptation methods, including adapters Houlsby et al. (2019), prefix tuning Li & Liang (2021) and QLoRA Dettmers et al. (2023). A recurring empirical observation is that fine-tuned adaptation appears to live in a much lower-dimensional space than pre-training, as also suggested by intrinsic-dimension studies of language-model fine-tuning Aghajanyan et al. (2021).

Recent theoretical work has begun to clarify some aspects of LoRA. Existing results mainly address its expressivity, optimization geometry or dynamics Zeng & Lee (2024); Zhang et al. (2025); Kim et al. (2025); Xu et al. (2025b); Liang et al. (2025); Mu & Klabjan (2025); Nwemadji et al. (2026); Steele (2026). Closer to our work, Kratsios et al. (2025) derives generalization bounds for asymmetric LoRA, where one factor of the LoRA is frozen at random and only the other is trained, while Jang et al. (2024) provides an NTK-based analysis that characterizes optimization and generalization in a lazy regime. These works shed light on important facets of LoRA, but leave open the question of a quantitative statistical theory for adaptation after pre-training. In particular, a characterization of how fine-tuning performance depends on both (large) pre-training and (small) fine-tuning sample sizes and pre-training quality is still lacking.

In this work we develop such a theory in a solvable high-dimensional model of LoRA fine-tuning for attention. We consider a single-head tied-attention layer that is first pre-trained on an extensive-rank attention-indexed (AIM) target Boncoraglio et al. (2025; 2026) with a large number of samples, and is then adapted to a downstream/fine-tuning task by only learning a rank-one LoRA component from a much smaller dataset. This yields a minimal setting in which pre-training a large model, freezing, and low-rank adaptation coexist in a form that remains analytically tractable. Our main result is an asymptotic characterization of the LoRA-fine-tuned estimator. In the high-dimensional limit, both the pre-training stage and the fine-tuning stage are described by a finite set of scalar order parameters, from which we obtain explicit low-dimensional predictions for the pre-training test error,

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.

the fine-tuning test error, and the overlaps with the ground truth. This provides, as far as we know, the first quantitative learning-curve theory for LoRA fine-tuning after attention pre-training.

Contributions. Our main contributions are:

1. We introduce a solvable high-dimensional setting of LoRA fine-tuning in attention, with a full-rank pre-training stage on a large dataset and a rank-one fine-tuning on a much more limited dataset.
2. We derive asymptotic predictions for the LoRA test error and reconstruction overlap in terms of finite-dimensional order parameters. Technically, we obtain a joint characterization of pre-training and fine-tuning by accounting for the statistical dependencies between the two stages, for fresh or reused inputs.
3. We identify an effective-noise mechanism that quantifies how pre-training quality controls fine-tuning performance. For linear attention, this single parameter guides pre-training and suggests inexpensive ways to reduce effective noise through recalibration of the pre-trained weights and an idealized active-learning rule. We show that these prescriptions extend qualitatively to softmax attention.
4. We identify a mismatch between fine-tuning test error and downstream signal reconstruction due to pre-training memorization when input sequences are reused for fine-tuning.

Further related work. Our work fits into a growing line of exact high-dimensional theories for feature learning, from multi-index models Arous et al. (2021); Damian et al. (2023); Arnaboldi et al. (2023); Berthier et al. (2025) to sequence multi-index models Cui (2025); Cui et al. (2024); Troiani et al. (2025); Arnaboldi et al. (2025); Duranthon et al. (2026) and attention architectures Boncoraglio et al. (2025; 2026). The specific contribution of the present paper is to bridge extensive-rank pre-training and low-rank adaptation, and thereby provide a statistical theory of LoRA in the feature-learning regime rather than in a kernel or purely optimization-based limit.

Notations. For a positive integer n we write $[n] = \{1, \dots, n\}$; I_n the $n \times n$ identity matrix; $\mathbb{S}(n)$ the set of symmetric $n \times n$ matrices. $||\cdot||_F$ is the Frobenius norm of a matrix. We denote a Gaussian law centered at ω with covariance V as $\mathcal{N}(\omega, V)$.

2 A TWO-STAGE MODEL FOR PRE-TRAINING AND LoRA FINE-TUNING

We consider sequences of T tokens of dimension D for a sequence-to-sequence task. We have access to a large dataset of pre-training samples and to a smaller dataset for fine-tuning. More precisely, we consider a set of N pre-training input and output sequences $\mathcal{S} = (X_\mu \in \mathbb{R}^{T\times D}, y_\mu \in \mathbb{R}^{T\times D})_{\mu\in[N]}$ and a set of N' fine-tuning input and output sequences $\mathcal{S}' = (X'_\mu \in \mathbb{R}^{T\times D}, y'_\mu \in \mathbb{R}^{T\times D})_{\mu\in[N']}$.

Model Architecture: we consider an attention layer with tied keys and queries and identity values, parameterized by the weights $W \in \mathbb{R}^{D\times P}$, with P the rank of W, and $w \in \mathbb{R}^D$. W is the large-rank part of the weights that are pre-trained, while w is the low-rank part that is fine-tuned. Given a sample X, y from $\mathcal{S}$ or $\mathcal{S}'$, the attention predicts the sequence

$$\hat{y}(X; W, w) = \sigma(A - \mathbb{E}_X A)X\,, \qquad A = D^{-1}X\left(P^{-1/2}WW^\top + ww^\top\right)X^\top\,, \tag{1}$$

where $\sigma : \mathbb{R}^{T\times T} \to \mathbb{R}^{T\times T}$ is the activation function. In the following we will consider it to be the row-wise softmax or the rescaled identity $\sigma : x \mapsto x/T$. A is the pre-activated attention matrix. $\mathbb{E}_X$ is the empirical expectation over the sequences of the set $\mathcal{S}$ or $\mathcal{S}'$ and the term $-\mathbb{E}_X A$ is a batch normalization. The scaling with respect to D and P is taken so that all the elements of A are of order one, and W and w contribute comparably to A.

Pre-training: we first pre-train the attention by minimizing the following empirical loss

$$\mathcal{L}(W) = \sum_{\mu=1}^{N} D^{-1}||\hat{y}(X_\mu; W, 0) - y_\mu||_F^2\,. \tag{2}$$

We fix $\hat{W} \in \arg\min_W(\mathcal{L} + \mathcal{R})$ a minimizer of the loss, with $\mathcal{R}(W) = \lambda||W||_F^2$ a regularization.

LoRA fine-tuning: we then fine-tune the attention on the loss

$$\mathcal{L}'(w) = \sum_{\mu=1}^{N'} D^{-1}||\hat{y}(X'_\mu; \hat{W}, w) - y'_\mu||_F^2 \tag{3}$$

taking $\hat{w} \in \arg\min_w(\mathcal{L}' + \mathcal{R}')$, with $\mathcal{R}'(w) = \lambda'||w||_2^2$. $\lambda, \lambda' > 0$ are the regularization strengths.

Measured performances: we are interested in the test errors of the attention with weights $\hat{W}, \hat{w}$ on test pre-training and fine-tuning samples

$$\mathcal{E}(\hat{W}) = \mathbb{E}_{(X,y)\sim\mathcal{P}}\left[D^{-1}||\hat{y}(X;\hat{W},0) - y||_F^2 \,\middle|\, \mathcal{S}\right] \tag{4}$$

$$\mathcal{E}'(\hat{W},\hat{w}) = \mathbb{E}_{(X',y')\sim\mathcal{P}'}\left[D^{-1}||\hat{y}(X';\hat{W},\hat{w}) - y'||_F^2 \,\middle|\, \mathcal{S},\mathcal{S}'\right] \tag{5}$$

We will also consider the reconstruction error of the LoRA task measured as the cosine similarity (or *overlap*) between the learned fine-tuning weights and the ground truth $o_w = |\hat{w}^\top w^*|/\|\hat{w}\|_2\|w^*\|_2$.

For a generic pre-training and fine-tuning dataset, the two-stage optimization above is too complex to admit a sharp analytical description. To make the problem tractable while preserving the key ingredients of pre-training followed by low-rank adaptation, we consider the data to be synthetic, generated by a target of the same type of the analyzed architecture.

Probabilistic model for synthetic data: let P_0 be the rank of the target and take $W^* \in \mathbb{R}^{D\times P_0}$, $w^* \in \mathbb{R}^D$ common to all samples, with $W^*_{i,j} \sim \mathcal{N}(0,1)$ and $w^*_i \sim \mathcal{N}(0,1)$ for all $i \in [D], j \in [P_0]$. We consider two possible scenarios, that we distinguish by the index $\mathrm{e} \in \{0,1\}$:
i) fine-tuning on different data: in this case the sequences X and X' of the pre-training and fine-tuning train sets are different and we draw them independently. We set $\mathrm{e} = 0$.
ii) fine-tuning on reused sequences: in this case the input sequences are shared between the pre-training and fine-tuning, while the output sequences correspond to different tasks. We set $\mathrm{e} = 1$.
We set $\mathcal{P}$ and $\mathcal{P}'$ to be the pre-training and fine-tuning distributions of the samples. They are jointly defined by

$$(\mathcal{P},\mathcal{P}') : \begin{cases} X \sim \mathcal{P}_0 \\ \xi \sim \mathcal{P}_\Delta \\ A^* = D^{-1}P_0^{-1/2}XW^*W^{*\top}X^\top \\ y = \sigma(A^* - \mathbb{E}_X A^* + \xi)X \end{cases}, \begin{cases} X' \sim \mathrm{e}\delta_X + (1-\mathrm{e})\mathcal{P}_0 \\ \xi' \sim \mathrm{e}\delta_\xi + (1-\mathrm{e})\mathcal{P}_\Delta \\ A'^* = D^{-1}X'\left(P_0^{-1/2}W^*W^{*\top} + w^*w^{*\top}\right)X'^\top \\ y' = \sigma(A'^* - \mathbb{E}_{X'}A'^* + \xi')X' \end{cases} \tag{6}$$

where $(\mathcal{P}_0)_{a,i} = \mathcal{N}(0,1)$ for all $a \in [T], i \in [D]$; $\xi, \xi' \in \mathcal{S}(T)$ are symmetric noise with strength $\Delta \geq 0$ and law $(\mathcal{P}_\Delta)_{a,b} = \mathcal{N}(0, \Delta/(2-\delta_{a,b}))$ for all $a \leq b$. The factor $1/(2-\delta_{a,b})$ takes into account the symmetrization. $\mathbb{E}_X$ and $\mathbb{E}_{X'}$ are expectations over the sequences. For $\mathrm{e} = 0$ the two distributions factorize over the samples, X, X', ξ and ξ' being independent; while for $\mathrm{e} = 1$ we match all the randomness, having $X = X'$ and $\xi = \xi'$ for each sample. $\mathcal{S}$ and $\mathcal{S}'$ are then constructed by independently drawing N' samples from $(\mathcal{P},\mathcal{P}')$ and the remaining $N - N'$ from $\mathcal{P}$.

Asymptotic limit: we consider a high-dimensional limit $D \to \infty$ together with

$$\frac{P_0}{D} \to \kappa_0 = \Theta(1)\,, \quad \frac{P}{D} \to \kappa = \Theta(1)\,, \quad \frac{N}{D^2} \to \alpha = \Theta(1)\,, \quad \frac{N'}{D} \to \alpha' = \Theta(1) \tag{7}$$

with $\kappa \geq 1$ and a finite number of tokens $T = \Theta(1)$. This means that, for the pre-training, we consider W and W^* to have extensive rank and we have access to a quadratically large number of pre-training samples; on the other hand, for the fine-tuning we consider w and w^* low-rank (rank one) and we only have access to a limited linear number of samples. We emphasize that $N \gg N'$, which corresponds to the common case where the fine-tuning samples are limited compared to the pre-training samples.

Comments and limitations: The shared extensive-rank target W^* allows us to isolate how the representation learned during pre-training affects downstream adaptation, while the noises ξ, ξ' capture residual task variability. We deliberately focus on tied query-key weights, identity values and a rank-one update: this gives a tractable setting with one extensive-rank matrix and one vector, while retaining the statistical coupling between pre-training and fine-tuning. Untied query-key weights, trainable values and higher-rank updates can be accommodated within the same general framework, but introduce additional coupled order parameters that are not needed to study this interaction. Similarly, the assumption $\kappa \geq 1$ avoids a rank constraint on the pre-trained positive-semidefinite matrix; treating $\kappa < 1$ would require retaining this constraint in the analysis. These choices keep the joint characterization explicit and let us focus on the transfer from the frozen representation to the learned low-rank component.

3 ASYMPTOTIC CHARACTERIZATION OF THE LORA

This section describes our main technical contribution: the asymptotic characterization of attention after pre-training and LoRA fine-tuning. Results 1 and 2 describe the distribution of test and train pre-activations through a few low-dimensional order parameters; Result 3 relates these distributions to prediction and reconstruction errors; and Results 4 and 5 determine the order parameters through stationarity equations.

We build on the extensive-rank pre-training analysis of Boncoraglio et al. (2026) and on the low-rank framework of Cui (2025). These works describe each of the stages of training in isolation. Our technical contribution is to couple the two in a way that remains asymptotically sharp, and to determine the coupling between the pre-training and the fine-tuning, with different laws for fresh and reused inputs (Appendices A.4–A.5). Appendix A gives the detailed derivation roadmap. The derivation relies on Gaussian equivalence to determine the effective pre-activations, random matrix theory to describe the spectrum of the extensive-rank component, and replica/approximate message passing (AMP) calculations for the stationarity equations. The final characterization is a replica-symmetric prediction, supported by simulations and by replicon stability checks, as detailed in Appendix A.6.

Order parameters: in the high-dimensional limit, the performances of the trained attention $\mathcal{L}$, $\mathcal{L}'$, $\mathcal{E}$, $\mathcal{E}'$ (in equations (2),(3) and (4)) and o_w can be solely described in terms of the following scalar order parameters (or *sufficient statistics*), setting $S^* = P_0^{-1/2} W^* W^{*\top}$ and $\hat{S} = P^{-1/2} \hat{W} \hat{W}^\top$:

$$Q_0 = D^{-2} \operatorname{Tr} S^* S^* \qquad Q = D^{-2} \operatorname{Tr} \hat{S} \hat{S} \qquad M = D^{-2} \operatorname{Tr} \hat{S} S^* \tag{8}$$

$$q_0 = D^{-1} w^{*\top} w^* \qquad q = D^{-1} \hat{w}^\top \hat{w} \qquad m = D^{-1} \hat{w}^\top w^* \tag{9}$$

Q and q quantify the norms of the learned weights after training, while M and m quantify how much the learned weights are aligned with the ground truth signals $W^\star, w^\star$. In the high-dimensional limit we have $Q_0 = 1 + \kappa_0$ and $q_0 = 1$. We moreover introduce the order parameters $V, v \in \mathbb{R}_+$, that quantify the inverse curvature of the train loss at the minimizer. Large V or large v means that the loss is flat, thus the minimizer is sensitive to perturbations and likely overfits data.

3.1 EFFECTIVE LOW-DIMENSIONAL PRE-ACTIVATIONS

In our setting, high-dimensional concentration allows us to describe the predictors $\hat{W}$ and $\hat{w}$ through a finite-dimensional effective problem. In particular, the pre-training and fine-tuning test errors can be expressed in terms of the joint distribution of a small number of scalar pre-activations associated with the target and the trained attention. We therefore introduce these effective variables first. For a sequence X, we define the pre-activations of the target z^*, χ^* and the trained attention z, χ as follows.

$$z^* = A^* - \mathbb{E}_X A^* + \xi \qquad z = A - \mathbb{E}_X A \tag{10}$$

$$A^* = D^{-1} P_0^{-1/2} X W^* W^{*\top} X^\top \qquad A = D^{-1} P^{-1/2} X \hat{W} \hat{W}^\top X^\top \tag{11}$$

$$\chi^* = D^{-1/2} X w^* \qquad \chi = D^{-1/2} X \hat{w} \tag{12}$$

with the $T \times T$ tied-attention logits $A, A^\star$. We characterize the distribution of these pre-activations firstly for X being a new test sequence.

Result 1 (Test samples)**.** *Let $X \sim \mathcal{P}_0$ be a fresh test sequence, independent of both the pre-training set $\mathcal{S}$ and the fine-tuning set $\mathcal{S}'$, and let $\xi \sim \mathcal{P}_\Delta$ be independent noise. In the high-dimensional limit, the pre-training pre-activations $(z^*, z) \in (\mathcal{S}(T))^2$ have joint law $\mathcal{Q}$, defined for each $1 \le a \le b \le T$ by*

$$\begin{pmatrix} z^*_{a,b} \\ z_{a,b} \end{pmatrix} = \begin{pmatrix} z^*_{b,a} \\ z_{b,a} \end{pmatrix} \sim \mathcal{Q}_{ab} = \mathcal{N}\left(0, (1+\delta_{a,b}) \begin{pmatrix} \Delta/2 + Q_0 & M \\ M & Q \end{pmatrix}\right) . \tag{13}$$

Distinct unordered token pairs are independent in this limiting law. Wick contractions give the factor $1 + \delta_{a,b}$, while the additive noise Δ contributes only to the target variance.

For the same test sequence X, the rank-one pre-activations $(\chi^*, \chi) \in (\mathbb{R}^T)^2$ have limiting joint law $\mathcal{Q}'$, defined for each $a \in [T]$ by

$$\begin{pmatrix} \chi^*_a \\ \chi_a \end{pmatrix} \sim \mathcal{Q}'_a = \mathcal{N}\left(0, \begin{pmatrix} q_0 & m \\ m & q \end{pmatrix}\right) . \tag{14}$$

The pairs (χ^*_a, χ_a) are independent across tokens.

We now turn to train samples. Because of the correlations between the trained weights and the train sequences, the distribution of the pre-activations is more involved. We introduce the proximal operators, for the pre-training and the fine-tuning. For symmetric $\tilde{z} \in \mathcal{S}(T)$ and for $\tilde{\chi} \in \mathbb{R}^T$ define the following scalar potentials

$$\Psi(z; z^*, \tilde{z}) = -||\sigma(z) - \sigma(z^*)||_F^2 - \sum_{a \leq b}^{T} \frac{1}{2V(1+\delta_{a,b})}(\tilde{z}_{a,b} - z_{a,b})^2 \tag{15}$$

$$\psi(\chi; \chi^*, \tilde{\chi}, z^*, z) = -||\sigma(z + \chi\chi^\top - qI_T) - \sigma(z^* + \chi^*\chi^{*\top} - q_0 I_T)||_F^2 - \sum_{a}^{T} \frac{1}{2v}(\tilde{\chi}_a - \chi_a)^2$$

We define the proximal operators as follows

$$\mathrm{Prox}(z^*, \tilde{z}) = \underset{z \in \mathcal{S}(T)}{\arg\max}\, \Psi(z; z^*, \tilde{z})\,, \qquad \mathrm{prox}(\chi^*, \tilde{\chi}, z^*, z) = \underset{\chi \in \mathbb{R}^T}{\arg\max}\, \psi(\chi; \chi^*, \tilde{\chi}, z^*, z)\,. \tag{16}$$

We can now state our result on the distribution of the pre-activations on the train samples.

Result 2 (Train samples)**.** *Consider a uniformly chosen sample from the pre-training set $\mathcal{S}$, with input X and associated noise ξ. Then (z^*, z) have limiting distribution $\mathcal{R}$, defined by*

$$\mathcal{R} : (z^*, \tilde{z}) \sim \mathcal{Q} \quad \text{and} \quad z = \mathrm{Prox}(z^*, \tilde{z})\,. \tag{17}$$

Consider X being uniformly sampled from the fine-tuning set $\mathcal{S}'$. Then (χ^*, χ, z^*, z) follow the distribution $\mathcal{R}'$ defined by

$$\mathcal{R}' : (\chi^*, \tilde{\chi}) \sim \mathcal{Q}'\,, \quad (z^*, z) \sim \begin{cases} \mathcal{Q} & \text{if e} = 0 \\ \mathcal{R} & \text{if e} = 1 \end{cases} \quad \text{and} \quad \chi = \mathrm{prox}(\chi^*, \tilde{\chi}, z^*, z)\,. \tag{18}$$

Here $(\chi^*, \tilde{\chi})$ is drawn independently of (z^*, z). Result 2 expresses the training pre-activations through the proximal operators defined above. To give some intuition on their role, we consider two limiting cases. For pre-training, as $V \to 0$, the quadratic penalty in equation (15) forces z towards $\tilde{z}$, recovering the test distribution $\mathcal{Q}$. As $V \to \infty$, the penalty vanishes and the proximal optimization favors fitting the sample, with zero loss whenever $\sigma(z) = \sigma(z^*)$. The fine-tuning proximal has the same interpretation: small v keeps χ close to $\tilde{\chi}$, while large v favors minimizing the sample loss for the fixed pre-training pre-activations z, z^*.

A consequence of the characterizations given in results 1 and 2 is that the performances of the trained attention can be entirely expressed in terms of the order parameters in equations (8).

Result 3 (Performances in terms of the order parameters)**.** *Take $(z^*, z) \sim \mathcal{Q}$ and $(\chi^*, \chi) \sim \mathcal{Q}'$. In the high-dimensional limit, the test errors after pre-training $\hat{W}$ and fine-tuning $\hat{w}$ are*

$$\mathcal{E}(\hat{W}) = \mathbb{E}\,||\sigma(z) - \sigma(z^*)||_F^2\,, \quad \mathcal{E}'(\hat{W}, \hat{w}) = \mathbb{E}\,||\sigma(z + \chi\chi^\top - qI_T) - \sigma(z^* + \chi^*\chi^{*\top} - q_0 I_T)||_F^2 \tag{19}$$

Take $(z^*, z) \sim \mathcal{R}$ or $(\chi^*, \chi, z^*, z) \sim \mathcal{R}'$. In the high-dimensional limit the train losses of the trained attention can be expressed as

$$\frac{\mathcal{L}(\hat{W})}{N} = \mathbb{E}_{\mathcal{R}}\,||\sigma(z) - \sigma(z^*)||_F^2\,, \quad \frac{\mathcal{L}'(\hat{W}, \hat{w})}{N'} = \mathbb{E}_{\mathcal{R}'}\,||\sigma(z + \chi\chi^\top - qI_T) - \sigma(z^* + \chi^*\chi^{*\top} - q_0 I_T)||_F^2$$

Moreover the cosine similarity between $\hat{w}$ and $w^\star$ converges to $o_w = m/\sqrt{q q_0}$.

3.2 Determination of the order parameters

The laws $\mathcal{Q}, \mathcal{Q}', \mathcal{R}, \mathcal{R}'$ depend on the order parameters Q, M, V, q, m, v. Determining these parameters completes our characterization of attention after pre-training and LoRA fine-tuning.

We introduce the spectral distribution of the pre-trained weights. Let μ_0 be the limiting Marchenko-Pastur spectral density of the rescaled target $S^*/\sqrt{D}$, $\mu_{\mathrm{sc},\eta}(x) = \sqrt{4\eta^2 - x^2}/(2\pi\eta^2)$ the semi-circle distribution of radius $2\eta > 0$, $\boxplus$ the free convolution and $\mu_\eta = \mu_0 \boxplus \mu_{\mathrm{sc},\eta}$.

Result 4 (Order parameters of the pre-training, taken from Boncoraglio et al. (2026))**.** *Let $\Theta = (\hat{M}, \hat{Q}, \hat{V}) \in \mathbb{R}_+^3$. Take $(z^*, \tilde{z}) \sim \mathcal{Q}$. Consider the following free entropy function*

$$\Phi = -\frac{1}{2\alpha} M\hat{M} + \frac{1}{4\alpha}(Q\hat{V} - V\hat{Q}) + \frac{\hat{M}^2}{4\alpha\hat{V}} J\left(\frac{\hat{Q}^{1/2}}{\hat{M}}, \frac{2\sqrt{\kappa}\lambda}{\hat{M}}\right) + \mathbb{E}_{z^*, \tilde{z}} \max_{z \in \mathcal{S}(T)} \Psi(z; z^*, \tilde{z}) \tag{20}$$

where $J(\eta, \epsilon) = \int_\epsilon^{+\infty} \mu_\eta(\mathrm{d}x)(x - \epsilon)^2$. Then the pre-training order parameters Q, M, V, together with their conjugates $\hat{Q}, \hat{M}, \hat{V}$ are determined by the physical stable extremum of Φ.

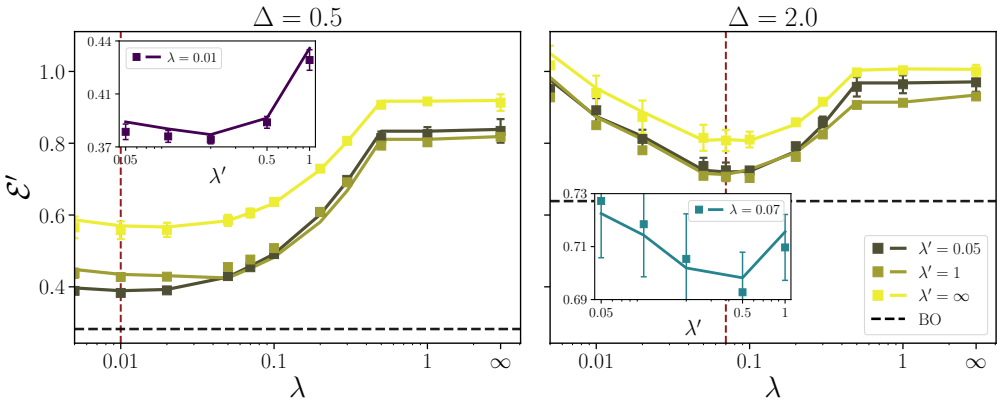


Figure 1: Fine-tuning test error $\mathcal{E}'$ versus pre-training regularization λ for $\lambda' \in \{0.05, 1, \infty\}$, at $\alpha = 0.1$, $\alpha' = 3$, $T = 3$, $\kappa_0 = \kappa = 1$, and noise levels $\Delta = 0.5$ (left) and $\Delta = 2$ (right). Insets show $\mathcal{E}'$ versus λ' at the optimal λ. Solid lines are theoretical predictions; square markers are simulations at $D = 150$ averaged over four runs. Vertical dark-red dashed lines mark curve minima, and the black dashed line is the Bayes-optimal benchmark. The cases $\lambda' = \infty$ and $\lambda = \infty$ correspond respectively to no LoRA adaptation ($\hat{w} = 0$) and LoRA-only fine-tuning ($\hat{W} = 0$).

Result 5 (Order parameters of the fine-tuning)**.** *Let $\theta = (\hat{m}, \hat{q}, \hat{v}) \in \mathbb{R}^3_+$. Take $(z^*, z) \sim \mathcal{Q}$ if $\mathrm{e} = 0$ else $(z^*, z) \sim \mathcal{R}$. Take $(\chi^*, \tilde{\chi}) \sim \mathcal{Q}'$. Consider the following free entropy function*

$$\phi = -\frac{1}{\alpha'} m\hat{m} + \frac{1}{2\alpha'}(q\hat{v} - v\hat{q}) + \frac{\hat{m}^2 + \hat{q}}{2\alpha'(2\lambda' + \hat{v})} + \mathbb{E}_{\chi^*, \tilde{\chi}, z^*, z} \max_{\chi \in \mathbb{R}^T} \psi(\chi; \chi^*, \tilde{\chi}, z^*, z) \,. \tag{21}$$

Then the fine-tuning order parameters q, m and v are determined as $(q, m, v, \hat{q}, \hat{m}, \hat{v}) = \operatorname{extr} \phi$.

Results 4 and 5 determine the values of the order parameters in terms of the extremization of two low-dimensional potentials Φ and ϕ. This extremization is equivalent to solving the systems of self-consistent equations described in Appendix A. Their solution fixes the covariances of the Gaussian laws $\mathcal{Q}$ and $\mathcal{Q}'$. Sampling pre-activations from these laws then allows us to evaluate the test errors in Result 3, while the reconstruction overlap follows directly from the order parameters.

In the following we will compare the performances of the trained attention to the *Bayes-optimal* (BO) performances. These are obtained by the best achievable reconstruction of W^* and w^* conditionally on $\mathcal{S}, \mathcal{S}'$, assuming the knowledge of the probabilistic data model. This constitutes the natural benchmark when having access to a limited number of samples. In the asymptotic limit we consider, the BO performances can be exactly characterized.

Result 6 (Bayes-optimal performances)**.** *In the high-dimensional limit, the BO pre-training and fine-tuning test errors $\mathcal{E}_{\mathrm{BO}}$ and $\mathcal{E}'_{\mathrm{BO}}$ are characterized by two scalar overlaps, $Q_{\mathrm{BO}} \in [0, Q_0]$ and $q_{\mathrm{BO}} \in [0, q_0]$. These four quantities are determined by the system of equations given in Appendix B.*

4 Consequences and Applications

We now discuss the main consequences of our asymptotic characterization. We start by validating the theory with numerical simulations of the full pre-training and LoRA fine-tuning pipeline. We then use our theory to isolate the effects of pre-training on the downstream task and we derive protocols to improve fine-tuning performance.

We compare our theoretical predictions to numerical simulations of the pre-training and fine-tuning of the attention at finite D on Fig. 1 for independent fine-tuning sequences, namely $\mathrm{e} = 0$ at $D = 150$, and also in Fig. 3 for shared sequences $\mathrm{e} = 1$ at $D = 10^3$. In both cases we show very good agreement. Our predictions are further supported by Figs. 2, and 4, which focus on the fine-tuning given an effective pre-training. All details on the numerics are given in App. D.

As a baseline, we consider the two limiting cases when there is no fine-tuning or when there is no pre-training. These are obtained by respectively taking the limit $\lambda' \to +\infty$, which set $\hat{w} = 0$, or $\lambda \to +\infty$, which sets $\hat{W} = 0$. Fig. 1 shows that, as expected, the LoRA consistently improves over the frozen pre-training baseline $\lambda' = \infty$, showing that the rank-one update $\hat{w}$ effectively exploits

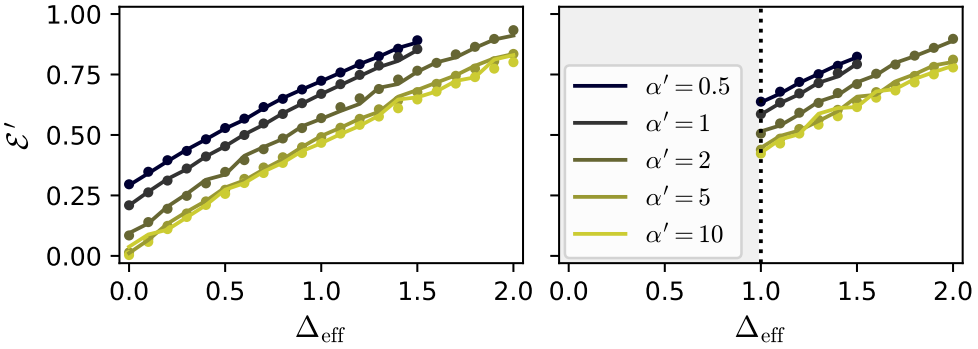


Figure 2: Effect of pre-training on the fine-tuning task. Fine-tuning test error $\mathcal{E}'$ vs $\Delta_{\text{eff}} = \Delta/2 + Q_0 - 2M + Q$, parameterized by fixing $M = 1$ and varying $Q_0 = Q$. Left: $\Delta = 0$; right: $\Delta = 2$. Lines are the theoretical predictions; dots are simulations at $D = 10^4$ over five independent runs. $T = 3$, $\lambda' = 0.5$, σ softmax activation.

information not captured by $\hat{W}$ alone. Conversely, taking λ too large does not produce an informative representation $\hat{W}$ and degrades fine-tuning test performances. Fig. 1 further informs the choice of the hyperparameters λ, λ' that are optimal for the test error.

4.1 Independent pre-training and fine-tuning samples

We first consider the case $\mathrm{e} = 0$ where the pre-training and fine-tuning input samples are independent. To gain insights on the effect of the pre-training on the fine-tuning performances, we consider a linear activation function $\sigma(u) = u/T$; we will then generalize our result to softmax activations. The LoRA training error, expressed in terms of the pre-activations in Result 3, is then

$$\mathcal{L}' = T^{-2}\mathbb{E}||\underbrace{z - z^*}_{h} + \chi\chi^\top - \chi^*\chi^{*\top} - (q - q_0)I_T||_F^2 \tag{22}$$

with $h_{ab} \sim \mathcal{N}(0, (1 + \delta_{ab})\Delta_{\text{eff}})$: pre-training affects fine-tuning only through the noise level

$$\Delta_{\text{eff}} = \frac{\Delta}{2} + Q_0 - 2M + Q\,. \tag{23}$$

The derivation is given in Appendix C. The effective noise Δ_{eff} combines the irreducible contribution $\Delta/2$ with the reconstruction error $Q_0 - 2M + Q$ of the pre-trained matrix. Better pre-training reconstruction reduces this contribution, reaching the minimum $\Delta_{\text{eff}} = \Delta/2$ when $\hat{S} = S^*$. For linear activation, hence, Δ_{eff} completely determines the impact of the pre-training impact on the fine-tuning stage. Lower effective noise thus facilitates recovery of w^* during fine-tuning.

For softmax activation, the channel depends separately on $Q_0 + \Delta/2$, Q or M, and Δ_{eff} is no longer a sufficient statistic; yet Fig. 2 shows similar qualitative trends along the considered parameters. Fig. 2 further shows the benefits of pre-training, in terms of the number of fine-tuning samples $N' = \alpha' D$ required to achieve a given test error. In the cases we consider, diminishing Δ_{eff} with a better pre-training is always beneficial and reduces the number of required samples to reach the same performance or to decrease the fine-tuning test error at constant α'. The same holds for a linear σ, as depicted in Fig. 5 in Appendix E.

Finally, the effective-noise description suggests a simple, inexpensive way to improve fine-tuning: recalibrating the frozen weights $\hat{W} \to \sqrt{\beta}\hat{W}$, with $\beta \geq 0$, to reduce this noise without repeating pre-training. For linear σ and $Q > 0$, minimizing over β gives

$$\Delta_{\text{eff}}(\beta) = \frac{\Delta}{2} + Q_0 + \beta^2 Q - 2\beta M, \qquad \beta_{\min} = \frac{M}{Q}, \qquad \Delta_{\text{eff}}(\beta_{\min}) = \frac{\Delta}{2} + Q_0 - \frac{M^2}{Q}. \tag{24}$$

The derivation is given in Appendix C. The optimum M/Q is an oracle value, since M depends on the unknown target $W^\star$. In practice, β can be tuned on held-out downstream data to improve transfer without repeating pre-training or retuning λ. This is related in spirit to the scale factors used in LoRA and its variants Hu et al. (2022); Kalajdzievski (2023); Bini et al. (2025). Appendix E examines recalibration for softmax attention and its interplay with pre-training regularization (Figure 6).

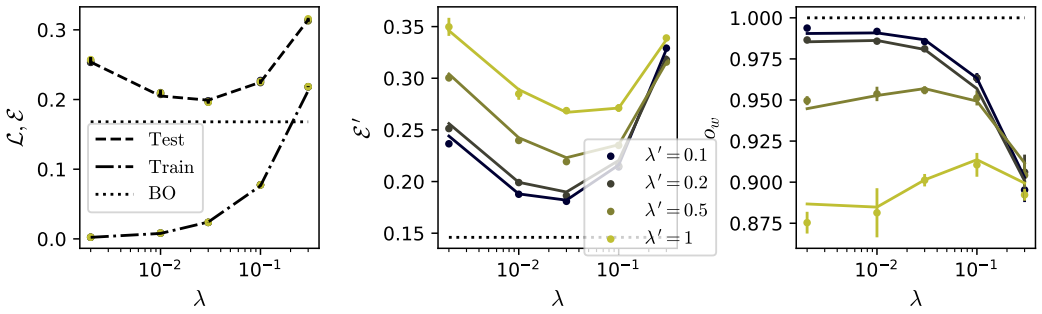


Figure 3: Fine-tuning input samples from the pre-training set $\mathrm{e} = 1$. Left: pre-training train error $\mathcal{L}$ and test error $\mathcal{E}$; center: fine-tuning test error $\mathcal{E}'$; right: fine-tuning overlap $o_w = |\hat{w}^\top w^*| / \|\hat{w}\|_2 \|w^*\|_2$; vs regularizations λ and various λ'. Lines are the theoretical predictions; dots are simulations at $D = 10^3$ over five independent runs. $T = 2, \Delta = 0.5, \kappa_0 = 0.5, \kappa = 1, \alpha = 0.2, \alpha' = 3, \sigma$ softmax.

4.2 Reused sequences

We now consider the case $\mathrm{e} = 1$, where fine-tuning reuses pre-training sequences. Unlike the independent-input case, minimizing the effective noise during fine-tuning need not minimize the test error $\mathcal{E}'$. Indeed, memorizing the pre-training data gives $z_\mu \approx z^*_\mu$ on the reused sequences, reducing the noise seen during fine-tuning without necessarily improving predictions on fresh inputs. Hence, the best pre-training procedure faces a trade-off between good fine-tuning test error $\mathcal{E}'$ and good fine-tuning reconstruction o_w. The effective noise in the fine-tuning stage is thus reduced to $\Delta_{\text{eff}}(\mathrm{e} = 1) \approx 0$ and this leads to a good reconstruction of w^*. Because of overfitting, the test errors $\mathcal{E}, \mathcal{E}'$ are not optimal in that regime. This is depicted in Fig. 3, which shows that, for the considered value of the model parameters, overfitting the pre-train data by taking $\lambda \to 0$ maximizes the fine-tuning overlap o_w; while there is a finite $\lambda \approx 0.3$ that minimizes $\mathcal{E}$ and $\mathcal{E}'$. We provide a precise description of the reduction of the noise obtained by reusing sequences: as detailed in App. C, for linear σ, we have

$$\Delta_{\text{eff}}(\mathrm{e} = 1) = (1 + 4VT^{-2})^{-2} \Delta_{\text{eff}}(\mathrm{e} = 0) \tag{25}$$

where $V \geq 0$ is given by Result 4. Consequently the effective noise is always smaller when reusing sequences, compared to training on new sequences at $\mathrm{e} = 0$. When $V \to \infty$, the pre-training fits data and the effective noise goes to zero.

Our model highlights that there can be a mismatch between good test error and good reconstruction, and raises the question of what is a relevant evaluation metric. When the goal is to recover w^* for further downstream tasks, the reconstruction overlap o_w may therefore be more relevant than the prediction error $\mathcal{E}'$ on the current task. A similar mismatch has been proven for unsupervised settings in Conde Mendes et al. (2026); Garnier-Brun et al. (2026); we extend these observations to a supervised setting.

4.3 Active fine-tuning

Active fine-tuning offers another way to reduce the effective noise in equation (23) and improve the fine-tuning test error $\mathcal{E}'$. We consider $\mathrm{e} = 0$ and a pool $\bar{\mathcal{S}}'$ of $\bar{N}' = \bar{\alpha}' D$ unlabeled inputs, from which we can label only $N' = \alpha' D$, with labeling ratio $\rho = N'/\bar{N}' < 1$. This corresponds to the common scenario where an unlabeled pool of size $\bar{N}'$ is available, but only N' samples can be labeled due to labeling budget. We seek a procedure to determine the best samples from $\bar{\mathcal{S}}'$ to label, compared to a uniform sampling of $\mathcal{S}'$ from $\bar{\mathcal{S}}'$. We focus on the pre-activations of the pre-trained attention.

For each candidate input X'_μ, let $z_\mu = A_\mu - \mathbb{E}_X A$ be the centered pre-training logits defined in Equation (10). Algorithm 1 selects the smallest scores $\|\sigma(z_\mu) - \sigma(0)\|_F^2$, favoring inputs whose pre-trained attention output is closest to $\sigma(0)$. Since z and the fine-tuning target projections χ^* are asymptotically independent, this selection leaves the distribution of the rank-one signal unchanged. For linear activation σ, the idealized limit $\rho \to 0$ corresponds to conditioning on $z = 0$ and gives the following effective noise. For any indices a, b

$$\mathrm{Var}_{\mathcal{Q}|z=0}(z - z^*)_{ab} = (1 + \delta_{a,b}) \Delta_{\text{eff,ActFT}} \,, \qquad \Delta_{\text{eff,ActFT}} = \frac{\Delta}{2} + Q_0 - \frac{M^2}{Q} \tag{26}$$

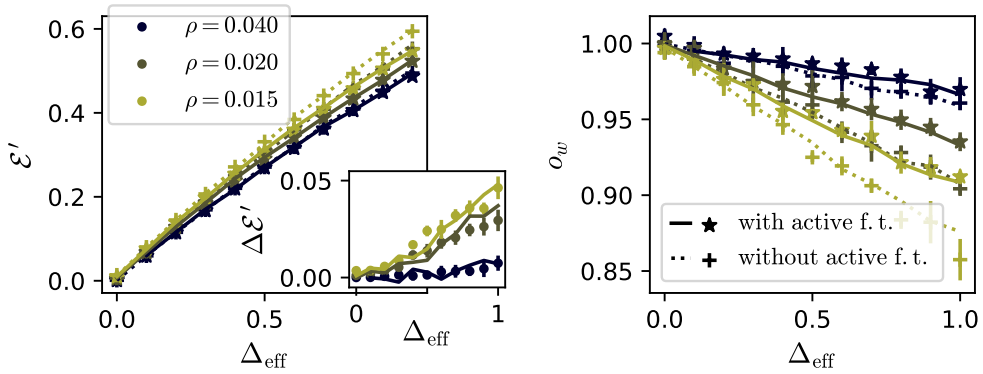


Figure 4: Active fine-tuning. Left: fine-tuning test errors $\mathcal{E}'(\hat{W}, \hat{w})$ and $\mathcal{E}'(\hat{W}, \hat{w}_{\text{ActFT}})$; inset: difference between the test errors $\Delta\mathcal{E}'$; right: overlaps o_w; vs $\Delta_{\text{eff}} = \Delta/2 + Q_0 - 2M + Q$, parameterized by fixing $M = 1$ and varying $Q_0 = Q$; for different labeling ratios ρ. Lines are the theoretical predictions; dots are simulations at $D = 10^4$ over five independent runs. We minimize $\mathcal{E}'$ over $\lambda' \in \{0.1, 0.2, 0.5, 1\}$ for each point. $\bar{\alpha}' = 100, T = 3, \Delta = 0$, σ softmax.

Details on the derivation are given in App. C. The resulting noise $\Delta_{\text{eff,ActFT}}$ is smaller than the initial noise Δ_{eff} in equation 23 for any pre-training order parameters M, Q.

Algorithm 1 Active fine-tuning.

Input: $\bar{N}'$ sequences X'_μ from $\mathcal{P}_0$; pre-trained attention.
1. Compute the pre-activation z_μ of the pre-trained attention for all X'_μ.
2. Compute the ordering $\mu(i)$ such that $(||\sigma(z_{\mu(i)}) - \sigma(0)||_F^2)_i$ is increasing.
3. Label $(X'_{\mu(i)})_{1 \le i \le N'}$.
4. Take $\mathcal{S}' = (X'_{\mu(i)}, y'_{\mu(i)})_{1 \le i \le N'}$.
5. Fine-tune the attention on $\mathcal{S}'$.

For softmax attention, Figure 4 shows the benefits of active fine-tuning over uniform sample selection, through the test-error reduction $\Delta\mathcal{E}' = \mathcal{E}'(\hat{W}, \hat{w}) - \mathcal{E}'(\hat{W}, \hat{w}_{\text{ActFT}})$ and the reconstruction overlap. For all noise levels we investigated, active learning improves the test error and allows for a better alignment with the ground truth. The effect is more pronounced when the effective noise Δ_{eff} to be reduced is large, and when the labeling ratio ρ is small, allowing more selective sampling. Our theoretical results 1 to 5 directly extend to this active fine-tuning procedure, by conditioning the distribution $\mathcal{Q}$ on z, z^* on the selection rule Alg. 1. The agreement with simulations depicted in Fig. 4 is good. While being effective, Alg. 1 relies on the independence between z and χ^* in our model: when these quantities are correlated, the same rule may discard inputs carrying useful information about the downstream task and degrade performances.

Conclusions. We introduced a solvable high-dimensional model for LoRA fine-tuning after attention pre-training, in which an extensive-rank component is learned from $N = \Theta(D^2)$ samples and a rank-one adaptation is learned from $N' = \Theta(D)$ samples. Within this setting, we characterized the predictor asymptotically and identified how pre-training affects fine-tuning through an effective noise mechanism. Our results also exhibit regimes where test error and representation quality are misaligned, extending to supervised settings observations previously made in unsupervised learning.

The single-head, rank-one analysis presented here serves as a first step toward more general settings. A natural extension is to higher-rank LoRA updates and multi-head attention, where multiple directions or heads may compete for the information contained in the frozen representation, potentially leading to a richer structure of effective noise across tasks.

More broadly, it will be important to understand how robust this effective-noise picture is beyond the present setting. This includes extending the analysis beyond Gaussian inputs to assess universality, as well as studying learning dynamics, where the interaction between pre-training and fine-tuning unfolds over time rather than at the level of the empirical risk minimizer. The flexibility of the high-dimensional framework also makes it possible to investigate related phenomena such as catastrophic forgetting (see App. A.8). Quantitative applicability to practical transformer fine-tuning remains open and requires validation on real data and architectures.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text, literature survey and reproducibility assessment of our results. We have not used generative AI tools for mathematical derivation of our results. We have reviewed all AI-assisted work, in particolar exact reproducibilty of our results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a synthetic theoretical model and does not involve any personal data or realistic dataset.

REPRODUCIBILITY STATEMENT

Section 2 specifies the data model and asymptotic scaling. Appendix A gives the derivation roadmap, effective channels and fixed-point equations; Appendix A.6 states the mathematical qualifications and numerical stability diagnostics. Appendix C derives the effective-noise formulas, and Appendix D describes the numerical procedures. Figure captions report the parameter settings, dimensions and repetitions. The supplementary material contains all the code used to reproduce our results.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 7319–7328, 2021.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional and mean-field dynamics to dimensionless odes: A unifying approach to SGD in two-layer networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1199–1227. PMLR, 2023.

Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborová. Asymptotics of sgd in sequence-single index models and single-layer attention networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20611–20645, 2025.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, 25(5):1627–1710, 2025.

Massimo Bini, Leander Girrbach, and Zeynep Akata. Decoupling angles and strength in low-rank adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 20216–20233, 2025.

Fabrizio Boncoraglio, Emanuele Troiani, Vittorio Erba, and Lenka Zdeborová. Bayes optimal learning of attention-indexed models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 105029–105074. Curran Associates, Inc., 2025. doi: 10.52202/085713-3509. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9740e3572538a1a9132d8d4ad2e8bf72-Paper-Conference.pdf.

Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. In *International Conference on Machine Learning*, 2026. arXiv:2509.24914.

Vicente Conde Mendes, Lorenzo Bardone, Cédric Koller, Jorge Medina Moreira, Vittorio Erba, Emanuele Troiani, and Lenka Zdeborová. A solvable high-dimensional model where nonlinear autoencoders learn structure invisible to PCA while test loss misaligns with generalization. In *International Conference on Machine Learning*, 2026. arXiv:2602.10680.

Hugo Cui. High-dimensional learning of narrow neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(2):023402, 2025.

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. *Advances in Neural Information Processing Systems*, 37:36342–36389, 2024.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: Optimal sample complexity for learning single index models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 752–784, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023.

O. Duranthon, P. Marion, C. Boyer, B. Loureiro, and L. Zdeborová. Statistical advantage of softmax attention: Insights from single-location regression. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.21936.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of erm in overparameterized quadratic networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 88862–88901, 2025.

Jerome Garnier-Brun, Luca Biggio, Davide Beltrame, Marc Mézard, and Luca Saglietti. Biased generalization in diffusion models, 2026. arXiv:2603.03469.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. LoRA training in the NTK regime has no spurious local minima. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 21306–21328. PMLR, 2024. arXiv:2402.11867.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora, 2023. arXiv:2312.03732.

Junsu Kim, Jaeyeon Kim, and Ernest K. Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (But it probably won't fail). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 30224–30247. PMLR, 2025. arXiv:2502.09376.

Anastasis Kratsios, Tin Sum Cheng, Aurelien Lucchi, and Haitz Sáez de Ocáriz Borde. Sharp generalization bounds for foundation models with asymmetric randomized low-rank adapters, 2025. arXiv:2506.14530.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597, 2021.

Zi Liang, Haibo Hu, Qingqing Ye, Yaxin Xiao, and Ronghua Li. Does low rank adaptation lead to lower robustness against training-time attacks? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 37181–37207. PMLR, 2025. arXiv:2505.12871.

Siqiao Mu and Diego Klabjan. On the convergence rate of lora gradient descent, 2025. arXiv preprint arXiv:2512.18248.

Gibbs Nwemadji, Bruno Loureiro, and Jean Barbier. When pre-training hurts LoRA fine-tuning: A dynamical analysis via single-index models, 2026. arXiv:2602.02855.

Brady Steele. Why LoRA resists label noise: A theoretical framework for noise-robust parameter-efficient fine-tuning, 2026. arXiv:2602.00084.

Emanuele Troiani, Hugo Chao Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: high-dimensional asymptotics and sharp thresholds. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Matteo Vilucchio, Yatin Dandi, Cedric Gerbelot, and Florent Krzakala. Asymptotics of non-convex generalized linear models in high dimensions: A proof of the replica formula, 2025. arXiv:2502.20003.

Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. In *Conference on Learning Theory*, 2025a.

Ziqing Xu, Hancheng Min, Lachlan Ewen MacDonald, Jinqi Luo, Salma Tarmoun, Enrique Mallada, and Rene Vidal. Understanding the learning dynamics of lora: A gradient flow perspective on low-rank adaptation in matrix factorization. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 4636–4644. PMLR, 2025b. arXiv:2503.06982.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, 2024. arXiv:2310.17513.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 75513–75574. PMLR, 2025. arXiv:2502.01235.

A Derivation of the main results

In this appendix, we derive the equations of Section 3 characterizing pre-training and LoRA fine-tuning, explaining how the analyses of the two stages are combined into a joint description. We build on the analyses of Boncoraglio et al. (2026); Cui (2025), using established techniques from Gaussian equivalence, random matrix theory, replica theory and approximate message passing (AMP). The resulting predictions show excellent agreement with numerical simulations. A complete rigorous proof is beyond the scope of this work and would require a substantial technical development that closely follows standard arguments from high-dimensional inference. Results of Section 3 are derived via the following roadmap:

1. App. A.1 shows the reduction of the problem from a sequence-to-sequence to an attention-to-attention inference. We show that for attention matrices A and A^* and tokens X the sequence loss $||AX - A^*X||_F^2/D$ is equivalent to the attention loss $||A - A^*||_F^2$, based on the law of large numbers, i.e. on the fact that $\frac{XX^\top}{D} \approx I_T$ in the high-dimensional limit $D \to \infty$ with T fixed.
2. App. A.2 shows the Gaussian equivalent for the pre-activations on test samples. We show that on test samples the pre-activations corresponding to the extensive-rank component (z, z^*) and to the low-rank part (χ, χ^*) are distributed like Gaussian random variables, with independent laws $\mathcal{Q}$ and $\mathcal{Q}'$ that depend on the order parameters M, Q and m, q respectively. This gives Results 1 and 3. This step would require to generalize the arguments of Xu et al. (2025a) from $T = 1$ to multiple tokens and is outside the scope of the current work.
3. App. A.3 characterizes the pre-training stage. The asymptotic characterization of the extensive-rank pre-training is established by Boncoraglio et al. (2026). For completeness we reproduce it here. The result is based on an AMP algorithm whose fixed-points solve the optimization problem of the pre-training. The AMP iterations can be tracked by state-evolution equations on the order parameters M, Q, V. At convergence, this provides the distribution $\mathcal{R}$ of the pre-activations (z, z^*) on train samples in terms of the order parameters M, Q, V (Result 2). The values of M, Q, V are determined as a fixed-point of some state-evolution equations, or equivalently by an extremization problem (as shown in Result 4). The state-evolution equations are used for the numerical analysis.
4. App. A.4 explains how the extensive-rank pre-training and rank-one fine-tuning are coupled. This is a central step of our analysis. We justify why this coupling can be described through the pre-training order parameters M, Q, V, without introducing additional mixed order parameters between the matrix and vector components. The pre-trained weights remain fixed during fine-tuning, and their contribution enters through the pre-activations (z, z^*), with law $\mathcal{Q}$ for fresh inputs or $\mathcal{R}$ for reused inputs. In the effective description, these pre-activations are independent of the Gaussian pair $(\chi^*, \tilde{\chi})$ that describes the target projection and the input to the fine-tuning proximal. The learned pre-activation χ is then obtained by applying this proximal with (z, z^*) held fixed. Thus, pre-training affects the learned rank-one component through the fine-tuning loss, even though the two pairs entering the proximal are independent.
5. App. A.5 characterizes the fine-tuning stage. The asymptotic characterization of the training of a low-rank component, in the complete absence of an extensive-rank part, is established by Cui (2025), based on an AMP algorithm whose fixed-points solve the fine-tuning optimization of the low-rank part. We extend their results to our model, taking in account the effect of the pre-training step. We provide the distribution $\mathcal{R}'$ of the pre-activations (χ, χ^*) on train samples in terms of the order parameters m, q, v and M, Q, V (Results 2 and 3) and determine the values of m, q, v as the solution of an extremization problem (Result 5). We additionally provide the associated state-evolution equations, used for the numerical analysis. For completeness, in App. A.5.1 we rederive the full asymptotic characterization of the fine-tuning with pre-training, using the replica method under the replica-symmetric (RS) assumption.
6. App. A.6 states the replicon conditions. In this part we discuss the validity of our equations. The main non-trivial part is that the convergence of the AMP algorithm and the validity of the replica symmetric (RS) ansatz is a priori not guaranteed for non-convex losses. For a simpler problem, corresponding to the case $T = 1$, Vilucchio et al. (2025) rigorously shows

that for the overall procedure to be correct a so-called replicon condition must be satisfied. We consider two distinct AMPs and we need to verify two different conditions. The replicon condition for the pre-training is given in Boncoraglio et al. (2026), while for the fine-tuning we extend the one given in Vilucchio et al. (2025) to take into account the pre-training and the multiple tokens. We numerically checked that these conditions are well satisfied for all the values of the parameters we considered.

A.1 Reduction to an attention-output problem

We first reduce the sequence-to-sequence loss to a loss on the attention matrices. For a fixed matrix $B \in \mathbb{R}^{T\times T}$ and $X \sim \mathcal{P}_0$, one has

$$D^{-1}\|BX\|_F^2 = \operatorname{Tr}\left(B\frac{XX^\top}{D}B^\top\right) \xrightarrow[D\to\infty]{} \|B\|_F^2, \tag{27}$$

since $XX^\top/D \to I_T$ for fixed T. Hence the sequence-to-sequence square loss is asymptotically equivalent to the square loss between the corresponding attention matrices. For instance,

$$\begin{aligned} D^{-1}\,&\|\sigma(A-\mathbb{E}_X A)X - \sigma(A^*-\mathbb{E}_X A^*+\xi)X\|_F^2 \\ &= \|\sigma(A-\mathbb{E}_X A) - \sigma(A^*-\mathbb{E}_X A^*+\xi)\|_F^2 + o_D(1). \end{aligned} \tag{28}$$

We use this reduction throughout the derivation.

A.2 Reduction to a Gaussian test channel, test samples

We show that the pre-activations of the attention are Gaussian over the test samples. Let

$$S^* = P_0^{-1/2}W^*W^{*\top}, \qquad \hat{S} = P^{-1/2}\hat{W}\hat{W}^\top. \tag{29}$$

For a fresh test sequence X, independent of the training sets and of $\hat{W}, \hat{w}$, define

$$z^* = A^* - \mathbb{E}_X A^* + \xi, \qquad z = A - \mathbb{E}_X A, \qquad \chi^* = D^{-1/2}Xw^*, \qquad \chi = D^{-1/2}X\hat{w}, \tag{30}$$

with

$$A^* = D^{-1}XS^*X^\top, \qquad A = D^{-1}X\hat{S}X^\top. \tag{31}$$

Since $\mathbb{E}_X X_a^\top B X_b = \delta_{ab}\operatorname{Tr}B$ for any deterministic matrix B, the centering term is

$$(\mathbb{E}_X A^*)_{ab} = \delta_{ab}D^{-1}\operatorname{Tr}S^*, \qquad (\mathbb{E}_X A)_{ab} = \delta_{ab}D^{-1}\operatorname{Tr}\hat{S}. \tag{32}$$

Thus

$$z^*_{ab} = D^{-1}X_a^\top S^* X_b - \delta_{ab}D^{-1}\operatorname{Tr}S^* + \xi_{ab}, \qquad z_{ab} = D^{-1}X_a^\top \hat{S} X_b - \delta_{ab}D^{-1}\operatorname{Tr}\hat{S}. \tag{33}$$

The vector pre-activations are Gaussian by construction:

$$\begin{pmatrix}\chi_a^*\\ \chi_a\end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix}q_0 & m\\ m & q\end{pmatrix}\right), \qquad q_0 = D^{-1}w^{*\top}w^*, \quad q = D^{-1}\hat{w}^\top\hat{w}, \quad m = D^{-1}\hat{w}^\top w^*. \tag{34}$$

For the quadratic pre-activations, we use the Gaussian equivalence principle for centered quadratic measurements of Gaussian data, as in Boncoraglio et al. (2026); Xu et al. (2025a). For fixed T, consider the centered quadratic pre-activations

$$Z_{ab}(B) = \frac{X_a^\top B X_b - \delta_{ab}\operatorname{Tr}B}{D}, \qquad B \in \{S^*, \hat{S}\}, \quad 1 \le a \le b \le T. \tag{35}$$

These give $z_{ab} = Z_{ab}(\hat{S})$ and $z^*_{ab} = Z_{ab}(S^*) + \xi_{ab}$. Assuming $\|S^*\|_{\mathrm{op}}, \|\hat{S}\|_{\mathrm{op}} = O(\sqrt{D})$ and convergence of Q_0, Q, M, the variables $Z_{ab}(B)$ converge jointly to a centered Gaussian vector. For $B, C \in \{S^*, \hat{S}\}$, Wick contractions give

$$\operatorname{Cov}_X(Z_{ab}(B), Z_{cd}(C)) = \frac{\operatorname{Tr}(BC)}{D^2}\left(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}\right). \tag{36}$$

Thus distinct unordered token pairs are independent in the Gaussian limit, while diagonal pairs carry the factor $1 + \delta_{ab} = 2$. In particular,

$$D^{-2} \operatorname{Cov}_X \left(\operatorname{Tr} S^*(X_a X_b^\top - \delta_{ab} I_D), \operatorname{Tr} S^*(X_{a'} X_{b'}^\top - \delta_{a'b'} I_D) \right) = \delta_{aa'}\delta_{bb'}(1+\delta_{ab}) D^{-2} \operatorname{Tr}(S^*)^2 + o_D(1), \tag{37}$$

and similarly for $\hat{S}$ and the cross-covariance between S^* and $\hat{S}$. Including the additive noise gives, for $1 \le a \le b \le T$,

$$\begin{pmatrix} z^*_{ab} \\ z_{ab} \end{pmatrix} = \begin{pmatrix} z^*_{ba} \\ z_{ba} \end{pmatrix} \sim \mathcal{Q}_{ab} = \mathcal{N}\left(0, (1+\delta_{ab}) \begin{pmatrix} \Delta/2 + Q_0 & M \\ M & Q \end{pmatrix}\right), \tag{38}$$

where

$$Q_0 = D^{-2} \operatorname{Tr}(S^*)^2, \qquad Q = D^{-2} \operatorname{Tr} \hat{S}^2, \qquad M = D^{-2} \operatorname{Tr} \hat{S} S^*. \tag{39}$$

Equations (34) and (38) give Result 1. Combining them with the reduction (28) gives the test-error part of Result 3.

A.3 Pre-training train channel and state evolution

The pre-training stage is unchanged with respect to Boncoraglio et al. (2026). We recall only the objects needed in the present paper: the train law $\mathcal{R}$, the free entropy Φ, and the fixed-point equations used in the numerical implementation.

The train pre-activation is obtained from the proximal map

$$z = \operatorname{Prox}(z^*, \tilde{z}) = \operatorname*{arg\,max}_{h \in \mathcal{S}(T)} \Psi(h; z^*, \tilde{z}), \tag{40}$$

where $(z^*, \tilde{z}) \sim \mathcal{Q}$ and

$$\Psi(h; z^*, \tilde{z}) = -\|\sigma(h) - \sigma(z^*)\|_F^2 - \sum_{a \le b} \frac{(\tilde{z}_{ab} - h_{ab})^2}{2V(1+\delta_{ab})}. \tag{41}$$

The law $\mathcal{R}$ is defined by

$$\mathcal{R}: \qquad (z^*, \tilde{z}) \sim \mathcal{Q}, \qquad z = \operatorname{Prox}(z^*, \tilde{z}). \tag{42}$$

This is the train-channel statement in Result 2 for the pre-training variables.

Let μ_0 be the limiting spectral law of $S^*/\sqrt{D}$, let $\mu_{\mathrm{sc},\eta}$ be the semicircle law of radius 2η, set

$$\mu_\eta = \mu_0 \boxplus \mu_{\mathrm{sc},\eta}, \qquad J(\eta, \epsilon) = \int_\epsilon^{+\infty} \mu_\eta(\mathrm{d}x)(x-\epsilon)^2. \tag{43}$$

The pre-training free entropy is

$$\Phi(Q, M, V; \hat{M}, \hat{Q}, \hat{V}) = -\frac{1}{2\alpha} M\hat{M} + \frac{1}{4\alpha}(Q\hat{V} - V\hat{Q}) + \frac{\hat{M}^2}{4\alpha \hat{V}} J\left(\frac{\hat{Q}^{1/2}}{\hat{M}}, \frac{2\sqrt{\kappa}\lambda}{\hat{M}}\right) + \mathbb{E}_{z^*, \tilde{z}} \max_{h \in \mathcal{S}(T)} \Psi(h; z^*, \tilde{z}). \tag{44}$$

Its stable extremizer gives the order parameters (Q, M, V) in Result 4.

State-evolution equations. We now give the equivalent fixed-point equations used to solve this extremization numerically. Set

$$A_0 = Q_0 + \frac{\Delta}{2}, \qquad B = Q - \frac{M^2}{A_0}, \qquad s_{ab} = 1 + \delta_{ab}. \tag{45}$$

For $1 \le a \le b \le T$, let g_{ab}, ζ_{ab} be independent standard Gaussians and write

$$z^*_{ab} = \sqrt{s_{ab} A_0}\, g_{ab}, \qquad \tilde{z}_{ab} = \sqrt{s_{ab}} \left(\frac{M}{\sqrt{A_0}} g_{ab} + \sqrt{B}\, \zeta_{ab} \right). \tag{46}$$

Let $h^{\mathrm{prox}} = \mathrm{Prox}(z^*, \tilde{z})$ and define

$$\eta = \frac{\sqrt{\hat{Q}}}{\hat{M}}, \qquad \epsilon = \frac{2\sqrt{\kappa}\lambda}{\hat{M}}, \qquad J_1 = \partial_\eta J(\eta, \epsilon), \qquad J_2 = \partial_\epsilon J(\eta, \epsilon). \tag{47}$$

The output-channel equations are

$$\hat{M} = \frac{2\alpha}{V}\mathbb{E}\sum_{a\le b}\frac{h_{ab}^{\mathrm{prox}} - \tilde{z}_{ab}}{\sqrt{s_{ab}}}\left(\frac{g_{ab}}{\sqrt{A_0}} - \frac{M}{A_0\sqrt{B}}\zeta_{ab}\right), \tag{48}$$

$$\hat{Q} = \frac{2\alpha}{V^2}\mathbb{E}\sum_{a\le b}\frac{(h_{ab}^{\mathrm{prox}} - \tilde{z}_{ab})^2}{s_{ab}}, \tag{49}$$

$$\hat{V} = -\frac{2\alpha}{V\sqrt{B}}\mathbb{E}\sum_{a\le b}\frac{h_{ab}^{\mathrm{prox}} - \tilde{z}_{ab}}{\sqrt{s_{ab}}}\zeta_{ab}. \tag{50}$$

The expectations are over the Gaussians in (46); if $B = 0$, the equations are understood as the limit $B \downarrow 0$. The spectral input-channel equations are

$$Q = \frac{\hat{M}^2}{\hat{V}^2}J(\eta, \epsilon), \tag{51}$$

$$V = \frac{1}{2\hat{V}\eta}J_1, \tag{52}$$

$$M = \frac{\hat{M}}{2\hat{V}}\left[2J(\eta, \epsilon) - \eta J_1 - \epsilon J_2\right]. \tag{53}$$

Equations (48)–(53) are the pre-training state-evolution equations used in the simulations. At convergence, the pre-training test error is

$$\mathcal{E}(\hat{W}) = \mathbb{E}_{(z^*, z)\sim\mathcal{Q}}\|\sigma(z) - \sigma(z^*)\|_F^2. \tag{54}$$

A.4 Decoupling of the extensive-rank and rank-one components

We now explain why the LoRA vector w learned during fine-tuning is described, in the high-dimensional limit, only by its overlap with the rank-one teacher w^*, and not by additional overlaps with the extensive-rank matrices S^* and $\hat{S}$.

Conditionally on the pre-training stage, S^* and $\hat{S}$ are frozen matrices with normalized traces of order one. In the Marchenko–Pastur setting considered here, their operator norms are of order $\sqrt{D}$ and their squared Frobenius norms are of order D^2.

Let B denote either S^* or $\hat{S}$. Let $u, v \in \mathbb{R}^D$ be deterministic vectors with $\|u\|_2, \|v\|_2 = O(\sqrt{D})$, independent of a fresh Gaussian sequence X. A direct Wick contraction over X gives, for any token indices a, b, c, d,

$$\begin{aligned}\mathrm{Cov}_X\left(D^{-1}X_a^\top B X_b,\ D^{-1/2}X_c^\top u\, D^{-1/2}X_d^\top v\right)\\ = D^{-2}\left[\delta_{ac}\delta_{bd}\, u^\top B v + \delta_{ad}\delta_{bc}\, v^\top B u\right].\end{aligned} \tag{55}$$

Since $\|B\|_{\mathrm{op}} = O(\sqrt{D})$ and $\|u\|_2, \|v\|_2 = O(\sqrt{D})$,

$$D^{-2}|u^\top B v| \le D^{-2}\|B\|_{\mathrm{op}}\|u\|_2\|v\|_2 = O(D^{-1/2}). \tag{56}$$

Thus the mixed covariance (55) vanishes as $D \to \infty$. Together with their joint Gaussian limit, this implies that the extensive-rank quadratic pre-activations (z^*, z) (in the pre-training) and the rank-one pre-activations (χ^*, χ) (in the fine-tuning) are asymptotically independent on fresh test inputs. The covariances within each pair are determined by (Q_0, Q, M) and (q_0, q, m), respectively.

This also rules out additional order parameters involving both w and the extensive-rank matrices. For example, for any vector w with $\|w\|_2 = O(\sqrt{D})$,

$$D^{-2}|w^\top S^* w| \le D^{-2}\|S^*\|_{\mathrm{op}}\|w\|_2^2 = O(D^{-1/2}), \tag{57}$$

and similarly for $\hat{S}$. Hence quantities such as

$$D^{-2}w^\top S^* w, \qquad D^{-2}w^\top \hat{S} w, \qquad D^{-2}w^{*\top}\hat{S} w \tag{58}$$

vanish in the normalization relevant for the extensive-rank pre-activations, and no order parameters mixing the pre-training with the fine-tuning stages appear in our analysis. The rank-one LoRA component can therefore learn the direction w^* at the scale $N' = \Theta(D)$, but it cannot learn the extensive-rank component, which remains a frozen Gaussian side field. The two stages of pre-training and fine-tuning are therefore coupled and interact at the level of the two nested empirical risk minimization processes, but not at the level of explicit mixed order parameters.

Consequently, conditionally on the pre-training order parameters (Q, M, V), the fine-tuning problem reduces to a rank-one generalized linear model with ground-truth vector w^*, trainable fine-tuning vector w, and sample-wise side information coming from the pre-training (z^*, z). For fresh fine-tuning samples, (z^*, z) has the test law $\mathcal{Q}$. For reused fine-tuning samples, we use the pre-training law $\mathcal{R}$ in the effective description. The fresh-input argument above does not by itself establish this training-channel decoupling. Here independence concerns (z^*, z) and the Gaussian pair $(\chi^*, \tilde{\chi})$ entering the fine-tuning proximal; the fitted χ depends on both pairs.

A.5 Fine-tuning free entropy and state evolution

We now turn to the rank-one fine-tuning stage. Without the frozen fields (z^*, z), this is a rank-one sequence multi-index ERM problem. We use the corresponding high-dimensional characterization of Cui et al. (2024); Cui (2025); Duranthon et al. (2026), with the modification that the output channel is conditioned on the pre-training fields (z^*, z). By the decoupling argument of Appendix A.4, these fields enter as side information and no additional overlaps with S^* or $\hat{S}$ are needed.

Conditionally on the pre-training stage,

$$(z^*, z) \sim \begin{cases} \mathcal{Q}, & \mathrm{e} = 0, \\ \mathcal{R}, & \mathrm{e} = 1. \end{cases} \tag{59}$$

The rank-one channel is parametrized by

$$\chi^* \sim \mathcal{N}(0, I_T), \qquad \zeta \sim \mathcal{N}(0, I_T), \qquad \tilde{\chi} = m\chi^* + \sqrt{q - m^2}\,\zeta, \tag{60}$$

with χ^* and ζ independent. For fixed $(\chi^*, \tilde{\chi}, z^*, z)$, define

$$\psi(\chi; \chi^*, \tilde{\chi}, z^*, z) = -\left\|\sigma\left(z + \chi\chi^\top - qI_T\right) - \sigma\left(z^* + \chi^*\chi^{*\top} - I_T\right)\right\|_F^2 - \sum_{a=1}^{T} \frac{(\tilde{\chi}_a - \chi_a)^2}{2v}, \tag{61}$$

and

$$\chi^{\mathrm{prox}} = \mathrm{prox}(\chi^*, \tilde{\chi}, z^*, z) = \arg\max_{\chi \in \mathbb{R}^T} \psi(\chi; \chi^*, \tilde{\chi}, z^*, z). \tag{62}$$

The fine-tuning free entropy is

$$\begin{aligned} \phi(m, q, v; \hat{m}, \hat{q}, \hat{v}) = -\frac{1}{\alpha'} m\hat{m} + \frac{1}{2\alpha'}(q\hat{v} - v\hat{q}) + \frac{\hat{m}^2 + \hat{q}}{2\alpha'(2\lambda' + \hat{v})} \\ + \mathbb{E}_{\chi^*, \zeta, z^*, z} \max_{\chi \in \mathbb{R}^T} \psi(\chi; \chi^*, \tilde{\chi}, z^*, z), \end{aligned} \tag{63}$$

which is the formula stated in Result 5. The three terms in the free entropy above have the standard generalized-linear-model interpretation. The first two terms enforce the order parameters q, m, v, the rational term comes from integrating out the microscopic vector w under the quadratic regularization $\lambda' \|w\|_2^2$, and the last term is the effective output-channel contribution. The only difference with a standard rank-one GLM is that the output channel depends on the frozen pre-training field (z^*, z).

State-evolution equations. Extremizing (63) with respect to $(\hat{m}, \hat{q}, \hat{v})$ gives the following stationarity equations used in the numerical implementation. Let

$$C(\chi^*, \tilde{\chi}, z^*, z) = -\left[\nabla_\chi^2 \psi(\chi^{\mathrm{prox}}; \chi^*, \tilde{\chi}, z^*, z)\right]^{-1}. \tag{64}$$

Then the fixed point is

$$m = \hat{m}v, \tag{65}$$

$$q = (\hat{m}^2 + \hat{q})v^2, \tag{66}$$

$$v = \frac{1}{2\lambda' + \hat{v}}, \tag{67}$$

$$\hat{m} = \frac{\alpha'}{v}\mathbb{E}\left[\chi^{*\top}\chi^{\mathrm{prox}} - \frac{m}{v}\operatorname{Tr} C(\chi^*, \tilde{\chi}, z^*, z)\right], \tag{68}$$

$$\hat{q} = \frac{\alpha'}{v^2}\mathbb{E}\left[\|\chi^{\mathrm{prox}} - \tilde{\chi}\|_2^2\right], \tag{69}$$

$$\hat{v} = \frac{\alpha' T}{v} - \frac{\alpha'}{v^2}\mathbb{E}\left[\operatorname{Tr} C(\chi^*, \tilde{\chi}, z^*, z)\right] - 2\alpha'\mathbb{E}\left[\partial_q \psi(\chi^{\mathrm{prox}}; \chi^*, \tilde{\chi}, z^*, z)\right]. \tag{70}$$

All expectations are taken over (59) and (60). The term $\partial_q \psi$ differentiates only the explicit dependence of (61) on the shift $-qI_T$; the dependence of χ^{prox} on q is not differentiated, by the envelope theorem.

At a fixed point, the fine-tuning test error is

$$\mathcal{E}'(\hat{W}, \hat{w}) = \mathbb{E}_{(z^*,z)\sim\mathcal{Q},\ (\chi^*,\chi)\sim\mathcal{Q}'}\left\|\sigma\left(z + \chi\chi^\top - qI_T\right) - \sigma\left(z^* + \chi^*\chi^{*\top} - I_T\right)\right\|_F^2 \tag{71}$$

and the overlap is given by $o_w = \frac{m}{\sqrt{q}}$. For softmax activation, the proximal map (62) is solved numerically for each Monte Carlo sample. In the next subsection we show the detailed replica derivation of the results presented in this section.

A.5.1 Replica derivation of the fine-tuning equations

We derive the fine-tuning potential (63) from the effective description in Appendix A.4, following the replica approach of Cui (2025). The pre-training order parameters (Q, M, V) are held fixed throughout. For each fine-tuning sample, (z_μ^*, z_μ) follows the law in (59): $\mathcal{Q}$ for fresh inputs and $\mathcal{R}$ for reused inputs. The fine-tuning sequences $X'_\mu \in \mathbb{R}^{T\times D}$ have independent standard Gaussian entries and are independent of w^*. In the replica calculation, we use the decoupling described in Appendix A.4 to average their rank-one projections independently of the pre-training pre-activations (z_μ^*, z_μ).

Effective objective and partition function. For a candidate vector w, write $r = \|w\|_2^2/D$. The loss of one sequence is

$$\ell_r(\chi; \chi^*, z^*, z) = \left\|\sigma(z + \chi\chi^\top - rI_T) - \sigma(z^* + \chi^*\chi^{*\top} - I_T)\right\|_F^2, \tag{72}$$

where we use the limiting target norm $q_0 = 1$. The shift $-rI_T$ is the population centering used in the effective equations. It varies with w and must therefore be included when differentiating the objective. The effective regularized loss is

$$H_D(w) = \sum_{\mu=1}^{N'} \ell_r\left(\frac{X'_\mu w}{\sqrt{D}}; \frac{X'_\mu w^*}{\sqrt{D}}, z_\mu^*, z_\mu\right) + \lambda'\|w\|_2^2. \tag{73}$$

Both contributions are of order D when $N'/D \to \alpha'$. The pre-training pre-activations (z_μ^*, z_μ) remain fixed during minimization; z_μ^* enters through the target and need not be observed separately.

Since $H_D(w) = H_D(-w)$, we choose the sign convention $w^\top w^* \geq 0$, corresponding to $m \geq 0$. Restricting the partition function to this half-space divides it by two and does not change its limiting normalized logarithm. For $\lambda' > 0$ and inverse temperature $b > 0$, define

$$Z_b = \int_{w^\top w^* \geq 0} \mathrm{d}w\, e^{-bH_D(w)}, \qquad f_b = \lim_{D\to\infty} \frac{1}{bD}\mathbb{E}\log Z_b. \tag{74}$$

We use b to distinguish inverse temperature from the recalibration parameter β in the main text.

Replica overlaps and averaging over samples. For an integer $n \geq 1$, introduce replicas $w^1, \dots, w^n$ and define

$$Q_{ij} = \frac{(w^i)^\top w^j}{D}, \qquad m_i = \frac{(w^i)^\top w^*}{D}, \qquad \mathbf{m} = (m_1, \dots, m_n)^\top. \tag{75}$$

For a fixed sample μ and token a, define the projections

$$\chi_a^* = \frac{(X'_\mu w^*)_a}{\sqrt{D}}, \qquad \chi_a^i = \frac{(X'_\mu w^i)_a}{\sqrt{D}}, \quad i = 1, \dots, n. \tag{76}$$

At fixed replica vectors, these projections are jointly Gaussian:

$$\begin{pmatrix} \chi_a^* \\ \chi_a^1 \\ \vdots \\ \chi_a^n \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \mathbf{m}^\top \\ \mathbf{m} & Q \end{pmatrix}\right). \tag{77}$$

Different tokens have independent projections before applying the sequence loss. The contribution of one sample to $\mathbb{E}Z_b^n$ is therefore

$$\mathcal{I}_n(Q, \mathbf{m}) = \mathbb{E}_{z^*, z, \chi^*, \chi^1, \dots, \chi^n} \exp\left\{-b \sum_{i=1}^n \ell_{Q_{ii}}(\chi^i; \chi^*, z^*, z)\right\}. \tag{78}$$

The same (z^*, z) and χ^* enter every replica, since all replicas are fitted to the same dataset. This retains the pre-training contribution inside the fine-tuning calculation.

After averaging over the inputs, the loss contribution depends on the replica vectors only through their overlaps. We must also account for the entropic factor associated with all replica configurations sharing these overlaps. To compute it, decompose $w^i = m_i w^* + u^i$, with $(u^i)^\top w^* = 0$, using $\|w^*\|_2^2/D \to 1$. The orthogonal components have Gram matrix $Q - \mathbf{m}\mathbf{m}^\top$, giving the entropy per dimension

$$s_{\text{ent}}(Q, \mathbf{m}) = \frac{1}{2} \log\det(\mathtt{Q} - \mathbf{m}\mathbf{m}^\top) + \frac{n}{2} \log(2\pi e) + o_D(1). \tag{79}$$

Combining this volume with the regularizer and the N' sample factors gives the large-D exponent

$$\frac{1}{2} \log\det(Q - \mathbf{m}\mathbf{m}^\top) - b\lambda' \operatorname{Tr} Q + \alpha' \log \mathcal{I}_n(Q, \mathbf{m}) + \frac{n}{2} \log(2\pi e) \tag{80}$$

in the overlap integral for $\mathbb{E}Z_b^n$.

Replica-symmetric ansatz. A standard ansatz in the context of statistical physics of machine learning calculations is to assume the order parameters of the model have a replica symmetric structure, namely

$$Q_{ii} = r, \qquad Q_{ij} = q \quad (i \neq j), \qquad m_i = m. \tag{81}$$

At finite temperature, r is the squared norm of one replica, whereas q is the overlap between two replicas. For $r > q \geq m^2$, the projections in equation (77) can be represented as

$$\chi^i = m\chi^* + \sqrt{q - m^2}\,\zeta + \sqrt{r - q}\,\eta^i, \qquad \chi^*, \zeta, \eta^1, \dots, \eta^n \sim \mathcal{N}(0, I_T), \tag{82}$$

where all Gaussian vectors are independent and independent of (z^*, z). The variables χ^* and ζ are shared by the replicas, while η^i varies between replicas. Writing $\tilde{\chi} = m\chi^* + \sqrt{q - m^2}\,\zeta$, define

$$\mathcal{Z}_{\text{out}} = \int \mathcal{D}\eta \, \exp\left\{-b\ell_r(\tilde{\chi} + \sqrt{r - q}\,\eta; \chi^*, z^*, z)\right\}, \qquad \mathcal{D}\eta = \frac{e^{-\|\eta\|_2^2/2}}{(2\pi)^{T/2}} \, \mathrm{d}\eta. \tag{83}$$

Then, from equation (78), we get $\mathcal{I}_n = \mathbb{E}[\mathcal{Z}_{\text{out}}^n]$, and using the notable replica trick we get the final expression that allows us to compute the finite-temperature free-entropy

$$\lim_{n \to 0} \frac{1}{n} \log \mathcal{I}_n = \mathbb{E} \log \mathcal{Z}_{\text{out}}. \tag{84}$$

In particular, the average over (z^*, z) is outside the logarithm: fine-tuning is performed for a fixed pre-training outcome before averaging over datasets.

The residual Gram determinant gives

$$\det(Q - \mathbf{m}\mathbf{m}^\top) = (r-q)^{n-1}\big[r - q + n(q - m^2)\big], \tag{85}$$

$$\lim_{n\to 0} \frac{1}{n} \log\det(Q - \mathbf{m}\mathbf{m}^\top) = \log(r-q) + \frac{q-m^2}{r-q}. \tag{86}$$

Applying the replica identity

$$f_b = \lim_{n\to 0} \lim_{D\to\infty} \frac{1}{bnD} \log \mathbb{E} Z_b^n \tag{87}$$

therefore yields, in the limit of zero number of replicas

$$f_b = \mathrm{stat}_{r,q,m} \left\{ -\lambda' r + \frac{\log(r-q) + (q-m^2)/(r-q) + \log(2\pi e)}{2b} + \frac{\alpha'}{b} \mathbb{E} \log \mathcal{Z}_{\mathrm{out}} \right\}. \tag{88}$$

Here stat denotes evaluation at the physical stationary saddle of the replica-symmetric calculation.

Zero-temperature limit. We take $b \to \infty$ with the rescaling

$$r = q + \frac{v}{b}, \qquad 0 < v < \infty. \tag{89}$$

Changing variables from η to χ in the sample integral gives

$$\mathcal{Z}_{\mathrm{out}} = \left(\frac{b}{2\pi v}\right)^{T/2} \int \mathrm{d}\chi \, \exp\left\{ -b \left[\ell_{q+v/b}(\chi; \chi^*, z^*, z) + \frac{\|\chi - \tilde{\chi}\|_2^2}{2v} \right] \right\}. \tag{90}$$

By the Laplace approximation,

$$\begin{aligned} \lim_{b\to\infty} \frac{1}{b} \log \mathcal{Z}_{\mathrm{out}} &= \max_{\chi \in \mathbb{R}^T} \left\{ -\ell_q(\chi; \chi^*, z^*, z) - \frac{\|\chi - \tilde{\chi}\|_2^2}{2v} \right\} \\ &= \max_{\chi \in \mathbb{R}^T} \psi(\chi; \chi^*, \tilde{\chi}, z^*, z), \end{aligned} \tag{91}$$

with ψ defined in (61). The quadratic proximal penalty thus comes from the Gaussian fluctuations between replicas. Taking the same limit in (88) gives

$$f_\infty = \mathrm{stat}_{q,m,v} \left\{ -\lambda' q + \frac{q-m^2}{2v} + \alpha' \mathbb{E} \max_{\chi \in \mathbb{R}^T} \psi(\chi; \chi^*, \tilde{\chi}, z^*, z) \right\}. \tag{92}$$

The replica norm r and overlap q now have the same limit, equal to the order parameter q in the main text. With our convention, f_∞ is minus the limiting minimum of H_D/D, including regularization.

Recovery of the fine-tuning potential. Introduce the conjugate parameters through

$$-\lambda' q + \frac{q-m^2}{2v} = \mathrm{stat}_{\hat{m},\hat{q},\hat{v}} \left\{ -m\hat{m} + \frac{q\hat{v} - v\hat{q}}{2} + \frac{\hat{m}^2 + \hat{q}}{2(2\lambda' + \hat{v})} \right\}, \tag{93}$$

with $\hat{q} \geq 0$ and $2\lambda' + \hat{v} > 0$. Indeed, differentiation with respect to the conjugate parameters gives

$$v = \frac{1}{2\lambda' + \hat{v}}, \qquad m = \hat{m} v, \qquad q = (\hat{m}^2 + \hat{q}) v^2, \tag{94}$$

and substituting these relations recovers the left-hand side.

The rational term also follows directly from the scalar Gaussian integral associated with the quadratic regularizer. For independent $U, Z \sim \mathcal{N}(0,1)$,

$$\begin{aligned} \lim_{b\to\infty} \frac{1}{b} \mathbb{E}_{U,Z} \log \int_{\mathbb{R}} \mathrm{d}t \, \exp\left\{ b \left[-\frac{2\lambda' + \hat{v}}{2} t^2 + (\hat{m} U + \sqrt{\hat{q}}\, Z) t \right] \right\} \\ = \frac{\hat{m}^2 + \hat{q}}{2(2\lambda' + \hat{v})}. \end{aligned} \tag{95}$$

Using (93) and dividing the potential in (92) by α' gives

$$\phi = -\frac{m\hat{m}}{\alpha'} + \frac{q\hat{v} - v\hat{q}}{2\alpha'} + \frac{\hat{m}^2 + \hat{q}}{2\alpha'(2\lambda' + \hat{v})} + \mathbb{E} \max_{\chi \in \mathbb{R}^T} \psi(\chi; \chi^*, \tilde{\chi}, z^*, z). \tag{96}$$

This is precisely the equation (63) and equation (21) presented in the main text, giving the replica-symmetric characterization of our fine-tuning model after frozen pre-training. Its stationary value is f_∞/α', so the normalization is per fine-tuning sample. The factor $2\lambda'$ follows from the regularizer $\lambda' \|w\|_2^2$.

Derivation of the fixed-point equations. For completeness, define

$$F(m, q, v) = \mathbb{E} \max_{\chi} \psi(\chi; \chi^*, \tilde{\chi}, z^*, z), \qquad d = \chi^{\text{prox}} - \tilde{\chi}, \qquad s = \sqrt{q - m^2}. \tag{97}$$

The pre-training law is held fixed in these derivatives. On a regular proximal branch, the envelope theorem and implicit differentiation give

$$\nabla_{\tilde{\chi}} \max_{\chi} \psi = \frac{d}{v}, \qquad \frac{\partial \chi^{\text{prox}}}{\partial \tilde{\chi}} = \frac{C}{v}, \tag{98}$$

where C is defined in (64). Since $\tilde{\chi} = m\chi^* + s\zeta$, Gaussian integration by parts in ζ yields

$$\mathbb{E}[\zeta^\top d] = s \left(\frac{\mathbb{E} \operatorname{Tr} C}{v} - T \right). \tag{99}$$

Combining this identity with the derivatives of $\tilde{\chi}$ gives

$$\partial_m F = \frac{1}{v} \mathbb{E}[\chi^{*\top} \chi^{\text{prox}}] - \frac{m}{v^2} \mathbb{E} \operatorname{Tr} C, \tag{100}$$

$$\partial_q F = \mathbb{E}[\partial_q \psi] + \frac{1}{2v^2} \mathbb{E} \operatorname{Tr} C - \frac{T}{2v}, \tag{101}$$

$$\partial_v F = \frac{1}{2v^2} \mathbb{E} \|d\|_2^2. \tag{102}$$

Here $\partial_q \psi$ denotes only the explicit derivative through the shift $-qI_T$ in the loss. The contribution from the q-dependence of $\tilde{\chi}$ is already included in the other terms of $\partial_q F$. These expressions are obtained for $q > m^2$; the boundary is treated by a continuous limit when it exists.

Stationarity of ϕ with respect to (m, q, v) gives

$$\hat{m} = \alpha' \partial_m F, \qquad \hat{v} = -2\alpha' \partial_q F, \qquad \hat{q} = 2\alpha' \partial_v F. \tag{103}$$

Substituting (102) recovers Equations (68)–(70). Together with (94), these are the six fixed-point equations of Appendix A.5.

Training and test pre-activations. For fixed $(\chi^*, \tilde{\chi}, z^*, z)$, the normalized integrand in (90) gives the effective training distribution of χ. As $b \to \infty$, it concentrates on the proximal optimizer χ^{prox} when this optimizer is unique. Thus, drawing (z^*, z) according to (59), independently drawing $(\chi^*, \tilde{\chi}) \sim \mathcal{Q}'$, and setting

$$\chi = \text{prox}(\chi^*, \tilde{\chi}, z^*, z) \tag{104}$$

recovers the training law $\mathcal{R}'$ in Result 2.

For a fresh test sequence, the pre-activations instead follow the independent laws $(z^*, z) \sim \mathcal{Q}$ and $(\chi^*, \chi) \sim \mathcal{Q}'$ at the learned order parameters, without applying a proximal map. This gives the test error in Result 3, including when the training sequences were reused. The reconstruction overlap is $o_w = |m|/\sqrt{q}$, or $m/\sqrt{q}$ with the sign convention adopted above.

A.6 Replicon stability conditions

Below we report the replicon conditions associated with the pre-training and fine-tuning characterizations. These conditions are necessary for our replica-symmetric predictions to be correct.

For the extensive-rank pre-training stage, the condition is the one derived in Boncoraglio et al. (2026); Erba et al. (2025). Let $(z^*, \tilde{z}) \sim \mathcal{Q}$, let $u_{ab} = z_{ab}/\sqrt{1 + \delta_{ab}}$ for $a \leq b$, and let H_u be the Hessian of

$\Psi(z(u); z^*, \tilde{z})$ with respect to these independent coordinates, evaluated at the proximal optimizer, with $L = T(T+1)/2$, and write

$$\mathrm{Prox} = \mathrm{Prox}(z^*, \tilde{z}). \tag{105}$$

Then the pre-training replicon condition reads

$$\frac{2\alpha}{V^2}\,\mathbb{E}\left\|(VH_u)^{-1} + I_L\right\|_F^2 \int \mu_{\hat{Q}^{1/2}/\hat{M}}(\mathrm{d}x)\mu_{\hat{Q}^{1/2}/\hat{M}}(\mathrm{d}y)\frac{(\zeta(x)-\zeta(y))^2}{\hat{V}^2(x-y)^2} < 1, \tag{106}$$

where

$$\zeta(x) = \left(x - \frac{2\sqrt{\kappa}\lambda}{\hat{M}}\right)_+ . \tag{107}$$

For rank-one fine-tuning, the local replicon criterion is as follows. Let $(z^*, z) \sim \mathcal{Q}$ if $\mathrm{e} = 0$ and $(z^*, z) \sim \mathcal{R}$ if $\mathrm{e} = 1$, let $(\chi^*, \tilde{\chi})$ be distributed as in (60), and write

$$\mathrm{prox} = \mathrm{prox}(\chi^*, \tilde{\chi}, z^*, z). \tag{108}$$

Then the fine-tuning replicon condition is

$$\alpha'\,\mathbb{E}\,\left\|(v\nabla^2\psi(\mathrm{prox}; \chi^*, \tilde{\chi}, z^*, z))^{-1} + I_T\right\|_F^2 < 1\,. \tag{109}$$

The derivation is done by considering the AMP algorithm associated to the distribution on w induced by the fine-tuning loss $\mathcal{L}'$. The AMP is stated in Cui (2025), for a generic class of sequence multi-index models. We specialize it to our case for the indices χ, χ^* and take into account the supplementary noise z, z^* from the pre-training. The replicon condition is then obtained by considering a linear perturbation of the AMP around its fixed-point and checking whether it is amplified or not. The computation follows steps similar to Erba et al. (2025). We stress that satisfaction of the replicon condition is only a necessary, and not by itself sufficient, condition for the RS characterization to correctly describe the global minima of a non-convex problem. As discussed in Vilucchio et al. (2025), a rigorous justification additionally requires suitable regularity assumptions on the loss and the associated proximal maps. We therefore regard the replicon condition here as a stability diagnostic rather than a proof of RS exactness. Nevertheless, its satisfaction throughout all the parameter regimes we investigate, together with the excellent agreement between the RS predictions and finite-dimensional simulations across the different settings considered, provides strong evidence that the characterization is sharp in the regimes studied.

A.7 Limiting baselines used in the figures

We briefly record the two limiting baselines shown in the numerical figures.

First, when $\lambda' \to +\infty$, the LoRA update is suppressed and $\hat{w} = 0$. Equivalently, $m, q, v \to 0$ in the fine-tuning state evolution. The fine-tuning test error reduces to the frozen-pretraining baseline

$$\mathcal{E}'_{\mathrm{frozen}}(\hat{W}) = \mathbb{E}_{(z^*,z)\sim\mathcal{Q},\ \chi^*\sim\mathcal{N}(0,I_T)}\left\|\sigma(z) - \sigma\left(z^* + \chi^*\chi^{*\top} - I_T\right)\right\|_F^2 . \tag{110}$$

For the rescaled identity activation $\sigma(x) = x/T$, this becomes

$$\mathcal{E}'_{\mathrm{frozen}}(\hat{W}) = \frac{(T^2+T)\Delta_{\mathrm{eff}} + T^2 + T}{T^2}, \qquad \Delta_{\mathrm{eff}} = \frac{\Delta}{2} + Q_0 + Q - 2M. \tag{111}$$

Indeed, $\mathbb{E}\|z - z^*\|_F^2 = (T^2+T)\Delta_{\mathrm{eff}}$ and $\mathbb{E}\|\chi^*\chi^{*\top} - I_T\|_F^2 = T^2 + T$; the cross term vanishes by independence and centering. Second, when $\lambda \to +\infty$, the pre-trained extensive-rank component is suppressed and $\hat{W} = 0$. Thus $Q = M = V = 0$, $z = 0$, and

$$z^*_{ab} \sim \mathcal{N}\left(0, (1+\delta_{ab})\left(Q_0 + \frac{\Delta}{2}\right)\right). \tag{112}$$

The fine-tuning state evolution is therefore (65)–(70) with $z = 0$ and z^* distributed as in (112). The corresponding LoRA-only test error is

$$\mathcal{E}'_{\mathrm{LoRA-only}} = \mathbb{E}_{z^*,\chi^*,\chi}\left\|\sigma\left(\chi\chi^\top - qI_T\right) - \sigma\left(z^* + \chi^*\chi^{*\top} - I_T\right)\right\|_F^2 , \tag{113}$$

with $\chi^* \sim \mathcal{N}(0, I_T)$, $\chi \sim \mathcal{N}(0, qI_T)$ and $\mathbb{E}[\chi^*\chi^\top] = mI_T$, evaluated at the reduced fixed point.

A.8 CATASTROPHIC FORGETTING: PERFORMANCES OF THE FINE-TUNED MODEL ON THE PRE-TRAINING DATA

One can evaluate the performances of the fine-tuned model on the original pre-training loss. Considering a new test pre-training sample X, y, it reads

$$\mathcal{E}''(\hat{W}, \hat{w}) = \mathbb{E}_{(X,y)\sim\mathcal{P}} \left[D^{-1} ||\hat{y}(X; \hat{W}, \hat{w}) - y||_F^2 \,\middle|\, \mathcal{S}, \mathcal{S}' \right] \tag{114}$$

In the frame of our theory, this quantity admits an asymptotic characterization, similarly to Result 3. Taking $(z^*, z) \sim \mathcal{Q}$ and $(\chi^*, \chi) \sim \mathcal{Q}'$ we have

$$\mathcal{E}''(\hat{W}, \hat{w}) = \mathbb{E}\, ||\sigma(z + \chi\chi^\top - qI_T) - \sigma(z^*)||_F^2 \,. \tag{115}$$

We expect this error $\mathcal{E}''$ to be higher than the pre-training test error $\mathcal{E}$ because of the supplementary learned term $\chi\chi^\top - qI_T$ that acts like noise. This is reminiscent of the catastrophic forgetting phenomenon.

B BAYES-OPTIMAL PERFORMANCES

In this appendix we give the Bayes-optimal benchmarks used in the main text. The extensive-rank pre-training BO benchmark is the single-layer AIM result of Boncoraglio et al. (2025). The fine-tuning BO benchmark follows the Bayes-optimal theory for sequence multi-index models Troiani et al. (2025), applied to the rank-one LoRA component and conditioned on the effective pre-training channel. This combination is justified by the same decoupling mechanism as in Appendix A.4: with only $N' = \Theta(D)$ fine-tuning samples, it is information-theoretically impossible to recover a finite fraction of the extensive-rank matrices W^* or $\hat{W}$, whose estimation requires a quadratic number of samples. Thus the fine-tuning BO problem only estimates the rank-one direction w^*, while the extensive-rank part enters as a Gaussian side field determined by the BO pre-training overlap.

B.1 BAYES-OPTIMAL PRE-TRAINING FOR THE EXTENSIVE-RANK AIM

We recall the single-layer BO characterization of Boncoraglio et al. (2025) in the notation of the present paper. Let $Q_{\rm BO} \in [0, Q_0]$ denote the BO overlap with the extensive-rank target S^*, and let $\hat{Q}_{\rm BO} \geq 0$ be its conjugate. For a general output channel, the BO free entropy has the form

$$\Phi_{\rm BO}(Q, \hat{Q}) = -\frac{Q\hat{Q}}{4} + I_{\rm in}(\hat{Q}) + \alpha I_{\rm out}(Q), \tag{116}$$

and the BO overlap is obtained from the corresponding stable saddle point

$$(Q_{\rm BO}, \hat{Q}_{\rm BO}) = \underset{Q\in[0,Q_0],\, \hat{Q}\geq 0}{\rm extr} \Phi_{\rm BO}(Q, \hat{Q}). \tag{117}$$

For the rotationally invariant prior considered here, the input-channel equation can be written as

$$Q_{\rm BO} = Q_0 - \frac{1}{\hat{Q}_{\rm BO}} + \frac{4\pi^2}{3\hat{Q}_{\rm BO}^2} \int \mu_{1/\hat{Q}_{\rm BO}}(t)^3 \, {\rm d}t, \tag{118}$$

where $\mu_{1/\hat{Q}_{\rm BO}}$ is the limiting spectral density of the noisy matrix channel

$$Y = S^* + \hat{Q}_{\rm BO}^{-1/2} Z, \qquad Z \sim {\rm GOE}(D). \tag{119}$$

For the softmax output channel, the output equation is explicit:

$$\hat{Q}_{\rm BO} = \alpha \, \frac{T^2 + T - 2}{Q_0 + \Delta/2 - Q_{\rm BO}}. \tag{120}$$

A symmetric logit matrix has $L = T(T+1)/2$ independent coordinates. Row-wise softmax on this subspace is invariant under the common shift $s\mathbf{1}_T\mathbf{1}_T^\top$, leaving $L-1$ informative directions. Hence the coefficient is $2(L-1) = T^2 + T - 2$, unlike the linear frozen baseline.

The associated effective test channel is, for $a \leq b$

$$\left(\begin{smallmatrix} z^*_{ab} \\ z_{ab} \end{smallmatrix}\right) \sim \mathcal{N}\left(0, (1+\delta_{ab}) \left(\begin{smallmatrix} Q_0+\Delta/2 & Q_{\rm BO} \\ Q_{\rm BO} & Q_{\rm BO} \end{smallmatrix}\right)\right) \tag{121}$$

where $z^*, z \in \mathcal{S}(T)$ are symmetrized. The BO pre-training test error is

$$\mathcal{E}_{\rm BO} = \mathbb{E}\, \|\sigma(z) - \sigma(z^*)\|_F^2 \,. \tag{122}$$

B.2 Bayes-optimal fine-tuning for the rank-one LoRA component

We now state the fine-tuning BO benchmark, following Troiani et al. (2025). Once $Q_{\rm BO}$ has been determined, the part of the extensive-rank teacher that is not known after pre-training is represented by the Gaussian residual in (121). The fine-tuning data are then informative only about the rank-one vector w^*, while the extensive-rank component remains a side noise (z^*, z) with law given by eq. 121 for $\mathrm{e} = 0$.

We first consider the case $\mathrm{e} = 0$. Let $q_{\rm BO} \in [0, q_0]$ denote the overlap of the BO estimator with the fine-tuning target vector w^*. Its input-channel equation is

$$q_{\rm BO} = \frac{\hat{q}_{\rm BO}}{1 + \hat{q}_{\rm BO}}, \tag{123}$$

where $\hat{q}_{\rm BO} \geq 0$ is the conjugate overlap.

To write the output-channel equation, let be

$$\left(\begin{smallmatrix} \chi^*_a \\ \chi_a \end{smallmatrix}\right) \sim \mathcal{N}\left(0, \left(\begin{smallmatrix} q_0 & q_{\rm BO} \\ q_{\rm BO} & q_{\rm BO} \end{smallmatrix}\right)\right) \tag{124}$$

for all $a \in [T]$. The output partition function is, conditionally on χ^*, χ:

$$Z'_{\rm out} = \int \prod_a \mathrm{d}\mathcal{N}(\tilde{\chi}_a; \omega_a, v_{\rm BO}) \mathbb{E}_{z^*,z}\, \delta(\sigma(z + \tilde{\chi}\tilde{\chi}^\top - q_0 I_T) - \sigma(z^* + \chi^*\chi^{*\top} - q_0 I_T)) \tag{125}$$

where $\omega_a = \chi_a$, $v_{\rm BO} = q_0 - q_{\rm BO}$ and δ an element-wise Dirac's delta. z^* and z are treated as noise in the output channel $\chi \mapsto \sigma(\chi\chi^\top - q_0 I)$. Define $g'_{\rm out} = \nabla_\omega \log Z'_{\rm out}$ the output denoising function. The output-channel equation is then

$$\hat{q}_{\rm BO} = \alpha' \mathbb{E}_{\chi^*,\chi} \left\| g'_{\rm out} \right\|_2^2 . \tag{126}$$

One can explicit $Z'_{\rm out}$ for numerical computations. For σ softmax, for $x, y \in \mathcal{S}(T)$, $\sigma(x) = \sigma(y)$ is equivalent to there is $s \in \mathbb{R}$ such that $x - y = s\mathbb{1}_H \mathbb{1}_H^\top$. One further has that $(z - z^*)_{ab} \sim \mathcal{N}(0, (1 + \delta_{ab}) V_{\rm BO})$ with $V_{\rm BO} = Q_0 + \Delta/2 - Q_{\rm BO}$. Integrating over s gives

$$Z'_{\rm out} = \int \prod_a \mathrm{d}\mathcal{N}(\tilde{\chi}_a; \omega_a, v_{\rm BO}) e^{-\frac{1}{4V_{\rm BO}} ||\tilde{\chi}\tilde{\chi}^\top - \chi^*\chi^{*\top}||_F^2 + \frac{1}{4T^2 V_{\rm BO}} ((\sum_a \tilde{\chi}_a)^2 - (\sum_a \chi^*_a)^2)^2} \tag{127}$$

Equations (123) and (126) determine $q_{\rm BO}$.

For $\mathrm{e} = 1$, the fine-tuning inputs are among those observed at pre-training time, i.e. we are given both $\sigma(z^*)$ and $\sigma(z^* + \chi^*\chi^{*\top} - q_0 I_T)$. Because of the invertibility of the function one has access to a noise-less χ^* and therefore can exactly infer w^* whenever $\alpha' \geq 1$. This automatically implies $q_{\rm BO} = 1$.

In both cases $\mathrm{e} = 0$ or $\mathrm{e} = 1$, the BO test error for fine-tuning is

$$\mathcal{E}'_{\rm BO} = \mathbb{E} \left\| \sigma\left(z + \chi\chi^\top - q_0 I_T\right) - \sigma\left(z^* + \chi^*\chi^{*\top} - q_0 I_T\right) \right\|_F^2 , \tag{128}$$

where (z, z^*, χ, χ^*) are distributed according to (121) and (124). The BO overlap with w^* is

$$o_{\rm BO} = \sqrt{q_{\rm BO}/q_0}. \tag{129}$$

C Details on the effective noises

We detail how the effective noises $\Delta_{\rm eff}$, $\Delta_{\rm eff}(\beta)$, $\Delta_{\rm eff}(\mathrm{e} = 1)$ and $\Delta_{\rm eff,ActFT}$ are obtained in Eq. 23, Eq. 24, Eq. 25 and Eq. 26.

Fresh sequences.
For $\mathrm{e} = 0$ the pre-activations over the fine-tuning sets $\mathcal{S}'$ or $\bar{\mathcal{S}}'$ (active learning) follow $z^*, z \sim \mathcal{Q}$ (Res. 1). $\mathcal{Q}$ factorizes over the token indices and is Gaussian. For some token indices $a \neq b$, the variance of $h_{ab} = z_{ab} - z^*_{ab}$ is

$$\Delta_{\rm eff} = \mathrm{Var}(z_{ab} - z^*_{ab}) \tag{130}$$

$$= \mathrm{Var}\, z_{ab} + \mathrm{Var}\, z^*_{ab} - 2\mathrm{Cov}(z_{ab}, z^*_{ab}) \tag{131}$$

$$= Q + \frac{\Delta}{2} + Q_0 - 2M\ . \tag{132}$$

The case $a = b$ is similar up to a global factor 2.

Recalibration. Regarding the recalibration of the frozen map proposed in Eq. 24 it is easy to see from the order parameter definition that upon the rescaling $\hat{W} \to \sqrt{\beta}\hat{W}$, this corresponds to sending $Q \to \beta^2 Q$ and $M \to \beta M$, which leads to the recalibrated effective noise:

$$\Delta_{\text{eff}}(\beta) = \frac{\Delta}{2} + Q_0 + \beta^2 Q - 2\beta M, \tag{133}$$

which is minimized for $\beta_{\min} = M/Q$, thus leading to the optimal effective noise upon recalibration:

$$\Delta_{\text{eff}}(\beta_{\min}) = \frac{\Delta}{2} + Q_0 - \frac{M^2}{Q}. \tag{134}$$

stated in the main text.

Active fine-tuning. For the active fine-tuning, we explicit the joint distribution of z and z^*, for some token indices $a \neq b$, by taking independent $\zeta_1, \zeta_2 \sim \mathcal{N}(0,1)$ and rewriting $\begin{pmatrix} z^*_{a,b} \\ z_{a,b} \end{pmatrix} \sim \mathcal{Q}_{ab} = \mathcal{N}\left(0, \begin{pmatrix} \Delta/2+Q_0 & M \\ M & Q \end{pmatrix}\right)$ as

$$z_{ab} = \sqrt{Q}\zeta_1 \ , \qquad z^*_{ab} = \frac{M}{\sqrt{Q}}\zeta_1 + \sqrt{\Delta/2 + Q_0 - M^2/Q}\zeta_2 \ . \tag{135}$$

We assume that we can choose the fine-tuning set $\mathcal{S}'$ such that $\operatorname{Var} z \approx 0$ on it. This means that $\operatorname{Var} \zeta_1 \approx 0$ and only remains the variance from ζ_2, i.e.

$$\operatorname{Var}(z_{ab} - z^*_{ab} | z_{ab} \approx 0) = \frac{\Delta}{2} + Q_0 - \frac{M^2}{Q} \tag{136}$$

which is $\Delta_{\text{eff,ActFT}}$. The case $a = b$ is similar up to a global factor 2. We have $\Delta_{\text{eff}} - \Delta_{\text{eff,ActFT}} = (M/\sqrt{Q} - \sqrt{Q})^2 \geq 0$ and the active fine-tuning reduces the effective noise.

Reused sequences.
For $\mathrm{e} = 1$ the pre-activations over the fine-tuning set $\mathcal{S}'$ follow $z^*, z \sim \mathcal{R}$ (Res. 2), i.e. $z = \operatorname{Prox}(z^*, \tilde{z})$ with $z^*, \tilde{z} \sim \mathcal{Q}$. One has to compute the proximal $\operatorname{Prox}(z^*, \tilde{z})$ defined as the optimizer of Ψ over $\mathcal{S}(T)$. For linear σ one has

$$\Psi(z; z^*, \tilde{z}) = -\frac{1}{T^2}||z - z^*||_F^2 - \sum_{a \leq b}^{T} \frac{1}{2V(1+\delta_{a,b})}(\tilde{z}_{a,b} - z_{a,b})^2 \ . \tag{137}$$

Ψ is quadratic in z and we obtain $\operatorname{Prox}(z^*, \tilde{z}) = (T^{-2} + (4V)^{-1})^{-1}(T^{-2}z^* + (4V)^{-1}\tilde{z})$. Consequently the noise term is $z - z^* = (1 + 4VT^{-2})^{-2}(\tilde{z} - z^*)$ and

$$\operatorname{Var}_{z^*, z \sim \mathcal{R}}(z_{ab} - z^*_{ab}) = \frac{1}{(1+4VT^{-2})^2} \operatorname{Var}_{z^*, \tilde{z} \sim \mathcal{Q}}(\tilde{z}_{ab} - z^*_{ab}) \tag{138}$$

$$= \frac{1}{(1+4VT^{-2})^2}(1+\delta_{a,b})\Delta_{\text{eff}} \ . \tag{139}$$

D Details on the numerics

We describe the numerical procedures used to produce the theoretical predictions and the finite-D simulations reported in the figures. The code used to generate the results is provided in the supplementary material.

The theoretical predictions are obtained by iterating until convergence the self-consistent equations in App. A.3 and A.5. The expectations are computed by Monte-Carlo sampling. We use $n_{\text{MC}} \geq 3 \times 10^4$ samples so the remaining sampling noise is comparable to the finite-D fluctuations of the simulations. The proximals are numerically extremized for each sample. The iterations are stable and we use small damping.

For the theoretical predictions for active learning, we sample the pre-activations z^*, z needed for the integration according to Algorithm 1. We first draw $\bar{\alpha}'/\alpha' n_{\rm MC}$ Monte-Carlo samples of z^*, z according to $\mathcal{Q}$, compute $\sigma(z)$ for each of them and select the $n_{\rm MC}$ samples with lowest $||\sigma(z) - \sigma(0)||_F^2$.

Regarding the simulations of the attention at given D, the losses of both stages are optimized using PyTorch L-BFGS with standard hyperparameters. In most cases, random initialization gives objective values and observables consistent with the theoretical predictions; this does not certify global optimality. In a few limited cases the optimization does not converge and we re-ran it with a different randomness. For Fig. 4, for the "without active fine-tuning" curves at large $\Delta_{\rm eff}$ and small λ', the optimization does not converge in most of the cases and we initialize it at $w = w^*$. The test set size is proportional to D and is taken as $10D$ samples. Error bars show the empirical standard deviation across random seeds.

Running one theoretical prediction takes a few hours and a few GBs on a local CPU. Running one simulation at $D = 10^3$ takes a few minutes and a few GBs on a local GPU. Overall we ran a hundred of predictions and a thousand of simulations.

E SUPPLEMENTARY FIGURES

Effective noise for the linear attetion. Figure 5 complements Fig. 2 by considering the rescaled identity activation $\sigma : x \mapsto x/T$. In this case the frozen pre-trained component enters the fine-tuning problem only through the scalar effective noise $\Delta_{\rm eff} = \Delta/2 + Q_0 - 2M + Q$, as derived in App. C. The right panel subtracts the value at $\alpha' \to \infty$ and shows the excess error due to the limited number of fine-tuning samples.

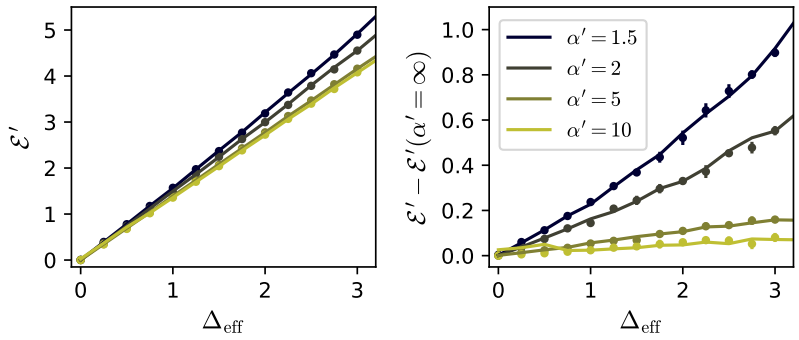


Figure 5: Effect of pre-training on the fine-tuning task for the rescaled identity activation $\sigma : x \mapsto x/T$. Fine-tuning test error $\mathcal{E}'$ versus $\Delta_{\rm eff} = \Delta/2 + Q_0 - 2M + Q$. Left: test error; right: excess test error with respect to the value at $\Delta_{\rm eff} = 0$. Lines are theoretical predictions; dots are simulations at $D = 5 \times 10^3$ over five independent runs. $T = 3$.

Finite-dimensional recalibration protocol. We consider the regime $\Delta = 0.5, \alpha = 0.1, \alpha' = 3, T = 3, \kappa_0 = \kappa = 1$, softmax attention and study a range of pre-training regularization strengths λ. For each value of λ, we first train the extensive-rank pre-training estimator $\hat{W}$ at finite dimension using LBFGS and compute the recalibration factor $\beta^\star = \frac{M}{Q}$, where (Q, M) are the corresponding pre-training state-evolution order parameters. The frozen representation is then rescaled according to $\hat{W} \mapsto \sqrt{\beta}\,\hat{W}$, with

$$\beta \in \{0.7\beta^\star,\ 0.85\beta^\star,\ \beta^\star,\ 1.15\beta^\star,\ 1.30\beta^\star\}\,. \tag{140}$$

For each pair (λ, β), we retrain the downstream LoRA vector while keeping the recalibrated matrix fixed and evaluate the corresponding generalization error. Finite-dimensional simulations are performed at $D = 800$ and averaged over 10 independent disorder realizations, with error bars indicating standard deviations. For each realization, optimization is repeated from four LBFGS initializations and the solution achieving the lowest objective value is retained. The estimation of the pre-training theoretical predictions is run for 250 iterations with Monte Carlo output-channel sampling, and order parameters are estimated by averaging the final 30 iterates to suppress residual sampling fluctuations.

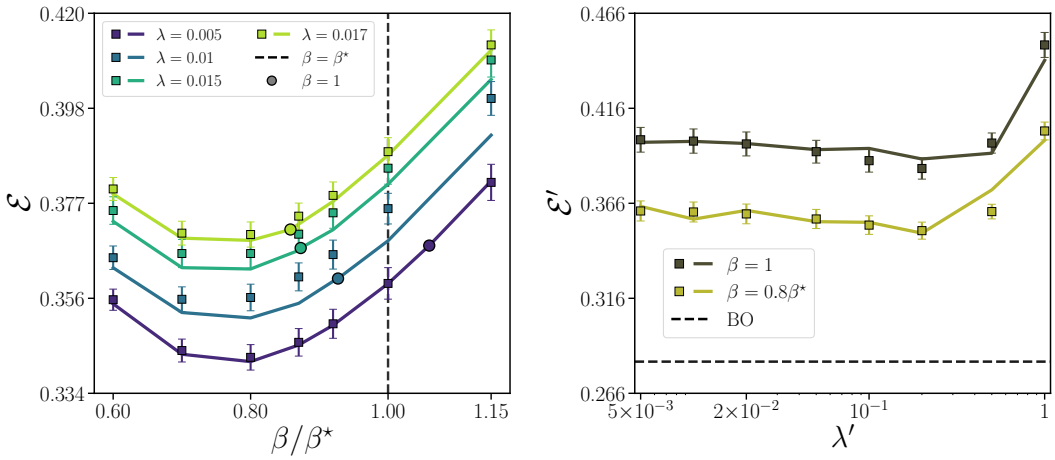


Figure 6: Recalibration of the frozen extensive-rank attention map improves downstream LoRA transfer. (Left:) pre-training test error $\mathcal{E}(\beta)$ as a function of the rescaling factor $\beta/\beta^\star$, where $\beta^\star = M/Q$. The dashed vertical line indicates $\beta = \beta^\star$, while circular markers denote the uncalibrated baseline $\beta = 1$. (Right:) fine-tuning test error $\mathcal{E}'$ as a function of the LoRA regularization λ' for the frozen representation obtained from the best pre-training configuration, i.e. $\lambda = 0.005$ and $\beta = 0.8\beta^\star$. We compare the original frozen representation ($\beta = 1$) with the recalibrated representation ($\beta = 0.8\beta^\star$). Solid lines denote state-evolution predictions and square markers correspond to finite-dimensional LBFGS simulations. The horizontal dashed line in the right panel indicates the Bayes-optimal benchmark.

The benefit of the recalibration procedure is shown in Fig. 6. This experiment uses $\beta^\star = M/Q$ as an informed reference.