Uprooting and Rerooting Higher-Order Graphical Models

Mark Rowland[*]
University of Cambridge
mr504@cam.ac.uk

Adrian Weller[*]
University of Cambridge and Alan Turing Institute
aw665@cam.ac.uk

Abstract

The idea of uprooting and rerooting graphical models was introduced specifically for binary pairwise models by Weller [19] as a way to transform a model to any of a whole equivalence class of related models, such that inference on any one model yields inference results for all others. This is very helpful since inference, or relevant bounds, may be much easier to obtain or more accurate for some model in the class. Here we introduce methods to extend the approach to models with higher-order potentials and develop theoretical insights. In particular, we show that the triplet-consistent polytope TRI is unique in being 'universally rooted'. We demonstrate empirically that rerooting can significantly improve accuracy of methods of inference for higher-order models at negligible computational cost.

1 Introduction

Undirected graphical models with discrete variables are a central tool in machine learning. In this paper, we focus on three canonical tasks of inference: identifying a configuration with highest probability (termed maximum a posteriori or MAP inference), computing marginal probabilities of subsets of variables (marginal inference) and calculating the normalizing constant (partition function). All three tasks are typically computationally intractable, leading to much work to identify settings where exact polynomial-time methods apply, or to develop approximate algorithms that perform well.

Weller [19] introduced an elegant method which first *uproots* and then *reroots* a given model M to any of a whole class of *rerooted* models $\{M_i\}$. The method relies on specific properties of binary pairwise models and makes use of an earlier construction which reduced MAP inference to the MAXCUT problem on the *suspension graph* ∇G (1; 2; 12; 19, see §3 for details). For many important inference tasks, the rerooted models are equivalent in the sense that results for any one model yield results for all others with negligible computational cost. This can be very helpful since various models in the class may present very different computational difficulties for inference.

Here we show how the idea may be generalized to apply to models with higher-order potentials over any number of variables. Such models have many important applications, for example in computer vision [6] or modeling protein interactions [5]. As for pairwise models, we again obtain significant benefits for inference. We also develop a deeper theoretical understanding and derive important new results. We highlight the following contributions:

- In §3-§4, we show how to achieve efficient uprooting and rerooting of binary graphical models with potentials of any order, while still allowing easy recovery of inference results.
- In §5, to simplify the subsequent analysis, we introduce *pure k-potentials* for any order k, which may be of independent interest. We show that there is essentially only one pure k-potential which we call the *even k-potential*, and that even k-potentials form a basis for all model potentials.
- In §6, we carefully analyze the effect of uprooting and rerooting on Sherali-Adams [11] relaxations $\mathbb{L}_r$ of the marginal polytope, for any order r. One surprising observation in §6.2 is that $\mathbb{L}_3$ (the

[*]Authors contributed equally.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

triplet-consistent polytope or TRI) is unique in being *universally rooted*, in the sense that there is an affine score-preserving bijection between $\mathbb{L}_3$ for a model and $\mathbb{L}_3$ for each of its rerootings.
- In §7, our empirical results demonstrate that rerooting can significantly improve accuracy of inference in higher-order models. We introduce effective heuristics to choose a helpful rerooting.

Our observations have further implications for the many variational methods of marginal inference which optimize the sum of score and an entropy approximation over a Sherali-Adams polytope relaxation. These include the Bethe approximation (intimately related to belief propagation) and cluster extensions, tree-reweighted (TRW) approaches and logdet methods [12; 14; 16; 22; 24].

1.1 Background and discussion of theoretical contributions

Based on earlier connections in [2], [19] showed the remarkable result for pairwise models that the triplet-consistent polytope ($\mathbb{L}_3$ or TRI) is universally rooted (in the restricted sense defined in [19, Theorem 3]). This observation allowed straightforward strengthening of previously known results, for example: it was previously shown [23] that the LP relaxation on TRI (LP+TRI) is always tight for an 'almost-balanced' binary pairwise model, that is a model which can be rendered balanced by removing one variable [17]. Given [19, Theorem 3], this earlier result could immediately be significantly strengthened to [19, Theorem 4], which showed that LP+TRI is tight for a binary pairwise model provided only that *some rerooting exists* such that the rerooted model is almost balanced.

Following [19], it was natural to suspect that the universal rootedness property might hold for all (or at least some) $\mathbb{L}_r, r \geq 3$. This would have impact on work such as [10] which examines which signed minors must be forbidden to guarantee tightness of LP+$\mathbb{L}_4$. If $\mathbb{L}_4$ were universally rooted, then it would be possible to simplify significantly the analysis in [10].

Considering this issue led to our analysis of the mappings to symmetrized uprooted polytopes given in our Theorem 17. We believe this is the natural generalization of the lower order relationships of $\mathbb{L}_2$ and $\mathbb{L}_3$ to RMET and MET described in [2], though this direction was not clear initially.

With this formalism, together with the use of even potentials, we demonstrate our Theorems 20 and 21, showing that in fact TRI is unique in being universally rooted (and indeed in a stronger sense than given in [19]). We suggest that this result is surprising and may have further implications.

As a consequence, it is not possible to generate some quick theoretical wins by generalizing previous results as [19] did to derive their Theorem 4, but on the other hand we observe that rerooting may be helpful in practice for any approach using a Sherali-Adams relaxation other than $\mathbb{L}_3$. We verify the potential for significant benefits experimentally in §7.

2 Graphical models

A *discrete graphical model* $M[G(V, E), (\theta_{\mathcal{E}})_{\mathcal{E}\in E}]$ consists of: a *hypergraph* $G = (V, E)$, which has n vertices $V = \{1, \dots, n\}$ corresponding to the variables of the model, and *hyperedges* $E \subseteq \mathcal{P}(V)$, where $\mathcal{P}(V)$ is the powerset of V; together with *potential* functions $(\theta_{\mathcal{E}})_{\mathcal{E}\in E}$ over the hyperedges $\mathcal{E} \in E$. We consider binary random variables $(X_v)_{v\in V}$ with each $X_v \in \mathbb{X}_v = \{0, 1\}$. For a subset $U \subseteq V$, $x_U \in \{0, 1\}^U$ is a configuration of those variables $(X_v)_{v\in U}$. We write $\overline{x}_U$ for the *flipping* of x_U, defined by $\overline{x}_i = 1 - x_i \ \forall i \in U$. The joint probability mass function factors as follows, where the normalizing constant $Z = \sum_{x_V\in\{0,1\}^V} \exp(\text{score}(x_V))$ is the *partition function*:

$$p(x_V) = \frac{1}{Z}\exp\left(\text{score}(x_V)\right), \qquad \text{score}(x_V) = \sum_{\mathcal{E}\in E} \theta_{\mathcal{E}}(x_{\mathcal{E}}). \tag{1}$$

3 Uprooting and rerooting

Our goal is to map a model M to any of a whole family of models $\{M_i\}$ in such a way that inference on any M_i will allow us easily to recover inference results on the original model M. In this section we provide our mapping, then in §4 we explain how to recover inference results for M.

The uprooting mechanism used by Weller [19] first reparametrizes edge potentials to the form $\theta_{ij}(x_i, x_j) = -\frac{1}{2}W_{ij}\mathbb{1}[x_i \neq x_j]$, where $\mathbb{1}[\cdot]$ is the indicator function (a reparameterization modifies

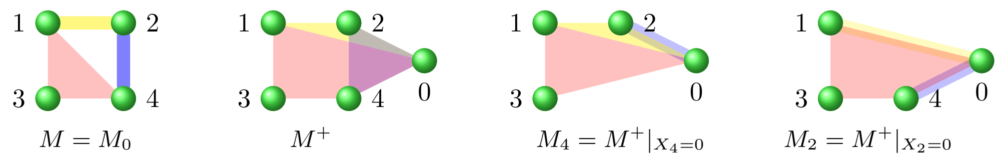


Figure 1: Left: The hypergraph G of a graphical model M over 4 variables, with potentials on the hyperedges $\{1,2\}$, $\{1,3,4\}$, and $\{2,4\}$. Center-left: The suspension hypergraph ∇G of the uprooted model M^+. Center-right: The hypergraph $\nabla G \setminus \{4\}$ of the rerooted model $M_4 = M^+|_{X_4=0}$, i.e. M^+ with X_4 clamped to 0. Right: The hypergraph $\nabla G \setminus \{2\}$ of the rerooted model $M_2 = M^+|_{X_2=0}$, i.e. M^+ with X_2 clamped to 0.

potential functions such that the complete score of each configuration is unchanged, see 15 for details). Next, singleton potentials are converted to edge potentials with this same form by connecting to an added variable X_0. This mechanism had been used previously to reduce MAP inference on M to MAXCUT on the converted model [1; 12], and applies specifically only to binary pairwise models.

We introduce a generalized construction which applies to models with potentials of any order. We first *uproot* a model M to a highly symmetric uprooted model M^+ where an extra variable X_0 is added, in such a way that the original model M is exactly M^+ with X_0 *clamped* to the value 0. Since X_0 is clamped to retrieve M, we may write $M = M_0 := M^+|_{X_0=0}$. Alternatively, we can choose instead to clamp a different variable X_i in M^+ which will lead to the *rerooted* model $M_i := M^+|_{X_i=0}$.

Definition 1 (**Clamping**). For a graphical model $M[G = (V, E), (\theta_\mathcal{E})_{\mathcal{E}\in E}]$, and $i \in V$, the model $M|_{X_i=a}$ obtained by clamping the variable X_i to the value $a \in \mathbb{X}_i$ is given by: the hypergraph $(V \setminus \{i\}, E_i)$, where $E_i = \{\mathcal{E} \setminus \{i\} | \mathcal{E} \in E\}$; and potentials which are unchanged for hyperedges which do not contain i, while if $i \in \mathcal{E}$ then $\theta_{\mathcal{E}\setminus\{i\}}(x_{\mathcal{E}\setminus\{i\}}) = \theta_\mathcal{E}(x_{\mathcal{E}\setminus\{i\}}, x_i = a)$.

Definition 2 (**Uprooting, suspension hypergraph**). Given a model $M[G(V, E), (\theta_\mathcal{E})_{\mathcal{E}\in E}]$, the *uprooted* model M^+ adds a variable X_0, which is added to every hyperedge of the original model. M^+ has hypergraph ∇G, with vertex set $V^+ = V \cup \{0\}$ and hyperedge set $E^+ = \{\mathcal{E}^+ = \mathcal{E} \cup \{0\} | \mathcal{E} \in E\}$. ∇G is the *suspension hypergraph* of G. M^+ has potential functions $(\theta^+_{\mathcal{E}\cup\{0\}})_{\mathcal{E}\in E}$ given by

$$\theta^+_{\mathcal{E}\cup\{0\}}(x_{\mathcal{E}\cup\{0\}}) = \begin{cases} \theta_\mathcal{E}(x_\mathcal{E}) & \text{if } x_0 = 0 \\ \theta_\mathcal{E}(\overline{x}_\mathcal{E}) & \text{if } x_0 = 1. \end{cases}$$

With this definition, all uprooted potentials are symmetric in that $\theta^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) = \theta^+_{\mathcal{E}^+}(\overline{x}_{\mathcal{E}^+})\ \forall \mathcal{E}^+ \in E^+$.

Definition 3 (**Rerooting**). From Definition 2, we see that given a model M, if we uproot to M^+ then clamp $X_0 = 0$, we recover the original model M. If instead in M^+ we clamp $X_i = 0$ for any $i = 1, \ldots, n$, then we obtain the *rerooted* model $M_i := M^+|_{X_i=0}$.

See Figure 1 and Table 1 for examples of uprooting and rerooting. We explore the question of how to choose a good variable for rerooting (i.e. how to choose a good variable to clamp in M^+) in §7.

4 Recovery of inference tasks

Here we demonstrate that the partition function, MAP score and configuration, and marginal distributions for a model M, can all be recovered from its uprooted model M^+ or any rerooted model M_i $i \in V$, with negligible computational cost. We write $V_i = \{0, 1, \ldots, n\} \setminus \{i\}$ for the variable set of rerooted model M_i; $\text{score}_i(x_{V_i})$ for the score of x_{V_i} in M_i; and p_i for the probability distribution for M_i. We use superscript $+$ to indicate the uprooted model. For example, the probability distribution for M^+ is given by $p^+(x_{V^+}) = \frac{1}{Z^+}\exp\left(\sum_{\mathcal{E}\in E^+}\theta_\mathcal{E}(x_\mathcal{E})\right)$. From the definitions of §3, we obtain the following key lemma, which is critical to enable recovery of inference results.

Lemma 4 (**Score-preserving map**). *Each configuration x_V of M maps to 2 configurations of the uprooted M^+ with the same score, i.e. from $M, x_V \rightarrow$ in M^+, both of $(x_0 = 0, x_V)$ and $(x_0 = 1, \overline{x}_V)$ with score(x_V) = score$^+(x_0 = 0, x_V)$ = score$^+(x_0 = 1, \overline{x}_V)$. For any $i \in V^+$, exactly one of the two uprooted configurations has $x_i = 0$, and just this one will be selected in M_i. Hence, there is a score-preserving bijection between configurations of M and those of M_i:*

$$\textit{For any } i \in V^+: \qquad \textit{in } M, x_V \quad \leftrightarrow \quad \textit{in } M_i, \begin{cases} (x_0 = 0, x_{V\setminus\{i\}}) & \textit{if } x_i = 0 \\ (x_0 = 1, \overline{x}_{V\setminus\{i\}}) & \textit{if } x_i = 1. \end{cases} \tag{2}$$

M config			M^+ configuration				M_4 config		
x_1	x_3	x_4	x_0	x_1	x_3	x_4	x_0	x_1	x_3
0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	1			
0	1	0	0	0	1	0	0	0	1
0	1	1	0	0	1	1			
1	0	0	0	1	0	0	0	1	0
1	0	1	0	1	0	1			
1	1	0	0	1	1	0	0	1	1
1	1	1	0	1	1	1			
			1	0	0	0	1	0	0
			1	0	0	1			
			1	0	1	0	1	0	1
			1	0	1	1			
			1	1	0	0	1	1	0
			1	1	0	1			
			1	1	1	0	1	1	1
			1	1	1	1			

Table 1: An illustration of how scores of potential θ_{134} on hyperedge $\{1, 3, 4\}$ in an original model M map to potential θ_{0134} in M^+ and then to θ_{013} in M_4. See Figure 1 for the hypergraphs. Each color indicates a value of $\theta_{134}(x_1, x_3, x_4)$ for a different configuration (x_1, x_3, x_4). Note that M^+ has 2 rows of each color, while after rerooting to M_4, we again have exactly one row of each color. The 1-1 score preserving map between configurations of M and any M_i is critical to enable recovery of inference results; see Lemma 4.

Table 1 illustrates this perhaps surprising result, from which the next two propositions follow.

Proposition 5 (**Recovering the partition function**). Given a model $M[G(V, E), (\theta_\mathcal{E})_{\mathcal{E} \in E}]$ with partition function Z as in (1), the partition function Z^+ of the uprooted model M^+ is twice Z, and the partition function of each rerooted model M_i is exactly Z, for any $i \in V$.

Proposition 6 (**Recovering a MAP configuration**). From M^+: x_V is an arg max for p iff $(x_0 = 0, x_V)$ is an arg max for p^+ iff $(x_0 = 1, \overline{x}_V)$ is an arg max for p^+. From a rerooted model M_i: $(x_{V\setminus\{i\}}, x_i = 0)$ is an arg max for p iff $(x_0 = 0, x_{V\setminus\{i\}})$ is an arg max for p_i; $(x_{V\setminus\{i\}}, x_i = 1)$ is an arg max for p iff $(x_0 = 1, \overline{x}_{V\setminus\{i\}})$ is an arg max for p_i.

We can recover marginals as shown in the following proposition, proof in the Appendix §9.1.

Proposition 7 (**Recovering marginals**). For a subset $\emptyset \neq U \subseteq V$, we can recover from M^+: $p(x_U) = p^+(x_0 = 0, x_U) + p^+(x_0 = 1, \overline{x}_U) \qquad = 2p^+(x_0 = 0, x_U) = 2p^+(x_0 = 1, \overline{x}_U)$.
To recover from a rerooted M_i: (i) For any $i \in V \setminus U$, $\;p(x_U) = p_i(x_0 = 0, x_U) + p_i(x_0 = 1, \overline{x}_U)$.
(ii) For any $i \in U$, $p(x_U) = \begin{cases} p_i(x_0 = 0, x_{U\setminus\{i\}}) & x_i = 0 \\ p_i(x_0 = 1, \overline{x}_{U\setminus\{i\}}) & x_i = 1. \end{cases}$

In §6, we provide a careful analysis of the impact of uprooting and rerooting on the Sherali-Adams hierarchy of relaxations of the marginal polytope [11]. We first introduce a way to parametrize potentials which will be particularly useful, and which may be of independent interest.

5 Pure k-potentials

We introduce the notion of *pure k-potentials*. These allow the specification of interactions which act 'purely' over a set of variables of a given size k, without influencing the distribution of any subsets. We show that in fact, there is essentially only one pure k-potential. Further, we show that one can express any $\theta_\mathcal{E}$ potential in terms of pure potentials over $\mathcal{E}$ and subsets of $\mathcal{E}$, and that pure potentials have appealing properties when uprooted and rerooted which help our subsequent analysis.

We say that a potential is a *k-potential* if k is the smallest number such that the score of the potential may be determined by considering the configuration of k variables. Usually a potential $\theta_\mathcal{E}$ is a k-potential with $k = |\mathcal{E}|$. For example, typically a singleton potential is a 1-potential, and an edge potential is a 2-potential. However, note that $k < |\mathcal{E}|$ is possible if one or more variables in $\mathcal{E}$ are not needed to establish the score (a simple example is $\theta_{12}(x_1, x_2) = x_1$, which clearly is a 1-potential).

In general, a k-potential will affect the marginal distributions of all subsets of the k variables. For example, one popular form of 2-potential is $\theta_{ij}(x_i, x_j) = W_{ij}x_ix_j$, which tends to pull X_i and X_j toward the same value, but also tends to increase each of $p(X_i = 1)$ and $p(X_j = 1)$. For pairwise models, a different reparameterization of potentials instead writes the score as

$$\text{score}(x_V) = \sum_{i \in V} \theta_i x_i + \frac{1}{2} \sum_{(i,j) \in E} W_{ij} \mathbb{1}[x_i = x_j]. \tag{3}$$

Expression (3) has the desirable feature that the $\theta_{ij}(x_i, x_j) = \frac{1}{2}W_{ij}\mathbb{1}[x_i = x_j]$ edge potentials affect only the pairwise marginals, without disturbing singleton marginals. This motivates the following definition.

Definition 8. Let $k \geq 2$, and let U be a set of size k. We say that a k-potential $\theta_U : \{0,1\}^U \to \mathbb{R}$ is a *pure k-potential* if the distribution induced by the potential, $p(x_U) \propto \exp(\theta_U(x_U))$, has the property that for any $\emptyset \neq W \subsetneq U$, the marginal distribution $p(x_W)$ is uniform.

We shall see in Proposition 10 that a pure k-potential must essentially be an *even k-potential*.

Definition 9. Let $k \in \mathbb{N}$, and $|U| = k$. An *even k-potential* is a k-potential $\theta_U : \{0,1\}^U \to \mathbb{R}$ of the form $\theta_U(x_U) = a\mathbb{1}[\ |\{i \in U | x_i = 1\}| \text{ is even}]$, for some $a \in \mathbb{R}$ which is its *coefficient*. In words, $\theta_U(x_U)$ takes value a if x_U has an even number of 1s, else it takes value 0.

As an example, the 2-potential $\theta_{ij}(x_i, x_j) = \frac{1}{2}W_{ij}\mathbb{1}[x_i = x_j]$ in (3) is an even 2-potential with $U = \{i, j\}$ and coefficient $W_{ij}/2$. The next two propositions are proved in the Appendix §9.2.

Proposition 10 (All pure potentials are essentially even potentials)**.** Let $k \geq 2$, and $|U| = k$. If $\theta_U : \{0,1\}^U \to \mathbb{R}$ is a pure k-potential then θ_U must be an affine function of the even k-potential, i.e. $\exists\, a, b \in \mathbb{R}$ s.t. $\theta_U(x_U) = a\mathbb{1}[\ |\{i \in U | x_i = 1\}| \text{ is even}] + b$.

Proposition 11 (Even k-potentials form a basis)**.** For a finite set U, the set of even k-potentials $\left(\mathbb{1}[\ |\{i \in W | X_i = 1\}| \text{ is even}]\right)_{W \subseteq U}$, indexed by subsets $W \subseteq U$, forms a basis for the vector space of all potential functions $\theta : \{0,1\}^U \to \mathbb{R}$.

Any constant in a potential will be absorbed into the partition function Z and does not affect the probability distribution, see (1). An even 2-potential with positive coefficient, e.g. as in (3) if $W_{ij} > 0$, is *supermodular*. Models with only supermodular potentials (equivalently, submodular cost functions) typically admit easier inference [3; 7]; if such a model is binary pairwise then it is called *attractive*. However, for $k > 2$, even k-potentials $\theta_\mathcal{E}$ are neither supermodular nor submodular. Yet if k is an even number, observe that $\theta_\mathcal{E}(x_\mathcal{E}) = \theta_\mathcal{E}(\overline{x}_\mathcal{E})$. We discuss this further in Appendix §10.4.

When a k-potential is uprooted, in general it may become a $(k+1)$-potential (recall Definition 2). The following property of even k-potentials is helpful for our analysis in §6, and is easily checked.

Lemma 12 (Uprooting an even k-potential)**.** *When an even k-potential $\theta_\mathcal{E}$ with $|\mathcal{E}| = k$ is uprooted: if k is an even number, then the uprooted potential is exactly the same even k-potential; if k is odd, then we obtain the even $(k+1)$-potential over $\mathcal{E} \cup \{0\}$ with the same coefficient as the original $\theta_\mathcal{E}$.*

6 Marginal polytope and Sherali-Adams relaxations

We saw in Lemma 4 that there is a score-preserving 1-2 mapping from configurations of M to those of M^+, and a bijection between configurations of M and any M_i. Here we examine the extent to which these score-preserving mappings extend to (pseudo-)marginal probability distributions over variables by considering the Sherali-Adams relaxations [11] of the respective marginal polytopes. These relaxations feature prominently in many approaches for MAP and marginal inference.

For $U \subseteq V$, we write μ_U for a probability distribution in $\mathscr{P}(\{0,1\}^U)$, the set of all probability distributions on $\{0,1\}^U$. Bold $\boldsymbol{\mu}$ will represent a collection of measures over various subsets of variables. Given (1), to compute an expected score, we need $(\mu_\mathcal{E})_{\mathcal{E} \in E}$. This motivates the following.

Definition 13. The *marginal polytope* $\mathbb{M}(G(V,E)) = \{(\mu_\mathcal{E})_{\mathcal{E} \in E} \big| \exists \mu_V \text{ s.t. } \mu_{V\downarrow\mathcal{E}} = \mu_\mathcal{E}\ \forall \mathcal{E} \in E\}$, where for $U_1 \subseteq U_2 \subseteq V$, $\mu_{U_2 \downarrow U_1}$ denotes the marginalization of $\mu_{U_2} \in \mathscr{P}(\{0,1\}^{U_2})$ onto $\{0,1\}^{U_1}$.

$\mathbb{M}(G)$ consists of marginal distributions for every hyperedge $\mathcal{E} \in E$ such that all the marginals are consistent with a global distribution over all variables V. Methods of variational inference typically optimize either the score (for MAP inference) or the score plus an entropy term (for marginal inference) over a relaxation of the marginal polytope [15]. This is because $\mathbb{M}(G)$ is computationally intractable, with an exponential number of facets [2]. Relaxations from the Sherali-Adams hierarchy [11] are often used, requiring consistency only over smaller clusters of variables.

Definition 14. Given an integer $r \geq 2$, if a hypergraph $G(V, E)$ satisfies $\max_{\mathcal{E}\in E} |\mathcal{E}| \leq r \leq |V|$, then we say that G is *r-admissible*, and define the *Sherali-Adams polytope of order r* on G by

$$\mathbb{L}_r(G) = \left\{ (\mu_{\mathcal{E}})_{\mathcal{E}\in E} \,\middle|\, \exists (\mu_U)_{\substack{U\subseteq V\\|U|=r}} \text{ locally consistent, s.t. } \mu_{U\downarrow\mathcal{E}} = \mu_{\mathcal{E}} \quad \forall\, \mathcal{E} \subseteq U \subseteq V,\ \ |U| = r \right\},$$

where a collection of measures $(\mu_A)_{A\in I}$ (for some set I of subsets of V) is *locally consistent*, or l.c., if for any $A_1, A_2 \in I$, we have $\mu_{A_1\downarrow A_1\cap A_2} = \mu_{A_2\downarrow A_1\cap A_2}$. Each element of $\mathbb{L}_r(G)$ is a set of locally consistent probability measures over the hyperedges. Note that $\mathbb{M}(G) \subseteq \mathbb{L}_r(G) \subseteq \mathbb{L}_{r-1}(G)$. The pairwise relaxation $\mathbb{L}_2(G)$ is commonly used but higher-order relaxations achieve greater accuracy, have received significant attention [10; 13; 18; 22; 23], and are required for higher-order potentials.

6.1 The impact of uprooting and rerooting on Sherali-Adams polytopes

We introduce two variants of the Sherali-Adams polytopes which will be helpful in analyzing uprooted models. For a measure $\mu_U \in \mathscr{P}(\{0,1\}^U)$, we define the *flipped measure* $\overline{\mu}_U$ as $\overline{\mu}_U(x_U) = \mu_U(\overline{x}_U)\ \forall x_U \in \{0,1\}^U$. A measure μ_U is *flipping-invariant* if $\mu_U = \overline{\mu}_U$.

Definition 15. The symmetrized Sherali-Adams polytopes for an uprooted hypergraph $\nabla G(V^+, E^+)$ (as given in Definition 2), is:

$$\widetilde{\mathbb{L}}_r(\nabla G) = \left\{ (\mu_{\mathcal{E}})_{\mathcal{E}\in E^+} \in \mathbb{L}_r(\nabla G) \,\middle|\, \mu_{\mathcal{E}} = \overline{\mu}_{\mathcal{E}}\ \forall \mathcal{E} \in E^+ \right\}.$$

Definition 16. For any $i \in V^+$, and any integer $r \geq 2$ such that $\max_{\mathcal{E}\in E^+} |\mathcal{E}| \leq r \leq |V^+|$, we define the symmetrized Sherali-Adams polytope of order r *uprooted at i* to be

$$\widetilde{\mathbb{L}}^i_r(\nabla G) = \left\{ (\mu_{\mathcal{E}})_{\mathcal{E}\in E^+} \,\middle|\, \exists (\mu_U)_{\substack{i\in U\subseteq V^+\\|U|=r}} \text{ l.c., s.t. } \begin{array}{ll} \mu_{U\downarrow\mathcal{E}} = \mu_{\mathcal{E}} & \forall\, \mathcal{E} \subseteq U \subseteq V,\ |U| = r, i \in U \\ \mu_U = \overline{\mu}_U & \forall U \subseteq V, |U| = r, i \in U \end{array} \right\}.$$

Thus, for each collection of measures over hyperedges in $\widetilde{\mathbb{L}}^i_r(\nabla G)$, there exist corresponding flipping-invariant, locally consistent measures on sets of size r which contain i (and their subsets). Note that for any hypergraph $G(V, E)$ and any $i \in V^+$, we have $\widetilde{\mathbb{L}}_{r+1}(\nabla G) \subseteq \widetilde{\mathbb{L}}^i_{r+1}(\nabla G) \subseteq \widetilde{\mathbb{L}}_r(\nabla G)$. We next extend the correspondence of Lemma 4 to collections of locally-consistent probability distributions on the hyperedges of G, see the Appendix §9.3 for proof.

Theorem 17. *For a hypergraph $G(V, E)$, and integer r such that $\max_{\mathcal{E}\in E} |\mathcal{E}| \leq r \leq |V|$, there is an affine score-preserving bijection*

$$\mathbb{L}_r(G) \underset{\textbf{RootAt0}}{\overset{\textbf{Uproot}}{\rightleftarrows}} \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)\,.$$

Theorem 17 establishes the following diagram of polytope inclusions and affine bijections:

$$\begin{array}{lccccc} \text{For } M = M_0: & \mathbb{L}_{r+1}(G) & \subseteq & \text{Unnamed} & \subseteq & \mathbb{L}_r(G) \\ & \scriptstyle\text{Uproot}\downarrow\uparrow\text{RootAt0} & & \scriptstyle\text{Uproot}\downarrow\uparrow\text{RootAt0} & & \scriptstyle\text{Uproot}\downarrow\uparrow\text{RootAt0} \\ \text{For } M^+: & \widetilde{\mathbb{L}}^0_{r+2}(\nabla G) & \subseteq & \widetilde{\mathbb{L}}_{r+1}(\nabla G) & \subseteq & \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)\,. \end{array} \tag{4}$$

A question of theoretical interest and practical importance is which of the inclusions in (4) are strict. Our perspective here generalizes earlier work. Using different language, Deza and Laurent [2] identified $\mathbb{L}_2(G)$ with $\widetilde{\mathbb{L}}^0_3(\nabla G)$, which was termed RMET, the *rooted semimetric polytope*; and $\widetilde{\mathbb{L}}_3(\nabla G)$ with MET, the *semimetric polytope*. Building on this, Weller [19] considered $\mathbb{L}_3(G)$, the triplet-consistent polytope or TRI, though only in the context of pairwise potentials, and showed that $\mathbb{L}_3(G)$ has the remarkable property that if it is used to optimize an LP for a model M on G, the exact same optimum is achieved for $\mathbb{L}_3(G_i)$ for any rerooting M_i. It was natural to conjecture that $\mathbb{L}_r(G)$ might have this same property for all $r > 3$, yet this was left as an open question.

6.2 $\mathbb{L}_3$ is unique in being universally rooted

We shall first strengthen [19] to show that $\mathbb{L}_3$ is *universally rooted* in the following stronger sense.

Definition 18. We say that the r^{th}-order Sherali-Adams relaxation is *universally rooted* (and write "$\mathbb{L}_r$ is universally rooted" for short) if for all admissible hypergraphs G, there is an affine score-preserving bijection between $\mathbb{L}_r(G)$ and $\mathbb{L}_r(G_i)$, for each rerooted hypergraph $(G_i)_{i\in V}$.

If $\mathbb{L}_r$ is universally rooted, this applies for potentials over up to r variables (the maximum which makes sense in this context), and clearly it implies that optimizing score over any rerooting (as in MAP inference) will attain the same objective. The following result is proved in the Appendix §9.3.

Lemma 19. *If $\mathbb{L}_r$ is universally rooted for hypergraphs of maximum hyperedge degree $p < r$ with p even, then $\mathbb{L}_r$ is also universally rooted for r-admissible hypergraphs with maximum degree $p + 1$.*

The proof relies on mapping to the symmetrized uprooted polytope $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G)$. Then by considering marginals using a basis equivalent to that described in Proposition 11 for even k-potentials, we observe that the symmetry of the polytope enforces only one possible marginal for $(p + 1)$-clusters.

Combining Lemma 19 with arguments which extend those used by [19] demonstrates the following result, proved in the Appendix.

Theorem 20. $\mathbb{L}_3$ *is universally rooted.*

We next provide a striking and rather surprising result, see the Appendix for proof and details.

Theorem 21. $\mathbb{L}_3$ *is unique in being universally rooted. Specifically, for any integer $r > 1$ other than $r = 3$, we constructively demonstrate a hypergraph $G(V, E)$ with $|V| = r + 1$ variables for which $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G) \neq \widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$ for any $i \in V$.*

Theorem 21 examines $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G)$ and $\widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$, which by Theorem 17 are the uprooted equivalents of $\mathbb{L}_r(G)$ and $\mathbb{L}_r(G_i)$. It might appear more satisfying to try to demonstrate the result directly for the rooted polytopes, i.e. to show $\mathbb{L}_r(G) \neq \mathbb{L}_r(G_i)$. However, in general the rooted polytopes are not comparable: an r-potential in M can map to an $(r + 1)$-potential in M^+ and then to an $(r + 1)$-potential in M_i which cannot be evaluated for an $\mathbb{L}_r$ polytope.

Theorem 21 shows that we may hope for benefits from rerooting for any inference method based on a Sherali-Adams relaxed polytope $\mathbb{L}_r$, unless $r = 3$.

7 Experiments

Here we show empirically the benefits of uprooting and rerooting for approximate inference methods in models with higher-order potentials. We introduce an efficient heuristic which can be used in practice to select a variable for rerooting, and demonstrate its effectiveness.

We compared performance after different rerootings of marginal inference (to guarantee convergence we used the double loop method of Heskes et al. [4], which relates to generalized belief propagation, 24) and MAP inference (using loopy belief propagation, LBP [9]). For true values, we used the junction tree algorithm. All methods were implemented using libDAI [8]. We ran experiments on complete hypergraphs (with 8 variables) and toroidal grid models (5×5 variables). Potentials up to order 4 were selected randomly, by drawing even k-potentials from $\text{Unif}([-W_{\text{max}}, W_{\text{max}}])$ distributions for a variety of W_{max} parameters, as shown in Figure 2, which highlights results for estimating $\log Z$. For each regime of maximum potential values, we plot results averaged over 20 runs. For additional details and results, including marginals, other potential choices and larger models, see Appendix §10.

We display average error of the inference method applied to: the original model M; the uprooted model M^+; then rerootings at: the *worst* variable, the *best* variable, the *K heuristic* variable, and the *G heuristic* 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 other measures).

7.1 Heuristics to pick a good variable for rerooting

From our Definition 3, a rerooted model M_i is obtained by clamping the uprooted model M^+ at variable X_i. Hence, selecting a good variable for rerooting is exactly the choice of a good variable to clamp in M^+. Considering pairwise models, Weller [19] refined the *maxW* method [20; 21] to introduce the *maxtW* heuristic, and showed that it was very effective empirically. *maxtW* selects the variable X_i with $\max \sum_{j\in\mathcal{N}(i)} \tanh |\frac{W_{ij}}{4}|$, where $\mathcal{N}(i)$ is the set of neighbors of i in the model graph, and W_{ij} is the strength of the pairwise interaction.

The intuition for *maxtW* is as follows. Pairwise methods of approximate inference such as Bethe are exact for models with no cycles. If we could, we would like to 'break' tight cycles with strong edge weights, since these lead to error. When a variable is clamped, it is effectively removed from the model. Hence, we would like to reroot at a variable that sits on many cycles with strong edge weights. Identifying such cycles is NP-hard, but the *maxtW* heuristic attempts to do this by looking only locally around each variable. Further, the effect of a strong edge weight saturates [21]: a very strong edge weight W_{ij} effectively 'locks' its end variables (either together or opposite depending on the sign of W_{ij}), and this effect cannot be significantly increased even by an extremely strong edge. Hence the $\tanh$ function was introduced to the earlier *maxW* method, leading to the *maxtW* heuristic.

As observed in §5, if we express our model potentials in terms of pure k-potentials, then the uprooted model will only have pure k-potentials for various values of k which are even numbers. Intuitively, the higher the coefficients on these potentials, the more tightly connected is the model leading to more challenging inference. Hence, a natural way to generalize the *maxtW* approach to handle higher-order potentials is to pick a variable X_i in M^+ which maximizes the following measure:

$$\text{clamp-heuristic-measure}(i) = \sum_{i\in\mathcal{E}:|\mathcal{E}|=2} c_2 \tanh |t_2 a_\mathcal{E}| + \sum_{i\in\mathcal{E}:|\mathcal{E}|=4} c_4 \tanh |t_4 a_\mathcal{E}|, \tag{5}$$

where $a_\mathcal{E}$ is the coefficient (weight) of the relevant pure k-potential, see Definition 9, and the $\{c_2, t_2\}, \{c_4, t_4\}$ terms are constants for pure 2-potentials and for pure 4-potentials respectively. This approach extends to potentials of higher orders by adding similar further terms. Since our goal is to rank the measures for each $i \in V^+$, without loss of generality we take $c_2 = 1$. We fit the t_2, c_4 and t_4 constants to the data from our experimental runs, see the Appendix for details. Our *K heuristic* was fit only to runs for complete hypergraphs while the *G heuristic* was fit only to runs for models on grids.

7.2 Observations on results

Considering all results across models and approximate methods for estimating $\log Z$, marginals and MAP inference (see Figure 2 and Appendix §10.3), we make the following observations. Both K and G heuristics perform well (in and out of sample): they never hurt materially and often significantly improve accuracy, attaining results close to the best possible rerooting. Since our two heuristics achieve similar performance, sensitivity to the exact constants in (5) appears low. We verified this by comparing to maxtW for pairwise models as in [19]: both K and G heuristics performed just slightly better than maxtW. For all our runs, inference on rerooted models took similar time as on the original model (time required to reroot and later to map back inference results is negligible), see §10.3.1.

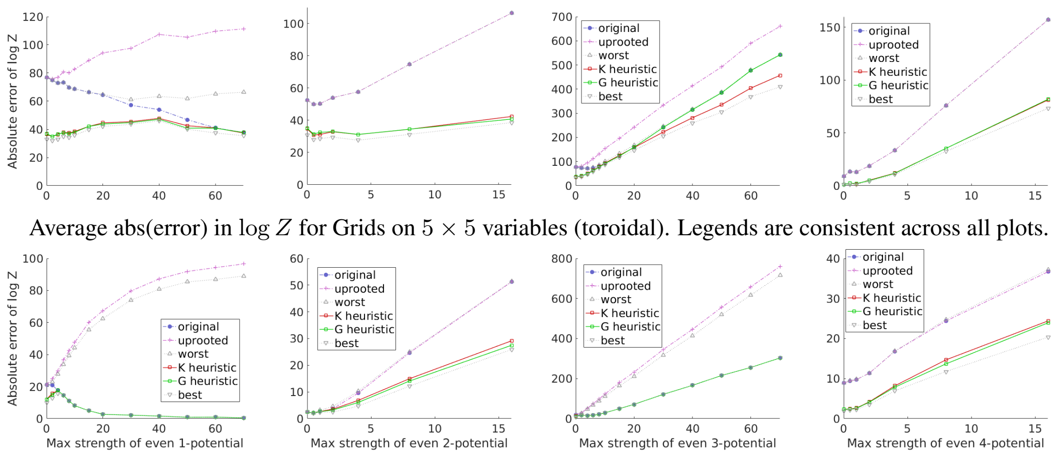
Observe that stronger 1-potentials tend to make inference *easier*, pulling each variable toward a specific setting, and reducing the benefits from rerooting (left column of Figure 2). Stronger pure k-potentials for $k > 1$ intertwine variables more tightly: this typically makes inference harder and increases the gains in accuracy from rerooting. The pure k-potential perspective facilitates this analysis.

When we examine larger models, or models with still higher order potentials, we observe qualitatively similar results, see Appendix §10.3.4 and 10.3.6.

8 Conclusion

We introduced methods which broaden the application of the uprooting and rerooting approach to binary models with higher-order potentials of any order. We demonstrated several important theoretical insights, including Theorems 20 and 21 which show that $\mathbb{L}_3$ is unique in being universally rooted. We developed the helpful tool of even k-potentials in §5, which may be of independent

Average abs(error) in $\log Z$ for K_8 complete hypergraphs (fully connected) on 8 variables.



vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

Figure 2: Error in estimating $\log Z$ for random models with various pure k-potentials over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. Where the red K heuristic curve is not visible, it coincides with the green G heuristic. Both K and G heuristics for selecting a rerooting work well: they never hurt and often yield large benefits. See §7 for details.

interest. We empirically demonstrated significant benefits for rerooting in higher-order models – particularly for the hard case of strong cluster potentials and weak 1-potentials – and provided an efficient heuristic to select a variable for rerooting. This heuristic is also useful to indicate when rerooting is unlikely to be helpful for a given model (if (5) is maximized by taking $i = 0$).

It is natural to compare the effect of rerooting M to M_i, against simply clamping X_i in the original model M. A key difference is that rerooting achieves the clamping at X_i for negligible computational cost. In contrast, if X_i is clamped in the original model then the inference method will have to be run twice: once clamping $X_i = 0$, and once clamping $X_i = 1$, then results must be combined. This is avoided with rerooting given the symmetry of M^+. Rerooting effectively replaces what may be a poor initial implicit choice of clamping at X_0 with a carefully selected choice of clamping variable almost for free. This is true even for large models where it may be advantageous to clamp a series of variables: by rerooting, one of the series is obtained for free, potentially gaining significant benefit with little work required. Note that each separate connected component may be handled independently, with its own added variable. This could be useful for (repeatedly) composing clamping and then rerooting each separated component to obtain an almost free clamping in each.

Acknowledgements

We thank the anonymous reviewers for helpful comments. 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. AW acknowledges support by the Alan Turing Institute under the EPSRC grant EP/N510129/1, and by the Leverhulme Trust via the CFI.

References

[1] 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.

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

[3] J. Djolonga and A. Krause. Scalable variational inference in log-supermodular models. In *ICML*, pages 1804–1813, 2015.

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

[5] A. Jaimovich, G. Elidan, H. Margalit, and N. Friedman. Towards an integrated protein–protein interaction network: A relational Markov network approach. *Journal of Computational Biology*, 13(2):145–164, 2006.

[6] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[7] 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.

[8] 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`.

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

[10] M. Rowland, A. Pacchiano, and A. Weller. Conditions beyond treewidth for tightness of higher-order LP relaxations. In *Artifical Intelligence and Statistics (AISTATS)*, 2017.

[11] 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.

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

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

[14] M. Wainwright and M. Jordan. Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing*, 2006.

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

[16] 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.

[17] A. Weller. Revisiting the limits of MAP inference by MWSS on perfect graphs. In *AISTATS*, 2015.

[18] A. Weller. Characterizing tightness of LP relaxations by forbidding signed minors. In *UAI*, 2016.

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

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

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

[22] 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.

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

[24] J. Yedidia, W. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Information Theory*, pages 2282–2312, 2005.

APPENDIX: Uprooting and Rerooting Higher-Order Graphical Models

In this Appendix, we provide:

- In §9, proofs of results appearing in the main paper, split into:
  - §9.1 Proofs of results from §4: Recovery of inference tasks
  - §9.2 Proofs of results from §5: Even k-potentials
  - §9.3 Proofs of results from §6: Sherali-Adams relaxations.
- In §10, additional experimental details and results.

Notation. A model $M[G(V,E),(\theta_\mathcal{E})_{\mathcal{E}\in E}]$ uproots to $M^+[G^+(V^+,E^+,(\theta_{\mathcal{E}^+})_{\mathcal{E}^+\in E^+}]$, where $G^+ = \nabla G$. Given a model M with hyperedges $\mathcal{E} \in E$ and potentials $(\theta_\mathcal{E})_{\mathcal{E}\in E}$, we adopt the convention that in the uprooted model M^+, each $\mathcal{E}^+ = \mathcal{E} \cup \{0\}$ and each $\theta_{\mathcal{E}^+}$ is the uprooted version of the respective $\theta_\mathcal{E}$, as given in Definition 2.

For a set S, we write $\#S = |S|$ for its cardinality. For example, $\#\{1,2,3\} = 3$.

9 Proofs of results from the main paper

9.1 Proofs of results from §4: Recovery of inference tasks

Proposition 5 (Recovering the partition function) Given a model $M[G(V,E),(\theta_\mathcal{E})_{\mathcal{E}\in E}]$ with partition function Z as in (1), the partition function Z^+ of the uprooted model M^+ is twice Z, and the partition function of each rerooted model M_i is exactly Z, for any $i \in V$.

Proof. Recall that for the model M, we have

$$Z = \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}\in E}\theta_\mathcal{E}(x_\mathcal{E})\right) .$$

Writing Z^+ for the partition function of M^+, by definition we have

$$\begin{aligned}
Z^+ &= \sum_{x_{V\cup\{0\}}\in\{0,1\}^{V\cup\{0\}}} \exp\left(\sum_{\mathcal{E}^+\in E^+}\theta_{\mathcal{E}^+}(x_{\mathcal{E}\cup\{0\}})\right)\\
&= \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}^+\in E^+}\theta_{\mathcal{E}^+}(x_0=0,x_\mathcal{E})\right) + \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}^+\in E^+}\theta_{\mathcal{E}^+}(x_0=1,x_\mathcal{E})\right)\\
&= \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}^+\in E^+}\theta_{\mathcal{E}^+}(x_0=0,x_\mathcal{E})\right) + \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}^+\in E^+}\theta_{\mathcal{E}^+}(x_0=1,x_\mathcal{E})\right)\\
&= \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}\in E}\theta_\mathcal{E}(x_\mathcal{E})\right) + \sum_{x_V\in\{0,1\}^V} \exp\left(\sum_{\mathcal{E}\in E}\theta_\mathcal{E}(\overline{x}_\mathcal{E})\right)\\
&= 2Z\,,
\end{aligned}$$

as required. Now, given $i \in V$, and noting that M^+ is also the uprooting of the model M_i, it immediately follows from the above that the partition function associated with M_i is Z, as required. $\square$

Proposition 6 (Recovering MAP configurations) From M^+: x_V is an arg max for p iff $(x_0 = 0, x_V)$ is an arg max for p^+ iff $(x_0 = 1, \overline{x}_V)$ is an arg max for p^+. From a rerooted model M_i: $(x_{V\setminus\{i\}}, x_i = 0)$ is an arg max for p iff $(x_0 = 0, x_{V\setminus\{i\}})$ is an arg max for p_i; $(x_{V\setminus\{i\}}, x_i = 1)$ is an arg max for p iff $(x_0 = 1, \overline{x}_{V\setminus\{i\}})$ is an arg max for p_i.

Proof. From M^+: we simply note that by construction of the uprooted potentials, for any $x_V \in \{0,1\}^V$ we have

$$\sum_{E\in E}\theta_E(x_E) = \sum_{E\in E}\theta^+_{E^+}(x_E, x_0=0) = \sum_{E\in E}\theta^+_{E^+}(\overline{x}_E, x_0=1),$$

from which the claim immediately follows.

From M_i: we have

$$p_i(x_{V\setminus\{i\}}, x_0) \propto p^+(x_{V\setminus\{i\}}, x_0, x_i = 0)\,,$$

which implies that

$$\begin{aligned}(x_{V\cup\{0\}\setminus\{i\}}) \in \arg\max p_i &\iff (x_{V\cup\{0\}\setminus\{i\}}, x_i = 0) \in \arg\max p^+ \\ &\iff (\overline{x}_{V\cup\{0\}\setminus\{i\}}, x_i = 1) \in \arg\max p^+ \\ &\iff \begin{cases}(x_{V\setminus\{i\}}, x_i = 0) \in \arg\max p & \text{if } x_0 = 0 \\ (\overline{x}_{V\setminus\{i\}}, x_i = 1) \in \arg\max p & \text{if } x_0 = 1.\end{cases}\end{aligned}$$

□

Proposition 7 (Recovering marginals) For a subset $\emptyset \neq U \subseteq V$, we can recover from M^+:

$$p(x_U) = p^+(x_0 = 0, x_U) + p^+(x_0 = 1, \overline{x}_U) \qquad = 2p^+(x_0 = 0, x_U) = 2p^+(x_0 = 1, \overline{x}_U).$$

To recover from a rerooted M_i: (i) For any $i \in V \setminus U$, $p(x_U) = p_i(x_0 = 0, x_U) + p_i(x_0 = 1, \overline{x}_U)$.
(ii) For any $i \in U$, $p(x_U) = \begin{cases}p_i(x_0 = 0, x_{U\setminus\{i\}}) & x_i = 0 \\ p_i(x_0 = 1, \overline{x}_{U\setminus\{i\}}) & x_i = 1.\end{cases}$

Proof. Let $x_U \in \{0,1\}^U$. Observe that

$$\begin{aligned}p(x_U) &= \frac{1}{Z}\sum_{x_{V\setminus U}} \exp\left(\sum_{\mathcal{E}\in E} \theta_{\mathcal{E}}(x_{\mathcal{E}})\right) \\ &= \frac{1}{Z}\sum_{x_{V\setminus U}} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0 = 0, x_{\mathcal{E}})\right) \\ &= \frac{1}{2Z}\left(\sum_{x_{V\setminus U}} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0 = 0, x_{\mathcal{E}})\right) + \sum_{x_{V\setminus U}} \exp\left(\sum_{\mathcal{E}^+\in E^+} \theta_{\mathcal{E}^+}(x_0 = 1, \overline{x}_{\mathcal{E}})\right)\right) \\ &= p^+(x_0 = 0, x_U) + p^+(x_0 = 1, \overline{x}_U) = 2p^+(x_0 = 0, x_U) = 2p^+(x_0 = 1, \overline{x}_U)\,.\end{aligned}$$

We next demonstrate recovery from a rerooted model M_i. Let $V_i = V \cup \{0\} \setminus \{i\}$. By the definition of rerooting and symmetry of M^+, $p_i(x_{V_i}) = p^+(x_{V_i}|x_i = 0) = p^+(\overline{x}_{V_i}|x_i = 1)$. Further, $p^+(x_i = 0) = p^+(x_i = 1) = \frac{1}{2}$ for any $i = 0, 1, \ldots, n$.

Case (i) $i \in V \setminus U$. Following the argument above, we obtain

$$\begin{aligned}p(x_U) &= p^+(x_0 = 0, x_U) + p^+(x_0 = 1, \overline{x_U}) \\ &= p^+(x_0 = 0, x_U, x_i = 0) + p^+(x_0 = 0, x_U, x_i = 1) \\ &\qquad + p^+(x_0 = 1, \overline{x_U}, x_i = 0) + p^+(x_0 = 1, \overline{x_U}, x_i = 1) \\ &= \frac{1}{2}\left[p^+(x_0 = 0, x_U|x_i = 0) + p^+(x_0 = 0, x_U|x_i = 1)\right] \\ &\qquad + \frac{1}{2}\left[p^+(x_0 = 1, \overline{x}_U|x_i = 0) + p^+(x_0 = 1, \overline{x}_U|x_i = 1)\right] \\ &= \frac{1}{2}\left[p^+(x_0 = 0, x_U|x_i = 0) + p^+(x_0 = 1, \overline{x}_U|x_i = 1)\right] \\ &\qquad + \frac{1}{2}\left[p^+(x_0 = 0, x_U|x_i = 1) + p^+(x_0 = 1, \overline{x}_U|x_i = 0)\right] \\ &= p_i(x_0 = 0, x_U) + p_i(x_0 = 1, \overline{x}_U).\end{aligned}$$

Case (ii) $i \in U$. Now we have

$$\begin{aligned}p(x_U) &= p^+(x_0 = 0, x_U) + p^+(x_0 = 1, \overline{x}_U) \\ &= \frac{1}{2}\left[p^+(x_0 = 0, x_U|x_i = 0) + p^+(x_0 = 0, x_U|x_i = 1)\right]\end{aligned}$$

$$+ \frac{1}{2}\left[p^{+}(x_0 = 1, \overline{x}_U | x_i = 0) + p^{+}(x_0 = 1, \overline{x}_U | x_i = 1)\right]$$
$$= \begin{cases} p_i(x_0 = 0, x_{U\setminus\{i\}}) & \text{if } x_i = 0 \\ p_i(x_0 = 1, \overline{x}_{U\setminus\{i\}}) & \text{if } x_i = 1. \end{cases} \qquad \square$$

9.2 Proofs of results from §5: Even k-potentials

Proposition 10 (All pure potentials are essentially even potentials) Let $k \geq 2$, and $|U| = k$. If $\theta_U : \{0,1\}^U \to \mathbb{R}$ is a pure k-potential then θ_U must be an affine function of the even k-potential, i.e. $\exists\, a, b \in \mathbb{R}$ s.t. $\theta_U(x_U) = a\mathbb{1}[\,|\{i \in U | x_i = 1\}| \text{ is even}] + b$.

Proof. It is sufficient to demonstrate that if, for two configurations $x_U, y_U \in \{0,1\}^U$, we have $\sum_{i\in U} x_i = \sum_{i\in U} y_i \bmod 2$, then $\theta_U(x_U) = \theta_U(y_U)$, since this demonstrates that θ_U depends on its input argument only through the quantity $\mathbb{1}_{\#\{i\in U|x_i=1\} \text{ is even}}$, and since this only takes on two possible values, θ_U may be expressed as an affine function of this indicator.

To demonstrate the claim above, it is sufficient to show that if $x_U \in \{0,1\}^U$, and $i, j \in V$ are two distinct indices, and $F_{ij}(x_U) \in \{0,1\}^U$ denotes the configuration obtained from x_U by flipping coordinates i and j, then $\theta_U(x_U) = \theta_U(F_{ij}(x_U))$. This is sufficient since given $x_U, y_U \in \{0,1\}^U$ with $\sum_{i\in U} x_i = \sum_{i\in U} y_i \bmod 2$, it is possible to obtain y_U from x_U by iteratively flipping pairs of distinct variables.

Let $F_i(x_U)$ denote the configuration obtained from x_U by flipping x_i. By the uniform marginalization property, we have

$$p(x_U) + p(F_i(x_U)) = p(F_j(x_U)) + p(F_{ij}(x_U))$$

and

$$p(F_i(x_U)) + p(F_{ij}(x_U)) = p(x_U) + p(F_j(x_U))\,.$$

Subtracting these equations from one another yields

$$p(x_U) = p(F_{ij}(x_U))\,.$$

Taking logarithms of this equations yields $\theta_U(x_U) = \theta_U(F_{ij}(x_U))$, as required. $\square$

Proposition 11 (Even k-potentials form a basis) For a finite set U, the set of even k-potentials $\left(\mathbb{1}[\,|\{i \in W | X_i = 1\}| \text{ is even}]\right)_{W\subseteq U}$, indexed by subsets $W \subseteq U$, forms a basis for the vector space of all potential functions $\theta : \{0,1\}^U \to \mathbb{R}$.

Proof. We show that the indicators $(\mathbb{1}[\#\{i \in W|x_i = 1\} \text{ is even}])_{W\subseteq U}$ form a basis for the vector space of functions $\mathbb{R}^{\{0,1\}^U}$; we interpret the indicator corresponding to the empty set as being the constant function equal to 1. Given this, we then note that $\mathscr{P}(\{0,1\}^U)$ is a convex subset of an affine subspace of $\mathbb{R}^{\{0,1\}^U}$ of co-dimension 1, and that the indicator corresponding to the empty set is orthogonal to this affine subspace. This is then sufficient to show that for any probability distribution $\mu \in \mathscr{P}(\{0,1\}^U)$, there is a unique set of parameters $(\eta_W)_{\emptyset\neq W\subseteq U}$ such that

$$\mu(x_U) = \sum_{\emptyset\neq W\subseteq U} \eta_W \mathbb{1}[\#\{i \in W|x_i = 1\} \text{ is even}]\,,$$

as required.

To demonstrate that $(\mathbb{1}[\#\{i \in W|x_i = 1\} \text{ is even}])_{W\subseteq U}$ form a basis for the vector space of functions $\mathbb{R}^{\{0,1\}^U}$, we first note that it has the correct number of elements to form a basis, and it is therefore sufficient to either demonstrate that it is a spanning set, or that it is a linearly independent set; we take the latter approach.

Suppose we have a collection of coefficients $(\alpha_W)_{W\subseteq U}$ such that

$$\sum_{W\subseteq U} \alpha_W \mathbb{1}[\#\{i \in W|x_i = 1\} \text{ is even}] = 0\,.$$

Given a subset $X \subseteq U$, note that we have

$$\Big(\mathbb{1}[\#\{i \in X | x_i = 1\} \text{ is even}] - \mathbb{1}[\#\{i \in X | x_i = 1\} \text{ is odd}]\Big) \cdot \Big(\sum_{W \subseteq U} \alpha_W \mathbb{1}[\#\{i \in W | x_i = 1\} \text{ is even}]\Big) = 0$$

$$\implies \sum_{W \subseteq U} \alpha_W \sum_{x \in \{0,1\}^U} \Bigg(\mathbb{1}[\#\{i \in W | x_i = 1\} \text{ is even}, \#\{i \in X | x_i = 1\} \text{ is even}] - \mathbb{1}[\#\{i \in W | x_i = 1\} \text{ is even}, \#\{i \in X | x_i = 1\} \text{ is odd}]\Bigg) = 0. \tag{6}$$

Considering the summand above for a fixed subset $W \subseteq U$, note that if $W = X$, then the result of summing over all configurations $x_U \in \{0,1\}^U$ is $2^{|U|-1}$. However, if $W \neq X$, the result of the sum is 0. From this it immediately follows that $\alpha_X = 0$, and the proof of linear independence is complete. An elegant perspective which demonstrates that the sum concerned above evaluates to 0 is to view $\{0,1\}^U$ as a vector space over the finite field with 2 elements $\mathbb{F}_2$, with addition defined componentwise. In this case, the set $\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\}\}$ is exactly the kernel of the linear form $\{0,1\}^U \ni x \mapsto \sum_{i \in W} x_i \in \mathbb{F}_2$ (where the addition is to be interpreted modulo 2). Considering the linear form $\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\}\} \ni x \mapsto \sum_{i \in X} x_i \in \mathbb{F}_2$, we observe that the two sets

$$\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\} \text{ is even}, \#\{i \in X | x_i = 1\} \text{ is even}\} \text{ and}$$
$$\{x \in \{0,1\}^U | \#\{i \in W | x_i = 1\} \text{ is even}, \#\{i \in X | x_i = 1\} \text{ is odd}\},$$

are the preimage of $0 \in \mathbb{F}_2$ and $1 \in \mathbb{F}_2$ under this linear form, respectively. Therefore, if the linear form is surjective, the two sets have the same size, and since they are clearly disjoint, the relevant term of (6) evaluates to 0. To see that the form is surjective, recall that by assumption $X \neq W$. If $X \setminus W$ is non-empty, then surjectivity is demonstrated by changing a single coordinate corresponding to an index in $X \setminus W$. If $X \setminus W$ is empty, then $W \setminus X$ is non-empty, and by simultaneously chaning a coordinate in $W \setminus X$ and X, surjectivity is demonstrated. □

9.3 Proofs of results from §6: Sherali-Adams relaxations

Theorem 17. For a hypergraph $G = (V, E)$, and integer r such that $\max_{\mathcal{E} \in E} |\mathcal{E}| \leq r \leq |V|$, there is an affine score-preserving bijection

$$\mathbb{L}_r(G) \underset{\textbf{RootAt0}}{\overset{\textbf{Uproot}}{\rightleftarrows}} \widetilde{\mathbb{L}}^0_{r+1}(\nabla G)\,.$$

Proof. The structure of the proof is as follows. We first construct the uprooting map $\textbf{Uproot}$, which we will denote by $\Psi : \mathbb{L}_k(G) \to \widetilde{\mathbb{L}}^0_{k+1}(\nabla G)$ for notational convenience, and show that it is bijective by exhibiting its double-sided inverse, $\textbf{RootAt}0$, which we will denote by $\Phi : \widetilde{\mathbb{L}}^0_{k+1}(\nabla G) \to \mathbb{L}_k(G)$. We then directly show that this bijection is affine and score-preserving.

To construct Ψ, let $\boldsymbol{\mu} \in \mathbb{L}_k(G)$, and define

$$\Psi(\boldsymbol{\mu}) = \boldsymbol{\mu}^+ = (\mu^+_U)_{\substack{U \subseteq V \\ |U \setminus \{0\}| \leq k}} \in \widetilde{\mathbb{L}}^0_{k+1}(\nabla G)$$

as follows. We begin defining the measures μ^+_U for subsets U not including the additional element $0 \in V^+$ in the suspension graph; let $U \subseteq V$ with $|U| \leq k$. We define the 'symmetrized' measures

$$\mu^+_U(x_U) = \frac{1}{2}\big[\mu_U(x_U) + \mu_U(\overline{x}_U)\big] \qquad \forall x_U \in \{0,1\}^U. \tag{7}$$

Now turning our attention to subsets that *do* contain the new element $0 \in V^+$, we write $U^+ = U \cup \{0\}$, and define:

$$\mu^+_{U^+}(x_{U^+}) = \begin{cases} \frac{1}{2}\mu_U(x_U) & \text{if } x_0 = 0 \\ \frac{1}{2}\mu_U(\overline{x}_U) & \text{if } x_0 = 1 \end{cases} \qquad \forall x_{U^+} \in \{0,1\}^{U^+}. \tag{8}$$

We define $\mu_{\{0\}}(x_0)$ to take value $1/2$ for $x_0 = 0$ and $x_0 = 1$. We have now defined the entire collection of probability measures $\boldsymbol{\mu}^+$. Note that by construction, each individual measure in the collection is flipping-invariant, and by observing the form of Equations (7) and (8), we observe that the map is affine. We now demonstrate consistency of these measures. Let $W \subset U \subseteq V \cup \{0\}$. We aim to demonstrate

$$\mu^+_W(x_W) = \sum_{\substack{y_U \in \{0,1\}^U \\ y_W = x_W}} \mu^+_U(y_U) \tag{9}$$

There are three cases to consider: (i) $W \not\subseteq V$ (i.e. both subsets contain 0), (ii) $U \subseteq V$ (i.e. neither subset contains 0), (iii) $0 \in U, 0 \notin W$. In the first two cases, the marginalization consistency condition of Equation (9) follows immediately from the definitions in Equations (7) and (8), and recalling the consistency of the collection of measures $\boldsymbol{\mu}$. For case (iii), we write $U = A \cup \{0\}$ for $A \subset V$ and directly calculate:

$$\begin{aligned}
\sum_{\substack{y_U \in \{0,1\}^U \\ y_W = x_W}} \mu^+_U(y_U) &= \sum_{\substack{y_U \in \{0,1\}^U \\ y_W = x_W \\ y_0 = 0}} \frac{1}{2}\mu_A(y_A) + \sum_{\substack{y_U \in \{0,1\}^U \\ y_W = x_W \\ y_0 = 1}} \frac{1}{2}\mu_A(F_A(y_A)) \\
&= \sum_{\substack{y_A \in \{0,1\}^A \\ y_W = x_W}} \frac{1}{2}\mu_A(y_A) + \sum_{\substack{y_A \in \{0,1\}^A \\ y_W = x_W}} \frac{1}{2}\mu_A(F_A(y_A)) \\
&= \sum_{\substack{y_A \in \{0,1\}^A \\ y_W = x_W}} \mu^+_A(y_A)\,,
\end{aligned}$$

so μ^+_U and μ^+_A are consistent. The consistency of μ^+_U and μ^+_W then follows from case (ii). Having checked consistency, we have verified that the map $\Psi : \mathbb{L}_k(G) \to \widetilde{\mathbb{L}}^0_{k+1}(\nabla G)$ is well-defined. We now exhibit its inverse. Given $\boldsymbol{\eta} \in \widetilde{\mathbb{L}}^0_{k+1}(\nabla G)$, we define $\Phi(\boldsymbol{\eta}) = \boldsymbol{\mu} = (\mu_U)_{|U| \leq k} \in \mathbb{L}_k(G)$ as follows. Given $U \subseteq V$, $|U| \leq k$, write $U^+ = U \cup \{0\}$, and define

$$\mu_U(x_U) = \eta_{U^+}(x_0 = 0, x_U) + \eta_{U^+}(x_0 = 1, \overline{x}_U)$$

We now directly show that for $\boldsymbol{\mu} \in \mathbb{L}_k(G)$, we have $\Phi(\Psi(\boldsymbol{\mu})) = \boldsymbol{\mu}$. We take $|U| \leq k$, and note that from our definitions of Ψ and Φ, we have for all $x_U \in \{0,1\}^U$ that

$$\Phi(\Psi(\boldsymbol{\mu}))_U(x_U) = \mu^+_U(x_U, x_0 = 0) + \mu^+_U(\overline{x}_U, x_0 = 1) = \frac{1}{2}\mu_U(x_U) + \frac{1}{2}\mu_U(\overline{\overline{x}}_U) = \mu_U(x_U)\,.$$

Now let $\boldsymbol{\mu} \in \widetilde{\mathbb{L}}^0_{k+1}(G)$. We demonstrate that $\boldsymbol{\mu}'' = \Psi(\Phi(\boldsymbol{\mu})) = \boldsymbol{\mu}$. First, for $U \subseteq V$, $|U| \leq k$, we have

$$\begin{aligned}
\Psi(\Phi(\boldsymbol{\mu}))_U(x_U) &= \frac{1}{2}\left(\mu^+_U(x_U) + \mu^+_U(\overline{x}_U)\right) \\
&= \frac{1}{2}\Big(\mu_{U\cup\{0\}}(x_U, x_0 = 0) + \mu_{U\cup\{0\}}(\overline{x}_U, x_0 = 1) \\
&\qquad + \mu_{U\cup\{0\}}(\overline{x}_U, x_0 = 0) + \mu_{U\cup\{0\}}(x_U, x_0 = 1)\Big) \\
&= \frac{1}{2}\left(\mu_U(x_U) + \mu_U(\overline{x}_U)\right) \\
&= \mu_U(x_U)\,,
\end{aligned}$$

where in the final equality we have used the flipping-invariance of μ_U. Secondly, for $U \subseteq V$, write $U^+ = U \cup \{0\}$, and note

$$\begin{aligned}
\Psi(\Phi(\boldsymbol{\mu}))_{U^+}(x_{U^+}) &= \frac{1}{2}\mu^+_U(x_U)\mathbb{1}[x_0 = 0] + \frac{1}{2}\mu^+_U(\overline{x}_U)\mathbb{1}[x_0 = 1] \\
&= \frac{1}{2}\left(\mu_{U^+}(x_U, x_0 = 0) + \mu_{U^+}(\overline{x}_U, x_0 = 1)\right)\mathbb{1}[x_0 = 0] \\
&\quad + \frac{1}{2}\left(\mu_{U^+}(\overline{x}_U, x_0 = 0) + \mu_{U^+}(\overline{\overline{x}}_U), x_0 = 1)\right)\mathbb{1}[x_0 = 1]
\end{aligned}$$

$$= \frac{1}{2}\left(\mu_{U^+}(x_U^+) + \mu_{U^+}(\overline{x}_{U^+})\right) \mathbb{1}[x_0 = 0]$$
$$+ \frac{1}{2}\left(\mu_{U^+}(\overline{x}_{U^+}) + \mu_{U^+}(x_{U^+}))\right) \mathbb{1}[x_0 = 1]$$
$$= \mu_{U^+}(x_{U^+}) \, ,$$

where again in the final equality we have used the flipping-invariance of μ_{U^+}.

Finally, to see that the map is score-preserving, let $(\theta_{\mathcal{E}})_{\mathcal{E}\in E}$ be a collection of potentials defining a model on $G = (V, E)$. Then for any $\boldsymbol{\mu}^+ \in \mathbb{L}^0_{k+1}(G)$, note that we have

$$\begin{aligned}
&\sum_{\mathcal{E}\in E} \mathbb{E}_{X_{\mathcal{E}\cup\{0\}}\sim\mu^+_{\mathcal{E}\cup\{0\}}}\left[\theta_{\mathcal{E}\cup\{0\}}(X_{\mathcal{E}\cup\{0\}})\right] \\
=&\sum_{\mathcal{E}\in E} \sum_{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+}} \theta_{C^+}(x_{C^+})\mu^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) \\
=&\sum_{\mathcal{E}\in E}\left[\sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+} \\ x_0=0}} \theta_{\mathcal{E}^+}(x_{\mathcal{E}^+})\mu^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) + \sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+} \\ x_0=1}} \theta_{\mathcal{E}^+}(x_{\mathcal{E}^+})\mu^+_{\mathcal{E}^+}(x_{\mathcal{E}^+}) \right] \\
=&\sum_{\mathcal{E}\in E}\left[\sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+} \\ x_0=0}} \theta_{\mathcal{E}}(x_{\mathcal{E}})\frac{1}{2}\mu_{\mathcal{E}}(x_{\mathcal{E}}) + \sum_{\substack{x_{\mathcal{E}^+}\in\{0,1\}^{\mathcal{E}^+} \\ x_0=1}} \theta_{\mathcal{E}}(\overline{x}_{\mathcal{E}})\frac{1}{2}\mu_{\mathcal{E}}(\overline{x}_{\mathcal{E}}) \right] \\
=&\sum_{\mathcal{E}\in E} \sum_{x_C\in\{0,1\}^C} \theta_{\mathcal{E}}(x_{\mathcal{E}})\mu_{\mathcal{E}}(x_{\mathcal{E}}) \\
=&\sum_{\mathcal{E}\in E} \mathbb{E}_{X_{\mathcal{E}}\sim\mu_{\mathcal{E}}}\left[\theta_{\mathcal{E}}(X_{\mathcal{E}})\right] \, ,
\end{aligned}$$

as required.

□

Lemma 19. If $\mathbb{L}_r$ is universally rooted for hypergraphs of maximum hyperedge degree $p < r$ with p even, then $\mathbb{L}_r$ is also universally rooted for r-admissible hypergraphs with maximum degree $p + 1$.

Proof. The key observation is that given some set of variables x_U of size $p + 1$, if we have a set of flipping-invariant probability measures on $\{0, 1\}^W$ for each subset $W \subseteq U$ of size p which are consistent, then by Proposition 11, then a flipping-invariant probability measure over $\{0, 1\}^U$ is specified by one additional parameter. The parameter corresponds to the even potential U, and is given by

$$\mathbb{P}(|\{i \in U | x_i = 1\}| \text{ is even})$$

But since $p + 1$ is odd, and we require the measure to be flipping-invariant, the only possible value for this parameter must be $1/2$. Moreover, taking the parameter to be $1/2$ must yield a valid distribution over $\{0, 1\}^U$, as we assumed that the measures on each of $\{0, 1\}^W$ ($W \subseteq U$, $|W| = p$) were consistent.

This demonstrates that given a hypergraph G with maximum hyperedge degree $p+1$, we can construct a new hypergraph $G' = (V, E')$, with the same vertex set as G, and hyperedge set defined by

$$E' = \{\mathcal{E} \in E | |\mathcal{E}| \leq p\} \cup \{U \subset V | U \subseteq \mathcal{E} \in E, |E| = p + 1\, , |U| = p\}$$

From our argument above, we have $\mathbb{L}_r(G)$ is in affine bijection with $\mathbb{L}_r(G')$, and since G' has maximum hyperedge degree p, the statement of the lemma follows. □

Theorem 20. $\mathbb{L}_3$ is universally rooted.

Proof. We observe that it is straightforward to extend the analysis in the Appendix of [19] to demonstrate that for any hypergraph of maximum hyperedge degree 2, there exists a score-preserving affine bijection between $\mathbb{L}_3(G)$ and each of its rerootings. We now combine this with the observation of Lemma 19, taking $p = 2$, from which the statement of the theorem immediately follows. $\square$

Theorem 21. $\mathbb{L}_3$ is unique in being universally rooted. Specifically, for any integer $r > 1$ other than $r = 3$, we constructively demonstrate a hypergraph $G = (V, E)$ with $|V| = r + 1$ variables for which $\widetilde{\mathbb{L}}^0_{r+1}(\nabla G) \neq \widetilde{\mathbb{L}}^i_{r+1}(\nabla G)$ for any $i \in V$.

Proof. For each $k \neq 3$, we shall constructively demonstrate a model M on hypergraph G as stated such that the LP relaxation over $\mathbb{L}_k(G)$ is not tight for M but the LP relaxation over $L_k(\nabla G \setminus \{i\})$ is tight for every rerooted model $M_i, i \in V$.

Case 1: k **is even.** Let $G = (V, E)$, with $V = \{1, \ldots, k+1\}$, and E the set of all subsets of V of size k. Consider a model with the following set of potentials on this hypergraph:

$$\theta_{\mathcal{E}}(x_{\mathcal{E}}) = -\mathbb{1}[\#\{i \in \mathcal{E} | x_i = 1\} \text{ is even}] \qquad \forall \mathcal{E} \in E\,. \tag{10}$$

The optimum score for a configuration $x_V \in \{0,1\}^V$ is -1. We show this by demonstrating (i) that the optimum is at most -1 (which is all we need here), then (ii) that the optimum is at least -1. For (i): Toward contradiction, assume that there exists a configuration that does not activate any of the $\theta_{\mathcal{E}}$ potentials, i.e. all k-clusters have an odd number of 1s. Pick one of the k-clusters and call it S. Since $k \geq 2$ is even, S contains at least one variable set to 0, call it x, and at least one set to 1, call it y. Now V has $k + 1$ variables consisting of S together with one more variable z. If $z = 0$ then consider the k-cluster $T = S \setminus \{y\} \cup \{z\}$. If $z = 1$ then let $T = S \setminus \{x\} \cup \{z\}$. In either case, T has an even number of 1s, contradiction. For (ii): Consider the setting $x_1 = 1$ with all other variables set to 0. All k-clusters including x_1 are inactive. There is just one k-cluster not including x_1, and this k-cluster has no 1s thus its potential is active. Hence, this configuration achieves a score of -1.

However, the set of pseudomarginal distributions in $\mathbb{L}_k(G)$ below attains a score of 0:

$$\mu_{\mathcal{E}}(x_{\mathcal{E}}) = \frac{1}{k} \sum_{i \in \mathcal{E}} \delta_{x_i = 1, x_{\mathcal{E} \setminus \{i\}} = 0} \qquad \forall \mathcal{E} \in E\,.$$

Now observe that when this model is uprooted, we have the hypergraph $\nabla G = (V^+, E)$, where $V^+ = \{0\} \cup V$, and the hyperedge set $E^+ = E$ as before with the same set of potentials as in (10), by Lemma 12. Therefore, rerooting at a variable $i \in \{1, \ldots, k+1\}$ will result in a graphical model on the graph $\nabla G \setminus \{i\}$ with vertices $\{0, 1, \ldots, k+1\} \setminus \{i\}$, and hyperedges given by one hyperedge of size k (the original hyperedge which did not include i), which is $\{1, \ldots, k+1\} \setminus \{i\}$, along with all subsets of $\{1, \ldots, k+1\} \setminus \{i\}$ of size $k-1$. In particular, the model consists of potentials over the set of k variables $\{1, \ldots, k+1\} \setminus \{i\}$, and the variable X_0 is independent from the rest of the variables, with symmetric distribution on its state space $\{0, 1\}$. Therefore, the polytope $\mathbb{L}_k(\nabla G \setminus \{i\})$ is tight for this potential since it is effectively a model over k variables, proving the claim.

Case 2: $k \geq 5$ **is odd.** Let $k \geq 5$ be odd, and again let $G = (V, E)$, with $V = \{1, \ldots, k+1\}$, this time letting E be the set of all subsets of V of size $k-1$ (an even number). Consider the following set of potentials on this hypergraph

$$\theta_{\mathcal{E}}(x_{\mathcal{E}}) = -\mathbb{1}[\#\{i \in \mathcal{E} | x_i = 1\} \text{ is even}] \qquad \forall \mathcal{E} \in E\,.$$

We note that the polytope $\mathbb{L}_k(G)$ is not tight for this polytope, by considering the following set of pseudomarginals over hyperedges of G:

$$\mu_{\mathcal{E}}(x_{\mathcal{E}}) = \frac{1}{k} \delta_{x_i = 0 \forall i \in \mathcal{E}} + \frac{1}{k} \sum_{i \in \mathcal{E}} \delta_{x_i = 1, x_{\mathcal{E} \setminus \{i\}} = 0} \qquad \forall \mathcal{E} \in E\,.$$

These are valid pseudomarginals in $\mathbb{L}_k(G)$, as the following distributions over k-clusters are consistent and marginalize down to the distributions over hyperedges:

$$\mu_U(x_U) = \frac{1}{k} \sum_{i \in U} \delta_{x_i = 1, x_{U \setminus \{i\}} = 0} \qquad \forall U \subseteq V\,, |U| = k\,.$$

Note that the score of this set of pseudomarginals is

$$\sum_{\mathcal{E} \in E} -\mu_{\mathcal{E}}(\#\{i \in \mathcal{E} | x_i = 1\} \text{ is even}) = -\binom{k+1}{k-1}\frac{1}{k} = -\frac{k+1}{2}$$

We now argue that this exceeds the maximum score obtainable by a configuration $x_V \in \{0,1\}^V$, demonstrating non-tightness of $\mathbb{L}_k(G)$ for this model. To see this, let $\ell \in \{0, \ldots, k+1\}$ be the number of non-zero coordinates of x_V. We count the number of subsets U of $\{1, \ldots, k+1\}$ of size $k-1$ for which x_U has an even number of non-zero coordinates, and show that this is greater than $(k+1)/2$, leading to a score less than $-(k+1)/2$. The number of such subsets is given by:

$$\sum_{p=0}^{\lfloor \ell/2 \rfloor} \binom{\ell}{2p}\binom{k+1-l}{k-1-2p} = \begin{cases} (k+1)(k-1)/2 & \ell = 0 \\ \binom{\ell}{\ell-2}\binom{k+1-l}{k+1-l} + \binom{\ell}{\ell}\binom{k+1-l}{k-1-\ell} = \frac{\ell(\ell-1)}{2} + \frac{(k+1-\ell)(k-\ell)}{2} & \ell \neq 0 \text{ even} \\ \binom{\ell}{\ell-1} + \binom{k+1-\ell}{k-\ell} = k+1 & \ell \text{ odd.} \end{cases}$$

For ℓ odd and $\ell = 0$ the conclusion is clear, and for ℓ even and non-zero, we observe that the quadratic expression in ℓ above is minimized at $\ell = (k+1)/2$ (which is an integer, as k is odd), and takes the value $(k^2-1)/4$, which is greater than $(k+1)/2$ for all odd $k \geq 5$ (though the two values are equal for $k = 3$).

Now observe that when this model is uprooted and subsequently rerooted at a new variable $i \in V$, we obtain a model on $k+1$ variables, but with the variable X_0, introduced by uprooting, independent from the rest. Therefore, the model is effectively over only k variables, and hence it follows that $\mathbb{L}_k(\nabla G \setminus \{i\})$ is tight for this rerooting, proving the claim. $\square$

10 Additional Experimental Details and Results

In this section, we expand on the Experiments Section 7 of the main paper to provide:

- §10.1: Model structures and parameters used for libDAI
- §10.2: How we fit constants of the clamp selection heuristics
- §10.3: Additional experimental results
  - §10.3.1: Timings
  - §10.3.2: MAP inference
  - §10.3.3: Marginals
  - §10.3.4: Higher-order potentials over clusters of 5 and 6 variables
  - §10.3.5: Comparison of our heuristics to the maxtW heuristic used in [19]
  - §10.3.6: Larger models
- §10.4: Additional discussion

10.1 Model structures and parameters used for libDAI

In this section we give further information about the model structures used in our experiments, as well as the methods of approximate inference used. All potentials are pure k-potentials, as in §5, which for brevity we may write simply as a k-potential.

Complete graphs For complete graph experiments, there is a pure k-potential for each subset of k variables, for $k = 1, 2, 3, 4$.

Grids All grids are square and toroidal. There is a 1-potential for each variable, and a 2-potential for each edge of the graph. There is a 3-potential for each possible "L-shaped" connected subgraph of size 3 (any of the four possible orientations), and a 4-potential for each cycle of size 4.

Potentials In our experiments, unless otherwise specified, the default is that all pure 2- and 4-potential coefficients are drawn independently from $\text{Unif}([-8, 8])$ distributions, while all pure 1- and 3-potential coefficients are set to 0. Using the notation of Section 7, in each experiment a parameter $W_{\max}$ is varied, and the default distribution of one class of pure potentials (either 1-, 2-, 3-, or 4-potentials) is overrided from the default specification to be replaced by coefficients from a $\text{Unif}([-W_{\max}, W_{\max}])$ distribution.

LibDAI settings In all cases, we use the junction tree algorithm with Hugin updates for exact inference. For approximate marginal inference, we use the LibDAI `HAK` implementation of [4], with precise parameters passed to MATLAB given by:

`'[doubleloop=1,clusters=BETHE,init=UNIFORM,tol=1e-9,maxiter=10000]'`.

For approximate marginal inference, we use the LibDAI `BP` loopy belief propagation implementation, with precise parameters passed to MATLAB given by:

`'[inference=MAXPROD,updates=SEQFIX,logdomain=1,tol=1e-9,maxiter=10000,damping=0.0]'`

10.2 How we fit constants of the clamp selection heuristics

In this section we give further details of how the K and G heuristics used in our experiments were fitted, expanding on the explanation given in Section 7. Using the notation developed in Section 7, the family of heuristics we consider maximize the following measure

$$\text{clamp-heuristic-measure}(i) = \sum_{i \in \mathcal{E}: |\mathcal{E}|=2} c_2 \tanh |t_2 a_{\mathcal{E}}| + \sum_{i \in \mathcal{E}: |\mathcal{E}|=4} c_4 \tanh |t_4 a_{\mathcal{E}}| \,, \qquad (11)$$

over $i \in V^+$, and are parametrized by the four scalars t_2, c_2, t_4, c_4. We first note that 11 is over-parametrized (since we are interested only in ranking the scores for each variable in M^+), so we take $c_2 = 1$. To fit the heuristic, we used gradient-free optimisation. For the K heuristic, we generated a collection of graphical models on K_8, and constructed a fitness function over the remaining parameters t_2, c_4, t_4, given by the average ranking of the rerooting selected by the heuristic for $\log Z$ estimation across our collection of complete graphs.

We then initialized the parameters $t_2 = 0, c_4 = 1, t_4 = 0$, and performed a local exploration of the parameter space dictated by a Gaussian random walk, updating our parameter settings when they led to an improvement in the value of the fitness function.

The G heuristic was constructed similarly, instead using a collection of grids to define the fitness function.

The precise values of the fitted heuristics are given below:

$$\text{K-heuristic-measure}(i) = \sum_{i \in \mathcal{E}: |\mathcal{E}|=2} c_2 \tanh |0.051 a_{\mathcal{E}}| + \sum_{i \in \mathcal{E}: |\mathcal{E}|=4} 0.091 \tanh |1.482 a_{\mathcal{E}}|,$$

$$\text{G-heuristic-measure}(i) = \sum_{i \in \mathcal{E}: |\mathcal{E}|=2} c_2 \tanh |0.019539 a_{\mathcal{E}}| + \sum_{i \in \mathcal{E}: |\mathcal{E}|=4} 0.3788 \tanh |0.033997 a_{\mathcal{E}}|.$$

The heuristic of [19], $maxtW$, applied only to pairwise models, and in the notation of our paper, was given by the following clamping score measure:

$$\text{clamp-heuristic-measure}(i) = \sum_{i \in \mathcal{E}: |\mathcal{E}|=2} \tanh |\frac{1}{2} a_{\mathcal{E}}|. \qquad (12)$$

Recognizing that the benefits of our heuristics appeared somewhat robust to exact parameter choice, when we extended analysis to 6-potentials in §10.3.4, we extended our K heuristic by eye (without fitting to any data, and before examining the results for higher order models), and explore a variant on the G heuristic. We used the following measures:

$$\text{K-heuristic-measure}(i) = \sum_{i \in \mathcal{E}: |\mathcal{E}|=2} \tanh |0.2 a_{\mathcal{E}}| + \sum_{i \in \mathcal{E}: |\mathcal{E}|=4} \frac{1}{3} \tanh |1.2 a_{\mathcal{E}}| + \sum_{i \in \mathcal{E}: |\mathcal{E}|=6} \frac{1}{5} \tanh |3 a_{\mathcal{E}}|,$$

$$\text{G-heuristic-measure}(i) = \sum_{i \in \mathcal{E}: |\mathcal{E}|=4} |a_{\mathcal{E}}|.$$

10.3 Additional experimental results

We provide the following:

- §10.3.1: Timings
- §10.3.2: MAP inference
- §10.3.3: Marginals

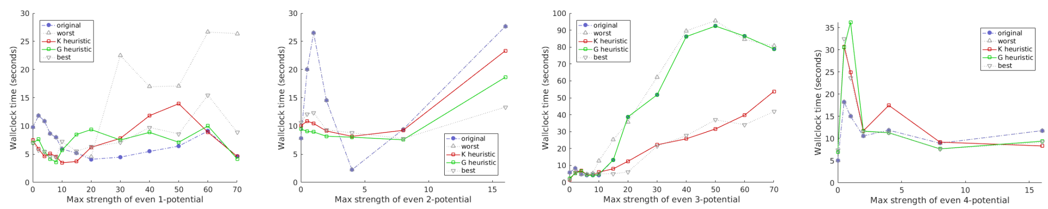
- §10.3.4: Higher-order potentials over clusters of 5 and 6 variables
- §10.3.5: Comparison of our heuristics to the maxtW heuristic used in [19]
- §10.3.6: Larger models

In all plots, if the red curve for the K heuristic is not visible, it coincides with the green curve for the G heuristic. We use consistent legends across all plots.

10.3.1 Timings

Times in seconds to run marginal inference (i.e. estimating $\log Z$ and marginals) using libDAI are shown in Figure 3. Inference for rerooted models took a similar amount of time as for the original model. We caution against relying heavily on the accuracy of these timings since we made no attempt to optimize our code for speed, and we ran our inference algorithms in a cluster environment beyond our control.

Time/sec to run marginal inference for K_8 complete hypergraphs (fully connected) on 8 variables.



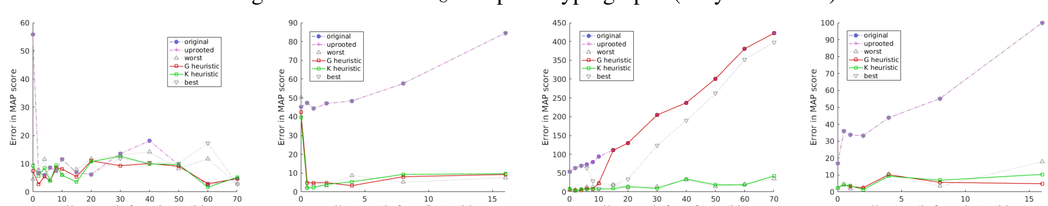
vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

Figure 3: Average time to perform marginal inference using libDAI over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which ex post gave the lowest error in estimating $\log Z$. See §10.3.1.

10.3.2 MAP inference

Results are shown in Figure 4. We observe here that rerooting does not help much when 1-pots are varied, but can provide great benefit for the other cases shown. The K heuristic (which was trained on complete graphs like the one we analyze here) performs well in all settings. Curiously, the G heuristic (which was trained only on grids) performs well when 2-pots or 4-pots are varied, but not when 3-pots are varied (though even here it does no worse than the original rooting). We aim to explore this further in future work.

Error in estimating MAP score for K_8 complete hypergraphs (fully connected) on 8 variables.



vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

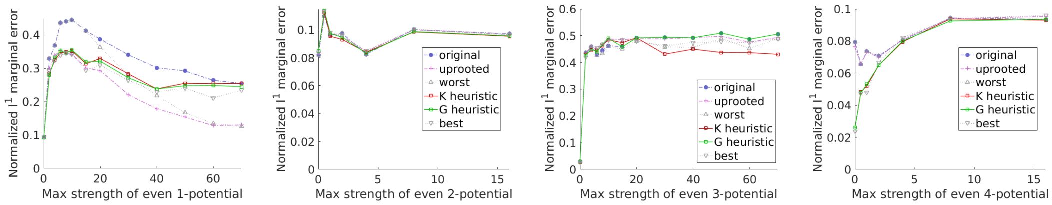
Figure 4: Average error in estimating MAP score using libDAI over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which ex post gave the lowest error in estimating $\log Z$. See §10.3.2.

10.3.3 Marginals

Results are shown in Figure 5. Our models were selected to present an interesting range of problems for partition function estimation, which led to marginals often being challenging to estimate. Still, results for marginal inference were often improved by rerooting.

We note that another natural way to estimate marginals is as the ratio of a clamped partition function to the original partition function. Since we have seen good evidence that rerooting can help significantly with partition function estimation, it is reasonable to hope that in future work, we may observe significant benefits to marginal inference via this approach by using rerooting.

Error in estimating 1-marginals for K_8 complete hypergraphs (fully connected) on 8 variables.



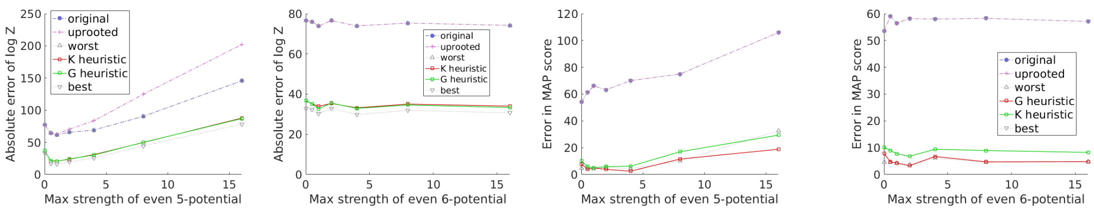
vary $W_{\max}$ for 1-pots | vary $W_{\max}$ for 2-pots | vary $W_{\max}$ for 3-pots | vary $W_{\max}$ for 4-pots

Figure 5: Average ℓ_1 error in estimating marginals (minimal representation corresponding to pure k-potentials, see §5) using libDAI over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which ex post gave the lowest error in estimating $\log Z$. See §10.3.3.

10.3.4 Higher-order potentials over clusters of 5 and 6 variables

Results for a complete hypergraph K_8 on 8 variables, this time with potentials up to order 6, are shown in Figure 6. In all cases, rerooting using our heuristics is very helpful.

Error in estimating $\log Z$ (left) and MAP score (right) for K_8 hypergraphs on 8 variables with **potentials up to order 6**.



vary $W_{\max}$ for 5-pots | vary $W_{\max}$ for 6-pots | vary $W_{\max}$ for 5-pots | vary $W_{\max}$ for 6-pots

Figure 6: Average error in estimating $\log Z$ (left) and MAP score (right) using libDAI over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. *Best* and *worst* refer to the rerootings which ex post gave the lowest error in estimating $\log Z$. See §10.3.4.

10.3.5 Comparison of our heuristics to the maxtW heuristic used in [19]

Results for a complete graph K_8 on 8 variables, this time with potentials only up to order 2, are shown in Figure 7. We have added the earlier maxtW heuristic used in [19], which using our notation corresponds to setting $t_2 = \frac{1}{2}$ in (5). Note that for the pairwise models considered here, the clamp heuristic constants for potentials of order higher than 2 are irrelevant.

We observe that our K and G heuristics (which were fit on different models with potentials up to order 4, so here are out of sample) achieve similar performance to the earlier maxtW heuristic, in fact yielding slightly better results. This is encouraging evidence for robustness of the simple form of heuristic score (5).

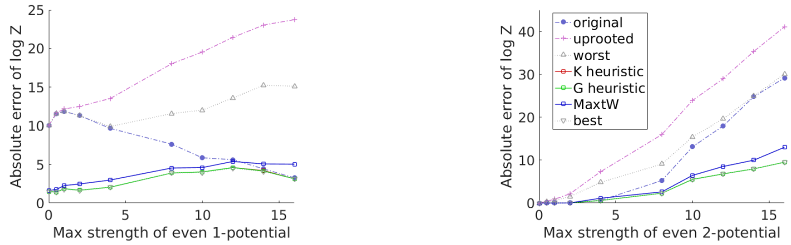
10.3.6 Larger models

Results for a complete hypergraph K_{10} on 10 variables (potentials up to order 4) are shown in Figure 8. Results are qualitatively similar to those for smaller models in §7 of the main paper.

10.4 Additional discussion

When discussing pure k-potentials in §5, we observed that for a pure k-potential (which we showed must essentially be an even k-potential) with k an even number, $\theta_{\mathcal{E}}(x_{\mathcal{E}}) = \theta_{\mathcal{E}}(\overline{x}_{\mathcal{E}})$. This means that the coefficient of any such k-potential is invariant with respect to a flipping of all variables of the

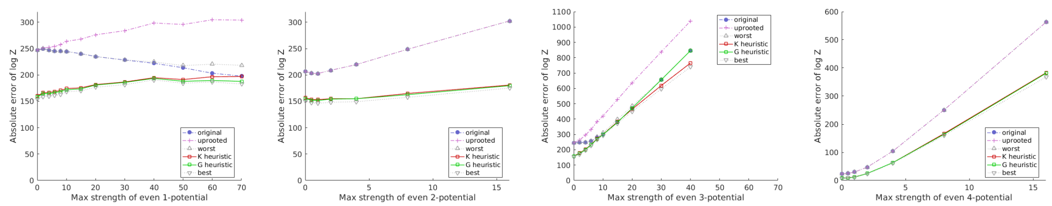
Average abs(error) in $\log Z$ for K_8 complete **pairwise** graphs (fully connected) on 8 variables:
adding earlier maxtW heuristic for comparison (our K and G heuristics coincide on these runs).



vary $W_{\max}$ for 1-pots vary $W_{\max}$ for 2-pots

Figure 7: Error in estimating $\log Z$ for random pairwise models with various pure k-potentials over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 8 for $k = 1$, and 8 for $k = 2$. K and G heuristics coincide. See §10.3.5.

Average abs(error) in $\log Z$ for K_{10} complete hypergraphs (fully connected) on 10 variables.



vary $W_{\max}$ for 1-pots vary $W_{\max}$ for 2-pots vary $W_{\max}$ for 3-pots vary $W_{\max}$ for 4-pots

Figure 8: Error in estimating $\log Z$ for random models with various pure k-potentials over 20 runs. If not shown, $W_{\max}$ max coefficients for pure k-potentials are 0 for $k = 1$, 8 for $k = 2$, 0 for $k = 3$, 8 for $k = 4$. See §10.3.6.

model (whereas the if k is an odd number, the coefficient will flip sign). Hence for k even, the sign of the coefficient may be regarded as a fundamental property of the potential.

When $k = 2$ this sign dicatates the submodularity or supermodularity of the 2-potential. If all potentials are pure 2-potentials with positive coefficients, then the model is *regular* or *ferromagnetic* and typically admits easier inference.

For higher k, this is no longer true. However, note that still if we represent a model's potentials in terms of pure k-potentials, and all have k even with a positive coefficient, then the model is special in the sense that:

- The configurations of all 0s and all 1s must be mode configurations, typically with significantly higher probabilities than others.
- Inference will typically be relatively straightforward.
- If the model is rerooted, then this will effectively clamp all variables close to 0 or 1 and the error of approximate inference should be low.