arXiv:2605.05189v2 [stat.ML] 19 Aug 2026

Sharp Capacity Thresholds in Linear Associative Memory: From Top-1 Retrieval to Tail-Average Learning

Nicholas Barnfield[1] Juno Kim[2] Eshaan Nichani[3]
Jason D. Lee[2] Yue M. Lu[1*]

[1]Harvard University [2]University of California, Berkeley [3]Princeton University

August 19, 2026

Abstract

How many key-value associations can a $d \times d$ linear memory store? The answer depends not only on the d^2 degrees of freedom in the memory matrix, but also on the retrieval criterion. Under isotropic Gaussian embeddings, we prove a sharp threshold for top-1 retrieval, where every signal must beat its largest distractor: the critical value of $d^2/(n \log n)$ is 2. Above the threshold, we explicitly construct a linear memory that retrieves all n associations with high probability; below it, no data-dependent linear memory can do so. The $\log n$ factor is therefore the unavoidable extreme-value cost of winner-take-all decoding.

Without the logarithmic factor—that is, when $n/d^2 \to \alpha \in (0, \infty)$—simultaneous top-1 retrieval is impossible. The matched target can nevertheless remain near the top of the ranking. We capture this weaker retrieval goal with the Tail-Average Margin (TAM), which, for list size k, compares each signal with the average of its k strongest competitors; a positive TAM margin certifies that the target belongs to the top-k candidate list. When $k/n \to r \in (0, 1)$, we learn the memory by empirical risk minimization with a smoothed TAM objective and derive an exact high-dimensional characterization through a two-parameter scalar variational problem. The result gives limiting laws for signal and competitor scores, margins, and percentile ranks. Sending the ridge parameter to zero after the high-dimensional limit yields a closed-form critical load $\alpha_c(r)$ separating vanishing from positive average loss.

The analysis in this work also develops a coupled leave-one-out method for matrix-valued empirical risk problems in which each sample enters many dependent comparisons, a tool that may be useful beyond associative memory.

Contents

1 Introduction **3**
1.1 Summary of main results 5
1.2 Related work 7

*Correspondence: yuelu@seas.harvard.edu.

2 Top-1 retrieval: the sharp threshold **9**
2.1 Correlation matrix memory and the logarithmic factor . . . 10
2.2 Achievability by adaptive sparse repair . . . 11
2.3 Impossibility via a dual certificate . . . 14
2.4 Finite-size scaling and a second-order conjecture . . . 21

3 Listwise retrieval and the Tail-Average Margin **22**
3.1 The Tail-Average Margin functional . . . 22
3.2 Learning under the TAM criterion . . . 24
3.3 Performance metrics . . . 25

4 Asymptotic theory of TAM **26**
4.1 Gradient and Hessian of the TAM loss . . . 27
4.2 Leave-one-out reduction . . . 28
4.2.1 LOO Hessian and rank-one reduction . . . 29
4.2.2 Proof of Proposition 3 . . . 30
4.3 Scalar characterization of the TAM optimizer . . . 32
4.4 Proof of Theorem 2 . . . 34
4.4.1 Gaussian law for competitor scores . . . 34
4.4.2 Proximal law for diagonal scores . . . 36
4.4.3 Finite-dimensional scalar equations and variational consistency . . . 37

5 Ridgeless limit and phase transition **40**
5.1 The critical load . . . 40
5.2 Score, margin, and percentile profiles . . . 43
5.2.1 The SAT phase . . . 44
5.2.2 The UNSAT phase . . . 46
5.3 Numerical validation . . . 47

6 Conclusion **50**

Disclosure of AI Assistance **50**

A Correlation matrix memory: score law and retrieval **56**
A.1 Law of one CMM score column . . . 56
A.2 CMM top-1 retrieval above constant 8 . . . 58

B Taylor expansion of the interior remainder **60**
B.1 Taylor identities . . . 60
B.2 Explicit form of the residual . . . 61
B.3 Entry-size bounds . . . 61

C Leave-one-out stability **63**
C.1 Leave-one-out expansion and Newton comparison . . . 63
C.2 Stability of the auxiliary variables . . . 71
C.3 Score stability and uniform boundedness . . . 74

D Spectral and self-averaging estimates **75**
D.1 Empirical-mean concentration . . . 76
D.2 Inverse-trace concentration . . . 79

E Bounds on the norms of $\widehat{W}$ **82**
E.1 Deterministic Frobenius upper bound . . . 82
E.2 High-probability Frobenius lower bound . . . 83
E.3 Spectral norm bound . . . 85

F A coupling lemma for bilinear Gaussian forms **88**
F.1 Deterministic matrix . . . 88
F.2 Random matrix with concentrated Frobenius norm . . . 89

G A trace estimate for the inverse Hessian **90**

H Variational formulation of the scalar equations **96**
H.1 The scalar TAM functional . . . 96
H.2 Moreau envelope identities . . . 97
H.3 Saddle-point equations . . . 97
H.4 Maximization in χ and convexity in ν . . . 98

1 Introduction

Associative memory refers broadly to the ability of a system to store and retrieve relationships between objects. Given a collection of input-output pairs, one would like the system to recover the correct target when presented with a corresponding cue. This basic mechanism appears in many forms, ranging from the first neural network models for pattern learning [Kohonen, 1972, Amari, 1972, Anderson, 1972, Hopfield, 1982, 1984], to modern key-value retrieval systems [Vaswani et al., 2017, Ramsauer et al., 2021, Gershman et al., 2025]. Associative memory has also recently attracted attention as a model of factual recall in large language models [Cabannes et al., 2024a,b, Jiang et al., 2024, Nichani et al., 2024] as trained transformer layers have been observed to function as key-value memories [Geva et al., 2021, Meng et al., 2022, Bietti et al., 2023].

A central theoretical question in associative memory is: given n associations, how large must the memory model be in order to retrieve them under a prescribed criterion? In this paper, we study this question in the simple but already nontrivial setting of *linear associative memory*. Given pairs of key vectors (embeddings) v_i and associated targets (unembeddings) u_i, we consider a matrix memory model $W \in \mathbb{R}^{d\times d}$ and define the score matrix $S = (s_{j,i})_{1\le i,j\le n}$ as

$$s_{j,i} = u_j^\top W v_i, \qquad \{(u_i, v_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^d, \qquad W \in \mathbb{R}^{d\times d}.$$

This score matrix encodes the predictions for the target when decoding to a fixed vocabulary of size n. Rather than requiring the output Wv_i of the memory model on an input v_i to equal its associated target u_i, as in classical models of associative memory [Hopfield, 1982], we compare the score of the matched target, $s_{i,i}$, with its competing scores $(s_{j,i})_{j\neq i}$. The capacity of a linear associative memory depends sharply on how retrieval is defined. We first study the strongest natural criterion, which requires the matched target to outrank every competitor.

This criterion is known as *top-1 retrieval*, or exact retrieval [Kohonen, 1972, Cabannes et al., 2024a,b, Nichani et al., 2024]. Under this criterion, each key must retrieve its matched target with the largest score:

$$s_{i,i} > s_{j,i}, \qquad \forall j \neq i, \quad \forall i \in [n]. \tag{1}$$

This is the natural criterion for exact winner-take-all decoding. It is also the hard-attention (low-temperature) limit of softmax attention mechanisms in transformer-based architectures [Vaswani et al., 2017].

To derive a sharp high-dimensional characterization of top-1 capacity, we study the isotropic Gaussian model

$$u_i, v_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d/d), \qquad i \in [n], \tag{2}$$

where all vectors in the two collections are mutually independent. This model isolates the effect of high-dimensional interference without imposing additional structure on the stored associations. As one of the main results of this work, Theorem 1 shows that, under (2), the asymptotic ratio $d^2/(n \log n)$ has the sharp critical value 2. Specifically, for every fixed $\epsilon > 0$,

$$\begin{aligned} d^2 \geq (2+\epsilon) n \log n \quad &\Longrightarrow \quad \text{top-1 retrieval is achievable with high probability,} \\ d^2 \leq (2-\epsilon) n \log n \quad &\Longrightarrow \quad \text{top-1 retrieval is impossible with high probability.} \end{aligned} \tag{3}$$

Here, achievability means that some (data-dependent) linear memory $W \in \mathbb{R}^{d \times d}$ satisfies (1), whereas impossibility means *no* linear memory can satisfy (1). The critical value 2 was conjectured in an earlier version of this paper and independently in the concurrent work of Giorlandino et al. [2026].

The logarithmic factor in (3) comes from an extreme value. To see this heuristically, fix a query i and suppose that its competing scores $\{s_{j,i}\}_{j \neq i}$ behave as independent centered Gaussians. Under the Gaussian model (2), the matched score has natural scale one, while an individual competing score has standard deviation $\sigma \approx \sqrt{n}/d$. The largest of the $n-1$ competitors is therefore about $\sigma\sqrt{2 \log n} \approx \sqrt{2n \log n}/d$. Equating this extreme competitor scale with the order-one matched score gives $\sqrt{2n \log n}/d \approx 1$, or $d^2 \approx 2n \log n$. Theorem 1 makes this prediction rigorous.

The logarithmic penalty is specific to the requirement that every target beat its largest competitor. Removing this factor brings us to the critical load $n/d^2 \to \alpha \in (0, \infty)$, where a linear memory has only a constant number of parameters per association and top-1 retrieval is impossible. Even so, the matched target need not disappear from the ranking: it may still outrank all but a fraction of its competitors. This residual ranking information is useful when a memory access is an intermediate computation whose output is refined by later layers, heads, or rerankers. We thus ask:

At critical load, how close to the top can a linear memory still place every matched target?

We call this coarser notion of retrieval *listwise retrieval*. Our second main result introduces a convex certificate for listwise retrieval and gives a sharp high-dimensional theory for learning a memory under this criterion.

For a fixed query i, order the $n-1$ competing scores as $s_{(1),i} \geq s_{(2),i} \geq \cdots \geq s_{(n-1),i}$, and, for a tail fraction $r \in (0, 1]$, let $k = \lceil r(n-1) \rceil$. Direct top-k retrieval asks that $s_{i,i} > s_{(k),i}$, so that at most $k-1$ competitors outrank the matched target. Except when $k = 1$, this rank constraint is nonconvex. Yet the average of the k largest competing scores is a convex upper bound on $s_{(k),i}$

[Rockafellar and Uryasev, 2000, 2002]. This motivates us to propose the *Tail-Average Margin* (TAM) criterion:

$$s_{i,i} > \mathrm{TAM}_r((s_{j,i})_{j\neq i}) := \frac{1}{k}\sum_{\ell=1}^{k} s_{(\ell),i}, \qquad \forall i \in [n]. \tag{4}$$

Because $\mathrm{TAM}_r((s_{j,i})_{j\neq i}) \geq s_{(k),i}$, (4) certifies that the matched target has rank at most k, and hence belongs to the top-k candidate list. The tail fraction r controls the resolution of retrieval. At the finest endpoint, $r = 1/(n-1)$, TAM is the largest competing score and the criterion recovers the top-1 condition (1); at the coarsest endpoint, $r = 1$, it is the mean competing score. For fixed $r \in (0, 1)$, TAM therefore controls the collective strength of the top r-fraction of competitors rather than a single extreme score.

This fractional-rank perspective is in the same spirit as list decoding in coding theory, where the decoder is allowed to retain a controlled list of plausible messages rather than make a unique decision [Elias, 1957, Guruswami, 2007]. It is also aligned with listwise ranking losses in information retrieval, which optimize criteria defined on ranked candidate lists rather than on a single pairwise comparison [Cao et al., 2007]. For fixed r, the certified list contains a fixed fraction of the vocabulary; the guarantee is therefore one of percentile rank rather than a sublinear shortlist.

1.1 Summary of main results

Sharp top-1 capacity. Our first main result is Theorem 1 in Section 2, which identifies the exact capacity of linear associative memory under top-1 retrieval as $d^2/(n\log n) = 2$. This critical threshold applies to the entire class of data-dependent linear memories, rather than to a particular construction or learning rule.

The achievability proof begins with a classical benchmark model, called the correlation matrix memory (CMM), which stores the associations as a superposition of outer products [Kohonen, 1972, Elhage et al., 2022]. Previous work of Nichani et al. [2024] showed that CMM succeeds when $d^2 \geq Cn(\log n)^4$ for a sufficiently large constant C. We sharpen this benchmark by proving the sufficient condition $\liminf d^2/(n\log n) > 8$ (Proposition 1). We then introduce an adaptive sparse repair, which uses CMM on a large coordinate block to store most associations, then corrects wrong (or small-margin) associations using a small reserve block of size $o(d)$. This reserve-and-repair principle is related in spirit to multiscale constructions for random perceptrons [Kim and Roche, 1998, Abbe et al., 2022, Ding and Sun, 2025]. Our construction succeeds throughout the regime $\liminf d^2/(n\log n) > 2$ (Proposition 2); see Sections 2.1–2.2 for details. Conversely, a dual certificate rules out every linear memory with high probability when $\limsup d^2/(n\log n) < 2$ (Theorem 1(ii)). Moreover, Section 2.4 presents Conjecture 1, suggested by a simpler surrogate model: a $-3n\log\log n$ second-order correction that better captures finite-size behavior.

Tail-Average Margin. Our second main result gives an exact high-dimensional theory for a convex learning rule based on the listwise criterion (4). With logistic loss $\ell(t) = \log(1+e^{-t})$, ridge parameter $\lambda \geq 0$, we consider the estimator

$$\widehat{W} \in \operatorname*{arg\,min}_{W\in\mathbb{R}^{d\times d}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell\left(s_{i,i}(W) - \mathrm{TAM}_{r,\beta}\big((s_{j,i}(W))_{j\neq i}\big)\right) + \frac{\lambda}{2d^2}\|W\|_{\mathsf{F}}^2 \right\}.$$

Here $\mathrm{TAM}_{r,\beta}$ is an exponentially smoothed version of the TAM functional TAM_r in (4), with $\beta > 0$ the smoothing parameter. For every fixed score vector x, $\mathrm{TAM}_{r,\beta}(x) \to \mathrm{TAM}_r(x)$ as $\beta \to \infty$. Its precise definition is given in Section 3.1, specifically in (17), and the corresponding convex learning formulation is developed in Section 3.2; see (19). The main goal of the TAM theory is to understand the asymptotic behavior of this ERM in the regime

$$n, d \to \infty, \qquad \frac{n}{d^2} \to \alpha \in (0, \infty), \qquad r \in (0, 1) \text{ fixed.}$$

Our main asymptotic conclusions are as follows:

1. *Scalar reduction.* The high-dimensional ERM admits a scalar variational description involving only two parameters, stated in Theorem 2 in Section 4.3;
2. *Score and margin laws.* The scalar description gives limiting predictions for the fitting loss, the laws of the true and competing scores, and the induced margin and percentile profiles, developed in Section 4 and Section 5;
3. *Ridgeless phase transition.* In the ridgeless limit, the scalar theory yields a sharp phase transition at a critical load

$$\alpha_c(r) = \frac{1}{(1+\kappa_r^2)\Phi(\kappa_r) + \kappa_r \varphi(\kappa_r)}, \qquad \kappa_r = \frac{\varphi(\Phi^{-1}(1-r))}{r},$$

where φ and Φ denote the standard normal density and distribution function. When $\alpha < \alpha_c(r)$, the limiting TAM fitting loss vanishes in the ridgeless limit; when $\alpha > \alpha_c(r)$, it remains positive. Thus $\alpha_c(r)$ is the sharp threshold for asymptotically perfect TAM recovery—vanishing average TAM loss and hence successful TAM retrieval for all but a vanishing fraction of keys.

A coupled leave-one-out method. Our third contribution is methodological: we develop a new coupled leave-one-out method for matrix-valued empirical-risk problems with shared samples. This method forms the technical core of our TAM theory. Classical leave-one-out analyses of high-dimensional M-estimators remove one independent loss term, leaving an optimizer that is nearly independent of the held-out feature [El Karoui et al., 2013, Lei et al., 2018]. That basic decoupling fails here: a stored pair supplies its own signal score, while its target also appears as a competitor for every other query. Removing one pair therefore perturbs order n loss terms and their data-dependent tail thresholds.

We develop a coupled leave-one-out expansion that tracks this global response. Its key conclusion is that, after all induced changes are accounted for, the leading perturbation of the optimizer is the inverse leave-one-out Hessian applied to the single rank-one feature $u_h v_h^\top$; the remaining terms are lower order. Iterating this expansion and combining it with new spectral and self-averaging estimates yields the Gaussian competitor law and the scalar description in Theorem 2. This method may be of independent interest for other matrix-valued or listwise ERMs in which each observation participates in many dependent comparisons rather than one separable loss term. Section 4.2 develops the coupled reduction, and the supporting estimates are proved in Appendices C–F.

1.2 Related work

Neural associative memory. Associative memory dates back to early neural models that store patterns through correlations in the weights [Amari, 1972, Anderson, 1972, Nakano, 1972, Kohonen, 1972, Hopfield, 1982, 1984]. The Hopfield network [Hopfield, 1982] is the symmetric case, where the goal is to memorize and retrieve the patterns themselves. We focus instead on associative recall: a key should retrieve its paired target, and the two need not coincide [Amari, 1972]. In particular, Kohonen [1972] define correlation matrix memory as a sum of rank-one key-target outer products, the construction analyzed in Section 2.1. Bidirectional associative memory is another paired-pattern model based on outer-product updates [Kosko, 1988].

Modern versions of neural networks with built-in associative memory capabilities include the neural Turing machine [Graves et al., 2014], memory networks [Weston et al., 2015], dense associative memory [Krotov and Hopfield, 2016, Lucibello and Mézard, 2024], MCHN [Ramsauer et al., 2021], MCSDM [Kanerva, 1988, Bricken and Pehlevan, 2021], and universal Hopfield networks [Millidge et al., 2022]. Binary variants of matrix memory have also been studied in the literature [Willshaw et al., 1969, Gardner, 1987, Hobson, 2011].

Transformers and factual recall. Transformer models trained on factual recall tasks have been empirically observed to implement associative memory through their weight matrices. Feedforward as well as attention layers have been shown to behave like key-value memories [Geva et al., 2021, Bietti et al., 2023], and the linear memory perspective has been used to study factual editing, induction head tasks, and latent space retrieval [Meng et al., 2022, Bietti et al., 2023, Jiang et al., 2024]. On the theory side, Nichani et al. [2024] study a sequence-based factual recall task where one token functions as the key, and show that a one-layer transformer solves the task with capacity scaling linearly with parameter count, up to logarithmic factors. This storage scaling has also been empirically observed in large language models [Allen-Zhu and Li, 2024]. Related theoretical analyses include optimization dynamics for a single-layer transformer on a token recall task [Vural et al., 2026], and a measure-theoretic formulation of transformers with implicit associative memory mechanisms [Kawata and Suzuki, 2026].

Storage capacity. Nichani et al. [2024] study the capacity of linear and MLP associative memories with random spherical embeddings, showing in particular that correlation matrix memory succeeds when $d^2 \gg n(\log n)^4$. Cabannes et al. [2024a] propose the exact retrieval criterion and study correlation matrix memory under a weighted data distribution. From an optimization perspective, Cabannes et al. [2024b] study gradient descent under a multiclass logistic formulation, and Kim et al. [2026] compare the capacity of memory matrices obtained by different optimizers under Gaussian embeddings. Compared with these works, our top-1 result identifies the exact statistical threshold $d^2 = 2n \log n$ for arbitrary linear memories. We also introduce the TAM criterion and, to our knowledge, give the first capacity theory for linear associative memory under a listwise retrieval rule.

We also briefly mention that storage capacity has been studied in detail for the Hopfield network and its variants. The capacity $d \asymp n$ of Hopfield networks is classical, see e.g., McEliece et al. [1987], Folli et al. [2017] and the works cited within, while modern extensions such as dense associative memory have been shown to achieve exponential capacity [Demircigil et al., 2017, Ramsauer et al., 2021, Lucibello and Mézard, 2024]. The scaling in these works differ from our

results since we decode to a fixed vocabulary of size n, while the classical Hopfield formulation decodes to the hypercube.

Concurrent work. The sharp top-1 threshold was conjectured in an earlier version of this paper and independently in the concurrent work of Giorlandino et al. [2026]. The latter studies the capacity of linear associative memory under the top-1 retrieval criterion through a *decoupled* model. In this simplified formulation, the output embeddings $\{u_j\}_{j\le n}$ are *resampled* across constraints, thus removing the statistical dependence structure among the retrieval inequalities in (1). Based on numerical evidence comparing training and test errors, and score distributions, they conjecture that the original and decoupled problems have the same critical capacity. For the decoupled model, they carry out a Gardner-type statistical physics calculation and predict the critical threshold $\frac{n\log n}{d^2}=\frac{1}{2}$ for top-1 retrieval, which we rigorously prove in Section 2. They further extend both the decoupling conjecture and the capacity calculation to rank-constrained memories, corresponding to two-layer linear MLPs.

Classification and leave-one-out methods. The TAM learning problem is also connected to high-dimensional classification through its logistic loss. The classical work of Cover [1965] computed the separability capacity of random patterns, and the perceptron capacity was later studied through statistical physics methods, including margin-constrained variants [Gardner, 1988]. In ordinary unregularized logistic regression, the existence of the maximum likelihood estimator is equivalent to nonseparability of the labeled data; equivalently, the failure of existence is a linear separability, or capacity, event. Sur and Candès [2019] established a high-dimensional phase transition for this event under Gaussian covariates. A broader line of work in statistical physics studies optimal errors and phase transitions in high-dimensional generalized linear models via the approximate message passing algorithm [Barbier et al., 2019, Aubin et al., 2020, Mignacco et al., 2020]. From a technical point of view, our TAM analysis uses a leave-one-out approach, which has played an important role in precise asymptotics for high-dimensional M-estimators [El Karoui et al., 2013, El Karoui, 2018, Lei et al., 2018]. At a conceptual level, this is also the cavity method from spin glass theory [Mézard et al., 1987].

The present problem differs from standard high-dimensional regression and classification in several ways. The unknown is a matrix rather than a vector; the effective covariates are rank-one tensors $u_j v_i^\top$ with strong dependencies; each sample contributes both a signal score and a full column of competitor scores; and the TAM threshold is endogenous, selected by the learned scores themselves. These couplings prevent existing leave-one-out theories from applying as black boxes; the resulting cavity calculation is matrix-valued, listwise, and self-consistent in a way that does not arise in standard high-dimensional regression.

Notation

We write

$$U=[u_1\ \cdots\ u_n]\in\mathbb{R}^{d\times n}\qquad\text{and}\qquad V=[v_1\ \cdots\ v_n]\in\mathbb{R}^{d\times n}$$

for the target and key matrices, respectively. Under the Gaussian model (2), $\sqrt{d}\,U$ and $\sqrt{d}\,V$ are independent standard Gaussian matrices in $\mathbb{R}^{d\times n}$. $\mathrm{vec}(\cdot)$ denotes column-wise vectorization.

We write $X \overset{(d)}{=} Y$ when two random variables, vectors, or matrices have the same distribution. For matrices, $\langle A, B\rangle$ is understood as the matrix inner product.

We use $X_n = O_{\mathbb{P}}(Y_n)$ and $X_n = o_{\mathbb{P}}(Y_n)$ in their standard meanings: X_n/Y_n is tight and converges to zero in probability, respectively. These notations record behavior at the displayed scale. In situations where we need tail control, especially uniformly over polynomial-size index sets, we use stochastic domination [Erdős and Yau, 2017]. Let $\{X_\alpha^{(n)}\}_{\alpha\in\mathcal{I}_n}$ be nonnegative random variables, where $\mathcal{I}_n$ may depend on n and has polynomial cardinality, and let $Y_n \geq 0$ be deterministic. We write

$$X_\alpha^{(n)} = \mathcal{O}_\prec(Y_n) \qquad \text{uniformly in } \alpha \in \mathcal{I}_n \tag{5}$$

if, for every $\varepsilon > 0$ and $D > 0$,

$$\mathbb{P}\left(\max_{\alpha\in\mathcal{I}_n} X_\alpha^{(n)} > n^\varepsilon Y_n\right) \leq n^{-D}$$

for all sufficiently large n. When $\mathcal{I}_n$ is a singleton, we suppress the uniformity qualifier. Thus, (5) permits an n^ε slack in exchange for n^{-D} tail control, uniformly over the indicated polynomial-size index set.

2 Top-1 retrieval: the sharp threshold

Throughout this section, we work under the Gaussian model (2) and the top-1 retrieval criterion (1). For a realization (U, V) and a matrix $W \in \mathbb{R}^{d\times d}$, define the perfect-retrieval condition

$$\mathsf{PR}(W; U, V) \quad\Longleftrightarrow\quad u_i^\top W v_i > \max_{j\neq i} u_j^\top W v_i \quad \text{for every } i \in [n].$$

We study the random feasibility event

$$\mathsf{PR}_n \coloneqq \left\{(U, V) : \text{there exists } W \in \mathbb{R}^{d\times d} \text{ such that } \mathsf{PR}(W; U, V) \text{ occurs}\right\}.$$

The matrix W in this definition may depend arbitrarily on the realized data (U, V). Accordingly, PR_n asks whether a random instance is retrievable by any linear memory, without imposing a particular learning rule.

Theorem 1 (Sharp top-1 threshold)**.** *Along any sequence $n, d \to \infty$:*

(i) *If* $\liminf \frac{d^2}{n\log n} > 2$, *then* $\mathbb{P}(\mathsf{PR}_n) \to 1$.

(ii) *If* $\limsup \frac{d^2}{n\log n} < 2$, *then* $\mathbb{P}(\mathsf{PR}_n) \to 0$.

The sharp threshold at $d^2 = 2n\log n$ was conjectured in an earlier version of this paper and independently in the concurrent work of Giorlandino et al. [2026]. We establish the achievability and impossibility directions in Section 2.2 and Section 2.3, respectively. Before turning to these proofs, Section 2.1 introduces the correlation matrix memory (CMM), a classical outer-product construction. Its analysis exposes the extreme-value origin of the logarithmic factor and motivates the sparse repair used in the achievability proof. We note that an order-wise $\Omega(n\log n)$ lower bound is also suggested in Nichani et al. [2024, Theorem 7] via an information-theoretic argument, but it does not identify the sharp constant and applies only to finite-bit learners. Finally, Section 2.4 studies finite-size scaling and proposes a second-order conjecture.

2.1 Correlation matrix memory and the logarithmic factor

The simplest linear memory stores each association by its outer product and superposes the results:

$$W_{\mathsf{CMM}} \coloneqq UV^\top = \sum_{k=1}^{n} u_k v_k^\top. \tag{6}$$

This is the classical *correlation matrix memory* (CMM) [Kohonen, 1972] and is the direct analogue of the classical self-organizing network [Amari, 1972] for asymmetric key and target pairs. When the key v_i is presented, the kth summand contributes $u_k\langle v_k, v_i\rangle$: the desired vector u_i appears with coefficient $\|v_i\|^2$, while the remaining terms create cross-talk. The rule is transparent and admits a rank-one update whenever an association is added or removed. It also illustrates *superposition*: the stored vectors need not be mutually orthogonal for retrieval to succeed [Elhage et al., 2022].

To understand the performance of this particular, data-dependent memory, write

$$\alpha_n \coloneqq \frac{n}{d^2}$$

and denote the CMM scores by $s_{j,i} = u_j^\top W_{\mathsf{CMM}} v_i$. Expanding (6) gives

$$s_{i,i} = \|u_i\|^2\|v_i\|^2 + \sum_{k\neq i}\langle u_i, u_k\rangle\langle v_k, v_i\rangle, \tag{7}$$

$$s_{j,i} = \langle u_j, u_i\rangle\|v_i\|^2 + \|u_j\|^2\langle v_j, v_i\rangle + \sum_{k\notin\{i,j\}}\langle u_j, u_k\rangle\langle v_k, v_i\rangle, \qquad j \neq i. \tag{8}$$

Under the Gaussian model, $\|u_i\|^2 = 1 + O_{\mathbb{P}}(d^{-1/2})$ and $\|v_i\|^2 = 1 + O_{\mathbb{P}}(d^{-1/2})$, so the first term in (7) is $1 + O_{\mathbb{P}}(d^{-1/2})$. Conditional on (u_i, v_i), the summands in the cross-talk term of (7) are independent and centered, each with variance $\|u_i\|^2\|v_i\|^2/d^2$. Their sum therefore has variance approximately $n/d^2 = \alpha_n$ and is approximately Gaussian by the central limit theorem.

Turning to a competitor score, the same conditional central limit argument applies to the cross-talk sum over $k \notin \{i, j\}$ in (8). The remaining $k = i$ and $k = j$ contributions, $\langle u_j, u_i\rangle\|v_i\|^2$ and $\|u_j\|^2\langle v_j, v_i\rangle$, respectively, are only $O_{\mathbb{P}}(d^{-1/2})$. At the scale $d^2 \asymp n\log n$, they and the norm fluctuations above are therefore $O_{\mathbb{P}}(d^{-1/2}) = o_{\mathbb{P}}(\sqrt{\alpha_n})$.

Combining these estimates suggests the following leading-order Gaussian picture for the scores associated with a fixed query:

$$s_{i,i} \approx 1 + \sqrt{\alpha_n}\, Z_i, \qquad s_{j,i} \approx \sqrt{\alpha_n}\, Z_{j,i}, \quad j \neq i. \tag{9}$$

Here the variables on the right are heuristically treated as independent standard Gaussians. The exact one-column representation in Lemma 3 provides a finite-(n, d) description underlying this picture. For a fixed ordered pair $i \neq j$, (9) predicts

$$s_{i,i} - s_{j,i} \approx 1 + \sqrt{2\alpha_n}\, G_{j,i}, \qquad G_{j,i} \sim \mathcal{N}(0, 1),$$

and hence, by standard Gaussian tail bounds,

$$\mathbb{P}(s_{i,i} \leq s_{j,i}) \approx \exp\left(-\frac{1}{4\alpha_n}\right).$$

Perfect retrieval requires all $n(n-1)$ such margins to be positive. A union bound therefore predicts

$$\mathbb{P}\big(W_{\mathsf{CMM}} \text{ fails perfect retrieval}\big) \lesssim n^2 \exp\left(-\frac{1}{4\alpha_n}\right) = \exp\left(2\log n - \frac{d^2}{4n}\right), \tag{10}$$

which vanishes whenever $\liminf d^2/(n\log n) > 8$. This also reveals the origin of the logarithmic factor: cross-talk has size $\sqrt{n}/d$, but the worst of quadratically many Gaussian-like margins incurs an additional $\sqrt{\log n}$ extreme-value factor.

Previously, Nichani et al. [2024, Theorem 1] proved that CMM succeeds with high probability when $d^2 \geq Cn(\log n)^4$ for a sufficiently large constant C. The calculation above sharpens this sufficient condition to the $n\log n$ scale and predicts the leading constant 8. The following proposition makes this improvement rigorous.

Proposition 1 (CMM succeeds above constant 8)**.** *Along any sequence $n, d \to \infty$,*

$$\liminf \frac{d^2}{n\log n} > 8 \quad \Longrightarrow \quad \mathbb{P}\big(\mathsf{PR}(W_{\mathsf{CMM}}; U, V)\big) \to 1.$$

Proof. See Appendix A.2. □

This sufficient constant 8 is nevertheless four times the optimal constant 2 in Theorem 1 (and one can show that this is unimprovable for CMM). Where does this gap come from? The calculation leading to (10) examines one pairwise margin at a time and then protects all $n(n-1)$ margins by a union bound. This does not require the margins to be independent, but it discards their strong dependence. For a fixed query v_i, the signal score $s_{i,i}$ is realized only once and is shared by every comparison, whereas, at leading order, the competitor scores $\{s_{j,i} : j \neq i\}$ behave as independent Gaussians. The exact distributional representation behind this claim is given in Appendix A.1. The natural query-wise object is therefore

$$s_{i,i} - \max_{j\neq i} s_{j,i} \approx 1 + \sqrt{\alpha_n} Z_i - \sqrt{2\alpha_n \log n}.$$

Here $\sqrt{2\alpha_n \log n}$ comes from the standard estimate for the maximum of n independent Gaussian variables with variance α_n. The constant 8 arises in the all-pairs argument because it repeatedly pays for fluctuations of the same signal score as though each comparison were a separate bad event; the query-wise picture pays for the maximum of the competitors only once.

If $d^2/(n\log n) > 2$ by a fixed amount, the display above predicts a positive margin for any *typical* query. Perfect retrieval, however, requires *all* n queries to succeed, and a typical-query calculation cannot rule out a small set with unusually low margins. Thus CMM correctly handles the bulk at constant 2, but by itself does not provide the required all-query guarantee. This naturally suggests a sparse correction: retain the CMM solution for most queries and modify it only to repair the exceptional ones. Section 2.2 makes this idea precise.

2.2 Achievability by adaptive sparse repair

We reserve a vanishing fraction of the coordinates for repair and use the rest to form a base CMM. Independence of the reserved block then allows us to repair only the deficient queries. Specifically, we construct a block-diagonal memory

$$W^\star := \mathrm{diag}(W^0, W^1),$$

Here $W^0 \in \mathbb{R}^{d_0 \times d_0}$ is the base memory and $W^1 \in \mathbb{R}^{d_1 \times d_1}$ is the repair; the block dimensions d_0, d_1 are specified below. Split every vector into independent coordinate blocks $u_i = (u_i^0, u_i^1)$ and $v_i = (v_i^0, v_i^1)$ of dimensions

$$d_1 = \left\lfloor d(\log n)^{-1/8} \right\rfloor, \qquad d_0 = d - d_1.$$

The large block stores all associations using the rescaled CMM

$$W^0 := \frac{d^2}{d_0^2} \sum_{k=1}^{n} u_k^0 (v_k^0)^\top, \qquad s_{j,i} := (u_j^0)^\top W^0 v_i^0.$$

Its retrieval margin for query i is

$$m_i := s_{i,i} - \max_{j \neq i} s_{j,i}.$$

For a tolerance $\gamma > 0$ (to be chosen later), call a query *deficient* when its margin is at most γ, and collect these queries in

$$\mathcal{B} := \{i : m_i \leq \gamma\}.$$

For each $i \in \mathcal{B}$, the coefficient $2\gamma - m_i$ is the amount by which we aim to raise its margin. We therefore use the reserved block to add the targeted rank-one corrections

$$W^1 := \frac{d^2}{d_1^2} \sum_{i \in \mathcal{B}} (2\gamma - m_i) u_i^1 (v_i^1)^\top.$$

Both $\mathcal{B}$ and the correction coefficients depend only on the large block and are thus independent of the coordinates used for repair.

Proposition 2 (Sparse repair succeeds above constant 2)**.** *Suppose* $\liminf_{n\to\infty} d^2/(n \log n) > 2$. *Choose constants* c *and* $\gamma \in (0, 1/2)$ *such that*

$$2 < c < \liminf_{n \to \infty} \frac{d^2}{n \log n}, \qquad c(1 - 2\gamma)^2 > 2,$$

and apply the construction above with this value of γ*. Then*

$$\mathbb{P}\big(\mathsf{PR}(W^\star; U, V)\big) \to 1.$$

Consequently, Theorem 1(i) *holds.*

Proof. Fix $i \neq j$ and set $\alpha = d_0/d \in (0, 1)$. Conditioning on $x = u_j^0$ and $y = v_i^0$, write

$$s_{j,i} = \frac{1}{\alpha^2} \left[\|x\|^2 (v_j^0)^\top y + (x^\top u_i^0)\|y\|^2 + \sum_{k \neq i,j} (x^\top u_k^0)(v_k^0)^\top y \right]$$

where all summands are conditionally independent. Then the log moment generating function of $s_{j,i}$ is

$$\log \mathbb{E}[e^{\lambda s_{j,i}} \mid x, y] = \frac{\lambda^2}{2\alpha^4 d} \|x\|^2 \|y\|^2 (\|x\|^2 + \|y\|^2) - \frac{n-2}{2} \log\left(1 - \frac{\lambda^2}{\alpha^4 d^2} \|x\|^2 \|y\|^2\right).$$

Taking $\lambda = \alpha^2 cz \log n$ and noting that $\|x\|^2, \|y\|^2 = \alpha(1+o(1))$ with probability at least $1 - n^{-C}$ for any constant $C > 0$, the first term is $o(1)$ and the second is

$$\frac{n-2}{2}\frac{\lambda^2}{\alpha^4 d^2}\|x\|^2\|y\|^2(1+o(1)) = \frac{\alpha^2 c^2 z^2 n(\log n)^2}{2d^2}(1+o(1)) < \frac{1}{2}\alpha^2 cz^2 \log n(1+o(1)).$$

By the Chernoff bound, we thus have

$$\begin{aligned}\log \Pr(s_{j,i} \geq z \mid x, y) &\leq -\lambda z + \log \mathbb{E}[e^{\lambda s_{j,i}} \mid x, y] \\ &\leq -\frac{1}{2}\alpha^2 cz^2 \log n + o(\log n),\end{aligned}$$

and so

$$\Pr(s_{j,i} \geq z) \leq n^{-\frac{1}{2}\alpha^2 cz^2 + o(1)} + n^{-C}. \tag{11}$$

Next, a straightforward computation gives

$$\mathbb{E}[s_{i,i}] = 1, \qquad \mathrm{Var}(s_{i,i}) = \frac{4}{\alpha d} + \frac{n+3}{\alpha^2 d^2} = O\left(\frac{1}{\log n}\right).$$

Writing

$$q_i := s_{i,i} - \alpha^{-2}\|u_i^0\|^2\|v_i^0\|^2,$$

a similar computation as above gives

$$\Pr(|q_i| \geq z) \leq 2n^{-\frac{1}{2}\alpha^2 cz^2 + o(1)} + n^{-C}. \tag{12}$$

Plugging $z = 2$ in (11) and $z = 1$ in (12) and choosing $C > 2$, by union bounding, it follows that

$$\min_i s_{i,i} \geq -1, \quad \max_{j \neq i} s_{j,i} \leq 2$$

with probability tending to 1.

Now recall the definition of deficient queries $\mathcal{B} = \{i : m_i \leq \gamma\}$. If $i \in \mathcal{B}$ and $s_{i,i} > 1 - \gamma$, then $\max_{j \neq i} s_{j,i} \geq 1 - 2\gamma$. Therefore

$$\begin{aligned}\Pr(i \in \mathcal{B}) &\leq \Pr(s_{i,i} \leq 1-\gamma) + \sum_{j \neq i} \Pr(s_{j,i} \geq 1 - 2\gamma) \\ &\leq \frac{\mathrm{Var}(s_{i,i})}{\gamma^2} + n^{1 - \frac{1}{2}c(1-2\gamma)^2 + o(1)} = O\left(\frac{1}{\log n}\right).\end{aligned}$$

Here, we have used $c(1-2\gamma)^2 > 2$ and Chebyshev's inequality. It follows from Markov's inequality that

$$\Pr\left(|\mathcal{B}| > \frac{n}{\sqrt{\log n}}\right) = o(1).$$

Write the repair block as

$$W^1 = (1-\alpha)^{-2} \sum_{i=1}^{n} w_i u_i^1 (v_i^1)^\top, \qquad w_i = (2\gamma - m_i) 1\{i \in \mathcal{B}\}, \quad t_{j,i} := (u_j^1)^\top W^1 v_i^1.$$

Conditional on the d_0 block and the above events, the weights w_i are fixed and satisfy

$$0 \le w_i \le 4, \qquad \|w\|_2 \le 4\sqrt{|\mathcal{B}|} \le 4n^{1/2}(\log n)^{-1/4}.$$

Then by chi-square concentration and Bernstein's inequality for subexponential random variables, it holds with probability $1 - o(1)$ that

$$\begin{aligned} \max_i |t_{i,i} - w_i| \vee \max_{j \neq i} |t_{j,i}| &= O\left(\|w\|_\infty \sqrt{\frac{\log n}{d_1}} + \frac{\|w\|_\infty \log n}{d_1} + \frac{\|w\|_2 \sqrt{\log n}}{d_1} \right) \\ &= O\left(n^{-1/4} (\log n)^{5/16} + n^{-1/2} (\log n)^{5/8} + (\log n)^{-1/8} \right). \end{aligned}$$

Choose n sufficiently large such that the right-hand side is less than $\gamma/2$. Then for every i and $j \neq i$,

$$\begin{aligned} u_i^\top W^\star v_i - u_j^\top W^\star v_i &= s_{i,i} - s_{j,i} + t_{i,i} - t_{j,i} \\ &\ge m_i + w_i - \gamma. \end{aligned}$$

If $i \notin \mathcal{B}$, then $m_i > \gamma$ and $w_i = 0$; if $i \in \mathcal{B}$, then $m_i + w_i = 2\gamma$. Thus all margins become positive under $W^\star$ and hence $\mathsf{PR}(W^\star; U, V)$ is satisfied with probability tending to one, as was to be shown. □

Remark 1 (Relation to perceptron constructions)**.** *The sparse-repair argument in our proof is related in spirit to multiscale majority constructions for random perceptrons, where successive fresh coordinate blocks are used to improve the constraints with the weakest current margins [Kim and Roche, 1998, Abbe et al., 2022]. In the related construction of Ding and Sun [2025], the deficits from a main construction are also filled using a small coordinate block. In contrast, our repair is one-shot: a single independent block targets the vanishing set of CMM-deficient queries through weighted rank-one corrections.*

2.3 Impossibility via a dual certificate

We now prove the impossibility direction of Theorem 1. The argument has a useful geometric interpretation. For each ordered comparison $i \neq j$, define the matrix

$$A_{ij} := (u_i - u_j) v_i^\top.$$

A memory W retrieves every association if and only if $\langle W, A_{ij} \rangle > 0$ for every $i \neq j$. Thus to show impossibility, it is enough to find nonnegative coefficients $\lambda = (\lambda_{ij})$, not all zero, such that

$$\sum_{i \neq j} \lambda_{ij} A_{ij} = 0.$$

Indeed, taking the Frobenius inner product with such a W would make the left-hand side both zero and strictly positive, a contradiction. Such a nontrivial nonnegative linear dependence among these comparison matrices is our *dual certificate*. The same principle underlies classical results on random half-spaces and linear separability [Wendel, 1962, Cover, 1965].

The proof constructs the certificate in two stages. First, we choose positive coefficients for which the above weighted sum is already close to zero; the threshold 2 ensures that there are enough comparisons for this approximate cancellation. Second, we show that small adjustments to these coefficients can produce any sufficiently small matrix perturbation, so that the sum can be made to be exactly zero while keeping all coefficients positive.

Proof of Theorem 1(ii). Since the threshold clearly must be monotonic in n, we may assume without loss of generality that

$$\frac{1}{2} \le \frac{d^2}{n \log n} \le c < 2.$$

We will use the index cutoff $r := \lfloor n(\log n)^{-2} \rfloor$ to construct the dual certificate. For an array $\lambda = (\lambda_{ij})_{i>r,\, j\le r}$, define the linear map

$$\mathcal{A}(\lambda) := \sum_{i=r+1}^{n} \sum_{j=1}^{r} \lambda_{ij}(u_i - u_j)v_i^\top.$$

If some $W \in \mathbb{R}^{d\times d}$ retrieves every association, it would follow that for every nonzero array $\lambda \ge 0$,

$$\langle W, \mathcal{A}(\lambda)\rangle_{\mathsf{F}} = \sum_{i=r+1}^{n} \sum_{j=1}^{r} \lambda_{ij}(u_i - u_j)^\top W v_i > 0.$$

It therefore suffices to construct an array λ with strictly positive entries such that $\mathcal{A}(\lambda) = 0$. We first construct a positive array for which $\mathcal{A}(\lambda)$ is small.

Lemma 1. *With probability tending to* 1*, there exists an array* λ *such that*

$$\min_{i>r, j\le r} \lambda_{ij} \ge \frac{1}{r \log n}, \qquad \|\mathcal{A}(\lambda)\|_{\mathsf{F}} = o\left(\frac{n}{d \log n}\right).$$

Proof. Write

$$U_1 := [u_{r+1} \ \cdots \ u_n], \qquad V_1 := [v_{r+1} \ \cdots \ v_n], \qquad M := U_1 V_1^\top, \qquad H := V_1 V_1^\top.$$

Since $n - r > d$ for sufficiently large n, H is invertible almost surely. For each $i > r$, define

$$x_i := M H^{-1} v_i, \qquad \Pi := V_1^\top H^{-1} V_1.$$

We have the identity

$$\sum_{i=r+1}^{n} x_i v_i^\top = M H^{-1} \sum_{i=r+1}^{n} v_i v_i^\top = M.$$

Also, Π is a uniformly distributed rank-d orthogonal projection in $\mathbb{R}^{n-r}$, and

$$x_i = U_1 \Pi e_{i-r}.$$

Set

$$\varepsilon_n := \sqrt{\frac{\log n}{d}} + \sqrt{\frac{d}{n}}.$$

Conditionally on V_1, the matrix U_1 is Gaussian and independent of Π. Standard concentration bounds give, with probability tending to one,

$$\max_{i>r} \left| \frac{n-r}{d} \|x_i\|^2 - 1 \right| \le C\varepsilon_n, \qquad \max_{\substack{i,k>r \\ i\neq k}} \frac{|x_i^\top x_k|}{\|x_i\|\|x_k\|} \le C\varepsilon_n, \qquad \max_{\substack{i,k>r \\ i\neq k}} |v_i^\top v_k| \le C\varepsilon_n,$$

and moreover

$$\max_{i>r} \|u_i\| \le 2, \qquad \max_{i>r} \|v_i\| \le 2, \qquad \|M\|_{\mathsf{F}} \le C\sqrt{n}, \qquad \left\| \sum_{i=r+1}^{n} v_i \right\| \le C\sqrt{n}.$$

Let $\mathcal{E}$ denote the intersection of these events, so that $\mathbb{P}(\mathcal{E}) \to 1$. We condition on this event.

Let ϕ, Φ denote the standard Gaussian density and cumulative density function, and define the (inverse) Mills ratio

$$R(z) := \frac{\phi(z)}{1 - \Phi(z)}.$$

For each $i > r$, let $z_i > 0$ be the unique solution of

$$bR(z_i) = \sqrt{d}\, \|x_i\|, \qquad b := 1 - \frac{1}{\log n},$$

and set $p_i := 1 - \Phi(z_i)$. Since $R(z) = z + O(z^{-1})$, it holds uniformly over $i > r$ that

$$\frac{z_i^2}{2\log n} = \frac{d\|x_i\|^2}{2b^2 \log n} + o(1) = \frac{1}{2(1-r/n)b^2} \frac{d^2}{n\log n} + o(1) \le \frac{c}{2} + o(1).$$

In particular, there are fixed $\theta_0, \theta < 1$ such that for all sufficiently large n, for every $i > r$,

$$\frac{z_i^2}{2\log n} < \theta_0 < \theta.$$

By Mills' bound, for sufficiently large n,

$$\frac{1}{p_i} \le C(1+z_i)e^{z_i^2/2} \le C\big(1 + \sqrt{2\log n}\big) n^{\theta_0} \le n^{\theta}.$$

Now define the array

$$\lambda_{ij} := \frac{1}{r}\left[\frac{1}{\log n} + \frac{b}{p_i} 1\left\{ \frac{\sqrt{d}\, u_j^\top x_i}{\|x_i\|} \ge z_i \right\} \right], \qquad i > r,\ j \le r.$$

We proceed to bound $\mathcal{A}(\lambda)$. Condition on $U_1, v_1, \ldots, v_n$ for the remainder of the proof. Let $u \sim \mathcal{N}(0, I_d/d)$ and define

$$\Xi(u) := \sum_{i=r+1}^{n} c_i(u)(u_i - u)v_i^\top, \qquad c_i(u) := \frac{1}{\log n} + \frac{b}{p_i} 1\left\{ \frac{\sqrt{d}\, u^\top x_i}{\|x_i\|} \geq z_i \right\}.$$

By construction,

$$\mathcal{A}(\lambda) = \frac{1}{r} \sum_{j=1}^{r} \Xi(u_j).$$

We first verify that $\Xi(u)$ is centered. Indeed,

$$\mathbb{E}[c_i(u)] = \frac{1}{\log n} + \frac{b}{p_i} \Pr(\mathcal{N}(0,1) \geq z_i) = 1$$

and

$$\mathbb{E}[u c_i(u)] = \frac{b}{p_i} \mathbb{E}\left[u 1\left\{ \frac{\sqrt{d}\, u^\top x_i}{\|x_i\|} \geq z_i \right\} \right] = \frac{b}{p_i} \frac{\phi(z_i) x_i}{\sqrt{d}\, \|x_i\|} = x_i,$$

hence $\mathbb{E}[\Xi(u)] = 0$.

Next, we bound the second moment. By direct calculation,

$$\mathbb{E}[c_i(u)^2] = 1 + b^2(p_i^{-1} - 1).$$

For $i \neq k$, it holds that

$$\left(\frac{\sqrt{d}\, u^\top x_i}{\|x_i\|}, \frac{\sqrt{d}\, u^\top x_k}{\|x_k\|} \right)^\top \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho_{ik} \\ \rho_{ik} & 1 \end{pmatrix} \right), \qquad \rho_{ik} := \frac{x_i^\top x_k}{\|x_i\| \|x_k\|}.$$

By Plackett's identity [Plackett, 1954],

$$\left| \mathbb{P}\left(\frac{\sqrt{d}\, u^\top x_i}{\|x_i\|} \geq z_i, \frac{\sqrt{d}\, u^\top x_k}{\|x_k\|} \geq z_k \right) - p_i p_k \right| \leq C |\rho_{ik}| \phi(z_i) \phi(z_k),$$

where we used $|\rho_{ik}|(z_i^2 + z_k^2) \leq C\varepsilon_n \log n = o(1)$. Hence

$$\begin{aligned} \left| \mathbb{E}[c_i(u) c_k(u)] - 1 \right| &\leq C |\rho_{ik}| \frac{\phi(z_i)\phi(z_k)}{p_i p_k} \\ &= C |\rho_{ik}| R(z_i) R(z_k) \\ &\leq C \varepsilon_n \log n. \end{aligned}$$

Expanding yields

$$\|\Xi(u)\|_{\mathsf{F}}^2 \leq 2 \left\| \sum_{i=r+1}^{n} c_i(u) u_i v_i^\top \right\|_{\mathsf{F}}^2 + 2\|u\|^2 \left\| \sum_{i=r+1}^{n} c_i(u) v_i \right\|^2,$$

where

$$\mathbb{E}\left[\left\|\sum_{i=r+1}^{n} c_i(u)u_iv_i^\top\right\|_{\mathsf{F}}^2\right] = \|M\|_{\mathsf{F}}^2 + b^2\sum_{i=r+1}^{n}(p_i^{-1}-1)\|u_i\|^2\|v_i\|^2 + \sum_{\substack{i,k>r\\ i\neq k}}\big(\mathbb{E}[c_i(u)c_k(u)]-1\big)(u_i^\top u_k)(v_i^\top v_k),$$

$$\mathbb{E}\left[\left\|\sum_{i=r+1}^{n} c_i(u)v_i\right\|^2\right] = \left\|\sum_{i=r+1}^{n} v_i\right\|^2 + b^2\sum_{i=r+1}^{n}(p_i^{-1}-1)\|v_i\|^2 + \sum_{\substack{i,k>r\\ i\neq k}}\big(\mathbb{E}[c_i(u)c_k(u)]-1\big)v_i^\top v_k.$$

On the event $\mathcal{E}$ and also $\{\|u\| \le 2\}$, using the above bounds and $\varepsilon_n^2 \log n \le C(\log n)^2/d$, we obtain

$$\mathbb{E}\left[\|\Xi(u)\|_{\mathsf{F}}^2 1\{\|u\|\le 2\}\right] \le C\left(n^{1+\theta} + n + \frac{n^2(\log n)^2}{d}\right).$$

Moreover if $\|u\| > 2$,

$$\left\|\sum_{i=r+1}^{n} c_i(u)v_i\right\| \le Cn^{1+\theta}, \qquad \mathbb{E}\left[\|u\|^2 1\{\|u\|>2\}\right] \le e^{-cd},$$

so this contribution is negligible. Hence the same bound holds in expectation. It follows that

$$\begin{aligned}\mathbb{E}[\|\mathcal{A}(\lambda)\|_{\mathsf{F}}^2] &= \frac{1}{r}\mathbb{E}[\|\Xi(u)\|_{\mathsf{F}}^2]\\ &\le C\left(n^\theta(\log n)^2 + (\log n)^2 + \frac{n(\log n)^4}{d}\right)\\ &= o\bigg(\left(\frac{n}{d\log n}\right)^2\bigg).\end{aligned}$$

By Markov's inequality conditioned on $\mathcal{E}$, we conclude with probability tending to 1 that

$$\|\mathcal{A}(\lambda)\|_{\mathsf{F}} = o\left(\frac{n}{d\log n}\right).$$

The lemma is proved. □

We next show the following uniform lower bound.

Lemma 2. *There is a constant $c_0 > 0$ such that, with probability tending to 1,*

$$\inf_{\|X\|_{\mathsf{F}}=1} \sum_{i=r+1}^{n} \sum_{j=1}^{r} |(u_i - u_j)^\top X v_i| \geq c_0 \frac{rn}{d}.$$

Moreover on this event,

$$c_0 \frac{rn}{d} \{X \in \mathbb{R}^{d\times d} : \|X\|_{\mathsf{F}} \leq 1\} \subseteq \{\mathcal{A}(\delta) : |\delta_{ij}| \leq 1\}.$$

Proof. We first prove a uniform lower bound using the first r vectors $u_1, \ldots, u_r$. By symmetrization and the contraction inequality,

$$\sup_{\|x\|^2+s^2=1} \left| \frac{1}{r} \sum_{j=1}^{r} |\sqrt{d}\, u_j^\top x - s| - \mathbb{E}_{g\sim\mathcal{N}(0,I_d)} \left[|g^\top x - s| \right] \right| = O_{\mathbb{P}}\left(\sqrt{d/r}\right) = o_{\mathbb{P}}(1).$$

Also,

$$\inf_{\|x\|^2+s^2=1} \mathbb{E}_{g\sim\mathcal{N}(0,I_d)} \left[|g^\top x - s| \right] = \inf_{a^2+s^2=1} \mathbb{E}_{g\sim\mathcal{N}(0,1)} \left[|ag - s| \right] > 0,$$

since the expectation is positive and continuous on the unit circle $a^2 + s^2 = 1$. Therefore, with probability tending to 1, simultaneously for every $x \in \mathbb{R}^d$ and $s \in \mathbb{R}$,

$$\frac{1}{r} \sum_{j=1}^{r} |\sqrt{d}\, u_j^\top x - s| \geq c_0 \sqrt{\|x\|^2 + s^2} \geq c_0 \|x\|.$$

Taking $s = \sqrt{d}\, u_i^\top x$ thus gives

$$\sum_{j=1}^{r} |(u_i - u_j)^\top x| \geq c_0 \frac{r}{\sqrt{d}} \|x\|, \qquad \forall x \in \mathbb{R}^d.$$

A similar argument for the vectors $v_{r+1}, \ldots, v_n$ also gives with probability tending to 1,

$$\sum_{i=r+1}^{n} |x^\top v_i| \geq c_0 \frac{n}{\sqrt{d}} \|x\| \qquad \forall x \in \mathbb{R}^d.$$

Now for any matrix X with $\|X\|_{\mathsf{F}} = 1$, write the spectral decomposition as

$$X^\top X = \sum_{k=1}^{d} \mu_k e_k e_k^\top, \qquad \sum_{k=1}^{d} \mu_k = 1.$$

For every $i > r$, by Cauchy–Schwarz,

$$\|X v_i\| = \left(\sum_{k=1}^{d} \mu_k (e_k^\top v_i)^2 \right)^{1/2} \geq \sum_{k=1}^{d} \mu_k |e_k^\top v_i|.$$

Summing over i, from the preceding bounds, we have that

$$\sum_{i=r+1}^{n}\sum_{j=1}^{r}|(u_i-u_j)^\top Xv_i| \geq c_0\frac{r}{\sqrt{d}}\sum_{i=r+1}^{n}\|Xv_i\| \\ \geq c_0\frac{r}{\sqrt{d}}\sum_{k=1}^{d}\mu_k\sum_{i=r+1}^{n}|e_k^\top v_i| \geq c_0\frac{rn}{d}.$$

This proves the first claim. For the second claim, define the set

$$\mathcal{K} := \{\mathcal{A}(\delta) : |\delta_{ij}| \leq 1\}.$$

$\mathcal{K}$ is convex and compact, and it holds that

$$\sup_{Y\in\mathcal{K}}\langle X, Y\rangle_{\mathsf{F}} = \sup_{|\delta_{ij}|\leq 1}\sum_{i=r+1}^{n}\sum_{j=1}^{r}\delta_{ij}(u_i-u_j)^\top Xv_i \\ = \sum_{i=r+1}^{n}\sum_{j=1}^{r}|(u_i-u_j)^\top Xv_i| \geq c_0\frac{rn}{d}\|X\|_{\mathsf{F}}\,.$$

If there exists a matrix $Z \notin \mathcal{K}$ such that $\|Z\|_{\mathsf{F}} \leq c_0 rn/d$, by the separating hyperplane theorem, there must exist X such that

$$\langle X, Z\rangle_{\mathsf{F}} > \sup_{Y\in\mathcal{K}}\langle X, Y\rangle_{\mathsf{F}} \geq \|X\|_{\mathsf{F}}\|Z\|_{\mathsf{F}}\,,$$

a contradiction. □

Now from Lemma 1, with probability tending to 1, there exists an array λ satisfying

$$\lambda_{ij} \geq \frac{1}{r\log n}, \qquad \|\mathcal{A}(\lambda)\|_{\mathsf{F}} = o\left(\frac{n}{d\log n}\right),$$

while Lemma 2 implies that, again with probability tending to 1,

$$\frac{c_0 n}{2d\log n}\{X : \|X\|_{\mathsf{F}} \leq 1\} \subseteq \left\{\mathcal{A}(\delta) : |\delta_{ij}| \leq \frac{1}{2r\log n}\right\}.$$

It follows that there exists an array δ satisfying

$$|\delta_{ij}| \leq \frac{1}{2r\log n}, \qquad \mathcal{A}(\delta) = -\mathcal{A}(\lambda).$$

Then $\mathcal{A}(\lambda+\delta) = 0$ by linearity and

$$\lambda_{ij}+\delta_{ij} \geq \frac{1}{r\log n} - \frac{1}{2r\log n} = \frac{1}{2r\log n} > 0.$$

This completes the proof. □

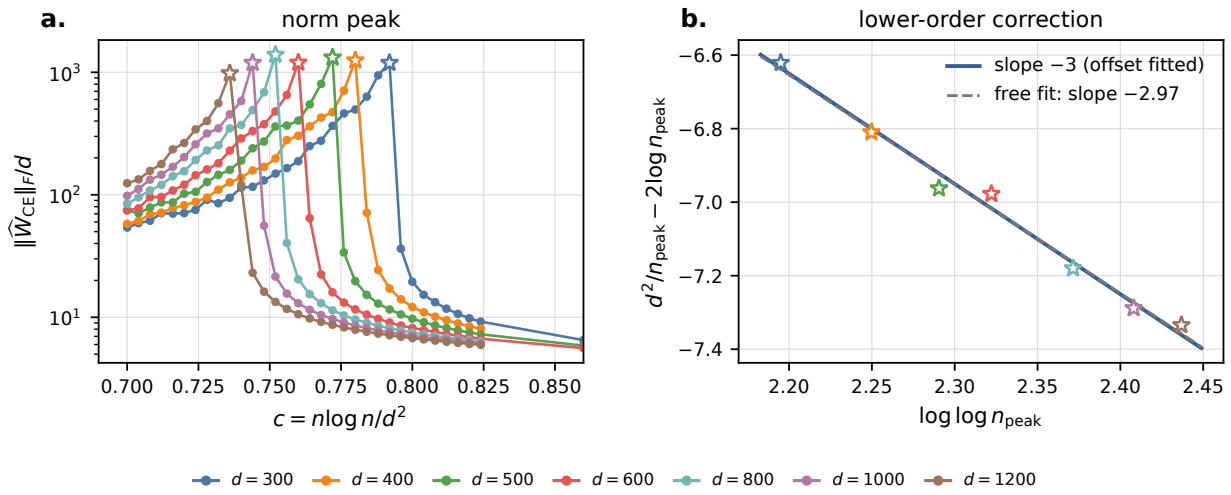

Figure 1: Finite-size transition diagnostic. **a.** Norm scale $\|\widehat{W}_{\mathrm{CE}}\|_{\mathsf{F}}/d$ along the cross-entropy sweep (13); stars mark the largest value on the sampled grid for each dimension. **b.** The same seven grid-selected peaks in the linearized coordinates used below; the solid and dashed lines are the fixed-slope and unconstrained fits, respectively.

2.4 Finite-size scaling and a second-order conjecture

We conclude this section with a numerical study of the top-1 transition. Our goal is to compare its location at moderate dimension with the leading-order threshold in Theorem 1. In our experiments, we learn the memory matrix using ridge-regularized multiclass softmax cross-entropy,

$$\widehat{W}_{\mathrm{CE}} := \operatorname*{arg\,min}_{W\in\mathbb{R}^{d\times d}} \left\{ \sum_{i=1}^{n} \left[-s_{i,i} + \log \sum_{j=1}^{n} e^{s_{j,i}} \right] + \frac{\lambda n}{2d^2}\|W\|_{\mathsf{F}}^2 \right\}, \qquad \lambda > 0. \tag{13}$$

This objective provides a constructive search for a retrieving memory. Indeed, if W has minimum retrieval margin $\gamma = \min_i \min_{j\neq i}(s_{i,i} - s_{j,i}) > 0$, then the unregularized cross-entropy at tW is at most $n(n-1)e^{-t\gamma}$ and hence tends to zero as $t \to \infty$. Consequently, as the effective ridge coefficient tends to zero, the minimizer in (13) retrieves whenever a strict retrieving memory exists: one can let $t \to \infty$ slowly enough that both the cross-entropy and ridge penalty vanish, whereas any nonretrieving matrix incurs cross-entropy at least $\log 2$.

On the feasible side, the norm needed to attain a fixed small loss grows as the best retrieval margin shrinks. We therefore use the observed peak of $\|\widehat{W}_{\mathrm{CE}}\|_{\mathsf{F}}/d$ under weak ridge regularization as an empirical proxy for the transition. For each dimension, we sweep the load $c = n\log n/d^2$, using the effective ridge coefficient $\lambda n/d^2 = 10^{-8}$. The results are shown in Figure 1.

Because c is the reciprocal of the ratio $d^2/(n\log n)$ used in Theorem 1, the critical value is $1/2$ in the present normalization rather than 2. Panel **a** shows a pronounced peak at every dimension. Denoting its location by $\widehat{c}(d)$, we find that $\widehat{c}(d)$ lies above $1/2$ and drifts only slowly toward it as d increases. Thus while Theorem 1 captures the leading location of the transition, the simulations exhibit a substantial finite-size correction. This motivates the following conjecture for the second-order correction to the threshold.

Conjecture 1 (Second-order correction to the top-1 threshold)**.** *Let* $d_c(n)$ *be the smallest integer* d *for which* $\mathbb{P}(\mathsf{PR}_n) \geq 1/2$*. Then*

$$d_c^2(n) = n\big(2\log n - 3\log\log n + o(\log\log n)\big). \tag{14}$$

Conjecture 1 is suggested by a simpler model. In the original problem, the n^2 score directions $\mathrm{vec}(u_j v_i^\top) = v_i \otimes u_j$ have a correlated Kronecker structure. Suppose instead that we replace them by n^2 independent isotropic Gaussian vectors in $\mathbb{R}^{d^2}$, while leaving the top-1 condition (1) unchanged. The score map then has independent Gaussian rows, which allows us to apply the standard conic kinematic formula [Amelunxen et al., 2014]. The resulting statistical-dimension calculation gives (14). We omit this calculation, since it does not apply directly to the original model. Proving the conjecture for linear associative memory is left to future work.

The conjecture also appears to be supported by the simulations. Conjecture 1 predicts the threshold to satisfy

$$\frac{d_c^2(n)}{n} - 2\log n = -3\log\log n + o(\log\log n).$$

Panel **b** plots the numerical peak locations according to the above formula. The fitted slope -2.97 gives an excellent fit with the predicted slope -3. In particular, (14) predicts the critical load c to converge to $1/2$ at a rate of $O(\frac{\log\log n}{\log n})$, explaining the slow convergence.

3 Listwise retrieval and the Tail-Average Margin

Top-1 retrieval has the finest possible resolution: each query must rank its matched target first. This is natural for final decoding. At an intermediate stage, however, it may be enough to keep the target among the strongest candidates and let later computation refine the choice. This motivates the coarser, listwise notion studied here.

A direct listwise requirement would ask the signal score to beat a high competitor quantile, thereby ensuring that the correct target belongs to a controlled candidate set. This rank condition is natural, but it is non-convex and combinatorial. We therefore use a convex certificate: we ask the signal score to beat the *average* of the strongest competitors. We call the resulting functional the *Tail-Average Margin* (TAM). TAM is a listwise retrieval criterion in its own right: it controls the mass of strong distractors rather than only the single largest one, while retaining enough convex structure for optimization and asymptotic analysis.

In this section we define the TAM functional, record its basic properties, and formulate the corresponding learning problem and performance criteria for linear associative memory. In the next section, we derive an asymptotic theory of TAM learning in the quadratic regime $d^2 \asymp n$, with a fixed tail fraction $r \in (0, 1)$.

3.1 The Tail-Average Margin functional

Let $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and fix a tail fraction $r \in (0, 1]$. Set $k := \lceil rm \rceil$.

Definition 1 (Tail-Average Margin)**.** *The* Tail-Average Margin *(TAM) at level* r *is*

$$\mathrm{TAM}_r(x) := \frac{1}{k}\sum_{j=1}^{k} x_{(j)}, \tag{15}$$

where $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(m)}$ are the order statistics of x.

We note two boundary cases: $\mathrm{TAM}_{1/m}(x) = \max_j x_j$ (take $k = 1$), and $\mathrm{TAM}_1(x) = \frac{1}{m}\sum_j x_j$ (take $k = m$). For fixed $r \in (0, 1)$, TAM then interpolates strictly between the maximum and the mean by averaging over the top r-fraction of entries.

The TAM criterion is a conservative certificate for top-k list retrieval. Indeed, since

$$\mathrm{TAM}_r(x) = \frac{1}{k}\sum_{j=1}^{k} x_{(j)} \geq x_{(k)},$$

the inequality $s_{i,i} > \mathrm{TAM}_r((s_{j,i})_{j\neq i})$ implies $s_{i,i} > s_{(k),i}$ for the ordered competitor scores in column i. Hence at most $k-1$ competitors can outrank the signal. In this sense, exact top-1 retrieval implies TAM retrieval, and TAM retrieval implies inclusion of the correct item in the top-k candidate list.

The TAM functional is essentially a reparametrization of the empirical *Conditional Value-at-Risk* (CVaR) studied extensively in the quantitative risk literature; see e.g., Rockafellar and Uryasev [2000]. A classical observation is that TAM admits a convex variational representation,

$$\mathrm{TAM}_r(x) \;=\; \min_{\mu\in\mathbb{R}} \left\{ \mu + \frac{1}{k}\sum_{j=1}^{m} (x_j - \mu)_+ \right\}, \tag{16}$$

where $(x)_+ = \max\{x, 0\}$ is the hinge function. The minimizer $\mu^\star = x_{(k)}$ of (16) is the empirical $(1-r)$-quantile of x, also known as the *Value-at-Risk*. The optimal value of (16) is then the mean of the exceedances of x above this quantile, which recovers the tail-average interpretation of (15). As an immediate consequence of (16), we see that $x \mapsto \mathrm{TAM}_r(x)$ is convex on $\mathbb{R}^m$, as it is the pointwise minimum over μ of functions jointly convex in (x, μ).

For optimization it is convenient to replace the piecewise-linear hinge $(\cdot)_+$ in (16) with a differentiable surrogate. For a smoothing parameter $\beta > 0$, let

$$\phi_\beta(t) \;:=\; \frac{1}{\beta}\log(1 + e^{\beta t})$$

denote the softplus function, and define the *smoothed* TAM

$$\mathrm{TAM}_{r,\beta}(x) \;:=\; \min_{\mu\in\mathbb{R}} \left\{ \mu + \frac{1}{k}\sum_{j=1}^{m} \phi_\beta(x_j - \mu) \right\}. \tag{17}$$

Since ϕ_β is convex, $\mathrm{TAM}_{r,\beta}$ is convex in x; moreover $\phi_\beta(t) \to (t)_+$ pointwise as $\beta \to \infty$, and hence $\mathrm{TAM}_{r,\beta}(x) \to \mathrm{TAM}_r(x)$ in the same limit. The exact TAM is recovered in the limiting notation $\mathrm{TAM}_{r,\infty} = \mathrm{TAM}_r$.

Throughout the remainder of the paper we specialize TAM to the linear associative memory setting: the collection $x_1, \ldots, x_m$ is taken to be the competing scores $(s_{j,i})_{j\neq i}$ associated with a query v_i, so that $m = n - 1$ and TAM_r quantifies the strength of the top r-fraction of competitors. The next subsection turns this into a convex learning objective.

3.2 Learning under the TAM criterion

With the TAM functional in hand, we now formulate retrieval and learning for the linear associative memory as a convex optimization problem.

Given stored pairs $\{(u_i, v_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^d$ and a memory or weight matrix $W \in \mathbb{R}^{d\times d}$, recall that the score matrix is $s_{j,i} = u_j^\top W v_i$. Fix a tail fraction $r \in (0,1)$ and set $k = \lceil r(n-1) \rceil$. The *TAM retrieval criterion* asks that each signal score exceed the TAM-aggregate of its competitors:

$$s_{i,i} \;>\; \mathrm{TAM}_r((s_{j,i})_{j\neq i}), \qquad \forall\, i \in [n]. \tag{18}$$

For $r = 1/(n-1)$, (18) reduces to the exact top-1 criterion (1) studied in Section 2; for fixed $r \in (0,1)$, it imposes an upper-tail margin requirement that controls the mass of strong competitors rather than a single extreme-value outlier.

Rather than demanding strict satisfaction of (18), we minimize a convex surrogate that penalizes violations softly. Let

$$\ell(t) \;:=\; \log(1+e^{-t})$$

denote the logistic loss, and let $\beta > 0$ be the smoothing parameter of $\mathrm{TAM}_{r,\beta}$. We consider the regularized objective

$$\min_{W\in\mathbb{R}^{d\times d}} \left\{ \sum_{i=1}^n \ell\big(s_{i,i} - \mathrm{TAM}_{r,\beta}((s_{j,i})_{j\neq i})\big) + \frac{\lambda n}{2d^2}\|W\|_{\mathsf{F}}^2 \right\}, \tag{19}$$

where $\lambda \geq 0$ is a ridge penalty. Since each score $s_{j,i}$ is linear in W and $\mathrm{TAM}_{r,\beta}$ is convex, the margin $s_{i,i} - \mathrm{TAM}_{r,\beta}((s_{j,i})_{j\neq i})$ is concave in W; composing with the decreasing convex ℓ yields a convex function of W, so (19) is a convex program (strongly convex whenever $\lambda > 0$).

The variational representation (17) of $\mathrm{TAM}_{r,\beta}$ suggests lifting (19) to a joint optimization over W and the auxiliary quantile variables $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$. Substituting (17) into (19) yields

$$\min_{W,\mu} \left\{ \sum_{i=1}^n \ell\left(s_{i,i} - \mu_i - \frac{1}{k}\sum_{j\neq i} \phi_\beta(s_{j,i} - \mu_i) \right) + \frac{\lambda n}{2d^2}\|W\|_{\mathsf{F}}^2 \right\}, \tag{20}$$

Minimizing (20) over μ at fixed W recovers (19). This is the workhorse formulation for the leave-one-out analysis of Section 4.

It is useful to compare (19) with the ridge-regularized multiclass softmax cross-entropy (13), the canonical convex surrogate for the top-1 criterion (1) introduced in Section 2.4. Denote the log-sum-exp aggregate of the competitor scores by $\mathrm{LSE}((s_{j,i})_{j\neq i}) := \log \sum_{j\neq i} e^{s_{j,i}}$. A short calculation rewrites each summand in (13) as

$$-\,s_{i,i} + \log \sum_{j=1}^n e^{s_{j,i}} \;=\; \ell\Big(s_{i,i} - \mathrm{LSE}\big((s_{j,i})_{j\neq i}\big)\Big),$$

with the same logistic loss ℓ used in (19). In this form, (13) has precisely the same structure as the TAM objective (19): a logistic loss applied to the margin between the signal score $s_{i,i}$ and a convex aggregate of the competitor scores $(s_{j,i})_{j\neq i}$. They differ only in the choice of aggregate. The log-sum-exp is a smoothed maximum, with $\mathrm{LSE}((s_{j,i})_{j\neq i}) = \max_{j\neq i} s_{j,i} + O(\log n)$, and therefore inherits the $\sqrt{2\log n}$ extreme-value statistics of the top-1 criterion. Replacing it with the tail average $\mathrm{TAM}_{r,\beta}$, which remains on the scale of a typical competitor at fixed r, enables retrieval at the quadratic model-size scale $d^2 \asymp n$ instead of $d^2 \asymp n\log n$.

3.3 Performance metrics

With the TAM learning problem in place, we consider three asymptotic quantities that together characterize retrieval performance: the empirical average fitting loss, the per-sample TAM margin, and the distribution of signal percentiles. Throughout this subsection, $(\widehat{W}, \widehat{\mu})$ denotes the unique minimizer of (20), with corresponding scores $\widehat{s}_{j,i} = u_j^\top \widehat{W} v_i$.

Empirical average fitting loss. Let

$$L_n(W, \mu) \;:=\; \frac{1}{n} \sum_{i=1}^{n} \ell \left(s_{i,i} - \mu_i - \frac{1}{k} \sum_{j \neq i} \phi_\beta(s_{j,i} - \mu_i) \right)$$

denote the normalized data-fitting part of (20), and write $\widehat{L}_n := L_n(\widehat{W}, \widehat{\mu})$ for its value at the optimizer. We will say that the TAM problem is *asymptotically fit at load* α if $\widehat{L}_n \to 0$ in probability as $n, d \to \infty$ with $n/d^2 \to \alpha$ and $\lambda \to 0$. Since ℓ is nonnegative and vanishes only in the large-margin limit, asymptotic fit is equivalent to the TAM retrieval criterion (18) holding with a positive margin for all but a vanishing fraction of keys.

TAM margin distribution. A finer summary than $\widehat{L}_n$ is the empirical distribution of the per-sample TAM margins

$$\widehat{m}_i \;:=\; \widehat{s}_{i,i} - \widehat{\mu}_i - \frac{1}{k} \sum_{j \neq i} \phi_\beta(\widehat{s}_{j,i} - \widehat{\mu}_i), \qquad i \in [n],$$

which are precisely the arguments of the logistic loss appearing in $\widehat{L}_n$. A positively-supported limiting distribution for $\widehat{m}_i$ indicates asymptotic fit; its shape shows how comfortably the signal beats the tail average.

Signal percentile distribution. As a complementary, rank-based summary, we also ask, for each sample, what fraction of competitors have score below the signal. Formally, define the *signal percentile*

$$\Omega_i \;:=\; \frac{1}{n-1} \sum_{j \neq i} \mathbf{1}\{\widehat{s}_{j,i} \leq \widehat{s}_{i,i}\}, \qquad i \in [n],$$

so that Ω_i close to 1 signals that the ith signal dominates nearly all of its competitors and $\Omega_i = 0$ that every competitor beats it. The random probability measure

$$\nu_n \;:=\; \frac{1}{n} \sum_{i=1}^{n} \delta_{\Omega_i}$$

on $[0, 1]$ is our object of study: when does ν_n converge in probability (say, in Wasserstein-1) to a deterministic limit?

CMM benchmark: from top-1 to percentiles. Section 2.1 used a Gaussian approximation to the scores of the non-optimized CMM W_{CMM} to explain its top-1 behavior. Here we return to the same picture, now in the proportional regime $n/d^2 \to \alpha \in (0, \infty)$, to predict the full distribution of signal percentiles. In particular, (9) suggests, for a fixed query,

$$s_{i,i} \approx 1 + \sqrt{\alpha}\, Z_i, \qquad s_{j,i} \approx \sqrt{\alpha}\, Z_{j,i}, \quad j \neq i,$$

where Z_i and $(Z_{j,i})_{j \neq i}$ are treated as independent standard Gaussians. As discussed in Section 2.1, Lemma 3 gives the exact one-column representation underlying this approximation: the competitor vector is a random radius times $g/\|g\|$, and $\sqrt{n}\, g/\|g\|$ is approximately Gaussian coordinatewise, while the radius concentrates at $\sqrt{\alpha}$.

Conditioning on Z_i and applying a law-of-large-numbers heuristic to the competitors gives

$$\Omega_i \approx \frac{1}{n-1} \sum_{j \neq i} \mathbf{1}\Big\{ Z_{j,i} \leq Z_i + \alpha^{-1/2} \Big\} \approx \Phi\Big(Z_i + \alpha^{-1/2} \Big).$$

Consequently, the predicted CDF of the CMM percentile is

$$F_\alpha(\omega) \;:=\; \Phi\Big(\Phi^{-1}(\omega) - \alpha^{-1/2} \Big), \qquad \omega \in (0,1), \tag{21}$$

with density $f_\alpha(\omega) = \exp(\Phi^{-1}(\omega)/\sqrt{\alpha} - 1/(2\alpha))$. The parameter $\alpha^{-1/2}$ plays the role of a signal-to-noise ratio: the law concentrates at 1 as $\alpha \to 0$ and tends to $\mathrm{Unif}(0,1)$ as $\alpha \to \infty$.

The next section develops a leave-one-out theory for these three observables under the TAM optimizer in the quadratic regime $n/d^2 \to \alpha$ with fixed $r \in (0,1)$.

4 Asymptotic theory of TAM

We now study the TAM optimizer in the quadratic regime $n, d \to \infty$ with $n/d^2 \to \alpha$ and $k/n \to r \in (0,1)$. Throughout this section the smoothing parameter $\beta < \infty$ and ridge parameter $\lambda > 0$ are fixed. The goal is to predict the empirical laws of the true scores $s_{i,i}(\widehat{W})$ and competitor scores $s_{j,i}(\widehat{W})$, and hence the limiting TAM loss, margin distribution, and signal-percentile profile introduced in Section 3.3.

The main technique is a leave-one-out (LOO) analysis. We compare the TAM optimizer with the optimizer obtained after removing a held-out pair (u_h, v_h), and use this comparison to separate the fresh randomness of the held-out sample from the remaining problem. The first three inputs below are LOO stability estimates, proved in Appendix C. The remaining three are spectral and self-averaging estimates: the spectral estimate is proved in Appendix E, while the two self-averaging estimates are proved in Appendix D. Together, the six estimates yield the two-parameter scalar variational description stated in Theorem 2.

The section is organized as follows. We first compute the gradient and Hessian of the smoothed TAM loss. We then formulate the LOO rank-one reduction, state the scalar variational prediction, and finally derive the finite-dimensional equations that lead to this prediction. The ridgeless limit $\lambda \downarrow 0$ and the associated phase transition are deferred to Section 5.

4.1 Gradient and Hessian of the TAM loss

Throughout this section the threshold variables are unregularized, and we work with the joint TAM objective of (20), written with the reparameterized ridge:

$$\Psi(W,\mu) = \sum_{i=1}^{n} \ell\Big(s_{i,i}(W) - \mu_i - \frac{1}{k}\sum_{j\neq i}\phi_\beta\big(s_{j,i}(W)-\mu_i\big)\Big) + \frac{\lambda n}{2d^2}\,\|W\|_{\mathsf{F}}^2\,,$$

where $s(W) = U^\top W V \in \mathbb{R}^{n\times n}$, with entries $s_{j,i}(W) = u_j^\top W v_i$, and $U = [u_1 \ \cdots \ u_n]$, $V = [v_1 \ \cdots \ v_n]$ are the data matrices from the Gaussian model (2). We use lowercase $s_{j,i}$ for finite-dimensional scores, reserving uppercase S for scalar limiting random variables. When the underlying matrix is clear from context, we write $s_{j,i}$ instead of $s_{j,i}(W)$; in particular, at the full optimizer this means $s_{j,i} = s_{j,i}(\widehat{W})$. Since Ψ depends on W only through $s(W)$, the chain rule reduces the computation of $\nabla_W\Psi$ to an $n\times n$ matrix of score derivatives,

$$\nabla_W\Psi \;=\; U\,G\,V^\top \;+\; \frac{\lambda n}{d^2}\,W, \qquad G_{j,i} := \frac{\partial\Psi}{\partial s_{j,i}}. \tag{22}$$

The remainder of this subsection evaluates the entries of G and rewrites $\nabla_W\Psi$ in a form convenient for the leave-one-out analysis in Section 4.2.

The i-th summand of Ψ equals $\ell(m_i)$, where the *TAM aggregate* and *margin* associated with sample i are

$$c_i := \mu_i + \frac{1}{k}\sum_{j\neq i}\phi_\beta\big(s_{j,i}-\mu_i\big), \qquad m_i := s_{i,i} - c_i.$$

The margin m_i is the only part of the i-th summand that depends on column i of the score matrix, and the i-th summand is the only summand that does. Differentiating $\ell(m_i)$ in $s_{j,i}$ therefore yields

$$G_{j,i} \;=\; \ell'(m_i)\,\frac{\partial m_i}{\partial s_{j,i}} \;=\; \begin{cases} \ell'(m_i), & j = i,\\ -\frac{1}{k}\,\ell'(m_i)\,\phi_\beta'(s_{j,i}-\mu_i), & j\neq i. \end{cases} \tag{23}$$

Both ℓ' and ϕ_β' are conveniently expressed through the sigmoid $\sigma(t) := (1+e^{-t})^{-1}$:

$$\ell'(t) \;=\; \sigma(t) - 1 \;=\; -\,\sigma(-t), \qquad \phi_\beta'(t) \;=\; \sigma(\beta t).$$

Substituting into (23) gives the two scalar weights

$$a_i \;:=\; \sigma(-m_i) \;=\; -\,\ell'(m_i) \;\in\; (0,1),$$

and, for $j\neq i$,

$$q_{j,i} \;:=\; \sigma\big(\beta\,(s_{j,i}-\mu_i)\big) \;=\; \phi_\beta'(s_{j,i}-\mu_i) \;\in\; (0,1), \qquad j\neq i.$$

In these variables (23) becomes $G_{j,i} = -a_i$ for $j=i$ and $G_{j,i} = \frac{1}{k}q_{j,i}a_i$ otherwise. Equivalently $G = Q\Lambda_a$, where $\Lambda_a := \mathrm{diag}(a_1,\dots,a_n)$ and

$$Q_{j,i} \;:=\; \begin{cases} -1, & j=i,\\ \dfrac{1}{k}\,q_{j,i}, & j\neq i. \end{cases}$$

Substituting $G = Q\Lambda_a$ into (22) and performing the analogous calculation for μ — where $\partial m_i/\partial \mu_i = -1 + \frac{1}{k}\sum_{j\neq i} q_{j,i} = -(Q^\top \mathbf{1}_n)_i$ — yields

$$\nabla_W \Psi(W,\mu) = U\, Q\Lambda_a\, V^\top \;+\; \frac{\lambda n}{d^2}\, W, \tag{24}$$

$$\nabla_\mu \Psi(W,\mu) = -\Lambda_a\, Q^\top \mathbf{1}_n,$$

where $\mathbf{1}_n \in \mathbb{R}^n$ is the all-ones vector. Expanding (24) by rows of Q separates the diagonal and off-diagonal contributions,

$$\nabla_W \Psi \;=\; -\sum_{i=1}^{n} a_i\, u_i v_i^\top \;+\; \frac{1}{k}\sum_{i=1}^{n} a_i \sum_{j\neq i} q_{j,i}\, u_j v_i^\top \;+\; \frac{\lambda n}{d^2}\, W. \tag{25}$$

For the Taylor expansion in Section 4.2 we will also need the second derivative of the outer loss at the margin, which is governed by a companion diagonal:

$$b_i \;:=\; \ell''(m_i) \;=\; a_i(1-a_i), \qquad \Lambda_b := \operatorname{diag}(b_1,\dots,b_n). \tag{26}$$

Definition 2 (Full Hessian)**.** *Let $H : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ denote the Hessian operator of Ψ in the W variable, evaluated at a fixed point (W,μ). Applied to a perturbation Δ, it is*

$$H[\Delta] \;:=\; \frac{\lambda n}{d^2}\,\Delta \;+\; U\, M\big(U^\top \Delta V\big)\, V^\top.$$

Here $M : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ acts columnwise,

$$M(X)_{:,i} \;=\; M_i X_{:,i},$$

where

$$M_i \;:=\; b_i Q_{:,i} Q_{:,i}^\top \;+\; a_i \operatorname{diag}(\widetilde{R}_{:,i}),$$

$$\widetilde{R}_{j,i} \;:=\; \begin{cases} \dfrac{\beta}{k}\, q_{j,i}(1-q_{j,i}), & j \neq i, \\ 0, & j = i. \end{cases}$$

The operator H is self-adjoint and strictly positive-definite, with smallest eigenvalue at least $\lambda n/d^2$. Under vectorization, writing $\mathcal{D} := \operatorname{diag}(M_1,\dots,M_n) \in \mathbb{R}^{n^2\times n^2}$, the same operator has matrix representation

$$H \;=\; \frac{\lambda n}{d^2}\, I_{d^2} \;+\; (V\otimes U)\,\mathcal{D}\,(V\otimes U)^\top.$$

4.2 Leave-one-out reduction

We relate the joint minimizer $(\widehat{W},\widehat{\mu})$ of (20) to a leave-one-out reference point by a first-order expansion. Fix a held-out index $h \in [n]$, and let $(\widehat{W}^{\setminus h},\widehat{\mu}^{\setminus h})$ denote the minimizer of the TAM objective with sample h removed from both the outer sum and from all inner summations over interferers:

$$\Psi^{\setminus h}(W,\mu) \;:=\; \sum_{i\neq h} \ell\Big(s_{i,i}(W) - \mu_i - \frac{1}{k}\sum_{\substack{j\neq i\\ j\neq h}} \phi_\beta\big(s_{j,i}(W)-\mu_i\big)\Big) + \frac{\lambda n}{2d^2}\,\|W\|_{\mathsf{F}}^2\,.$$

Write $\Delta W := \widehat{W} - \widehat{W}^{\setminus h}$, $\Delta\mu := \widehat{\mu} - \widehat{\mu}^{\setminus h}$, and $\Delta S := U^\top(\Delta W)V \in \mathbb{R}^{n\times n}$ for the corresponding perturbations. A hat decoration (e.g., $\hat{a}_i, \hat{q}_{j,i}, \hat{b}_i, \hat{m}_i$) denotes a quantity from Section 4.1 evaluated at $(\widehat{W}, \widehat{\mu})$; a superscript $\setminus h$ denotes the same quantity evaluated at $(\widehat{W}^{\setminus h}, \widehat{\mu}^{\setminus h})$. We adopt the conventions $a_h^{\setminus h} := 0$ and $q_{j,h}^{\setminus h} := q_{h,i}^{\setminus h} := 0$, so that $Q^{\setminus h}, \Lambda_{a^{\setminus h}}$ extend naturally to $n \times n$ arrays.

4.2.1 LOO Hessian and rank-one reduction

The LOO Hessian has the same form as the full Hessian in Definition 2, with the h-th sample removed and all coefficients evaluated at $(\widehat{W}^{\setminus h}, \widehat{\mu}^{\setminus h})$. Explicitly, $H^{\setminus h} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is defined by

$$H^{\setminus h}[\Delta] \;:=\; \frac{\lambda n}{d^2}\,\Delta \;+\; U\,M^{\setminus h}\big(U^\top \Delta\, V\big)\,V^\top, \tag{27}$$

where $M^{\setminus h} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ acts columnwise as

$$M^{\setminus h}(X)_{:,i} \;=\; M_i^{\setminus h}\,X_{:,i} \quad (i \neq h), \qquad M^{\setminus h}(X)_{:,h} \;=\; 0. \tag{28}$$

In the above display, $M_i^{\setminus h} \in \mathbb{R}^{n\times n}$ is a symmetric matrix with a rank-one-plus-diagonal form

$$\begin{aligned} M_i^{\setminus h} \;&:=\; b_i^{\setminus h}\,Q_{:,i}^{\setminus h}(Q_{:,i}^{\setminus h})^\top \;+\; a_i^{\setminus h}\,\operatorname{diag}\big(\widetilde{R}_{:,i}^{\setminus h}\big) \\ \widetilde{R}_{j,i}^{\setminus h} \;&:=\; \begin{cases} \frac{\beta}{k}\,q_{j,i}^{\setminus h}\big(1 - q_{j,i}^{\setminus h}\big) & j \neq i,\; j \neq h \\ 0, & \text{otherwise} \end{cases}. \end{aligned} \tag{29}$$

Here $b_i^{\setminus h}$ is the quantity in (26) evaluated at $(\widehat{W}^{\setminus h}, \widehat{\mu}^{\setminus h})$. The operator $H^{\setminus h}$ is self-adjoint and strictly positive-definite on $\mathbb{R}^{d\times d}$: its smallest eigenvalue is at least $\lambda n/d^2 = \Theta(1)$ in the regime $d^2 \asymp n$, so that $\big\|(H^{\setminus h})^{-1}\big\|_{\mathrm{op}} = \mathcal{O}_\prec(1)$.

Rank-one reduction. The leave-one-out perturbation is governed, to first order, by the held-out feature pair (u_h, v_h).

Proposition 3 (LOO rank-one reduction)**.** *Under the LOO stability estimates (A1)–(A3) stated below, with high probability,*

$$\widehat{W} - \widehat{W}^{\setminus h} \;=\; \sigma(-\hat{m}_h)\,\big(H^{\setminus h}\big)^{-1}\big[u_h v_h^\top\big] \;+\; R_h, \qquad \big\|R_h\big\|_{\mathsf{F}} \;=\; \mathcal{O}_\prec(d^{-1/2}). \tag{30}$$

Here $\hat{m}_h$ is the TAM margin at the held-out sample evaluated at the full minimizer $(\widehat{W}, \widehat{\mu})$:

$$\hat{m}_h \;=\; s_{h,h}(\widehat{W}) \;-\; \hat{\mu}_h \;-\; \frac{1}{k}\sum_{j\neq h} \phi_\beta\big(s_{j,h}(\widehat{W}) - \hat{\mu}_h\big).$$

Thus the leading change in the weight matrix is the inverse LOO Hessian applied to $\sigma(-\hat{m}_h)\,u_h v_h^\top$, with a Frobenius-norm correction of lower order. In Section 4.3, this formula is combined with trace concentration for $(H^{\setminus h})^{-1}$ to obtain the susceptibility equation.

4.2.2 Proof of Proposition 3

We derive (30) from the following LOO stability proposition.

Proposition 4 (LOO stability)**.** *Uniformly after removing any fixed finite set of samples,*

(A1) $\max_{i\neq h} \left|(\Delta\mu)_i\right| \;=\; \mathcal{O}_{\prec}(d^{-3/2})$;

(A2) $\max_{j,i\neq h} \left|(\Delta S)_{j,i}\right| \;=\; \mathcal{O}_{\prec}(d^{-1})$;

(A3) $\max_{j,i\in[n]} \left|s_{j,i}(\widehat{W})\right| \;=\; \mathcal{O}_{\prec}(1)$.

The proposition is proved in Appendix C. The estimates are in the sense of stochastic domination, absorbing d^{ε} factors. In particular, (A3) applies to the optimizer of each finite-removal problem. These estimates control the Taylor remainder and the coefficient perturbations in the LOO expansion (30).

Step 1: Subtract the stationarity conditions. The W-components of the stationarity conditions (24) at the full and LOO optimizers, subtracted, yield

$$\frac{\lambda n}{d^2}\,\Delta W \;+\; U\left(\hat{Q}\Lambda_{\hat{a}} - Q^{\setminus h}\Lambda_{a^{\setminus h}}\right)V^\top \;=\; 0. \tag{31}$$

Step 2: Separate the terms involving the held-out sample. With the extension conventions on $a^{\setminus h}$ and $q^{\setminus h}$, the matrix $\hat{Q}\Lambda_{\hat{a}} - Q^{\setminus h}\Lambda_{a^{\setminus h}}$ decomposes entrywise as

$$\left(\hat{Q}\Lambda_{\hat{a}} - Q^{\setminus h}\Lambda_{a^{\setminus h}}\right)_{j,i} \;=\; \begin{cases} -\,\hat{a}_h & (j,i) = (h,h), \\ \hat{a}_h\,\hat{q}_{j,h}/k & j\neq h,\; i = h, \\ \hat{a}_i\,\hat{q}_{h,i}/k & j = h,\; i\neq h, \\ \mathcal{R}^{\setminus h}_{j,i} & j,i\neq h, \end{cases}$$

where the first three cases are the entries involving the held-out sample: the held-out outer term ($i = h$) and the held-out interferer row ($j = h$, across interior columns) are present in the full objective but absent in the LOO objective. The remaining entries are collected in $\mathcal{R}^{\setminus h} \in \mathbb{R}^{n\times n}$, supported on $\{j, i \neq h\}$:

$$\mathcal{R}^{\setminus h}_{j,i} \;:=\; \begin{cases} \dfrac{\hat{a}_i\,\hat{q}_{j,i} \;-\; a_i^{\setminus h}\,q_{j,i}^{\setminus h}}{k} & j\neq i,\; j,i\neq h, \\ a_i^{\setminus h} \;-\; \hat{a}_i & j = i\neq h. \end{cases} \tag{32}$$

Separating these entries inside (31),

$$\frac{\lambda n}{d^2}\,\Delta W \;+\; U\,\mathcal{R}^{\setminus h}\,V^\top \;=\; \hat{a}_h\,u_h v_h^\top \;-\; \hat{a}_h\,\eta_h\,v_h^\top \;-\; u_h\,\xi_h^\top, \tag{33}$$

where

$$\eta_h \;:=\; \frac{1}{k}\sum_{j\neq h}\hat{q}_{j,h}\,u_j, \qquad \xi_h \;:=\; \frac{1}{k}\sum_{i\neq h}\hat{a}_i\,\hat{q}_{h,i}\,v_i.$$

Step 3: Linearize the common part. In Appendix B, we carry out a first-order Taylor expansion of $\mathcal{R}^{\setminus h}$ in the perturbations ΔS and $\Delta\mu$, with all higher-order terms collected into a residual matrix $\hat{E}^{\setminus h}$ supported on $\{j, i \neq h\}$. This yields

$$\mathcal{R}^{\setminus h}_{:,i} \;=\; M_i^{\setminus h}\,\Delta S_{:,i} \;+\; \hat{E}^{\setminus h}_{:,i}, \qquad i \neq h, \tag{34}$$

with $M_i^{\setminus h}$ as in (29). Combining (34) with (33) and the definition of $H^{\setminus h}$ yields

$$H^{\setminus h}[\Delta W] \;=\; \hat{a}_h\, u_h v_h^\top \;-\; \hat{a}_h\, \eta_h\, v_h^\top \;-\; u_h\, \xi_h^\top \;-\; U\hat{E}^{\setminus h}V^\top. \tag{35}$$

Moreover, we show in Appendix B that, under (A1)–(A3),

$$\left|\hat{E}^{\setminus h}_{i,i}\right| \;=\; \mathcal{O}_\prec(n^{-1}) \text{ for } i \neq h, \qquad \left|\hat{E}^{\setminus h}_{j,i}\right| \;=\; \mathcal{O}_\prec(n^{-2}) \text{ for } j \neq i,\; j, i \neq h. \tag{36}$$

Step 4: Invert and bound the correction. Applying $(H^{\setminus h})^{-1}$ to (35) produces (30) with

$$R_h \;=\; \left(H^{\setminus h}\right)^{-1}\Big[-\hat{a}_h\, \eta_h\, v_h^\top \;-\; u_h\, \xi_h^\top \;-\; U\hat{E}^{\setminus h}V^\top\Big]. \tag{37}$$

Since $\left\|(H^{\setminus h})^{-1}\right\|_{\mathrm{op}} = \mathcal{O}_\prec(1)$, it suffices to bound each of the three remaining terms in (37) in Frobenius norm at scale $\mathcal{O}_\prec(d^{-1/2})$.

Writing $k\,\eta_h = U\,\hat{q}_{:,h}$ (with the h-th entry of $\hat{q}_{:,h}$ understood to be zero), the Marchenko–Pastur bound $\|U\|_{\mathrm{op}} = \mathcal{O}_\prec(\sqrt{d})$ in the regime $d^2 \asymp n$, together with the deterministic entrywise bound $\left\|\hat{q}_{:,h}\right\|_2 \le \sqrt{n-1}$, gives $\|\eta_h\|_2 \le \|U\|_{\mathrm{op}} \left\|\hat{q}_{:,h}\right\|_2 / k = \mathcal{O}_\prec(\sqrt{nd}/n) = \mathcal{O}_\prec(d^{-1/2})$; the same argument applies to ξ_h. Consequently

$$\left\|\hat{a}_h\, \eta_h\, v_h^\top\right\|_{\mathsf{F}} \;=\; |\hat{a}_h|\, \|\eta_h\|_2\, \|v_h\|_2 \;=\; \mathcal{O}_\prec(d^{-1/2}), \qquad \left\|u_h\, \xi_h^\top\right\|_{\mathsf{F}} \;=\; \mathcal{O}_\prec(d^{-1/2}). \tag{38}$$

Next, we consider the term $U\hat{E}^{\setminus h}V^\top$ in (37). Let $\hat{E}^{\mathsf{d}}$ and $\hat{E}^{\circ}$ denote the diagonal and off-diagonal parts of $\hat{E}^{\setminus h}$, respectively. The diagonal part is a diagonal matrix, so $\left\|\hat{E}^{\mathsf{d}}\right\|_{\mathrm{op}} = \max_{i\neq h} |\hat{E}^{\setminus h}_{i,i}| = \mathcal{O}_\prec(n^{-1})$ by (36). The off-diagonal part is controlled through its Frobenius norm,

$$\left\|\hat{E}^{\circ}\right\|_{\mathrm{op}} \;\le\; \left\|\hat{E}^{\circ}\right\|_{\mathsf{F}} \;=\; \Big(\sum_{\substack{j\neq i\\ j,i\neq h}} |\hat{E}^{\setminus h}_{j,i}|^2\Big)^{1/2} \;\le\; \sqrt{n^2 \cdot \mathcal{O}_\prec(n^{-4})} \;=\; \mathcal{O}_\prec(n^{-1}),$$

again by (36). Combining, $\left\|\hat{E}^{\setminus h}\right\|_{\mathrm{op}} = \mathcal{O}_\prec(n^{-1}) = \mathcal{O}_\prec(d^{-2})$. Submultiplicativity with $\|U\|_{\mathrm{op}}, \|V\|_{\mathrm{op}} = \mathcal{O}_\prec(\sqrt{d})$ yields $\left\|U\hat{E}^{\setminus h}V^\top\right\|_{\mathrm{op}} \le \mathcal{O}_\prec(d \cdot d^{-2}) = \mathcal{O}_\prec(d^{-1})$. Since $U\hat{E}^{\setminus h}V^\top \in \mathbb{R}^{d\times d}$ has rank at most d, $\|\cdot\|_{\mathsf{F}} \le \sqrt{d}\,\|\cdot\|_{\mathrm{op}}$, and

$$\left\|U\hat{E}^{\setminus h}V^\top\right\|_{\mathsf{F}} \;\le\; \sqrt{d}\left\|U\hat{E}^{\setminus h}V^\top\right\|_{\mathrm{op}} \;=\; \mathcal{O}_\prec(d^{-1/2}). \tag{39}$$

Combining (38) and (39) and applying $(H^{\setminus h})^{-1}$ establishes (30). □

4.3 Scalar characterization of the TAM optimizer

We now state the asymptotic characterization of the TAM minimizer. In addition to Proposition 4, we use the following spectral and self-averaging proposition. The derivation is given in Section 4.4.

Proposition 5 (Spectral and self-averaging estimates)**.** *We have the following:*

(A4) *(Spectrum bound)*

$$\left\|\widehat{W}\right\|_{\mathrm{op}} \;=\; \mathcal{O}_{\prec}(\sqrt{d})$$

(A5) *(Empirical-mean concentration) Let $\psi \in C^1(\mathbb{R})$ be pseudo-Lipschitz of order two. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\psi\big(s_{i,i}(\widehat{W})\big) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\psi\big(s_{i,i}(\widehat{W})\big)\right] \;=\; \mathcal{O}_{\prec}(d^{-1/2}),$$

and

$$\frac{1}{n(n-1)}\sum_{i\neq j}\psi\big(s_{j,i}(\widehat{W})\big) - \mathbb{E}\left[\frac{1}{n(n-1)}\sum_{i\neq j}\psi\big(s_{j,i}(\widehat{W})\big)\right] \;=\; \mathcal{O}_{\prec}(d^{-1/2}).$$

(A6) *(Trace concentration)*

$$\frac{1}{d^2}\,\mathrm{tr}\big(H^{-1}\big) - \mathbb{E}\left[\frac{1}{d^2}\,\mathrm{tr}\big(H^{-1}\big)\right] \;=\; \mathcal{O}_{\prec}(d^{-1/2}).$$

The spectral estimate (A4) is proved in Appendix E; the empirical-mean and trace concentration estimates (A5) and (A6) are proved in Appendix D. Here (A4) is a spectral estimate on the optimizer $\widehat{W}$, while (A5) and (A6) are self-averaging estimates at the common scale $\mathcal{O}_{\prec}(d^{-1/2})$ for empirical means and normalized inverse traces, respectively.

For $\nu > 0$, define the scalar TAM functional

$$c_r(\nu) := \min_{\mu\in\mathbb{R}}\left\{\mu + \frac{1}{r}\mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\phi_\beta(\nu Z - \mu)\right]\right\}. \tag{40}$$

Let $\mu_r(\nu)$ denote the minimizer in (40). It is characterized by

$$r = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\sigma\big(\beta(\nu Z - \mu_r(\nu))\big)\right]. \tag{41}$$

Definition 3 (Scalar saddle and reduced potentials)**.** *For $\nu, \chi > 0$, define the scalar Moreau envelope*

$$M_r(\nu,\chi) := \mathbb{E}_{G\sim\mathcal{N}(0,1)}\left[\min_{s\in\mathbb{R}}\left[\frac{(s-\nu G)^2}{2\chi} + \ell\big(s - c_r(\nu)\big)\right]\right]. \tag{42}$$

The scalar saddle potential is

$$\mathcal{P}_{\alpha,r}(\nu,\chi) := \frac{\lambda}{2}\nu^2 - \frac{\nu^2}{2\alpha\chi} + M_r(\nu,\chi),$$

and the reduced scalar potential is

$$\mathcal{V}_{\alpha,r}(\nu) := \sup_{\chi>0}\mathcal{P}_{\alpha,r}(\nu,\chi). \tag{43}$$

Proposition 10, stated in Section 4.4, shows that, for each fixed $\nu > 0$, the maximization over χ is well posed and has a unique maximizer, and that the reduced potential $\mathcal{V}_{\alpha,r}$ is strongly convex in ν.

For $\chi > 0$, define

$$\mathrm{Prox}_\chi(x) := \arg\min_{y\in\mathbb{R}} \left\{ \frac{(y-x)^2}{2\chi} + \ell(y) \right\}. \tag{44}$$

Equivalently, $y = \mathrm{Prox}_\chi(x)$ if and only if

$$y - \chi\sigma(-y) = x.$$

If $G \sim \mathcal{N}(0,1)$, define

$$S_{\nu,\chi} := c_r(\nu) + \mathrm{Prox}_\chi\big(\nu G - c_r(\nu)\big). \tag{45}$$

Let ν_* be the minimizer of $\mathcal{V}_{\alpha,r}$, and let χ_* be the maximizer of $\mathcal{P}_{\alpha,r}(\nu_*, \cdot)$. Define μ_* as the minimizer in (40) at $\nu = \nu_*$, and set

$$c_* := c_r(\nu_*).$$

The variational facts used here, including uniqueness and the equivalence with the finite-dimensional equations derived below, are recorded in Proposition 10.

Theorem 2 (Scalar description of the joint TAM minimizer)**.** *Fix $(\alpha, r, \beta, \lambda)$ with $\beta, \lambda > 0$. Assume that the minimizer ν_* of $\mathcal{V}_{\alpha,r}$ is positive. In the joint limit $n, d \to \infty$ with $n/d^2 \to \alpha$ and $k/n \to r$, the following hold in probability:*

(i) (Order-parameter concentration.)

$$\frac{1}{d}\,\|\widehat{W}\|_{\mathsf{F}} \longrightarrow \nu_*, \qquad \frac{1}{d^2}\,\mathrm{tr}\big(H^{-1}\big) \longrightarrow \chi_*. \tag{46}$$

(ii) (Gaussian law for competitor scores.) *For each $i \in [n]$, the empirical distribution $\dfrac{1}{n-1}\displaystyle\sum_{j\neq i} \delta_{s_{j,i}(\widehat{W})}$ converges weakly to $\mathcal{N}(0, \nu_*^2)$.*

(iii) (Proximal law for diagonal scores.) *The empirical distribution $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \delta_{s_{i,i}(\widehat{W})}$ converges weakly to the law of S_{ν_*,χ_*} in (45).*

Remark 2 (Reading off the retrieval metrics)**.** *Theorem 2 directly answers the three performance questions posed in Section 3.3. Since the TAM aggregate converges to c_*, the limiting margin is*

$$M_* := S_{\nu_*,\chi_*} - c_*,$$

and hence the average fitting loss converges to

$$\mathbb{E}\big[\ell(M_*)\big]\,. \tag{47}$$

The empirical margin distribution converges to the law of M_*. Finally, writing Φ for the standard normal CDF, the signal percentile distribution converges to the law of

$$\Phi\left(\frac{S_{\nu_*,\chi_*}}{\nu_*}\right),$$

because the competitor scores have limiting law $\mathcal{N}(0, \nu_*^2)$.

4.4 Proof of Theorem 2

Define

$$\nu_d^2 \;:=\; \mathbb{E}\left[\frac{\|\widehat{W}\|_{\mathsf{F}}^2}{d^2}\right]. \tag{48}$$

With H denoting the full Hessian operator from Definition 2, evaluated at $(\widehat{W}, \widehat{\mu})$, define

$$\chi_d \;:=\; \mathbb{E}\left[\frac{1}{d^2}\,\mathrm{tr}\big(H^{-1}\big)\right]. \tag{49}$$

In this section, we prove Theorem 2 by deriving the finite-dimensional equations for ν_d and χ_d. The proof proceeds in four steps. First, using (A4) and the bilinear coupling estimate in Appendix F, we identify the off-diagonal competitor scores with Gaussians of variance ν_d^2. Second, this Gaussian approximation and the trace input (A6) give the proximal characterization of the diagonal score $s_{i,i}(\widehat{W})$. Third, W-stationarity, together with empirical concentration (A5), yields a finite-dimensional equation for ν_d. Fourth, a random matrix calculation for a simplified Hessian gives the companion equation for χ_d. Combining the two equations gives an approximate residual system, whose zero-residual limit is the saddle-point condition for the scalar potential in Definition 3.

4.4.1 Gaussian law for competitor scores

We begin with the off-diagonal scores. The statement below gives a finite-dimensional joint Gaussian coupling for any fixed number of competitor scores in a single column, using an iterated leave-one-out reduction and the bilinear coupling estimate of Corollary 1.

Proposition 6 (Joint Gaussian limit for competitor scores)**.** *Fix $i \in [n]$, $m \geq 1$, and distinct $j_1, \ldots, j_m \in [n] \setminus \{i\}$. Assume (A1)–(A6). Then, in the joint regime $n, d \to \infty$ with $n/d^2 \to \alpha$ and $k/n \to r$: on a common probability space, there exist i.i.d. Gaussians $Y_1, \ldots, Y_m \sim \mathcal{N}(0, \nu_d^2)$ such that*

$$\big\|\big(s_{j_1,i}(\widehat{W}), \ldots, s_{j_m,i}(\widehat{W})\big) - (Y_1, \ldots, Y_m)\big\|_2 \;=\; \mathcal{O}_{\prec}(d^{-1/2}). \tag{50}$$

Remark 3 (Connection with the limiting statement)**.** *Proposition 6 is the finite-dimensional form of the off-diagonal Gaussian law in Theorem 2. Empirical-measure consequences, such as weak convergence of $\frac{1}{n-1}\sum_{j\neq i}\delta_{s_{j,i}(\widehat{W})}$ to $\mathcal{N}(0,\nu_d^2)$, follow by combining exchangeability with concentration of the optimizer. The expectation-level replacements used below are obtained from this coupling for the finite list of bounded or C^1 pseudo-Lipschitz observables appearing in the stationarity identities, together with the moment bounds implicit in (A5).*

Proof. Throughout, write $X_k := s_{j_k,i}(\widehat{W})$ for $k = 1, \ldots, m$. Set $\widehat{W}_{\text{fresh}} := \widehat{W}^{\setminus\{i,j_1,\ldots,j_m\}}$ and $V := \|\widehat{W}_{\text{fresh}} v_i\|^2/d$.

Step 1: Iterated leave-one-out. Apply (30) successively after removing the samples $i, j_1, \ldots, j_m$. The finite-removal version of the LOO stability estimates, together with the spectral and self-averaging estimates (A4)–(A5), gives the same rank-one expansion at each level. The resulting

iterated identity expresses

$$\widehat{W} \;=\; \widehat{W}_{\text{fresh}} \;+\; \sum_{\ell=0}^{m} \Big[\, \hat{a}_{i_\ell}^{\setminus I_\ell}\, (H^{\setminus I_\ell})^{-1}\big[u_{i_\ell} v_{i_\ell}^\top\big] + R_\ell \,\Big], \tag{51}$$

where $i_0 = i$, $i_\ell = j_\ell$ for $\ell \geq 1$, $I_\ell := \{i, j_1, \ldots, j_\ell\}$, and each $\|R_\ell\|_{\mathsf{F}} = \mathcal{O}_\prec(d^{-1/2})$ by Section 4.2 applied at the corresponding LOO level.

Step 2: Score decomposition. For each $k \in [m]$, sandwiching (51) with $u_{j_k}^\top \cdot v_i$ yields

$$X_k \;=\; u_{j_k}^\top \widehat{W}_{\text{fresh}} v_i \;+\; \mathcal{E}_k, \tag{52}$$

where $\mathcal{E}_k$ is the sum of $m+1$ rank-one bilinear contributions plus residual terms.

Step 3: Bound the corrections. Each rank-one contribution at level ℓ

$$\hat{a}_{i_\ell}^{\setminus I_\ell}\, u_{j_k}^\top (H^{\setminus I_\ell})^{-1}[u_{i_\ell} v_{i_\ell}^\top]\, v_i$$

admits a self-adjoint rewrite that places one of the held-out vectors — u_i for $\ell = 0$ and v_i for $\ell \geq 1$ — on the outer test side, independent of the LOO operator $(H^{\setminus I_\ell})^{-1}$ (which sees no sample in I_ℓ). Combined with the operator-norm bound $\|(H^{\setminus I_\ell})^{-1}\|_{\text{op}} \leq d^2/(\lambda n) = O(1)$ from the ridge floor and $\|u_a v_b^\top\|_{\mathsf{F}} = O(1)$, the unmatched-Gaussian inner product gives an $\mathcal{O}_\prec(d^{-1/2})$ bound on each rank-one contribution. The residuals $u_{j_k}^\top R_\ell v_i$ are bounded by $\|R_\ell\|_{\mathsf{F}}\, \|u_{j_k}\| \|v_i\| = \mathcal{O}_\prec(d^{-1/2})$. Summing $m+1$ such terms,

$$|\mathcal{E}_k| \;=\; \mathcal{O}_\prec(d^{-1/2}) \quad \text{uniformly in } k. \tag{53}$$

Step 4: Apply the bilinear coupling. Since $\widehat{W}_{\text{fresh}}$ is a function of $\{(u_l, v_l) : l \notin \{i, j_1, \ldots, j_m\}\}$, it is independent of $u_{j_1}, \ldots, u_{j_m}, v_i$. The bilinear coupling Corollary 1 requires a Frobenius self-averaging estimate and an operator-norm input for $\widehat{W}_{\text{fresh}}$; both are transferred from $\widehat{W}$ by the same finite-removal stability used above. In particular, the stochastic-dominance deviation $\|\widehat{W}_{\text{fresh}}\|_{\mathsf{F}}^2/d^2 - \nu_d^2 = \mathcal{O}_\prec(d^{-1/2})$ follows from the Frobenius consequence of (A5), while the operator-norm input $\|(\widehat{W}_{\text{fresh}})^\top \widehat{W}_{\text{fresh}}\|_{\mathsf{F}}/d^2 = \mathcal{O}_\prec(d^{-1/2})$ follows from (A4). Apply Corollary 1 with $W = \widehat{W}_{\text{fresh}}$, $v_0 = v_i$, and $\rho_f = \rho_{\text{op}} = O(d^{-1/2})$: there exist i.i.d. $Y_1, \ldots, Y_m \sim \mathcal{N}(0, \nu_d^2)$ with

$$\big\|(u_{j_1}^\top \widehat{W}_{\text{fresh}} v_i, \ldots, u_{j_m}^\top \widehat{W}_{\text{fresh}} v_i) - (Y_1, \ldots, Y_m)\big\|_2 \;=\; \mathcal{O}_\prec(d^{-1/2}), \tag{54}$$

with $Y_1, \ldots, Y_m$ i.i.d. $\mathcal{N}(0, \nu_d^2)$.

Step 5: Combine. Triangle inequality applied to (52), (53), and (54) yields (50). □

4.4.2 Proximal law for diagonal scores

We next record the finite-dimensional proximal law for the diagonal scores. Set

$$\mu_d := \mu_r(\nu_d), \qquad c_d := c_r(\nu_d),$$

where μ_r and c_r are defined in (40)–(41).

Proposition 7 (Coupled proximal law for s_{ii})**.** *Assume (A1)–(A6). Fix $i \in [n]$. On a common probability space, there is a standard Gaussian random variable $G_i \sim \mathcal{N}(0,1)$ such that*

$$s_{i,i}(\widehat{W}) = c_d + \mathrm{Prox}_{\chi_d}(\nu_d G_i - c_d) + \mathcal{O}_{\prec}(d^{-1/2}), \tag{55}$$

where Prox_χ is defined in (44).

Proof. Apply the LOO rank-one expansion (30) with $h = i$ and contract with $u_i^\top(\cdot)v_i$. This gives

$$s_{i,i}(\widehat{W}) = g_i + \hat{a}_i \big\langle u_i v_i^\top,\, (H^{\setminus i})^{-1}[u_i v_i^\top] \big\rangle + \mathcal{O}_{\prec}(d^{-1/2}), \tag{56}$$

where

$$g_i := u_i^\top \widehat{W}^{\setminus i} v_i, \qquad \hat{a}_i = \sigma(-\hat{m}_i).$$

The LOO Hessian $H^{\setminus i}$ is independent of (u_i, v_i). Lemma 21, (A6), and the finite-removal trace stability therefore give

$$\big\langle u_i v_i^\top,\, (H^{\setminus i})^{-1}[u_i v_i^\top] \big\rangle = \chi_d + \mathcal{O}_{\prec}(d^{-1/2}).$$

Since $\hat{a}_i$ is bounded, (56) becomes

$$s_{i,i}(\widehat{W}) = g_i + \chi_d\, \sigma\big(-s_{i,i}(\widehat{W}) + \hat{c}_i\big) + \mathcal{O}_{\prec}(d^{-1/2}), \tag{57}$$

where

$$\hat{c}_i := \hat{\mu}_i + \frac{1}{k} \sum_{j \neq i} \phi_\beta\big(s_{j,i}(\widehat{W}) - \hat{\mu}_i\big).$$

We now replace g_i and $\hat{c}_i$ by their finite-dimensional deterministic counterparts. The bilinear coupling estimate, applied to the leave-one-out matrix $\widehat{W}^{\setminus i}$ and the fresh pair (u_i, v_i), gives a coupling

$$g_i = \nu_d G_i + \mathcal{O}_{\prec}(d^{-1/2}), \qquad G_i \sim \mathcal{N}(0,1). \tag{58}$$

Next, μ-stationarity gives

$$\frac{1}{n} \sum_{j \neq i} \sigma\big(\beta(s_{j,i}(\widehat{W}) - \hat{\mu}_i)\big) = r + \mathcal{O}_{\prec}(d^{-1/2}).$$

The off-diagonal Gaussian law from Proposition 6, together with the empirical-mean concentration estimate, implies that the left-hand side is

$$\mathbb{E}_{Z \sim \mathcal{N}(0,1)} \Big[\sigma\big(\beta(\nu_d Z - \hat{\mu}_i)\big)\Big] + \mathcal{O}_{\prec}(d^{-1/2}).$$

For fixed $r \in (0,1)$ and fixed β, the bounded-score event keeps $\hat{\mu}_i$ and μ_d in a compact set; on this set the derivative in μ of the preceding expectation is bounded away from zero. Hence (41) yields

$$\hat{\mu}_i = \mu_d + \mathcal{O}_{\prec}(d^{-1/2}).$$

Applying the same Gaussian approximation to the pseudo-Lipschitz test function $\phi_\beta(\cdot - \mu)$, and using the Lipschitz dependence on μ, gives

$$\hat{c}_i = c_d + \mathcal{O}_{\prec}(d^{-1/2}). \tag{59}$$

Combining (57), (58), and (59), we obtain

$$s_{i,i}(\widehat{W}) = \nu_d G_i + \chi_d\, \sigma\big(-s_{i,i}(\widehat{W}) + c_d\big) + \mathcal{O}_{\prec}(d^{-1/2}).$$

Let $y_i := s_{i,i}(\widehat{W}) - c_d$. The last display is equivalent to

$$y_i - \chi_d \sigma(-y_i) = \nu_d G_i - c_d + \mathcal{O}_{\prec}(d^{-1/2}).$$

The map $y \mapsto y - \chi_d \sigma(-y)$ is strictly increasing with derivative at least one, so its inverse Prox_{χ_d} is 1-Lipschitz. This proves (55). □

4.4.3 Finite-dimensional scalar equations and variational consistency

We now derive the two finite-dimensional scalar equations for ν_d and χ_d, and then identify their zero-residual limit with the saddle-point equations of the scalar variational problem. The variance equation is obtained directly from W-stationarity. Let G and Z be independent standard Gaussian variables, and define

$$S_d := c_d + \mathrm{Prox}_{\chi_d}(\nu_d G - c_d)\,, \qquad A_d := \sigma(-S_d + c_d). \tag{60}$$

Proposition 8 (Finite-dimensional variance equation)**.** *Assume (A1)–(A6), and use the coupled diagonal score law of Proposition 7. Then*

$$\nu_d^2 \left[\lambda + \frac{1}{r}\,\mathbb{E}[A_d]\,\mathbb{E}\Big[\phi_\beta''\big(\nu_d Z - \mu_d\big)\Big]\right] = \mathbb{E}[A_d S_d] + O(d^{-1/2}). \tag{61}$$

Proof. At the full minimizer, the W-stationarity condition (24) gives

$$UQ\Lambda_a V^\top + \frac{\lambda n}{d^2}\widehat{W} = 0.$$

Taking the Frobenius inner product with $\widehat{W}$ and using $s_{j,i}(\widehat{W}) = u_j^\top \widehat{W} v_i$ gives the exact identity

$$\lambda\,\frac{\|\widehat{W}\|_{\mathsf{F}}^2}{d^2} = \frac{1}{n}\sum_{i=1}^{n} a_i \left(s_{i,i}(\widehat{W}) - \frac{1}{k}\sum_{j\neq i} q_{j,i}\, s_{j,i}(\widehat{W})\right). \tag{62}$$

Taking expectations turns the left-hand side into $\lambda\nu_d^2$.

The first empirical average in (62) is a smooth test function of the diagonal score, after replacing c_i by c_d at the $d^{-1/2}$ scale. Proposition 7 and (A5) therefore yield

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} a_i s_{i,i}(\widehat{W})\right] = \mathbb{E}[A_d S_d] + O(d^{-1/2}).$$

For the off-diagonal contribution, the competitor Gaussian law Proposition 6, the same empirical-mean concentration, and the leave-one-out independence between the diagonal factor and the competitor score give

$$Q_d := \sigma\big(\beta(\nu_d Z - \mu_d)\big).$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} a_i \frac{1}{k}\sum_{j\neq i} q_{j,i} s_{j,i}(\widehat{W})\right] = \frac{1}{r}\,\mathbb{E}[A_d]\mathbb{E}[\nu_d Z Q_d] + O(d^{-1/2}). \tag{63}$$

Here the distinction between k/n and $k/(n-1)$ contributes only $O(n^{-1})$, which is absorbed into $O(d^{-1/2})$. Combining (62)–(63) gives

$$\lambda\nu_d^2 = \mathbb{E}[A_d S_d] - \frac{1}{r}\,\mathbb{E}[A_d]\mathbb{E}[\nu_d Z Q_d] + O(d^{-1/2}).$$

Finally, Stein's identity for the standard Gaussian Z gives

$$\mathbb{E}[\nu_d Z Q_d] = \nu_d^2 \mathbb{E}\Big[\phi_\beta''\big(\nu_d Z - \mu_d\big)\Big],$$

which gives (61). □

We next derive the companion equation for the scalar χ_d defined in (49). It is the expectation of the normalized inverse trace of the full Hessian, and is often referred to as the susceptibility in the statistical physics literature. Set

$$B_d := A_d(1 - A_d).$$

Proposition 9 (Finite-dimensional susceptibility equation)**.** *Assume (A1)–(A6). Then the finite-dimensional susceptibility χ_d satisfies*

$$1 = \alpha\lambda\chi_d + \alpha\,\mathbb{E}\left[\frac{\chi_d B_d}{1+\chi_d B_d}\right] + \frac{\alpha\chi_d}{r}\,\mathbb{E}[A_d]\,\mathbb{E}\Big[\phi_\beta''\big(\nu_d Z - \mu_d\big)\Big] + O(d^{-1/2}). \tag{64}$$

In the above display, A_d is the random variable defined in (60).

Proof. For $j, i \in [n]$, write

$$f_{ji} := \mathrm{vec}(u_j v_i^\top) \in \mathbb{R}^{d^2}.$$

Also let

$$r_{j,i} := \phi_\beta''\big(s_{j,i} - \mu_i\big) = \beta\, q_{j,i}(1 - q_{j,i}), \qquad q_{j,i} = \sigma\big(\beta(s_{j,i} - \mu_i)\big),$$

so that $\widetilde{R}_{j,i} = r_{j,i}/k$. Define the simplified Hessian

$$\widetilde{H} = \frac{\lambda n}{d^2} I + \sum_{i=1}^{n} b_i f_{ii} f_{ii}^\top + \sum_{i=1}^{n}\sum_{j\neq i} \frac{a_i r_{j,i}}{k} f_{ji} f_{ji}^\top. \tag{65}$$

By Lemmas 22 and 23 in Appendix G, the normalized inverse trace of the original Hessian may be replaced, up to $\mathcal{O}_\prec(d^{-1/2})$, by the normalized inverse trace of $\widetilde{H}$. Thus, by the trace concentration estimate (A6),

$$\frac{1}{d^2}\operatorname{tr}(\widetilde{H}^{-1}) = \chi_d + \mathcal{O}_\prec(d^{-1/2}). \tag{66}$$

We now use the identity

$$1 = \frac{1}{d^2}\operatorname{tr}\Big(\widetilde{H}\widetilde{H}^{-1}\Big).$$

Expanding the three terms in (65) gives

$$1 = \frac{\lambda n}{d^2}\frac{1}{d^2}\operatorname{tr}(\widetilde{H}^{-1}) + \frac{1}{d^2}\sum_{i=1}^{n} b_i f_{ii}^\top \widetilde{H}^{-1} f_{ii} + \frac{1}{d^2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{a_i r_{j,i}}{k} f_{ji}^\top \widetilde{H}^{-1} f_{ji}.$$

The first term is evaluated by (66). Lemmas 25 and 26 in Appendix G evaluate the two families of quadratic forms:

$$f_{ii}^\top \widetilde{H}^{-1} f_{ii} = \frac{\chi_d}{1+b_i\chi_d} + \mathcal{O}_\prec(d^{-1/2}), \qquad f_{ji}^\top \widetilde{H}^{-1} f_{ji} = \chi_d + \mathcal{O}_\prec(d^{-1/2}).$$

Substitution gives the finite-dimensional trace identity

$$1 = \frac{\lambda n}{d^2}\chi_d + \frac{1}{d^2}\sum_{i=1}^{n} b_i \frac{\chi_d}{1+b_i\chi_d} + \frac{\chi_d}{d^2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{a_i r_{j,i}}{k} + \mathcal{O}_\prec(d^{-1/2}). \tag{67}$$

Taking expectations under the moment bounds established in the appendices, and then applying the diagonal proximal law, the off-diagonal Gaussian law, and (A5), gives

$$\mathbb{E}\left[b_i \frac{\chi_d}{1+b_i\chi_d}\right] = \mathbb{E}\left[\frac{\chi_d B_d}{1+\chi_d B_d}\right] + O(d^{-1/2}),$$

and

$$\mathbb{E}\left[a_i \frac{1}{k}\sum_{j\neq i} r_{j,i}\right] = \frac{1}{r}\mathbb{E}[A_d]\,\mathbb{E}\Big[\phi_\beta''\big(\nu_d Z - \mu_d\big)\Big] + O(d^{-1/2}).$$

Substituting these two replacements into (67), and replacing n/d^2 by α at the same error scale, yields (64). □

Combining Propositions 8 and 9 gives the finite-dimensional residual system

$$\begin{aligned} \nu_d^2\left[\lambda + \frac{1}{r}\mathbb{E}[A_d]\,\mathbb{E}\Big[\phi_\beta''\big(\nu_d Z - \mu_d\big)\Big]\right] - \mathbb{E}[A_d S_d] &= O(d^{-1/2}), \\ 1 - \alpha\lambda\chi_d - \alpha\,\mathbb{E}\left[\frac{\chi_d B_d}{1+\chi_d B_d}\right] - \frac{\alpha}{r}\mathbb{E}[A_d]\,\mathbb{E}\Big[\phi_\beta''\big(\nu_d Z - \mu_d\big)\Big]\chi_d &= O(d^{-1/2}). \end{aligned} \tag{68}$$

Here $\mu_d = \mu_r(\nu_d)$ is the deterministic threshold defined in (41), evaluated at the finite-dimensional variance parameter ν_d. Thus (ν_d, χ_d) is an *approximate* solution of the nonlinear scalar equations whose zero-residual limit turns out to be exactly equal to the saddle-point equations for the potential $\mathcal{P}_{\alpha,r}$ in Definition 3. This variational connection is established in the next proposition.

Proposition 10 (Variational form of the scalar equations)**.** *For every $\nu > 0$, $\chi \mapsto \mathcal{P}_{\alpha,r}(\nu, \chi)$ has a unique maximizer. Moreover, the reduced potential $\mathcal{V}_{\alpha,r}$ is λ-strongly convex in ν. If ν_* is the unique minimizer of $\mathcal{V}_{\alpha,r}$ and χ_* is the unique maximizer of $\mathcal{P}_{\alpha,r}(\nu_*, \cdot)$, then the saddle-point equations of*

$$\inf_{\nu>0} \sup_{\chi>0} \mathcal{P}_{\alpha,r}(\nu, \chi)$$

are exactly the zero-residual limit of (68), with the finite-dimensional proximal channel replaced by (45) and with c_r given by (40).

Proof. See Appendix H. □

5 Ridgeless limit and phase transition

The scalar theory in Section 4.3 gives a limiting variational problem for each fixed ridge parameter $\lambda > 0$ and fixed smoothing parameter $\beta < \infty$. In this section we take the sequential ridgeless limit: first $n, d \to \infty$ with $n/d^2 \to \alpha$ and $k/n \to r$, and then $\lambda \downarrow 0$ while keeping $\beta \in (0, \infty)$ fixed. The limit separates two regimes. Below the critical load $\alpha_c(r)$ defined in (69) below, the scalar norm ν escapes to infinity as $\lambda \downarrow 0$ and the limiting fitting loss vanishes; we call this the satisfiable, or SAT, phase. Above this critical load the scalar norm remains finite and the limiting fitting loss is positive; we call this the unsatisfiable, or UNSAT, phase. After deriving the transition from the variational principle, we describe the score, margin, and percentile profiles on both sides.

All results in this section concern the scalar saddle problem of Definition 3. They translate into statements about the original TAM minimizer $(\widehat{W}, \widehat{\mu})$ through the scalar identification in Theorem 2.

5.1 The critical load

Definition 4 (Ridgeless threshold)**.** *Let $G \sim N(0,1)$, and define*

$$\kappa_r := \frac{\varphi(\Phi^{-1}(1-r))}{r},$$

where Φ and φ are the CDF and density of a standard normal random variable. The critical load is

$$\alpha_c(r) = \frac{1}{\mathbb{E}[(\kappa_r - G)_+^2]} = \frac{1}{(1+\kappa_r^2)\Phi(\kappa_r) + \kappa_r \varphi(\kappa_r)}. \tag{69}$$

In what follows, we first explain the origin and underlying ideas behind the transition at $\alpha_c(r)$. The formal proof is then given in Proposition 11.

We start from the dual form of the reduced potential in (219), writing the ridge parameter explicitly:

$$\begin{aligned}\mathcal{V}_{\alpha,r,\lambda}(\nu) = \frac{\lambda}{2}\nu^2 + \sup_{0 \le a(\cdot) \le 1} \Big\{ &\mathbb{E}\Big[a(G)\big(c_r(\nu) - \nu G\big) - I(a(G))\Big] \\ &- \frac{\nu}{\sqrt{\alpha}} \left(\mathbb{E}[a(G)^2]\right)^{1/2} \Big\},\end{aligned} \tag{70}$$

where $G \sim N(0,1)$ and $I(a) = a\log a + (1-a)\log(1-a)$. The supremum is over measurable functions $a : \mathbb{R} \to [0,1]$. We write $\mathcal{V}_{\alpha,r,0}$ for the ridgeless potential obtained by setting $\lambda = 0$; explicitly,

$$\mathcal{V}_{\alpha,r,0}(\nu) = \sup_{0\le a(\cdot)\le 1} \left\{ \mathbb{E}\Big[a(G)\big(c_r(\nu) - \nu G\big) - I(a(G))\Big] - \frac{\nu}{\sqrt{\alpha}} \left(\mathbb{E}[a(G)^2]\right)^{1/2} \right\}. \tag{71}$$

The phase transition is determined by the large-ν behavior of the ridgeless potential $\mathcal{V}_{\alpha,r,0}$. The only input from the TAM threshold $c_r(\nu)$ needed for the leading-order calculation is the asymptotic relation

$$\frac{c_r(\nu)}{\nu} \longrightarrow \kappa_r. \tag{72}$$

We defer the verification of (72) to the proof of Proposition 11. Now set $m_r(\nu) := c_r(\nu)/\nu$. For fixed $a(\cdot)$, the linear-in-ν part of (70) is

$$\nu \left[\mathbb{E}\big[a(G)(m_r(\nu) - G)\big] - \frac{1}{\sqrt{\alpha}}\|a\|_2 \right], \qquad \|a\|_2 := \left(\mathbb{E}[a(G)^2]\right)^{1/2}.$$

The entropy term $I(a)$ is not multiplied by ν. Thus the large-norm behavior is governed by a simple question: can we choose a nonzero dual profile a so that the coefficient of the leading linear term in ν is positive?

Letting $\nu \to \infty$ in this coefficient and using (72), the relevant expression becomes

$$\mathbb{E}[a(G)(\kappa_r - G)] - \frac{1}{\sqrt{\alpha}}\|a\|_2.$$

Since $a \ge 0$, the negative part of $\kappa_r - G$ can only hurt, so the best profile should live on the event $\{G < \kappa_r\}$. More precisely, by Cauchy–Schwarz, for any $0 \le a \le 1$,

$$\mathbb{E}[a(G)(\kappa_r - G)] \le \mathbb{E}[a(G)(\kappa_r - G)_+] \le \|a\|_2 \left(\mathbb{E}[(\kappa_r - G)_+^2]\right)^{1/2}.$$

The upper bound is sharp despite the pointwise constraint $a \le 1$. Indeed, equality in Cauchy–Schwarz would ask for a proportional to $(\kappa_r - G)_+$, which is unbounded. We can approximate this direction by the admissible truncations

$$a_M(G) := \min\left\{ \frac{(\kappa_r - G)_+}{M}, 1 \right\}, \qquad M \to \infty.$$

The ratio $\mathbb{E}[a_M(G)(\kappa_r - G)]/\|a_M\|_2$ converges to $\left(\mathbb{E}[(\kappa_r - G)_+^2]\right)^{1/2}$. Hence a positive large-ν direction exists exactly when $\alpha > \alpha_c(r)$, with $\alpha_c(r)$ as in (69).

Proposition 11 (Ridgeless scalar phase transition)**.** *Fix $r \in (0,1)$ and $\beta \in (0,\infty)$. For each $\lambda > 0$, let ν_λ be the unique minimizer of $\mathcal{V}_{\alpha,r,\lambda}$.*

(i) *If $\alpha < \alpha_c(r)$, then*

$$\nu_\lambda \to \infty \qquad \textit{and} \qquad \mathcal{V}_{\alpha,r,0}(\nu_\lambda) \to 0$$

as $\lambda \downarrow 0$.

(ii) *If $\alpha > \alpha_c(r)$, then ν_λ remains bounded as $\lambda \downarrow 0$ and converges to the unique minimizer $\nu_0 < \infty$ of $\mathcal{V}_{\alpha,r,0}$. The ridgeless fitting loss $\mathcal{V}_{\alpha,r,0}(\nu_0)$ is strictly positive.*

Proof. We first justify (72). Rescaling the definition of c_r by writing $\mu = \nu t$ gives

$$\frac{c_r(\nu)}{\nu} = \min_{t\in\mathbb{R}} \left\{ t + \frac{1}{r}\mathbb{E}\left[\frac{\phi_\beta(\nu(G-t))}{\nu}\right]\right\}.$$

For each fixed y, the function $\nu \mapsto \phi_\beta(\nu y)/\nu$ is nonincreasing and converges to y_+. Thus the limiting problem is

$$\min_{t\in\mathbb{R}} \left\{ t + \frac{1}{r}\mathbb{E}[(G-t)_+]\right\}.$$

Its minimizer is $t = \Phi^{-1}(1-r)$ and its value is κ_r, which proves (72). In fact the convergence is monotone decreasing.

SAT case: $\alpha < \alpha_c(r)$. By (72) and the definition (69), there are $\delta > 0$ and $\nu_0 < \infty$ such that, for all $\nu \geq \nu_0$,

$$\left(\mathbb{E}[(m_r(\nu) - G)_+^2]\right)^{1/2} \leq \frac{1}{\sqrt{\alpha}} - \delta.$$

Therefore, uniformly over $0 \leq a \leq 1$,

$$\mathbb{E}[a(G)(m_r(\nu) - G)] - \frac{1}{\sqrt{\alpha}}\|a\|_2 \leq -\delta\|a\|_2.$$

The remaining terms in the dual representation are lower order. More precisely, for any $p \in (0,1)$ there is $C_p < \infty$ such that $-I(x) \leq C_p x^p$ for $0 \leq x \leq 1$, while $\mathbb{E}[a^p] \leq \|a\|_2^p$. Hence

$$\mathcal{V}_{\alpha,r,0}(\nu) \leq \sup_{0\leq x\leq 1} \left\{-\delta\nu x + C_p x^p\right\} \longrightarrow 0 \quad \text{as } \nu \to \infty.$$

Since $a = 0$ is admissible, $\mathcal{V}_{\alpha,r,0}(\nu) \geq 0$, and therefore $\mathcal{V}_{\alpha,r,0}(\nu) \to 0$ as $\nu \to \infty$.

To identify the variational value with the scalar fitting loss, fix ν and let $\chi(\nu)$ be the unique maximizer of $\mathcal{P}_{\alpha,r,\lambda}(\nu,\cdot)$, whose existence and uniqueness are given by Proposition 10. With the scalar proximal channel $S_{\nu,\chi}$ defined in (45), the identity

$$\mathcal{V}_{\alpha,r,\lambda}(\nu) = \frac{\lambda}{2}\nu^2 + \mathbb{E}\big[\ell(S_{\nu,\chi} - c_r(\nu))\big]$$

follows from a short cancellation. Writing $A := \sigma(-S_{\nu,\chi} + c_r(\nu))$, the proximal equation gives $S_{\nu,\chi} - \nu G = \chi A$, and hence

$$M_r(\nu,\chi) = \frac{\chi}{2}\mathbb{E}[A^2] + \mathbb{E}\big[\ell(S_{\nu,\chi} - c_r(\nu))\big].$$

At the maximizing χ, the saddle equation $\nu^2 = \alpha\chi^2\mathbb{E}[A^2]$ gives $\nu^2/(2\alpha\chi) = (\chi/2)\mathbb{E}[A^2]$, so the two quadratic terms in $\mathcal{P}_{\alpha,r,\lambda}$ cancel. Thus the ridgeless potential is the scalar fitting loss and is strictly positive at every finite ν. For any $\varepsilon > 0$, choose N so large that $\mathcal{V}_{\alpha,r,0}(N) \leq \varepsilon$, and then choose λ so small that $\lambda N^2/2 \leq \varepsilon$. This gives

$$\mathcal{V}_{\alpha,r,\lambda}(\nu_\lambda) \leq 2\varepsilon.$$

Hence the optimal value tends to zero. If ν_λ remained bounded along a subsequence, lower semi-continuity would force a finite point with zero ridgeless fitting loss, contradicting strict positivity. Therefore $\nu_\lambda \to \infty$, and the displayed identity implies that the per-sample fitting loss tends to zero.

UNSAT case: $\alpha > \alpha_c(r)$. By the sharpness of the Cauchy–Schwarz bound, there is a bounded measurable $a : \mathbb{R} \to [0, 1]$ and a number $\delta > 0$ such that

$$\mathbb{E}[a(G)(\kappa_r - G)] - \frac{1}{\sqrt{\alpha}}\|a\|_2 \geq 2\delta.$$

Using $m_r(\nu) \to \kappa_r$ and bounded convergence, the same a satisfies, for all sufficiently large ν,

$$\mathbb{E}[a(G)(m_r(\nu) - G)] - \frac{1}{\sqrt{\alpha}}\|a\|_2 \geq \delta.$$

Substituting this fixed test function into (70) gives

$$\mathcal{V}_{\alpha,r,0}(\nu) \geq \delta\nu - O(1),$$

so the ridgeless potential is coercive. The minimizers of $\mathcal{V}_{\alpha,r,\lambda} = \mathcal{V}_{\alpha,r,0} + \lambda\nu^2/2$ therefore remain in a compact set as $\lambda \downarrow 0$. Any limit point minimizes $\mathcal{V}_{\alpha,r,0}$ by the standard variational argument: compare ν_λ against an arbitrary fixed ν and let $\lambda \downarrow 0$.

We now show that the ridgeless minimizer is unique. For finite β, the TAM threshold c_r is strictly convex. Indeed, with

$$w_\nu(Z) := \phi''_\beta\big(\nu Z - \mu_r(\nu)\big),$$

the envelope theorem and the defining equation for $\mu_r(\nu)$ give

$$c''_r(\nu) = \frac{1}{r}\left\{\mathbb{E}[w_\nu(Z)Z^2] - \frac{\mathbb{E}[w_\nu(Z)Z]^2}{\mathbb{E}[w_\nu(Z)]}\right\} > 0.$$

For each finite ν, the dual maximizer in (71) has positive expectation. In the primal variables this maximizer is $a(G) = \sigma(-S_{\nu,\chi} + c_r(\nu))$, and hence $0 < a(G) < 1$ almost surely. Therefore $\mathcal{V}_{\alpha,r,0}$ is strictly convex: if $\nu_\theta = \theta\nu_1 + (1-\theta)\nu_2$ with $0 < \theta < 1$ and $\nu_1 \neq \nu_2$, evaluating the dual formula at the maximizer for ν_θ and using the strict convexity of c_r gives

$$\mathcal{V}_{\alpha,r,0}(\nu_\theta) < \theta\mathcal{V}_{\alpha,r,0}(\nu_1) + (1-\theta)\mathcal{V}_{\alpha,r,0}(\nu_2).$$

Thus the ridgeless minimizer is unique; call it ν_0. Hence $\nu_\lambda \to \nu_0$. Since ν_0 is finite, the corresponding scalar logistic loss is strictly positive. □

Remark 4 (No dependence on the smoothing parameter)**.** *The critical load $\alpha_c(r)$ does not depend on the smoothing parameter $\beta \in (0, \infty)$. The parameter β affects the finite-λ scalar channel and the finite ridgeless minimizer in the UNSAT phase, but it disappears from the large-ν limit* (72)*, where the smoothed logistic tail $\phi_\beta(\nu y)/\nu$ converges to y_+.*

5.2 Score, margin, and percentile profiles

We next translate the ridgeless phase transition into predictions for the true scores, the competing scores, the margins, and the percentile rank of the true score among its competitors. The two phases have different scaling behavior. In the SAT phase, Proposition 11 shows that $\nu_\lambda \to \infty$. Thus the true scores and the competitor scores both diverge on the original scale. The meaningful objects are the normalized scores, obtained by dividing by ν_λ, and these have a non-degenerate limit. In the UNSAT phase, by contrast, ν_λ remains finite, so the ridgeless profiles live on the original scale.

5.2.1 The SAT phase

We begin with the SAT phase, where the limiting normalized profiles are explicit. The single parameter controlling them is the number $\rho_\alpha > \kappa_r$, defined as the unique solution of

$$\mathbb{E}[(\rho_\alpha - G)_+^2] = \frac{1}{\alpha}, \qquad G \sim N(0,1). \tag{73}$$

The left-hand side is strictly increasing in ρ_α, and the boundary value $\rho_\alpha = \kappa_r$ is exactly (69). Hence ρ_α is monotone decreasing in α and satisfies $\rho_\alpha \downarrow \kappa_r$ as $\alpha \uparrow \alpha_c(r)$.

Proposition 12 (SAT score, margin, and percentile profiles)**.** *Assume $\alpha < \alpha_c(r)$, and let $\rho_\alpha > \kappa_r$ be the unique solution of (73). In the sequential ridgeless limit $\lambda \downarrow 0$, the normalized true score satisfies*

$$\frac{S_{\nu_\lambda,\chi_\lambda}}{\nu_\lambda} \Rightarrow \max\{G, \rho_\alpha\}, \qquad G \sim N(0,1). \tag{74}$$

Consequently, the normalized TAM margin satisfies

$$\frac{S_{\nu_\lambda,\chi_\lambda} - c_r(\nu_\lambda)}{\nu_\lambda} \Rightarrow \max\{G, \rho_\alpha\} - \kappa_r, \qquad G \sim N(0,1). \tag{75}$$

Here and below, $\Rightarrow$ denotes weak convergence of laws as $\lambda \downarrow 0$. The associated percentile variable satisfies

$$\Omega \Rightarrow \Phi\big(\max\{G, \rho_\alpha\}\big)\,. \tag{76}$$

Equivalently, if $p_\alpha := \Phi(\rho_\alpha)$, then

$$\mathbb{P}(\Omega \le \omega) = \begin{cases} 0, & \omega < p_\alpha, \\ \omega, & p_\alpha \le \omega \le 1. \end{cases} \tag{77}$$

Proof. By Proposition 11, in the SAT phase $\nu_\lambda \to \infty$ and the scalar fitting loss tends to zero. The proof has three steps. First, we rewrite the proximal channel in terms of the normalized score and the correction term created by the loss. Second, we introduce the scalar quantity ρ_λ and use the saddle-point equation to show that it is bounded. Third, we identify every subsequential limit of ρ_λ; uniqueness of the equation defining ρ_α then gives the stated weak convergence.

Set $c_\lambda := c_r(\nu_\lambda)$. For each deterministic scalar input $g \in \mathbb{R}$, define

$$S_\lambda(g) := c_\lambda + \mathrm{Prox}_{\chi_\lambda}(\nu_\lambda g - c_\lambda), \qquad X_\lambda(g) := \frac{S_\lambda(g)}{\nu_\lambda}.$$

The proximal first-order condition is the pointwise identity

$$S_\lambda(g) = \nu_\lambda g + \chi_\lambda \sigma(-S_\lambda(g) + c_\lambda).$$

We also define the pointwise correction

$$B_\lambda(g) := X_\lambda(g) - g = \frac{\chi_\lambda}{\nu_\lambda}\sigma(-S_\lambda(g) + c_\lambda) \ge 0. \tag{78}$$

Only now do we insert the Gaussian variable: when $G \sim N(0,1)$, the random variables in the scalar channel are $S_\lambda(G)$, $X_\lambda(G)$, and $B_\lambda(G)$. The χ saddle-point equation for the scalar potential in Definition 3, recorded in (215) of Appendix H, gives

$$\mathbb{E}[B_\lambda(G)^2] = \frac{1}{\alpha}. \tag{79}$$

Moreover, the zero-loss statement implies that the normalized score does not fall below the TAM threshold. Indeed, with $M_\lambda(g) := S_\lambda(g) - c_\lambda$, we have pointwise

$$\left(\frac{c_\lambda}{\nu_\lambda} - X_\lambda(g)\right)_+ = \frac{(-M_\lambda(g))_+}{\nu_\lambda}.$$

For every $m \in \mathbb{R}$,

$$(-m)_+ \le \ell(m).$$

Therefore, since the fitting loss $\mathbb{E}[\ell(M_\lambda)]$ tends to zero and $\nu_\lambda \to \infty$,

$$\mathbb{E}\left[\left(\frac{c_\lambda}{\nu_\lambda} - X_\lambda(G)\right)_+\right] \longrightarrow 0.$$

Since $c_\lambda/\nu_\lambda \to \kappa_r$ by (72), the 1-Lipschitz property of $x \mapsto x_+$ gives

$$\mathbb{E}\left[\big(\kappa_r - X_\lambda(G)\big)_+\right] \longrightarrow 0. \tag{80}$$

We now identify the limit of the channel. The useful normalization is the level at which the sigmoid correction in (78) changes from negligible to order one. Define

$$\rho_\lambda := \frac{c_\lambda + \log(\chi_\lambda/\nu_\lambda)}{\nu_\lambda}.$$

Combining this definition with (78) and the identity $\sigma(x) = (1+\exp\{-x\})^{-1}$ gives the exact formula

$$B_\lambda(g) = \frac{\exp\{\nu_\lambda(\rho_\lambda - X_\lambda(g))\}}{1+\exp\{c_\lambda - \nu_\lambda X_\lambda(g)\}}. \tag{81}$$

The sequence ρ_λ is bounded. Indeed, if $\rho_\lambda \to -\infty$ along a subsequence, then $\chi_\lambda/\nu_\lambda = \exp\{\nu_\lambda\rho_\lambda - c_\lambda\} \to 0$, and hence $B_\lambda(g) \le \chi_\lambda/\nu_\lambda \to 0$ uniformly in g, contradicting (79). Next suppose that $\rho_\lambda \to +\infty$ along a subsequence. We claim that, for all sufficiently small λ,

$$B_\lambda(g) \ge \frac{\rho_\lambda}{2} \qquad \text{for every } g \in [-1,0].$$

Indeed, if the reverse inequality held at some $g \in [-1,0]$, then $X_\lambda(g) = g + B_\lambda(g) \le \rho_\lambda/2$. Using (81) and the boundedness of c_λ/ν_λ, we would get

$$B_\lambda(g) \ge \exp\{\nu_\lambda\rho_\lambda/3\}$$

for all sufficiently small λ, contradicting $B_\lambda(g) < \rho_\lambda/2$. Hence the claim holds. Since $\mathbb{P}(-1 \le G \le 0) = \Phi(0) - \Phi(-1) > 0$, the claim implies

$$\mathbb{E}[B_\lambda(G)^2] \ge \big(\Phi(0) - \Phi(-1)\big)\frac{\rho_\lambda^2}{4} \longrightarrow \infty,$$

contradicting (79).

It remains to identify the possible limits of ρ_λ. Take a subsequence along which $\rho_\lambda \to \rho$. Passing to a further subsequence, we may assume from (80) that

$$\liminf_{\lambda \downarrow 0} X_\lambda(G) \geq \kappa_r \qquad \text{almost surely.} \tag{82}$$

The rest of the argument is pointwise in g. First, if $g > \rho$, then $B_\lambda(g) \to 0$ and $X_\lambda(g) \to g$. Indeed, any positive subsequential limit of $B_\lambda(g)$ would force $X_\lambda(g) = g + B_\lambda(g) > \rho_\lambda$ eventually, and the right-hand side of (81) would then converge to zero. This excludes $\rho < \kappa_r$, because on the positive-probability event $\{\rho < G < \kappa_r\}$ it would imply $X_\lambda(G) \to G < \kappa_r$, contradicting (82). Hence $\rho \geq \kappa_r$.

We now identify the channel limit for any subsequential limit $\rho \geq \kappa_r$. Along any bounded subsequence of ρ_λ, (81) gives the domination

$$B_\lambda(g)^2 \leq 1 + \left(\sup_\lambda \rho_\lambda - g\right)_+^2. \tag{83}$$

Indeed, if $B_\lambda(g) \leq 1$ there is nothing to prove; if $B_\lambda(g) > 1$, then (81) implies $X_\lambda(g) \leq \rho_\lambda$, and so $B_\lambda(g) = X_\lambda(g) - g \leq (\rho_\lambda - g)_+$.

For $g > \rho$, we already proved $X_\lambda(g) \to g$. Now fix $g < \rho$. For almost every such g, every subsequential limit x of $X_\lambda(g)$ satisfies $x \geq \kappa_r$ by (82). If $x < \rho$, then the exponential numerator in (81) grows like $\exp\{\nu_\lambda(\rho - x + o(1))\}$, while the denominator grows at most like $\exp\{\nu_\lambda(\kappa_r - x)_+ + o(\nu_\lambda)\}$. Since $\rho \geq \kappa_r$ and $x \geq \kappa_r$, this is impossible. If $x > \rho$, the correction term vanishes by (81), forcing $x = g$, also impossible because $g < \rho$. Hence $x = \rho$. Thus, with $G \sim N(0,1)$,

$$X_\lambda(G) \longrightarrow \max\{G, \rho\}, \qquad B_\lambda(G) \longrightarrow (\rho - G)_+.$$

The domination bound (83) allows us to pass to the second moment in (79), so

$$\mathbb{E}[(\rho - G)_+^2] = \frac{1}{\alpha}.$$

By uniqueness of the solution to (73), every subsequential limit is ρ_α. Hence the whole sequence satisfies $X_\lambda(G) \Rightarrow \max\{G, \rho_\alpha\}$. Subtracting $c_\lambda/\nu_\lambda \to \kappa_r$ proves (74) and (75).

The percentile distribution is obtained by comparing the diagonal score with the limiting competitor field $N(0, \nu_\lambda^2)$. Hence (76). The CDF formula (77) follows by noting that $\Phi(\max\{G, \rho_\alpha\})$ has an atom of mass $p_\alpha = \Phi(\rho_\alpha)$ at p_α, followed by a uniform tail on $(p_\alpha, 1)$. □

5.2.2 The UNSAT phase

Assume $\alpha > \alpha_c(r)$. In this regime the ridgeless limit does not escape to infinite norm. By Proposition 11, the minimizers ν_λ converge to the unique finite minimizer ν_0 of the ridgeless reduced potential. The next proposition records the corresponding finite-scale profiles. We omit the proof, since the statement is obtained by specializing the scalar channel from Section 4.3 at the ridgeless minimizer.

Proposition 13 (UNSAT score, margin, and percentile profiles)**.** *Let χ_0 be the maximizer in the scalar saddle problem corresponding to ν_0, and define*

$$c_0 := c_r(\nu_0).$$

Then the limiting true score is the finite-scale scalar random variable

$$S_0 := S_{\nu_0,\chi_0} = c_0 + \mathrm{Prox}_{\chi_0}\big(\nu_0 G - c_0\big), \qquad G \sim N(0,1).$$

The competing scores remain on the same finite scale, with limiting Gaussian scale ν_0. The limiting TAM margin and fitting loss are

$$M_0 := S_0 - c_0, \qquad L_0 := \mathbb{E}\big[\ell(M_0)\big] > 0.$$

Finally, the limiting percentile rank of the true score is

$$\Omega_0 := \Phi\left(\frac{S_0}{\nu_0}\right).$$

Remark 5. *Unlike the phase boundary and the SAT normalized profiles, these UNSAT quantities generally retain a dependence on the fixed value of β through the scalar functions $c_r(\nu)$ and $S_{\nu,\chi}$.*

5.3 Numerical validation

We now give finite-dimensional numerical checks of the scalar predictions. The most direct prediction is the phase boundary of Proposition 11: as (α, r) crosses the curve $\alpha = \alpha_c(r)$, the optimized fitting loss should separate from zero and the norm scale $\nu = \|\widehat{W}\|_{\mathsf{F}}/d$ should change sharply. Figure 2 tests this prediction in a two-dimensional sweep over the load $\alpha = n/d^2$ and the tail scale $r = k/(n-1)$. The empirical transition extracted from the simulations is closely aligned with the theoretical curve (69), both in the loss landscape and in the growth of the norm scale.

The phase diagram gives a global view of the transition. We next look more locally, along a one-dimensional slice in α (with $r = 0.15$), and compare the finite-dimensional optimizer with the scalar predictions on both sides of the transition. This second comparison provides more information: it tests not only the location of the phase boundary, but also the predicted loss, norm scale, margin profile, and percentile distribution.

Figure 3 shows good agreement between the scalar theory and the finite-dimensional simulations. The upper panels compare two scalar observables: the fitting loss and the inverse norm scale $1/\nu$. The loss remains close to zero throughout the SAT phase and becomes positive beyond the predicted critical load. At the same transition, $1/\nu$ collapses toward zero on the SAT side, while it remains positive in the UNSAT phase, as predicted by Proposition 11.

The lower panels compare distributional predictions in the SAT phase. The empirical normalized margin distribution follows the limiting law (75), with the visible rounding caused by finite dimension. The percentile panel is particularly informative. The TAM prediction (77) has a hard lower edge at $p_\alpha = \Phi(\rho_\alpha)$, meaning that in the ridgeless SAT limit the true item is never predicted to fall below this percentile among its competitors. The finite-dimensional curve tracks this edge closely. The same panel also shows the correlation-memory heuristic benchmark (21); compared with this non-optimized baseline, the TAM optimizer shifts the signal percentiles sharply toward the upper tail, reflecting the improvement obtained by learning the memory matrix for the listwise objective.

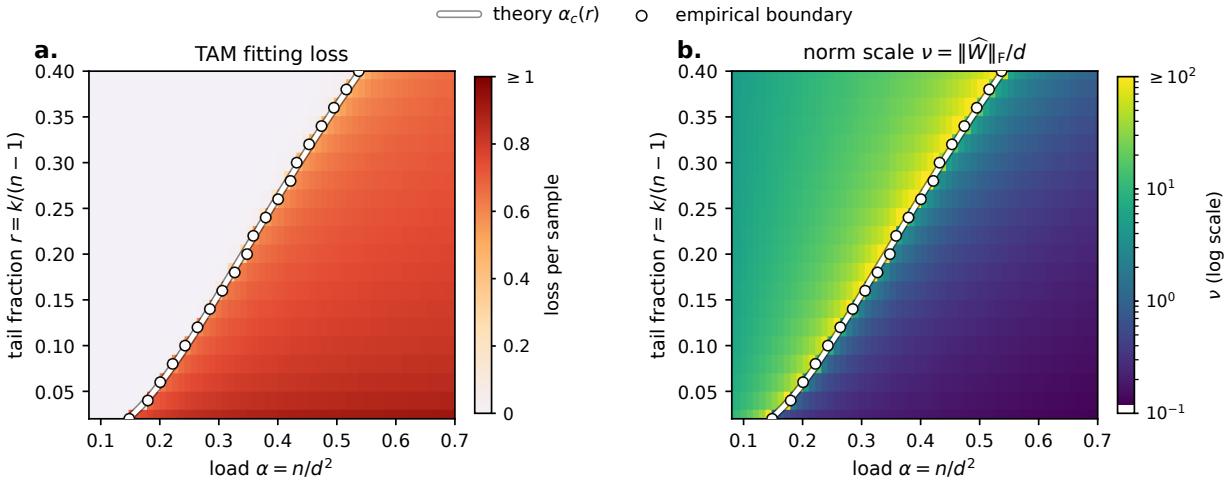

Figure 2: Finite-dimensional validation of the predicted phase boundary. The solid curve is the theoretical threshold $\alpha_c(r)$ from (69), and open circles show empirical boundary estimates from the numerical sweep. **a.** Optimized TAM fitting loss per sample, after subtracting the ridge contribution; colors are clipped at one. **b.** Norm scale $\nu = \|\widehat{W}\|_F/d$ on a logarithmic color scale; colors are clipped at 100. Both panels use $d = 400$ and small ridge regularization $\lambda = 1e-6$. The observed transition is sharply aligned with the scalar prediction.

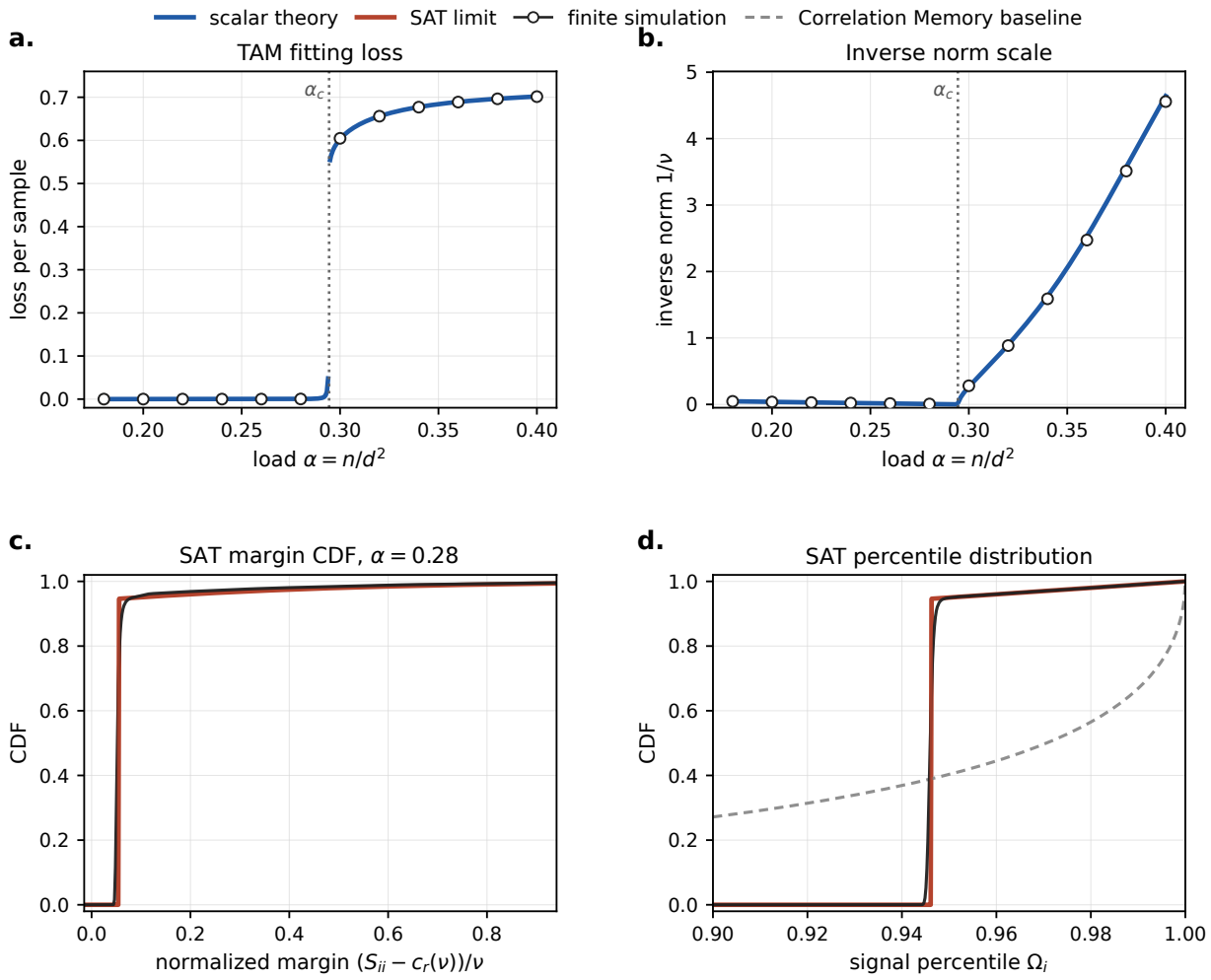

Figure 3: Scalar theory compared with finite-dimensional simulations along a one-dimensional slice at $r = 0.15$. **a.** Optimized TAM fitting loss per sample as a function of the load α, after subtracting the ridge contribution. **b.** Inverse norm scale $1/\nu$, showing the collapse of $1/\nu$ in the SAT phase and its positive value in the UNSAT phase. In both upper panels, the solid curve is obtained from the scalar prediction (43), with the norm and loss read off through (46) and (47), at $\beta = 30$ and ridge $\lambda = 10^{-7}$. Open circles are finite-dimensional simulations at $d = 600$, and the dashed vertical line marks the ridgeless threshold $\alpha_c(r)$. **c.** Empirical CDF of the normalized TAM margin $(S_{ii} - c_r(\nu))/\nu$ in the SAT phase, compared with (75). **d.** Empirical CDF of the signal percentile Ω_i, compared with (77) and the correlation-memory heuristic (21). The lower panels use the representative SAT load $\alpha = 0.28$.

6 Conclusion

The capacity of a linear associative memory depends as much on how it is read as on how many parameters it has. Under isotropic Gaussian embeddings, top-1 retrieval has a sharp threshold: the critical value of $d^2/(n \log n)$ is 2. If this ratio stays above 2, our adaptive sparse-repair construction retrieves all n associations with high probability; if it stays below 2, no data-dependent linear memory can do so. The logarithmic factor is therefore not a weakness of a particular storage rule. It is the price of requiring every query to beat its strongest competitor. The critical value was conjectured in an earlier version of this paper and independently in the concurrent work of Giorlandino et al. [2026].

The picture changes when the correct target need only remain within a fixed fraction of the candidates. TAM captures this regime by comparing the signal with the average of its $k = \lceil r(n-1) \rceil$ strongest competitors. For fixed $r \in (0,1)$, the list size is $k = \Theta(n)$, so this is a percentile-rank guarantee rather than a sublinear shortlist. At any fixed finite smoothing parameter and positive ridge, empirical risk minimization with the smoothed TAM objective has a high-dimensional limit governed by two scalar parameters. These parameters determine the limiting score, margin, and percentile laws. Taking the high-dimensional limit first and then sending the ridge parameter to zero gives the critical load $\alpha_c(r)$ for $\alpha = n/d^2$. Below this load, the average fitting loss vanishes and the exact TAM criterion holds for all but a vanishing fraction of keys; above it, the limiting average fitting loss remains positive.

The TAM proof required a leave-one-out argument adapted to the strong dependence among the losses. Removing one association does two things at once: it changes that association's signal term, and it changes its role as a competitor for every other query. Our coupled leave-one-out expansion tracks both effects and reduces the leading response to an inverse Hessian applied to a single rank-one feature. The accompanying stability, spectral, and self-averaging estimates then lead to the scalar theory. This mechanism is not specific to TAM and may also be useful for other matrix-valued empirical-risk problems in which every sample appears throughout the loss.

Between the two retrieval regimes lies a natural open question: what happens when the list fraction r tends to zero with n? This limit should clarify how the quadratic TAM scale crosses over to the extra logarithmic cost of top-1 retrieval. It also remains to prove the second-order correction proposed in Conjecture 1, which appears to capture the slow finite-size drift, and to understand how much of the theory persists beyond isotropic Gaussian embeddings.

Disclosure of AI Assistance

We used AI tools, specifically ChatGPT 5.5 and 5.6, in this work. The problem formulations and central ideas—including the overall approach to top-1 retrieval, the Tail-Average Margin criterion, and the leave-one-out framework for its analysis—originated with the authors. For top-1 retrieval, the authors first established nonmatching achievability and impossibility bounds, based respectively on CMM and a trace comparison using uniform Gaussian anticoncentration, and reported them in an earlier version of this paper. Starting from these arguments, iterative collaboration with AI tools helped sharpen both sides to the exact constant 2. Building on our earlier CMM analysis, the sharp achievability argument augments CMM with an adaptive sparse correction that repairs the exceptional low-margin queries. This construction, formalized in Proposition 2, drew inspiration from multiscale majority and deficit-repair constructions for random perceptrons [Kim and

Roche, 1998, Abbe et al., 2022, Ding and Sun, 2025]; AI tools assisted in refining the argument. For the converse, the dual-certificate approach was motivated by classical separation and perceptron methods [Wendel, 1962, Cover, 1965], and AI tools assisted substantially in completing the technical arguments. We view this as an encouraging example of how human–AI collaboration can contribute to mathematical research.

The TAM formulation and leave-one-out analysis, including the rank-one reduction in Proposition 3 and the resulting scalar characterization, were developed by the authors and reported in an earlier version of this paper. AI tools assisted in completing and checking several supporting technical arguments, particularly the leave-one-out stability estimates in Proposition 4 and the spectral and self-averaging concentration estimates in Proposition 5. AI tools also assisted with numerical experiments and with editing and polishing the manuscript. The authors carefully verified all results and take full responsibility for the content of this paper.

References

Emmanuel Abbe, Shuangping Li, and Allan Sly. Binary perceptron: efficient algorithms can find solutions in a rare well-connected cluster. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 860–873, 2022. doi: 10.1145/3519935.3519975.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: part 3.3, knowledge capacity scaling laws, 2024. URL https://arxiv.org/abs/2404.05405.

S.-I. Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, C-21(11):1197–1206, 1972. doi: 10.1109/T-C.1972.223477.

Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014. doi: 10.1093/imaiai/iau005. URL https://doi.org/10.1093/imaiai/iau005.

James A. Anderson. A simple neural network generating an interactive memory. *Mathematical Biosciences*, 14(3):197–220, 1972. ISSN 0025-5564. doi: 10.1016/0025-5564(72)90075-2.

Benjamin Aubin, Florent Krzakala, Yue M. Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization. In *Advances in Neural Information Processing Systems*, 2020.

Afonso S Bandeira, Kevin Lucca, Petar Nizic-Nikolac, and Ramon Van Handel. Matrix chaos inequalities and chaos of combinatorial type. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 795–805, 2025.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. doi: 10.1073/pnas.1802705116.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: a memory viewpoint, 2023. URL https://arxiv.org/abs/2306.00802.

Trenton Bricken and Cengiz Pehlevan. Attention approximates sparse distributed memory. In *Advances in Neural Information Processing Systems*, 2021.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories, 2024a. URL https://arxiv.org/abs/2310.02984.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent, 2024b. URL https://arxiv.org/abs/2402.18724.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136. ACM, 2007. doi: 10.1145/1273496.1273513.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3): 326–334, 1965. doi: 10.1109/PGEC.1965.264137.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017.

Jian Ding and Nike Sun. Capacity lower bound for the Ising perceptron. *Probability Theory and Related Fields*, 193:627–715, 2025. doi: 10.1007/s00440-025-01364-x.

Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1–2):95–175, 2018. doi: 10.1007/s00440-016-0754-9.

Noureddine El Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013. doi: 10.1073/pnas.1307842110.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Peter Elias. List decoding for noisy channels. Technical Report 335, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1957.

László Erdős and Horng-Tzer Yau. *A dynamical approach to random matrix theory*, volume 28 of *Courant Lecture Notes in Mathematics*. American Mathematical Society, 2017.

Viola Folli, Marco Leonetti, and Giancarlo Ruocco. On the maximum storage capacity of the Hopfield model. *Frontiers in Computational Neuroscience*, 10:144, 2017. ISSN 1662-5188. doi: 10.3389/fncom.2016.00144.

E. Gardner. Maximum storage capacity in neural networks. *Europhysics Letters*, 4(4):481–485, aug 1987. doi: 10.1209/0295-5075/4/4/016.

E. Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, 1988. doi: 10.1088/0305-4470/21/1/030.

Samuel J. Gershman, Ila Fiete, and Kazuki Irie. Key-value memory in the brain, 2025. URL https://arxiv.org/abs/2501.02950.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories, 2021. URL https://arxiv.org/abs/2012.14913.

Alessio Giorlandino, Sebastian Goldt, and Antoine Maillard. Factual recall in linear associative memories: sharp asymptotics and mechanistic insights, 2026. arXiv preprint.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines, 2014. URL https://arxiv.org/abs/1410.5401.

Venkatesan Guruswami. Algorithmic results in list decoding. *Foundations and Trends in Theoretical Computer Science*, 2(2):107–195, 2007. doi: 10.1561/0400000007.

Stephen Hobson. *Correlation matrix memories: improving performance for capacity and generalisation*. PhD thesis, University of York, September 2011. URL https://etheses.whiterose.ac.uk/id/eprint/2211/.

J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554–8, April 1982. doi: 10.1073/pnas.79.8.2554.

J.J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81:3088–92, May 1984. doi: 10.1073/pnas.81.10.3088.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Do LLMs dream of elephants (when told not to)? Latent concept association and associative memory in transformers, 2024. URL https://arxiv.org/abs/2406.18400.

Pentti Kanerva. *Sparse distributed memory*. MIT Press, Cambridge, MA, 1988.

Ryotaro Kawata and Taiji Suzuki. Transformers as measure-theoretic associative memory: a statistical perspective and minimax optimality, 2026. URL https://arxiv.org/abs/2602.01863.

Jeong Han Kim and James R. Roche. Covering cubes by random half cubes, with applications to binary neural networks. *Journal of Computer and System Sciences*, 56(2):223–252, 1998. doi: 10.1006/jcss.1997.1560.

Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D. Lee. Sharp capacity scaling of spectral optimizers in learning associative memory, 2026. URL https://arxiv.org/abs/2603.26554.

Teuvo Kohonen. Correlation matrix memories. *IEEE Transactions on Computers*, C-21:353–359, 1972. doi: 10.1109/TC.1972.5008975.

B. Kosko. Bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):49–60, 1988. doi: 10.1109/21.87054.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition, 2016. URL https://arxiv.org/abs/1606.01164.

Lihua Lei, Peter J. Bickel, and Noureddine El Karoui. Asymptotics for high dimensional regression M-estimates: fixed design results. *Probability Theory and Related Fields*, 172(3–4):983–1079, 2018. doi: 10.1007/s00440-017-0824-7.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7):077301, February 2024. doi: 10.1103/PhysRevLett.132.077301.

Robert J. McEliece, Edward C. Posner, Eugene R. Rodemich, and Santosh S. Venkatesh. The capacity of the hopfield associative memory. *IEEE Transactions on Information Theory*, 33: 461–482, 1987.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin Glass Theory and Beyond*. World Scientific, 1987. doi: 10.1142/0271.

Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pages 6874–6883, 2020.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal Hopfield networks: a general framework for single-shot associative memory models, 2022. URL https://arxiv.org/abs/2202.04557.

Kaoru Nakano. Associatron-A model of associative memory. *IEEE Transactions on Systems, Man Cybernetics*, SMC-2:380–388, July 1972. doi: 10.1109/TSMC.1972.4309133.

Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in Transformers via associative memories, 2024. URL https://arxiv.org/abs/2412.06538.

Gilles Pisier. Probabilistic methods in the geometry of banach spaces. In *Probability and Analysis*, volume 1206 of *Lecture Notes in Mathematics*, pages 167–241. Springer, 2006.

R. L. Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41(3/4):351–360, 1954. doi: 10.1093/biomet/41.3-4.351.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need, 2021. URL https://arxiv.org/abs/2008.02217.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. doi: 10.1016/S0378-4266(02)00271-6.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019. doi: 10.1073/pnas.1810420116.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, 2017.

Nuri Mert Vural, Alberto Bietti, Mahdi Soltanolkotabi, and Denny Wu. Learning to recall with Transformers beyond orthogonal embeddings, 2026. URL https://arxiv.org/abs/2603.15923.

J. G. Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11:109–112, 1962. doi: 10.7146/math.scand.a-10655.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks, 2015. URL https://arxiv.org/abs/1410.3916.

D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222:960–962, 1969. doi: 10.1038/222960a0.

A Correlation matrix memory: score law and retrieval

This appendix gives the CMM score-law and retrieval calculations used in Section 2.1. Section A.1 derives the exact one-query law underlying (9), and Section A.2 proves Proposition 1 using a conditional moment-generating function.

Under the Gaussian model (2), let $U = [u_1 \ \cdots \ u_n]$ and $V = [v_1 \ \cdots \ v_n]$ in $\mathbb{R}^{d\times n}$; their columns are mutually independent $\mathcal{N}(0, I_d/d)$ vectors. For the correlation matrix memory $W_{\mathsf{CMM}} = UV^\top$ from (6), throughout this appendix we write

$$S := U^\top W_{\mathsf{CMM}} V = U^\top U V^\top V \in \mathbb{R}^{n\times n}, \qquad S_{j,i} = s_{j,i} = u_j^\top W_{\mathsf{CMM}} v_i.$$

Thus column i contains all decoder scores for query v_i: $s_{i,i}$ is its signal score, while $(s_{j,i})_{j\neq i}$ are its competitor scores.

Moreover, $\chi(k)$ denotes the chi distribution with k degrees of freedom.

A.1 Law of one CMM score column

Lemma 3 (Exact law of one CMM score column)**.** *Assume $n \geq 3$ and $d \geq 2$. Let $e_1^{(n)}$ be the first standard basis vector in $\mathbb{R}^n$ and write*

$$\begin{bmatrix}\theta\\ \xi\end{bmatrix} := Se_1^{(n)}, \qquad \theta = s_{1,1}, \qquad \xi = (s_{2,1},\dots,s_{n,1})^\top \in \mathbb{R}^{n-1}.$$

Let $a \sim \mathcal{N}(0,1)$, $g \sim \mathcal{N}(0, I_{n-1})$, and

$$\rho_1, \rho_2 \sim \frac{\chi(d)}{\sqrt{d}}, \quad \rho_3 \sim \frac{\chi(n-1)}{\sqrt{n}}, \quad \rho_4 \sim \frac{\chi(d-1)}{\sqrt{d}}, \quad \rho_5 \sim \frac{\chi(n-2)}{\sqrt{n}}.$$

Assume that $a, g, \rho_1, \dots, \rho_5$ are mutually independent. Then

$$\begin{bmatrix}\theta\\ \xi\end{bmatrix} \overset{(d)}{=} \rho_1 \begin{bmatrix}\rho_2\tau\\ \frac{\sigma}{\|g\|/\sqrt{n}}\, g\end{bmatrix},$$

where

$$\tau = \rho_1\rho_2 + \sqrt{\frac{n}{d^2}}\,\rho_3\, a, \qquad \sigma = \left[\frac{(a\tau/\sqrt{n} + \rho_3\rho_4^2)^2}{d} + \frac{\tau^2\rho_5^2}{d} + \frac{n}{d^2}\,\rho_3^2\rho_4^2\rho_5^2\right]^{1/2}. \tag{84}$$

Proof. We derive the distributional representation (84) for the first CMM score column. It is easy to see that, for any deterministic orthogonal matrix $O \in O(n-1)$,

$$\begin{bmatrix}\theta\\ \xi\end{bmatrix} \overset{(d)}{=} \begin{bmatrix}\theta\\ O\xi\end{bmatrix}. \tag{85}$$

To see this, writing

$$Q := \begin{bmatrix}1 & 0\\ 0 & O\end{bmatrix} \in O(n),$$

we have $(UQ^\top, VQ^\top) \overset{(d)}{=} (U, V)$. Defining

$$w := U^\top U V^\top V e_1^{(n)}, \qquad w' := (UQ^\top)^\top (UQ^\top)(VQ^\top)^\top (VQ^\top) e_1^{(n)},$$

it holds that $w' = Qw$ since $Q^\top e_1^{(n)} = e_1^{(n)}$. Since $w' \overset{(d)}{=} w$, we have $w \overset{(d)}{=} Qw$ and so (85) holds. From rotational invariance, it follows that

$$\begin{bmatrix} \theta \\ \xi \end{bmatrix} \overset{(d)}{=} \begin{bmatrix} \theta \\ \|\xi\| \frac{g}{\|g\|} \end{bmatrix}, \tag{86}$$

where $g \sim \mathcal{N}(0, I_{n-1})$ is a Gaussian vector independent of θ and $\|\xi\|$.

Let $v_1 \in \mathbb{R}^d$ denote the first column of V. It is straightforward to check that

$$V^\top V e_1^{(n)} \overset{(d)}{=} \|v_1\| \begin{bmatrix} \|v_1\| \\ z \end{bmatrix},$$

where $z \sim \mathcal{N}(0, I_{n-1}/d)$ is independent of v_1. We may also write

$$U = \begin{bmatrix} u_1 & H_{u_1} M H_z^\top \end{bmatrix} = H_{u_1} \begin{bmatrix} \|u_1\| e_1^{(d)} & M H_z^\top \end{bmatrix},$$

where $H_{u_1} \in \mathbb{R}^{d \times d}$ and $H_z \in \mathbb{R}^{(n-1)\times(n-1)}$ denote orthogonal matrices whose first columns are aligned with u_1 and z respectively, and

$$M = \begin{bmatrix} \tilde{a} & q^\top \\ p & G \end{bmatrix} \in \mathbb{R}^{d \times (n-1)}, \qquad \tilde{a} \in \mathbb{R},\ q \in \mathbb{R}^{n-2},\ p \in \mathbb{R}^{d-1},\ G \in \mathbb{R}^{(d-1)\times(n-2)}$$

is a Gaussian matrix with i.i.d. $\mathcal{N}(0, 1/d)$ entries. Substituting the above representations in

$$U^\top U V^\top V e_1^{(n)} = \begin{bmatrix} \theta \\ \xi \end{bmatrix}$$

yields

$$\theta = \|v_1\| \|u_1\| \tau, \qquad \text{where} \qquad \tau := \|u_1\| \|v_1\| + \tilde{a} \|z\|.$$

Moreover,

$$\xi = \|v_1\| H_z \begin{bmatrix} \tilde{a}\tau + \|z\| \|p\|^2 \\ \tau q + \|z\| G^\top p \end{bmatrix} \overset{(d)}{=} \|v_1\| H_z \begin{bmatrix} \tilde{a}\tau + \|z\| \|p\|^2 \\ \sqrt{\tau^2 + \|z\|^2 \|p\|^2}\, \tilde{z} \end{bmatrix}, \tag{87}$$

where $\tilde{z} \sim \mathcal{N}(0, I_{n-2}/d)$ is independent of (τ, z, p). To reach the claim of the lemma, set

$$\rho_1 = \|v_1\|, \quad \rho_2 = \|u_1\|, \quad \rho_3 = \sqrt{\frac{d}{n}} \|z\|, \quad \rho_4 = \|p\|, \quad \rho_5 = \sqrt{\frac{d}{n}} \|\tilde{z}\|.$$

Moreover, the standard Gaussian variable a in the statement of the lemma is $\tilde{a}\sqrt{d}$. The claim of the lemma then follows readily by applying (86) with (87). □

Remark 6 (Gaussian consequence of the one-column law)**.** *Suppose that $n, d \to \infty$ with $n/d^2 \to \alpha \in (0, \infty)$. Then, for every fixed m,*

$$(\theta, \xi_1, \dots, \xi_m) \Longrightarrow \big(1 + \sqrt{\alpha}\, G_0, \sqrt{\alpha}\, G_1, \dots, \sqrt{\alpha}\, G_m\big),$$

where $G_0, \dots, G_m$ are independent standard Gaussians. To see this, apply Lemma 3: each ρ_ℓ, $\ell \in [5]$, and $\|g\|/\sqrt{n}$ converge in probability to 1, whereas $\sigma \to \sqrt{\alpha}$.

The same representation yields a more general standardized formulation. Set $\alpha_n := n/d^2$. If $d/n \to 0$, then, for every fixed m,

$$\left(\frac{\theta - 1}{\sqrt{\alpha_n}}, \frac{\xi_1}{\sqrt{\alpha_n}}, \dots, \frac{\xi_m}{\sqrt{\alpha_n}}\right) \Longrightarrow (G_0, G_1, \dots, G_m),$$

where $G_0, G_1, \dots, G_m$ are independent standard Gaussians, and the convergence is joint in $\mathbb{R}^{m+1}$.

Thus, for a fixed query, the signal score and any fixed number of competitor scores have precisely the leading-order Gaussian form suggested by (9), including their asymptotic independence. This finite-dimensional consequence does not by itself control the maximum over all $n - 1$ competitors; the uniform tail estimate needed for perfect retrieval is proved in Appendix A.2.

A.2 CMM top-1 retrieval above constant 8

Proof of Proposition 1. For an ordered pair $i \neq j$, write

$$D_{j,i} := s_{i,i} - s_{j,i}, \qquad x_{j,i} := u_i - u_j.$$

The CMM succeeds precisely when $D_{j,i} > 0$ for every ordered pair. We bound the lower tail of one such margin directly.

Exact conditional moment-generating function. Fix $i \neq j$ and condition on (u_i, u_j, v_i). With $x = x_{j,i}$ and $y = v_i$, expanding $W_{\mathsf{CMM}} = \sum_{k=1}^n u_k v_k^\top$ gives

$$D_{j,i} = a_{j,i} + b_{j,i}\langle v_j, y\rangle + \sum_{k \notin \{i,j\}} \langle x, u_k\rangle\langle v_k, y\rangle, \tag{88}$$

where

$$a_{j,i} := \langle x, u_i\rangle \|y\|^2, \qquad b_{j,i} := \langle x, u_j\rangle.$$

The terms after $a_{j,i}$ in (88) are conditionally independent and centered. Moreover, $\langle v_j, y\rangle \sim \mathcal{N}(0, \|y\|^2/d)$, while for $k \notin \{i, j\}$ the two factors $\langle x, u_k\rangle$ and $\langle v_k, y\rangle$ are independent centered Gaussians with variances $\|x\|^2/d$ and $\|y\|^2/d$. Thus, if $R_{j,i} := D_{j,i} - a_{j,i}$, then

$$\mathbb{E}\left[e^{\lambda R_{j,i}} \mid u_i, u_j, v_i\right] = \exp\left(\frac{\lambda^2 b_{j,i}^2 \|y\|^2}{2d}\right)\left(1 - \frac{\lambda^2 \|x\|^2 \|y\|^2}{d^2}\right)^{-(n-2)/2} \tag{89}$$

whenever $\lambda^2 \|x\|^2 \|y\|^2 < d^2$. This identity uses $\mathbb{E}e^{tGH} = (1 - t^2\sigma_G^2\sigma_H^2)^{-1/2}$ for independent centered Gaussians G, H.

Uniform concentration. The condition $\liminf d^2/(n\log n) > 8$ implies $d/\log n \to \infty$. Standard chi-square and Gaussian tail bounds, followed by a union bound, therefore give a deterministic sequence $\varepsilon_n \to 0$ such that, with probability $1 - o(1)$,

$$\max_{i\le n}\Big(\big|\|u_i\|^2 - 1\big| \vee \big|\|v_i\|^2 - 1\big|\Big) \le \varepsilon_n,$$
$$\max_{i\neq j} |\langle u_i, u_j\rangle| \le \varepsilon_n.$$

Call this event $\mathcal{H}_n$, and let $\mathcal{H}_{j,i}$ denote the corresponding restrictions involving only (u_i, u_j, v_i). On $\mathcal{H}_{j,i}$, uniformly over $i \neq j$,

$$a_{j,i} = 1 + o(1), \qquad b_{j,i} = -1 + o(1), \qquad \|x_{j,i}\|^2\|v_i\|^2 = 2 + o(1). \tag{90}$$

It remains to apply Chernoff's bound to (89). We do so along an arbitrary subsequence. Passing to a further subsequence, either $d/n \to 0$, or d/n is bounded below by a positive constant.

Suppose first that $d/n \to 0$, and set $\alpha_n = n/d^2$. Conditional on $(u_i, u_j, v_i) \in \mathcal{H}_{j,i}$, choose $\lambda = a_{j,i}/(2\alpha_n)$. Then $\lambda/d = O(d/n) = o(1)$, and $-\log(1-t) = t + O(t^2)$ in (89) yields, uniformly over $i \neq j$,

$$\log \mathbb{E}\left[e^{-\lambda R_{j,i}} \mid u_i, u_j, v_i\right] = (1 + o(1))\lambda^2\alpha_n.$$

Here the Gaussian term in (89) is negligible because $1/d = o(\alpha_n)$. Since $R_{j,i}$ is symmetric, Chernoff's bound and (90) give

$$\sup_{(u_i,u_j,v_i)\in\mathcal{H}_{j,i}} \mathbb{P}\left(D_{j,i} \le 0 \mid u_i, u_j, v_i\right) \le \exp\left(-\frac{1-o(1)}{4\alpha_n}\right). \tag{91}$$

Choose $\eta > 0$ so that $d^2/(n\log n) \ge 8 + \eta$ eventually. The exponent on the right-hand side of (91) is then at least $(2 + c_\eta)\log n$ for some $c_\eta > 0$ and all sufficiently large n.

Suppose instead that $d/n \ge c_0 > 0$. Taking $\lambda = \kappa d$ in (89), with $\kappa > 0$ sufficiently small depending only on c_0, gives

$$\sup_{(u_i,u_j,v_i)\in\mathcal{H}_{j,i}} \mathbb{P}\left(D_{j,i} \le 0 \mid u_i, u_j, v_i\right) \le e^{-c_1 d} \tag{92}$$

for some $c_1 > 0$. Indeed, the Chernoff exponent contributes $-\kappa d(1 + o(1))$, whereas the logarithm of the moment-generating function is $O(\kappa^2(d+n)) = O(\kappa^2 d)$. Since $d/\log n \to \infty$, the right-hand side of (92) is $o(n^{-2})$.

Finally, a union bound over the $n(n-1)$ ordered pairs gives

$$\begin{aligned}\mathbb{P}\big(\mathsf{PR}(W_{\mathsf{CMM}}; U, V)^c\big) &\le \mathbb{P}(\mathcal{H}_n^c) + \sum_{i\neq j}\mathbb{P}\big(D_{j,i} \le 0, \mathcal{H}_{j,i}\big)\\ &= o(1)\end{aligned}$$

along either type of further subsequence. Every subsequence admits such a further subsequence, so the same convergence holds along the original sequence. □

B Taylor expansion of the interior remainder

This appendix carries out the first-order Taylor expansion of the interior remainder $\mathcal{R}^{\setminus h}$ defined in (32) of Section 4.2, derives the explicit form of the residual $\hat{E}$ appearing in the main identity (35), and establishes the entry-size bounds (36) used there. All notation is as in Section 4.2: $(\widehat{W}, \widehat{\mu})$ and $(\widehat{W}^{\setminus h}, \widehat{\mu}^{\setminus h})$ are the full and leave-one-out minimizers, $\Delta S := U^\top(\widehat{W} - \widehat{W}^{\setminus h})V$, $(\Delta\mu)_i := \hat{\mu}_i - \mu_i^{\setminus h}$; a hat denotes a quantity evaluated at $(\widehat{W}, \widehat{\mu})$ and a superscript $\setminus h$ one evaluated at $(\widehat{W}^{\setminus h}, \widehat{\mu}^{\setminus h})$. The analysis is conditional on the LOO postulates (A1)–(A3).

B.1 Taylor identities

We record three Taylor identities around the LOO point. Throughout, for a scalar quantity $X(W, \mu)$ we write $\Delta X := X(\widehat{W}, \widehat{\mu}) - X(\widehat{W}^{\setminus h}, \widehat{\mu}^{\setminus h})$.

Distractor sigmoid. Since $q_{j,i} = \phi'_\beta(s_{j,i} - \mu_i) = \sigma(\beta(s_{j,i} - \mu_i))$ and $\sigma'(t) = \sigma(t)(1 - \sigma(t))$, a Taylor expansion in the single variable $s_{j,i} - \mu_i$ gives, for $j, i \neq h$ and $j \neq i$,

$$\begin{aligned} \hat{q}_{j,i} - q_{j,i}^{\setminus h} &= k\, \widetilde{R}_{j,i}^{\setminus h} \left(\Delta S_{j,i} - (\Delta\mu)_i\right) + \rho_{j,i}^{(q)}, \\ \rho_{j,i}^{(q)} &= \mathcal{O}_\prec\left(\beta^2 \left|\Delta S_{j,i} - (\Delta\mu)_i\right|^2\right) = \mathcal{O}_\prec(\beta^2/d^2), \end{aligned} \tag{93}$$

where $k\, \widetilde{R}_{j,i}^{\setminus h} = \beta\, q_{j,i}^{\setminus h}(1 - q_{j,i}^{\setminus h}) = \phi''_\beta(s_{j,i}^{\setminus h} - \mu_i^{\setminus h})$ and the last bound uses $\max_{j,i \neq h} |\Delta S_{j,i}| = \mathcal{O}_\prec(d^{-1})$ by (A2) and $|\Delta\mu_i| = \mathcal{O}_\prec(n^{-1}) = \mathcal{O}_\prec(d^{-2})$ by (A1).

Margin. For $i \neq h$, the full margin $\hat{m}_i = s_{i,i}(\widehat{W}) - \hat{c}_i$ and the LOO margin $m_i^{\setminus h} = s_{i,i}(\widehat{W}^{\setminus h}) - c_i^{\setminus h}$ differ both through the shifts $(\Delta S_{i,i}, (\Delta\mu)_i)$ and through the missing $j = h$ interferer in the LOO inner sum. Taylor-expanding the inner softplus summation in the column-i score perturbations and peeling off the missing interferer term,

$$\hat{m}_i - m_i^{\setminus h} = -\left(Q_{:,i}^{\setminus h}\right)^\top \Delta S_{:,i} + \left[\,(Q_{:,i}^{\setminus h})^\top \mathbf{1}_n\,\right](\Delta\mu)_i - \tfrac{1}{k}\,\phi_\beta\left(s_{h,i}^{\setminus h} - \mu_i^{\setminus h}\right) + \rho_i^{(m)}, \tag{94}$$

where the first two terms assemble the linear dependence on $(\Delta S_{:,i}, (\Delta\mu)_i)$ (using $\partial m_i / \partial s_{j,i} = -Q_{j,i}$ with the convention $Q_{i,i} = -1$, $Q_{j,i} = q_{j,i}/k$ for $j \neq i$), the third term is the removed-interferer contribution, and $\rho_i^{(m)}$ collects second-order Taylor remainders of the ϕ_β terms and cross contributions. Using the second-order softplus bound $|\phi_\beta(x+\delta) - \phi_\beta(x) - \phi'_\beta(x)\delta| \leq \frac{\beta}{8}\,\delta^2$ together with (A1)–(A2),

$$\begin{aligned} |\rho_i^{(m)}| &\leq \frac{\beta}{8k} \sum_{j \neq i,\, j \neq h} \left(\Delta S_{j,i} - (\Delta\mu)_i\right)^2 + \mathcal{O}_\prec\left(|\Delta S_{h,i}| + |(\Delta\mu)_i|\right) \cdot \tfrac{1}{k} \\ &= \mathcal{O}_\prec(\beta/d^2) + \mathcal{O}_\prec(1/(nd)) = \mathcal{O}_\prec(1/d^2). \end{aligned} \tag{95}$$

Activation. Since $\ell'(t) = -\sigma(-t)$, $\ell''(t) = \sigma'(-t) = a(1-a) = b$, and ℓ''' is uniformly bounded,

$$\begin{aligned} \hat{a}_i - a_i^{\setminus h} &= -\,b_i^{\setminus h}\,(\hat{m}_i - m_i^{\setminus h}) + \rho_i^{(a)}, \\ |\rho_i^{(a)}| &= \mathcal{O}_\prec\left((\hat{m}_i - m_i^{\setminus h})^2\right) = \mathcal{O}_\prec(1/d^2) = \mathcal{O}_\prec(1/n). \end{aligned} \tag{96}$$

B.2 Explicit form of the residual

We now unfold the two cases of $\mathcal{R}^{\setminus h}$ in (32) and identify the linear part with $(M_i^{\setminus h}\Delta S_{:,i})_j$ from (28), so that $\hat{E}_{j,i}$ is what remains.

Off-diagonal case ($j \neq i$, $j, i \neq h$). Using the algebraic identity

$$\hat{a}_i\hat{q}_{j,i} - a_i^{\setminus h}q_{j,i}^{\setminus h} \;=\; a_i^{\setminus h}\,(\hat{q}_{j,i} - q_{j,i}^{\setminus h}) \;+\; (\hat{a}_i - a_i^{\setminus h})\,q_{j,i}^{\setminus h} \;+\; (\hat{a}_i - a_i^{\setminus h})\,(\hat{q}_{j,i} - q_{j,i}^{\setminus h}),$$

and substituting (93), (94), (96),

$$\begin{aligned}\mathcal{R}_{j,i}^{\setminus h} \;&=\; \frac{a_i^{\setminus h}\cdot k\,\widetilde{R}_{j,i}^{\setminus h}}{k}\big(\Delta S_{j,i} - (\Delta\mu)_i\big) \;+\; \frac{q_{j,i}^{\setminus h}}{k}\cdot\big[-b_i^{\setminus h}(\hat{m}_i - m_i^{\setminus h})\big] \;+\; \hat{E}_{j,i}^{(\mathrm{off})}\\ &=\; a_i^{\setminus h}\,\widetilde{R}_{j,i}^{\setminus h}\big(\Delta S_{j,i} - (\Delta\mu)_i\big) \;+\; b_i^{\setminus h}\,Q_{j,i}^{\setminus h}\,(Q_{:,i}^{\setminus h})^\top\Delta S_{:,i} \;+\; \hat{E}_{j,i}^{(\mathrm{off})},\end{aligned}$$

where in the second equality we used $Q_{j,i}^{\setminus h} = q_{j,i}^{\setminus h}/k$ for $j \neq i$, $j \neq h$, and absorbed the $(\Delta\mu)_i$-dependent piece of $\hat{m}_i - m_i^{\setminus h}$ from (94) into $\hat{E}_{j,i}^{(\mathrm{off})}$. The first two terms match $(M_i^{\setminus h}\Delta S_{:,i})_j$ for $j \neq i$ from (28), provided we recognize that the $a_i^{\setminus h}\widetilde{R}$ piece also absorbs the $-a_i^{\setminus h}\widetilde{R}_{j,i}^{\setminus h}(\Delta\mu)_i$ contribution into $\hat{E}$. Collecting the residual pieces,

$$\begin{aligned}\hat{E}_{j,i}^{(\mathrm{off})} \;=\; &-\frac{b_i^{\setminus h}\,q_{j,i}^{\setminus h}}{k^2}\,\phi_\beta\big(s_{h,i}^{\setminus h} - \mu_i^{\setminus h}\big) \;+\; \frac{a_i^{\setminus h}}{k}\,\rho_{j,i}^{(q)} \;-\; a_i^{\setminus h}\,\widetilde{R}_{j,i}^{\setminus h}\,(\Delta\mu)_i\\ &+\; \frac{q_{j,i}^{\setminus h}}{k}\,\big[\,b_i^{\setminus h}(Q_{:,i}^{\setminus h})^\top\mathbf{1}_n\,(\Delta\mu)_i - \rho_i^{(a)}\,\big] \;+\; \frac{(\hat{a}_i - a_i^{\setminus h})(\hat{q}_{j,i} - q_{j,i}^{\setminus h})}{k}.\end{aligned} \tag{97}$$

Diagonal case ($j = i \neq h$). We have $\mathcal{R}_{i,i}^{\setminus h} = a_i^{\setminus h} - \hat{a}_i = b_i^{\setminus h}(\hat{m}_i - m_i^{\setminus h}) - \rho_i^{(a)}$. Substituting (94),

$$\begin{aligned}\mathcal{R}_{i,i}^{\setminus h} \;&=\; b_i^{\setminus h}\cdot\Big[-(Q_{:,i}^{\setminus h})^\top\Delta S_{:,i} \;+\; \big[(Q_{:,i}^{\setminus h})^\top\mathbf{1}_n\big](\Delta\mu)_i \;-\; \tfrac{1}{k}\,\phi_\beta\big(s_{h,i}^{\setminus h} - \mu_i^{\setminus h}\big) \;+\; \rho_i^{(m)}\Big] \;-\; \rho_i^{(a)}\\ &=\; -\,b_i^{\setminus h}\,(Q_{:,i}^{\setminus h})^\top\Delta S_{:,i} \;+\; \hat{E}_{i,i},\end{aligned}$$

where the first term equals $(M_i^{\setminus h}\Delta S_{:,i})_i = b_i^{\setminus h}\,Q_{i,i}^{\setminus h}\,(Q_{:,i}^{\setminus h})^\top\Delta S_{:,i} = -b_i^{\setminus h}\,(Q_{:,i}^{\setminus h})^\top\Delta S_{:,i}$ (using $Q_{i,i}^{\setminus h} = -1$ and $\widetilde{R}_{i,i}^{\setminus h} = 0$), and

$$\hat{E}_{i,i} \;=\; -\,\frac{b_i^{\setminus h}}{k}\,\phi_\beta\big(s_{h,i}^{\setminus h} - \mu_i^{\setminus h}\big) \;+\; b_i^{\setminus h}\big[(Q_{:,i}^{\setminus h})^\top\mathbf{1}_n\big](\Delta\mu)_i \;+\; b_i^{\setminus h}\,\rho_i^{(m)} \;-\; \rho_i^{(a)}. \tag{98}$$

B.3 Entry-size bounds

We now bound each term in (97) and (98) under (A1)–(A3), arriving at the claim (36). Throughout, we use the deterministic bounds $a_i, b_i, q_{j,i} \in [0,1]$ (with $b_i \leq 1/4$) and $|\widetilde{R}_{j,i}^{\setminus h}| \leq \beta/(4k)$.

We first record the only pointwise consequence of (A3) needed below. The μ-stationarity condition for $\Psi^{\setminus h}$ gives $\frac{1}{n-1}\sum_{j\neq i,h}\sigma(\beta(s_{j,i}^{\setminus h} - \mu_i^{\setminus h})) \approx r_d$ for each $i \neq h$. Since the full scores are

bounded by (A3), the LOO scores with $i \neq h$ are bounded by (A2). As $r_d \in (0,1)$, this implies $\mu_i^{\setminus h} = \mathcal{O}_\prec(1)$ uniformly. Therefore, by Lipschitz continuity of the softplus,

$$\left|\phi_\beta\left(s_{j,i}^{\setminus h} - \mu_i^{\setminus h}\right)\right| \;\leq\; \frac{\log 2}{\beta} + \left|s_{j,i}^{\setminus h} - \mu_i^{\setminus h}\right| \;=\; \mathcal{O}_\prec(1) \tag{99}$$

uniformly in (j,i).

Off-diagonal entries. For $j \neq i$, $j, i \neq h$, each term in (97):

(i) $\left|\frac{b_i^{\setminus h} q_{j,i}^{\setminus h}}{k^2}\,\phi_\beta(\cdot)\right| \leq \frac{1}{4k^2}\cdot\mathcal{O}_\prec(1) = \mathcal{O}_\prec(n^{-2})$ by (99);

(ii) $\left|\frac{a_i^{\setminus h}}{k}\,\rho_{j,i}^{(q)}\right| \leq \frac{1}{k}\cdot\mathcal{O}_\prec(\beta^2/d^2) = \mathcal{O}_\prec(\beta^2/(nd^2)) = \mathcal{O}_\prec(n^{-2})$, using $\beta = \mathcal{O}_\prec(1)$ in the retrieval regime, or absorbing β as a constant;

(iii) $\left|a_i^{\setminus h}\,\widetilde{R}_{j,i}^{\setminus h}\,(\Delta\mu)_i\right| \leq \frac{\beta}{4k}\cdot\mathcal{O}_\prec(n^{-1}) = \mathcal{O}_\prec(n^{-2})$ by (A1);

(iv) $\left|\frac{q_{j,i}^{\setminus h}}{k}\,b_i^{\setminus h}(Q_{:,i}^{\setminus h})^\top\mathbf{1}_n\,(\Delta\mu)_i\right| \leq \frac{1}{k}\cdot\frac{1}{4}\cdot\mathcal{O}_\prec(1)\cdot\mathcal{O}_\prec(n^{-1}) = \mathcal{O}_\prec(n^{-2})$, using $|(Q_{:,i}^{\setminus h})^\top\mathbf{1}_n| = |-1 + (1/k)\sum q| = \mathcal{O}_\prec(1)$;

(v) $\left|\frac{q_{j,i}^{\setminus h}}{k}\,\rho_i^{(a)}\right| \leq \frac{1}{k}\cdot\mathcal{O}_\prec(1/n) = \mathcal{O}_\prec(n^{-2})$ by (96);

(vi) $\left|\frac{(\hat{a}_i - a_i^{\setminus h})(\hat{q}_{j,i} - q_{j,i}^{\setminus h})}{k}\right| \leq \frac{1}{k}\cdot\mathcal{O}_\prec(d^{-1})\cdot\mathcal{O}_\prec(d^{-1}) = \mathcal{O}_\prec(1/(nd^2)) = \mathcal{O}_\prec(n^{-2})$, using (93), (96) and $|\hat{m} - m^{\setminus h}| = \mathcal{O}_\prec(d^{-1})$.

Summing, $|\hat{E}_{j,i}^{(\mathrm{off})}| = \mathcal{O}_\prec(n^{-2})$.

Diagonal entries. For $i \neq h$, each term in (98):

(i) $\left|\frac{b_i^{\setminus h}}{k}\,\phi_\beta(\cdot)\right| \leq \frac{1}{4k}\cdot\mathcal{O}_\prec(1) = \mathcal{O}_\prec(n^{-1})$ by (99);

(ii) $\left|b_i^{\setminus h}\,(Q_{:,i}^{\setminus h})^\top\mathbf{1}_n\,(\Delta\mu)_i\right| \leq \frac{1}{4}\cdot\mathcal{O}_\prec(1)\cdot\mathcal{O}_\prec(n^{-1}) = \mathcal{O}_\prec(n^{-1})$ by (A1);

(iii) $\left|b_i^{\setminus h}\,\rho_i^{(m)}\right| \leq \frac{1}{4}\cdot\mathcal{O}_\prec(d^{-2}) = \mathcal{O}_\prec(n^{-1})$ by (95);

(iv) $|\rho_i^{(a)}| = \mathcal{O}_\prec(n^{-1})$ by (96).

Summing, $|\hat{E}_{i,i}| = \mathcal{O}_\prec(n^{-1})$.

This completes the proof of (36) under (A1)–(A3).

C Leave-one-out stability

We use the notation and standing assumptions of Section 4 and Section 4.2: $n/d^2 \to \alpha \in (0,\infty)$, $r \in (0,1)$, and $\beta, \lambda > 0$ are fixed. We prove the three stability inputs (A1)–(A3).

The proof has two main steps. First, after profiling out the threshold variables, we construct a one-step Newton candidate for the full optimizer from the leave-one-out optimizer. Freshness of the held-out pair controls the candidate on all surviving scores, while a residual estimate and strong convexity show that the candidate is $O_\prec(d^{-1})$-close to the true optimizer. Second, the exact threshold equation reduces the sharper threshold estimate to a weighted average of these score perturbations. A second use of freshness gains an additional factor $d^{-1/2}$ in this average, yielding (A1). The remaining score-stability and boundedness conclusions then follow directly from the same finite-deletion comparison.

C.1 Leave-one-out expansion and Newton comparison

For $J \subset [n]$ with $|J| > k$, define

$$\mathcal{T}_J(z) := \min_{\mu\in\mathbb{R}} \left\{ \mu + \frac{1}{k}\sum_{j\in J} \phi_\beta(z_j - \mu) \right\}.$$

Its unique minimizer $\mu_J(z)$ satisfies

$$\sum_{j\in J} q_j = k, \qquad q_j = \sigma\big(\beta(z_j - \mu_J(z))\big), \qquad p_j = \frac{q_j}{k}, \tag{100}$$

so $0 < p_j \le k^{-1}$, $\sum_j p_j = 1$, and $\|p\|_2 \le k^{-1/2}$.

The next lemma collects the two inputs used throughout the appendix: the first group of estimates is deterministic profiled-tail calculus, while the last group supplies the Gaussian bulk curvature and delocalization needed for uniform leave-one-out bounds.

Lemma 4 (Profiled-tail identities and bulk geometry)**.** *The map $\mathcal{T}_J$ is smooth, convex, and translation equivariant with $\nabla\mathcal{T}_J = p$. Put $w_j = (\beta/k)q_j(1-q_j)$, $D = \mathrm{diag}(w_j)$, and $s = \sum_j w_j$. Then*

$$\frac{\partial\mu_J}{\partial z_\ell} = \frac{w_\ell}{s}, \qquad \nabla^2\mathcal{T}_J(z) = B_J(z) := D - \frac{(D\mathbf{1})(D\mathbf{1})^\top}{\mathbf{1}^\top D\mathbf{1}}, \tag{101}$$

with

$$0 \preceq B_J(z) \preceq D \preceq \frac{\beta}{4k} I. \tag{102}$$

Consequently, for $H = \|h\|_\infty$,

$$\left|\mathcal{T}_J(z+h) - \mathcal{T}_J(z) - p^\top h\right| \le \frac{\beta}{8k}\|h\|_2^2, \tag{103}$$

$$\|p(z+h) - p(z)\|_2 \le \frac{\beta}{4k}\|h\|_2, \tag{104}$$

$$\max_j \left|p_j(z+h) - p_j(z) - (B_J(z)h)_j\right| \le \frac{C_\beta}{k}H^2. \tag{105}$$

Moreover, uniformly under fixed finite deletions, if $z_j = u_j^\top x$, $\rho = \|x\|/\sqrt{d} \le C\sqrt{d}$, *and* $\sum_{j\in J}\sigma(\beta(z_j-\mu)) = k + O(1)$, *then*

$$|\mu| \le C(1+\rho), \qquad \sum_{j\in J} q_j(1-q_j) \ge c\frac{n}{1+\rho}, \qquad \left\| \frac{(q_j(1-q_j))_{j\in J}}{\sum_\ell q_\ell(1-q_\ell)} \right\|_2^2 \le C\frac{1+\rho}{n}. \tag{106}$$

If one competitor is inserted, the old tail weights decrease and their change $r = p^{\mathrm{old}} - p^{\mathrm{new}}$ *obeys*

$$r \ge 0, \qquad \|r\|_2 \le \frac{C}{k}\sqrt{\frac{1+\rho}{n}}. \tag{107}$$

Proof. The envelope theorem gives $\partial_{z_j}\mathcal{T}_J = q_j/k$. Differentiating (100) in direction h gives $\dot\mu = \sum_j w_j h_j / \sum_j w_j$ and $\dot p_j = w_j(h_j - \dot\mu)$, which proves (101)–(102); integration gives (103) and (104). Along $z + th$, $|\dot w_j| \le 2\beta H w_j$ and the normalized weights have derivative at most $4\beta H$ times themselves, so a second differentiation gives (105).

For the bulk estimates, the quantile equation forces a fixed fraction of $u_j^\top x$ to lie on each side of μ up to an $O(1)$ window, while the Gaussian frame bound gives $m^{-1}\sum_j (u_j^\top x)^2 \lesssim \rho^2$; thus $|\mu| \lesssim 1+\rho$. For $g(t) = \sigma(\beta t)(1-\sigma(\beta t))$, uniformly on this range, $\mathbb{E}g(\rho Z - \mu) \gtrsim (1+\rho)^{-1}$. Bernstein's inequality on a d^{-2} net in $(x/\sqrt{d}, \mu)$, followed by the frame bound and a union bound, yields the curvature estimate in (106); ℓ_2 delocalization follows from $\|\pi\|_2^2 \le \|\pi\|_\infty \|\pi\|_1$. Finally, interpolate the new competitor with weight $\vartheta \in [0,1]$. The threshold increases, so every old weight decreases, and differentiating the old weights gives an ℓ_2 derivative bounded by $Ck^{-1}\sqrt{(1+\rho)/n}$; integration proves (107). □

Profiling the full and leave-one-out objectives gives

$$\Psi(W) = \sum_{i=1}^n \mathcal{L}_i(W) + \frac{\lambda n}{2d^2}\|W\|_F^2, \tag{108}$$

$$\Psi^{\setminus h}(W) = \sum_{i\ne h} \mathcal{L}_i^{\setminus h}(W) + \frac{\lambda n}{2d^2}\|W\|_F^2. \tag{109}$$

For one column, with $t_i = \mathcal{T}_J((s_{j,i})_{j\in J})$, $m_i = s_{i,i} - t_i$, $a_i = \sigma(-m_i)$, $b_i = a_i(1-a_i)$, and tail weights $p_i = (p_{j,i})$, set

$$x_i(W) = \left(u_i - \sum_{j\in J} p_{j,i}u_j\right) v_i^\top.$$

Then $\nabla\mathcal{L}_i(W) = -a_i x_i(W)$, and for a perturbation Δ, with $\delta s_i = u_i^\top \Delta v_i$ and $\delta z_i = (u_j^\top \Delta v_i)_{j\in J}$,

$$\big\langle \Delta, \nabla^2\mathcal{L}_i(W)[\Delta]\big\rangle_F = b_i(\delta s_i - p_i^\top \delta z_i)^2 + a_i \delta z_i^\top B_i \delta z_i \ge 0. \tag{110}$$

Thus the profiled objectives are γ_d-strongly convex with $\gamma_d = \lambda n/d^2 \to \lambda\alpha > 0$. We also use repeatedly

$$\|p_i\|_2 \le k^{-1/2}, \qquad \left\|\sum_{j\in J} p_{j,i}u_j\right\| \le \|U\|_{\mathsf{op}} k^{-1/2} = O_\prec(d^{-1/2}). \tag{111}$$

Fix $h \in [n]$ throughout the one-deletion construction. We use the standard Gaussian frame, Khatri–Rao matrix bounds, and fresh-contraction bounds, uniformly over polynomially many fixed finite deletions.

The deterministic upper bound of Lemma 18 applies unchanged after any fixed finite deletion; hence $\left\|\widehat{W}^{\setminus h}\right\|_F + \left\|\widehat{W}\right\|_F = O(d)$.

For $i \neq h$, let $\mathcal{L}_i$ and $\mathcal{L}_i^{\setminus h}$ denote the full and leave-one-out column losses in (108)–(109). The insertion perturbation is

$$\mathcal{P}_h(W) := \sum_{i \neq h} \big(\mathcal{L}_i(W) - \mathcal{L}_i^{\setminus h}(W)\big), \qquad G_h(W) := \nabla \mathcal{P}_h(W).$$

For one fixed $i \neq h$, unadorned quantities refer to $\mathcal{L}_i$, and superscript $\setminus h$ refers to $\mathcal{L}_i^{\setminus h}$. Let

$$\delta t_i = t_i - t_i^{\setminus h} \geq 0, \qquad A_i = a_i - a_i^{\setminus h} \geq 0,$$

let $p_i^{\setminus h}$ and $p_{i,\mathrm{old}}$ be the leave-one-out and full old-coordinate tail weights, and put

$$r_i = p_i^{\setminus h} - p_{i,\mathrm{old}} \geq 0, \qquad p_{h,i} = \frac{q_{h,i}}{k}.$$

Define

$$\eta_{i,\mathrm{old}} = \sum_{j \neq i,h} p_{j,i} u_j.$$

A direct subtraction of the two column gradients gives the exact identity

$$G_h(W) = \sum_{i \neq h} \left[-A_i u_i + A_i \eta_{i,\mathrm{old}} - a_i^{\setminus h} U r_i + a_i p_{h,i} u_h \right] v_i^\top. \tag{112}$$

Here Ur_i denotes the old-target linear combination, with zeros in the removed coordinates.

Lemma 5 (Size of the insertion gap)**.** *Let W satisfy $\|W\|_F = O_\prec(d)$, and set*

$$\rho_i(W) = \frac{\|Wv_i\|}{\sqrt{d}}.$$

Assume additionally that

$$\sum_{i \neq h} s_{h,i}(W)^2 = O_\prec(n).$$

Then

$$\big\|(\delta t_i)_{i \neq h}\big\|_2 = O_\prec(d^{-1}), \qquad \big\|(A_i)_{i \neq h}\big\|_2 = O_\prec(d^{-1}). \tag{113}$$

Proof. Adding one nonnegative softplus term and using the quantile-location bound give

$$0 \leq \delta t_i \leq \frac{1}{k} \phi_\beta\big(s_{h,i}(W) - \mu_i^{\setminus h}(W)\big), \qquad |\mu_i^{\setminus h}| \leq C(1 + \rho_i).$$

Since $\phi_\beta(t) \leq (\log 2)/\beta + |t|$,

$$\sum_{i \neq h} \delta t_i^2 \leq \frac{C}{k^2} \left[n + \sum_{i \neq h} s_{h,i}(W)^2 + \sum_{i \neq h} \rho_i(W)^2 \right].$$

Moreover, $\sum_{i \neq h} \rho_i(W)^2 = d^{-1} \big\|W V_{\setminus h}\big\|_F^2 \leq d^{-1} \|W\|_F^2 \|V\|_{\mathsf{op}}^2 = O_\prec(n)$. Since $k \asymp n \asymp d^2$, this proves the first estimate in (113); the second follows from $0 \leq A_i \leq \delta t_i / 4$. □

Lemma 6 (Insertion-gradient bound)**.** *At* $\widehat{W}^{\setminus h}$,

$$\left\|G_h(\widehat{W}^{\setminus h})\right\|_F = O_\prec(d^{-1/2}).$$

More precisely, in the decomposition (112)*, the first three matrix sums are* $O_\prec(d^{-1})$ *in Frobenius norm, while the final term has the form*

$$u_h\xi_h^\top, \qquad \|\xi_h\| = O_\prec(d^{-1/2}).$$

Proof. Since $\widehat{W}^{\setminus h}$ is independent of u_h, conditional Gaussian quadratic-form bounds and the deterministic Frobenius estimate give

$$\sum_{i\neq h} s_{h,i}(\widehat{W}^{\setminus h})^2 = u_h^\top \widehat{W}^{\setminus h} V_{\setminus h} V_{\setminus h}^\top (\widehat{W}^{\setminus h})^\top u_h = O_\prec(n)$$

uniformly in h. Indeed, the conditional trace scale is at most $d^{-1}\left\|\widehat{W}^{\setminus h}\right\|_F^2 \|V\|_{\mathsf{op}}^2 = O_\prec(n)$. Hence Lemma 5 applies.

For the first sum in (112), the standard Khatri–Rao bound gives

$$\left\|\sum_{i\neq h} A_i u_i v_i^\top\right\|_F \leq O_\prec(1)\|A\|_2 = O_\prec(d^{-1}).$$

For the second, (111) gives $\max_i \left\|\eta_{i,\mathrm{old}}\right\| = O_\prec(d^{-1/2})$; therefore, if Y has columns $A_i\eta_{i,\mathrm{old}}$, then

$$\|Y\|_F = O_\prec(d^{-3/2}), \qquad \left\|YV^\top\right\|_F = O_\prec(d^{-1}).$$

For the third, (107) and the bound $\sum_i(1+\rho_i(\widehat{W}^{\setminus h})) = O_\prec(n)$ imply $\sum_i \|r_i\|_2^2 = O_\prec(d^{-4})$. Thus, for Y with columns $Ur_i, \|Y\|_F^2 \leq \|U\|_{\mathsf{op}}^2 \sum_i \|r_i\|_2^2 = O_\prec(d^{-3})$, and again $\left\|YV^\top\right\|_F = O_\prec(d^{-1})$.

Finally, the fourth sum is $u_h\xi_h^\top$ with

$$\xi_h = \sum_{i\neq h} a_i p_{h,i} v_i, \qquad \|\xi_h\| \leq \|V\|_{\mathsf{op}} \frac{\sqrt{n}}{k} = O_\prec(d^{-1/2}),$$

since $0 \leq a_i p_{h,i} \leq k^{-1}$. This proves the lemma. □

Newton candidate. We now build one candidate for the full optimizer from the leave-one-out optimizer. The implicit scalar equation absorbs the held-out outer loss, and freshness makes the candidate's surviving scores small.

Let $\mathcal{H}_h := \nabla^2\Psi^{\setminus h}(\widehat{W}^{\setminus h})$ and $T_h := \mathcal{H}_h^{-1}$. By (110) and the ridge floor, $\|T_h\|_{\mathsf{op}} \leq \gamma_d^{-1} = O(1)$. The operator is self-adjoint and is measurable with respect to the data excluding (u_h, v_h).

Evaluate the held-out outer loss at $\widehat{W}^{\setminus h}$. Let

$$t_h^{\mathrm{loo}} = \mathcal{T}_{[n]\setminus\{h\}}\big((s_{j,h}(\widehat{W}^{\setminus h}))_{j\neq h}\big), \qquad m_h^{\mathrm{loo}} = s_{h,h}(\widehat{W}^{\setminus h}) - t_h^{\mathrm{loo}},$$

and let $p_{j,h}^{\mathrm{loo}}$ be the corresponding tail weights. Define

$$\eta_h^{\mathrm{loo}} := \sum_{j \neq h} p_{j,h}^{\mathrm{loo}} u_j, \qquad X_h^{\mathrm{loo}} := (u_h - \eta_h^{\mathrm{loo}}) v_h^\top.$$

Then $\nabla m_h(\widehat{W}^{\setminus h}) = X_h^{\mathrm{loo}}$ and the outer-loss gradient is $-\sigma(-m_h^{\mathrm{loo}}) X_h^{\mathrm{loo}}$. The Gaussian frame bound and (111) give

$$\left\|\eta_h^{\mathrm{loo}}\right\| = O_\prec(d^{-1/2}), \qquad \left\|X_h^{\mathrm{loo}}\right\|_F = O_\prec(1). \tag{114}$$

Set

$$B_h := -T_h[G_h(\widehat{W}^{\setminus h})], \qquad \theta_h := \left\langle X_h^{\mathrm{loo}}, T_h[X_h^{\mathrm{loo}}] \right\rangle_{\mathsf{F}}, \qquad \zeta_h := \left\langle X_h^{\mathrm{loo}}, B_h \right\rangle_{\mathsf{F}}.$$

Let $a_h^\star \in (0,1)$ be the unique solution of

$$a_h^\star = \sigma\left(-m_h^{\mathrm{loo}} - \zeta_h - \theta_h a_h^\star\right). \tag{115}$$

Finally, define

$$\widetilde{W}_h := \widehat{W}^{\setminus h} + B_h + a_h^\star T_h[X_h^{\mathrm{loo}}].$$

Theorem 3 (One-deletion Newton comparison)**.** *Uniformly in h,*

$$\max_{i,j \neq h} |u_j^\top (\widetilde{W}_h - \widehat{W}^{\setminus h}) v_i| = O_\prec(d^{-1}), \tag{116}$$

$$\left\|\nabla \Psi(\widetilde{W}_h)\right\|_F = O_\prec(d^{-1}), \tag{117}$$

$$\left\|\widehat{W} - \widetilde{W}_h\right\|_F = O_\prec(d^{-1}). \tag{118}$$

The score estimate and stationarity residual yield (A2).

The solution in (115) is unique because the left side is strictly increasing and the right side is decreasing in a. Also, $\theta_h \geq 0$ and $\theta_h = O_\prec(1)$. By Lemma 6,

$$\|B_h\|_F = O_\prec(d^{-1/2}), \qquad \left\|\widetilde{W}_h - \widehat{W}^{\setminus h}\right\|_F = O_\prec(1). \tag{119}$$

Lemma 7 (Surviving candidate scores)**.** *Uniformly in h and $i, j \neq h$,*

$$|u_j^\top (\widetilde{W}_h - \widehat{W}^{\setminus h}) v_i| = O_\prec(d^{-1}). \tag{120}$$

Proof. Fix $i, j \neq h$ and set $A_{j,i}^{(h)} = T_h[u_j v_i^\top]$. This matrix is independent of (u_h, v_h) and has Frobenius norm $O_\prec(1)$. By self-adjointness,

$$u_j^\top T_h[X_h^{\mathrm{loo}}] v_i = \left\langle A_{j,i}^{(h)}, (u_h - \eta_h^{\mathrm{loo}}) v_h^\top \right\rangle_{\mathsf{F}}.$$

Fresh two-sided contraction gives $u_h^\top A_{j,i}^{(h)} v_h = O_\prec(d^{-1})$, while (114) and one-sided contraction give $|(\eta_h^{\mathrm{loo}})^\top A_{j,i}^{(h)} v_h| = O_\prec(d^{-1})$.

For $B_h = -T_h G_h(\widehat{W}^{\setminus h})$, Lemma 6 separates G_h into an $O_\prec(d^{-1})$ Frobenius term and $u_h \xi_h^\top$ with $\|\xi_h\| = O_\prec(d^{-1/2})$. The former pairs with $A_{j,i}^{(h)}$ at scale d^{-1}, while the latter is bounded by $\left\|(A_{j,i}^{(h)})^\top u_h\right\| \|\xi_h\| = O_\prec(d^{-1})$ by fresh one-sided contraction. This proves (120). □

Lemma 8 (Held-out row and column increments)**.** *Let* $\Delta_h = \widetilde{W}_h - \widehat{W}^{\setminus h}$. *Uniformly in* h,

$$\max_{i \neq h} |u_h^\top \Delta_h v_i| = O_\prec(d^{-1/2}), \tag{121}$$

$$\left\| (u_j^\top \Delta_h v_h)_{j \neq h} \right\|_2 = O_\prec(\sqrt{d}), \tag{122}$$

$$\sum_{i \neq h} s_{h,i}(\widetilde{W}_h)^2 = O_\prec(n). \tag{123}$$

Proof. By (119), the B_h contribution is $O_\prec(d^{-1/2})$. For the outer Newton direction, self-adjointness gives

$$u_h^\top T_h[u_h v_h^\top] v_i = \left\langle T_h[u_h v_i^\top], u_h v_h^\top \right\rangle_{\mathsf{F}}.$$

Conditionally on u_h and the surviving data, the first factor is independent of v_h and has Frobenius norm $O_\prec(1)$, so this is $O_\prec(d^{-1/2})$; replacing u_h in the forcing by η_h^{loo} gives $O_\prec(d^{-1})$. This proves (121).

The remaining estimates use only operator norms:

$$\left\| U_{\setminus h}^\top \Delta_h v_h \right\|_2 \leq \|U\|_{\mathsf{op}} \|\Delta_h\|_F \|v_h\| = O_\prec(\sqrt{d}),$$

which gives (122). Also,

$$\left\| (s_{h,i}(\widetilde{W}_h))_{i \neq h} \right\|_2 \leq \left\| (s_{h,i}(\widehat{W}^{\setminus h}))_{i \neq h} \right\|_2 + \|V\|_{\mathsf{op}} \|\Delta_h\|_F \|u_h\|.$$

The first term is $O_\prec(\sqrt{n})$ by the conditional quadratic-form argument used in Lemma 6, while the second is $O_\prec(\sqrt{d})$; since $n \asymp d^2$, (123) follows. □

Stationarity residual. Write

$$\Delta_h = \widetilde{W}_h - \widehat{W}^{\setminus h}.$$

The full reduced objective decomposes as

$$\Psi(W) = \Psi^{\setminus h}(W) + \mathcal{P}_h(W) + \mathcal{L}_h(W).$$

By construction,

$$\mathcal{H}_h[\Delta_h] = a_h^\star X_h^{\mathrm{loo}} - G_h(\widehat{W}^{\setminus h}).$$

Therefore

$$\nabla \Psi(\widetilde{W}_h) = R_{0,h} + R_{\mathrm{in},h} + R_{\mathrm{out},h}, \tag{124}$$

where

$$\begin{aligned} R_{0,h} &:= \nabla \Psi^{\setminus h}(\widehat{W}^{\setminus h} + \Delta_h) - \mathcal{H}_h[\Delta_h], \\ R_{\mathrm{in},h} &:= G_h(\widehat{W}^{\setminus h} + \Delta_h) - G_h(\widehat{W}^{\setminus h}), \\ R_{\mathrm{out},h} &:= a_h^\star X_h^{\mathrm{loo}} - a_h(\widetilde{W}_h) X_h(\widetilde{W}_h). \end{aligned}$$

We bound the three terms separately.

Common leave-one-out remainder.

Lemma 9 (Common residual)**.** *Uniformly in h,*

$$\left\|R_{0,h}\right\|_F = O_\prec(d^{-1}). \tag{125}$$

Proof. By Lemma 7, every surviving score perturbation is $\varepsilon = O_\prec(d^{-1})$. The second-order profiled-tail bounds (103) and (105) therefore produce an $n \times n$ coefficient matrix E, with row and column h zero, such that

$$\max_{i\neq h} |E_{ii}| = O_\prec(d^{-2}), \qquad \max_{\substack{i,j\neq h\\ i\neq j}} |E_{ji}| = O_\prec(d^{-2}/k).$$

For its diagonal part, Khatri–Rao synthesis gives $\left\|UE_{\mathrm{d}}V^\top\right\|_F \le O_\prec(1)\sqrt{n}\, d^{-2} = O_\prec(d^{-1})$. For the off-diagonal part, $\|E_{\mathrm{o}}\|_F = O_\prec(d^{-2})$, hence $\left\|UE_{\mathrm{o}}V^\top\right\|_F \le \|U\|_{\mathsf{op}}\|E_{\mathrm{o}}\|_F\|V\|_{\mathsf{op}} = O_\prec(d^{-1})$. The ridge term is exactly linear, proving (125). □

Insertion-gradient stability.

Lemma 10 (Insertion-gradient stability along the candidate step)**.** *Uniformly in h,*

$$\left\|R_{\mathrm{in},h}\right\|_F = \left\|G_h(\widetilde{W}_h) - G_h(\widehat{W}^{\setminus h})\right\|_F = O_\prec(d^{-1}). \tag{126}$$

Proof. At both $\widehat{W}^{\setminus h}$ and $\widetilde{W}_h$, the deterministic Frobenius bound and Lemma 8 give the hypotheses of Lemma 5. Hence the first three sums in (112) are $O_\prec(d^{-1})$ at each endpoint, by the proof of Lemma 6, and so is their difference.

It remains to compare the structured fourth terms. Write $c_i(W) = a_i(W)p_{h,i}(W)$ and $\xi(W) = \sum_{i\neq h} c_i(W)v_i$. Along $W_t = \widehat{W}^{\setminus h} + t\Delta_h$, surviving scores move by $\varepsilon = O_\prec(d^{-1})$, whereas the newly inserted score moves by $\delta = O_\prec(d^{-1/2})$. Since the tail weights sum to one and $p_{h,i} \le k^{-1}$, $|\frac{d}{dt}t_i(W_t)| \le \varepsilon + \delta/k$. Therefore

$$|a_i(\widetilde{W}_h) - a_i(\widehat{W}^{\setminus h})| \le C\left(\varepsilon + \frac{\delta}{k}\right) = O_\prec(d^{-1}),$$

while (101) gives

$$|p_{h,i}(\widetilde{W}_h) - p_{h,i}(\widehat{W}^{\setminus h})| \le \frac{C}{k}(\delta + \varepsilon) = O_\prec(d^{-5/2}).$$

Combining these bounds with $p_{h,i} \le k^{-1}$ yields $|c_i(\widetilde{W}_h) - c_i(\widehat{W}^{\setminus h})| = O_\prec(d^{-5/2})$. Thus the coefficient difference has ℓ_2 norm $O_\prec(d^{-3/2})$ and

$$\left\|u_h(\xi(\widetilde{W}_h) - \xi(\widehat{W}^{\setminus h}))^\top\right\|_F \le \|u_h\|\|V\|_{\mathsf{op}}\, O_\prec(d^{-3/2}) = O_\prec(d^{-1}).$$

This proves (126). □

Held-out outer loss. Let $m_h(W)$ and $X_h(W) = \nabla m_h(W)$ denote the held-out margin and its effective feature. Thus

$$\nabla \mathcal{L}_h(W) = -a_h(W)X_h(W), \qquad a_h(W) = \sigma(-m_h(W)).$$

Lemma 11 (Outer-loss residual)**.** *Uniformly in h,*

$$\left\|R_{\mathrm{out},h}\right\|_F = \left\|a_h^\star X_h^{\mathrm{loo}} - a_h(\widetilde{W}_h)X_h(\widetilde{W}_h)\right\|_F = O_\prec(d^{-1}). \tag{127}$$

Proof. Let $\delta z_h = (u_j^\top \Delta_h v_h)_{j\neq h}$. By (122), $\|\delta z_h\|_2 = O_\prec(\sqrt{d})$, so the scalar tail remainder gives

$$\left| m_h(\widetilde{W}_h) - m_h^{\mathrm{loo}} - \left\langle X_h^{\mathrm{loo}}, \Delta_h \right\rangle_{\mathsf{F}} \right| \le \frac{\beta}{8k}\|\delta z_h\|_2^2 = O_\prec(d^{-1}). \tag{128}$$

By construction, $\langle X_h^{\mathrm{loo}}, \Delta_h\rangle_{\mathsf{F}} = \zeta_h + \theta_h a_h^\star$, and (115) chooses $a_h^\star$ so that the linearized margin is matched exactly. Since the sigmoid is $1/4$-Lipschitz, (128) implies

$$|a_h(\widetilde{W}_h) - a_h^\star| = O_\prec(d^{-1}). \tag{129}$$

Finally, (104) yields $\left\|p_h(\widetilde{W}_h) - p_h(\widehat{W}^{\setminus h})\right\|_2 = O_\prec(d^{-3/2})$, and therefore

$$\left\|X_h(\widetilde{W}_h) - X_h^{\mathrm{loo}}\right\|_F \le \|U\|_{\mathsf{op}}\, O_\prec(d^{-3/2})\|v_h\| = O_\prec(d^{-1}).$$

Together with (129) and $\left\|X_h^{\mathrm{loo}}\right\|_F = O_\prec(1)$, this proves (127). □

Completion of the residual estimate.

Proof of Theorem 3, residual and optimizer parts. By (124) and Lemmas 9–11,

$$\left\|\nabla\Psi(\widetilde{W}_h)\right\|_F = O_\prec(d^{-1}),$$

which is (117). Strong convexity of Ψ therefore yields

$$\left\|\widehat{W} - \widetilde{W}_h\right\|_F \le \gamma_d^{-1}\left\|\nabla\Psi(\widetilde{W}_h)\right\|_F = O_\prec(d^{-1}),$$

which is (118). Equation (116) is Lemma 7. □

Finite-deletion consequences.

Proposition 14 (One- and two-deletion consequences)**.** *Uniformly in distinct $h, i \in [n]$,*

$$\left\|\widehat{W}\right\|_F + \left\|\widehat{W}^{\setminus h}\right\|_F + \left\|\widehat{W}^{\setminus\{h,i\}}\right\|_F = O_\prec(d), \tag{130}$$

$$\left\|\widehat{W} - \widehat{W}^{\setminus h}\right\|_F + \left\|\widehat{W}^{\setminus h} - \widehat{W}^{\setminus\{h,i\}}\right\|_F = O_\prec(1). \tag{131}$$

Moreover,

$$\widehat{W} - \widehat{W}^{\setminus h} = B_h + a_h^\star T_h[X_h^{\mathrm{loo}}] + \mathcal{R}_h, \qquad \|\mathcal{R}_h\|_F = O_\prec(d^{-1}), \tag{132}$$

where B_h, T_h, X_h^{loo}, and $a_h^\star$ are defined above and $\mathcal{R}_h = \widehat{W} - \widetilde{W}_h$.

Proof. The finite-deletion version of Lemma 18 gives (130). The first term in (131) follows from this bound, (119), and (118); the second follows by repeating the one-deletion construction after first deleting i. There are only polynomially many index pairs. Finally, (132) follows immediately from the candidate definition and (118). □

C.2 Stability of the auxiliary variables

We now sharpen the threshold perturbation beyond what follows from the pointwise score bound alone. The exact threshold equation below expresses $\Delta\mu_i$ through an averaged score perturbation. The denominator is kept uniformly nondegenerate by the bulk curvature estimates, while the Newton decomposition from the preceding subsection makes the numerator smaller by a further factor $d^{-1/2}$.

Let $(\widehat{W}, \widehat{\mu})$ be the full lifted optimizer and $(\widehat{W}^{\setminus h}, \widehat{\mu}^{\setminus h})$ the optimizer after deleting h. Define

$$\Delta W = \widehat{W} - \widehat{W}^{\setminus h}, \qquad \Delta\mu_i = \widehat{\mu}_i - \widehat{\mu}_i^{\setminus h}, \qquad \Delta s_{j,i} = u_j^\top \Delta W v_i.$$

Exact threshold identity. Fix h and $i \neq h$. The labels *full* and *loo* denote evaluation at the full and leave-one-out optimizers, respectively. Thus

$$s_{j,i}^{\mathrm{full}} = u_j^\top \widehat{W} v_i, \qquad s_{j,i}^{\mathrm{loo}} = u_j^\top \widehat{W}^{\setminus h} v_i.$$

For $t \in [0,1]$, set

$$s_{j,i}^{(t)} = s_{j,i}^{\mathrm{loo}} + t\Delta s_{j,i}, \qquad \mu_i^{(t)} = \widehat{\mu}_i^{\setminus h} + t\Delta\mu_i,$$

$$q_{j,i}^{(t)} = \sigma\Big(\beta(s_{j,i}^{(t)} - \mu_i^{(t)})\Big), \qquad c_{j,i}^{(t)} = \beta q_{j,i}^{(t)}(1 - q_{j,i}^{(t)}), \qquad \bar{c}_{j,i} = \int_0^1 c_{j,i}^{(t)}\, dt.$$

Lemma 12 (Exact threshold identity)**.** *For every h and $i \neq h$,*

$$\Delta\mu_i = \frac{q_{h,i}^{\mathrm{full}} + \sum_{j \neq i,h} \bar{c}_{j,i} \Delta s_{j,i}}{\sum_{j \neq i,h} \bar{c}_{j,i}}. \tag{133}$$

Proof. Subtracting the full and leave-one-out threshold equations gives $0 = q_{h,i}^{\mathrm{full}} + \sum_{j\neq i,h}(q_{j,i}^{\mathrm{full}} - q_{j,i}^{\mathrm{loo}})$. Along the interpolation above, the fundamental theorem of calculus gives

$$q_{j,i}^{\mathrm{full}} - q_{j,i}^{\mathrm{loo}} = \bar{c}_{j,i}(\Delta s_{j,i} - \Delta\mu_i).$$

Substitution and rearrangement prove (133). □

Uniform quantile curvature.

Lemma 13 (Output scale and endpoint curvature)**.** *Uniformly in h and $i \neq h$,*

$$\left\|\widehat{W}^{\setminus h} v_i\right\| = O_\prec(\sqrt{d}), \qquad \left\|\widehat{W} v_i\right\| = O_\prec(\sqrt{d}). \tag{134}$$

If

$$c_{j,i}^{\mathrm{loo}} = \beta q_{j,i}^{\mathrm{loo}}(1 - q_{j,i}^{\mathrm{loo}}), \qquad D_{h,i}^{\mathrm{loo}} = \sum_{j \neq i,h} c_{j,i}^{\mathrm{loo}},$$

then

$$\max_h \max_{i \neq h} \frac{k}{D_{h,i}^{\mathrm{loo}}} = O_\prec(1). \tag{135}$$

Proof. The two-deletion optimizer $\widehat{W}^{\setminus\{h,i\}}$ is independent of v_i and has Frobenius norm $O_\prec(d)$ by (130); one-sided Gaussian contraction therefore gives $\left\|\widehat{W}^{\setminus\{h,i\}}v_i\right\| = O_\prec(\sqrt{d})$. The stability estimate (131) transfers this bound to $\widehat{W}^{\setminus h}v_i$ and then to $\widehat{W}v_i$, proving (134).

Set $x_{h,i} = \widehat{W}^{\setminus h}v_i$ and $\rho_{h,i} = \left\|x_{h,i}\right\|/\sqrt{d} = O_\prec(1)$. Applying the bulk-curvature part of Lemma 4 to the leave-one-out quantile equation gives

$$\sum_{j\neq i,h} q_{j,i}^{\mathrm{loo}}(1-q_{j,i}^{\mathrm{loo}}) \geq \frac{cn}{1+\rho_{h,i}}.$$

Since $k \asymp n$, this is exactly (135). □

Lemma 14 (Rough threshold stability and path curvature)**.** *Uniformly in h and $i \neq h$,*

$$|\Delta\mu_i| = O_\prec(d^{-1}), \tag{136}$$

and

$$\left(\frac{1}{k}\sum_{j\neq i,h}\bar{c}_{j,i}\right)^{-1} = O_\prec(1). \tag{137}$$

Proof. Put $J = [n]\setminus\{i,h\}$ and

$$F_{\mathrm{loo}}(x) = \sum_{j\in J}\sigma(\beta(s_{j,i}^{\mathrm{loo}}-x)), \qquad F_{\mathrm{full}}(x) = \sum_{j\in J}\sigma(\beta(s_{j,i}^{\mathrm{full}}-x)).$$

Then $F_{\mathrm{loo}}(\widehat{\mu}_i^{\setminus h}) = k$ and $F_{\mathrm{full}}(\widehat{\mu}_i) = k - q_{h,i}^{\mathrm{full}} \in (k-1,k)$. The candidate-score and candidate-error estimates in Theorem 3 imply

$$\delta_i := \max_{j\in J}|\Delta s_{j,i}| = O_\prec(d^{-1}).$$

Monotonicity gives

$$F_{\mathrm{loo}}(x+\delta_i) \leq F_{\mathrm{full}}(x) \leq F_{\mathrm{loo}}(x-\delta_i).$$

The left inequality implies $\widehat{\mu}_i \geq \widehat{\mu}_i^{\setminus h} - \delta_i$.

For $c(z) = \beta\sigma(\beta z)(1-\sigma(\beta z))$, one has $|(\log c)'(z)| \leq \beta$. Therefore, for $0 \leq t \leq 1$,

$$-F'_{\mathrm{loo}}(\widehat{\mu}_i^{\setminus h}+t) \geq e^{-\beta}D_{h,i}^{\mathrm{loo}}.$$

Set $\tau_{h,i} = e^{\beta}/D_{h,i}^{\mathrm{loo}} = O_\prec(n^{-1})$. Integrating gives $F_{\mathrm{loo}}(\widehat{\mu}_i^{\setminus h}+\tau_{h,i}) \leq k-1$, and the other monotonicity inequality implies $\widehat{\mu}_i \leq \widehat{\mu}_i^{\setminus h} + \delta_i + \tau_{h,i}$. This proves (136).

Now

$$\varepsilon_{h,i} := \max_{j\neq i,h}|\Delta s_{j,i} - \Delta\mu_i| = O_\prec(d^{-1}).$$

The same logarithmic derivative bound yields $c_{j,i}^{(t)} \geq e^{-\beta\varepsilon_{h,i}}c_{j,i}^{\mathrm{loo}}$ for every t. Sum over j, use (135), and integrate in t to obtain (137). □

Define

$$g_{h,i} = \frac{1}{k}\sum_{j\neq i,h} \bar{c}_{j,i} u_j, \qquad g^{\mathrm{loo}}_{h,i} = \frac{1}{k}\sum_{j\neq i,h} c^{\mathrm{loo}}_{j,i} u_j.$$

Then

$$\frac{1}{k}\sum_{j\neq i,h} \bar{c}_{j,i}\Delta s_{j,i} = g_{h,i}^\top \Delta W v_i. \tag{138}$$

Lemma 15 (Derivative-weight average bounds)**.** *Uniformly in h and $i \neq h$,*

$$\left\| g^{\mathrm{loo}}_{h,i} \right\| = O_\prec(d^{-1/2}), \tag{139}$$

$$\left\| g_{h,i} - g^{\mathrm{loo}}_{h,i} \right\| = O_\prec(d^{-3/2}). \tag{140}$$

Moreover, $g^{\mathrm{loo}}_{h,i}$ is measurable with respect to the data excluding (u_h, v_h).

Proof. Since $0 \le c^{\mathrm{loo}}_{j,i} \le \beta/4$, $\left\|(c^{\mathrm{loo}}_{j,i})_{j\neq i,h}\right\|_2 \le C\sqrt{n}$; hence

$$\left\| g^{\mathrm{loo}}_{h,i} \right\| \le \frac{\|U\|_{\mathsf{op}}}{k} C\sqrt{n} = O_\prec(d^{-1/2}).$$

The map $z \mapsto \beta\sigma(\beta z)(1-\sigma(\beta z))$ is globally Lipschitz. The candidate score bound together with (136) therefore yields $\max_{j\neq i,h} |\bar{c}_{j,i} - c^{\mathrm{loo}}_{j,i}| = O_\prec(d^{-1})$. Applying the same operator-norm estimate gives

$$\left\| g_{h,i} - g^{\mathrm{loo}}_{h,i} \right\| \le \frac{\|U\|_{\mathsf{op}}}{k}\sqrt{n}\, O_\prec(d^{-1}) = O_\prec(d^{-3/2}).$$

Measurability follows from the leave-one-out definition. □

Averaged Newton contraction. The following is the key gain in the threshold argument: although each surviving score changes at scale d^{-1}, the curvature-weighted average in (138) is smaller by an additional factor $d^{-1/2}$.

Lemma 16 (Averaged score perturbation)**.** *Uniformly in h and $i \neq h$,*

$$\frac{1}{k}\sum_{j\neq i,h} \bar{c}_{j,i}\Delta s_{j,i} = O_\prec(d^{-3/2}). \tag{141}$$

Proof. Use the one-deletion decomposition (132):

$$\Delta W = B_h + a_h^\star T_h[X_h^{\mathrm{loo}}] + \mathcal{R}_h.$$

First, by (140), (131), and the Gaussian norm bound,

$$|(g_{h,i} - g^{\mathrm{loo}}_{h,i})^\top \Delta W v_i| = O_\prec(d^{-3/2}).$$

The residual contributes

$$|(g^{\mathrm{loo}}_{h,i})^\top \mathcal{R}_h v_i| \le \left\| g^{\mathrm{loo}}_{h,i} \right\| \|\mathcal{R}_h\|_F \|v_i\| = O_\prec(d^{-3/2}).$$

Define

$$C_{h,i} = T_h[g_{h,i}^{\mathrm{loo}} v_i^\top].$$

By self-adjointness and (139),

$$\left\|C_{h,i}\right\|_F = O_\prec(d^{-1/2}).$$

The matrix $C_{h,i}$ is independent of the held-out pair. Since $X_h^{\mathrm{loo}} = (u_h - \eta_h^{\mathrm{loo}})v_h^\top$ with $\left\|\eta_h^{\mathrm{loo}}\right\| = O_\prec(d^{-1/2})$,

$$\begin{aligned}(g_{h,i}^{\mathrm{loo}})^\top T_h[X_h^{\mathrm{loo}}]v_i &= \left\langle C_{h,i}, (u_h - \eta_h^{\mathrm{loo}})v_h^\top \right\rangle_{\mathsf{F}} \\ &= u_h^\top C_{h,i} v_h - (\eta_h^{\mathrm{loo}})^\top C_{h,i} v_h.\end{aligned}$$

The two-sided fresh contraction gives $u_h^\top C_{h,i} v_h = O_\prec(d^{-3/2})$, while $\left\|C_{h,i}v_h\right\| = O_\prec(d^{-1})$ and hence the second term is also $O_\prec(d^{-3/2})$.

It remains to treat $B_h = -T_h[G_h(\widehat{W}^{\setminus h})]$. By Lemma 6, write

$$G_h(\widehat{W}^{\setminus h}) = G_h^{\mathrm{sm}} + u_h \xi_h^\top, \qquad \|G_h^{\mathrm{sm}}\|_F = O_\prec(d^{-1}), \qquad \|\xi_h\| = O_\prec(d^{-1/2}).$$

Self-adjointness gives

$$(g_{h,i}^{\mathrm{loo}})^\top B_h v_i = -\left\langle C_{h,i}, G_h^{\mathrm{sm}} \right\rangle_{\mathsf{F}} - u_h^\top C_{h,i} \xi_h.$$

The first term is $O_\prec(d^{-3/2})$. For the second, freshness gives $\left\|C_{h,i}^\top u_h\right\| = O_\prec(d^{-1})$, and the pathwise bound on ξ_h gives the same $d^{-3/2}$ rate. Combining the estimates with (138) proves (141). □

Proof of Proposition 4, (A1). Divide the identity (133) by k. The denominator is bounded below at order one by (137); the removed-competitor term satisfies $q_{h,i}^{\mathrm{full}}/k = O(k^{-1}) = O(d^{-2})$; and the averaged score term is $O_\prec(d^{-3/2})$ by Lemma 16. Therefore

$$|\Delta\mu_i| = O_\prec(d^{-3/2}).$$

All estimates are uniform over the $O(n^2)$ pairs (h, i). □

C.3 Score stability and uniform boundedness

Proof of Proposition 4, (A2). Fix h and $i, j \neq h$. Add and subtract the Newton candidate:

$$\left|u_j^\top(\widehat{W} - \widehat{W}^{\setminus h})v_i\right| \le \left|u_j^\top(\widetilde{W}_h - \widehat{W}^{\setminus h})v_i\right| + \left|u_j^\top(\widehat{W} - \widetilde{W}_h)v_i\right|.$$

The first term is $O_\prec(d^{-1})$ by Lemma 7. For the second,

$$\left|u_j^\top(\widehat{W} - \widetilde{W}_h)v_i\right| \le \left\|u_j\right\|\left\|v_i\right\|\left\|\widehat{W} - \widetilde{W}_h\right\|_F = O_\prec(d^{-1})$$

by (118) and the Gaussian norm bound. All estimates were uniform over h, i, j, and the total number of index triples is $O(n^3) = O(d^6)$, which is polynomial. Therefore

$$\max_h \max_{i,j\neq h} \left|u_j^\top(\widehat{W} - \widehat{W}^{\setminus h})v_i\right| = O_\prec(d^{-1}).$$

This proves the claim. □

Uniform score boundedness.

Lemma 17 (Fresh leave-two-out score)**.** *Uniformly in* $i, j \in [n]$,

$$\left| u_j^\top \widehat{W}^{\setminus\{i,j\}} v_i \right| = O_\prec(1), \tag{142}$$

where $\{i,j\} = \{i\}$ *when* $i = j$.

Proof. After deleting the pair(s) indexed by i and j, the optimizer is independent of u_j and v_i, so its $O_\prec(d)$ Frobenius norm from (130), conditional two-sided Gaussianity, and a union bound over the polynomial-size family of pairs give (142). □

Proof of Proposition 4, (A3). For fixed i, j, insert $\widehat{W}^{\setminus\{i,j\}}$. The fresh term is $O_\prec(1)$ by Lemma 17, while

$$\left| u_j^\top \big(\widehat{W} - \widehat{W}^{\setminus\{i,j\}}\big) v_i \right| \le \|u_j\| \|v_i\| \left\| \widehat{W} - \widehat{W}^{\setminus\{i,j\}} \right\|_F = O_\prec(1)$$

by (131); the claim follows uniformly over the $O(n^2)$ pairs. □

D Spectral and self-averaging estimates

This appendix proves (A5) and (A6). The proof of the spectral estimate (A4) is given in Appendix E. We use the notation of Section 4, in particular the score matrix $S(W) = U^\top W V$, the joint optimizer $(\widehat{W}, \widehat{\mu})$, and the coefficients $a_i, b_i, q_{j,i}$ from Section 4.1.

The proofs of (A5) and (A6) use the same sensitivity–concentration scheme. After profiling out the threshold variables μ, we regard the statistic of interest as a function of the Gaussian data (U, V). Strong convexity of the profiled objective permits implicit differentiation of the optimizer, and a self-adjoint adjoint solve supplies this optimizer response into a single matrix direction P. The resulting data-gradient is controlled using the profiled Hessian quadratic form, after which the Gaussian L^p Poincare inequality gives concentration. For (A6) there is one additional preliminary step: we first replace the lifted Hessian by the simpler rank-one-feature Hessian of (65), and then apply the same adjoint argument used for (A5).

For $J_i = [n] \setminus \{i\}$, write $\mathcal{T}_i = \mathcal{T}_{J_i}$ and $\mu_i = \mu_{J_i}$ in the notation of the common profiled-tail calculus in Appendix C, Section C.1. The profiled objective is

$$\Phi(W) := \sum_{i=1}^n \ell\big(s_{i,i}(W) - \mathcal{T}_i(S_{-i,i}(W))\big) + \frac{\gamma_d}{2}\|W\|_{\mathsf{F}}^2, \qquad \gamma_d := \frac{\lambda n}{d^2}. \tag{143}$$

Minimizing (143) is equivalent to minimizing the joint objective (20), and its unique minimizer is $\widehat{W}$.

For the score-level bounds used below, set

$$h_{j,i} := \phi_\beta''(s_{j,i} - \mu_i), \qquad H_i^\mu := \sum_{j \ne i} h_{j,i}.$$

Let $B_i := B_{J_i}(S_{-i,i}(\widehat{W}))$. Thus, by (101)–(102),

$$0 \preceq B_i \preceq \frac{1}{k}\operatorname{diag}(h_{-i,i}), \qquad \sum_{j \ne i} \left| \frac{\partial q_{j,i}}{\partial s_{\ell,i}} \right| + \sum_{j \ne i} \left| \frac{\partial h_{j,i}}{\partial s_{\ell,i}} \right| \le C. \tag{144}$$

These are the only threshold derivatives needed below. In particular, we never require a lower bound on H_i^μ: differentiation is carried out through the profiled coefficient fields $q_{j,i}, h_{j,i}$ rather than through the derivative of μ_i in isolation. Let $L(S)$ denote the unregularized profiled loss and $G := \nabla_S L(S)$. As in (25),

$$G_{i,i} = -a_i, \qquad G_{j,i} = \frac{a_i q_{j,i}}{k} \quad (j \neq i), \qquad \|G\|_{\mathrm{op}} \leq C. \tag{145}$$

The last bound follows from Schur's test, since the absolute column sums are at most 2 and the row sums are bounded by $1 + (n-1)/k$.

For a score perturbation $T \in \mathbb{R}^{n \times n}$, define

$$\delta_i(T) := T_{i,i} - \frac{1}{k} \sum_{j \neq i} q_{j,i} T_{j,i}, \qquad \mathcal{Q}(T) := \sum_{i=1}^n b_i \delta_i(T)^2 + \sum_{i=1}^n a_i T_{-i,i}^\top B_i T_{-i,i}.$$

Note that $\mathcal{Q}(T)$ is the Hessian quadratic form of L. In particular,

$$0 \leq \mathcal{Q}(T) \leq C \sum_i T_{i,i}^2 + \frac{C}{k} \sum_{i=1}^n \sum_{j \neq i} T_{j,i}^2. \tag{146}$$

Let

$$\mathcal{J} := \nabla_W^2 \Phi(\widehat{W}). \tag{147}$$

Then $\mathcal{J}$ is self-adjoint and

$$\mathcal{J} \succeq \gamma_d I \succeq \lambda \alpha_- I, \qquad \left\|\mathcal{J}^{-1}\right\|_{\mathrm{op}} \leq C. \tag{148}$$

These facts provide polynomial envelopes for all derivatives used below. Consequently, high-probability derivative estimates may be upgraded to fixed-moment estimates by truncating to the standard Gaussian norm event and applying Hölder's inequality.

In both proofs of (A5) and (A6) we use the following Gaussian L^p Poincare inequality which follows from [Pisier, 2006, Theorem 2.2]: if $Z \sim \mathcal{N}(0, I_N/d)$ and f is continuously differentiable with the required moments, then, for fixed $p \geq 2$,

$$\left\|f(Z) - \mathbb{E}f(Z)\right\|_{L^p} \leq C\sqrt{\frac{p}{d}} \left\|\left\|\nabla f(Z)\right\|_2\right\|_{L^p}. \tag{149}$$

D.1 Empirical-mean concentration

We prove a slightly stronger statement. Let $\psi \in C^1(\mathbb{R})$ be pseudo-Lipschitz of order two and set

$$\Gamma_{\mathrm{d}} := \frac{1}{n} \sum_{i=1}^n \psi(s_{i,i}), \qquad \Gamma_{\mathrm{o}} := \frac{1}{n(n-1)} \sum_{i \neq j} \psi(s_{j,i}). \tag{150}$$

Then, for every fixed $p \geq 2$ and $\varepsilon > 0$,

$$\|\Gamma_{\mathrm{d}} - \mathbb{E}\Gamma_{\mathrm{d}}\|_{L^p} \leq C_{p,\varepsilon} d^{-1/2+\varepsilon}, \qquad \|\Gamma_{\mathrm{o}} - \mathbb{E}\Gamma_{\mathrm{o}}\|_{L^p} \leq C_{p,\varepsilon} d^{-1+\varepsilon}. \tag{151}$$

We first give a common sensitivity identity. Let Γ denote either statistic in (150), and let $B = \nabla_S \Gamma$ at $S = S(\widehat{W})$; explicitly,

$$B^{\mathrm{d}}_{j,i} = \frac{\psi'(s_{i,i})}{n} \mathbf{1}_{\{j=i\}}, \qquad B^{\mathrm{o}}_{j,i} = \frac{\psi'(s_{j,i})}{n(n-1)} \mathbf{1}_{\{j\neq i\}}.$$

Let P solve

$$\mathcal{J}[P] = UBV^\top, \qquad T := U^\top P V. \tag{152}$$

The matrix P is the adjoint sensitivity direction associated with Γ: it absorbs the response of $\widehat{W}$ to the data, avoiding any entrywise differentiation of the d^2 coordinates of $\widehat{W}$. Let C be the score-Hessian action

$$C := \nabla_S \big(DL(S)[T] \big).$$

Equivalently,

$$C_{i,i} = b_i \delta_i(T), \qquad C_{j,i} = -\frac{b_i \delta_i(T) q_{j,i}}{k} + a_i (D_i T_{-i,i})_j \quad (j \neq i).$$

Implicit differentiation of the first-order optimality equation, followed by the self-adjoint adjoint substitution (152), gives

$$\nabla_U \Gamma = \widehat{W} V (B - C)^\top - P V G^\top, \tag{153}$$

$$\nabla_V \Gamma = \widehat{W}^\top U (B - C) - P^\top U G. \tag{154}$$

These formulas contain all direct and optimizer-response terms; no entrywise differentiation of $\widehat{W}$ is needed.

Pairing (152) with P yields the identity

$$\mathcal{Q}(T) + \gamma_d \|P\|_{\mathsf{F}}^2 = \langle P, UBV^\top \rangle_{\mathsf{F}}. \tag{155}$$

Consequently,

$$\|P\|_{\mathsf{F}} \leq C \left\| UBV^\top \right\|_{\mathsf{F}}, \qquad \mathcal{Q}(T) \leq C \left\| UBV^\top \right\|_{\mathsf{F}}^2. \tag{156}$$

For the C-terms, Frobenius duality and Cauchy–Schwarz for the positive semidefinite bilinear form associated with $\mathcal{Q}$ give

$$\left\| \widehat{W} V C^\top \right\|_{\mathsf{F}} + \left\| \widehat{W}^\top U C \right\|_{\mathsf{F}} \leq d^{1/2+o(1)} \left\| UBV^\top \right\|_{\mathsf{F}}. \tag{157}$$

To see the first bound, write $\mathcal{B}_{\mathcal{Q}}$ for the symmetric bilinear form associated with $\mathcal{Q}$ and test against A with $\|A\|_{\mathsf{F}} = 1$. With $T_A = A^\top \widehat{W} V$,

$$\langle \widehat{W} V C^\top, A \rangle_{\mathsf{F}} = \langle C, T_A \rangle_{\mathsf{F}} = \mathcal{B}_{\mathcal{Q}}(T, T_A),$$

so positivity gives $|\mathcal{B}_{\mathcal{Q}}(T, T_A)| \leq \mathcal{Q}(T)^{1/2} \mathcal{Q}(T_A)^{1/2}$. The first factor is at most $C \left\| UBV^\top \right\|_{\mathsf{F}}$ by (156). For the second, (146), (A4), and the Gaussian operator- and column-norm bounds give

$$\mathcal{Q}(T_A) \leq C \sum_i (\alpha_i^\top \widehat{W} v_i)^2 + \frac{C}{k} \left\| A^\top \widehat{W} V \right\|_{\mathsf{F}}^2 \leq d^{1+o(1)},$$

where $A = [\alpha_1, \ldots, \alpha_n]$. Taking the supremum over A gives the first term in (157); the second is identical. Together with (153)–(154), (156), and $\|G\|_{\mathrm{op}} \le C$, this reduces the proof to elementary bounds on B.

For the diagonal statistic, let

$$\mathcal{F}_{\mathrm{d}} := [\mathrm{vec}(u_1 v_1^\top), \ldots, \mathrm{vec}(u_n v_n^\top)].$$

The columns are independent Gaussian Khatri–Rao features, and the standard spectral-norm bound for this matrix-chaos model (see e.g. [Bandeira et al., 2025]) gives

$$\|\mathcal{F}_{\mathrm{d}}\|_{\mathrm{op}} = O_\prec(1). \tag{158}$$

Hence,

$$\sum_i s_{i,i}^2 \le \|\mathcal{F}_{\mathrm{d}}\|_{\mathrm{op}}^2 \left\|\widehat{W}\right\|_{\mathsf{F}}^2 = O_\prec(d^2).$$

Since $|\psi'(t)| \le C(1+|t|)$, it follows that $\sum_i \psi'(s_{i,i})^2 = O_\prec(d^2)$. Substituting the diagonal matrix B^{d} into the preceding bounds gives

$$\left\|UB^{\mathrm{d}}V^\top\right\|_{\mathsf{F}} = O_\prec(d^{-1/2}), \qquad \left\|\widehat{W}V(B^{\mathrm{d}})^\top\right\|_{\mathsf{F}} + \left\|\widehat{W}^\top UB^{\mathrm{d}}\right\|_{\mathsf{F}} = O_\prec(d^{-1/2}).$$

Equations (153)–(157) therefore imply

$$\left\|\nabla_{(U,V)}\Gamma_{\mathrm{d}}\right\|_2 = O_\prec(1). \tag{159}$$

For the off-diagonal statistic,

$$\sum_{i,j} s_{j,i}^2 = \left\|U^\top \widehat{W} V\right\|_{\mathsf{F}}^2 \le \|U\|_{\mathrm{op}}^2 \|V\|_{\mathrm{op}}^2 \left\|\widehat{W}\right\|_{\mathsf{F}}^2 = O_\prec(d^4).$$

Thus $\sum_{i\neq j} \psi'(s_{j,i})^2 = O_\prec(d^4)$ and

$$\|B^{\mathrm{o}}\|_{\mathsf{F}} = O_\prec(d^{-2}).$$

Using ((A4)) and the Gaussian operator-norm bounds,

$$\left\|UB^{\mathrm{o}}V^\top\right\|_{\mathsf{F}} = O_\prec(d^{-1}), \qquad \left\|\widehat{W}V(B^{\mathrm{o}})^\top\right\|_{\mathsf{F}} + \left\|\widehat{W}^\top UB^{\mathrm{o}}\right\|_{\mathsf{F}} = O_\prec(d^{-1}).$$

The master estimates then give

$$\left\|\nabla_{(U,V)}\Gamma_{\mathrm{o}}\right\|_2 = O_\prec(d^{-1/2}). \tag{160}$$

The deterministic bound (148) provides polynomial envelopes in $1 + \|U\|_{\mathsf{F}} + \|V\|_{\mathsf{F}}$. Hence (159)–(160) hold in fixed L^p, with an arbitrary factor d^ε. Thus the diagonal and off-diagonal data-gradient scales are respectively $O(1)$ and $O(d^{-1/2})$; the additional factor $d^{-1/2}$ in the Gaussian Poincare inequality is exactly what produces the two concentration scales in (151). Applying (149) to the $2dn$ Gaussian entries of (U, V) gives (151). Markov's inequality, with p arbitrarily large, proves

$$\Gamma_{\mathrm{d}} - \mathbb{E}\Gamma_{\mathrm{d}} = O_\prec(d^{-1/2}), \qquad \Gamma_{\mathrm{o}} - \mathbb{E}\Gamma_{\mathrm{o}} = O_\prec(d^{-1}),$$

which establishes (A5) (and is stronger than required in the off-diagonal case).

D.2 Inverse-trace concentration

We distinguish two Hessians. The operator H in (A6) is the W-Hessian of the lifted objective with μ held fixed, evaluated at $(\widehat{W}, \widehat{\mu})$. The operator $\mathcal{J}$ in (147) is instead the Hessian of the profiled objective and governs the response of $\widehat{W}$ to the data. Both have the same ridge floor, but only $\mathcal{J}$ may be used in implicit differentiation.

Write $f_{j,i} := \mathrm{vec}(u_j v_i^\top)$ and

$$\xi_i := \mathrm{vec}\left(\left[-u_i + \frac{1}{k}\sum_{j\neq i} q_{j,i} u_j\right] v_i^\top\right).$$

Then

$$H = \gamma_d I + \sum_i b_i \xi_i \xi_i^\top + \sum_i \sum_{j\neq i} \frac{a_i h_{j,i}}{k} f_{j,i} f_{j,i}^\top.$$

Recall the simplified Hessian $\widetilde{H}$ defined in (65), which is also used in Appendix G:

$$\widetilde{H} = \gamma_d I + \sum_i b_i f_{i,i} f_{i,i}^\top + \sum_i \sum_{j\neq i} \frac{a_i h_{j,i}}{k} f_{j,i} f_{j,i}^\top,$$

and set

$$\Theta := \frac{1}{d^2}\,\mathrm{tr}(H^{-1}), \qquad \widetilde{\Theta} := \frac{1}{d^2}\,\mathrm{tr}(\widetilde{H}^{-1}).$$

The trace-reduction estimate of Lemma 22 applies verbatim without a cavity index. Here

$$p_i := \frac{1}{k}\sum_{j\neq i} q_{j,i} u_j, \qquad \|p_i\|_2 \le \|U\|_{\mathrm{op}} \left(\frac{n-1}{k^2}\right)^{1/2} = O_\prec(d^{-1/2})$$

uniformly in i. Thus $\widetilde{H}$ replaces the mixed signal direction $(-u_i + p_i)v_i^\top$ by the matched rank-one feature $-u_i v_i^\top$. The discarded component has size $O_\prec(d^{-1/2})$, while all off-diagonal curvature terms are left unchanged; this is the reason for introducing $\widetilde{H}$, whose feature dependence can be differentiated directly. Together with Gaussian column-norm bounds, the resolvent estimate gives, for every fixed p and $\varepsilon > 0$,

$$\left\|\Theta - \widetilde{\Theta}\right\|_{L^p} \le C_{p,\varepsilon} d^{-1/2+\varepsilon}. \tag{161}$$

It remains to prove the stronger concentration of the simplified trace.

Let $R := \widetilde{H}^{-2}$. Since $\widetilde{H} \succeq \gamma_d I, \|R\|_{\mathrm{op}} \le C$. The resolvent derivative

$$D\widetilde{\Theta}[\dot{\widetilde{H}}] = -\frac{1}{d^2}\,\mathrm{tr}(\widetilde{H}^{-1}\dot{\widetilde{H}}\widetilde{H}^{-1}) = -\frac{1}{d^2}\langle R, \dot{\widetilde{H}}\rangle_{\mathrm{HS}} \tag{162}$$

separates into direct feature variation and variation of the coefficients $a_i, b_i, h_{j,i}$. Define

$$\vartheta_{j,i} := \langle f_{j,i}, R f_{j,i}\rangle,$$

and, holding the $\vartheta_{j,i}$ fixed under score differentiation,

$$B^{\mathrm{tr}}_{:,i} := -\frac{1}{d^2}\nabla_{S_{:,i}}\left\{\vartheta_{i,i}b_i + \frac{a_i}{k}\sum_{j\neq i}\vartheta_{j,i}h_{j,i}\right\}. \tag{163}$$

The threshold derivative bounds (144), the bounded coefficient derivatives, and $|\vartheta_{j,i}| \le C\left\|u_j\right\|^2\|v_i\|^2$ give the following scales. Differentiation with respect to the signal score $s_{i,i}$ acts on the margin with unit strength, so the prefactor d^{-2} in (163) gives a diagonal entry of order d^{-2}. In contrast, for $\ell \neq i$,

$$\frac{\partial}{\partial s_{\ell,i}}\big(s_{i,i} - \mathcal{T}_i(S_{-i,i})\big) = -\frac{q_{\ell,i}}{k}.$$

Hence the derivatives of a_i, b_i contribute an additional factor k^{-1}, while the h-derivative term is also $O(k^{-1})$ after using (144). Since $k \asymp d^2$,

$$|B^{\mathrm{tr}}_{i,i}| = O_{\prec}(d^{-2}), \qquad |B^{\mathrm{tr}}_{j,i}| = O_{\prec}(d^{-4}) \quad (j \neq i).$$

Consequently, using (158) for the diagonal part,

$$\left\|UB^{\mathrm{tr}}V^\top\right\|_{\mathsf{F}} = O_{\prec}(d^{-1}), \tag{164}$$

$$\left\|\widehat{W}V(B^{\mathrm{tr}})^\top\right\|_{\mathsf{F}} + \left\|\widehat{W}^\top UB^{\mathrm{tr}}\right\|_{\mathsf{F}} = O_{\prec}(d^{-1/2}). \tag{165}$$

The direct feature derivative obtained from (162) is

$$D\widetilde{\Theta}_{\mathrm{feat}} = -\frac{2}{d^2}\sum_{i,j}\gamma_{j,i}\langle Rf_{j,i},\, Df_{j,i}\rangle,$$

where $\gamma_{i,i} = b_i$ and $\gamma_{j,i} = a_i h_{j,i}/k$ for $j \neq i$. A representative coordinate makes the scaling explicit. Let $X_{j,i} \in \mathbb{R}^{d\times d}$ satisfy $\mathrm{vec}(X_{j,i}) = Rf_{j,i}$. Then

$$\nabla_{u_\ell}\widetilde{\Theta}_{\mathrm{feat}} = -\frac{2}{d^2}\sum_i \gamma_{\ell,i}X_{\ell,i}v_i, \qquad \left\|X_{\ell,i}\right\|_{\mathsf{F}} \le C\|u_\ell\|\|v_i\|.$$

The analogous identity holds for ∇_{v_ℓ}. Since the coefficient row and column sums are bounded, squaring and summing over ℓ gives

$$\left\|\nabla_{(U,V)}\widetilde{\Theta}_{\mathrm{feat}}\right\|_2 = O_{\prec}(d^{-1}). \tag{166}$$

After this explicit feature term has been separated, the remaining score and optimizer dependence has exactly the same form as in the proof of (A5), with B^{tr} in place of B. We therefore reuse the adjoint system, energy identity, and $\mathcal{Q}$-duality estimate from above. To account for the motion of $\widehat{W}$, solve the profiled adjoint system

$$\mathcal{J}[P] = UB^{\mathrm{tr}}V^\top, \qquad T = U^\top PV,$$

and set

$$C^{\mathrm{resp}} := \nabla_S(DL(S)[T]).$$

The same implicit differentiation used for (153)–(154) gives

$$\begin{aligned}\nabla_U\widetilde{\Theta} &= \nabla_U\widetilde{\Theta}_{\mathrm{feat}} + \widehat{W}V(B^{\mathrm{tr}} - C^{\mathrm{resp}})^\top - PVG^\top,\\ \nabla_V\widetilde{\Theta} &= \nabla_V\widetilde{\Theta}_{\mathrm{feat}} + \widehat{W}^\top U(B^{\mathrm{tr}} - C^{\mathrm{resp}}) - P^\top UG.\end{aligned} \tag{167}$$

The identity (155), now with $B = B^{\mathrm{tr}}$, and (164) imply

$$\|P\|_{\mathsf{F}} = O_\prec(d^{-1}), \qquad \mathcal{Q}(T) = O_\prec(d^{-2}).$$

The duality argument leading to (157) then yields

$$\left\|\widehat{W}V(C^{\mathrm{resp}})^\top\right\|_{\mathsf{F}} + \left\|\widehat{W}^\top UC^{\mathrm{resp}}\right\|_{\mathsf{F}} + \left\|PVG^\top\right\|_{\mathsf{F}} + \left\|P^\top UG\right\|_{\mathsf{F}} = O_\prec(d^{-1/2}). \tag{168}$$

Combining (165), (166), and (167)–(168) gives

$$\left\|\nabla_{(U,V)}\widetilde{\Theta}\right\|_2 = O_\prec(d^{-1/2}). \tag{169}$$

As before, the ridge floor and the smooth coefficient calculus give a polynomial deterministic envelope, so (169) upgrades to

$$\left\|\left\|\nabla_{(U,V)}\widetilde{\Theta}\right\|_2\right\|_{L^p} \le C_{p,\varepsilon}d^{-1/2+\varepsilon}.$$

Applying (149) therefore gives

$$\left\|\widetilde{\Theta} - \mathbb{E}\widetilde{\Theta}\right\|_{L^p} \le C_{p,\varepsilon}d^{-1+\varepsilon}. \tag{170}$$

Thus the simplified trace self-averages at the stronger scale d^{-1}; the final $d^{-1/2}$ rate is set by the trace-reduction error (161), not by the concentration of $\widetilde{\Theta}$ itself. Finally,

$$\Theta - \mathbb{E}\Theta = (\Theta - \widetilde{\Theta}) + (\widetilde{\Theta} - \mathbb{E}\widetilde{\Theta}) + \mathbb{E}(\widetilde{\Theta} - \Theta).$$

Equations (161) and (170) imply

$$\|\Theta - \mathbb{E}\Theta\|_{L^p} \le C_{p,\varepsilon}d^{-1/2+\varepsilon}.$$

Markov's inequality with arbitrarily large fixed p proves

$$\frac{1}{d^2}\operatorname{tr}(H^{-1}) - \mathbb{E}\left[\frac{1}{d^2}\operatorname{tr}(H^{-1})\right] = O_\prec(d^{-1/2}),$$

which establishes (A6). Repeating the Poincare estimate (149) for a fixed finite-removal objective and using its stability gives the corresponding estimate for every fixed finite LOO Hessian.

E Bounds on the norms of $\widehat{W}$

This appendix establishes two-sided control on $\|\widehat{W}\|_{\mathsf{F}}/d$ for the joint TAM minimizer of 20 as well as (A4). For the former, the upper bound is deterministic and uniform in the data; the lower bound holds with overwhelming probability ($1-d^{-D}$ for every $D>0$) in the regime $n/d^2 \to \alpha \in (0,\infty)$. Together, they pin

$$c_0 - O(d^{-D}) \;\le\; \nu_d \;\le\; C_0,$$

for deterministic constants $0 < c_0 \le C_0 < \infty$ depending only on the parameters $(\alpha, r, \beta, \lambda)$, where $\nu_d^2 = \mathbb{E}[\|\widehat{W}\|_{\mathsf{F}}^2/d^2]$ as in (48).

The two Frobenius bounds provide the a priori scale for $\|\widehat{W}\|_{\mathsf{F}}$ needed in Section 4.3 and Corollary 1. The corresponding Frobenius concentration estimate is treated there as a consequence of the empirical concentration postulate (A5), applied to the observables in the energy identity. The spectral estimate (A4) is proved at the end of this appendix.

Throughout the appendix, C, c denote positive constants that may depend on $(\alpha, r, \beta, \lambda)$ but not on d, and may vary line to line.

E.1 Deterministic Frobenius upper bound

Lemma 18 (Deterministic upper bound)**.** *For any $\lambda > 0$, $\beta > 0$, $r \in (0,1)$, and any data matrices $U, V \in \mathbb{R}^{d\times n}$,*

$$\frac{\|\widehat{W}\|_{\mathsf{F}}^2}{d^2} \;\le\; \frac{2}{\lambda}\Big[\,\log 2 \;+\; \tfrac{\log 2}{r\beta}\,\Big] \;=:\; C_0. \tag{171}$$

In particular, $\nu_d^2 \le C_0$.

Proof. Evaluate Ψ at the test point $W = 0$, $\mu = 0$. All scores vanish, so

$$\Psi(0,0) \;=\; n\cdot \ell\Big(-\tfrac{1}{k}\,(n-1)\,\phi_\beta(0)\Big) \;=\; n\cdot\ell\Big(-\tfrac{\log 2}{r\beta}\Big),$$

using $\phi_\beta(0) = (\log 2)/\beta$ and $r = k/(n-1)$. The function ℓ is convex with $|\ell'| \le 1$, so

$$\ell(t) \;\le\; \ell(0) + |t| \;=\; \log 2 + |t|. \tag{172}$$

Applying (172) at $t = -(\log 2)/(r\beta)$ gives

$$\Psi(0,0) \;\le\; n\Big[\,\log 2 + \tfrac{\log 2}{r\beta}\,\Big]. \tag{173}$$

By optimality of $(\widehat{W}, \widehat{\mu})$ and nonnegativity of ℓ,

$$\frac{\lambda n}{2d^2}\,\|\widehat{W}\|_{\mathsf{F}}^2 \;\le\; \Psi(\widehat{W},\widehat{\mu}) \;\le\; \Psi(0,0). \tag{174}$$

Combining (173) and (174) and dividing by $\lambda n/(2d^2)$ yields (171). The bound on ν_d^2 follows from $\nu_d^2 = \mathbb{E}\|\widehat{W}\|_{\mathsf{F}}^2/d^2$. □ □

Remark 7 (Dependence on β)**.** *The dependence on β enters as $1/\beta$, which vanishes in the large-β limit, leaving the cleaner asymptotic bound $C_0 \to 2\log 2/\lambda$.*

E.2 High-probability Frobenius lower bound

Lemma 19 (High probability lower bound)**.** *Fix $r, \alpha \in (0,1) \times (0,\infty)$ and $\beta, \lambda > 0$, and consider the regime $n/d^2 \to \alpha$. There exists a constant $c_0 > 0$ such that, for every $D > 0$, there is $d_0 = d_0(D)$ for which*

$$\mathbb{P}\left(\frac{\|\widehat{W}\|_{\mathsf{F}}}{d} \;\geq\; c_0 \right) \;\geq\; 1 - d^{-D}, \qquad d \geq d_0. \tag{175}$$

In particular, $\nu_d^2 \geq c_0^2 - O(d^{-D})$ for every $D > 0$.

Remark 8 (Strategy)**.** *The argument is by objective gap. We exhibit a one-dimensional family $W_t = t\, W^{\times}$ along the direction*

$$W^{\times} \;:=\; UV^{\top}$$

along which $\Psi(W_t, \mu_0 \mathbf{1})$ decreases by $\Theta(n) = \Theta(d^2)$ at small $t > 0$, and complement this with the convex linearization $\Psi(W, \mu_0\mathbf{1}) \geq \Psi(0, \mu_0\mathbf{1}) + \langle \nabla_W \Psi(0, \mu_0\mathbf{1}), W \rangle$, whose linear term is bounded by $\|\nabla_W \Psi(0, \mu_0\mathbf{1})\|_{\mathsf{F}} \|W\|_{\mathsf{F}} = O(d\,\|W\|_{\mathsf{F}})$. Comparing the two excludes a Frobenius ball of radius ηd for small η.

Proof. The variable μ is unregularized; we choose a fixed $\mu_0 \in \mathbb{R}$ and track the objective along the slice $\mu = \mu_0 \mathbf{1}_n$ throughout. Let

$$\mu_0 \;:=\; -\tfrac{1}{\beta} \log \tfrac{r}{1-r}, \qquad q_0 := \sigma(-\beta\mu_0) = r,$$

so that, with $f(\mu) := -\mu - \frac{1}{r}\phi_\beta(-\mu)$, we have $f'(\mu_0) = -1 + \frac{1}{r} q_0 = 0$. Define $a_0 := \sigma(-f(\mu_0)) \in (0,1)$, the activation associated with the test profile $(0, \mu_0 \mathbf{1}_n)$.

Step 1: Gradient at the test point. By the gradient formula (25) evaluated at $W = 0$, $\mu = \mu_0 \mathbf{1}_n$, all $a_i = a_0$ and $q_{j,i} = q_0 = r$, hence

$$\nabla_W \Psi(0, \mu_0 \mathbf{1}_n) \;=\; -a_0\, U\, M\, V^{\top}, \qquad M_{ii} = 1, \quad M_{ij} = -\tfrac{1}{n-1} \text{ for } j \neq i. \tag{176}$$

The matrix M is symmetric with eigenvalues 0 (on $\mathbf{1}_n$) and $n/(n-1)$ (on $\mathbf{1}_n^{\perp}$, with multiplicity $n-1$); in particular, $\|M\|_{\mathsf{F}}^2 = n^2/(n-1)$ and $\|M\|_{\mathrm{op}} = n/(n-1)$.

By independence of U and V together with $\mathbb{E}[U^{\top}U] = \mathbb{E}[V^{\top}V] = I_n$ (since each column has variance I_d/d),

$$\mathbb{E}\,\|UMV^{\top}\|_{\mathsf{F}}^2 \;=\; \mathrm{tr}\big(M\,\mathbb{E}[U^{\top}U]\,M\,\mathbb{E}[V^{\top}V]\big) \;=\; \mathrm{tr}(M^2) \;=\; \tfrac{n^2}{n-1}. \tag{177}$$

Hanson–Wright applied to the Gaussian quadratic form $\|UMV^{\top}\|_{\mathsf{F}}^2$, viewed as a degree-4 polynomial in the i.i.d. $\mathcal{N}(0,1)$ entries of $\sqrt{d}\,U$ and $\sqrt{d}\,V$, has variance $O(n)$ and operator-norm scale $O(1)$, yielding sub-exponential concentration: for any $\delta > 0$,

$$\mathbb{P}\big(\|UMV^{\top}\|_{\mathsf{F}}^2 > (1+\delta)\ \mathrm{tr}(M^2)\big) \;\leq\; \exp(-c\,d^2),$$

for a constant $c > 0$ depending only on δ. Hence there exists a constant $K = K(\alpha) > 0$ such that, for every $D > 0$, the event

$$\Omega_d^{(1)} \;:=\; \big\{\, \|\nabla_W \Psi(0, \mu_0 \mathbf{1}_n)\|_{\mathsf{F}} \;\leq\; a_0\, K\, d \,\big\} \tag{178}$$

satisfies $\mathbb{P}(\Omega_d^{(1)}) \geq 1 - d^{-D}$ for all sufficiently large d.

Step 2: Directional derivative along $W^\times$. Define $g(t) := \Psi(tW^\times, \mu_0 \mathbf{1}_n)$, viewed as a one-dimensional convex function of $t \geq 0$. By (176) and the symmetry of M,

$$g'(0) \;=\; \langle \nabla_W \Psi(0, \mu_0 \mathbf{1}_n),\, W^\times \rangle \;=\; -a_0 \,\mathrm{tr}\big(M\, V^\top V\, U^\top U\big).$$

The quantity $\mathrm{tr}(MV^\top VU^\top U)$ is a degree-4 polynomial in the i.i.d. $\mathcal{N}(0,1)$ entries of $(\sqrt{d}\,U, \sqrt{d}\,V)$, with mean $\mathrm{tr}(M) = n$ (by independence of U, V and $\mathbb{E}[U^\top U] = \mathbb{E}[V^\top V] = I_n$) and variance $O(n)$. Sub-exponential concentration via Hanson–Wright (or, equivalently, Gaussian hypercontractivity for finite-degree chaos) yields

$$\Omega_d^{(2)} \;:=\; \big\{\, g'(0) \;\leq\; -\tfrac{1}{2}\, a_0\, n \,\big\}, \qquad \mathbb{P}(\Omega_d^{(2)}) \;\geq\; 1 - d^{-D}, \tag{179}$$

for every $D > 0$ and all sufficiently large d.

Step 3: Curvature along the path. Computing the curvature $g''(t)$ for the smooth function $\Psi(\,\cdot\,, \mu_0 \mathbf{1}_n)$ and bounding $|\ell''| \leq 1/4$, $|\phi_\beta''| \leq \beta/4$:

$$g''(t) \;\leq\; \tfrac{1}{4} \sum_{i=1}^n (t_i'(t))^2 \;+\; \tfrac{\beta}{4k} \sum_{i=1}^n \sum_{j \neq i} (S_{ji}^\times)^2 \;+\; \tfrac{\lambda n}{d^2}\, \|W^\times\|_{\mathsf{F}}^2,$$

where $t_i(t)$ is the loss argument at sample i along the path. Using $|t_i'(t)|^2 \leq 2(S_{ii}^\times)^2 + (2/(rk)) \sum_{j \neq i} (S_{ji}^\times)^2$, the first two terms collapse into

$$\tfrac{1}{2} \sum_i (S_{ii}^\times)^2 \;+\; \Big(\tfrac{1}{2rk} + \tfrac{\beta}{4k}\Big) \sum_{i, j \neq i} (S_{ji}^\times)^2. \tag{180}$$

The summed second moments of $S^\times$ admit the matching expectations

$$\mathbb{E} \sum_i (S_{ii}^\times)^2 \;\asymp\; n, \qquad \mathbb{E} \sum_{i, j \neq i} (S_{ji}^\times)^2 \;\asymp\; n^2,$$

both inherited from the Wishart structure of $U^\top U$ and $V^\top V$. The ridge contribution $\|W^\times\|_{\mathsf{F}}^2 = \|UV^\top\|_{\mathsf{F}}^2$ has mean n (by the same calculation as (177) with $M = I$), so that $(\lambda n/d^2)\|W^\times\|_{\mathsf{F}}^2 \asymp \lambda \alpha n$. Each of the three sums is a polynomial of bounded degree (at most 8) in the i.i.d. $\mathcal{N}(0,1)$ entries of $\sqrt{d}\,U$ and $\sqrt{d}\,V$ with mean of order n^a for $a \in \{1, 2, 1\}$, respectively, and variance of strictly smaller order. Gaussian hypercontractivity for finite-degree chaos (or, equivalently, Latala's moment inequalities) yields stretched-exponential concentration $\mathbb{P}(|X - \mathbb{E}X| > \delta\, \mathbb{E}X) \leq \exp(-c\, d^{2/q})$ for any constant $\delta > 0$ and degree-q chaos. Combining the three concentration statements via a union bound, there is a constant $L = L(\alpha, r, \beta, \lambda) > 0$ such that, for every $D > 0$, the event

$$\Omega_d^{(3)} \;:=\; \Big\{ \sup_{t \in [0,1]} g''(t) \;\leq\; L\, n \Big\} \tag{181}$$

satisfies $\mathbb{P}(\Omega_d^{(3)}) \geq 1 - d^{-D}$ for all sufficiently large d. (The supremum over t is harmless: the data-dependent factors in (180) are independent of t, while the loss curvatures $|\ell''|, |\phi_\beta''|$ admit deterministic bounds $1/4, \beta/4$ that hold uniformly in t.)

Step 4: Objective gap. Set $t_\star := a_0/(2L) \wedge 1$. By Taylor's theorem with remainder,

$$g(t_\star) \;\le\; g(0) \;+\; g'(0)\,t_\star \;+\; \tfrac{1}{2}\sup_{[0,t_\star]} g''\cdot t_\star^2. \tag{182}$$

On $\Omega_d^{(2)} \cap \Omega_d^{(3)}$, the bounds (179) and (181) reduce (182) to

$$g(t_\star) \;\le\; g(0) \;-\; \tfrac{1}{2}\,a_0\,n\,t_\star \;+\; \tfrac{1}{2}\,L\,n\,t_\star^2 \;\le\; g(0) \;-\; \tfrac{1}{4}\,a_0\,n\,t_\star \;=:\; g(0) - c_1 n,$$

with $c_1 := a_0\,t_\star/4 > 0$.

Step 5: Convex linearization and contradiction. By optimality of $(\widehat{W}, \widehat{\mu})$ in Ψ and convexity of $\Psi(\,\cdot\,, \mu_0\mathbf{1}_n)$ in W,

$$\begin{aligned} g(0) - c_1 n \;\ge\; g(t_\star) \;\ge\; \Psi(\widehat{W}, \widehat{\mu}) \;\ge\; \Psi(0, \mu_0\mathbf{1}_n) \;+\; \langle \nabla_W \Psi(0, \mu_0\mathbf{1}_n),\, \widehat{W}\rangle \\ \ge\; g(0) \;-\; \|\nabla_W \Psi(0, \mu_0\mathbf{1}_n)\|_{\mathsf{F}} \cdot \|\widehat{W}\|_{\mathsf{F}}, \end{aligned}$$

where the third inequality follows from the convex subgradient inequality around $(0, \mu_0\mathbf{1}_n)$, and the optimality $\nabla_\mu \Psi(0, \mu_0\mathbf{1}_n) = 0$ at μ_0 (which follows from $f'(\mu_0) = 0$) eliminates the μ-component. Rearranging, on $\Omega_d := \bigcap_{j=1}^3 \Omega_d^{(j)}$,

$$\|\widehat{W}\|_{\mathsf{F}} \;\ge\; \frac{c_1 n}{\|\nabla_W \Psi(0, \mu_0\mathbf{1}_n)\|_{\mathsf{F}}} \;\ge\; \frac{c_1 n}{a_0 K d} \;=:\; c_0\, d,$$

using (178) and $n \asymp \alpha d^2$, with $c_0 := c_1\alpha/(a_0 K) > 0$. This is (175).

A union bound over $\Omega_d^{(1)}, \Omega_d^{(2)}, \Omega_d^{(3)}$ gives $\mathbb{P}(\Omega_d^c) \le 3d^{-D}$, which is at most $d^{-(D-1)}$ for $d \ge 3$; since D is arbitrary, the rate d^{-D} is preserved.

For the conclusion on ν_d^2: combining Lemma 18 with the bound just established,

$$\|\widehat{W}\|_{\mathsf{F}}^2/d^2 \;\ge\; c_0^2\,\mathbf{1}_{\Omega_d}, \qquad \|\widehat{W}\|_{\mathsf{F}}^2/d^2 \;\le\; C_0 \text{ everywhere},$$

so $\nu_d^2 \ge c_0^2\,\mathbb{P}(\Omega_d) \ge c_0^2(1 - d^{-D})$ for every $D > 0$. □

Together, Lemma 18 and Lemma 19 certify that ν_d is bounded away from 0 and ∞ for all sufficiently large d. This is the *a priori* input that subsequent analyses—in particular the bilinear coupling Corollary 1—use implicitly. The Frobenius-norm concentration statement $\|\widehat{W}\|_{\mathsf{F}}^2/d^2 - \nu_d^2 = \mathcal{O}_\prec(d^{-1/2})$ is not a separate postulate; it is folded into (A5) through the energy identity in Section 4.3.

E.3 Spectral norm bound

The Frobenius bound of Lemma 18 alone yields only $\left\|\widehat{W}\right\|_{\mathrm{op}} = O_\prec(d)$. To recover the square-root scale, we profile the optimization problem along every pair of test directions, thereby separating the adaptive score-gradient matrix from the Gaussian coordinates being tested. Recall that the ridge coefficient in the profiled objective is $\gamma_d = \lambda n/d^2 \asymp 1$.

Fix deterministic $x, y \in S^{d-1}$ and define the orthogonal subspaces

$$\mathcal{C}_{x,y} := \{C : x^\top C = 0,\ Cy = 0\}, \qquad \mathcal{L}_{x,y} := \mathcal{C}_{x,y}^{\perp}.$$

The orthogonal projection onto $\mathcal{L}_{x,y}$ is

$$\Pi_{x,y} M = xx^\top M + Myy^\top - xx^\top Myy^\top. \tag{183}$$

In particular, $xy^\top \in \mathcal{L}_{x,y}$.

Let $C_{x,y}$ minimize Φ over $\mathcal{C}_{x,y}$ and consider the profile

$$F_{x,y}(L) := \min_{C \in \mathcal{C}_{x,y}} \Phi(C + L), \qquad L \in \mathcal{L}_{x,y}.$$

For $C \perp L$, the ridge term splits as

$$\frac{\gamma_d}{2}\|C + L\|_{\mathsf{F}}^2 = \frac{\gamma_d}{2}\|C\|_{\mathsf{F}}^2 + \frac{\gamma_d}{2}\|L\|_{\mathsf{F}}^2 .$$

Thus partial minimization preserves convexity while retaining the γ_d-quadratic term in L, so $F_{x,y}$ is γ_d-strongly convex. If $\widehat{W} = C^\star + L^\star$ is the corresponding orthogonal decomposition, then global optimality of $\widehat{W}$ gives $L^\star = \Pi_{x,y}\widehat{W} = \arg\min F_{x,y}$. The envelope theorem gives

$$\nabla F_{x,y}(0) = \Pi_{x,y} \nabla \Phi(C_{x,y}) = \Pi_{x,y}\big[UG(C_{x,y})V^\top\big],$$

since $\Pi_{x,y} C_{x,y} = 0$. Moreover $\nabla F_{x,y}(L^\star) = 0$, and strong monotonicity yields

$$\gamma_d \|L^\star\|_{\mathsf{F}}^2 \le \big\langle \nabla F_{x,y}(0), -L^\star \big\rangle \le \big\|\nabla F_{x,y}(0)\big\|_{\mathsf{F}} \|L^\star\|_{\mathsf{F}} .$$

Hence $\|L^\star\|_{\mathsf{F}} \le \gamma_d^{-1}\big\|\nabla F_{x,y}(0)\big\|_{\mathsf{F}}$. Since $xy^\top \in \mathcal{L}_{x,y}$ and $\big\|xy^\top\big\|_{\mathsf{F}} = 1$,

$$|x^\top \widehat{W} y| = \big|\langle \Pi_{x,y}\widehat{W}, xy^\top\rangle\big| \le \Big\|\Pi_{x,y}\widehat{W}\Big\|_{\mathsf{F}} \le \gamma_d^{-1}\Big\|\Pi_{x,y}[UG(C_{x,y})V^\top]\Big\|_{\mathsf{F}} . \tag{184}$$

It remains to bound the right-hand side. Decompose

$$u_j = p_j x + \bar{u}_j, \qquad v_i = r_i y + \bar{v}_i,$$

where $p, r \sim \mathcal{N}(0, I_n/d)$ are mutually independent and independent of $(\bar{U}, \bar{V})$. If $C \in \mathcal{C}_{x,y}$, then

$$u_j^\top C v_i = \bar{u}_j^\top C \bar{v}_i.$$

Thus every score in the objective restricted to $\mathcal{C}_{x,y}$ depends only on $(\bar{U}, \bar{V})$. By uniqueness of the restricted minimizer, both $C_{x,y}$ and $G_{x,y} := G(C_{x,y})$ are measurable with respect to $(\bar{U}, \bar{V})$ and hence independent of the fresh Gaussian variables (p, r). This is the key decoupling in the argument. By (145), $\big\|G_{x,y}\big\|_{\mathrm{op}} \le C$ deterministically.

Writing $M = UG_{x,y}V^\top$, (183) gives

$$\big\|\Pi_{x,y} M\big\|_{\mathsf{F}} \le 3|p^\top G_{x,y} r| + \Big\|p^\top G_{x,y}\bar{V}^\top\Big\|_2 + \big\|\bar{U}G_{x,y}r\big\|_2 . \tag{185}$$

Let $\mathcal{E}_d$ denote the standard event

$$\|U\|_{\mathrm{op}} + \|V\|_{\mathrm{op}} \le C\sqrt{d}, \qquad \|U\|_{\mathsf{F}} + \|V\|_{\mathsf{F}} \le C\sqrt{n},$$

which fails with probability at most e^{-cd}. Since we condition on $(\bar{U}, \bar{V})$ below, note that $\mathcal{E}_d$ implies the residual bounds $\|\bar{U}\|_{\mathrm{op}} + \|\bar{V}\|_{\mathrm{op}} \le C\sqrt{d}$ and $\|\bar{U}\|_{\mathsf{F}} + \|\bar{V}\|_{\mathsf{F}} \le C\sqrt{n}$, as well as $\|p\|_2 + \|r\|_2 \le C\sqrt{d}$.

Condition on $(\bar{U}, \bar{V})$, so that $G_{x,y}$ is fixed. For $A := G_{x,y}\bar{V}^\top$, the preceding residual bounds give $\|A\|_{\mathsf{F}} \le C\sqrt{n}$ and $\|A\|_{\mathrm{op}} \le C\sqrt{d}$. Writing $p = d^{-1/2}g$ with $g \sim \mathcal{N}(0, I_n)$,

$$\mathbb{E}\left[\left\|p^\top A\right\|_2^2 \mid \bar{U}, \bar{V}\right] = \frac{1}{d}\|A\|_{\mathsf{F}}^2 = O(d), \qquad \mathrm{Lip}\left(g \mapsto d^{-1/2}\left\|g^\top A\right\|_2\right) \le \frac{\|A\|_{\mathrm{op}}}{\sqrt{d}} = O(1).$$

Thus $\left\|p^\top G_{x,y}\bar{V}^\top\right\|_2$ has conditional mean $O(\sqrt{d})$ and Gaussian deviations on an $O(1)$ scale. The same argument with $B := \bar{U}G_{x,y}$ controls $\left\|\bar{U}G_{x,y}r\right\|_2$. Finally, conditional on $(\bar{U}, \bar{V}, r)$,

$$p^\top G_{x,y}r \sim \mathcal{N}\left(0, \frac{1}{d}\left\|G_{x,y}r\right\|_2^2\right), \qquad \frac{1}{d}\left\|G_{x,y}r\right\|_2^2 \le \frac{C}{d}\|r\|_2^2 = O(1)$$

on $\mathcal{E}_d$. Applying Gaussian concentration to the three terms in (185) therefore gives, uniformly over deterministic x, y,

$$\mathbb{P}\left(\mathcal{E}_d \cap \left\{\left\|\Pi_{x,y}M\right\|_{\mathsf{F}} > K\sqrt{d}\right\}\right) \le Ce^{-cK^2 d} \tag{186}$$

for all sufficiently large fixed K. Here the use of $\mathcal{E}_d$ causes no issue under the conditioning: we only invoke its residual consequences above, which are measurable with respect to $(\bar{U}, \bar{V})$, together with the bound on $\|r\|_2$ for the scalar term.

Let $\mathcal{N}$ be a deterministic $1/4$-net of S^{d-1} with $|\mathcal{N}| \le 9^d$. Taking a union bound in (186),

$$\begin{aligned}\mathbb{P}\left(\max_{x,y\in\mathcal{N}}\left\|\Pi_{x,y}[UG(C_{x,y})V^\top]\right\|_{\mathsf{F}} > K\sqrt{d}\right) \\ \le e^{-c_0 d} + 81^d Ce^{-cK^2 d}.\end{aligned}$$

Choosing K sufficiently large makes the right-hand side exponentially small in d, and hence

$$\max_{x,y\in\mathcal{N}}\left\|\Pi_{x,y}[UG(C_{x,y})V^\top]\right\|_{\mathsf{F}} = O_\prec(\sqrt{d}).$$

Combining this with (184), $\gamma_d^{-1} = O(1)$, and the standard net inequality $\|W\|_{\mathrm{op}} \le 2\max_{x,y\in\mathcal{N}}|x^\top W y|$ proves

$$\left\|\widehat{W}\right\|_{\mathrm{op}} = O_\prec(\sqrt{d}).$$

This establishes (A4). The same proof applies after removing any fixed number of samples: the Gaussian residual decomposition and deterministic bound on G are unchanged, while $n - O(1) \asymp d^2$.

F A coupling lemma for bilinear Gaussian forms

This appendix establishes a coupling estimate for the joint law of bilinear forms $u_j^\top W v_0$ that share a common right-side Gaussian vector v_0. The estimate is the technical input behind the joint Gaussian limit for competitor scores derived in Section 4.3, where it is applied with $W = \widehat{W}^{\setminus i,j_1,\dots,j_m}$ and $v_0 = v_i$.

We first prove the estimate for a deterministic matrix W, then transfer it to the random case via concentration of $\|W\|_{\mathsf{F}}^2/d^2$.

F.1 Deterministic matrix

Lemma 20 (Bilinear Gaussian coupling, deterministic matrix)**.** *Let $W \in \mathbb{R}^{d\times d}$ be a deterministic matrix with $V_0 := \|W\|_{\mathsf{F}}^2/d^2 > 0$, and let $u_1,\dots,u_m, v_0 \in \mathbb{R}^d$ be independent Gaussian vectors with $u_k, v_0 \sim \mathcal{N}(0, I_d/d)$. Set*

$$X_j \;:=\; u_j^\top W v_0, \qquad j = 1,\dots,m.$$

On a common probability space, there exist i.i.d. random variables $Y_1,\dots,Y_m \sim \mathcal{N}(0, V_0)$ such that

$$\big\|(X_1,\dots,X_m) - (Y_1,\dots,Y_m)\big\|_2 \;=\; \mathcal{O}_{\prec}\big(\|W^\top W\|_{\mathsf{F}}/(d^2\sqrt{V_0})\big). \tag{187}$$

Remark 9 (Interpretation)**.** *The deviation $\|W^\top W\|_{\mathsf{F}}/d^2$ measures the spread of the eigenvalues of $W^\top W/d^2$: when the eigenvalues are concentrated, $\|W^\top W\|_{\mathsf{F}}/d^2$ is small relative to $\|W\|_{\mathsf{F}}^2/d^2 = \operatorname{tr}(W^\top W)/d^2$. In the proof below it is exactly the Hanson–Wright scale for the random conditional variance $\|Wv_0\|^2/d$.*

Proof. The argument exploits the conditional Gaussian structure of $\{X_j\}$ given v_0.

Step 1: Conditional Gaussian. Set $w := Wv_0 \in \mathbb{R}^d$. Conditional on v_0, w is deterministic, and $X_j = u_j^\top w$ is a Gaussian linear functional of u_j. Since $u_1,\dots,u_m \sim \mathcal{N}(0, I_d/d)$ are independent across j,

$$(X_1,\dots,X_m)\,\big|\,v_0 \;\sim\; \mathcal{N}\big(0,\, V\cdot I_m\big), \qquad V \;:=\; \frac{\|Wv_0\|^2}{d}.$$

Step 2: Hanson–Wright on the conditional variance. We have

$$\mathbb{E}[V] \;=\; \frac{1}{d}\,\mathbb{E}[v_0^\top W^\top W v_0] \;=\; \frac{1}{d^2}\,\operatorname{tr}(W^\top W) \;=\; \frac{\|W\|_{\mathsf{F}}^2}{d^2} \;=:\; V_0,$$

using $\mathbb{E}[v_0 v_0^\top] = I_d/d$. The Hanson–Wright inequality applied to the centered quadratic form $V - V_0 = (1/d)(v_0^\top W^\top W v_0 - \|W\|_{\mathsf{F}}^2/d)$ in the Gaussian vector v_0 gives

$$V - V_0 \;=\; \mathcal{O}_{\prec}\big(\|W^\top W\|_{\mathsf{F}}/d^2\big), \qquad \operatorname{Var} V \;=\; \frac{2}{d^4}\,\|W^\top W\|_{\mathsf{F}}^2. \tag{188}$$

Step 3: Coupling. Define $Y_j := X_j \cdot \sqrt{V_0/V}$ for each j. Conditional on v_0, $\{Y_j\}$ inherit the Gaussianity and independence of $\{X_j\}$, with conditional variance $V \cdot (V_0/V) = V_0$. Since the conditional law $\mathcal{N}(0, V_0 \cdot I_m)$ is independent of v_0, the unconditional joint law of $(Y_1, \dots, Y_m)$ is i.i.d. $\mathcal{N}(0, V_0)$ as required.

The deviation factors as

$$X_j - Y_j \;=\; X_j\big(1 - \sqrt{V_0/V}\big) \;=\; X_j \cdot \frac{\sqrt{V} - \sqrt{V_0}}{\sqrt{V}}.$$

Two facts: conditional on v_0, the standardized form $X_j/\sqrt{V} \sim \mathcal{N}(0,1)$, so $|X_j/\sqrt{V}| = \mathcal{O}_\prec(1)$ uniformly in j; and $|\sqrt{V} - \sqrt{V_0}| \le |V - V_0|/\sqrt{V_0}$, by $\sqrt{V} + \sqrt{V_0} \ge \sqrt{V_0}$. Combined with (188),

$$|X_j - Y_j| \;=\; |X_j/\sqrt{V}| \cdot |\sqrt{V} - \sqrt{V_0}| \;=\; \mathcal{O}_\prec\big(\|W^\top W\|_{\mathsf{F}}/(d^2\sqrt{V_0})\big),$$

uniformly in j. For fixed m, we have

$$\|(X) - (Y)\|_2^2 = \sum_{j=1}^m |X_j - Y_j|^2 = \mathcal{O}_\prec\big((\|W^\top W\|_{\mathsf{F}}/(d^2\sqrt{V_0}))^2\big),$$

which gives (187). □

F.2 Random matrix with concentrated Frobenius norm

Corollary 1 (Bilinear Gaussian coupling, random matrix)**.** *Let W be a random $d \times d$ matrix with $\|W\|_{\mathsf{F}} > 0$, independent of $u_1, \dots, u_m, v_0$ which are i.i.d. $\mathcal{N}(0, I_d/d)$. Let $\nu_d^2 > 0$ be a deterministic constant. Suppose $\rho_f, \rho_{\mathrm{op}} > 0$ are deterministic rates such that*

$$\frac{\|W\|_{\mathsf{F}}^2}{d^2} - \nu_d^2 \;=\; \mathcal{O}_\prec(\rho_f), \qquad \frac{\|W^\top W\|_{\mathsf{F}}}{d^2} \;=\; \mathcal{O}_\prec(\rho_{\mathrm{op}}). \tag{189}$$

Set $X_j := u_j^\top W v_0$ for $j = 1, \dots, m$. On a common probability space, there exist i.i.d. random variables $Y_1, \dots, Y_m \sim \mathcal{N}(0, \nu_d^2)$ such that

$$\big\|(X_1, \dots, X_m) - (Y_1, \dots, Y_m)\big\|_2 \;=\; \mathcal{O}_\prec\big((\rho_{\mathrm{op}} + \rho_f)/\nu_d\big). \tag{190}$$

Proof. The key observation is that, conditional on (W, v_0), $X_j/\sqrt{V} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ across j, where $V := \|Wv_0\|^2/d$ is the conditional variance.

Define

$$Y_j \;:=\; X_j \sqrt{\nu_d^2/V}, \qquad j = 1, \dots, m.$$

Conditional on (W, v_0), $\{Y_j\}$ inherit the Gaussianity and independence of $\{X_j\}$ with variance $V \cdot \nu_d^2/V = \nu_d^2$, so the unconditional joint law of $(Y_1, \dots, Y_m)$ is i.i.d. $\mathcal{N}(0, \nu_d^2)$ as required. The deviation factors as $X_j - Y_j = X_j(1 - \sqrt{\nu_d^2/V}) = X_j\,(\sqrt{V} - \sqrt{\nu_d^2})/\sqrt{V}$, and using $|X_j/\sqrt{V}| = \mathcal{O}_\prec(1)$ together with $|\sqrt{V} - \sqrt{\nu_d^2}| \le |V - \nu_d^2|/\sqrt{\nu_d^2}$,

$$|X_j - Y_j| \;=\; |X_j/\sqrt{V}| \cdot |\sqrt{V} - \sqrt{\nu_d^2}| \;=\; \mathcal{O}_\prec\big(|V - \nu_d^2|/\sqrt{\nu_d^2}\big). \tag{191}$$

Decompose

$$V - \nu_d^2 \;=\; (V - V_0(W)) \;+\; (V_0(W) - \nu_d^2), \qquad V_0(W) := \|W\|_{\mathsf{F}}^2/d^2.$$

Hanson–Wright applied to $V - V_0(W)$ conditional on W gives

$$V - V_0(W) = \mathcal{O}_{\prec}\Big(\|W^\top W\|_{\mathsf{F}}/d^2\Big) = \mathcal{O}_{\prec}(\rho_{\mathrm{op}}),$$

using (188). The second piece is $\mathcal{O}_{\prec}(\rho_f)$ by (189). Combining with (191) and summing over j yields (190). □

G A trace estimate for the inverse Hessian

This appendix records several elementary estimates that will be used in the characterization of the susceptibility parameter χ. The first says that a fresh rank-one Gaussian direction probes the normalized trace of a bounded linear operator on $\mathbb{R}^{d\times d}$. The next two show that the normalized inverse trace is stable under the perturbations needed to simplify the TAM Hessian.

We will use the following moment notation. For a sequence of random variables X_d and deterministic positive numbers a_d, we write

$$X_d = O_{\mathcal{L}}(a_d)$$

if, for every fixed $p < \infty$, there is a constant $C_p < \infty$ such that $\|X_d\|_{L^p} \le C_p a_d$ for all sufficiently large d.

Lemma 21 (Fresh rank-one trace estimate)**.** *Let $u, v \in \mathbb{R}^d$ be independent Gaussian vectors with distribution $\mathcal{N}(0, I_d/d)$. Let $T : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ be a random symmetric linear operator, independent of (u, v), and assume that*

$$\|T\|_{\mathrm{op}} = O_{\mathcal{L}}(1).$$

Then

$$\Big\langle uv^\top,\, T[uv^\top]\Big\rangle = \frac{\mathrm{tr}(T)}{d^2} + O_{\mathcal{L}}(d^{-1/2}).$$

Equivalently, for every fixed $p < \infty$,

$$\left\| \Big\langle uv^\top,\, T[uv^\top]\Big\rangle - \frac{\mathrm{tr}(T)}{d^2} \right\|_{L^p} \le \frac{C_p}{\sqrt{d}}. \tag{192}$$

Proof. Write $u = g/\sqrt{d}$ and $v = h/\sqrt{d}$, where $g, h \in \mathbb{R}^d$ are independent standard Gaussian vectors. Define

$$F(g, h, T) \;:=\; \Big\langle uv^\top,\, T[uv^\top]\Big\rangle = \frac{1}{d^2}\big\langle g \otimes h,\, T[g \otimes h]\big\rangle.$$

Since T is independent of (g, h), we have

$$\mathbb{E}\big[F(g, h, T) \mid T\big] = \frac{\mathrm{tr}(T)}{d^2}.$$

We first bound the fluctuation in L^2. By the Gaussian Poincare inequality applied to the variables (g, h), with T kept fixed,

$$\mathrm{Var}\big(F(g,h,T) \mid T\big) \le \mathbb{E}\big[\|\nabla_g F\|^2 + \|\nabla_h F\|^2 \mid T\big]. \tag{193}$$

For fixed h, define the compressed operator

$$T_h := (I \otimes h)^* T (I \otimes h) \qquad \text{on } \mathbb{R}^d,$$

where $(I \otimes h)g = g \otimes h$. Then

$$F(g,h,T) = \frac{1}{d^2} g^\top T_h g, \qquad \nabla_g F = \frac{2}{d^2} T_h g,$$

using the symmetry of T. It follows that

$$\mathbb{E}\big[\|\nabla_g F\|^2 \mid T\big] = \frac{4}{d^4}\,\mathbb{E}\big[\mathrm{tr}(T_h^2) \mid T\big]. \tag{194}$$

Moreover,

$$\|T_h\|_{\mathrm{op}} \le \|T\|_{\mathrm{op}}\,\|h\|^2,$$

and hence

$$\mathrm{tr}(T_h^2) \le d\,\|T_h\|_{\mathrm{op}}^2 \le d\,\|T\|_{\mathrm{op}}^2\,\|h\|^4.$$

Since $\mathbb{E}\|h\|^4 = d^2 + 2d$, (194) gives

$$\mathbb{E}\big[\|\nabla_g F\|^2 \mid T\big] \le \frac{C\|T\|_{\mathrm{op}}^2}{d}. \tag{195}$$

The same argument, with the roles of g and h interchanged, gives

$$\mathbb{E}\big[\|\nabla_h F\|^2 \mid T\big] \le \frac{C\|T\|_{\mathrm{op}}^2}{d}. \tag{196}$$

Combining (193), (195), and (196), we obtain

$$\mathrm{Var}\big(F(g,h,T) \mid T\big) \le \frac{C\|T\|_{\mathrm{op}}^2}{d}. \tag{197}$$

We now upgrade this L^2 bound to fixed moments. The centered random variable

$$F(g,h,T) - \frac{\mathrm{tr}(T)}{d^2}$$

is, for fixed T, a Gaussian polynomial of degree at most 4 in (g, h). By Gaussian hypercontractivity, for every fixed $p < \infty$,

$$\left\| F(g,h,T) - \frac{\mathrm{tr}(T)}{d^2} \right\|_{L^p(g,h|T)} \le C_p \left\| F(g,h,T) - \frac{\mathrm{tr}(T)}{d^2} \right\|_{L^2(g,h|T)} \le \frac{C_p\|T\|_{\mathrm{op}}}{\sqrt{d}},$$

where the last step uses (197). Taking expectation over T and using $\|T\|_{\mathrm{op}} = O_{\mathcal{L}}(1)$ yields (192). This proves the lemma. □

We next record two deterministic perturbation estimates for the Hessian operators in (27)–(29). We identify $\mathbb{R}^{d\times d}$ with $\mathbb{R}^{d^2}$ throughout this discussion. If $A, B : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ are positive-definite self-adjoint operators, equivalently $d^2\times d^2$ symmetric matrices after vectorization, and $E := A-B$, then the resolvent identity gives the following bound: whenever $\|A^{-1}\|_{\mathrm{op}}, \|B^{-1}\|_{\mathrm{op}} \le \tau^{-1}$,

$$\left|\frac{1}{d^2}\operatorname{tr}(A^{-1}) - \frac{1}{d^2}\operatorname{tr}(B^{-1})\right| \le \frac{1}{\tau^2 d^2}\|E\|_* = \frac{1}{\tau^2 d^2}\|A-B\|_*, \tag{198}$$

where $\|\cdot\|_*$ denotes the nuclear norm on this d^2-dimensional operator space.

For a fixed cavity index h, let $H^{\setminus h}$ be the Hessian in (27). Define the simplified Hessian $\widetilde{H}^{\setminus h}$ by replacing the block $M_i^{\setminus h}$ in (29), for every $i \neq h$, by

$$\widetilde{M}_i^{\setminus h} := b_i^{\setminus h} e_i e_i^\top + a_i^{\setminus h}\operatorname{diag}\big(\widetilde{R}_{:,i}^{\setminus h}\big),$$

and keeping the h-th block equal to zero. Equivalently, $\widetilde{H}^{\setminus h}$ is obtained from $H^{\setminus h}$ by replacing $b_i^{\setminus h} Q_{:,i}^{\setminus h}(Q_{:,i}^{\setminus h})^\top$ by $b_i^{\setminus h} e_i e_i^\top$ in each active block.

Lemma 22 (Trace reduction to the simplified Hessian)**.** *Assume $\lambda n/d^2 \ge \tau > 0$ and $\max_i b_i^{\setminus h} \le B$. For $i \neq h$, write*

$$Q_{:,i}^{\setminus h} = -e_i + \rho_i^{\setminus h}, \qquad p_i^{\setminus h} := U\rho_i^{\setminus h}.$$

Then

$$\left|\frac{1}{d^2}\operatorname{tr}\Big((H^{\setminus h})^{-1}\Big) - \frac{1}{d^2}\operatorname{tr}\Big((\widetilde{H}^{\setminus h})^{-1}\Big)\right| \le \frac{B}{\tau^2 d^2}\sum_{i\neq h}\|v_i\|^2\left(2\|u_i\|\|p_i^{\setminus h}\| + \|p_i^{\setminus h}\|^2\right). \tag{199}$$

Consequently, if $n/d^2 = O(1)$, $\max_i\{\|u_i\|, \|v_i\|\} = O(1)$, and $\max_{i\neq h}\|p_i^{\setminus h}\| \le \eta$, then the normalized inverse trace of $H^{\setminus h}$ differs from that of the simplified Hessian $\widetilde{H}^{\setminus h}$ by $O(\eta)$.

Proof. Since

$$Q_{:,i}^{\setminus h}(Q_{:,i}^{\setminus h})^\top - e_i e_i^\top = -e_i(\rho_i^{\setminus h})^\top - \rho_i^{\setminus h} e_i^\top + \rho_i^{\setminus h}(\rho_i^{\setminus h})^\top,$$

the ith summand in $H^{\setminus h} - \widetilde{H}^{\setminus h}$ is

$$\Delta \mapsto b_i^{\setminus h}\left[-u_i(p_i^{\setminus h})^\top - p_i^{\setminus h}u_i^\top + p_i^{\setminus h}(p_i^{\setminus h})^\top\right]\Delta\, v_i v_i^\top.$$

Each of the three displayed terms is rank one as an operator on $\mathbb{R}^{d\times d}$. For example,

$$u_i(p_i^{\setminus h})^\top \Delta\, v_i v_i^\top = \big\langle p_i^{\setminus h} v_i^\top, \Delta\big\rangle\, u_i v_i^\top,$$

and hence its nuclear norm is $\|u_i v_i^\top\|_{\mathsf{F}}\|p_i^{\setminus h} v_i^\top\|_{\mathsf{F}} = \|u_i\|\|p_i^{\setminus h}\|\|v_i\|^2$. The same bound holds for the transposed cross term, while the $p_i^{\setminus h}(p_i^{\setminus h})^\top$ term contributes $\|p_i^{\setminus h}\|^2\|v_i\|^2$. Thus

$$\|H^{\setminus h} - \widetilde{H}^{\setminus h}\|_* \le B\sum_{i\neq h}\|v_i\|^2\left(2\|u_i\|\|p_i^{\setminus h}\| + \|p_i^{\setminus h}\|^2\right).$$

Both $H^{\setminus h}$ and $\widetilde{H}^{\setminus h}$ are bounded below by the ridge term $(\lambda n/d^2)I$, so their inverse operator norms are at most τ^{-1}. Applying (198) gives (199). □

Lemma 23 (Stability at the natural coefficient scale)**.** *Let $\widetilde{H}^{\setminus h}$ be the simplified Hessian above. Let $\bar{H}^{\setminus h}$ be the simplified Hessian obtained by replacing, for $i \neq h$, the coefficients $a_i^{\setminus h}$ and $\widetilde{R}_{j,i}^{\setminus h}$ by $\bar{a}_i$ and $\bar{R}_{j,i}$, while keeping the coefficients $b_i^{\setminus h}$ unchanged. Assume*

$$\frac{\lambda n}{d^2} \geq \tau > 0, \qquad 0 \leq a_i^{\setminus h}, \bar{a}_i \leq B, \qquad \bar{R}_{j,i} \geq 0,$$

and

$$\max_{i \neq h} |\bar{a}_i - a_i^{\setminus h}| \leq \delta_a, \qquad \max_{i \neq h} \sum_{j \neq i,h} \widetilde{R}_{j,i}^{\setminus h} \leq C_R, \qquad \max_{j \neq i,h} |\bar{R}_{j,i} - \widetilde{R}_{j,i}^{\setminus h}| \leq \frac{\delta_R}{k}.$$

Then

$$\begin{aligned} &\left| \frac{1}{d^2} \operatorname{tr}\Big((\bar{H}^{\setminus h})^{-1}\Big) - \frac{1}{d^2} \operatorname{tr}\Big((\widetilde{H}^{\setminus h})^{-1}\Big) \right| \\ &\qquad \leq \frac{1}{\tau^2 d^2} \sum_{i \neq h} \|v_i\|^2 \sum_{j \neq i,h} \left(\delta_a \widetilde{R}_{j,i}^{\setminus h} + \frac{B \delta_R}{k} \right) \|u_j\|^2. \end{aligned} \tag{200}$$

Consequently, if $n/d^2 = O(1)$, $n/k = O(1)$, $\max_i\{\|u_i\|, \|v_i\|\} = O(1)$, and $C_R = O(1)$, then the normalized inverse trace changes by

$$O(\delta_a + \delta_R).$$

Proof. Each elementary map

$$\Delta \mapsto u_j u_j^\top \Delta \, v_i v_i^\top = \big\langle u_j v_i^\top, \Delta \big\rangle \, u_j v_i^\top$$

has rank one and nuclear norm $\|u_j\|^2 \|v_i\|^2$. Therefore

$$\|\bar{H}^{\setminus h} - \widetilde{H}^{\setminus h}\|_* \leq \sum_{i \neq h} \sum_{j \neq i,h} |\bar{a}_i \bar{R}_{j,i} - a_i^{\setminus h} \widetilde{R}_{j,i}^{\setminus h}| \, \|u_j\|^2 \|v_i\|^2. \tag{201}$$

The coefficient difference satisfies

$$|\bar{a}_i \bar{R}_{j,i} - a_i^{\setminus h} \widetilde{R}_{j,i}^{\setminus h}| \leq |\bar{a}_i - a_i^{\setminus h}| \, \widetilde{R}_{j,i}^{\setminus h} + \bar{a}_i \, |\bar{R}_{j,i} - \widetilde{R}_{j,i}^{\setminus h}| \leq \delta_a \widetilde{R}_{j,i}^{\setminus h} + \frac{B \delta_R}{k}.$$

Substituting this into (201), and using that both Hessians are bounded below by the ridge term $(\lambda n/d^2) I$, (198) proves (200). □

Quadratic forms of the simplified Hessian

We now use the simplified Hessian to extract the scalar equation for χ. In this subsection we suppress the leave-one-out superscript and write

$$f_{ji} := \operatorname{vec}(u_j v_i^\top), \qquad r_{j,i} := \beta q_{j,i}(1 - q_{j,i}), \qquad \widetilde{R}_{j,i} = \frac{r_{j,i}}{k}.$$

The full simplified Hessian is

$$\widetilde{H} = \frac{\lambda n}{d^2} I + \sum_{i=1}^{n} b_i f_{ii} f_{ii}^\top + \sum_{i=1}^{n} \sum_{j \neq i} \frac{a_i r_{j,i}}{k} f_{ji} f_{ji}^\top.$$

Here $a_i, b_i, q_{j,i}$ are evaluated at the full minimizer. For the i-cavity problem, superscript (i) denotes the corresponding quantities after removing sample i from both tensor legs.

Feature spectral bounds. We use the following spectral estimates for the feature matrices: uniformly in i and ℓ,

$$\left\| [f_{\ell\ell}]_{\ell\neq i} \right\|_{\mathrm{op}} = \mathcal{O}_{\prec}(1), \qquad \left\| [f_{j\ell}]_{j\neq \ell,i} \right\|_{\mathrm{op}} = \mathcal{O}_{\prec}(\sqrt{d}). \tag{202}$$

Lemma 24 (LOO coefficient stability)**.** *Assume (A1)–(A3). Then, uniformly in i,*

$$\max_{\ell\neq i} |b_\ell - b_\ell^{(i)}| = \mathcal{O}_{\prec}(d^{-1}), \tag{203}$$

and

$$\max_{\ell\neq i} \max_{j\neq \ell,i} \left| a_\ell r_{j,\ell} - a_\ell^{(i)} r_{j,\ell}^{(i)} \right| = \mathcal{O}_{\prec}(d^{-1}). \tag{204}$$

Proof. The functions $a = \sigma(-m)$, $b = a(1-a)$, and $r = \beta q(1-q)$ are uniformly Lipschitz on the bounded-score event supplied by (A3); the corresponding LOO scores are bounded by (A2). The LOO perturbations of the relevant scores and thresholds are $\mathcal{O}_{\prec}(d^{-1})$ by (A1)–(A2), since $n \asymp d^2$ and the μ perturbation is smaller. Applying the Lipschitz bounds entrywise gives (203)–(204). □

Lemma 25 (Diagonal quadratic form)**.** *Assume (A1)–(A6) and the feature spectral bounds (202). Let χ_d be as in (49). Then, uniformly in i,*

$$f_{ii}^\top \widetilde{H}^{-1} f_{ii} = \frac{\chi_d}{1 + b_i \chi_d} + \mathcal{O}_{\prec}(d^{-1/2}). \tag{205}$$

Proof. Let H_i be the simplified i-cavity Hessian that keeps only feature directions $f_{j\ell}$ with $j, \ell \neq i$, and uses the i-cavity coefficients:

$$H_i = \frac{\lambda n}{d^2} I + \sum_{\ell\neq i} b_\ell^{(i)} f_{\ell\ell} f_{\ell\ell}^\top + \sum_{\ell\neq i} \sum_{j\neq \ell,i} \frac{a_\ell^{(i)} r_{j,\ell}^{(i)}}{k} f_{j\ell} f_{j\ell}^\top.$$

This operator is independent of (u_i, v_i). Set

$$A_i := H_i + b_i f_{ii} f_{ii}^\top.$$

We first compare $\widetilde{H}$ with A_i. Write

$$\widetilde{H} = A_i + C_i^{\mathrm{feat}} + C_i^{\mathrm{coef}},$$

where

$$C_i^{\mathrm{feat}} = \sum_{\ell\neq i} \frac{a_\ell r_{i,\ell}}{k} f_{i\ell} f_{i\ell}^\top + \sum_{j\neq i} \frac{a_i r_{j,i}}{k} f_{ji} f_{ji}^\top$$

contains the feature directions involving u_i or v_i, except for the diagonal rank-one term, and

$$C_i^{\mathrm{coef}} = \sum_{\ell\neq i} (b_\ell - b_\ell^{(i)}) f_{\ell\ell} f_{\ell\ell}^\top + \sum_{\ell\neq i} \sum_{j\neq \ell,i} \frac{a_\ell r_{j,\ell} - a_\ell^{(i)} r_{j,\ell}^{(i)}}{k} f_{j\ell} f_{j\ell}^\top.$$

The resolvent identity gives

$$f_{ii}^\top \widetilde{H}^{-1} f_{ii} - f_{ii}^\top A_i^{-1} f_{ii} = -f_{ii}^\top \widetilde{H}^{-1}(C_i^{\mathrm{feat}} + C_i^{\mathrm{coef}}) A_i^{-1} f_{ii}.$$

By Woodbury,

$$A_i^{-1} f_{ii} = \frac{H_i^{-1} f_{ii}}{1 + b_i f_{ii}^\top H_i^{-1} f_{ii}},$$

and the denominator is at least one. Thus it suffices to bound

$$f_{ii}^\top \widetilde{H}^{-1}(C_i^{\mathrm{feat}} + C_i^{\mathrm{coef}}) H_i^{-1} f_{ii}. \tag{206}$$

Let $\xi_i := \widetilde{H}^{-1} f_{ii}$. The ridge lower bound gives $\|\xi_i\| = \mathcal{O}_\prec(1)$. For the two terms in C_i^{feat}, each summand has a factor of the form

$$f_{i\ell}^\top H_i^{-1} f_{ii} \quad \text{or} \quad f_{ji}^\top H_i^{-1} f_{ii}.$$

In the first expression v_i is an unmatched Gaussian vector, and in the second expression u_i is unmatched. Equivalently, if $M = \mathrm{mat}(H_i^{-1} f_{i\ell})$ or $M = \mathrm{mat}(H_i^{-1} f_{ji})$, then $\|M\|_{\mathsf{F}} = \mathcal{O}_\prec(1)$ and the expression is $u_i^\top M v_i$ with one Gaussian leg independent of M and the other leg. Hence each such factor is $\mathcal{O}_\prec(d^{-1/2})$. Since the coefficients have total scale $n/k = O(1)$, the contribution of C_i^{feat} to (206) is $\mathcal{O}_\prec(d^{-1/2})$.

We next consider C_i^{coef}. The diagonal coefficient part can be written as

$$\big[f_{\ell\ell}\big]_{\ell\neq i} \,\mathrm{diag}(b_\ell - b_\ell^{(i)})_{\ell\neq i} \big[f_{\ell\ell}\big]_{\ell\neq i}^\top.$$

By Lemma 24 and (202), its operator norm is $\mathcal{O}_\prec(d^{-1})$, so its contribution to (206) is $\mathcal{O}_\prec(d^{-1})$.

For the off-diagonal coefficient part, fix $\ell \neq i$ and write

$$F_\ell^{(i)} := \big[f_{j\ell}\big]_{j\neq \ell, i}, \qquad D_\ell^{(i)} := \mathrm{diag}\Big(a_\ell r_{j,\ell} - a_\ell^{(i)} r_{j,\ell}^{(i)}\Big)_{j\neq \ell, i}.$$

The corresponding block is $k^{-1} F_\ell^{(i)} D_\ell^{(i)} (F_\ell^{(i)})^\top$. Lemma 24 gives $\|D_\ell^{(i)}\|_{\mathrm{op}} = \mathcal{O}_\prec(d^{-1})$, while (202) gives $\|F_\ell^{(i)}\|_{\mathrm{op}} = \mathcal{O}_\prec(\sqrt{d})$. Moreover,

$$\Big\| (F_\ell^{(i)})^\top H_i^{-1} f_{ii} \Big\| = \mathcal{O}_\prec(1),$$

because H_i and $F_\ell^{(i)}$ are independent of f_{ii} and

$$\frac{1}{d^2} \mathrm{tr}\Big(H_i^{-1} F_\ell^{(i)} (F_\ell^{(i)})^\top H_i^{-1} \Big) = \mathcal{O}_\prec(1).$$

Therefore the fixed-ℓ contribution is $\mathcal{O}_\prec(1/(k\sqrt{d}))$, and summing over ℓ gives $\mathcal{O}_\prec(d^{-1/2})$. We have proved

$$f_{ii}^\top \widetilde{H}^{-1} f_{ii} = f_{ii}^\top A_i^{-1} f_{ii} + \mathcal{O}_\prec(d^{-1/2}). \tag{207}$$

It remains to evaluate the right-hand side. Woodbury gives

$$f_{ii}^\top A_i^{-1} f_{ii} = \frac{x_i}{1 + b_i x_i}, \qquad x_i := f_{ii}^\top H_i^{-1} f_{ii}.$$

Since H_i is independent of f_{ii}, Lemma 21, (A6), and the finite-removal trace stability transferred to the simplified Hessian by the perturbation estimates above give

$$x_i = \chi_d + \mathcal{O}_{\prec}(d^{-1/2}).$$

The map $x \mapsto x/(1 + b_i x)$ is uniformly Lipschitz for $b_i \geq 0$, and (205) follows from (207). □

Lemma 26 (Off-diagonal quadratic form)**.** *Under the assumptions of Lemma 25, uniformly for $j \neq i$,*

$$f_{ji}^\top \widetilde{H}^{-1} f_{ji} = \chi_d + \mathcal{O}_{\prec}(d^{-1/2}).$$

Proof. The proof is the same as that of Lemma 25. One removes all terms involving u_j or v_i to create a cavity independent of f_{ji}. The direct Woodbury correction along f_{ji} has coefficient $a_i r_{j,i}/k = O(k^{-1})$ and is therefore absorbed into the $\mathcal{O}_{\prec}(d^{-1/2})$ error. The feature-removal and coefficient-removal errors are controlled by the same unmatched-Gaussian and feature-matrix estimates used above. □

Remark 10 (Use for χ)**.** *In the leave-one-out analysis, the operator T will be a Hessian inverse, for example $T = (H^{\setminus i})^{-1}$, evaluated on a sigma-field that is independent of a fresh pair (u_i, v_i). Lemma 21 then gives*

$$\left\langle u_i v_i^\top,\, (H^{\setminus i})^{-1}[u_i v_i^\top] \right\rangle = \frac{1}{d^2} \operatorname{tr}\Big((H^{\setminus i})^{-1}\Big) + O_{\mathcal{L}}(d^{-1/2}),$$

provided $\|(H^{\setminus i})^{-1}\|_{\mathrm{op}} = O_{\mathcal{L}}(1)$. Thus the susceptibility can be characterized through the normalized trace of the inverse cavity Hessian. Lemmas 22 and 23 then justify replacing the original Hessian by simpler approximations without changing this normalized trace at leading order.

H Variational formulation of the scalar equations

This appendix verifies the variational representation used in Section 4.3. Throughout, Z and G denote independent standard normal variables, and $\nu, \chi > 0$.

H.1 The scalar TAM functional

Recall

$$c_r(\nu) = \min_{\mu \in \mathbb{R}} \left\{ \mu + \frac{1}{r} \mathbb{E}_Z \left[\phi_\beta(\nu Z - \mu) \right] \right\}.$$

The minimizer, denoted $\mu_r(\nu)$, is characterized by

$$r = \mathbb{E}_Z \left[\sigma\big(\beta(\nu Z - \mu_r(\nu))\big) \right].$$

$$c_r'(\nu) = \frac{1}{r} \mathbb{E}_Z \left[Z\, \phi_\beta'\big(\nu Z - \mu_r(\nu)\big) \right].$$

Stein's identity gives

$$c_r'(\nu) = \frac{\nu}{r} \mathbb{E}_Z \left[\phi_\beta''\big(\nu Z - \mu_r(\nu)\big) \right]. \tag{208}$$

We will also use that c_r is convex in ν. Indeed, the function

$$(\nu, \mu) \mapsto \mu + \frac{1}{r}\mathbb{E}_Z\left[\phi_\beta(\nu Z - \mu)\right]$$

is jointly convex, and partial minimization over μ preserves convexity.

H.2 Moreau envelope identities

Let $S = S_{\nu,\chi}(G)$ be the minimizer in (42), and set

$$A := \sigma\big(-S + c_r(\nu)\big), \qquad B := A(1 - A).$$

The first-order condition in s is

$$S = \nu G + \chi A. \tag{209}$$

Envelope differentiation in χ gives

$$\partial_\chi M_r(\nu, \chi) = -\frac{1}{2}\mathbb{E}[A^2]. \tag{210}$$

For the derivative in ν, the envelope theorem gives

$$\partial_\nu M_r(\nu, \chi) = -\mathbb{E}[AG] + \mathbb{E}[A]c_r'(\nu). \tag{211}$$

Differentiating (209) with respect to the standard Gaussian coordinate G yields

$$\frac{dA}{dG} = -\frac{\nu B}{1 + \chi B}.$$

Therefore Stein's identity gives

$$\mathbb{E}[AG] = -\nu\,\mathbb{E}\left[\frac{B}{1 + \chi B}\right]. \tag{212}$$

Combining (208), (211), and (212),

$$\partial_\nu M_r(\nu, \chi) = \nu\,\mathbb{E}\left[\frac{B}{1 + \chi B}\right] + \frac{\nu}{r}\mathbb{E}[A]\,\mathbb{E}_Z\left[\phi_\beta''\big(\nu Z - \mu_r(\nu)\big)\right]. \tag{213}$$

H.3 Saddle-point equations

The saddle potential is

$$\mathcal{P}_{\alpha,r}(\nu, \chi) = \frac{\lambda}{2}\nu^2 - \frac{\nu^2}{2\alpha\chi} + M_r(\nu, \chi).$$

Using (210),

$$\partial_\chi \mathcal{P}_{\alpha,r}(\nu, \chi) = \frac{\nu^2}{2\alpha\chi^2} - \frac{1}{2}\mathbb{E}[A^2]. \tag{214}$$

Hence the χ saddle-point equation is

$$\nu^2 = \alpha\chi^2\mathbb{E}[A^2]. \tag{215}$$

Using (213),

$$\partial_\nu \mathcal{P}_{\alpha,r}(\nu,\chi) = \lambda\nu - \frac{\nu}{\alpha\chi} + \nu\,\mathbb{E}\left[\frac{B}{1+\chi B}\right] + \frac{\nu}{r}\mathbb{E}[A]\,\mathbb{E}_Z\left[\phi_\beta''\big(\nu Z - \mu_r(\nu)\big)\right].$$

Since $\nu > 0$, the ν saddle-point equation is

$$\frac{1}{\alpha} = \lambda\chi + \chi\,\mathbb{E}\left[\frac{B}{1+\chi B}\right] + \frac{\chi}{r}\mathbb{E}[A]\,\mathbb{E}_Z\left[\phi_\beta''\big(\nu Z - \mu_r(\nu)\big)\right]. \tag{216}$$

This is exactly the χ-equation obtained from the zero-residual limit of (68).

It remains to connect (215) with the variance equation. From (209),

$$\mathbb{E}[AS] = \nu\mathbb{E}[AG] + \chi\mathbb{E}[A^2].$$

Using (212) and (215),

$$\mathbb{E}[AS] = -\nu^2\mathbb{E}\left[\frac{B}{1+\chi B}\right] + \frac{\nu^2}{\alpha\chi}.$$

Substituting (216) into the last display gives

$$\mathbb{E}[AS] = \nu^2\left[\lambda + \frac{1}{r}\mathbb{E}[A]\,\mathbb{E}_Z\left[\phi_\beta''\big(\nu Z - \mu_r(\nu)\big)\right]\right],$$

which is the corresponding ν-equation.

H.4 Maximization in χ and convexity in ν

For fixed ν, the inner problem is concave after the change of variables $t = 1/\chi$. Indeed,

$$\mathcal{P}_{\alpha,r}(\nu, 1/t) = \frac{\lambda}{2}\nu^2 - \frac{\nu^2}{2\alpha}t + \mathbb{E}\min_{s\in\mathbb{R}}\left\{\frac{t}{2}(s-\nu G)^2 + \ell(s - c_r(\nu))\right\}.$$

The last term is the infimum of affine functions of t, and hence is concave in t. Thus the inner extremization may equivalently be viewed as a one-dimensional concave maximization in $1/\chi$.

For fixed ν, differentiating $A = \sigma(-S + c_r(\nu))$ with respect to χ gives

$$\partial_\chi A = -\frac{AB}{1+\chi B}.$$

Consequently,

$$\frac{d}{d\chi}\left(\chi^2\mathbb{E}[A^2]\right) = 2\chi\,\mathbb{E}\left[\frac{A^2}{1+\chi B}\right] > 0.$$

Moreover $\chi^2\mathbb{E}[A^2] \to 0$ as $\chi \downarrow 0$, while $\chi^2\mathbb{E}[A^2] \to \infty$ as $\chi \to \infty$. By (214), $\partial_\chi\mathcal{P}$ changes sign exactly once, from positive to negative. Hence $\chi \mapsto \mathcal{P}_{\alpha,r}(\nu,\chi)$ has a unique global maximizer.

Finally we prove strong convexity of the reduced potential $\mathcal{V}_{\alpha,r}$ defined in (43). The logistic loss has the Fenchel representation

$$\ell(t) = \sup_{a\in[0,1]}\left\{-at - I(a)\right\}, \qquad I(a) := a\log a + (1-a)\log(1-a), \tag{217}$$

with the convention $0 \log 0 = 0$. Applying this pointwise in the Moreau envelope and minimizing over s gives

$$M_r(\nu, \chi) = \sup_{a(\cdot)\in[0,1]} \mathbb{E}\left[a(G)\big(c_r(\nu) - \nu G\big) - I(a(G)) - \frac{\chi}{2}a(G)^2\right]. \tag{218}$$

Taking the supremum over $\chi > 0$ yields

$$\mathcal{V}_{\alpha,r}(\nu) = \frac{\lambda}{2}\nu^2 + \sup_{a(\cdot)\in[0,1]} \left\{\mathbb{E}\left[a(G)\big(c_r(\nu) - \nu G\big) - I(a(G))\right] - \frac{\nu}{\sqrt{\alpha}}\left(\mathbb{E}[a(G)^2]\right)^{1/2}\right\}. \tag{219}$$

For each fixed measurable $a(\cdot)$, the expression inside the supremum is a convex function of ν, because c_r is convex and $\mathbb{E}[a(G)] \geq 0$; adding $\lambda\nu^2/2$ makes it λ-strongly convex. The supremum of λ-strongly convex functions is λ-strongly convex. This proves the convexity and uniqueness assertions in Proposition 10.

Remark 11 (Meaning of the dual optimizer)**.** *The optimizer $a(G)$ in (218) is the scalar limit of the sample-wise logistic weights. Indeed, equality in the Fenchel representation (217) is attained at*

$$a = \sigma(-t).$$

Taking $t = S - c_r(\nu)$, where $S = S_{\nu,\chi}(G)$ is the proximal minimizer, gives

$$a_*(G) = \sigma\big(-S_{\nu,\chi}(G) + c_r(\nu)\big).$$

Equivalently, the proximal first-order condition gives

$$S_{\nu,\chi}(G) = \nu G + \chi a_*(G).$$

Thus the dual optimizer is the limiting counterpart of the finite-dimensional coefficients $a_i = \sigma(-m_i)$.