Collision Resistance of Single-Layer Neural Nets

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]Bocconi University
[2]University of Ottawa

June 2, 2026

Abstract

We initiate the study of the algorithmic complexity of finding collisions in single-layer binary neural networks. Given a random matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, an input $\mathbf{x} \in \{-1,1\}^n$ is mapped to a binary output vector $\varphi(\mathbf{Ax}) \in \{-1,1\}^m$, where φ is an activation function with constant behavior on $[\kappa,\infty)$ for some threshold $\kappa \geq 0$.

We identify the threshold scale $\kappa = \Theta(1/\sqrt{\alpha})$, where $\alpha = m/n$, as separating two complementary phenomena. When $\kappa \ll 1/\sqrt{\alpha}$, we give a simple online algorithm that efficiently produces extensive collisions. When $\kappa \gg 1/\sqrt{\alpha}$, for a natural *randomized* non-periodic activation and suitable oscillation complexity, we prove that the extensive-collision space exhibits an overlap gap property (OGP), yielding an exponential lower bound against online algorithms.

Ours is the first work to use the overlap gap property as a rigorous criterion for collision resistance. The key difference between collision finding and average-case search is that collision finding has a new "worst-case" aspect: the collision finder has full control over the choice of colliding pairs. Our lower bound is proved in the online model; extending such guarantees to broader classes of algorithms, including spectral, algebraic, lattice-based, or quantum methods, remains an open direction.

Keywords: statistical physics, cryptography, perceptron, multi overlap gap, teacher-student problem, online algorithms, second preimage attack, collision resistance

1 Introduction

Is it feasible to find two different inputs that map to the same output in a neural net? Finding such collisions is sometimes easy [14, 18], but the question is whether there exist combinations of input domains and activation functions where collision finding becomes hard.

*Part of the work was done while visiting Bocconi University, supported by European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (Grant agreement No. 101019547).

†Work supported by European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (Grant agreement No. 101019547) and Cariplo CRYPTONOMEX grant.

‡Most of the work was done while at Bocconi University, supported by European Research Council (ERC) under the EU's Horizon 2020 research and innovation programme (Grant agreement No. 101019547).

We consider a standard setup of a single-layer neural network with n binary inputs $\mathbf{x} \in \{-1,1\}^n$ and m binary outputs [19, 20, 16]. The outputs are obtained by multiplying $\mathbf{x}$ with a weight matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, followed by applying an activation map $\varphi : \mathbb{R}^m \to \{-1,1\}^m$:

$$f(\mathbf{x}) = \varphi(\mathbf{A}\mathbf{x}) \in \{-1,1\}^m .$$

A *collision* is a pair $\mathbf{x} \neq \mathbf{x}'$ such that $\varphi(\mathbf{A}\mathbf{x}) = \varphi(\mathbf{A}\mathbf{x}')$. We focus on the shrinking regime $m < n$, where collisions must exist by the pigeonhole principle. The question remains interesting even when $m \geq n$ as long as collisions persist. We restrict to discrete inputs since continuous domains permit trivial linear-algebraic collisions via the kernel of $\mathbf{A}$. We also assume that $\mathbf{A} \sim \mathcal{N}(0,1/n)^{m\times n}$ is given as input to, but not controlled by, the collision finder. It is sufficient to focus on a single-layer network since any collision found in one layer propagates through subsequent layers.

Collision resistance in single-layer binary networks depends both on the activation function and the constraint density $\alpha = m/n$. Our results identify a scaling window for collision finding, governed by the interaction between the activation's margin width κ and the density α. We analyze three representative activations under this lens:

- The *asymmetric binary perceptron* (ABP), $\varphi(z) = \mathrm{sign}(z)$.
- The *symmetric binary perceptron* (SBP), $\varphi(z) = \mathrm{sign}(\kappa - |z|)$.
- A *randomized oscillating activation*, where each neuron i uses an odd number K of independent random thresholds $t_{ij} \sim \mathrm{Unif}[-\kappa,\kappa]$,

$$\varphi_i(x) = \prod_{j=1}^{K} \mathrm{sign}(x - t_{ij}) .$$

  Here $\kappa > 0$ determines the interval of oscillation and the odd integer $K \geq 1$ determines its complexity. The oddness makes the activation change sign between the two tails, rather than returning to the same value on both sides of the oscillating window. This is the regime we analyze: the OGP and online lower-bound proofs use odd K in the row-wise contraction argument, and we do not treat even K. Each row i of the output uses its own randomly chosen thresholds, so the activations φ_i vary across coordinates. The thresholds are sampled once together with the instance and are then fixed and revealed; the activation is randomized only in the choice of the instance, not during evaluation.

When $\kappa \ll 1/\sqrt{\alpha}$, we design an online algorithm that finds *extensive collisions* with a sufficiently small constant overlap gap, for any activation that is constant on the positive tail $[\kappa,\infty)$. This includes both ABP ($\kappa = 0$) and SBP in the small-κ regime (for SBP, sufficiently positive margins all lie on the same constant tail, so the outputs coincide).

For the randomized oscillating activation, in the complementary large-width regime, we show that the collision space undergoes an r-wise *overlap gap property* (r-OGP). The OGP is a geometric criterion linked to algorithmic barriers in random inference and optimization [10, 8, 7, 21, 6]. In our setting, the relevant form is an *ensemble* r-OGP, which applies to several correlated instances sharing a common prefix; this rules out online algorithms for finding extensive collisions in the corresponding parameter regime. In related models, such gaps also align with known limitations of low-degree polynomial methods and message-passing heuristics [8, 7]. Our formal lower bound

here is for online algorithms; lower bounds for broader algorithmic models, including low-degree, message-passing, and quantum algorithms, remain open.

The SBP lies between these two regimes. For small κ, it falls within our positive-tail algorithmic window; for very large κ, the activation is nearly constant on typical Gaussian margins, and collisions become trivial. The difficult case is the intermediate window, where the positive-tail escape route no longer applies but no OGP-based hardness result is known for collision finding. Resolving whether SBP is collision resistant in this middle regime, for some $\alpha < 1$ and some κ, remains a central open challenge.

1.1 A Statistical Physics Perspective

The ABP is susceptible to a second preimage attack: sample $\mathbf{y} = \text{sign}(\mathbf{Ax})$ for random $\mathbf{x}$ and try to solve the so-called *teacher-student problem* on input $(\mathbf{A}, \mathbf{y})$. The output will be an $\mathbf{x}'$ such that, $\text{sign}(\mathbf{Ax}) = \mathbf{y} = \text{sign}(\mathbf{Ax}')$, which is a collision whenever $\mathbf{x}' \neq \mathbf{x}$. This occurs with probability at least $1/2$ for $m < n$ and with noticeable probability up to $\alpha \leq 1.1$. Approximate message passing (AMP) algorithms are known to heuristically solve the teacher-student problem in this regime [1].

Fact 1.1. *The ABP is (heuristically) not collision resistant for any $\alpha < 1.1$.*

The perceptron and its random variants have long served as a model system in statistical physics [11, 12]; see also the monograph [17]. A recent line of work studies the *square-wave perceptron* (SWP), where a periodic activation induces structured oscillations across the input domain [4, 3]. Using non-rigorous replica methods alongside first-moment arguments, it predicts an overlap gap at large densities and suggests a phase transition in collision-finding complexity.

Our work is complementary: we study non-periodic, *randomized* oscillating activations with odd K and prove a rigorous r-OGP, while also giving a positive online algorithm for any activation that is constant on the positive tail $[\kappa, \infty)$ in the small-width regime and for sufficiently small overlap gap. In particular, the randomized (non-periodic) structure we consider avoids the lattice-like regularity of the SWP. This motivates the conjecture that a single-layer neural network with random weights, binary inputs, and a randomized oscillating activation is collision resistant in the parameter regime where the OGP first moment is negative. In this regime, our ensemble OGP result proves an online lower bound.

1.2 Result I: An Online Collision Finder for Positive-Tail Activations

For activations that are constant on a positive tail, collision finding is algorithmically easy at low densities. Indeed, if $\varphi(x)$ is constant for all $x \geq \kappa$, then it suffices to drive all margins above κ; once there, the activation is flat and the outputs coincide. We design a simple online algorithm that efficiently produces two distinct inputs mapping to the same output pattern whenever α is below a constant threshold.

The algorithm operates in the *extensive regime*, where the two colliding inputs differ on a linear fraction of their coordinates. This is also the regime relevant for our hardness results later: both our positive and negative results concern collisions that are extensive rather than so-called *subextensive* collisions that differ in a vanishing fraction of coordinates.

Theorem 1.1 (Online Collision Finding at Low Density)**.** *There exists $\delta_0 > 0$ such that for every fixed $\delta \in (0, \delta_0)$ there are constants $\alpha_0 = \alpha_0(\delta) > 0$ and $c > 0$ such that, for every $\alpha < \alpha_0$ and every*

margin parameter $\kappa \le c/\sqrt{\alpha}$, there is an efficient online algorithm which, given a random matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m = \alpha n$, outputs (with constant probability) two inputs $\mathbf{x}, \mathbf{y} \in \{\pm 1\}^n$ satisfying:

$$\left|\mathbf{x}^\top \mathbf{y}\right| \le (1-\delta)\, n\,, \qquad \min_i (\mathbf{A}\mathbf{x})_i \ge \kappa\,, \qquad \min_i (\mathbf{A}\mathbf{y})_i \ge \kappa\,.$$

In particular, for any activation function φ that is constant on $[\kappa, \infty)$, we have $\varphi(\mathbf{A}\mathbf{x}) = \varphi(\mathbf{A}\mathbf{y})$, so $\mathbf{x}$ and $\mathbf{y}$ form a δ-extensive collision for φ.

Our algorithm works for both the ABP and the SBP for sufficiently small densities. For ABP, the same procedure, when run on a single vector, also finds a preimage of the all-positive output $\mathbf{1}^m$.

The algorithm's evolution is governed by an exponential potential that gives a single global measure of how many coordinates still have poor margin. In the collision-finding variant, we first create the desired overlap gap by assigning opposite signs to $\mathbf{x}$ and $\mathbf{y}$ on an initial block of coordinates. From that point on, every new coordinate is assigned the same sign in both vectors, so the overlap remains fixed while the margins continue to improve. The analysis then follows the pointwise minimum of the two margin vectors: a coordinate is considered safe only once both $\mathbf{A}\mathbf{x}$ and $\mathbf{A}\mathbf{y}$ are safely above the target margin. The *Push* phase steadily reduces the exponential potential of this minimum state, shrinking the low-margin tail. A final refinement phase focuses on the remaining bad coordinates and raises them above the threshold, ensuring that every entry of both $\mathbf{A}\mathbf{x}$ and $\mathbf{A}\mathbf{y}$ is at least κ.

This potential-based analysis gives a different way to study sequential margin-building algorithms. An earlier algorithm of Kim and Roche [13] also constructs signs in stages: at each stage it examines a fresh block of columns and chooses the corresponding signs by majority vote among the rows with the smallest current partial margins. Their proof tracks the evolution of this low-margin tail through the stages and shows that all margins become positive when $\alpha = O(1/\log^2 n)$, achieving minimum margin $\Omega(1/\log n)$. Our exponential-potential argument serves a similar purpose, but reduces the low-margin tail to a single quantity that is easier to adapt to the collision-finding setting.

A recent algorithm of Schramm et al. [15] gives a sharper result for the single-preimage margin problem using tools from discrepancy minimization. Their method first solves a linear program over the relaxed cube $[-1,1]^n$, with a random objective direction and the desired margin constraints, and then rounds the remaining fractional coordinates while preserving feasibility. For every constant $c < 2/\pi$ and all sufficiently large κ, it finds $\mathbf{x}$ with $\mathbf{A}\mathbf{x} \ge \kappa\mathbf{1}$ whenever $\alpha \le c/\kappa^2$, matching the large-margin statistical threshold up to the sharp constant. Our potential-based method has a less optimized constant for this single-preimage task, but it is tailored to collision finding: by tracking the pointwise minimum of two evolving margin vectors, it constructs two correlated inputs with prescribed extensive overlap, rather than a single preimage of $\mathbf{1}^m$. Whether the sharper discrepancy-minimization approach can be adapted to this collision-finding setting remains open.

1.3 The Symmetric Binary Perceptron

The SBP exposes the main limitation of the positive-tail framework. Our online algorithm applies whenever $\kappa \ll 1/\sqrt{\alpha}$, yielding efficient collisions with a sufficiently small constant overlap gap in the small-κ regime. Separately, in extreme parameter regimes where the output can be saturated to $\pm\mathbf{1}^m$, earlier margin-based algorithms [2, 13, 15] also produce collisions. Beyond these extremes,

however, the collision resistance of SBP remains unresolved. Neither known algorithmic techniques nor hardness arguments currently apply when κ lies in an intermediate window, such as $\kappa \approx 0.675$ and $\alpha \approx 0.9$, where statistical physics suggests nontrivial structure may emerge.

One insight into the structure of SBP is provided by Gamarnik et al. [9], who show that in the teacher–student version of SBP, the solution space satisfies an m-overlap gap property (m-OGP) at very small κ and α. This implies that, in this setting, *second preimage attacks*, where one first samples $\mathbf{x}$, computes $\varphi(A\mathbf{x})$, and seeks a new $\mathbf{x}'$ mapping to the same output, cannot succeed efficiently. However, this does not rule out collision finding in the standard formulation, where the collision-finder has full control over the two inputs and is not restricted to second-preimage attacks. Nor does their result extend to moderate or large values of α.

A related hardness indication comes from recent work of Vafa and Vaikuntanathan [22], who give a worst-case-to-average-case reduction from standard approximate lattice problems to the SBP search problem. Translating their notation to ours, their polynomial-time hardness result applies to margins of order $\kappa = \alpha^{1/2+\epsilon}$ for every constant $\epsilon > 0$, under polynomial-factor worst-case lattice hardness assumptions; under stronger subexponential lattice hardness assumptions, they obtain hardness nearly at the Bansal–Spencer scale $\kappa \approx \sqrt{\alpha}$, up to logarithmic factors. Their reduction applies in a different regime, where the number of variables is polynomially larger than the number of constraints, and they explicitly leave the proportional regime $m = \Theta(n)$ open. Thus their result does not directly address collision finding or the intermediate-density SBP window considered here, but it provides complementary evidence for SBP hardness from worst-case lattice assumptions.

> **Open Question.** *For some $\alpha < 1$, does there exist an intermediate window of κ with $1/\log n \ll \kappa \ll 1/\sqrt{\alpha}$ in which SBP is collision resistant against polynomial-time algorithms?*

We therefore introduce a randomized activation with the same kind of non-monotone, windowed behavior as SBP, but with a collision space that can be analyzed by a first-moment argument. The randomized oscillating activation replaces the fixed SBP boundary by independent random thresholds at each neuron. This extra randomness removes the rigid alignments that make the SBP collision space difficult to analyze, while preserving the obstruction to the positive-tail algorithm. In this model we prove an overlap gap property for extensive collisions, and hence an online lower bound, in a regime where our algorithmic approach no longer applies.

1.4 Result II: Overlap Gaps for Randomized Oscillating Activation

To obtain a model in which the collision geometry can be analyzed by a first-moment argument, we turn to the randomized oscillating activation introduced above. The oscillations persist across a wide region of inputs, and when κ is sufficiently large the positive-tail algorithm cannot drive all margins above κ within the available n columns. The row-wise random thresholds make the activation non-monotone and heterogeneous while keeping the one-row collision probability tractable.

We prove that for this randomized model, extensive collision pairs exhibit an *overlap gap property* (OGP). The theorem considers collision pairs that are internally well separated, and shows that the mutual overlap between two such pairs cannot lie in an intermediate range. In this sense, the collision space has a forbidden band of external overlaps: valid pairs can be close to one another or far apart, but not at intermediate separation. Such overlap gaps are known to create algorithmic barriers in spin-glass models and random constraint satisfaction problems.

Theorem 1.2 (Overlap Gap Property)**.** *For every constant density $\alpha \in (0,1)$, for suitable constants $r \geq 2$, $\delta > 0$, and $C > 0$ depending on α, whenever K is odd, n is sufficiently large, and the parameters satisfy*

$$\kappa^2 \geq \frac{C}{\alpha} \qquad \text{and} \qquad K \geq C\,\frac{\kappa}{\sqrt{\delta}},$$

the randomized oscillating activation exhibits an r-wise overlap gap property for δ-extensive collisions with high probability over $\mathbf{A}$ and the random thresholds.

The formal statement in Theorem 5.1 gives the sharper first-moment condition, expressed through the row exponent $\Lambda_{r,\delta}(\kappa, K)$. The constant $C(\alpha)$ above comes from choosing $r = O(1/\alpha)$ and from the constants needed for this exponent to dominate the entropy term; in particular, it absorbs the factor $2^r r^3$ appearing in the sufficient lower bound on K in Corollary 5.2.

Theorem 1.3 (Online Collision Resistance)**.** *For every constant density $\alpha \in (0,1)$, for suitable constants $\delta > 0$ and $C > 0$ depending on α, whenever K is odd, n is sufficiently large, and the parameters satisfy*

$$\kappa^2 \geq \frac{C}{\alpha} \qquad \text{and} \qquad K \geq C\,\frac{\kappa}{\sqrt{\delta}},$$

every online algorithm has success probability at most $2^{-\Omega(\alpha n)}$ of outputting a δ-extensive collision on a single random instance. In particular, the randomized family $\mathbf{x} \mapsto \varphi(\mathbf{A}\mathbf{x})$ is online δ-extensive collision resistant at density α.

The constant C above comes from the same first-moment and ensemble-OGP requirements as in the preceding theorem; the precise dependence is given by the row exponent $\Lambda_{r,\delta}(\kappa, K)$ in Theorem 5.1. The formal online bound is $2^{-\Omega(n/r)}$; for fixed α the proof takes $r = O(1/\alpha)$, which gives the displayed $2^{-\Omega(\alpha n)}$.

Our formal lower bound is stated for extensive collisions. For cryptographic collision resistance in the usual sense, one may first encode the input with a code $E : \{0,1\}^\ell \to \{\pm 1\}^n$ whose distinct codewords have absolute inner product at most $(1-\delta)n$. Any collision in the composed function $\varphi \circ \mathbf{A} \circ E$ then gives a δ-extensive collision in the neural network. If the code has rate $R = \ell/n$, the composed function is compressing precisely when $\alpha < R$; hence this interpretation applies in the density regimes covered by our lower bound whenever such a code exists with rate exceeding α.

The randomized oscillating activation differs from the square-wave perceptron in that the thresholds are sampled independently for each row, rather than placed on a fixed period shared across all neurons.

The online lower bound provides evidence for a broader hardness phenomenon. The square-wave model studied in [4] suggested that OGP can serve as a geometric indicator of collision resistance in neural networks, supported there by first-moment calculations and AMP experiments. Our randomized oscillating model removes the shared periodic structure of the square wave and proves the corresponding online statement: once the row exponent dominates entropy, the collision space develops an overlap gap, and online algorithms fail. Our theorem proves this online implication for odd K; the conjectural next step is that the same geometric obstruction reflects hardness beyond the online model. In particular, even for fixed κ and fixed odd K, sufficiently high density should make it computationally hard to coordinate two extensive preimages through the row-wise random thresholds. We do not prove lower bounds against spectral, algebraic, lattice-based, low-degree, message-passing, or quantum algorithms, but the OGP theorem motivates the following conjecture.

Conjecture 1.4 (Hardness of Collision Finding)**.** *For fixed constants $K \in \mathbb{N}$, $\kappa > 0$, and $\delta > 0$, there exists a threshold $\alpha_\star < 1$ such that for all $\alpha \geq \alpha_\star$, no polynomial-time algorithm can find a δ-extensive collision with non-negligible probability.*

We leave confirming this conjecture, or finding algorithmic counterexamples, as an open direction for future work.

Paper Organization. The preliminaries fix notation and the formal collision and OGP definitions. Section 4 proves the positive result via a one-sided potential method and a refinement procedure for positive-tail activations. Section 5 proves the first-moment OGP bound for the randomized oscillating activation; the subsequent subsections extend it to correlated-prefix ensembles and derive the online lower bound.

2 Technical Overview

The two main technical components of this paper are a positive algorithmic result for positive-tail activations, including the asymmetric binary perceptron (ABP), and a negative online-hardness result via the overlap gap property (OGP) for a randomized activation function. We outline the core ideas behind both directions.

2.1 A Simple Online Algorithm

For any activation that is constant on the positive tail $[\kappa, \infty)$, we only need to drive all margin values $(\mathbf{Ax})_i$ and $(\mathbf{Ay})_i$ above κ. Once there, the activation φ stabilizes and outputs coincide. The drift analysis is stated in the normalization $\mathbf{B} = \mathbf{A}/\sqrt{\alpha}$, whose columns have entrywise variance $1/m$; reaching margin $\rho = \kappa/\sqrt{\alpha}$ for $\mathbf{B}$ gives margin κ for $\mathbf{A}$. The basic margin-building routine processes columns online, maintaining the current normalized margin vector $\mathbf{s} = \mathbf{B}_{\leq t}\mathbf{x}_{\leq t}$. At step $t+1$, it chooses the next bit x_{t+1} so as to reduce an exponential potential $\Phi(\mathbf{s})$, which gives large weight to coordinates whose margins are still small. This local sign choice creates a systematic drift in the potential: although each normalized column changes every margin by only $O(1/\sqrt{m})$, the exponential weighting turns these small updates into a $\Theta(1/m)$ decrease in Φ per step in expectation.

We show that when the activation φ is constant on $[\kappa, \infty)$, one can efficiently construct extensive collisions with a sufficiently small constant overlap gap in the shrinking regime $\alpha = m/n < 1$. The algorithm processes the columns of $\mathbf{A}$ sequentially, choosing the signs of two vectors $\mathbf{x}, \mathbf{y} \in \{\pm 1\}^n$ in an online fashion as each column is revealed. The goal is to produce two inputs that differ on a linear fraction of their coordinates, yet agree under the activation map $\varphi(\mathbf{A}\cdot)$.

For a single margin vector, the one-step drift is captured by the following potential. Given the current state $\mathbf{s}_t$, the algorithm chooses the next sign $x_t \in \{\pm 1\}$ to minimize

$$\Phi(\mathbf{s}) = \frac{1}{m}\sum_{i=1}^{m} e^{-\lambda s_i},$$

which penalizes small margins and rewards those that already lie far above ρ. The exponential weights make the potential sensitive to precisely the coordinates that are still problematic: a coordinate with small margin contributes much more to Φ than a coordinate already far above the target. When a fresh Gaussian column arrives, the algorithm chooses the sign that is better aligned

with this weighted set of low-margin coordinates. Concretely, after seeing the fresh column $\mathbf{a}$, the choice $x_t = \operatorname{sign}\big(\frac{1}{m}\sum_i a_i e^{-\lambda s_i}\big)$ makes the first-order term in the Taylor expansion negative and of order $\Theta(1/m)$ in expectation over $\mathbf{a}$. The second-order correction is also of order $1/m$, but with a smaller constant at our choice of λ, and the remaining terms are $O(1/m^2)$. This yields a clean multiplicative drift for $\mathbb{E}[\Phi]$. Quantitatively,

$$\mathbb{E}[\Phi(\mathbf{s}_{t+1}) \,|\, \mathbf{s}_t] \le \Big(1 - \frac{\gamma}{m}\Big)\,\Phi(\mathbf{s}_t), \quad \gamma = \Theta(1)\,,$$

because $\sqrt{2/\pi}\,\sqrt{\mathbb{E}\left[e^{-2\lambda s}\right]} - \frac{1}{2}\lambda\,\mathbb{E}\left[e^{-\lambda s}\right] \ge c\,\mathbb{E}\left[e^{-\lambda s}\right]$ for $\lambda = \sqrt{2/\pi}$ and some universal $c > 0$. The first term is the gain from choosing the better sign against a Gaussian column, while the second term is the variance cost from the quadratic Taylor expansion. Thus, Φ decays exponentially in t/m. After $O((1+\rho)m)$ steps, only a small fraction of coordinates of $\mathbf{Bx}$ remain below the normalized target margin ρ; a short refinement phase completes the separation, and rescaling by $\sqrt{\alpha}$ gives margins above κ for $\mathbf{A}$.

For collision finding, we first create the required overlap gap by assigning opposite signs to $\mathbf{x}$ and $\mathbf{y}$ on $\beta_1 m$ coordinates, with $\beta_1 = \delta/(2\alpha)$. Since $m = \alpha n$, this gives $\delta n/2 + o(n)$ initial disagreements. After this initialization, every remaining column receives the same sign in $\mathbf{x}$ and $\mathbf{y}$, so the overlap is preserved. The proof tracks the pointwise minimum $\mathbf{s} = \min(\mathbf{Bx}, \mathbf{By})$; because $\min(u+za, v+za) = \min(u,v) + za$, the same one-sided potential argument applies to this minimum state. A refinement phase based on $FocusedPush$ uses $O(m)$ additional columns to clear the remaining low-margin coordinates. The total number of used columns is at most n once δ, $\kappa\sqrt{\alpha}$, and α are chosen sufficiently small: reaching normalized margin $\rho = \kappa/\sqrt{\alpha}$ costs $O((1+\rho)m) = O((\alpha + \kappa\sqrt{\alpha})n)$ drift columns, and the initialization/refinement budgets are also linear in m and δn. This is why the theorem is stated only for sufficiently small constant overlap gap. Once all entries of $\mathbf{Ax}$ and $\mathbf{Ay}$ lie above κ, the component-wise activations stabilize, giving,

$$\varphi(\mathbf{Ax}) = \varphi(\mathbf{Ay})\,,$$

certifying a δ-extensive collision. Informally, the exponential potential converts the low-margin tail into a single scalar quantity: coordinates far below the target contribute exponentially more than coordinates already safely above it. The $1/m$ drift therefore does not just improve an average margin; it steadily reduces the weighted mass of problematic coordinates. After $O((1+\rho)m)$ drift columns this leaves only a small residual set, which the refinement phase targets directly.

2.2 Overlap Gaps for Collision Finding

To certify online hardness for more expressive, non-monotone activations, we establish an r-wise overlap gap property (OGP) for a randomized multi-threshold activation. For each row $i \in [m]$, we fix an odd integer $K \ge 1$, so that the activation changes sign between the two tails, draw independent thresholds $t_{ij} \sim \mathrm{Unif}[-\kappa, \kappa]$ for $j \in [K]$, and define

$$\varphi_i(x) = \prod_{j=1}^{K} \operatorname{sign}(x - t_{ij})\,, \qquad \varphi(\mathbf{Ax})_i = \varphi_i((\mathbf{Ax})_i)\,.$$

The independent thresholds make the row calculation depend on the Gaussian positions of the candidate margins, rather than on a common period shared by all rows. For a fixed admissible r-tuple, a single threshold has some chance of separating the signs in the selected Gaussian subfamily.

Using K independent thresholds amplifies the resulting row-wise contraction, while increasing κ spreads the thresholds over a wider interval and weakens the contribution of each one. The exponent $\Lambda_{r,\delta}(\kappa, K)$ is the resulting row-wise gain. The OGP proof then compares this gain with the entropy cost of choosing an admissible r-tuple.

Intuitively, the OGP states that no valid r-tuple of collision pairs $(\mathbf{x}_i, \mathbf{y}_i)$ can have external overlaps,

$$\frac{1}{2n}\left(\mathbf{x}_i^\top \mathbf{x}_j + \mathbf{y}_i^\top \mathbf{y}_j\right),$$

lying in a fixed forbidden interval. The proof certifies this gap by a first-moment bound. Fix r collision pairs obeying the internal overlap constraint and consider a random row $\mathbf{a}^\top$ of $\mathbf{A}$. For each pair,

$$(X_i, Y_i) = (\mathbf{a}^\top \mathbf{x}_i, \mathbf{a}^\top \mathbf{y}_i),$$

is jointly Gaussian with covariance determined by overlaps. For a single pair, the signs of X_i and Y_i differ exactly when the threshold lies between them. Conditional on $|X_i|, |Y_i| \le \kappa$, this has probability $\Theta(|X_i - Y_i|/\kappa)$. The proof uses a higher-order version of this observation: once a selected subfamily of Gaussian values is well separated inside $[-\kappa, \kappa]$, a random threshold has nontrivial probability of making the relevant product of signs negative. The K independent thresholds then amplify this one-threshold contraction. By Lemma 3.4, for every $p \in (0, 1)$ the selected Gaussian variables are separated by order $p\sqrt{\delta}$ except with probability of order $r^2 p$, while the event that some selected Gaussian lies outside $[-\kappa, \kappa]$ costs order $r\, e^{-\kappa^2/2}$. Combining this with the K independent thresholds gives the formal row-wise exponent $\Lambda_{r,\delta}(\kappa, K)$ defined in Section 5: for a fixed admissible r-tuple, the one-row collision probability is at most $2^{-\Lambda_{r,\delta}(\kappa,K)}$. By the entropy count in Lemma 5.3, summing over all families of r pairs consistent with the overlap constraints gives the per-coordinate first-moment bound,

$$\frac{1}{n}\log_2 \mathbb{E}[\#r\text{-tuples of } \delta\text{-admissible collisions}] \le 2 + 2(r-1)H(\delta/(2c_1)) - \alpha\Lambda_{r,\delta}(\kappa, K) + o(1).$$

Thus the first-moment exponent is negative once $\alpha\Lambda_{r,\delta}(\kappa, K)$ exceeds $2 + 2(r-1)H(\delta/(2c_1))$ by a constant amount. By Markov's inequality, in this negative-exponent regime no such r-tuple exists with exponentially high probability, giving the formal OGP result in Theorem 5.1 and Corollary 5.2. The overlap assumptions enter through a selection lemma. Applied to the $2r$ Gaussian projections from one row, it shows that admissibility forces an $(r+1)$-variable subfamily whose covariances are bounded away from one. This separated subfamily is what drives the one-row contraction; the anti-concentration parameter p converts covariance separation into actual separation of Gaussian values and is optimized in the definition of $\Lambda_{r,\delta}(\kappa, K)$.

Online hardness uses a separate ensemble version of the OGP. In the ensemble setting, we consider correlated instances $(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(r)})$ sharing their first t columns and having independent suffixes, with

$$\frac{\delta}{c_2} < 1 - \frac{t}{n} < \frac{\delta}{c_1}.$$

For prefix-aligned families, the first-moment count becomes $2 + 2(r-1)(1-t/n) + o(1)$, and the same row-wise contraction applies after selecting an explicitly separated subfamily. This gives ensemble r-OGP. If a single-instance online algorithm succeeds with probability p, then, after conditioning on the shared prefix, thresholds, and internal coins, the r suffixes remain independent with a common conditional success probability q. Jensen's inequality gives $\mathbb{E}[q^r] \ge \mathbb{E}[q]^r = p^r$, so the coupled

experiment succeeds on all r instances with probability at least p^r. Because the algorithm is online and the coupled instances have the same prefix, the outputs produced on the first t coordinates are identical across the r runs. On the suffix, we independently flip columns in each coupled instance and then undo the flips in the output; this preserves the collision equations, while making the suffix contributions to external overlaps behave like sums of independent mean-zero signs. Choosing t inside the forbidden window therefore turns simultaneous success into an ensemble-OGP-forbidden configuration except with exponentially small probability. Hence $p \le \exp(-\Omega(n/r))$.

In contrast to the ABP, where monotonicity and a simple exponential potential enable efficient collision finding, sufficiently strong randomized oscillation introduces a geometric barrier for online algorithms.

3 Preliminaries and Definitions

We will use capital letters $A, B, C, \ldots$ to denote sets and use standard notation $\cap, \cup, \setminus$ to denote common set-theoretic operations. We denote vectors using boldface lower-case letters $\mathbf{x}, \mathbf{y}, \ldots, \in \mathbb{R}^n$, and reserve boldface upper-case letters $\mathbf{A}, \mathbf{B}, \ldots$ for matrices. For a vector $\mathbf{x}$, write $\mathbf{x}_{\le t}$ for its first t coordinates; for a matrix $\mathbf{B}$, write $\mathbf{B}_{\le t}$ for its first t columns. We use $\mathbf{0}_k$ or $\mathbf{1}_k$ to denote the k-vectors containing only 0 or 1 respectively. Inner products between vectors $\mathbf{x}$ and $\mathbf{y}$ are written as $\mathbf{x}^\top \mathbf{y}$, or occasionally as $\langle \mathbf{x}, \mathbf{y} \rangle$ when the latter is more convenient. We may omit the subscript if its length is implicit from the context. We use $\mathbf{1}\{\cdot\}$ to denote the indicator variable which is 1 if its argument holds true. For $\mathbf{x}, \mathbf{y} \in \{-1, 1\}^n$ we write $\Delta_{\mathrm{Ham}}(\mathbf{x}, \mathbf{y})$ for their Hamming distance; equivalently, $\mathbf{x}^\top \mathbf{y} = n - 2\Delta_{\mathrm{Ham}}(\mathbf{x}, \mathbf{y})$. We write $[n] = \{1, \ldots, n\}$ for positive integers n.

We denote by $\mathcal{N}(\mu, \sigma^2)$ the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Similarly, we denote by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ the multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. Denote by $H(p) = -p \log_2(p) - (1-p) \log_2(1-p)$ the binary entropy function. Throughout the paper, $\alpha = m/n$ is treated as a fixed constant unless stated otherwise. When a fixed density α is specified, all asymptotic statements are over integer pairs (m, n) with $n \to \infty$ and $m/n \to \alpha$; equivalently, one may take $m = \lfloor \alpha n \rfloor$, with rounding effects absorbed in the $o(1)$ terms. We use the convention $\mathrm{sign}(0) = 1$.

Collision Finding. Fix an activation map $\varphi : \mathbb{R}^m \to \{-1, 1\}^m$, deterministic coordinate-wise or row-wise randomized. The instance consists of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with entries $\mathbf{A}_{ij} \sim \mathcal{N}(0, 1/n)$, drawn independently. In the randomized case, the activation is sampled once and fully revealed together with $\mathbf{A}$. When randomized, all thresholds are drawn once at the start and then treated as fixed.

Definition 3.1 (Collision Problem)**.** *Given $\delta \in (0, 1)$, a (randomized) algorithm C* solves the collision problem *for φ at distance δ if, on input $\mathbf{A}$ and any sampled activation thresholds, it outputs $\mathbf{x}, \mathbf{y} \in \{\pm 1\}^n$ with*

$$\varphi(\mathbf{A}\mathbf{x}) = \varphi(\mathbf{A}\mathbf{y}), \quad \textit{and} \quad \left|\mathbf{x}^\top \mathbf{y}\right| \le (1 - \delta)\, n, \tag{1}$$

with non-negligible probability over the joint randomness of $\mathbf{A}$ and C. We say φ is δ-extensive collision resistant *if no probabilistic polynomial-time algorithm solves this distance-δ collision problem.*

The preceding definition is the full probabilistic-polynomial-time notion for extensive collisions, stated for context. The rigorous lower bound proved in this paper is for online algorithms; when

referring to proved hardness below, we use *online extensive collision resistance* to mean resistance within that restricted online model.

We refer to any pair $(\mathbf{x}, \mathbf{y})$ satisfying Eq. (1) as a δ-*extensive collision*: a valid collision whose two inputs have absolute inner product at most $(1-\delta)n$. When $\delta = \Omega(1)$ is a positive constant we call it simply an *extensive collision*.

The extensive-distance requirement is weaker than standard cryptographic collision resistance, since it only rules out collisions whose inputs have small absolute overlap. This distinction can be removed by placing an error-correcting code before the neural network, provided the code itself has the corresponding absolute-overlap separation and the resulting function is still compressing. Indeed, if $E : \{0,1\}^\ell \to \{\pm 1\}^n$ satisfies

$$|\langle E(u), E(v)\rangle| \le (1-\delta)n \qquad \text{for all } u \neq v,$$

then any collision

$$\varphi(\mathbf{A}E(u)) = \varphi(\mathbf{A}E(v)), \qquad u \neq v,$$

induces a δ-extensive collision between the codewords $E(u)$ and $E(v)$. If the code has rate $R = \ell/n$, the composed function $\varphi \circ \mathbf{A} \circ E$ maps ℓ bits to $m = \alpha n$ bits and is compressing exactly when $\alpha < R$. Since such absolute-overlap codes have rates bounded away from zero for every fixed small enough δ, the ECC interpretation applies to the randomized-activation online lower bound for any fixed α below such a rate, by taking $\kappa = \Omega(1/\sqrt{\alpha})$ and an odd $K = \Omega_\alpha(\kappa/\sqrt{\delta})$.

Definition 3.2 (Overlap Gap Property)**.** *Let $r \in \mathbb{N}$, $0 < \delta < 1$, and let $\varphi : \mathbb{R}^m \to \{-1,1\}^m$ be an activation map, deterministic coordinate-wise or sampled once and then fixed. Let $2 < c_1 < c_2 < 2c_1$ be two constants. Define $T(r, \delta, c_1, c_2) \subseteq \{-1,1\}^{2rn}$ as the set of all r pairs $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^r$ with $\mathbf{x}_i, \mathbf{y}_i \in \{-1,1\}^n$ that satisfy the following two conditions.*

1. (Internal Overlap). *For every $i = 1 \dots r$, it holds that,*

$$\left|\mathbf{x}_i^\top \mathbf{y}_i\right| \le (1-\delta)\, n\,. \tag{2}$$

2. (External Overlap). *For every $i, j = 1 \dots r$ with $i \neq j$, we have that,*

$$1 - \frac{\delta}{c_1} \le \frac{\mathbf{x}_i^\top \mathbf{x}_j + \mathbf{y}_i^\top \mathbf{y}_j}{2n} \le 1 - \frac{\delta}{c_2}\,. \tag{3}$$

A set of r pairs $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^r$ satisfying the internal and external overlap requirements is said to be δ-admissible.

Let $\alpha > 0$ be a density parameter, and let $m = m(n)$ satisfy $m/n \to \alpha$. Let $\mathbf{A} = (\mathbf{a}_j)_{j=1}^n \in \mathbb{R}^{m \times n}$ have columns $\mathbf{a}_j \in \mathbb{R}^m$, and entries $\mathbf{A}_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$. Denote by $S(r, \delta, c_1, c_2; \mathbf{A}) \subseteq T(r, \delta, c_1, c_2)$ the subset of δ-admissible sets of r pairs that also satisfy a third property.

3. (Collision). *For every $i = 1 \dots r$, it holds that,*

$$\varphi(\mathbf{A}\mathbf{x}_i) = \varphi(\mathbf{A}\mathbf{y}_i)\,. \tag{4}$$

For fixed $\mathbf{A}$ and fixed activation thresholds, the function $f_{\mathbf{A}}(\mathbf{x}) = \varphi(\mathbf{A}\mathbf{x})$ satisfies the r-overlap gap property *(r-OGP) with parameters $0 < \delta < 1$ and $2 < c_1 < c_2 < 2c_1$ if*

$$S(r, \delta, c_1, c_2; \mathbf{A}) = \emptyset,$$

i.e. if there is no r-tuple of collision pairs satisfying the internal overlap condition whose pairwise external overlaps all lie in the forbidden window.

Probability Conventions. Unless stated otherwise, all probabilities and expectations are over the joint randomness of: (i) the Gaussian matrix $\mathbf{A}$ (and, in the ensemble setting, the correlated draw $(\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(r)})$), (ii) any random thresholds defining the activation (sampled once and then revealed together with $\mathbf{A}$), and (iii) the internal randomness of the algorithm (when present). When we write $\Pr[\cdot \mid \mathbf{A}]$ (or $\mathbb{E}[\cdot \mid \mathbf{A}]$) we condition on the realized instance(s) and thresholds, and keep only algorithmic randomness. When we write $\mathbb{E}[\mathsf{Coll}_{r,\delta}]$ we mean expectation over the draw of $\mathbf{A}$ and the activation thresholds (and over the ensemble when applicable).

Lemma 3.3 (First-Moment Bound)**.** *If $X \geq 0$, then $\Pr[X \geq a] \leq \mathbb{E}[X]/a$ for every $a > 0$. In particular, if N counts forbidden configurations, then $\Pr[N \geq 1] \leq \mathbb{E}[N]$.*

Lemma 3.4 (Gaussian Estimates)**.** *If $G \sim \mathcal{N}(0, \sigma^2)$, then for all $u \geq 0$,*

$$\Pr[|G| \geq u\sigma] \leq 2\, e^{-u^2/2}, \qquad \Pr[|G| \leq u\sigma] \leq \sqrt{\frac{2}{\pi}}\, u.$$

Moreover, for every fixed integer $k \geq 1$, $\mathbb{E}\left[|G|^k\right] = \Theta_k(\sigma^k)$.

Lemma 3.5 (Hoeffding Bound)**.** *If $\xi_1, \dots, \xi_s$ are independent mean-zero random variables with $\xi_i \in [-1, 1]$, then for all $\lambda > 0$,*

$$\Pr\left[\left|\sum_{i=1}^{s} \xi_i\right| \geq \lambda s\right] \leq 2\, e^{-\lambda^2 s/2}.$$

These standard forms can be found, for example, in [23, 5].

4 Finding Collisions for Positive-Tail Activations

We demonstrate that extensive collisions with sufficiently small constant overlap gap can be found at sufficiently low aspect ratios α for activations that are constant on a positive tail. To measure quality, we define the *algorithmic rate* to be the smaller of the absolute-overlap gap δ, where $|\langle x, x' \rangle| \leq (1-\delta)n$, and the collision ratio $\alpha = m/n$ as n increases.

We describe two algorithms. The first one works well in practice, but possibly not so well in theory. We prove its correctness at inverse logarithmic algorithmic rates. We leave open its performance at constant algorithmic rates. We use it as a component in a more complicated algorithm that provably works at fixed positive algorithmic rates, for sufficiently small constants δ and α. The main result of this section is the following:

Theorem 4.1 (Collision Finding for Positive-Tail Activations)**.** *There exists $\delta_0 > 0$ such that for every fixed $\delta \in (0, \delta_0)$ there exist constants $\alpha_0 = \alpha_0(\delta) > 0$ and $c > 0$, and an efficient online algorithm, such that for every pair of parameters (α, κ) satisfying $\alpha < \alpha_0$ and $\kappa \leq c/\sqrt{\alpha}$ and for all sufficiently large n, the algorithm outputs with constant probability two vectors $\mathbf{x}, \mathbf{y} \in \{\pm 1\}^n$ such that*

$$\left|\mathbf{x}^\top \mathbf{y}\right| \leq (1-\delta)\, n\,, \qquad \min_i (\mathbf{A}\mathbf{x})_i \geq \kappa\,, \qquad \min_i (\mathbf{A}\mathbf{y})_i \geq \kappa\,.$$

In particular, for any activation function φ that is constant on $[\kappa, \infty)$, $\mathbf{x}$ and $\mathbf{y}$ form a δ-extensive collision for φ.

The proof works in the normalized matrix $\mathbf{B} = \mathbf{A}/\sqrt{\alpha}$ and targets margin $\rho = \kappa/\sqrt{\alpha}$. The collision finder first creates the desired overlap gap by assigning opposite signs on an initial block, and then assigns the same signs to $\mathbf{x}$ and $\mathbf{y}$ on all remaining coordinates. The state tracked after initialization is the pointwise minimum $\min(\mathbf{Bx}, \mathbf{By})$. A drift phase applies $Push$ to reduce the one-sided exponential potential of this minimum state, and $FocusedPush$ clears the remaining low-margin coordinates. The column budget is kept below n by taking δ, $\kappa\sqrt{\alpha}$, and α sufficiently small.

Because the drift analysis acts row-by-row, the algorithmic bookkeeping is measured relative to the output dimension m. Since $m = \alpha n$, a disagreement on a βm-sized set of coordinates corresponds to an overlap deficit of $2\alpha\beta + o(1)$ on the n-scale.

The drift analysis is most naturally stated in the normalization where the online columns have law $\mathcal{N}(\mathbf{0}, \mathbf{I}_m/m)$. To apply it to the paper's convention $\mathbf{A}_{ij} \sim \mathcal{N}(0, 1/n)$, write $\mathbf{B} = \mathbf{A}/\sqrt{\alpha}$, so that $\mathbf{B}_{ij} \sim \mathcal{N}(0, 1/m)$. If the algorithm constructs $\mathbf{x}, \mathbf{y}$ with $\min_i(\mathbf{Bx})_i, \min_i(\mathbf{By})_i \geq \rho$, then the original matrix satisfies $\min_i(\mathbf{Ax})_i, \min_i(\mathbf{Ay})_i \geq \sqrt{\alpha}\rho$. Thus, to obtain margin κ for $\mathbf{A}$, the $\mathbf{B}$-normalized analysis targets $\rho = \kappa/\sqrt{\alpha}$. Since the normalized process has $n = m/\alpha$ available columns, a target ρ costs $O(\rho m)$ drift steps, which is feasible as long as $\alpha\rho = \kappa\sqrt{\alpha}$ is sufficiently small.

4.1 A Simple Online Algorithm

At the heart is an online search algorithm in the normalized model. Let $\mathbf{M} \in \mathbb{R}^{m \times n}$ have independent columns with law $\mathcal{N}(\mathbf{0}, \mathbf{I}_m/m)$. The goal is to choose $\mathbf{x} \in \{-1, 1\}^n$ so that $\mathbf{Mx}$ is not only entrywise positive, but bounded away from zero. The algorithm maintains a current state of the form $\mathbf{s} = \mathbf{Mx}$ (and possibly a few extra bits of information). Upon seeing the next column $\mathbf{a}$ of $\mathbf{M}$ it decides on the corresponding sign in $\mathbf{x}$. The sign is effectively chosen to minimize the potential,

$$\Phi(\mathbf{s}) = \mathbb{E}_{I \sim \mathrm{Unif}([m])}\big[\psi(s_I)\big] = \frac{1}{m}\sum_{i=1}^{m} \psi(s_i)\,,$$

for a suitable "energy function" ψ, where the expectation is over a uniformly random coordinate $I \in [m]$. The wisdom of our choice $\psi(s) = e^{-\lambda s}$, where $\lambda = \sqrt{2/\pi}$, will become apparent shortly. Given the current state $\mathbf{s} \in \mathbb{R}^m$, upon reading input $\mathbf{a} \in \mathbb{R}^m$, the potential updates to,

$$\Phi(\mathbf{s}') = \Phi(\mathbf{s} \pm \mathbf{a}) = \mathbb{E}[\psi(s \pm a)] = \Phi(\mathbf{s}) \pm \mathbb{E}\left[\psi'(s)\, a\right] + \frac{1}{2}\,\mathbb{E}\left[\psi''(s)\, a^2\right] \pm \frac{1}{6}\,\mathbb{E}\left[\psi'''(s)\, a^3\right] + \cdots\,,$$

assuming ψ is analytic at s. All other expectations $\mathbb{E}[\cdot]$ in this section are over the randomness of the current Gaussian column $\mathbf{a}$ (and any internal randomness of the algorithm), conditioned on the current state $\mathbf{s}$ unless stated otherwise. To simplify the analysis, we will choose the sign so as to minimize the leading term $\mathbb{E}[\psi'(s)a]$ instead of $\Phi(\mathbf{s}')$ itself. That choice is $x_j = -\operatorname{sign}\mathbb{E}[\psi'(s)a]$. The change in potential is,

$$\Phi(\mathbf{s}') - \Phi(\mathbf{s}) = -\big|\mathbb{E}[\psi'(s)\, a]\big| + \frac{1}{2}\,\mathbb{E}[\psi''(s)\, a^2] + \sum_{i=3}^{\infty} (\pm 1)^i \frac{1}{i!}\,\mathbb{E}[\psi^{(i)}(s)\, a^i]\,. \tag{5}$$

To understand the typical behavior, we first analyze the expected change given $\mathbf{s}$. Heuristically, choosing the sign to minimize the linear term exploits the $1/\sqrt{m}$ fluctuations of a to produce a systematic $1/m$ drop in Φ, while the quadratic and higher-order terms contribute only smaller,

controlled corrections. As $\mathbf{a}$ is m-dimensional normal with entrywise variance $1/m$, the key quantity $\mathbb{E}[\psi'(s)a]$ is a centered normal with variance $(1/m^2)\,\mathbb{E}[\psi'(s)^2]$. The leading term is the absolute value of this variable with a minus sign in front. Its mean value is,

$$-\,\mathbb{E}\big[\big|\mathbb{E}[\psi'(s)\,a \mid \mathbf{s}]\big|\big] = -\sqrt{\frac{2}{\pi}}\,\frac{\sqrt{\mathbb{E}\,[\psi'(s)^2]}}{m}\,.$$

The next term $\mathbb{E}[\psi''(s)\,a^2]/2$ contributes $+\frac{1}{2m}\,\mathbb{E}[\psi''(s)]$, since $\mathbb{E}[a^2] = 1/m$ and s is independent of a. For analytic ψ with subexponential growth, the remainder is $O\big(\mathbb{E}[|\psi^{(4)}(s)|]/m^2\big)$ by Lemma 3.4, using $\mathbb{E}\,|a|^k = \Theta(m^{-k/2})$. Intuitively, the linear term in the expansion enforces a negative drift (pulling the margins outward), while the quadratic term captures variance from the Gaussian increments and is controlled by moment bounds; the remaining higher-order terms are $O(1/m^2)$ for our choice of ψ. In summary, the average change in potential is,

$$\mathbb{E}\big[\Phi(\mathbf{s}') - \Phi(\mathbf{s})\,\big|\,\mathbf{s}\big] = \frac{1}{m}\left(-\sqrt{\frac{2}{\pi}}\sqrt{\mathbb{E}\,[\psi'(s)^2]} + \frac{1}{2}\,\mathbb{E}\,\big[\psi''(s)\big]\right) + O\left(\frac{\mathbb{E}\,\big[\psi^{(4)}(s)\big]}{m^2}\right).$$

Intuitively, the first term captures the outward pull from small margins, the second is a variance correction that stays strictly smaller at our choice of ψ, and the remainder vanishes at rate $1/m$. The potential ψ should be decreasing to reward positivity. To control outliers, it should be positive. Thus its derivative ψ' should be decreasing, and the second derivative ψ'' should be positive. Then progress is made as long as the term $\sqrt{2/\pi}\sqrt{\mathbb{E}\,[\psi'(s)^2]}$ wins over $\mathbb{E}\,[\psi''(s)]$ in the battle of drifts.

For exponential potentials $\psi(x) = e^{-\lambda x}$, this is guaranteed as long as $0 < \lambda < \sqrt{8/\pi}$ because,

$$\begin{aligned}&\sqrt{\frac{2}{\pi}}\sqrt{\mathbb{E}\,[\psi'(s)^2]} - \frac{1}{2}\,\mathbb{E}\,\big[\psi''(s)\big]\\ &\quad= \sqrt{\frac{2}{\pi}}\lambda\sqrt{\mathbb{E}\,[e^{-2\lambda s}]} - \frac{1}{2}\lambda^2\,\mathbb{E}\left[e^{-\lambda s}\right] \geq (\sqrt{2/\pi} - \lambda/2)\lambda\,\mathbb{E}\left[e^{-\lambda s}\right],\end{aligned}$$

by convexity of moments. The drift is optimized when $\lambda = \sqrt{2/\pi}$. In effect, the updates harvest many tiny $1/\sqrt{m}$ nudges into a uniform multiplicative decay of Φ, which is exactly what ultimately clears the low-margin coordinates.

The algorithms below use three related routines. $Push(m)$ is the basic one-step potential descent rule in an m-dimensional normalized state. $Push(m, m_0)$ is the same rule when the active set has size m but the ambient Gaussian increments have variance $1/m_0$. $FocusedPush(m_0)$ repeatedly applies this scaled rule to the currently smallest coordinates, while updating all coordinates with the chosen signs.

It is useful to isolate the one-step sign rule used by $Push$. For a state $\mathbf{s} \in \mathbb{R}^m$ and a fresh column $\mathbf{a} \in \mathbb{R}^m$, define

$$PushSign(\mathbf{s}, \mathbf{a}) = \operatorname{sign}\left(\frac{1}{m}\sum_{i=1}^{m} a_i e^{-\lambda s_i}\right).$$

The corresponding update is $\mathbf{s} \leftarrow \mathbf{s} + PushSign(\mathbf{s}, \mathbf{a})\mathbf{a}$.

Algorithm $Push(m)$
State $\mathbf{s} = (\dots, s, \dots) \in \mathbb{R}^m$
On input $\mathbf{a} \sim (1/\sqrt{m})\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$,
1 Output $x = PushSign(\mathbf{s}, \mathbf{a})$.
2 Update $\mathbf{s}$ to $\mathbf{s} + x\mathbf{a}$.

The average potential Φ_λ then satisfies the following recurrent inequality:

Claim 4.2. *When $\psi(s) = e^{-\lambda s}$, assuming $0 < \lambda \le \sqrt{2/\pi}$,*

$$\mathbb{E}\big[\Phi_\lambda(\mathbf{s}') \,\big|\, \mathbf{s}\big] \le \left(1 - \frac{\gamma\lambda}{m}\right)\Phi_\lambda(\mathbf{s})\,,$$

where $\gamma = \sqrt{2/\pi} - \sqrt{\pi/2}(\exp(1/\pi) - 1) > 0.328$.

Proof. For the exponential potential, set $w_i = e^{-\lambda s_i}$ and

$$Z = \frac{1}{m}\sum_{i=1}^{m} w_i a_i, \qquad x = \operatorname{sign}(Z).$$

The linear term in $\Phi_\lambda(\mathbf{s} + x\mathbf{a}) - \Phi_\lambda(\mathbf{s})$ is $-\lambda|Z|$. We next justify that the adaptive sign does not increase the nonlinear remainder. Fix i and $k \ge 1$. Since (a_i, Z) is jointly Gaussian and $\operatorname{Cov}(a_i, Z) = w_i/m^2 \ge 0$, we may write

$$a_i = \rho_i m^{-1/2} G + \sqrt{1 - \rho_i^2}\, m^{-1/2} H, \qquad \operatorname{sign}(Z) = \operatorname{sign}(G),$$

where $\rho_i \in [0, 1]$ and G, H are independent standard Gaussians. Expanding in powers of H, only even powers of H survive, and $\operatorname{sign}(G)G^{2\ell+1} = |G|^{2\ell+1}$, so

$$\mathbb{E}[\operatorname{sign}(Z) a_i^{2k+1} \mid \mathbf{s}] \ge 0.$$

Thus all odd Taylor terms beyond the linear one have nonpositive contribution. Dropping them and summing the even terms gives

$$\begin{aligned}\mathbb{E}\big[\Phi_\lambda(\mathbf{s}') - \Phi_\lambda(\mathbf{s}) \,\big|\, \mathbf{s}\big] &\le -\lambda\, \mathbb{E}[|Z| \mid \mathbf{s}] + \frac{1}{m}\sum_{i=1}^{m} w_i \sum_{k\ge 1} \frac{\lambda^{2k}}{(2k)!}\, \mathbb{E}[a_i^{2k}] \\ &= -\lambda\, \mathbb{E}[|Z| \mid \mathbf{s}] + \big(\exp(\lambda^2/2m) - 1\big)\, \Phi_\lambda(\mathbf{s}).\end{aligned}$$

The first term on the right hand side evaluates to,

$$-\lambda\, \mathbb{E}[|Z| \mid \mathbf{s}] = -\sqrt{\frac{2}{\pi}}\,\frac{\lambda}{m}\,\sqrt{\frac{1}{m}\sum_{i=1}^{m} e^{-2\lambda s_i}} \le -\sqrt{\frac{2}{\pi}}\,\frac{\lambda}{m}\,\Phi_\lambda(\mathbf{s})\,,$$

by Cauchy–Schwarz. Overall,

$$\mathbb{E}\big[\Phi_\lambda(\mathbf{s}') - \Phi_\lambda(\mathbf{s}) \,\big|\, \mathbf{s}\big] \le \left(-\sqrt{\frac{2}{\pi}}\,\frac{\lambda}{m} + \exp\big(\lambda^2/2m\big) - 1\right)\Phi_\lambda(\mathbf{s})\,.$$

As $\lambda^2/2 \le 1/\pi$ and $m \ge 1$, $\exp(\lambda^2/2m) - 1 \le \pi(\exp(1/\pi) - 1)(\lambda^2/2m) \le \sqrt{\pi/2}(\exp(1/\pi) - 1)(\lambda/m)$. The desired bound follows. □

Corollary 4.3. $\mathbb{E}\,[\Phi_\lambda(\mathbf{s}_t)] \le (1 - \gamma\lambda/m)^t \Phi_\lambda(\mathbf{s}_0) \le \exp(-\gamma\lambda t/m)\, \Phi_\lambda(\mathbf{s}_0)$, *where $\mathbf{s}_t$ is the state after t steps, assuming $\lambda^2 \le 2/\pi$.*

Proof. By Claim 4.2, $\mathbb{E}[M_{t+1} \mid M_1, \ldots, M_t] \le M_t$, i.e. $M_t = (1 - \gamma\lambda/m)^{-t}\,\Phi_\lambda(\mathbf{s}_t)$ is a supermartingale. Applying the tower property of conditional expectation repeatedly gives $\mathbb{E}[M_t] \le \mathbb{E}[M_0] = M_0$. □

Claim 4.4. *Let $\mathbf{s}_T$ be the state obtained by running $Push(m)$ for T steps from an initial state $\mathbf{s}_0$, with $\lambda = \sqrt{2/\pi}$. Then,*

$$\mathbb{E}[\Phi(\mathbf{s}_T)] \le \exp\bigl(-\gamma' T/m\bigr)\,\Phi(\mathbf{s}_0).$$

Consequently, for every threshold $b \in \mathbb{R}$, the expected fraction of coordinates below b is at most,

$$\exp\bigl(\lambda b - \gamma' T/m\bigr)\,\Phi(\mathbf{s}_0).$$

Proof. The first inequality is Corollary 4.3. If $s_i < b$, then $e^{-\lambda s_i} > e^{-\lambda b}$. Hence the fraction of coordinates below b is at most $e^{\lambda b}\Phi(\mathbf{s}_T)$. Taking expectations gives the second bound. □

For example, from Claim 4.4 with $\mathbf{s}_0 = \mathbf{0}$ and $b = 0$, the expected fraction of negative entries after t steps is at most $\exp(-\gamma' t/m)$. Thus plain *Push* clears all negative coordinates with constant probability after $O(m \log m)$ steps. Similarly, taking $b = 1$ shows that after $O(m)$ steps all but a constant fraction of the coordinates have margin at least 1 in expectation. The limitation is that this still leaves open the possibility of a linear number of bad coordinates after only $O(m)$ steps.

4.2 A Constant Rate Algorithm

The preceding analysis shows that plain *Push* steadily reduces the exponential potential, but by itself it gives only an inverse-logarithmic algorithmic rate if we insist that every coordinate becomes positive. The obstruction is that, after $O(m)$ steps, a small but still linear set of low-margin coordinates may remain. The algorithm for constant algorithmic rate removes this loss by focusing the potential descent on the current low-margin tail.

In each stage, the entries of $\mathbf{s}$ are split into critical and safe coordinates. The critical set consists of the m smallest entries at the current halving scale, and *Push* is applied only to this set when choosing the next sign. The same sign is then applied to all coordinates. Thus the algorithm spends most of its effort on the coordinates that still need help, while the safe coordinates are allowed to fluctuate.

As the size of the critical set changes from the initial m_0 to the "current" m, the potential function has to be scaled accordingly.

Algorithm $Push(m, m_0)$
State $\mathbf{s} \in \mathbb{R}^m$
On input $\mathbf{a} \sim (1/\sqrt{m_0})\mathcal{N}(\mathbf{0}, \mathbf{I}_m)$,
1 Set $\lambda_{m,m_0} = \sqrt{2/\pi}\sqrt{m/m_0}$.
2 Set $x \leftarrow \operatorname{sign}(W)$, where
3 $$W = \frac{1}{m}\sum_{i=1}^{m} \sqrt{\frac{m_0}{m}}\, a_i \exp\left(-\lambda_{m,m_0}\sqrt{\frac{m_0}{m}}\, s_i\right).$$
4 Update $\mathbf{s} \leftarrow \mathbf{s} + x\mathbf{a}$.

Equivalently, this is the one-step rule of $Push(m)$ applied to the rescaled state $\sqrt{m_0/m}\mathbf{s}$ and rescaled input $\sqrt{m_0/m}\mathbf{a}$. The state $\mathbf{s}_t$ of $Push(m, m_0)$ in step t satisfies

$$\mathbb{E}\left[\Phi(\mathbf{s}_t)\right] \le \exp\bigl(-\gamma' t/\sqrt{m m_0}\bigr)\ \Phi(\mathbf{s}_0). \tag{6}$$

Proof of (6). Let $\lambda = \sqrt{2/\pi}\sqrt{m/m_0}$. By Corollary 4.3,

$$\mathbb{E}\left[\Phi_\lambda\left(\sqrt{m_0/m}\,\mathbf{s}_t\right)\right] \le \exp\left(-\gamma' t/\sqrt{m m_0}\right)\ \Phi_\lambda\left(\sqrt{m_0/m}\,\mathbf{s}_0\right).$$

By definition of Φ, $\Phi_\lambda\left(\sqrt{m_0/m}\,\mathbf{s}\right) = \Phi_{\lambda\sqrt{m_0/m}}(\mathbf{s}) = \Phi_{\sqrt{2/\pi}}(\mathbf{s})$. □

For a vector $\mathbf{x} \in \mathbb{R}^{m_0}$ and subset of entries S, $\mathbf{x}|_S \in \mathbb{R}^S$ will denote the projection of $\mathbf{x}$ onto S.
In the description and analysis of *FocusedPush*, m denotes the current scale in the halving sequence (ranging from m_0 down to 2); the original output dimension $m_0 = \alpha n$ is fixed throughout.
For notational clarity in the description of *FocusedPush*, assume m_0 is a power of two, so that the active set size can be halved cleanly. For general m_0, one may round the active set sizes to integers; this changes only universal constants and none of the asymptotic conclusions.
We distinguish between the nonterminal scales $m = m_0, m_0/2, \ldots, 4$, where the proof maintains an invariant for the next critical set of size $m/2$, and the terminal scale $m = 2$, where the remaining two critical coordinates are pushed above zero directly.

Algorithm *FocusedPush*(m_0)
State $\mathbf{s} \in \mathbb{R}^{m_0}$
1 For $m = m_0, m_0/2, \ldots, 2$,
2 Let S be the coordinates of the m smallest entries in $\mathbf{s}$.
3 For $t = \lceil 6\gamma^{-1} m_0^{5/8} m^{3/8} \rceil$ steps,
4 On input $\mathbf{a} \sim (1/\sqrt{m_0})\mathcal{N}(\mathbf{0}, \mathbf{I}_{m_0})$,
5 Run $Push(m, m_0)$ on $\mathbf{s}|_S, \mathbf{a}|_S$, obtaining sign x.
6 Update all coordinates with the same sign: $\mathbf{s} \leftarrow \mathbf{s} + x\mathbf{a}$.

Let $L_{\mathrm{FP}}(m_0)$ denote the deterministic number of update steps used by *FocusedPush*(m_0) with the rounded loop counts above. Under the power-of-two convention,

$$\begin{aligned} L_{\mathrm{FP}}(m_0) &= \sum_{k=0}^{\log_2 m_0 - 1} \left\lceil 6\gamma^{-1} m_0^{5/8} (m_0 2^{-k})^{3/8} \right\rceil \\ &\le \sum_{k=0}^{\log_2 m_0 - 1} \left(6\gamma^{-1} m_0\, 2^{-3k/8} + 1 \right) \le C m_0 \end{aligned}$$

for all sufficiently large m_0, where $C < \infty$ is a universal constant.
For the rest of this section, λ is fixed to $\sqrt{2/\pi}$.

Theorem 4.5 (*FocusedPush*)**.** *Assuming $\Phi(\mathbf{s}_0) \le \exp(-499\lambda)$, all entries of the final state of FocusedPush are positive with constant probability.*

Call an entry s of $\mathbf{s}$ *b-safe* if $s > b$. In the following two claims $\mathbf{s}_0$ and $\mathbf{s}_t$ are the initial and final state in iteration m of loop 1.

Claim 4.6. *Assuming $\Phi(\mathbf{s}_0|_S) \le \exp(-\lambda(b-1))$, $\Phi(\mathbf{s}_t|_S) \le \exp(-\lambda(b+2))$, except with probability at most $\exp\left(-3\lambda(m_0/m)^{1/8}\right)$.*

Proof. By (6),

$$\begin{aligned}\mathbb{E}\left[\Phi(\mathbf{s}_t|_S)\right] &\le \exp\Big(-\lambda\left(b-1+6(m_0/m)^{1/8}\right)\Big)\\ &\le \exp(-\lambda(b+2))\,\exp\Big(-3\lambda(m_0/m)^{1/8}\Big)\,.\end{aligned}$$

The conclusion follows from Lemma 3.3. □

Claim 4.7. *For a nonterminal scale $m \ge 4$, assuming all entries of $\mathbf{s}_0$ outside S are $b = 500(m/m_0)^{1/8}$-safe, (1) at most $m/4$ entries of $\mathbf{s}_t$ outside S are not $500(m/2m_0)^{1/8}$-safe and (2) the unnormalized total potential of these unsafe entries is at most $m\exp(-\lambda(b+2))$ after the iteration, except with probability at most $0.007\exp\big(-(m_0/m)^{1/8}\big)$.*

Proof. The sign chosen by *Push* in this iteration depends only on $\mathbf{a}|_S$. Conditional on that sign, the entries of $\mathbf{a}$ outside S remain independent centered Gaussians, so the relevant entries are shifted by independent centered normals of variance t/m_0, where

$$6\gamma^{-1}m_0^{5/8}m^{3/8} \le t \le (6+\gamma)\gamma^{-1}m_0^{5/8}m^{3/8}.$$

Let

$$u = 500\left(1-2^{-1/8}\right)(m/m_0)^{1/8}\,.$$

For an entry not to be $500(m/2m_0)^{1/8}$-safe, the shift must take value at most $-u$. By Lemma 3.4 and the upper bound on t, for any given entry this probability is at most $\exp\big(-44(m_0/m)^{1/8}\big)$. By a union bound the probability that at least $m/4$ offending entries exist is at most, writing $R = (m_0/m)^{1/8} \ge 1$,

$$\begin{aligned}\binom{m_0}{m/4}\exp(-44R)^{m/4} &\le \big(4eR^8\exp(-44R)\big)^{m/4}\\ &\le \big(4eR^8\exp(-40R)\big)^{1/4}\exp(-R)\\ &\le 0.005\exp(-R)\,.\end{aligned}$$

The average unnormalized total potential of the unsafe entries is at most,

$$m_0\,\mathbb{E}\left[\exp(-\lambda X)\,\mathbf{1}\{X \le b-u\}\right]\,,$$

for a normal X of mean b and variance t/m_0. Applying Cauchy–Schwarz as

$$\mathbb{E}[\exp(-\lambda X)\,\mathbf{1}\{X \le b-u\}] \le \sqrt{\mathbb{E}[\exp(-2\lambda X)]}\,\sqrt{\Pr(X \le b-u)},$$

and using the upper bound on t and the Gaussian tail estimate, this is at most

$$\begin{aligned}m_0\sqrt{\mathbb{E}[\exp(-2\lambda X)]}\,\sqrt{\Pr(X \le b-u)} &\le m_0\exp(-\lambda b)\exp\big(\lambda^2 t/m_0\big)\exp(-22R)\\ &\le m_0\exp(-\lambda b)\exp\big(13R^{-3}\big)\exp(-22R)\\ &\le m\exp(-\lambda(b+2))\;0.002\exp(-R)\,.\end{aligned}$$

In the second line we used $\lambda^2 = 2/\pi$, $\gamma > 0.328$, and $t \le (6+\gamma)\gamma^{-1}m_0^{5/8}m^{3/8}$. The last inequality has numerical slack: for $R \ge 1$,

$$R^8\exp\big(2\lambda + 13R^{-3} - 21R\big) \le 0.002.$$

The claim follows from Lemma 3.3. □

Claim 4.8. *For a nonterminal scale $m \geq 4$, assume that all entries of $\mathbf{s}_0$ outside S are $b(m)$-safe and $\Phi(\mathbf{s}_0|_S) \leq \exp(-\lambda(b(m)-1))$ before iteration m, with $b(m) = 500(m/m_0)^{1/8}$. Let S' be the set of $m/2$ smallest entries of $\mathbf{s}$ after this iteration. Then, except with probability $\exp\bigl(-3\lambda(m_0/m)^{1/8}\bigr) + 0.007\exp\bigl(-(m_0/m)^{1/8}\bigr)$, (1) all $b(m/2)$-unsafe entries are contained in S', and (2) $\Phi(\mathbf{s}|_{S'}) \leq \exp(-\lambda\,(b(m/2)-1))$.*

Proof. Assume the exceptional events from the above two claims do not occur. By Claim 4.6, at the end of iteration m, the total (unnormalized) potential in S is at most $m\exp(-\lambda(b(m)+2))$. An entry in S is $b(m/2)$-unsafe only if $s_i < b(m/2) < b(m)$, hence its potential exceeds $\exp(-\lambda b(m))$. If more than $me^{-2\lambda}$ entries of S were $b(m/2)$-unsafe, their potential alone would exceed $me^{-2\lambda} \cdot \exp(-\lambda b(m)) = m\exp(-\lambda(b(m)+2))$, contradicting the total potential bound; so at most $me^{-2\lambda} < m/4$ entries of S are $b(m/2)$-unsafe. By Claim 4.7, at most $m/4$ entries of $\mathbf{s}_t$ outside S are not $b(m/2)$-safe. Altogether there can be at most $m/2$ unsafe entries at the end of iteration m; since S' is the set of the $m/2$ smallest entries, it contains all of them, establishing (1). For (2), S' is contained in the union of the entries of S, the unsafe entries outside S, and $b(m/2)$-safe filler entries needed to bring the size to $m/2$. Their total potential is at most the contribution from S, at most $m\exp(-\lambda(b(m)+2))$, plus the contribution of the unsafe entries outside S, also at most $m\exp(-\lambda(b(m)+2))$ by Claim 4.7, plus the potential of the safe filler entries, at most $(m/2)\exp(-\lambda b(m/2))$. As $b(m) > b(m/2)$,

$$\frac{m}{2}\,\Phi(\mathbf{s}|_{S'}) \;\leq\; 2m\,\exp(-\lambda\,(b(m/2)+2)) + \frac{m}{2}\,\exp(-\lambda b(m/2)) \;\leq\; \frac{m}{2}\,\exp(-\lambda\,(b(m/2)-1))\,,$$

where the last step uses $4e^{-2\lambda} + 1 \leq e^{\lambda}$ (numerically $4 \cdot 0.20 + 1 = 1.80 < e^{0.798} \approx 2.22$). Dividing by $m/2$ gives $\Phi(\mathbf{s}|_{S'}) \leq \exp(-\lambda\,(b(m/2)-1))$. □

Proof of Theorem 4.5. Apply Claim 4.8 inductively over the halving scales $m = m_0, m_0/2, \ldots, 4$, with invariant: all $b(m)$-unsafe entries are contained in the active set S, and $\Phi(\mathbf{s}|_S) \leq \exp(-\lambda(b(m)-1))$. At $m = m_0$ the invariant holds: S consists of all m_0 entries so the outside-S condition is vacuous, and $\Phi(\mathbf{s}_0) \leq \exp(-499\lambda) = \exp(-\lambda(b(m_0)-1))$ by hypothesis. Before the final scale $m = 2$, the active set S of the two smallest coordinates contains all $b(2)$-unsafe entries and satisfies

$$\Phi(\mathbf{s}|_S) \leq \exp(-\lambda(b(2)-1))\,.$$

The final push clears these remaining unsafe entries. Claim 4.6 with $m = 2$ gives

$$\Phi(\mathbf{s}_t|_S) \leq \exp(-\lambda(b(2)+2))\,,$$

except with probability at most $\exp\bigl(-3\lambda(m_0/2)^{1/8}\bigr)$. Since $|S| = 2$, each of these two coordinates is at least $b(2) + 2 - \lambda^{-1}\log 2 > 0$.

It remains to verify that the coordinates outside S stay safe. At the final scale the sign depends only on $\mathbf{a}|_S$ with $|S| = 2$; since $\mathbf{a}|_S$ is independent of the entries of $\mathbf{a}$ outside S, the sign is independent of those increments, which are therefore independent centered Gaussians of variance t/m_0, where

$$t = \left\lceil 6\gamma^{-1}m_0^{5/8}2^{3/8} \right\rceil \leq (6+\gamma)\gamma^{-1}m_0^{5/8}2^{3/8}.$$

Each coordinate starts above $b(2) = 500(2/m_0)^{1/8}$. Hence, by the same Gaussian tail estimate used in Claim 4.7 and a union bound, the probability that any coordinate outside S falls below zero

during the final scale is at most $m_0 \exp\left(-cm_0^{1/8}\right)$ for a universal constant $c > 0$. This is $o(1)$, and is at most 0.001 for all sufficiently large m_0. Thus, outside the exceptional events, every coordinate of the final state is positive.

The failure probability from the inductive scales and the final critical coordinates is at most the sum of,

$$\exp(-3\lambda) + \exp\left(-3\lambda\, 2^{1/8}\right) + \exp\left(-3\lambda\, 2^{1/4}\right) + \cdots \le 0.379\,,$$

and

$$0.007\left(\exp(-1) + \exp\left(-2^{1/8}\right) + \exp\left(-2^{1/4}\right) + \cdots\right) \le 0.020\,.$$

Together with the $o(1)$ final-scale complement error, the success probability is bounded below by a positive universal constant for all sufficiently large m_0. □

4.3 Finding Collisions

We now adapt the margin-building procedure to collision finding. The algorithm first creates the desired Hamming separation by assigning opposite signs to $\mathbf{x}$ and $\mathbf{y}$ on an initial block of coordinates. After this block, every new coordinate receives the same sign in both vectors, so the overlap remains fixed. The relevant state is then the pointwise minimum,

$$\mathbf{s} = \min(\mathbf{Mx}, \mathbf{My}),$$

which evolves by the same update rule as a single margin vector whenever the two assignments use the same sign. Since Corollary 4.3 applies even from a moderately large starting potential, we can run *Push* on this minimum state and then use *FocusedPush* for the final refinement.

Algorithm *CollisionFinder*$(\mathbf{M}, \theta; \beta_1)$
Input columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of $\mathbf{M} \in \mathbb{R}^{m \times n}$
State $\mathbf{s} \in \mathbb{R}^m$, the pointwise minimum of the two current partial margin vectors, initially $\mathbf{s} = \mathbf{0}$

Let $L_{\mathrm{FP}} = L_{\mathrm{FP}}(m)$, and reserve the final L_{FP} columns for refinement.
Initialization:
    For $t = 1, \ldots, \lfloor \beta_1 m \rfloor$, set $x_t \leftarrow -1$ and $y_t \leftarrow 1$.
    Update $\mathbf{s} \leftarrow \min(\mathbf{Mx}, \mathbf{My})$.
Drift phase:
    For $t = \lfloor \beta_1 m \rfloor + 1, \ldots, n - L_{\mathrm{FP}}$,
        On input column $\mathbf{a}_t$, set $z_t \leftarrow \mathit{PushSign}(\mathbf{s}, \mathbf{a}_t)$.
        Set $x_t \leftarrow z_t$ and $y_t \leftarrow z_t$, and update $\mathbf{s} \leftarrow \mathbf{s} + z_t \mathbf{a}_t$.
Refinement:
    Update $\mathbf{s} \leftarrow \mathbf{s} - \theta \mathbf{1}$.
    Run *FocusedPush* from $\mathbf{s}$ on columns $n - L_{\mathrm{FP}} + 1, \ldots, n$,
    using each output sign for both $\mathbf{x}$ and $\mathbf{y}$.

Output $(\mathbf{x}, \mathbf{y})$.

In the ABP case $\theta = 0$, if every entry of $\mathbf{Mx}$ and $\mathbf{My}$ is positive at termination then $\mathrm{sign}(\mathbf{Mx}) = \mathrm{sign}(\mathbf{My}) = \mathbf{1}_m$, so $(\mathbf{x}, \mathbf{y})$ is a collision for ABP.

Proof of Theorem 4.1. Let $\mathbf{B} = \mathbf{A}/\sqrt{\alpha}$ and set $\rho = \kappa/\sqrt{\alpha}$. Set $D = \lceil \delta n/2 \rceil$ and $\beta_1 = D/m$, and run $CollisionFinder(\mathbf{B}, \rho; \beta_1)$. Since the algorithm initializes for $\lfloor \beta_1 m \rfloor = D$ coordinates, the initial inner product is exactly $\mathbf{x}^\top \mathbf{y} = n - 2D$. For all sufficiently large n, $D \le n/2$, so $|\mathbf{x}^\top \mathbf{y}| = n - 2D \le (1-\delta)n$.

Let $\mathbf{s}$ be the pointwise minimum of $\mathbf{Bx}$ and $\mathbf{By}$ after these assignments. For each coordinate, the assigned part of $(\mathbf{Bx})_i$ is a Gaussian $G \sim \mathcal{N}(0, \beta_1)$ and the assigned part of $(\mathbf{By})_i$ is $-G$, so $s_i = -|G|$. Hence,

$$\mathbb{E}[\Phi(\mathbf{s})] = \mathbb{E}[\exp(\lambda |G|)] \le 2 \exp\bigl(\lambda^2 \beta_1 / 2\bigr) , \qquad \lambda = \sqrt{2/\pi} .$$

The parameter β_2 below is used only in the analysis, as a lower bound on the number of drift steps available before the reserved refinement block. Take

$$\beta_2 = C\,(1 + \beta_1 + \rho) ,$$

for a sufficiently large universal constant C. Let $L_{\mathrm{FP}} = L_{\mathrm{FP}}(m)$. The number of columns needed for initialization, $\beta_2 m$ drift steps, and the reserved refinement block is at most,

$$D + \beta_2 m + L_{\mathrm{FP}} \le D + \beta_2 m + Cm \le \left(\frac{\delta}{2} + \alpha(\beta_2 + C)\right) n + 1 \le \left(\frac{\delta}{2} + C'\delta + C'\kappa\sqrt{\alpha} + C'\alpha\right) n,$$

for a universal constant C', with the additive $+1$ absorbed for all sufficiently large n. Choose $\delta_0 > 0$ small enough that $\delta/2 + C'\delta < 1/3$ for all $\delta < \delta_0$. Then choose $c > 0$ small enough that $C'c < 1/3$, and finally choose $\alpha_0(\delta) > 0$ small enough that $C'\alpha < 1/3$ for all $\alpha < \alpha_0(\delta)$. Since $\kappa \le c/\sqrt{\alpha}$, the number of used columns is at most n. Thus the actual drift phase in $CollisionFinder$ has length at least $\beta_2 m$, because it runs through column $n - L_{\mathrm{FP}}$. Applying $Push$ for this many or more steps gives,

$$\mathbb{E}[\Phi] \le 2 \exp\bigl(\lambda^2 \beta_1/2 - \gamma' \beta_2\bigr) ,$$

where $\gamma' = \lambda\gamma > 0$ is the fixed decay constant from Corollary 4.3. The choice of C above ensures,

$$\mathbb{E}[\Phi] \le \frac{1}{2} \exp(-499\lambda - \lambda\rho) .$$

Since the expectation is at most half of $\exp(-499\lambda - \lambda\rho)$, Lemma 3.3 gives,

$$\Pr[\Phi > \exp(-499\lambda - \lambda\rho)] \le \frac{1}{2}.$$

Thus, with probability at least $1/2$, we have $\Phi \le \exp(-499\lambda - \lambda\rho)$ before the refinement. During the drift phase and the refinement, every new column receives the same sign in $\mathbf{x}$ and $\mathbf{y}$, so the pointwise minimum evolves according to the same update rule since $\min(u + za, v + za) = \min(u, v) + za$. The shift $\mathbf{s} \leftarrow \mathbf{s} - \rho\mathbf{1}$ multiplies the potential by $e^{\lambda\rho}$, so after the shift we have $\Phi \le \exp(-499\lambda)$, and Theorem 4.5 applies to the reserved final block.

The algorithm assigns every one of the n coordinates: the initialization and drift phases assign the first $n - L_{\mathrm{FP}}$ coordinates, and $FocusedPush$ assigns the reserved final L_{FP} coordinates. After initialization, every assigned coordinate receives the same sign in $\mathbf{x}$ and $\mathbf{y}$, so the final inner product remains exactly $n - 2D \le (1-\delta)n$. If $c_{\mathrm{FP}} > 0$ denotes the success probability in Theorem 4.5, then conditioning on the Markov event, Theorem 4.5 gives success probability at least c_{FP}; hence the

intersection has probability at least $c_{\mathrm{FP}}/2$. On this event, *CollisionFinder* returns a pair $(\mathbf{x}, \mathbf{y})$ satisfying,

$$\min_i (\mathbf{B}\mathbf{x})_i \geq \rho, \qquad \min_i (\mathbf{B}\mathbf{y})_i \geq \rho, \qquad \left|\mathbf{x}^\top \mathbf{y}\right| \leq (1-\delta)\, n.$$

Multiplying the first two inequalities by $\sqrt{\alpha}$ gives,

$$\min_i (\mathbf{A}\mathbf{x})_i \geq \kappa, \qquad \min_i (\mathbf{A}\mathbf{y})_i \geq \kappa,$$

as required. □

4.4 Simulations

This subsection is purely empirical and is not used in any of the formal results below.

We implemented a version of the *CollisionFinder* algorithm to explore the empirical transition. The implementation processes columns of $\mathbf{A} \in \mathbb{R}^{m \times n}$ online, uses the one-sided potential $\Phi(s) = \frac{1}{m}\sum_{i=1}^{m} e^{-\lambda s_i/\sqrt{\alpha}}$ with $\lambda = \sqrt{2/\pi}$, and applies the drift-based decision rule described in Section 4. All experiments use $\delta = 0.1$ and Gaussian matrices $\mathbf{A}_{ij} \sim \mathcal{N}(0, 1/n)$. The implementation includes practical choices, so success probabilities should be interpreted as empirical evidence for the scaling rather than as a direct verification of the proof.

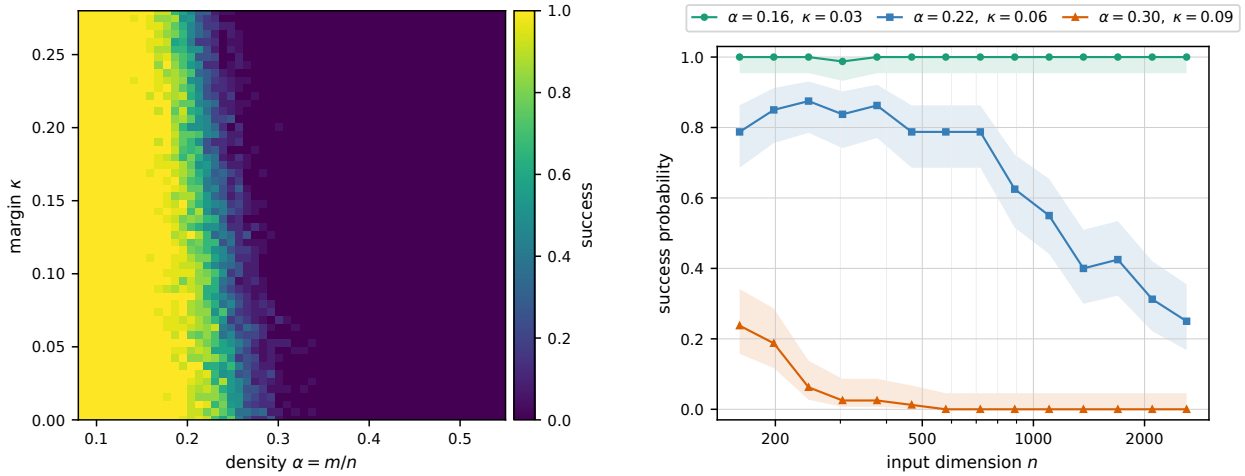

Figure 1: Empirical behavior of the online collision heuristic. Left: Phase diagram on a 54×54 grid with $n = 420$. The observed transition is consistent with dependence on $\kappa\sqrt{\alpha}$. Right: Success probability versus n for three configurations. The easy regime $(\alpha, \kappa) = (0.16, 0.03)$ remains near 1, the transition regime $(0.22, 0.06)$ decreases with n, and the low-success regime $(0.30, 0.09)$ quickly decays to 0.

Figure 1 shows two empirical patterns. The phase diagram shows the success probability over a 54×54 grid spanning $\alpha \in [0.08, 0.55]$ and $\kappa \in [0, 0.28]$ for $n = 420$. While the exact transition curve fluctuates mildly due to implementation choices, a clear inverse relationship between α and κ emerges: the algorithm succeeds almost always in the lower-left region, where $\kappa\sqrt{\alpha}$ is smaller, and almost never in the upper-right region, where $\kappa\sqrt{\alpha}$ is larger. This diagonal pattern is consistent with the scaling in Theorem 4.1, suggesting that the empirical boundary is controlled by the dimensionless parameter $\kappa\sqrt{\alpha}$.

To test whether this empirical transition persists as n grows, we ran the algorithm for $n \in [160, 2600]$ across three representative configurations. In the easy regime $(\alpha, \kappa) = (0.16, 0.03)$, the success probability remains near 1 across all n. In the transition regime $(0.22, 0.06)$ it decreases with n, while in the low-success regime $(0.30, 0.09)$ it quickly decays toward 0.

Overall, while precise probabilities are sensitive to implementation details, the experiments support the prediction that the empirical transition is controlled by the product $\kappa\sqrt{\alpha}$.

5 Establishing OGP

In this section we analyze a randomized activation function. For each coordinate $i \in [m]$ we fix an odd integer $K \geq 1$ and draw K i.i.d. thresholds t_{ij} uniformly at random from the range $[-\kappa, \kappa]$, for $j \in [K]$, and define the coordinate-wise activation,

$$\varphi_i(x) \;=\; \prod_{j=1}^{K} \operatorname{sign}(x - t_{ij})\,, \qquad \varphi(\mathbf{A}\mathbf{x})_i \;=\; \varphi_i\left((\mathbf{A}\mathbf{x})_i\right)\,.$$

The thresholds are independent across i and j, and independent of $\mathbf{A}$. Note that the activation is constant on each tail: for $x > \kappa$ all factors are $+1$, while for $x < -\kappa$ all factors are -1, so the product is $(-1)^K = -1$ for odd K. We restrict to odd K for simplicity; the oddness is used in the row-wise contraction argument below, and we do not analyze the even-K case. The proof is a first-moment argument over admissible r-tuples of collision pairs. For a fixed row, admissibility and a selection lemma produce $r+1$ Gaussian projections whose covariances are separated from one. The random thresholds then give a row-wise contraction, summarized by the exponent $\Lambda_{r,\delta}(\kappa, K)$. Combining this row bound with entropy counting and independence across rows proves Theorem 5.1. The ensemble version repeats the row-wise argument for correlated-prefix instances, where admissibility is controlled by $1 - t/n$, and the online lower bound follows by coupling r runs of an online algorithm and symmetrizing the suffix signs.

Fix $r \geq 2$ and $\delta \in (0, 1]$, and let $2 < c_1 < c_2 < 2c_1$ be constants. Throughout this section, we will let $(X_i, Y_i)_{i=1}^r$ be a set of centered Gaussian variables with unit variances. Let Z_i, Z_j be two random variables from the set. For Gaussian projections arising from Boolean vectors in a common instance,

$$d(\langle a, u\rangle, \langle a, v\rangle) = \frac{1 - \operatorname{Cov}(\langle a, u\rangle, \langle a, v\rangle)}{2} = \frac{1}{n}\Delta_{\mathrm{Ham}}(u, v),$$

and hence d satisfies the triangle inequality on the variables considered below. More generally, the selection argument below is used only for Gaussian families for which this quantity defines a metric.

Let $a, b, \tau > 0$ be constants. A set of centered Gaussian variables $(X_i, Y_i)_{i=1}^r$ with unit variances is said to be an *admissible set* if it satisfies the following two conditions:

1. *(Internal overlap).* For every i, $d(X_i, Y_i) \geq \tau$.
2. *(External overlap).* For every $i \neq j$, $a < d(X_i, X_j) + d(Y_i, Y_j) < b$.

For the Gaussian projections of a δ-admissible family from Definition 3.2, we have,

$$\tau = \delta/2, \qquad a = \delta/c_2, \qquad b = \delta/c_1.$$

Conditioned on a pair (X, Y), a single threshold falls between them with probability $\Theta(|X-Y|/\kappa)$ while both lie in $[-\kappa, \kappa]$; K independent thresholds then amplify this effect multiplicatively. Let

$$\eta_{\text{si}} = \min\left\{1 - \frac{2}{c_1},\, \frac{1}{2}\left(\frac{2}{c_2} - \frac{1}{c_1}\right)\right\}, \qquad \eta_{\text{ens}} = \min\left\{\frac{1}{c_2},\, 1 - \frac{1}{c_1},\, 1\right\}, \qquad \eta_0 = \frac{1}{2}\min\{\eta_{\text{si}}, \eta_{\text{ens}}\}.$$

Thus the selection lemma below, applied with $\gamma = \eta_0\delta/2$, gives covariance separation at least $\eta_0\delta$, and the same η_0 is no larger than the separated-subfamily constants used in the ensemble argument. Let $\bar{c} > 0$ be the absolute contraction constant from Claim 5.4. Fix constants $c_0, C_0, C_1 > 0$, depending only on c_1, c_2, with $c_0 = \bar{c}\sqrt{\eta_0}$ and C_0, C_1 large enough for Lemma 5.6, and set

$$p_{\max} = \min\left\{1, \frac{\kappa}{c_0\sqrt{\delta}}\right\}.$$

For compactness, define

$$\Lambda_{r,\delta}(\kappa, K) = -\log_2\left(2^{-r} + \inf_{p\in(0,p_{\max}]}\left[\left(1 - c_0 p\sqrt{\delta}/\kappa\right)^K + C_0\, r^2 p + C_1\, r\, e^{-\kappa^2/2}\right]\right).$$

The main result of this section is the following.

Theorem 5.1 (First-Moment OGP Bound)**.** *Fix constants $r \geq 2$, $\delta \in (0, 1]$, and $2 < c_1 < c_2 < 2c_1$, a parameter $\kappa > 0$, an odd integer $K \geq 1$, and $\alpha \in (0, 1]$. Let $\mathsf{Coll}_{r,\delta} = |S(r, \delta, c_1, c_2; \mathbf{A})|$ be the number of δ-admissible r-tuples of collisions (as in Definition 3.2). The expectation $\mathbb{E}[\cdot]$ is over the draw of the Gaussian matrix $\mathbf{A}$ and the randomized thresholds of φ. Then,*

$$\frac{1}{n}\log_2 \mathbb{E}[\mathsf{Coll}_{r,\delta}] \leq 2 + 2(r-1)H(\delta/(2c_1)) - \alpha\,\Lambda_{r,\delta}(\kappa, K) + o(1). \tag{7}$$

In particular, once

$$\alpha\Lambda_{r,\delta}(\kappa, K)$$

exceeds $2 + 2(r-1)H(\delta/(2c_1))$ by a fixed positive constant, the exponent in (7) is negative and the first moment of r-collisions vanishes exponentially in n.

Corollary 5.2 (Single-Instance r-OGP)**.** *There exist constants $c > 0$ and $C < \infty$, depending only on c_1, c_2, such that the following holds for every $\alpha \in (0, 1]$, $\kappa > 0$, and odd integer $K \geq 1$. Assume $H(\delta/(2c_1)) \leq c\alpha$, $\alpha r \geq C$, and $\Lambda_{r,\delta}(\kappa, K) \geq (1-c)r$. Then no r-collisions exist except with probability $\exp(-\Omega(\alpha r n))$. Consequently, with probability at least $1 - 2^{-\Omega(n)}$, the randomized activation function φ satisfies r-OGP.*

Proof. By Theorem 5.1, the first-moment exponent is at most

$$2 + 2(r-1)H(\delta/(2c_1)) - \alpha\Lambda_{r,\delta}(\kappa, K) + o(1) \leq -\Omega(\alpha r)\,.$$

The expected number of r-collisions is then at most $\exp(-\Omega(\alpha r n))$. By Lemma 3.3, no r-collision exists except with such probability. □

The quantity $\Lambda_{r,\delta}(\kappa, K)$ is the row-wise collision exponent: the proof below shows that a fixed admissible r-tuple collides on one row with probability at most $2^{-\Lambda_{r,\delta}(\kappa,K)}$. A simple sufficient

regime is obtained by choosing $p = \Theta(2^{-r}/r^2)$: for constants $c > 0$ and $C < \infty$ depending only on c_1, c_2, the conditions

$$\kappa^2 \geq C(r + \log r), \qquad K \text{ is odd and } K \geq C\, 2^r r^3\, \frac{\kappa}{\sqrt{\delta}}$$

imply

$$\Lambda_{r,\delta}(\kappa, K) \geq (1 - c)r.$$

The extra factor of r in the lower bound on K comes from making $\left(1 - c_0 p\sqrt{\delta}/\kappa\right)^K$ comparable to 2^{-r}, while the choice $p = \Theta(2^{-r}/r^2)$ keeps the $C_0\, r^2 p$ error at the same scale.

The proof of Theorem 5.1 has two ingredients. First, the overlap constraints give the entropy count below. Second, for every fixed admissible family, the Gaussian variables seen by one row contain a large separated subfamily; this forces a uniform contraction under the randomized thresholds. Independence across rows then gives the stated first moment.

Lemma 5.3 (Entropy Count for Admissible Families)**.** *For fixed constants $r \geq 2$, $\delta \in (0, 1]$, and $2 < c_1 < c_2 < 2c_1$,*

$$\frac{1}{n} \log_2 |T(r, \delta, c_1, c_2)| \leq 2 + 2(r - 1)H(\delta/(2c_1)) + o(1).$$

Proof. The first pair $(\mathbf{x}_1, \mathbf{y}_1)$ has at most 2^{2n} choices. For each $i > 1$, the *lower* external-overlap bound $(\mathbf{x}_i^\top \mathbf{x}_1 + \mathbf{y}_i^\top \mathbf{y}_1)/(2n) \geq 1 - \delta/c_1$, combined with $u^\top v = n - 2\Delta_{\mathrm{Ham}}(u, v)$, gives

$$\Delta_{\mathrm{Ham}}(\mathbf{x}_i, \mathbf{x}_1) + \Delta_{\mathrm{Ham}}(\mathbf{y}_i, \mathbf{y}_1) \leq \frac{\delta n}{c_1}.$$

Thus, after fixing $(\mathbf{x}_1, \mathbf{y}_1)$, the logarithm of the number of choices for $(\mathbf{x}_i, \mathbf{y}_i)$ is at most

$$\log_2 \sum_{\ell \leq \delta n/c_1} \binom{2n}{\ell} \leq (2H(\delta/(2c_1)) + o(1))\, n.$$

Multiplying this bound over $i = 2, \ldots, r$ proves the claim. The internal-overlap condition and the remaining external-overlap conditions can only reduce the count. □

Claim 5.4. *Let $Z_1, \ldots, Z_{2s}$ be centered Gaussian random variables with unit variances. Suppose that among them there are $s + 1$ variables such that, for every distinct pair Z_i, Z_j in this subfamily,*

$$\mathrm{Cov}(Z_i, Z_j) \leq 1 - \eta.$$

Then there is an absolute constant $\bar{c} > 0$ such that, with φ the randomized activation function and for every $p \in (0, 1)$ with $\bar{c}p\sqrt{\eta}/\kappa \leq 1$,

$$\mathbb{E}\left[\prod_{i=1}^{2s} \varphi(Z_i)\right] \leq (1 - \bar{c}p\sqrt{\eta}/\kappa)^K + \binom{s+1}{2} p + 2s\, e^{-\kappa^2/2}.$$

Proof. By a union bound and Lemma 3.4, $|Z_i| \le \kappa$ for all i except with probability at most $2s\, e^{-\kappa^2/2}$. By another union bound and Lemma 3.4, since each selected difference has variance at least 2η, $|Z_i - Z_j| \ge p\sqrt{\eta} =: q$ for all selected pairs except with probability at most $\binom{s+1}{2}p$.

On this good event all Z_i lie in $[-\kappa, \kappa]$, and the selected $s+1$ variables are pairwise separated by at least q. We show that $\operatorname{sign}(Z_1 - T) \cdots \operatorname{sign}(Z_{2s} - T)$ takes value -1 with probability at least $q/2\kappa$ for a threshold T chosen uniformly from $[-\kappa, \kappa]$.

Let π order $Z_1, \ldots, Z_{2s}$ increasingly. For any threshold $T \in (Z_{\pi(j)}, Z_{\pi(j+1)})$, exactly j of the quantities $Z_i - T$ are negative, so the sign product equals $(-1)^j$ on this interval. Since all $Z_i \in [-\kappa, \kappa]$, these adjacent intervals are not clipped by the threshold range. Thus, up to endpoints of measure zero, the set of thresholds for which the product equals -1 is precisely the union of the odd-parity intervals

$$(Z_{\pi(1)}, Z_{\pi(2)}),\ (Z_{\pi(3)}, Z_{\pi(4)}),\ \ldots,\ (Z_{\pi(2s-1)}, Z_{\pi(2s)}),$$

and

$$\Pr_T\left[\prod_{i=1}^{2s} \operatorname{sign}(Z_i - T) = -1\right] = \frac{1}{2\kappa}\sum_{j=1}^{s}(Z_{\pi(2j)} - Z_{\pi(2j-1)}).$$

If every odd-parity interval had length less than q, then each pair of ordered positions $(\pi(2j-1), \pi(2j))$ could contain at most one selected variable, because selected variables are pairwise separated by at least q. These s pairs partition the $2s$ positions, so they would contain at most s selected variables in total, contradicting the existence of $s+1$ selected variables. Hence some odd-parity interval has length at least q, and it contributes at least $q/2\kappa$ to the probability above.

In summary, $\operatorname{sign}(Z_1 - T) \cdots \operatorname{sign}(Z_{2s} - T)$ takes value -1 with probability at least $q/(2\kappa)$, and hence its one-threshold expectation satisfies

$$g := \mathbb{E}_T\left[\prod_{i=1}^{2s} \operatorname{sign}(Z_i - T)\right] \le 1 - q/\kappa \le 1 - \bar{c}p\sqrt{\eta}/\kappa$$

for a fixed absolute $\bar{c} > 0$. Here $g = g(Z_1, \ldots, Z_{2s})$ is a function of the Gaussian values. Since the K thresholds $T_1, \ldots, T_K$ are independent of each other and of the Z's, the conditional expectation over thresholds given the Z values satisfies

$$\mathbb{E}\left[\prod_{i=1}^{2s}\varphi(Z_i)\,\middle|\, Z_1, \ldots, Z_{2s}\right] = \mathbb{E}\left[\prod_{j=1}^{K}\prod_{i=1}^{2s}\operatorname{sign}(Z_i - T_j)\,\middle|\, Z_1, \ldots, Z_{2s}\right] = g(Z_1, \ldots, Z_{2s})^K.$$

On the good event, $g \le 1 - \bar{c}p\sqrt{\eta}/\kappa$. Since K is odd the map $x \mapsto x^K$ is increasing, so on the good event $g^K \le (1 - \bar{c}p\sqrt{\eta}/\kappa)^K$. Using $|g| \le 1$ on the bad event, the bound on the unconditional expectation follows from the law of total probability. □

Lemma 5.5 (Selection Lemma)**.** *Let $(X_i, Y_i)_{i=1}^r$ be an admissible set, and assume that $d(Z, Z') = (1 - \operatorname{Cov}(Z, Z'))/2$ is a metric on $\{X_i, Y_i : i \in [r]\}$. If $\tau > b$ and $2a > b$, then for every $\gamma < \min\{\tau - b, (2a - b)/4\}$, there exists a subset of $r+1$ variables $Z_1, \ldots Z_{r+1}$ satisfying $d(Z_i, Z_j) \ge \gamma$.*

Proof. Fix any $\gamma < \min\{\tau - b, (2a - b)/4\}$, and define the set,

$$S = \{i \in [r] : d(X_i, X_j) \ge \gamma \text{ for all } j \ne i\}.$$

To show the statement we condition on the set S.

- If $S = \{1, \ldots, r\}$, then we may take all X_i and one arbitrary Y_j. The statement follows since for all i, j,
$$d(X_i, Y_j) \geq d(X_i, Y_i) - d(Y_i, Y_j) \geq \tau - b > \gamma,$$
where we used the triangle inequality and the overlap conditions.

- If $S \neq \{1, \ldots r\}$, we take all X_i for $i \in S$, all Y_j for $j \notin S$, and an arbitrary X_{j^*} for $j^* \notin S$. By definition of S, the condition is satisfied for any pairs of X's. Distances between X's and Y's are again bounded by $d(X_i, Y_j) \geq \tau - b$. It remains to bound the distances between the Y's.

  We claim that if $d(X_i, X_j) < \gamma$, then $d(Y_j, Y_k) \geq \gamma$ for every $k \neq j$. For $k = i$, the external lower bound gives $d(Y_j, Y_i) \geq a - d(X_i, X_j) > a - \gamma > \gamma$, since $\gamma < a/2$. For distinct i, j, k, we will show it is impossible to have both $d(X_i, X_j) < \gamma$ and $d(Y_j, Y_k) < \gamma$, provided $4\gamma < 2a - b$. Assume for contradiction it is possible. By the external overlap condition, we get $d(Y_i, Y_j) \geq a - \gamma$. By the triangle inequality, $d(Y_i, Y_k) \geq a - 2\gamma$. From external overlap, we obtain $d(X_i, X_k) \leq b - a + 2\gamma$. By triangle inequality again, we get $d(X_j, X_k) \leq b - a + 3\gamma$. Adding $d(Y_j, Y_k) < \gamma$, we finally get,
$$d(X_j, X_k) + d(Y_j, Y_k) < b - a + 4\gamma\,,$$
which contradicts the lower bound $a \leq d(X_j, X_k) + d(Y_j, Y_k)$. Therefore, if $d(X_i, X_j) < \gamma$, then $d(Y_j, Y_k) \geq \gamma$ for all $k \neq j$. It follows that all selected Y's are pairwise at distance at least γ, and hence all pairwise distances in the chosen set are at least γ, completing the proof. □

Lemma 5.6 (Uniform Row Contraction)**.** *There exist constants $c_0, C_0, C_1 > 0$, depending only on c_1, c_2, such that the following holds. Let $(X_i, Y_i)_{i=1}^r$ be an admissible set, and assume that $d(Z, Z') = (1 - \mathrm{Cov}(Z, Z'))/2$ is a metric on $\{X_i, Y_i : i \in [r]\}$. Then in the randomized oscillating activation model with odd K, for every $p \in (0, p_{\max}]$,*
$$\Pr[\varphi(X_j) = \varphi(Y_j)\ \forall j] \leq 2^{-r} + \left(1 - c_0 p\sqrt{\delta}/\kappa\right)^K + C_0\, r^2 p + C_1\, r\, e^{-\kappa^2/2}.$$

Proof. We arithmetize the collision probability as follows:
$$\Pr[\varphi(X_j) = \varphi(Y_j)\ \forall j] = \mathbb{E}\left[\prod_{j=1}^{r} \frac{1 + \varphi(X_j)\varphi(Y_j)}{2}\right] = 2^{-r} \sum_{S \subseteq [r]} \mathbb{E}\left[\prod_{j \in S} \varphi(X_j)\varphi(Y_j)\right].$$

The term $S = \emptyset$ is 1. Fix any nonempty S of size s. The restricted family $(X_j, Y_j)_{j \in S}$ remains admissible with the same parameters τ, a, b. By Lemma 5.5, applied to this restricted family with $\gamma = \eta_0\delta/2$, there exist $s + 1$ variables among $\{X_j, Y_j : j \in S\}$ such that all pairwise covariances are at most $1 - \eta_0\delta$. Applying Claim 5.4 with $\eta = \eta_0\delta$ yields
$$\mathbb{E}\left[\prod_{j \in S} \varphi(X_j)\varphi(Y_j)\right] \leq \left(1 - c_0 p\sqrt{\delta}/\kappa\right)^K + C_0\, s^2 p + C_1\, s\, e^{-\kappa^2/2},$$
where $c_0 = \bar{c}\sqrt{\eta_0}$, and C_0, C_1 absorb the absolute constants in Claim 5.4. Since $s \leq r$, each nonempty term is at most
$$\left(1 - c_0 p\sqrt{\delta}/\kappa\right)^K + C_0\, r^2 p + C_1\, r\, e^{-\kappa^2/2}.$$

Therefore

$$\Pr[\varphi(X_j) = \varphi(Y_j)\ \forall j] \le 2^{-r} + \left(1 - c_0 p\sqrt{\delta}/\kappa\right)^K + C_0\, r^2 p + C_1\, r\, e^{-\kappa^2/2},$$

which is the claimed bound. Taking the infimum over $p \in (0, p_{\max}]$ gives the row bound $2^{-\Lambda_{r,\delta}(\kappa,K)}$ used below. □

Proof of Theorem 5.1. Consider a fixed δ-admissible family of r pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^r$ and a fixed row $\mathbf{a}$ of $\mathbf{A}$, where $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/n)$. For each $i \in [r]$, let

$$X_i = \langle \mathbf{a}, \mathbf{x}_i \rangle, \qquad Y_i = \langle \mathbf{a}, \mathbf{y}_i \rangle.$$

Then $(X_i, Y_i)_{i=1}^r$ is a jointly Gaussian family with unit variances. Moreover, since $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^r$ is δ-admissible in the sense of Definition 3.2, the Gaussian family $(X_i, Y_i)_{i=1}^r$ is admissible in the above sense, with

$$\tau = \delta/2, \qquad a = \delta/c_2, \qquad b = \delta/c_1.$$

For Gaussian projections from Boolean vectors, $d(\langle \mathbf{a}, u\rangle, \langle \mathbf{a}, v\rangle) = \Delta_{\mathrm{Ham}}(u, v)/n$, so the metric assumption in Lemma 5.6 holds. Since $2 < c_1 < c_2 < 2c_1$, Lemma 5.5 applies, and hence so does Lemma 5.6. Therefore the single-row collision probability is bounded by

$$\Pr[\varphi(X_i) = \varphi(Y_i)\ \forall i \in [r]] \le 2^{-\Lambda_{r,\delta}(\kappa,K)}.$$

Since the rows of $\mathbf{A}$ are independent, the probability that this fixed δ-admissible family collides across all $m = \alpha n$ rows is at most

$$2^{-\alpha n \Lambda_{r,\delta}(\kappa,K)}.$$

Therefore,

$$\log_2 \mathbb{E}[\mathsf{Coll}_{r,\delta}] \le \log_2 |T(r, \delta, c_1, c_2)| - \alpha n \Lambda_{r,\delta}(\kappa, K).$$

The desired bound follows after applying Lemma 5.3 and dividing by n. □

5.1 Ensemble r-OGP

The r-OGP established above concerns configurations within a single random instance. For the hardness argument, we require the corresponding *ensemble* r-OGP, which forbids such configurations across a correlated family of instances that share a common prefix and have independent suffixes: for any $t \in \{0, \dots, n\}$ consider r correlated instances that share the first t columns and have independent suffixes (the randomized thresholds of the activation, when present, are shared across instances).

Definition 5.7 (Ensemble r-OGP)**.** *Fix $r \ge 2$ and $\delta \in (0, 1]$, and let (c_1, c_2) be the forbidden external-overlap window from Definition 3.2. For $t \in \{0, 1, \dots, n\}$, let $\mathsf{A}^{(t)} = (\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(r)})$ be the correlated ensemble in which the first t columns coincide across all $\mathbf{A}^{(j)}$ and the remaining $n - t$ columns are independent (with the same randomized thresholds shared across all instances, when present). A family $\{(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})\}_{j=1}^r \subseteq (\{-1, 1\}^n)^2$ is called* t-prefix aligned *if*

$$\mathbf{x}^{(1)}_{\le t} = \cdots = \mathbf{x}^{(r)}_{\le t} \qquad \text{and} \qquad \mathbf{y}^{(1)}_{\le t} = \cdots = \mathbf{y}^{(r)}_{\le t}.$$

It is called ensemble δ-admissible *if it satisfies the following conditions.*

1. (Internal Overlap). *For every* $j \in [r]$, $\frac{1}{n}\left|\langle \mathbf{x}^{(j)}, \mathbf{y}^{(j)} \rangle\right| \le 1 - \delta$;

2. (External Overlap). *For every distinct* $j \neq j'$,

$$1 - \frac{\delta}{c_1} \le \frac{1}{2n}(\langle \mathbf{x}^{(j)}, \mathbf{x}^{(j')} \rangle + \langle \mathbf{y}^{(j)}, \mathbf{y}^{(j')} \rangle) \le 1 - \frac{\delta}{c_2}.$$

Let $\mathsf{Coll}^{\mathrm{ens}}_{r,\delta}(t)$ *denote the number of* t*-prefix aligned, ensemble* δ*-admissible families such that each pair is a collision on its own instance:*

$$\varphi(\mathbf{A}^{(j)}\mathbf{x}^{(j)}) = \varphi(\mathbf{A}^{(j)}\mathbf{y}^{(j)}) \qquad \text{for all } j \in [r].$$

We say ensemble r-OGP holds at (α, δ) *if, for every fixed* $t \in \{0, 1, \ldots, n\}$ *satisfying*

$$\frac{\delta}{c_2} < 1 - \frac{t}{n} < \frac{\delta}{c_1},$$

the corresponding correlated-prefix ensemble satisfies

$$\Pr_{\mathbf{A}^{(t)}}\left[\mathsf{Coll}^{\mathrm{ens}}_{r,\delta}(t) = 0\right] \ge 1 - 2^{-\Omega(n)}.$$

This is a fixed-t *statement, uniform over choices of* t *in the displayed window; it does not assert a simultaneous union bound over all* t*.*

Lemma 5.8 (Ensemble First-Moment Bound)**.** *For the odd-integer* K *randomized activation above, fix* $r \ge 2$, $\delta \in (0, 1]$, $2 < c_1 < c_2 < 2c_1$, $\alpha \in (0, 1]$, $\kappa > 0$, *and* $t \in \{0, \ldots, n\}$ *satisfying*

$$\frac{\delta}{c_2} < 1 - \frac{t}{n} < \frac{\delta}{c_1}.$$

Then,

$$\frac{1}{n}\log_2 \mathbb{E}\left[\mathsf{Coll}^{\mathrm{ens}}_{r,\delta}(t)\right] \le 2 + 2(r-1)\left(1 - \frac{t}{n}\right) - \alpha\,\Lambda_{r,\delta}(\kappa, K) + o(1).$$

Proof. We follow the same first-moment argument as in the proof of Theorem 5.1.

Fix a candidate family $(x^{(j)}, y^{(j)})_{j=1}^{r}$ satisfying the ensemble constraints. For each row of the matrices, define the Gaussian projections

$$X_j = \langle a^{(j)}, x^{(j)} \rangle, \qquad Y_j = \langle a^{(j)}, y^{(j)} \rangle,$$

where the row vectors $a^{(j)}$ are jointly Gaussian according to the correlated ensemble.

The internal overlap condition implies

$$d(X_j, Y_j) = \frac{1 - \mathrm{Cov}(X_j, Y_j)}{2} \ge \delta/2.$$

For distinct $j \neq j'$, the suffix rows of $a^{(j)}$ and $a^{(j')}$ are independent, and only the shared prefix contributes to the covariance. Since the family is t-prefix aligned,

$$\mathrm{Cov}(X_j, X_{j'}) = \frac{t}{n}, \qquad \mathrm{Cov}(Y_j, Y_{j'}) = \frac{t}{n}.$$

This holds because $\mathbf{a}^{(j)}$ and $\mathbf{a}^{(j')}$ share their first t columns, have independent suffixes, and prefix alignment gives $\mathbf{x}^{(j)}_{\le t} = \mathbf{x}^{(j')}_{\le t}$; therefore $\mathrm{Cov}(X_j, X_{j'}) = (1/n)\,\mathbf{x}^{(j)\top}_{\le t}\,\mathbf{x}^{(j')}_{\le t} = t/n$, and similarly for $Y_j, Y_{j'}$. Therefore

$$d(X_j, X_{j'}) + d(Y_j, Y_{j'}) = 1 - \frac{t}{n}.$$

We now verify the separated subfamily needed in the row-wise arithmetization, using the same η_0 fixed before the definition of $\Lambda_{r,\delta}$. Fix a nonempty $S \subseteq [r]$, write $s = |S|$, and choose $j_0 \in S$. The variables $\{X_j : j \in S\} \cup \{Y_{j_0}\}$ form a separated $(s+1)$-subfamily. Indeed, for distinct $j \neq j'$ in S,

$$d(X_j, X_{j'}) = \frac{1 - t/n}{2} \ge \frac{\delta}{2c_2},$$

so their covariance separation is at least $\delta/c_2 \ge \eta_0\delta$. For $j \neq j_0$, the covariance between X_j and Y_{j_0} comes only from the shared prefix, and prefix alignment gives $x^{(j)}_{\le t} = x^{(j_0)}_{\le t}$:

$$\mathrm{Cov}(X_j, Y_{j_0}) = \frac{1}{n}\left\langle x^{(j_0)}_{\le t}, y^{(j_0)}_{\le t}\right\rangle.$$

If $D_{\mathrm{pre}} = \Delta_{\mathrm{Ham}}(x^{(j_0)}_{\le t}, y^{(j_0)}_{\le t})$, then

$$d(X_j, Y_{j_0}) = \frac{1 - t/n}{2} + \frac{D_{\mathrm{pre}}}{n}.$$

The internal-overlap condition gives

$$D_{\mathrm{pre}} \ge \delta n/2 - (n - t),$$

and hence

$$d(X_j, Y_{j_0}) \ge \frac{\delta}{2} - \frac{1 - t/n}{2} > \frac{\delta}{2}\left(1 - \frac{1}{c_1}\right).$$

Thus the covariance separation between X_j and Y_{j_0} is greater than $\delta(1 - 1/c_1) \ge \eta_0\delta$, while $d(X_{j_0}, Y_{j_0}) \ge \delta/2$ gives covariance separation at least δ. Therefore every pair in this $(s+1)$-subfamily has covariance at most $1 - \eta_0\delta$.

For the fixed row under consideration, all r ensemble instances use the same thresholds for that row. Thus the collision event on this row has the same arithmetization as in Lemma 5.6:

$$\Pr[\varphi(X_j) = \varphi(Y_j)\ \forall j] = 2^{-r} \sum_{S \subseteq [r]} \mathbb{E}\left[\prod_{j \in S} \varphi(X_j)\varphi(Y_j)\right].$$

For each nonempty S, the separated $(s+1)$-subfamily exhibited above replaces the use of Lemma 5.5; applying Claim 5.4 with $\eta = \eta_0\delta$ gives the same bound for the corresponding moment term. Therefore

$$\Pr[\varphi(X_j) = \varphi(Y_j)\ \forall j] \le 2^{-\Lambda_{r,\delta}(\kappa,K)}.$$

Since the rows are independent, the probability that all r pairs collide across all $m = \alpha n$ rows is at most,

$$2^{-\alpha n \Lambda_{r,\delta}(\kappa,K)}.$$

The number of t-prefix aligned families is at most

$$2^{(2+2(r-1)(1-t/n))n}.$$

Indeed, the first pair has at most 2^{2n} choices, and each later pair is fixed on the first t coordinates and arbitrary on the remaining $n-t$ coordinates.

Finally, taking expectations and combining the bounds yields the stated exponent. The conclusion then follows by Lemma 3.3. □

Corollary 5.9 (Ensemble r-OGP)**.** *Fix $\alpha \in (0,1]$, $\kappa > 0$, $\delta \in (0,1]$, $r \geq 2$, $2 < c_1 < c_2 < 2c_1$, and an odd integer $K \geq 1$. Suppose that*

$$2 + 2(r-1)\frac{\delta}{c_1} \; - \; \alpha\, \Lambda_{r,\delta}(\kappa, K) < 0.$$

Then, for every $t \in \{0, 1, \ldots, n\}$ satisfying

$$\frac{\delta}{c_2} < 1 - \frac{t}{n} < \frac{\delta}{c_1},$$

$$\Pr\big[\mathsf{Coll}^{\mathrm{ens}}_{r,\delta}(t) \geq 1\big] \leq 2^{-\Omega(n)}.$$

Consequently, ensemble r-OGP holds at (α, δ) in the sense of Definition 5.7.

Proof. For t in the admissible window, Lemma 5.8 gives the counting term $2 + 2(r-1)(1-t/n)$, which is at most $2 + 2(r-1)\delta/c_1$. Thus, for each fixed t we have $\mathbb{E}[\mathsf{Coll}^{\mathrm{ens}}_{r,\delta}(t)] \leq 2^{-\Omega(n)}$ in the stated negative-exponent regime. By Lemma 3.3, $\Pr[\mathsf{Coll}^{\mathrm{ens}}_{r,\delta}(t) \geq 1] \leq 2^{-\Omega(n)}$. The exponent is uniform over all fixed t in the window, so this proves the stated ensemble r-OGP. □

5.2 Hardness for Online Algorithms

An online algorithm observes the columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of $\mathbf{A}$ sequentially and, after seeing the current column, must commit to the corresponding pair of input bits before seeing future columns. In the randomized activation model, the full threshold array $\mathcal{T}$ is auxiliary input available before the column stream begins.

Formally, after $\mathcal{T}$ is revealed, a randomized online algorithm $\mathcal{A}$ samples internal randomness ω at the start. At time t, it is given the threshold array, its coins, the observed column prefix, and its previous decisions, and outputs

$$(x_t, y_t) = \mathcal{A}_t(\mathcal{T}, \omega, \mathbf{a}_1, \ldots, \mathbf{a}_t, (x_1, y_1), \ldots, (x_{t-1}, y_{t-1})) \in \{-1, 1\}^2.$$

Equivalently, the algorithm may maintain an arbitrary internal state updated from the observed prefix, but it cannot inspect columns $\mathbf{a}_{t+1}, \ldots, \mathbf{a}_n$ before committing to (x_t, y_t).

For $t \in \{0, 1, \ldots, n\}$, let $\mathsf{A}^{(t)} = (\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(r)})$ be an ensemble whose first t columns coincide and whose remaining $n-t$ columns are independent, with the same thresholds shared across all $\mathbf{A}^{(j)}$.

Theorem 5.10. *Fix $2 < c_1 < c_2 < 2c_1$. Assume that ensemble r-OGP holds at parameters (α, δ) with this forbidden-overlap window. Then, over the random matrix, thresholds, and internal coins, every online algorithm $\mathcal{A}$ satisfies*

$$\Pr\left[\begin{array}{c}\mathcal{A} \text{ outputs } \mathbf{x}, \mathbf{y} \in \{\pm 1\}^n \text{ such that } \varphi(\mathbf{A}\mathbf{x}) = \varphi(\mathbf{A}\mathbf{y}) \\ \text{and } |\langle \mathbf{x}, \mathbf{y}\rangle| \le (1-\delta)n\end{array}\right] \le 2^{-\Omega(n/r)}.$$

Thus any constant choice of r for which ensemble r-OGP holds gives an exponentially small online success probability.

Proof. Let $\xi = \frac{1}{3}\delta(1/c_1 - 1/c_2)$. The interval below has positive constant width, so for all sufficiently large n we may choose $t \in \{0, \dots, n\}$ with constant slack in the ensemble-OGP window:

$$\frac{\delta}{c_2} + \xi < 1 - \frac{t}{n} < \frac{\delta}{c_1} - \xi.$$

Since δ, c_1, c_2 are fixed constants, this choice has $n - t = \Theta(n)$. Let p denote the single-instance success probability of $\mathcal{A}$ for producing a δ-extensive collision (unrelated to the anti-concentration parameter of the same name in Theorem 5.1). The coupling is designed so that online-ness forces the r outputs to agree on the shared prefix, while independent sign flips on the unseen suffix randomize the cross-overlaps without changing whether each output pair is a collision. We use the following symmetrized coupled experiment. Sample the correlated ensemble $\mathsf{A}^{(t)} = (\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(r)})$. For each $j \in [r]$ and each $i > t$, draw an independent sign $\sigma_i^{(j)} \in \{\pm 1\}$ and let $\widetilde{\mathbf{A}}^{(j)}$ be obtained from $\mathbf{A}^{(j)}$ by replacing the suffix column $\mathbf{a}_i^{(j)}$ with $\sigma_i^{(j)}\mathbf{a}_i^{(j)}$. Run $\mathcal{A}$ on the matrices $\widetilde{\mathbf{A}}^{(1)}, \dots, \widetilde{\mathbf{A}}^{(r)}$, using the same internal coins and the same thresholds across all instances. Since the signed matrices have the same marginal law as the original matrices, each run succeeds with probability p.

Condition on the common prefix, the thresholds, and the internal coins. Given this information, the signed suffixes are independent across the r instances and have identical conditional laws. Since the algorithm is online and the thresholds and coins are already fixed, the r success events are conditionally independent and have the same conditional success probability, say q. Hence the probability that all r runs succeed in the symmetrized coupled experiment is $\mathbb{E}[q^r] \ge (\mathbb{E}[q])^r = p^r$, by Jensen's inequality since $x \mapsto x^r$ is convex for $r \ge 2$.

Let $(\mathbf{u}^{(j)}, \mathbf{v}^{(j)})$ be the raw output of $\mathcal{A}$ on $\widetilde{\mathbf{A}}^{(j)}$. Since the first t columns, the thresholds, and the internal coins are identical across the coupled runs, online-ness implies that on every outcome the raw prefixes

$$\mathbf{u}_{\le t}^{(1)} = \cdots = \mathbf{u}_{\le t}^{(r)}, \qquad \mathbf{v}_{\le t}^{(1)} = \cdots = \mathbf{v}_{\le t}^{(r)}$$

are identical across all $j \in [r]$.

For $i > t$, set

$$x_i^{(j)} = \sigma_i^{(j)} u_i^{(j)}, \qquad y_i^{(j)} = \sigma_i^{(j)} v_i^{(j)}.$$

On the prefix, set $x_i^{(j)} = u_i^{(j)}$ and $y_i^{(j)} = v_i^{(j)}$. Then

$$\mathbf{A}^{(j)}\mathbf{x}^{(j)} = \widetilde{\mathbf{A}}^{(j)}\mathbf{u}^{(j)}, \qquad \mathbf{A}^{(j)}\mathbf{y}^{(j)} = \widetilde{\mathbf{A}}^{(j)}\mathbf{v}^{(j)},$$

where $\widetilde{\mathbf{A}}^{(j)}$ is the signed matrix seen by the algorithm. Thus this post-processing preserves collision equations and internal overlaps: since $(\sigma_i^{(j)})^2 = 1$ for all i, the inner product $\langle \mathbf{x}^{(j)}, \mathbf{y}^{(j)}\rangle = \sum_i (\sigma_i^{(j)})^2 u_i^{(j)} v_i^{(j)} = \langle \mathbf{u}^{(j)}, \mathbf{v}^{(j)}\rangle$ is unchanged.

By Gaussian sign symmetry, the signs $\sigma_i^{(j)}$ are independent of the signed matrices $\widetilde{\mathbf{A}}^{(j)}$. The raw outputs are functions of the signed matrices, the thresholds, and the fixed coins, so after conditioning on the signed matrices and raw outputs the suffix flip signs remain independent. For every distinct j, j', the suffix products

$$x_i^{(j)} x_i^{(j')} = \sigma_i^{(j)} \sigma_i^{(j')} u_i^{(j)} u_i^{(j')}$$

are independent mean-zero signs over $i > t$, and the same holds for the y-vectors. Therefore, by Lemma 3.5 with $\lambda = \xi n/(4(n-t))$ and a union bound over pairs $j \neq j'$,

$$\left| \frac{1}{n} \langle \mathbf{x}^{(j)}, \mathbf{x}^{(j')} \rangle - \frac{t}{n} \right| \leq \xi/4, \qquad \left| \frac{1}{n} \langle \mathbf{y}^{(j)}, \mathbf{y}^{(j')} \rangle - \frac{t}{n} \right| \leq \xi/4$$

for all distinct j, j' except with probability $2^{-\Omega(n)}$. Since $1 - t/n$ has slack ξ in the admissible window, these inequalities imply

$$1 - \frac{\delta}{c_1} \leq \frac{\langle \mathbf{x}^{(j)}, \mathbf{x}^{(j')} \rangle + \langle \mathbf{y}^{(j)}, \mathbf{y}^{(j')} \rangle}{2n} \leq 1 - \frac{\delta}{c_2}$$

for all distinct j, j'.

On the event that all r runs succeed and the above concentration event holds, the outputs form a t-prefix aligned, ensemble δ-admissible family of collisions. Hence

$$p^r \leq \Pr[\mathsf{Coll}_{r,\delta}^{\mathrm{ens}}(t) \geq 1] + 2^{-\Omega(n)} \leq 2^{-\Omega(n)},$$

where the last inequality is ensemble r-OGP. Therefore $p \leq 2^{-\Omega(n/r)}$. □

Corollary 5.11 (Online Collision Resistance)**.** *Fix $\alpha \in (0,1)$, $\kappa > 0$, $r \geq 2$, $\delta \in (0,1]$, $2 < c_1 < c_2 < 2c_1$, and an odd integer $K \geq 1$. If*

$$2 + 2(r-1)\frac{\delta}{c_1} - \alpha \Lambda_{r,\delta}(\kappa, K) < 0,$$

then every online algorithm has success probability at most $2^{-\Omega(n/r)}$ for outputting a δ-extensive collision. In particular, any constant choice of r satisfying the condition gives online δ-extensive collision resistance at density α. (The ensemble condition uses δ/c_1 in place of the entropy $H(\delta/(2c_1))$ appearing in Corollary 5.2; this is because prefix alignment restricts the external-overlap count to the suffix, contributing $(r-1)(1-t/n) \leq (r-1)\delta/c_1$ linearly rather than entropically.)

Proof. The condition implies ensemble r-OGP by Corollary 5.9. The conclusion follows from Theorem 5.10. □

References

[1] Baldassi, C., Ingrosso, A., Lucibello, C., Saglietti, L., Zecchina, R.: Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses. Physical Review Letters **115**, 128101 (2015). https://doi.org/10.1103/PhysRevLett.115.128101

[2] Bansal, N., Spencer, J.H.: On-Line Balancing of Random Inputs. Random Structures & Algorithms **57**(4), 879–891 (2020). https://doi.org/10.1002/rsa.20955

[3] Benedetti, M., Bogdanov, A., Malatesta, E., Mézard, M., Perrupato, G., Rosen, A., Schwartzbach, N.I., Zecchina, R.: Are Neural Networks Collision Resistant? (2025), `https://eprint.iacr.org/2025/1756`, cryptology ePrint Archive, Paper 2025/1756

[4] Benedetti, M., Bogdanov, A., Malatesta, E., Mézard, M., Perrupato, G., Rosen, A., Schwartzbach, N.I., Zecchina, R.: Overlap Gap and Computational Thresholds in the Square Wave Perceptron. Journal of Statistical Mechanics: Theory and Experiment **2025**(12), 123303 (2025). https://doi.org/10.1088/1742-5468/ae23be

[5] Dubhashi, D.P., Panconesi, A.: Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge University Press (2009). https://doi.org/10.1017/CBO9780511581274

[6] Gamarnik, D.: The Overlap Gap Property: A Topological Barrier to Optimizing over Random Structures. Proceedings of the National Academy of Sciences **118**(41), e2108492118 (2021). https://doi.org/10.1073/pnas.2108492118

[7] Gamarnik, D., Jagannath, A.: The Overlap Gap Property and Approximate Message Passing Algorithms for p-Spin Models. The Annals of Probability **49**(1), 180–205 (2021). https://doi.org/10.1214/20-AOP1448

[8] Gamarnik, D., Jagannath, A., Wein, A.: Low-Degree Hardness of Random Optimization Problems. In: IEEE Symposium on Foundations of Computer Science (FOCS). pp. 131–140. IEEE (2020). https://doi.org/10.1109/FOCS46700.2020.00021

[9] Gamarnik, D., Kızıldağ, E.C., Perkins, W., Xu, C.: Algorithms and Barriers in the Symmetric Binary Perceptron Model. In: IEEE Symposium on Foundations of Computer Science (FOCS). pp. 576–587. IEEE (2022). https://doi.org/10.1109/FOCS54457.2022.00061

[10] Gamarnik, D., Sudan, M.: Limits of Local Algorithms over Sparse Random Graphs. The Annals of Probability **45**(4), 2353–2376 (2017). https://doi.org/10.1214/16-AOP1114

[11] Gardner, E.: The Space of Interactions in Neural Network Models. Journal of Physics A: Mathematical and General **21**(1), 257–270 (1988). https://doi.org/10.1088/0305-4470/21/1/030

[12] Gardner, E., Derrida, B.: Three Unfinished Works on the Optimal Storage Capacity of Networks. Journal of Physics A: Mathematical and General **22**(12), 1983–1994 (1989). https://doi.org/10.1088/0305-4470/22/12/004

[13] Kim, J.H., Roche, J.R.: On the Optimal Capacity of Binary Neural Networks: Rigorous Combinatorial Approaches. In: Proceedings of the Conference on Computational Learning Theory (COLT). pp. 240–249. ACM (1995). https://doi.org/10.1145/225298.225327

[14] Li, K., Zhang, T., Malik, J.: Approximate Feature Collisions in Neural Nets. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 15816–15824 (2019)

[15] Li, S., Schramm, T., Zhou, K.: Discrepancy algorithms for the binary perceptron. In: Proceedings of the 57th Annual ACM Symposium on Theory of Computing. pp. 1668–1679. STOC '25 (2025). https://doi.org/10.1145/3717823.3718124

[16] McCulloch, W.S., Pitts, W.: A Logical Calculus of the Ideas Immanent in Nervous Activity. The Bulletin of Mathematical Biophysics **5**(4), 115–133 (1943). https://doi.org/10.1007/BF02478259

[17] Mézard, M., Montanari, A.: Information, Physics, and Computation. Oxford University Press (2009). https://doi.org/10.1093/acprof:oso/9780198570837.001.0001

[18] Ozbulak, U., Rao, S., De Neve, W., Vankerschaver, J., Van Messem, A., Gasparyan, M.: Exact Feature Collisions in Neural Networks. Scientific Reports **16**, 10139 (2026). https://doi.org/10.1038/s41598-026-40605-4

[19] Rosenblatt, F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review **65**(6), 386–408 (1958). https://doi.org/10.1037/h0042519

[20] Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan, Washington, DC (1962)

[21] Subag, E.: The Geometry of the Gibbs Measure of Pure Spherical Spin Glasses. Inventiones Mathematicae **210**(1), 135–209 (2017). https://doi.org/10.1007/s00222-017-0726-4

[22] Vafa, N., Vaikuntanathan, V.: Symmetric Perceptrons, Number Partitioning and Lattices. In: Proceedings of the ACM Symposium on Theory of Computing (STOC). pp. 2191–2202. ACM (2025). https://doi.org/10.1145/3717823.3718263

[23] Vershynin, R.: High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press (2018). https://doi.org/10.1017/9781108231596