Optimal information storage and the distribution of synaptic weights: Perceptron *vs.* Purkinje cell

Supplemental Material: Computation of perceptron capacity and synaptic weight distribution

Nicolas Brunel[1], Vincent Hakim[2], Philippe Isope[3], Jean-Pierre Nadal[2] and Boris Barbour[4]

August 23, 2004

[1] Neurophysique et Physiologie, Université René Descartes, 45 rue des Saints Pères, 75270 Paris Cedex 06, France. brunel@biomedicale.univ-paris5.fr

[2] Laboratoire de Physique Statistique, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.

[3] Dept. of Psychiatry, Kinsmen Lab, 4N11-2255 Wesbrook Mall, Vancouver, BC, V6T 1Z3, Canada.

[4] Laboratoire de Neurobiologie, Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris Cedex 05, France.

The capacity and the synaptic weight distribution are computed using statistical mechanics techniques developed in a classical paper by Gardner (1988) for perceptrons with unconstrained continuous weights and random patterns. The results obtained in the literature are first translated to the situation of the Purkinje cell, by calculating its storage capacity. In particular, this requires imposing a sign constraint on synapses (because parallel fibre inputs are excitatory), and introducing different coding levels for the input and output (to account for different levels of activity in granule cells and Purkinje cells). Secondly, the distribution of synaptic weights is computed, using a generalisation of the approach of Kohler and Widmaier (1991).

The model

A large number N (with $N \sim 150000$) of parallel fibres make synaptic contacts with a given Purkinje cell. A subset of active granule cells (or parallel fibres) constitute what we call a *pattern*. It is mathematically described by a set of N numbers G_j taking the value 0 or 1 depending on whether the j-th granule cell is active (*i.e.*, emits a spike in an interval of the order of 1 (a few) ms, $G_j = 1$) or inactive (*i.e.*, does not emit a spike, $G_j = 0$).

The Purkinje cell should give the desired response, that is emit or not emit an action potential in the $\sim$1ms interval when it receives a given pattern. Mathematically, the activity of the Purkinje cell is given by the number P with $P = 1$ if the Purkinje cell is active or $P = 0$ if the Purkinje cell is inactive. We call an *association* the couple $(\{G_j\}, P)$ of a given pattern with the appropriate response.

We assume that the input patterns of activity and the associated outputs are not correlated. This is modelled by drawing randomly and independently the granule cell activities for each pattern with probabilities, $\text{Prob}(G_j^\mu = 1) = f$, $\text{Prob}(G_j^\mu = 0) = 1 - f$, where G_j^μ represents the activity (spike emission or not) of the Granule cell j ($j = 1, \ldots, N$ where N is the number of granule cells) in pattern μ ($\mu = 1, \ldots, p$ where p is the number of patterns). Similarly, the Purkinje cell activity is randomly drawn with probability f' in each association.

An active parallel fibre produces a depolarisation given by the strength of its synapse with the Purkinje cell. In the perceptron model, the total Purkinje cell depolarisation produced by a given pattern is simply obtained by the sum of these depolarisations. When the total depolarisation is larger than a threshold depolarisation θ, an action potential is emitted. Noise (due to stochastic ion channel dynamics, variability in single synaptic responses, errors in the input pattern, or other synaptic inputs) could blur the dis-

tinction between the two classes of patterns (*i.e.*, those with or without spike emission). Hence, we introduce a stability requirement: the synaptic weights should be such that the total depolarisation is $\kappa\,\mathrm{mV}$ below threshold for the patterns that should not elicit a spike and $\kappa\,\mathrm{mV}$ above threshold for those that elicit a spike. In mathematical terms, the conditions for storing associations can be expressed as,

$$\begin{aligned} \sum_j w_j G_j - \theta > \kappa \text{ when } P = 1, \\ \sum_j w_j G_j - \theta < -\kappa \text{ when } P = 0. \end{aligned} \tag{1}$$

In the statistical-mechanics approach to learning, results for large but finite devices are obtained by considering the mathematical limit of larger and larger devices, the so-called "thermodynamic limit". In the present case, this corresponds to considering perceptrons with larger and larger numbers N of synapses. In order to obtain a well-defined limiting behaviour, one should thus prescribe how the perceptron parameters change as N is increased. We take the threshold θ to be a fixed, finite quantity as $N \to \infty$. Since about fN parallel fibres are active in each pattern, synaptic strengths should scale as θ/fN as $N \to \infty$. Hence, we define $w_i = W_i/N$ where W_i remains finite, of order θ/f, when $N \to \infty$. Fluctuations in the total depolarisation from pattern to pattern will then be of order $\theta/\sqrt{fN}$. Hence, we define $\kappa = K/\sqrt{N}$. In principle, the w_i's and κ should be vanishingly small in the thermodynamic limit. With the parameters $N = 150000$ and $f \simeq 0.004$, one obtains $fN \simeq 600$, which is reasonably large. The corresponding value of the stability constant κ is found to be approximately $2\,\theta/\sqrt{fN} \simeq 0.8\,\mathrm{mV}$, that is, about 10 % of threshold. As reported in the main text, simulations confirm that the theoretical results are applicable in this regime, with the small difference between the theoretical and numerical results providing an estimate of finite size corrections.

Using a condensed form of Eq. (1) valid both for $P = 0$ and $P = 1$ and rewritten in terms of the scaled parameters, we form an indicator function that takes the value of 1 if the storage conditions for a particular association are satisfied and 0 if not

$$S(P, \{W_j\}, \{G_j\}) = \Theta\left[(2P-1)\left(\sum_j \frac{W_j}{\sqrt{N}} G_j - \theta\sqrt{N}\right) - K\right] \tag{2}$$

where $\Theta(x)$ is the Heaviside function, $\Theta(x) = 1$ if $x > 0$ and $\Theta(x) = 0$ otherwise.

Synaptic weights W_j are allowed positive or zero values with *a priori* distribution $dr(W)$ normalised to one for convenience. The fraction of the N-dimensional space of all possible weights, where p given associations $(\{G_j^\mu\}, P^\mu, \mu = 1, \ldots, p)$ are satisfied (2) is

$$V = \int \prod_j dr(W_j) \prod_\mu S(P^\mu, \{W_j\}, \{G_j^\mu\}) \tag{3}$$

If $V = 0$, no synaptic strengths can be found such that all the constraints are satisfied. The largest number of patterns p_c for which $V \neq 0$ gives the perceptron maximal capacity (per synapse) $\alpha_c = p_c/N$. It is also called the *critical capacity* in the literature since below p_c synaptic weights can be found such that the perceptron does not make errors while this is no longer possible for $p > p_c$.

Mathematical formalism

The main mathematical task consists of obtaining V [Eq. (3)] given p randomly drawn associations. One important point is that one should compute the typical value of V, which here is different from the average value of V. The typical value is obtained by computing the average of $\log V$, using the so called 'replica trick' borrowed from the statistical mechanics of disordered media (Mézard et al., 1987). The identity

$$\log V = \lim_{n \to 0} \frac{V^n - 1}{n}$$

leads to the computation being performed in three steps:
(i) One first calculates the average of V^n where n is an integer (thus the volume of n 'replicas' of the system),
(ii) One then makes an analytical continuation to non-integer n;
(iii) Finally, one performs the limit $n \to 0$.

In the present case, one should calculate

$$\langle V^n \rangle = \langle \int \prod dr(W_j^a) \prod_{\mu,a} S(P^\mu, \{W_j^a\}, \{G_j^\mu\}) \rangle \tag{4}$$

where $a = 1, \ldots, n$ is a 'replica' index and the angle brackets mean that the average over p randomly drawn associations should be performed.

The calculation follows standard steps (Gardner, 1988; Engel and van den Broeck, 2001). The large N limit is taken advantage of by defining global

quantities, the 'order parameters', which have well-defined limits when $N \to +\infty$ [see e.g. Mézard et al. (1987)]. Here, we define

$$\frac{1}{N}\sum_j W_j^a = \frac{\theta}{f} + \frac{M^a}{\sqrt{N}} \equiv \overline{W} + \frac{M^a}{\sqrt{N}} \tag{5}$$

$$\frac{1}{N}\sum_j W_j^a W_j^b = q^{ab}, \tag{6}$$

together with conjugate parameters $\hat{M}^a$ and $\hat{q}^{ab}$. The first term in the r. h. s. of Eq. (5) represents the average scaled synaptic weight $\overline{W}$ ($\equiv N\overline{w}$) in the large N limit, while the second term represents a $1/\sqrt{N}$ correction that is needed for fine tuning the average synaptic weight relative to the threshold when there are unequal numbers of patterns yielding 1 and 0 outputs, *i.e.*, $f' \neq 0.5$ (see below). q^{ab} represents the typical overlap between two weight vectors W_j that are both solutions to the storage problem, *i.e.*, that both obey Eqs. (2). Introduction of the order parameters replaces the integral over the weights in Eq. (4) by an integral over order parameters.

The evaluation of $\langle V^n \rangle$ can be done in the large N limit by the method of steepest descent: when $N \to +\infty$, the value of the integral comes dominantly from the neighbourhood of a saddle-point, where the integrand has a vanishing gradient. In the case considered here of a convex synaptic weight space in the error-free regime (*i.e.*, below the critical capacity), the saddle-point can be searched under the replica-symmetric form,

$$M^a = M, \quad q^{aa} = Q, \quad q^{ab} = q \text{ if } a \neq b \tag{7}$$

and similarly for the conjugate variables. Q represents the typical norm of a weight vector that is a solution of the storage problem. It also represents the variance of the distribution of synaptic weights. q represents the typical overlap between two vectors in solution space. As the space of solutions becomes small, q approaches Q.

Using the replica-symmetric ansatz, the integral over order parameters becomes

$$\langle V^n \rangle = \int d\hat{q}dqd\hat{Q}dQd\hat{M}dM \exp\left[\mathcal{F}(\hat{q}, q, \hat{Q}, Q, \hat{M}, M)\right], \tag{8}$$

where $\mathcal{F}$ is, in the limit $n \to 0$,

$$\mathcal{F}/n = N\left[-\hat{Q}Q + \frac{1}{2}\hat{q}q + \overline{W}\hat{M} + \alpha\mathcal{Z}_A(Q, q, M) + \mathcal{Z}_W(\hat{Q}, \hat{q}, \hat{M})\right] + O(\sqrt{N}) \tag{9}$$

The functions $\mathcal{Z}_W, \mathcal{Z}_A$ are given by the explicit expressions

$$\mathcal{Z}_W = \int_{-\infty}^{+\infty} Du \ln \left(\int dr(W) \exp \left[(\hat{Q} - \frac{\hat{q}}{2}) W^2 + W(u\sqrt{\hat{q}} - \hat{M}) \right] \right) \quad (10)$$

$$\mathcal{Z}_A = \int_{-\infty}^{+\infty} Du \left(f' \ln \left\{ H \left[\frac{K - fM + u\sqrt{qf(1-f)}}{\sqrt{f(1-f)(Q-q)}} \right] \right\} + (1-f') \ln \left\{ H \left[\frac{K + fM + u\sqrt{qf(1-f)}}{\sqrt{f(1-f)(Q-q)}} \right] \right\} \right). \quad (11)$$

Conventional notations have been introduced for $Du \equiv du/\sqrt{2\pi} \exp(-u^2/2)$ and for the function H defined by

$$H(x) = \int_x^{+\infty} \frac{dt}{\sqrt{2\pi}} \exp(-t^2/2) \quad (12)$$

As stated above, the computation of the integral in Eq. (4) is completed by using the steepest descent technique in the large N limit. Six saddle point equations are obtained by differentiating Eq. (9) with respect to the order parameters $Q, q, M, \hat{Q}, \hat{q}$ and $\hat{M}$.

These six equations are essentially those obtained in Gutfreund and Stein (1990), with two differences (apart from notational choices). A minor difference is that f' is assumed equal to f in Gutfreund and Stein (1990). A more important difference is that threshold is assumed in Gutfreund and Stein (1990) to scale like the stability parameter. Hence, $d\mathcal{Z}_W/d\hat{M} = 0$ is obtained in Gutfreund and Stein (1990), instead of the correct formula $d\mathcal{Z}_W/d\hat{M} = \overline{W}$ that follows from Eq. (9).

For $f' = 0.5$, $M = 0$ is an exact solution of the obtained equations. This means that for $f' = 0.5$, the average synaptic weight is equal to the threshold divided by fN, up to $1/N^2$ corrections. On the other hand, when $f' < 0.5$ we find that $M < 0$. The synaptic weights have to be on average slightly smaller than θ/fN (by a correction proportional to $1/N^{3/2}$), to accommodate the fact that more patterns lead to 0 output than to 1 output.

Capacity

The six saddle-point equations determine the six unknowns Q, q, v and $\hat{Q}$, $\hat{q}$, $\hat{v}$ as a function of the capacity per synapse and the physiological parameters f, f', θ, K. We consider the case of a flat measure $dr(W) \propto dw$ on real positive weights since the parallel fibre/Purkinje cell synapses are excitatory. For $\alpha = 0$ one finds that the order parameters are $\hat{M} = 1/\overline{w}$, $\hat{Q} = 0$, $\hat{q} = 0$,

$Q = 2\overline{W}^2$, $q = \overline{W}^2$. For $0 < \alpha < \alpha_c$ where α_c is the critical capacity, the six saddle-point equations were solved numerically.

As α approaches the critical capacity α_c, the set of possible synaptic weights shrinks to a single point and q tends toward Q. This allows for a simplification of the saddle-point equations and an analysis of the limiting solution when $\alpha \to \alpha_c$. The quantities $\hat{Q}, \hat{q}$ and $\hat{M}$ diverge when $q \to Q$,

$$2\hat{Q} \sim \hat{q} \sim \frac{C}{(Q-q)^2}, \hat{q} - 2\hat{Q} \sim \frac{A}{Q-q} \text{ and } \hat{M} \sim \frac{B\sqrt{C}}{Q-q}. \tag{13}$$

The three saddle-point equations obtained by varying $\mathcal{F}$ with respect to the order parameters provide the following expressions for the prefactors of the diverging terms

$$\begin{aligned} C &= \alpha_c Q \left\{ f'[(1+\tau_-^2)H(\tau_-) - \tau_- G(\tau_-)] \right. \\ &\quad \left. +(1-f')[(1+\tau_+^2)H(\tau_+) - \tau_+ G(\tau_+)] \right\} && (14) \\ A &= \alpha_c \left[f' H(\tau_-) + (1-f') H(\tau_+) \right] && (15) \\ 0 &= f'[G(\tau_-) - \tau_- H(\tau_-)] - (1-f')[G(\tau_+) - \tau_+ H(\tau_+)] && (16) \end{aligned}$$

where the function H is defined by Eq. (12), G is the Gaussian $G(x) = \exp(-x^2/2)/\sqrt{2\pi}$, and the quantity $\tau_\pm$ denotes

$$\tau_\pm = -\frac{K \pm fM}{\sqrt{Qf(1-f)}} \tag{17}$$

The divergence of $(\hat{Q}, \hat{q}, \hat{M})$ at the critical capacity allows for an evaluation of the remaining three saddle-point equations and provides the three supplementary relations,

$$\begin{aligned} Q &= \overline{W}^2 \frac{(1+B^2)H(B) - BG(B)}{[G(B) - BH(B)]^2} && (18) \\ A &= H(B) && (19) \\ C &= \overline{W}^2 \frac{H(B)^2}{[G(B) - BH(B)]^2} && (20) \end{aligned}$$

Eqs. (14-16,18-20) are the six equations that need to be solved to determine the five unknowns A, B, C, Q, M and the critical capacity α_c as functions of the perceptron characteristics N, θ, K, f and f'.

To relate B to the perceptron parameters, we first consider the three Eqs. (14-16) and write $\tau_- = -y + z$, $\tau_+ = -y - z$ with

$$y = \frac{K}{\sqrt{Qf(1-f)}}, \; z = \frac{fM}{\sqrt{Qf(1-f)}} \tag{21}$$

A numerical solution of Eq. (16) determines z as a function of y. Then, Eqs. (14,15) give the ratio AQ/C as of function of y,

$$\frac{AQ}{C} = F_1(y) \tag{22}$$

with

$$F_1(y) = \frac{f'H(\tau_-) + (1-f')H(\tau_+)}{f'[(1+\tau_-^2)H(\tau_-) - \tau_- G(\tau_-)] + (1-f')[(1+\tau_+^2)H(\tau_+) - \tau_+ G(\tau_+)]} \tag{23}$$

This ratio can also be computed from Eqs. (18-20) as a function of the parameter B. The comparison of these two determinations determines y as a function of B.

$$y_1(B) = F_1^{-1}\left[(1+B^2) - B\frac{G(B)}{H(B)}\right] \tag{24}$$

From its very definition and the formula (18) for Q, y can also be expressed as a function of B and the reliability parameter ρ. The compatibility of these two formulas for y relates B to the reliability parameter,

$$\rho = \frac{K}{\overline{W}\sqrt{f(1-f)}} = \frac{\kappa}{\theta}\sqrt{\frac{fN}{1-f}} = \frac{y_1(B)}{y_2(B)} \tag{25}$$

with

$$y_2(B) = \frac{G(B) - BH(B)}{\sqrt{(1+B^2)H(B) - BG(B)}} \tag{26}$$

$B = 0$ for $\rho = 0$, and B increases monotonically with ρ. Note that B also depends upon f' through $F_1(y)$, [Eq. (23)]. However, this dependence is weak.

Finally, the critical capacity can be obtained by comparing the two expressions of A, Eqs. (15,19)

$$\alpha_c = \frac{H(B)}{[f'H(\tau_-) + (1-f')H(\tau_+)]} \tag{27}$$

Figure 7A of the main text shows how the critical capacity depends on the reliability parameter ρ. Note that the critical capacity depends on the reliability parameter mostly through the numerator $H(B)$, which represents the fraction of non-silent synapses, which goes from 0.5 when $\rho = 0$ to 0 as ρ becomes large. It depends on the output coding level f' mostly through the denominator $f'H(\tau_-) + (1-f')H(\tau_+)$. Decreasing f' increases the critical

capacity, but the information stored in the system stays essentially constant as f' is decreased [when f' becomes very small the information drops significantly (Gardner, 1988)]. This is due to the fact that the average information per pattern decreases when f' decreases.

Distribution of synaptic weights

The probability for a given synapse (here given the label 1) to take a particular value W can be computed by looking at the fraction of the volume in weight space that stores a given set of patterns, given by Eq. (3), and for which $W_1 = W$. Thus, the probability distribution $P(W)$ can be written

$$P(W) = \frac{1}{V} \int \prod_j dr(W_j) \delta(W - W_1) \prod_\mu S(P^\mu, \{W_j\}, \{G_j^\mu\}), \tag{28}$$

where V is given by Eq. (3). It is difficult to compute the average, for p randomly drawn associations, of $P(W)$, which is the ratio of two integrals. The difficulty can be circumvented by using the replica formalism again, under a slightly different form. One introduces $n = 1 + (n-1)$ replicas, where in the first replica we impose the constraint that the first synaptic weight is equal to W_1, while no constraint is imposed on the other $n-1$ replicas. In the limit $n \to 0$, the ratio of Eq. (28) is recovered. This leads to

$$P(W) = \lim_{n \to 0} V^{n-1} \int \prod_j dr(W_j)\ \delta(W - W_1) \prod_\mu S(P^\mu, \{W_j\}, \{G_j^\mu\}) \tag{29}$$

Thus, one computes the average over the p random associations of

$$\langle P(W) \rangle = \langle \int \prod dr(W_j^a) \delta(W - W_1^{a=1}) \prod_{\mu,a} S(P^\mu, \{W_j^a\}, \{G_j^\mu\}) \rangle \tag{30}$$

The computation then proceeds along the lines of the computation of the volume fraction. The new δ function term only affects the computation of Z_W. If the measure $dr(W)$ is flat on $[0, +\infty[$, we obtain, in the limit $n \to 0$,

$$\langle P(W) \rangle = \Theta(W) \int_{-\infty}^{+\infty} Du \frac{\exp\left[+(\hat{Q} - \frac{\hat{q}}{2}) W^2 + W(u\sqrt{\hat{q}} - \hat{M})\right]}{\int_0^{+\infty} dW' \exp\left[+(\hat{Q} - \frac{\hat{q}}{2}) W'^2 + W'(u\sqrt{\hat{q}} - \hat{M})\right]} \tag{31}$$

One can verify that the integral of $\langle P(W) \rangle$ over W is equal to one, as it should be: after permutation of the u and the w integrals, the fraction disappears and only the integral over Du remains.

For $\alpha = 0$ (the number of patterns is small compared to N), the distribution is exponential, $\langle P(W) \rangle = \exp(-W/\overline{W})$. This is the distribution of individual synaptic weights when subject to the single constraint that the average weight should be equal to θ/fN. When more and more patterns are added, the distribution 'stretches out' to accommodate these patterns. Figure 5D shows how the distribution changes as the capacity α is increased. As α increases, more and more weights accumulate at values close to zero.

A more explicit form for the weight distribution function can be obtained at the critical capacity by taking advantage of Eq. (13). When $q \to Q$, Eq. (31) simplifies to

$$P(W) = H(-B)\delta(W) + \frac{A}{\sqrt{2\pi C}} \exp\left[-\frac{A^2}{2C}\left(W + \frac{B\sqrt{C}}{A}\right)^2\right] \Theta(W) \quad (32)$$

The distribution of weights depends only on two parameters: B and a scale factor

$$W_s = \frac{\sqrt{C}}{A} = \frac{\overline{W}}{G(B) - BH(B)} \quad (33)$$

Using these parameters the distribution can be rewritten as

$$P(W) = H(-B)\delta(W) + \frac{1}{\sqrt{2\pi} W_s} \exp\left[-\frac{1}{2W_s^2}\left(W + BW_s\right)^2\right] \Theta(W) \quad (34)$$

The right-hand term of the r. h. s. can be identified as the positive portion of a normal distribution with mean $-BW_s$ and standard deviation W_s. The silent synapses are represented by the left-hand term of the r. h. s.: the Dirac delta generates a spike of zero-weight synapses. The fraction of silent synapses is $H(-B)$. Given the dependence of B on ρ (Eq. 25), this fraction increases from 50% when $K = 0$ to 100% when $K \to \infty$.

Distribution of compound EPSP amplitudes induced by input patterns

The compound EPSP amplitude induced by a particular input pattern $\{G_j\}$ is

$$v = \sum_j w_j G_j$$

In the large N limit, we compute the distribution of $V \equiv \sqrt{N}(v - \theta)$. The calculation is a generalisation of calculations by Krauth et al. (1988) and

Abbott and Kepler (1989). At critical capacity, the distribution is

$$\begin{aligned} P(V) &= \frac{f'}{\sqrt{2\pi f(1-f)Q}} \exp\left(-\frac{(V-fM)^2}{2Qf(1-f)}\right) \Theta(V-K) \\ &+ f' H\left(\frac{-V+fM}{\sqrt{f(1-f)Q}}\right) \delta(V-K) \\ &+ \frac{1-f'}{\sqrt{2\pi f(1-f)Q}} \exp\left(-\frac{(V-fM)^2}{2Qf(1-f)}\right) \Theta(-V-K) \\ &+ (1-f') H\left(\frac{V-fM}{\sqrt{f(1-f)Q}}\right) \delta(V+K) \end{aligned} \tag{35}$$

The distribution using the best-fit parameters (at critical capacity) is shown in Fig. 8 both without and with additional Gaussian noise of standard deviation σ.

Distribution of synaptic weights in a perceptron with non-linear summation of inputs

We take a model with a simple form of non-linearity. The individual granule cell inputs $w_j G_j$ are first summed linearly, and the total depolarisation is taken to be the a nonlinear function Φ of this sum. In this scenario, an individual connection makes an EPSP from rest of $\Phi(w_j)$, while a compound EPSP resulting from a given granule cell input is $\Phi(\sum_j w_j G_j)$. The calculation of the distribution of synaptic weights in this case follows the derivation in the linear case. The distribution of synaptic weights has the same functional form, except that one must apply the non-linear transformation $w \to \Phi(w)$. For a monotonically increasing function Φ for which $\Phi(0) = 0$, the two main conclusions of the linear model are unchanged: there is a majority of silent synapses, and the distribution is decreases monotonically with weight. What changes is that the functional form of the distribution at positive weights is no longer a Gaussian.

An alternative formulation: minimising the fraction of errors for a given noise level.

We assume here that a Gaussian noise with zero mean and standard deviation (SD) σ is added to the total Purkinje cell depolarisation each time an input pattern is presented. The fraction of errors performed by the system in

this situation is given by $(1/p)\sum_\mu H(\Delta_\mu/\sigma)$, where H is the complementary error function defined in Eq. (12), and

$$\Delta_\mu = (2P^\mu - 1)\left(\sum_j \frac{W_j}{\sqrt{N}} G_j^\mu - \theta\sqrt{N}\right)$$

In this section, we define a cost function that is proportional to this fraction of errors, i.e. $H(\Delta/\sigma)$. Calculation of the minimum of this cost function in the space of couplings is done along the lines of Gardner and Derrida (1988); Griniasty and Gutfreund (1991).

The calculation leads to equations similar to the previous Eqs. (9-11), but with $\mathcal{Z}_A$ given by

$$\begin{aligned}\mathcal{Z}_A &= \int_{-\infty}^{+\infty} Du \left(f' \ln \int dz \exp\left(-\beta F_+(z,t,Q,q,M)\right) \right. \\ &\quad \left. + (1-f') \ln \int dz \exp\left(-\beta F_-(z,t,Q,q,M)\right) \right). \end{aligned} \tag{36}$$

where β is an 'inverse temperature' parameter that will be sent to ∞, and $F_\pm$ is given by

$$\begin{aligned}F_\pm(z,t,Q,q,M) &= \frac{1}{2(Q-q)\beta}\left(z\sqrt{Q-q} \mp \frac{fM}{\sqrt{f(1-f)}} + \sqrt{Q}t\right)^2 \\ &\quad + H\left(\frac{z\sqrt{f(1-f)(Q-q)}}{\sigma}\right) \end{aligned} \tag{37}$$

We are interested in the limit $q \to Q$, $\beta \to \infty$, but with $x \equiv \beta(Q-q)/Q$ finite. x monitors the allowed fraction of errors that is necessarily nonzero in this case. F can be rewritten as a function of x, $\lambda = z\sqrt{Q/(Q-q)}$, $\tilde{M} = fM/\sqrt{f(1-f)Q}$ and $\tilde{\sigma} = \sigma/\sqrt{f(1-f)Q}$ (the latter parameter plays the role of ρ in the previous calculation)

$$F_\pm(\lambda,t,x,\tilde{M}) = \frac{1}{2x}\left(\lambda \mp \tilde{M} + t\right)^2 + H\left(\frac{\lambda}{\tilde{\sigma}}\right) \tag{38}$$

In the limit $\beta \to \infty$, integrals of the type $\int d\lambda \ldots \exp\left(-\beta F_\pm(\lambda,t,x,\tilde{M})\right)$ are dominated by the minimum of F over λ, defined as

$$Z_\pm(t,x,\tilde{M}) = \min_\lambda F_\pm(\lambda,t,x,\tilde{M}))$$

Eqs. (14-16) are replaced by

$$C = QA + \alpha_c Q \frac{x}{\tilde{\sigma}} \int Dt \left[f' Z_+(t,x,\tilde{M})) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z_+(t,x,\tilde{M})^2}{2\tilde{\sigma}^2} \right) + (1-f') Z_-(t,x,\tilde{M})) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{Z_-(t,x,\tilde{M})^2}{2\tilde{\sigma}^2} \right) \right] \quad (39)$$

$$A = \alpha_c \int Dt\, t \left[f'(Z_+(t,x,\tilde{M}) + t - \tilde{M}) + (1-f')(Z_-(t,x,\tilde{M}) + t + \tilde{M}) \right] \quad (40)$$

$$0 = \int Dt \left[f'(Z_+(t,x,\tilde{M}) + t - \tilde{M}) - (1-f')(Z_-(t,x,\tilde{M}) + t + \tilde{M}) \right], \quad (41)$$

while the three other saddle-point equations, Eqs. (18-20) are unchanged.

The fraction of errors is

$$E_0 = \int Dt \left[f' H\left(\frac{Z_+(t,x,\tilde{M})}{\tilde{\sigma}} \right) + (1-f') H\left(\frac{Z_-(t,x,\tilde{M})}{\tilde{\sigma}} \right) \right]$$

The strategy to solve these equations is then similar to the previous scenario. The distribution of synaptic weights is unchanged. One obtains all parameters of interest (fraction of silent synapses, capacity) as a function of the noise parameter σ and the fraction of errors E_0. The fraction of silent synapses is shown in Fig. 1. Note the similarity with Fig. 5B inset (fraction of silent synapses vs ρ).

Learning algorithm

For unbounded weights, the perceptron learning algorithm (Rosenblatt, 1962; Minsky and Papert, 1988) is guaranteed to find a synaptic vector that gives the desired response for all patterns, provided such a synaptic vector exists. This algorithm can be generalised to the situation of excitatory weights and a fixed threshold, and a convergence proof can be derived in this situation as well (not shown).

The initialisation is as follows:

- Choose a threshold θ;
- Generate p random associations (G_i^μ, P^μ);

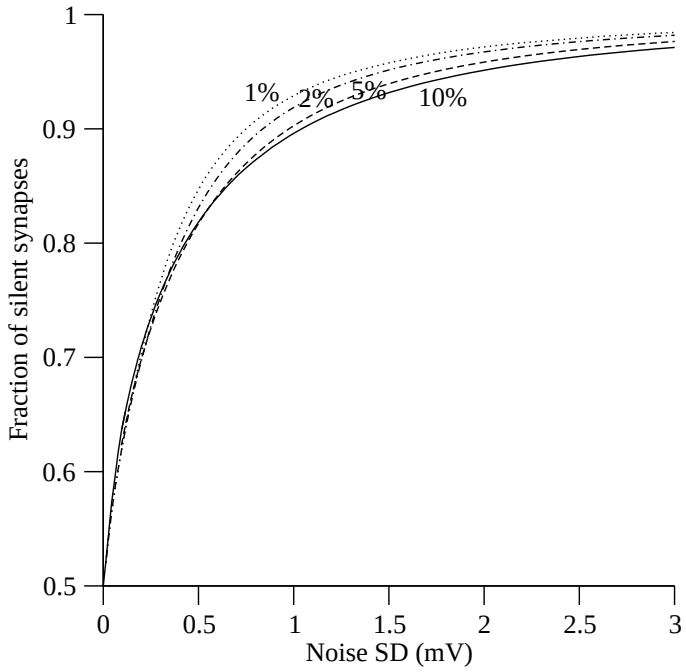

Figure 1: Fraction of silent synapses vs SD of noise, for several values of the fraction of errors, with f and f' given by their standard 'best-fit' values. Note that the fraction of silent synapses depends only weakly on the fraction of errors. The fraction of silent synapses is about 80% for noise of amplitude 0.5 mV, and about 90% for noise of amplitude 1 mV.

- Generate a random initial synaptic matrix. We choose a uniform distribution between 0 and $2\theta/(fN)$. In this way, the average synaptic value is initially close to its optimal value;

The modified perceptron algorithm then goes as follows:

1. Choose an association μ;
2. Compute $\Delta = \sum_i w_i G_i^\mu - \theta$.
3. If $(2P^\mu - 1)\Delta > \kappa$, the association is already learned with stability κ; go to 1. Else, modify the synaptic vector according to

$$w_i \to w_i + dw G_i^\mu (2P^\mu - 1), \quad i = 1, \ldots, N$$

If the resulting weight is negative, then set the weight to zero, $w_i = 0$. Go to 1.

Adapting the convergence proof of the standard perceptron algorithm (Rosenblatt, 1962; Minsky and Papert, 1988), one can show that, below critical capacity, the algorithm converges provided dw is smaller than some critical value $dw_c > 0$. The fact that dw has to be small enough come from learning with an *a priori* given threshold. To reach critical capacity, the simulation is done as follows: We start with a single pattern in the learning set, and $dw = 0.001\theta$. Patterns are added one by one to the learning set once the previous set has been learned. If the algorithm does not converge after a fixed number of iterations (1,000,000), dw is reduced by a factor 2. Learning stops once dw reaches $10^{-6}\theta$. In particular, we applied the algorithm on a perceptron with $N = 2000$, $f = 0.1$, $f' = 0.25$. For $\rho = 2.1$, the algorithm learns an average of 620 associations ($\alpha = 0.31$, 10 samples) compared to a predicted capacity of 0.33. The fraction of silent synapses is 78%, to be compared to the theoretical prediction of 80%. The distribution in this case is shown in Fig. 5A.

References

L. F. Abbott and T. B. Kepler. Universality in the space of interactions for network models. *J Phys A: Math Gen*, 2:2031–8, 1989.

A. Engel and C. van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

E. Gardner. The phase space of interactions in neural network models. *J. Phys. A: Math. Gen.*, 21:257–270, 1988.

E. Gardner and B. Derrida. Optimal storage properties of neural network models. *J Phys A: Math Gen*, 21:271–84, 1988.

M. Griniasty and H. Gutfreund. Learning and retrieval in attractor neural networks above saturation. *J Phys A: Math Gen*, 24:715–34, 1991.

H. Gutfreund and Y. Stein. Capacity of neural networks with discrete synaptic couplings. *J. Phys. A: Math. Gen.*, 23:2613–2630, 1990.

H. M. Kohler and D. Widmaier. Sign-constrained linear learning and diluting in neural networks. *J. Phys. A.*, 24:L495–L502, 1991.

W. Krauth, M. Mézard, and J.-P. Nadal. Basins of attraction in perceptron-like neural networks. *Complex Syst.*, 2:387, 1988.

M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and beyond*. World Scientific: Singapore, 1987.

M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry. Expanded Edition*. MIT Press, Cambridge, Ma, 1988.

F. Rosenblatt. *Principles of neurodynamics*. Spartan Books, New York, 1962.