Uprooting and Rerooting Graphical Models

Adrian Weller ADRIAN.WELLER@ENG.CAM.AC.UK
Department of Engineering, University of Cambridge, United Kingdom

Abstract

We show how any binary pairwise model may be 'uprooted' to a fully symmetric model, wherein original singleton potentials are transformed to potentials on edges to an added variable, and then 'rerooted' to a new model on the original number of variables. The new model is essentially equivalent to the original model, with the same partition function and allowing recovery of the original marginals or a MAP configuration, yet may have very different computational properties that allow much more efficient inference. This meta-approach deepens our understanding, may be applied to any existing algorithm to yield improved methods in practice, generalizes earlier theoretical results, and reveals a remarkable interpretation of the triplet-consistent polytope.

1. Introduction

Undirected graphical models, also called Markov random fields (MRFs), have become a central tool in machine learning, providing a powerful and compact way to model relationships between variables. However, many key problems, such as identifying a configuration with highest probability (termed maximum a posteriori or MAP inference), estimating marginal probabilities of subsets of variables (marginal inference) or calculating the normalizing partition function, are typically computationally intractable, leading to much work to identify settings where exact polynomial-time methods apply, or to create approximate algorithms that perform well.

Focusing on binary pairwise models (see §2 for definitions and details), we provide a general meta-method for inference that generalizes and strengthens existing theoretical results, deepens our understanding, and can help significantly in practice. Suppose a model M has n variables $X_1, \dots X_n$ with various singleton and edge potentials. We start by *uprooting* this model to a uniquely determined 'parent' model M^+ where all previous singleton potentials are converted to edge potentials to a newly added variable X_0. This M^+ model is elegantly symmetric: no singleton potentials, and all edge potentials give a score only if incident variables are different. This uprooting is not a novel idea for MAP inference (Barahona et al., 1988; Sontag, 2007) but we believe the other ideas presented here are new.

The uprooted M^+ model is interesting in itself; for example, its partition function is exactly twice that of the original model M, which we may consider as the parent M^+ model *rooted* at X_0. A key idea is that we can *reroot* M^+ at any other variable, for example X_2, to yield an equivalent model on n variables $X_0, X_1, X_3, \dots, X_n$, which has new singleton potentials determined by the edge potentials to X_2 in M^+. In effect, this is a different view or 'crystallization' of the parent symmetric M^+ model. Call this X_2-rooted model M_2 (we could root at any variable X_i to obtain M_i; note that the original model M is M_0).

We make the following observations.

- M_2 indeed represents essentially the same model as M. It lies in the equivalence class of models that map to the same unique symmetric representation M^+.
- M_2 has the same partition function as M but may have very different computational properties. The original model M might be hard but M_2 could be easy.
- Using any existing inference method for M_2, it is not hard also to recover all the original singleton marginals or a MAP configuration of M, see §4.1.
- Hence we have a general meta-method for inference: given any inference approach, instead of applying it to M, we can instead consider various equivalent rerooted models and apply the approach to one of them instead, which may work much better.
- Many existing methods and bounds apply only to particular ranges of edge and singleton potentials, which are changed after rerooting. Hence, we can generalize existing approaches. We discuss various implications in §5. For example, we can use the very efficient max flow/min cut method for MAP inference in a model if all edges are attractive with no conditions on singleton potentials (more generally if the model is *balanced*, see §2.1). This might not be possible in the original model M but will

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

be possible in some rerooted model iff there exists some variable X_i in M^+ s.t. after rooting at X_i, the remainder M_i is balanced. This is equivalent to the condition that M^+ is *almost balanced*. This can be tested efficiently.

- Understanding that singleton and pairwise potentials appear different only due to a particular choice of root in M^+ provides an important fresh perspective, leading to a re-evaluation of existing methods, and a remarkable interpretation of the triplet-consistent polytope, see §5.

Binary pairwise models play an important role in many fields such as computer vision (Blake et al., 2011). Further, any discrete graphical model may essentially be converted into an equivalent binary pairwise model, though this may require a large increase in the number of variables.[1]

Contributions. After providing background in §2, we explain the details of the uprooting and rerooting approach in §3-4, showing how inference on a rerooted model allows recovery of information about the original model. This includes a discussion in §4.2 of the relation to clamping, where we introduce a new clamping heuristic that performs particularly well in settings that are likely to arise for rerooting. In §5, we discuss implications of rerooting, showing how it generalizes theoretical results, may be used as a meta-algorithm for inference methods, and provides a fascinating perspective on the triplet-consistent polytope. We provide an empirical evaluation in §6, showing that rerooting can be particularly effective for models with dense, strong edges and weak singleton potentials.

Related Work. What we call uprooting has been described previously as a way to reduce MAP inference of M to the MAXCUT problem in M^+ (Barahona et al., 1988). As we discuss in §4.2, uprooting to M^+ may be viewed as a de-clamping of the model at X_0, while a rerooting may be considered a re-clamping at a different variable. Hence, rerooting replaces an initial implicit clamp choice at X_0 with another. The choice of which root to choose is essentially the question of which variable in M^+ to clamp. Methods to select a variable to clamp have been explored by Eaton and Ghahramani (2009) and Weller and Domke (2016).

2. Preliminaries

We focus on undirected graphical models with n binary variables $X_1, \dots, X_n \in \{0, 1\}$. We consider a probability distribution $p(x) = e^{\theta(x)}/Z(\theta)$ where $x = (x_1, \dots, x_n)$ is one particular configuration of all variables and $\theta(x)$ is the score of configuration x, which decomposes into singleton and pairwise (log) potential terms. The topology of the model is the graph $(\mathcal{V}, \mathcal{E})$, that is $\mathcal{V} = \{1, \dots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains an edge for each pairwise score relationship. We assume a reparameterization to the minimal representation (Wainwright and Jordan, 2008) wherein the score may be written

$$\theta(x) = \sum_{i \in \mathcal{V}} \theta_i x_i - \frac{1}{2} \sum_{(i,j) \in \mathcal{E}} W_{ij} \mathbb{1}[x_i \neq x_j], \qquad (1)$$

where $\mathbb{1}[\cdot]$ is the indicator function. $Z(\theta)$ is a normalizing constant, called the *partition function*, which ensures that probabilities sum to 1, i.e. $Z(\theta) = \sum_{x \in \{0,1\}^n} e^{\theta(x)}$.

Note that (1) gives a score to an edge only if its incident variables are different. The factor of $-\frac{1}{2}$ before the edge potentials means that the signs and scaling of our parameters are consistent with earlier work such as (Welling and Teh, 2001; Weller and Domke, 2016). If $W_{ij} > 0$ then the edge (i, j) is *attractive* and tends to pull its incident variables towards the same value; if $W_{ij} < 0$ then the edge is *repulsive* and tends to push apart the values of its variables.

2.1. Attractive, Mixed and Balanced Models

If all edges of a model are attractive, i.e. if $W_{ij} \geq 0 \; \forall (i, j) \in \mathcal{E}$, then we say that the model is attractive, else it is mixed. Sometimes a mixed model may be converted to an equivalent attractive model by flipping (also called switching) a subset of variables S, which reverses the signs of their singleton potentials and of the edge potentials between variables in S and variables in $\mathcal{V} \setminus S$; if possible, such a mixed model is called *balanced*. Harary (1953) showed that a model is balanced iff it does not contain a *frustrated cycle*, which is a cycle containing an odd number of repulsive edges. This may be checked and, if balanced, then a flipping set S found, in time linear in $|\mathcal{E}|$ (Harary and Kabell, 1980). Hence, results for attractive models readily extend to the larger class of balanced models.

Notation. For any $a \in \{0, 1\}$, let $\bar{a} = 1 - a$ (this flips $0 \leftrightarrow 1$). Similarly, for a vector $x = (x_1, \dots, x_n)$, let $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n) = (1 - x_1, \dots, 1 - x_n)$. For a configuration $y = (y_0, y_1, \dots, y_n)$ of M^+, and $a \in \{0, 1\}$, we may write $y = (a, x)$ to mean $y = (a, x_1, \dots, x_n)$.

3. Uprooting a Model

We show how any model M on n variables $X_1, \dots, X_n$ with singleton potentials may be uprooted to a unique symmetric (i.e. no singleton potentials) model M^+ on $n + 1$ variables $X_0, X_1, \dots, X_n$. Edges to the extra variable X_0 encode the original singleton potentials.[2]

[1] Eaton and Ghahramani (2013) show that this is strictly true if all model states have probability strictly > 0, otherwise an arbitrarily good approximation is possible.

[2] If the original model M has no singleton potentials, then it may be regarded as already in M^+ form. It may still be rooted at any variable, as described in §4.

M^+ configuration				edges: score ✓ if ends are different					
x_0	x_1	x_2	x_3	e_{01}	e_{02}	e_{03}	e_{12}	e_{13}	e_{23}
0	0	0	0						
0	**0**	**0**	**1**			✓		✓	✓
0	0	1	0		✓		✓		✓
0	0	1	1		✓	✓	✓	✓	
0	1	0	0	✓			✓	✓	
0	**1**	**0**	**1**	✓		✓	✓		✓
0	1	1	0	✓	✓			✓	✓
0	1	1	1	✓	✓	✓			
1	**0**	**0**	**0**	✓	✓	✓			
1	0	0	1	✓	✓			✓	✓
1	0	1	0	✓		✓	✓		✓
1	0	1	1	✓			✓	✓	
1	**1**	**0**	**0**		✓	✓	✓	✓	
1	1	0	1		✓		✓		✓
1	1	1	0			✓		✓	✓
1	1	1	1						

Table 1. An example showing all configurations of an uprooted M^+ model on 4 variables. The original model M has 3 variables X_1, X_2, X_3 then X_0 was added. Each configuration of M features twice: once as $(0, x_1, x_2, x_3)$ in the top half, and then again with all settings flipped as $(1, \bar{x}_1, \bar{x}_2, \bar{x}_3)$ in the bottom. Each has the same score in M^+, with the score determined only by the edges which are activated: see the check marks on the right and note their reflective symmetry across the horizontal line in the middle of the table.
The pink shaded rows indicate the configurations for the rerooted model M_2 where $X_2 = 0$. Observe that given the symmetry, these correspond 1-1 with the configurations of M. Hence, we can recover the partition function, marginal probabilities or a MAP configuration for M by inference on M_2. For example, $p_0(X_3 = 1)$ for M may be computed as $p_2(X_3 \neq X_0)$ for M_2, i.e. sum over the rows shown in bold. Each of the rows in the top half with $x_3 = 1$ which is missing from M_2 (that is, not shaded pink) has an exactly corresponding row in the bottom half, as indicated by blue curves in the table. See §3 and §4 for details.

Let $y = (y_0, y_1, \ldots, y_n)$ be a configuration in M^+ of its $n+1$ variables, and let $\phi(y)$ be its score in M^+. Requiring $\phi(y)$ to be in the same form as (1) but with no singleton potentials, and to match the scores of configurations in M when $x_0 = 0$, i.e. requiring $\phi(0, x) = \theta(x)$, implies

$$\phi(y) = -\frac{1}{2}\sum_{\mathcal{E}'} W_{ij}\mathbb{1}[y_i \neq y_j], \tag{2}$$

where the edges of M^+ are $\mathcal{E}' = \mathcal{E} \cup \mathcal{F}$ consisting of the original edges $\mathcal{E}$ of M, together with new edges $\mathcal{F}$ which are added to the new variable X_0, given by $\mathcal{F} = \{(0, i) : \theta_i \neq 0\}$. Weights for edges in $\mathcal{E}$ remain unchanged. Weights for the new edges in $\mathcal{F}$ are set as $W_{0i} = -2\theta_i$. To see this, note that the singleton potentials in (1) are already in the form $\theta_i \mathbb{1}[x_i \neq x_0 | x_0 = 0]$.

Note the sign flip when a singleton potential is converted to an edge potential, e.g. $\theta_i > 0$ becomes a repulsive edge in M^+ with $W_{0i} < 0$. This is an unavoidable consequence of choosing parameters in (1) to match earlier work. It may be helpful to think of $\theta_i > 0$ as encouraging X_i to be different to 0, i.e. a repulsive edge from $X_0 = 0$.

Observe that each configuration x of M maps to *two* configurations y_0 and y_1 in M^+,

$$M : x = (x_1, \ldots, x_n) \to M^+ : \begin{cases} y_0 = (0, x) \\ y_1 = \bar{y}_0 = (1, \bar{x}), \end{cases} \tag{3}$$

i.e. $y_0 = (0, x_1, \ldots, x_n)$ and $y_1 = \bar{y}_0 = (1, \bar{x}_1, \ldots, \bar{x}_n)$. Given the symmetry of (2), it is clear that $\phi(y_0) = \phi(y_1)$. See Table 1 and Figure 1 for an example.

The partition function for M^+ is clearly twice that of M, i.e. $Z(\phi) = 2Z(\theta)$. The original model M is exactly M^+ conditioned on $X_0 = 0$, and we can write $M = M_0$.

4. Rerooting a Model

The symmetric model M^+ with form (2) may be rooted at any variable X_i by conditioning on $X_i = 0$ to yield a model on n variables consisting of $\{X_0, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$, which we write as M_i.[3] See Table 1 and Figure 1 for an example.

Considering (3), for any i, there is a score-preserving 1-1 correspondence between configurations in M and those in M_i which matches x in M with whichever of y_0 or y_1 has $x_i = 0$ (the x_i coordinate is removed to give the configuration in M_i). Equivalently, if x in M has $x_i = 0$, then it matches to $(0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ in M_i, otherwise it matches to the same configuration but with all settings flipped, i.e. $(1, \bar{x}_1, \ldots, \bar{x}_{i-1}, \bar{x}_{i+1}, \ldots, \bar{x}_n)$.

4.1. Recovery of Original MAP Configuration, Partition Function or Marginals

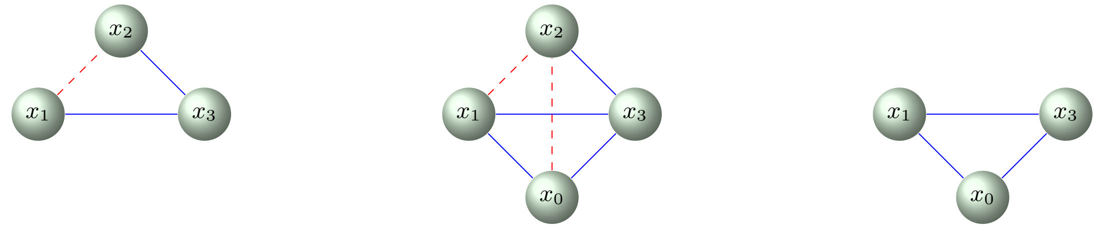
In this Section, we show that if inference can be performed on a rerooted model M_i, then we can recover results for the original model M.

4.1.1. MAP inference

For MAP inference, given the score-preserving 1-1 correspondence of configurations noted above in §4, if a MAP configuration is determined for M_i, then the corresponding configuration in M is a MAP configuration for M with the same score. Specifically, we have the following result.

Lemma 1. *If $(x_0^*, \ldots, x_{i-1}^*, x_{i+1}^*, \ldots, x_n^*)$ is a MAP con-*

[3]Given the symmetry, one could instead equivalently consider M^+ conditioned on $X_i = 1$ but then one would need to flip variables after performing inference in order to match the original interpretation in M.



(a) Original model $M = M_0$
$n = 3$ variables: x_1, x_2, x_3
edge + singleton potentials: $\theta_1, \theta_3 < 0; \theta_2 > 0$
frustrated cycle

(b) *Uprooted* symmetric model M^+
$n + 1 = 4$ variables: x_0, x_1, x_2, x_3
only edge, no singleton potentials
almost balanced

(c) Equivalent model M_2 *rerooted* at x_2
$n = 3$ variables: x_0, x_1, x_3
edge + singleton potentials: $\theta'_0, \theta'_1 > 0; \theta'_3 < 0$
attractive

Figure 1. Examples of (a) an original model $M = M_0$ on three variables, together with (b) its unique uprooted model M^+, and (c) a different rooting of M^+ at x_2 to yield M_2, where now all edges are attractive. Solid blue (dashed red) edges are attractive (repulsive).

figuration for M_i, then the corresponding MAP configuration for M, with the same score, is:

$$\begin{cases} m = (x_1^*, \ldots, x_{i-1}^*, x_i = 0, x_{i+1}^*, \ldots, x_n^*) & \textit{if } x_0^* = 0 \\ \bar{m} = (\bar{x}_1^*, \ldots, \bar{x}_{i-1}^*, x_i = 1, \bar{x}_{i+1}^*, \ldots, \bar{x}_n^*) & \textit{if } x_0^* = 1. \end{cases}$$

4.1.2. Marginal inference and computing Z

Since M_i and M have corresponding configurations with equal scores, they have the same partition function.

In order to differentiate between probabilities obtained for different models, we use the following notation: let p_i be the probability distribution in model M_i, in particular p_0 is the distribution for model M_0 which is the original model M; let p_+ be the distribution in the uprooted model M^+.

Each model M_i is the result of conditioning on $X_i = 0$ in M^+. We would like to perform (exact or approximate) inference on M_i to obtain p_i, then use this to recover marginals $p_0(X_j = 1)$ $\forall j \in \{1, \ldots, n\}$. The following result achieves this. See Table 1 for an example.

Lemma 2. *Given distribution p_i for any $i \in \{1, \ldots, n\}$, the marginals $p_0(X_j = 1)$ for the original model $M = M_0$ may be recovered as follows:*

$$p_0(X_j = 1) = \begin{cases} p_i(X_0 = 1) & j = i \\ p_i(X_j \neq X_0) & j \neq i. \end{cases}$$

Proof. This follows from the symmetry of M^+, see the Appendix for details. □

4.2. Relation to Clamping, How to Choose a Root?

Conditioning a model on a variable taking a particular value is sometimes called *clamping* (Eaton and Ghahramani, 2009; Weller and Jebara, 2014). Since M_i is M^+ conditioned on $X_i = 0$, we may view uprooting from $M = M_0$ to M^+ as *de-clamping* X_0 back to a parent model; then rerooting at variable X_i is a *re-clamping* at $X_i = 0$ to obtain M_i.

In typical clamping for inference, one must condition a variable to each of its values and combine all results (for example, if estimating Z, one must sum the approximate sub-partition functions). For binary variables, this requires running your inference algorithm twice. In contrast, a rooted model gets a 'clamping for free' at the root variable, with just one inference run required.

A natural question is how to choose a good root when rerooting a model? Given the interpretation of rooting as clamping, we can draw on earlier work. Weller and Domke (2016) examined a range of heuristics and concluded that a fast method called maxW typically performs very well for approximate inference.[4] The idea behind maxW is that many popular methods of approximate inference, such as belief propagation, are exact on acyclic models but can perform poorly when there are cycles composed of strong edge weights. It is NP-hard to identify heavy cycles but the following simple heuristic was shown to be empirically effective. For each variable, a sum of absolute values of incident edge weights is computed, then the variable with the highest sum is selected to clamp. When it is clamped, this variable is effectively removed from the model, thereby eliminating any cycles which ran through it.

4.2.1. A new method: maxtW

In §6, we explore the value of rerooting using the earlier maxW heuristic to select the root variable. We observe that maxW sometimes performs well, but one setting where it can perform poorly is if a choice must be made between one variable that has a few strong edges and another that has many weak edges. When rerooting, this may happen frequently. For example, consider an initial model M with a grid topology, and singleton potentials that are low relative to edge potentials: in M^+ this leads to X_0 having a weak

[4] Weller and Domke (2016) showed that a more sophisticated variant called maxW+core+TRE performed slightly better in general, but TRE is redundant for the fully symmetric M^+ model, and the core idea makes no difference in the experiments we run.

edge to every other variable, whereas other variables have few strong edges. maxW simply picks the variable i with highest $\sum_{j \in \mathcal{N}(i)} |W_{ij}|$, where $\mathcal{N}(i)$ is the set of variables adjacent to i. However, the direct influence of a strong edge weight does not keep increasing linearly with its weight, rather it reaches a hard saturation level (Weller and Jebara, 2014, Supplement, Lemma 12). Here we address this by introducing an alternative heuristic we call *maxtW*, which picks the variable i with $\max \sum_{j \in \mathcal{N}(i)} \tanh |\frac{W_{ij}}{4}|$. This form was chosen based on earlier work on loop series expansions (Weller et al., 2014; Sudderth et al., 2007). See results in §6 where maxtW can dramatically outperform maxW on grids.

5. Implications of Rerooting

Rerooting provides a conceptual framework to view singleton and edge potentials as essentially equivalent except for a choice of rooting of the symmetric uprooted M^+ parent model. After rerooting, it may be possible to apply many methods or bounds that were unavailable for the original model M. We consider important examples below.

5.1. MAP Inference

The success of many existing methods of MAP inference depends critically on the nature of the edge potentials of a model, but can be relatively insensitive to the singleton potentials. For example, both the max flow/min cut method (Greig et al., 1989) and the basic linear programming (LP) relaxation over the local polytope LOC (Wainwright and Jordan, 2008) provide an exact solution in polynomial-time if the model is attractive. These approaches generalize to balanced models, see §2.1.

With rerooting, these methods can now be used on the significantly larger class of model where some rooting M_i exists which is balanced. This holds iff the uprooted model M^+ is *almost balanced*, which means it contains a variable such that removing it renders the remaining model balanced. See Figure 1 for an example.

Almost balanced models have received recent attention. Jebara (2009) introduced a method for MAP inference via a reduction to the graph-theoretic challenge of identifying a *maximum weight stable set* (MWSS) in a derived weighted graph, which if *perfect*, allows an exact solution to be obtained efficiently. Weller (2015b) showed that this method applies iff the *block decomposition* of the model M yields blocks (maximal 2-connected components) which are all almost balanced. With rerooting, we can extend this method to models M that have uprooted models M^+ that are *2-almost balanced*, i.e. models which can be rendered balanced by deleting 2 variables (since by rooting at either of these variables, the rooted model is almost balanced).

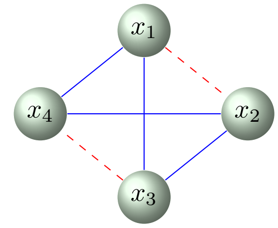


Figure 2. An example of a model M which is not almost balanced, hence does not satisfy the conditions of Weller et al. (2016) for tightness of LP+TRI. Nevertheless, by Theorem 4, it is sufficient if its uprooted M^+ is 2-almost balanced. Hence LP+TRI will always be tight for (any rerooting of) this model provided singleton potentials are not of the form: θ_1, θ_2 take one sign (either positive or negative); and θ_3, θ_4 take the other. Solid blue (dashed red) edges are attractive (repulsive). See §5.2 for details.

5.2. Local and Triplet Polytopes, Why 'Rooting'?

Weller et al. (2016) showed that the LP relaxation on the triplet-consistent polytope, LP+TRI, yields an exact MAP configuration of a model M provided it is almost balanced (for any singleton potentials). As above this can now be generalized to be used for any model if its uprooted model M^+ is 2-almost balanced, since then a rooting exists which is almost balanced. In fact, we can achieve a much stronger result due to the following remarkable property of TRI.

Theorem 3 (TRI is 'universally rooted')**.** *LP+TRI yields the same optimum score for M as for any rerooting M_i; hence LP+TRI is either tight for all rerootings or for none.*

Theorem 3 immediately yields the following new result.

Theorem 4. *LP+TRI is tight for (any rerooting of) a model M whose uprooted model M^+ is 2-almost balanced.*

See the Appendix §8 for details and proofs. This beautifully shows the common nature of edge and singleton potentials for TRI, examining the signs of all edges in M^+ in the same symmetric way.

Theorem 4 helps us to understand tightness of LP+TRI on real-world vision tasks, where learned models are close to attractive due to contiguity of objects. As a small example, Theorem 4 shows that LP+TRI is tight for the model shown in Figure 2, despite it not being almost balanced, provided the signs of singleton potentials leave M^+ 2-almost balanced (this holds for all values unless: θ_1, θ_2 take one sign (positive or negative); and θ_3, θ_4 take the other).

Here we sketch the reasoning. The following polytopes are equivalent (see Deza and Laurent, 1997):

n variables + $\binom{n}{2}$ edges		$\binom{n+1}{2}$ edges
Marginal polytope of M	$\leftrightarrow$	Cut polytope of M^+
TRI relaxation	$\leftrightarrow$	MET relaxation
LOC relaxation	$\leftrightarrow$	RMET relaxation

MET, the semimetric polytope relaxation of the cut polytope, enforces triplet constraints on *every triplet* of $\{X_0, \ldots, X_n\}$. In contrast, RMET, the *rooted* semimetric polytope, enforces these same triplet constraints *only on triplets that include the root X_0 variable*.

This explains the name *rooting*. LOC is equivalent to a specifically rooted RMET polytope, which is why approaches over LOC (including many message-passing algorithms) deal differently with singleton and edge potentials, and might benefit significantly from rerooting. TRI, however, is equivalent to MET, which deals symmetrically with all variables in M^+ and corresponds to *a simultaneous rooting at every variable*. This intriguing observation likely has further theoretical and algorithmic consequences, which we aim to explore in future work.

5.3. Belief Propagation

Belief propagation (BP, Pearl, 1988), or more generally the Bethe approximation (Yedidia et al., 2000), is a widely used approach for approximate inference, guaranteed to yield exact results in linear time for models without cycles. When applied to models with cycles, it often yields strikingly accurate results but may fail to converge altogether.

Much work has analyzed the convergence of BP, and the uniqueness of a fixed point, relying either on the strength of edge interactions (Heskes, 2004; Mooij and Kappen, 2005), or just on their signs (Watanabe, 2011). Mooij and Kappen (2007) refined their earlier result by considering also singleton potentials, but these are incorporated quite differently to the edge potentials. Hence, by rerooting it may be possible to provide theoretical guarantees on performance that are not available on the initial model.

As one example, it is known that if a model has one cycle, then the Bethe free energy is convex and BP has a unique fixed point (Pakzad and Anantharam, 2002). Consider the model shown in Figure 1. The original model (a) is a frustrated cycle, hence the BP estimate of Z will be too high, with unbounded high error as edge weights increase (Weller, 2015a, §6.3). In contrast, the rerooted model (c) is attractive, hence the BP estimate is always in the range $[Z/2, Z]$ for any potentials (Weller and Jebara, 2014).

5.4. FPRAS, Bounds

Jerrum and Sinclair (1993) devised a *fully polynomial-time randomized approximation scheme* (FPRAS) for the partition function of a model M provided it is attractive and all singleton potentials are consistent in taking the same sign (positive or negative). This generalizes to any model with uprooted model M^+ which is balanced (see §2.1).

Various methods have been developed to bound the partition function or marginals of a model (Leisink and Kappen, 2003; Ihler, 2007; Mooij and Kappen, 2008). These treat singleton and edge potentials differently, hence may be generalized by considering rerootings.

5.5. Remarks

Comparison to clamping. Some of the benefits of rerooting could also be obtained by usual clamping of M. For example, if a model can be rendered balanced by rerooting at X_i, then this could also be achieved by clamping X_i in the original model. However, this would require performing multiple inference runs and combining results, rather than using the 'free clamping' available with a rerooting, see §4.2. Further, several results, including Theorem 4 and the observations in §5.2 on the triplet polytope, are not possible without considering rerooting.

Evaluation of inference methods. Approximate inference methods are typically evaluated empirically on a range of models, where singleton and edge potentials are treated quite differently. Often singleton potentials are drawn from some fixed narrow range while edge potentials are drawn from a range whose width is varied widely. From an uprooted model perspective, singleton and edge potentials are equivalent. Hence: (i) Varying singleton and edge potentials differently in empirical evaluations may be a peculiar assumption, though it could be justified as reflecting typical patterns in the real world; (ii) The implicit choice of root may be poor in some cases (i.e. results might be improved significantly by rerooting), which will obscure the underlying performance attributes of the inference method. We examine the extent of this effect in §6.

6. Experiments

Following the observation in §5.5, we are interested in the effect of rerooting in standard settings for empirical evaluation. We compared performance of estimating the partition function and singleton marginals after different rerootings of three popular approximation methods: Bethe (using the approach of Heskes et al., 2003 to ensure convergence), tree-reweighted (TRW, Wainwright et al., 2005) and naive mean field. For true values, we used the junction tree algorithm. All methods were implemented using libDAI (Mooij, 2010), see the Appendix §9 for details.

We ran experiments on the following topologies and model sizes: complete graphs on 10 and 15 variables; grids of size 5×5 and 9×9. All potentials were drawn randomly: mixed models used $W_{ij} \sim U[-W_{\max}, W_{\max}]$, attractive models used $W_{ij} \sim U[0, W_{\max}]$, as $W_{\max}$ was varied; singleton potentials were drawn either from a low range $\theta_i \sim [-0.1, 0.1]$, medium range $\theta_i \sim [-2, 2]$, or from a range commensurate with edge potentials, i.e. $\theta_i \sim U[-W_{\max}/2, W_{\max}/2]$, with the factor of 2 needed given the form of (1). These settings allow direct comparison to earlier work such as by Weller and Domke (2016) or Weller et al. (2014). Others (Meshi et al., 2009; Sontag and Jaakkola, 2007) use binary variables with values in $\{-1, 1\}$ instead of $\{0, 1\}$, hence their edge (singleton) potentials should be multiplied by 4 (2, respectively) when making comparisons. We plot average error over 100 random runs for each setting. All results are in the Appendix.

As in §4.2, any rooting of a model M may be considered a clamping of the uprooted model M^+. The original model $M = M_0$ implicitly reflects the decision to clamp at X_0, which might be a good or bad choice depending on the setting. Recall from §4.2 that maxW often performs well for selecting a variable to clamp, picking one with highest sum of incident edge strengths (taking absolute values). However, if a choice must be made between variable A with many weak edges, or B with few strong edges, maxW may make a poor choice by not recognizing that A is often better since the influence of strong edges saturates. Hence we introduced the maxtW heuristic in §4.2.1, which selects variable X_i with $\max \sum_{j \in \mathcal{N}(i)} \tanh |\frac{W_{ij}}{4}|$.

Our plots show average error when applying the approximate inference method to: the original model M; the uprooted model M^+; then rerootings at: the *worst* variable, the *best* variable, the *maxW* variable, and the *maxtW* variable. *Best* and *worst* always refer to the variable at which rerooting gave with hindsight the best and worst error for the partition function (even in plots for marginals).

6.1. Results

Figure 3 summarizes results for Bethe, typically the most accurate method. Looking across all results (see Appendix §9), we make the following observations.

For complete graphs, maxW and maxtW perform well. Rerooting is very effective as edge strength grows, both at low and medium levels of singleton potentials. This makes sense, since in this setting, the default rooting at X_0 has relatively weak edges, and all variables in M^+ have the same number of edges, hence it is likely to be very beneficial to switch to a different root with stronger edges. When singleton and edge potentials vary together, edges in M^+ are all similar, but X_0 is an average variable to clamp, whereas we do somewhat better by choosing a good variable.

For grids, maxtW is much better than maxW (maxW performs very poorly in some cases), appearing to handle uneven edge weights in M^+ well. At low singleton potentials, rerooting is very helpful but this benefit disappears for stronger singleton potentials, where the original rooting performs equally to maxtW.

Results for MF and TRW are qualitatively similar to Bethe, with Bethe typically performing best. For mixed models with strong edges, MF performs very well. This is likely due to MF optimizing within the marginal polytope, whereas Bethe and TRW use the local polytope, in which strong frustrated cycles can lead to high error.

Based on maxtW, we can suggest a guideline for when rerooting is likely to be helpful. For example, for a 4-way grid with n variables, constant singleton potentials T and edge weights W: $4 \tanh \frac{W}{4} + \tanh \frac{2T}{4} > n \tanh \frac{2T}{4} \Leftrightarrow 4 \tanh \frac{W}{4} > (n-1) \tanh \frac{T}{2}$. This is conservative since more randomness increases the value of rerooting by raising the chance of a better root. Demonstrating this, observe in Figure 3 that when singleton potentials are low, the improvement in $\log Z$ estimate from rerooting using maxtW is about the same for 9×9 grids as for smaller 5×5 grids.

7. Conclusion

We introduced the idea of uprooting and then rerooting any binary pairwise graphical model. This immediately leads to a meta-algorithm for inference into which any existing approach may be slotted, and generalizes important theoretical results. Further, it provides an elegant conceptual framework for rethinking singleton and edge potentials with methodological consequences for how we evaluate models and methods. One application in §5.2 leads to Theorem 4, a strong result for tightness of LP relaxations on the triplet-consistent polytope TRI, and a remarkable interpretation of TRI as universally rooted.

Rerooting switches an implicit clamp choice in the uprooted model at X_0 (perhaps a poor choice), instead to a carefully selected clamp choice, almost for free. This applies even for large models where it is desirable to clamp a series of variables: by rerooting, we may obtain one of the series for free, sometimes achieving dramatic improvements in accuracy with little work required. If there are multiple connected components, each should be handled separately, with its own X_0-type variable. This could be useful for (repeatedly) composing clamping and then rerooting each separated component.

Rerooting is particularly effective when a model has dense, strong edge weights and weak singleton potentials (a difficult setting for many existing methods). Our new maxtW heuristic performs particularly well in this setting (and should also be helpful for standard clamping approaches), sometimes dramatically outperforming the earlier maxW method. maxtW also provides a useful guideline for when uprooting is likely to be helpful, see the last paragraph of §6.1.

It will be interesting in future work to study further consequences of our interpretation of the triplet-consistent polytope, to consider the value of rerooting for approaches to learning graphical models, and to explore the benefits of rerooting when variables have a higher number of labels.

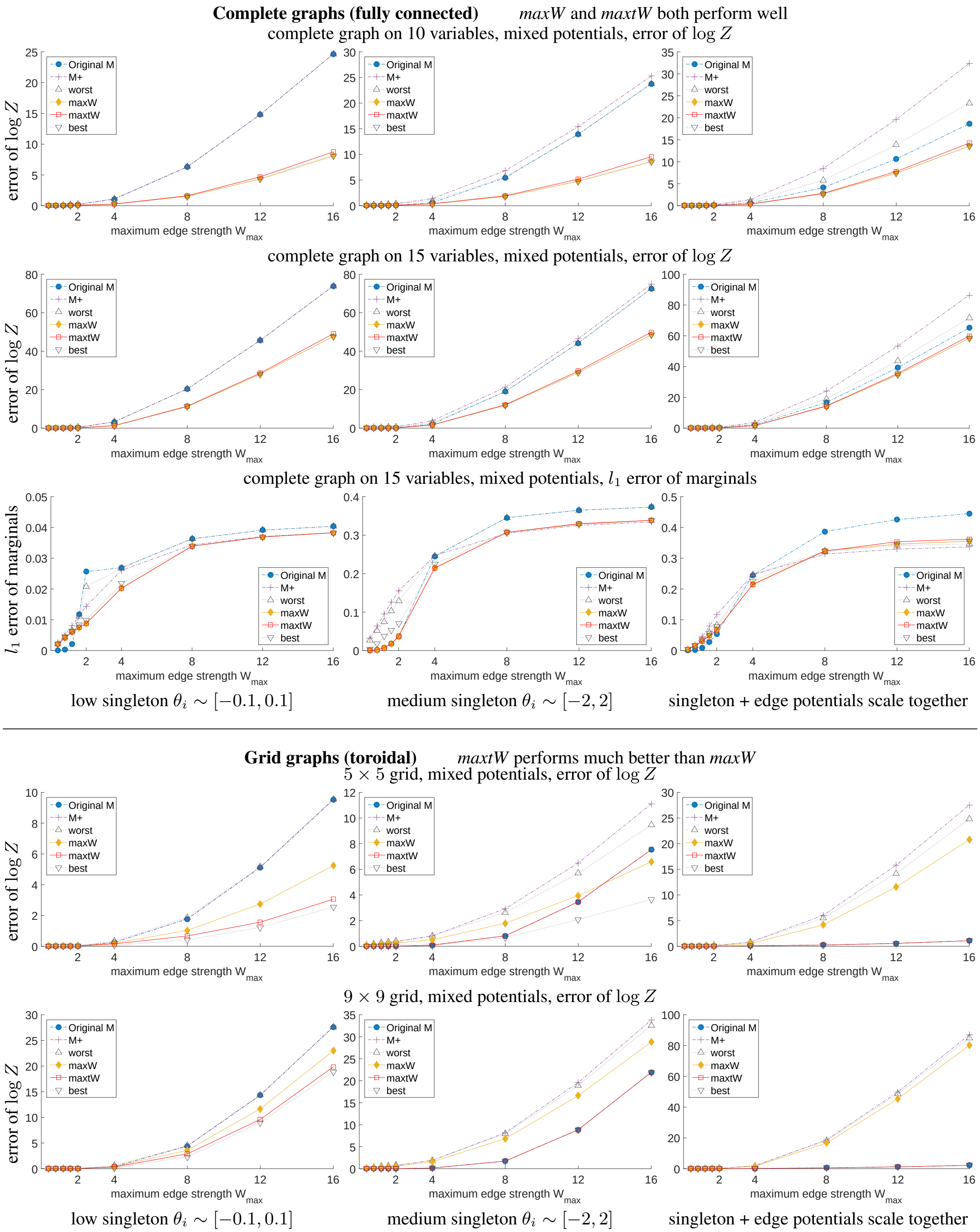


Figure 3. Average error of Bethe approximation for models with mixed potentials over 100 runs, showing smaller and larger models for comparison. Top: complete graphs (10 and 15 variables). Bottom: toroidal grid graphs (5×5 and 9×9). Each column shows different settings for singleton potentials: left is low range; centre is medium range; right varies singleton and edge potentials together. See §6.

Acknowledgments

The author thanks Justin Domke, Tony Jebara, Mark Rowland, Nilesh Tripuraneni, David Sontag and the anonymous reviewers for helpful comments.

References

F. Barahona and A. Mahjoub. On the cut polytope. *Mathematical Programming*, 36(2):157–173, 1986.

F. Barahona, M. Grötschel, M. Jünger, and G. Reinelt. An application of combinatorial optimization to statistical physics and circuit layout design. *Operations Research*, 36(3):493–513, 1988.

A. Blake, P. Kohli, and C. Rother, editors. *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer Publishing Company, Incorporated, 1st edition, 1997. ISBN 978-3-642-04294-2.

F. Eaton and Z. Ghahramani. Choosing a variable to clamp: Approximate inference using conditioned belief propagation. In *Artificial Intelligence and Statistics*, 2009.

F. Eaton and Z. Ghahramani. Model reductions for inference: Generality of pairwise, binary, and planar factor graphs. *Neural Computation*, 25(5):1213–1260, 2013.

D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *J. Royal Statistical Soc., Series B*, 51(2):271–279, 1989.

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

F. Harary and J. Kabell. A simple algorithm to detect balance in signed graphs. *Mathematical Social Sciences*, 1(1):131–136, 1980.

T. Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16(11):2379–2413, 2004.

T. Heskes, K. Albers, and B. Kappen. Approximate inference and constrained optimization. In *UAI*, pages 313–320, 2003.

A. Ihler. Accuracy bounds for belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*, 2007.

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

M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.*, 22(5):1087–1116, 1993.

M. Leisink and H. Kappen. Bound propagation. *J. Artif. Intell. Res. (JAIR)*, 19:139–154, 2003.

O. Meshi, A. Jaimovich, A. Globerson, and N. Friedman. Convexifying the Bethe free energy. In *UAI*, pages 402–410, 2009.

J. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, August 2010. URL http://www.jmlr.org/papers/volume11/mooij10a/mooij10a.pdf.

J. Mooij and H. Kappen. On the properties of the Bethe approximation and loopy belief propagation on binary networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005.

J. Mooij and H. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, December 2007.

J. M. Mooij and H. J. Kappen. Bounds on marginal probability distributions. In *Neural Information Processing Systems*, pages 1105–1112, 2008.

P. Pakzad and V. Anantharam. Belief propagation and statistical physics. In *Princeton University*, 2002.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

D. Sontag. Cutting plane algorithms for variational inference in graphical models. Master's thesis, MIT, EECS, 2007.

D. Sontag and T. Jaakkola. New outer bounds on the marginal polytope. In *NIPS*, 2007.

E. Sudderth, M. Wainwright, and A. Willsky. Loop series and Bethe variational bounds in attractive graphical models. In *NIPS*, 2007.

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

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

Y. Watanabe. Uniqueness of belief propagation on signed graphs. In *Neural Information Processing Systems*, 2011.

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 *UAI*, 2016.

A. Weller and J. Domke. Clamping improves TRW and mean field approximations. In *Artificial Intelligence and Statistics (AISTATS)*, 2016.

A. Weller and T. Jebara. Clamping variables and approximate inference. In *Neural Information Processing Systems (NIPS)*, 2014.

A. Weller, K. Tang, D. Sontag, and T. Jebara. Understanding the Bethe approximation: When and how can it go wrong? In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

A. Weller, M. Rowland, and D. Sontag. Tightness of LP relaxations for almost balanced models. In *Artificial Intelligence and Statistics (AISTATS)*, 2016.

M. Welling and Y. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*, 2001.

J. Yedidia, W. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, 2000.