Entropy landscape of solutions in the binary perceptron problem

Haiping Huang[1,2], K. Y. Michael Wong[2] and Yoshiyuki Kabashima[1]

[1] Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 226-8502, Japan

[2] Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

Abstract. The statistical picture of the solution space for a binary perceptron is studied. The binary perceptron learns a random classification of input random patterns by a set of binary synaptic weights. The learning of this network is difficult especially when the pattern (constraint) density is close to the capacity, which is supposed to be intimately related to the structure of the solution space. The geometrical organization is elucidated by the entropy landscape from a reference configuration and of solution-pairs separated by a given Hamming distance in the solution space. We evaluate the entropy at the annealed level as well as replica symmetric level and the mean field result is confirmed by the numerical simulations on single instances using the proposed message passing algorithms. From the first landscape (a random configuration as a reference), we see clearly how the solution space shrinks as more constraints are added. From the second landscape of solution-pairs, we deduce the coexistence of clustering and freezing in the solution space.

PACS numbers: 89.75.Fb, 87.19.L-, 75.10.Nr

1. Introduction

Learning in a single layer of feed forward neural network with binary synapses has been studied either based on statistical mechanics analysis [1, 2, 3, 4] or in algorithmic aspects [5, 6, 7, 8, 9, 10, 11, 12]. This network can learn an extensive number $P = \alpha N$ of random patterns, where N is the number of synapses and α denotes the constraint density. The critical α (also called the capacity) separating the learnable phase from unlearnable phase is predicted to be $\alpha_s \simeq 0.833$ where the entropy vanishes [1]. A solution is defined as a configuration of synaptic weights to implement the correct classification of P random input patterns. Above α_s, no solutions can be found with high probability (converging to 1 in the thermodynamic limit). The replica symmetric solution presented in Ref. [1] has been shown to be stable up to the capacity, which is in accordance with the convexity of the solution space [3]. Note that the solutions disappear at the threshold α_s still maintaining a typical finite value of Hamming distance between them, which is quite distinct from the case in the continuous perceptron with real-valued synapses. In the continuous perceptron, this distance tends to zero when the solutions disappear at the corresponding threshold [13]. On the other hand, many local search algorithms [6, 7, 11, 12] were proposed to find solutions of the perceptron learning problem, however, the search process slows down with increasing α, and the critical α for the local search algorithm [12] decreases when the number of synapses increases. This typical behavior of the stochastic local search algorithm is conjectured to be related to the geometrical organization of the solution space [7, 4]. In order to acquire a better understanding for the failure of the local search strategy, we compute the entropy landscape both from a reference configuration and for solution-pairs with a given distance in the solution space. Both distance landscapes contain rich information about the detailed structure of the solution space and then can help us understand the observed glassy behavior of the local search algorithms. Throughout the paper, the term distance refers to the Hamming distance. The distance landscape has been well studied in random constraint satisfaction problems defined on diluted or sparse random graphs [14, 15, 16, 17].

Learning in the binary perceptron can be mapped onto a bipartite graph where variable node represents synaptic weight and function node represents the input random pattern to be learned (see figure 1 (b)). This graph is also called graphical model or factor graph [18]. The efficient message passing learning algorithm for the binary perceptronal learning problem has been derived using the cavity method and this factor graph representation [9]. In this paper, we focus on the typical property of the solution space in random ensembles of the binary perceptronal learning problem. We apply the replica trick widely used to study disordered systems [19] to compute the statistical properties in the thermodynamic limit. To confirm the mean field result computed using the replica approach, we derive the message passing equations in the cavity context which can be applied on single random instances of the current problem. In this context, we apply the decorrelation assumption as well as the central limit theorem to derive the

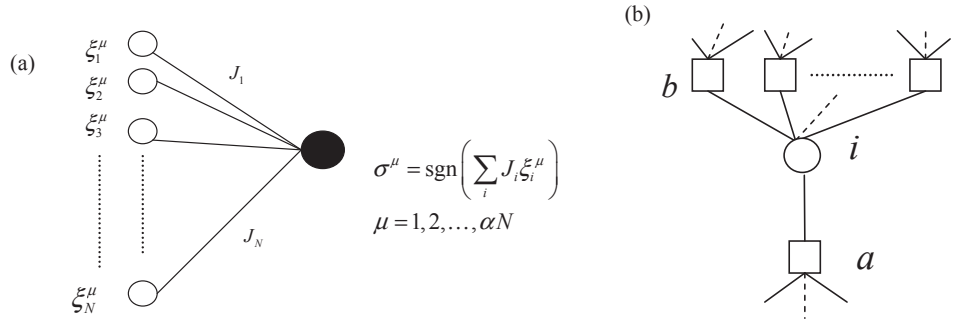

Figure 1. The sketch of the binary perceptron and the factor graph representation. (a) N input units (open circles) feed directly to a single output unit (solid circle). A binary input pattern $(\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu)$ of length N is mapped through a sign function to a binary output σ^μ, i.e., $\sigma^\mu = \mathrm{sgn}\left(\sum_{i=1}^N J_i \xi_i^\mu\right)$. The set of N binary synaptic weights $\{J_i\}$ is regarded as a solution of the perceptron problem if the output $\sigma^\mu = \sigma_0^\mu$ for each of the $P = \alpha N$ input patterns $\mu \in [1, P]$, where σ_0^μ is a preset binary value. (b) Each circle denotes the variable node whose value takes J_i with i its index. The square is the function node denoting a random binary pattern to be learned. If the pattern is learned by the synaptic vector $\boldsymbol{J}$, the value of the corresponding function node takes zero. The dotted line represents other $P-4$ function nodes while the dashed line for variable node i means i is connected to other $P-4$ function nodes and that for function node means the function node (e.g., b) is connected to other $N-3$ variable nodes.

formula at the replica symmetric level. This assumption arises from the weak correlation among synaptic weights (within one pure state [19]) and among input patterns [20]. The efficiency of the inspired message passing algorithms in loopy systems has been observed in Refs. [21, 22, 9, 23] while the underlying mechanism still needs to be fully understood. However, our cavity method focuses on the physical content and yields the same result as that obtained using replica approach [20, 24, 25].

The remainder of this paper is organized as follows. The random classification by the binary perceptron is defined in Sec. 2. In Sec. 3, we derive the self-consistent equations to compute the distance landscape (entropy landscape) from a reference configuration, i.e., to count the number of solutions at a distance from the reference configuration. Both the annealed and replica symmetric (RS) computations of this entropy landscape are presented. We also derive the message passing equations for single instances in this section using the cavity method and factor graph representation. In Sec. 4, the landscape of Hamming distances between pairs of solutions is evaluated at both annealed approximation and RS ansatz, and the associated message passing equations are proposed as well. Discussion and conclusion are given in Sec. 5.

2. Problem definition

The binary perceptron realizes a random classification of P random input patterns (see figure 1(a)). To be more precise, the learning task is to find an optimal set of binary synaptic weights (solution) $\{J_i\}_{i=1}^N$ that could map correctly each of random

input patterns $\{\xi_i^\mu\}(\mu = 1, \ldots, P)$ to the desired output σ_0^μ which is assigned a value ± 1 at random. P is proportional to N with the coefficient α defining the constraint density (each input pattern serves as a constraint for all synaptic weights, see figure 1 (b)). The critical value is $\alpha_s \simeq 0.833$ below which the solution space is non-empty [1]. Given the input pattern $\boldsymbol{\xi}^\mu$, the actual output σ^μ of the perceptron is $\sigma^\mu = \mathrm{sgn}\left(\sum_{i=1}^N J_i \xi_i^\mu\right)$ where J_i takes ± 1 and ξ_i^μ takes ± 1 with equal probabilities. If $\sigma^\mu = \sigma_0^\mu$, we say that the synaptic weight vector $\boldsymbol{J}$ has learned the μ-th pattern. Therefore we define the number of patterns mapped incorrectly as the energy cost

$$E(\boldsymbol{J}) = \sum_\mu \Theta\left(-\frac{\sigma_0^\mu}{\sqrt{N}} \sum_{i=1}^N J_i \xi_i^\mu\right) \tag{1}$$

where $\Theta(x)$ is a step function with the convention that $\Theta(x) = 0$ if $x \leq 0$ and $\Theta(x) = 1$ otherwise. The prefactor $N^{-1/2}$ is introduced to ensure that the argument of the step function remains at the order of unity, for the sake of the following statistical mechanical analysis in the thermodynamic limit. In the current setting, both $\{\xi_i^\mu\}$ and the desired output $\{\sigma_0^\mu\}$ are generated randomly independently. Without loss of generality, we assume $\sigma_0^\mu = +1$ for any input pattern in the remaining part of this paper, since one can perform a gauge transformation $\xi_i^\mu \to \xi_i^\mu \sigma_0^\mu$ to each input pattern without affecting the result.

3. Distance landscape from a reference configuration

In this section, we consider the entropy landscape from a reference configuration (which is not a solution). This entropy counts the number of solutions at a distance Nd from the reference configuration $\boldsymbol{J}^*$. The behavior of this entropy landscape reflects the geometrical organization of the solution space. Since we concentrate on the ground state ($E = 0$), we take the inverse temperature $\beta \to \infty$ and introduce a coupling field x to control the distance between solutions and the reference configuration. The partition function for this setting is

$$Z = \sum_{\boldsymbol{J}} \prod_\mu \Theta\left(\frac{1}{\sqrt{N}} \sum_i J_i \xi_i^\mu\right) \exp\left[x \sum_i J_i J_i^*\right] \tag{2}$$

where the sum $\sum_{\boldsymbol{J}}$ goes over all possible synaptic weight vectors and $\sum_i$ means the sum over all variable nodes. Under the definition of the overlap $\tilde{q} \equiv \frac{1}{N}\sum_i J_i J_i^*$, the partition function can be written as

$$Z = \sum_{\tilde{q}} \exp\left[N(s(\tilde{q}) + x\tilde{q})\right] \tag{3}$$

where $e^{Ns(\tilde{q})}$ is the number of solutions with the overlap $\tilde{q}$. In the thermodynamic limit $N \to \infty$, the saddle point analysis leads to $f(x) \equiv \frac{1}{N}\log Z = \max_{\tilde{q}}\left[s(\tilde{q}) + x\tilde{q}\right]$ where $f(x)$ is defined as the free energy density. Therefore, we can determine the entropy $s(\tilde{q})$

by a Legendre transform [16, 17]

$$s(\tilde{q}) = \min_x \left[f(x) - x\tilde{q} \right], \tag{4}$$

$$\tilde{q}(x) = \frac{df(x)}{dx} \tag{5}$$

where $\tilde{q}$ is related to d through $d = \frac{1-\tilde{q}}{2}$ and then the entropy density can be expressed as a function of the distance d which can be understood as the probability that a synaptic weight takes different values in $\boldsymbol{J}$ and $\boldsymbol{J}^*$. One recovers the total number of solutions by setting $x = 0$ in Eq. (2).

3.1. Annealed approximation for $s(d)$

We first calculate the annealed entropy density $s_{\rm ann}(d)$ which serves as the upper bound (Jensen's inequality [26]) for the true value of the entropy density. Actually, the free energy $\log Z$ should be averaged over the random input patterns. However, the annealed approximation alternatively performs the average of the partition function first and then takes the logarithmic operation as $f_{\rm ann} \equiv \frac{1}{N} \log \langle Z \rangle$ where the average is taken over the distribution of the random input patterns. This can be computed as

$$\begin{aligned} \langle Z \rangle &= \left\langle \sum_{\boldsymbol{J}} \prod_{\mu} \Theta\left(\frac{1}{\sqrt{N}} \sum_i J_i \xi_i^\mu \right) \exp\left[x \sum_i J_i J_i^* \right] \right\rangle \\ &= \int d\tilde{q} \int \frac{d\hat{q}}{2\pi \mathrm{i}/N} \exp\left[N \left(-\hat{q}\tilde{q} + x\tilde{q} - \alpha \log 2 + \log(2\cosh \hat{q}) \right) \right] \end{aligned} \tag{6}$$

where the integral representation of $\Theta(\cdot)$ is used and the conjugated counterpart $\hat{q}$ of the overlap $\tilde{q}$ is introduced as a Dirac delta function $\delta\left(\tilde{q} - \frac{1}{N} \sum_i J_i J_i^* \right)$ is inserted [27]. A saddle point analysis results in

$$f_{\rm ann} = \max_{\tilde{q}, \hat{q}} \{ -\hat{q}\tilde{q} + x\tilde{q} - \alpha \log 2 + \log(2\cosh \hat{q}) \} \tag{7}$$

where the saddle point equation reads $\hat{q} = x, \tilde{q} = \tanh \hat{q}$. Using Eq. (4) and the saddle point equation, we get the annealed entropy density

$$s_{\rm ann}(d) = -\alpha \log 2 - d \log d - (1-d) \log(1-d). \tag{8}$$

3.2. Replica symmetric computation of $s(d)$

The free energy density $f(x)$ is a self-averaging quantity whose value concentrates in probability around its expectation in the thermodynamic limit [28], and its average over the random input patterns is very difficult to compute because the logarithm appears inside the average. The replica trick bypasses this difficulty by using the identity $\log Z = \lim_{n \to 0} \frac{Z^n - 1}{n}$. Then the disorder averaged free energy density can be computed by first averaging an integer power of the partition function and then letting $n \to 0$ as

$$f = \lim_{n \to 0, N \to \infty} \frac{\log \langle Z^n \rangle}{nN}. \tag{9}$$

Although the replica method is not generally rigorous, the obtained theoretical result can be checked by numerical simulations. To compute $\langle Z^n \rangle$, we introduce n replicated synaptic weight vectors $\boldsymbol{J}^a (a = 1, \ldots, n)$ as follows.

$$\langle Z^n \rangle = \left\langle \sum_{\{\boldsymbol{J}^a\}} \prod_{a,\mu} \Theta\left(\frac{1}{\sqrt{N}} \sum_i J_i^a \xi_i^\mu \right) \exp\left[x \sum_{i,a} J_i^a J_i^* \right] \right\rangle$$

$$= \int \prod_{a<b} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi \mathrm{i}/N} \exp\left[-N \sum_{a<b} q^{ab} \hat{q}^{ab} + N\alpha \log G_0(\{q^{ab}\}) + N G_1(\{\hat{q}^{ab}\}) \right], \tag{10}$$

where G_0 and G_1 are expressed respectively as

$$G_0(\{q^{ab}\}) = \prod_a \left[\int \frac{d\lambda^a}{2\pi} \int_0^\infty dt^a \right] e^{\mathrm{i} \sum_a \lambda^a t^a - \sum_{a<b} \lambda^a \lambda^b q^{ab} - \frac{1}{2} \sum_a (\lambda^a)^2}, \tag{11}$$

$$G_1(\{\hat{q}^{ab}\}) = \log \sum_{J^a : a=1,\ldots,n} e^{\sum_{a<b} \hat{q}^{ab} J^a J^b + x \sum_a J^a J^*}, \tag{12}$$

where we have introduced the replica overlap $q^{ab} \equiv \frac{1}{N} \sum_i J_i^a J_i^b$ and its associated conjugated counterpart $\hat{q}^{ab}$. The replica symmetric ansatz assumes $q^{ab} = q, \hat{q}^{ab} = \hat{q}$ for $a \neq b$. Now using the saddle point analysis, we finally arrive at the formula of the free energy density and the corresponding saddle point equations,

$$f(x) = \frac{\hat{q}}{2}(q-1) + \alpha \int Dz \log H\left(\sqrt{\frac{q}{1-q}} z \right) + \int Dz \log\left[2\cosh(\sqrt{\hat{q}} z + x) \right], \tag{13}$$

$$q = \int Dz \tanh^2(\sqrt{\hat{q}} z + x), \tag{14}$$

$$\hat{q} = \frac{\alpha}{1-q} \int Dz \left[G\left(\sqrt{\frac{q}{1-q}} z \right) / H\left(\sqrt{\frac{q}{1-q}} z \right) \right]^2, \tag{15}$$

where $G(x) = \exp(-x^2/2)/\sqrt{2\pi}$ and $H(x) \equiv \int_x^\infty Dz$ with the Gaussian measure $Dz \equiv G(z)dz$. After the fixed point of the self-consistent equations (14) and (15) is obtained, the entropy landscape $s(d)$ is computed as

$$s(x) = f(x) - x \int Dz \tanh(\sqrt{\hat{q}} z + x). \tag{16}$$

Note that the final expression of $s(x)$ does not depend on the reference configuration and the integral in the second term of Eq. (16) is $\tilde{q}(x)$ defined in Eq. (5).

3.3. Message passing equations for single instances

In this section, we derive the message passing equations to compute the entropy landscape for single instances under the replica symmetric ansatz. To derive the self-consistent equation, we apply the cavity method [20, 9] and first define two kinds of cavity probabilities. One is the probability $p_{i\to a}^{J_i}$ that variable node i in figure 1 (b) takes value J_i in the absence of constraint a. The other is $\hat{p}_{b\to i}^{J_i}$ staying for the probability

that constraint b is satisfied (pattern $\mu = b$ is learned) if synaptic weight i takes J_i. According to the above definitions, the self-consistent equation for these two kinds of probabilities is readily obtained as

$$p_{i\to a}^{J_i} = \frac{1}{Z_{i\to a}} e^{xJ_iJ_i^*} \prod_{b\in\partial i\backslash a} \hat{p}_{b\to i}^{J_i}, \tag{17}$$

$$\hat{p}_{b\to i}^{J_i} = \sum_{\{J_j, j\in\partial b\backslash i\}} \Theta\left(\frac{1}{\sqrt{N}}\sum_j J_j\xi_j^b\right) \prod_{j\in\partial b\backslash i} p_{j\to b}^{J_j}, \tag{18}$$

where $Z_{i\to a}$ is a normalization constant, $\partial i\backslash a$ denotes the neighbors of node i except constraint a and $\partial b\backslash i$ denotes the neighbors of constraint b except variable node i. Eqs. (17) and (18) are actually the belief propagation equations [20, 9]. For the binary perceptron, directly solving the belief propagation equations is impossible. To reduce the computational complexity, we define $w_{b\to i} \equiv \frac{1}{\sqrt{N}}\sum_{j\neq i} J_j\xi_j^b$. Note that the sum involves $N-1$ independent random terms, as a result, the central limit theorem implies that $w_{b\to i}$ follows a Gaussian distribution with mean $\langle w_{b\to i}\rangle$ and variance $\sigma^2_{w_{b\to i}}$ where $\langle w_{b\to i}\rangle = \frac{1}{\sqrt{N}}\sum_{j\neq i} m_j\xi_j^b$ and $\sigma^2_{w_{b\to i}} = \frac{1}{N}\sum_{j\neq i}(1-m_j^2)$. Within the RS ansatz, the clustering property $\langle J_iJ_j\rangle - \langle J_i\rangle\langle J_j\rangle \simeq 0$ for $i\neq j$ in the thermodynamic limit is used to get the variance [9]. $m_j \equiv \langle J_j\rangle$ is the magnetization in statistical physics language. By separating the term $\frac{1}{\sqrt{N}}J_i\xi_i^b$ from the sum in the $\Theta(\cdot)$ of Eq. (18), and approximating the sum $\sum_{\{J_j, j\in\partial b\backslash i\}}$ by an integral over $w_{b\to i}$, we get finally

$$\hat{p}_{b\to i}^{J_i} = H\left(-\frac{J_i\xi_i^b + \hat{w}_{b\to i}}{\sqrt{\hat{\sigma}_{b\to i}}}\right) \tag{19}$$

where $\hat{w}_{b\to i} = \sum_{j\in\partial b\backslash i} m_{j\to b}\xi_j^b$ and $\hat{\sigma}_{b\to i} = \sum_{j\in\partial b\backslash i}(1-m_{j\to b}^2)$ in which the cavity magnetization $m_{j\to b} \equiv \tanh h_{j\to b}$. Using Eqs. (17) and (19), the cavity field $h_{j\to b}$ can be obtained in the log-likelihood representation

$$h_{j\to b} = \frac{1}{2}\log\frac{p_{j\to b}^{+1}}{p_{j\to b}^{-1}} = xJ_j^* + \sum_{a\in\partial j\backslash b} u_{a\to j}, \tag{20}$$

$$u_{a\to j} = \frac{1}{2}\log\frac{\hat{p}_{a\to j}^{+1}}{\hat{p}_{a\to j}^{-1}} = \frac{1}{2}\left[\log H\left(-\frac{\xi_j^a + \hat{w}_{a\to j}}{\sqrt{\hat{\sigma}_{a\to j}}}\right) - \log H\left(\frac{\xi_j^a - \hat{w}_{a\to j}}{\sqrt{\hat{\sigma}_{a\to j}}}\right)\right]. \tag{21}$$

Notice that the cavity bias $u_{a\to j}$ can be approximated by $\frac{\xi_j^a G\left(\frac{\hat{w}_{a\to j}}{\sqrt{\hat{\sigma}_{a\to j}}}\right)}{H\left(-\frac{\hat{w}_{a\to j}}{\sqrt{\hat{\sigma}_{a\to j}}}\right)\sqrt{\hat{\sigma}_{a\to j}}}$ in the large N limit. Eqs. (20) and (21) constitute the recursive equations to compute the free energy density in the Bethe approximation [29]

$$f(x) = \frac{1}{N}\sum_i \Delta f_i - \frac{N-1}{N}\sum_a \Delta f_a, \tag{22}$$

$$\Delta f_i = \log\left[e^{xJ_i^*}\prod_{b\in\partial i} H\left(-\frac{\xi_i^b + \hat{w}_{b\to i}}{\sqrt{\hat{\sigma}_{b\to i}}}\right) + e^{-xJ_i^*}\prod_{b\in\partial i} H\left(\frac{\xi_i^b - \hat{w}_{b\to i}}{\sqrt{\hat{\sigma}_{b\to i}}}\right)\right], \tag{23}$$

$$\Delta f_a = \log H\left(-\frac{\hat{w}_a}{\sqrt{\hat{\sigma}_a}}\right), \tag{24}$$

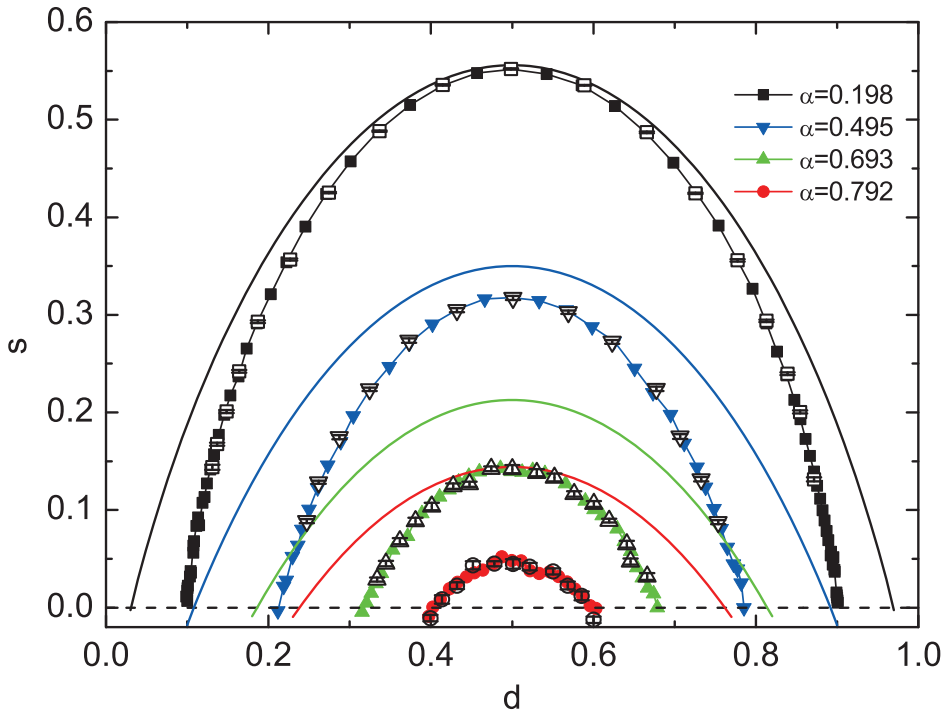

Figure 2. (Color online) Distance landscape from a reference configuration. The solid lines are the analytic annealed approximation (Eq. (8)) for $\alpha = 0.198, 0.495, 0.693, 0.792$ (from the top to the bottom) respectively. The horizontal dashed line indicates the zero entropy value. The line connecting symbols is a guide to the eye. The empty symbols stay for the numerical simulation results on systems with $(N, P) = (1001, 198), (1001, 495), (1001, 694), (1001, 793)$ (from the top to the bottom) using message passing algorithms. The result is the average over 20 random instances. Solid symbols are the replica symmetric results computed numerically by solving the saddle point equations.

where $\Delta f_i = \log Z_i$ and $\Delta f_a = \log Z_a$ are the free energy shifts due to variable node (i) addition (and all its function nodes) and function node (a) addition [16] respectively. Actually Z_i is the normalization constant of the full probability $p_i^{J_i}$ and Z_a the normalization constant of $\hat{p}_a^{\boldsymbol{J}}$ [29]. $\hat{w}_a = \sum_{j\in\partial a} m_{j\to a}\xi_j^a$ and $\hat{\sigma}_a = \sum_{j\in\partial a}(1 - m_{j\to a}^2)$. Equations (20) and (21) can be solved by an iterative procedure with a random initialization of the corresponding messages. After the iteration converges, the entropy landscape $s(d)$ from the fixed reference configuration $\boldsymbol{J}^*$ can be computed according to the Legendre transform Eqs. (4) and (5). The computational complexity is of the order $\mathcal{O}(N^2)$ for this densely connected graphical model. $f(x)$ computed based on Eq. (22) does not depend on the reference configuration since the change of $\xi_i^b \to -\xi_i^b$ does not affect the final result, consistent with Eq. (13).

The distance landscape from a reference configuration is reported in figure 2. We choose the reference configuration $\boldsymbol{J}^* = \{J_i^* = 1\}_{i=1}^N$ for simplicity. Other choices of the reference configuration still yield the same behavior of the landscape. Note that the annealed entropy provides an upper bound for the RS one, and it roughly coincides

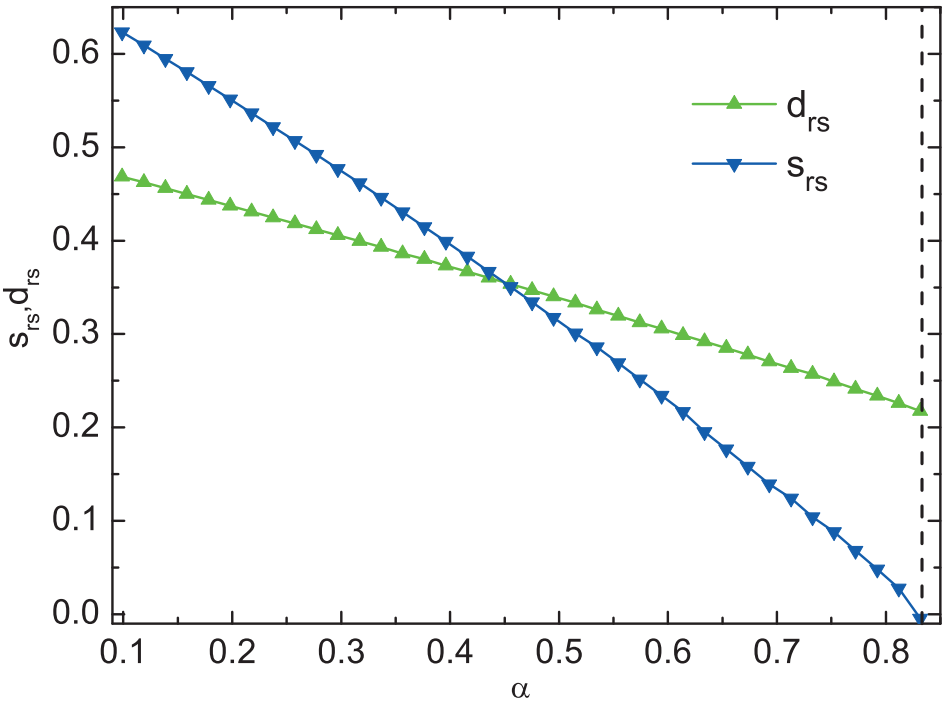

Figure 3. (Color online) Entropy density and typical distance between any two solutions as a function of constraint density. The vertical dashed line indicates the capacity for the binary perceptron.

with the RS one at low α (around the maximal point) while the large deviation is observed when α further increases. It is clear that most of the solutions concentrate around the dominant point where the maximum of the entropy is reached. When the given distance is larger or less than certain values ($d > d_{\max}$ or $d < d_{\min}$), the number of solutions at those distances becomes exponentially small in N. In the intermediate range ($d \in [d_{\min}, d_{\max}]$), as the distance increases, the number of solutions separated by the distance from the reference point in the solution space increases first and then reaches the maximum which dominates the typical value of the entropy in the original systems (by setting $x = 0$, see figure 3). The maximum is then followed by a decreasing trend as the distance is further increased. This mean field behavior is confirmed by the numerical simulations on large-size single random instances using the message passing algorithms derived in Sec. 3.3. The consistency between the mean field result obtained by replica approach and the simulation result obtained on single random instances is clearly shown in figure 2. The bell shape in figure 2 is similar to that observed in calculating the growth rate of expected distance enumerator for a random code ensemble [30]. Note that as the constraint density increases, the distance range where solutions exist shrinks, which illustrates clearly how the solution space changes as more patterns are presented to the binary perceptron.

We also compute the typical value of the entropy in the original system (by setting $x = 0$) and of the distance between any two solutions as a function of the constraint

density using the replica method. The result is reported in figure 3. Here we define the typical value of the distance between any two solutions as $d_{\rm rs} = \frac{1-q}{2}$ where q is obtained from the stationary value of Eq. (14). The entropy vanishes at $\alpha_s \simeq 0.833$ with a finite typical value of distance [1]. This typical distance is also in accordance with that computed on single instances by sampling a finite number of solutions [11]. Note that this distance is evaluated here based on the RS ansatz. One can further check its stability by the population dynamics [31] on the one-step replica symmetry breaking (1RSB) solution [29] where we define two typical distances: one is inter-cluster distance $d_0 = \frac{1-\left[\langle \tanh h\rangle^2\right]}{2}$ where $\langle\cdot\rangle$ means the average over clusters and $[\cdot]$ over the disorder, and the other is intra-cluster distance defined by $d_1 = \frac{1-\left[\left\langle \tanh^2 h\right\rangle\right]}{2}$ [32] where h is the local field defined in Eq. (20) by including all contributions of patterns around the synaptic weight ($x = 0$). In general, solutions within a single cluster are separated by a sub-extensive number of synaptic weights while any two clusters are separated by an extensive number of synaptic weights. Our numerical simulations confirmed that d_0 and d_1 will turn out to be identical (equal to $d_{\rm rs}$) after sufficient iterations implying that the RS ansatz is unbroken below the capacity. However, for constraint density close to the capacity, one needs a much larger sampling interval (by way of the Metropolis importance sampling method) [33] in the population dynamics algorithm. To probe the fine structure of the connection pattern in the solution space, we study the distance landscape of solution-pairs in the following section. This is similar to the study of the spherical p-spin model in the presence of an attractive coupling with quenched configuration in Ref. [34], however, the rich information about the solution space structure of the current problem can also be attained by calculating the distance landscape of solution-pairs.

4. Distance landscape of solution-pairs

The geometrical property of the solution space can also be studied by counting the number of solution-pairs with a predefined distance d, equivalently an overlap of value $\tilde{q}$. Actually this entropy value may be much larger than the entropy density of the original problem (which is obtained by setting $x = 0$ in Eq. (22)). As we shall present later, this case becomes more involved for the binary perceptron with an increasing computational cost.

Considering distance between solutions, we write the partition function as

$$Z = \sum_{\boldsymbol{J}^1,\boldsymbol{J}^2} \prod_{\mu} \Theta\left(\frac{1}{\sqrt{N}} \sum_i J_i^1 \xi_i^\mu\right) \Theta\left(\frac{1}{\sqrt{N}} \sum_i J_i^2 \xi_i^\mu\right) \exp\left[x \sum_i J_i^1 J_i^2\right] \quad (25)$$

where the coupling field x is used to control the distance between a pair of solutions $(\boldsymbol{J}^1, \boldsymbol{J}^2)$ and the associated overlap $\tilde{q} \equiv \frac{1}{N}\sum_i J_i^1 J_i^2$. This partition function has been used to predict optimal coupling field for a multiple random walking strategy to find a solution for the perceptronal learning problem [12]. In the following sections, we present an annealed computation as well as RS computation of the distance landscape

$s(d)$. Note that in this setting, Eqs. (3), (4) and (5) can also be used but here d should be understood as the distance separating two solutions in the weight space.

4.1. Annealed approximation for $s(d)$

Following the same techniques used in Sec. 3.1, we obtain the annealed free energy density as (see also Ref. [12])

$$f_{\rm ann} = \max_{\tilde{q},\hat{q}} \left\{ -\hat{q}\tilde{q} + x\tilde{q} + \log(4\cosh\hat{q}) + \alpha \log \int_0^\infty DtH\left(-\frac{\tilde{q}t}{\sqrt{1-\tilde{q}^2}}\right)\right\}. \quad (26)$$

The maximization with respect to $\tilde{q}$ and $\hat{q}$ leads to the following saddle point equation

$$\tilde{q} = \tanh\hat{q}, \quad (27)$$

$$\hat{q} = x + \frac{\alpha}{\sqrt{1-\tilde{q}^2}\,\mathrm{arccot}\left(-\frac{\tilde{q}}{\sqrt{1-\tilde{q}^2}}\right)}, \quad (28)$$

where the identity $\int_0^\infty DtH\left(-\frac{\tilde{q}t}{\sqrt{1-\tilde{q}^2}}\right) = \frac{1}{2\pi}\mathrm{arccot}\left(-\frac{\tilde{q}}{\sqrt{1-\tilde{q}^2}}\right)$ has been used. Using Eq. (4) and the above saddle point equation, we get the final expression for $s_{\rm ann}(d)$:

$$s_{\rm ann}(d) = \log 2 - (1-d)\log(1-d) - d\log d + \alpha \log\left[\frac{1}{2\pi}\mathrm{arccot}\left(-\frac{1-2d}{2\sqrt{d(1-d)}}\right)\right], \quad (29)$$

where $s_{\rm ann}(0) = (1-\alpha)\log 2$ which is actually the annealed entropy density of the original system [1]. If $\alpha = 0$, then $s_{\rm ann}(0) = \log 2$ which is in accord with the fact that the number of solution-pairs with a distance $d = 0$ should be the total number of solutions 2^N if no constraints are present.

4.2. Replica symmetric computation of $s(d)$

In this section, we derive the free energy density $f(x)$ for the landscape of solution-pairs under the replica symmetric approximation, using the replica trick introduced in Sec. 3.2. Since the partition function in this case involves a sum of all possible configurations of two synaptic weight vectors, the computation becomes a bit complicated. The disorder average of the integer power of the partition function can be evaluated as

$$\begin{aligned}\langle Z^n\rangle = \sum_{\{\boldsymbol{J}^{1,a},\boldsymbol{J}^{2,a}\}} \prod_{a<b} \int dq_1^{ab} dq_2^{ab} dr^{ab} \delta\left(q_1^{ab} - \frac{1}{N}\sum_i J_i^{1,a}J_i^{1,b}\right)\delta\left(q_2^{ab} - \frac{1}{N}\sum_i J_i^{2,a}J_i^{2,b}\right) \\ \times\delta\left(r^{ab} - \frac{1}{N}\sum_i J_i^{1,a}J_i^{2,b}\right)\prod_a \int dR^{aa}\delta\left(R^{aa} - \frac{1}{N}\sum_i J_i^{1,a}J_i^{2,a}\right) \\ \times\left\langle \prod_{\mu,a}\Theta(w_1^{\mu,a})\Theta(w_2^{\mu,a})\right\rangle e^{x\sum_{a,i}J_i^{1,a}J_i^{2,a}}, \end{aligned} \quad (30)$$

where $w_1^{\mu,a} \equiv \frac{1}{\sqrt{N}}\sum_i J_i^{1,a}\xi_i^\mu$ and $w_2^{\mu,a} \equiv \frac{1}{\sqrt{N}}\sum_i J_i^{2,a}\xi_i^\mu$. Under the replica symmetric ansatz, the disorder average is carried out as

$$\left\langle \prod_a \Theta(w_1^{\mu,a})\Theta(w_2^{\mu,a}) \right\rangle_{\boldsymbol{\xi}^\mu} = \int D\boldsymbol{z} \left[\int DyH(y_1)H(y_2)\right]^n, \tag{31}$$

where $\int D\boldsymbol{z} \equiv \int Dz_1 \int Dz_2 \int Dt$, $y_x = -\frac{\sqrt{R-r}y+\sqrt{q-r}z_x+\sqrt{r}t}{\sqrt{1-q-R+r}}(x=1,2)$ and we have used $q_1^{ab} = q_2^{ab} = q, r^{ab} = r, R^{aa} = R$ under the RS ansatz. After the computation of the summation in Eq. (30) by using $\sum_{a,b} J_i^a J_i^b = \frac{\left(\sum_a J_i^{1,a}+\sum_b J_i^{2,b}\right)^2 - \left(\sum_a J_i^{1,a}\right)^2 - \left(\sum_b J_i^{2,b}\right)^2}{2}$ and the Hubbard-stratonovich transform, we get the replica symmetric free energy density

$$f(x) = \alpha \int D\boldsymbol{z} \log F_1(q,R,r) + xR + \hat{q}(q-1) + \frac{1}{2}r\hat{r} - R\hat{R} + \int D\hat{\boldsymbol{z}} \log F_2(\hat{q},\hat{R},\hat{r}), \tag{32}$$

where $\int D\hat{\boldsymbol{z}} \equiv \int D\hat{z}_1 \int D\hat{z}_2 \int D\hat{z}_3$, $F_1(q,R,r) = \int DyH(y_1)H(y_2)$ and $F_2(\hat{q},\hat{R},\hat{r}) = 2e^{a_3}\cosh(a_1+a_2) + 2e^{-a_3}\cosh(a_1-a_2)$ where $a_x = \sqrt{\hat{q}-\hat{r}/2}\hat{z}_x + \sqrt{\hat{r}/2}\hat{z}_3 (x=1,2), a_3 = \hat{R}-\hat{r}/2$. The RS order parameters $(q,R,r,\hat{q},\hat{R},\hat{r})$ are determined by the following self-consistent equations

$$\hat{R} = x + \frac{\alpha}{1-q-R+r}\int D\boldsymbol{z}\frac{\int DyG(y_1)G(y_2)}{\int DyH(y_1)H(y_2)}, \tag{33}$$

$$\hat{r} = \frac{2\alpha}{1-q-R+r}\int D\boldsymbol{z}\frac{\int DyG(y_1)H(y_2)\int DyG(y_2)H(y_1)}{\left[\int DyH(y_1)H(y_2)\right]^2}, \tag{34}$$

$$\hat{q} = \frac{\hat{r}}{2} + \frac{\alpha}{2(1-q-R+r)}\int D\boldsymbol{z}\left[\frac{\int Dy\left[G(y_1)H(y_2)-G(y_2)H(y_1)\right]}{\int DyH(y_1)H(y_2)}\right]^2, \tag{35}$$

$$R = \int D\hat{\boldsymbol{z}}\frac{\tanh a_3 + \tanh a_1 \tanh a_2}{1+\tanh a_3 \tanh a_1 \tanh a_2}, \tag{36}$$

$$r = \int D\hat{\boldsymbol{z}}\frac{\tanh a_3(\tanh^2 a_1 + \tanh^2 a_2) + \tanh a_1 \tanh a_2(1+\tanh^2 a_3)}{(1+\tanh a_3 \tanh a_1 \tanh a_2)^2}, \tag{37}$$

$$q = r + \int D\hat{\boldsymbol{z}}\frac{(\tanh a_3 - 1)^2(\tanh a_1 - \tanh a_2)^2}{2(1+\tanh a_3 \tanh a_1 \tanh a_2)^2}. \tag{38}$$

In the derivation of the above saddle point equations, we have used a useful property of the Gaussian measure $\int DzzF(z) = \int DzF'(z)$ where $F'(z)$ is the derivative of the function $F(z)$ with respect to z. After the fixed point of the above saddle point equations is obtained, one can compute the entropy density $s = f(x) - xR$ with $d = \frac{1-R}{2}$. Note that $R-r$ may become negative, in this case we replace R and r by $-R$ and $-r$ respectively, y by $-y$ only for y_2 in Eqs. (33) to (35).

4.3. Message passing equations for single instances

By analogy with definitions in Sec. 3.3, we define by $p_{i\to a}^{J_i^1,J_i^2}$ the probability that the synaptic weight i takes a two-component vector state (J_i^1, J_i^2) in the absence of constraint a and by $\hat{p}_{b\to i}^{J_i^1,J_i^2}$ the probability that constraint b is satisfied given the vector state (J_i^1, J_i^2)

of weight i. These two cavity probabilities obey the following recursive equations

$$p_{i\to a}^{J_i^1,J_i^2} = \frac{1}{Z_{i\to a}} e^{xJ_i^1J_i^2} \prod_{b\in\partial i\backslash a} \hat{p}_{b\to i}^{J_i^1,J_i^2}, \tag{39}$$

$$\hat{p}_{b\to i}^{J_i^1,J_i^2} = \sum_{\{J_j^1,J_j^2,j\in\partial b\backslash i\}} \Theta\left(\frac{1}{\sqrt{N}}\sum_j J_j^1\xi_j^b\right)\Theta\left(\frac{1}{\sqrt{N}}\sum_j J_j^2\xi_j^b\right) \prod_{j\in\partial b\backslash i} p_{j\to b}^{J_j^1,J_j^2}, \tag{40}$$

where $Z_{i\to a}$ is a normalization constant. In fact the belief propagation equations (39) and (40) correspond to the stationary point of the Bethe free energy function of the current system [35, 36]. The exchange of $\boldsymbol{J}^1$ and $\boldsymbol{J}^2$ does not change the partition function Eq. (25), thus the cavity probabilities have the property that $p_{i\to a}^{+1,-1} = p_{i\to a}^{-1,+1}$ and $\hat{p}_{b\to i}^{+1,-1} = \hat{p}_{b\to i}^{-1,+1}$. This symmetry property will simplify the following analysis a lot.

To simplify Eqs. (39) and (40), we need the joint distribution of $w_{b\to i}^1$ and $w_{b\to i}^2$ where $w_{b\to i}^1 \equiv \frac{1}{\sqrt{N}}\sum_{j\neq i} J_j^1\xi_j^b$ and $w_{b\to i}^2 \equiv \frac{1}{\sqrt{N}}\sum_{j\neq i} J_j^2\xi_j^b$. Since we impose a distance constraint upon two solutions $\boldsymbol{J}^1$ and $\boldsymbol{J}^2$ in Eq. (25), there exists correlation between these two normally distributed random numbers and this correlation is characterized by the correlation coefficient

$$\hat{\rho}_{b\to i} = \frac{\sum_{j\in\partial b\backslash i}(q_{j\to b} - m_{j\to b}^2)}{\hat{\sigma}_{b\to i}} \tag{41}$$

due to the symmetry property. Based on Eq. (39), messages $q_{j\to b}$ and $m_{j\to b}$ are determined respectively by the following equations,

$$q_{j\to b} = \frac{e^x\left[\prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{+1,+1} + \prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{-1,-1}\right] - 2e^{-x}\prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{+1,-1}}{e^x\left[\prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{+1,+1} + \prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{-1,-1}\right] + 2e^{-x}\prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{+1,-1}}, \tag{42}$$

$$m_{j\to b} = \frac{e^x\left[\prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{+1,+1} - \prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{-1,-1}\right]}{e^x\left[\prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{+1,+1} + \prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{-1,-1}\right] + 2e^{-x}\prod_{a\in\partial j\backslash b}\hat{p}_{a\to j}^{+1,-1}}. \tag{43}$$

Therefore, both $w_{b\to i}^1$ and $w_{b\to i}^2$ obey a bivariate normal distribution and $\hat{p}_{b\to i}^{J_i^1,J_i^2}$ is reduced to be

$$\hat{p}_{b\to i}^{J_i^1,J_i^2} = \int_{-\frac{J_i^2\xi_i^b+\hat{w}_{b\to i}}{\sqrt{\hat{\sigma}_{b\to i}}}}^{\infty} DtH\left(-\frac{J_i^1\xi_i^b + \hat{w}_{b\to i}}{\sqrt{(1-\hat{\rho}_{b\to i}^2)\hat{\sigma}_{b\to i}}} - \frac{\hat{\rho}_{b\to i}t}{\sqrt{1-\hat{\rho}_{b\to i}^2}}\right) \tag{44}$$

where $\hat{w}_{b\to i} = \sum_{j\in\partial b\backslash i} m_{j\to b}\xi_j^b$ and $\hat{\sigma}_{b\to i} = \sum_{j\in\partial b\backslash i}(1-m_{j\to b}^2)$. The overlap $\tilde{q}$ is determined by $\tilde{q}(x) = \frac{1}{N}\sum_i \tilde{q}_i$ where $\tilde{q}_i$ is given by

$$\tilde{q}_i = \frac{e^x\left[\prod_{b\in\partial i}\hat{p}_{b\to i}^{+1,+1} + \prod_{b\in\partial i}\hat{p}_{b\to i}^{-1,-1}\right] - 2e^{-x}\prod_{b\in\partial i}\hat{p}_{b\to i}^{+1,-1}}{e^x\left[\prod_{b\in\partial i}\hat{p}_{b\to i}^{+1,+1} + \prod_{b\in\partial i}\hat{p}_{b\to i}^{-1,-1}\right] + 2e^{-x}\prod_{b\in\partial i}\hat{p}_{b\to i}^{+1,-1}}. \tag{45}$$

Eq. (44) is more computationally demanding than Eq. (19) since an additional numerical integral is required to compute $\hat{p}$ here. However, the integral in Eq. (44) can be approximated by $c_0 H\left(\frac{c+c_1}{\sqrt{c_2}}\right)$ if we write the right hand side of Eq. (44) as $\int_c^\infty Dt e^{\log H(a-bt)}$ and expand $H(a-bt)$ up to the second order in bt. The constants c_0,

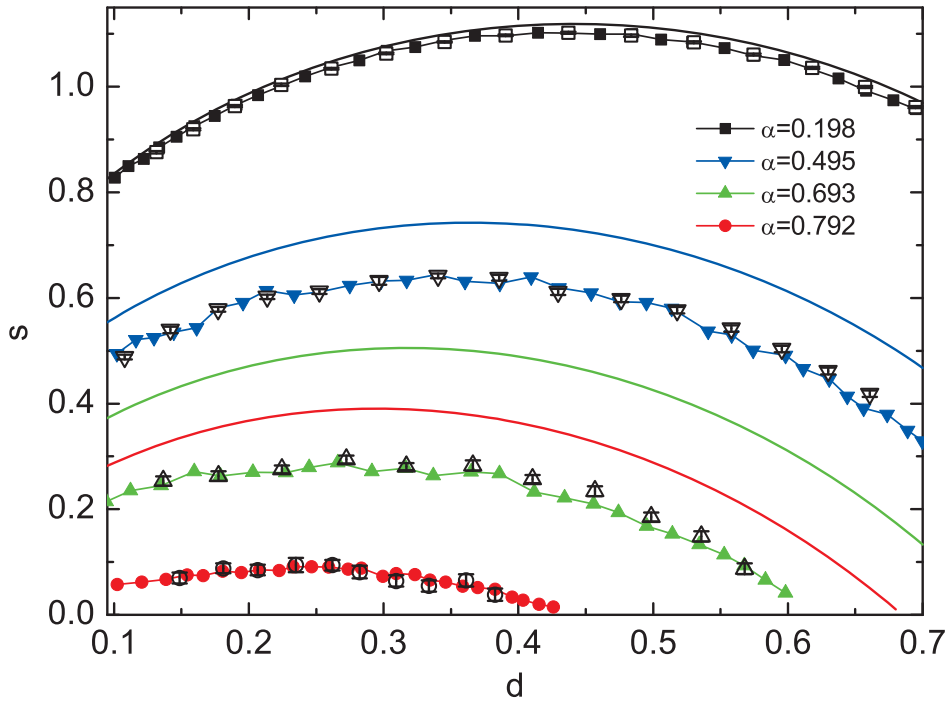

Figure 4. (Color online) Distance landscape of solution-pairs with a predefined distance d. The solid lines are the analytic annealed approximation (Eq. (29)) for $\alpha = 0.198, 0.495, 0.693, 0.792$ (from the top to the bottom) respectively. The line connecting symbols is a guide to the eye. The empty symbols stay for the numerical simulation results on systems with $(N, P) = (501, 99), (501, 248), (501, 347), (501, 397)$ (from the top to the bottom) using message passing algorithms. The result is the average over 20 random instances. Solid symbols are the replica symmetric results computed numerically by solving the saddle point equations.

c_1 and c_2 can be determined as a function of a and b. Therefore, this approximation is accurate only when large bt has vanishing contribution to the integral.

The free energy shift due to variable node addition (and all its adjacent constraints) can be obtained as $\Delta f_i = \log Z_i$ and the free energy shift due to constraint addition $\Delta f_a = \log Z_a$ where

$$Z_i = e^x \left[\prod_{b\in\partial i} \hat{p}_{b\to i}^{+1,+1} + \prod_{b\in\partial i} \hat{p}_{b\to i}^{-1,-1} \right] + 2e^{-x} \prod_{b\in\partial i} \hat{p}_{b\to i}^{+1,-1}, \tag{46}$$

$$Z_a = \int_{-\frac{\hat{w}_a}{\sqrt{\hat{\sigma}_a}}}^{\infty} DtH\left(-\frac{\hat{w}_a}{\sqrt{(1-\hat{\rho}_a^2)\hat{\sigma}_a}} - \frac{\hat{\rho}_a t}{\sqrt{1-\hat{\rho}_a^2}} \right), \tag{47}$$

where $\hat{w}_a = \sum_{j\in\partial a} m_{j\to a}\xi_j^a$, $\hat{\sigma}_a = \sum_{j\in\partial a}(1-m_{j\to a}^2)$ and $\hat{\rho}_a = \frac{\sum_{i\in\partial a}(q_{i\to a}-m_{i\to a}^2)}{\hat{\sigma}_a}$. The free energy density can then be obtained using Eq. (22) and the entropy landscape can be obtained correspondingly. The recursive equations Eqs. (41), (42), (43) and (44) can be solved by an iterative procedure similar to that used in Sec. 3.3.

As is seen from figure 4, the entropy density increases smoothly until a maximum

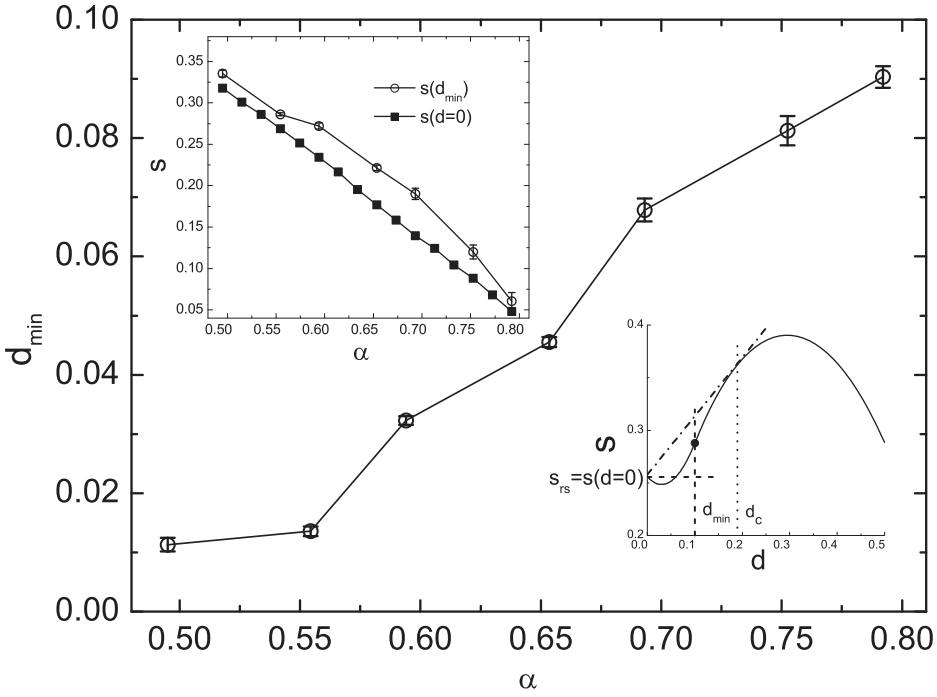

Figure 5. Minimal distance $d_{\rm min}$ as a function of the constraint density. The data points are computed by solving numerically the saddle point equations in Sec. 4.2. The error bars characterize the numerical fluctuations from ten different random initializations. The upper inset shows the corresponding entropy values. The lower inset shows *schematically* the typical concave and non-concave behavior of the entropy landscape, where the horizontal dashed line indicates $s_{\rm rs}$ and the vertical dashed line denotes $d_{\rm min}$, and the black point $(d_{\rm min}, s(d_{\rm min}))$ marks the first change of the concavity, i.e., $\frac{\partial^2 s(d)}{\partial d^2} = 0$. The vertical dotted line marks the first order thermodynamic transition point d_c, where the dash-dotted line going through $(0, s_{\rm rs})$ touches the concave part of $s(d)$ and has the slope $2x_c$. Note that $d_{\rm min}$ corresponds to the spinodal point x_s $(\geq x_c)$, i.e., $\frac{\partial s(d)}{\partial d}|_{d=d_{\rm min}} = 2x_s$.

is reached for $\alpha = 0.198$ and then decreases as the distance further grows. Interestingly, this behavior observed in figure 4 can be well fitted by the annealed approximation keeping the concavity of the entropy function. However, as α increases, large deviation from the annealed approximation occurs. The mean field calculations are supported by the numerical simulations on single instances using the proposed message passing algorithms, as shown in figure 4. The distance corresponding to the maximum of the entropy landscape curve in figure 4 is actually the typical distance $d_{\rm rs}$ calculated in figure 3, and $s(d = 0)$ recovers the typical entropy density of the original problem. By taking the limit $R \to 1$ in Eq. (32), one can show that $s(d = 0) = f(x = 0)$ where $f(x)$ is given by Eq. (13). As the constraint density increases, the maximal point of the entropy curve moves to the left, however, solution-pairs still maintain a relatively broad distribution in the solution space when α approaches α_s, e.g., $d_{\rm max} = {\rm argmax}_d\{s(d) = 0\} \simeq 0.332$ at $\alpha = 0.82$ ($d_{\rm rs} \simeq 0.222$), which may be responsible for the algorithmic hardness in this region.

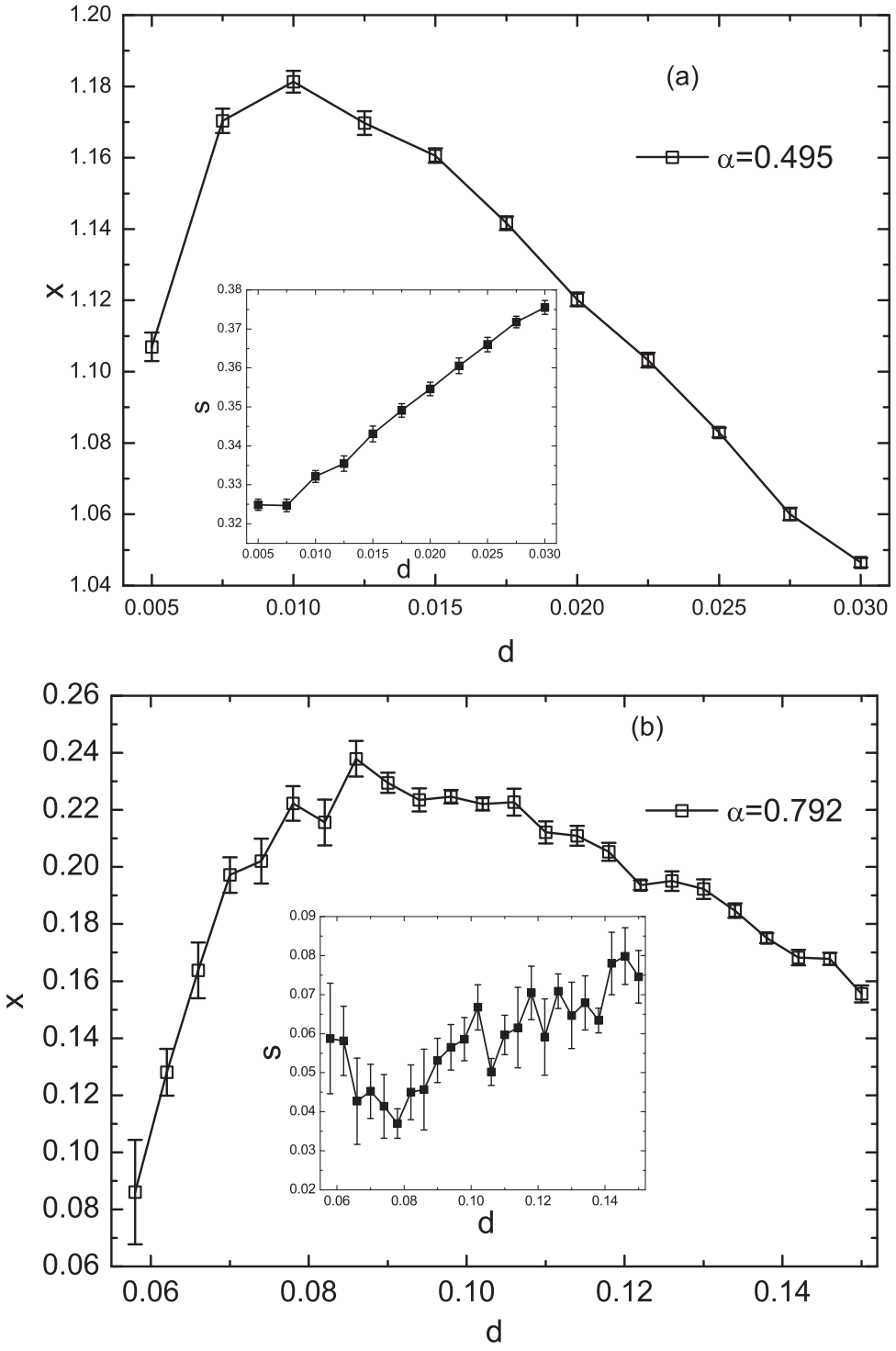

Figure 6. Compatible coupling field x for fixed d. The data points are computed by solving numerically the saddle point equations in Sec. 4.2. The error bars characterize the numerical fluctuations from ten different random initializations. Inset: the corresponding entropy curve as a function of d. (a) $\alpha = 0.495$. (b) $\alpha = 0.792$.

As α increases, the message passing algorithm requires a large number of iteration steps to converge (especially at small distances) and additionally a computationally expensive Monte Carlo integral involved in Eq. (44) cannot be avoided. On the other hand, when α is large enough, one can easily observe a rapid growth of the order parameter R to unity, i.e., at some critical coupling field x_c, R changes sharply from a value smaller than one to one. This implies that at x_c, $R = 1$ becomes a globally stable solution of the saddle point equations in Sec. 4.2. The first order thermodynamic transition is signalled by the change of the concavity at $d = d_{\min} > 0$. We define $d_{\min}$ as the minimal distance before $R = 1$ becomes a unique stable solution. Figure 5 shows the entropy gap and $d_{\min}$ as a function of the constraint density. The corresponding coupling field x_s marks the point where the concavity starts to change, i.e., $\frac{\partial^2 s(d)}{\partial d^2} = 0$, as shown in the lower inset of figure 5 ($x_s \geq x_c$). After x_s, $R = 1$ becomes the unique stable solution of the saddle point equations. Note that in the entropy gap, there exists a non-concave part of the entropy curve (for small distances), which can only be obtained by fixing d instead of x and searching for a compatible x (by the secant method). The result is shown in figure 6 for $\alpha = 0.495$ and 0.792. The compatible x for small distances (the left branch) is smaller than x_s. When $x > x_c$, the right branch is no longer globally stable solution but becomes metastable solution of the saddle point equations until $x = x_s$, i.e., the spinodal point is reached. By fixing $x_c \leq x \leq x_s$, one typically observes the right branch or $R = 1$, which describes the equilibrium properties of the Boltzmann measure in Eq. (25). Thus, the non-concave behavior observed in $d \in (0, d_{\min})$ is thermodynamically non-dominant and unstable, suggesting that the solution space is made of isolated solutions instead of separated clusters of exponentially many close-by solutions, and this behavior becomes much more evident as α increases. This explains why the multiple random walking strategy is extremely difficult to find a solution by tuning the coupling field at high α and large N [12].

As shown in figure 5, $d_{\min}$ increases as α grows, making a uniform sampling of solutions extremely hard. In addition, $d_{\min}$ seems to grow continuously, being the order of $\mathcal{O}(10^{-3})$ or less for $\alpha < 0.5$. The isolations of solutions can be explained by the nature of the hard constraints [37] for the binary perceptron. Unlike the random K-SAT and graph-coloring problems [38, 39], the hard constraint in the binary perceptron problem implies that the synaptic weight on one node in the factor graph is completely determined by the values of other nodes. But for finite N, solutions may not be strictly isolated. This explains why some local search heuristics can find a solution when N and α is not large enough. As α increases, some frozen solutions are more prone to disappear, thus solutions become much more far apart from each other, as shown by increasing $d_{\min}$ in figure 5. However, the thermodynamic properties can still be derived from the RS solution before α_s. We conjecture that clustering and freezing coexist for $\alpha < \alpha_s$, which is consistent with the computation in Ref. [4] in the sense that the total entropy (displayed in figure 3) $s_{\rm rs} = \Sigma(s) + s$ where $\Sigma(s)$ is the complexity of clusters of entropy density s and $s = 0$ for the current problem. We can say that the solution space is simple in the sense that it is made of isolated solutions instead of separated

clusters of exponentially many solutions; however, it becomes rather difficult to find a solution via stochastic local search algorithm. Below α_s, $s_{\rm rs} > 0$, meaning that there exist exponentially many solutions, but they are widely dispersed (much more apparent at large α). In other words, solution-pairs maintain a relatively broad distribution.

5. Discussion and Conclusion

The typical property of the distance landscape either from a reference configuration or for pairs of solutions is studied. For the first distance landscape, as the distance increases, the number of associated solutions grows first and then reaches its maximum (dominating the typical value of the entropy in the original system) followed by a decreasing behavior. This typical trend is confirmed by the numerical simulations on single instances using the proposed message passing algorithms. This behavior suggests that most of the solutions concentrate around the dominant point (the maximum in the distance landscape) in the N-dimensional weight space. It is clear that as the constraint density increases, the distance landscape shows larger and larger deviation from the analytic annealed approximation. We also calculate the second distance landscape characterizing the number of solution-pairs separated by a given distance. In this case, the replica symmetric result is in good agreement with the annealed computation at low α, while the large deviation is observed between the replica symmetric approximation and annealed computation for high α. Both landscapes are evaluated in the thermodynamic limit and confirmed by message passing simulations on large-size single instances.

In this paper, we calculate the whole picture of the distance (entropy) landscape and show that the entropy value rises to a maximum before declining at higher values of distance at certain range of distances. From the first landscape (a random configuration as a reference), we see clearly how the solution space shrinks as more constraints are added. From the second landscape of solution-pairs, we deduce a picture in which each global minimum (referred to as a canyon) is occupied by a single solution with zero ground state energy, and is surrounded by local minima (referred to as valleys) with positive energy [40]. This is also known as the valleys-dominated energy landscape [39]. The isolation of solutions implies that one cannot expect to satisfy all constraints by flipping a few synapses. The necessary number of synapses to be flipped should be proportional to N. The distance between the isolated solutions increases as the constraint density grows. This is the very reason why some simple local search heuristics cannot find a solution at high α or large N [11, 12] and the critical α for the local search algorithm decreases when the number of synapses increases. Simulated annealing process used in Refs. [6, 7] suffers from a critical slowing down when approaching a certain temperature, therefore, it would be interesting to study this picture within a finite temperature framework by focusing on the metastable states around the isolated solutions. The structure of these states should also be responsible for the algorithmic hardness. This issue will be addressed in our future work.

The distance landscape evaluated here is very similar to the weight enumerator in coding theory [41] and the method can be extended to consider the landscape analysis for low-density parity-check codes or code-division multiple access multiuser detection problems [21, 42], which will help to clarify what role the distance landscape plays with respect to the decoding performance.

Acknowledgments

We are grateful to Haijun Zhou for helpful comments on earlier versions of this paper. This work was partially supported by the Research Council of Hong Kong (Grant No. HKUST 605010) and the JSPS Fellowship for Foreign Researchers Grant No. $24 \cdot 02049$ (HH) and JSPS KAKENHI Nos. 22300003 and 22300098 (YK).

References

[1] W. Krauth and M. Mézard. Storage capacity of memory networks with binary couplings. *J. Phys. (France)*, 50:3057, 1989.

[2] H. Gutfreund and Y. Stein. Capacity of neural networks with discrete synaptic couplings. *J. Phys. A*, 23:2613, 1990.

[3] E. Barkai, D. Hansel, and H. Sompolinsky. Broken symmetries in multilayered perceptrons. *Phys. Rev. A*, 45:4146, 1992.

[4] T. Obuchi and Y. Kabashima. Weight space structure and analysis using a finite replica number in the ising perceptron. *J. Stat. Mech.*, P12014, 2009.

[5] H. M. Köhler. Adaptive genetic algorithm for the binary perceptron problem. *J. Phys. A*, 23:L1265, 1990.

[6] H. K. Patel. Computational complexity, learning rules and storage capacities: A monte carlo study for the binary perceptron. *Z. Phys. B*, 91:257, 1993.

[7] H. Horner. Dynamics of learning for the binary perceptron problem. *Z. Phys. B*, 86:291, 1992.

[8] M. Bouten, L. Reimers, and B. Van Rompaey. Learning in the hypercube: A stepping stone to the binary perceptron. *Phys. Rev. E*, 58:2378, 1998.

[9] A. Braunstein and R. Zecchina. Learning by message passing in networks of discrete synapses. *Phys. Rev. Lett*, 96:030201, 2006.

[10] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina. Efficient supervised learning in networks with binary synapses. *Proc. Natl. Acad. Sci. USA*, 104:11079, 2007.

[11] H. Huang and H. Zhou. Learning by random walks in the weight space of the ising perceptron. *J. Stat. Mech.: Theory Exp*, P08014, 2010.

[12] H. Huang and H. Zhou. Combined local search strategy for learning in networks of binary synapses. *Europhys. Lett*, 96:58003, 2011.

[13] E. Gardner. The space of interactions in neural network models. *J. Phys. A*, 21:257, 1988.

[14] T. Mora and M. Mézard. Geometrical organization of solutions to random linear boolean equations. *J. Stat. Mech*, P10007, 2006.

[15] H. Daudé, M. Mézard, T. Mora, and R. Zecchina. Pairs of sat assignment in random boolean formula. *Theor. Comput. Sci*, 393:260, 2008.

[16] L. Dall'Asta, A. Ramezanpour, and R. Zecchina. Entropy landscape and non-gibbs solutions in constraint satisfaction problems. *Phys. Rev. E*, 77:031118, 2008.

[17] H. Zhou and C. Wang. Ground-state configuration space heterogeneity of random finite-connectivity spin glasses and random constraint satisfaction problems. *J. Stat. Mech.*, P10010, 2010.

[18] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theory*, 47:498, 2001.

[19] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

[20] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *J. Phys. A*, 22:2181, 1989.

[21] Y. Kabashima. A cdma multiuser detection algorithm on the basis of belief propagation. *J. Phys. A*, 36:11111, 2003.

[22] J. P. Neirotti and D. Saad. Improved message passing for inference in densely connected systems. *Europhys. Lett*, 71:866, 2005.

[23] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans Inf Theory*, 57:764, 2011.

[24] K. Y. Michael Wong. Microscopic equations and stability conditions in optimal neural networks. *Europhys. Lett*, 30:245–250, 1995.

[25] A. Barra. Irreducible free energy expansion and overlaps locking in mean field spin glasses. *J. Stat. Phys*, 123:601, 2006.

[26] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

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

[28] F. Guerra and F. L. Toninelli. The thermodynamic limit in mean field spin glass models. *Commun. Math. Phys.*, 230:71, 2002.

[29] M. Mézard and A. Montanari. *Information, Physics, and Computation*. Oxford University Press, Oxford, 2009.

[30] A. Barg and Jr G. David Forney. Random codes: minimum distances and error exponents. *IEEE. Trans. Inf. Theory*, 48:2568, 2002.

[31] M. Mézard and G. Parisi. The bethe lattice spin glass revisited. *Eur. Phys. J. B*, 20:217, 2001.

[32] G. Biroli, R. Monasson, and M. Weigt. A variational description of the ground state structure in random satisfiability problems. *Eur. Phys. J. B*, 14:551, 2000.

[33] H. Zhou. T→ 0 mean-field population dynamics approach for the random 3-satisfiability problem. *Phys. Rev. E*, 77:066102, 2008.

[34] S. Franz and G. Parisi. Phase diagram of coupled glassy systems: a mean-field study. *Phys. Rev. Lett*, 79:2486, 1997.

[35] M. Mézard and G. Parisi. The cavity method at zero temperature. *J. Stat. Phys*, 111:1, 2003.

[36] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Trans Inf Theory*, 51:2282–2312, 2005.

[37] O. C. Martin, M. Mézard, and O. Rivoire. Frozen glass phase in the multi-index matching problem. *Phys. Rev. Lett*, 93:217205, 2004.

[38] H. Zhou. Glassy behavior and jamming of a random walk process for sequentially satisfying a constraint satisfaction formula. *Eur. Phys. J. B*, 73:617, 2010.

[39] F. Krzakala and L. Zdeborová. Phase transitions and computational difficulty in random constraint satisfaction problems. *J. Phys.: Conf. Ser*, 95:012012, 2008.

[40] L. Zdeborová and F. Krzakala. Generalization of the cavity method for adiabatic evolution of gibbs states. *Phys. Rev. B*, 81:224205, 2010.

[41] C. Di, A. Montanari, and R. Urbanke. Weight distribution of ldpc codes: Combinatorics meets statistical physics,. In *Proc. IEEE Int. Symp. Information Theory*, page 102, Chicago, 2004.

[42] Y. Kabashima and D. Saad. Statistical mechanics of low-density parity-check codes. *J. Phys. A*, 37:R1–R43, 2004.