Frozen 1-RSB structure of the symmetric Ising perceptron

Will Perkins¹  |  Changji Xu²

¹School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia, USA
²Center of Mathematical Sciences and Applications, Harvard University, Cambridge, Massachusetts, USA

Correspondence
Will Perkins, School of Computer Science, Georgia Institute of Technology, Atlanta, GA, USA.
Email: math@willperkins.org

Abstract
We prove, under an assumption on the critical points of a real-valued function, that the symmetric Ising perceptron exhibits the ‘frozen 1-RSB’ structure conjectured by Krauth and Mézard in the physics literature; that is, typical solutions of the model lie in clusters of vanishing entropy density. Moreover, we prove this in a very strong form conjectured by Huang, Wong, and Kabashima: a typical solution of the model is isolated with high probability and the Hamming distance to all other solutions is linear in the dimension. The frozen 1-RSB scenario is part of a recent and intriguing explanation of the performance of learning algorithms by Baldassi, Ingrosso, Lucibello, Saglietti, and Zecchina. We prove this structural result by comparing the symmetric Ising perceptron model to a planted model and proving a comparison result between the two models. Our main technical tool towards this comparison is an inductive argument for the concentration of the logarithm of number of solutions in the model.

KEYWORDS
frozen solutions, learning algorithms, perceptron, planted model, solution space

1 | INTRODUCTION

The perceptron model is a simple model of a neural network storing random patterns. It has been studied in several fields including information theory [19], statistical physics [24, 25, 30, 40], and probability theory [32, 41, 42].
There are several variants of the model, grouped into two main categories: spherical perceptrons in which patterns are $N$-dimensional vectors on the unit sphere and Ising perceptrons in which patterns are $\pm 1$ vectors of length $N$. In each case, we want to understand how many random patterns can be ‘stored’ by a neural network formed by taking random synapses $J_{ij}$ and applying a given activation function. Here we will study the Ising perceptron.

Let $\Sigma_N = \{\pm 1\}^N$ and let $\{X_i\}_{i \geq 1}$ be a sequence of independent $N$-dimensional standard Gaussian vectors. For a real valued function $\phi$, a real number $\kappa$, and $X \in \mathbb{R}^N$, define

$$H_{\phi,\kappa}(X) = \{\sigma \in \Sigma_N : \phi(\langle X, \sigma \rangle / \sqrt{N}) \leq \kappa\}.$$ 

The solution space of the Ising perceptron with Gaussian disorder, activation function $\phi$, threshold $\kappa$, and $m$ constraints is the random subset of $\Sigma_N$,

$$S = S_{\phi,\kappa,N,m} = \bigcap_{i=1}^m H_{\phi,\kappa}(X_i).$$

Thus $S$ is a random subset of the Hamming cube $\Sigma_N$. We call the vectors $X_i$ constraint vectors and elements in $S$ solutions. The constraints depend on the set of constraint vectors, the activation function $\phi$, and the threshold $\kappa$. The classic Ising perceptron corresponds to the choice $\phi(x) = x$ where the most studied case is $\kappa = 0$ (e.g., [21, 30]).

In another variant of the model the constraint vectors $X_i$ are given by independent samples from $\Sigma_N$ (Bernoulli disorder). While there are significant differences between the spherical and Ising perceptrons, the choice of Gaussian or Bernoulli disorder is insignificant for the properties discussed here.

1.1 Structure of the solution space

We will be concerned with the typical structure of the solution space $S$ as a function of $\phi$, $\kappa$, and the constraint density $\alpha := m/N$, in the regime where $\phi$, $\kappa$, and $\alpha$ are fixed as $N \to \infty$.

The most basic structural question is whether $S$ is empty or not. The capacity of the perceptron is defined as the random variable

$$M_{\phi,\kappa}(N) = \max\{m : S_{\phi,\kappa,N,m} \neq \emptyset\},$$ 

and the critical capacity density

$$\alpha_c,\phi(\kappa) = \inf \left\{ \alpha : \liminf_{N \to \infty} \mathbb{P}(S_{\phi,\kappa,N,\lfloor\alpha N\rfloor} = \emptyset) = 1 \right\}$$

is the typical constraint density of the capacity.

For densities below the critical capacity density, when the solution space is typically non-empty, we can ask about its structure, and how this structure affects the performance of learning algorithms: algorithms that find some solution in $S$ given the instance defined by $X = (X_1, \ldots, X_m)$.

Basic structural questions include whether solutions appear in connected clusters or are isolated; and what the typical distance is from a solution to the next nearest solution. For these structural properties we regard $\Sigma_N$ as the Hamming cube endowed with Hamming distance: for $\sigma, \sigma' \in \Sigma_N$,

$$\text{dist}(\sigma, \sigma') = |\{i : \sigma_i \neq \sigma'_i\}| = \frac{N - \langle \sigma, \sigma' \rangle}{2}. $$
The perspective taken in recent work on Ising perceptrons in both statistical physics and computer science is to view the Ising perceptron as a random constraint satisfaction problem (CSP): \( \Sigma_N \) is the set of possible solutions, and each random vector \( X_i \) defines a constraint \( \phi \left( \langle X, \sigma \rangle / \sqrt{N} \right) \leq \kappa \) on possible solutions \( \sigma \in \Sigma_N \). The critical capacity density \( \alpha_{c, \phi} \) is then the satisfiability threshold of the model.

Just as in the random \( k \)-SAT, random \( k \)-NAE-SAT, or random \( k \)-XOR-SAT problems, each constraint rules out a constant fraction of all solutions in \( \Sigma_N \). Where the perceptron differs from these other models is that in the perceptron each constraint involves all of the \( N \) coordinates, while in the other models each constraint only involves \( k \) coordinates, with \( k \) held constant as \( N \to \infty \).

Random CSP’s have been studied extensively in computer science, statistical physics, probability, and combinatorics since Mitchell, Selman, and Levesque [37] observed empirically that random \( k \)-SAT formulae at certain densities proved extremely challenging for widely used SAT solvers. Understanding this phenomenon is a major ongoing challenge that has led to the development of a new field of inquiry at the intersection of computer science, physics, and mathematics.

A key to the current understanding of random CSP’s is understanding the typical structure of the solution space at different constraint densities. A beautifully detailed but non-rigorous picture was put forth by Krzakala, Montanari, Ricci-Tersenghi, Semerjian, and Zdeborová [31], based on the cavity method from statistical physics.

In a paper that transformed the computer science perspective on random computational problems, Achlioptas and Coja-Oghlan [2] proved the existence of a ‘clustering threshold’ at which the solution space of certain random CSP’s breaks apart into exponentially many clusters separated by linear Hamming distance. They also observed that this clustering threshold coincides asymptotically with the threshold at which known efficient search algorithms for these problems fail.

A further structural property is freezing. Given a solution \( \sigma \), a variable (or coordinate) is free if flipping that coordinate results in another solution \( \sigma' \). Variables that are not free are frozen. The freezing threshold for a random CSP is the threshold at which typical solutions have a linear number of frozen variables [34, 46].

These two properties—clustering and freezing—are conjectured to be the source of computational hardness in random CSP’s. Identifying thresholds for the onset of these and other structural phenomena and rigorously connecting them to the performance of algorithms have been a main focus of the field of random computational problems in the last decade.

For the Ising perceptron, the conjectured structural picture looks strikingly different. Krauth and Mezard [30] conjectured, by means of the replica method, that at all densities below the critical capacity, the solution space of the perceptron is dominated by clusters of vanishing entropy density (i.e., of size \( e^{o(n)} \)), each cluster separated from the others by linear Hamming distance. Wong, Kabashima, and Huang [29] and Huang and Kabashima [27] refined these conjectures and posited that in fact typical solutions in the Ising perceptron are completely frozen; that is, all coordinates are frozen and the solutions lie in clusters of size 1, separated from all other solutions by linear Hamming distance. Thus the Ising perceptron exhibits clustering and freezing in the strongest possible form throughout the entire satisfiability regime. Based on the current conjectural understanding of random CSP’s (see e.g., [48]), one might venture a guess that finding a solution in the Ising perceptron is computationally hard at all positive densities.

However, this theory is at odds with other work in physics and mathematics on learning algorithms for the perceptron. Kim and Roche gave a polynomial time algorithm to find solutions with high probability at very small but positive constraint densities [32]. Braunstein and Zecchina [16] also observed empirically that a simple message-passing algorithm is able to find solutions at positive densities in the Ising perceptron (further algorithms followed in [7, 8]). Attempting to reconcile this apparent contradiction, Baldassi, Ingrosso, Lucibello, Saglietti, and Zecchina [13] conjectured that these successful
learning algorithms were in fact finding solutions belonging to rare clusters of positive entropy density. That is, although a $1 - o(1)$ fraction of solutions belong to isolated, frozen clusters, an exponentially small fractions of solutions belong to clusters that are exponentially large; strikingly, the authors observed that learning algorithms find solutions in these rare clusters. Specifically, the solutions that contribute to the dominant portion of the partition function (number of solutions) and determine the equilibrium properties of the model are completely distinct from those solutions that efficient algorithms find. This work was followed by the proposal of several different algorithms to target these subdominant clusters [9, 14].

In [6], a symmetric Ising perceptron, with activation function $\phi(x) = |x|$, was studied as a model conjectured to exhibit the same structural and algorithmic properties, but more amenable to mathematical analysis. Baldassi, Della Vecchia, Lucibello, and Zecchina [12] confirmed that on the level of the physics predictions this model has the same qualitative behavior as the Ising perceptron with activation function $\phi(x) = x$.

In summary, the Ising perceptron, and its symmetric variants, are conjectured to exhibit ‘frozen 1-RSB behavior’ at all positive densities below the critical capacity density. The current understanding of the link between clustering, freezing, and the performance of algorithms would suggest that finding a solution in these models is therefore intractable. This, however, is seemingly in contradiction with theoretical results and empirical observations, and one hypothesis suggests that learning algorithms find exponentially rare solutions with atypical structural properties.

Resolving these questions is a pressing problem since the hypothesis about subdominant clusters calls into question the link between the equilibrium properties of these models and algorithmic tractability. In this work, we take a first step in addressing this problem rigorously by establishing the frozen 1-RSB picture for the symmetric Ising perceptron (Theorem 1 below), under an assumption on the critical points of a real-valued function (Assumption 1).

1.2 Previous results

There are few rigorous results on the Ising perceptron, and most are concerned with bounds on the critical capacity.

For the classic Ising perceptron Krauth and Mézard [30] predicted, using the replica method, that $\alpha_c(\kappa) = \alpha_{KM}(\kappa)$ for a complicated but explicit function $\alpha_{KM}$ (with $\alpha_{KM}(0) \approx .83$). Following some previous bounds of Kim and Roche and Talagrand [32, 41], Ding and Sun [21] recently proved that $\alpha_c(\kappa) \geq \alpha_{KM}(\kappa)$ using a sophisticated form of the second-moment method guided by the Thouless–Anderson–Palmer (TAP) equations [44]. Their result assumes a technical condition on a certain real-valued function, akin to Assumption 1 below.

Much of the technical difficulty of [21] comes from the asymmetry inherent in the activation function $\phi(x) = x$; this necessitates a conditioning argument and the sophisticated second-moment calculation. On the other hand, Aubin, Perkins, and Zdeborová [6] considered two symmetric activation functions: $\phi_r(x) = |x|$ and $\phi_u(x) = -|x|$, which they called the rectangular and ‘u’ activation functions respectively. Studying symmetric constraints has a long history in the random CSP literature: the random $k$-NAE-SAT model is a symmetric variant of the random $k$-SAT model. While the qualitative properties of the two models are expected to be very similar, the symmetric model is often more amenable to rigorous analysis, and thus a clearer understanding can be obtained (see e.g., [15, 22, 39] for recent work on the $k$-NAE-SAT model). Studying symmetric perceptrons allows us to prove stronger and more detailed results than are currently attainable for the classic perceptron, but the phenomena studied are expected to be universal.
Aubin, Perkins, and Zdeborová [6] determine the critical capacity density for the symmetric perceptron with rectangular activation function:

\[ \alpha_{r,c}(\kappa) = -\frac{\log 2}{\log p(\kappa)} \]

where \( p(\kappa) = \mathbb{P}(|Z| \leq \kappa) \) for a standard Gaussian random variable \( Z \). This result, like that of [21], is contingent on an assumption about a certain real-valued function. This function will also prove useful in our work. Let \( H(\beta) = -\beta \log \beta - (1 - \beta) \log(1 - \beta) \) be the Shannon entropy function (all logarithms in this paper are base \( e \)) and let

\[ q_{\kappa}(\beta) = \mathbb{P}(|Z_1| \leq \kappa, |Z_2| \leq \kappa) \]

where \((Z_1, Z_2)\) is a jointly Gaussian vector with means 0, variances 1, and covariance \( 2\beta - 1 \).

**Assumption 1** ([6]). The function

\[ F_{a}(\beta) = H(\beta) + a \log q_{\kappa}(\beta). \]

has a single critical point for \( \beta \in (0, 1/2) \) whenever \( F''_{a}(1/2) < 0 \).

**Remark 1.** Since the appearance of this paper as a preprint, this assumption has been proved rigorously by Abbe, Li, and Sly in [4].

Xu [45] proved a general sharp threshold result (an analogue of Friedgut’s sharp threshold result for random graphs and CSP’s [23]), which, combined with [6], gives a sharp threshold for the existence of solutions: for any \( \epsilon > 0 \),

\[
\mathbb{P}(S_{\phi, \kappa, N}(a_{r,c} + \epsilon N) = \emptyset) \to 1 \text{ as } N \to \infty, \\
\mathbb{P}(S_{\phi, \kappa, N}(a_{r,c} - \epsilon N) = \emptyset) \to 0 \text{ as } N \to \infty.
\]

Both statements hold also for the \( u \)-function in a range of \( \kappa \) values, and the results of [21, 45] prove that the second statement holds for the classic perceptron. Proving the matching upper bound on the critical capacity for the classic perceptron remains a challenging open problem.

Baldassi, Della Vecchia, Lucibello, and Zecchina [12] used the second-moment method to show the existence of pairs of solutions at arbitrary distances in the symmetric Ising perceptron.

### 1.3 Main results

We will study properties of typical solutions in the symmetric Ising perceptron (with the rectangular activation function \( \phi(x) = |x| \)). We now specialize and simplify the notation from above.

For \( X \in \mathbb{R}^N \) and \( \kappa > 0 \), define

\[ H_{\kappa}(X) := \{ \sigma \in \Sigma_N : |\langle X, \sigma \rangle| \leq \kappa \sqrt{N} \}. \]

Let \( \{X_i\}_{i \geq 1} \) be a sequence of i.i.d. \( N \)-dimensional standard Gaussians, and define the solution space

\[ S = S_{a}(N) = \bigcap_{i=1}^{\lfloor aN \rfloor} H_{\kappa}(X_i). \]

The critical capacity density, determined in [6], is \( \alpha_{c} = \alpha_{c}(\kappa) = -\log 2 / \log p(\kappa) \).
The main result of this paper confirms the frozen 1-RSB scenario in the symmetric Ising perceptron: typical solutions are completely frozen with high probability for \( \alpha < \alpha_c \). Let

\[
\beta_c = \beta_c(\kappa, \alpha) = \beta \in (0, 1/2) : F_\alpha(\beta) - \alpha \log p(\kappa) = 0.
\]  

(3)

See Figure 1 for a depiction of \( \beta_c \). We show in Lemma 6 that for \( \alpha < \alpha_c \), \( \beta_c > 0 \) exists and is uniquely defined.

**Theorem 1.** Let \( \kappa > 0 \) and \( \alpha < \alpha_c(\kappa) \). Let \( \sigma \) be uniformly sampled from \( S \) conditioned on the event \( S \neq \emptyset \). Under Assumption 1, for any \( \delta \in (0, \beta_c) \),

\[
\{ \sigma' \in S : \text{dist}(\sigma, \sigma') \leq (\beta_c - \delta)N \} = \{ \sigma \}.
\]

with probability \( 1 - o(1) \) as \( N \to \infty \). In particular, \( \sigma \) is completely frozen with probability \( 1 - o(1) \).

Note that \( \sigma \) is selected according to two sources of randomness: the randomness of the perceptron instance \( X \) and the random choice of \( \sigma \) drawn uniformly from \( S \). With probability \( 1 - o(1) \) over both sources of randomness the solution \( \sigma \) is completely frozen.

The next result shows that the logarithm of the number of number of solutions in the rectangular Ising perceptron is tightly concentrated below the critical density.

**Theorem 2.** Under Assumption 1, for \( \alpha < \alpha_c \)

\[
\frac{\log |S|}{N} = \log 2 + \alpha \log p(\kappa) + O_p\left(\frac{\log N}{N}\right) \quad \text{as } N \to \infty \text{ in probability.}
\]

In particular, for \( \alpha < \alpha_c \), \( S \) is non-empty with probability \( 1 - o(1) \).
The second statement of the Theorem 2 proves that the symmetric Ising perceptron undergoes a sharp satisfiability phase transition at $\alpha_c$, answering an open question from [6] (where the complementary statement that for $\alpha > \alpha_c$, $S = \emptyset$ with high probability is proved). This could also be proved by adapting the sharp threshold result of [45] to the symmetric perceptron.

The following result is a major ingredient in proving Theorem 2, which gives a quantitative description of the fact that two independent random samples from the solution space are almost orthogonal below the critical density.

**Lemma 3.** Let $\kappa > 0$ and $\alpha < \alpha_c(\kappa)$. Let $\sigma_1, \sigma_2$ be uniformly and independently sampled from $S$ conditioned on the event $S \neq \emptyset$. Under Assumption 1,

$$\frac{\langle \sigma_1, \sigma_2 \rangle}{N} = O_p\left(\sqrt{\frac{\log N}{N}}\right)$$

as $N \to \infty$ in probability.

### 1.4 Overview of the techniques

We study the properties of a typical solution drawn from $S$ by way of the *planted model*: the experiment of first selecting a uniformly random solution from $\Sigma_N$, then choosing a random configuration of constraints consistent with this solution. Planted models have been studied extensively in the random CSP literature and beyond. They are used as a toy model for statistical inference: for example, the ‘teacher–student model’ [47] or the stochastic block model [1]. They are used to understand the condensation threshold in random CSP’s [11, 17, 18, 20]. They are used to understand the structure of the solution space in random CSP’s [2, 3, 34, 35].

Given $N, m \in \mathbb{N}$, $\kappa > 0$, and following [2], we define two probability distributions on pairs $(\sigma^*, X) \in \Sigma_N \times (\mathbb{R}^N)^m$ of solutions and configurations of $m$ constraint vectors.

In the *random model* we:

1. Sample $m$ i.i.d. $N$-dimensional standard Gaussian constraint vectors $X = (X_1, \ldots, X_m)$, conditioned on the event that $S(X) = \bigcap_{i=1}^m H_\kappa(X_i) \neq \emptyset$.
2. Sample $\sigma^*$ uniformly at random from $S$.

We denote the law of the random model with $\mathbb{P}_r, \mathbb{E}_r$ to distinguish the law from both the unconditional perceptron model and the planted model below. The random model is simply the experiment of selecting a uniformly random solution from the symmetric Ising perceptron conditioned on satisfiability.

In the *planted model* we:

1. Sample $\sigma^*$ uniformly at random from $\Sigma_N$.
2. Sample a configuration of $m$ i.i.d. constraint vectors $X = (X_1, \ldots, X_m)$, with each $X_i$ distributed as a standard $N$-dimensional Gaussian vector conditioned on the event that $\sigma^* \in H_\kappa(X_i)$.

We denote the law of the planted model with $\mathbb{P}_{pl}, \mathbb{E}_{pl}$.

The key to using the planted model to understand the original model is to show that at low enough constraint densities, the two distributions on $(\sigma^*, X)$ are close. Proving that the distributions are close, as we do below in Lemma 14, amounts to proving that the number of solutions, $|S|$, is typically not too far from its expectation, $\mathbb{E}|S|$. The better concentration of $|S|$ one can prove, the more one can deduce about the original model from the planted model. In [2] it is shown (in the case of $q$-colorings of a random graph) that if $\log |S| = \log \mathbb{E}|S| + o(N)$, then events that occur with probability at most $\exp(-\theta(N))$ in the planted model occur with probability $o(1)$ as $N \to \infty$ in the random model. This
notion of closeness is ‘quiet planting’ [33] and it suffices to prove some structural results on the solution spaces such as clustering [2]. On the other hand, much stronger notions of closeness have been proved: ‘silent planting’ [10] which implies the two distributions are mutually contiguous: any event with probability $o(1)$ in the planted model has probability $o(1)$ in the random model. This has been used to prove stronger structural results [34]. Proving contiguity requires much stronger concentration of $\log|S|$. A very general result on the contiguity of the planted and random model for symmetric CSP’s [17] involves a rigorous implementation of the cavity method and the small subgraph conditioning method. In the setting of the perceptron, neither of these tools exist and so we must prove concentration via another route.

We prove Theorem 1 in three steps.

In Section 2, we prove that the planted solution is isolated and the next nearest solution is at linear Hamming distance with high probability in the planted model (Lemma 7).

In Section 3 we prove Theorem 2, showing that for $\alpha < \alpha_c$, the logarithm of the number of solutions in the random model is concentrated around the logarithm of the expected number of solutions. We use an inductive martingale argument to show this. In [43], Talagrand also employed a martingale approach to prove concentration for $\log|S|$, which applies to general $\phi$ but only for small constraint density and has less precision. The main difference with our approach is that we are able to inductively provide and then extract more precise information on the overlap of two random samples from solution space.

In Section 4 we transfer our results about the planted model to the random model by showing that events that occur with probability at most $N^{-o(1)}$ in the planted model occur with probability $o(1)$ in the standard model (Lemma 14). This relies on the concentration properties of the logarithm of the number of solutions.

1.5 Extensions and future work

Both Theorems 1 and 2 can be extended verbatim to the $u$-function Ising perceptron studied in [6], for $\kappa \in (0, .817)$ (the same range of $\kappa$ for which the second-moment method works there).

The main open problem in this area is to resolve the conceptual dilemma described in Section 1.1 and answer the questions raised in [13]. Are there efficient learning algorithms that always find out-of-equilibrium solutions in subdominant clusters? This would raise a serious challenge to the belief that an understanding of the associated equilibrium statistical mechanics model (on the level of the free energy) can explain computational tractability or intractability.

Concretely, now that we have verified the frozen 1-RSB scenario, we can ask for provably efficient learning algorithms for the symmetric Ising perceptron.

Question 4. Is there a polynomial-time algorithm that, with probability $1 - o(1)$, finds a solution to the symmetric Ising perceptron for some density $\alpha \in (0, \alpha_c)$?

One likely candidate would be adapting the algorithm of Kim and Roche [32] from the perceptron to the symmetric perceptron; this algorithm, however, only finds solutions at very small densities.

Alternatively, one could leverage the structural results we have proved here and rule out some class of learning algorithms. For instance, what can one say about the performance of stable algorithms (discussed in, e.g., [26])?

We leave the following additional open problems for future work.

1. Prove that the classic perceptron with activation function $\phi(x) = x$ exhibits the frozen 1-RSB property. It is not at all clear how to extend the method of this paper to this case. As discussed in [Sec. 2.4] [17], for asymmetric random CSP’s, like the random $k$-SAT model, the planted
model, at least in its straightforward implementation, is not useful to compare to the random
model (in particular it is not contiguous with the random model at any positive density). We
expect the same with the classic perceptron, and so our strategy of arguing via the planted model
will not work.

2. Prove full contiguity between the random and planted models for \( \alpha < \alpha_c \). Our comparison result
(Lemma 14) suffices for our purposes here, but it is natural to ask for more (as is the case for a
large class of symmetric random CSP’s [17]).

**Conjecture 5.** For \( \alpha < \alpha_c \), the random and planted models of the symmetric Ising per-
ceptron are mutually contiguous. That is, if \( P_{pl}(A) = o(1) \) then \( P_r(A) = o(1) \) and vice
versa.

**Remark 2.** In [4], Abbe, Li, and Sly have proved Conjecture 5 by adapting the small
subgraph conditioning method to the symmetric perceptron, thus improving the concen-
tration results of Theorem 2. Likewise, Abbe, Li, and Sly have answered Question 4 in the
affirmative in [5].

2 | THE PLANTED MODEL

Consider the planted model with planted solution \( \sigma^* \) and constraints \( X = (X_1, X_2, \ldots, X_m) \). Define
\( S(X) = \bigcap_{i=1}^{m} H_\alpha(X_i) \) as in (2). Recall the definition of \( \beta_c \) from (3). We show that \( \beta_c \) exists and is unique
for \( \alpha \in (0, \alpha_c) \).

**Lemma 6.** Under Assumption 1, for \( \alpha \in (0, \alpha_c) \), there exists a unique \( \beta \in (0, 1/2) \) so that

\[ F_\alpha(\beta) - \alpha \log p(\kappa) = 0. \]

Also, for any \( \delta \in (0, \beta_c/2) \),

\[ \sup_{\delta < \beta < \beta_c - \delta} F_\alpha(\beta) - \alpha \log p(\kappa) < 0. \]  \hspace{1cm} (4)

**Proof.** Let \( G(\beta) = F_\alpha(\beta) - \alpha \log p(\kappa) \). Then since \( q_\kappa(0) = p(\kappa) \) and \( H(0) = 0 \), we have
\( G(0) = 0 \). As observed in [6], \( G'(x) = -\infty \) and so for some \( \varepsilon > 0 \), \( G(x) < 0 \) for \( x \in (0, \varepsilon) \).
Moreover, since \( \alpha < \alpha_c \), \( G(1/2) > 0 \), and so by continuity there exists \( \beta \in (0, 1/2) \) with
\( G(\beta) = 0 \). By Assumption 1 this \( \beta \) is unique. In addition, Assumption 1 implies that \( G(\beta) \)
is first strictly decreasing and then strictly increasing on \( \beta \in (0, 1/2) \). This gives (4). \( \blacksquare \)

The following result says the planted solution is completely frozen with high probability in the
planted model.

**Lemma 7.** Under Assumption 1, for any \( \alpha \in (0, \alpha_c) \) and any \( \delta \in (0, \beta_c) \), there exists a
constant \( c_\delta > 0 \) such that

\[ P_{pl}(\{ \sigma \in S : \text{dist}(\sigma, \sigma^*) \leq (\beta_c - \delta)N \} \neq \{ \sigma^* \}) \leq \exp \left\{ -c_\delta \sqrt{N} \right\}. \]

**Proof.** Let \( q(m) = P(\sigma, \sigma' \in H_\alpha(X)) \), where \( \sigma, \sigma' \in \Sigma_N \) are two arbitrary vectors with
\( \langle \sigma, \sigma' \rangle = m \) and \( X \) is a standard \( N \)-dimensional Gaussian vector. Then

\[ q(m) = q_\kappa \left( \frac{1}{2} + \frac{m}{2N} \right) \],  \hspace{1cm} (5)
where \( q_\kappa \) is defined in (1). We will compute the expected number of solutions with a given Hamming distance from \( \sigma^* \).

**Case 1.** We first consider the case that dist(\( \sigma, \sigma^* \)) is close to \( N \) or 0. We claim that there exists a constant \( \epsilon > 0 \) sufficiently small such that for all \( m \leq \epsilon N \),

\[
\log \mathbb{E}_{pl}[|\{ \sigma \in S : |\langle \sigma, \sigma^* \rangle| = N - m \}|] \leq -c\sqrt{mN}.
\]

To prove (6), we let \( \sigma_0 \) be an arbitrary vector in \( \Sigma_N \). Note that

\[
\mathbb{E}_{pl}[|\{ \sigma \in S : \langle \sigma, \sigma^* \rangle = N - m \}|] = \sum_{\sigma : \langle \sigma, \sigma_0 \rangle = N - m} \mathbb{P}(\sigma \in S|\sigma_0 \in S) = \binom{N}{m} \frac{q(m)}{p(\kappa)} a^N.
\]

We claim that uniformly over all \( |m| \leq \epsilon N \) with sufficiently small \( \epsilon \),

\[
q(N - m) \leq p(\kappa) - c\sqrt{m/N}
\]

Provided with (8), we have

\[
(7) \leq \exp \left\{ m \log(2N/m)/2 - c\sqrt{mN} \right\} \leq \exp \left\{ -c\sqrt{mN} \right\}.
\]

Hence (6) follows. Now it remains to prove (8). To this end, let \( 1 \) be the all 1’s vector of length \( N \) and \( 1_m \) be an \( N \)-dimensional vector with the first \( (N - m) \) coordinates +1 and the remaining \( m \) coordinates –1. We write

\[
p(\kappa) - q(N - m) = \mathbb{P}\left( |\langle 1, X \rangle| \leq \kappa\sqrt{N}, \left| \frac{1_m}{\sqrt{N}} , X \right| > \kappa\sqrt{N} \right)
\]

\[
\geq \mathbb{P}\left( \sum_{i=1}^{N-\frac{m}{2}} X_i \in (\kappa\sqrt{N}, \kappa\sqrt{N} + \sqrt{m}), \quad \sum_{i=N-\frac{m}{2}+1}^{N} X_i \in (-3\sqrt{m}, -2\sqrt{m}) \right)
\]

\[
= \mathbb{P}\left( Z_1 \in \left( \kappa\sqrt{\frac{N}{N-m}} , \kappa\sqrt{\frac{N}{N-m}} + \sqrt{\frac{m}{N-m}} \right), Z_2 \in (-3\sqrt{2}, -2\sqrt{2}) \right)
\]

\[
\geq c\sqrt{m/N}
\]

where \( Z_1, Z_2 \) are two independent standard Gaussian random variables. This proves (8).

**Case 2.** Next, we consider the case that dist(\( \sigma, \sigma^* \)) is away from 0 or \( N \). We claim that for any \( \epsilon > 0 \), uniformly in \( |m| \leq N - \epsilon N \),

\[
\lim_{N \to \infty} \frac{1}{N} \log \mathbb{E}_{pl}\left[ |\{ \sigma \in S : \langle \sigma, \sigma^* \rangle = m \}| \right] = \left[ F\left( \frac{1}{2} + \frac{m}{2N} \right) - \alpha \log p(\kappa) \right] = 0.
\]

To prove (9), we note that

\[
\mathbb{E}_{pl}\left[ |\{ \sigma \in S : \langle \sigma, \sigma^* \rangle = m \}| \right] = \left( \binom{N}{N+m/2} \right) \frac{q(m)}{p(\kappa)} a^N
\]
and that for any $\epsilon > 0$, uniformly in $\epsilon N \leq k \leq N - \epsilon N$,

$$\frac{1}{N} \log \binom{N}{k} \to 1 \quad \text{as } N \to \infty.$$ Combining these with (5) yields (9).

Now, let $\delta \in (0, \beta_c)$. To prove Lemma 7, it suffices to show that

$$\mathbb{E}_{pl} \left[ \sigma \in S : 0 < \text{dist}(\sigma, \sigma^*) \leq (\beta_c - \delta)N \right] \leq \exp \left\{ -c_\delta \sqrt{N} \right\}.$$ This bound follows from (6), (9), and (4).

3 \ | \ CONCENTRATION OF THE NUMBER OF SOLUTIONS

Fixing $N$, consider the symmetric perceptron as a discrete-time stochastic process with one constraint vector added at each time step. The solution space at time $t \in \mathbb{N}$ is defined as

$$S_t := \bigcap_{i=1}^t H_\kappa(X_i)$$

which is the intersection of $t$ random rectangles.

The following strengthening of Theorem 2 is the main result of this section.

**Theorem 2'.** Under Assumption 1, for every $\epsilon > 0$ there exists $M = M(\epsilon)$ such that for any $\alpha < \alpha_c$,

$$\limsup_{N \to \infty} \sup_{0 \leq t \leq aN} \mathbb{P} \left( \left| \log \left( \frac{|S_t|}{\mathbb{E}[|S_t|]} \right) \right| \geq M \log N \right) \leq \epsilon. \quad (11)$$

Theorem 2 follows immediately since $\frac{1}{N} \log \mathbb{E}|S_t| = \log 2 + \frac{t}{n} \log p(\kappa)$.

Theorem 2' says that the cardinality of the solution space will only deviate from its expectation slightly after adding $aN$ random constraints. To prove this theorem, we will look at the change in this deviation at each time when a new constraint is added. Write

$$Q_t := \log \left( \frac{|S_t|}{\mathbb{E}[|S_t|]} \right) = \sum_{i=1}^t \log \left( \frac{|S_t|/|S_{t-1}|}{\mathbb{E}[|S_t|]/\mathbb{E}[|S_{t-1}|]} \right).$$

Note that $\mathbb{E}[|S_t|]/\mathbb{E}[|S_{t-1}|] = p(\kappa)$. Let

$$Y_t := \frac{1}{p(\kappa)} \left( \frac{|S_t|}{|S_{t-1}|} - p(\kappa) \right),$$

so that

$$Q_t = \sum_{i=1}^t \log(1 + Y_i). \quad (12)$$
Since $0 \leq |S_t| \leq |S_{t-1}|$, we have that $-1 \leq Y_t \leq (1 - p(\kappa))/p(\kappa)$; however, we expect that the $Y_t$’s are very close to zero with high probability. Hence by a Taylor expansion, as $N \to \infty$,

$$Q_t = \sum_{i=1}^{t} Y_i - \frac{Y_t^2}{2} + o(1).$$

Notice that $(\sum_{i=0}^{t} Y_i)_{t \geq 0}$ is a martingale with respect to the filtration $\mathcal{F}_t = \sigma(S_1, S_2, \ldots, S_t)$, $t \geq 1$. If we know that $Y_t$’s are of order $N^{-1/2}$, we expect $Q_t$ to be of order 1.

The key ingredient here is to provide good bounds on $Y_t$’s. This is done by an inductive procedure: We prove that $Y_t$ has an exponential tail provided the condition that two random samples from $S_{t-1}$ are almost orthogonal with high probability. This condition, as well as the bound on $Y_t$, depends on the deviation in the previous step $|Q_{t-1}|$. As a result, a bound on $|Q_{t-1}|$ yields the orthogonality property of random solution pairs, which then provide a bound on $Y_t$ and hence a bound on the next $|Q_t|$.

**Definition 8.** For each $t \geq 0$, we let $(\sigma_i^{(t)})_{i \geq 1}$ be independent uniform random samples in $S_t$ and denote $\mathbb{P}_t(\cdot) := \mathbb{P}(\cdot | \mathcal{F}_t)$, $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_t]$. With $C_2 > 0$ a constant to be determined later in Lemma 11, we say $S_t$ is regular if

$$\mathbb{P}_t \left( \left| \left\langle \sigma_1^{(t)}, \sigma_2^{(t)} \right\rangle \right| \leq C_2 \sqrt{N} \sqrt{|Q_t| + \log N} \right) \geq 1 - N^{-10}.$$ 

Roughly speaking, $S_t$ is said to be regular if two random samples from $S_t$ are almost orthogonal with high probability. Define the stopping time

$$\tau_S := \inf \{ t \geq 0 : S_t \text{ is not regular} \}.$$ 

The following lemma says that for regular $S_t$, $Y_{t+1}$ is roughly of order $N^{-1/2}$.

**Proposition 9.** There exists a constant $C > 0$ such that for all $t \geq 0$,

$$\mathbb{P}_t \left( |Y_{t+1}| \geq C \sqrt{|Q_t| + \log N} \right) \leq C \exp(-x).$$

By (12), Proposition 9 provides an upper bound on $Y_{t+1}$, which can be used to control the increment of $|Q_t|$. This will be one of the key ingredients in proving Lemma 10, which gives an upper bound on $|Q_t|$ inductively for time $t$ before a stopping time defined below.

With $C_3 > 0$ a constant to be determined later in Lemma 11, we define

$$\tau_Y := \inf \left\{ t \geq 1 : |Y_t| \geq C_3 \sqrt{|Q_{t-1}| + \log N} \log N \right\},$$

$$\tau_Q := \inf \{ t \geq 0 : |Q_t| \geq (\log N)^2 \},$$

and

$$\tau = \tau_S \wedge \tau_Y \wedge \tau_Q.$$ (14)

**Lemma 10.** There exists a constant $C > 0$ such that for all $t \geq 1$,

$$\mathbb{E}[|Q_{t \wedge \tau - 1}|] \leq \exp(Ct/N) \log N.$$ (15)
Lemma 11 says that the stopping time $\tau$ will occur later than time $\alpha N$ with high probability. Theorem 2′ will be a direct consequence of Lemmas 10 and 11.

**Lemma 11.** Under Assumption 1, there exists constants $C_2, C_3 > 0$ such that for any $\alpha < \alpha_c$,

$$\lim_{N \to \infty} \mathbb{P}(\tau < \alpha N) = 0.$$  

Proof of Theorem 2′ and Lemma 3. By Lemmas 10 and 11, the probability bound (11) follows from Markov’s inequality. Combining (11) with Lemma 11 gives Lemma 3. □

3.1 Proof of Proposition 9

In this section, we consider an arbitrary subset $A \subset \Sigma_N$ and define the probability measure $\mathbb{P}_A$ under which $(\sigma_i)_{i \geq 1}$ are independent uniformly random vectors in $A$.

Suppose $S_{t-1}$ is given and we consider $A = S_{t-1}$. Proposition 9 says if two independent samples from $\mathbb{P}_A$ are almost orthogonal, then $\mathbb{P}_A(\sigma_1 \in H_\psi(X))$ should be close to $p(\psi)$. To prove this, we first fix $X$ and show in Lemma 12 that if $\mathbb{P}_A(\sigma_1 \notin H_\psi(X))$ is large, then there is a non-negligible probability that $k$ independent samples from $A$ are pairwise almost orthogonal and none of them is in $H_\psi(X)$. Comparing it with the expectation (over $X$) of such a probability, in Proposition 9′ we show that $\mathbb{P}_A(\sigma_1 \in H_\psi(X))$ is unlikely to be much smaller than $p(\psi)$. Proposition 9 is a direct consequence of Proposition 9′.

**Lemma 12.** For all $N$ sufficiently large and every $B \subset A \subset \Sigma_N$, if we denote

$$\delta_A = \mathbb{P}_A(|\langle \sigma_1, \sigma_2 \rangle| > \varphi_N \sqrt{N}) \quad \text{and} \quad q = \mathbb{P}_A(\sigma_1 \notin B),$$

then for all $1 \leq k \leq \frac{q}{4 \delta_A}$

$$\mathbb{P}_A(E_k(B)) \geq \left(q - \frac{2k \delta_A}{q}\right)^k,$$

where

$$E_k(B) := \{ \sigma \notin B, \forall 1 \leq i \leq k \} \cap \left\{ |\langle \sigma_i, \sigma_j \rangle| \leq \varphi_N \sqrt{N} \quad \forall 1 \leq i, j \leq k, i \neq j \right\}. \quad (16)$$

Proof. We will prove the following by induction

$$\mathbb{P}_A(E_{t+1}|E_t) \geq q - \frac{2t \delta_A}{q} \quad \text{for} \quad 1 \leq t \leq \frac{q^2}{4 \delta_A}. \quad (17)$$

First, since $\mathbb{P}_A(E_1) = \mathbb{P}_A(\sigma_1 \notin B) = 1 - q$, we see that (17) holds for $t = 1$.

Next, we suppose (17) is true for $t = k$ with $k \leq \frac{q^2}{4 \delta_A} - 1$. Since $\sigma_{k+1}$ is independent of $E_k$, we have

$$\mathbb{P}_A(E_{k+1}|E_k) \geq \mathbb{P}_A(\sigma_{k+1} \notin B|E_k) - \sum_{i=1}^{k} \mathbb{P}_A(|\langle \sigma_i, \sigma_{k+1} \rangle| > \varphi_N \sqrt{N}|E_k)$$

$$= q - k \mathbb{P}_A(|\langle \sigma_k, \sigma_{k+1} \rangle| > \varphi_N \sqrt{N}|E_k).$$
Also, since $(\sigma_k, \sigma_{k+1})$ is independent of $E_{k-1}$ and $E_k \supset E_k$,
\[
P_A(|\langle \sigma_k, \sigma_{k+1} \rangle| > \varphi_N \sqrt{N} |E_k) \leq \frac{P_A(\{ |\langle \sigma_k, \sigma_{k+1} \rangle| > \varphi_N \sqrt{N} \} \cap E_{k-1})}{P_A(E_k)}
= P_A(|\langle \sigma_k, \sigma_{k+1} \rangle| > \varphi_N \sqrt{N}) \cdot \frac{P(E_{k-1})}{P(E_k)}.
\]

By the induction hypothesis
\[
\frac{P(E_{k-1})}{P(E_k)} = \frac{1}{P(E_k |E_{k-1})} \leq \frac{q}{2},
\]
where we used the assumption $k \leq \frac{q^2}{4\delta_A} - 1$. Combining with the previous two inequalities, we get
\[
P_A(E_{k+1} |E_k) \geq q - \frac{2k\delta_A}{q}.
\]
Hence (17) holds.

**Proposition 9'.** There exists constant $C, c > 0$ depending only on $\kappa$ such that for all $N$ sufficiently large, all $A \subset \Sigma_N$, and $\varphi_N \in \mathbb{R}$ satisfying
\[
\delta_A := P_A(|\langle \sigma_1, \sigma_2 \rangle| > \varphi_N \sqrt{N}) \leq \varphi_N^{-2} N^{-2},
\]
we have
\[
P\left(\left|\frac{|H_k(X) \cap A|}{|A|} - p(\kappa)\right| > \frac{\varphi_N x}{\sqrt{N}}\right) \leq C \exp(-cx),
\]
where $X$ is a standard $N$-dimensional Gaussian vector.

**Proof.** We will prove that
\[
P\left(\left|\frac{|H_k(X) \cap A|}{|A|} < p(\kappa) - \frac{\varphi_N x}{\sqrt{N}}\right) \leq C \exp(-cx).
\]
The other direction can be proved similarly.

Recall $E_k(\cdot)$ as in (16) and denote
\[
A_k^+ := \{(\sigma_i)_{i=1}^k \in A^k : |\langle \sigma_i, \sigma_j \rangle| \leq \varphi_N \sqrt{N} \quad \forall 1 \leq i, j \leq k, i \neq j\}.
\]
Then
\[
\mathbb{E}\left[P_A(E_k(H_k(X)))\right] = \sum_{(\sigma_i)_{i=1}^k \in A_k^+} |A|^{-k} P(\sigma_i \notin H_k(X), \forall 1 \leq i \leq k).
\]
To bound the probabilities on the right hand side, we use the following lemma which controls the joint probabilities of weakly correlated Gaussian variables. Its proof appears in Appendix A.
Lemma 13. For any $\kappa > 0$ and $m \in \mathbb{N}$, there exists constants $c_{\kappa}, C > 0$ such that if we let $(X_i)_{i \geq 1}$ be a Gaussian vector with

$$
E[X_i] = 0, \quad E[X_i^2] = 1, \quad |E[X_iX_j]| \leq c_{\kappa}m^{-1},
$$

then

$$
\mathbb{P}\left( \min_{1 \leq i \leq m} |X_i| > \kappa \right) \leq C \mathbb{P}(|X_1| > \kappa)^m, \quad \text{(20)}
$$

$$
\mathbb{P}\left( \max_{1 \leq i \leq m} |X_i| \leq \kappa \right) \leq C \mathbb{P}(|X_1| \leq \kappa)^m. \quad \text{(21)}
$$

By Lemma 13, we have that for $k = \lfloor c_{\kappa}q_{N}^{-1}\sqrt{N} \rfloor$ and all $(\sigma_i)_{i=1}^k \in A_k^+$,

$$
\mathbb{P}(\sigma_i \notin H_k(X), \forall 1 \leq i \leq k) \leq C \mathbb{P}(|Z| > \kappa)^k.
$$

Hence

$$
\mathbb{E}\left[ \mathbb{P}_A(E_k(H_k(X))) \right] \leq C \mathbb{P}(|Z| > \kappa)^k. \quad \text{(22)}
$$

On the other hand, we set

$$
\Delta = \frac{q_{N}X}{\sqrt{N}}.
$$

Then (18) implies $\Delta \geq \frac{4k\Delta}{q}$ for our choice of $k$. Therefore, it follows from Lemma 12 that on the event \{ $\mathbb{P}_A(\sigma_1 \notin H_k(X)|H_k(X)) \geq \mathbb{P}(|Z| > \kappa) + \Delta$ \}, we have

$$
\mathbb{P}_A(E_k[H_k(X)])H_k(X) \geq \mathbb{P}_A(\sigma_1 \notin H_k(X)|H_k(X)) - \Delta/2)^k
$$

$$
\geq \mathbb{P}(|Z| > \kappa)^k e^{k\Delta/2}.
$$

Combined with (22), this yields (19).

\[ \blacksquare \]

3.2 | Proof of Lemma 10

By the definitions (13) and (14), we see that for all $0 \leq i \leq \tau - 1$,

$$
|Y_i| \leq C \sqrt{\frac{|Q_{i-1}| + \log N}{N}} \leq C \sqrt{\frac{(\log N)^2 + \log N}{N}}.
$$

Hence it follows from $|\log(1+x) - x| \leq x^2$ for all $x \geq -3/5$ that

$$
\left| \sum_{j=1}^{i\wedge \tau - 1} \log(1 + Y_i) \right| \leq \left| \sum_{j=1}^{i\wedge \tau - 1} Y_i \right| + \sum_{j=1}^{i\wedge \tau - 1} Y_i^2 \leq \left| \sum_{j=1}^{i\wedge \tau} Y_i \right| + \sum_{j=1}^{i\wedge \tau} Y_i^2 + |Y_{i\wedge \tau}|.
$$

Since $\tau \leq \tau_5$, Proposition 9 implies

$$
\mathbb{E}[Y_i^2 1_{i\wedge \tau}] = \mathbb{E}[E_i[1_{i\wedge \tau} 1_{i\geq i-1}]] \leq C' \mathbb{E}[|Q_{i-1}| 1_{i\geq i-1}] + \log N.
$$
Therefore, since $(\sum_{i=1}^{t \wedge \tau} Y_i)_{t \geq 1}$ is a martingale and $t \land \tau \leq \tau_S$, we have

$$E \left| \sum_{i=1}^{t \wedge \tau - 1} \log(1 + Y_i) \right| \leq \sqrt{E \sum_{i=1}^{t \wedge \tau} Y_i^2 + E \sum_{i=1}^{t \wedge \tau} Y_i^2 + E[Y_{t \wedge \tau}]}
$$

$$\leq 2E \sum_{i=1}^{t \wedge \tau} Y_i^2 1_{t \geq 1} + 1 + E[Y_{t \wedge \tau}]
$$

$$\leq 2C' \cdot t \log N + \sum_{i=1}^{t \wedge \tau} \log(1 + Yi)
$$

where $C'$ is a constant. Hence

$$E[|Q_{t \wedge \tau - 1}|] \leq 3C' \left( \frac{t \log N + \sum_{i=1}^{t-1} b_i}{N} + \frac{1}{p(\kappa)} \right).
$$

We claim that this yields (15). To prove this, we define $b_1 := E[|Q_{1 \wedge \tau - 1}|]$ and for $t \geq 2$ define

$$b_t := 3C' \left( \frac{t \log N + \sum_{i=1}^{t-1} b_i}{N} + \frac{1}{p(\kappa)} \right). \quad (23)
$$

Then a straightforward induction argument shows that

$$b_t \geq E[|Q_{t \wedge \tau - 1}|] \quad \text{for } t \geq 1. \quad (24)
$$

On the other hand, (23) implies

$$b_t = 3C' \left( \frac{\log N + b_{t-1}}{N} + \frac{(t-1) \log N + \sum_{i=1}^{t-2} b_i}{N} + \frac{1}{p(\kappa)} \right)
$$

$$= 3C' \left( \frac{\log N + b_{t-1}}{N} + \frac{b_{t-1}}{3C'} \right).
$$

This implies

$$b_t + \log N = \left( 1 + \frac{3C'}{N} \right) (b_{t-1} + \log N) = \left( 1 + \frac{3C'}{N} \right)^{t-1} (b_1 + \log N). \quad (25)
$$

Note that $S_0 = \{-1, 1\}^N$, $Q_0 = 0$ and $P(\tau \geq 1) = 1$. We get $b_1 = 0$. Hence (15) follows from (24) and (25). We complete the proof of Lemma 10.

### 3.3 Proof of Lemma 11

Write

$$P(\tau \geq t) \leq P(\tau_S \leq aN) + P(\tau_Y \leq \tau_S, \tau_Y \leq aN) + P(\tau_Q \leq t, \tau_Q = \tau, |Q_{\tau_Q-1}| \leq (\log N)^2 / 2)
$$

$$+ P(\tau_Q \leq t, \tau_Q = \tau, |Q_{\tau_Q-1}| \geq (\log N)^2 / 2).$$
Lemma 10 and the Markov inequality yield
\[ P(\tau_Q \leq t, \tau_Q = \tau, |Q_{\tau_Q - 1}| \geq (\log N)^2/2) \leq ((\log N)^2/2)^{-1}E[|Q_{\tau_Q - 1}| : \tau_Q \leq t, \tau_Q = \tau, |Q_{\tau_Q - 1}| \geq (\log N)^2/2] \leq ((\log N)^2/2)^{-1}E[|Q_{\tau_Q - 1}|] \leq (\log N)^{-1/2}. \]

Hence, to prove Lemma 11, it suffices to prove there exist constants \( C_2, C_3 > 0 \) such that
\[
\begin{align*}
P(\tau_Y \leq \tau_S, \tau_Y \leq aN) &\leq N^{-8}, \quad (26) \\
P(|Q_{\tau_Q - 1}| \leq (\log N)^2/2, \tau_Q \leq \tau_S, \tau_Q = t + 1) &\leq N^{-10}. \quad (27)
\end{align*}
\]

First, (26) follows from Proposition 9 and a union bound. Next, Proposition 9 gives
\[
P(|Q_{\tau_Q - 1}| \leq (\log N)^2/2, \tau_Q \leq \tau_S, \tau_Q = t + 1) \\
= E \left[ P_t(|Q_t + \log(1 + Y_t+1)| \geq (\log N)^2) \mathbf{1}_{|Q_t| \leq (\log N)^2/2, \tau_t \geq t+1} \right] \\
\leq E \left[ P_t(Y_t \leq -1/2) \mathbf{1}_{|Q_t| \leq (\log N)^2/2, \tau_t \geq t+1} \right] \\
\leq N^{-10}.
\]

This yields (27). Finally, we prove (28). By definition, it suffices to prove that there exists a constant \( C_2 > 0 \) such that for every \( 1 \leq t \leq aN \),
\[
P \left( \left| \left\langle \sigma_1^{(t)}, \sigma_2^{(t)} \right\rangle \right| \geq C_2 \sqrt{N} \sqrt{|Q_t| + \log N} \geq N^{-10} \right) \leq N^{-10}.
\]

By the Markov inequality, this follows from
\[
P \left( \left| \left\langle \sigma_1^{(t)}, \sigma_2^{(t)} \right\rangle \right| \geq C \sqrt{N} \sqrt{|Q_t| + \log N} \right) \leq N^{-100} \quad (29)
\]
for some constant \( C > 0 \). In the rest of the proof, we will prove (29) by bounding the exponential moment of \( \left\langle \sigma_1^{(t)}, \sigma_2^{(t)} \right\rangle^2 / N \).

By (10) and \( \binom{n}{k} \leq \sqrt{\frac{n}{2\pi k(n-k)}} \exp(nH(k/n)) \) (see e.g., [Chap. 10, lemma 7] [36]), we have that
\[
E[|\{(\sigma_1, \sigma_2) \in S_t : |\langle \sigma_1, \sigma_2 \rangle| \geq \lambda \sqrt{N} \}|] \\
\leq 2E[|S_t|] + 2^{-N} \sum_{\lambda \sqrt{N} \leq |m| < N} C \sqrt{\frac{N}{N^2 - m^2}} \exp \left\{ NF \left( \frac{1}{2} + \frac{m}{2N} \right) - t \log p(\kappa) \right\}.
\]

It is proved in [6] that Assumption 1 implies \( F'(1/2) = 0, F''(1/2) < 0 \) and \( F(1/2) > F(\beta) \) for all \( \beta \neq 1/2 \). Hence there exists a small constant \( c > 0 \) such that \( F\left( \frac{1}{2} + \lambda \right) \leq F\left( \frac{1}{2} \right) - c\lambda^2 \) for all \( \lambda \in [-1, 1] \). Hence the previous display is
\[
\leq 2E[|S_t|] + C2^{-N} \exp \left\{ NF \left( \frac{1}{2} \right) - t \log p(\kappa) \right\} \sum_{\lambda \sqrt{N} \leq |m| < N} \left( \frac{N - m^2}{N} \right)^{-1/2} \exp \left\{ -c \frac{m^2}{4N} \right\} \\
\leq 2E[|S_t|] + C2^{-N} \exp \left\{ NF \left( \frac{1}{2} \right) - t \log p(\kappa) - \frac{c\lambda^2}{8} \right\}.
\]
Note that $2^{-N} \exp \left\{ NF \left( \frac{1}{2} - t \log p(\kappa) \right) \right\} = \mathbb{E}[|S_t|^2]$. There exist $C', c' > 0$ such that

$$\mathbb{E} \left[ \mathbb{P}_t \left( \left| \left\langle \sigma^{(i)}_1, \sigma^{(i)}_2 \right\rangle \right| \geq \lambda \sqrt{N} \right) \frac{|S_t|^2}{\mathbb{E}[|S_t|^2]} \right] \leq C' \exp \left( - c' \lambda^2 \right).$$

Therefore,

$$\mathbb{E} \left[ \exp \left( \frac{c'}{2} \frac{\left| \left\langle \sigma^{(i)}_1, \sigma^{(i)}_2 \right\rangle \right|^2}{N} + 2Q_t \right) \right] = \mathbb{E} \left[ \int_0^\infty c' \lambda e^{-\frac{c'}{2} \frac{\left| \left\langle \sigma^{(i)}_1, \sigma^{(i)}_2 \right\rangle \right|^2}{N}} \mathbb{P}_t \left( \left| \left\langle \sigma^{(i)}_1, \sigma^{(i)}_2 \right\rangle \right| \geq \lambda \right) d\lambda \cdot \frac{|S_t|^2}{\mathbb{E}[|S_t|^2]} \right] \leq C' \int_0^\infty c' \lambda e^{-\frac{c'}{2} \frac{\left| \left\langle \sigma^{(i)}_1, \sigma^{(i)}_2 \right\rangle \right|^2}{N}} d\lambda \leq C.$$

This yields (29).

### 4 FROM THE PLANTED MODEL TO THE RANDOM MODEL

We will use Theorem 2 to prove the following lemma, which says events that occur with probability at most $N^{-o(1)}$ in the planted model occur with probability $o(1)$ in the standard model. Then, Theorem 1 will follow in conjunction with Lemma 7.

**Lemma 14.** Suppose $A \subseteq \{ (\sigma, U) : \sigma \in U \subseteq \Sigma_N \}$, and suppose $\alpha < \alpha_c$. If $\mathbb{P}_{pl}( (\sigma^*, S(X)) \in A) \leq N^{-m_N}$ for some $m_N \to \infty$, then

$$\mathbb{P}_t( (\sigma^*, S(X)) \in A) \to 0 \quad \text{as } N \to \infty.$$

**Proof.** Theorem 2 implies the event

$$G = \left\{ \frac{|S|}{\mathbb{E}[|S|]} \geq \exp \{-m_N \log N\} \right\}$$

has probability $1 - o(1)$. We will use the following relation between random model and planted model:

$$\mathbb{P}_{pl}( (\sigma^*, S(X)) \in A) = \mathbb{E}_r \left[ 1_{(\sigma^*, S(X)) \in A} \cdot \frac{|S(X)|}{\mathbb{E}[|S(X)|]} \right] \cdot \mathbb{P}(S(X) \neq \emptyset). \quad (30)$$

Provided with (30), we have, as $N \to \infty$,

$$\mathbb{P}_t( (\sigma^*, S) \in A) \leq \mathbb{P}_t( (\sigma^*, S) \in A \cap G) + \mathbb{P}(G^c) / \mathbb{P}(S \neq \emptyset)$$

$$\leq \exp \{m_N \log N\} \mathbb{E} \left[ 1_{(\sigma^*, S) \in A} \cdot \frac{|S|}{\mathbb{E}[|S|]} \right] + o(1)$$

$$= \exp \{m_N \log N\} \mathbb{P}_{pl}( (\sigma^*, S) \in A) / \mathbb{P}(S \neq \emptyset) + o(1)$$

$$= o(1).$$

where we used (30) in the third step. This yields Lemma 14.
It remains to prove (30). To this end, it suffices to prove that for any \( \sigma \in \Sigma_N \) and \( U \subset \Sigma_N \) such that \( \sigma \in U \), we have

\[
\mathbb{P}_{\text{pl}}(\sigma^* = \sigma, S(\mathbf{X}) = U) = \mathbb{P}_i(\sigma^* = \sigma, S(\mathbf{X}) = U) \frac{|U|}{\mathbb{E}[|S(\mathbf{X})|]} \cdot \mathbb{P}(S(\mathbf{X}) \neq \emptyset).
\]

(31)

Now, write

\[
\mathbb{P}_{\text{pl}}(\sigma^* = \sigma, S(\mathbf{X}) = U) = |\Sigma_N|^{-1} \mathbb{P}(S(\mathbf{X}) = U | \sigma \in S(\mathbf{X}))
\]

\[
= \frac{\mathbb{P}(S(\mathbf{X}) = U, \sigma \in S(\mathbf{X}))}{|\Sigma_N| \mathbb{P}(\sigma \in S(\mathbf{X}))}
\]

\[
= \frac{\mathbb{P}(S(\mathbf{X}) = U)}{|\Sigma_N| \mathbb{P}(\sigma \in S(\mathbf{X}))}.
\]

Note that \( \mathbb{E}[|S(\mathbf{X})|] = |\Sigma_N| \mathbb{P}(\sigma \in S(\mathbf{X})) \) and that

\[
\mathbb{P}_i(\sigma^* = \sigma, S(\mathbf{X}) = U) \cdot \mathbb{P}(S(\mathbf{X}) \neq \emptyset) = |U|^{-1} \mathbb{P}(S(\mathbf{X}) = U).
\]

We get (31), and thus complete the proof of Lemma 14.

Finally we prove Theorem 1.

Proof of Theorem 1. Let

\[
A = \{(\tau, U) : \{\sigma \in U : \langle \sigma, \tau \rangle \geq (d_c + \delta)N \} \neq \{\tau\}\}.
\]

Then Lemma 7 implies that

\[
\mathbb{P}_{\text{pl}}((\sigma^*, S(\mathbf{X})) \in A) \leq \exp \left\{ -c \sqrt{N} \right\}
\]

Combined with Lemma 14, this yields

\[
\mathbb{P}_i((\sigma^*, S(\mathbf{X})) \in A) = o(1),
\]

and thus Theorem 1.

ACKNOWLEDGMENTS
We thank Benjamin Aubin and Lenka Zdeborová for inspiring discussions and introducing us to this problem. Will Perkins was supported in part by NSF grant DMS-2309958.

REFERENCES
1. E. Abbe, Community detection and stochastic block models: Recent developments, J Mach Learn Res 18 (2017), no. 1, 6446–6531.
2. D. Achlioptas and A. Coja-Oghlan. Algorithmic barriers from phase transitions. Paper presented at: 2008 49th annual IEEE symposium on foundations of computer science, IEEE. 2008 793–802.
3. D. Achlioptas, A. Coja-Oghlan, and F. Ricci-Tersenghi, *On the solution-space geometry of random constraint satisfaction problems*, Random Struct. Algorithm. 38 (2011), no. 3, 251–268.

4. E. Abbe, S. Li, and A. Sly, Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron. Paper presented at: 2021 IEEE 62nd annual symposium on foundations of computer science (FOCS), IEEE. 2022 327–338.

5. E. Abbe, S. Li, and A. Sly, Binary perceptron: Efficient algorithms can find solutions in a rare well-connected cluster, Proceedings of the 54th annual ACM SIGACT symposium on theory of computing. 2022 860–873.

6. B. Aubin, W. Perkins, and L. Zdeborová, *Storage capacity in symmetric binary perceptrons*, J. Phys. A Math. Theor. 52 (2019), no. 29, 294003.

7. C. Baldassi, *Generalization learning in a perceptron with binary synapses*, J. Stat. Phys. 136 (2009), no. 5, 902–916.

8. C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina, *Efficient supervised learning in networks with binary synapses*, Proc. Natl. Acad. Sci. 104 (2007), no. 26, 11079–11084.

9. C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, *Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes*, Proc Natl Acad Sci 113 (2016), no. 48, E7655–E7662.

10. V. Bapst, A. Coja-Oghlan, and C. Efthymiou, *Planting colourings silently*, Comb. Probab. Comput. 26 (2017), no. 3, 338–366.

11. V. Bapst, A. Coja-Oghlan, S. Hetterich, F. Raßmann, and D. Vilenchik, The condensation phase transition in random graph coloring, Commun. Math. Phys. 341 (2016), no. 2, 543–606.

12. C. Baldassi, R. D. Vecchia, C. Lucibello, and R. Zecchina, *Clustering of solutions in the symmetric binary perceptron*, J Stat Mech: Theory Exp 2020 (2020), no. 7, 073303.

13. C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses, Phys. Rev. Lett. 115 (2015), no. 12, 128101.

14. C. Baldassi, A. Coja-Oghlan, C. Efthymiou, N. Jaafari, M. Kang, and T. Kapetanopoulos, Charting the replica symmetric phase, Commun. Math. Phys. 359 (2018), no. 2, 603–698.

15. C. Baldassi, A. Coja-Oghlan, F. Krzakala, W. Perkins, and L. Zdeborová, Information-theoretic thresholds from the cavity method, Adv. Math. 333 (2018), 694–795.

16. T. M. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE Trans Electron Comput 3 (1965), 326–334.

17. A. Coja-Oghlan and L. Zdeborová. The condensation transition in random hypergraph 2-coloring. Proceedings of the twenty-third annual ACM-SIAM symposium on discrete algorithms, SIAM. 2012 241–250.

18. J. Ding and N. Sun. Capacity lower bound for the Ising perceptron. Proceedings of the 51st annual ACM SIGACT symposium on theory of computing. 2019 816–827.

19. J. Ding, A. Sly, and N. Sun. Satisfiability threshold for random regular NAE-SAT. Proceedings of the forty-sixth annual ACM symposium on theory of computing. 2014 814–822.

20. E. Friedgut, Sharp thresholds of graph properties, and the k-sat problem, J. Am. Math. Soc. 12 (1999), no. 4, 1017–1054.

21. E. Gardner, *Maximum storage capacity in neural networks*, Europhys. Lett. 4 (1987), no. 4, 481.

22. E. Gardner and B. Derrida, *Optimal storage properties of neural network models*, J. Phys. A Math. Gen. 21 (1988), no. 1, 271.

23. D. Gamarnik and E. C. K˝ozlida˘g, *Algorithmic obstructions in the random number partitioning problem*, Ann. Appl. Probab. (2021), preprint, arXiv:2103.01369.

24. H. Huang and Y. Kabashima, Origin of the computational hardness for learning with binary synapses, Phys. Rev. E 90 (2014), no. 5, 052813.

25. D. L. Hanson and F. T. Wright, A bound on tail probabilities for quadratic forms in independent random variables, Ann. Math. Statist. 42 (1971), 1079–1083.

26. K. Y. Haiping Huang, M. Wong, and Y. Kabashima, Entropy landscape of solutions in the binary perceptron problem, J. Phys. A Math. Theor. 46 (2013), no. 37, 375002.

27. W. Krauth and M. Mézard, Storage capacity of memory networks with binary couplings, J. Phys. 50 (1989), no. 20, 3057–3066.

28. F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborová, Gibbs states and the set of solutions of random constraint satisfaction problems, Proc. Natl. Acad. Sci. 104 (2007), no. 25, 10318–10323.

29. J. H. Kim and J. R. Roche, Covering cubes by random half cubes, with applications to binary neural networks, J. Comput. Syst. Sci. 56 (1998), no. 2, 223–252.
33. F. Krzakala and L. Zdeborová, *Hiding quiet solutions in random constraint satisfaction problems*, Phys. Rev. Lett. 102 (2009), no. 23, 238701.

34. M. Molloy, *The freezing threshold for k-colourings of a random graph*, J. ACM 65 (2018), no. 2, 1–62.

35. A. Montanari, R. Restrepo, and P. Tetali, *Reconstruction and clustering in random constraint satisfaction problems*, SIAM J. Discret. Math. 25 (2011), no. 2, 771–808.

36. F. J. MacWilliams and N. J. A. Sloane, *The theory of error correcting codes*, Vol. 16, Elsevier, Amsterdam, 1977.

37. D. Mitchell, B. Selman, and H. Levesque. *Hard and easy distributions of SAT problems*. Aaai, volume 92. 1992 459–465.

38. A. M. Ostrowski, *Sur l’approximation du determinant de fredholm par les determinants des systèmes d’équations linéaires*, Ark. Math. Stockholm 26A (1938), 1–15.

39. A. Sly, N. Sun, and Y. Zhang. *The number of solutions for random regular NAE-SAT*. Paper presented at: 2016 IEEE 57th annual symposium on foundations of computer science (FOCS), IEEE. 2016 724–731.

40. H. Sompolinsky, N. Tishby, and H. S. Seung, *Learning from examples in large neural networks*, Phys. Rev. Lett. 65 (1990), no. 13, 1683.

41. M. Talagrand, *Intersecting random half cubes*, Random Struct. Algoritm. 15 (1999), no. 3-4, 436–449.

42. M. Talagrand, Mean field models for spin glasses: Volume I: Basic examples, Vol 54, Springer Science & Business Media, Heidelberg, Berlin, 2010.

43. M. Talagrand, Mean field models for spin glasses: Volume II: Advanced replica-symmetry and low temperature, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics, Springer, Heidelberg, Berlin 2011.

44. D. J. Thouless, P. W. Anderson, and R. G. Palmer, *Solution of ‘Solvable model of a spin glass’*, Philos. Mag. 35 (1977), no. 3, 593–601.

45. X. Changji, *Sharp threshold for the Ising perceptron model*, Ann. Probab. 49 (2021), no. 5, 2399–2415.

46. L. Zdeborová and F. Krzakala, *Phase transitions in the coloring of random graphs*, Phys. Rev. E 76 (2007), no. 3, 031131.

47. L. Zdeborová and F. Krzakala, *Statistical physics of inference: Thresholds and algorithms*, Adv. Phys. 65 (2016), no. 5, 453–552.

48. L. Zdeborová and M. Mézard, *Constraint satisfaction problems with isolated solutions are hard*, J Stat Mech: Theory Exp 2008 (2008), no. 12, P12004.

**APPENDIX A: PROOF OF LEMMA 13**

Proof of Lemma 13. We first prove (20). Let $\Sigma = (\Sigma_{ij}) = (\mathbb{E}[X_iX_j]) \in \mathbb{R}^{m \times m}$, then

$$
\mathbb{P}\left( \min_{1 \leq i \leq m} |X_i| > \kappa \right) = \sqrt{\frac{1}{(2\pi)^m |\Sigma|}} \int_{\min_{1 \leq i \leq m} |x_i| > \kappa} \exp\left\{-\frac{x^T (\Sigma^{-1} - I)x + x^T x}{2}\right\} \, dx
$$

$$
= |\Sigma|^{-\frac{1}{2}} \mathbb{E} \left[ \exp\left\{-\frac{Z^T (\Sigma^{-1} - I)Z}{2}\right\} ; \min_{1 \leq i \leq m} |Z_i| > \kappa \right]
$$

where $Z = (Z_1, \ldots, Z_m)$ is a standard Gaussian vector. Since $(Z_1, \ldots, Z_m)$ and $(e_1Z_1, \ldots, e_mZ_m)$ has the same distribution for every $e_1 \in \{-1, 1\}$, we have

$$
\mathbb{P}\left( \min_{1 \leq i \leq m} |X_i| > \kappa \right) = |\Sigma|^{-\frac{1}{2}} \mathbb{E} \left[ 2^{-m} \sum_{(e_i)_{i=1}^m \in \{-1, 1\}^m} \exp\left\{-\frac{1}{2} \sum_{1 \leq i,j \leq m} g_{ij} e_i e_j\right\} ; \min_{1 \leq i \leq m} |Z_i| > \kappa \right]
$$
where \( g_{ij} = \delta_{ij} Z_i Z_j \) and \( \delta_{ij} \) is the \((i,j)\)-th element of the matrix \( \Sigma^{-1} - I \). Note that \( \Sigma_{ii} = 1 \) and \( |\Sigma_{ij}| \leq c_k m^{-1} \) for \( i \neq j \) implies

\[
\max_{ij} |\delta_{ij}| \leq C c_k m^{-1}, \quad \text{and} \quad \|\Sigma^{-1} - I\|_2 \leq C c_k. \quad \text{(A1)}
\]

Hence

\[
\sum_{ij} g_{ij}^2 \leq C c_k^2 m^{-2} |Z|^4. \quad \text{(A2)}
\]

Now, we first note that by (A1), for any \( q > 0 \)

\[
\mathbb{E} \left[ \exp \left\{ -\frac{1}{2} Z^T (\Sigma^{-1} - I) Z \right\}; |Z|^2 \geq q m \right] \leq \mathbb{E} \left[ \exp \left\{ \frac{C c_k |Z|^2}{2} \right\}; |Z|^2 \geq q m \right] = (1 - C c_k)^\frac{m}{2} \mathbb{P}(|Z|^2 \geq (1 - C c_k) q m). \]

Since the tail probability of a Chi-square random variable decays exponentially, we see that for \( c_k \) sufficiently small and \( q \) sufficiently large,

\[
\mathbb{E} \left[ \exp \left\{ -\frac{1}{2} Z^T (\Sigma^{-1} - I) Z \right\}; |Z|^2 \geq q m \right] \leq \mathbb{P}(|Z| > \kappa)^m. \quad \text{(A3)}
\]

Next, by (4) [28], we have that if \( \sum_{ij} g_{ij}^2 \) is sufficiently small (hence the \( \ell_2 \) operator norm of the matrix \((|g_{ij}|)\) is also small), then

\[
2^{-m} \sum_{(e_i)_{i=1}^m \in \{-1,1\}^m} \exp \left\{ -\frac{1}{2} \sum_{1 \leq i,j \leq m} g_{ij} e_i e_j \right\} \leq C.
\]

But we see from (A2) that if \( c_k \) is sufficiently small, then on the event \( \{|Z|^2 \leq q m\} \) the condition that \( |\sum_i g_{ii}| \) and \( \sum_{ij} g_{ij}^2 \) are sufficiently small holds. Combining with (A3), we see that

\[
\mathbb{E} \left[ 2^{-m} \sum_{(e_i)_{i=1}^m \in \{-1,1\}^m} \exp \left\{ -\frac{1}{2} \sum_{1 \leq i,j \leq m} g_{ij} e_i e_j \right\}; \min_{1 \leq i \leq m} |Z_i| > \kappa \right] \leq C \mathbb{P}(|Z| > \kappa)^m.
\]

Finally, [38] gives \( |\Sigma| \geq 1 - c_k \). We complete the proof of (20).

The other inequality (21) can be proved verbatim by changing all occurrences of \( \min_{1 \leq i \leq m} |X_i| > \kappa \) and \( \min_{1 \leq i \leq m} |Z_i| > \kappa \) to \( \max_{1 \leq i \leq m} |X_i| \leq \kappa \) and \( \max_{1 \leq i \leq m} |Z_i| \leq \kappa \). In fact, it would be even easier as \( \max_{1 \leq i \leq m} |Z_i| \leq \kappa \) implies \( |Z|^2 \leq \kappa m \), and thus there is no need to consider the case \( \{|Z|^2 \geq q m\} \).