Factual recall in linear associative memories: sharp asymptotics and mechanistic insights

Alessio Giorlandino
International School of Advanced Studies (SISSA)
Trieste, Italy
agiorlan@sissa.it

Sebastian Goldt
International School of Advanced Studies (SISSA)
Trieste, Italy
sgoldt@sissa.it

Antoine Maillard
INRIA Paris & DI ENS
PSL University
Paris, France
antoine.maillard@inria.fr

Abstract

Large language models excel at factual recall, yet the fundamental limits of storing input–output associations with neural networks remain unclear. We study these limits in a minimal setting: a linear associative memory that maps p input embeddings in $\mathbb{R}^d$ to their d-dimensional targets via a linear layer, requiring each mapped input to score its own target above all other outputs. Since all inputs compete for the same outputs, these constraints are strongly correlated, which makes the storage capacity p_c — the largest number of associations that can be stored — hard to characterise. We address this in three steps. First, we introduce a decoupled model, in which each input competes with its own independent set of outputs, and provide numerical and analytical evidence that it is equivalent to the original model, both in capacity and in the structure of the learnt solution. Second, using tools from statistical physics, we show that the decoupled model has capacity $p_c \log p_c / d^2 = 1/2$, extend this computation to linear two-layer networks, and go beyond the capacity by predicting the spectrum of the learnt weights and the storage mechanism. The latter explains how the optimal solution outperforms a naïve Hebbian rule: rather than boosting input–output alignments with broad fluctuations, it raises the correct scores just above the extreme-value threshold set by the competing outputs. Third, we prove the decoupled model's capacity thresholds rigorously, including in the non-convex two-layer case; combined with a recent proof for the original model, this shows that both models have the same capacity for full-rank weights. These findings give a sharp characterisation of factual storage in linear networks, and suggest decoupling as a general tool for memorisation and learning problems in which each input must be matched to one of an extensive number of outputs.

1 Introduction

Large language models have a remarkable capacity to memorise factual associations from their training data [Petroni et al., 2019]. This observation has motivated a growing body of work aimed at localising and quantifying memorised facts within such models [Geva et al., 2021, Meng et al., 2022]. These empirical findings also raise fundamental theoretical questions, among them: how many distinct associations can neural networks reliably store, and what are the absolute limits governing factual memory in modern learning systems [Roberts et al., 2020, Allen-Zhu and Li, 2025]?

In this paper, we seek to understand the fundamental limits of learning a set of associations with a simple neural network. We illustrate the task on the left of Fig. 1: the task is to learn the mapping of N input tokens to M output tokens according to some unknown ground-truth rule $f^* : [N] \to [M]$. We are given embeddings $\{e_x\}_{x\in[N]} \subset \mathbb{R}^d$ and $\{u_y\}_{y\in[M]} \subset \mathbb{R}^d$ for the input and output vocabularies respectively, and the goal is to learn a parametric model $F_W : \mathbb{R}^d \to \mathbb{R}^d$ that, given e_x, correctly

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

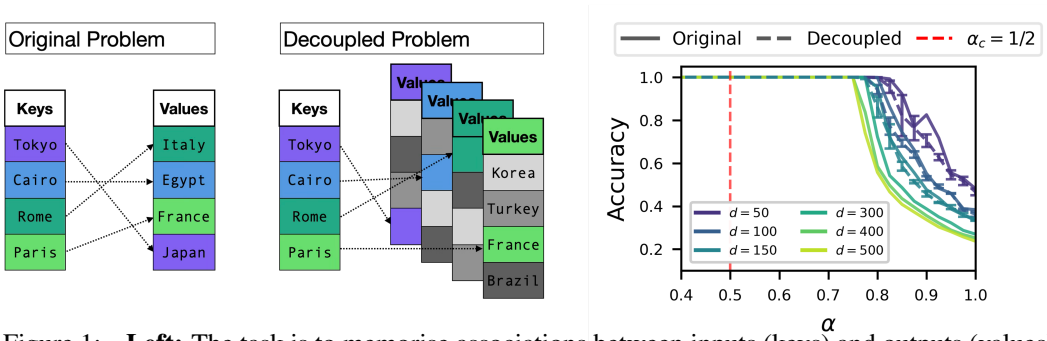


Figure 1: **Left:** The task is to memorise associations between inputs (keys) and outputs (values), illustrated here as cities and their countries. In the *original problem*, all inputs share a common set of outputs. In the *decoupled problem*, each input has its own independent set of competing outputs. **Right:** Empirical accuracy as a function of the load parameter $\alpha = p \log p / d^2$, shown for the original problem (1) (solid), the decoupled problem (2) (dashed), across several embedding dimensions d. The original and decoupled problems appear consistently equally hard, suggesting a common capacity threshold. Models are trained with Adam on the cross-entropy loss (see Section A). The red dashed line marks the asymptotic threshold $\alpha_c = 1/2$ of both problems (Corollary 1), which the transitions approach slowly due to finite-size effects of order $O((\log p)^{-1})$, see Section 5.

associates the input to the corresponding output token, via the rule $\arg\max_{y \in [M]} u_y^\top F_W(e_x) = f^*(x)$. This setting was first introduced by Cabannes et al. [2024a], and the simplest instance, which we refer to as *linear associative memory*, consists in taking F_W as a linear map, i.e. $F_W(e_x) = W e_x$ with $W \in \mathbb{R}^{d \times d}$. This simple setting, which we consider in our work, already exhibits some of the core phenomenology of associative memories, and is the subject of a very active line of research [Cabannes et al., 2024a,b, Nichani et al., 2025, Vural et al., 2026, Kim et al., 2026].

More precisely, Cabannes et al. [2024a,b] study scaling laws and training dynamics in the regime where the number of input patterns N scales with the embedding dimension d, while the number of possible outputs is $M = O(1)$, which is closer to a typical classification setting with high-dimensional inputs and a finite number of classes. Vural et al. [2026] and Kim et al. [2026] study the early-stage learning dynamics of linear associative memories with various models for the embeddings, but do not address the fundamental capacity limits of these architectures. To the best of our knowledge, Nichani et al. [2025] were the first to study the capacity of associative memories in the more realistic case of a large number of output tokens M in a *high-dimensional limit*, where we have an injective rule $f^\star$, Gaussian input and output embeddings with dimension d, and a number of associations $p := N = M$. They show that if $d^2 \gtrsim p \operatorname{polylog}(p)$, then with high probability over the draw of the embeddings, the *Hebbian ansatz* for the weight matrix $W_{\text{Hebb}} := \sum_{z \in [p]} u_{f^*(z)} e_z^\top$ satisfies all the association constraints, and that it can be obtained from a single step of gradient descent on the correlation loss. They also provide numerical results for the capacity achieved by the minimiser of the cross-entropy loss, which suggest that the storage capacity transition occurs in the regime $d^2 \sim p \log p$.

In this work, we consider the same setting as Nichani et al. [2025] and crucially go beyond the Hebbian ansatz. Our goal is to provide sharp thresholds for the maximum number of associations that can be stored as a function of the embedding dimension d. Since we are interested in the optimal weights W beyond a specific ansatz, we have to characterise the volume of the space of weights that satisfy a given set of associations. This perspective is inspired by classical works on the storage capacity of single-layer feed-forward neural networks for binary classification, see Gardner and Derrida [1988] and the further references discussed below. Compared to storing binary labels for a set of high-dimensional inputs in a single neuron, the associative memory problem raises significant new challenges. Rather than assigning binary labels, each input must be associated with one among many candidate outputs, and the learnable parameters form a weight matrix rather than a weight vector. The main difficulty in determining the optimal storage capacity lies in the strong correlations between the constraints of different associations.

Our main contribution is the introduction of a "decoupled" version of the storage problem, which allows us to derive a sharp capacity threshold of $p_c \log p_c / d^2 = 1/2$ for linear associative memories, and to show that this limit is achieved by a storage mechanism that is fundamentally different than the

Hebbian construction. We prove this threshold rigorously for the decoupled problem, together with its generalisation to two-layer linear networks; combined with the proof of Barnfield et al. [2026] for the original problem, which appeared after the first version of this work, this shows that the two problems have the same capacity for full-rank matrices.

More precisely, our main contributions are as follows:

1. **Decoupled formulation:** We introduce a decoupled variant of the associative memory problem in which the constraints are independent in Section 2, see Fig. 1 for an illustration.
2. **Evidence for the equivalence**: We conjecture that this variant is equivalent to the original associative memory in the high-dimensional limit, based on three strands of evidence:
   - (i) Both models exhibit the same capacity, see Section 3.1 and Fig. 2;
   - (ii) The optimal solution in both models has the same asymptotic singular value distribution, see Section 3.2 and Fig. 3;
   - (iii) Both models exhibit the same storage mechanism, see Section 3.3 and Fig. 4.
3. **Mechanistic insights**: We explain how the optimal solution beats the Hebbian solution by boosting the correct output scores just above a deterministic threshold, see Section 4.
4. **Sharp capacity characterisation:** Using analytic techniques from statistical physics, we derive an exact expression for the optimal storage capacity of the linear associative memory in terms of the load parameter $\alpha := p \log p / d^2$ (recall p is the number of associations and d the embedding dimension) and show good agreement with numerical simulations, see Section 5.
5. **Two-layer (rank-constrained) case:** We study matrices $W = QR^\top$ with $Q, R \in \mathbb{R}^{d \times m}$, corresponding to two linear layers with a smaller hidden dimension of size $m = \kappa d$. For generic $\kappa \in (0, 1]$, we derive the sharp capacity threshold, generalising the full-rank case, and quantify how the inner layer size affects the memorisation capacity, see Section 5.3. We also give a theoretical prediction for the singular value distribution of the optimal solution close to capacity, showcasing excellent agreement with numerical simulations, see Fig. 3.
6. **Rigorous capacity threshold:** We prove that, for every $\kappa \in (0, 1]$, the capacity threshold of the decoupled problem is exactly the one predicted by the replica-symmetric computation, although for $\kappa < 1$ the problem is non-convex, see Theorem 1. The achievability part also holds for the original problem, which can therefore store at least as many associations as the decoupled one.

Collectively, our results offer the first precise characterisation of the capacity of linear associative memories in the high-dimensional regime, providing new insights into the mechanisms and the fundamental limits of factual memory in neural networks.

Further related work

Storage capacity in supervised learning A classical line of work characterises the storage capacity of neural networks. The capacity of a perceptron, a single neuron used for binary classification, can be sharply characterised and is linear in the input dimension, $p \sim d$, for a variety of rules [Cover, 1965, Gardner and Derrida, 1988, Krauth and Mézard, 1989]. More recently, non-convex variants of the problem have been studied in both statistical physics [Franz et al., 2017] and probability theory [Montanari et al., 2024, Huang, 2024]. In factual recall, rather than assigning binary labels, the learnable parameters form a weight matrix rather than a vector, meaning that in our setting the maximal number of associations p scales with the embedding dimension as $d^2 \sim p \log p$. The quadratic dependence reflects the d^2 degrees of freedom of the matrix W, while the $\log p$ factor arises because each constraint involves an optimisation over p competing outputs.

Storage capacity of associative memories The canonical example of an auto-associative memory are Hopfield networks [Hopfield, 1982], where the goal is to store a number of patterns as fixed points of the network dynamics. The Hopfield model can store $p \simeq 0.14d$ patterns, a result that was first obtained using techniques from statistical physics by Amit et al. [1985]. More recently, it was shown that *dense* associative memories with non-polynomial activation functions can store an exponential number of patterns [Krotov and Hopfield, 2016, Demircigil et al., 2017, Krotov and Hopfield, 2021, Ramsauer et al., 2021]; notably, Lucibello and Mézard [2024] provided sharp asymptotics on the capacity using tools from statistical physics. Rather than simply storing patterns, here we instead

seek to map inputs to distinct outputs as in classical correlation-matrix memories [Anderson, 1972, Kohonen, 1972, Kosko, 1988] and modern key–value retrieval architectures [Weston et al., 2015].

Parallel work This work was conducted in parallel with an independent study by Barnfield et al. [2026] that also explores the optimal capacity of linear associative memories. In the first version of their work, Barnfield et al. [2026] prove that the capacity threshold scales as $d^2 \sim p \log p$, obtain explicit bounds for the critical capacity on this scale, and conjecture its exact value $\alpha_c = 1/2$. In a revised version, which appeared after the first version of the present work, they prove rigorously that $\alpha_c = 1/2$ for the original problem, confirming the threshold we derive from the decoupled model. Building on their approach, we prove the capacity threshold for rank-constrained weights in the decoupled problem, and its achievability in the original one (Theorem 1). Beyond the argmax objective, they also introduce a convex Tail-Average Margin relaxation in the regime $p \asymp d^2$, for which they derive exact asymptotic predictions and an explicit phase transition in the ridgeless limit. In our work, we take another approach, and instead argue that the mechanism behind memorisation remains intact when decoupling the different constraints. Beyond the capacity threshold, this allows us to characterise properties of the optimal memorisation matrix (such as its asymptotic spectrum and the storage mechanism) and to generalise our results to two-layer linear models. Taken together, both works shed new light on the mechanisms driving the memorisation capacity of linear associative memories.

2 A decoupled variant of associative memory

In this section, we recall the original associative memory problem in its precise form, following Cabannes et al. [2024a], Nichani et al. [2025]. We then propose a decoupled variant of the problem and conjecture that the two problems share the same storage capacity.

Original Problem (OP) – Let $E \coloneqq \{e_\mu\}_{\mu\in[p]} \subseteq \mathbb{R}^d$ and $U \coloneqq \{u_\rho\}_{\rho\in[p]} \subseteq \mathbb{R}^d$ denote sets of input and output embeddings, each drawn independently from a standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$.[1] We are given p associations specified by an injective mapping $f^* : [p] \to [p]$, yielding the paired set $\{(e_\mu, u_{f^*(\mu)})\}_{\mu=1}^p$. The objective is to learn a matrix $W \in \mathbb{R}^{d\times d}$ such that, for every $\mu \in [p]$, the correct output vector achieves the highest score:

$$(\mathrm{OP}): \quad \forall \mu \in [p], \quad \arg\max_{\rho\in[p]} \; u_\rho^\top W e_\mu \;=\; f^*(\mu). \tag{1}$$

Without loss of generality, we assume throughout the remainder of the manuscript that the associations are ordered, i.e., $f^*(\mu) = \mu$ for all $\mu \in [p]$. We focus on two classes of weight matrices: general $W \in \mathbb{R}^{d\times d}$, and the rank-constrained case $W = QR^\top$ with $Q, R \in \mathbb{R}^{d\times m}$ and hidden dimension $m = \kappa d$ for $\kappa \in (0, 1]$, corresponding to a two-layer linear network [Nichani et al., 2025].

Decoupled Problem (DP) – The difficulty in analysing the original problem of eq. (1) directly is that the constraints are coupled across examples, in sharp contrast with classical perceptron-type models à la Gardner and Derrida [1988], where each input vector is mapped to a binary label that is independent of all other labels and inputs. In contrast, in factual recall the model must select the correct target for each input vector e_μ from the *same* set targets U. This observation motivates the study of a less correlated variant in which each input e_μ is associated with its own independent set of candidate outputs $U^{(\mu)}$. In this *decoupled* setting, which we formally define below, the competition set differs across inputs, substantially decoupling the constraints and simplifying greatly the analytical treatment, while preserving the essential structure of the associative memory task.

As in the original problem, let $E \coloneqq \{e_\mu\}_{\mu\in[p]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbb{I}_d)$ denote a set of input vectors. In contrast to the original problem, we introduce *for each* $\mu \in [p]$ a set of output vectors

$$U^{(\mu)} \coloneqq \{u_\rho^{(\mu)}\}_{\rho\in[p]} \subseteq \mathbb{R}^d,$$

where all vectors are drawn i.i.d. from $\mathcal{N}(0, \mathbb{I}_d)$ and are *independent* across different values of $\mu \in [p]$. This modification is the only change with respect to the original problem; the objective remains to

[1] This choice is without loss of generality for Gaussian embeddings with full-rank covariances Σ_e, Σ_u: writing $e_\mu = \Sigma_e^{1/2}\tilde{e}_\mu$ and $u_\rho = \Sigma_u^{1/2}\tilde{u}_\rho$, one has $u_\rho^\top W e_\mu = \tilde{u}_\rho^\top \tilde{W} \tilde{e}_\mu$ with $\tilde{W} \coloneqq \Sigma_u^{1/2} W \Sigma_e^{1/2}$, and $W \mapsto \tilde{W}$ is a rank-preserving bijection of $\mathbb{R}^{d\times d}$. The capacity threshold is thus the same as for isotropic embeddings.

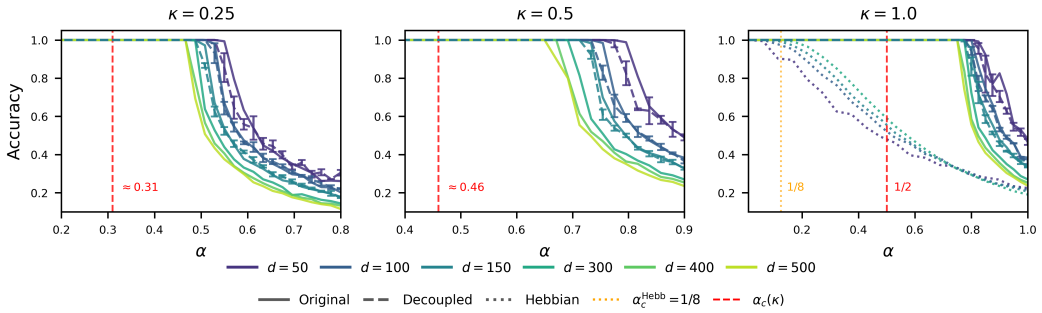


Figure 2: Empirical accuracy as a function of the load parameter $\alpha = p \log p / d^2$, for three rank fractions $\kappa \in \{0.25, 0.5, 1.0\}$. In each panel, curves correspond to the original problem (1) (solid) and the decoupled problem (2) (dashed), across several embedding dimensions d. In the full-rank case ($\kappa = 1$) the Hebbian ansatz (dotted) is also shown. Red dashed lines mark the asymptotic threshold $\alpha_c(\kappa)$ given by Theorem 1 ($\alpha_c \approx 0.31, 0.46, 0.50$). The transitions for the original and decoupled problems are indistinguishable across all κ, supporting the equivalence conjecture also in the rank-constrained case, and both substantially outperform the Hebbian ansatz at $\kappa = 1$. Models are trained with Adam on the cross-entropy loss with $W = QR^\top$, $Q, R \in \mathbb{R}^{d \times (\kappa d)}$; see Section A.

learn a matrix $W \in \mathbb{R}^{d \times d}$ such that, for every input index μ, the designated output $u_\mu^{(\mu)}$ achieves the largest score among its local candidate set:

$$(\mathrm{DP}): \quad \forall \mu \in [p], \quad \arg\max_{\rho \in [p]} \; u_\rho^{(\mu)\top} W e_\mu \;=\; f^*(\mu). \tag{2}$$

As in the original problem, we assume without loss of generality that $f^*(\mu) = \mu$ for all $\mu \in [p]$.

On a technical level, we expect the decoupled problem to fall within the scope of Gaussian universality: each constraint depends on W only through bulk statistics of the weight matrix, so the high-dimensional behaviour is governed primarily by its covariance structure, in line with prior work on universality in high-dimensional inference [Hu and Lu, 2023, Bandeira and Maillard, 2025, Goldt et al., 2022, Montanari and Saeed, 2022] (see Section C). In contrast, the shared outputs in the original problem induce strong correlations between constraints, which make a similar analysis of (OP) intractable with these tools (see Section B, where we also formulate the equivalence of the two problems as a statement on orthant probabilities).

Equivalence between original and decoupled problems – Our first main result is the following conjecture on the equivalence between the original and the decoupled problems. Here, we state the conjecture; in Section 3, we present different strands of evidence.

Conjecture 1 (Capacity Equivalence)**.** *Assume that $d, p \to \infty$ with a fixed ratio*

$$\alpha := \lim_{d \to \infty} \frac{p \log p}{d^2} \in (0, \infty), \tag{3}$$

and fix a rank fraction $\kappa \in (0, 1]$, the case $\kappa = 1$ corresponding to general matrices $W \in \mathbb{R}^{d \times d}$. Let $\mathcal{P} \in \{(\mathrm{OP}), (\mathrm{DP})\}$ denote either problem formulation. For each $\mathcal{P}$, there exists a critical capacity threshold $\alpha_c^{\mathcal{P}}(\kappa)$ such that, with high probability over the random draw of the input and output vectors, there exists a solution $W^\star$ of rank at most $m = \kappa d$ storing all associations if $\alpha < \alpha_c^{\mathcal{P}}(\kappa)$, whereas for $\alpha > \alpha_c^{\mathcal{P}}(\kappa)$ no such solution exists. Moreover we have:

$$\alpha_c^{\mathrm{OP}}(\kappa) = \alpha_c^{\mathrm{DP}}(\kappa).$$

We write $\alpha_c^{\mathcal{P}} := \alpha_c^{\mathcal{P}}(1)$ in the full-rank case. Conjecture 1 is supported by two sets of evidence. The equivalence between (DP) and (OP) is corroborated by numerical and analytical evidence presented below, for both full-rank and rank-constrained matrices. The threshold of the decoupled problem is then characterised sharply: the replica method yields $\alpha_c^{\mathrm{DP}} = 1/2$ in the full-rank case (Result 1) and an explicit formula for $\alpha_c^{\mathrm{DP}}(\kappa)$ (Claim 2), which we prove rigorously for every κ (Theorem 1).

Status of Conjecture 1 – For full-rank matrices, Conjecture 1 is now proven (Corollary 1): since the first version of this work appeared, Barnfield et al. [2026] have proved that $\alpha_c^{\mathrm{OP}} = 1/2$, and

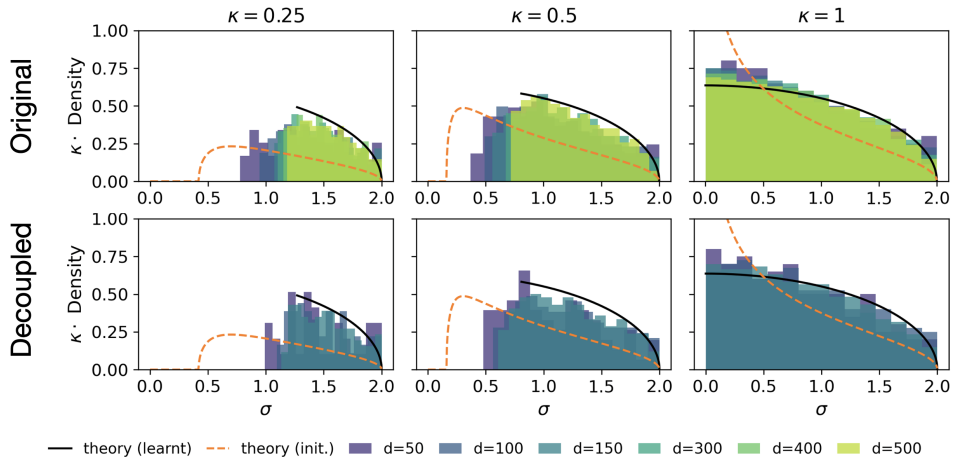


Figure 3: Distribution of the non-zero singular values at the capacity threshold in the (OP) (top) and (DP) (bottom) for various d and κ, normalised to have top singular value equal to 2. Each panel shows histograms from simulation data (shaded) alongside the theoretical prediction at capacity from Claim 2 (solid line) and the spectrum at initialisation (dashed line) for the two-layer parameterisation $W = QR^\top$ with i.i.d. Gaussian $Q, R \in \mathbb{R}^{d\times m}$, $m = \kappa d$. The strong similarity between the spectra of the optimal solutions in the two problems provides further evidence for their equivalence.

Theorem 1 gives $\alpha_c^{\mathrm{DP}} = 1/2$. For $\kappa < 1$, Theorem 1 determines $\alpha_c^{\mathrm{DP}}(\kappa)$ rigorously, so that Conjecture 1 predicts the same capacity for the original problem, as supported by our numerical experiments (Section 3.1). The same theorem shows that this capacity is achievable in the original problem as well; the matching impossibility bound remains open. Finally, our evidence indicates that the equivalence extends beyond the capacity threshold: the two problems also share the spectrum of the learnt weights and the storage mechanism (Sections 3.2 and 3.3).

3 Evidence for Conjecture 1

3.1 Evidence I: Retrieval accuracy and capacity threshold

We first validate numerically that the original problem (OP) and the decoupled problem (DP) share the same critical capacity. In Fig. 2 (right), we show the empirical retrieval accuracy as a function of the load parameter $\alpha = p \log p/d^2$ for a linear model trained using the cross-entropy loss, for different rank fractions κ and embedding dimensions d. For both (OP) and (DP), the accuracy undergoes a sharp transition from perfect memorisation (and an accuracy equal to 1) to an extensive number of violated constraints. The two transition points are indistinguishable across all dimensions tested, supporting the conjectured equivalence $\alpha_c^{\mathrm{OP}}(\kappa) = \alpha_c^{\mathrm{DP}}(\kappa)$ in terms of their critical capacity.

For the considered dimensions, the networks store more patterns than predicted by the high-dimensional theory, consistently with the finite-size corrections of size $\widetilde{\mathcal{O}}(1/\log p)$ predicted by our derivation (see Section C.2), which make the threshold converge slowly to its asymptotic value, as confirmed by the finite-size analysis of Fig. 7. The agreement between the two problems goes beyond the capacity threshold: already in Fig. 1, the full curves of the accuracy as a function of α match closely between (OP) and (DP), not just at the transition point. The cross-entropy loss curves, shown in Fig. 5, further confirm this agreement.

3.2 Evidence II: Structure of the learnt weights

Our second line of evidence concerns the structure of the learnt weights. Figure 3 shows the singular value distribution of the optimal weight matrices at the capacity threshold for different rank ratios κ and input dimensions d. The first observation is that the spectra show a clear departure from the spectra at initialisation, which we indicate with the orange dashed line. Second, we observe again a good match between the spectra of the original and the decoupled problem. Finally, we note that the

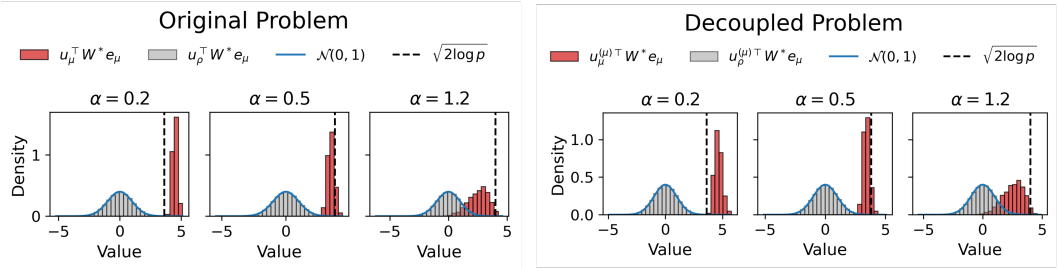


Figure 4: Histograms of the target scores (red) and non-target scores (grey) aggregated over indices $\mu \in [p]$, for (OP)((1);left) and (DP)((2);right), shown for three representative values of the load parameter α ($\alpha = 0.2, 0.5, 1.2$), with $d = 150$. After training, the learned weight matrix W^* is normalised so that the non-target scores $s_{\mu\rho} = u_\rho^\top W^* e_\mu$ have unit variance (the problem being invariant under rescaling of W, this fixes the scale); target scores are rescaled accordingly.

spectral density is well predicted by our theory; we plot the prediction of Claim 2 with the black solid line. Interestingly, the strong agreement with the theory suggests that the singular value distribution at capacity converges much faster than the capacity threshold itself.

3.3 Evidence III: Scores histograms

Our final line of evidence for the equivalence between (OP) and (DP) comes from an analysis of the scores $u_\rho^\top W e_\mu$ (and analogously $u_\rho^{(\mu)\top} W e_\mu$ in the decoupled problem), which also sheds light on the mechanism of pattern storage in the associative memory. Successful storage can in principle be achieved through multiple mechanisms: one may (1) increase the target scores where $\rho = \mu$ while leaving the non-target scores $\rho \neq \mu$ essentially unchanged, (2) decrease the non-target scores while keeping the target scores fixed, or (3) combine both effects.

We found that the associative memories opt for the strategy (1). In Fig. 4, we find that scores for non-target pairs follow a standard Gaussian distribution for various loads. In contrast, for the target pair $\rho = \mu$, learning drives the corresponding score towards the right tail of the Gaussian distribution of non-target scores, reaching values that must be larger than the maximum of the other $p-1$ Gaussian random variables, which is around $\sqrt{2\log p}$. This appears to be the case for both the original and decoupled problems. Note that Fig. 4 aggregates all target scores and all non-target scores across indices: a representative single-constraint histogram is shown in Fig. 6.

4 Mechanistic insights into memory storage and retrieval

The distribution of the scores gives additional insights into the storage mechanisms of optimal associative memories. For a Hebbian weight matrix $W_{\mathrm{Hebb}} := \sum_{z\in[p]} u_{f^*(z)} e_z^\top$, off-target scores are approximately normally distributed with mean zero and variance 1 (in our units), while target scores are normally distributed with mean $\sqrt{\alpha^{-1}\log p}$ and variance 1. This approximation suggests that the model fails to store all associations when the distributions overlap too much, in a sense that we make precise in Section E, and that allows to derive an heuristic for the capacity threshold of the Hebbian rule at $\alpha_c^{\mathrm{Hebb}} = 1/8$ (see Theorem 5), which matches well numerical observations and agrees with the precise analysis of Barnfield et al. [2026]. The Hebbian weight matrix thus has a significantly lower capacity than the optimal weight, cf. Fig. 2.

The optimal weight matrix achieves the higher capacity $\alpha_c = 1/2$ by adopting a different strategy close to the capacity threshold. Interestingly, as we detail in Section C.2, our theory predicts that below capacity, the diagonal scores rather concentrate (to leading order as $p \to \infty$) around a deterministic value that is just enough to allow for correct mapping of inputs to outputs. This seems to match the observations of Fig. 4, where we see the variance of the red histograms increase significantly as we cross capacity, while in the Hebbian matrix the variance of the diagonal scores is always large (Fig. 9). This effect occurs in both problems, hinting at a common mechanism behind optimal capacity.

5 Sharp capacity thresholds of the decoupled problem

Motivated by the evidence of Section 3, we now outline our computation of the memorisation capacity of the decoupled problem, with and without rank constraints; its decoupled structure is what makes this approach tractable (see Sections B and C).

5.1 Statistical physics approach

Following a classical statistical physics approach, we compute the *fractional volume of the solution space* for the decoupled problem, conditioned on fixed realisations of the input and output embeddings, (E, U). Recall that the embeddings $E = (e_\mu)_{\mu\in[p]}$ and $U = (u_\rho^{(\mu)})_{\mu,\rho\in[p]}$ are drawn independently in $\mathbb{R}^d$ from $\mathcal{N}(0, \mathbb{I}_d)$. We seek a weight matrix $W \in \mathbb{R}^{d\times d}$ with rank at most $m = \kappa d$ (for a constant $\kappa \in (0, 1]$) such that $W \in \Gamma(E, U)$, with

$$\begin{cases} \Gamma(E,U) & \coloneqq \left\{ W \in \mathbb{R}^{d\times d} \,:\, \forall \mu \in [p],\, \max_{\rho\in[p]} u_\rho^{(\mu)\top} W e_\mu \le u_\mu^{(\mu)\top} W e_\mu \right\}, \\ \mathcal{V}_{\mathrm{DP}}(E,U) & \coloneqq P_m\left(\Gamma(E,U)\right). \end{cases} \tag{4}$$

where P_m is a probability measure on W supported on the set of matrices of rank at most m, Γ is the (convex) solution space, and $\mathcal{V}_{\mathrm{DP}}(E, U)$ its volume under P_m. In practice, for $m = d$, we will take $P_m = \mathcal{N}(0, \mathbb{I}_{d^2}/d)$ the standard Gaussian distribution with variance $1/d$, while for $m = \kappa d$ with $\kappa \in (0, 1]$ we take P_m as the law of $W = QR^\top/\sqrt{dm}$, for $Q, R \in \mathbb{R}^{d\times m}$ independently sampled with i.i.d. $\mathcal{N}(0, 1)$ elements. While the choice of P_m will yield some analytic simplifications, the asymptotic capacity threshold is naturally agnostic to the choice of this prior.

Scaling insights – For a fixed W, by permutation invariance a given association $\mu \in [p]$ is satisfied with probability $1/p$, and so the probability that all the independent associations are satisfied is $(1/p)^p = e^{-p\log p}$. In particular, we have $\mathbb{E}[\mathcal{V}_{\mathrm{DP}}(E, U)] = e^{-p\log p}$. On the other hand, the parameter space has d^2 degree of freedom: a balancing with the entropic cost of the constraints suggests that the capacity phase transition occurs for a finite value of the load parameter

$$\alpha \coloneqq \frac{p\log p}{d^2}. \tag{5}$$

Notice that in the original problem, even the first moment of the volume of solutions is hard to characterise exactly, as we discuss in Section B. Beyond this heuristic argument, the scaling of eq. (5) is naturally predicted by our formal capacity analysis, which we now outline.

5.2 Full-rank linear associative memory

Our main results are sharp computations of the asymptotics of the (normalised) volume:

$$\varphi_d \coloneqq d^{-2}\mathbb{E}\log \mathcal{V}_{\mathrm{DP}}(E, U), \tag{6}$$

where the average is over $(e_\mu, u_\rho^{(\mu)})_{\rho,\mu} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbb{I}_d)$. φ_d is also called *free entropy* in statistical physics. Our first results concerns the case $\kappa = 1$ ($m = d$), i.e. without any rank constraint.

Result 1 (Storage capacity of a decoupled linear associative memory)**.** *Consider φ_d as defined in eq.* (6)*, for $P_m = \mathcal{N}(0, \mathbb{I}_{d^2}/d)$. In the limit $d, p \to \infty$ with $(p\log p)/d^2 \to \alpha > 0$:*

- *If $\alpha < 1/2$, then (up to additive constants independent of α)*

$$\lim_{d\to\infty} \varphi_d = \inf_{q\in[0,1)} \left\{ \frac{1}{2}\log(1-q) + \frac{q}{2(1-q)} + \alpha \lim_{p\to\infty} \frac{1}{\log p} \mathbb{E}_{\eta\sim\mathcal{N}(0,\mathbb{I}_p)}\left[\log f_p(q;\eta)\right] \right\}, \tag{7}$$

  where $f_p(q; \eta)$ is defined in eq. (37)*. In particular, $\varphi(\alpha) \coloneqq \lim_{d\to\infty} \varphi_d > -\infty$.*

- *As $\alpha \uparrow 1/2$, the minimiser $q^\star(\alpha)$ of eq.* (7) *satisfies $q^\star(\alpha) \uparrow 1$, and $\varphi(\alpha) \to -\infty$.*

In particular, Result 1 predicts the storage capacity threshold $\alpha_c^{\mathrm{DP}} = 1/2$, which we prove rigorously in Theorem 1. The derivation of Result 1 is detailed in Section C.1.

The replica method – Result 1 is based on an application of the *replica method* from statistical physics, a non-rigorous technique which has had a wide range of applications, notably in statistical

learning and the theory of neural networks [Mézard et al., 1987, Gardner and Derrida, 1988, Gabrié, 2020, Charbonneau et al., 2023, Maillard et al., 2024, Barbier et al., 2026]. The replica method is widely believed to yield exact asymptotic predictions, as was established rigorously in a large range of problems [Guerra and Toninelli, 2002, Talagrand, 2006, Panchenko, 2013, Barbier et al., 2019, Gerbelot et al., 2023, Vilucchio et al., 2025]. More precisely, Result 1 relies on a so-called *replica-symmetry* assumption, which is known to hold in a large class of convex constraint satisfaction problems [Talagrand, 2010, Barbier et al., 2022]. Since the full-rank problem is convex, we expect this assumption to be exact here, and we therefore present this prediction as a Result; by contrast, the rank-constrained prediction of Section 5.3 concerns a non-convex problem, for which replica symmetry rests on less firm grounds, and we state it as a Claim (Claim 2). As we show in Section 5.3, the capacity threshold it predicts is nevertheless exact (Theorem 1).

Finally, while we do not prove it here, the statistical physics analysis predicts that $d^{-2}\log \mathcal{V}_{\mathrm{DP}}(E,U)$ concentrates on its average φ_d in the asymptotic limit $d,p\to\infty$. This concentration of the log-volume has been established in related models [Talagrand, 2010], and Result 1 should thus be understood as characterising the *typical* size of the solution space.

The parameter q – One can associate to the volume of eq. (4) a Gibbs-type measure $\langle\cdot\rangle$, supported on $\Gamma(E,U)$, and with density proportional to $P_m(W)$. Geometrically, the value $q^\star(\alpha)\in[0,1)$ achieving the infimum in eq. (7) is such that, in the high-dimensional limit, $(W\cdot W')/(\|W\|_F\|W'\|_F)\to q^\star(\alpha)$ for $W,W'\sim\langle\cdot\rangle$ two samples under this measure. The set of solutions $\Gamma(E,U)$ is a convex cone, and $(1-q)$ measures thus the "angular width" of this cone: as the load α increases towards the capacity threshold, q increases as well, marking the shrinking of the set of solutions. q approaches 1 in the limit $\alpha\uparrow 1/2$, marking the critical threshold beyond which no solutions exist.

Finite-size effects, and slow convergence to the asymptotic limit – Interestingly, the derivation detailed in Section C predicts that eq. (7) holds, up to $o(1)$ error terms, for the values of (d,p) corresponding to the embedding dimension and number of patterns, that is:

$$\varphi_d=\frac{1}{2}\log(1-q^\star)+\frac{q^\star}{2(1-q^\star)}+\frac{\alpha}{\log p}\mathbb{E}_\eta\big[\log f_p(q^\star;\eta)\big]+o(1). \tag{8}$$

Our derivation (cf Section C.2) suggests that the last term in the right-hand-side of this equation converges to its asymptotic limit very slowly, with typical corrections of size $\widetilde{\mathcal{O}}(1/\log p)$ (up to multiplicative $\log\log p$ terms): in turn, this creates strong finite-size corrections to the capacity threshold with respect to its asymptotic limit $\alpha_c=1/2$. This mechanism explains the very slow convergence observed in Fig. 1: in Fig. 7, we show that these observations are compatible with a $\widetilde{\mathcal{O}}(1/\log p)$ convergence rate of the capacity threshold. This slow convergence is a generic feature of the extreme-value statistics underlying the problem: each constraint requires the target score to exceed the maximum of $p-1$ Gaussian competitors, whose typical value approaches its asymptotic behaviour $\sqrt{2\log p}$ with relative corrections of order $\widetilde{\mathcal{O}}(1/\log p)$.

5.3 Two-layer linear MLPs

We finally turn to characterising the optimal storage capacity in the setting where the matrix is constrained to have rank $m=\kappa d$ with $\kappa\in(0,1]$. Notably, this setting encompasses the full-rank case studied above when $\kappa=1$, so that Claim 2 and Theorem 1 below provide a unified treatment of both cases. Recall that φ_d was defined in eq. (6), and that we take P_m as the law of $W=QR^\top/\sqrt{dm}$ with $Q,R\in\mathbb{R}^{d\times m}$ with i.i.d. elements sampled from $\mathcal{N}(0,1)$. Equivalently, we consider a two-layer linear network with hidden dimension m as our input-to-output map in embedding space.

Claim 2 (Replica-symmetric prediction of the 2-layer linear MLP capacity)**.** *Let $m=\kappa d$ with $\kappa\in(0,1]$ denote the rank of the weight matrix, and recall the limit $d,p\to\infty$ with $(p\log p)/d^2\to\alpha>0$. Within the so-called replica-symmetric assumption discussed below:*

- *There exists $\alpha_c(\kappa)>0$ such that for $\alpha<\alpha_c(\kappa)$, $\varphi(\alpha)=\lim_{d\to\infty}\varphi_d(\alpha)>-\infty$ is given by a formula similar to eq.* (7)*, while $\varphi(\alpha)\to-\infty$ as $\alpha\uparrow\alpha_c(\kappa)$. Moreover, we have*

$$\alpha_c(\kappa)=\frac{1}{2}\int_{X_{\mathrm{q.c.}}(1-\kappa)}^{2}\rho_{\mathrm{q.c.}}(\sigma)\,\sigma^2\,d\sigma, \tag{9}$$

where $\rho_{\mathrm{q.c.}}(\sigma)\coloneqq(1/\pi)\sqrt{4-\sigma^2}$ is the quarter-circle law, and $X_{\mathrm{q.c.}}:[0,1]\to[0,2]$ denotes its quantile function, i.e. $X_{\mathrm{q.c.}}(p)=\inf\{\sigma:\int_0^\sigma\rho_{\mathrm{q.c.}}(\sigma')\,\mathrm{d}\sigma'\geq p\}$.

- *Recall that we define $\langle\cdot\rangle$ as the measure supported on $\Gamma(E,U)$, and with density proportional to $P_m(W)$. For $\alpha < \alpha_c(\kappa)$, let $W \sim \langle\cdot\rangle$, and denote $\widetilde{W} \coloneqq 2W/\|W\|_{\rm op}$, where $\|W\|_{\rm op}$ is the top singular value of W. Then the empirical singular value distribution of $\widetilde{W}$ converges as $d \to \infty$ to a well-defined distribution $\rho[\kappa,\alpha]$, and further for $\alpha \uparrow \alpha_c(\kappa)$, $\rho[\kappa,\alpha] \to \rho_c[\kappa]$, given by*

$$\rho_c[\kappa](\sigma) \coloneqq (1-\kappa)\,\delta(\sigma) + \kappa\,\rho_{\rm q.c.}(\sigma)\,\mathbf{1}_{[X_{\rm q.c.}(1-\kappa),\,2]}(\sigma). \tag{10}$$

Notice that $\alpha_c(1) = 1/2$, in accordance with Result 1; the variational formula for the asymptotic volume is given in eq. (70). Beyond the capacity, our derivation characterises the asymptotic singular value distribution of the optimal solution close to the capacity threshold, in excellent agreement with numerical experiments (black curves in Fig. 3).

Rigorous capacity threshold – Beyond the replica-symmetric predictions of Claim 2, we also provide a formal proof that $\alpha_c(\kappa)$ is the capacity threshold of the decoupled problem for every $\kappa \in (0,1]$, including the non-convex case $\kappa < 1$, and that it is achievable in the original problem.

Theorem 1 (Rank-constrained capacity threshold)**.** *Let $d,p,m \to \infty$ with $m/d \to \kappa \in (0,1]$ and $(p\log p)/d^2 \to \alpha > 0$, and let $\alpha_c(\kappa)$ be given by eq.* (9)*. With probability tending to one, if $\alpha < \alpha_c(\kappa)$ both the decoupled problem* (2) *and the original problem* (1) *admit a solution W of rank at most m, while if $\alpha > \alpha_c(\kappa)$ the decoupled problem admits none.*

We prove Theorem 1 in Section D, building on the approach of Barnfield et al. [2026] for the full-rank original problem. For $\kappa = 1$, it gives $\alpha_c^{\rm DP} = 1/2$; together with their result for the original problem, this settles the full-rank case of Conjecture 1.

Corollary 1 (Full-rank capacity equivalence)**.** *For $\kappa = 1$, $\alpha_c^{\rm OP} = \alpha_c^{\rm DP} = 1/2$.*

Remark 1 (Rank-constrained weights in the original problem)*.* For $\kappa < 1$, the impossibility bound for (1), and thus Conjecture 1, remains open. The dual-certificate argument of Barnfield et al. [2026] handles the coupled constraints but, being convex, is blind to the rank constraint; ours handles the rank constraint but relies on the decoupling.

Convexity and replica symmetry – It is important to notice that for $\kappa < 1$ the set of solutions $\Gamma_m(E,U) \coloneqq \Gamma(E,U) \cap \{W \in \mathbb{R}^{d\times d} : \mathrm{rk}(W) \le m\}$ is non-convex, due to the rank constraint. In non-convex problems the replica-symmetric assumption generically yields only an upper bound on the capacity threshold (see e.g. Maillard et al. [2025]); here, Theorem 1 shows that it is exact, while the spectral prediction of Claim 2 remains derived *within the replica-symmetric approximation*. Non-convexity also means that gradient-based optimisation is not guaranteed to reach optimal capacity; we nevertheless find excellent agreement between theory and experiments (see e.g. Fig. 3).

6 Outlook and open questions

A natural open direction stemming from our work is a rigorous proof of Conjecture 1 for rank-constrained matrices. Since achievability holds for the original problem (Theorem 1), this requires an impossibility argument that handles both the coupling of the constraints and the non-convexity of the rank constraint. More broadly, we believe that the equivalence between coupled and decoupled formulations may extend well beyond the linear associative memory studied here, and it would be interesting to understand in which problems it holds and where it breaks down. Already within our model, our derivation indicates that it should hold for extensive-rank matrices ($m = \kappa d$), for which each constraint depends on the weights only through their bulk statistics (see Section C.1.1), while it clearly fails for finite-rank ones. Establishing this equivalence rigorously might require new probabilistic tools to disentangle the strong correlations between the constraints. Another intriguing question arises from the validity of the replica-symmetric assumption in the non-convex rank-constrained problem, which Theorem 1 establishes at the level of the capacity threshold: interestingly, the validity of the replica-symmetric assumption and the ability of gradient descent to reach the global optimum of a non-convex loss have also been observed in neural networks with similar rank constraints [Maillard et al., 2024, Erba et al., 2026, Barbier et al., 2026, Martin et al., 2026], possibly for a common reason. Finally, extending our analysis to non-linear two-layer models or to the more realistic embedding distributions of Kim et al. [2026] is an exciting prospect.

Acknowledgments and Disclosure of Funding

AM is particularly grateful to Lenka Zdeborová for introducing him to this problem as well as for stimulating discussions. AG wishes to thank Antoine Maillard and the entire Inria ARGO team for their hospitality during the visit in which this project was carried out. AG gratefully acknowledges funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Missione 4, Componente 1, Investimento 4.1 "Estensione del numero di dottorati di ricerca e dottorati innovativi per la Pubblica Amministrazione e il patrimonio culturale" (CUP G93C23000620003), and also gratefully acknowledges financial support from Fondazione Zegna. SG gratefully acknowledges funding from the European Research Council (ERC) for the project "beyond2", ID 101166056; from the European Union–NextGenerationEU, in the framework of the PRIN Project SELF-MADE (code 2022E3WYTY – CUP G53D23000780001), and from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 – Project FAIR "Future Artificial Intelligence Research" (CUP G53C22000440006).

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 14937–14946, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/26d3c9a66836ded8f34a944f2bfe868e-Paper-Conference.pdf.

Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

James A. Anderson. A simple neural network generating an interactive memory. *Mathematical Biosciences*, 14(3):197–220, 1972. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(72)90075-2. URL https://www.sciencedirect.com/science/article/pii/0025556472900752.

Keith Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31 of *Mathematical Sciences Research Institute Publications*, pages 1–58. Cambridge University Press, 1997.

Afonso S Bandeira and Antoine Maillard. Exact threshold for approximate ellipsoid fitting of random points. *Electronic Journal of Probability*, 30:1–46, 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.

Jean Barbier, Dmitry Panchenko, and Manuel Sáenz. Strong replica symmetry for high-dimensional disordered log-concave gibbs measures. *Information and Inference: A Journal of the IMA*, 11(3): 1079–1108, 2022.

Jean Barbier, Francesco Camilli, Minh-Toan Nguyen, Mauro Pastore, and Rudy Skerk. Statistical physics of deep learning: Optimal learning of a multi-layer perceptron near interpolation. *Phys. Rev. X*, 2026. doi: 10.1103/56sb-pdh6. URL https://link.aps.org/doi/10.1103/56sb-pdh6.

Nicholas Barnfield, Juno Kim, Eshaan Nichani, Jason D. Lee, and Yue M. Lu. Sharp capacity thresholds in linear associative memory: From top-1 retrieval to tail-average learning, 2026. URL https://arxiv.org/abs/2605.05189.

Joel Bun, Jean-Philippe Bouchaud, Satya N Majumdar, and Marc Potters. Instanton approach to large n harish-chandra-itzykson-zuber integrals. *Physical Review Letters*, 113:070201, 2014.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 9718–9747, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/28f734334f9f325b515933e70b6b412c-Paper-Conference.pdf.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5114–5134. PMLR, 2024b. URL https://proceedings.mlr.press/v235/cabannes24a.html.

Patrick Charbonneau, Enzo Marinari, Giorgio Parisi, Federico Ricci-tersenghi, Gabriele Sicuro, Francesco Zamponi, and Marc Mezard. *Spin Glass Theory and Far Beyond: Replica Symmetry Breaking after 40 Years*. World Scientific, 2023.

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, May 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1806-y. URL http://dx.doi.org/10.1007/s10955-017-1806-y.

A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

Vittorio Erba, Emanuele Troiani, Luca Biggio, Antoine Maillard, and Lenka Zdeborová. Bilinear sequence regression: A model for learning from long sequences of high-dimensional tokens. *Physical Review X*, 15(2):021092, 2025.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborova, and Florent Krzakala. The nuclear route: Sharp asymptotics of ERM in overparameterized quadratic networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=MNqBFIGT8l.

Silvio Franz, Giorgio Parisi, Maxime Sevelev, Pierfrancesco Urbani, and Francesco Zamponi. Universality of the sat-unsat (jamming) threshold in non-convex continuous constraint satisfaction problems. *SciPost Physics*, 2(3):019, 2017.

Marylou Gabrié. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53(22):223002, 2020.

E Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and Theoretical*, 21(1):271–284, 1988. doi: 10.1088/0305-4470/21/1/031. URL https://hal.science/hal-03285587.

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of perceptrons with random labels. *Physical Review E*, 109(3):034305, 2024.

Cedric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove kabashima's replica formula). *IEEE Transactions on Information Theory*, 69(3):1824–1852, 2023. doi: 10.1109/TIT.2022.3222913.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. PMLR, 2022. URL https://proceedings.mlr.press/v145/goldt22a.html.

Francesco Guerra and Fabio Lucio Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230(1):71–79, 2002.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023. doi: 10.1109/TIT.2022.3217698.

Brice Huang. Capacity threshold for the ising perceptron. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1126–1136. IEEE, 2024.

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*, 21(04):353–359, April 1972. ISSN 1557-9956. doi: 10.1109/TC.1972.5008975. URL https://doi.ieeecomputersociety.org/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.

Werner Krauth and Marc Mézard. Storage capacity of memory networks with binary couplings. *Journal de Physique*, 50(20):3057–3066, 1989.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=X4y_10OX-hX.

Giacomo Livan, Marcel Novaes, and Pierpaolo Vivo. Introduction to random matrices theory and practice. *Monograph Award*, 63(54):914, 2018.

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

Antoine Maillard and Dmitriy Kunisky. Fitting an ellipsoid to random points: predictions using the replica method. *IEEE Transactions on Information Theory*, 2024.

Antoine Maillard, Emanuele Troiani, Simon Martin, Florent Krzakala, and Lenka Zdeborová. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. *Advances in Neural Information Processing Systems*, 37:82085–82132, 2024.

Antoine Maillard, Afonso S Bandeira, David Belius, Ivan Dokmanić, and Shuta Nakajima. Injectivity of relu networks: perspectives from statistical physics. *Applied and Computational Harmonic Analysis*, 76:101736, 2025.

Simon Martin, Giulio Biroli, and Francis Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *arXiv preprint arXiv:2601.10483*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

Andrea Montanari and Subhabrata Sen. A friendly tutorial on mean-field spin glass techniques for non-physicists. *Foundations and Trends in Machine Learning*, 17(1):1–173, 2024.

Andrea Montanari, Yiqiao Zhong, and Kangjie Zhou. Tractability from overparametrization: The example of the negative perceptron. *Probability Theory and Related Fields*, 188(3):805–910, 2024.

Eshaan Nichani, Jason Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 4166–4207, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0bf9f909c24c8879d1b7f86fa50a9e49-Paper-Conference.pdf.

Dmitry Panchenko. The parisi ultrametricity conjecture. *Annals of Mathematics*, pages 383–393, 2013.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: for Physicists, Engineers and Data Scientists*. Cambridge University Press, 2020.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437/.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 01 2013. doi: 10.1214/ECP.v18-2865.

Michel Talagrand. The parisi formula. *Annals of mathematics*, pages 221–263, 2006.

Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

Nuri Mert Vural, Alberto Bietti, Mahdi Soltanolkotabi, and Denny Wu. Learning to recall with transformers beyond orthogonal embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CfFj68C9Cn.

Garrett G Wen, Hong Hu, Yue M Lu, Zhou Fan, and Theodor Misiakiewicz. When does gaussian equivalence fail and how to fix it: Non-universal behavior of random features with quadratic scaling. *arXiv preprint arXiv:2512.03325*, 2025.

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. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1410.3916.

Yizhou Xu, Antoine Maillard, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. In *38th Annual Conference on Learning Theory*, 2025.

Contents of the appendix

A Numerical Experiments **16**

A.1 Experiments in the Main Text 17

A.2 Additional Experiments 17

B Difficulties in the analysis of the original problem (OP) **19**

C Statistical physics analysis of the decoupled problem (DP) **20**

C.1 Derivation of Result 1 20

C.2 Expansion of the energetic term 25

C.3 Rank-constrained weights: derivation of Claim 2 28

D Proof of the capacity threshold of the rank-constrained decoupled problem **33**

D.1 Setting and main result 34

D.2 Truncated-SVD preliminaries 35

D.3 Achievability for both problems 41

D.4 Sharp rank-constrained impossibility 44

D.5 Technical lemmas 49

E Heuristic derivation of the capacity of the Hebbian ansatz **54**

A Numerical Experiments

Common training strategy. All experiments reported in this manuscript follow the same training protocol. Inputs and outputs (in both the original and decoupled problems) are drawn independently from a d-dimensional Gaussian distribution with zero mean and identity covariance. The goal is to satisfy the association constraints in Eqs. (1) and (2). To this end, we train a model $F_W(e)$ with the cross-entropy loss on both variants of the memory task.

Original problem. The cross-entropy loss with a shared set of outputs reads

$$\mathcal{L}_p^{(\mathrm{OP})}(W) = -\frac{1}{p}\sum_{\mu=1}^{p} \log \frac{\exp\big(u_\mu^\top F_W(e_\mu)\big)}{\sum_{\rho=1}^{p}\exp\big(u_\rho^\top F_W(e_\mu)\big)}. \tag{11}$$

Decoupled problem. The variant with replicated sets of outputs is trained on

$$\mathcal{L}_p^{(\mathrm{DP})}(W) = -\frac{1}{p}\sum_{\mu=1}^{p} \log \frac{\exp\Big((u_\mu^{(\mu)})^\top F_W(e_\mu)\Big)}{\sum_{\rho=1}^{p}\exp\Big((u_\rho^{(\mu)})^\top F_W(e_\mu)\Big)}. \tag{12}$$

Both objectives are optimised with Adam at learning rate $\eta = 10^{-2}$ for at most 512 full-batch steps, using linear warmup followed by cosine decay; training stops once 99.9% accuracy is reached. In the full-rank case, we use a linear model $F_W(x) = Wx$ with $W \in \mathbb{R}^{d\times d}$ initialised with i.i.d. Gaussian entries of standard deviation d^{-1}. In the rank-constrained case, we train a two-layer linear network $F_W(e) = QR^\top e$ with $Q, R \in \mathbb{R}^{d\times m}$ and $m \le d$, both initialised at the same scale.

All experiments were run on a single NVIDIA A100-PCIE-40GB GPU and each completes within a few hours. Given the small scale, reproduction is feasible on less powerful hardware, including standard CPUs.

A.1 Experiments in the Main Text

Figure 1(Right) is obtained by following the common training protocol described above for 25 values of the load parameter $\alpha = p \log p / d^2$, evenly spaced between 0.4 and 1.0. For the original problem we report results for $d = 50, 100, 150, 300, 400, 500$. For the decoupled problem, simulations were limited to $d \leq 150$ due to memory constraints: in this setting, each input is associated with its own set of p competing outputs, each in $\mathbb{R}^d$, requiring storage of order $O(dp^2)$. Since $p \log p$ scales as d^2 at fixed load, this quickly becomes prohibitive at larger dimensions. In contrast, for the original problem only a single shared set of p outputs is stored, resulting in a memory footprint of order $O(dp)$. Each experiment is repeated 5 times with independent random draws of the inputs and outputs. The lines report the empirical mean across repetitions and the standard deviation is reported only for the decoupled problem to avoid cluttering the figure.

Figure 2 The same experiment of Fig. 1 is reported also for $\kappa \in \{0.25, 0.5\}$. We additionally provide the accuracy lines for the Hebbian ansatz in the case $\kappa = 1$. The asymptotic thresholds are the ones of Theorem 1.

Figure 3 reports the distribution of the non-zero singular values of the experiments of Fig. 2. Since the task is invariant under renormalisation of the weight matrix, after training we renormalise it to have $\|W\|_{\text{op}} = 2$.

Figure 4 is obtained by training the full-rank linear models according to the common training protocol with $d = 150$ and for three values of the load parameter α (0.2, 0.5, and 1.2), corresponding respectively to regimes below, near, and above the capacity threshold. The choice $d = 150$ reflects the largest dimension that can be simulated for the decoupled problem; for consistency, the same dimension is used for the original problem.

After training, we extract the learned weight matrix W^* and normalise it so that the non-target scores $s_{\mu\rho} = u_\rho^\top W^* e_\mu$ (and equivalently for the decoupled problem) have unit variance (the problem being invariant under re-normalisation of W, this fixes the scale); the target scores are rescaled accordingly. Target scores are shown in red and non-target scores in grey.

Note that overlap between the aggregated target and non-target distributions does not necessarily imply classification errors, since correctness depends on comparing each target score only with its own competing non-target scores rather than with the full set of scores. A representative single-instance, single-constraint histogram is shown in Fig. 6.

A.2 Additional Experiments

Figure 5 reports the cross-entropy loss for the experiments in Fig. 2, and the empirical standard deviation is reported for the decoupled problem lines.

Figure 6. This figure shows the single-constraint analogue of Fig. 4.

Figure 7 reports the finite-size scaling of the empirical capacity threshold, for $d \in \{24, 32, 45, 64, 90, 128, 180, 256, 362, 512\}$ in the original problem and $d \leq 128$ in the decoupled one. For each d, the critical load $\alpha_c = p_c \log p_c / d^2$ is obtained by bisection on p, where p_c is the largest number of patterns retrieved without errors after training (full-batch Adam, learning rate 10^{-2}, at most 2^{15} epochs). Results are averaged over 3 to 10 realisations of the patterns, with more realisations at smaller d.

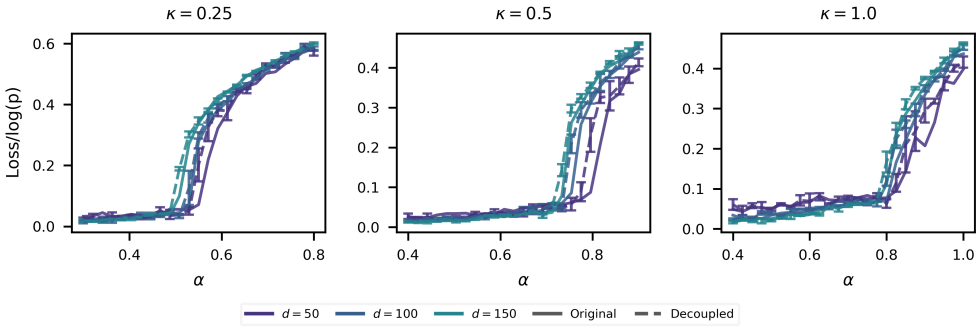


Figure 5: Cross-entropy loss for the experiment in Fig. 2. Empirical standard deviation is reported for the decoupled problem lines.

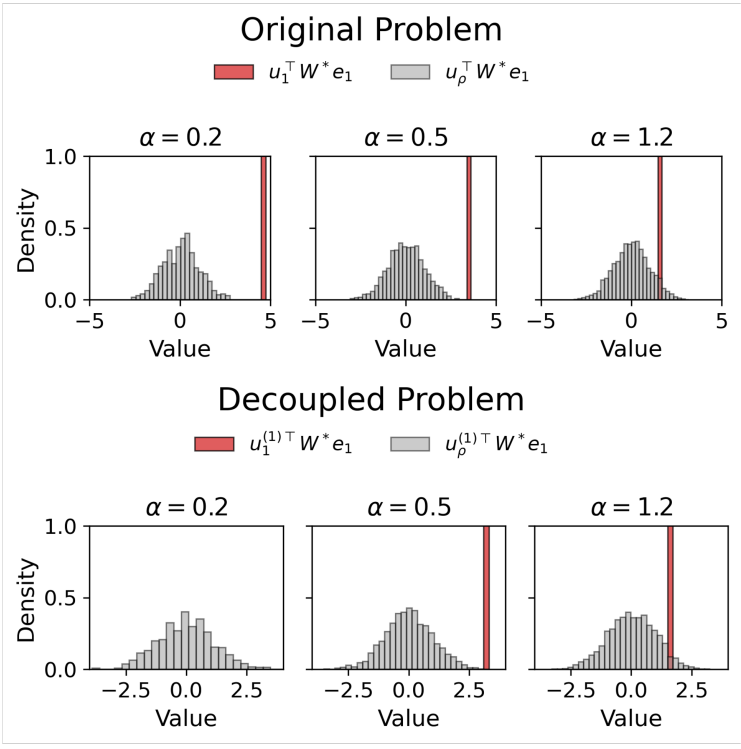


Figure 6: Single-constraint counterpart of Fig. 4. The constraint is violated whenever the target score (red) fails to exceed the maximum of its competing non-target scores (grey).

B Difficulties in the analysis of the original problem (OP)

Let us consider the case $m = d$ ($\kappa = 1$) for simplicity. Like eq. (4), we can associate a volume of solutions to the original problem (OP) defined in eq. (1):

$$\begin{aligned}\mathcal{V}_{\rm OP}(E,U) &\coloneqq \int P_d(\mathrm{d}W)\prod_{\mu=1}^{p}\mathbb{1}\left(\arg\max_{\rho\in[p]} u_\rho^\top W e_\mu = \mu\right), \\ &= \int P_d(\mathrm{d}W)\prod_{\mu=1}^{p}\prod_{\rho(\neq\mu)}\Theta\big[(u_\mu-u_\rho)^\top W e_\mu\big],\end{aligned} \tag{13}$$

where $\Theta(x) \coloneqq \mathbb{1}\{x \geq 0\}$ is the Heaviside step function, and $P_d = \mathcal{N}(0, \mathbb{I}_{d^2}/d)$. Already, one can see that computing even the first moment of eq. (13) (with respect to the realisation of (E, U)) leads to important difficulties. Recall that the patterns $U = \{u_\mu\}_{\mu=1}^p$ and $\{e_\mu\}_{\mu=1}^p$ are drawn independently from the d-dimensional standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$. We thus have

$$\mathbb{E}\mathcal{V}_{\rm OP}(E,U) = \int P_d(\mathrm{d}W)\ \underbrace{\mathbb{E}_{E,U}\left[\prod_{\mu=1}^{p}\prod_{\rho(\neq\mu)}\Theta\left((u_\mu-u_\rho)^\top W e_\mu\right)\right]}_{=:J(W)}. \tag{14}$$

While in the decoupled problem a simple permutation invariance argument yields $J(W) = e^{-p\log p}$ (see the main text), in the original problem such a simple argument does not apply. One can notice that conditionally on W and U, the variables

$$z_\rho^\mu \coloneqq u_\rho^\top W e_\mu$$

are jointly Gaussian with zero mean and covariance

$$\mathbb{E}_{E|W,U}[z_\rho^\mu z_\sigma^\nu] = \delta_{\mu\nu} u_\rho^\top W W^\top u_\sigma =: (\Sigma^U)_{\rho\sigma}\delta_{\mu\nu}.$$

Thus

$$J(W) = \mathbb{E}_U \prod_{\mu=1}^{p}\left[\mathbb{E}_{z^{(\mu)}|W,U}\prod_{\rho\neq\mu}\Theta\left(z_\mu^\mu - z_\rho^\mu\right)\right], \tag{15}$$

where $z^{(\mu)} \sim \mathcal{N}(0, \Sigma^U)$. Eq. (15) is the probability that a multivariate Gaussian vector with non-trivial covariance falls within a given orthant. This does not admit a closed-form expressions for generic covariance matrices, which prevents us from a simple computation even for the first moment of the volume of solutions in the original problem.

The equivalence as a statement on orthant probabilities – We can pinpoint exactly where, at least formally, a difference between the two problems arises. For n replicas $W^1, \dots, W^n$, define the scores $z_{\mu\rho}^a \coloneqq d^{-1/2}\, u_\rho^\top W^a e_\mu$ and the event

$$\mathcal{E}_{\rm OP}(\{W^a\}) \coloneqq \bigcap_{a=1}^{n}\bigcap_{\mu=1}^{p}\bigcap_{\rho\neq\mu}\{z_{\mu\mu}^a > z_{\mu\rho}^a\}.$$

Define $\mathcal{E}_{\rm DP}(\{W^a\})$ analogously, using an independent output set $\{u_\rho^{(\mu)}\}_{\rho\in[p]}$ for each input μ. The replica computations for (OP) and (DP) coincide at leading order, and in particular give the same capacity threshold, if, uniformly over the replica configurations contributing to the saddle point,

$$\frac{1}{d^2}\log\mathbb{P}\big(\mathcal{E}_{\rm OP}(\{W^a\})\big) = \frac{1}{d^2}\log\mathbb{P}\big(\mathcal{E}_{\rm DP}(\{W^a\})\big) + o(1). \tag{16}$$

For the decoupled scores, this probability factorises over $\mu \in [p]$ and reduces to the Gaussian computation of Section C (see the discussion of Gaussian equivalence in Section C.1.1). For the original problem, conditionally on U the scores are centred Gaussian with covariance

$$\mathbb{E}_{E|U}\left[z_{\mu\rho}^a\, z_{\nu\sigma}^b\right] = \delta_{\mu\nu}\,\frac{1}{d}\,u_\rho^\top W^a W^{b\top} u_\sigma,$$

so that $\mathbb{P}(\mathcal{E}_{\rm OP})$ is an orthant probability of a Gaussian vector whose covariance still depends on U (for $n=1$, this is eq. (15)). If $\|W^a\|_{\rm op} = \Theta(1)$, each entry of this covariance concentrates on its average over U, namely $Q_{ab}\,\delta_{\rho\sigma}$ with Q_{ab} defined in eq. (21), with fluctuations of order $d^{-1/2}$. Replacing the covariance by its average yields exactly eq. (16). Given the evidence of Section 3, we expect that only this average contributes to the leading order of the orthant probability; assuming eq. (16) is then enough to extend our derivation, and hence Result 1 and Claim 2, directly to the original problem. Note that entrywise concentration does not by itself imply eq. (16), as the covariance matrix, of rank at most $d \ll p$, remains far from its average in operator norm. Unfortunately, orthant probabilities of Gaussian vectors do not admit closed-form expressions in high dimension, and we have not been able to establish eq. (16) in this way.

C Statistical physics analysis of the decoupled problem (DP)

In this section we compute the asymptotic limit of φ_d (called "free entropy" in the statistical physics language), as defined in eq. (6), using the heuristic replica method. We first detail the computation in the case $\kappa = 1$ ($m = d$) for a Gaussian prior $P_d = \mathcal{N}(0, \mathbb{I}_{d^2}/d)$ in Section C.1, and then extend these results to arbitrary $\kappa \leq 1$ ($m \leq d$), i.e. the two-layer linear model, in Section C.3).

C.1 Derivation of Result 1

Here we present the full derivation supporting Result 1 on the optimal storage capacity. Throughout this section, we take $P_d = \mathcal{N}(0, \mathbb{I}_{d^2}/d)$. As detailed in the main text, we leverage the replica method from statistical physics [Mézard et al., 1987, Charbonneau et al., 2023]. It starts from the so-called "replica trick":

$$\mathbb{E}\log \mathcal{V}_{\rm DP} = \left[\frac{\partial}{\partial n}\log \mathbb{E}[\mathcal{V}_{\rm DP}^n]\right]_{n=0}. \tag{17}$$

In the replica method, one proceeds to compute the moments $\mathbb{E}[\mathcal{V}_{\rm DP}^n]$ for *integer* $n \geq 0$, and then analytically expands the result to all $n > 0$, allowing to take the limit $n \to 0$ in eq. (17). The replica method is a widely-discussed topic in the statistical physics of learning: we thus refrain from giving a full introduction to its mechanisms, and refer instead the reader to Engel and Van den Broeck [2001] for foundational applications, and to Montanari and Sen [2024], Maillard et al. [2025] for mathematically-friendly introductions. We emphasize that the replica method is inherently non-rigorous, and therefore the following computations are performed at a level of rigour which is standard in the literature on the statistical physics of learning.

C.1.1 Computation of the moments and Gaussian equivalence

Following the replica method, we start from the computations of the moments of $\mathcal{V}_{\rm DP}(E,U)$. By independence of the constraints, we have the crucial decoupling:

$$\begin{aligned}
&\mathbb{E}\mathcal{V}_{\rm DP}^n(E,U)\\
&= \int \prod_{a=1}^n P_d(\mathrm{d}W^a) \prod_{\mu=1}^p \left[\mathbb{E}_{U^{(\mu)},e_\mu} \prod_{a=1}^n \prod_{\rho(\neq\mu)} \Theta\left[\frac{1}{\sqrt{d}}(u_\mu^{(\mu)} - u_\rho^{(\mu)})^\top W^a e_\mu\right]\right],\\
&= \int \prod_{a=1}^n P_d(\mathrm{d}W^a) \left[\underbrace{\mathbb{E}_{U,e} \prod_{a=1}^n \prod_{\rho=2}^p \Theta\left[\frac{1}{\sqrt{d}}(u_1 - u_\rho)^\top W^a e\right]}_{=:J(\{W^a\})}\right]^p.
\end{aligned} \tag{18}$$

Gaussian universality – Let us introduce

$$z_\rho^a \coloneqq \frac{u_\rho^\top W^a e}{\sqrt{d}}. \tag{19}$$

Conditionally on e and $(W^a)_{a=1}^n$, z_ρ^a are Gaussian random variables, with zero mean and covariance

$$\mathbb{E}_{U|\{W^a\}_{a\in[n]},e}[z_\rho^a z_\sigma^b] = \delta_{\rho\sigma}\frac{1}{d}e^\top W^a W^{b\top} e =: \delta_{\rho\sigma} Q_{ab}^e,$$

With respect to the randomness of e, we expect that for generic matrices[2] $(W^a)_{a=1}^n$ the variables $(Q^e_{ab})_{1\leq a,b\leq n}$ concentrate as $d\to\infty$ around their expectation $Q_{ab} \coloneqq (1/d)\mathrm{Tr}[W^a W^b]$. Under this assumption, we thus reach that to leading order

$$\Phi_d(n) \coloneqq \frac{1}{d^2}\log \mathbb{E}[\mathcal{V}_{\mathrm{DP}}(E,U)^n] = \frac{1}{d^2}\log \int \prod_{a=1}^n P_d(\mathrm{d}W^a)\, F_{n,p}(Q)^p + o(1), \tag{20}$$

with

$$Q_{ab} \coloneqq \frac{1}{d}\mathrm{Tr}[W^a W^{b\top}], \tag{21}$$

and where we introduced for a generic symmetric positive semi-definite $Q\in\mathbb{R}^{n\times n}$:

$$F_{n,p}(Q) \coloneqq \mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_p\otimes Q)} \prod_{a=1}^n \prod_{\rho=2}^p \Theta\left[z_1^a - z_\rho^a\right], \tag{22}$$

$$= \mathbb{E}_{z_1\sim\mathcal{N}(0,Q)}\left[\left(\mathbb{P}_{z\sim\mathcal{N}(0,Q)}\left\{\forall a\in[n], z^a\leq z_1^a\right\}\right)^{p-1}\right]. \tag{23}$$

The equivalence in high dimension of the law of the random projections z_ρ^a with a Gaussian distribution with same first and second moment structure is a phenomenon known as *Gaussian equivalence* in the statistics literature [Goldt et al., 2022, Hu and Lu, 2023, Montanari and Saeed, 2022, Gerace et al., 2024, Wen et al., 2025]. It has been shown for similar rank-one matrix projections (a transposition of eq. (19) with symmetric rank-one projections) in the context of ellipsoid fitting and of learning in large neural networks [Bandeira and Maillard, 2025, Maillard et al., 2024, Xu et al., 2025]. At the statistical physics level of rigour, one typically postulates this equivalence as we did above (see e.g. Maillard and Kunisky [2024]). Nevertheless, for completeness we sketch in what follows an approach towards a rigorous establishment of this property.

Towards a formal establishment of Gaussian equivalence – As we detail below, in order to formally control the concentration of Q^e, one must require that the matrices over which we integrate in eq. (4) have bounded operator norm $\|W\|_{\mathrm{op}}$ (i.e. largest singular value) as $d\to\infty$. This can be done by a different choice of prior distribution P_d, supported on such matrices (as opposed to the Gaussian distribution, which allows arbitrarily large $\|W\|_{\mathrm{op}}$, albeit with exponentially small probability [Anderson et al., 2009]). We note however that the support of these distributions avoids by definition matrices with very large operator norm (with respect to their Euclidean norm, which can be fixed by scale-invariance of the problem), i.e. approximately low-rank matrices. This is a known difficulty in computing the free entropy of matrix problems, see e.g. the example of ellipsoid fitting [Bandeira and Maillard, 2025]. While we do not expect such low-rank matrices to be solutions of the memorisation problem (both from a mathematical[3] and numerical viewpoint), in a mathematically rigorous approach one must rule out their existence with different methods, which we leave to future work focusing on a mathematical establishment of our predictions.

Let us thus assume that we chose such a distribution P_d, so that when considering $J(\{W^a\})$, all matrices W^a are such that $\|W^a\|_{\mathrm{op}}\leq M$, for some fixed $M>0$. Recall now the Hanson-Wright inequality (see e.g. Vershynin [2018]), which we state for Gaussian vectors.

Theorem 2 (Hanson–Wright Inequality, Gaussian vectors)**.** *Let $d\geq 1$ and $X\sim\mathcal{N}(0,\mathbb{I}_d)$. There exists $C>0$ such that for all $A\in\mathbb{R}^{d\times d}$ and $t>0$:*

$$\mathbb{P}\Big(\left|X^\top AX - \mathbb{E}\, X^\top AX\right| > t\Big) \leq 2\exp\left[-C\,\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|_{\mathrm{op}}}\right)\right].$$

Applied to the random variables Q^e_{ab}, Theorem 2 yields:

$$\mathbb{P}_e(|Q^e_{ab} - Q_{ab}| > t) \leq 2\exp\left[-C\min\left(\frac{t^2d^2}{\|W^aW^{b\top}\|_F^2}, \frac{td}{\|W^aW^{b\top}\|_{\mathrm{op}}}\right)\right].$$

We can bound the norms:

$$\begin{cases} \|W^aW^{b\top}\|_{\mathrm{op}} & \leq M^2, \\ \|W^aW^{b\top}\|_F & \leq \sqrt{d}\|W^aW^{b\top}\|_{\mathrm{op}} \leq \sqrt{d}M^2. \end{cases}$$

[2] As we detail below, this holds except for matrices W^a with atypically large operator norm with respect to their Euclidean norm (typically low-rank matrices), which this assumption effectively neglects to compute the asymptotics of $\mathcal{V}_{\mathrm{DP}}$.

[3] In particular we predict that optimal solutions close to the capacity threshold have bounded operator norm, see Claim 2.

Putting all together, we conclude using on top the union bound that for some constant $c = c(M) > 0$:

$$\mathbb{P}_e\left(\max_{1\le a,b\le n}\left|Q^e_{ab} - Q_{ab}\right| > t\right) \le 2n^2\, e^{-cdt}. \tag{24}$$

Coming back to eq. (18), we have

$$J(\{W^a\}) = \mathbb{E}_e[F_{n,p}(Q^e)],$$

and recall the definition of $F_{n,p}(Q)$ in eq. (22). As we will detail later on (see Section C.2), for the matrices Q that dominate the integral in eq. (20), $F_{n,p}$ is in the scale $\exp\{\Theta(\log p)\}$, and we can define a finite function $G_n(Q) \coloneqq \lim_{p\to\infty} \log F_{n,p}(Q)/\log p$. This separation of scales between the concentration of Q^e (exponentially fast in d) and the polynomial scale of J implies that

$$\frac{1}{\log p}\log J(\{W^a\}) = \frac{1}{\log p}\log F_{n,p}(Q) + o(1). \tag{25}$$

Indeed, denoting $\varphi_e(Q^e)$ the law of Q^e under the randomness of Q and using Laplace's method, we have:

$$\begin{aligned}\frac{1}{\log p}\log J(\{W^a\}) &\simeq \frac{1}{\log p}\int \mathrm{d}Q^e e^{(\log p)G_n(Q^e)}\varphi_e(Q^e),\\ &\simeq \sup_{Q^e}\left[G_n(Q^e) + \lim_{p\to\infty}\frac{1}{\log p}\log\varphi_e(Q^e)\right].\end{aligned}$$

By eq. (24), we must have since $\log p \ll d$ that $\lim_{p\to\infty}\log\varphi_e(Q^e)/\log p = -\infty$ except if $Q^e = Q$, which yields the result. Eq. (25) then allows to replace $J(\{W^a\})$ by $F_{n,p}(Q)$ to leading order in $\exp\{\Theta(d^2)\}$ in eq. (18). While this argument can be formalised rigorously using standard large deviation theory, our derivation primarily proceeds at the theoretical physics level of rigour and we therefore leave a rigorous treatment to a future mathematical work.

C.1.2 Replica-symmetric assumption

Going back to eq. (20) and applying Laplace's method, we get:

$$\lim_{d\to\infty}\Phi_d(n) = \sup_{Q\in\mathcal{S}_n^+}\left[J(Q) + \alpha\lim_{p\to\infty}\frac{1}{\log p}\log F_{n,p}(Q)\right], \tag{26}$$

where $\mathcal{S}_n^+$ represents the set of $n\times n$ symmetric positive semidefinite matrices, and where

$$J(Q) \coloneqq \lim_{d\to\infty}\frac{1}{d^2}\log\int\prod_{a=1}^n \mathrm{d}W^a \frac{e^{-\frac{\|W^a\|_F^2}{2}}}{(2\pi)^{d^2/2}}\prod_{1\le a\le b\le n}\delta(d^2Q_{ab} - \mathrm{Tr}[W^aW^{b\top}]),$$

where we rescaled the matrices W^a. Introducing Lagrange multipliers $\hat{Q}_{ab}$, we get:

$$\begin{aligned}J(Q) = \mathrm{extr}_{\hat{Q}\in\mathcal{S}_n}\Bigg[&\frac{1}{2}\sum_{a,b}Q_{ab}\hat{Q}_{ab}\\ &+\lim_{d\to\infty}\frac{1}{d^2}\log\int\prod_{a=1}^n\mathrm{d}W^a\frac{e^{-\frac{\|W^a\|_F^2}{2}}}{(2\pi)^{d^2/2}}e^{-\frac{1}{2}\sum_{1\le a,b\le n}\hat{Q}_{ab}\mathrm{Tr}[W^aW^{b\top}]}\Bigg],\end{aligned} \tag{27}$$

where we wrote the variational problem as an extremum condition over $\hat{Q}$. By simple Gaussian integration, we obtain:

$$\begin{aligned}&\frac{1}{d^2}\log\int\prod_{a=1}^n\mathrm{d}W^a\frac{e^{-\frac{\|W^a\|_F^2}{2}}}{(2\pi)^{d^2/2}}e^{-\frac{1}{2}\sum_{1\le a,b\le n}\hat{Q}_{ab}\mathrm{Tr}[W^aW^{b\top}]}\\ &= \log\int_{\mathbb{R}}\prod_{a=1}^n\frac{\mathrm{d}z^a}{\sqrt{2\pi}}e^{-\frac{(z^a)^2}{2}-\frac{1}{2}\sum_{1\le a,b\le n}\hat{Q}_{ab}z^az^b]},\\ &= -\frac{1}{2}\log\det[\mathbb{I}_n + \hat{Q}].\end{aligned}$$

And therefore we have in eq. (27) that the optimal $\hat{Q}$ satisfies $Q = (\mathbb{I}_n + \hat{Q})^{-1}$, and thus

$$J(Q) = \frac{n}{2} + \frac{1}{2}\log\det Q - \frac{\mathrm{Tr}(Q)}{2}. \tag{28}$$

The replica-symmetric assumption [Mézard et al., 1987], justified here by the convexity of the problem [Talagrand, 2010, Barbier et al., 2022], postulates the form of (Q_{ab}) extremizing the variational principle in eq. (26). It reads:

$$Q_{aa}^{\mathrm{RS}} = Q, \quad Q_{ab}^{\mathrm{RS}} = q, \qquad \forall\, a \neq b, \tag{29}$$

for some parameters $0 \leq q < Q$. Under this assumption, we have from eq. (28)

$$J(Q^{\mathrm{RS}}) = \frac{n(1-Q)}{2} + \frac{n-1}{2}\log(Q-q) + \frac{1}{2}\log[Q + (n-1)q]. \tag{30}$$

C.1.3 The limit $n \to 0$ and the final result

From eq. (26) and the replica symmetric assumption we finally obtain:

$$\begin{aligned}&\lim_{d\to\infty} \Phi_d(n) \qquad (31)\\ &= \sup_{0\leq q<Q}\left[\frac{n(1-Q)}{2} + \frac{n-1}{2}\log(Q-q) + \frac{1}{2}\log[Q+(n-1)q] + \alpha \lim_{p\to\infty}\frac{1}{\log p}\log F_{n,p}(Q^{\mathrm{RS}})\right].\end{aligned}$$

The "energetic" term – Recall that $F_{n,p}$ was defined in eq. (22):

$$F_{n,p}(Q) \coloneqq \mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_p\otimes Q)} \prod_{a=1}^{n}\prod_{\rho=2}^{p} \Theta\left[z_1^a - z_\rho^a\right].$$

For $Q = Q^{\mathrm{RS}}$ we can simplify this drastically. Indeed, noting that (z_ρ^a) are zero-mean Gaussian variables with covariance $\mathbb{E}[z_\rho^a z_\sigma^b] = [(Q-q)\delta_{ab} + q]\delta_{\rho\sigma}$, such a correlation structure can be realised by the representation

$$z_\rho^a = \sqrt{q}\,\eta_\rho + \sqrt{Q-q}\,x_\rho^a, \tag{32}$$

where $\{\eta_\rho\}_{\rho\geq 1}$, $\{x_\rho^a\}_{a,\rho}$, are independent standard Gaussian random variables. We then reach, with $t \coloneqq q/Q \in [0,1)$:

$$F_{n,p}(Q^{\mathrm{RS}}) = G_{n,p}(t) \coloneqq \mathbb{E}_\eta\left[\left(\mathbb{E}_x\left[\prod_{\rho=2}^{p}\Theta\left(x_1 - x_\rho + \sqrt{\frac{t}{1-t}}(\eta_1 - \eta_\rho)\right)\right]\right)^n\right]. \tag{33}$$

In particular we can solve the supremum over Q in eq. (31):

$$\begin{aligned}&\lim_{d\to\infty}\Phi_d(n)\\ &= \sup_{t\in[0,1)}\sup_{Q\geq 0}\left[\frac{n(1-Q)}{2} + \frac{n}{2}\log Q + \frac{n-1}{2}\log(1-t) + \frac{1}{2}\log[1+(n-1)t]\right.\\ &\qquad\left. + \alpha\lim_{p\to\infty}\frac{1}{\log p}\log G_{n,p}(t)\right], \qquad (34)\\ &= \sup_{t\in[0,1)}\left[\frac{1}{2}\log[1+(n-1)t] + \frac{n-1}{2}\log(1-t) + \alpha\lim_{p\to\infty}\frac{1}{\log p}\log G_{n,p}(t)\right]. \qquad (35)\end{aligned}$$

We now take the limit $n \to 0$ in eq. (34). Focusing on the last term and using eq. (33), we get:

$$\lim_{p\to\infty}\frac{1}{\log p}\log G_{n,p}(t) = n\lim_{p\to\infty}\frac{1}{\log p}\mathbb{E}_\eta\big[\log f(t;\eta)\big] + O(n^2), \tag{36}$$

where $\beta_t \coloneqq \sqrt{t/(1-t)}$, $\eta \sim \mathcal{N}(0,\mathbb{I}_p)$, and

$$f_p(t;\eta) \coloneqq \mathbb{E}_{x\sim\mathcal{N}(0,\mathbb{I}_p)}\prod_{\rho=2}^{p}\Theta(x_1 - x_\rho + \beta_t(\eta_1 - \eta_\rho)). \tag{37}$$

All in all, we reach[4]

$$\lim_{d\to\infty}\frac{1}{d^2}\mathbb{E}\log\mathcal{V}_{\rm DP}(E,U)=\inf_{q\in[0,1)}\left[\frac{1}{2}\log(1-q)+\frac{q}{2(1-q)}+\alpha\lim_{p\to\infty}\frac{\mathbb{E}_\eta[\log f_p(q;\eta)]}{\log p}\right], \quad (38)$$

which ends the first part of Result 1, more precisely the derivation of eq. (7).

C.1.4 The capacity threshold

In order to deduce the capacity threshold from eq. (38), we analyse when the convex space of solution shrinks, i.e. when $q=1$ becomes the global infimum of the functional. Let us denote

$$G(q)\coloneqq\lim_{p\to\infty}\frac{\mathbb{E}_\eta[\log f_p(q;\eta)]}{\log p}. \quad (39)$$

The optimal q in eq. (38) satisfies

$$\frac{q}{2(1-q)^2}+\alpha G'(q)=0.$$

Thus we can probe the capacity threshold as the following limit

$$\alpha_c^{-1}=\lim_{q\to1}-2(1-q)^2G'(q)=\lim_{q\to1}-2(1-q)G(q), \quad (40)$$

using L'Hôpital's Rule. As we show in Section C.2, as $q\to1$, we have $G(q)\sim-(1-q)^{-1}$, which finally yields the prediction $\alpha_c=1/2$, and ends the derivation of Result 1.

Finite-size considerations – In Section C.2 we show that the convergence as $p\to\infty$ of $G(q)$ in eq. (39) is very slow, with an error of order $\widetilde{\mathcal{O}}(1/\log p)$ (where $\widetilde{\mathcal{O}}$ hides possible multiplicative $\log\log p$ terms). In turn, via eq. (40), we expect the finite-size corrections to the critical threshold to scale as

$$\alpha_c(p)=\frac{1}{2}+\widetilde{\mathcal{O}}\left(\frac{1}{\log p}\right). \quad (41)$$

Since $p\log p=\alpha d^2$, we have $\log p=2\log d+\mathcal{O}(\log\log d)$, so that, up to the $\log\log$ factors already hidden in $\widetilde{\mathcal{O}}$, the correction can equivalently be written as $\widetilde{\mathcal{O}}(1/\log d)$. To test eq. (41), we measure the empirical critical threshold $\alpha_c(d)$, defined as the largest load at which all patterns are retrieved without errors (Fig. 7).

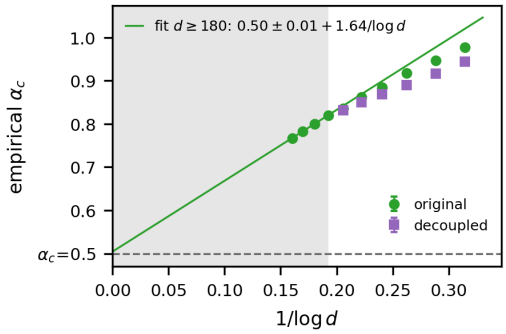


Figure 7: Finite-size scaling of the empirical critical threshold α_c for the original (green) and decoupled (purple) problems, plotted against $1/\log d$ (see Section A for the numerical procedure). As d increases, the decoupled thresholds converge onto the original ones, with which they coincide within error at $d=128$. The line is a linear fit to the original thresholds with $d\geq180$ (shaded); its extrapolation to $d\to\infty$, $\alpha_c=0.50\pm0.01$, agrees with the theoretical value $1/2$ (dashed).

[4]We change back notations from t to $q\in[0,1]$. Notice that the transformation of the extremization over q from a supremum to an infimum in the limit $n\to0$ is also a prediction of the replica method, see e.g. Maillard et al. [2025].

C.2 Expansion of the energetic term

In this appendix we analyse the function $G(q)$ of eq. (39), close to $q = 1$. Recall

$$\begin{cases} G(t) & \coloneqq \lim_{p\to\infty} \dfrac{\mathbb{E}_\eta[\log f_p(t;\eta)]}{\log p}, \\ f_p(t;\eta) & \coloneqq \mathbb{E}_{x\sim\mathcal{N}(0,\mathbb{I}_p)} \displaystyle\prod_{\rho=2}^{p} \Theta(x_1 - x_\rho + \beta_t(\eta_1 - \eta_\rho))\,. \end{cases} \tag{42}$$

Let us first derive a very simple heuristic analysis of the behaviour of $G(t)$ as $t \uparrow 1$. This will also give us insight into the mechanism driving this behaviour. We will then derive mathematically sound lower and upper bounds on $G(t)$ as $t \to 1$, which will validate this heuristic.

C.2.1 Heuristic derivation and mechanistic insight

Heuristic derivation – Notice that

$$G(t) = \lim_{p\to\infty} \frac{1}{\log p} \mathbb{E}_\eta \left[\log \mathbb{P}_{x\sim\mathcal{N}(0,\mathbb{I}_p)} \left(x_1 + \beta_t\eta_1 \geq \max_{\rho\geq 2}[x_\rho + \beta_t\eta_\rho] \right) \right],$$

For $p \to \infty$ and $\beta_t \gg 1$, the random variable $\max_{\rho\geq 2}[x_\rho + \beta_t\eta_\rho] \simeq \beta_t \max_{\rho\geq 2} \eta_\rho \simeq \beta_t\sqrt{2\log p}$, by classical results on the maxima of independent Gaussian variables. We thus approximate in this limit

$$\begin{aligned} &\mathbb{E}_\eta \left[\log \mathbb{P}_{x\sim\mathcal{N}(0,\mathbb{I}_p)} \left(x_1 + \beta_t\eta_1 \geq \max_{\rho\geq 2}[x_\rho + \beta_t\eta_\rho] \right) \right] \\ &\simeq \mathbb{E}_{\eta_1} \left[\log \mathbb{P}_{x_1\sim\mathcal{N}(0,1)} \left(x_1 + \beta_t\eta_1 \geq \beta_t\sqrt{2\log p} \right) \right], \\ &\overset{(a)}{\simeq} \log \mathbb{P}_{x_1\sim\mathcal{N}(0,1)} \left(x_1 \geq \beta_t\sqrt{2\log p} \right), \end{aligned}$$

where in (a) we use that $\eta_1 \ll \sqrt{2\log p}$ with high probability. The value of this last probability is simply a Gaussian tail:

$$\log \mathbb{P}_{x_1\sim\mathcal{N}(0,1)} \left(x_1 \geq \beta_t\sqrt{2\log p} \right) \simeq -\beta_t^2 \log p, \tag{43}$$

and this probability is dominated by the boundary value $x_1 \simeq \beta_t\sqrt{2\log p}$. All in all, this simple heuristic predicts (recall $\beta_t = \sqrt{t/(1-t)}$) that:

$$G(t) \sim_{t\uparrow 1} -\frac{1}{1-t}. \tag{44}$$

Insights on the storing mechanism – While heuristic, the derivation of eq. (44) suggests an interesting insight into the mechanism with which the associations are stored. Recall that one can interpret the variables in eq. (42) as the scores of a fixed pattern association $\mu \in [p]$ (which we took to be $\mu = 1$): for $\rho \in [p]$, the score (diagonal if $\rho = 1$, off-diagonal if $\rho \geq 2$) is equal to $z_\rho = x_\rho + \beta_t\eta_\rho$, as was defined in eq. (32).

The heuristic derivation above (see eq. (43)) suggests that, to leading order, the score $z_1 = x_1 + \beta_t\eta_1$, is dominated by $x_1 \simeq \beta_t\sqrt{2\log p}$ and concentrates on the deterministic value $\beta_t\sqrt{2\log p}$, with fluctuations of sub-leading order (e.g. of order $\mathcal{O}(\beta_t)$ via the fluctuations of η_1). This behaviour is validated by our refined analysis in Section C.2.2, and is strikingly different from the one exhibited by the diagonal scores of the Hebbian ansatz, which have strong fluctuations, as we discuss in Section E. In contrast, our derivation suggests that, when memorisation is possible, the diagonal scores of the optimal matrix are, to leading order, concentrated around a deterministic value as $d, p \to \infty$. This prediction is consistent with the observations of Fig. 4, where we see that below the capacity threshold the variance of the diagonal scores is much smaller than the one of the off-diagonal scores.

C.2.2 Mathematical justification

We now provide a mathematical derivation of eq. (44), using matching upper and lower bounds. Denoting $\lambda \coloneqq -\eta_1$ and $\xi \coloneqq x_1$, we can write f_p equivalently as:

$$f_p(t;\lambda,\eta) = \mathbb{E}_\xi \prod_{\rho=2}^{p} \Phi(\xi - \beta_t(\eta_\rho + \lambda))\,, \tag{45}$$

where Φ denotes the standard normal cumulative distribution function and $\eta = (\eta_2, \cdots, \eta_p)$.

Lower bound – Let us define

$$a_p \coloneqq \max_{\rho \in \{2,\ldots,p\}} (\lambda + \eta_\rho) \tag{46}$$

It holds that, for all $\rho \in [p]$,

$$\Phi\big(\xi - \beta_t(\lambda + \eta_\rho)\big) \geq \Phi\big(\xi - \beta_t a_p\big).$$

In particular,

$$\prod_{\rho=2}^{p} \Phi\big(\xi - \beta_t(\lambda + \eta_\rho)\big) \geq \Phi\big(\xi - \beta_t a_p\big)^{p-1}. \tag{47}$$

Therefore we obtain the lower bound

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \geq \mathbb{E}_{a_p} \log\Big[\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_p\big)^{p-1}\Big]. \tag{48}$$

Let us compute the inner expectation. For any $\varepsilon \in (0,1)$,

$$\begin{aligned}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_p\big)^{p-1} &\geq \int_{\beta_t a_p + (1-\varepsilon)a_p}^{\beta_t a_p + (1+\varepsilon)a_p} d\xi\, \frac{e^{-\xi^2/2}}{\sqrt{2\pi}}\, \Phi\big(\xi - \beta_t a_p\big)^{p-1} \\ &= a_p \int_{1-\varepsilon}^{1+\varepsilon} dx\, \frac{e^{-(\beta_t+x)^2 a_p^2/2}}{\sqrt{2\pi}}\, \Phi(x a_p)^{p-1}.\end{aligned} \tag{49}$$

Since $a_p \sim \sqrt{2\log p}$, we use the approximation for any $x \in (1-\varepsilon, 1+\varepsilon)$

$$\Phi(x a_p) \approx 1 - \frac{e^{-x^2 a_p^2/2}}{\sqrt{2\pi}\, x a_p} \gtrsim 1 - p^{-x^2},$$

which yields

$$\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_p\big)^{p} \gtrsim a_p \int_{1-\varepsilon}^{1+\varepsilon} dx\, p^{-(\beta_t+x)^2} \left(1 - \frac{1}{p^{x^2}}\right)^{p-1}.$$

Since the integrand is positive, we can lower-bound the integral by restricting the integration domain to $[1, 1+\varepsilon]$. On this subinterval, $p^{-(\beta_t+x)^2} \geq p^{-(\beta_t+1+\varepsilon)^2}$. Furthermore, we observe that for any $r \geq 1$:

$$\lim_{p\to\infty} \left(1 - \frac{1}{p^r}\right)^{p-1} = \begin{cases} e^{-1}, & r = 1, \\ 1, & r > 1, \end{cases}$$

so for $x \in [1, 1+\varepsilon]$ the factor $\left(1 - p^{-x^2}\right)^p$ is bounded below by $1/e$, which we absorb into the $\gtrsim$ symbol. Combining these two bounds yields

$$\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\Phi(\xi - \beta_t a_p)^p] \;\gtrsim\; \varepsilon\, a_p\, p^{-(\beta_t+1+\varepsilon)^2}.$$

In particular, we are interested in the behaviour as $t \to 1$, where $\beta_t = \sqrt{t/(1-t)}$ becomes large. Taking logarithms yields

$$\lim_{p\to\infty} \frac{1}{\log p} \mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \geq -(\beta_t + 1 + \varepsilon)^2, \tag{50}$$

and the convergence occurs at a rate at most $O(\log\log p/\log p)$. Since $\varepsilon > 0$ is arbitrary, we conclude

$$G(t) = \lim_{p\to\infty} \frac{1}{\log p} \mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \geq -\left[1 + \sqrt{\frac{t}{1-t}}\right]^2,$$

In particular, as $t \to 1$, we have

$$\lim_{t\to 1}(1-t)G(t) \geq -1. \tag{51}$$

Upper Bound. Fix an integer $k \geq 1$, and let $z_\rho \coloneqq \lambda + \eta_\rho$. Since $\Phi \in (0,1)$ and is decreasing, we can bound the product by:

$$\prod_{\rho=2}^{p} \Phi\big(\xi - \beta_t z_\rho\big) \leq \Phi\big(\xi - \beta_t a_{p-k}\big)^k, \tag{52}$$

where a_{p-k} denotes the $(p-k)$-th order statistic of the variables $\{z_\rho\}$, i.e.

$$a_1 \leq \cdots \leq a_{p-k} \leq a_p.$$

This implies the upper bound

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \leq \mathbb{E}\log\Big[\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_{p-k}\big)^k\Big]. \tag{53}$$

We now tackle the inner expectation. Observe that

$$\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_{p-k}\big)^k = \mathbb{E}_\xi \prod_{i=1}^{k} \mathbb{E}_{\xi_i\sim\mathcal{N}(0,1)}\mathbb{I}\big(0 \leq \xi - \xi_i - \beta_t a_{p-k}\big).$$

Introducing the variables $y_i = \xi - \xi_i - \beta_t a_{p-k}$, the vector $y = (y_1, \ldots, y_k)$ is (conditionally on a_{p-k}) Gaussian with mean $\mathbb{E}[y_i] = -\beta_t a_{p-k}$, and covariance matrix $\Sigma_{ij} = 1 + \delta_{ij}$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_{p-k}\big)^k \\
&= \int_{[0,\infty)^k} d^k y\, \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\Big[-\frac{1}{2}\big(y + \mathbf{1}\beta_t a_{p-k}\big)^\top \Sigma^{-1}\big(y + \mathbf{1}\beta_t a_{p-k}\big)\Big], \\
&= \frac{e^{-\frac{1}{2}\beta_t^2 a_{p-k}^2 \mathbf{1}^\top\Sigma^{-1}\mathbf{1}}}{\sqrt{\det(2\pi\Sigma)}} \int_{[0,\infty)^k} d^k y\, \exp\Big[-\frac{1}{2}y^\top\Sigma^{-1}y - \beta_t a_{p-k}\mathbf{1}^\top\Sigma^{-1}y\Big].
\end{aligned}$$

Rescaling $y' = \beta_t a_{p-k} y$, we obtain (we use that $a_{p-k} > 0$ with overwhelming probability for any fixed k as $p \to \infty$)

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_{p-k}\big)^k \\
&= \frac{e^{-\frac{1}{2}\beta_t^2 a_{p-k}^2 \mathbf{1}^\top\Sigma^{-1}\mathbf{1}}}{(\beta_t a_{p-k})^k\sqrt{\det(2\pi\Sigma)}} \int_{[0,\infty)^k} d^k y\, \exp\Big[-\frac{1}{2(\beta_t a_{p-k})^2}y^\top\Sigma^{-1}y - \mathbf{1}^\top\Sigma^{-1}y\Big].
\end{aligned}$$

In the limit $p \to \infty$, we have $a_{p-k} \sim \sqrt{2\log p}$ for any fixed $k \geq 1$. We thus reach in this limit

$$\int_{[0,\infty)^k} d^k y\, \exp\Big[-\frac{1}{2(\beta_t a_{p-k})^2}y^\top\Sigma^{-1}y - \mathbf{1}^\top\Sigma^{-1}y\Big] \simeq \int_{[0,\infty)^k} d^k y\, \exp\big[-\mathbf{1}^\top\Sigma^{-1}y\big] =: C_k,$$

a constant only depending on k. Using the Sherman–Morrison formula, we get

$$\Sigma^{-1} = I_k - \frac{1}{k+1}\mathbf{1}\mathbf{1}^\top,$$

so that

$$\mathbf{1}^\top\Sigma^{-1}\mathbf{1} = \frac{k}{k+1}.$$

Therefore, we reach:

$$\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_{p-k}\big)^k \simeq \frac{e^{-\frac{k}{2(k+1)}\beta_t^2 a_{p-k}^2}}{(\beta_t a_{p-k})^k}\,\widetilde{C}_k,$$

where $\widetilde{C}_k$ is a finite constant depending only on k. Since $a_{p-k} \sim \sqrt{2\log p}$ for finite $k \geq 1$, we reach by combining this with eq. (53):

$$G(t) = \lim_{p\to\infty}\frac{1}{\log p}\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \leq -\frac{k}{k+1}\frac{t}{1-t}. \tag{54}$$

Similarly to the upper bound, the corrections of order $\mathcal{O}(\log\log p/\log p)$. Since eq. (54) holds for any $k \geq 1$, we get by combining it with eq. (51):

$$\lim_{t\to 1}(1-t)G(t) = -1. \tag{55}$$

Slow convergence in p – Our derivation suggests that the typical corrections to $G(t)$ for finite p are of order $\mathcal{O}(\log\log p/\log p)$. This is validated by the slow convergence that we observe for the capacity threshold, which is compatible with this convergence rate Fig. 7.

Finally, we emphasize that our approach essentially made mathematically sound the intuition laid out in Section C.2.1, by showing that the inner expectation is dominated by values of ξ tightly concentrated around $\beta_t\sqrt{2\log p}$.

C.3 Rank-constrained weights: derivation of Claim 2

We now consider weight matrices constrained to have rank at most $m = \kappa d$ for some $\kappa \in (0, 1]$. We assume that the prior $P_m(\mathrm{d}W)$ is given by $W \overset{\mathrm{d}}{=} QR^\top/\sqrt{dm}$ for $Q, R \in \mathbb{R}^{d\times m}$ with i.i.d. $\mathcal{N}(0,1)$ elements. Notice in particular that $\mathbb{E}[\mathrm{Tr}[WW^\top]] = d$, and that the law of W is invariant by left and right orthogonal transformations: in particular if $\sigma_1, \cdots, \sigma_m$ are the non-zero singular values of W, P_m only depends on W through the empirical distribution $\hat{\rho} \coloneqq (1/m)\sum_{i=1}^m \delta_{\sigma_i}$. While the precise form of P_m chosen here does not bear importance in the derivation of Claim 2, we will use that the law of $\hat{\rho}$ satisfies, under P_m, a large deviations principle in the scale $\mathcal{O}(d^2)$, i.e. that

$$P_m(\hat{\rho} \simeq \mu) \sim \exp\{-d^2 J_\kappa(\rho)\}, \tag{56}$$

for some function J_κ : this is widely expected to hold for a large class of random matrix ensembles [Anderson et al., 2009, Potters and Bouchaud, 2020], and in our heuristic computation we will assume P_m satisfies such a property.

C.3.1 The replica method

The computation of the moments of $\mathcal{V}_{\mathrm{DP}}(E, U)$ in this setting is fairly similar to the one of Section C. In particular, a completely similar Gaussian equivalence phenomenon can be postulated in the high-dimensional limit (see the discussion in Section C.1.1), and leads to:

$$\frac{1}{d^2}\log\mathbb{E}\mathcal{V}_{\mathrm{DP}}(E,U) = \frac{1}{d^2}\log\int\prod_{a=1}^n P_m(\mathrm{d}W^a)\, F_{n,p}(Q)^p + o(1), \tag{57}$$

again with

$$Q_{ab} \coloneqq \frac{1}{d}\mathrm{Tr}[W^a W^{b\top}]. \tag{58}$$

Non-convexity and the replica-symmetric ansatz – Note that the support of the rank-constrained prior $P_m(W)$ is non-convex for $\kappa < 1$. The convexity of the problem was important in justifying the validity of the replica-symmetric assumption. Here, we nonetheless work under the same assumption: it is generically known that it yields an upper bound on the limiting volume (see e.g. [Maillard et al., 2025]), and its predictions are in close agreement with numerical experiments, as detailed in the main text of this paper. Likewise, when the rank constraint is enforced through the two-layer parametrisation $W = QR^\top$ (as in our experiments), the loss becomes non-convex in (Q, R), so gradient-based optimisation is not guaranteed to find the global optimum.

One applies then the replica method to eq. (57), in a very similar way to Section C. The only difference arises from the "entropic" term

$$J(Q) \coloneqq \lim_{d\to\infty}\frac{1}{d^2}\log\int\prod_{a=1}^n P_m(\mathrm{d}W^a)\prod_{1\leq a\leq b\leq n}\delta(d^2 Q_{ab} - d\mathrm{Tr}[W^a W^{b\top}]),$$

which under the replica-symmetric assumption reads:

$$J(Q,q) = \mathrm{extr}_{\hat{Q}\in\mathcal{S}_n}\left[\frac{1}{2}\sum_{a,b} Q_{ab}\hat{Q}_{ab} + \lim_{d\to\infty}\frac{1}{d^2}\log\int\prod_{a=1}^{n} P_m(\mathrm{d}W^a)\, e^{-\frac{d}{2}\sum_{1\le a,b\le n}\hat{Q}_{ab}\mathrm{Tr}[W^a W^{b\top}]}\right],$$

$$= \mathrm{extr}_{\hat{Q},\hat{q}}\left[\frac{nQ\hat{Q}}{2} - \frac{n(n-1)}{2}q\hat{q}\right.$$

$$\left. + \lim_{d\to\infty}\frac{1}{d^2}\log\int\prod_{a=1}^{n} P_m(\mathrm{d}W^a)\, e^{-d\frac{(\hat{Q}+\hat{q})}{2}\sum_{a=1}^{n}\|W^a\|_F^2 + d\frac{\hat{q}}{2}\left\|\sum_{a=1}^{n} W^a\right\|_F^2}\right],$$

where we used the replica-symmetric assumption $\hat{Q}_{ab} = -\hat{q}$, $\hat{Q}_{aa} = \hat{Q}$ for $a \neq b$. We use the identity:

$$e^{d\frac{\hat{q}}{2}\left\|\sum_a W^a\right\|_F^2} = \int\frac{dZ}{(2\pi/d)^{d^2/2}}\, e^{-\frac{d}{2}\|Z\|_F^2 + d\sqrt{\hat{q}}\,\mathrm{Tr}\left(Z^\top\sum_a W^a\right)},$$

This decouples the replicas: each copy of W^a now interacts independently with the shared auxiliary field Z, and the entropic term takes the form

$$J(Q,q) = \mathrm{extr}_{\hat{Q},\hat{q}}\left[\frac{nQ\hat{Q}}{2} - \frac{n(n-1)}{2}q\hat{q} + \lim_{d\to\infty}\frac{1}{d^2}\log\,\mathbb{E}_{Z\sim\mathrm{Gin}(d)}[I(Z)]^n\right], \tag{59}$$

where $Z \sim \mathrm{Gin}(d)$ denotes a $d\times d$ matrix with i.i.d. $\mathcal{N}(0,1/d)$ entries, and

$$I(Z) := \int P_m(\mathrm{d}W)\, e^{-\frac{d}{2}(\hat{Q}+\hat{q})\|W\|_F^2/2 + d\sqrt{\hat{q}}\,\mathrm{Tr}(Z^\top W)}. \tag{60}$$

All in all we reach:

$$\lim_{d\to\infty}\frac{1}{d^2}\log\mathbb{E}\mathcal{V}_{\mathrm{DP}}(E,U)^n = \mathrm{extr}_{Q,q,\hat{Q},\hat{q}}\left[\frac{nQ\hat{Q}}{2} - \frac{n(n-1)}{2}q\hat{q}\right. \tag{61}$$

$$\left. + \lim_{d\to\infty}\frac{1}{d^2}\log\,\mathbb{E}_{Z\sim\mathrm{Gin}(d)}[I(Z)]^n + \alpha\lim_{p\to\infty}\frac{\log G_{n,p}(q/Q)}{\log p}\right],$$

where recall that we defined $G_{n,p}$ in eq. (33). Taking then the limit $n\downarrow 0$ in the replica method, we get a similar expression with respect to eq. (38), with the first entropic term modified:

$$\lim_{d\to\infty}\frac{1}{d^2}\mathbb{E}\log\mathcal{V}_{\mathrm{DP}}(E,U) = \mathrm{extr}_{Q,q,\hat{Q},\hat{q}}\left[\frac{Q\hat{Q}}{2} + \frac{q\hat{q}}{2} + \lim_{d\to\infty}\frac{1}{d^2}\,\mathbb{E}_{Z\sim\mathrm{Gin}(d)}\log[I(Z)]\right. \tag{62}$$

$$\left. + \alpha\lim_{p\to\infty}\frac{\mathbb{E}_\eta[\log f_p(q/Q;\eta)]}{\log p}\right],$$

We now focus on analysing this entropic term. We note that this analysis is very close to the ones performed in Maillard and Kunisky [2024], Erba et al. [2025].

SVD reduction– The auxiliary field $Z \sim \mathrm{Gin}(d)$ also enjoys bi-orthogonal invariance, sharing the same rotational symmetry as $P_m(W)$. This allows us to integrate out the orthogonal degrees of freedom and reduce the problem to one involving only the singular values of W and Z. We decompose W via its compact SVD: $W = U\Sigma V^\top$, where $U, V \in \mathrm{St}(d,m)$ and $\Sigma = \mathrm{diag}(\sigma_1,\dots,\sigma_m)$ with $\sigma_i > 0$. The change of variables becomes [Anderson et al., 2009]:

$$\mathrm{d}W = c_{d,m}\,|\Delta(s^2)|\,d\mu(U)\,d\mu(V)\prod_{i=1}^{m}\sigma_i^{d-m}\,d\sigma_i, \tag{63}$$

where $d\mu(U)$, $d\mu(V)$ are the Haar measures on $\mathrm{St}(d,m)$, and $|\Delta(s^2)| = \prod_{1\le i<j\le m}|\sigma_i^2 - \sigma_j^2|$ is the Vandermonde determinant. Since $P(W)$ depends only on the singular values and $\|W\|_F^2 = \sum_i\sigma_i^2$, the integral $I(Z)$ of eq. (60) becomes

$$\begin{aligned} I(Z) = c_{d,m}\int_{\mathbb{R}_+^m}\prod_{i=1}^{m} d\sigma_i\,\sigma_i^{d-m}\,|\Delta(\sigma^2)|\,P_m(\Sigma)\,e^{-\frac{d}{2}(\hat{Q}+\hat{q})\sum_i\sigma_i^2} \\ \times\int_{\mathrm{St}(d,m)} d\mu(U)\int_{\mathrm{St}(d,m)} d\mu(V)\,\exp\Big\{d\sqrt{\hat{q}}\,\mathrm{Tr}\big(U^\top Z V\Sigma\big)\Big\}. \end{aligned} \tag{64}$$

We introduce the empirical singular value density

$$\hat{\rho}(\sigma) = \frac{1}{m}\sum_{i=1}^{m} \delta(\sigma - \sigma_i),$$

and change integration variables from $(\sigma_1, \dots, \sigma_m)$ to a generic probability density ρ. All terms in the integrand depend on the singular values only through $\hat{\rho}$:

$$\begin{cases} \dfrac{1}{2}\displaystyle\sum_{i\neq j} \log|\sigma_i^2 - \sigma_j^2| & = \dfrac{m^2}{2}\iint \hat{\rho}(\sigma)\hat{\rho}(\sigma')\log|\sigma^2 - \sigma'^2|\, d\sigma\, d\sigma', \\ \displaystyle\sum_{i=1}^{m} \log\sigma_i & = m\int \hat{\rho}(\sigma)\, \log\sigma\, d\sigma, \\ \displaystyle\sum_{i=1}^{m} \sigma_i^2 & = m\int \hat{\rho}(\sigma)\, \sigma^2\, d\sigma. \end{cases} \tag{65}$$

Finally, recall the large deviations assumption stated at the beginning of this section in eq. (56). The change of variables from the m-dimensional empirical density to a generic density ρ introduces a Jacobian of order $\exp(O(m)) = \exp(O(d))$, which is subleading with respect to the $\exp(O(d^2))$ terms and can therefore be neglected[5]. Using $m = \kappa d$ and performing Laplace's method in the space of probability distributions, we finally reach that

$$\begin{aligned} \lim_{d\to\infty} \frac{1}{d^2}\mathbb{E}_Z \log I(Z) = \sup_{\rho\in\mathcal{M}_1^+(\mathbb{R}_+)} \Bigg[& \frac{\kappa^2}{2}\iint \rho(\sigma)\rho(\sigma')\log|\sigma^2 - \sigma'^2|\, d\sigma\, d\sigma' \\ & + \kappa(1-\kappa)\int \rho(\sigma)\log\sigma\, d\sigma \\ & - \frac{\kappa(\hat{Q}+\hat{q})}{2}\int \rho(\sigma)\,\sigma^2\, d\sigma \\ & - J_\kappa(\rho) + \mathcal{I}_{\mathrm{HCIZ}}(\hat{q};\rho,\rho_z)\Bigg]. \end{aligned} \tag{66}$$

where $\mathcal{M}_1^+(\mathbb{R}_+)$ denotes probability distributions supported on $\mathbb{R}_+$, and we defined the rectangular *HCIZ integral* (see Anderson et al. [2009])

$$\mathcal{I}_{\mathrm{HCIZ}}(\hat{q};\rho,\rho_Z) \coloneqq \lim_{d\to\infty} d^{-2}\log\int_{\mathrm{St}(d,m)^2} d\mu(U)d\mu(V)\, \exp\Big\{d\sqrt{\hat{q}}\,\mathrm{Tr}\big(U^\top Z V \Sigma\big)\Big\}, \tag{67}$$

which only depends on the asymptotic empirical distribution ρ_Z of the singular values of Z. The later is known to be the *quarter-circle law*:

$$\rho_z(\sigma) = \rho_{\mathrm{q.c.}}(\sigma) \coloneqq \frac{1}{\pi}\sqrt{4-\sigma^2}, \qquad \sigma\in[0,2]. \tag{68}$$

Note that since W has rank at most $m = \kappa d$, only m singular values are non-zero; the overall singular value distribution of W is therefore

$$\rho_W(\sigma) = (1-\kappa)\,\delta(\sigma) + \kappa\,\rho(\sigma), \tag{69}$$

[5] See Livan et al. [2018] for a general justification of this point, and a a similar discussion in Maillard and Kunisky [2024].

where ρ is the density of the non-zero singular values obtained from the saddle-point equation above. Overall, we reach the final prediction of the replica method (rescaling $t = q/Q \in [0, 1)$):

$$\lim_{d\to\infty} \frac{1}{d^2}\mathbb{E}\log \mathcal{V}_{\mathrm{DP}}(E, U) \tag{70}$$
$$= \underset{Q,\, t,\, \hat{Q},\, \hat{q}}{\mathrm{extr}} \sup_{\rho\in\mathcal{M}_1^+(\mathbb{R}_+)} \left\{ \begin{array}{l} \dfrac{Q\hat{Q}}{2} + \dfrac{1}{2}tQ\hat{q} \\ + \dfrac{\kappa^2}{2}\displaystyle\iint \rho(\sigma)\rho(\sigma')\log|\sigma^2 - \sigma'^2|\, d\sigma\, d\sigma' \\ + \kappa(1-\kappa)\displaystyle\int \rho(\sigma)\, \log\sigma\, d\sigma - J_\kappa(\rho) \\ - \dfrac{\kappa}{2}\Big(\hat{Q} + \hat{q}\Big)\displaystyle\int \rho(\sigma)\, \sigma^2\, d\sigma + \mathcal{I}_{\mathrm{HCIZ}}(\hat{q}; \rho, \rho_{\mathrm{q.c.}}) \\ + \alpha G(t) \end{array} \right\}.$$

where recall that we defined

$$G(t) \coloneqq \lim_{p\to\infty} \frac{\mathbb{E}_\eta[\log f_p(t;\eta)]}{\log p}.$$

C.3.2 Computation of the capacity threshold

With respect to the non-rank-constrained problem (cf eq. (38)), the difference resides in the "entropic" contribution to the formula. While it is an involved term in eq. (70), we can evaluate it close to the capacity threshold thanks to explicit expansions of the HCIZ integral. Interestingly, this also gives us access to the value of the maximising density ρ, corresponding to the asymptotic singular value density of solutions close to the capacity transition. This derivation is in essence close to ones performed in Maillard and Kunisky [2024], Erba et al. [2025], and we refer to these works for more details on some technical points. Let us first write the saddle point equations associated to eq. (70):

$$\left\{ \begin{array}{ll} \hat{q} = -\dfrac{2\alpha}{QG'(t)} & \text{(71a)} \\ \hat{Q} + t\hat{q} = 0 & \text{(71b)} \\ Q = \kappa \displaystyle\int \rho(\sigma)\sigma^2 d\sigma & \text{(71c)} \\ Q(1-t) = 2\dfrac{\partial}{\partial\hat{q}} I_{\mathrm{HCIZ}}(\hat{q}; \rho, \rho_{\mathrm{q.c.}}) & \text{(71d)} \\ \displaystyle\int \rho(\sigma) d\sigma = 1 & \text{(71e)} \\ \kappa^2 \displaystyle\iint \rho(\sigma')\log|\sigma^2 - \sigma'^2| d\sigma' + \kappa(1-\kappa)\log\sigma - \dfrac{\kappa(\hat{Q}+\hat{q})}{2}\sigma^2 & \\ \quad + \dfrac{\delta}{\delta\rho(\sigma)}\left[I_{\mathrm{HCIZ}}(\hat{q}; \rho, \rho_{\mathrm{q.c.}}) - J_\kappa(\rho)\right] - \lambda_0 = 0 & \text{(71f)} \end{array} \right.$$

where the last equation holds for all $\sigma \in \mathrm{supp}(\rho)$, with λ_0 a Lagrange multiplier enforcing the normalisation of the probability density.

Dilute limit ($\hat{q} \to \infty$) As we approach the capacity threshold, we have $t \to 1$. As we have seen in Section C.2, in this limit $G'(t) \sim -(1-t)^{-2}$, and thus $\hat{q} \sim \hat{q}_0/(1-t)^2 \to \infty$, with $\hat{q}_0 = 2\alpha/Q$. We can exploit this, to compute the leading order of the HCIZ integral when the argument becomes very large. Indeed, in this case it is dominated by the maximiser of the integrand in eq. (67), see e.g. Bun et al. [2014]:

$$\mathcal{I}_{\mathrm{HCIZ}}(\hat{q}; \rho, \rho_{\mathrm{q.c.}}) \;\sim_{\hat{q}\to\infty}\; d^{-1}\sqrt{\hat{q}} \max_{U,V\in\mathrm{St}(d,m)} \mathrm{Tr}\big(U^\top Z\, V\, \Sigma\big). \tag{72}$$

The maximum is achieved when U and V align the m largest singular values of Z with the m non-zero singular values of Σ:

$$\max_{U,V\in\mathrm{St}(d,m)} \mathrm{Tr}\big(U^\top Z\, V\, \Sigma\big) = \sum_{i=1}^{m} z^{\downarrow}_{(i)}\ \sigma^{\downarrow}_{(i)} = \sum_{i=1}^{d} z^{\downarrow}_{(i)}\ \sigma^{\downarrow}_{W,(i)}, \tag{73}$$

where $z^{\downarrow}_{(1)} \geq \cdots \geq z^{\downarrow}_{(m)}$ are the m largest singular values of Z, and $\sigma^{\downarrow}_{W,(i)}$ denotes the i-th largest singular value of W (with $\sigma^{\downarrow}_{W,(i)} = 0$ for $i > m$). As $d \to \infty$, the leading order of this sum is simply an integral given by the quantile functions of ρ_W and $\rho_{\mathrm{q.c.}}$. To sum up:

$$\mathcal{I}_{\mathrm{HCIZ}}(\hat{q};\rho,\rho_{\mathrm{q.c.}})\ \sim_{q\to\infty}\ \sqrt{\hat{q}}\ \int_{1-\kappa}^{1} X_{\mathrm{q.c.}}(p)\ X_W(p)\ dp \tag{74}$$

where X_W is the quantile function of $\rho_W = (1-\kappa)\delta_0 + \kappa\rho$. Notice that since the zero singular values of W do not contribute, the integral is restricted to $p \in [1-\kappa, 1]$.

Saddle-point equations close to the SAT/UNSAT transition. We can simplify eq. (71) close to the transition, using that $G'(t) \sim -(1-t)^2$ as $t \uparrow 1$, and using eq. (74). Notice that the terms in the last equation of eq. (71) that do not depend on $\hat{q}, \hat{Q}$ become subdominant[6] as $t \to 1$, and we get to leading order:

$$\begin{cases} \hat{q} \sim \dfrac{2\alpha}{Q(1-t)^2} & \text{(75a)} \\ \hat{Q} + \hat{q} = \hat{Q} + t\hat{q} + (1-t)\hat{q} = (1-t)\hat{q} \sim \dfrac{2\alpha}{Q(1-t)} & \text{(75b)} \\ Q = \kappa \displaystyle\int \rho(\sigma)\sigma^2 d\sigma & \text{(75c)} \\ Q(1-t) \sim \dfrac{1}{\sqrt{\hat{q}}} \displaystyle\int_{1-\kappa}^{1} X_{\mathrm{q.c.}}(p) X_W(p) dp & \text{(75d)} \\ \displaystyle\int \rho(\sigma) d\sigma = 1 & \text{(75e)} \\ \kappa \dfrac{\hat{Q}+\hat{q}}{2}\sigma^2 = \sqrt{\hat{q}} \dfrac{\delta}{\delta\rho(\sigma)} \left[\displaystyle\int_{1-\kappa}^{1} X_{\mathrm{q.c.}}(p) X_W(p) dp \right] & \text{(75f)} \end{cases}$$

Letting $\hat{q}_0 \coloneqq \frac{2\alpha}{Q}$, we reach from eq. (75d).

$$\sqrt{\hat{q}_0}\, Q = \int_{1-\kappa}^{1} dp\ X_{\mathrm{q.c.}}(p)\, X_W(p).$$

Recall that $X_W(p)$ is the quantile function of the singular value distribution of W, i.e. the inverse of the cumulative distribution corresponding to the density in Eq. (69):

$$F_W(x) = \int_{-\infty}^{x} d\sigma' \Big((1-\kappa)\, \delta(\sigma') + \kappa\, \rho(\sigma') \Big), \tag{76}$$

For $p \in [1-\kappa, 1]$, differentiating the identity $F_W(X_W(p)) = p$ with respect to $\rho(\sigma)$ gives:

$$\frac{\delta F_W(X_W(p))}{\delta\rho(\sigma)} = \kappa\, \rho(X_W(p))\, \frac{\delta X_W(p)}{\delta\rho(\sigma)} + \kappa\, \mathbf{1}[\sigma \leq X_W(p)] = 0. \tag{77}$$

This gives access to the functional derivative of the quantile function:

$$\frac{\delta X_W(p)}{\delta\rho(\sigma)} = -\frac{\mathbf{1}[\sigma \leq X_W(p)]}{\rho(X_W(p))}. \tag{78}$$

Substituting into eq. (75f) and performing the change of variables $p = F_W(u)$ for $p \in (1-\kappa, 1)$, we obtain to leading order as $t \to 1$:

$$\frac{\hat{q}_0}{2}\, \sigma^2 + \sqrt{\hat{q}_0} \int_{\sigma}^{\max \mathrm{supp}\rho} X_{\mathrm{q.c.}}(F_W(u))\, du = 0, \tag{79}$$

[6] It is at this point in the derivation that the specific choice of the prior P_m becomes irrelevant when considering the capacity threshold.

which holds for all $\sigma \in \mathrm{supp}(\rho)$. Differentiating with respect to σ yields

$$\sqrt{\hat{q}_0}\,\sigma = X_{\mathrm{q.c.}}\Big((1-\kappa) + \kappa\, F_\rho(\sigma)\Big), \tag{80}$$

from which we read off the cumulative distribution $F_\rho(\sigma)$ of non-zero singular values:

$$F_\rho(\sigma) = \frac{1}{\kappa}\left(F_{\mathrm{q.c.}}\left(\sqrt{\hat{q}_0}\,\sigma\right) - (1-\kappa)\right). \tag{81}$$

From eq. (81), we can see that ρ is a quarter-circle law with variance $1/\hat{q}_0$, and truncated from below: it is supported in $(\sigma_{\min}, \sigma_{\max})$, with

$$\sigma_{\min} = \frac{X_{\mathrm{q.c.}}(1-\kappa)}{\sqrt{\hat{q}_0}}, \text{ and } \sigma_{\max} = \frac{2}{\sqrt{\hat{q}_0}}, \tag{82}$$

and for $\sigma \in (\sigma_{\min}, \sigma_{\max})$ we have

$$\rho(\sigma) = \frac{\sqrt{\hat{q}_0}}{\kappa}\rho_{\mathrm{q.c.}}(\sqrt{\hat{q}_0}\sigma). \tag{83}$$

Substituting the above singular value distribution into eq. (75c), we get:

$$Q = \frac{1}{\hat{q}_0}\int_{X_{\mathrm{q.c.}}(1-\kappa)}^{2} \rho_{\mathrm{q.c.}}(\sigma)\,\sigma^2\,d\sigma. \tag{84}$$

Recalling the definition $\hat{q}_0 = 2\alpha/Q$, we obtain the equation satisfied by the value $\alpha_c(\kappa)$ of the critical threshold:

$$\alpha_c(\kappa) = \frac{1}{2}\int_{X_{\mathrm{q.c.}}(1-\kappa)}^{2} \rho_{\mathrm{q.c.}}(\sigma)\,\sigma^2\,d\sigma. \tag{85}$$

Notably, $X_{\mathrm{q.c.}}(0) = 0$ and $\int_0^2 \rho_{\mathrm{q.c.}}(\sigma)\,\sigma^2\,d\sigma = 1$, we recover $\alpha_c = \frac{1}{2}$ when $\kappa = 1$.

Notice that the equations above do not constrain the value of Q, which is the norm of solutions: this is consistent with the scale invariance of the problem, as the set of solutions is always a cone. Figure 3 compares the singular value spectra (up to a global normalisation) at capacity for the decoupled and original problems, showing excellent agreement between simulations and theoretical predictions and highlighting the equivalence of the two setups.

Comparison with spectrum at initialisation It is instructive to contrast the singular value distribution of the optimal weight matrix at the capacity threshold, given by Claim 2, with the one at initialisation. The natural baseline depends on the parametrisation. For the full-rank case $\kappa = 1$, one typically initialises $W \in \mathbb{R}^{d\times d}$ directly with i.i.d. Gaussian entries; in this case the empirical singular value distribution of W converges to the quarter-circle law $\rho_{\mathrm{q.c.}}$ on $[0, 2]$, which already coincides with the spectrum predicted by Claim 2 at the capacity threshold. This coincidence is, however, specific to $\kappa = 1$: for $\kappa < 1$ the rank constraint cannot be enforced through such an initialisation, and one resorts instead to the two-layer parametrisation $W = UV^\top$ with $U, V \in \mathbb{R}^{d\times m}$, $m = \kappa d$, and i.i.d. Gaussian entries. Standard results in free probability (see e.g. Potters and Bouchaud [2020]) imply that, in the high-dimensional limit, the empirical singular value distribution of W then converges to the free product $\mathrm{MP}_\kappa \boxtimes \mathrm{MP}_\kappa$ of two Marchenko–Pastur laws of shape κ. Equivalently, denoting by G the Stieltjes transform of the law of the squared singular values of W, one has the cubic equation

$$\kappa^2 x^2\, G(x)^3 + 2\kappa(1-\kappa)\, x\, G(x)^2 + \big((1-\kappa)^2 - x\big)G(x) + 1 = 0, \tag{86}$$

from which the bulk density follows by the inverse Stieltjes transformation. The spectra at initialisation (Fig. 8) differ substantially from those of trained matrices in the $UV^\top$ parametrisation (Fig. 3).

D Proof of the capacity threshold of the rank-constrained decoupled problem

In this appendix, we prove Theorem 1: the capacity threshold of the decoupled problem (2) with weights of rank at most $m = \kappa d$ is $\alpha_c^{\mathrm{DP}}(\kappa) = \alpha_c(\kappa)$, the value derived with the replica method in Claim 2. For $\kappa = 1$, this gives $\alpha_c^{\mathrm{DP}} = 1/2$, which, combined with the result of Barnfield et al. [2026] for the original problem, proves Conjecture 1 for full-rank matrices (Corollary 1). The achievability part of the proof also applies to the original problem (1): for every $\kappa \in (0, 1]$ and $\alpha < \alpha_c(\kappa)$, it admits a solution of rank at most m with probability tending to one (Theorem 3).

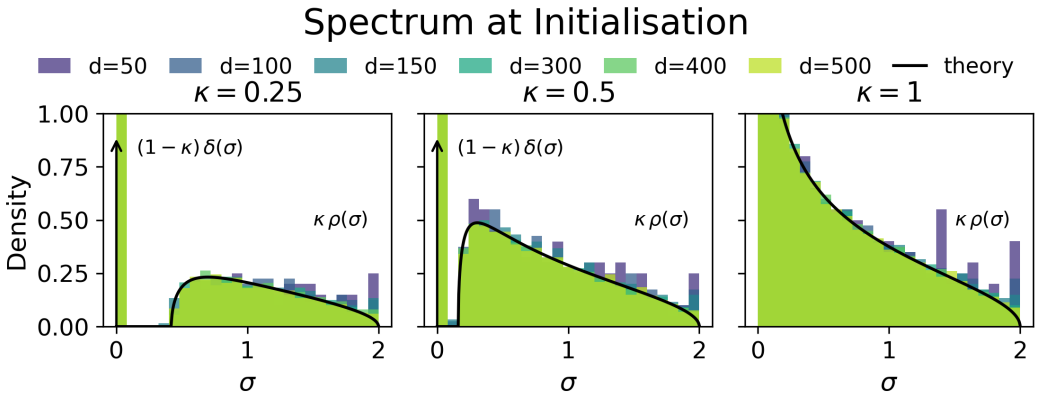


Figure 8: Distribution of singular values of $W = UV^\top$ at initialisation, with $U, V \in \mathbb{R}^{d\times\kappa d}$ having i.i.d. Gaussian entries, for $\kappa \in \{0.25, 0.5, 1\}$, rescaled so that the largest singular value equals 2. Histograms from simulations (shaded) are overlaid with the theoretical prediction (solid line) given by the free multiplicative convolution $\mathrm{MP}_\kappa \boxtimes \mathrm{MP}_\kappa$ [Potters and Bouchaud, 2020]. For $\kappa < 1$, a delta peak at zero (arrow) accounts for the fraction $(1-\kappa)$ of vanishing singular values.

D.1 Setting and main result

Setting. The events in (1) and (2) are invariant under a common rescaling of all embeddings, so we normalise them to have law $\mathcal{N}(0, \mathbb{I}_d/d)$ rather than $\mathcal{N}(0, \mathbb{I}_d)$. Sample indices are denoted by $\mu, \rho \in [p]$, while $i, j \in [d]$ are reserved for vector and matrix coordinates. We treat the two problems in a common notation. For each query $\mu \in [p]$, we call u_μ its *target* and $(u_\rho^{(\mu)})_{\rho\neq\mu}$ its *competitors*. In both problems, the inputs and the targets

$$(e_\mu)_{\mu\in[p]}, \quad (u_\mu)_{\mu\in[p]} \overset{\text{independent}}{\sim} \mathcal{N}(0, \mathbb{I}_d/d)$$

are mutually independent, while the competitors differ:

- in the decoupled problem (2), $u_\mu := u_\mu^{(\mu)}$, and the competitors $\big(u_\rho^{(\mu)}\big)_{\mu,\rho\in[p],\,\rho\neq\mu}$ are independent $\mathcal{N}(0, \mathbb{I}_d/d)$ vectors, independent of the inputs and the targets, including across queries;
- in the original problem (1), all queries share their competitors, which are the other targets: $u_\rho^{(\mu)} := u_\rho$ for all μ and $\rho \neq \mu$.

For $\mathcal{P} \in \{\mathrm{OP}, \mathrm{DP}\}$, the rank-constrained retrieval event is

$$\mathrm{PR}^{\mathcal{P}}_{p,d,m} := \Big\{\exists W \in \mathbb{R}^{d\times d} : \mathrm{rank}(W) \le m, \quad u_\mu^\top W e_\mu > \big(u_\rho^{(\mu)}\big)^\top W e_\mu \ \ \forall\mu,\ \forall\rho\neq\mu\Big\}. \tag{87}$$

A matrix has rank at most m if and only if it factorises through a hidden layer of width m, so (87) is the event that a two-layer linear network with hidden dimension m solves problem $\mathcal{P}$, as in Section 5.3. We assume $1 \le m \le d$ without loss of generality, since a larger rank budget is equivalent to $m = d$. We work in the joint limit $p, d, m \to \infty$ with

$$\frac{m}{d} \longrightarrow \kappa \in (0, 1], \qquad \frac{p\log p}{d^2} \longrightarrow \alpha \in (0, \infty).$$

As in Claim 2, let $\rho_{\mathrm{q.c.}}(s) = \frac{1}{\pi}\sqrt{4-s^2}\,\mathbf{1}_{[0,2]}(s)$ be the quarter-circle density and $X_{\mathrm{q.c.}}$ its quantile function, and define the top-κ spectral mass

$$E_\kappa := \int_{X_{\mathrm{q.c.}}(1-\kappa)}^{2} \rho_{\mathrm{q.c.}}(s) s^2\, ds. \tag{88}$$

Main result. The threshold (9) of Claim 2 is $\alpha_c(\kappa) = E_\kappa/2$, so in this notation Theorem 1 states that

$$\mathbb{P}\big(\mathrm{PR}^{\mathrm{DP}}_{p,d,m}\big) \longrightarrow \begin{cases} 1 & \text{if } \alpha < E_\kappa/2, \\ 0 & \text{if } \alpha > E_\kappa/2. \end{cases} \tag{89}$$

and that $\mathbb{P}(\mathrm{PR}^{\mathrm{OP}}_{p,d,m}) \to 1$ for $\alpha < E_\kappa/2$. For $\kappa = 1$, $X_{\mathrm{q.c.}}(0) = 0$ and $E_1 = \frac{1}{\pi}\int_0^2 s^2\sqrt{4-s^2}\,ds = 1$, so that $\alpha_c^{\mathrm{DP}} = 1/2$. Since Barnfield et al. [2026, Theorem 1] prove $\alpha_c^{\mathrm{OP}} = 1/2$ under the same Gaussian model, this gives Corollary 1.

For the original problem above the threshold, the obstacle is the rank constraint. The result of Barnfield et al. [2026] gives $\mathbb{P}(\mathrm{PR}^{\mathrm{OP}}_{p,d,m}) \to 0$ only for $\alpha > 1/2$, where no matrix of any rank is a solution, because their argument cannot account for the rank constraint (see below). For $\kappa = 1$ this matches the threshold $E_1/2 = 1/2$. For $\kappa < 1$, however, $E_\kappa < 1$, so it remains open whether (1) admits a solution of rank at most m for $E_\kappa/2 < \alpha < 1/2$.

Relation to Barnfield et al. [2026]. The achievability part follows the sparse-repair construction of Barnfield et al. [2026] for the full-rank original problem. Their memory is block-diagonal: a large block stores all associations in the correlation matrix $C = \sum_\mu u_\mu e_\mu^\top$ (the Hebbian ansatz up to normalisation), and a block of vanishing relative dimension d_1, of order $d(\log p)^{-1/8}$, adds weighted rank-one corrections for the few queries whose margin is small. We replace the correlation matrix in the large block by its rank-m_0 truncated singular value decomposition, with $m_0 = m - d_1$. This lowers the target scores from 1 to E_κ and the competitor variances from $\alpha/\log p$ to $\alpha E_\kappa/\log p$ (Lemmas 2 and 3), which moves the threshold from $1/2$ to $E_\kappa/2$. For $\kappa = 1$, no truncation occurs and the construction coincides with theirs. The impossibility part uses a different argument. Barnfield et al. [2026] construct a dual certificate: a nonnegative combination of the constraint matrices, not all zero, that vanishes. Such a certificate cannot account for the rank constraint: it excludes every matrix W at once, and a convex argument of this kind cannot distinguish the matrices of rank at most m from their convex hull, which is the whole of $\mathbb{R}^{d\times d}$. It can therefore only prove impossibility when no matrix of any rank is a solution, i.e. for $\alpha > 1/2$. We give instead a direct argument: summing the constraints over the queries yields a necessary condition, comparing the total target score with the sum of the competitor maxima, and we bound both sides uniformly over all matrices of rank at most m (Lemmas 4 and 5). Although summing discards information, the resulting condition is tight: it fails for $\alpha > E_\kappa/2$, exactly the threshold of the achievability part of the proof, making the threshold sharp.

Where the decoupling is used. Only the impossibility part of the proof uses the decoupling.

- *Achievability* (Theorem 3) holds for both problems. It handles one query at a time: besides properties of the inputs and the targets, it only uses that the competitor scores of each query have a rotation-invariant law (Lemma 3). This holds in both problems, even though in the original problem the competitors also enter the memory. The dependence between the competitor scores of different queries plays no role.
- *Impossibility* (Theorem 4) must handle all queries jointly, uniformly over all matrices of rank at most m. In Lemma 4, the competitors of different queries are independent, which gives the product tail bound (148) needed for a union bound over all such matrices. In the original problem all queries share the same competitors, and this product structure is lost.

For full-rank matrices, Barnfield et al. [2026] prove impossibility for the original problem through convexity, with a dual certificate that, as explained above, cannot see the rank constraint. Our argument instead accommodates the rank constraint, but relies on the decoupling. An impossibility argument that handles both the coupling of the original problem and the non-convexity of the rank constraint is the missing step towards Conjecture 1 for $\kappa < 1$.

Organisation. Section D.2 collects properties of the truncated correlation matrix and of the competitor scores, for both problems. Section D.3 proves achievability for both problems and Section D.4 proves impossibility for the decoupled problem; together they prove Theorem 1. Section D.5 contains the remaining technical lemmas.

D.2 Truncated-SVD preliminaries

The correlation memory below is the starting point of the full-rank construction of Barnfield et al. [2026]; here we use its rank-m truncation. Except for Lemma 3, which treats the competitors of both problems, this subsection involves only the inputs and the targets, and therefore applies to both problems.

Set

$$E = [e_1, \dots, e_p], \qquad U = [u_1, \dots, u_p], \qquad C = UE^\top = \sum_{\mu=1}^{p} u_\mu e_\mu^\top. \tag{90}$$

Note that U contains only the targets. In the decoupled problem, it is independent of the competitors; in the original problem, its columns are also the competitors. Choose a singular value decomposition $C = \sum_{r=1}^{d} s_r(C) a_r b_r^\top$, where $s_1(C) \geq \cdots \geq s_d(C) \geq 0$ and (a_r) and (b_r) are orthonormal bases of $\mathbb{R}^d$. Define the rank-m truncation by

$$T_m(C) \coloneqq \sum_{r=1}^{m} s_r(C) a_r b_r^\top. \tag{91}$$

Conditionally on U, the entries of C are jointly Gaussian with a covariance that is almost surely non-degenerate (see (97) below), so C almost surely has distinct singular values, and $T_m(C)$ does not depend on the choice of the decomposition.

Lemma 1 (Conditional Gaussian structure of the correlation memory)**.** *Set*

$$S_U \coloneqq \frac{d}{p} U U^\top, \qquad S_E \coloneqq \frac{d}{p} E E^\top. \tag{92}$$

If $G \in \mathbb{R}^{d \times d}$ has independent $\mathcal{N}(0, 1/d)$ entries and is independent of U, then, conditionally on U,

$$\sqrt{\frac{d}{p}}\, C \overset{\text{law}}{=} S_U^{1/2} G. \tag{93}$$

Moreover, for a universal constant $c > 0$, each of the two bounds

$$\|S_U - \mathbb{I}_d\|_{\text{op}}, \ \|S_E - \mathbb{I}_d\|_{\text{op}} \leq c\left(\sqrt{\frac{d}{p}} + \frac{d}{p}\right) = O\left(\sqrt{\frac{\log p}{d}}\right) = o(1) \tag{94}$$

holds with probability at least $1 - 2e^{-d}$. Consequently, for all sufficiently large p,

$$\left\| \sqrt{\frac{d}{p}}\, C \right\|_{\text{op}} \leq 3 \tag{95}$$

with probability at least $1 - 4e^{-c'd}$, for a universal constant $c' > 0$.

Proof. For $i, j \in [d]$,

$$C_{ij} = \sum_{\mu=1}^{p} u_{i\mu} e_{j\mu}. \tag{96}$$

Conditional on U, the entries of C are jointly centered Gaussian, and independence of the entries of E gives

$$\begin{aligned} \operatorname{Cov}(C_{ij}, C_{i'j'} \mid U) &= \sum_{\mu,\nu=1}^{p} u_{i\mu} u_{i'\nu} \mathbb{E}[e_{j\mu} e_{j'\nu}] \\ &= \frac{\mathbf{1}_{\{j=j'\}}}{d} (UU^\top)_{ii'}. \end{aligned} \tag{97}$$

Hence the columns of $\sqrt{d/p}\, C$ are conditionally independent centered Gaussian vectors with covariance S_U/d, which proves (93).

Writing $x_\mu = \sqrt{d}\, u_\mu$, we have $S_U = p^{-1} \sum_{\mu=1}^{p} x_\mu x_\mu^\top$ with $x_\mu \sim \mathcal{N}(0, \mathbb{I}_d)$ independently. Let $A \in \mathbb{R}^{p \times d}$ have rows $x_\mu^\top$, so $S_U = A^\top A / p$. The two-sided sub-Gaussian matrix bound [Vershynin, 2018, Theorem 4.6.1], applied with $t = \sqrt{d}$, gives

$$\mathbb{P}\left(\|S_U - \mathbb{I}_d\|_{\text{op}} > c\left(\sqrt{\frac{d}{p}} + \frac{d}{p}\right) \right) \leq 2e^{-d},$$

which is (94) for S_U. Since E has the same law as U, the same bound holds for S_E. Finally, the Gaussian operator-norm bound [Vershynin, 2018, Theorem 7.3.1 and Corollary 7.3.3] gives $\|G\|_{\mathrm{op}} \le 5/2$ except with probability $2e^{-c'd}$. On the intersection of this event with (94) for S_U, we have $\left\|S_U^{1/2} G\right\|_{\mathrm{op}} \le \|S_U\|_{\mathrm{op}}^{1/2} \|G\|_{\mathrm{op}} \le (1+o(1))\, 5/2 \le 3$ for all large p, and (93) proves (95). $\square$

Lemma 2 (Truncated-SVD self-averaging properties)**.** *Let $m/d \to \kappa \in (0,1]$ and $p \log p/d^2 \to \alpha \in (0,\infty)$, and set $T = T_m(C)$.*

(a) There is a universal constant $c > 0$ such that, for every fixed $\eta > 0$ and all sufficiently large p,

$$\mathbb{P}\left(\left| \frac{1}{p} \|T\|_{\mathrm{F}}^2 - E_\kappa \right| > \eta \right) \le e^{-cd}. \tag{98}$$

(b) For every fixed $\varepsilon > 0$,

$$\max_{\mu \le p} \mathbb{P}\big(\big| u_\mu^\top T e_\mu - E_\kappa \big| > \varepsilon \big) = O_\varepsilon\left(\frac{1}{\log p} \right). \tag{99}$$

(c) In probability,

$$\max_{\mu \le p} \left| \frac{d}{p} \|T e_\mu\|^2 - E_\kappa \right| \longrightarrow 0. \tag{100}$$

(d) For some fixed $A_0 = A_0(\alpha) < \infty$,

$$\mathbb{P}\left(\max_{\mu \le p} \big| u_\mu^\top T e_\mu \big| \le A_0 \right) \longrightarrow 1. \tag{101}$$

Proof. Throughout the proof, $c, c' > 0$ denote universal constants whose values may change from line to line. Let $b_d \coloneqq c(\sqrt{d/p} + d/p) = o(1)$ be the bound in (94), and define the event

$$\Omega_0 \coloneqq \left\{ \begin{array}{ll} \|S_E - \mathbb{I}_d\|_{\mathrm{op}} \le b_d, & \|E\|_{\mathrm{op}} \le 2\sqrt{p/d}, \\ \|U\|_{\mathrm{op}} \le 2\sqrt{p/d}, & \|C\|_{\mathrm{op}} \le 3\sqrt{p/d} \end{array} \right\}. \tag{102}$$

The Gaussian operator-norm bound [Vershynin, 2018, Theorem 7.3.1 and Corollary 7.3.3] gives $\|E\|_{\mathrm{op}} \le \sqrt{p/d} + 2 \le 2\sqrt{p/d}$ for all large p, except with probability $2e^{-cd}$, and likewise for U. Together with Lemma 1, this gives $\mathbb{P}(\Omega_0^c) \le e^{-cd}$ for all large p. On Ω_0, since $\mathrm{rank}(T) \le m \le d$,

$$\|T\|_{\mathrm{op}} \le \|C\|_{\mathrm{op}} \le 3\sqrt{\frac{p}{d}}, \qquad \|T\|_{\mathrm{F}}^2 \le m \|C\|_{\mathrm{op}}^2 \le 9p. \tag{103}$$

Rotation invariance. Let Q be Haar distributed on the orthogonal group $O(p)$ and independent of (E,U), and let q_μ denote the μth column of Q. Each q_μ is uniform on $\mathbb{S}^{p-1}$. The rows of E and of U are independent $\mathcal{N}(0, \mathbb{I}_p/d)$ vectors, so $(EQ, UQ) \overset{\text{law}}{=} (E,U)$. The substitution $(E,U) \mapsto (EQ, UQ)$ leaves $C = (UQ)(EQ)^\top$, T, S_U, S_E and Ω_0 unchanged, and maps (e_μ, u_μ) to (Eq_μ, Uq_μ). Hence, setting

$$K \coloneqq U^\top T E, \qquad L \coloneqq E^\top T^\top T E \in \mathbb{R}^{p \times p},$$

which are functions of (E,U) only, we obtain the identity in law

$$\Big(\big(u_\mu^\top T e_\mu\big)_{\mu \le p},\ \big(\|T e_\mu\|^2\big)_{\mu \le p} \Big) \overset{\text{law}}{=} \Big(\big(q_\mu^\top K q_\mu\big)_{\mu \le p},\ \big(q_\mu^\top L q_\mu\big)_{\mu \le p} \Big). \tag{104}$$

In Steps 2–4 we bound the right-hand side conditionally on (E,U), using only the randomness of Q. We will use the identities

$$\mathrm{tr}\, K = \langle T, C \rangle_{\mathrm{F}} = \|T\|_{\mathrm{F}}^2, \qquad \mathrm{tr}\, L = \frac{p}{d} \mathrm{tr}\big(T^\top T S_E\big), \tag{105}$$

where the first holds because $\langle T_m(C), C \rangle_{\mathrm{F}} = \sum_{r \le m} s_r(C)^2$.

1. Proof of (98)**.** By (93), $\sqrt{d/p}\,C$ has the same law as $S_U^{1/2}G$, with G as in Lemma 1. Since the probability in (98) depends only on the law of C, it suffices to prove (98) with $\sqrt{d/p}\,C$ replaced by $S_U^{1/2}G$, which we do from now on in this step. Let $\Omega_1 \coloneqq \{\|S_U - \mathbb{I}_d\|_{\mathrm{op}} \le b_d,\ \|G\|_{\mathrm{op}} \le 3\}$. By (94) and the Gaussian operator-norm bound, $\mathbb{P}(\Omega_1^c) \le e^{-cd}$. On Ω_1, the singular-value form of Weyl's inequality, which applies even though the matrices need not be symmetric,[7] gives

$$\max_{r\le d}\left|\sqrt{\frac{d}{p}}\,s_r(C) - s_r(G)\right| \le \left\|(S_U^{1/2}-\mathbb{I}_d)G\right\|_{\mathrm{op}} \le \|G\|_{\mathrm{op}}\,\|S_U-\mathbb{I}_d\|_{\mathrm{op}} \le 3b_d,$$

where we used $\left\|S_U^{1/2}-\mathbb{I}_d\right\|_{\mathrm{op}} \le \|S_U-\mathbb{I}_d\|_{\mathrm{op}}$, which follows from $S_U \succeq 0$ and $\left|\sqrt{\lambda}-1\right| \le |\lambda-1|$ for $\lambda \ge 0$. All the singular values compared here are at most 4, since $\left\|S_U^{1/2}G\right\|_{\mathrm{op}} \le 3(1+b_d)^{1/2} \le 4$, so

$$\left|\frac{1}{p}\,\|T\|_{\mathrm{F}}^2 - \frac{1}{d}\,\|T_m(G)\|_{\mathrm{F}}^2\right| = \frac{1}{d}\left|\sum_{r=1}^{m}\left(\frac{d}{p}\,s_r(C)^2 - s_r(G)^2\right)\right| \le 24b_d = o(1).$$

It remains to show that $\frac{1}{d}\,\|T_m(G)\|_{\mathrm{F}}^2$ concentrates around E_κ. Write $G = Z/\sqrt{d}$, with Z standard Gaussian, and set

$$F(Z) \coloneqq \frac{\|T_m(Z)\|_{\mathrm{F}}}{d} = \frac{1}{d}\sup_{\substack{\mathrm{rank}(B)\le m\\ \|B\|_{\mathrm{F}}\le 1}} \langle Z, B\rangle_{\mathrm{F}}.$$

The variational formula makes F $1/d$-Lipschitz. Gaussian concentration [Vershynin, 2018, Theorem 5.2.2] therefore gives $\mathbb{P}(|F-\mathbb{E}F| > t) \le 2e^{-cd^2t^2}$. The quarter-circle law for the singular values of G, i.e. the Marchenko–Pastur law with unit aspect ratio for the eigenvalues of $GG^\top$ [Anderson et al., 2009], gives[8]

$$F^2 = \frac{1}{d}\,\|T_m(G)\|_{\mathrm{F}}^2 = \frac{1}{d}\sum_{r=1}^{m} s_r(G)^2 \longrightarrow \int_{X_{\mathrm{q.c.}}(1-\kappa)}^{2} s^2\rho_{\mathrm{q.c.}}(s)\,ds = E_\kappa$$

in probability. Since $\mathbb{E}F^2 \le \mathbb{E}\,\|Z\|_{\mathrm{F}}^2/d^2 = 1$, the family F is uniformly integrable, so $\mathbb{E}F \to \sqrt{E_\kappa}$. Hence, for each fixed $\eta > 0$ and all large d,

$$\mathbb{P}\left(\left|\frac{1}{d}\,\|T_m(G)\|_{\mathrm{F}}^2 - E_\kappa\right| > \frac{\eta}{2}\right) \le 2e^{-c_\eta d^2}.$$

For all large p, $24b_d \le \eta/2$, and the estimates above give

$$\mathbb{P}\left(\left|\frac{1}{p}\,\|T\|_{\mathrm{F}}^2 - E_\kappa\right| > \eta\right) \le \mathbb{P}(\Omega_1^c) + 2e^{-c_\eta d^2} \le e^{-cd} + 2e^{-c_\eta d^2},$$

which is at most $e^{-c'd}$ for all large p. This proves (98).

2. Proof of (99)**.** Fix μ and write $q = q_\mu$. By (104) and (105),

$$\mathbb{P}\left(\left|u_\mu^\top T e_\mu - \frac{\|T\|_{\mathrm{F}}^2}{p}\right| > \frac{\varepsilon}{2}\right) = \mathbb{P}\left(\left|q^\top K q - \frac{\mathrm{tr}\,K}{p}\right| > \frac{\varepsilon}{2}\right).$$

[7] For arbitrary real $d\times d$ matrices A, B, set $\mathcal{H}(A) = \left(\begin{smallmatrix} 0 & A \\ A^\top & 0\end{smallmatrix}\right)$. The d largest eigenvalues of the symmetric matrix $\mathcal{H}(A)$ are $s_1(A),\dots,s_d(A)$, while $\mathcal{H}(A-B)^2 = \mathrm{diag}((A-B)(A-B)^\top, (A-B)^\top(A-B))$ gives $\|\mathcal{H}(A)-\mathcal{H}(B)\|_{\mathrm{op}} = \|A-B\|_{\mathrm{op}}$. Applying the eigenvalue version of Weyl's inequality to $\mathcal{H}(A)$ and $\mathcal{H}(B)$ yields $|s_r(A)-s_r(B)| \le \|A-B\|_{\mathrm{op}}$.

[8] The empirical singular value distribution of G converges weakly in probability to $\rho_{\mathrm{q.c.}}$, whose distribution function is continuous and strictly increasing on $[0,2]$, and $\|G\|_{\mathrm{op}} \le 5/2$ with probability tending to one. Hence $s_m(G) \to X_{\mathrm{q.c.}}(1-\kappa)$ in probability, and for every fixed $\delta > 0$, with probability tending to one, $\frac{1}{d}\sum_{r\le m} s_r(G)^2$ lies between $\int s^2\mathbf{1}_{\{s\ge X_{\mathrm{q.c.}}(1-\kappa)\pm\delta\}}\,d\mu_G(s)$, where μ_G is the empirical singular value distribution. Both bounds converge to within $O(\delta)$ of E_κ.

Haar invariance gives $\mathbb{E}_Q[qq^\top] = \mathbb{I}_p/p$, hence

$$\mathbb{E}_Q[q^\top Kq] = \operatorname{tr}\big(K\mathbb{E}_Q[qq^\top]\big) = \frac{\operatorname{tr} K}{p}. \tag{106}$$

Write $K_s = (K + K^\top)/2$, so that $q^\top Kq = q^\top K_s q$. The fourth moments of a uniform $q \in \mathbb{S}^{p-1}$ are

$$\mathbb{E}_Q[q_i q_j q_k q_\ell] = \frac{\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}}{p(p+2)}.$$

Expanding $(q^\top K_s q)^2$ gives $\mathbb{E}_Q[(q^\top Kq)^2] = ((\operatorname{tr} K_s)^2 + 2\,\|K_s\|_{\mathrm{F}}^2)/(p(p+2))$, so

$$\operatorname{Var}_Q(q^\top Kq) = \frac{2}{p(p+2)}\left(\|K_s\|_{\mathrm{F}}^2 - \frac{(\operatorname{tr} K_s)^2}{p}\right) \le \frac{2\,\|E\|_{\mathrm{op}}^2\,\|U\|_{\mathrm{op}}^2\,\|T\|_{\mathrm{F}}^2}{p^2}, \tag{107}$$

using $\|K_s\|_{\mathrm{F}} \le \|K\|_{\mathrm{F}} \le \|U\|_{\mathrm{op}}\,\|T\|_{\mathrm{F}}\,\|E\|_{\mathrm{op}}$. On the event Ω_0 of (102), by (103), the right-hand side is $O(p/d^2)$. Since Ω_0 depends only on (E, U), Chebyshev's inequality conditional on (E, U) gives

$$\mathbb{P}\left(\left|q^\top Kq - \frac{\operatorname{tr} K}{p}\right| > \frac{\varepsilon}{2}\right) \le \mathbb{P}(\Omega_0^c) + O\Big(\frac{p}{\varepsilon^2 d^2}\Big) = O\left(\frac{1}{\varepsilon^2 \log p}\right), \tag{108}$$

where we used $p/d^2 = (p\log p/d^2)/\log p$. Together with (98) at $\eta = \varepsilon/2$ and the triangle inequality, this proves (99); none of the bounds depends on μ.

3. Proof of (100)**.** By (104), it suffices to prove that $\max_{\mu\le p}\left|\frac{d}{p}q_\mu^\top L q_\mu - E_\kappa\right| \to 0$ in probability. We first locate the center. By (105), on Ω_0,

$$\left|\frac{d}{p^2}\operatorname{tr} L - \frac{\|T\|_{\mathrm{F}}^2}{p}\right| = \frac{1}{p}\left|\operatorname{tr}\big(T^\top T(S_E - \mathbb{I}_d)\big)\right| \le \|S_E - \mathbb{I}_d\|_{\mathrm{op}}\,\frac{\|T\|_{\mathrm{F}}^2}{p} \le 9b_d, \tag{109}$$

so $\frac{d}{p^2}\operatorname{tr} L \to E_\kappa$ in probability by (98). It remains to control the deviations of the diagonal entries $q_\mu^\top L q_\mu$ from their Haar mean $\operatorname{tr} L/p$, simultaneously for all μ. On the event Ω_0 of (102), (103) gives

$$\|L\|_{\mathrm{op}} \le \|E\|_{\mathrm{op}}^2\,\|T\|_{\mathrm{op}}^2 \le \frac{36p^2}{d^2}, \qquad \operatorname{tr} L \le \|E\|_{\mathrm{op}}^2\,\|T\|_{\mathrm{F}}^2 \le \frac{36p^2}{d},$$

hence $\|L\|_{\mathrm{F}}^2 \le \|L\|_{\mathrm{op}}\operatorname{tr} L \le 36^2p^4/d^3$, since $L \succeq 0$. Conditionally on $(E, U) \in \Omega_0$, Lemma 6 applied to $H = L$ gives, except with probability $4/p$,

$$\frac{d}{p}\max_{\mu\le p}\left|q_\mu^\top L q_\mu - \frac{\operatorname{tr} L}{p}\right| \le C_1\frac{d}{p^2}\left(\|L\|_{\mathrm{F}}\sqrt{\log p} + \|L\|_{\mathrm{op}}\log p\right) \le 36C_1\left(\sqrt{\frac{\log p}{d}} + \frac{\log p}{d}\right),$$

which tends to zero. Since $\mathbb{P}(\Omega_0^c) \to 0$, this proves (100).

4. Proof of (101)**.** By (104), $\max_{\mu\le p}\left|u_\mu^\top T e_\mu\right|$ has the law of $\max_{\mu\le p}\left|q_\mu^\top K_s q_\mu\right|$, with K_s as in Step 2. On the event Ω_0 of (102), (105) and (103) give $\operatorname{tr} K_s/p = \|T\|_{\mathrm{F}}^2/p \le 9$ and, since symmetrisation does not increase the Frobenius or the operator norm,

$$\|K_s\|_{\mathrm{F}} \le \|U\|_{\mathrm{op}}\,\|T\|_{\mathrm{F}}\,\|E\|_{\mathrm{op}} \le \frac{12p^{3/2}}{d}, \qquad \|K_s\|_{\mathrm{op}} \le \|U\|_{\mathrm{op}}\,\|T\|_{\mathrm{op}}\,\|E\|_{\mathrm{op}} \le 12\left(\frac{p}{d}\right)^{3/2}.$$

Conditionally on $(E, U) \in \Omega_0$, Lemma 6 applied to $H = K_s$ gives, except with probability $4/p$,

$$\max_{\mu\le p}\left|q_\mu^\top K_s q_\mu\right| \le 9 + 12C_1\left(\frac{\sqrt{p\log p}}{d} + \frac{\sqrt{p}\log p}{d^{3/2}}\right).$$

The first ratio tends to $\sqrt{\alpha}$, while the second equals $\frac{\sqrt{p\log p}}{d}\sqrt{\frac{\log p}{d}} \to 0$. Hence (101) holds with any $A_0 > 9 + 12C_1\sqrt{\alpha}$. Unlike (99), this is only a crude bound: uniformly over μ, the fluctuations of $u_\mu^\top T e_\mu$ are of order $\sqrt{p\log p}/d \to \sqrt{\alpha}$, which is why the achievability proof needs a repair step. □

Lemma 3 (Competitor scores of the truncated memory)**.** *Let $m/d \to \kappa \in (0,1]$ and $p\log p/d^2 \to \alpha \in (0,\infty)$, and set $T = T_m(C)$. In either problem, collect the competitors of query μ in $U^{(\mu)} := \big[u_\rho^{(\mu)}\big]_{\rho\neq\mu} \in \mathbb{R}^{d\times(p-1)}$, and write*

$$Z_{\rho\mu} := \big(u_\rho^{(\mu)}\big)^\top T e_\mu \quad (\rho\neq\mu), \qquad Z_\mu := \big(Z_{\rho\mu}\big)_{\rho\neq\mu} = \big(U^{(\mu)}\big)^\top T e_\mu \in \mathbb{R}^{p-1}$$

for the competitor scores of query μ. In the decoupled problem, $U^{(\mu)}$ is a Gaussian matrix independent of (E,U); in the original problem, it is U with its μth column removed. Then, in both problems, the following hold.

(a) For every μ, the law of Z_μ is invariant under all orthogonal transformations of $\mathbb{R}^{p-1}$. Consequently, for all $t, r > 0$,

$$\mathbb{P}\Big(\max_{\rho\neq\mu} Z_{\rho\mu} \ge t,\ \|Z_\mu\|^2 \le r(p-1)\Big) \le (p-1)\, e^{-t^2/(2r)}. \tag{110}$$

(b) In probability,

$$\max_{\mu\le p}\left|\frac{\log p}{p-1}\|Z_\mu\|^2 - \alpha E_\kappa\right| \longrightarrow 0. \tag{111}$$

Proof. **1. Rotation invariance.** In the decoupled problem, conditionally on (E,U), the vector Te_μ is fixed and the columns of $U^{(\mu)}$ are still independent $\mathcal{N}(0,\mathbb{I}_d/d)$ vectors, so $Z_\mu = (U^{(\mu)})^\top Te_\mu$ is conditionally $\mathcal{N}\big(0, \frac{\|Te_\mu\|^2}{d}\mathbb{I}_{p-1}\big)$. This conditional law is rotation invariant, hence so is the law of Z_μ.

In the original problem, $U^{(\mu)}$ is U without its μth column. Fix $R \in O(p-1)$, and let $\Pi \in O(p)$ act as R on the coordinates $\rho \neq \mu$ and leave the μth coordinate unchanged. The rows of E and of U are independent $\mathcal{N}(0,\mathbb{I}_p/d)$ vectors, so $(E\Pi, U\Pi) \overset{\text{law}}{=} (E,U)$. The substitution $(E,U)\mapsto(E\Pi,U\Pi)$ leaves $C = (U\Pi)(E\Pi)^\top$, hence T, unchanged, and it also leaves e_μ, the μth column of E, unchanged. It maps $U^{(\mu)}$ to $U^{(\mu)}R$, and therefore $Z_\mu = (U^{(\mu)})^\top Te_\mu$ to $R^\top Z_\mu$. Thus $R^\top Z_\mu \overset{\text{law}}{=} Z_\mu$ for every $R \in O(p-1)$.

2. Proof of (110)**.** We use the spherical-cap bound [Ball, 1997, Lemma 2.2]: if θ is uniformly distributed on the unit sphere $\mathbb{S}^{n-1}$ of $\mathbb{R}^n$ and $v \in \mathbb{S}^{n-1}$, then

$$\mathbb{P}\big(\theta^\top v \ge s\big) \le e^{-ns^2/2} \qquad \text{for all } s \ge 0 \tag{112}$$

(for $s \ge 1$, the left-hand side is zero). Let $\theta_\mu := Z_\mu/\|Z_\mu\|$ on $\{Z_\mu \neq 0\}$, let v_ρ be the coordinate vector of $\mathbb{R}^{p-1}$ indexed by $\rho\neq\mu$, and set $s := t/\sqrt{r(p-1)}$. Then

$$\begin{aligned}
\mathbb{P}\Big(\max_{\rho\neq\mu} Z_{\rho\mu} \ge t,\ \|Z_\mu\|^2 \le r(p-1)\Big) &\overset{\text{(i)}}{\le} \mathbb{P}\Big(Z_\mu \neq 0,\ \exists \rho\neq\mu:\ \theta_\mu^\top v_\rho \ge s\Big) \\
&\overset{\text{(ii)}}{\le} \sum_{\rho\neq\mu} \mathbb{P}\big(\theta_\mu^\top v_\rho \ge s \,\big|\, Z_\mu \neq 0\big) \\
&\overset{\text{(iii)}}{\le} (p-1)\, e^{-(p-1)s^2/2} = (p-1)\, e^{-t^2/(2r)}.
\end{aligned}$$

(i) If $Z_{\rho\mu} \ge t > 0$ and $\|Z_\mu\| \le \sqrt{r(p-1)}$, then $Z_\mu \neq 0$ and $\theta_\mu^\top v_\rho = Z_{\rho\mu}/\|Z_\mu\| \ge s$. (ii) Union bound over $\rho\neq\mu$. (iii) By Step 1, given $Z_\mu \neq 0$, θ_μ is uniform on $\mathbb{S}^{p-2}$; apply (112) with $n = p-1$.

3. Proof of (111)**.** Since $\frac{\log p}{d}\|Te_\mu\|^2 = \frac{p\log p}{d^2}\cdot\frac{d}{p}\|Te_\mu\|^2$, (100) and $p\log p/d^2 \to \alpha$ give

$$\max_{\mu\le p}\left|\frac{\log p}{d}\|Te_\mu\|^2 - \alpha E_\kappa\right| \longrightarrow 0 \qquad \text{in probability.} \tag{113}$$

It therefore suffices to show that

$$\max_{\mu\le p}\left|\frac{\log p}{p-1}\|Z_\mu\|^2 - \frac{\log p}{d}\|Te_\mu\|^2\right| \longrightarrow 0 \qquad \text{in probability.}$$

In the decoupled problem, by Step 1, conditionally on (E, U),

$$\|Z_\mu\|^2 = \sum_{\rho \neq \mu} Z_{\rho\mu}^2$$

is a sum of $p-1$ independent squares of $\mathcal{N}\big(0, \frac{\|Te_\mu\|^2}{d}\big)$ variables. Bernstein's inequality [Vershynin, 2018, Theorem 2.8.1] and a union bound over μ give, for every fixed $\delta \in (0, 1]$, with probability at least $1 - 2pe^{-c(p-1)\delta^2}$,

$$\left| \frac{\log p}{p-1} \|Z_\mu\|^2 - \frac{\log p}{d} \|Te_\mu\|^2 \right| \leq \delta \frac{\log p}{d} \|Te_\mu\|^2 \qquad \text{for every } \mu \leq p.$$

Since δ is arbitrary, (113) gives (111).

In the original problem, $U^{(\mu)}(U^{(\mu)})^\top = UU^\top - u_\mu u_\mu^\top = \frac{p}{d} S_U - u_\mu u_\mu^\top$, so

$$\begin{aligned} &\left| \frac{\log p}{p-1} \|Z_\mu\|^2 - \frac{\log p}{d} \|Te_\mu\|^2 \right| \\ &\quad \leq \left(\frac{p}{p-1} \|S_U - \mathbb{I}_d\|_{\mathrm{op}} + \frac{1}{p-1} \right) \frac{\log p}{d} \|Te_\mu\|^2 + \frac{\log p}{p-1} \big(u_\mu^\top T e_\mu\big)^2. \end{aligned}$$

The first term tends to zero uniformly in μ by (94) and (113). By (101), the second is at most $A_0^2 \log p/(p-1) \to 0$ with probability tending to one. This proves (111). □

D.3 Achievability for both problems

Theorem 3 (Rank-constrained achievability)**.** *If*

$$\alpha < \frac{E_\kappa}{2}, \tag{114}$$

then, for both $\mathcal{P} \in \{\mathrm{OP}, \mathrm{DP}\}$*,*

$$\mathbb{P}\big(\mathrm{PR}_{p,d,m}^{\mathcal{P}}\big) \longrightarrow 1.$$

Proof. Set

$$d_1 = \left\lfloor \frac{d}{(\log p)^{1/8}} \right\rfloor, \qquad d_0 = d - d_1, \qquad m_0 = m - d_1. \tag{115}$$

For all sufficiently large p, $d_1 \leq m$, and $m_0/d_0 \to \kappa$ and $p \log p/d_0^2 \to \alpha$. We reserve the first d_0 coordinates for the standard truncated-SVD construction and the remaining d_1 coordinates for a correction.

Let $P_a : \mathbb{R}^d \to \mathbb{R}^{d_a}$ denote the coordinate projection onto block $a \in \{0, 1\}$. For every input, target, and competitor, define the normalized blocks

$$e_\mu^a \coloneqq \sqrt{\frac{d}{d_a}} P_a e_\mu, \qquad u_\mu^a \coloneqq \sqrt{\frac{d}{d_a}} P_a u_\mu, \qquad u_\rho^{(\mu),a} \coloneqq \sqrt{\frac{d}{d_a}} P_a u_\rho^{(\mu)}.$$

Thus the two blocks are independent, and each block has the standard $\mathcal{N}(0, \mathbb{I}_{d_a}/d_a)$ normalization. Write

$$E_0 = [e_1^0, \ldots, e_p^0], \qquad U_0 = [u_1^0, \ldots, u_p^0],$$

and take the standard construction on the large block to be

$$W_0 \coloneqq T_{m_0}(U_0 E_0^\top) = T_{m_0}\left(\sum_{\mu=1}^p u_\mu^0 (e_\mu^0)^\top \right).$$

For each query μ, its margin supplied by this base block is

$$M_\mu^{(0)} \coloneqq (u_\mu^0)^\top W_0 e_\mu^0 - \max_{\rho \neq \mu} (u_\rho^{(\mu),0})^\top W_0 e_\mu^0. \tag{116}$$

Fix a margin level $\gamma > 0$, to be chosen later, and define the set of queries not safely handled by the base construction by

$$\mathcal{B} \coloneqq \{\mu \in [p] : M_\mu^{(0)} \leq \gamma\}. \tag{117}$$

Bounding the size of the bad set.

For $\mu \in [p]$, write

$$S_\mu^{(0)} \coloneqq (u_\mu^0)^\top W_0 e_\mu^0, \qquad Z_{\rho\mu}^{(0)} \coloneqq (u_\rho^{(\mu),0})^\top W_0 e_\mu^0 \quad (\rho \neq \mu), \qquad Z_\mu^{(0)} \coloneqq \big(Z_{\rho\mu}^{(0)}\big)_{\rho \neq \mu}.$$

Thus $M_\mu^{(0)} = S_\mu^{(0)} - \max_{\rho\neq\mu} Z_{\rho\mu}^{(0)}$. These are the target and competitor scores of Lemmas 2 and 3 for the base block, which satisfies the assumptions of both lemmas by (115). Choose $\gamma \in (0, E_\kappa/2)$ sufficiently small that

$$\lambda \coloneqq \frac{(E_\kappa - 2\gamma)^2}{2\alpha E_\kappa} > 1; \tag{118}$$

this is possible by (114). Then

$$\mathbb{P}(\mu \in \mathcal{B}) \le \mathbb{P}\big(S_\mu^{(0)} \le E_\kappa - \gamma\big) + \mathbb{P}\left(\max_{\rho\neq\mu} Z_{\rho\mu}^{(0)} \ge E_\kappa - 2\gamma\right). \tag{119}$$

The quantitative bound (99) of Lemma 2 gives, uniformly in μ,

$$\mathbb{P}\big(S_\mu^{(0)} \le E_\kappa - \gamma\big) = O_\gamma\left(\frac{1}{\log p}\right). \tag{120}$$

For the competitors, fix $\lambda' \in (1, \lambda)$, put $r_\star \coloneqq (E_\kappa - 2\gamma)^2/(2\lambda')$, and define

$$\mathcal{H} \coloneqq \left\{\max_{\mu\le p} \frac{\log p}{p-1} \left\|Z_\mu^{(0)}\right\|^2 \le r_\star\right\}. \tag{121}$$

Since $\lambda' < \lambda$, we have $r_\star > \alpha E_\kappa$, so (111) gives

$$\mathbb{P}(\mathcal{H}) \longrightarrow 1. \tag{122}$$

On $\mathcal{H}$, $\left\|Z_\mu^{(0)}\right\|^2 \le r_\star(p-1)/\log p$ for every μ. Hence (110), applied with $t = E_\kappa - 2\gamma$ and $r = r_\star/\log p$, gives

$$\mathbb{P}\left(\max_{\rho\neq\mu} Z_{\rho\mu}^{(0)} \ge E_\kappa - 2\gamma,\ \mathcal{H}\right) \le (p-1)\exp\left(-\frac{(E_\kappa - 2\gamma)^2 \log p}{2r_\star}\right) \le p^{1-\lambda'}. \tag{123}$$

Combining (119), (120), and (123) gives, uniformly in μ,

$$\mathbb{P}(\mu \in \mathcal{B},\ \mathcal{H}) \le O\left(\frac{1}{\log p}\right) + p^{1-\lambda'}. \tag{124}$$

Summing over μ and applying Markov's inequality,

$$\begin{aligned}\mathbb{P}\left(|\mathcal{B}| > \frac{p}{\sqrt{\log p}},\ \mathcal{H}\right) &\le \frac{\sqrt{\log p}}{p} \sum_{\mu\le p} \mathbb{P}(\mu \in \mathcal{B},\ \mathcal{H}) \\ &\le O\Big((\log p)^{-1/2}\Big) + \sqrt{\log p}\, p^{1-\lambda'} = o(1).\end{aligned} \tag{125}$$

Together with (122), this proves

$$|\mathcal{B}| \le \frac{p}{\sqrt{\log p}} \qquad \text{with probability tending to one.} \tag{126}$$

Uniform control of the base margins.

Put $B_0 \coloneqq 3\sqrt{r_\star}$.[9] The uniform target-score bound (101) of Lemma 2 gives

$$\mathbb{P}\left(\min_{\mu\le p} S_\mu^{(0)} < -A_0\right) \longrightarrow 0. \tag{127}$$

[9]The union bound below ranges over fewer than p^2 scores and vanishes for any fixed $B_0 > 2\sqrt{r_\star}$. The factor 3 gives the exponent $2 - 9/2 = -5/2$.

Moreover, (110) with $t = B_0$ and $r = r_\star/\log p$, together with a union bound over μ, gives

$$\mathbb{P}\left(\max_{\substack{\mu\le p\\ \rho\neq\mu}} Z^{(0)}_{\rho\mu} > B_0,\ \mathcal{H}\right) \le p(p-1)\exp\left(-\frac{B_0^2\log p}{2r_\star}\right)$$
$$\le p^{-5/2} = o(1). \tag{128}$$

Since $\mathbb{P}(\mathcal{H}^c)\to 0$, the competitor bound also holds without intersecting with $\mathcal{H}$. Together with (126) and (127), it follows that the base-block event

$$\mathcal{G} \coloneqq \left\{|\mathcal{B}| \le \frac{p}{\sqrt{\log p}}\right\} \cap \left\{\min_{\mu\le p} S^{(0)}_\mu \ge -A_0\right\} \cap \left\{\max_{\substack{\mu\le p\\ \rho\neq\mu}} Z^{(0)}_{\rho\mu} \le B_0\right\} \tag{129}$$

satisfies $\mathbb{P}(\mathcal{G})\to 1$. Notice that $\mathcal{G}$ is measurable with respect to the first coordinate block of the data (the inputs, the targets and, in the decoupled problem, the competitors), and is therefore independent of all the variables in the repair block.

Independent sparse repair.

For $\mu\in[p]$, define the nonnegative base-margin deficit

$$w_\mu \coloneqq (2\gamma - M^{(0)}_\mu)\mathbf{1}_{\{\mu\in\mathcal{B}\}}, \qquad W_1 \coloneqq \sum_{\mu=1}^{p} w_\mu u^1_\mu (e^1_\mu)^\top. \tag{130}$$

On $\mathcal{G}$, $M^{(0)}_\mu = S^{(0)}_\mu - \max_{\rho\neq\mu} Z^{(0)}_{\rho\mu} \ge -A_0 - B_0$ for every μ. Set the fixed constant $\beta \coloneqq 2\gamma + A_0 + B_0$. Then

$$\|w\|_\infty \le \beta, \qquad \|w\|_2 \le \beta\sqrt{|\mathcal{B}|} \le \frac{\beta\sqrt{p}}{(\log p)^{1/4}}. \tag{131}$$

Condition on all first-block variables. The weights (130) are then deterministic, and the repair block is independent of them. On $\mathcal{G}$, apply Lemma 7, for the problem at hand, to the repair block, i.e. with $n = d_1$ and with $(e_\nu, u_\nu, u^{(\mu)}_\rho)$ replaced by $(e^1_\nu, u^1_\nu, u^{(\mu),1}_\rho)$, so that $R(w) = W_1$, and with β as defined above. Since the error scale $\Delta_{d_1}(w,\beta)$ is increasing in $\|w\|_2$, the bound (131) yields $C_0\Delta_{d_1}(w,\beta)\le \bar\Delta_p$, where

$$\bar\Delta_p \coloneqq C_0\beta\sqrt{\frac{\log p}{d_1}} + \frac{C_0\beta\sqrt{p\log p}}{d_1(\log p)^{1/4}}$$
$$= o(1). \tag{132}$$

For the last equality, use (115) and $p\log p/d^2\to\alpha$: the final term is $O((\log p)^{-1/8})$, while the first term vanishes more rapidly. Since the failure bound of Lemma 7 does not depend on the deterministic weights or on β, it holds uniformly over the conditioning. Thus, with conditional probability $1-o(1)$ on $\mathcal{G}$,

$$\max_{\mu\le p}\left|(u^1_\mu)^\top W_1 e^1_\mu - w_\mu\right| \le \bar\Delta_p, \tag{133}$$

$$\max_{\substack{\mu\le p\\ \rho\neq\mu}}\left|(u^{(\mu),1}_\rho)^\top W_1 e^1_\mu\right| \le \bar\Delta_p. \tag{134}$$

Embed the two matrices into the original coordinates and compensate for the block normalization by setting

$$W_\star \coloneqq \frac{d}{d_0}P_0^\top W_0 P_0 + \frac{d}{d_1}P_1^\top W_1 P_1. \tag{135}$$

The two summands have orthogonal coordinate supports. Hence

$$\operatorname{rank}(W_\star) \le \operatorname{rank}(W_0) + \operatorname{rank}(W_1) \le m_0 + d_1 = m. \tag{136}$$

Since $P_a x = \sqrt{d_a/d}\, x^a$ for every input, target and competitor x, each block contributes $\frac{d}{d_a}(P_a x)^\top W_a (P_a e_\mu) = (x^a)^\top W_a e_\mu^a$ to the score of x under $W_\star$. Hence, for every μ and $\rho \neq \mu$, we have the exact margin identity

$$\begin{aligned} u_\mu^\top W_\star e_\mu - (u_\rho^{(\mu)})^\top W_\star e_\mu &= (S_\mu^{(0)} - Z_{\rho\mu}^{(0)}) + (u_\mu^1)^\top W_1 e_\mu^1 - (u_\rho^{(\mu),1})^\top W_1 e_\mu^1 \\ &\geq M_\mu^{(0)} + w_\mu - 2\bar{\Delta}_p. \end{aligned} \tag{137}$$

For $\mu \notin \mathcal{B}$, the right-hand side is strictly positive for all large p because $M_\mu^{(0)} > \gamma$ and $w_\mu = 0$. For $\mu \in \mathcal{B}$, it equals $2\gamma - 2\bar{\Delta}_p > 0$. Therefore, on an event of probability tending to one, $W_\star$ has rank at most m and retrieves every association. □

D.4 Sharp rank-constrained impossibility

This subsection concerns the decoupled problem only. We first bound the summed competitor scores simultaneously for all W.

Lemma 4 (Uniform lower bound for the summed competitor scores)**.** *For $W \in \mathbb{R}^{d\times d}$, write*

$$D_E(W) \coloneqq \frac{\sqrt{d}}{p} \sum_{\mu=1}^{p} \|We_\mu\|.$$

With probability tending to one, simultaneously for every $W \in \mathbb{R}^{d\times d}$,

$$\sum_{\mu=1}^{p} \max_{\rho\neq\mu} (u_\rho^{(\mu)})^\top W e_\mu \geq (1 - o(1)) \frac{p\sqrt{2\log p}}{d} D_E(W), \tag{138}$$

where the $o(1)$ is deterministic and uniform in W.

Proof. **1. The computation for a fixed matrix.**

Let $S(W)$ denote the summed competitor score:

$$S(W) \coloneqq \sum_{\mu=1}^{p} \max_{\rho\neq\mu} (u_\rho^{(\mu)})^\top W e_\mu.$$

Both $S(W)$ and $D_E(W)$ are positively homogeneous in W, so it is enough to consider $\|W\|_{\mathrm{F}} = 1$. Fix such a deterministic W and condition on $(e_\mu)_{\mu\leq p}$. The calculation below gives a strong estimate for this fixed W; the only difficulty is to make it simultaneous over all W.

For each query, set

$$\sigma_\mu \coloneqq \frac{\|We_\mu\|}{\sqrt{d}}, \qquad \sum_{\mu=1}^{p} \sigma_\mu = \frac{p}{d} D_E(W). \tag{139}$$

Since $u_\rho^{(\mu)} \sim \mathcal{N}(0, \mathbb{I}_d/d)$ is independent of $(e_\nu)_{\nu\leq p}$, each competitor score has, conditionally on $(e_\nu)_{\nu\leq p}$, mean zero and variance

$$\mathrm{Var}\Big((u_\rho^{(\mu)})^\top W e_\mu \,\Big|\, (e_\nu)_{\nu\leq p} \Big) = \frac{\|We_\mu\|^2}{d} = \sigma_\mu^2. \tag{140}$$

When $\sigma_\mu > 0$, the normalized scores $(u_\rho^{(\mu)})^\top W e_\mu/\sigma_\mu$ are therefore independent standard Gaussians as ρ varies. Define their maximum by

$$Y_\mu(W) \coloneqq \max_{\rho\neq\mu} \frac{(u_\rho^{(\mu)})^\top W e_\mu}{\sigma_\mu}.$$

If $\sigma_\mu = 0$, every competitor score for query μ is zero. For $\sigma_\mu > 0$, since each query has its own competitors, the maxima $Y_\mu(W)$ are also independent across μ; this is where the decoupling enters the impossibility proof. Moreover,

$$\max_{\rho\neq\mu} (u_\rho^{(\mu)})^\top W e_\mu = \sigma_\mu Y_\mu(W) \qquad (\sigma_\mu > 0). \tag{141}$$

Put

$$t_p \coloneqq \left(1 - \frac{4\log\log p}{\log p}\right)\sqrt{2\log p}. \tag{142}$$

For every μ with $\sigma_\mu > 0$, Mills' lower bound $1 - \Phi(t) \geq ct^{-1}e^{-t^2/2}$, valid for $t \geq 1$, yields

$$\begin{aligned}\mathbb{P}(Y_\mu(W) < t_p \,|\, (e_\nu)_{\nu\leq p}) &= \Phi(t_p)^{p-1} \leq \exp\{-(p-1)(1-\Phi(t_p))\} \\ &\leq \exp\{-(\log p)^7\}\end{aligned} \tag{143}$$

for all sufficiently large p.[10] Indeed,

$$p\,e^{-t_p^2/2} = p^{1-(1-4\log\log p/\log p)^2} = (\log p)^{8-o(1)},$$

so that $(p-1)(1-\Phi(t_p)) \geq c(\log p)^{8-o(1)}/\sqrt{2\log p} = (\log p)^{15/2-o(1)}$, which exceeds $(\log p)^7$ for all large p. Only the crude exponent 7 is used below.

The next step explains why we aggregate over queries.

2. The uniformity budget.

Define the unit Frobenius sphere and the net mesh by

$$\mathcal{S} \coloneqq \{W \in \mathbb{R}^{d\times d} : \|W\|_{\mathrm{F}} = 1\}, \qquad \epsilon_p \coloneqq \frac{1}{\sqrt{d}\,(\log p)^{3/2}}. \tag{144}$$

Choose a deterministic Frobenius ϵ_p-net $\mathcal{N}_p$ of $\mathcal{S}$. The Euclidean volumetric bound in dimension d^2 gives

$$|\mathcal{N}_p| \leq \left(\frac{3}{\epsilon_p}\right)^{d^2}, \qquad \log|\mathcal{N}_p| = O(d^2\log d) = O(p(\log p)^2). \tag{145}$$

Thus the single-query failure probability in (143), whose exponent is only $(\log p)^7$, cannot be union-bounded over $\mathcal{N}_p$. We need a failure exponent of order p times a power of $\log p$. This is why we aggregate the maxima across the p queries.

3. Queries with a small maximum have negligible total weight.

Let

$$\mathcal{I}(W) \coloneqq \{\mu : \sigma_\mu > 0,\ Y_\mu(W) < t_p\}.$$

On the event of Lemma 8, equations (175) and (176) give

$$\sum_{\mu=1}^{p}\sigma_\mu = \frac{p}{d}D_E(W) \geq \frac{p}{d\sqrt{2\pi}}, \qquad \sum_{\mu=1}^{p}\sigma_\mu^2 = \frac{1}{d}\sum_{\mu=1}^{p}\|We_\mu\|^2 \leq \frac{2p}{d^2}. \tag{146}$$

Therefore, if the indices in $\mathcal{I}(W)$ carry more than a $1/\log p$ fraction of the total weight, then Cauchy–Schwarz gives

$$\frac{p}{d\sqrt{2\pi}\log p} < \sum_{\mu\in\mathcal{I}(W)}\sigma_\mu \leq \sqrt{|\mathcal{I}(W)|}\left(\sum_{\mu=1}^{p}\sigma_\mu^2\right)^{1/2} \leq \frac{\sqrt{2p\,|\mathcal{I}(W)|}}{d}.$$

Hence

$$|\mathcal{I}(W)| \geq \frac{c_1 p}{(\log p)^2}. \tag{147}$$

For any fixed set I of indices with positive weights, conditional independence and (143) give

$$\mathbb{P}\big(Y_\mu(W) < t_p \text{ for every } \mu \in I \;\big|\; (e_\mu)_{\mu\leq p}\big) \leq \exp\{-(\log p)^7\,|I|\}.$$

For realizations of $(e_\mu)_{\mu\leq p}$ on the event of Lemma 8, a union bound over the at most 2^p possible sets I, together with (147), therefore yields

$$\begin{aligned}\mathbb{P}\left(\sum_{\mu\in\mathcal{I}(W)}\sigma_\mu > \frac{1}{\log p}\sum_{\mu=1}^{p}\sigma_\mu \;\middle|\; (e_\mu)_{\mu\leq p}\right) &\leq 2^p\exp\{-c_1 p(\log p)^5\} \\ &\leq \exp\{-c_2 p(\log p)^5\}.\end{aligned} \tag{148}$$

[10] The first inequality uses $x \leq e^{x-1}$ for $x \in [0,1]$, with $x = \Phi(t_p)$.

Thus, for each fixed W and conditional on such a realization of $(e_\mu)_{\mu\le p}$, with probability at least $1-\exp\{-c_2 p(\log p)^5\}$ over the competitors,

$$\sum_{\mu\in\mathcal{I}(W)}\sigma_\mu \le \frac{1}{\log p}\sum_{\mu=1}^{p}\sigma_\mu = \frac{p}{d\log p}D_E(W).$$

4. Discarding the bad queries.

By definition, for $\mu\in\mathcal{I}(W)$, we cannot use $Y_\mu(W)\ge t_p$. We want to omit these summands from a lower bound on $S(W)$, which is valid only if they are nonnegative. We prove the stronger statement that every competitor maximum is nonnegative, simultaneously for all μ and W. For each fixed μ, the $p-1$ points $(u_\rho^{(\mu)})_{\rho\neq\mu}$ are i.i.d. with a symmetric law and almost surely in general position. Wendel's theorem [Wendel, 1962] computes the probability that they lie in a common half-space through the origin, which almost surely is the event that 0 is not in their convex hull:

$$\mathbb{P}\left(0\notin\operatorname{conv}\{u_\rho^{(\mu)}:\rho\neq\mu\}\right) = 2^{-p+2}\sum_{r=0}^{d-1}\binom{p-2}{r}. \tag{149}$$

Since $\sum_{r<d}\binom{p-2}{r}\le(ep/d)^d$ and $d\log(ep/d)=o(p)$ because $d/p\to 0$, the right-hand side is at most $2^{-p+2}e^{o(p)}\le e^{-cp}$ for all large p. A union bound over μ therefore shows that, with probability tending to one,

$$\max_{\rho\neq\mu}(u_\rho^{(\mu)})^\top x\ge 0 \quad \text{for every } \mu\in[p] \text{ and every } x\in\mathbb{R}^d. \tag{150}$$

Indeed, if zero is a convex combination of the competitors of query μ, their inner products with any x cannot all be negative. Notice that (150) is already simultaneous in x, so it requires no union bound over W.

On (150), we may omit the summands indexed by $\mathcal{I}(W)$. For each remaining index with $\sigma_\mu>0$, (141) is at least $t_p\sigma_\mu$; if $\sigma_\mu=0$, the summand is zero. Together with (148), this gives

$$S(W)\ge t_p\sum_{\mu\notin\mathcal{I}(W)}\sigma_\mu \ge \frac{t_p(1-1/\log p)p}{d}D_E(W). \tag{151}$$

5. The union bound and transfer from the net.

The exponent $p(\log p)^5$ in (148) now exceeds the net budget in (145). Indeed, condition on inputs $(e_\mu)_{\mu\le p}$ in the event of Lemma 8. For each $V\in\mathcal{N}_p$, let $\mathcal{F}_V$ be the event that (150) holds but (151) fails for V. By Step 4, $\mathcal{F}_V$ is contained in the event in (148), so a union bound gives

$$\mathbb{P}\Big(\bigcup_{V\in\mathcal{N}_p}\mathcal{F}_V\;\Big|\;(e_\mu)_{\mu\le p}\Big)\le|\mathcal{N}_p|\;e^{-c_2p(\log p)^5}\le e^{O(p(\log p)^2)-c_2p(\log p)^5},$$

which tends to zero. Since the event of Lemma 8 and (150) have probability tending to one, (151) holds simultaneously for every $V\in\mathcal{N}_p$ with probability tending to one.

It remains to transfer this estimate to all of $\mathcal{S}$. First, the triangle inequality and Lemma 8 give, for any W,V,

$$|D_E(W)-D_E(V)|\le D_E(W-V)\le\sqrt{2}\,\|W-V\|_{\mathrm{F}}\,. \tag{152}$$

For the competitor scores, Gaussian norm concentration and a union bound over fewer than p^2 vectors give

$$\max_{\mu,\rho:\rho\neq\mu}\left\|u_\rho^{(\mu)}\right\|\le 2 \tag{153}$$

with probability tending to one; here we use $\log p = o(d)$. On this event,

$$\begin{aligned}
|S(W) - S(V)| &\le \sum_{\mu=1}^{p} \max_{\rho \neq \mu} \left| (u_\rho^{(\mu)})^\top (W - V) e_\mu \right| \\
&\le 2 \sum_{\mu=1}^{p} \|(W - V) e_\mu\| \\
&= \frac{2p}{\sqrt{d}} D_E(W - V) \\
&\le \frac{2\sqrt{2}p}{\sqrt{d}} \|W - V\|_{\mathrm{F}} . \qquad (154)
\end{aligned}$$

For $W \in \mathcal{S}$, choose $V \in \mathcal{N}_p$ with $\|W - V\|_{\mathrm{F}} \le \epsilon_p$. Combining (151)–(154) gives

$$\begin{aligned}
S(W) &\ge S(V) - \frac{2\sqrt{2}p}{\sqrt{d}} \epsilon_p \\
&\ge \frac{t_p(1 - 1/\log p)p}{d} D_E(V) - \frac{2\sqrt{2}p}{\sqrt{d}} \epsilon_p \\
&\ge \frac{t_p(1 - 1/\log p)p}{d} \big(D_E(W) - \sqrt{2}\epsilon_p \big) - \frac{2\sqrt{2}p}{\sqrt{d}} \epsilon_p .
\end{aligned}$$

By Lemma 8, $D_E(W) \ge 1/\sqrt{2\pi}$ on $\mathcal{S}$. Dividing the last bound by $p\sqrt{\log p}\, D_E(W)/d$ gives

$$\begin{aligned}
\frac{dS(W)}{p\sqrt{\log p}\, D_E(W)} &\ge \frac{t_p(1 - 1/\log p)}{\sqrt{\log p}} \left(1 - \frac{\sqrt{2}\epsilon_p}{D_E(W)} \right) - \frac{2\sqrt{2}\epsilon_p}{D_E(W)} \sqrt{\frac{d}{\log p}} \\
&\ge \sqrt{2} - o(1),
\end{aligned}$$

uniformly for $W \in \mathcal{S}$. Indeed, $t_p/\sqrt{\log p} \to \sqrt{2}$, $\epsilon_p \to 0$, and $\epsilon_p \sqrt{d/\log p} = 1/(\log p)^2 \to 0$; the lower bound on $D_E(W)$ makes both error terms uniform. This proves (138) on $\mathcal{S}$. Positive homogeneity extends it to every nonzero W; the case $W = 0$ is immediate. □

We next bound the summed target score simultaneously for all W of rank at most m.

Lemma 5 (Uniform rank-m bound for $\langle C, W \rangle_{\mathrm{F}}$)**.** *With probability tending to one, simultaneously for every W with $\mathrm{rank}(W) \le m$,*

$$\frac{1}{\sqrt{p}} \sum_{\mu=1}^{p} u_\mu^\top W e_\mu = \frac{\langle C, W \rangle_{\mathrm{F}}}{\sqrt{p}} \le \left(\frac{\|T_m(C)\|_{\mathrm{F}}}{\sqrt{p}} + o(1) \right) D_E(W), \qquad (155)$$

where the $o(1)$ is deterministic and uniform in W.

Proof. Both sides of (155) scale by t under $W \mapsto tW$ for $t > 0$. It is therefore enough to prove the inequality for $\|W\|_{\mathrm{F}} = 1$. Write an ordered singular value decomposition

$$W = \sum_{j=1}^{d} s_j a_j b_j^\top , \qquad s_1 \ge \cdots \ge s_d \ge 0,$$

where $s_j = 0$ for $j > \mathrm{rank}(W)$. Let $k = \lceil (\log p)^6 \rceil$. For all sufficiently large p, $k < m$. Split

$$W_{\le k} := \sum_{j \le k} s_j a_j b_j^\top , \qquad W_{>k} := W - W_{\le k} = \sum_{j > k} s_j a_j b_j^\top .$$

The column spaces of $W_{\le k}$ and $W_{>k}$ are orthogonal. Hence, pointwise in e,

$$\|We\|^2 = \|W_{\le k} e\|^2 + \|W_{>k} e\|^2 , \qquad D_E(W) \ge D_E(W_{>k}). \qquad (156)$$

The largest singular value of $W_{>k}$ is s_{k+1}. Moreover, monotonicity and $\sum_{j=1}^{d} s_j^2 = \|W\|_{\mathrm{F}}^2 = 1$ give $k s_{k+1}^2 \le \sum_{j=1}^{k} s_j^2 \le 1$. Therefore,

$$\|W_{>k}\|_{\mathrm{op}} = s_{k+1} \le k^{-1/2} \le (\log p)^{-3}. \tag{157}$$

By Lemmas 1, 8, and 9, all the estimates below hold simultaneously on an event of probability tending to one. We work on this event. In particular,

$$\|C\|_{\mathrm{op}} \le 3\sqrt{\frac{p}{d}}, \qquad D_E(W) \ge \frac{\|W\|_{\mathrm{F}}}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}.$$

We distinguish two cases according to the size of the tail $W_{>k}$.

Case 1: $\|W_{>k}\|_{\mathrm{F}} \ge (\log p)^{-1}$. By (157),

$$\frac{\|W_{>k}\|_{\mathrm{op}}}{\|W_{>k}\|_{\mathrm{F}}} \le \frac{(\log p)^{-3}}{(\log p)^{-1}} = (\log p)^{-2}.$$

Thus $W_{>k}/\|W_{>k}\|_{\mathrm{F}}$ has Frobenius norm one and operator norm at most $(\log p)^{-2}$. Lemma 9, the homogeneity of D_E, and (156) therefore give

$$D_E(W) \ge D_E(W_{>k}) = \|W_{>k}\|_{\mathrm{F}}\, D_E\!\left(\frac{W_{>k}}{\|W_{>k}\|_{\mathrm{F}}}\right) \ge (1 - o(1))\, \|W_{>k}\|_{\mathrm{F}}\,. \tag{158}$$

We next bound separately the contributions of $W_{\le k}$ and $W_{>k}$. For the head, whose singular values are $s_1, \dots, s_k$, von Neumann's trace inequality, $s_j(C) \le \|C\|_{\mathrm{op}}$, and Cauchy–Schwarz give

$$|\langle C, W_{\le k}\rangle_{\mathrm{F}}| \le \sum_{j=1}^{k} s_j(C) s_j \le \|C\|_{\mathrm{op}} \sqrt{k} \left(\sum_{j=1}^{k} s_j^2\right)^{1/2} \le 3\sqrt{\frac{pk}{d}},$$

where we used $\sum_{j\le k} s_j^2 \le \|W\|_{\mathrm{F}}^2 = 1$ in the last step. Moreover, $\operatorname{rank}(W_{>k}) \le \operatorname{rank}(W) \le m$. Von Neumann's trace inequality, followed by Cauchy–Schwarz, gives

$$\begin{aligned} |\langle C, W_{>k}\rangle_{\mathrm{F}}| &\le \sum_{j=1}^{m} s_j(C) s_j(W_{>k}) \\ &\le \left(\sum_{j=1}^{m} s_j(C)^2\right)^{1/2} \|W_{>k}\|_{\mathrm{F}} = \|T_m(C)\|_{\mathrm{F}}\, \|W_{>k}\|_{\mathrm{F}}\,. \end{aligned}$$

Adding these two estimates and using $W = W_{\le k} + W_{>k}$, we obtain

$$\langle C, W\rangle_{\mathrm{F}} \le |\langle C, W_{\le k}\rangle_{\mathrm{F}}| + |\langle C, W_{>k}\rangle_{\mathrm{F}}| \le 3\sqrt{\frac{pk}{d}} + \|T_m(C)\|_{\mathrm{F}}\, \|W_{>k}\|_{\mathrm{F}}\,. \tag{159}$$

Dividing by (158) gives

$$\begin{aligned} \frac{\langle C, W\rangle_{\mathrm{F}}}{D_E(W)} &\le \frac{3\sqrt{pk/d}}{(1 - o(1))\, \|W_{>k}\|_{\mathrm{F}}} + \frac{\|T_m(C)\|_{\mathrm{F}}}{1 - o(1)} \\ &\le \|T_m(C)\|_{\mathrm{F}} + o(\sqrt{p}). \end{aligned} \tag{160}$$

Indeed, $\|W_{>k}\|_{\mathrm{F}}^{-1} \le \log p$ and $\sqrt{k} = O((\log p)^3)$ make the first term $O(\sqrt{p/d}(\log p)^4) = o(\sqrt{p})$. Moreover,

$$\|T_m(C)\|_{\mathrm{F}} \le \sqrt{m}\, \|C\|_{\mathrm{op}} \le 3\sqrt{p},$$

so replacing its factor $(1-o(1))^{-1}$ by 1 costs only $o(\sqrt{p})$. Here $(\log p)^4 = o(\sqrt{d})$ since $p \log p / d^2 \to \alpha \in (0, \infty)$.

Case 2: $\|W_{>k}\|_{\mathrm{F}} < (\log p)^{-1}$. In this case we do not need the stable-rank estimate for $W_{>k}$. Indeed, (159), $\sqrt{k} = O((\log p)^3)$, and $\|T_m(C)\|_{\mathrm{F}} \le 3\sqrt{p}$ give

$$\langle C, W\rangle_{\mathrm{F}} \le O\!\left(\sqrt{\frac{p}{d}} (\log p)^3\right) + O(\sqrt{p}/\log p) = o(\sqrt{p}).$$

The displayed $o(\sqrt{p})$ is deterministic and uniform over the matrices in this case. The input-frame bound of Lemma 8 and $\|W\|_{\mathrm{F}} = 1$ give $D_E(W) \geq 1/\sqrt{2\pi}$. Dividing by this lower bound yields

$$\frac{\langle C, W\rangle_{\mathrm{F}}}{D_E(W)} \leq \sqrt{2\pi}\, o(\sqrt{p}) = o(\sqrt{p}).$$

Since $\|T_m(C)\|_{\mathrm{F}} \geq 0$, this implies (160) also in Case 2. The errors in both cases are deterministic and uniform in W; after division by $\sqrt{p}$, they give the $o(1)$ in (155). Rescaling W proves the claim; $W = 0$ is immediate. □

Summing the retrieval constraints over the queries gives a necessary condition for feasibility, and the two lemmas show that it fails for every W of rank at most m when $\alpha > E_\kappa/2$.

Theorem 4 (Rank-constrained impossibility)**.** *Suppose that*

$$\alpha > \frac{E_\kappa}{2}. \tag{161}$$

Then

$$\mathbb{P}\big(\mathrm{PR}^{\mathrm{DP}}_{p,d,m}\big) \longrightarrow 0.$$

Proof. Suppose that W is feasible. It must be nonzero because the inequalities in (87) are strict. For each query μ, feasibility gives

$$u_\mu^\top W e_\mu > \max_{\rho \neq \mu} (u_\rho^{(\mu)})^\top W e_\mu.$$

Summing over the queries and using the definition of C yields the deterministic necessary condition

$$\langle C, W\rangle_{\mathrm{F}} > \sum_{\mu=1}^{p} \max_{\rho \neq \mu} (u_\rho^{(\mu)})^\top W e_\mu. \tag{162}$$

By (161), we may fix $\delta > 0$ with $\sqrt{E_\kappa} + \delta < \sqrt{2\alpha} - \delta$. By (98), $\|T_m(C)\|_{\mathrm{F}}/\sqrt{p} \leq \sqrt{E_\kappa} + \delta/2$ with probability tending to one. On this event, Lemma 5 gives, for all large p and simultaneously for every matrix W of rank at most m,

$$\langle C, W\rangle_{\mathrm{F}} \leq \sqrt{p}\left(\sqrt{E_\kappa} + \delta\right) D_E(W). \tag{163}$$

By Lemma 4 and $p \log p/d^2 \to \alpha$, the right-hand side of (162) satisfies, for all large p,

$$\begin{aligned}\sum_{\mu=1}^{p} \max_{\rho \neq \mu} (u_\rho^{(\mu)})^\top W e_\mu &\geq (1 - o(1)) \frac{p\sqrt{2\log p}}{d} D_E(W) \\ &\geq \sqrt{p}\left(\sqrt{2\alpha} - \delta\right) D_E(W).\end{aligned} \tag{164}$$

By Lemma 8, $D_E(W) \geq \|W\|_{\mathrm{F}}/\sqrt{2\pi} > 0$, so the lower bound (164) is strictly larger than the upper bound (163). This contradicts (162). The events used above hold simultaneously with probability tending to one, which proves the theorem. □

Proof of Theorem 1. The first case of (89), together with the statement for (1), is Theorem 3, and the second case is Theorem 4. □

D.5 Technical lemmas

Lemma 6 (Diagonal of a Haar-rotated matrix)**.** *There is a universal constant $C_1 < \infty$ such that the following holds for all $p \geq 3$. Let $H \in \mathbb{R}^{p\times p}$ be symmetric and deterministic, let Q be Haar distributed on $O(p)$, and let q_μ denote the μth column of Q. Then, with probability at least $1 - 4/p$,*

$$\max_{\mu \leq p} \left| q_\mu^\top H q_\mu - \frac{\operatorname{tr} H}{p} \right| \leq C_1 \, \frac{\|H\|_{\mathrm{F}} \sqrt{\log p} + \|H\|_{\mathrm{op}} \log p}{p}. \tag{165}$$

Proof. The constants C, c, c' below are universal. Fix μ. Since q_μ is uniform on $\mathbb{S}^{p-1}$, we may write $q_\mu = g/\|g\|$ with $g \sim \mathcal{N}(0, \mathbb{I}_p)$. Then

$$q_\mu^\top H q_\mu - \frac{\operatorname{tr} H}{p} = \frac{g^\top H g - \operatorname{tr} H}{\|g\|^2} + \frac{\operatorname{tr} H}{p}\,\frac{p - \|g\|^2}{\|g\|^2}.$$

Let $t \geq 1$. The Hanson–Wright inequality [Rudelson and Vershynin, 2013] and Bernstein's inequality [Vershynin, 2018, Theorem 2.8.1] give, except with probability $4e^{-ct}$,

$$\left|g^\top H g - \operatorname{tr} H\right| \leq C\big(\|H\|_{\mathrm{F}}\sqrt{t} + \|H\|_{\mathrm{op}}\, t\big), \qquad \left|\|g\|^2 - p\right| \leq C\big(\sqrt{pt} + t\big).$$

If $t \leq c'p$, with c' small enough that $C(\sqrt{c'} + c') \leq 1/2$, the second bound gives $\|g\|^2 \geq p/2$. Using also $|\operatorname{tr} H| \leq \sqrt{p}\,\|H\|_{\mathrm{F}}$ and $|\operatorname{tr} H| \leq p\,\|H\|_{\mathrm{op}}$, both terms of the decomposition are then at most $2C(\|H\|_{\mathrm{F}}\sqrt{t} + \|H\|_{\mathrm{op}}\, t)/p$. If $t > c'p$, the left-hand side is at most $2\,\|H\|_{\mathrm{op}} < (2/c')\,\|H\|_{\mathrm{op}}\, t/p$. Hence, with $C' \coloneqq \max(4C, 2/c')$, for every $t \geq 1$,

$$\mathbb{P}\left(\left|q_\mu^\top H q_\mu - \frac{\operatorname{tr} H}{p}\right| > C'\,\frac{\|H\|_{\mathrm{F}}\sqrt{t} + \|H\|_{\mathrm{op}}\, t}{p}\right) \leq 4e^{-ct}.$$

Take $t = A\log p$ with $A \coloneqq \max(1, 2/c)$, so that $t \geq 1$ and $\sqrt{t} \leq A\sqrt{\log p}$. A union bound over $\mu \leq p$, which does not require the columns of Q to be independent, gives (165) with $C_1 \coloneqq AC'$, except with probability at most $4p\,e^{-cA\log p} \leq 4/p$. □

Lemma 7 (Uniform control of an independent weighted repair)**.** *Let $n = n_p$ satisfy $\log p = o(n)$. Let the inputs $(e_\nu)_{\nu\in[p]}$, the targets $(u_\nu)_{\nu\in[p]}$ and the competitors $\big(u_\rho^{(\mu)}\big)_{\mu,\rho\in[p],\,\rho\neq\mu}$ be as in either problem of Section D.1, but in dimension n: the inputs and the targets are independent $\mathcal{N}(0, \mathbb{I}_n/n)$ vectors, and the competitors are either independent $\mathcal{N}(0, \mathbb{I}_n/n)$ vectors, independent of them (decoupled problem), or $u_\rho^{(\mu)} = u_\rho$ (original problem). Let $w \in \mathbb{R}^p$ be deterministic, let $\beta \geq \|w\|_\infty$ be a deterministic bound, and put*

$$R(w) \coloneqq \sum_{\nu=1}^{p} w_\nu u_\nu e_\nu^\top.$$

Define the error scale

$$\Delta_n(w, \beta) \coloneqq \underbrace{\beta\sqrt{\frac{\log p}{n}}}_{\text{self-terms } \nu=\mu} + \underbrace{\frac{\|w\|_2\,\sqrt{\log p}}{n}}_{\text{cross-terms } \nu\neq\mu}. \tag{166}$$

There are universal constants $C_0, c_0 > 0$ such that the two bounds

$$\max_{\mu\leq p}\left|u_\mu^\top R(w) e_\mu - w_\mu\right| \leq C_0\Delta_n(w,\beta), \tag{167}$$

$$\max_{\substack{\mu\leq p\\ \rho\neq\mu}}\left|(u_\rho^{(\mu)})^\top R(w) e_\mu\right| \leq C_0\Delta_n(w,\beta) \tag{168}$$

hold jointly with probability at least $1 - C_0 p^{-3} - C_0 p^2 e^{-c_0 n}$.

Proof. The case $w = 0$ is trivial, so assume $\|w\|_\infty > 0$, hence $\beta > 0$. The constants C, c below are universal and may change from line to line.

1. Exceptional norm events.

Bernstein's inequality [Vershynin, 2018, Theorem 2.8.1], applied to the centered sub-exponential variables $g_i^2 - 1$ of a standard Gaussian vector $g \in \mathbb{R}^n$, gives for every $t > 0$

$$\mathbb{P}\left(\left|\frac{\|g\|^2}{n} - 1\right| > C\left(\sqrt{\frac{t}{n}} + \frac{t}{n}\right)\right) \leq 2e^{-t}.$$

Put

$$\delta_n \coloneqq C\left(\sqrt{\frac{\log p}{n}} + \frac{\log p}{n}\right) = o(1),$$

the last equality by $\log p = o(n)$. Taking $t = 4\log p$ and a union bound over the $2p$ vectors $(e_\mu)_{\mu\le p}$, $(u_\mu)_{\mu\le p}$ gives

$$\max_{\mu\le p}\left|\|e_\mu\|^2 - 1\right| \ \vee\ \max_{\mu\le p}\left|\|u_\mu\|^2 - 1\right| \le \delta_n \tag{169}$$

with failure probability at most $4p^{-3}$. In particular, on this event $\|e_\mu\|, \|u_\mu\| \le 2$ for every μ and all large p. Separately, in the decoupled problem, the same tail applied to the fewer than p^2 competitors gives

$$\max_{\substack{\mu\le p\\ \rho\ne\mu}}\left\|u_\rho^{(\mu)}\right\| \le 2 \tag{170}$$

with failure probability at most Cp^2e^{-cn}. In the original problem, the competitors are targets, and (170) holds on (169).

Conditioning on (169) or (170) would destroy the Gaussian conditional laws used below, so we never do so. Instead, each estimate in Steps 2 and 3 is a conditional tail bound evaluated at fixed conditioning vectors of norm at most 2; integrating it therefore controls the probability of the corresponding bad event *intersected with* that norm restriction. The probability that some norm restriction fails is added once, in Step 4.

2. Diagonal scores.

Fix μ and separate the term $\nu = \mu$:

$$u_\mu^\top R(w)e_\mu - w_\mu = \underbrace{w_\mu\big(\|e_\mu\|^2\|u_\mu\|^2 - 1\big)}_{=:D_\mu} + \underbrace{\sum_{\nu\ne\mu} w_\nu (e_\mu^\top e_\nu)(u_\nu^\top u_\mu)}_{=:X_\mu}. \tag{171}$$

On the event (169),

$$\left|\|e_\mu\|^2\|u_\mu\|^2 - 1\right| \le \left|\|e_\mu\|^2 - 1\right|\|u_\mu\|^2 + \left|\|u_\mu\|^2 - 1\right| \le 3\delta_n,$$

so $\max_{\mu\le p}|D_\mu| \le 3\beta\delta_n$, which is the self-term of (166).

For X_μ, condition on (e_μ, u_μ). The pairs $(e_\nu, u_\nu)_{\nu\ne\mu}$ are independent of one another and of the conditioning, so the summands are conditionally independent; they are centered because e_ν and u_ν are independent, whence

$$\mathbb{E}\big[(e_\mu^\top e_\nu)(u_\nu^\top u_\mu) \mid e_\mu, u_\mu\big] = \mathbb{E}\big[e_\mu^\top e_\nu \mid e_\mu\big]\,\mathbb{E}\big[u_\nu^\top u_\mu \mid u_\mu\big] = 0.$$

For conditioning vectors of norm at most 2, the factors $e_\mu^\top e_\nu \sim \mathcal{N}(0, \|e_\mu\|^2/n)$ and $u_\nu^\top u_\mu \sim \mathcal{N}(0, \|u_\mu\|^2/n)$ are independent with sub-gaussian norm at most $C/\sqrt{n}$, so each unweighted product has sub-exponential norm at most C/n [Vershynin, 2018, Lemma 2.7.7]. The weighted Bernstein inequality [Vershynin, 2018, Theorem 2.8.2], whose ℓ^∞ branch we relax using $\|w\|_\infty \le \beta$, therefore gives for every $t > 0$

$$\mathbb{P}\big(|X_\mu| > t \mid e_\mu, u_\mu\big) \le 2\exp\left[-c\min\left\{\frac{n^2t^2}{\|w\|_2^2}, \frac{nt}{\beta}\right\}\right]. \tag{172}$$

Take

$$\tau_p \coloneqq C\left(\frac{\|w\|_2\sqrt{\log p}}{n} + \frac{\beta\log p}{n}\right). \tag{173}$$

The first summand is the cross-term scale in (166); the second is Bernstein's linear-tail scale. Its first summand alone forces $n^2\tau_p^2/\|w\|_2^2 \ge C^2\log p$ and its second alone forces $n\tau_p/\beta \ge C\log p$, so for C large enough the right-hand side of (172) at $t = \tau_p$ is at most $2p^{-5}$. Integrating and summing over the p indices,

$$\mathbb{P}\Big(\max_{\mu\le p}|X_\mu| > \tau_p \text{ and (169) holds}\Big) \le 2p^{-4}.$$

Since $\log p/n = o(1)$, we have $\beta \log p/n \le \beta\sqrt{\log p/n}$ for all large p. Thus $3\beta\delta_n + \tau_p \le C\Delta_n(w,\beta)$, proving (167).

3. Off-diagonal scores.

Fix μ and $\rho \neq \mu$, write $v = u_\rho^{(\mu)}$, and separate $\nu = \mu$ again:

$$v^\top R(w) e_\mu = \underbrace{w_\mu \left\|e_\mu\right\|^2 (u_\mu^\top v)}_{=:D'_{\mu\rho}} + \underbrace{\sum_{\nu\neq\mu} w_\nu (e_\mu^\top e_\nu)(u_\nu^\top v)}_{=:X'_{\mu\rho}}. \tag{174}$$

Condition on (e_μ, v) only, so that u_μ keeps its law $\mathcal{N}(0, \mathbb{I}_n/n)$. For conditioning vectors of norm at most 2, $D'_{\mu\rho}$ is centered Gaussian with standard deviation $|w_\mu| \left\|e_\mu\right\|^2 \|v\| / \sqrt{n} \le 8\beta/\sqrt{n}$, so $\left|D'_{\mu\rho}\right| \le C\beta\sqrt{\log p/n}$ with conditional failure probability at most $2p^{-5}$. This again reproduces the self-term of (166).

In the decoupled problem, the competitors $\left(u_\rho^{(\mu)}\right)$ form a family independent of $(e_\nu)_{\nu\le p}$ and $(u_\nu)_{\nu\le p}$. Under the same conditioning the summands of $X'_{\mu\rho}$ are therefore independent and centered, with sub-exponential norm at most C/n, and the analogue of (172) bounds $\left|X'_{\mu\rho}\right|$ by τ_p with conditional failure probability at most $2p^{-5}$.

In the original problem, $v = u_\rho$ also enters $R(w)$, through the summand $\nu = \rho$ of $X'_{\mu\rho}$, which equals $w_\rho \left\|u_\rho\right\|^2 (e_\mu^\top e_\rho)$. Under the same conditioning, e_ρ keeps its law, so this summand is centered Gaussian with standard deviation at most $8\beta/\sqrt{n}$, and it is bounded by $C\beta\sqrt{\log p/n}$ exactly like $D'_{\mu\rho}$. The remaining summands, $\nu \notin \{\mu, \rho\}$, are conditionally independent and centered, with sub-exponential norm at most C/n, and are bounded by τ_p as in the decoupled problem.

In both problems, integrating and summing over the fewer than p^2 pairs (μ, ρ) costs at most $6p^{-3}$. Since $2C\beta\sqrt{\log p/n} + \tau_p \le C\Delta_n(w,\beta)$, this proves (168).

4. Total failure probability.

Adding the four contributions—$4p^{-3}$ from (169), Cp^2e^{-cn} from (170), $2p^{-4}$ from Step 2 and $6p^{-3}$ from Step 3—gives the bound stated in the lemma. The constants in this probability estimate do not depend on the deterministic weights w or their bound β. Hence the estimate also applies after conditioning on weights chosen independently of the Gaussian family above. □

Lemma 8 (Uniform Gaussian input-frame bounds)**.** *With probability tending to one, simultaneously for every $W \in \mathbb{R}^{d\times d}$,*

$$\frac{\|W\|_{\mathrm{F}}}{\sqrt{2\pi}} \le D_E(W) \le \sqrt{2}\,\|W\|_{\mathrm{F}}, \tag{175}$$

$$\frac{d}{p}\sum_{\mu=1}^{p} \|We_\mu\|^2 \le 2\,\|W\|_{\mathrm{F}}^2. \tag{176}$$

Recall that

$$D_E(W) = \frac{\sqrt{d}}{p}\sum_{\mu=1}^{p} \|We_\mu\|.$$

Proof. For fixed $v \in \mathbb{S}^{d-1}$ and $e \sim \mathcal{N}(0, \mathbb{I}_d/d)$, $\sqrt{d}\, v^\top e \sim \mathcal{N}(0,1)$. Set

$$F(v) \coloneqq \frac{\sqrt{d}}{p}\sum_{\mu=1}^{p} \left|v^\top e_\mu\right|, \qquad a \coloneqq \sqrt{d}\,\mathbb{E}_e\left|v^\top e\right| = \sqrt{2/\pi}.$$

For this fixed v, $F(v)$ is $p^{-1/2}$-Lipschitz as a function of the standard Gaussian vectors $(\sqrt{d}\, e_\mu)_{\mu\le p}$. Gaussian concentration [Vershynin, 2018, Theorem 5.2.2] therefore gives

$$\mathbb{P}\left(|F(v) - a| > \frac{a}{4}\right) \le 2e^{-cp}. \tag{177}$$

Let $\mathcal{N}$ be a fixed ε_0-net of $\mathbb{S}^{d-1}$, where $\varepsilon_0 = 1/(8\sqrt{\pi}) = a/(8\sqrt{2})$. The volumetric net bound [Vershynin, 2018, Corollary 4.2.13] gives $|\mathcal{N}| \le (3/\varepsilon_0)^d$. A union bound in (177) gives

$$\mathbb{P}\big(F(w) < 3a/4 \text{ for some } w \in \mathcal{N}\big) \le 2(3/\varepsilon_0)^d e^{-cp} = o(1),$$

where the last equality uses $p/d \to \infty$. Thus, with probability tending to one, $F(w) \ge 3a/4$ for every net point w.

Recall $S_E = (d/p)\sum_{\mu=1}^p e_\mu e_\mu^\top$ from (92). By (94), $\|S_E\|_{\mathrm{op}} \le 2$ for all large p, except with probability $2e^{-d}$. On this covariance event, the reverse triangle inequality and Cauchy–Schwarz give, for any $v, w \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
|F(v) - F(w)| &\le \frac{\sqrt{d}}{p}\sum_{\mu=1}^p \big|(v-w)^\top e_\mu\big| \\
&\le \left(\frac{d}{p}\sum_{\mu=1}^p ((v-w)^\top e_\mu)^2\right)^{1/2} \\
&= \sqrt{(v-w)^\top S_E (v-w)} \\
&\le \sqrt{\|S_E\|_{\mathrm{op}}}\,\|v-w\| \le \sqrt{2}\,\|v-w\|\,.
\end{aligned}$$

On the intersection of the two high-probability events, take any $v \in \mathbb{S}^{d-1}$ and choose $w \in \mathcal{N}$ with $\|v-w\| \le \varepsilon_0$. Then

$$F(v) \ge F(w) - \sqrt{2}\varepsilon_0 \ge \frac{3a}{4} - \frac{a}{8} = \frac{5a}{8} \ge \frac{a}{2} = \frac{1}{\sqrt{2\pi}}.$$

The two bounds just proved hold on the same event, with probability tending to one.

It remains to derive the matrix consequences. Write

$$W^\top W = \sum_{r=1}^d s_r^2 v_r v_r^\top,$$

where $(v_r)_{r\le d}$ is an orthonormal basis of right singular vectors. If $W \neq 0$, the numbers $\theta_r = s_r^2/\|W\|_{\mathrm{F}}^2$ are nonnegative and sum to one. The root-mean-square–arithmetic-mean inequality gives

$$\sqrt{d}\,\|We_\mu\| = \|W\|_{\mathrm{F}}\left(\sum_r \theta_r d (v_r^\top e_\mu)^2\right)^{1/2} \ge \|W\|_{\mathrm{F}} \sum_r \theta_r \sqrt{d}\,\big|v_r^\top e_\mu\big|\,.$$

Averaging over μ and using the uniform bound on F gives $D_E(W) \ge \|W\|_{\mathrm{F}} \sum_r \theta_r F(v_r) \ge \|W\|_{\mathrm{F}}/\sqrt{2\pi}$, the first inequality in (175). Moreover,

$$\frac{d}{p}\sum_\mu \|We_\mu\|^2 = \mathrm{tr}(W^\top W S_E) \le \|S_E\|_{\mathrm{op}}\,\|W\|_{\mathrm{F}}^2 \le 2\,\|W\|_{\mathrm{F}}^2\,.$$

This is (176). Cauchy–Schwarz now gives

$$D_E(W) = \frac{1}{p}\sum_{\mu=1}^p \sqrt{d}\,\|We_\mu\| \le \left(\frac{d}{p}\sum_{\mu=1}^p \|We_\mu\|^2\right)^{1/2} \le \sqrt{2}\,\|W\|_{\mathrm{F}}\,,$$

which is the upper bound in (175). The case $W = 0$ is immediate. □

Lemma 9 (Uniform estimate for D_E at large stable rank)**.** *With D_E as in Lemma 4, with probability tending to one,*

$$\sup_{\substack{W\in\mathbb{R}^{d\times d}:\ \|W\|_{\mathrm{F}}=1 \\ \|W\|_{\mathrm{op}} \le (\log p)^{-2}}} |D_E(W) - 1| = o(1), \tag{178}$$

where the $o(1)$ is deterministic and uniform in W.

Proof. Fix a W in the displayed class and let $e \sim \mathcal{N}(0, \mathbb{I}_d/d)$. For $X = d\,\|We\|^2$,

$$\mathbb{E}X = 1, \qquad \mathrm{Var}(X) = 2\,\mathrm{tr}((W^\top W)^2) \le 2\,\|W\|_{\mathrm{op}}^2 \le 2(\log p)^{-4}. \tag{179}$$

Jensen's inequality and $|\sqrt{x} - 1| \le |x - 1|$ for $x \ge 0$ therefore give

$$0 \le 1 - \sqrt{d}\,\mathbb{E}\,\|We\| \le \mathbb{E}\,|X - 1| \le \sqrt{2}(\log p)^{-2}. \tag{180}$$

As a function of the p standard Gaussian vectors $(\sqrt{d}\,e_\mu)_{\mu \le p}$, $D_E(W)$ is $\|W\|_{\mathrm{op}}/\sqrt{p}$-Lipschitz. Gaussian concentration [Vershynin, 2018, Theorem 5.2.2] hence gives

$$\mathbb{P}\left(|D_E(W) - \mathbb{E}D_E(W)| > (\log p)^{-1/2}\right) \le 2\exp\{-cp(\log p)^3\}. \tag{181}$$

Choose a maximal Frobenius $(\log p)^{-1/2}/(4\sqrt{2})$-separated subset of the class. It is a $(\log p)^{-1/2}/(4\sqrt{2})$-net of the class by maximality, and its elements belong to the class, so (181) applies to each of them. The Euclidean volumetric bound in dimension d^2 gives it cardinality at most

$$\left(C\sqrt{\log p}\right)^{d^2} = \exp\{O(d^2 \log\log p)\}.$$

Since $d^2 = O(p \log p)$, a union bound in (181) closes. Finally, on the event in (175),

$$|D_E(W) - D_E(W')| \le D_E(W - W') \le \sqrt{2}\,\|W - W'\|_{\mathrm{F}}.$$

This transfers the estimate from the net to the entire class. Together with (180), it gives, with probability tending to one, $|D_E(W) - 1| \le \frac{5}{4}(\log p)^{-1/2} + \sqrt{2}(\log p)^{-2}$ for every W in the class, which proves (178). □

E Heuristic derivation of the capacity of the Hebbian ansatz

Recall that the (normalised) Hebbian ansatz reads:

$$W_{\mathrm{Hebb}} \coloneqq \frac{1}{d}\sum_{\mu=1}^{p} u_\mu e_\mu^\top. \tag{182}$$

Let us consider the normalised scores

$$s_{\mu\rho} \coloneqq \frac{1}{d} u_\mu^\top W_{\mathrm{Hebb}} e_\rho. \tag{183}$$

For any given μ, ρ, one can easily compute that for $p \gg d$:

$$\begin{cases} \mathbb{E}[s_{\mu\rho}] & = \delta_{\mu\rho}, \\ \mathrm{Var}[s_{\mu\rho}] & = \dfrac{p}{d^2}(1 + o(1)) \simeq \dfrac{\alpha}{\log p}. \end{cases} \tag{184}$$

While the central limit theorem ensures that the marginal distribution of any given score is close to a Gaussian distribution in the high-dimensional limit, they are correlated with one another. Nevertheless, a very simple heuristic is to approximate the joint distribution of these scores by independent Gaussians. Interestingly, this heuristic predicts a threshold for the capacity of this ansatz at $\alpha = 1/8$, which seems compatible with numerical simulations presented in Fig. 2. We formalise it in the following theorem.

Theorem 5. *Let $d, p \ge 1$, $\alpha > 0$, and $(s_{\mu\rho})_{\mu,\rho\in[p]}$ be independent Gaussian random variables, with mean and covariance structure given by:*

$$\begin{cases} \mathbb{E}[s_{\mu\rho}] & = \delta_{\mu\rho}, \\ \mathrm{Var}[s_{\mu\rho}] & = \dfrac{\alpha}{\log p}. \end{cases} \tag{185}$$

Then, as $d, p \to \infty$ with $(p \log p)/d^2 \to \alpha$:

$$\lim_{d\to\infty} \mathbb{P}[\forall \mu \in [p],\ s_{\mu\mu} \ge \max_{\rho(\neq\mu)} s_{\mu\rho}] = \begin{cases} 1 & \text{if } \alpha < 1/8, \\ 0 & \text{if } \alpha > 1/8. \end{cases}$$

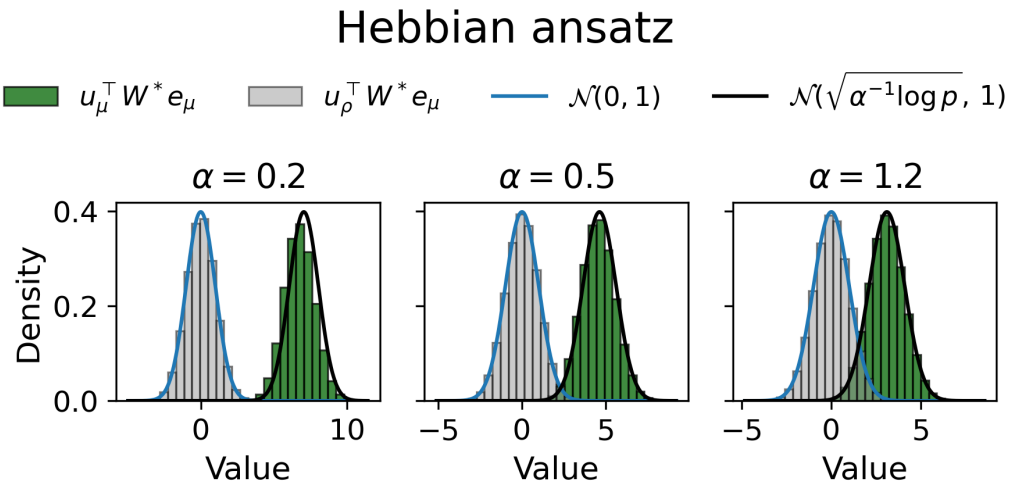


Figure 9: Distribution of the diagonal and off-diagonal scores of eq. (183), for $d = 1000$. We renormalise the scores so that the off-diagonal scores have variance 1, similarly to Fig. 4. We also compare to the Gaussian distributions with theoretical mean and covariance as $d, p \to \infty$.

In Fig. 9, we show the histograms of the (normalised) scores of the Hebbian ansatz, validating the heuristic picture of Theorem 5: the diagonal and off-diagonal scores both are approximately distributed as Gaussians. Both have unit variance (once normalised), and whether memorisation is possible or not is determined by the distance of their means, which decreases as α increases[11]. We emphasise that this mechanism is very distinct from the one of the optimal solution as shown in Fig. 4: the variance of the diagonal scores is there much smaller than the one of the off-diagonal ones, consistently with our theoretical analysis (see Section C.2).

We sketch the proof of Theorem 5 below.

Sketch of proof of Theorem 5. Let $\sigma_p^2 = \alpha/\log p$. Let $Z_\mu := \max_{\rho(\neq\mu)} s_{\mu\rho}$ denote the maximum of the off-diagonal elements in the μ-th row. Let E_μ be the event $\{s_{\mu\mu} \geq Z_\mu\}$. By assumption, the elements $(s_{\mu\rho})$ are independent. Thus, the sets of random variables $\{s_{\mu\rho}\}_{\rho=1}^p$ are mutually independent across different rows μ, which makes the events E_μ mutually independent. Therefore:

$$\mathbb{P}\Big[\forall \mu \in [p],\, s_{\mu\mu} \geq \max_{\rho(\neq\mu)} s_{\mu\rho}\Big] = \mathbb{P}\left[\bigcap_{\mu=1}^{p} E_\mu\right] = \big(1 - \mathbb{P}[E_1^c]\big)^p,$$

where $E_1^c = \{s_{11} < Z_1\}$ is the failure event for the first row. Note that as $p \to \infty$, the limit is entirely determined by the asymptotic behaviour of $p\mathbb{P}[E_1^c]$: more precisely, if $p\mathbb{P}[E_1^c] \to 0$, the probability converges to 1; if $p\mathbb{P}[E_1^c] \to \infty$, the probability converges to 0. We evaluate $\mathbb{P}[E_1^c]$ by conditioning on the diagonal signal s_{11}:

$$\mathbb{P}[E_1^c] = \int_{\mathbb{R}} \mathbb{P}[Z_1 > x] f_{s_{11}}(x)\mathrm{d}x = \int_{\mathbb{R}} \mathbb{P}[Z_1 > x] \frac{e^{-\frac{(x-1)^2}{2\sigma_p^2}}}{\sqrt{2\pi\sigma_p^2}}\mathrm{d}x. \tag{186}$$

with $f_{s_{11}}$ the distribution function of s_{11}. Notice that terms in the integral are of the order $\exp\{\Theta(\log p)\}$ since $\sigma_p^2 = \alpha/\log p$. More precisely, we have for any $x \in \mathbb{R}$ the following elementary lemma.

Lemma 10. *For Z_1 defined as above:*

$$\lim_{p\to\infty} \frac{1}{\log p} \log \mathbb{P}[Z_1 > x] = \min\left(1 - \frac{x^2}{2\alpha}, 0\right). \tag{187}$$

[11] In eq. (184) the mean is always equal to 1 while the variance varies with α, which is another equivalent normalisation choice.

Lemma 10 is shown later on. Using Laplace's method in eq. (186) we get:

$$\lim_{p\to\infty} \frac{\log(p\mathbb{P}[E_1^c])}{\log p} = \sup_{x\in\mathbb{R}} \left[1 + \min\left(1 - \frac{x^2}{2\alpha}, 0\right) - \frac{(x-1)^2}{2\alpha}\right] =: -\inf_{x\in\mathbb{R}} [J_\alpha(x) - 1]. \tag{188}$$

One can easily evaluate the right-hand-side of eq. (188). Notice that the rate function $J_\alpha(x)$ has two regions.

Region 1 ($x \le \sqrt{2\alpha}$): Here, $J_\alpha(x) = (x-1)^2/2\alpha$. Its minimum value is either 0 if $1 \le \sqrt{2\alpha}$, or $(\sqrt{2\alpha}-1)^2/(2\alpha)$ otherwise.

Region 2 ($x > \sqrt{2\alpha}$): Here

$$J_\alpha(x) = \frac{(x-1)^2}{2\alpha} + \frac{x^2}{2\alpha} - 1 = \frac{2x^2 - 2x + 1}{2\alpha} - 1.$$

$J'_\alpha(x) = (4x-2)/(2\alpha)$ has a zero at $x_0 = 1/2$, with $J_\alpha(1/2) = (4\alpha)^{-1} - 1$.

We can then separate cases.

Case 1 ($\alpha < 1/8$): In this regime, $\sqrt{2\alpha} < 1/2$, and $(4\alpha)^{-1} - 1 \le (\sqrt{2\alpha}-1)^2/(2\alpha)$. Then $x_0 = 1/2$ is the global minimum, and $J_\alpha(x_0) = (4\alpha)^{-1} - 1 > 1$. Therefore by eq. (188), we have

$$\lim_{p\to\infty} \frac{\log(p\mathbb{P}[E_1^c])}{\log p} < 0,$$

and thus $p\mathbb{P}[E_1^c] \to 0$, which implies the sought result.

Case 2 ($\alpha > 1/2$): Then the analysis of Region 1 shows that $\inf_x J_\alpha(x) = 0$, which implies that $p\mathbb{P}[E_1^c] \to \infty$, and thus the sought result.

Case 3 ($1/2 > \alpha > 1/8$): Here, $1 > \sqrt{2\alpha} > 1/2$. J_α is here decreasing strictly in $(0, \sqrt{2\alpha})$ (Region 1) and strictly increasing in $(\sqrt{2\alpha}, \infty)$ (Region 2). By continuity, its global minimum is reached in $x = \sqrt{2\alpha}$, with value

$$\frac{(\sqrt{2\alpha}-1)^2}{2\alpha} > 1 \iff 2\alpha - 2\sqrt{2\alpha} + 1 > 2\alpha \iff 1 > 2\sqrt{2\alpha} \iff \alpha < \frac{1}{8}.$$

Therefore the minimum value of J_α is $\inf_x J_\alpha(x) < 1$, and we reach from eq. (188) that $p\mathbb{P}[E_1^c] \to \infty$, ending the proof of Theorem 5.

It simply remains to show Lemma 10. □

Proof of Lemma 10. Let $q_p(x) = \mathbb{P}[s_{1\rho} > x]$ for $\rho \ge 2$. Denoting $\Phi(t) \coloneqq (1/\sqrt{2\pi}) \int_t^\infty e^{-u^2/2} \mathrm{d}u$, we have:

$$q_p(x) = \Phi\left(x\sqrt{\frac{\log p}{\alpha}}\right).$$

For any $t > 0$, the standard Gaussian tail is bounded by:

$$\frac{t}{1+t^2}\frac{1}{\sqrt{2\pi}} e^{-t^2/2} \le \Phi(t) \le \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}.$$

Therefore:

$$\log \Phi(t) = -\frac{t^2}{2} - \log t - \frac{1}{2}\log(2\pi) + O(t^{-2}).$$

Substituting $t = x\sqrt{\log p/\alpha}$, the logarithmic behaviour of the marginal tail is:

$$\log q_p(x) = -\frac{x^2}{2\alpha}\log p - \frac{1}{2}\log(\log p) - \log\left(\frac{x}{\sqrt{\alpha}}\right) - \frac{1}{2}\log(2\pi) + o(1). \tag{189}$$

Because Z_1 is the maximum of $p-1$ independent and identically distributed variables, we have exactly:

$$\mathbb{P}[Z_1 > x] = 1 - \big(1 - q_p(x)\big)^{p-1}. \tag{190}$$

We proceed by establishing matching upper and lower bounds for the limit.

Upper Bound: By the union bound:

$$\mathbb{P}[Z_1 > x] \leq (p-1)q_p(x) < p\, q_p(x).$$

Therefore, using eq. (189):

$$\begin{aligned}\frac{1}{\log p}\log \mathbb{P}[Z_1 > x] &\leq \frac{1}{\log p}\big(\log p + \log q_p(x)\big) \\ &= 1 - \frac{x^2}{2\alpha} - \frac{\log(\log p)}{2\log p} + O\left(\frac{1}{\log p}\right).\end{aligned}$$

Taking the limit as $p \to \infty$, we reach (notice that the limit must be trivially non-positive):

$$\limsup_{p\to\infty} \frac{1}{\log p}\log \mathbb{P}[Z_1 > x] \leq \min\left(1 - \frac{x^2}{2\alpha}, 0\right). \tag{191}$$

Lower Bound: Let us assume first $x \geq \sqrt{2\alpha}$. From eq. (189) we have:

$$(p-1)q_p(x) \asymp p^{1-\frac{x^2}{2\alpha}}(\log p)^{-1/2}.$$

This implies that $(p-1)q_p(x) \to 0$ as $p \to \infty$. For any $y > 0$, standard inequalities give $1 - e^{-y} \geq y - y^2/2$. Using the fact that $(1-q)^{p-1} \leq e^{-(p-1)q}$, we have from eq. (190):

$$\mathbb{P}[Z_1 > x] \geq 1 - e^{-(p-1)q_p(x)} \geq (p-1)q_p(x)\left(1 - \frac{(p-1)q_p(x)}{2}\right).$$

Taking the logarithm:

$$\log \mathbb{P}[Z_1 > x] \geq \log(p-1) + \log q_p(x) + \log\left(1 - \frac{(p-1)q_p(x)}{2}\right).$$

Because $(p-1)q_p(x) \to 0$, the last term converges to 0. Dividing by $\log p$ gives:

$$\frac{1}{\log p}\log \mathbb{P}[Z_1 > x] \geq \frac{\log(p-1)}{\log p} + \frac{\log q_p(x)}{\log p} + o(1) = 1 - \frac{x^2}{2\alpha} + o(1),$$

where we used eq. (189). Taking the limit as $p \to \infty$ yields :

$$\liminf_{p\to\infty} \frac{1}{\log p}\log \mathbb{P}[Z_1 > x] \geq 1 - \frac{x^2}{2\alpha}. \tag{192}$$

It remains to consider the case $x < \sqrt{2\alpha}$. We will show that $\mathbb{P}[Z_1 > x] \to 1$ in this case. Using the standard analytic inequality $1 - y \leq e^{-y}$ for all $y \in \mathbb{R}$, we obtain:

$$1 - \mathbb{P}[Z_1 > x] = \mathbb{P}[Z_1 \leq x] \leq \exp\big(-(p-1)q_p(x)\big). \tag{193}$$

If $x \leq 0$, then $q_p(x) \geq 1/2$. Consequently, $(p-1)q_p(x) \to \infty$ trivially as $p \to \infty$. If $0 < x < \sqrt{2\alpha}$, we exponentiate the precise tail expansion from eq. (189) to find:

$$q_p(x) = \exp\left(-\frac{x^2}{2\alpha}\log p - \frac{1}{2}\log(\log p) + O(1)\right) \asymp \frac{1}{\sqrt{\log p}} p^{-\frac{x^2}{2\alpha}}.$$

Therefore

$$(p-1)q_p(x) \asymp \frac{1}{\sqrt{\log p}} p^{1-\frac{x^2}{2\alpha}} \to \infty,$$

since $1 - x^2/(2\alpha) > 0$. In all cases, this shows from eq. (193) that $\lim_{p\to\infty} \mathbb{P}[Z_1 > x] = 1$, and in particular

$$\liminf_{p\to\infty} \frac{1}{\log p}\log \mathbb{P}[Z_1 > x] = \lim_{p\to\infty} \frac{o(1)}{\log p} = 0.$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions: the decoupled formulation and its conjectured equivalence with the original problem; evidence supporting this conjecture; mechanistic insights into both the Hebbian and the optimal storing mechanisms; and a sharp capacity characterisation for both the full-rank and rank-constrained models in the decoupled framework.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: While we provide strong evidence, we acknowledge that a proof of the equivalence of the two problems remains an open question. The sharp characterisation of the capacity threshold and the asymptotic spectral distributions relies on the replica method from the statistical physics of disordered systems, which is not a rigorous method, as we note in the paper. Moreover, we discuss how our high-dimensional predictions are affected by finite-size effects. Regarding the derivation of the Hebbian threshold, we explicitly state that we treat the non-target scores as independent, although they are in fact correlated.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The complete proof of Theorem 1, with all its assumptions, is given in Section D. Predictions obtained with the non-rigorous replica method are stated as a Result and a Claim (Result 1 and Claim 2), and their full derivations are provided in the appendices (Sections B, C and E). Throughout, we aim to provide intuition for the results and to mention the tools used.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The training protocol is fully described in Section A and is easily reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not provided, and our work is mostly theoretical. However, all experiments are fully described in Section A and are straightforward to reproduce from the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: It is all reported in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In all figures (except histograms), we report error bars from the experiments on the decoupled model."

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We discuss this in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is purely theoretical ML research with no data collection, human subjects, or deployment concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is theoretical research on the storage capacity of linear associative memories. It has no direct path to negative societal applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no models, datasets, or generative systems that could pose misuse risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

Justification: The paper does not use any external datasets, pretrained models, or third-party code assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new datasets or models are released with this paper.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects research is conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology (the replica computations and the resulting predictions) did not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.