SHARP STORAGE CAPACITY OF A SIMPLIFIED MODEL OF LINEAR ASSOCIATIVE MEMORY

Alessio Giorlandino[1] **Sebastian Goldt**[1] **Antoine Maillard**[2]

[1] International School of Advanced Studies (SISSA), Trieste, Italy
[2] INRIA Paris & DI ENS, PSL University, Paris, France
{agiorlan, sgoldt}@sissa.it antoine.maillard@inria.fr

ABSTRACT

Large language models demonstrate remarkable ability in factual recall, but the fundamental limits of memorizing and retrieving many input–output associations remain unclear. We study these limits in a minimal setting: a linear associative memory that must map p input embeddings in $\mathbb{R}^d$ to their corresponding d-dimensional targets via a single linear map, while keeping each mapped input well separated from all other targets. Unlike in supervised classification, this strict separation condition induces p constraints per association and produces correlations between constraints through the shared outputs. Here, we characterise the storage *capacity* $p_c(d)$ of a linear associative memory, i.e. the maximum number of input-output patterns it can store reliably, in the following ways. We introduce a simpler "decoupled" capacity problem in which, for each input, the full set of competing output patterns is independently re-sampled. We find numerically that the original and the decoupled problem exhibit a striking similarity. We then characterise the capacity $p_c(d)$ of this decoupled model using tools from the statistical physics of disordered systems. Our results clarify the fundamental scaling law governing linear associative memories, and provide a starting point for the analysis of more complex models.

1 INTRODUCTION

Large language models have been shown to exhibit a remarkable capacity to memorize factual associations from their training data (Petroni et al., 2019). This observation has motivated a growing body of work aimed at localizing and quantifying memorized facts within such models (Geva et al., 2021; Meng et al., 2022). More broadly, these empirical findings raise fundamental theoretical questions: how many distinct associations can neural networks reliably store, and what are the absolute limits governing factual memory in modern learning systems (Allen-Zhu & Li, 2024; Roberts et al., 2020)?

In particular, we seek to understand, from a fundamental perspective, the limiting capabilities of neural networks to learn and store associations. In a very general way, this reduces to associating 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 $F_W : \mathbb{R}^d \to \mathbb{R}^d$ such that, given e_x, it correctly associates the input to the corresponding output token, in the sense that $\arg\max_{y\in[M]} u_y^\top F_W(e_x) = f^*(x)$. This setting was first introduced by Cabannes et al. (2023), and the simplest instance — which we refer to as *linear associative memory* — is when F_W is a single linear layer, i.e. $F_W(e_x) = We_x$ with $W \in \mathbb{R}^{d\times d}$, which is the case considered by a line of research (Cabannes et al., 2023; 2024; Nichani et al., 2024). In this work, we consider the specific case $N = M = p$, where the number of input and output tokens coincide with the number of associations p, and study the high-dimensional limit of large p and d.

In particular, Cabannes et al. (2023) focuses on the case where the number of output tokens M is fixed and finite, i.e. it does not scale with N. In this setting, they consider input tokens drawn from a Zipf distribution $x \sim \mu_x$ over $[N]$ and a deterministic teacher $f^*(x) = x \mod M$, obtaining

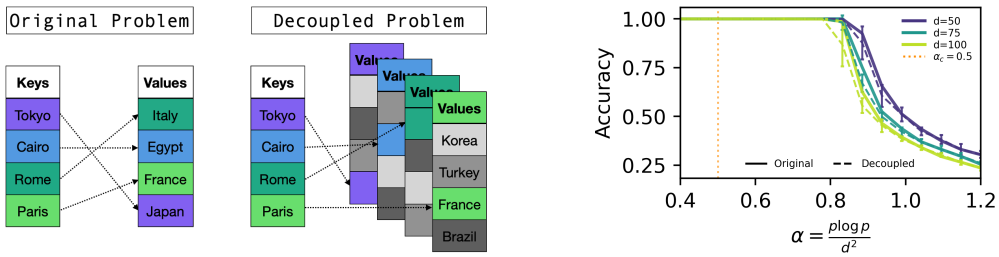

Figure 1: **Left:** The factual recall task consists of memorising associations between inputs (keys) and outputs (values), illustrated here as cities and countries, respectively. In the original problem, the set of outputs is shared across all inputs. In the decoupled problem, each input is associated with its own independent set of competing outputs. **Right:** Empirical accuracy as a function of the load parameter $\alpha = p \log p / d^2$ for the original problem (1) (solid lines) and the decoupled problem (2) (dashed lines), across several embedding dimensions d. Models are trained using Adam on Cross-Entropy loss; details are provided in section A. Convergence to the theoretical storage capacity α_c as $p \to \infty$ is slow, due to finite-size corrections which decay only logarithmically with p.

scaling laws for the generalisation error as a function of the number of training samples and the embedding dimension d. Cabannes et al. (2024) then studies the training dynamics of the same model. They both focus on the regime where N scales with d while $M = O(1)$, whereas the number of parameters of the model is d^2.

Nichani et al. (2024) is the first to also consider a large number of output tokens: in particular, they study the case $N = M = p$ with f^* injective, where both vocabularies scale together. They show that if $d^2 \gtrsim p\,\mathrm{polylog}(p)$, then with high probability over the draw of the embeddings, the *Hebbian ansatz* $W^* = \sum_{z \in [p]} u_{f^*(z)} e_z^\top$ satisfies all the association constraints, while corresponding to a single step of gradient descent on the correlation loss. They also provide numerics for the scaling of the optimal W, which suggest that the right scaling is $d^2 \sim p \log p$.

In this work, we go beyond the Hebbian ansatz, aiming to precisely characterise in the high-dimensional limit the optimal W and provide sharp thresholds on the number of associations that can be stored as a function of the embedding dimension d.

While this problem is reminiscent of the classical storage capacity of the perceptron (Gardner & Derrida, 1988), it differs in that each input must be associated with one among many candidate outputs via a weight matrix rather than a vector, leading to a quadratic rather than linear scaling in d, with an additional logarithmic correction from the competition among an extensive number of outputs.

The main difficulty lies in the correlations between constraints induced by the shared output set, which makes an exact analytical characterization challenging. We therefore introduce a *decoupled* variant in which each input is assigned an independent set of competing outputs, rendering the problem analytically tractable while preserving its essential structure (see figure 1). Within this setting, we sharply characterize the storage capacity α_c in terms of the load parameter $\alpha = p \log p / d^2$, confirming the scaling suggested by Nichani et al. (2024).

2 Storing associations with linear associative memories: original problem and its decoupled variant

In this section, we restate the associative memory problem introduced by Cabannes et al. (2023) for general N inputs and M outputs, specialising to the case $N = M = p$ with injective associations f^* also considered by Nichani et al. (2024), which we term the *original problem*. We then introduce a variant in which all constraints are fully decoupled, which we term the *decoupled problem*, and for which we are able to determine sharply the storage capacity.

Original Problem Let $E = \{e_\mu \in \mathbb{R}^d\}_{\mu\in[p]}$ denote a set of input vectors and $U = \{u_\rho \in \mathbb{R}^d\}_{\rho\in[p]}$ a set of output vectors, where each vector is drawn independently from a multivariate Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$. 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:

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

Without loss of generality, we assume throughout the remainder of the manuscript that the associations are ordered, i.e., $f^*(\mu) = \mu$ for all $\mu \in [p]$.

In addition to the differences in storage capacity discussed above, the original problem also differs from the perceptron in that its constraints are coupled through the shared set of outputs, whereas in the perceptron setting constraints are independent across examples. This coupling motivates the study of a less correlated variant in which each input e_μ is associated with its own independent set of candidate outputs $U^{(\mu)}$, thereby substantially decoupling the constraints.

Decoupled Problem. As in the original problem, let $E = \{e_\mu \in \mathbb{R}^d\}_{\mu\in[p]}$ denote a set of input vectors, with $e_\mu \sim \mathcal{N}(0, \mathbb{I}_d)$ i.i.d. In contrast to the original formulation, for each $\mu \in [p]$ we introduce an independent set of output vectors

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

where all vectors are drawn i.i.d. from $\mathcal{N}(0, \mathbb{I}_d)$ and are independent across different values of μ. Apart from this change, all assumptions, notation, and modelling choices coincide with those of the original problem.

The objective is to 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:

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

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

Interestingly, despite these modeling differences, training linear associative memories on the original and decoupled problems yields very similar empirical performance (see figure 1 and figure 2).

3 Storage capacity of linear associative memories in the decoupled problem

Motivated by the structural similarity between the original formulation and its decoupled counterpart (illustrated in the left panel of figure 1), by the close agreement in empirical accuracy between the two settings (right panel of figure 1), and by the analytical tractability afforded by decoupled constraints, we now turn to the computation of the storage capacity of the linear model in the decoupled problem. In this section, we outline a strategy to characterise the maximal load at which perfect retrieval remains feasible in the decoupled setting, based on a Gardner-type calculation of the volume of the solution space (Gardner & Derrida, 1988).

3.1 The volume of the solution space

The key quantity we need to characterise to derive fundamental limits on the capacity of the linear associative memory is the *volume of the solutions* at fixed realisations of the input set E and of the output sets $U = \{U^{(\mu)}\}_{\mu=1}^p$, where each vector is drawn independently from a d-dimensional Gaussian distribution with zero mean and covariance $\mathbb{I}_d$. The volume of solutions for the decoupled problem is given by

$$\mathcal{V}_{\mathrm{DP}}(E, U) = \int dP(W) \prod_{\mu=1}^{p} \mathbb{1}\Big(\arg\max_{\rho\in[p]} u_\rho^{(\mu)\top} W e_\mu = \mu\Big). \tag{3}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, and $P(W)$ denotes the prior measure on W, which we take to be the i.i.d. Gaussian distribution with zero mean and unit variance over the entries of W. An analogous definition can be introduced for the original problem by replacing the independent output sets with a single shared set of outputs.

Insights on the High-Dimensional Scaling. At a heuristic level, if W were chosen at random, a given association would be satisfied with probability of order $1/p$, since the correct output must be selected among p candidates. Therefore, the probability that it satisfies all constraints is of order $(1/p)^p = e^{-p\log p}$. On the other hand, the volume of admissible configurations scales as $O\big(e^{d^2}\big)$, reflecting the d^2 degrees of freedom of the parameter space. Balancing these two competing effects suggests that the relevant scaling regime is $d^2 \sim p\log p$, which motivates the definition of the load parameter

$$\alpha := \frac{p\log p}{d^2}. \tag{4}$$

A more formal justification of the $\log p$ factor is given later in this section, in particular in lemma 1.

3.2 Gardner Analysis of the Decoupled Problem

A natural first approach is the so-called *annealed* average, which consists of taking the expectation of the volume in equation (3) directly. While this encounters substantial difficulties in the original problem due to correlations between constraints (see section B.1), it can be carried out exactly for the decoupled problem (see section C.1). Since the volume itself is not a concentrating quantity, we instead work with its logarithm, normalised by the number of parameters d^2:

$$\varphi(E,U) = d^{-2}\log \mathcal{V}_{\mathrm{DP}}(E,U). \tag{5}$$

We study this quantity in the high-dimensional limit $d,p \to \infty$ at fixed load $\alpha = p\log p/d^2$, in which $\varphi(E,U)$ concentrates around its expectation—the so-called *quenched* average—and can be expressed, via the replica method, in terms of the overlap matrix between replicated solutions,

$$Q_{ab} = \frac{1}{d^2}\mathrm{Tr}(W^a W^{b\top}), \qquad a,b \in [n].$$

Since the set of solutions is convex, we expect replica symmetry to hold and adopt the replica-symmetric ansatz Q^{RS}, with unit diagonal entries and constant off-diagonal entries q, corresponding to the typical overlap between two distinct solutions.

Under this ansatz, the limiting free entropy admits the variational representation

$$\varphi(\alpha) = \max_{q\in[0,1)} \frac{1}{2}\log(1-q) + \frac{q}{2(1-q)} - \alpha \lim_{n\to 0}\lim_{p\to\infty} \log F_{n,p}(Q^{\mathrm{RS}}), \tag{6}$$

where $F_{n,p}(Q)$ is defined in equation (11) and reduces to the expression given in equation (56) under the replica-symmetric ansatz. Since the problem is convex, we expect that $q \to 1$ as we approach the SAT threshold, as is standard in convex constraint satisfaction problems (see, e.g., Maillard & Kunisky (2024)). Our main result is the following.

Claim 1 (Storage Capacity of a Linear Associative Memory in the Decoupled Problem)**.** *As $q \to 1$, the free entropy admits the expansion*

$$\varphi(\alpha, q) = \frac{1}{2}\log(1-q) + \frac{1}{2(1-q)} - \frac{\alpha}{1-q}. \tag{7}$$

Consequently, the storage capacity in the high-dimensional limit is

$$\alpha_c = \frac{1}{2}. \tag{8}$$

3.3 Insights into the Derivation of the Capacity Threshold

The analysis relies on the replica method (Mézard et al., 1987), a classical tool from the statistical physics of disordered systems, originally developed in the study of spin glasses and later applied

extensively to learning problems. While non-rigorous, the replica approach has provided remarkably accurate predictions for the storage capacity and generalization properties of high-dimensional models, including perceptrons and related associative memory architectures (see, e.g., Gardner & Derrida (1988); Nishimori (2001); Engel & Van den Broeck (2001)). In recent years, replica-based techniques have also been successfully applied to modern machine learning models, offering sharp asymptotic characterisations in a variety of settings (Aubin et al., 2018; Maillard et al., 2024; Barbier et al., 2025).

The derivation of Claim 1 is based on the replica method and is reported in detail in section C.2. We exploit the replica trick to compute

$$\mathbb{E}\log\mathcal{V} = \lim_{n\to 0}\frac{\mathbb{E}\,\mathcal{V}^n - 1}{n}. \tag{9}$$

The order parameter that naturally stems out from the quenched computation is the $n\times n$ overlap matrix between the replicated weights:

$$Q = (Q_{ab})_{a,b\in[n]}, \quad Q_{ab} = \tfrac{1}{d^2}\mathrm{Tr}(W^a W^b) \tag{10}$$

The energetic contribution of a single constraint can be written solely in terms of the overlap matrix:

$$F_{n,p}(Q) = \mathbb{E}_{z_1\sim\mathcal{N}(0,Q)}\left[\Big(\mathbb{P}_{z\sim\mathcal{N}(0,Q)}(z_1 > z)\Big)^{p-1}\right], \tag{11}$$

The volume of the solution, which are determined by the competition of the energetic term involving all the p constraints and the the entropic, can then be written as:

$$\mathbb{E}\,\mathcal{V}_{\mathrm{DP}}^n = \int\prod_{a\le b} dQ_{ab}\int\prod_{a=1}^n dW^a\,P(W^a)\prod_{a\le b}\delta\Big(\tfrac{1}{d^2}\mathrm{Tr}(W^a W^{b\top}) - Q_{ab}\Big)\big(F_{n,p}(Q)\big)^p. \tag{12}$$

We note that, in the decoupled problem, thanks to the independence of the constraints, this term appears independently p times, whereas in the original problem the p constraints are correlated (see section B). To further motivate the scaling of the load parameter α in equation (4), we establish the following lower bound on the constraint-enforcing term.

Lemma 1 (Lower Bound on $F_{n,p}(Q)$)**.** *Let $Q\in\mathbb{R}^{n\times n}$ be a symmetric positive semidefinite matrix Then the following bound holds:*

$$F_{n,p}(Q) \ge e^{-n\log p}. \tag{13}$$

The proof is given in section C.2.1. The bound gives a natural intuition on the origin of the $\log p$ factor in the scaling of the energetic contribution and motivates the choice of the load parameter.

Evaluating this energetic term constitutes the main technical difference with respect to the standard Gardner computation of the perceptron capacity. In particular, in section D we analyse this term in the limit approaching the storage capacity, namely $q\to 1$ and $\varphi(\alpha)\to-\infty$. Combining this energetic contribution with the corresponding entropic terms, we obtain the leading-order action in equation (6). Optimizing this expression shows that $q\to 1$ as $\alpha\to 1/2$, from which we conclude that the critical capacity is $\alpha_c = 1/2$.

4 Concluding perspectives

We studied the storage capacity of a simplified linear associative memory model motivated by factual recall. By introducing a decoupled formulation, we obtained an analytically tractable setting and established a sharp capacity threshold at $\alpha_c = 1/2$, corresponding to the scaling $p\sim d^2/\log p$. This yields a theoretical characterization of linear associative memory capacity in the high-dimensional limit, with implications for building models that reliably store important factual associations.

Looking ahead, we hope to leverage this decoupled formulation as a tractable framework for studying more complex settings, including multi-layer models and structured embeddings, with the broader goal of better understanding beneficial memorisation, as opposed to overfitting, in modern representation learning systems.

REFERENCES

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

Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

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. *arXiv preprint arXiv:2510.24616*, 2025.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. *arXiv preprint arXiv:2310.02984*, 2023.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent. *arXiv preprint arXiv:2402.18724*, 2024.

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

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.

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*, pp. 5484–5495, 2021.

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.

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.

Eshaan Nichani, Jason D Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. *arXiv preprint arXiv:2412.06538*, 2024.

H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. International series of monographs on physics. Oxford University Press, 2001. ISBN 9780198509400. URL https://books.google.fr/books?id=nO0T1VzfhZcC.

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 (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

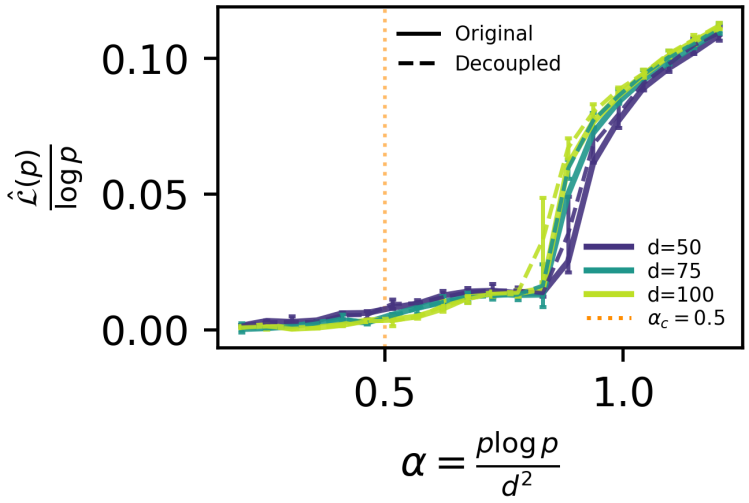

Figure 2: Training loss $\hat{\mathcal{L}}(p)/\log p$ as a function of the load parameter $\alpha = p \log p / d^2$ for both the original (solid lines) and decoupled (dashed lines) problems, across embedding dimensions $d \in \{50, 75, 100\}$. The vertical dotted line indicates the predicted critical load $\alpha_c = 0.5$ in the high-dimensional limit.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

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.

A Numerical Experiments

All inputs E and outputs U (for 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 all association constraints, namely equations (1) and (2). To this end, the weight matrix W is initialized with i.i.d. Gaussian entries of variance d^{-1} and trained using the following objective.

A linear model $f_W(x) = Wx$ is trained on randomly normalized data using the empirical loss

$$\hat{\mathcal{L}}_\gamma(W; E, U) = \frac{1}{p} \sum_{\mu=1}^{p} \frac{1}{\gamma} \log \left(1 + \sum_{\rho(\neq\mu)} \exp\Big(\gamma \, (u_\rho^{(\mu)} - u_\mu^{(\mu)})^\top f_W(e_\mu) \Big) \right). \tag{14}$$

where in the original problem we have $u_\rho^{(\mu)} = u_\rho$ for all $\mu, \rho \in [p]$, i.e., the same output set is shared across inputs, whereas in the decoupled problem the vectors $\{u_\rho^{(\mu)}\}_{\rho \in [p]}$ form an independent set for each μ.

For $\gamma = 1$, this recovers the standard cross-entropy loss.In the limit $\gamma \to \infty$, the empirical loss converges to

$$\hat{\mathcal{L}}_\infty(W; E, U) = \frac{1}{p} \sum_{\mu=1}^{p} \max\Big\{ 0, \max_{\rho(\neq\mu)} (u_\rho - u_\mu)^\top f_W(e_\mu) \Big\}, \tag{15}$$

so that $\hat{\mathcal{L}}_\infty(W; E, U) = 0$ if all association constraints in equations (1) and (2) are satisfied, and is strictly positive otherwise. We remark that in the large-p limit this loss may still vanish even if $O_p(1)$ associations are violated, since the contribution of finitely many errors becomes negligible when averaged over p.

In the experiments, the loss is optimized using Adam with learning rate $\eta = 10^{-3}$ for a maximum of 2048 full-batch optimization steps and $\gamma = 5$. The training loss, rescaled by $\log(p)$, is reported in figure 2.

The results in figure 1 are obtained by following the common training protocol described above for 20 values of the load parameter $\alpha = p \log p / d^2$, evenly spaced between 0.2 and 1.2. For the original problem we report results for $d = 50, 75, 100$.

Conducting experiments at larger dimensions is challenging due to memory constraints arising in the decoupled setting: in this formulation, each input is associated with its own set of p competing outputs 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 as d increases. In contrast, in 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 10 times with independent random draws of the inputs and outputs. Solid lines report the empirical mean across repetitions, while error bars indicate the corresponding standard deviation. For clarity, error bars are displayed only for the original problem.

B Capacity Formulation of the Original Problem

The original problem (OP) is defined in equation (1) and consists of correctly associating each input pattern with its corresponding output. The volume of solutions, for a given prior on the weight matrix W, is given by

$$\mathcal{V}_{\text{OP}}(d, p) = \int dW\, P(W) \prod_{\mu=1}^{p} \mathbb{1}\left(\arg\max_{\rho \in [p]} u_\rho^\top W e_\mu = \mu \right). \tag{16}$$

Or, equivalently,

$$\mathcal{V}_{\text{OP}}(d, p) = \int dW\, P(W) \prod_{\mu=1}^{p} \prod_{\rho \neq \mu} \Theta\left[(u_\mu - u_\rho)^\top W e_\mu \right]. \tag{17}$$

Here $\Theta(\cdot)$ denotes the Heaviside step function.

B.1 Difficulties of the original problem

Already at the annealed level, it is tricky to take the expectation over both the set of inputs and outputs.

$$\mathbb{E}\mathcal{V}_{\text{OP}}(d, p) = \int dW\, P(W)\, \mathbb{E}_U \prod_{\mu=1}^{p} \mathbb{E}_{e_\mu} \prod_{\rho(\neq\mu)} \Theta\left[(u_\mu - u_\rho)^\top W e_\mu \right] \tag{18}$$

Conditionally on W and $U = \{u_\rho\}_{\rho \in [p]}$, the *scores*:

$$u_\rho^\top W e_\mu =: z_\rho^\mu \tag{19}$$

are Gaussian random variables with zero mean and covariance:

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

So we get:

$$\mathbb{E}\mathcal{V}_{\text{OP}}(d, p) = \int dW\, P(W)\, \mathbb{E}_U \prod_{\mu=1}^{p} \left[\mathbb{E}_{z^{(\mu)}|W,U \sim \mathcal{N}(0,\Sigma^U)} \prod_{\rho(\neq\mu)} \Theta\left(z_\mu^\mu - z_\rho^\mu \right) \right]$$

Since orthant probabilities for Gaussian vectors with generic covariance matrices are not available in closed form, the problem must be simplified in order to proceed. A similar difficulty arises, in an even more pronounced form, in the quenched computation, where more complicated orthant probabilities would appear.

C Gardner Analysis of the Decoupled Problem

The decoupled problem (DP) is defined in equation (2). In this appendix section, we provide a complete replica derivation of both the annealed volume, $\mathbb{E}_{E,U}\,\mathcal{V}_{\mathrm{DP}}(d,p;E,U)$, and the quenched volume, $\mathbb{E}_{E,U}\,\log\mathcal{V}_{\mathrm{DP}}(d,p;E,U)$.

The approach closely follows the classical analysis of the storage capacity of the perceptron; see Nishimori (2001); Engel & Van den Broeck (2001) for pedagogical treatments, as well as the original pioneering work Gardner & Derrida (1988). As highlighted throughout the main text, the key difference between the model we consider and the standard perceptron is that each constraint involves an extensive number of conditions.

We state here a classical result, which we will leverage in both commutations to argue the concentration of the covariance of the scores.

Theorem C.1 (Hanson–Wright Inequality (Rudelson & Vershynin, 2013))**.** *Let $X = (X_1,\dots,X_d)\in\mathbb{R}^d$ be a random vector with independent components X_i which satisfy*

$$\mathbb{E}X_i = 0 \quad \textit{and} \quad \|X_i\|_{\psi_2}\le K.$$

Let A be an $n\times n$ matrix, and consider a constant $c>0$. Then, for every $t\ge 0$,

$$\mathbb{P}\Big(\big|X^\top AX - \mathbb{E}\,X^\top AX\big| > t\Big) \le 2\exp\left[-c\,\min\left(\frac{t^2}{K^4\|A\|_{\mathrm{F}}^2}, \frac{t}{K^2\|A\|}\right)\right].$$

C.1 Annealed Computation

The problem doesn't depend on the norm of W, so let's introduce:

$$\widetilde{W} = \frac{W}{\sqrt{d\,\|WW^\top\|}} \tag{20}$$

then, the decoupled problem takes the following form:

$$\mathcal{V}_{\mathrm{DP}}(d,p) = \int dW\,P(W)\,\prod_{\mu=1}^{p}\mathbb{1}\left(\arg\max_{\rho\in[p]} u_\rho^{(\mu)\top}\widetilde{W}e_\mu = \mu\right). \tag{21}$$

The independence between output sets corresponding to different inputs, allows decouple the constraints: while in the original problem we had equation (18), here instead we can bring in the expectation over the outputs set $U^{(\mu)}$:

$$\mathbb{E}\mathcal{V}_{\mathrm{DP}}(d,p) = \int dW\,P(W)\,\prod_{\mu=1}^{p}\left[\mathbb{E}_{U^{(\mu)}}\mathbb{E}_{e_\mu}\prod_{\rho(\neq\mu)}\Theta\left[(u_\mu^{(\mu)}-u_\rho^{(\mu)})^\top\widetilde{W}e_\mu\right]\right]$$

that is, decoupling the constraint, we have p independent problems:

$$\mathbb{E}\mathcal{V}_{\mathrm{DP}}(d,p) = \int dW\,P(W)\,\left[\mathbb{E}_U\mathbb{E}_e\prod_{\rho=2}^{p}\Theta\left[(u_1-u_\rho)^\top\widetilde{W}e\right]\right]^p. \tag{22}$$

Like in the original problem, we consider the *scores*:

$$z_\rho := u_\rho^\top\widetilde{W}e \tag{23}$$

which conditionally on W and e are Gaussian random variables with zero mean and correlation:

$$\mathbb{E}_{U|W,e}[z_\rho z_\sigma] = \delta_{\rho\sigma} \, \|\widetilde{W} e\|^2 =: \delta_{\rho\sigma} \, Q^e.$$

We rewrite the decoupled annealed volume as:

$$\mathbb{E}\mathcal{V}_{\text{DP}}(d,p) = \int dW \, P(W) \, \left[\mathbb{E}_e \mathbb{E}_{z|W,e\sim\mathcal{N}(0,Q^e)} \prod_{\rho=2}^{p} \Theta\left(z_1 - z_\rho\right) \right]^p. \tag{24}$$

The problem now is that the variance Q^e is itself random because it depends on e.

We want to exploit concentration of quadratic forms to argue that Q^e concentrates to its expectation value:

$$Q := \mathbb{E}_{e|W}[Q^e] = \text{Tr}(\widetilde{W}\widetilde{W}^\top) \tag{25}$$

By the Hanson–Wright inequality (see theorem C.1), we have

$$\mathbb{P}\Big(\left| e^\top \widetilde{W}\widetilde{W}^\top e - \text{Tr}(\widetilde{W}\widetilde{W}^\top) \right| > t \Big) \leq 2 \exp\left[-c \min\left(\frac{t^2}{\|\widetilde{W}\widetilde{W}^\top\|_F^2}, \, \frac{t}{\|\widetilde{W}\widetilde{W}^\top\|} \right) \right].$$

We now record two simple observations. By the chosen normalization of $\widetilde{W}$, we have

$$\|\widetilde{W}\widetilde{W}^\top\| = d^{-1}.$$

Moreover,

$$\|\widetilde{W}\widetilde{W}^\top\|_F^2 = \frac{\|WW^\top\|_F^2}{d^2 \, \|WW^\top\|^2}.$$

Substituting these identities into the Hanson–Wright bound yields, independently of the choice of the prior,

$$\mathbb{P}\Big(\left| e^\top \widetilde{W}\widetilde{W}^\top e - Q \right| > t \Big) \leq 2 \exp\left[-c \min\left(d^2 \, t^2 \frac{\|WW^\top\|^2}{\|WW^\top\|_F^2}, \, d\, t \right) \right]. \tag{26}$$

No additional assumptions on W are required, since the inequality

$$\frac{\|WW^\top\|_F^2}{\|WW^\top\|^2} \leq d$$

always holds. So, we conclude:

$$\mathbb{P}(|Q^e - Q| > t) \leq 2e^{-cdt}, \quad c > 0. \tag{27}$$

$$\mathbb{E}\mathcal{V}_{\text{DP}}(d,p) = \int_0^\infty dQ \int dW \, P(W) \, \delta\left(Q - \frac{\text{Tr}(WW^\top)}{d\|WW^\top\|} \right) \left[\int dQ^e P(Q^e) \mathbb{E}_{z\sim\mathcal{N}(0,Q^e\mathbb{I}_p)} \prod_{\rho=2}^{p} \Theta(z_1 - z_\rho) \right]^p. \tag{28}$$

Let us note that

$$\mathbb{E}_{z\sim\mathcal{N}(0,Q^e\mathbb{I}_p)} \left[\prod_{\rho=2}^{p} \Theta(z_1 - z_\rho) \right] = \mathbb{E}_{z\sim\mathcal{N}(0,1)} \left[\Phi(z)^{p-1} \right] =: F(p),$$

which is independent of Q^e. Here Φ denotes the cumulative distribution function of a standard normal random variable.

The function $F(p)$ can be evaluated explicitly by observing that if $z \sim \mathcal{N}(0,1)$, then $\Phi(z) \sim \mathcal{U}(0,1)$, yielding

$$F(p) = \mathbb{E}_{x\sim\mathcal{U}(0,1)}\left[x^{p-1}\right] = \frac{1}{p}. \tag{29}$$

Although the dependence on the empirical covariance Q^e drops out in this annealed computation, it is instructive to compare scales: the energetic contribution decays only polynomially in p, whereas the concentration of Q^e around its expectation occurs at an exponential rate (cf. equation (27)). This separation of scales indicates that fluctuations of Q^e do not affect the leading-order behaviour of the volume in the high-dimensional limit. In the quenched computation this cancellation no longer occurs, but a closely related argument is required to control the convergence of Q^e.

We therefore obtain

$$\mathbb{E}\,\mathcal{V}_{\mathrm{DP}}(d,p) = p^{-p} = e^{-p\log p},$$

which already at the annealed level identifies the correct scaling regime

$$d^2 \sim p\log p.$$

C.2 Quenched Computation: Derivation of Claim 1.

Here we present the full derivation supporting our claim on the optimal storage capacity in section 3.2. The strategy follows the outline discussed in section 3.3. We begin by computing the expected replicated volume of the decoupled problem.

The key advantage of the DP is that the constraints are independent:

$$\mathbb{E}\mathcal{V}^n_{\mathrm{DP}}(d,p) = \int\left(\prod_{a=1}^{n} dW^a\, P(W^a)\right)\prod_{\mu=1}^{p}\left[\mathbb{E}_{U^{(\mu)}}\mathbb{E}_{e_\mu}\prod_{a=1}^{n}\prod_{\rho(\neq\mu)}\Theta\left[(u_\mu^{(\mu)}-u_\rho^{(\mu)})^\top\widetilde{W}^a e_\mu\right]\right].$$

Consequently, thanks to this independence we simply have p decoupled constraints:

$$\mathbb{E}\mathcal{V}^n_{\mathrm{DP}}(d,p) = \int\left(\prod_{a=1}^{n} dW^a\, P(W^a)\right)\left[\mathbb{E}_U\mathbb{E}_e\prod_{a=1}^{n}\prod_{\rho=2}^{p}\Theta\left[(u_1-u_\rho)^\top\widetilde{W}^a e\right]\right]^p. \tag{30}$$

The constraints are independent of the norm of the replicated weight matrices. Therefore, we define:

$$\widetilde{W}^a := \frac{2\,W^a}{\sqrt{d}\,\|W^a\|} \tag{31}$$

Let's proceed similarly to the annealed computation. We introduce the variables:

$$z^a_\rho = u_\rho^\top\widetilde{W}^a e, \tag{32}$$

which conditionally on e and W^a is a Gaussian random variable; but now it's correlated with its replicated counterparts:

$$\mathbb{E}_{U|\{W^a\}_{a\in[n]},e}[z^a_\rho z^b_\sigma] = \delta_{\rho\sigma}e^\top\widetilde{W}^a\widetilde{W}^{b\top}e =: \delta_{\rho\sigma}\widetilde{Q}^e_{ab}. \tag{33}$$

These overlaps are themselves random, with they're expectation over e being:

$$\widetilde{Q}_{ab} := \mathbb{E}_{e|\{W^a\}_{a\in[n]}}[\widetilde{Q}^e_{ab}] = \mathrm{Tr}(\widetilde{W}^a\widetilde{W}^{b\top}). \tag{34}$$

Again, we argue that these random variables concentrate to their expected value.

The Hanson-Wright inequality in this case reads:

$$\mathbb{P}\Big(\left|e^\top\widetilde{W}^a\widetilde{W}^{b\top}e - \widetilde{Q}_{ab}\right| > t\Big) \le 2\exp\left[-c\min\left(\frac{t^2}{\|\widetilde{W}^a\widetilde{W}^{b\top}\|_F^2},\ \frac{t}{\|\widetilde{W}^a\widetilde{W}^{b\top}\|}\right)\right].$$

Now we make two important observations. First of all we can bound the operator norm:

$$\|\widetilde{W}^a\widetilde{W}^{b\top}\| = \frac{4\,\|W^aW^{b\top}\|}{d\|W^a\|\|W^{b\top}\|} \leq 4\,d^{-1} \lesssim d^{-1}.$$

Moreover, we can also bound the Frobenius norm:

$$\|\widetilde{W}^a\widetilde{W}^{b\top}\|_F^2 = \frac{4\,\|W^aW^{b\top}\|_F^2}{d^2\|W^a\|^2\|W^b\|^2} \leq \frac{4d\,\|W^aW^{b\top}\|^2}{d^2\|W^a\|^2\|W^b\|^2} \leq 4d^{-1} \lesssim d^{-1}.$$

Putting all together, we conclude for a single overlap that:

$$\mathbb{P}(|\widetilde{Q}^e_{ab} - \widetilde{Q}_{ab}| > t) \leq 2e^{-cdt}, \quad c > 0. \tag{35}$$

Now, to argue the concentration of the whole overlap matrix we leverage the union bound and we get:

$$\mathbb{P}\left(\max_{1\leq a,b\leq n}\left|\widetilde{Q}^e_{ab} - \widetilde{Q}_{ab}\right| > t\right) \leq 2n^2\,e^{-cdt}, \qquad c > 0. \tag{36}$$

The replicated decoupled problem becomes:

$$\mathbb{E}\mathcal{V}^n_{\text{DP}}(d,p) = \int\left(\prod_{a=1}^n dW^a\,P(W^a)\right)\left[\left(\int\prod_{a\geq b}^n dQ^e\,P(Q^e_{ab})\right)\mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_p\otimes\widetilde{Q}^e)}\prod_{a=1}^n\prod_{\rho=2}^p\Theta\left[z_1^a - z_\rho^a\right]\right]^p. \tag{37}$$

Let's introduce, for a generic overlap matrix Q, the following quantity:

$$F_{n,p}(Q) := \mathbb{E}_{z\sim\mathcal{N}(0,\mathbb{I}_p\otimes Q)}\prod_{a=1}^n\prod_{\rho=2}^p\Theta\left[z_1^a - z_\rho^a\right], \tag{38}$$

or, equivalently, by exploiting the independence over the index ρ, it can be written in the form of equation (11).

We aim to show that this quantity, which involves orthant probabilities of correlated Gaussian random vectors, decays only polynomially in p.

C.2.1 Considerations on the Orthant Probabilities (Proof of lemma 1)

The function $F_{n,p}(Q)$ enforcing the memory constraint of this problem for a given input, can be rewritten in this way:

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

Let's find a lower bound for a generic Q with strictly positive diagonal elements.

$$F_{n,p}(Q) \geq \mathbb{E}_{z_1\sim\mathcal{N}(0,Q)}\left[\mathbb{I}\left\{\forall a\in[n], z_1^a \geq t_a\right\}\left(\mathbb{P}_{z\sim\mathcal{N}(0,Q)}\left\{\forall a\in[n], z^a < z_1^a\right\}\right)^{p-1}\right]$$

for any $t_1,\dots,t_n\in\mathbb{R}$. Trivially this other bound follows :

$$F_{n,p}(Q) \geq \mathbb{E}_{z_1\sim\mathcal{N}(0,Q)}\left[\mathbb{I}\left\{\forall a\in[n], z_1^a \geq t_a\right\}\left(\mathbb{P}_{z\sim\mathcal{N}(0,Q)}\left\{\forall a\in[n], z^a < t_a\right\}\right)^{p-1}\right]$$

Now let's take $t_a = t\sqrt{Q_{aa}}$, for some $t\in\mathbb{R}$. Let $D = \text{diag}(Q)$ and let $\Sigma = D^{-1/2}QD^{1/2}$

$$F_{n,p}(Q) \geq \mathbb{P}_{\tilde{z}_1\sim\mathcal{N}(0,\Sigma)}\left\{\min_{a\in[n]}\tilde{z}_1^a \geq t\right\}\left(\mathbb{P}_{\tilde{z}\sim\mathcal{N}(0,\Sigma)}\left\{\max_{a\in[n]}\tilde{z}^a < t\right\}\right)^{p-1}$$

By Slepian inequality (Vershynin, 2018), we can lower-bound these probabilities by less correlated variables:

$$F_{n,p}(Q) \geq \mathbb{P}_{\tilde{z}_1 \sim \mathcal{N}(0,\mathbb{I}_n)}\Big\{ \min_{a\in[n]} \tilde{z}_1^a \geq t \Big\} \left(\mathbb{P}_{\tilde{z}\sim\mathcal{N}(0,\mathbb{I}_n)}\big\{ \max_{a\in[n]} \tilde{z}^a < t \big\} \right)^{p-1}$$

where we used that $\mathrm{diag}(\Sigma) = \mathbb{I}_n$.

Now, these are just the tails of Gaussian random variables.

Let $\Phi(x)$ be the cumulative distribution function of a standard Gaussian. Then,

$$F_{n,p}(Q) \geq \Big(\Phi(-t)(1-\Phi(-t))^{p-1} \Big)^n$$

In particular, this is valid for any t, then it's valid for the t that maximises this expression. Let $x_t = \Phi(-t) \in [0,1]$, then the lower bound is maximised by $x_t = p^{-1}$. That is, asymptotically:

$$F_{n,p}(Q) \geq e^{-n\log(p)}$$

As a note, this lower bound corresponds to the value that this function take in the annealed calculation (setting $n = 1$), see equation (29).

C.2.2 Further considerations on the overlaps

We now define the overlap matrix as

$$Q_{ab} = d^{-2}\mathrm{Tr}\big(W^a W^{b\top}\big). \tag{39}$$

By definition 31, the corresponding normalised overlap reads

$$\widetilde{Q}_{ab} = \mathrm{Tr}\big(\widetilde{W}^a \widetilde{W}^{b\top}\big) = \frac{4d\, Q_{ab}}{\|W^a\|\,\|W^b\|}.$$

Since the prior on the weight matrices has i.i.d. standard Gaussian entries, the operator norms concentrate around their typical value (Vershynin, 2018), and therefore

$$\widetilde{Q}_{ab} = Q_{ab} + o_d(1),$$

where $o_d(1)$ denotes a term vanishing as $d \to \infty$.

Moreover, having shown that the energetic term decays only polynomially in p, namely that

$$F_{n,p}(\widetilde{Q}^e) \gtrsim e^{-n\log p},$$

we can safely approximate

$$F_{n,p}(\widetilde{Q}^e) \approx F_{n,p}(Q). \tag{40}$$

Indeed, the overlaps $\widetilde{Q}^e$ concentrate exponentially fast in the embedding dimension d (see equation (35)), whereas $p\log p \sim d^2$ in the scaling regime of interest. As a consequence, fluctuations of $\widetilde{Q}^e$ do not affect the leading-order behaviour of the energetic term.

Finally we can write the replicated volume as:

$$\mathbb{E}\mathcal{V}_{\mathrm{DP}}^n(d,p) = \int \left(\prod_{a\geq b}^{n} dQ_{ab} \right) \int \left(\prod_{a=1}^{n} dW^a\, P(W^a) \right) \prod_{a\geq b}^{n} \delta\left(\mathrm{Tr}(W^a W^{b\top}) - d^2 Q_{ab} \right) e^{p\log(F_{n,p}(Q))}. \tag{41}$$

Since the solution space for this problem is convex—that is, the set of configurations of W satisfying all constraints—we expect that, for any number of patterns p for which a solution exists, the solution is replica symmetric. Moreover, note that the problem is purely a vectorial problem.

C.2.3 Replica Symmetric Ansatz

Let's use the Fourier representation of the delta function to rewrite the replicated volume.

$$\mathbb{E}\mathcal{V}_{\mathrm{DP}}^{n}(d,p) = \int \prod_{a\geq b}^{n} dQ_{ab}\, d\hat{Q}_{ab} \int \prod_{a=1}^{n} dW^{a}\, P(W^{a})\, \exp\Big\{ i \sum_{a\geq b}^{n} \hat{Q}_{ab}\big(\mathrm{Tr}(W^{a}W^{b\top}) - d^{2}Q_{ab}\big) + p \log F_{n,p}(Q) \Big\}.$$

Since the constraints are linear in W, the solution set is convex. By log-concavity of the Gibbs measure, this implies that the problem is replica-symmetric (Talagrand, 2010; Barbier et al., 2022). Accordingly, we adopt the replica-symmetric ansatz for Q and $\hat{Q}$:

$$Q_{aa}^{\mathrm{RS}} = Q, \quad Q_{ab}^{\mathrm{RS}} = q, \qquad \hat{Q}_{aa}^{\mathrm{RS}} = i\hat{Q}, \quad \hat{Q}_{ab}^{\mathrm{RS}} = i\hat{q}, \qquad \forall\, a \neq b. \tag{42}$$

Let us define the entropic term

$$J(\hat{Q},\hat{q}) := \log \int \left(\prod_{a=1}^{n} dW^{a}\, P(W^{a}) \right) \exp\left\{ -\hat{Q} \sum_{a=1}^{n} \|W^{a}\|_{F}^{2} - \hat{q} \sum_{a>b}^{n} \mathrm{Tr}\big(W^{a}W^{b\top}\big) \right\}. \tag{43}$$

Then the replicated volume takes the form

$$\mathbb{E}\,\mathcal{V}_{\mathrm{DP}}^{n}(d,p) = \int dQ\, dq\, d\hat{Q}\, d\hat{q}\ \exp\left\{ -i\, d^{2} \sum_{a\geq b}^{n} Q_{ab}^{\mathrm{RS}} \hat{Q}_{ab}^{\mathrm{RS}} + p \log F_{n,p}\big(Q^{\mathrm{RS}}\big) + J(\hat{Q},\hat{q}) \right\}. \tag{44}$$

The term coupling the overlaps with their conjugate variables can be expanded as

$$-i \sum_{a\geq b}^{n} Q_{ab}^{\mathrm{RS}} \hat{Q}_{ab}^{\mathrm{RS}} = n\, Q\hat{Q} + \frac{n(n-1)}{2}\, q\, \hat{q} = n\, Q\hat{Q} - \frac{n}{2}\, q\, \hat{q} + O(n^{2}), \tag{45}$$

so that we can express the RS replicated volume as

$$\mathbb{E}\,\mathcal{V}_{\mathrm{DP}}^{n}(d,p) = \int dQ\, dq\, d\hat{Q}\, d\hat{q}\ \exp\left\{ -i\, d^{2} \left(n\, Q\hat{Q} - \frac{n}{2}\, q\, \hat{q} \right) + p \log F_{n,p}\big(Q^{\mathrm{RS}}\big) + J(\hat{Q},\hat{q}) + O(n^{2}) \right\}. \tag{46}$$

C.2.4 Entropic Term: Gaussian Prior

To evaluate the entropic term, we need to specify a prior on the weight matrices. As a natural choice for full-rank matrices, we consider i.i.d. Gaussian entries:

$$P(W^{a}) = \frac{e^{-\frac{1}{2}\|W^{a}\|_{F}^{2}}}{(2\pi)^{d^{2}/2}}.$$

We can now compute the entropic term $J(\hat{Q},\hat{q})$. The trace decomposes entry-wise, and the integral factorises into d^{2} identical copies of an n-dimensional Gaussian integral:

$$J(\hat{Q},\hat{q}) = \log \int \prod_{a=1}^{n} dW^{a}\, P(W^{a})\, \exp\Big\{ i \sum_{a\geq b}^{n} \hat{Q}_{ab}^{\mathrm{RS}}\, \mathrm{Tr}(W^{a}W^{b\top}) \Big\} = d^{2} \log \int_{\mathbb{R}^{n}} \prod_{a} \frac{dw_{a}}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\boldsymbol{w}^{\top}\boldsymbol{M}\,\boldsymbol{w}} = -\frac{d^{2}}{2} \log\det(\boldsymbol{M}),$$

where $\boldsymbol{w} \in \mathbb{R}^{n}$ and $\boldsymbol{M}_{ab} = \big(1 + 2\hat{Q} - \hat{q}\big)\delta_{ab} + \hat{q}$. The eigenvalues of $\boldsymbol{M}$ are $1 + 2\hat{Q} - \hat{q}$ with multiplicity $n-1$ and $1 + 2\hat{Q} + (n-1)\hat{q}$ with multiplicity 1, giving

$$J(\hat{Q},\hat{q}) = -\frac{d^{2}}{2}\Big[\log\big(1 + 2\hat{Q} + (n-1)\hat{q}\big) + (n-1)\log\big(1 + 2\hat{Q} - \hat{q}\big) \Big].$$

Expanding for $n \to 0$ and keeping only the terms linear in n:

$$J(\hat{Q},\hat{q}) = -\frac{nd^{2}}{2}\left[\log\big(1 + 2\hat{Q} - \hat{q}\big) + \frac{\hat{q}}{1 + 2\hat{Q} - \hat{q}} \right].$$

We are left with determining the energetic term. In section D, we evaluate it at the replica-symmetric ansatz in the limit $q \uparrow Q$, corresponding to the capacity threshold, obtaining

$$\log F_{n,p}(Q^{\mathrm{RS}}) = -n \log(p)\, \frac{q}{Q - q}. \tag{47}$$

We will take this limit explicitly after deriving the saddle-point equations in the high-dimensional regime $d, p \to \infty$.

Saddle point equations The replica-symmetric calculation of the replicated volume of the decoupled problem leads to the expression

$$\mathbb{E}\,\mathcal{V}_{\text{DP}}^{n}(d,p) = \int dQ\, dq\, d\hat{Q}\, d\hat{q}\ \exp\left\{nd^2\Big(Q\hat{Q} - \tfrac{1}{2}q\hat{q}\Big) + \log F_{n,p}(Q^{\text{RS}}) - \frac{nd^2}{2}\left(\log(1+2\hat{Q}-\hat{q}) + \frac{\hat{q}}{1+2\hat{Q}-\hat{q}}\right)\right\}.$$

In the high-dimensional regime $d, p \to \infty$, the integral is dominated by its saddle point, so that by the Laplace method

$$\log \mathbb{E}\,\mathcal{V}_{\text{DP}}^{n}(d,p) \sim \underset{Q,\,q,\,\hat{Q},\,\hat{q}}{\text{extr}} \left\{nd^2\Big(Q\hat{Q} - \tfrac{1}{2}q\hat{q}\Big) + \log F_{n,p}(Q^{\text{RS}}) - \frac{nd^2}{2}\left(\log(1+2\hat{Q}-\hat{q}) + \frac{\hat{q}}{1+2\hat{Q}-\hat{q}}\right)\right\}.$$

In the capacity threshold limit $q \uparrow Q$, the energetic term takes the form $\log F_{n,p}(Q^{\text{RS}}) = -n\log(p)\,\frac{q}{Q-q}$, from which the natural scaling of the load emerges:

$$\alpha := \frac{p\log p}{d^2}. \tag{48}$$

Substituting the expression for the energetic term and writing the problem in terms of the load α, the extremisation becomes

$$\log \mathbb{E}\,\mathcal{V}_{\text{DP}}^{n}(d,\alpha) \sim \underset{Q,\,q,\,\hat{Q},\,\hat{q}}{\text{extr}}\ nd^2\left\{Q\hat{Q} - \tfrac{1}{2}q\hat{q} - \alpha\frac{q}{Q-q} - \frac{1}{2}\left(\log(1+2\hat{Q}-\hat{q}) + \frac{\hat{q}}{1+2\hat{Q}-\hat{q}}\right)\right\}.$$

Extremising over the auxiliary variables $\hat{Q}$ and $\hat{q}$ yields

$$\hat{Q} = \frac{Q-2q}{2(Q-q)^2} - \frac{1}{2}, \quad \hat{q} = -\frac{q}{(Q-q)^2}. \tag{49}$$

Substituting back, the problem reduces to

$$\log \mathbb{E}\,\mathcal{V}_{\text{DP}}^{n}(d,\alpha) \sim \underset{Q,\,q}{\text{extr}}\ nd^2\,\varphi(Q,q,\alpha), \tag{50}$$

where the resulting action is

$$\varphi(Q,q,\alpha) = \frac{1}{2} - \frac{Q}{2} + \frac{1}{2}\log(Q-q) + \frac{q}{2(Q-q)} - \alpha\frac{q}{Q-q}.$$

which only depends on the ratio $t = q/Q$. To see this, we substitute $q = tQ$ to obtain

$$\varphi(Q,t,\alpha) = \frac{1}{2} - \frac{Q}{2} + \frac{1}{2}\log Q + \frac{1}{2}\log(1-t) + \frac{t}{2(1-t)} - \alpha\frac{t}{1-t}.$$

Extremising over Q yields $\partial_Q\varphi = -\frac{1}{2} + \frac{1}{2Q} = 0$, so that $Q = 1$. Setting $Q = 1$ and renaming $t \to q$, the action reduces to

$$\varphi(q,\alpha) = \frac{1}{2}\log(1-q) + \frac{q}{2(1-q)} - \alpha\frac{q}{1-q}. \tag{51}$$

As $q \to 1$ the free entropy diverges. In this limit the action, to leading order, is:

$$\varphi(q,\alpha) = \frac{1}{2}\log(1-q) + \frac{1}{2(1-q)} - \frac{\alpha}{1-q} \tag{52}$$

Now, extremising also with respect to q yields:

$$1 - q = 2\big(\tfrac{1}{2} - \alpha\big),$$

from which we conclude that the critical load is $\alpha_c = \frac{1}{2}$. Finally, taking the replica limit $n \to 0$, the expected log-volume of the solution space is

$$\frac{1}{d^2}\,\mathbb{E}\,\log \mathcal{V}_{\text{DP}}(d,\alpha) = \lim_{n\to 0}\lim_{d\to\infty}\frac{1}{nd^2}\left(\mathbb{E}\,\mathcal{V}_{\text{DP}}^{n} - 1\right) = \varphi(q^\star,\alpha), \tag{53}$$

which is positive for $\alpha < \frac{1}{2}$ and diverges to $-\infty$ as $\alpha \uparrow \frac{1}{2}$, signalling the disappearance of solutions.

D About the Energetic Term.

In this appendix we analyse the energetic term enforcing the constraints. We recall its definition in equation (11). Equivalently, in the quenched calculation it appears in the form given in equation (38):

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

Computing this quantity for a general covariance matrix Q is intractable, as no closed-form expression is available. However, for specific choices of Q it can be evaluated explicitly. In particular, we compute $F_{n,p}(Q)$ at the replica-symmetric ansatz

$$Q_{aa}^{\mathrm{RS}} = Q, \qquad Q_{ab}^{\mathrm{RS}} = q, \qquad \forall\, a \neq b \in [n]. \tag{54}$$

We perform the change of variables

$$Z_\rho^a = z_1^a - z_\rho^a, \qquad \rho = 2, \dots, p.$$

These variables are Gaussian with zero mean and covariance

$$\mathbb{E}\big[Z_\rho^a Z_\sigma^b\big] = Q_{ab}^{\mathrm{RS}}\,(1+\delta_{\rho\sigma}) = \big(q + (Q-q)\delta_{ab}\big)\,(1+\delta_{\rho\sigma}). \tag{55}$$

Such a correlation structure can be realised, for instance, by the representation

$$Z_\rho^a = \sqrt{q}\,(\lambda + \eta_\rho) + \sqrt{Q-q}\,\big(\xi^a + x_\rho^a\big),$$

where $\{\eta_\rho\}_{\rho\geq 2}$, $\{\xi^a\}_{a\in[n]}$, $\{x_\rho^a\}$, and λ, are independent standard Gaussian random variables.

With this parametrisation, taking the expectation with respect to x_ρ^a yields

$$F_{n,p}(Q^{\mathrm{RS}}) = \mathbb{E}_{\lambda,\eta}\left[\left(\mathbb{E}_\xi \prod_{\rho=2}^{p} \Phi\left(\xi + \frac{\sqrt{q}}{\sqrt{Q-q}}(\eta_\rho + \lambda)\right)\right)^n\right], \tag{56}$$

where Φ denotes the standard normal cumulative distribution function, $\eta \sim \mathcal{N}(0, \mathbb{I}_{p-1})$ and we called $t \coloneqq q/Q$ and $\beta_t \coloneqq \sqrt{\frac{t}{1-t}}$.

Since the quenched computation requires taking the limit $n \to 0$, we expand $F_{n,p}(Q^{\mathrm{RS}})$ to first order in n. This yields

$$F_{n,p}(Q^{\mathrm{RS}}) = 1 + n\,\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] + O(n^2), \tag{57}$$

where

$$f(t;\lambda,\eta) := \mathbb{E}_\xi \prod_{\rho=2}^{p} \Phi(\xi + \beta_t(\eta_\rho + \lambda))\,. \tag{58}$$

Computing this expectation analytically is challenging. Our strategy is therefore to derive matching upper and lower bounds and show that they coincide at leading order.

In both cases, a useful quantity is the maximum of the shifted Gaussian variables $\lambda + \eta_\rho$, namely

$$a_p = \max_{\rho\in[p]}(\lambda + \eta_\rho) = \lambda + \max_{\rho\in[p]}\eta_\rho \sim \sqrt{2\log p}, \tag{59}$$

where we used that $p - 1 \simeq p$ for large p.

Lower Bound. 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,

$$\prod_{\rho=1}^{p} \Phi\big(\xi - \beta_t(\lambda + \eta_\rho)\big) \geq \Phi\big(\xi - \beta_t a_p\big)^p. \tag{60}$$

Therefore we obtain the lower bound

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \geq \mathbb{E}_{a_p} \log\big[\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_p\big)^p\big]\,. \tag{61}$$

Let us compute the inner expectation. For any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi - \beta_t a_p\big)^p &\geq \int_{\beta_t a_p + (1-\varepsilon)a_p}^{\beta_t a_p + (1+\varepsilon)a_p} d\xi\, \frac{e^{-\xi^2/2}}{\sqrt{2\pi}}\, \Phi\big(\xi - \beta_t a_p\big)^p \\ &= a_p \int_{1-\varepsilon}^{1+\varepsilon} dx\, \frac{e^{-(\beta_t + x)^2 a_p^2/2}}{\sqrt{2\pi}}\, \Phi(x a_p)^p.\end{aligned} \tag{62}$$

Since a_p is large and $a_p \sim \sqrt{2\log p}$, we may use the approximation

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

which yields

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

Taking the limit $p \to \infty$, we observe that

$$\lim_{p\to\infty} \left(1 - \frac{1}{p^x}\right)^p = \begin{cases} 0, & x < 1, \\ e^{-1}, & x = 1, \\ 1, & x > 1. \end{cases}$$

From this we obtain the lower bound

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

In particular, we are interested in the behaviour as $t \to 1$, where β_t becomes large. Taking logarithms yields

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \gtrsim -\frac{t}{1-t}\, \log p, \tag{63}$$

where we have neglected $O(\log\log p)$ terms.

Upper Bound. A first naive bound follows from

$$\prod_{\rho=1}^{p} \Phi\big(\xi - \beta_t z_\rho\big) \leq \Phi\big(\xi - \beta_t a_p\big),$$

which yields the rough estimate

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \lesssim -\frac{1}{2}\, \frac{t}{1-t},$$

and does not match the lower bound. However, one can do better.

Fix an arbitrary integer k which does not diverge with p. Then

$$\prod_{\rho=1}^{p} \Phi\big(\xi - \beta_t z_\rho\big) \leq \Phi\big(\xi - \beta_t a_{p-k}\big)^k, \tag{64}$$

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

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

This implies the upper bound

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \le \mathbb{E}_{a_{p-k}} \log\Big[\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\Phi\big(\xi-\beta_t a_{p-k}\big)^k\Big]. \tag{65}$$

We now compute the inner expectation explicitly. Observe that

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

Introducing the variables $y_i = \xi - \xi_i - \beta_t a_{p-k}$, the vector $y=(y_1,\dots,y_k)$ is Gaussian with mean

$$\mathbb{E}[y_i] = -\beta_t a_{p-k},$$

and covariance matrix $\Sigma_{ij} = 1+\delta_{ij}$. Therefore,

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

Expanding the quadratic form yields

$$= \frac{e^{-\frac{1}{2}\beta_t^2 a_{p-k}^2 \mathbf{1}^\top\Sigma^{-1}\mathbf{1}}}{\sqrt{\det(2\pi\Sigma)}} \int_{[0,\infty)^k} d^k y\, \exp\Big[-\frac{1}{2}y^\top\Sigma^{-1}y - \beta_t a_{p-k}\mathbf{1}^\top\Sigma^{-1}y\Big].$$

Rescaling $y\mapsto(\beta_t a_{p-k})y$, we obtain

$$= \frac{e^{-\frac{1}{2}\beta_t^2 a_{p-k}^2 \mathbf{1}^\top\Sigma^{-1}\mathbf{1}}}{\beta_t a_{p-k}\sqrt{\det(2\pi\Sigma)}} \int_{[0,\infty)^k} d^k y\, \exp\Big[-\frac{1}{2(\beta_t a_{p-k})^2}y^\top\Sigma^{-1}y - \mathbf{1}^\top\Sigma^{-1}y\Big].$$

In the limit $\beta_t a_{p-k}\to\infty$, the first term in the exponential becomes negligible. Using the Sherman–Morrison formula,

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

so that

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

Therefore,

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

where C_k is a finite constant depending only on k.

Since k is fixed and $a_{p-k}\sim\sqrt{2\log p}$ to leading order, taking logarithms yields

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \lesssim -\frac{k}{k+1}\,\frac{t}{1-t}\,\log p.$$

As this bound holds for any fixed k, we may take k large so that $\frac{k}{k+1}\uparrow 1$, which matches the lower bound.

Conclusion. We conclude that, in the joint limit of large p and large β_t (equivalently $q\to 1$),

$$\mathbb{E}_{\lambda,\eta}\big[\log f(t;\lambda,\eta)\big] \approx -\frac{t}{1-t}\,\log p, \tag{66}$$

up to $O(\log\log p)$ corrections.

Putting everything together, the energetic term in equation (67) becomes

$$F_{n,p}(Q^{\mathrm{RS}}) \approx 1 - n\log p\,\frac{t}{1-t} + O(n^2), \tag{67}$$

and therefore

$$\log F_{n,p}(Q^{\mathrm{RS}}) \approx -n\log p\,\frac{t}{1-t} + O(n^2). \tag{68}$$