How Much Rank Does LoRA Need? Rank–Error Bounds for Transformer Attention

Gerard Conangla Planes*
Aily Labs

Abstract

Choosing the rank of a low-rank adaptation (LoRA) update is usually an empirical task. In this paper, we provide a task-dependent theory of the approximation error achievable at each LoRA rank for Transformer attention. We fix a pretrained attention head, a target attention function, and a distribution over inputs from the downstream task, and bound the smallest expected Kullback–Leibler (KL) error achievable by a rank-r query LoRA update. When target attention probabilities are bounded away from zero, we prove a lower bound of the error proportional to $\psi(\|d\|_2)$, where d is the difference between candidate and target attention scores and $\psi(t) = \min\{t^2, t\}$. We also prove an unconditional upper bound $\min\{\|d\|_2^2/4, \sqrt{2}\|d\|_2\}$. Under explicit realizability, geometry, and moment conditions, we then bound the best rank-r error between an explicit multiple of $\psi(\sqrt{T_r})$ and $\min\{T_r/4, \sqrt{2T_r}\}$, where T_r is the downstream-weighted tail energy of the target update. We also provide target-Fisher bounds when candidate scores remain within a fixed range of the target scores, and an unrestricted lower bound when a subset of tokens carries most of the probability mass. These spectral bounds describe finite-score approximation. We then construct explicit families in which softmax saturation makes the rank required to match the attention function strictly smaller than the rank required to match the finite logits. Finally, we extend the analysis to fused multi-head LoRA and joint query/key updates, exposing the effects of rank sharing and query/key factorization constraints.

1 Introduction

LoRA [11] makes it cheaper to fine-tune a pretrained Transformer [22] by replacing a dense weight update with a low-rank one. Such a low rank is both an advantage and a restriction, reducing the number of trainable parameters and computation, but also limiting the representational power of the adapter. In practice, rank is commonly chosen using rules of thumb or by training several adapters of different rank and comparing their downstream performance [11, 27, 21]. Such a sweep reveals which rank worked in a particular run, but it doesn't tell us whether a smaller adapter lacked capacity or was simply harder to train.

To isolate the representation question, suppose a dense or high-rank target adapter has already been obtained. We then ask: for each rank budget r, how closely can an adapter reproduce the target on inputs from the downstream task? The answer cannot be read from the singular values of the target weight update ΔW alone. For instance, a large singular direction may have no effect if inputs from the task never activate it, while a smaller direction may matter on nearly every input because it repeatedly changes which tokens receive attention. The softmax function in the attention mechanism introduces two further effects. First, adding the same constant to every score leaves the attention probabilities unchanged. Second, softmax saturates, so a score error

*Correspondence: gerardpc@gmail.com

with large magnitude need not produce a proportionally large attention error. The rank needed to reproduce the target function can therefore differ from the rank suggested by the raw singular values of its weight update.

In this paper, we begin with a LoRA update to the query projection of a single attention head. That update changes the scores assigned to the available keys, which softmax converts into attention probabilities. We compare these probabilities with those of a dense or high-rank target adapter, ask for the smallest expected KL error achievable at rank r, and obtain bounds on the error, some unconditional and some conditional on explicit assumptions. Attention KL therefore measures how much the query update changes the attention distribution itself, before later layers can modify that change.

We stress that our results bound the attention KL described above, but we do not provide bounds on the head output or final task loss. To do so would require additional assumptions about the architecture, including the value and output projections, residual connections, feed-forward network, and later layers. These components determine how a difference in one head is propagated and may amplify, suppress, or cancel it. Without such architecture-specific assumptions, attention KL alone does not determine output or task error. Figure 1 shows the scope of our analysis within a Transformer block in more detail, including extensions to fused multi-head and joint query/key LoRA.

Overview of the results. For each rank budget r, let $\mathcal{E}_r$ denote the smallest expected attention KL over all rank-r candidates; Section 2 defines this objective formally. Our analysis first connects attention KL to score approximation, then connects score approximation to LoRA rank.

Theorem 3.1 gives the first connection. If d is the difference between candidate and target scores after removing their common shift, define

$$\psi(t) = \min\{t^2, t\}, \qquad \psi_{\mathrm{up}}(t) = \min\{t^2/4, \sqrt{2}\,t\}. \tag{1}$$

The theorem proves that the pointwise attention KL is comparable to $\psi(\|d\|_2)$ when the target probabilities are bounded away from zero, and is at most $\psi_{\mathrm{up}}(\|d\|_2)$ without this assumption. Thus the error is quadratic for small score differences and linear for large ones. Corollary 3.2 lifts this pointwise statement to the expected, rank-constrained objective $\mathcal{E}_r$.

Theorem 4.1 supplies the second connection. Under its realizability, geometry, probability-floor, and moment assumptions, it proves

$$c_{\mathrm{lo}}\,\psi(\sqrt{T_r}) \le \mathcal{E}_r \le \psi_{\mathrm{up}}(\sqrt{T_r}). \tag{2}$$

Here T_r is the residual tail energy after the target update has been weighted by the queries and keys occurring on the downstream task. Directions that the task never uses therefore do not contribute, and c_{lo} is the explicit assumption-dependent constant in Theorem 4.1. The upper bound constructs a candidate whose error does not exceed its curve; the lower bound applies to every candidate. Together they bracket the best achievable error at each rank. Section 9 explains how to estimate this rank–error curve from downstream data and compare it with a desired error tolerance.

When the probability floor in the main theorem is too small to give a useful lower bound, Section 5 provides two alternatives with different assumptions (Figure 2 shows a decision diagram to clarify the choices, and Table 2 gives an overview of the three laws of approximation). Theorem 5.1 gives target-Fisher upper and lower bounds for candidates whose scores remain within a fixed range of the target scores. Because that candidate class is restricted, its lower bound does not apply to the unrestricted $\mathcal{E}_r$. Theorem 5.2 instead gives a lower bound on the unrestricted $\mathcal{E}_r$ by focusing on a subset of tokens carrying most of the target probability mass.

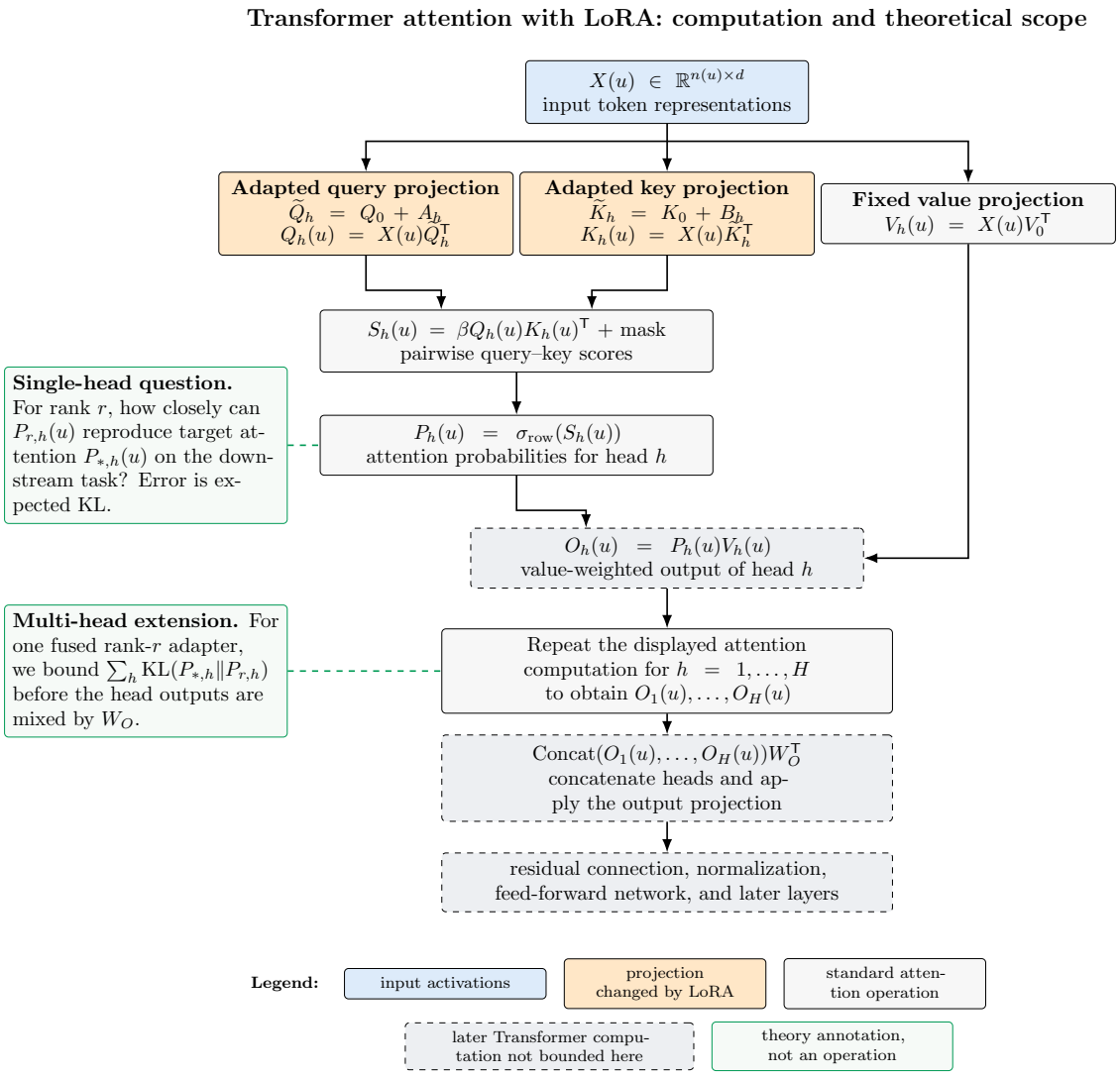

Figure 1: Transformer attention with LoRA and the scope of our analysis. Read from top to bottom, the diagram traces an input through the query, key, and value projections, attention scores, row-wise softmax, and multi-head aggregation. Our central question is how closely a rank-r LoRA update can reproduce a target attention function on inputs from the downstream task. The main theorem treats query-only adaptation in one head; later results cover a fused update shared across heads and simultaneous query/key updates. For an input u, $X(u)$ contains $n(u)$ token representations of width d. For each head index $h \in \{1,\dots,H\}$, query/key width is p and value width is p_v, with $Q_h(u), K_h(u) \in \mathbb{R}^{n(u)\times p}$, $V_h(u) \in \mathbb{R}^{n(u)\times p_v}$, and $S_h(u), P_h(u) \in \mathbb{R}^{n(u)\times n(u)}$. Here σ_{row} applies softmax row-wise and β is typically $1/\sqrt{p}$. The pretrained projections Q_0, K_0, V_0 are fixed, while A_h and B_h are rank-r LoRA updates. The multi-head result bounds the sum of headwise attention KL errors before W_O; the remaining Transformer computation and the final model output are outside this guarantee.

The preceding results concern approximation at finite scores. Theorem 6.1 treats a different effect: after softmax saturates, a lower-rank sequence of logits can approach an attention function that requires higher rank to realize with finite logits. It constructs an explicit family with a constant-factor separation between these two ranks. Finally, Theorem 7.1 extends the rank–KL bounds to a fused update shared across attention heads, and Theorem 8.1 extends them to simultaneous query/key LoRA, where the score update must also satisfy a query/key factorization constraint.

Contributions.

1. Under explicit assumptions, we characterize the best task-dependent attention approximation available at each LoRA rank. Global softmax bounds convert centered score error into KL, and a downstream-weighted spectral theorem converts the remaining score error into an explicit function of rank.
2. We give two alternatives when the main lower-bound constant is weak: target-Fisher bounds for a score-restricted candidate class, and a high-mass lower bound for the original unrestricted class.
3. We prove an exact separation for an explicit Walsh family: matching its finite logits requires rank k, but after softmax saturation the same limiting attention can be recovered to arbitrary accuracy at rank $k - \lfloor k/3 \rfloor$ within the score space reachable by query adaptation.
4. We extend the rank–KL analysis to a fused update shared across heads and to joint query/key LoRA, where a separate factorization gap measures whether the best effective score update can be realized by the two factors.

Positioning relative to prior work. Prior work studies adaptive rank allocation during training [27, 18], activation-aware approximation [24, 4, 23], finite-sample LoRA rank selection [12], and the transformations that LoRA can express [26]. Our contribution is different: for a fixed pretrained head and a known target attention function, we bound the best attention KL attainable at every rank, averaged over $u \sim P$, the downstream task distribution (e.g. text-to-SQL examples). The underlying weighted approximation theorem is classical [7, 16]; the new step is to connect that task-weighted approximation problem to global upper and lower bounds on attention KL. Section 10 gives a more detailed comparison with these and the related attention-rank and softmax-saturation literature.

2 Problem formulation

We begin with one query position in one attention head, so that every matrix and function can be defined explicitly. Sections 7 and 8 extend the formulation.

Let P denote the downstream input distribution and draw $u \sim P$, with $n(u)$ available token positions. Fix pretrained logits $z_0(u) \in \mathbb{R}^{n(u)}$, a pretrained key matrix $K(u) \in \mathbb{R}^{n(u)\times d_k}$, a query activation $h(u) \in \mathbb{R}^{d_h}$, and the attention scale $\beta > 0$. A query LoRA update is a matrix $M \in \mathbb{R}^{d_k \times d_h}$ with $\operatorname{rank}(M) \le r$. It produces the score vector and attention distribution

$$z_M(u) = z_0(u) + \beta K(u) M h(u), \qquad p_M(\cdot \mid u) = \operatorname{softmax}(z_M(u)). \tag{3}$$

Here $h(u)$ is the query representation entering the adapted projection, and the rows of $K(u)$ are the key vectors against which that query is scored. M is one stored matrix; the attention distribution $p_M(\cdot \mid u)$ changes with the input u.

Let $z_*(u)$ be fixed target logits and $p_*(\cdot \mid u) = \operatorname{softmax}(z_*(u))$ the target attention function. The target may be arbitrary for the score-space results in Section 3. The spectral specialization in Theorem 4.1 requires it to be generated exactly by a dense query update Δ_*. A generic fully fine-tuned Transformer can also change keys, values, earlier layers, and the activations entering this head, so it need not satisfy that assumption. For such a misspecified target, Equation (4) remains well defined, but the query-only spectral formula would need an extra irreducible approximation term. For rank budget r, define

$$\boxed{\mathcal{E}_r = \inf_{\operatorname{rank}(M)\le r} \mathbb{E}_{u\sim P} \mathrm{KL}(p_*(\cdot \mid u) \,\|\, p_M(\cdot \mid u)) \,.} \tag{4}$$

Thus $\mathcal{E}_r$ is the best error attainable by the entire rank-r candidate class. An upper bound exhibits a candidate with small error. A lower bound applies to every candidate and certifies unavoidable error. Inverting the bounds for a tolerance ϵ then yields sufficient and necessary rank conditions. Section 9 turns this interpretation into a calibration procedure.

Softmax is invariant to adding a constant to all scores. We therefore center score differences with

$$\Pi_n = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad d_M(u) = \Pi_{n(u)}\big(z_M(u) - z_*(u)\big). \tag{5}$$

The associated approximation objective, robust to both small and large score errors, is

$$\Psi_r = \inf_{\mathrm{rank}(M)\le r} \mathbb{E}_{u\sim P}\psi(\|d_M(u)\|_2)\,. \tag{6}$$

For the sharper upper bound, also define

$$\Phi_r = \inf_{\mathrm{rank}(M)\le r} \mathbb{E}_{u\sim P}\psi_{\mathrm{up}}(\|d_M(u)\|_2)\,. \tag{7}$$

Object	Meaning	What changes?
u	One input from the downstream task	Varies across examples
$z_0(u)$	Attention scores produced by the pretrained head	Depends on u
$p_*(\cdot \mid u)$	Target attention probabilities to be reproduced	Depends on u; the target model is fixed
M	Candidate query update with $\mathrm{rank}(M) \le r$	One fixed matrix for all inputs
$p_M(\cdot \mid u)$	Attention probabilities produced with update M	Depends on u and on the chosen M
$d_M(u)$	Candidate–target score difference after removing a common shift	Depends on u and on the chosen M
$\mathcal{E}_r$	Smallest average target-to-candidate KL achievable at rank r	One number for each rank budget

Table 1: Objects in the target–candidate comparison. For each downstream input u, the pretrained head and fixed target determine the behavior to be matched; M ranges over candidate updates of rank at most r.

Example 1 grounds this objective in a concrete downstream task.

Example 1: Interpreting attention KL in text-to-SQL

An input may contain the user's question, the database schema, instructions, and previously generated SQL tokens. The target and candidate heads each assign attention probabilities to those token positions. Our objective asks how much rank is needed for the candidate to reproduce the target's attention on such inputs. This gives a head-level approximation measure that can be studied alongside end-to-end measures such as SQL execution accuracy.

3 From score error to attention KL

The first result is independent of LoRA: it relates centered score error to probability error.

Theorem 3.1 (Global robust softmax bounds)**.** *Let $d \in \mathbb{R}^n$ satisfy $\mathbf{1}^\top d = 0$, let $p_* = \mathrm{softmax}(z_*)$, and put $a = \min_i p_{*,i} > 0$. Then*

$$\frac{a}{2e^2}\psi(\|d\|_2) \le \mathrm{KL}(p_* \,\|\, \mathrm{softmax}(z_* + d)) \le \psi_{\mathrm{up}}(\|d\|_2). \tag{8}$$

The upper bound does not require a positive lower bound on a.

The quadratic part of ψ comes from local curvature of log-sum-exp. The linear part is unavoidable globally: a rare input can contain a very large score error while contributing only linear or bounded KL. A purely quadratic global lower bound therefore cannot hold; Proposition B.1 gives an explicit fixed-probability-floor construction.

Corollary 3.2 (Global functional bounds)**.** *If* $\min_i p_{*,i}(u) \geq a > 0$ *almost surely, then*

$$\frac{a}{2e^2}\Psi_r \leq \mathcal{E}_r \leq \Phi_r. \tag{9}$$

The proof of Theorem 3.1 is short and appears in Appendix A. The rare-context obstruction to a quadratic law is given in Appendix B.

4 Task-weighted spectral rank–KL bounds

Assume in this section that the target is exactly realizable by a dense query update Δ_*. For $A = M - \Delta_*$, define

$$G(u) = \beta^2 K(u)^\mathsf{T} \Pi_{n(u)} K(u), \qquad \Sigma = \mathbb{E}[h(u)h(u)^\mathsf{T}]. \tag{10}$$

$G(u)$ records which update directions alter centered attention scores in context u; Σ records which query directions appear on the task.

In the separable case, the random key Gram matrix $G(u)$ is independent of the query activation $h(u)$. Then

$$\mathbb{E}\|\beta \Pi_{n(u)} K(u) A h(u)\|_2^2 = \|G^{1/2} A \Sigma^{1/2}\|_\mathrm{F}^2, \qquad G = \mathbb{E}G(u). \tag{11}$$

The matrix whose spectrum matters is therefore

$$D_* = G^{1/2} \Delta_* \Sigma^{1/2}, \qquad T_r = \sum_{j>r} \sigma_j(D_*)^2. \tag{12}$$

T_r is the residual squared score error after the best rank-r update. Directions annihilated by the keys or never activated by $h(u)$ disappear automatically, a phenomenon that is illustrated in Example 2.

Example 2: A large update direction that the task never uses

Suppose the target update has two singular directions. The first has the larger singular value, but every query activation from the downstream task is orthogonal to it, so it never changes an attention score. The second is smaller but active on nearly every input. A raw truncated SVD keeps the first direction; the weighted spectrum above removes it and keeps the direction that actually changes attention. Theorem 4.1 turns this distinction into upper and lower error bounds.

To turn this quadratic tail into a global robust lower bound, assume that, for every deterministic compatible matrix C,

$$\mathbb{E}\|Ch(u)\|_2^4 \leq \kappa_h \big(\mathbb{E}\|Ch(u)\|_2^2\big)^2, \qquad G(u) \preceq \Lambda G \quad \text{almost surely.} \tag{13}$$

These uniform moment conditions prevent the quadratic error from being carried entirely by extremely rare contexts.

Theorem 4.1 (Downstream spectral rank–KL bounds)**.** *Suppose the target attention function is generated by a dense query update* Δ_*, $\min_i p_{*,i}(u) \geq a > 0$ *almost surely,* $G(u)$ *and* $h(u)$ *are independent, and the two conditions in Equation* (13) *hold. Then*

$$\boxed{\frac{a}{2e^2(1+\Lambda\sqrt{\kappa_h})}\psi(\sqrt{T_r}) \leq \mathcal{E}_r \leq \psi_{\mathrm{up}}(\sqrt{T_r}).} \tag{14}$$

The upper bound is achieved by truncated SVD of D_*, which already weights the target update by the task queries and keys. The lower bound applies to every rank-r candidate. Both sides are quadratic in the score scale near zero and linear at large scale, but their constants can be far apart when the target probability floor a is small.

For a desired error tolerance ϵ, the upper bound gives the sufficient rank

$$r_{\mathrm{suff}}(\epsilon) = \min\{r : \psi_{\mathrm{up}}(\sqrt{T_r}) \leq \epsilon\}. \tag{15}$$

Conversely, any rank attaining error at most ϵ must satisfy

$$\psi(\sqrt{T_r}) \leq \frac{2e^2(1+\Lambda\sqrt{\kappa_h})}{a}\epsilon. \tag{16}$$

Inverting the error bounds in this way yields sufficient and necessary rank conditions.

Remark 4.2 (Size of the bracket). The worst-case ratio between the upper and lower constants is at most

$$\frac{2\sqrt{2}e^2(1+\Lambda\sqrt{\kappa_h})}{a}. \tag{17}$$

For example, it is about 2.0×10^4 when $a = 10^{-2}$, $\Lambda = 5$, and $\kappa_h = 3$. This is a sensitivity calculation from the theorem, not an empirical estimate. For long, peaked attention vectors, the probability floor can make the global lower bound numerically weak; Section 5 gives alternatives.

The proof is in Appendix C. In ordinary self-attention, $G(u)$ and $h(u)$ are computed from the same input and need not be independent. The following result replaces independence by direct comparison between the true mean squared score error and the reference quadratic form used to define T_r.

For $A = M - \Delta_*$, write

$$Q(A) = \mathbb{E}\|\beta\Pi_{n(u)}K(u)Ah(u)\|_2^2, \qquad S(A) = \|G^{1/2}A\Sigma^{1/2}\|_{\mathrm{F}}^2, \tag{18}$$

and let A_r^{svd} be the weighted-SVD candidate used in the upper bound.

Theorem 4.3 (Dependence-allowing spectral bounds)**.** *Suppose, for every feasible* $A = M - \Delta_*$,

$$Q(A) \geq c_r S(A), \qquad \mathbb{E}\|\beta\Pi_{n(u)}K(u)Ah(u)\|_2^4 \leq \kappa_r Q(A)^2, \tag{19}$$

and suppose $Q(A_r^{\mathrm{svd}}) \leq C_r^+ T_r$*. If the target probability floor is at least* $a > 0$*, then*

$$\frac{a}{2e^2(1+\sqrt{\kappa_r})}\psi(\sqrt{c_r T_r}) \leq \mathcal{E}_r \leq \psi_{\mathrm{up}}(\sqrt{C_r^+ T_r}). \tag{20}$$

This theorem allows arbitrary dependence between keys and queries; its price is that the comparison and moment constants must hold on the candidate class. For example, if $\lambda G \preceq G(u) \preceq \Lambda G$ almost surely and the activation moment condition in Equation (13) holds, it applies with $c_r = \lambda$, $C_r^+ = \Lambda$, and $\kappa_r = \kappa_h(\Lambda/\lambda)^2$. The proof is in Appendix D.

5 Which rank–KL bound should be used?

Because the minimum target probability can be very small in long, peaked attention vectors, we provide two refinements that address different needs. The target-Fisher theorem gives upper and lower bounds when score differences stay inside a fixed range. The high-mass theorem gives a lower bound for the original unrestricted class, using a set of tokens, chosen from the target, that carries most probability mass. Figure 2 and Table 2 make the distinction explicit:

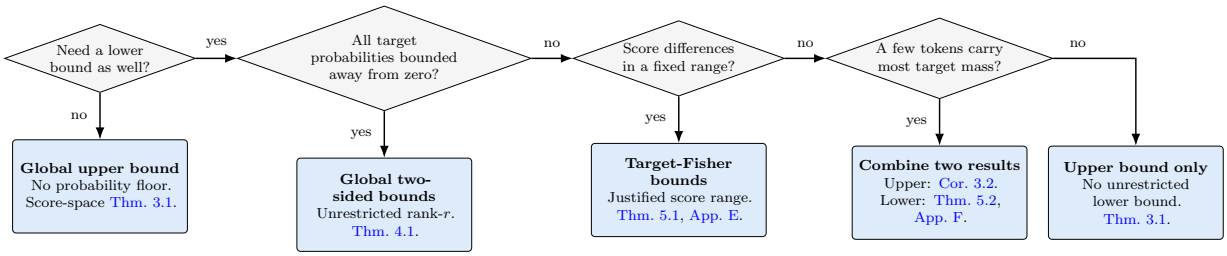

Figure 2: Which rank–KL theorem to use. Start at the left and stop at the first result whose assumptions you can justify. An upper bound shows that some rank-r update achieves at most the stated error; a lower bound shows that every rank-r update incurs at least the stated error. The Fisher bounds apply only to candidates whose scores remain within a fixed range of the target scores. Consequently, their lower bound does not apply to the unrestricted optimum $\mathcal{E}_r$.

Route	Upper bound	Lower bound	Candidate class	Main price
Global	✓	✓	unrestricted rank-r	minimum target probability and moment assumptions on keys/queries
Target Fisher	✓	✓	rank-r with a justified range of score differences	span-dependent constants and the target Fisher inner product
High mass	×	✓	unrestricted rank-r	tokens carrying most target mass, a conditional probability floor, and the key/query Gram on that set

Table 2: The three rank–KL routes and their logical scope.

5.1 Target-Fisher bounds under a controlled score span

Let $H_* = \operatorname{diag}(p_*) - p_* p_*^\mathsf{T}$ and let $R(d) = \max_i d_i - \min_i d_i$. Define

$$c_-(R) = \frac{R - 1 + e^{-R}}{R^2}, \qquad c_+(R) = \frac{e^R - 1 - R}{R^2}, \tag{21}$$

with both values set to $1/2$ at $R = 0$.

Theorem 5.1 (Target-Fisher score bounds)**.** *For every score displacement d,*

$$c_-(R(d)) d^\mathsf{T} H_* d \le \mathrm{KL}(p_* \| \operatorname{softmax}(z_* + d)) \le c_+(R(d)) d^\mathsf{T} H_* d. \tag{22}$$

Consequently, on the candidate class whose score differences from the target have range at most R_0, the optimal KL is bounded above and below by $c_+(R_0)$ and $c_-(R_0)$ times the corresponding target-Fisher quadratic optimum.

There is no explicit minimum-probability constant; small target probabilities enter through the Fisher matrix H_*. More importantly, the optimized class is restricted by the span condition. Because that class is a subset of the unrestricted rank-r class, its lower bound cannot be transferred to $\mathcal{E}_r$. The spectral statements, including the extra span condition needed for a matching upper bound, and the proofs are in Appendix E.

5.2 An unrestricted high-mass lower bound

For each context, let $S(u)$ be a measurable token set chosen from the target before optimizing the candidate. Suppose it carries at least $1-\delta$ of the target mass and the conditional target distribution on $S(u)$ has minimum probability at least a_{core}.

Theorem 5.2 (High-mass lower bound)**.** *Under the preceding assumptions,*

$$\mathcal{E}_r \geq \frac{(1-\delta)a_{\mathrm{core}}}{2e^2} \inf_{\mathrm{rank}(M)\leq r} \mathbb{E}\psi\Big(\big\|\Pi_{S(u)}(z_M(u)-z_*(u))_{S(u)}\big\|_2\Big). \tag{23}$$

This is a lower bound on the original unrestricted $\mathcal{E}_r$. Under additional control of the key/query Gram and moments on $S(u)$, the right-hand side reduces to the same spectral tail, but only on those tokens. It does not give a full-KL upper bound: matching conditional scores inside $S(u)$ need not match the total mass assigned to $S(u)$. The proof and spectral forms are in Appendix F.

6 Saturation changes the relevant notion of rank

The main theorem describes approximation of finite logits, but softmax also has a boundary regime in which logits can diverge while their distributions converge. Example 3 shows the basic saturation mechanism before the rank separation is stated formally.

Example 3: Different logits, the same saturated attention

For two tokens,

$$\mathrm{softmax}(T,0) \longrightarrow (1,0), \qquad \mathrm{softmax}(2T,0) \longrightarrow (1,0) \quad \text{as } T\to\infty.$$

The difference between the two logit vectors grows with T, but their attention distributions converge to the same limit. The results below prove a stronger effect: for explicit families of targets, saturation reduces the rank needed to approximate the entire attention function.

For a boundary target P_∞ and an allowed centered score space L, define its *L-relative softmax closure rank* $r_{\mathrm{cl}}^{L}(P_\infty)$ as the smallest r for which there is a sequence of score matrices of rank at most r, with every score column in L, whose columnwise softmax distributions converge to P_∞. For query adaptation, L is the score space reachable through the fixed key matrix, so the restriction is part of the LoRA model. This rank can be smaller than the rank required to match every finite target logit exactly.

Theorem 6.1 (Exact closure rank of an isolated-triple Walsh family)**.** *For every $k\geq 3$, there is a query-only Walsh-attention family with k contexts and fewer than $4k^2+8$ token positions such that every exact realization with finite logits has update rank k. For the limiting target $P_\infty^{(k)}$ of this family and its Walsh score space L_k, the exact relative closure rank is*

$$r_{\mathrm{cl}}^{L_k}(P_\infty^{(k)}) = k - \lfloor k/3 \rfloor. \tag{24}$$

This equality is for the isolated-triple family constructed in Appendix G; other target families can have smaller closure rank, as the construction below demonstrates.

The construction groups Walsh characters into isolated triples. A rank-two path per triple gives the upper bound. For the lower bound, restricted Fourier identities and columnwise normalization force every limiting triple block to retain rank at least two. Saturation therefore yields a genuine constant-factor reduction, but in this family it cannot collapse the rank of the attention function to $o(k)$.

A separate linear-token family gives a stronger achieved ratio: with fewer than $8(k+1)$ token positions, rank at most

$$k - \max\{\lfloor k/3 \rfloor, 3\lfloor k/7 \rfloor\} \tag{25}$$

attains vanishing KL, reaching ratio $4/7$ on complete seven-context blocks. That construction concerns a different target family and does not identify its minimum closure rank. The two results and their proofs are kept separate in Appendix G.

7 Fused multi-head LoRA

Standard query LoRA for multi-head attention often constrains one fused projection matrix, vertically divided into head blocks. This is not the same as assigning an independent rank to every head: one fused rank direction can serve several heads.

For $j = 1, \dots, H$, let $K_j(u) \in \mathbb{R}^{n_j(u) \times d_{k,j}}$ and let all heads share the query activation $h(u) \in \mathbb{R}^{d_h}$. Stack the head updates as $M = (M_1^\mathsf{T}, \dots, M_H^\mathsf{T})^\mathsf{T}$ and impose the single fused constraint $\operatorname{rank}(M) \le r$. Assume the target is generated by the similarly stacked dense update Δ_*^{MH}. Define

$$d_{M,j}(u) = \beta_j \Pi_{n_j(u)} K_j(u)(M_j - \Delta_{*,j})h(u), \tag{26}$$

$$\mathcal{E}_r^{\mathrm{MH}} = \inf_{\operatorname{rank}(M) \le r} \mathbb{E} \sum_{j=1}^{H} \mathrm{KL}(p_{*,j}(\cdot \mid u) \| p_{M,j}(\cdot \mid u)), \tag{27}$$

$$G_{\mathrm{MH}}(u) = \operatorname{blockdiag}\Big(\beta_j^2 K_j(u)^\mathsf{T} \Pi_{n_j(u)} K_j(u) : j = 1, \dots, H\Big), \qquad G_{\mathrm{MH}} = \mathbb{E} G_{\mathrm{MH}}(u), \tag{28}$$

$$T_r^{\mathrm{MH}} = \sum_{j>r} \sigma_j \Big(G_{\mathrm{MH}}^{1/2} \Delta_*^{\mathrm{MH}} \Sigma^{1/2}\Big)^2, \qquad \Sigma = \mathbb{E}[h(u)h(u)^\mathsf{T}]. \tag{29}$$

Theorem 7.1 (Fused multi-head rank–KL bounds)**.** *Suppose every target probability is at least $a_{\min} > 0$ almost surely, $G_{\mathrm{MH}}(u)$ is independent of $h(u)$, $G_{\mathrm{MH}}(u) \preceq \Lambda_{\mathrm{MH}} G_{\mathrm{MH}}$ almost surely, and*

$$\mathbb{E}\|Ch(u)\|_2^4 \le \kappa_h \big(\mathbb{E}\|Ch(u)\|_2^2\big)^2 \quad \text{for every compatible deterministic } C. \tag{30}$$

Then

$$\frac{a_{\min}}{2e^2(1 + \Lambda_{\mathrm{MH}}\sqrt{\kappa_h})} \psi\big(\sqrt{T_r^{\mathrm{MH}}}\big) \le \mathcal{E}_r^{\mathrm{MH}} \le \min\{T_r^{\mathrm{MH}}/4, \sqrt{2HT_r^{\mathrm{MH}}}\}. \tag{31}$$

The upper/lower bracket can widen by a factor of order $\sqrt{H}$ because the objective sums KL over heads; the aggregation inequality producing this factor is tight without further assumptions on how error is distributed across heads. If each head instead has its own adapter and integer rank r_h, the feasible set separates after the ranks are fixed; allocating a total budget R is the discrete problem $\min_{\sum_h r_h \le R} \sum_h \mathcal{E}_{h,r_h}$. The proof and the underlying pointwise aggregation inequality are in Appendix H.

8 Joint query/key LoRA

Practical LoRA may update both queries and keys. Let $Q_0, K_0 \in \mathbb{R}^{p\times d}$ be pretrained factors and let $A, B \in \mathbb{R}^{p\times d}$ be query and key updates with ranks at most r_Q, r_K. On ordinary dot-product scores they change the scores by the effective update

$$C(A,B) = (K_0+B)^\mathsf{T}(Q_0+A) - K_0^\mathsf{T}Q_0 = K_0^\mathsf{T}A + B^\mathsf{T}(Q_0+A). \tag{32}$$

The bound $\operatorname{rank}(C(A,B)) \le r_Q + r_K$ is known [26]: grouping the bilinear term gives the sharp containment

$$\operatorname{rank}(C(A,B)) \le r_Q + r_K. \tag{33}$$

Let a context provide token representations $X(u) \in \mathbb{R}^{n(u)\times d}$ and a query representation $h(u) \in \mathbb{R}^d$. For target factors (A_*, B_*), put $C_* = C(A_*, B_*)$ and define the centered score error and actual factor-class objective

$$d_{A,B}(u) = \beta \Pi_{n(u)} X(u)[C(A,B) - C_*]h(u), \tag{34}$$

$$\mathcal{E}^{\mathrm{QK}}_{r_Q,r_K} = \inf_{\substack{\operatorname{rank}(A)\le r_Q \\ \operatorname{rank}(B)\le r_K}} \mathbb{E}\mathrm{KL}(p_*\|p_{A,B}). \tag{35}$$

Let

$$G_X(u) = \beta^2 X(u)^\mathsf{T}\Pi_{n(u)}X(u), \qquad G_X = \mathbb{E}G_X(u), \qquad \Sigma = \mathbb{E}[h(u)h(u)^\mathsf{T}]. \tag{36}$$

If $G_X(u)$ and $h(u)$ are independent, the mean squared score error equals $\|G_X^{1/2}[C(A,B)-C_*]\Sigma^{1/2}\|_{\mathrm{F}}^2$. Define

$$D_*^{\mathrm{eff}} = G_X^{1/2} C_* \Sigma^{1/2}, \qquad T_s^{\mathrm{eff}} = \sum_{j>s} \sigma_j(D_*^{\mathrm{eff}})^2, \qquad s = r_Q + r_K. \tag{37}$$

Let $\mathcal{S}_s^{\mathrm{svd}}$ be the set of rank-s matrices attaining the weighted approximation error T_s^{eff}. An element of this set need not be writable as query and key updates of size $p \times d$ with the separate rank budgets. Define the extra error from that factorization gap by

$$\rho_{r_Q,r_K} = \inf_{C_s \in \mathcal{S}_s^{\mathrm{svd}}} \inf_{\substack{\operatorname{rank}(A)\le r_Q \\ \operatorname{rank}(B)\le r_K}} \|G_X^{1/2}[C(A,B) - C_s]\Sigma^{1/2}\|_{\mathrm{F}}. \tag{38}$$

Theorem 8.1 (Joint Q/K rank–KL bounds)**.** *Suppose* $\min_i p_{*,i}(u) \ge a > 0$ *almost surely,* $G_X(u)$ *is independent of* $h(u)$ *so that the mean-squared identity above holds, and, for every feasible pair* (A,B)*,*

$$\mathbb{E}\|d_{A,B}(u)\|_2^4 \le \kappa\big(\mathbb{E}\|d_{A,B}(u)\|_2^2\big)^2. \tag{39}$$

Then

$$\frac{a}{2e^2(1+\sqrt{\kappa})}\psi\big(\sqrt{T_s^{\mathrm{eff}}}\big) \le \mathcal{E}^{\mathrm{QK}}_{r_Q,r_K} \le \psi_{\mathrm{up}}\left(\sqrt{T_s^{\mathrm{eff}}} + \rho_{r_Q,r_K}\right). \tag{40}$$

The lower bound follows from containment in the effective rank-s class. The reverse containment generally fails: every updated width-p head also satisfies $\operatorname{rank}(K_0^\mathsf{T}Q_0 + C) \le p$. Exact one-sided support conditions or a sequential two-sided factorization make $\rho_{r_Q,r_K} = 0$; otherwise that extra error can be large. The realization lemmas, a constructive target-factor upper bound, and a fused multi-head lower bound are proved in Appendix I.

The two sides form a useful bracket only when ρ_{r_Q,r_K} is comparable to, or smaller than, $\sqrt{T_s^{\mathrm{eff}}}$. In general, computing ρ_{r_Q,r_K} is a nonconvex factorization problem over the weighted rank-s optimizers; the theorem does not provide a tractable procedure or a universal upper bound for it. The cases with $\rho_{r_Q,r_K} = 0$ therefore identify the cleanest regime for this extension.

RoPE inserts a relative rotation between the query and key factors. Every relative-position slice still has rank at most $r_Q + r_K$, so the global robust bound remains valid for the assembled scores. The candidate is then a coupled family of effective matrices, however, and the single-matrix spectral tail in Theorem 8.1 does not apply automatically.

This RoPE-specific obstruction is not universal. Architectures that omit explicit positional embeddings and use NoPE, such as Kimi K3 [13], keep token order only through causal computation. For those heads, the ordinary dot-product formulation is the more direct starting point, subject to the model's other attention mechanisms.

9 Estimating LoRA rank from downstream data

The theory can be used once a dense or high-rank target update and a calibration set from the downstream task are available. Its output is a pair of error curves over rank: an upper curve achieved by a constructed candidate and a lower curve that no candidate in the stated class can beat. These curves may identify one rank, or they may leave an interval that the theory does not resolve.

For query-only adaptation, the procedure is:

1. **Check the target class.** Verify that the target attention is generated by a query update Δ_*. If both queries and keys change, use Section 8; a target produced by unrestricted fine-tuning may also contain an irreducible query-only approximation error.

2. **Collect calibration inputs.** Sample complete inputs from the downstream task and run the fixed pretrained and target heads. Store the query activations $h(u)$, centered key Gram matrices $G(u)$, and target attention probabilities. Complete inputs, rather than individual token positions treated as independent observations, are the sampling units.

3. **Choose the applicable bound.** Use Figure 2. If the separability assumptions of Theorem 4.1 are not credible, use the dependence-allowing geometry in Appendix D. If the target probability floor is too small, evaluate the target-Fisher or high-mass route instead.

4. **Estimate the spectral tail.** For the main route, estimate $G = \mathbb{E}G(u)$ and $\Sigma = \mathbb{E}[h(u)h(u)^\mathsf{T}]$, form $\widehat{D}_* = \widehat{G}^{1/2}\Delta_*\widehat{\Sigma}^{1/2}$, and compute $\widehat{T}_r = \sum_{j>r}\sigma_j(\widehat{D}_*)^2$ for every rank of interest. The truncated SVD of $\widehat{D}_*$ gives the corresponding upper-bound candidate in the weighted coordinates. On the supported subspaces, the corresponding query update is $\widehat{M}_r = \widehat{G}^{\dagger/2}(\widehat{D}_*)_r\widehat{\Sigma}^{\dagger/2}$, where $(\widehat{D}_*)_r$ is the rank-r truncated SVD.

5. **Supply and stress-test the constants.** The probability floor, almost-sure geometry bound, and uniform moment constant are assumptions of the population theorem, not quantities certified by a finite calibration set. Use analytic bounds when available, or report L_r over a range of assumed values. Sample estimates are diagnostics, not finite-sample certificates. The upper curve U_r comes from the constructed candidate and does not require these lower-bound constants.

6. **Compare with the desired tolerance.** If $U_r \le \epsilon$, the upper bound exhibits a rank-r candidate within tolerance. If $L_r > \epsilon$, no rank-r candidate in the relevant class can meet the tolerance. The smallest rank certified by the upper curve is sufficient; ranks rejected by the lower curve are impossible under the theorem's assumptions. Any gap between the two remains unresolved.

Example 4: Estimating rank on a given calibration set

Suppose a high-rank query adapter has already been trained, for instance for a text-to-SQL task as in Example 1. A calibration set contains complete prompts with the question, schema, instructions, and SQL prefix available at each attention call. The upper and conditional lower curves divide the tested ranks into three sets: ranks whose constructed candidate meets the KL tolerance, ranks excluded by the conditional lower bound, and ranks for which the two bounds do not decide. The width of the unresolved set depends on the gap between the upper and lower constants and should be reported rather than replaced by a single selected rank.

The population theorems do not provide confidence intervals for these plug-in estimates. A held-out calibration split can test whether the predicted rank–KL curve tracks measured target-to-candidate KL, but it does not turn the current bounds into a finite-sample guarantee. The relevant comparison is against the raw spectrum of Δ_* and activation-only rank criteria: the question is whether task-weighted tails predict attainable attention KL more accurately, not whether a particular optimizer produces a U-shaped validation curve.

10 Related work

LoRA and expressivity. LoRA introduced low-rank updates as a parameter-efficient adaptation mechanism [11]. Zeng and Lee [26] analyze LoRA expressivity, including products of adapted matrices and exact Transformer score matching. We use their bound $\operatorname{rank}(C) \le r_Q + r_K$ in the joint Q/K analysis and then quantify the additional error caused when the best effective score update cannot be realized by factors with the separate rank budgets. Duranthon et al. [6] give a statistical theory of rank-one LoRA fine-tuning in a solvable attention model; their object is learning-curve asymptotics rather than error as a function of rank for a prescribed target. Arunan [12] studies finite-sample estimation and rank selection for empirical risk minimization over rank-constrained LoRA. In a well-specified locally quadratic model, that work derives matching $\widetilde{\Theta}(rd/n)$ estimation rates and an under-ranking bias governed by the raw singular values of the target update. That work studies statistical estimation from finite training data. We hold the target fixed and study the approximation error remaining at each rank, with a spectrum weighted by the task queries and keys and a global error scale determined by softmax.

Weighted approximation. The weighted low-rank approximation theorem is a classic result. [7, 16]. Methods such as DRONE and SVD-LLM use related data-aware decompositions for model compression [4, 23]. CorDA orients a weight decomposition by downstream activation covariance, while EVA initializes and reallocates LoRA rank using activation variance [24, 18]. These methods motivate task-aware approximation directions. Our analysis places the weighted approximation inside the target-to-candidate attention KL and derives both upper and lower bounds, including the global scale ψ needed for small and large score errors.

Task-intrinsic attention rank and low-rank logits. Yoon [25] defines attention-native intrinsic rank as the minimum query–key kernel rank realizing a task and studies controlled deficiency and recovery phenomena. We study the narrower class of LoRA updates of a pretrained head, and a prescribed attention target under KL. Golowich et al. [8] study low-rank approximation of extended language-model logit matrices under average KL, without pretrained LoRA attention or separate Q/K update constraints.

Attention rank and saturation. Attention head dimension and score-matrix rank have been studied as expressivity constraints [3]. Sparse-target cross-entropy and diverging-margin geometry are adjacent to our saturation branch [28]; low pre-softmax rank can also produce high post-softmax rank [15]. Boundary supports of fixed discrete exponential families are classical [5, 20]. Sign rank and rounding rank concern finite, exact realization of prescribed sign or threshold patterns [19, 17], whereas our closure rank allows a sequence of scores to diverge while its softmax converges. Our lower bound uses restricted Fourier identities, columnwise normalization, and the closedness of a bounded-rank matrix set. We do not know a direct implication in either direction between the cited Walsh sign-rank bounds and the L-relative closure-rank formula in Theorem 6.1. Low-rank softmax models have also been studied through argmaxability [9, 10]. Basri and Jacobs [2] show that low-dimensional softmax embeddings can preserve ratios among selected top-token probabilities while off-support mass vanishes. These results establish the broad support-separation and diverging-margin mechanisms. Our saturation theorem makes a different comparison within one explicit multi-context family: the rank required to match finite logits versus the exact softmax closure rank.

Softmax curvature. The Hessian comparison behind our target-Fisher bounds is related to generalized self-concordant analyses of logistic and softmax losses [1, 14]. We specialize this curvature principle through a finite-dimensional Bregman calculation for KL from the target attention to the candidate.

11 Limitations

The theory starts from an available target update. It can assess compression of that target or the rank needed to reproduce it, but it does not prescribe a rank before any target behavior is known.

The bounds are population statements. From a calibration set, one can plug in estimates of the task geometry (G, Σ) and the corresponding spectral tails, although we do not provide finite-sample confidence intervals for the resulting rank–KL curves. Other inputs to the lower bounds—the probability floor, the almost-sure geometry constant, and the uniform moment constant—enter as theorem assumptions rather than as quantities certified by a finite sample, so a practical lower curve must rely on analytic bounds or report sensitivity to those values. The spectral specialization likewise assumes that the target attention is realizable within the stated query-only or joint query/key adapter class; unrestricted fine-tuning may leave an additional approximation error outside that class.

The results describe the best candidate in a rank-constrained class, not the path taken by a training algorithm. They therefore separate representational capacity from optimization but do not guarantee that SGD or another optimizer finds the candidate attaining the upper bound. The constants can also be loose for long, sharply peaked attention vectors. The target-Fisher and high-mass results offer alternatives in that regime, but their own assumptions must still be checked. We have not established that any of these constants are numerically tight on trained, real-model attention heads (for more information on computational checks, see Appendix 12).

Finally, the guarantees concern attention probabilities. Converting them into bounds on layer outputs or task loss requires assumptions about values, output projections, residual paths, feed-forward networks, and later layers. The saturation families isolate a genuine nonlinear effect but are explicit constructions rather than models of typical language data. For joint query/key LoRA, RoPE couples the factors across relative positions, so the single-matrix spectral specialization does not apply directly, although it is worth noting that this particular obstruction is absent in NoPE architectures such as Kimi K3 [13].

12 Conclusion

We studied how much LoRA rank is needed to reproduce a known attention function on inputs from a downstream task. The main message is that this question has a task-dependent answer: it is not settled by the raw rank or singular values of the target weight update alone, but by the part of that update that the task's queries and keys actually activate. Under explicit assumptions, we connect that activated component to the best expected attention KL attainable at each rank. A global softmax comparison turns centered score error into KL, and a downstream-weighted spectral theorem turns the remaining score error into an explicit function of rank through the scale $\psi(t) = \min\{t^2, t\}$. Together, these steps replace rank heuristics and post-hoc sweeps with a two-sided, rank-indexed bracket on representational error at the attention layer itself.

The same framework extends when the standard constants are weak or the adapter is richer than query-only LoRA. Target-Fisher and high-mass lower bounds provide alternative routes when the global probability-floor assumption is uninformative, and an explicit saturation family shows that softmax can separate the rank needed to match finite logits from the rank needed to match the limiting attention function. Fused multi-head analysis accounts for a shared rank budget across heads, and joint query/key LoRA isolates the extra error from factorizing an effective score update into separate low-rank factors.

Given a target adapter and a calibration set drawn from the task, the theory therefore yields upper and lower error curves over rank and makes explicit where the available assumptions certify a rank, exclude it, or leave the decision unresolved.

References

[1] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[2] Ronen Basri and David W. Jacobs. The softmax bottleneck does not limit the probabilities of the most likely tokens. In *International Conference on Learning Representations*, 2026.

[3] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International Conference on Machine Learning*, 2020.

[4] Patrick Chen, Hsiang-Fu Yu, Inderjit S. Dhillon, and Cho-Jui Hsieh. DRONE: Data-aware low-rank compression for large NLP models. In *Advances in Neural Information Processing Systems*, 2021.

[5] Imre Csiszár and František Matúš. Closures of exponential families. *The Annals of Probability*, 2005.

[6] O. Duranthon, F. Boncoraglio, and Lenka Zdeborová. High-dimensional theory of LoRA fine-tuning in a solvable attention model. *arXiv preprint arXiv:2606.05899*, 2026.

[7] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[8] Noah Golowich, Allen Liu, and Abhishek Shetty. Sequences of logits reveal the low rank structure of language models. In *International Conference on Learning Representations*, 2026.

[9] Andreas Grivas, Nikolay Bogoychev, and Adam Lopez. Low-rank softmax can have unargmaxable classes in theory but rarely in practice. *arXiv preprint arXiv:2203.06462*, 2022.

[10] Andreas Grivas, Antonio Vergari, and Adam Lopez. Taming the sigmoid bottleneck: Provably argmaxable sparse multi-label classification. *arXiv preprint arXiv:2310.10443*, 2023.

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

[12] Arunan J. Tight sample complexity for low-rank adaptation: Matching bounds and rank selection. *arXiv preprint arXiv:2607.27680*, 2026.

[13] Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

[14] Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Globally convergent newton methods for ill-conditioned generalized self-concordant losses. In *Advances in Neural Information Processing Systems*, 2019.

[15] Wojciech Masarczyk, Mateusz Ostaszewski, Tin Sum Cheng, Tomasz Trzciński, Aurélien Lucchi, and Răzvan Pascanu. Unpacking softmax: How temperature drives representation collapse, compression, and generalization. *arXiv preprint arXiv:2506.01562*, 2025.

[16] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[17] Stefan Neumann, Rainer Gemulla, and Pauli Miettinen. What you will gain by rounding: Theory and algorithms for rounding rank. *arXiv preprint arXiv:1609.05034*, 2016.

[18] Fabian Paischer, Lukas Hauzenberger, Thomas Schmied, Benedikt Alkin, Marc Peter Deisenroth, and Sepp Hochreiter. Parameter efficient fine-tuning via explained variance adaptation. In *Advances in Neural Information Processing Systems*, 2025.

[19] Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986.

[20] Johannes Rauh, Thomas Kahle, and Nihat Ay. Support sets in exponential families and oriented matroid theory. *International Journal of Approximate Reasoning*, 52(5):613–626, 2011.

[21] John Schulman and Thinking Machines Lab. LoRA without regret. Thinking Machines Lab blog, September 2025. thinkingmachines.ai/blog/lora; accessed 2026-08-18.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[23] Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. SVD-LLM: Truncation-aware singular value decomposition for large language model compression. *arXiv preprint arXiv:2403.07378*, 2024.

[24] Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems*, 2024.

[25] Byeong Hoon Yoon. The entropic bound for transformers: Why static rank fails and attention-native rank recovers. *arXiv preprint arXiv:2607.23050*, 2026.

[26] Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, 2024.

[27] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

[28] Zihao Zhao, Tina Behnia, Ali Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. *arXiv preprint arXiv:2408.15417*, 2024.

Guide to the appendix

Appendices A–D prove the global softmax, moment, spectral, and dependence-allowing bounds. Appendices E and F contain the two alternative lower-bound routes. Appendix G gives the saturation constructions and closure-rank lower bound, and Appendices H–I prove the multi-head and joint query/key extensions.

Computational checks. The accompanying code is available at `https://github.com/gerardpc/lora-rank-project` and verifies the displayed finite constructions, rank calculations, and selected inequalities on small instances. These checks are provided for reproducibility and debugging and are not used in the proofs.

Appendix contents

A Proof of the global softmax bounds

For $z \in \mathbb{R}^n$, write $\mathrm{lse}(z) = \log \sum_i e^{z_i}$ and $H(z) = \mathrm{diag}(p(z)) - p(z)p(z)^\mathsf{T}$, where $p(z) = \mathrm{softmax}(z)$.

Proof of Theorem 3.1. The KL divergence is the log-sum-exp Bregman divergence

$$F(d) = \mathrm{lse}(z_* + d) - \mathrm{lse}(z_*) - \langle p_*, d\rangle. \tag{41}$$

Taylor's formula with integral remainder gives

$$F(d) = \int_0^1 (1-t) d^\mathsf{T} H(z_* + td) d \, dt. \tag{42}$$

Since $\|H(z)\|_{\mathrm{op}} \le 1/2$, $F(d) \le \|d\|_2^2/4$. Also both $\mathrm{lse}(z_* + d) - \mathrm{lse}(z_*)$ and $\langle p_*, d\rangle$ lie in $[\min_i d_i, \max_i d_i]$. Therefore

$$F(d) \le \max_i d_i - \min_i d_i \le \sqrt{2}\|d\|_2. \tag{43}$$

Taking the smaller of the quadratic and linear bounds proves $F(d) \le \psi_{\mathrm{up}}(\|d\|_2)$.

For the lower bound, first suppose $\|d\|_2 \le 1$. For every $t \in [0,1]$,

$$p_i(z_* + td) = \frac{p_{*,i} e^{td_i}}{\sum_j p_{*,j} e^{td_j}} \ge a e^{-2}. \tag{44}$$

For any probability vector q satisfying $\min_i q_i \geq \alpha$ and any centered vector v,

$$v^\mathsf{T}(\mathrm{diag}(q) - qq^\mathsf{T})v = \sum_i q_i(v_i - \mu_q)^2 \tag{45}$$

$$\geq \alpha \sum_i (v_i - \mu_q)^2 \geq \alpha \|v\|_2^2, \tag{46}$$

where $\mu_q = \sum_i q_i v_i$ and the last inequality uses that zero is the Euclidean-optimal centering constant for a centered v. Inserting $\alpha = ae^{-2}$ into (42) yields

$$F(d) \geq \frac{ae^{-2}}{2}\|d\|_2^2, \qquad \|d\|_2 \leq 1. \tag{47}$$

If $\|d\|_2 = s \geq 1$, put $v = d/s$ and $g(t) = F(tv)$. Convexity and $g(0) = 0$ imply that $g(t)/t$ is nondecreasing. Hence

$$F(d) = g(s) \geq sg(1) \geq \frac{ae^{-2}}{2}s. \tag{48}$$

Combining this with (47) proves the lower bound. □

Proof of Corollary 3.2. Apply Theorem 3.1 pointwise to the centered displacement $d_M(u)$, take expectations, and then take the infimum over the same rank-constrained candidate class on both sides. □

B Why a purely quadratic global lower bound is impossible

Proposition B.1 (Fixed-floor rare-context obstruction)**.** *There is a family with target probability floor* $1/3$ *and mean key Gram* $G = I$ *for which the best rank-one quadratic error tends to one while the best rank-one expected attention KL tends to zero.*

Proof. Let $d = 2$, $n = 3$, and choose $K \in \mathbb{R}^{3\times 2}$ with orthonormal columns in $\mathbf{1}^\perp$. Draw

$$h = e_1 \quad \text{with probability } 1-p, \qquad h = Te_2 \quad \text{with probability } p, \qquad T = \sqrt{(1-p)/p}. \tag{49}$$

Take the pretrained query map to be $-I_2$ and the dense target update to be I_2. The target query map is therefore zero, so the target attention is uniform and its floor is $1/3$. The activation covariance is

$$\Sigma = (1-p)e_1e_1^\mathsf{T} + pT^2e_2e_2^\mathsf{T} = (1-p)I_2. \tag{50}$$

The best quadratic rank-one approximation to I_2 has error $1 - p \to 1$.

Choose the rank-one candidate $M = e_1e_1^\mathsf{T}$. It is exact when $h = e_1$. When $h = Te_2$, its centered scores are $-Tq_2$, where q_2 is a centered unit column of K. Thus

$$\mathrm{KL}(\mathrm{unif}_3 \| \,\mathrm{softmax}(-Tq_2)) = \mathrm{lse}(-Tq_2) - \log 3 \leq T. \tag{51}$$

The population KL is at most $pT = \sqrt{p(1-p)} \to 0$. □

C Moment and spectral lemmas

For a deterministic error matrix A, define

$$X_A(u) = \|\beta \Pi_{n(u)} K(u) A h(u)\|_2, \qquad Q(A) = \mathbb{E}X_A(u)^2. \tag{52}$$

Lemma C.1 (Fourth-moment bridge). *If a nonnegative random variable X satisfies $\mathbb{E}X^4 \le \kappa(\mathbb{E}X^2)^2$, then, with $\mu = \mathbb{E}X^2$,*

$$\mathbb{E}\psi(X) \ge \frac{\psi(\sqrt{\mu})}{1+\sqrt{\kappa}}. \tag{53}$$

Proof. For $x \ge 0$, $\psi(x) \ge x^2/(1+x)$. Cauchy–Schwarz gives

$$\mu^2 \le \mathbb{E}\left[\frac{X^2}{1+X}\right] \mathbb{E}[X^2(1+X)]. \tag{54}$$

Moreover, $\mathbb{E}X^3 \le \sqrt{\mathbb{E}X^2\mathbb{E}X^4} \le \sqrt{\kappa}\mu^{3/2}$. Therefore

$$\mathbb{E}\psi(X) \ge \frac{\mu}{1+\sqrt{\kappa\mu}} \ge \frac{\psi(\sqrt{\mu})}{1+\sqrt{\kappa}}. \tag{55}$$

For the last inequality: if $\mu \le 1$, divide by μ and use $\sqrt{\mu} \le 1$; if $\mu \ge 1$, divide by $\sqrt{\mu}$ and use $1/\sqrt{\mu} \le 1$. □

Lemma C.2 (Uniform moment condition). *Assume $G(u)$ and $h(u)$ are independent, that for every deterministic compatible matrix C,*

$$\mathbb{E}\|Ch\|_2^4 \le \kappa_h(\mathbb{E}\|Ch\|_2^2)^2, \tag{56}$$

and that $G(u) \preceq \Lambda G$ almost surely, where $G = \mathbb{E}G(u)$. Then every deterministic A satisfies

$$\mathbb{E}X_A^4 \le \kappa_h\Lambda^2(\mathbb{E}X_A^2)^2. \tag{57}$$

Proof. Condition on $G(u) = G_u$. Independence and (56), applied to $C = G_u^{1/2}A$, give

$$\mathbb{E}_h[X_A^4 \mid G_u] \le \kappa_h\{\mathrm{tr}(G_uA\Sigma A^\mathsf{T})\}^2. \tag{58}$$

Because $A\Sigma A^\mathsf{T} \succeq 0$ and $G_u \preceq \Lambda G$, this is at most $\kappa_h\Lambda^2\{\mathrm{tr}(GA\Sigma A^\mathsf{T})\}^2$. Independence also gives $\mathbb{E}X_A^2 = \mathrm{tr}(GA\Sigma A^\mathsf{T})$. Combine the two identities. □

Lemma C.3 (Singular weighted spectral optimizer). *Let $G, \Sigma \succeq 0$, let Δ_* be any compatible matrix, and put $D_* = G^{1/2}\Delta_*\Sigma^{1/2}$. Then*

$$\inf_{\mathrm{rank}(M)\le r} \|G^{1/2}(M-\Delta_*)\Sigma^{1/2}\|_\mathrm{F}^2 = \sum_{j>r}\sigma_j(D_*)^2 = T_r. \tag{59}$$

This remains true when G or Σ is singular.

Proof. For every feasible M, the matrix $G^{1/2}M\Sigma^{1/2}$ has rank at most r, so Eckart–Young–Mirsky gives the lower bound. Let $(D_*)_r$ be a rank-r truncated SVD and define

$$M_r = G^{\dagger/2}(D_*)_r\Sigma^{\dagger/2}. \tag{60}$$

The left and right singular vectors of $(D_*)_r$ lie in the supports of G and Σ, respectively. Hence $G^{1/2}M_r\Sigma^{1/2} = (D_*)_r$ and $\mathrm{rank}(M_r) \le r$. This candidate attains the tail. □

Proof of Theorem 4.1. Exact dense-target realizability gives $d_M(u) = \beta\Pi_{n(u)}K(u)(M-\Delta_*)h(u)$. Independence yields

$$Q(A) = \|G^{1/2}A\Sigma^{1/2}\|_\mathrm{F}^2. \tag{61}$$

By Lemma C.3, every feasible $A = M - \Delta_*$ satisfies $Q(A) \geq T_r$. Lemma C.2 and Lemma C.1 therefore imply, uniformly over the complete candidate class,

$$\mathbb{E}\psi(X_A) \geq \frac{\psi(\sqrt{Q(A)})}{1 + \Lambda\sqrt{\kappa_h}} \geq \frac{\psi(\sqrt{T_r})}{1 + \Lambda\sqrt{\kappa_h}}. \tag{62}$$

Taking the infimum gives the robust-objective lower bound.

For any nonnegative X,

$$\mathbb{E}\psi_{\mathrm{up}}(X) \leq \min\{\mathbb{E}X^2/4, \sqrt{2}\,\mathbb{E}X\} \leq \psi_{\mathrm{up}}(\sqrt{\mathbb{E}X^2}). \tag{63}$$

Apply this to the explicit weighted-SVD candidate M_r from Lemma C.3, for which $Q(M_r - \Delta_*) = T_r$. Thus $\Phi_r \leq \psi_{\mathrm{up}}(\sqrt{T_r})$. Corollary 3.2 converts the two robust-objective bounds into (14). □

D Dependence-allowing rank–KL bounds

Without assuming independence, retain the definitions

$$Q(A) = \mathbb{E}X_A^2, \qquad S(A) = \|G^{1/2}A\Sigma^{1/2}\|_{\mathrm{F}}^2, \qquad G = \mathbb{E}G(u). \tag{64}$$

Proof of Theorem 4.3. Every feasible A satisfies $S(A) \geq T_r$. Applying Lemma C.1 and (19) gives

$$\mathbb{E}\psi(X_A) \geq \frac{\psi(\sqrt{Q(A)})}{1 + \sqrt{\kappa_r}} \geq \frac{\psi(\sqrt{c_r T_r})}{1 + \sqrt{\kappa_r}}. \tag{65}$$

This holds for the entire candidate class and survives the infimum. For the upper bound, insert A_r^{svd} into (63). Apply Corollary 3.2. □

E Target-Fisher bounds

Proof of Theorem 5.1. Let $q_t = \mathrm{softmax}(z_* + td)$. Coordinatewise,

$$e^{-tR(d)}p_{*,i} \leq q_{t,i} \leq e^{tR(d)}p_{*,i}. \tag{66}$$

For any probability vector q,

$$v^{\mathsf{T}}H(q)v = \min_c \sum_i q_i(v_i - c)^2. \tag{67}$$

Consequently, if $\alpha p_i \leq q_i \leq \gamma p_i$, then $\alpha H(p) \preceq H(q) \preceq \gamma H(p)$. The lower inequality follows by applying the coordinatewise lower bound before minimizing over c; for the upper inequality, evaluate the q-weighted variance at the p-optimal centering constant.

Insert (66) into the Bregman integral

$$\mathrm{KL}(p_* \| \,\mathrm{softmax}(z_* + d)) = \int_0^1 (1-t)d^{\mathsf{T}}H(q_t)d\,dt. \tag{68}$$

The scalar integrals are

$$\int_0^1 (1-t)e^{-tR}dt = \frac{R - 1 + e^{-R}}{R^2}, \qquad \int_0^1 (1-t)e^{tR}dt = \frac{e^R - 1 - R}{R^2}. \tag{69}$$

This proves (22), with the value $1/2$ obtained by continuity at $R = 0$. □

The same integral representations show that c_- is nonincreasing and c_+ is nondecreasing on $[0,\infty)$. For every fixed $t \in [0,1]$, $(1-t)e^{-tR}$ decreases with R, whereas $(1-t)e^{tR}$ increases; integration preserves these inequalities. This justifies the monotonicity step below.

For query-only adaptation, define

$$G_{\mathrm{F}}(u) = \beta^2 K(u)^{\mathsf{T}}[\mathrm{diag}(p_*(u)) - p_*(u)p_*(u)^{\mathsf{T}}]K(u) \tag{70}$$

and $Q_{\mathrm{F}}(A) = \mathbb{E}h^{\mathsf{T}}A^{\mathsf{T}}G_{\mathrm{F}}(u)Ah$. Let $\mathcal{C}_{r,R_0}$ be the rank-r candidates whose score differences from the target have range almost surely at most R_0, and define

$$\mathcal{E}_{r,R_0} = \inf_{M \in \mathcal{C}_{r,R_0}} \mathbb{E}\mathrm{KL}(p_* \| p_M), \quad T^{\mathrm{F}}_{r,R_0} = \inf_{M \in \mathcal{C}_{r,R_0}} Q_{\mathrm{F}}(M - \Delta_*). \tag{71}$$

Corollary E.1 (Constrained Fisher rank–KL bounds)**.** *If $\mathcal{C}_{r,R_0}$ is nonempty, then*

$$c_-(R_0)T^{\mathrm{F}}_{r,R_0} \le \mathcal{E}_{r,R_0} \le c_+(R_0)T^{\mathrm{F}}_{r,R_0}. \tag{72}$$

Proof. Apply Theorem 5.1 pointwise to every admissible candidate and use monotonicity of c_- and c_+. Taking the infimum gives the lower bound. For the upper bound, apply the inequality to an η-minimizing sequence for T^{F}_{r,R_0} and let $\eta \downarrow 0$. □

Under independence, put $G_{\mathrm{F}} = \mathbb{E}G_{\mathrm{F}}(u)$, $D_{\mathrm{F}} = G_{\mathrm{F}}^{1/2}\Delta_*\Sigma^{1/2}$, and $T^{\mathrm{F}}_r = \sum_{j>r}\sigma_j(D_{\mathrm{F}})^2$. Lemma C.3 gives T^{F}_r as the unrestricted rank-r Fisher tail. Since $\mathcal{C}_{r,R_0}$ is a subset of the unrestricted class,

$$\mathcal{E}_{r,R_0} \ge c_-(R_0)T^{\mathrm{F}}_r. \tag{73}$$

If the Fisher-weighted SVD optimizer belongs to $\mathcal{C}_{r,R_0}$, it also attains the constrained quadratic infimum, yielding the matching upper bound $\mathcal{E}_{r,R_0} \le c_+(R_0)T^{\mathrm{F}}_r$.

F High-mass lower bounds

Lemma F.1 (Conditional KL decomposition)**.** *For positive distributions p, q and nonempty S, put $s = p(S)$ and $t = q(S)$. Then*

$$\mathrm{KL}(p\|q) = \mathrm{kl}_{\mathrm{Bern}}(s\|t) + s\mathrm{KL}(p^S\|q^S) + (1-s)\mathrm{KL}(p^{S^c}\|q^{S^c}). \tag{74}$$

If $p = \mathrm{softmax}(z)$ and $q = \mathrm{softmax}(w)$, then $p^S = \mathrm{softmax}(z_S)$ and $q^S = \mathrm{softmax}(w_S)$.

Proof. For $i \in S$, write $\log(p_i/q_i) = \log(s/t) + \log(p_i^S/q_i^S)$ and sum with weights p_i. Repeat on S^c. Conditional softmax follows because the full normalizer cancels after conditioning. □

Proof of Theorem 5.2. By Lemma F.1, the full KL is at least $s_*\mathrm{KL}(p_*^S\|p_M^S)$, where $s_* = p_*(S)$. Apply Theorem 3.1 to the restricted logits. Their target floor is

$$a_S = \min_{i \in S} \frac{p_{*,i}}{s_*}. \tag{75}$$

Thus, for each candidate,

$$\mathrm{KL}(p_*\|p_M) \ge \frac{s_* a_S}{2e^2}\psi(\|\Pi_S(z_M - z_*)_S\|_2). \tag{76}$$

Use $s_* \ge 1-\delta$ and $a_S \ge a_{\mathrm{core}}$, take expectations, and then take the infimum over the complete unrestricted rank-r class. □

For the spectral form, let $K_S(u)$ contain the selected key rows and define

$$G_{\mathrm{core}}(u) = \beta^2 K_S(u)^\top \Pi_{S(u)} K_S(u), \qquad D_{\mathrm{core}} = G_{\mathrm{core}}^{1/2} \Delta_* \Sigma^{1/2}, \tag{77}$$

where $G_{\mathrm{core}} = \mathbb{E} G_{\mathrm{core}}(u)$. Under independence, activation hypercontractivity, and $G_{\mathrm{core}}(u) \preceq \Lambda_{\mathrm{core}} G_{\mathrm{core}}$, the same proof as Theorem 4.1 gives

$$\mathcal{E}_r \geq \frac{(1-\delta) a_{\mathrm{core}}}{2e^2(1 + \Lambda_{\mathrm{core}}\sqrt{\kappa_h})} \psi\Big(\sqrt{T_r^{\mathrm{core}}}\Big), \tag{78}$$

where $T_r^{\mathrm{core}} = \sum_{j>r} \sigma_j(D_{\mathrm{core}})^2$. Under two-sided dependent geometry $\lambda_{\mathrm{core}} \bar{G}_{\mathrm{core}} \preceq G_{\mathrm{core}}(u) \preceq \Lambda_{\mathrm{core}} \bar{G}_{\mathrm{core}}$, apply Theorem 4.3 to obtain

$$\mathcal{E}_r \geq \frac{(1-\delta) a_{\mathrm{core}}}{2e^2(1 + (\Lambda_{\mathrm{core}}/\lambda_{\mathrm{core}})\sqrt{\kappa_h})} \psi\Big(\sqrt{\lambda_{\mathrm{core}} \bar{T}_r^{\mathrm{core}}}\Big). \tag{79}$$

G Saturation constructions and lower bound

For a boundary target matrix P_∞ and an allowed centered score space L, define its softmax closure rank as the least r for which there is a sequence Z_m satisfying $\mathrm{rank}(Z_m) \leq r$, every column of Z_m lies in L, and columnwise softmax converges to P_∞.

G.1 The linear-token achievable construction

Write $k = 3q + s$, $s \in \{0, 1, 2\}$, and choose the least t such that $n = 4^t \geq k + 1$. Identify $\mathbb{F}_2^{2t}$ with $\mathbb{F}_4^t$. The one-dimensional $\mathbb{F}_4$ subspaces partition nonzero vectors into triples $\{a, b, a + b\}$. Select q complete triples and s characters from one further triple. For token row x, set $\chi_a(x) = (-1)^{\langle a, x\rangle}$ and form the normalized character matrix K. Its columns are centered and orthonormal. Context j has $h_j = e_j$, zero pretrained logits, and target logits TKe_j.

At every finite T, centered-softmax injectivity and $K^\top K = I$ force an exact update to equal TI_k, hence to have rank k.

On a complete triple, write the characters as x, y, xy, put $g = T^{1/2}$ and $\alpha = T^{-3/4}$, and define

$$b_2 = (T, T + g, -T)^\top, \quad b_3 = (T, -T, T + g)^\top, \quad b_1 = \alpha(b_2 + b_3), \quad B_T = [b_1\ b_2\ b_3]. \tag{80}$$

B_T has rank two. On the first context's target support $x = 1$, its scores are $2T^{1/4} + 2T^{-1/4}y$, whereas on $x = -1$ the score is $-2T^{1/4}$. The within-support spread vanishes and the support gap diverges. For the second context the score is $xT + y(T + g) - xyT$: it equals $T + g$ on $y = 1$ and is at most $T - g$ on $y = -1$. The third context is symmetric. Scalar $T^{1/4}$ blocks handle leftovers. The block-diagonal update has rank $2q + s = k - \lfloor k/3 \rfloor$.

For the seven-character strengthening, take the seven nonzero characters of $\mathbb{F}_2^3$, set $\rho = T^{-2}$ and $L = T^5$, and define

$$C_\rho = \begin{bmatrix} 1 & \rho^2 & 0 & 0 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho^2 & \rho^2 & 0 & 0 \\ 0 & 0 & 0 & \rho & \rho^2 & 1 & 0 \\ 0 & \rho^2 & 0 & \rho^2 & 0 & 0 & 1 \end{bmatrix} \tag{81}$$

and

$$F_\rho = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -\rho & -\rho^2 & 0 & -\rho^2 \\ 0 & -\rho & -\rho^2 & -\rho^2 \\ -1 & 0 & -\rho^2 & 0 \\ 0 & 0 & -1 & -\rho \\ -\rho^2 & -\rho^2 & -\rho & 0 \\ 0 & -1 & 0 & 0 \\ -\rho^2 & 0 & 0 & -1 \end{bmatrix}, \qquad Z_T = LF_\rho C_\rho. \tag{82}$$

The factorization gives $\operatorname{rank}(Z_T) \le 4$. For completeness, the product before the scalar factor L is

$$F_\rho C_\rho = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\rho & -2\rho^3-\rho^4 & -\rho^2 & -2\rho^4 & -\rho^2-\rho^4 & 0 & -\rho^2 \\ 0 & -\rho^2-\rho^4 & -\rho & -2\rho^3-\rho^4 & -\rho^3-\rho^4 & -\rho^2 & -\rho^2 \\ -1 & -\rho^2 & 0 & -\rho^3 & -\rho-\rho^4 & -\rho^2 & 0 \\ 0 & -\rho^3 & 0 & -\rho-\rho^3 & -\rho^2 & -1 & -\rho \\ -\rho^2 & -\rho^3-\rho^4 & -\rho^2 & -\rho^2-\rho^4 & -2\rho^3-\rho^4 & -\rho & 0 \\ 0 & -\rho & -1 & -\rho^2 & -\rho^2 & 0 & 0 \\ -\rho^2 & -\rho^2-\rho^4 & 0 & -\rho^2 & -\rho^3 & 0 & -1 \end{bmatrix}. \tag{83}$$

In column a, the four rows in the positive halfspace S_a are precisely those with entries between $-3\rho^3$ and 0; every remaining entry is at most $-\rho^2$ for $0 < \rho < 1/3$. Multiplying by L gives, for the positive halfspace S_a of every character a,

$$-3\rho^3 L \le Z_T(x,a) \le 0 \quad (x \in S_a), \qquad Z_T(x,a) \le -\rho^2 L \quad (x \notin S_a). \tag{84}$$

Thus the within-support spread is at most $3/T$ and the support gap is at least $T - 3/T$. Column centering preserves softmax and does not increase rank.

Identify $\mathbb{F}_2^{3t}$ with $\mathbb{F}_8^t$ and pack its nonzero vectors into seven-character one-dimensional $\mathbb{F}_8$ subspaces. If $k = 7m + s$, $0 \le s < 7$, the resulting path has rank at most $4m + s = k - 3\lfloor k/7 \rfloor$ and uses $n < 8(k+1)$ tokens.

It remains to verify KL convergence rather than only weak convergence of supports. For each target halfspace, the target off-support mass is $\epsilon_T = (1 + e^{2T/\sqrt{n}})^{-1}$. Let η_T be the candidate off-support mass. In both constructions, $\eta_T \to 0$, within-support KL vanishes, and the candidate score range is polynomial in T. By Lemma F.1, the only nontrivial term satisfies

$$\epsilon_T \log(\epsilon_T/\eta_T) \le \epsilon_T[-\log \eta_T] = \epsilon_T \operatorname{poly}(T) \longrightarrow 0. \tag{85}$$

The conditional complement contribution obeys the same bound. Taking the better of the triple and septuple packings proves the linear-token statement in Section 6.

G.2 Proof of the exact closure-rank theorem

Proof of Theorem 6.1. Write $k = 3q + s$, $s \in \{0, 1, 2\}$. Work in a binary vector space V whose size is the least power of two above $2k^2 + 2$. We construct

$$\mathcal{A} = T_1 \dot\cup \cdots \dot\cup T_q \dot\cup R, \quad T_\ell = \{a_\ell, b_\ell, a_\ell + b_\ell\}, \quad |R| = s, \tag{86}$$

so these are the only additive triples in $\mathcal{A}$. Given a current set A_j, let $F_j = A_j + A_j$. Choose $u \notin F_j$ and then $v \notin F_j \cup (u + F_j) \cup \{0, u\}$. Fewer than $2k^2 + 2$ vectors are forbidden, so the construction continues. At most two leftovers can be added outside the current sumset. The number of token rows satisfies $n < 4k^2 + 8$.

Use normalized Walsh characters χ_a for $a \in \mathcal{A}$ as centered orthonormal key columns. The finite targets are again TKe_a, so every exact finite-logits update is TI_k and has rank k. Applying

the rank-two path in (80) independently to each T_ℓ, with scalar paths on leftovers, gives closure rank at most $2q + s$.

For the reverse inequality, consider any coefficient sequence M_m whose score softmaxes converge to the target boundary distributions. Absorb the fixed Walsh normalization into M_m. For context a, write

$$f_{m,a}(x) = \sum_{b \in \mathcal{A}} m_{b,a}^{(m)} \chi_b(x), \qquad t_{m,a} = m_{a,a}^{(m)}. \tag{87}$$

On $S_a = \{x : \chi_a(x) = 1\}$, the restrictions of χ_b and χ_{b+a} agree. Orthogonality of the restricted characters and convergence to the uniform target give

$$m_{b,a}^{(m)} + m_{b+a,a}^{(m)} \longrightarrow 0 \quad \text{for } b \notin \{0, a\}. \tag{88}$$

All support scores equal $t_{m,a} + o(1)$, whereas the mean complement score is $-t_{m,a}$. The diverging support gap forces $t_{m,a} \to \infty$.

Triple isolation and (88) make every cross-block coefficient $o(1)$. If $\{a, b, c = a + b\}$ is the block containing a, put $s_{m,a} = m_{b,a}^{(m)}$. Then $m_{c,a}^{(m)} = -s_{m,a} + o(1)$. On the negative halfspace, $\chi_c = -\chi_b$ and both signs occur, so

$$\max_{S_a^c} f_{m,a} = -t_{m,a} + 2|s_{m,a}| + o(1). \tag{89}$$

The diverging gap implies $|s_{m,a}| \le t_{m,a} + o(t_{m,a})$.

Normalize each column by its positive diagonal coefficient:

$$C_m = M_m \operatorname{diag}(t_{m,a}^{-1} : a \in \mathcal{A}). \tag{90}$$

This preserves rank. Diagonal entries become one, cross-block entries converge to zero, and within-block entries remain bounded. Pass to a convergent subsequence. The limit is block diagonal, with singleton blocks [1] and triple blocks of the form

$$B(\alpha, \beta, \gamma) = \begin{bmatrix} 1 & \beta & \gamma \\ \alpha & 1 & -\gamma \\ -\alpha & -\beta & 1 \end{bmatrix}. \tag{91}$$

Every real block (91) has rank at least two. If it had rank one, its 2×2 minors would force $\alpha\beta = 1$, $-\alpha\gamma = 1$, and $\beta\gamma = 1$. The first and third equalities imply $\alpha = \gamma$, contradicting $-\alpha\gamma = 1$ over the reals. The limiting normalized matrix therefore has rank at least $2q + s = k - \lfloor k/3 \rfloor$. Since the set of matrices of rank at most r is closed, every approximating sequence has rank at least $2q + s$. This matches the construction. □

H Fused multi-head proofs

For nonnegative numbers $x_1, \dots, x_H$, put $s = (\sum_h x_h^2)^{1/2}$. We first prove

$$\psi\left(\sqrt{\sum_h x_h^2}\right) \le \sum_h \psi(x_h) \le \sqrt{H}\, \psi\left(\sqrt{\sum_h x_h^2}\right) \tag{92}$$

for all $x_h \ge 0$. If $s \le 1$, every $x_h \le 1$ and $\sum_h \psi(x_h) = s^2 = \psi(s)$. If $s > 1$, then $\psi(x_h) \ge x_h^2/s$: for $x_h \le 1$ this follows from $s \ge 1$, and for $x_h > 1$ it follows from $x_h \le s$. Summing gives the left inequality. Moreover, $\psi(x_h) \le x_h$, so Cauchy–Schwarz gives $\sum_h \psi(x_h) \le \sum_h x_h \le \sqrt{H} s = \sqrt{H}\psi(s)$.

Proof of Theorem 7.1. Applying Theorem 3.1 headwise and summing, then using Equation (92), gives the lower bound

$$\frac{a_{\min}}{2e^2}\psi(X_{\mathrm{MH}}) \le \sum_h D_h, \qquad X_{\mathrm{MH}}^2 = \sum_h X_h^2. \tag{93}$$

For the upper bound, the headwise inequalities give directly

$$\sum_h D_h \le \sum_h \min\{X_h^2/4, \sqrt{2}X_h\} \le \min\{X_{\mathrm{MH}}^2/4, \sqrt{2H}X_{\mathrm{MH}}\}. \tag{94}$$

Both constants in the final aggregation inequality are attained: the quadratic branch when all X_h are small, and the linear branch when the errors are equal across heads. The block definitions in (29) and independence give

$$\mathbb{E}X_{\mathrm{MH}}^2 = \|G_{\mathrm{MH}}^{1/2}(M - \Delta_*^{\mathrm{MH}})\Sigma^{1/2}\|_{\mathrm{F}}^2. \tag{95}$$

Weighted Eckart–Young–Mirsky makes the infimum of this quadratic objective equal to T_r^{MH}. The leverage and fourth-moment assumptions imply $\mathbb{E}X_{\mathrm{MH}}^4 \le \Lambda_{\mathrm{MH}}^2\kappa_h(\mathbb{E}X_{\mathrm{MH}}^2)^2$ by the argument of Lemma C.2. Lemma C.1 therefore gives the robust lower bound

$$\mathbb{E}\psi(X_{\mathrm{MH}}) \ge \frac{\psi(\sqrt{T_r^{\mathrm{MH}}})}{1 + \Lambda_{\mathrm{MH}}\sqrt{\kappa_h}}. \tag{96}$$

For the weighted-SVD candidate, take expectations in Equation (94), use Cauchy–Schwarz on $\mathbb{E}X_{\mathrm{MH}}$, and substitute $\mathbb{E}X_{\mathrm{MH}}^2 = T_r^{\mathrm{MH}}$. This gives the stated upper bound and completes the proof of (31). □

For independently parameterized head adapters, the feasible set is a Cartesian product once integer ranks r_h are fixed, so

$$\inf_{\substack{\sum_h r_h \le R \\ \mathrm{rank}(M_h) \le r_h}} \sum_h \mathbb{E}D_h = \min_{\sum_h r_h \le R} \sum_h \mathcal{E}_{h,r_h}. \tag{97}$$

This proves the allocation identity.

Finally, let $y_{M,h}$ and $y_{*,h}$ be the value-weighted head outputs, let $\mathcal{D}_h(u)$ be the diameter of the head's value vectors, and let O_h be its output-projection block. Total variation and Pinsker give

$$\|y_{M,h} - y_{*,h}\|_2 \le \mathcal{D}_h(u)\,\mathrm{TV}(p_{M,h}, p_{*,h}) \le \mathcal{D}_h(u)\sqrt{D_h/2}. \tag{98}$$

By Cauchy–Schwarz across heads,

$$\left\|\sum_h O_h(y_{M,h} - y_{*,h})\right\|_2^2 \le \frac{H}{2}\sum_h \|O_h\|_{\mathrm{op}}^2 \mathcal{D}_h(u)^2 D_h(u). \tag{99}$$

I Joint query/key proofs

For ordinary dot-product attention, define

$$C(A, B) = K_0^{\mathsf{T}} A + B^{\mathsf{T}}(Q_0 + A). \tag{100}$$

The two summands have ranks at most $\operatorname{rank}(A)$ and $\operatorname{rank}(B)$, proving Equation (33). The bound is sharp: for $p = r$, $d = 2r$, take

$$K_0 = A = [I_r\ 0], \qquad Q_0 = B = [0\ I_r]. \tag{101}$$

Then $C(A, B)$ has rank $2r$.

Every joint candidate also satisfies

$$\operatorname{rank}(K_0^{\mathsf{T}} Q_0 + C(A, B)) = \operatorname{rank}((K_0 + B)^{\mathsf{T}}(Q_0 + A)) \le p. \tag{102}$$

Thus the actual joint class is contained in the effective rank-s class, $s = r_Q + r_K$, and in its width-aware subclass.

Let

$$d_{A,B}(u) = \Pi_{n(u)} \beta X(u)[C(A, B) - C_*]h(u) \tag{103}$$

and let $\Psi^{\mathrm{QK}}_{r_Q, r_K}$ be the infimum of $\mathbb{E}\psi(\|d_{A,B}\|_2)$ over the actual factor class. Applying the lower half of Theorem 3.1 to each candidate gives

$$\frac{a}{2e^2} \Psi^{\mathrm{QK}}_{r_Q, r_K} \le \mathcal{E}^{\mathrm{QK}}_{r_Q, r_K}. \tag{104}$$

Proof of Theorem 8.1. By rank containment, every effective error $C(A, B) - C_*$ is a rank-s candidate relative to C_*. Weighted Eckart–Young–Mirsky therefore gives

$$\mathbb{E}\|d_{A,B}(u)\|_2^2 \ge T_s^{\mathrm{eff}} \tag{105}$$

for every feasible pair. The assumed fourth-moment condition and Lemma C.1 imply

$$\mathbb{E}\psi(\|d_{A,B}\|_2) \ge \frac{\psi(\sqrt{T_s^{\mathrm{eff}}})}{1 + \sqrt{\kappa}}. \tag{106}$$

Take the infimum and use (104) for the lower bound.

Let $\mathcal{S}_s^{\mathrm{svd}}$ be the nonempty set of weighted rank-s optimizers and define

$$\rho_{r_Q, r_K} = \inf_{C_s \in \mathcal{S}_s^{\mathrm{svd}}} \inf_{\substack{\operatorname{rank}(A) \le r_Q \\ \operatorname{rank}(B) \le r_K}} \|G_X^{1/2}[C(A, B) - C_s]\Sigma^{1/2}\|_{\mathrm{F}}. \tag{107}$$

Choose jointly η-minimizing sequences in these two infima. The weighted seminorm triangle inequality gives

$$\left\{\mathbb{E}\|d_{A,B}(u)\|_2^2\right\}^{1/2} \le \sqrt{T_s^{\mathrm{eff}}} + \rho_{r_Q, r_K} + o(1). \tag{108}$$

Apply the pointwise global upper bound and Equation (63), then let $\eta \downarrow 0$. □

The realizability price vanishes in several checkable cases. If $\operatorname{rank}(C) \le r_Q$ and $\operatorname{col}(C) \subseteq \operatorname{col}(K_0^{\mathsf{T}})$, then

$$A = (K_0^{\mathsf{T}})^{\dagger} C, \qquad C(A, 0) = C, \qquad \operatorname{rank}(A) \le \operatorname{rank}(C). \tag{109}$$

The symmetric statement for keys uses $B^{\mathsf{T}} = C Q_0^{\dagger}$ when $\operatorname{row}(C) \subseteq \operatorname{row}(Q_0)$. More generally, suppose $C = C_Q + C_K$, choose A_Q with $K_0^{\mathsf{T}} A_Q = C_Q$ and $\operatorname{rank}(A_Q) \le r_Q$, and assume $\operatorname{row}(C_K) \subseteq \operatorname{row}(Q_0 + A_Q)$ with $\operatorname{rank}(C_K) \le r_K$. Then

$$B_K^{\mathsf{T}} = C_K (Q_0 + A_Q)^{\dagger} \tag{110}$$

has rank at most r_K and satisfies $C(A_Q, B_K) = C_Q + C_K$. These projector identities prove the one-sided and sequential realization claims.

An always-valid factor-dependent upper bound follows from

$$C(A,B) - C(A_*,B_*) = (K_0 + B)^{\mathsf{T}}(A - A_*) + (B - B_*)^{\mathsf{T}}(Q_0 + A_*). \tag{111}$$

Submultiplicativity yields

$$\begin{aligned} &\|G_X^{1/2}[C(A,B) - C_*]\Sigma^{1/2}\|_{\mathrm{F}} && (112)\\ &\quad \le \|G_X^{1/2}(K_0 + B)^{\mathsf{T}}\|_{\mathrm{op}}\|(A - A_*)\Sigma^{1/2}\|_{\mathrm{F}} && (113)\\ &\qquad + \|(Q_0 + A_*)\Sigma^{1/2}\|_{\mathrm{op}}\|G_X^{1/2}(B - B_*)^{\mathsf{T}}\|_{\mathrm{F}}. && (114) \end{aligned}$$

Combining this with (63) and the global softmax upper bound gives the constructive target-factor upper bound. Weighted truncated SVDs of the actual target factors make the two Frobenius errors equal to their corresponding spectral tails.

For fused multi-head joint adaptation, every head block satisfies $\operatorname{rank} C_h(A_h, B_h) \le s$. Applying weighted Eckart–Young headwise and then the aggregate moment bridge gives

$$\mathcal{E}^{\mathrm{MH,QK}}_{r_Q,r_K} \ge \frac{a_{\min}}{2e^2(1 + \sqrt{\kappa_{\mathrm{MH}}})}\psi\left(\sqrt{\sum_h T^{\mathrm{eff}}_{h,s}}\right). \tag{115}$$

This bound deliberately grants every head the full allowance s and is therefore safe but potentially loose.