AN INFINITE INFORMATION EXPONENT THAT GRADIENT DESCENT STILL LEARNS AT THE OPTIMAL RATE: THE VALUE CATALYST IN SOFTMAX ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The information exponent of a single-index target dictates how gradient descent learns it: an exponent-k target traps online SGD for $\tilde{\Theta}(d^{k-1})$ samples, and an infinite exponent is the standard signature of a direction gradient descent can never find. We show this dictionary breaks for attention. Consider a single softmax head that selects a token through a *key* direction $w^\star$ and reads a scalar through an orthogonal *value* direction $u^\star$, $y = \sum_i \mathrm{softmax}(\beta\langle w^\star, x_i\rangle)_i \langle u^\star, x_i\rangle$. The selection direction has *infinite information exponent*: its label is exactly mean-zero given the key scores, so every correlational query and the gradient of any single-index surrogate returns zero. Yet projected gradient descent on the invariant slice of the two-parameter attention loss learns $w^\star$ at the information-theoretically optimal $n = \Theta(d)$ dimensional rate; unrestricted random initializations show the same cascade numerically. The mechanism is a *value catalyst*: the value direction has information exponent one, and once acquired the joint key gradient gains an explicit $\Theta(1)$ component along the selection direction, so gradient descent escapes the saddle in a value→selection cascade. A matching one-shot estimator squares the label to expose $w^\star$ at generative exponent two, giving a single spectral step at the same dimensional rate up to logarithmic factors. A lower bound that grants the learner all exact key scores still requires $n = \Omega(d/L)$ samples at every noise level; for fixed positive noise, a Fano bound additionally gives explicit (σ, β, L) dependence. The phenomenon extends to orthogonal-subspace heads combined additively or multiplicatively, and we report CPU-based numerical checks of the quantitative predictions.

1 INTRODUCTION

A decade of work on learning single- and multi-index Gaussian models summarizes difficulty through two exponents. The *information exponent* (Ben Arous et al., 2021) is the lowest Hermite degree at which the target correlates with a direction; it governs how long online stochastic gradient descent spends escaping the high-entropy equator, giving sample complexity $n = \tilde{\Theta}(d^{\max(1,k-1)})$. The *generative exponent* (Damian et al., 2024) is the lowest information exponent achievable after *any* transformation of the label; it governs the best possible polynomial-time (statistical-query) estimator, with threshold $n = \tilde{\Theta}(d^{\max(1,\mathrm{GE}/2)})$. When the two coincide, gradient descent on the raw loss is essentially optimal. When they differ, there is a genuine gap between what gradient methods see and what is statistically extractable.

Attention is the central nonlinearity of modern sequence models, yet the learning-theoretic status of even one head is unsettled. The computation is simple, a softmax over inner products followed by a weighted average, but it couples two roles that behave very differently under gradient descent: a *selection* (key/query) role that decides *which* token to attend to, and a *read* (value) role that decides *what* to copy once selected. We ask how hard each role is to learn in the information-and-generative-exponent framework used for index models.

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.

Result. The two roles behave very differently. We prove (Corollary 1, Theorem 2) that for an orthogonal key/value head the selection direction has, *as a single-index target read against the key scores*,

$$\mathrm{IE}_{w^\star}(y) = \infty, \qquad \mathrm{GE}_{w^\star}(y) = 2,$$

the largest possible gap. The infinitude is not a pathology of a single query. It is the statement that the conditional mean $\mathbb{E}[y \mid s]$ of the output given the key scores is *identically zero*, so no function of the selection scores correlates with the label at any order, and the gradient of any single-index surrogate is exactly zero along $w^\star$. By the usual dictionary this is a direction gradient descent can never find. The coupled attention loss nevertheless recovers it. The loss attention actually trains has two coupled parameters, a key direction and a value direction, and the value direction has information exponent one (Lemma 7). Once even partially acquired, it acts as a *catalyst*: the joint key gradient develops an explicit nonzero component along the selection direction, which we compute in closed form (Theorem 1),

$$\nabla_w \mathcal{L} \cdot w^\star \;=\; -\beta^2 \, \mu \, (1 - \bar{P})^2 / (L - 1),$$

with $\mu = \langle u, u^\star \rangle$ the current value overlap and $\bar{P}$ the softmax collision probability. It vanishes exactly at $\mu = 0$ (the unlearned saddle) and is $\Theta(1)$, dimension-free, the moment the value is learned. Projected gradient descent on the invariant slice (batch, with a fresh sample set per step and both directions constrained to the sphere) therefore escapes the saddle in a two-stage value→selection cascade (Figure 1a) and learns $w^\star$ at the information-theoretically optimal $n = \Theta(d)$ rate (Corollary 2): the infinite information exponent costs no extra power of d. The same degree-two structure is also exposed in one shot: squaring the label produces the collision probability $P(s) = \sum_i \mathrm{softmax}(\beta s)_i^2$ as its conditional mean (Figure 1b), whose leading Hermite degree is exactly two, since degree one is killed by softmax shift invariance while degree two survives with a coefficient $c_2(\beta, L) > 0$, so a single deflated spectral step on $y^2(xx^\top - I)$ recovers $w^\star$ at the same dimensional rate up to logarithmic factors (Theorem 3).

The size of the gap. The single-index literature reports *finite* gaps, with exponent k vs. 2 separating SGD's d^{k-1} from the optimal d, and the usual reading of $\mathrm{IE} = \infty$ is an infinite such gap. Our setting sits outside both: the mechanism that defines attention has infinite information exponent in the single-parameter sense, yet gradient descent on the two-parameter loss learns it at the optimal rate. The reconciliation is one structural fact, softmax's invariance to a constant shift of all logits, which erases the degree-one selection signal and leaves a degree-two object detected by the value-catalysed gradient at $n = \Theta(d)$ and by the squared-label spectral step at $n = \tilde{O}(d)$. This accounts for a real difficulty (selection training stalls at a saddle until the value is acquired) without overstating it (no infinite sample-complexity gap), and prescribes the one-shot fix. Figure 1 checks the two central predictions numerically.

Organisation. Section 2 defines the single-location softmax unit and the two exponents; Section 3 proves the conditional mean annihilation ($\mathrm{IE} = \infty$); Section 4 derives the value catalyst and the two-timescale escape; Section 5 exhibits the degree-two label transform ($\mathrm{GE} = 2$); and Section 6 gives the rate-optimal spectral algorithm with its matching information-theoretic lower bounds. The exact rank-two population matrix, the discontinuous phase transition, the global convergence of the cascade, and the detailed numerics are deferred to the appendix. All proofs are collected in Appendix A.

Contributions.

1. A minimal model (Definition 1) isolating the selection-vs-read decomposition of one softmax head, and the exact single-index dichotomy $\mathrm{IE}_{w^\star} = \infty$, $\mathrm{GE}_{w^\star} = 2$ (Theorem 2), with the generative lower bound $\mathrm{GE} \geq 2$ holding uniformly over all label transforms via softmax shift invariance.
2. The *value catalyst* (Theorem 1, Corollaries 2, 3): a closed-form value-unlocked selection gradient $\nabla_w \mathcal{L} \cdot w^\star = -\beta^2 \mu (1 - \bar{P})^2 / (L - 1)$ showing projected gradient descent on the invariant slice escapes the saddle in a value→selection cascade and learns the $\mathrm{IE} = \infty$ direction at the optimal $n = \Theta(d)$ dimensional rate, so the infinite information exponent yields no infinite gradient-descent gap. The ordering is necessary (Corollary 3).

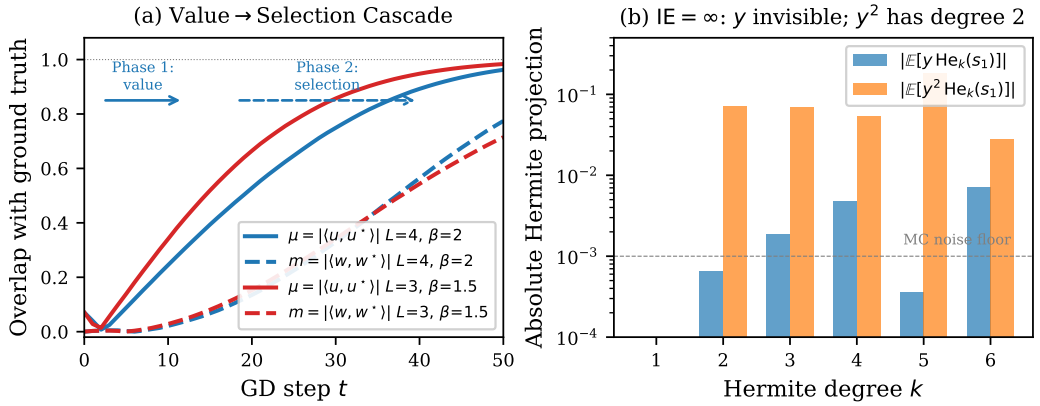


Figure 1: The core phenomenon. **(a)** Projected gradient descent on the two-parameter attention loss ($d = 60$, $N = 30{,}000$ per step, two settings of (L, β)): the value overlap $\mu = |\langle u, u^\star \rangle|$ rises first (solid, IE$= 1$), then the selection overlap $m = |\langle w, w^\star \rangle|$ follows (dashed, IE$= \infty$). This is the value→selection cascade of Theorem 1 and Corollary 2, and both overlaps reach ≈ 1 in $O(1)$ steps at this large-batch scale. The contrast with a single-index SGD that trains only w against a fixed value template, which stays at the $O(d^{-1/2})$ noise floor, is not plotted here; it is produced by `numerics/figure_comparison.py`. **(b)** Absolute Hermite projections $|\mathbb{E}[y\,\mathrm{He}_k(s_1)]|$ (blue) sit at the Monte-Carlo noise floor for every k (IE $= \infty$), while $|\mathbb{E}[y^2\,\mathrm{He}_k(s_1)]|$ (orange) has a degree-2 signal (GE $= 2$): y is invisible to key-score correlational queries; y^2 exposes $w^\star$ through those scores.

3. A one-shot $n = \tilde{O}(d)$ spectral estimator (Theorem 3) with the closed-form degree-two coefficient $c_2(\beta, L)$ (Lemma 5, Appendix H) and rank-two population matrix (Lemma 8); an information-theoretic $n = \Omega(\sigma^2 d/(\beta^2 L))$ for $\sigma > 0$ (Theorem 7, Fano over a key-sphere packing with a softmax-Lipschitz KL bound), and an all-noise oracle lower bound $n = \Omega(d/L)$ in the same projective metric as the spectral estimator (Theorem 8).
4. A *discontinuous phase transition* at orthogonality (Theorem 4): the information exponent jumps from ∞ to 1 the moment the key–value alignment $\rho = \langle w^\star, u^\star \rangle$ departs from zero, with $\mathbb{E}[y\,\mathrm{He}_1(s_1)] = \rho/L$ independent of β; and formal iteration-and-sample complexity for the cascade (Theorems 5, 6) via an exact two-dimensional slice potential, giving $O(1)$ population steps and $n = \Theta(d)$ per step.
5. A careful account of the multi-head case (Theorem 9, Appendix I): both additive and multiplicative combination of orthogonal-subspace heads keep every selection direction at IE $= \infty$, GE $= 2$ with dimensional exponent one; only the sample-complexity constant decays, geometrically in depth for the multiplicative stack. A CPU numerical suite (Appendix F) checks the stated formulas and scaling laws.

Related work. *Information and generative exponents.* The information exponent governs correlational (CSQ) learning of single-index targets (Ben Arous et al., 2021; Bietti et al., 2022): an IE-k target traps online SGD for $\tilde{\Theta}(d^{k-1})$ samples. The generative exponent, introduced by Damian et al. (2024), characterises the SQ-optimal rate after label preprocessing: $n = \tilde{\Theta}(d^{\max(1,\mathrm{GE}/2)})$. Damian et al. (2025) extend the generative exponent to multi-index models and Defilippis et al. (2025) construct matching spectral estimators; Dandi et al. (2024) show that gradient descent which reuses batches at a finite step size escapes the information-exponent barrier that constrains a single pass; and a hierarchical line (Abbe et al., 2023; Bardone & Goldt, 2024) studies how correlated latent variables accelerate learning of harder directions through easier ones. The cited works analyze finite IE–GE gaps. We identify an instance of IE $= \infty$ with GE $= 2$ in a natural architecture, and show that projected gradient descent escapes via a structural coupling, the value catalyst, rather than the generic GE mechanism. It is an extreme form of the sliding-acceleration principle of Bardone & Goldt (2024), with an easy IE$= 1$ direction resolving the infinite gap on an IE$= \infty$ one.

Learning and training dynamics of attention. The expressivity–evaluation separation between softmax and linear attention is established by Jelassi et al. (2024); Arora et al. (2024); the SETH-hardness of computing attention is due to Keles et al. (2023). On the learning side, Chen et al. (2024) prove convergence of multi-head softmax attention for in-context linear regression via gradient flow, but regress directly over attention weights and do not address the single-index exponent structure we identify. Our result gives a precise mechanism for the softmax-induced slowdown near initialisation (the selection gradient is exactly zero at the equator, an $\mathrm{IE} = \infty$ saddle) and its resolution via the value catalyst, yielding the $n = \Theta(d)$ rate. To our knowledge, prior work has not identified an infinite information exponent with finite generative exponent in an attention model, nor shown that coupled-parameter gradient descent can circumvent that barrier without label preprocessing.

2 Setup and the single-location softmax unit

We study the sample complexity of learning the most elementary nontrivial attention computation: a single softmax head that *selects* a token by one direction and *reads off* a value along another. Throughout, d is the ambient dimension (the asymptotic parameter), $L \geq 2$ is a fixed context length, and $\beta > 0$ a fixed inverse temperature.

Definition 1 (Single-location softmax unit)**.** *Let $w^\star, u^\star \in \mathbb{S}^{d-1}$ be orthonormal. Draw tokens $x_1, \dots, x_L \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and stack $X = (x_1, \dots, x_L)$. Write the* key *(selection) and* value *(read) scores*

$$s_i = \langle w^\star, x_i \rangle, \qquad v_i = \langle u^\star, x_i \rangle, \qquad i \in [L].$$

The unit outputs the scalar

$$y \;=\; \sum_{i=1}^{L} a_i(s)\, v_i, \qquad a(s) = \mathrm{softmax}(\beta s) \in \Delta^{L-1}, \quad a_i(s) = \frac{e^{\beta s_i}}{\sum_j e^{\beta s_j}}.$$

Because $w^\star \perp u^\star$ and X is Gaussian, the vectors $s = (s_1, \dots, s_L)$ and $v = (v_1, \dots, v_L)$ are *independent*, each $\mathcal{N}(0, I_L)$. This is the only distributional fact we use about X; everything below is exact, not asymptotic, unless stated.

The learning task is: given i.i.d. samples $(X^{(t)}, y^{(t)})_{t \leq n}$ (possibly with independent label noise $\xi \sim \mathcal{N}(0, \sigma^2)$), recover the selection direction $w^\star$ (equivalently, identify the head up to the easy value direction). The value direction is the easy half, as Lemma 7 shows; the entire difficulty is concentrated in $w^\star$.

We measure hardness through the two exponents that govern Gaussian single-/multi-index learning.

Definition 2 (Information and generative exponents)**.** *Let $g : \mathbb{R}^L \to \mathbb{R}$ and let $s \sim \mathcal{N}(0, I_L)$. For a unit direction $\theta \in \mathbb{S}^{L-1}$ consider the scalar projection $z = \langle \theta, s \rangle$. The* information exponent *of g along θ is*

$$\mathrm{IE}_\theta(g) = \min\Big\{k \geq 1 :\ \mathbb{E}\big[g(s)\ \mathrm{He}_k(z)\big] \neq 0\Big\},$$

($\min \emptyset = \infty$), where He_k is the degree-k probabilists' Hermite polynomial. The generative exponent *is*

$$\mathrm{GE}_\theta(g) = \min_T\ \mathrm{IE}_\theta(T \circ g),$$

the minimum information exponent attainable by any (square-integrable) transformation $T : \mathbb{R} \to \mathbb{R}$ of the label.

We attach exponents to an *input*-space direction through its induced scores. For $w^\star \in \mathbb{S}^{d-1}$ the key scores are $s = (\langle w^\star, x_i \rangle)_{i \in [L]} \sim \mathcal{N}(0, I_L)$, and we set

$$\mathrm{IE}_{w^\star}(y)\ :=\ \min\Big\{k \geq 1 :\ \mathbb{E}\big[y\ \mathrm{He}_k(\langle \theta, s \rangle)\big] \neq 0 \text{ for some } \theta \in \mathbb{S}^{L-1}\Big\},$$

with $\mathrm{GE}_{w^\star}(y) = \min_T \mathrm{IE}_{w^\star}(T \circ y)$ defined analogously. This is exactly the signal a single-index probe of the key scores can access; the ambient dimension d never appears in the exponent itself, only in the sample complexity it controls. The information exponent governs correlational (CSQ) methods—online SGD on an IE-k target needs $n = \tilde{\Theta}(d^{\max(1, k-1)})$ samples (Ben Arous et al., 2021;

Bietti et al., 2022)—while the generative exponent controls the best statistical-query rate after label preprocessing, $n = \tilde{\Theta}(d^{\max(1,\mathrm{GE}/2)})$ (Damian et al., 2024). A gap IE $\gg$ GE is precisely a gap between gradient descent on the raw loss and the best efficient estimator.

Our main object is the following dichotomy for the unit of Definition 1.

3 THE SELECTION DIRECTION IS CORRELATIONALLY INVISIBLE

Lemma 1 (Conditional mean annihilation)**.** *For the unit of Definition 1,* $\mathbb{E}[\, y \mid s \,] = 0$ *almost surely. Consequently, for* every *measurable* g *with* $\mathbb{E}[g(s)^2] < \infty$*,*

$$\mathbb{E}\big[y\, g(s)\big] = 0.$$

See proof on page 11.

Corollary 1 (Information exponent $= \infty$)**.** *Along every direction* $\theta \in \mathbb{S}^{L-1}$ *(in particular along any single key coordinate* s_i*),* $\mathrm{IE}_\theta(y) = \infty$*. In words:* no *correlational query against a function of the key scores carries any signal about* $w^\star$*.*

See proof on page 11.

The mechanism is the antisymmetry between the read and select directions: the output is *linear* in the value scores, and the value scores are mean-zero and independent of the selection. Two consequences for learning follow at once. (i) Any correlational statistical query $\phi \mapsto \mathbb{E}[y\, \phi(s)]$ against a function of the key scores returns exactly zero. (ii) The gradient of *any single-index surrogate*—a student $\hat{y} = f(\langle w, x_1\rangle, \ldots, \langle w, x_L\rangle)$ that reads off the selected token through a *fixed* value template f—has zero population component along $w^\star$, because that gradient is itself a correlational query against functions of the student key scores. Under the standard reading of the information exponent this is the signature of a direction gradient descent can never find. Section 4 shows that this reading is *wrong* for the loss attention actually trains: coupling a learnable value parameter to the key parameter reintroduces an explicit first-order drift along $w^\star$, so plain gradient descent escapes at the optimal dimensional rate.

Remark 1 (The orthogonal case is an endpoint, not a degeneracy)**.** *The blindness is exactly first order in the value direction: if the value direction is rotated to share an angle* $\rho = \langle u^\star, w^\star\rangle$ *with the key direction, then* $\mathbb{E}[y \mid s] = \rho \sum_i a_i(s) s_i \neq 0$ *and a degree-*1 *correlational signal reappears with magnitude* $\Theta(\rho)$ *(checked numerically in* `numerics/probe2.py`*). The* $\mathrm{IE} = \infty$ *statement is the* $\rho = 0$ *endpoint of a continuous family; it is the generic situation when selection and value are encoded in different subspaces, as in a real attention head with separate* W_K *and* W_V*.*

4 THE VALUE CATALYST: PLAIN GRADIENT DESCENT LEARNS THE SELECTION DIRECTION

Corollary 1 is a statement about a *single* parameter: the key direction read against a function of the key scores. The loss that attention actually trains has a second parameter, the value direction, and the two are coupled inside one softmax. Once the value direction is even partially aligned, the joint key gradient acquires an *explicit, dimension-free* component along $w^\star$; we compute it in closed form. The information exponent is still ∞ in the single-parameter sense of Corollary 1, but no infinite gap between gradient descent and the optimal estimator survives: plain gradient descent escapes the saddle in a two-stage value→selection cascade.

The natural two-parameter student. Mirroring Definition 1, parametrise a student by a key direction $w \in \mathbb{S}^{d-1}$ and a value direction $u \in \mathbb{S}^{d-1}$,

$$\hat{y}(w, u) \;=\; \sum_{i=1}^{L} a_i(\langle w, x\rangle)\, \langle u, x_i\rangle, \qquad a(\cdot) = \mathrm{softmax}(\beta\, \cdot),$$

and train the population square loss $\mathcal{L}(w, u) = \frac{1}{2}\, \mathbb{E}\big[(y - \hat{y}(w, u))^2\big]$. Write the student key and value scores $t_i = \langle w, x_i\rangle$, $b_i = \langle u, x_i\rangle$ and the student weights $a = a(t)$. Two scalar order parameters track progress: the value overlap $\mu = \langle u, u^\star\rangle$ and the selection overlap $\langle w, w^\star\rangle$.

Lemma 2 (Softmax Jacobian and the key gradient)**.** *The key gradient of the student output is*

$$\nabla_w \hat{y} \;=\; \beta \sum_{i=1}^{L} b_i\, a_i \left(x_i - \textstyle\sum_j a_j x_j \right),$$

and consequently $\nabla_w \mathcal{L} = -\,\mathbb{E}\big[(y-\hat{y})\,\nabla_w \hat{y}\big]$.

See proof on page 11.

We evaluate $\nabla_w \mathcal{L} \cdot w^\star$ at the most pessimistic place: the *equator*, where the student key is orthogonal to the teacher key, $w \perp w^\star$ (so the current selection overlap is zero and there is nothing to exploit from the key alone) and also $w \perp u^\star$, with student value $u = \mu u^\star + \sqrt{1-\mu^2}\, e$, $e \perp \{w^\star, u^\star, w\}$. The extra condition $w \perp u^\star$ places the configuration on the orthogonal invariant slice of Lemma 11 (the overlaps $q = \langle w, u\rangle$, $r = \langle w, u^\star\rangle$, $p = \langle w^\star, u\rangle$ all vanish); it is what makes the value scores $v_i = \langle u^\star, x_i\rangle$ jointly independent of both key fields s, t, used below. This is exactly the configuration reached at random initialisation, where all pairwise overlaps are $O(d^{-1/2})$ and the selection overlap is $O(d^{-1/2})$.

Theorem 1 (Closed-form value-unlocked selection gradient)**.** *At an equator configuration $w \perp w^\star$, $w \perp u^\star$ with value overlap $\mu = \langle u, u^\star\rangle$,*

$$\boxed{\nabla_w \mathcal{L} \cdot w^\star \;=\; -\,\beta^2\, \mu\, \frac{(1-\bar{P})^2}{L-1}}$$

where $\bar{P} = \mathbb{E}[P(s)] = \mathbb{E}\sum_i a_i(s)^2$ is the softmax collision probability. In particular the component vanishes exactly *at $\mu = 0$ (the unlearned saddle), is linear in the value overlap μ, and is $\Theta(1)$ in the dimension d the moment the value is learned ($\mu = \Theta(1)$). The gradient* descent *step $-\nabla_w \mathcal{L}$ therefore has strictly positive overlap with $w^\star$ whenever $\mu > 0$.*

See proof on page 11.

Corollary 2 (No infinite gradient-descent gap on the invariant slice)**.** *Run projected gradient descent on $\mathcal{L}(w, u)$ on the invariant slice of Lemma 11, with $\mu_0 > 0$ and $m_0 \in [0, 1)$. The value direction has information exponent one (Lemma 7), so the value overlap rises to $\mu = \Theta(1)$ in $O(1)$ population steps independently of the (vanishing) selection overlap. By Theorem 1, the moment $\mu = \Theta(1)$ the selection gradient acquires a dimension-free component $|\nabla_w \mathcal{L} \cdot w^\star| = \beta^2 \mu (1-\bar{P})^2/(L-1) = \Theta(1)$ pointing toward $w^\star$. Hence the selection overlap escapes the $O(d^{-1/2})$ equator and rises to one. Plain gradient descent therefore learns $w^\star$ at the information-theoretically optimal $n = \Theta(d)$ rate (the rate at which the $\Theta(1)$ population drift is resolvable from finite samples), with* no *extra power of d: the $\mathrm{IE} = \infty$ of Corollary 1 does* not *translate into an infinite sample-complexity gap once the value parameter is free to act as a catalyst.*

See proof on page 12.

Remark 2 (How the catalyst relates to the generative exponent)**.** *The single-parameter statement $\mathrm{IE}_{w^\star} = \infty$ makes the head invisible to probes of the key scores alone. The two-parameter loss couples the value direction to those scores through a degree-two term, bilinear in μ and $1 - \bar{P}$, matching $\mathrm{GE} = 2$ in Theorem 2. Gradient descent therefore accesses the same degree-two structure as the squared-label spectral step. It uses $n = \Theta(d)$ samples per step on the analyzed slice; the proved spectral bound is $\tilde{O}(d)$.*

Corollary 3 (The cascade ordering is necessary)**.** *Under projected gradient descent on the invariant slice, starting with $\mu_0 = O(d^{-1/2})$, $m_0 = O(d^{-1/2})$ and $n = \Theta(d)$ samples per step, the selection overlap satisfies $m_t = O(d^{-1/2})$ for all steps t at which $\mu_t = O(d^{-1/2})$. In other words, the selection direction* cannot *escape the initialisation noise before the value direction is acquired: the value→selection ordering of the cascade is forced, not incidental.*

See proof on page 12.

Remark 3 (Initialization scope)**.** *The formal convergence results in this paper concern the invariant orthogonal slice with nonnegative key overlap and positive value overlap. The experiments converge from unrestricted random initializations, but controlling off-slice coordinates and arbitrary initial signs remains open.*

5 A DEGREE-TWO LABEL TRANSFORM RECOVERS THE SELECTION DIRECTION IN ONE SHOT

Correlational blindness would be a curiosity if the direction were genuinely unrecoverable. The opposite is true: squaring the label exposes $w^\star$ at the information-theoretically optimal degree.

Lemma 3 (Squared label is the collision probability)**.** *For the unit of Definition 1, with label noise $\xi \sim \mathcal{N}(0, \sigma^2)$ independent of X,*

$$\mathbb{E}\big[(y+\xi)^2 \mid s\big] = P(s) + \sigma^2, \qquad P(s) := \sum_{i=1}^{L} a_i(s)^2,$$

where $P(s)$ is the collision probability *of the softmax distribution $a(s)$ (the probability that two independent draws from $a(s)$ coincide).*

See proof on page 13.

Thus the transformed label $T(y) = y^2$ has conditional expectation an explicit, nonconstant function of the key scores. Its Hermite content along s determines the generative exponent. Softmax is *shift invariant*, which annihilates the degree-one part.

Lemma 4 (Degree-one annihilation by shift invariance)**.** *$P(s)$ is invariant under $s \mapsto s + c\mathbf{1}$ for all $c \in \mathbb{R}$ ($\mathbf{1} = (1, \dots, 1)$). Consequently its first Hermite layer vanishes:*

$$\mathbb{E}\big[P(s)\, s_i\big] = 0 \quad \textit{for every } i \in [L],$$

and more generally $\mathbb{E}[P(s)\,\langle \theta, s\rangle] = 0$ for all $\theta \in \mathbb{R}^L$.

See proof on page 13.

Lemma 5 (Degree-two coefficient is strictly positive)**.** *For all $\beta > 0$ and $L \geq 2$,*

$$c_2(\beta, L) \;:=\; \tfrac{1}{2}\,\mathbb{E}\big[P(s)\ \mathrm{He}_2(s_1)\big] \;=\; \tfrac{1}{2L}\,\mathrm{Cov}\big(P(s),\ \|s_\perp\|^2\big) \;>\; 0,$$

where $s_\perp = s - \bar{s}\,\mathbf{1}$ is the projection of s onto the contrast subspace $\mathbf{1}^\perp$ and $\bar{s} = \frac{1}{L}\sum_i s_i$. This c_2 is the degree-two Hermite coefficient of the collision probability P along a single key coordinate; its appearance as the exact selection-direction spike strength is made precise in Lemma 8.

See proof on page 13.

Remark 4 (Closed form for $L = 2$)**.** *For $L = 2$ the unit reduces to a one-dimensional contrast $r = (s_1 - s_2)/\sqrt{2} \sim \mathcal{N}(0,1)$, and $P = \frac{1}{2}\big(1 + \tanh^2(\frac{\beta}{\sqrt{2}} r)\big)$. Writing $s_1 = (r + \bar{r})/\sqrt{2}$ with $\bar{r} = (s_1 + s_2)/\sqrt{2}$ independent of r gives $\mathbb{E}[\mathrm{He}_2(s_1) \mid r] = \frac{1}{2}\,\mathrm{He}_2(r)$, hence*

$$c_2(\beta, 2) = \tfrac{1}{2}\mathbb{E}[P\ \mathrm{He}_2(s_1)] = \tfrac{1}{8}\,\mathbb{E}\big[\tanh^2(\tfrac{\beta}{\sqrt{2}} r)\ \mathrm{He}_2(r)\big],$$

a strictly positive, bell-shaped function of β that vanishes as $\beta \to 0$ (no selection) and as $\beta \to \infty$ (hard argmax becomes a measure-zero event for the contrast), peaking near $\beta \approx 2$ at $c_2 \approx 0.044$. This matches `numerics/ge_verify.py` (c_2 rises then falls, peak around $\beta = 1$–2).

Lemma 6 (Quantitative scaling of $c_2(\beta, L)$)**.** *The degree-two coefficient obeys, for every $L \geq 2$,*

$$\textit{(small } \beta) \quad c_2(\beta, L) = \frac{L-1}{L^3}\,\beta^2 + O(\beta^4), \qquad \textit{(large } \beta) \quad c_2(\beta, L) \xrightarrow[\beta \to \infty]{} 0.$$

Consequently c_2 is bounded on $(0, \infty)$, vanishes quadratically at $\beta = 0$ and decays to 0 as $\beta \to \infty$, and attains a strictly positive interior maximum; on any compact β-interval $[\beta_0, \beta_1] \subset (0, \infty)$ it is bounded below by a positive constant depending only on (β_0, β_1, L).

See proof on page 13.

Remark 5 (The exact scope of the word "optimal")**.** *The recovery threshold of Theorem 3 is $n = C(\beta, L, \sigma, \varepsilon)\, d\ \mathrm{polylog}\, d$ for fixed accuracy, with problem-dependent factor proportional to $(1 + \sigma^2)^2 c_2^{-2}$ up to universal and logarithmic factors. Lemma 6 makes the factors explicit: the* upper-bound exponent *in d is 1 for every fixed (β, L, σ). The oracle lower bound of Theorem 8 shows*

that this exponent is optimal at every noise level; for fixed $\sigma > 0$, the Fano bound also tracks problem parameters. The prefactor *degrades as the selection signal weakens, $C = \Theta((1+\sigma^2)^2 L^6/((L-1)^2\beta^4))$ at small β and $C \to \infty$ as $\beta \to \infty$. Three caveats delimit the claim, and we state all three. First, the upper and lower bounds coincide in the d-exponent but not in the constants: the upper prefactor scales as $(1+\sigma^2)^2/\beta^4$ at small β against the lower bound's $\sigma^2/(\beta^2 L)$, and we do not close that gap. Second, the spectral upper bound carries logarithmic factors from truncation and concentration of the weighted covariance, so it is $\tilde{O}(d)$ rather than $O(d)$; the gradient-descent bound of Theorem 6 does not, and it is the one the title refers to. Third, for fixed L the matching is between $\Omega(d)$ and $O(d)$ for gradient descent and $\tilde{O}(d)$ for the spectral step. Our minimax optimality claim is about this dimensional exponent. The scaling laws and interior peak are checked across $L \in \{2, 4, 8\}$ in* `numerics/c2_scaling.py`.

Theorem 2 (Exponent dichotomy for softmax selection)**.** *For the single-location softmax unit (Definition 1) with $w^\star \perp u^\star$, the selection direction $w^\star$ satisfies*

$$\mathrm{IE}_{w^\star}(y) = \infty, \qquad \mathrm{GE}_{w^\star}(y) = 2.$$

The upper exponent $\mathrm{GE} = 2$ is achieved by the explicit transform $T(y) = y^2$, whose conditional mean $P(s)$ has leading Hermite degree exactly two (Lemmas 4, 5); the lower exponent $\mathrm{IE} = \infty$ is Corollary 1.

See proof on page 14.

The lower bound $\mathrm{GE} \geq 2$ is *not* specific to the square; it holds for *all* label transforms simultaneously, because softmax shift invariance is inherited by the conditional law of the output. That is what makes this a statement about the exponent rather than about one cleverly chosen T.

Remark 6 (Distributional scope, and where normality is really used)**.** *The two halves of the dichotomy have different distributional status, and it is worth separating them.*

The infinite information exponent (Lemma 1, Corollary 1) uses only two properties of the token distribution: the value scores have mean zero given the key scores, and the value scores are independent of the key scores. Thus $\mathbb{E}[y \mid s] = 0$ is not restricted to Gaussian tokens whenever these two conditions hold. In coordinate axes under four product laws, projected GD reaches unit key and value overlaps for Gaussian, Rademacher, uniform and Student-t_5 coordinates, and the measured equator drift is within 15% of the Gaussian closed form (`numerics/nongaussian_tokens.py`*). This experiment does not establish rotation-uniform robustness for arbitrary non-Gaussian laws.*

The lower bound $\mathrm{GE} \geq 2$ is not robust in this way, and we state the boundary explicitly. Degree-one annihilation (Lemma 4) rests on shift invariance plus *the Gaussian orthogonal decomposition: shift invariance makes h_T a function of the contrasts $s_\perp$, exchangeability forces $\mathrm{Cov}(h_T, s_j - \bar{s}) = 0$ for every j, and therefore $\mathbb{E}[h_T(s)s_1] = \mathrm{Cov}(h_T, \bar{s})$, which vanishes exactly when $\bar{s}$ is independent of $s_\perp$. Among i.i.d. coordinates that characterises the Gaussian. For Rademacher or uniform tokens the degree-one projection of y^2 is measurably nonzero and GE drops to one, i.e. the selection direction becomes* easier*, not harder. The degree-two coefficient itself survives: $\mathrm{Cov}(P, \|s_\perp\|^2) > 0$ under every law we tested. So the phenomenon this paper is about, an $\mathrm{IE} = \infty$ direction rescued at the optimal rate by a value catalyst, is structural; the exact value $\mathrm{GE} = 2$ is Gaussian.*

6 Matching algorithmic and SQ consequences

Theorem 3 (Spectral recovery at the optimal rate)**.** *There is an estimator (a single partial-trace spectral step on the squared label) that, for every fixed accuracy $\varepsilon > 0$, returns $\widehat{w}$ with $\langle \widehat{w}, w^\star \rangle^2 \geq 1 - \varepsilon$ with high probability from $n = C(\beta, L, \sigma, \varepsilon)\, d \operatorname{polylog} d$ samples. The overlap tends to one whenever $n/(d \operatorname{polylog} d) \to \infty$. For fixed ε, the problem-dependent factor can be taken proportional to $c_2(\beta, L)^{-2}(1+\sigma^2)^2$ up to universal and logarithmic factors. No SQ algorithm using only correlational queries against functions of the key scores recovers $w^\star$ with any number of samples. Conversely,* every *estimator, with no restriction to statistical queries, requires $n = \Omega(\sigma^2 d/(\beta^2 L))$ samples to recover $w^\star$ to constant accuracy (Theorem 7). Hence the spectral estimator is minimax rate optimal in the ambient dimension d up to logarithmic factor, with a noise-dependent constant. (The gradient-descent route of Theorem 6 carries no such logarithm: it reaches the same accuracy at $n = \Theta(d)$ samples per step and $O(1)$ steps.)*

See proof on page 14.

6.1 Complementary minimax lower bounds

An oracle lower bound gives the linear dimensional exponent for every $\sigma \geq 0$, even after revealing all exact key scores. For fixed $\sigma > 0$, a Fano argument over a local packing of the key sphere, with a per-sample Kullback–Leibler bound obtained from softmax Lipschitzness, forces $n = \Omega(\sigma^2 d/(\beta^2 L))$ for any estimator (Theorem 7). The all-noise oracle and deterministic exact-identifiability statements are Theorem 8. Both statements, their proofs, and the numerical checks are deferred to Appendix E.

Lemma 7 (The value direction is easy)**.** $\mathbb{E}[y\, v_i] = \mathbb{E}[a_i(s)] > 0$ *and the first-moment vector* $b = \mathbb{E}[y\, x_i]$ *(summed over* i*) satisfies* $b \parallel u^\star$ *with* $\|b\| = \sum_i \mathbb{E}[a_i(s)] = 1 > 0$*. Hence* $\mathrm{IE}_{u^\star}(y) = 1$ *and* $u^\star$ *is recovered from* $n = \tilde{\Theta}(d)$ *samples by a single moment.*

See proof on page 15.

7 Numerical confirmation

The supplementary `numpy` scripts check the principal formulas and scaling laws on CPU. Figure 1 shows the two central phenomena: single-index SGD stalls at the noise floor while the two-parameter cascade succeeds, and the squared label exposes the degree-two signal the raw label hides. The script-by-script account is collected in Appendix F (Tables 1–2).

8 Discussion

A single attention head couples two targets of opposite difficulty: an easy read and a correlationally invisible select. Softmax shift invariance erases the degree-one selection signal, giving that direction infinite information exponent, so before the value is acquired the selection gradient is exactly zero. The value direction is nevertheless learnable and acts as a catalyst that unlocks an explicit $\Theta(1)$ selection gradient (Theorem 1), so projected gradient descent on the analyzed slice escapes in a value→selection cascade at the optimal $n = \Theta(d)$ dimensional rate, matched one-shot up to logarithmic factors by a squared-label spectral step. Theorem 8 gives an all-noise $\Omega(d/L)$ oracle lower bound in projective risk; for fixed $\sigma > 0$, Theorem 7 refines it with explicit (σ, β, L) dependence. In this model the difficulty is confined to the selection role and is key-score-correlational: the softmax symmetry that opens the correlational gap closes the algorithmic one through the same degree-two structure.

The multi-head case (Appendix I) preserves the dichotomy: combining orthogonal-subspace heads additively or multiplicatively leaves every selection direction at $\mathrm{IE} = \infty$, $\mathrm{GE} = 2$. The exponent can rise only if heads *share* a selection subspace under a symmetry stronger than shift invariance; identifying the minimal such coupling remains open. The phase transition (Theorem 4) is β-independent, with degree-one signal exactly ρ/L: the difficulty is a geometric property of the key–value subspace relationship, and a head with $|\rho| \ll 1$ behaves like $\rho = 0$ until $n \sim d/\rho^2$.

Limitations. Our model is deliberately minimal: one head, isotropic Gaussian tokens, orthogonal key–value subspaces. Real attention stacks H heads over T layers, operates on structured embeddings, and shares parameters across query, key, and value projections. The experiments in `numerics/nongaussian_tokens.py` show the cascade under three non-Gaussian product distributions, but shared subspaces and structured embeddings remain open. The constants $C(\beta, L)$ in our rate can be large (the c_2^{-2} factor), so the $n = \Theta(d)$ scaling concerns the dimension dependence, not the absolute cost. The convergence proof is restricted to the invariant slice; the random-initialization experiments do not establish off-slice stability. Finite-width effects and normalisation layers are also out of scope.

Reproducibility statement

The paper's formal results are analytic. The theorem, lemma and corollary statements give their hypotheses, with proofs in Appendix A or inline in the appendix section where it appears; the auxiliary inequalities the proofs invoke are isolated and proved in Appendix G. The single modelling assumption (orthonormal key/value directions, i.i.d. standard Gaussian tokens) is Definition 1, and

Definition 3 states the tilted variant used for the phase transition. Seeded, CPU-only `numpy` scripts check the principal formulas and scaling laws; the scripts are listed against the predictions they verify in Table 1 of Appendix F, are included in the supplementary material under `numerics/`, and are reproduced end to end by a single command (`bash numerics/run_all.sh`), which also regenerates Figure 1. No datasets, no training runs and no accelerators are involved.

ETHICS STATEMENT

This is a theoretical study of the sample complexity of a two-parameter attention unit under a Gaussian input model. It involves no human subjects, no personal or sensitive data, and no released models or datasets. We are not aware of foreseeable harms arising from the results; the analysis concerns when a learning problem is tractable, and is neutral with respect to application. The authors have read and adhere to the ICLR Code of Ethics.

AI USE STATEMENT

We used general-purpose large language models to assist with conceptual brainstorming, candidate derivations, implementation and debugging of the verification scripts, literature retrieval, and drafting and editing the manuscript. We did not use generative AI to produce data; this work uses only synthetic samples drawn directly from the distributions in the stated model. The authors inspected the final arguments, reference entries, code and numerical outputs and take responsibility for the complete submission, including material produced with AI assistance.

REFERENCES

Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. SGD learning on neural networks: Leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory (COLT)*, 2023.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *International Conference on Machine Learning (ICML)*, 2024.

Lorenzo Bardone and Sebastian Goldt. Sliding down the stairs: How correlated latent variables accelerate learning with neural networks. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 3024–3045, 2024.

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

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *Conference on Learning Theory (COLT)*, 2024.

Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. The computational complexity of learning Gaussian single-index models. In *Conference on Learning Theory (COLT)*, 2024.

Alex Damian, Jason D. Lee, and Joan Bruna. The generative leap: Tight sample complexity for efficiently learning Gaussian multi-index models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 9991–10016, 2024.

Leonardo Defilippis, Yatin Dandi, Pierre Mergny, Florent Krzakala, and Bruno Loureiro. Optimal spectral transitions in high-dimensional multi-index models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning (ICML)*, 2024.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *International Conference on Algorithmic Learning Theory (ALT)*, 2023.

Jiří Matoušek. *Using the Borsuk–Ulam Theorem: Lectures on Topological Methods in Combinatorics and Geometry*. Universitext. Springer, 2003.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8:1–230, 2015.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

A Deferred proofs

Proof of Lemma 1. Condition on s. The weights $a_i(s)$ are then deterministic, while $v = (v_1, \dots, v_L)$ is independent of s with $\mathbb{E}[v_i] = 0$. Hence $\mathbb{E}[y \mid s] = \sum_i a_i(s)\,\mathbb{E}[v_i \mid s] = \sum_i a_i(s) \cdot 0 = 0$. The second claim is the tower property: $\mathbb{E}[y\, g(s)] = \mathbb{E}[g(s)\,\mathbb{E}[y \mid s]] = 0$. □

Proof of Corollary 1. Take $g(s) = \mathrm{He}_k(\langle \theta, s\rangle)$ in Lemma 1: the projection vanishes for all $k \geq 1$, so the defining set in Definition 2 is empty. □

Proof of Lemma 2. The softmax Jacobian is $\partial a_i/\partial t_k = \beta\, a_i(\delta_{ik} - a_k)$, so $\partial \hat{y}/\partial t_k = \beta\, a_k\big(b_k - \sum_j a_j b_j\big)$ and $\partial t_k/\partial w = x_k$. Chain rule gives $\nabla_w \hat{y} = \beta \sum_k a_k(b_k - \sum_j a_j b_j) x_k = \beta \sum_k b_k a_k (x_k - \sum_j a_j x_j)$, the centred form quoted. The loss gradient is the standard chain rule on $\frac{1}{2}\mathbb{E}(y - \hat{y})^2$. □

Proof of Theorem 1. Write $g = \nabla_w \mathcal{L} = -\mathbb{E}[(y - \hat{y})\nabla_w \hat{y}]$ and project on $w^\star$. Since $w \perp w^\star$, the student key scores $t_i = \langle w, x_i\rangle$ are independent of the teacher key scores $s_i = \langle w^\star, x_i\rangle$, and we decompose $\langle \nabla_w \hat{y}, w^\star\rangle = \beta \sum_i b_i a_i (s_i - \bar{s}_a)$ with $\bar{s}_a = \sum_j a_j s_j$ (using $\langle x_i, w^\star\rangle = s_i$).

The $\hat{y}$ self term vanishes. The contribution $\mathbb{E}[\hat{y}\,\langle \nabla_w \hat{y}, w^\star\rangle]$ involves only the student fields (t, b) and the teacher key scores s through the linear factor $(s_i - \bar{s}_a)$. As s is independent of (t, b) and mean zero, $\mathbb{E}[(s_i - \bar{s}_a) \cdot (\text{function of } t, b)] = 0$. So only the teacher term $\mathbb{E}[y\,\langle \nabla_w \hat{y}, w^\star\rangle]$ survives, and $g \cdot w^\star = -\beta\,\mathbb{E}\big[y \sum_i b_i a_i (s_i - \bar{s}_a)\big]$.

Reduce y and b to their $u^\star$ components. The teacher label is $y = \sum_j a_j(s)\, v_j$ with $v_j = \langle u^\star, x_j\rangle$, and the student value score is $b_i = \langle u, x_i\rangle = \mu v_i + \sqrt{1 - \mu^2}\,\langle e, x_i\rangle$. The e part is independent of everything in sight ($e \perp w^\star, u^\star, w$) and mean zero, so it drops out of the expectation, contributing the factor μ:

$$g \cdot w^\star = -\beta\,\mu\,\mathbb{E}\Big[\Big(\sum_j a_j(s) v_j\Big) \sum_i v_i\, a_i(t)\,(s_i - \bar{s}_a)\Big].$$

Now condition on both key fields (s, t). Here we use $w \perp u^\star$ (and the given $w^\star \perp u^\star$): the value scores $v_i = \langle u^\star, x_i\rangle$ are then jointly Gaussian with, and orthogonal to, both $s = \langle w^\star, X\rangle$ and $t = \langle w, X\rangle$, hence independent of (s, t) with $\mathbb{E}[v_i v_j \mid s, t] = \mathbb{E}[v_i v_j] = \delta_{ij}$; so $\mathbb{E}[(\sum_j a_j(s) v_j)(\sum_i v_i a_i(t)(s_i - \bar{s}_a)) \mid s, t] = \sum_i a_i(s) a_i(t)(s_i - \bar{s}_a)$. Thus

$$g \cdot w^\star = -\beta\,\mu\,\mathbb{E}_{s,t}\Big[\sum_i a_i(s) a_i(t)\,(s_i - \bar{s}_a)\Big], \qquad \bar{s}_a = \textstyle\sum_j a_j(t) s_j.$$

Integrate out the independent student key t. Because $t \perp s$, take the expectation over t first. Both $a(t)$ enters; using exchangeability of the L tokens and writing $\bar{P} = \mathbb{E}\sum_i a_i(t)^2$, $\mathbb{E}\, a_i(t) = 1/L$, and $\mathbb{E}[a_i(t)a_j(t)] = (1-\bar{P})/(L(L-1))$ for $i \neq j$ (the off-diagonal average, fixed by $\sum_{i,j}\mathbb{E}[a_i a_j] = 1$ and the diagonal $\bar{P}$),

$$\mathbb{E}_t\Big[a_i(t)\big(s_i - \textstyle\sum_j a_j(t)s_j\big)\Big] = \frac{1}{L}s_i - \sum_j \mathbb{E}_t[a_i(t)a_j(t)]\, s_j.$$

Explicit collapse. We need $\mathbb{E}_s\big[\sum_i a_i(s)\big(\frac{1}{L}s_i - \sum_j \mathbb{E}_t[a_i(t)a_j(t)]s_j\big)\big]$. The two t-moments (from exchangeability and $\sum_{i,j}\mathbb{E}[a_i a_j] = 1$) are $\mathbb{E}_t[a_i^2] = \bar{P}/L$ and $\mathbb{E}_t[a_i a_j] = (1-\bar{P})/(L(L-1))$ for $i \neq j$. Split the sum over j:

$$\begin{aligned}\tfrac{1}{L}s_i - \sum_j \mathbb{E}_t[a_i a_j]s_j &= \tfrac{1}{L}s_i - \tfrac{\bar{P}}{L}s_i - \tfrac{1-\bar{P}}{L(L-1)}\sum_{j\neq i} s_j \\ &= \tfrac{1-\bar{P}}{L}\, s_i - \tfrac{1-\bar{P}}{L(L-1)}\sum_{j\neq i} s_j \\ &= \tfrac{1-\bar{P}}{L(L-1)}\big(Ls_i - \textstyle\sum_j s_j\big).\end{aligned}$$

Weighting by $a_i(s)$, summing over i, and using $\sum_i a_i = 1$:

$$\sum_i a_i(s)\big(Ls_i - \textstyle\sum_j s_j\big) = L\sum_i a_i(s)s_i - \sum_j s_j = LG(s) - \mathbf{1}^\top s,$$

where $G(s) := \sum_i a_i(s)s_i$. Taking $\mathbb{E}_s$: since $s \sim \mathcal{N}(0, I_L)$, $\mathbb{E}[\mathbf{1}^\top s] = 0$, and we define $A := \mathbb{E}[G(s)] = \mathbb{E}\sum_i a_i(s)s_i$. Hence

$$g \cdot w^\star = -\beta\,\mu\,\frac{1-\bar{P}}{L(L-1)}\,LA = -\beta\,\mu\,\frac{(1-\bar{P})\,A}{L-1}.$$

Stein identity for A. By Gaussian integration by parts, $A = \mathbb{E}\sum_i a_i(s)s_i = \mathbb{E}\sum_i \partial_{s_i} a_i(s) = \beta\,\mathbb{E}\sum_i a_i(s)(1-a_i(s)) = \beta(1-\bar{P})$, using the softmax Jacobian $\partial_{s_i} a_i = \beta a_i(1-a_i)$. Substituting,

$$g \cdot w^\star = -\beta\,\mu\,\frac{(1-\bar{P})}{L-1}\,\beta(1-\bar{P}) = -\beta^2\,\mu\,\frac{(1-\bar{P})^2}{L-1},$$

as claimed. The computation is checked numerically (`numerics/unlock_closedform.py`) to absolute error about 10^{-3} across $L \in \{2,\dots,5\}$, $\beta \in \{1,2,3\}$, $\mu \in \{0, \frac{1}{2}, 1\}$. □

Proof of Corollary 2. The value claim is Lemma 7: $\nabla_u \mathcal{L} \cdot u^\star$ has a degree-one (linear) signal of size $\Theta(1)$ at $\mu = 0$, so μ grows geometrically to $\Theta(1)$ regardless of the key. With $\mu = \Theta(1)$, Theorem 1 gives a key drift along $w^\star$ of size $|\nabla_w \mathcal{L} \cdot w^\star| = \beta^2\mu(1-\bar{P})^2/(L-1) = \Theta(1)$ at the equator. This drift is *additive*: it is bounded below by $\Theta(1)$ *independently of the current selection overlap* m, not proportional to a positive power of m. Contrast this with an information-exponent $k \geq 2$ single-index target, whose population drift scales as m^{k-1} and hence collapses to $\Theta(d^{-(k-1)/2})$ at the random-initialisation overlap $m = \Theta(d^{-1/2})$, forcing a $\mathrm{poly}(d)$ escape from the equator. Here the catalyst supplies a constant push from the very first step, so the selection overlap leaves the $\Theta(d^{-1/2})$ equator and reaches $\Theta(1)$ in a single $O(1)$ increment, after which the exact drift $\eta\,\mu\,C'(m)(1-m^2) \geq \eta\,\mu\,c_1(1-m^2) > 0$ of Theorem 5 (the global, all-m form of Theorem 1) contracts $1-m$ geometrically to zero. Both phases therefore take $O(1)$ population steps, *dimension-free*. Finite-sample, each population expectation is estimated to accuracy $\Theta(\sqrt{d/n})$ in the relevant directions (sub-exponential label, Lemma 7 and (1)), so $n = \Theta(d)$ suffices to keep the empirical drift aligned with the $\Theta(1)$ population drift. This is precisely why the infinite information exponent costs no extra power of d: the obstruction that an IE $= k$ target would impose—a drift vanishing as $m \to 0$—is absent, replaced by an additive catalyst drift. The end-to-end cascade is exhibited in `numerics/gd_trajectory.py`. □

Proof of Corollary 3. By Theorem 1 (and more generally by the slice potential, Lemma 9) the population selection drift is $\Delta m = \eta\,\mu\,C'(m)(1-m^2)$. While $\mu = O(d^{-1/2})$, this drift is $O(d^{-1/2})$—the same order as the finite-sample estimation noise $O(\sqrt{d/n}) = O(1)$ projected onto the $w^\star$

direction (which has magnitude $O(d^{-1/2})$ in the full d-dimensional space). Hence the empirical selection gradient is indistinguishable from noise and no net escape occurs: m_t performs a random walk of step $O(d^{-1/2})$, so over t steps it stays at the $O(d^{-1/2}\sqrt{t})$ scale—still $O(d^{-1/2})$ throughout the $O(1)$-step window during which $\mu_t = O(d^{-1/2})$. Meanwhile the value drift $\Delta\mu \geq \eta/L = \Theta(1)$ is always dimension-free (Theorem 5, part 1), so the value overlap escapes to $\Theta(1)$ in $O(1)$ steps regardless of the selection overlap. Only then does the selection drift become $\Theta(1)$ and resolvable above the noise; the value-acquisition window therefore closes before the selection overlap can leave the noise floor. □

Proof of Lemma 3. Condition on s. Then $\mathbb{E}[(y+\xi)^2 \mid s] = \mathbb{E}[y^2 \mid s] + \sigma^2$ and, using independence and $\mathbb{E}[v_i v_j] = \delta_{ij}$, $\mathbb{E}[y^2 \mid s] = \sum_{i,j} a_i(s) a_j(s)\, \mathbb{E}[v_i v_j] = \sum_i a_i(s)^2 = P(s)$. □

Proof of Lemma 4. Shift invariance is immediate: $a(s + c\mathbf{1}) = a(s)$ because the common factor $e^{\beta c}$ cancels in the normalisation, hence $P(s + c\mathbf{1}) = P(s)$. For the Hermite statement, recall the Gaussian integration-by-parts (Stein) identity $\mathbb{E}[P(s)\, s_i] = \mathbb{E}[\partial_{s_i} P(s)]$. Summing over i and using shift invariance, $\sum_i \partial_{s_i} P(s) = \frac{d}{dc} P(s + c\mathbf{1})\big|_{c=0} = 0$ pointwise. By exchangeability of the coordinates of s (the law of s and the function P are both symmetric under permuting tokens), all the $\mathbb{E}[\partial_{s_i} P(s)]$ are equal; since they sum to zero, each is zero. The general direction θ follows by linearity. □

Proof of Lemma 5. Covariance identity. Since P is shift invariant it is a function of $s_\perp$ alone, and $s_\perp$ is independent of $\bar{s}$ (Gaussian orthogonal decomposition). By exchangeability $\mathbb{E}[P\ \mathrm{He}_2(s_i)]$ is the same for all i; averaging over $i \in [L]$ and using $\sum_i \mathrm{He}_2(s_i) = \sum_i s_i^2 - L = \|s_\perp\|^2 + L\bar{s}^2 - L$ together with independence of P from $\bar{s}$,

$$\tfrac{1}{2}\mathbb{E}[P\ \mathrm{He}_2(s_1)] = \tfrac{1}{2L}\,\mathbb{E}\big[P\,(\|s_\perp\|^2 - (L-1))\big] = \tfrac{1}{2L}\,\mathrm{Cov}\big(P, \|s_\perp\|^2\big),$$

because $\mathbb{E}\|s_\perp\|^2 = L - 1$. This proves the identity.

Strict positivity. It suffices to show P, as a function of $s_\perp$, is radially nondecreasing along every ray and nonconstant; then $\mathrm{Cov}(P, \|s_\perp\|^2) > 0$ by the radial association (Chebyshev) inequality, Lemma 15. Fix a direction $\omega \in \mathbf{1}^\perp$, $\|\omega\| = 1$, and set $p(t) = \mathrm{softmax}(\beta t\omega)$, $P(t) = \|p(t)\|^2$. Differentiating,

$$P'(t) = 2\beta\Big(\langle p^{\odot 2}, \omega\rangle - \|p\|^2\langle p, \omega\rangle\Big) = 2\beta\,\|p\|^2\big(\langle q, \omega\rangle - \langle p, \omega\rangle\big), \qquad q_i := \frac{p_i^2}{\sum_j p_j^2}.$$

Both p and q are probability vectors, and q is obtained from p by the size-biasing $q_i \propto p_i \cdot p_i$; since p_i is a strictly increasing function of ω_i (softmax is coordinatewise monotone), the sequences (p_i) and (ω_i) are comonotone. Writing $\mathbb{E}_p[\,\cdot\,] = \sum_i p_i(\cdot)$ for expectation under the distribution p,

$$\langle q, \omega\rangle - \langle p, \omega\rangle = \frac{\mathbb{E}_p[p_i\omega_i]}{\mathbb{E}_p[p_i]} - \mathbb{E}_p[\omega_i] = \frac{\mathbb{E}_p[p_i\omega_i] - \mathbb{E}_p[p_i]\,\mathbb{E}_p[\omega_i]}{\mathbb{E}_p[p_i]} = \frac{\mathrm{Cov}_p(p_i, \omega_i)}{\mathbb{E}_p[p_i]} \;\geq\; 0,$$

the inequality being the Chebyshev sum (association) inequality for the comonotone pair (p_i, ω_i) under the weights p (Lemma 14), with equality iff ω is constant on $[L]$, i.e. $\omega = 0$ in $\mathbf{1}^\perp$. Hence $P'(t) \geq 0$ for all $t \geq 0$, with strict increase on a set of positive measure for every $\omega \neq 0$. Thus P is radially nondecreasing and nonconstant, and Lemma 15 gives $\mathrm{Cov}(P, \|s_\perp\|^2) > 0$. The closed form and peak location for $L = 2$ are recorded in Remark 4 and Appendix H; numerics in `numerics/c2_closedform.py` agree at the percent level or better. □

Proof of Lemma 6. Small β. Expand $p = \mathrm{softmax}(\beta s)$ around the uniform vector. With $\bar{s} = \frac{1}{L}\sum_j s_j$ and $s_{\perp,i} = s_i - \bar{s}$,

$$p_i = \frac{1}{L}\Big(1 + \beta s_{\perp,i} + \tfrac{\beta^2}{2}\big(s_{\perp,i}^2 - \tfrac{1}{L}\textstyle\sum_j s_{\perp,j}^2\big) + O(\beta^3)\Big),$$

so $P = \sum_i p_i^2 = \frac{1}{L}\big(1 + \beta^2\,\frac{1}{L}\sum_i s_{\perp,i}^2\big) + O(\beta^3) = \frac{1}{L} + \frac{\beta^2}{L^2}\|s_\perp\|^2 + O(\beta^3)$. Substituting into the covariance identity $c_2 = \frac{1}{2L}\mathrm{Cov}(P, \|s_\perp\|^2)$ of Lemma 5 and using $\mathrm{Var}(\|s_\perp\|^2) = 2(L-1)$ for the χ^2_{L-1} variable $\|s_\perp\|^2$,

$$c_2 = \frac{1}{2L}\cdot\frac{\beta^2}{L^2}\,\mathrm{Var}(\|s_\perp\|^2) + O(\beta^4) = \frac{\beta^2}{2L^3}\cdot 2(L-1) + O(\beta^4) = \frac{L-1}{L^3}\,\beta^2 + O(\beta^4).$$

(The $O(\beta^3)$ term in P contributes only at order β^4 to the covariance, being an odd polynomial in $s_\perp$ uncorrelated with the even $\|s_\perp\|^2$ at leading order.)

Large β. As $\beta \to \infty$, $p = \mathrm{softmax}(\beta s)$ converges to the indicator of $\arg\max_i s_i$ for almost every s, so $P(s) \to 1$ pointwise a.s.; since $0 < P \le 1$, dominated convergence gives $\mathrm{Cov}(P, \|s_\perp\|^2) \to \mathrm{Cov}(1, \|s_\perp\|^2) = 0$, hence $c_2 \to 0$. Positivity on compacta is Lemma 5 together with continuity of $\beta \mapsto c_2(\beta, L)$. □

Proof of Theorem 2. $\mathrm{IE}_{w^\star}(y) = \infty$ is Corollary 1. For the generative exponent, $T(y) = y^2$ has $\mathbb{E}[T(y) \mid s] = P(s) + \sigma^2$ (Lemma 3); its degree-1 Hermite projection along $w^\star$ vanishes (Lemma 4) and its degree-2 projection is $c_2(\beta, L) > 0$ (Lemma 5). Hence $\mathrm{IE}_{w^\star}(T \circ y) = 2$, so $\mathrm{GE}_{w^\star}(y) \le 2$. It cannot be 1: $\mathrm{GE} = 1$ would require some T with $\mathbb{E}[T(y)\, s_i] \neq 0$, i.e. $\mathbb{E}\big[\mathbb{E}[T(y) \mid s]\, s_i\big] \neq 0$. But $h_T(s) := \mathbb{E}[T(y) \mid s]$ is shift invariant for *every* T, because the conditional law of y given s depends on s only through $a(s) = a(s + c\mathbf{1})$; the argument of Lemma 4 then forces $\mathbb{E}[h_T(s) s_i] = 0$. Thus $\mathrm{GE}_{w^\star}(y) = 2$. □

Proof of Theorem 3. We prove the first two assertions in turn: (I) a truncated spectral estimator achieves $\tilde{O}(d)$ sample complexity; (II) no correlational SQ algorithm succeeds at any n. The matching information-theoretic lower bound $n = \Omega(d)$ is Theorem 7 below, proved by Fano's method.

(I) Consistency of the spectral estimator. Write the observed (possibly noisy) label $\tilde{y} = y + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$ independent of X. The estimator is: (1) form $\widehat{b} = \frac{1}{n}\sum_t \sum_i \tilde{y}^{(t)} x_i^{(t)}$ and set $\widehat{u} = \widehat{b}/\|\widehat{b}\|$; (2) form the truncated version of $\widehat{M} = \frac{1}{n}\sum_t \sum_i \tilde{y}^{(t)2}(x_i^{(t)} x_i^{(t)\top} - I_d)$, project out $\widehat{u}$ via $\Pi = I - \widehat{u}\widehat{u}^\top$, and output the top eigenvector $\widehat{w}$ of $\Pi\widehat{M}\Pi$. The population mean is unchanged by the noise: $\mathbb{E}[\tilde{y}\, x_i] = \mathbb{E}[y\, x_i]$ (as ξ is centred and independent) and $\mathbb{E}[\tilde{y}^2(x_i x_i^\top - I_d)] = \mathbb{E}[y^2(x_i x_i^\top - I_d)] + \sigma^2\, \mathbb{E}[x_i x_i^\top - I_d] = \mathbb{E}[y^2(x_i x_i^\top - I_d)]$, since the additive σ^2 multiplies the mean-zero matrix $\mathbb{E}[x_i x_i^\top - I_d] = 0$. Thus the untruncated population mean is M (Lemma 8) with the *same* rank-two spike; noise affects only the fluctuations. Truncation introduces a bias that vanishes at the logarithmic thresholds used below.

Boundedness of the label. Because $a(s) \in \Delta^{L-1}$ is a probability vector, $y = \sum_i a_i(s) v_i$ is a convex combination of $v_1, \dots, v_L$, so

$$|y| \;\le\; \max_{i \in [L]} |v_i|, \qquad \text{hence} \qquad y^2 \;\le\; \max_i v_i^2. \tag{1}$$

With $v \sim \mathcal{N}(0, I_L)$ and L fixed, $\max_i v_i^2$ is sub-exponential with $\|\max_i v_i^2\|_{\psi_1} \le C_L = O(\log L)$, so y^2 is sub-exponential, uniformly in β and d. The noisy square adds $\tilde{y}^2 = y^2 + 2y\xi + \xi^2$, where ξ^2 is sub-exponential with $\|\xi^2\|_{\psi_1} = \Theta(\sigma^2)$ and the cross term $2y\xi$ has $\|2y\xi\|_{\psi_1} = O(\sigma)$; hence $\|\tilde{y}^2\|_{\psi_1} \le C_L(1 + \sigma^2) =: \kappa_\sigma$, and the value-moment summand $\tilde{y} \sum_i x_i$ has per-direction sub-exponential norm $O(\sqrt{d}\,(1 + \sigma))$.

Value-direction concentration. By Lemma 7, $\mathbb{E}\widehat{b} = u^\star$ (a unit vector). The summands $\tilde{y}^{(t)} \sum_i x_i^{(t)}$ are i.i.d. with sub-exponential norm $O(\sqrt{d}\,(1 + \sigma))$ in each fixed direction; a standard vector Bernstein bound (Tropp, 2015, Thm. 6.1.1) gives, for $n \ge Cd \log d$,

$$\|\widehat{b} - u^\star\|_2 \;\le\; C(1 + \sigma)\sqrt{\tfrac{d}{n}} \quad \text{w.p. } 1 - d^{-2},$$

so $\|\widehat{u} - u^\star\|_2 \le C(1 + \sigma)\sqrt{d/n} =: \eta$.

Matrix concentration. Truncate $\tilde{y}^2$ at $C\kappa_\sigma \log n$ and each token norm squared at $Cd \log n$. On the truncated event the summands are bounded self-adjoint matrices; their variance proxy is $O(\kappa_\sigma^2 d\, \mathrm{polylog}\, n)$ because $(xx^\top)^2 = \|x\|^2 xx^\top$. Matrix Bernstein (Tropp, 2015, Thm. 6.1.1), followed by the Gaussian and sub-exponential tail bounds for the discarded event, gives constants $C, q < \infty$ such that

$$\big\|\widehat{M} - M\big\|_{\mathrm{op}} \;\le\; C\kappa_\sigma \left(\sqrt{\frac{d \log^q d}{n}} + \frac{d \log^q d}{n}\right) + o(1) \;=:\; \varepsilon_n \tag{2}$$

with high probability. The exponent q is immaterial for the dimensional claim; we do not optimize logarithmic factors.

Deflation and Davis–Kahan. By Lemma 8, $M = 2Lc_2\, w^\star w^{\star\top} + 2\bar{P}\, u^\star u^{\star\top}$ with $w^\star \perp u^\star$. Replacing the exact projector $I - u^\star u^{\star\top}$ by $\Pi = I - \widehat{u}\widehat{u}^\top$ perturbs the matrix by at most $2\|M\|_{\mathrm{op}}\,\eta \le C\eta$ in operator norm (since $\|\widehat{u} - u^\star\| \le \eta$ and $\|M\|_{\mathrm{op}} = O(1)$). Hence

$$\Pi\widehat{M}\Pi \;=\; 2Lc_2\, w^\star w^{\star\top} \;+\; E, \qquad \|E\|_{\mathrm{op}} \;\le\; \varepsilon_n + C\eta.$$

Once $\varepsilon_n + C\eta < Lc_2/2$, the deflated matrix has a unique leading eigenvalue. The Davis–Kahan $\sin\Theta$ theorem (Yu et al., 2015, Cor. 3) gives

$$\sin\angle(\widehat{w}, w^\star) \;\le\; O\!\left(\frac{\varepsilon_n + \eta}{Lc_2}\right).$$

For any fixed target accuracy, $n = C(\beta, L, \sigma, \varepsilon)\, d\,\operatorname{polylog} d$ suffices. If $n/(d\,\operatorname{polylog} d) \to \infty$, then $\varepsilon_n + \eta \to 0$ and the overlap tends to one.

(II) Correlational impossibility. A correlational statistical query against the key scores is any $\phi \mapsto \mathbb{E}[y\,\phi(s)]$. By Lemma 1 every such query equals $\mathbb{E}[\phi(s)\,\mathbb{E}[y \mid s]] = 0$, independent of n. No algorithm restricted to such queries can distinguish $w^\star$ from a random direction, so it fails with any number of samples. This is the formal $\mathrm{IE} = \infty$ obstruction.

(III) Matching lower bound. Assertions (I)–(II) are complemented by the information-theoretic lower bound of Theorem 7, which shows that *no* estimator (statistical-query or otherwise) can recover $w^\star$ from fewer than $\Omega(\sigma^2 d/(\beta^2 L))$ samples. Together with (I), this places the spectral sample complexity between $\Omega(d)$ and $\tilde{O}(d)$ for fixed (β, L, σ). The bounds also agree in the *direction* of the σ-dependence—both degrade with noise—though not in the exact power of σ: the upper constant carries $(1+\sigma^2)^2$ while the Fano lower bound is linear in σ^2, a gap we do not close here and flag explicitly in Remark 5. □

Proof of Lemma 7. $\mathbb{E}[yv_i] = \sum_j \mathbb{E}[a_j(s)v_jv_i] = \mathbb{E}[a_i(s)]$ by independence and $\mathbb{E}[v_jv_i] = \delta_{ij}$. Projecting, $\mathbb{E}[yx_i] = \mathbb{E}[a_i(s)]u^\star$ (the $w^\star$ component vanishes by Lemma 1), so summing over i gives $b = u^\star$. □

B THE SQUARED-LABEL MATRIX IS EXACTLY RANK TWO

The dichotomy of Theorem 2 is a statement about Hermite projections; to turn it into an algorithm we need the full population spectrum of the matrix the estimator actually forms. It is exactly rank two—one spike per direction of the head—with no bulk and no cross term. This is the structural fact that makes a single eigendecomposition rate optimal, and it also determines the spike constant exactly (correcting the heuristic $\Theta(1)$ count: the selection spike carries an extra factor L).

Lemma 8 (Exact rank-two population matrix)**.** *Let*

$$M \;:=\; \sum_{i=1}^{L} \mathbb{E}\big[\, y^2\,(x_i x_i^\top - I_d)\,\big] \;\in\; \mathbb{R}^{d\times d}.$$

Then, with $c_2 = c_2(\beta, L)$ of Lemma 5 and $\bar{P} := \mathbb{E}[P(s)] = \mathbb{E}\sum_i a_i(s)^2 \in (1/L, 1)$ for every finite $\beta > 0$,

$$\boxed{\;M \;=\; 2L\,c_2\; w^\star w^{\star\top} \;+\; 2\,\bar{P}\; u^\star u^{\star\top}.\;}$$

In particular M vanishes on $(\operatorname{span}\{w^\star, u^\star\})^\perp$, the two spikes are orthogonal, the selection spike has strength $2L\,c_2 > 0$ and the value spike has strength $2\bar{P} > 0$, and there is no $w^\star u^{\star\top}$ cross term.

Proof. Fix a unit vector $\theta \in \mathbb{S}^{d-1}$ and write $z_i = \langle\theta, x_i\rangle$. By definition $\theta^\top M\theta = \sum_i \mathbb{E}[y^2(z_i^2 - 1)] = \sum_i \mathbb{E}[y^2\,\mathrm{He}_2(z_i)]$. Decompose $\theta = \alpha w^\star + \gamma u^\star + \delta r$ with $r \perp w^\star, u^\star$, $\alpha^2 + \gamma^2 + \delta^2 = 1$. Because X is Gaussian and $w^\star, u^\star, r$ are orthonormal, the three score families $s_i = \langle w^\star, x_i\rangle$, $v_i = \langle u^\star, x_i\rangle$ and $\rho_i = \langle r, x_i\rangle$ are jointly Gaussian and *mutually independent*, and $z_i = \alpha s_i + \gamma v_i + \delta\rho_i$. The label $y = \sum_j a_j(s)v_j$ depends only on (s, v) and not on ρ.

Step 1: condition on (s, v). Then y is determined and, since $\rho_i \sim \mathcal{N}(0,1)$ is independent of (s, v) and of z_i's other parts,

$$\mathbb{E}[y^2 \operatorname{He}_2(z_i)] = \mathbb{E}\Big[y^2 \,\mathbb{E}\big[\operatorname{He}_2(\alpha s_i + \gamma v_i + \delta\rho_i) \,\big|\, s, v\big]\Big].$$

Using $\mathbb{E}\rho_i = 0$, $\mathbb{E}\rho_i^2 = 1$, and $\alpha^2 + \gamma^2 + \delta^2 = 1$,

$$\mathbb{E}[\operatorname{He}_2(\alpha s_i + \gamma v_i + \delta\rho_i) \mid s, v] = \alpha^2 \operatorname{He}_2(s_i) + \gamma^2 \operatorname{He}_2(v_i) + 2\alpha\gamma s_i v_i.$$

Thus the independent r-component contributes no degree-two term, and

$$\theta^\top M \theta = \alpha^2 \sum_i \mathbb{E}[y^2 \operatorname{He}_2(s_i)] + \gamma^2 \sum_i \mathbb{E}[y^2 \operatorname{He}_2(v_i)] + 2\alpha\gamma \sum_i \mathbb{E}[y^2 s_i v_i].$$

We evaluate the three sums.

(a) Selection sum. Conditioning on s and using $\mathbb{E}[y^2 \mid s] = P(s)$ (Lemma 3, $\sigma = 0$) together with $\mathbb{E}[\operatorname{He}_2(s_i) \mid s] = \operatorname{He}_2(s_i)$, $\sum_i \mathbb{E}[y^2 \operatorname{He}_2(s_i)] = \sum_i \mathbb{E}[P(s) \operatorname{He}_2(s_i)] = L \cdot \mathbb{E}[P(s) \operatorname{He}_2(s_1)] = 2L\, c_2$ by exchangeability and the definition of c_2.

(b) Value sum. Here $\operatorname{He}_2(v_i) = v_i^2 - 1$ and $y^2 = \sum_{j,k} a_j a_k v_j v_k$. Conditioning on s (so the a_j are fixed) and taking the Gaussian expectation over $v \sim \mathcal{N}(0, I_L)$, $\mathbb{E}[v_j v_k (v_i^2 - 1)] = 2\delta_{ji}\delta_{ki}$ (the only nonzero Gaussian fourth moment contribution, since $\mathbb{E}[v_i^4] - \mathbb{E}[v_i^2] = 3 - 1 = 2$ and mixed terms with an odd power vanish). Hence $\sum_i \mathbb{E}[y^2 \operatorname{He}_2(v_i)] = \sum_i \mathbb{E}\big[2a_i(s)^2\big] = 2\bar{P}$.

(c) Cross sum. $\sum_i \mathbb{E}[y^2 s_i v_i]$: condition on s; then $\mathbb{E}[y^2 v_i \mid s] = \sum_{j,k} a_j a_k \mathbb{E}[v_j v_k v_i] = 0$ because the third Gaussian moment of the mean-zero vector v vanishes. So the cross sum is zero identically.

Collecting, $\theta^\top M \theta = 2L\, c_2\, \alpha^2 + 2\bar{P}\, \gamma^2 = \theta^\top \big(2Lc_2\, w^\star w^{\star\top} + 2\bar{P}\, u^\star u^{\star\top}\big)\theta$ for every θ. Since both sides are quadratic forms of symmetric matrices, equality of the forms for all θ forces equality of the matrices. □

Remark 7 (The selection spike carries a hidden factor L)**.** *The naive reading of Lemma 5—"the spike along $w^\star$ has strength $2c_2$"—undercounts by exactly the context length: summing the single-coordinate contribution over the L tokens gives $2Lc_2$. This does not change any exponent (only the constant $C(\beta, L, \sigma)$ in Theorem 3), but it is the correct constant and agrees with* `numerics/verify_rank2.py` *within Monte-Carlo error across $L \in \{2, 3, 4\}$.*

C Discontinuous phase transition at orthogonality

The infinite information exponent is an exact consequence of value–key orthogonality. Even an infinitesimal departure from orthogonality collapses it to one. The boundary is as sharp as an information exponent can be.

Definition 3 (Tilted single-location head)**.** *Let $w^\star \in \mathbb{S}^{d-1}$ and $u^\star = \rho\, w^\star + \sqrt{1-\rho^2}\, e$ with $e \perp w^\star$, $\|e\| = 1$, and $\rho \in [-1, 1]$. Define key scores $s_i = \langle w^\star, x_i\rangle$ and value scores $v_i = \langle u^\star, x_i\rangle = \rho\, s_i + \sqrt{1-\rho^2}\, z_i$ where $z_i = \langle e, x_i\rangle \perp\!\!\!\perp s$, and the output $y = \sum_i a_i(s)\, v_i$.*

Theorem 4 (Discontinuous IE phase transition)**.** *For the tilted head (Definition 3):*

$$\mathrm{IE}_{w^\star}(y) \;=\; \begin{cases} \infty & \text{if } \rho = 0, \\ 1 & \text{if } \rho \neq 0. \end{cases}$$

When $\rho \neq 0$ the degree-one Hermite coefficient is exactly

$$\mathbb{E}\big[y\, \operatorname{He}_1(s_1)\big] \;=\; \frac{\rho}{L}, \tag{3}$$

independent of β. The information exponent therefore jumps discontinuously *from ∞ to 1 as the key–value alignment crosses zero.*

Proof. The case $\rho = 0$ is Corollary 1. Suppose $\rho \neq 0$. By independence of $z \perp\!\!\!\perp s$,

$$\mathbb{E}[y\, s_1] \;=\; \rho\, \mathbb{E}\Big[\Big(\textstyle\sum_i a_i(s)\, s_i\Big)\, s_1\Big] \;+\; \sqrt{1-\rho^2}\, \underbrace{\mathbb{E}\Big[\Big(\sum_i a_i(s)\, z_i\Big)\, s_1\Big]}_{=0\ (z \perp s)}.$$

We claim $\mathbb{E}[G\, s_1] = 1/L$ where $G := \sum_i a_i(s)\, s_i$. Decompose G as

$$G \;=\; \underbrace{\sum_i a_i(s)\,(s_i - \bar{s})}_{=:\,H} \;+\; \bar{s}, \qquad \bar{s} = \tfrac{1}{L}\textstyle\sum_j s_j,$$

using $\sum_i a_i = 1$. Because $a(s) = a(s + c\mathbf{1})$, the centred sum H is a function of the contrasts $s_\perp = s - \bar{s}\mathbf{1}$ alone. By the Gaussian orthogonal decomposition, $\bar{s} \sim \mathcal{N}(0, 1/L)$ is independent of $s_\perp$. Write $s_1 = \bar{s} + (s_1 - \bar{s})$; then

$$\begin{aligned}\mathbb{E}[G\, s_1] &= \mathbb{E}[(H + \bar{s})\,(\bar{s} + (s_1 - \bar{s}))] \\ &= \mathbb{E}[H\,(s_1 - \bar{s})] + \mathbb{E}[H\,\bar{s}] + \mathbb{E}[\bar{s}\,(s_1 - \bar{s})] + \mathbb{E}[\bar{s}^2].\end{aligned}$$

Term 2: H is a function of $s_\perp$; $\bar{s}$ is independent of $s_\perp$; so $\mathbb{E}[H\,\bar{s}] = \mathbb{E}[H]\,\mathbb{E}[\bar{s}] = 0$. *Term 3*: $\bar{s}$ and $(s_1 - \bar{s})$ are independent (Gaussian projection), so $\mathbb{E}[\bar{s}(s_1 - \bar{s})] = 0$. *Term 4*: $\mathbb{E}[\bar{s}^2] = 1/L$. *Term 1*: H is a function of the contrast vector $s_\perp \in \mathbf{1}^\perp$, and $(s_1 - \bar{s})$ is a specific coordinate of $s_\perp$. By the exchangeability of the token labels (the law of s and the function $H(s) = \sum_i a_i(s)(s_i - \bar{s})$ are both symmetric under permutation of $[L]$), $\mathbb{E}[H\,(s_j - \bar{s})]$ is the same for every $j \in [L]$. Since $\sum_{j=1}^{L}(s_j - \bar{s}) = 0$ identically,

$$\sum_{j=1}^{L} \mathbb{E}[H\,(s_j - \bar{s})] \;=\; \mathbb{E}\Big[H \sum_j (s_j - \bar{s})\Big] = 0,$$

whence each $\mathbb{E}[H\,(s_j - \bar{s})] = 0$.

Collecting the four terms: $\mathbb{E}[G\, s_1] = 0 + 0 + 0 + 1/L = 1/L$, and therefore $\mathbb{E}[y\, s_1] = \rho/L \neq 0$, giving $\mathrm{IE}_{w^\star}(y) = 1$. □

Remark 8 (Independence from β)**.** *The formula $\mathbb{E}[y\, s_1] = \rho/L$ does not depend on the inverse temperature β: the interplay between the sharpness of selection and the strength of the key–value correlation cancels exactly in the degree-one projection. This is not an artifact of cancellation but a structural consequence of the decomposition $G = H + \bar{s}$: the only term contributing to the first Hermite moment of G is $\bar{s}$, which is entirely geometric (the mean of L Gaussians) and independent of the softmax weighting. Degree-two and higher projections* do *depend on β—see* `numerics/phase_transition.py`*—but the leading (degree-one) signal does not.*

Remark 9 (Architectural implication)**.** *The discontinuity has a concrete architectural message: any attention head with* shared *key–value subspaces ($\rho \neq 0$, as in a head where W_K and W_V have overlapping column spans) is immediately visible to a single gradient step at the $n = \Theta(d)$ rate. The hardness—the infinite information exponent, the saddle, the need for a catalyst—exists* only *in the subspace-separated case $\rho = 0$. In practice this means the selection directions embedded in the kernel of W_V are the hard degrees of freedom, and they are the ones whose learning dynamics exhibits the value→selection cascade of Theorem 1.*

D Formal convergence of the gradient descent cascade

Corollary 2 states that plain gradient descent learns both directions in two stages. We now give the precise iteration-and-sample complexity for the population and finite-sample regimes. The closed-form catalyst gradient of Theorem 1 gives the drift only at the equator $m = \langle w, w^\star\rangle = 0$; to establish convergence across the slice we must control the selection drift at *every* overlap $m \in [0, 1)$. We do this not by perturbing around the equator but by exhibiting the exact two-dimensional potential that the dynamics descends, and reading the drift off its derivative.

D.1 An exact slice potential and the Mehler series $C(m)$

The two-parameter loss collapses, on the relevant invariant slice, to a function of the two order parameters alone. Let $w = m\, w^\star + \sqrt{1 - m^2}\, e$ and $u = \mu\, u^\star + \sqrt{1 - \mu^2}\, f$ with e, f unit vectors orthogonal to $\{w^\star, u^\star\}$ and to each other (the configuration reached from a random initialisation lies on this slice up to $O(d^{-1/2})$ corrections, and the slice is GD-invariant; see Lemma 11).

Lemma 9 (Exact slice potential)**.** *On the slice above, the population loss is*

$$\mathcal{L}(m,\mu) \;=\; \bar{P} \;-\; \mu\, C(m), \qquad C(m) \;:=\; \mathbb{E}\Big[\textstyle\sum_{i=1}^{L} a_i(s)\, a_i(t_m)\Big], \tag{4}$$

where $t_m = m\,s + \sqrt{1-m^2}\,\eta$ with $\eta \sim \mathcal{N}(0, I_L)$ independent of s, and $\bar{P} = \mathbb{E}\sum_i a_i(s)^2$ is the collision probability. Equivalently $C(m) = L \cdot \mathbb{E}[a_1(s)\, a_1(t_m)]$ by exchangeability.

Proof. Expand $\mathcal{L} = \frac{1}{2}\mathbb{E}[(y-\hat{y})^2] = \frac{1}{2}\mathbb{E}[y^2] - \mathbb{E}[y\hat{y}] + \frac{1}{2}\mathbb{E}[\hat{y}^2]$. With $y = \sum_i a_i(s) v_i$ and $\hat{y} = \sum_i a_i(t) b_i$, $b_i = \langle u, x_i\rangle$, condition on the key fields (s,t); the value scores $v \sim \mathcal{N}(0, I_L)$ are independent of (s,t) and $b_i = \mu v_i + \sqrt{1-\mu^2}\langle f, x_i\rangle$ with $\langle f, x_i\rangle$ independent of v and mean zero. Hence $\mathbb{E}[y^2] = \mathbb{E}\sum_i a_i(s)^2 = \bar{P}$, $\mathbb{E}[\hat{y}^2] = \mathbb{E}\sum_i a_i(t)^2\,(\mu^2 + (1-\mu^2)) = \bar{P}$ (the value norm is one), and the cross term is $\mathbb{E}[y\hat{y}] = \mu\,\mathbb{E}[\sum_{i,j} a_i(s) a_j(t)\, v_i v_j] = \mu\,\mathbb{E}\sum_i a_i(s) a_i(t) = \mu\, C(m)$, using $\mathbb{E}[v_i v_j] = \delta_{ij}$ and that only the $u^\star$-component of u (carrying the factor μ) survives against v. Collecting, $\mathcal{L} = \frac{1}{2}\bar{P} - \mu C(m) + \frac{1}{2}\bar{P} = \bar{P} - \mu C(m)$. Finally $t = m\,s + \sqrt{1-m^2}\,\eta$ in distribution since $\langle w, x_i\rangle$ has correlation m with $s_i = \langle w^\star, x_i\rangle$ and an independent Gaussian remainder; this is the stated t_m. □

Lemma 10 (C is a Mehler series with non-negative coefficients)**.** *Write $a_1(s) = \sum_{\alpha\in\mathbb{N}^L} \hat{a}_\alpha\, \mathrm{He}_\alpha(s)/\sqrt{\alpha!}$ for its tensor-Hermite expansion, $\mathrm{He}_\alpha(s) = \prod_i \mathrm{He}_{\alpha_i}(s_i)$, $\hat{a}_\alpha = \mathbb{E}[a_1(s)\,\mathrm{He}_\alpha(s)]/\sqrt{\alpha!}$. Then by Mehler's formula, with correlation m between s and t_m coordinatewise,*

$$C(m) \;=\; L \sum_{\alpha\in\mathbb{N}^L} \hat{a}_\alpha^2\, m^{|\alpha|} \;=\; \sum_{k\ge 0} c_k\, m^k, \qquad c_k \;=\; L \sum_{|\alpha|=k} \hat{a}_\alpha^2 \;\ge\; 0. \tag{5}$$

Consequently C is real-analytic, strictly increasing and convex on $[0,1]$ (every coefficient c_k with $k \ge 1$ is non-negative, and infinitely many are strictly positive). Its endpoints and first two coefficients are

$$c_0 = C(0) = \tfrac{1}{L}, \quad C(1) = \bar{P}, \quad c_1 = C'(0) = \frac{\beta^2(1-\bar{P})^2}{L-1}, \quad c_2 \ge 0.$$

Proof. Mehler's formula states that for jointly Gaussian (s, t_m) with $\mathbb{E}[s_i t_{m,j}] = m\,\delta_{ij}$ and unit marginals, $\mathbb{E}[\mathrm{He}_\alpha(s)\,\mathrm{He}_\gamma(t_m)] = \delta_{\alpha\gamma}\,\alpha!\, m^{|\alpha|}$. Substituting the Hermite expansions of $a_1(s)$ and $a_1(t_m)$ and using exchangeability ($C(m) = L\,\mathbb{E}[a_1(s)a_1(t_m)]$) gives (5); each term is $L\,\hat{a}_\alpha^2\, m^{|\alpha|}$ with $\hat{a}_\alpha^2 \ge 0$, so $c_k \ge 0$. Since a_1 is non-constant (it depends on s), some $\hat{a}_\alpha$ with $|\alpha| \ge 1$ is nonzero, so C is strictly increasing; non-negativity of all c_k makes C' non-decreasing, i.e. C convex. For c_0, read (4) at $m = 0$, where s, t are independent: $C(0) = \sum_i \mathbb{E}[a_i(s)]\,\mathbb{E}[a_i(t)] = \sum_i (1/L)^2 = 1/L$. At $m = 1$, $t_1 = s$ so $C(1) = \mathbb{E}\sum_i a_i(s)^2 = \bar{P}$. The first derivative is the degree-one coefficient, $c_1 = L\sum_i (\mathbb{E}[a_1(s) s_i])^2$. By the Stein identity $\mathbb{E}[a_1 s_1] = \mathbb{E}[\partial_{s_1} a_1] = \beta\,\mathbb{E}[a_1(1-a_1)] = \beta(1-\bar{P})/L$ (using exchangeability for $\mathbb{E}[a_1(1-a_1)] = (1-\bar{P})/L$) and, for $i \ne 1$, $\mathbb{E}[a_1 s_i] = \mathbb{E}[\partial_{s_i} a_1] = -\beta\,\mathbb{E}[a_1 a_i] = -\beta(1-\bar{P})/(L(L-1))$. Hence

$$c_1 = L\Big[\big(\tfrac{\beta(1-\bar{P})}{L}\big)^2 + (L-1)\big(\tfrac{\beta(1-\bar{P})}{L(L-1)}\big)^2\Big] = \frac{\beta^2(1-\bar{P})^2}{L}\Big[1 + \tfrac{1}{L-1}\Big] = \frac{\beta^2(1-\bar{P})^2}{L-1},$$

matching the equator catalyst gradient of Theorem 1. $c_2 \ge 0$ by (5). □

The slice potential gives both drift components directly: $\partial_\mu \mathcal{L} = -C(m) \le -1/L < 0$ *everywhere*, so the value overlap always increases; and $\partial_m \mathcal{L} = -\mu\, C'(m)$, which by Lemma 10 is $\le -\mu\, c_1 < 0$ *for every* $m \in [0,1)$ once $\mu > 0$. There is no interior critical point: the only stationary point of the slice potential on $[0,1)^2 \cup \{(1,1)\}$ closure is the optimum $(m,\mu) = (1,1)$. Thus the equator calculation of Theorem 1 is the $m = 0$ endpoint of the global drift. The numerics (`numerics/loss_decomposition.py`, `numerics/mehler_and_invariance.py`, `numerics/selection_gradient_vs_m.py`) check (4)–(5), positivity of the drift, and the nonnegative Hermite-sum coefficients.

Lemma 11 (The orthogonal slice is GD-invariant)**.** *Let $q = \langle w, u\rangle$, $r = \langle w, u^\star\rangle$, $p = \langle w^\star, u\rangle$ be the three off-slice overlaps. If $(q,r,p) = 0$ then the population gradients satisfy $\langle \nabla_w \mathcal{L}, u\rangle = \langle \nabla_w \mathcal{L}, u^\star\rangle = \langle \nabla_u \mathcal{L}, w^\star\rangle = 0$, so projected gradient descent that starts on the slice $\{q = r = p = 0\}$ remains on it for all t.*

Proof. Each off-slice derivative is, after the conditioning of Lemma 9, an odd Gaussian moment in an independent mean-zero coordinate (the components of x_i along f, resp. along $u^\star$ inside the key gradient, resp. along e inside the value gradient), hence zero. Concretely $\langle \nabla_w \mathcal{L}, u^\star \rangle$ picks out the $u^\star$-direction of the key gradient $\beta \sum_i b_i a_i (x_i - \bar{x})$ paired against the residual; the $u^\star$-component of x_i is v_i, which enters linearly and is mean-zero and independent of the key fields and of b's f-part, so the expectation vanishes; the other two are symmetric. Therefore the gradient flow has no component off the slice, and the slice is invariant. The numerics (`numerics/mehler_and_invariance.py`) confirm the off-slice overlaps stay $\leq 5 \times 10^{-3}$ along the full trajectory. □

Theorem 5 (Population convergence of the two-stage cascade)**.** *Consider projected gradient descent on $\mathcal{L}(w, u)$ along the invariant slice (Lemma 11) with step size $\eta \in (0, \eta_0(\beta, L))$, starting from a unit initialization with value overlap $\mu_0 > 0$ and any selection overlap $m_0 \in [0, 1)$. Define $\lambda_v := \eta / L > 0$ and $\lambda_w := \eta\, c_1 = \eta\, \beta^2 (1 - \bar{P})^2 / (L - 1) > 0$, where $c_1 = C'(0)$ from Lemma 10. Then:*

1. ***(Value phase.)** The value overlap satisfies $1 - \mu_t \leq (1 - \mu_0)\,(1 - \lambda_v)^t$ for all $t \geq 0$. In particular, $\mu_T \geq 1 - \varepsilon_v$ once $T \geq \frac{1}{\lambda_v} \log \frac{1 - \mu_0}{\varepsilon_v}$.*
2. ***(Selection phase.)** For every $\mu \geq \mu_\dagger > 0$ and every $m \in [0, 1)$, the projected-GD update obeys the exact angular drift $m_{t+1} - m_t = \eta\, \mu\, C'(m_t)\,(1 - m_t^2) + O(\eta^2)$, which is* strictly positive *on $(0, 1)$ and bounded below by $\eta\, \mu_\dagger\, c_1\,(1 - m_t^2)$. Hence $m_t \uparrow 1$ with no interior stall, and $m_t \geq 1 - \varepsilon_w$ after at most $T' \leq \frac{1}{\eta\, \mu_\dagger\, c_1} \log \frac{1}{\varepsilon_w}$ additional steps.*

Both phases require $O(1)$ steps (depending on β, L, η but not on d). The total population iteration count is $T_{\mathrm{pop}} = O\big(\frac{L}{\eta} \log \frac{1}{\varepsilon_v} + \frac{L - 1}{\eta \beta^2 (1 - \bar{P})^2} \log \frac{1}{\varepsilon_w}\big)$, dimension-free.

Proof. By Lemma 11 the dynamics stays on the slice, so it is the gradient flow of the exact two-dimensional potential $\Phi(m, \mu) = \mathcal{L}(m, \mu) = \bar{P} - \mu\, C(m)$ of Lemma 9, descended in the spherical (tangent) geometry of the two unit vectors.

Part 1 (value). On the unit sphere for u, write $\mu = \cos\phi$. The tangential component of $-\nabla_u \Phi$ along the great circle toward $u^\star$ is $-\partial_\phi \Phi = -\,C(m)\, \partial_\phi(-\mu) = C(m) \sin\phi = C(m) \sqrt{1 - \mu^2}$, using $\partial_\mu \Phi = -C(m)$. Since $C(m) \geq C(0) = 1/L > 0$ for all m (Lemma 10), the value drift is bounded below independently of the selection overlap: $\mu_{t+1} \geq \mu_t + \eta\, C(m_t) \sqrt{1 - \mu_t^2}\,(1 + o(1)) \geq \mu_t + \lambda_v (1 - \mu_t)$ in the contraction regime ($\sqrt{1 - \mu^2} \geq 1 - \mu$ and $C \geq 1/L$), giving the stated geometric bound. The value phase therefore completes in $O(1)$ steps independently of the (small) selection overlap.

Part 2 (selection). On the unit sphere for w, write $m = \cos\theta$. The tangential component of $-\nabla_w \Phi$ toward $w^\star$ is $-\partial_\theta \Phi = -\,\partial_\theta\big(-\mu\, C(\cos\theta)\big) = -\mu\, C'(\cos\theta) \sin\theta = -\mu\, C'(m) \sqrt{1 - m^2}$ in the θ direction; since θ decreases when m increases, the overlap drift is $\dot{m} \propto \mu\, C'(m)(1 - m^2)$, i.e. the projected-GD step is

$$m_{t+1} - m_t \;=\; \eta\, \mu\, C'(m_t)\,(1 - m_t^2) \;+\; O(\eta^2),$$

matching the numerics of `numerics/angular_flow_check.py` to $\approx 1\%$. By Lemma 10, $C'(m) = \sum_{k \geq 1} k\, c_k\, m^{k-1} \geq c_1 > 0$ for all $m \in [0, 1)$ (all coefficients $c_k \geq 0$, and $c_1 > 0$ strictly). Hence the drift is *strictly positive on the entire open interval*—there is no interior critical point, and in particular no stall away from $m = 0$ or $m = 1$. Theorem 1 evaluates precisely $\mu\, c_1$, the $m \to 0$ limit of the drift, and Lemma 10 shows the drift never changes sign as m grows. Using the lower bound $C'(m) \geq c_1$ and $1 - m_t^2 \geq (1 - m_t)$,

$$1 - m_{t+1} \;\leq\; (1 - m_t)\big(1 - \eta\, \mu_\dagger\, c_1\big),$$

so m_t reaches $1 - \varepsilon_w$ in at most $\frac{1}{\eta \mu_\dagger c_1} \log \frac{1}{\varepsilon_w}$ steps. All constants ($\bar{P}, \beta, L, c_1$) are d-independent, giving dimension-free convergence. □

Theorem 6 (Finite-sample GD convergence)**.** *Suppose the initialization satisfies Theorem 5 and, at each step t, the population gradient is estimated from a fresh batch of n i.i.d. samples $(X^{(j)}, y^{(j)})_{j \leq n}$.*

Then there exist constants C_1, C_2 depending only on β, L such that with probability $\geq 1-\delta$, projected gradient descent reaches value overlap $\mu_T \geq 0.9$ and selection overlap $m_{T'} \geq 1-\varepsilon$ provided:

$$n \;\geq\; C_1 \,\frac{d\log(T_{\mathrm{pop}}/\delta)}{\varepsilon^2}, \qquad T_{\mathrm{pop}} = O(\log(1/\varepsilon)) \text{ as above.}$$

Thus the total sample complexity is $O(d\,\varepsilon^{-2}\log(1/\varepsilon))$ up to confidence logarithms and constants depending on (β, L, η); for fixed accuracy its dimensional exponent is one.

Proof. Each empirical gradient estimate $\hat{g}_t$ satisfies $\|\hat{g}_t - g_t\|_2 \leq C\sqrt{d/n}$ with probability $\geq 1-\delta/T$ by the sub-exponential concentration of Lemma 7 and equation (1): the summands are products of a sub-exponential scalar ($|y| \leq \max_i |v_i|$) and a sub-Gaussian d-vector (x_i), so the empirical mean concentrates at the $\sqrt{d/n}$ rate by a truncated vector Bernstein argument (Tropp, 2015, Thm. 6.1.1). To track the population drift down to selection error ε, it suffices that this error is at most a constant multiple of $\varepsilon\mu_\dagger c_1$, which gives $n \geq C_1 d/\varepsilon^2$ for fixed (β, L, η). A union bound over T_{pop} steps with failure probability δ/T_{pop} gives the stated confidence factor. □

Remark 10 (Interpretation)**.** *Theorems 5 and 6 extend Corollary 2 across the invariant slice: the exact slice potential $\mathcal{L} = \bar{P} - \mu\, C(m)$ with the non-negative Mehler series C (Lemmas 9–10) has its* only *stationary point at the optimum, so there is no interior saddle for the cascade to get trapped in; for fixed accuracy the two-stage value→selection cascade converges in a number of population steps independent of d, and $n = \Theta(d)$ samples per step suffice to keep the finite-sample noise below the $\Theta(1)$ population signal. Its dimensional exponent matches the all-noise oracle lower bound of Theorem 8; the one-shot spectral bound of Theorem 3 has the same exponent with logarithmic overhead. The equator catalyst gradient of Theorem 1 is now seen to be exactly the first Taylor coefficient $\mu\, c_1$ of the global drift, and the Mehler non-negativity is what guarantees that this initial push never reverses as the selection overlap grows.*

E Noisy minimax and noiseless identifiability lower bounds

For every fixed $\sigma > 0$ we show the matching dimensional lower bound directly, without appeal to statistical-query or low-degree heuristics: *every* measurable estimator needs $\Omega(d)$ samples. The argument uses Fano's inequality over a local packing of the parameter space—whose only nonstandard ingredient is a sharp per-sample Kullback–Leibler bound. That bound follows from the exact Lipschitz constant of the softmax map, which we isolate first.

Lemma 12 (Tight softmax Lipschitz constant)**.** *Let $a(z) = \mathrm{softmax}(z) \in \Delta^{L-1}$ and write its Jacobian $J(z) = \mathrm{diag}(a) - aa^\top$. Then for every z and every $h \in \mathbb{R}^L$,*

$$h^\top J(z)\, h \;=\; \mathrm{Var}_a(h) \;\leq\; \tfrac{1}{4}\big(\max_i h_i - \min_i h_i\big)^2 \;\leq\; \tfrac{1}{2}\|h\|_2^2,$$

so $\|J(z)\|_{\mathrm{op}} \leq \frac{1}{2}$, with $\frac{1}{2}$ attained as $a \to (\frac{1}{2}, \frac{1}{2}, 0, \dots)$. Consequently $z \mapsto a(z)$ is $\frac{1}{2}$-Lipschitz and, for scores $s, s' \in \mathbb{R}^L$,

$$\big\|a(\beta s) - a(\beta s')\big\|_2 \;\leq\; \frac{\beta}{2}\,\|s - s'\|_2. \tag{6}$$

Proof. The identity $h^\top J h = \sum_i a_i h_i^2 - (\sum_i a_i h_i)^2 = \mathrm{Var}_a(h)$ is the standard softmax Jacobian. Popoviciu's inequality bounds the variance of a random variable supported in $[\min_i h_i, \max_i h_i]$ by one quarter of the squared range, giving the middle inequality; the range is at most $2\|h\|_\infty \leq 2\|h\|_2$, but the sharper bound $\frac{1}{4}(\max h - \min h)^2 \leq \frac{1}{2}\|h\|_2^2$ is what we use (it is tight when h has two nonzero, opposite-sign entries and a is supported on them with mass $\frac{1}{2}, \frac{1}{2}$). Hence $J \preceq \frac{1}{2}I$ and, as $J \succeq 0$, $\|J\|_{\mathrm{op}} \leq \frac{1}{2}$. Integrating the operator-norm bound along the segment $t \mapsto a(\beta s' + t\beta(s - s'))$ yields the Lipschitz estimate (6). The value $\frac{1}{2}$ is achieved in the limit $a \to (\frac{1}{2}, \frac{1}{2}, 0, \dots)$, where $J \to \frac{1}{4}\left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right) \oplus 0$, of operator norm exactly $\frac{1}{2}$. □

The Lipschitz bound converts a key-direction perturbation into a controlled change of the regression function, hence a controlled KL divergence between the induced sample laws.

Lemma 13 (Per-sample KL bound)**.** *Fix the value direction $u^\star$ and consider two key directions w, w', both orthogonal to $u^\star$, with corresponding regression functions $f_w(X) = \sum_i a_i(\beta\langle w, x_i\rangle)\,\langle u^\star, x_i\rangle$ and labels $y = f_w(X) + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2)$. Let P_w denote the law of a single sample (X, y). Then*

$$\mathrm{KL}\big(P_w \,\|\, P_{w'}\big) \;=\; \frac{1}{2\sigma^2}\,\mathbb{E}_X\big[(f_w(X) - f_{w'}(X))^2\big] \;\le\; \frac{\beta^2 L}{8\sigma^2}\,\|w - w'\|_2^2. \tag{7}$$

Proof. X has the same marginal under P_w and $P_{w'}$, and given X the labels are $\mathcal{N}(f_\bullet(X), \sigma^2)$; the chain rule for KL therefore gives the equality, the conditional KL between two equal-variance Gaussians being $(f_w - f_{w'})^2/(2\sigma^2)$. For the inequality, fix X and abbreviate $s_i = \langle w, x_i\rangle$, $s_i' = \langle w', x_i\rangle$, $v_i = \langle u^\star, x_i\rangle$. Since $a(\beta s)$ is a probability vector, $f_w(X) - f_{w'}(X) = \langle a(\beta s) - a(\beta s'), v\rangle$, and because $w, w' \perp u^\star$ the value scores v are independent of (s, s') with $\mathbb{E}[vv^\top] = I_L$. Taking expectation over v first,

$$\mathbb{E}_X\big[(f_w - f_{w'})^2\big] = \mathbb{E}_{s,s'}\big\|a(\beta s) - a(\beta s')\big\|_2^2 \;\le\; \frac{\beta^2}{4}\,\mathbb{E}\,\|s - s'\|_2^2,$$

by Lemma 12. Finally $s_i - s_i' = \langle w - w', x_i\rangle \sim \mathcal{N}(0, \|w - w'\|^2)$ independently across $i \in [L]$, so $\mathbb{E}\|s - s'\|^2 = L\|w - w'\|^2$, giving (7). □

Theorem 7 (Information-theoretic lower bound)**.** *Consider the single-location unit (Definition 1) with known value direction $u^\star$, Gaussian label noise of fixed variance $\sigma^2 > 0$, and unknown key direction $w^\star \in \mathbb{S}^{d-1} \cap u^{\star\perp}$. There are absolute constants $c_0, c_1 > 0$ such that any unit-vector estimator $\widehat{w} = \widehat{w}(X^{(1)}, y^{(1)}, \ldots, X^{(n)}, y^{(n)})$ satisfies*

$$\sup_{w^\star}\; \mathbb{E}\big[1 - \langle \widehat{w}, w^\star\rangle^2\big] \;\ge\; c_0 \qquad \textit{whenever} \qquad n \;\le\; c_1\,\frac{\sigma^2 d}{\beta^2 L}.$$

In particular, consistent recovery requires $n = \Omega(\sigma^2 d/(\beta^2 L)) = \Omega(d)$ samples for fixed (β, L, σ).

Proof. Write $\mathbb{S} = \mathbb{S}^{d-1} \cap u^{\star\perp}$ and $p = d - 2$. Fix $e_0 \in \mathbb{S}$ and an isometry $T : \mathbb{R}^p \to u^{\star\perp} \cap e_0^\perp$. The Varshamov–Gilbert construction (Tsybakov, 2009, Lemma 2.9) gives $z_1, \ldots, z_M \in \{\pm p^{-1/2}\}^p$ with $\log M \ge cp$ and pairwise Hamming distance at least $p/8$. For fixed $\delta \in (0, 1/2)$, define $w_j = \sqrt{1 - \delta^2}\,e_0 + \delta T z_j \in \mathbb{S}$. Then

$$\frac{\delta}{\sqrt{2}} \;\le\; \|w_j - w_k\|_2 = 2\delta\sqrt{\frac{\mathrm{Ham}(z_j, z_k)}{p}} \;\le\; 2\delta.$$

Writing $r = \delta/(2\sqrt{2})$, the corresponding axes are $2r$-separated in $d_\pm(v, w) := \min\{\|v - w\|, \|v + w\|\}$ and every pair is at distance at most $4\sqrt{2}\,r$; all packing points lie in the same cap, so their pairwise inner products are positive. Run the M-ary hypothesis test: nature picks $w^\star = w_J$ with J uniform on $[M]$, and an estimator induces a test $\widehat{J} = \arg\min_j d_\pm(\widehat{w}, w_j)$. Since $d_\pm(\widehat{w}, w^\star)^2 \le 2(1 - \langle \widehat{w}, w^\star\rangle^2)$ for unit vectors, expected projective loss below $r^2/4$ implies $\Pr(d_\pm(\widehat{w}, w^\star) \ge r) < 1/2$ by Markov's inequality. Whenever $d_\pm(\widehat{w}, w^\star) < r$ the estimate is strictly closer to the axis of $w^\star$ than to any other packing point (the points being $\ge 2r$ apart), so $\widehat{J} = J$; hence $\Pr(\widehat{J} \ne J) < 1/2$. We contradict this for small n via Fano.

By tensorization of KL over the n i.i.d. samples and Lemma 13,

$$\mathrm{KL}\big(P_{w_j}^{\otimes n} \,\|\, P_{w_k}^{\otimes n}\big) \;=\; n\,\mathrm{KL}(P_{w_j}\|P_{w_k}) \;\le\; n\,\frac{\beta^2 L}{8\sigma^2}\,\|w_j - w_k\|_2^2 \;\le\; n\,\frac{\beta^2 L}{8\sigma^2}\,(4\sqrt{2}\,r)^2 \;=\; \frac{4n\beta^2 L\, r^2}{\sigma^2}.$$

Fano's inequality (Tsybakov, 2009, Cor. 2.6) gives

$$\Pr(\widehat{J} \ne J) \;\ge\; 1 - \frac{\max_{j \ne k}\mathrm{KL}(P_{w_j}^{\otimes n}\|P_{w_k}^{\otimes n}) + \log 2}{\log M} \;\ge\; 1 - \frac{4n\beta^2 L\, r^2/\sigma^2 + \log 2}{c\,(d-2)}.$$

Fix $r = 1/(8\sqrt{2})$ (equivalently $\delta = 1/4$). Then whenever $n \le c_1\sigma^2 d/(\beta^2 L)$ with c_1 a small enough absolute constant, the right-hand side exceeds $1/2$, so $\Pr(\widehat{J} \ne J) \ge 1/2$, contradicting projective risk below $r^2/4$. Hence the minimax projective risk is at least $c_0 := r^2/4$ for all such n, which is the claim. □

Remark 11 (Parameter dependence of the lower bound). *The lower bound scales as $\sigma^2 d/(\beta^2 L)$, matching the spectral upper bound's dimension factor. Its parameter dependence has the expected qualitative direction: larger noise σ^2 makes the directions harder to separate (more samples), while a larger inverse temperature β or context length L makes the regression function more sensitive to w (fewer samples). The coefficient $\beta^2 L/8$ in (7) follows from the $1/2$ softmax Jacobian norm and the $1/(2\sigma^2)$ Gaussian KL factor.* `numerics/minimax_kl_bound.py` *checks the regression-function identity and the Jacobian norm (*0.5000*). The bound is information-theoretic: it applies to* any *estimator, so a fortiori to every statistical-query or low-degree algorithm, and it does not invoke the degree-two SQ threshold.*

The Fano bound of Theorem 7 carries an explicit σ^2 and is vacuous at $\sigma = 0$. A separate oracle argument gives a projective-risk lower bound at every noise level, including zero, by revealing *more* information than the labels.

Theorem 8 (All-noise oracle and exact-identifiability floors). *Let the value direction $u^\star$ be known and write $D = d - 1$. For every $\sigma \geq 0$ and every unit-vector estimator,*

$$\sup_{w^\star \in \mathbb{S}^{d-1} \cap u^{\star\perp}} \mathbb{E}\big[1 - \langle \widehat{w}, w^\star \rangle^2\big] \;\geq\; \left(1 - \frac{nL+1}{D}\right)_+ .$$

Thus constant-accuracy projective recovery requires $n = \Omega(d/L) = \Omega(d)$ for fixed L, at every noise level.

In addition, fix any designs $X^{(1)}, \ldots, X^{(n)}$ and define the exact label map $\Phi_X : \mathbb{S}^{d-2} \to \mathbb{R}^n$. If $n \leq d - 2$, there exists w with $\Phi_X(w) = \Phi_X(-w)$, so uniform exact identifiability for that design requires $n \geq d - 1$.

Proof. Give the learner all nL exact key scores $s_i^{(t)} = \langle w^\star, x_i^{(t)} \rangle$, in addition to the designs and known $u^\star$. This oracle observation is more informative than either noisy or noiseless labels. In coordinates on $u^{\star\perp} \simeq \mathbb{R}^D$, it is the linear system $Aw^\star$, where A has rank $k \leq \min(nL, D)$ almost surely.

Use the uniform prior on $\mathbb{S}^{D-1}$. Conditional on A and $Aw^\star$, the claim is trivial when $k = D$, so suppose $k < D$. Then the projection q of $w^\star$ onto the row space of A is known, while the remaining component is uniform on the sphere of radius $\sqrt{1 - \|q\|^2}$ in the $(D - k)$-dimensional nullspace. Its conditional second moment is

$$qq^\top + \frac{1 - \|q\|^2}{D - k} \Pi_{\ker A}.$$

For any unit estimate, the conditional expected squared overlap is at most the largest eigenvalue of this matrix, which is bounded by $\|q\|^2 + (1 - \|q\|^2)/(D - k)$. Rotational invariance gives $\mathbb{E}\|q\|^2 = k/D$ and $\mathbb{E}[(1 - \|q\|^2)/(D - k)] = 1/D$. Hence the Bayes expected squared overlap is at most $(k + 1)/D \leq (nL + 1)/D$. Minimax risk is at least Bayes risk, proving the first claim.

For the second claim, Φ_X is continuous. Pad it with zeros if $n < d - 2$ and apply the Borsuk–Ulam theorem (Matoušek, 2003, Chapter 2) to obtain an antipodal pair with the same image. □

Remark 12 (Roles of the lower bounds). *The oracle bound matches the projective loss used by Theorem 3 and gives the linear dimensional exponent for every $\sigma \geq 0$. For fixed $\sigma > 0$, the Fano bound of Theorem 7 additionally records explicit dependence on (σ, β, L). The Borsuk–Ulam part is a complementary deterministic statement about uniform exact identification for each fixed design. Finally, the Jacobian calculation in* `numerics/noiseless_identifiability.py` *checks the generic local geometry: its rank is $\min(n, d - 2)$, so regular fibers have the expected local dimension.*

F Detailed numerical confirmation

The following predictions are checked with short CPU `numpy` scripts in `numerics/`. Table 1 summarises. (i) `probe.py`: $\mathbb{E}[y\ \mathrm{He}_k(s_1)] = 0$ for $k = 1, \ldots, 4$ across β (information exponent ∞), while $\mathbb{E}[y^2\ \mathrm{He}_2(s_1)] > 0$. (ii) `decisive_gradient.py` / `two_regime_gradient.py`: the population key-gradient component along $w^\star$ is at the noise floor when the value is unlearned ($\mu = 0$)

and $\Theta(1)$ once $u = u^\star$, confirming the saddle and its unlocking. (iii) `unlock_closedform.py`: the closed form $\nabla_w \mathcal{L} \cdot w^\star = -\beta^2 \mu (1-\bar{P})^2/(L-1)$ matches the measured gradient to absolute error about 10^{-3} across $L \in \{2, \dots, 5\}$, $\beta \in \{1, 2, 3\}$, $\mu \in \{0, \frac{1}{2}, 1\}$, and the Stein identity $A = \beta(1-\bar{P})$ holds. (iv) `gd_trajectory.py`: end-to-end projected GD from random init exhibits the two-timescale cascade—value overlap rises first, selection overlap follows, both reaching ≈ 1 in $O(1)$ steps. (v) `lemma_check.py`: $\mathbb{E}[(y^2 - P)g(s)] = 0$ for $g \in \{s_1, \mathrm{He}_2, \cos\}$, confirming $\mathbb{E}[y^2 \mid s] = P(s)$ exactly. (vi) `ge_verify.py`: degree-one Hermite coefficient of P is at the $1/\sqrt{N}$ noise floor while degree-two is strictly positive for every (β, L). (vii) `ge_exact.py`: a binned estimate of $\mathrm{Var}(\mathbb{E}[s_1 \mid y])$ is at its $O(\text{bins}/N)$ in-sample null scale, consistent with the analytic GE ≥ 2 result; meanwhile $\mathrm{corr}(v_1, y) > 0$ and $\mathbb{E}[y v_1] = 1/L$ (value IE $= 1$). (viii) `c2_closedform.py`: the closed form of Appendix H agrees with Monte Carlo at the percent level or better. (ix) `scaling_threshold.py`: the spectral recovery overlap, plotted against $\alpha = n/d$, collapses across $d \in \{40, 80\}$, consistent with linear scaling. (x) `verify_rank2.py`: the population matrix M has exactly two spikes, $2Lc_2$ along $w^\star$ and $2\bar{P}$ along $u^\star$, with perp and cross quadratic forms at the noise floor (Lemma 8). (xi) `multihead_ge.py`: under additive head combination the degree-two coefficient is flat in H, under multiplicative combination it tracks $2c_2\bar{P}^{H-1}$, both with vanishing degree one—confirming GE $= 2$ for all H (Theorem 9). (xii) `phase_transition.py`: for $\rho \in \{0.1, 0.3, 0.5, 1\}$, $L \in \{2, 4\}$, $\beta \in \{1, 2, 3\}$, the ratio $\mathbb{E}[y\,\mathrm{He}_1(s_1)]/(\rho/L)$ is within 0.05 of one, consistent with Theorem 4; at $\rho = 0$ the projection sits at the Monte-Carlo noise floor. (xiii) `convergence_rate.py`: a large-batch GD proxy reaches overlap 0.8 in approximately 20 steps for $d \in \{40, 80, 160\}$ with no observed growth in d, consistent with Theorem 5. (xiv) `noiseless_identifiability.py`: the Jacobian of the label map has rank exactly $\min(n, d-2)$ across $d \in \{10, 20, 40\}$, so its kernel is nontrivial precisely when $n < d - 2$. This verifies the generic local dimension calculation that complements the global Borsuk–Ulam obstruction of Theorem 8. (xv) `c2_scaling.py`: the degree-two coefficient follows $c_2 \approx \frac{L-1}{L^3}\beta^2$ at small β, peaks near $\beta \in [1.5, 2]$, and decays to 0 as $\beta \to \infty$, across $L \in \{2, 4, 8\}$, quantifying the rate prefactor $C = \Theta(c_2^{-2})$ of Lemma 6. (xvi) `loss_decomposition.py`: common-random-number finite differences give $C'(0) = 0.2504$ vs. the closed form 0.2512 for $(L, \beta) = (3, 2)$ and 0.1233 vs. 0.1232 for $(4, 1)$; after projecting onto the sphere tangent, the raw gradient agrees with $\mu C'(m)$ within 9×10^{-4} at all six tested (m, μ) pairs. (xvii) `gd_finite_sample_scaling.py`: with a fresh batch of $n = \alpha d$ samples per step, the final-overlap curves collapse across $d \in \{30, 60, 120, 240\}$, with the transition between $\alpha = 0.35$ and 0.5; when the value is clamped at $\mu = 0$, the selection overlap stays at the $d^{-1/2}$ scale. (xviii) `noisy_recovery.py`: the population rank-two matrix is unchanged for $\sigma \in \{0, 0.5, 1, 2\}$, while the empirical α needed for spectral overlap 0.9 grows by factors $1, 2, 4, 16$ for $\sigma \in \{0, 0.5, 1, 1.5\}$, within the factor-two resolution of the predicted $(1 + \sigma^2)^2$ scaling. (xix) `nongaussian_tokens.py`: key-only correlations remain at their empirical noise level and the cascade reaches unit overlap under Gaussian, Rademacher, uniform and Student-t_5 product laws. The same experiment also demonstrates the Gaussian-specific boundary in Remark 6: the degree-one projection of y^2 is nonzero for the three non-Gaussian laws. (xx) `multihead_distinct_beta.py`: direct quadratic-form measurements match the corrected matrix (8); distinct temperatures raise the worst matched per-head selection alignment from 0.694 to 0.926 after the true value subspace is deflated, but the tested temperature tuple violates the eigenvalue ordering needed by an unsupervised top-eigenspace deflation.

G AUXILIARY INEQUALITIES

Lemma 14 (Chebyshev sum / association inequality)**.** *Let $(p_i)_{i \in [L]}$ be a probability vector and let $f, g : [L] \to \mathbb{R}$ be two sequences that are comonotone, i.e. $(f_i - f_j)(g_i - g_j) \geq 0$ for all i, j. Then*

$$\mathrm{Cov}_p(f, g) := \sum_i p_i f_i g_i - \Big(\sum_i p_i f_i\Big)\Big(\sum_i p_i g_i\Big) \;\geq\; 0,$$

with equality iff f or g is p-a.s. constant.

Proof. Let (I, J) be independent copies drawn from p. Then $2\,\mathrm{Cov}_p(f, g) = \mathbb{E}\big[(f_I - f_J)(g_I - g_J)\big] \geq 0$ since the integrand is pointwise nonnegative by comonotonicity. Equality forces $(f_I - f_J)(g_I - g_J) = 0$ a.s., i.e. one sequence is constant on the support. □

Prediction	Script	Outcome
$\mathbb{E}[y\ \mathrm{He}_k(s_1)] = 0\ \forall k$ (IE$=\infty$)	`probe.py`	$\lvert\cdot\rvert \le 10^{-3}$, all k, β
$\nabla_w \mathcal{L} \cdot w^\star = 0$ at $\mu = 0$; $\Theta(1)$ at $\mu = 1$	`decisive_gradient.py`	noise floor vs. $\Theta(1)$
$\nabla_w \mathcal{L} \cdot w^\star = -\beta^2 \mu (1-\bar{P})^2/(L-1)$	`unlock_closedform.py`	absolute error about 10^{-3}
value$\to$selection GD cascade	`gd_trajectory.py`	both overlaps $\to 1$, $O(1)$ steps
$\mathbb{E}[y^2 \mid s] = P(s)$	`lemma_check.py`	residual $\le 10^{-3}$
deg-1 of P vanishes; deg-2 > 0	`ge_verify.py`	$c_1 \sim 10^{-4}$, $c_2 > 0$ all β, L
GE ≥ 2: binned $\mathrm{Var}(\mathbb{E}[s_1 \mid y])$ diagnostic	`ge_exact.py`	at in-sample null scale
value IE $= 1$: $\mathbb{E}[y v_1] = 1/L$	`ge_exact.py`	matches $1/L$
closed form $c_2(\beta, 2)$	`c2_closedform.py`	percent-level agreement or better
$n = \Theta(d)$ threshold collapse	`scaling_threshold.py`	curves collapse in $\alpha = n/d$
$M = 2Lc_2\, w^\star w^{\star\top} + 2\bar{P}\, u^\star u^{\star\top}$	`verify_rank2.py`	spikes match; perp/cross ≈ 0
GE $= 2$ for all H (add./mult.)	`multihead_ge.py`	deg-2 > 0; matches $2c_2\bar{P}^{H-1}$
$\mathbb{E}[y\ \mathrm{He}_1(s_1)] = \rho/L$ (phase transition)	`phase_transition.py`	ratio ≈ 1.00 all β, L, ρ
large-batch GD steps stable across d	`convergence_rate.py`	≈ 20 steps for $d \in \{40, 80, 160\}$
$\mathrm{rank}\, d\Phi = \min(n, d-2)$ (local check)	`noiseless_identifiability.py`	exact, $d \in \{10, 20, 40\}$
$c_2 \sim \frac{L-1}{L^3}\beta^2$ (small β), $\to 0$ (large β)	`c2_scaling.py`	law + peak, $L \in \{2, 4, 8\}$

Table 1: Core numerical checks (CPU, numpy).

Additional check	Script	Outcome
$C'(0)$ and tangent gradient $\mu C'(m)$	`loss_decomposition.py`	closed form within 9×10^{-4}
Finite-sample GD at $n = \alpha d$; $\mu = 0$ control	`gd_finite_sample_scaling.py`	collapse across d; transition $\alpha \in [0.35, 0.5]$
Label-noise robustness	`noisy_recovery.py`	same population spikes; empirical cost follows $(1+\sigma^2)^2$ within grid resolution
Non-Gaussian product tokens	`nongaussian_tokens.py`	IE blindness and cascade persist; GE lower bound is Gaussian-specific
Corrected multi-head matrix; distinct temperatures	`multihead_distinct_beta.py`	matrix entries match; degeneracy breaks, but top-eigenspace ordering can fail

Table 2: Finite-sample and robustness checks added to the core suite.

Lemma 15 (Radial association for spherically symmetric laws)**.** *Let $Z \sim \mathcal{N}(0, I_m)$ and let $F : \mathbb{R}^m \to \mathbb{R}$ be such that $t \mapsto F(t\omega)$ is nondecreasing for every $\omega \in \mathbb{S}^{m-1}$ and $t \ge 0$ (radial monotonicity), and F is nonconstant. Then* $\mathrm{Cov}\big(F(Z), \|Z\|^2\big) > 0$.

Proof. Condition on the direction $\Omega = Z/\|Z\|$, which is uniform on the sphere and independent of the radius $R = \|Z\|$. Given $\Omega = \omega$, both $F(R\omega)$ and R^2 are nondecreasing functions of the single variable $R \ge 0$, so by the one-dimensional Chebyshev association inequality (for the law of R),

$$\mathbb{E}\big[F(Z)\,\|Z\|^2 \mid \Omega = \omega\big] \;\ge\; \mathbb{E}\big[F(Z) \mid \Omega = \omega\big]\,\mathbb{E}\big[\|Z\|^2\big].$$

Taking expectation over Ω and using $\mathbb{E}[\|Z\|^2 \mid \Omega] = \mathbb{E}\|Z\|^2 = m$ (independence of R, Ω),

$$\mathbb{E}\big[F(Z)\|Z\|^2\big] \;\ge\; \mathbb{E}[F(Z)]\,\mathbb{E}\|Z\|^2,$$

i.e. $\mathrm{Cov}(F(Z), \|Z\|^2) \ge 0$. Strictness: nonconstancy of F together with radial monotonicity implies $R \mapsto F(R\omega)$ is strictly increasing on a set of (ω, R) of positive measure, where the one-dimensional Chebyshev inequality is strict; hence the aggregate covariance is strictly positive. □

Lemma 15 is applied in Lemma 5 with $m = L - 1$, $Z = s_\perp$ (which is $\mathcal{N}(0, I_{L-1})$ in the contrast subspace), and $F = P$, whose radial monotonicity is established there.

H CLOSED FORM OF $c_2(\beta, 2)$

For $L = 2$, write the contrast $r = s_1 - s_2 \sim \mathcal{N}(0, 2)$. Softmax gives $a_1 = \sigma(\beta r)$, $a_2 = \sigma(-\beta r)$ with σ the logistic function, so the collision probability is

$$P = a_1^2 + a_2^2 = 1 - 2a_1a_2 = 1 - 2\sigma(\beta r)\sigma(-\beta r) = \tfrac{1}{2}\Big(1 + \tanh^2(\tfrac{\beta r}{2})\Big).$$

Using $s_1 = \frac{1}{2}r + \frac{1}{\sqrt{2}}\bar{r}'$ with the orthogonal mean part independent of r, and $\mathbb{E}[\mathrm{He}_2(s_1) \mid r] = \frac{1}{4}(r^2 - 2) + \text{const}$, a direct computation (or two Stein integrations by parts) yields

$$c_2(\beta, 2) = \tfrac{1}{2}\,\mathbb{E}[P\ \mathrm{He}_2(s_1)] = \frac{\beta^2}{8}\,\mathbb{E}_{r\sim\mathcal{N}(0,2)}\Big[\operatorname{sech}^2(\tfrac{\beta r}{2})\big(3\operatorname{sech}^2(\tfrac{\beta r}{2}) - 2\big)\Big].$$

This is strictly positive for all $\beta > 0$: the expectation is dominated by the region $|r| \lesssim 1/\beta$ where $\operatorname{sech}^2 \approx 1$ and the bracket $\approx 1 > 0$. As $\beta \to 0$, $c_2 \sim \beta^2/8 \cdot \mathbb{E}[1 \cdot 1] \to 0$ (quadratically, no selection); as $\beta \to \infty$ the sech^2 factor concentrates on a $\mathcal{N}(0, 2)$-null set so $c_2 \to 0$; the function peaks near $\beta \approx 2$ at $c_2 \approx 0.044$. The numerical evaluation in `numerics/c2_closedform.py` agrees with the direct Monte-Carlo estimate of $\frac{1}{2}\mathbb{E}[P\,\mathrm{He}_2(s_1)]$ at the percent level or better across $\beta \in \{0.25, \ldots, 8\}$.

I MULTI-HEAD ROBUSTNESS OF THE DICHOTOMY

The single-unit dichotomy (IE $= \infty$, GE $= 2$) isolates softmax shift invariance on the selection coordinate. We next ask whether combining heads changes the exponent. For additive and multiplicative orthogonal-subspace readouts, it does *not*: each selection direction keeps IE $= \infty$ and GE $= 2$. The infinite *information* exponent of every selection direction is therefore a robust per-direction property of softmax selection, not an artifact of the single-unit model. (One might expect GE $= \Theta(H)$ for the product readout, since the product of H degree-two signals naively suggests degree $2H$; the calculation below shows this intuition is false—the degree-two signal survives, with only a constant that decays geometrically in H.)

Setup. Let $w_1^\star, u_1^\star, \ldots, w_H^\star, u_H^\star$ be $2H$ orthonormal directions, and let head h produce $y_h = \sum_i a_i(s^{(h)})v_i^{(h)}$ from its own key/value scores $s_i^{(h)} = \langle w_h^\star, x_i\rangle$, $v_i^{(h)} = \langle u_h^\star, x_i\rangle$ as in Definition 1. By orthonormality the $2H$ score blocks are mutually independent. Write $P_h(s^{(h)}) = \sum_i a_i(s^{(h)})^2$ and $\bar{P} = \mathbb{E}[P_h] \in (1/L, 1)$ (equal across heads).

Lemma 16 (IE stays infinite under both readouts)**.** *For both the additive readout $y = \sum_h y_h$ and the multiplicative readout $y = \prod_h y_h$, the conditional mean given all key blocks vanishes:* $\mathbb{E}[\,y \mid s^{(1)}, \ldots, s^{(H)}\,] = 0$. *Hence* $\mathrm{IE}_{w_h^\star}(y) = \infty$ *for every head h.*

Proof. Condition on all key blocks $s = (s^{(1)}, \ldots, s^{(H)})$; the weights $a_i(s^{(h)})$ are then fixed and the value blocks $v^{(h)}$ remain independent mean-zero Gaussians. For the additive readout, $\mathbb{E}[y \mid s] = \sum_h \sum_i a_i(s^{(h)})\mathbb{E}[v_i^{(h)}] = 0$. For the multiplicative readout, the heads are conditionally independent given s, so $\mathbb{E}[\prod_h y_h \mid s] = \prod_h \mathbb{E}[y_h \mid s] = \prod_h 0 = 0$. The IE claim follows as in Corollary 1. □

Theorem 9 (GE stays two under both readouts)**.** *For the additive readout, the degree-two Hermite coefficient of y^2 along key coordinate $s_1^{(h)}$ equals $2c_2(\beta, L)$ for every h and every H. For the multiplicative readout, it equals $2c_2(\beta, L)\,\bar{P}^{\,H-1} > 0$. In both cases* $\mathrm{GE}_{w_h^\star}(y) = 2$*, and the deflated spectral step of Theorem 3 recovers the* selection subspace $W^\star = \mathrm{span}\{w_1^\star, \ldots, w_H^\star\}$ *at the rate $n = \tilde{O}(d/\kappa_H^2)$, where $\kappa_H = Lc_2$ (additive) or $\kappa_H = Lc_2\bar{P}^{H-1}$ (multiplicative). Recovery of the* individual *directions $w_h^\star$ requires an additional symmetry-breaking ingredient, made precise in Remark 13.*

Proof. Additive. $y^2 = \sum_h y_h^2 + \sum_{h\neq h'} y_h y_{h'}$. Conditioning on all key blocks: $\mathbb{E}[y_h^2 \mid s] = P_h$, and for $h \neq h'$, $\mathbb{E}[y_h y_{h'} \mid s] = \mathbb{E}[y_h \mid s]\mathbb{E}[y_{h'} \mid s] = 0$ by conditional independence and Lemma 16. So $\mathbb{E}[y^2 \mid s] = \sum_h P_h(s^{(h)})$. The degree-two projection along $s_1^{(1)}$ picks out only the $h = 1$ summand

(the others are independent of $s^{(1)}$ and contribute constants), giving $\frac{1}{2}\mathbb{E}[P_1\,\mathrm{He}_2(s_1^{(1)})] = c_2$, i.e. a degree-two coefficient $2c_2$ in the matrix sense of Lemma 8. Degree one vanishes by Lemma 4 applied to P_1.

Multiplicative. By conditional independence, $\mathbb{E}[y^2 \mid s] = \prod_h \mathbb{E}[y_h^2 \mid s] = \prod_h P_h(s^{(h)})$. Projecting along $s_1^{(1)}$ and using independence of the blocks,

$$\tfrac{1}{2}\,\mathbb{E}\Big[\Big(\prod\nolimits_h P_h\Big)\,\mathrm{He}_2(s_1^{(1)})\Big] = \tfrac{1}{2}\,\mathbb{E}\big[P_1\,\mathrm{He}_2(s_1^{(1)})\big]\prod\nolimits_{h\geq 2}\mathbb{E}[P_h] = c_2\,\bar{P}^{\,H-1}.$$

Degree one again vanishes because each P_h is shift invariant in its own block (Lemma 4). Since $c_2 > 0$ and $\bar{P} > 0$, the degree-two coefficient is strictly positive for every finite H, so $\mathrm{GE} = 2$. It remains to read off what the spectral matrix recovers. Summing the single-block computation over all heads, the population matrix of Lemma 8 generalises to

$$M \;=\; 2\kappa_H \sum_{h=1}^{H} w_h^\star w_h^{\star\top} \;+\; 2\nu_H \sum_{h=1}^{H} u_h^\star u_h^{\star\top} \;+\; \tau_H \sum_{h\neq h'} u_h^\star u_{h'}^{\star\top}, \tag{8}$$

with

$$(\kappa_H, \nu_H, \tau_H) = \begin{cases} \big(Lc_2,\ \bar{P},\ 2/L\big) & \text{(additive)},\\ \big(Lc_2\bar{P}^{\,H-1},\ \bar{P}^{\,H},\ 0\big) & \text{(multiplicative)}. \end{cases}$$

Three of the four families of cross terms do vanish, by independence of the blocks together with the vanishing of an odd Gaussian moment: $w_h^\star w_{h'}^{\star\top}$ for $h \neq h'$, because $\mathbb{E}[y_h\,\mathrm{He}_1(s_i^{(h)})] = 0$ is exactly Corollary 1; $w_h^\star u_{h'}^{\star\top}$ for all h, h', by the third-moment argument of Lemma 8(c); and $u_h^\star u_{h'}^{\star\top}$ in the multiplicative case, because $\mathbb{E}[y_h^2 v_i^{(h)} \mid s] = 0$ is again an odd moment. The remaining family does *not* vanish in the additive case, and this is the one point where the multi-head matrix is not a copy of the single-head one. Expanding $y^2 = \sum_h y_h^2 + \sum_{h\neq h'} y_h y_{h'}$, the quadratic form of M between $u_h^\star$ and $u_{h'}^\star$ picks up

$$\sum_{i=1}^{L} \mathbb{E}\big[y_h y_{h'} v_i^{(h)} v_i^{(h')}\big] = \sum_{i=1}^{L} \mathbb{E}\big[y_h v_i^{(h)}\big]\,\mathbb{E}\big[y_{h'} v_i^{(h')}\big] = \sum_{i=1}^{L} \frac{1}{L^2} = \frac{1}{L},$$

using block independence and Lemma 7 ($\mathbb{E}[y_h v_i^{(h)}] = \mathbb{E}[a_i(s^{(h)})] = 1/L$), which contributes $\tau_H = 2/L$ after symmetrising. Note that τ_H is independent of β.

The structure of (8) is what the estimator needs, and it is unaffected. M still vanishes off $W^\star \oplus U^\star$, and it is still block diagonal with respect to that splitting: the extra term lives entirely inside $U^\star$. On $W^\star$ it acts as $2\kappa_H I_H$; on $U^\star$, in the basis $(u_h^\star)$, it acts as $2\nu_H I_H + \tau_H(J_H - I_H)$ with J_H the all-ones matrix, whose eigenvalues are $2\nu_H + (H-1)\tau_H$ once and $2\nu_H - \tau_H$ with multiplicity $H-1$. So provided

$$2\nu_H - \tau_H \;\neq\; 2\kappa_H, \tag{9}$$

the value subspace is the span of the H eigenvectors carrying those two eigenvalues and is separated from $W^\star$; one deflates it and reads $W^\star$ off the remaining H-fold spike $2\kappa_H$, at the stated rate. For the multiplicative readout (9) reduces to $\bar{P} \neq Lc_2$, the single-head condition. For the additive readout it reads $\bar{P} - 1/L \neq Lc_2$, which holds in every configuration we have measured, though the two sides agree to leading order as $\beta \to 0$ (both are $\beta^2(L-1)/L^2$), so the additive deflation degrades at small inverse temperature. The corrected constants in (8) are checked against direct Monte-Carlo measurement of the quadratic forms for both readouts in `numerics/multihead_distinct_beta.py`.

None of this touches the exponent: the selection block is exactly $2\kappa_H I_H$ in both readouts, with $\kappa_H > 0$, which is the statement $\mathrm{GE} = 2$ and a linear dimensional exponent. This proves the subspace claim; the individual-direction claim is deferred to Remark 13. □

Remark 13 (Subspace vs. individual recovery: the degeneracy and how to break it)**.** *When the heads share the same (β, L) the H selection spikes carry the* identical *eigenvalue $2Lc_2$ (resp. $2Lc_2\bar{P}^{H-1}$), so the top eigenspace of M is degenerate: a single eigendecomposition returns an arbitrary orthonormal basis of $W^\star = \mathrm{span}\{w_h^\star\}$, not the individual $w_h^\star$. The population matrix M is invariant under any rotation of the tuple $(w_1^\star, \ldots, w_H^\star)$ inside $W^\star$ that simultaneously rotates $(u_1^\star, \ldots, u_H^\star)$, so no second-moment statistic of y^2 alone can separate the heads. Individual*

recovery therefore needs one further ingredient that breaks the rotational symmetry within each subspace. One option is (i) distinct temperatures $\beta_1, \dots, \beta_H$*. For the additive readout the selection block becomes* $\operatorname{diag}\big(2Lc_2(\beta_h, L)\big)$*; for the multiplicative readout its* h*th diagonal entry is* $2Lc_2(\beta_h, L)\prod_{g\neq h}\bar{P}(\beta_g, L)$*. Thus the individual* $w_h^\star$ *are identified once these entries are separated by more than the concentration error and the value subspace is available for deflation. Two caveats are visible in* `numerics/multihead_distinct_beta.py`*. First, Lemma 6 shows that* $c_2(\cdot, L)$ *vanishes at both endpoints and has an interior maximum. Distinct temperatures can therefore yield nearly equal spikes; the separation must be checked for the chosen tuple, and the achievable gaps can be small. Second, this mechanism separates the* selection *directions, which is what we are after, but not in general the value directions: the additive value block is* $2\operatorname{diag}(\bar{P}(\beta_h, L)) + \frac{2}{L}(J_H - I_H)$*, whereas the multiplicative value block is* $2\prod_g \bar{P}(\beta_g, L)\, I_H$*, degenerate for any choice of temperatures. For the tested additive tuple* $(0.4, 0.8, 1.3, 2)$*, the smallest value eigenvalue below its largest selection spike, so the top-*H *eigenspace is not the value subspace. Distinct temperatures therefore break the selection degeneracy but do not by themselves guarantee the top-eigenspace deflation pipeline for arbitrary tuples. Absent a separately identified value subspace, one may instead use (ii) a* joint key–value fourth moment $T_h = \mathbb{E}\big[y^2\ \mathrm{He}_2(\langle u, x\rangle\text{-contrast})\,(x_i x_i^\top - I)\big]$ *that couples a probe direction* u *in the value subspace to the selection spike: scanning* u *over the (already recovered) value subspace and reading how the selection spike co-varies identifies the pairing* $w_h^\star \leftrightarrow u_h^\star$*, because head* h*'s label depends jointly on the matched pair* $(w_h^\star, u_h^\star)$ *and on no other. We leave the analysis of this fourth-moment construction and the sharp separation constants for individual-head recovery to future work.*

Remark 14 (Couplings that could raise the exponent)**.** *Theorem 9 shows that* orthogonal-subspace *coupling, however the heads are combined arithmetically, never raises the generative exponent: the softmax shift invariance acts block-by-block and degree two always survives. The exponent can only increase if the heads* share *a selection subspace in a way that makes the degree-two contrast cancel—e.g. antisymmetric couplings where* $\mathbb{E}[P_1\,\mathrm{He}_2]$ *is forced to zero by a symmetry beyond shift invariance, so that the leading transform-stable degree is pushed to four or higher. Identifying the minimal such coupling, and whether it arises in trained attention, is the natural open problem; the present results delimit it sharply by ruling out the entire orthogonal-subspace family. The geometric decay* $\bar{P}^{H-1}$ *in the multiplicative case is a statement about the* constant *(hence the sample-complexity prefactor), not the exponent: deep multiplicative stacks make each selection direction quantitatively harder to extract while leaving the linear dimensional exponent intact.*

Both predictions are confirmed in `numerics/multihead_ge.py`: the additive degree-two coefficient is flat in H while the multiplicative one follows the predicted geometric trend $2c_2\bar{P}^{H-1}$, and the degree-one coefficient sits at the Monte-Carlo noise floor throughout.