Can a student learn optimally from two different teachers?

J. Phys. A: Math. Theor. **43** (2010) 015101 (16pp) doi:10.1088/1751-8113/43/1/015101

Can a student learn optimally from two different teachers?

J P Neirotti

The Neural Computing Research Group, Aston University, Birmingham, UK

E-mail: j.p.neirotti@aston.ac.uk

Received 2 July 2009, in final form 14 October 2009
Published 7 December 2009
Online at stacks.iop.org/JPhysA/43/015101

Abstract

We explore the effects of over-specificity in learning algorithms by investigating the behavior of a student, suited to learn optimally from a teacher $\mathbf{B}$, learning from a teacher $\mathbf{B}' \neq \mathbf{B}$. We only considered the supervised, on-line learning scenario with teachers selected from a particular family. We found that, in the general case, the application of the optimal algorithm to the wrong teacher produces a residual generalization error, even if the right teacher is harder. By imposing mild conditions to the learning algorithm form, we obtained an approximation for the residual generalization error. Simulations carried out in finite networks validate the estimate found.

PACS numbers: 89.70.Eg, 84.35.+i, 87.23.Kg

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Neural networks are connectivist models inspired by the dynamical behavior of the brain [1]. They are not only theoretically interesting models, they can also be used in a number of applications, from voice recognition systems to curve fitting software. Some of the properties that make neural networks most useful are probably their potential to store patterns and their capability for learning tasks.

One of the most well-studied types of networks is feed-forward. What characterizes a feed-forward network is that the flux of information follows a non-loopy path from input to output nodes, making the information processing much faster. Perceptrons [2] are feed-forward networks with no internal nodes and only one output; they have been utilized for a number of theoretical studies and applications of statistical mechanics techniques [3]. In particular, the knowledge of Hebbian learning algorithms in an online scenario is quite complete.

Due to their simplicity, perceptrons are excellent systems to test new ideas that could lead to applications for more sophisticated and realistic systems. This has probably been the main motivation for the research focused on a mismatched student–teacher scenario, like when learning from a noisy teacher [4, 5]. This scenario has recently revisited and extended to the situation of a student learning from two teachers [6, 7]. The common factor in all these studies is that the student's learning algorithm is meant to be used for learning from a *generic* teacher (represented mostly by an N-dimensional vector randomly selected from the sphere of radius $\sqrt{N}$). Our aim, in contrast, was to study the mismatched situation where the student applies an algorithm specific for a particular teacher, to learn from a different one. In this situation, it is natural to ask if a student prepared to learn from a *difficult* teacher would be able to learn from an *easier* one. To formally analyze this problem we need to quantify the hardness of the teachers, set up the scenario where the learning process would take place and thus quantify the student's performance.

Attempts to quantify hardness as an inherent property of the observed object have given origin to many formal definitions of complexity [8–12]. Recently [13] L Franco has proposed to quantify a (Boolean) function's hardness by the size of the minimal set of examples needed to train a feed-forward network, with a predetermined architecture until reaching zero prediction error. He also found [14, 15] that in this minimal set there are many pairs of examples that, although only differing in a finite number $P = 1, 2, \ldots$ of entries, have different outputs, implying that these examples are located at each side of the classification boundary (similar to the support vectors for SVMs [16]). Further investigation showed that the average discrepancy of the function's outputs (measure over neighboring pairs) is correlated to the generalization ability of the network implementing the function. In order to contour the use of the neural network and its minimal training set, Franco proposed to use the average distance sensitivity directly as a measure of the function's hardness. This is probably the most suitable measure for our study given that the nature of the measure itself is linked to the concept of generalization ability.

In this paper, the systems we studied the most were diluted perceptrons. These systems have been widely studied using statistical mechanics techniques [17–22] and have also been used for approximating Boolean functions [23, 24]. In general, dilution gives rise to networks with fewer connections, which can be more efficient in solving tasks and can be more easily implemented in hardware. Some of the most important features of diluted perceptrons related to the present work are the existence of analytical expressions for the sensitivity component (1) and the associated optimal learning algorithm (see below).

The paper is organized as follows. In section 2, we introduce the concepts used for obtaining our results. In section 3, we present an analytical study of the mismatched scenario. Our main result is expression (17) where we present the maximum overlap between student and teacher as a function of the teacher's true dilution parameter $m_{\mathbf{B}}$ and m, the dilution parameter the student supposes the teacher presents, for sufficiently large values of the overlap R. In section 4, we present numerical results that validate our analytical expressions, and in section 5, we present our conclusions and future work.

2. Background

2.1. Measure of hardness

In the present paper, we study the ability of a student $\mathbf{J}$, using an algorithm specific for teacher $\mathbf{B}$, to learn from teacher $\mathbf{B}'$. The hardness measure we will use is the average output discrepancy taken over all pairs of inputs at a given Hamming distance P. Formally, for a given

Boolean function $f : \{\pm 1\}^N \to \{\pm 1\}$, the Pth distance sensitivity component $\mathfrak{d}_P^N[f]$ is the functional

$$\mathfrak{d}_P^N[f] = 2^{-N} \sum_{\mathbf{S}\in\{\pm 1\}^N} \binom{N}{P}^{-1} \sum_{\mathbf{S}'\in\Omega_P(\mathbf{S})} \frac{1 - f(\mathbf{S}) f(\mathbf{S}')}{2}, \tag{1}$$

where $\binom{N}{P}$ is the binomial factor and $\Omega_P(\mathbf{S}) = \left\{\mathbf{S}' \in \{\pm 1\}^N \,|\, \sum_{j=1}^N \Theta(-S_j S_j') = P\right\}$, where $\Theta(x) = 1$ iff $x \geqslant 0$ and 0 otherwise. $\Omega_P(\mathbf{S})$ is the set of inputs $\mathbf{S}'$ that differ from $\mathbf{S}$ in P entries.

Consider a perceptron characterized by a synaptic vector $\mathbf{B}^{(m)} \in \mathbb{R}^N$ that classifies binary vectors $\mathbf{S} \in \{\pm 1\}^N$ with labels $\sigma_\mathbf{B} \in \{\pm 1\}$ according to the rule $\sigma_\mathbf{B} = \mathrm{sgn}(\mathbf{B}^{(m)} \cdot \mathbf{S})$. If the i-th entry of the synaptic vector $[\mathbf{B}^{(m)}]_i = \delta(i \in \mathbb{I}_m)\, O(\sqrt{N/m}) + \delta(i \notin \mathbb{I}_m)\, o(\sqrt{m/N})$ where $\mathbb{I}_m \subset \{1, 2, \ldots, N\}$ is a set of m (odd) different indexes $1 \leqslant i \leqslant N$ and $\delta(i \in \mathbb{I}_m) = 1$ iff $i \in \mathbb{I}_m$ and 0 otherwise, then $\mathbf{B}^{(m)}$ is a *diluted binary perceptron*. In our calculations, we will consider $[\mathbf{B}^{(m)}]_i = \delta(i \in \mathbb{I}_m)\sqrt{N/m}$ where $m \ll N$ will be kept finite.

For the binary perceptron $\mathbf{B}^{(m)}$, the distance sensitivity component (1) in the large system limit ($P < N \to \infty$ with $p \equiv P/N < \infty$) $\mathfrak{d}^{(m)}(p)$ is given by (A.8)

$$\mathfrak{d}^{(m)}(p) = \frac{1}{2} - \frac{1}{2} \sum_{n=0}^{(m-1)/2} a_n^m (1 - 2p)^{2n+1}$$

$$a_n^m = \frac{1}{4^{m-1}} \binom{m}{2n+1} \left[\binom{2n}{n} \binom{m-1-2n}{(m-1)/2 - n} \binom{(m-1)/2}{n}^{-1} \right]^2.$$

As is shown in Appendix A, and following [24], $\mathfrak{d}^{(m)}(p)$ are a family of concave functions, ordered according to $\mathfrak{d}^{(m)}(p) < \mathfrak{d}^{(m+2)}(p)\ \forall p \in (0, \frac{1}{2})$. Therefore, the order given by the hardness measure coincides with the order given by m; thus the larger m the harder the Teacher.

2.2. *Supervised, online learning*

Another reason that appeals for using a diluted perceptron as a teacher is that it is possible to obtain the correspondent optimal learning algorithm analytically. In a supervised online learning scenario, the synaptic vector of the student perceptron $\mathbf{J}$ is adjusted after receiving new information in the form of the pair $(\mathbf{S}, \sigma_\mathbf{B})$, following the rule

$$\mathbf{J}_{\mathrm{new}} = \mathbf{J}_{\mathrm{old}} + F \frac{\sigma_{\mathbf{B}\mathrm{new}} \mathbf{S}_{\mathrm{new}}}{\sqrt{N}}, \tag{2}$$

where $\mathbf{J} \in \mathbb{R}^N$, $\sigma_\mathbf{B} = \mathrm{sgn}(\mathbf{B} \cdot \mathbf{S})$ is the classification given to the example by the teacher $\mathbf{B}$, F is the learning amplitude or algorithm and $\mathbf{S}$ is an input vector with entries drawn from the distribution $\mathcal{P}_\mathbf{S}(S_i = 1) = 1 - \mathcal{P}_\mathbf{S}(S_i = -1) = \frac{1}{2}$. The parameters of the problem are

$$h \equiv \frac{\mathbf{J} \cdot \mathbf{S}}{|\mathbf{J}|}, \qquad b \equiv \frac{\mathbf{B} \cdot \mathbf{S}}{|\mathbf{B}|}, \qquad Q \equiv \frac{\mathbf{J} \cdot \mathbf{J}}{N}, \qquad R \equiv \frac{\mathbf{B} \cdot \mathbf{J}}{|\mathbf{B}||\mathbf{J}|}$$

where h is known as the student's post-synaptic field, b is the teacher's post-synaptic field, i.e. $\mathrm{sgn}(b) = \sigma_\mathbf{B}$, Q is the normalized length of $\mathbf{J}$ and R is the overlap between teacher and student.

Following [3], we found that the equation of motion for the overlap R in terms of the total number of examples received αN, in the large size limit $N \to \infty$, is

$$\frac{\mathrm{d}R}{\mathrm{d}\alpha} = \left\langle \frac{F}{\sqrt{Q}} [\langle |b| \rangle_{b|\phi} - R\phi] - \frac{RF^2}{2Q} \right\rangle_\phi, \tag{3}$$

where $\langle\cdot\rangle_\phi$ represents an average over the distribution $\mathcal{P}(\phi)$ and $\phi \equiv \sigma_{\mathbf{B}} h$. The solution of this equation represents the evolution of the overlap R as a function of the *time* α.

The generalization error is defined as the average over a set $\mathscr{L}$ of αN-sequences of examples $e_g(\alpha) = \langle\Theta(-\sigma_{\mathbf{J}}\sigma_{\mathbf{B}})\rangle_{\mathscr{L}}$, which in the typical case, i.e. when the teacher is drawn from a uniform distribution over the N-dimensional sphere, can be expressed as $e_g(\alpha) = \arccos(R)/\pi$. We define the residual error as the asymptotic value of the learning curve at large values of α, i.e. $e_g^\star = \lim_{\alpha\to\infty} e_g(\alpha)$.

By the application of a variational technique, it is possible to obtain an expression for the optimal algorithm F_{op}. The optimal algorithm is the algorithm that produces the fastest decaying learning curve and can be generically expressed as

$$F_{\mathrm{op}} = \frac{\sqrt{Q}}{R}[\langle|b|\rangle_{b|\phi} - R\phi],$$

where $\langle\,\cdot\,\rangle_{b|\phi}$ is the average over the conditional distribution $\mathcal{P}(b|\phi)$.

3. Analytical results

Following [26], we can prove (see Appendix B) that, for a perceptron with dilution m

$$\mathcal{P}(\phi|m) = \frac{1}{2^{m-1}}\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mathcal{N}(\phi|R\mu_k, 1-R^2) \tag{4a}$$

$$\langle|b|\rangle_{b|\phi,m} = \frac{\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mu_k\,\mathcal{N}(\phi|R\mu_k, 1-R^2)}{\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mathcal{N}(\phi|R\mu_k, 1-R^2)} \tag{4b}$$

$$F_{\mathrm{op}}^{(m)} = \frac{\sqrt{Q}}{R}\left[\frac{\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mu_k\,\mathcal{N}(\phi|R\mu_k, 1-R^2)}{\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mathcal{N}(\phi|R\mu_k, 1-R^2)} - R\phi\right], \tag{4c}$$

where $\mu_k = (2k+1)/\sqrt{m}$ and $\mathcal{N}(x|\mu,\sigma^2)$ is a normal distribution centered at μ with variance σ^2. Observe that (4*b*) is needed for computing the evolution (3), and (4*c*) represents the optimal learning algorithm.

Suppose that the teacher is characterized by a dilution $m_{\mathbf{B}}$ and the student implements an algorithm (4*c*) for learning a Teacher perceptron with dilution m. This is equivalent to having prepared a student to learn optimally from $\mathbf{B}^{(m)}$ and now exposing it to $\mathbf{B}^{(m_{\mathbf{B}})} \neq \mathbf{B}^{(m)}$. Let us define the quantity

$$\Upsilon(\phi|R,m) \equiv \langle|b|\rangle_{b|\phi,m} - R\phi. \tag{5}$$

In this settings, the algorithm has the form $F^{(m)} = \frac{\sqrt{Q}}{R}\Upsilon(\phi|R,m)$ and the distribution of ϕ is a function of $m_{\mathbf{B}}$. The evolution of the overlap R is given now by the equation (3)

$$\frac{\mathrm{d}R}{\mathrm{d}\alpha} = \left\langle\frac{1}{R}\Upsilon(\phi|R,m)\Upsilon(\phi|R,m_{\mathbf{B}}) - \frac{1}{2R}\Upsilon^2(\phi|R,m)\right\rangle_{\phi|m_{\mathbf{B}}}$$

which can be reduced to

$$\frac{\mathrm{d}R^2}{\mathrm{d}\alpha} = 2\langle\Upsilon(\phi|R,m)\,\Upsilon(\phi|R,m_{\mathbf{B}})\rangle_{\phi|m_{\mathbf{B}}} - \langle\Upsilon^2(\phi|R,m)\rangle_{\phi|m_{\mathbf{B}}} \tag{6a}$$

$$= \langle\Upsilon^2(\phi|R,m_{\mathbf{B}})\rangle_{\phi|m_{\mathbf{B}}} - \langle[\Upsilon(\phi|R,m_{\mathbf{B}}) - \Upsilon(\phi|R,m)]^2\rangle_{\phi|m_{\mathbf{B}}}. \tag{6b}$$

The overlap R grows from zero to a stationary value; thus, we expect the second term at the rhs of (6*b*) to be smaller than the first one. In the asymptotic regime ($\alpha \to \infty$), the derivative

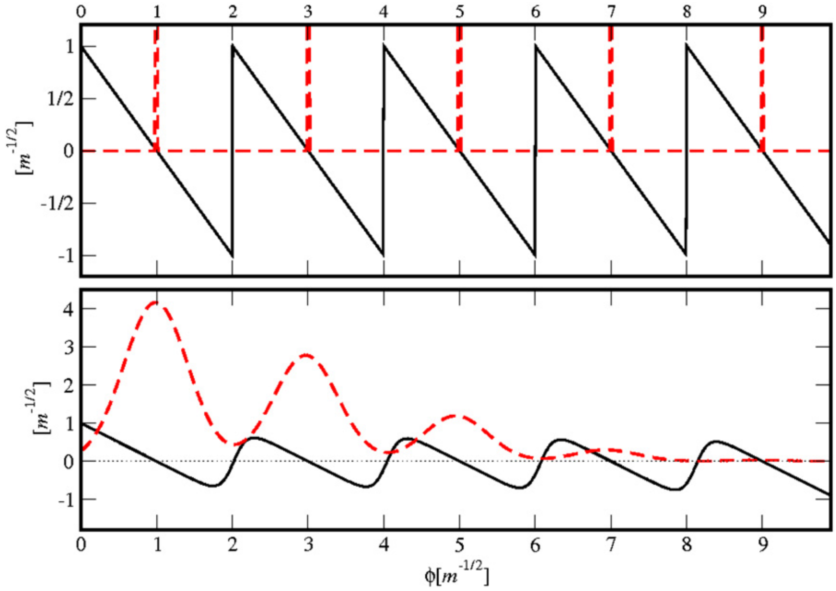

Figure 1. $\Upsilon(\phi|R, m)$ (full curve) and the probability of the stability $\mathcal{P}(\phi|m)$ (dashed curve) against ϕ in units of $1/\sqrt{m}$ for $R = 1$ (upper panel) and $R = 0.99$ (lower panel) for $m = 9$. Observe that for $R = 1$ (upper panel), the average of the LHS (7) involves only the points at which $\Upsilon(\phi|1, m)$ is zero, whilst for $R < 1$ (lower panel) the same average requires a more intensive calculation.

is zero, implying that no further changes are expected in the overlap, and then we have that

$$\langle \Upsilon^2(\phi|R^\star, m_{\mathbf{B}})\rangle_{\phi|m_{\mathbf{B}}} = \langle [\Upsilon(\phi|R^\star, m) - \Upsilon(\phi|R^\star, m_{\mathbf{B}})]^2\rangle_{\phi|m_{\mathbf{B}}} \tag{7}$$

where $R^\star \equiv \lim_{\alpha\uparrow\infty} R(\alpha)$.

Observe that if $m = m_{\mathbf{B}}$, the second term of the rhs of (6*a*) is zero, the algorithm applied is optimal and the overlap reaches $R^\star = 1$ with the smallest possible set of examples. If perfect learning implies $R^\star = 1$, it is natural to ask for what values of m the student can learn a teacher with dilution $m_{\mathbf{B}}$ without errors. From (4*a*) and (4*b*), we have that, for $R = 1$,

$$\mathcal{P}(\phi|m_{\mathbf{B}}) = \frac{\sqrt{m_{\mathbf{B}}}}{2^{m_{\mathbf{B}}-1}} \sum_{k=0}^{(m_{\mathbf{B}}-1)/2} \binom{m_{\mathbf{B}}}{(m_{\mathbf{B}}-1)/2-k} \delta(\sqrt{m_{\mathbf{B}}}\,\phi - (2k+1)) \tag{8a}$$

$$\Upsilon(\phi|1, m_{\mathbf{B}}) = \frac{1}{\sqrt{m_{\mathbf{B}}}}\left[1 + \sum_{k=1}^{(m_{\mathbf{B}}-1)/2} \Theta(\sqrt{m_{\mathbf{B}}}\,\phi - 2k)\right] - \phi \tag{8b}$$

$$\Upsilon(\phi|1, m) = \frac{1}{\sqrt{m}}\left[1 + \sum_{k=1}^{(m-1)/2} \Theta(\sqrt{m}\,\phi - 2k)\right] - \phi. \tag{8c}$$

The lhs of (7), averaged over (8*a*) is zero (see figure 1). This is due to the fact that $\Upsilon((2k+1)/\sqrt{m_{\mathbf{B}}}|1, m_{\mathbf{B}}) = 0$. Therefore, in order to satisfy (7), we also need that $\Upsilon((2k+1)/\sqrt{m_{\mathbf{B}}}|1, m) = 0$. Particularly, for $k = 0$ these two equation imply that

$$\sqrt{\frac{m}{m_{\mathbf{B}}}} = 1 + \sum_{k=1}^{(m-1)/2} \Theta\left(\sqrt{\frac{m}{m_{\mathbf{B}}}} - 2k\right).$$

Therefore

$$\sqrt{\frac{m}{m_{\mathbf{B}}}} = 2q + 1, \tag{9}$$

where q is a suitable, non-negative integer. Thus, the condition for $R = 1$ to be a solution of (7) is that there exist $q \in \mathbb{N} \cup \{0\}$ such that $m = (2q + 1)^2 m_{\mathbf{B}}$.

If this is not true, the solution of (7) is at $R^\star < 1$. We will present an approach based on the assumption that the root $R^\star$ occurs in a regime where the Gaussian distributions $\mathcal{N}(\phi|R^\star\mu_k, 1 - R^{\star 2})$ in (4*a*) and (4*b*) have a small overlap. This could be ensured if the separation of two adjacent Gaussian components was larger than two standard deviations, i.e.

$$R^\star|\mu_k - \mu_{k+1}| = \frac{2R^\star}{\sqrt{m}} \gg 2\sqrt{1 - R^{\star 2}} \tag{10a}$$

$$1 \gg m\,\frac{1 - R^2}{R^2} \tag{10b}$$

At $R = 1$, the curve $\Upsilon(\phi|1, m)$ is discontinuous at $\phi \equiv 2k/\sqrt{m}$ and the probability $\mathcal{P}(\phi|m)$ is a linear combination of delta functions centered at $\phi = (2k + 1)/\sqrt{m}$ (upper panel of figure 1). For $R < 1$ (figure 1, lower panel), $\Upsilon(\phi|R, m)$ is continuous and $\mathcal{P}(\phi|m)$ is a linear combination of Gaussian distributions centered at $\phi = R(2k + 1)/\sqrt{m}$ with variance $1 - R^2$. In both cases, $\Upsilon(\phi|R, m)$ appears to be a periodic function of ϕ with period $\phi_T \equiv 2R/\sqrt{m}$, in the support of $\mathcal{P}(\phi|m)\mathscr{D}_\phi \subset \mathbb{R}$, i.e.

$$\Upsilon(\phi|R, m) \simeq \Upsilon(\phi + n\phi_T|R, m),$$

and particularly for $R = 1$, we have that

$$\Upsilon(\phi|1, m) = \sum_{\ell=0}^{(m-1)/2} \Theta[(2(\ell + 1) - \sqrt{m}\phi)(\sqrt{m}\phi - 2\ell)]\left(\frac{2\ell + 1}{\sqrt{m}} - \phi\right). \tag{11}$$

We can approximate $\Upsilon(\phi|R, m)$ by a suitable superposition of normal distributions. Consider the superposition

$$\tilde{\Upsilon}(\phi|R, m) \equiv \int_{-\infty}^{\infty} \mathrm{d}r\,g(r)\mathcal{N}(\phi|r, 1 - R^2). \tag{12}$$

To determine the function $g(r)$, we perform a variational calculation to minimize the error functional

$$\varepsilon[g] \equiv \frac{1}{2}\int_{\mathscr{D}_\phi} \mathrm{d}\phi[\Upsilon(\phi|R, m) - \tilde{\Upsilon}(\phi|R, m)]^2.$$

Observe that the optimal function $g_o(r)$ is the solution of the equation $\frac{\delta\varepsilon}{\delta g}|_{g_o} = 0$, which implies that for all $r_0 \in \mathbb{R}$ we have that

$$\int_{\mathscr{D}_\phi} \mathrm{d}\phi[\Upsilon(\phi|R, m) - \tilde{\Upsilon}(\phi|R, m)]\mathcal{N}(\phi|r_0, 1 - R^2) = 0,$$

in particular if $R = 1$ (we assume that $g_o(r)$ is independent of R)

$$0 = \int_{\mathscr{D}_\phi} \mathrm{d}\phi\left[\Upsilon(\phi|1, m) - \int_{-\infty}^{\infty} \mathrm{d}r\,g_o(r)\delta(\phi - r)\right]\delta(\phi - r_0)$$
$$g_o(r_0) = \Upsilon(r_0|1, m).$$

Therefore

$$\begin{aligned}\Upsilon(\phi|R,m) &\simeq \int_{\mathscr{D}_\phi} \mathrm{d}r\,\Upsilon(r|1,m)\,\mathcal{N}(\phi|r,1-R^2)\\ &= \sum_{\ell=0}^{(m-1)/2}\int_{2\ell R/\sqrt{m}}^{2(\ell+1)R/\sqrt{m}} \mathrm{d}r\,\left[(2\ell+1)\frac{R}{\sqrt{m}}-r\right]\mathcal{N}(\phi|r,1-R^2)\\ &= \frac{R^2}{m}\int_{-1}^{1}\mathrm{d}t\,t\sum_{\ell=0}^{(m-1)/2}\mathcal{N}(\phi|R(2\ell+1-t)/\sqrt{m},1-R^2). \qquad (13)\end{aligned}$$

Let us define the integral

$$\mathscr{I}_{m_1,m_2} \equiv \int \mathrm{d}\phi\,\mathcal{P}(\phi|m_{\mathbf{B}})\,\Upsilon(\phi|R,m_1)\,\Upsilon(\phi|R,m_2). \qquad (14)$$

Following the development of Appendix C, we have that

$$\mathscr{I}_{m_1,m_2} \simeq 1-R^2\left[1-\frac{1}{2^{m_{\mathbf{B}}-1}}\sum_{k=0}^{(m_{\mathbf{B}}-1)/2}\binom{m_{\mathbf{B}}}{(m_{\mathbf{B}}-1)/2-k}\delta^\star_{m_1,k}\delta^\star_{m_2,k}\right], \qquad (15)$$

where $\delta^\star_{m_j,k}$ are given by

$$\delta^\star_{m_j,k} = \frac{2}{\sqrt{m_j}}\left[\!\left[\sqrt{\frac{m_j}{m_{\mathbf{B}}}}\frac{2k+1}{2}-\frac{1}{2}\right]\!\right]+\frac{1}{\sqrt{m_j}}-\frac{2k+1}{\sqrt{m_{\mathbf{B}}}}, \qquad (16)$$

where $[\![r]\!]$ is the closest integer to $r \in \mathbb{R}$.

Observe that from (6*a*), we have that in the asymptotic regime

$$\begin{aligned}0 &= 2\langle\Upsilon(\phi|R^\star,m_{\mathbf{B}})\Upsilon(\phi|R^\star,m)\rangle_{\phi|m_{\mathbf{B}}}-\langle\Upsilon^2(\phi|R^\star,m)\rangle_{\phi|m_{\mathbf{B}}}\\ &= 2\mathscr{I}_{m_{\mathbf{B}},m}-\mathscr{I}_{m,m},\end{aligned}$$

and observing that $\mathscr{I}_{m_{\mathbf{B}},m} = 1-R^2$ (given that $\delta^\star_{m_{\mathbf{B}},k} = 0\ \forall k$) and

$$\mathscr{I}_{m,m} = 1-R^2+\frac{R^2}{2^{m_{\mathbf{B}}-1}}\sum_{k=0}^{(m_{\mathbf{B}}-1)/2}\binom{m_{\mathbf{B}}}{(m_{\mathbf{B}}-1)/2-k}\delta^{\star 2}_{m,k}$$

then

$$R^{\star 2} = \left[1+\frac{1}{2^{m_{\mathbf{B}}-1}}\sum_{k=0}^{(m_{\mathbf{B}}-1)/2}\binom{m_{\mathbf{B}}}{(m_{\mathbf{B}}-1)/2-k}\delta^{\star 2}_{m,k}\right]^{-1} \qquad (17)$$

and observe that $\delta^\star_{m,k} = 0$ if $m = (2q+1)^2 m_{\mathbf{B}},\ q \in \mathbb{N}$, which is consistent with (9).

4. Numerical results

Using (17), we plot $e_g^\star = e_g(R^\star)$ as a function of $\sqrt{m/m_{\mathbf{B}}}$ (see figure 2).

To validate our result shown in (17), we run a series of numerical experiments consisting of a student learning from a Teacher with only one bit ($m_{\mathbf{B}} = 1$). The student updates its synaptic vector following (2) using a learning algorithm given by (4*c*) with $m = 1, 3, \ldots, N$. To compute the generalization error, we average over 50 realizations of the learning curve. The maximum number of examples considered was 16 000. In figure 3, we present the e_g as a function of $\alpha^{\frac{1}{4}}$ for $m = 1, 5, 9, 13, 25, 27$ and network size $N = 51$. We have chosen the exponent $\frac{1}{4}$ to better show the curve features at short times and the approach to the asymptotic regime. It is clear from the picture that for $m = 1^2, 3^2, 5^2$, the generalization error for large

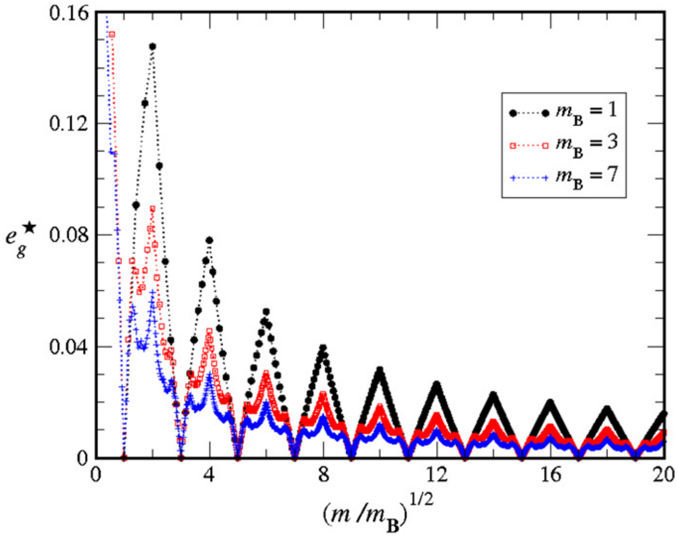

Figure 2. Generalization error in the asymptotic regime $e_g^\star$ as a function of $\sqrt{m/m_{\mathbf{B}}}$, for $m_{\mathbf{B}} = 1, 3, 7$. We have use (17) to compute the overlap $R^\star$.

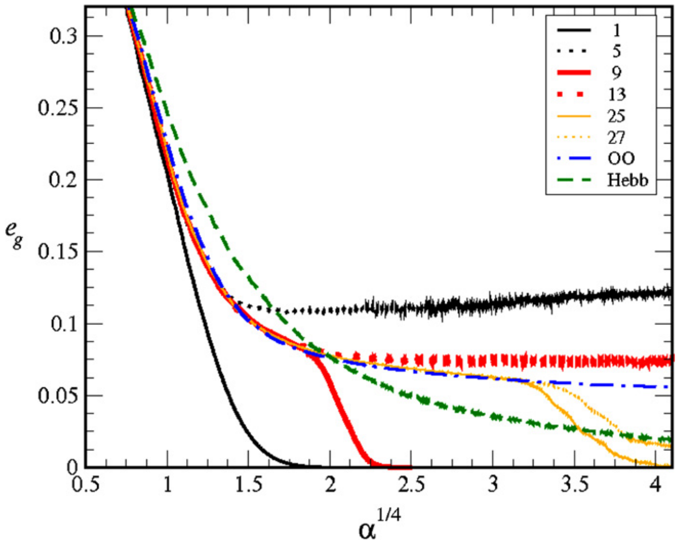

Figure 3. Generalization error as a function of $\alpha^{\frac{1}{4}}$, for a teacher with dilution $m_{\mathbf{B}} = 1$ and students with $m = 1, 5, 9, 13, 25, 27$, for a network with $N = 51$. The curves that corresponds to the Hebb algorithm ($F = 1$, long dashed) and $m = \infty$ (dot dashed) are presented as a reference.

α drops to zero as predicted. In order to extract the asymptotic behaviour of the curves, we applied the Bulirsch–Stoer algorithm [25].

In figure 4, we present the extrapolated values of the learning curves together with the values estimated by the application of (17) as a function of $\sqrt{m}$. The error bars are estimates obtained also by the application of the Bulirsch–Stoer algorithm.

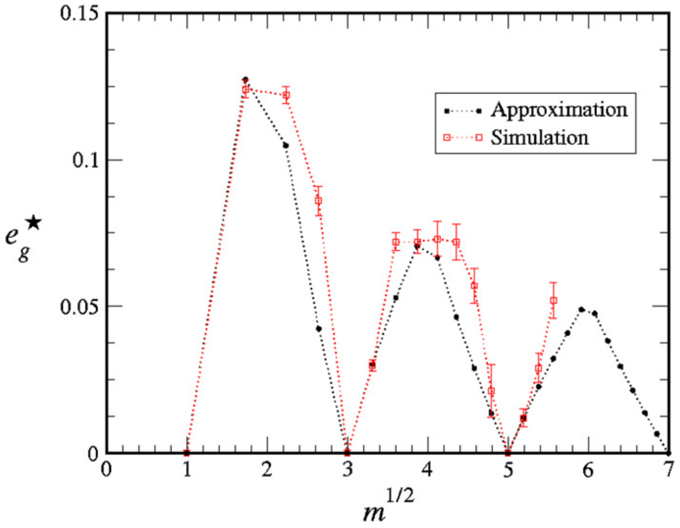

Figure 4. Comparison of the asymptotic value of the generalization error using (17) and the extrapolated values of the curvespresented in figure 3.

5. Conclusions

We studied the generalization capabilities of a student optimally adapted for learning from a teacher $\mathbf{B}$, when learning from a teacher $\mathbf{B}' \neq \mathbf{B}$. We observed that, although the algorithm the student uses may be suited for learning from a harder teacher (as defined by Franco), that does not guarantee the success of the process, as revealed by (17). This behavior is due to the extreme specialization implied by the algorithm (4*c*). When this algorithm (with parameter m) is applied to learn from a teacher with $m_{\mathbf{B}} < m$, the student tries to extract information from bits that the teacher does not use for producing the correct classification. These interference effects produce mostly bad results, originating a residual error in the asymptotic regime. In this sense, the algorithm $F_{\mathrm{op}}^{(m)}$ is worse than the Hebb algorithm $F_{\mathrm{Hebb}} = 1$.

Despite the discrepancies shown in figure 4, our estimate (17) faithfully reproduces the qualitative behaviour observed in the simulations. There are two sources of uncertainty that may account for the observed discrepancies: the (finite) size of the network used and a not sufficiently large α.

From figure 2, the algorithm obtained by taking the limit $m \to \infty$ in (4*c*)

$$F_{\mathrm{op}}^{(\infty)} = \sqrt{\frac{Q(1-R^2)}{2\pi R^2}}\,\frac{\exp\left(-\frac{1}{2}\frac{R^2\phi^2}{1-R^2}\right)}{\mathcal{H}\left(-\frac{R\phi}{\sqrt{1-R^2}}\right)},$$

where $\mathcal{H}(x) = \int_x^\infty \mathrm{d}y\,\mathrm{e}^{-\frac{y^2}{2}}/\sqrt{2\pi}$, as reported by [26], produces the zero residual error for all $m_{\mathbf{B}}$. The Hebb algorithm $F_{\mathrm{Hebb}} = 1$ also produces learning curves with the zero residual error. In figure 3, we observe that the Hebb algorithm performs better than $F_{\mathrm{op}}^{(\infty)}$. This is not a contradictory result. $F_{\mathrm{op}}^{(\infty)}$ is the algorithm that has the best average performance considering a homogeneous distribution of teachers over the N-sphere. For a measure zero subset of vectors embedded in the N-sphere, like the perceptrons with finite dilution m, $F_{\mathrm{op}}^{(\infty)}$ could perform worse than the Hebb algorithm, as seems to be the case here.

In order to obtain the fastest decaying learning curve, a student has to infer the correct dilution of the teacher for choosing the appropriate learning algorithm. Developing an efficient technique for inferring the correct dilution parameter will be the subject of our future research.

Acknowledgments

I would like to acknowledge the fruitful discussions with Dr L Rebollo-Neira, Dr R C Alamino, Dr L Franco, Dr J van Mourik and Professor N Caticha which have enriched the content of this article.

Appendix A. Distance sensitivity

$\mathbf{S}$ and $\mathbf{S}'$ are vectors that differ in exactly P bits, i.e. $\sum_{j=1}^{P}\Theta(-S_{\sigma_j}S'_{\sigma_j}) = P$. Taken $\mathbf{S}$ as a reference, we can construct a Pth neighbor $\mathbf{S}'$ by choosing without replacement P indexes from 1 to N and flipping the correspondent entries in $\mathbf{S}$. There are $\binom{N}{P}$ different ways to choose P indexes, each one creating a different set of indexes $\mathbb{I}_P$. Introducing the scaled variables $\mu = \mathbf{w}\cdot\mathbf{S}/\sqrt{N}$ and $\mu' = \mathbf{w}\cdot\mathbf{S}'/\sqrt{N}$ by means of Dirac delta functions and adding up over all possible configurations $\mathbf{S}$, we can express the discrepancy component as

$$\eth_P^N(\mathbf{B}^{(m)}) = \binom{N}{P}^{-1}\int_{-\infty}^{\infty}\frac{\mathrm{d}\mu\,\mathrm{d}\hat{\mu}}{2\pi}\frac{\mathrm{d}\mu'\,\mathrm{d}\hat{\mu}'}{2\pi}\Theta(-\mu\mu')\mathrm{e}^{-\mathrm{i}(\mu\hat{\mu}+\mu'\hat{\mu}')} \times\sum_{\mathbb{I}_P}\prod_{j\in\mathbb{I}_P}\cos\left(\frac{\hat{\mu}-\hat{\mu}'}{\sqrt{N}}B_j^{(m)}\right)\prod_{j\notin\mathbb{I}_P}\cos\left(\frac{\hat{\mu}+\hat{\mu}'}{\sqrt{N}}B_j^{(m)}\right). \qquad (\text{A.1})$$

The fraction of sets $\mathbb{I}_P$ with $n \leqslant m$ indexes $\ell \leqslant m$ is $\binom{m}{n}\binom{N-m}{P-n}/\binom{N}{P}$ and observing that in the limit $P \leqslant N \to \infty$ with $P/N = p \leqslant 1$ we have that

$$\lim_{P\leqslant N\uparrow\infty}\binom{N}{P}^{-1}\binom{N-m}{P-n} = p^n(1-p)^{m-n}.$$

From equation (A.1), we have that

$$\eth^{(m)}(p) = \int_{-\infty}^{\infty}\frac{\mathrm{d}\mu\,\mathrm{d}\hat{\mu}}{2\pi}\frac{\mathrm{d}\mu'\,\mathrm{d}\hat{\mu}'}{2\pi}\Theta(-\mu\mu')\,\mathrm{e}^{-\mathrm{i}(\mu\hat{\mu}+\mu'\hat{\mu}')} \times\sum_{n=0}^{m}\binom{m}{n}p^n(1-p)^{m-n}\cos\left(\frac{\hat{\mu}-\hat{\mu}'}{\sqrt{m}}\right)^n\cos\left(\frac{\hat{\mu}+\hat{\mu}'}{\sqrt{m}}\right)^{m-n}.$$

By adding up the sum, opening up the cosines and applying the identity $\Theta(ab) = \Theta(a)\Theta(b) + \Theta(-a)\Theta(-b)$, the expression for the sensitivity gets reduced to

$$\eth^{(m)}(p) = 2\sum_{n=0}^{m}\binom{m}{n}(1-2p)^n\left[\int\mathcal{D}(\mu,\hat{\mu})\cos(\hat{\mu}/\sqrt{m})^{m-n}\sin(\hat{\mu}/\sqrt{m})^n\right]^2,$$

where the notation $\int\mathcal{D}(\mu,\hat{\mu})\,f(\mu,\hat{\mu})$ stands for $(2\pi)^{-1}\int_0^\infty\mathrm{d}\mu\int_{-\infty}^{\infty}\mathrm{d}\hat{\mu}\,\mathrm{e}^{-\mathrm{i}\mu\hat{\mu}}\,f(\mu,\hat{\mu})$. The integrals to be solved are

$$b_0^m \equiv \int\mathcal{D}(\eta,\hat{\eta})\cos(\hat{\eta})^m$$

$$b_n^m \equiv \int\mathcal{D}(\eta,\hat{\eta})\cos(\hat{\eta})^{m-2n}\sin(\hat{\eta})^{2n}$$

$$c_n^m \equiv \int\mathcal{D}(\eta,\hat{\eta})\cos(\hat{\eta})^{m-(2n+1)}\sin(\hat{\eta})^{2n+1}.$$

Before computing the integrals, observe that for all $A > 0$ and $B \geqslant 0$

$$\begin{aligned}\int \mathcal{D}(\eta, \hat{\eta}) \sin(A\hat{\eta}) &= -\frac{\mathrm{i}}{4\pi} \int_0^{\infty} \mathrm{d}\eta \int_{-\infty}^{\infty} \mathrm{d}\hat{\eta} \exp(-\mathrm{i}\hat{\eta}\eta)[\exp(\mathrm{i}\hat{\eta}A) - \exp(-\mathrm{i}\hat{\eta}A)] \\ &= -\frac{\mathrm{i}}{4\pi} \int_0^{\infty} \mathrm{d}\eta \int_{-\infty}^{\infty} \mathrm{d}\hat{\eta}[\exp[-\mathrm{i}\hat{\eta}(\eta - A)] - \exp[-\mathrm{i}\hat{\eta}(\eta + A)]] \\ &= -\frac{\mathrm{i}}{2}[\Theta(A) - \Theta(-A)] \\ &= -\frac{\mathrm{i}}{2}, \end{aligned} \tag{A.2}$$

similarly

$$\begin{aligned}\int \mathcal{D}(\eta, \hat{\eta}) \cos(A\hat{\eta}) \cos(B\hat{\eta}) &= \frac{1}{4}[\Theta(A + B) + \Theta(-A - B) + \Theta(A - B) + \Theta(-A + B)] \\ &= \frac{1}{2} \end{aligned} \tag{A.3}$$

and

$$\begin{aligned}\int \mathcal{D}(\eta, \hat{\eta}) \cos(A\hat{\eta}) \sin(B\hat{\eta}) &= -\frac{\mathrm{i}}{4}[\Theta(A + B) - \Theta(A - B) + \Theta(-A + B) - \Theta(-A - B)] \\ &= -\frac{\mathrm{i}}{2}\Theta(B - A). \end{aligned} \tag{A.4}$$

The first integral is (remember that m is odd)

$$\begin{aligned} b_0^m = \int \mathcal{D}(\eta, \hat{\eta}) \cos(\hat{\eta})^m &= \frac{1}{2^{m-1}} \sum_{k=0}^{(m-1)/2} \binom{m}{k} \int \mathcal{D}(\eta, \hat{\eta}) \cos[(m - 2k)\hat{\eta}] \\ &= \frac{1}{2^m} \sum_{k=0}^{(m-1)/2} \binom{m}{k} = \frac{1}{2}. \end{aligned} \tag{A.5}$$

The second integral is

$$\begin{aligned} b_n^m &= \int \mathcal{D}(\eta, \hat{\eta}) \cos(\hat{\eta})^{m-2n} \sin(\hat{\eta})^{2n} \\ &= \int \mathcal{D}(\eta, \hat{\eta}) \frac{1}{2^{m-2n-1}} \sum_{k=0}^{(m-1)/2-n} \binom{m-2n}{k} \cos[(m - 2(k + n))\hat{\eta}] \\ &\quad \times \frac{1}{2^{2n}} \left\{ 2 \sum_{j=0}^{n-1} (-1)^{n-j} \binom{2n}{j} \cos[(2n - 2j)\hat{\eta}] + \binom{2n}{n} \right\} \\ &= \frac{1}{2^{m-1}} \sum_{k=0}^{(m-1)/2-n} \binom{m-2n}{k} \sum_{j=0}^{n-1} (-1)^{n-j} \binom{2n}{j} + \frac{1}{2^{2n+1}} \binom{2n}{n} \\ &= \frac{1}{2^{m-1}} \frac{2^{m-2n}}{2} \left[-\frac{1}{2} \binom{2n}{n} \right] + \frac{1}{2^{2n+1}} \binom{2n}{n} \\ &= 0. \end{aligned} \tag{A.6}$$

And the last integral is then

$$\begin{aligned} c_n^m &= \int \mathcal{D}(\eta, \hat{\eta}) \cos(\hat{\eta})^{m-(2n+1)} \sin(\hat{\eta})^{2n+1} \\ &= \int \mathcal{D}(\eta, \hat{\eta}) \cos(\hat{\eta})^{m-(2n+1)} [1 - \cos^2(\hat{\eta})]^n \sin(\hat{\eta}) \end{aligned}$$

$$
\begin{aligned}
&= \int \mathcal{D}(\eta,\hat{\eta}) \sum_{\ell=0}^{n} (-1)^{\ell} \binom{n}{\ell} \cos(\hat{\eta})^{m-(2n+1)+2\ell} \sin(\hat{\eta}) \\
&= \sum_{\ell=0}^{n} (-1)^{\ell} \binom{n}{\ell} \frac{(-\mathrm{i})}{2^{m-1+2(\ell-n)}} \left\{ \frac{1}{2} \binom{m-1+2(\ell-n)}{(m-1)/2+\ell-n} \right. \\
&\quad \left. + \sum_{k=0}^{(m-1)/2+\ell-n-1} \binom{m-1+2(\ell-n)}{k} \Theta[1-(m-1+2(\ell-n)-k)] \right\} \\
&= -\frac{\mathrm{i}}{2^{m-2n}} \sum_{\ell=0}^{n} \left(-\frac{1}{4}\right)^{\ell} \binom{n}{\ell} \binom{m-1+2(\ell-n))}{(m-1)/2+\ell-n} \\
&= -\frac{\mathrm{i}}{2^{m}} \binom{2n}{n} \binom{m-1-2n}{(m-1)/2-n} \binom{(m-1)/2}{n}^{-1}. \qquad \text{(A.7)}
\end{aligned}
$$

We have that for all m odd

$$\mathfrak{d}^{(m)}(p) = \frac{1}{2} - \frac{1}{2} \sum_{n=0}^{(m-1)/2} a_n^m \, (1-2p)^{2n+1}, \qquad \text{(A.8)}$$

where

$$a_n^m \equiv \frac{1}{4^{m-1}} \binom{m}{2n+1} \left[\binom{2n}{n} \binom{m-1-2n}{(m-1)/2-n} \binom{(m-1)/2}{n}^{-1} \right]^2. \qquad \text{(A.9)}$$

Observe that $\mathfrak{d}^{(m)}(p)$ is concave in $p \in [0, \frac{1}{2}]$ (it is simply a sum of an affine plus concave functions) and $\mathfrak{d}^{(m)}(p) < \mathfrak{d}^{(m+2)}(p)$ for all $p \in (0, \frac{1}{2})$. To demonstrate the latter, we use that $\mathfrak{d}^{(m)}(0) = 0$ and $a_s^m > 0\ \forall s$. Thus, from (A.8) at $p = 0$, we have that

$$\sum_{s=0}^{(m-1)/2} a_s^m = 1 \qquad \forall m \geqslant 1. \qquad \text{(A.10)}$$

Therefore

$$a_{(m+1)/2}^{m+2} = \sum_{s=0}^{(m-1)/2} \left(a_s^m - a_s^{m+2}\right)$$

simply by applying (A.10) to m and to $m + 2$. Observe that $(1-2p)^n < (1-2p)^{n'}$ for all $n > n'$ and $p \in (0, \frac{1}{2})$, thus

$$a_{(m+1)/2}^{m+2}(1-2p)^{m+2} < \sum_{s=0}^{(m-1)/2} \left(a_s^m - a_s^{m+2}\right)(1-2p)^{2s+1}$$

$$\sum_{s=0}^{(m+1)/2} a_s^{m+2}(1-2p)^{2s+1} < \sum_{s=0}^{(m-1)/2} a_s^m (1-2p)^{2s+1}$$

and thus $\mathfrak{d}^{(m)}(p) < \mathfrak{d}^{(m+2)}(p)$.

In the large m limit, we have that

$$\lim_{m\uparrow\infty} \mathfrak{d}^{(m)}(p) = \frac{1}{\pi} \arccos(1-2p),$$

which is the expected result [27].

Appendix B. Optimal learning algorithm

The basic ingredient to compute the optimal learning algorithm is the joint probability distribution of the variables $\sigma_{\mathbf{B}}$, h and b. Given that $\mathcal{P}(\sigma_{\mathbf{B}}, h, b|m) = \Theta(\sigma_{\mathbf{B}} b)\,\mathcal{P}(h, b|m)$, we will start our inference task by computing the distribution of the post-synaptic fields:

$$\begin{aligned}\mathcal{P}(h, b|m) &= \langle \delta(h - \mathbf{J}\cdot\mathbf{S}/|\mathbf{J}|)\delta(b - \mathbf{B}\cdot\mathbf{S}/|\mathbf{B}|)\rangle_{\mathbf{S}} \\ &= \int_{-\infty}^{\infty} \frac{\mathrm{d}\hat{h}}{2\pi}\mathrm{e}^{-\mathrm{i}\hat{h}h} \int_{-\infty}^{\infty} \frac{\mathrm{d}\hat{b}}{2\pi}\mathrm{e}^{-\mathrm{i}\hat{b}b} \left\langle \exp\left(\mathrm{i}\hat{h}\frac{\mathbf{J}\cdot\mathbf{S}}{|\mathbf{J}|} + \mathrm{i}\hat{b}\frac{\mathbf{B}\cdot\mathbf{S}}{|\mathbf{B}|}\right)\right\rangle_{\mathbf{S}}\end{aligned}$$

and assuming that $[\mathbf{B}]_j = \sqrt{\frac{N}{m}}\,\Theta(m + 1 - j)$ we can suppose that the student *learns* this rule in such a way that $[\mathbf{J}]_j \simeq \mathcal{J}\Theta(m + 1 - j) + \varepsilon_j$, where $\varepsilon_j \ll |\mathbf{J}|$ are i.i.d. variables. Therefore

$$\frac{\mathcal{J}}{|\mathbf{J}|} = \frac{R}{\sqrt{m}} - \frac{\overline{\varepsilon}}{|\mathbf{J}|},$$

where R is the teacher–student overlap and $\overline{\varepsilon} \equiv \sum_{j=1}^{m} \varepsilon_j/m$. Let us define the variables

$$\varphi_j \equiv \frac{\varepsilon_j - \overline{\varepsilon}}{|\mathbf{J}|},$$

with the properties of $\sum_{j=1}^{m} \varphi_j = 0$ and

$$\sum_{j>m} \frac{\varepsilon_j^2}{|\mathbf{J}|^2} = 1 - R^2 - \sum_{j=1}^{m} \varphi_j^2.$$

Thus the trace over the spin variables gives

$$\begin{aligned}\left\langle \exp\left(\mathrm{i}\hat{h}\frac{\mathbf{J}\cdot\mathbf{S}}{|\mathbf{J}|} + \mathrm{i}\hat{b}\frac{\mathbf{B}\cdot\mathbf{S}}{|\mathbf{B}|}\right)\right\rangle_{\mathbf{S}} &= \prod_{j=1}^{N} \frac{1}{2} \sum_{S=\pm 1} \exp\left[\mathrm{i}\left(\frac{\hat{h}[\mathbf{J}]_j}{|\mathbf{J}|} + \frac{\hat{b}[\mathbf{B}]_j}{|\mathbf{B}|}\right) S\right] \\ &= \prod_{j=1}^{m} \cos\left(\hat{h}\frac{\mathcal{J} + \varepsilon_j}{|\mathbf{J}|} + \frac{\hat{b}}{\sqrt{m}}\right) \prod_{j>m} \cos\left(\frac{\hat{h}\varepsilon_j}{|\mathbf{J}|}\right) \\ &= \prod_{j=1}^{m} \cos\left(\frac{\hat{h}R + \hat{b}}{\sqrt{m}} + \hat{h}\varphi_j\right) \prod_{j>m} \cos\left(\frac{\hat{h}\varepsilon_j}{|\mathbf{J}|}\right) \\ &\simeq \cos\left(\frac{\hat{h}R + \hat{b}}{\sqrt{m}}\right)^m \prod_{j=1}^{m}\left[1 - \hat{h}\varphi_j \tan\left(\frac{\hat{h}R + \hat{b}}{\sqrt{m}}\right) + O\left(\varphi_j^2\right)\right] \\ &\quad \times \exp\left(-\frac{\hat{h}^2}{2}\sum_{j>m}\frac{\varepsilon_j^2}{|\mathbf{J}|^2}\right)\left[1 + O\left(\sum_{j>m}\frac{\varepsilon_j^4}{|\mathbf{J}|^4}\right)\right] \\ &\simeq \cos\left(\frac{\hat{h}R + \hat{b}}{\sqrt{m}}\right)^m \exp\left(-\frac{1 - R^2}{2}\hat{h}^2\right) + O\left(\sum_{j=1}^{m}\varphi_j^2\right).\end{aligned}$$

Therefore, and using that m is odd,

$$\begin{aligned}\mathcal{P}(h,b|m) &\simeq \frac{1}{2^{m-1}}\sum_{k=0}^{(m-1)/2}\binom{m}{k}\int_{-\infty}^{\infty}\frac{\mathrm{d}\hat{h}}{2\pi}\frac{\mathrm{d}\hat{b}}{2\pi}\exp\left(-\frac{1-R^2}{2}\hat{h}^2-\mathrm{i}h\hat{h}-\mathrm{i}b\hat{b}\right)\\ &\quad\times\cos\left[(m-2k)\frac{\hat{b}+R\hat{h}}{\sqrt{m}}\right]\\ &=\mathcal{N}(h|Rb,1-R^2)\,\frac{1}{2^m}\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}[\delta(b-\mu_k)+\delta(b+\mu_k)],\end{aligned}\tag{B.1}$$

where $\mu_k=(2k+1)/\sqrt{m}$ and $\mathcal{N}(x|\mu,\sigma^2)$ is a normal distribution in x, centered at μ, with variance σ^2.

From (B.1), we can compute the joint distribution of the variables h and $\sigma_{\mathbf{B}}$

$$\mathcal{P}(\sigma_{\mathbf{B}},h|m)=\frac{1}{2^m}\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mathcal{N}(\sigma_{\mathbf{B}}h|R\mu_k,1-R^2),\tag{B.2}$$

which implies that

$$\begin{aligned}\mathcal{P}(\phi|m)&=\sum_{\sigma_{\mathbf{B}}=\pm1}\int_{-\infty}^{\infty}\mathrm{d}h\,\mathcal{P}(\sigma_{\mathbf{B}},h|m)\,\delta(\phi-\sigma_{\mathbf{B}}h)\\ &=\frac{1}{2^{m-1}}\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mathcal{N}(\phi|R\mu_k,1-R^2).\end{aligned}\tag{B.3}$$

The conditional probability of the field b given $\sigma_{\mathbf{B}}$ and h can be obtained from (B.1) and (B.2).

It is a simple inference exercise to find the conditional distribution of the field b given the stability ϕ

$$\mathcal{P}(b|\phi,m)=\frac{1}{2}\frac{\mathcal{N}(\phi|R|b|,1-R^2)\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\delta(|b|-\mu_k)}{\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mathcal{N}(\phi|R\mu_k,1-R^2)}.$$

The conditional expectation of the field $|b|$ is

$$\begin{aligned}\langle|b|\rangle_{b|\phi,m}&=\int_{-\infty}^{\infty}\mathrm{d}b|b|\mathcal{P}(b|\phi)\\ &=\frac{\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mu_k\mathcal{N}(\phi|R\mu_k,1-R^2)}{\sum_{k=0}^{(m-1)/2}\binom{m}{(m-1)/2-k}\mathcal{N}(\phi|R\mu_k,1-R^2)}.\end{aligned}\tag{B.4}$$

Appendix C. Derivation of (15)

In this appendix, we continue the development of (14)

$$\begin{aligned}\mathscr{I}_{m_1,m_2}&=\int\mathrm{d}\phi\mathcal{P}(\phi|m_{\mathbf{B}})\Upsilon(\phi|R,m_1)\Upsilon(\phi|R,m_2)\\ &=\frac{1}{2^{m_{\mathbf{B}}-1}}\sum_{k=0}^{(m_{\mathbf{B}}-1)/2}\binom{m_{\mathbf{B}}}{(m_{\mathbf{B}}-1)/2-k}\sum_{\ell_1=0}^{(m_1-1)/2}\sum_{\ell_2=0}^{(m_2-1)/2}\frac{R^4}{m_1m_2}\int_{-1}^{1}\mathrm{d}t_1t_1\int_{-1}^{1}\mathrm{d}t_2t_2\\ &\quad\times\int\mathrm{d}\phi\mathcal{N}(\phi|R\mu_k)\mathcal{N}\left(\phi\left|\frac{R}{\sqrt{m_1}}(2\ell_1+1-t_1)\right.\right)\mathcal{N}\left(\phi\left|\frac{R}{\sqrt{m_2}}(2\ell_2+1-t_2)\right.\right),\end{aligned}$$

where all the normal distributions have exactly the same variance $1 - R^2$. The integral over ϕ is simple and produces a bi-variate Gaussian distribution in t_1 and t_2

$$\mathscr{I}_{m_1,m_2} = \frac{1}{2^{m_\mathbf{B}-1}} \sum_{k=0}^{(m_\mathbf{B}-1)/2} \binom{m_\mathbf{B}}{(m_\mathbf{B}-1)/2-k} \sum_{\ell_1=0}^{(m_1-1)/2} \sum_{\ell_2=0}^{(m_2-1)/2} \frac{R^2}{\sqrt{m_1 m_2}} \times \int_{\mathscr{D}_\mathbf{t}} \mathrm{d}\mathbf{t}\, t_1 t_2 \mathscr{N}(\mathbf{t}|\mathbf{t}_{\ell_1,\ell_2,k};\, \boldsymbol{\Sigma}) \tag{C.1}$$

where $\mathbf{t} = (t_1, t_2)^\mathsf{T}$, $\mathscr{D}_\mathbf{t} \equiv (-1, 1) \times (-1, 1)$, $\mathbf{t}_{\ell_1,\ell_2,k} = (\sqrt{m_1}\,\delta_{\ell_1,k}, \sqrt{m_2}\,\delta_{\ell_2,k})^\mathsf{T}$ and

$$\delta_{\ell_j,k} \equiv \frac{2\ell_j+1}{\sqrt{m_j}} - \frac{2k+1}{\sqrt{m_\mathbf{B}}}, \tag{C.2}$$

$$\boldsymbol{\Sigma} \equiv 2\,\frac{1-R^2}{R^2} \begin{pmatrix} m_1 & \frac{1}{2}\sqrt{m_1 m_2} \\ \frac{1}{2}\sqrt{m_1 m_2} & m_2 \end{pmatrix}. \tag{C.3}$$

From (10*a*), all the entries of the covariance matrix (C.3) are small; therefore all the distributions are concentrated around $\mathbf{t}_{\ell_1,\ell_2,k}$. Let $\mathbf{t}_k^\star$ be the vector that corresponds to the largest term in (C.1). Its components are

$$\ell^\star_{m_j,k} \equiv \left[\!\left[\sqrt{\frac{m_j}{m_\mathbf{B}}}\,\frac{2k+1}{2} - \frac{1}{2} \right]\!\right] \tag{C.4}$$

$$\sqrt{m_j}\,\delta^\star_{m_j,k} = 2\ell^\star_{m_j,k} + 1 - \sqrt{\frac{m_j}{m_\mathbf{B}}}\,(2k+1), \tag{C.5}$$

where $[\![r]\!]$ is the closest integer to $r \in \mathbb{R}$; thus $\sqrt{m_j}\,\delta^\star_{m_j,k} \in (-1, 1)$. All the other vectors can be expressed as $\mathbf{t}_{\ell_1,\ell_2,k} = \mathbf{t}_k^\star + 2\mathbf{n}$, where $\mathbf{n} = (n_1, n_2)^\mathsf{T}$, $n_j = -\ell^\star_{m_j,k}, -\ell^\star_{m_j,k} + 1, \ldots, -1, 1, \ldots, -\ell^\star_{m_j,k} + (m_j - 1)/2$. We have that

$$\mathscr{I}_{m_1,m_2} = \frac{1}{2^{m_\mathbf{B}-1}} \sum_{k=0}^{(m_\mathbf{B}-1)/2} \binom{m_\mathbf{B}}{(m_\mathbf{B}-1)/2-k} \frac{R^2}{\sqrt{m_1 m_2}} \times \left\{ \int_{\mathscr{D}_\mathbf{t}} \mathrm{d}\mathbf{t}\, t_1 t_2 \mathscr{N}\left(\mathbf{t}\,\middle|\mathbf{t}_k^\star;\, \boldsymbol{\Sigma}\right) + \sum_{\mathbf{n}} \int_{\mathscr{D}_\mathbf{t}} \mathrm{d}\mathbf{t}\, t_1 t_2 \mathscr{N}\left(\mathbf{t}\,\middle|\mathbf{t}_k^\star + 2\mathbf{n};\, \boldsymbol{\Sigma}\right) \right\}.$$

Observe that the vectors $\mathbf{t}_k^\star$ are always strictly inside the domain $\mathscr{D}_\mathbf{t}$. They can never be in the boundary of the domain given that m_j is odd, and then the argument of the rhs of (C.4) is never in $\mathbb{Z}_{1/2}$ (which would produce the largest possible value of $\sqrt{m_j}\,\delta^\star_{m_j,k}$). Thus, the largest contribution to the sum over $\mathbf{n}$ is of $O[\exp(-\epsilon^2/\max\{\lambda \in \mathrm{spec}(\boldsymbol{\Sigma})\})]$, where $\epsilon \sim 1 - |\sqrt{m_j}\,\delta^\star_{m_j,k}| > \frac{1}{2}$ and

$$\max\{\lambda \in \mathrm{spec}(\boldsymbol{\Sigma})\} = \frac{1-R^2}{R^2}\left[m_1 + m_2 + \sqrt{m_1^2 + m_2^2 - m_1 m_2}\right] \ll 1,$$

according to (10*a*). Within the same approximation error, we can suppose that the centre of the zeroth normal distribution is located inside the domain and sufficiently farther from the boundary. Thus,

$$\mathscr{I}_{m_1,m_2} \simeq \frac{1}{2^{m_\mathbf{B}-1}} \sum_{k=0}^{(m_\mathbf{B}-1)/2} \binom{m_\mathbf{B}}{(m_\mathbf{B}-1)/2-k} \frac{R^2}{\sqrt{m_1 m_2}} \int_{\mathbb{R}^2} \mathrm{d}\mathbf{t}\, t_1 t_2 \mathscr{N}\left(\mathbf{t}\,\middle|\mathbf{t}_k^\star;\, \boldsymbol{\Sigma}\right) + O\left(\exp\left[-\frac{\epsilon^2}{\max\left\{\lambda \in \mathrm{spec}(\boldsymbol{\Sigma})\right\}}\right]\right).$$

Thus

$$\begin{aligned}\mathscr{I}_{m_1,m_2} &\simeq \frac{1}{2^{m_{\mathbf{B}}-1}} \sum_{k=0}^{(m_{\mathbf{B}}-1)/2} \binom{m_{\mathbf{B}}}{(m_{\mathbf{B}}-1)/2-k} \frac{R^2}{\sqrt{m_1 m_2}} \int_{-\infty}^{\infty} \mathrm{d}t_2 t_2 \mathcal{N}\left(t_2 \,\middle|\, \delta^\star_{m_2,k};\, 2m_2 \frac{1-R^2}{R^2}\right) \\ &\quad \times \int_{-\infty}^{\infty} \mathrm{d}t_1 t_1 \mathcal{N}\left(t_1 \,\middle|\, \delta^\star_{m_1,k} + \frac{1}{2}\sqrt{\frac{m_1}{m_2}}(t_2 - \delta^\star_{m_2,k});\, \frac{3}{2} m_1 \frac{1-R^2}{R^2}\right) \\ &= 1 - R^2 \left[1 - \frac{1}{2^{m_{\mathbf{B}}-1}} \sum_{k=0}^{(m_{\mathbf{B}}-1)/2} \binom{m_{\mathbf{B}}}{(m_{\mathbf{B}}-1)/2-k} \delta^\star_{m_1,k} \delta^\star_{m_2,k}\right].\end{aligned}$$

References

[1] Rumelhart D E and Mac Clelland J L 1986 *Parallel Distributed Processing* vol I (Cambridge, MA: MIT Press)

[2] Minsky M and Papert S A 1988/1989 *Perceptrons: An Introduction to Computational Geometry* (Cambridge, MA: MIT Press) (expanded edition)

[3] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)

[4] Copelli M, Kinouchi O and Caticha N 1996 *Phys. Rev.* E **53** 6341

[5] Copelli M, Eichhorn R, Kinouchi O, Biehl M, Simonetti R, Riegler P and Caticha N 1997 *Europhys. Lett.* **37** 427

[6] Uezu T, Maeda Y and Yamaguchi S 2006 *J. Phys. Soc. Japan* **75** 114007

[7] Uezu T, Miyoshi S, Izuo M and Okada M 2007 *J. Phys. Soc. Japan* **76** 114006

[8] Church A 1936 *Am. J. Math.* **58** 345

[9] Turing A M 1973 *Proc. London Math. Soc.* **42** 230

[10] Hartmanis J and Stearns R E 1965 *Trans. Am. Math. Soc.* **117** 285

[11] Kolmogorov A N 1965 *Probl. Inf. Transm.* **1** 1

[12] Li M and Vitányi P 1997 *An Introduction to Kolmogorov Complexity and its Applications* 2nd edn (Berlin: Springer)

[13] Franco L and Cannas S 2000 *Neural Comput.* **12** 2405

[14] Franco L and Cannas S 2004 *Physica* A **332** 337

[15] Franco L and Anthony M 2004 *Proc. IEEE Int. Joint Conf. Neural Networks (Budapest, Hungary)* p 973

[16] Coolen A C C, Kühn R and Sollich P 2005 *Theory of Neural Information Processing Systems* (Oxford: Oxford University Press)

[17] Canning A and Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 3275

[18] Bouten M, Engel A, Komoda A and Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643

[19] Bollé D and van Mourik J 1994 *J. Phys. A: Math. Gen.* **27** 1151

[20] Kuhlmann P and Müller K R 1994 *J. Phys. A: Math. Gen.* **27** 3759

[21] López B and Kinzel W 1997 *J. Phys. A: Math. Gen.* **30** 7753

[22] Malzahn D 2000 *Phys. Rev.* E **61** 6261

[23] Kalai G 2002 *Adv. Appl. Math.* **29** 412

[24] O'Donnell R W 2003 *Computational Applications of Noise Sensitivity MIT thesis*

[25] Stoer J and Bulirsch R 1980 *Introduction to Numerical Analysis* (New York: Springer)

[26] Kinouchi O and Caticha N 1996 *Phys. Rev.* E **54** R54

[27] Feller W 1957 *An Introduction to Probability Theory and Its Applications* 2nd edn (New York: Wiley)