Artificial Neural Networks

Lecture 2: Hebbian, Supervised Learning in Perceptrons

Dr Juan Neirotti

j.p.neirotti@aston.ac.uk

2023

Supervised Learning in Perceptrons

- To understand the principles behind Hebb's update rule for supervised learning.
- To understand and apply the principles of modeling biological system and its statistical treatment.
- To understand the basic mechanisms of optimization.

Supervised Hebbian Learning in Perceptrons

- Problem: Classification of binary vectors $\boldsymbol{x}_\ell \in \{-1,+1\}^N = \mathscr{X}$ with large N according to $t_\ell \equiv \operatorname{sgn}(\langle \boldsymbol{B} \,|\boldsymbol{x}_\ell \rangle)$ for a given $\boldsymbol{B} \in \{-1,+1\}^N$, with $\|\boldsymbol{B}\|_2 := \sqrt{\langle \boldsymbol{B} \,|\boldsymbol{B} \rangle} = \sqrt{N}$. $|\boldsymbol{B}\rangle$ is the *supervisor* or *teacher* that indicates whether an example is correctly classified or not.
- Settings: $\mathcal{D} = \{(\boldsymbol{x}_\ell, t_\ell)\}_{\ell=1}^{p}$. The loss functions to be considered are:

$$\varepsilon_T(|\boldsymbol{w}\rangle) := \frac{1}{p} \sum_{\ell=1}^{p} \Theta\left(-t_\ell \langle \boldsymbol{w} \,|\boldsymbol{x}_\ell \rangle\right)$$

Hebb's Algorithm

- Hebb's algorithm is based on Pavlov's coincidental training applied at the level of a single neuron.
- Suppose $|\boldsymbol{x}\rangle = |x_1, x_2, \ldots, x_N\rangle$, where x_i is the input to channel w_i of the neuron. If the required output $t = \operatorname{sgn}\left(\langle \boldsymbol{B} | \boldsymbol{x} \rangle\right)$ has the same sign as x_i, the i-th connection has to be strengthen: $w_{i,\mathrm{new}} = w_{i,\mathrm{old}} + 1$, if they have different sign, the connection is weakened: $w_{i,\mathrm{new}} = w_{i,\mathrm{old}} - 1$.
- Hebb's algorithm can be written as:

$$|\boldsymbol{w}_{\ell+1}\rangle = |\boldsymbol{w}_{\ell}\rangle + \frac{t_\ell |\boldsymbol{x}_\ell\rangle}{\sqrt{N}}$$

$$|\boldsymbol{w}_p\rangle = \frac{1}{\sqrt{N}} \sum_{\ell=1}^{p} t_\ell |\boldsymbol{x}_\ell\rangle \tag{1}$$

where the $1/\sqrt{N}$ is an appropriate normalization factor.

Aligning Field

- We define the stability, or aligning field, as

$$\phi_\ell := \frac{t_\ell \langle \boldsymbol{w} | \boldsymbol{x}_\ell \rangle}{\sqrt{N}}. \tag{2}$$

- A correct classification of $|\boldsymbol{x}_\ell\rangle$ by $|\boldsymbol{w}\rangle$ implies that $\phi_\ell > 0$. We would like to evaluate the probability of this event.
- Observe that by using (1) in (2) we have that for a given $1 \leq m \leq p$:

$$\phi_m = 1 + \frac{1}{N} \sum_{\ell \neq m} t_m t_\ell \langle \boldsymbol{x}_m | \boldsymbol{x}_\ell \rangle .$$

Aligning Field

- Observe that, for all $\ell = 1, \ldots, p$ we can write $|\boldsymbol{x}_\ell\rangle = |\boldsymbol{x}_{\ell\perp}\rangle + |\boldsymbol{x}_{\ell\|}\rangle$, where $|\boldsymbol{x}_{\perp(\|)}\rangle$ is the component of the vector $|\boldsymbol{x}\rangle$ perpendicular (parallel) to the vector $|\boldsymbol{B}\rangle = \|\boldsymbol{B}\|_2 |\boldsymbol{b}\rangle$:

$$|\boldsymbol{x}_{\|}\rangle := \langle \boldsymbol{b} | \boldsymbol{x} \rangle |\boldsymbol{b}\rangle \qquad |\boldsymbol{x}_{\perp}\rangle := |\boldsymbol{x}\rangle - |\boldsymbol{x}_{\|}\rangle$$

- Then

$$t_m t_\ell \langle \boldsymbol{x}_m | \boldsymbol{x}_\ell \rangle = t_m t_\ell \langle \boldsymbol{x}_{m\perp} | \boldsymbol{x}_{\ell\perp} \rangle + t_m t_\ell \langle \boldsymbol{x}_{m\|} | \boldsymbol{x}_{\ell\|} \rangle \tag{3}$$

Statistical properties of the variables

- The following two variables have important statistical properties that need to be expressed before proceeding:

$$x_{\parallel} := \langle \boldsymbol{b} \,|\boldsymbol{x}\, \rangle \tag{4}$$

$$x_{\perp n} := x_n - x_{\parallel} \frac{B_n}{\sqrt{N}} \tag{5}$$

  which are the parallel and n-th perpendicular components of $\boldsymbol{x}$.
- Consider the following identity for Dirac's delta distribution:

$$\delta(x-y) = \int \frac{\mathrm{d}z}{2\pi}\, \mathrm{e}^{-iz(x-y)}. \tag{6}$$

- The probability distribution of $x_{\parallel}$ can be expressed as:

$$\mathcal{P}_{X_{\parallel}}(x_{\parallel}) = \sum_{\boldsymbol{x}\in\mathscr{X}} \mathcal{P}_{X_{\parallel},\boldsymbol{X}}(x_{\parallel},\boldsymbol{x}) = \sum_{\boldsymbol{x}\in\mathscr{X}} \mathcal{P}_{X_{\parallel}|\boldsymbol{X}}(x_{\parallel}|\boldsymbol{x})\,\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x})$$

.

- By using that $\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{n=1}^N \mathcal{P}_X(x_n) = \prod_{n=1}^N \frac{1}{2}\left(\delta_{x_n,1} + \delta_{x_n,-1}\right)$, where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise, we have that $\sum_{\boldsymbol{x}\in\mathscr{X}} = \prod_{n=1}^N \sum_{x_n=\pm 1}$ and:

$$\begin{aligned}
\mathcal{P}_{X_\parallel}(x_\parallel) &= \prod_{n=1}^N \sum_{x_n=\pm1} \frac{1}{2}\left(\delta_{x_n,1} + \delta_{x_n,-1}\right) \delta\left(x_\parallel - \frac{\sum_{n=1}^N x_n B_n}{\sqrt{N}}\right) \\
&= \int \frac{\mathrm{d}z}{2\pi} \mathrm{e}^{-izx_\parallel} \prod_{n=1}^N \sum_{x_n=\pm1} \frac{1}{2}\left(\delta_{x_n,1} + \delta_{x_n,-1}\right) \exp\left(iz\frac{x_n B_n}{\sqrt{N}}\right) \\
&= \int \frac{\mathrm{d}z}{2\pi} \mathrm{e}^{-izx_\parallel} \prod_{n=1}^N \cos\left(z\frac{B_n}{\sqrt{N}}\right) \qquad (\text{observe that } B_n = \pm 1) \\
&= \int \frac{\mathrm{d}z}{2\pi} \mathrm{e}^{-izx_\parallel} \left[\cos\left(\frac{z}{\sqrt{N}}\right)\right]^N \\
&= \int \frac{\mathrm{d}z}{2\pi} \exp\left(-\frac{z^2}{2} - izx_\parallel\right) + O(N^{-\frac{1}{2}}) = \frac{\mathrm{e}^{-x_\parallel^2/2}}{\sqrt{2\pi}} + O(N^{-\frac{1}{2}})
\end{aligned} \tag{7}$$

- Therefore, $x_{\|}$ is a Gaussian distributed variable with zero mean and unit variance, i.e. $\mathcal{P}_{X_{\|}}(x) = \mathcal{N}(x|0,1)$.
- For $X_{\perp n} = s$ we have that:

$$\mathcal{P}_{X_{\perp n}}(s) = \int \mathrm{d}r \sum_{x=\pm 1} \mathcal{P}_{X_{\perp n}|X,X_{\|}}(s|x,r)\mathcal{P}_X(x)\mathcal{P}_{X_{\|}}(r)$$

where

$$\mathcal{P}_{X_{\perp n}|X,X_{\|}}(s|x,r) = \delta\left(s - \left(x - N^{-1/2}B_n r\right)\right)$$
$$\mathcal{P}_X(x) = \frac{1}{2}\left\{\delta_{x,1} + \delta_{x,-1}\right\}$$
$$\mathcal{P}_{X_{\|}}(r) = \mathcal{N}(r|0,1)$$

therefore

$$\mathcal{P}_{X_{\perp n}}(s) = \mathcal{P}_{X_{\perp}}(s) = \int \frac{\mathrm{d}z}{2\pi} \exp\left(-\frac{z^2}{2N} - izs\right)\cos z.$$

.

- The probability $\mathcal{P}_{X_\perp}(s)$ becomes:

$$\mathcal{P}_{X_\perp}(s) = \sqrt{\frac{N}{2\pi}} \left\{ \frac{1}{2} \exp\left[-\frac{N}{2}(s-1)^2 \right] + \frac{1}{2} \exp\left[-\frac{N}{2}(s+1)^2 \right] \right\} \tag{8}$$

- From (8) we obtain that

$$\mathbb{E}_{X_\perp} = 0 \qquad \mathbb{V}_{X_\perp} = 1 + \frac{1}{N}$$

thus

$$\mathcal{P}_{X_\perp}(s) = \frac{1}{2} \{\delta_{s,1} + \delta_{s,-1}\} + O(N^{-1}).$$

Second Term

- The second term of (3) can be expressed as

$$t_m t_\ell \left\langle \boldsymbol{x}_{m\|} \middle| \boldsymbol{x}_{\ell\|} \right\rangle = \left\| \boldsymbol{x}_{m\|} \right\|_2 \left\| \boldsymbol{x}_{\ell\|} \right\|_2 = |x_{\ell\|}||x_{\|m}|.$$

- Observe that, from (7), $x_\|$ is a Gaussian deviate with zero mean and unit variance. Observe also that $x_{\|\ell}$ and $x_{\|m}$ are independent if $\ell \neq m$. Thus, by applying the law of large numbers we have that (remember $p = \alpha N$):

$$\frac{1}{N} \sum_{m \neq \ell} t_m t_\ell \left\langle \boldsymbol{x}_{m\|} \middle| \boldsymbol{x}_{\ell\|} \right\rangle = \frac{\alpha N - 1}{N} |x_{\ell\|}| \frac{1}{\alpha N - 1} \sum_{m \neq \ell} |x_{m\|}|$$

$$= \alpha |x_{\ell\|}| \int \mathrm{d}z \mathcal{N}(z|0,1)|z| + O(N^{-1})$$

$$= \sqrt{\frac{2}{\pi}} \alpha |x_{\ell\|}| + O(N^{-1})$$

First Term

- The contributions from the first term of (3) add up to:

$$\frac{1}{N}\sum_{m\neq\ell} t_m t_\ell \langle \boldsymbol{x}_{m\perp} | \boldsymbol{x}_{\ell\perp} \rangle \approx \sqrt{\frac{\alpha}{N(p-1)}} \sum_{m\neq\ell}\sum_{n=1}^{N} t_\ell x_{\ell\perp n} t_m x_{m\perp n} \qquad (9)$$

- The right hand side (RHS) of (9) is the sum of $N(p-1)$ terms that are independent and identically distributed (iid). Each term has a zero average and almost unit variance. By applying the Central Limit Theorem we have that:

$$\frac{1}{N}\sum_{m\neq\ell} t_m t_\ell \langle \boldsymbol{x}_{m\perp} | \boldsymbol{x}_{\ell\perp} \rangle = \sqrt{\alpha} y + O(N^{-1}),$$

where y is a Gaussian variable with zero mean and unit variance.

First Term

- We have then:

$$\lim_{N\to\infty} \phi_m = 1 + \sqrt{\frac{2}{\pi}}\alpha|z| + \sqrt{\alpha}y,$$

where $\mathcal{P}_Z(z) = \mathcal{N}(z|0,1)$ and $\mathcal{P}_Y(y) = \mathcal{N}(y|0,1)$,with $\mathcal{P}_{Z,Y}(z,y) = \mathcal{P}_Z(z)\mathcal{P}_Y(y)$.

Training error

- The training error is hence given by the sum of the contributions that make $\phi_m < 0$ (i.e. $y < -1/\sqrt{\alpha} - \sqrt{2\alpha/\pi}|z|$):

$$
\begin{aligned}
\varepsilon_T(|\boldsymbol{w}\rangle) &= \frac{1}{p}\sum_{m=1}^{p}\Theta(-\phi_m) \\
&= \int_{\mathbb{R}} \mathrm{d}z \mathcal{P}_Z(z) \int_{\mathbb{R}} \mathrm{d}y \mathcal{P}_Y(y) \int_{\mathbb{R}} \mathrm{d}\phi \mathcal{P}_{\Phi|Z,Y}(\phi|z,y)\Theta(-\phi) \\
&= \int_{\mathbb{R}} \mathrm{d}z \mathcal{N}(z|0,1) \int_{\mathbb{R}} \mathrm{d}y \mathcal{N}(y|0,1)\Theta\left(-1 - \sqrt{\frac{2}{\pi}}\alpha|z| - \sqrt{\alpha}y\right) \\
&= 2\int_0^\infty \mathrm{d}z \mathcal{N}(z|0,1) \int_{-\infty}^{-\frac{1}{\sqrt{\alpha}} - \sqrt{\frac{2\alpha}{\pi}}z} \mathrm{d}y \mathcal{N}(y|0,1) \\
&= 2\int_0^\infty \mathrm{d}z \mathcal{N}(z|0,1)\mathcal{H}\left(\frac{1}{\sqrt{\alpha}} + \sqrt{\frac{2\alpha}{\pi}}z\right)
\end{aligned}
$$

where $\mathcal{H}(x) \equiv \int_x^\infty \mathrm{d}y\, \mathcal{N}(y,|0,1)$.

Training error

- Limits:

$$\varepsilon_T(\alpha) \overset{\alpha\to 0}{\approx} \frac{\pi-2}{\pi}\sqrt{\frac{2\alpha}{\pi}}\,\exp\left(-\frac{1}{2\alpha}\right)$$

$$\varepsilon_T(\alpha) \overset{\alpha\to\infty}{\approx} \frac{1}{\sqrt{2\pi\alpha}}.$$

Generalization error

- The generalization error is the probability of misclassifying a new example, given the training set. Thus:

$$\varepsilon_G(|\boldsymbol{w}\rangle) = \int \mathrm{d}\boldsymbol{\xi}\mathcal{P}_{\boldsymbol{\Xi}}(\boldsymbol{\xi})\Theta\left(-\frac{\langle \boldsymbol{B}\,|\boldsymbol{\xi}\rangle\,\langle \boldsymbol{w}\,|\boldsymbol{\xi}\rangle}{N}\right)$$
$$= \int \mathrm{d}\boldsymbol{\xi}\mathrm{d}\phi\mathrm{d}\beta\,\mathcal{P}_{\boldsymbol{\Phi},B|\boldsymbol{\Xi}}(\phi,\beta|\boldsymbol{\xi})\mathcal{P}_{\boldsymbol{\Xi}}(\boldsymbol{\xi})\Theta(-\beta\phi)$$
$$= \int \mathrm{d}\phi\mathrm{d}\beta\,\mathcal{P}_{\boldsymbol{\Phi},B}(\phi,\beta)\Theta(-\beta\phi)$$

where

$$\phi := \frac{\langle \boldsymbol{w}\,|\boldsymbol{\xi}\rangle}{\sqrt{N}} \qquad \beta := \frac{\langle \boldsymbol{B}\,|\boldsymbol{\xi}\rangle}{\sqrt{N}}$$

and $\mathcal{P}_{\boldsymbol{\Phi},B}(\phi,\beta)$ must be inferred from the properties of the pattern $\boldsymbol{\xi}$.

- The joint distribution can be obtained by:

$$
\begin{aligned}
\mathcal{P}_{\Phi,B}(\phi,\beta) &= \int \mathrm{d}\boldsymbol{\xi} \mathcal{P}_{\Phi,B|\boldsymbol{\Xi}}(\phi,\beta|\boldsymbol{\xi})\mathcal{P}_{\boldsymbol{\Xi}}(\boldsymbol{\xi}) \\
&= \int \mathrm{d}\boldsymbol{\xi} \mathcal{P}_{\boldsymbol{\Xi}}(\boldsymbol{\xi})\, \delta\left(\beta - \frac{\langle \boldsymbol{B}\,|\boldsymbol{\xi}\,\rangle}{\sqrt{N}}\right)\, \delta\left(\phi - \frac{\langle \boldsymbol{w}\,|\boldsymbol{\xi}\,\rangle}{\sqrt{N}}\right) \\
&= \int \frac{\mathrm{d}\hat{\phi}\,\mathrm{d}\hat{\beta}}{4\pi^2} \mathrm{e}^{-i\hat{\beta}\beta - i\hat{\phi}\phi} \times \\
&\times \prod_{n=1}^{N} \sum_{\xi=\pm 1} \mathcal{P}_{\Xi}(\xi) \exp\left[\left(i\hat{\beta}B_n + i\hat{\phi}w_n\right)\frac{\xi}{\sqrt{N}}\right] \\
&= \int \frac{\mathrm{d}\hat{\phi}\,\mathrm{d}\hat{\beta}}{4\pi^2} \mathrm{e}^{-i\hat{\beta}\beta - i\hat{\phi}\phi} \prod_{n=1}^{N} \cos\left(\frac{\hat{\beta}B_n + \hat{\phi}w_n}{\sqrt{N}}\right) \\
&\approx \int \frac{\mathrm{d}\hat{\phi}\,\mathrm{d}\hat{\beta}}{4\pi^2} \mathrm{e}^{-i\hat{\beta}\beta - i\hat{\phi}\phi} \exp\left\{-\frac{1}{2N}\sum_{n=1}^{N}(\hat{\beta}B_n + \hat{\phi}w_n)^2\right\}
\end{aligned}
$$

- The argument of the exponential can be worked up as:

$$\frac{1}{N}\sum_{n=1}^{N}(\hat{\beta}B_n + \hat{\phi}w_n)^2 = \hat{\beta}^2 + 2\hat{\beta}\,\hat{\phi}\frac{\langle \boldsymbol{w}\,|\boldsymbol{B}\,\rangle}{N} + \hat{\phi}^2\frac{\langle \boldsymbol{w}\,|\boldsymbol{w}\,\rangle}{N}, \quad (10)$$

where $\langle \boldsymbol{B}\,|\boldsymbol{B}\,\rangle = \|\boldsymbol{B}\|_2^2 = N$.

- The inner product between supervisor and student is:

$$\begin{aligned}\frac{\langle \boldsymbol{w}\,|\boldsymbol{B}\,\rangle}{N} &= \frac{1}{N}\sum_{\ell=1}^{p} t_\ell \frac{\langle \boldsymbol{B}\,|\boldsymbol{x}_\ell\,\rangle}{\sqrt{N}} \\ &= \frac{\alpha}{p}\sum_{\ell=1}^{p}|x_{\ell\|}| \\ &\approx \sqrt{\frac{2}{\pi}}\alpha \qquad (11)\end{aligned}$$

.

- The length of the student's vector is:

$$\frac{\langle \boldsymbol{w} \,|\, \boldsymbol{w} \rangle}{N} = \left(\frac{\alpha}{p} \sum_{\ell=1}^{p} |x_{\ell \|}| \right)^2 + \frac{1}{N} \sum_{n=1}^{N} \left(\sqrt{\frac{\alpha}{p}} \sum_{\ell=1}^{p} t_\ell x_{\ell \perp n} \right)^2 . \tag{12}$$

- The first term at the RHS of (12) is the square of the integration of the absolute value of a Gaussian variable with zero mean and unit variance, thus:

$$\left(\frac{\alpha}{p} \sum_{\ell=1}^{p} |x_{\ell \|}| \right)^2 \approx \frac{2\alpha^2}{\pi}.$$

- The second term at the RHS of (12) is the integration of the square of the sum of p iid variables (with zero mean and unit variance) divided by the square root of p,which is in itself a Gaussian variable with zero mean and unit variance. Thus:

$$\frac{1}{N}\sum_{n=1}^{N}\left(\sqrt{\frac{\alpha}{p}}\sum_{\ell=1}^{p} t_\ell x_{\ell\perp n}\right)^2 \approx \frac{\alpha}{N}\sum_{n=1}^{N} s_n^2 \approx \alpha$$

  where $s_n \equiv p^{-1/2}\sum_{\ell=1}^{p} t_\ell x_{\ell\perp n}$ are iid Gaussian variables with zero mean and unit variance.

- Thus

$$\frac{\langle \boldsymbol{w} | \boldsymbol{w} \rangle}{N} \approx \alpha\left(1+\frac{2\alpha}{\pi}\right) \tag{13}$$

- From (11) and (13) we have that (10) is:

$$\frac{1}{N}\sum_{n=1}^{N}(\hat{\beta}B_n + \hat{\phi}w_n)^2 \approx \hat{\beta}^2 + 2\sqrt{\frac{2}{\pi}}\alpha\,\hat{\beta}\,\hat{\phi} + \alpha\left(1 + \frac{2\alpha}{\pi}\right)\hat{\phi}^2.$$

- Thus

$$\begin{aligned}
\mathcal{P}_{\Phi,B}(\phi,\beta) &\approx \int \frac{\mathrm{d}\hat{\phi}\,\mathrm{d}\hat{\beta}}{4\pi^2}\times \\
&\quad\times \exp\left\{-i\hat{\beta}\beta - i\hat{\phi}\phi - \frac{\hat{\beta}^2}{2} - \sqrt{\frac{2}{\pi}}\alpha\,\hat{\beta}\,\hat{\phi} - \alpha\left(1+\frac{2\alpha}{\pi}\right)\frac{\hat{\phi}^2}{2}\right\} \\
&= \mathcal{N}(\beta|0,1)\,\mathcal{N}\left(\phi\,\middle|\,\sqrt{\frac{2}{\pi}}\alpha\,\beta, \alpha\right).
\end{aligned}$$

.

- The generalization error becomes:

$$\begin{aligned}
\varepsilon_G(\alpha) &= 2\int_0^\infty \mathrm{d}\beta\,\mathcal{N}(\beta|0,1)\int_{\sqrt{\frac{2\alpha}{\pi}}\beta}^\infty \mathrm{d}\phi\,\mathcal{N}(\phi|0,1) \\
&= 2\int_0^\infty \mathrm{d}\beta\,\mathcal{N}(\beta|0,1)\mathcal{H}\left(\sqrt{\frac{2\alpha}{\pi}}\beta\right) \\
&= \frac{1}{\pi}\arccos\left(\sqrt{\frac{2\alpha}{2\alpha+\pi}}\right).
\end{aligned}$$

- Limits:

$$\begin{aligned}
\varepsilon_G(\alpha) &\overset{\alpha\to 0}{\approx} \frac{1}{2} - \frac{\sqrt{2\alpha}}{\pi^{3/2}} \\
\varepsilon_G(\alpha) &\overset{\alpha\to\infty}{\approx} \frac{1}{\sqrt{2\pi\alpha}}.
\end{aligned}$$