An Analysis of the Exponential Correlation Associative Memory

Edwin R. Hancock
Department of Computer Science
University of York
YO1 5DD York, England

Marcello Pelillo
Dip. di Matematica Applicata e Informatica
Università "Ca' Foscari" di Venezia
Via Torino 155, 30173 Venezia Mestre, Italy

Abstract

The exponential correlation associative memory (ECAM) is a recurrent neural network model which has large storage capacity and is particularly suited for VLSI hardware implementation. Our aim in this paper is to show how the ECAM model can be entirely derived within a Bayesian framework, thereby providing more insight into the behaviour of this algorithm. The framework for our study is a novel relaxation method which involves direct probabilistic modelling of the pattern corruption mechanism. The parameter of this model is the memoryless probability of error on nodes of the network. This bit-error probability is not only important for the interpretation of the ECAM model, but allows also us to understand some more general properties of Bayesian pattern reconstruction by relaxation. To study the dynamical behaviour of our relaxation model, we use the Hamming distance picture of Kanerva which allows us to understand how the bit-error probability evolves during the relaxation process. We also derive a parameter-free expression for the storage capacity of the model which, like a previous result of Chiueh and Goodman, scales exponentially with the number of nodes in the network.

1 Introduction

Since Hopfield's influential work [1], there has been an explosion of interest in the study of neural network models of associative memory. These are parallel computational networks in which idealised versions of corrupted patterns are recalled by locally updating individual nodes in the network to optimise a global cost function. The original model proposed by Hopfield, however, was soon recognised to have severe limitations associated with its restricted storage capacity [2], and this has then motivated a number of investigators to develop alternative memory models in an attempt to overcome this fundamental drawback. One of the most successful such attempts has recently been reported by Chiueh and Goodman [3] who developed an exponential correlation associative memory (ECAM) model which not only turns out to have large storage capacity, but can also be easily implemented in VLSI hardware.

Viewed as systems that achieve global configurational optimisation via local iterative computations, these associative memories have many features in common with relaxation algorithms widely employed in the field of computer vision and pattern recognition [4, 5]. Our aim in this paper is to demonstrate not only that there is a strong conceptual similarity between the two approaches, but that the ECAM model can be regarded as a particular instance of discrete relaxation [5], providing it with a direct Bayesian interpretation. The analysis provides an empirical understanding of the dynamical properties of this model.

Basic to the methodology described in this paper is the direct modelling of the pattern corruption mechanism which is objective yet simple [5]. We assume that each node in the network is subject to memoryless corruption, and adopt as our criterion for global labelling the joint probability of the binary node configuration on the network. This turns out to be a compound exponential function of the Hamming distances to the idealised stored patterns, the parameter of these exponentials being the memoryless probability of bit-errors acting on the binary patterns. Node updating takes place so as to perform gradient ascent on the global configurational probability. The ECAM, too, implements an exponential update rule, and we will show how our relaxation scheme is identical to the ECAM model when the base of its exponentials is chosen as a certain function of the bit-error probability. This corresponds also to the best possible choice in terms of storage capacity. This analysis complements in fact a recent study by Hancock and Kittler [6], where it has been shown that the Hopfield [1] and the Aleksander [7] memories can be regarded as limits of the presented relaxation approach with precise physical meaning in terms of the bit-error probability.

In order to understand some of the dynamical properties of our discrete relaxation model of associative memory, we use the Hamming distance picture of Kanerva [8]. This model allows us to predict how the bit-error probability evolves as the corrupted patterns undergo iterative associative recall. In particular, the fixed points of the bit-error probability play a crucial role in limiting effective pattern recovery. If the bit-error probability exceeds a certain value termed the "critical distance" by Kanerva, then iterations only result in disruption of memory patterns. Moreover, if the critical distance altogether vanishes then the memory capacity of the network has been exceeded. We use this latter property to show that the storage capacity of our memory is exponential with the number of nodes. This is exactly the finding of Chiueh and Goodman [3]. However, our result for the storage capacity depends only on the number of nodes in the network, i.e., the length of memory patterns. This contrasts with the theoretical result of Chiueh and Goodman which relates the storage capacity to a threshold parameter which is effectively a trade-off between storage capacity and bit-error probability. Furthermore, it is interesting to note that our Bayesian framework naturally leads us to a choice of the exponential base that gives the ultimate upper bound for the asymptotic capacity of an associative memory.

2 ECAM

The ECAM is an instance of a more general associative memory model which Chiueh and Goodman [3] called the recursive correlation associative memory (RCAM). The network is composed of N computational nodes, and at any particular stage of updating each node will be in one of the binary states denoted by $\Omega \equiv \{-1, 1\}$. The particular realisation of the labelling of the node indexed j is denoted by s_j. With this notation, the global state of the network is represented by the configuration of binary values $S = \{s_j \in \Omega, \forall j = 1 \ldots N\}$.

Now, suppose that we have access to a set of training patterns. Typically, these would be configurations of binary labels which we want to recover from an initial inconsistent state of the network. Assume that there are Z such global patterns denoted by $\Lambda^\mu = \{\lambda_j^\mu \in \Omega, \forall j = 1 \ldots N\}$. According to this notation, μ is the pattern index and λ_j^μ is the binary value assigned to the site indexed j by the μ^{th} training pattern.

The dynamical behaviour of an RCAM is governed by the following updating rule

$$s_j = \text{sgn}\left\{ \sum_{\mu=1}^{Z} \lambda_j^\mu \, f\left(\sum_{i=1}^{N} s_i \lambda_i^\mu \right) \right\} \tag{1}$$

for all $j = 1 \ldots N$, where $f(\cdot)$ is an appropriate weighting function. In [3] it was shown that the dynamical system described by equation (1) is asymptotically stable both in the asynchronous and synchronous update modes, provided that the weighting function f is continuous and monotone non-decreasing over the interval $[-N, N]$. This was proved by showing that the system has an associated Liapunov (or "energy") function that governs its dynamical behavior. In addition, Chiueh and Goodman [3] showed how some known associative memory models can be viewed as a special case of the RCAM by just choosing a suitable weighting function. For example, if f is chosen to be the identity function, i.e. $f(x) = x$, then the model becomes essentially identical to the original Hopfield memory [1], with the weights initialised by the Hebb rule. Of particular interest is the case when the weighting function has an exponential form, i.e., $f(x) = a^x$, where a is a predetermined constant greater than unity. This class of memories constitutes the ECAM original model. This was shown to have large storage capacity and turns out to be especially suited for VLSI implementation [3].

3 Bayesian Derivation of the ECAM

We commence our analysis of the ECAM model by considering an arbitrary input pattern S. The basic information relevant to the "consistency" of the configuration S is conveyed by the set of Hamming distances to the prototype patterns Λ^μ, $\mu = 1 \ldots Z$. With the binary node variables defined in the previous section, the Hamming distance to the prototype pattern indexed μ can be written as

$$H_\mu = \frac{1}{2}\left(N - \sum_{i=1}^{N} s_i \lambda_i^\mu \right) . \tag{2}$$

The configurational relaxation procedure is based on maximising the joint probability of the binary label configuration, i.e., $P(S)$. It is therefore necessary to find a way of enumerating $P(S)$ when the label configuration is highly inconsistent, i.e., when there are no prototype patterns for which the Hamming distance is zero. The approach is to adopt a Bayesian viewpoint in which it is assumed that only prototype patterns are legal and have uniform non-zero *a priori* probabilities of occurrence $P(\Lambda^\mu)$. Other configurations do not occur *a priori* but are the corrupted realisations of the prototype patterns. This idea is realised by applying the axiomatic property of joint probability to expand $P(S)$ over the space of consistent configurations

$$P(S) = \sum_{\mu=1}^{Z} P(S|\Lambda^\mu) P(\Lambda^\mu) . \tag{3}$$

To develop this idea into a useful update rule requires a model of the bit corruption process, that is of the conditional probabilities of the inconsistent configurations given the Z possible prototypes $P(S|\Lambda^\mu)$. We adopt a very simple viewpoint: bit errors are assumed to be memoryless and to occur with uniform probability $p \leq \frac{1}{2}$.

The first consequence of the assumed absence of memory is that the errors are independent. As a result we can factorise the conditional probabilities over the individual nodes in the network, i.e.,

$$P(S|\Lambda^\mu) = \prod_{i=1}^{N} P(s_i|\lambda_i^\mu) . \tag{4}$$

Our next step is to propose a model for the bit corruption mechanism at each node in the network. Again, taking recourse to the memoryless assumption, the probability of error on individual nodes is independent of the class of label. This leads us to the following assignment of probability

$$P(s_i|\lambda_i^\mu) = \begin{cases} 1 - p & \text{if } s_i = \lambda_i^\mu \\ p & \text{otherwise} \end{cases} \tag{5}$$

The model components given in (4) and (5) naturally lead to the following expression for $P(S)$ in terms of the set of Hamming distances to the prototype patterns

$$P(S) = \frac{\beta}{Z} \sum_{\mu=1}^{Z} \alpha^{H_\mu} \tag{6}$$

where $\alpha = \frac{p}{1-p}$ and $\beta = (1-p)^N$. We can re-write the above expression for the configurational probability in terms of the more familiar exponential function

$$P(S) = \frac{\beta}{Z} \sum_{\mu=1}^{Z} \exp\left(H_\mu \ln \frac{1}{\alpha}\right) . \tag{7}$$

Our aim in making node updates is to locate an optimum value of the configurational probability $P(S)$. The optimisation procedure is based on a conventional gradient ascent approach. When rather than being binary values the node variables are real-valued quantities in the range $[-1,1]$, the step-size for updating s_j is

$$\frac{\partial P(S)}{\partial s_j} = \frac{\beta}{Z} \ln \alpha \sum_{\mu=1}^{Z} \alpha^{H_\mu} \frac{\partial H_\mu}{\partial s_j} \tag{8}$$

where, from (2), we have $\frac{\partial H_\mu}{\partial s_j} = -\frac{1}{2}\lambda_j^\mu$. In the case of binary values, the update rule for gradient ascent becomes

$$s_j = \mathrm{sgn}\left\{\sum_{\mu=1}^{Z} \lambda_j^\mu \alpha^{H_\mu}\right\} . \tag{9}$$

Now, observe that from (2) this update rule is equivalent to

$$s_j = \mathrm{sgn}\left\{\sum_{\mu=1}^{Z} \lambda_j^\mu \alpha^{-\frac{1}{2}\sum_{i=1}^{N} s_i \lambda_i^\mu}\right\} \tag{10}$$

which is identical to the ECAM update rule defined by (1) when the base of exponentiation a is taken to be $\alpha^{-1/2}$ or, in other words, when $a = \sqrt{\frac{1-p}{p}}$. Note that under the assumption that the error probability p is less than $\frac{1}{2}$, the base constant a is greater than unity as required by the ECAM model. This provides, therefore, a direct Bayesian interpretation for the ECAM model. When the base of the exponentials a is chosen in this way, the ECAM can be viewed as a method for maximising the configurational probability $P(S)$, thereby acting as a type of Bayesian pattern reconstruction process.

It may be of some interest to discuss how our Bayesian framework relates to the storage capacity result derived by Chiueh and Goodman in [3]. At this point our analysis will be rather informal; a more rigorous discussion concerning the storage capacity of the ECAM will be presented in Section 4. Consider an input pattern S that is k bits away from the nearest memory pattern from which S is assumed to be derived through a process of memoryless bit corruption. Let $k = \kappa N$, where κ is the fraction of incorrect bits in the pattern S, and define κ' as $\kappa + \frac{1}{N}$. In [3], it was shown that when $\kappa' \geq (1+a^2)^{-1}$ the asymptotic storage capacity of the ECAM is the largest possible for an associative memory, as determined in [9]. Observe that, under our memoryless corruption assumption, the bit-error probability p represents the expected fraction of incorrect bits in any input pattern, and is therefore intimately related to the parameter κ defined above. Specifically, κ is the actual realisation of a random variable whose expected value is p. Now, from the definition of the base of the ECAM, we obtain $p = (1+a^2)^{-1}$, and thus $p' \equiv p + \frac{1}{N} > (1+a^2)^{-1}$ for all N, which means that Chiueh and Goodman's condition for meeting the largest possible capacity is expected to hold, at least in a statistical sense. We can therefore conclude that our Bayesian framework naturally leads to a choice of the exponentiation base a which is (on the average) the best possible: it allows us to achieve the ultimate upper bound for the asymptotic storage capacity of an associative memory.

4 The Hamming Distance Picture

The ECAM model uses a function of Hamming distance as the basis for node update decisions. To understand the dynamic properties of the memory, it will prove instructive to study how the Hamming distance is distributed for realistic pattern classification problems. The distribution has two components. The first reflects the effect of noise corruption on perfect patterns producing departures from zero Hamming distance. The second is due to the distribution of Hamming distance between the different stored patterns—we refer to this as the inter-pattern Hamming distance.

Consider a noise corrupted pattern indexed ν. If the bit corruption is memoryless, then the Hamming distance to the true uncorrupted prototype of this pattern is equal to the number of bit errors. In this case H_μ follows the binomial distribution

$$P_s(H_\nu = t) = \frac{N!}{(N-t)!t!} p^t (1-p)^{N-t} . \tag{11}$$

To describe the Hamming distribution between the competing stored patterns requires a model of the structure of the pattern space. The basic assumption is that the patterns are random bit patterns. Actual bit values at different configuration sites can be treated as independent events occurring with a uniform probability distribution. According to this model the high and low bits are assigned with probabilities q and $1-q$, respectively. The probability that the bits on the same configuration site of any two patterns disagree is $r = 2q(1-q)$. It is easy to show using the independence and uniformity assumptions that the probability distribution for the inter-pattern Hamming distance is again binomial with mean Nr and variance $Nr(1-r)$

$$P_b(H_\mu = t) = \frac{N!}{(N-t)!t!} r^t (1-r)^{N-t} . \tag{12}$$

In what follows we will confine our attention to the case when the binary labels occur with equal probability, i.e., $r = \frac{1}{2}$.

The two-component model described above allows us to view the ECAM as a decision rule in the measurement domain of Hamming distance. A similar framework for studying the dynamical behaviour of associative memories has already been explored intuitively by Kanerva [8] and has been formalized by Wong and Sherrington [10]. We are interested in using this picture to understand how the bit-error probability evolves with the iterative node update process. The first step towards meeting this aim is to use the signal and background Hamming distributions to compute the probability that any single stored pattern has a Hamming distance smaller than that for the correct pattern which is indexed ν, i.e., $P(H_\mu \leq H_\nu)$. This goal is achieved by performing a discrete convolution of binomial distributions given in (11) and (12)

$$P(H_\mu \leq H_\nu) = \sum_{t=0}^{N} P_s(H_\nu = t) \sum_{u=0}^{t} P_b(H_\mu = u) \ . \quad (13)$$

If we confine our attention to the case when the binary labels occur with equal probability, i.e. $r = \frac{1}{2}$, then with the binomial forms of $P_s(H_\mu = t)$ and $P_b(H_\mu = u)$ we obtain

$$P(H_\mu \leq H_\nu) = \frac{1}{2^N} \sum_{t=0}^{N} \frac{N! p^t (1-p)^{N-t}}{(N-t)! t!} \sum_{u=0}^{t} \frac{N!}{(N-u)! u!} \ . \quad (14)$$

Our aim is to compute the probability of bit-errors after updating has taken place. The condition for error is that at least one of the $Z-1$ competing background patterns produces a configuration with Hamming distance less than H_ν. The background patterns can be regarded as independent events each of which has a uniform probability $P(H_\mu \leq H_\nu)$ of causing an error. These conditions specify a binomial distribution for the number of potential error-producing patterns. As a result, the probability of misclassifying the true pattern as any of the remaining $Z-1$ background patterns, i.e. $P(\min_\mu H_\mu \leq H_\nu)$, is equal to

$$P_{error}(p) \simeq P(\min_{\mu \neq \nu} H_\mu \leq H_\nu) = 1 - (1 - P(H_\mu \leq H_\nu))^{Z-1} \ . \quad (15)$$

Substituting (14) into (15), and applying the Taylor expansion for small values of p we obtain the following expression for the bit-error probability which applies in the vicinity of the origin

$$P_{error}(p) \simeq \frac{Z-1}{2^N} \left[1 + N^2 p + \frac{1}{4} N^2 (N-1)(N-3) p^2 + \ldots \right] \ . \quad (16)$$

The effect of performing configurational relaxation on the network is equivalent to modifying the bit-error probability. This can be viewed as an iterative process of the form $P_{error}(p) \mapsto p$. It is the fixed points of this iteration scheme that determine the behaviour of the relaxation process and the feasibility of pattern recovery. These fixed points are the values of the bit-error probability for which $P_{error}(p) = p$. A detailed numerical analysis of (16) reveals the existence of three such points; the lower one is close to the origin, i.e. $p \simeq 0$, the upper one occurs at $p = 1$ while the third occurs at an intermediate value of p.

An important property of the fixed points of the iterative scheme is whether they are convergent or divergent. Convergent points are those towards which the iterative scheme migrates; divergent points are those from which it migrates. The condition for a fixed point to exhibit the convergent property is

$$\left| \frac{\partial P_{error}(p)}{\partial p} \right|_{p = P_{error}(p)} \leq 1 \ . \quad (17)$$

Numerical analysis reveals that the lower and upper fixed points are both convergent while the intermediate fixed point is divergent. According to Kanerva's terminology, the position of the intermediate fixed point is referred to as the "critical distance." As a concrete example, Figure 1 shows the curve $P_{error}(p)$ superimposed on the line of unit gradient for a configuration of 15 nodes; moving from left to right the individual curves correspond to Z=300, 150, 50, 25, and 5 stored patterns.

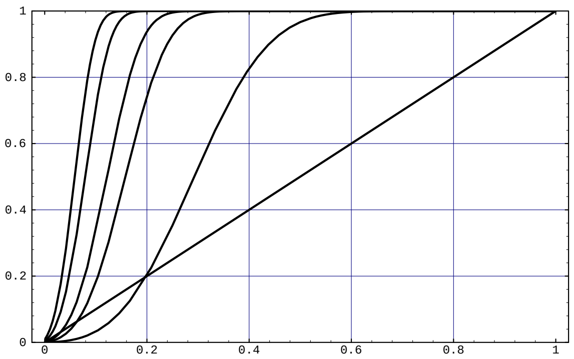

Figure 1: Fixed points of bit-error probability.

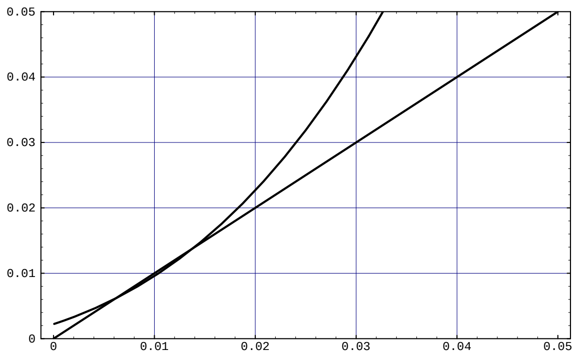

Figure 2: Degeneracy property for a network with 15 nodes. The curve corresponds to the behaviour of $P_{error}(p)$ with $Z = 73$, the theoretical maximum storage capacity for a network with 15 nodes.

The intersections of curve and line are the fixed points of the iterative scheme $P_{error}(p) \mapsto p$. For successful pattern recovery the relaxation scheme must be initialised below the intermediate fixed point in order for the bit-error probability to monotonically decrease until coming to rest at the lower fixed point. In other words, the initial Hamming distance must be less than the critical distance. If this is not the case, then the bit-error probability will grow iteration by iteration, eventually coming to rest at the upper fixed point; this will have catastrophic results on quality of the recovered configuration.

The Kanerva picture of associative recall has been formally used by Wong and Sherrington [10] to compute the storage capacity of Aleksander's memory [7]. The basic idea is to compute the conditions under which the lower and intermediate fixed points become degenerate. Under these circumstances, there is no range of initial Hamming distance for which iterative improvement in the bit-error probability is possible. This degeneracy condition is illustrated in Figure 2. Following Wong and Sherrington [10], the conditions for degeneracy may be determined by approximating the polynomial expression for $P_{error}(p)$ appearing in equation (16) to second-order in p. The fixed points are the two solutions of the quadratic equation

$$1 + \left(N^2 - \frac{2^N}{Z-1} \right) p + \frac{1}{4} N^2 (N-1)(N-3) p^2 = 0 \ . \quad (18)$$

The degeneracy condition for this quadratic gives the following expression for the number of stored patterns Z in terms of the numbers of nodes of the network N

$$Z = 1 + \frac{2^N}{N^2 \pm N\sqrt{(N-1)(N-3)}} \ . \quad (19)$$

In the limit of large N, the storage capacity is approximately equal to $Z \simeq \frac{2^{N-1}}{N^2}$.

It should be noted that our approach to computing storage capacity is somewhat different to that adopted by Chiueh and Goodman [3]. Rather than imposing a threshold on acceptable error-rate, we draw on Kanerva's picture which attempts to understand associative recall in terms of a critical Hamming distance above which successful pattern recovery is impossible. Using a simple model of the pattern-space in which the associative memory operates, we effectively compute the condition for the critical distance to vanish. In consequence, our storage result depends only upon the size of the network. By contrast, Chiueh and Goodman's result depends on the prespecified error threshold. However, both results share the common feature of being exponential in the number of nodes.

5 Conclusions

We have demonstrated how Chiueh and Goodman's ECAM model can be entirely derived within a Bayesian framework, and can therefore be interpreted as performing a type of Bayesian pattern reconstruction process. This expands on a previous study [6] which showed how the Hopfield and the Aleksander memories can effectively be given a Bayesian interpretation. The basis for this work has been a novel relaxation scheme developed in the context of pattern recognition and computer vision, which involves direct probabilistic modelling of the pattern corruption process. Based on a simple model of the pattern space suggested by Kanerva [8], we have theoretically explored the behaviour of the bit-error probability under iterative node updating. This leads us to several conclusions. Firstly, the ECAM iterative scheme has convergent fixed points for values of the bit-error probability close to zero and unity. A further interesting property is that the model has a saturated error which originates from the effect of interpattern competition. Finally, the ECAM approach is completely inoperable when its fixed points become degenerate. This degeneracy condition led us to derive an expression for the storage capacity of the model which turns out to be exponential in the number of nodes in the network. A similar result was also derived by Chiueh and Goodman [3] but, in contrast with theirs, our result does not depend on an adjustable parameter which affects both the bit-error probability and the storage capacity of the model.

References

[1] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Natl. Acad. Sci. USA*, vol. 79, pp. 2554–2558, 1982.

[2] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, "The capacity of the Hopfield associative memory," *IEEE Trans. Inform. Theory*, vol. 33, pp. 461–482, 1987.

[3] T. D. Chiueh and R. M. Goodman, "Recurrent correlation associative memories," *IEEE Trans. Neural Networks*, vol. 2, pp. 275–284, 1991.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, pp. 721–741, 1984.

[5] E. R. Hancock and J. Kittler, "Discrete relaxation," *Pattern Recognition*, vol. 23, pp. 711–733, 1990.

[6] E. R. Hancock and J. Kittler, "A Bayesian interpretation for the Hopfield network," in *Proc. IEEE Int. Conf. Neural Networks*, San Francisco, CA, 1993, pp. 341–346.

[7] I. Aleksander, "The logic of connectionist systems," in I. Aleksander, Ed., *Neural Computing Architectures: The Design of Brain-Like Machines*. London, UK: North Oxford Academic Publishers, 1989, pp. 133–155.

[8] P. Kanerva, *Sparse Distributed Memory*. MIT Press, Cambridge, MA, 1988.

[9] P. A. Chou, "The capacity of the Kanerva associative memory is exponential," in D. Z. Anderson, Ed., *Neural Information Processing Systems*. New York, NY: American Institute of Physics, 1988, pp. 184–191.

[10] K. Y. M. Wong and D. Sherrington, "Theory of associative memory in randomly connected Boolean neural networks," *J. Phys. A*, vol. 22, pp. 2233–2264, 1989.