LETTER

Stability of the replica symmetric solution in diluted perceptron learning

Alejandro Lage-Castellanos\textsuperscript{1,2,4}, Andrea Pagnani\textsuperscript{3} and Gretel Quintero Angulo\textsuperscript{1,2}

1 Department of Theoretical Physics, Physics Faculty, University of Havana, La Habana, CP 10400, Cuba
2 ‘Henri-Poincaré’ Group of Complex Systems and Department of Theoretical Physics, Physics Faculty, University of Havana, La Habana, Cuba
3 Human Genetic Foundation (HuGeF-Torino), Via Nizza 52, I-10122, Torino, Italy
E-mail: lage@fisica.uh.cu, andrea.pagnani@gmail.com
and gquintero@fisica.uh.cu

Received 21 September 2012
Accepted 11 January 2013
Published 12 February 2013

Online at stacks.iop.org/JSTAT/2013/L02002
doi:10.1088/1742-5468/2013/02/L02002

Abstract. We study the role played by dilution in the average behavior of a perceptron model with continuous coupling with the replica method. We analyze the stability of the replica symmetric solution as a function of the dilution field for the generalization and memorization problems. Thanks to a Gardner-like stability analysis we show that at any fixed ratio $\alpha$ between the number of patterns $M$ and the dimension $N$ of the perceptron ($\alpha = M/N$), there exists a critical dilution field $h_c$ above which the replica symmetric ansatz becomes unstable.

Keywords: message-passing algorithms, cognitive dynamical networks, learning theory, statistical inference

\textsuperscript{4} Author to whom any correspondence should be addressed.
Introduction

Neural networks [1] have become among the most studied and successful models in the field of artificial intelligence. In spite of more than 50 years of research, the field is still thriving [2]–[4]. In the past few years, thanks to the advances of the new-generation high-throughput technologies in molecular biology, the field has experienced a renewed interest with problems coming from the analysis of high-dimensional data in biology, where sparsity in the underlying model is a general feature [3]–[8].

In this work we address the issue of the stability of the replica symmetric solution presented in [9, 10] which has important implications for the implementation of many algorithmic strategies. This work extends thus the analysis presented in [9, 10] and generalizes the stability results of [11] to the external dilution case.

A standard problem in artificial intelligence is that of generalization [11, 12]. A set of $M$ patterns $\vec{x}^\mu$ is given together with a binary output variable $y^\mu$ for each of them ($\mu \in \{1, 2, \ldots, M\}$). A pattern $\vec{x}$ is an $N$-dimensional vector and we are interested in learning the hidden relation among its components and the classification value $y^\mu$ related to pattern $\mu$. In the simplest setup, a linear perceptron tries to encode such a relation in a vector $\vec{J}$ (called student) such that the binary classification $y^\mu = \pm 1$ is reproduced by

$$
y^\mu = \text{Sign}(\vec{J} \cdot \vec{x}^\mu),$$

We will consider that the classification is actually generated by an unknown teacher $\vec{J}^0$

$$
y^\mu = \text{Sign}(\vec{J}^0 \cdot \vec{x}^\mu + \eta^\mu)$$

and therefore, the aim of the student $\vec{J}$ is to be as close as possible to the teacher $\vec{J}^0$. A Gaussian noise $\eta^\mu \sim N(0, N \gamma^2)$ is added to the classification function to account for

doi:10.1088/1742-5468/2013/02/L02002
Experimental noise in the data. Two interesting limits are studied: no noise $\gamma = 0$, and random classification $\gamma = \infty$.

As a first step in analyzing the problem, we define an energy function counting the number of patterns that are wrongly classified by the student $\vec{J}$:

$$E(\vec{J}) = \sum_{\mu} \Theta(- (\vec{J}_0 \cdot \vec{x}^\mu + \eta^\mu) \vec{J} \cdot \vec{x}^\mu).$$

(2)

In the following we will consider only the case $J \in \mathbb{R}^N$. Since the energy expression depends only on the angle of $\vec{J}$ and not on its length, i.e. $E(c\vec{J}) = E(\vec{J})$ for all $c > 0$, we restrict the student to the surface of the sphere $\sum_{i=1}^N J_i^2 = N$.

Standard statistical mechanics methods [3, 6], [13]–[15] have been largely used to study the thermodynamic properties of this problem. In a recent work [9, 10] a slightly different point of view has been analyzed: which is the $\vec{J}$ that minimizes the energy function with the largest possible number of coordinates equal to zero or, in other words, which is the sparsest $\vec{J}$ compatible with a correct classification?

The problem of extracting a sparse signature from high-dimensional datasets is ubiquitous in many different fields ranging across computer science, combinatorial chemistry, text processing and finally to bioinformatics, and different approaches have been proposed so far [16, 17]. The literature concerning feature selection has so far concentrated around two different strategies: (i) the wrapper strategy which utilizes learning to score signatures according to their predictive value; (ii) using filters that fix the signature as a preprocessing step independent from the classification strategy used in the second step.

This issue has been addressed in [9, 10] where the performance of a perceptron in presence of an external dilution field $h$ coupled to the classification vector $\vec{J}$ was studied. To this end the following Hamiltonian was considered:

$$\beta \mathcal{H}(\vec{J}) = \beta E(\vec{J}) + h \|\vec{J}\|_p$$

(3)

where the dilution field $h$, in the last term, acts as a chemical potential on $\vec{J}$.

This way of addressing the inference of a sparse teacher can be considered as a wrapper strategy of sparse signature extraction, and in the statistical mechanics of disordered systems field it is known as the annealed dilution problem [18], since the cutting of the weights is done during the learning process and depends on the experiments: the weights that are set to zero correspond to variables that are not relevant to the classification. Another way to impose a dilution is by setting a fraction of the weights to zero at random (quenched dilution) before or after the learning process, but in that case the weights that are set to zero are independent from the patterns and cannot give any information about which components are relevant to the classification [18].

Concerning the norm used for dilution, two interesting cases were studied at the replica symmetric level: the $L_1$ norm ($\|\vec{J}\|_1 = \sum_{i=1}^N |J_i|$) and the $L_0$ norm ($\|\vec{J}\|_0 = \lim_{p \to 0} \sum_{i=1}^N |J_i|^p = \sum_{i=1}^N (1 - \delta_{0J_i})$).

The use of dilution has special practical interest in the case where one knows a priori that the teacher is actually sparse. In many real life problems [3]–[5], [7] the multidimensional patterns $\vec{x}$ contain irrelevant information, in the sense that only few of the components are actually considered in the classification process. To represent this situation we will consider that only a fraction $n_{\text{eff}}^0$ of the teacher’s components are nonzero.

doi:10.1088/1742-5468/2013/02/L02002
Therefore each teacher’s component is extracted independently from the distribution
\[ \rho(J_0) = (1 - n_{\text{eff}}^0) \delta_{0,h} + n_{\text{eff}}^0 \rho'(J_0). \] (4)

It has been shown [5, 8] that dilution with \( L_p \) norm, with \( p \leq 1 \), forces a fraction of the components of \( \vec{J} \) to be exactly zero, and therefore, allows a selection of the components that actively participate in the classification process. It is worth pointing out that the aim of this work is to analytically characterize the stability of the replica symmetric solution of the problem through a replica computation, and that the dilution of the teacher will always be considered as a fraction of \( N \) as in equation (4). The extreme dilution case (i.e. the case where only \( O(1) \) elements of the teacher’s vector are different from zero) is beyond the scope of a replica computation.

The minimization of the cost function \( H(\vec{J}) \) raises severe computational problems due to the scale invariance of the energy (\( \vec{J} \) lives on the surface of the \( N \)-hypersphere) which makes the optimization problem non-convex in general. At odds with problems like compressed sensing [19], for which optimization methods of order \( O(N^3) \) can be used for the \( L_1 \) norm [5] whereas the most effective \( L_0 \) norm [5] makes the minimization an NP-hard optimization problem [3,5], in our perceptron-like case we are not aware of any ad hoc numerical technique for finding efficiently the minimum of the cost function in equation (3), apart from Monte Carlo based optimization methods.

From a theoretical perspective, we are interested in the average case properties of the structure of the solution space in either of cases \( L_1 \) and \( L_0 \). The work is organized in the following way: section 1 fixes the notation and sketches the replica symmetric results presented in [9]; in section 2 we study the stability of this replica symmetric solution; in section 3 we show that at \( h = 0 \) the replica symmetric solution is always stable, while at \( h \to \infty \) it is always unstable, and therefore a critical value of the dilution field marks the threshold between two phases, computed in section 4; and finally, our results are summarized in section 5, and the main results of [9] are re-evaluated in the present context. Technical details about the stability are reported in the appendix.

1. The replica symmetric solution

To get the average properties of the solution space we need to compute the expected free energy over the experiments:
\[ \beta F = \lim_{N \to \infty} (1/N) \log Z, \]
where
\[ Z(\beta, h) = \int \prod_{i=1}^{N} dJ_i \exp \left( -\beta E(\vec{J}) - h \| \vec{J} \|_p \right) \delta(J^2 - N) \]
(5)
is the partition function. It depends implicitly on the teacher \( \vec{J}^0 \) and the patterns presented to the perceptron \( (y^\mu, \vec{x}^\mu) \), and the average \( \log Z \) is taken with respect to both. The expected value \( \bar{F} \) should give the typical situation of any experiment in the thermodynamic limit.

The teacher and the patterns act as frozen disorder in this problem, and we use the replica method [13, 20] to average over them. This consists in computing the intensive free energy by using this equivalent expression:
\[ \beta F = \lim_{n \to 0} \lim_{N \to \infty} \frac{1}{nN} \log Z^n. \]
For integer $n$, $Z^n$ is the average partition function of $n$ identical systems (the replicas of the original one):

$$Z^n = 2^{-MN} \sum_{x_i^a = \pm 1} \int \cdots \int \prod_{\mu=1}^M D_x[\eta_\mu] \prod_{i=1}^N \prod_{a=1}^N dJ_i^0 \prod_{i=1}^N \rho(J_i^0) \int \cdots \int \prod_{a=1}^N \prod_{i=1}^N dJ_i^a$$

$$\times \exp \left\{ -\beta \sum_{a=1}^n \sum_{\mu=1}^N \Theta \left( - \left[ \sum_{i=1}^N J_i^0 x_i^a + \eta_\mu \right] \left[ \sum_{i=1}^N J_i^a x_i^\mu \right] \right) - h \sum_{a=1}^n \| \hat{J}_i^a \|_p \right\} .$$  

Then the averaging can be done by introducing the variables $Q_{ab}, a, b = 0, 1, 2, \ldots, n$ via integrals of the delta functions $\delta(NQ_{ab} - \sum_i J_i^a J_i^b)$ and their Fourier conjugates $\hat{Q}_{ab}, a, b = 1, 2, \ldots, n$. The $Q_{ab}$ are known as the overlaps between two replicas $a$ and $b$. Both the $Q_{ab}$s and the $\hat{Q}_{ab}$s form symmetric $(n+1) \times (n+1)$ matrices since $Q_{ab} = Q_{ba}$ and consequently $\hat{Q}_{ab} = \hat{Q}_{ba}$.

In order to take the limit $n \to 0$ in the average free energy we need to make an assumption about the values of these new variables. The replica symmetric (RS) ansatz assumes that any two replicas of the system have the same overlap, and therefore the corresponding overlap matrix and its Fourier conjugate have the following structure:

$$Q = \begin{pmatrix} t & 1 & \ldots & 1 \\ r & q & 1 & \ldots & 1 \\ r & q & \ldots & q & 1 \end{pmatrix}$$

$$-i\hat{Q} = \begin{pmatrix} \lambda^0 & \hat{\lambda} & \ldots & \hat{\lambda} \\ \hat{\lambda} & \lambda & \ldots & \lambda \\ \hat{\lambda} & \hat{\lambda} & \ldots & \lambda \end{pmatrix}. \quad (7)$$

The first column (row) contains the teacher’s parameters. For instance, $t$ is the variance of the teacher and is related to the distribution (4), and is not a variational parameter. On the other hand, $r = \langle (1/N)J \cdot \hat{J}^0 \rangle$ is the average overlap between the teacher and the student, and $q = \langle (1/N)\hat{J}^a \cdot \hat{J}^b \rangle$ is the average overlap between two replicas (see [9] for details).

The correct value for the free energy is obtained by minimizing the variational RS free energy

$$-\beta \bar{f} = -r\hat{r} + \frac{1}{2}q\hat{q} - \lambda + G_J + \alpha G_X$$

with respect to the variational parameters $(q, r, \hat{q}, \hat{r}, \lambda)$, where

$$G_J = \int Dx \int dJ \rho(J^0) \log \int dJ e^{-\frac{1}{2}J^0 J - h||J||_p + \langle \hat{J}^0 - \sqrt{q}x \rangle J}$$

$$G_X = 2 \int Dx \left( \frac{q}{\sqrt{q}^2 + q - r^2} \right) \log \left( e^{-\beta - 1} \right). \quad (8)$$

From now on we will define $Dx = dx \exp(-x^2/2)/\sqrt{2\pi}$, and the function $H(y) = \int_y^\infty Dx$.

A key role is played by $\alpha = M/N$, the ratio between the number of patterns $M$ and the space dimensionality $N$, that measures the amount of available information and, together
with \( h \), is the fundamental parameter controlling the quality of the generalization. We assume that in the limit \( N \to \infty \), \( \alpha \) remains finite.

2. Stability à la Gardner

The main goal of this work is to analyze the stability of the replica symmetric solution. Other structures for the overlap matrices \((7)\) are indeed possible. From the analogy with the organization of the thermodynamic states in the low temperature phase of mean field spin glass-like models [20], we expect that the space of solutions of our model could be described by a hierarchical (ultrametric) organization of replicas, or, in other words, that our system might undergo a spontaneous replica symmetry breaking (RSB). From a purely geometric point of view this process is related to the fragmentation of the space of zero-energy configurations into unconnected regions, or equivalently, the fragmentation of the Gibbs measure into many pure states (see figure 1).

We will now apply Gardner’s method for the stability analysis of the replica symmetric solution in [14], studying the eigenvalues of the Hessian matrix:

\[
\|\partial^2(\beta n f_n)\| = \begin{pmatrix}
\|\partial^2(\beta n f_n)\|_{\partial Q_{ab}\partial Q_{cd}} \\
\|\partial^2(\beta n f_n)\|_{\partial Q_{ab}\partial (-i\hat{Q}_{cd})}
\end{pmatrix} \begin{pmatrix}
\|\partial^2(\beta n f_n)\|_{\partial Q_{ab}\partial (-i\hat{Q}_{cd})} \\
\|\partial(-i\hat{Q}_{ab})\partial(-i\hat{Q}_{cd})\|
\end{pmatrix}
\]

evaluated at the RS fixed point. Details of the calculation are given in the appendix.

The eigenvalues of the \( n(n+1) \times n(n+1) \) Hessian matrix are simply related to the spectrum of the four sub-matrices. Among the eigenvectors, only those with eigenvalues \( \delta_1 \) and \( \delta_2 \) signal an instability of the replica symmetric solution:

\[
\delta_1 \delta_2 = \gamma_1 \gamma_2 - 1
\]

where \( \gamma_1 \) and \( \gamma_2 \) are eigenvalues of the inner matrices \( \|\partial^2(\beta n f_n)\|/\partial Q_{ab}\partial Q_{cd}\| \) and

\[\text{doi:10.1088/1742-5468/2013/02/L02002}\]
Stability of the replica symmetric solution in diluted perceptron learning

\[ \left\| \partial^2(\beta n T_n) / \partial (-i Q_{ab}) \partial (-i Q_{cd}) \right\|, \text{ given by} \]

\[ \gamma_1 = - \frac{2\alpha}{(1 - q)^2} \int \text{Dy} H \left( \frac{yr}{\sqrt{q r^2 + qr - r^2}} \right) \left\{ \left( \frac{e^{-h_0^2/2}}{\sqrt{2\pi} (H(h_0) - 1)} \right)^2 - \frac{h_0 e^{-h_0^2/2}}{\sqrt{2\pi} (H(h_0) - 1)} \right\}^2 \]

\[ \gamma_2 = - \int \text{Dz} \int \text{d}J^0 \rho(J^0) \left\{ \left( \frac{\int \text{d}J \, J^2 e^{H_J}}{\int \text{d}J \, e^{H_J}} \right)^2 - \left( \frac{\int \text{d}J \, J e^{H_J}}{\int \text{d}J \, e^{H_J}} \right)^2 \right\}^2 \]

where \( H_J = -(q/2 - \lambda) J^2 + (r J_0 - \sqrt{q} x) J - h|J|^p \).

The sign of the product \( \delta_1 \delta_2 \) determines the stability of the extremal point: when \( \delta_1 \delta_2 < 0 \) the point is stable, and it is unstable otherwise.

3. Extreme cases: \( h = 0 \) and \( h \to \infty \)

In the zero-temperature case, the Gibbs measure concentrates over the states of lower energy, i.e. over those vectors \( \vec{J} \) that correctly classify most patterns. In the absence of dilution \( (h = 0) \), the measure is uniform over these states, and the free energy is thus a measure of their entropy (we are working at zero temperature). At the other extreme, we have the high dilution limit \( h \to \infty \) that we studied in [9] at the replica symmetric level. Next we check the stability of the replica symmetric solution for both the memorization and the generalization problems in these two extreme situations. As we will see, the replica symmetric solution is stable for \( h = 0 \) and unstable for \( h \to \infty \), and therefore, there is a critical value \( h_c(\alpha) \) dividing the RS phase from the non-RS one.

3.1. Stability of the nondiluted perceptron, \( h = 0 \)

In the absence of dilution \( (h = 0) \) we still have two cases: memorization \( (\gamma \to \infty) \) and generalization \( (\gamma = 0) \). In the high noise limit \( \gamma \to \infty \) (see equation (1)) there is no rule to infer because the experiments are randomly classified, and the perceptron tries to memorize the output variable \( y^\mu \). In the \( \gamma = 0 \) case, the patterns are classified according to the teacher and this rule can be generalized.

In the memorization limit \( (\gamma \to \infty) \), an RS solution of zero energy always exists for \( 0 < \alpha < 2 \) [14]. For these values of \( \alpha \) the student is capable of memorizing the experiments, whereas for \( \alpha > 2 \) there are no zero-energy solutions, and solutions are not replica symmetric. In agreement with this known behavior (see [14]), the product of the eigenvalues \( \delta_1 \delta_2 \), evaluated in the nondiluted memorization replica symmetric solution, is negative for \( \alpha < 2 \). Its value grows from \(-1\) at \( \alpha = 0 \) to 0 at \( \alpha = 2 \) (figure 2), an indication that the zero-energy replica symmetric solution is stable in this range.

In the generalization limit \( (\gamma = 0) \) the perceptron can asymptotically learn the classification rule provided with a big enough amount of data as shown in [9]. However, for any finite amount of information \( \alpha \), there is not just one zero-cost student, but a continuum of them filling a bounded region of the \( J_s \) space (see the left panel of figure 1). An indication of the size of the solution space is given by the student–student overlap \( q = \langle J^a J^b \rangle \). As in the memorization case (for \( \alpha < 2 \)), the solution space is connected, and, consistently, the product \( \delta_1 \delta_2 \) remains always negative (see figure 2), which means that...
Stability of the replica symmetric solution in diluted perceptron learning

Figure 2. The eigenvalue product (left) evaluated in the replica symmetric solution for the generalization and memorization not diluted, and the corresponding self-overlap $q$ of the student (right).

The replica symmetric solution is stable for every $\alpha$ as in [11]. As $\alpha \to \infty$ the solution space shrinks around the correct value $\vec{J}^0$, and $q \to 1$ (see figure 2).

As expected, as little information is given to the student (low values of $\alpha$), there is little difference between the memorization and the generalization. Figure 2 shows that also the stability, given by the product $\delta_1 \delta_2$ for the nondiluted generalization and memorization, coincides for small $\alpha$ and has the same limit for $\alpha \to 0^+$.

3.2. Stability of the diluted perceptron, $h = \infty$

The dilution field selects those solutions with the lowest values of the norm. In particular, $L_0$ and $L_1$ norms are known to force the sparsity of the solutions [8], pushing a fraction of the student components to be zero. Seeking for sparse solutions can enhance the efficiency in the use of available information [5]–[8], in particular when the teacher is actually sparse. In the following we will consider a 95% sparse teacher, i.e. $n_{\text{eff}}^0 = 0.05$ in equation (4).

The behavior of the large dilution field $h \to \infty$ limit depends on the norm used and so does the learning behavior of the perceptron [9]. We restrict ourselves to the study of the space of $E = 0$ solutions. To do so we take the limit $T \to 0$ first, and then the limit $h \to \infty$. Again in [9] it was shown that, in the replica symmetric case, $h \to \infty$ pushes $q$ to 1 at any $\alpha$, concentrating the Gibbs measure over a single point, i.e. the most diluted zero-energy $\vec{J}$. However, the question remains as to whether the replica symmetric ansatz is still valid or, for the strong dilution field, the Gibbs measure breaks into unconnected components.

In figure 3 we show the eigenvalue product for the $L_0$ and $L_1$ dilutions in the $h \to \infty$ limit for both the memorization and the generalization cases. At variance with the nondilated case, the $h = \infty$ one is always unstable.

4. The phase diagram

In the region of $\alpha$ where perfect memorization/generalization is possible, the replica symmetric ansatz is stable in absence of dilution and unstable in presence of a very strong dilution field $h \to \infty$. Therefore there should exist a critical value for the dilution
Stability of the replica symmetric solution in diluted perceptron learning

Figure 3. The eigenvalue product evaluated for the replica symmetric solution for the $L_0$ and the $L_1$ norms in generalization and memorization.

Figure 4. The phase diagram for the $L_1$ (left) and the $p = 0.1$ (right) norms with several teacher dilutions $n_{\text{eff}}$. The replica symmetric solution is stable for values of $\alpha$ and $h$ below the lines.

field $h_c(\alpha)$ separating the RS from the RSB phase. In figure 4 we show $h_c(\alpha)$ for the $L_1$ and the $p = 0.1$ norms applied to models with different teacher dilutions.

The similarity of the stability curves near $\alpha = 0$ arises because with so little information there is not a big difference between memorizing and generalizing. The curves separate around $\alpha \gtrsim 0.4$. The value of the critical dilution field increases with $\alpha$. This can be rationalized as follows: more information implies a reduction in the size of the $E = 0$ solution space, and therefore different solutions are closer, so a higher dilution field is needed to ‘clusterize’ the set of zero-energy students. In generalization, when $\alpha \to \infty$ the solution space is composed of only one vector which is precisely the teacher, and the replica–replica overlap $q$ achieves its maximum value 1; consequently the critical curves go to infinity too. For the memorization case, the value $q = 1$ is achieved for $\alpha = 2$, and that is why the critical curve diverges here.

How to draw the critical lines for the $L_0$ norm is not obvious because setting $p = 0$ in the equations, before taking the strong dilution limit $h \to \infty$, eliminates the effect of any finite dilution field, so there is no way of finding $h_c$ (see [9]). What we have done instead is to study the critical lines for a value of the norm exponent $p$ that is small enough. The

doi:10.1088/1742-5468/2013/02/L02002
results obtained using $p = 0.1$ are shown in the right panel of figure 4. Again, around $\alpha = 0$, memorization and generalization behave similarly. The memorization critical line also diverges when $\alpha \to 2$, while the generalization one increases with $\alpha$ as for the $L_1$ norm. In both cases, $L_1$ and $L_{0,1}$, the critical field $h_c(\alpha)$ depends on the sparsity $n_{\text{eff}}^0$ of the teacher (equation (4)).

5. Conclusions

We have performed a full stability analysis of the replica symmetric solution for the nondiluted and diluted generalization and memorization problems for a learning perceptron. Imposing a dilution should improve the results obtained in real algorithms, but comes at a price. We showed that even in the satisfiable phase (where zero-cost solutions exist) an infinite dilution breaks the symmetry of the replicas, whereas no dilution at all keeps the replica symmetric solution stable. The breaking of the symmetry is usually connected with convergence problems in algorithms like belief propagation. Yet, it is always possible to find a dilution field weak enough for the replica symmetric solution to be stable, provided it exists. The critical dilution field $h_c(\alpha)$ depends on the actual sparseness of the teacher. Our results are in agreement with the analysis presented in [21] where a perceptron with diluted couplings was studied although without an external dilution field.

Figure 5, partially taken from [9], shows the generalization error achieved in the average case by a perceptron without a dilution field, and with a very strong ($h \to \infty$) dilution field. The teacher used is quite sparse ($n_{\text{eff}}^0 = 5\%$), and therefore the learning process is enhanced by the use of dilution. Restricting ourselves to the replica symmetric space, the $h \to \infty$ limit is not achievable, and $h_c(\alpha)$ is the best we can do. The gain in accuracy (lower error) obtained with an $h_c$-diluted perceptron is not as impressive as that for the $h \to \infty$ case, but still improves the results without dilution.

doi:10.1088/1742-5468/2013/02/L02002
The present work is useful as it defines the phase space of the replica symmetric solution in the information–dilution ($\alpha$–$h$) plane. The actual relation of these predictions to the behavior of algorithms remains as an open project for the future. Also left for the future is the use of 1RSB (or higher) parameters for making average predictions for the RSB phase.

Acknowledgment

The authors would like to thank Martin Weigt, of the Statistical Genomics and Biological Physics group at the Université Pierre et Marie Curie, for useful comments.

Appendix. Details of the stability calculation

The variational free energy in terms of all the $Q_{ab}$ and their Fourier counterparts is (see [9]), after taking the $N \rightarrow \infty$ limit, given by

$$
\beta n \bar{f}_n = -i \sum_{a \leq b} Q_{ab} \hat{Q}_{ab} 
- \log \int \prod_{a=1}^{n} dJ^a \int dJ^0 \rho(J^0) e^{-h \sum_{a=1}^{n} |J^a|^p - h \sum_{a=1}^{n} \sum_{b=1}^{n} Q_{ab} J^a J^b} 
- \alpha \log \int D\eta \prod_{a=0}^{n} \frac{dx^a}{2\pi} e^{-\beta \sum_{a=1}^{n} \theta(-x^a(0^+ + \eta)) + i \sum_{a=0}^{n} x^a - (1/2) \sum_{a,b=0}^{n} Q_{ab} x^a x^b}.
$$

Differentiating with respect to its $(n+1)(n+2)$ parameters $Q_{ab}$ and $\hat{Q}_{ab}$, one can construct the second-derivative matrix or Hessian of $\beta n \bar{f}_n$. The sign of the eigenvalues of this matrix evaluated at an extremal point will give us all the information about its stability.

The free energy second derivatives are

$$
\frac{\partial^2 (\beta n \bar{f}_n)}{\partial Q_{ab} \partial Q_{cd}} = \alpha \left(1 - \frac{\delta_{ab}}{2}\right) \left(1 - \frac{\delta_{cd}}{2}\right) (\langle \hat{x}^a \hat{x}^b \rangle \langle \hat{x}^c \hat{x}^d \rangle - \langle \hat{x}^a \hat{x}^b \hat{x}^c \hat{x}^d \rangle) 
- \beta \sum_{a=1}^{n} x^a - (1/2) \sum_{a,b=0}^{n} Q_{ab} x^a x^b,
$$

where

$$
\langle g(x^a, x^b, \ldots) \rangle = \frac{\int D\eta \prod_{a=0}^{n} (dx^a dx^a / 2\pi) g(x^a, x^b, \ldots) e^{-\beta \sum_{a=1}^{n} \theta(-x^a(0^+ + \eta)) + i \sum_{a=0}^{n} x^a - (1/2) \sum_{a,b=0}^{n} Q_{ab} x^a x^b}}{\int D\eta \prod_{a=0}^{n} (dx^a dx^a / 2\pi) e^{-\beta \sum_{a=1}^{n} \theta(-x^a(0^+ + \eta)) + i \sum_{a=0}^{n} x^a - (1/2) \sum_{a,b=0}^{n} Q_{ab} x^a x^b}}
$$

$$
\langle g(J^a, J^b, \ldots) \rangle = \frac{\int \prod_{a=1}^{n} dJ^a \int dJ^0 \rho(J^0) g(J^a, J^b, \ldots) e^{-h \sum_{a=1}^{n} |J^a|^p - h \sum_{a=1}^{n} \sum_{b=1}^{n} Q_{ab} J^a J^b}}{\int \prod_{a=1}^{n} dJ^a \int dJ^0 \rho(J^0) e^{-h \sum_{a=1}^{n} |J^a|^p - h \sum_{a=1}^{n} \sum_{b=1}^{n} Q_{ab} J^a J^b}}.
$$
The structure of the Hessian matrix is

$$
\|\partial^2(\beta n M_{ab})\| = \begin{pmatrix}
\left|\frac{\partial^2(\beta n M_{ab})}{\partial Q_{ab}\partial Q_{cd}}\right|
& \left|\frac{\partial^2(\beta n M_{ab})}{\partial Q_{cd}\partial(-iQ_{cd})}\right|

\left|\frac{\partial^2(\beta n M_{ab})}{\partial Q_{cd}\partial(-iQ_{cd})}\right|
& \left|\frac{\partial^2(\beta n M_{ab})}{\partial(-iQ_{ab})\partial(-iQ_{cd})}\right|
\end{pmatrix}.
\tag{A.1}
$$

In order to find the eigenvalues we need to determine first the eigenvalues of the inner matrices. The matrices $\|\partial^2(\beta n M_{ab})/\partial Q_{ab}\partial(-iQ_{cd})\|$ and $\|\partial^2(\beta n M_{ab})/\partial Q_{cd}\partial(-iQ_{cd})\|$ are identity matrices of order $(n+1)(n+2)/2$. The matrices $\|\partial^2(\beta n M_{ab})/\partial Q_{cd}\partial Q_{cd}\|$ and $\|\partial^2(\beta n M_{ab})/\partial(-iQ_{ab})\partial(-iQ_{cd})\|$ are of the same order, but each of them has only seven different elements: $A, B, G, H, M, N, R$ and $\dot{A}, \dot{B}, \dot{G}, \dot{H}, \dot{M}, \dot{N}, \dot{R}$ respectively, depending on their position on the matrices determined by the values of the pair of replica indices $ab$ and $cd$. The distribution of these elements in the matrices is the same; this implies that their eigenvectors are the same too.

We begin by looking for the eigenvalues of $\|\partial^2(\beta n M_{ab})/\partial Q_{ab}\partial Q_{cd}\|$. To find them we use another property that these matrices share: the columns can be divided in two groups according to the value of the sum of the elements in a column. More precisely, the first $n$ columns have a common value for the sum of their elements, while the remaining columns have a different value but also common to them. For that reason we begin looking for the eigenvectors $v = (v_0, v_1, \ldots, v_{n-1})$ with the structure [13]

$$
v_{ab} = x, \quad a = 0 \quad v_{ab} = y, \quad a > 0
$$

that have the first $n$ components equal to $x$ and the others equal to $y$. There are two nondegenerate eigenvalues associated with such eigenvectors. It is easy to see that these eigenvectors are RS in the sense that they do not take the $Q_{ab}$ and $Q_{cd}$ matrices away from the RS subspace (they keep the structure (7)). The reason is that the change in the first $n$ components, corresponding to the teacher–student overlap $r$, is the same for each student, and the change in the rest of the components, corresponding to the student–student overlap, is the same for all pairs of students too. Consequently, the eigenvalues associated with these eigenvectors do not determine instabilities of the RS saddle point.

To take away the RS saddle point from the RS subspace at least one of the replicas must be special (its overlap with the teacher and the rest of the replicas must be different from the other overlaps). Let us suppose that effectively, one replica is special, for instance the one with index $c$. The corresponding eigenvectors have the form

$$
v_{ab} = x, \quad a = 0 \land b = c
$$
$$
v_{ab} = y, \quad a = 0 \land b \neq c
$$
$$
v_{ab} = z, \quad a > 0 \land a = c \lor b = c
$$
$$
v_{ab} = w, \quad a > 0 \land a \neq c \land b \neq c.
$$

The corresponding eigenvalues are two in number, each $n - 1$ times degenerate. However, in the limit $n \to 0$ they tend to the eigenvalues of the RS eigenvectors, so they do not determine instabilities of the RS saddle point either.
Stability of the replica symmetric solution in diluted perceptron learning

The rest of the eigenvectors can be found by assuming that two replicas are special, \(c\) and \(d\):

\[
\begin{align*}
\mathbf{v}_{ab} &= x, \quad a = 0 \land b = c \lor b = d \\
\mathbf{v}_{ab} &= y, \quad a = 0 \land b \neq c \land b \neq d \\
\mathbf{v}_{ab} &= z, \quad a > 0 \land a = c \land b = d \\
\mathbf{v}_{ab} &= w, \quad a > 0 \land a \neq c \land b \neq d \land a \neq d \land b \neq c.
\end{align*}
\]

The corresponding eigenvalue is

\[
\gamma_1 = -2M + N + R,
\]

\(n(n-1)/2\) times degenerate with

\[
\begin{align*}
M &= \alpha \left( \langle \hat{x}^a x^b \rangle^2 - \langle \hat{x}^a x^a \hat{x}^c x^d \rangle \right) \\
N &= \alpha \left( \langle \hat{x}^a x^b \rangle^2 - \langle \hat{x}^a x^a x^b \hat{x}^b \rangle \right) \\
R &= \alpha \left( \langle \hat{x}^a x^b \rangle^2 - \langle \hat{x}^a x^b x^c x^d \rangle \right).
\end{align*}
\]

This eigenvector is different from the preceding ones even in the \(n \to 0\) limit and is the one related to instabilities of the RS saddle point. In a similar way, the matrix

\[
\| \partial^2 (\beta n f_n) / \partial (-i \hat{Q}_{ab}) \partial (-i \hat{Q}_{cd}) \|
\]

has a non-RS eigenvalue:

\[
\gamma_2 = -2M + \tilde{N} + \tilde{R}
\]

with

\[
\begin{align*}
\tilde{M} &= \langle \hat{J}^a \hat{J}^b \rangle^2 - \langle \hat{J}^a \hat{J}^a \hat{J}^c \hat{J}^d \rangle \\
\tilde{N} &= \langle \hat{J}^a \hat{J}^b \rangle^2 - \langle \hat{J}^a \hat{J}^a \hat{J}^b \hat{J}^b \rangle \\
\tilde{R} &= \langle \hat{J}^a \hat{J}^b \rangle^2 - \langle \hat{J}^a \hat{J}^b \hat{J}^c \hat{J}^d \rangle.
\end{align*}
\]

Using the RS ansatz and taking into account the elements of the matrices involved, one can reach expressions (11) from expressions (A.2) and (A.6).

In accordance with [13, 14, 22], the eigenvector space of the Hessian matrix can be divided in two—one subspace corresponding to instabilities inside the replica symmetric space (and whose eigenvalue signs are those determining, among all the possible replica symmetric solutions, the one which actually minimizes the free energy), and the other corresponding to instabilities that take our extremal point outside the replica symmetric space. The first of these subspaces has no interest for our stability analysis, because solving the replica symmetric fixed point equations is equivalent to searching for the stable extremal replica symmetric point, so we are going to focus on the eigenvalues of the eigenvectors belonging to the second subspace. Given an eigenvector \(\mathbf{v}\) of the inner matrices, it is easy to show that the eigenvectors of the Hessian matrix are a linear combination of the vectors \(\mathbf{v}, -\mathbf{v}\) and \(\mathbf{v}, -\mathbf{v}\) [13]. The corresponding eigenvalues of the Hessian matrix are related to those of the inner matrices in such a way that their product can be expressed as in equation (10). \(\delta_1\) and \(\delta_2\) are not the unique eigenvalues of the Hessian matrix, but they are the ones related to the non-RS eigenvalues of the inner matrices.

doi:10.1088/1742-5468/2013/02/L02002
Stability of the replica symmetric solution in diluted perceptron learning

References

[1] Hastie T, Tibshirani R and Friedman J, 2009 The Elements of Statistical Learning (Berlin: Springer)
[2] Braunstein A and Zecchina R, Learning by message passing in networks of discrete synapses, 2006 Phys. Rev. Lett. 96 030201
[3] Braunstein A, Pagani A, Weigt M and Zecchina R, Inference algorithms for gene networks: a statistical-mechanics analysis, 2008 J. Stat. Mech. P12001
[4] Pagani A, Tria F and Weigt M, Classification and sparse-signature extraction from gene-expression data, 2009 J. Stat. Mech. P05001
[5] Kabashima Y, Wadayama T and Tanaka T, A typical reconstruction limit for compressed sensing based on l 1-p-norm minimization, 2009 J. Stat. Mech. L00003
[6] Wadayama T, Kabashima Y and Tanaka T, Statistical mechanics analysis of a typical reconstruction limit of compressed sensing, 2010 2010 IEEE Information Theory Proc. pp 1533–7
[7] Krzakala F, Mézard M, Sausset F, Sun Y F and Zdeborová L, Statistical-physics-based reconstruction in compressed sensing, 2012 Phys. Rev. X 2 021005
[8] Tibshirani R, Regression shrinkage and selection via the lasso, 1994 J. R. Stat. Soc. B 58 267
[9] Lage-Castellanos A, Pagani A and Weigt M, Statistical mechanics of sparse generalization and graphical model selection, 2009 J. Stat. Mech. P10009
[10] Lage-Castellanos A, Pagani A and Weigt M, The importance of dilution in the inference of biological networks, 2009 47th Annual Allerton Conf. on Communication, Control and Computing (Monticello, IL, Sept. 30–Oct. 02, 2009) vols 1 and 2 (Piscataway, NJ: IEEE) pp 531–8
[11] Derényi I, Geszti T and Győrgyi G, Generalization in the programed teaching of a perceptron, 1994 Phys. Rev. E 50 3192
[12] Vallet F, The Hebb rule for learning separable Boolean functions: learning and generalization, 1989 Europhys. Lett. 8 747
[13] Engel A and Van der Broeck C, 2001 Statistical Mechanics of Learning (Cambridge: Cambridge University Press)
[14] Gardner E and Derrida B, Optimal storage properties of neural network models, 1988 J. Phys. A: Math. Gen. 21 271
[15] Gardner E, The space of interactions in neural network models, 1988 J. Phys. A: Math. Gen. 21 257
[16] Guyon I and Elisseeff A, An introduction to variable and feature selection, 2003 J. Mach. Learn. Res. 3 1157
[17] Guyon I, Gunn S, Nikravesh M and Zadeh L A, 2006 Feature Extraction: Foundations and Applications vol 207 (Berlin: Springer)
[18] Komoda A, Boutent M, Engel A and Serneels R, Quenched versus annealed dilution in neural networks, 1990 J. Phys. A: Math. Gen. 23 4643
[19] Donoho D L, Compressed sensing, 2006 IEEE Trans. Inf. Theory 52 1289
[20] Mézard M, Parisi G and Virasoro M A, 1987 Spin Glass Theory and Beyond (Singapore: World Scientific)
[21] Bolle D and van Mourik J, Capacity of diluted multi-state neural networks, 1994 J. Phys. A: Math. Gen. 27 1151
[22] de Almeida J R L and Thouless D J, Stability of the Sherrington–Kirkpatrick solution of a spin glass model, 1978 J. Phys. A: Math. Gen. 11 983

doi:10.1088/1742-5468/2013/02/L02002