Tightness of LP Relaxations for Almost Balanced Models

Adrian Weller
University of Cambridge

Mark Rowland
University of Cambridge

David Sontag
New York University

Abstract

Linear programming (LP) relaxations are widely used to attempt to identify a most likely configuration of a discrete graphical model. In some cases, the LP relaxation attains an optimum vertex at an integral location and thus guarantees an exact solution to the original optimization problem. When this occurs, we say that the LP relaxation is tight. Here we consider binary pairwise models and derive sufficient conditions for guaranteed tightness of (i) the standard LP relaxation on the local polytope LP+LOC, and (ii) the LP relaxation on the triplet-consistent polytope LP+TRI (the next level in the Sherali-Adams hierarchy). We provide simple new proofs of earlier results and derive significant novel results including that LP+TRI is tight for any model where each block is balanced or almost balanced, and a decomposition theorem that may be used to break apart complex models into smaller pieces. An almost balanced (sub-)model is one that contains no frustrated cycles except through one privileged variable.

1 INTRODUCTION

Undirected graphical models, also called Markov random fields (MRFs), are a compact and powerful way to model dependencies among variables, and have become a central tool in machine learning. A fundamental problem is to identify a configuration of all variables that has highest probability, termed *maximum a posteriori* (MAP) inference. For discrete graphical models, this is a classical combinatorial optimization problem. A popular approach is to express the problem as an integer program, then to relax this to a linear program (LP). If the LP is solved over the convex hull of marginals corresponding to all global settings, termed the *marginal polytope*, then this would solve the original problem (since an LP attains an optimum at a vertex). However, the marginal polytope has exponentially many facets (Deza and Laurent, 2009), hence this space is typically relaxed to the *local polytope* (LOC), which enforces only pairwise consistency using a linear number of constraints, which we term the LP+LOC approach. If this identifies an optimum at an integer location, then this must be an optimum of the original problem and we say that LP+LOC is *tight*.

Sontag et al. (2008) demonstrated that using higher-order cluster constraints to tighten LOC to a more constrained yet still tractable polytope, enables many real world examples to be exactly solved efficiently. Using triplets, i.e. clusters of size 3, which leads to the triplet-consistent polytope (TRI), is already very helpful. However, aside from purely topological conditions bounding treewidth, to date there has been little theoretical understanding of when these cluster methods will be effective. In this paper, we focus on binary pairwise models, and provide an important contribution by proving that LP+TRI is guaranteed to be tight for the significant class of models that satisfy the hybrid condition (combining restrictions on topology and potentials) that each *block* is *almost balanced* (see §2 for definitions).

We provide background and preliminaries in §2, then in §3, begin by analyzing LP+LOC. By applying a simple yet powerful primal perturbation argument, we first provide new, short proofs of existing results, then go on to derive novel results on how the optimum varies if one particular marginal is held to various values. These may have applications in other domains, e.g. they may be incorporated into the method of Weller and Jebara (2014) to yield more efficient approximation of the Bethe partition function. Next in §4, we consider the triplet-consistent polytope TRI. A significant result is that LP+TRI is tight for any model which is almost balanced. In §5, we provide a general decomposition result which may be of independent interest. By combining this with the result in §4, we are able to extend and demonstrate that LP+TRI is tight for any model in which every block is almost balanced. This result is of additional interest since Weller (2015b) recently demonstrated that a different 'MWSS' approach can be used for efficient MAP inference for any valid potentials iff each block of a model is almost balanced. We are able to show that LP+TRI dom-

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

inates that approach, in the sense that it is guaranteed to be able to solve a superset of models for any potentials.

This subject area has received considerable attention from several communities. We discuss related work throughout the text; for a more comprehensive treatment, see (Wainwright and Jordan, 2008, §8) or (Deza and Laurent, 2009). Note that for binary models (with potentials of any arity), Sontag (2010) has shown that TRI is equivalent to the *cycle polytope*, which enforces consistency over all cycles.

2 PRELIMINARIES

For binary pairwise models, MAP inference may be framed in a minimal representation (Wainwright and Jordan, 2008) as the discrete optimization problem to identify

$$x^* \in \arg\max_{x\in\{0,1\}^n} \left(\sum_{i\in\mathcal{V}} \theta_i x_i + \sum_{(i,j)\in\mathcal{E}} W_{ij} x_i x_j \right), \quad (1)$$

where the model's topology is given by the graph $(\mathcal{V}, \mathcal{E})$, with $n = |\mathcal{V}|$ variables and $m = |\mathcal{E}| \leq \binom{n}{2}$ edge relationships between the variables. The n θ_i singleton parameters and m W_{ij} edge weights define the potentials, and may take any real value. Sometimes we may assume all $\binom{n}{2}$ edges (i, j) are present, allowing for some to have zero weight $W_{ij} = 0$, where the context will make this clear. Whenever discussing the topology of a model, we mean the graph $(\mathcal{V}, \mathcal{E})$.

If $W_{ij} \geq 0$, the edge (i, j) tends to pull X_i and X_j toward the same value and is called *attractive*. If $W_{ij} < 0$, the edge is *repulsive*. We may concatenate the potential parameters together into a vector $w \in \mathbb{R}^d$, where $d = n + m$. Similarly, we may define $y_{ij} = x_i x_j$, then concatenate the n x_i and m y_{ij} terms into a vector $z = (x_1, \dots, x_n, \dots, x_i x_j, \dots) \in \{0, 1\}^d$. This yields the following equivalent integer programming formulation, to identify

$$z^* \in \arg\max_{z:x\in\{0,1\}^n} w \cdot z \quad (2)$$

The convex hull of the 2^n possible integer solutions is called the *marginal polytope* $\mathbb{M}$. Regarding the convex coefficients as a probability distribution p over all possible states, $\mathbb{M}$ may be considered the space of all singleton and pairwise mean marginals that are consistent with some global distribution p over the 2^n states, that is

$$\begin{aligned}\mathbb{M} = \{ & z = (z_1, \dots, z_n, z_{12}, z_{13}, \dots, z_{(n-1)n}) \\ & \text{s.t. } \exists p : z_i = \mathbb{E}_p(X_i)\ \forall i, z_{ij} = \mathbb{E}_p(X_i X_j)\ \forall (i,j)\}.\end{aligned}$$

A standard approach is to relax (2) to a linear program (LP). However, this remains intractable over $\mathbb{M}$ (we use tractable to mean solvable in polynomial time) since the number of facets (and hence the number of LP constraints) grows extremely rapidly with n (Deza and Laurent, 2009). Hence, a simpler, relaxed constraint set is typically employed, yielding an upper bound on the original optimum. This set is often chosen as the *local polytope* (LOC or $\mathbb{L}$), defined as the polytope over $q = (q_1, \dots, q_n, \dots, q_{ij}, \dots) \in \mathbb{R}^d$ subject to the following linear constraints (see Figure 2):

$$\begin{aligned} 0 \leq q_i \leq 1 \quad & \forall i \in \mathcal{V}, \quad (3) \\ \max(0, q_i + q_j - 1) \leq q_{ij} \leq \min(q_i, q_j) \quad & \forall (i,j) \in \mathcal{E}.\end{aligned}$$

It is easily checked that these are exactly the requirements to ensure that q gives rise to valid singleton and pairwise marginals (nonnegative values summing to 1) that are locally consistent (marginalizing a pairwise marginal yields the appropriate singleton marginal), given by

$$\begin{aligned} &\text{singletons} \quad q(X_i = 0) = 1 - q_i, \quad q(X_i = 1) = q_i, \\ &\text{edges} \quad \begin{pmatrix} q(X_i=0, X_j=0) & q(X_i=0, X_j=1) \\ q(X_i=1, X_j=0) & q(X_i=1, X_j=1) \end{pmatrix} \\ &\qquad = \begin{pmatrix} 1 + q_{ij} - q_i - q_j & q_j - q_{ij} \\ q_i - q_{ij} & q_{ij} \end{pmatrix}. \quad (4)\end{aligned}$$

Hence, $\mathbb{M} \subseteq \mathbb{L}$ though $q \in \mathbb{L}$ may not be consistent with any global probability distribution, thus q is termed a *pseudomarginal* vector. $\mathbb{L}$ is defined by a polynomial number of constraints, thus it is tractable (Schrijver, 1998) to solve the relaxation LP+LOC given by

$$q^* \in \arg\max_{q\in\mathbb{L}} \left(\sum_{i=1}^{n} \theta_i q_i + \sum_{(i,j)\in\mathcal{E}} W_{ij} q_{ij} \right) = \arg\max_{q\in\mathbb{L}} w \cdot q \quad (5)$$

If an optimum vertex is achieved at an integer solution, then this must be an optimum of the original discrete problem (2), in which case we say that the relaxation is *tight*.

Starting with LOC, an intuitively appealing series of successively more restrictive relaxations was established by Sherali and Adams (1990). At order r, the $\mathcal{L}_r$ polytope enforces consistency over all clusters of variables of size r. Hence, $\mathcal{L}_2$ is the local polytope LOC. Next, $\mathcal{L}_3$ enforces consistency over all triplets of variables, which we denote by TRI, and so on. Since $\mathcal{L}_n = \mathbb{M}$, it is clear that LP+$\mathcal{L}_n$ is tight. Building on the junction tree theorem (Cowell et al., 1999), Wainwright and Jordan (2004) demonstrated that a topological sufficient condition for LP+$\mathcal{L}_r$ to be tight, is if a model has treewidth[1] $\leq r - 1$. Note that this holds for any potentials, whereas looser requirements may suffice given certain restrictions on the potential functions, as we shall show in later Sections.

2.1 Flipping, Balanced and Almost Balanced Models, Block Decomposition and the MWSS Method

If a model has only attractive edges, it is an *attractive model*, whereas a *general* model may have any edge types.

[1]The treewidth of a graph is one less than the smallest possible size of a largest clique in a triangulation of the graph. As examples: a tree has treewidth 1; an $n \times n$ grid has treewidth n.

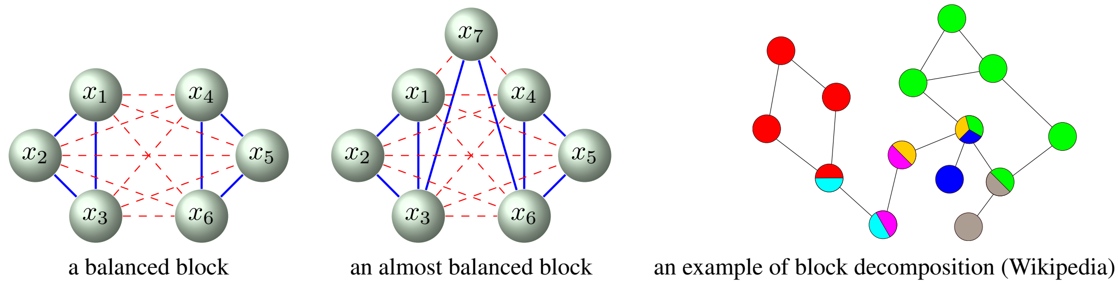


Figure 1: The left two figures show example model blocks (maximal 2-connected components), with solid blue (dashed red) edges indicating attractive (repulsive) edges. In the balanced block, flipping either $\{x_1, x_2, x_3\}$ or $\{x_4, x_5, x_6\}$ partition renders the block attractive. The almost balanced block adds x_7 creating frustrated cycles. On the right, each color indicates a different block of a graph; multi-colored vertices are cut vertices (if these are removed, the graph becomes disconnected), hence belong to multiple blocks.

If a model is not attractive, in some cases it is still possible to render it attractive by *flipping* (sometimes called *switching*) a subset of variables, as follows. Partition the variable indices into two subsets, $A \subseteq [n] = \{1, \dots, n\}$ and $B = [n] \setminus A$. Consider the model with new variables $Y_1, \dots, Y_n$ where $Y_i = X_i \; \forall i \in A$, and $Y_i = 1 - X_i \; \forall i \in B$. As described in (Weller, 2015a, §2.4), new potential parameters $\{\theta'_i, W'_{ij}\}$ may be determined such that the scores over states are unchanged up to a constant (and hence the distribution is unchanged). In particular, edge weights $W'_{ij} = \pm W_{ij}$, where the sign changes iff exactly one of X_i and X_j is flipped. Harary (1953) showed that $\exists$ a subset $A \subseteq [n]$ such that flipping those variables renders the model attractive iff there is no cycle with an odd number of repulsive edges. Such a cycle is called a *frustrated cycle*. Checking for a frustrated cycle may be performed efficiently, and models without frustrated cycles are called *balanced*. Thus, many results that apply to attractive models may be extended to the wider class of balanced models.

An interesting approach to MAP inference was introduced by Jebara (2009), via a reduction to the *maximum weight stable set* (MWSS) problem on a derived weighted graph (see Diestel, 2010 for all terms from graph theory). Weller (2015b) considered binary pairwise models and proved that this method is guaranteed to yield an efficient optimum configuration for any valid potentials (because the derived graph is perfect) iff each *block* of the model is *almost balanced*. A block is a maximal 2-connected subgraph, thus a graph may be repeatedly broken apart at cut vertices to yield its unique block decomposition. A (sub-)model is *almost balanced* if it may be rendered balanced by deleting one variable (hence, in particular, a balanced (sub-)model is almost balanced). Checking to see if all blocks of a model are almost balanced may be performed efficiently (Weller, 2015b). Our new results show that LP+TRI dominates this MWSS approach, see §5.1. Figure 1 shows examples of a balanced block, an almost balanced block, and block decomposition.

3 RESULTS FOR LOCAL POLYTOPE

The following perturbation argument will be central in our analysis. Recall that an optimum of an LP is always attained at a vertex (extreme point) of the polytope (Schrijver, 1998). Suppose we wish to show that an optimum vertex may be found with certain properties. Toward contradiction, suppose that all optimum vertices do not have the properties and let q^* be any such vertex. We shall explicitly construct q^+ and q^- which lie in the polytope under consideration, such that $q^* = \frac{1}{2}(q^+ + q^-)$, hence q^* is not a vertex, and the result follows. This constructively demonstrates a direction in which the score is nondecreasing (a similar approach was used by Taskar et al., 2004). To construct appropriate q^+ and q^-, we shall typically perturb the singleton marginals by symmetric small distances from q^*, and the difficulty will be to ensure that the edge marginal terms can also be perturbed symmetrically.

For the local polytope LOC, given singleton terms $\{q_i\}$, all pairwise terms $\{q_{ij}\}$ may be optimized independently. From the constraints (3), optimum edge terms are

$$q^*_{ij}(q_i, q_j) = \begin{cases} \min(q_i, q_j) & \text{if } W_{ij} > 0 \\ \max(0, q_i + q_j - 1) & \text{if } W_{ij} < 0 \end{cases}. \quad (6)$$

Figure 2 indicates the feasible range of q_{ij} values for ways that q_i and q_j might vary together.

Problem cases. If optimum edge terms q_{ij} are always recomputed, then if q_i is perturbed up then down by ϵ, while q_j is moved by ϵ in the same way, in contrary direction or not at all, then the edge term q_{ij} will always move symmetrically, except in the following two problem cases: (i) $q_i = q_j$ with an attractive edge $W_{ij} > 0$, in which case we call i and j *locked*, and q_i and q_j must move together; or (ii) $q_i = 1 - q_j$ with a repulsive edge $W_{ij} < 0$, in which case we call i and j *anti-locked*, and q_i and q_j must move in opposite directions. Observe that case (ii) may be seen as equivalent to case (i) after flipping either variable, see §2.1, or §8 in the Appendix for more comments on symmetry.

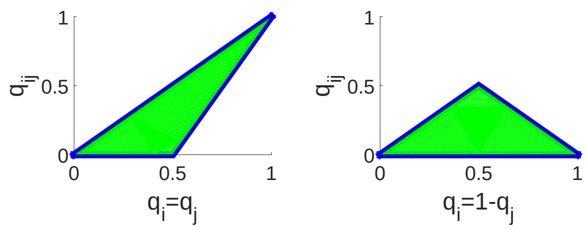


Figure 2: Feasible ranges in the local polytope for edge marginal q_{ij} given singleton marginals q_i and q_j moving together (left) or in opposite directions (right), illustrating the vertex at $1/2$. If the edge is attractive, then the optimal q_{ij} will be in the upper envelope (leading to a possible vertex if q_i and q_j move in opposite directions); if repulsive, then the optimal q_{ij} will lie in the lower envelope (possible vertex if q_i and q_j move together).

The results in §3.1 were shown previously by other methods (Padberg, 1989), but we provide new, intuitive, short proofs. We believe results in §3.2 and thereafter are new.

3.1 New Short Proofs of Earlier Results for LOC

Theorem 1. *For an attractive model, LP+LOC is tight.*

Proof. Toward contradiction, suppose all optimum vertices have some non-integer coordinate. Let $q^* \in \mathbb{L}$ be such an optimum vertex. Let $\mathcal{I} = \{i : q_i^* \notin \{0,1\}\}$. From (6), $\forall (i,j) \in \mathcal{E}, q_{ij}^* = \min(q_i^*, q_j^*)$. Define $q^+ = (q_1^+, \ldots, q_n^+, \ldots, q_{ij}^+, \ldots)$ as follows:

$$q_i^+ = \begin{cases} q_i^* + \epsilon & i \in \mathcal{I} \\ q_i^* & i \notin \mathcal{I} \end{cases} \quad q_{ij}^+ = \min(q_i^+, q_j^+),\ (i,j) \in \mathcal{E}.$$

Note these are optimum edge terms. Similarly, define q^-,

$$q_i^- = \begin{cases} q_i^* - \epsilon & i \in \mathcal{I} \\ q_i^* & i \notin \mathcal{I} \end{cases} \quad q_{ij}^- = \min(q_i^-, q_j^-),\ (i,j) \in \mathcal{E}.$$

Here $\epsilon > 0$ is sufficiently small such that both $q^+, q^- \in \mathbb{L}$. More precisely, let $a = \min_{i \in \mathcal{I}} q_i^*$, $b = \min_{i \in \mathcal{I}}(1 - q_i^*)$ then we may take any $\epsilon < \min(a, b)$.[2] It is easily checked that $q^* = \frac{1}{2}(q^+ + q^-)$, hence q^* is not a vertex. □

The particular choice of q^+ and q^- in the proof above works by ensuring that all edge terms $\{q_{ij}\}$ move symmetrically, i.e. each edge term either does not move for both q^+ and q^-, or moves up for one and down for the other.

Theorem 2. *For a balanced model, LP+LOC is tight.*

Proof A. Since the model is balanced, a subset of variables may be identified such that flipping them renders the model attractive (Harary, 1953, see §2.1); then apply Theorem 1 (if a model is tight then so too is any flipping of it). □

[2] More generally, going forward, take $\epsilon > 0$ sufficiently small s.t. any polytope constraint which was not tight initially, remains so after perturbing q^* by $\pm\epsilon$.

Proof B. We provide an alternative derivation which essentially incorporates the flipping into the proof. Recall the two possible problem cases described above in §3.

Split $\mathcal{I} = \{i : q_i^* \notin \{0,1\}\}$ into two groups, A and B, such that all intra-group edges are attractive and all inter-group edges are repulsive (flipping either group renders the model attractive, see §2.1). Observe that $q^* = \frac{1}{2}(q^+ + q^-)$ if we define

$$q_i^+ = \begin{cases} q_i^* + \epsilon & i \in A \\ q_i^* - \epsilon & i \in B \\ q_i^* & i \notin \mathcal{I} \end{cases}, \quad q_i^- = \begin{cases} q_i^* - \epsilon & i \in A \\ q_i^* + \epsilon & i \in B \\ q_i^* & i \notin \mathcal{I} \end{cases}, \quad (7)$$

with both using optimum edge terms $\{q_{ij}^+, q_{ij}^-\}$, see (6). □

Theorem 3. *For a general model (any potentials, attractive or not), LP+LOC is half-integral.*

Proof. Let $A = \{i : 0 < q_i^* < \frac{1}{2}\}$, let $B = \{i : \frac{1}{2} < q_i^* < 1\}$. Set q^+ and q^- as in (7), with $\mathcal{I} = A \cup B$. □

Since Theorem 3 considers optimizing an arbitrary linear function over the polytope LOC, an immediate corollary is that all vertices of LOC are half-integral.

3.2 New Results for LOC, Fixing One Variable and Optimizing Over the Others

Results in this Section may be of independent interest, and also serve as a warm-up for our approach for TRI in §4.

Theorem 4. *For an attractive model, if we fix one variable's marginal $q_i = x \in [0,1]$, and optimize over all others $\{q_j : j \neq i\}$, then an optimum vertex is achieved with $q_j \in \{0, x, 1\}\ \forall j$.*

Proof. Toward contradiction, if all optima have some $q_j^* \notin \{0, x, 1\}$ then construct q^+ and q^- by moving these variables up/down together by ϵ, i.e. the same construction for q^+ and q^- as in the proof of Theorem 1, setting positive $\epsilon <$ min distance to any member of $\{0, x, 1\}$. □

We define the following constrained optimum function for any polytope $\mathbb{P}$ which is a relaxation of $\mathbb{M}$,

$$F_{\mathbb{P}}^i(x) = \max_{q \in \mathbb{P}: q_i = x} w \cdot q, \qquad x \in [0,1]. \qquad (8)$$

First we provide the following simple Lemma.

Lemma 5. *For any $\mathbb{P}$, $F_{\mathbb{P}}^i(x)$ is a concave function for $x \in [0,1]$.*

Proof. Given any $x_0, x_1 \in [0,1]$, let $q^0, q^1 \in \mathbb{R}^d$ be arg max locations for $F_{\mathbb{P}}^i(x_0)$ and $F_{\mathbb{P}}^i(x_1)$ respectively. For any $\lambda \in [0,1]$, let $\hat{x} = \lambda x_1 + (1-\lambda)x_0$ and $\hat{q} = \lambda q^1 + (1-\lambda)q^0$. Now $F_{\mathbb{P}}^i(\hat{x}) = \max_{q \in \mathbb{P}: q_i = \hat{x}} w \cdot q \geq w \cdot \hat{q} = \lambda F_{\mathbb{P}}^i(x_1) + (1-\lambda)F_{\mathbb{P}}^i(x_0)$. □

Using Theorem 4 and Lemma 5, we shall show how $F_{\mathbb{L}}^i(x)$, the constrained optimum on LOC, varies with x.

Theorem 6. *For a balanced model, $F_{\mathbb{L}}^i(x)$ is linear.*

Proof. We assume an attractive model. The result will then extend to a balanced model by first flipping an appropriate subset of variables, see §2.1. We shall show here that $F_{\mathbb{L}}^i(x)$ is convex, then linearity follows from Lemma 5.

For any $y \in [0,1]$, consider an $\arg\max$ of $F_{\mathbb{L}}^i(y)$ as given by Theorem 4. Partition the variables into 3 exhaustive sets: $A_y = \{j : q_j = 0\}, B_y = \{j : q_j = y\}$ and $C_y = \{j : q_j = 1\}$. Define the function $f_y : [0,1] \to \mathbb{R}$ given by $f_y(x) = f(q(x;y))$ where $q(x;y)$ is defined by:

$$q_j(x;y) = \begin{cases} 0 & j \in A_y \\ x & j \in B_y \\ 1 & j \in C_y \end{cases},$$

using optimum terms $q_{jk}(x;y) = \min[q_j(x;y), q_k(x;y)]$ for all edges. Observe that $f_y(x)$ is the linear function achieved by holding fixed the partition of variables A_y, B_y, C_y that was determined for the $\arg\max$ of the constrained optimum at $q_i = y$. Now $F_{\mathbb{L}}^i(x) = \sup_{y\in[0,1]} f_y(x)$, hence is convex. □

Note that since $F_{\mathbb{L}}^i(x)$ is linear, it must be that each of the linear $f_y(x)$ functions from the proof are equal, so as an immediate corollary, we may take the A, B, C sets to be constant with the same variables in them, independent of y.

For a general model, we can show an analog of Theorem 4.

Theorem 7. *For a general model, if one variable's marginal $q_i = x \in [0,1]$ is fixed and we optimize over all others $\{q_j : j \neq i\}$, then an optimum is achieved with $q_j \in \{0, x, \frac{1}{2}, 1-x, 1\}\ \forall j$.*

Proof. Fix $q_i = x$ and optimize over all other variables. Let $\mathcal{I} = \{j : q_j \notin \{0, x, \frac{1}{2}, 1-x, 1\}\}$. If $\exists j \in \mathcal{I}$, take A to be all variables in $\mathcal{I}$ equal to q_j and B to be all variables in $\mathcal{I}$ equal to $1 - q_j$. Perturb up A and down B, then vice versa, i.e. set q^+ and q^- as in (7). □

Observe that (because of the fixed $\frac{1}{2}$ in its statement) Theorem 7 does *not* allow an argument as in the proof of Theorem 6 to yield the (false) conclusion that $F_{\mathbb{L}}^i(x)$ is linear for a general model.

4 RESULTS FOR TRIPLET POLYTOPE

The triplet-consistent polytope TRI is defined by the constraints of the local polytope $\mathbb{L}$ (3), together with the following additional *triangle inequalities* (4 per triplet):

$$\forall \text{ distinct } i,j,k, \qquad q_i + q_{jk} \geq q_{ij} + q_{ik}, \qquad (9)$$

$$q_{ij} + q_{ik} + q_{jk} \geq q_i + q_j + q_k - 1. \qquad (10)$$

These enforce consistency over any triplet of variables, as may be derived by the lift-and-project method. Hence, $\mathbb{M} \subseteq \text{TRI} \subseteq \mathbb{L}$. For the purpose of these inequalities, if an edge $(i,j) \notin \mathcal{E}$ then assume it is present with $W_{ij} = 0$. See Appendix §7 for a derivation of the inequalities, and §8 for a discussion of their symmetry.

In this Section, we shall show that, somewhat remarkably, an almost balanced model on TRI behaves in many ways just like a balanced model on LOC. A key result is the following analog of Theorem 1.

Theorem 8. *For an almost balanced model, LP+TRI (the LP relaxation over TRI) is tight.*

To prove Theorem 8, we shall show the following analog of Theorem 6, where s is the special variable in an almost balanced model such that when removed, the remainder is balanced (see §2.1).

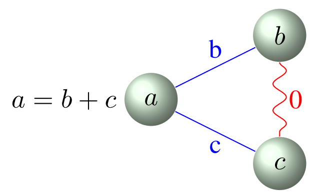
Theorem 9. *For an almost balanced model with special variable s, $F_{TRI}^s(x)$ is a linear function.*

If we can prove Theorem 9, then an optimum occurs at $s = 0$ or $s = 1$. Conditioning on this value of s yields a balanced model; then Theorem 8 follows by Theorem 2 (since TRI $\subseteq \mathbb{L}$). Full details and proofs are provided in the Appendix. Here we provide a sketch.

Just as the proof of Theorem 6 holds one singleton marginal fixed and relies on Theorem 4 to provide all other optimum marginals, here we shall hold fixed q_s, the singleton marginal of the special variable s, and develop Theorem 11 to provide all other optimum marginals.

For our perturbation method, on LOC, once we condition on a set of singleton marginals, the edge marginals are independent and easily computed. On TRI, in contrast, edges interact. We call any edge where the optimum edge marginal takes its maximum possible value on LOC (behaving 'like an attractive edge', though the edge may be repulsive), a *strong up* edge. Similarly, we call an edge where the optimum marginal takes its minimum possible value on LOC (behaving 'like a repulsive edge'), a *strong down* edge. Generalizing from §3, 2 variables are *locked up* (*locked down*) if they have $q_i = q_j$ ($q_i = 1 - q_j$) and are joined by a strong up (strong down) edge; in either case (up or down) the edge is *locking*. A cycle of strong (up or down) edges is *strong frustrated* if it contains an odd number of strong down edges. If not strong, an edge is *weak*.

Problem triangles. In addition to the earlier problem cases for LOC in §3 involving 2 variables, from which we observe that if we have locked up (locked down) variables, they must move together (opposite), when we consider TRI, we must also respect all TRI constraints (9), Lower case letters such as a may be overloaded for variable names and their singleton marginals.



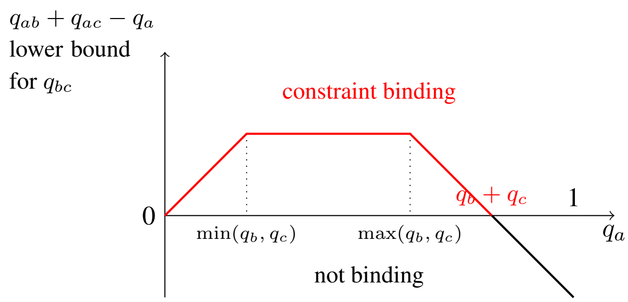


Figure 3: Above: an illustration of strong 'problem triangle' type (i). Blue edges are strong up, the red wavy edge is strong down. Below: a plot showing the relevant triangle constraint (others are always satisfied) $q_a + q_{bc} \geq q_{ab} + q_{ac}$ as q_a is varied, holding fixed q_b and q_c while recomputing LOC-optimum edge marginals for q_{ab} and q_{ac}. The TRI constraint is binding where the plot is red, and not where it is black. Here we consider $q_b + q_c < 1$, hence on LOC, $q_{bc} = 0$, and $q_{ab} = \min(q_a, q_b), q_{ac} = \min(q_a, q_c)$. $q_a = q_b + q_c$ is the new problem case (e.g. if just q_a is perturbed, the constraint becomes binding just on one side leading to an asymmetric response). There may also be problems at $q_a \in \{\min(q_b, q_c), \max(q_b, q_c)\}$ but these are already covered since they would form locking edges from a to b or c.

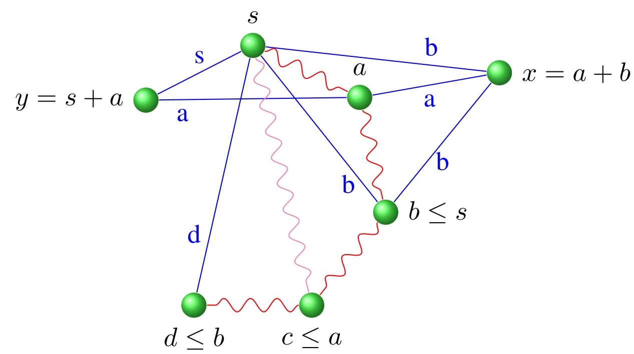


Figure 4: Illustration of how marginals must behave for an almost attractive model in TRI to obtain a path of strong down edges $s - a - b - c - d$, shown in wavy red, for the case $\boldsymbol{s + a < 1}$ (hence all strong down edges shown will have edge marginal 0). Singleton marginals are in black. The lighter pink wavy edge $s - c$ is implied also to have 0 edge marginal. The other edges (straight blue) are forced to be strong up edges, and show their edge marginal in blue. Note that as we move along any strong down path from s, the edge marginals to s alternate between 0 (for an odd distance by wavy edges) and the respective singleton marginal > 0 (for an even distance by wavy edges); in particular it is not possible to have a cycle with 3 strong down edges. Two problem triangles of type (i) are shown: x, a, b and y, s, a. See Appendix §11.

(10). We call any triplet with binding TRI constraints a 'problem triangle'. We must ensure that our perturbation maintains all binding constraints for all problem triangles, otherwise one direction of the perturbation will lead to a constraint violation, i.e. moving outside TRI.

Full details are in the Appendix. For illustration, we consider here the case of a problem triangle where all edges are strong. There are four subcases to consider, each is strong frustrated: (i) One strong down edge $b - c$ with $b + c < 1$ and $a = b + c$, see Figure 3; (ii) One strong down edge $b - c$ with $b + c > 1$ and $a = b + c - 1$; (iii) Three strong down edges with $a + b + c = 1$ (this implies that each pair sums to less than 1); (iv) Three strong down edges with $a + b + c = 2$ (this implies that each pair sums to more than 1). In each of the four cases, only certain combined perturbations of variables will result in symmetric edge marginal perturbations. In all cases, it works if exactly 2 (of the 3) variables are perturbed, to move in opposite directions, with the 2 variables being on either end of a strong down edge.

Locked (up or down) variables must move appropriately. See Appendix §9 for details of the following: Variables connected by paths of locking edges form in TRI a *locking component*, in which all variables are adjacent by locking edges and there is no strong frustrated cycle. If we know the edge marginal from any member of a locking component to a variable outside it, we can uniquely determine all edge marginals (which move together/opposite) to that outside variable from all members of the locking component. Hence, we may 'contract' any locking component to a single variable for analysis purposes on a reduced model where we may assume we have no locking edges. Once we have analyzed the reduced model, it is straightforward to 'expand' the analysis back up to the original model.

In Appendix §10, we show: If any variable has singleton marginal 0 or 1, then this uniquely determines incident edge marginals, which will always satisfy the TRI constraints and move symmetrically. Hence we may assume no variables with 0 or 1 singleton marginal.

To further simplify analysis, without loss of generality, by flipping an appropriate set of variables in $\mathcal{V} \setminus \{s\}$ (see §2.1), we may assume that we have an 'almost attractive' model, with all edges attractive, except for some edges incident to s; results then extend to almost balanced models.

With these observations, we provide a key result on the structure of strong down and weak edges (see Theorem 25 in the Appendix for the full version).

Lemma 10. *In an almost attractive model with special variable s (i.e. the model on $\mathcal{V} \setminus \{s\}$ is attractive), if all edge marginals have been optimized in TRI given a set of singleton marginals, then any strong down or weak edge $x - y$ with $s \notin \{x, y\}$ must form a problem triangle with s.*

Using Lemma 10, we show the following analog of Theorem 4, which will enable us to prove Theorem 9.

Theorem 11. *In an almost balanced model with special*

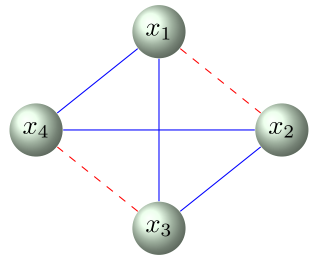


Figure 5: A minimal example of a block that is not *almost balanced*, also a minimal example of a block that has treewidth > 2, hence models with this topology might not be tight for LP+TRI. Solid blue (dashed red) edges are attractive (repulsive). All triangles are frustrated with an odd number of repulsive edges.

variable s, if we fix $q_s = x \in [0,1]$ and optimize in TRI over all other marginals, then an optimum is achieved with: $q_j \in \{0, x, 1-x, 1\}\ \forall j$; all edges (other than to variables which have 0 or 1 singleton marginal) are locked up or locked down, with no strong frustrated cycles.

Proof. We shall show that any variables $\notin \{0,1\}$ that are not locked up or locked down to s may be perturbed with symmetric edge marginals, demonstrating that we are not at an optimum vertex. As above, we may assume an almost attractive model with no locking components and no variables $\in \{0,1\}$. Using the structural result of Lemma 10, we may construct a symmetric perturbation, as required, see Appendix 12 for details. □

Using Theorem 11, Theorem 9 may be proved in the same way as was shown for Theorem 6 (recall Lemma 5 applies to TRI; also we obtain a similar corollary that the partition of variables holding each value may be taken to be constant). This now also proves Theorem 8.

4.1 Remarks

A minimal example of a block that is not almost balanced is shown in Figure 5. If there are no singleton potentials, then by the analysis of the cut polytope by Barahona (1983), TRI= $\mathbb{M}$ and hence LP+TRI is tight. However, potentials exist s.t. LP+TRI is not tight for models with this topology.

Non-integral Vertices of TRI. Padberg (1989) proved that LOC= $\mathcal{L}_2$ is $\frac{1}{2}$-integral (we showed a new, short proof, see Theorem 3), and also showed that TRI= $\mathcal{L}_3$ has no $\frac{1}{2}$-integral vertex. Hence, the triangle inequalities are sufficient to cut off all fractional vertices of LOC. It is natural to wonder if perhaps TRI= $\mathcal{L}_3$ is $\frac{1}{3}$-integral. Laurent and Poljak (1995) considered this by analyzing the *metric polytope* (which is equivalent via the covariance mapping, Hammer, 1965; Deza, 1973; De Simone, 1989/90). Translating their results to our context, they proved that indeed TRI is $\frac{1}{3}$-integral for $n \leq 5$, but as n grows, vertices of TRI at fractions with arbitrarily large denominator are possible.

5 MODEL DECOMPOSITION RESULTS

In this section we show a general result that an LP relaxation of a component-structured graphical model is tight whenever the LP relaxations on the components are tight and consistency is enforced on the variables in common between adjacent components. Consider a graphical model with variables $\mathcal{V} = A \cup B$, and let $C = A \cap B$ be the variables in common between A and B. Specifically, let $p(\vec{x}, \vec{y}, \vec{z})$ be an exponential family distribution with sufficient statistic vector $\phi(\vec{x}, \vec{y}, \vec{z}) = [\phi_x(\vec{x}), \phi_y(\vec{y}), \phi_z(\vec{z})]$, and let $A = X \cup Y$ and $B = Y \cup Z$.

Let M be the marginal polytope corresponding to $\phi(\vec{x}, \vec{y}, \vec{z})$, i.e. the convex hull of $\phi(\vec{x}, \vec{y}, \vec{z})$ for every assignment to X, Y, Z. Similarly, let M_A and M_B be the marginal polytopes corresponding to sufficient statistic vectors $[\phi_x(\vec{x}), \phi_y(\vec{y})]$ and $[\phi_y(\vec{y}), \phi_z(\vec{z})]$, respectively. Every polytope can be equivalently defined as the intersection of linear inequalities (the polytope's maximal facets). Let $M_I = M_A \cap M_B$ be the polytope defined by combining the linear inequalities making up both M_A and M_B.

Theorem 12 (Decomposition result for graphical models)**.** *Suppose we have two polytopes M_A and M_B for models with variables A and B, where $C = A \cap B$ are the variables in common. Suppose we have LP relaxations for M_A and M_B which are known to be tight for any objective $\theta_A \in \Theta_A$ and $\theta_B \in \Theta_B$, respectively. If the sets Θ_A and Θ_B are closed under the addition of an arbitrary potential function θ_C, then $M_I = M_A \cap M_B$ (defined just above) is tight on the combined model over variables $A \cup B$, i.e. $M_I = M$.*

Proof. Clearly M_I is a polytope and $M \subseteq M_I$, i.e. M_I is a relaxation, which we shall demonstrate is tight. We do this by showing that for every weight vector $\vec{w}$, the optimal value of $\vec{w} \cdot \mu$ is the same for $\mu \in M_I$ as for $\mu \in M$. To do that, we consider the Lagrangian relaxation and demonstrate a dual witness.

For any $\vec{w} = [w_{\vec{x}}, w_{\vec{y}}, w_{\vec{z}}]$, let $\theta_{\vec{w}}(\vec{x}, \vec{y}, \vec{z}) = \vec{w} \cdot \phi(\vec{x}, \vec{y}, \vec{z}) = \theta_y(\vec{x}, \vec{y}) + \theta_z(\vec{x}, \vec{z})$, where $\theta_y(\vec{x}, \vec{y}) = [w_{\vec{x}}, w_{\vec{y}}] \cdot [\phi_x(\vec{x}), \phi_y(\vec{y})]$ and $\theta_z(\vec{x}, \vec{z}) = [0, w_{\vec{z}}] \cdot [\phi_x(\vec{x}), \phi_z(\vec{z})]$. Consider the following:

$$\begin{aligned}
&\max_{\vec{x},\vec{y},\vec{z}} \theta(\vec{x}, \vec{y}, \vec{z}) = \max_{\mu \in M} \vec{w} \cdot \mu \leq \max_{\mu \in M_I} \vec{w} \cdot \mu \\
&= \max_{\mu_1 \in M_A, \mu_2 \in M_B : \mu_1(\vec{x}) = \mu_2(\vec{x})} [w_{\vec{x}}, w_{\vec{y}}] \cdot \mu_1 + [0, w_{\vec{z}}] \cdot \mu_2 \\
&= \min_{\lambda_{\vec{x}}} \Bigg(\max_{\mu_1 \in M_A, \mu_2 \in M_B} [w_{\vec{x}}, w_{\vec{y}}] \cdot \mu_1 + [0, w_{\vec{z}}] \cdot \mu_2 \\
&\qquad\qquad + \lambda_{\vec{x}}(\mu_1(\vec{x}) - \mu_2(\vec{x})) \Bigg) \\
&= \min_{\lambda_{\vec{x}}} \left(\max_{\vec{x},\vec{y}} [\theta_y(\vec{x}, \vec{y}) + \lambda_{\vec{x}}] + \max_{\vec{x},\vec{z}} [\theta_z(\vec{x}, \vec{z}) - \lambda_{\vec{x}}] \right),
\end{aligned}$$

where in the last step we use the assumption that M_A and M_B are tight for any potential $\theta(\vec{x})$.

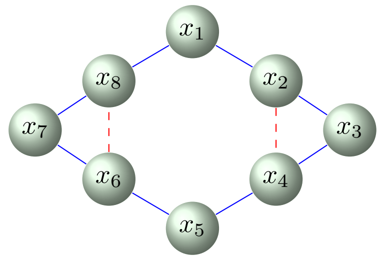


Figure 6: Illustration of a 2-connected model with treewidth 2, hence LP+TRI is tight for any potentials; but it is not almost balanced (since it contains two disjoint frustrated cycles $x_2 - x_3 - x_4$ and $x_6 - x_7 - x_8$), thus it is not always solvable by the MWSS approach. Solid blue (dashed red) edges are attractive (repulsive).

Now plug in $\lambda_{\vec{x}} = [\max_{\vec{z}} \theta_z(\vec{x}, \vec{z}) - \max_{\vec{y}} \theta_y(\vec{x}, \vec{y})] / 2$, and one can verify that the last term is equal to $\max_{\vec{x},\vec{y},\vec{z}} \theta(\vec{x}, \vec{y}, \vec{z})$, and thus the inequality must be an equality, which proves that the relaxation is tight. □

As a special case, for Sherali-Adams relaxations we have

Corollary 13. *If LP+$\mathcal{L}_r$ (clusters of up to r variables) is tight for model A, and similarly LP+$\mathcal{L}_s$ is tight for model B, in each case no matter what the single-node potentials are, and with the two models having exactly one variable in common, then LP+$\mathcal{L}_t$ is tight on the combined MRF over all the variables, where $t = \max(r, s)$.*

5.1 Application to LP+TRI, Comparison to MWSS Approach

Wainwright and Jordan (2004) showed that LP+TRI is tight for any model that has treewidth ≤ 2. Theorem 8 shows that LP+TRI is tight for any model that is almost balanced. Applying Corollary 13, we deduce that LP+TRI is tight for any model with block structure such that each block is either almost balanced or has treewidth 2 (a model with treewidth 1 is a tree hence is balanced).

An interesting approach to MAP inference was introduced by Jebara (2009) and Sanghavi et al. (2009), which reduces the problem to the graph theoretic challenge of identifying a *maximum weight stable set* (MWSS) in a derived weighted graph termed a *nand Markov random field* (NMRF). For binary pairwise models, Weller (2015b) demonstrated that this method will yield an exact solution (via a perfect graph) in polynomial time for any valid potentials iff each block of the model is almost balanced.

Our result demonstrates that the LP+TRI approach can handle all these models and more. For example, Figure 6 shows a 2-connected model that is not almost balanced (since it contains two disjoint frustrated cycles), hence for some potentials, the MWSS approach will fail on this model; yet LP+TRI is guaranteed to solve MAP inference efficiently for any potentials, since the treewidth is 2.

6 DISCUSSION

We have analyzed the tightness of LP relaxations on LOC and TRI, the first two levels of the Sherali-Adams hierarchy, for MAP inference in binary pairwise graphical models, demonstrating novel techniques and insights, and significant results. The subject is of great theoretical interest and has been studied extensively by several communities. It is also of great practical importance given the widespread use of LP relaxations in real-world problems. The relaxation on the local polytope is very popular, though recently tighter relaxations have been implemented with impressive results (Komodakis and Paragios, 2008; Batra et al., 2011).

We have provided intuitive proofs and derived new results that deepen our understanding and may help to provide guidance in practice, including a general decomposition result (Theorem 12). Theorem 8 on hybrid conditions (combining restrictions on topology and potentials) for tightness of LP+TRI is interesting for several reasons. It improves our understanding of why and when the relaxation will perform well. It supports the interesting characterization of almost balanced models, which, to our knowledge, was not much considered prior to Weller (2015b). It shows that LP+TRI dominates the MWSS approach, in the sense that LP+TRI is guaranteed to solve a strict superset of MAP inference problems for any valid potentials in polynomial time. Finally, it provides an important step into hybrid characterizations, which remains an exciting uncharted field following success in characterizations of tractability using only topological constraints (Chandrasekaran et al., 2008), or only families of potentials (Kolmogorov et al., 2015; Thapper and Živný, 2015).

Note that by combining Theorems 8 and 12, an even larger class of models may be shown to be tight for LP+TRI by pasting almost balanced models together on edges in certain settings: for each submodel, the pasted edge must include its special variable.

In future work, we plan to examine higher order relaxations in the Sherali-Adams hierarchy, which impose consistency over larger clusters. LP+LOC=$\mathcal{L}_2$ is tight for any balanced model and we now know that LP+TRI=$\mathcal{L}_3$ is tight for any almost balanced model. It will be interesting to explore whether LP+$\mathcal{L}_4$ is tight for any model that can be rendered balanced by deleting two variables.

It may be tempting to conjecture that if LP+$\mathcal{L}_r$ is tight over a model class for some r, then if an extra variable is added with arbitrary interactions, LP+$\mathcal{L}_{r+1}$ will be tight on the larger model. However, this is false. Consider a planar binary pairwise model with no singleton potentials. LP+TRI is tight for such models (Barahona, 1983); yet if one adds a new variable connected to all of the original ones, the MAP inference task becomes NP-hard (Barahona, 1982).

Acknowledgements

We thank Tomaž Slivnik and the anonymous referees for helpful comments and suggestions. MR acknowledges support by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016516/1 for the University of Cambridge Centre for Doctoral Training, the Cambridge Centre for Analysis. DS was supported by NSF CAREER award #1350965.

References

F. Barahona. On the computational complexity of Ising spin glass models. *Journal of Physics A: Mathematical and General*, 15(10):3241, 1982.

F. Barahona. The max-cut problem on graphs not contractible to K_5. *Operations Research Letters*, 2(3):107–111, 1983.

D. Batra, S. Nowozin, and P. Kohli. Tighter relaxations for MAP-MRF inference: A local primal-dual gap based separation algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 146–154, 2011.

V. Chandrasekaran, N. Srebro, and P. Harsha. Complexity of inference in graphical models. In D. McAllester and P. Myllymäki, editors, *UAI*, pages 70–78. AUAI Press, 2008. ISBN 0-9749039-4-9.

R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.

C. De Simone. The cut polytope and the Boolean quadric polytope. *Discrete Mathematics*, 79(1):71–75, 1989/90.

M. Deza. Matrices des formes quadratiques non négatives pour des arguments binaires. *Comptes rendus de l'Académie des Sciences de Paris*, 277(A):873–875, 1973.

M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 3642042945, 9783642042942.

R. Diestel. *Graph Theory*. Springer, fourth edition, 2010.

P. Hammer. Some network flow problems solved with pseudo-Boolean programming. In *Operations Research*, volume 13, pages 388–399, 1965.

F. Harary. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2:143–146, 1953.

T. Jebara. MAP estimation, message passing, and perfect graphs. In *Uncertainty in Artificial Intelligence*, 2009.

V. Kolmogorov, J. Thapper, and S. Živný. The power of linear programming for general-valued CSPs. *SIAM Journal on Computing*, 44(1):1–36, 2015. doi: 10.1137/130945648.

N. Komodakis and N. Paragios. Beyond loose LP-relaxations: Optimizing MRFs by repairing cycles. In *European Conference on Computer Vision*, pages 806–820, 2008.

M. Laurent and S. Poljak. One-third-integrality in the max-cut problem. *Mathematical Programming*, 71(1):29–50, 1995.

M. Padberg. The Boolean quadric polytope: some characteristics, facets and relatives. *Mathematical Programming*, 45(1-3):139–172, 1989.

S. Sanghavi, D. Shah, and A. Willsky. Message passing for maximum weight independent set. *IEEE Transactions on Information Theory*, 55(11):4822–4834, 2009.

A. Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

H. Sherali and W. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430, 1990.

D. Sontag. *Approximate Inference in Graphical Models using LP Relaxations*. PhD thesis, MIT, EECS, 2010.

D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *UAI*, pages 503–510, 2008.

B. Taskar, V. Chatalbashev, and D. Koller. Learning associative Markov networks. In *ICML*, 2004.

J. Thapper and S. Živný. The complexity of finite-valued CSPs. Technical report, February 2015. URL `http://arxiv.org/abs/1210.2987v3`. arXiv:1210.2987v3.

M. Wainwright and M. Jordan. Treewidth-based conditions for exactness of the Sherali-Adams and Lasserre relaxations. *Univ. California, Berkeley, Technical Report*, 671:4, 2004.

M. Wainwright and M. Jordan. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

A. Weller. Bethe and related pairwise entropy approximations. In *Uncertainty in Artificial Intelligence (UAI)*, 2015a.

A. Weller. Revisiting the limits of MAP inference by MWSS on perfect graphs. In *Artificial Intelligence and Statistics (AISTATS)*, 2015b.

A. Weller. Characterizing tightness of LP relaxations by forbidding signed minors. In *Uncertainty in Artifical Intelligence (UAI)*, 2016a.

A. Weller. Uprooting and rerooting graphical models. In *International Conference on Machine Learning (ICML)*, 2016b.

A. Weller and T. Jebara. Approximating the Bethe partition function. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.