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

Anonymous authors
Paper under double-blind review

Abstract

We study how accurately a rank-r low-rank adaptation (LoRA) update can approximate a target attention function on downstream inputs. For a fixed pretrained attention head, target attention function, and downstream input distribution, we bound the minimum expected Kullback-Leibler (KL) error achievable by any rank-r query update. Our main results relate this error to the target update's spectral tail, weighted by its effect on downstream attention scores. We also derive alternative bounds under different assumptions on the target attention probabilities and candidate scores. To show that finite-logit rank can overestimate the rank needed to approximate attention, we construct target families that require rank k for exact realization at every finite logit scale, but whose attention probabilities can be approximated arbitrarily well with rank $k - \lfloor k/3 \rfloor$. Finally, we extend the error analysis to fused multi-head LoRA and joint query/key updates, accounting for shared rank budgets and query/key factorization constraints.

1 Introduction

LoRA (Hu et al., 2022) makes it cheaper to fine-tune a pretrained Transformer (Vaswani et al., 2017) by replacing a dense weight update with a low-rank one. The rank of a LoRA adapter controls both the cost of training and the expressive capacity of the adapter, and is therefore an important training hyperparameter. In practice, rank is commonly selected through heuristics or by training adapters of several ranks and comparing downstream performance (Hu et al., 2022; Zhang et al., 2023b; Schulman & Thinking Machines Lab, 2025). Such a sweep reveals what worked in particular optimization runs, but provides limited information about the approximation error inherent to the rank constraint.

This motivates our central question: given a dense or high-rank target adapter, how closely can a rank-r adapter reproduce its attention function on downstream inputs? It is tempting to think that the raw singular values of the target update would answer this question, but a large singular direction may have little effect if the task never activates it, whereas a smaller direction may repeatedly change which tokens receive attention. Moreover, softmax removes common score shifts and saturates at large score gaps. The rank needed to reproduce attention probabilities can therefore differ from the rank suggested by the target weight matrix or its finite logits.

We first consider query LoRA in a single attention head and study the expected KL divergence between the target and rank-r candidate attention distributions. Our results therefore concern attention probabilities rather than head outputs or task loss. Those quantities also depend on the values, output projection, residual connections, subsequent Transformer blocks, and other architectural choices. Figure 1 shows this scope and the two extensions considered later.

Results and contributions. Section 2 defines $\mathcal{E}_r$, the minimum expected KL divergence from the target attention distribution to any rank-r candidate distribution. The paper then follows four steps.

1. Theorem 3.1 relates this KL divergence to the difference between target and candidate attention scores, after removing score shifts that do not affect softmax. Theorem 4.1 then bounds $\mathcal{E}_r$ by a task-weighted spectral tail: the sum of squared singular values beyond the first r after weighting

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 target update by the downstream query and key geometry. Appendix K explains how to estimate the resulting rank–error curves.

2. When very small target attention probabilities make the global lower bound uninformative, Section 5 gives a target-Fisher comparison for rank-r candidates whose attention scores remain close to the target scores, and an unrestricted lower bound on tokens carrying most target mass.
3. To show that finite-logit rank can overestimate the rank needed to approximate attention, Theorem 6.1 constructs target families whose logits require rank k to reproduce exactly, but whose limiting attention probabilities can be approximated arbitrarily well with rank $k - \lfloor k/3 \rfloor$.
4. Theorems 7.1 and 8.1 extend the error bounds to adapters that share one rank budget across multiple heads, and also to simultaneous low-rank query and key adaptation. In the joint query–key setting, some low-rank changes to the attention scores cannot be produced by any query and key updates satisfying their respective rank constraints. Our bounds account for this additional approximation error.

Positioning relative to prior work. Prior work studies adaptive rank allocation during training (Zhang et al., 2023b; Valipour et al., 2023; Zhang et al., 2023a; Paischer et al., 2025), low-rank initialization (Meng et al., 2024), activation-aware approximation (Yang et al., 2024; Chen et al., 2021; Wang et al., 2024), and the transformations that LoRA can express (Zeng & Lee, 2024). We instead fix a pretrained head and a target attention function and bound the best attention KL attainable at every rank on the downstream distribution. The weighted approximation step is classical (Eckart & Young, 1936; Mirsky, 1960); the new step is to connect it to global rank-indexed bounds on attention KL. Section 9 compares the closest results in detail.

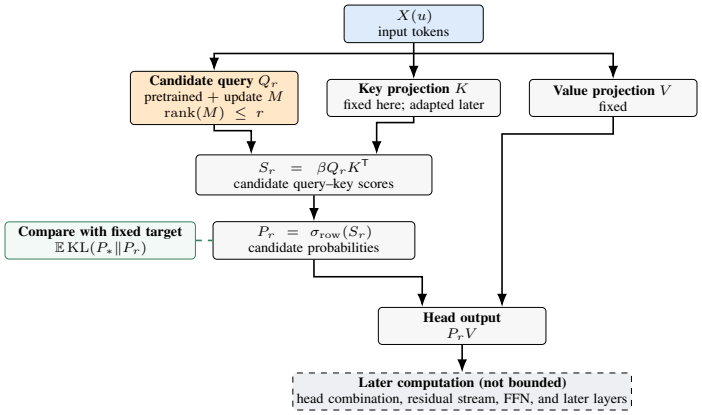


Figure 1: **LoRA in an attention head and the scope of our error bounds.** The main result adapts the query projection of one head while keeping its keys and values fixed. It bounds the expected KL divergence between the resulting candidate probabilities P_r and fixed target probabilities P_*. Here β is the attention scaling factor and σ_{row} applies softmax row by row. The bound ends at the attention probabilities and does not control the head output or later Transformer computation (dashed grey). Section 7 treats an adapter shared across heads, and Section 8 allows both queries and keys to be adapted.

2 Problem Formulation

Let $x = (x_1, \ldots, x_L)$ be a tokenized input. At position t, an attention head assigns probabilities to the token positions visible under the attention mask. We call the resulting probability vector the attention row associated with $u = (x, t)$. As a concrete example of a downstream task, consider a sentiment-classification dataset containing N movie-review texts, where the model predicts "positive" or "negative" from the final position of each text. If $x^{(i)}$ is the i-th tokenized text and L_i is its final position, then $\mathcal{D}$ assigns weight $1/N$ to each pair $(x^{(i)}, L_i)$. The approximation error is the KL divergence between the target attention distribution and the distribution produced by the rank-r update, averaged over the N examples. This quantity measures how closely the update reproduces the target attention distributions. It does not measure agreement in class predictions, output logits,

or task accuracy. More generally, $\mathcal{D}$ can assign different weights to different input texts and query positions. We first derive rank-dependent upper and lower bounds on this approximation error for query-only LoRA in a single attention head. Sections 7 and 8 extend these bounds to multiple heads and joint query/key updates.

Fix $u = (x, t)$, and let $n(u)$ be the number of token positions that t may attend to. Let $h(u) \in \mathbb{R}^{d_h}$ be the hidden representation at position t, before the query projection. Let $K(u) \in \mathbb{R}^{n(u)\times d_k}$ contain the corresponding projected key vectors, one per row. The pretrained head assigns these keys the score vector $z_0(u) \in \mathbb{R}^{n(u)}$ before applying softmax. Here d_h is the hidden-state dimension and d_k is the query/key dimension of the head. Causal and padding masks are incorporated by retaining only the allowed token positions in $K(u)$ and $z_0(u)$.

A rank-r query update is a matrix $M \in \mathbb{R}^{d_k\times d_h}$ satisfying $\operatorname{rank}(M) \le r$ (the matrix usually denoted ΔW_Q in LoRA notation). It maps $h(u)$ to a change $Mh(u)$ in the query vector. Comparing this change with the visible keys adds $\beta K(u)Mh(u)$ to the pretrained attention scores, where $\beta > 0$ is the usual attention scaling factor. The updated scores and probabilities are therefore

$$z_M(u) = z_0(u) + \beta K(u)Mh(u), \qquad p_M(\cdot \mid u) = \operatorname{softmax}(z_M(u)). \tag{1}$$

We evaluate the rank-constrained update M by comparing its attention distribution with a fixed target attention function. For each u, let $z_*(u) \in \mathbb{R}^{n(u)}$ be the target score vector, with one score for each token position that t may attend to, and let $p_*(\cdot \mid u) = \operatorname{softmax}(z_*(u)) \in \mathbb{R}^{n(u)}$ be the corresponding target attention probability vector. Given a rank budget r, we seek one query-update matrix M, shared across all sequences and positions, whose attention distributions are as close as possible to this target. Although M is fixed, its effect depends on the hidden representation and keys in each sampled attention row. We therefore define

$$\boxed{\mathcal{E}_r = \inf_{\operatorname{rank}(M)\le r} \mathbb{E}_{u\sim\mathcal{D}} \mathrm{KL}(p_*(\cdot \mid u) \,\|\, p_M(\cdot \mid u))\,.} \tag{2}$$

Thus $\mathcal{E}_r$ is the infimum of the expected attention KL over all query updates of rank at most r.

We analyze this objective at two levels. First, the score-to-KL results apply to any fixed target score function z_*. To express the error directly in terms of the spectrum of a target update, we then make the stronger assumption that the target itself is generated by a dense query update $\Delta_* \in \mathbb{R}^{d_k\times d_h}$:

$$z_*(u) = z_0(u) + \beta K(u)\Delta_* h(u). \tag{3}$$

This assumption holds when the target differs from the pretrained head only through its query projection. It need not hold for a fully fine-tuned Transformer, which may also change the keys, values, earlier layers, and the representations entering this head.

Our upper and lower bounds bracket the best achievable error $\mathcal{E}_r$ at each rank. For a desired tolerance ϵ, the upper bound can certify that a given rank is sufficient, while the lower bound can rule out ranks that cannot attain error at most ϵ. Appendix K explains how to estimate these rank–error curves from calibration data.

Because adding the same constant to every score does not change the softmax probabilities, we remove this direction before measuring the score error. Let $\mathbf{1} \in \mathbb{R}^n$ be the all-ones vector and define

$$\Pi_n = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad d_M(u) = \Pi_{n(u)}\big(z_M(u) - z_*(u)\big). \tag{4}$$

The vector $d_M(u)$ is therefore the candidate–target score difference after subtracting its mean. We call its Euclidean norm $\|d_M(u)\|_2$ the *centered score error*. Near the target, KL has zero first derivative with respect to the scores, so it grows quadratically with this error. For large score differences, softmax saturates and KL can grow only linearly. Section 3 will bound $\mathcal{E}_r$ above and below using two score-error objectives that capture both regimes. To define them, for $s \ge 0$ set

$$\psi(s) = \min\{s^2, s\}, \qquad \psi_{\mathrm{up}}(s) = \min\{s^2/4, \sqrt{2}\,s\}. \tag{5}$$

We define the optimal lower and upper score surrogates at rank r:

$$\Psi_r = \inf_{\operatorname{rank}(M)\le r} \mathbb{E}_{u\sim\mathcal{D}}\psi(\|d_M(u)\|_2)\,, \qquad \Phi_r = \inf_{\operatorname{rank}(M)\le r} \mathbb{E}_{u\sim\mathcal{D}}\psi_{\mathrm{up}}(\|d_M(u)\|_2)\,. \tag{6}$$

Appendix A summarizes the notation, and Appendix K develops a text-to-SQL example.

3 GLOBAL BOUNDS FOR SOFTMAX KL

We first isolate the effect of softmax from the low-rank constraint. Consider one attention row with target scores $z_* \in \mathbb{R}^n$ and target probabilities $p_* = \mathrm{softmax}(z_*)$. A candidate whose centered score difference is $d \in \mathbb{R}^n$ produces probabilities $\mathrm{softmax}(z_* + d)$. The next theorem bounds their KL divergence using only $\|d\|_2$ and the smallest target probability.

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

$$\frac{a}{2e^2}\psi(\|d\|_2) \le \mathrm{KL}(p_* \,\|\, \mathrm{softmax}(z_* + d)) \le \psi_{\mathrm{up}}(\|d\|_2). \tag{7}$$

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

For small score differences, both bounds are quadratic. For large differences, the global dependence must become linear: a rare input can have a very large score error while making only a linear or bounded contribution to expected KL. Proposition C.1 gives an explicit construction showing why a purely quadratic global lower bound cannot hold, even when every target probability has a fixed positive lower bound.

Corollary 3.2 (Global functional bounds). *For every target attention function, $\mathcal{E}_r \le \Phi_r$. If in addition $\min_i p_{*,i}(u) \ge a > 0$ almost surely, then*

$$\frac{a}{2e^2}\Psi_r \le \mathcal{E}_r \le \Phi_r. \tag{8}$$

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

4 TASK-WEIGHTED SPECTRAL RANK–KL BOUNDS

Section 3 related centered score error to attention KL. We now ask how the rank of a query update controls that score error. Throughout this section, assume that the target is generated by a dense query update Δ_* as in Equation (3). For a candidate M, write $E_M = M - \Delta_*$ for its parameter error. Its centered score error is $\beta\Pi_{n(u)}K(u)E_M h(u)$.

Two matrices describe how the downstream distribution weights this parameter error. Define

$$G(u) = \beta^2 K(u)^\top \Pi_{n(u)} K(u), \qquad \Sigma = \mathbb{E}[h(u)h(u)^\top]. \tag{9}$$

$G(u)$ measures which query directions change the centered scores for the visible keys in row u, while Σ measures which hidden-state directions occur under $\mathcal{D}$. All expectations in this section are over $u \sim \mathcal{D}$.

To obtain a closed-form spectrum, first consider the case in which $G(u)$ and $h(u)$ are independent. Their contributions to the mean squared score error then separate:

$$\mathbb{E}\|\beta\Pi_{n(u)}K(u)E_M h(u)\|_2^2 = \|G^{1/2}E_M\Sigma^{1/2}\|_{\mathrm{F}}^2, \qquad G = \mathbb{E}G(u). \tag{10}$$

This identity shows that the relevant spectrum is not the raw spectrum of Δ_*. It is the spectrum of the downstream-weighted target update

$$D_* = G^{1/2}\Delta_*\Sigma^{1/2}, \qquad T_r = \sum_{\ell > r}\sigma_\ell(D_*)^2. \tag{11}$$

By Equation (10), the mean squared score error of M is $\|G^{1/2}M\Sigma^{1/2} - D_*\|_{\mathrm{F}}^2$. For every rank-$r$ update M, the matrix $G^{1/2}M\Sigma^{1/2}$ also has rank at most r, so Eckart–Young–Mirsky says the smallest attainable value is the singular-value tail T_r. The candidate

$$M_r^{\mathrm{svd}} = G^{\dagger/2}(D_*)_r\Sigma^{\dagger/2}, \qquad E_r^{\mathrm{svd}} = M_r^{\mathrm{svd}} - \Delta_*, \tag{12}$$

attains it: $(D_*)_r$ is supported on the ranges of G and Σ, so $G^{1/2}M_r^{\mathrm{svd}}\Sigma^{1/2} = (D_*)_r$ and the residual Frobenius error is exactly T_r (Lemma D.3). Here $(D_*)_r$ is a rank-r truncated SVD and $\dagger$ denotes

the Moore–Penrose pseudoinverse. The formula remains valid when G or Σ is singular because it uses only the subspaces on which they are supported. Directions ignored by the visible keys or never activated by $h(u)$ disappear from T_r. Thus a large singular direction of Δ_* can contribute nothing, while a smaller direction that repeatedly changes attention can remain. Example 2 in Appendix K makes this contrast explicit.

The spectral tail controls a mean squared error, whereas the global KL bound uses ψ, which becomes linear for large errors. A lower bound therefore also needs to rule out the possibility that almost all squared error is concentrated on extremely rare inputs. We impose the following conditions for constants $\kappa_h \geq 1$ and $\Lambda \geq 1$ and every deterministic matrix $C \in \mathbb{R}^{d_k \times d_h}$:

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

Here $A \preceq B$ means $B - A$ is positive semidefinite (Loewner order), so $G(u) \preceq \Lambda G$ says that no attention row has key geometry larger than Λ times the average G in any direction. The first condition controls the tails of the hidden representations uniformly over directions. Together they keep the squared score error from being concentrated on extremely rare rows.

Theorem 4.1 (Downstream spectral rank–KL bounds)**.** *Suppose the target attention function is generated by a dense query update Δ_* and $G(u)$ and $h(u)$ are independent. Then $\mathcal{E}_r \leq \psi_{\mathrm{up}}(\sqrt{T_r})$. If, in addition, $\min_i p_{*,i}(u) \geq a > 0$ almost surely 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 the weighted-SVD candidate in Equation (12). The lower bound applies to every rank-r candidate. Both bounds scale quadratically when the residual score error is small and linearly when it is large, although their constants can be far apart when the target probability floor a is small.

Given an error tolerance ϵ, any rank satisfying $\psi_{\mathrm{up}}(\sqrt{T_r}) \leq \epsilon$ is sufficient according to the upper bound. Conversely, a rank can achieve error at most ϵ only if the lower expression in Equation (14) is no greater than ϵ. These two tests generally leave an interval of unresolved ranks. Appendix K, with a worked decision in Example 3, explains how to compute this interval and analyze its sensitivity to the constants. In particular, a small probability floor a can make the lower bound weak for long, peaked attention vectors.

The proof is in Appendix D. Ordinary self-attention need not satisfy the independence assumption because $G(u)$ and $h(u)$ come from the same input. Theorem E.1 in Appendix E replaces independence by direct comparisons between the true mean squared score error and the quadratic form defining T_r. It permits arbitrary dependence, but the comparison and moment constants must hold over the rank-constrained candidate class.

5 Which rank–KL bound should be used?

The lower bound of Theorem 4.1 is multiplied by a uniform probability floor $a > 0$ satisfying $\min_i p_{*,i}(u) \geq a$ almost surely. When attention is peaked on a few tokens, the remaining entries can be of order $1/n$ or smaller, so the largest valid a can be tiny and the lower bound weak even if T_r is not. We therefore provide two alternative bounds. The target-Fisher result gives upper and lower bounds when candidate scores remain within a specified range of the target scores. The high-mass result instead gives an unrestricted lower bound by focusing on a target-selected set of tokens that carries most of the probability. Figure 2 shows which result applies, and Table 2 in the appendix compares their conclusions and assumptions.

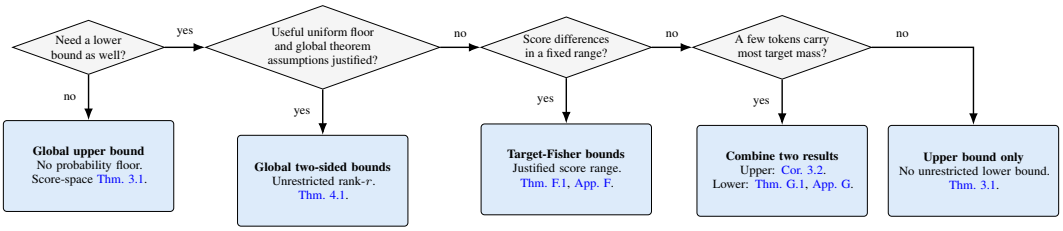


Figure 2: **Which rank–KL theorem to use**. Start at the left and stop at the first result whose assumptions you can justify.

The two alternatives differ in an important way. The target-Fisher result in Appendix F avoids a global probability floor by restricting to candidates whose score differences from the target have range at most some R_0; it therefore lower-bounds the restricted optimum $\mathcal{E}_{r,R_0}$, not the unrestricted $\mathcal{E}_r$. The high-mass result in Appendix G remains an unrestricted lower bound because its token set is chosen from the target before optimizing the candidate, but it provides no matching upper bound. The appendices state the precise conditions and explain when each result can be reduced to a spectral calculation. The Fisher route is most informative when candidate–target score differences can be uniformly bounded by a modest R_0. As R_0 grows, its upper and lower constants move rapidly apart; if no rank-r candidate satisfies the bound, the restricted class is empty. For the high-mass route, a row contributes nothing when its selected core contains only one token, because centering removes every score difference on that core. Its lower bound also does not pair symmetrically with the global upper bound: the two use different downstream-weighted matrices and therefore different spectral tails.

6 Rank reduction under softmax saturation

The preceding bounds start from finite target scores. Softmax can nevertheless converge even when those scores diverge: for example, $\mathrm{softmax}(T, 0)$ and $\mathrm{softmax}(2T, 0)$ both tend to $(1, 0)$ as $T \to \infty$. The limiting probabilities no longer retain the scale of the score gap. We now show that this loss of information can also reduce the rank needed to approximate an entire attention function.

To state this precisely, let P_∞ collect the limiting target probabilities across a finite set of contexts, and let L be the subspace of centered score vectors reachable through the fixed keys. The *L-relative softmax closure rank*, denoted $r_{\mathrm{cl}}^{L}(P_\infty)$, is the smallest r for which a sequence of rank-r score matrices, with every column in L, converges to P_∞ after applying softmax to each column. This definition allows the scores to diverge while their probabilities converge, but it preserves the score-space restriction imposed by query-only LoRA.

Theorem 6.1 (Exact closure rank of an isolated-triple Walsh family)**.** *For every $k \geq 3$, there is a query-only attention family, constructed from Walsh characters, with k contexts and fewer than $4k^2 + 8$ token positions. At every finite score scale, reproducing its target scores exactly requires update rank k. Let L_k be the centered score space reachable through the keys of this construction. For its limiting target probabilities $P_\infty^{(k)}$, the exact relative closure rank is*

$$r_{\mathrm{cl}}^{L_k}(P_\infty^{(k)}) = k - \lfloor k/3 \rfloor. \tag{15}$$

This equality is for the isolated-triple family constructed in Appendix H; other families can have smaller closure rank.

The construction divides the k contexts into groups of three. Each group can be approximated along a sequence of rank-two score matrices, which proves the upper bound. Restricted Fourier identities then show that no limiting score sequence can use rank below two per group. Saturation therefore gives a constant-factor reduction for this family, but not a sublinear rank.

Appendix H also gives a different, linear-token family that achieves ratio $4/7$ on complete seven-context blocks. That construction is an upper bound for another family, not an exact closure-rank result, so it is kept separate from Theorem 6.1.

7 FUSED MULTI-HEAD LORA

Multi-head implementations often apply LoRA to one fused query projection whose row blocks correspond to the individual heads. The rank constraint then applies to the stacked update, not to each head separately. A single fused rank direction may therefore contribute to several heads.

Let the heads be indexed by $j = 1, \dots, H$. Head j has visible-key matrix $K_j(u) \in \mathbb{R}^{n_j(u) \times d_{k,j}}$, attention scale β_j, and update block $M_j \in \mathbb{R}^{d_{k,j} \times d_h}$. All heads receive the same hidden representation $h(u) \in \mathbb{R}^{d_h}$. Write $d_{\text{stack}} = \sum_{j=1}^{H} d_{k,j}$ for the number of rows in the fused projection. We stack the update blocks vertically,

$$M = (M_1^\mathsf{T}, \dots, M_H^\mathsf{T})^\mathsf{T}, \qquad \operatorname{rank}(M) \le r, \tag{16}$$

and assume that the target is generated by a dense stacked update Δ_*^{MH} with blocks $\Delta_{*,j}$. Applying the query-only model from Section 2 to each head gives target probabilities $p_{*,j}(\cdot \mid u)$, candidate probabilities $p_{M,j}(\cdot \mid u)$, and centered score errors

$$d_{M,j}(u) = \beta_j \Pi_{n_j(u)} K_j(u)(M_j - \Delta_{*,j})h(u). \tag{17}$$

We measure the sum of the headwise KL divergences and optimize it under the single fused rank budget:

$$\mathcal{E}_r^{\text{MH}} = \inf_{\operatorname{rank}(M) \le r} \mathbb{E}_{u \sim \mathcal{D}} \sum_{j=1}^{H} \mathrm{KL}(p_{*,j}(\cdot \mid u) \| p_{M,j}(\cdot \mid u)). \tag{18}$$

The corresponding key geometry is block diagonal because each score error is computed within one head:

$$G_{\text{MH}}(u) = \operatorname{blockdiag}\left(\beta_j^2 K_j(u)^\mathsf{T} \Pi_{n_j(u)} K_j(u) : j = 1, \dots, H\right), \quad G_{\text{MH}} = \mathbb{E} G_{\text{MH}}(u), \tag{19}$$

$$\Sigma = \mathbb{E}[h(u)h(u)^\mathsf{T}]. \tag{20}$$

The fused weighted spectral tail is then

$$T_r^{\text{MH}} = \sum_{\ell > r} \sigma_\ell \left(G_{\text{MH}}^{1/2} \Delta_*^{\text{MH}} \Sigma^{1/2} \right)^2. \tag{21}$$

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

$$\mathbb{E}\|Ch(u)\|_2^4 \le \kappa_h \left(\mathbb{E}\|Ch(u)\|_2^2\right)^2 \quad \textit{for every deterministic } C \in \mathbb{R}^{d_{\text{stack}} \times d_h}. \tag{22}$$

Then

$$\frac{a_{\min}}{2e^2(1 + \Lambda_{\text{MH}}\sqrt{\kappa_h})} \psi\left(\sqrt{T_r^{\text{MH}}}\right) \le \mathcal{E}_r^{\text{MH}} \le \min\{T_r^{\text{MH}}/4, \sqrt{2HT_r^{\text{MH}}}\}. \tag{23}$$

The factor $\sqrt{H}$ appears only when the upper bound is on its linear branch. When the residual score errors are small, the headwise quadratic errors add without an additional factor depending on the number of heads. The global lower constant can nevertheless be weak if even one head assigns an extremely small target probability to some token. In that case, headwise Fisher or high-mass bounds can be combined instead. If each head instead has its own adapter of rank r_h, fixing those ranks separates the optimization across heads. Allocating a total rank budget R then reduces to 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 I.

Converting score error to KL does not need independence of $G_{\text{MH}}(u)$ and $h(u)$. Independence is used in Theorem 7.1 only to obtain the closed-form spectral tail T_r^{MH}. If $G_{\text{MH}}(u)$ and $h(u)$ are dependent, Appendix I still yields the same style of bound, provided that for every rank-r update the true mean squared score error is comparable to $\|G_{\text{MH}}^{1/2} E \Sigma^{1/2}\|_{\text{F}}^2$ (the quadratic form behind T_r^{MH}).

8 JOINT QUERY/KEY LORA

We close the analysis by extending the same rank–KL bounds from query-only updates to simultaneous query and key adaptation. Practical LoRA may update both projections at once. Let d be the hidden dimension and p the query and key dimension of the head. Write $Q_0, K_0 \in \mathbb{R}^{p\times d}$ for the pretrained projection matrices and $A, B \in \mathbb{R}^{p\times d}$ for their respective updates, with $\mathrm{rank}(A) \le r_Q$ and $\mathrm{rank}(B) \le r_K$.

Attention scores depend on the product of the key and query projections. The change to this product is the effective score-update matrix

$$C(A,B) = (K_0+B)^\top(Q_0+A) - K_0^\top Q_0 = K_0^\top A + B^\top(Q_0+A). \tag{24}$$

Although $C(A,B)$ contains the bilinear term $B^\top A$, it can be grouped with the key contribution as shown above. Consequently, as observed by Zeng & Lee (2024),

$$\mathrm{rank}(C(A,B)) \le r_Q + r_K. \tag{25}$$

For a sampled attention row u, let $X(u) \in \mathbb{R}^{n(u)\times d}$ contain the hidden representations of the visible tokens, one per row, and let $h(u) \in \mathbb{R}^d$ be the hidden representation at the query position. The effective matrix changes the scores by $\beta X(u)C(A,B)h(u)$. For prescribed target updates (A_*, B_*), define $C_* = C(A_*, B_*)$. The centered candidate–target score error is

$$d_{A,B}(u) = \beta \Pi_{n(u)} X(u)[C(A,B) - C_*]h(u). \tag{26}$$

Let $p_{A,B}(\cdot \mid u)$ and $p_*(\cdot \mid u)$ be the candidate and target attention distributions obtained from these scores. The best error under the two separate rank budgets is

$$\mathcal{E}^{\mathrm{QK}}_{r_Q,r_K} = \inf_{\substack{\mathrm{rank}(A)\le r_Q \\ \mathrm{rank}(B)\le r_K}} \mathbb{E}_{u\sim\mathcal{D}} \mathrm{KL}(p_*(\cdot \mid u) \| p_{A,B}(\cdot \mid u)). \tag{27}$$

As before, the downstream token and query geometry is summarized by

$$G_X(u) = \beta^2 X(u)^\top \Pi_{n(u)} X(u), \qquad G_X = \mathbb{E}G_X(u), \qquad \Sigma = \mathbb{E}[h(u)h(u)^\top]. \tag{28}$$

If $G_X(u)$ and $h(u)$ are independent, the mean squared score error equals $\|G_X^{1/2}[C(A,B) - C_*]\Sigma^{1/2}\|_{\mathrm{F}}^2$. Because every feasible $C(A,B)$ has rank at most $s = r_Q + r_K$, we first relax the problem to all effective matrices of rank at most s. The weighted spectral tail of the target in this larger class is

$$D_*^{\mathrm{eff}} = G_X^{1/2} C_* \Sigma^{1/2}, \qquad T_s^{\mathrm{eff}} = \sum_{\ell > s} \sigma_\ell(D_*^{\mathrm{eff}})^2, \qquad s = r_Q + r_K. \tag{29}$$

This relaxation gives a lower bound, but its optimal effective matrix need not factor into query and key updates satisfying the separate rank budgets. Let $\mathcal{S}_s^{\mathrm{svd}}$ be the set of rank-s effective matrices attaining T_s^{eff}. We measure the distance from this set to the matrices that the actual query/key factors can realize by

$$\rho_{r_Q,r_K} = \inf_{C_s \in \mathcal{S}_s^{\mathrm{svd}}} \inf_{\substack{\mathrm{rank}(A)\le r_Q \\ \mathrm{rank}(B)\le r_K}} \|G_X^{1/2}[C(A,B) - C_s]\Sigma^{1/2}\|_{\mathrm{F}}. \tag{30}$$

Thus $\rho_{r_Q,r_K} = 0$ whenever at least one optimal effective matrix is realizable by the permitted factors. A positive value is the additional price of enforcing the query/key factorization. The relaxed problem allows any rank-s matrix C, but a width-p attention head imposes the additional constraint $\mathrm{rank}(K_0^\top Q_0 + C) \le p$. The relaxed class can therefore contain effective updates that no width-p query/key factors can realize. Optimizing over this larger class preserves the lower bound, but can make it loose.

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)*, there is a constant* $\kappa \ge 1$ *such that*

$$\mathbb{E}\|d_{A,B}(u)\|_2^4 \le \kappa\big(\mathbb{E}\|d_{A,B}(u)\|_2^2\big)^2. \tag{31}$$

Then

$$\frac{a}{2e^2(1+\sqrt{\kappa})}\psi\Big(\sqrt{T_s^{\mathrm{eff}}}\Big) \le \mathcal{E}^{\mathrm{QK}}_{r_Q,r_K} \le \psi_{\mathrm{up}}\left(\sqrt{T_s^{\mathrm{eff}}} + \rho_{r_Q,r_K}\right). \tag{32}$$

The lower bound follows because every feasible query/key pair belongs to the larger effective rank-s class. The converse is false: not every rank-s effective matrix can be produced under the separate query and key budgets. The upper bound therefore includes ρ_{r_Q,r_K}. This term vanishes under the one-sided and sequential realizability conditions in Appendix J; otherwise it may be large and is generally nonconvex to compute. The appendix also gives a directly computable, but potentially looser, certificate obtained by approximating the target query and key factors separately. Dependence between $X(u)$ and $h(u)$ requires candidate-class comparisons as in Theorem E.1. With RoPE, the pointwise rank containment still holds for each relative-position slice, but the single-matrix spectral tail above does not apply automatically; Appendix J gives the details.

9 Related work

LoRA rank and expressivity. Adaptive methods allocate or vary rank during training (Zhang et al., 2023b; Valipour et al., 2023; Zhang et al., 2023a; Paischer et al., 2025), while PiSSA changes the initialization (Meng et al., 2024). These methods choose how to train when the target is unknown; we fix a target and isolate the error imposed by rank. Zeng & Lee (2024) study exact LoRA expressivity; we use their joint Q/K rank containment and quantify the error caused by separate factor budgets. Statistical analyses instead address learning curves (Duranthon et al., 2026).

Task-weighted approximation. Weighted low-rank approximation is classical (Eckart & Young, 1936; Mirsky, 1960). Data-aware decompositions also appear in model compression (Chen et al., 2021; Wang et al., 2024), while CorDA and EVA use activation statistics to orient or allocate LoRA rank (Yang et al., 2024; Paischer et al., 2025). We instead connect weighted approximation to attention KL in the quadratic and saturated regimes.

Attention rank and softmax limits. Prior work studies score-matrix expressivity (Bhojanapalli et al., 2020), task-intrinsic attention rank (Yoon, 2026), KL approximation of language-model logits (Golowich et al., 2026), and pre- versus post-softmax rank (Masarczyk et al., 2025). Related theories cover closures of exponential families (Csiszár & Matúš, 2005; Rauh et al., 2011), diverging margins (Zhao et al., 2024), and threshold notions of rank (Paturi & Simon, 1986; Neumann et al., 2016; Grivas et al., 2022; 2023). Unlike the probability-ratio construction of Basri & Jacobs (2026), we compute an exact closure rank and compare it with finite-logit rank. The Fisher argument is related to self-concordant loss analyses (Bach, 2010; Marteau-Ferey et al., 2019).

10 Limitations

The bounds assume that the target behavior is known beforehand. The spectral results also assume a dense query-only or jointly factorized Q/K target update. Thus, they quantify how well a specified target can be approximated, not how to choose rank before the target is known. Estimates of G, Σ, and the spectral tails can guide rank selection, but our guarantees are population-level: they give no finite-sample confidence intervals and do not verify the probability, geometry, or moment assumptions. We study representational capacity, not optimization, training dynamics, initialization, or sample complexity. For peaked attention, the constants may be pessimistic; the Fisher and high-mass alternatives replace the global floor a with other restrictions. Finally, our error metric concerns attention probabilities, not their propagation through later layers or to task loss, which depends on architecture-specific details. The saturation examples show only that finite-logit rank can exceed rank after saturation, not that this is typical of language-model data.

11 Conclusion

LoRA rank is not an intrinsic property of a weight update alone. It depends on how the target update interacts with the downstream input distribution and the invariances and saturation of softmax. This suggests viewing rank selection as a task-dependent approximation problem: preserve the update directions that matter for attention on downstream inputs, rather than those that are merely large in the raw parameter spectrum. Our results provide a theoretical basis for this view across single-head, multi-head, and joint query/key adaptation. Turning it into a practical rank-selection method will require finite-sample guarantees and an account of optimization and downstream task loss.

AI USE STATEMENT

Generative AI tools were used to help develop and refine the mathematical claims, assist with proofs, check related literature, write numerical verification code and edit the manuscript. All AI-assisted content was reviewed by the authors. We take responsibility for the final text, proofs, code, and figures.

ETHICS STATEMENT

This work is theoretical and does not involve human participants, personal or private data, or deployment of machine learning models. We do not identify any ethical risks beyond those associated with the development and evaluation of machine-learning methods.

REPRODUCIBILITY STATEMENT

The main text states the assumptions and conclusions of each result, and Appendices B–J contain complete proofs. Appendix K describes the calibration workflow and separates plug-in estimates from population assumptions. The supplementary code reproduces the constructions, rank calculations, and numerical checks. The paper contains no empirical experiments or dataset-dependent numerical claims.

REFERENCES

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010. doi: 10.1214/09-EJS521.

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. URL https://openreview.net/forum?id=DgJqQk6y19.

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. URL https://proceedings.mlr.press/v119/bhojanapalli20a.html.

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. URL https://proceedings.neurips.cc/paper/2021/hash/f56de5ef149cf0aedcc8f4797031e229-Abstract.html.

Imre Csiszár and František Matúš. Closures of exponential families. *The Annals of Probability*, 2005. URL https://arxiv.org/abs/math/0503653.

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. URL https://arxiv.org/abs/2606.05899.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

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. URL https://arxiv.org/abs/2510.24966.

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. URL https://arxiv.org/abs/2203.06462.

Andreas Grivas, Antonio Vergari, and Adam Lopez. Taming the sigmoid bottleneck: Provably argmaxable sparse multi-label classification. *arXiv preprint arXiv:2310.10443*, 2023. URL https://arxiv.org/abs/2310.10443.

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. URL https://arxiv.org/abs/2106.09685.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

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. URL https://arxiv.org/abs/1907.01771.

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. URL https://arxiv.org/abs/2506.01562.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.02948.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50.

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. URL https://arxiv.org/abs/1609.05034.

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. URL https://arxiv.org/abs/2410.07170.

Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986. doi: 10.1016/0022-0000(86)90046-2.

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. doi: 10.1016/j.ijar.2011.02.002.

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

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2210.07558.

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.

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. URL https://arxiv.org/abs/2403.07378.

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. URL https://arxiv.org/abs/2406.05223.

Byeong Hoon Yoon. The entropic bound for transformers: Why static rank fails and attention-native rank recovers. *arXiv preprint arXiv:2607.23050*, 2026. URL https://arxiv.org/abs/2607.23050.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.17513.

Feiyu Zhang, Liangzhi Li, Junhao Chen, Zhouqiang Jiang, Bowen Wang, and Yiming Qian. IncreLoRA: Incremental parameter allocation method for parameter-efficient fine-tuning. *arXiv preprint arXiv:2308.12043*, 2023a. URL https://arxiv.org/abs/2308.12043.

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*, 2023b. URL https://arxiv.org/abs/2303.10512.

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. URL https://arxiv.org/abs/2408.15417.

GUIDE TO THE APPENDIX

Appendix A summarizes the notation used in the bounds. Appendices B–E prove the global softmax, moment, spectral, and dependence-allowing bounds. Appendices F and G contain the two alternative lower-bound routes. Appendix H gives the saturation constructions and closure-rank lower bound, and Appendices I–J prove the multi-head and joint query/key extensions. Appendix K collects the practical workflow and worked examples.

Computational checks. The accompanying code (available as a zip file in the Supplementary material) verifies the displayed finite constructions, rank calculations, and selected inequalities on small instances. These checks are provided for reproducibility and debugging, but are not used in the proofs.

APPENDIX CONTENTS

A. Notation glossary
B. Proof of the global softmax bounds
C. Why a purely quadratic global lower bound is impossible
D. Moment and spectral lemmas
E. Dependence-allowing rank–KL bounds
F. Target-Fisher bounds
G. High-mass lower bounds
H. Saturation constructions and lower bound
I. Fused multi-head proofs
J. Joint query/key proofs
K. Using the bounds on downstream data

A NOTATION GLOSSARY

Table 1 collects the notation used for the query-only rank–KL bounds. Quantities that depend on u are evaluated separately for each downstream attention row; Δ_* and M are single matrices shared across all rows.

Table 1: Notation for the target–candidate comparison and its spectral bounds.

Symbol	Definition
$u = (x, t) \sim \mathcal{D}$	A downstream sequence x and query position t, drawn from the evaluation distribution $\mathcal{D}$.
$n(u)$	Number of key positions visible from the query position in row u.
$h(u)$	Query-side activation entering the adapted query projection.
$K(u)$	Matrix of pretrained key vectors at the visible positions.
β	Attention-score scaling factor.
$\Pi_{n(u)}$	Projection that subtracts the mean of a score vector, removing the softmax-invariant constant direction.
$z_0(u)$	Attention scores produced by the pretrained head.
Δ_*	Fixed dense or high-rank query update that generates the target attention function.
$z_*(u)$, $p_*(\cdot \mid u)$	Target scores and their softmax probabilities.
M	One candidate query update, constrained by $\mathrm{rank}(M) \le r$ and used for every u.
$z_M(u)$, $p_M(\cdot \mid u)$	Candidate scores and their softmax probabilities.
$d_M(u)$	Centered score difference $\Pi_{n(u)}(z_M(u) - z_*(u))$.
$\mathcal{E}_r$	Smallest population expected KL from p_* to p_M over all candidates with rank at most r.
$G(u)$, G	Rowwise centered key Gram $\beta^2 K(u)^\mathsf{T} \Pi_{n(u)} K(u)$ and its mean $G = \mathbb{E}G(u)$.
Σ	Activation second moment $\mathbb{E}[h(u)h(u)^\mathsf{T}]$.
D_*	Downstream-weighted target $G^{1/2}\Delta_*\Sigma^{1/2}$ under the separable geometry.
T_r	Squared singular-value tail $\sum_{\ell > r} \sigma_\ell(D_*)^2$.
ψ, ψ_{up}	Robust lower and upper score-error functions defined in Equation (5).
a	Almost-sure lower bound on the target probabilities in the global lower bound.
κ_h, Λ	Uniform activation-moment and rowwise key-geometry constants from Equation (13).

B 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

$$F(d) = \mathrm{lse}(z_* + d) - \mathrm{lse}(z_*) - \langle p_*, d \rangle. \tag{33}$$

Taylor's formula with integral remainder gives

$$F(d) = \int_0^1 (1 - t) d^\mathsf{T} H(z_* + td) d \, dt. \tag{34}$$

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

$$F(d) \le \max_i d_i - \min_i d_i \le \sqrt{2}\|d\|_2. \tag{35}$$

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 $t \in [0, 1]$,

$$p_i(z_* + td) = \frac{p_{*,i} e^{td_i}}{\sum_j p_{*,j} e^{td_j}} \ge a e^{-2}. \tag{36}$$

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

$$v^\mathsf{T}(\mathrm{diag}(q) - qq^\mathsf{T})v = \sum_i q_i (v_i - \mu_q)^2 \tag{37}$$

$$\ge \alpha \sum_i (v_i - \mu_q)^2 \ge \alpha \|v\|_2^2, \tag{38}$$

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 (34) yields

$$F(d) \geq \frac{ae^{-2}}{2}\|d\|_2^2, \qquad \|d\|_2 \leq 1. \tag{39}$$

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

$$F(d) = g(s) \geq sg(1) \geq \frac{ae^{-2}}{2}s. \tag{40}$$

Combining this with (39) 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. □

C Why a purely quadratic global lower bound is impossible

Proposition C.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

$$h = e_1 \quad \text{with probability } 1-p, \qquad h = Te_2 \quad \text{with probability } p, \qquad T = \sqrt{(1-p)/p}. \tag{41}$$

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

$$\Sigma = (1-p)e_1e_1^\mathsf{T} + pT^2e_2e_2^\mathsf{T} = (1-p)I_2. \tag{42}$$

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

$$\mathrm{KL}(\mathrm{unif}_3 \| \operatorname{softmax}(-Tq_2)) = \operatorname{lse}(-Tq_2) - \log 3 \leq T. \tag{43}$$

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

D Moment and spectral lemmas

For a deterministic error matrix A, define

$$X_A(u) = \|\beta\Pi_{n(u)}K(u)Ah(u)\|_2, \qquad Q(A) = \mathbb{E}X_A(u)^2. \tag{44}$$

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

$$\mathbb{E}\psi(X) \geq \frac{\psi(\sqrt{\mu})}{1+\sqrt{\kappa}}. \tag{45}$$

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

$$\mu^2 \leq \mathbb{E}\left[\frac{X^2}{1+X}\right]\mathbb{E}[X^2(1+X)]. \tag{46}$$

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

$$\mathbb{E}\psi(X) \geq \frac{\mu}{1+\sqrt{\kappa\mu}} \geq \frac{\psi(\sqrt{\mu})}{1+\sqrt{\kappa}}. \tag{47}$$

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

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

$$\mathbb{E}\|Ch\|_2^4 \le \kappa_h(\mathbb{E}\|Ch\|_2^2)^2, \tag{48}$$

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

$$\mathbb{E}X_A^4 \le \kappa_h\Lambda^2(\mathbb{E}X_A^2)^2. \tag{49}$$

Proof. Condition on $G(u) = G_u$. Independence and (48), applied to $C = G_u^{1/2}A$, give

$$\mathbb{E}_h[X_A^4 \mid G_u] \le \kappa_h\{\mathrm{tr}(G_uA\Sigma A^\mathsf{T})\}^2. \tag{50}$$

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 D.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*

$$\inf_{\mathrm{rank}(M)\le r} \|G^{1/2}(M-\Delta_*)\Sigma^{1/2}\|_\mathrm{F}^2 = \sum_{j>r}\sigma_j(D_*)^2 = T_r. \tag{51}$$

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

$$M_r = G^{\dagger/2}(D_*)_r\Sigma^{\dagger/2}. \tag{52}$$

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

$$Q(A) = \|G^{1/2}A\Sigma^{1/2}\|_\mathrm{F}^2. \tag{53}$$

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

$$\mathbb{E}\psi(X_A) \ge \frac{\psi(\sqrt{Q(A)})}{1+\Lambda\sqrt{\kappa_h}} \ge \frac{\psi(\sqrt{T_r})}{1+\Lambda\sqrt{\kappa_h}}. \tag{54}$$

Taking the infimum gives the robust-objective lower bound.

For any nonnegative X,

$$\mathbb{E}\psi_{\mathrm{up}}(X) \le \min\{\mathbb{E}X^2/4, \sqrt{2}\,\mathbb{E}X\} \le \psi_{\mathrm{up}}(\sqrt{\mathbb{E}X^2}). \tag{55}$$

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

E Dependence-allowing rank–KL bounds

The closed-form tail T_r in the main theorem uses independence to separate the key geometry from the query activation. When these quantities are dependent, we compare the true coupled mean squared error with the same reference quadratic form. For an error matrix $E = M - \Delta_*$, define

$$X_E(u) = \|\beta\Pi_{n(u)}K(u)Eh(u)\|_2, \qquad Q(E) = \mathbb{E}X_E(u)^2, \tag{56}$$

$$S(E) = \|G^{1/2}E\Sigma^{1/2}\|_\mathrm{F}^2, \qquad G = \mathbb{E}G(u). \tag{57}$$

Here $Q(E)$ is the actual mean squared score error under the joint downstream distribution, whereas $S(E)$ is the quadratic form obtained from the separate averages G and Σ. Let E_r^{svd} be the error of the weighted-SVD candidate in Equation (12).

Theorem E.1 (Dependence-allowing spectral bounds). *Suppose, for every feasible $E = M - \Delta_*$,*

$$Q(E) \geq c_r S(E), \qquad \mathbb{E}\|\beta \Pi_{n(u)} K(u) E h(u)\|_2^4 \leq \kappa_r Q(E)^2, \tag{58}$$

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

$$\frac{a}{2e^2(1+\sqrt{\kappa_r})} \psi(\sqrt{c_r T_r}) \leq \mathcal{E}_r \leq \psi_{\mathrm{up}}(\sqrt{C_r^+ T_r}). \tag{59}$$

This result allows arbitrary dependence between keys and queries. Its main assumption is the average candidate-class comparison $Q(E) \geq c_r S(E)$, which may hold even when every individual $G(u)$ is singular. A simpler but much stronger sufficient condition is $\lambda G \preceq G(u) \preceq \Lambda G$ almost surely. It requires every context to expose every direction present in G and cannot hold with $\lambda > 0$ and full-rank G when $n(u) - 1 < d_k$. Together with Equation (13), it gives $c_r = \lambda$, $C_r^+ = \Lambda$, and $\kappa_r = \kappa_h(\Lambda/\lambda)^2$.

Proof of Theorem E.1. Every feasible E satisfies $S(E) \geq T_r$. Applying Lemma D.1 and (58) gives

$$\mathbb{E}\psi(X_E) \geq \frac{\psi(\sqrt{Q(E)})}{1+\sqrt{\kappa_r}} \geq \frac{\psi(\sqrt{c_r T_r})}{1+\sqrt{\kappa_r}}. \tag{60}$$

This holds for the entire candidate class and survives the infimum. For the upper bound, insert E_r^{svd} into (55). Apply Corollary 3.2. □

F Target-Fisher bounds

For one target attention distribution p_*, let $H_* = \mathrm{diag}(p_*) - p_* p_*^{\mathsf{T}}$. This matrix is both the covariance of a one-hot draw from p_* and the local curvature of softmax KL at the target. For a score displacement d, let $R(d) = \max_i d_i - \min_i d_i$ be its range. Define

$$c_-(R) = \frac{R - 1 + e^{-R}}{R^2}, \qquad c_+(R) = \frac{e^R - 1 - R}{R^2}, \tag{61}$$

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

Theorem F.1 (Target-Fisher score bounds). *For every score displacement d,*

$$c_-(R(d)) d^{\mathsf{T}} H_* d \leq \mathrm{KL}(p_* \| \,\mathrm{softmax}(z_* + d)) \leq c_+(R(d)) d^{\mathsf{T}} H_* d. \tag{62}$$

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

Proof of Theorem F.1. Let $q_t = \mathrm{softmax}(z_* + td)$. Coordinatewise,

$$e^{-tR(d)} p_{*,i} \leq q_{t,i} \leq e^{tR(d)} p_{*,i}. \tag{63}$$

For any probability vector q,

$$v^{\mathsf{T}} H(q) v = \min_c \sum_i q_i (v_i - c)^2. \tag{64}$$

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 (63) into the Bregman integral

$$\mathrm{KL}(p_* \| \,\mathrm{softmax}(z_* + d)) = \int_0^1 (1-t) d^{\mathsf{T}} H(q_t) d \, dt. \tag{65}$$

The scalar integrals are

$$\int_0^1 (1-t) e^{-tR} dt = \frac{R - 1 + e^{-R}}{R^2}, \qquad \int_0^1 (1-t) e^{tR} dt = \frac{e^R - 1 - R}{R^2}. \tag{66}$$

This proves (62), 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

$$G_{\mathrm{F}}(u) = \beta^2 K(u)^{\mathsf{T}}[\mathrm{diag}(p_*(u)) - p_*(u)p_*(u)^{\mathsf{T}}]K(u) \tag{67}$$

and, for an error matrix E, use the trace form

$$Q_{\mathrm{F}}(E) = \mathbb{E}\big[\mathrm{tr}\big(G_{\mathrm{F}}(u)Eh(u)h(u)^{\mathsf{T}}E^{\mathsf{T}}\big)\big]\,. \tag{68}$$

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

$$\mathcal{E}_{r,R_0} = \inf_{M \in \mathcal{C}_{r,R_0}} \mathbb{E}\mathrm{KL}(p_* \| p_M), \quad T^{\mathrm{F}}_{r,R_0} = \inf_{M \in \mathcal{C}_{r,R_0}} Q_{\mathrm{F}}(M - \Delta_*). \tag{69}$$

Corollary F.2 (Constrained Fisher rank–KL bounds)**.** *If $\mathcal{C}_{r,R_0}$ is nonempty, then*

$$c_-(R_0)T^{\mathrm{F}}_{r,R_0} \le \mathcal{E}_{r,R_0} \le c_+(R_0)T^{\mathrm{F}}_{r,R_0}. \tag{70}$$

Proof. Apply Theorem F.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 $\eta \downarrow 0$. □

The nonemptiness condition is substantive. With unbounded keys or activations, every inexact rank-r candidate may have unbounded score range, making $\mathcal{C}_{r,R_0}$ empty for every finite R_0. One simple sufficient check is available when $\|K(u)\|_{\mathrm{op}} \le B_K$ and $\|h(u)\|_2 \le B_h$ almost surely. For $E = M - \Delta_*$,

$$R\big(\beta\Pi_{n(u)}K(u)Eh(u)\big) \le \sqrt{2}\,\beta B_K B_h \|E\|_{\mathrm{op}}, \tag{71}$$

so a candidate satisfying the right-hand side at most R_0 belongs to the restricted class. Other support conditions can be used, but the range envelope must be verified rather than inferred from a mean-square error. The constants also limit this route to a local regime: as $R \to \infty$, $c_+(R) \asymp e^R/R^2$ and $c_-(R) \asymp 1/R$, so their ratio grows as e^R/R.

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 $T^{\mathrm{F}}_r = \sum_{j>r}\sigma_j(D_{\mathrm{F}})^2$. The quadratic then factorizes in the same trace notation as Appendix D:

$$Q_{\mathrm{F}}(E) = \mathrm{tr}(G_{\mathrm{F}}E\Sigma E^{\mathsf{T}}) = \|G_{\mathrm{F}}^{1/2}E\Sigma^{1/2}\|_{\mathrm{F}}^2. \tag{72}$$

Lemma D.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 $\mathcal{E}_{r,R_0} \le c_+(R_0)T^{\mathrm{F}}_r$. Here independence is a genuine extra assumption: $G_{\mathrm{F}}(u)$ contains the target probabilities and therefore generally depends on the same input as $h(u)$. Without independence, the constrained score-space result remains valid. A spectral specialization instead requires direct comparison between $Q_{\mathrm{F}}(E)$ and a chosen reference quadratic form, as in Theorem E.1.

G High-mass lower bounds

For each sampled attention row, let $S(u)$ be a token set chosen from the target before the candidate is optimized. Suppose this set carries at least $1-\delta$ of the target probability. After conditioning the target on $S(u)$, assume that every retained token has probability at least a_{core}. The theorem below discards the nonnegative KL contributions outside this set and applies the global lower bound to the conditional distribution that remains.

Theorem G.1 (High-mass lower bound)**.** *Under these assumptions,*

$$\mathcal{E}_r \ge \frac{(1-\delta)a_{\mathrm{core}}}{2e^2} \inf_{\mathrm{rank}(M)\le r} \mathbb{E}\psi\big(\big\|\Pi_{S(u)}(z_M(u) - z_*(u))_{S(u)}\big\|_2\big)\,. \tag{74}$$

The theorem is valid for every nonempty core, but a row with a singleton core contributes zero to its right-hand side. Indeed, if $|S(u)| = 1$, then $\Pi_{S(u)} = 0$. In the conditional decomposition below, any mismatch in the total mass assigned to that singleton lies in the Bernoulli partition term, which the theorem discards. The population lower bound can still receive contributions from other rows with larger cores, but it is identically vacuous when the cores are singleton almost surely. For nearly one-sparse targets, one must control the partition term or accept that this lower bound may be weak.

Lemma G.2 (Conditional KL decomposition)**.** *For positive distributions p, q and nonempty S, put $s = p(S)$ and $t = q(S)$. Then*

$$\mathrm{KL}(p\|q) = \mathrm{kl}_{\mathrm{Bern}}(s\|t) + s\mathrm{KL}(p^S\|q^S) + (1-s)\mathrm{KL}(p^{S^c}\|q^{S^c}). \tag{75}$$

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 G.1. By Lemma G.2, 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

$$a_S = \min_{i\in S} \frac{p_{*,i}}{s_*}. \tag{76}$$

Thus, for each candidate,

$$\mathrm{KL}(p_*\|p_M) \geq \frac{s_* a_S}{2e^2}\psi(\|\Pi_S(z_M - z_*)_S\|_2). \tag{77}$$

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

For a dependence-aware spectral form, let $K_S(u)$ contain the selected key rows and define, for $E = M - \Delta_*$,

$$G_{\mathrm{core}}(u) = \beta^2 K_S(u)^\mathsf{T}\Pi_{S(u)}K_S(u), \qquad Q_{\mathrm{core}}(E) = \mathbb{E}\, h(u)^\mathsf{T}E^\mathsf{T}G_{\mathrm{core}}(u)Eh(u). \tag{78}$$

Choose a reference matrix $\bar{G}_{\mathrm{core}} \succeq 0$, put

$$S_{\mathrm{core}}(E) = \|\bar{G}_{\mathrm{core}}^{1/2}E\Sigma^{1/2}\|_{\mathrm{F}}^2, \qquad \bar{T}_r^{\mathrm{core}} = \inf_{\mathrm{rank}(M)\leq r} S_{\mathrm{core}}(M-\Delta_*), \tag{79}$$

and suppose that, over the complete rank-r candidate class,

$$Q_{\mathrm{core}}(E) \geq c_{r,\mathrm{core}}S_{\mathrm{core}}(E), \qquad \mathbb{E}\|\Pi_{S(u)}\beta K_S(u)Eh(u)\|_2^4 \leq \kappa_{r,\mathrm{core}}Q_{\mathrm{core}}(E)^2. \tag{80}$$

The weighted approximation lemma identifies $\bar{T}_r^{\mathrm{core}} = \sum_{\ell>r}\sigma_\ell(\bar{G}_{\mathrm{core}}^{1/2}\Delta_*\Sigma^{1/2})^2$. Applying the moment bridge to the score-space high-mass bound gives

$$\mathcal{E}_r \geq \frac{(1-\delta)a_{\mathrm{core}}}{2e^2(1+\sqrt{\kappa_{r,\mathrm{core}}})}\psi\left(\sqrt{c_{r,\mathrm{core}}\bar{T}_r^{\mathrm{core}}}\right). \tag{81}$$

This formulation permits arbitrary dependence between the selected core, keys, and query activation. If $G_{\mathrm{core}}(u)$ is independent of $h(u)$, one may take $\bar{G}_{\mathrm{core}} = \mathbb{E}G_{\mathrm{core}}(u)$ and obtain $c_{r,\mathrm{core}} = 1$; activation hypercontractivity together with $G_{\mathrm{core}}(u) \preceq \Lambda_{\mathrm{core}}\bar{G}_{\mathrm{core}}$ gives $\kappa_{r,\mathrm{core}} = \Lambda_{\mathrm{core}}^2\kappa_h$. When $S(u)$ is selected from the target probabilities, however, that independence generally fails, and the candidate-class conditions in (80) must be checked directly.

When this lower bound is paired with the global upper bound, the two sides use different geometries. The upper curve is built from $G = \mathbb{E}G(u)$ and T_r, whereas the core lower curve is built from $\bar{G}_{\mathrm{core}}$ and $\bar{T}_r^{\mathrm{core}}$. Their singular subspaces and decay rates need not agree. Thus, the upper and lower bounds cannot generally be interpreted as constant-factor bounds on a single spectral tail. Computing both requires separate weighted decompositions using G and $\bar{G}_{\mathrm{core}}$.

Table 2: Comparison of the three rank–KL routes. Each conclusion applies only under the requirements in the last column.

Route	Upper-bound conclusion	Lower-bound conclusion	Additional requirements
Global	Explicit weighted-SVD candidate; no probability floor needed	All rank-r candidates	Probability floor a and moment/geometry assumptions for the lower bound
Target Fisher	Weighted-SVD candidate, after verifying its score span	Candidates in the same bounded-span class	Target-Fisher geometry and a justified span bound R_0
High mass	No core-based upper bound; retain the global upper bound	All rank-r candidates	Target-selected $(1-\delta)$-mass core and a conditional probability floor

H 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 $\mathrm{rank}(Z_m) \le r$, every column of Z_m lies in L, and columnwise softmax converges to P_∞.

We give two related constructions. The first uses only linearly many token positions and shows that ranks 2 and 4 suffice for blocks of three and seven contexts, respectively. These are upper bounds for that family. The second construction uses isolated three-context blocks and proves the matching lower bound required by Theorem 6.1.

H.1 THE LINEAR-TOKEN ACHIEVABLE CONSTRUCTION

Write $k = 3q + s$, $s \in \{0, 1, 2\}$, and choose the least t such that $n = 4^t \ge k + 1$. 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 set $K_{x,a} = n^{-1/2}\chi_a(x)$. The nontrivial Walsh characters have zero mean, and character orthogonality gives $K^\mathsf{T}K = I_k$. Context a has $h_a = e_a$, zero pretrained logits, and target logits TKe_a. These logits equal $T/\sqrt{n}$ on $S_a = \{x : \chi_a(x) = 1\}$ and $-T/\sqrt{n}$ on its complement. The target attention is therefore uniform within each halfspace and converges to the uniform distribution on S_a as $T \to \infty$.

We first check the finite-score rank. Let M be a coefficient update that reproduces all target attention columns at a fixed T. Its score in context a is KMe_a. If two vectors have the same softmax, they differ by a constant multiple of the all-ones vector. Both KMe_a and TKe_a are centered, so that constant is zero:

$$KMe_a = TKe_a. \tag{82}$$

Multiplication by K^T now gives $Me_a = Te_a$. This holds for every context, hence $M = TI_k$ and every exact finite-score update has rank k.

We now construct lower-rank scores with the same limiting attention. Consider one complete character triple, written as x, y, xy. Put $g = T^{1/2}$ and $\alpha = T^{-3/4}$, and define

$$b_2 = (T, T+g, -T)^\mathsf{T}, \quad b_3 = (T, -T, T+g)^\mathsf{T}, \quad b_1 = \alpha(b_2 + b_3), \quad B_T = [b_1\ b_2\ b_3]. \tag{83}$$

B_T has rank two because its first column is a multiple of the sum of the other two; b_2 and b_3 are linearly independent, so its rank is exactly two. The first column can be written explicitly as

$$b_1 = (2T^{1/4}, T^{-1/4}, T^{-1/4})^\mathsf{T}. \tag{84}$$

Consequently, its score at token (x, y) is

$$2T^{1/4}x + T^{-1/4}y + T^{-1/4}xy = \begin{cases} 2T^{1/4} + 2T^{-1/4}y, & x = 1, \\ -2T^{1/4}, & x = -1. \end{cases} \tag{85}$$

The score spread on the target support $\{x = 1\}$ is $4T^{-1/4} \to 0$, while the smallest support score exceeds the off-support score by $4T^{1/4} - 2T^{-1/4} \to \infty$.

For the second column, the score is

$$xT + y(T+g) - xyT = \begin{cases} T+g, & y = 1, \\ 2xT - T - g, & y = -1. \end{cases} \tag{86}$$

It is constant on the target support $\{y = 1\}$, and its largest off-support value is $T - g$, leaving a gap of $2g = 2T^{1/2}$. The third column has the same calculation with y replaced by xy. In each column, the within-support spread tends to zero and the gap to the complement diverges. On a finite token set, these two facts imply softmax convergence to the uniform distribution on the support. A leftover character needs only the scalar coefficient $T^{1/4}$, whose two scores are $\pm T^{1/4}$. Combining the independent blocks gives rank $2q + s = k - \lfloor k/3 \rfloor$.

The same idea can achieve a smaller ratio on blocks of seven contexts. Take the token rows in the order $000, 001, \ldots, 111$ and the seven nonzero character indices in the order $001, 010, \ldots, 111$. Set $\rho = T^{-2}$ and $L = T^5$, and define

$$C_\rho = \begin{bmatrix} 1 & \rho^2 & 0 & 0 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho^2 & \rho^2 & 0 & 0 \\ 0 & 0 & 0 & \rho & \rho^2 & 1 & 0 \\ 0 & \rho^2 & 0 & \rho^2 & 0 & 0 & 1 \end{bmatrix} \tag{87}$$

and

$$F_\rho = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -\rho & -\rho^2 & 0 & -\rho^2 \\ 0 & -\rho & -\rho^2 & -\rho^2 \\ -1 & 0 & -\rho^2 & 0 \\ 0 & 0 & -1 & -\rho \\ -\rho^2 & -\rho^2 & -\rho & 0 \\ 0 & -1 & 0 & 0 \\ -\rho^2 & 0 & 0 & -1 \end{bmatrix}, \qquad Z_T = LF_\rho C_\rho. \tag{88}$$

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

$$F_\rho C_\rho = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\rho & -2\rho^3 - \rho^4 & -\rho^2 & -2\rho^4 & -\rho^2 - \rho^4 & 0 & -\rho^2 \\ 0 & -\rho^2 - \rho^4 & -\rho & -2\rho^3 - \rho^4 & -\rho^3 - \rho^4 & -\rho^2 & -\rho^2 \\ -1 & -\rho^2 & 0 & -\rho^3 & -\rho - \rho^4 & -\rho^2 & 0 \\ 0 & -\rho^3 & 0 & -\rho - \rho^3 & -\rho^2 & -1 & -\rho \\ -\rho^2 & -\rho^3 - \rho^4 & -\rho^2 & -\rho^2 - \rho^4 & -2\rho^3 - \rho^4 & -\rho & 0 \\ 0 & -\rho & -1 & -\rho^2 & -\rho^2 & 0 & 0 \\ -\rho^2 & -\rho^2 - \rho^4 & 0 & -\rho^2 & -\rho^3 & 0 & -1 \end{bmatrix}. \tag{89}$$

Direct inspection of column a shows that the four rows in its positive halfspace $S_a = \{x : \langle a, x \rangle = 0\}$ have entries between $-3\rho^3$ and 0. Indeed, the terms in these positions are zero or sums of at most two ρ^3 terms and one ρ^4 term; since $\rho < 1/3$, their magnitude is below $3\rho^3$. Every off-support position contains a term of magnitude at least ρ^2 and is therefore at most $-\rho^2$. Multiplying by L gives, for every character a,

$$-3\rho^3 L \le Z_T(x, a) \le 0 \quad (x \in S_a), \qquad Z_T(x, a) \le -\rho^2 L \quad (x \notin S_a). \tag{90}$$

Because $\rho^3 L = T^{-1}$ and $\rho^2 L = T$, the within-support spread is at most $3/T$ and the gap between the smallest support score and largest off-support score is at least $T - 3/T$. Thus every column again converges to the uniform distribution on its target halfspace.

The columns of Z_T need not initially be centered. Subtracting each column mean leaves its softmax unchanged and amounts to left multiplication by the centering projection, so it cannot increase rank. Every centered vector on $\mathbb{F}_2^3$ lies in the span of the seven nontrivial Walsh characters. The centered columns are therefore valid scores in the allowed character space.

To obtain any number of contexts, identify $\mathbb{F}_2^{3t}$ with $\mathbb{F}_8^t$, where t is least such that $n = 8^t \ge k + 1$. Partition the nonzero character indices into one-dimensional $\mathbb{F}_8$ subspaces. Each such subspace consists of seven nonzero vectors and, as a binary space, has dimension three. The seven-character construction above therefore lifts to it, with each of the eight sign patterns repeated equally often across the n token rows. If $k = 7m + s$, $0 \le s < 7$, the resulting path has rank at most $4m + s = k - 3\lfloor k/7 \rfloor$; the incomplete block uses s scalar paths. Minimality of t gives $n < 8(k + 1)$.

The score calculations above show convergence to the correct supports. We also need convergence in KL. 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. Splitting both distributions between S_a and S_a^c decomposes the KL divergence into a binary-mass term and the two conditional KL terms, as in Lemma G.2. The target conditional distribution is uniform on each halfspace. The candidate becomes uniform on S_a, so the support conditional term vanishes.

In both constructions, $\eta_T \to 0$ and the candidate score range is polynomial in T. Hence every candidate probability is at least $\exp[-\operatorname{poly}(T)]/n$, which gives $-\log \eta_T = \operatorname{poly}(T)$. The potentially large part of the binary-mass term is therefore bounded by

$$\epsilon_T \log(\epsilon_T/\eta_T) \le \epsilon_T[-\log \eta_T] = \epsilon_T \operatorname{poly}(T) \longrightarrow 0. \tag{91}$$

The remaining part of the binary KL tends to zero because $\epsilon_T, \eta_T \to 0$. The complement conditional KL is at most the candidate score range plus $\log(n/2)$, so after multiplication by ϵ_T it obeys the same exponentially small bound. This proves KL, rather than only support, convergence. Taking the better of the triple and septuple packings proves the linear-token statement in Section 6.

H.2 Proof of the exact closure-rank theorem

Proof of Theorem 6.1. Isolated character blocks. Write $k = 3q + s$, $s \in \{0, 1, 2\}$. We first choose the character indices so that each intended triple is isolated from all the others. Work in a binary vector space V whose size is the least power of two above $2k^2 + 2$, and construct

$$\mathcal{A} = T_1 \dot\cup \cdots \dot\cup T_q \dot\cup R, \quad T_\ell = \{a_\ell, b_\ell, a_\ell + b_\ell\}, \quad |R| = s, \tag{92}$$

so that the T_ℓ are the only additive triples in $\mathcal{A}$. To add a new triple to a current set A_j, let $F_j = A_j + A_j$, choose $u \notin F_j$, and then choose $v \notin F_j \cup (u + F_j) \cup \{0, u\}$. Fewer than $2k^2 + 2$ vectors are forbidden at each step: $|F_j| \le |A_j|^2 \le k^2$, the translate $u + F_j$ has the same size, and the final set has size at most two. The choice is therefore possible because $|V| > 2k^2 + 2$.

To see why this isolates the new block, set $w = u + v$. Neither u nor v is the sum of two old elements because both lie outside F_j, and $w \notin F_j$ because $v \notin u + F_j$. Thus no two old elements sum to a new one. Conversely, the sum of any two elements of $\{u, v, w\}$ is the third, so it cannot be an old element. The only new additive triple is therefore $\{u, v, w\}$. (Every element already in A_j belongs to a complete triple and hence lies in F_j, so the new elements are also distinct from the old ones.) Finally, add each of the at most two elements of R outside the updated sumset; doing this sequentially creates no additional triple. Since $|V|$ is the least power of two above $2k^2 + 2$, the number of token rows satisfies $n = |V| < 4k^2 + 8$.

Finite-rank requirement and closure upper bound. Use normalized Walsh characters χ_a for $a \in \mathcal{A}$ as centered orthonormal key columns. The finite targets are again TKe_a. The centered-softmax argument above applies without change: an exact coefficient matrix M must satisfy $KMe_a = TKe_a$ for every a, and orthonormality gives $M = TI_k$. Thus exact reproduction at any finite T requires rank k.

For the limiting target, order the coefficient matrix by the blocks $T_1, \ldots, T_q, R$. Place one copy of B_T from (83) on each triple block and a scalar $T^{1/4}$ on each leftover coordinate. The resulting matrix is block diagonal. Each triple block has rank two, each leftover block has rank one, and the score calculation above gives the required limiting distribution in every context. Hence the closure rank is at most $2q + s$.

Fourier constraints for the lower bound. Now consider any coefficient sequence M_m whose score softmaxes converge to the target boundary distributions. We show that its rank cannot be below $2q + s$. Absorb the fixed Walsh normalization into M_m. For context a, write

$$f_{m,a}(x) = \sum_{b \in \mathcal{A}} m_{b,a}^{(m)} \chi_b(x), \qquad t_{m,a} = m_{a,a}^{(m)}. \tag{93}$$

On $S_a = \{x : \chi_a(x) = 1\}$, the restrictions of χ_b and χ_{b+a} agree because $\chi_{b+a}(x) = \chi_b(x)\chi_a(x) = \chi_b(x)$. Softmax convergence to the uniform distribution on S_a implies that all score differences within S_a vanish: for $x, y \in S_a$,

$$\exp(f_{m,a}(x) - f_{m,a}(y)) = \frac{p_{m,a}(x)}{p_{m,a}(y)} \longrightarrow 1. \tag{94}$$

Since the set is finite, $f_{m,a}$ is therefore its mean on S_a plus $o(1)$ uniformly over S_a.

The restricted Fourier coefficient indexed by b is

$$\frac{2}{n}\sum_{x\in S_a} f_{m,a}(x)\chi_b(x) = m_{b,a}^{(m)} + m_{b+a,a}^{(m)}. \tag{95}$$

Here we used $\mathbf{1}_{S_a} = (1+\chi_a)/2$ and Walsh orthogonality. Every nonconstant restricted coefficient must tend to zero, which gives

$$m_{b,a}^{(m)} + m_{b+a,a}^{(m)} \longrightarrow 0 \quad \text{for } b \notin \{0, a\}. \tag{96}$$

Here a coefficient whose index is not in $\mathcal{A}$ is understood to be zero. The constant restricted coefficient is

$$\frac{2}{n}\sum_{x\in S_a} f_{m,a}(x) = m_{0,a}^{(m)} + m_{a,a}^{(m)} = t_{m,a}, \tag{97}$$

because the constant character is not in $\mathcal{A}$. Thus $f_{m,a}(x) = t_{m,a} + o(1)$ uniformly on S_a. Similarly,

$$\frac{2}{n}\sum_{x\in S_a^c} f_{m,a}(x) = m_{0,a}^{(m)} - m_{a,a}^{(m)} = -t_{m,a}. \tag{98}$$

For every $x \in S_a^c$ and $y \in S_a$, convergence to the boundary target gives $p_{m,a}(x)/p_{m,a}(y) \to 0$, and hence $f_{m,a}(x) - f_{m,a}(y) \to -\infty$. Averaging over the two halfspaces and using the preceding identities yields $-t_{m,a} - t_{m,a} \to -\infty$. Therefore $t_{m,a} \to \infty$.

If b belongs to another block, isolation ensures that $b+a \notin \mathcal{A}$. Equation (96) then gives $m_{b,a}^{(m)} = o(1)$, so all cross-block coefficients vanish. Within the triple $\{a, b, c = a+b\}$, put $s_{m,a} = m_{b,a}^{(m)}$. The same identity gives $m_{c,a}^{(m)} = -s_{m,a} + o(1)$. On the negative halfspace, $\chi_c = \chi_a\chi_b = -\chi_b$. All cross-block terms are $o(1)$, while χ_b takes both signs on S_a^c, so

$$\max_{S_a^c} f_{m,a} = -t_{m,a} + 2|s_{m,a}| + o(1). \tag{99}$$

The smallest support score is $t_{m,a} + o(1)$. Since its gap above the largest complement score must diverge,

$$2(t_{m,a} - |s_{m,a}|) + o(1) \longrightarrow \infty. \tag{100}$$

In particular, $|s_{m,a}|/t_{m,a}$ is bounded (indeed, eventually at most one).

Normalization and rank. Normalize each column by its positive diagonal coefficient:

$$C_m = M_m \operatorname{diag}(t_{m,a}^{-1} : a \in \mathcal{A}). \tag{101}$$

This preserves rank. Diagonal entries become one, cross-block entries converge to zero because their numerators are $o(1)$ while $t_{m,a} \to \infty$, and the preceding bound makes all within-block entries bounded. There are only finitely many entries, so a single subsequence makes all within-block ratios converge. The limit is block diagonal, with singleton blocks $[1]$ and, after ordering a triple as (a, b, c), blocks of the form

$$B(\alpha, \beta, \gamma) = \begin{bmatrix} 1 & \beta & \gamma \\ \alpha & 1 & -\gamma \\ -\alpha & -\beta & 1 \end{bmatrix}. \tag{102}$$

Here α, β, and γ are the limiting normalized off-diagonal coefficients in columns a, b, and c, respectively; the signs follow from (96). If such a block had rank one, the minors on row/column pairs $(1,2)$, $(1,3)$, and $(2,3)$ would give

$$\alpha\beta = 1, \qquad -\alpha\gamma = 1, \qquad \beta\gamma = 1. \tag{103}$$

The first and third equalities imply $\alpha = \gamma$ (and $\beta \neq 0$), while the second would then require $-\alpha^2 = 1$, which is impossible over the reals. Every triple block therefore has rank at least two. Rank adds across the diagonal blocks, so the limiting normalized matrix has rank at least $2q + s = k - \lfloor k/3 \rfloor$.

If the original sequence had rank at most r, then every C_m would also have rank at most r. Because the set of matrices of rank at most r is closed, its limit would have rank at most r as well. The preceding block calculation therefore forces $r \geq 2q + s$. Together with the constructed upper bound, this proves the claimed equality. □

I FUSED MULTI-HEAD PROOFS

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

$$\psi\left(\sqrt{\sum_h x_h^2}\right) \le \sum_h \psi(x_h) \le \sqrt{H}\,\psi\left(\sqrt{\sum_h x_h^2}\right) \tag{104}$$

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, still in the case $s > 1$, $\psi(x_h) \le x_h$, 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 (104), gives the lower bound

$$\frac{a_{\min}}{2e^2}\psi(X_{\mathrm{MH}}) \le \sum_h D_h, \qquad X_{\mathrm{MH}}^2 = \sum_h X_h^2. \tag{105}$$

For the upper bound, the headwise inequalities give directly

$$\sum_h D_h \le \sum_h \min\{X_h^2/4, \sqrt{2}X_h\} \le \min\{X_{\mathrm{MH}}^2/4, \sqrt{2H}X_{\mathrm{MH}}\}. \tag{106}$$

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 (21) and independence give

$$\mathbb{E}X_{\mathrm{MH}}^2 = \|G_{\mathrm{MH}}^{1/2}(M - \Delta_*^{\mathrm{MH}})\Sigma^{1/2}\|_{\mathrm{F}}^2. \tag{107}$$

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 D.2. Lemma D.1 therefore gives the robust lower bound

$$\mathbb{E}\psi(X_{\mathrm{MH}}) \ge \frac{\psi(\sqrt{T_r^{\mathrm{MH}}})}{1 + \Lambda_{\mathrm{MH}}\sqrt{\kappa_h}}. \tag{108}$$

For the weighted-SVD candidate, take expectations in Equation (106), 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 (23). □

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

$$\inf_{\substack{\sum_h r_h \le R \\ \mathrm{rank}(M_h) \le r_h}} \sum_h \mathbb{E}D_h = \min_{\sum_h r_h \le R} \sum_h \mathcal{E}_{h,r_h}. \tag{109}$$

This proves the allocation identity.

I.1 DEPENDENCE-AWARE FUSED BOUNDS

The independence assumption is not needed for the score-to-KL step. Let $E = M - \Delta_*^{\mathrm{MH}}$ and define

$$Q_{\mathrm{MH}}(E) = \mathbb{E}\sum_{h=1}^{H} \|d_{M,h}(u)\|_2^2, \qquad S_{\mathrm{MH}}(E) = \|G_{\mathrm{MH}}^{1/2}E\Sigma^{1/2}\|_{\mathrm{F}}^2. \tag{110}$$

If $Q_{\mathrm{MH}}(E) \ge c_r S_{\mathrm{MH}}(E)$ and $\mathbb{E}(\sum_h \|d_{M,h}(u)\|_2^2)^2 \le \kappa_r Q_{\mathrm{MH}}(E)^2$ for every feasible E, while the weighted-SVD error E_r^{svd} satisfies $Q_{\mathrm{MH}}(E_r^{\mathrm{svd}}) \le C_r^+ T_r^{\mathrm{MH}}$, then the proof above gives

$$\frac{a_{\min}}{2e^2(1+\sqrt{\kappa_r})}\psi(\sqrt{c_r T_r^{\mathrm{MH}}}) \le \mathcal{E}_r^{\mathrm{MH}} \le \min\left\{\frac{C_r^+ T_r^{\mathrm{MH}}}{4}, \sqrt{2HC_r^+ T_r^{\mathrm{MH}}}\right\}. \tag{111}$$

Thus the fused result survives dependent keys and queries whenever these candidate-class comparisons can be established.

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

$$\|y_{M,h} - y_{*,h}\|_2 \le \mathrm{diam}_{V,h}(u)\,\mathrm{TV}(p_{M,h}, p_{*,h}) \le \mathrm{diam}_{V,h}(u)\sqrt{D_h/2}. \tag{112}$$

By Cauchy–Schwarz across heads,

$$\left\|\sum_h O_h(y_{M,h} - y_{*,h})\right\|_2^2 \le \frac{H}{2}\sum_h \|O_h\|_{\mathrm{op}}^2\,\mathrm{diam}_{V,h}(u)^2 D_h(u). \tag{113}$$

J JOINT QUERY/KEY PROOFS

For ordinary dot-product attention, define

$$C(A,B) = K_0^\top A + B^\top(Q_0 + A). \tag{114}$$

The two summands have ranks at most $\mathrm{rank}(A)$ and $\mathrm{rank}(B)$, proving Equation (25). The bound is sharp: for $p = r$, $d = 2r$, take

$$K_0 = A = [I_r\ 0], \qquad Q_0 = B = [0\ I_r]. \tag{115}$$

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

Every joint candidate also satisfies

$$\mathrm{rank}(K_0^\top Q_0 + C(A,B)) = \mathrm{rank}((K_0+B)^\top(Q_0+A)) \le p. \tag{116}$$

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

$$d_{A,B}(u) = \Pi_{n(u)}\beta X(u)[C(A,B) - C_*]h(u) \tag{117}$$

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

$$\frac{a}{2e^2}\Psi^{\mathrm{QK}}_{r_Q,r_K} \le \mathcal{E}^{\mathrm{QK}}_{r_Q,r_K}. \tag{118}$$

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

$$\mathbb{E}\|d_{A,B}(u)\|_2^2 \ge T_s^{\mathrm{eff}} \tag{119}$$

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

$$\mathbb{E}\psi(\|d_{A,B}\|_2) \ge \frac{\psi(\sqrt{T_s^{\mathrm{eff}}})}{1+\sqrt{\kappa}}. \tag{120}$$

Take the infimum and use (118) for the lower bound.

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

$$\rho_{r_Q,r_K} = \inf_{C_s \in \mathcal{S}_s^{\mathrm{svd}}}\ \inf_{\substack{\mathrm{rank}(A)\le r_Q \\ \mathrm{rank}(B)\le r_K}} \|G_X^{1/2}[C(A,B) - C_s]\Sigma^{1/2}\|_{\mathrm{F}}. \tag{121}$$

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

$$\left\{\mathbb{E}\|d_{A,B}(u)\|_2^2\right\}^{1/2} \le \sqrt{T_s^{\mathrm{eff}}} + \rho_{r_Q,r_K} + o(1). \tag{122}$$

Apply the pointwise global upper bound and Equation (55), then let $\eta \downarrow 0$. □

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

$$A = (K_0^\mathsf{T})^\dagger C, \qquad C(A, 0) = C, \qquad \mathrm{rank}(A) \le \mathrm{rank}(C). \tag{123}$$

The symmetric statement for keys uses $B^\mathsf{T} = CQ_0^\dagger$ when $\mathrm{row}(C) \subseteq \mathrm{row}(Q_0)$. More generally, suppose $C = C_Q + C_K$, choose A_Q with $K_0^\mathsf{T} A_Q = C_Q$ and $\mathrm{rank}(A_Q) \le r_Q$, and assume $\mathrm{row}(C_K) \subseteq \mathrm{row}(Q_0 + A_Q)$ with $\mathrm{rank}(C_K) \le r_K$. Then

$$B_K^\mathsf{T} = C_K (Q_0 + A_Q)^\dagger \tag{124}$$

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

$$C(A, B) - C(A_*, B_*) = (K_0 + B)^\mathsf{T}(A - A_*) + (B - B_*)^\mathsf{T}(Q_0 + A_*). \tag{125}$$

Submultiplicativity yields

$$\begin{aligned} &\|G_X^{1/2}[C(A, B) - C_*]\Sigma^{1/2}\|_\mathrm{F} && (126)\\ &\quad \le \|G_X^{1/2}(K_0 + B)^\mathsf{T}\|_\mathrm{op}\|(A - A_*)\Sigma^{1/2}\|_\mathrm{F} && (127)\\ &\qquad + \|(Q_0 + A_*)\Sigma^{1/2}\|_\mathrm{op}\|G_X^{1/2}(B - B_*)^\mathsf{T}\|_\mathrm{F}. && (128) \end{aligned}$$

Combining this with (55) and the global softmax upper bound gives the constructive target-factor upper bound. Weighted truncated SVDs make this certificate explicit. Define

$$A_{r_Q} = (A_*\Sigma^{1/2})_{r_Q}\Sigma^{\dagger/2}, \qquad B_{r_K}^\mathsf{T} = G_X^{\dagger/2}(G_X^{1/2}B_*^\mathsf{T})_{r_K}, \tag{129}$$

and let

$$\begin{aligned} \eta_{r_Q, r_K} &= \|G_X^{1/2}(K_0 + B_{r_K})^\mathsf{T}\|_\mathrm{op}\|(A_{r_Q} - A_*)\Sigma^{1/2}\|_\mathrm{F} \\ &\quad + \|(Q_0 + A_*)\Sigma^{1/2}\|_\mathrm{op}\|G_X^{1/2}(B_{r_K} - B_*)^\mathsf{T}\|_\mathrm{F}. \end{aligned} \tag{130}$$

These candidates satisfy the two rank budgets, and, under the separable geometry used in Theorem 8.1,

$$\mathcal{E}_{r_Q, r_K}^{\mathrm{QK}} \le \psi_\mathrm{up}(\eta_{r_Q, r_K}). \tag{131}$$

Unlike ρ_{r_Q, r_K}, every quantity in this upper curve follows from two ordinary weighted truncated SVDs and operator norms. It may be looser because the factors are approximated separately.

Dependence and positional encoding. When $X(u)$ and $h(u)$ come from the same hidden states, a spectral result requires candidate-class comparisons between the true coupled mean-square error and a reference quadratic form, as in Theorem E.1. RoPE introduces a relative rotation between query and key factors. Each relative-position slice still has rank at most $r_Q + r_K$, so the pointwise robust bound survives, but the candidate is a coupled family of effective matrices and the single-matrix tail T_s^{eff} does not apply automatically. This obstruction is absent for ordinary dot-product attention in NoPE architectures such as Kimi K3 (Kimi Team, 2026), subject to the model's other attention mechanisms.

K USING THE BOUNDS ON DOWNSTREAM DATA

This section describes how to estimate the bounds from downstream calibration data when a dense or high-rank target update is already available. The target attention function and downstream distribution must be specified, and the appropriate bound must be chosen according to the assumptions summarized in Table 2. The method applies only after the target update has been learned.

Fix an error tolerance ϵ. The final subsection explains how to compare the upper and lower curves with this tolerance, and distinguishes population guarantees from conclusions based on calibration estimates.

K.1 Define the target–candidate comparison

The target update Δ_* and candidate update M are fixed matrices, but their scores and probabilities vary with the sampled sequence and query position $u = (x, t)$. Calibration therefore averages row-level errors over the chosen distribution $\mathcal{D}$, while using the same M for every row. Appendix A lists the notation.

The following running example makes the sampling unit and error measure concrete.

Example 1: Attention KL for text-to-SQL

The token sequence x may contain the instructions, database schema, question, and SQL prefix generated so far. The index t selects the position whose attention row is evaluated. At that position, the target and candidate heads assign probabilities to the visible tokens. The objective asks how much rank is needed to reproduce the target attention distribution across the sampled sequence–position pairs. It is a head-level approximation measure to be considered alongside end-to-end measures such as execution accuracy.

K.2 Why the downstream geometry matters

The raw singular values of Δ_* measure parameter-space size, not the effect of the update on downstream attention. The matrices G and Σ remove directions that are invisible to the observed keys or query activations and emphasize directions that repeatedly change scores. The small example below isolates this distinction.

Example 2: Why the raw spectrum can fail

Suppose the target update has two singular directions. The larger one is orthogonal to every query activation in the task and never changes a score; the smaller one changes attention on nearly every input. A raw truncated SVD keeps the first direction, whereas the weighted spectrum removes it and keeps the direction that affects the target function.

This is why the calibration target is $D_* = G^{1/2}\Delta_*\Sigma^{1/2}$ rather than Δ_* itself. Its residual tail T_r measures the best rank-r score approximation in the downstream quadratic geometry. Converting that tail into KL bounds requires one of the three routes below.

K.3 Choose the applicable bound

The routes answer different questions and should be chosen before their empirical quantities are computed. The global route gives both sides for all rank-r candidates, but can be weak when the target probability floor is small. The Fisher route replaces that floor by target curvature, at the cost of restricting candidates to a bounded score span. The high-mass route gives an unrestricted lower bound on target-selected token cores, but no matching upper bound.

K.4 Calibration workflow

The following steps keep the population assumptions of the theorems separate from quantities estimated on a finite calibration set.

For query-only adaptation:

1. **Identify the target class.** Verify that the target attention is generated by a query update Δ_*. If both queries and keys change, use Theorem 8.1; unrestricted fine-tuning may not be representable by any query-only update.
2. **Define the downstream distribution.** Specify which attention rows are evaluated and how they are weighted. Run the pretrained and target heads on complete downstream inputs, storing $h(x, t)$, the masked centered key Gram $G(x, t)$, and target probabilities. Weight positions within each sequence according to $\mathcal{D}$, then average across sequences. Complete sequences, not individual rows, are the independent sampling units.

3. **Choose the route.** Use Figure 2 and Table 2. Decide whether the global probability floor is usable, whether a Fisher score-span restriction is justified, or whether a target-selected high-mass core is needed. If key geometry and query activations are not plausibly separable, use the coupled comparison in Theorem E.1.
4. **Compute the global weighted target.** Estimate $G = \mathbb{E}G(u)$ and $\Sigma = \mathbb{E}[h(u)h(u)^\mathsf{T}]$. Under a reported support or ridge convention, form $\widehat{D}_{*,\tau} = \widehat{G}_\tau^{1/2}\Delta_*\widehat{\Sigma}_\tau^{1/2}$ and its tails $\widehat{T}_r = \sum_{\ell>r}\sigma_\ell(\widehat{D}_{*,\tau})^2$. The same convention must be used in the explicit candidate $\widehat{M}_r = \widehat{G}_\tau^{\dagger/2}(\widehat{D}_{*,\tau})_r\widehat{\Sigma}_\tau^{\dagger/2}$.
5. **Compute route-specific quantities.** For the Fisher route, replace the global Gram by the target-curvature Gram and check the score-span envelope of the resulting candidate. For the high-mass route, fix a target-only core rule before evaluation, record singleton cores, and form the core Gram. The formulas are given below.
6. **Separate estimates from assumptions.** The probability floor, geometry comparison, and uniform moment constant are population assumptions, not facts certified by the calibration set. Use analytic values when available or report sensitivity over an explicit range. Report held-out KL of each constructed candidate separately from its analytic upper curve.
7. **Compare with the tolerance.** Interpret the curves as described in the final subsection. In particular, distinguish theorem conclusions based on population quantities from plug-in calculations based on calibration estimates.

K.5 Computing the empirical curves

The fourth and fifth steps above share a numerical issue: empirical Gram matrices are often singular or poorly conditioned. The square roots and inverse square roots used to construct a candidate must therefore follow one consistent support or regularization convention. Otherwise the reported tail and the reconstructed update no longer describe the same approximation.

A common support convention. Here is one explicit truncation rule. If $H = U\,\mathrm{diag}(\lambda_i)U^\mathsf{T} \succeq 0$ and $\tau_H > 0$, define

$$H_\tau^{1/2} = U\,\mathrm{diag}\Big(\lambda_i^{1/2}\mathbf{1}\{\lambda_i \ge \tau_H\}\Big)\,U^\mathsf{T},$$
$$H_\tau^{\dagger/2} = U\,\mathrm{diag}\Big(\lambda_i^{-1/2}\mathbf{1}\{\lambda_i \ge \tau_H\}\Big)\,U^\mathsf{T}. \tag{132}$$

The threshold τ_H may be absolute or relative to $\lambda_{\max}(H)$, but that choice and the retained dimension must be reported for both $\widehat{G}$ and $\widehat{\Sigma}$. Using the same retained support in the square root, weighted target, and inverse factors preserves the weighted-SVD identity on that support. A ridge alternative replaces each empirical Gram consistently by $H + \lambda_H I$ when forming both the weighted target and its inverse factors. Either modified curve is a stability diagnostic rather than the unregularized population curve.

Global route. Under the separable global assumptions, the truncated weighted target from the workflow gives the plug-in curves

$$\widehat{U}_r^{\mathrm{th}} = \psi_{\mathrm{up}}(\sqrt{\widehat{T}_r}), \qquad \widehat{L}_r^{\mathrm{global}} = \frac{a}{2e^2(1+\Lambda\sqrt{\kappa_h})}\psi(\sqrt{\widehat{T}_r}). \tag{133}$$

The upper curve corresponds to the explicit weighted-SVD candidate $\widehat{M}_r$ in the empirical geometry. The lower curve is the plug-in analogue of the population lower bound; using it to exclude a rank requires treating the estimated tail and the assumed constants as their population values. The two curves use the same spectral tail; their separation comes from the constants in the softmax comparison.

Target-Fisher route. For calibration rows u_m with weights w_m summing to one, the Fisher plug-in Gram is

$$\widehat{G}_{\mathrm{F}} = \sum_m w_m\beta^2 K(u_m)^\mathsf{T}\left[\mathrm{diag}(p_*(u_m)) - p_*(u_m)p_*(u_m)^\mathsf{T}\right]K(u_m). \tag{134}$$

If separability is credible, form $\widehat{D}_{\mathrm{F},\tau} = \widehat{G}_{\mathrm{F},\tau}^{1/2}\Delta_*\widehat{\Sigma}_\tau^{1/2}$ using the same support convention, and compute its residual tails. The lower curve applies to the bounded-span candidate class; a matching

spectral upper curve is justified only after checking the span envelope for its weighted-SVD candidate. Calibration and held-out checks of that envelope do not certify an almost-sure population bound.

High-mass route. For the high-mass route, choose $S(u_m)$ using only the target probabilities, for example the smallest set whose cumulative target mass is at least $1 - \delta$, with a fixed tie-breaking rule. Then compute

$$\widehat{G}_{\mathrm{core}} = \sum_m w_m \beta^2 K_{S(u_m)}(u_m)^{\mathsf{T}} \Pi_{S(u_m)} K_{S(u_m)}(u_m). \tag{135}$$

Report the frequency of singleton cores and the empirical conditional floor. Under a justified separability model, apply the same support convention to $\widehat{G}_{\mathrm{core}}$ and $\widehat{\Sigma}$ when forming the core spectral tail; otherwise evaluate the coupled score-space quantities in Equations (78)–(80). This lower curve has no matching core upper curve, so a hybrid decision must report it separately from the global upper curve and retain both spectra.

K.6 Interpreting the rank interval

Fix a KL tolerance ϵ. When the curves are computed from exact population quantities, an upper curve at most ϵ certifies the constructed rank-r candidate, and a lower curve above ϵ excludes every rank-r candidate covered by that theorem. Neither conclusion follows when the tolerance lies between the curves.

Curves computed from $\widehat{G}$, $\widehat{\Sigma}$, and other calibration estimates should instead be labeled as plug-in results. The same comparisons give a plug-in sufficient region, a plug-in excluded region, and an unresolved region, conditional on the empirical quantities and assumed constants. They are not population guarantees.

The qualification "valid" is essential. A Fisher lower curve excludes only candidates satisfying its score-span restriction. A high-mass lower curve can be paired with a global upper curve, but the two use different weighted spectra and must be displayed separately. Even the symmetric global route can leave a wide interval: the ratio of its upper and lower constants can be as large as $2\sqrt{2}e^2(1 + \Lambda\sqrt{\kappa_h})/a$.

Example 3: A calibration-set decision

For a high-rank text-to-SQL query adapter, collect complete prompts containing the question, schema, instructions, and SQL prefix, treating each prompt as an independent sampling unit. Choose and report how attention rows are weighted within a prompt. From those rows, estimate $\widehat{G}$, $\widehat{\Sigma}$, and the weighted tails, then construct $\widehat{M}_r$ for each tested rank.
Next inspect the assumptions rather than selecting the most favorable curve. If the global probability floor and moment assumptions are defensible, use the global pair. If the floor is too small, the global upper curve may still provide a candidate while a high-mass construction supplies a separate lower curve; its core rule, singleton frequency, and conditional floor must then be reported. Report the resulting plug-in rank interval together with held-out target-to-candidate KL for the constructed candidates.

Finally, all curves in this section are plug-in population calculations, not finite-sample confidence intervals. A held-out split can test whether the weighted tail predicts target-to-candidate KL better than the raw spectrum of Δ_* or activation-only criteria. It can also reveal instability under support thresholds, ridge values, or core choices. It does not verify the population assumptions or provide a finite-sample guarantee.