MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1580 September 6, 1996
C.B.C.L Paper No. 138

LEARNING LINEAR, SPARSE, FACTORIAL CODES

BRUNO A. OLSHAUSEN

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1580.ps.Z

Abstract

In previous work (Olshausen & Field 1996), an algorithm was described for learning linear sparse codes which, when trained on natural images, produces a set of basis functions that are spatially localized, oriented, and bandpass (i.e., wavelet-like). This note shows how the algorithm may be interpreted within a maximum-likelihood framework. Several useful insights emerge from this connection: it makes explicit the relation to statistical independence (i.e., factorial coding), it shows a formal relationship to the algorithm of Bell and Sejnowski (1995), and it suggests how to adapt parameters that were previously fixed.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology. This research is sponsored by an Individual National Research Service Award to B.A.O. (NIMH F32-MH11062) and by a grant from the National Science Foundation under contract ASC-9217041 (this award includes funds from ARPA provided under the HPCC program) to CBCL.

1 Introduction

There has been much interest in recent years in unsupervised learning algorithms for finding efficient representations of data. Among these are algorithms for sparse or minimum entropy coding (Foldiak 1990; Zemel 1993; Olshausen & Field 1996; Harpur & Prager 1996), independent component analysis (Comon 1994; Bell & Sejnowski 1995; Amari et al. 1995; Pearlmutter & Parra 1996), and hierarchical generative modeling (Dayan 1995; Hinton et al. 1995). One finds common threads among many of these techniques, and this note is an attempt to tie some of them together. In particular, I will focus on the sparse coding algorithm of Olshausen and Field (1996) and its relation to maximum-likelihood techniques. As we shall see, forming this link enables one to see a formal relationship to the independent component analysis algorithm of Bell and Sejnowski (1995), which although not originally described in terms of maximum-likelihood may be understood in this light. I shall also show how the algorithm may be cast in terms of mean-field theory techniques in order to obtain a lower bound on the log-likelihood, which shares some similarity to the use of a "recognition distribution" in the Helmholtz machine of Dayan et al. What emerges from this process is a better understanding of the algorithm and how it may be improved.

2 Learning linear sparse codes

In the sparse coding learning algorithm of Olshausen and Field (1996), a set of basis functions, $\phi_i(\vec{x})$, is sought such that when an image, $I(\vec{x})$, is linearly decomposed via these basis functions,

$$I(\vec{x}) = \sum_i a_i \phi_i(\vec{x})\,, \tag{1}$$

the resulting coefficient values, a_i, are rarely active (non-zero). In other words, the probability distribution over the a_i should be unimodal and peaked at zero with heavy tails (positive kurtosis). This is accomplished by constructing an energy function of the form

$$E(I,a|\phi) = \sum_{\vec{x}} \left[I(\vec{x}) - \sum_i a_i \phi_i(\vec{x}) \right]^2 + \lambda \sum_i S(a_i)\,, \tag{2}$$

and then minimizing it with respect to the a_i and ϕ_i. The first term in Equation 2 ensures that information is preserved (i.e., that the ϕ_i span the input space), while the second term incurs a penalty on activity so as to encourage sparseness. The intuition behind the choice of S is that it should favor among activity states with equal variance ($|a|^2$) those with the fewest number of non-zero (or not-close-to-zero) components. The choices experimented with include $|a_i|$, $\log(1+a_i^2)$, and $-e^{a_i^2}$.

Gradient descent on E is performed in two phases, one nested inside the other: For each image presentation, E is minimized with respect to the a_i; the ϕ_i then evolve by gradient descent on E averaged over many image presentations. Stated more formally, we seek a set of basis functions, ϕ^*, such that

$$\phi^* = \arg\min_{\phi} \left\langle \min_a E(I,a|\phi) \right\rangle \tag{3}$$

where $\langle\cdot\rangle$ denotes an ensemble average over the images. Note that in this expression and in the rest that follow, I refers to the vector with components $I(\vec{x}_j)$, a refers to the vector with components a_i, and ϕ refers to the matrix with components $\phi_i(\vec{x}_j)$.

The intuition behind the algorithm is that on each image presentation, the gradient of S "sparsifies" the activity on the a_i by differentially reducing the value of low-activity coefficients more than high-activity coefficients. This weeds out the low-activity units. The ϕ_i then learn on the error induced by this sparsification process. The result is a set of ϕ_i that can tolerate sparsification with minimum mean-square reconstruction error. A virtually identical algorithm was developed independently by Harpur and Prager (1996).

3 Maximum-likelihood framework

While the energy function framework provides a useful, intuitive way of formulating the sparse coding problem, a probabilistic approach could provide a more general framework. Harpur and Prager (1996) point out that the first term on the right-hand side of Equation 2 may be interpreted as the negative log-likelihood of the image given ϕ and a (assuming a gaussian noise model), while the second term may be interpreted as specifying a particular log-prior on a. That is,

$$P(I|a,\phi) = \frac{1}{Z_{\sigma_N}} e^{-\frac{|I-a\phi|^2}{2\sigma_N^2}} \tag{4}$$

$$P(a) = \Pi_i \frac{1}{Z_\beta} e^{-\beta S(a_i)} \tag{5}$$

with $\lambda = 2\sigma_N^2\beta$. Thus, we may interpret E as being proportional to $-\log P(I,a|\phi)$, since

$$P(I,a|\phi) = P(I|a,\phi)\,P(a) \tag{6}$$

$$\propto e^{-\frac{1}{2\sigma_N^2}E(I,a|\phi)} \tag{7}$$

How can we use this insight to improve our understanding of the algorithm?

Under the maximum-likelihood approach, we would try to find the set of basis functions, ϕ^*, such that

$$\phi^* = \arg\max_{\phi} \langle \log P(I|\phi) \rangle \tag{8}$$

$$P(I|\phi) = \int P(I|a,\phi)P(a)da \tag{9}$$

In other words, we are trying to find a set of ϕ_i that maximize the log-likelihood that the set of images could have arisen from a random process in which the ϕ_i are linearly mixed with statistically independent amplitudes distributed according to $\frac{1}{Z_\beta}e^{-\beta S(a_i)}$, with additive gaussian image noise. This is formally equivalent to minimizing the Kullback-Leibler (KL) distance between the actual joint probability of the images, $P^*(I)$, and our

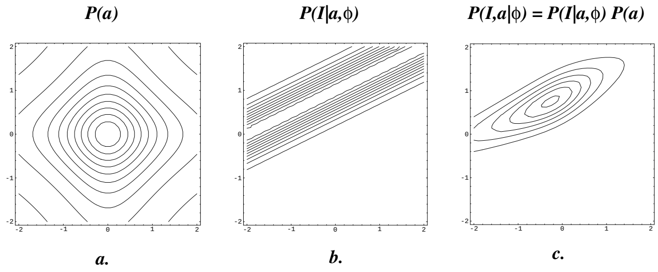

Figure 1: Two-dimensional iso-probability plots of a, Cauchy prior, b, Gaussian likelihood, and c, their product. The axes on each plot are a_1, a_2.

model of the joint probability based on independent causes, $P(I|\phi)$, since

$$\begin{aligned} \mathrm{KL}[P^*(I), P(I|\phi)] &= \int P^*(I) \log \frac{P^*(I)}{P(I|\phi)} dI \quad (10) \\ &= -H_{P^*} - \langle \log P(I|\phi) \rangle \quad (11) \end{aligned}$$

and $H_{P^*} \doteq -\int P^* \log P^*$ is fixed, so maximizing $\langle \log P(I|\phi) \rangle$ minimizes KL.

Unfortunately, all of this is easier said than done because we have to integrate over the entire set of a_i in Equation 9, which is computationally intractable. A reasonable approximation may be to assume that σ_N is small, in which case the dominant contribution to the integral is at the maximum of $P(I, a|\phi)$. Thus,

$$\phi^* \cong \arg\max_{\phi} \left\langle \log[\max_a P(I|a,\phi) P(a)] \right\rangle . \quad (12)$$

This is equivalent to the algorithm of Olshausen and Field (1996), as can be seen by comparing to Equation 3 and using the definitions of Equations 4 and 5. The intuition for why this approximation works in practice is shown in Figure 1. The prior, $P(a)$, is a product of 1-D "sparse" distributions, such as $\frac{1}{1+a_i^2}$, which are unimodal and peaked at zero. The likelihood, $P(I|a,\phi)$, is a multivariate gaussian, and since we are usually working in the overcomplete case (the number of basis functions exceeds the dimensionality of the input) this will take the form of a gaussian ridge (or sandwich) that has its maximum along the line (or plane, etc.) given by $I = a\phi$. The product of these two functions, $P(I|a,\phi)P(a)$, will have its maximum displaced away from the maximum along the gaussian ridge (i.e., away from the "perfect solution") and towards the origin, but also towards the ridges of the prior. Thus, the gradient with respect to ϕ will tend to steer the gaussian ridge towards the ridges of the prior, which will in turn increase the volume of their product, or $P(I|\phi)$. The reason we can get by with this approximation in this case is because we are working with a product of two fairly smooth, unimodal functions. If the functions were not so well behaved, then one can see that such an approximation might produce problems.

4 Relation to Bell and Sejnowski

Bell and Sejnowski (1995) describe an algorithm for "independent component analysis" based on maximizing the mutual information between the inputs and outputs of a neural network. Here, we show that this algorithm may be understood as solving the same maximum-likelihood problem described above (Section 3), except by making a different simplifying assumption. This has also been shown recently by Pearlmutter & Parra (1996) and Mackay (1996).

Bell and Sejnowski examine the case where the number of basis functions is equal to the number of inputs, and where the ϕ_i are linearly independent. In this case, there is a unique set of a_i for which $|I - a\phi|^2$ equals zero for any given image, I. In terms of the previous discussion, $P(I|a,\phi)$ is now a gaussian hump with a single maximum at $a = I\phi^{-1}$, rather than a gaussian ridge as in Figure 1b. If we let σ_N go to zero in Equation 4, then $P(I|a,\phi)$ becomes like a delta function and the integral of Equation 9 becomes

$$\begin{aligned} P(I|\phi) &= \int \delta(I - a\phi) P(a) da \quad (13) \\ &= P(I\phi^{-1}) \times |\det \phi^{-1}| \quad (14) \end{aligned}$$

and so

$$\begin{aligned} \phi^* &= \arg\max_{\phi} \left[\langle \log P(I\phi^{-1}) \rangle + \log|\det \phi^{-1}| \right] \quad (15) \\ &= \arg\min_{\phi} \left[\langle \lambda \sum_i S((\phi^{-1})_i \cdot I) \rangle - \log|\det \phi^{-1}| \right] . \quad (16) \end{aligned}$$

By making the following definitions according to the convention of Bell and Sejnowski (1995),

$$\begin{aligned} \mathbf{W} &= \phi^{-1} \quad (17) \\ u_i &= \mathbf{W}_i \cdot I \quad (18) \end{aligned}$$

then, the gradient descent learning rule for $\mathbf{W}$ becomes

$$\Delta W_{ij} \propto -\lambda S'(u_i) I_j + \frac{\mathrm{cof}\, W_{ij}}{\det \mathbf{W}} . \quad (19)$$

This is precisely Bell and Sejnowski's learning rule when the output non-linearity of their network, $g(x)$, is equal to the cdf (cumulative density function) of the prior on the a_i, i.e.,

$$\begin{aligned} y_i &= g(u_i) \quad (20) \\ g(u_i) &= \int_{-\infty}^{u_i} \frac{1}{Z_\beta} e^{-\beta S(x)} dx . \quad (21) \end{aligned}$$

Thus, the independent component analysis algorithm of Bell and Sejnowski (1995) is formally equivalent to maximum likelihood in the case of no noise and a square system (dimensionality of output = dimensionality of input). It is easy to generalize this to the case when the number of outputs is less than the number of inputs, but not the other way around. When the number of outputs is greater than the effective dimensionality of the input (# of non-zero eigenvalues of the input covariance matrix), then the extra dimensions of the output will simply drop out. While this does not pose a problem for blind separation problems where the number of independent sources (dimensionality of a) is less than or equal to the number of mixed signals (dimensionality of I), it will become a concern in the representation of images, where overcompleteness is a desirable feature (Simoncelli et al., 1992).

5 Lower-bound maximization

A central idea behind the Helmholtz machine of Dayan et al. (1995), as well as the "mean field" theory of Saul et al. (1996), is the construction of an alternative probability distribution, $Q(a|I)$, that is used to obtain a lower-bound on $\log P(I|\phi)$. First, we rewrite $\log P(I|\phi)$ as

$$\log P(I|\phi) = \log \int Q(a|I) \frac{P(I|a,\phi)P(a)}{Q(a|I)} da \qquad (22)$$

Then, as long as Q is a probability (i.e., $\int Q = 1, Q > 0$), we obtain by Jensen's inequality

$$\begin{aligned} \log P(I|\phi) &\geq \int Q(a|I) \log \frac{P(I|a,\phi)P(a)}{Q(a|I)} da \quad (23) \\ &= H_Q - \frac{1}{2\sigma_N^2} \langle E(I,a|\phi) \rangle_Q + \text{const} \quad (24) \end{aligned}$$

where $H_Q \doteq -\int Q(a|I) \log Q(a|I) da$. Thus, if we can construct $Q(a|I)$ so that the integral is tractable, then we can do gradient ascent on a lower bound of $\langle \log P(I|\phi) \rangle$. How good the bound is, though, depends on the Kullback-Leibler distance between $Q(a|I)$ and $P(a|I)$, or in other words, on how closely we can approximate $P(a|I)$ with our tractable choice of Q. Typically, Q is chosen to be factorial, $Q(a|I) = \Pi_i q_i(a_i|I)$, in which case

$$\begin{aligned} H_Q &= \sum_i H_{q_i} \quad (25) \\ \langle E(I,a|\phi) \rangle_Q &= |I - \mu\phi|^2 + \sum_i \sigma_i^2 |\phi_i|^2 + \\ & \quad \lambda \sum_i \int q_i(a_i) S(a_i) da_i \quad (26) \end{aligned}$$

where $\mu_i = \int q_i(a_i) a_i da_i$ and $\sigma_i^2 = \int q_i(a_i)(a_i - \mu_i)^2 da_i$.

Comparing Equation 26 to Equation 2, one can see that the sparse coding learning algorithm of Olshausen and Field (1996) effectively uses $q_i(a_i) = \delta(a_i - \mu_i)$, with μ_i chosen so as to minimize E (and hence maximize the lower bound of Equation 24). This choice would seem suboptimal, though, because we are getting zero entropy out of H_Q (actually $H_Q = -\infty$, but we are ignoring the infinities here because it is the derivatives we really care about). If we could find a q_i with higher entropy which also lowers the energy, then we could move the bound closer to the true log-likelihood. However, broadening q_i (for example, by making it gaussian with adjustable μ_i and σ_i) only affects the solution for μ insofar as it low-pass filters the cost function, S, which has a similar effect to simply lowering λ. So, it is difficult to see that adding this extra complexity will improve matters. One apparent benefit of having non-zero σ_i is that there is now a growth-limiting term on the ϕ_i (second term on the right side of Eq. 26). Without such a term, the ϕ_i will grow without bound, and so it is necessary in the algorithm of Olshausen and Field (1996) to keep the ϕ_i normalized (which is rather *ad hoc* by comparison). Preliminary investigation using a Gaussian q_i and minimizing E with respect to both μ_i and σ_i for each image (but still keeping the ϕ_i normalized) does not reveal significant differences in the solution, but it deserves further study. It may also be worthwhile to try using a $Q(a|I)$ that is defined by pairwise statistics (i.e., a covariance matrix on the a_i).

It should be noted that what is important here is the location of the maximum of whatever approximating function we use, not the absolute value of the bound per se. If the maximum of the lower-bound occurs at a significantly different point than the maximum of the true log-likelihood, then the approximation is not much help to us.

6 Discussion

What I think has been gained from this process is a better understanding of both the sparse coding algorithm of Olshausen and Field (1996) and the independent component analysis algorithm of Bell and Sejnowski (1995). Although neither of these algorithms was originally cast in maximum-likelihood terms, they are both essentially solving the same problem. The main difference between them is in the simplifying assumptions they make in order to deal with the intractable integration problem posed by Equation 9: Olshausen and Field's algorithm assumes low-noise (small σ_N) and thus a peaky, unimodal distribution on $P(I,a|\phi)$ in order to justify evaluating it at the maximum, whereas Bell and Sejnowski limit the dimensionality of the a_i to equal the dimensionality of the input and also assume no noise so that the integral becomes tractable. The maximum-likelihood framework also makes possible the link to techniques used in the Helmholtz machine (Dayan et al., 1995), which reveals that a better choice of approximating distribution, Q, could potentially lead to improvements.

A practical advantage of looking at the problem within this framework is that it suggests we could adapt the shape of the prior. For example, the prior on the a_i need not be i.i.d., but could be shaped differently for each a_i, e.g., $P(a_i) = \frac{1}{Z_{\beta_i}} e^{-\beta_i S(a_i)}$, in order to best fit the data. Adapting β_i would be accomplished by letting it evolve along the gradient of $\langle \log P(I|\phi) \rangle$. Using the approximation of Equation 12, this yields the learning rule:

$$\dot{\beta}_i \propto -\langle S(a_i) \rangle - \frac{1}{Z_{\beta_i}} \frac{\partial Z_{\beta_i}}{\partial \beta_i}. \qquad (27)$$

A problem that may arise here, due to the fact that the full integral in Equation 9 is not being computed, is that there may be a bias toward non-informative flat priors (since these will yield perfect reconstruction on each trial). An advantage of Bell and Sejnowski's algorithm in this case is that it essentially computes the full integral in Equation 9 and so does not have this problem. For their algorithm, the maximum-likelihood framework prescribes a method for adapting the "generalized sigmoid" parameters p and r for shaping the prior (see pp. 1137-8 of their paper), again by doing gradient ascent on the average log-likelihood. (See also Mackay, 1996, for other methods of parameterizing and adapting a factorial prior.) In cases where a statistically independent linear code may not be achieved (e.g., natural images), it may be advantageous to alter the prior so that information about pairwise or higher-order statistical dependencies among the a_i may by incorporated into our model of $P(a)$, for example using a Markov random field type model.

References

Amari S, Cichocki A, Yang HH (1996) A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems, 8*, MIT Press.

Bell AJ, Sejnowski TJ (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*: 1129-1159.

Comon P (1994) Independent component analysis, a new concept? *Signal Processing, 36*: 287-314.

Dayan P, Hinton GE, Neal RM, Zemel RS (1995) The Helmholtz machine. *Neural Computation, 7*: 889-904.

Foldiak P (1990) Forming sparse representations by local anti-Hebbian learning. *Biol. Cybernetics, 64*: 165-170.

Harpur GF, Prager RW (1996) Development of low entropy coding in a recurrent network, *Network, 7*.

Hinton GE, Dayan P, Frey BJ, Neal RM (1995) The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268*: 1158-1161.

Mackay DJC (1996) Maximum likelihood and covariant algorithms for independent component analysis. Available via `ftp://wol.ra.phy.cam.ac.uk/pub/www/mackay/ica.ps.gz`

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*: 607-609.

Pearlmutter BA, Parra LC (1996) A context-sensitive generalization of ICA. International Conference on Neural Information Processing, September 1996, Hong Kong. In press.

Saul LK, Jaakkola T, Jordan MI (1996) Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research, 4*: 61-76.

Simoncelli EP, Freeman WT, Adelson EH, Heeger DJ (1992) "Shiftable multiscale transforms," *IEEE Transactions on Information Theory, 38(2)*: 587-607.

Zemel RS (1993) A minimum description length framework for unsupervised learning. Ph.D. Thesis, University of Toronto, Dept. of Computer Science.

Acknowledgments

This note grew out of discussions with Chris Lee, Peter Dayan, Federico Girosi, Max Riesenhuber, Tony Bell, George Harpur, Mike Lewicki, and Dan Ruderman over the past several months. I thank Tommy Poggio for making possible the visit to MIT which fostered many of these interactions.