Parallel strategy for optimal learning in perceptrons

J. P. Neirotti

NCRG, Aston University, Birmingham, United Kingdom

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

Abstract. We developed a parallel strategy for learning optimally specific realizable rules by perceptrons, in an on-line learning scenario. Our result is a generalisation of the Caticha-Kinouchi (CK) algorithm developed for learning a perceptron with a synaptic vector drawn from a uniform distribution over the N-dimensional sphere, so called the typical case. Our method outperforms the CK algorithm in almost all possible situations, failing only in a denumerable set of cases. The algorithm is optimal in the sense that it saturates Bayesian bounds when it succeeds.

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

Submitted to: *J. Phys. A: Math. Gen.*

1. Introduction

One of today's challenges in the area of Artificial Intelligences (AI) is the development of autonomous intelligent agents. In general terms an autonomous agent is a system situated within an environment, which the agent senses and acts over [1]. These agents need mechanisms for assimilating and processing environmental information acquired through their sensors. Data acquisition and information processing are some of the most characteristic features of neural networks.

During the past decade great effort has been applied in the research of the on-line learning scenario in artificial systems. In the on-line scenario information, represented by strings of bits drawn from a given distribution, is presented to the network for processing and then discarded [2, 3, 4, 5]. This scenario is particularly appealing for the development of autonomous agents that have to interpret, adapt and react to ever changing environmental conditions.

In the statistical mechanics approach to the learning from examples and generalisation by neural networks, the single-layered perceptron has been the preferred laboratory. 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 [6, 7], which signifies a real challenge for the adaptability of the system modelled by the network. This scenario has been recently revisited and extended to the situation of a student learning from two teachers [8, 9]. The common factor in all these studies is that the teacher is a *typical* perceptron, with a synaptic vector drawn from a uniform distribution over the N-sphere of radius $\sqrt{N}$.

In a previous article [10], we studied the mismatched scenario where a student uses an algorithm suited from learning optimally from a teacher different from the one the student is currently learning from. We demonstrated that in such cases the student mostly fails to learn even when the algorithm applied is suitable for learning from a teacher harder than the one currently in use. We have also proven that if the rule to be learned is the simplest possible (the one-bit diluted perceptron) the algorithm developed for learning optimally the typical teacher [3] is outperformed by the simplest possible algorithm (the pure Hebb rule). These results naturally triggered the question whether it is possible to tailored an algorithm specific for learning a particular realizable rule.

We present in this paper an algorithm developed for learning from almost any perceptron teacher, with performance not worse than the Caticha-Kinouchi (CK) algorithm [3]. In the next section we present the background needed for the main development of the algorithm. In section 3 we present the algorithm based on an estimate for the distribution of the teacher's post-synaptic field. In section 4 we present numerical estimates of the learning curve for different cases, including the particular synaptic vectors where the algorithm fails. Finally, in Section 5, we present our conclusions and a brief description of our future work.

2. Background

In the supervised, on-line learning scenario, the student learns to classify input vectors like a teacher. The input vectors are drawn according to a given distribution, presented to the student one by one and then discarded. The measure of the student's performance is given by the estimate of the expected mismatch between teacher's and student's classifications. For computing these estimates it is necessary to obtain the distribution of the relevant variables of the problem.

By the development presented in Appendix A we may suppose, without loss of generality, that any teacher perceptron has a synaptic vector $\mathbf{B}$ with non-negative, decreasingly ordered entries and norm B. Let $\mathbf{J}$ be the student's synaptic vector learning from $\mathbf{B}$. The norm of $\mathbf{J}$ is denoted by J. Let

$$b = \frac{\mathbf{B}^{\mathrm{T}}\mathbf{S}}{B}\,, \qquad h = \frac{\mathbf{J}^{\mathrm{T}}\mathbf{S}}{J}$$

be the teacher's and student's post synaptic fields. Observe that we have opted for the matrix notation of the inner product (i.e. $\forall\,\mathbf{U},\,\mathbf{V} \in \mathbb{R}^N\ \mathbf{U}\cdot\mathbf{V} = \mathbf{U}^{\mathrm{T}}\mathbf{V}$, where T indicates the transpose). The input $\mathbf{S}$ is binary, unless said otherwise. It can be demonstrated (see Appendix A) that the joint distribution of the post synaptic fields can be expressed as

$$\mathcal{P}(b,h) \simeq \mathcal{N}(h|bR, 1-R^2)\,\mathcal{P}_b(b), \tag{1}$$

where $\mathcal{N}(x|\mu,\sigma^2)$ is a normal distribution in x, centred at μ and variance σ^2. The marginal distribution of the field b is

$$\mathcal{P}_b(b) = \lim_{N\to\infty}\int_{-\infty}^{\infty}\frac{\mathrm{d}\hat{b}}{2\pi}\,\mathrm{e}^{-i\hat{b}b}\prod_{k=1}^{N}\cos(\hat{b}\beta_k)\,, \tag{2}$$

where $\boldsymbol{\beta} = \mathbf{B}/B$ is a unit vector with positive, decreasingly ordered entries. The product of cosines can be rewritten as:

$$\Phi_N(\hat{b}) \equiv \prod_{k=1}^{N}\cos(\hat{b}\beta_k) = \frac{1}{2^N}\sum_{\{\mathbf{T}\in\{\pm1\}^N\}}\cos(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{T}) \tag{3}$$

and thus

$$\mathcal{P}_b(b) = \lim_{N\to\infty}\frac{1}{2^N}\sum_{\{\mathbf{T}\in\{\pm1\}^N\}}\delta(b-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{T})\,,$$

where $\delta(x)$ is Dirac's delta function. Thus the field b can only be equated to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{T} = \sum_{k=1}^{N}\beta_k T_k$ which is the length of a random path with decreasing step sizes β_k. An interesting study on random walks with decreasing steps can be found in [11]. It is important to note that if the entries β_k depend on the size of the path N such that $\forall\,k\ \lim_{N\to\infty}\beta_k = 0$ then:

(i) $\mathcal{P}_b(b) = \mathcal{N}(b|0,1)$, and then the optimal learning algorithm is the one found by Caticha and Kinouchi [3],

(ii) $b_{\max} \equiv \lim_{N\to\infty} \sum_{k=1}^{N} \beta_k = \infty$.

If the entries of $\boldsymbol{\beta}$ are taken from a sequence $\{\beta_k\}_{k=1}^{\infty}$ in ℓ_2 (i.e. the space of sequences $\{a_k\}$ such that $\sum_{k=1}^{\infty} a_k^2 < \infty$), with not all of its elements equal to zero, then the following hold,

(i) The product $\Phi_N(\hat{b})$ converges absolutely for all $\hat{b}$.

(ii) The product $\Phi_N(\hat{b})$ converges uniformly on compact sets.

(iii) The product $\Phi_N(\hat{b})$ is uniformly continuous.

(iv) The product $\Phi_N(\hat{b})$ has a Fourier transform in the distribution sense.

About point (iv) above, the Fourier transform of the product $\Phi(\hat{b})$ is the measure $\mathcal{P}_b(b)$ which may be singular with respect to the Lebesgue measure (we will explore this case with a particular example in 4.5). If the measure is not singular then the following algorithm can be applied to learn the teacher $\mathbf{B}$.

3. The Parallel Algorithm

A Hebbian-like algorithm has the following form

$$\mathbf{J}_{\text{new}} = \mathbf{J}_{\text{old}} + F\,\frac{\sigma_{\mathbf{B}}}{\sqrt{N}}\,\mathbf{S} \tag{4}$$

where $\sigma_{\mathbf{B}} \equiv \text{sgn}(\mathbf{B}^{\text{T}}\mathbf{S})$ is the classification given by the teacher and F is the learning rate, which can be a function of the variables available to the student, the pair $(\sigma_{\mathbf{B}}, \mathbf{S})$ and the state of the student, represented by $\mathbf{J}_{\text{old}}$. It has been demonstrated [3] that the learning rate that produces the lowest expected error has the form:

$$F_{\text{op}} = \frac{\sqrt{Q}}{R}\,\left[\langle|b|\rangle_{b|\phi} - R\phi\right] \tag{5}$$

where $Q \equiv J^2/N$ is the normalised size of the student's synaptic vector, $\phi \equiv \sigma_{\mathbf{B}} h$ is the stability or *surprise* parameter, $R \equiv \mathbf{B}^{\text{T}}\mathbf{J}/(BJ)$ is the student-teacher overlap and

$$\langle|b|\rangle_{b|\phi} \equiv \int \text{d}b\,|b|\,\mathcal{P}(b|\phi)$$

is the conditional expected value of the absolute value of the teacher's synaptic field given the knowledge available to the student conveyed by the variable ϕ. It is a simple exercise to show that the conditional probability can be obtained from (1)

$$\mathcal{P}(b|\phi) = \frac{\mathcal{N}(\phi||b|R, 1-R^2)\,\mathcal{P}_b(b)}{\int \text{d}b\,\mathcal{N}(\phi||b|R, 1-R^2)\,\mathcal{P}_b(b)},$$

thus

$$\langle|b|\rangle_{b|\phi} = \frac{\int_0^{\infty} \text{d}b\,b\,\mathcal{N}(\phi|bR, 1-R^2)\,\mathcal{P}_b(b)}{\int_0^{\infty} \text{d}b\,\mathcal{N}(\phi|bR, 1-R^2)\,\mathcal{P}_b(b)}. \tag{6}$$

The optimal algorithm relies on the knowledge of the overlap R and the distribution $\mathcal{P}_b$. To obtain an appropriate estimate for the overlap R we rely on the measurement of the *time averaged* generalisation error

$$e_{\text{g}} \equiv \langle\Theta(-\phi)\rangle_{\mathscr{L}_M} \tag{7}$$

where $\Theta(x) = 1$ if $x \geq 1$ and 0 otherwise and $\mathscr{L}_M = \{\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_M\}$ is a collection of M sets of input samples $\mathscr{S}_m = \left\{\mathbf{S}_1^{(m)}, \mathbf{S}_2^{(m)}, \ldots, \mathbf{S}_P^{(m)}\right\}$. Each one of these sets is used in a particular realization of the learning process, and the average over realizations provides the estimate for the generalisation error. In terms of the joint probability (1) we have that the *ensemble average* of the generalisation error is

$$\begin{aligned} e_{\mathrm{g}}(R) &= \int_{-\infty}^{\infty} \mathrm{d}b\, \mathrm{d}h\, \mathcal{P}(b,h)\, \Theta(-bh) \\ &= 2\int_{0}^{\infty} \mathrm{d}b\, \mathcal{P}_b(b)\, \mathcal{H}(bR/\sqrt{1-R^2}), \end{aligned} \tag{8}$$

where $\mathcal{H}(x) \equiv \int_x^\infty \mathrm{d}u\, \exp(-u^2/2)/\sqrt{2\pi}$.

One way to estimate the LHS of (7) is by considering an ensemble of M students learning in parallel, all following the update rule (4). Let us denote such an ensemble as $\left\{\mathbf{J}_p^{(i)}, F_p^{(i)}\right\}$ where $\mathbf{J}_p^{(i)}$ and $F_p^{(i)}$ are the synaptic vector and the learning rate of the i-th student after p updates. The natural initial condition is by supposing the students start from the *tabula rasa* state, i.e. $\mathbf{J}_0^{(i)} = \mathbf{0}$ and learning rates set to pure Hebb algorithms $F_0^{(i)} = 1$. The first element of $\mathscr{S}_i$, i.e. $\mathbf{S}_1^{(i)}$ is classified according to $\sigma_{\mathbf{B},1}^{(i)} \equiv \mathrm{sgn}(\mathbf{B}^{\mathrm{T}}\mathbf{S}_1^{(i)})$. Given that all the students are assigned a null synaptic vector, the initial estimate for the generalisation error is set to $\tilde{e}_0 = \frac{1}{2}$, consistent with a $\tilde{R}_0 = 0$. The first update for the i-th student is

$$\mathbf{J}_1^{(i)} = \frac{\sigma_{\mathbf{B},1}^{(i)}}{\sqrt{N}}\, \mathbf{S}_1^{(i)}. \tag{9}$$

Next, the second inputs from the sets $\mathscr{S}_i$ are classified by the teacher producing the pairs $(\sigma_{\mathbf{B},2}^{(i)}, \mathbf{S}_2^{(i)})$. With these inputs we can compute the stabilities

$$\phi_1^{(i)} \equiv \sigma_{\mathbf{B},2}\, \frac{\mathbf{J}_1^{(i)\mathrm{T}}\mathbf{S}_2^{(i)}}{J_1^{(i)}} \tag{10}$$

and the generalisation error

$$\tilde{e}_1 \equiv \frac{1}{M}\sum_i \Theta(-\phi_1^{(i)}). \tag{11}$$

Following the Ansatz (A.2) we set $\mathbf{J}^{(i)} = J_{\mathbf{B}}^{(i)}\boldsymbol{\beta} + J_{\perp}^{(i)}\boldsymbol{\beta}_{\perp}^{(i)}$ where $\boldsymbol{\beta}_{\perp}^{(i)}$ is a random unit vector in the hyper-plane perpendicular to $\boldsymbol{\beta}$. To estimate the teacher's synaptic vector we use the arithmetic average over the ensemble of students

$$\tilde{\boldsymbol{\beta}}_1 \equiv \frac{\sum_i \mathbf{J}_1^{(i)}}{|\sum_i \mathbf{J}_1^{(i)}|}\,; \tag{12}$$

if M is sufficiently large, the perpendicular component of the students synaptic vectors cancel each other. If M is large enough and there is no correlation between inputs from different sets (i.e. $\langle \mathbf{S}_p^{(i)\mathrm{T}}\mathbf{S}_p^{(j)}\rangle \simeq N\delta_{i,j}$) then we would expect $\hat{\boldsymbol{\beta}}_1$ to be parallel to $\boldsymbol{\beta}$ with corrections of $O(1/\sqrt{M})$.

The existence of the fast Fourier transform (FFT) algorithm [12] makes practical the numerical estimation of the density $\mathcal{P}_b$. This technique produces better results when

applied to grids of a size equal to a power of two, 2^G. The FFT of the function $f(\hat{x})$, i.e. $\mathsf{FFT}[f(\hat{x})]_{2^G}$, produces a 2^G-dimensional vector $\mathbf{F}$, with entries equal to the Fourier transform of $f(\hat{x})$, evaluated at the points $x_k = (k-1)2^{-G}\, x_{\max}$, for a suitable value of the cutoff $x_{\max}$. Thus

$$F_k = \int_{-\infty}^{\infty} \frac{\mathrm{d}\hat{x}}{2\pi}\, \mathrm{e}^{-i\hat{x}x_k} f(\hat{x}) \quad \forall\, k = 1, 2, \ldots, 2^G\,.$$

In order to compute the estimate of $\mathcal{P}_b$ we need first to compute the cutoff $b_{\max,1}$, the grid vector $\mathbf{b}_1$ and finally the Fourier transform $\tilde{\mathbf{P}}_1$:

$$b_{\max,1} = \sum_{k=1}^{N} \tilde{\beta}_{1,k} \tag{13a}$$

$$\mathbf{b}_1 \quad = \frac{b_{\max,1}}{2^G}(0, 1, 2, \ldots, 2^G - 1)^{\mathrm{T}} \tag{13b}$$

$$\tilde{\mathbf{P}}_1 \quad = \mathsf{FFT}\left[\prod_{k=1}^{N} \cos(\hat{b}\tilde{\beta}_{1,k})\right]_{2^G}\,. \tag{13c}$$

With the estimate of the probability density stored in a vector, the expectation values take the form of an inner product.

In order to estimate the overlap $\tilde{R}_1$ we use the estimate of the error obtained by (11) and the expression (8). To estimate this last one we define the vectors $\mathbf{H}(\mathbf{b}, R)$ and $\boldsymbol{\Gamma}(\mathbf{b}, R)$ with entries

$$H_i(\mathbf{b}, R) \equiv \mathcal{H}(b_i\, R/\sqrt{1-R^2}) \tag{14}$$

and

$$\Gamma_i(\mathbf{b}, R) \equiv b_i\, \mathcal{N}(b_i\, R|0, 1-R^2) \tag{15}$$

To determine $\tilde{R}_1$ we appeal to Newton's method, which provides the following iterative equation

$$\tilde{R}_1 \leftarrow \left\lceil \tilde{R}_{1,n} + (1-\tilde{R}_{1,n}^2)\, \frac{2\,\tilde{\mathbf{P}}_1^{\mathrm{T}}\, \mathbf{H}(\mathbf{b}_1, \tilde{R}_{1,n}) - \tilde{e}_1}{2\,\tilde{\mathbf{P}}_1^{\mathrm{T}}\, \boldsymbol{\Gamma}(\mathbf{b}_1, \tilde{R}_{1,n})} \right\rfloor_{n|\delta, N_{\max}}\,, \tag{16}$$

where $\tilde{R}_{1,0} \equiv \cos(\pi\tilde{e}_1)$ and $x \leftarrow \lceil f(x_n) \rfloor_{n|\delta,N_{\max}}$ represents the iterative map $x_{n+1} = f(x_n)$ that stops when either $|x_{n+1} - x_n| < \delta$ or $n > N_{\max}$ for suitable, prefixed $0 < \delta \in \mathbb{R}$ and $N_{\max} \in \mathbb{N}$. In such a case $x \equiv x_n$.

Let us define now the vectors $\mathbf{N}(\phi, \mathbf{b}, R)$ and $\boldsymbol{\Upsilon}(\phi, \mathbf{b}, R)$ with entries

$$N_i(\phi, \mathbf{b}, R) \equiv \mathcal{N}(\phi|b_i\, R, 1-R^2) \tag{17}$$

and

$$\Upsilon_i(\phi, \mathbf{b}, R) \equiv b_i\, \mathcal{N}(\phi|b_i\, R, 1-R^2) \tag{18}$$

such that the estimate for the conditional average of the teacher's post-synaptic field becomes

$$\tilde{b}_1^{(i)} \equiv \frac{\tilde{\mathbf{P}}_1^{\mathrm{T}}\, \boldsymbol{\Upsilon}(\phi_1^{(i)}, \mathbf{b}_1, \tilde{R}_1)}{\tilde{\mathbf{P}}_1^{\mathrm{T}}\, \mathbf{N}(\phi_1^{(i)}, \mathbf{b}_1, \tilde{R}_1)} \tag{19}$$

and the learning rates

$$F_1^{(i)} \equiv \frac{\sqrt{Q_1^{(i)}}}{\tilde{R}_1}\,(\tilde{b}_1^{(i)} - \tilde{R}_1\phi_1^{(i)}) \tag{20}$$

where $Q_1^{(i)} \equiv \mathbf{J}_1^{(i)\mathrm{T}}\mathbf{J}_1^{(i)}/N$. With the M inputs generated to compute the estimate for the generalisation error ($\mathbf{S}_2^{(i)}$) and their correct labels ($\sigma_{\mathbf{B},2}^{(i)}$) we can compute the updates

$$\mathbf{J}_2^{(i)} = \mathbf{J}_1^{(i)} + F_1^{(i)}\,\frac{\sigma_{\mathbf{B},2}^{(i)}}{\sqrt{N}}\,\mathbf{S}_2^{(i)}.$$

This procedure is then iterated. The algorithm can be expressed as a pseudo code in the following way:

(i) $\forall\, i$ make $\mathbf{J}_0^{(i)} = \mathbf{0}$ and $F_0^{(i)} = 1$. Set $\tilde{e}_0 = \frac{1}{2}$, $\tilde{R}_0 = 0$ and $p = 1$.

(ii) $\forall\, i$ make $\mathbf{J}_p^{(i)} = \mathbf{J}_{p-1}^{(i)} + F_{p-1}^{(i)}\,\sigma_{\mathbf{B},p}^{(i)}\,\mathbf{S}_p^{(i)}/\sqrt{N}$.

(iii) $\forall\, i$ make $\phi_p^{(i)} = \sigma_{\mathbf{B},p+1}\,\mathbf{J}_p^{(i)\mathrm{T}}\mathbf{S}_{p+1}^{(i)}/J_p^{(i)}$

(iv) Make $\tilde{e}_p = \frac{1}{M}\sum_i \Theta(-\phi_p^{(i)})$

(v) Make $\tilde{\boldsymbol{\beta}}_p = \sum_i \mathbf{J}_p^{(i)}/|\sum_i \mathbf{J}_p^{(i)}|$

(vi) Compute $b_{\max,p}$, $\mathbf{b}_p$ and $\tilde{\mathbf{P}}_p(b)$ using (13a), (13b) and (13c)

(vii) Set $\tilde{R}_{p,0} = \cos(\pi\tilde{e}_p)$ (or $\tilde{R}_{p-1}$)

(viii) Using (14) and (15), compute

$$\tilde{R}_p \leftarrow \left[\tilde{R}_{p,n} + (1-\tilde{R}_{p,n}^2)\,\frac{2\,\tilde{\mathbf{P}}_p^{\mathrm{T}}\,\mathbf{H}(\mathbf{b}_p,\tilde{R}_{p,n}) - \tilde{e}_p}{2\,\tilde{\mathbf{P}}_p^{\mathrm{T}}\,\boldsymbol{\Gamma}(\mathbf{b}_p,\tilde{R}_{p,n})}\right]_{n|\delta,N_{\max}}$$

(ix) Using (17) and (18), compute

$$\tilde{b}_p^{(i)} = \frac{\tilde{\mathbf{P}}_p^{\mathrm{T}}\,\boldsymbol{\Upsilon}(\phi_p^{(i)},\mathbf{b}_p,\tilde{R}_p)}{\tilde{\mathbf{P}}_p^{\mathrm{T}}\,\mathbf{N}(\phi_p^{(i)},\mathbf{b}_p,\tilde{R}_p)}$$

(x) Make $F_p^{(i)} = \left(\sqrt{Q_p^{(i)}}/\tilde{R}_p\right)\,(\tilde{b}_p^{(i)} - \tilde{R}_p\phi_p^{(i)})$

(xi) IF $p < P$ THEN set $p = p + 1$ and GO TO (ii), else STOP.

4. Results

The curves presented as follows have been computed following the algorithm presented in section 3, considering an ensemble with $M = 4\,000$ students and networks of size $N = 51$. In all cases, the Fast Fourier Transform algorithm was ran considering a grid of size 2^8.

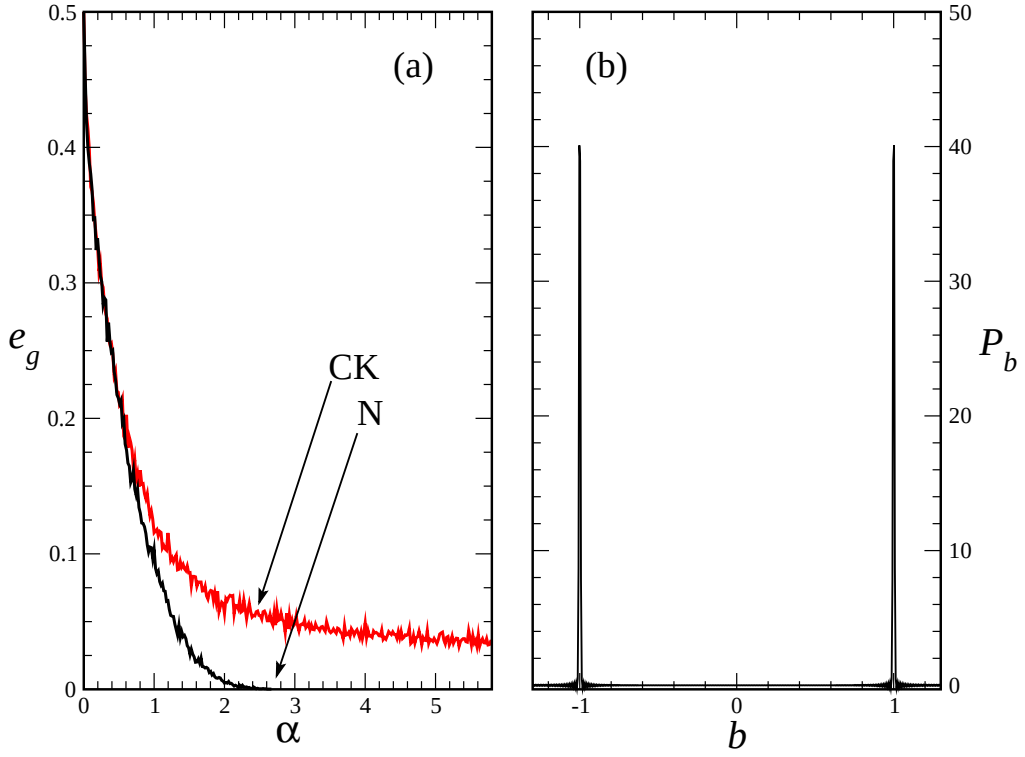

Figure 1. (a) Learning curve for the 1-bit diluted teacher obtained by applying our algorithm (N) and the Caticha-Kinouchi algorithm (CK, red in the on-line version). (b) Plot of the estimate for the density distribution of the teacher's post-synaptic field. The agreement with the analytical solution, $\mathcal{P}_b(b) = \frac{1}{2}\delta(|b|-1)$ is excellent.

4.1. Diluted teachers

The first case we analyse is for the diluted teacher perceptrons with dilutions $m = 1, 5$ (the m-diluted teacher has a synaptic vector with components $B_j = 1$ for all $j \leq m$ and 0 otherwise). These instances were analysed also in [10] and in both cases the CK algorithm did not converge to zero within the time considered. In figure 1 (a) we present the learning curves obtained by our algorithm (N) and the CK algorithm (red in the on-line version). Defining the parameter $\alpha \equiv p/N$ where p is the number of examples presented, it is observed that our algorithm converges after $\alpha = 2$, whilst the CK algorithm still presents an error of 4% even for $\alpha > 5$. In panel (b) we present the estimate $\tilde{\mathcal{P}}_b(b)$ which matches the analytical expression of the probability $\mathcal{P}_b(b) = \frac{1}{2}\delta(|b|-1)$. A similar result has been obtained for $m = 5$ (figure 2). The analytical expression of the probability $\mathcal{P}_b(b) = \frac{10}{32}\delta(|b|-1/\sqrt{5}) + \frac{5}{32}\delta(|b|-3/\sqrt{5}) + \frac{1}{32}\delta(|b|-\sqrt{5})$ is very well approximated by our estimate.

4.2. Teachers constructed from geometric series

Suppose that $B_k \propto r^{-k}$ for any $2 \leq r \in \mathbb{R}$. Given that $\mathrm{sgn}(\sum_{k=1}^{N} S_k r^{-k}) = S_1$ these synaptic vectors will lead to the same algorithm as the 1-bit diluted teacher. If we

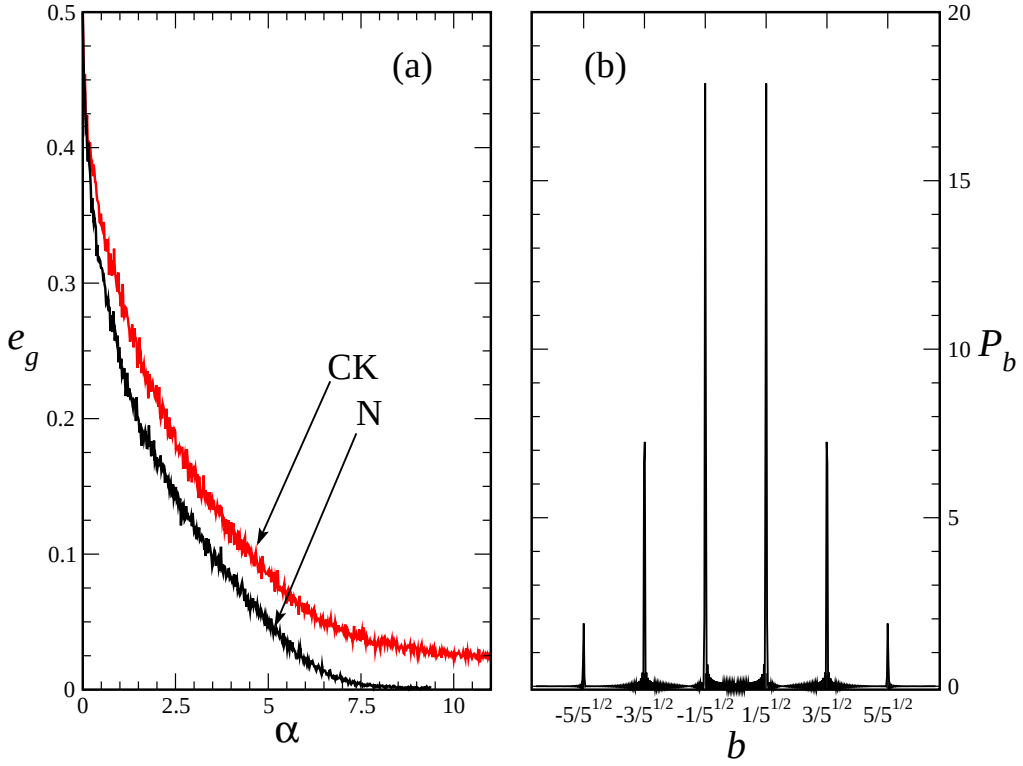

Figure 2. (a) Learning curve for the 5-bit diluted teacher obtained by applying our algorithm (N) and the Caticha-Kinouchi algorithm (CK, red in the on-line version). (b) Plot of the estimate for the density distribution of the teacher's post-synaptic field. The peaks' positions and relative heights are in agreement with the analytical expression $\mathcal{P}_b(b) = \frac{10}{32}\delta(|b| - 1/\sqrt{5}) + \frac{5}{32}\delta(|b| - 3/\sqrt{5}) + \frac{1}{32}\delta(|b| - \sqrt{5})$. The oscillations observed around the peaks are effects due to the finite size of the grid.

consider the vector $\mathbf{B} \propto (1, 1, 2^{-1}, 2^{-1}, 2^{-2}, 2^{-2}, \ldots)^{\mathrm{T}}$ instead, the results obtained are different. Observe that this vector is not diluted and, although the two first entries are fifty percent larger than the second largest, all the entries play a role in the input classification. The limit of the characteristic function is

$$\lim_{N\to\infty} \Phi_N(\hat{b}) = \mathrm{sinc}^2\left(\sqrt{\frac{3}{2}}\,\hat{b}\right)$$

which corresponds to the triangular density function $\mathcal{P}_b(b) = \frac{1}{6}\Theta(\sqrt{6} - |b|)(\sqrt{6} - |b|)$ (where $\mathrm{sinc}(x) \equiv \sin(x)/x$). In figure 3 (a) we present the correspondent learning curves considering our algorithm (N) and the CK algorithm (red line in the on-line version). Even after a long number of examples ($\alpha \simeq 60$) the generic algorithm does not perform as well as the specific algorithm. In panel (b) we present the distributions of post-synaptic fields. Observe the agreement between of the estimate (full line) and the exact value (dashed line, red in the on-line version).

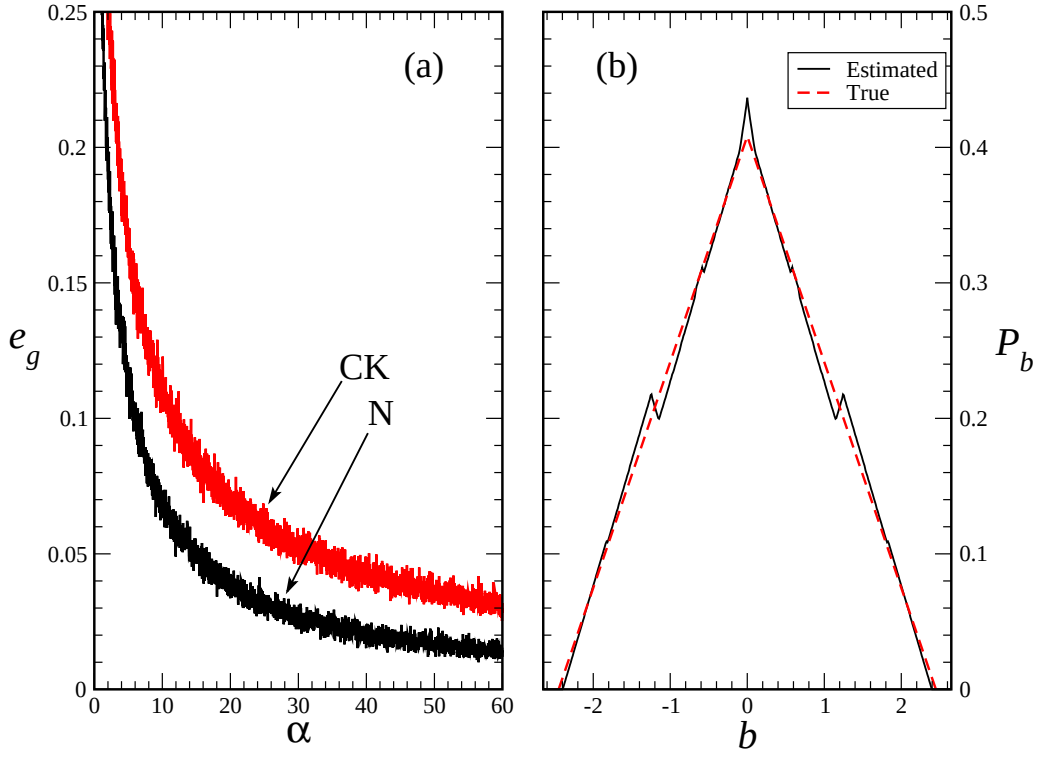

Figure 3. (a) Learning curve for the teacher $\mathbf{B} \propto (1, 1, 2^{-1}, 2^{-1}, 2^{-2}, 2^{-2}, \ldots)^{\mathrm{T}}$ obtained by applying our algorithm (N) and the Caticha-Kinouchi algorithm (CK, red line in the on-line version). (b) Plot of the estimated (full line) and true density distribution of the teacher's post-synaptic field $\mathcal{P}_b(b) = \frac{1}{6}\Theta(\sqrt{6} - |b|)$ (dashed line, red in the on-line version).

4.3. Marginal case: The harmonic sequence

The vector constructed from the harmonic sequence has the components $B_k \propto 1/k$. $\mathcal{P}_b$ cannot be obtained analytically but, according to [11], we know that it is absolutely continuous. For this particular case, the algorithm for the typical case and ours produce indistinguishable results. To illustrate this point we define the variable $X(\alpha) \equiv (e_g^{\mathrm{N}}(\alpha) - e_g^{\mathrm{CK}}(\alpha))/\sigma$, where e_g^{N} is the learning curve obtained by the application of our method, e_g^{CK} is the learning curve obtained by the application of the Caticha-Kinouchi method and $\sigma \simeq 1/\sqrt{M}$ is a parameter associated with the level of noise inherent of the measurement process (a more thorough discussion about this point is presented in the conclusions). In figure 4(a) we present the curve $X(\alpha)$ which is, after a short initial period, bounded in the interval (-1,1). The straightforward conclusion extracted from this result is that the differences between learning curves is of the order of the noise. The only advantage in the application of our method is that, as a byproduct, we obtained a good estimate for the distribution of the teacher's post-synaptic field (panel (b) in full line). We also present in panel (b) the numerically computed distribution $\mathcal{P}_b$, obtained from $\mathsf{FFT}\left[\prod_{k=1}^{51}\cos(\frac{\sqrt{6}}{\pi}\hat{b}/k)\right]_{2^8}$ (dashed line, red in the on-line version).

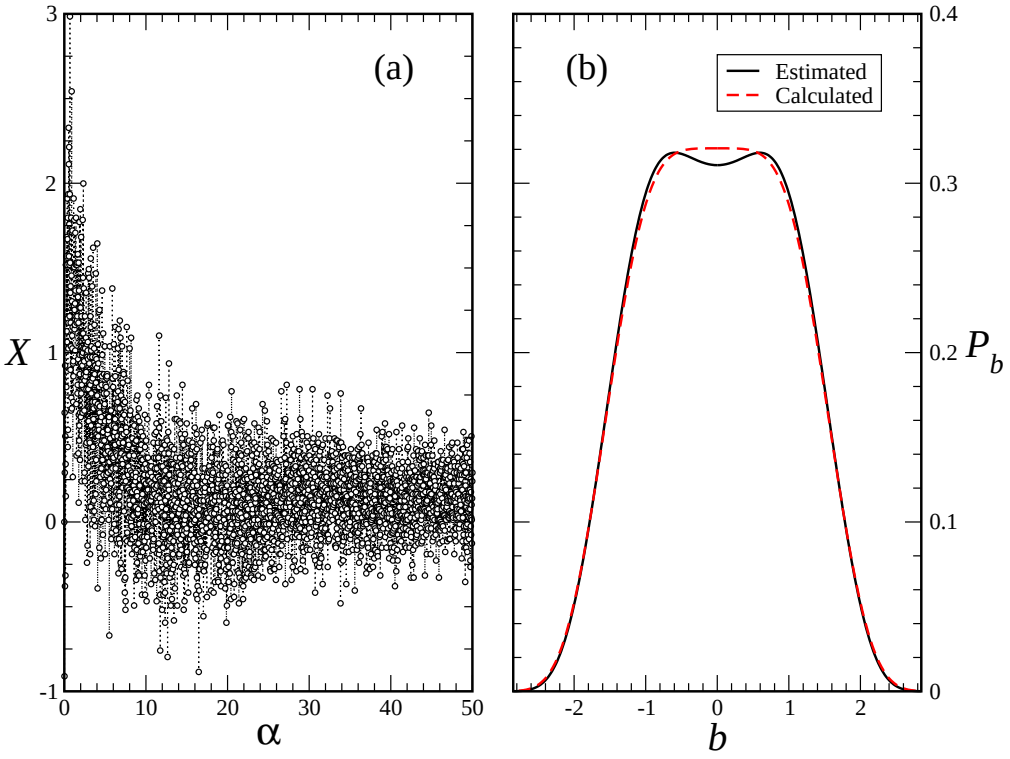

Figure 4. (a) Plot of the variable $X(\alpha) \equiv (e_g^{\rm N}(\alpha) - e_g^{\rm CK}(\alpha))/\sigma$, where $e_g^{\rm N}$ is the learning curve obtained by the application of our method, $e_g^{\rm CK}$ is the learning curve obtained by the application of the Caticha-Kinouchi method and $\sigma \simeq 1/\sqrt{M}$. (b) Plot of the estimated density (full line) and $\mathsf{FFT}\left[\prod_{k=1}^{51}\cos(\frac{\sqrt{6}}{\pi}\hat{b}/k)\right]_{2^8}$ (dashed line, red in the on-line version).

4.4. Typical case

We place under the title *typical case* the teachers whose synaptic vectors have been drawn from a uniform distribution over the N-sphere, i.e. vectors $\mathbf{B}$ whose components are i.i.d. variables. This implies that the components of the unit vector will be at most of $O(1/\sqrt{N})$. If that is the case, the characteristic function of the distribution of post-synaptic fields can be expressed as:

$$\Phi_N(\hat{b}) \simeq \exp\left(-\frac{1}{2}\hat{b}^2\sum_{k=1}^{N}\beta_k^2\right) + O(N^{-1})$$

which is, disregarding corrections of $O(N^{-1})$, $\sqrt{2\pi}$ times a Normal distribution in $\hat{b}$ with unit variance and centred at 0. Trivially, $\mathcal{P}_b(b) = \mathcal{N}(b|0,1)$, which is Caticha and Kinouchi's result. We ran our algorithm on several teachers satisfying these conditions with results indistinguishable (in the sense explained in the previous subsection) to the results obtained by the application of the CK algorithm.

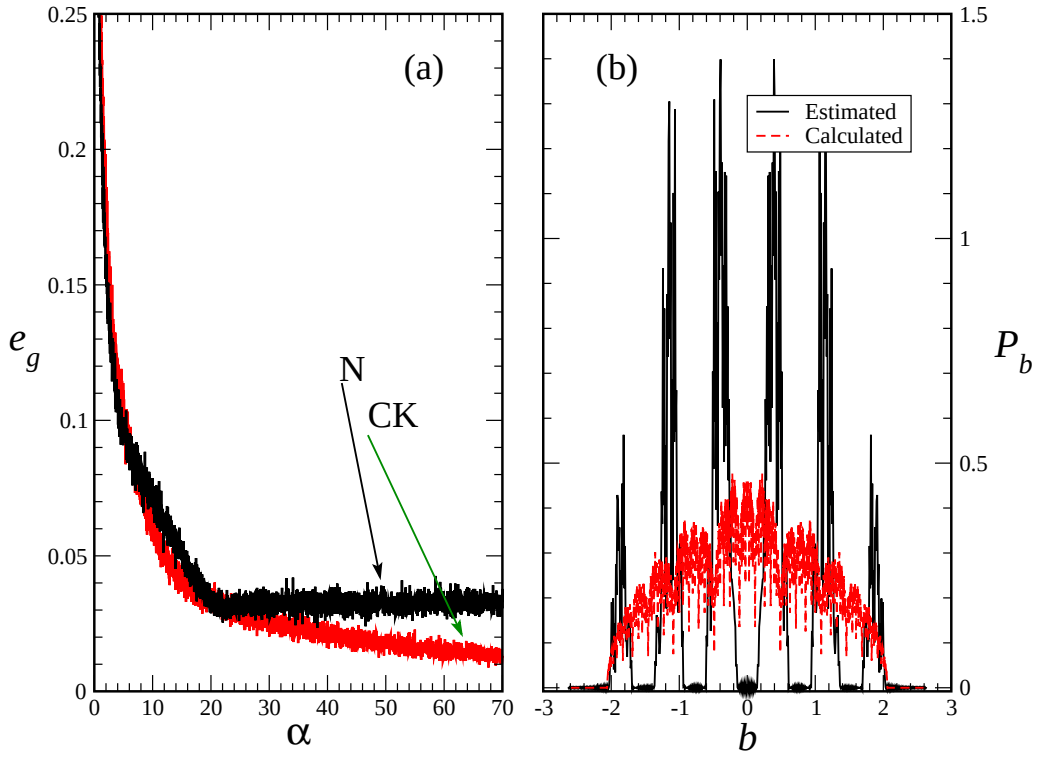

Figure 5. (a) Learning curves for the teacher $B_k \propto r^{-k}$ with $r = (1+\sqrt{5})/2$, obtained by applying our algorithm (N) and the Caticha-Kinouchi algorithm (CK, red line in the on-line version). (b) Plot of the estimated density (full line) and the FFT of the characteristic function computed using the teacher's synaptic vector (dashed line, red in the on-line version). Observe that the *true* value of the density is not a smooth curve of b and that there is no match between this curve and the estimate.

4.5. PV teachers

Our algorithm relies on the estimation of the teacher's post-synaptic field distribution based on a Fourier transform method. If the Fourier transform of the characteristic function $\prod_{k=1}^{N} \cos(\hat{b}\beta_k)$ is singular for a particular vector $\boldsymbol{\beta}$, then the method could produce meaningless results.

Following [11] (and references therein) we found that a geometric sequence $B_k \propto r^{-k}$ with r equal to the reciprocal of a PV number produces a distribution of the field b that is singular. A PV number (or Pisot-Vijayaraghavan number) is an algebraic integer whose Galois conjugates are all less than one in absolute value.

We computed the learning curve for the teacher with entries taken from a geometric series with basis equal to a particular PV number $r = (1+\sqrt{5})/2$, also known as the *Golden Section*. The results are presented in figure 5. It is clear that the CK algorithm produces a better behaved curve (CK, red in the on-line version) than our algorithm (N).

5. Conclusions

We developed and tested a new and improved algorithm for learning realizable rules in perceptrons. The algorithm works in an on-line scenario, using optimally the information available to the student. The updates of the student's synaptic vector are based on an estimate of the distribution of the teacher's post-synaptic field, computed with the aid of an ensemble of students learning in parallel. The algorithm performs better than the one developed for learning optimally a typical rule when the student learns from a diluted teacher. In marginal (harmonic sequence) and typical cases the algorithm matches the performance of the CK algorithm. The algorithm produces less competitive results only when the estimate of the density distribution of the field b is singular. It has been conjectured that this occurs only for geometric sequences with a base equal to a PV number. Given that PV numbers are denumerable, it is expected that the occurrence of one of these cases to be extremely rare.

Observe that both algorithms (CK's and ours) produce an outcome, per example presented, that is either a 0 or a 1, depending on whether the student has produced the correct classification or not. Therefore, the learning curve over one realization of the learning process, i.e. over only on set $\mathscr{S}$ of P examples, produces a discontinuous curve (a simple sequence of 0s and 1s). If the process is repeated M times (like the usual serial version of the algorithms) the averaged curve so obtained is still discontinuous, but with discontinuities of $O(1/\sqrt{M})$. That is why our curves, for both algorithms, look noisy with fluctuations of order $1/\sqrt{4000} \simeq 0.015$ around an average. If $M \to \infty$ the averaged curves finally obtained are continuous. There is no extra cost on running the algorithms in parallel, but there is an important advantage for both algorithms alike. By running in parallel we can generate an estimate for the overlap R as a function of the number of examples the student has received so far.

Our algorithm is more time consuming than CK's only because of the estimation of the distribution $\mathcal{P}_b$ and the quantities that depend upon it. The FFT algorithm has a complexity of $O(G2^G)$, and the averages depending on the distribution are calculated with $O(2^G)$ operations. In our experiments we kept a value of $G = 8$; this value granted estimates of good quality in a reasonable time.

With respect to the chosen size of the system $N = 51$, we found that for this value the curves were produced in a reasonable time and the behaviour of the distribution of synaptic fields mimic closely the asymptotic behaviour expected at the thermodynamic limit. A more comprehensive study on the dependencies over the system size are left for a future work.

Observe that this generalisation of Caticha-Kinouchi's algorithm occurs because we present binary inputs to the network. If the input vectors were formed by real components, drawn from a Normal distribution with zero mean and unit variance, the distribution of the teacher's post-synaptic field becomes Normal and Caticha-Kinouchi's result is recovered.

In all the cases studied we consider the entries of the input vector to be i.i.d

variables. The case when there is some structure in the input vectors will be a subject of future work.

Acknowledgments

I would like to acknowledge the fruitful discussions with Dr R. C. Alamino, Prof. N. Caticha and Prof. K. E. Morrison which have enriched the contents of this article .

Appendix A. Proof of (1)

Consider the synaptic vector $\mathbf{B} \in \mathbb{R}^N$ and the input vectors $\mathbf{S} \in \{\pm 1\}^N$ with i.i.d. entries, distributed according to $\mathcal{P}_{\mathbf{S}}(\mathbf{S}) = \prod_{j=1}^{N} \mathcal{P}(S_j)$ where $\mathcal{P}(S_j = 1) = \mathcal{P}(S_j = -1) = \frac{1}{2}$. We dub a *gauge transformation* any linear transformation that leaves invariant the form of the input vectors and the inner products averaged over $\mathcal{P}_{\mathbf{S}}(\mathbf{S})$, i.e. K is a gauge transformation if

(i) $\forall\, \mathbf{S} \in \{\pm 1\}^N \; K(\mathbf{S}) \in \{\pm 1\}^N$.

(ii) $\left\langle \mathbf{B}^{\mathrm{T}}\mathbf{S} \right\rangle_{\mathbf{S}} = \left\langle K(\mathbf{B})^{\mathrm{T}} K(\mathbf{S}) \right\rangle_{K(\mathbf{S})}$, where $\langle \cdot \rangle_{\mathbf{S}} = \sum_{\{\mathbf{S}\}} \cdot\, \mathcal{P}_{\mathbf{S}}(\mathbf{S})$.

Consider the following transformations T_i and E_{ij} with the following actions

- $T_i\mathbf{B} = (B_1, \ldots, -B_i, \ldots, B_N)^{\mathrm{T}}$
- $E_{ij}(B_1, \ldots, B_i, \ldots, B_j, \ldots\ldots, B_N)^{\mathrm{T}} = (B_1, \ldots, B_j, \ldots, B_i, \ldots\ldots, B_N)^{\mathrm{T}}$.

It is very simple to prove that these transformations, and their products, satisfy (i) and (ii) above and, therefore, they are gauge transformations. We can then transform any vector $\mathbf{B} \in \mathbb{R}^N$ into $\mathbf{B}' = \prod_{j \in \mathscr{N}} T_j \prod_{(j,k) \in \mathscr{O}} E_{jk}(\mathbf{B})$, where $\mathscr{N} = \{1 \leq j \leq N | B_j < 0\}$ is the set of indexes corresponding to negative entries of $\mathbf{B}$, $\mathscr{O} = \{(i,j)\,,\, 1 \leq i < j \leq N |\, |B_i| < |B_j|\}$ is the set of all index pairs linking entries that are not yet decreasingly ordered. In this form the vector $\mathbf{B}'$ so created has entries that satisfy $B'_k \geq B'_l \geq 0$ for all pair of indexes $N \geq l > k \geq 1$.

The joint distribution of the post synaptic fields can be written as

$$\begin{aligned} \mathcal{P}(b,h) &= \sum_{\{\mathbf{S}\}} \mathcal{P}(b,h,\mathbf{S}) = \left\langle \delta\left(b - \frac{\mathbf{B}^{\mathrm{T}}\mathbf{S}}{B}\right) \delta\left(h - \frac{\mathbf{J}^{\mathrm{T}}\mathbf{S}}{J}\right) \right\rangle_{\mathbf{S}} \\ &= \int_{-\infty}^{\infty} \frac{\mathrm{d}\hat{b}}{2\pi}\, \mathrm{e}^{-ib\hat{b}} \int_{-\infty}^{\infty} \frac{\mathrm{d}\hat{h}}{2\pi}\, \mathrm{e}^{-ih\hat{h}} \left\langle \exp\left(i\hat{b}\frac{\mathbf{B}^{\mathrm{T}}\mathbf{S}}{B} + i\hat{h}\frac{\mathbf{J}^{\mathrm{T}}\mathbf{S}}{J} \right) \right\rangle_{\mathbf{S}} . \end{aligned} \tag{A.1}$$

Let us decompose the synaptic vector of the student

$$\mathbf{J} = J_{\mathbf{B}}\boldsymbol{\beta} + J_{\perp}\boldsymbol{\beta}_{\perp} = J_{\mathbf{B}}\left(\boldsymbol{\beta} + \epsilon\boldsymbol{\beta}_{\perp}\right), \tag{A.2}$$

where $\epsilon \equiv J_{\perp}/J_{\mathbf{B}}$, $\boldsymbol{\beta} \equiv \mathbf{B}/B$ and $\boldsymbol{\beta}_{\perp}$ is a random unit vector laying on the hyper-plane perpendicular to $\mathbf{B}$. If the student learns, we can expect that $\epsilon \ll 1$. Using (A.2) we have that

$$R = \frac{\mathbf{J}^{\mathrm{T}}\mathbf{B}}{J\,B} = \frac{J_{\mathbf{B}}}{\sqrt{J_{\mathbf{B}}^2 + J_{\perp}^2}} \simeq 1 - \frac{1}{2}\epsilon^2 + O(\epsilon^4).$$

It is easy to demonstrate that the frequently used quantity $1-R^2$, related to the projection of the student's synaptic vector into the hyper-plane perpendicular to the teacher's synaptic vector is

$$1-R^2 \simeq \epsilon^2 + O(\epsilon^3)$$

or equivalently

$$\epsilon \simeq \sqrt{1-R^2} + O\left[(1-R^2)^{\frac{3}{2}}\right].$$

Each component of the unit vector $\boldsymbol{\eta} \equiv \mathbf{J}/J$ can be approximated by

$$\begin{aligned}\eta_k &\equiv \frac{J_{\mathbf{B}}\beta_k + J_\perp \nu_k}{\sqrt{J_{\mathbf{B}}^2 + J_\perp^2}} \\ &\simeq R\beta_k + \epsilon\nu_k + O(\epsilon^3),\end{aligned} \tag{A.3}$$

where $\nu_k \equiv [\boldsymbol{\beta}_\perp]_k$

The expectation in (A.1) is

$$\begin{aligned}\left\langle \exp\left(i\hat{b}\frac{\mathbf{B}^{\mathrm{T}}\mathbf{S}}{B} + i\hat{h}\frac{\mathbf{J}^{\mathrm{T}}\mathbf{S}}{J}\right)\right\rangle_{\mathbf{S}} &= \prod_{k=1}^{N}\frac{1}{2}\sum_{s=\pm 1}\exp(i\hat{b}\beta_k s + i\hat{h}\eta_k s) \\ &= \prod_{k=1}^{N}\cos(\hat{b}\beta_k + \hat{h}\eta_k)\end{aligned}$$

and by using (A.3) we have that

$$\hat{b}\beta_k + \hat{h}\eta_k \simeq (\hat{b} + \hat{h}R)\beta_k + \hat{h}\epsilon\nu_k + O(\epsilon^3).$$

Up to $O(\epsilon^3)$ we have that

$$\begin{aligned}\cos(\hat{b}\beta_k + \hat{h}\eta_k) &\simeq \cos((\hat{b} + \hat{h}R)\beta_k + \hat{h}\epsilon\nu_k) + O(\epsilon^3) \\ &\simeq \cos((\hat{b} + \hat{h}R)\beta_k)\exp\left(-\frac{\hat{h}^2}{2}\epsilon^2\nu_k^2\right)\left[1 - \hat{h}\epsilon\,\tan((\hat{b} + \hat{h}R)\beta_k)\nu_k\right] + O(\epsilon^3)\end{aligned}$$

where we used that $\epsilon\exp(\hat{h}^2\epsilon^2\nu_k^2/2) \simeq \epsilon + O(\epsilon^3)$. Thus, by applying the change of variables $\hat{b} + \hat{h}R \to \hat{b}$ and disregarding terms of order ϵ^3, we obtain

$$\mathcal{P}(b,h) \simeq \int_{-\infty}^{\infty}\frac{\mathrm{d}\hat{h}\,\mathrm{d}\hat{b}}{4\pi^2}\,\exp\left(-\frac{1-R^2}{2}\hat{h}^2 - i\hat{h}(h - bR) - i\hat{b}b\right)\prod_{k=1}^{N}\left[1 - \epsilon\hat{h}\tan(\hat{b}\beta_k)\nu_k\right]\cos(\hat{b}\beta_k)$$

Observe that

$$\begin{aligned}\prod_{k=1}^{N}\left[1 - \epsilon\hat{h}\,\tan(\hat{b}\beta_k)\nu_k\right] &\simeq 1 - \epsilon\hat{h}\sum_{k=1}^{N}\tan(\hat{b}\beta_k)\nu_k + \epsilon^2\hat{h}^2\sum_{j<k}\tan(\hat{b}\beta_j)\nu_j\tan(\hat{b}\beta_k)\nu_k + O(\epsilon^3) \\ &\simeq 1 - \epsilon\hat{h}\sum_{k=1}^{N}\tan(\hat{b}\beta_k)\nu_k + \frac{\epsilon^2\hat{h}^2}{2}\left[\left(\sum_{k=1}^{N}\tan(\hat{b}\beta_k)\nu_k\right)^2 - \sum_{k=1}^{N}\tan(\hat{b}\beta_k)^2\nu_k^2\right] + O(\epsilon^3)\,.\end{aligned}$$

Without lose of generality we can suppose that the entries of the vector $\boldsymbol{\beta}_\perp$ satisfy the equation $\nu_k = \kappa_k/(\sigma_N\sqrt{N})$ where κ_k are random deviates distributed in [-1,1] according

to $\mathcal{P}_\kappa(\boldsymbol{\kappa}) \propto \delta(\boldsymbol{\kappa}^{\mathrm{T}}\boldsymbol{\beta}) \prod_{k=1}^{N} \Theta(1-\kappa_k^2)$ and $0 < \sigma_N^2 \equiv \frac{1}{N}\sum_{j=1}^{N}\kappa_j^2 \leq 1$. To bound the parameter σ_N^2 observe that the expected value of κ_k^2 and κ_k^4 are

$$\langle \kappa_k^2 \rangle = \frac{1}{\mathscr{N}} \int_{-\infty}^{\infty} \mathrm{d}x \prod_{j=1}^{N} \mathrm{sinc}(x\beta_j) \left(1 + 2\,\frac{\cot(x\beta_k)}{x\beta_k} - \frac{2}{x^2\beta_k^2}\right)$$

$$\langle \kappa_k^4 \rangle = \frac{1}{\mathscr{N}} \int_{-\infty}^{\infty} \mathrm{d}x \prod_{j=1}^{N} \mathrm{sinc}(x\beta_j) \left(1 + 4\,\frac{\cot(x\beta_k)}{x\beta_k} - \frac{12}{x^2\beta_k^2} - 24\,\frac{\cot(x\beta_k)}{x^3\beta_k^3} + \frac{24}{x^4\beta_k^4}\right)$$

where $\mathscr{N} \equiv \int_{-\infty}^{\infty} \mathrm{d}x \prod_{j=1}^{N} \mathrm{sinc}(x\beta_j)$ is the normalisation constant. Therefore the following additions can be approached by:

$$\begin{aligned}
\frac{1}{N}\sum_{k=1}^{N}\langle \kappa_k^2 \rangle &\simeq \frac{1}{\mathscr{N}} \int_{-\infty}^{\infty} \mathrm{d}x \prod_{j=1}^{N} \mathrm{sinc}(x\beta_j) \frac{1}{N}\sum_{k=1}^{N} \left(\frac{1}{3} - \frac{2}{45}\,x^2\beta_k^2 + O(\beta_k^4)\right) \\
&\simeq \frac{1}{3} + O(N^{-1}) \\
\frac{1}{N}\sum_{k=1}^{N}\langle \kappa_k^4 \rangle &\simeq \frac{1}{\mathscr{N}} \int_{-\infty}^{\infty} \mathrm{d}x \prod_{j=1}^{N} \mathrm{sinc}(x\beta_j) \frac{1}{N}\sum_{k=1}^{N} \left(\frac{1}{5} - \frac{4}{105}\,x^2\beta_k^2 + O(\beta_k^4)\right) \\
&\simeq \frac{1}{5} + O(N^{-1})
\end{aligned}$$

and thus $\sigma_N^2 \simeq \frac{1}{3} \pm \sqrt{\frac{4}{45}} + O(N^{-1})$. We can conclude that the parameter σ_N^2 is strictly positive and expected to be close to $\frac{1}{3}$ independently from $\boldsymbol{\beta}$. Thus

$$\begin{aligned}
\prod_{k=1}^{N}(1 - \epsilon\hat{h}\,\tan(\hat{b}\beta_k)\nu_k) \simeq 1 - \frac{\epsilon\hat{h}}{\sigma_N\sqrt{N}} \sum_{k=1}^{N} \tan(\hat{b}\beta_k)\kappa_k + \\
+ \frac{\epsilon^2\hat{h}^2}{2\sigma_N^2 N} \left[\left(\sum_{k=1}^{N} \tan(\hat{b}\beta_k)\kappa_k\right)^2 - \sum_{k=1}^{N} \tan(\hat{b}\beta_k)^2\kappa_k^2\right] + O(\epsilon^3)\,.
\end{aligned}$$

From the Taylor expansion of the tangent we have that

$$\sum_{j=1}^{N} \tan(\hat{b}\beta_j)\kappa_j = \sum_{\ell=0}^{\infty} C_\ell\, \hat{b}^{2\ell+1} \sum_{j=1}^{N} \beta_j^{2\ell+1}\kappa_j$$

where $C_\ell > 0$ and observe that for $\ell = 0$, the first term, $0 = \sum_{j=1}^{N} \beta_j\kappa_j < 1$ just because $\boldsymbol{\beta}$ and $\boldsymbol{\kappa} = \sqrt{N}\sigma_N\boldsymbol{\beta}_\perp$ are perpendicular, and the other terms, $\ell \geq 1$, can be bound by

$$\left|\sum_{j=1}^{N} \beta_j^{2\ell+1}\kappa_j\right| < \sum_{j=1}^{N} \beta_j^{2\ell+1}|\kappa_j| < \sum_{j=1}^{N} \beta_j^2 = 1\,,$$

due to the facts that $1 \geq \beta_k \geq \beta_{k+1} \geq 0$ and $|\kappa_j| < 1$, thus

$$\left|\sum_{j=1}^{N} \tan(\hat{b}\beta_j)\kappa_j\right| < \sum_{\ell=0}^{\infty} C_\ell\, |\hat{b}^{2\ell+1}| = \tan(|\hat{b}|).$$

In a similar fashion

$$\sum_{j=1}^{N} \tan(\hat{b}\beta_j)^2 \kappa_j^2 = \sum_{\ell=1}^{\infty} D_\ell \, \hat{b}^{2\ell} \sum_{j=1}^{N} \beta_j^{2\ell} \kappa_j^2 < \sum_{\ell=1}^{\infty} D_\ell \, \hat{b}^{2\ell} \sum_{j=1}^{N} \beta_j^2 = \tan(\hat{b})^2,$$

where $D_\ell > 0$ for all $\ell = 1, 2, \ldots$. Putting all things together and disregarding terms of order ϵ^3, we have that, for a sufficiently large N,

$$\left| 1 - \prod_{k=1}^{N} [1 - \epsilon \hat{h} \, \tan(\hat{b}\beta_k)\nu_k] \right| < 3 \, |\hat{h} \tan(\hat{b})| \, \frac{\epsilon}{\sqrt{N}}$$

We finally have, disregarding corrections of $O(\epsilon^3, \epsilon/\sqrt{N})$, the estimate to the joint probability is

$$\mathcal{P}(b, h) \simeq \mathcal{N}(h|bR, 1 - R^2) \, \mathcal{P}_b(b), \tag{A.4}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ is a Gaussian distribution in x, centred at μ with variance σ^2 and $\mathcal{P}_b(b)$ is the Fourier transform of $\lim_{N\to\infty} \prod_{k=1}^{N} \cos(\hat{b}\beta_k)$. In other words, $\lim_{N\to\infty} \prod_{k=1}^{N} \cos(\hat{b}\beta_k)$ is the *characteristic function* of $\mathcal{P}_b(b)$.

References

[1] Franklin S and Graesser A 1996 *Proceedings of the third international workshop on agent theories, architecture and language* (Springer-Verlag)
[2] Seung H S, Sompolinsky H and Tishby N 1992 *Phys Rev A* **45** 6056
[3] Kinouchi O and Caticha N 1992 *J Phys A* **25** 6243
[4] Watkin L H T, Rau A and Biehl M 1993 *Rev Mod Phys* **65** 499
[5] Reents G and Urbanczik R 1998 *Phys Rev Lett* **80** 5445
[6] Copelli M, Kinouchi O and Caticha N 1996 *Phys Rev E* **53** 6341
[7] Copelli M, Eichhorn R, Kinouchi O, Biehl M, Simonetti R, Riegler P and Caticha N 1997 *Europhys Lett* **37** 427
[8] Uezu T, Maeda Y and Yamaguchi S 2006 *J Phys Soc Japan* **75** 114007
[9] Uezu T, Miyoshi S, Izuo M, Okada M 2007 *J Phys Soc Japan* **76** 114006
[10] Neirotti J 2010 *J Phys A* **43** 015101
[11] Morrison K E, *Random Walks with Decreasing Steps*, 1998
[12] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical Recipes* (Cambridge Univ Press)