Symmetric Perceptron with Random Labels

Eren C. Kızıldağ
Columbia University
eck2170@columbia.edu

Tanay Wakhare
Massachusetts Institute of Technology
twakhare@mit.edu

***Abstract*—The symmetric binary perceptron (SBP) is a random constraint satisfaction problem (CSP) and a single-layer neural network; it exhibits intriguing features, most notably a sharp phase transition regarding the existence of satisfying solutions. In this paper, we propose two novel generalizations of the SBP by incorporating random labels. Our proposals admit a natural machine learning interpretation: any satisfying solution to the random CSP is a minimizer of a certain empirical risk. We establish that the expected number of solutions for both models undergoes a sharp phase transition and calculate the location of this transition, which corresponds to the annealed capacity in statistical physics. We then establish a universality result: the location of this transition does not depend on the underlying distribution. We conjecture that both models in fact exhibit an even stronger phase transition akin to the SBP and give rigorous evidence towards this conjecture through the second moment method.**

I. Introduction

The focus of this paper is on the *perceptron* model, a natural model in high-dimensional probability and a toy shallow neural network, which stores random patterns [1, 2, 3]. Given patterns $X_i \in \mathbb{R}^n$, $1 \le i \le M$, *storage* corresponds to finding a vector $\boldsymbol{\sigma} \in \mathbb{R}^n$ of *synaptic weights* consistent with all X_i: $\langle \boldsymbol{\sigma}, X_i \rangle \ge 0$ for $1 \le i \le M$. Our focus is on the *binary* case where $\boldsymbol{\sigma} \in \Sigma_n \triangleq \{-1,1\}^n$, see [4, 5, 6, 7, 8] for the *spherical* case where $\|\boldsymbol{\sigma}\|_2 = \sqrt{n}$. Statistical physics literature provided a very detailed yet non-rigorous characterization of the *storage capacity*, i.e. the maximum number of patterns one can store via a suitable $\boldsymbol{\sigma}$—see [9, 4, 10]. More general perceptron models considered recently involve an *activation function* $U : \mathbb{R} \to \{0,1\}$. Here, an $X_i \in \mathbb{R}^n$ is stored with respect to U if $U(\langle \boldsymbol{\sigma}, X \rangle) = 1$. Our particular focus is on *symmetric binary perceptron* [11] defined by $U(x) = \mathbb{1}\{|x| \le \kappa\sqrt{n}\}$ (where $U(x) = 1$ iff $|x| \le \kappa\sqrt{n}$ and $U(x) = 0$ otherwise), see below. For even more general variants, see [12, 13].

A. Symmetric Binary Perceptron (SBP)

Fix $\kappa > 0$, $\alpha > 0$, and let $M = \lfloor n\alpha \rfloor \in \mathbb{N}$. Generate i.i.d. random vectors $X_i \sim \mathcal{N}(0, I_n)$, $1 \le i \le M$, where $\mathcal{N}(0, I_n)$ is the centered multivariate normal distribution on $\mathbb{R}^n$ with identity covariance. Consider the random set

$$S_\alpha(\kappa) \triangleq \left\{ \boldsymbol{\sigma} \in \Sigma_n : |\langle \boldsymbol{\sigma}, X_i \rangle| \le \kappa\sqrt{n}, 1 \le i \le M \right\}. \quad (1)$$

Observe that $S_\alpha(\kappa)$ is indeed *symmetric* about the origin: $\boldsymbol{\sigma} \in \Sigma_n$ iff $-\boldsymbol{\sigma} \in \Sigma_n$. Proposed by Aubin, Perkins, and Zdeborová [11], the SBP is a symmetrized analogue of the much studied *asymmetric binary perceptron* (ABP), where the constraints are instead of the form $\langle \boldsymbol{\sigma}, X_i \rangle \ge \kappa\sqrt{n}$, $1 \le i \le M$. The rigorous study of ABP is an ongoing and difficult mathematical quest, see [10, 14, 15, 16, 17, 18] for related work. On the other hand, the SBP exhibits relevant structural properties conjectured for ABP [19] (see below); at the same time, it is more amenable to analysis.

Strikingly, the SBP exhibits a certain *sharp phase transition*, conjectured in [11] and verified independently by Perkins and Xu [20] and Abbe, Li, and Sly [21]. Let $\alpha_c(\kappa) = -1/\log_2 \mathbb{P}[|\mathcal{N}(0,1)| \le \kappa]$. Then,

$$\lim_{n\to\infty} \mathbb{P}[S_\alpha(\kappa) \ne \varnothing] = \begin{cases} 0, & \text{if } \alpha > \alpha_c(\kappa) \\ 1, & \text{if } \alpha < \alpha_c(\kappa) \end{cases}. \quad (2)$$

The part $\alpha > \alpha_c(\kappa)$ is established in [11] through the *first moment method*: when $\alpha > \alpha_c(\kappa)$, $\mathbb{E}|S_\alpha(\kappa)| = o(1)$ and therefore $S_\alpha(\kappa) = \varnothing$ w.h.p. by Markov's inequality, where $|S_\alpha(\kappa)|$ is the cardinality of $S_\alpha(\kappa)$. The same paper also considers $\alpha < \alpha_c(\kappa)$ and shows that $\liminf_{n\to\infty} \mathbb{P}[S_\alpha(\kappa) \ne \varnothing] \ge \delta$ for some $0 < \delta < 1$. This is based on the *second moment method*; one requires more advanced tools for the high probability guarantee (i.e. for boosting δ to one), see [20, 21]. Furthermore, [22, 23] showed that the aforementioned phase transition is very sharp: the critical window around $\alpha_c(\kappa)$ where the probability increases quickly from $o(1)$ to $1-o(1)$ is of constant width. So, the first moment method correctly predicts the phase transition point in SBP. This is in stark contrast with the ABP as the conjectured phase transition point [10] differs substantially from the first moment prediction, see [16].

Recalling that $S_\alpha(\kappa)$ is non-empty when α is below the critical $\alpha_c(\kappa)$ threshold, a natural goal is algorithmically finding a $\boldsymbol{\sigma} \in S_\alpha(\kappa)$. The best known polynomial-time algorithm for the SBP is due to Bansal and Spencer [24] from combinatorial discrepancy literature[1], see also [18] for a different algorithm. However, both of these algorithms work at densities substantially below $\alpha_c(\kappa)$, highlighting a *statistical-to-computational gap*: for any $\kappa > 0$, there exists an $\alpha_{\rm ALG}(\kappa) \ll \alpha_c(\kappa)$ such that finding a $\boldsymbol{\sigma} \in S_\alpha(\kappa)$ is likely to be computationally intractable when $\alpha_{\rm ALG}(\kappa) < \alpha < \alpha_c(\kappa)$. Limits of efficient algorithms were recently

[1] See [25, Section 1.3] for details on the connection between the SBP and combinatorial discrepancy.

explored in [26, 25] and tight lower bounds against stable and online algorithms were obtained. For a more elaborate discussion on SBP, see [20, 26, 25].

Notation. Given any $p \in [0,1]$, $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p. For any $M \in \mathbb{N}$, $[M]$ denotes the set $\{1,\dots,M\}$. For any proposition E, $\mathbb{1}\{E\} \in \{0,1\}$ denotes its indicator. Given a set S, $|S|$ denotes its cardinality. For any $\mathbf{\Sigma}$, $\mathcal{N}(0,\mathbf{\Sigma})$ denotes the centered multivariate normal distribution with covariance $\mathbf{\Sigma}$; the cases $\mathbf{\Sigma} = I_n$ (the identity matrix in $\mathbb{R}^n$) and $\mathbf{\Sigma} = \sigma^2$ ($\sigma \in \mathbb{R}^+$) are of particular relevance. For any $r > 0$, $\log_r(\cdot)$ and $\exp_r(\cdot)$ respectively denote the logarithm and the exponential functions base r; we omit the subscript when $r = e$. We omit all floor/ceiling operators. We use the standard asymptotic notation, e.g. $\Theta(\cdot), O(\cdot), o(\cdot), \omega(\cdot)$, where the underlying asymptotics are with respect to $n \to \infty$.

II. Models and Main Results

In this section, we propose two novel generalizations of the SBP by incorporating random labels.

Definition II.1. *Fix $\kappa > 0$, $\alpha > 0$, $p \in [0,1]$, and set $M = n\alpha \in \mathbb{N}$. Let $X_i \sim \mathcal{N}(0, I_n)$, $1 \le i \le M$ be i.i.d. random vectors and $U(x) = \mathbb{1}\{|x| \le \kappa\sqrt{n}\}$ be the activation.*

- *Let $Y_i \sim \mathrm{Ber}(p)$, $1 \le i \le M$ be i.i.d. Set*
$$S_\alpha(\kappa,p) = \big\{\boldsymbol{\sigma} \in \Sigma_n : Y_i = U(\langle \boldsymbol{\sigma}, X_i\rangle), \forall i \in [M]\big\}.$$
- *Draw a $\mathcal{I} \subset \{1,2,\dots,M\}$ with $|\mathcal{I}| = Mp$ uniformly at random and let $Y_i = \mathbb{1}\{i \in \mathcal{I}\}$, $1 \le i \le M$. Set*
$$\widetilde{S}_\alpha(\kappa,p) = \big\{\boldsymbol{\sigma} \in \Sigma_n : Y_i = U(\langle \boldsymbol{\sigma}, X_i\rangle), \forall i \in [M]\big\}.$$

Several remarks are in order. Note that the SBP is indeed a special case of the models arising in Definition II.1, corresponding to the extreme case of $p = 1$. Furthermore, our model also captures the activation $\mathbb{1}\{|x| > \kappa\sqrt{n}\}$ by considering the labels $Y_i' = 1 - Y_i$ instead (equivalently replacing p by $1-p$). This is dubbed the u-function binary perceptron (UBP), see [11] for details. We now highlight some fundamental differences between our models and both the SBP and the UBP. Note that for the SBP (resp. UBP), the solution space gets larger (resp. smaller) as $\kappa \to \infty$ and smaller (resp. larger) as $\kappa \to 0$. Importantly though, for $p \in (0,1)$, the sets $S_\alpha(\kappa,p)$ and $\widetilde{S}_\alpha(\kappa,p)$ shrink both as $\kappa \to 0$ as well as $\kappa \to \infty$.

We next compare the two models. On the one hand, they are somewhat similar: if $Y_i \sim \mathrm{Ber}(p)$, $1 \le i \le M$, are i.i.d., then $|\{i : Y_i = 1\}| = Mp + O(\sqrt{M})$ w.h.p. due to concentration of measure. On the other hand, the labels are not independent under the second model. Indeed, while $\mathbb{P}[i \in \mathcal{I}] = p$ for any $i \in [M]$, we have that for any $j \neq i$,
$$\mathbb{P}[j \in \mathcal{I} | i \in \mathcal{I}] = \frac{\binom{M-1}{Mp-1}}{\binom{M}{Mp}} = \frac{Mp-1}{M-1} < p = \mathbb{P}[j \in \mathcal{I}],$$
provided $p < 1$. In the next section, we show that breaking the independence in fact lowers the *critical threshold*.

We now provide two interpretations of our models.

a) Random CSP Interpretation: Both the SBP and its generalizations in Definition II.1 can be viewed as a random constraint satisfaction problem (CSP): each pair (X_i, Y_i) defines a random constraint $Y_i = \mathbb{1}\{|\langle \boldsymbol{\sigma}, X_i\rangle| \le \kappa\sqrt{n}\}$ and any $\boldsymbol{\sigma} \in S_\alpha(\kappa,p)$ is a satisfying solution to the induced CSP. Random CSPs have been thoroughly studied through various angles, ranging from the existence of solutions to the solution space geometry and the limits of polynomial-time algorithms, see [20] for pointers to relevant literature.

b) Machine Learning Interpretation: Given data consisting of feature/label pairs $(X_i, Y_i) \in \mathbb{R}^n \times \{0,1\}$, $1 \le i \le M$, a canonical task in machine learning is to find a model $f(\cdot, \boldsymbol{\sigma})$, $\boldsymbol{\sigma} \in \theta$ 'accurately explaining' these data, where θ is some domain. This often entails solving the empirical risk minimization (ERM) problem:
$$\min_{\boldsymbol{\sigma}\in\theta} \widehat{\mathcal{L}}(\boldsymbol{\sigma}), \text{ where } \widehat{\mathcal{L}}(\boldsymbol{\sigma}) = \frac{1}{M}\sum_{1\le i\le M} \ell\big(Y_i; f(X_i,\boldsymbol{\sigma})\big).$$
Here, $\ell : \mathbb{R}^2 \to \mathbb{R}_{\ge 0}$ is a loss function. Note that when $\theta = \Sigma_n$, $\ell(y;x) = \mathbb{1}\{y \neq x\}$ and $f(X_i,\boldsymbol{\sigma}) = U(\langle \boldsymbol{\sigma}, X_i\rangle)$, $S_\alpha(\kappa,p)$ is simply the set of *interpolators*:
$$S_\alpha(\kappa,p) = \big\{\boldsymbol{\sigma} \in \Sigma_n : \widehat{\mathcal{L}}(\boldsymbol{\sigma}) = 0\big\}.$$
The case of random labels as we do here is important both from an optimization viewpoint and as a theoretical toy model in statistics. Closely related to this is the negative spherical perceptron with random labels, where $\|\boldsymbol{\sigma}\|_2 = 1$ and the constraints are of the form $Y_i\langle \boldsymbol{\sigma}, X_i\rangle \ge \kappa$ (note that since $\|\boldsymbol{\sigma}\|_2 = 1$, the right hand side scales as κ instead of $\kappa\sqrt{n}$). See Montanari et al. [27] for a thorough study of this model, including a rigorous phase transition and the analysis of a certain linear program.

Annealed and Quenched Free Energies. In the next section, we apply the *first moment method* to show that the expected size of $S_\alpha(\kappa,p)$ (resp. $\widetilde{S}_\alpha(\kappa,p)$) undergoes a phase transition as α crosses an explicit threshold $\alpha_c(\kappa,p)$ (resp. $\widetilde{\alpha}_c(\kappa,p)$). More precisely, we show that for $S_\alpha(\kappa,p)$,
$$\lim_{n\to\infty} \frac{\log \mathbb{E}\big[|S_\alpha(\kappa,p)|\big]}{n} > 0, \quad \forall \alpha < \alpha_c(\kappa,p) \quad (3)$$
$$\lim_{n\to\infty} \frac{\log \mathbb{E}\big[|S_\alpha(\kappa,p)|\big]}{n} < 0, \quad \forall \alpha > \alpha_c(\kappa,p), \quad (4)$$
and analogously for $\widetilde{S}_\alpha(\kappa,p)$. This result pertains $n^{-1}\log\mathbb{E}\big[|S_\alpha(\kappa,p)|\big]$, which is known as the *annealed free energy* in statistical physics literature, see e.g. [28, 12]. This should be contrasted with the *quenched free energy*, $n^{-1}\mathbb{E}\big[\log|S_\alpha(\kappa,p)|\big]$ (which is upper bounded by the annealed free energy via Jensen's inequality). An ultimate goal towards which we give some rigorous evidence in Theorem II.7 is to show that (a) $S_\alpha(\kappa,p) \neq \varnothing$ (w.h.p.) if $\alpha < \alpha_c(\kappa)$ and (b) $S_\alpha(\kappa,p) = \varnothing$ (w.h.p.) if $\alpha > \alpha_c(\kappa)$. Note that when $\alpha > \alpha_c(\kappa,p)$, (4) yields $S_\alpha(\kappa,p) = \varnothing$ (w.h.p.) via Markov's inequality, see Theorems II.2-II.3 for details. However for $\alpha < \alpha_c(\kappa,p)$, (3) does not necessarily

imply $S_\alpha(\kappa,p)\neq\varnothing$: it is possible that $\mathbb{E}[|S_\alpha(\kappa,p)|]$ is large, while $|S_\alpha(\kappa,p)|$ is in fact zero w.h.p. To establish $S_\alpha(\kappa,p)\neq\varnothing$ for $\alpha<\alpha_c(\kappa,p)$, it might help studying the quenched free energy instead, e.g. if $n^{-1}\log|S_\alpha(\kappa,p)|$ concentrates around its mean. For the SBP this was done in [20]. For our models, this is left for future work. For more on the annealed and quenched energies, see [28, 11]. In light of the preceding discussion, the quantities $\alpha_c(\kappa,p)$ and $\widetilde{\alpha}_c(\kappa,p)$ are dubbed as the *annealed capacity*.

A. Main Results

Throughout this section, $q(\kappa)$ denotes $\mathbb{P}[|\mathcal{N}(0,1)|\le\kappa]$. Our first main result addresses the case of i.i.d. labels.

Theorem II.2. *Recall $S_\alpha(\kappa,p)$ from Definition II.1 and let*

$$\alpha_c(\kappa,p)=-1/\log_2\bigl(pq(\kappa)+(1-p)(1-q(\kappa))\bigr). \quad (5)$$

Then

$$\mathbb{E}\bigl[|S_\alpha(\kappa,p)|\bigr]=\begin{cases}\exp(-\Theta(n)), & \text{if } \alpha>\alpha_c(\kappa,p)\\ \exp(\Theta(n)), & \text{if } \alpha<\alpha_c(\kappa,p)\end{cases}.$$

In particular, $\mathbb{P}[S_\alpha(\kappa,p)=\varnothing]\ge 1-e^{-\Theta(n)}$ if $\alpha>\alpha_c(\kappa)$.

Proof of Theorem II.2. Our proof is based on the *first moment method*: note that by Markov's inequality,

$$\mathbb{P}\bigl[|S_\alpha(\kappa,p)|\ge 1\bigr]\le\mathbb{E}\bigl[|S_\alpha(\kappa,p)|\bigr],$$

so that $S_\alpha(\kappa,p)=\varnothing$ w.h.p. if $\mathbb{E}\bigl[|S_\alpha(\kappa,p)|\bigr]=o(1)$. So, the remainder of proof estimates $\mathbb{E}\bigl[|S_\alpha(\kappa,p)|\bigr]$. Fix any $\boldsymbol{\sigma}\in\Sigma_n$ and let $Z_i(\boldsymbol{\sigma})=\mathbb{1}\bigl\{Y_i=U(\langle\boldsymbol{\sigma},X_i\rangle)\bigr\}$. Then,

$$|S_\alpha(\kappa,p)|=\sum_{\boldsymbol{\sigma}\in\Sigma_n}Z(\boldsymbol{\sigma}),\quad\text{where}\quad Z(\boldsymbol{\sigma})=\prod_{1\le i\le M}Z_i(\boldsymbol{\sigma}).$$

Now fix any $\boldsymbol{\sigma}\in\Sigma_n$ and observe that $Z_1(\boldsymbol{\sigma}),\dots,Z_M(\boldsymbol{\sigma})$ are i.i.d. Bernoulli. Moreover, $\langle\boldsymbol{\sigma},X_i\rangle\sim\mathcal{N}(0,n)$. So,

$$\begin{aligned}\mathbb{P}[Z_i(\boldsymbol{\sigma})=1]&=\mathbb{P}[Z_i(\boldsymbol{\sigma})=1|Y_i=1]\mathbb{P}[Y_i=1]\\&+\mathbb{P}[Z_i(\boldsymbol{\sigma})=1|Y_i=0]\mathbb{P}[Y_i=0]\\&=p\mathbb{P}[|\langle\boldsymbol{\sigma},X_i\rangle|\le\kappa\sqrt{n}]\\&+(1-p)\mathbb{P}[|\langle\boldsymbol{\sigma},X_i\rangle|>\kappa\sqrt{n}]\\&=pq(\kappa)+(1-p)(1-q(\kappa)).\end{aligned}$$

Thus, $\mathbb{E}\bigl[|S_\alpha(\kappa,p)|\bigr]=\exp_2\bigl(nf(\alpha,p,\kappa)\bigr)$ where

$$f(\alpha,p,\kappa)=1+\alpha\log_2\bigl(pq(\kappa)+(1-p)(1-q(\kappa))\bigr).$$

As $f(\alpha,p,\kappa)>0$ iff $\alpha<\alpha_c(\kappa)$ the proof is complete. □

Our second main result addresses the case where the set $\{i:Y_i=1\}$ is drawn uniformly at random.

Theorem II.3. *Recall $\widetilde{S}_\alpha(\kappa,p)$ from Definition II.1 and let*

$$\widetilde{\alpha}_c(\kappa,p)=-1/\bigl(p\log_2 q(\kappa)+(1-p)\log_2(1-q(\kappa))\bigr). \quad (6)$$

Then,

$$\mathbb{E}\bigl[|\widetilde{S}_\alpha(\kappa,p)|\bigr]=\begin{cases}\exp(-\Theta(n)), & \text{if } \alpha>\widetilde{\alpha}_c(\kappa,p)\\ \exp(\Theta(n)), & \text{if } \alpha<\widetilde{\alpha}_c(\kappa,p)\end{cases}.$$

In particular, $\mathbb{P}[\widetilde{S}_\alpha(\kappa,p)=\varnothing]\ge 1-e^{-\Theta(n)}$ if $\alpha>\widetilde{\alpha}_c(\kappa,p)$.

Proof of Theorem II.3. The proof is quite similar to that of Theorem II.2; we only point out necessary modifications. Define $\widetilde{Z}(\boldsymbol{\sigma})=\prod_{1\le i\le M}\widetilde{Z}_i(\boldsymbol{\sigma})$, where $\widetilde{Z}_i(\boldsymbol{\sigma})=\mathbb{1}\{Y_i=U(\langle\boldsymbol{\sigma},X_i\rangle)\}$ for $1\le i\le M$. Let $\mathcal{I}_t$, $1\le t\le\binom{M}{Mp}$ be the subsets of $[M]$ of size Mp. Notice that $\mathbb{P}[\widetilde{Z}(\boldsymbol{\sigma})=1|\mathcal{I}=\mathcal{I}_t]=\prod_{i\in\mathcal{I}_t}\mathbb{P}[|\langle\boldsymbol{\sigma},X_i\rangle|\le\kappa\sqrt{n}]\cdot\prod_{i\in[M]\setminus\mathcal{I}_t}\mathbb{P}[|\langle\boldsymbol{\sigma},X_i\rangle|>\kappa\sqrt{n}]=q(\kappa)^{Mp}(1-q(\kappa))^{M(1-p)}$, using the fact $\langle\boldsymbol{\sigma},X_i\rangle\sim\mathcal{N}(0,n)$ and the independence of $X_1,\dots,X_M$. Hence,

$$\begin{aligned}\mathbb{P}[\widetilde{Z}(\boldsymbol{\sigma})=1]&=\sum_{t=1}^{\binom{M}{Mp}}\binom{M}{Mp}^{-1}\mathbb{P}[\widetilde{Z}(\boldsymbol{\sigma})=1|\mathcal{I}=\mathcal{I}_t]\\&=q(\kappa)^{Mp}(1-q(\kappa))^{M(1-p)}.\end{aligned}$$

As $M=\alpha n$, we immediately obtain $\mathbb{E}\bigl[|\widetilde{S}_\alpha(\kappa,p)|\bigr]=\exp_2\bigl(n\tilde{f}(\alpha,p,\kappa)\bigr)$, where

$$\tilde{f}(\alpha,p,\kappa)=1+\alpha\bigl(p\log_2 q(\kappa)+(1-p)\log_2(1-q(\kappa))\bigr).$$

This yields Theorem II.3. □

a) Universality: We next result establish a *universality* result: under mild assumptions, the quantities $\alpha_c(\kappa,p)$ and $\widetilde{\alpha}_c(\kappa,p)$ do not depend on the distribution of X_i.

Theorem II.4. *Theorems II.2-II.3 still hold if $X_i=(X_i(j):j\in[n])\in\mathbb{R}^n$ consists of i.i.d. coordinates with $\mathbb{E}[X_i(1)]=0$, $\mathbb{E}[X_i(1)^2]>0$ and $\mathbb{E}[|X_i(1)|^3]<\infty$.*

We note that several related universality results appeared in the literature. In particular, [26] establishes the universality of a certain intricate geometrical property in the solution space of the SBP and [29] establishes the universality of the training error for linear classification with random inputs.

Proof of Theorem II.5. We show the extension for Theorem II.2; that of Theorem II.3 is analogous. Our argument is based on the Berry-Esseen inequality [30, 31], reproduced below for convenience.

Theorem II.5. *There exists an absolute constant $C>0$ such that the following holds. Let $T_1,\dots,T_n$ be i.i.d. random variables with $\mathbb{E}[T_1]=0$, $\mathbb{E}[T_1^2]=\sigma^2>0$ and $\mathbb{E}[|T_1|^3]=\rho<\infty$. Then, for $Z\sim\mathcal{N}(0,1)$,*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left[\frac{T_1+\cdots+T_n}{\sigma\sqrt{n}}\le x\right]-\mathbb{P}[Z\le x]\right|\le\frac{C\rho}{\sigma^3\sqrt{n}}.$$

Equipped with Theorem II.5, fix any $i\in[M]$ and let $X_i=(X_i(j):j\in[n])$ with $\mathbb{E}[X_i(j)]=0$, $\mathbb{E}[X_i(j)^2]=\sigma^2$ (where $\sigma>0$) and $\mathbb{E}[|X_i(j)|^3]=\rho<\infty$. Note that

$$\begin{aligned}&\mathbb{P}[Y_i=U(\langle\boldsymbol{\sigma},X_i\rangle)]\\&=\mathbb{P}[Y_i=U(\langle\boldsymbol{\sigma},X_i\rangle)|Y_i=1]\mathbb{P}[Y_i=1]\\&+\mathbb{P}[Y_i=U(\langle\boldsymbol{\sigma},X_i\rangle)|Y_i=0]\mathbb{P}[Y_i=0]\\&=\mathbb{P}\bigl[|\langle\boldsymbol{\sigma},X_i\rangle|\le\kappa\sqrt{n}\bigr]p+\mathbb{P}\bigl[|\langle\boldsymbol{\sigma},X_i\rangle|>\kappa\sqrt{n}\bigr](1-p).\end{aligned} \quad (7)$$

Let $q(\kappa) = \mathbb{P}[-\kappa \le Z \le \kappa]$ where $Z \sim \mathcal{N}(0,1)$. Applying Theorem II.5 to $X_i(j)$, $j \in [n]$, together with triangle inequality, we obtain

$$\left|\mathbb{P}\left[|\langle \boldsymbol{\sigma}, X_i\rangle| \le \kappa\sqrt{n}\right] - q(\kappa)\right| \le \frac{2C\rho}{\sigma^3\sqrt{n}} \triangleq \frac{\mathcal{C}}{\sqrt{n}}, \quad (8)$$

where $\mathcal{C} = \frac{2C\rho}{\sigma^3} = O(1)$. Combining (7) and (8), we obtain

$$\mathbb{P}[Y_i = U(\langle \boldsymbol{\sigma}, X_i\rangle)] \le q(\kappa)p + (1-q(\kappa))(1-p) + \frac{\mathcal{C}}{\sqrt{n}}. \quad (9)$$

Recall now the Taylor expansion for logarithm: as $x \to 0$,

$$\log_2(1+x) = -\frac{x}{\log 2} + O(x^2). \quad (10)$$

We now combine (9) with (10) to obtain

$$\begin{aligned}
&\mathbb{P}[Y_i = U(\langle \boldsymbol{\sigma}, X_i\rangle)]^{\alpha n} \\
&\le \big(q(\kappa)p + (1-q(\kappa))(1-p)\big)^{\alpha n} \\
&\times \left(1 + \frac{\mathcal{C}}{\sqrt{n}\big(q(\kappa)p + (1-q(\kappa))(1-p)\big)}\right)^{\alpha n} \\
&= \exp_2\Big(\alpha n \log_2\big(q(\kappa)p + (1-q(\kappa))(1-p)\big) + \\
&\alpha n \log_2\left(1 + \frac{\mathcal{C}}{\sqrt{n}\big(q(\kappa)p + (1-q(\kappa))(1-p)\big)}\right)\Big) \\
&= \exp_2\Big(\alpha n \log_2\big(q(\kappa)p + (1-q(\kappa))(1-p)\big) + \Theta(\sqrt{n})\Big).
\end{aligned}$$

With this, we obtain immediately that

$$\mathbb{E}\big|S_\alpha(\kappa,p)\big| = \exp_2\big(nf(\alpha,p,\kappa) + \Theta(\sqrt{n})\big).$$

The extension for Theorem II.3 is similar. □

b) Comparison of Thresholds in Theorems II.2-II.3: Inspecting (5) and (6), observe that Jensen's inequality and the concavity of the map $x \mapsto \log_2 x$ on $(0,\infty)$ collectively yield $\alpha_c(\kappa,p) \ge \widetilde{\alpha}_c(\kappa,p)$. We found it quite remarkable that breaking the independence lowers the critical threshold: the model with independent labels has a higher annealed capacity. We are unaware of any prior work in the random CSP literature that investigates whether and how the critical threshold changes with the dependence structure. We believe that this direction merits further investigation.

c) A Sharp Phase Transition Conjecture: Recall that the prior works [20, 21] establish a sharp phase transition (2) for the SBP, and show that the first moment method correctly predicts the location of this transition. Further, Theorems II.2-II.3 collectively yield a phase transition for the first moment itself. In light of these, we conjecture an analogous phase transition for the models we propose.

Conjecture II.6. *There exists a $\kappa^* > 0$ such that the following holds for every $\kappa < \kappa^*$. The quantity $\mathbb{P}\big[S_\alpha(\kappa,p) \ne \varnothing\big]$ (resp. $\mathbb{P}\big[\widetilde{S}_\alpha(\kappa,p) \ne \varnothing\big]$) undergoes a phase transition at value $\alpha_c(\kappa,p)$ (resp. $\widetilde{\alpha}_c(\kappa,p)$) as $n \to \infty$:*

$$\lim_{n\to\infty} \mathbb{P}[S_\alpha(\kappa,p) \ne \varnothing] = \begin{cases} 0, & \textit{if } \alpha > \alpha_c(\kappa,p) \\ 1, & \textit{if } \alpha < \alpha_c(\kappa,p), \end{cases}$$

$$\lim_{n\to\infty} \mathbb{P}[\widetilde{S}_\alpha(\kappa,p) \ne \varnothing] = \begin{cases} 0, & \textit{if } \alpha > \widetilde{\alpha}_c(\kappa,p) \\ 1, & \textit{if } \alpha < \widetilde{\alpha}_c(\kappa,p). \end{cases}$$

For the UBP (corresponding to $p = 0$), [11] shows that the moment method works only for $\kappa < \kappa^* \approx 0.817$. Remarkably, the value 0.817 corresponds to the onset of replica symmetry breaking, see [11] for details. In light of this, we anticipate Conjecture II.6 to be valid for small κ, more concretely for $\kappa < \kappa^* \approx 0.817$. The behaviour of our models beyond κ^* is a very interesting open question.

Towards Conjecture II.6, we establish the following result, contingent on a certain assumption.

Theorem II.7. *For any $\kappa > 0$, there exists a $p_\kappa^* < 1$ such that the following holds. Fix any $p \in [p_\kappa^*, 1]$ and any $\alpha < \widetilde{\alpha}_c(\kappa,p)$. Then, $\liminf_{n\to\infty} \mathbb{P}\big[\widetilde{S}_\alpha(\kappa,p) \ne \varnothing\big] > 0$.*

Moreover, for any $\kappa \in (0, 0.817)$, there exists a $p_\kappa^{**} > 0$ such that the following holds. Fix any $p \in [0, p_\kappa^{**}]$ and any $\alpha < \widetilde{\alpha}_c(\kappa,p)$. Then, $\liminf_{n\to\infty} \mathbb{P}\big[\widetilde{S}_\alpha(\kappa,p) \ne \varnothing\big] > 0$.

We highlight that our proof is contingent on an assumption regarding (the critical points of) a certain real function, akin to [11, Hypothesis 3]. See the supplementary material for details. Theorem II.7 covers the cases when p is close to 1 (corresponding to SBP) and close to 0 (corresponding to UBP). We prove Theorem II.7 by adapting the *second moment* argument of [11] with few extra steps.

III. Open Problems

a) Sharp Phase Transition: In light of earlier discussion, we conjecture that both models exhibit a sharp phase transition (Conjecture II.6). It is plausible that Conjecture II.6 can be resolved by employing an argument similar to [20, 21]; we leave this as an open problem.

b) Interplay between the Critical Threshold and Dependence Structure: Recall that $\alpha_c(\kappa,p) \ge \widetilde{\alpha}_c(\kappa,p)$ for any $\kappa > 0$ and $p \in [0,1]$. The interplay between the critical threshold and the dependence structure in the context of other random CSPs or neural network models (such as the Hopfield model) is an interesting question for future work.

c) Other Perceptron Models: It would be very interesting to extend our results to the spherical case ($\|\boldsymbol{\sigma}\|_2 = 1$). We believe that the arguments of [27] may transfer. Similarly, it would be interesting to consider different activations $U(x)$ and more general perceptron models [12, 13].

d) Algorithms: While [24] and [18] devise efficient algorithms for finding solutions of the SBP and the UBP at sufficiently low densities, it is not clear whether they apply to our models. Let $\mathcal{I} = \{i : Y_i = 1\}$, $\mathcal{M} \in \mathbb{R}^{|\mathcal{I}|\times n}$ with rows $X_i \in \mathbb{R}^n$, $i \in \mathcal{I}$, and $\overline{\mathcal{M}} \in \mathbb{R}^{(M-|\mathcal{I}|)\times n}$ with rows $X_i \in \mathbb{R}^n$, $i \in [M] \setminus \mathcal{I}$. Note that when $0 < p < 1$ holds strictly, both $\mathcal{I}$ and $\mathcal{I}^c$ are w.h.p. non-empty. Observe that finding a $\boldsymbol{\sigma} \in S_\alpha(\kappa,p)$ (or a $\boldsymbol{\sigma} \in \widetilde{S}_\alpha(\kappa,p)$) amounts to finding a $\boldsymbol{\sigma}$ such that both $\|\mathcal{M}\boldsymbol{\sigma}\|_\infty \le \kappa\sqrt{n}$ and $\min_i\big|\big(\overline{\mathcal{M}}\boldsymbol{\sigma}\big)_i\big| > \kappa\sqrt{n}$ hold. To that end, one can potentially run (a) the discrepancy minimization algorithm to find a $\boldsymbol{\sigma}_1 \in \Sigma_n$ with $\|\mathcal{M}\boldsymbol{\sigma}\|_\infty \le \kappa\sqrt{n}$ and (b) the algorithm of Abbe, Li, and Sly [18] to find a $\boldsymbol{\sigma}_2 \in \Sigma_n$ with $\min_i \left|\left(\mathcal{M}\boldsymbol{\sigma}\right)_i\right| > \kappa\sqrt{n}$. It is, however, unclear if these algorithms return the same solution (i.e. $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2$) even at very low densities. Assuming that solutions do exist for densities below the critical threshold, it is thus a very interesting open question to find efficient algorithms finding these solutions at certain densities.

e) Solution Space Geometry: A large body of literature on random CSPs is devoted to the study of their solution space geometry [20]. Intricate geometrical properties of their solution spaces are linked to the failure of algorithms, see [32, 33, 20, 26, 25] for a discussion. Gamarnik, Kızıldağ, Perkins and Xu [26] studied the solution space geometry of the SBP and established the presence of the multi Overlap Gap Property (m-OGP) in order to obtain nearly tight lower bounds against the class of stable algorithms. More recently, the same authors established a different intricate geometrical property and leveraged it to obtain tight hardness guarantees against online algorithms, see [25]. The class of online algorithms captures, in particular, the best known algorithm for the SBP [24]. It would be very interesting to study the solution space geometry of these models via the m-OGP to obtain algorithmic lower bounds. We anticipate that the fact $S_\alpha(\kappa, p)$ shrinks both as $\kappa \to 0$ and as $\kappa \to \infty$ may simplify the analysis. For more on the OGP, see the survey by Gamarnik [34].

Acknowledgments. We thank anonymous reviewers for their helpful feedback. The first author is supported by a Columbia University, Distinguished Postdoctoral Fellowship in Statistics. The second author is supported by a National Science Foundation Graduate Research Fellowship.

REFERENCES

[1] R. O. Winder, "Single stage threshold logic," in *2nd Annual Symposium on Switching Circuit Theory and Logical Design (SWCT 1961)*. IEEE, 1961, pp. 321–332.

[2] J. G. Wendel, "A problem in geometric probability," *Mathematica Scandinavica*, vol. 11, no. 1, pp. 109–111, 1962.

[3] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE transactions on electronic computers*, no. 3, pp. 326–334, 1965.

[4] E. Gardner, "The space of interactions in neural network models," *Journal of physics A: Mathematical and general*, vol. 21, no. 1, p. 257, 1988.

[5] M. Shcherbina and B. Tirozzi, "Rigorous solution of the Gardner problem," *Communications in mathematical physics*, vol. 234, no. 3, pp. 383–422, 2003.

[6] M. Stojnic, "Another look at the Gardner problem," *arXiv preprint arXiv:1306.3979*, 2013.

[7] M. Talagrand, *Mean Field Models for Spin Glasses: Advanced replica-symmetry and low temperature*. Springer, 2011.

[8] A. E. Alaoui and M. Sellke, "Algorithmic pure states for the negative spherical perceptron," *arXiv preprint arXiv:2010.15811*, 2020.

[9] E. Gardner and B. Derrida, "Optimal storage properties of neural network models," *Journal of Physics A: Mathematical and general*, vol. 21, no. 1, p. 271, 1988.

[10] W. Krauth and M. Mézard, "Storage capacity of memory networks with binary couplings," *Journal de Physique*, vol. 50, no. 20, pp. 3057–3066, 1989.

[11] B. Aubin, W. Perkins, and L. Zdeborová, "Storage capacity in symmetric binary perceptrons," *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 29, p. 294003, 2019.

[12] E. Bolthausen, S. Nakajima, N. Sun, and C. Xu, "Gardner formula for ising perceptron models at small densities," in *Conference on Learning Theory*. PMLR, 2022, pp. 1787–1911.

[13] S. Nakajima and N. Sun, "Sharp threshold sequence and universality for ising perceptron models," in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 2023, pp. 638–674.

[14] J. H. Kim and J. R. Roche, "Covering cubes by random half cubes, with applications to binary neural networks," *Journal of Computer and System Sciences*, vol. 56, no. 2, pp. 223–252, 1998.

[15] M. Talagrand, "Intersecting random half cubes," *Random Structures & Algorithms*, vol. 15, no. 3-4, pp. 436–449, 1999.

[16] J. Ding and N. Sun, "Capacity lower bound for the Ising perceptron," in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019, pp. 816–827.

[17] C. Xu, "Sharp threshold for the ising perceptron model," *The Annals of Probability*, vol. 49, no. 5, pp. 2399–2415, 2021.

[18] E. Abbe, S. Li, and A. Sly, "Binary perceptron: efficient algorithms can find solutions in a rare well-connected cluster," *arXiv preprint arXiv:2111.03084*, 2021.

[19] C. Baldassi, R. Della Vecchia, C. Lucibello, and R. Zecchina, "Clustering of solutions in the symmetric binary perceptron," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2020, no. 7, p. 073303, 2020.

[20] W. Perkins and C. Xu, "Frozen 1-RSB structure of the symmetric Ising perceptron," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 1579–1588.

[21] E. Abbe, S. Li, and A. Sly, "Proof of the contiguity conjecture and lognormal limit for the symmetric perceptron," *arXiv preprint arXiv:2102.13069*, 2021.

[22] A. Sah and M. Sawhney, "Distribution of the threshold for the symmetric perceptron," *arXiv preprint arXiv:2301.10701*, 2023.

[23] D. J. Altschuler, "Fluctuations of the symmetric perceptron," *arXiv preprint arXiv:2205.02319*, 2022.

[24] N. Bansal and J. Spencer, "On-line balancing of random inputs," *Random Structures and Algorithms*, vol. 57, no. 4, pp. 879–891, Dec. 2020.

[25] D. Gamarnik, E. C. Kizildağ, W. Perkins, and C. Xu, "Geometric barriers for stable and online algorithms for discrepancy minimization," in *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2023, pp. 3231–3263.

[26] D. Gamarnik, E. C. Kızıldağ, W. Perkins, and C. Xu, "Algorithms and barriers in the symmetric binary perceptron model," in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2022, pp. 576–587.

[27] A. Montanari, Y. Zhong, and K. Zhou, "Tractability from overparametrization: The example of the negative perceptron," *arXiv preprint arXiv:2110.15824*, 2021.

[28] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.

[29] F. Gerace, F. Krzakala, B. Loureiro, L. Stephan, and L. Zdeborová, "Gaussian universality of linear classifiers with random labels in high-dimension," *arXiv preprint arXiv:2205.13303*, 2022.

[30] A. C. Berry, "The accuracy of the gaussian approximation to the sum of independent variates," *Transactions of the american mathematical society*, vol. 49, no. 1, pp. 122–136, 1941.

[31] C.-G. Esseen, "On the liapunov limit error in the theory of probability," *Ark. Mat. Astr. Fys.*, vol. 28, pp. 1–19, 1942.

[32] D. Achlioptas and F. Ricci-Tersenghi, "On the solution-space geometry of random constraint satisfaction problems," in *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, 2006, pp. 130–139.

[33] D. Achlioptas and A. Coja-Oghlan, "Algorithmic barriers from phase transitions," in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, 2008, pp. 793–802.

[34] D. Gamarnik, "The overlap gap property: A topological barrier to optimizing over random structures," *Proceedings of the National Academy of Sciences*, vol. 118, no. 41, 2021.