Are Neural Networks Collision Resistant?

Marco Benedetti[1], Andrej Bogdanov[2], Enrico M. Malatesta[1], Marc Mézard[1], Gianmarco Perrupato[1], Alon Rosen[1], Nikolaj I. Schwartzbach[1], and Riccardo Zecchina[1]

[1]Department of Computing Sciences, Bocconi University, Milano, Italy
[2]University of Ottawa, Canada

Abstract

When neural networks are trained to classify a dataset, one finds a set of weights from which the network produces a label for each data point. We study the algorithmic complexity of finding a collision in a single-layer neural net, where a collision is defined as two distinct sets of weights that assign the same labels to all data. For binary perceptrons with oscillating activation functions, we establish the emergence of an overlap gap property in the space of collisions. This is a topological property believed to be a barrier to the performance of efficient algorithms. The hardness is supported by numerical experiments using approximate message passing algorithms, for which the algorithms stop working well below the value predicted by our analysis. Neural networks provide a new category of candidate collision resistant functions, which for some parameter setting depart from constructions based on lattices. Beyond relevance to cryptography, our work uncovers new forms of computational hardness emerging in large neural networks which may be of independent interest.

1 Introduction

Neural networks are models of computation at the core of modern machine learning. In feed-forward networks used as classifiers, given an input vector of size N, the network computes an output by alternating layers of linear transformations based on 'weight' matrices with non-linear activation functions. Arguably, the simplest neural network is the *binary perceptron* [31, 13], consisting of a single layer and binary weights $\mathbf{x} \in \{-1, 1\}^N$. Given a matrix of P data points in N-dimensional space, $\mathbf{A} \in \mathbb{R}^{P \times N}$ (also called the *disorder*), the perceptron outputs the labels $f_{\mathbf{A}}(\mathbf{x}) = \varphi(\mathbf{A}\mathbf{x})$, where the activation function $\varphi : \mathbb{R} \to \{-1, 1\}$ is applied element-wise.

Here we consider the perceptron from a dual point of view. Given the disorder $\mathbf{A}$, we use as inputs the weights $\mathbf{x}$ and study the function $f_{\mathbf{A}}(\mathbf{x})$ in a regime of large dimensions where $N, P \to \infty$ with fixed $\alpha = P/N$. Several constraint satisfaction problems (CSPs) arise in this context, listed in order of non-increasing algorithmic difficulty:

- *inversion:* given a matrix $\mathbf{A} \in \mathbb{R}^{P \times N}$ and a vector of labels $\boldsymbol{\ell} \in \{-1, 1\}^P$, find any set of inputs $\mathbf{x} \in \{-1, 1\}^N$ such that $f_{\mathbf{A}}(\mathbf{x}) = \boldsymbol{\ell}$.
- *second-preimage:* given a matrix $\mathbf{A} \in \mathbb{R}^{P \times N}$, and a 'teacher' $\mathbf{x} \in \{-1, 1\}^N$ generating labels $\boldsymbol{\ell} = f_{\mathbf{A}}(\mathbf{x}) \in \{-1, 1\}^P$, find any 'student' $\mathbf{x}' \in \{-1, 1\}^N$ with $\mathbf{x}' \neq \mathbf{x}$ such that $f_{\mathbf{A}}(\mathbf{x}') = \boldsymbol{\ell}$.

- *collision-finding:* given a matrix $\mathbf{A} \in \mathbb{R}^{P\times N}$, find any two inputs $\mathbf{x} \neq \mathbf{x}' \in \{-1,1\}^N$ such that $f_{\mathbf{A}}(\mathbf{x}) = f_{\mathbf{A}}(\mathbf{x}')$.

We are interested in the typical case complexity of CSPs, where the matrix $\mathbf{A}$ has random identically distributed independent (i.i.d.) entries, and also the labels are i.i.d. random in the inversion problem. Statistical physics has a long tradition of studying random CSPs, such as inversion and second-preimage variants [27, 35, 29, 32], yet the collision-finding problem itself had not been addressed within this framework until now. Notice the difference between second-preimage and collision-finding: in the latter, there is full freedom to choose both $\mathbf{x}$ and $\mathbf{x}'$ in a way that might depend on $\mathbf{A}$, while in the former one of the two is fixed. A function f for which there are no polynomial time algorithms that find collisions[1] is said to be collision resistant.

The collision-finding problem, on the other hand, has been extensively studied in cryptography. A function for which the collision-finding problem is hard is said to be *collision resistant*. If in addition, the function is *shrinking*, i.e. the output size P is smaller than the input size N, it is known as a *collision resistant hash function* (CRH). CRHs are of fundamental importance and form the basis of many cryptographic protocols and security guarantees. In fact, they are an integral part to securing data privacy [25], in applications including commitment schemes [11], encryption [14] and secure computation [44]. In this paper, our main goal is to understand whether the collision-finding problem in neural networks with a suitably chosen activation function is algorithmically hard.

Statistical physics studies on random Boolean satisfiability [35, 36, 33, 37] have linked typical-case algorithmic hardness to the presence of peculiar geometrical properties in the space of solutions of CSPs. Building upon this geometric approach [45, 6, 5, 7], it has been rigorously shown that the presence of a disconnectivity feature known as the overlap gap property (OGP) in the solution space of CSPs leads to the failure of stable algorithms [17, 18] and larger classes including online algorithms and low-degree polynomials. In many CSPs, the most efficient solvers identified to date belong to this class, giving rise to the conjecture that OGP is a marker of computational hardness[2]. Examples are the symmetric binary perceptron problem [23], number partitioning [21], maximum independent set on random graphs [22], and the optimization of p-spin glass Hamiltonians [19, 20].

Our main result is the identification of a novel candidate for collision resistant hash function based on a family of neural networks, where the OGP is used as a criterion for collision resistance between pairs of inputs of large Hamming distance. We then argue that by applying an error-correcting code upstream of a perceptron with an oscillating activation one obtains a still shrinking and collision resistant function.

These new candidate CRHs are markedly different from existing constructions from e.g. lattices, and hint at a new source of hardness for use in cryptography.

[1] Formally, collision resistance is defined as: for every efficient algorithm $C(\cdot)$ and any constant $c > 0$, the probability that $C(\mathbf{A})$ outputs $\mathbf{x} \neq \mathbf{y}$ with $f_{\mathbf{A}}(\mathbf{x}) = f_{\mathbf{A}}(\mathbf{y})$ is smaller than N^{-c} for any sufficiently large N (where the randomness is taken over the coins used by C)

[2] CSPs with a linear structure are considered somewhat an exception to this hardness conjecture, since despite the disconnectivity properties of their space of solutions, they can be always solved in polynomial time by Gaussian elimination.

2 Summary of Our Contributions

We study here the collision resistance of a family of perceptrons, where the activation function is a periodic square wave[3]. Specifically, we consider the following activation function, called the Square Wave Perceptron (SWP):

$$\varphi_\delta(h) = \operatorname{sgn}\left(\frac{h}{2\delta} \bmod 1\right) , \tag{1}$$

where $x \bmod 1$ is the unique real number in $[-\frac{1}{2}, \frac{1}{2})$ that differs from x by an integer, and $\delta \in \mathbb{R}^+$ is a parameter that controls the oscillation period. The function (1) is a square-wave with period 2δ. The choice of activation is ultimately guided by the goal of obtaining a shrinking collision resistant function which, as we show, is crucially dependent on δ.

Some properties of the SWP have been studied in [10] in average-case variants of inversion and second-preimage problems. Here we focus on the collision problem. In Section 3, we study the existence of collisions in the large size limit, in the regime where the inputs are at an extensive Hamming distance, or equivalently, when the overlap $q_1 = \mathbf{x}^\top \mathbf{y}/N$ between the two inputs $\mathbf{x}, \mathbf{y} \in \{-1, 1\}^N$ forming a collision is strictly smaller than one, $|q_1| \leq 1 - \Omega(1)$. We show that above a certain threshold of α, depending on q_1 and δ, with high probability collisions do not exist. The value of this threshold computed by using the replica method with a replica-symmetric Ansatz [34] coincides with a first-moment computation, leading us to conjecture that the replica-symmetric estimate coincides with the actual SAT/UNSAT transition in the collision problem. This conjecture is supported by numerical simulations based on exhaustive search of collisions on finite-size systems.

In Section 4 we move to the main focus of the paper, namely determining if there is a region of α where collisions exist but are hard to find. Using the first moment method we show that, conditioning to $|q_1| < 1$ the space of solutions of the collision-finding problem presents an OGP, which implies the failure of stable algorithms. We conjecture that in this regime, finding collisions with $|q_1| < 1$ is infeasible.

By tuning the parameter δ, we show that the conjectured hard region extends to values of α smaller than one, which is a request to construct a hash function. More precisely, we show that there exist α_*, δ_* and $q_* < 1$ such that for $\delta < \delta_*$ the space of collisions with overlap $q_1 < q_*$ exhibits OGP for $\alpha_\star < \alpha < 1$. Still, a solver could target collisions with overlap $q_* < q < 1$, where our analysis does not imply collision resistance for $\alpha < 1$. To remedy this, in Section 5 we propose a collision resistent hash function obtained by composing an error correcting code with the square wave perceptron. The error correcting code removes collisions with overlaps in the region $q_* < q < 1$, leading to a bona-fide collision resistant hash function.

In Section 6 we provide some numerical evidence supporting our conjectured CRHs. We study numerically two algorithmic attacks to find collisions, showing that both of them stop working at values of α smaller than the value α_* where we predict the presence of an OGP.

[3] As shown in the appendices, our analysis holds for a general φ. We concentrate mostly on the family of periodic square-wave activations merely because of their potential relevance in cryptography.

3 Existence of Collisions

Consider the number of collisions with internal overlap q_1:

$$Z(q_1; \boldsymbol{A}) := \sum_{\boldsymbol{c}\in\{\pm 1\}^{2N}} \mathbb{X}_{\boldsymbol{A}}(\boldsymbol{c})\, \delta\left(q_i(\boldsymbol{c}) - q_1\right), \tag{2}$$

where $\mathbb{X}_{\boldsymbol{A}}(\boldsymbol{c})$ is an indicator function, equal to one if the pair of inputs $\boldsymbol{c} = (\mathbf{x}, \mathbf{y})$ forms a collision and zero otherwise:

$$\mathbb{X}_{\mathbf{A}}(\boldsymbol{c}) = \prod_{\mu=1}^{P} \Theta\left[f_{\mathbf{A}}(\mathbf{x})_\mu f_{\mathbf{A}}(\mathbf{y})_\mu\right]. \tag{3}$$

In this notation $\Theta(\cdot)$ is the Heaviside step function, and $f_{\mathbf{A}}(\cdot)_\mu$ the μ-th element of vector $f_{\mathbf{A}}(\cdot)$. We have also defined the '*internal*' overlap of a collision $\boldsymbol{c}$ as the normalized scalar product between the colliding inputs

$$q_i(\boldsymbol{c}) = \frac{1}{N}\mathbf{x}^\top \mathbf{y}\,. \tag{4}$$

In the limit $N, P \to \infty$ with $\alpha = P/N$ fixed, the random variable $1/N \log Z$ concentrates on its average value, the so called *quenched* free entropy [42]:

$$\Phi(q_1) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\boldsymbol{A}} \log Z(q_1; \boldsymbol{A}). \tag{5}$$

Hence, $Z(q_1; \boldsymbol{A}) = \exp(N\Phi(q_1) + o(N))$, and if $\Phi < 0$ one has $Z(\boldsymbol{A}) = 0$ with high probability in the large N limit. Therefore, collisions with internal overlap q_1 exist up to a threshold $\alpha_c(q_1)$ where $\Phi = 0$. We estimated $\alpha_c(q_1)$ using two complementary approaches. First, we derived an upper bound $\alpha_c^a(q_1)$ to $\alpha_c(q_1)$ by using the first moment method. Because of Markov's inequality, the probability that the random variable $Z(q_1; \boldsymbol{A}) \geq 1$ is bounded by

$$P\left[Z(q_1; \boldsymbol{A}) \geq 1\right] \leq \mathbb{E}_{\boldsymbol{A}} Z(q_1; \boldsymbol{A}) = e^{N\Phi^{\mathrm{a}}(q_1)}, \tag{6}$$

where the *annealed* free entropy $\Phi^{\mathrm{a}}(q_1)$ is given by

$$\Phi^{\mathrm{a}}(q_1) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}_{\boldsymbol{A}} Z(q_1; \boldsymbol{A})\,. \tag{7}$$

The value of α for which the annealed free entropy vanishes therefore gives an upper bound to $\alpha_c(q_1)$. As we show in the Supplementary material, this is given by

$$\alpha_c^{\mathrm{a}}(q_1) = -\frac{\log(2) + H_B(q_1)}{\log \int Dz\, F_\varphi\left(\sqrt{q_1} z; \sqrt{1-q_1}\right)}, \tag{8}$$

where

$$H_B(q_1) \equiv -\frac{1-q_1}{2} \log\left(\frac{1-q_1}{2}\right) - \frac{1+q_1}{2} \log\left(\frac{1+q_1}{2}\right), \tag{9}$$

$$F_\varphi(x; \sigma) \equiv 1 - 2 I_\varphi(x; \sigma)\left(1 - I_\varphi(x; \sigma)\right), \tag{10}$$

$$I_\varphi(x; \sigma) \equiv \int Dh\, \Theta\left[\varphi\left(\sigma h + x\right)\right]. \tag{11}$$

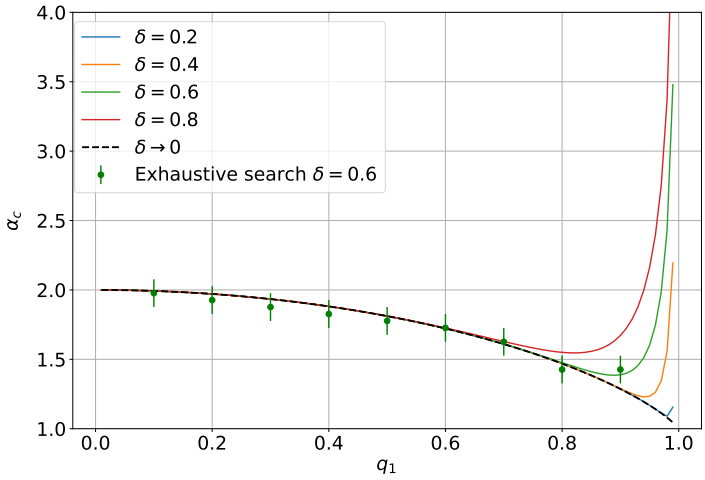

Figure 1: Full lines: analytic prediction for the transition α_c above which no collision exists versus the internal overlap q_1, for different values of δ. Dots: numerical estimates from exhaustive search.

Second, we applied the replica method [34] to compute the Replica Symmetric (RS) upper bound $\Phi^{RS} \geq \Phi$ to the quenched free entropy (5), and the corresponding upper bound $\alpha_c^{RS} \geq \alpha_c$. Details can be found in the Supplementary Material. Here we mention that previous applications of the replica method using the RS approximation [28] have already been shown to give the correct result for the storage capacity in inversion problems [15, 26] with binary inputs, whereas for continuous inputs one needs Replica Symmetry Breaking [9, 16, 4, 2].

Interestingly, we found that the more refined RS computation of the quenched free entropy (5), does not change the result found using the first moment method $\Phi^{\rm RS}(q_1) = \Phi^{\rm a}(q_1)$. This leads also to the same prediction for the value of α up to which collision exist $\alpha_c^{\rm RS}(q_1) = \alpha_c^{\rm a}(q_1)$, suggesting that the annealed bound is tight and, in this case, exact.

Figure 1 displays $\alpha_c^{\rm RS}$ as a function of q_1 for various values of δ. For $\alpha > \alpha_c^{RS}(q_1)$ the number of collisions is exponentially small. For all δ strictly larger than zero, $\alpha_c(q_1)$ diverges in the limit $q_1 \to 1$ as $\sim (1-q_1)^{-1/2}$. This is to be expected: trivial collisions with $q_1 = 1$, i.e., $\mathbf{x} = \mathbf{y}$ are always present. The divergence around $q_1 \sim 1$ means that sub-extensive collisions exist for all $\alpha = O(1)$ and independent of the activation function. In the SM we show that the computation of the annealed free entropy for $q_1 \approx 1 - 1/N$ consistently predicts $\alpha_c^{\rm a} \propto N^{1/2}$.

Finally, notice that for each value of q_1, decreasing δ leads to a smaller value of α_c^{RS}. In particular in the small δ limit, which we call *strong hashing limit*, one obtains

$$\alpha_c^{\rm a}(q_1) = 1 + \frac{H_B(q_1)}{\log 2}, \tag{12}$$

which is a monotonically decreasing function of q_1. This curve is shown in black dashed in Figure 1. Dots in Fig. 1 present a numerical estimate of α_c, obtained by exhaustively searching the space of weights and enumerating how many collisions exist with a given value of q_1 (see the SM for details). The procedure takes an exponential time in N, so the estimate is based on data with small $N = 14, 16, ..., 24$. Still, finite size effects are small, and the extrapolation are in excellent agreement with the analytic prediction.

4 Algorithmic Barriers to Collision-Finding

4.1 Multi Overlap Gap Property.

The overlap gap property is a geometric property of the solution space of CSPs, which is proven to imply the failure of stable algorithms. Informally, the stability of an algorithm $\mathcal{A}$, defined as a map from the instance space to the solution space of a CSP, requires that a small perturbation of the input results in a small perturbation of the output. In various CSPs, the best known solvers have been shown to belong to this class, fostering the belief that the presence of OGP is a signature of general algorithmic hardness [17]. Examples of algorithms proven to satisfy this stability property are the Kim-Roche algorithm for the symmetric binary perceptron [23], approximate message passing (AMP) [19], low-degree polynomials, Langevin dynamics and low-depth Boolean circuits [20, 43].

In this section we discuss a variant of OGP called multi overlap gap property (m-OGP) [17, 18]. In order to define it, we need to introduce a distance between two arbitrary collisions. Take two arbitrary colliding pairs $\boldsymbol{c}_a, \boldsymbol{c}_b \in \{-1,1\}^{2N}$; we define their '*external*' overlap as

$$q_e(\boldsymbol{c}_a, \boldsymbol{c}_b) \equiv \frac{1}{2N}\left(\mathbf{x}_a^\top \mathbf{x}_b + \mathbf{y}_a^\top \mathbf{y}_b\right). \tag{13}$$

The overlaps range from -1 to 1. Let $m \geq 2$ be an integer. We say that the function $f_{\mathbf{A}}(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}^P$ exhibits m-OGP for the collision-finding problem, with parameters $-1 \leq q_1 \leq 1$ and $-1 \leq \varepsilon_1 < \varepsilon_2 \leq 1$ if, given a random $\mathbf{A}$, it holds with high probability that there is no set of m pairs $\boldsymbol{c}_a = (\mathbf{x}_a, \mathbf{y}_a)$, $a = 1, \dots, m$ such that the following conditions are met:

1. $f_{\mathbf{A}}(\mathbf{x}_a) = f_{\mathbf{A}}(\mathbf{y}_a)$ with $\mathbf{x}_a \neq \mathbf{y}_a$, for every $a = 1 \dots m$;
2. $q_i(\boldsymbol{c}_a) = q_1$, for every $a = 1, \dots, m$;
3. $\varepsilon_1 \leq q_e(\boldsymbol{c}_a, \boldsymbol{c}_b) \leq \varepsilon_2$ for all $a \neq b$.

The first condition imposes that each $\boldsymbol{c}_a$ is a collision. The parameter q_1 in the second condition controls the internal overlap of each collision as defined in (4); the third condition imposes a gap on the external overlap between each collision pair. We define, respectively, $q_1 = O(1) < 1$ and $q_1 = 1 - o(1)$ as the 'extensive' and 'sub-extensive' regimes. It is important to note that obstruction to stable algorithms does not require a specific value of m [17]. In other words, independent of $m \geq 2$, $\alpha^m_{\mathrm{OGP}}(q_1)$ serves as an upper bound on the values of α where stable algorithms can find collisions with internal overlap q_1. Furthermore, if $m' > m$, then m-OGP implies m'-OGP, meaning that larger values of m can provide stronger bounds on the region of algorithmic hardness.

The $m-$OGP has a simple geometric interpretation. It requires that for every m-tuple of collisions with a certain internal overlap, there are at least two collisions that are 'close' or 'far apart' in the sense of definition (13).

In the following, we study the presence of $m-$OGP in the collision-finding problem for the function $f_{\mathbf{A}}$, in the limit where both the dimensions of the input N and the output P of the function scale to infinity, with their ratio fixed to $\alpha = P/N = O(1)$. In this high-dimensional limit we show the existence of a region of α delimited by $\alpha^m_{\mathrm{OGP}}(q_1)$, such that for $\alpha > \alpha^m_{\mathrm{OGP}}(q_1)$ the m-OGP holds for the collision problem with internal overlap q_1. Moreover, $\alpha^m_{\mathrm{OGP}}(q_1)$ is strictly

lower than the value $\alpha_c(q_1)$ above which no collisions with internal overlap q_1 can be found. We emphasize that the analytical formulas we derive in the Supplementary Material are valid for generic non-linearity φ. However, in the main text we focus specifically on the SWP activation function, due to its appealing cryptographic properties.

4.2 Existence of m-OGP in the Collision-Finding problem

The quantity of interest in establishing the presence of $m-$OGP is the number of m-tuples of collisions $\{\boldsymbol{c}_a = (\mathbf{x}_a, \mathbf{y}_a)\}_{a=1}^m$ with reciprocal external overlap p, and internal overlap q_1:

$$\mathcal{N}_m(p, q_1; \boldsymbol{A}) := \sum_{\{\boldsymbol{c}_a\}_{a=1}^m} \prod_{a=1}^m \left[\mathbb{X}_{\boldsymbol{A}}(\boldsymbol{c}_a)\, \delta(q_i(\boldsymbol{c}_a) - q_1)\right] \times \prod_{a<b} \delta\left(q_e(\boldsymbol{c}^a, \boldsymbol{c}^b) - p\right) . \tag{14}$$

We say that the problem has an m-OGP when, for a given value of the internal overlap q_1, there exists a range of external overlaps p where $\mathcal{N}_m(p, q_1; \boldsymbol{A}) = 0$ w.h.p. over the realization of the disorder. Note that this is analogous to condition 3 in the definition of m-OGP (see Sec. 4.1). The existence of such an excluded region can be proved by noticing that, since $\mathcal{N}_m(p, q_1; \mathbf{A})$ is a non-negative integer, Markov inequality gives

$$P(\mathcal{N}_m(p, q_1; \mathbf{A}) > 0) \leq \mathbb{E}_{\mathbf{A}} \mathcal{N}_m(p, q_1; \mathbf{A}) = e^{mN\phi_m^{\mathrm{a}}(p, q_1)}. \tag{15}$$

where the annealed entropy ϕ_m^{a} is defined as

$$\phi_m^{\mathrm{a}}(p, q_1) = \lim_{N\to\infty} \frac{1}{mN} \ln \mathbb{E}_{\boldsymbol{A}} \mathcal{N}_m(p, q_1; \boldsymbol{A}). \tag{16}$$

If, given q_1, for some value of α there exists an interval of p where $\phi_m^{\mathrm{a}}(p, q_1) \leq 0$, then Equation (15) proves that w.h.p. there are no collisions within such an external overlap range. We define $\alpha_{\mathrm{OGP}}^m(q_1)$ as the lowest value of α such that this condition is met. Notice that, because of Markov's inequality, $\alpha_{\mathrm{OGP}}^m(q_1)$ computed using the annealed entropy in (16) is only an upper bound to the value α_{OGP} above which an excluded region actually exists.

The computation of (16) is detailed in the SM, here we only state the final result. One finds that $\phi_m^a(p, q_1)$ can be computed by maximizing a function $\phi_m(Q)$ over a suitable space of matrices $Q \in \mathbf{F}(p, q_1)$

$$\phi_m^a(p, q_1) = \max_{Q\in\mathcal{F}(p,q_1)} \phi_m(Q)\,. \tag{17}$$

The detailed expression of $\phi_m(Q)$ is reported in the SM. Q is a $2m \times 2m$ matrix that represents the covariance matrix of m-clones of colliding inputs. Namely denoting by α an index that runs over the $2m$ inputs of the m collision clones, Q is defined as

$$Q_{\alpha\beta} = \frac{1}{N} \sum_{k=1}^N w_{k\alpha} w_{k\beta}\,. \tag{18}$$

The maximization over the entries of Q in (17) is performed over $\mathcal{F}(p, q_1)$, which is the set of covariance matrices Q that fix the internal overlap q_1 of the m collisions and the external overlap among them to p. In other words the entries of Q should satisfy the following set of constraints

$$\frac{1}{2}\left(Q_{2s-1,2t-1} + Q_{2s,2t}\right) = p\,, \qquad \forall s < t \in [m] \tag{19a}$$

$$Q_{2s-1,2s} = q_1\,, \qquad \forall s \in [m] \tag{19b}$$

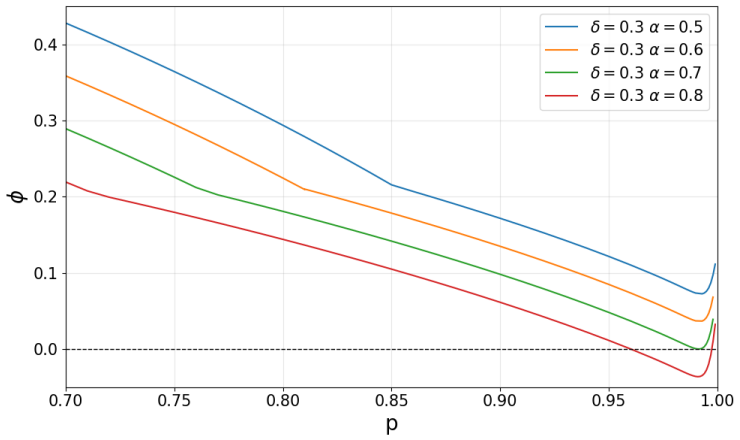

Figure 2: Typical behavior of $\phi^{\mathrm{a}}_m(p)$, for various values of α, at $q_1 = 0.6$ and $m = 5$. The activation is a square wave, with $\delta = 0.3$. The onset value of the OGP is $\alpha = 0.7$. For $\alpha > 0.7$, an interval of external overlaps p where $\phi < 0$ is present.

Finding the maximum in (17) over the space of covariance matrices Q can be done numerically; however, the complexity of such maximization increases considerably with m. We have therefore made an educated guess on the structure of the covariance matrix Q achieving the global maximum of the function $\phi_m(Q)$. We considered the so-called *symmetric* ansatz, which imposes

$$Q_{2s-1,2t-1} = Q_{2s,2t} = p\,, \qquad \forall s < t \in [m] \tag{20a}$$

$$Q_{2s-1,2t} = Q_{2s,2t-1} = q_0\,, \qquad \forall s < t \in [m] \tag{20b}$$

This ansatz considerably restricts the number of optimization parameters to just one (i.e. the parameter q_0). Moreover, this ansatz can be proved to be the right one in certain cases (see next section).

Figure 2 exemplifies the typical behavior of $\phi^{\mathrm{a}}_m(p)$, for various values of α, including the onset value of the OGP at $\alpha \sim 0.7$. Such curves are continuous, but not everywhere differentiable (see the SM for the details). Figure 3 displays α^m_{OGP} as a function of q_1 for $m = 5$ and several values of δ. Decreasing δ leads to a decrease of α_{OGP} for all q_1. For any δ strictly larger than zero, the α^m_{OGP} curve is non-monotonic in q_1, and start increasing when $q_1 \lesssim 1$, as observed for $\alpha_c(q_1)$. We stress that this threshold is strictly below the one corresponding to the existence of collisions. Between $\alpha^m_{\mathrm{OGP}}(q_1)$ and $\alpha_c(q_1)$, despite the presence of exponentially many collisions we conjecture that finding one of them is algorithmically hard in the average case.

4.3 Strong Hashing Limit

The strong hashing limit (SHL) $\delta \to 0$ can be explicitly studied analytically. Moreover, it is a simple case in which the validity of the symmetric ansatz can be theoretically tested. In particular, as we show in the SM, a detailed study of the case $m = 2$ reveals that: i) the global maximization procedure (17) leads to a symmetric (20) covariance matrix Q; ii) in the SHL, the optimum is reached at q_0 satisfying:

$$q_0 = p + q_1 - 1. \tag{21}$$

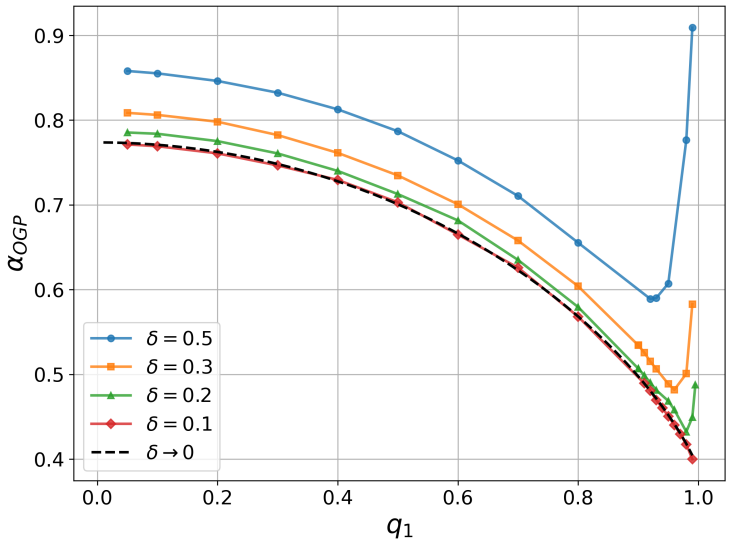

Figure 3: α^m_{OGP} versus the internal overlap q_1 for the Square Wave activation function, for various values of δ, $m = 5$.

Condition (21) implies that Q has a vanishing eigenvalue. In the SM we show that, under the symmetric ansatz, the matrix Q achieving the global maximum in the SHL satisfies the constraint (21) even for $m > 2$, implying that Q is not full rank in this case as well.

In the SHL, the annealed entropy curve becomes a monotonic function of the external overlap p. This allows us to explicitly find the value of $\alpha^m_{\mathrm{OGP}}(q_1)$ by performing the limit $p \to 1$ for every m. One finds:

$$\alpha^m_{\mathrm{OGP}}(q_1) = \frac{\log 2 + H_B(q_1)}{\log(1+m)} \,. \tag{22}$$

Note that this threshold is monotonically decreasing in m and goes to 0 for large m. In Figure 4 we show the OGP threshold given by (22) as a function of q_1 and for different values of m.

In Figure 3 we plot the $\delta \to 0$ OGP transition in black dashed for $m = 5$ together with predictions for finite values of δ. Notice that for $\delta = 0.1$ already, the curve is practically indistinguishable from the SHL.

Since we have established the validity of the symmetric ansatz only for $m = 2$ in the SHL, this does not rule out the possibility that the true global maximum is not symmetric, especially for $\delta > 0$. We have therefore compared the analytical prediction of the free entropy under the symmetric ansatz, $\phi^{\mathrm{a}}_m(p, q_1)$, with a numerical estimate obtained by an exhaustive search in the space of m-tuples of collisions. The results presented in the SM are in agreement with the theory.

5 Hash Functions from Collision Resistant Neural Networks

5.1 Collision Resistant Function

As already observed, when $\delta > 0$ all m-OGP curves diverge in the limit $q_1 \to 1$. Since an adversary is allowed to look for collisions with arbitrary value of q_1, the region $q_1 \approx 1$ represents a potential flaw in the security of the hash function. To remedy this, recall that an *error-correcting code* (ECC)

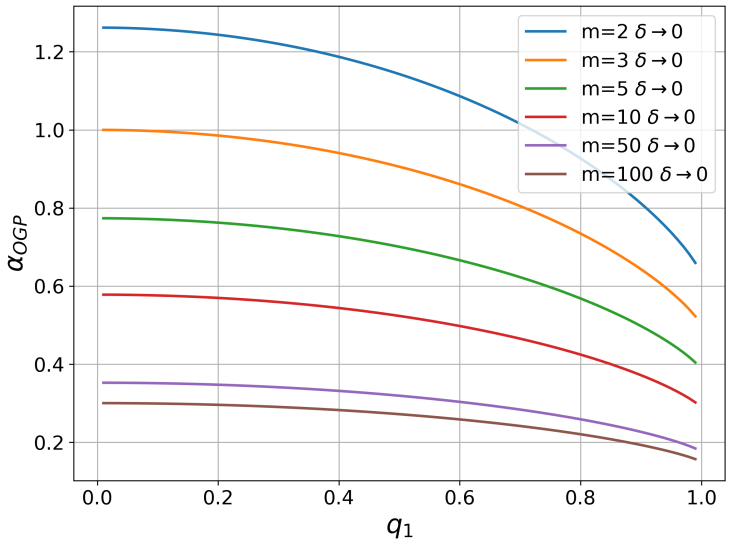

Figure 4: α^m_{OGP} versus the internal overlap q_1 for the Square Wave activation function in the limit $\delta \to 0$ for several values of m.

is a function that maps distinct inputs into 'codewords', separated by large Hamming distance. In order to obtain a collision resistant function, one can apply an ECC upstream of the network, restricting the space of inputs to $q_1 \leq q_{\text{code}}$. The results of Section 4.2 then imply that the composition $f_{\boldsymbol{A}} \circ ECC$ is collision resistant as long as α is higher than the OGP threshold for an adversarially chosen $q_1 \leq q_{\text{code}}$:

$$\alpha > \alpha_{\text{adv}}(q_{\text{code}}) := \max_{q_1 \leq q_{\text{code}}} \alpha_{\text{OGP}}(q_1). \tag{23}$$

5.2 Collision Resistant Hash Function

We now show how, by composing the neural network with an ECC with appropriate q_{code} value, it is possible to build a CRH. We define the code rate r as the ratio between the input and output bits of the ECC. The compression ratio of the neural network composed with the code is then $\tilde{\alpha} \equiv \alpha / r$. The Gilbert-Varshamov (GV) bound [32] shows the existence of a binary code satisfying $r(q_{\text{code}}) = 1 - H_2(q_{\text{code}})$, where

$$H_2(q_{\text{code}}) = -\tfrac{1-q_{\text{code}}}{2} \log_2 \tfrac{1-q_{\text{code}}}{2} - \tfrac{q_{\text{code}}+1}{2} \log_2 \tfrac{q_{\text{code}}+1}{2}.$$

The lowest compression rate compatible with collision resistance is then

$$\tilde{\alpha}(q_{\text{code}}) = \frac{\alpha_{\text{adv}}(q_{\text{code}})}{r(q_{\text{code}})} \tag{24}$$

obtained using a code matching the GV bound, and selecting the smallest α that guarantees hardness of collision-finding for $q_1 \leq q_{\text{code}}$. Figure 5 shows $\tilde{\alpha}(q_{\text{code}})$ as a function of q_{code}, for $m = 5$ and various δ values. In the interval of q_{code} values where $\tilde{\alpha}(q_{\text{code}}) < 1$, we conjecture that $f_{\boldsymbol{A}} \circ ECC$ is a CRH.

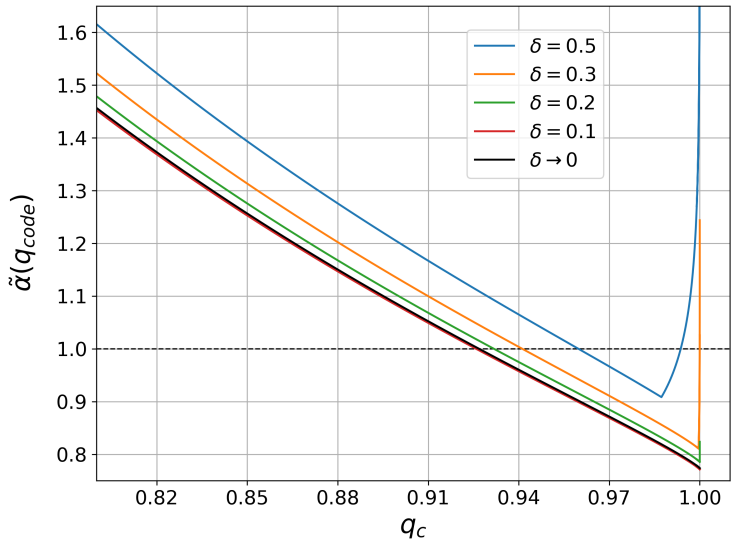

Figure 5: The lowest compression rate of $NN \circ ECC$, i.e. $\alpha_{\text{adv}}(q_{\text{code}})/r(q_{\text{code}})$, as a function of q_{code}, for different values of δ. $\alpha_{\text{adv}}(q_{\text{code}})$ is given by a m-OGP computation with $m = 5$.

5.3 Comparison with Ajtai's Function.

Post-quantum cryptography seeks primitives whose security can be based on problems that are believed to remain intractable even for quantum algorithms. Among the various approaches, lattice-based cryptography has emerged as one of the most promising directions, offering efficient CRHs constructions whose security lies on worst-case to average case reductions.

Ajtai's function [1] stands as a landmark in lattice-based cryptography, as it laid the foundation for subsequent, more efficient lattice-based constructions. Formally, it acts on $\mathbf{x} \in \{0,1\}^N$ as:

$$f_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{x} \pmod{q}. \tag{25}$$

Here, $q \in \mathbb{N}$ is a parameter and $\mathbf{A} \in \mathbb{F}_q^{P \times N}$ is chosen uniformly at random, where $\mathbb{F}_q$ is a Galois field with q elements. This function is shrinking when $P \log_2 q < N$. For certain values of $q, P = \text{poly}(N)$ satisfying this bound, Ajtai's function can be shown to be a CRH, assuming it is hard to find short vectors in certain integer lattices [24, 1]. Namely, the security relies on the worst-case hardness of the *short integer solution* (SIS) problem [39] that for a matrix $\mathbf{A} \in \mathbb{F}_q^{n \times m}$ asks to find a 'short' vector $\mathbf{x} \in \mathbb{F}_q^m$ such that $\mathbf{A}\mathbf{x} = 0$, a problem for which, for certain values of q, M, N, the only known algorithms run in time $2^{\Omega(n)}$ [38, 41], even with quantum computers.

Note that while our function bears some resemblance to Ajtai's function, as the oscillating nature of our activation is reminiscent of the mod operation. So our parameter δ plays a role similar to q in Ajtai's function. However, there are also important differences. First of all, in Ajtai's case the worst to average case reduction, which is a crucial request for its collision resistance, is guaranteed when q scales as $\text{poly}(N)$, whereas in our case we can maintain $\delta = O(1)$. Secondly, the output spaces differ: our function maps to $\pm 1^P$, while Ajtai's maps to $\mathbb{F}_q^P$. Moreover, in our case the entries of $\boldsymbol{A}$ are independent random variables which are drawn from an *arbitrary* distribution. Note also that our security guarantee via the OGP criterion, only depends on the first two moments of such distribution. Finally, and most importantly, the periodicity of Ajtai's function is believed to be

a central aspect of its hardness, while our results do not rely on the periodicity of the activation. Indeed, we have repeated the OGP analysis for an activation with a finite number $2K$ of oscillations, followed by a constant plateau, i.e. a 'truncated' version of the SWP. Specifically, oscillations span a range $[-\gamma, \gamma]$, where γ is a constant with N. As we show in the SM it is possible to find a region for q_{code} where our function composed with the ECC is a CRH. Preliminary checks indicate that qualitatively similar results also arise in a randomized version of the SWP, where the periodicity is broken by randomizing the points at which the activation function switches sign.

These results suggest that what matters the most for our hardness criterion is not the periodicity of the function, but rather the frequency of sign switches of the activation, even if they are located in a limited region near the origin.

Qualitatively, our function can be considered a generalization of Ajtai's function, where one considers a specific bit of its output, when outputs in $\mathbb{Z}_q$ are represented as binary strings, rather than the least significant bit. The corresponding modification of Ajtai's function would not be secure due to the performance of the best-known lattice algorithms [30, 40].

6 Algorithmic Attacks

In this section, we present two algorithmic strategies for finding collisions: one involves a local search starting from a random point on the hypercube, where local moves are performed to explore potential collisions, while the other is based on approximate message passing.

6.1 Local Search From a Random Reference

Perhaps the most straightforward strategy to look for a collision in a binary perceptron is the following. Take a random binary configuration $\boldsymbol{w}$. To construct a collision, one can attempt a sequence of single bit-flips on the elements of $\boldsymbol{w}$, stopping as soon as the resulting vector $\boldsymbol{w}'$ collides with $\boldsymbol{w}$. In this section, we argue that this strategy is unfeasible if the number of patterns is proportional to N. In order to do so, let us count the average number $\mathcal{N}_t$ of collisions obtained by flipping $t = O(1)$ bits from the reference $\boldsymbol{w}$. One gets (see the SM):

$$\mathcal{N}_t \approx \binom{N}{t} e^{-a(\delta)\, t^{1/2}\, P\, N^{-1/2}}, \tag{26}$$

where $a(\delta)$ is a positive constant in N, which depends on δ. (26) implies that if $P = O(N)$, in the limit $N \to \infty$ there are no collisions in the finite-size neighborhood of typical vertices of the hypercube, indicating that a local search from a random reference $\boldsymbol{w}$ is an unfeasible strategy.

6.2 Extensive Distances From a Random Reference

Consider again a random vertex of the hypercube $\boldsymbol{w}$. In this section, instead of looking at a finite-distance neighborhood of $\boldsymbol{w}$, we are interested in the algorithmic problem of finding collisions at an extensive distance. Note that in this case the probability of finding a collision is exponentially small in N, but there is an exponential number of configurations and the two can compensate (see (26)). In order to do so, we study the performance of a message-passing algorithm inspired by statistical physics called reinforced approximate message passing (rAMP). Variants of the rAMP algorithms are known to be effective polynomial-time heuristics to solve inversion and second-preimage problems in perceptrons with binary synapses [12, 6, 3, 8, 7]. We write in the SI the

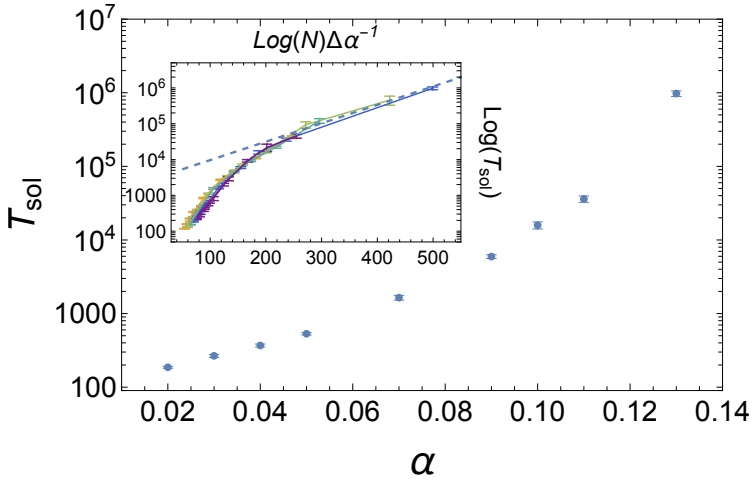

Figure 6: Number of iterations $T_{\rm sol}$ required by rAMP to find a solution for $N = 8 \times 10^3$ and $\delta = 0.6$, as a function of the constraint density α. *Inset*: $T_{\rm sol}$ has an exponential growth of the form $T_{\rm sol} \approx N^{b(\alpha)}$, with $b(\alpha) \propto (\alpha_r - \alpha)^{-1} \equiv \Delta\alpha^{-1}$ in the proximity of $\alpha_r \approx 0.15$. To test this behaviour, we show $\log T_{\rm sol}$ as a function of $\log N/\Delta\alpha$, for different values of N. The curves collapse for large $\log N/\Delta\alpha$. From bottom left to bottom right $N = n \times 10^3$, with $n = 1, 2, 4, 8, 16, 32$. The dashed line is a scaling function of the form $\exp(a\,x + b)$.

details of the implementation and the analysis of the performance with square wave activation. The complexity of rAMP is $O(T_{\rm sol} NP)$, where $NP \propto N^2$ comes from the fact that the algorithm involves matrix vector multiplications with the pattern matrix $\boldsymbol{A}$, and $T_{\rm sol}$ is the number of iterations required to find a solution. The quantity $T_{\rm sol}$ displays a power-law behavior $T_{\rm sol} \approx N^{b(\alpha)}$ in the system size, with a critical exponent $b(\alpha) \approx 1/(\alpha_r - \alpha)$. Therefore, the total complexity is $O\left(N^{2+b(\alpha)}\right)$. The exponent $b(\alpha)$ is usually extrapolated numerically (see [3]).

In order to test the hardness of the collision-finding problem, we run rAMP on the SWP with $\delta = 0.6$, where the annealed computation of the previous sections predicts the presence of an OGP. We find that rAMP outputs collisions having internal normalized overlap compatible with zero up to $\alpha_{\rm rAMP} \approx 0.15$, where $\alpha_{\rm rAMP}$ is estimated by a numerical characterization of the exponent $b(\alpha)$ (see Fig. 6). We note that $\alpha_{\rm rAMP}$ is strictly smaller than the OGP estimate for $q_1 = 0$, that for $m = 5$ and $\delta = 0.6$ is $\alpha^5_{OGP} \approx 0.9$ (extrapolating from Fig. 3).

7 Conclusions

The square-wave perceptron, together with its truncated variant, defines a class of neural-network-based functions that, in a suitable regime, admit collisions while exhibiting features consistent with collision-resistance, as suggested by the presence of an overlap gap property. By combining these functions with error-correcting codes, we obtain a novel family of candidate collision-resistant hash functions. To our knowledge, this is the first instance where a geometric criterion rooted in statistical physics is employed to argue the security of cryptographic primitives. A direction for future work is to investigate whether more complex neural network architectures also display this property.

Acknowledgements E.M.M. acknowledges the MUR-Prin 2022 funding Prot. 20229T9EAT, financed by the European Union (Next Generation EU). This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program. N.I.S. was supported by European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (Grant agreement No. 101019547). A.R. supported by European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (Grant agreement No. 101019547) and Cariplo CRYPTONOMEX grant.

References

[1] Miklós Ajtai. "Generating hard instances of lattice problems". In: *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. 1996, pp. 99–108.

[2] Brandon L. Annesi, Enrico M. Malatesta, and Francesco Zamponi. "Exact full-RSB SAT/UNSAT transition in infinitely wide two-layer neural networks". In: *SciPost Phys.* 18 (2025), p. 118. DOI: 10.21468/SciPostPhys.18.4.118. URL: https://scipost.org/10.21468/SciPostPhys.18.4.118.

[3] Carlo Baldassi and Alfredo Braunstein. "A max-sum algorithm for training discrete neural networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.8 (2015), P08008.

[4] Carlo Baldassi, Enrico M. Malatesta, and Riccardo Zecchina. "Properties of the Geometry of Solutions and Capacity of Multilayer Neural Networks with Rectified Linear Unit Activations". In: *Phys. Rev. Lett.* 123 (17 Oct. 2019), p. 170602. DOI: 10.1103/PhysRevLett.123.170602. URL: https://link.aps.org/doi/10.1103/PhysRevLett.123.170602.

[5] Carlo Baldassi et al. "Learning through atypical phase transitions in overparameterized neural networks". In: *Phys. Rev. E* 106 (1 July 2022), p. 014116. DOI: 10.1103/PhysRevE.106.014116. URL: https://link.aps.org/doi/10.1103/PhysRevE.106.014116.

[6] Carlo Baldassi et al. "Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses". In: *Physical review letters* 115.12 (2015), p. 128101.

[7] Carlo Baldassi et al. "Typical and atypical solutions in nonconvex neural networks with discrete and continuous weights". In: *Phys. Rev. E* 108 (2 Aug. 2023), p. 024310. DOI: 10.1103/PhysRevE.108.024310. URL: https://link.aps.org/doi/10.1103/PhysRevE.108.024310.

[8] Carlo Baldassi et al. "Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes". In: *Proceedings of the National Academy of Sciences* 113.48 (2016), E7655–E7662.

[9] E. Barkai, D. Hansel, and H. Sompolinsky. "Broken symmetries in multilayered perceptrons". In: *Phys. Rev. A* 45 (6 Mar. 1992), pp. 4146–4161. DOI: 10.1103/PhysRevA.45.4146. URL: https://link.aps.org/doi/10.1103/PhysRevA.45.4146.

[10] Marco Benedetti et al. "Overlap Gap and Computational Thresholds in the Square Wave Perceptron". In: *arXiv preprint arXiv:2506.05197* (2025).

[11] Manuel Blum. "Coin flipping by telephone a protocol for solving impossible problems". In: *ACM SIGACT News* 15.1 (1983), pp. 23–27.

[12] Alfredo Braunstein and Riccardo Zecchina. "Learning by message passing in networks of discrete synapses". In: *Physical review letters* 96.3 (2006), p. 030201.

[13] Thomas M Cover. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition". In: *IEEE transactions on electronic computers* 3 (1965), pp. 326–334.

[14] Whitfield Diffie and Martin E Hellman. "New directions in cryptography". In: *Democratizing Cryptography: The Work of Whitfield Diffie and Martin Hellman*. 2022, pp. 365–390.

[15] Jian Ding and Nike Sun. "Capacity lower bound for the Ising perceptron". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 816–827.

[16] A. Engel et al. "Storage capacity and learning algorithms for two-layer neural networks". In: *Phys. Rev. A* 45 (10 May 1992), pp. 7590–7609. DOI: 10.1103/PhysRevA.45.7590. URL: https://link.aps.org/doi/10.1103/PhysRevA.45.7590.

[17] David Gamarnik. "The overlap gap property: A topological barrier to optimizing over random structures". In: *Proceedings of the National Academy of Sciences* 118.41 (2021), e2108492118.

[18] David Gamarnik. "Turing in the shadows of Nobel and Abel: an algorithmic story behind two recent prizes". In: *arXiv preprint arXiv:2501.15312* (2025).

[19] David Gamarnik and Aukosh Jagannath. "The overlap gap property and approximate message passing algorithms for p-spin models". In: *The Annals of Probability* 49.1 (2021), pp. 180–205.

[20] David Gamarnik, Aukosh Jagannath, and Alexander S Wein. "Hardness of random optimization problems for Boolean circuits, low-degree polynomials, and Langevin dynamics". In: *SIAM Journal on Computing* 53.1 (2024), pp. 1–46.

[21] David Gamarnik and Eren C Kızıldağ. "Algorithmic obstructions in the random number partitioning problem". In: *The Annals of Applied Probability* 33.6B (2023), pp. 5497–5563.

[22] David Gamarnik, Eren C Kızıldağ, and Lutz Warnke. "Optimal hardness of online algorithms for large independent sets". In: *arXiv preprint arXiv:2504.11450* (2025).

[23] David Gamarnik et al. "Algorithms and barriers in the symmetric binary perceptron model". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 576–587.

[24] Oded Goldreich, Shafi Goldwasser, and Shai Halevi. "Collision-free hashing from lattice problems". In: *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation: In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman* (2011), pp. 30–39.

[25] Shafi Goldwasser, Silvio Micali, and Chales Rackoff. "The knowledge complexity of interactive proof-systems". In: *Providing sound foundations for cryptography: On the work of shafi goldwasser and silvio micali*. 2019, pp. 203–225.

[26] Brice Huang. "Capacity threshold for the Ising perceptron". In: *arXiv preprint arXiv:2404.18902* (2024).

[27] Scott Kirkpatrick and Bart Selman. "Critical behavior in the satisfiability of random boolean expressions". In: *Science* 264.5163 (1994), pp. 1297–1301.

[28] Werner Krauth and Marc Mézard. "Storage capacity of memory networks with binary couplings". In: *Journal de Physique* 50.20 (1989), pp. 3057–3066.

[29] Florent Krzakała et al. "Gibbs states and the set of solutions of random constraint satisfaction problems". In: *Proceedings of the National Academy of Sciences* 104.25 (2007), pp. 10318–10323. DOI: https://doi.org/10.1073/pnas.070368510.

[30] Arjen K Lenstra, Hendrik Willem Lenstra, and László Lovász. "Factoring polynomials with rational coefficients". In: *Mathematische annalen* 261 (1982), pp. 515–534.

[31] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.

[32] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[33] Marc Mézard, Thierry Mora, and Riccardo Zecchina. "Clustering of solutions in the random satisfiability problem". In: *Physical Review Letters* 94.19 (2005), p. 197205.

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

[35] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. "Analytic and algorithmic solution of random satisfiability problems". In: *Science* 297.5582 (2002), pp. 812–815. DOI: 10.1126/science.1073287.

[36] Marc Mézard and Riccardo Zecchina. "Random k-satisfiability problem: From an analytic solution to an efficient algorithm". In: *Physical Review E* 66.5 (2002), p. 056126.

[37] Rémi Monasson et al. "Determining computational complexity from characteristic 'phase transitions'". In: *Nature* 400.6740 (1999), pp. 133–137.

[38] Xavier Pujol and Damien Stehlé. "Solving the shortest lattice vector problem in time 2ˆ 2.465 n". In: *Cryptology ePrint Archive* (2009).

[39] Oded Regev. "On lattices, learning with errors, random linear codes, and cryptography". In: *Journal of the ACM (JACM)* 56.6 (2009), pp. 1–40.

[40] Claus-Peter Schnorr. "A more efficient algorithm for lattice basis reduction". In: *Journal of algorithms* 9.1 (1988), pp. 47–62.

[41] Claus-Peter Schnorr and Martin Euchner. "Lattice basis reduction: Improved practical algorithms and solving subset sum problems". In: *Mathematical programming* 66 (1994), pp. 181–199.

[42] Michel Talagrand. "Self averaging and the space of interactions in neural networks". In: *Random Structures & Algorithms* 14.3 (1999), pp. 199–213. DOI: https://doi.org/10.1002/(SICI)1098-2418(199905)14:3<199::AID-RSA1>3.0.CO;2-6.

[43] Alexander S Wein. "Optimal low-degree hardness of maximum independent set". In: *Mathematical Statistics and Learning* 4.3 (2022), pp. 221–251. DOI: 10.4171/msl/25.

[44] Andrew C Yao. "Protocols for secure computations". In: *23rd annual symposium on foundations of computer science (sfcs 1982)*. IEEE. 1982, pp. 160–164.

[45] Lenka Zdeborová and Florent Krzakala. "Phase transitions in the coloring of random graphs". In: *Phys. Rev. E* 76 (3 Sept. 2007), p. 031131. DOI: 10.1103/PhysRevE.76.031131. URL: https://link.aps.org/doi/10.1103/PhysRevE.76.031131.

A The Entropy of the $m-$Cloned System

A.1 Recap of the Main Definitions

As discussed in the main text, we are interested in computing the following free entropy:

$$\phi_m^{\mathrm{a}}(p, q_1) = \lim_{N\to\infty} \frac{1}{mN} \log \mathbb{E}_{\boldsymbol{A}} \mathcal{N}_m(p, q_1; \boldsymbol{A}) \tag{27}$$

where

$$\mathcal{N}_m(p, q_1; \boldsymbol{A}) \equiv \sum_{\{\boldsymbol{c}_a\}_{a=1}^m} \prod_{a=1}^m \mathbb{X}_{\boldsymbol{A}}(\boldsymbol{c}_a) \prod_{a<b} \delta\left(q_e(\boldsymbol{c}_a, \boldsymbol{c}_b) - p\right) \prod_{a=1}^m \delta\left(q_i(\boldsymbol{c}_a) - q_1\right) \tag{28}$$

represents the number of m-uples of collisions $\boldsymbol{c}_a \equiv (\mathbf{x}_a, \mathbf{y}_a)$, $a \in [m]$, that are at a mutual overlap $p \in [0, 1]$ and with an *internal overlap* $q_1 \in [-1, 1]$. We consider as internal overlap of a collision simply the scalar product:

$$q_i(\boldsymbol{c}_a) = \frac{1}{N} \mathbf{x}_a \cdot \mathbf{y}_a. \tag{29}$$

Note that the function $\mathbb{X}_{\boldsymbol{A}}(\boldsymbol{c})$

$$\mathbb{X}_{\mathbf{A}}(\boldsymbol{c}) = \prod_{\mu=1}^P \Theta\left[f_{\mathbf{A}}(\mathbf{x})_\mu f_{\mathbf{A}}(\mathbf{y})_\mu\right] = \prod_{\mu=1}^P \Theta\left[\varphi\left(\frac{1}{\sqrt{N}} \mathbf{x} \cdot \boldsymbol{A}^\mu\right) \varphi\left(\frac{1}{\sqrt{N}} \mathbf{y} \cdot \boldsymbol{A}^\mu\right)\right] \tag{30}$$

imposes that $\boldsymbol{c} = (\mathbf{x}, \mathbf{y})$ is a collision. With $\boldsymbol{A}^\mu$ we denoted the $\mu-$th row of the data matrix $\boldsymbol{A} \in \mathbb{R}^{P\times N}$, and with $f_{\mathbf{A}}(\bullet)_\mu$ the μ-th element of vector $f_{\mathbf{A}}(\bullet)$.

As stated in the main text, the *annealed* entropy in (27) gives an upper bound to the existence of such constrained m-uplets of collisions, thanks to Markov's inequality.

A.2 Overlap Among Collisions

We consider an overlap function between two collisions $\boldsymbol{c}_a = (\mathbf{x}_a, \mathbf{y}_a)$ and $\boldsymbol{c}_b = (\mathbf{x}_b, \mathbf{y}_b)$ of the form

$$q_e(\boldsymbol{c}_a, \boldsymbol{c}_b) = \frac{1}{2N}\left(\mathbf{x}_a^\top \mathbf{x}_b + \mathbf{y}_a^\top \mathbf{y}_b\right). \tag{31}$$

For simplicity, we will name this constraint as the *external overlap* between two collisions.

A.3 Computation of $\phi_m^{\mathrm{a}}(p, q_1)$

The computation of $\mathbb{E}_{\boldsymbol{A}} \mathcal{N}_m(p, q_1; \boldsymbol{A})$ follows from simple statistical physics methods. With a slight abuse of notation we will denote from now on the elements forming a collision with the same letter $\boldsymbol{w}$ as $\boldsymbol{c}_a = (\boldsymbol{w}_a^{(1)}, \boldsymbol{w}_a^{(2)})$. We will also use the index $a \in [m]$ as the one running over the number of clones m, whereas with $s = 1, 2$ as the one that runs over the two weight components composing each collision. We can use the identity

$$1 = \prod_{\mu,a,s} \int dh_s^{\mu a}\, \delta\left(h_s^{\mu a} - \frac{1}{\sqrt{N}} \mathbf{A}^\mu \cdot \boldsymbol{w}_a^{(s)}\right) = \prod_{\mu,a,s} \int \frac{dh_s^{\mu a} d\hat{h}_s^{\mu a}}{2\pi}\, e^{i\hat{h}_s^{\mu a}\left(h_s^{\mu a} - \frac{1}{\sqrt{N}} \mathbf{A}^\mu \cdot \boldsymbol{w}_a^{(s)}\right)} \tag{32}$$

to extract the dependence on the matrix $\mathbf{A}$ in $\prod_a \mathbb{X}_{\boldsymbol{A}}(\boldsymbol{c}_a)$. The average over $\mathbf{A}$ is then straightforward

$$\prod_{k=1}^{N}\prod_{\mu=1}^{P}\mathbb{E}_{A_k^\mu}e^{-i\frac{A_k^\mu}{\sqrt{N}}\left(\sum_{as}\hat{h}_s^{\mu a}w_{ak}^{(s)}\right)}=\prod_{k=1}^{N}\prod_{\mu=1}^{P}e^{-\frac{1}{2N}\left(\sum_\alpha \hat{h}_\alpha^\mu w_{k\alpha}\right)^2}=\prod_{\mu=1}^{P}e^{-\frac{1}{2}\sum_{\alpha\beta}Q_{\alpha\beta}\hat{h}_\alpha^\mu\hat{h}_\beta^\mu} \tag{33}$$

where we flattened the indices (a,s) into a collision index $\alpha\in[2m]$. We have also introduced the $2m\times 2m$-dimensional overlap matrix $Q_{\alpha\beta}$ which is defined as

$$Q_{\alpha\beta}\equiv\frac{1}{N}\sum_{k=1}^{N}w_{k\alpha}w_{k\beta} \tag{34}$$

which can be equivalently expressed in terms of indices $\alpha,\beta\in[2m]$ as a matrix that we will call $Q_{\alpha\beta}$. The corresponding $2y\times 2y$-dimensional matrix will be called as Q. Enforcing the definition in (34) via a delta function

$$1=\prod_{\alpha<\beta}\int dQ_{\alpha\beta}\,\delta\left(Q_{\alpha\beta}-\frac{1}{N}\boldsymbol{w}_\alpha\cdot\boldsymbol{w}_\beta\right) \tag{35}$$

one finally gets that the index $k\in[N]$ and $\mu\in[P]=[\alpha N]$ become decoupled. Notice that, the matrix $Q_{\alpha\beta}$ is symmetric and $Q_{\alpha\alpha}=1$ for any $\alpha\in[2m]$ as the weights are binary $w_{k\alpha}=\pm1$. This is the reason why we need to impose $2y(2y-1)$ constraints in equation (35). The expected value of $\mathcal{N}_m(p,q_1;\mathbf{A})$ can be therefore written as

$$\mathbb{E}_{\mathbf{A}}\left[\mathcal{N}_m(p,q_1;\mathbf{A})\right]=\int_{Q\in\mathcal{F}(p,q_1)}\prod_{\alpha<\beta}dQ_{\alpha\beta}\,e^{Nm\phi_m(Q)} \tag{36}$$

with

$$\phi_m(Q)\equiv G_S(Q)+\alpha G_E(Q) \tag{37a}$$

$$G_S(Q)\equiv\lim_{N\to\infty}\frac{1}{N}\log\left[\sum_{\boldsymbol{w}_\alpha}\prod_{\alpha<\beta}\delta\left(Q_{\alpha\beta}-\frac{1}{N}\boldsymbol{w}_\alpha\cdot\boldsymbol{w}_\beta\right)\right] \tag{37b}$$

$$G_E(Q)\equiv\frac{1}{m}\ln\int\mathcal{D}_Q(u_1,v_1,\dots u^m,v^m)\prod_{a=1}^{m}\Theta\left[\varphi(u^a)\varphi(v^a)\right]=\frac{1}{m}\ln\int\mathcal{D}_Q\boldsymbol{h}\prod_{a=1}^{m}\Theta\left[\varphi(h^{2a-1})\varphi(h^{2a})\right], \tag{37c}$$

where $\mathcal{D}_Q$ is a multivariate Gaussian measure over the variables $\boldsymbol{h}=(u_1,v_1,\dots u^m,v^m)$ with zero mean ad covariance given by the $2m\times 2m$-dimensional matrix of overlaps Q. $\mathcal{F}(p,q_1)$ represents the set of $2m\times 2m$ overlap matrices Q whose elements satisfy the constraint that the external (31) and internal overlaps (29) of the collisions are respectively p and q_1. The integral in Equation (36) can be computed at leading order in the limit $N\to\infty$ through the Saddle Point method, i.e.

$$\mathbb{E}_{\mathbf{A}}\left[\mathcal{N}_m(p,q_1;\mathbf{A})\right]=e^{Nm\phi_m^\star(p,q_1)}, \tag{38}$$

where

$$\phi_m^\star(p,q_1)\equiv\max_{Q\in\mathcal{F}(p,q_1)}\mathcal{G}_m(Q). \tag{39}$$

A.4 The Admissible Set of Covariance Matrices

We say that an overlap matrix $Q \in \mathcal{F}(p, q_1)$ is *admissible* if there exists a sequence of binary inputs $\{w_{k\alpha} = \pm 1\}$, $k \in [N]$, $\alpha \in [2m]$ such that

$$Q_{\alpha\beta} = \frac{1}{N} \sum_{k=1}^{N} w_{k\alpha} w_{k\beta} \,. \tag{40}$$

Admissible matrix are symmetric, have all elements in the interval $[-1, 1]$ and have the diagonal elements equal to 1. Moreover another necessary condition for Q to be admissible is that it is positive semi-definite, indeed for any vector u:

$$\sum_{\alpha\beta} u_\alpha Q_{\alpha\beta} u_\beta = \frac{1}{N} \sum_{k=1}^{N} \left(\sum_{\alpha} w_{k\alpha} u_\alpha \right)^2 \geq 0 \,. \tag{41}$$

In addition being $Q \in \mathcal{F}(p, q_1)$ the matrix element should satisfy the following constraints

$$\frac{1}{2} \left(Q_{2s-1,2t-1} + Q_{2s,2t} \right) = p \,, \qquad \forall s < t \in [m] \tag{42a}$$

$$Q_{2s-1,2s} = q_1 \,, \qquad \forall s \in [m] \tag{42b}$$

With a slight abuse of language, we will simply say that the set $\mathcal{F}(p, q_1)$ denotes the set of admissible covariance matrices.

Another important aspect of admissible covariance matrices concerns the possibility of some eigenvalues being zero. If the matrix Q has an eigenvector u with eigenvalue 0, (41) implies that there must exist a set of $2m$ spins $\tau_\alpha = \pm 1$ such that

$$\sum_{\alpha=1}^{2m} \tau_\alpha u_\alpha = 0 \,. \tag{43}$$

A.5 The Entropic Term

By introducing the Fourier integral representation of the delta function, the entropic term can be written as

$$G_S(Q) = \min_{\hat{Q}} \left[-\frac{1}{2m} \mathrm{Tr}[Q\hat{Q}] + \frac{1}{m} \log \sum\nolimits_{\boldsymbol{w}} e^{\frac{1}{2} \boldsymbol{w}^T \hat{Q} \boldsymbol{w}} \right] , \tag{44}$$

where we have introduced a conjugated variable matrix $\hat{Q}$ which enforces each of the constraints $Q_{\alpha\beta} = \frac{1}{N} \sum_{k=1}^{N} w_{i\alpha} w_{i\beta}$ with $\alpha < \beta$. Note that due to the binary nature of the inputs, we do not need to impose any constraint for $Q_{\alpha\alpha} = 1$ $\alpha \in [2m]$; so we consider $\hat{Q}_{\alpha\alpha} = 0$ when needed.

From this representation we can find a nice property of the $G_S(Q)$ function, namely that it is a *concave* function. Indeed, note that the function to be minimized in G_S is convex. The concave property therefore follows from the fact that $G_S(Q)$ is the Legendre transform of this convex function.

A.6 Energetic Term for the Square Wave Activation Function

We restrict here the discussion to the main activation function which we have analyzed in this paper, that is the square wave, which is

$$\varphi(h) \equiv \mathrm{sgn}\left(\frac{h}{2\delta} \bmod 1\right) = \mathrm{sign}\left[\sin\left(\frac{\pi}{\delta}h\right)\right], \tag{45}$$

where δ tunes the width of the oscillations. Since this activation is periodic with period 2δ we can conveniently evaluate the energetic term via Fourier transform.

Let $\mathcal{P}([m])$ be the set of subset of the set $[m]$. For each element $\gamma \in \mathcal{P}([m])$ we will denote by $|\gamma|$ its cardinality. The integrand in the energetic term can be written for the square wave activation as

$$\prod_{a=1}^{m} \Theta\left[\varphi(h^{2a-1})\varphi(h^{2a})\right] = \frac{1}{2^m}\prod_{a=1}^{m}\left[1+\varphi(h^{2a-1})\varphi(h^{2a})\right] = \frac{1}{2^m} + \frac{1}{2^m}\sum_{r=1}^{m}\sum_{\substack{\gamma\in\mathcal{P}([m])\\|\gamma|=r}}\prod_{g\in\gamma}\varphi(h^{2g-1})\varphi(h^{2g}). \tag{46}$$

In the following, we will denote by $\boldsymbol{h}_\gamma$ the $2r$ dimensional vector containing the variables h^{2g-1} and h^{2g}, $\forall g \in \gamma$. We define the function of $2r$ variables as

$$f(\boldsymbol{h}_\gamma) = \prod_{g\in\gamma}\varphi(h^{2g-1})\varphi(h^{2g}), \tag{47}$$

so that the energetic term can be expressed as

$$G_E = \frac{1}{m}\ln\left[\frac{1}{2^m} + \frac{1}{2^m}\sum_{r=1}^{m}\sum_{\substack{\gamma\in\mathcal{P}([m])\\|\gamma|=r}}\int \mathcal{D}_{Q_\gamma}\boldsymbol{h}_\gamma\, f(\boldsymbol{h}_\gamma)\right]. \tag{48}$$

In the previous expression, we have integrated over the variables h not appearing in the integrand that depends on $\boldsymbol{h}_\gamma$. We also defined Q_γ as the $2r \times 2r$ matrix obtained from Q by removing the $(2g-1)$-th and $2g$-th rows, as well as the corresponding columns, for all $g \notin \gamma$.

The $2r$ dimensional multivariate Gaussian integration can be computed analytically by using the Fourier transform of the function $f(\boldsymbol{h}_\gamma)$

$$\int \mathcal{D}_{Q_\gamma}\boldsymbol{h}_\gamma\, f(\boldsymbol{h}_\gamma) = \sum_{\boldsymbol{n}\in\mathbb{Z}^{2r}}\hat{f}(\boldsymbol{n})\int \mathcal{D}_{Q_\gamma}\boldsymbol{h}\, e^{-i\frac{\pi}{\delta}\boldsymbol{n}^T\boldsymbol{h}_\gamma} = \sum_{\boldsymbol{n}\in\mathbb{Z}^{2r}}\hat{f}(\boldsymbol{n})\, e^{-\frac{\pi^2}{2\delta^2}\boldsymbol{n}Q_\gamma\boldsymbol{n}}, \tag{49}$$

where $\hat{f}$ is the Fourier transform of f which reads

$$\hat{f}(\boldsymbol{n}) = \prod_{i=1}^{2r}\hat{\varphi}(n_i) \tag{50a}$$

$$\hat{\varphi}(n) = \frac{1}{\delta}\int_{-\delta}^{\delta} dh\,\varphi(h)\sin\left(\frac{\pi}{\delta}nh\right) = \frac{4\cos(n\pi)\sin^2\left(\frac{n\pi}{2}\right)}{n\pi}, \tag{50b}$$

which is equal to $-4/(\pi n)$ if n is odd and 0 otherwise. We therefore get the following representation of the energetic term for the square wave activation

$$G_E = \frac{1}{m} \ln \left[\frac{1}{2^m} + \frac{1}{2^m} \sum_{r=1}^{m} \sum_{\substack{\gamma \in \mathcal{P}([m]) \\ |\gamma|=r}} \left(-\frac{4}{\pi^2} \right)^r \sum_{\boldsymbol{n} \in \mathbb{Z}^{2r}} \frac{1}{\prod_{i=1}^{2r} (2n_i+1)} \, e^{-\frac{\pi^2}{2\delta^2} (2\boldsymbol{n}+1)^T Q_\gamma (2\boldsymbol{n}+1)} \right]. \quad (51)$$

A.7 Symmetric Ansatz and Hubbard-Stratonovich Representation

We now proceed under the assumption that such optimum can be found among the class of symmetric matrices

$$Q_{2s-1,2t-1} = Q_{2s,2t} = p\,, \qquad \forall s < t \in [m] \quad (52a)$$
$$Q_{2s-1,2t} = Q_{2s,2t-1} = q_0\,, \qquad \forall s < t \in [m] \quad (52b)$$

Similarly, we will consider a symmetric ansatz over the conjugated matrix

$$\hat{Q}_{2s-1,2t-1} = \hat{Q}_{2s,2t} = \hat{p}\,, \qquad \forall s < t \in [m] \quad (53a)$$
$$\hat{Q}_{2s-1,2t} = \hat{Q}_{2s,2t-1} = \hat{q}_0\,, \qquad \forall s < t \in [m] \quad (53b)$$
$$\hat{Q}_{2s-1,2s} = \hat{q}_1\,, \qquad \forall s \in [m]. \quad (53c)$$

For example, for $m = 3$, one has:

$$Q = \left(\begin{array}{cc|cc|cc} 1 & q_1 & p & q_0 & p & q_0 \\ q_1 & 1 & q_0 & p & q_0 & p \\ \hline p & p_0 & 1 & q_1 & p & q_0 \\ q_0 & p & q_1 & 1 & q_0 & p \\ \hline p & p_0 & p & q_0 & 1 & q_1 \\ q_0 & p & q_0 & p & q_1 & 1 \end{array} \right), \quad \hat{Q} = \left(\begin{array}{cc|cc|cc} 0 & \hat{q}_1 & \hat{p} & \hat{q}_0 & \hat{p} & \hat{q}_0 \\ \hat{q}_1 & 0 & \hat{q}_0 & \hat{p} & \hat{q}_0 & \hat{p} \\ \hline \hat{p} & \hat{p}_0 & 0 & \hat{q}_1 & \hat{p} & \hat{q}_0 \\ \hat{q}_0 & \hat{p} & \hat{q}_1 & 0 & \hat{q}_0 & \hat{p} \\ \hline \hat{p} & \hat{p}_0 & \hat{p} & \hat{q}_0 & 0 & \hat{q}_1 \\ \hat{q}_0 & \hat{p} & \hat{q}_0 & \hat{p} & \hat{q}_1 & 0 \end{array} \right). \quad (54)$$

This reasonable assumption can be verified for small m numerically by solving the optimization problem for general matrices and checking that symmetry holds at the optimum. Furthermore, in the SHL and $m = 2$ it can be proved rigorously (see Appendix C.1). Within this ansatz, one can perform a Hubbard-Stratonovich transformation to decouple terms in G_S and G_E that depend on different clone indexes. As a result, the parameter m appears as an exponent, rather as the number of integration variables in the expression, and can be promoted to a real value as in standard replica computations. For example, the entropic term reads

$$\begin{aligned} G_S =& - q_1\hat{q}_1 - (m-1)\, p\,\hat{p} - (m-1)\, q_0\,\hat{q}_0 \\ & - (\hat{p} + \hat{q}_1 - \hat{q}_0) \\ & + \frac{1}{m} \log \int Du_1 Du_2 \left[2\cosh\left(\sqrt{2(\hat{p}-\hat{q}_0)}\, u_1 \right) + 2e^{2(\hat{q}_1-\hat{q}_0)} \cosh\left(\sqrt{2(\hat{p}+\hat{q}_0)}\, u_2 \right) \right]^m \\ =& - q_1\hat{q}_1 - (m-1)p\hat{p} - (m-1)q_0\hat{q}_0 - (\hat{p} + \hat{q}_1 - \hat{q}_0) \\ & + \frac{1}{m} \log \sum_{k=0}^{m} \binom{m}{k} e^{2(m-k)(\hat{q}_1-\hat{q}_0)} \left[\sum_{s=0}^{k} \binom{k}{s} e^{(k-2s)^2(\hat{p}-\hat{q}_0)} \right] \left[\sum_{s=0}^{m-k} \binom{m-k}{s} e^{(m-k-2s)^2(\hat{p}+\hat{q}_0)} \right] \end{aligned}$$

where $Dz \equiv dz\,\frac{e^{-z^2/2}}{\sqrt{2\pi}}$. The energetic term can be written instead as

$$\begin{aligned}
G_E &\equiv \frac{1}{m}\log\int\prod_{as}\frac{dh_s^a d\hat{h}_s^a}{2\pi}e^{ih_s^a\hat{h}_s^a}\prod_{a=1}^{m}\Theta\left[\prod_{s=1}^{2}\varphi(h_s^a)\right]e^{-\frac{1}{2}\sum_{ab,st}q_{st}^{ab}\hat{h}_s^a\hat{h}_t^b} \\
&= \frac{1}{m}\log\int\prod_{as}\frac{dh_s^a d\hat{h}_s^a}{2\pi}e^{ih_s^a\hat{h}_s^a}\prod_{a=1}^{m}\Theta\left[\prod_{s=1}^{2}\varphi(h_s^a)\right] \\
&\quad\times e^{-\frac{1-q_1-p+q_0}{2}\sum_{as}(\hat{h}_s^a)^2-\frac{q_1-q_0}{2}\sum_a\left(\sum_s\hat{h}_s^a\right)^2-\frac{p-q_0}{2}\sum_s\left(\sum_a\hat{h}_s^a\right)^2-\frac{q_0}{2}\left(\sum_{as}\hat{h}_s^a\right)^2} \qquad (55)\\
&= \frac{1}{m}\log\int\prod_s Du_s\,Dx\left[\int Dz\prod_s Dh_s\,\Theta\Big(\prod_s\varphi\Big(\sqrt{1-q_1-p+q_0}\,h_s+\sqrt{p-q_0}\,u_s\right. \\
&\qquad\qquad\left.+\sqrt{q_0}\,x+\sqrt{q_1-q_0}\,z\Big)\Big)\right]^m.
\end{aligned}$$

Using the property of Heaviside function $\Theta\left(\prod_{s=1}^{2}\varphi(a_s)\right)=\prod_{s=1}^{2}\Theta\left(\varphi(a_s)\right)+\prod_{s=1}^{2}\Theta\left(-\varphi(a_s)\right)$ and performing some 2D rotations of variables one gets a simplified expression of the energetic term

$$G_E = \frac{1}{m}\log\int\prod_s Du_s\left[\int Dz\,F_\varphi(a(u_1,u_2,z),a(u_1,-u_2,z);\sigma)\right]^m, \qquad (56)$$

where we have defined the quantities

$$a(u_1,u_2,z)\equiv\sqrt{\frac{p+q_0}{2}}u_1-\sqrt{\frac{p-q_0}{2}}u_2+\sqrt{q_1-q_0}z \qquad (57a)$$

$$\sigma\equiv\sqrt{1-q_1-p+q_0} \qquad (57b)$$

and the φ-dependent functions

$$F_\varphi(x,y;\sigma)\equiv 1-I_\varphi(x;\sigma)-I_\varphi(y;\sigma)+2I_\varphi(x;\sigma)I_\varphi(y;\sigma) \qquad (58a)$$

$$I_\varphi(x;\sigma)\equiv\int Dh\,\Theta\left[\varphi\left(\sigma h+x\right)\right]. \qquad (58b)$$

With such simplifications, $\phi_m(Q,\hat{Q})$ can be optimized numerically, for each value of the external parameters m, α, p.

A.7.1 Explicit Formulas for Some Activation Functions

We list here the expression of the Kernel $I_\varphi(x;\sigma)$ defined in (58b) for some case of interest.

- Asymmetric Binary Perceptron (ABP): $\varphi(h)=h$. One gets

$$I_{\text{ABP}}(x;\sigma)=H\left(-\frac{x}{\sigma}\right) \qquad (59)$$

- Symmetric Binary Perceptron (SBP): $\varphi_\kappa(h)=\kappa-|h|$. One finds

$$I_{\text{SBP}}(x;\sigma)=H\left(-\frac{\kappa+x}{\sigma}\right)-H\left(\frac{\kappa-x}{\sigma}\right) \qquad (60)$$

- Square Wave (SWP): $\varphi_\delta(h) \equiv \text{sign}\left[\sin\left(\frac{\pi}{\delta}h\right)\right]$. The kernel I_φ can be efficiently computed by Fourier expansion

$$I_{\text{SWP}}(x;\sigma) = \frac{1}{2} + \frac{2}{\pi}\sum_{n=0}^{\infty} \frac{\mathrm{e}^{-\frac{\pi^2\sigma^2}{2\delta^2}(2n+1)^2}}{2n+1} \sin\left(\frac{\pi(2n+1)}{\delta}x\right). \tag{61}$$

  Note that for $\lim_{\delta\to\infty}\varphi_\delta(h) = h$, so that the Kernel tends in this limit to the one of the ABP.

B On the Existence of Collisions

Recall the definitions of the entropy of collisions with internal overlap q_1

$$\Phi(q_1) = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\boldsymbol{A}} \log Z(q_1;\boldsymbol{A}) \tag{62}$$

where

$$Z(q_1;\boldsymbol{A}) \equiv \sum_{\boldsymbol{c}\in\{\pm1\}^{2N}} \mathbb{X}_{\boldsymbol{A}}(\boldsymbol{c}) \prod_{k=1}^{m} \delta\left(q_i(\boldsymbol{c}) - q_1\right). \tag{63}$$

This section deals with the computation of the annealed and RS upper bounds $\Phi \le \Phi^{\text{RS}} \le \Phi^{\text{a}}$. As we will see, they coincide. This serves as an indication that the RS estimate is exact $\Phi = \Phi^{RS}$, although a proof would require a second moment calculation, which is outside the scope of this work.

B.1 First Moment Computation

The first moment bound on the existence of collisions can be obtained by computing the so called *annealed* free entropy:

$$\Phi^{\text{a}}(q_1) \equiv \frac{1}{N}\log \mathbb{E}_{\boldsymbol{A}} Z(q_1;\boldsymbol{A}) \ge \Phi(q_1). \tag{64}$$

This quantity can be directly obtained from the expression of $\phi^{\text{a}}_m(p,q_1)$ in Section A.3 by setting $m = 1$. In this limit, the dependence on the overlap p between the m collision clones disappears. One therefore gets for large N

$$\Phi^{\text{a}}(q_1) = \min_{\hat{q}_1}\left[G_S(q_1,\hat{q}_1) + \alpha G_E(q_1)\right] \tag{65a}$$

$$G_S(q_1,\hat{q}_1) = -q_1\hat{q}_1 + 2\log 2 + \log\cosh(\hat{q}_1) \tag{65b}$$

$$G_E(q_1) = \log\int Dz\, F_\varphi\left(\sqrt{q_1}z, \sqrt{q_1}z; \sqrt{1-q_1}\right). \tag{65c}$$

The saddle point equation for $\hat{q}_1$ can be easily solved:

$$\hat{q}_1 = \operatorname{atanh} q_1 = \frac{1}{2}\log\left(\frac{1+q_1}{1-q_1}\right) \tag{66}$$

which substituted in (65) gives

$$\Phi^{\text{a}}(q_1) = \log 2 + H_B(q_1) + \alpha\log\int Dz\, F_\varphi\left(\sqrt{q_1}z, \sqrt{q_1}z; \sqrt{1-q_1}\right), \tag{67}$$

$H_B(q_1)$ being the Bernoulli entropy with magnetization q_1 defined as

$$H_B(q_1) = -\frac{1-q_1}{2}\log\left(\frac{1-q_1}{2}\right) - \frac{1+q_1}{2}\log\left(\frac{1+q_1}{2}\right). \tag{68}$$

Because of the Markov's inequality the probability that the random variable $Z(\boldsymbol{A}) \geq 1$ is bounded by

$$P\left[Z(q_1;\boldsymbol{A}) \geq 1\right] \leq \mathbb{E}_{\boldsymbol{A}} Z(q_1;\boldsymbol{A}) = e^{N\Phi^{\rm a}(q_1)}. \tag{69}$$

The right hand side goes to zero for $N \to 0$ when $\Phi^{\rm a}(q_1) < 0$, i.e. when $\alpha \geq \alpha_c^{\rm A}(q_1)$ where $\alpha_c^{\rm a}(q_1)$ is given by

$$\alpha_c^{\rm a}(q_1) = -\frac{\log(2) + H_B(q_1)}{\log \int Dz\, F_\varphi\left(\sqrt{q_1}z, \sqrt{q_1}z; \sqrt{1-q_1}\right)}. \tag{70}$$

This means that for $\alpha \geq \alpha_c^{\rm a}(q_1)$ with high probability there is no collision. Due to Markov's inequality this is only *upper bound* on the existence of collisions. We analyze below the behavior of $\alpha_c^{\rm a}(q_1)$ for two functions $\varphi(h)$ of interest:

- Asymmetric Binary Perceptron (ABP): $\varphi(h) = h$. One obtains

$$\alpha_c^{\rm a}(q_1) = -\frac{\log 2 + H_B(q_1)}{\log\left(\frac{1}{\pi}\arccos(-q_1)\right)} \tag{71}$$

  In the neighborhood of $q_1 \simeq 1$, this behaves as $\alpha_c^{\rm a} \simeq \frac{\pi \log 2}{\sqrt{2dq_1}}$ where $dq_1 = 1 - q_1$.

- Square Wave (SWP): If one fixes $\delta > 0$ one numerically finds $\alpha_c^{\rm a} \propto \frac{1}{\sqrt{1-q_1}}$ as $q_1 \to 1$. If one performs first the limit $\delta \to 0$ *before* $q_1 \to 1$ the divergence observed at strictly positive δ near $q_1 \simeq 1$ is removed. Indeed, since $I(x;\sigma) \to 1/2$ one simply finds

$$\alpha_c^{\rm a}(q_1) = 1 + \frac{H_B(q_1)}{\log 2} \tag{72}$$

  which is a monotonically decreasing function of q_1. In particular, for $q_1 \to 1$ one finds $\alpha_c^{\rm a} \to 1$.

B.2 Upper Bound on the Existence of Collisions in the Sub-Extensive Regime

Consider an activation function $\varphi(\bullet)$ that is always ± 1 except for a set of discontinuity points $\{\xi_0, \xi_1, \dots\}$. We assume that the distances between the discontinuity points remain greater than zero in the thermodynamic limit.

For large N, the number of collisions with inputs differing for a single bit flip is given by:

$$\mathbb{E}_{\boldsymbol{A}} Z\left(q_1;\boldsymbol{A}\right)|_{q_1=1-\frac{1}{N}} \approx N\, 2^N J_\varphi^P \approx N\, 2^N \left(1 - \frac{a_\varphi}{\sqrt{N}}\right)^P, \tag{73}$$

where

$$\begin{aligned} J_\varphi &= \int Dz_0 Dz_1 \Theta\Big(\varphi\left(z_0 + z_1/\sqrt{N}\right)\varphi\left(z_0 - z_1/\sqrt{N}\right)\Big) \approx 2\int_0^\infty Dz_1 \sum_n \int_{\xi_n + z_1/\sqrt{N}}^{\xi_{n+1} - z_1/\sqrt{N}} Dz_0 \approx \\ &\approx 1 - \frac{1}{\sqrt{N}}\frac{2}{\pi}\sum_n e^{-\xi_n^2/2}, \end{aligned} \tag{74}$$

and

$$a_\varphi = \frac{2}{\pi} \sum_n e^{-\xi_n^2/2}. \tag{75}$$

Therefore, the annealed free entropy is

$$\Phi^{\mathrm{a}} = \log 2 - a_\varphi \, P \, N^{-3/2}, \tag{76}$$

which leads to the following estimate for the annealed capacity:

$$\alpha_c^{\mathrm{a}} = \frac{\log 2}{a_\varphi} N^{1/2}. \tag{77}$$

B.3 The Replica Symmetric Ansatz

The quenched entropy (62) can be computed by using the *replica trick*

$$\mathbb{E}_{\boldsymbol{A}} \log Z(q_1; \boldsymbol{A}) = \lim_{m \to 0} \frac{1}{m} \log \mathbb{E}_{\boldsymbol{A}} Z^m(q_1; \boldsymbol{A}). \tag{78}$$

The computation proceeds by considering m to be an integer, averaging over the disorder $\boldsymbol{A}$ and then analytically continuing the result to $m \to 0$. All the calculations can therefore be reconducted to the one presented in Section A.3 for the m-cloned entropy with two key differences. Firstly, the number of clones is not fixed to an integer but it is sent to 0. Such an analytical continuation is usually performed by doing an '*ansatz*' over the structure of the overlap matrix q_{st}^{ab}. In this section we will employ the same ansatz as in Section A.7, which in the context of the replica method is called *Replica Symmetric ansatz* (RS). Secondly, because of the definition in (63), there is no constraint on the distance between the clones. As a result, within the RS ansatz, the overlap p should not be fixed but instead treated as a variational parameter to be optimized. One therefore finally gets to

$$\Phi^{\mathrm{RS}}(q_1) = \max_{p, q_0} \min_{\hat{q}_0, \hat{q}_1, \hat{p}} [G_S(q_0, q_1, p, \hat{q}_0, \hat{q}_1, \hat{p}) + \alpha G_E(q_0, q_1, p)] \tag{79a}$$

$$G_S = -q_1 \hat{q}_1 - p\hat{p} - q_0 \hat{q}_0 - (\hat{p} + \hat{q}_1 - \hat{q}_0) \tag{79b}$$

$$+ \int Du_1 Du_2 \log \left[2 \cosh \left(\sqrt{2(\hat{p} - \hat{q}_0)} \, u_1 \right) + 2 e^{2(\hat{q}_1 - \hat{q}_0)} \cosh \left(\sqrt{2(\hat{p} + \hat{q}_0)} \, u_2 \right) \right] \tag{79c}$$

$$G_E = \int \prod_s Du_s \log \int Dz \, F_\varphi(a(u_1, u_2, z), a(u_1, -u_2, z); \sigma) \, , \tag{79d}$$

where $a(u_1, u_2, z)$ and σ are the same quantities defined in (57).

Interestingly, we have found numerically that the optimization in Equation (79) always leads to $p = q_0 = 0$ for any value of q_1. One then obtains $\Phi^{\mathrm{RS}}(q_1) = \Phi^{\mathrm{a}}(q_1)$, which gives the same predictions for the threshold corresponding to the existence of collisions $\alpha_c^{\mathrm{RS}}(q_1) = \alpha_c^{\mathrm{a}}(q_1)$. This result does not in principle preclude the possibility that the global optimum of the free entropy in the $m \to 0$ limit requires Replica Symmetry Breaking (RSB). However, we have ultimately validated the critical capacity predictions by exhaustive search of the collisions with a given value of q_1. Despite the small values of N, the numerical results are in very good agreement with the theoretical predictions.

C Overlap Gap Transition

Given an instance of the data matrix $\boldsymbol{A}$, define as $\mathcal{S}_m(p, \epsilon; q_1, \boldsymbol{A})$ the set of all the m configurations $\boldsymbol{c}_1, \dots, \boldsymbol{c}_m \in \{\pm 1\}^{2N}$, such that the following properties hold:

1. for each $a \in [m]$, $\boldsymbol{c}_a$ satisfies the constraints (30), i.e. it is a collision;
2. each collision $\boldsymbol{c}_a$, $a \in [m]$ has an internal overlap q_1;
3. for each $a \neq b \in [m]$
$$p < q_e(\boldsymbol{c}_a, \boldsymbol{c}_b) < p + \epsilon \,. \tag{80}$$

We say that the $m-$OGP property holds if it is possible to find p and ϵ such that:

$$\Pr_{\boldsymbol{A}} \left[\mathcal{S}_m(p, \epsilon; q_1, \boldsymbol{A}) \neq \emptyset\right] \overset{N\uparrow\infty}{\longrightarrow} 0 \,. \tag{81}$$

We call by $\alpha_{\rm OGP}(m; q_1)$ the constraint density above which the m-OGP begins to hold for collisions constrained to have an internal overlap q_1.

The general procedure to be adopted to obtain the m-OGP transition from our analytical framework is the following. Fix the value of α, p and q_1. Solve the corresponding maximization procedure of equation (39). Then impose the following additional constraints on the values of p and α

$$p_m^\star(\alpha; q_1) = \text{argmin}_p \phi_m(p, q_1) \tag{82a}$$
$$\phi_m(p_m^\star(\alpha; q_1), q_1) = 0. \tag{82b}$$

Namely, the value of $p_m^\star$ at the $m-$OGP transition represents the value of external overlap at which m collisions should be disposed to have zero entropy. For $\alpha > \alpha_{\rm OGP}(m; q_1)$, therefore, since the entropy is a decreasing function of α, there exists a region of external overlap between collisions p around $p_m^\star(\alpha; q_1)$ for which the entropy is negative. In other words, it is possible to find $\epsilon > 0$ s.t. condition (81) is satisfied, signaling the presence of a *gap* in the overlaps.

We stress here that our procedure is only guaranteed to give an upper bound to the true value of the OGP transition, as we are using a first moment to estimate the entropy of the $m-$cloned system.

We also underline that the maximization in equation (39) should be carried over a generic structure of the admissible correlation matrices Q. This procedure is difficult to deal analytically for generic non-linearities φ and number of clones m. The next subsections are therefore organized as follows. In section C.1 we compute the 2-OGP transition in the SHL by computing the global maximum of the saddle point equations. We will see that this maximum is actually attained by a correlation matrix which satisfies the symmetric ansatz introduced in A.7. Finally in section C.2 we restrict the maximization procedure on matrices Q satisfying the symmetric ansatz, and find the upper bound to the OGP threshold for generic m for the SHL. The same procedure has been applied to compute the m-OGP threshold at finite δ.

C.1 2-OGP in the SHL

We start here to examine the emergence of 2-OGP in the SHL i.e. for $\delta \to 0$. We will compute the 2-OGP threshold for an arbitrary structure of the covariance matrix Q.

We shall decompose $\mathcal{F}(p, q_1)$ according to the number of 0 eigenvalues: we write

$$\mathcal{F}(p, q_1) = \bigcup_{k=0}^{2m-1} \mathcal{F}_k(p, q_1) \tag{83}$$

where $\mathcal{F}_k(p, q_1)$ the set of admissible matrices with exactly k eigenvalues equal to zero, the other $2m - k$ eigenvalues being larger than zero. We do not include $\mathcal{F}_{2m}(p, q_1)$ as this is obviously an empty set. We therefore write equation (39) as

$$\phi_m^\star(p, q_1) = \max_{k \in \{0, \dots, 2m-1\}} G_m^{(k)}(p, q_1) \tag{84}$$

$$G_m^{(k)}(p, q_1) = \max_{Q \in \mathcal{F}_k(p, q_1)} \phi_m(Q). \tag{85}$$

In the next subsections, we study the case $m = 2$ in detail, analyzing the form of $Q_k^\star$, i.e. the maximum on each of the sets $\mathcal{F}_k$. In the $m = 2$ case we parametrize the covariance matrix Q by

$$Q = \begin{pmatrix} 1 & q_1 & r_{xx} & r_{xy} \\ q_1 & 1 & r_{yx} & r_{yy} \\ r_{xx} & r_{yx} & 1 & q_1 \\ r_{xy} & r_{yy} & q_1 & 1 \end{pmatrix} \tag{86}$$

with the fixed distance over collisions that imposes $r_{xx} + r_{yy} = 2p$.

C.1.1 The Maximum on the $\mathcal{F}_0$ Set

When all the eigenvalues of Q are positive, in the strong hashing limit we have

$$G_E(Q_0^\star) \equiv G_E^{k=0} = -\frac{1}{2} \log 2 \tag{87}$$

for all admissible Q. One has to find Q by maximizing the $G_S(Q)$ term only. The maximum is found at

$$Q_0^\star(p, q_1) = \begin{pmatrix} 1 & q_1 & p & q_0 \\ q_1 & 1 & q_0 & p \\ p & q_0 & 1 & q_1 \\ q_0 & p & q_1 & 1 \end{pmatrix} \tag{88}$$

with a conjugate matrix

$$\hat{Q}_0^\star(p, q_1) = \begin{pmatrix} 0 & \hat{q}_1 & \hat{p} & 0 \\ \hat{q}_1 & 0 & 0 & \hat{p} \\ \hat{p} & 0 & 0 & \hat{q}_1 \\ 0 & \hat{p} & \hat{q}_1 & 0 \end{pmatrix} \tag{89}$$

where $\hat{p}$, $\hat{q}_1$ and q_0 are functions of p and q_1 that can be obtained by solving the following coupled equations

$$p = \frac{\cosh(2\hat{q}_1)\sinh(2\hat{p})}{1+\cosh(2\hat{p})\cosh(2\hat{q}_1)} \tag{90a}$$

$$q_1 = \frac{\cosh(2\hat{p})\sinh(2\hat{q}_1)}{1+\cosh(2\hat{p})\cosh(2\hat{q}_1)} \tag{90b}$$

$$q_0 = \frac{\sinh(2\hat{q}_1)\sinh(2\hat{p})}{1+\cosh(2\hat{p})\cosh(2\hat{q}_1)} \tag{90c}$$

Note that $Q_0^\star$ is a matrix that shares a structure similar to that obtained under a Symmetric ansatz, see section A.7.

Let's show that this is a maximum. We first notice that

$$\frac{\partial G_S}{\partial Q_{\alpha\beta}} = \hat{Q}_{\alpha\beta}\,, \tag{91}$$

which implies

$$\frac{\partial G_S}{\partial Q_{14}} = \frac{\partial G_S}{\partial Q_{23}} = 0\,. \tag{92}$$

Then we also have

$$\frac{\partial G_S}{\partial Q_{13}} = \frac{\partial G_S}{\partial Q_{24}} = \hat{p} \tag{93}$$

Writing $Q_{13} = r$ and $Q_{24} = 2p - r$ we have

$$\left.\frac{\partial G_S}{\partial r}\right|_{r=p} = \frac{\partial G_S}{\partial Q_{13}} - \frac{\partial G_S}{\partial Q_{24}} = 0 \tag{94}$$

Therefore $Q_0^\star$ is a maximum. As G_S is concave, this is the maximum. (90) can be found by taking the expression found for the entropic term G_S in the Symmetric ansatz, see section A.7, and imposing vanishing derivatives with respect to $\hat{p}$, $\hat{q}_1$ and q_0, with $m = 2$ and $\hat{q}_0 = 0$.

C.1.2 The Maximum on the $\mathcal{F}_1$ Set

If one eigenvalue of Q vanishes, the integration measure $\langle\bullet\rangle$ in (37c) is constrained on the subspace orthogonal to the zero mode. This means that it exists one linear combination of the variable of integration imposing such a constraint, i.e.

$$au_1 + bv_1 + cu_2 + dv_2 = 0\,. \tag{95}$$

A dangerous case could be $c = d = 0$ (or equivalently $a = b = 0$), then $v_1 = tu_1$: for some values of t the hashing in u_1 and v_1 is in phase and the integral $\int du_1 \Theta(\varphi(u_1)\varphi(tu_1))$ turns out to be different from $1/2$, even in the strong hashing limit (for example, take the easy case $t = 1$). However we note that, if $v_1 = tu_1$ the fact that $\langle(v_1)^2\rangle = \langle(u_1)^2\rangle = 1$ implies that $t = \pm 1$. But then the internal overlap $q_1 = \langle u_1 v_1\rangle = \pm 1$, which is violated by our constraint on q_1.

The other potentially dangerous case is $v_2 = au_1 + bv_1 + cu_2$. The constraint in (43) then imposes that there exist three spins $\tau_1, \tau_2, \tau_3 = \pm 1$ such that $c = \tau_1 + a\tau_2 + b\tau_3$. One can search

the most dangerous case by computing the small δ limit of

$$I(a,b) = \max_{\tau_1,\tau_2,\tau_3} \int Du_1 Du_2 Dv_1 Dv_2 \,\Theta\left(\varphi(u_1)\varphi(v_1)\right) \times \\ \Theta\left(\varphi(u_2)\varphi(v_2)\right) \delta\left(v_2 - a u_1 - b v_1 - (\tau_1 + a\tau_2 + b\tau_3) u_2\right). \quad (96)$$

Note that here we use independent Gaussian for u_1, u_2, v_1. In principle we should use the covariance matrix of these three variables, but as long as this covariance matrix is not singular, in the strong hashing limit $\delta \to 0$ the result is the same.

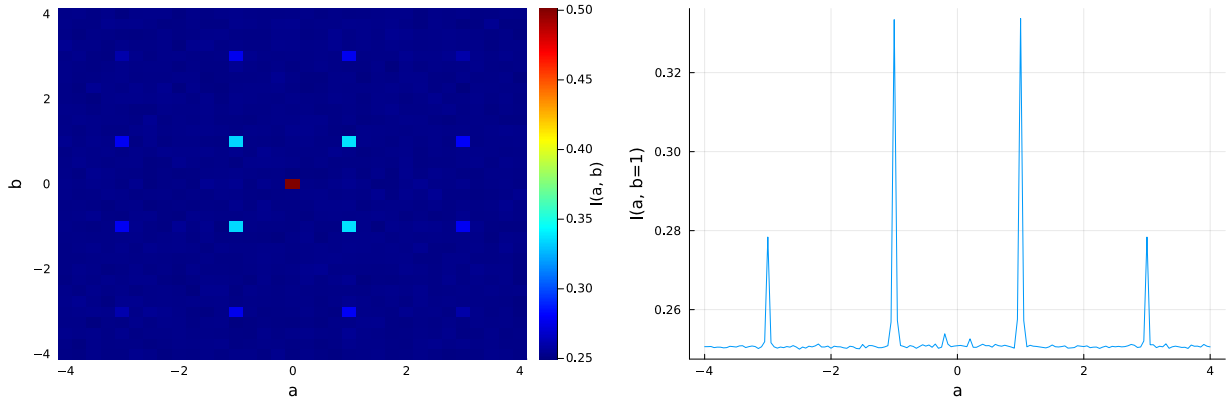

Figure 7: Left: heatmap plot of $I(a,b)$. The function is equal to $1/4$ almost everywhere except at some discrete value of a and b. Right: plot of $I(a,b)$ as a function of a with $b = 1$. $I(a,b)$ is evaluated by sampling 10^6 random Gaussian variables u_1, u_2 and u_3 for each of the possible combinations of τ_1, τ_2 and τ_3 and then computing the maximum over those variables. Here we have also considered $\delta = 0.1$. Note that $I(a,b) = 1/2$ for $a = b = 0$ and $1/3$ for $a = \pm 1$ and $b = \pm 1$.

A numerical study of this integral, see Figure 7, allows us to conjecture that the maximum value is obtained when $a = \pm 1$, $b = \pm 1$ (the global maximum located at $a = b = 0$ is again not admissible as examined before). By symmetry we can study one of these four cases, say $a = -1$, $b = 1$, $c = 1$. The singular eigenvalue enforces the constraint $v_2 = u_2 + v_1 - u_1$ or

$$v_2 - u_2 = v_1 - u_1. \quad (97)$$

We can compute the value of the energetic term corresponding to $Q_1^\star$. In this case calling $g = u_1 - v_1$ one finds

$$I(-1,1) = \int Dg\, \mathcal{K}^2(g) \quad (98)$$

with

$$\mathcal{K}(g) \equiv \int Du\, \Theta(\varphi(u)\varphi(u+g)). \quad (99)$$

The function $\mathcal{K}(g)$ is periodic of period 2δ, and its value in the interval $[-\delta, \delta]$ is $K(g) = 1 - \frac{|g|}{\delta}$. Therefore $I(-1,1) = 1/3$, and

$$G_E^{k=1} = \frac{1}{2} \log I(-1,1) = -\frac{1}{2} \log 3. \quad (100)$$

We need to check whether there exists an admissible covariance matrix with this singular eigenvector, and what is the entropy associated with this matrix. We use the parametrization in (86) for the covariance matrix. The constraint $v_2 = u_2 + v_1 - u_1$ imposes that the elements in the last column $r_{xy} = \langle u_1 v_2 \rangle$, $r_{yy} = \langle v_1 v_2 \rangle$, $q_1 = \langle u_2 v_2 \rangle$, $\langle (v_2)^2 \rangle = 1$ of the matrix can be expressed in terms of r_{xx}, r_{yx}. This gives:

$$r_{xy} = r_{xx} + q_1 - 1 \tag{101a}$$
$$r_{yy} = r_{yx} + 1 - q_1 \tag{101b}$$
$$q_1 = 1 + r_{yx} - r_{xx} \tag{101c}$$
$$1 = q_1 + r_{yy} - r_{xy} \tag{101d}$$

which give

$$r_{xx} = r_{yy} = p \tag{102a}$$
$$r_{xy} = r_{yx} = p + q_1 - 1. \tag{102b}$$

Altogether, we have the following family of dangerous admissible covariance matrices parametrized by q_1 and p which have one eigenvalue equal to 0, implementing the constraint in (97)

$$Q_1^\star(p, q_1) = \begin{pmatrix} 1 & q_1 & p & p+q_1-1 \\ q_1 & 1 & p+q_1-1 & p \\ p & p+q_1-1 & 1 & q_1 \\ p+q_1-1 & p & q_1 & 1 \end{pmatrix} \tag{103}$$

which again has the same structure of the ones of the symmetric ansatz.

C.1.3 The Maximum on the $\mathcal{F}_2$ Set

When two eigenvalues are zero, we integrate over a plane in the space u_1, v_1, u_2, v_2. We first study the case when this plane is in a generic direction, so that

$$a u_1 + b v_1 - u_2 = 0 \tag{104a}$$
$$c u_1 + d v_1 - v_2 = 0. \tag{104b}$$

Using $\langle (u_1)^2 \rangle = \langle (v_1)^2 \rangle = 1$ and $\langle (u_1) \rangle = q_1$ we get

$$r_{xx} = a + b q_1 \tag{105a}$$
$$r_{yy} = c q_1 + d \tag{105b}$$
$$r_{xy} = c + d q_1 \tag{105c}$$
$$r_{yx} = a q_1 + b \tag{105d}$$
$$q_1 = ac + bd + (bc + ad) q_1 \tag{105e}$$
$$1 = a^2 + b^2 + 2ab q_1 \tag{105f}$$
$$1 = c^2 + d^2 + 2cd q_1. \tag{105g}$$

We also need to implement the constraints in (43) for each of the eigenvectors with zero eigenvalue. In the present case, there are two such eigenvectors characterized by (104). Let us start with the

first one, i.e. $(a, b, -1, 0)$; there should exist τ_1, $\tau_2 \in \pm 1$ such that $a = \tau_1 + b\tau_2$. Using (105f) one finds that either $b = 0$ and $a = \pm 1$ which implies $r_{xx} = \pm 1$, $r_{yx} = \pm q_1$, or $a = 0$, $b = \pm 1$ which gives $r_{xx} = \pm q_1$, $r_{yx} = \pm 1$.

A similar analysis can be carried out for the second eigenvector $(c, d, 0, -1)$. Using (105g) one finds that either $c = 0$, $d = \pm 1$ which gives $r_{yy} = \pm 1$, $r_{xy} = \pm q_1$ or $d = 0$, $c = \pm 1$ which gives $r_{yy} = \pm q_1$ and $r_{xy} = \pm 1$. Putting those results together and using (105e) one finds either that $q_1 = \pm 1$, or that p needs to assume special values $p = 1$ which are forbidden.

Planes in a non-generic direction are parallel to some of the axes, but then they need to have at least one of the components which vanishes, say $u_1 = 0$. But this is excluded by $\langle (u_1)^2 \rangle = 1$. Therefore $\mathcal{F}_2 = \emptyset$.

C.1.4 The Maximum on the $\mathcal{F}_3$ Set

When three eigenvalues are zero, we integrate over a line. We can parametrize it as:

$$u_1 = t\,, \quad v_1 = at\,, \quad u_2 = bt\,, \quad v_2 = ct \tag{106}$$

$\langle (u_1)^2 \rangle = \langle (v_1)^2 \rangle = 1$ implies that $a = \pm 1$, but then we get $q_1 = a = \pm 1$ which is excluded. Therefore $\mathcal{F}_3 = \emptyset$.

C.1.5 The Global Maximum

In order to study the global maximum, we have to compare the entropy attained by using the two matrices $Q_0^\star$ and $Q_1^\star$ respectively obtained imposing zero and one vanishing eigevalues. Notice that $G_E^{k=0} \geq G_E^{k=1}$ but since $r_{xy} = r_{yx} \neq 0$ for $Q_1^\star$, the entropic terms $G_S^{k=0} \leq G_S^{k=1}$. However, in order to find the value of $\alpha_{\rm OGP}(q_1)$ one has to find the value of p where the entropy is minimal. This is attained in both the $k = 0, 1$ cases when $p \to 1$. This should be expected, as the entropic term is always a monotonically decreasing function of the overlap between collisions p; indeed if one constraints more the distance (namely p is larger) one has less entropy. When there is no dependence on the energetic term on p, therefore, the minimal entropy is found for $p \to 1$. In this limit one finds

$$Q_0^\star = Q_1^\star = \begin{pmatrix} 1 & q_1 & p & q_1 \\ q_1 & 1 & q_1 & p \\ p & q_1 & 1 & q_1 \\ q_1 & p & q_1 & 1 \end{pmatrix} \tag{107}$$

Indeed for $k = 0$ equations (90) in the limit $p \to 1$ give $\hat{p} \to \infty$, $\hat{q}_1 = \frac{1}{2}\text{arctanh}(q_1)$ and $q_0 = \tanh(2\hat{q}_1) \to q_1$. Therefore the entropic terms become equal in this limit to

$$\lim_{p \to 1} G_S^{k=0}(p, q_1) = \lim_{p \to 1} G_S^{k=1}(p, q_1) = \frac{1}{2}\log(2) + \frac{1}{2} H_B(q_1) \tag{108}$$

Therefore the global maximum in the $p \to 1$ is given by covariance matrices $Q \in \mathcal{F}_1$. The $2-$OGP transition is finally given by

$$\lim_{p \to 1} \phi_{m=2}^\star(p, q_1) = 0 \implies \alpha_{\rm OGP}(m = 2; q_1) = \frac{\log(2) + H_B(q_1)}{\log 3} \tag{109}$$

C.2 Analysis of the SHL in the Symmetric Ansatz for Generic m

C.2.1 Hubbard-Stratonovich

Here we compute the energetic term in the symmetric ansatz for generic m, in the limit $\delta \to 0$. We will use the Hubbard-Stratonovich representation given in (55) that we rewrite for convenience:

$$G_E = \frac{1}{m} \log \int \prod_s Du_s Dx \left[\int Dz \prod_s Dh_s \, \Theta \left(\prod_s \varphi(\sqrt{1-q_1-p+q_0} h_s + \sqrt{p-q_0} u_s + \sqrt{q_0} x + \sqrt{q_1-q_0} z) \right) \right]^m . \tag{110}$$

Note that the term $1 - q_1 - p + q_0$ appears explicitly as one of the terms under the square roots. When this quantity vanishes, it gives exactly the condition on q_0 that was explicitly derived for the case $m = 2$, see (102b), and that was related to the presence of a singular eigenvalue in the covariance matrix Q, as given in (97). The case of $\sigma^2 = 1 - q_1 - p + q_0 \sim 0$ must therefore be handled with care.

If $1 - q_1 - p + q_0 > 0$, we can actually use the G_E given in (56) with the kernel $I(m, \sigma)$ given by (61). As $\sigma^2 = 1 - q_1 - p + q_0 > 0$ all the terms in the series vanish exponentially fast and give $\lim_{\delta \to 0} I_{\mathrm{SWV}}(m, \sigma) = 1/2$. This leads to a constant G_E

$$G_E = -\log 2, \quad 1 - q_1 - p + q_0 > 0. \tag{111}$$

When $1 - q_1 - p + q_0 = 0$, the determinant of Q vanishes, the Gaussian measure in (37c) becomes singular, and the integration is effectively reduced to an integration over a lower-dimensional space. Plugging into (110) the expression $q_0 = q_1 + p - 1$, one obtains

$$G_E = \frac{1}{m} \log \int \prod_s Du_s Dx \left[\int Dz \, \Theta \left(\prod_s \varphi(m_s + \sqrt{1-p} z) \right) \right]^m , \tag{112}$$

where $m_s := \sqrt{1-q_1} u_s + \sqrt{p+q_1-1}\, x$. The kernel $\mathcal{K}(m_1, m_2) := \int Dz \, \Theta \left(\prod_s \varphi(m_s + \sqrt{1-p}\, z) \right)$ is translationally invariant due to the periodicity of the SWP activation and therefore depends only on $g := m_2 - m_1$. Moreover, it is periodic in g with period 2δ. Defining

$$\mathcal{K}(g) := \int Dz \, \Theta \left(\varphi(\sqrt{1-p} z) \varphi(g + \sqrt{1-p} z) \right)$$

one obtains,

$$G_E = \frac{1}{m} \log \int Dg \, \mathcal{K}^m \left(\sqrt{2(1-q_1)} g \right). \tag{113}$$

Within the Kernel, the Θ function is also periodic in z, with period δ. To compute the Kernel in the small δ limit, it is convenient to expand it in Fourier series

$$\Theta \left(\varphi(t) \varphi(g+t) \right) = \frac{1}{2} a_0(g) + \sum_{i=1}^{\infty} a_i(g) \cos \left(\frac{2\pi n}{\delta} t \right) + \sum_{i=1}^{\infty} b_i(g) \sin \left(\frac{2\pi n}{\delta} t \right), \tag{114}$$

where

$$a_0(g) = \frac{2}{\delta}\int_0^\delta \Theta\left(\varphi(t)\varphi(g+t)\right) dt\,, \tag{115a}$$

$$a_i(g) = \frac{2}{\delta}\int_0^\delta \Theta\left(\varphi(t)\varphi(g+t)\right) \cos\left(\frac{2\pi n}{\delta}t\right) dt\,, \tag{115b}$$

$$b_i(g) = \frac{2}{\delta}\int_0^\delta \Theta\left(\varphi(t)\varphi(g+t)\right) \sin\left(\frac{2\pi n}{\delta}t\right) dt. \tag{115c}$$

Upon integration with the Gaussian, only the zero mode gives a finite contribution in the $\delta \to 0$ limit, so that

$$\mathcal{K}(g) = \int Dz\,\Theta\left[\varphi\left(\sqrt{1-p}z\right)\varphi\left(g+\sqrt{1-p}z\right)\right] = \frac{1}{2}a_0(g) = 1 - \frac{|g|}{\delta}\,, \qquad g \in [-\delta,\delta]\,. \tag{116}$$

This profile repeats periodically with a period 2δ. The Gaussian integral in Equation (113) can be similarly computed by decomposing $\mathcal{K}(g)$ in Fourier modes, and keeping only the zero mode. This leads to a different constant value of G_E in the $\delta \to 0$ limit, for points on the boundary $1 - q_1 - p + q_0 = 0$:

$$G_E = -\frac{1}{m}\log(1+m), \quad 1 - q_1 - p + q_0 = 0. \tag{117}$$

Since both the energetic term are going to a constant, the maximum of the entropy is attained for values of the overlaps on the boundary $1-q_1-p+q_0 = 0$. The optimum of the entropic contribution $G_I + G_S$ is found at $p = 1$, implying that $\hat{p} \to \infty$, $q_0 \to q_1$ and $\hat{q}_0 \to \hat{q}_1$, giving

$$G_S = -mq_1\hat{q}_1 + \frac{2\log 2}{m} + \frac{1}{m}\log\cosh(m^2\hat{q}_1). \tag{118}$$

One can then find the solution of $\hat{q}_1$ as a function of q_1

$$\hat{q}_1 = \frac{1}{m^2}\,\mathrm{atanh}(q_1) = \frac{1}{2m^2}\log\left(\frac{1+q_1}{1-q_1}\right). \tag{119}$$

Putting all together, the result of the optimization in the $\delta \to 0$ limit is

$$\phi_m(q_1) = \frac{\log 2 + H_B(q_1)}{m} - \alpha\frac{\log(1+m)}{m} \tag{120}$$

which gives

$$\alpha^m_{\mathrm{OGP}}(q_1) = \frac{\log 2 + H_B(q_1)}{\log(1+m)} \tag{121}$$

which is monotonically decreasing in m and vanishes for $m \to \infty$ for every q_1.

C.2.2 Fourier Representation

In this section, we show that in the SHL the same result for the energetic term (117) that was derived from the Hubbard-Stratonovich representation, can be obtained starting from its Fourier

representation (51), whichcovers all the domain where the covariance matrix Q is positive definite. In the symmetric ansatz, the Fourier representation of the energetic term (51) becomes

$$G_E = \frac{1}{m} \ln \left[\frac{1}{2^m} + \frac{1}{2^m} \sum_{r=1}^{m} \binom{m}{r} \left(-\frac{4}{\pi^2} \right)^r \sum_{\boldsymbol{n} \in \mathbb{Z}_{\text{odd}}^{2r}} \frac{1}{\prod_{i=1}^{2r} n_i} e^{-\frac{\pi^2}{2\delta^2} \boldsymbol{n}^T Q_r \boldsymbol{n}} \right], \tag{122}$$

where Q_r is the covariance matrix of r clones in the symmetric ansatz and we have denoted by $\mathbb{Z}_{\text{odd}}^{2r}$ the set $\mathbb{Z}^{2r}$ where each entry is an odd number. In the strong hashing limit $\delta \to 0$, for each r the sum over $\boldsymbol{n}$ is dominated by the eigenvectors of Q_r. When $r = 1$, $Q_{r=1}$ has eigenvalues $1 - q_1$ and $1 + q_1$. Since we are assuming $1 < q_1 < 1$, the $r = 1$ will always vanish in the strong hashing limit.

For $r > 1$, Q_r has eigenvalues

$$\lambda_1 = 1 - p + q_0 - q_1 \,, \qquad d_1 = r - 1 \tag{123a}$$

$$\lambda_2 = 1 - p - q_0 + q_1 \,, \qquad d_2 = r - 1 \tag{123b}$$

$$\lambda_3 = 1 + (r-1)p - (m-1)q_0 - q_1 \,, \qquad d_3 = 1 \tag{123c}$$

$$\lambda_4 = 1 + (r-1)p + (m-1)q_0 + q_1 \,, \qquad d_4 = 1. \tag{123d}$$

We call the corresponding eigenvectors respectively as $\boldsymbol{\eta}^{1,s}$, $\boldsymbol{\eta}^{2,s}$, with $s \in [r-1]$ and $\boldsymbol{\eta}^3$ and $\boldsymbol{\eta}^4$. Their expression is

$$\eta_i^{1,s} = \begin{cases} (-1)^{i+1} & i = 1, 2, 2s+1, 2s+2 \\ (-1)^i & i = 2s+1, 2s+2 \\ 0 & \text{otherwise} \end{cases} \tag{124a}$$

$$\eta_i^{2,s} = \begin{cases} -1 & i = 1, 2 \\ 1 & i = 2s+1, 2s+2 \\ 0 & \text{otherwise} \end{cases} \tag{124b}$$

$$\eta_i^3 = (-1)^i \,, \qquad i \in [2r] \tag{124c}$$

$$\eta_i^4 = 1 \,, \qquad i \in [2r]. \tag{124d}$$

Depending on which eigenvalue vanishes, the energetic term goes to a different finite limit as $\delta \to 0$. We therefore analyze the contribution of each eigenvalue and finally select the one that gives the maximum contribution.

We start by considering the non-degenerate eigenvalues, which are easier to analyze. If we suppose for example that $\lambda_3 = 0$, the index $\boldsymbol{n} \in \mathbb{Z}_{\text{odd}}^{2r}$ can be written as $\boldsymbol{n} = k\boldsymbol{\eta}^3$, where $k \in \mathbb{Z}_{\text{odd}}$. The argument of the series (122) is therefore $\prod_{i=1}^{2r} n_i = k^{2r} \prod_{i=1}^{2r} (-1)^i = k^{2r}$. It is easy to see that the case $\lambda_4 = 0$, gives the same result.

Therefore if either λ_3 or λ_4 vanish, in the SHL one obtains

$$\lim_{\delta \to 0} \sum_{\boldsymbol{n} \in \mathbb{Z}_{\text{odd}}^{2r}} \frac{1}{\prod_{i=1}^{2r} n_i} e^{-\frac{\pi^2}{2\delta^2} \boldsymbol{n}^T Q_r \boldsymbol{n}} = \sum_{k \neq 0} \frac{1}{(2k+1)^{2r}} = 2(1 - 4^{-r}) \zeta(2r), \tag{125}$$

where ζ is the Riemann zeta function. The energetic term reads in this case as

$$\lim_{\delta \to 0} G_E(q_1) = \frac{1}{m} \ln \left[\frac{1}{2^m} + \frac{1}{2^{m-1}} \sum_{r=2}^{m} \binom{m}{r} \frac{(-1)^r (4^r - 1)}{\pi^{2r}} \zeta(2r) \right]. \tag{126}$$

Let us now continue analyzing just the case $\lambda_1 = 0$. The only difference with respect to the non-degenerate eigenvalues is that now the index $\boldsymbol{n}$ can be written as a linear combination of the $r-1$ eigenvectors:

$$\boldsymbol{n} = \sum_{s=1}^{r-1} k_s \boldsymbol{\eta}^{3,s} \tag{127}$$

k_s, $s \in [r-1]$ should be chosen in such a way that *every entry* in $\boldsymbol{n}$ is an odd number. Due to the structure of the $r-1$ eigenvectors $\boldsymbol{\eta}^{3,s}$ it is easy to see that we have to use at least once each eigenvector, namely $k_s \neq 0$, $\forall s \in [r-1]$.

If r is odd, however it is impossible to satisfy this constraint and at the same time to maintain all the components of $\boldsymbol{n}$ even. We therefore conclude that if r is odd the contribution to the sum always vanishes.

If r is even the constraint that $\boldsymbol{n} \in \mathbb{Z}^{2r}_{\text{odd}}$ simply imposes that k_s is odd for all $s \in [r-1]$. If r is even, then we get

$$\lim_{\delta\to 0} \sum_{\boldsymbol{n}\in\mathbb{Z}^{2r}_{\text{odd}}} \frac{1}{\prod_{i=1}^{2r} n_i} e^{-\frac{\pi^2}{2\delta^2}\boldsymbol{n}^T Q_r \boldsymbol{n}} = \sum_{\boldsymbol{k}\in\mathbb{Z}^{r-1}_{\text{odd}}} \frac{1}{\left(\prod_{i=1}^{r-1} k_i^2\right)\left(\sum_{i=1}^{r-1} k_i\right)^2} = \sum_{\boldsymbol{k}\in\mathbb{Z}^{r}_{\text{odd}}} \frac{\delta\left(\sum_{i=1}^{r} k_i\right)}{\prod_{i=1}^{r} k_i^2}. \tag{128}$$

Using the Fourier representation of the Kronecker delta function, we find

$$\sum_{\boldsymbol{k}\in\mathbb{Z}^{r}_{\text{odd}}} \frac{\delta\left(\sum_{i=1}^{r} k_i\right)}{\prod_{i=1}^{r} k_i^2} = \int_{-\pi}^{\pi} \frac{dx}{2\pi} \left(\sum_{k\in\mathbb{Z}_{\text{odd}}} \frac{e^{ikx}}{k^2}\right)^r = \frac{\pi^{2r}}{4^r(r+1)}\,, \tag{129}$$

where in the last step, we have used the following identity

$$\sum_{k\in\mathbb{Z}_{\text{odd}}} \frac{e^{ikx}}{k^2} = \frac{\pi^2}{4} - \frac{\pi}{2}|x|\,, \qquad x \in [-\pi,\pi]\,. \tag{130}$$

We therefore get:

$$\lim_{\delta\to 0} G_E(q_1) = \frac{1}{m} \ln\left[\frac{1}{2^m} + \frac{1}{2^m}\sum_{r=1}^{m} \binom{m}{2r} \frac{1}{2r+1}\right] = -\frac{1}{m}\ln(1+m). \tag{131}$$

The same argument can be done for λ_2; it is easy to check that it gives the same contribution as λ_1. In Fig. 8 we show that the maximum value of the energetic term is always attained by the degenerate eigenvalues λ_1 and λ_2.

D Generalized Reverse Wedge Perceptron

It is interesting to see to what extent our results rely on the activation function being periodic. We therefore applied the same formalism to the 'Generalized Reverse Wedge Perceptron' (GRWP) activation function, namely a Square Wave Perceptron where oscillations are limited within a finite window. The precise shape of the activation we are considering is:

$$\varphi(h) = \prod_{l=-K}^{K} \left(h + \frac{l\gamma}{K}\right), \tag{132}$$

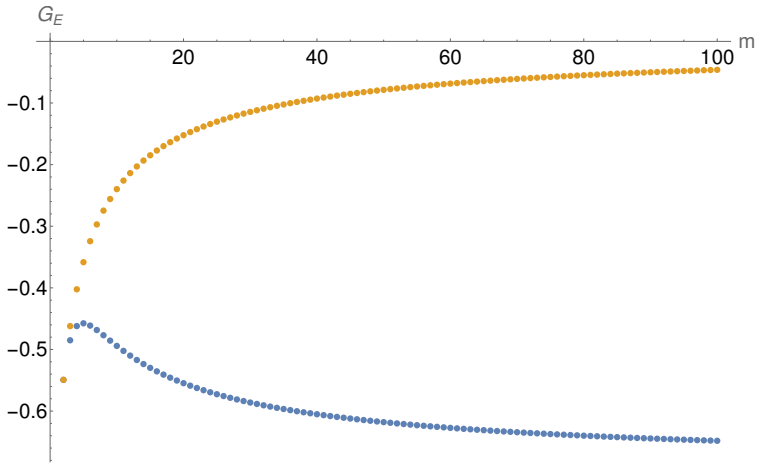

Figure 8: Energetic term corresponding to the vanishing of λ_1 or λ_2 (orange) and to the vanishing of λ_3 or λ_4 (blue) as a function of m. Note that for $m = 2$ is the only case where all the contributions coincide.

where $2K$ is the number of oscillations within $[-\gamma, \gamma]$. In Figure 9 you can find the comparison between the compression rate achievable using a GRWP with $K = 5$ and $\gamma = 2.5$ and its fully periodic counterpart, namely a SWP with $\delta = 0.5$. As intuitive, having an activation with constant plateaus makes finding collisions easier, leading to a higher (worse) compression rate. Still, the difference between the two curves is small, they both render a CRH when composed with a code with q_c in a window around $q_c = 0.98$.

In Figure 10 you can find the values of α_{OGP} for the collision-finding problem for $q_1 = 0.05$ and $K = 5$, as a function of γ. Notice that these low values of q_1 determine whether composing the network with an appropriate code will lead to a CRH or not. For the analyzed values, decreasing γ at fixed K, i.e. increasing the frequency of the oscillation, leads to a decrease of α_{OGP}. For $\gamma < 2.5$ α_{OGP} is sufficiently low to give a CRH (compare it with the corresponding values for the SWP in Figure 3).

Notice that the standard deviation of the Gaussian fields h is normalized to one in our setting. Hence, $\gamma < 2.5$ means that a substantial fraction of the Gaussian mass lies outside of the oscillating region of the activation function, and as a result the activation can be considered as bona fide non-periodic. This has interesting consequences in terms of comparisons with Ajtai's function, as detailed in Section 5.3.

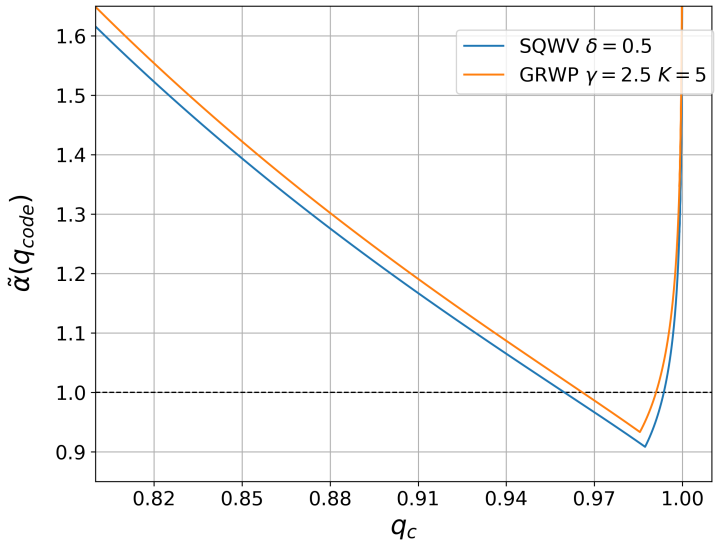

Figure 9: Comparison between the compression rate achievable using a GRWP with $K = 5$ and $\gamma = 2.5$ and a SWP with $\delta = 0.5$. An activation with constant plateaus leads to a higher compression rate. Both activations render a CRH when composed with a code with q_c in a window around $q_c = 0.98$.

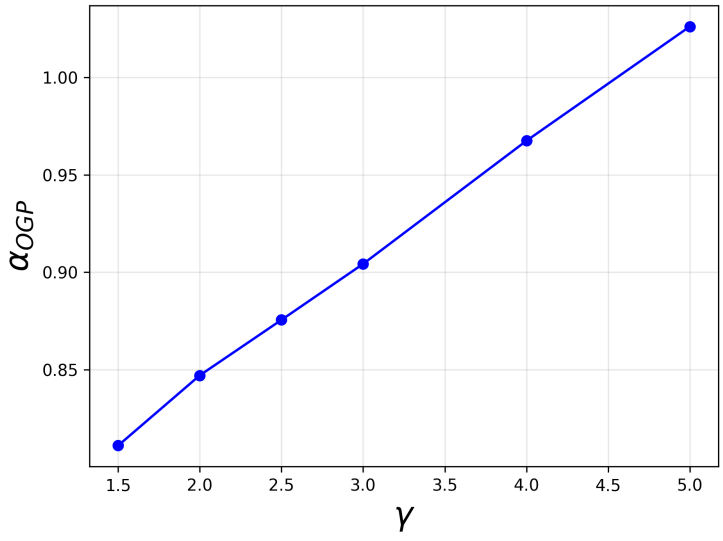

Figure 10: For the GRWP, α_{OGP} for the collision-finding problem for $q_1 = 0.05$ and $K = 5$, as a function of γ. In the range of γ we explored, decreasing γ leads to a decrease of α_{OGP}. Note that since K fixes the number of oscillations, reducing γ increases their frequencies.