Neural Nets Inference and Content Addressable Memory

Christian Mazza, Member IEEE
Section de Mathématiques, 2-4 Rue du Lièvre, Case Postale 240, CH-1211 Genève 24, Switzerland
This work was supported by the Swiss National Science Foundation
May 1996

***Abstract*—We** consider Whittle's probabilistic content addressable memory, the antiphon, which is designed for recovering stored patterns from nonlinear distorsions of the input messages. We give an application to content-based image retrieval systems, and propose canonical ways of choosing similarity thresholds ensuring statistical consistency.

I. Introduction

Let $C_{-}^{n} = \{-1,1\}^n$ be the hypercube of dimension $n \in \mathbb{N}$, from which we extract a set of words $S = \{a^1,..,a^m\}$, $m \in \mathbb{N}$. A device or memory which is able to recall a stored word of the above list without reference to its physical location is called content addressable, and the retrieval process is based only on semantic content. One well-known neural network of this type is the Hopfield network, with Hebbian learning. Let $A := [a^1,..,a^m] \in \{-1,1\}^{nxm}$. Given a distorded version $\eta \in C_{-}^{n}$ of one of the a^r, $r = 1,..,m$, the retrieval process is essentially governed by iterated rules of the form $\eta^* = \mathrm{Sgn}(AA^T\eta - \theta)$, where θ is a threshold vector. Baum, Moody and Wilczek[2] proposed a different associative and content addressable memory (ACAM), based on the notion of *grandmother cell* ([1] [25]). The idea is to define a string $\xi \in C_0^m := \{0,1\}^m$ of binary variables, each of them firing when a given word a^r is recognized. Given a partial or distorded version $y \in \mathbb{R}^n$ of one of the memories a^r, or of a combination of these memories, the inference problem consists in testing the hypothesis $z = a^r$, $r = 1,..,m$, on the basis of the observation y, where z is the intended memory behind y. The excitation pattern is estimated with the help of a test function:

$$\xi^* := H_\theta(A^T y) \in C_0^m,$$

where A^T denotes the matrix transpose, $H_\theta(x) \equiv H_0(x-\theta)$, and $H_0(\cdot)$ is the heaviside function, which is taken componentwise. This simple inference rule is based only on correlation, which is measured by the overlap between patterns . The threshold $\theta \in \mathbb{R}^m$ indicates a level of correlation, and we take two patterns a and $\tilde{a}$ to be similar when their overlap $a^T\tilde{a}$ is larger than a threshold value θ_r. The intended trace is estimated by computing $AH_\theta(A^Ty)$. In practice synapses have a finite accuracy, and the presented cue y is no longer only a superposition of memories of the list $\{a^1,..,a^m\}$, but a stochastically distorded version of it. Whittle's model[29], the *antiphon* (AP), proceeds by introducing noise in the above procedure, and proposes a device for realising reliable memory from unreliable components, by making the analogy of circulating a code word around a noisy channel. Consider a network with two set of nodes ([31])): m α-nodes and n β-nodes. An excitation pattern $\xi \in C_0^m$ is presented to the α-nodes, which is a code for the presence or the absence of a given memory trace, that is $\xi_r = 1$ means that the memory trace $a^r \in C_{-}^{n}$ is present. ξ is then presented to the β-nodes as an input, is decoded as $u := A\xi \in \mathbb{R}^n$, and is perturbed through a noisy channel with statistic $P(y|u) = \prod_{k=1}^{n} p(y_k|u_k)$. The α-nodes receive then $x = A^Ty$ as an input, and, using the Heaviside test function, give the excitation pattern $\xi^* := H_\theta(x) \in C_0^m$ as an output. More schematically, we write (AP)

$$\xi \longrightarrow A\xi \longrightarrow y \longrightarrow A^Ty \longrightarrow H_\theta(A^Ty) = \xi^*$$

the last part of the (AP), $y \longrightarrow A^Ty \longrightarrow H_\theta(A^Ty)$, corresponding to an inference step. The probability of error is simply given by $P_e := P(\xi^* \neq \xi)$, where we assume that the pattern matrix is chosen at random with a given distribution. The *capacity* problem consists in determinating the maximal rate of growth of m as function of n ensuring that $\lim_{n\to\infty} P_e = 0$. In what follows, we consider mainly two families of distributions. The first consists in choosing ξ uniformly over C_0^m, and the second is the family of uniform distributions on the Hamming spheres of radius $k \in \mathbb{N}$. Let $N(\xi,n,m)$ be the number of stored patterns. Then $N(\xi,n,m) = 2^m$ in the first case and $N(\xi,n,m) = \binom{m}{k}$ in the second case. In both situations, we say that the memory size is a positive capacity when $\lim_{n\to\infty} P_e = 0$ and $\lim_{n\to\infty} n^{-1}\log(N(\xi,n,m)) > 0$. When $A \in \{0,1\}^{nxm}$, Whittle ([29],[30]) considers the following channel statistics: $y_k \in \{0,1\}$ with transitions $p(y_k|u_k)$ of the form $P(y_k = 0|u_k) = \rho\gamma^{u_k}$, where $0 < \rho \leq 1$ and $0 < \gamma < 1$. In his model, the pattern matrix A is chosen to be the adjacency matrix associated with a random graph. For a well chosen threshold θ , it is shown that $\lim_{n\to\infty} P_e = 0$ when $m(n) \leq const n/\log(n)$, a lower bound comparable to that associated with Hopfield networks ([16], [17] , [20]). Both (AP) and the maximum likelihood (ML) can realize a positive capacity (see [31]), but this also requires an exponential growth of the number of computation steps. In what follows, we investigate the asymptotic behavior of the antiphon for thresholds of

the form $\theta_L = (1/2)W1_m$, where $W := A^TA$ and 1_m is the m-dimensional vector composed only of ones, and for thresholds of the form $\theta_s = sn1_m$, $s \in \mathbb{R}$. Our topic will consist in studying the domain of applicability of these various neural networks. The first threshold is used for the linear model $y = A\xi + R$, and the second for nonlinear model of the form $y = G(A\xi)+R$, where $G : \mathbb{R} \longrightarrow \mathbb{R}$ denotes a distorsion function. We also give an application to image analysis and pattern recognition.

II. The linear model

In this section, the channel statistics is the linear model $y = A\xi + R$, where $A \in \mathbb{R}^{nxm}$ is the memory matrix, ξ the excitation pattern and R a n-dimensional random vector with independent and identically distributed components. Formally, the inference step of the (AP) could be viewed as a way of making a regression, with the a priori knowledge that the intended parameter ξ lies in C_0^m. Let us denote by a_i, $1 \leq i \leq n$ the column vector in $\mathbb{R}^m$ whose coordinates form the ith row of the design matrix, and consider the least squares equation $0 = \sum_{i=1}^n a_{ij}(y_i - a_i^T\xi)$, $j = 1,..,m$, with solution given by $\hat{\xi}_{LS}(y) := (A^TA)^{-1}A^Ty$, $m < n$, $\mathrm{rank}(A) = m$, which suggests the inference rule, denoted by (LSAP), $y \longrightarrow H_\theta(\hat{\xi}_{LS}(y))$. Following Portnoy [23], the consistency of the least squares can be obtained when the distribution of z^Ta (typical column of A) does not depend too strongly on z. Let $y = A\xi_0 + R$. Suppose that $m(n)\log(n)/n \longrightarrow 0$ as $n \to \infty$. Under some conditions on the deterministic design matrix A, Portnoy proved that $\hat{\xi}_{LS}(y)$ is such that $||\hat{\xi}_{LS} - \xi_0||^2 = \mathcal{O}_m(m/n)$.

Consider the following antiphon, denoted by (LNAP),

$$\xi \longrightarrow z := 2y - A1_m \longrightarrow H_0(A^Tz) =: \xi_{LN}(y),$$

with

$$H_0(A^Tz) = H_{\theta_L}(A^Ty), \quad \theta_L := \frac{1}{2}A^TA1_m. \qquad \text{(II..1)}$$

When $\xi \in C_0^m$, $2y - A1_m = 2R + A(\xi - \bar{\xi})$, where $\bar{\xi}$ is given by $\xi + \bar{\xi} \equiv 1 (\mathrm{mod}\ 2)$. Let $W := A^TA \in \mathbb{R}^{mxm}$. When $R = 0$, the identification $C_0^m \sim C_-^m$, $\xi \longrightarrow \eta := \xi - \bar{\xi} \in C_-^m$, permits to get $P_e = P(H_0(A^Tz) \neq \xi) = P(\eta \neq \mathrm{Sgn}(W\eta))$, which is similar to the typical error probability associated with Hopfield neural networks. Note that for arbitrary threshold $\theta \in \mathbb{R}^m$, the bipolar antiphon becomes

$$\eta \longrightarrow z = A\eta + F \longrightarrow Sgn_{\theta'}(A^Tz) = \eta^*,$$

where $F = 2R$ and

$$\theta' = 2(\theta - \theta_L). \qquad \text{(II..2)}$$

Lemma 1 *Assume that $A \in \mathbb{R}^{nxm}$ is chosen at random, with i.i.d. entries a_i^r, $1 \leq i \leq n$, $1 \leq r \leq m$, such that $\mathrm{E}(a_i^r) = 0$ and $|a_i^r| < M$, for some positive constant M, and let $\mu_A^4 := \mathrm{E}(a_1^1)^4 < \infty$. Let $R \in \mathbb{R}^n$ be a zero-mean random vector independent of A, with i.i.d. components R_i such that $\mathrm{E}(R_1^2) := \sigma_R^2 < \infty$. Let ξ be chosen uniformly in C_0^m, independently of both A and R, and set $y(\xi) := A\xi + R$. Suppose that $P(A^Ty - \theta = 0) = 0$. (a) Assume that $m = m(n)$ is such that*

$$m(n) < \frac{\delta n}{2\log(n)} \frac{(\sigma_A^2 - \epsilon_1^2 - \epsilon_2^2)^2}{M^4},$$

for positive constants ϵ_1 and ϵ_2 such that $\epsilon_1^2 + \epsilon_2^2 < \sigma_A^2$. Then

$$P_e = P(\xi \neq H_{\theta_L}(A^Ty(\xi))) \leq \left(1 + \frac{4\sigma_R^2\sigma_A^2}{\epsilon_2^2} + \frac{\mu_A^4 + \sigma_A^4}{\epsilon_1^2}\right)\frac{m}{n}.$$

(b) Suppose moreover that $P(a = 1) = P(a = -1) = 1/2$, and that the law $\mathcal{L}(R_1)$ of R_1 is such that $\mathcal{L}(R_1) = \mathcal{L}(-R_1)$. Let $m = m(n) \to \infty$ as $n \to \infty$ be such that

$$m(n) < \frac{\delta n}{2\log(n)},$$

for some constant $0 < \delta < 1$. Then

$$P_e \leq \left(\frac{4\sigma_R^2}{(1 - \sqrt{\delta})^2} + \frac{1}{\sqrt{2\pi\delta}}\sqrt{\frac{m}{n}}\right)\frac{m}{n} = O(\frac{m}{n}),$$

Under the assumptions of Lemma 1, the memory size is thus larger that *const* $n/\log(n)$. We suspect that (b) gives the best possible rate (see [17]).

Classical statistical methods preassume consistency when $R = 0$; being designed to facilitate computations, neural nets of the form $H_\theta(Lz)$, $L \in \mathbb{R}^{mxn}$ provide non consistent estimators when $R = 0$, but are asymptotically consistent in large dimensions. A way of finding a ($R = 0$) consistent estimator is to impose that $\eta = LA\eta$, $\forall\eta \in C_-^m$, to get the error probability $P_e = P(\mathrm{Sgn}(\eta) \neq \eta) = 0$. In this case, we necessarily have $A\eta = ALA\eta$, $\forall\eta \in C_-^m$, which is satisfied when $A = ALA$. The solutions are generated by the projection learning rule, or Penrose's pseudo-inverse (see e.g. [14] or [19]) $L = A^+ := (A^TA)^{-1}A^T$, $m < n$, $\mathrm{rank}(A) = m$. The associated inference rule $\eta^* = \mathrm{Sgn}(Lz)$ becomes $\mathrm{Sgn}(A^+z) = \mathrm{Sgn}(\hat{\eta}_{LS}(z))$.

In fact, the object is different of the usual regression task: the regressors are restricted to be binary variables, indicating only the presence or the absence of a given explanatory variable behind the data, and the inference problem is more close to a m-class problem.

III. Classification

Given a space Ω, a decision rule partitions Ω into components Ω_i, $i = 1,..,N$, representing classes ω_i, $i = 1,..,N$. An observation y is classified as coming from class ω_i when y is in Ω_i. We assume a uniform prior probability. Given an observation y, a natural way of discriminating

between the possible classes ω_i is to choose the class ω_i maximizing the posterior probability $P(\omega_i|y)$. In general, the above probabilities can not be computed since the exact statistical model is not available, and should be estimated from the design set. Using Bayes's theorem, the above decision rule becomes

$$P(y|\omega_i) > P(y|\omega_j)\ \forall j \neq i\ \Rightarrow\ y \in \Omega_i.$$

Such classification procedure are currently developped in the (ANN) litterature, e.g. in engineering and physics [2], in handwriting recognition systems (see e.g.[6], [26]), in psychology ([10], [15]) and in various other fields, providing an interesting link between statistical techniques (see e.g. [11]) and neural nets methods.

In what follows, we set $N = N(\xi, n, m)$ (which depends of the distribution of ξ, see section I.), each class being associated with a code $\xi \in C_0^m$. Consider the Gaussian case, where each class Ω_ξ is statistically represented by a normal distribution $\mathrm{N}(\mu_\xi, \sigma^2\mathrm{Id})$, with $\mu_\xi := A\xi \in \mathbb{R}^n$. When $S = \{a^1, .., a^m\} \subset C_-^n$ and $|\xi| = 1$, the (ML) estimate is obtained by maximizing the scalar product $y^T\mu_\xi$, and the intended trace behind the data y is picked from the set

$$\mathrm{argmax}\ y^T\mu_\xi. \qquad \text{(III..1)}$$

Assume that $y \in S$, that is $R \equiv 0$. Then (III..1) can be realized as a two-layer neural network, $(MLAP)$, (input of n neurons (y) and output of m neurons $(\bar{\xi}^*)$) with activation rule given by

$$\bar{\xi}^* = H_{\bar{\theta}}(A^T y), \quad \bar{\theta} := (n-1)1_m. \qquad \text{(III..2)}$$

When $R \neq 0$, consider the event $\{\bar{\xi}_1^* \neq 1\}$. Then $P(\bar{\xi}_1^* \neq 1 | y = a^1 + R) = P(< a^1, R >< -1)$, and therefore $\liminf_{n\to\infty} P(\bar{\xi}_1^* \neq 1 | y = a^1 + R) > 0$, showing that the above neural network permits perfect discrimination when $R \equiv 0$ but is not consistent when $R \neq 0$. A natural way of avoiding the above problem consists in choosing $\theta_s = sn1_m$ with $s > 0$ such that $\lim_{n\to\infty} P(H_{\theta_s}(A^T y) \neq e^1 | y = a^1 + R) = 0$, $e^1 = (1, 0, .., 0)^T$, which implies that $P(||a^1||^2 + < a^1, R >< sn) \longrightarrow 0$, as $n \to \infty$. Approximating $||a^1||^2$ by $n\sigma_A^2$, we obtain from the law of large numbers that we must choose s with $0 < s < \sigma_A^2$. We will therefore consider the new antiphon, denoted by (LAP),

$$\xi \longrightarrow A\xi \longrightarrow y = A\xi + R \longrightarrow H_{\theta_s}(A^T y)\ =\ \xi^*,$$

$\theta_s = sn1_m$, $0 < s < \sigma_A^2$, (when $\mathrm{E}(a) = 0$). In some sense, (LAP) derives from (MLAP) and is obtained by picking ξ in a larger subset.

Nonlinear distorsions: The models encoutered in the previous section are linear, perturbated by additive noise. Here, we investigate the case where the transmitted signal $A\xi$ is distorded by a function $G : \mathbb{R} \longrightarrow \mathbb{R}$, and perturbated by additive noise, that is, let

$$y\ =\ G(A\xi) + R,$$

where G is applied componentwise. As we will see, inference rules of the generic form $\xi^*\ =\ H_\theta(A^T y)$ still work, but under assumptions on the form of the underlying excitation. Given $G : \mathbb{R} \longrightarrow \mathbb{R}$ and $A \in \mathbb{R}^{nxm}$, let

$$\Lambda(G|a)\ :=\ \{k \in \mathbb{N}_+;\ \mathrm{Cov}(a^1, G(a^1 + .. + a^k)) > 0\} \qquad \text{(III..3)}$$

where the sequence $a^1, ..$ is i.i. distributed according to the law $\mathcal{L}(a)$ of a typical entry of A. $k \notin \Lambda(G|a)$ when the associated expectation does not exist. Let $\mu_k := \mathrm{E}(a^1 G(a^1 + .. + a^k))$, $\mu_{G,k} := \mathrm{E}(G(a^1 + .. + a^k)$, $\sigma_k^2 := \mathrm{E}((a^1 G(a^1 + .. + a^k))^2)$, and $\sigma_{G,k}^2 := \mathrm{E}(G(a^1 + .. + a^k)^2)$.

Theorem 1 *Assume that $A \in \mathbb{R}^{nxm}$ has i.i.d. entries a_i^r, $1 \leq i \leq n$, $1 \leq r \leq m$. Assume that the random vector $R \in \mathbb{R}^n$ has i.i.d. zero-mean components R_i with finite second moment $\mathrm{E}(R_1)^2 = \sigma_R^2 < \infty$, and is independent of A . Let $G : \mathbb{R} \longrightarrow \mathbb{R}$, and consider the random variable $y = G(A\xi) + R$. Suppose that $P(A^T y - sn1_m = 0) = 0$, $\forall s \in \mathbb{R}$.*

(a): Assume that $\Lambda(G|a) \neq \emptyset$, and let $k \in \mathbb{N}$ be such that $k \in \Lambda(G|a)$, $\sigma_{G,k}^2 < \infty$ and $\sigma_k^2 < \infty$. Then, choosing ξ uniformly in $\{\xi \in C_0^m;\ |\xi| = k\}$, we have

$$P(\xi \neq H_0(A^T y - sn1_m)) \leq \frac{k}{n}\Big(\frac{\mathrm{E}(a^2)\sigma_R^2}{(\mu_k - s - \epsilon)^2} + \frac{\sigma_k^2 - \mu_k^2}{\epsilon^2}\Big)$$

$$+\frac{(m-k)}{n}\frac{\big(\mathrm{E}(a^2)(\sigma_{G,k}^2 + \sigma_R^2) - (\mathrm{E}(a)\mu_{G,k})^2\big)}{(s - E(a)\mu_{G,k})^2},$$

$\forall s$ such that $E(a^1)E(G(a^1 + .. + a^k)) < s < E(a^1 G(a^1 + .. + a^k))$, $\forall 0 < \epsilon < \mu_k - s$.

(b): On the other hand, let $k \in \mathbb{N}_+$ be such that $\mu_k \leq \mathrm{E}(a)\mu_{G,k}$, $\sigma_k^2 < \infty$ and $\sigma_{G,k}^2 < \infty$, and let ξ be chosen uniformly in the Hamming sphere of radius k. Then , $\forall s \in \mathbb{R}$,

$$\liminf_{n\to\infty,\ m\to\infty} P(\xi \neq H_0(A^T y - sn1_m)) > 0,$$

Illustrations: When a has finite range, the (FKG) inequality implies that $Cov(a^1, G(a^1 + .. + a^k)) \geq 0$, for every bounded increasing function G. Another situation is the normal case. Let N denote a standard normal random variable, and assume that G is differentiable with bounded derivative. We have $\mathrm{E}(NG(N)) = \mathrm{E}(G'(N))$. When $A \in \{-1, +1\}^{nxm}$, with $P(a = -1) = P(a = +1) = 1/2$, we can see that $\Lambda(G|a) = \emptyset$, for every even function G.

Example 1 (Random linear codes) *Assume that $P(a = 1) = p < 1$, $P(a = 0) = 1 - p < 1$, and choose ξ at random in the Hamming sphere of radius $k \in \mathbb{N}$. Let $G(x) \equiv x \bmod 2$. Then $\mathrm{Cov}(a_1, G(a_1 + .. + a_k)) = p(1-p)(1-2p)^{k-1}$, showing that $\Lambda(G|a)$ is equals to $\mathbb{N}_+$ when $p < 1/2$, to $\emptyset$ when $p = 1/2$, and to $2\mathbb{N} + 1$ when $p > 1/2$. The set $\{G(A\xi);\ \xi \in C_0^m, |\xi| = k\}$ is a subset of the linear code of GF_2^n generated by the random set $S = \{a^1, .., a^m\}$.*

Example 2 (RCAM) *The recurrent correlation associative memory [3] is a recursive network with evolution*

equation given by $x(t+1) = Sgn(AF(A^T x(t)))$, $t \in \mathbb{N}$, $x(t) \in C_-^n$, where F is a weighting function assumed to be monotone nondecreasing (see also [5], [22] and [24]). For example, the Hopfield recursive network corresponds to $F(x) \equiv x$. To make a link with (AP), we choose $F \equiv H_{\theta_s}$ and $G \equiv Sgn$. We then get the double recursion $\xi(t) = F(A^T x(t))$, $x(t+1) = G(A\xi(t))$, which gives $\xi(t+1) = F(A^T G(A\xi(t)))$, a recursive antiphon (given $\xi(0)$, set $\xi(1) = \xi(0)^*$ and so on). (LAP) is thus seen as a one-shot version of this particular (RCAM). Assume that $P(a=1) = p$, $P(a=-1) = 1-p$, for some $0 < p < 1$. Then $Cov(a, G(a)) = 1 - (1-2p)^2 > 0$.

An application to image recognition: Suppose that the patterns $a^1, .., a^m$ represent discretized images. The discretization is obtained by restricting the image to a two dimensional grid of side L, with $n = L^2$ pixels. The deffects occuring in the images are modeled as being the results of three main causes: blurring, distorsion and random noise. The input image $a = (a_{ij})_{1 \leq i,j \leq L}$ is perturbed according to the following model

$$y \ = \ G(B(a)) + R,$$

where $G : \mathbb{R} \longrightarrow \mathbb{R}$ models the effect of transmission, B represents the blurring and R is the random noise. Typically, B is chosen to be a convolution through a small window of the form $B(a)_{ij} \ := \ \sum_{kl} B_{i-k,j-l} a_{kl}$ (see [8] for more details). Usually, the basic images are modeled to be realizations of a Markov random field; here, we suppose independence between sites .

Lemma 2 *Assume that the set of stored images $S = \{a^1, .., a^m\}$ is drawn at random with i.i.d. entries a_{ij} such that $\mathrm{E}(a_{ij})^2 < \infty$, and that the random perturbation R is independent of A with i.i.d. entries R_{ij} such that $\mathrm{E}(R_{ij}) = 0$ and $\mathrm{E}(R_{ij})^2 < \infty$. Let the convolution window be such that $B_{00} > 0$, $B_{k,l} \geq 0$ when $|k| \leq 1$, $|l| \leq 1$, $(k,l) \neq (0,0)$, and $B_{k,l} = 0$ otherwise. Suppose that $y = G(Ba) + R$, is such that $P(A^T y - s 1_m = 0) = 0$, $\forall s \in \mathbb{R}$. Assume moreover that $\mathrm{E}(a_{ij} G(Ba)_{ij})^2 < \infty$, $\mathrm{E}(G(Ba)_{ij})^2 < \infty$, and that*

$$\mathrm{Cov}(a_{ij}, G(Ba)_{ij}) > 0, \quad \forall (i,j). \qquad \text{(III..4)}$$

Then $\lim_{L \to \infty} P(\xi \neq H_{\theta_s}(A^T y) || \xi | = 1) = 0$, when $m = o(n)$ and

$$\mathrm{E}(a_{ij})\mathrm{E}(G(Ba)_{ij}) < s < \mathrm{E}(a_{ij} G(Ba)_{ij}). \qquad \text{(III..5)}$$

The above Lemma shows that (LAP) detects positive correlations, and can be used for retrieving images in a large database. Note that the main difference with the previous results is that the channel statistics is no longer of the form $\prod_i p(y_i | u_i)$ (see section I.), the convolution window introducing interactions between components.

IV. The Perceptron

The classical perceptron (P) is a single formal neuron with bipolar output, and activation rule given by

$$\mu^* \ = \ Sgn_\lambda(J^T y) \in \{-1, 1\}, \quad y \in \mathbb{R}^n,$$

with connection vector $J \in \mathbb{R}^n$ and threshold $\lambda \in \mathbb{R}$. Given a set of m patterns $S = \{a^1, .., a^m\}$, and a set of desired outputs $\{\mu_1, .., \mu_m\}$, one for each pattern a^r, the object is to find $J \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ in such a way that

$$\mu_r \ = \ Sgn_\lambda(J^T a^r), \quad r = 1, .., m.$$

Let $\mu := (\mu_1, .., \mu_m)^T \in C_-^m$. Let $\Omega_P[S, \mu]$ be the solution set associated with the perceptron, given by $\Omega_P[S, \mu] = \{(J, \lambda) \in \mathbb{R}^n \mathrm{x} \mathbb{R}; \ Sgn_0(A^T J - \lambda 1_m) = \mu\} \subset \mathbb{R}^{n+1}$. Standard perceptron learning theory shows that $\Omega_P[S, \mu] \neq \emptyset$ if and only if $T = (S, \mu)$ is linearly separable. Let $\theta' := \lambda 1_m$.

For given θ', we can consider the *dual* perceptron (P_m^*) (see [18]) which is a two-layer neural network (output of m bipolar neurons (μ) and input of n neurons (J)), given by $\mu = Sgn_0(A^T J - \theta')$. The m hyperplanes representing the elements of S decompose $\mathbb{R}^n$ in polygonal domains, each of them corresponding to a bipolar vector. A dichotomy $T = (S, \mu)$ is separable if and only if there exists a domain $D_\mu \subset \mathbb{R}^n$ with output $\mu \in C_-^m$.

The *Vapnik dimension* of (P_m^*) is equal to n (see [18] or [21]) and thus, when $m > n$, only a fraction of the m-cube C_-^m is realized by the mosaic. The randomized Perceptron problem consists in studying $\Omega_P[S, \mu]$ when S has a given distribution and μ is uniformly distributed on C_-^m. Let m be of the form $m = [\alpha n]$, for some constant $\alpha > 0$. When $\theta' = 0_m$ and $0 < \alpha < 2$, Cover[4] proved that $\lim_{n \to \infty} P(\Omega_P[S, \mu] \neq \emptyset) = 1$, showing that the portion of C_-^m which is not realized remains vanishingly up to the *critical storage capacity* $\alpha_c = 2$. However, when $\alpha > \alpha_c$, this probability converges to 0.

When $\theta' = 0_m$, (P_m^*) can be viewed as the neural network associated with the inference step of (LNAP) (see II..1, II..2), which can also be seen as a dual multi-layer perceptron, with one hidden layer of n neurons, the input (η) and the output (η^*) layers containing m neurons. This also permits to give a partial solution to the randomized perceptron problem. Under the assumptions of Lemma 1, $\lim_{n \to \infty} P((A\mu, 0) \in \Omega_P[S, \mu]) \longrightarrow 1$ when $m(n) < const \ n / \log(n)$, giving in this way a consistency result for *the correlation learning rule* (see e.g.[12]). Practically, the components of the vector $J = A\mu$ have a very broad distribution, and it would be suitable to have bounded weights (see [27] or [28]). The following is a consequence of Theorems 2 and 3 of [16].

Theorem 2 *Assume that A has i.i.d. entries a_{ij} with $P(a=-1) = P(a=+1) = 1/2$, and that ξ is independent of A, uniformly distributed on C_0^m. Let N denote a standard normal random variable, and let $G : \mathbb{R} \longrightarrow \mathbb{R}$ be bounded with bounded continuous derivative G'. Let*

$J_G(\xi) := G(\frac{1}{\sqrt{m}} A\eta)$, $\eta := 2\xi - 1_m \in C_-^m$. *i) If G is such that* $\mathrm{E}(NG(N)) = \mathrm{E}(G'(N)) > 0$, *then* $\lim_{n\to\infty} P(\xi \neq H_0(A^T J_G(\xi)) = 0$ *when* $m(n) < (n/(2\log(n)))q_G$, *where* $0 < q_G := \mathrm{E}(NG(N)))^2/(\mathrm{E}(G^2(N))) \leq 1$. *ii) Let G be such that* $\mathrm{E}(NG(N)) < 0$. *Then,* $\forall 0 < \delta < (1/2)$, $\exists n_0 \in \mathbb{N}$ *with* $P(\xi = H_0(A^T J_G(\xi))) < \frac{1}{2} + \delta$, $\forall n \geq n_0$.

We suspect that the $n/(2\log(n))q_G$ is the best possible, and that this result should extend to more general situations.

Preprocessing: The above considerations give us a new way of sending data through a noisy channel. Given a message (S, η), find a vector $J[S, \eta] \in \mathbb{R}^n$ such that

$$\eta = \mathrm{Sgn}_0(A^T J[S, \eta]). \qquad \text{(IV..1)}$$

Send $J[S, \eta]$ through the noisy channel to get the output $z \in \mathbb{R}^n$. Decode the message according to the antiphon inference rule $\eta^* = \mathrm{Sgn}_0(A^T z)$. Cover's result says nothing but that $P(\eta^* \neq \eta | R = 0) \longrightarrow 0$, $n \to \infty$ when $m = m(n) < 2n$. When $R = 0$, this preprocessing phase permits to increase the *const* $n/\log(n)$ rate to a $2n$ rate (i.e. a positive capacity), allowing thus the storage of 2^{2n} superposed messages.

One way of finding a vector $J[S, \eta]$ satisfying (IV..1) consists in using Rosenblatt's Perceptron learning algorithm, which is defined as follows: Given an arbitrary initial vector $J(0) \in \mathbb{R}^n$, consider the dynamical system $J(t+1) = J(t) + I(J(t)^T \eta_{r(t+1)}\, a^{r(t+1)} \leq 0)\, \eta_{r(t+1)} a^{r(t+1)}$, $t \in \mathbb{N}$, where the sequence $\{r(t)\}_{t\in\mathbb{N}}$ is i.i.d. uniformly distributed over the set $\{1, 2, .., m\}$. When the dichotomy (S, η) is linearly separable, Rosenblatt proved that the above algorithm converges in a finite number of steps to a solution of equation (IV..1).

Let us perhaps give an intuitive description of the above algorithm for a start with $J(0) = A\eta$. By construction, $J(1) \neq J(0) = A\eta$ if and only if the antiphon inference rule makes an error at the $r(1)$th component, that is if $\eta_{r(1)} \neq \mathrm{Sgn}_0(A^T(A\eta))_{r(1)}$. The Perceptron learning algorithm is thus seen as a way of *coding the message by simulating step by step the behavior of the inference rule.* The practical drawback of this preprocessing step is the time needed to learn (S, η) with the Perceptron algorithm, and almost nothing is known concerning its average convergence time. It would be interesting to check the behavior of the above coding-denoising scheme when $R \neq 0$. This seems to be a non obvious problem. The study of the volume $D_\eta \subset \mathbb{R}^n$ consisting of those vectors J such that $\eta = Sgn(A^T J)$ is very hard to handle, and the only known results are due to Gardner[7], where the volume is considered through the replica mean field method.

V. Concluding remark

We have considered a generic inference rule for recovering stored patterns from noisy data, and explored the associated memory sizes. The basic architecture of the networks is already used in content-based image retrieval systems (see e.g. [9]) where the structure is ad hoc. Our results (Theorem 1, Lemmas 1 and 2) give a canonical way of choosing similarity thresholds ensuring statistical consistency. The asymptotic normality of the (LS) has been investigated in [13] under some conditions on the growth of $m(n)$, and the statistical validation of the neural procedure should also be envisaged, at least to allow a comparison of the inference rules based on θ_L and θ_s.

VI. Appendix

Proof of Lemma 1: (a): We have

$$P_e \leq \sum_{r=1}^{m} P(\eta_r(\sum_{l\neq r}^{m} w_{rl}\eta_l + 2\sum_{i=1}^{n} R_i a_i^r) < -||a^r||^2).$$

Using the independence between η_r and the other variables, we get that $P_e \leq (1/2)\sum_{r=1}^{m}(P_{e,r}^- + P_{e,r}^+)$, where

$$P_{e,r}^- := P(\sum_{l\neq r}^{m} w_{rl}\eta_l + 2\sum_{i=1}^{n} R_i a_i^r < -||a^r||^2),$$

and $P_{e,r}^+ := P(\sum_{l\neq r}^{m} w_{rl}\eta_l + 2\sum_{i=1}^{n} R_i a_i^r > ||a^r||^2)$. Consider first $P_{e,r}^+$:

$$P_{e,r}^+ = \int_{\mathbb{R}^n \mathrm{x} C_-^{m-1}} P(\sum_{l\neq r}\sum_{i=1}^{n} \bar{\eta}_l a_i^r a_i^l + 2\sum_{i=1}^{n} R_i a_i^r > ||a^r||^2) dF_n(a^r) dG_m(\bar{\eta}),$$

where F_n and G_m denote the repartition functions of a^r and $\bar{\eta} = (\eta_1, .., \eta_{r-1}, \eta_{r+1}, .., \eta_m) \in C_-^{m-1}$. The above probability is bounded by

$$P(A_{\epsilon_1}^c) + P(A_{\epsilon_2}^c) + P(\sum_{l\neq r}\sum_{i=1}^{n} \bar{\eta}_l a_i^r a_i^l n^{-1}\epsilon_2 > \sigma_A^2 - \epsilon_1)$$

, where $A_{\epsilon_1} := \{|||a^r||^2/n - \sigma_A^2| < \epsilon_1\}$, and $A_{\epsilon_2} := \{|2\sum_{i=1}^{n} R_i a_i^r/n| < \epsilon_2\}$. Using Chebyshev's inequality we get that $P(A_{\epsilon_2}^c) \leq 4\sigma_A^2\sigma_R^2/(n\epsilon_2^2)$ and $P(A_{\epsilon_1}^c) \leq (\mathrm{E}(a_1^1)^4 + \sigma_A^4)/(n\epsilon_1^2)$, and it remains to consider the probability

$$P(\sum_{l\neq r}\sum_{i=1}^{n} \bar{\eta}_l a_i^r a_i^l > n(\sigma_A^2 - \epsilon_1 - \epsilon_2)).$$

Conditional on a^r and $\bar{\eta}$, the variables $y_{il} := \bar{\eta}_l a_i^r a_i^l$ are independent. Using Hoeffding's inequality we obtain that

$$P(\sum_{l\neq r}\sum_{i=1}^{n} y_{il} > n(\sigma_A^2 - -\epsilon_1 - \epsilon_2)) \leq \exp(-\frac{n}{m}\frac{(\sigma_A^2 - -\epsilon_1 - \epsilon_2)^2}{2M^4}).$$

The same arguments also hold for $P_{e,r}^-$, and we get the result.

Concerning (b), see [16] or [17]. □

Proof of Theorem 1: We must consider the error probability

$$P_e \;=\; \binom{m}{k}^{-1} \sum_{|\xi|=k} P(\xi \neq H_0(A^T(G(A\xi)+R) - sn1_m)).$$

Assume without loss of generality that $\xi = (1,..,1,0,..,0)^T$, the first k components of ξ equal to 1 and the remaining $m-k$ equal to 0. We have

$$P(\xi \neq H_0(A^T(G(A\xi)+R) - sn1_m))$$
$$\leq \sum_{l=1}^{k} P(< a^l, G(A\xi)+R >< sn)$$
$$+ \sum_{l>k} P(< a^l, G(A\xi)+R >> sn).$$

Assume first that $1 \leq l \leq k$: We have

$$P(\sum_{i=1}^{n} a_i^l \big(G(\sum_{\nu=1}^{m} a_i^\nu \xi_\nu) + R_i\big) < sn)$$
$$= P(\sum_{i=1}^{n} a_i^l \big(G(\sum_{\nu=1}^{k} a_i^\nu) + R_i\big) < sn),$$

which is probability of large deviations for a sum of i.i.d. random variables, which becomes

$$P(\sum_{i=1}^{n} a_i^l R_i < sn - \sum_{i=1}^{n} a_i^l G(\sum_{\nu=1}^{k} a_i^\nu)).$$

Consider the event $A_\epsilon := \{|n^{-1}\sum_{i=1}^n a_i^l G(\sum_{\nu=1}^k a_i^\nu) - \mu_k| \leq \epsilon\}$, where $\epsilon > 0$. The above probability is then bounded by

$$P(A_\epsilon^c) + P(\{-n^{-1}\sum_{i=1}^{n} a_i^l R_i > (-s + n^{-1}\sum_{i=1}^{n} a_i^l G(\sum_{\nu=1}^{k} a_i^\nu))\} \cap A_\epsilon),$$

and therefore, choosing ϵ such that $0 < \epsilon < \mu_k - s$, we obtain the upper bound

$$P(A_\epsilon^c) + P(-n^{-1}\sum_{i=1}^{n} a_i^l R_i > \mu_k - s - \epsilon).$$

Of course we assume that $k \in \Lambda(G|a)$ and that $s < \mu_k$. Using Chebyshev's inequality, the second term is bounded by $\mathrm{E}(a^2)\sigma_R^2/(n(\mu_k - s - \epsilon)^2)$, and it remains to check the behavior of the first term. Using the same argument, we obtain that $P(A_\epsilon^c) \leq (\sigma_k^2 - \mu_k^2)/(n\epsilon^2)$. In summary, we have proved that $\sum_{l=1}^k P(< a^l, G(A\xi)+R >< sn)$ is bounded by

$$\frac{k}{n}\big(\frac{\mathrm{E}(a^2)\sigma_R^2}{(\mu_k - s - \epsilon)^2} + \frac{\sigma_k^2 - \mu_k^2}{\epsilon^2}\big).$$

Let us now consider the case $l > k$, and more precisely the probability

$$P(\sum_{i=1}^{n} \big(a_i^l (G(\sum_{\nu=1}^{k} a_i^\nu) + R_i) - \mathrm{E}(a)\mu_{G,k}\big) > (s - \mathrm{E}(a)\mu_{G,k})n),\ k < l,$$

which is bounded by

$$\frac{\mathrm{E}(a^2)(\sigma_{G,k}^2 + \sigma_R^2) - (\mathrm{E}(a)\mu_{G,k})^2}{n(s - E(a)\mu_{G,k})^2},$$

as required.

Let us now consider b). Without loss of generality, asssume as above that $\xi = (1,..,1,0,..,0)^T$. Then

$$P(\xi \neq H_0(A^T y - sn1_m) \geq$$
$$P(\xi_l \neq H_0(< a^l, G(A\xi) + R > -sn)),\ \forall l.$$

First assume that $s < \mathrm{E}(a)\mu_{G,k}$, and let $l > k$ (this is possible since the asymptotic concerns the case $m \to \infty$, and k is kept fixed). Then the probability $P(\xi_l \neq H_0(< a^l, G(A\xi) + R > -sn)$ becomes

$$P(n^{-1}\sum_{i=1}^{n} a_i^l (G(\sum_{\nu=1}^{k} \big(a_i^\nu) + R_i) - \mathrm{E}(a)\mu_{G,k}\big) > s - \mathrm{E}(a)\mu_{G,k}),$$

which converges to 1 by the law of large numbers. Assume now that $s > \mathrm{E}(a)\mu_{G,k}$, and let l be such that $1 \leq l \leq k$. Then

$$P(\xi_l \neq H_0(< a^l, G(A\xi) + R > -sn))$$
$$\geq P(\{-\sum_{i=1}^{n} \frac{a_i^l R_i}{n} > -s + \sum_{i=1}^{n} \frac{a_i^l G(\sum_{\nu=1}^k a_i^\nu)}{n}\} \cap A_\epsilon)$$
$$\geq P(\{-\frac{\sum_{i=1}^n a_i^l R_i}{n} > -s + \mu_k + \epsilon\} \cap A_\epsilon)$$
$$\geq P(-\sum_{i=1}^{n} \frac{a_i^l R_i}{n} > -s + \mu_k + \epsilon) - P(A_\epsilon^c).$$

For well chosen $\epsilon > 0$, $-s + \mu_k + \epsilon < 0$, and the result is a consequence of the law of large numbers.

When $s = \mathrm{E}(a)\mu_{G,k}$, choose $l \in \mathbb{N}_+$ with $l > k$. Then

$$P(\xi_l \neq H_0(< a^l, G(A\xi) + R > -sn))$$
$$= P(\sum_{i=1}^{n} a_i^l \big(G(\sum_{\nu=1}^{k} a_i^\nu) + R_i\big) > n\mathrm{E}(a)\mu_{G,k}),$$

which converges to $1/2$ by the central limit theorem. □

Proof of Lemma 2: Without loss of generality, assume that $\xi = (1,0,..,0)^T$. The error probability is bounded by

$$P(\sum_{(ij)} z_{ij} < m(s - \mu(n))) + (m-1)P(\sum_{(ij)} w_{ij} > n(s - \beta(n))),$$

where we set

$$z_{ij} \;:=\; a_{ij}^1 \big(G(\sum_{(kl)} B_{i-k,j-l} a_{kl}^1) + R_{ij}\big) - \mu_{ij},$$
$$w_{ij} \;:=\; a_{ij}^2 \big(G(\sum_{(kl)} B_{i-k,j-l}) + R_{ij}\big) - \beta_{ij},$$
$$\mu_{ij} \;:=\; \mathrm{E}(z_{ij}),\ \beta_{ij} \;:=\; \mathrm{E}(w_{ij}),$$

$$\mu(n) \;:=\; n^{-1}\sum_{(ij)}\mu_{ij}, \quad \text{and} \quad \beta(n) \;:=\; n^{-1}\sum_{(ij)}\beta_{ij}.$$

Using Chebyshev's inequality, we get the upper bound

$$\frac{\mathrm{E}(\sum_{(ij)} z_{ij})^2}{n^2(s-\mu(n))^2} + \frac{\mathrm{E}(\sum_{(ij)} w_{ij})^2}{n^2(s-\beta(n))^2}.$$

First note that $\mu(n) = n^{-1}\sum_{(ij)\in\partial\Lambda_n}\mu_{ij} + n^{-1}\sum_{(ij)\in\Lambda_n^0}\mu_{ij}$, where Λ_n denotes the grid, $\partial\Lambda_n$ its boundary and Λ_n^0 its interior. So

$$\mu(n) \longrightarrow \mathrm{E}(a_{ij}G(Ba)_{ij}), \quad L\to\infty, \quad (ij)\in\Lambda_\infty^o.$$

Similary we have $\lim_{L\to\infty}\beta(n) = \mathrm{E}(a_{ij}^2 G(Ba^1)_{ij})$, $(ij)\in\Lambda_\infty^o$. Next consider the sum

$$\mathrm{E}(\sum_{(ij)} z_{ij})^2 \;=\; \sum_{(ij)\in\Lambda_n^0} z_{ij}^2 + \sum_{(ij)\in\partial\Lambda_n} z_{ij}^2 + \sum_{(ij)\neq(kl)} \mathrm{E}(z_{ij}z_{kl}).$$

By assumption, there exist constants C_1 and C_2 such that

$$\mathrm{E}(\sum_{(ij)\in\partial\Lambda_n} z_{ij}^2) \le C_1 L, \quad \mathrm{E}(\sum_{(ij)\in\Lambda_n^0} z_{ij}^2) \le C_2 L^2.$$

It remains to consider the sum $\mathrm{E}(\sum_{(ij)\neq(kl)} z_{ij}z_{kl})$. But $\mathrm{E}(z_{ij}z_{kl}) = 0$ as soon as $|i-k|+|j-l|>2$, and there exists a positive constant C_3 with $\mathrm{E}(\sum_{(ij)\neq(kl)} z_{ij}z_{kl}) \le C_3L^2$. There also exists a positive constant C such that

$$P(\sum_{(ij)} z_{ij} < n(s-\mu(n))) \le \frac{C}{L^2(s-\mu(\infty))^2}.$$

Applying the same arguments to the second probability, we get that there exists a positive constant D such that

$$mP(\sum_{(ij)} w_{ij} > n(s-\beta(n))) \le \frac{mD}{L^2(s-\beta(\infty))^2}.$$

□

Acknowledgment: The author is grateful to an anonymous referee for his careful reading and for several constructive remarks.

References

[1] Barlow, H.B.(1972). Single units and sensation: a neuron doctrine for perceptual psychology. *Perception* 1:371.

[2] Baum, E.B., Moody, J. and Wilczek, F.(1988). Internal Representations for Associative Memory. *Biological Cybernetics***59** 217-228.

[3] Chiueh, T., Goodman, R.(1991). Recurrent Correlation Associative Memories. *IEEE Trans. N. N.***2** 275-284.

[4] Cover, T.M.(1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers***14**, 326-334.

[5] Dembo, A., Zeitouni, O. (1988). High density associative memories, in *Neural Information Processing Systems*, New York, American Institute of Physics, 211-218.

[6] Dreyfus, G., Knerr, S., Personnaz, L. and Price, D.(1994). Pairwise Neural Network Classifiers with Probabilistic Outputs. *Neural Information Processing Systems.*

[7] Gardner E. (1988). The space of interactions in neural networks models. *J. Phys. A: Math. Gen.* **21**, p. 257-270.

[8] Geman S. and Geman D.(1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern and Machine Intelligence***6** pp. 721-741.

[9] Gudivada V.N. and Raghavan V.V. Content-Based Retrieval Systems. *Computer*, september 1995.

[10] Gluck, M.A. and Bower, G.H.(1988). From Conditioning to Category Learning: An Adaptive Network Model. *Journal of Experimental Psychology: General* **117** (3) 227-247.

[11] Hand, D.J. *Discrimination and Classification.* Wiley, 1981.

[12] Hassoun M. *Fundamentals of artificial intelligence*, Bradford, MIT Press 1995.

[13] Huber P.(1973). Robust Regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.***1** pp. 799-821.

[14] Kamp, Y. and Hasler, M. *Recursive Neural Networks for Associative Memory* Wiley, 1990.

[15] Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* **99** 22-44.

[16] Mazza, C.(1995). On the storage capacity of nonlinear neural networks. To appear in *Neural Networks.*

[17] McEliece, R.J., Posner, E.C., Rodemich, E.R and Venkatesh, S.S.(1987). The Capacity of the Hopfield Associative Memory. *IEEE Transactions on Information Theory*, **IT-33**(4) 461-482.

[18] Nadal, J-P. Formal Neural Networks: From Supervised to Unsupervised Learning. In *Cellular Automata, Dynamical Systems and Neural Networks*, Eds E. Goles and S. Martinez, Kluwer Academic Publishers, 1994.

[19] Personnaz, L., Guyon, I. and Dreyfus, G.(1985). Information storage and retrieval in spin-glass like neural networks. *J. Physique Lett.***46** 359-365.

[20] Piret, P.(1990). Analysis of a Modified Hebbian Rule. *IEEE Transactions on Information Theory*, **36**(6).

[21] Pollard, D. *Convergence of Stochastic Processes*, Springer (1989).

[22] Sayeh, R., Han, J.(1987) Pattern recognition using a neural network. In *Proc. SPIE Cambridge Symp. Opt. and Optoelec. Eng.* (Cambridge, MA)

[23] Portnoy, S.(1984). Asymptotic Behavior of M-Estimators of p Regression Parameter when p^2/n is Large: Consistency. *The Annals of Statistics*, **12**(4) 1298-1309.

[24] Psaltis D., Park, C. (1986). Nonlinear discriminant functions and associative memories, in *Neural networks for Computing*, J.S. Denker. Ed. New York. American Institute of Physics, 370-375.

[25] Sherrington, C.S. *Man on his nature.* Cambridge University Press, Cambridge, 1941.

[26] Simon, J.C.(1992). Off-Line Cursive Word Recognition. *Proceedings of the IEEE*, **80**(7) 1150-1161.

[27] Sompolinsky, H.(1986). Neural networks with nonlinear synapses and a static noise. Physical Review A **34**, No. 3, p. 2571-2574.

[28] Van Hemmen, J.L. and Kuhn, R. Collective Phenomena in Neural Networks, in *Models of Neural Networks*, Springer, 1992.

[29] Whittle, P.(1989). The antiphon: a device for reliable memory from unreliable components. I. *Proc. R. Soc. Lond.* **A 423** 201-218.

[30] Whittle, P.(1990). The antiphon. II. The exact evaluation of memory capacity. *Proc. R. Soc. Lond.* **A 429** 45-60.

[31] Whittle, P.(1991). Neyman Lecture: Neural Nets and Implicit Inference. *The Annals of Applied Probability.* **1**(2) 173-188.