Queremos demostrar que el método del perceptrón converge

libro Hertz

**Teorema de convergencia** (perceptrón) ($\Delta w =$

Consideremos $\bar{w}_j$ después de cierto nº de pasos:

$$w_j = \eta \sum_{\mu=1}^{P} M_\mu z_j^\mu$$

(P → # de patrones; $M_\mu$ → cantidad de veces que se modificó el patrón $\mu$)

(despreciamos $w_j$ inicial)

Queremos probar que $M = \sum_{\mu=1}^{P} M_\mu$ es finito (es decir que el algoritmo converge en un nº finito de pasos)

Queremos encontrar $\bar{w}^0$ / $\sum_{j=1} w_j^0 x_j^\mu S^\mu > NK$ óptimo:

ie aquel que maximiza el mínimo sobre $\mu$ de el pto más cercano al pto de seguridad

$$\min_\mu \left( \sum_{j=1}^{N} w_j^0 x_j^\mu S^\mu \right)$$

(quiero que el elemento más cerca a la línea de seguridad esté lejos)

$$\vec{w} \cdot \vec{w}^{opt} = \eta \sum_{\mu=1}^{P} n_\mu \vec{z}^\mu \vec{w}^{opt} \geq \eta M \min_\mu \left( \vec{z} \cdot \vec{w}^{opt} \right) = \eta M D(\vec{z}, \vec{w}_{opt}) |\vec{w}^{opt}|$$

(D → distancia)

$$D(\vec{z}, \vec{w}_{opt}) = \frac{\vec{z} \cdot \vec{w}_{opt}}{|\vec{w}_{opt}|}$$

Queremos ver como crecía $|\vec{w}^{opt}|$ en cada paso

(1) $\boxed{\vec{w} \cdot \vec{w}^{opt} \geq \eta M D |\vec{w}^{opt}|}$ ( $\Delta \vec{w}^{\alpha \text{ patrón}} = \eta \vec{z}^\alpha$ )

$\vec{w} \to \vec{w} + \Delta \vec{w}$ $\quad \Delta |\vec{w}|^2 = |\vec{w} + \eta \vec{z}^\alpha|^2 - |\vec{w}|^2$

$= \cancel{|\vec{w}|^2} + \eta^2 \underbrace{|\vec{z}^\alpha|^2}_{N} + 2\eta \underbrace{\vec{w} \cdot \vec{z}}_{\geq NK \text{ (si hay corrección)}} - \cancel{|\vec{w}|^2}$

$\leq \eta^2 N + 2\eta NK = \eta N [\eta + 2K]$

$\Rightarrow \boxed{|\vec{w}|^2 \leq \eta N M [\eta + 2K]}$ (2)

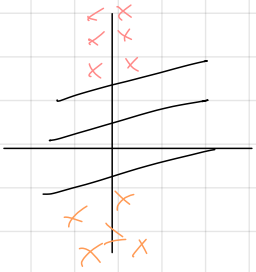
$\phi = \frac{(\vec{w} \cdot \vec{w}_{opt})^2}{|\vec{w}|^2 |w_{opt}|^2}$ $\quad (0 \leq \phi \leq 1)$

$\Rightarrow$ $1 \geq \phi \geq$ (de (1) y (2)) $\frac{\eta^2 M^2 D^2 (\vec{z}, \vec{w}_{opt}) |\vec{w}^{opt}|^2}{\eta N M [\eta + 2K] |w_{opt}|^2}$

$$\Rightarrow \quad M \leq \frac{\overset{\leq 1}{\phi}\, N\,[\eta + 2K]}{\eta\, D^2(\vec{z}, \vec{w}_{opt})} \leq \frac{N[\eta+2K]}{\eta\, D^2(\vec{z}, \vec{w}_{opt})}$$

$N$ + grande $\rightarrow$ $M$ mayor, lleva más tiempo converger

$D^2$ + grande $\rightarrow$



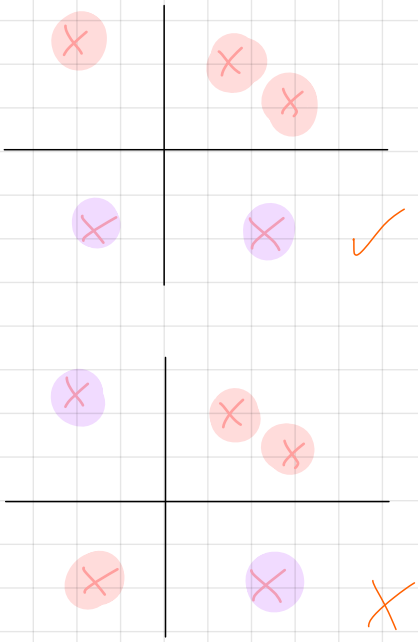
separación entre nubes de ptos mayor, se llega más tiempo?

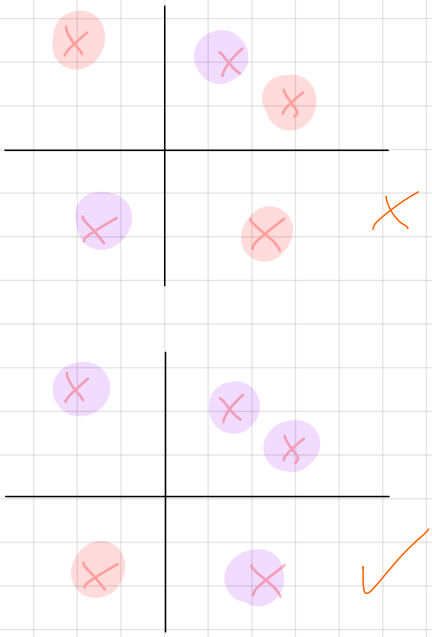
Si $\eta \gg 2K \rightarrow M \propto \frac{1}{\eta}$

---

¿ En qué condiciones el problema es linealmente separable ? (tener nubes de ptos)

Tenemos $P$ vectores $\vec{x}^{\mu}$ en $N$ dimensiones
(en el ejemplo $P=5$, $N=2$)





✓ ✗ son o no linealmente separables

● : 1

● : −1



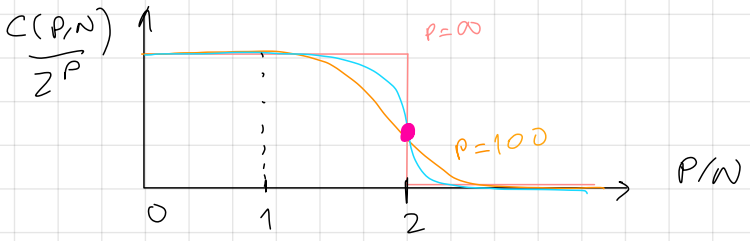


Manteniendo fijo los posiciones de los $x$, se pueden resolver varios problemas del tipo perceptron
Hay $2^P$ posibles problemas a resolver ($\mu = 1, \dots P$ puede tomar valor $+1$ o $-1$)
No todos los problemas son linealmente separables

Definimos $C(P,N)$ : # de problemas linealmente sep.

$\Rightarrow \frac{C(P,N)}{2^P}$ : fracción de problemas linealmente sep.

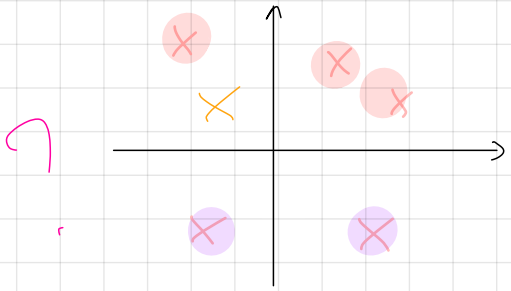
Esto se puede calcular. Se encuentra que




$P < 2N$ cualq. problema es L.S
$P > 2N$ ninguno lo es

Sup. que conozco $C(P,N) \to C(P+1,N)$ y además ↓ si conozco
conocemos $C(1,N) = 2$ podemos resolver el problema recursivamente

Encontremos entonces la relación entre $C(P,N)$



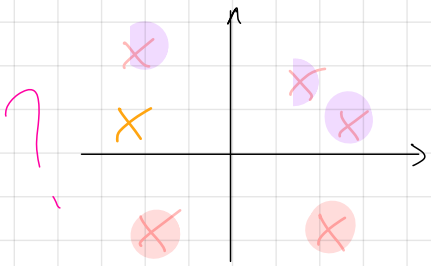
Supongamos que agregamos un punto (naranja)

↓

$C(P+1,N)$ es mayor o igual a $C(P,N)$ (siempre hay un coloreado que lo deja en el lado correcto

$$C(P+1,N) = C(P,N) + \dots$$

En este caso, el pto



Si estamos en 1 o 2 lo define si hay en el problema $C(P,N)$ un hiperplano de separación que pase por el pto nuevo

" tienen que estar en "posición gral"

Si hay n ptos en $\mathbb{R}^2$. Por 2 ptos siempre pasa una recta. $\Rightarrow$ los ptos están en posición general, si al unir dos ptos con la recta, no hay otro pto más que pase por ella

$P < N$

Se puede probar que $C(P,N) = \binom{P-1}{0} C(1,N) + \binom{P-1}{1} C(1,N-1)$

$$+ \cdots \binom{p-1}{p-1} C(1, N-p+1)$$

$$= 2 \sum_{j=0}^{p-1} \binom{p-1}{j} = 2\, 2^{p-1} = 2p$$

$C(1,n) = 0$ si $n \leq 0$

si $p = 2N$

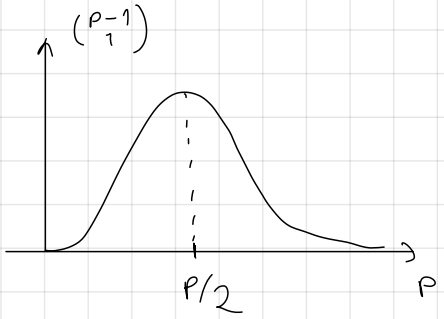
$$C(p,N) = \binom{p-1}{0} 2 + \binom{p-1}{1} 2 + \cdots \binom{p-1}{p/2}$$

$$= 2 \sum_{j=0}^{p/2} \binom{p-1}{j} = 2^{p-2}$$

↓ binomial simétrica

$$= 2^{p-1}$$

$$\Rightarrow \frac{C(p,N)}{2^p} = 0{,}5$$ → todas las curvas se cruzan en ese valor




$$C(p,N) = \sum_{j=0}^{p-1} \binom{p-1}{i}$$

The change in $\langle E \rangle$ in one cycle of weight updatings is thus

$$\begin{aligned}\Delta\langle E\rangle &= \sum_{ik} \frac{\partial\langle E\rangle}{\partial w_{ik}} \Delta w_{ik} \\ &= -\sum_{i\mu k} \Delta w_{ik} \zeta_i^\mu \frac{\partial}{\partial w_{ik}} \tanh(\beta h_i^\mu) \\ &= -\sum_{i\mu k} \eta[1 - \zeta_i^\mu \tanh(\beta h_i^\mu)]\beta \mathrm{sech}^2(\beta h_i^\mu) \end{aligned} \tag{5.62}$$

using[6] $d\tanh(x)/dx = \mathrm{sech}^2 x$. The result (5.62) is clearly always negative (recall $\tanh(x) < 1$), so the procedure always improves the average performance.

## 5.7 Capacity of the Simple Perceptron ⋆

In the case of the associative network in Chapter 2 we were able to find the **capacity** $p_{\max}$ of a network of $N$ units; for random patterns we found $p_{\max} = 0.138N$ for large $N$ if we used the standard Hebb rule. If we tried to store $p$ patterns with $p > p_{\max}$ the performance became terrible.

Similar questions can be asked for simple perceptrons:

- How many *random* input-output pairs can we expect to store reliably in a network of given size?
- How many of these can we expect to *learn* using a particular learning rule?

The answer to the second question may well be smaller than the first (e.g., for nonlinear units), but is presently unknown in general. The first question, which this section deals with, gives the maximum capacity that *any* learning algorithm can hope to achieve.

For continuous-valued units (linear or nonlinear) we already know the answer, because the condition is simply linear independence. If we choose $p$ *random* patterns, then they will be linearly independent if $p \leq N$ (except for cases with very small probability). So the capacity is $p_{\max} = N$.

The case of threshold units depends on linear separability, which is harder to deal with. The answer for random continuous-valued inputs was derived by Cover [1965] (see also Mitchison and Durbin [1989]) and is remarkably simple:

$$p_{\max} = 2N. \tag{5.63}$$

As usual $N$ is the number of *input* units, and is presumed large. The number of *output* units must be small and fixed (independent of $N$). Equation (5.63) is strictly true in the $N \to \infty$ limit.

[6] The function $\mathrm{sech}^2 x = 1 - \tanh^2 x$ is a bell-shaped curve with peak at $x = 0$.

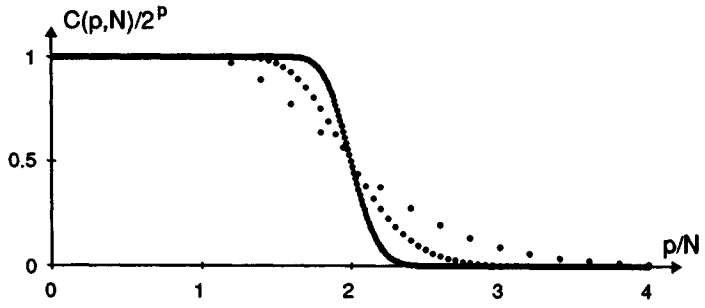



FIGURE 5.11 The function $C(p, N)/2^p$ given by (5.67) plotted versus $p/N$ for $N = 5$, 20, and 100.

The rest of this section is concerned with proving (5.63), and may be omitted on first reading. We follow the approach of Cover [1965]. A more general (but much more difficult) method for answering this sort of question was given by Gardner [1987] and is discussed in Chapter 10.

We consider a perceptron with $N$ continuous-valued inputs and one $\pm 1$ output unit, using the deterministic threshold limit. The extension to several output units is trivial since output units and their connections are independent—the result (5.63) applies separately to each. For convenience we take the thresholds to be zero, but they could be reinserted at the expense of one extra input unit, as in (5.2).

In (5.11) we showed that the perceptron divides the $N$-dimensional input space into two regions separated by an $(N-1)$-dimensional hyperplane. For the case of zero threshold this plane goes through the origin. All the points on one side give an output of $+1$ and all those on the other side give $-1$. Let us think of these as red $(+1)$ and black $(-1)$ points respectively. Then the question we need to answer is: how many points can we expect to put randomly in an $N$-dimensional space, some red and some black, and then find a hyperplane through the origin that divides the red points from the black points?

Let us consider a slightly different question. For a given set of $p$ randomly placed points in an $N$-dimensional space, for how many out of the $2^p$ possible red and black colorings of the points can we find a hyperplane dividing red from black? Call the answer $C(p, N)$. For $p$ small we expect $C(p, N) = 2^p$, because we should be able to find a suitable hyperplane for *any* possible coloring; consider $N = p = 2$ for example. For $p$ large we expect $C(p, N)$ to drop well below $2^p$, so an arbitrarily chosen coloring will *not* possess a dividing hyperplane. The transition between these regimes turns out to be sharp for large $N$, and gives us $p_{\max}$.

We will calculate $C(p, N)$ shortly, but let us first examine the result. Figure 5.11 shows a graph of $C(p, N)/2^p$ against $p/N$ for $N = 5$, 20, and 100. Our expectations for small and large $p$ are fulfilled, and we see that the transition occurs quite rapidly in the neighborhood of $p = 2N$, in agreement with (5.63). As $N$ is made larger and

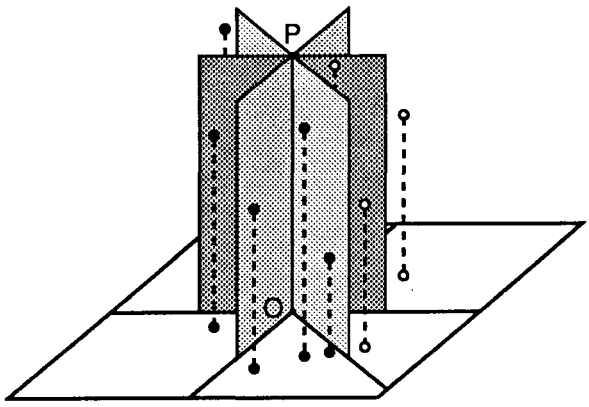



FIGURE 5.12 Finding separating hyperplanes constrained to go through a point P as well as the origin O is equivalent to projecting onto one lower dimension.

larger the transition becomes more and more sharp. Thus (5.63) is justified if we can demonstrate that Fig. 5.11 is correct.

The random placement of points is not actually necessary.[7] All that we need is that the points be in **general position**. As discussed on page 97, this means (for the no threshold case) that all subsets of $N$ (or fewer) points must be linearly independent. As an example consider $N = 2$: a set of $p$ points in a two-dimensional plane is in general position if no two lie on the same line through the origin. A set of points chosen from a continuous random distribution will obviously be in general position except for coincidences that have zero probability.

We can now calculate $C(p, N)$ by induction. Let us call a coloring that *can* be divided by a hyperplane a **dichotomy**. Suppose we start with $p$ points and add a new point $P$. Then the $C(p, N)$ old dichotomies fall into two classes:

- For those previous dichotomies where the dividing hyperplane could have been drawn *through* point $P$, there'll be *two* new dichotomies, one with $P$ red and one with it black. This is because when the points are in general position any hyperplane through $P$ can be shifted infinitesimally to go either side of it, without changing the side of any of the other $p$ points.
- For the remainder of the previous dichotomies only one color of point $P$ will fit, so there'll be *one* new dichotomy for each old one.

Thus

$$C(p+1, N) = C(p, N) + D \tag{5.64}$$

where $D$ is the number of the previous $C(p, N)$ dichotomies that could have had the dividing hyperplane drawn through $P$ as well as the origin $O$. But this number is simply $C(p, N-1)$, because constraining the hyperplanes to go through a particular point $P$ makes the problem effectively $(N-1)$-dimensional; as illustrated in Fig. 5.12, we can project the whole problem onto an $(N-1)$-dimensional plane

[7]Nor is it well defined unless a distribution function is specified.

perpendicular to $OP$, since any displacement of a point along the $OP$ direction cannot affect which side it is of any hyperplane containing $OP$.

We thereby obtain the **recursion relation**

$$C(p+1,N) = C(p,N) + C(p,N-1). \tag{5.65}$$

Iterating this equation for $p$, $p-1$, $p-2$, ..., 1 yields

$$C(p,N) = \binom{p-1}{0}C(1,N) + \binom{p-1}{1}C(1,N-1) + \ldots + \binom{p-1}{p-1}C(1,N-p+1). \tag{5.66}$$

For $p \leq N$ this is easy to handle, because $C(1,N) = 2$ for all $N$; one point can be colored red or black. For $p > N$ the second argument of $C$ becomes 0 or negative in some terms, but these terms can be eliminated by taking $C(p,N) = 0$ for $N \leq 0$. It is easy to check that this choice is consistent with the recursion relation (5.65), and with $C(p,1) = 2$ (in one dimension the only "hyperplane" is a point at the origin, allowing two dichotomies). Thus (5.66) makes sense for all values of $p$ and $N$ and can be written as

$$C(p,N) = 2\sum_{i=0}^{N-1}\binom{p-1}{i} \tag{5.67}$$

if we use the standard convention that $\binom{n}{m} = 0$ for $m > n$. Equation (5.67) was used to plot Fig. 5.11, thus completing the demonstration.

It is actually easy to show from the symmetry $\binom{2n}{n-m} = \binom{2n}{n+m}$ of binomial coefficients that

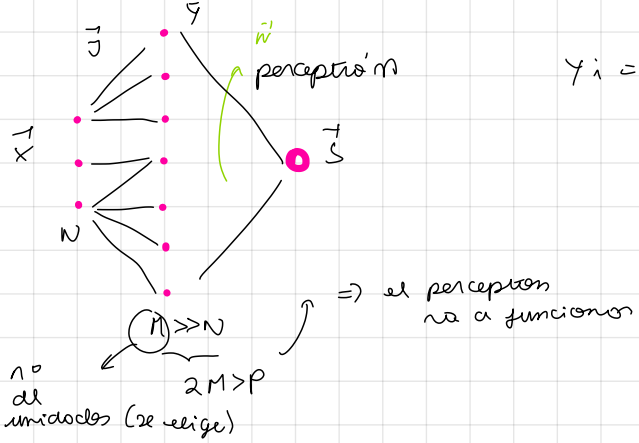
$$C(2N,N) = 2^{p-1} \tag{5.68}$$

so the curve goes through 1/2 at $p = 2N$. To show analytically that the transition sharpens up for increasing $N$, one can appeal to the large $N$ Gaussian limit of the binomial coefficients, which leads to

$$C(p,N)/2^p \approx \frac{1}{2}\left[1 + \operatorname{erf}\left(\sqrt{\frac{p}{2}}\left(\frac{2N}{p} - 1\right)\right)\right] \tag{5.69}$$

for large $N$.

It is worth noting that $C(p,N) = 2^p$ if $p \leq N$ (this is shown on page 155). So *any* coloring of up to $N$ points is linearly separable, provided only that the points are in general position. For $N$ or fewer points general position is equivalent to linear independence, so the sufficient conditions for a solution are exactly the same in the threshold and continuous-valued networks. But this is not true, of course, for $p > N$.




$$y_i = g\left(\sum_i J_{ij} x_j\right)$$

⇒ el perceptron no a funcionado

los pts fueron mapeados de manera tal que el problema es SI

↳ base de una familia de técnicos --- machine