Chapter 7

Upper bounds based on convex combinations

7.1 Introduction

A fundamental quantity associated with any graph-structured distribution is the log partition function. In most cases, actually computing the log partition function, though a straightforward summation in principle, is NP-hard [40, 48] due to the exponential number of terms. Therefore, an important problem is either to approximate or obtain bounds on the log partition function. There is a large literature on approximation algorithms for the log partition function [e.g., 103,140,141]. A related goal is to obtain upper and lower bounds [e.g., 12,93,94,106,154,179]. The applicability of such bounds on the log partition function is wide; possible uses include approximate inference [e.g., 95,97], model fitting [e.g., 106], decision theory and large deviations analysis [e.g., 158].

An important property of the log partition function is its convexity (see Section 2.2). Mean field theory [e.g., 106], as presented in Section 2.3.1, can be viewed as exploiting one property of a convex function: namely, that the first order tangent approximation is always an underestimate [20]. In Chapter 3, we exploited another fundamental property of convex functions — namely, Jensen's inequality [42] — in order to derive a new class of upper bounds applicable to an arbitrary undirected graphical model. These upper bounds were based on taking a particular convex combination of exponential parameter vectors.

In this chapter, we analyze this new class of bounds in more detail, focusing on the case where the exponential parameter vectors are drawn from some tractable class for which exact computations can be performed efficiently. The canonical example of such a tractable substructure is a tree embedded within the original graph. The weights in the convex combination are specified by a probability distribution $\vec{\mu}$ over the set of tractable substructures.

For any given log partition function, there is an entire family of upper bounds, indexed by the choice of exponential parameters as well as the probability distribution $\vec{\mu}$. It is therefore natural to consider the problem of optimizing both the choice of exponential parameters, and the distribution over tractable subgraphs, so as to obtain the tightest possible bounds on the log partition function. This optimization problem turns out to have an interesting structure. When the distribution $\vec{\mu}$ is fixed, the objective function is strictly convex in the exponential parameters, and so has a unique global minimum. This uniqueness is in sharp contrast to mean field theory, where the associated optimization problem is well-known to suffer from multiple local minima, even for relatively simple problems [e.g., 93].

The line of analysis that we develop here is quite general, in that it applies to discrete random variables assuming an arbitrary number of states m, and in principle to arbitrary sizes of clique potentials. The only restriction that our analysis imposes on the approximating structures themselves is that they correspond to a triangulated graph. However, in order to bring our development into sharp focus, this chapter treats the special case of binary nodes ($m = 2$) and pairwise clique potentials; moreover, we assume that the set of tractable substructures, denoted by $\mathfrak{T}$, corresponds to the set of all spanning trees of the graph G. Based on an understanding of this case, the modifications necessary to deal with higher state numbers ($m > 2$), larger clique sizes, and more complex approximating structures will be clear.

A major challenge to be overcome lies in the dimension of the problem. The length of $\vec{\mu}$ is equal to the number of spanning trees in the graph $G = (V, E)$. Reasonably complex graphs tend to have a very large number of spanning trees — one which grows prohibitively quickly in the number of nodes $N = |V|$. The collection of exponential parameter vectors is even larger by an additional factor of $\mathcal{O}(N)$. Fortunately, we are able to sidestep this combinatorial explosion by applying the theory of Lagrangian duality [20]. Indeed, in the dual formulation, the entire collection of exponential parameters is replaced by a single vector of length $N + |E|$. In addition, the dual function consists of a convex combination of entropy terms of tree-structured distributions, each of which can be decomposed as a sum of single node and edge terms. This local decomposition is the crucial property that permits efficient optimization of the bounds. In particular, for a fixed distribution over spanning trees of the graph, we develop a constrained Newton's method to optimize efficiently the choice of exponential parameters. Simultaneously optimizing both the choice of the exponential parameters and the distribution $\vec{\mu}$ over spanning trees requires a more intensive but computationally tractable algorithm with an inner and outer loop. Interestingly, steps in the outer loop correspond to solving a maximum weight spanning tree problem [108], which can be interpreted as finding the tree that best fits the current data [see 37].

This chapter is organized in the following manner. In Section 7.1.1, we introduce the notation and definitions required for analysis. In Section 7.1.2, we derive the basic form of the upper bounds to be studied in this chapter. The dual formulation of these bounds, which is essential in avoiding the combinatorial explosion described above, is developed in Section 7.2. Section 7.3 builds on this dual formulation by stating and characterizing the optimal form of the upper bounds, first for the case of a fixed distribution, and secondly when both the distribution $\vec{\mu}$ and the collection of exponential parameter vectors are allowed to vary. Section 7.4 is devoted to more practical issues: we present algorithms for computing the optimal form of the bounds specified in Section 7.3. We also present the results of these techniques in application to bounding the log partition function of randomly generated distributions. We finish up in Section 7.5 with a summary and extensions to the work described here.

■ 7.1.1 Set-up

The analysis of this chapter makes heavy use of exponential representations of distributions, and the associated Legendre transform between exponential and mean parameters. The reader should consult Section 2.2 for the relevant background. In this section, we set up the notation necessary for subsequent analysis. So as to provide a relatively self-contained and more readable presentation, we duplicate some material from Section 3.3.3 of Chapter 3, in which upper bounds of this nature were first introduced.

Let $\mathfrak{T} = \mathfrak{T}(G)$ denote the set of all spanning trees of G. We use the symbol $\mathcal{T}$ to refer to a spanning tree in $\mathfrak{T}$. The number $|\mathfrak{T}(G)|$ of spanning trees in a graph G is typically quite large; for instance, a well-known result of Cayley [168] states that the complete graph K_N on N nodes has N^{N-2} spanning trees. More generally, the number of spanning trees in a graph can be computed via the Matrix-Tree theorem [168].

We now define a probability distribution over the set of spanning trees $\mathfrak{T} = \mathfrak{T}(G)$:

$$\vec{\boldsymbol{\mu}} = \{\, \mu(\mathcal{T}),\ \mathcal{T} \in \mathfrak{T} \mid \mu(\mathcal{T}) \geq 0;\ \ \sum_{\mathcal{T}\in\mathfrak{T}} \mu(\mathcal{T}) = 1 \,\} \tag{7.1}$$

The *support* of $\vec{\boldsymbol{\mu}}$ is defined as

$$\operatorname{supp}(\vec{\boldsymbol{\mu}}) \triangleq \{\, \mathcal{T} \in \mathfrak{T} \mid \mu(\mathcal{T}) > 0 \,\} \tag{7.2}$$

In the sequel, it will also be of interest to consider the probability that a given edge $e \in E$ appears in a spanning tree $\mathcal{T}$ chosen randomly under the distribution $\vec{\boldsymbol{\mu}}$.

Definition 7.1.1. Given a distribution $\vec{\boldsymbol{\mu}}$ over spanning trees, the *edge appearance probability* of an edge $e \in E$ is defined as follows:

$$\mu_e \ \triangleq\ \mathbb{E}_{\vec{\boldsymbol{\mu}}}\big\{\delta[e \in \mathcal{T}]\big\} \ =\ \Pr_{\vec{\boldsymbol{\mu}}}\{e \in \mathcal{T}\} \tag{7.3}$$

where $\delta[e \in \mathcal{T}]$ is the indicator function for edge e to appear in tree $\mathcal{T}$. I.e., this is the probability that edge e belongs to a spanning tree chosen randomly under distribution $\vec{\boldsymbol{\mu}}$.

For a random vector $\mathbf{x}$ taking values in $\{0,1\}^N$, let θ^* denote the minimal exponential parameter of a distribution $p(\mathbf{x};\theta^*)$ defined on the graph $G = (V, E)$:

$$p(\mathbf{x};\theta^*) = \exp\Big\{ \sum_{s\in V} \theta^*_s x_s + \sum_{(s,t)\in E} \theta^*_{st} x_s\, x_t - \Phi(\theta^*) \Big\} \tag{7.4}$$

We refer to this quantity as the *target distribution.*

For each spanning tree $\mathcal{T} \in \mathfrak{T}$, let $\theta(\mathcal{T})$ be an exponential parameter vector of the same dimension as θ^* that respects the structure of $\mathcal{T}$. To be explicit, if $\mathcal{T}$ is defined by an edge set $E(\mathcal{T}) \subset E$, then $\theta(\mathcal{T})$ must have zeros in all elements corresponding to edges not in $E(\mathcal{T})$. For a binary process, such a tree-structured distribution has the form:

$$p(\mathbf{x}; \theta(\mathcal{T})) = \exp\big\{ \sum_{s \in V} \theta(\mathcal{T})_s x_s + \sum_{(s,t) \in E(\mathcal{T})} \theta(\mathcal{T})_{st}\, x_s x_t - \Phi(\theta(\mathcal{T}))\big\} \tag{7.5}$$

Since any spanning tree on a connected graph with N nodes has $N-1$ edges, the parameter vector $\theta(\mathcal{T})$ has $d(\theta(\mathcal{T})) \triangleq 2N-1$ non-zero elements for a binary-valued process.

For compactness in notation, let

$$\boldsymbol{\theta} = \{\theta(\mathcal{T}) \mid \mathcal{T} \in \mathfrak{T}\} \tag{7.6}$$

denote the full collection of tree-structured exponential parameter vectors. This quantity can be viewed as a large vector with $[(2N-1)\,|\mathfrak{T}(G)|]$ non-zero elements. The notation $\theta(\mathcal{T})$ specifies those subelements of $\boldsymbol{\theta}$ corresponding to spanning tree $\mathcal{T}$.

■ 7.1.2 Basic form of bounds

The central idea is that of a convex combination of tree-structured parameter vectors:

Definition 7.1.2. Given a distribution $\vec{\boldsymbol{\mu}}$ and a collection of exponential vectors $\boldsymbol{\theta}$, a *convex combination* of exponential parameter vectors is defined via the expectation:

$$\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] \triangleq \sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T})\theta(\mathcal{T}) \tag{7.7}$$

We are especially interested in sets of approximating points $\boldsymbol{\theta}$ for which there exists a convex combination that is equal to θ^*. Accordingly, we define the following set of pairs $(\boldsymbol{\theta}; \vec{\boldsymbol{\mu}})$:

$$\mathcal{A}(\theta^*) \triangleq \Big\{ (\boldsymbol{\theta}; \vec{\boldsymbol{\mu}}) \;\Big|\; \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] = \theta^* \Big\} \tag{7.8}$$

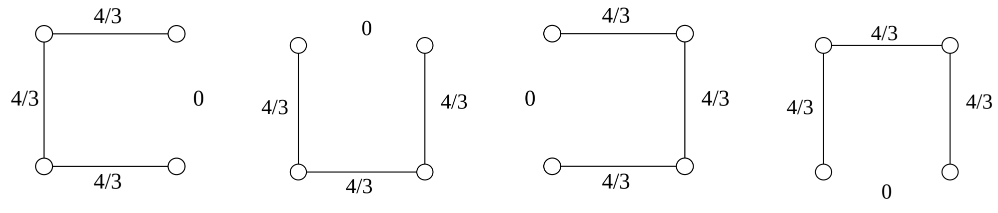
It is not difficult to see that $\mathcal{A}(\theta^*)$ is never empty.

Example 7.1.1. To illustrate these definitions, consider a binary distribution defined by a single cycle on 4 nodes, as shown in Figure 7.1. Consider a target distribution of the form

$$p(\mathbf{x}; \theta^*) = \exp\{x_1x_2 + x_2x_3 + x_3x_4 + x_4x_1 - \Phi(\theta^*)\}$$

That is, the target distribution is specified by the exponential parameter

$$\theta^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$



(a) $\theta(\mathcal{T}_1)$ (b) $\theta(\mathcal{T}_2)$ (c) $\theta(\mathcal{T}_3)$ (d) $\theta(\mathcal{T}_4)$

Figure 7.1. A convex combination of four distributions $p(\mathbf{x};\theta(\mathcal{T}_i))$, each defined by a spanning tree $\mathcal{T}_i$, is used to approximate the target distribution $p(\mathbf{x};\theta^*)$ on the single-cycle graph.

where the zeros represent the fact that $\theta^*_s = 0$ for all $s \in V$. The tractable class consists of the four possible spanning trees $\mathfrak{T} = \{\mathcal{T}_i \mid i = 1, \ldots, 4\}$ on a single cycle on four nodes. We define a set of associated exponential parameters $\boldsymbol{\theta} = \{\theta(\mathcal{T}_i)\}$ as follows:

$$\begin{aligned}
\theta(\mathcal{T}_1) &= (4/3)\begin{bmatrix}0 & 0 & 0 & 0 & 1 & 1 & 1 & 0\end{bmatrix}\\
\theta(\mathcal{T}_2) &= (4/3)\begin{bmatrix}0 & 0 & 0 & 0 & 1 & 1 & 0 & 1\end{bmatrix}\\
\theta(\mathcal{T}_3) &= (4/3)\begin{bmatrix}0 & 0 & 0 & 0 & 1 & 0 & 1 & 1\end{bmatrix}\\
\theta(\mathcal{T}_4) &= (4/3)\begin{bmatrix}0 & 0 & 0 & 0 & 0 & 1 & 1 & 1\end{bmatrix}
\end{aligned}$$

Finally, we choose $\mu(\mathcal{T}_i) = 1/4$ for all $\mathcal{T}_i \in \mathfrak{T}$. It is not difficult to check that this choice of a uniform distribution ensures that $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] = \theta^*$; that is, the specified pair $(\boldsymbol{\theta}; \vec{\boldsymbol{\mu}})$ belongs to $\mathcal{A}(\theta^*)$.

Recall from Lemma 2.2.1 that the log partition function Φ is convex as a function of θ. This property allows us to apply Jensen's inequality [42] to a convex combination specified by a pair $(\boldsymbol{\theta}, \vec{\boldsymbol{\mu}}) \in \mathcal{A}(\theta^*)$; doing so yields the following result:

Proposition 7.1.1. *For any pair $(\boldsymbol{\theta}, \vec{\boldsymbol{\mu}}) \in \mathcal{A}(\theta^*)$, the following upper bound is valid:*

$$\Phi(\theta^*) \;\leq\; \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Phi(\theta(\mathcal{T}))] \triangleq \sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T})\Phi(\theta(\mathcal{T})) \tag{7.9}$$

Note that the bound of equation (7.9) is a function of both the distribution $\vec{\boldsymbol{\mu}}$ over spanning trees; and the collection of tree-structured exponential parameter vectors $\boldsymbol{\theta}$. The primary goal of this chapter is to optimize these choices so as to minimize the RHS of equation (7.9), thereby obtaining the tightest possible upper bound of the form in Proposition 7.1.1. We shall consider first the problem of optimizing the choice of $\boldsymbol{\theta}$ for a fixed $\vec{\boldsymbol{\mu}}$; and then the joint optimization of $\boldsymbol{\theta}$ and $\vec{\boldsymbol{\mu}}$. Despite the relatively simple form of equation (7.9), these optimization problems turn out to have a rich and interesting structure.

7.2 Dual formulation with fixed $\vec{\boldsymbol{\mu}}$

In this section, we develop a Lagrangian dual formulation of the bounds of equation (7.9). For a fixed distribution $\vec{\boldsymbol{\mu}}$, consider the following constrained optimization problem:

$$\begin{cases} \min_{\boldsymbol{\theta}} \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Phi(\theta(\mathcal{T}))] \\ \text{s. t} \qquad \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\boldsymbol{\theta}] = \theta^* \end{cases} \tag{7.10}$$

With $\vec{\boldsymbol{\mu}}$ fixed, the upper bound $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Phi(\theta(\mathcal{T})]$ is strictly convex as a function of $\boldsymbol{\theta}$, and the associated constraint is linear in $\boldsymbol{\theta}$.

We assume that $\vec{\boldsymbol{\mu}}$ is chosen such that the associated edge appearance probabilities $\mu_e = \Pr_{\vec{\boldsymbol{\mu}}}\{e \in \mathcal{T}\}$ are all strictly positive. I.e., all edges $e \in E$ appear in at least one tree $\mathcal{T} \in \text{supp}(\vec{\boldsymbol{\mu}})$. This assumption is necessary to ensure that constraint set $\{ \boldsymbol{\theta} \mid (\boldsymbol{\theta}, \vec{\boldsymbol{\mu}}) \in \mathcal{A}(\theta^*) \}$ is non-empty. By standard results in nonlinear programming [20], problem (7.10) has a unique global minimum, attained at $\widehat{\boldsymbol{\theta}} \equiv \widehat{\boldsymbol{\theta}}(\vec{\boldsymbol{\mu}})$. In principle, a variety of methods [e.g., 20] could be used to solve the convex program (7.10). However, an obvious concern is the dimension of the parameter vector $\boldsymbol{\theta}$; in particular, it is directly proportional to $|\mathfrak{T}|$, the number of spanning trees in G, which is typically very large.

As we will show in this section, the theory of convex duality allows us to neatly avoid this combinatorial explosion. In particular, we show that the Lagrangian dual of problem (7.10) depends on a vector λ of length $N + |E|$, which has the form:

$$\lambda = \{\lambda_s,\ s \in V; \quad \lambda_{st},\ (s,t) \in E\} \tag{7.11}$$

This vector can be viewed as a set of parameters defining the local marginal distributions of a binary process on the single nodes and edges of the original graph G as follows:

$$p(x_s; \lambda) \triangleq [1 - \lambda_s;\ \lambda_s]' \tag{7.12a}$$

$$p(x_s, x_t; \lambda) \triangleq \begin{pmatrix} [1 + \lambda_{st} - \lambda_s - \lambda_t] & \lambda_{st} - \lambda_t \\ \lambda_{st} - \lambda_s & \lambda_{st} \end{pmatrix} \tag{7.12b}$$

To ensure that these definitions make sense as marginal distributions (i.e., their elements lie between zero and one), the vector λ must belong to the following polytope:

$$\mathbb{L}(G) = \{ \lambda \mid 0 \leq \lambda_{st} \leq \lambda_s \leq 1; \quad \lambda_s + \lambda_t \leq 1 + \lambda_{st} \quad \forall\, s \in V,\ (s,t) \in E \} \tag{7.13}$$

Let $\widehat{\boldsymbol{\theta}} = \{ \widehat{\theta}(\mathcal{T}) \mid \mathcal{T} \in \mathfrak{T} \}$ denote the optimum of problem (7.10). The significance of λ is in specifying this optimum in a very compact fashion. For each tree $\mathcal{T} \in \mathfrak{T}$, let

$\Pi^{\mathcal{T}}(\lambda)$ denote the projection of λ onto the spanning tree $\mathcal{T}$. Explicitly,

$$\Pi^{\mathcal{T}}(\lambda) \triangleq \{\lambda_s,\ \lambda_{st} \mid s \in V; \quad (s,t) \in E(\mathcal{T})\ \} \tag{7.14}$$

consists only of those elements of λ belonging to single nodes, or elements of the edge set $E(\mathcal{T}) \subset E$ of the tree $\mathcal{T}$.

Via equation (7.12), any such vector $\Pi^{\mathcal{T}}(\lambda)$ defines a set of marginal distributions for each node $s \in V$ and edge $(s,t) \in E(\mathcal{T})$. These marginals provide an explicit construction of a distribution $p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda))$ by the usual factorization of tree-structured distributions implied by the junction tree representation (see Section 2.1.5) — viz.:

$$p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda)) \triangleq \prod_{s \in V} p(x_s; \lambda) \prod_{(s,t) \in E(\mathcal{T})} \frac{p(x_s, x_t; \lambda)}{p(x_s; \lambda)\ p(x_t; \lambda)} \tag{7.15}$$

The proof of Proposition 7.2.1 below shows that the optimal dual parameter $\widehat{\lambda}$ specifies the full collection of optimal exponential parameters $\widehat{\boldsymbol{\theta}}$ via the relation:

$$p(\mathbf{x}; \widehat{\theta}(\mathcal{T})) = p(\mathbf{x}; \Pi^{\mathcal{T}}(\widehat{\lambda})) \qquad \text{for all} \ \ \mathcal{T} \in \mathfrak{T} \tag{7.16}$$

That is, at the optimum, a single vector $\widehat{\lambda}$ of length $N + |E|$ suffices to specify the full collection $\widehat{\boldsymbol{\theta}} = \{\ \widehat{\theta}(\mathcal{T}) \mid \mathcal{T} \in \mathfrak{T}\ \}$. Consequently, the dual formulation reduces the problem dimension from the size of $\boldsymbol{\theta}$ (which is proportional to $|\mathfrak{T}|$) down to the dimension of λ (namely, $N + |E|$). It is this massive reduction in the problem dimension that permits efficient optimization.

An insight that emerges from our analysis is that the collection of tree-structured distributions $\{\ p(\mathbf{x}; \widehat{\theta}(\mathcal{T})) \mid \mathcal{T} \in \mathfrak{T}\ \}$ has the following remarkable property:

(a) For every $p(\mathbf{x}; \widehat{\theta}(\mathcal{T}))$, the single node marginal probability $p(x_s = 1; \widehat{\theta}(\mathcal{T}))$ is equal to the same constant $\widehat{\lambda}_s$, for all vertices $s \in V$.

(b) For every tree-structured distribution $p(\mathbf{x}; \widehat{\theta}(\mathcal{T}))$ for which the tree $\mathcal{T}$ includes edge (s,t), the corresponding marginal probability $p(x_s = 1, x_t = 1; \widehat{\theta}(\mathcal{T}))$ is equal to the same constant $\widehat{\lambda}_{st}$.

These conditions are very similar to the consistency conditions satisfied by any fixed point of tree-based reparameterization (see Chapter 5). Not surprisingly then, the dual function of Proposition 7.2.1 has a very close relation with the Bethe free energy, as we point out in Section 7.3.4.

■ 7.2.1 Explicit form of dual function

We now state and derive the dual form of problem (7.10). Let $\Psi(\Pi^{\mathcal{T}}(\lambda))$ be the negative entropy of the tree-structured distribution $p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda))$ defined in equation (7.15), and recall the definition of the polytope $\mathbb{L}(G)$ given in equation (7.13).

Proposition 7.2.1 (Dual formulation). *For a fixed weight vector $\vec{\boldsymbol{\mu}}$, we have the equivalent dual formulation of problem* (7.10)*:*

$$\min_{\boldsymbol{\theta}\,\text{s.t}\,\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})]=\theta^*} \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Phi(\theta(\mathcal{T}))] \;=\; \max_{\lambda\in\mathbb{L}(G)} \mathcal{Q}(\lambda;\vec{\boldsymbol{\mu}};\theta^*) \tag{7.17}$$

where

$$\mathcal{Q}(\lambda;\vec{\boldsymbol{\mu}};\theta^*) \triangleq -\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Psi(\Pi^{\mathcal{T}}(\lambda))] + \sum_{\alpha} \lambda_\alpha \theta^*_\alpha \tag{7.18}$$

Proof. For the fixed $\vec{\boldsymbol{\mu}}$, we consider the Lagrangian dual function associated with problem (7.10):

$$\mathcal{Q}(\lambda;\vec{\boldsymbol{\mu}};\theta^*) \;=\; \inf_{\boldsymbol{\theta}} \Big\{\mathcal{L}(\boldsymbol{\theta};\lambda;\vec{\boldsymbol{\mu}};\theta^*)\Big\} \tag{7.19}$$

where the Lagrangian is defined by

$$\mathcal{L}(\boldsymbol{\theta};\lambda;\vec{\boldsymbol{\mu}};\theta^*) = \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Phi(\theta(\mathcal{T})] + \sum_{\alpha} \lambda_\alpha \big\{\theta^*_\alpha - \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})_\alpha]\big\} \tag{7.20}$$

The function $\mathcal{Q}(\lambda;\vec{\boldsymbol{\mu}};\theta^*)$ is a function of the dual variables λ; in particular, λ_α is a Lagrange multiplier associated with the constraint $\theta^*_\alpha - \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})_\alpha] = 0$. In addition to these constraints, each $\theta(\mathcal{T})$ is restricted to correspond to a tree-structured distribution, meaning that certain elements $\theta(\mathcal{T})_\beta$ must be zero. We enforce these zero constraints explicitly without Lagrange multipliers.

Now the Lagrangian is also strictly convex as a function of $\boldsymbol{\theta}$, so that the infimum of equation (7.19) is attained at some value $\widehat{\boldsymbol{\theta}} = \{\widehat{\theta}(\mathcal{T})\}$. By taking derivatives of the Lagrangian with respect to $\boldsymbol{\theta}$, we obtain the stationary conditions for the optimum:

$$\mu(\mathcal{T})\big\{\mathbb{E}_{\widehat{\theta}(\mathcal{T})}[\phi_\alpha] - \widehat{\lambda}_\alpha\big\} = 0 \tag{7.21}$$

where

$$\phi_\alpha(x) = \begin{cases} x_s & \text{if } \alpha = s \in V \\ x_s x_t & \text{if } \alpha = (s,t) \in E \end{cases}$$

If $\mu(\mathcal{T}) = 0$, then the approximating parameter $\theta(\mathcal{T})$ plays no role in the problem, so that we can simply ignore it. Otherwise, if $\mu(\mathcal{T}) > 0$, equation (7.21) implies that for all indices $\alpha \in V \cup E(\mathcal{T})$, the Lagrange multipliers are connected to the optimal approximating parameters $\widehat{\theta}(\mathcal{T})$ via the relation:

$$\mathbb{E}_{\widehat{\theta}(\mathcal{T})}[\phi_\alpha] = \widehat{\lambda}_\alpha \tag{7.22}$$

Recall from Section 2.2.4 that the expectations $\mathbb{E}_{\widehat{\theta}(\mathcal{T})}[\phi_\alpha]$ define a set of mean parameters $\eta(\widehat{\theta}(\mathcal{T}))$ that are dually coupled via the Legendre transform to the exponential parameter $\widehat{\theta}(\mathcal{T})$. Therefore, equation (7.22) has two important implications:

(a) for all $\mathcal{T} \in \text{supp}(\vec{\mu})$ and nodes $s \in V$, the mean parameters $\mathbb{E}_{\widehat{\theta}(\mathcal{T})}[x_s]$ are all equal to a common value $\widehat{\lambda}_s$

(b) similarly, for all $\mathcal{T} \in \text{supp}(\vec{\mu})$ that include edge (s,t), the mean parameters $\mathbb{E}_{\widehat{\theta}(\mathcal{T})}[x_s\, x_t]$ are all equal to a common value $\widehat{\lambda}_{st}$

Since $\widehat{\theta}(\mathcal{T})$ and $\Pi^{\mathcal{T}}(\widehat{\lambda})$ are coupled via the Legendre transform, they correspond to the exponential parameter and mean parameter respectively of the tree-structured distribution $p(\mathbf{x}; \widehat{\theta}(\mathcal{T})) \equiv p(\mathbf{x}; \Pi^{\mathcal{T}}(\widehat{\lambda}))$. Here, as is common in exponential families, we are using the exponential parameter $\widehat{\theta}(\mathcal{T})$ and the mean parameter $\Pi^{\mathcal{T}}(\lambda)$ interchangeably to index the same distribution.[1]

By the Legendre duality between the log partition function and the negative entropy function (see equation (2.32) in Chapter 2), we have the relation:

$$\Phi(\widehat{\theta}(\mathcal{T})) = \sum_{\alpha} \widehat{\theta}(\mathcal{T})_\alpha \widehat{\lambda}_\alpha - \Psi(\Pi^{\mathcal{T}}(\widehat{\lambda})) \tag{7.23}$$

where $\Psi(\Pi^{\mathcal{T}}(\widehat{\lambda}))$ is the negative entropy of $p(\mathbf{x}; \widehat{\theta}(\mathcal{T})) \equiv p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda))$. Substituting equation (7.23) into equation (7.20) yields an explicit expression for the Lagrangian dual function:

$$\mathcal{Q}(\widehat{\lambda}; \vec{\mu}; \theta^*) = -\mathbb{E}_{\vec{\mu}}[\Psi(\Pi^{\mathcal{T}}(\widehat{\lambda}))] + \sum_{\alpha} \widehat{\lambda}_\alpha \theta^*_\alpha$$

Since λ must correspond to a set of mean parameters valid for each node and edge, it is restricted to the polytope $\mathbb{L}(G)$ defined in equation (7.13). The cost function is strictly convex and the constraints are linear, so that strong duality holds [20]; therefore, the optimum dual value $\mathcal{Q}^*(\vec{\mu}; \theta^*) = \max_{\lambda \in \mathbb{L}(G)} \mathcal{Q}(\lambda; \vec{\mu}; \theta^*)$ is equivalent to the global minimum of the primal problem (7.10). $\square$

7.2.2 Characterization of optimal points

In this section, we characterize both the optimizing argument $\widehat{\lambda}$ of the dual problem in Proposition 7.2.1, as well as the corresponding optimum $\widehat{\boldsymbol{\theta}}$ of the original primal problem (7.10). We begin by showing that for finite θ^*, the optimal $\widehat{\lambda}$ always occurs at interior points of the constraint set $\mathbb{L}(G)$. Next, we provide an explicit construction of these optima, specified in terms of θ^*, and the edge appearance probabilities μ_e associated with the distribution $\vec{\mu}$.

Lemma 7.2.1. *If $\|\theta^*\| < \infty$, the optimum $\widehat{\lambda}$ of problem* (7.17) *is always attained at an interior point of* $\mathbb{L}(G)$.

[1] Strictly speaking, we should write $p(\mathbf{x}; \Lambda^{-1}(\Pi^{\mathcal{T}}(\lambda)))$ to mean $p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda))$, where Λ^{-1} is the inverse Legendre mapping from mean parameters to exponential parameters. (See Section 2.2 for more details).

Proof. Consider the dual function of equation (7.18) as a function of λ and θ^*. Observe that the function $-\mathbb{E}_{\vec{\mu}}[\Psi(\Pi^{\mathcal{T}}(\lambda))]$ is strictly concave. Therefore, from the form of equation (7.18), the optimum $\widehat{\lambda} = \widehat{\lambda}(\theta^*)$ and θ^* can be put into one-to-one correspondence via the invertible and continuously differentiable Legendre transform [151]. The domains of $\widehat{\lambda}$ and θ^* are $\mathbb{L}(G)$ and $\mathbb{R}^{N+|E|}$ respectively. Therefore, extreme points of the polyhedron $\mathbb{L}(G)$ are attained only as $\|\theta^*\| \to \infty$. □

The significance of Lemma 7.2.1 is in allowing us to characterize optima of the dual function in equation (7.18) in terms of ordinary gradient conditions. That is, it obviates the need to consider Lagrange multipliers associated with the constraints defining $\mathbb{L}(G)$ in equation (7.13).

Proposition 7.2.2 (Characterization of optimum). *For any $\|\theta^*\| < \infty$, the optimal pair $\widehat{\boldsymbol{\theta}}$ and $\widehat{\lambda}$ are characterized by the relations:*

$$\theta^*_s \;=\; \mathbb{E}_{\vec{\mu}}[\widehat{\theta}(\mathcal{T})_s] \;=\; \log\left[\frac{\widehat{\lambda}_s}{(1-\widehat{\lambda}_s)}\right] + \sum_{t \in \mathcal{N}(s)} \mu_{st} \log\left[\frac{(\widehat{\lambda}_s - \widehat{\lambda}_{st})}{(1+\widehat{\lambda}_{st} - \widehat{\lambda}_s - \widehat{\lambda}_t)}\right] \tag{7.24a}$$

$$\theta^*_{st} \;=\; \mathbb{E}_{\vec{\mu}}[\widehat{\theta}(\mathcal{T})_{st}] \;=\; \mu_{st}\widehat{\theta}_{st} \;=\; \mu_{st}\,\log\left[\frac{(\widehat{\lambda}_{st})\;(1+\widehat{\lambda}_{st} - \widehat{\lambda}_s - \widehat{\lambda}_t)}{(\widehat{\lambda}_s - \widehat{\lambda}_{st})(\widehat{\lambda}_t - \widehat{\lambda}_{st})}\right] \tag{7.24b}$$

where $\mathcal{N}(s) = \{\, t \in V \mid (s,t) \in E \,\}$ is the set of neighbors of node s in G.

Proof. See Appendix D.1. □

Equations (7.24a) and (7.24b) can be viewed as an alternative and more explicit statement of the fact that $p(\mathbf{x}; \widehat{\theta}(\mathcal{T})) = p(\mathbf{x}; \Pi^{\mathcal{T}}(\widehat{\lambda}))$ for all spanning trees $\mathcal{T} \in \mathfrak{T}$. An important implication of equation (7.24b) is that that the optimal exponential edge parameters $\widehat{\theta}(\mathcal{T})_e$ are equal for all spanning trees $\mathcal{T} \in \mathfrak{T}$. From equation (7.24b) and our assumption that $\mu_e > 0$ for all $e \in E$, this common value is given by θ^*_e/μ_e. As a consequence, the only remaining degree of freedom is in the single node parameters $\widehat{\theta}(\mathcal{T})_s$.

Example 7.2.1 (Single cycle). To illustrate Proposition 7.2.2, we return to the single cycle of Example 7.1.1, which has four spanning trees in total. Each edge in the graph appears in 3 of these 4 spanning trees. As a result, the edge appearance probabilities under the uniform distribution (i.e., $\mu(\mathcal{T}_i) = 1/4$ for all $i = 1, \ldots, 4$) are given by $\mu_e = 3/4$ for all edges $e \in E$. From equation (7.24b) and the fact that

$$\theta^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

the optimal exponential edge parameters are given by $\widehat{\theta}_e = \theta^*_e \,/\, (3/4) = 4/3$ for all edges. We now must choose the single node parameters $\widehat{\theta}(\mathcal{T}_i)_s$ of each spanning tree $\mathcal{T}_i$

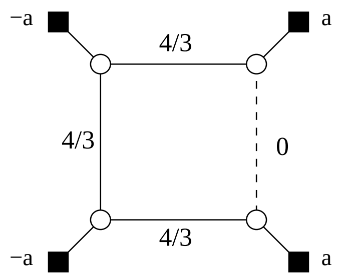


Figure 7.2. Illustration of optimal exponential tree parameter $\widehat{\theta}(\mathcal{T}_1)$ on a single cycle. All edges are assigned weight $(4/3)$ and single node parameters are specified in terms of $a \approx 0.5043$. Other optimal solutions $\widehat{\theta}(\mathcal{T}_i)$ are rotated versions of this one.

to ensure that the mean parameters are all equal as well. Doing so numerically yields optimal solutions of the form:

$$\widehat{\theta}(\mathcal{T}_1) \quad = \quad \begin{bmatrix} -a & -a & +a & +a & (4/3) & (4/3) & (4/3) & 0 \end{bmatrix}$$

where $a \approx 0.5043$. Figure 7.2 gives a graphical illustration of the structure of this solution. Other optimal solutions $\widehat{\theta}(\mathcal{T}_i)$, $i = 2, \ldots, 4$ are obtained by rotating $\widehat{\theta}(\mathcal{T}_1)$. With this specification of optimal solutions, it can be verified that

$$\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\widehat{\theta}(\mathcal{T})] = (1/4)\sum_{\mathcal{T}_i} \widehat{\theta}(\mathcal{T}_i) = \theta^*$$

as required.

■ 7.2.3 Decomposition of entropy terms

A major advantage of the dual formulation of Proposition 7.2.1 is the attendant reduction in the dimension of the problem — namely, from the $\mathcal{O}(N\,|\mathfrak{T}(G)|)$ vector $\boldsymbol{\theta}$ to the $(N + |E|)$-dimensional vector λ. However, a remaining concern with the formulation of Proposition 7.2.1 is the apparent need to calculate an entropy term for all structures $\mathcal{T} \in \operatorname{supp}(\vec{\boldsymbol{\mu}})$. Fortunately, this problem can also be circumvented by using the fact that for a tree-structured distribution, the negative entropy decomposes into a sum of node and edge terms.

In particular, using the form in $p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda))$ in equation (7.15), we can write

$$\begin{aligned} \Psi(\Pi^{\mathcal{T}}(\lambda)) \quad &= \quad \sum_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda)) \log p(\mathbf{x}; \Pi^{\mathcal{T}}(\lambda)) \\ &= \quad -\sum_{s \in V} H_s(\lambda) + \sum_{(s,t) \in E(\mathcal{T})} I_{st}(\lambda) \end{aligned} \qquad (7.25)$$

where

$$\begin{aligned}
H_s(\lambda) &= -\lambda_s \log \lambda_s - (1-\lambda_s)\log(1-\lambda_s) && \text{(7.26a)}\\
H_{st}(\lambda) &= -\lambda_{st}\log\lambda_{st} - (1+\lambda_{st}-\lambda_s-\lambda_t)\log(1+\lambda_{st}-\lambda_s-\lambda_t) \\
&\quad -(\lambda_s-\lambda_{st})\log(\lambda_s-\lambda_{st}) - (\lambda_t-\lambda_{st})\log(\lambda_t-\lambda_{st}) && \text{(7.26b)}\\
I_{st}(\lambda) &= H_s(\lambda) + H_t(\lambda) - H_{st}(\lambda) && \text{(7.26c)}
\end{aligned}$$

are the entropy of the single node distribution $p(x_s;\lambda)$, the joint marginal entropy of $p(x_s, x_t;\lambda)$, and the mutual information between x_s and x_t under $p(x_s, x_t;\lambda)$ respectively.

Using the decomposition of equation (7.25), we obtain:

$$\begin{aligned}
\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Psi(\Pi^{\mathcal{T}}(\lambda))] &= \sum_{\mathcal{T}\in\mathfrak{T}} \mu(\mathcal{T}) \Big\{ -\sum_{s\in V} H_s(\lambda) + \sum_{(s,t)\in E(\mathcal{T})} I_{st}(\lambda) \Big\} \\
&= -\sum_{s\in V} H_s(\lambda) + \sum_{(s,t)\in E} \mu_{st} I_{st}(\lambda) && \text{(7.27)}
\end{aligned}$$

where $\mu_{st} = \Pr_{\vec{\boldsymbol{\mu}}}\{(s,t) \in \mathcal{T}\}$ is the edge appearance probability defined in Definition 7.1.1.

As a consequence, the optimal value of the upper bound depends on the distribution $\vec{\boldsymbol{\mu}}$ *only* via the vector of edge appearance probabilities $\boldsymbol{\mu_e} = \{\mu_e \mid e \in E\}$. In principle, this result allows us to consider optimizing the choice of distribution $\vec{\boldsymbol{\mu}}$ over all spanning trees by appropriately adjusting the vector $\boldsymbol{\mu_e}$. The potential reduction in complexity is significant, since the vector $\boldsymbol{\mu_e}$ has only $|E|$ entries, as opposed to the $|\mathfrak{T}(G)|$ entries of $\vec{\boldsymbol{\mu}}$.

7.2.4 Spanning tree polytope

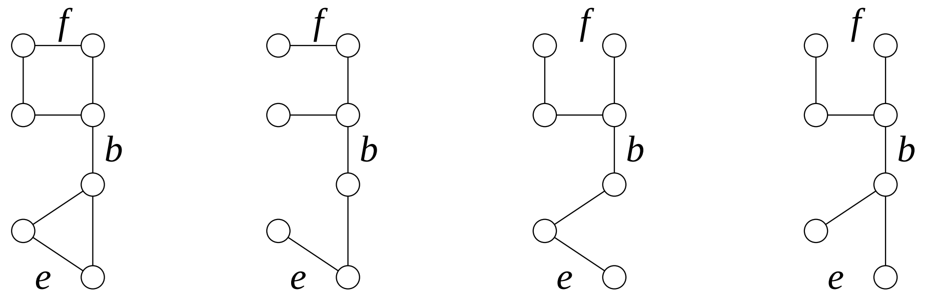
Our focus up to this point has been on the optimization of the tree-structured exponential parameters $\boldsymbol{\theta} = \{\theta(\mathcal{T})\}$ with the distribution $\vec{\boldsymbol{\mu}}$ fixed. We now consider the problem of optimizing $\vec{\boldsymbol{\mu}}$, for which it suffice to optimize the vector of edge appearance probabilities $\boldsymbol{\mu_e}$ (as established by the analysis of the previous section). Any procedure for adjusting the elements of $\boldsymbol{\mu_e}$ needs to ensure that they still correspond to the edge appearance probabilities of a valid distribution $\vec{\boldsymbol{\mu}}$ over spanning trees. I.e., they must belong to the set

$$\mathbb{T}(G) = \{\boldsymbol{\mu_e} \mid \mu_e = \mathbb{E}_{\vec{\boldsymbol{\mu}}}\{\delta[e \in \mathcal{T}]\} \text{ for some } \vec{\boldsymbol{\mu}}; \quad \forall\, e \in E\,\} \tag{7.28}$$

where $\delta[e \in \mathcal{T}]$ is an indicator function for the event that edge e belongs to spanning tree $\mathcal{T}$. Therefore, we need to gain an understanding of the constraints that govern membership in $\mathbb{T}(G)$.

A simple example helps to provide some intuition for the constraints defining the edge appearance probabilities:

Example 7.2.2 (Edge appearance probabilities). Figure 7.3(a) shows a graph, and panels (b) through (d) show three of its spanning trees $\{\mathcal{T}^1, \mathcal{T}^2, \mathcal{T}^3\}$. Suppose that we form a uniform distribution $\vec{\mu}$ over these trees by assigning probability $\mu(\mathcal{T}^i) = 1/3$ to each $\mathcal{T}^i$, $i = 1, 2, 3$. Consider the edge with label f; notice that it appears in $\mathcal{T}^1$, but not in either of $\mathcal{T}^2$ and $\mathcal{T}^3$. Therefore, under the uniform distribution $\vec{\mu}$, the associated edge appearance probability is $\mu_f = 1/3$. Since edge e appears in two of the three spanning trees, similar reasoning establishes that $\mu_e = 2/3$. Finally, observe that edge b



(a) Original (b) $\mu(\mathcal{T}^1) = \frac{1}{3}$ (c) $\mu(\mathcal{T}^2) = \frac{1}{3}$ (d) $\mu(\mathcal{T}^3) = \frac{1}{3}$

Figure 7.3. (a) Original graph. (b)–(d) Three spanning trees of the graph. Each is assigned probability $\mu(\mathcal{T}^i) = 1/3$. Under this uniform distribution, it can be seen that $\mu_f = 1/3$, $\mu_e = 2/3$ and $\mu_b = 1$.

appears in all three spanning trees; therefore, it must have edge appearance probability $\mu_b = 1$, which reflects the fact that it will always appear, no matter which of the three spanning trees is chosen. □

Proceeding with our development, we use the delta function $\delta[e \in \mathcal{T}]$ to define the *spanning tree incidence* vector $\boldsymbol{\nu}(\mathcal{T})$ as follows. For a given spanning tree $\mathcal{T}$, the quantity $\boldsymbol{\nu}(\mathcal{T})$ is a binary-valued vector of length $|E|$ with elements

$$\boldsymbol{\nu}(\mathcal{T})_e = \delta[e \in \mathcal{T}] \tag{7.29}$$

With this definition, we can rewrite the equations $\mu_e = \mathbb{E}_{\vec{\mu}}\{\delta[e \in \mathcal{T}]\}$ that define membership in $\mathbb{T}(G)$ in a vector form as follows:

$$\boldsymbol{\mu_e} = \sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T}) \boldsymbol{\nu}(\mathcal{T}) \tag{7.30}$$

Equation (7.30) shows that the set $\mathbb{T}(G)$ is the convex hull of the set of spanning tree incidence vectors $\{\ \boldsymbol{\nu}(\mathcal{T}) \mid \mathcal{T} \in \mathfrak{T}\ \}$. For this reason, we refer to $\mathbb{T}(G)$ as the *spanning tree polytope*.

By standard results on polyhedra [20], the set $\mathbb{T}(G)$ must have an equivalent characterization in terms of a set of linear inequalities. Fortunately, the spanning tree polytope is a well-studied object in combinatorial optimization and matroid theory [e.g., 35, 60]. The following lemma, based on a result of Edmonds [60], provides such a characterization of $\mathbb{T}(G)$ in terms of linear relations:

Lemma 7.2.2 (Spanning tree polytope). *Given a graph $G = (V, E)$, and any subset $F \subseteq E$, define the rank of F as $r(F) = v(F) - c(F)$, where $v(F)$ is the number of vertices covered by edges in F, and $c(F)$ is the number of connected components of the subgraph induced by F. Then the spanning tree polytope $\mathbb{T}(G)$ is characterized by the following linear relations:*

$$\begin{aligned} \sum_{e \in F} \mu_e &\leq r(F) \qquad \forall \quad F \subseteq E && (7.31a) \\ \sum_{e \in E} \mu_e &= N - 1 && (7.31b) \\ \mu_e &\geq 0 \qquad \forall \quad e \in E && (7.31c) \end{aligned}$$

□

Note that as defined, the rank function $r(F)$ of a subset $F \subseteq E$ corresponds to the number of edges in the largest acyclic subgraph of $G(F)$. For example, when $F = E$, then the largest acyclic subgraph is a spanning tree. Since any spanning tree has $N - 1$ edges, we have $r(E) = N - 1$. The term *rank function* comes from matroid theory [178]; for those readers familiar with matroids, spanning trees are bases in the so-called graphic matroid.

In order to gain some intuition for the constraints in equation (7.31), we consider some particular cases. The necessity of the non-negativity constraints in equation (7.31c) is clear, since the μ_e correspond to edge appearance probabilities. The corresponding upper bounds $\mu_e \leq 1$ are obtained by choosing a single edge set $F = \{e\}$. In this case, we have $v(F) = 2$ and $c(F) = 1$, so that $r(F) = 1$. Equation (7.31a) thus reduces to the constraint $\mu_e \leq 1$. Next, equation (7.31b) can be deduced with the following reasoning. Let $\vec{\boldsymbol{\mu}} = \{\mu(\mathcal{T})\}$ be the distribution giving rise to the edge appearance probabilities $\boldsymbol{\mu_e}$. Then

$$\begin{aligned} \sum_{e \in E} \mu_e &= \sum_{e \in E} \sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T}) \boldsymbol{\nu}(\mathcal{T})_e \\ &= \sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T}) \sum_{e \in E} \boldsymbol{\nu}(\mathcal{T})_e \\ &= N - 1 \end{aligned}$$

where we have used the fact that $\sum_{e \in E} \boldsymbol{\nu}(\mathcal{T})_e = N - 1$ (since any spanning tree $\mathcal{T}$ on N nodes has $N - 1$ edges); and the fact that $\sum_{\mathcal{T}} \mu(\mathcal{T}) = 1$.

Note that equation (7.31a) captures a large number of linear inequalities — one for each subset F of the edge set. For certain choices of F, such inequalities capture more subtle constraints that arise from the particulars of graph structure. For instance, consider a connected graph in which the edge b is a bridge. I.e. removing the edge b breaks the graph into two components. As an illustration, edge b in the graph of Figure 7.3(a) is a bridge. Clearly, any spanning tree $\mathcal{T}$ of G must include the edge b, which implies that $\mu_b = 1$ for any[2] valid spanning tree distribution over this graph.

For the bridge b, the constraint $\mu_b = 1$ is captured by setting $F = E \backslash b$ in equation (7.31a). Note that $r(E\backslash b) = N - 2$, since the graph induced by $E\backslash b$ covers N vertices, and has two connected components. Therefore, equation (7.31a) indicates that

$$\sum_{e \in E\backslash b} \mu_e \leq r(E\backslash b) = N - 2$$

At the same time, from equation (7.31b), we also have that $\sum_{e\in E\backslash b} \mu_e + \mu_b = N - 1$. These two equations, along with $\mu_b \leq 1$, imply that $\mu_b = 1$ as necessary.

Some of the constraints in equation (7.31a) turn out to be redundant; indeed, it suffices to impose bounds of the form in equation (7.31a) only for a certain collection of subsets F of the edge set [142]. To characterize these subsets, recall the following definitions from the background in Section 2.1.1. Given a subset $S \subseteq V$ of the vertex set, the node-induced subgraph $G[S]$ is the subgraph of G induced by S. That is, $G[S] = (S, E[S])$ where

$$E[S] \triangleq \{\ (s,t) \in E \mid s,t \in S\ \}$$

A *cut vertex* or *cut node* in a graph G is a member of V whose removal from G increases the number of components.

With this terminology, we can now define the relevant collection of edge subsets:

Definition 7.2.1. A subset $F \subset E$ is a *critical subset* means that:

(a) F corresponds to the edge set $E[S]$ of the induced graph $G[S]$, for some subset $S \subseteq V$ of the vertex set, and

(b) the induced graph $G[S]$ is connected and contain no cut nodes.

Any singleton set $\{e\} \subset E$ is also critical.

An important result in polyhedral combinatorics [see 142] asserts that a polytope of the form of $\mathbb{T}(G)$ can be characterized using constraints of equation (7.31a) only for critical subsets $F \subseteq E$. The number of such critical subsets is at most 2^N, corresponding to the edge sets $E[S]$ of the graphs induced by all 2^N possible subsets S of the vertex set V. Since $|E| \geq N$ for any connected graph with cycles, this may be a substantial reduction relative to the total number of edge subsets ($2^{|E|}$). However, it is still an intractable number of constraints in general.

[2] The analysis of Example 7.2.2 showed that $\mu_b = 1$ for a particular distribution that placed positive probability on only three spanning trees. Since b is a bridge in the graph, this statement holds regardless of the choice of distribution.

Example 7.2.3. In certain cases, condition (b) of Definition 7.2.1 can lead to a substantial reduction in the number of critical subsets (relative to 2^N). For a graph consisting of a single cycle, consider any node-induced subgraph $G[S]$ with $3 \leq |S| < N$. It can be seen such a graph either has a cut node (Figure 7.4(a)), or is not connected (Figure 7.4(b)). Therefore, the only critical subsets for a single cycle are the singleton

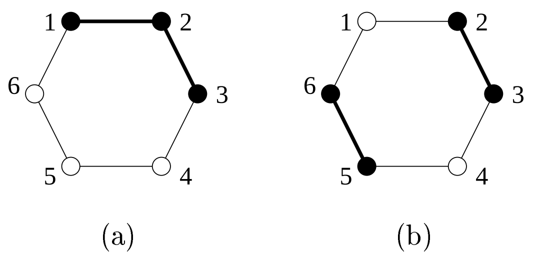


Figure 7.4. For a single cycle, each node-induced subgraph $G[S]$ with $3 \leq |S| < N$ either has cut node, or is not connected. (a) With $S = \{1, 2, 3\}$, node 2 is a cut node in the induced subgraph $G[S]$. (b) With $S = \{2, 3, 5, 6\}$, the induced subgraph $G[S]$ is not connected.

edge sets $\{e\}$, and the full edge set E. Consequently, the conditions $0 \leq \mu_e \leq 1$ and $\sum_{e \in E} \mu_e = N - 1$ are sufficient to characterize $\mathbb{T}(G)$ for a single cycle.

■ 7.3 Jointly optimal upper bounds

In this section, we begin by specifying the form of the optimal upper bounds on the log partition function $\Phi(\theta^*)$, where the optimization takes place both over the dual variables λ and the set of edge appearance probabilities $\boldsymbol{\mu_e}$. We then turn to characterization of the optimizing arguments $(\widehat{\lambda}, \widehat{\boldsymbol{\mu_e}})$. Finally, we point out some connections between the cost function central to our bounds, and the Bethe free energy of statistical physics [183].

■ 7.3.1 Optimal upper bounds on $\Phi(\theta^*)$

In order to make links with the Bethe free energy, it is convenient to define a function $\mathcal{F}$ that corresponds to the negative of the dual function $\mathcal{Q}$ defined in equation (7.18). The key insight of Section 7.2.3 is that the expectation $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Psi(\Pi^{\mathcal{T}}(\lambda))]$ depends on $\vec{\boldsymbol{\mu}}$ only via the *edge appearance probabilities* $\boldsymbol{\mu_e}$ (see Definition 7.1.1). Thus, using the decomposition of entropy terms given in equation (7.27), we can express $-\mathcal{Q}(\lambda; \vec{\boldsymbol{\mu}}; \theta^*)$ as follows:

$$\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*) = -\sum_{s \in V} H_s(\lambda) + \sum_{(s,t) \in E} \mu_{st} I_{st}(\lambda) - \sum_{\alpha} \lambda_\alpha \theta^*_\alpha \tag{7.32}$$

where the entropy H_s and mutual information I_{st} are defined in equations (7.26a) and (7.26c) respectively. All of our upper bounds will be expressed in terms of this function evaluated at particular values of $\lambda \in \mathbb{L}(G)$ (as defined in equation (7.13)) and $\boldsymbol{\mu_e} \in \mathbb{T}(G)$ (as defined in equation (7.28)).

Theorem 7.3.1 (Optimal upper bounds).
For an arbitrary $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, define the function

$$\mathcal{H}(\boldsymbol{\mu_e}; \theta^*) \triangleq \min_{\lambda \in \mathbb{L}(G)} \left\{ \mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*) \right\} \tag{7.33}$$

(a) For any $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, the log partition function is bounded above as follows:

$$\Phi(\theta^*) \quad \leq \quad -\mathcal{H}(\boldsymbol{\mu_e}; \theta^*) \tag{7.34}$$

Moreover, for all $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, the minimum defining $\mathcal{H}$ in equation (7.33) is attained at a unique vector $\widehat{\lambda} \equiv \widehat{\lambda}(\boldsymbol{\mu_e}) \in \mathbb{L}(G)$.

(b) In addition, we have an upper bound, jointly optimal over both λ and $\boldsymbol{\mu_e}$, of the form:

$$\Phi(\theta^*) \quad \leq \quad - \max_{\boldsymbol{\mu_e} \in \mathbb{T}(G)} \mathcal{H}(\boldsymbol{\mu_e}; \theta^*) \tag{7.35}$$

Proof. (a) For each $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, there exists a corresponding distribution $\vec{\boldsymbol{\mu}}$ that realizes the edge appearance probabilities $\boldsymbol{\mu_e}$. Therefore, the function $\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*)$ is equivalent to a function of the form $-\mathcal{Q}(\lambda; \vec{\boldsymbol{\mu}}; \theta^*) = \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\Psi(\Pi^{\mathcal{T}}(\lambda))] - \sum_\alpha \lambda_\alpha \theta^*_\alpha$ for some distribution $\vec{\boldsymbol{\mu}}$ over spanning trees $\mathcal{T} \in \mathfrak{T}$. Based on this relation, the upper bound of equation (7.34) follows from Proposition 7.2.1. The negative entropy $\Psi(\Pi^{\mathcal{T}}(\lambda))$ is strictly concave as a function of λ, so that $\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*) = -\mathcal{Q}(\lambda; \vec{\boldsymbol{\mu}}; \theta^*)$ is strictly convex as a function of λ. Consequently, the associated minimization problem (with linear constraints on λ) has a unique global minimum $\widehat{\lambda}(\boldsymbol{\mu_e})$.

(b) The bound of equation (7.34) holds for all $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, from which equation (7.35) follows. Observe that $\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*)$ is linear in $\boldsymbol{\mu_e}$. Therefore, $\mathcal{H}(\boldsymbol{\mu_e}; \theta^*)$ is the minimum of a collection of linear functions, and so is concave as a function of $\boldsymbol{\mu_e}$ [20]. Consequently, $\mathcal{H}(\boldsymbol{\mu_e}; \theta^*)$ has a global maximum attained at some point $\widehat{\boldsymbol{\mu_e}}$, which yields the optimal value of the upper bound in equation (7.35). The corresponding optimal λ given by $\widehat{\lambda} \equiv \widehat{\lambda}(\widehat{\boldsymbol{\mu_e}})$. □

We illustrate Theorem 7.3.1 by following up the single cycle case of Example 7.2.1.

Example 7.3.1. Consider a single cycle on four nodes, as shown in Figure 7.5(a).

(a) We begin with the choice of exponential parameter

$$\theta^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

as in Example 7.2.1, and fix uniform edge appearance probabilities $\mu_e = (3/4)$ for all edges $e \in E$. The optimal mean parameter for the bound of Theorem 7.3.1(a) can be calculated as

$$\widehat{\lambda}(\boldsymbol{\mu_e}) = \begin{bmatrix} b_1 & b_1 & b_1 & b_1 & \quad b_2 & b_2 & b_2 & b_2 \end{bmatrix}$$

where $b_1 \approx 0.8195$ and $b_2 \approx 0.7069$. This yields the optimal upper bound of $-\mathcal{F}(\widehat{\lambda}; 3/4; \theta^*) \approx 4.6422$ on the true log partition function $\Phi(\theta^*) \approx 4.6252$.

By symmetry of the problem, it can be inferred that the uniform choice of edge appearance probabilities is indeed optimal in the sense of Theorem 7.3.1(b).

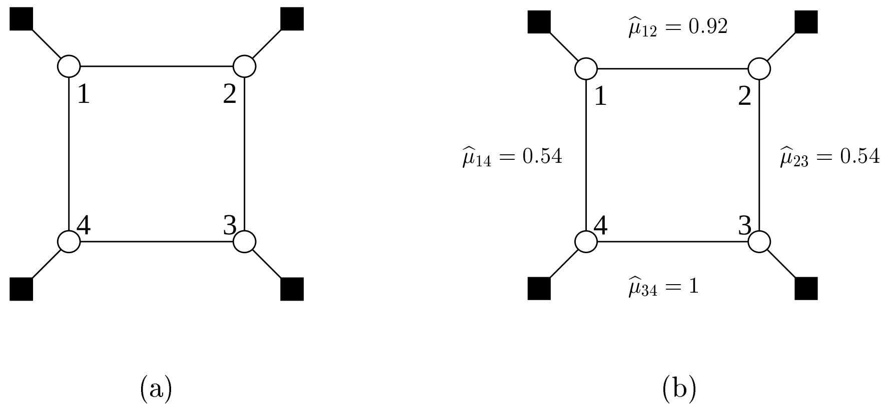


Figure 7.5. Illustration of optimality conditions on a single cycle. (a) Graph consisting of a single cycle. (b) Optimal edge appearance probabilities for $\theta^* = [0\ 0\ 0\ 0\ 1\ 1\ 1\ 3]^T$.

(b) Now consider the same single cycle, but the non-symmetric choice of exponential parameter

$$\theta^* = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 3 \end{bmatrix}$$

If we choose uniform (3/4) edge appearance probabilities, then we obtain an upper bound $-\mathcal{F}(\widehat{\lambda}; 3/4; \theta^*) \approx 6.3451$, optimal in the sense of Theorem 7.3.1(a), on the log partition function $\Phi(\theta^*) \approx 6.3326$.

Given the inhomogeneous nature of θ^*, it is appropriate to consider joint optimization over both λ and $\boldsymbol{\mu_e}$, as dictated by Theorem 7.3.1(b). Performing this optimization using Algorithm 7.4.2 of Section 7.4, we obtain the following optimal edge appearance probabilities:

$$\widehat{\boldsymbol{\mu_e}} \approx \begin{bmatrix} 0.92 & 0.54 & 0.54 & 1 \end{bmatrix} \tag{7.36}$$

Note that the optimum assigns edge appearance probability of one to the edge with largest weight (i.e., the single edge with weight 3). As a result, this edge must appear in any spanning tree in the support of the optimizing distribution $\widehat{\vec{\boldsymbol{\mu}}}$. This set of edge appearance probabilities, combined with the associated $\widehat{\lambda}(\widehat{\boldsymbol{\mu}_e})$, yields the upper bound $\mathcal{H}(\widehat{\boldsymbol{\mu}_e}; \theta^*) \approx 6.3387$ on the true log partition function $\Phi(\theta^*) \approx 6.3326$. This upper bound is tighter than the previous bound (≈ 6.3451) based on uniform edge appearance probabilities.

■ 7.3.2 Alternative proof

In this section, we present an alternative proof of part (a) of Theorem 7.3.1, which provides a complementary perspective to the approach of taking convex combinations in the exponential domain. We begin with the variational representation of $\Phi(\theta^*)$, as guaranteed by the Legendre transform (see Section 2.2.4):

$$\Phi(\theta^*) = \max_{\eta \in \mathbb{M}(G)} \left\{ \eta^T \theta^* - \Psi(\eta) \right\} \tag{7.37}$$

where $\mathbb{M}(G)$ is the set of valid mean parameters. (I.e., $\mathbb{M}(G)$ is given by the range $Ra(\Lambda)$ of the Legendre mapping).

By the convexity of the negative entropy Ψ, for any tree-structured set of mean parameters η^{tree}, we have

$$\Psi(\eta) \quad \geq \quad \Psi(\eta^{\text{tree}}) + \sum_{\alpha} \theta^{\text{tree}}_\alpha (\eta_\alpha - \eta^{\text{tree}}_\alpha) \tag{7.38}$$

Here we have used the fact that $\frac{\partial \Psi}{\partial \eta_\alpha}(\eta^{\text{tree}}) = \theta^{\text{tree}}_\alpha$. For a fixed tree $\mathcal{T}$ specified by edge set $E(\mathcal{T}) \subset E$, the lower bound of equation (7.38) is tightest, in fact, for the *moment-matched* tree distribution

$$\eta^{\text{tree}} = \Pi^{\mathcal{T}}(\eta) = \{ \eta_s, \;\; \eta_{st} \;\mid\; s \in V; \quad (s,t) \in E(\mathcal{T}) \} \tag{7.39}$$

This fact is most easily seen by noting that the difference between the LHS and RHS of equation (7.38) is equivalent to the KL divergence $D(\eta \parallel \eta^{\text{tree}})$ between the distributions $p(\mathbf{x}; \eta)$ and $p(\mathbf{x}; \eta^{\text{tree}})$. (See equation (2.33) for this dual representation of the KL divergence). Therefore, the problem of maximizing the lower bound on the RHS of equation (7.38) is equivalent to minimizing this KL divergence, which corresponds to an I-projection onto the e-flat manifold of tree-structured distributions. (See Section 2.2.7 for details on these notions from information geometry). Therefore, the optimal tree parameter is given by the moment matching procedure specified in equation (7.39).

For this choice of η^{tree}, equation (7.38) reduces to the simpler form

$$\Psi(\eta) \quad \geq \quad \Psi(\Pi^{\mathcal{T}}(\eta)) \tag{7.40}$$

since the mean parameters of η and η^{tree} are equal for indices corresponding to single nodes or edges in the tree, and $\theta^{\text{tree}}_\alpha = 0$ for all other indices (corresponding to edges not in the tree). This is a statement of the fact that any distribution on a graph has lower[3] entropy than a moment-matched distribution structured according to any of its spanning trees.

Since equation (7.40) holds for any spanning tree $\mathcal{T}$, we can consider taking a convex combination of such bounds, each weighted by some $\mu(\mathcal{T}) \geq 0$. This yields the weighted lower bound:

$$\Psi(\eta) \;\geq\; \sum_{\mathcal{T}} \mu(\mathcal{T})\Psi(\Pi^{\mathcal{T}}(\eta)) \;=\; \mathbb{E}_{\vec{\mu}}[\Psi(\Pi^{\mathcal{T}}(\eta))] \tag{7.41}$$

Finally, applying the bound of equation (7.41) to the original variational formulation of equation (7.37) yields:

$$\begin{aligned} \Phi(\theta^*) &\leq \max_{\eta \in \mathbb{M}(G)} \left\{ \eta^T \theta^* - \mathbb{E}_{\vec{\mu}}[\Psi(\Pi^{\mathcal{T}}(\eta))] \right\} && (7.42\text{a}) \\ &\leq \max_{\eta \in \mathbb{L}(G)} \left\{ \eta^T \theta^* - \mathbb{E}_{\vec{\mu}}[\Psi(\Pi^{\mathcal{T}}(\eta))] \right\} && (7.42\text{b}) \\ &= -\min_{\eta \in \mathbb{L}(G)} \mathcal{F}(\eta; \boldsymbol{\mu_e}; \theta^*) && (7.42\text{c}) \end{aligned}$$

where equation (7.42a) follows from the bound of equation (7.41); and equation (7.42b) follows because the set of tree-consistent mean parameters $\mathbb{L}(G)$ is a superset of the set $\mathbb{M}(G)$ of globally G-consistent mean parameters. Finally, equation (7.42c) is obtained in a straightforward manner by decomposing the negative entropies into node and edge terms, and using the definition of $\mathcal{F}$ in equation (7.32). This final equation (7.42c) is equivalent to equation (7.34) in the statement of Theorem 7.3.1(a).

This alternative proof shows why moment-matched tree distributions arise in the optimal form of the bounds. From this alternative derivative, it is easily seen how to extend our analysis from spanning trees (i.e., acyclic graphs) to hypertrees (i.e., acyclic hypergraphs). More background on hypergraphs and hypertrees can be found in the book [17] and survey article [25]. Indeed, a lower bound analogous to that of equation (7.38) can be derived for any triangulated subgraph of the original graph. A similar argument will establish that moment-matching again yields the optimal form of the bound, as in equation (7.40). Finally, we can use a weighted combination of negative entropies, thereby yielding upper bounds analogous to that of equation (7.42).

■ 7.3.3 Characterization of joint optima

In this section, we provide a number of results characterizing the joint optima $(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}}), \widehat{\boldsymbol{\mu_e}})$ of Theorem 7.3.1(b). In particular, we show that these optima can be characterized in terms of a balancing of mutual information terms $I_{st}(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}}))$ on each edge (s,t) of the graph. Moreover, we develop a geometric relation between this mutual information vector, and spanning tree incidence vectors $\boldsymbol{\nu}(\mathcal{T})$. Finally, these results lead to a minimax result that has a game-theoretic flavor [170].

[3]Remember that Ψ is *negative* entropy.

Our first result is a characterization of the mutual information terms $I_{st}(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}}))$, which arise from the optimal pair $(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}}), \widehat{\boldsymbol{\mu_e}})$.

Proposition 7.3.1 (Characterization of joint optima). *Let I_{st} be the mutual information as defined in equation (7.26c), and let F range over all critical subsets (of the edge set).*

(a) There exist Lagrange multipliers ξ_0 and $\xi(F) \geq 0$ such that for each $(s,t) \in E$, the optimum $(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}}), \widehat{\boldsymbol{\mu_e}})$ is characterized by the following conditions:

$$I_{st}(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}})) \quad = \quad \xi_0 + \sum_{F \ni (s,t)} \xi(F) \tag{7.43}$$

The sum $\sum_{F \ni (s,t)}$ ranges over all critical subsets F that include edge (s,t). Moreover, for each critical subset F, we have $\xi(F) > 0$ only if the rank constraint $\sum_{e \in F} \mu_e \leq r(F)$ of equation (7.31a) is met with equality.

(b) Letting $\langle \, , \, \rangle$ denote the Euclidean inner product, we have:

$$\langle I(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}})), \, \boldsymbol{\nu}(\mathcal{T}) - \widehat{\boldsymbol{\mu_e}} \rangle \leq 0 \qquad \forall \quad \mathcal{T} \in \mathfrak{T} \tag{7.44}$$

for the incidence vector $\boldsymbol{\nu}(\mathcal{T})$ of any spanning tree $\mathcal{T} \in \mathfrak{T}$.

Since $\widehat{\boldsymbol{\mu_e}} \in \mathbb{T}(G)$, there exists a distribution over spanning trees $\widehat{\vec{\boldsymbol{\mu}}}$ that gives rise to this set of edge appearance probabilities. Inequality (7.44) holds with equality for all spanning trees $\mathcal{T}$ in the support of $\widehat{\vec{\boldsymbol{\mu}}}$ (i.e., spanning trees for which $\widehat{\mu}(\mathcal{T}) > 0$).

Proof. See Appendix D.2. □

Equation (7.43) of Proposition 7.3.1(a) specifies that at the optimum $(\widehat{\lambda}, \widehat{\boldsymbol{\mu_e}})$, the mutual information on each edge (s,t) of the graph (i.e., between the random variables x_s and x_t) is balanced in a certain sense. The mutual information on edge (s,t) consists of a baseline amount ξ_0, to which we add varying amounts of additional information (i.e., $\sum_{F \ni (s,t)} \xi(F) \geq 0$), depending on how many critical sets F corresponding to an active constraint involve edge (s,t).

Example 7.3.2 (Optimal information terms for a single cycle). The conditions of Proposition 7.3.1(a) take a particularly simple form for a single cycle, where the only critical subsets F are formed of single edges. In this case, we have

$$I_{st}(\widehat{\lambda}(\widehat{\boldsymbol{\mu_e}})) = \begin{cases} \xi_0 & \text{if } \widehat{\mu}_{st} < 1 \\ \xi_0 + \xi[(s,t)] & \text{if } \widehat{\mu}_{st} = 1 \end{cases} \tag{7.45}$$

Thus, the Lagrangian conditions correspond to an equalization of mutual information on edges. The mutual information is equal to a constant for all edges (s,t) with appearance probability $\widehat{\mu}_{st} < 1$; the mutual information on any edge (s,t) with appearance probability $\widehat{\mu}_{st} = 1$ is boosted by some quantity $\xi[(s,t)] \geq 0$.

To follow up Example 7.3.1(b), the optimal information terms for this problem (with the same ordering of edges as in equation (7.36)) are given by:

$$I_e(\widehat{\lambda}(\widehat{\boldsymbol{\mu}_e}) = \begin{bmatrix}\xi_0 & \xi_0 & \xi_0 & (\xi_0 + \xi_{\{34\}})\end{bmatrix}$$

where $\xi_0 \approx 0.012$ and $\xi_{\{34\}} \approx 0.007$.

The geometric interpretation of Proposition 7.3.1(b) is interesting, as illustrated in Figure 7.6. Each spanning tree incidence vector $\boldsymbol{\nu}(\mathcal{T})$ is an extreme point of the

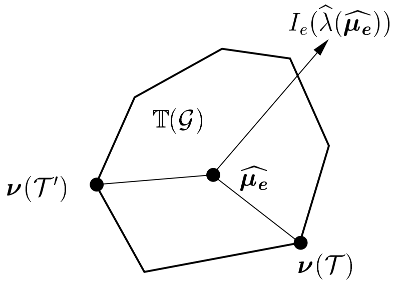


Figure 7.6. Geometry of the optimal edge appearance vector $\widehat{\boldsymbol{\mu}_e}$ in the spanning tree polytope $\mathbb{T}(G)$. Extreme points of this polytope are spanning tree incidence vectors $\boldsymbol{\nu}(\mathcal{T})$. The vector $I_{st}(\widehat{\lambda}(\widehat{\boldsymbol{\mu}_e}))$ forms an obtuse angle with $\boldsymbol{\nu}(\mathcal{T}') - \widehat{\boldsymbol{\mu}_e}$ for all $\mathcal{T}' \in \mathfrak{T}$. It is orthogonal to $\boldsymbol{\nu}(\mathcal{T}) - \widehat{\boldsymbol{\mu}_e}$ whenever $\mathcal{T} \in \text{supp}(\widehat{\boldsymbol{\mu}_e})$.

spanning tree polytope $\mathbb{T}(G)$. The inequality (7.44) indicates that the angle between the information vector $I(\widehat{\lambda}(\widehat{\boldsymbol{\mu}_e}))$ and the difference vector $\boldsymbol{\nu}(\mathcal{T}') - \widehat{\boldsymbol{\mu}_e}$ is obtuse for all spanning trees $\mathcal{T}' \in \mathfrak{T}$. Moreover, this angle is orthogonal for all spanning tree incidence vectors $\boldsymbol{\nu}(\mathcal{T})$ in the support of $\widehat{\vec{\boldsymbol{\mu}}}$ (i.e., trees for which $\widehat{\mu}(\mathcal{T}) > 0$).

Proposition 7.3.1(b) also leads to the following result:

Proposition 7.3.2 (Minimax relation).
For all $\|\theta^*\| < \infty$, the following minimax relation holds:

$$\max_{\boldsymbol{\mu}_e \in \mathbb{T}(G)} \min_{\lambda \in \mathbb{L}(G)} \mathcal{F}(\lambda; \boldsymbol{\mu}_e; \theta^*) = \min_{\lambda \in \mathbb{L}(G)} \max_{\boldsymbol{\mu}_e \in \mathbb{T}(G)} \mathcal{F}(\lambda; \boldsymbol{\mu}_e; \theta^*) \tag{7.46}$$

Proof. The function $\mathcal{F}(\lambda; \boldsymbol{\mu}_e; \theta^*)$ is convex in λ and linear (hence concave) in $\boldsymbol{\mu}_e$. Moreover, the constraint sets $\mathbb{L}(G)$ and $\mathbb{T}(G)$ are both convex and compact. Equation (7.46) therefore follows from standard minimax results [63]. □

Proposition 7.3.2 has an interesting game-theoretic interpretation. Imagine a two-person game specified by the payoff function $\mathcal{F}(\lambda; \boldsymbol{\mu}_e; \theta^*)$. The goal of player 1 is to choose $\lambda \in \mathbb{L}(G)$ so as to minimize this function, whereas the goal of player 2 is to choose a spanning tree (or set of spanning trees) so as to maximize this function. Choosing a single spanning tree can be viewed as a *pure strategy* in the game-theoretic sense [170]. Equation (7.46) specifies an equilibrium condition for the optimal solution pair. In contrast to a pure strategy, this equilibrium involves choosing a distribution $\widehat{\vec{\boldsymbol{\mu}}}$ over spanning trees, which gives rise to the optimal edge appearance vector $\widehat{\boldsymbol{\mu}_e}$. Consequently, the optimal strategy for player 2 is not the deterministic choice of a single spanning tree, but rather the *mixed strategy* of randomly choosing a spanning tree from the specified distribution $\widehat{\vec{\boldsymbol{\mu}}}$.

7.3.4 Relation to Bethe free energy

Recall the cost function that (when optimized) gives rise to bounds on the log partition function:

$$\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*) = -\sum_{s \in V} H_s(\lambda) + \sum_{(s,t) \in E} \mu_{st} I_{st}(\lambda) \; - \sum_{\alpha} \lambda_\alpha \theta^*_\alpha \tag{7.47}$$

The first two terms are a sum of (negative) entropy terms at each node, and a sum of (weighted) mutual information terms for each pair of random variables joined by an edge. Borrowing terminology from statistical physics [136], the final term $\sum_\alpha \lambda_\alpha \theta^*_\alpha$ can be viewed as an "average energy".

The function of equation (7.47) is very closely related to the Bethe free energy [183], as can be seen by comparing equation (7.47) to equation (6.17) of Chapter 6. In fact, the Bethe free energy is a special case of equation (7.47), in which all the edge appearance probabilities μ_e are set equal to 1. This particular choice of $\boldsymbol{\mu_e}$ does *not* belong to the spanning tree polytope $\mathbb{T}(G)$, unless of course G is tree-structured. Therefore, our analysis does not guarantee that the Bethe free energy is convex; indeed, the Bethe free energy fails to be convex for many graphs with cycles, which leads to failures of convergence and multiple local minima.

Nonetheless, this link is interesting. As shown by Yedidia et al. [183], belief propagation (BP) can be viewed as attempting to perform a constrained minimization of the Bethe free energy. It is not surprising, then, that the optimality conditions of our variational formulation (see Proposition 7.2.2) are very closely related to the optimality conditions of tree-based reparameterization or BP (see Theorem 5.4.2 of Chapter 5). This link suggests that our variational formulations might be useful for inference as well. That is, just like the minimizing arguments computed by BP, the optimal $\widehat{\lambda}$ can be taken as approximations to the exact marginals of the distribution $p(\mathbf{x}; \theta^*)$. Open questions include assessing the relative accuracy of $\widehat{\lambda}$ and the BP approximate marginals, and understanding how the edge appearance probabilities $\boldsymbol{\mu_e}$ control the approximation.

7.4 Algorithms and simulation results

In this section, we develop algorithms for carrying out the optimizations specified in Theorem 7.3.1. We then present the results of applying these algorithms to compute upper bounds on the log partition function of randomly chosen instances of certain problems.

■ 7.4.1 Inner minimization over λ

We begin by describing an algorithm for carrying out the inner minimization step (i.e., computing $\min_{\lambda \in \mathbb{L}(G)} \{\mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*)\}$) required to evaluate the upper bound of Theorem 7.3.1(a). The function $\mathcal{F}$ is strictly convex in λ, and the constraint set $\mathbb{L}(G)$ is formed by a set of $\mathcal{O}(N + |E|)$ linear constraints (see equation (7.13)). Therefore, the problem is a suitable candidate for *constrained Newton's method* [20], wherein we take Newton steps projected back onto the constraint set.

The steps involved are given in Algorithm 7.4.1. Computing both the gradient $\nabla \mathcal{F}(\lambda^n; \boldsymbol{\mu_e}; \theta^*)$ and Hessian $\nabla^2 \mathcal{F}(\lambda^n; \boldsymbol{\mu_e}; \theta^*)$ are straightforward tasks. Indeed, since $\mathcal{F}$ decouples into a sum of node and edge terms, the Hessian has a structure that reflects the edge structure of G. (Hence, for sparse graphs, the Hessian will also be sparse). The computation of the descent direction $\widetilde{\lambda}^{n+1}$ in step 2 of Algorithm 7.4.1 is a quadratic program (i.e., minimizing a quadratic function subject to linear constraints); and can be solved efficiently. With suitable choice of step sizes α^n (e.g., via the Armijo rule, or limited minimization rule [see 20]), this algorithm is guaranteed to converge to the unique global minimum $\widehat{\lambda}$. The convergence of Algorithm 7.4.1 is guaranteed to be superlinear in a neighborhood of the optimum with unity step size [20].

Algorithm 7.4.1 (Constrained Newton's method).

1. Initialize $\lambda^0 \in \mathbb{L}(G)$.
2. For iterations $n = 0, 1, 2, \ldots$, compute the descent direction:
$$\widetilde{\lambda}^{n+1} = \arg \min_{\lambda \in \mathbb{L}(G)} \Big\{ \nabla \mathcal{F}(\lambda^n; \boldsymbol{\mu_e}; \theta^*)'(\lambda - \lambda^n) + \frac{1}{2}(\lambda - \lambda^n) \nabla^2 \mathcal{F}(\lambda^n; \boldsymbol{\mu_e}; \theta^*)(\lambda - \lambda^n) \Big\}$$
3. Form the new iterate $\lambda^{n+1} = (1 - \alpha^n)\lambda^n + \alpha^n \widetilde{\lambda}^{n+1}$, where $\alpha^n \in (0, 1]$ is a step size parameter.

■ 7.4.2 Outer maximization over μ_e

We now consider the maximization $\max_{\boldsymbol{\mu_e} \in \mathbb{T}(G)} \mathcal{H}(\boldsymbol{\mu_e}; \theta^*)$ required to compute the upper bound of equation (7.35) in Theorem 7.3.1. Neither the Hessian nor the gradient of $\mathcal{H}$ are difficult to compute. It is therefore tempting to apply a constrained Newton's method once again. However, recall from Lemma 7.2.2 that the spanning tree polytope $\mathbb{T}(G)$ is defined by a very large number ($\mathcal{O}(2^N)$) of linear inequalities. For this reason, solving a constrained quadratic program over $\mathbb{T}(G)$ (as in step 2 of Algorithm 7.4.1) is intractable for large enough graphs.

Fortunately, despite the exponential number of constraints characterizing $\mathbb{T}(G)$, optimizing a linear function subject to the constraint $\boldsymbol{\mu_e} \in \mathbb{T}(G)$ turns out to be straightforward. Two observations are key. First of all, from standard results on linear programming [20], the optimal value of a feasible linear program is always attained at an extreme point of the linear polyhedron formed by the constraints.[4] Secondly, extreme points of the spanning tree polytope $\mathbb{T}(G)$ are given by spanning tree incidence vectors $\boldsymbol{\nu}(\mathcal{T})$, as defined in equation (7.29) [60].

Algorithm 7.4.2 (Conditional gradient).

1. Initialize $\boldsymbol{\mu_e}^0 \in \mathbb{T}(G)$.

2. For iterations $n = 0, 1, 2, \ldots$, compute the ascent direction as follows:

$$\widetilde{\boldsymbol{\mu_e}}^{n+1} = \arg \max_{\boldsymbol{\mu_e} \in \mathbb{T}(G)} \Big\{ \big\langle \nabla \mathcal{H}(\boldsymbol{\mu_e}^n; \theta^*),\ (\boldsymbol{\mu_e} - \boldsymbol{\mu_e}^n) \big\rangle \Big\} \tag{7.48}$$

3. Form the new iterate $\boldsymbol{\mu_e}^{n+1} = (1-\alpha^n)\boldsymbol{\mu_e}^n + \alpha^n \widetilde{\boldsymbol{\mu_e}}^{n+1}$, where $\alpha^n \in (0,1]$ is a step size parameter.

As a consequence, maximizing a linear function over the spanning tree polytope is equivalent to solving a *maximum weight spanning tree* problem [see 108]. Using these facts, it can be seen that the conditional gradient method [20], as specified in Algorithm 7.4.2, is a computationally feasible proposal.

It is helpful to consider the steps of Algorithm 7.4.2 in more detail. Due to the exponential number of constraints defining $\mathbb{T}(G)$, even the first step — that of assessing whether a given vector belongs to $\mathbb{T}(G)$ — is non-trivial. For instance, the uniform assignment $\mu_e = (N-1)/|E|$ need not belong to $\mathbb{T}(G)$. If we are given a distribution $\vec{\boldsymbol{\mu}}$ with a limited support, it is possible to compute the expectations that define the edge appearance probabilities $\boldsymbol{\mu_e}$ (see equation (7.3)) by direct summation. More generally, it turns out to be useful to consider a particular class of distributions over spanning trees, defined by:

$$\vec{\boldsymbol{\mu}}(\mathcal{T}; \omega) \quad \propto \quad \prod_{e \in \mathcal{T}} \omega_e \tag{7.49}$$

where $\omega_e \geq 0$ is a weight assigned to each edge $e \in E$. That is, the probability of a given spanning tree $\mathcal{T}$ is proportional to the product of the weights on all its edges. For

[4]The optimal value may be attained at more than one extreme point, or at interior points as well.

a distribution in this class, Jaakkola et al. [98] showed how a weighted variant of the matrix-tree theorem [22, 168] could be used to compute expectations under $\vec{\boldsymbol{\mu}}(\mathcal{T};\omega)$. This method can be used to compute a feasible starting point $\boldsymbol{\mu_e}^0 \in \mathbb{T}(G)$.

In the second step of Algorithm 7.4.2, given a fixed $\boldsymbol{\mu_e}^n$, we first solve a problem of the form in Theorem 7.3.1(a) using Algorithm 7.4.1 to obtain the optimal $\widehat{\lambda}(\boldsymbol{\mu_e}^n)$. Having obtained this optimal point, computing the ascent direction of equation (7.48) is equivalent to solving a maximum weight spanning tree problem. It can be shown (see Appendix D.2) that

$$\frac{\partial \mathcal{H}}{\partial \mu_e}(\boldsymbol{\mu_e}^n;\theta^*) = I_e(\widehat{\lambda}(\boldsymbol{\mu_e}^n)) \tag{7.50}$$

so that the edge weights in the maximum spanning tree problem are mutual information terms. This computation can be performed efficiently using Kruskal's algorithm [117]; see also [108]. In the third step, the step size $\alpha^n \in (0,1]$ can be chosen via the Armijo rule. (See Bertsekas [20] for details on this and other stepsize selection rules.)

Given equation (7.50), the second and third steps have an interesting interpretation. In particular, let us view the vector of mean parameters $\widehat{\lambda}(\boldsymbol{\mu_e}^n)$ as a set of data, which is used to specify mutual information terms $I_e(\widehat{\lambda}(\boldsymbol{\mu_e}^n))$ on each edge. In this case, finding the corresponding maximum weight spanning tree is equivalent to finding the tree distribution that best fits the data in the maximum likelihood sense (or KL divergence between the empirical distribution specified by the data, and the tree distribution). See Chow and Liu [37] for more details on this interpretation of the maximum weight spanning tree procedure. Therefore, at each iteration, the algorithm moves towards the spanning tree that best fits the current data.[5]

■ 7.4.3 Empirical simulations

In this section, we present the results of applying the previously described algorithms to compute the upper bounds specified in Theorem 7.3.1. We performed simulations for a binary-valued vector $\mathbf{x}$ (taking values in $\{0,1\}$) for two different types of graphs (square grids and a fully connected graph) under two different types of interactions (attractive or mixed potentials). For the purposes of comparison, we also calculated lower bounds using the naive mean field approximation. See Section 2.3.1 for details on mean field.

For each trial on a given graph, we defined a distribution $p(\mathbf{x};\theta^*)$ by randomly choosing an exponential parameter vector θ^* from either the uniform attractive ensemble or the uniform mixed ensemble. See Section 2.2.1 for the definitions of these ensembles of distributions.

Grids of varying sizes

We first performed simulations for square 2-D grids of varying sizes; the number of nodes N was either 9, 36, or 81. For each of these grid sizes and each of the two conditions

[5]There is one minor caveat with this interpretation — namely, as noted previously, the vector $\widehat{\lambda}(\boldsymbol{\mu_e}^n)$ may not correspond to a valid set of marginals for any distribution.

(attractive or mixed), we ran simulations with edge strengths d ranging[6] from 0 to $\frac{4}{\sqrt{N}}$.For each setting of the edge strength, we performed 30 trials for the $N = 9$ grids, and 10 trials for $N = 36$ or 81. The inner minimization $\min_{\lambda \in \mathbb{L}(G)} \mathcal{F}(\lambda; \boldsymbol{\mu_e}; \theta^*)$ was performed using the constrained Newton's method (Algorithm 7.4.1), whereas the outer maximization was performed with the conditional gradient method (Algorithm 7.4.2). In all cases, step size choices were made by the Armijo rule [20]. The value of the actual partition function $\Phi(\theta^*)$ was computed by forming a junction tree for each grid, and performing exact computations on this junction tree.[7]

Shown in Figure 7.7 are plots of the *relative error* [Bound $-\Phi(\theta^*)]/\Phi(\theta^*)$ versus the edge strength (normalized by $4/\sqrt{N}$ for each N so that it falls in the interval $[0, 1]$). The three rows of this figure corresponds to problem sizes $N = 9$, 36 and 81 respectively. Panels on the left correspond to the attractive condition, whereas those on the right correspond to the mixed condition. Each panel displays the relative error in two types of upper bounds. The "unoptimized" curve shows the bound of Theorem 7.3.1(a) with the fixed choice of uniform edge appearance probabilities $\mu_e = (N-1)/|E|$. The "optimized" curve corresponds to the jointly optimal (over both λ and $\vec{\boldsymbol{\mu}}$) upper bounds of Theorem 7.3.1(b). The lower curve in each panel corresponds to the relative error in the naive mean field lower bound.

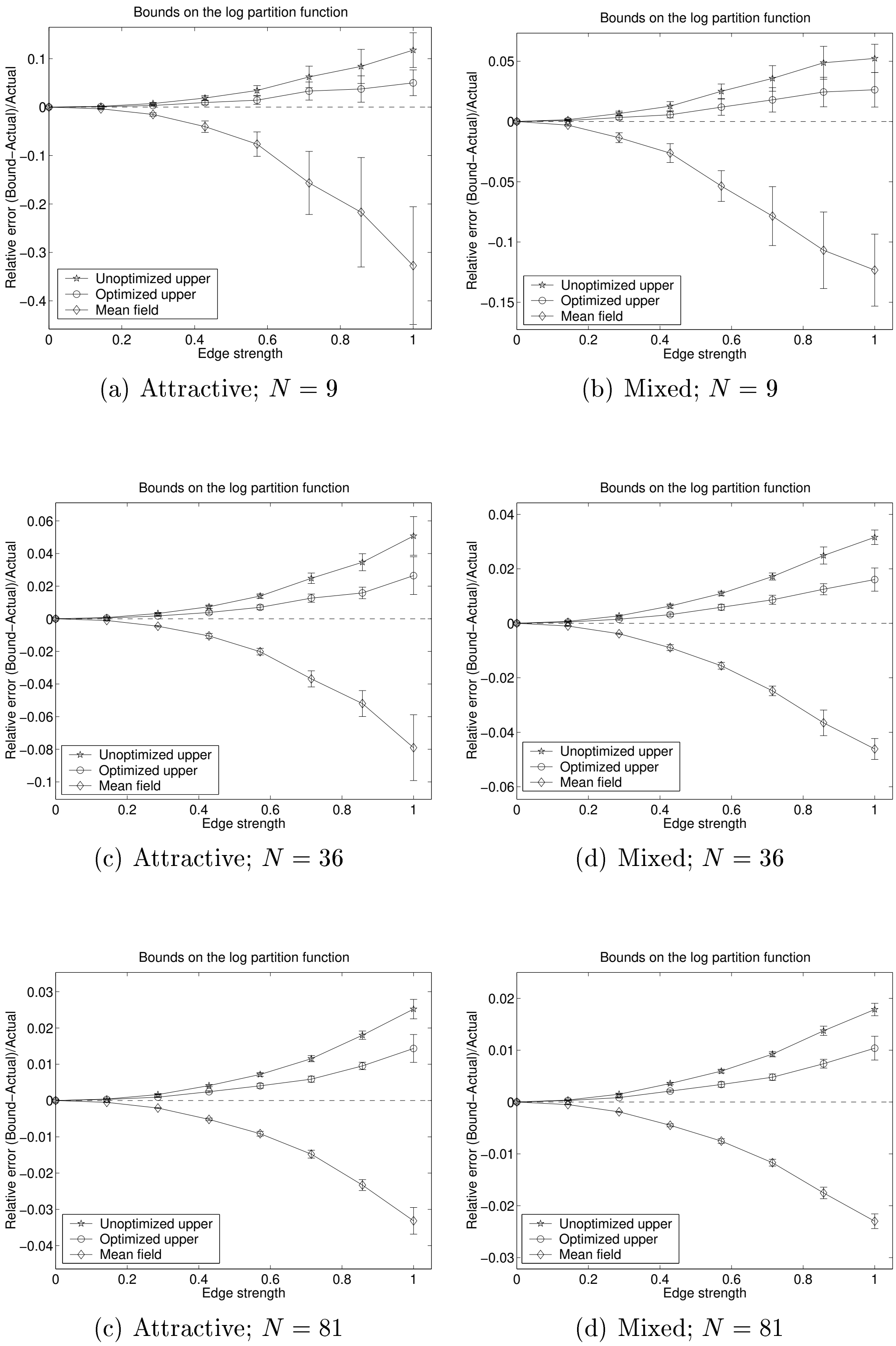
The bounds are tightest for low edge strengths d; their tightness decreases as the edge strength is increased. Optimizing the edge appearance probabilities can lead to significantly better upper bounds. This effect is especially pronounced as the edge strength is increased, in which case the distribution of edge weights θ^*_{st} becomes more inhomogeneous. For these square grids, the tightness of the upper bounds of Theorem 7.3.1 decreases more slowly than the corresponding mean field lower bound. In terms of the relative error plotted here, the upper bounds are superior to the mean field bound by factors of roughly 3 in the attractive case, and roughly 2 in the mixed case. The tightness of the bounds, measured in terms of relative error, decreases slightly as the problem size (number of nodes N) is increased and the edge strengths are rescaled in terms of $1/\sqrt{N}$.

It is worthwhile emphasizing the importance of the dual formulation of our bounds. Indeed, the naive approach of attempting to optimize the primal formulation of the bounds (e.g., see problem (7.10)) would require dealing with a number[8] of spanning trees that ranges from 192 for $N = 9$ nodes, all the way up to the astronomical number $\approx 8.33 \times 10^{33}$ for $N = 81$ nodes.

[6] The normalization by $1/\sqrt{N}$ guarantees that the effects impinging on a given node are scale-invariant; they converge to a fixed distribution as the problem size N tends to infinity.

[7] Thank you to Yee Whye Teh and Max Welling for generosity in sharing their code for performing exact inference on grids.

[8] These numbers can be calculated by applying the Matrix-Tree theorem [168] to the square grids.



(a) Attractive; $N = 9$

(b) Mixed; $N = 9$

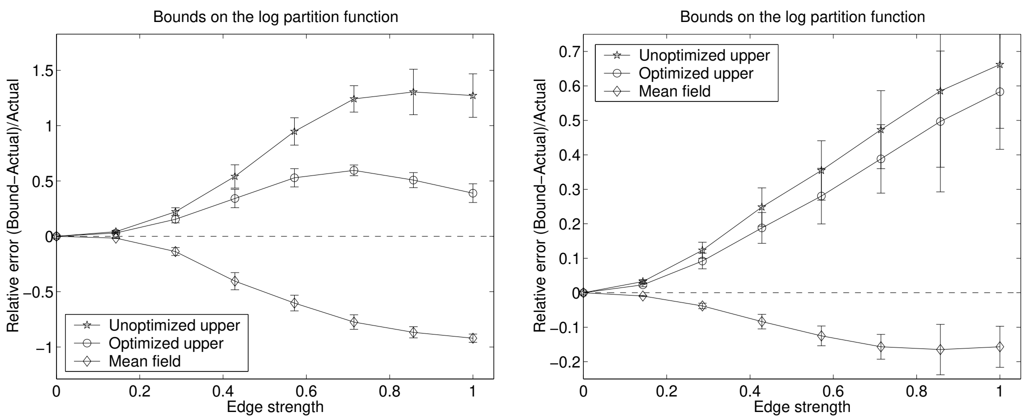
(c) Attractive; $N = 36$

(d) Mixed; $N = 36$

(c) Attractive; $N = 81$

(d) Mixed; $N = 81$

Figure 7.7. Upper and lower bounds on $\Phi(\theta^*)$ for a randomly chosen distribution $p(\mathbf{x}; \theta^*)$ on grids of various sizes: $N = 9$ nodes (first row), $N = 36$ (middle row) or $N = 81$ (bottom row). Panels on the left (respectively right) correspond to the attractive (respectively mixed) condition. Each panel shows the mean relative error $[\text{Bound} - \Phi(\theta^*)]/\Phi(\theta^*)$ versus a normalized measure of edge strength; error bars correspond to plus/minus one standard deviation.



(a) (b)

Figure 7.8. Upper and lower bounds on $\Phi(\theta^*)$ for a randomly chosen distribution $p(\mathbf{x}; \theta^*)$ on the complete graph K_9. (a) Attractive condition. (b) Mixed condition.

Fully connected graph

We also performed simulations on the fully connected graph K_9 on $N = 9$ nodes, with edge strengths d ranging from 0 to $4/3$. The results are plotted in Figure 7.8. Both types of bounds (upper convex and mean field) are much poorer for this densely connected graph (as compared to the grids). Moreover, in contrast to the grids, there is a striking disparity between the attractive and mixed conditions. In the attractive condition, none of the bounds are good; however, in a relative sense, the optimized upper bounds of Theorem 7.3.1(b) are better than the mean field lower bound. We also note that optimizing the edge appearance probabilities leads to significant improvements; indeed, the unoptimized upper bound is worse than the mean field lower bound. In the mixed condition, the mean field lower bound is of mediocre quality, whereas the upper bounds are very poor. Thus, the quality of the bounds in Theorem 7.3.1 appears to degrade for the case of mixed potentials on densely connected graphs. Of course, at least in the limit of large problem sizes, mean field would behave very well for densely connected graphs with mixed potentials, since the aggregation of multiple effects converges to a mean effect [see 93].

■ 7.5 Discussion

In this chapter, we have developed and analyzed a new class of upper bounds for log partition functions in graphical models. These bounds are based on convex combinations of distributions defined on spanning trees of the graph with cycles. We proved that there is a unique distribution over spanning trees and an associated set of exponential parameters that yield the tightest possible upper bound. Despite the prohibitively large number of spanning trees in a general graph, we developed a technique for optimizing the bound efficiently — though implicitly — over all spanning trees.

It would be interesting to compare the quality of the upper bounds of this chapter to the upper bound of Jaakkola and Jordan [95]. In particular, for the Ising model (i.e., a binary process on a graph with pairwise potentials), they developed a recursive procedure, in which the contribution of a given node is bounded above (quadratically), and then eliminated from the graph. As a side effect, all neighbors of the eliminated node are joined together by a modified set of potentials. The procedure then continues by selecting a node from this modified graph. For relatively weak choices of potentials, these bounds are much superior to the (linear) mean field lower bound [95]. However, they are unlikely to behave as well for very strong potentials.

The bounds that we developed in this chapter have a number of potential uses. Of particular relevance to this thesis are their application to computing bounds on the error of the TRP/BP algorithm (Chapter 5), as well as error bounds for more structured approximating algorithms (Chapter 6). More generally, the techniques in this chapter can be used to compute bounds on arbitrary expectations $\mathbb{E}_{\theta^*}[f]$, as described in Chapter 3.

Another potential application of our techniques is in large deviations analysis [e.g., 55,158]. It is well-known that the log partition function plays the role of a rate function: that is, it specifies the exponential rate at which the probability of certain events — namely, large deviations — decays as the number of samples is increased. In cases for which exact computation of these error exponents is infeasible, it would be of interest to obtain bounds. See Chapter 8 for further discussion of this issue.

For clarity in exposition, this chapter focused on the special case of binary-valued nodes, and graphs with pairwise cliques. However, the line of analysis outlined here is broadly applicable, in that it extends naturally to more general state spaces (e.g., $\mathcal{X} = \mathbb{R}$, or the discrete space $\mathcal{X} = \{0, 1, \ldots, m-1\}$) and larger cliques. Analogs of the results given in this chapter hold for these cases. It is also possible to consider more complex approximating structures — more specifically, hypertrees as opposed to spanning trees [17,25]. One caveat is relevant here: unlike the case of spanning trees, the optimization of a distribution $\vec{\boldsymbol{\mu}}$ over hypertrees would not be straightforward. Although solving the maximum weight spanning tree problem required as part of Algorithm 7.4.2 is easy, its analog for hypertrees of width $k \geq 2$ is NP-hard [112, 162].