The Latent Maximum Entropy Principle

Shaojun Wang*‡ Ronald Rosenfeld* Yunxin Zhao† Dale Schuurmans‡

*School of Computer Science, Carnegie Mellon University, USA

†Department of Computer Engineering and Computer Science, University of Missouri at Columbia, USA

‡Department of Computer Science, University of Waterloo, Canada

Abstract

In this paper, we present an extension of Jaynes' maximum entropy principle to handle latent variables. We use an EM algorithm that incorporates nested iterative scaling to calculate maximum entropy solutions for this principle, and give a proof of its convergence.

1 Introduction

Learning about the world requires a system to extract useful sensory features and then form a model for how they interact, perhaps by using abstract concepts. The maximum entropy (ME) principle [8] is an effective method for combining sources of evidence from complex but structured natural systems that has wide application in science, engineering, and economics [6, 7]. However, many of the natural patterns that we wish to classify are the result of causal processes that have hidden hierarchical structure; yielding data that does not report the value of latent variables. For example, in natural lanugage data, the hidden semantic information or syntactic structure is rarely given explicitly [9]. Unfortunately, the standard ME principle only models observed data and ignores possible latent variable structure. In this paper, we propose a latent maximum entropy principle (LME) that explicitly handles latent variables and thereby extends Jaynes' original ME principle. We first formulate the problem so that latent variables are encoded in the model, and then propose an EM algorithm that incorporates nested iterative scaling to solve the problem. The monotone convergence property of this algorithm is then proved.

2 Motivation

In 1957, Jaynes [8] proposed the maximum entropy (ME) principle for statistical inference, which states that data should be summarized by a model that is maximally noncommittal with respect to missing information. That is, if one must estimate a probability distribution on the data source which satisfies known constraints, then, of those distributions that are consistent with the constraints, one should choose the distribution that has maximum entropy. A simple example can be given for a probabilistic source with a single real variable:

Example: Assume Y corresponds to people's heights. Given data $\tilde{Y} = (y_1, ..., y_T)$, one might trust simple statistics such as the sample mean and sample mean square of Y. By Jaynes' ME principle, we should maximize the entropy of Y subject to $EY = a_1$ and $EY^2 = a_2$, where $a_1 = \frac{1}{T}\sum_{t=1}^{T} y_t$ and $a_2 = \frac{1}{T}\sum_{t=1}^{T} y_t^2$. In this case, the ME principle says that one should estimate the probability density $p(Y)$ as a Gaussian with mean a_1 and variance $a_2 - a_1^2$, $N(a_1, a_2 - a_1^2)$. This is a consequence of the well-known fact that a Gaussian random variable has the largest differential entropy of any random variable for a specified mean and variance [2].

However, assume that after observing the data histogram, we find that there are two peaks in the empirical data. Obviously the standard ME solution is not the best suited for such bi-modal data. The cause of two peaks in this case could be due to the fact that there are two data clusters, namely male and female, and this information is missing from the measured data. How can such latent knowledge be incorporated into the ME framework? One way is to explicitly re-label the measurements. That is, let $X = (Y, C)$, where Y denotes people's height and C indicates male or female. Then the labeled measurements are $(y_1, c_1, ..., y_T, c_T)$. Define $N_i = \sum_{t=1}^{T} \delta(c_t, i)$, $p(c_t = i) = \frac{N_i}{T}$, and $p(y_t | c_t = i) = \frac{1}{N_i}$, for $i = 1, 2$. Then we can formulate the ME problem as

$$\max_{p(X)} \; H(X) = H(C) + H(Y|C) \quad \text{subject to} \quad \int_X p(X) Y \delta(C, i) \mu(dX) = \sum_{c,y} \tilde{p}(c)\tilde{p}(y|c)\tilde{Y}\delta(C, i)$$

$$\text{and} \quad \int_X p(X) Y^2 \delta(C, i) \mu(dX) = \sum_{c,y} \tilde{p}(c)\tilde{p}(y|c)\tilde{Y}^2\delta(C, i)$$

where $\delta(C,i)$ is the indicator function for $i=1,2$. In this fully observed data case where we witness the gender label C we have a unique solution: $p(c)=\frac{N_c}{T}$ and $p(y|c)=N(\mu_c,\sigma_c^2)$ for $c=1,2$.

However, in most situations, obtaining fully labelled data is tedious or impossible. In such cases, Jaynes' ME principle, which is maximally noncommittal with respect to missing information, becomes insufficient. To cope with missing but non-arbitrary hidden structure, we must extend the ME principle to account for the underlying causal structure in our data model.

3 A Training Algorithm for LME

Let $X\in\mathcal{X}$ be the complete data and let $Y\in\mathcal{Y}$ be the observed incomplete data. Thus $Y=Y(X)$ is a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. For example, Y might be observed natural language in the form of text, and X might be the text along with its missing syntactic and semantic information. If we let $p(X)$ and $p(Y)$ denote the densities of X and Y respectively, and let $p(X|Y)$ denote the conditional density of X given Y, then $p(Y)=\int_{\mathcal{X}(Y)}p(X)\mu(dX)$, where $\mathcal{X}(Y)=\{X: X\in\mathcal{X}, Y(X)=Y\}$ and $p(X)=p(Y)p(X|Y)$. In this paper, μ denotes a given σ-finite measure on $\mathcal{X}$. If $\mathcal{X}$ is finite or countably infinite, then μ is the counting measure, and integrals reduce to sums. If $\mathcal{X}$ is a subset of a finite dimensional space, μ is the Lebesgue measure. If $\mathcal{X}$ is a combination of both cases, μ will be a combination of both measures.

The problem of maximizing entropy while respecting a given latent variable structure is to select a model $p_*\subseteq\mathcal{P}$ from a set of allowed probability distributions to maximize the entropy

$$H(p)=-\int_{x\in\mathcal{X}}p(x)\log p(x)\mu(dx) \tag{1}$$

subject to

$$\int_{x\in\mathcal{X}}p(x)f_i(x)\mu(dx)=\sum_{y\in\mathcal{Y}}\tilde{p}(y)\int_{x\in X(Y)}p(x|Y=y)f_i(x)\mu(dx),\quad i=1...N \tag{2}$$

Here $\tilde{p}(y)$ is the empirical distribution of a set of observable training samples $y_1,...,y_C$, and is thus given by $\tilde{p}(y)=\frac{C(y)}{C}$; $C(y)=\sum_{i=1}^{C}\delta(y,y_i)$ is the occurrence count of y among the training samples; and $p(x|Y=y)$ gives the hierarchical dependency structure of the model. The features $f_i(X)$, $i=1...N$, encode the constraints that we would like to match from the data (e.g. corresponding to the sufficient statistics in exponential models). That is, we require that the expectations of $f_i(X)$ in the complete model match their empirical expectations on the incomplete data Y, taking into account the structure of the dependence of X on Y. Note that if there is no missing data, i.e. $X=Y$, then the problem is reduced to Jaynes' model where the constraints are given by $\int_{y\in\mathcal{Y}}p(y)f_j(y)\mu(dy)=\sum_{y\in\mathcal{Y}}\tilde{p}(Y=y)f_j(Y=y)$. In this sense, the principle (2) is more general than ME.

Note that due to the nonlinear mapping $p(x|Y=y)$, Eqs. (2) are nonlinear constraints of $p(x)$ and the feasible set is no longer convex. Therefore, even though the objective function (1) is concave, no unique maximum can be guaranteed to exist. In fact, minima and saddle points may exist. Nevertheless, we proceed to derive an iterative training procedure that finds approximate local solutions to the LME problem.

Define the Lagrangian $\Lambda(p,\lambda)$ by

$$\Lambda(p,\lambda)=H(p)+\sum_{i=1}^{N}\lambda_i\left[\int_{x\in\mathcal{X}}p(x)f_i(x)\mu(dx)-\sum_{y\in\mathcal{Y}}\tilde{p}(y)\int_{x\in X(Y)}p(x|Y=y)f_i(x)\mu(dx)\right] \tag{3}$$

Then the optimization procedure can be carried out by holding λ fixed, and computing the unconstrained maximum of the Lagrangian over $p\in\mathcal{P}$. Let

$$p_\lambda = \arg\max_{p\in\mathcal{P}}\Lambda(p,\lambda) \tag{4}$$

$$\Psi(\lambda) = \Lambda(p_\lambda,\lambda) \tag{5}$$

We call $\Psi(\lambda)$ the dual function. The functions p_λ and $\Psi(\lambda)$ may be calculated explicitly using simple variational calculus. First, take the derivative of $\Lambda(p,\lambda)$ with respect to p, to obtain

$$\frac{\partial\Lambda(p,\lambda)}{\partial p(x)} = -\log p(x)-1+\sum_{i=1}^{N}\left\{\lambda_i\left[f_i(x)-\tilde{p}(y)\left(\frac{f_i(x)}{p(y)}-\frac{\int_{x'\in X(Y)}p(x')f_i(x')\mu(dx')}{(\int_{x''\in X(Y)}p(x'')^2\mu(dx''))}\right)\right]\right\}$$

$$= \quad -\log p(x) - 1 + \sum_{i=1}^{N} \lambda_i f_i(x) + \sum_{i=1}^{N} \lambda_i \tilde{p}(y) \frac{\int_{x' \in X(Y)} p(x')[f_i(x') - f_i(x)]\mu(dx')}{p(y)^2}$$

Ignoring the last term of the above equation and setting the remainder to zero, we find

$$p_\lambda(x) \quad \approx \quad Z_\lambda^{-1} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(x)\right) \tag{6}$$

where $Z_\lambda = \int_{x \in \mathcal{X}} e^{\sum_{i=1}^{N} \lambda_i f_i(x)} \mu(dx)$ is a normalizing constant that ensures $\int_{x \in \mathcal{X}} p_\lambda(x)\mu(dx) = 1$. Thus

$$\Psi(\lambda) \quad \approx \quad \log(Z_\lambda) - \sum_{i=1}^{N} \lambda_i \left(\sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{x \in X(Y)} p_\lambda(x|Y = y) f_i(x)\mu(dx)\right) \tag{7}$$

Finally, we pose the unconstrained *dual* problem:

$$\text{Find } \lambda^* = \arg\min_{\lambda} \Psi(\lambda) \tag{8}$$

To solve this problem, we restrict p_λ to have the form of a log-linear model with incomplete data, $p_\lambda(x) = Z_\lambda^{-1} \exp(\sum_{i=1}^{N} \lambda_i f_i(x))$, since this is the approximate form derived in (6). In this case, $p_\lambda(y) = \int_{x \in X(Y)} p_\lambda(x)\mu(dx)$ and the loglikelihood function for the observed incomplete data will be

$$L(\lambda) \; = \; \log \prod_y p_\lambda(y)^{\tilde{p}(y)} \; = \; \sum_y \tilde{p}(y) \log p_\lambda(y) \tag{9}$$

Taking the derivative of $L(\lambda)$ with respect to λ_i, $i = 1...N$, we obtain

$$\begin{aligned}
\frac{\partial L(\lambda)}{\partial \lambda_i} \; &= \; \sum_y \tilde{p}(y) \frac{1}{p_\lambda(y)} \int_{x \in X(Y)} \left(-\frac{1}{Z_\lambda^{-2}} \int_{x \in \mathcal{X}} f_i(x) e^{\sum_{i=1}^{N} \lambda_i f_i(x)} \mu(dx)\right) e^{\sum_{i=1}^{N} \lambda_i f_i(x)} \mu(dx) \\
&\quad + \sum_y \tilde{p}(y) \int_{X \in X(Y)} \frac{\frac{1}{Z_\lambda^{-1}} e^{\sum_{i=1}^{N} \lambda_i f_i(x)}}{p_\lambda(y)} f_i(X)\mu(dx) \\
&= \; -\int_x p_\lambda(x) f_i(x)\mu(dx) + \sum_{y \in Y} \tilde{p}(y) \int_{x \in X(Y)} p_\lambda(x|Y = y) f_i(x)\mu(dx)
\end{aligned}$$

By setting the above derivative to zero, we obtain Eqs. (2). This shows that Eqs. (2) are the necessary conditions for the stationary points of the loglikelihood function. However, unlike the complete data case, $L(\lambda) \neq \Lambda(p, \lambda)$.

We now resort to the EM algorithm [5] to solve the problem of maximizing the likelihood of the observed incomplete data (9). Decompose $L(\lambda)$ into

$$L(\lambda) \; = \; \sum_{y \in \mathcal{Y}} \tilde{p}(y) \log p_\lambda(y) \; = \; Q(\lambda, \lambda') - K(\lambda, \lambda') \tag{10}$$

where $Q(\lambda, \lambda') = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{x \in X(y)} p_{\lambda'}(x|y) \log p_\lambda(x)\mu(dx)$ is the conditional expected complete-data loglikelihood, and $K(\lambda, \lambda') = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{x \in X(y)} p_{\lambda'}(x|y) \log p_\lambda(x|y)\mu(dx)$ is the conditional expected missing-data loglikelihood. The EM algorithm maximizes $L(\lambda)$ by iteratively maximizing $Q(\lambda, \lambda')$ over λ. The jth iteration $\lambda^{(j)} \to \lambda^{(j+1)}$ of the EM algorithm is defined by an expectation step E, which computes $Q(\lambda, \lambda^{(j)})$ as a function of λ, followed by a maximization step M, which finds $\lambda = \lambda^{(j+1)}$ to maximize $Q(\lambda, \lambda^{(j)})$. Each iteration of EM increases $L(\lambda)$, and very generally, if EM converges to λ^*, λ^* is a local maximum of $L(\lambda)$ [5, 10].

For this particular log-linear model, we have

$$Q(\lambda, \lambda^{(j)}) = \sum_{y \in \mathcal{Y}} \tilde{p}(y) \int_{x \in X(Y)} p_{\lambda^{(j)}}(x|Y = y) \log p_\lambda(x)\mu(dx) \tag{11}$$

Surprisingly, maximizing $Q(\lambda, \lambda^{(j)})$ is equivalent to maximizing the dual function of the complete data entropy:

$$\max_p H(p) = -\int_{x\in\mathcal{X}} p(x)\log p(x)\mu(dx) \tag{12}$$
$$\text{subject to} \quad \int_{x\in\mathcal{X}} p(x)f_i(x)\mu(dx) = \sum_{y\in\mathcal{Y}} \tilde{p}(y)\int_{x\in X(Y)} p_{\lambda^{(j)}}(x|Y=y)f_i(X)\mu(dx), \quad i = 1...N$$

This is because, plugging Eq. (6) into Eq. (11), we have

$$Q(\lambda, \lambda^{(j)}) = -\log(Z_\lambda) + \sum_{i=1}^{N} \lambda_i \left(\sum_{y\in\mathcal{Y}} \tilde{p}(y) \int_{x\in X(Y)} p_{\lambda^{(j)}}(x|Y=y) f_i(x)\mu(dx) \right)$$

which is exactly the negative of the dual function of Eq. (12).

In the case where the feature functions $f_i(x)$ are all nonegative, the generalized iterative scaling algorithm (GIS) [4] or improved generalized iterative scaling algorithm (IIS) [1, 3] can be used to maximize $Q(\lambda, \lambda')$. Usually only one or several GIS or IIS steps are needed for the M step.

Thus the proposed EM algorithm for maximum entropy with latent variables (LME-EM) is the following:

LME-EM:

E step, compute $\sum_{y\in\mathcal{Y}} \tilde{p}(y) \int_{x\in X(Y)} p_{\lambda^{(j)}}(x|Y=y)f_i(x)\mu(dx)$, for $i = 1...N$;

M step, perform K iterations of full parallel update of parameter values λ_i, $i = 1...N$, by GIS or IIS.

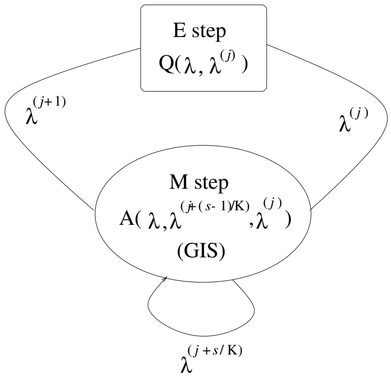

Figure 1: *LME-EM, an EM procedure embedding an iterative scaling loop, where* $A(\lambda^{(j+s/K)}, \lambda^{(j+(s-1)/K)}, \lambda^{(j)})$ *is the auxiliary function in* IIS, s *denotes the index of one cycle of full parallel update of* $\lambda_i, i = 1...N$ *and* K *denotes the number of cycles of full parallel updates.*

A natural interpretation of this iterative procedure is the following: If the right hand side of Eq. (2) is constant, then the optimal soultion of $p_\lambda(x)$ is a log-linear model with parameters provided by the GIS/IIS algorithms. Once we get $p_\lambda(x)$, we could calculate the value of the right hand side of Eq. (2). If this value matches the constant we assigned before, by the optimality condition, we have reached a stationary point of the likelihood function; otherwise, we iterate until the constraints are met.

4 Convergence

The convergence proof for the proposed EM algorithm is quite similar to the GEM algorithm [10].

Theorem: The LME-EM algorithm monotonically increases the likelihood function $L(\lambda)$. All limit points of any LME-EM sequence $\{\lambda^{(j)}, j \geq 0\}$ belong to the set

$$\Gamma = \left\{ \lambda \in \Re^N : \frac{\partial L(\lambda)}{\partial \lambda} = 0 \right\} \tag{13}$$

Proof: First, we show that LME-EM is a generalized EM procedure. To do this we define the auxiliary function A in the same way as in [1, 3]. More specifically, given two parameter settings λ' and λ, we bound from below the change in the objective functions $Q(\lambda, \lambda^{(j)})$ and $Q(\lambda', \lambda^{(j)})$ with an auxiliary function $A(\lambda', \lambda, \lambda^{(j)})$.

$$\begin{aligned}
Q(\lambda, \lambda^{(j)}) - Q(\lambda', \lambda^{(j)}) &= \sum_{i=1}^{N}(\lambda_i - \lambda'_i)\left(\sum_{y\in\mathcal{Y}} \tilde{p}(y)\int_{x\in X(Y)} p_{\lambda^{(j)}}(x|Y=y) f_i(x)\mu(dx)\right) - \log\left(\frac{Z_\lambda}{Z_{\lambda'}}\right) \\
&\geq \sum_{i=1}^{N}(\lambda_i - \lambda'_i)\left(\sum_{y\in\mathcal{Y}} \tilde{p}(y)\int_{x\in X(Y)} p_{\lambda^{(j)}}(x|Y=y) f_i(x)\mu(dx)\right) + 1 - \frac{Z_\lambda}{Z_{\lambda'}} \\
&= \sum_{i=1}^{N}(\lambda_i - \lambda'_i)\left(\sum_{y\in\mathcal{Y}} \tilde{p}(y)\int_{x\in X(Y)} p_{\lambda^{(j)}}(x|Y=y) f_i(x)\mu(dx)\right) + 1 \\
&\quad - \int_{x\in\mathcal{X}} p_{\lambda'}(x) e^{\sum_{i=1}^{N}(\lambda_i - \lambda'_i) f_i(x)}\mu(dx) \\
&\geq \sum_{i=1}^{N}(\lambda_i - \lambda'_i)\left(\sum_{y\in\mathcal{Y}} \tilde{p}(y)\int_{x\in X(Y)} p_{\lambda^{(j)}}(x|Y=y) f_i(x)\mu(dx)\right) + 1 \\
&\quad - \int_{x\in\mathcal{X}} p_{\lambda'}(x) \sum_{i=1}^{N}\left(\frac{f_i(x)}{f(x)}\right) e^{(\lambda_i - \lambda'_i) f(x)}\mu(dx) \\
&= A(\lambda, \lambda', \lambda^{(j)}) \qquad (14)
\end{aligned}$$

where the inequalities follow from the convexity of $-log$ and exp, and $f(x) = \sum_{i=1}^{N} f_i(x)$.

Let s be the index of one cycle of full parallel update of $\lambda_i, i = 1...N$ and assume we perform K cycles of full parallel updates. Then from Eq. (14), we have:

$$Q(\lambda^{(j+s/K)}, \lambda^{(j)}) - Q(\lambda^{(j+(s-1)/K)}, \lambda^{(j)}) \geq A(\lambda^{(j+s/K)}, \lambda^{(j+(s-1)/K)}, \lambda^{(j)}) \qquad (15)$$

By the property of auxiliary function A, it is clear that each step of this procedure increases Q, and thus the LME-EM algorithm monotonically increases the likelihood function $L(\lambda)$.

To show the convergence of λ to the stationary points of the likelihood function, by [10, Theorem 1], we need to show that:

(i) the mapping defined by GIS or IIS is a closed mapping; and

(ii) if $\lambda^{(j)} \notin \Gamma$, then $Q(\lambda^{(j+1)}, \lambda^{(j)}) > Q(\lambda^{(j)}, \lambda^{(j)})$.

First, under the compactness condition [10, (6)] and continuity condition [10, (10)], assertion (i) can be verified directly using $\lambda \in \mathcal{R}^N$. Second, it can be shown that $\frac{\partial Q(\lambda, \lambda^{(j)})}{\partial \lambda} = \frac{\partial A(\lambda, \lambda', \lambda^{(j)})}{\partial \lambda}$. Therefore, if $\lambda^{(j)} \notin \Gamma$, then $\frac{\partial L(\lambda)}{\partial \lambda} \neq 0$, which implies that $\frac{\partial Q(\lambda, \lambda^{(j)})}{\partial \lambda} \neq 0$ and hence $\frac{\partial A(\lambda, \lambda', \lambda^{(j)})}{\partial \lambda} \neq 0$. Therefore we cannot be at a maximum of A. Hence given that $\lambda^{(j+1)}$ maximizes $A(\lambda, \lambda^{(j+(s-1)/M)}, \lambda^{(j)})$ we have $Q(\lambda^{(j+1)}, \lambda^{(j)}) > Q(\lambda^{(j)}, \lambda^{(j)})$ as required.

5 Example Revisited

Recall the simple example in Section 2 where Jaynes' ME principle failed to incorporate latent variables in an underlying model (that had two hidden classes). Here we will find that the extended LME principle will yield a Gaussian mixture as a solution to this problem.

Let $X = (Y, C)$, where C denotes the cluster index. Then the LME formulation will be

$$\max_{p(X)} \quad H(X) = H(C) + H(Y|C) \qquad (16)$$

$$\text{subject to} \quad \int_x p(x) y \delta(C, i)\mu(dx) = \sum_{y\in Y} \tilde{p}(y) \sum_{C\in C(y)} p(x|y) y \delta(C, i), \quad i = 1, 2 \qquad (17)$$

$$\int_x p(x) y^2 \delta(C, i)\mu(dx) = \sum_{y\in Y} \tilde{p}(y) \sum_{C\in C(y)} p(x|y) y^2 \delta(C, i), \quad i = 1, 2 \qquad (18)$$

where $\delta(C,i)$ is the indicator function. Due to the nonlinear mapping caused by $p(x|y)$ there is no closed form solution. However, by applying the LME-EM algorithm, we use (12) to obtain the following ME problem

$$\max_{p(X)} \quad H(X) = H(C) + H(Y|C) \tag{19}$$

$$\text{subject to} \quad \int_x p(x)y\delta(C,i)\mu(dx) = \frac{\sum_{t=1}^T y_t \tau_t^i}{T}, \quad i = 1,2 \tag{20}$$

$$\int_x p(x)y^2\delta(C,i)\mu(dx) = \frac{\sum_{t=1}^T y_t^2 \tau_t^i}{T}, \quad i = 1,2 \tag{21}$$

where $\tau_t^i = p(c = i|y_t, \lambda^{(j)})$. Solving this ME problem, which has a unique solution, we obtain $p(c) = \frac{\sum_{t=1}^T \tau_t^c}{T}$ and $p(y|c) = N(\mu_c, \sigma_c^2)$, where $\mu_c = \frac{\sum_{t=1}^T \tau_t^c y_t}{\sum_{t=1}^T \tau_t^c}$ and $\sigma_c^2 = \frac{\sum_{t=1}^T \tau_t^c (y_t - \mu_c)^2}{\sum_{t=1}^T \tau_t^c}$. These formulae are exactly the same iterative equations for parameter estimation of Gaussian mixture by the EM algorithm. Thus we conclude that the LME principle will give a solution in the format of a Gaussian mixture. This example shows that the LME principle can encode more expressive power than Jaynes' ME principle.

6 Conclusion

We have proposed a latent maximum entropy principle, LME, which incorporates latent variables in the model and yields more expressive power than the standard maximum entropy principle of Jaynes. An EM algorithm that incorporates a nested iterative scaling algorithm is used to solve the problem of approximately calculating a local LME solution, and the local convergence properties of this algorithm are proved. For future work, we are planning to investigate several theoretical questions, including: the asymptotic convergence properties of the LME-ME algorithm, the information geometric properties of the LME principle, incorporating Bayesian methods, investigating the relationship to graphical models, and incorporating Bregman distances. We are also planning to investigate several practical applications of the LME principle, including problems in language modeling, text classification, information retrieval, image analysis, and computer vision. In [9], we have applied this technique to a complex statistical language model where various aspects of natural language—such as local word interaction, syntactic structure, and semantic document information—can be modeled by mixtures of exponential families with a rich expressive power, all in a unified framework.

References

[1] A. Berger, S. Della Pietra and V. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, Vol. 22, No. 1, pp. 39-71, 1996

[2] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991

[3] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing Features of Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 4, pp. 380-393, April 1997

[4] J. Darroch and D. Ratchliff, "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1470-1480, 1972

[5] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," *Journal of Royal Statistical Society, Series B*, Vol. 39, pp 1-38, 1977

[6] S. Fang, J. Rajasekera, H. Tsao, *Entropy Optimization and Mathematical Programming*, Kluwer Academic Publishers, 1997

[7] A. Golan, D. Miller and G. Judge, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, John Wiley & Son, 1996

[8] E. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, edited by R. Rosenkrantz, D. Reidel Publishing Company, 1983

[9] S. Wang, R. Rosenfeld and Y. Zhao, "Latent Maximum Entropy Principle for Statistical Language Modeling," *IEEE Workshop on Automatic Speech Recognition and Understanding*, December 2001

[10] C. Wu, "On the Convergence Properties of the EM Algorithm," *Annals of Statistics*, Vol. 11, pp. 95-103, 1983