Equivalence of Entropy Regularization and Relative-Entropy Proximal Method

Jason K. Johnson

May 16, 2008

Abstract

We consider two entropy-based interior point methods that solve LP relaxations of MAP estimation in graphical models: (1) an entropy-regularization method and (2) a relative-entropy proximal method. Using the fact that relative-entropy is the Bregman distance induced by entropy, we show that the two approaches are actually equivalent. The purpose of this note is to show one connection between the two approaches described in [1, 2]. Another connection between these two works is that both use distributed iterative-scaling/Bregman-projections algorithms [3, 4] to solve the "inner-loop" optimizations required by the methods summarized below. This second connection, however, is not explored in this present note.

Introduction For the sake of this note, we consider the exponential family of probability distributions on n binary variables $x = (x_1, \ldots, x_n) \in \{0,1\}^n$:

$$P(x) = \exp\{\theta^T \phi(x) - \Phi(\theta)\} = \frac{1}{Z(\theta)} \exp\{\sum_{E \in \mathcal{G}} \theta_E \phi_E(x)\} \tag{1}$$

where $\theta \in \mathbb{R}^d$ are model parameters, $\phi : \mathbb{R}^n \to \mathbb{R}^d$ are model features (sufficient statistics) and $\Phi(\theta) = \log \sum_x \exp\{\theta^T \phi(x)\}$ is the log-partition function. We specify a graphical model based on a hypergraph $\mathcal{G} \subset [V]$ based on vertices $V = \{1, \ldots, n\}$ that are identified with the variables $(x_1, \ldots, x_n)$. We use $[V] = 2^V$ to denote the power set of V. The hypergraph $\mathcal{G}$ may be defined as the set of all cliques of some graph. The features of this model may be defined as products of variables taken over each edge/clique $E \in \mathcal{G}$:

$$\phi_E(x) = \prod_{v \in E} x_v \quad \text{for all} \quad E \in \mathcal{G}. \tag{2}$$

The moments $\eta = P[\phi] = \sum_x P(x)\phi(x)$ define another parameterization of the model. There is one-to-one correspondence between θ and η. The set all realizable moments is defined $\mathcal{M} \triangleq \{\eta = P[\phi]\} \subset \mathbb{R}^d$. This is also the convex hull of $\{\phi(x), x \in \{0,1\}^n\}$. The subset of moments $\eta[E] = (\eta_{E'}, E' \subseteq E)$ determines the corresponding marginal distribution $P(x_E)$. Thus, $\mathcal{M}$ as defined here is essentially equivalent to the *marginal polytope* that is usually defined in the graphical modeling literature [5, 6].

For a specified model $\theta \in \mathbb{R}^d$, we seek the *maximum a posteriori* (MAP) estimate $x^* \in \{0,1\}^n$ to maximize $P(x)$. This problem is equivalent to the following linear program over the set $\mathcal{M}$:

$$\begin{array}{ll} \text{maximize} & \theta^T \eta \\ \text{subject to} & \eta \in \mathcal{M} \end{array}$$

Given η^* maximizing this objective it is straight-forward to recover x^* as $x^*_v = \eta^*_v$ for all $v \in V$. However, it is difficult to characterize the polytope $\mathcal{M}$ in general graphs and with large n.

To demonstrate the stated equivalence, lets begin by considering direct solution of this LP using two interior point methods. Initially, we consider methods assuming it is tractable to evaluate entropy and relative entropy for a specified $\eta \in \mathcal{M}$. Later, we consider relaxations of this method for intractable models (which is our real objective). But discussing the tractable case first will simplify presentation.

Entropy Regularization Let $H(\eta)$ denote the entropy of the probability distribution P_η with moment parameters η, defined by $H(P) = P[\log 1/P(x)]$. We remark that $\nabla H(\eta) = -\theta$ if and only $P_\theta[\phi] = \eta$.

Then, consider the following algorithm. For a decreasing sequence of "temperature" parameters $\{\alpha_k\}$, with $\alpha_k > 0$ and $\alpha_k \to 0$, we solve for η_k that solves the following convex optimization problem:

$$\begin{array}{ll} \text{maximize} & \theta^T \eta + \alpha_k H(\eta) \\ \text{subject to} & \eta \in \mathcal{M} \end{array}$$

Such entropic approaches to solution of LPs have been considered in earlier work [7, 8]. This procedure is illustrated in Figure 1(a). Note that using entropy as a barrier function insures that the solution to this problem will always be in the interior of $\mathcal{M}$ (hence, the constraint $\eta \in \mathcal{M}$ does not actually need to be explicitly enforced). This occurs because the *gradient* of entropy becomes infinite as one approaches the boundary of $\mathcal{M}$ and points into the set (even though the entropy itself is finite). Naturally, as one varies α we obtain a homotopy of solutions $\eta(\alpha)$ such that $\eta(\alpha) \to \eta^*$ (the solution of the LP) as $\alpha \to 0$. We can track the solution down to zero temperature by using η_k as an initial guess for η_{k+1}. One could also extrapolate from previous solutions to get a better initial guess. This will work best if we slowly decrease α.

Relative-Entropy Proximal Method Next, we consider another method which turns out to actually be equivalent. Let $D(\eta, \eta')$ denote the relative entropy between P_η and $P_{\eta'}$, defined by $D(P, P') = P[\log P(x)/P'(x)]$. This also is the *Bregman distance* [3] based on the concave entropy function $H(\eta)$:

$$D(\eta, \eta') = \{H(\eta') + \nabla H(\eta')^T (\eta - \eta')\} - H(\eta) \tag{3}$$

Using this as a measure of distance, we now consider a second "proximal" method to solve the LP. Now, we generate a sequence of iterates $\{\eta_k\}$ by solving a sequence of relative-entropy regularized problems, where each problem is regularized by the previous iterate. That is η_k is produced by solving:

$$\begin{array}{ll} \text{maximize} & \theta^T \eta - \beta_k D(\eta, \eta_{k-1}) \\ \text{subject to} & \eta \in \mathcal{M} \end{array} \tag{4}$$

or, equivalently (using the Bregman distance formula and dividing through by β_k),

$$\begin{array}{ll} \text{maximize} & (\beta_k^{-1}\theta + \theta_{k-1})^T \eta + H(\eta) \\ \text{subject to} & \eta \in \mathcal{M} \end{array} \tag{5}$$

where $\theta_{k-1} = -\nabla H(\eta_{k-1})$. This procedure is illustrated in Figure 1(b). This is similar to the entropy regularization method, except that the θ vector is perturbed by the previous solution. It leads to algorithms similar to one based on an augmented Lagrange multiplier method [9]. Here, we only require that $\beta_k > 0$. We do not require that $\{\beta_k\}$ is decreasing (in fact it may be set to $\beta_k = 1$ for all k).

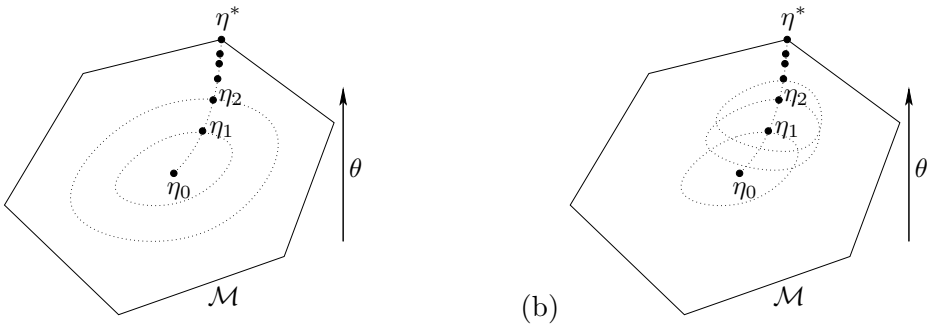

Figure 1: Geometric illustration of two interior-point approaches to solving an LP over the (local) marginal polytope $\mathcal{M}$ with parameter vector θ: (a) the entropic regularization method, and (b) the proximal method using relative entropy between iterates as distance function. Both methods generate point $\{\eta_k\}$ along the same curve $\eta(\alpha)$ connecting the uniform distribution η_0 to the LP solution η^*.

Equivalence Principle Although these two approaches may seem rather different, they are in fact essentially equivalent (provided we initialize the proximal method from the uniform distribution). Let η_0 be the moments corresponding to $\theta_0 = 0$, or equivalently $\nabla H(\eta_0) = 0$.

First, η_1 is obtained by maximizing:

$$f_1(\eta) = \beta_1^{-1}\theta^T\eta + H(\eta) \tag{6}$$

We have used $\nabla H(\eta_0) = 0$. Note that η_1 must satisfy $\nabla f_1(\eta_1) = 0$. Thus, $\theta_1 = -\nabla H(\eta_1) = \beta_1^{-1}\theta$. Next, η_2 is obtained by maximizing:

$$f_2(\eta) = (\beta_1^{-1} + \beta_2^{-1})\theta^T\eta + H(\eta) \tag{7}$$

The solution η_2 must satisfy $\theta_2 \triangleq -\nabla H(\eta_2) = (\beta_1^{-1}+\beta_2^{-1})\theta$. Next, η_3 is obtained by maximizing:

$$f_3(\eta) = (\beta_1^{-1} + \beta_2^{-1} + \beta_3^{-1})\theta^T\eta + H(\eta) \tag{8}$$

(and so on...) By induction, we find that η_k is obtained by maximizing:

$$f_k(\eta) = (\beta_1^{-1} + \cdots + \beta_k^{-1})\theta^T\eta + H(\eta) \tag{9}$$

Clearly, this is equivalent to the entropy-regularized approach with $\alpha_k = (\beta_1^{-1} + \cdots + \beta_k^{-1})^{-1}$. Any sequence $\{\beta_k\}$ will generate a sequence of points along the curve $\eta(\alpha)$ of solutions to the entropy-regularized problem. For instance, the proximal method with $\beta_k = 1$ for all k is equivalent to the entropy-regularized method with $\alpha_k = \frac{1}{k}$.

Tractable Relaxations of MAP The approach described above will be intractable for large n, because $H(\eta)$ is itself difficult to evaluate (unless $\mathcal{G}$ is a thin chordal graph, such as a tree). This suggests using a tractable substitute for the entropy barrier function, such as an edge-wise entropy function:

$$\hat{H}(\eta) = \sum_{E\in\mathcal{G}} H_E(\eta[E]) \tag{10}$$

Here, each term H_E is the marginal entropy of the $P(x_E)$ determined by the marginal moments $\eta[E]$. This function is defined over the larger set $\hat{\mathcal{M}}$ of all η vectors that are *edgewise realizable*, that is, $\hat{\mathcal{M}} = \cap_{E\in\mathcal{G}}\mathcal{M}_E$ where $\mathcal{M}_E$ denotes the set of all η vectors such that the subset of parameters $\eta[E]$ are realizable. Note, this does not assert that the entire vector η is jointly

realizable, but only that each subset $\eta[E]$ ($E \in \mathcal{G}$) are each independently realizable. This provides an outer bound on the marginal polytope $\mathcal{M} \subset \hat{\mathcal{M}}$. This outer bound is called the *local marginal polytope* in the graphical model literature [5, 6]. Then, we seek to solve the relaxed problem:

$$\begin{array}{ll} \text{maximize} & \theta^T \eta + \alpha_k \hat{H}(\eta) \\ \text{subject to} & \eta \in \hat{\mathcal{M}} \end{array}$$

We again track the solution down to zero temperature to obtain the solution of the relaxed LP over $\hat{\mathcal{M}}$. We could also use the proximal method based on the Bregman distance induced by $\hat{H}$:

$$\hat{D}(\eta, \eta') = \sum_{E \in \mathcal{G}} D_E(\eta[E], \eta'[E]) \tag{11}$$

This too is now tractable to compute, involving local calculations of relative entropy between marginals of η and η'. Again, the two methods are equivalent. These algorithm provide a convex optimization approach to solution of the MAP problem using continuous optimization methods. As is well-known, if the solution is integral it provides the optimal MAP. Otherwise, we obtain an upper-bound on the MAP value. Also, by including larger "blocks" of nodes in $\hat{H}$, we obtain tighter approximations to $\hat{\mathcal{M}}$ and reduce (and hopefully eliminate) the integrality gap.

Note that one could also define other entropic barrier functions using non-negative (e.g., convex) combinations of entropies defined on tractable subgraphs (thin chordal graphs). For instance such entropy functions arise in the dual method of Wainwright et al [6], which used convex combinations of trees to approximate the log-parition function of a graphical model. However, the final solution $\eta^* = \lim_{\alpha \to 0} \eta(\alpha)$ is determined solely by the set $\hat{\mathcal{M}} = \cap_E \mathcal{M}_E$, where E ranges over the set of maximal cliques of these graphs. Thus, it will lead to the same solution as using the basic block decompositions. However, these different choices of barrier functions do effect the *path* of solutions $\eta(\alpha)$ one follows to obtain η^*. Similarly, one could use a different starting point η_0 in the proximal method, and this is equivalent to an annealed relative-entropy regularization method, replacing $\alpha_k H(\eta)$ by $-\alpha_k D(\eta, \eta_0)$ in the entropy-regularization method. This also changes the path $\eta(\alpha)$ but does not alter the final solution η^*.

References

[1] Johnson, Malioutov, and Willsky. Lagrangian relaxation for MAP estimation in graphical models. In *Proceedings of the 45th Allerton Conference on Communication, Control and Computing*, September 2007.

[2] Ravikumar, Agarwal, and Wainwright. Message-passing for graph-structured linear programs: proximal projections, convergence and rounding schemes. In *Proceedings of 25th the International Conference on Machine Learning (ICML)*, July 2008 (to appear).

[3] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, 7:200–217, 1967.

[4] I. Csiszár and P.C. Shields. *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communication and Information Theory. NOW Publisher, Inc., 2004.

[5] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. A new class of upper bound on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335, 2005.

[6] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. MAP estimation via agreement on trees: Message-passing and linear programming. *IEEE Transactions of Information Theory*, 51:3697–3717, 2005.

[7] X. Li and S. Fang. On the entropic regularization method for solving min-max problems with applications. *Mathematical methods of operations research*, 46:119–130, 1997.

[8] M. Diasparra and H. Gzyl. Entropic approach to interior point solution of linear programs. *Applied Mathematics and Computation*, 10:339–347, November 2003.

[9] D.P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.