The Latent Maximum Entropy Principle

Shaojun Wang* Dale Schuurmans† Yunxin Zhao‡

*Department of Statistics, University of Toronto, Canada
†School of Computer Science, University of Waterloo, Canada
‡Department of Computer Engineering and Computer Science, University of Missouri at Columbia, USA

Abstract

We present an extension to Jaynes' maximum entropy principle that handles latent variables. The principle of *latent maximum entropy* we propose is different from both Jaynes' maximum entropy principle and maximum likelihood estimation, but often yields better estimates in the presence of hidden variables and limited training data. We first show that solving for a latent maximum entropy model poses a hard nonlinear constrained optimization problem in general. However, we then show that feasible solutions to this problem can be obtained efficiently for the special case of log-linear models—which forms the basis for an efficient approximation to the latent maximum entropy principle. We derive an algorithm that combines expectation-maximization with iterative scaling to produce feasible log-linear solutions. This algorithm can be interpreted as an alternating minimization algorithm in the information divergence, and reveals an intimate connection between the latent maximum entropy and maximum likelihood principles. To select a final model, we generate a series of feasible candidates, calculate the entropy of each, and choose the model that attains the highest entropy. Our experimental results show that estimation based on the latent maximum entropy principle generally gives better results than maximum likelihood when estimating latent variable models on small observed data samples.

Index Terms — Maximum entropy, iterative scaling, expectation maximization, latent variable models, information geometry, alternating minimization, graphical models, random fields, probabilistic inference, statistical machine learning

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 [25] is an effective method for combining sources of evidence from complex but structured natural systems which has had wide application in science, engineering, and economics [23, 24]. The effectiveness of the ME principle arises from its ability to model distributions over many random variables by combining only a few critical features (i.e., functions of random variables) in a log-linear form. This can yield a succinct representation of a complex joint distribution and thereby allow for effective generalization and practical inference to be realized; as with standard graphical models like Bayesian networks and Markov random fields. However, unlike standard graphical models, instead of making direct conditional independence assumptions about the domain, the ME principle only requires the specification of certain properties in the data that the model should respect; for example, that the marginal means in the model should match the marginal means in the data. In many applications, specifying constraints on the model in this form is easier than proposing conditional independence properties [19].

However, one weakness with the standard ME approach is that it only handles constraints over the *observed* data, and does not directly model latent variable structure. That is, the standard ME principle does not allow for any missing data in its constraints and therefore never infers the existence of hidden variables. This weakness is problematic because in practice many of the natural patterns 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, natural language data rarely reports the value of hidden semantic variables or syntactic structure [42].

In this paper, we propose a latent maximum entropy principle (LME) that explicitly handles latent variables and thus extends Jaynes' original ME principle to the case where some data components are missing.

We first formulate the problem so that latent variables are explicitly encoded in the model. Although the constrained optimization problem that results is complex, we introduce a log-linear assumption that allows us to derive a practical algorithm (EM-IS) for obtaining feasible solutions. The EM-IS algorithm is an iterative technique that combines expectation-maximization (EM) with iterative scaling (IS) to yield a convergent procedure that is guaranteed to produce log-linear models satisfying desired feature expectations. To develop EM-IS, we show an intimate connection between the latent maximum entropy principle and maximum likelihood estimation (MLE). However, the latent maximum entropy and maximum likelihood principles remain distinct in the sense that, among feasible solutions, LME chooses the model that maximizes entropy, whereas MLE selects the model that maximizes likelihood. To compare these two different approaches for estimating hidden variable models, we then present our main estimation algorithm, ME-EM-IS, which repeatedly solves for different feasible log-linear models, calculates the entropy of each, and selects the model that attains highest entropy. In order to implement this algorithm, we exploit the fact that the entropy can be efficiently determined for the feasible log-linear models produced by EM-IS. Our experimental results show that the LME principle (implemented by the ME-EM-IS algorithm) consistently achieves better estimates than maximum likelihood estimation when estimating hidden variable models from small samples of observed data.

2 Motivation

In 1957, Jaynes [25] 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 infer a probability distribution from data where the distribution should satisfy known constraints, then among distributions consistent with the constraints, one should choose the distribution that has maximum entropy. This principle can be understood clearly by considering the case of modeling a single real variable:

2.1 A simple example

Assume we observe a random variable Y that reports people's heights in a population. Given sample data $\tilde{Y} = (y_1, ..., y_T)$, one might trust that simple statistics such as the sample mean and sample mean square of Y are well represented in the data. If so, then Jaynes' ME principle suggests that one should infer a distribution for Y that has maximum entropy, subject to the constraints that the mean and mean square values of Y match the sample values; that is, that $EY = m_1$ and $EY^2 = m_2$, where $m_1 = \frac{1}{T}\sum_{t=1}^{T} y_t$ and $m_2 = \frac{1}{T}\sum_{t=1}^{T} y_t^2$ respectively. In this case, it is known that the maximum entropy solution is a Gaussian density with mean m_1 and variance $m_2 - m_1^2$, $p(y) = N(y; m_1, m_2 - m_1^2)$; 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 [12].

However, assume further that after observing the data histogram we find that there are actually two peaks in the empirical data. Obviously the standard ME solution would not be the most appropriate model for such bi-modal data, because it will continue to postulate a uni-modal distribution. However, the existence of the two peaks in the data might not be accidental. For example, there could be two sub-populations represented in the data, male and female, each of which have different height distributions. In this case, each height measurement Y has an accompanying (hidden) gender label C that indicates the sub-population the measurement is taken from. How can such additional knowledge be incorporated in the ME framework? One way is to *explicitly* add the missing label data. That is, we could let $X = (Y, C)$, where Y denotes a person's height and C is the gender label, and then obtain *labeled* measurements $(y_1, c_1, ..., y_T, c_T)$. In this case we can formulate the ME problem as follows. Let $\delta_k(c)$ be the indicator function where $\delta_k(c) = 1$ if $c = k$ and $\delta_k(c) = 0$ otherwise. Then let $N_k = \sum_{t=1}^{T} \delta_k(c_t)$, $\tilde{p}(C = k) = \frac{N_k}{T}$, $\tilde{p}(y_t|C = k) = \frac{\delta_k(c_t)}{N_k}$, for $k = 1, 2$, and let $\tilde{\mathcal{Y}}$ denote the set of observed heights $(y_1, ..., y_T)$. With these definitions, then formulate the ME problem as

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

$$\begin{aligned}
\text{subject to} \quad \int_{x\in\mathcal{X}} \delta_k(c)\;\; p(x)\,\mu(dx) &= \sum_{c\in\{1,2\}} \delta_k(c)\;\; \tilde{p}(c) \\
\int_{x\in\mathcal{X}} y\,\delta_k(c)\;\; p(x)\,\mu(dx) &= \sum_{y\in\tilde{\mathcal{Y}}}\;\sum_{c\in\{1,2\}} y\,\delta_k(c)\;\; \tilde{p}(c)\,\tilde{p}(y|c) \qquad (1)\\
\int_{x\in\mathcal{X}} y^2\,\delta_k(c)\;\; p(x)\,\mu(dx) &= \sum_{y\in\tilde{\mathcal{Y}}}\;\sum_{c\in\{1,2\}} y^2\,\delta_k(c)\;\; \tilde{p}(c)\,\tilde{p}(y|c) \qquad \text{for } k=1,2
\end{aligned}$$

The problem then is to find a joint model $p(x) = p(y, c)$ that maximizes entropy while matching the expectations over $\delta_k(c)$, $y\,\delta_k(c)$, and $y^2\,\delta_k(c)$, for $k = 1, 2$. In this fully observed data case, *where we witness the gender label* C, we obtain a separable optimization problem that has a unique solution. In this case, the maximum entropy solution $p(x) = p(y, c)$ is a mixture of two Gaussian distributions specified by $p(c) = \theta_c = \frac{N_c}{T}$ and $p(y|c) = N(y; \mu_c, \sigma_c^2)$, where $\mu_c = \frac{1}{N_c}\sum_{t=1}^{T} y_t\,\delta_c(c_t)$ and $\sigma_c^2 = \frac{1}{N_c}\sum_{t=1}^{T}(y_t - \mu_c)^2\,\delta_c(c_t)$ for $c = 1, 2$.

Unfortunately, obtaining fully labeled data is tedious or impossible in most realistic situations. In cases where variables are unobserved, Jaynes' ME principle, which is maximally noncommittal with respect to missing information, becomes insufficient. For example, if the gender label were unobserved, one would still be reduced to inferring a single uni-modal Gaussian, as above. To cope with missing but non-arbitrary hidden structure, we must extend the ME principle to account for the underlying causal structure in the data model.

3 The LME principle

To formulate the latent maximum entropy (LME) principle, let $X \in \mathcal{X}$ be a random variable denoting the complete data, $Y \in \mathcal{Y}$ be the observed incomplete data and $Z \in \mathcal{Z}$ be the missing data. That is, $X = (Y, Z)$. 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 Z. If we let $p(x)$ and $p(y)$ denote the densities of X and Y respectively, and let $p(z|y)$ denote the conditional density of Z given Y, then $p(y) = \int_{z\in\mathcal{Z}} p(x)\;\mu(dz)$, and $p(x) = p(y)p(z|y)$.[1] Given this notation we propose the latent maximum entropy principle as follows.

LME principle Given features $f_1, ..., f_N$, specifying the properties that we would like to match in the data, select a joint probability model $p(x)$ from the space of all probability distributions, $\mathcal{P}$, over $\mathcal{X}$, to maximize the entropy

$$H(p) \;=\; -\int_{x\in\mathcal{X}} p(x)\log p(x)\;\mu(dx) \qquad (2)$$

subject to the constraints

$$\int_{x\in\mathcal{X}} f_i(x)\;\; p(x)\,\mu(dx) \;=\; \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} f_i(x)\;\; p(z|y)\,\mu(dz), \qquad \text{for } i = 1, ..., N \qquad (3)$$

where $x = (y, z)$. ▮

Here $\tilde{p}(y)$ is the empirical distribution of the observed data, $\tilde{\mathcal{Y}}$ denotes the set of observed Y values, and $p(z|y)$ is the conditional distribution of latent variables given the observed data. Intuitively, the constraints specify that we require the expectations of $f_i(X)$ in the joint model to match their empirical expectations on the incomplete data Y, taking into account the structure of the implied dependence of the unobserved component Z on Y.

Note that the conditional distribution $p(z|y)$ implicitly encodes the latent structure and is a nonlinear mapping of $p(x)$. That is, $p(z|y) = p(y, z)/\int_{z'\in\mathcal{Z}} p(y, z')\mu(dz) = p(x)/\int_{x'=(y,z')} p(x')\mu(dx')$ where $x = (y, z)$

[1] 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.

and $x' = (y, z')$ by definition. Clearly, $p(z|y)$ is a nonlinear function of $p(x)$ because of the division. If there is no missing data, i.e. $X = Y$, then the problem is reduced to Jaynes' model where the constraints are given by $\int_{y\in\mathcal{Y}} p(y)f_j(y)\ \mu(dy) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)f_j(y)$. However, this is not a requirement in our framework, and in this sense, the LME principle given by (2) and (3) is more general than ME.

Unfortunately, we will find that the most straightforward formulation of LME does not yield a simple closed form solution for the optimal distribution. Nevertheless, by further constraining the distribution to have an exponential (log-linear) form, we will be able to show the equivalence between satisfying the constraints (i.e. achieving feasibility) and locally maximizing likelihood. This equivalence will allow us to derive a practical algorithm for finding feasible solutions in Section 4.

3.1 Finding LME solutions

Consider the problem of finding a joint distribution $p(x)$ that satisfies the LME principle for a given set of features and data (where, for example, the features could specify sufficient statistics for a desired exponential model). This problem amounts to solving the constrained optimization problem (2,3). Unfortunately, due to the mapping $p(z|y)$, the constraints (3) are *nonlinear* in $p(x)$ and the feasible set is no longer convex. Therefore, even though the objective function (2) is concave, no unique maximum can be guaranteed to exist. In fact, minima and saddle points may exist. Nevertheless, one can still attempt to derive an iterative training procedure that finds approximate local solutions to the LME problem.

First, define the Lagrangian $\Lambda(p,\lambda)$ by

$$\Lambda(p,\lambda) \quad = \quad H(p) \ + \ \sum_{i=1}^{N} \lambda_i \left(\int_{x\in\mathcal{X}} f_i(x)\, p(x)\, \mu(dx) \ - \ \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} f_i(x)\, p(z|y)\, \mu(dz) \right) \tag{4}$$

A natural way to proceed with the optimization is to iteratively hold λ fixed and compute the unconstrained maximum of the Lagrangian over $p \in \mathcal{P}$. To do so let

$$\begin{aligned} p_\lambda &= \arg\max_{p\in\mathcal{P}} \ \Lambda(p,\lambda) \\ \Upsilon(\lambda) &= \Lambda(p_\lambda,\lambda) \end{aligned}$$

We refer to $\Upsilon(\lambda)$ as the *dual function*. Note that by weak duality the dual function provides upper bounds on the optimal value H^* of the original LME problem:

$$\Upsilon(\lambda) \ = \ \Lambda(p_\lambda,\lambda) \ = \ \max_{p\in\mathcal{P}} \ \Lambda(p,\lambda) \ \geq \ H^* \quad \text{for all } \lambda$$

If strong duality holds, we have

$$\min_\lambda \ \Upsilon(\lambda) \ = \ \min_\lambda \ \Lambda(p_\lambda,\lambda) \ = \ \min_\lambda \max_{p\in\mathcal{P}} \ \Lambda(p,\lambda) \ = \ H^*$$

Therefore, if one could obtain a closed form solution for p_λ in terms of λ, one could then plug p_λ into $\Lambda(p_\lambda,\lambda)$ and reduce the constrained optimization to the *unconstrained minimization* of $\Upsilon(\lambda)$ with respect to λ. However, in attempting to solve for p_λ we still run into difficulty.

To attempt to solve for p_λ, one can take the derivative of $\Lambda(p,\lambda)$ with respect to $p(x)$ and try to set this to 0 for all $p(x)$:

$$\begin{aligned} \frac{\partial\Lambda(p,\lambda)}{\partial p(x)} &= -\log p(x) \ - \ 1 \ + \ \sum_{i=1}^{N} \lambda_i \left[f_i(x) - \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \left(\frac{f_i(x)}{p(y)} - \frac{\int_{z'\in\mathcal{Z}} f_i(x')\, p(x')\, \mu(dz')}{\left(\int_{z''\in\mathcal{Z}} p(x'')^2\, \mu(dz'')\right)} \right) \right] \\ &= -\log p(x) \ - \ 1 \ + \ \sum_{i=1}^{N} \lambda_i f_i(x) \ + \ \sum_{i=1}^{N} \lambda_i \left(\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \frac{\int_{z'\in\mathcal{Z}} [f_i(x') - f_i(x)]\, p(x')\, \mu(dz')}{p(y)^2} \right) \end{aligned} \tag{5}$$

where $x = (y,z)$, $x' = (y,z')$ and $x'' = (y,z'')$. Unfortunately the resulting system $\partial\Lambda/\partial p(x) = 0$ is nonlinear in $p(x)$ and there is no simple closed form solution for p_λ.

3.2 Approximating LME solutions: Restriction to log-linear form

Since the original LME principle does not yield a simple closed form solution for p_λ, we instead look for an approximate solution. By ignoring the last term of Equation (5) and setting the remainder to zero, we find

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

where $\Phi_\lambda = \int_{x\in\mathcal{X}} \exp\left(\sum_{i=1}^N \lambda_i f_i(x)\right) \mu(dx)$ is a normalizing constant that ensures $\int_{x\in\mathcal{X}} p_\lambda(x)\ \mu(dx) = 1$. Thus, one could hope that p_λ is at least approximately log-linear. Note that if we impose the additional constraint that p_λ is indeed log-linear, (6), and plug this back into the definition of the Lagrangian (4) we can obtain a closed form for an approximation to the dual function

$$\Upsilon(\lambda) \quad \approx \quad \log(\Phi_\lambda) \quad - \quad \sum_{i=1}^N \lambda_i \left(\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} f_i(x)\ p_\lambda(z|y)\, \mu(dz)\right) \tag{7}$$

That is, under the assumption of a log-linear model p_λ we can approximately reduce the original constrained optimization to a much simpler unconstrained minimization problem

$$\lambda^* \quad = \quad \arg\min_\lambda \Upsilon(\lambda) \tag{8}$$

where Υ is given as in (7). Assuming λ^* can be found, we can easily recover p_{λ^*} from (6), up to the normalization constant Φ_λ^{-1}.

Now to attempt to solve for λ^*, take the derivative of $\Upsilon(\lambda)$ with respect to λ, and obtain

$$\begin{aligned}
\frac{\partial\Upsilon(\lambda)}{\partial\lambda_i} \quad = \quad & \int_{x\in\mathcal{X}} f_i(x)\ p_\lambda(x)\, \mu(dx) \quad - \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} f_i(x)\ p_\lambda(z|y)\, \mu(dz) \\
& - \quad \sum_{j=1}^N \lambda_j \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \Bigg(\int_{z\in\mathcal{Z}} f_i(x) f_j(x)\ p_\lambda(z|y)\, \mu(dz) \\
& \qquad - \int_{z\in\mathcal{Z}} f_i(x)\ p_\lambda(z|y)\, \mu(dz) \int_{z\in\mathcal{Z}} f_j(x)\ p_\lambda(z|y)\, \mu(dz) \Bigg)
\end{aligned} \tag{9}$$

Unfortunately, once again, the system of equations $\partial\Upsilon(\lambda)/\partial\lambda_i = 0$ is nonlinear due to the $p_\lambda(z|y)$ terms, and therefore this does not yield a simple closed form solution for λ^*. Even under the log-linear assumption, it is still not easy to satisfy the LME principle! Nevertheless, we have made valuable progress toward formulating a practical algorithm for approximately satisfying the LME principle under the assumption of log-linearity. In fact, at this point we can show an intimate connection between the LME principle and maximum likelihood estimation (MLE) principle under log-linear models.

Theorem 1 *Under the log-linear assumption, maximizing the likelihood of log-linear models on incomplete data is equivalent to satisfying the feasibility constraints of the LME principle. That is, the only distinction between MLE and LME in log-linear models is that, among local maxima (feasible solutions), LME selects the model with the maximum entropy, whereas MLE selects the model with the maximum likelihood.*

Proof: By assuming a log-linear model p_λ, we first prove that satisfying the constraints (3) of the LME principle is equivalent to achieving a local maxima in log-likelihood. Restrict the complete model p_λ to have a log-linear form $p_\lambda(x) = \Phi_\lambda^{-1} \exp(\sum_{i=1}^N \lambda_i f_i(x))$. Then we have $p_\lambda(y) = \int_{z\in\mathcal{Z}} p_\lambda(x)\ \mu(dz)$ and the log-likelihood function for the observed incomplete data is given by

$$L(\lambda) \quad = \quad \log \prod_{y\in\tilde{\mathcal{Y}}} p_\lambda(y)^{\tilde{p}(y)} \quad = \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \log p_\lambda(y) \tag{10}$$

Taking the derivative of $L(\lambda)$ with respect to λ_i yields

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

By setting $\partial L(\lambda)/\partial\lambda_i = 0$, for $i = 1, ..., N$, we obtain the original constraints (3). Therefore the feasible solutions of (3) satisfy the conditions for the stationary points of the log-likelihood function. This establishes the first part of the theorem.

All that remains is to show that the MLE and LME principles remain distinct for log-linear models. We prove this by proving that the log-likelihood function $L(\lambda)$ and entropy $H(p_\lambda)$ are related by the equation $L(\lambda) = H(p_\lambda) + H(\lambda, \lambda)$, where $H(\lambda,\lambda)$ is a non-constant function of λ whose maxima generally do not coincide with $L(\lambda)$ or $H(p_\lambda)$. This fact is proved in Theorem 4 in Section 5 below. Given this result, we conclude that among feasible log-linear solutions, MLE and LME do not maximize the same objective, and hence produce different solutions. ∎

Although the problem of maximum likelihood estimation of log-linear models with missing data has previously been studied by Lauritzen [28] and Riezler [36], it had not been previously observed that locally maximizing the likelihood of a log-linear model is equivalent to satisfying the feasibility constraints for a latent maximum entropy problem.

3.3 Example revisited

To illustrate the relationship between the MLE and LME principles more concretely, consider the simple example introduced in Section 2.1. In the circumstance where the gender labels are unobserved, Jaynes' ME principle fails to incorporate the effect of these latent variables. However, the LME principle can capture the influence of the latent gender information by considering a joint model that includes a hidden two-valued variable. Let $X = (Y, C)$, where $C \in \{1, 2\}$ denotes the hidden gender index. In this case, given the observed data $\tilde{\mathcal{Y}} = (y_1, ..., y_T)$, the *latent* maximum entropy principle (LME) can be formulated as

$$\begin{aligned}
&\max_{p(x)} \ H(X) &&= H(C) + H(Y|C) \\
\text{subject to} \quad &\int_{x\in\mathcal{X}} \delta_k(c)\ \ p(x)\,\mu(dx) &&= \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c\in\{1,2\}} \delta_k(c)\ \ p(c|y) \\
&\int_{x\in\mathcal{X}} y\,\delta_k(c)\ \ p(x)\,\mu(dx) &&= \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c\in\{1,2\}} y\,\delta_k(c)\ \ p(c|y) \\
&\int_{x\in\mathcal{X}} y^2\,\delta_k(c)\ \ p(x)\,\mu(dx) &&= \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c\in\{1,2\}} y^2\,\delta_k(c)\ \ p(c|y) \qquad \text{for } k = 1, 2
\end{aligned} \tag{11}$$

So here we are trying to maximize the joint entropy while matching the expectations over the features

$$f_0^k(x) \ = \ \delta_k(c), \quad f_1^k(x) \ = \ y\,\delta_k(c), \quad \text{and} \quad f_2^k(x) \ = \ y^2\,\delta_k(c), \quad \text{for } k = 1, 2 \tag{12}$$

where $x = (y, c)$, and $\delta_k(c)$ denotes the indicator function of the event $c = k$. Comparing the constraints (11) with those in the complete data case (1), one can see that the only difference is that here we use the conditional probability of the complete model instead of the empirical conditional probability. However, due to the nonlinear mapping imposed by $p(c|y)$ a simple closed form solution no longer exists. Nevertheless, a common log-linear model gives a convenient approximation.

Imagine that, instead of attempting to satisfy the LME principle directly, one was instead interested in finding a maximum likelihood model for the observed data $\tilde{\mathcal{Y}} = (y_1, ..., y_T)$. Consider a distribution $p(x)$ that is a mixture of two Gaussians; i.e., $p(x) = p(y, c) = \theta_c N(y; \mu_c, \sigma_c^2)$ for parameters $\theta_c, \mu_c, \sigma_c^2$, where $\theta_c = p(c)$, and μ_c, σ_c^2 are the means and variances for the respective classes $c = 1, 2$. This distribution has the marginal density $p(y) = \theta_1 N(y; \mu_1, \sigma_1^2) + \theta_2 N(y; \mu_2, \sigma_2^2)$ on Y. In this case, the joint distribution of $X = (Y, C)$ can be written

$$p(y, c) \quad = \quad \prod_{k \in \{1,2\}} \left[\theta_k \, \frac{1}{\sqrt{2\pi\sigma_k^2}} \, \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2} \right) \right]^{\delta_k(c)}$$

If we use the natural (canonical) parameters $\lambda = (\lambda_0^k, \lambda_1^k, \lambda_2^k)$ for the corresponding features f_0^k, f_1^k and f_2^k given in (12), $k = 1, 2$, we can then re-write this distribution in a log-linear form [3]

$$\begin{aligned} p(y, c) \quad &= \quad \prod_{k \in \{1,2\}} \left(\frac{1}{\Phi_{\lambda_0^1 \lambda_0^2}} e^{\lambda_0^k} \frac{1}{\Phi_{\lambda_1^k \lambda_2^k}} e^{\lambda_1^k \, y \, + \, \lambda_2^k \, y^2} \right)^{\delta_k(c)} \\ &= \quad \frac{1}{\Phi_\lambda} \exp\left(\sum_{k=1}^{2} \left(\lambda_0^k \, \delta_k(c) + \lambda_1^k \, y \, \delta_k(c) + \lambda_2^k \, y^2 \, \delta_k(c) \right) \right) \end{aligned} \tag{13}$$

where the canonical parameters are related to the standard parameters by $\lambda_0^k = \log \theta_k$, $\lambda_1^k = \mu_k / \sigma_k^2$, and $\lambda_2^k = -1/(2\sigma_k^2)$ for $k = 1, 2$. The normalization constant is given by $\Phi_\lambda = \Phi_{\lambda_0^1 \lambda_0^2} \Phi_{\lambda_1^1 \lambda_2^1} \Phi_{\lambda_1^2 \lambda_2^2}$, where $\Phi_{\lambda_0^1 \lambda_0^2} = 1/(e^{\lambda_0^1} + e^{\lambda_0^2})$ and $\Phi_{\lambda_1^k \lambda_2^k} = \exp(-(\lambda_1^k)^2/(4\lambda_2^k))\sqrt{2\sigma_k^2 \pi}$ for $k = 1, 2$. For this model, the log-likelihood, as a function of λ, can be written

$$\begin{aligned} L(\lambda) \quad &= \quad \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \log p(y) \\ &= \quad \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \; \log \sum_{c \in \{1,2\}} \frac{1}{\Phi_\lambda} \exp\left(\sum_{k=1}^{2} \left(\lambda_0^k \, \delta_k(c) + \lambda_1^k \, y \, \delta_k(c) + \lambda_2^k \, y^2 \, \delta_k(c) \right) \right) \end{aligned}$$

Therefore, to solve for the maximum likelihood solution one can calculate the derivatives to obtain

$$\begin{aligned} \frac{\partial L(\lambda)}{\lambda_0^k} \quad &= \quad \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c \in \{1,2\}} \delta_k(c) \;\; p(c|y) \;\; - \;\; \int_{y \in \mathcal{Y}} \sum_{c \in \{1,2\}} \delta_k(c) \;\; p(y, c) \, dy \\ \frac{\partial L(\lambda)}{\lambda_1^k} \quad &= \quad \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c \in \{1,2\}} y \, \delta_k(c) \;\; p(c|y) \;\; - \;\; \int_{y \in \mathcal{Y}} \sum_{c \in \{1,2\}} y \, \delta_k(c) \;\; p(y, c) \, dy \\ \frac{\partial L(\lambda)}{\lambda_2^k} \quad &= \quad \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c \in \{1,2\}} y^2 \, \delta_k(c) \;\; p(c|y) \;\; - \;\; \int_{y \in \mathcal{Y}} \sum_{c \in \{1,2\}} y^2 \, \delta_k(c) \;\; p(y, c) \, dy \qquad \text{for } k = 1, 2 \end{aligned} \tag{14}$$

The key result is that setting these quantities to zero results in precisely the same constraints as (11). That is, a locally maximum likelihood Gaussian mixture is also a feasible solution of the LME principle, and conversely, a feasible log-linear solution for the LME principle will be a critical point of the log-likelihood function $L(\lambda)$ (and have the form of a Gaussian mixture). This example provides a concrete demonstration that the log-linear model parameterized with the stationary points of the incomplete data likelihood function will give a feasible solution to the original LME principle.

4 A general algorithm for finding feasible log-linear solutions

We can now exploit the observation of Theorem 1 to derive a practical training algorithm for obtaining feasible solutions to the LME principle under the log-linear assumption. Obviously, since Theorem 1 shows that locally maximizing the likelihood of observed incomplete data will satisfy the constraints of the LME principle (3), the most natural strategy is to derive an EM algorithm for log-linear models. In so doing, we will be able to guarantee that we recover feasible solutions to the original constrained optimization problem, by Theorem 1.

4.1 Derivation of the EM-IS iterative algorithm

Recall that a log-linear model is determined by its parameter vector λ (6). Therefore to derive the EM algorithm [21] one typically decomposes the log-likelihood $L(\lambda)$ as a function of λ into

$$\begin{aligned} L(\lambda) &= \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\log p_\lambda(y) \\ &= Q(\lambda,\lambda') \;+\; H(\lambda,\lambda') \qquad \text{for all } \lambda' \end{aligned} \tag{15}$$

$$\text{where} \quad Q(\lambda,\lambda') = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} p_{\lambda'}(z|y)\log p_\lambda(x)\;\mu(dz) \tag{16}$$

$$\text{and} \quad H(\lambda,\lambda') = -\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} p_{\lambda'}(z|y)\log p_\lambda(z|y)\;\mu(dz) \tag{17}$$

Here $x=(y,z)$, $Q(\lambda,\lambda')$ is the conditional expected complete-data log-likelihood, and $H(\lambda,\lambda')$ is the conditional expected missing data log-likelihood, which measures the uncertainty due to missing data. Note that in the case where $\lambda'=\lambda$, $H(\lambda,\lambda)$ becomes the empirical conditional entropy on latent variables.

The EM algorithm maximizes $L(\lambda)$ by iteratively maximizing $Q(\lambda,\lambda')$ over λ. The jth iteration $\lambda^{(j)}\rightarrow\lambda^{(j+1)}$ of EM 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 monotonically non-decreases $L(\lambda)$, and very generally, if EM converges to a fixed point λ^*, then λ^* is a stationary point of $L(\lambda)$ which is usually a local maximum [21, 43].[2]

For log-linear models in particular, we have

$$Q(\lambda,\lambda^{(j)}) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} p_{\lambda^{(j)}}(z|y)\log p_\lambda(x)\;\mu(dz) \tag{18}$$

$$\begin{aligned} &= \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} p_{\lambda^{(j)}}(z|y)\left[\left(\sum_{i=1}^N \lambda_i f_i(x)\right) - \log(\Phi_\lambda)\right]\;\mu(dz) \\ &= -\log(\Phi_\lambda) \;+\; \sum_{i=1}^N \lambda_i \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} f_i(x)\;p_{\lambda^{(j)}}(z|y)\,\mu(dz) \end{aligned} \tag{19}$$

by plugging the log-linear form (6) into (18) and recalling that $x=(y,z)$. Crucially, it turns out that maximizing $Q(\lambda,\lambda^{(j)})$ as a function of λ for fixed $\lambda^{(j)}$ (the M step) is equivalent to solving another constrained optimization problem corresponding to a maximum entropy principle; but a much simpler one than before.

Theorem 2 *Maximizing $Q(\lambda,\lambda^{(j)})$ as a function of λ for fixed $\lambda^{(j)}$ is equivalent to solving*

$$\max_{p_\lambda} \quad H(p_\lambda) = -\int_{x\in\mathcal{X}} p_\lambda(x)\log p_\lambda(x)\;\mu(dx) \tag{20}$$

$$\textit{subject to} \quad \int_{x\in\mathcal{X}} f_i(x)\;p_\lambda(x)\,\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} f_i(x)\;p_{\lambda^{(j)}}(z|y)\,\mu(dz), \quad i=1,...,N \tag{21}$$

where $x=(y,z)$.

[2] It is usually possible to check whether the stationary point is in fact a local maximum [21, 43].

Proof: Define the Lagrangian $\Lambda(p, \lambda, \lambda^{(j)})$ by

$$\Lambda(p, \lambda, \lambda^{(j)}) \;\; = \;\; H(p) \; + \; \sum_{i=1}^{N} \lambda_i \left(\int_{x \in \mathcal{X}} p(x) f_i(x) \mu(dx) - \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} p_{\lambda^{(j)}}(z|y) f_i(x) \mu(dz) \right) \tag{22}$$

Holding $\lambda^{(j)}$ fixed, compute the unconstrained maximum of the Lagrangian over $p \in \mathcal{P}$, to get

$$\begin{aligned} p_\lambda &= \arg\max_{p \in \mathcal{P}} \; \Lambda(p, \lambda, \lambda^{(j)}) \\ &= \Phi_\lambda^{-1} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(x) \right) \end{aligned}$$

(This result is obtained by taking the derivative of (22) with respect to $p(x)$ and setting it to zero.) Now by plugging p_λ into $\Lambda(p_\lambda, \lambda, \lambda^{(j)})$, we obtain the dual function

$$\Upsilon(\lambda, \lambda^{(j)}) \;\; = \;\; \Lambda(p_\lambda, \lambda, \lambda^{(j)}) \;\; = \;\; \log(\Phi_\lambda) \; - \; \sum_{i=1}^{N} \lambda_i \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) \, p_{\lambda^{(j)}}(z|y) \, \mu(dz)$$

which is exactly the negative of $Q(\lambda, \lambda^{(j)})$ as given in (19). If we denote the optimal value of (20) subject to (21) as $H^*(\lambda^{(j)})$, then under the conditions where strong duality holds [5, 31] we have

$$\begin{aligned} \max_\lambda \; Q(\lambda, \lambda^{(j)}) &= -\min_\lambda \; \Upsilon(\lambda, \lambda^{(j)}) \\ &= -\min_\lambda \; \Lambda(p_\lambda, \lambda, \lambda^{(j)}) \\ &= -\min_\lambda \max_{p \in \mathcal{P}} \; \Lambda(p, \lambda, \lambda^{(j)}) \\ &= -H^*(\lambda^{(j)}) \end{aligned} \tag{23}$$

∎

It is important to realize that the new constrained optimization problem in Theorem 2 is much easier than maximizing (2) subject to (3) for log-linear models, because the right hand side of the constraints (21) no longer depend on λ but on the previous fixed $\lambda^{(j)}$. That means maximizing (20) subject to (21) is now a convex optimization problem with *linear* constraints in p_λ. Unfortunately, there is no closed form solution to (20,21) in general, which means that iterative algorithms are usually necessary. However, the maximizer is unique if it exists. For such problems there are a large number of iterative algorithms available, including Bregman's balancing method, the multiplicative algebraic reconstruction technique (MART), Newton's method, conjugate gradient, and interior-point methods [9, 23]. In the case where the feature functions $f_i(x)$ are all non-negative, the generalized iterative scaling algorithm (GIS) [18] or improved iterative scaling algorithm (IIS) [4, 19] can be used to maximize $Q(\lambda, \lambda')$ very efficiently. Usually only a few GIS or IIS iterations are needed for the M step.

Given these observations, we propose maximizing the entropy of log-linear models with latent variables by using an algorithm that combines EM with nested iterative scaling (either IIS or GIS) to calculate the M step; see Figure 1.

EM-IS algorithm:

Initialization: Randomly choose initial guesses for the parameters, $\lambda^{(0)}$.

E step: Given the current model $\lambda^{(j)}$, for each feature f_i, $i = 1, ..., N$, calculate its current expectation $\eta_i^{(j)}$ with respect to $\lambda^{(j)}$ by

$$\eta_i^{(j)} \;\; = \;\; \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x) \;\; p_{\lambda^{(j)}}(z|y) \, \mu(dz) \tag{24}$$

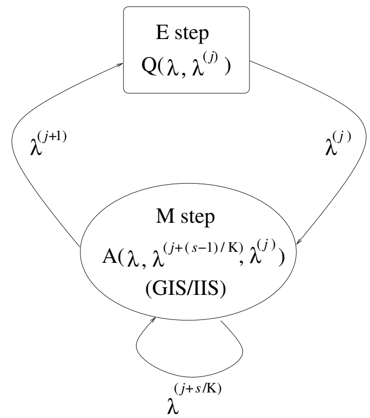


Figure 1: *EM-IS, an EM procedure embedding an iterative scaling loop, where* $A\left(\lambda^{(j+s/K)}, \lambda^{(j+(s-1)/K)}, \lambda^{(j)}\right)$ *is the auxiliary function in* GIS/IIS*,* s *denotes the index of one cycle of full parallel update of* $\lambda_i, i=1,...,N$ *and* K *denotes the number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

These quantities will form the right hand side of the constraints in (21).

M step: Let $f(x) = \sum_{i=1}^N f_i(x)$. To attempt to solve (21) (or equivalently, maximize $Q(\lambda, \lambda^{(j)})$ with respect to λ): initialize λ to $\lambda^{(j)}$ and perform K iterations of a full parallel update of the parameter values λ_i, $i = 1,...,N$, either by GIS or IIS, as follows. Each update is given by

$$\lambda_i^{(j+s/K)} \quad = \quad \lambda_i^{(j+(s-1)/K)} + \gamma_i^{(j+s/K)} \tag{25}$$

where $\gamma_i^{(j+s/K)}$ satisfies

$$\int_{x\in\mathcal{X}} f_i(x)\, e^{\gamma_i^{(j+s/K)} f(x)}\; p_{\lambda^{(j+(s-1)/K)}}(x)\, \mu(dx) \quad = \quad \eta_i^{(j)} \tag{26}$$

In the special case where $f(x)$ is a constant, i.e., $f(x) = a$ for all x, $\gamma_i^{(j+s/K)}$ is given explicitly by

$$\gamma_i^{(j+s/K)} \quad = \quad \frac{1}{a}\log\left(\frac{\eta_i^{(j)}}{\int_{x\in\mathcal{X}} f_i(x)\; p_{\lambda^{(j+(s-1)/K)}}(x)\,\mu(dx)}\right) \qquad \text{for } s=1,...,K \tag{27}$$

If $f(x)$ is not constant, then the value of $\gamma_i^{(j+s/K)}$ has to be computed numerically, for example, by solving the nonlinear equation (26) using Newton-Raphson:

$$\gamma_i^{(j+s/K)}(\text{new}) \quad = \quad \gamma_i^{(j+s/K)}(\text{old}) \;\; - \;\; \frac{\int_{x\in\mathcal{X}} f_i(x)\, e^{\gamma_i^{(j+s/K)}(\text{old}) f(x)}\; p_{\lambda^{(j+(s-1)/K)}}(x)\,\mu(dx) \;-\; \eta_i^{(j)}}{\int_{x\in\mathcal{X}} f_i(x)\, f(x)\, e^{\gamma_i^{(j+s/K)}(\text{old}) f(x)}\; p_{\lambda^{(j+(s-1)/K)}}(x)\,\mu(dx)}$$

It is also possible to use a bisection method for this purpose.

Repeat until: $\lambda^{(j+1)} \approx \lambda^{(j)}$. ∎

Note that in implementing this algorithm, as with any EM or IS algorithm, one must be able to calculate various expectations with respect to the underlying log-linear model p_λ. In particular, we need to calculate expectations of the form $\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} g(x)\; p_\lambda(z|y)\,\mu(dz)$ and $\int_{x\in\mathcal{X}} g(x)\; p_\lambda(x)\,\mu(dx)$ for a given λ. In structured models, such as Gaussian mixtures or other simple log-linear models, these expectations can be calculated directly and efficiently (in time polynomial in the number of features N and the number of observations T). However, in other log-linear models, such efficient algorithms for calculating expectations do not exist, and one must resort to Monte Carlo methods or approximation methods in these cases [19]. We will demonstrate both kinds of models in Section 7 below.

A natural interpretation of the iterative EM-IS procedure is the following: If the right hand side of Equation (3) is constant, then the optimal solution of p_λ is a log-linear model with parameters provided by the GIS/IIS algorithm. Once we obtain p_λ, we can calculate the value of the right hand side of Equation (3). If this value matches the constant assigned previously, by the optimality condition, we have reached a stationary point of the likelihood function and hence a feasible solution of maximizing the entropy for the complete model subject to the required nonlinear constraints. Otherwise, we iterate until the constraints are met.

We note that Lauritzen [28] has suggested similar algorithm for maximum likelihood estimation of log-linear models with incomplete data. However, he did not supply a proof of convergence (which we provide below). More recently, Riezler [36] has also proposed a similar algorithm, but disfavored the doubly iterative approach of nesting iterative scaling inside an EM loop. Instead, Riezler proposed a single loop procedure by repeatedly applying the auxiliary function to obtain a closed form solution for the parameter estimates. However, it turns out that Riezler's algorithm is a special case of our EM-IS algorithm by setting $K = 1$. Although the nested iteration of EM-IS might appear to be an unnecessary complication, we will see in Section 7 that setting $K > 1$ is important for obtaining rapid convergence.

4.2 Example

To demonstrate how EM-IS can be applied, consider the simple example from Sections 2.1 and 3.3. Given a joint model $X = (Y, C)$ representing heights and gender labels, where we only observe height measurements $\tilde{\mathcal{Y}} = (y_1, ..., y_T)$, the LME principle can be formulated as shown in (11). To solve for a feasible log-linear model, we apply EM-IS as follows: First, start with some initial guess for the parameters $\lambda^{(0)}$, where we use the canonical parameterization $\lambda = (\lambda_0^k, \lambda_1^k, \lambda_2^k)$, $k = 1, 2$, for the features specified in (12). To execute the E step, we then calculate the feature expectations according to (24)

$$\begin{aligned}
\eta_0^{k,(j)} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{c\in\{1,2\}} \delta_k(c)\,\rho_t^{k,(j)} \\
\eta_1^{k,(j)} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{c\in\{1,2\}} y_t\,\delta_k(c)\,\rho_t^{k,(j)} \\
\eta_2^{k,(j)} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{c\in\{1,2\}} y_t^2\,\delta_k(c)\,\rho_t^{k,(j)} \qquad \text{for } k = 1, 2
\end{aligned}$$

where here $\rho_t^{k,(j)} = p_{\lambda^{(j)}}(C = k|y_t) = p_{\lambda^{(j)}}(y_t|C = k)\ p_{\lambda^{(j)}}(C = k) \,/\, \sum_{c\in\{1,2\}} p_{\lambda^{(j)}}(y_t|c)\ p_{\lambda^{(j)}}(c)$. To execute the M step we then formulate the simpler maximum entropy problem with linear constraints, as in (20) and (21), obtaining

$$\begin{aligned}
&\max_{p(x)} \quad H(X) \;=\; H(C) + H(Y|C) \\
&\text{subject to} \quad \int_{x\in\mathcal{X}} \delta_k(c)\ \ p(x)\,\mu(dx) \;=\; \eta_0^{k,(j)} \\
&\qquad\qquad \int_{x\in\mathcal{X}} y\,\delta_k(c)\ \ p(x)\,\mu(dx) \;=\; \eta_1^{k,(j)} \\
&\qquad\qquad \int_{x\in\mathcal{X}} y^2\,\delta_k(c)\ \ p(x)\,\mu(dx) \;=\; \eta_2^{k,(j)} \qquad \text{for } k = 1, 2
\end{aligned} \tag{28}$$

where $x = (y, c)$. Similarly to Section 2.1, we can solve this ME problem analytically and avoid the use of GIS/IIS in performing the M step. That is, for problem (28) we can directly obtain the unique log-linear solution $p(x) = p(y, c)$, where $p(c) = \frac{1}{T}\sum_{t=1}^{T}\rho_t^{c,(j)}$ and $p(y|c) = N(y; \mu_c, \sigma_c^2)$ with $\mu_c = \sum_{t=1}^{T} y_t\rho_t^{c,(j)} \,/\, \sum_{t=1}^{T}\rho_t^{c,(j)}$ and $\sigma_c^2 = \sum_{t=1}^{T}(y_t - \mu_c)^2\rho_t^{c,(j)} \,/\, \sum_{t=1}^{T}\rho_t^{c,(j)}$ for $c = 1, 2$. We then set $p_{\lambda^{(j+1)}} = p$ and repeat until convergence.

Thus, EM-IS produces a model that has the form of a Gaussian mixture. In this case, LME is more general than Jaynes' ME principle, because it can postulate a bi-modal distribution over the observed component

Y, whereas standard ME is reduced to producing a uni-modal Gaussian in this situation. Interestingly, the update formula we obtain for $p_{\lambda^{(j)}} \rightarrow p_{\lambda^{(j+1)}}$ is equivalent to the standard EM update for estimating Gaussian mixture distributions. In fact, we find that in many natural situations, EM-IS recovers standard EM updates as a special case. However, it turns out that there are other situations where EM-IS yields new iterative update procedures that converge faster than standard parameter estimation formulas. We demonstrate both cases in Section 7 below.

We now establish the key result that EM-IS is guaranteed to converge to a feasible LME solution for log-linear models.

4.3 Proof of correctness

To prove that EM-IS converges to log-linear models that are feasible solutions of the LME principle (3), Theorem 1 can be exploited to reduce this question to showing that EM-IS converges to a critical point of the log-likelihood function. The convergence proof for EM-IS then becomes similar to that for the GEM algorithm [43].

Theorem 3 *The EM-IS algorithm monotonically increases the likelihood function $L(\lambda)$, and all limit points of any EM-IS sequence $\{\lambda^{(j+s/K)}, j \geq 0\}$, $s = 1, ..., K$, belong to the set*

$$\Theta \quad = \quad \left\{ \lambda \in \Re^N : \frac{\partial L(\lambda)}{\partial \lambda} = 0 \right\} \tag{29}$$

Therefore, EM-IS asymptotically yields feasible solutions to the LME principle for log-linear models.

Proof: As discussed in the previous section, it is obvious that if the EM-IS algorithm converges to a local maximum in likelihood, it yields a feasible solution of the LME principle by Theorem 1. To prove the convergence, we first show that EM-IS is a generalized EM procedure. To do this we define the auxiliary function A in the same way as in [4, 19]. 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 $A(\lambda, \lambda', \lambda^{(j)})$.

$$\begin{aligned}
Q\left(\lambda, \lambda^{(j)}\right) - Q\left(\lambda', \lambda^{(j)}\right) \quad &= \quad \sum_{i=1}^{N} (\lambda_i - \lambda_i') \left(\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)\, p_{\lambda^{(j)}}(z|y)\, \mu(dz) \right) - \log \left(\frac{\Phi_\lambda}{\Phi_{\lambda'}} \right) \\
&\geq \quad \sum_{i=1}^{N} (\lambda_i - \lambda_i') \left(\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)\, p_{\lambda^{(j)}}(z|y)\, \mu(dz) \right) + 1 - \frac{\Phi_\lambda}{\Phi_{\lambda'}} \\
&= \quad \sum_{i=1}^{N} (\lambda_i - \lambda_i') \left(\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)\, p_{\lambda^{(j)}}(z|y)\, \mu(dz) \right) + 1 \\
&\quad\quad - \int_{x \in \mathcal{X}} e^{\sum_{i=1}^{N} (\lambda_i - \lambda_i')\, f_i(x)}\, p_{\lambda'}(x)\, \mu(dx) \\
&\geq \quad \sum_{i=1}^{N} (\lambda_i - \lambda_i') \left(\sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_i(x)\, p_{\lambda^{(j)}}(z|y)\, \mu(dz) \right) + 1 \\
&\quad\quad - \int_{x \in \mathcal{X}} p_{\lambda'}(x) \sum_{i=1}^{N} \frac{f_i(x)}{f(x)} e^{(\lambda_i - \lambda_i') f(x)}\, \mu(dx) \\
&= \quad A(\lambda, \lambda', \lambda^{(j)})
\end{aligned} \tag{30}$$

where the inequalities follow from the convexity of $-\log$ and exp.

Now let s be the index of one cycle of a full parallel update of λ and assume we perform K cycles of full parallel updates, $s = 1, ..., K$. Then from Equation (30), we have

$$Q\left(\lambda^{(j+s/K)}, \lambda^{(j)}\right) - Q\left(\lambda^{(j+(s-1)/K)}, \lambda^{(j)}\right) \quad \geq \quad A\left(\lambda^{(j+s/K)}, \lambda^{(j+(s-1)/K)}, \lambda^{(j)}\right)$$

for each s. It is true by inspection that $A\left(\lambda^{(j+(s-1)/K)}, \lambda^{(j+(s-1)/K)}, \lambda^{(j)}\right) = 0$ and $A\left(\lambda, \lambda^{(j+(s-1)/K)}, \lambda^{(j)}\right)$ is concave in λ. Moreover, the new update $\lambda^{(j+s/K)}$ is the stationary point of $A\left(\lambda, \lambda^{(j+(s-1)/K)}, \lambda^{(j)}\right)$. Therefore, we have the result that $A\left(\lambda^{(j+s/K)}, \lambda^{(j+(s-1)/K)}, \lambda^{(j)}\right) > 0$, and each step of this procedure increases Q. Thus the EM-IS algorithm monotonically increases the likelihood function $L(\lambda)$.

Next, to show the convergence of $\{\lambda^{(j+s/K)}, j \geq 0\}$, $s = 1, ..., K$, to the stationary points of the likelihood function, we first show the convergence of $\{\lambda^{(j)}, j \geq 0\}$ when we just consider successive phases at the stage $s = 0$. By [43, Theorem 1], we must show that:

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

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

First, under the compactness condition [43, (6)] and continuity condition [43, (10)], assertion (i) can be verified directly using $\lambda \in \mathcal{R}^N$. Second, to establish assertion (ii), it can be shown that $\partial Q(\lambda, \lambda^{(j)})/\partial\lambda = \partial A(\lambda, \lambda', \lambda^{(j)})/\partial\lambda$. Therefore, if $\lambda^{(j)} \notin \Theta$, then $\partial L(\lambda)/\partial\lambda \neq 0$, which implies that $\partial Q(\lambda, \lambda^{(j)})/\partial\lambda \neq 0$ and hence $\partial A(\lambda, \lambda', \lambda^{(j)})/\partial\lambda \neq 0$. So if $\lambda^{(j)} \notin \Theta$, we cannot be at a maximum of A. Therefore, 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.

Finally, to show the convergence of $\{\lambda^{(j+s/K)}, j \geq 0\}$ for the cases of $s = 1, ..., K-1$ respectively, we argue similarly to above. Therefore, we conclude that all limit points of any EM-IS sequence $\{\lambda^{(j+s/K)}, j \geq 0\}$ for $s = 0, ..., K-1$ belong to the set Θ. ∎

Appendix A gives a detailed characterization of the information geometry of EM-IS that provides further insight into its behavior, as well as the behavior of EM and IS algorithms more generally.

5 Finding high-entropy solutions

We can now exploit the EM-IS algorithm to develop a practical approximation to the LME principle. As noted in Section 3.1 it is difficult to solve for an optimal latent maximum entropy model in general. In fact, Section 3.2 points out that it is hard to solve for an optimal LME model even if we restrict our attention to log-linear models. However, the EM-IS algorithm of Section 4 provides an effective technique for finding *feasible*, but not necessarily optimal solutions of the LME principle. (Appendix A illustrates how there can be multiple distinct feasible solutions in general.) Our approach to using EM-IS to approximate the LME principle then is very simple: we first generate several candidate feasible solutions by running EM-IS to convergence from different initial points $\lambda^{(0)}$, then evaluate the entropy of each candidate model, and finally select the model that has the highest entropy.

ME-EM-IS algorithm:

Initialization: Randomly choose initial guesses for the parameters λ.

EM-IS: Run EM-IS to convergence, to obtain a feasible solution λ^*.

Entropy calculation: Calculate the entropy of p_{λ^*}.

Model selection: Repeat the above steps several times to produce a set of distinct feasible candidates. Choose as the final estimate the candidate that achieves the highest entropy. ∎

An apparent difficulty in implementing this algorithm is that one needs to calculate the entropies of the candidate models produced by EM-IS. One might suppose that the entropy has to be calculated explicitly for each candidate model by evaluating the expectation

$$H(p_\lambda) \quad = \quad \int_{x\in\mathcal{X}} p_\lambda(x) \log p_\lambda(x)\, \mu(dx) \quad = \quad -\log(\Phi_\lambda) \; + \; \sum_{i=1}^{N} \lambda_i \int_{x\in\mathcal{X}} f_i(x)\; p_\lambda(x)\, \mu(dx) \tag{31}$$

However, it turns out that we do not need to perform this calculation explicitly. In fact, we can easily recover the entropy of a feasible log-linear model merely as a byproduct of running EM-IS to convergence. Recall the decomposition from (15) that $L(\lambda) = Q(\lambda, \lambda') + H(\lambda, \lambda')$ for all λ', where $Q(\lambda, \lambda')$ and $H(\lambda, \lambda')$ are given by (16) and (17) respectively. In the case where λ is a feasible solution according to (3) (and hence (29)) we obtain the following relationship.

Theorem 4 *If λ is in the set of feasible solutions, i.e., $\lambda \in \Theta$ as defined by (29), then*

$$\begin{aligned} Q(\lambda, \lambda) &= -H(p_\lambda) \\ L(\lambda) &= -H(p_\lambda) + H(\lambda, \lambda) \end{aligned} \tag{32}$$

Proof: By (15), we know that $L(\lambda) = Q(\lambda, \lambda) + H(\lambda, \lambda)$ for all $\lambda \in \Theta$. Let $\lambda^{(j+1)} = \arg\max_\lambda Q(\lambda, \lambda^{(j)})$. Then from (23) we obtain $Q(\lambda^{(j+1)}, \lambda^{(j)}) = \max_\lambda Q(\lambda, \lambda^{(j)}) = -H^*(\lambda^{(j)})$. Now, using the same argument as in the proof of Theorem 3, we can show that all limit points of the sequence $\{\lambda^{(j+1)}, j \geq 0\}$ belong to the set Θ, and therefore $Q(\lambda, \lambda) = -H(p_\lambda)$ for all $\lambda \in \Theta$. Thus we have $L(\lambda) = -H(p_\lambda) + H(\lambda, \lambda)$ for all $\lambda \in \Theta$. ∎

This theorem provides the needed result for establishing the latter half of Theorem 1 in Section 3. Interestingly, it also provides a simplification of the entropy calculation, (31), when λ^* is a feasible solution found by EM-IS, because at convergence we will have the relationship $Q(\lambda^*, \lambda^*) = -H(p^*_\lambda)$. All one has to do is calculate $-Q(\lambda^*, \lambda^*)$ for a given feasible solution $\lambda^* \in \Theta$, since combining (19) with (24) we have

$$H(p_{\lambda^*}) \;=\; -Q(\lambda^*, \lambda^*) \;=\; \log(\Phi_{\lambda^*}) \;-\; \sum_{i=1}^{N} \lambda_i^* \eta_i^*$$

Therefore, the entropy of p_{λ^*} can be easily determined: the η_i^* values for $i = 1, ..., N$ are already calculated in the E step of EM-IS (24), and the normalization constant Φ_{λ^*} already needs to have been determined as part of the M step for solving (26).

There are a few other observations that follow from Theorem 4. First, note that in the special case where there is no missing data, i.e., $X = Y$, we have $H(\lambda, \lambda) = 0$ and Theorem 4 shows that $L(\lambda) = -H(p_\lambda)$ for a feasible solution $\lambda \in \Theta$; a well known result of standard maximum entropy theory [4, 19]. One can also draw a clear distinction between the LME and MLE principles from (32). Assume the term $H(\lambda, \lambda)$ is constant for different feasible solutions. In this case MLE (which maximizes likelihood) will choose the model that has lowest entropy, whereas LME (which maximizes entropy) will choose the model that has least likelihood. Of course, $H(\lambda, \lambda)$ will not be constant among different feasible λ in practice and the comparison between MLE and LME is not so straightforward, but this example does highlight their difference. The difference between these two principles raises the question of which method is the most effective when inferring a model from sample data. To address this question we turn to a brief experimental comparison of LME and MLE.

6 An experimental comparison

We conducted a simple experiment to ascertain whether LME or MLE yields better estimates when inferring models from sample data that has missing components. We considered a simple three component mixture model as a case study, where the mixing component C is unobserved but a two dimensional vector $Y \in \Re^2$ is observed. Thus, the features (sufficient statistics) we try to match in the data are the same as in Sections 3.3 and 4.2, except that in this case there are three rather than two mixture components and the observed data Y is two dimensional rather than one dimensional. Given sample data $\tilde{\mathcal{Y}} = (y_1, ..., y_T)$ the idea is to infer a log-linear model $p(x) = p(y, c)$ such that $c \in \{1, 2, 3\}$.

The basis for comparison between LME and MLE is to realize that by the discussion in Section 3.3, any feasible solution to the LME principle (11) corresponds to a locally maximum likelihood Gaussian mixture as specified by (14). Therefore, we can implement EM-IS as outlined in Section 4.2 and generate feasible candidates for the LME and MLE principles simultaneously (although as noted in Section 4.2, EM-IS reduces to the standard EM algorithm for estimating Gaussian mixtures in this case). From Theorem 1 we know that LME and MLE consider the same set of feasible candidates, except that among feasible solutions, LME selects the model with the highest entropy, whereas MLE selects the model with the highest likelihood. Theorem 4 shows that these are not equivalent.

We are interested in determining which method yields better estimates of various underlying models p^* used to generate the data. We measure the quality of an estimate p_λ by calculating the *cross entropy* from the correct marginal distribution $p^*(y)$ to the estimated marginal distribution $p_\lambda(y)$ on the *observed* data component Y

$$D(p^*(y)\|p_\lambda(y)) \quad = \quad \int_{y\in\mathcal{Y}} p^*(y) \log \frac{p^*(y)}{p_\lambda(y)} \; \mu(dy)$$

The goal is to minimize the cross entropy between the marginal distribution of the estimated model p_λ and the correct marginal p^*. A cross entropy of zero is obtained only when $p_\lambda(y)$ matches $p^*(y)$.

We consider a series of experiments with different models and different sample sizes to test the robustness of both LME and MLE to sparse training data, high variance data, and deviations from log-linearity in the underlying model. In particular, we used the following experimental design.

1. Fix a generative model $p^*(x) = p^*(y, c)$.
2. Generate a sample of observed data $\tilde{\mathcal{Y}} = (y_1, ..., y_T)$ according to $p^*(y)$.
3. Run EM-IS to generate multiple feasible solutions by restarting from 300 random initial vectors λ. We generated initial vectors λ by generating mixture weights θ_c from a uniform prior, and independently generating each component of the mean vectors μ_c and covariance matrices σ_c^2 by choosing numbers uniformly from $\{-4, -2, 0, 2, 4\}$ (see Section 4.2 for the relation between the $\theta_c, \mu_c, \sigma_c^2$ parameters and λ).
4. Calculate the entropy and likelihood for each feasible candidate.
5. Select the maximum entropy candidate $p_{\scriptscriptstyle LME}$ as the LME estimate, and the maximum likelihood candidate $p_{\scriptscriptstyle MLE}$ as the MLE estimate.
6. Calculate the cross entropy from $p^*(y)$ to the marginals $p_{\scriptscriptstyle LME}(y)$ and $p_{\scriptscriptstyle MLE}(y)$ respectively.
7. Repeat Steps 2 to 6 500 times and compute the average of the respective cross entropies. That is, average the cross entropy over 500 repeated trials for each sample size and each method, in each experiment.
8. Repeat Steps 2 to 7 for different sample sizes T.
9. Repeat Steps 1 to 8 for different generative models $p^*(x)$.

Scenario 1

In the first experiment, we generated the data according to a three component Gaussian mixture model that has the form expected by the estimators. Specifically, we used a uniform mixture distribution $\theta_c = \frac{1}{3}$ for $c = 1, 2, 3$, where the component Gaussians were specified by the mean vectors $\begin{bmatrix} 0 \\ -3 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \end{bmatrix}$ and covariance matrices $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ respectively.

Figures 2 and 3 first show that the average log-likelihoods and average entropies of the models produced by LME and MLE respectively behave as expected. MLE clearly achieves higher log-likelihood than LME, however LME clearly produces models that have significantly higher entropy than MLE. The interesting outcome is that the two estimation strategies obtain significantly different cross entropies. Figure 4 reports the average cross entropy obtained by MLE and LME as a function of sample size, and shows the somewhat surprising result that LME achieves substantially lower cross entropy than MLE. LME's advantage

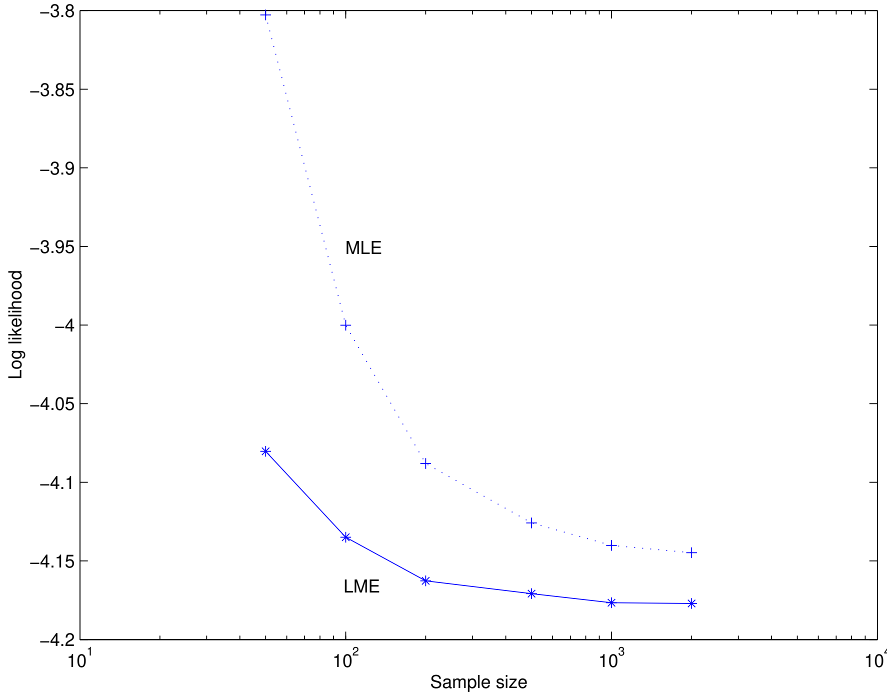


Figure 2: Average log-likelihood of the MLE estimates versus the LME estimates in Experiment 1.

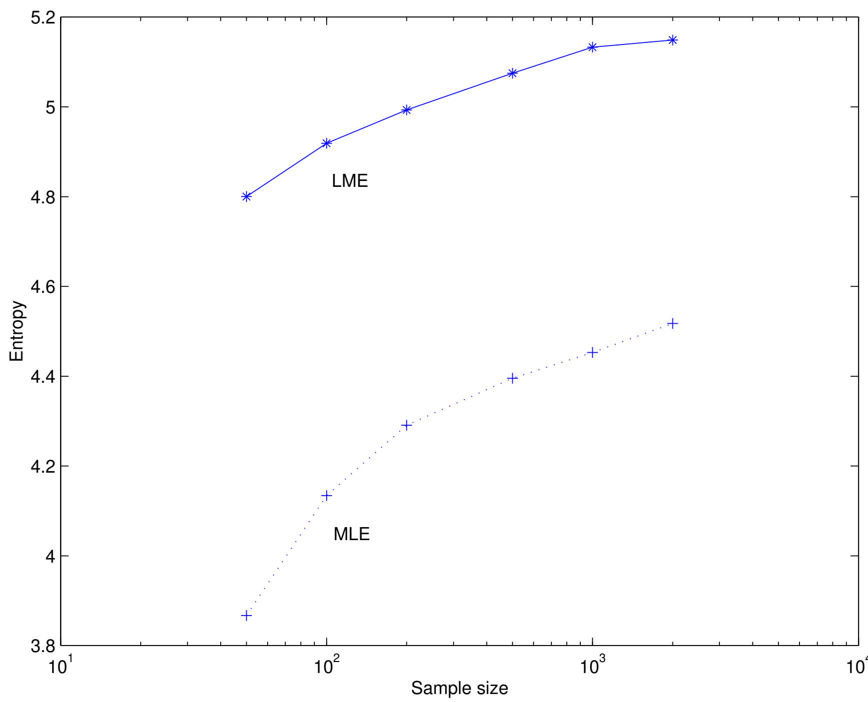
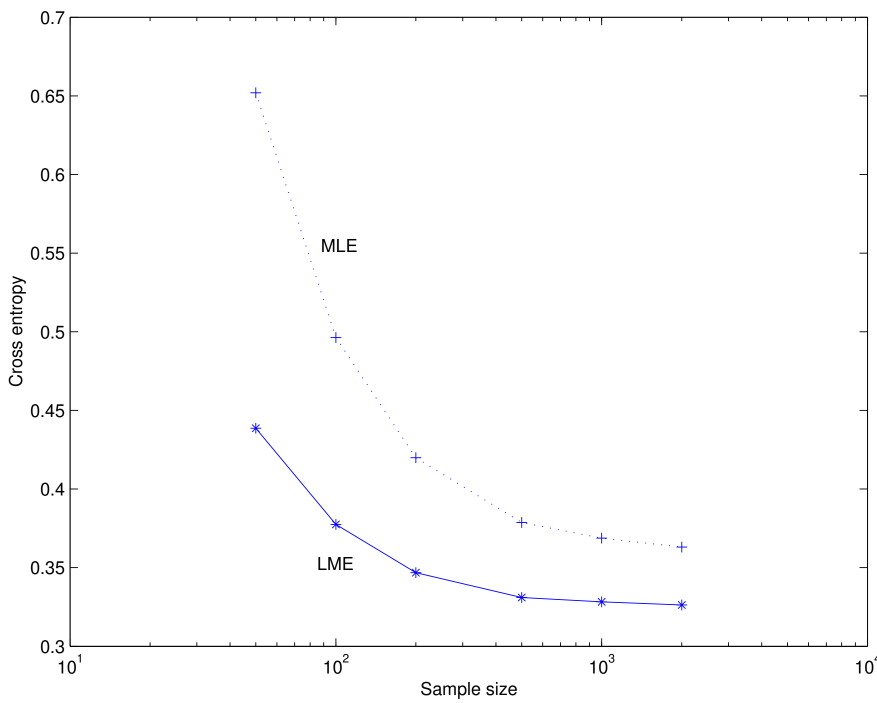


Figure 3: Average entropy of the MLE estimates versus the LME estimates in Experiment 1.



sample size	50	100	200	500	1000	2000	5000	10000
MLE	0.6520	0.4963	0.4199	0.3788	0.3688	0.3631		
LME	0.4386	0.3775	0.3468	0.3310	0.3285	0.3264		

Figure 4: Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 1.

is especially pronounced at small sample sizes, and persists even when sample sizes as large as 10,000 are considered (Figure 4).

Although one might have expected an advantage for LME because of a "regularization" effect, this does not completely explain LME's superior performance at large sample sizes. However, before discussing the regularization properties of LME in detail, let us first consider alternative scenarios where the observed relationship between MLE and LME is different. This first experiment considered a favorable scenario where the underlying generative model p^* has the same form as the distributional assumptions made by the estimators. We next consider situations where these structural assumptions are violated.

Scenario 2

In our second experiment we used a generative model that was a mixture of five Gaussian distributions over $\Re^2$. Specifically, we generated data by sampling from a uniform distribution over mixture components $\theta_c = \frac{1}{5}$ for $c = 1, ..., 5$, and then generated the observed data $Y \in \Re^2$ by sampling from the corresponding Gaussian distribution, where these distributions had means $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\begin{bmatrix} -2 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ -2 \end{bmatrix}$ and covariances $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ respectively. The LME and MLE estimators still only inferred three component mixtures in this case, and hence were each making an incorrect assumption about the underlying model.

Figure 5 shows that LME still obtained a significantly lower cross entropy than MLE at small sample sizes, but lost its advantage at larger sample sizes. At a crossover point of $T = 1000$ data points, MLE began to produce slightly better estimates than LME, but only marginally so. Overall, LME still appears to be a safer estimator for this problem, but it is not uniformly dominant.

Scenario 3

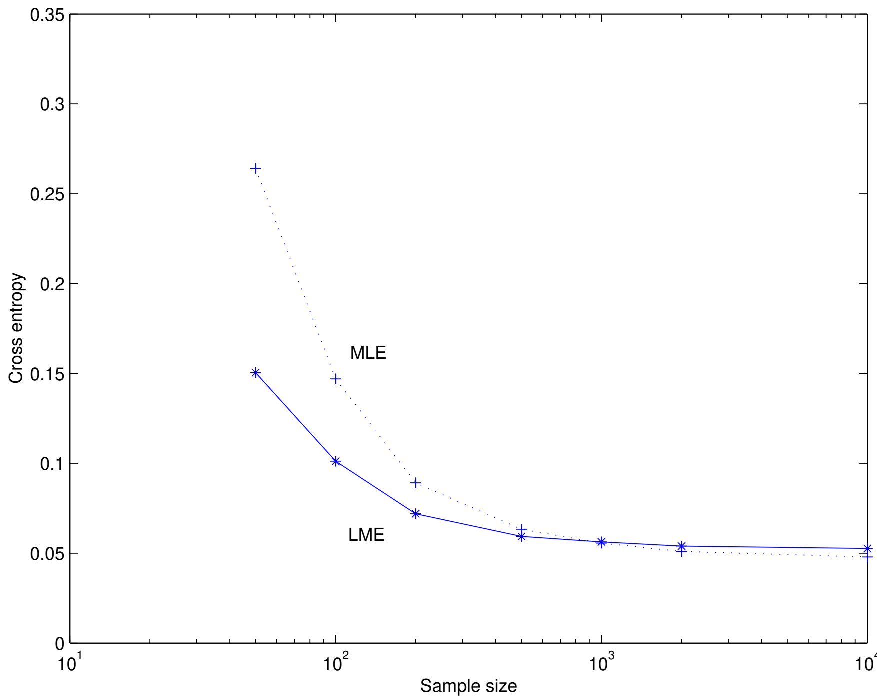
Our third experiment attempted to test how robust the estimators were to high variance data generated by a heavy tailed distribution. This experiment yielded our most dramatic results. We generated data according to a three component mixture (which was correctly assumed by the estimators) but then used a Laplacian distribution instead of a Gaussian distribution to generate the Y observations. This model generated data that was much more variable than data generated by a Gaussian mixture, and challenged the estimators significantly. The specific parameters we used in this experiment were $\theta_c = \frac{1}{3}$ for $c = 1, 2, 3$, and means $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$ and "covariances" $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ for the Laplacians.

Figure 6 shows that LME produces significantly better estimates than MLE in this case, and even improved its advantage at larger sample sizes. Clearly, MLE is not a stable estimator when subjected to heavy tailed data when this is not expected. LME proves to be far more robust in such circumstances and clearly dominates MLE.

Scenario 4

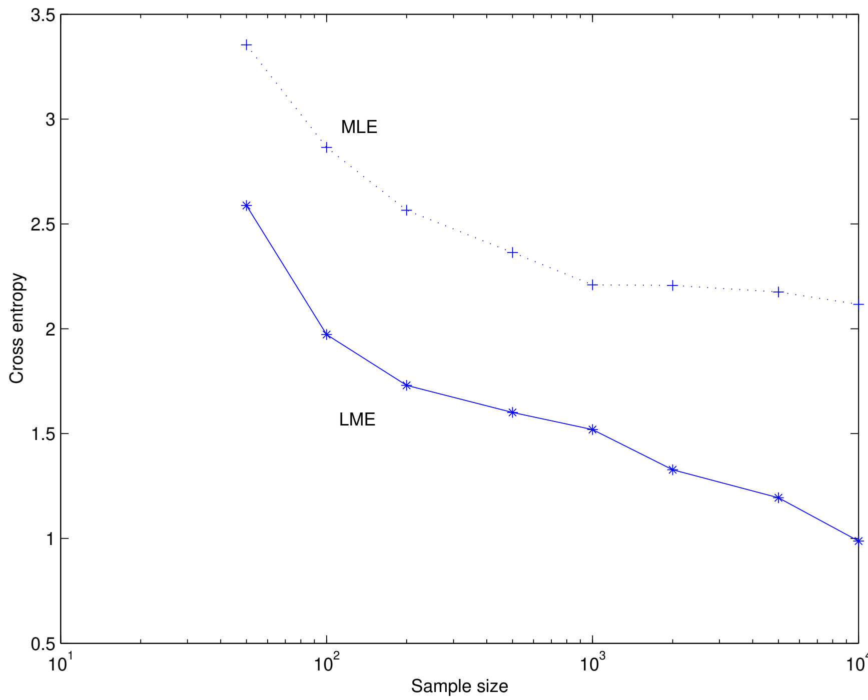
However, there are other situations where MLE appears to be a slightly better estimator than LME when sufficient data is available. Figure 7 shows the results of subjecting the estimators to data generated from a three component Gaussian mixture, $\theta = \frac{1}{3}$, $c = 1, 2, 3$, with means $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$ and covariances $\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ respectively. In this case, LME still retains a sizeable advantage at small sample sizes, but after a sample size of $T = 500$, MLE begins to demonstrate a persistent although modest advantage.

Overall, these results suggest that maximum likelihood estimation (MLE) is effective at large sample sizes, so long as the presumed model is close to the underlying data source. If there is a mismatch between the assumption and reality however, or if there is limited training data, then LME appears to offer a significantly



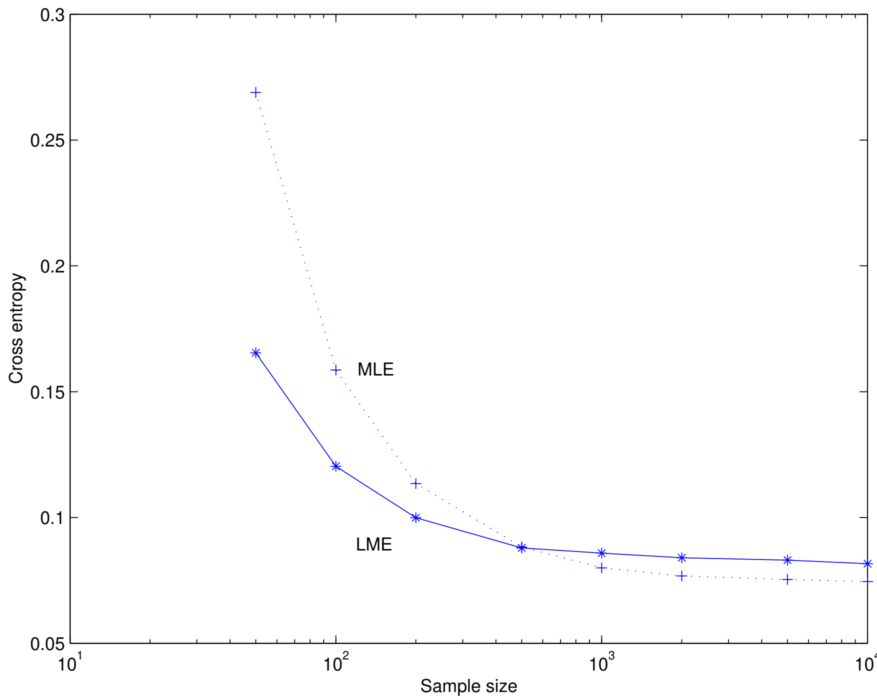
sample size	50	100	200	500	1000	2000	5000	10000
MLE	0.2641	0.1470	0.0892	0.0633	0.0557	0.0510		
LME	0.1505	0.1012	0.0719	0.0594	0.0563	0.0540		

Figure 5: Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 2.



sample size	50	100	200	500	1000	2000	5000	10000
MLE	3.3539	2.8648	2.5654	2.3634	2.2098	2.2067	2.1751	2.1164
LME	2.5881	1.9723	1.7301	1.6009	1.5196	1.3276	1.1941	0.9871

Figure 6: Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 3.



sample size	50	100	200	500	1000	2000	5000	10000
MLE	0.2689	0.1586	0.1135	0.0883	0.0800	0.0768	0.0754	0.0745
LME	0.1654	0.1203	0.0999	0.0879	0.0858	0.0851	0.0840	0.0816

Figure 7: Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Experiment 4.

safer and more effective alternative. Of course, these results are far from definitive, and further experimental and theoretical analysis is required to give completely authoritative answers.

A few comments are in order. It appears that LME adds more than just a fixed regularization effect to MLE. In fact, as we demonstrate in Section 8.1 below, one can add a regularization term to the LME principle in the same way one can add a regularization term to the MLE principle. LME behaves more like an *adaptive* rather than fixed regularizer, because we see no real under-fitting from LME on large data samples, even though LME chooses far "smoother" models than MLE at smaller sample sizes. In fact, LME can demonstrate a far stronger regularization effect than any standard penalization method: In the well known case where EM-IS converges to a degenerate solution (i.e., such that the determinant of the covariance matrix goes to zero) no finite penalty can counteract the resulting unbounded likelihood. However, the LME principle can automatically filter out degenerate models, because such models have zero entropy and any non-degenerate model will be preferred. Eliminating degenerate models by the LME principle solves one of the main practical problems with Gaussian mixture estimation.

Another observation is that all of our experiments show that MLE and LME reduce cross entropy error when the sample size is increased. However, we have not yet proved that the LME principle is statistically consistent; that is, that it is guaranteed to converge to zero cross entropy in the limit of large samples—when the underlying model has a log-linear form in the same features considered by the estimator. We are actually interested in a stronger form of consistency that requires the estimator to converge to the best *representable* log-linear model (i.e., the one with minimum cross entropy error) for any underlying distribution, even if the minimum achievable cross entropy is nonzero. Determining the statistical consistency of LME, in either sense, remains an important topic for future research.

7 Application to other models

Clearly the LME principle is more general than Gaussian mixture models. In this section we demonstrate how LME can be applied to other important estimation problems involving latent variables. Our aim in this section is not to present a full fledged study of each problem, but to merely illustrate how the LME principle can be applied in each case. Specifically, we focus on the application of the EM-IS algorithm to finding feasible solutions, and point out cases where it yields faster converging algorithms than standard maximum likelihood training algorithms.

7.1 Mixtures of Dirichlet distributions

The first model we consider is a mixture of Dirichlet distributions, which has applications in natural language modeling and other areas [6, 32]. In this problem, the observed data has the form of an M dimensional probability vector $y = (y_1, ..., y_M)$ such that $0 \le y_\ell \le 1$ for $\ell = 1, ..., M$ and $\sum_{\ell=1}^{M} y_\ell = 1$. That is, the observed variable is a random vector $Y = (Y_1, ..., Y_M) \in [0,1]^M$ which happens to be normalized. There is also an underlying class variable $C \in \{1, 2\}$ that is unobservable. Let $X = (Y, C)$. Given an observed sequence of T M-dimensional probability vectors $\tilde{\mathcal{Y}} = (y^1, ..., y^T)$, where $y^t = (y_1^t, ..., y_M^t)$ for $t = 1, ..., T$, we attempt to infer a latent maximum entropy model that matches expectations on the features $f_0^k(x) = \delta_k(c)$ and $f_\ell^k(x) = (-\log y_\ell)\delta_k(c)$ for $\ell = 1, ..., M$ and $k = 1, 2$, where $x = (y, c)$. In this case, the LME principle can be formulated as

$$\begin{aligned}
\max_{p(x)} \; H(X) \quad &= \quad H(C) + H(Y|C) \\
\text{subject to} \qquad \int_{x\in\mathcal{X}} \delta_k(c)\;\; p(x)\,\mu(dx) \quad &= \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_c \delta_k(c)\;\; p(c|y)\,\mu(dx) \\
\int_{x\in\mathcal{X}} (-\log y_\ell)\,\delta_k(c)\;\; p(x)\,\mu(dx) \quad &= \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_c (-\log y_\ell)\,\delta_k(c)\;\; p(c|y)\,\mu(dx) \\
&\qquad \text{for } \ell = 1, ..., M \text{ and } k = 1, 2
\end{aligned}$$

where $\delta_k(c)$ indicates whether $c = k$ and $\tilde{p}(y) = \frac{1}{T}$. Due to the nonlinear mapping caused by $p(c|y)$ there is no closed form solution. However, as for Gaussian mixtures, we can apply EM-IS to obtain a feasible log-linear model for this problem. To perform the E step, one can calculate the feature expectations according to (24)

$$\eta_0^{k,(j)} = \frac{1}{T}\sum_{t=1}^{T}\sum_{c\in\{1,2\}} \delta_k(c)\, \rho_t^{k,(j)}$$

$$\eta_\ell^{k,(j)} = \frac{1}{T}\sum_{t=1}^{T}\sum_{c\in\{1,2\}} (-\log y_\ell^t)\, \delta_k(c)\, \rho_t^{k,(j)} \qquad \text{for } \ell = 1,...,M \text{ and } k = 1,2$$

where $\rho_t^{k,(j)} = p_{\lambda^{(j)}}(C=k|y^t) = p_{\lambda^{(j)}}(y^t|C=k)\; p_{\lambda^{(j)}}(C=k) \,/\, \sum_{c\in\{1,2\}} p_{\lambda^{(j)}}(y^t|c)\; p_{\lambda^{(j)}}(c)$. Note that these expectations can be calculated efficiently, like the Gaussian mixture case.

To execute the M step we then formulate the simpler maximum entropy problem with linear constraints, as in (20) and (21), to obtain

$$\begin{aligned} \max_{p(x)} \; H(X) &= H(C) + H(Y|C) \\ \text{subject to} \quad \int_{x\in\mathcal{X}} \delta_k(c)\;\, p(x)\, \mu(dx) &= \eta_0^{k,(j)} \\ \int_{x\in\mathcal{X}} (-\log y_\ell)\, \delta_k(c)\;\, p(x)\, \mu(dx) &= \eta_\ell^{k,(j)} \qquad \text{for } \ell = 1,...,M \text{ and } k = 1,2 \end{aligned}$$

For this problem we can obtain a log-linear solution of the form $p(x) = p(y,c)$ where $p(c) = \frac{1}{T}\sum_{t=1}^{T}\rho_k^t$ and the class conditional model $p(y|c)$ is a Dirichlet distribution with parameters $\alpha_\ell^c = 1 - \lambda_\ell^c$; that is $p(y|c) = \Gamma\left(\sum_{\ell=1}^{M}\alpha_\ell^c\right)\left(\prod_{\ell=1}^{M}\Gamma(\alpha_\ell^c)\right)^{-1}\prod_{\ell=1}^{M} y_\ell^{\alpha_\ell^c-1}$. However, we still need to solve for the Lagrange multipliers λ_ℓ^c by using iterative scaling. (This is unlike the Gaussian mixture case where we could solve for the Lagrange multipliers directly.) As specified in (25), the update parameters $\gamma^{(j+s/K)}$ to be added to $\lambda_\ell^{c,(j+(s-1)/K)}$ in iterative scaling phase s of M step j are obtained by solving the equations

$$\int_{y\in\mathcal{Y}} (\log y_\ell)\left(\prod_{m=1}^{M} y_m\right)^{-\gamma_\ell^{(j+s/K)}} p_{\lambda^{(j+(s-1)/K)}}(y|C=k)\;\mu(dy) \;=\; \eta_\ell^{k,(j)}$$

for $\ell = 1,...,M$ and $k = 1,2$. This is equivalent to solving for $\gamma^{(j+s/K)}$ in

$$\Psi\left(1 - \lambda_\ell^{c,(j+(s-1)/K)} - \gamma_\ell^{(j+s/K)}\right) - \Psi\left(\sum_{m=1}^{M}\left(1 - \lambda_m^{c,(j+(s-1)/K)} - \gamma_\ell^{(j+s/K)}\right)\right) = \eta_\ell^{k,(j)}$$

for $\ell = 1,...,M$ and $k = 1,2$, where Ψ is the digamma function. The solution can be obtained by iterating the fixed-point equations

$$\Psi\left(1 - \lambda_\ell^{c,(j+(s-1)/K)} - \gamma_\ell^{(j+s/K)}\right) = \Psi\left(\sum_{m=1}^{M}\left(1 - \lambda_m^{c,(j+(s-1)/K)} - \gamma_\ell^{(j+s/K)}\right)\right) + \eta_\ell^{k,(j)}$$

for $\ell = 1,...,M$ and $k = 1,2$. This iteration corresponds to a well known technique for locally maximizing the likelihood of a Dirichlet mixture [34]. Thus, EM-IS recovers a classical training algorithm as a special case.

7.2 Hidden Markov models

Assume the observed data is a sequence $\tilde{\mathcal{Y}} = (y_1,...,y_t,...,y_T)$ where each y_t takes value on a finite set of integer values from $\mathcal{Y} = \{1,...,V\}$. Also assume that the observed sequence is related to a hidden state sequence $(s_1,...,s_t,...,s_T)$, where s_t takes value on a finite set of integer values from $\mathcal{S} = \{1,...,L\}$. Thus we have the complete data $X = (Y,S)$ where Y and S are random sequences $Y = (Y_1,...,Y_T)$, $S = (S_1,...,S_T)$ respectively.

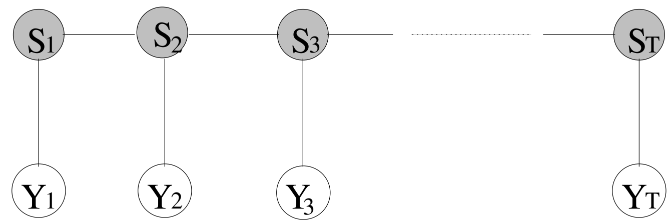


Figure 8: Hidden Markov model

The LME approach to estimation does not make conditional independence assumptions directly, but instead specifies a set of features whose expectations we force the model to satisfy. In this case we assume the features are defined over contiguous pairs of hidden state variables and concurrent observed and hidden state variables. That is, we consider the features $f_{kv}(y,s) = \delta_v(y)\delta_k(s)$ and $f_{k\ell}(s,s') = \delta_k(s)\delta_\ell(s')$ for $v = 1,...,V$, $k = 1,...,L$, and $\ell = 1,...,L$, where s and y are values for concurrent hidden state and observed variables respectively, and s and s' are values for consecutive hidden state variables. Note that all features are binary and therefore one can represent the structure of the log-linear model by a graph, where nodes depict variables and edges connect each pair of variables that co-occur in some feature function. The resulting graphical structure, illustrated in Figure 8, corresponds to the standard graphical representation of a hidden Markov model (HMM).

Now, given a sequence of observed data $\tilde{\mathcal{Y}} = (y_1,...,y_T)$ we can formulate the LME principle as

$$\begin{aligned}
\max_{p(x)} \; H(X) &= H(S) + H(Y|S) \\
\text{subject to} \quad \sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} \delta_v(y)\,\delta_k(s)\;\; p(y,s) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \delta_v(y_t)\,\delta_k(s)\;\; p(s|y_t) \\
\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \delta_k(s)\,\delta_\ell(s')\;\; p(s,s') &= \sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \delta_k(s)\delta_\ell(s')\;\; p(s,s'|\tilde{\mathcal{Y}}) \\
&\quad \text{for } k=1,...,L,\; \ell=1,...,L,\; v=1,...,V
\end{aligned}$$

where $x = (y_1,...,y_T,s_1,...,s_T)$. To find feasible log-linear solutions to this problem, one can apply the EM-IS algorithm. First, the E step can be performed by calculating feature expectations according to (24)

$$\begin{aligned}
\eta_{kv}^{(j)} &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in\mathcal{S}} \delta_v(y_t)\,\delta_k(s)\;\; p_{\lambda^{(j)}}(s|y_t) \\
\eta_{k\ell}^{(j)} &= \sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \delta_k(s)\,\delta_\ell(s')\;\; p_{\lambda^{(j)}}(s,s'|\tilde{\mathcal{Y}})
\end{aligned}$$

for $k = 1,...,L$, $\ell = 1,...,L$ and $v = 1,...,V$. Below we will prove that the log-linear model is equivalent to a standard HMM, and therefore these terms can be efficiently calculated by the forward-backward algorithm.

To execute the M step, we then formulate the simpler maximum entropy problem with linear constraints, as in (20) and (21)

$$\begin{aligned}
\max_{p(x)} \; H(X) &= H(S) + H(Y|S) \\
\text{subject to} \quad \sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{S}} \delta_v(y)\,\delta_k(s)\;\; p(y,s) &= \eta_{kv}^{(j)} \\
\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \delta_k(s)\,\delta_\ell(s')\;\; p(s,s') &= \eta_{k\ell}^{(j)} \\
\text{for } k=1,...,L,\; \ell=1,...,L,\; v=1,...,V &
\end{aligned}$$

This problem can be solved by considering the equivalent problem of finding the parameters λ that maximize the likelihood of the observed data (by Theorem 1).

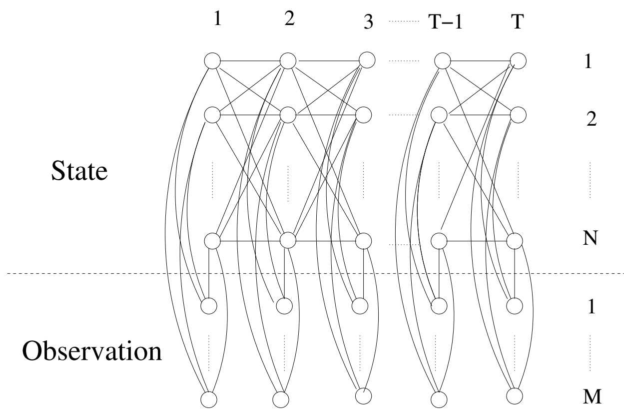


Figure 9: Trellis to calculate the normalization factor and feature expectation, the weight of each link is the exponential of the corresponding Lagrangian multiplier

These computational problems become easier once we prove that the marginal distribution over the hidden state sequence has the structure of a Markov chain. To do this, we use the fact that a log-linear model defines a decomposable graphical model [29]. Following [29], we can rewrite the full joint probability as a product of maximal clique potentials, divided by the potentials of every nonempty intersection of maximal cliques. (Here variables are thought of as nodes in a graph where there is an edge between two variables if they share a feature function.) Accordingly, we obtain

$$\begin{aligned} p(y_1,...,y_T,s_1,...,s_T) &= g_1(s_1,s_2)\prod_{t=1}^{T}\frac{g(y_t,s_t)}{g(s_t)}\prod_{t=2}^{T-1}\frac{g(s_t,s_{t+1})}{g(s_t)} \\ p(s_1,...,s_T) &= \sum_{(y_1,...,y_T)\in\mathcal{Y}^T} p(y_1,...,y_T,s_1,...,s_T) = g(s_1,s_2)\prod_{t=2}^{T-1}\frac{g(s_t,s_{t+1})}{g(s_t)} \end{aligned}$$

This equation shows that the sequence $s_1,...,s_T$ is a first order Markov chain. Therefore, we have shown that the LME principle infers a distribution that has the form of a hidden Markov model in this problem. To perform the M step in EM-IS, note that $f(x) = 2T$ in this case, where $x = (y_1,...,y_T,s_1,...,s_T)$, so we can use Equation (27) to directly calculate the update $\lambda^{(j)} \to \lambda^{(j+1)}$. Contrary to the assertion in [27], the normalization factor

$$\Phi_\lambda = \exp\left(\sum_{k=1}^{L}\sum_{v=1}^{V}\sum_{s\in\mathcal{S}}\sum_{y\in\mathcal{Y}}\lambda_{kv}\,\delta_v(y)\,\delta_k(s) + \sum_{k=1}^{L}\sum_{\ell=1}^{L}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\lambda_{k\ell}\,\delta_k(s)\,\delta_\ell(s')\right)$$

can be calculated efficiently by the sum-product algorithm. That is , summing over all the links at each time slice and passing through the trellis nodes with the product of the weight to the ongoing nodes, the feature expectations can be calculated explicitly as

$$\begin{aligned} \sum_{x\in\mathcal{X}}\delta_v(y_t)\,\delta_k(s_t)\,p(x) &= 1-\sum_{x\in\mathcal{X}}\delta_{\neq v}(y_t)\,\delta_{\neq i}(s_t)\,p(x) \\ \sum_{x\in\mathcal{X}}\delta_k(s_t)\,\delta_\ell(s_{t+1})\,p(x) &= 1-\sum_{x\in\mathcal{X}}\delta_{\neq k}(s_t)\,\delta_{\neq \ell}(s_{t+1})\,p(x) \end{aligned}$$

where the right hand sides can be calculated efficiently by summing over all the links except the corresponding feature at each time slice and passing through the trellis nodes, finally dividing by the normalization factor. (See Figure 9 for an illustration.) Thus the computational complexity of EM-IS in this problem is at the same order as the Baum-Welch algorithm.

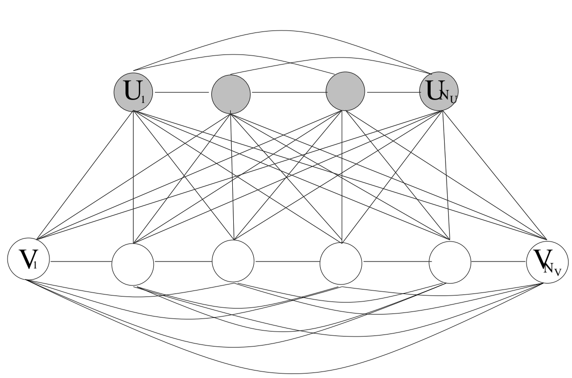


Figure 10: Boltzmann machine model: nodes Y are observable, nodes U are unobservable.

7.3 Boltzmann machines

Consider a graphical model with M binary nodes taking values either 0 or 1. Assume that among these nodes there are J observable nodes $Y = (Y_1, ..., Y_J)$, and $L = M - J$ unobservable nodes $U = (U_1, ..., U_L)$. Let $X = (Y, U)$. Thus, $\mathcal{Y} = \{0,1\}^J$, $\mathcal{U} = \{0,1\}^L$ and $\mathcal{X} = \{0,1\}^{J+L} = \{0,1\}^M$. For this problem, the observed data has the form of a J dimensional vector $y = (y_1, ..., y_J) \in \{0,1\}^J$. Given an observed sequence of T J-dimensional vectors $\mathcal{Y} = (y^1, ..., y^T)$, where $y^t \in \{0,1\}^J$ for $t = 1, ..., T$, we attempt to infer a latent maximum entropy model that matches expectations on features defined between every pair of variables in the model. Specifically, we consider the features $f_{k\ell}(x) = y_k y_\ell$, $f_{km}(x) = y_k u_m$, $f_{mn}(x) = u_m u_n$, for $1 \le k < \ell \le J$ and $1 \le m < n \le L$, where $x = (y, u) = (y_1, ..., y_J, u_1, ..., u_L)$. Note that once again the features are all binary, and therefore we can represent the structure of the log-linear model by a graph, as shown in Figure 10.

Given a sequence of observed data $\tilde{\mathcal{Y}} = (y^1, ..., y^T)$, we formulate the LME principle as

$$\begin{aligned}
\max_{p(x)} H(X) &= H(Y) + H(U|Y) \\
\text{subject to} \qquad \sum_{x\in\mathcal{X}} y_k y_\ell \; p(x) &= \sum_{y\in\tilde{\mathcal{Y}}} y_k y_\ell \; \tilde{p}(y) \\
\sum_{x\in\mathcal{X}} y_k u_m \; p(x) &= \sum_{y\in\tilde{\mathcal{Y}}} y_k \; \tilde{p}(y) \sum_{u\in\{0,1\}^L} u_m \; p(u|y) \\
\sum_{x\in\mathcal{X}} u_m u_n \; p(x) &= \sum_{u\in\{0,1\}^L} u_m u_n \; p(u) \qquad \text{for } 1 \le k < \ell \le J \text{ and } 1 \le m < n \le L
\end{aligned}$$

where $x = (y, u) = (y_1, ..., y_J, u_1, ..., u_L)$. Again we can apply EM-IS to find a feasible log-linear model. To execute the E step, calculate the feature expectations according to (24)

$$\begin{aligned}
\eta^{(j)}_{k,\ell} &= \frac{1}{T} \sum_{t=1}^{T} y^t_k y^t_\ell \\
\eta^{(j)}_{k,m} &= \frac{1}{T} \sum_{t=1}^{T} y^t_k \sum_{u\in\{0,1\}^L} u_m \; p(u|y^t) \\
\eta^{(j)}_{m,n} &= \sum_{u\in\{0,1\}^L} u_m u_n \; p(u) \qquad \text{for } 1 \le k < \ell \le J \text{ and } 1 \le m < n \le L
\end{aligned}$$

To execute the M step we then formulate the simpler maximum entropy problem with linear constraints, as in (20) and (21)

$$\max_{p(x)} H(X) = H(Y) + H(U|Y)$$

$$\begin{aligned}
\text{subject to } \sum_{x\in\mathcal{X}} y_k y_\ell\, p(x) &= \eta^{(j)}_{k,\ell} \\
\sum_{x\in\mathcal{X}} y_k u_m\, p(x) &= \eta^{(j)}_{k,m} \\
\sum_{x\in\mathcal{X}} u_m u_n\, p(x) &= \eta^{(j)}_{m,n} \qquad \text{for } 1\le k<\ell\le J \text{ and } 1\le m<n\le L
\end{aligned}$$

where $x = (y,u) = (y_1, ..., y_J, u_1, ..., u_L)$. In this case, the probability distribution for the complete data model can be written

$$p_\Lambda(x) \;=\; p_\Lambda(u,y) \;=\; \frac{1}{\Phi_\Lambda} e^{\frac{1}{2}y^\top \Lambda_Y y + \frac{1}{2}u^\top \Lambda_U u + y^\top \Lambda_{YU} u} \;=\; \frac{1}{\Phi_\Lambda} e^{\frac{1}{2}x^\top \Lambda x}$$

where $\Lambda = \begin{bmatrix} \Lambda_Y & \Lambda_{YU} \\ \Lambda_{YU} & \Lambda_U \end{bmatrix}$ is the $M\times M$ symmetric matrix of λ parameters corresponding to the features over the variable pairs (with the diagonal elements of Λ equal to zero), and $\Phi_\Lambda = \sum_{x\in\{0,1\}^M} e^{\frac{1}{2}x^\top \Lambda x}$ is the normalization factor. This graphical model corresponds to a Boltzmann machine [1]. To solve for the optimal Lagrange multipliers $\Lambda^{(j)}$ in the M step we once again need to use iterative scaling. Following (25), we iteratively improve $\Lambda^{(j)}$ by adding the update parameters $\gamma^{(j+s/K)}$ that satisfy (26). These can be calculated by by using Newton's method or the bisection method to solve for $\gamma^{(j+s/K)}$ in

$$\begin{aligned}
\sum_{x\in\{0,1\}^M} \frac{1}{\Phi_{\Lambda^{(j+(s-1)/K)}}}\, y_k y_\ell\, \exp\left(\frac{1}{2}x^\top \left[\Lambda^{(j+(s-1)/K)} + \gamma^{(j+s/K)}_{k,\ell}(\mathbf{1}^\top\mathbf{1} - I_M)\right] x\right) &= \eta^{(j)}_{k,\ell} \\
\sum_{x\in\{0,1\}^M} \frac{1}{\Phi_{\Lambda^{(j+(s-1)/K)}}}\, y_k u_m\, \exp\left(\frac{1}{2}x^\top \left[\Lambda^{(j+(s-1)/K)} + \gamma^{(j+s/K)}_{k,i}(\mathbf{1}^\top\mathbf{1} - I_M)\right] x\right) &= \eta^{(j)}_{k,m} \\
\sum_{x\in\{0,1\}^M} \frac{1}{\Phi_{\Lambda^{(j+(s-1)/K)}}}\, u_m u_n\, \exp\left(\frac{1}{2}x^\top \left[\Lambda^{(j+(s-1)/K)} + \gamma^{(j+s/K)}_{i,j}(\mathbf{1}^\top\mathbf{1} - I_M)\right] x\right) &= \eta^{(j)}_{m,n} \\
\text{for } 1\le k<\ell\le J \text{ and } 1\le m<n\le L&
\end{aligned}$$

Here $\mathbf{1}$ is the M dimensional vector with all 1 elements, and I_M is the $M\times M$ identity matrix. The required expectations can be calculated by direct enumeration when M is small, or approximated by generalized belief propagation [44] or Monte Carlo estimation [1] when M is large.

Byrne [8] used a sequential update algorithm for the M step in a Boltzmann machine parameter estimation algorithm. However, to maintain monotonic convergence, Byrne's algorithm requires a large number of iterations in the M step to ensure a maximum is achieved, otherwise monotonic convergence property can be violated for the sequential updates he proposes. In our case, EM-IS uses a parallel update that avoids this difficulty. A sequential algorithm that maintains the monotonic convergence property can also be adapted as described in [11].

To compare EM-IS to standard Boltzmann machine estimation techniques, first consider the derivation of a direct EM approach. In standard EM, given the previous parameters $\Lambda^{(j)}$, one solves for new parameters Λ by maximizing the auxiliary Q function with respect to Λ

$$\begin{aligned}
Q(\Lambda, \Lambda') &= \frac{1}{T}\sum_{t=1}^{T}\sum_{u\in\{0,1\}^L} p_{\Lambda'}(u|y^t)\log p_\Lambda(y^t,u) \\
&= -\log(\Phi_\Lambda) \;+\; \frac{1}{2T}\sum_{t=1}^{T}\sum_{u\in\{0,1\}^L} x^\top \Lambda\, x \;\; p_{\Lambda'}(u|y^t)
\end{aligned}$$

Taking derivatives with respect to Λ gives

$$\frac{\partial}{\partial\Lambda}Q(\Lambda,\Lambda') \;=\; -\frac{1}{2}\,E_{p_\Lambda}[xx^\top] \;+\; \frac{1}{2T}\sum_{t=1}^{T}\sum_{u\in\{0,1\}^L} xx^\top \;\; p_{\Lambda'}(u|y^t)$$

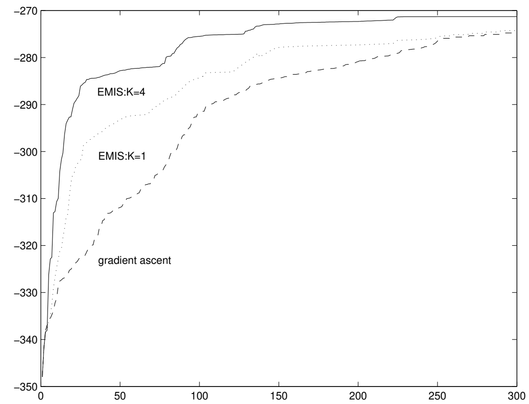


Figure 11: Convergence evaluation: log-likelihood versus iteration, solid curve denotes EM-IS with k=4, dotted curve denotes EM-IS with k=1, and dashed curve denotes gradient ascent.

Apparently there is no closed form solution to the M step and a generalized EM algorithm has to be used in this case. The standard approach is to use a gradient ascent to approximately solve the M step. However, the step size needs to be controlled to ensure a monotonic improvement in Q.

By comparison, EM-IS has distinct advantages over the standard gradient ascent EM approach. First, EM-IS completely avoids the use of tuning parameters while still guaranteeing monotonic improvement. Moreover, we have found that EM-IS converges faster than gradient ascent EM. Figure 11 shows the result of a simple experiment that compares the rate of convergence of M step optimization techniques on a small Boltzmann machine with five visible nodes and three hidden nodes. Comparing EM-IS to the gradient ascent EM algorithm proposed in [1], we find that EM-IS obtains substantially faster convergence. Figure 11 also shows that using several IS iterations in the inner loop, $K = 4$, yields faster convergence than taking a single IS step, $K = 1$ (which corresponds to Riezler's proposed algorithm [36]).

8 Extensions

We briefly outline some useful extensions and relaxations of the basic LME principle.

8.1 A Bayesian extension

In many statistical modeling situations, the constraints themselves are subject to error due to small sample size effects—particularly in domains where there are a large number of features. One way to mitigate the sensitivity to constraint errors is to relax the LME principle by introducing slack variables [10, 17, 30]. That is, we can augment the LME principle to be

$$\max_{p,\varepsilon} \; H(p) - U(\epsilon)$$

subject to the constraints

$$\int_{x\in\mathcal{X}} f_i(x)\, p(x)\, \mu(dx) \quad = \quad \epsilon_i \; + \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} f_i(x)\, p(z|y)\, \mu(dz) \qquad\qquad i = 1, ..., N$$

where the ϵ_i, for $i = 1, ..., N$, are slack variables that allow for errors on the constraints, and $U : \Re^N \to R$ is a convex function that has its minimum at 0. The regularization term $U(\epsilon)$ penalizes violations in reliably observed constraints to a greater degree than deviations in less reliably observed constraints. This establishes a Bayesian framework for exponential models in which a prior distribution on feature parameters can be naturally incorporated.

To solve the reformulated LME problem, we again restrict p to be a log-linear model and develop an iterative algorithm for finding feasible solutions. The key to developing such an algorithm is to note that the stationary points of the penalized log-likelihood of the observed data, $R(\lambda, \sigma) = \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \log p_\lambda(y) + U^*(\lambda)$, are among the feasible set of the relaxed constraints, where $U^*(\lambda)$ is the convex conjugate of U. For example, given a quadratic penalty $U(\epsilon) = \sum_{i=1}^N \frac{1}{2}\sigma_i^2 \epsilon_i^2$ with $\epsilon_i = \frac{\lambda_i}{\sigma_i^2}$ we obtain $U^*(\lambda) = \sum_{i=1}^N \frac{\lambda_i^2}{2\sigma_i^2}$, the Gaussian prior. In this case, the EM-IS algorithm remains almost the same except that the parameter update (26) in the M step needs to modified to

$$\int_{x \in \mathcal{X}} f_i(x)\, e^{\gamma_i^{(j+s/K)} f(x)}\, p_{\lambda^{(j+(s-1)/K)}}(x)\, \mu(dx) \quad + \quad \frac{\lambda_i^{(j+(s-1)/K)} + \gamma_i^{(j+s/K)}}{\sigma_i^2} \quad = \quad \eta_i^{(j)}$$

8.2 Automated feature selection for model construction

By using the LME principle, it becomes possible to extend the incremental feature induction paradigm [19] to reveal hidden structure. Assume there is a pool of candidate features $\mathcal{F}$. After the nth step of the induction algorithm, we have added n features with weights $\underline{\lambda}_n$ to the model. In the $n+1$st step, we consider adding a single feature $f_\alpha \in \mathcal{F}$ with weight α that has the largest information gain

$$\begin{aligned} G_{\underline{\lambda}_n}(\alpha^*, f_{\alpha^*}) &= \arg\min_{f_\alpha \in \mathcal{F}} \; H(\underline{\lambda}_n) - H(\alpha, \underline{\lambda}_n) \\ &= \arg\max_{f_\alpha \in \mathcal{F}} \; Q(\alpha, \underline{\lambda}_n; \alpha, \underline{\lambda}_n) \\ &= \arg\max_{f_\alpha \in \mathcal{F}} \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} p_{\alpha, \underline{\lambda}_n}(z|y) \log p_{\alpha, \underline{\lambda}_n}(x)\; \mu(dz) \end{aligned}$$

The value of $Q(\alpha, \underline{\lambda}_n; \alpha, \underline{\lambda}_n)$ can be obtained by an iterative maximization, $Q(\alpha, \underline{\lambda}_n; \alpha^{(j)}, \underline{\lambda}_n)$, which is equivalent to maximizing the dual function of the complete data Kullback-Leibler divergence problem

$$\begin{aligned} \max_{p_\alpha} \; D(p_{\alpha, \underline{\lambda}_n} \| p_{\underline{\lambda}_n}) &= -\int_{x \in \mathcal{X}} p(x) \log \frac{p_{\alpha, \underline{\lambda}_n}(x)}{p_{\underline{\lambda}_n}(x)} \mu(dx) \\ \text{subject to} \quad \int_{x \in \mathcal{X}} f_\alpha(x)\, p_{\alpha, \underline{\lambda}_n}(x)\, \mu(dx) &= \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} f_\alpha(x)\, p_{\alpha^{(j)}, \underline{\lambda}_n}(z|y)\, \mu(dz) \end{aligned}$$

This can be solved by Newton's method [4, 19].

8.3 Using general Bregman divergences

The LME principle can be easily extended in terms of minimizing *Bregman distances* [7, 9, 20, 26], a class of generalized entropy measures that is associated with convex functions for a discrete-state distribution. Let $\phi : \Re^M \to \Re$ be a strictly convex function defined on a closed convex set $\mathcal{S} \subset \Re^M$, where $\mathcal{S}$ is typically assumed to be the set of probability distributions (or positive measures) $\underline{p} = (p(x_1), ..., p(x_M))$ over a given finite set of M points, $x_1, ..., x_M$. Assume ϕ is differentiable at all points of $\mathcal{S}_{\text{int}}$, the interior of $\mathcal{S}$. The Bregman divergence for $\underline{p} \in \mathcal{S}$ and $\underline{q} \in \mathcal{S}_{\text{int}}$ is defined to be

$$B_\phi(\underline{p}; \underline{q}) \quad = \quad \phi(\underline{p}) - \phi(\underline{q}) \quad - \quad \left\langle \nabla\phi\left(\underline{q}\right) \cdot \left(\underline{p} - \underline{q}\right) \right\rangle$$

That is, B_ϕ measures the discrepancy between two distributions $\underline{p}$ and $\underline{q}$ by taking the difference between ϕ evaluated at $\underline{p}$ and ϕ's first-order Taylor expansion about $\underline{q}$, evaluated at $\underline{p}$.

Given a finite set of data points $x_1, ..., x_M$ and a set of *features* $f^{(1)}, ..., f^{(N)}$, let F denote the $N \times M$ matrix of feature values at each point. That is, $F(j, x) = f^{(j)}(x)$. Here the columns of F correspond to the finite set of points $x_1, ..., x_M$ and the rows of F correspond to the features $f^{(j)}$. We use the notation $F(x)$ to denote the column of values corresponding to $(f^{(1)}(x), ..., f^{(N)}(x))^\top$. Features in this case correspond to "weak learners" in boosting, or sufficient statistics in a log-linear model. Given a set of data points we can propose a generalization of the LME principle based on minimizing Bregman divergence.

Latent minimum Bregman divergence principle Let $\underline{q}_0 \in \mathcal{S}$ be a default distribution chosen so that $\nabla\phi(\underline{q}_0) = 0$. Given features specifying the properties we would like to match in the data, choose a distribution $\underline{p}$ to minimize

$$\min_{\underline{p}\in\mathcal{S}} \qquad B_\phi(\underline{p}; \underline{q}_0)$$

subject to the constraints

$$F(x)\ \underline{p}(x) \quad = \quad F(x)\ \tilde{\underline{p}}(y)\ \underline{p}(z|y) \quad \text{for all } x$$

where each x decomposes as $x = (y, z)$ for some y and z. ∎

To find feasible solutions to this principle, one can use an analogous alternating minimization procedure with nested iterative scaling as before.

9 Conclusion

We have proposed a latent maximum entropy principle, LME, which incorporates latent variables in its inferred models to obtain more expressive power than the maximum entropy principle of Jaynes. An EM algorithm that incorporates nested iterative scaling, EM-IS, is used to solve the problem of finding feasible solutions for the LME principle. EM-IS retains the main virtues of the EM algorithm—its guarantee of monotonic improvement of the likelihood function, and its absence of tuning parameters. We have shown that many familiar models can be recovered by the LME principle, and that EM-IS recovers many standard iterative training procedures for these models. In one case we have seen that EM-IS leads to a new training procedure that has superior convergence properties to standard methods. We then used EM-IS to develop the ME-EM-IS algorithm for approximately realizing the LME principle. This algorithm exploits EM-IS to generate feasible solutions, but then evaluates the entropy of the candidates and selects a highest entropy feasible solution. A few experiments show the advantage of LME over standard maximum likelihood estimation (MLE) in estimating a data source with hidden variables. Finally, we presented several applications of the LME principle to highlight its generality and to show some useful extensions.

For future work, we are planning to investigate several theoretical questions, including the statistical consistency of LME, bounds on its generalization error, techniques for automatic model complexity control, and the relationship between LME and graphical models. We are also investigating ideas for relaxing the log-linear assumption, and using the LME principle for unsupervised boosting [30].

We have begun to use LME to build models of complex natural phenomenon. In [42], we have applied this method to build a sophisticated mixed chain/tree/table graphical model for statistical language modeling, 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. LME allows us to combine these models, effectively, in a unified framework. We are also planning to investigate several practical applications of the LME principle, including problems in machine translation, text classification, information retrieval, image analysis, computer vision and bioinformatics.

A The information geometry of EM-IS

We give an information geometric interpretation of the EM-IS algorithm by using the information divergence and the technique of alternating minimization on probability manifolds. This interpretation will provide a clear illustration on how the EM-IS algorithm converges to a stationary point of the likelihood function. Our analysis also clarifies some of the properties of EM algorithms more generally.

Define the Kullback-Leibler divergence: $D(p\|q) = \int_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\ \mu(dx)$, (where $0\log 0 = 0 \log \frac{0}{0} = 0, c\log\frac{c}{0} = \infty$ if $c > 0$), which is a measure of distance p from q. It is non-negative, equals 0 if and only if $p = q$, but non-symmetric and does not satisfy triangle inequality.

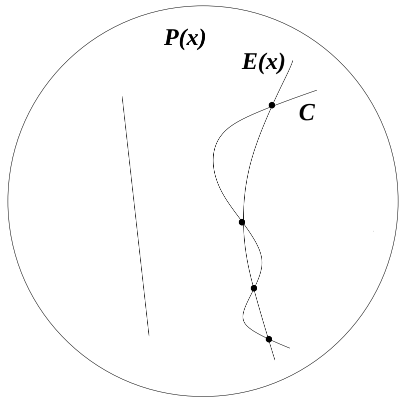


Figure 12: *In the space of all probability distribution on the complete data $\mathcal{P}$, curve $\mathcal{C}$ denotes the set which satisfies the nonlinear LME constraints, curve $\mathcal{E}$ denotes the set of exponential models, and the intersection of $\mathcal{C}$ and $\mathcal{E}$ is the set of the stationary points of the log-likelihood function of the observed data.*

To understand the relationship between maximum likelihood and LME models, note that, unlike the complete data case, we have $L(\lambda) \neq \Lambda(p, \lambda)$ if there are missing data components. However, the stationary points of the log-likelihood function (10) are the approximate solution for (8) under the log-linear assumption, because ignoring the last two terms of (9) we have $\frac{\partial \Upsilon(\lambda)}{\partial \lambda_i} \approx \frac{\partial L(\lambda)}{\partial \lambda_i}$. To illustrate the relationship between maximum likelihood models and LME models, consider the manifolds of the stationary points of the log-likelihood on incomplete data (10) for a general model, and the feasible solutions of the LME principle (3) under the log-linear assumption respectively:

$$\mathcal{C} \;=\; \left\{ p \in \mathcal{P} : \int_{x \in \mathcal{X}} p(x) f_i(x)\, \mu(dx) \;=\; \sum_{y \in \tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z \in \mathcal{Z}} p(z|y) f_i(x)\, \mu(dz),\ i = 1, ..., N \right\} \tag{33}$$

$$\mathcal{E} \;=\; \left\{ p_\lambda \in \mathcal{P} : p_\lambda(x) = \frac{1}{\Phi_\lambda} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(x) \right),\ \lambda \in \Omega \right\} \tag{34}$$

where

$$\Omega \;=\; \left\{ \lambda \in \Re^N : \int_{x \in \mathcal{X}} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(x) \right) \mu(dx) \;<\; \infty \right\} \tag{35}$$

The restriction $\lambda \in \Omega$ will guarantee that the maximum likelihood estimate is an interior point of set of λ's for which $p_\lambda(y)$ is defined.

Figure 12 illustrates that the two manifolds intersect at the set of log-linear models that are also stationary points of the log-likelihood function of the incomplete data.

Now define manifolds $\mathcal{M}$ and $\mathcal{G}_a$ as

$$\mathcal{M} \;=\; \left\{ p \in \mathcal{P} : \int_{z \in \mathcal{Z}} p(x)\, \mu(dz) \;=\; \tilde{p}(y),\quad y \in \mathcal{Y} \right\} \tag{36}$$

$$\mathcal{G}_a \;=\; \left\{ p \in \mathcal{P} : \int_{x \in \mathcal{X}} p(x) f_i(x)\, \mu(dx) \;=\; a_i,\quad i = 1, ..., N \right\} \tag{37}$$

where a is some given vector of constants, $a = (a_1, ..., a_N)$. Then we have

Lemma 1 *$\mathcal{M}$ is a linear submanifold of $\mathcal{C}$.*

Proof: Assume $p_1 \in \mathcal{M}$ and $p_2 \in \mathcal{M}$, and let $p(x) = \theta p_1(x) + (1 - \theta) p_2(x)$ for $\theta \in [0, 1]$. Then $\int_{z \in \mathcal{Z}} p(x)\mu(dz) = \theta \int_{z \in \mathcal{Z}} p_1(x)\mu(dz) + (1 - \theta) \int_{z \in \mathcal{Z}} p_2(x)\mu(dz) = \tilde{p}(y)$. Therefore $p \in \mathcal{M}$, and $\mathcal{M}$ is a linear manifold. Also, for all $p \in \mathcal{M}$ we have $p(x) = \tilde{p}(y)p(z|y)$, and therefore $\int_{x\in\mathcal{X}} p(x)f_i(x)\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} p(z|y)f_i(x)\mu(dz)$, $i = 1, ..., N$. Thus $\mathcal{M} \subset \mathcal{C}$. So we conclude that $\mathcal{M}$ is a linear submanifold of $\mathcal{C}$. ∎

One alternating minimization step [8, 14] starts from a given distribution $p_{\lambda^{(j)}} \in \mathcal{E}$, and finds the backward I-projection, $p_{(j)}$, of $p_{\lambda^{(j)}}$ onto $\mathcal{M}$; that is, $p_{(j)} = \arg\min_{p\in\mathcal{M}} D(p\|p_{\lambda^{(j)}})$. Then by fixing $p_{(j)}$ we next find the forward I-projection, $p_{\lambda^{(j+1)}}$, of $p_{(j)}$ onto $\mathcal{E}$; that is $p_{\lambda^{(j+1)}} = \arg\min_{p_\lambda\in\mathcal{E}} D(p_{(j)}\|p_\lambda)$. It is possible to establish a well known result that an alternating backward I-projection, forward I-projection step leads to the EM update of the auxiliary function $Q(\lambda, \lambda^{(j)})$. We include a proof here to make this paper self-contained.

Lemma 2 *One alternating minimization step between $\mathcal{M}$ and $\mathcal{E}$ is equivalent to an EM update:*

$$\lambda^{(j+1)} \quad = \quad \arg\max_{\lambda\in\Omega} \, Q(\lambda, \lambda^{(j)}) \tag{38}$$

Proof: Given $p_{\lambda^{(j)}} \in \mathcal{E}$, for all $p \in \mathcal{M}$, we have

$$\begin{aligned} D(p\|p_{\lambda^{(j)}}) &= \int_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{p_{\lambda^{(j)}}(x)}\,\mu(dx) \\ &= \sum_{y\in\tilde{\mathcal{Y}}} \int_{z\in\mathcal{Z}} \tilde{p}(y)p(z|y)\log\frac{\tilde{p}(y)p(z|y)}{p_{\lambda^{(j)}}(y)p_{\lambda^{(j)}}(z|y)}\,\mu(dz) \\ &= \underbrace{\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\log\frac{\tilde{p}(y)}{p_{\lambda^{(j)}}(y)}}_{\text{independent of } p(z|y)} + \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\,\underbrace{D(p(z|y)\|p_{\lambda^{(j)}}(z|y))}_{\geq 0} \end{aligned} \tag{39}$$

which implies that

$$\min_{p\in\mathcal{M}} D(p\|p_{\lambda^{(j)}}) \quad = \quad \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\log\frac{\tilde{p}(y)}{p_{\lambda^{(j)}}(y)} \quad = \quad D(\tilde{p}(y)\|p_{\lambda^{(j)}}(y))$$

is achieved by setting $p_{(j)}(x) = \tilde{p}(y)p_{\lambda^{(j)}}(z|y)$.

Now fixing $p_{(j)}$ we seek the $p_\lambda \in \mathcal{E}$ that minimizes $D(p_{(j)}\|p_\lambda)$:

$$\begin{aligned} D(p_{(j)}\|p_\lambda) &= \int_{x\in\mathcal{X}} p_{(j)}(x)\log\frac{p_{(j)}(x)}{p_\lambda(x)}\,\mu(dx) \\ &= \sum_{y\in\tilde{\mathcal{Y}}} \int_{z\in\mathcal{Z}} \tilde{p}(y)p_{\lambda^{(j)}}(z|y)\log\frac{\tilde{p}(y)p_{\lambda^{(j)}}(z|y)}{p_\lambda(x)}\,\mu(dz) \\ &= \underbrace{\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\log\tilde{p}(y) + \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} \tilde{p}(y)p_{\lambda^{(j)}}(z|y)\log p_{\lambda^{(j)}}(z|y)\,\mu(dz)}_{\text{independent of } p_\lambda(x)} \\ &\quad - \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} p_{\lambda^{(j)}}(z|y)\log p_\lambda(x)\,\mu(dz) \end{aligned} \tag{40}$$

The last term is exactly the auxiliary function $Q(\lambda, \lambda^{(j)})$. Thus $p_{\lambda^{(j+1)}} = \arg\min_{p_\lambda\in\mathcal{E}} D(p_{(j)}\|p_\lambda)$ is equivalent to finding $\lambda^{(j+1)}$ in (38). ∎

This equivalence enables us to establish an information geometric interpretation of EM-IS algorithm as follows (see Figure 13 for an illustration): In the space of all probability distributions on the complete data, $\mathcal{P}$, curve $\mathcal{C}$ denotes the set which satisfies the nonlinear LME constraints, curve $\mathcal{E}$ denotes the set of exponential models, and the intersection of $\mathcal{C}$ and $\mathcal{E}$ is the set of stationary points of the log-likelihood function of the observed data. Line $\mathcal{M}$ denotes the set of distributions whose marginal on y matches the empirical distribution.

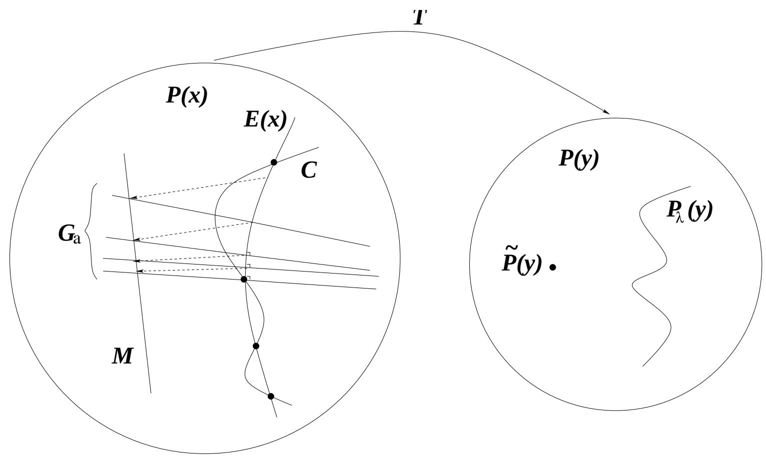


Figure 13: *The information geometry of alternating minimization procedures. Here the straight line $\mathcal{M}$ denotes the set of distributions whose marginal distribution matches the empirical distribution, $\mathcal{M} \subset \mathcal{C}$. The nonlinear operator T denotes marginalization of $p(x)$ over z, and maps the entire space of $p(x)$ into $p(y)$, $\mathcal{M}$ into a singleton $\tilde{p}(y)$, and $\mathcal{E}$ into $p_\lambda(y)$. The intersection of $\mathcal{C}$ and $\mathcal{E}$ is the set of distributions for which the alternating minimization procedure reaches a fixed point.*

Starting from $p_{\lambda^{(j)}} \in \mathcal{E}$, line $\mathcal{G}_a$ denotes the set whose feature expectations match the constant a. The intersection of $\mathcal{M}$ and $\mathcal{G}_a$ is the point $p_{(j)}(x) = \tilde{p}(y)p_{\lambda^{(j)}}(z|y)$ such that $\sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} p_{\lambda^{(J)}}(z|y) f_i(x)\ \mu(dz) = a_i, i = 1, ..., N$. That is, it is the backward I-projection of $p_{\lambda^{(j)}} \in \mathcal{E}$ to $\mathcal{M}$, given by $p_{(j)} = \arg\min_{p\in\mathcal{M}} D(p\|p_{\lambda^{(j)}})$. The E step determines the value of a. The M step finds the intersection of $\mathcal{E}$ and $\mathcal{G}_a$. This is achieved by a forward I-projection of $p_{(j)}$ onto $\mathcal{E}$, given by $p_{\lambda^{(j+1)}} = \arg\min_{p_\lambda\in\mathcal{E}} D(p_{(j)}\|p_\lambda)$; which is equivalent to the I-projection of the uniform distribution $\mathcal{U}$ onto $\mathcal{G}_a$, $p_{\lambda^{(j+1)}} = \arg\min_{p\in\mathcal{G}_a} D(p\|\mathcal{U})$. This alternating procedure will halt at a point where the three manifolds $\mathcal{C}$, $\mathcal{E}$ and $\mathcal{G}_a$ have a common intersection because we will reach a stationary point in that case. Due to the nonlinearity of the manifold $\mathcal{C}$, the intersection is not unique.

Note that in the EM-IS algorithm, each update $\lambda^{(j+s/K)}$ after an iterative scaling phase increases $Q(\lambda, \lambda^{(j)})$, and therefore decreases the divergence $D(p_{(j)}\|p_\lambda)$ between $p_{(j)}$ and p_λ. Instead of finding a final forward I-projection $p_{\lambda^{(j+1)}}$ for each M step, EM-IS only finds an approximation solution after K iterations of the iterative scaling procedure.

Also, note that in the case where there is no unobserved training data, the manifold $\mathcal{M}$ shrinks to a singleton $\tilde{p}(x)$, and $\mathcal{C}$ stretches to match $\mathcal{G}$. In this case, the manifolds $\mathcal{C}, \mathcal{G}$ and $\mathcal{E}$ intersect at a unique point.

Previously, Amari [2], Byrne [8], and Csiszar and Tusnady [14] have given information geometric interpretations of the EM algorithm for log-linear models. However, they did not explicitly consider the constraints imposed by the nonlinear manifold $\mathcal{C}$, and subsequently their explanations of why EM can converge to different solutions depending on the initial point were unclear and hampered by this omission.

We gain further insight by considering the well known Pythagorean theorem [19] for log-linear models, which in the complete data case states that if there exists $p_{\lambda^*} \in \mathcal{G}_a \cap \mathcal{E}$, then

$$D(p\|p_\lambda) \quad = \quad D(p\|p_{\lambda^*}) + D(p_{\lambda^*}\|p_\lambda) \quad \text{for all } p \in \mathcal{G}_a,\ p_\lambda \in \mathcal{E}$$

In the incomplete data case, this theorem needs to be modified to reflect the effect of latent variables.

Theorem 5 *Pythagorean Property: for all $p_\lambda \in \mathcal{E}$ and all $p_{\lambda^*} \in \mathcal{C} \cap \mathcal{E}$, there exists a $p \in \mathcal{C}$ such that*

$$D(p\|p_\lambda) \quad = \quad D(p\|p_{\lambda^*}) + D(p_{\lambda^*}\|p_\lambda) \tag{41}$$

Proof: For all $p_{\lambda^*} \in \mathcal{C} \cap \mathcal{E}$, pick $p(x) = \tilde{p}(y)p_{\lambda^*}(z|y)$. Obviously $p \in \mathcal{M} \subset \mathcal{C}$. Now we show that for all $p_\lambda \in \mathcal{E}$

$$D(\tilde{p}(y)p_{\lambda^*}(z|y)\|p_\lambda(x)) \quad = \quad D(\tilde{p}(y)p_{\lambda^*}(z|y)\|p_{\lambda^*}(x)) + D(p_{\lambda^*}(x)\|p_\lambda(x)) \tag{42}$$

Establishing (42) is equivalent to showing

$$\sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\int_{z\in\mathcal{Z}}p_{\lambda^*}(z|y)\log p_\lambda(x)\mu(dz) = \sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\int_{z\in\mathcal{Z}}p_{\lambda^*}(z|y)\log p_{\lambda^*}(x)\mu(dz)+H(p_{\lambda^*}(x)) + \int_{x\in\mathcal{X}}p_{\lambda^*}(x)\log p_\lambda(x)\mu(dx) \tag{43}$$

The first and second terms on the right hand side cancel because $Q(\lambda^*,\lambda^*)=-H(p_{\lambda^*})$ for all $\lambda^*\in\Theta$ and $p_{\lambda^*}\in\mathcal{C}\cap\mathcal{E}$, by Theorem 4. Plugging the exponential form of p_λ into the remaining terms yields

$$\begin{aligned}&\sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\int_{z\in\mathcal{Z}}p_{\lambda^*}(z|y)\log p_\lambda(x)\mu(dz)-\int_{x\in\mathcal{X}}p_{\lambda^*}(x)\log p_\lambda(x)\mu(dx)\\ =\;&\sum_{i=1}^N\lambda_i\left(\sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\int_{z\in\mathcal{Z}}p_{\lambda^*}(z|y)f_i(x)\mu(dz)-\int_{x\in\mathcal{X}}p_{\lambda^*}(x)f_i(x)\mu(dx)\right) = 0\end{aligned}$$

The term inside the brackets is 0 since $p_{\lambda^*}\in\mathcal{C}\cap\mathcal{E}$. ∎

In the incomplete data case, for each point $p_{\lambda^*}\in\mathcal{C}\cap\mathcal{E}$ there is a unique point $p(x)=\tilde{p}(y)p_{\lambda^*}(z|y)\in\mathcal{C}$ such that $(p,p_{\lambda^*},p_\lambda)$ forms a right triangle for all $p_\lambda\in\mathcal{E}$. However, unlike the complete data case, in the incomplete data case we now have we have multiple points $p_{\lambda^*}\in\mathcal{C}\cap\mathcal{E}$.

Acknowledgments

This work was performed while the first author was at the University of Waterloo. Thanks to Roni Rosenfeld, Fuchun Peng and Ali Ghodsi for their kind assistance. Research supported by MITACS, BUL and NSERC.

References

[1] D. Ackley, G. Hinton and T. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, Vol. 9, pp. 147-169, 1985

[2] S. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, Vol.8, No. 9, pp. 1379-1408, 1995

[3] S. Amari and H. Nagaoka, *Methods of Information Geometry*, American Mathematical Society, 2000

[4] 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

[5] D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999

[6] D. Blei, A. Ng and M. Jordan, "Latent Dirichlet allocation," *Advances in Neural Information Processing Systems 14*, (NIPS), 2002

[7] L. Bregman, "The relaxation method of finding the common point of convex sets and its applications to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, Vol. 7, pp. 200-217, 1967

[8] W. Byrne, "Alternating minimization and Boltzmann machine learning," *IEEE Trans. on Neural Networks*, Vol. 3, No. 4, pp. 612-620, July 1992

[9] Y. Censor and S. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, 1997

[10] S. Chen and R. Rosenfeld, "A survey of smoothing techniques for ME models," *IEEE Trans. on Speech and Audio Processing*, Vol. 88, No. 1, pp. 37–50, 2000

[11] M. Collins, R. Schapire and Y. Singer, "Logistic regression, AdaBoost and Bregman distances," *Machine Learning*, Vol. 48, No. 1-3, pp. 253-285, 2002

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

[13] I. Csiszar, "I-Divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, Vol.3, pp. 146-158, 1975

[14] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statistics and Decisions*, Supplement Issue 1, pp. 205-237, 1984

[15] I. Csiszar, "A geometric interpretation of Darroch and Ratchliff's generalized iterative scaling," *The Annals of Statisics*, Vol. 17, No. 3, pp. 1409-1413, 1989

[16] I. Csiszar, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *The Annals of Statisics*, Vol. 19, No. 4, pp. 2032-2066, 1991

[17] I. Csiszar, "Maxent, mathematics, and information theory," *Maximum Entropy and Bayesian Methods*, Edited by K. Hanson and R. Silver, pp. 35-50, Kluwer Academic Publishers, 1996

[18] 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

[19] 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

[20] S. Della Pietra, V. Della Pietra and J. Lafferty, "Duality and auxiliary functions for Bregman distances," Technical Report CMU-CS-01-109, School of Computer Science, CMU, 2001

[21] 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

[22] G. Elidan, M. Ninio, N. Friedman and D. Schuurmans, "Data perturbation for escaping local maxima in learning," *Proceedings of National Conference on Artificial Intelligence (AAAI)*, 2002.

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

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

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

[26] J. Lafferty, S. Della Pietra and V. Della Pietra, "Statistical learning algorithms based on Bregman distances," *Proceedings of Canadian Workshop on Information Theory*, pp. 77-80, 1997

[27] J. Lafferty, F. Pereira and A. McCallum, "Conditional random fields: probabilistic models for segmenting and labeling sequence data", *Proceedings of International Conference on Machine Learning*, 2001

[28] S. Lauritzen, "The EM-algorithm for graphical association models with missing data", *Computational Statistics and Data Analysis*, Vol. 1, pp. 191-201, 1995

[29] S. Lauritzen, *Graphical Models*, Clarendon Press, 1996

[30] G. Lebanon and J. Lafferty, "Boosting and maximum likelihood for exponential models," In *Advances in Neural Information Processing Systems* (NIPS), 14, 2002

[31] D. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, 1969

[32] D. MacKay and L. Peto, "A hierarchical Dirichlet language model," *Natural Language Engineering*, Vol. 1, No. 3, pp. 289-307, 1995

[33] X. Meng, "On the rate of convergence of the ECM algorithm," *The Annals of Statistics*, Vol. 22, No. 1, pp. 326-339, 1994

[34] T. Minka, "Estimating a Dirichlet distribution,", manuscript, 2000

[35] J. O'Sullivan, "Alternating minimization algorithms: from Blahut-Arimoto to expectation-maximization," In *Codes, Curves and Signals: Common Threads in Communications*, A. Vardy, ed., Kluwer, 1998

[36] S. Riezler, *Probabilistic Constraint Logic Programming*, Ph.D. Dissertation, University of Stuttgart, Germany, 1999

[37] Y. Rubinstein and T. Hastie, "Discriminative vs Informative Learning," *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1997

[38] C. Small, J. Wang and Z. Yang, "Eliminating multiple root problems in estimation," *Statistical Science*, Vol. 15, No. 4, pp. 313–340, 2000

[39] V. Vapnik, *The Natural of Statistical Learning Theory*, Springer, 2000

[40] M. Wainwright, T. Jaakkola and A. Willsky, "Tree-based reparameterization framework for analysis of belief propagation and related algorithms," to appear on IEEE Trans. Information Theory, 2003

[41] S. Wang, *Statistical Recursive Estimation Algorithms for Speaker Adaptation*, Ph.D. Dissertation, University of Illinois at Urbana-Champaign, 2000

[42] 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

[43] C. Wu, "On the convergence properties of the EM algorithm," *Annals of Statistics*, Vol. 11, pp. 95-103, 1983

[44] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring Artificial Intelligence in the New Millenium*, G. Lakemeyer and B. Nebel (editors), Morgan Kaufmann, 2002.