Approximate inference using conditional entropy decompositions

Amir Globerson, Tommi Jaakkola
Computer Science and Artificial Intelligence Laboratory
MIT
Cambridge, MA 02139

Abstract

We introduce a novel method for estimating the partition function and marginals of distributions defined using graphical models. The method uses the entropy chain rule to obtain an upper bound on the entropy of a distribution given marginal distributions of variable subsets. The structure of the bound is determined by a permutation, or elimination order, of the model variables. Optimizing this bound results in an upper bound on the log partition function, and also yields an approximation to the model marginals. The optimization problem is convex, and is in fact a dual of a geometric program. We evaluate the method on a 2D Ising model with a wide range of parameters, and show that it compares favorably with previous methods in terms of both partition function bound, and accuracy of marginals.

Graphical models are a powerful tool for representing multivariate distributions, and have been used with considerable success in numerous domains from coding algorithms to image processing. Although graphical models yield compact representations of distributions, it is often very difficult to infer simple properties of these distributions, such as the marginals over single variables, or the MAP assignment. This difficulty stems from the fact that these problems involve enumeration over an exponential number of assignments, and has motivated extensive research into approximate inference algorithms. Another problem, which turns out to have a key role in developing inference algorithms, is the calculation of the partition function. Recent works (Wainwright & Jordan, 2003; Yedidia et al., 2005) have illustrated that a variational view of partition function estimation can be used to analyze most of the previously introduced approximate inference algorithms, such as mean field, belief propagation (BP) and the tree re-weighting (TRW) framework (Wainwright et al., 2005).

The above analyzes emphasize that a key ingredient in most approximate inference algorithms is the estimation of the entropy of a graphical model given marginals over subsets of its variables. This approximation may be an upper bound on the true entropy, as in the TRW framework, or one which is not guaranteed to be a bound as in the Kikuchi entropies used in Generalized Belief Propagation (GBP) (Yedidia et al., 2005). Another important property of entropy approximation is convexity. The TRW entropies are convex whereas those of GBP are not necessarily convex.

In the current work, we introduce a novel upper bound on graphical model entropy, which results in a convex upper bound on the partition function. The bound is constructed by decomposing the full model entropy into a sum of conditional entropies using the entropy chain rule (Cover & Thomas, 1991), and then discarding some of the conditioning variables, thus potentially increasing the entropy. This entropy bound is then *plugged into* the variational formulation, resulting in a convex optimization problem that yields an upper bound on the partition function. As with previous methods (Yedidia et al., 2005; Wainwright et al., 2005), a byproduct of this optimization problem is a set of *pseudo-marginals* which can be used to approximate the true model marginals.

We evaluate our Conditional Entropy Decomposition (CED) method on a two dimensional Ising grid, and show that it performs well for a wide range of parameters, improving on both TRW and belief propagation.

1 Definitions and Notation

We shall be interested in multivariate distributions over a set of variables $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Consider a set $\mathcal{C}$ of subsets $C \subseteq \{1, \ldots, n\}$. Denote by x_C an assignment to the variables x_i such that $i \in C$. A dis-

tribution over $\boldsymbol{x}$ will be parameterized using functions $\theta(x_C)$. We denote by $\boldsymbol{\theta}$ the vector of all parameters for $C \in \mathcal{C}$. These can be used to define an exponential distribution over $\boldsymbol{x}$ given by

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{\sum_{C\in\mathcal{C}} \theta(x_C)} \ ,$$

where $Z(\boldsymbol{\theta})$ is the partition function defined as $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} e^{\sum_{C\in\mathcal{C}} \theta(x_C)}$. When the sets C are pairs of nodes connected by edges in a graph G, the model is referred to as a pairwise Markov random field.

The marginals of the variables x_C turn out to have a key role in the theory of inference. Specifically, we shall be interested in the set of marginals that can be achieved under *some* distribution.

$$\mathcal{M}(\mathcal{C}) = \{\boldsymbol{\mu} |\ \exists p(\cdot) \ \text{ s.t. } \sum_{\hat{x}:\hat{x}_C = x_C} p(\hat{x}) = \mu(x_C) \ \ \forall C, x_C\} \tag{1}$$

The set $\mathcal{M}(\mathcal{C})$ is known to as the *Marginal Polytope*. The notation $\boldsymbol{\mu}$ indicates a vector comprised of the marginals for all $C \in \mathcal{C}$.

The definition of $\mathcal{M}(\mathcal{C})$ does not restrict the form of the distribution $p(\cdot)$. However, it turns out that any point $\boldsymbol{\mu}$ in the relative interior of $\mathcal{M}(\mathcal{C})$ can actually be achieved using an exponential distribution of the form in Eq. 1 for some parameter vector $\boldsymbol{\theta}(\boldsymbol{\mu})$ (Wainwright & Jordan, 2003)[1]. We shall be specifically interested in the entropy of the exponential distribution corresponding to a given $\boldsymbol{\mu}$. Following (Wainwright & Jordan, 2003), we define the following function

$$\mathcal{A}^*(\boldsymbol{\mu}) = \begin{cases} -H(p(\boldsymbol{x};\boldsymbol{\theta}(\boldsymbol{\mu}))) & \text{if } \boldsymbol{\mu} \in \text{ri}(\mathcal{M}(\mathcal{C})) \\ \infty & \text{if } \boldsymbol{\mu} \notin \text{cl}(\mathcal{M}(\mathcal{C})) \end{cases} \tag{2}$$

where $\text{ri}(\mathcal{M})$ denotes the relative interior of $\mathcal{M}$, $\text{cl}(\mathcal{M})$ is its closure and H is the entropy functional. The function $\mathcal{A}^*(\boldsymbol{\mu})$ returns the negative entropy for points in $\text{ri}(\mathcal{M}(\mathcal{C}))$ and is infinite otherwise. The next section will discuss its importance in approximate inference.

2 A Variational View of Inference

The goal of inference algorithms is to calculate marginals of variables for a given distribution $p(\boldsymbol{x};\boldsymbol{\theta})$. The estimation of the partition function turns out to have a key role in such algorithms. In (Wainwright & Jordan, 2003) the authors show how most known inference algorithms can be analyzed using a variational view of partition function estimation. We review their formalism here, and later use it to introduce our new inference algorithm.

The log partition function can be shown to be the solution of an optimization problem, as described in the following theorem, which results from the convex dual of the partition function (see e.g., (Wainwright & Jordan, 2003; Yedidia et al., 2005))

Theorem 1 *For a given parameter vector $\boldsymbol{\theta}$, the log partition function is given by*

$$\log Z(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu}\in\mathcal{M}(\mathcal{C})} \boldsymbol{\theta}\cdot\boldsymbol{\mu} - \mathcal{A}^*(\boldsymbol{\mu}) \tag{3}$$

Furthermore, the marginals $\boldsymbol{\mu}$ that achieve the optimum are the marginals of the distribution $p(\boldsymbol{x};\boldsymbol{\theta})$.

There are two key difficulties in solving the above problem. One is that the polytope $\mathcal{M}(\mathcal{C})$ does not typically have an explicit description, e.g. in terms of a polynomial number of linear constraints. Notable exceptions to this case are when the set $\mathcal{C}$ corresponds to edges of a tree structured graph, and for restricted classes of functions on planar graphs (see (Deza & Laurent, 1997) page 434). A common approach to addressing this problem is to define a set $\mathcal{M}'$ such that $\mathcal{M}' \supseteq \mathcal{M}(\mathcal{C})$, i.e., it is an outer bound on $\mathcal{M}(\mathcal{C})$. One way of constructing this set is by considering properties that any point $\boldsymbol{\mu} \in \mathcal{M}(\mathcal{C})$ must satisfy, such as consistency between marginals on overlapping sets of variables. Another approach is to use an inner bound of $\mathcal{M}(\mathcal{C})$ by considering a subset of the exponential distributions, for example those in the factored form, as in naive mean field approximations.

The second difficulty is the estimation of the function $\mathcal{A}^*(\boldsymbol{\mu})$. The mapping from $\boldsymbol{\mu} \in \mathcal{M}(\mathcal{C})$ to $\boldsymbol{\theta}(\boldsymbol{\mu})$ is usually hard to state explicitly (again, with the exception of tree graphs). This seems to be true even in cases where the log-partition function can be calculated in polynomial time, such as planar graphs (with binary variables and no field). Also, even if $\boldsymbol{\theta}(\boldsymbol{\mu})$ is known, evaluating $H(p(\boldsymbol{x};\boldsymbol{\theta}(\boldsymbol{\mu})))$ is typically as hard as calculating the partition function. A common approach to the problem of expressing $\mathcal{A}^*(\boldsymbol{\mu})$ is to use approximations which become exact in the tree structured case.

Note that if one replaces $\mathcal{M}(\mathcal{C})$ and -$\mathcal{A}^*(\boldsymbol{\mu})$ in Eq. 3 by an outer bound *and* an upper bound respectively, the resulting optimization problem yields an upper bound on $\log Z(\boldsymbol{\theta})$. This is the case in the TRW method, and also in the approach we present here.

3 The Conditional Entropy Decomposition Approach

We are now ready to describe our new approximate inference algorithm. Our approach builds on an upper

[1] In the general case there will be infinitely many parameter vectors that yield a given exponential distribution and therefore achieve a given $\boldsymbol{\mu}$. Thus the notation $\boldsymbol{\theta}(\boldsymbol{\mu})$ should be understood as some $\boldsymbol{\theta}$ in this set.

bound on $-\mathcal{A}^*(\boldsymbol{\mu})$ which we describe below. Consider an *elimination order* on the variables $x_1, \ldots, x_n$, i.e. a permutation on the set $\{1, \ldots, n\}$. We denote the permutation by $\mathbf{e}$ and its i^{th} element by $e(i)$. By the chain rule for entropy (Cover & Thomas, 1991) we have

$$H(X_1, \ldots, X_n) = \sum_i H(X_{e(i)} | X_{e(i+1)}, \ldots, X_{e(n)})$$

Since conditioning reduces entropy (Cover & Thomas, 1991) removing some of the conditioning variables (e.g., removing Y in $H(X|Y)$) cannot decrease the entropy, thus yielding an upper bound on $H(X_1, \ldots, X_n)$. To construct such a bound that is also efficient to compute, we would like to restrict the number of variables conditioned on. This can be done by defining a subset $C_i \subset \{1, \ldots, n\}$ for each i, and restricting the conditioning variables to be in this set. Formally, for a given elimination order $\mathbf{e}$ define the restricted elimination neighborhood of X_i to be the set of indices[2]

$$N(\mathbf{e}, i) = \{j : j \in C_i \cap \{e(i+1), \ldots, e(n)\}\}$$

The resulting upper bound on the full joint entropy is

$$H(X_1, \ldots, X_n) \leq \sum_i H(X_{e(i)} | X_{N(\mathbf{e},i)}) \qquad (4)$$

The key property of the above bound is that it only requires the marginals of the sets $\boldsymbol{\mu}(x_{e(i)}, x_{N(\mathbf{e},i)})$. Since the size of these is exponential in $|C_{e(i)}|$ the bound is tractable when $|C_{e(i)}|$ is reasonably small. We introduce a notation $C_i^+ = \{i, C_i\}$ for the set which includes both C_i and i itself, since it is the marginals over this set which are relevant for estimating the entropy

Next, we wish to incorporate the above entropy bound in the variational problem in Theorem 1, for a given $\boldsymbol{\theta}$ defined over a set $\mathcal{C}$. A simple way of doing so is to define the clusters C_i^+ such that for every subset $\hat{C}_j \in \mathcal{C}$ there exists a C_i^+ such that $C_i^+ \supseteq \hat{C}_j$. The parameter vector $\boldsymbol{\theta}$ can be redefined over the clusters C_i^+ to yield the same distribution as the original one in Eq. 1. From now on, we thus consider distributions over the clusters C_i^+. A natural outer bound on $\mathcal{M}(\mathcal{C})$ in this case is the set of all marginals over C_i^+ that *agree* on the marginals of variables in their intersection

$$\mathcal{M}_L(\mathcal{C}) = \left\{\boldsymbol{\mu} : \mu_{C_i^+}(x_{C_i^+ \cap C_j^+}) = \mu_{C_j^+}(x_{C_i^+ \cap C_j^+}) \ \forall i, j\right\}$$

Clearly $\mathcal{M}_L(\mathcal{C}) \supseteq \mathcal{M}(\mathcal{C})$ since every set of achievable marginals must satisfy the above consistency constraints. Since the bound in Eq. 4 depends only on marginals of the clusters C_i^+, we may interpret it as a function of these marginals

$$g(\mathbf{e}, \boldsymbol{\mu}) = \sum_i H(\mu(x_{e(i)} | x_{N(\mathbf{e},i)})) \qquad (5)$$

Note that this function is well defined even for $\boldsymbol{\mu} \in \mathcal{M}_L(\mathcal{C})$ that do not correspond to *any* distribution. It is easy to see that the function $g(\mathbf{e}, \boldsymbol{\mu})$ is an upper bound on $-\mathcal{A}^*(\boldsymbol{\mu})$: for $\boldsymbol{\mu} \in \mathcal{M}(\mathcal{C})$ we have $g(\mathbf{e}, \boldsymbol{\mu}) \geq -\mathcal{A}^*(\boldsymbol{\mu})$ due to the entropy inequality in Eq. 4, and for $\boldsymbol{\mu} \notin \mathcal{M}(\mathcal{C})$ the function $-\mathcal{A}^*(\boldsymbol{\mu})$ is $-\infty$ so the bound trivially holds.

Since we have an outer bound on $\mathcal{M}(\mathcal{C})$ and an upper bound on the entropy we can define an optimization problem whose optimum is always an upper bound on the partition function. Define

$$f(\boldsymbol{\theta}, \mathbf{e}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}_L(\mathcal{C})} \boldsymbol{\theta} \cdot \boldsymbol{\mu} + g(\mathbf{e}, \boldsymbol{\mu}) \qquad (6)$$

Then

$$\begin{aligned} \log Z(\boldsymbol{\theta}) &= \sup_{\boldsymbol{\mu} \in \mathcal{M}(\mathcal{C})} \boldsymbol{\theta} \cdot \boldsymbol{\mu} - \mathcal{A}^*(\boldsymbol{\mu}) \\ &\leq \sup_{\boldsymbol{\mu} \in \mathcal{M}(\mathcal{C})} \boldsymbol{\theta} \cdot \boldsymbol{\mu} + g(\mathbf{e}, \boldsymbol{\mu}) \\ &\leq \sup_{\boldsymbol{\mu} \in \mathcal{M}_L(\mathcal{C})} \boldsymbol{\theta} \cdot \boldsymbol{\mu} + g(\mathbf{e}, \boldsymbol{\mu}) = f(\boldsymbol{\theta}, \mathbf{e}) \end{aligned}$$

Thus for all $\boldsymbol{\theta}$ and $\mathbf{e}$ we have $f(\boldsymbol{\theta}, \mathbf{e}) \geq \log Z(\boldsymbol{\theta})$. The above optimization problem in fact maximizes a concave function over a convex set, and thus local optima are global ones (Bertsekas, 1995). The constraint set is linear and hence defines a convex set. The objective is a sum of conditional entropies and is therefore concave. This follows from the fact that the conditional entropy $H(X|Y)$ is a concave function of the joint distribution $p(x, y)$. We provide a proof of this concavity in the appendix.

3.1 Multiple Elimination Orders

The previous section considered a bound based on a single elimination order. However, since the bound is true for any elimination order, we may extend it to use multiple orders. Consider a set of elimination orders $\mathcal{E}$, and a distribution $q(\mathbf{e})$ on this set. Now consider the convex combination of the bounds $g(\mathbf{e}, \boldsymbol{\mu})$

$$g(\boldsymbol{\mu}, q) = \sum_{\mathbf{e}} q(\mathbf{e}) g(\mathbf{e}, \boldsymbol{\mu}) \qquad (7)$$

Again we have a bound $g(\boldsymbol{\mu}, q) \geq -\mathcal{A}^*(\boldsymbol{\mu})$. since $g(\boldsymbol{\mu}, \mathbf{e}) \geq -\mathcal{A}^*(\boldsymbol{\mu})$ for all $\mathbf{e}$, and we take their convex combination using $q(\mathbf{e})$. Since the function $g(\mathbf{e}, \boldsymbol{\mu})$ depends only on subsets of the variables $x_1, \ldots, x_n$ it will

[2]The use of C_i do define $N(\mathbf{e}, i)$ may appear redundant for a single elimination order, since we can define $N(\mathbf{e}, i)$ directly. However this will become useful in Section 3.1 where we consider multiple elimination orders.

not depend directly on $q(\mathbf{e})$ but rather on the probabilities of these subsets appearing in a given elimination order. Denote by $\rho(S|i)$ the probability that a set $S \subseteq C_i$ satisfies $S = N(\mathbf{e}, i)$ for a permutation drawn from $\mathcal{E}$. The notation $\rho(S|i)$ is justified by the fact that $\sum_{S \subseteq C_i} \rho(S|i) = 1$ since every elimination order corresponds to some choice of a subset of C_i. Then

$$g(\boldsymbol{\mu}, \boldsymbol{\rho}) = \sum_{i=1}^{n} \sum_{S \subseteq C_i} \rho(S|i) H(\mu(x_i|x_S)) \tag{8}$$

We can therefore introduce the following optimization problem as another bound on the log partition function

$$f(\boldsymbol{\theta}, \boldsymbol{\rho}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}_L(C)} \boldsymbol{\theta} \cdot \boldsymbol{\mu} + g(\boldsymbol{\mu}, \boldsymbol{\rho}) \tag{9}$$

The above bound only holds if the distributions $\boldsymbol{\rho}$ indeed correspond to some distribution $q(\mathbf{e})$ over elimination orders. One way of obtaining a valid $\boldsymbol{\rho}$ is by considering some finite set of elimination orders and calculating $\boldsymbol{\rho}$ for a distribution over those (e.g., uniform).

We conclude with a specific example of setting the values of $\boldsymbol{\rho}$. Consider a distribution over four variables $x_1, \ldots, x_4$, with clusters $C_1 = \{2, 4\}, C_2 = \{1, 3\}, C_3 = \{2\}, C_4 = \{3\}$. We assume that the distribution $q(\mathbf{e})$ is uniform over two elimination orders $\mathbf{e}_1 = \{1, 2, 3, 4\}$ and $\mathbf{e}_2 = \{1, 2, 4, 3\}$ i.e. $q(i) = \frac{1}{2}$ for $i = 1, 2$. The elimination order $\mathbf{e}_1$ contributes $\frac{1}{2}$ to $\rho(2, 4|1), \rho(3|2), \rho(\emptyset|3), \rho(\emptyset|4)$, where $\emptyset$ is the empty set. Similarly, for $\mathbf{e}_2$ we have a contribution of $\frac{1}{2}$ to $\rho(2, 4|1), \rho(3|2), \rho(3|4)$ and $\rho(\emptyset|3)$. Thus $\rho(2, 4|1) = 1, \rho(3|2) = 1, \rho(\emptyset|3) = 1, \rho(3|4) = \frac{1}{2}, \rho(\emptyset|4) = \frac{1}{2}$ and zero probability for all the other subsets.

4 Exact Decompositions

We now study the conditions under which our approximation procedure becomes exact. Recall that the approximation is generated by *deleting* variables from the conditional entropy decomposition of Eq. 4. Such deletion would usually result in an increase in entropy. However, it will be exact as long as the variables that are not deleted provide all the relevant information about the eliminated variable $X_{\mathbf{e}(i)}$. More formally, consider the conditional entropy $H(X_1|X_R)$ where X_R is some set of variables. Denote the Markov blanket of X_1 by X_U. The condition for X_U being a Markov blanket is that X_1 and $X_{R \setminus U}$ are conditionally independent given X_U, or equivalently $p(x_1|x_U, x_{R \setminus U}) = p(x_1|x_U)$. The above condition implies that $H(X_1|X_R) = H(X_1|X_U)$. This is also true for any U' such that $U' \supseteq U$. Thus as long as we do not delete variables that are in the Markov blanket of X_1 our entropy decomposition is exact. It is important to stress that X_U needs to be a Markov blanket for X_1 under the distribution for $p(x_1, x_R)$ which is already a marginal of the complete distribution. The Markov blanket of X_1 may thus be larger than its blanket under the full distribution $p(x_1, \ldots, x_n)$.

The discussion above suggests a scenario where the decomposition is exact. Assume the original set $\mathcal{C}$ corresponds to edges in a graph G. Consider a given elimination order $\mathbf{e}$, w.l.o.g $\mathbf{e} = \{1, \ldots, n\}$. Now proceed by triangulating the graph using $\mathbf{e}$: i.e., eliminating variables according to $\mathbf{e}$ and after eliminating a variable, connecting all its neighbors in G to each other. Set the cluster C_i to be the set of neighbors of X_i in this process. The C_i constructed in this fashion are clearly a Markov blanket for the variable X_i. Thus, as long as we use marginals such that $\boldsymbol{\mu} \in \mathcal{M}(\mathcal{C})$ the decomposition is exact. In order to enforce $\boldsymbol{\mu} \in \mathcal{M}(\mathcal{C})$ it suffices to require that the cliques corresponding to clusters in the junction tree corresponding to G are consistent. The clusters C_i are necessarily subsets of these cliques so marginals over C_i will also be in $\mathcal{M}(\mathcal{C})$. Thus, CED is similar to other inference algorithms (e.g., generalized belief propagation (Yedidia et al., 2005)) in that it becomes exact when constructed over a junction tree.

5 A Dual Geometric Program

The optimization problem in Eq. 9 is convex and therefore has an equivalent convex dual. It is known that such problems of maximum conditional entropy are duals of convex optimization problems known as geometric programs (GP) (Chiang, 2005). Large scale geometric programs can be solved very efficiently[3], and it is thus worthwhile to explore their duality to the current problem. A geometric problem in convex form is given by

$$\begin{array}{lll} \min & \log \sum_{k=1}^{K_0} e^{\boldsymbol{a}_{0k}^T \boldsymbol{x} + b_{0k}} & \\ \text{So that} & \log \sum_{k=1}^{K_i} e^{\boldsymbol{a}_{ik}^T \boldsymbol{x} + b_{ik}} \leq 0 & i = 1 \ldots, m \\ & \boldsymbol{a}_l^T \boldsymbol{x} + b_l = 0 & l = 1, \ldots, M \end{array}$$

where optimization is over the variable $\boldsymbol{x}$, and $\boldsymbol{a}_{ik}$ and $\boldsymbol{a}_l$ are vectors of the size of $\boldsymbol{x}$ and b_{ik} are scalars. To define the GP dual of CED we introduce the following dual variables

- Variables β_i, t_i for $i = 1, \ldots, n$.
- Subset variables $\lambda^i(x_i, x_S)$ for every $i = 1, \ldots, n$ and $S \subset C_i$ (note that the full set is not included), and every assignment to the variables x_i, x_S.

[3]One efficient implementation is available in the commercial MOSEK optimization package (see www.mosek.com).

- Overlap variables $\gamma^i(x_S)$ for any subset S that is the intersection of C_i^+ and some other C_j^+, and every assignment to x_S. We also denote by $\mathcal{O}^i$ the set of overlap subsets for cluster C_i^+.

Define the following factor for every variable x_i and an assignment to $x_{C_i^+}$

$$h(x_{C_i^+}) = \boldsymbol{\theta}(x_{C_i^+}) + \sum_{S \in \mathcal{O}^i} \gamma^i(x_S) + \sum_{S \subset C_i} \lambda^i(x_i, x_S) + \beta_i$$

The dual problem is then[4][5]

$$\begin{array}{ll} \min & -\sum_i \beta_i + \sum_i \rho(\emptyset|i) t_i \\ \text{So that:} & \sum_{x_i} e^{\rho^{-1}(C_i|i) h(x_{C_i^+})} \leq 1 \\ & \sum_{x_i} e^{-\rho^{-1}(S|i)\lambda^i(x_i,x_S)} \leq 1 \quad S \subset C_i, S \neq \emptyset \\ & \sum_{x_i} e^{-\rho^{-1}(\emptyset|i)\lambda^i(x_i) - t_i} \leq 1 \\ & \sum_{i:S \subset C_i^+} \gamma^i(x_S) = 0 \end{array} \tag{10}$$

The above dual yields some insight into the structure of the primal solution as we show below. By deriving the primal Lagrangian the following characterization of the conditional $\mu(x_i|x_{C_i}), \mu(x_i|x_S)$ is obtained

$$\begin{aligned} \mu(x_i|x_{C_i}) &= e^{\rho^{-1}(C_i|i) h(x_{C_i^+})} \\ \mu(x_i|x_S) &= e^{-\rho^{-1}(S|i)\lambda^i(x_i,x_S)} \end{aligned}$$

These characterizations hold as long as the corresponding marginals $\mu(x_{C_i}), \mu(x_S)$ are not zero. The first two constraints of the dual may thus be interpreted as *sub normalization* constraints for the above conditional distributions. Complementary slackness implies that as long as the marginals are not zero, these constraints hold with equality, but when marginals are zero the constraints may not be active. Comparing the number of constraints to the number of variables, it becomes evident that not all equalities can be satisfied in the general case, and indeed we empirically observed that the primal solution has zero marginal probabilities for some assignments.

6 Optimization Issues

The variational problem described above may be solved by optimizing either the primal (Eq. 9) or the dual (Eq. 10). One shortcoming of the dual formulation is that it requires variables for all assignments to all subsets, and thus may become quite large, albeit sparse. The primal problem is considerably smaller since it has variables only for assignment to the sets C_i^+. Thus in the experimental section we chose to solve via the primal. There are many different algorithms for solving convex problems. Here we used the conditional gradient method (Bertsekas, 1995), which uses linear programming to find feasible search directions.

Another approach to optimization may be via a message passing algorithm as in GBP. Since the primal objective may be written as a weighted sum of entropies (see Eq. 12) one may derive message passing algorithms similar to the GBP ones. Such updates are designed so that their fixed point is a local optimum of the Lagrangian. Since the current problem is convex, convergence of such an algorithm would therefore guarantee a global optimum. However, as in GBP and TRW, such algorithms are not guaranteed to converge. Since in this work we are primarily interested in the quality of the approximation, we defer further study of message passing algorithms to future work.

7 Related Methods

The CED method presented above constructs an approximation of the free energy and uses it to estimate the partition function. It is thus a variational based approach, and as such is related to previously introduced variational methods. The key difference between these methods is in the way they approximate the entropy term $\mathcal{A}^*(\boldsymbol{\mu})$. We next highlight the difference between our approach and previous ones.

7.1 The Tree Re-weighting Framework

Wainwright et al. (2005) construct an upper bound on $-\mathcal{A}^*(\boldsymbol{\mu})$ using spanning trees of the graph G. For a spanning tree T, define the function

$$\hat{g}(\boldsymbol{\mu}, T) = \sum_{i=1}^{n} H(X_i|Pa_T(X_i)) \tag{11}$$

where $Pa_T(X_i)$ is the parent of variable X_i in the tree T, and the entropy is calculated using $\boldsymbol{\mu}$. Then $\hat{g}(\boldsymbol{\mu}, T)$ is the entropy of a distribution on the tree T with marginals $\boldsymbol{\mu}$ and can be shown to be an upper bound on $-\mathcal{A}^*(\boldsymbol{\mu})$. A convex combination of such bounds is then used to obtain an upper bound on $\log Z(\boldsymbol{\theta})$. Interestingly, the above bound may be viewed as an instance of the bound in Eq. 4 if one considers an elimination order where children in the tree T are always eliminated before their parents. The CED bound is however more general, since it does not require a tree property, and can condition on nodes that do not correspond to a tree (or a junction tree). Importantly, the bound in Eq. 4 does not necessarily correspond to an entropy of *any* distribution in the general case, and thus generalizes the TRW approach.

[4] We use $\rho(\emptyset|i)$ to denote the probability of the elimination neighborhood of X_i being empty. Note that for a single elimination order this is always one for the last eliminated variable

[5] Note that the objective is in fact linear

An elegant property of TRW is that one can consider a distribution $\boldsymbol{\rho}$ over *all* spanning trees, and find the $\boldsymbol{\rho}$ that yields the optimal bound. For CED, we can also optimize $\boldsymbol{\rho}$ for a fixed and small enough set of elimination orders. However, it seems hard to optimize over *all* possible elimination orders (see Discussion).

7.2 Generalized Belief Propagation

Yedidia et al. (2005) observed that the belief propagation algorithm is closely related to the variational problem in Theorem 1 with $\mathcal{A}^*(\boldsymbol{\mu})$ given by the Bethe entropy. This motivated the use of higher order approximations such as the one introduced by Kikuchi. Such approximations are weighted combinations of entropies that generally neither upper or lower bound the true entropy. The bound in Eq. 8 can also be viewed as a sum over entropies by rewriting it as [6]

$$g(\boldsymbol{\mu}, \boldsymbol{\rho}) \quad = \quad \sum_{i=1}^{n} \sum_{S \subseteq C_i} \rho(S|i)(H(X_i, X_S) - H(X_S))$$

As in (Yedidia et al., 2005), we also have positive and negative entropy contributions. However, unlike Kikuchi approximations, the current expression yields an upper bound on the entropy, and is also a convex function of its parameters. Convex variants of the Kikuchi entropy were also studied by Heskes et al. (2003) in the context of optimizing a non-convex Kikuchi entropy more efficiently. The entropy decomposition in Heskes et al. (2003) is not necessarily an upper bound on the true entropy. Furthermore, the convexity of the entropy approximation in (Heskes et al., 2003) holds only for consistent marginals whereas the entropy decomposition presented here is convex over all marginals, consistent or not.

8 Experiments

To evaluate the performance of our conditional entropy decomposition method we apply it to an Ising model on a two dimensional grid. A grid of size 10×10 is used to allow comparison with the exact partition function and marginals. The distribution has the form $p(\boldsymbol{x}) \propto e^{\sum_{ij \in E} \theta_{ij} x_i x_j + \theta_i x_i}$, where θ_{ij}, θ_i are parameters, $x_i \in \{\pm 1\}$, and E are edges of the 2D grid. The clusters C_i were chosen as all the neighbors of node i in the graph, so that $|C_i| \leq 4$. The subset probabilities $\rho(S|i)$ were constructed by considering a uniform distribution over elimination orders shown in Figure 1. We considered these four elimination orders, all their cyclical orderings, and their reverse orderings and cyclical reverse orderings. The rationale behind this choice is that those elimination orders yield clique trees with width equal to the tree width of the grid. We also experimented with other elimination distributions but these yielded inferior results.

The parameters θ_i were drawn uniformly from $\mathcal{U}[-d_f, d_f]$ where $d_f \in \{0.05, 1\}$ The parameters θ_{ij} were drawn from $\mathcal{U}[-d_o, d_o]$ or $\mathcal{U}[0, d_o]$ to obtain *mixed* or *attractive* interaction potential respectively. The interaction levels were $d_o \in \{0.2, 0.4, \ldots, 4\}$. The following algorithms were used to estimate both the partition function of the distribution and its singleton marginals

- The conditional entropy decomposition (CED) method with the clusters and elimination orders defined above. Marginals are given by the $\boldsymbol{\mu}$ which maximizes the CED optimization problem.
- A variant of the conditional entropy method, where $\boldsymbol{\rho}$ is not a *legal* vector, but rather is defined as $\rho(C_i|i) = 1$ for the full cluster, and zero otherwise. We denote this variant by CEDF. Here the optimization is still convex, but does not necessarily give a bound on the partition function.
- The TRW method of Wainwright et al. (2005)[7], using a uniform distribution over spanning trees (denoted by TRWUNI in the results). TRW provides an upper bound on the partition function and a set of estimated (*pseudo*) marginals.
- Loopy belief propagation (BP).[8] The marginals obtained by the BP algorithm were used to calculate the Bethe free energy, which approximates the partition function. For some settings of the parameters BP did not converge.[9]

For each setting of the parameters and each algorithm we calculated the following measures: 1) The normalized error in the log partition function $\frac{1}{n}|\log Z^{alg} - \log Z^{true}|$. For TRW and CED Z^{alg} is always larger than the true one. For BP,CEDF this does not necessarily hold. To emphasize this difference, the results for BP,CEDF are shown with a negative sign in the figures. 2) The mean L1 error in the marginals $\frac{1}{n}\sum_i |p^{(alg}(x_i = 1) - p^{(true)}(x_i = 1)|$.

Since CED is essentially a cluster based method, we were also interested in comparing it with GBP. When running GBP with the same clusters C_i^+ that were used for CED, convergence was reached only for very low interaction levels, and therefore we do not report these results here. However, GBP with *square clusters* had good convergence and marginals performance. In order to incorporate such *square clusters* into CED one

[6] Here we take both entropies w.r.t the distribution $\mu(x_{C_i^+})$

[7] We used M. Wainwright's TRW implementation.

[8] We used the inference package by Talya Meltzer available at http://www.cs.huji.ac.il/∼talyam/.

[9] Results are not shown in these cases.

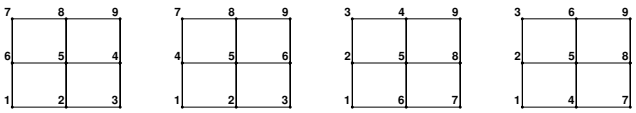

Figure 1: Illustration of the elimination orders considered in generating the distribution $\rho(S|i)$, shown here for a 3×3 two dimensional grid. The four *basic* orderings shown eliminate rows or columns in a sequential fashion.

should include them in the description of $\mathcal{M}(\mathcal{C})$. We expect this to improve CED results further, but have not performed this evaluation yet.

Results for the partition function are shown in Figure 2 and those for marginals in Figure 3. It can be seen that CED yields a much better partition function bound than TRW, and also results in improved marginals. BP yields good marginal estimates for low interaction levels, but breaks down at higher ones, even when it does converge. The performance of CEDF, which is not a upper bound, is good for $d_f = 1$ but deteriorates at high interaction levels for $d_f = 0.05$.

9 Discussion

The CED method provides an intuitive way of generating entropy bounds by relating them to elimination orders of variables. The quality of the approximation is governed by the fraction of the Markov blanket that is preserved in the conditional entropy. Since the neighbors of a node in the graph are clearly a subset of its Markov blanket this constitutes a natural choice for clusters, and indeed the one we followed in the experimental evaluations. However, choosing a larger set would result in a tighter bound. It should be noted however, that one can in fact choose a subset of these neighbors and the bound would remain valid. Such a choice may be reasonable for nodes with high degree.

In the current evaluations we considered a uniform distribution over a set of elimination orders. An improved bound may be obtained by optimizing over this distribution. The main difficulty in such an optimization is that one needs a characterization of which vectors $\boldsymbol{\rho}$ are *valid* subset probabilities. While it is easy to specify properties which these $\boldsymbol{\rho}$ must satisfy, it is not immediately clear how to obtain a complete description of this set. Interestingly, Grotschel et al. (1985) obtained results which characterize this *elimination polytope* if one only considers subsets of size two (i.e., edges) and the graph is planar. It will be interesting to further study the relation between their work and ours.

Acknowledgments The authors wish to thank Martin Wainwright and Talya Meltzer for providing their inference codes, and acknowledge support from the Defense

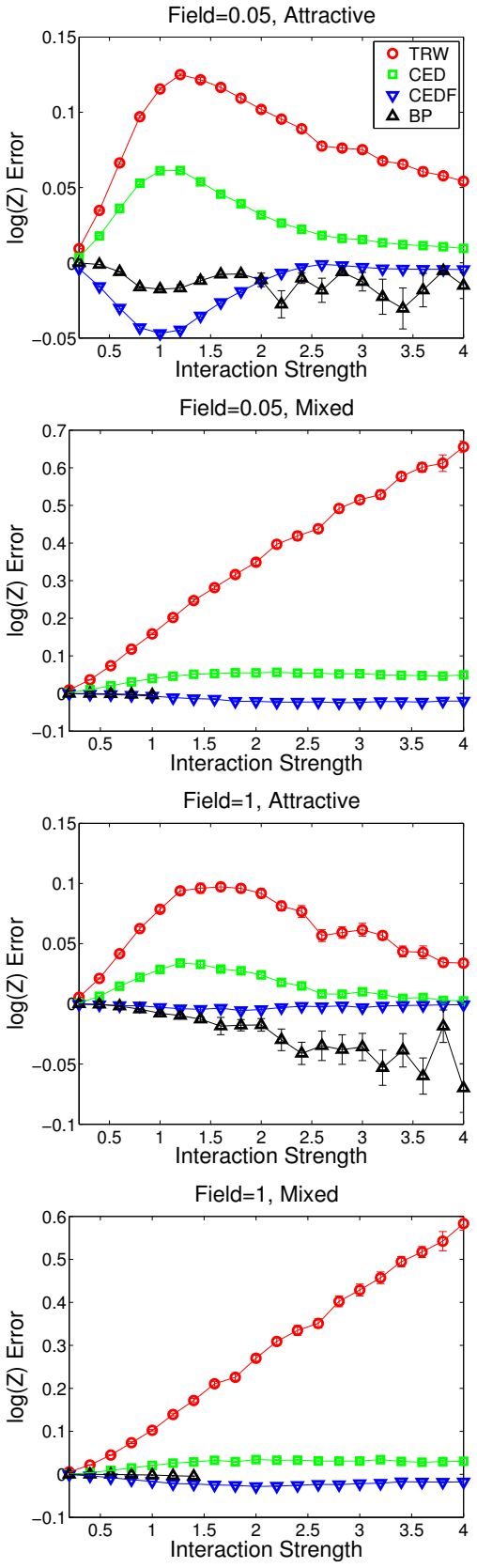

Figure 2: Comparison of partition function estimates for CED and other approximate inference algorithms. Results shown for different settings of the field parameter d_f and interaction parameter setting (Mixed or Attractive). Results for BP are shown only for the convergent cases. Mean and STE are shown for 20 random trials.

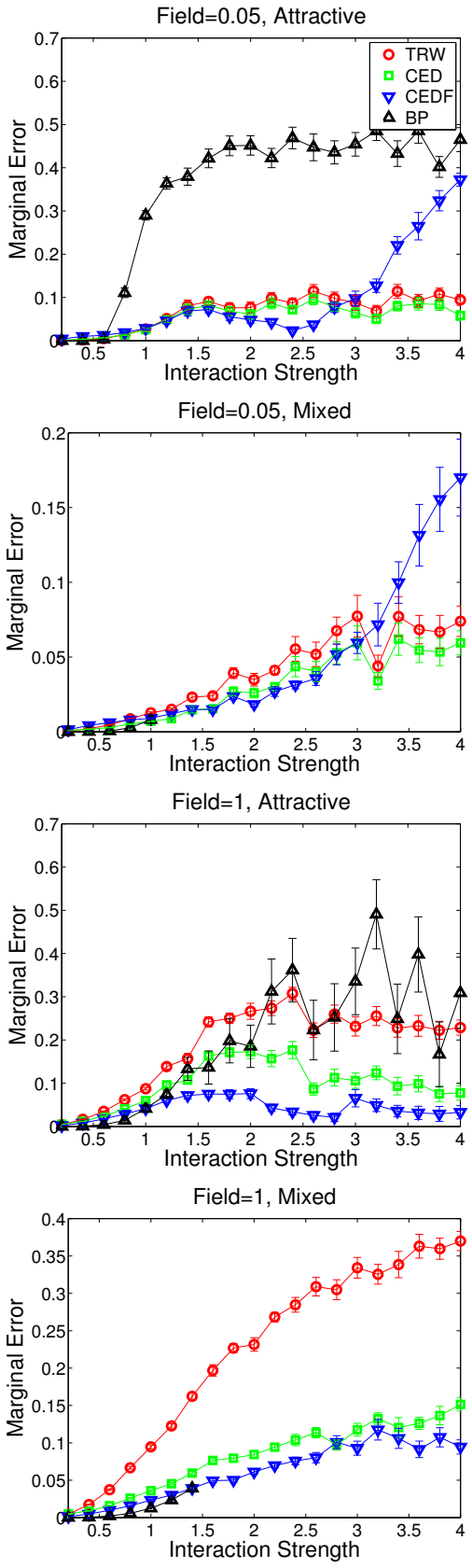

Figure 3: Comparison of error in marginal estimation for CED and other approximate inference algorithms.

Advanced Research Projects Agency (Transfer Learning program). AG is also supported by a fellowship from the Rothschild Foundation - Yad Hanadiv.

A Concavity of Conditional Entropy

Consider the function $H(X|Y)$ as a function of $p(x,y)$. We want to show it is a concave function. Define $u(x) = \frac{1}{|X|}$ as the uniform distribution over X. Write $H(X|Y)$ as a KL divergence (up to a constant)

$$D_{KL}[p(x,y)|p(y)u(x)] = -H(X|Y) + \log|X| \quad (12)$$

Since $D_{KL}[p|q]$ is convex in (p,q) (Cover & Thomas, 1991), we can use it to address the convexity of $H(X|Y)$. Given two distributions $p_1(x,y), p_2(x,y)$, define two *factored* distributions $q_i(x,y) = p_i(y)u(x)$. For $0 \le \lambda \le 1$ define

$$p_\lambda(x,y) = \lambda p_1(x,y) + (1-\lambda)p_2(x,y)$$
$$q_\lambda(x,y) = \lambda q_1(x,y) + (1-\lambda)q_2(x,y)$$

By the definition of q_i we have $q_\lambda(x,y) = (\lambda p_1(y) + (1-\lambda)p_2(y))u(x) = p_\lambda(y)u(x)$ To establish concavity we need to show that $H_{p_\lambda}(X|Y) \ge \lambda H_{p_1}(X|Y) + (1-\lambda)H_{p_2}(X|Y)$. From Eq. 12 we have $H_{p_\lambda}(X|Y) = -D_{KL}[p_\lambda|q_\lambda] + \log|X|$, and convexity of D_{KL} yields

$$\begin{aligned} H_{p_\lambda}(X|Y) &\ge -\lambda D_{KL}[p_1(x,y)|q_1(x,y)] - \\ &\quad (1-\lambda)D_{KL}[p_2(x,y)|q_2(x,y)] + \log|X| \\ &= \lambda H_{p_1}(X|Y)) + (1-\lambda)H_{p_2}(X|Y) \end{aligned}$$

References

Bertsekas, D. P. (1995). *Nonlinear programming.* Belmont, MA: Athena Scientific.

Chiang, M. (2005). Geometric programming for communication systems. *Foundations and Trends in Communications and Information Theory, 2*, 1–154.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* Wiley.

Deza, M., & Laurent, M. (1997). *Geometry of cuts and metrics.* Springe-Verlag.

Grotschel, M., Junger, M., & Reinelt, G. (1985). On the acyclic subgraph polytope. *Math. Prog., 33*, 28–42.

Heskes, T., Albers, K., & Kappen, B. (2003). Approximate inference and constrained optimization. *UAI.*

Wainwright, M., & Jordan, M. (2003). *Graphical models, exponential families, and variational inference* (Technical Report). UC Berkeley Dept. of Statistics.

Wainwright, M. J., Jaakkola, T., & Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Trans. on Information Theory, 51*, 2313–2335.

Yedidia, J., W.T. Freeman, W., & Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Information Theory, 51*, 2282–2312.