The Latent Maximum Entropy Principle

SHAOJUN WANG, Wright State University
DALE SCHUURMANS, University of Alberta
YUNXIN ZHAO, University of Missouri at Columbia

We present an extension to Jaynes' maximum entropy principle that incorporates latent variables. The principle of *latent maximum entropy* we propose is different from both Jaynes' maximum entropy principle and maximum likelihood estimation, but can yield 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.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Maximum entropy, iterative scaling, expectation maximization, latent variable models, information geometry

ACM Reference Format:
Wang, S., Schuurmans, D., and Zhao, Y. 2012. The latent maximum entropy principle. ACM Trans. Knowl. Discov. Data 6, 2, Article 8 (July 2012), 42 pages.
DOI = 10.1145/2297456.2297460 http://doi.acm.org/10.1145/2297456.2297460

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 [Jaynes 1983] 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 [Fang et al. 1997; Golan et al. 1996]. 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

Authors' addresses: S. Wang, Department of Computer Science and Engineering, Wright State University, Dayton, OH 45435; D. Schuurmans, Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8 Canada; Y. Zhao, Department of Computer Engineering and Computer Science, University of Missouri at Columbia, Columbia, MO 65211-2060.

DOI 10.1145/2297456.2297460 http://doi.acm.org/10.1145/2297456.2297460

joint distribution, and thereby allow for effective generalization and practical inference to be realized; as with standard graphical models such as 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 [Della et al. 1997].

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 [Wang et al. 2001].

In this article, 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 that satisfy 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) often achieves better estimates than maximum likelihood estimation when estimating hidden variable models from small samples of observed data.

Learning probabilistic models with latent variables have been extensively studied in machine learning and statistics for many decades. For both directed and undirected graphical models, model parameters are learned by maximum likelihood estimation where the latent variables are marginalzing out to obtain the likelihood over observed data. A key difference between directed graphical models and undirected graphical models is that a directed graphical model requires many local normalization constraints, whereas an undirected graphical model has a global normalization factor. In this article, we show an intimate connection between the latent maximum entropy principle and maximum likelihood estimation (MLE) for undirected graphical models is that the feasible solutions in LME are equivalent to the set of stationary points of the likelihood in MLE. However, the LME and MLE 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 for undirected graphical models. Another important relevant work on incorporating hidden variables in a maximum entropy philosophy is the maximum entropy discrimination (MED) model proposed by Jaakkola et al. [1999] where hidden variables are considered in Jebara's thesis [2000], and its later extensions to structured prediction by Zhu et al. [2008] and Zhu and Xing [2009]. Basically, maximum entropy discrimination (and its structured extensions) has the same objective function (with a uniform prior, the KL-divergence is equivalent to the ME) as the ME principle but with a different set of constraints. The methods to consider hidden variables are similar, that is, learning a joint distribution over all the random variables and taking the averaging (expectations) over hidden variables to define the constraints. However, the motivations and problem formulations for ME and MED are completely different. Fist of all, ME is motivated for density estimation and the observed data samples are given as training data; MED is motivated for classification and the pairwise observed data samples as well as its labels are given as training data. Second, in ME, the observable and hidden variables are random variables, and the task is to look for the joint distribution of both observable and hidden variables that maximizes the joint entropy subject to nonlinear constraints that model's feature expectation match empirical feature expectation; but in MED, the prediction is made by averaging a parametric discriminant function, which is a linear model of a set of features and their weights, and the weights of features are treated as random variables. The joint distribution of the weights and hidden variables are learned by maximizing the entropy of the joint distribution, subject to margin constraints where the hidden variables are marginalized out. Due to the hidden variables, both have to perform EM type iterative procedures to obtain the feasible or locally optimal solutions. Another important relevant work on incorporating hidden variables is the posterior regularization (PR) for latent variable models proposed by Ganchev et al. [2010] and Graca et al. [2007]. PR is a variant of EM algorithm where, in E step, prior knowledges are encoded as constraints that posterior probability has to satisfy, and the objective PR maximizes is log-likelihood penalized by average Kullback–Leibler divergence of posteriors from the set of constraints. Thus PR applies to both directed graphical models and undirected graphical models, but LME only applies to undirected graphical models; both PR and LME are penalized log-likelihood methods, but the penalization terms are different.

2. MOTIVATION

In 1957, Jaynes [1983] 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 we must infer a probability distribution from data where the distribution should satisfy known constraints, then among distributions consistent with the constraints, we 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 $\tilde{Y} = (y_1, ..., y_T)$, we 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 we 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 [Cover and Thomas 1991].

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 bimodal data because it will continue to postulate a unimodal distribution. However, the existence of the two peaks in the data might not be accidental. For example, there could be two subpopulations 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 subpopulation 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

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

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, we would still be reduced to inferring a single unimodal Gaussian, as above. To cope with missing but nonarbitrary 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 as 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,

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

subject to the constraints

$$\int_{x\in\mathcal{X}} f_i(x)\;\; p(x)\,\mu(dx) \;=\; \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y)\int_{z\in\mathcal{Z}} f_i(x)\;\; p(z|y)\,\mu(dz),\; i = 1...N, \tag{3}$$
$$Y \text{ and } Z \text{ not independent,}$$

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, $p(z|y) = p(y,z)/\int_{z'\in\mathcal{Z}} p(y,z')\mu(dz) = p(x)/\int_{x'=(y,z')} p(x')\mu(dz')$, where $x = (y,z)$ 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, that is, $X = Y$, then the problem is reduced to Jaynes' model where the constraints are given by $\int_{y\in\mathcal{Y}} p(y) f_i(y)\;\mu(dy) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) f_i(y)$. 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,

[1] In this article, 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.

we 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

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

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

$$p_\lambda \;=\; \arg\max_{p\in\mathcal{P}} \;\Lambda(p,\lambda),$$
$$\Upsilon(\lambda) \;=\; \Lambda(p_\lambda, \lambda).$$

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 we could obtain a closed form solution for p_λ in terms of λ, we 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, we can take the derivative of $\Lambda(p, \lambda)$ with respect to $p(x)$ and try to set this to 0 for all $p(x)$:

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

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 Eq. (5) and setting the remainder to zero, we find

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

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, we could hope that p_λ is at least approximately log-linear in the feature values f_i. 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

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

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 of

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

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

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

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 3.1. *Under the log-linear assumption, locally 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

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

(This quantity is actually $1/T$ times the standard log-likelihood where T is the sample size; but this additional factor is not relevant for our purposes.) 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 nonconstant function of λ whose maxima generally do not coincide with $L(\lambda)$ or $H(p_\lambda)$. This fact is proved in Theorem 5.1 in Section 5. 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 [1995] and Riezler [1999], 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

$$\begin{aligned} &\max_{p(x)}\ H(X) = H(C) + H(Y|C), \\ \text{subject to}\quad &\int_{x\in\mathcal{X}} \delta_k(c)\ \ p(x)\,\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c\in\{1,2\}} \delta_k(c)\ \ p(c|y), \\ &\int_{x\in\mathcal{X}} y\,\delta_k(c)\ \ p(x)\,\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c\in\{1,2\}} y\,\delta_k(c)\ \ p(c|y), \\ &\int_{x\in\mathcal{X}} y^2\,\delta_k(c)\ \ p(x)\,\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_{c\in\{1,2\}} y^2\,\delta_k(c)\ \ p(c|y) \qquad \text{for } k = 1, 2, \\ &\qquad Y \text{ and } C \text{ not independent.} \end{aligned} \tag{11}$$

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

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

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), we 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, we were 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; that is, $p(x) = p(y, c) = \theta_c N(y; \mu_c, \sigma_c^2)$ 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 as

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

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 rewrite this distribution in a log-linear form [Amari and Nagaoka 2000],

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

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 as

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

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

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

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 3.1 to derive a practical training algorithm for obtaining feasible solutions to the LME principle under the log-linear assumption. Obviously, since Theorem 3.1 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 3.1.

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 [Dempster et al. 1977], we typically decomposes the log-likelihood $L(\lambda)$ as a function of λ into

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

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

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

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 $\lambda^{(j)} \to \lambda^{(j+1)}$ 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 nondecreases $L(\lambda)$, 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 [Dempster et al. 1977; Wu 1983].[2]

For log-linear models in particular, we have

$$\begin{aligned} Q\left(\lambda, \lambda^{(j)}\right) &= \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} p_{\lambda^{(j)}}(z|y) \log p_\lambda(x)\, \mu(dz) && (18)\\ &= \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} p_{\lambda^{(j)}}(z|y) \left[\left(\sum_{i=1}^{N} \lambda_i f_i(x) \right) - \log(\Phi_\lambda) \right] \mu(dz) \\ &= -\log(\Phi_\lambda) \; + \; \sum_{i=1}^{N} \lambda_i \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} f_i(x)\; p_{\lambda^{(j)}}(z|y)\, \mu(dz). && (19) \end{aligned}$$

by plugging the log-linear form (6) into (18) and recalling that $x = (y, z)$. Crucially, it turns out that maximizing $Q\left(\lambda, \lambda^{(j)}\right)$ 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 4.1. *Maximizing $Q\left(\lambda, \lambda^{(j)}\right)$ as a function of λ for fixed $\lambda^{(j)}$ is equivalent to solving*

$$\max_{p} \quad H(p) = -\int_{x\in\mathcal{X}} p(x) \log p(x)\, \mu(dx), \qquad (20)$$

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

where $x = (y, z)$.

PROOF. Define the Lagrangian $\Lambda\left(p, \lambda, \lambda^{(j)}\right)$ by

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

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

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

[2] It is usually possible to check whether the stationary point is in fact a local maximum [Dempster et al. 1977; Wu 1983].

(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

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

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 [Bertsekas 1999] we have

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

□

It is important to realize that the new constrained optimization problem in Theorem 4.1 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, coordinate descent [Huang et al. 2010], conjugate gradient [Malouf 2002; Minka 2003], and interior-point methods [Censor and Zenios 1997; Fang et al. 1997]. In the case where the feature functions $f_i(x)$ are all non-negative, the generalized iterative scaling algorithm (GIS) [Darroch and Ratchliff 1972] or improved iterative scaling algorithm (IIS) [Berger et al. 1996; Della et al. 1997] 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.

Note that in implementing this algorithm, as with any EM or IS algorithm, we 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 we must resort to Monte Carlo methods or approximation methods in these cases [Della et al. 1997]. We will demonstrate both kinds of models in Section 7.

A natural interpretation of the iterative EM-IS procedure is the following: If the right-hand side of Eq. (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 Eq. (3). If this value matches the constant

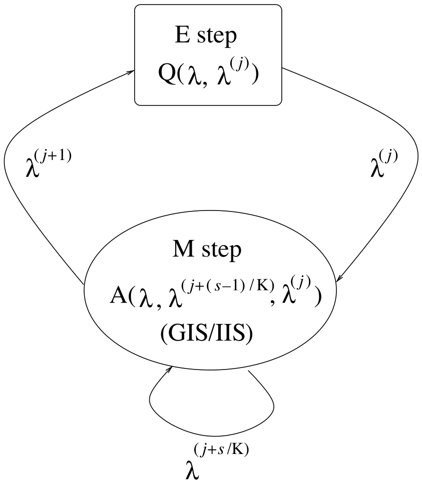

E step
$Q(\lambda, \lambda^{(j)})$
$\lambda^{(j+1)}$
$\lambda^{(j)}$
M step
$A(\lambda, \lambda^{(j+(s-1)/K)}, \lambda^{(j)})$
(GIS/IIS)
$\lambda^{(j+s/K)}$

Fig. 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 λ_i, $i = 1, ..., N$, and K denotes the number of cycles of full parallel updates.

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 approaches of maximum likelihood estimation estimation for log-linear models with incomplete data, and even its general theory, similar to what we presented in this article, have been presented earlier [Hagenaars 1993; Little and Rubin 2002; Meng and Rubin 1993] by combinations of the EM algorithm with iterative proportional fitting techniques. Special instances of the combination of EM-IS have been developed in the context of applications such as natural language parsing [Riezler et al. 2000], text segmentation and labeling [Lafferty et al. 2001] and finite-state processing [Eisner 2002]. Lauritzen [1995] has suggested a similar EM-IS 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). Riezler [1999] has also proposed a similar algorithm and provied the general theory of the EM-IS algorithm, convergence of the EM-IS algorithm, Theorem 3 in this article, follows directly from the proof of convergence given in Riezler [1999]. There, convergence is shown for a GEM algorithm that is a special case of the EM-IS algorithm where only one iteration of IS in applied in the M-step. From convergence of this GEM algorithm, convergence of a corresponding GEM algorithm that employs more than one IS iteration, or a corresponding EM algorithm that iterates IS until convergence to achieve full maximization in the M-step, follows directly. But Riezler 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.

Sequential update variants for iterative scaling have been presented by Darroch and Ratchliff [1972] and extended by Goodman [2002]. The experiments conducted by Goodman clearly show that sequential update in iterative scaling can improve

ALGORITHM 1. EM-IS

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

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

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

$$\lambda_i^{(j+s/K)} \;=\; \lambda_i^{(j+(s-1)/K)} + \gamma_i^{(j+s/K)}, \qquad (25)$$

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

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

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

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

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:

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

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

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

convergence speed over parallel updates. Moreover, for maximum entropy models, the experiments conducted by Minka and Malouf show an even more impressive improvement of convergence speed of conjugate-gradient techniques over iterative scaling techniques. This motivates us to employ conjugate gradient techniques in the M-step of an "EM-CG" algorithm to directly optimize the incomplete data log-likelihood for log-linear models. This could possibly yield more efficient approximations to the LME principle than EM-IS. Unfortunately, these approaches are not scalable to large-scale data sets, since these optimization methods are not parallel/distributed algorithms and have to be done at one machine. However, for some problems such as language modeling in Section 8, there are too many parameters to be stored in a single machine, iterative scaling with parallel update is an ideal optimization technique.

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

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

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

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 bimodal distribution over the observed component Y, whereas standard ME is reduced to producing a unimodal Gaussian in this situation.[3] Interestingly, the update formula we obtain for $p_{\lambda^{(j)}} \to p_{\lambda^{(j+1)}}$ 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.

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 3.1 can be exploited to reduce this question to showing

[3] Radford Neal has observed that dropping the dependence constraint between Y and C allows the unimodal ME Gaussian solution with a uniform mixing distribution to be a feasible global solution in this specific case. However, this model is ruled out by the dependence requirement.

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 [Wu 1983].

THEOREM 4.2. *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*

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

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 3.1. 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 [Berger et al. 1996; Della et al. 1997]. 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 $A(\lambda, \lambda', \lambda^{(j)})$.

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

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 Eq. (30), we have

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

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 Theorem 1 of Wu [1983], 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 (6) of Wu [1983] and Wu's continuity condition (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 the 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 is then 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.

ALGORITHM 2. ME-EM-IS

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.

Although this is not a sophisticated optimization approach, we have found it sufficient to demonstrate the potential benefits of the LME principle, and therefore have left the problem of refining the optimization technique to future research. Nevertheless, despite its simplicity, an apparent difficulty in implementing ME-EM-IS remains: we need to calculate the entropies of the candidate models produced by EM-IS. We might suppose that the entropy has to be calculated explicitly for each candidate model by evaluating the expectation,

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

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 5.1. *If λ is in the set of feasible solutions, that is, $\lambda \in \Theta$ 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 Theorem 2, 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 4.2, 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 3.1 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 we have 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^*} needs to have been determined already as part of the M step for solving (26).

There are a few other observations that follow from Theorem 5.1. First, note that in the special case where there is no missing data, that is, $X = Y$, we have $H(\lambda, \lambda) = 0$ and Theorem 5.1 shows that $L(\lambda) = -H(p_\lambda)$ for a feasible solution $\lambda \in \Theta$; a well-known result of standard maximum entropy theory [Berger et al. 1996; Della et al. 1997]. We 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 the 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 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 series of simple experiments to ascertain whether LME or MLE yields better estimates when inferring models from sample data that has missing components [Wang et al. 2003]. In the first instance, 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 $\bar{\dagger} = (y_1, ..., y_T)$ 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 3.1 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 5.1 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

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

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 θ_c, μ_c, σ_c^2 parameters and λ).
(4) Calculate the entropy and likelihood for each feasible candidate.
(5) Select the maximum entropy candidate p_{LME} as the LME estimate, and the maximum likelihood candidate p_{MLE} as the MLE estimate.
(6) Calculate the cross entropy from $p^*(y)$ to the marginals $p_{LME}(y)$ and $p_{MLE}(y)$, 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

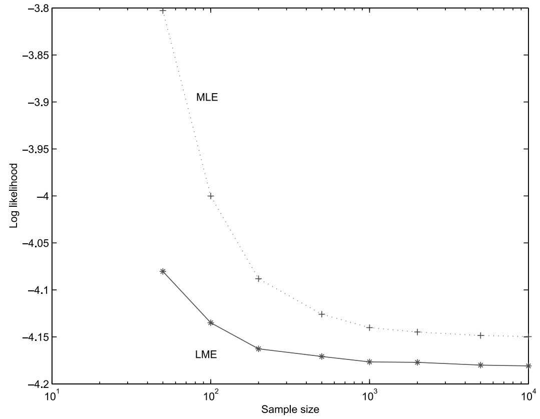


Fig. 2. Average log-likelihood of the MLE estimates versus the LME estimates in Gaussian mixture experiment 1.

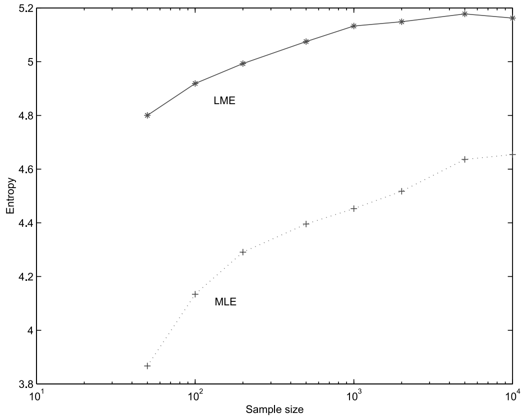
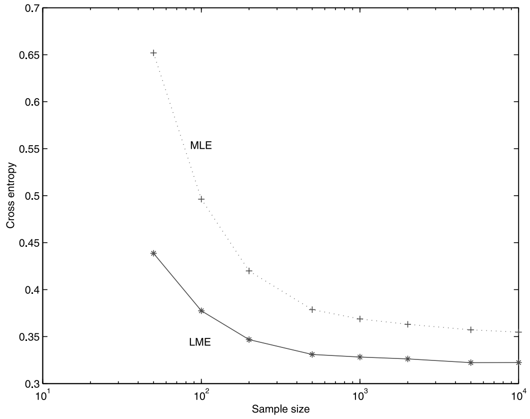


Fig. 3. Average entropy of the MLE estimates versus the LME estimates in Gaussian mixture experiment 1.

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 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. (In fact, in Section 8 we show that LME can be regularized in exactly



sample size	10	50	100	200	500	1000	2000	5000	10000
MLE	3.6656	0.6520	0.4963	0.4199	0.3788	0.3688	0.3631	0.3572	0.3548
LME	1.4325	0.4386	0.3775	0.3468	0.3310	0.3285	0.3264	0.3223	0.3224

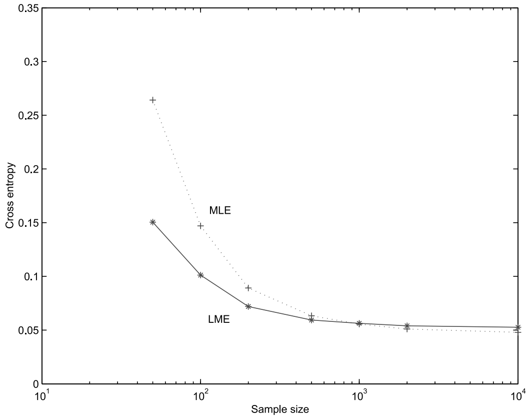
Fig. 4. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Gaussian mixture experiment 1.

the same way as MLE by incorporating a prior on parameters. It still maintains an empirical advantage in this case.) 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.

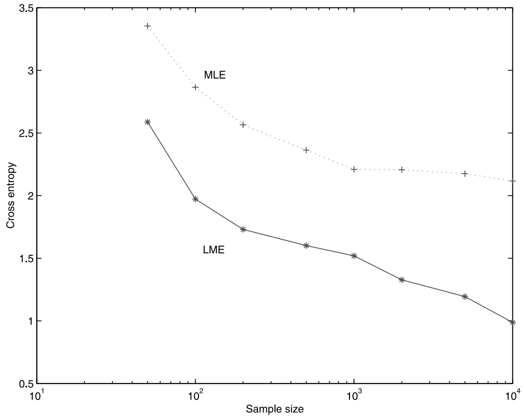
Figure 5 shows that LME still obtained a significantly lower cross entropy than MLE at small sample sizes, but lost its advantage at larger sample sizes. At a crossover point of $T = 1000$ data points, MLE began to produce slightly better estimates than LME, but only marginally so. Overall, LME still 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



sample size	10	50	100	200	500	1000	2000	5000	10000
MLE	3.0388	0.2641	0.1470	0.0892	0.0633	0.0557	0.0510	0.0487	0.0479
LME	0.5695	0.1505	0.1012	0.0719	0.0594	0.0563	0.0540	0.0529	0.0526

Fig. 5. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Gaussian mixture experiment 2.

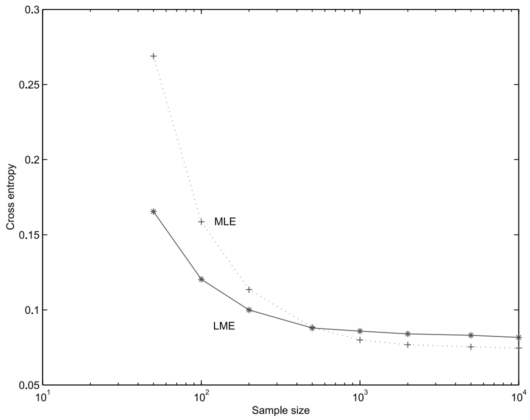


sample size	50	100	200	500	1000	2000	5000	10000
MLE	3.3539	2.8648	2.5654	2.3634	2.2098	2.2067	2.1751	2.1164
LME	2.5881	1.9723	1.7301	1.6009	1.5196	1.3276	1.1941	0.9871

Fig. 6. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Gaussian mixture experiment 3.

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.



sample size	10	50	100	200	500	1000	2000	5000	10000
MLE	4.4644	0.2689	0.1586	0.1135	0.0883	0.0800	0.0768	0.0754	0.0745
LME	0.3865	0.1654	0.1203	0.0999	0.0879	0.0858	0.0851	0.0840	0.0816

Fig. 7. Average cross entropy between the true distribution and the MLE estimates versus the LME estimates in Gaussian mixture experiment 4.

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 sizable 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 as 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 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.

Experiment on Iris Data. To further confirm our observations, we consider a classification problem on the well-known set of *Iris* data as originally collected by Anderson and first analyzed by Fisher [1936]. The data consists of measurements of the length and width of both sepals and petals of 50 plants for each of three types of *Iris* species *setosa, versicolor,* and *virginica*. In our experiments, we intentionally ignore the types of species, and use the data for unsupervised learning and clustering of multivariate Gaussian mixture models. Among 150 samples, we uniformly chose 100 samples as training data, and the rest of the 50 samples as test data. Again, we started from 300 initial points, where each initial point is chosen as follows: first, we calculate the sample mean and covariance matrix of the training data, then perturb the sample mean using the sample variance as the initial mean, and take sample covariance as the covariance for each class. To measure the performance of the estimates, we use the empirical test set likelihood and clustering error rate. We repeat this procedure 100 times. Table I shows the averaged results. We see that the test data is

Table I. Comparison of LME and MLE on *Iris* Data Set

	log-likelihood	error rate
LME	5.58886	0.1220
MLE	5.37704	0.2446

more likely under the LME estimates, and also that the clustering error rate is cut in half.

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, we can add a regularization term to the LME principle in the same way we 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 a differential entropy of $-\infty$ and any nondegenerate 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. In fact, this leads to a question of whether 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. In Section 9 we give an answer to this important topic.

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 merely to 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 [Wang and Schuurmans 2003], which has applications in natural language modeling and other areas [Blei et al. 2002; MacKay and Peto 1995]. In this problem, the observed data has the form of an M dimensional probability vector $y = (y_1, ..., y_M)$ such that $0 \leq y_\ell \leq 1$ 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

$$\begin{aligned} & \max_{p(x)} \; H(X) = H(C) + H(Y|C), \\ \text{subject to} \quad & \int_{x\in\mathcal{X}} \delta_k(c) \;\; p(x)\,\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_c \delta_k(c) \;\; p(c|y)\,\mu(dx) \\ & \int_{x\in\mathcal{X}} (-\log y_\ell)\,\delta_k(c) \;\; p(x)\,\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \sum_c (-\log y_\ell)\,\delta_k(c) \;\; p(c|y)\,\mu(dx) \\ & Y \text{ and } C \text{ not independent} \qquad \text{for } \ell = 1, ..., M \text{ and } k = 1, 2, \end{aligned}$$

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, we can calculate the feature expectations according to (24),

$$\begin{aligned} \eta_0^{k,(j)} &= \frac{1}{T}\sum_{t=1}^{T} \sum_{c\in\{1,2\}} \delta_k(c)\; \rho_t^{k,(j)}, \\ \eta_\ell^{k,(j)} &= \frac{1}{T}\sum_{t=1}^{T} \sum_{c\in\{1,2\}} (-\log y_\ell^t)\,\delta_k(c)\; \rho_t^{k,(j)} \qquad \text{for } \ell = 1, ..., M \text{ and } k = 1, 2, \end{aligned}$$

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 $\alpha_\ell^c = 1 - \lambda_l^c$; 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 parameters α_ℓ^c. (This is unlike the Gaussian mixture case

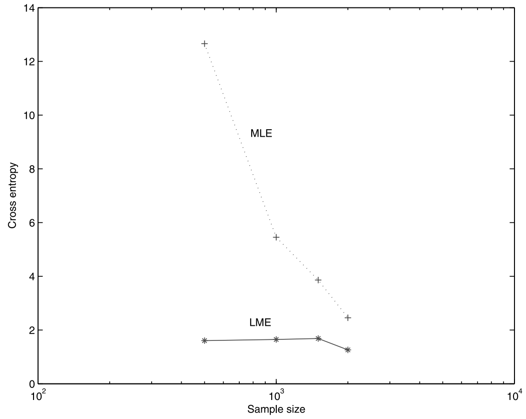


Fig. 8. Average cross entropy between true distribution and MLE versus LME estimates in Dirichlet mixture experiment.

where we could solve for the Lagrange multipliers directly.) By plugging in the form of Dirichlet distribution, the feature expectation will have an explicit formula, thus the constraints that the parameters α_ℓ^c should satisfy become

$$-\Psi(\alpha_l^{c,(j)}) + \Psi\left(\sum_{m=1}^{M} \alpha_m^{c,(j)}\right) = \eta_\ell^{k,(j)}$$

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

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

for $\ell = 1, ..., M$ and $k = 1, 2$. This iteration corresponds to a well-known technique for locally monotonic maximizing the likelihood of a Dirichlet mixture [Minka 2003]. Thus, EM-IS recovers a classical training algorithm as a special case.

Dirichlet Mixture Experiment. To compare model selection based on the LME versus MLE principles for this problem, we conducted an experiment on a mixture of Dirichlet sources. In this experiment, we generate the data according to a three-component Dirichlet mixture, with mixing weights $\theta_c = \frac{1}{6}, \frac{1}{2}, \frac{1}{3}$ and component Dirichlets specified by the α parameters $[1\ 2]^\top$, $[3\ 1]^\top$, and $[5\ 2]^\top$, respectively. The initial mixture weights were generated from a uniform prior, and each α was generated by choosing numbers uniformly from $\{0.1, 0.5, 1, 2.5, 5\}$. Figure 8 shows the cross entropy results of LME and MLE averaged over 10 repeated trials for each fixed training sample size. The outcome in this case shows a significant advantage for LME.

7.2 Boltzmann Machines

Interestingly, the LME principle leads to fundamentally new training algorithms for Boltzmann machine learning [Wang and Schuurmans 2003]. 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

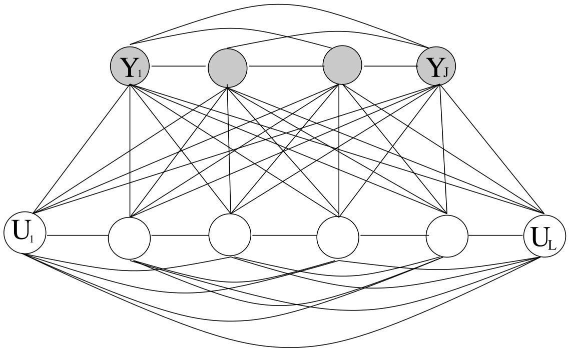


Fig. 9. Boltzmann machine model: nodes Y are observable, nodes U are unobservable.

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 $1 \leq k < \ell \leq J$ and $1 \leq m < n \leq L$, 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 9.

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

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

subject to

$$\begin{aligned}
\sum_{x \in \mathcal{X}} y_k y_\ell \; p(x) &= \sum_{y \in \tilde{\mathcal{Y}}} y_k y_\ell \; \tilde{p}(y) \\
\sum_{x \in \mathcal{X}} y_k u_m \; p(x) &= \sum_{y \in \tilde{\mathcal{Y}}} y_k \; \tilde{p}(y) \sum_{u \in \{0,1\}^L} u_m \; p(u|y) \\
\sum_{x \in \mathcal{X}} u_m u_n \; p(x) &= \sum_{u \in \{0,1\}^L} u_m u_n \; p(u) \\
&\text{for } 1 \leq k < \ell \leq J \text{ and } 1 \leq m < n \leq L \\
&Y \text{ and } U \text{ not independent},
\end{aligned}$$

where $x = (y, u) = (y_1, ..., y_J, u_1, ..., u_L)$ and $\tilde{p}(y) = \frac{1}{T}$. 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):

$$\begin{aligned}
\eta_{k,\ell}^{(j)} &= \frac{1}{T} \sum_{t=1}^{T} y_k^t y_\ell^t \\
\eta_{k,m}^{(j)} &= \frac{1}{T} \sum_{t=1}^{T} y_k^t \sum_{u \in \{0,1\}^L} u_m \; p(u|y^t) \\
\eta_{m,n}^{(j)} &= \sum_{u \in \{0,1\}^L} u_m u_n \; p(u) \qquad \text{for } 1 \leq k < \ell \leq J \text{ and } 1 \leq m < n \leq L.
\end{aligned}$$

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

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

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 as

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

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 [Ackley et al. 1985]. 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

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

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 [Wainwright et al. 2003; Yedidia et al. 2005] or Monte Carlo estimation [Ackley et al. 1985] when M is large.

Byrne [1992] 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 [Collins et al. 2002].

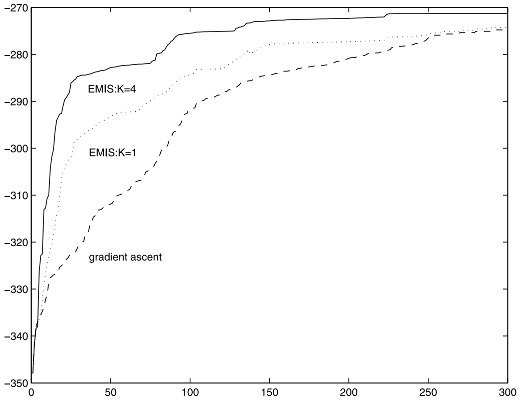


Fig. 10. Convergence evaluation for Boltzmann machine training: log-likelihood versus iteration; solid curve denotes EM-IS with k = 4; dotted curve denotes EM-IS with k = 1; and dashed curve denotes gradient ascent.

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 $\Lambda^{(j)}$, we solve for new parameters Λ by maximizing the auxiliary Q function with respect to Λ:

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

Taking derivatives with respect to Λ gives

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

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 10 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 Ackley et al. [1985], we find that EM-IS obtains substantially faster convergence. Figure 10 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 [Riezler 1999]).

Experiments on Learning Boltzmann Machines. Even assuming that we have an effective algorithm for local parameter optimization, there remains the issue of coping with multiple local maxima. To ascertain whether LME or MLE yields better estimates

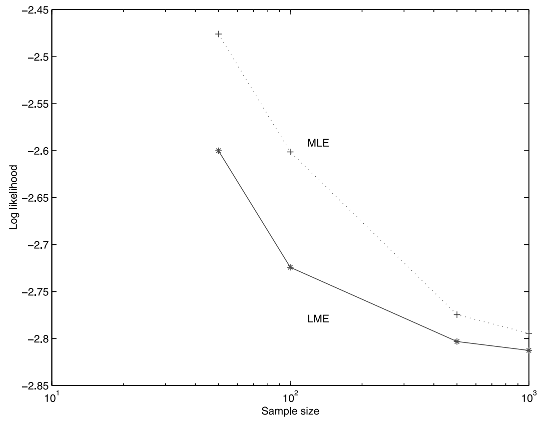


Fig. 11. Average log-likelihood of the MLE estimate versus the LME estimates in Boltzmann machine experiment 1 over 10 runs.

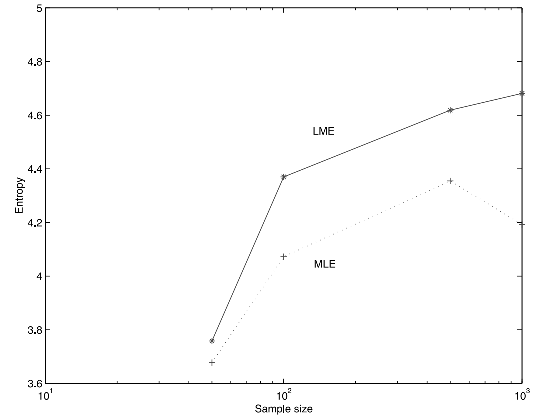


Fig. 12. Average entropy of the MLE estimate versus the LME estimates in Boltzmann machine experiment 1 over 10 runs.

when inferring models from sample data that has a missing component, we conducted a series of simple experiments. In particular, we considered inferring a simple Boltzmann machine model from data that, in each case, consisted of eight nodes with five observable and three hidden units.

In the first experiment, we generated the data according to the assumed model: a Boltzmann machine with five observable and three hidden units, and attempted to learn the parameters for a Boltzmann machine that assumed the same architecture. Figures 11 and 12 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 13 reports the average cross entropy obtained by MLE and LME as a function of sample size, and shows the result that LME achieves substantially lower cross entropy than MLE. LME's advantage is especially pronounced at small sample sizes, and persists even when sample sizes as large as 1,000 are considered (Figure 13).

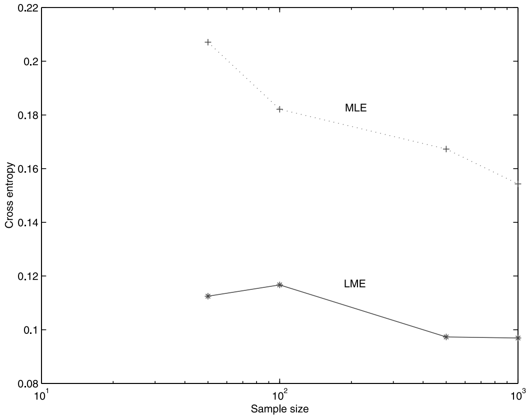


Fig. 13. Average cross entropy between the true distribution and the MLE estimate versus the LME estimates in Boltzmann machine experiment 1 over 10 runs.

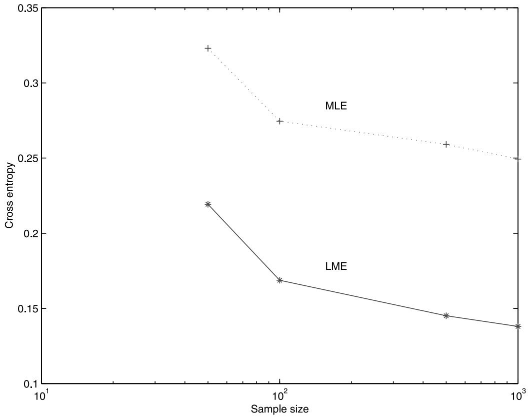


Fig. 14. Average cross entropy between the true distribution and the MLE estimate versus the LME estimates in Boltzmann machine experiment 2 over 10 runs.

In our second experiment, we used a generative model that was a Boltzmann machine with five observable and five hidden units. Specifically, we generated data with this architecture. The LME and MLE estimators still only inferred a Boltzmann machine with five observable and three hidden in this case, and hence were making an incorrect “undercomplete” assumption about the underlying model. Figure 14 shows that LME obtained a significantly lower cross entropy than MLE.

In our third experiment, we used a generative model that was a Boltzmann machine with five observable and one hidden, and the data were generated by this architecture. Again, the LME and MLE estimators inferred Boltzmann machine with five observable and three hidden in this case, and hence were making an incorrect “overcomplete” assumption about the underlying model. Figure 15 shows that LME still obtained a significantly lower cross entropy than MLE.

Although these results are anecdotal, we have witnessed a similar outcome on several other models. Nevertheless, wider experimentation on synthetic and real Boltzmann machine applications and theoretical analysis are necessary to confirm this as a general conclusion.

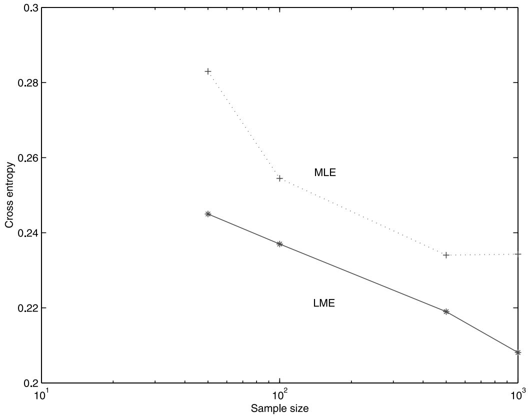


Fig. 15. Average cross entropy between the true distribution and the MLE estimate versus the LME estimates in Boltzmann machine experiment 3 over 10 runs.

8. A REGULARIZED 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 [Chen and Rosenfeld 2000; Csiszar 1996; Lebanon and Lafferty 2002]. That is, we can augment the LME principle to be

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

subject to the constraints

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

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

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

Gaussian Mixture Example

To demonstrate the difference for regularized LME with the penalized maximum likehood estimate, we first consider a learning simple Guassian mixture in Scenario 3 in Section 6. As in Gauvain and Lee [1994], we take the Dirichlet density to model the prior knowledge about the mixture weights

$$p(w_1, \cdots, w_K|\nu, \cdots, \nu_K) \propto \prod_{k=1}^{K} w_k^{\nu_k - 1}. \tag{33}$$

Then, for the mean and covariance of each Gaussian component, we use the joint conjugate prior density, a normal-Wishart density of the form

$$p(\mu, \Sigma|\tau, m, \alpha, V) \propto |\Sigma|^{(\alpha-n)/2} \exp\left(-\frac{\tau}{2}(\mu - m)^T \Sigma(\mu - m)\right) \exp\left(-\frac{1}{2} tr(V\Sigma)\right), \tag{34}$$

where (τ, m, α, V) are the prior density parameters such that $\alpha > n - 1, \tau > 0, \mu$ is an n-dimensional vector and V is $n \times n$ positive definite matrix. Thus, the joint prior density is the product of the prior density defined in (33) and (34).

The EM re-estimation formulas can be derived as follows.

$$w_k = \frac{(\nu_k - 1) + \sum_{t=1}^{T} \rho_t^k}{\sum_{k=1}^{K} \left(\nu_k - 1 + \sum_{t=1}^{T} \rho_t^k\right)} \tag{35}$$

$$\mu_k = \frac{\tau \mu_k + \sum_{t=1}^{T} \rho_t^k y_t}{\tau_k + \sum_{t=1}^{T} \rho_t^k} \tag{36}$$

$$\Sigma_k = \frac{\mu_k + \sum_{t=1}^{T} \rho_t^k (y_t - \mu_k)(y_t - \mu_k)' + \tau_t (m_k - \mu_k)(m_k - \mu_k)'}{(\alpha_k - n) + \sum_{t=1}^{T} \rho_t^k}. \tag{37}$$

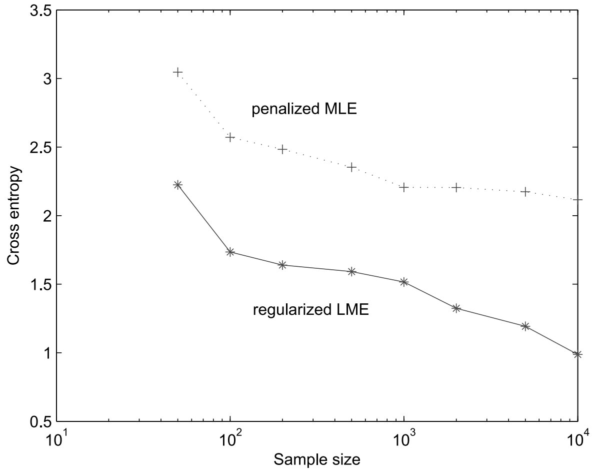
Once we obtain the estimates of $w_k, \mu_k, \Sigma_k,$ for $k = 1, \cdots, K$, we can then transform them into the natural parameterization and calculate the regularized entropy and penalized likelihood. We then choose the highest regularized entropy estimate as the final regularized LME estimate and highest penalized likelihood estimate as the final penalized MLE estimate. (Note that when we calculate the regularized entropy, we use the negative value of auxilary function, since the negative value of the auxilary function is equal to the regularized entropy at the fixed point.)

Figure 16 shows that the regularized LME still produces significantly better estimates than the penalized MLE in this case. Comparing with Figure 6, we notice that when the data is small, the regularization term causes the estimates to be closer to the true distribution, however, when the sample size gets large, this effect diminishes.

Language Modeling Example

The maximum entropy approach has been a key method for language modeling since the 1990s [Jelinek 1998; Lau et al. 1993; Rosenfeld 1996]. In this section we briefly illustrate how to use the regularized LME principle to combine the trigram Markov model with probabilistic latent semantic analysis (PLSA) [Hofmann 2001] to form a stronger language model.

Define the complete data as $x = (W_{-2}, W_{-1}, W_0, D, T_{-2}, T_{-1}, T_0)$, where W_0, W_{-1}, W_{-2} are the current and two previous words, T_{-2}, T_1, T_0 are the hidden "topic" values associated with these words, and D is a document identifier. Thus, $y = (W_{-2}, W_{-1}, W_0, D)$ is the observed data and $z = (T_{-2}, T_{-1}, T_0)$ is unobserved. Typically, the number of documents, words in the vocabulary, and latent class variables are on the order of



sample size	50	100	200	500	1000	2000	5000	10000
MLE	3.0456	2.5713	2.4831	2.3527	2.2061	2.2046	2.1742	2.1158
LME	2.2245	1.7343	1.6391	1.5917	1.5162	1.3241	1.1928	0.9867

Fig. 16. Average cross entropy between the true distribution and the penalized MLE estimates versus the regularized LME estimates in Gaussian mixture experiment 3.

100,000, 10,000, and 100, respectively. A graphical representation of a semantic node interacting with a trigram is illustrated in Figure 17.

We choose n-gram (n = 1,2,3), co-occured n-gram (n = 1,2,3), and the corresponding topic, as well as co-occured topic document as the features. Then, constraints that $p(x)$ should respect are

$$\sum_x p(x)\delta(W_{-2}=w_i,\, W_{-1}=w_j,\, W_0=w_k) = \sum_d \tilde{p}(d)\tilde{p}(W_{-2}=w_i,\, W_{-1}=w_j,\, W_0=w_k|d) \ \ \forall\, i,j,k \tag{38}$$

$$\sum_x p(x)\sum_{\ell=-1}^{0}\delta(W_{\ell-1}=w_i,\, W_\ell=w_j) = \sum_d \tilde{p}(d)\sum_{\ell=-1}^{0}\tilde{p}(W_{\ell-1}=w_i,\, W_\ell=w_j|d) \ \ \forall\, i,j \tag{39}$$

$$\sum_x p(x)\sum_{\ell=-2}^{0}\delta(W_\ell=w_i) = \sum_d \tilde{p}(d)\sum_{\ell=-2}^{0}\tilde{p}(W_\ell=w_i|d) \ \ \forall i \tag{40}$$

$$\begin{aligned}\sum_x p(x)\delta(T_0=t,\, W_{-2}=w_i,\, W_{-1}=w_j,\, W_0=w_k) = {}&\sum_d \tilde{p}(d)\tilde{p}(W_{-2}=w_i,\, W_{-1}=w_j,\, W_0=w_k|d) \ \forall\, i,j,k,t\\ &p(T_0=t|W_{-2}=w_i,\, W_{-1}=w_j,\, W_0=w_k,\, D=d)\end{aligned} \tag{41}$$

$$\begin{aligned}\sum_x p(x)\sum_{\ell=-1}^{0}\delta(T_\ell=t,\, W_{\ell-1}=w_i,\, W_\ell=w_j) = {}&\sum_d \tilde{p}(d)\sum_{\ell=-1}^{0}\tilde{p}(W_{\ell-1}=w_i,\, W_\ell=w_j|d) \ \ \forall\, i,j,t\\ &p(T_\ell=t|W_{\ell-1}=w_i,\, W_\ell=w_jD=d)\end{aligned} \tag{42}$$

$$\begin{aligned}\sum_x p(x)\sum_{\ell=-2}^{0}\delta(T_\ell=t,\, W_\ell=w_i) = {}&\sum_d \tilde{p}(d)\sum_{\ell=-2}^{0}\tilde{p}(W_\ell=w_i|d) \ \ \forall\, i,t\\ &p(T_\ell=t|W_\ell=w_i,\, D=d)\end{aligned} \tag{43}$$

$$\sum_x p(x)\sum_{\ell=-2}^{0}\delta(T_\ell=t,\, D=d) = \sum_d \tilde{p}(d)\sum_{\ell=-2}^{0}p(T_\ell=t|D=d) \ \ \forall\, t, \tag{44}$$

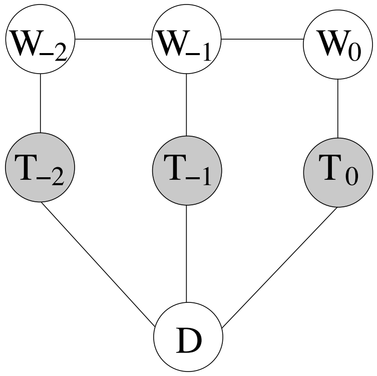


Fig. 17. A graphical representation of the semantic tri-gram model, many arcs share the same parameters and many features are not reflected by arcs.

where $\tilde{p}$ denotes the empirical distribution actually seen in the training corpus, and $\delta(.)$ is an indicator that returns 1 if the event is active, and 0 otherwise. Note the δ functions specify the features that the learned model $p(x)$ should respect. Equations (36 to 38) specify the trigram, bigram, and unigram constraints, which are linear. Equations (39 to 41) speficy the co-occured topic-trigram, topic-bigram, and topic-unigram constraints, which involve the hidden topic variables T, thus they are nonlinear. Finally, Eq. (42) specifies the co-occured document-topic constraints, which again involve the hidden topic variables T; thus they are nonlinear.

The corpus used to train our model was taken from the WSJ portion of the NAB corpus, and was composed of about 150,000 documents spanning the years 1987 to 1989, comprising approximately 42 millions words. The vocabulary was constructed by taking the 60,000 most frequent words of the training data. We split another, separate set of data consisting of 325,000 words, taken from the year 1989, into two parts: one part with 68,000 words used as development data and another part with 257,000 words for testing. There are approximately 12 million types of trigrams from the training data set, if we choose the topic to be 200, then the constraints for Eq. (39) will be 1.2 billion, which is too big to store. Thus, we first ran PLSA on the training data set, then, for each document, we chose the most likely 5 topics from a total of 125 toipcs, and all the other 195 topics were pruned. This procedure significantly reduces the number of constraints for Eq. (39) to approximately 120 million. Unfortunately, this number of constraints leads to the same number of parameters that can be stored on a single machine. So we use a set of machines to store and update the parameters via IIS; use another set of machines to compute feature expectation; and use MPI for message passing, scheduling, and synchronization and so on. In the experiment below, we chose a Gaussian prior with a variance of 1 for each constraint to serve as a regularizer. We set the number of EM iterations to 5 and the number of internal IIS loop iterations to 20.

To control for the effects of maximizing regularized entropy (RLME) versus maximizing *a posteriori* probability (MAP), we first omitted the outer ME-EM-IS procedure and instead just initialize the parameters to zero and execute a single run of EM-IS. We then perturbed the parameters randomly and ran a single EM-IS to find a single locally MAP model (or, equivalently, a single feasible model for the RLME principle). Then, using these results as a control, we reran the procedures with the outer ME-EM-IS procedure reintroduced, to find higher regularized entropy (RLME) solutions and higher penalized likelihood (MAP) solutions. Specifically, we used 20 random starting points for λ, ran EM-IS from each, and then selected the highest regularized entropy solution as the RLME estimate, and the highest penalized maximum likelihood solution as the MAP estimate. The perplexity of the baseline trigram with linear interpolation smoothing technique is 132, while the perplexity of the composite trigram/PLSA trained by RLME is 106, a 19% reduction over baseline: the perplexity of the composite trigram/PLSA trained by MAP is 110, a 16% reduction over baseline.

9. CONSISTENCY AND GENERALIZATION BOUNDS

The MLE method has been extensively studied in the statistics literature and has good statistical properties, such as asymptotic consistency. What we are shown in Wang et al. [2005] and summarized below is that under certain necessary conditions, the latent maximum entropy density estimate $p_{\lambda^\diamond}(y)$ is also consistent.

THEOREM 9.1. *Let* $p_{\lambda^\diamond}(y)$ *denote the maximum* entropy *estimate over the exponential family* $\mathcal{E}$. *Assume for all* $\lambda \in \Omega$ *and for all* $y \in \mathcal{Y}$, *we have* $0 < a \leq \mathcal{F}(y) \leq b$. *Then there exist* $0 < \zeta < \alpha < \infty$ *such that with probability at least* $1 - \eta$

$$D(p_0(y)\|p_{\lambda^\diamond}(y)) - D(p_0(y)\|p_{\hat{\lambda}}(y)) \;\leq\; \frac{4C_3}{\sqrt{M}} E_{\tilde{\mathcal{Y}}}\left[\int_\zeta^\alpha \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)}\, d\epsilon\;\right]$$
$$+C_4\sqrt{\frac{2\log\left(\frac{1}{\eta}\right)}{M}} + E_{\tilde{p}(y)} \log \frac{p_{\hat{\lambda}}(y)}{p_{\lambda^\diamond}(y)},$$

where $p_{\hat{\lambda}}(x)$ *is the information projection* [*Csiszar* 1975] *of (unknown) true distribution* $p_0(y)$ *to the marginal exponential family* $\mathcal{E}(y)$, $\mathcal{N}(\mathcal{F}(y), \epsilon, d_y)$ *is the random covering number of the marginal feature functions* $\mathcal{F}(y) = \int_{z \in \mathcal{Z}} \exp\Big(\langle \lambda,\ f(y,z)\rangle\Big)\mu(dz)$ *at scale* ϵ *with empirical Euclidean distance* d_y *on sample data* $\tilde{\mathcal{Y}}$.

Using this result, we can then easily establish the following consistency property.

COROLLARY 9.2. *Universal consistency: If* $\int_\zeta^\alpha \sqrt{\log \mathcal{N}(\mathcal{F}(y), \epsilon, d_y)}\, d\epsilon$ *is bounded, and also* $E_{\tilde{p}(y)} \log p_{\hat{\lambda}}(y) \;\leq\; E_{\tilde{p}(y)} \log p_{\lambda^\diamond}(y)$, *then* $p_{\lambda^\diamond}(y)$ *will converge to* $p_{\hat{\lambda}}(y)$ *(in terms of the difference of Kullback–Leibler divergence to the true distribution* $p_0(y)$*) with rate* $O(\frac{1}{\sqrt{M}})$, *for any true distribution* $p_0(y)$.

Corollary 9.2 gives a sufficient condition, that is, $E_{\tilde{p}(y)} \log p_{\hat{\lambda}}(y) \;\leq\; E_{\tilde{p}(y)} \log p_{\lambda^\diamond}(y)$, which leads to the universal consistency of latent maximum entropy estimation. This, perhaps, partially explains our observations of experimental results on synthetic data conducted above, that is, in some cases, as the sample size goes to ∞, LME is consistent and does converge to the same point as MLE.

Note that in the proof of Theorem 9.1 and Corollary 9.2, it is not necessary to restrict $p_{\lambda^\diamond}$ to be the model that has global maximum joint entropy over all feasible log-linear solutions. It turns out that the conclusion still holds for all feasible log-linear models $p_\lambda(y)$ which have greater empirical loglikelihood, $E_{\tilde{p}(y)} \log p_\lambda(y)$, than the empirical loglikelihood, $E_{\tilde{p}(y)} \log p_{\hat{\lambda}}(y)$, of the optimal expected loglikelihood estimate $p_{\hat{\lambda}}(y)$. That is, as the sample size grows, any of these feasible log-linear models will converge to $p_{\hat{\lambda}}(y)$ (in terms of the difference of Kullback–Leibler divergence to the true distribution $p_0(y)$) with rate $O(\frac{1}{\sqrt{M}})$.

10. CONCLUSION

We have presented an extension of Jaynes' maximum entropy principle to incomplete data or latent variable estimation problems. It is shown that in contrast to the well-known duality between entropy and likelihood maximization for log-linear models, for latent variable problems, a weaker correlation between maximum entropy and maximum likelihood holds. For the parametric family of log-linear probability distributions, the solutions to local likelihood maximization satisfy the constraints on matching empirical expectations to conditional model expectations, given incomplete data in latent entropy maximization. Among those feasible log-linear solutions, maximization of likelihood and entropy produce different results. 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 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. Some experiments show the advantage of LME over standard maximum likelihood estimation (MLE) in estimating a data source with hidden variables, particularly from small amounts of data.

APPENDIX 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 is nonsymmetric and does not satisfy triangle inequality.

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.

$$\mathcal{C} \ = \ \left\{ p \in \mathcal{P} : \int_{x\in\mathcal{X}} p(x) f_i(x)\mu(dx) = \sum_{y\in\tilde{\mathcal{Y}}} \tilde{p}(y) \int_{z\in\mathcal{Z}} p(z|y) f_i(x)\mu(dz), i = 1, ..., N \right\} \quad (45)$$

$$\mathcal{E} \ = \ \left\{ p_\lambda \in \mathcal{P} : p_\lambda(x) = \frac{1}{\Phi_\lambda} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(x) \right), \ \lambda \in \Omega \right\}, \quad (46)$$

where

$$\Omega \ = \ \left\{ \lambda \in \Re^N : \int_{x\in\mathcal{X}} \exp\left(\sum_{i=1}^{N} \lambda_i f_i(x) \right) \ \mu(dx) \ < \ \infty \right\}. \quad (47)$$

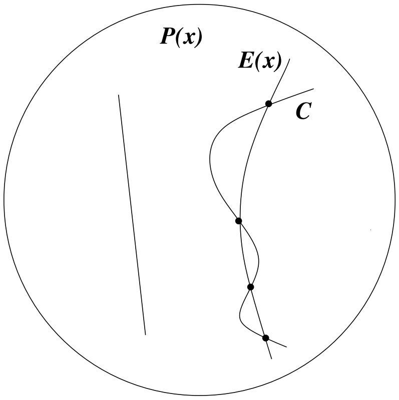


Fig. 18. In the space of all probability distribution on the complete data $\mathcal{P}$, curve $\mathcal{C}$ denotes the set which satisfies the nonlinear LME constraints; curve $\mathcal{E}$ denotes the set of exponential models; and the intersection of $\mathcal{C}$ and $\mathcal{E}$ is the set of the stationary points of the log-likelihood function of the observed data.

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 18 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.

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

$$\mathcal{M} = \left\{ p \in \mathcal{P} : \int_{z\in\mathcal{Z}} p(x)\,\mu(dz) = \tilde{p}(y), \quad y \in \mathcal{Y} \right\} \tag{48}$$

$$\mathcal{G}_a = \left\{ p \in \mathcal{P} : \int_{x\in\mathcal{X}} p(x) f_i(x)\,\mu(dx) = a_i, \quad i = 1, ..., N \right\}, \tag{49}$$

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

LEMMA A.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 [Byrne 1992; Csiszar and Tusnady 1984] 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 article self-contained.

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

$$\lambda^{(j+1)} = \arg\max_{\lambda\in\Omega} Q\left(\lambda, \lambda^{(j)}\right) \tag{50}$$

This equivalence enables us to establish an information geometric interpretation of EM-IS algorithm, as follows (see Figure 19 for an illustration): In the space of all

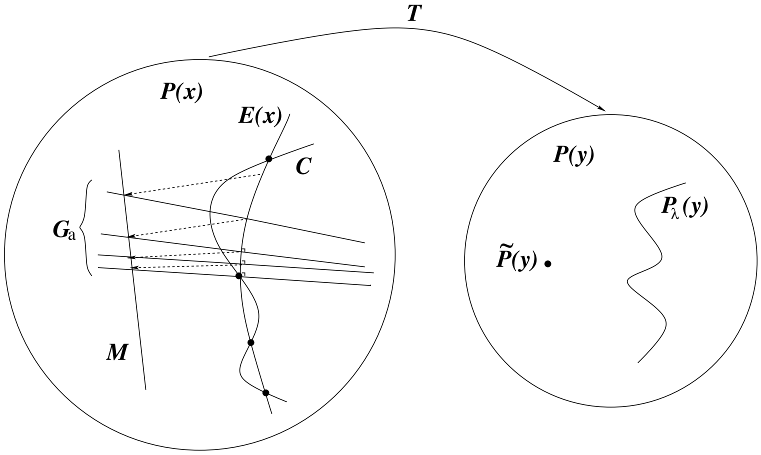


Fig. 19. The information geometry of alternating minimization procedures. Here the straight line $\mathcal{M}$ denotes the set of distributions whose marginal distribution matches the empirical distribution, $\mathcal{M} \subset \mathcal{C}$. The nonlinear operator T denotes marginalization of $p(x)$ 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.

probability distributions on the complete data, $\mathcal{P}$, curve $\mathcal{C}$ denotes the set that 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 margin on y matches the empirical distribution.

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 $p_{\lambda^{(j+1)}} = \arg\min_{p_\lambda \in \mathcal{E}} D(p_{(j)} \| p_\lambda)$; this 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, since 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 [1995], Byrne [1992], and Csiszar and Tusnady [1984] have given an 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 [Della et al. 1997] 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

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

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

THEOREM .3. *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*

$$D(p\|p_\lambda) \;=\; D(p\|p_{\lambda^*}) + D(p_{\lambda^*}\|p_\lambda). \tag{51}$$

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 $p_\lambda \in \mathcal{E}$ that

$$D(\tilde{p}(y)p_{\lambda^*}(z|y)\|p_\lambda(x)) \;=\; D(\tilde{p}(y)p_{\lambda^*}(z|y)\|p_{\lambda^*}(x)) + D(p_{\lambda^*}(x)\|p_\lambda(x)). \tag{52}$$

Establishing (52) is equivalent to showing

$$\begin{aligned}\sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\int_{z\in\mathcal{Z}} p_{\lambda^*}(z|y)\log p_\lambda(x)\mu(dz) \;=\; &\sum_{y\in\tilde{\mathcal{Y}}}\tilde{p}(y)\int_{z\in\mathcal{Z}} p_{\lambda^*}(z|y)\log p_{\lambda^*}(x)\mu(dz) + H(p_{\lambda^*}(x))\\ &+\int_{x\in\mathcal{X}} p_{\lambda^*}(x)\log p_\lambda(x)\mu(dx).\end{aligned} \tag{53}$$

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 5.1. 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 multiple points $p_{\lambda^*} \in \mathcal{C} \cap \mathcal{E}$.

REFERENCES

ACKLEY, D., HINTON, G., AND SEJNOWSKI, T. 1985. A learning algorithm for Boltzmann machines. *Cognitive Sci. 9*, 147–169.

AMARI, S. 1995. Information geometry of the EM and em algorithms for neural networks. *Neural Netw.* 8, 9, 1379–1408.

AMARI, S. AND NAGAOKA, H. 2000. *Methods of Information Geometry*. American Mathematical Society.

BERGER, A., DELLA PIETRA, S., AND DELLA PIETRA, V. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist. 22*, 1, 39–71.

BERTSEKAS, D. 1999. *Nonlinear Programming*. Athena Scientific.

BLEI, D., NG, A., AND JORDAN, M. 2002. Latent Dirichlet allocation. *Advances Neural Inf. Process. Syst. 14*.

BYRNE, W. 1992. Alternating minimization and Boltzmann machine learning. *IEEE Trans. Neural Netw. 3*, 4, 612–620.

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

Collins, M., Schapire, R., and Singer, Y. 2002. Logistic regression, AdaBoost and Bregman distances. *Mach. Learn. 48*, 3, 253–285.

Cover, T. and Thomas, J. 1991. *Elements of Information Theory*. Wiley.

Csiszar, I. 1975. I-Divergence geometry of probability distributions and minimization problems. *Ann. Probab. 3*, 146–158.

Csiszar, I. and Tusnady, G. 1984. Information geometry and alternating minimization procedures. *Statistics and Decisions*. Supplement Issue 1, 205–237.

Darroch, J. and Ratchliff, D. 1972. Generalized iterative scaling for log-linear models. *Ann. Math. Stat. 43*, 5, 1470–1480.

Della Pietra, S., Della Pietra, V., and Lafferty, J. 1997. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell. 19*, 4, 380–393.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Royal Stat. Soc. Series B*, 39, 1–38.

Eisner, J. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Fang, S., Rajasekera, J., and Tsao, H. 1997. *Entropy Optimization and Mathematical Programming*, Kluwer.

Fisher, R. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugenics 7*, II, 179–188.

Ganchev, K., Grafia, J., Gillenwater, J., and Taskar, B. 2010. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res. 11*, 2001–2049.

Gauvain, J. and Lee, C.-H. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. Speech Audio Process.* 2, 2, 291–298.

Golan, A., Miller, D., and Judge, G. 1996. *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* Wiley.

Goodman, J. 2002. Sequential conditional generalized iterative scaling. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 9–16.

Graca, J., Ganchev, K., and Taskar, B. 2007. Expectation maximization and posterior constraints. In *Advances in Neural Information Processing Systems (NIPS)*.

Hagenaars, J. 1993. *Loglinear Models with Latent Variables*. Sage Publications.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* 42, 1, 177–196.

Huang, F., Hsieh, C., Chang, K., and Lin, C. 2010. Iterative scaling and coordinate descent methods for maximum entropy models. *J. Mach. Learn. Res.* 11, 815–848.

Jaakkola, T., Meila, M., and Jebara, T. 1999. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems*.

Jaynes, E. 1983. *Papers on Probability, Statistics, and Statistical Physics*. R. Rosenkrantz Ed., D. Reidel Publishing.

Jebara, T. 2000. Discriminative, generative and imitative learning. Ph.D. dissertation, MIT.

Jelinek, F. 1998. *Statistical Methods for Speech Recognition*. MIT Press.

Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In *Proceedings of the International Conference on Machine Learning (ICML)*. 282–289.

Lau, R., Rosenfeld, R., and Roukos, S. 1993. Trigger-based language models: A maximum entropy approach. In *Proceedings of the 18th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. II, 45–48.

Lauritzen, S. 1995. The EM-algorithm for graphical association models with missing data. *Comput. Stat. Data Anal.19*, 2, 191–201.

Lauritzen, S. 1996. *Graphical Models*. Clarendon Press.

Little, R. and Rubin, D. 2002. *Statistical Analysis with Missing Data* 2nd Ed., Wiley-Interscience.

MacKay, D. and Peto, L. 1995. A hierarchical Dirichlet language model. *Natural Lang. Eng. 1*, 3, 289–307.

Malouf, R. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Natural Language Learning*. 49–55.

Meng, X. and Rubin, D. 1993. Maximum likelihood estimation via the ECL algorithm: A general framework. *Biometrika 80*, 2, 267–278.

Minka, T. 2003. A comparison of numerical optimizers for logistic regression. Manuscript.

Riezler, S. 1999. Probabilistic constraint logic programming. Ph.D. dissertation, University of Stuttgart.

Riezler, et al., 2000. Lexicalized stochastic modeling of constraint-based grammars using log-linear measures and EM training. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Rosenfeld, R. 1996. A maximum entropy approach to adaptive statistical language modeling. *Comput. Speech Lang. 10*, 2, 187–228.

Wainwright, M., Jaakkola, T., and Willsky, A. 2003. Tree-based reparameterization framework for analysis of belief propagation and related algorithms. *IEEE Trans. Inf. Theory 49*, 5, 1120–1146.

Wang, S. and Schuurmans, D. 2003. Learning continuous latent variable models with Bregman divergences. In *Proceedings of the 14th International Conference on Algorithmic Learning Theory (ALT)*.

Wang, S., Rosenfeld, R., and Zhao, Y. 2001. Latent maximum entropy principle for statistical language modeling. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*.

Wang, S., Schuurmans, D., Peng, F., and Zhao, Y. 2005. Combining statistical language models via the latent maximum entropy principle. *Mach. Learn. J. 60* (Special Issue on Learning in Speech and Language Technologies), 229–250.

Wang, S., Greiner, R., and Wang, S. 2009. Consistency and generalization bounds for maximum entropy density. Manuscript.

Wu, C. 1983. On the convergence properties of the EM algorithm. *Ann. Stat. 11*, 95–103.

Yedidia, J., Freeman, W., and Weiss, Y. 2005. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory 51*, 7, 2282–2312.

Zhu, J., Xing, E., and Zhang, B. 2008. Partially observed maximum entropy discrimination Markov networks. In *Advances in Neural Information Processing Systems (NIPS)*.

Zhu, J. and Xing, E. 2009. Maximum entropy discrimination Markov networks. *J. Mach. Learn. Res. 10*, 2531–2569.

Received September 2011; revised October 2011; accepted October 2011