Communicated by Radford Neal

Statistical Physics, Mixtures of Distributions, and the EM Algorithm

Alan L. Yuille
*Division of Applied Sciences, Harvard University,
Cambridge, MA 02138 USA*

Paul Stolorz
*Jet Propulsion Laboratory, MS 198-219, Pasadena, CA 91109 and
Santa Fe Institute, Santa Fe, NM 87501 USA*

Joachim Utans
*International Computer Science Institute,
1947 Center Street, Suite 600, Berkeley, CA 94704 USA*

We show that there are strong relationships between approaches to optmization and learning based on statistical physics or mixtures of experts. In particular, the EM algorithm can be interpreted as converging either to a local maximum of the mixtures model or to a saddle point solution to the statistical physics system. An advantage of the statistical physics approach is that it naturally gives rise to a heuristic continuation method, deterministic annealing, for finding good solutions.

In recent years there has been considerable interest in formulating optimization problems in terms of statistical physics. This has led to the development of powerful optimization algorithms, such as deterministic annealing.

At the same time good results have been attained by formulating learning theory in terms of mixtures of distributions (Jacobs *et al.* 1991) and using the EM algorithm (Jordan and Jacobs 1993).

The aim of this note is to show that there are close connections between the mixture of distributions and the statistical physics approaches. The EM algorithm can be, and has been, used in conjunction with deterministic annealing. This equivalence has previously been mentioned for some specific cases (Yuille *et al.* 1991; Stolorz 1991; Utans 1993) but, to our knowledge, its generality does not seem to be widely appreciated. We will demonstrate these equivalences by examining the elastic net algorithm for the Traveling Salesman Problem (TSP). Then we will discuss the generalization to other problems.

The elastic net (Durbin and Willshaw 1987) attempts to fit an elastic net, consisting of cities $\{\mathbf{y}_j : j = 1, \ldots, N\}$ joined together by elastic strings, to a set of cities $\{\mathbf{x}_\mu : \mu = 1, \ldots, M\}$ where $N \geq M$. The intuition is that the elastic forces will cause the net to find the shortest possible tour. This corresponds to minimizing an energy function:

$$E_{\text{eff}}[\{\mathbf{y}_j\}; \beta] = \frac{-1}{\beta} \sum_{\mu=1}^{M} \log \left\{ \sum_{j=1}^{N} e^{-\beta |\mathbf{x}_\mu - \mathbf{y}_j|^2} \right\} + \gamma \sum_{k=1}^{N} |\mathbf{y}_k - \mathbf{y}_{k+1}|^2 \tag{1}$$

where the net is circular so that $\mathbf{y}_{N+1} = \mathbf{y}_1$.

Here β is a parameter that characterizes the inverse scale. The idea is to minimize $E_{\text{eff}}[\{\mathbf{y}_j\}; \beta]$ at large scale, small β, and then track the solution as β increases.

It was shown (Durbin *et al.* 1989) that this could be interpreted in a Bayesian framework. We can write the Gibbs distribution $P(Y \mid X) = (1/Z) \exp\{-\beta E[Y]\}$ (where Z is a normalization constant) and express this in terms of Bayes' formula as

$$P(Y \mid X) = \frac{P(X \mid Y) P(Y)}{P(X)} \tag{2}$$

We interpret $P(Y) = (1/Z_1) \prod_i e^{-\beta\gamma |\mathbf{y}_i - \mathbf{y}_{i+1}|^2}$ as the prior (conditional) probability for the tour (Z_1 is a normalization constant). The distribution

$$P(X \mid Y) = (1/Z_2) \prod_{\mu} \left\{ \sum_j e^{-\beta |\mathbf{x}_\mu - \mathbf{y}_j|^2} \right\} \tag{3}$$

corresponds to the product of a mixture of gaussian distribution (Z_2 is a normalization constant). More specifically, each data point $\mathbf{x}_\mu$ is assumed to be produced by a mixture of gaussians centered on the points $\{\mathbf{y}_j\}$. Thus the elastic net corresponds to a mixture model of data generation combined with a prior model.

It was then shown (Simic 1990; Yuille 1990) that the elastic net could be derived from a statistical physics system as a saddle point approximation. The derivation in Yuille (1990) started from an energy

$$E[V, Y] = \sum_{\mu j} V_{\mu j} \left| \mathbf{x}_\mu - \mathbf{y}_j \right|^2 + \gamma \sum_i |\mathbf{y}_{i+1} - \mathbf{y}_i|^2 \tag{4}$$

where the $\{V_{\mu j}\}$ are binary (0,1) variables which obey the constraint $\sum_j V_{\mu j} = 1, \ \forall \mu$.

The partition function of the corresponding Gibbs distribution can be written as $Z = \sum_V \int [d\mathbf{y}] e^{-\beta E[V, \mathbf{y}]}$. The sum over the V variables can be done explicitly (Yuille 1990) while imposing the constraints, to yield $Z = \int [d\mathbf{y}] e^{-\beta E_{\text{eff}}[\mathbf{y}; \beta]}$, where $E_{\text{eff}}[\mathbf{y}; \beta]$ is given by equation 1. By an identical calculation we can compute the marginal distribution $P_M(\mathbf{y}) = \sum_V P(V, \mathbf{y})$, where $P(\mathbf{y})$ is a Gibbs distribution with energy $E_{\text{eff}}[\mathbf{y};\beta]$, corresponding to the mixture distribution (equation 3).

Thus extremizing $E_{\text{eff}}[\mathbf{y};\beta]$ corresponds to performing a saddle point approximation[1] to the partition function and hence to finding the mean field approximation to the system (Amit 1989). It can also be considered to be maximizing $P_M(\mathbf{y})$. This equivalence between finding the saddle point approximation and maximizing P_M is the key reason why the statistical physics and mixture approaches correspond.

It is important to emphasize that there are many possible algorithms for attempting to minimize $E_{\text{eff}}[\mathbf{y};\beta]$. The original algorithm (Durbin and Willshaw 1987) was a discretized version of steepest descent:

$$\frac{d\mathbf{y}_i}{dt} = -\frac{\partial E_{\text{eff}}}{\partial \mathbf{y}_i} = -2\sum_{\mu} \frac{e^{-\beta|\mathbf{x}_\mu - \mathbf{y}_i|^2}}{\sum_k e^{-\beta|\mathbf{x}_\mu - \mathbf{y}_k|^2}}(\mathbf{y}_i - \mathbf{x}_\mu) - 2\gamma\{2\mathbf{y}_i - \mathbf{y}_{i+1} - \mathbf{y}_{i-1}\} \quad (5)$$

To obtain the algorithm used in Durbin and Willshaw (1987) approximate $d\mathbf{y}/dt$ by $[\mathbf{y}(t+1) - \mathbf{y}(t)]/K$, where K is a constant, and set $\mathbf{y} = \mathbf{y}(t)$ in the right-hand side of equation 5.

However, the EM algorithm has also been successfully applied to find the saddle point solutions (Durbin, private communication) with results reported in Peterson (1990). An EM algorithm assumes there are two types of parameters, in this case the V and the $\mathbf{y}$. An E-step estimates the V with the $\mathbf{y}$ fixed. An M-step maximizes to find the $\mathbf{y}$ with the V fixed. The E-step and the M-step alternate until convergence. (Observe that the M-step finds a single value for y while the E-step finds an expectation over a distribution of values for V. The EM algorithm finds a probability distribution for the V but, because they are binary variables, this is equivalent to finding their mean values.) For the TSP the E-step is

$$\overline{V}_{\mu j} = \sum_{V} V_{\mu j} P(V,\mathbf{y}) = \frac{e^{-\beta|\mathbf{x}_\mu - \mathbf{y}_j|^2}}{\sum_k e^{-\beta|\mathbf{x}_\mu - \mathbf{y}_k|^2}} \quad (6)$$

and the the M-step corresponds to maximizing $P(V,\mathbf{y})$ with V given by $\overline{V}$. This is equivalent to solving the linear equations for $\mathbf{y}$:

$$\sum_{\mu} \overline{V}_{\mu j}(\mathbf{y}_j - \mathbf{x}_\mu) + \gamma\{2\mathbf{y}_j - \mathbf{y}_{j+1} - \mathbf{y}_{j-1}\} = 0, \quad \forall j \quad (7)$$

These can be solved by a variety of algorithms including the dynamic system:

$$\frac{d\mathbf{y}_j}{dt} = -\sum_{\mu} \overline{V}_{\mu j}(\mathbf{y}_j - \mathbf{x}_\mu) - \gamma\{2\mathbf{y}_j - \mathbf{y}_{j+1} - \mathbf{y}_{j-1}\}. \quad \forall j \quad (8)$$

[1] This corresponds to approximating the integral for Z by the maximum of the integrand. The more peaked the integrand, as $\beta \mapsto \infty$, the better the approximation.

Observe that the previous steepest descent algorithm can be written, using equation 5, as

$$\frac{d\mathbf{y}_j}{dt} = -2\sum_{\mu} \overline{V}_{\mu j}(\mathbf{y}_j - \mathbf{x}_\mu) - 2\gamma\{2\mathbf{y}_j - \mathbf{y}_{j+1} - \mathbf{y}_{j-1}\}, \; \forall j \tag{9}$$

Thus the only difference between EM (equations 6, 8) and steepest descent (equation 5) is that for EM the $\overline{V}$ and $\mathbf{y}$ are estimated in turn, while for steepest descent they are estimated together. Both algorithms will converge to a local mimimum of $E_{\text{eff}}[\{\mathbf{y}_j\} : \beta]$. It appears that, at fixed temperature, EM converges faster than steepest descent (Durbin, private communication) probably because, for this specific energy function, the E-step can be computed directly (see equation 6), and the M-step corresponds to solving linear equations (see equation 7). But the quality of the results on the TSP (Peterson 1990) decreased badly as the annealing schedule was increased, demonstrating that EM was effective only when used in conjunction with annealing.

In summary, we can regard the elastic net as two types of system: (1) a mixture of distributions that can be solved, at fixed β, by an EM algorithm, or (2) a statistical physics system whose mean fields can be estimated by a variety of algorithms including steepest descent and EM. In both cases deterministic annealing requires that the solutions are found at low β and then tracked as β increases. This continuation method is a heuristic technique for finding the global minimum of the effective energy. By contrast the EM algorithm applied at fixed β is guaranteed to find only a local minimum.

The basic ideas here are straightforward to generalize. A problem posed in terms of a mixture of distributions can be reformulated as a statistical physics problem and vice versa. An EM algorithm can be applied and can be thought of as either a way to obtain a maximum a posteriori estimate of the mixture distribution or as a solution to the saddle point equations, the mean field equations, for the statistical physics system. In addition there is a simple relationship between a steepest descent algorithm to estimate the mean fields and the EM algorithm.

Thus the EM algorithm should not be thought of as a rival to deterministic annealing. It is simply one way to solve the mean field equations. The key idea of deterministic annealing, which takes it beyond EM, is the continuation strategy of finding the solution at small β and tracking it as β increases.

We now briefly discuss how these results can be extended to other problems such as learning/adaptive experts (Jacobs *et al.* 1991). For a general mixtures model one assumes that the data $\{x_\mu\}$ are generated by a mixture of distributions:

$$P(\{x_\mu\}|\{\alpha_i\}) = \prod_{\mu}\sum_{i} a_i P_i(x_\mu|\alpha_i) \tag{10}$$

where the set $\{a_i\}$ consists of nonnegative numbers such that $\sum_i a_i = 1$,[2] and the set $\{\alpha_i\}$ characterizes the (continuous) parameters of the distributions $\{P_i\}$. For example, we might let P_i be a gaussian with parameters $\alpha_i = (\mu, \sigma)$.

In practice, we are interested in determining the parameters $\{\alpha_i\}$ from the data. This can be done by applying Bayes' theorem to obtain

$$P(\{\alpha_i\}|\{x_\mu\}) = P_p(\alpha) \prod_\mu \frac{P(x_\mu|\{\alpha_i\})}{P(x)} \tag{11}$$

where $P_p(\alpha)$ is the prior probability of the parameters and $P(x)$ is a normalization constant. The $\{\alpha_i\}$ can be chosen by an estimator for this distribution, for example, the *maximum a posteriori* estimator $\{\alpha_i^*\} = \arg\max_{\{\alpha_i\}} P(\{\alpha_i\} \mid \{x_\mu\})$.

The nature of x will depend on the problem being modeled: (1) for supervised learning it is an input–output training pair, (2) for unsupervised learning it is an input, (3) it is the data for an optimization problem, for example, see our discussion of the TSP, and (4) for a signal processing or vision problem it is some processed version of the input signal or image. Observe that the V are interpreted differently for these cases.

To formulate this as a mixtures problem, or in terms of statistical physics, we introduce binary decision variables $\{V_{\mu i}\}$, as for the TSP, with $\sum_i V_{\mu i} = 1$, $\forall\ \mu$. This gives rise to $P(V, \alpha \mid \mathbf{x}) = e^{-\beta E(V,\alpha)}/Z$ where

$$E(V, \alpha) = -\frac{1}{\beta} \sum_{\mu i} V_{\mu i}\{\log P_i(x_\mu \mid \alpha_i) + \log a_i\} - \frac{1}{\beta} \log P_p(\alpha) \tag{12}$$

As before we can define an EM algorithm for V and α. The E-step gives

$$\overline{V}_{\mu i} = \frac{e^{\log P_i(x_\mu|\alpha_i) + \log a_i}}{\sum_j \{e^{\log P_j(x_\mu|\alpha_j) + \log a_j}\}},\ \forall\ \mu, i \tag{13}$$

and the M-step corresponds to solving

$$\sum_\mu \overline{V}_{\mu i} \frac{d \log P_i(x_\mu \mid \alpha_i)}{d\alpha_i} + \frac{d \log P_p(\alpha)}{d\alpha_i} = 0,\ \forall\ i \tag{14}$$

This will converge to a local maximum of $P(\alpha \mid x)$ or equivalently to a solution of the saddle point equations of the statistical physics system. Deterministic annealing will give a heuristic continuation method for solving these equations which, in general, will be preferable to using EM at fixed β.

While completing this work we learnt of an interesting result by Neal and Hinton (1993; see also Hathaway 1986), which states that both the E- and the M-steps of the EM algorithm can be interpreted as minimizing an effective energy, or equivalently as maximizing the associated Gibbs

[2] These can be considered as hyperpriors for the models.

distribution. To understand this result from our perspective, observe that we can use saddle point techniques (see, for example, Simic 1990) to obtain an effective energy for the TSP energy without eliminating the V variables. This gives

$$\begin{aligned} E_{\text{eff}}[\bar{V}, \mathbf{y}] &= \sum_{\mu j} \bar{V}_{\mu j} \left| \mathbf{x}_\mu - \mathbf{y}_j \right|^2 + \lambda \sum_k |\mathbf{y}_{k+1} - \mathbf{y}_k|^2 \\ &+ \sum_\mu Q_\mu \left(\sum_j \bar{V}_{\mu j} - 1 \right) + T \sum_{\mu j} \bar{V}_{\mu j} \log \bar{V}_{\mu j} \end{aligned}$$

where the $\{Q_\mu\}$ are Lagrange multipliers and $\{\bar{V}_{\mu j}\}$ correspond to the expected value of the $\{V_{\mu j}\}$. Minimizing $E_{\text{eff}}[\bar{V}, \mathbf{y}]$ with respect to either $\bar{V}$ or $\mathbf{y}$ (keeping the other fixed) will yield the E- and the M-steps. (Observe that minimizing $E_{\text{eff}}[\bar{V}, \mathbf{y}]$ with respect to $\bar{V}$, solving for $\bar{V}(\mathbf{y})$, and substutiting back gives the effective energy $E_{\text{eff}}[\mathbf{y}]$ of equation 1.)

Acknowledgments

We would like to thank Eric Mjolsness and Anand Rangarajan for helpful conversations and encouragement. One of us (A.L.Y.) thanks DARPA and the Air Force for support under contract F49620-92-J-0466 and Geoffrey Hinton for a helpful conversation.

References

Amit, D. J. 1989. *Modeling Brain Function*. Cambridge University Press, Cambridge, England.

Durbin, R., and Willshaw, D. 1987. An analog approach to the travelling salesman problem using an elastic net method. *Nature (London)* **326**, 689–691.

Durbin, R., Szeliski, R., and Yuille, A. L. 1989. An analysis of the elastic net approach to the travelling salesman problem. *Neural Comp.* **1**, 348–358.

Hathaway, R. J. 1986. Another interpretation of the EM algorithm for mixture distributions. *Stat. Prob. Lett.* **4**, 53–56.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural Comp.* **3**, 79–87.

Jordan, M. I., and Jacobs, R. A. 1993. Hierarchical mixtures of experts and the EM algorithm. MIT Dept. of Brain and Cognitive Science preprint.

Neal, R. M., and Hinton, G. E. 1993. A new view of the EM algorithm that justifies incremental and other variants. *Biometrica*, submitted.

Peterson, C. 1990. Parallel distributed approaches to combinatorial optimization. *Neural Comp.* **2**, 261.

Simic, P. 1990. Statistical mechanics as the underlying theory of "elastic" and "neural" optimization. *NETWORK: Comp. Neural Syst.* **I**(1), 1–15.

Stolorz, P. 1991. Abusing statistical mechanics to do adaptive learning and combinatorial optimization. Los Alamos preprint.

Utans, J. 1993. Mixture models and the EM algorithm for object recognition within compositional hierarchies. Part 1: Recognition. TR-93-004, ICSI, 1947 Center St., Berkeley, CA 94704.

Yuille, A. L. 1990. Generalized deformable models, statistical physics and matching problems. *Neural Comp.* **2**, 1–24.

Yuille, A. L., Peterson, C., and Honda, K. 1991. Deformable templates, robust statistics, and Hough transforms. *Proceedings SPIE Geometric Methods in Computer Vision*, San Diego, CA.

Received March 29, 1993; accepted July 15, 1993.