Methods for Inference in Graphical Models

Adrian Weller

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Methods for Inference in Graphical Models

Adrian Weller

Graphical models provide a flexible, powerful and compact way to model relationships between random variables, and have been applied with great success in many domains.

Combining prior beliefs with observed evidence to form a prediction is called *inference*. Problems of great interest include finding a configuration with highest probability (MAP inference) or solving for the distribution over a subset of variables (marginal inference). Further, these methods are often critical subroutines for learning the relationships. However, inference is computationally intractable in general. Hence, much effort has focused on two themes: finding subdomains where exact inference is solvable efficiently, or identifying approximate methods that work well. We explore both these themes, restricting attention to undirected graphical models with discrete variables.

First we address exact MAP inference by advancing the recent method of reducing the problem to finding a maximum weight stable set (MWSS) on a derived graph, which, if perfect, admits polynomial time inference. We derive new results for this approach, including a general decomposition theorem for models of any order and number of labels, extensions of results for binary pairwise models with submodular cost functions to higher order, and a characterization of which binary pairwise models can be efficiently solved with this method. This clarifies the power of the approach on this class of models, improves our toolbox and provides insight into the range of tractable models.

Next we consider methods of approximate inference, with particular emphasis on the Bethe approximation, which is in widespread use and has proved remarkably effective, yet is still far from being completely understood. We derive new formulations and properties of the derivatives of the Bethe free energy, then use these to establish an algorithm to compute $\log$ of the optimum Bethe partition function to arbitrary ϵ-accuracy. Further, if the model is attractive, we demonstrate a fully polynomial-time approximation scheme (FPTAS), which is an important theoretical result, and demonstrate its practical applications. Next we explore ways to tease apart the two aspects of

the Bethe approximation, i.e. the polytope relaxation and the entropy approximation. We derive analytic results, show how optimization may be explored over various polytopes in practice, even for large models, and remark on the observed performance compared to the true distribution and the tree-reweighted (TRW) approximation. This reveals important novel observations and helps guide inference in practice. Finally, we present results related to clamping a selection of variables in a model. We derive novel lower bounds on an array of approximate partition functions based only on the model's topology. Further, we show that in an attractive binary pairwise model, clamping any variable and summing over the approximate sub-partition functions can only increase (hence improve) the Bethe approximation, then use this to provide a new, short proof that the Bethe partition function lower bounds the true value for this class of models.

The bulk of this work focuses on the class of binary, pairwise models, but several results apply more generally.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I owe many debts of thanks. First to my advisor, Tony Jebara, for taking me under his wing and giving me great freedom to explore, while also sharing thoughtful and calm guidance throughout. In addition, Tony provided the opportunity for me to teach the Machine Learning class at Columbia, which was a terrific experience.

Everyone on my thesis committee has been a superb teacher, supporter, friend and advisor. Each is a leader in their field, yet has taken time to provide patient help and guidance. Many thanks to Tony Jebara, Maria Chudnovsky, Al Aho, David Sontag and Amir Globerson.

I've had the privilege to collaborate with wonderful people. Many thanks to Tony, Dan Ellis, Kui Tang, David Sontag and Eli Vovsha.

To everyone who has put up with sharing an office with me - Delbert Dueck, Michele Merler, Anna Chromanska, Kui Tang and Nick Ruozzi - thank you for your patience and friendship. And to everyone else that has been part of the Machine Learning Lab during my time - Max Danisch, Andrew Howard, Bert Huang, Pannaga Shivaswamy, Blake Shaw, Yingbo Song, Pannaga Shivaswamy and Kapil Thadani - thank you all for your help and many valuable discussions.

Among many others that have brightened my time at Columbia, I thank Mihalis Yannakakis, Tal Malkin, Rocco Servedio, John Zheng, Daniel Guetta, Ido Rosen, Akiva Bamberger, Berk Birand, Paul Vishayanuroj, Samina Rahman, and Team Upbeat - Peyton Sherwood, Matt d'Zmura, Fan Lin and Miles Ulrich.

Several researchers generously responded to emails and provided wise guidance. These include Aryeh Kontorovich, Simon Baker, Joris Mooij, Simon Prince, Shai Bagon, Sri Ramalingam, Boris Flach, Pushmeet Kohli, Uri Heinemann, Martin Szummer, Vladimir Kolmogorov and Erik Sudderth. Thank you for your help.

I thank all my friends, particularly Tomaž Slivnik and Barney Pell, and, most importantly, all my family for their constant patience, wisdom and warm support.

Some of the material herein is based upon work supported by the National Science Foundation under Grant No. 1117631.

For my family

Chapter 1

Introduction

A *graphical model* is a structured probabilistic model for a system of random variables whose joint distribution may be compactly specified by the product of *potential functions* over subsets of the variables. In this thesis, we focus on discrete *undirected graphical models*, also termed *Markov random fields* (MRFs), wherein each potential function returns a nonnegative measure of compatibility of the settings of its variables. These potential functions are typically unnormalized, hence, in order to provide a probability distribution which sums to 1, a normalizing constant called the *partition function* must be computed. The complete set of variables, together with the set of the potential function subsets, specify a topology which is naturally represented by a hypergraph.

This formulation provides a natural platform over which to seek efficient algorithms that lever the conditional independence properties of the variables, which are captured via graph separation, and is well-suited to help make meaningful sense out of large, complex data sets, as are increasingly familiar across many domains. There is a rich history of contributions from, and applications to, many fields, including:

- Modeling phase transitions in statistical physics (Bethe, 1935; Peierls and Born, 1936; Yeomans, 1992). Indeed, Bethe's ideas from the 1930s are still finding remarkably fresh application today, as we shall explore in Part III.
- Signal processing and speech recognition (Viterbi, 1967; Baum et al., 1970; Forney, 1973; Rabiner, 1990).
- Communication and coding theory (Frey, 1998; McEliece et al., 1998).

- Combinatorics (Barahona, 1982; Barahona and Mahjoub, 1986; Chandrasekaran et al., 2011).

- Computer vision and image processing (Woods, 1976; Geman and Geman, 1984; Felzenszwalb and Huttenlocher, 2004; Li, 1995; Blake et al., 2011).
- Modeling protein structure (Yanover and Weiss, 2002; Kamisetty et al., 2007; Yanover et al., 2008).

Given a graphical model with its potential functions, sometimes together with a set of some observed variables, there are three canonical problems of *inference*:

1. Maximum a posteriori (MAP) inference, which is the task of identifying a setting of all the unobserved variables with maximum probability. An example is image reconstruction, where a noisy image is received and we try to reconstruct the most likely image which was initially sent.
2. Marginal inference, which is computing the probability distribution of a given subset of variables. For example, a medical diagnosis system might observe many characteristics of a patient, such as temperature, blood pressure and other symptoms; given this information, we would like to estimate the probability that the patient has various diseases.
3. Evaluating the partition function, which involves summing over all possible configurations. For example, counting the number of independent sets of a graph.

All are computationally challenging. Given the numerous applications, there is an extensive literature of the many approaches that have been explored. In this thesis, we shall examine and make contributions to each of the three problems. In Part I, we introduce notation and provide general background, then present our contributions in Parts II and III.

1.1 Summary of Contributions

In Part II, we approach the problem of MAP inference by building on a recent method introduced by Jebara (2009) and Sanghavi et al. (2009). This reduces the problem to the graph theoretic challenge of finding a *maximum weight stable set* (MWSS) in a derived weighted graph, termed a *nand Markov random field* (NMRF). Terms and earlier results used from graph theory are provided in Section 3.3.

- In Chapter 4, we derive new results for this approach, including a general decomposition theorem for MRFs of any order and number of labels, extensions of results for binary pairwise models with submodular cost functions to higher order, and a characterization of which binary pairwise MRFs can be efficiently solved with this method. This helps to define the power of the approach, improves our toolbox and provides insight into the range of tractable models.

In Part III, we consider approximate inference techniques to estimate marginal probabilities and the partition function, focusing on the Bethe approximation. This has a long history beginning in statistical physics in the 1930s, and is in widespread use through the belief propagation algorithm, yet is still far from being completely understood. We make contributions in several ways:

- In Chapter 6, we derive novel bounds and results on marginals of the Bethe approximation and derivatives of the Bethe free energy, building on the work of Korč et al. (2012), including a new characterization of the Hessian matrix of second partial derivatives. This demonstrates interesting qualitative properties: the main diagonal is always positive (hence the function is convex if all but one variables are held fixed), and importantly, all off-diagonal entries are negative for an attractive edge or positive for a repulsive edge. Thus, an attractive model has *submodular* energy. By bounding the terms of the Hessian, we are able to derive an optimization algorithm over a discrete mesh, which may be framed as multi-label MAP inference, to solve for $\log$ of the optimal Bethe partition function to arbitrary ϵ-accuracy.
- By also analyzing properties of the first derivatives of the Bethe free energy, we show how the mesh efficiency may be dramatically improved for almost all problems, unless ϵ is extremely small. Applying the method to attractive binary pairwise models, the earlier submodularity result means that the derived mesh optimization problem is submodular for any discretization and hence can be solved in polynomial time using graph cuts. This leads to a fully polynomial-time approximation scheme (FPTAS) to approximate $\log$ of the Bethe partition function for any attractive binary pairwise model (any topology).
- In Chapter 7, we contribute to a deeper understanding of the Bethe approximation by separately analyzing the two aspects of its approximation: the polytope relaxation from the marginal polytope to the local polytope, and the Bethe entropy approximation. We demonstrate novel theoretical insights and also investigate empirically the merits of optimizing over

the local, cycle or marginal polytopes, comparing results to the true values and those obtained using the tree-reweighted (TRW) approach.

- In Chapter 8, we make further use of the properties of the derivatives of the Bethe free energy derived in Chapter 6. We are able to demonstrate that, in an attractive binary pairwise model, clamping any variable to each of its two values and summing over the optimal Bethe sub-partition functions obtained, can only increase (and hence improve) the partition function approximation obtained. In deriving this result, we show an interesting stronger result on how the optimal Bethe partition function varies as one variable is held to various fixed values. We combine this result with an observation on clamping to derive a new proof from first principles of a recent, important result, that the Bethe partition function for an attractive binary pairwise model is always a lower bound for the true value (Ruozzi, 2012). Further, we are able to derive a related lower bound for the approximate partition function of a range of approximation methods, which applies in the multi-label case.

Appendices are provided at the end with additional proofs and technical material.

1.2 Publications

Much of the work in this thesis is based upon material appearing in the publications below, all are available at `www.cs.columbia.edu/~adrian`.

- Weller and Jebara (2013a)

  A. Weller and T. Jebara. Bethe bounds and approximating the global optimum. In *Artificial Intelligence and Statistics*, 2013.

- Weller and Jebara (2013b)

  A. Weller and T. Jebara. On MAP inference by MWSS on perfect graphs. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

- Weller et al. (2014)

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

- Weller and Jebara (2014a)

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

And in the following paper to appear in NIPS 2014:

- Weller and Jebara (2014b)

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

Related work appears in

- Tang et al. (2013) K. Tang, A. Weller, and T. Jebara. Network ranking with Bethe pseudomarginals. In *NIPS Workshop on Discrete Optimization in Machine Learning*, December 2013.

1.2.1 Earlier work

Earlier work explored the related theme of structured prediction methods for chord classification in music:

- (Weller et al., 2009) A. Weller, D. Ellis, and T. Jebara. Structured prediction models for chord transcription of music audio. In *International Conference on Machine Learning and Applications*, 2009.

When combined with Dan Ellis' already-powerful approach to chord transcription, we submitted the best overall entry to the MIREX open competition that year, see results at `http://www.music-ir.org/mirex/wiki/2009:Audio_Chord_Detection_Results`.

An overview of the complete 2010 LabROSA chord recognition system is available here `http://www.ee.columbia.edu/~dpwe/pubs/Ellis10-chords`.

Part I

General Background

Chapter 2

Notation and Preliminaries

In this Chapter, we provide a brief survey of undirected graphical models, introducing notation and preliminary results. For further reading, several excellent texts are available such as those by Wainwright and Jordan (2008), Koller and Friedman (2009) or Murphy (2012).

2.1 Markov Random Fields

Markov random fields (MRFs), also termed undirected probabilistic graphical models, are a central tool in machine learning with wide use in many areas including speech recognition (Lafferty et al., 2001), vision (Li, 1995) and computational biology (Yanover and Weiss, 2002). A model (V, Π) is specified by a set of n random variables $V = \{X_1, \ldots, X_n\}$ together with *potential functions* over subsets c of V, $\Pi = \{\pi_c : c \in C \subseteq \mathcal{P}(V)\}$, where $\mathcal{P}(V)$ is the powerset of V.[1] Throughout this thesis, we deal exclusively with *finite* and *discrete* MRFs, where each variable X_i may take finite k_i possible values which we label $\chi_i = \{0, \ldots, k_i - 1\}$. Let $x = (x_1, \ldots, x_n)$ be one particular complete configuration from the set of all possible configurations $\mathcal{X} = \prod_{i=1}^{n} \chi_i$, and x_c be a configuration of just the variables in c, which we write as X_c. A potential function π_c maps each possible setting x_c of its variables X_c to a non-negative real number $\pi_c(x_c)$. The joint probability

[1] In the literature, the potential functions are more commonly labeled ϕ or ψ but we have saved these symbols for other related purposes, which are also in common use.

distribution is given by

$$p(x) = \frac{1}{Z} \prod_{c \in C} \pi_c(x_c), \text{ where } Z = \sum_{x \in \mathcal{X}} \prod_{c \in C} \pi_c(x_c). \tag{2.1}$$

Z is the normalizing constant, termed the *partition function*, which ensures that the distribution sums to 1.[2] Treating the variables as nodes, the potential functions' argument sets define hyperedges in a hypergraph topology (see 3.3 for definitions of terms from graph theory).

An MRF embodies conditional independence relationships through graph separation. Let X_A, X_B and X_C be three sets of variables from V. From (2.1), it may easily be shown that if X_C separates X_A from X_B, i.e. if removing X_C leaves no path in the MRF hypergraph from any node of X_A to any node of X_B, then X_A is conditionally independent of X_B given X_C. The Hammersley-Clifford theorem (unpublished, 1971) shows that the converse is true in the following sense. Provided $p(x) > 0 \; \forall x \in \mathcal{X}$, if graph separation implies conditional independence, then the probability distribution may be written as a product of factors as in (2.1), with one function for each maximal clique c. For details, see (Koller and Friedman, 2009, §4.3).

In this thesis, we shall also often require $p(x) > 0 \; \forall x \in \mathcal{X}$. One reason for this is that we typically write (2.1) using *log-potential functions* given by $\psi_c(x_c) = \log \pi_c(x_c)$. With this form, (2.1) becomes

$$p(x) = \frac{1}{Z} \exp\left(\sum_{c \in C} \psi_c(x_c) \right), \tag{2.2}$$

where we require $p(x) > 0$ in order to avoid infinite exponents. This is reminiscent of a similar expression from statistical physics, where the probability of a configuration of the system is given by a *Gibbs* (or *Boltzmann*) distribution, written

$$p(x) = \frac{1}{Z} \exp\left(-\beta E(x)\right), \text{ where } \beta = \frac{1}{kT}. \tag{2.3}$$

k is *Boltzmann's constant*, which we may take to be 1, T is the *temperature* of the system, and $E(x)$ is the *energy* of the particular configuration (Yeomans, 1992). If we take $T = 1$, (2.2) and (2.3) have the same form provided the energy of a configuration is defined as

$$E(x) = -\sum_{c \in C} \psi_c(x_c). \tag{2.4}$$

[2]The symbol Z stands for *zustandssumme*, which means *sum over states* in German.

Since it is sensible for energy to be the sum of potential functions, sometimes the ψ_c functions are referred to as potential functions instead of the exponentiated form π_c.

A physical interpretation considers a high-dimensional energy landscape, where lower energy states are more desirable and have higher probability. Accordingly, the energy is also sometimes described as a *cost function*, with lower values being better, i.e. more likely. Alternatively, we shall sometimes discuss the *score* or a configuration, defined to be the negative of its energy. Higher score configurations have higher probability.

2.1.1 Alternative factor graph representation

A popular equivalent representation of the hypergraph topology is a *factor graph* (Kschischang et al., 1998). This is a bipartite graph where the variables V form one stable partition and each subset or *factor* $c \in C$ is a node in the other partition, with an edge from c to each variable it contains.

2.1.2 Conditioning on observed variables

Suppose V is split into *observed* variables $Y = y$ and *unobserved* variables X_U so $x = (x_u, y)$ with $x_u \in \mathcal{X}_u$. Then $p(x_u|y) = \frac{p(x_u,y)}{p(y)} = \frac{p(x_u,y)}{\sum_{x'_u \in \mathcal{X}_u} p(x'_u,y)}$.

This is just a new smaller MRF with modified potentials on the variable set X_U, with a new partition function to normalize the new distribution.

Hence the MRF framework is rich enough to handle conditioning. Henceforth, when we discuss MRFs, they might or might not have been based on conditioning on observed variables.

2.2 Problems of Inference

Given a model specified by a set of variables V and their log-potential compatibility functions ψ_c, there are 3 canonical inference tasks:

1. Maximum a posteriori (MAP) inference, which is the task of identifying a setting of all the unobserved variables with maximum probability.[3] From (2.2), we see that in our notation,

[3]Some authors use MAP inference to mean the more general, and typically harder, problem of identifying a setting of a *given subset* of the unobserved variables with maximum probability. A phrase that unambiguously means a MAP

this is the combinatorial problem of identifying

$$x^* \in \arg\max_{x \in \mathcal{X}} \sum_{c \in C} \psi_c(x_c). \tag{2.5}$$

Given the definition of energy (2.4), this problem is also described as *energy minimization*.

2. Marginal inference, which is computing the probability distribution of a given subset of variables. Let X_c be the given subset, then

$$p(x_c) = \sum_{x \in \mathcal{X}: X_c = x_c} p(x) = \frac{\sum_{x \in \mathcal{X}: X_c = x_c} \exp\left(\sum_{c \in C} \psi_c(x_c)\right)}{\sum_{x \in \mathcal{X}} \exp\left(\sum_{c \in C} \psi_c(x_c)\right)}. \tag{2.6}$$

3. Evaluating the partition function,

$$Z = \sum_{x \in \mathcal{X}} \exp\left(\sum_{c \in C} \psi_c(x_c)\right). \tag{2.7}$$

It is clear that problems 2 and 3 are closely related, since marginal inference may be viewed as the ratio of a sub-partition function to the full partition function. These problems, which involve summing over an exponential number of states, are typically more challenging than problem 1, where only a single optimal configuration must be identified. Yet the problems are more closely related than they may initially appear.

Considering (2.3), note that in the limit as the temperature $\to 0$, all the probability mass is focused on just the MAP state(s), hence marginal inference approaches MAP inference. Recently, a fascinating, different relationship was shown by applying MAP inference to randomly perturbed models, from which conclusions may be drawn about the partition function (Papandreou and Yuille, 2011; Hazan and Jaakkola, 2012; Ermon et al., 2013).

The work in Chapter 6 adds a further link between inference problems by demonstrating that finding a MAP configuration over a derived multi-label MRF may be used to approximate to arbitrary precision the Bethe partition function approximation of a binary pairwise model. In addition, in Appendix C.4, we explore the consequences for this method as the temperature $\to 0$.

---

configuration of *all* the unobserved variables is a *most probable explanation* (MPE) (Darwiche, 2009).

2.2.1 Reparameterizations

A *reparameterization* is a transformation of the log-potential functions $\{\psi_c\} \to \{\psi'_c\}$ such that $\forall x \in \mathcal{X}, \sum_{c\in C} \psi'_c(x_c) = \sum_{c\in C} \psi_c(x_c)+$ a constant. Considering (2.2), the constant is absorbed into the new partition function, i.e. $Z' = Z \exp(\text{constant})$, resulting in the same probability distribution. Further, the rank ordering of configurations is clearly unchanged, thus a MAP solution is unaffected. However, this type of transformation can significantly simplify subsequent problems of inference, as we shall see in Parts II and III.

2.3 Parameterization and the Exponential Family

One way to specify the log-potential functions ψ_c is a log-linear form using real-valued *sufficient statistics* of the variables $\phi_l(x)$, for $l = 1, \ldots, d$, together with a parameter vector $\theta \in \mathbb{R}^d$ so that equation 2.2 for the probability of a state becomes

$$p_\theta(x) = p(x|\theta) = \frac{1}{Z(\theta)} \exp\left(\theta \cdot \phi(x)\right), \text{ or equivalently } p(x|\theta) = \exp\left(\theta \cdot \phi(x) - \log Z(\theta)\right).$$

A general form known as the *standard overcomplete representation* provides a sufficient statistic as an indicator function for every possible configuration of each subset $c \in C$, resulting in $d = \sum_{c\in C} \prod_{i\in c} \chi_i$. This simplifies some methods of analysis but in any overcomplete representation, there is redundancy in that multiple linear combinations $a\cdot\phi(x)$ are equal to a constant, and thus give rise to the same probability distribution. A concise alternative is a *minimal representation* where the parameter vector θ is unique for each distribution. In an overcomplete representation, as described in Section 2.2.1, any transformation $\theta \to \theta'$ such that for any $x \in \mathcal{X}, \theta' \cdot \phi(x) = \theta \cdot \phi(x)+$ a constant, is a reparameterization. The constant is absorbed into the new partition function, i.e. $Z(\theta') = Z(\theta)\exp(\text{constant})$, resulting in an identical probability distribution.

The representation in the exponential family reveals many insights, as discussed in (Wainwright and Jordan, 2008, §3). We provide a few key observations, on which we shall expand in Chapter 5.

2.3.1 Derivatives of $\log Z$

The log-potential function $\log Z(\theta)$ has several useful properties:

- The derivatives yield expected values of the sufficient statistics, specifically:

$$\frac{\partial \log Z(\theta)}{\partial \theta_l} = \mathbb{E}_\theta[\phi_l(x)] = \sum_{x \in X} \phi_l(x) p_\theta(x),$$
$$\frac{\partial^2 \log Z(\theta)}{\partial \theta_{l_1} \partial \theta_{l_2}} = \mathbb{E}_\theta[\phi_{l_1}(x)\phi_{l_2}(x)] - \mathbb{E}_\theta[\phi_{l_1}(x)]\mathbb{E}_\theta[\phi_{l_2}(x)]. \tag{2.8}$$

- $\log Z$ is a convex function of θ.

2.3.2 Maximum entropy

The log-linear representation has a remarkable property with respect to the expected values of the sufficient statistics: of all probability distributions consistent with a given set of expected values, it is the one with *maximum (Shannon) entropy*, where this entropy is defined by

$$S(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{2.9}$$

From this perspective, the θ parameters emerge in the optimization process as Lagrange multipliers associated with the constraints.

There is a deep relationship between entropy and the log-partition function, in that they are *conjugate duals* of each other, see (Wainwright and Jordan, 2008, §3.6.1) for details.

2.4 Binary Pairwise Models and The Ising Model

We now turn to the particular case of *binary pairwise* models, which form the main subject of this thesis.

A *binary* model is one in which every variable has just two states, i.e. $\chi_i = \{0, 1\}\ \forall i \in V$. A *pairwise* model is one where all potential functions are defined over at most two variables, i.e. $|c| \leq 2\ \forall c \in C$. Provided $n > 1$, note that potential functions over just one variable may always be absorbed or converted into a pairwise function over two variables. Hence, the topology of a pairwise model may be described by a simple graph (that is with no loops or double edges). A *planar* model is one whose topology is a planar graph. See Section 3.3 for definitions from graph theory.

Focusing on binary pairwise models may appear restrictive but the framework is rich enough to exhibit the behavior of a complex interacting system. These models play a key role in many

applications, both directly and as critical subroutines in solving more elaborate problems (Blake et al., 2011; Pletscher and Kohli, 2012). Further, in the sense described below, a general MRF may be converted into an equivalent binary pairwise model, though this may lead to an exponential increase in the state space. Recent work by Eaton and Ghahramani (2013) has clarified the nature of the relationship. We summarize their results:

- Binary pairwise MRFs are not *universal*, in the sense that there exist models that cannot be *simply* reduced to the binary pairwise form.
- However, pairwise MRFs (without the restriction of binary), binary 3-wise MRFs (potential functions over triplets of variables, i.e. arity 3), planar pairwise, and also planar binary MRFs, all *are* universal.
- Further, binary pairwise MRFs (even if restricted to planar) are *positive universal*, meaning that any general model where strictly $p(x) > 0 \; \forall x \in \mathcal{X}$ can be simply reduced to this form.
- Binary pairwise MRFs (even if restricted to planar) are *almost universal*, in the sense that one can construct a binary pairwise model that simply reduces any general model (even if it has configurations with 0 probability) to within an arbitrarily small approximation error.

2.4.1 Parameterizations, reparameterizations and associativity

For the bulk of this thesis, we shall be focused on binary pairwise MRFs with $p(x) > 0 \; \forall x \in \mathcal{X}$. Let $n = |V|$ be the number of variables, each of which takes values in $\mathbb{B} = \{0, 1\}$. Let $\mathcal{E}$ represent the edge relationships, that is $\mathcal{E} = \{c \in C : |c| = 2\}$. Let $m = |\mathcal{E}|$ be the number of edges. Let $\mathcal{N}(i) = \{j \in V : (i, j) \in \mathcal{E}\}$ be the neighbors of variable X_i.

As introduced in 2.3, one way to specify such a model is using the standard overcomplete representation, which here means a θ vector with $2n + 4m$ dimensions, that is θ is a vector with $2n$ dimensions, $(\theta_{1:0}, \theta_{1:1}, \ldots, \theta_{n:0}, \theta_{n:1})$, concatenated together with a four element vector $(\theta_{ij:00}, \theta_{ij:01}, \theta_{ij:10}, \theta_{ij:11})$ for each edge $(i, j) \in \mathcal{E}$. Alternatively, we may write the parameters in functional form as: $\theta_i(a) = \theta_{i:a}$ for $i \in V, a \in \mathbb{B}$, which we term the *singleton potentials*; and $\theta_{ij}(a, b) = \theta_{ij:ab}$ for $(i, j) \in \mathcal{E}$ and $a, b \in \mathbb{B}$, which we term the *edge* or *pairwise potentials*.

Reparameterizations were introduced in Section 2.2.1. One simple reparameterization is just to add or subtract a constant from any log-potential function. Hence, without loss of generality,

we may assume that $\theta_i(0) = 0 \ \forall i \in V$. The same idea can be used to require any one particular element of $\theta_{ij}(a, b)$ to be 0. We can go further using the following idea.

A *singleton transformation* is a change in one or more singleton potentials, together with a corresponding change to a pairwise potential which brings it to a convenient form. It is easily checked that a reparameterization of an edge potential via singleton transformations, $\begin{pmatrix} \theta_{ij:00} & \theta_{ij:01} \\ \theta_{ij:10} & \theta_{ij:11} \end{pmatrix} \rightarrow \begin{pmatrix} \theta'_{ij:00} & \theta'_{ij:01} \\ \theta'_{ij:10} & \theta'_{ij:11} \end{pmatrix}$ is valid if and only if $\theta_{ij:00}+\theta_{ij:11}-\theta_{ij:01}-\theta_{ij:10} = \theta'_{ij:00}+\theta'_{ij:11}-\theta'_{ij:01}-\theta'_{ij:10}$. Hence this one quantity, which we call the *associativity* or *weight* of the edge and write W_{ij}, may be used to describe the edge, and is invariant with respect to any singleton transformation (hence is well-defined).

If the associativity of an edge is positive, we describe that edge as *attractive* (equivalently *associative*, *ferromagnetic* or *regular*). This is equivalent to θ_{ij} for the edge being supermodular, or the edge having submodular cost function. In this case, the edge tends to pull the variables corresponding to its two end vertices toward the same value. If the associativity of an edge is negative, we describe it as *repulsive*, in which case it tends to push the variables corresponding to its two end vertices apart to different values. An edge with 0 associativity may be removed since we may transform its edge potential to the zero matrix. A binary pairwise model is attractive if and only if every one of its edges is attractive. Inference in attractive models is much easier than for the general case, as we shall see in Section 2.5.1.

Accordingly, all but one of the four values of an edge potential function may be taken to be 0. A minimal (unique) form which we shall use extensively in Part III uses parameters $\{\theta_i \ \forall i \in V; \ W_{ij} \ \forall (i, j) \in \mathcal{E}\}$ so that the energy of a configuration has the form

$$E(x) = -\sum_{i \in V} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} W_{ij} x_i x_j. \tag{2.10}$$

Although this form facilitates analysis, a disadvantage is that if W_{ij} is increased in order to increase the attractive pull between variables X_i and X_j, it also has the effect of increasing the probability that each variable will be equal to 1 rather than 0. Accordingly, a different reparameterization which we shall sometimes use, particularly when specifying input parameters of a model, instead makes

use of the symmetric form

$$E(x) = -\sum_{i \in V} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} \frac{W_{ij}}{2}[x_i x_j + (1 - x_i)(1 - x_j)], \tag{2.11}$$

where the associativity of an edge continues to be W_{ij}. It is easy to see that the reparameterization required to map from the form of (2.11) to that of (2.10) takes $\theta_i' \leftarrow \theta_i - \sum_{j \in \mathcal{N}(i)} W_{ij}$, where $\mathcal{N}(i)$ is the set of neighbors of X_i, while leaving W_{ij} unchanged.

2.4.2 The Ising model

In the early 1920s, Ernst Ising was studying phase transitions in statistical physics, specifically the macro magnetic properties of materials as the temperature is varied, under the supervision of his PhD advisor, Wilhelm Lenz. In a rare turn of events, what has come to be named the Ising model was in fact proposed by his advisor. As described by Prof Barry Simon:[4]

> *This model was suggested to Ising by his thesis adviser, Lenz. Ising solved the one-dimensional model, . . . , and on the basis of the fact that the one-dimensional model had no phase transition, he asserted that there was no phase transition in any dimension. As we shall see, this is false. It is ironic that on the basis of an elementary calculation and erroneous conclusion, Ising's name has become among the most commonly mentioned in the theoretical physics literature. But history has had its revenge. Ising's name, which is correctly pronounced "E-zing," is almost universally mispronounced "I-zing."*

In the classical Ising model, there is a system of n magnetic atoms with spins $\sigma_i \in \{-1, +1\}$. Neighboring atoms, as specified by an edge set $\mathcal{E}$, influence each other *ferromagnetically*, meaning that the energy of a configuration is lower by a positive amount βJ for each pair of neighboring atoms that are aligned with the same spin, where $\beta = \frac{1}{kT}$ as in (2.3). In addition there is an *external field* with strength h so that the total energy is given by

$$E = -\beta \left(J \sum_{(i,j) \in \mathcal{E}} \sigma_i \sigma_j + h \sum_i \sigma_i \right).$$

Physicists are interested in identifying the *ground state*, which is the lowest energy state, of the system, which corresponds to MAP inference. They also derive much useful information from computing the partition function.

[4]This quote appears at http://math.arizona.edu/ tgk/541/chap1.pdf .

The problem can be generalized to allow different J_{ij} parameters for each edge interaction, and individual *local fields* h_i for each atom. It is easy to see that this may be mapped via $W_{ij} = 4\beta J_{ij}, \theta_i = 2\beta(h_i - \sum_{j \in \mathcal{N}(i)} J_{ij})$ into the canonical binary pairwise MRF form of (2.10).

In statistical physics, a topic of great interest is to identify possible *phase transitions* in macro behavior of the system as the temperature T, or other parameters are varied. Often the typical behavior over ensembles of systems, as the size tends to infinity, is studied. The Ising model led to a prediction of a phase transition that was later experimentally observed (Nobel prize to Onsager).

In contrast, in machine learning, computer science and combinatorics, we are typically interested in the properties of individual, finite systems. Even in these settings, however, phase transitions in behavior can be observed as parameters vary. Examples of work in this area include (Kanefsky and Taylor, 1991; Hogg et al., 1996; Zhang, 2004; Coppersmith et al., 2004). We provide a novel perspective on one form of this phenomenon in Section 7.5.

2.5 Related Problems and Complexity Results

Computing Z belongs to the class of counting problems #P (Valiant, 1979). A *fully polynomial-time randomized approximation scheme* (FPRAS) was derived for binary pairwise models by Jerrum and Sinclair (1993), but only when singleton potentials are uniform (i.e. a uniform external field), and the resulting runtime is high at $O(\epsilon^{-2} m^3 n^{11} \log n)$. Marginal inference is NP-hard (Cooper, 1990), even to approximate (Dagum and Luby, 1993). The MAP problem is typically easier, yet is still NP-hard (Shimony, 1994), even to approximate (Abdelbar and Hedetniemi, 1998).

MAP inference may be reduced to finding a *maximum weight stable set* (MWSS) in a derived weighted graph (Sanghavi et al., 2009; Jebara, 2009), as we shall explore in Part II. Further, the MWSS problem may easily be reduced to MAP inference on a binary pairwise MRF. Hence, binary pairwise MRFs are universal for MAP. This was also demonstrated by showing that optimization of pseudo-Boolean functions may be reduced to optimization of quadratic pseudo-Boolean functions (Boros and Hammer, 2002).

MAP inference is also easily seen to be at least as general as the classic NP-hard MAXCUT problem (Karp, 1972). Each variable is a node and is assigned to a partition based on its value. Goemans and Williamson (1994) provided a polynomial-time approximation algorithm based on a

semidefinite program relaxation guaranteed to be within a ratio of 0.878 of the optimum solution. Khot et al. (2007) showed that if the unique games conjecture is true, then this result is optimal. However, if the model is attractive (all edge weights are positive) then the problem is a MINCUT problem, solvable in polynomial time. Further, MAXCUT is solvable in polynomial time for a planar graph by a reduction to the matching problem. This applies to planar binary pairwise MRFs with no local fields (singleton potentials). Further, for this subclass of models, the partition function may be computed in polynomial time (Kastelyn, 1963; Fisher, 1966; Globerson and Jaakkola, 2006; Schraudolph and Kamenetsky, 2009). Since singleton potentials may be emulated by adding an extra variable with appropriate pairwise terms, this shows that problems with such singleton potentials are tractable provided the resulting graph is planar. Hence, in particular, models with *outerplanar* graphs are tractable (Batra et al., 2010)[5]. However, Barahona (1982) demonstrated that even MAP inference on general planar binary pairwise models (with arbitrary singleton potentials) is NP-hard via a reduction to planar MWSS, which is NP-hard (Garey and Johnson, 1979).

Inference may be performed in polynomial time using the *junction tree algorithm* (Lauritzen and Spiegelhalter, 1988a; Lauritzen, 1996; Cowell et al., 1999) provided the *treewidth* of the model's graph is bounded (the runtime is exponential in the treewidth). The treewidth may be defined to be one less than the optimal (minimum) cardinality of a maximum clique in a triangulation of the graph.[6] Further, under mild assumptions, this was shown to be the only restriction which will allow efficient inference for any potential functions (Chandrasekaran et al., 2008).

2.5.1 Models with submodular energies

In order to identify subclasses of problems that may be solved in polynomial time, restrictions may be placed either on the nature of the potential functions, or on the topology of the model. An important subclass of MRFs are those that restrict the log-potential functions to be supermodular, or equivalently, to have energy (cost) functions which are *submodular*.

For V a finite set, let $\mathcal{P}(V)$ be its power set (the set of all subsets). A set function $f : \mathcal{P}(V) \to \mathbb{R}$

[5]Since outerplanar graphs have treewidth (see next paragraph for a definition) at most two, this result is less strong than it may appear.

[6]The concept of treewidth was introduced by Halin (1976) while investigating the Hadwiger number. Later, it was rediscovered and presented by Robertson and Seymour (1984), after which it came into widespread use.

is submodular iff $\forall A, B \subseteq V, f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$. An equivalent definition captures the idea of 'diminishing returns': $f : \mathcal{P}(V) \to \mathbb{R}$ is submodular iff $\forall A \subset B \subset V, x \in V \setminus B, f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B)$.

This may be generalized to a *lattice*, which is a partially ordered set L in which any two elements $a, b \in L$ have a *join* (or least upper bound) $a \vee b$, and a *meet* (or largest lower bound) $a \wedge b$. A function on a lattice is submodular iff $\forall a, b \in L, f(a \vee b) + f(a \wedge b) \leq f(a) + f(b)$. A function f is supermodular iff $-f$ is submodular.

Submodular functions behave in some ways like convex functions, and in some ways like concave functions. They have attracted attention in combinatorics (Lovász, 1983), economics (Topkis, 1998; Milgrom and Roberts, 1990), and increasingly in machine learning (Bach, 2013; Bilmes, 2014).

In our context, a pairwise multi-label function on a set of ordered labels $X_{ij} = \{0, \ldots, k_i - 1\} \times \{0, \ldots, k_j - 1\}$ is submodular iff

$$\forall x, y \in X_{ij},\ f(x \wedge y) + f(x \vee y) \leq f(x) + f(y) \tag{2.12}$$

where for $x = (x_1, x_2)$ and $y = (y_1, y_2)$, $(x \wedge y) = (\min(x_1, y_1), \min(x_2, y_2))$ and $(x \vee y) = (\max(x_1, y_1), \max(x_2, y_2))$. For binary variables this is equivalent to the edge being attractive (see Section 2.4.1). A function over more than two variables is submodular iff every projection onto any two variables is submodular.

While this is a subclass of all general models, it is still rich enough to be of great interest, and has direct application in areas where it is reasonable to have a prior that neighboring variables will take similar values, such as image denoising (Greig et al., 1989). Further, methods have been explored which decompose a general model into a number of submodular problems on the same topology (Osokin et al., 2011).

MAP inference is solvable in $O(n^3)$ time for attractive binary pairwise models, for example via graph cuts (Greig et al., 1989; Goldberg and Tarjan, 1988), though note that marginal inference and computing Z are not (Jerrum and Sinclair, 1993).

Given the tractability of attractive binary pairwise models, i.e. binary pairwise models with submodular cost functions, much work has focused on understanding which other models can be reduced to these. Building on the 'battleship' construction of Ishikawa (2003), a key paper by

Schlesinger and Flach (2006) showed that this can be done for any *pairwise multi-label* model with submodular cost functions, which is important for our work in Chapter 6.10. Observe that the definition of submodularity relies on a particular ordering of the labels for each variable. Schlesinger (2007) introduced the notion of *permuted submodular* to mean a model where there exists some permutation of the labels of variables such that the resulting cost functions are submodular; and further demonstrated that testing for the existence of such a permutation, and finding one if it exsits, may be performed in polynomial time.

Zivny et al. (2009) showed that models with submodular cost functions of arity 3 (i.e. potential functions over triplets) can always be mapped to the attractive binary pairwise case, but not models with submodular functions of arity 4 or higher, unless other conditions are also satisfied. We reach a related, similar, conclusion in Section 4.6. However, Arora et al. (2012) recently demonstrated a novel graph cuts method for submodular cost functions of any order over binary variables, though the time is exponential in the order of the potentials.

2.6 Approaches to Analyzing Graphical Models

There is a vast literature on approaches to analyzing graphical models. Here we provide a very brief history to provide context.

The junction tree algorithm (Pearl, 1988; Lauritzen and Spiegelhalter, 1988a; Lauritzen, 1996; Cowell et al., 1999) is a dynamic programming approach which may be used either in sum-product form for marginal inference, or max-product form for MAP inference.[7] The sum-product form will return the true marginal distributions over all factors, and the max-product form is guaranteed to return a correct MAP assignment, but in either case, the model's topology must first be triangulated, then messages passed between resulting cliques, leading to runtime that is exponential in the treewidth (see Section 2.5). A popular variant is simply to omit the triangulation phase, and proceed using cliques of size 2, i.e. each edge of the model. This approach is termed *belief propagation* (BP). If the topology is a tree then this is efficient and exact. If not, then messages pass around cycles and the algorithm is known as *loopy belief propagation* (LBP), which often produces excellent results (McEliece et al., 1998; Murphy et al., 1999) but in general has no guarantees on accuracy or

[7] More generally, the algorithm may be used for any commutative semi-ring.

even convergence.

Historically, close variants of belief propagation were derived independently in different disciplines. These include the Kalman filter for signal processing (Kalman, 1960), the Viterbi algorithm (Viterbi, 1967), decoding algorithms for recent error-correcting codes including low-density parity check codes (Gallager, 1962) and turbo codes (Berrou and Glavieux, 1996; MacKay and Neal, 1996; McEliece et al., 1998), and the transfer-matrix method in statistical mechanics (Baxter, 1982).

A different approach for MAP inference considers instead the appropriate integer program. This is then typically relaxed in two ways: (i) the integer program is relaxed to a linear program (LP); and (ii) rather than optimize over the space of all valid distributions, termed the *marginal polytope*, instead this is relaxed to the *local polytope*, which enforces only pairwise (rather than global) consistencies and is computationally easier to handle. As shown by Yedidia et al. (2001), this LP relaxation is in fact closely connected to the LBP algorithm. Indeed, the solution to the dual of the LP is a fixed point of LBP, which can therefore be considered an efficient and easily parallelizable computational approach to try to solve the dual problem.

The dual problem can be solved by other methods such as block coordinate descent, which leads to the max-product linear programming (MPLP) algorithm (Globerson and Jaakkola, 2007). This is guaranteed to converge monotonically (though perhaps to a local optimum). This idea was developed further by Sontag and Jaakkola (2009), where the coordinate descent is performed on spanning trees of the graphical model, leading to improved performance.

2.6.1 Variational methods for marginal inference

The *Kullback-Leibler divergence* (KL-divergence) between two discrete probability distributions $q(x), p(x)$, is defined by $D(q||p) = \sum_x q(x) \log \frac{q(x)}{p(x)}$. It is easily shown by Jensen's inequality that $D(q||p) \geq 0$ with equality iff $q = p$. Consider (2.2) which gives the true distribution p. For any distribution $q(x)$, let $S(q(x))$ be its Shannon entropy. It is easily seen that

$$0 \leq D(q||p) = \sum_x q(x) \log \frac{q(x)}{p(x)} = -\sum_{c \in C} \mathbb{E}_q \left(\psi_c(x_c)\right) + \log Z - S(q(x)).$$

Hence $\log Z = \max_q \sum_{c \in C} \mathbb{E}_q \left(\psi_c(x_c)\right) + S(q(x))$ or equivalently $-\log Z = \min_q \mathcal{F}_G(q)$, where $\mathcal{F}_G(q) = \mathbb{E}_q \left(-\psi_c(x_c)\right) - S(q(x))$ is the expected energy minus the entropy of the distribution, termed the (Gibbs) *free energy*, with the optimum occurring when $q = p$, i.e. the true distribution.

This optimization is to be performed over all valid probability distributions, that is over the marginal polytope. However, this problem is intractable. Various approximations have been introduced, most famously the Bethe approximation, as discussed in Part III, which relaxes the marginal polytope to the local polytope, and uses the Bethe entropy approximation (which in general is neither an upper nor lower bound on the true entropy, though it is exact for acyclic models).

If instead, a concave upper bound on the entropy is employed, a concave upper bound on the log-partition function is obtained, which may be efficiently optimized to yield an upper bound on the true partition function. A well-known example is the *tree-reweighted* approximation (TRW) (Wainwright et al., 2005; Wainwright and Jordan, 2008).

For both MAP and marginal inference, efficient ways to optimize over tighter relaxations of the marginal polytope have been explored (Sontag and Jaakkola, 2007; Sontag et al., 2008; Sontag, 2010). The practicalities of dealing with this for the Bethe approximation, and the associated potential benefits and drawbacks, are addressed in Chapter 7.

Part II

Exact MAP Inference by MWSS on Perfect Graphs

This Part builds on methods introduced by Jebara (2009) and Sanghavi et al. (2009), and is based on work that appeared in (Weller and Jebara, 2013b).

Chapter 3

Additional Background

3.1 MAP Inference and Tractable Cases

Finding MAP assignments of MRFs has been an intense area of research for many years. An early example is the Viterbi algorithm for hidden Markov models (Viterbi, 1967). The belief propagation algorithm and junction tree generalizations (Pearl, 1988; Lauritzen and Spiegelhalter, 1988a) showed that the topological restriction of bounded treewidth allows polynomial time inference. Further, under mild assumptions, this was shown to be the only restriction which will allow efficient inference for any score functions (Chandrasekaran et al., 2008). Max-product belief propagation on graphs with cycles proved to be extremely helpful in the context of turbo-decoding (McEliece et al., 1998).

Binary pairwise graphical models with more general (and often dense) topologies yet whose potentials are all attractive were shown to be solvable efficiently using graph-cuts or network flow (Greig et al., 1989; Goldberg and Tarjan, 1988).[1] More recently, MAP estimation for graphical models with cycles involving matching and b-matching problems[2] was shown to be solvable efficiently using the max-product algorithm (Bayati et al., 2005; Huang and Jebara, 2007; Sanghavi et al., 2008; Bayati et al., 2008). In previous work, these known cases were all shown to compile to a maximum weight stable set problem on a perfect graph, which is known to be solvable in

[1] This result has recently been generalized to higher order potentials by Arora et al. (2012).

[2] These graphical models involve topological constraints as well as various constraints on the potential functions (not simply associativity or submodularity).

polynomial time (Jebara, 2009, 2014). This Part of the thesis derives new results for this approach, first described in (Jebara, 2009; Sanghavi et al., 2009), and examines which other models may be handled in this manner, see Chapter 4.

An earlier method examining triangulated[3] micro-structure graphs was presented (Jégou, 1993) in the context of constraint satisfaction problems (CSPs). Valued CSPs (VCSPs) use soft constraints with explicit costs, and are closely related to MAP inference problems, see (Dechter, 2003) for a survey. Many techniques have been developed, including optimal soft arc consistency (Cooper et al., 2010), belief propagation (Weiss et al., 2007) and linear program relaxations (Sontag et al., 2008), which may be considered to proceed through identifying helpful reparameterizations (see Section 3.7).

In general, many different methods are available, see Kappes et al. (2013) for a recent survey. Some, such as dual approaches, may provide a helpful bound even if the optimum is not found.

3.2 Notation and Preliminaries

As described in Part I, we shall consider only discrete, finite MRFs (V, Ψ), which may be specified by a set of n variables $V = \{X_1, \ldots, X_n\}$ together with (log) potential functions over subsets c of V, $\Psi = \{\psi_c : c \in C \subseteq \mathcal{P}(V)\}$, where $\mathcal{P}(V)$ is the powerset of V. Each variable X_i may take finite k_i possible values which we label $\{0, \ldots, k_i - 1\}$. Write $x = (x_1, \ldots, x_n)$ for one particular complete configuration and x_c for a configuration just of the variables in c. A potential function ψ_c maps each possible setting x_c of its variables c to a real number $\psi_c(x_c)$.

The probability of each configuration x is given by the equation below, which also defines the notion of *energy* $E(x)$,[4]

$$p(x) = \frac{e^{\sum_{c \in C} \psi_c(x_c)}}{Z} = \frac{e^{-E(x)}}{Z},\ E = -\sum_{c \in C} \psi_c(x_c), \tag{3.1}$$

[3]Triangulated, or chordal, graphs are a subclass of perfect graphs.

[4]Throughout this Part, we assume $p(x) > 0\ \forall x$, with finite $\psi_c(x_c)$ terms. There are reasonable distributions where this does not hold, i.e. distributions where $\exists x : p(x) = 0$, but this can often be handled by assigning such configurations a sufficiently small positive probability ϵ. Also *cost* functions are the negative of our ψs, thus submodular cost functions are equivalent to supermodular ψs.

where the *partition function* $Z := \sum_x e^{-E(x)}$ is the normalizing constant to ensure that probabilities sum to 1.

Identifying a configuration of variables that is most likely, termed *maximum a posteriori* or MAP inference, is very useful in many contexts, yet in general is NP-hard (Shimony, 1994). Given (3.1), it is equivalent to *energy minimization*. In our notation this is the combinatorial problem of identifying

$$x^* \in \operatorname*{arg\,max}_{x=(x_1,\ldots,x_n)} \sum_{c \in C} \psi_c(x_c). \tag{3.2}$$

In general, an MRF may be considered a hypergraph together with associated ψ_c functions (see Section 3.3 for definitions). A popular alternative representation is a factor graph, which is a bipartite graph where the variables V form one stable partition and each $c \in C$ is a node in the other partition, with an edge from c to each variable it contains. In the special case that all variables X_i take values only in $\mathbb{B} = \{0, 1\}$, the model is said to be *binary*. If $|c| \leq 2\ \forall c \in C$ then the model is *pairwise*. Binary pairwise models play a key role in computer vision both directly and as critical subroutines in solving more complex problems (Pletscher and Kohli, 2012). Further, it is possible to convert a general MRF into an equivalent binary pairwise model, see Section 2.4.

3.3 Terms from Graph Theory

We follow standard definitions and omit some familiar terms, see (Diestel, 2010). For further information on related topics from graph theory, including a sketch of recent work on claw-free graphs, and on recognizing perfect graphs, see Appendix A.

A *graph* $G(V, E)$ is a set of vertices V, and edges $E \subseteq V \times V$. Let $n = |V|$ and $m = |E|$. Throughout this paper, unless otherwise specified, all graphs are finite and *simple*, that is a vertex may not be adjacent to itself (no loops) and each pair of vertices may have at most one edge (no multiple edges).

The *complete* graph on n vertices, written K_n, has all $\binom{n}{2}$ edges. A *path* of length n is a graph P_n with n edges connecting $n + 1$ vertices as $v_1 - v_2 - \cdots - v_n - v_{n+1}$. An *induced subgraph* $H(U, F)$ of a graph $G(V, E)$ is a graph on some subset of the vertices $U \subseteq V$, inheriting all edges with both ends in U, so $F = \{(v, w) \in E : v, w \in U\}$. The union of two subgraphs, $H_1(V_1, E_1)$ and $H_2(V_2, E_2)$ of a graph $G(V, E)$, written $H_1 + H_2$, is the induced subgraph of G on $V_1 \cup V_2$.

A *hypergraph* (V, E) is a generalization of a graph where the elements of E are any non-empty subsets of V, not necessarily of size two. A general MRF may be regarded as a hypergraph (V, C) together with functions $\{\psi_c\}$ $\forall c \in C$. For the special case of a pairwise model, the structural relationships are naturally interpreted as a graph.

A graph is *connected* if there is a path connecting any two vertices. A *cut vertex* of a connected graph G is a vertex $v \in V$ such that deleting v disconnects G. A graph is *2-connected*, equivalently *biconnected*, if it is connected and contains no cut vertex. A *block* is a maximal connected subgraph with no cut vertex of the subgraph. Every block is either K_2 (two vertices joined by an edge) or a maximal 2-connected subgraph containing a cycle. Different blocks of G overlap on at most one vertex, which must be a cut vertex. Hence G can be written as the union of its blocks with every edge in exactly one block. These blocks are connected without cycles in the *block tree* for each connected component of G.

A *cutset* S of a graph G is a set of vertices $S \subseteq V(G)$ s.t. $G \setminus S$ is disconnected. A *star-cutset* S of G is a cutset s.t. $\exists$ some $x \in S$ s.t. x is complete to $S \setminus \{x\}$.

A *stable* set in a graph is a set of vertices, no two of which are adjacent. A *weighted graph* (V, E, w) is a graph with a nonnegative real value for each vertex, called its *weight* $w(v)$. A *maximum weight stable set (MWSS)* is a stable set with maximum possible weight. A *maximal maximum weight stable set (MMWSS)* is a MWSS of maximal cardinality (this is useful in our context since, after reparameterization, we may have many nodes with 0 weight, see Sections 3.6 and 3.7).

A *clique* in a graph is a set of vertices, of which every pair is adjacent. The *clique number* of a graph G, written $\omega(G)$, is the maximum size of a clique in G.

The *complement* of a graph $G(V, E)$ is the graph $\bar{G}(V, F)$ on the same vertices with an edge in F iff it is not in E. Hence a clique is the complement of a stable set and vice versa.

A *coloring* of a graph is a map from its vertices to the integers (considered the colors of the vertices) such that no two adjacent vertices share the same color. The *chromatic number* of a graph G, written $\chi(G)$, is the minimum number of colors required to color it. Observe that clearly $\chi(G) \geq \omega(G)$ for any graph G.

A graph G is *perfect* iff $\chi(H) = \omega(H)$ for all induced subgraphs H of G. As examples, any bipartite or chordal graph is perfect. Related concepts (see Theorem 3.5.5) are: a *hole* in a graph G is an induced subgraph which is a cycle of length ≥ 4 (note this means the cycle must be chordless);

an *antihole* is an induced subgraph whose complement is a hole. A hole or antihole is *odd* if it has an odd number of vertices. Note that, as a special case, a hole with 5 vertices is isomorphic to an antihole of the same size. It is easily shown that odd holes and antiholes are not perfect. A graph is *Berge* if it contains no odd holes or antiholes (equivalently, if neither the graph nor its complement contains an odd hole).

3.4 Further Terms

This Section may be skipped on a first reading, and referred to later for definitions.

A *nand Markov random field* (NMRF, Jebara, 2009) may be considered a particular kind of binary pairwise MRF, where each edge potential enforces a nand operation. Each edge log-potential is of the form $\psi_{ij}(X_i, X_j) = \begin{cases} -\infty & X_i = X_j = 1 \\ 0 & \text{otherwise} \end{cases}$. Hence, when considered in its exponentiated form as a potential function, see equations (2.1) and (2.2), $\pi_{ij}(X_i, X_j) = \begin{cases} 0 & X_i = X_j = 1 \\ 1 & \text{otherwise} \end{cases}$. A MAP configuration of the NMRF must have at most one variable from each edge set to 1, hence is equivalent to finding a maximum weight stable set (MWSS).

A *clique group* for a set of variables c is a clique in an NMRF corresponding to all possible settings x_c of those variables of its MRF, see Section 3.6.

An *snode* is a node in an NMRF relating to a setting of a single variable from its MRF. Equivalently, it is a node from a clique group deriving from $c = \{X_i\}$ for some i. An *enode* is a node from a clique group deriving from some $c \in C$ with $|c| \geq 2$. For example, when considering binary pairwise models, an enode derives from an edge of the MRF.

Given a graph $G(V, E)$, its *line graph* L is the graph that takes E as its vertices, with $e, f \in E$ adjacent in L iff they share an end vertex in G.

For a graph (V, E), if $X \subseteq V$ and $v \in V \setminus X$ then v is *complete* to X if v is adjacent to every member of X. If $X, Y \subseteq V$ are disjoint, then X is *complete* to Y if every vertex in X is complete to Y.

A *cutset* S of a graph G is a set of vertices $S \subseteq V(G)$ such that $G \setminus S$ is disconnected. A *star-cutset* S of G is a cutset such that $\exists$ some $x \in S$ such that x is complete to $S \setminus \{x\}$.

A *signed graph* (Harary, 1953) is a graph (V, E) together with one of two possible signs for each edge. This is a natural structure when considering binary pairwise models, where we characterize edges as either *associative* or *repulsive*, see Section 3.8. When discussing signed graphs, we use the notation $\oplus$ to show an associative edge, and $\ominus$ for a repulsive edge. For example, $x \oplus y \ominus z$ is a graph with 3 vertices x, y and z, and two edges, where x and y are adjacent via an associative edge, and y and z are adjacent via a repulsive edge.

A *frustrated cycle* in a signed graph is a cycle with an odd number of repulsive edges.

A B_R structure (see Figure 3.1 for an example) is a signed graph over variables V with associative edges E_A and repulsive edges E_R such that (V, E_R) is bipartite and $\exists$ a disjoint bipartition $V = V_1 \cup V_2$ with all E_R crossing between the partitions $V_1 - V_2$, and no E_A crossing between them. Either E_A or E_R may be empty, so for example, this includes any signed graph with only associative edges. See Lemma 4.5.1 for equivalent definitions.

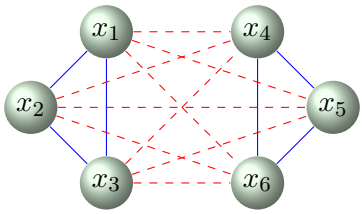


Figure 3.1: An example B_R structure. Solid blue (dashed red) edges are associative (repulsive). Deleting any edges maintains the B_R property.

A $T_{m,n}$ structure (see Figure 3.2 for an example) is a 2-connected signed graph containing $m + n \geq 1$ triangles on a common base given by: 2 base vertices s, t connected via a repulsive edge, so $s \ominus t$; together with $m \geq 0$ vertices r_i, each adjacent only to s and t via repulsive edges, so $s \ominus r_i \ominus t$; and $n \geq 0$ vertices a_i, each adjacent only to s and t via associative edges, so $s \oplus a_i \oplus t$. Note $T_{m,n}$ would be bipartite, with $\{s, t\}$ as one partition and all other vertices in the other, except that we have the repulsive edge $s \ominus t$.

A U_n structure (see Figure 3.3 for an example) is a 2-connected signed graph containing $n \geq 1$ triangles on a common base given by: 2 base vertices s, t connected via an associative edge, so $s \oplus t$; together with $n \geq 1$ vertices v_i, each adjacent only to s and t via one associative and one repulsive edge (either way), so either $s \oplus v_i \ominus t$ or $s \ominus v_i \oplus t$.

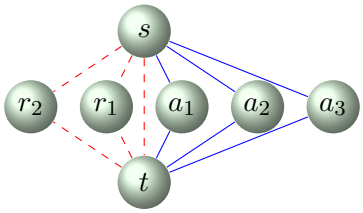


Figure 3.2: An example $T_{m,n}$ structure with $m = 2$ and $n = 3$. Solid blue (dashed red) edges are associative (repulsive).

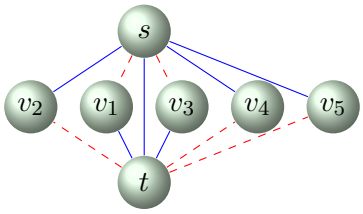


Figure 3.3: An example U_n structure with $n = 5$. Solid blue (dashed red) edges are associative (repulsive).

Note that U_1 is the same as $T_{0,1}$ but this is the only overlap. In Lemma 4.5.5, we show that, subject to the singleton node assumption of Section 4.1, $T_{m,n}$ and U_n structures are the only 2-connected signed graphs containing a frustrated cycle that map to a perfect NMRF.

3.5 Properties of Perfect Graphs

3.5.1 Complexity of MWSS

Our approach to MAP inference is to reduce the problem to finding a maximum weight stable set on a derived weighted graph, as described in Section 3.6. This is helpful only if we can find a MWSS efficiently, yet in general this is still an NP-hard problem for a graph with N vertices. However, if the derived graph is perfect[5], then a MWSS may be found in polynomial time via the ellipsoid method (Grötschel et al., 1984).

Faster exact methods (Yildirim and Fan-Orzechowski, 2006) based on semidefinite program-

[5]There are a few other classes of graphs that also admit efficient MWSS, such as *claw-free* graphs, where significant recent advances have been made (Faenza et al., 2011), but so far these have not been useful in analyzing MRFs.

ming are possible in $O(N^6)$ and are improved using primal-dual methods (Chan et al., 2009). Alternatively, linear programming can solve MWSS problems but requires $O(N^3\sqrt{n_K})$ time where n_K is the number of maximal cliques in the graph (Jebara, 2009, 2014). Clearly, whenever n_K is small, linear programming can be more efficient than semidefinite programming. However, in the worst case, n_K may be exponentially large in N which makes linear programming useful only in some cases. Message-passing methods can also be applied for finding the maximum weight stable set in a perfect graph though they too become inefficient for graphs with many cliques (Foulds et al., 2011; Jebara, 2014).

Where other methods exist for solving exact MAP inference, the reduction to MWSS is typically not the fastest method, yet there is hope for improvement since the field is advancing rapidly, with significant breakthroughs in recent years (Chudnovsky et al., 2006; Faenza et al., 2011).

3.5.2 Other properties

There is a rich literature on perfect graphs. We highlight key results used later in this thesis. See Appendix A for more background on related graph theory.

Theorem 3.5.1 (Gallai, 1962)**.** *The graph obtained by pasting two perfect graphs on a clique is perfect.*

Theorem 3.5.2 (Chvátal, 1985)**.** *The graph obtained by pasting two perfect graphs on a star-cutset is perfect.*

Theorem 3.5.3 (Substitution Lemma, Lovász, 1972)**.** *The graph obtained by substituting one perfect graph for a vertex of another perfect graph is also perfect.*

Here, substituting H for x in G means deleting x and joining every vertex of H to those vertices of G which were adjacent to x.

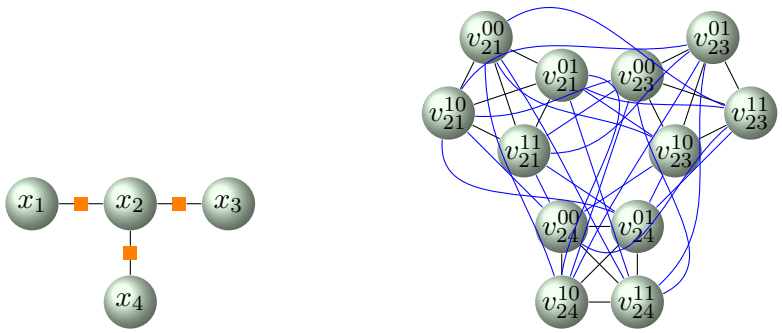
Theorem 3.5.4 (Weak Perfect Graph Theorem, Lovász, 1972)**.** *A graph is perfect iff its complement is perfect.*

Theorem 3.5.5 (Strong Perfect Graph Theorem 'SPGT', Chudnovsky et al., 2006)**.** *A graph is perfect iff it contains no odd hole or antihole (equivalently, iff it is Berge; equivalently, iff neither the graph nor its complement contains an odd hole).*

3.6 Reduction of MAP inference to MWSS on a NMRF

Given an MRF model (V, Ψ), construct a *nand Markov random field (NMRF)*, see Jebara (2009):

- A weighted graph $N(V_N, E_N, w)$ with vertices V_N, edges E_N and a weight function $w : V_N \to \mathbb{R}_{\geq 0}$.
- Each $c \in C$ of the original model maps to a *clique group* of N which contains one node for each possible configuration x_c, all pairwise adjacent.
- Nodes in N are adjacent iff they have inconsistent settings for any variable X_i.
- Nonnegative weights of each node in N are set as $\psi_c(x_c) - \min_{x_c} \psi_c(x_c)$, see Section 3.7 for an explanation of the subtraction.



(a) Input MRF (b) Derived NMRF

Figure 3.4: An example of mapping an MRF with binary variables (shown as a factor graph) to an NMRF (subscripts denote the factor variables c and superscripts denote the configuration x_c; black edges are between nodes in the same clique group, blue edges go between different clique groups).

See Figure 3.4 for an example. Jebara (2014) proved that a maximal cardinality set of consistent configuration nodes in N with greatest total weight, i.e. a MMWSS of N (see Section 3.3), will identify a globally consistent configuration of all variables of the original MRF that solves the MAP inference problem (3.2).

Sketch proof: (Slightly different to Jebara (2014), this will allow us to extend the result after discussing pruning in Section 3.7.) A MMWSS S is consistent by construction and clearly contains at most one node from each clique group. It remains to show it has at least one from each clique group. Suppose a clique group has no representative. Identify a member of this group which could

be added to S, establishing a contradiction since S is maximal, as follows: the group overlaps with some variables of S, copy the settings of these; for all other variables in the group, pick any setting. Note that if we do not insist on a maximal MWSS, it is possible that we do not get a representative for some clique groups and hence do not obtain a complete MAP configuration for the initial MRF.

3.7 Reparameterizations and Pruning

A *reparameterization* is a transformation

$$\{\psi_c\} \rightarrow \{\psi'_c\} \text{ such that } \forall x, \sum_{c \in C} \psi_c(x_c) = \sum_{c \in C} \psi'_c(x_c) + \text{constant}.$$

This clearly does not modify (3.2) but can be helpful to simplify the problem.

One particular reparameterization is to add a constant just to any ψ_c function, since any consistent configuration has exactly one setting for each group of variables c. Hence we may subtract the minimum $\psi_c(x_c)$ and assume that in each clique group of N, the minimum weight of a node is exactly zero. The earlier reduction result in Section 3.6 holds provided we insist on a *maximal* MWSS (MMWSS). To find a MMWSS, it is sufficient first to remove or *prune* the zero weight nodes, find a MWSS on the remaining graph, then reintroduce a maximal number of the zero weight nodes while maintaining stability of the set. Different reparameterizations will yield different pruned NMRFs. By the earlier argument: MWSS will find one member from each of some of the clique groups, then we can always find one of the zero weight nodes to add from each of the remaining groups using any greedy method. Hence we have shown the following result.

Lemma 3.7.1. *MAP inference on an MRF is tractable provided $\exists$ an efficiently identifiable efficient reparameterization such that the MRF maps to a perfect pruned NMRF.*

3.8 Singleton Transformations, Binary Pairwise MRFs and Associativity

Another useful reparameterization is what we term a *singleton transformation*, which is a change in one or more ψ functions for a single variable, with corresponding changes to a higher order term which brings it to a convenient form.

Considering binary pairwise models only, it is easily shown that a reparameterization of an edge via singleton transformations, $\begin{pmatrix} \psi_{00} & \psi_{01} \\ \psi_{10} & \psi_{11} \end{pmatrix} \rightarrow \begin{pmatrix} \psi'_{00} & \psi'_{01} \\ \psi'_{10} & \psi'_{11} \end{pmatrix}$ is valid iff $\psi_{00} + \psi_{11} - \psi_{01} - \psi_{10} = \psi'_{00} + \psi'_{11} - \psi'_{01} - \psi'_{10}$. Hence this quantity, which we call the *associativity* of the edge, is invariant with respect to any singleton transformation, and thus is well defined.

We describe an edge as either *associative*[6], in which case it tends to pull its two end vertices toward the same value, or *repulsive*, in which case it tends to push its two end vertices apart to different values, according to whether its associativity is > 0 or < 0. An edge with 0 associativity may be removed since we may transform its edge potential to the zero matrix. A binary pairwise model is associative iff every one of its edges is associative.

An associative edge may be reparameterized such that three of its entries are 0, and therefore may be pruned, leaving only either ψ'_{00} or ψ'_{11} (or both, though for our purposes of mapping to a perfect NMRF, it is always easier to prune more nodes) with a positive value. Similarly, we may reparameterize a repulsive edge $x \ominus y$ to leave only a $(x = 0, y = 1)$ or $(x = 1, y = 0)$ node.[7]

[6]Other equivalent terms used are *attractive*, *ferro-magnetic* or *regular*. This is equivalent to ψ for the edge being supermodular, or having submodular cost function.

[7]For repulsive edges, selecting one or other form is exactly analogous to choosing an *orientation* of the edge, $x \rightarrow y$ or $x \leftarrow y$. Further, such enodes from repulsive edges are adjacent iff their directed edges connect 'head to tail', hence the induced subgraph of an NMRF on these repulsive enodes is exactly a directed line graph of (V, E_R).

Chapter 4

New Results

Here we provide our main new results for the MWSS approach to MAP inference.

4.1 Singleton Clique Groups

Since typically we would like to allow any finite values for singleton potential functions, and singleton transformations as described in Section 3.8 without restriction, in some of this Chapter (specifically, in Sections 4.4 and 4.5) we assume that any NMRF includes the complete clique group for each of the single variables of its MRF.

In particular contexts, however, one may drop this requirement, and since this would remove nodes from the NMRF, it might be thought that this can only help to show perfection (since any induced subgraph of a perfect graph is perfect). Typically, however, if singleton nodes are removed through reparameterization (as described in Section 3.7), it is at the cost of adding additional edge nodes, such that the gain of fewer singleton nodes is more than offset by the disadvantage of adding extra edge nodes, which usually have more neighbors in the NMRF, and thus often leads to a greater risk of forming an odd hole or antihole in the NMRF. In addition, care must then be taken to confirm the decomposition result of Theorem 4.2.1. Nevertheless, in Appendix B we explore the possible benefits of relaxing the assumption of the presence of all singleton nodes, and observe that there are situations where it can be helpful.

4.2 New Results for All MRFs

Theorem 4.2.1 (MRF Decomposition)**.** *If $MRF_A(V_A, \Psi_A)$ and $MRF_B(V_B, \Psi_B)$ both map to perfect NMRFs N_A and N_B, and have exactly one variable s in common, i.e. $V_A \cap V_B = \{s\}$, with consistent ψ_s, then the combined $MRF'(V_A \cup V_B, \Psi_A \cup \Psi_B)$ maps to an NMRF N' which is also perfect. The converse is true by the definition of perfect graphs.*

Proof. [1] See Section 3.4 for notation. We may assume both Ψ_A and Ψ_B contain the same ψ_s forming the complete s clique group K_s in N_A and N_B (see Section 4.1, though in fact this Theorem holds more generally, provided only that both N_A and N_B have the same nodes from the clique group for s).

Let the possible values of s be $\{0, \ldots, k-1\}$, and s_i be the snode corresponding to $(s = i)$. Let A_i be all those vertices of $N_A \setminus \{s_i\}$ which have setting $s = i$, similarly define B_i for N_B. Observe that A_i is complete to A_j for all $i \neq j$, and similarly for B_i. N' is the result of pasting N_A and N_B on K_s, together with all edges from A_i to B_j if $i \neq j$.

Hence N' admits a star-cutset given by $X = K_s + A_0 + \cdots + A_{k-1} + B_0 + \cdots + B_{k-1}$ with s_0 complete to $X \setminus \{s_0\}$. Thus by Theorem 3.5.2, it is sufficient to show that $N_A + X$ and $N_B + X$ are each perfect. But this is true by Theorem 3.5.3, since $N_A + X = N_A + B_0 + \cdots + B_{k-1}$ may be obtained from N_A by substituting (via Theorem 3.5.3) $B_i + s_i$ for s_i, $i = 0, \ldots, k-1$; and similarly for N_B.

□

4.2.1 Block decomposition

Theorem 4.2.1 is a powerful tool for analyzing MRFs of any order and number of labels. As a special case, we have an immediate corollary.

Theorem 4.2.2. *A pairwise MRF maps to a perfect NMRF for all valid ψ iff each of its blocks maps to a perfect NMRF.*

This provides an elegant way to derive a previous result (Jebara, 2009):

[1] This proof, due to Maria Chudnovsky, is shorter and neater than the authors' original.

Theorem 4.2.3. *A pairwise MRF whose graph structure is a tree (i.e. no cycles) maps to a perfect NMRF.*

Proof. By Theorem 4.2.2, we need only consider one edge together with its two end vertices (then use induction). The edge clique group together with each one of the singleton clique groups is the complement of a bipartite graph, hence is perfect (by Theorem 3.5.4). Now paste the two together on the edge clique group to show the whole is perfect (by Theorem 3.5.1). □

We show the following further general result.

Lemma 4.2.4. *Neither a hole H nor an antihole A of size* ≥ 5 *in a NMRF can contain* ≥ 2 *members, say* s_1 *and* s_2*, of any singleton clique group.*

Proof. In H, s_1 and s_2 must be next to each other, then moving out round H one node in each direction, we cannot avoid a chord, contradiction. In A, there must be at least 2 nodes between s_1 and s_2 in at least one direction. Taking this way round A, the node next to s_1 must be adjacent to s_2 but not s_1, so has setting $s = 1$. Continuing round A, the next node must be adjacent to s_1, so must have an s value $\neq 1$ but then it is adjacent to its predecessor, contradiction. □

4.2.2 Remarks on the decomposition result

An interesting question is whether it might be possible to extend Theorem 4.2.1 to allow pasting on more than one variable. Unfortunately, we illustrate why this is not possible, except for restricted settings which are unlikely to be of interest.

Using similar notation to Theorem 4.2.1, suppose we paste on the edge $s - t$ and let us assume that at least one of the original NMRFs, say N_A, contains a hole running through the clique groups of s, $s - t$ and t, connecting back via path P (for example, the hole might be $\{s1t1, t0, P, s0\}$). Since N_A is perfect, P must have an even number of edges. But now, apart from restricted cases, the combined NMRF will contain a hole of the following form: $\{s0t1, t0, P, s0\}$ in N_A, together with $s1$ from N_B. This is an odd hole, hence the combined NMRF is not perfect (by Theorem 3.5.5).

4.3 New Result for Pairwise MRFs (any number of labels)

Lemma 4.3.1. *An antihole A of size ≥ 7 (odd or even) in an NMRF N derived from a pairwise MRF M (any number of labels) cannot contain an snode.*

Proof. Suppose A contains snode s_{i_0}, where $\{i_r : r = 0, \dots\}$ are distinct indices to labels of s. Consider 4 nodes furthest from s_{i_0} around A (the furthest 4 if A is odd), label these in order around A as x, x', y', y. All 4 must be adjacent to s_{i_0}, hence all have setting $s \neq i_0$. Let x have $s = i_1$, then moving around A we observe that x', y' and y must in turn all similarly have setting $s = i_1$ to avoid disallowed adjacencies between neighbors in A. By Lemma 4.2.4, all must be enodes. x and y must be adjacent, hence have different settings for some other variable t. Let x have settings $(s = i_1, t = j_0)$ and y have settings $(s = i_1, t = j_1)$, where $\{j_r : r = 0, \dots\}$ are distinct indices to labels of t. Now consider x': it must be adjacent to y so has some t setting, call it x'_t, which $\neq j_1$. But all such settings will lead to x' being adjacent to x, contradiction, unless $x'_t = j_0$ but then $x = x'$, again contradiction. □

4.4 New Results for Binary Pairwise MRFs

The results of this Section 4.4 and the next Section 4.5 are subject to the singleton node assumption of Section 4.1.

Lemma 4.4.1. *Let M be a binary pairwise MRF. $\exists$ a reparameterization such that M maps to perfect pruned NMRF $\Leftrightarrow$ $\exists$ a reparameterization with just one enode per edge in the pruned NMRF which is perfect.*

Proof. ($\Leftarrow$) is clear. ($\Rightarrow$) see Section 3.8. With a standard reparameterization, we may always achieve just one pruned enode (either 00 or 11 for associative, 01 or 10 for repulsive) from those already present. The result follows from the definition of a perfect graph. □

Therefore henceforth, when referring to a pruned NMRF of a binary pairwise MRF, we may assume just one enode per edge.

Lemma 4.4.2. *An antihole A of size ≥ 7 can never occur in a pruned NMRF N from a binary pairwise MRF M.*

Proof. Suppose A exists containing an snode, WLOG say s_0. This must be adjacent to ≥ 4 nodes in A, all of which must have $s = 1$ settings. The 2 closest to s_0 around A one way must both be adjacent to the closest to s_0 around A the other way, which cannot be achieved, hence A must contain only enodes. By Lemma 4.4.1, we have only one enode per edge of M. Two enodes are adjacent in N if they have one end in common with different settings - since only 2 settings are possible, a triangle in N must derive from edges in M that formed a triangle. Given 2 enodes which are adjacent, there is exactly one possible third enode with which they can form a triangle (e.g. for $s0t1$ and $t0u0$, $s1u1$ is the unique third possible enode). Yet A must contain ≥ 2 triangles which have the same 2 members but a different third member, contradiction. □

Since an antihole of size 5 is equivalent to a hole of the same size, SPGT (Theorem 3.5.5) gives the following.

Lemma 4.4.3. *For a binary pairwise MRF, a pruned NMRF is perfect ⇔ it contains no odd hole.*

4.5 Which Binary Pairwise MRFs Yield Perfect MRFs

The results of this Section 4.5 and the previous Section 4.4 are subject to the singleton node assumption of Section 4.1.

By Theorem 4.2.2, we need only consider 2-connected graphs G (considering both associative and repulsive edges), and by Lemma 4.4.3 we need only check for odd holes. G either contains a frustrated cycle or does not. If it does, we shall see that G must have the form $T_{m,n}$ or U_n. If not, we show G must have the form B_R. See Section 3.4 for definitions.

Lemma 4.5.1 (Harary, 1953)**.** *The following are equivalent properties for a signed graph G on vertices V:*

1. *G contains no frustrated cycle*
2. *G is of the form B_R*
3. *G is flippable to fully associative*

(1)⇔(2) by a variant of the standard proof that a graph is bipartite iff it has no odd cycle, considering repulsive edges. (3) means ∃ some subset $S \subseteq V$ such that if we replace each $X_i \in S$ by $Y_i = 1 - X_i$, and modify potential functions accordingly, thereby flipping the nature of each

edge incident to X_i between associative and repulsive, then all edges of G can be made associative. (2)$\Leftrightarrow$(3) by setting S as either partition.

Theorem 4.5.2. *A binary pairwise MRF with the form B_R maps efficiently to a bipartite NMRF N.*

Proof. Let the partitions of the variables be S and T with snodes $\{s_i^0, s_i^1\}$ from S, and $\{t_j^0, t_j^1\}$ from T. Choose a reparameterization such that any associative edge $x \oplus y$ maps to an enode $(x = 0, y = 0)$, and for any repulsive edge pick either form. Hence in N we have:

$\{e_i\}$ associative enodes from S, form $(s_i = 0, s_j = 0)$,

$\{f_i\}$ associative enodes from T, form $(t_i = 0, t_j = 0)$,

$\{a_i\}$ repulsive enodes $S \to T$, form $(s_i = 0, t_j = 1)$,

$\{b_i\}$ repulsive enodes $S \leftarrow T$, form $(s_i = 1, t_j = 0)$.

Observe N is bipartite with partitions $\{a_i, s_i^0, t_j^1, e_i\}$ and $\{b_i, s_i^1, t_j^0, f_i\}$. □

We now explore the case that G has a frustrated cycle.

Lemma 4.5.3. *Any cycle C in a binary pairwise MRF generates an induced (chordless) cycle H in its NMRF N with size at least as great, and with the same parity (odd/even number of vertices) as the number of repulsive edges (odd/even) in the MRF's cycle.*
In particular, if M contains any frustrated cycle with ≥ 4 edges, or with 3 edges requiring any snode to link the enodes in N, then this maps to an odd hole in N.

Proof. By Lemma 4.4.1, we may assume just one enode in N per edge in G. Form a cycle H in N using the enodes corresponding to the edges of C, together with connecting snodes as required (if two enodes meet at a variable and have the same setting, add an snode with the opposite setting). Clearly H is chordless and $|H| \geq |C|$.

Pick some enode e_1 and orientation around H. Consider the *end parity* of e_1, that is the setting for the next variable around H. For subsequent enodes, to maintain end parity requires an even (odd) total number of nodes, including possible snodes, for associative (repulsive) edges, and the reverse to flip end parity. Let a_m and a_f be the number of times end parity is maintained and flipped respectively using all associative edges around H, and similarly define r_m and r_f for all repulsive edges. In order to connect to the other end of e_1 after traversing H requires in total (including e_1) an odd number of flips, hence $a_f + r_f \equiv 1 \pmod 2$. The total number of nodes in H is comprised

of the first enode together with all subsequent nodes, hence

$$\begin{aligned} |H| &\equiv 1 + 0.a_m + 1.a_f + 1.r_m + 0.r_f \pmod 2 \\ &\equiv a_f + r_m + 1 \pmod 2 \equiv r_f + r_m \pmod 2. \end{aligned}$$

□

Using Lemmas 4.4.3 and 4.5.3 we show the following result.

Lemma 4.5.4. *Let M be a binary pairwise MRF that maps to an NMRF N. If N is not perfect then $\exists$ a frustrated cycle in M that maps to an odd hole in N. Hence, N is perfect $\Leftrightarrow \nexists$ such a cycle in M.*

Proof. By Lemma 4.4.3, N contains an odd hole H. By Lemma 4.2.4, any snode in H is adjacent to two enodes, and hence H must have derived from a cycle in M. Lemma 4.5.3 completes the proof. □

Lemma 4.5.5. *The only 2-connected binary pairwise MRFs containing a frustrated cycle, that map to a perfect NMRF, are of the form $T_{m,n}$ or U_n.*

Proof. See Section 3.4 for definitions. By Lemmas 4.5.3 and 4.5.4, we need only consider a frustrated triangle in M whose enodes in N require no connecting snodes. This triangle may have either (1) one repulsive and two associative edges, which we shall show must be of the form U_n or $T_{m,n}$ with $n \geq 1$, or (2) three repulsive edges, which we shall show must be of the form $T_{m,n}$.

It is simple to check that, in either case, a fourth vertex adjacent to all 3 vertices of the triangle, resulting in a K_4 clique, does not admit a reparameterization that avoids a frustrated cycle requiring connecting snodes.

Case 1: Triangle with one repulsive edge. We have a U_1 structure. Let the configuration in the MRF be $s \oplus t \ominus v_1 \oplus s$. In order to avoid connecting snodes in N, we must have one of the following two reparameterizations: $\{(s = 0, t = 0), (t = 1, v_1 = 0), (v_1 = 1, s = 1)\}$ or $\{(s = 1, t = 1), (t = 0, v_1 = 1), (v_1 = 0, s = 0)\}$. Once one edge has been selected, the others can follow in only one way. Consider what may be added to this graph while remaining 2-connected and avoiding a frustrated cycle with ≥ 4 edges. Any additional vertex v_2 must be attached by disjoint paths to at least 2 vertices x and y of the triangle. If either path has length ≥ 2 then, by choosing

one or other path in the original U_1 from x to y, we always find a frustrated cycle with ≥ 4 edges, leading to an odd hole. Using the argument from the preceding paragraph, v_2 must be adjacent to exactly 2 vertices of U_1. If these vertices are connected by an associative edge, we now have U_2; otherwise we have $T_{0,2}$. Checking cases now shows that the only way to add further vertices results in U_n or $T_{m,n}$ structures, with any $m \geq 0, n \geq 1$ allowed.

Case 2: Triangle with three repulsive edges. We have $T_{1,0}$. Similar reasoning to case 1 shows that the only possibilities are $T_{m,n}$ for any $m \geq 1, n \geq 0$.

□

Taking the results of this Section together, we have the following characterization.

Theorem 4.5.6. *A binary pairwise MRF maps to a perfect NMRF for all valid ψ_c iff each of its blocks (using all edges) has the form B_R, $T_{m,n}$ or U_n.*

4.5.1 Remarks

Theorem 4.5.6 certainly has theoretical value in establishing the boundaries of the MWSS approach for this class of MRFs. Further, it appears to broaden the landscape of tractable models. Each of the three block categories is itself tractable by other methods *in isolation*: QPBO (Rother et al., 2007) is guaranteed to be able to handle a B_R structure (though not $T_{m,n}$ or U_n), or indeed a B_R structure may be flipped to yield a fully associative model which can be solved with any appropriate technique such as graph cuts; and each $T_{m,n}$ or U_n has low tree width so admits traditional inference methods.

Initially, we believed our approach was the first to be able to handle an MRF containing $\Omega(n)$ of these structures, including high tree width B_R sections, automatically in polynomial time. However, as pointed out by David Sontag (private correspondence), the LP relaxation on the triplet-consistent polytope[2] will solve all our cases exactly in polynomial time. In addition, it can go further to handle any block with treewidth 2. For a fuller discussion of this result, along with a discussion of which additional binary pairwise MRFs are tractable via a reparameterization to a perfect pruned NMRF if the singleton node assumption of Section 4.1 is relaxed, see Appendix B.

[2]Whereas the usual local polytope enforces consistency between all pairs of variables, the triplet-consistent polytope, denoted TRI, enforces consistency across all triplets of variables.

4.5.1.1 Efficient detection

Detecting if a binary pairwise MRF with topology (V, E) satisfies our conditions may be performed in time $O(|E|)$: identifying block structure is an application of DFS, then each block type may be efficiently checked. The $T_{m,n}$ and U_n structures are straightforward. For B_R, first test if it is bipartite using just E_R (an application of BFS). Next check each component by E_R to see that no E_A cross partitions. Then stitch together partitions from different components (if more than one) using E_A. If any E_A cross partitions then it is easy to see $\exists$ a frustrated cycle with ≥ 4 edges which would lead to an odd hole in the NMRF.

4.6 Higher Order Submodular Cost Functions

As noted in the introduction, Jebara (2014) has shown that a fully associative binary pairwise model, which is equivalent to a model with supermodular pairwise ψ functions (submodular cost functions), can always be reparameterized so as to yield a bipartite pruned NMRF. Indeed, we have seen in Section 3.8 that, for each associative edge $x \oplus y$, one may reparameterize and prune the edge clique group so as to leave only either form $(x = 0, y = 0)$ or $(x = 1, y = 1)$. Here we extend the analysis to consider higher order models, still focusing on submodular cost functions over binary variables. We shall show that for potentials over 3 variables, a bipartite pruned NMRF is obtained for any topology iff all cost functions are submodular. Further, we show that submodularity is a necessary but strictly insufficient condition to obtain a bipartite pruned NMRF for all orders higher than 3.

Considering other approaches, this is similar to the result of Zivny et al. (2009) that all order 3 submodular functions over Boolean variables can be represented by order 2 submodular functions using auxiliary variables, but this is not always true when the order > 3. Also, Kolmogorov and Zabih (2004) showed that submodularity was necessary for a function to be graph-representable. However, Arora et al. (2012) recently demonstrated a novel graph cuts method for submodular cost functions of any order[3] over binary variables. Still, our result usefully clarifies the boundaries of our approach if we restrict to bipartite NMRFs only, and there is hope yet that a broader class of models may map to the wider class of perfect NMRFs.

[3]The time is exponential in the order of the potentials.

4.6.1 Notation

Let ψ be an order k potential function over k binary variables $X = \{X_1, \ldots, X_k\}$. Let one setting be $x = (x_1, \ldots, x_k)$. Let $x-ij$ be a setting for all variables other than X_i and X_j. Let $\psi_x = \psi(X = x)$. Define the supermodularity s of ψ with respect to X_i, X_j on the projection given by $x - ij$, as $s^{ij}_{x-ij} = \psi(X_i = 0, X_j = 0) + \psi(X_i = 1, X_j = 1) - \psi(X_i = 1, X_j = 0) - \psi(X_i = 0, X_j = 1)$ where all other variables in $X \setminus \{X_i, X_j\}$ are held fixed at $x - ij$.

Define $\alpha_k = \sum_{\text{all } 2^k \text{ settings of } x} (-1)^{\#0\text{s in } x} \psi_x$. Observe that for $k = 2$, this is the supermodularity s term. For $k = 3$, this is the difference between s with (any) one variable set to 0 and that with the same variable set to 1. For $k = 4$, we have the sum of two s terms minus two others, etc.

$\forall Y \subseteq \mathcal{P}(X)$, let O_Y and I_Y be weighted indicator functions. The O functions are 0 unless all of Y are 0. The I functions are 0 unless all of Y are 1. Otherwise, O_Y and I_Y take values Z_Y and A_Y, respectively. $Y = b$ means fix variables Y at value b where $b \in \mathbb{B} = \{0, 1\}$.

In order to map to a bipartite pruned NMRF for any topology at order k, we must be able to represent every ψ_x as the sum of a constant term and nonnegative[4] O and I indicator functions over all subsets of X, which correspond exactly to the nodes in the pruned NMRF (which is then clearly bipartite with stable sets corresponding to the $\{O_Y\}$ and $\{I_Y\}$).

4.6.2 Results

Theorem 4.6.1. *For $k \geq 2$, mapping to a bipartite pruned NMRF for any topology $\Rightarrow$ ψ is supermodular, equivalently every projection of ψ onto two variables is supermodular.*

Proof. Given the ψ_x representation from the previous paragraph, consider which A_Y, Z_Y terms survive when a general supermodularity term s^{ij}_{x-ij} is computed. For some Y, analyze A_Y terms (a similar result holds for Z_Y terms): Y will include either none, one or two of the variables $\{X_i, X_j\}$. Consider the cases: If none, then A_Y does not feature in the s^{ij}_{x-ij} computation. If one, then we get plus A_Y (from the $X_i = X_j = 1$ term) minus A_Y (from the appropriate other term), so they cancel. Finally, if two, then we simply get plus the A_Y term. Hence for every s^{ij}_{x-ij}, it must be equal to the sum of some A_{Y_i} and Z_{Y_j} terms, all of which are constrained to be ≥ 0. Hence all supermodularity

[4] It is critical that the functions be nonnegative in order that the corresponding nodes in the NMRF are the only ones not pruned.

terms are ≥ 0. □

Further, for $k = 4$, it is easily checked that $\alpha_k = A_X + Z_X$, where we require $A_X, Z_X \geq 0$, yet it also equals $s^{ij}_{x-ij=00} + s^{ij}_{x-ij=11} - s^{ij}_{x-ij=01} - s^{ij}_{x-ij=10}$ (for any 2 variables X_i, X_j), which may be positive but equally may be negative.[5] Similarly for all $k > 4$, we are not able to represent all supermodular ψ functions.

Theorem 4.6.2. *For general interactions over $k = 3$ variables, ψ is supermodular $\Leftrightarrow$ we obtain a bipartite pruned NMRF for any topology.*

Proof. ($\Leftarrow$) follows from Theorem 4.6.1. ($\Rightarrow$) we provide a constructive proof:[6]

If $\alpha_k \geq 0$, use only O_Y for $|Y| \geq 2$. Set $Z_X = \alpha_k$. For $|Y| = 2$, set $Z_Y = s^Y_1$. For $|Y| = 1$, set $Z_Y = \psi(Y = 0, (X \setminus Y) = 1) - \psi_{111}$. Set constant to ψ_{111} to observe we match ψ_x values $\forall x$. Now reparameterize all singleton terms and prune as usual, see Section 3.7.

If $\alpha_k \leq 0$, use only I_Y for $|Y| \geq 2$. Set $A_X = -\alpha_k$. For $|Y| = 2$, set $A_Y = s^Y_0$. For $|Y| = 1$, set $A_Y = \psi(Y = 1, (X \setminus Y) = 0) - \psi_{000}$. As before, set constant to ψ_{000} to check values, then reparameterize all singleton terms and prune, see Section 3.7. □

4.7 Conclusions

The MWSS approach to MAP inference is an exciting, recent approach, leveraging the rapid progress in combinatorics. Here we have derived new general tools (Section 4.2), defined the scope of the approach in an important, broad setting (Sections 4.4 and 4.5), and clarified the power of mapping to bipartite NMRFs (Section 4.6).

Future areas to explore include the open questions in Appendix B (where we consider reparameterizations that absorb singleton nodes), non-bipartite perfect NMRFs for higher order potentials, and variables with a greater number of labels.

---

[5] An example of supermodular ψ for $k = 4$ where $\alpha_k < 0$: $\psi(x_1, x_2, x_3, x_4) = 0$ except $\psi(0, 0, 0, 0) = 2, \psi(1, 0, 0, 0) = \psi(0, 1, 0, 0) = \psi(0, 0, 1, 0) = \psi(0, 0, 0, 1) = 1$.

[6] In fact, as shown, we need use only either exclusively O_Y or I_Y nodes for $|Y| \geq 2$, which may further improve efficiency.

Part III

On the Bethe Approximation

The material in this Part is based upon work in (Weller and Jebara, 2013a, 2014a; Weller et al., 2014) and to appear in (Weller and Jebara, 2014b). Related code is available at `http://www.cs.columbia.edu/~jebara/code/betheCleanUAI.tar`.

Chapter 5

Additional Background

This Chapter provides additional background to the Bethe approximation.

5.1 Marginal Inference and Estimating the Partition Function

One popular method to tackle the problems of estimating the partition function and marginal inference is the 'sum-product' version of the message-passing algorithm called belief propagation (Pearl, 1988). If the topology of the model is a tree, this will return the exact solution in linear time in n, the number of variables. If the method is applied to general topologies, termed loopy belief propagation (LBP), results are often strikingly good (McEliece et al., 1998; Murphy et al., 1999), though there are cases where it performs poorly, typically when there are many short cycles with strong edge interactions (Wainwright and Jordan, 2008, § 4.1), and in general it may not converge at all.

Coming from a seemingly different perspective, variational methods show that the partition function may be obtained by minimizing the free energy over the marginal polytope. Recall Section 2.6.1, where it was shown that

$$-\log Z = \min_{q \in \mathbb{M}} \mathcal{F}_G(q) = \min_{q \in \mathbb{M}} \mathbb{E}_q(E) - S(q(x)).$$

$\mathbb{M}$ is the marginal polytope which comprises all globally valid probability distributions over all the variables, i.e. for binary variables, it is the convex hull of all 2^n configurations. $\mathcal{F}_G$ is the Gibbs free energy, with the global optimum $\arg\min$ occurring at the true distribution.

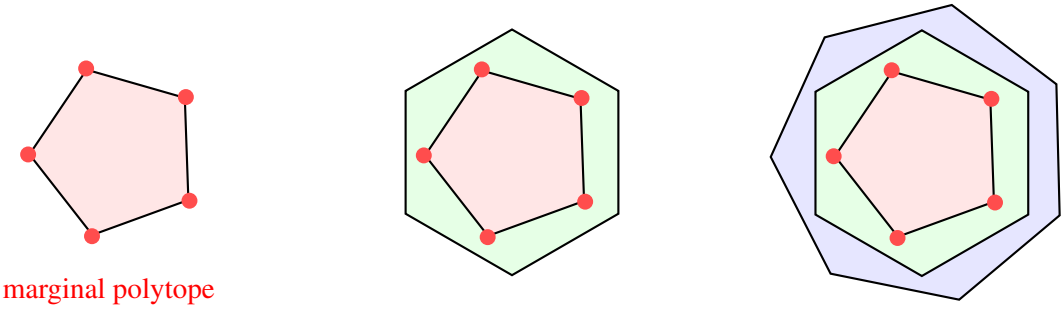
The Bethe approximation (Bethe, 1935) has two aspects, both pairwise approximations:

1. Relax the marginal polytope $\mathbb{M}$ to the local polytope $\mathbb{L}$ which enforces only pairwise consistency (thus we may obtain a *pseudo-marginal* solution that is not globally valid; see Section 6.2 for more details); and
2. Use the Bethe entropy approximation $S_B = \sum_{i \in V} S_i + \sum_{(i,j) \in \mathcal{E}} S_{ij} - S_i - S_j$. Here S_i is the entropy of the singleton distribution of X_i, and S_{ij} is the entropy of the pairwise pseudo-marginal of X_i and X_j. Note that $S_{ij} - S_i - S_j = -I_{ij} \leq 0$, where I_{ij} is the mutual information for an edge (see Wainwright and Jordan, 2008, §4 for details).

This yields the *Bethe partition function* Z_B at the global optimum of the Bethe free energy $\mathcal{F}$,

$$-\log Z_B = \min_{q \in \mathbb{L}} \mathcal{F}(q) = \min_{q \in \mathbb{L}} \mathbb{E}_q(E) - S_B(q(x)).$$

An illustration of the marginal and local polytopes is provided in Figure 5.1, which also shows the *cycle* polytope, which will be explored in Chapter 7.



marginal polytope

global consistency

cycle polytope

cycle consistency

local polytope

local consistency

Figure 5.1: Illustration of marginal, cycle and local polytopes. The marginal polytope is always within the cycle polytope which is always within the local polytope but in some cases they can be equal (for example, if the model has a tree topology). This is a stylistic representation, note that vertices of the marginal polytope are also vertices of the cycle and local polytopes, which can add further vertices (Sontag, 2010).

Yedidia et al. (2001) demonstrated a remarkable connection between the minimization of $\mathcal{F}$ and LBP, in that any fixed point of LBP corresponds to a stationary point of the Bethe free energy. This

was refined by Heskes (2002), who showed that every stable fixed point of LBP is a local minimum of the Bethe free energy.

Much work has focused on understanding conditions under which LBP is guaranteed to converge to the global optimum (Heskes, 2004; Mooij and Kappen, 2007; Watanabe, 2011), but outside these restricted settings, until recently, there were no methods even to approximate Z_B. Some conjectured that when LBP behaves poorly, it is likely that the Bethe approximation, as given by the global minimum, also performs poorly, but it has not previously been possible to test this.

Other approaches to minimizing the Bethe free energy such as gradient descent (Welling and Teh, 2001), double-loop methods (Yuille, 2002) or Frank-Wolfe (Frank and Wolfe, 1956; Belanger et al., 2013), will converge but only to a local minimum, and with no runtime guarantee. Focusing on the important class of binary pairwise models, Shin (2012) introduced an algorithm guaranteed to return an approximately stationary point of $\mathcal{F}$ in polynomial time,[1] though with a bound on the maximum degree, $\Delta = O(\log n)$, where n is the number of variables.

Ruozzi (2012) recently proved that $Z_B \leq Z$ for attractive models, using the technique of graph covers. Similarly, for graphical models whose partition function is the permanent of a non-negative matrix, Z_B is recoverable via convex optimization and, here too, $Z_B \leq Z$ (Huang and Jebara, 2009; Vontobel, 2010; Watanabe and Chertkov, 2010; Gurvits, 2011). Otherwise, beyond cases where the graph is acyclic, efficiently computing or approximating Z_B remains an active research topic.

An interesting, recent example of the quality of the Bethe approximation is the analysis of Chandrasekaran et al. (2011), where the approximation is shown to be very useful to count independent sets of a graph. Further, it is demonstrated that if the shortest cycle cover conjecture of Alon and Tarsi (1985) is true, then the Bethe approximation is very good indeed for a random 3-regular graph. As one further motivation for algorithms related to message passing, it was recognized that some form of this paradigm might be a useful simplified model for how neurons in the brain communicate (Doya, 2007), though this is highly speculative.

[1] An approximately stationary point is a pseudo-marginal vector where the magnitude of the derivative of the Bethe free energy is guaranteed to be below a given level. The value of $\mathcal{F}$ at such a point could be far from the optimum.

Chapter 6

Discrete Methods to Approximate the Partition Function

In this Chapter, we show how discrete MAP inference techniques may be used to find $\log Z_B$, that is $\log$ of the optimal Bethe partition function Z_B, of a binary pairwise model to within an arbitrary level of accuracy, specified by ϵ. For attractive models, we derive a fully polynomial-time approximation scheme (FPTAS), which addresses a long-standing theoretical question. Our approach is practical for use on small real-world problems. Further, since other methods which approximate the Bethe optimum, such as LBP or CCCP (Yuille, 2002), may converge only to a local optimum, our new method provides a new benchmark against which to compare other approaches.

The overall approach is to construct a provably sufficient mesh $\mathcal{M}(\epsilon)$, i.e. to discretize the problem in such a way that we guarantee that the optimum discretized point q^* will have Bethe free energy $\mathcal{F}(q^*)$ within ϵ of the true optimum. Two approaches for constructing this sufficient mesh are considered: one called *curvMesh*, based on bounding second derivatives of $\mathcal{F}$, as was introduced in (Weller and Jebara, 2013a); and another, typically much more efficient method called *gradMesh*, introduced in (Weller and Jebara, 2014a), based on bounding first derivatives. In either case, we first preprocess in order to bound the possible locations of any minima of the Bethe free energy $\mathcal{F}$ away from the extreme values of 0 or 1, where derivatives become infinite, inside an orthotope we term the *Bethe box*. Then we consider how best to solve the resulting discrete optimization problem. The analysis of second derivatives, extending the work of Korč et al. (2012), also crucially

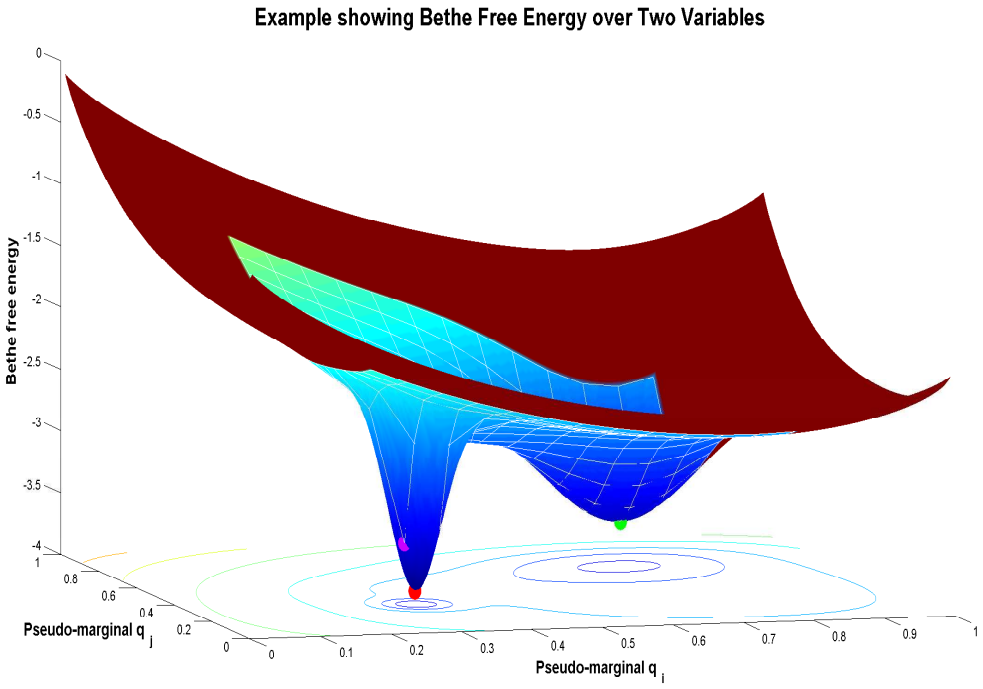


Figure 6.1: Stylized example showing Bethe free energy over two variables. We preprocess in order to rule out the region shown in red. This leaves the area shown in blue which we term the *Bethe box*. This is discretized using a *sufficient mesh*, shown in white. The red dot indicates the (continuous) global minimum. The purple dot is the mesh point with closest location. The green dot is the lowest point on the mesh, hence this is the discretized optimum returned.

demonstrates that if the initial binary model is fully attractive (i.e. has submodular cost functions), then the resulting multi-label pairwise MAP problem is submodular, hence can be solved efficiently by graph cuts (Schlesinger and Flach, 2006; Greig et al., 1989; Goldberg and Tarjan, 1988). See Figure 6.1 for a stylized example, and Algorithm 1 for a high level summary of the approach.

This Chapter is organized as follows. In Section 6.1 we describe related work. In Sections 6.2-6.5, we establish notation and present various preliminary results, including various bounds and important properties of the derivatives of the Bethe free energy. We apply these in Section 6.6 to derive a mesh construction approach based on bounding first derivatives, which we call *gradMesh*. In Section 6.7, we provide an alternative approach based on bounding second derivatives, which we call *curvMesh*. This is typically significantly less efficient then gradMesh but is superior for very

small values of ϵ. In Section 6.8, we discuss the resulting discrete optimization problem. In certain settings this is tractable, and in general we mention several features that can make it easier to find a satisfactory solution, or at least to bound its value. Experiments are described in Section 6.9, where we compare the efficiency of the various methods for mesh construction, and demonstrate practical application of the algorithm. In particular, interestingly we show an example of a model where LBP fails to converge yet the Bethe approximation, as obtained using our algorithm, returns reasonable results. Conclusions are presented in Section 6.10.

Algorithm 1 Mesh method to return ϵ-approximate global optimum $\log Z_B$ for a general binary pairwise model

Input: ϵ, model parameters (convert using Section 6.2.1 if required)

Output: estimate of global optimum $\log Z_B$ guaranteed to be in range $[\log Z_B - \epsilon, \log Z_B]$, together with corresponding pseudo-marginal as $\arg$ for the discrete optimum

Preprocess by computing bounds on the locations of minima, see Section 6.4.

Construct a sufficient mesh using one of the methods in this Chapter, see Sections 6.6 and 6.7 (all approaches are fast, so several may be used and the most efficient mesh selected).

Attempt to solve the resulting multi-label MAP inference problem, see Section 6.8.

If unsuccessful, but a strongly persistent partial solution was obtained, then improved location bounds may be generated (see Section 6.8.2.1), repeat from 2.

At anytime, one may stop and compute bounds on $\log Z_B$, see Section 6.8.2.

6.1 Related Work

Jerrum and Sinclair (1993) derived a fully polynomial-time randomized approximation scheme (FPRAS) for the true partition function, but only when singleton potentials are uniform (i.e. a uniform external field), and the runtime is high at $O(\epsilon^{-2}m^3n^{11}\log n)$. Heinemann and Globerson (2011) have shown that models exist such that the true marginal probability cannot possibly be the location of a minimum of the Bethe free energy. Approaches have been developed to solve related convex problems but results are typically less good (Meshi et al., 2009). Our work demonstrates an interesting connection between MAP inference techniques (NP-hard) and estimating the partition function Z (#P-hard). A different connection was shown by using MAP inference on randomly

perturbed models to approximate and bound Z (Hazan and Jaakkola, 2012).

6.2 Preliminaries

We focus on a binary pairwise MRF over n variables $X_1, \ldots, X_n \in \mathbb{B} = \{0, 1\}$ with graph topology $(\mathcal{V}, \mathcal{E})$ and follow notation similar to Welling and Teh (2001). Let $x = (x_1, \ldots, x_n)$ be one particular configuration. We assume[1]

$$p(x) = \frac{e^{-E(x)}}{Z},\ E = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} W_{ij} x_i x_j, \tag{6.1}$$

where the partition function $Z = \sum_x e^{-E(x)}$ is the normalizing constant. Let $m = |\mathcal{E}|$ be the number of edges, $\mathcal{N}(i)$ be the neighbors of i in the topology, and $d_i = |\mathcal{N}(i)|$ be the degree of i.

Given any joint probability distribution $q(X_1, \ldots, X_n)$ over all variables, the (Gibbs) free energy is defined as $\mathcal{F}_G(q) = \mathbb{E}_q(E) - S(q)$, where $S(q)$ is the (Shannon) entropy of the distribution. Using variational methods, a remarkable result is easily shown (see Section 2.6.1 or Wainwright and Jordan, 2008): minimizing $\mathcal{F}_G$ over the set of all globally valid distributions (termed the *marginal polytope*) yields a value of $-\log Z$ at the true marginal distribution p, given in (6.1).

This minimization is, however, computationally intractable, hence the approach of minimizing the Bethe free energy $\mathcal{F}$ makes two approximations: (i) the marginal polytope is relaxed to the *local polytope*, where we require only *local* consistency, that is we deal with a *pseudo-marginal* vector q, which here may be considered the set of all singleton and pairwise marginals $\{q_i = q(X_i = 1) \in [0, 1]\ \forall i \in \mathcal{V}, \mu_{ij} = q(x_i, x_j) \in [0, 1]^{2\times 2}\ \forall (i, j) \in \mathcal{E}\}$ subject to the usual conditions for probability distributions (non-negative, sum to 1), together with pairwise consistency requirements $q_i = \sum_{j \in \mathcal{N}(i)} \mu_{ij}, q_j = \sum_{i \in \mathcal{N}(j)} \mu_{ij}\ \forall i, j \in \mathcal{V}$; and (ii) the entropy S is approximated by the Bethe entropy $S_B = \sum_{(i,j) \in \mathcal{E}} S_{ij} + \sum_{i \in \mathcal{V}} (1 - d_i) S_i$, where S_{ij} is the entropy of μ_{ij}, and S_i is the entropy of the singleton distribution with probabilities $\{1 - q_i, q_i\}$.

[1] The energy E can always be thus reparameterized with finite θ_i and W_{ij} terms provided $p(x) > 0\ \forall x$, see Section 2.4.1. There are reasonable distributions where this does not hold, i.e. $\exists x : p(x) = 0$ but this can often be handled by assigning such configurations a sufficiently small positive probability ϵ.

The local polytope constraints imply that, given q_i and q_j,

$$\mu_{ij} = \begin{pmatrix} 1 + \xi_{ij} - q_i - q_j & q_j - \xi_{ij} \\ q_i - \xi_{ij} & \xi_{ij} \end{pmatrix}, \tag{6.2}$$

where $\mu_{ij}(a, b) = q(X_i = a, X_j = b)$, and we must have $\xi_{ij} \in [\max(0, q_i + q_j - 1), \min(q_i, q_j)]$ in order that all terms are non-negative.

Hence, the global optimum of the Bethe free energy,

$$\begin{aligned} \mathcal{F}(q) &= \mathbb{E}_q(E) - S_B(q) \\ &= \sum_{(i,j)\in\mathcal{E}} -\big(W_{ij}\xi_{ij} + S_{ij}(q_i, q_j)\big) \\ &\quad + \sum_{i\in\mathcal{V}} \big(- \theta_i q_i + (d_i - 1)S_i(q_i)\big), \end{aligned} \tag{6.3}$$

is achieved by minimizing $\mathcal{F}$ over the local polytope, with Z_B defined such that the result obtained equals $-\log Z_B$. See (Wainwright and Jordan, 2008) for details. Let $\alpha_{ij} = e^{W_{ij}} - 1$. $\alpha_{ij} = 0 \Leftrightarrow W_{ij} = 0$ may be assumed not to occur else the edge (i, j) may be deleted. α_{ij} has the same sign as W_{ij}, if positive then the edge (i, j) is *attractive* or *associative*; if negative then the edge is *repulsive*. The MRF is attractive if all edges are attractive. As shown by Welling and Teh (2001), one can solve for the optimum ξ_{ij} explicitly in terms of q_i and q_j by minimizing $\mathcal{F}$, leading to a quadratic with real roots,

$$\alpha_{ij}\xi_{ij}^2 - [1 + \alpha_{ij}(q_i + q_j)]\xi_{ij} + (1 + \alpha_{ij})q_i q_j = 0. \tag{6.4}$$

For $\alpha_{ij} > 0$, $\xi_{ij}(q_i, q_j)$ is the lower root, for $\alpha_{ij} < 0$ it is the higher. Thus we may consider the minimization of $\mathcal{F}$ over $q = (q_1, \ldots, q_n) \in [0, 1]^n$, that is we must search over n dimensions, which is still hard, but much easier than also having to search over pairwise marginals for every edge.

Collecting the pairwise terms of $\mathcal{F}$ from (6.3) for one edge, define

$$f_{ij}(q_i, q_j) = -W_{ij}\xi_{ij}(q_i, q_j) - S_{ij}(q_i, q_j). \tag{6.5}$$

We are interested in *discretized pseudo-marginals* where for each q_i, we restrict its possible values to a discrete mesh $\mathcal{M}_i$ of points in $[0, 1]$. The points may be spaced unevenly and we may have $\mathcal{M}_i \neq \mathcal{M}_j$. Let $N_i = |\mathcal{M}_i|$, and define $N = \sum_{i\in\mathcal{V}} N_i$ and $\Pi = \prod_{i\in\mathcal{V}} N_i$, the sum and product respectively of the number of mesh points in each dimension. Write $\mathcal{M}$ for the entire mesh.

Let $\hat{q}$ be the location of a global optimum of $\mathcal{F}$. We say that a mesh construction $\mathcal{M}(\epsilon)$ is *sufficient* if, given $\epsilon > 0$, it can be guaranteed that $\exists$ a mesh point $q^* \in \prod_{i \in \mathcal{V}} \mathcal{M}_i$ such that $\mathcal{F}(q^*) - \mathcal{F}(\hat{q}) \leq \epsilon$. The resulting discrete optimization problem may be framed as MAP inference in a multi-label MRF, where variable i takes values in $\mathcal{M}_i$, with the same topology (see Section 6.8).

We shall use the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$ in deriving bounds. We write A_i for the lower bound of q_i and B_i for the lower bound of $1 - q_i$ in the Bethe box, so the box is given by $A_i \leq q_i \leq (1 - B_i)\ \forall i \in \mathcal{V}$. Define $\eta_i = \min(A_i, B_i)$.

6.2.1 Input model specification

To be consistent with Welling and Teh (2001), for all theoretical analysis in this Chapter, we assume the reparameterization in (6.1). However, when an input model is specified, in order to avoid bias, we use singleton terms θ_i as in (6.1), but instead use pairwise energy terms given by $-\frac{W_{ij}}{2} x_i x_j - \frac{W_{ij}}{2}(1 - x_i)(1 - x_j)$. With this form, varying W_{ij} alters only the degree of association between i and j. We assume maximum possible values W and T are known with $|\theta_i| \leq T\ \forall i \in \mathcal{V}$, and $|W_{ij}| \leq W\ \forall (i,j) \in \mathcal{E}$. The required transformation to convert from input model to the format of (6.1), simply takes $\theta_i \leftarrow \theta_i - \sum_{j \in \mathcal{N}(i)} W_{ij}/2$, leaving W_{ij} unaffected.

6.2.2 Submodularity

In our context, a pairwise multi-label function on a set of ordered labels $X_{ij} = \{1, \ldots, K_i\} \times \{1, \ldots, K_j\}$ is *submodular* iff

$$\forall x, y \in X_{ij},\ f(x \wedge y) + f(x \vee y) \leq f(x) + f(y) \tag{6.6}$$

where for $x = (x_1, x_2)$ and $y = (y_1, y_2)$, $(x \wedge y) = (\min(x_1, y_1), \min(x_2, y_2))$ and $(x \vee y) = (\max(x_1, y_1), \max(x_2, y_2))$. For binary variables, submodular energy is equivalent to being attractive. For a more general discussion, see Section 2.5.1.

The key property in this Chapter is that if all pairwise cost functions f_{ij} over $\mathcal{M}_i \times \mathcal{M}_j$ from (6.5) are submodular, then the global discretized optimum may be found efficiently using graph cuts (Schlesinger and Flach, 2006). In analyzing second derivatives of $\mathcal{F}$ in Section 6.5, we show that this condition always holds provided the initial model is fully attractive.

6.3 Flipping Variables

The technique of flipping variables will be very useful for our analysis. Given a model on binary variables $\{X_i\}$, we can consider a new model with variables $\{X'_i\}$, where $X'_i = 1 - X_i$ for some selection of i. Flipping a variable flips the parity of all its incident edges so attractive $\leftrightarrow$ repulsive. Flipping both ends of an edge leaves its parity unchanged.

6.3.1 Flipping all variables

Consider a new model with variables $\{X'_i = 1 - X_i, i = 1, \ldots, n\}$ and the same edges. Instead of θ_i and W_{ij} parameters, let those of the new model be θ'_i and W'_{ij}. Identify values such that the energies of all states are maintained up to a constant[2]:

$$E = -\sum_{i\in\mathcal{V}} \theta_i X_i - \sum_{(i,j)\in\mathcal{E}} W_{ij} X_i X_j = const - \sum_{i\in\mathcal{V}} \theta'_i (1 - X_i) - \sum_{(i,j)\in\mathcal{E}} W'_{ij} (1 - X_i)(1 - X_j).$$

$$\text{Matching coefficients gives} \qquad W'_{ij} = W_{ij},\ \theta'_i = -\theta_i - \sum_{j\in\mathcal{N}(i)} W_{ij}. \tag{6.7}$$

If the original model was attractive, so too is the new.

6.3.2 Flipping some variables

Sometimes it is helpful to flip only a subset $\mathcal{R} \subseteq \mathcal{V}$ of the variables. This can be useful, for example, to make the model locally attractive around a variable, which can always be achieved by flipping just those neighbors to which it has a repulsive edge. Let $X'_i = 1 - X_i$ if $i \in \mathcal{R}$, else $X'_i = X_i$ for $i \in \mathcal{S}$, where $\mathcal{S} = \mathcal{V} \setminus \mathcal{R}$. Let $\mathcal{E}_t = \{\text{edges with exactly } t \text{ ends in } \mathcal{R}\}$ for $t = 0, 1, 2$.

As in 6.3.1, solving for W'_{ij} and θ'_i such that energies are unchanged up to a constant,

$$W'_{ij} = \begin{cases} W_{ij} & (i,j) \in \mathcal{E}_0 \cup \mathcal{E}_2, \\ -W_{ij} & (i,j) \in \mathcal{E}_1 \end{cases} \qquad \theta'_i = \begin{cases} \theta_i + \sum_{(i,j)\in\mathcal{E}_1} W_{ij} & i \in \mathcal{S}, \\ -\theta_i - \sum_{(i,j)\in\mathcal{E}_2} W_{ij} & i \in \mathcal{R}. \end{cases} \tag{6.8}$$

Lemma 6.3.1. *Flipping variables changes affected pseudo-marginal matrix entries' locations but not values. $\mathcal{F}$ is unchanged up to a constant, hence the locations of stationary points are unaffected.*

[2] Any constant difference will be absorbed into the partition function and leave probabilities unchanged.

Proof. By construction, energies are the same up to a constant. The singleton entropies are symmetric functions of q_i and $1 - q_i$ so are unaffected. The impact on pseudo-marginal matrix entries follows directly from definitions. Thus Bethe entropy is unaffected. □

6.4 Preliminary Bounds

We derive the following results, bounding the locations of stationary points of the Bethe free energy. We write ξ_{ij} for the optimum Bethe pairwise marginal parameter described in Section 6.2.

Lemma 6.4.1. $\alpha_{ij} \geq 0 \Rightarrow \xi_{ij} \geq q_i q_j, \alpha_{ij} \leq 0 \Rightarrow \xi_{ij} \leq q_i q_j$

Proof. The quadratic equation (6.4) for ξ_{ij} may be rewritten $\xi_{ij} - q_i q_j = \alpha_{ij}(q_i - \xi_{ij})(q_j - \xi_{ij})$. Both terms in parentheses on the right are elements of the pseudo-marginal matrix μ_{ij} so are constrained to be ≥ 0. □

For each variable X_i, we define the sum of the magnitude of incident attractive edge weights $W_i = \sum_{j \in \mathcal{N}(i): W_{ij} > 0} W_{ij}$, and similarly for incident repulsive edge weights, let $V_i = -\sum_{j \in \mathcal{N}(i): W_{ij} < 0} W_{ij}$.

Theorem 6.4.2. *For general edge types (associative or repulsive), at any stationary point of the Bethe free energy, $\sigma(\theta_i - V_i) \leq q_i \leq \sigma(\theta_i + W_i)$. Proof in Appendix C.*

Let the *Bethe box* be the smallest closed orthotope we can identify that must contain a global optimum of $\mathcal{F}$. We define A_i and B_i to be the minimum values in the Bethe box for q_i and $1 - q_i$ respectively, hence the Bethe box is given by $\prod_{i \in \mathcal{V}} [A_i, 1 - B_i]$, and by Theorem 6.4.2, we may take $A_i = \sigma(\theta_i - V_i), B_i = 1 - \sigma(\theta_i + W_i)\ \forall i$. These bounds alone are sufficient for all our theoretical analysis. Improved $\{A_i, B_i\}$ bounds may, however, be found by various methods, including Bethe bound propagation (BBP, a new approach we developed, see Section C.1 in the Appendix), which returns ranges guaranteed to include any stationary points of $\mathcal{F}$. An alternative approach, which we term MK, was derived in (Mooij and Kappen, 2007), based on considering the set of possible beliefs after iterating LBP, starting from any initial values. Since any minimum of $\mathcal{F}$ corresponds to a fixed point of LBP (Yedidia et al., 2001), this method may be used as an alternative to BBP. MK considers cavity fields around each variable, which requires more time, but the bounds obtained are no worse, and sometimes significantly better. Both BBP and MK require $O(m)$ time per iteration in

an efficient implementation, with each A_i and B_i term monotonically nondecreasing at each step. They typically converge within 50 iterations, even for large, dense models, and may be stopped at any time.

Define $\eta_i = \min(A_i, B_i)$, i.e. the closest that q_i can come to the extreme values of 0 or 1.

Lemma 6.4.3 (Upper bound for ξ_{ij} for an attractive edge)**.** *If $\alpha_{ij} > 0$, then*
$q_j - \xi_{ij} \geq \frac{q_j(1-q_i)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \geq \frac{q_j(1-q_i)}{1+\alpha_{ij}}, \quad q_i - \xi_{ij} \geq \frac{q_i(1-q_j)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \geq \frac{q_i(1-q_j)}{1+\alpha_{ij}}$.
Also $\xi_{ij} \leq m(\alpha_{ij}+M)/(1+\alpha_{ij}) \Rightarrow \xi_{ij} - q_iq_j \leq \frac{\alpha_{ij}m(1-M)}{1+\alpha_{ij}}$, where $m = \min(q_i, q_j)$ and $M = \max(q_i, q_j)$. Proof in Appendix C.

6.5 Derivatives of $\mathcal{F}$

In (Welling and Teh, 2001), first partial derivatives of the Bethe free energy are derived as

$$\frac{\partial \mathcal{F}}{\partial q_i} = -\theta_i + \log Q_i \text{ , where } Q_i = \frac{(1-q_i)^{d_i-1}}{q_i^{d_i-1}} \frac{\prod_{j\in\mathcal{N}(i)}(q_i - \xi_{ij})}{\prod_{j\in\mathcal{N}(i)}(1+\xi_{ij}-q_i-q_j)}. \tag{6.9}$$

Using the tools of convex analysis, extending the approach of Korč et al. (2012), we derive novel formulations of the second derivatives of the Bethe free energy, which will be used in this Chapter to prove Theorem 6.5.2 (submodularity of any discretization of an attractive MRF), bound the maximum possible curvature as needed for curvMesh (see Section 6.7), and will also be required for the methods used in Chapter 8 for analyzing the behavior of the Bethe optimum as variables are clamped.

Theorem 6.5.1 (Second derivatives for each edge)**.** *For any edge (i, j), for any α_{ij},*

$$\frac{\partial^2 f_{ij}}{\partial q_i^2} = \frac{1}{T_{ij}} q_j(1-q_j), \quad \frac{\partial^2 f_{ij}}{\partial q_i \partial q_j} = \frac{\partial^2 f_{ij}}{\partial q_j \partial q_i} = \frac{1}{T_{ij}}(q_iq_j - \xi_{ij}), \quad \frac{\partial^2 f_{ij}}{\partial q_j^2} = \frac{1}{T_{ij}} q_i(1-q_i)$$

$$\text{where } T_{ij} = q_iq_j(1-q_i)(1-q_j) - (\xi_{ij} - q_iq_j)^2 \geq 0 \text{ with equality iff } q_i \text{ or } q_j \in \{0, 1\}. \tag{6.10}$$

Proof in Appendix C.

Using Theorem 6.5.1 and Lemma 6.4.1, it is easy to show the following important result.

Theorem 6.5.2 (Submodularity for any discretization of an attractive model)**.** *If a binary pairwise MRF is submodular on an edge (i, j), i.e. $W_{ij} > 0$, then the multi-label discretized MRF for any mesh $\mathcal{M}$ is submodular for that edge. In particular, if the MRF is fully attractive, i.e. $W_{ij} >$*

$0\ \forall (i,j) \in \mathcal{E}$, *then the multi-label discretized MRF is fully submodular for any discretization. Proof in Appendix C.*

The on-diagonal Hessian terms are easily derived by considering the contribution from singleton terms $f_i(q_i)$ from (6.3). The only non-zero derivatives are with respect to q_i.

$$\begin{aligned} f_i(q_i) &= -\theta_i q_i + (d_i - 1) S_i(q_i), \\ \frac{\partial f_i}{\partial q_i} &= -\theta_i - (d_i - 1)[\log q_i - \log(1 - q_i)], \\ \frac{\partial^2 f_i}{\partial q_i^2} &= -(d_i - 1)\frac{1}{q_i(1 - q_i)} \leq 0 \text{ for a connected graph.} \end{aligned}$$

Incorporating all singleton and edge terms gives the following result.

Theorem 6.5.3 (All terms of the Hessian)**.** *Let H be the Hessian of $\mathcal{F}$ for a binary pairwise model, with $H_{ij} = \frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j}$, and $d_i = |\mathcal{N}(i)|$ be the degree of variable X_i, then*

$$H_{ii} = -\frac{d_i - 1}{q_i(1 - q_i)} + \sum_{j \in \mathcal{N}(i)} \frac{q_j(1 - q_j)}{T_{ij}} \geq \frac{1}{q_i(1 - q_i)}, \qquad H_{ij} = \begin{cases} \frac{q_i q_j - \xi_{ij}}{T_{ij}} & (i,j) \in \mathcal{E} \\ 0 & (i,j) \notin \mathcal{E}, i \neq j \end{cases}, \tag{6.11}$$

where $T_{ij} = q_i q_j (1 - q_i)(1 - q_j) - (\xi_{ij} - q_i q_j)^2 \geq 0$ with equality iff q_i or $q_j \in \{0, 1\}$.

Proof. Combine singleton terms from above with edge terms from Theorem 6.5.1. Observe that $T_{ij} \leq q_i q_j (1 - q_i)(1 - q_j)$. □

Remark. Lemma 6.4.1 shows that $q_i q_j - \xi_{ij} \leq 0$ for an attractive edge, hence in an attractive model, $H_{ij} = \frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j} \leq 0\ \forall i, j \in \mathcal{V}$.

6.6 *gradMesh* Approach Based on Bounding First Derivatives

We construct a sufficient mesh $\mathcal{M}$ by analyzing bounds on the first derivatives of $\mathcal{F}$ (applying a variant of the analysis that was used to derive the BBP algorithm, see the Appendix C.1). To help distinguish between methods, we call this first derivative approach *gradMesh*, and the second derivative approach *curvMesh*, described in Section 6.7. The gradMesh approach has several attractive features:

- For attractive models, we obtain a FPTAS with worst case runtime $O(\epsilon^{-3}n^3m^3W^3)$ and no restriction on topology, unlike with curvMesh (see Section 6.7) that requires $\max$ degree $\Delta = O(\log n)$ to guarantee polynomial runtime.

- The sufficient mesh is typically dramatically coarser than that achieved with curvMesh, leading to a much smaller subsequent MAP problem, unless ϵ is very small. For gradMesh, the sum of the number of discretizing points in each dimension, $N = O\left(\frac{nmW}{\epsilon}\right)$. For comparison, curvMesh forms a mesh with $N = O\left(\epsilon^{-1/2}n^{7/4}\Delta^{3/4}\exp\left[\frac{1}{2}(W(1+\Delta/2)+T)\right]\right)$. See Section 6.9.1 for examples.

- The approach immediately handles a general model with both attractive and repulsive edges. Hence approximating $\log Z_B$ may be reduced to a discrete multi-label MAP inference problem. This is valuable due to the availability of many MAP techniques, see Section 6.8.

First consider a model which is fully attractive around variable X_i, i.e. $W_{ij} > 0\ \forall j \in \mathcal{N}(i)$. From (6.9) and Lemma 6.4.1, we obtain

$$\frac{\partial \mathcal{F}}{\partial q_i} = -\theta_i + \log Q_i \leq -\theta_i + \log \frac{q_i}{1-q_i}. \tag{6.12}$$

Flip all variables (see Section 6.3). Write $'$ for the parameters of the new flipped model, which is also fully attractive, then using (6.7) and (6.12),

$$\begin{aligned} \frac{\partial \mathcal{F}'}{\partial q_i'} &\leq -\theta_i' + \log \frac{q_i'}{1-q_i'} \\ \Leftrightarrow -\theta_i - W_i + \log \frac{q_i}{1-q_i} &\leq \frac{\partial \mathcal{F}}{\partial q_i}. \end{aligned}$$

Combining this with (6.12) yields the sandwich result

$$-\theta_i - W_i + \log \frac{q_i}{1-q_i} \leq \frac{\partial \mathcal{F}}{\partial q_i} \leq -\theta_i + \log \frac{q_i}{1-q_i}.$$

Now generalize to consider the case that i has some neighbors $\mathcal{R}$ to which it is adjacent by repulsive edges. In this case, flip those nodes $\mathcal{R}$ (see Section 6.3) to yield a model, which we denote by $''$, which is fully attractive around i, hence we may apply the above result. By (6.8) we have $\theta_i'' = \theta_i - V_i$, and using $W_i'' = W_i + V_i$, we obtain that for a general model,

$$-\theta_i - W_i + \log \frac{q_i}{1-q_i} \leq \frac{\partial \mathcal{F}}{\partial q_i} \leq -\theta_i + V_i + \log \frac{q_i}{1-q_i}. \tag{6.13}$$

This bounds each first derivative $\frac{\partial \mathcal{F}}{\partial q_i}$ within a range of width $V_i + W_i = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$, which is sufficient for the main theoretical result, see (6.17). We take the opportunity, however, to describe a method which sometimes significantly narrows this range, thereby improving the result in practice.

Using one $O(m)$ iteration of the belief propagation algorithm (BBP, see the Appendix), allows us to refine the bounds for variable X_i of (6.13) based on the $[A_j, 1 - B_j]$ location bounds on its neighbors $j \in \mathcal{N}(i)$, to show

$$f_i^L(q_i) \leq \frac{\partial \mathcal{F}}{\partial q_i} \leq f_i^U(q_i), \text{ where}$$
$$f_i^L(q_i) = -\theta_i - W_i + \log \frac{q_i}{1 - q_i} + \log U_i$$
$$f_i^U(q_i) = -\theta_i + V_i + \log \frac{q_i}{1 - q_i} - \log L_i. \tag{6.14}$$

L_i, U_i are each > 1 with $\log L_i + \log U_i \leq V_i + W_i$. They are computed as $L_i = \prod_{j \in \mathcal{N}(i)} L_{ij}$, $U_i = \prod_{j \in \mathcal{N}(i)} U_{ij}$, with $L_{ij} = \begin{cases} 1 + \frac{\alpha_{ij} A_j}{1 + \alpha_{ij}(1 - B_i)(1 - A_j)} & \text{if } W_{ij} > 0 \\ 1 + \frac{\alpha_{ij} B_j}{1 + \alpha_{ij}(1 - B_i)(1 - B_j)} & \text{if } W_{ij} < 0 \end{cases}$,

$U_{ij} = \begin{cases} 1 + \frac{\alpha_{ij} B_j}{1 + \alpha_{ij}(1 - A_i)(1 - B_j)} & \text{if } W_{ij} > 0 \\ 1 + \frac{\alpha_{ij} A_j}{1 + \alpha_{ij}(1 - A_i)(1 - A_j)} & \text{if } W_{ij} < 0 \end{cases}$.

See Figure 6.2 for an example. We make the following observations:

- The upper bound is equal to the lower bound plus the constant $D_i = V_i + W_i - \log L_i - \log U_i \geq 0$.
- The bound curves are monotonically increasing with q_i, ranging from $-\infty$ to $+\infty$ as q_i ranges from 0 to 1.
- A necessary condition to be within the Bethe box is that the upper bound is ≥ 0 and the lower bound is ≤ 0. Hence, anywhere within the Bethe box, we must have bounded derivative, $|\frac{\partial \mathcal{F}}{\partial q_i}| \leq D_i$. BBP generates $\{[A_i, 1 - B_i]\}$ bounds by iteratively updating with L_i, U_i terms. In general, however, we may have better bounds from any other method, such as MK, which lead to higher L_i and U_i parameters and lower D_i.

$\mathcal{F}$ is continuous on $[0, 1]^n$ and differentiable everywhere in $(0, 1)^n$ with partial derivatives satisfying (6.14). $f_i^L(q_i)$ and $f_i^U(q_i)$ are continuous and integrable. Indeed, using the notation

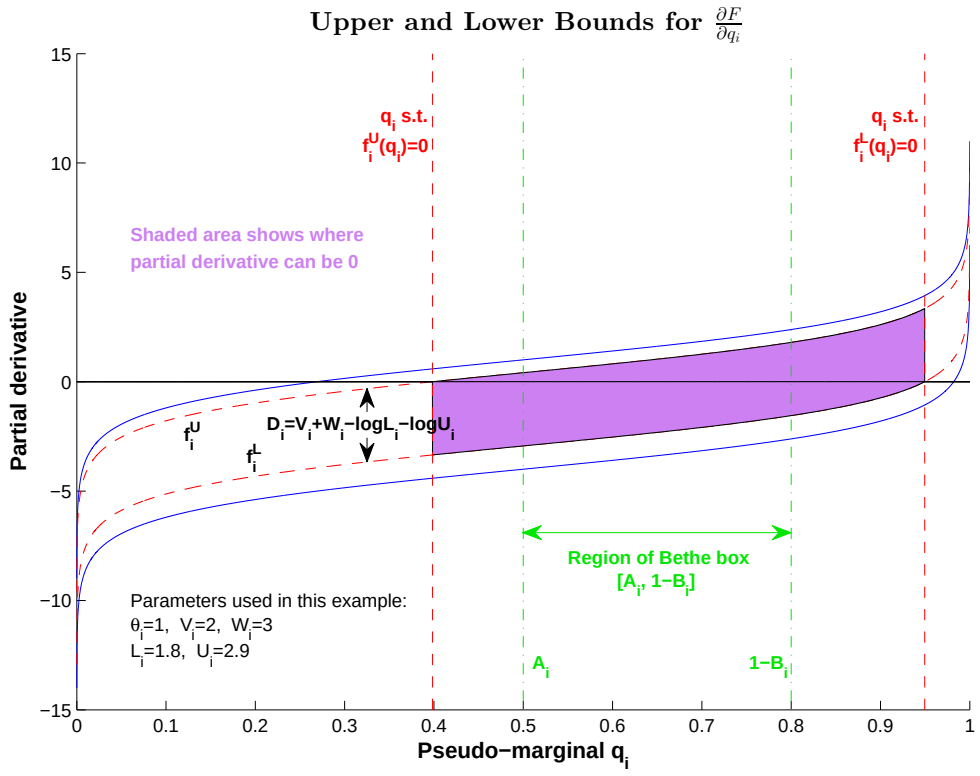


Figure 6.2: Upper and Lower Bounds for $\frac{\partial \mathcal{F}}{\partial q_i}$. Solid blue curves show worst case bounds (6.13) as functions of q_i, and are different by a constant $V_i + W_i = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$. Dashed red curves show the upper $f_i^U(q_i)$ and lower $f_i^L(q_i)$ bounds (6.14) after being lowered by $\log L_i$ and raised by $\log U_i$ respectively, which incorporate the information from the bounds of neighboring variables. All bounding curves are strictly monotonic. The Bethe box region for q_i must lie within the shaded region demarcated by vertical red dashed lines, but we may have better bounds available, e.g. from MK, as shown by A_i and $1 - B_i$.

$\left[\phi(x)\right]_{x=a}^{x=b} = \phi(b) - \phi(a)$,

$$\int_a^b \log \frac{q_i}{1-q_i} dq_i = \Big[q_i \log q_i + (1-q_i)\log(1-q_i)\Big]_{q_i=a}^{q_i=b} \tag{6.15}$$

for $0 \le a \le b \le 1$, which relates to the binary entropy function $H(p) = -p\log p - (1-p)\log(1-p)$, recall the definition of $\mathcal{F}$. We remark that although $\frac{\partial \mathcal{F}}{\partial q_i}$ tends to $-\infty$ or $+\infty$ as q_i tends to 0 or 1, the integral converges (taking $0 \log 0 = 0$).

Hence if $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_n)$ is the location of a global minimum, then for any $q = (q_1, \ldots, q_n)$ in the Bethe box,

$$\mathcal{F}(q) - \mathcal{F}(\hat{q}) \le \sum_{i:\hat{q}_i \le q_i} \int_{\hat{q}_i}^{q_i} f_i^U(q_i) dq_i + \sum_{i:q_i < \hat{q}_i} \int_{q_i}^{\hat{q}_i} -f_i^L(q_i) dq_i. \tag{6.16}$$

To construct a sufficient mesh, a simple initial bound relies on $|\frac{\partial \mathcal{F}}{\partial q_i}| \le D_i$. If mesh points $\mathcal{M}_i$ are chosen such that in dimension i there must be a point q^* within γ_i of a global minimum (which can be achieved using a mesh width in each dimension of $2\gamma_i$), then by setting $\gamma_i = \frac{\epsilon}{nD_i}$, we obtain $\mathcal{F}(q^*) - \mathcal{F}(\hat{q}) \le \sum_i D_i \frac{\epsilon}{nD_i} = \epsilon$. It is easily seen that $N_i \le 1 + \lceil \frac{1}{2\gamma_i} \rceil$, hence the total number of mesh points, $N = \sum_{i \in \mathcal{V}} N_i$, satisfies

$$\begin{aligned} N &\le 2n + \frac{n}{2\epsilon} \sum_i D_i \le 2n + \frac{n}{\epsilon} \sum_{(i,j)\in\mathcal{E}} |W_{ij}| \\ &= O\left(\frac{n}{\epsilon} \sum_{(i,j)\in\mathcal{E}} |W_{ij}| \right) = O\left(\frac{nmW}{\epsilon} \right), \end{aligned} \tag{6.17}$$

since $D_i \le V_i + W_i = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$. Here $W = \max_{(i,j)\in\mathcal{E}} |W_{ij}|$ and $m = |\mathcal{E}|$ is the number of edges.

If the initial model is fully attractive, then by Theorem 6.5.2, we obtain a submodular multi-label MAP problem which is solvable using graph cuts with worst case runtime $O(N^3) = O(\epsilon^{-3} n^3 m^3 W^3)$ (Schlesinger and Flach, 2006; Greig et al., 1989; Goldberg and Tarjan, 1988).

Note from the first expression in (6.17) that if we have information on individual edge weights then we have a better bound using $\sum_{(i,j)\in\mathcal{E}} |W_{ij}|$ rather than just mW.

For comparison, the second derivative curvMesh approach (Section 6.7) has runtime $O(\epsilon^{-\frac{3}{2}} n^6 \Sigma^{\frac{3}{4}} \Omega^{\frac{3}{2}})$, where $\Omega = O(\Delta e^{W(1+\Delta/2)+T})$. Unless ϵ is very small, the new first derivative approach is typically dramatically more efficient and useful in practice. Further, it naturally handles both attractive and repulsive edge weights in the same way.

6.6.1 Refinements and adaptive methods

Since the resulting multi-label MAP inference problem (which is not submodular in general) is NP-hard (Shimony, 1994), it is helpful to minimize its size. As noted above, setting $\gamma_i = \frac{\epsilon}{nD_i}$, which we term the *simple method*, yields a sufficient mesh, where $|\frac{\partial \mathcal{F}}{\partial q_i}| \leq D_i = V_i + W_i - \log L_i - \log U_i$. However, since the bounding curves are monotonic with $f_i^U \geq 0$ and $f_i^L \leq 0$, a better bound for the magnitude of the derivative is available by setting $D_i = \max\{f_i^U(1 - B_i), -f_i^L(A_i)\}$.

6.6.1.1 The *minsum* method

We define $N_i =$ the number of mesh points in dimension i, with sum $N = \sum_{i \in \mathcal{V}} N_i$ and product $\Pi = \prod_{i \in \mathcal{V}} N_i$. For a fully attractive model, the resulting MAP problem may be solved in time $O(N^3)$ by graph cuts (Theorem 6.5.2, (Schlesinger and Flach, 2006; Greig et al., 1989; Goldberg and Tarjan, 1988)), so it is sensible to minimize N. In other cases, however, it is less clear what to minimize. For example, a brute force search over all points would take time $\Theta(\Pi)$.

Define the spread of possible values in dimension i as $S_i = 1 - B_i - A_i$ and note $N_i = 1 + \lceil \frac{S_i}{2\gamma_i} \rceil$ is required to cover the whole range. To minimize N while ensuring the mesh is sufficient, consider the Lagrangian $\mathcal{L} = \sum_{i \in \mathcal{V}} \frac{S_i}{2\gamma_i} - \lambda(\epsilon - \sum_{i \in V} \gamma_i D_i)$, where D_i is set as in the simple method (Section 6.6.1). Optimizing gives

$$\gamma_i = \frac{\epsilon}{\sum_{j \in \mathcal{V}} \sqrt{S_j D_j}} \sqrt{\frac{S_i}{D_i}}, \text{and } N \leq 2n + \frac{1}{2\epsilon} \left(\sum_{i \in V} \sqrt{S_i D_i} \right)^2 \tag{6.18}$$

which we term the *minsum method*. Note $D_i \leq d_i W$ where d_i is the degree of X_i, hence $\left(\sum_{i \in V} \sqrt{S_i D_i}\right)^2 \leq W \left(\sum_{i \in V} \sqrt{d_i}\right)^2$. By Cauchy-Schwartz and the handshake lemma, $\left(\sum_{i \in V} \sqrt{d_i}\right)^2 \leq n \sum_{i \in \mathcal{V}} d_i = 2mn$, with equality iff the d_i are constant, i.e. the graph is regular.

If instead Π is minimized, rather than N, a similar argument shows that the simple method (Section 6.6.1) is optimal.

6.6.1.2 Adaptive methods

The previous methods rely on one bound D_i for $|\frac{\partial \mathcal{F}}{\partial q_i}|$ over the whole range $[A_i, 1 - B_i]$. However, we may increase efficiency by using local bounds to vary the mesh width across the range. A bound

on the maximum magnitude of the derivative over any sub-range may be found by checking just $-f_i^L$ at the lower end and f_i^U at the upper end.

This may be improved by using the exact integral as in (6.16). First, constant proportions $k_i > 0$ should be chosen with $\sum_i k_i = 1$. Next, the first or smallest mesh point $\gamma_1^i \in \mathcal{M}_i$ should be set such that $\int_{A_i}^{\gamma_1^i} f_i^U(q_i)dq_i = k_i\epsilon$. This will ensure that γ_1^i covers all points to its left in the sense that $\mathcal{F}[q_i = \gamma_1^i] - \mathcal{F}[q_i \in [A_i, \gamma_1^i]] \leq k_i\epsilon$ where all other variables $q_j, j \neq i$, are held constant at any values within the Bethe box. γ_1^i also covers all points to its right up to what we term its *reach*, i.e. the point r_1^i such that $\int_{\gamma_1^i}^{r_1^i} -f_i^L(q_i)dq_i = k_i\epsilon$. Next, γ_2^i is chosen as before, using r_1^i as the left extreme rather than A_i, and so on, until the final mesh point is computed with reach $\geq 1 - B_i$. This yields an optimal mesh for the choice of $\{k_i\}$.

If $k_i = \frac{1}{n}$, we achieve an optimized *adaptive simple* method. If $k_i = \frac{\sqrt{S_i D_i}}{\sum_{j \in \mathcal{V}} \sqrt{S_j D_j}}$, we achieve an *adaptive minsum* method. For many problems, this adaptive minsum method will be the most efficient.

Integrals are easily computed using (6.15). To our knowledge, computing optimal points $\{\gamma_s^i\}$ is not possible analytically, but each may be found with high accuracy in just a few iterations using a search method, hence total time to compute the mesh is $O(N)$, which is negligible compared to solving the subsequent MAP problem.

6.7 *curvMesh* Approach Based on Bounding Second Derivatives

In this Section, we describe the *curvMesh* approach to constructing a sufficient mesh. This is typically much less efficient than gradMesh (see Section 6.9.1 for a comparison of methods) but is included here because it has better ϵ dependency, so for extremely small ϵ, it can produce a more efficient mesh than gradMesh. It is also interesting in its own right and was developed first (Weller and Jebara, 2013a).

The approach is as follows: As for gradMesh, the possible location of a global minimum $\hat{q}$ is first bounded in the Bethe box given by $\prod_{i \in \mathcal{V}}[A_i, 1 - B_i]$ by preprocessing with BBP or MK (see Section 6.4). Next an upper bound Λ is derived on the maximum possible eigenvalue of the Hessian H of $\mathcal{F}$ anywhere within the Bethe box. Then a mesh of constant width in each dimension is introduced such that the nearest mesh point q^* to $\hat{q}$ is at most γ away in each dimension. Hence the

ℓ_2 distance δ satisfies $\delta^2 \leq n\gamma^2$ and by Taylor's theorem, $\mathcal{F}(q^*) \leq F(\hat{q}) + \frac{1}{2}\Lambda\delta^2$. Λ is computed by bounding the maximum magnitude of any element of H. Considering Theorem 6.5.3, this involves separate analysis of diagonal H_{ii} terms, which are positive and bounded above by the term b; and edge H_{ij} terms, which are negative (positive) for attractive (repulsive) edges, whose magnitude is bounded above by a. Then Ω is set as $\max(a, b)$, and Σ as the proportion of non-zero entries in H. Finally, $\Lambda \leq \sqrt{\text{tr}(H^T H)} \leq \sqrt{\Sigma n^2 \Omega^2} = n\Omega\sqrt{\Sigma}$ (it may be possible to derive a tighter bound on Λ using more sophisticated techniques, for example the method of Zhan, 2005, Corollary 2).

The approach is more natural for fully attractive models, but we show in Section 6.7.3 how a general model may be handled. The mesh generated has $N = \sum_{i \in \mathcal{V}} N_i = O(\epsilon^{-\frac{1}{2}} n^2 \Sigma^{\frac{1}{4}} \Omega^{\frac{1}{2}})$, where $\Omega = O(\Delta e^{W(1+\Delta/2)+T})$ and $\Delta = \max_{i \in \mathcal{V}} d_i$ is the maximum degree of the topology. Hence, curvMesh leads to a FPTAS for attractive models only if $\Delta = O(\log n)$, interestingly the same restriction as Shin (2012) required to find an approximately stationary point, whereas gradMesh has no such restriction.

We remark on two reasons why curvMesh is less efficient than gradMesh. First, when computing the mesh width based on an upper bound Λ on curvature in any direction, we must consider the worst case throughout the entire Bethe box. Since the derivatives $\to \infty$ as one approaches the edges of 0 or 1, this depends critically on $\min_{i \in \mathcal{V}}\{A_i, B_i\}$, so poor $\{A_i, B_i\}$ bounds lead to a very fine mesh. Secondly, the same, worst case, mesh width must be used isotropically throughout the entire Bethe box. Both aspects are in contrast to gradMesh.

6.7.1 Bounding off-diagonal terms H_{ij} for attractive edges

Here we derive an expression for a, an upper bound on $-H_{ij}$ for attractive edges, and show that $a = O(e^{W(1+\Delta/2)+T})$. This is a stronger result than was derived in (Weller and Jebara, 2013a): essentially, a more careful analysis allows a potentially small term in the numerator and denominator of a fraction to be canceled before bounding.

Using Theorem 6.5.3, equation (6.10) and Lemma 6.4.3,

$$-H_{ij} = (\xi_{ij} - q_i q_j)\frac{1}{T_{ij}} \leq \frac{m(1-M)\alpha_{ij}}{1+\alpha_{ij}} \frac{1}{m(1-M)\left[(1-m)M - m(1-M)\left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2\right]}$$
$$= \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right) \frac{1}{(1-m)M - m(1-M)\left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2} \tag{6.19}$$

where $m = \min(q_i, q_j), M = \max(q_i, q_j)$. Now we use the following result.

Lemma 6.7.1. *For any $k \in (0,1)$, let $y = \min_{q_i \in [A_i, 1-B_i], q_j \in [A_j, 1-Bj]} (1-m)M - m(1-M)k$, then*

$$y = \begin{cases} B_i A_j - (1-B_i)(1-A_j)k & \text{if } (1-B_i) \leq A_j \quad \textit{i range} \leq \textit{j range} \\ (1-k)\min\{A_j(1-A_j), B_i(1-B_i)\} & \text{if } A_i \leq A_j \leq 1-B_i \leq 1-B_j \quad \textit{ranges overlap, i lower} \\ (1-k)\min\{A_j(1-A_j), B_j(1-B_j)\} & \text{if } A_i \leq A_j \leq 1-B_j \leq 1-B_i \quad \textit{j range} \subseteq \textit{i range} \\ (1-k)\min\{A_i(1-A_i), B_i(1-B_i)\} & \text{if } A_j \leq A_i \leq 1-B_i \leq 1-B_j \quad \textit{i range} \subseteq \textit{j range} \\ (1-k)\min\{A_i(1-A_i), B_j(1-B_j)\} & \text{if } A_j \leq A_i \leq 1-B_j \leq 1-B_i \quad \textit{ranges overlap, j lower} \\ B_j A_i - (1-B_j)(1-A_i)k & \text{if } (1-B_j) \leq A_i \quad \textit{j range} \leq \textit{i range.} \end{cases}$$

Proof. The minimum is achieved by minimizing the larger and maximizing the smaller of q_i and q_j. The result follows for cases where their ranges are disjoint. If ranges overlap, then the minimum is achieved at some $q_i = q_j$ in the overlap, with value $q_i(1-q_i)(1-k)$, which is concave and minimized at an extreme of the overlap range. □

Lemma 6.7.1 is useful in practice, and should be used to compute $a = \max_{(i,j)\in\mathcal{E}}$ of the bound above. To analyze theoretical worst case, it is straightforward to see the corollary that $y \geq (1-k)\bar{\eta}$, where $\bar{\eta} = \min_{i\in\mathcal{V}} \eta_i(1-\eta_i)$. This bound can be met, for example, if all ranges coincide. Considering Section 6.2.1 on input model specification, we see that $\frac{1}{\eta_i(1-\eta_i)} = O(e^{T+\Delta W/2})$. Hence, from (6.19), and using $\alpha_{ij} = e^{W_{ij}} - 1$, we obtain

$$-H_{ij} \leq \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right) \Big/ \bar{\eta}\left(1 - \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2\right) = O(e^{W(1+\Delta/2)+T}). \tag{6.20}$$

This compares favorably to the earlier bound in (Weller and Jebara, 2013a), where it was shown that $a = O(e^{W(1+\Delta)+2T})$.

6.7.2 Bounding on-diagonal terms H_{ii} for attractive models

Here we derive b, an upper bound on H_{ii} for attractive models, and show that $b = O(\Delta e^{W(1+\Delta/2)+T})$. Since $\Omega = \max(a, b)$, this shows that $\Omega = O(\Delta e^{W(1+\Delta/2)+T})$.

First we derive a lower bound for T_{ij} at any point in the Bethe box. Let $K_{ij} = \eta_i\eta_j(1-\eta_i)(1-\eta_j)\frac{2\alpha_{ij}+1}{(\alpha_{ij}+1)^2}$. All terms are known from the data prior to the discrete optimization.

Lemma 6.7.2. *At any point in the Bethe box, $T_{ij} \geq K_{ij}$.*

Proof. Using Theorem 6.5.1 and Lemma 6.4.3,

$$\begin{aligned} T_{ij} &\geq q_iq_j(1-q_i)(1-q_j) - \Big(\frac{\alpha_{ij}m(1-M)}{1+\alpha_{ij}}\Big)^2 \\ &\geq q_iq_j(1-q_i)(1-q_j)\Big[1-\Big(\frac{\alpha_{ij}}{1+\alpha_{ij}}\Big)^2\Big]. \end{aligned}$$ □

Now using (6.11) and the expression from the proof of Lemma 6.7.2,

$$\begin{aligned} H_{ii} &\leq \frac{1-z_i}{\eta_i(1-\eta_i)} + \sum_{j\in\mathcal{N}(i)} \frac{1}{q_i(1-q_i)\Big[1-\Big(\frac{\alpha_{ij}}{1+\alpha_{ij}}\Big)^2\Big]} \\ &\leq \frac{1}{\eta_i(1-\eta_i)}\Big(1-z_i+\sum_{j\in\mathcal{N}(i)}\frac{(\alpha_{ij}+1)^2}{2\alpha_{ij}+1}\Big). \end{aligned}$$

Since $\alpha_{ij}+1 = e^{W_{ij}}$ and as above, $\frac{1}{\eta_i(1-\eta_i)} = O(e^{T+\Delta W/2})$, we obtain $b = O(\Delta e^{W(1+\Delta/2)+T})$.

We remark that $\alpha_{ij}+1 < 2\alpha_{ij}+1$, hence we have the corollary that $H_{ii} < \frac{1+\sum_{j\in\mathcal{N}(i)}\alpha_{ij}}{\eta_i(1-\eta_i)}$. At any minimum of the Bethe free energy, all eigenvalues are ≥ 0 so at these locations, the maximum eigenvalue $\leq \text{Tr}\, H < \sum_{i\in\mathcal{V}}\frac{1}{\eta_i(1-\eta_i)} + \sum_{(i,j)\in\mathcal{E}}\alpha_{ij}\Big(\frac{1}{\eta_i(1-\eta_i)}+\frac{1}{\eta_j(1-\eta_j)}\Big)$.

6.7.3 Extending the second derivative approach to a general model

Using flipping arguments from Section 6.3, we are able to extend the curvMesh method to apply to general (non-attractive) models. Interestingly, the bounds derived for $\Omega = \max(a, b)$ take exactly the same form as for the purely attractive case, except that now $-W \leq W_{ij} \leq W$, whereas previously it was required that $0 \leq W_{ij} \leq W$. Details and derivations are in the Appendix C.2.

6.8 The Derived Multi-Label MAP Inference Problem

After computing a sufficient mesh, it remains to solve the multi-label MAP inference problem on a MRF with the same topology as the initial model, where each q_i takes values in $\mathcal{M}_i$. In general, this is NP-hard (Shimony, 1994).

6.8.1 Tractable cases

If it happens that all cost functions are submodular (as is always the case if the initial model is fully attractive by Theorem 6.5.2), then as already noted, it may be solved efficiently using graph cut methods, which rely on solving a max flow/min cut problem on a related graph, with worst case runtime $O(N^3)$ (Schlesinger and Flach, 2006; Greig et al., 1989; Goldberg and Tarjan, 1988). Using the algorithm of Boykov and Kolmogorov (2004), performance is typically much faster, sometimes approaching $O(N)$. This submodular setting is the only known class of problem which is solvable for any topology.

Alternatively, the topological restriction of bounded tree-width allows tractable inference (Pearl, 1988). Further, under mild assumptions, this was shown to be the only restriction which will allow efficient inference for any cost functions (Chandrasekaran et al., 2008). We note that if the problem has bounded tree-width, then so too does the original binary pairwise model, hence exact inference (to yield the true marginals or the true partition function Z) on the original model is tractable using the junction tree algorithm, making our approximation result less interesting for this class. In contrast, although MAP inference is tractable for any attractive binary pairwise model, marginal inference and computing Z are not (Jerrum and Sinclair, 1993).

A recent approach reducing MAP inference to identifying a maximum weight stable set in a derived weighted graph, as described in Part II of this thesis or in (Jebara, 2014; Weller and Jebara, 2013b), shows promise, allowing efficient inference if the derived graph is perfect. Further, testing if this graph is perfect can be performed in polynomial time (Jebara, 2014; Chudnovsky et al., 2005b).

6.8.2 Intractable MAP cases

Many different methods are available, see Kappes et al. (2013) for a recent survey. Some, such as dual approaches, may provide a helpful bound even if the optimum is not found. Indeed, a LP

relaxation will run in polynomial time and return an upper bound on $\log Z_B$ that may be useful. A lower bound may be found from any discrete point, and this may be improved using local search methods.

Note that the Bethe box bounds on each $q_i \in [A_i, 1 - B_i]$ are worst case, irrespective of other variables. However, given a particular value for one or more q_j, $j \in \mathcal{N}(i)$, either BBP (see Appendix C.1) or MK (Mooij and Kappen, 2007) can produce better bounds on q_i, which may be helpful for pruning the solution space.

6.8.2.1 Persistent partial optimization approaches

The multi-label implementation of quadratic pseudo-Boolean optimization (Kohli et al., 2008, MQPBO), and the method of Kovtun (2003), are examples of this class. Both consider LP-relaxations and run in polynomial time. In our context, the output consists of ranges (which in the best case could be one point) of settings for some subset of the variables. If any such ranges are returned, the strong persistence property ensures that *any* MAP solution satisfies the ranges. Hence, these may be used to update $\{A_i, B_i\}$ bounds (padding the discretized range to the full continuous range covered by the end points if needed), compute a new, smaller, sufficient mesh and repeat until no improvement is obtained.

6.9 Experiments

6.9.1 Comparison of methods

We compared the efficiency of the various mesh construction methods, see Figure 6.3. We considered two values of ϵ: 1 (medium resolution) and 0.1 (fine resolution). For each value, we generated random MRFs on n variables, all pairwise connected, where $\theta_i \sim U[-2, 2]$ and $W_{ij} \sim U[-W, W]$, using the input convention of Section 6.2.1. We show results first for fixed $W = 5$ as n is varied from 3 to 20, then for fixed $n = 10$ as W is varied from 1 to 10, generating 10 random models for each value. Of the various first derivative gradMesh methods, only minsum is shown since the others would not be sufficiently distinguishable on these plots.[3] In addition to the methods described in

[3] In practice, the adaptive methods typically produce a mesh with about half the number of points in each dimension.

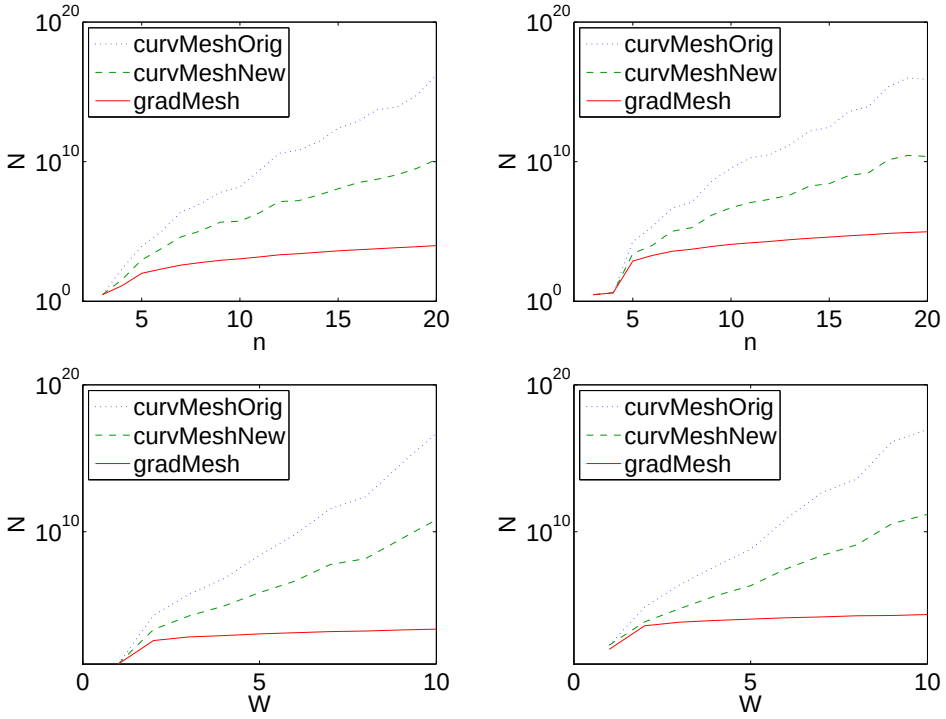


Figure 6.3: Variation in N = sum of number of mesh points in each dimension, **log scale**, as: (top) n = number of variables is changed, keeping $W = 5$ fixed; (bottom) W = maximum coupling strength is changed, keeping $n = 10$ fixed. On the left, $\epsilon = 1$ (medium resolution); on the right, $\epsilon = 0.1$ (fine resolution). In each case the topology is a complete graph, edge weights are chosen $W_{ij} \sim U[-W, W]$ and $\theta_i \sim U[-2, 2]$. Average over 10 random models for each value. *curvMeshOrig* is the original method of Weller and Jebara (2013a); *curvMeshNew* is the refinement described here in see Section 6.7; *gradMesh* is the first derivative minsum method, see Section 6.6. For more details, see text of Section 6.9.1.

this thesis, we also show results for the original curvMesh method described in (Weller and Jebara, 2013a), before it was improved as described in Section 6.7.[4]

Note that N is shown on a $\log$ axis, thus we observe that the new methods dramatically outperform the original curvMesh method of Weller and Jebara (2013a) by many orders of magnitude for most cases of interest, even for small ϵ. Further, recall that $N = \sum_i N_i$ is the sum of the number of mesh points in each dimension. The runtime of the overall algorithm is certainly $\Omega(N)$, even for attractive models[5], and for general models is typically a significantly higher power, thus further demonstrating the benefit of the new methods.

6.9.2 Power network

As a first step toward applying our algorithm to explore the usefulness of the global optimum of the Bethe approximation, here we identify one setting where LBP fails to converge, yet still we achieve reasonable results.

We aim to predict transformer failures in a power network (Rudin et al., 2012). Since the real data is sensitive, our experiments use synthetic data. Let $X_i \in \{0, 1\}$ indicate if transformer i has failed or not. Each transformer has a probability of failure on its own which is represented by a singleton potential θ_i. However, when connected in a network, a transformer can propagate its failure to nearby nodes (as in viral contagion) since the edges in the network form associative dependencies. We assume that homogeneous attractive pairwise potentials couple all transformers that are connected by an edge, i.e. $W_{ij} = W \; \forall (i, j) \in \mathcal{E}$. The network topology creates a Markov random field specifying the distribution $p(X_1, \ldots, X_n)$. Our goal is to compute the marginal probability of failure of each transformer within the network (not simply in isolation as in Rudin et al. (2012)). Since recovering $p(X_i)$ is hard, we estimate Bethe pseudo-marginals $q_i = q(X_i = 1)$ through our algorithm, which emerge as the $\arg\min$ when optimizing the Bethe free energy.

A single simulated sub-network of 55 connected transformers was generated using a random preferential attachment model, resulting in average degree 2 (see Figure C.1 in the Appendix C.3).

[4]The original method of (Weller and Jebara, 2013a) could only handle attractive models but we augment it in a manner similar to Section 6.7.3. Plots for attractive models, where $W_{ij} \sim U[0, W]$ are very similar to those shown.

[5]In our experiments on attractive models, the Boykov-Kolmogorov algorithm typically runs in time $O(N^{1.5})$ to $O(N^{2.5})$.

Typical settings of $\theta_i = -2$ and $W = 4$ were specified (using the input model specification of Section 6.2.1). We attempted to run BP using the libDAI package (Mooij, 2010) but were unable to achieve convergence, even with multiple initial values, using various sequential or parallel settings and with damping. However, running our gradMesh adaptive minsum algorithm with $\epsilon = 1$ achieved reasonable results as shown in Table 6.1, where true values were obtained with the junction tree algorithm.

$\epsilon = 1$ PTAS for $\log Z_B$	Error from true value
Mean ℓ_1 error of single marginals	0.003
Log-partition function	0.26

Table 6.1: Results on simulated power network

It has been suggested that the Bethe approximation is poor when BP fails to converge (Mooij and Kappen, 2005b). Our new method will allow this to be explored rigorously in future work. The initial result above is a promising first step and justifies further investigation.

6.10 Discussion and Future Work

To our knowledge, we have derived the first ϵ-approximation algorithm for $\log Z_B$ for a general binary pairwise model. From experiments run, we note that the ϵ bounds for the adaptive minsum first derivative *gradMesh* approach appear to be close to tight since we have found models where the optimum returned when run with $\epsilon = 1$ is more than 0.5 different to that for $\epsilon = 0.1$. When applied to attractive models, we guarantee a FPTAS with no degree restriction.

As described in Section 6.9.2, Bethe pseudomarginals may be recovered from our approach by taking the q^* that is returned as the $\arg\min$ of $\mathcal{F}$ over the discrete mesh. However, although $\mathcal{F}(q^*)$ is guaranteed within ϵ of the optimum, there is no guarantee that q^* will necessarily be close to a true Bethe optimum pseudo-marginal. For example, the surface could be very flat over a wide region, or the true optimum might be $\frac{\epsilon}{2}$ better at a location far from q^*. We sketch out how our approach may be used to bound the location of a global optimum pseudo-marginal, though note that there is no runtime guarantee. First pick an initial ϵ_1 and run the main algorithm to find q_1^*. Now use any method to solve for the second best discretized mesh point q_2^*. If it happens that $\mathcal{F}(q_2^*) \geq \mathcal{F}(q_1^*) + \epsilon_1$

then, by the nature of the mesh construction, there must be a global minimum within the orthotope given by the neighboring mesh points of q_1^* in each dimension[6] and we terminate. On the other hand, if $\mathcal{F}(q_2^*) < \mathcal{F}(q_1^*) + \epsilon_1$ then reduce ϵ_1, for example to $\frac{\epsilon_1}{2}$ and repeat until successful.

Future work includes further reducing the size of the mesh, considering how it should be selected to simplify the subsequent discrete optimization problem, and exploring applications. See Section C.4 for how the mesh size may be improved dramatically for many models when W is very high. Importantly, we now have the opportunity to examine rigorously the performance of the global Bethe optimum. In addition, this will provide a benchmark against which to compare other (non-global) Bethe approaches that typically run more quickly, such as LBP or CCCP (Yuille, 2002). Another interesting avenue is to use our algorithm as a subroutine in a dual decomposition approach to optimize over a tighter relaxation of the marginal polytope, as we explore in Chapter 7.

[6] In fact, the optimum must be within a tighter orthotope based on the *reach* down and up, in each dimension, of q_1^*.

Chapter 7

Understanding the Bethe Approximation: Polytope and Entropy

Belief propagation is a remarkably effective tool for inference, even when applied to networks with cycles. It may be viewed as a way to seek the minimum of the Bethe free energy, though with no convergence guarantee in general. A variational perspective shows that, compared to exact inference, this minimization employs two forms of approximation: (i) the true entropy is approximated by the Bethe entropy, and (ii) the minimization is performed over a relaxation of the marginal polytope termed the local polytope. Here we explore when and how the Bethe approximation can fail for binary pairwise models by examining each aspect of the approximation, deriving results both analytically and with new experimental methods.

7.1 Introduction

Graphical models are a central tool in machine learning. However, the task of inferring the marginal distribution of a subset of variables, termed *marginal inference*, is NP-hard (Cooper, 1990), even to approximate (Dagum and Luby, 1993), and the closely related problem of computing the normalizing partition function is #P-hard (Valiant, 1979). Hence, much work has focused on finding efficient approximate methods. The sum-product message-passing algorithm termed belief propagation is guaranteed to return exact solutions if the underlying topology is a tree. Further, when applied to models with cycles, known as loopy belief propagation (LBP), the method is popular and often

strikingly accurate (McEliece et al., 1998; Murphy et al., 1999).

A variational perspective shows that the true partition function and marginal distributions may be obtained by minimizing the true free energy over the marginal polytope. The standard Bethe approximation instead minimizes the Bethe free energy, which incorporates the Bethe pairwise approximation to the true entropy, over a relaxed pseudo-marginal set termed the local polytope. A fascinating link to LBP was shown (Yedidia et al., 2001), in that fixed points of LBP correspond to stationary points of the Bethe free energy $\mathcal{F}$. Further, stable fixed points of LBP correspond to minima of $\mathcal{F}$ (Heskes, 2002). Werner (2010) demonstrated a further equivalence to stationary points of an alternate function on the space of homogeneous reparameterizations.

In general, LBP may converge only to a local optimum or not converge at all. Various sufficient conditions have been derived for the uniqueness of stationary points (Mooij and Kappen, 2007; Watanabe, 2011), though convergence is often still not guaranteed (Heskes, 2004). Convergent methods based on analyzing derivatives of the Bethe free energy (Welling and Teh, 2001) and double-loop techniques (Heskes et al., 2003) have been developed. Recently, algorithms have been devised that are guaranteed to return an approximately stationary point (Shin, 2012) or a point with value ϵ-close to the optimum (Weller and Jebara, 2013a).

However, there is still much to learn about when and why the Bethe approximation performs well or badly. We shall explore both aspects of the approximation in this paper. Interestingly, sometimes they have opposing effects such that together, the result is better than with just one (see §7.4 for an example). We shall examine minima of the Bethe free energy over three different polytopes: marginal, local and cycle (see §7.2 for definitions). For experiments, we explore two methods, dual decomposition and Frank-Wolfe, which may be of independent interest. To provide another benchmark and isolate the entropy component, we also examine the tree-reweighted (TRW) approximation (Wainwright et al., 2005). Sometimes we shall focus on models where all edges are *attractive*, that is neighboring variables are pulled toward the same value; in this case it is known that the Bethe approximation is a lower bound for the true partition function (Ruozzi, 2012).

Questions we shall address include:

- In attractive models, why does the Bethe approximation perform well for the partition function but, when local potentials are low and coupling high, poorly for marginals?
- In models with both attractive and repulsive edges, for low couplings, the Bethe approxima-

tion performs much better than TRW, yet as coupling increases, this advantage disappears. Can this be repaired by tightening the relaxation of the marginal polytope?

- Does tightening the relaxation of the marginal polytope always improve the Bethe approximation? In particular, is this true for attractive models?

This Chapter is organized as follows. Notation and preliminary results are presented in §7.2. In §7.3-7.4 we derive instructive analytic results, first focusing on the simplest topology that is not a tree, i.e. a single cycle. Already we observe interesting effects from both the entropy and polytope approximations. For example, even for attractive models, the Bethe optimum may lie outside the marginal polytope and tightening the relaxation leads to a worse approximation to the partition function. In §7.5 we examine more densely connected topologies, demonstrating a dramatic phase transition in attractive models as a consequence of the entropy approximation that leads to poor singleton marginals. Experiments are described in §7.6, where we examine test cases. Conclusions are discussed in §7.7. Related work is discussed throughout the text. An Appendix with technical details and proofs is attached at the back.

7.2 Notation and Preliminaries

Throughout this Chapter, we restrict attention to binary pairwise Markov random fields (MRFs). We consider a model with n variables $X_1, \ldots, X_n \in \mathbb{B} = \{0, 1\}$ and graph topology $(\mathcal{V}, \mathcal{E})$; that is $\mathcal{V}$ contains nodes $\{1, \ldots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains an edge for each pairwise relationship. Let $x = (x_1, \ldots, x_n)$ be a configuration of all the variables, and $\mathcal{N}(i)$ be the neighbors of i. Primarily we focus on models with no 'hard' constraints, i.e. $p(x) > 0 \; \forall x$, though many of our results extend to this case. We may reparameterize the potential functions (Wainwright and Jordan, 2008) and define the energy E such that $p(x) = \frac{e^{-E(x)}}{Z}$ with

$$E = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} \frac{W_{ij}}{2} \left[x_i x_j + (1 - x_i)(1 - x_j) \right]. \tag{7.1}$$

This form allows edge coupling weights W_{ij} to be varied independently of the singleton potentials θ_i. If $W_{ij} > 0$ then an edge is *attractive*, if $W_{ij} < 0$ then it is *repulsive*. If all edges are attractive, then the model is attractive. We write μ_{ij} for pairwise marginals and, collecting together the θ_i

and W_{ij} potential terms into a vector θ, with a slight abuse of notation, sometimes write (7.1) as $E = -\theta \cdot \mu$.

7.2.1 Free energy, variational approach

Given any joint probability distribution $q(x)$ over all variables, the (Gibbs) free energy is defined as $\mathcal{F}_G(q) = \mathbb{E}_q(E) - S(q)$, where $S(q)$ is the (Shannon) entropy of the distribution.

It is easily shown (Wainwright and Jordan, 2008) that $-\log Z(\theta) = \min_q \mathcal{F}_G$, with the optimum when $q = p(\theta)$, the true distribution. This optimization is to be performed over all valid probability distributions, that is over the *marginal polytope*. However, this problem is intractable due to the difficulty of both computing the exact entropy S, and characterizing the polytope (Deza and Laurent, 2009).

7.2.2 Bethe approximation

The standard approach of minimizing the *Bethe free energy* $\mathcal{F}$ makes two approximations:

1. The entropy S is approximated by the *Bethe entropy*

$$S_B(\mu) = \sum_{(i,j)\in\mathcal{E}} S_{ij}(\mu_{ij}) + \sum_{i\in\mathcal{V}} (1 - d_i) S_i(\mu_i), \tag{7.2}$$

   where S_{ij} is the entropy of μ_{ij}, S_i is the entropy of the singleton distribution of X_i and $d_i = |\mathcal{N}(i)|$ is the degree of i; and

2. The marginal polytope is relaxed to the *local polytope*, where we require only local (pairwise) consistency, that is we deal with a *pseudo-marginal* vector q, that may not be globally consistent, which consists of $\{q_i = q(X_i = 1)\ \forall i \in \mathcal{V}, \mu_{ij} = q(x_i, x_j)\ \forall (i,j) \in \mathcal{E}\}$ subject to the constraints $q_i = \sum_{j\in\mathcal{N}(i)} \mu_{ij}, q_j = \sum_{i\in\mathcal{N}(j)} \mu_{ij}\ \forall i, j \in \mathcal{V}$.

In general, the Bethe entropy S_B is not concave and hence, the Bethe free energy $\mathcal{F} = E - S_B$ is not convex.

The global optimum of the Bethe free energy $\mathcal{F} = \mathbb{E}_q(E) - S_B(q)$ is achieved by minimizing $\mathcal{F}$ over the local polytope, with the *Bethe partition function* Z_B defined such that the global minimum obtained equals $-\log Z_B$.

The local polytope constraints imply that, given q_i and q_j,

$$\mu_{ij} = \begin{pmatrix} 1 + \xi_{ij} - q_i - q_j & q_j - \xi_{ij} \\ q_i - \xi_{ij} & \xi_{ij} \end{pmatrix} \tag{7.3}$$

for some $\xi_{ij} \in [0, \min(q_i, q_j)]$, where $\mu_{ij}(a, b) = q(X_i = a, X_j = b)$.

As in (Welling and Teh, 2001), one can solve for the Bethe optimal ξ_{ij} explicitly in terms of q_i and q_j by minimizing $\mathcal{F}$, leading to

$$\xi^*_{ij}(q_i, q_j) = \frac{1}{2\alpha_{ij}} \left(Q_{ij} - \sqrt{Q^2_{ij} - 4\alpha_{ij}(1 + \alpha_{ij}) q_i q_j} \right), \tag{7.4}$$

where $\alpha_{ij} = e^{W_{ij}} - 1, Q_{ij} = 1 + \alpha_{ij}(q_i + q_j)$.

Thus, we may consider the Bethe approximation as minimizing $\mathcal{F}$ over $q = (q_1, \ldots, q_n) \in [0, 1]^n$. Further, the derivatives are given by

$$\frac{\partial \mathcal{F}}{\partial q_i} = -\phi_i + \log \left[\frac{(1 - q_i)^{d_i - 1}}{q_i^{d_i - 1}} \prod_{j \in \mathcal{N}(i)} \frac{(q_i - \xi^*_{ij})}{(1 + \xi^*_{ij} - q_i - q_j)} \right], \tag{7.5}$$

where $\phi_i = \theta_i - \frac{1}{2} \sum_{j \in \mathcal{N}(i)} W_{ij}$.

7.2.3 Tree-reweighted approximation

Our primary focus in this paper is on the Bethe approximation but we shall find it helpful to compare results to another form of approximate inference. The *tree-reweighted* (TRW, Wainwright et al., 2005) approach may be regarded as a family of variational methods, where first one selects a point from the *spanning tree polytope*, that is the convex hull of all spanning trees of the model, represented as a weighting for each edge. Given this selection, the corresponding TRW entropy is the weighted combination of entropies on each of the possible trees. This is then combined with the energy and optimized over the local polytope, similarly to the Bethe approximation. Hence it provides an interesting contrast to the Bethe method, allowing us to focus on the difference in the entropy approximation. An important feature of TRW is that its entropy is concave and always upper bounds the true entropy (neither property is true in general for the Bethe entropy). Hence minimizing the TRW free energy is a convex problem and yields an upper bound on the true partition function. Sometimes we shall consider the optimal upper bound, i.e. the lowest upper bound achievable over all possible selections from the spanning tree polytope. For more details, see (Wainwright and Jordan, 2008, §7.2.1).

7.2.4 Cycle polytope

We shall consider an additional relaxation of the marginal polytope termed the *cycle polytope*. This inherits all constraints of the local polytope, hence is at least as tight, and in addition enforces consistency around any cycle. A polyhedral approach characterizes this by requiring the following *cycle inequalities* to be satisfied (Barahona, 1993; Deza and Laurent, 2009; Sontag, 2010) for all cycles C and every subset of edges $F \subseteq C$ with $|F|$ odd:

$$\sum_{(i,j)\in F} (\mu_{ij}(0,0) + \mu_{ij}(1,1)) + \sum_{(i,j)\in C\setminus F} (\mu_{ij}(1,0) + \mu_{ij}(0,1)) \geq 1. \tag{7.6}$$

Each cycle inequality describes a facet of the marginal polytope (Barahona and Mahjoub, 1986). It is typically easier to optimize over the cycle polytope than the marginal polytope, and earlier work has shown that results are often similar (Sontag and Jaakkola, 2007).

7.2.5 Symmetric and homogeneous MRFs

For analytic tractability, we shall often focus on particular forms of MRFs. We say a MRF is *homogeneous* if all singleton potentials are equal, all edge potentials are equal, and its graph has just one vertex and edge orbit.[1]

A MRF is *symmetric* if it has no singleton potentials, hence flipping all variables $0 \leftrightarrow 1$ leaves the energy unchanged, and the true marginals for each variable are $(\frac{1}{2}, \frac{1}{2})$. For symmetric, planar binary pairwise MRFs, it is known that the cycle polytope is equal to the marginal polytope (Barahona and Mahjoub, 1986). Using (7.4) and (7.5), it is easy to show the following result.

Lemma 7.2.1. *The Bethe free energy of any symmetric MRF has a stationary point at* $q_i = \frac{1}{2}\ \forall i$.

We remark that this is *not* always a minimum (see §7.5).

7.2.6 Derivatives and marginals

It is known that the derivatives of $\log Z$ with respect to the potentials are the marginals, and that this also holds for any convex free energy, where pseudo-marginals replace marginals if a polytope

[1] This means there is a graph isomorphism mapping any edge to any other, and the same for any vertex.

other than the marginal is used (Wainwright, 2006). Using Danskin's theorem (Bertsekas, 1995), this can be generalized as follows.

Lemma 7.2.2. *Let $\hat{F} = E - \hat{S}(\mu)$ be any free energy approximation, X be a compact space, and $\hat{A} = -\min_{\mu \in X} \hat{F}$ be the corresponding approximation to* $\log Z$.
If the $\arg\min$ *is unique at pseudo-marginals τ,*
then $\frac{\partial \hat{A}}{\partial \theta_i} = \tau_i(1), \frac{\partial \hat{A}}{\partial W_{ij}} = \tau_{ij}(0,0) + \tau_{ij}(1,1)$.
If the $\arg\min$ *is not unique then let $Q(\theta)$ be the set of* $\arg\min$*s; the directional derivative of $\hat{A}$ in direction*
$\theta \leftarrow \theta + y$ is given by $\nabla_y \hat{A} = \max_{\tau \in Q(\theta)} \tau \cdot y$.

In the next Section we begin to apply these results to analyze the locations and values of the minima of the Bethe free energy.

7.3 Homogeneous Cycles

Since the Bethe approximation is exact for models with no cycles, it is instructive first to consider the case of one cycle on n variables, which we write as C_n. Earlier analysis considered the perspective of belief updates (Weiss, 2000; Aji, 2000). Here we examine the Bethe free energy, which in this context is convex (Pakzad and Anantharam, 2002) with a unique optimum.[2] We consider symmetric models, initially analyzing the homogeneous case.

With Lemma 7.2.1, we see that singleton marginals are $\frac{1}{2}$ across all approximation methods. For pairwise marginals, the following result holds due to convexity.

Lemma 7.3.1. *For any symmetric MRF and a free energy that is convex, the optimum occurs at uniform pseudo-marginals across all pairs of variables, either where the derivative is zero or at an extreme point of the range.*

The uniformity of the optimal edge pseudo-marginals, together with Lemma 7.2.1, shows that all are $\mu_{ij} = \begin{pmatrix} x & \frac{1}{2} - x \\ \frac{1}{2} - x & x \end{pmatrix}$ $\forall (i,j) \in \mathcal{E}$, where just x remains to be identified. The optimum

[2]This follows by considering (7.2) and observing that $S_{ij} - S_i$ (conditional entropy) is concave over the local consistency constraints, hence by appropriate counting, the total Bethe entropy is concave provided an MRF has at most one cycle.

x with zero derivative is always contained within the local polytope but we shall see that this is not always the case when we consider the cycle relaxation. Using (7.4), it is straightforward to derive the following result for the Bethe pairwise marginals.

Lemma 7.3.2. *For a symmetric homogeneous cycle, the Bethe optimum over the local polytope is at* $x = x_B(W) = \frac{1}{2}\sigma(W/2)$, *where we use standard sigmoid* $\sigma(y) := \frac{1}{1+e^{-y}}$. *Observe that* $x_B(-W) = 1/2 - x_B(W)$.

Further, we can derive the error of the Bethe pairwise marginals by using the loop series result given in Lemma 7.4.1 of §7.4, taking $\log$, differentiating and using Lemma 7.2.2, to give the difference between true x and Bethe x_B as

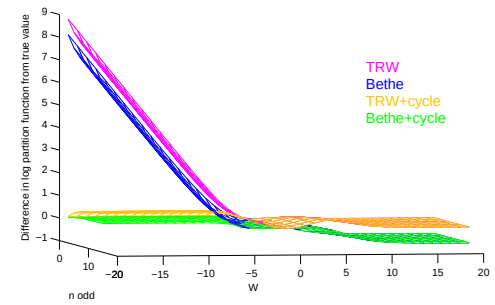
$$x - x_B = \frac{1}{4}\frac{\operatorname{sech}^2 \frac{W}{4} \tanh^{n-1} \frac{W}{4}}{1 + \tanh^n \frac{W}{4}}. \tag{7.7}$$

Remarks: Observe that at $W = 0$, $x - x_B = 0$; as $W \to \pm\infty$, $x - x_B \to 0$. For $W \neq 0$, $x - x_B$ is always > 0 unless n is even and $W < 0$, in which case it is negative. Differentiating (7.7) and solving for where x and x_B are most apart gives empirically $W \approx 2\log n + 0.9$ with corresponding $\max$ value of $x - x_B \approx \frac{1}{5n}$ for large n.

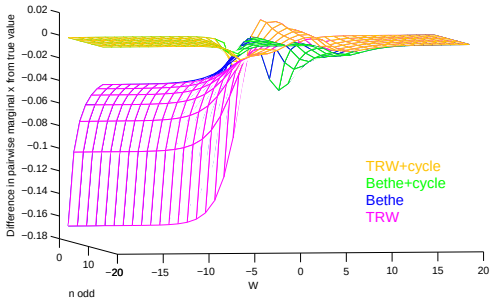
See Figure 7.1 for plots, where, for TRW, values were computed using optimal edge weights, as derived in the Appendix. Observe that at $W = 0$, all methods are exact. As W increases, the Bethe approximations to both $\log Z$ and the marginal x rise more slowly than the true values, though once W is high enough that x is large and cannot rise much further, then the Bethe x_B begins to catch up until they are both close to $\frac{1}{2}$ for large W. We remark that since the Bethe approximation is always a lower bound on the partition function for an attractive model (Ruozzi, 2012), and both the partition functions and marginals are equal at $W = 0$, we know from Lemma 7.2.2 that x_B must rise more slowly than x, as seen.

For $W > 0$, tightening the polytope makes no difference. The picture is different for negative W if n is odd, in which case we have a *frustrated cycle*, that is a cycle with an odd number of repulsive edges, which often causes difficulties with inference methods (Weller and Jebara, 2013b). In this case, (7.6) is binding for $W < -2\log(n-1)$ and prevents the Bethe+cycle marginal x_{BC} from falling below $\frac{1}{2n}$. As $W \to -\infty$, the true x also does not fall below $\frac{1}{2n}$.[3] Thus, as $W \to -\infty$, the

[3]To see this, note there are $2n$ configurations whose probabilities dominate as $W \to -\infty$: $01\ldots0$, its inverse flipping



(a) Errors of $\log Z$ approximations



(b) Errors of pairwise marginal x

Figure 7.1: Homogeneous cycle C_n, n odd, edge weights W. By Lemma 7.2.2, the slope of the error of $\log Z$ wrt W is twice the error of x. For $W > 0$, local and cycle polytopes have the same values.

score (negative energy) and hence $\log Z \to -\infty$ for the true distribution. This also holds for Bethe or TRW on the cycle polytope, but on the local polytope, their energy and $\log Z \to 0$. Observe that for $W < 0$, Bethe generally outperforms TRW over both polytopes.

Tables 7.1 and 7.2 summarize results as $W \to \pm\infty$, again using optimal edge weights for TRW.

7.4 Nonhomogeneous Cycles

The loop series method (Chertkov and Chernyak, 2006; Sudderth et al., 2007) provides a powerful tool to analyze the ratio of the true partition function to its Bethe approximation. In symmetric models with at most one cycle, by Lemma 7.3.1, we know that the unique Bethe optimum is at uniform marginals $q_i = \frac{1}{2}$. Using this and (7.4), and substituting into the loop series result yields the following.

$0 \leftrightarrow 1$, and all n rotations; of these, just one has 00 and one has 11 for a specific edge.

Model	$W \to -\infty$		$W \to \infty$	
	$\log Z'$	x	$\log \frac{Z'}{Z}$	x
Bethe	0	0	$-\log 2$	1/2
Bethe+cycle	0	0	$-\log 2$	1/2
TRW	$\log 2$	0	0	1/2
TRW+cycle	$\log 2$	0	0	1/2
True distribution	$\log 2$	0	0	1/2

Table 7.1: Analytic results for homogenous cycle C_n, n even. As $W \to \infty, \log Z'$ and $\log Z \to \infty$ so the difference is shown.

Model	$W \to -\infty$		$W \to \infty$	
	$\log Z'$	x	$\log \frac{Z'}{Z}$	x
Bethe	0	0	$-\log 2$	1/2
Bethe+cycle	$-\infty$	$1/(2n)$	$-\log 2$	1/2
TRW	$\log 2$	0	0	1/2
TRW+cycle	$-\infty$	$1/(2n)$	0	1/2
True distribution	$-\infty$	$1/(2n)$	0	1/2

Table 7.2: Analytic results for homogeneous cycle C_n, n odd. As $W \to \infty, \log Z'$ and $\log Z \to \infty$ so the difference is shown.

Lemma 7.4.1. *For a symmetric MRF which includes exactly one cycle C_n, with edge weights $W_1, \ldots, W_n$, then $Z/Z_B = 1 + \prod_{i=1}^{n} \tanh \frac{W_i}{4}$.*

Remarks: In this setting, the ratio Z/Z_B is always ≤ 2 and ≈ 1 if even one cycle edge is weak, as might be expected since then the model is almost a tree. The ratio has no dependence on edges not in the cycle and those pairwise marginals will be exact. Further, since the Bethe entropy is concave, by Lemma 7.2.1, all singleton marginals are exact at $\frac{1}{2}$. Errors of pairwise pseudo-marginals on the cycle can be derived by using the expression for Z/Z_B from Lemma 7.4.1, taking $\log$ then differentiating and using Lemma 7.2.2.

Several principles are illustrated by considering 3 variables, A, B and C, connected in a triangle.

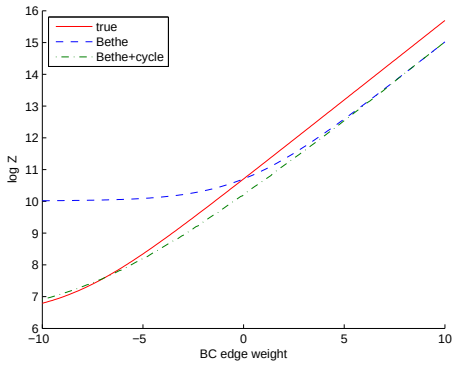


Figure 7.2: Log partition function and approximations for ABC triangle, see §7.4. Edge weights for AB and AC are 10 (strongly attractive) while BC is varied as shown. Near 0: Bethe is a better approximation to $\log Z$ but Bethe+cycle has better derivative, hence better marginals by Lemma 7.2.2; since Bethe+cycle is below Bethe in this region, its optimum does not lie in the local polytope.

Suppose AB and AC have strongly attractive edges with weight $W = 10$. We examine the effect of varying the weight of the third edge BC, see Figure 7.2.

It was recently proved (Ruozzi, 2012) that $Z_B \leq Z$ for attractive models. A natural conjecture is that the Bethe optimum pseudo-marginal in the local polytope must lie inside the marginal polytope. However, our example, when BC is weakly attractive, proves this conjecture to be false. As a consequence, tightening the local polytope to the marginal polytope for the Bethe free energy in this case worsens the approximation of the log-partition function (though it improves the marginals), see Figure 7.2 near 0 BC edge weight. For this model, the two aspects of the Bethe approximation to $\log Z$ act in opposing directions - the result is more accurate with both than with either one alone. For intuition, note that via the path $B - A - C$, in the globally consistent probability distribution, B and C are overwhelmingly likely to take the same value. Given that singleton marginals are $\frac{1}{2}$, the Bethe approximation, however, decomposes into a separate optimization for each edge, which for the weak edge BC, yields that B and C are almost independent, leading to a conflict with the true marginal. This causes the Bethe optimum over the local polytope to lie outside the marginal polytope. The same conclusion may be drawn rigorously by considering the cycle inequality (7.6), taking the edge set $F = \{BC\}$ and observing that the terms are approximately $\frac{1}{4} + \frac{1}{4} + 2(0 + 0) \approx \frac{1}{2} < 1$. Recall that here the cycle and marginal polytopes are the same (see §7.2.5). The same phenomenon can also be shown to occur for the TRW approximation with uniform edge appearance

probabilities.

Notice in Figure 7.2 that as the BC edge strength rises above 0, the Bethe marginals (given by the derivative) improve while the $\log Z$ approximation deteriorates. We remark that the exactness of the Bethe approximation on a tree can be very fragile in the sense that adding a very weak edge between variables to complete a cycle may expose that pairwise marginal as being (perhaps highly) inaccurate.

7.5 General Homogeneous Graphs

We discuss how the Bethe entropy approximation leads to a 'phase shift' in behavior for graphs with more than one cycle when W is above a positive threshold.

The true entropy is always maximized at $q_i = \frac{1}{2}$ for all variables. This also holds for the TRW approximation. However, in densely connected attractive models, the Bethe approximation pulls singleton marginals towards 0 or 1. This behavior has been discussed previously (Heskes, 2004; Mooij and Kappen, 2005a) and described in terms of algorithmic stability (Wainwright and Jordan, 2008, §7.4), or heuristically as a result of LBP over-counting information when going around cycles (Ihler, 2007), but here we explain it as a consequence of the Bethe entropy approximation.

We focus on symmetric homogeneous models which are *d-regular*, i.e. each node has the same degree d. One example is the complete graph on n variables, K_n. For this model, $d = n - 1$. The following result is proved in the Appendix, using properties of the Hessian from (Weller and Jebara, 2013a).

Lemma 7.5.1. *Consider a symmetric homogeneous MRF on n vertices with $d-$regular topology and edge weights W. $q = (\frac{1}{2}, \ldots, \frac{1}{2})$ is a stationary point of the Bethe free energy but for W above a critical value, this is not a minimum. Specifically, let H be the Hessian of the Bethe free energy at q, x_B be the value from Lemma 7.3.2 and $\mathbf{1}$ be the vector of length n with 1 in each dimension; then $\mathbf{1}^T H \mathbf{1} = n[d - 4x_B(d-1)]/x_B < 0$ if $x_B > \frac{1}{4}\frac{d}{d-1} \Leftrightarrow W > 2\log\frac{d}{d-2}$.*

To help understand this result, consider (7.2) for the Bethe entropy S_B, and recall that $\sum_i d_i = 2m$ (m is the number of edges, handshake lemma), hence $S_B = mS_{ij} - (2m - n)S_i$. For large W, all the probability mass for each edge is pulled onto the main diagonal, thus $S_{ij} \approx S_i$. For $m > n$, which interestingly is exactly the case of more than one cycle, in order to achieve the

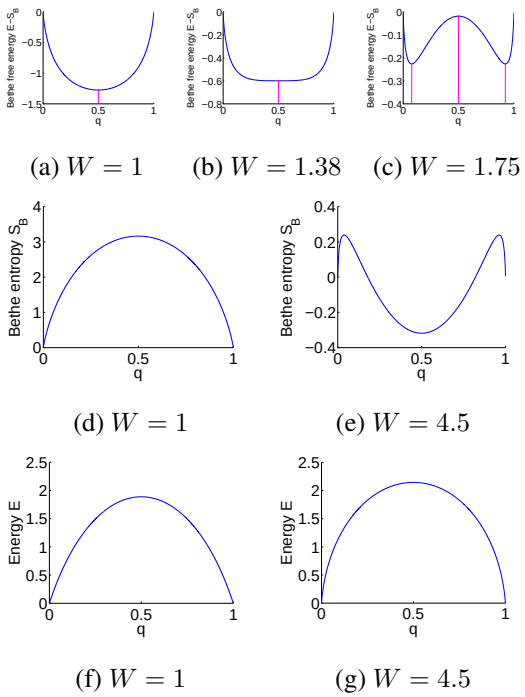
optimum S_B, each entropy term $\to 0$ by tending to pairwise marginal $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ or symmetrically $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. See the second row of Figure 7.3 for an illustration of how the Bethe entropy surface changes dramatically as W rises, even sometimes going negative, and the top row to see how the Bethe free energy surfaces changes rapidly as W moves through the critical threshold.

Reinforcing this pull of singleton marginals away from $\frac{1}{2}$ is the shape of the energy surface, when optimized for free energy subject to given singleton marginals. In the Bethe approximation, this is achieved by computing ξ_{ij} terms according to (7.4), as illustrated in the bottom row of Figure 7.3, but for any reasonable entropy term (including TRW), always $\xi_{ij} < \min(q_i, q_j)$, hence the energy is lower towards the extreme values 0 or 1.

Remarks: (i) This effect is specifically due to the Bethe entropy approximation, and is not affected by tightening the polytope relaxation, as we shall see in §7.6. (ii) To help appreciate the consequences of Lemma 7.5.1, observe that $\log \frac{d}{d-2}$ is positive, monotonically decreasing to 0 as d increases. Thus, for larger, more densely connected topologies, the threshold for this effect is at lower positive edge weights. Above the threshold, $q_i = \frac{1}{2}$ is no longer a minimum but becomes a saddle point.[4] (iii) This explains the observation made after (Heinemann and Globerson, 2011, Lemma 3), where it is pointed out that for an attractive model as $n \to \infty$, if $n/m \to 0$, a marginal distribution (other than the extreme of all 0 or all 1) is unlearnable by the Bethe approximation (because the effect we have described pushes all singleton marginals to 0 or 1). (iv) As W rises, although the Bethe singleton marginals can be poor, the Bethe partition function does not perform badly: For a symmetric model, as $W \to \infty$, there are 2 dominating MAP states (all 0 or all 1) with equal probability. The true marginals are at $q_i = \frac{1}{2}$ which picks up the benefit of $\log 2$ entropy, whereas the Bethe approximation converges to one or other of the MAP states with 0 entropy, hence has $\log 2$ error.

To see why a similar effect does not occur as $W \to -\infty$, note that for $W < 0$ around a frustrated cycle, the minimum energy solution on the local polytope is at $q_i = \frac{1}{2}$. Indeed, this

[4]The Hessian at $q_i = \frac{1}{2}$ is neither positive nor negative definite. Moving away from the valley where all q_i are equal, the Bethe free energy rises quickly.



(a) $W = 1$ (b) $W = 1.38$ (c) $W = 1.75$

(d) $W = 1$ (e) $W = 4.5$

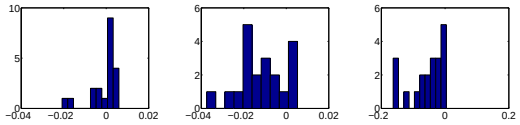
(f) $W = 1$ (g) $W = 4.5$

Figure 7.3: Bethe free energy $E - S_B$ with stationary points highlighted (top), then entropy S_B (middle) and energy E (bottom) vs $q_i = q \ \forall i$ for symmetric homogeneous complete graph K_5. **All quantities are evaluated at the optimum over pairwise marginals**, i.e. $\{\xi_{ij}\}$ are computed as in (7.4). These figures are described in Lemma 7.5.1 and the text thereafter. $W \approx 1.38$ is the critical threshold, above which Bethe singleton marginals are rapidly pulled toward 0 or 1. $W = 4.5$ is sufficiently high that the Bethe entropy becomes negative at $q = \frac{1}{2}$ (middle row).

can pull singleton Bethe marginals *toward* $\frac{1}{2}$ in this case. See §7.5.1 in the Appendix for further analysis.

7.6 Experiments

We are interested in the empirical performance of the optimum Bethe marginals and partition function, as the relaxation of the marginal polytope is tightened. Many methods have been developed to attempt the optimization over the local polytope, primarily addressing its non-convexity, though none is guaranteed to return the global optimum. Recently, an algorithm was derived to return an ϵ-approximation to the optimum $\log Z_B$ based on constructing a discretized mesh of pseudo-marginals (Weller and Jebara, 2013a, 2014a). One method for optimizing over tighter relaxations is



(a) local (b) cycle (c) marginal

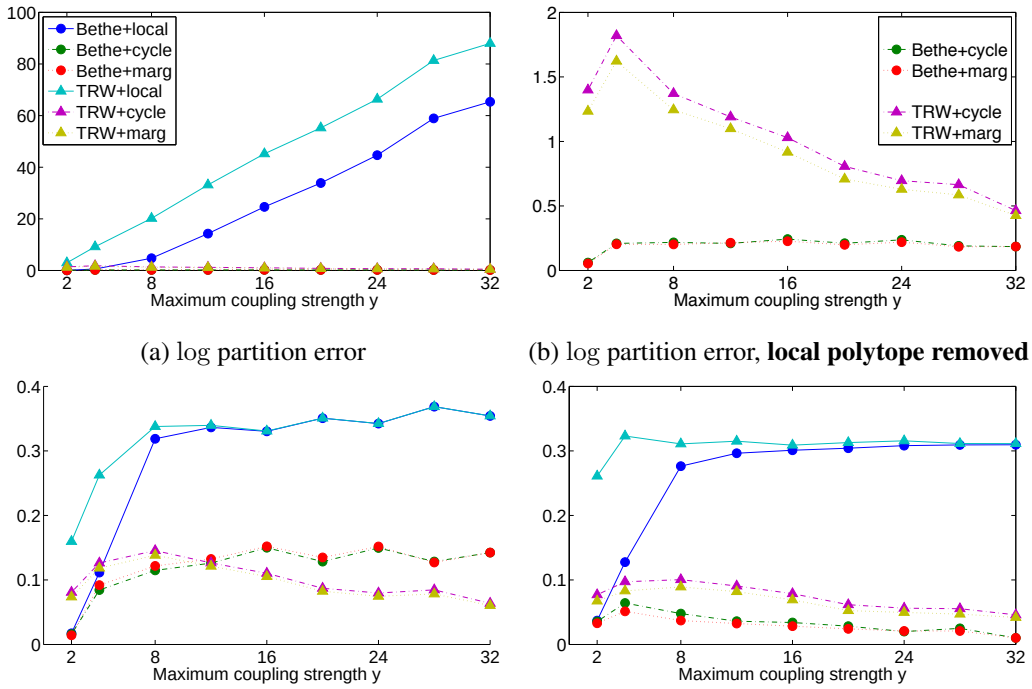
Figure 7.4: Histogram of differences observed in optimum returned Bethe free energy, FW-mesh primal, over the 20 models in the validation set (mesh using $\epsilon = 0.1$, less than ϵ is insignificant). Negative numbers indicate FW outperformed mesh.

to use this algorithm as an inner solver in an iterative dual decomposition approach with subgradient updates (Sontag, 2010; Sontag et al., 2011), where it can be shown that, when minimizing the Bethe free energy, the dual returned less ϵ lower bounds $-\log Z_B$ over the tighter polytope. This would be our preferred approach, but for the models on which we would like to run experiments, the runtime is prohibitive.

Hence we explored two other methods: (i) We replaced the inner solver with a faster, convergent double-loop method, the HAK-BETHE option in libDAI (Heskes et al., 2003; Mooij, 2010), though this is guaranteed only to return a local optimum at each iteration, hence we have no guarantee on the quality of the final result; (ii) We applied the Frank-Wolfe algorithm (FW) (Frank and Wolfe, 1956; Jaggi, 2013; Belanger et al., 2013). At each iteration, a tangent hyperplane is computed at the current point, then a move is made to the best computed point along the line to the vertex (of the appropriate polytope) with the optimum score on the hyperplane. This proceeds monotonically, even on a non-convex surface such as the Bethe free energy, hence will converge (since it is bounded), though runtime is guaranteed only for a convex surface as in TRW.

FW can be applied directly to optimize over marginal, cycle or local polytopes, and performed much better than HAK: runtime was orders of magnitude faster, and the energy found was in line with HAK.[5] To further justify using FW, which may only reach a local optimum, on our main test cases, we compared its performance on a small validation set against the benchmark of dual decomposition using the guaranteed ϵ-approximate mesh method (Weller and Jebara, 2014a) as an inner solver.

[5]The average difference between energies found was < 0.1.



(a) log partition error

(b) log partition error, **local polytope removed**

(c) Singleton marginals, average ℓ_1 error

(d) Pairwise marginals, average ℓ_1 error

Figure 7.5: Results for general models showing error vs true values. $\theta_{\mathbf{i}} \sim \mathbf{U}[-\mathbf{2}, \mathbf{2}]$. **The legend is consistent across plots.** These may be compared to plots in (Sontag and Jaakkola, 2007).

7.6.1 Implementation and validation

To validate FW for the Bethe approximations on each polytope, we compared log partition functions and pairwise marginals across 20 MRFs, each on a complete graph with 5 variables. Each edge potential was drawn $W_{ij} \sim [-8, 8]$ and each singleton potential $\theta_i \sim [-2, 2]$. To handle the tighter polytope relaxations using the mesh method, we used a dual decomposition approach as follows. For the cycle polytope, one Lagrangian variable was introduced for each cycle constraint (7.6) with projected subgradient descent updates. For the marginal polytope, rather than imposing each facet constraint, which would quickly become unmanageable[6], instead a lift-and-project method was employed (Sontag, 2010). These algorithms may be of independent interest and are provided in the Supplement.

For all mesh runs, we used $\epsilon = 0.1$. Note that strong duality is not guaranteed for Bethe since

[6]The number of facets of the marginal polytope grows extremely rapidly (Deza and Laurent, 2009).

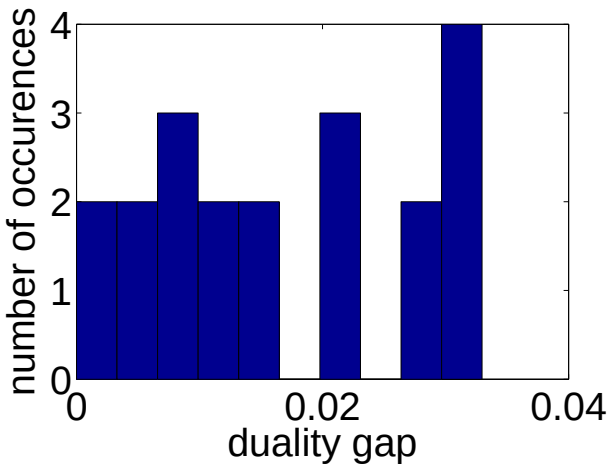


Figure 7.6: Duality gaps observed on the validation set using mesh approach + dual decomposition over 20 models, cycle polytope, $\epsilon = 0.1$. See text in §7.6.1

the objective is non-convex, hence we are guaranteed only an upper bound on $\log Z_B$; yet we were able to monitor the duality gap by using rounded primals and observed that the realized gaps were typically within ϵ, see Figure 7.6.

For FW, we always initialized at the uniform distribution, i.e. $\mu_{ij} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} \; \forall (i,j) \in \mathcal{E}$, note this is always within the marginal polytope. At each iteration, to determine how far to go along the line to the optimum vertex, we used Matlab's fminbnd function. This induces a minimum move of 10^{-6} along the line to the optimum vertex, which was helpful in escaping from local minima. When we tried allowing zero step size, performance became worse. Our stopping criterion was to run for $10,000$ iterations (which did not take long) or until the objective value changed by $< 10^{-6}$, at which point we output the best value found so far, and the corresponding pseudo-marginals.

Results on the validation set are shown in Figure 7.4, indicating that FW performed well compared to mesh + dual decomposition (the best standard we have for the Bethe optimum). Note, however, that good performance on $\log Z_B$ estimation does not necessarily imply that the Bethe optimal marginals were being returned for either method. There may be several local optima where the Bethe free energy has value close to the global optimum, and methods may return different locations. This is a feature of the non-convex surface which should be borne in mind when considering later results, hence we should not be surprised that in the validation set, although $17/20$ of the runs had ℓ_1 error in singleton marginals under 0.05, there were 3 runs with larger differences, in one case as high as 0.7 (not shown).[7]

[7]Recall the example from §7.5, where a symmetric homogeneous MRF with complete graph K_n topology and high

Given this performance, we used FW for all Bethe optimizations on the test cases. FW was also used for all TRW runs, where edge appearance probabilities were obtained using the matrix-tree theorem with weights proportional to each edge's coupling strength $|W_{ij}|$, as was used in (Sontag and Jaakkola, 2007).

7.6.2 Test sets

Models with 10 variables connected in a complete graph were drawn with random potentials. This allows comparison to earlier work such as (Sontag and Jaakkola, 2007) and (Meshi et al., 2009, Appendix). In addition to examining error in $\log$ partition functions and singleton marginals as was done in earlier work, given our theoretical observations in §7.3-7.5, we also explored the error in pairwise marginals. To do this, we report the ℓ_1 error in the estimated probability that a pair of variables is equal, averaged over all edges, i.e. we report average ℓ_1 error of $\mu_{ij}(0,0)+\mu_{ij}(1,1)$. We used FW to minimize the Bethe and TRW free energies over each of the local, cycle and marginal polytopes. For each maximum coupling value used, 100 models were generated and results averaged for plotting. Given the theoretical observations of §7.3-7.5, we are interested in behavior both for attractive and general (non-attractive) models.

For general models, potentials were drawn for single variables $\theta_i \sim U[-2,2]$ and edges $W_{ij} \sim U[-y,y]$ where y was varied to observe the impact of coupling strength.[8] Results are shown in Figure 7.5. Tightening the relaxation of the polytope from local to cycle or marginal, dramatically improves both Bethe and TRW approximations on all measures, with little difference between the cycle or marginal polytopes. This confirms observations in (Sontag and Jaakkola, 2007).

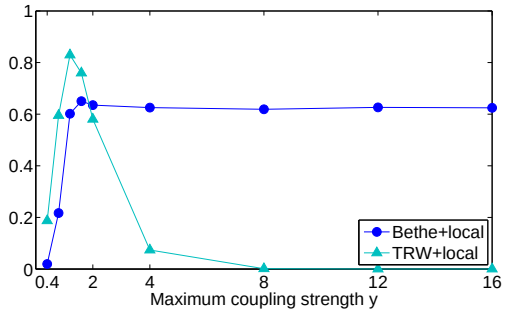
The relative performance of Bethe compared to TRW depends on the criteria used. Looking at the error of singleton marginals, Bethe is better than TRW for low coupling strengths, but for high coupling strengths the methods perform equally well on the local polytope, whereas on the cycle or marginal polytopes, TRW outperforms Bethe (though Bethe is still competitive). Thus, tightening

edge weights was shown to have 2 locations at the global minimum, with average ℓ_1 distance between them approaching 1.

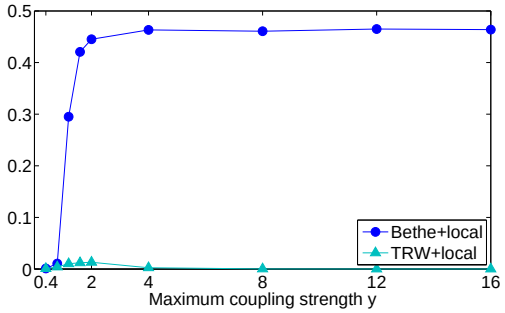
[8] These settings were chosen to facilitate comparison with the results of (Sontag and Jaakkola, 2007), though in that paper, variables take values in $\{-1,1\}$ so the equivalent singleton potential ranges coincide. To compare couplings, our y values should be divided by 4.

the relaxation of the local polytope at high coupling does not lead to Bethe being superior on all measures. However, in terms of partition function and pairwise marginals, which are important in many applications, Bethe does consistently outperform TRW in all settings, and over all polytopes.

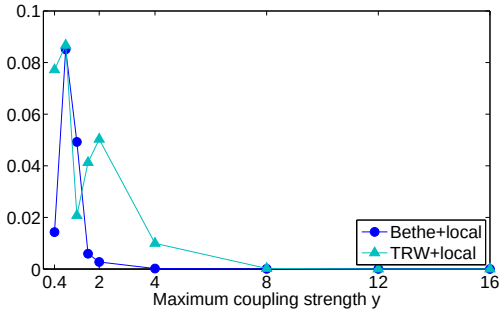
For attractive models, in order to explore our observations in §7.5, much lower singleton potentials were used. We drew $\theta_i \sim U[-0.1, 0.1]$ and $W_{ij} \sim U[0, y]$ where y is varied. This is consistent with parameters used by Meshi et al. (2009). Results are shown in Figure 7.7. When coupling is high, the Bethe entropy approximation pushes singleton marginals away from $\frac{1}{2}$. This effect quickly becomes strong above a threshold. Hence, when singleton potentials are very low, i.e. true marginals are close to $\frac{1}{2}$, the Bethe approximation will perform poorly irrespective of polytope, as observed in our attractive experiments. We note, however, that this effect rarely causes singleton marginals to cross over to the other side of $\frac{1}{2}$. Further, as discussed in §7.5, the partition function approximation is not observed to deviate by more than $\log 2$ on average.



(a) log partition error



(b) Singleton marginals, average ℓ_1 error



(c) Pairwise marginals, average ℓ_1 error. **Note small scale.**

Figure 7.7: Results for attractive models showing error vs true values. $\theta_i \sim \mathbf{U}[-\mathbf{0.1}, \mathbf{0.1}]$. Only local polytope shown, **results for other polytopes are almost identical**.

7.7 Conclusions

We have used analytic and empirical methods to explore the two aspects of the Bethe approximation: the polytope relaxation and the entropy approximation. We found Frank-Wolfe to be an effective method for optimization, and note that for the cycle polytope, the runtime of each iteration scales polynomially with the number of variables (see §7.6.1.3 in the Appendix for further details).

For general models with both attractive and repulsive edges, tightening the relaxation of the polytope from local to cycle or marginal, dramatically improves both Bethe and TRW approximations on all measures, with little difference between the cycle or marginal polytopes. For singleton marginals, except when coupling is low, there does not appear to be a significant advantage to solving the non-convex Bethe free energy formulation compared to convex variational approaches such as TRW. However, for log-partition function estimation, Bethe does provide significant benefits. Empirically, in both attractive and mixed models, Bethe pairwise marginals appear consistently better than TRW.

In our experiments with attractive models, the polytope approximation appears to makes little difference. However, we have shown theoretically that in some cases it can cause a significant effect. In particular, our discussion of nonhomogeneous attractive cycles in §7.4 shows that even in the attractive setting, tightening the polytope can affect the Bethe approximation - improving marginals but worsening the partition function. It is possible that to observe this phenomenon empirically, one needs a different distribution over models.

Chapter 8

Clamping Variables and Approximate Inference

In this Chapter, we apply the earlier analysis on derivatives of the Bethe free energy (Chapter 6) to derive new results. It was recently proved using graph covers (Ruozzi, 2012) that the Bethe partition function is upper bounded by the true partition function for a binary pairwise model that is attractive. Here we provide a new, arguably simpler proof from first principles. We make use of the idea of clamping a variable to a particular value. For an attractive model, we show that summing over the Bethe partition functions for each sub-model obtained after clamping any variable can only raise (and hence improve) the approximation. In fact, we derive a stronger result that may have other useful implications. Repeatedly clamping until we obtain a model with no cycles, where the Bethe approximation is exact, yields the result. We also provide a related lower bound on approximate partition functions of general pairwise multi-label models that depends only on the topology. We demonstrate that clamping a few wisely chosen variables can be of practical value by dramatically reducing approximation error.

8.1 Introduction

Marginal inference and estimating the partition function for undirected graphical models, also called Markov random fields (MRFs), are fundamental problems in machine learning. It is well-known that exact solutions may be obtained via variable elimination or the junction tree method, but unless the

treewidth is bounded, this takes exponential time in general (Pearl, 1988; Lauritzen and Spiegelhalter, 1988b; Wainwright and Jordan, 2008). Hence, much attention has focused on approximate methods, where many approaches have been developed.

Of particular note is the Bethe approximation, which is widely used via the *loopy belief propagation* algorithm (LBP). Though this is typically fast and results are often accurate, in general it may converge only to a local optimum of the Bethe free energy, or may not converge at all (McEliece et al., 1998; Murphy et al., 1999). Another drawback is that, until recently, there were no guarantees on whether the returned approximation to the partition function was higher or lower than the true value. Both aspects are in contrast to methods such as the *tree-reweighted* approximation (TRW, Wainwright et al., 2005), which features a convex free energy and is guaranteed to return an upper bound on the true partition function. Nevertheless, empirically, LBP or convergent implementations of the Bethe approximation often outperform other methods (Meshi et al., 2009; Weller et al., 2014).

Using the method of graph covers (Vontobel, 2013), Ruozzi (2012) recently proved that the optimum Bethe partition function provides a lower bound for the true value, i.e. $Z_B \leq Z$, for discrete binary MRFs with submodular log potential cost functions of any arity. Here we provide an alternative proof for attractive binary pairwise models. Our proof does not rely on any methods of loop series (Sudderth et al., 2007) or graph covers, but rather builds on fundamental properties of the derivatives of the Bethe free energy. Our approach applies only to binary models (whereas Ruozzi, 2012 applies to any arity), but we obtain stronger results for this class, from which $Z_B \leq Z$ easily follows. We use the idea of *clamping* a variable and considering the approximate sub-partition functions over the remaining variables, as the clamped variable takes each of its possible values.

Notation and preliminaries are presented in §8.2. In §8.3, we derive a lower bound, not just for the standard Bethe partition function, but for a range of approximate partition functions over multi-label variables that may be defined from a variational perspective as an optimization problem, based only on the topology of the model. In §8.4, we consider the Bethe approximation for attractive binary pairwise models. We show that clamping any variable and summing the Bethe sub-partition functions over the remaining variables can only increase (hence improve) the approximation. Together with a similar argument to that used in §8.3, this proves that $Z_B \leq Z$ for this class of model. To derive the result, we analyze how the optimum of the Bethe free energy varies as the singleton marginal of one particular variable is fixed to different values in $[0, 1]$. Remarkably, we show that

the negative of this optimum, less the singleton entropy of the variable, is a convex function of the singleton marginal. This may have further interesting implications. We present experiments in §8.5, demonstrating that clamping even a single variable selected using a simple heuristic can be very beneficial.

8.1.1 Related work

Branching or conditioning on a variable (or set of variables) and approximating over the remaining variables has a fruitful history in algorithms such as branch-and-cut (Padberg and Rinaldi, 1991; Mitchell, 2002), work on resolution versus search (Rish and Dechter, 2000) and various approaches of (Darwiche, 2009, Chapter 8). Cutset conditioning was discussed by Pearl (1988) and refined by Peot and Shachter (1991) as a method to render the remaining topology acyclic in preparation for belief propagation. Eaton and Ghahramani (2009) developed this further, introducing the *conditioned belief propagation* algorithm together with *back-belief-propagation* as a way to help identify which variables to clamp. Liu et al. (2012) discussed feedback message passing for inference in Gaussian (not discrete) models, deriving strong results for the particular class of attractive models. Choi and Darwiche (2008) examined methods to approximate the partition function by deleting edges.

8.2 Preliminaries

We consider a pairwise model with n variables $X_1, \ldots, X_n$ and graph topology $(\mathcal{V}, \mathcal{E})$: $\mathcal{V}$ contains nodes $\{1, \ldots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains an edge for each pairwise relationship. We sometimes consider multi-label models where each variable X_i takes values in $\{0, \ldots, L_i - 1\}$, and sometimes restrict attention to binary models where $X_i \in \mathbb{B} = \{0, 1\}\ \forall i$. Let $x = (x_1, \ldots, x_n)$ be a configuration of all the variables, and $\mathcal{N}(i)$ be the neighbors of i. For all analysis of binary models, to be consistent with Welling and Teh (2001) and Weller and Jebara (2013a), we assume a reparameterization such that $p(x) = \frac{e^{-E(x)}}{Z}$, where the energy of a configuration, $E = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} W_{ij} x_i x_j$, with singleton potentials θ_i and edge weights W_{ij}.

8.2.1 Clamping a variable and related definitions

We shall find it useful to examine sub-partition functions obtained by *clamping* one particular variable X_i, that is we consider the model on the $n-1$ variables $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ obtained by setting X_i equal to one of its possible values.

Let $Z|_{X_i=a}$ be the sub-partition function on the model obtained by setting $X_i = a, a \in \{0, \ldots, L_i - 1\}$. Observe that true partition functions and marginals are self-consistent in the following sense:

$$Z = \sum_{j=0}^{L_i-1} Z|_{X_i=j} \; \forall i \in \mathcal{V}, \qquad p(X_i = a) = \frac{Z|_{X_i=a}}{\sum_{j=0}^{L_i-1} Z|_{X_i=j}}. \tag{8.1}$$

This is not true in general for approximate forms of inference,[1] but if the model has no cycles, then in many cases of interest, (8.1) does hold, motivating the following definition.

Definition 8.2.1. We say an approximation to the log-partition function Z_A is *ExactOnTrees* if it may be specified by the variational formula $-\log Z_A = \min_{q \in Q} F_A(q)$ where: (1) Q is some compact space that includes the marginal polytope; (2) F_A is a function of the (pseudo-)distribution q (typically a free energy approximation); and (3) For any model, whenever a subset of variables $\mathcal{V}' \subseteq \mathcal{V}$ is clamped to particular values $P = \{p_i \in \{0, \ldots, L_i - 1\}, \; \forall X_i \in \mathcal{V}'\}$, i.e. $\forall X_i \in \mathcal{V}'$, we constrain $X_i = p_i$, which we write as $\mathcal{V}' \leftarrow P$, and the remaining induced graph on $\mathcal{V} \setminus \mathcal{V}'$ is acyclic, then the approximation is exact, i.e. $Z_A|_{\mathcal{V}' \leftarrow P} = Z|_{\mathcal{V}' \leftarrow P}$. Similarly, define an approximation to be in the broader class of *NotSmallerOnTrees* if it satisfies all of the above properties except that condition (3) is relaxed to $Z_A|_{\mathcal{V}' \leftarrow P} \geq Z|_{\mathcal{V}' \leftarrow P}$. Note that the Bethe approximation is ExactOnTrees, and approximations such as TRW are NotSmallerOnTrees, in both cases whether using the marginal polytope or any relaxation thereof, such as the cycle or local polytope (Weller et al., 2014).

We shall derive bounds on Z_A with the following idea: Obtain upper or lower bounds on the approximation achieved by clamping and summing over the approximate sub-partition functions; Repeat until an acyclic graph is reached, where the approximation is either exact or bounded. We introduce the following related concept from graph theory.

[1] For example, consider a single cycle with positive edge weights. This has $Z_B < Z$ (Weller et al., 2014), yet after clamping any variable, each resulting sub-model is a tree hence the Bethe approximation is exact.

Definition 8.2.2. A *feedback vertex set* (FVS) of a graph is a set of vertices whose removal leaves a graph without cycles. Determining if there exists a feedback vertex set of a given size is a classical NP-hard problem (Karp, 1972). There is a significant literature on determining the minimum cardinality of an FVS of a graph G, which we write as $\nu(G)$. Further, if vertices are assigned non-negative weights, then a natural problem is to find an FVS with minimum weight, which we write as $\nu_w(G)$. An FVS with a factor 2 approximation to $\nu_w(G)$ may be found in time $O(|\mathcal{V}| + |\mathcal{E}| \log |\mathcal{E}|)$ (Bafna et al., 1999). For pairwise multi-label MRFs, we may create a weighted graph from the topology by assigning each node i a weight of $\log L_i$, and then compute the corresponding $\nu_w(G)$.

8.3 Lower Bound on Approximate Partition Functions

We obtain a lower bound on any approximation that is NotSmallerOnTrees by observing that $Z_A \geq Z_A|_{X_n=j}\ \forall j$ from the definition (the sub-partition functions optimize over a subset).

Theorem 8.3.1. *If a pairwise MRF has topology with an FVS of size n and corresponding values $L_1, \ldots, L_n$, then for any approximation that is NotSmallerOnTrees, $Z_A \geq \frac{Z}{\prod_{i=1}^{n} L_i}$.*

Proof. We proceed by induction on n. The base case $n = 0$ holds by the assumption that Z_A is NotSmallerOnTrees. Now assume the result holds for $n - 1$ and consider a MRF which requires n vertices to be deleted to become acyclic. Clamp variable X_n at each of its L_n values to create the approximation $Z_A^{(n)} := \sum_{j=0}^{L_n-1} Z_A|_{X_n=j}$. By the definition of NotSmallerOnTrees, $Z_A \geq Z_A|_{X_n=j}\ \forall j$; and by the inductive hypothesis, $Z_A|_{X_n=j} \geq \frac{Z|_{X_n=j}}{\prod_{i=1}^{n-1} L_i}$.
Hence, $L_n Z_A \geq Z_A^{(n)} = \sum_{j=0}^{L_n-1} Z_A|_{X_n=j} \geq \frac{1}{\prod_{i=1}^{n-1} L_i} \sum_{j=0}^{L_n-1} Z|_{X_n=j} = \frac{Z}{\prod_{i=1}^{n-1} L_i}$. □

By considering an FVS with minimum $\prod_{i=1}^{n} L_i$, Theorem 8.3.1 is equivalent to the following result.

Theorem 8.3.2. *For any approximation that is NotSmallerOnTrees, $Z_A \geq Ze^{-\nu_w}$.*

This bound applies to general multi-label models with any pairwise and singleton potentials (no need for attractive). The bound is trivial for a tree, but already for a binary model with one cycle we obtain that $Z_B \geq Z/2$ for any potentials, even over the marginal polytope. The bound is tight, at

least for uniform $L_i = L \; \forall i$.[2] The bound depends only on the vertices that must be deleted to yield a graph with no cycles, not on the number of cycles (which clearly upper bounds $\nu(G)$). For binary models, exact inference takes time $\Theta((|\mathcal{V}| - |\nu(G)|)2^{\nu(G)})$.

8.4 Attractive Binary Pairwise Models

In this Section, we restrict attention to the standard Bethe approximation. We shall use results derived in (Welling and Teh, 2001) and Chapter 6, and adopt similar notation. The Bethe partition function, Z_B, is defined as in Definition 8.2.1, where Q is set as the *local polytope* relaxation and F_A is the Bethe free energy, given by $\mathcal{F}(q) = \mathbb{E}_q(E) - S_B(q)$, where E is the energy and S_B is the Bethe pairwise entropy approximation (see Chapter 5 for details). We consider attractive binary pairwise models and apply similar clamping ideas to those used in §8.3. In §8.4.1 we show that clamping can never decrease the approximate Bethe partition function, then use this result in §8.4.2 to prove that $Z_B \leq Z$ for this class of model. In deriving the clamping result of §8.4.1, in Theorem 8.4.3 we show an interesting, stronger result on how the optimum Bethe free energy changes as the singleton marginal q_i is varied over $[0, 1]$.

8.4.1 Clamping a variable can only increase the Bethe partition function

Let Z_B be the Bethe partition function for the original model. Clamp variable X_i and form the new approximation $Z_B^{(i)} = \sum_{j=0}^{1} Z_B|_{X_i=j}$. In this Section, we shall prove the following Theorem.

Theorem 8.4.1. *For an attractive binary pairwise model and any variable X_i, $Z_B^{(i)} \geq Z_B$.*

We first introduce notation and derive preliminary results, which build to Theorem 8.4.3, our strongest result, from which Theorem 8.4.1 easily follows. Let $q = (q_1, \ldots, q_n)$ be a location in n-dimensional pseudomarginal space, i.e. q_i is the singleton pseudomarginal $q(X_i = 1)$ in the local polytope. Let $\mathcal{F}(q)$ be the Bethe free energy computed at q using Bethe optimum pairwise pseudomarginals given by the formula for $q(X_i = 1, X_j = 1) = \xi_{ij}(q_i, q_j, W_{ij})$ in (Welling and

[2] For example, in the binary case: consider a sub-MRF on a cycle with no singleton potentials and uniform, very high edge weights. This can be shown to have $Z_B \approx Z/2$ (see Section 7.4). Now connect ν of these together in a chain using very weak edges (this construction is due to Nicholas Ruozzi).

Teh, 2001), i.e. for an attractive model, for edge (i, j), ξ_{ij} is the lower root of

$$\alpha_{ij}\xi_{ij}^2 - [1 + \alpha_{ij}(q_i + q_j)]\xi_{ij} + (1 + \alpha_{ij})q_i q_j = 0, \tag{8.2}$$

where $\alpha_{ij} = e^{W_{ij}} - 1$, and $W_{ij} > 0$ is the strength (associativity) of the log-potential edge weight.

Let $\mathcal{G}(q) = -\mathcal{F}(q)$. Note that $\log Z_B = \max_{q\in[0,1]^n} \mathcal{G}(q)$. For any $x \in [0, 1]$, consider the optimum constrained by holding $q_i = x$ fixed, i.e. let $\log Z_{Bi}(x) = \max_{q\in[0,1]^n : q_i = x} \mathcal{G}(q)$. Let $r^*(x) = (r_1^*(x), \ldots, r_{i-1}^*(x), r_{i+1}^*(x), \ldots, r_n^*(x))$ with corresponding pairwise terms $\{\xi_{ij}^*\}$, be an $\arg\max$ for where this optimum occurs. Observe that $\log Z_{Bi}(0) = \log Z_B|_{X_i=0}, \log Z_{Bi}(1) = \log Z_B|_{X_i=1}$ and $\log Z_B = \log Z_{Bi}(q_i^*) = \max_{q\in[0,1]^n} \mathcal{G}(q)$, where q_i^* is a location of X_i at which the global optimum is achieved.

To prove Theorem 8.4.1, we need a sufficiently good upper bound on $\log Z_{Bi}(q_i^*)$ compared to $\log Z_{Bi}(0)$ and $\log Z_{Bi}(1)$. First we demonstrate what such a bound could be, then prove that this holds. Let $S_i(x) = -x \log x - (1 - x) \log(1 - x)$ be the standard singleton entropy.

Lemma 8.4.2 (Demonstrating what would be a sufficiently good upper bound on $\log Z_B$)**.** *If* $\exists x \in [0, 1]$ *such that* $\log Z_B \le x \log Z_{Bi}(1) + (1 - x) \log Z_{Bi}(0) + S_i(x)$*, then:*
(i) $Z_{Bi}(0) + Z_{Bi}(1) - Z_B \ge e^m f_c(x)$ *where* $f_c(x) = 1 + e^c - e^{xc+S_i(x)}$,
$m = \min(\log Z_{Bi}(0), \log Z_{Bi}(1))$ *and* $c = |\log Z_{Bi}(1) - \log Z_{Bi}(0)|$*; and*
(ii) $\forall x \in [0, 1]$, $f_c(x) \ge 0$ *with equality iff* $x = \sigma(c) = 1/(1 + \exp(-c))$*, the sigmoid function.*

Proof. (i) This follows easily from the assumption. (ii) This is easily checked by differentiating. It is also given in (Koller and Friedman, 2009, Proposition 11.8). □

See Figure E.1 in the Supplement for example plots of the function $f_c(x)$. Lemma 8.4.2 motivates us to consider if perhaps $\log Z_{Bi}(x)$ might be upper bounded by $x \log Z_{Bi}(1) + (1 - x) \log Z_{Bi}(0) + S_i(x)$, i.e. the linear interpolation between $\log Z_{Bi}(0)$ and $\log Z_{Bi}(1)$, plus the singleton entropy term $S_i(x)$. It is easily seen that this would be true if $r^*(q_i)$ were constant. In fact, we shall show that $r^*(q_i)$ varies in a particular way which yields the following, stronger result, which, together with Lemma 8.4.2, will prove Theorem 8.4.1.

Theorem 8.4.3. *Let* $A_i(q_i) = \log Z_{Bi}(q_i) - S_i(q_i)$*. For an attractive binary pairwise model,* $A_i(q_i)$ *is a convex function.*

Proof. We outline the main points of the proof. Observe that $A_i(x) = \max_{q\in[0,1]^n:q_i=x} \mathcal{G}(q) - S_i(x)$, where $\mathcal{G}(q) = -\mathcal{F}(q)$. Note that there may be multiple $\arg\max$ locations $r^*(x)$. As shown in Chapter 6 and Appendix C, $\mathcal{F}$ is at least thrice differentiable in $(0,1)^n$ and all stationary points lie in the interior $(0,1)^n$. Given our conditions, the 'envelope theorem' of (Milgrom, 1999, Theorem 1) applies, showing that A_i is continuous in $[0,1]$ with right derivative[3]

$$A'_{i+}(x) = \max_{r*(q_i=x)} \frac{\partial}{\partial x}\left[\mathcal{G}(q_i = x, r^*(x)) - S_i(x)\right] = \max_{r*(q_i=x)} \frac{\partial}{\partial x}\left[\mathcal{G}(q_i = x, r^*(x))\right] - \frac{dS_i(x)}{dx}. \tag{8.3}$$

We shall show that this is non-decreasing, which is sufficient to show the convexity result of Theorem 8.4.3. To evaluate the right hand side of (8.3), we use the derivative shown by Welling and Teh (2001):

$$\frac{\partial \mathcal{F}}{\partial q_i} = -\theta_i + \log Q_i,$$

$$\text{where } \log Q_i = \log \frac{(1-q_i)^{d_i-1}}{q_i^{d_i-1}} \frac{\prod_{j\in\mathcal{N}(i)}(q_i - \xi_{ij})}{\prod_{j\in\mathcal{N}(i)}(1+\xi_{ij}-q_i-q_j)} \text{ (as in Section 6.5)}$$

$$= \log \frac{q_i}{1-q_i} + \log \prod_{j\in\mathcal{N}(i)} Q_{ij}, \text{ here defining } Q_{ij} = \left(\frac{q_i - \xi_{ij}}{1+\xi_{ij}-q_i-q_j}\right)\left(\frac{1-q_i}{q_i}\right).$$

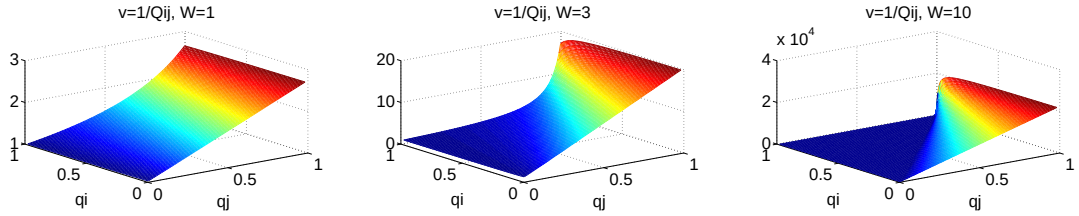
A key observation is that the $\log \frac{q_i}{1-q_i}$ term is exactly $-\frac{dS_i(q_i)}{dq_i}$, and thus cancels the $-\frac{dS_i(x)}{dx}$ term at the end of (8.3). Hence, $A'_{i+}(q_i) = \max_{r*(q_i)} \left[-\sum_{j\in\mathcal{N}(i)} \log Q_{ij}(q_i, r^*_j, \xi^*_{ij})\right]$. [4]

It remains to show that this expression is non-decreasing with q_i. We shall show something stronger, that at every $\arg\max r^*(q_i)$, and for all $j \in \mathcal{N}(i)$, $-\log Q_{ij}$ is non-decreasing $\Leftrightarrow$ $v_{ij} = Q_{ij}^{-1}$ is non-decreasing. The result then follows since the $\max$ of non-decreasing functions is non-decreasing.

See Figure 8.1 for example plots of the v_{ij} function, and observe that v_{ij} appears to decrease with q_i (which is unhelpful here) while it increases with q_j. Now, in an attractive model, the Bethe free energy is *submodular*, i.e. $\frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j} \leq 0$ (Section 6.5), hence as q_i increases, $r^*_j(q_i)$ can only increase (Topkis, 1978). For our purpose, we must show that $\frac{dr^*_j}{dq_i}$ is sufficiently large such that $\frac{dv_{ij}}{dq_i} \geq 0$. This forms the remainder of the proof.

[3] This result is similar to Danskin's theorem (Bertsekas, 1995). Intuitively, for multiple $\arg\max$ locations, each may increase at a different rate, so here we must take the $\max$ of the derivatives over all the $\arg\max$.

[4] We remark that Q_{ij} is the ratio $\left(\frac{p(X_i=1,X_j=0)}{p(X_i=0,X_j=0)}\right) \Big/ \left(\frac{p(X_i=1)}{p(X_i=0)}\right) = \frac{p(X_j=0|X_i=1)}{p(X_j=0|X_i=0)}$.



(a) W=1 (b) W=3 (c) W=10

Figure 8.1: 3d plots of $v_{ij} = Q_{ij}^{-1}$, using $\xi_{ij}(q_i, q_j, W)$ from (Welling and Teh, 2001).

At any particular $\arg\max r^*(q_i)$, writing $v = v_{ij}[q_i, r_j^*(q_i), \xi_{ij}^*(q_i, r_j^*(q_i))]$, we have

$$\begin{aligned}\frac{dv}{dq_i} &= \frac{\partial v}{\partial q_i} + \frac{\partial v}{\partial \xi_{ij}}\frac{d\xi_{ij}^*}{dq_i} + \frac{\partial v}{\partial q_j}\frac{dr_j^*}{dq_i} \\ &= \frac{\partial v}{\partial q_i} + \frac{\partial v}{\partial \xi_{ij}}\frac{\partial \xi_{ij}^*}{\partial q_i} + \frac{dr_j^*}{dq_i}\left(\frac{\partial v}{\partial \xi_{ij}}\frac{\partial \xi_{ij}^*}{\partial q_j} + \frac{\partial v}{\partial q_j}\right). \end{aligned} \tag{8.4}$$

From Lemma C.0.4, $\frac{\partial \xi_{ij}}{\partial q_i} = \frac{\alpha_{ij}(q_j - \xi_{ij}) + q_j}{1 + \alpha_{ij}(q_i - \xi_{ij} + q_j - \xi_{ij})}$ and similarly, $\frac{\partial \xi_{ij}}{\partial q_j} = \frac{\alpha_{ij}(q_i - \xi_{ij}) + q_i}{1 + \alpha_{ij}(q_j - \xi_{ij} + q_i - \xi_{ij})}$, where $\alpha_{ij} = e^{W_{ij}} - 1$. The other partial derivatives are easily derived: $\frac{\partial v}{\partial q_i} = \frac{q_i(q_j - 1)(1 - q_i) + (1 + \xi_{ij} - q_i - q_j)(q_i - \xi_{ij})}{(1 - q_i)^2(q_i - \xi_{ij})^2}$, $\frac{\partial v}{\partial \xi_{ij}} = \frac{q_i(1 - q_j)}{(1 - q_i)(q_i - \xi_{ij})^2}$, and $\frac{\partial v}{\partial q_j} = \frac{-q_i}{(1 - q_i)(q_i - \xi_{ij})}$.

The only remaining term needed for (8.4) is $\frac{dr_j^*}{dq_i}$. The following results are proved in the Appendix, subject to a technical requirement that at an $\arg\max$, the reduced Hessian $H_{\setminus i}$, i.e. the matrix of second partial derivatives of $\mathcal{F}$ after removing the ith row and column, must be non-singular in order to have an invertible locally linear function. Call this required property $\mathcal{P}$. By nature, each $H_{\setminus i}$ is positive semi-definite. If needed, a small perturbation argument allows us to assume that no eigenvalue is 0, then in the limit as the perturbation tends to 0, Theorem 8.4.3 holds since the limit of convex functions is convex. Let $[n] = \{1, \ldots, n\}$ and G be the topology of the MRF.

Theorem 8.4.4. *For any $k \in [n] \setminus i$, let C_k be the connected component of $G \setminus i$ that contains X_k. If $C_k + i$ is a tree, then $\frac{dr_k^*}{dq_i} = \prod_{(s \to t) \in P(i \rightsquigarrow k)} \frac{\xi_{st}^* - r_s^* r_t^*}{r_s^*(1 - r_s^*)}$, where $P(i \rightsquigarrow k)$ is the unique path from i to k in $C_k + i$, and for notational convenience, define $r_i^* = q_i$. Proof in Appendix (subject to $\mathcal{P}$).*

In fact, this result applies for any combination of attractive and repulsive edges. The result is remarkable, yet also intuitive. In the numerator, $\xi_{st} - q_s q_t = \mathrm{Cov}_q(X_s, X_t)$, increasing with W_{ij} and equal to 0 at $W_{ij} = 0$, and in the denominator, $q_s(1 - q_s) = \mathrm{Var}_q(X_s)$, hence the ratio is

exactly what is called in finance the beta of X_t with respect to X_s.[5]

In particular, Theorem 8.4.4 shows that for any $j \in \mathcal{N}(i)$ whose component is a tree, $\frac{dr_j^*}{dq_i} = \frac{\xi_{ij}^* - q_i r_j^*}{q_i(1-q_i)}$. The next result shows that in an attractive model, additional edges can only reinforce this sensitivity.

Theorem 8.4.5. *In an attractive model with edge* (i,j), $\frac{dr_j^*(q_i)}{dq_i} \geq \frac{\xi_{ij}^* - q_i r_j^*}{q_i(1-q_i)}$. *Proof in Appendix (subject to* $\mathcal{P}$*).*

Now collecting all terms, substituting into (8.4), and using (8.2), after some algebra yields that $\frac{dv}{dq_i} \geq 0$, as required to prove Theorem 8.4.3. This now also proves Theorem 8.4.1. □

8.4.2 The Bethe partition function lower bounds the true partition function

Theorem 8.4.1, together with an argument similar to the proof of Theorem 8.3.1, easily yields a new proof that $Z_B \leq Z$ for an attractive binary pairwise model.

Theorem 8.4.6 (first proved by Ruozzi, 2012)**.** *For an attractive binary pairwise model,* $Z_B \leq Z$.

Proof. We shall use induction on n to show that the following statement holds for all n:
If a MRF may be rendered acyclic by deleting n vertices $v_1, \ldots, v_n$, then $Z_B \leq Z$.
The base case $n = 0$ holds since the Bethe approximation is ExactOnTrees. Now assume the result holds for $n-1$ and consider a MRF which requires n vertices to be deleted to become acyclic. Clamp variable X_n and consider $Z_B^{(n)} = \sum_{j=0}^{1} Z_B|_{X_n=j}$. By Theorem 8.4.1, $Z_B \leq Z_B^{(n)}$; and by the inductive hypothesis, $Z_B|_{X_n=j} \leq Z|_{X_n=j}\ \forall j$. Hence, $Z_B \leq \sum_{j=0}^{1} Z_B|_{X_n=j} \leq \sum_{j=0}^{1} Z|_{X_n=j} = Z$. □

8.5 Experiments

For an approximation which is ExactOnTrees, it is natural to try clamping a few variables to remove cycles from the topology. Here we run experiments on binary pairwise models to explore the potential benefit of clamping even just one variable, though the procedure can be repeated. For exact inference, we used the junction tree algorithm. For approximate inference, we used Frank-Wolfe (FW) (Frank and Wolfe, 1956): At each iteration, a tangent hyperplane to the approximate free

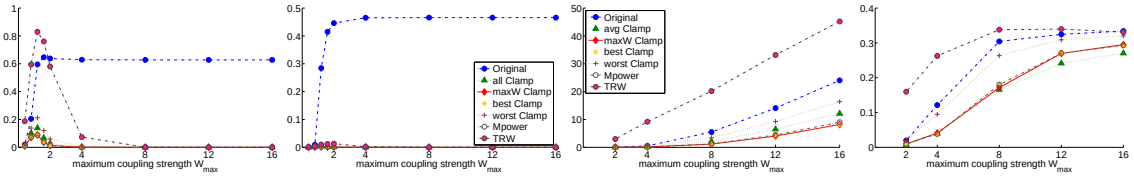
[5] Sudderth et al. (2007) defined a different, symmetric $\beta_{st} = \frac{\xi_{st} - q_s q_t}{q_s(1-q_s)q_t(1-q_t)}$ for analyzing loop series.

energy is computed at the current point, then a move is made to the best computed point along the line to the vertex of the local polytope with the optimum score on the hyperplane. This proceeds monotonically, even on a non-convex surface, hence will converge (since it is bounded), though it may be only to a local optimum and runtime is not guaranteed. This method typically produces good solutions in reasonable time compared to other approaches (Belanger et al., 2013; Weller et al., 2014) and allows direct comparison to earlier results (Meshi et al., 2009; Weller et al., 2014). To further facilitate comparison, in this Section we use the same unbiased reparameterization used by Weller et al. (2014), with $E = -\sum_{i\in\mathcal{V}} \theta_i x_i - \sum_{(i,j)\in\mathcal{E}} \frac{W_{ij}}{2} [x_i x_j + (1-x_i)(1-x_j)]$.

Test models were constructed as follows: For n variables, singleton potentials were drawn $\theta_i \sim U[-T_{max}, T_{max}]$; edge weights were drawn $W_{ij} \sim U[0, W_{max}]$ for attractive models, or $W_{ij} \sim U[-W_{max}, W_{max}]$ for general models. For models with random edges, we constructed Erdős-Renyi random graphs (rejecting disconnected samples), where each edge has independent probability p of being present. To observe the effect of increasing n while maintaining approximately the same average degree, we examined $n = 10, p = 0.5$ and $n = 50, p = 0.1$. We also examined models on a complete graph topology with 10 variables for comparison with TRW in (Weller et al., 2014). 100 models were generated for each set of parameters with varying T_{max} and W_{max} values.

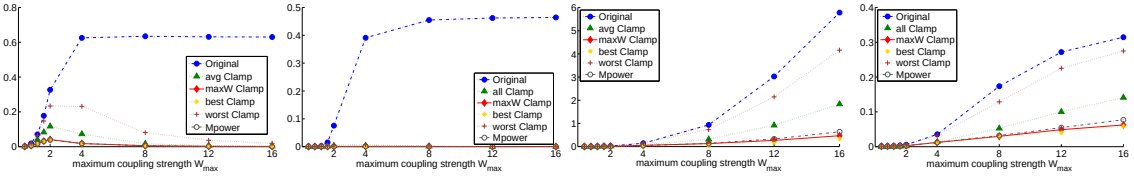
Results are displayed in Figures 8.2 to 8.4 showing average absolute error of $\log Z_B$ vs $\log Z$ and average ℓ_1 error of singleton marginals. The legend indicates the different methods used: *Original* is FW on the initial model; then various methods were used to select the variable to clamp, before running FW on the 2 resulting submodels and combining those results. *avg Clamp* for $\log Z$ means average over all possible clampings, whereas *all Clamp* for marginals computes each singleton marginal as the estimated $\hat{p}_i = Z_B|_{X_i=1}/(Z_B|_{X_i=0} + Z_B|_{X_i=1})$. *best Clamp* uses the variable which with hindsight gave the best improvement in $\log Z$ estimate, thereby showing the best possible result for $\log Z$. Similarly, *worst Clamp* picks the variable which showed worst performance. Where one variable is clamped, the respective marginals are computed thus: for the clamped variable X_i, use $\hat{p}_i$ as before; for all others, take the weighted average over the estimated Bethe pseudomarginals on each sub-model using weights $1 - \hat{p}_i$ and $\hat{p}_i$ for sub-models with $X_i = 0$ and $X_i = 1$ respectively.

maxW and Mpower are heuristics to try to pick a good variable in advance. Ideally, we would like to break heavy cycles, but searching for these is NP-hard. maxW is a simple $O(|\mathcal{E}|)$ method which picks a variable X_i with $\max_{i\in\mathcal{V}} \sum_{j\in\mathcal{N}(i)} |W_{ij}|$, and can be seen to perform well (Liu et al.,



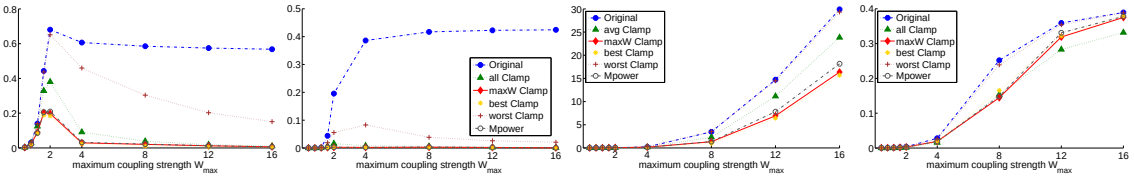
(a) attractive $\log Z, T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z, T_{max} = 2$ (d) general margs, $T_{max} = 2$

Figure 8.2: Average errors vs true, **complete graph on $n = 10$. TRW in pink**. Consistent legend throughout.



(a) attractive $\log Z, T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z, T_{max} = 2$ (d) general margs, $T_{max} = 2$

Figure 8.3: Average errors vs true, **random graph on $n = 10, p = 0.5$**. Consistent legend throughout.

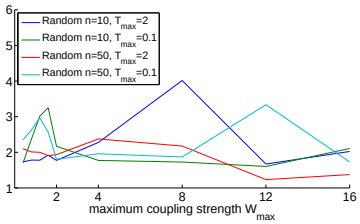


(a) attractive $\log Z, T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z, T_{max} = 2$ (d) general margs, $T_{max} = 2$

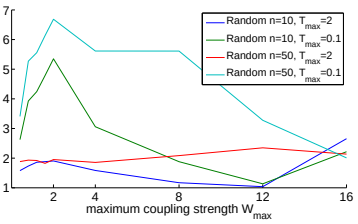
Figure 8.4: Average errors vs true, **random graph on $n = 50, p = 0.1$**. Consistent legend throughout.

2012 proposed the same maxW approach for inference in Gaussian models). One way in which maxW can make a poor selection is to choose a variable at the centre of a large star configuration but far from any cycle. Mpower attempts to avoid this by considering the convergent series of powers of a modified W matrix, but on the examples shown, this did not perform significantly better. See §E.2.1 in the Appendix for more details on Mpower and further experimental results.

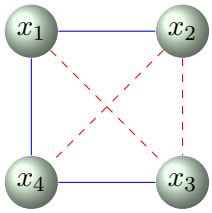
FW provides no runtime guarantee when optimizing over a non-convex surface such as the Bethe free energy, but across all parameters, the average combined runtimes on the two clamped sub-models was the same order of magnitude as that for the original model, see Figure 8.5.



(a) attractive random graphs



(b) general random graphs



(c) Blue (dashed red) edges are attractive (repulsive) with edge weight $+2$ (-2). No singleton potentials.

Figure 8.5: Left: Average ratio of combined sub-model runtimes to original runtime (using maxW, other choices are similar). Right: Example model where *clamping any variable worsens* the Bethe approximation to $\log Z$.

8.6 Discussion

The results of §8.4 immediately also apply to any binary pairwise model where a subset of variables may be flipped to yield an attractive model, i.e. where the topology has no frustrated cycle (Weller et al., 2014), and also to any model that may be reduced to an attractive binary pairwise model (Schlesinger and Flach, 2006; Zivny et al., 2009). For this class, together with the lower bound of §8.3, we have sandwiched the range of Z_B (equivalently, given Z_B, we have sandwiched the range of the true partition function Z) and bounded its error; further, clamping any variable, solving for optimum $\log Z_B$ on sub-models and summing is guaranteed to be more accurate than solving on the original model. In some cases, it may also be faster; indeed, some algorithms such as LBP may fail on the original model but perform well on clamped sub-models.

Methods presented may prove useful for analyzing general (non-attractive) models, or for other applications. As one example, it is known that the Bethe free energy is convex for a MRF whose topology has at most one cycle (Pakzad and Anantharam, 2002). In analyzing the Hessian of the Bethe free energy, we are able to leverage this to show the following result, which may be useful for optimization (proof in Appendix).

Lemma 8.6.1. *In a binary pairwise MRF (attractive or repulsive edges, any topology), for any subset of variables $S \subseteq \mathcal{V}$ whose induced topology contains at most one cycle, the Bethe free energy (using optimum pairwise marginals) over S, holding variables $\mathcal{V} \setminus S$ at fixed singleton marginals, is convex.*

In §8.5, clamping appears to be very helpful, especially for attractive models with low singleton potentials where results are excellent (overcoming TRW's advantage in this context), but also

for general models, particularly with the simple maxW selection heuristic. We can observe some decline in benefit as n grows but this is not surprising when clamping just a single variable. Note, however, that non-attractive models exist such that clamping and summing over *any variable* can lead to a *worse* Bethe approximation of $\log Z$, see Figure 8.5c for a simple example on four variables.

It will be interesting to explore the extent to which our results may be generalized beyond binary pairwise models. Further, it is tempting to speculate that similar results may be found for other approximations. For example, some methods that upper bound the partition function, such as TRW, might always yield a lower (hence better) approximation when a variable is clamped.

Part IV

Conclusions

Chapter 9

Conclusions

Graphical models are a powerful tool for dealing with relationships between variables across a wide array of disciplines including computer vision, speech recognition and computational biology. Core algorithmic tools are needed to address the key challenges of inference. Since these problems are NP-hard, much attention has focused on identifying subclasses of problem where efficient algorithms may be applied, or constructing approximate algorithms with good performance.

In this thesis we have advanced the state of the art in both domains. In Part II, we developed a fascinating, recent link between exact MAP inference and the problem of finding a maximum weight stable set in a derived weighted graph, thus marrying statistical methods of machine learning with recent developments in graph theory. We characterized the power of this approach on the important class of binary pairwise models, provided contributions to the toolbox of methods in this domain, and clarified the range of tractable models. In Appendix B, we have suggested interesting avenues for future exploration.

In Part III, we turned to methods of approximate inference, with particular focus on the Bethe approximation, which is in widespread use through the belief propagation algorithm and its convergent cousins, such as CCCP. Results are often extremely accurate yet, although the ideas behind this approximation arose many decades ago, they are still not properly understood. We made contributions to this understanding by identifying key properties of the approximation, then using these to construct a discrete algorithm guaranteed to return an ϵ-approximation to the Bethe $\log$-partition function (which, to our knowledge, was not previously possible). This will allow the merits of the approximation finally to be tested rigorously. For the important subclass of attractive binary pair-

wise models, our methods provide a fully polynomial time approximation scheme (FPTAS), thus answering a longstanding theoretical question. Further, we explored where and why the two aspects of the approximation (the entropy approximation and the relaxation of the marginal polytope to the local polytope) can lead to error, drawing surprising theoretical conclusions, and providing helpful guidance for practitioners on which tools to apply to real-world problems. Additionally, by making further explorations into the nature of the derivatives of the Bethe free energy and their consequences for clamping methods, we derived a useful general lower bound on a broad class of partition function approximations, and have been able to provide a new, arguably simpler and more intuitive proof, of a landmark result. In doing so, we have derived a stronger result that may have important further implications. There is hope that some of these results may generalize to a broader class of models and entropy approximations.

Part V

Bibliography

Bibliography

A. Abdelbar and S. Hedetniemi. Approximating MAPs for belief networks is NP-hard and other theorems. *Artificial Intelligence*, 102(1):21–38, 1998.

S. Aji. *Graphical models and iterative decoding*. PhD thesis, California Institute of Technology, 2000.

D. Aldous. The ζ (2) limit in the random assignment problem. *Random Structures & Algorithms*, 18(4):381–418, 2001.

N. Alon and M. Tarsi. Covering multigraphs by simple circuits. *SIAM Journal on Algebraic Discrete Methods*, 6:345–350, 1985.

C. Arora, S. Banerjee, P. Kalra, and S. Maheshwari. Generic cuts: An efficient algorithm for optimal inference in higher order MRF-MAP. In *ECCV (5)*, pages 17–30, 2012.

F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

V. Bafna, P. Berman, and T. Fujito. A 2-approximation algorithm for the undirected feedback vertex set problem. *SIAM Journal on Discrete Mathematics*, 12(3):289–9, 1999.

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

F. Barahona. On cuts and matchings in planar graphs. *Math. Program.*, 60:53–68, 1993.

F. Barahona and A. Mahjoub. On the cut polytope. *Mathematical Programming*, 36(2):157–173, 1986. ISSN 0025-5610. doi: 10.1007/BF02592023.

D. Batra, A. Gallagher, D. Parikh, and T. Chen. Beyond trees: MRF inference via outer-planar decomposition. In *CVPR*, pages 2496–2503, 2010.

L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

R. Baxter. Exactly solved models in statistical physics. *Academic, New York*, 1982.

M. Bayati, D. Shah, and M. Sharma. Maximum weight matching via max-product belief propagation. In *IEEE International Symposium on Information Theory*, 2005.

M. Bayati, C. Borgs, J. Chayes, and R. Zecchina. On the exactness of the cavity method for weighted b-matchings on arbitrary graphs and its relation to linear programs. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(06):L06001 (10pp), 2008.

D. Belanger, D. Sheldon, and A. McCallum. Marginal inference in MRFs using Frank-Wolfe. In *NIPS Workshop on Greedy Optimization, Frank-Wolfe and Friends*, December 2013.

C. Berrou and A. Glavieux. Near optimum error correcting coding and decoding: Turbo-codes. *Communications, IEEE Transactions on*, 44(10):1261–1271, 1996.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

H. Bethe. Statistical theory of superlattices. *Proc. R. Soc. Lond. A*, 150(871):552–575, 1935.

J. Bilmes. Mathematical properties of submodularity with applications to machine learning. Machine Learning Summer School Tutorial, Reykjavik, Iceland, May 2014.

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

E. Boros and P. Hammer. Pseudo-boolean optimization. *Discrete Appl. Math.*, 123(1-3):155–225, November 2002. ISSN 0166-218X. doi: 10.1016/S0166-218X(01)00341-9. URL `http://dx.doi.org/10.1016/S0166-218X(01)00341-9`.

S. Boyd and A. Mutapcic. Subgradient Methods, notes for EE364b, Jan 2007. `http://www.stanford.edu/class/ee364b/notes/subgrad_method_notes.pdf`, 2007.

Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

A. Brandstädt and F. Dragan. On linear and circular structure of (claw, net)-free graphs. *Discrete Applied Mathematics*, 129(2-3):285–303, 2003.

T-H. Chan, K. Chang, and R. Raman. An SDP primal-dual algorithm for approximating the Lovász-theta function. In *IEEE International Symposium on Information Theory*, 2009.

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

V. Chandrasekaran, M. Chertkov, D. Gamarnik, D. Shah, and J. Shin. Counting independent sets using the Bethe approximation. *SIAM J. Discrete Math.*, 25(2):1012–1034, 2011.

M. Chertkov and M. Chernyak. Loop series for discrete statistical models on graphs. *J. Stat. Mech.*, 2006.

A. Choi and A. Darwiche. Approximating the partition function by deleting and then correcting for model edges. In *Uncertainty in Artificial Intelligence (UAI)*, 2008.

M. Chudnovsky and P. Seymour. The structure of claw-free graphs. In *London Mathematical Society Lecture Note Series*, volume 324. Cambridge University Press, 2005.

M. Chudnovsky and P. Seymour. Claw-free graphs. i. orientable prismatic graphs. *J. Comb. Theory, Ser. B*, 97(6):867–903, 2007.

M. Chudnovsky and P. Seymour. Claw-free graphs. ii. non-orientable prismatic graphs. *J. Comb. Theory, Ser. B*, 98(2):249–290, 2008a.

M. Chudnovsky and P. Seymour. Claw-free graphs. iii. circular interval graphs. *J. Comb. Theory, Ser. B*, 98(4):812–834, 2008b.

M. Chudnovsky and P. Seymour. Claw-free graphs. iv. decomposition theorem. *J. Comb. Theory, Ser. B*, 98(5):839–938, 2008c.

M. Chudnovsky and P. Seymour. Claw-free graphs. v. global structure. *J. Comb. Theory, Ser. B*, 98 (6):1373–1410, 2008d.

M. Chudnovsky and P. Seymour. Claw-free graphs vi. colouring. *J. Comb. Theory, Ser. B*, 100(6): 560–572, 2010.

M. Chudnovsky and P. Seymour. Claw-free graphs. vii. quasi-line graphs. *J. Comb. Theory, Ser. B*, 102(6):1267–1294, 2012.

M. Chudnovsky, G. Cornuéjols, X. Liu, P. Seymour, and K. Vusković. Recognizing Berge graphs. *Combinatorica*, 25:143–186, 2005a.

M. Chudnovsky, G. Cornuéjols, X. Liu, P. Seymour, and K. Vuskovic. Recognizing Berge graphs. *Combinatorica*, 25(2):143–186, 2005b.

M. Chudnovsky, N. Robertson, P.D. Seymour, and R. Thomas. The strong perfect graph theorem. *Ann. Math*, 164:51–229, 2006.

V. Chvátal. Star-cutsets and perfect graphs. *J. Comb. Theory, Ser. B*, 39(3):189–199, 1985.

V. Chvátal and N. Sbihi. Recognizing claw-free perfect graphs. *J. Comb. Theory, Ser. B*, 44(2): 154–176, 1988.

M. Conforti, G. Cornuéjols, X. Liu, K. Vuskovic, and G. Zambelli. Odd hole recognition in graphs of bounded clique size. *SIAM Journal on Discrete Mathematics*, 20(1):42–48, January 2006.

G. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

M. Cooper, S. de Givry, M. Sanchez, T. Schiex, M. Zytnicki, and T. Werner. Soft arc consistency revisited. *Artif. Intell.*, 174(7-8):449–478, 2010.

D. Coppersmith, D. Gamarnik, M. Hajiaghayi, and G. Sorkin. Random MAX SAT, random MAX CUT, and their phase transitions. *Random Structures & Algorithms*, 24(4):502–545, 2004.

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

P. Dagum and M. Luby. Approximate probabilistic reasoning in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60:141–153, 1993.

A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

R. Dechter. *Constraint processing*. Elsevier Morgan Kaufmann, 2003. ISBN 978-1-55860-890-0.

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

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

K. Doya. *Bayesian brain: Probabilistic approaches to neural coding*. MIT Press, 2007.

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.

J. Edmonds. Paths, trees and flowers. *Canadian Journal of Mathematics*, 17, 1965.

S. Ermon, C. Gomes, A. Sabharwal, and B. Selman. Optimization with parity constraints: From binary codes to discrete integration. In *UAI*, 2013.

Y. Faenza. Solving the maximum weighted stable set problem in claw-free graphs via decomposition. `http://rutcor.rutgers.edu/~mkaminski/AGT/AGTslides/Faenza.pdf`, 2011.

Y. Faenza, G. Oriolo, and G. Stauffer. An algorithmic decomposition of claw-free graphs leading to an $O(n^3)$-algorithm for the weighted stable set problem. In *SODA*, pages 630–646, 2011.

K. Fan. Topological proofs for certain theorems on matrices with non-negative elements. *Monatshefte fr Mathematik*, 62:219–237, 1958.

F. Felzenszwalb and D. Huttenlocher. Efficient belief propagation for early vision. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer*

Society Conference on, volume 1, pages I–261–I–268 Vol.1. IEEE, June 2004. ISBN 0-7695-2158-4. URL http://dx.doi.org/10.1109/cvpr.2004.1315041.

M. Fisher. On the dimer solution of planar Ising models. *Journal of Mathematical Physics*, 7(10): 1776–1781, 1966. doi: http://dx.doi.org/10.1063/1.1704825.

G. Forney. The viterbi algorithm. *Proc. of the IEEE*, 61:268 – 278, March 1973.

J. Foulds, N. Navaroli, P. Smyth, and A. Ihler. Revisiting MAP estimation, message passing, and perfect graphs. In *Artificial Intelligence and Statistics*, 2011.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. ISSN 1931-9193. doi: 10.1002/nav.3800030109.

B. Frey. *Graphical Models for Machine Learning and Digital Communication*. MIT Press, 1998.

R. Gallager. Low-density parity-check codes. *Information Theory, IRE Transactions on*, 8(1):21–28, 1962.

T. Gallai. Graphen mit triangulierbaren ungeraden Vielecken. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 7:3–36, 1962.

M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979. ISBN 0-7167-1044-7.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 1984.

A. Globerson and T. Jaakkola. Approximate inference using planar graph decomposition. In *NIPS*, pages 473–480, 2006.

A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *NIPS*, 2007.

M. Goemans and D. Williamson. .878-approximation algorithms for max cut and max 2sat. In *Proceedings of the Twenty-sixth Annual ACM Symposium on Theory of Computing*, STOC '94, pages 422–431, New York, NY, USA, 1994. ACM. ISBN 0-89791-663-8. doi: 10.1145/195058.195216. URL http://doi.acm.org/10.1145/195058.195216.

A. Goldberg and R. Tarjan. A new approach to the maximum flow problem. *Journal of the ACM*, 35:921–940, 1988.

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.

M. Grötschel, L. Lovász, and A. Schrijver. *Topics on perfect graphs*, chapter Polynomial algorithms for perfect graphs. North-Hollannd, Amsterdam, 1984.

L. Gurvits. Unleashing the power of Schrijver's permanental inequality with the help of the Bethe approximation. *Elec. Coll. Comp. Compl.*, 2011.

R. Halin. S-functions for graphs. *Journal of Geometry*, 8(1-2):171–186, 1976. ISSN 0047-2468. doi: 10.1007/BF01917434.

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

T. Hazan and T. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *ICML*, 2012.

U. Heinemann and A. Globerson. What cannot be learned with Bethe approximations. In *UAI*, pages 319–326, 2011.

H. Hempel and D. Kratsch. On claw-free asteroidal triple-free graphs. *Discrete Appl. Math.*, 121 (1-3):155–180, September 2002. ISSN 0166-218X.

T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Neural Information Processing Systems*, 2002.

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.

T. Hogg, B. Huberman, and C. Williams. Phase transitions and the search problem. *Artificial intelligence*, 81(1):1–15, 1996.

B. Huang and T. Jebara. Loopy belief propagation for bipartite maximum weight b-matching. In *Artificial Intelligence and Statistics*, 2007.

B. Huang and T. Jebara. Approximating the permanent with belief propagation. Technical report, arXiv:0908.1769, 2009.

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

H. Ishikawa. Exact optimization for Markov random fields with convex priors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(10):1333–1336, 2003.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

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

T. Jebara. *Tractability: Practical Approaches to Hard Problems*, chapter Perfect graphs and graphical modeling. Cambridge Press, 2014.

P. Jégou. Decomposition of domains based on the micro-structure of finite constraint-satisfaction problems. In *AAAI*, pages 731–736, 1993.

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

R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

H. Kamisetty, E. Xing, and C. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. In T. Speed and H. Huang, editors, *RECOMB*, volume 4453 of *Lecture Notes in Computer Science*, pages 366–380. Springer, 2007. ISBN 3-540-71680-7. URL `http://dblp.uni-trier.de/db/conf/recomb/recomb2007.html#KamisettyXL07`.

B. Kanefsky and W. Taylor. Where the really hard problems are. In *Proceedings of IJCAI*, volume 91, pages 163–169, 1991.

J. Kappes, B. Andres, F. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. Kausler, J. Lellmann, N. Komodakis, and C. Rother. A comparative study of modern inference techniques for discrete energy minimization problems. In *CVPR*, 2013.

R. Karp. *Complexity of Computer Computations*, chapter Reducibility Among Combinatorial Problems, pages 85–103. New York: Plenum., 1972.

P. Kastelyn. Dimer statistics and phase transitions. *Journal of Math. Physics*, 4:287–293, 1963.

S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–39, 2007.

P. Kohli, A. Shekhovtsov, C. Rother, V. Kolmogorov, and P. Torr. On partial optimality in multi-label MRFs. In W. Cohen, A. McCallum, and S. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 480–487. ACM, 2008. ISBN 978-1-60558-205-4.

D. Koller and N. Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

F. Korč, V. Kolmogorov, and C. Lampert. Approximating marginals using discrete energy minimization. Technical report, IST Austria, 2012.

I. Kovtun. Partial optimal labeling search for a NP-hard subclass of (max, +) problems. In B. Michaelis and G. Krell, editors, *DAGM-Symposium*, volume 2781 of *Lecture Notes in Computer Science*, pages 402–409. Springer, 2003. ISBN 3-540-40861-4.

F. Kschischang, B. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. *IEEE TRANSACTIONS ON INFORMATION THEORY*, 47:498–519, 1998.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

S. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN 0-19-852219-3.

S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society series B*, 50:157–224, 1988a.

S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society series B*, 50:157–224, 1988b.

S. Li. *Markov random field modeling in computer vision*. Springer-Verlag, 1995.

Y. Liu, V. Chandrasekaran, A. Anandkumar, and A. Willsky. Feedback message passing for inference in Gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(8):4135–4150, 2012.

L. Lovász. Normal hypergraphs and the perfect graph conjecture. *Discrete Mathematics*, 2(3): 253–267, 1972.

L. Lovász. Submodular functions and convexity. In A. Bachem, M. Grötschel, and B. Korte, editors, *Mathematical Programming – The State of the Art*, pages 235–257, Berlin, 1983. Springer-Verlag.

D. MacKay and R. Neal. Near Shannon limit performance of low density parity check codes. *Electronics letters*, 32(18):1645–1646, 1996.

R. McEliece, D. MacKay, and J. Cheng. Turbo decoding as an instance of Pearl's "Belief Propagation" algorithm. *IEEE Journal on Selected Areas in Communications*, 16(2):140–152, 1998.

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

P. Milgrom. The envelope theorems. *Department of Economics, Standford University, Mimeo*, 1999. URL `http://www-siepr.stanford.edu/workp/swp99016.pdf`.

P. Milgrom and J. Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, 58(6):1255–1277, 1990.

G. Minty. On maximal independent sets of vertices in claw-free graphs. *J. Comb. Theory, Ser. B*, 28 (3):284–304, 1980.

J. Mitchell. Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of Applied Optimization*, pages 65–77, 2002.

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*, 2005a.

J. Mooij and H. Kappen. Sufficient conditions for convergence of loopy belief propagation. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 396–403. AUAI Press, 2005b.

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.

K. Murphy. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT Press, August 2012. ISBN 0262018020.

K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Uncertainty in Artificial Intelligence (UAI)*, 1999.

G. Oriolo, U. Pietropaoli, and G. Stauffer. A new algorithm for the maximum weighted stable set problem in claw-free graphs. In *IPCO*, pages 77–96, 2008.

A. Osokin, D. Vetrov, and V. Kolmogorov. Submodular decomposition framework for inference in associative Markov networks with global constraints. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1889–1896. IEEE, 2011.

M. Padberg and G. Rinaldi. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. *SIAM review*, 33(1):60–100, 1991.

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

G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 193–200. IEEE, 2011.

K. Parthasarathy and G. Ravindra. The strong perfect-graph conjecture is true for K1,3-free graphs. *J. Comb. Theory, Ser. B*, 21(3):212–223, 1976.

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

R. Peierls and M. Born. On Ising's model of ferromagnetism. *Proc. Camb. Phil. Soc.*, 32(3):477, 1936.

M. Peot and R. Shachter. Fusion and propagation with multiple observations in belief networks. *Artificial Intelligence*, 48(3):299–318, 1991.

P. Pletscher and P. Kohli. Learning low-order models for enforcing high-order statistics. In *Artificial Intelligence and Statistics*, 2012.

W. Pulleyblank and F. Shepherd. Formulations for the stable set polytope of a claw-free graph. In G. Rinaldi and L. Wolsey, editors, *IPCO*, pages 267–279. CIACO, 1993.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 267–296. Morgan Kaufmann Publishers Inc., 1990. ISBN 1-55860-124-4.

I. Rish and R. Dechter. Resolution versus search: Two strategies for SAT. *Journal of Automated Reasoning*, 24(1-2):225–275, 2000.

N. Robertson and P. Seymour. Graph minors. iii. planar tree-width. *Journal of Combinatorial Theory, Series B*, 36(1):49 – 64, 1984. ISSN 0095-8956. doi: http://dx.doi.org/10.1016/0095-8956(84)90013-3.

C. Rother, V. Kolmogorov, V. S. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality. In *CVPR*, 2007.

C. Rudin, D. Waltz, R. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. Gross, B. Huang, and S. Ierome. Machine learning for the New York City power grid. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):328–345, February 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.108.

N. Ruozzi. The Bethe partition function of log-supermodular graphical models. In *Neural Information Processing Systems*, 2012.

S. Sanghavi, D. Malioutov, and A. Willsky. Linear programming analysis of loopy belief propagation for weighted matching. In *Neural Information Processing Systems*, 2008.

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

D. Schlesinger. Exact solution of permuted submodular minsum problems. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 28–38. Springer, 2007.

D. Schlesinger and B. Flach. Transforming an arbitrary minsum problem into a binary one. Technical report, Dresden University of Technology, 2006.

N. Schraudolph and D. Kamenetsky. Efficient exact inference in planar Ising models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1417–1424. Curran Associates, Inc., 2009. URL `http://papers.nips.cc/paper/3390-efficient-exact-inference-in-planar-ising-models.pdf`.

S. Shimony. Finding MAPs for belief networks is NP-hard. *Aritifical Intelligence*, 68(2):399–410, 1994.

J. Shin. Complexity of Bethe approximation. In *Artificial Intelligence and Statistics*, 2012.

D. Sontag. *Approximate Inference in Graphical Models using LP Relaxations*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2010.

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

D. Sontag and T. Jaakkola. Tree block coordinate descent for MAP in graphical models. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AI-STATS)*, volume 8, pages 544–551. JMLR: W&CP, 2009.

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

D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

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

K. Tang, A. Weller, and T. Jebara. Network ranking with Bethe pseudomarginals. In *NIPS Workshop on Discrete Optimization in Machine Learning*, December 2013.

D. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26(2):305–321, 1978.

D. Topkis. *Supermodularity and complementarity*. Princeton University Press, 1998.

L. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2): 189–201, 1979.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13(2):260–269, April 1967.

P. Vontobel. The Bethe permanent of a non-negative matrix. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 341–346. IEEE, 2010.

P. Vontobel. Counting in graph covers: A combinatorial characterization of the Bethe entropy function. *Information Theory, IEEE Transactions on*, 59(9):6018–6048, Sept 2013. ISSN 0018-9448.

M. Wainwright. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7:1829–1859, 2006.

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

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

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.

Johan Wästlund. An easy proof of the ζ (2) limit in the random assignment problem. *manuscript available at the authors webpage*, 2009.

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

Y. Watanabe and M. Chertkov. Belief propagation and loop calculus for the permanent of a non-negative matrix. *Journal of Physics A: Mathematical and Theoretical*, 43(24):242002, 2010.

Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, 2000.

Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *UAI*, pages 416–425, 2007.

A. Weller and T. Jebara. Bethe bounds and approximating the global optimum. In *Artificial Intelligence and Statistics (AISTATS)*, 2013a.

A. Weller and T. Jebara. On MAP inference by MWSS on perfect graphs. In *Uncertainty in Artificial Intelligence (UAI)*, 2013b.

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

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

A. Weller, D. Ellis, and T. Jebara. Structured prediction models for chord transcription of music audio. In *International Conference on Machine Learning and Applications*, 2009.

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.

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

T. Werner. Primal view on belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*, 2010.

J. Woods. Markov image modeling. In *Decision and Control including the 15th Symposium on Adaptive Processes, 1976 IEEE Conference on*, volume 15, pages 596–600. IEEE, 1976.

C. Yanover and Y. Weiss. Approximate inference and protein folding. In *NIPS*, 2002.

C. Yanover, O. Schueler-Furman, and Y. Weiss. Minimizing and learning energy functions for side-chain prediction. *Journal of Computational Biology*, 15(7):899–911, 2008.

J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *International Joint Conference on Artificial Intelligence, Distinguished Lecture Track*, 2001.

J Yeomans. *Statistical mechanics of phase transitions*. Oxford Univ., Oxford, 1992.

E. Yildirim and X. Fan-Orzechowski. On extracting maximum stable sets in perfect graphs using Lovász's theta function. *Computational Optimization and Applications*, 33(2-3):229–247, 2006.

A. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14:1691–1722, 2002.

X. Zhan. Extremal eigenvalues of real symmetric matrices with entries in an interval. *SIAM J. Matrix Analysis Applications*, 27(3):851–860, 2005. URL `http://dx.doi.org/10.1137/050627812`.

W. Zhang. Phase transitions and backbones of the asymmetric traveling salesman problem. *J. Artif. Intell. Res.(JAIR)*, 21:471–497, 2004.

S. Zivny, D. Cohen, and P. Jeavons. The expressive power of binary submodular functions. *Discrete Applied Mathematics*, 157(15):3347–3358, 2009.

Part VI

Appendices

Appendix A

Related Graph Theory

Graphical models provide a rich field in which to explore connections between algorithm development, combinatorial optimization and graph theory. In Part II, a promising, recent approach to exact MAP inference was explored based on reducing the problem to finding a *maximum weight stable set* (MWSS) on a derived weighted graph called a *nand Markov random field* (NMRF) (Jebara, 2009; Sanghavi et al., 2009; Jebara, 2014). In general, finding a MWSS is NP-hard, but if the NMRF is perfect, then a MWSS may be found in polynomial time via the ellipsoid method (Grötschel et al., 1984), thereby efficiently yielding a MAP configuration for the original MRF. In this Appendix, we describe related results from graph theory which may be of interest. Relevant terms and properties are provided in Sections 3.4. For terms not defined, see the papers referenced.

A.1 Recognizing Berge Graphs

Shortly after the Strong Perfect Graph Theorem ('SPGT') was proved (Chudnovsky et al., 2006), Chudnovsky et al. (2005a) solved another open problem by providing a constructive proof that checking if a graph G is Berge (and hence perfect by SPGT) can be carried out in polynomial time, specifically $O(n^9)$. Here we provide a sketch of the algorithm:

- Check in $O(n^9)$ if G or $\bar{G}$ contains any of the following induced subgraphs, each of which can be shown to contain an odd hole: a *pyramid*; a *jewel*; any of the structures designated of type $\mathcal{T}_1, \mathcal{T}_2$ or $\mathcal{T}_3$

- If any of the above is found, output NOT BERGE and stop; else it is proved that every shortest odd hole in G and $\bar{G}$ is *amenable* (hence easy to find)
- Enumerate the $O(n^5)$ possible cleaners X of G
- For each cleaner X, check if $G \setminus X$ contains an amenable hole
- If an amenable hole is found, output NOT BERGE and stop; else
- Repeat the last 3 steps for $\bar{G}$, if still no amenable hole is found, output BERGE

The speed bottleneck here is the $O(n^9)$ time required to test for pyramids. Note it remains an open problem to test in polynomial time if a graph contains an odd hole. For the restricted case of graphs with largest clique size bounded by a constant k, using similar ideas to Chudnovsky et al. (2005a), Conforti et al. (2006) provided a polynomial time algorithm to test for an odd hole in time $O(n^{8k})$, with the speed bottleneck due to the cleaning algorithm used.

A.2 Line, Quasi-line and Claw-free Graphs

Another important class of graphs where a MWSS can be found in polynomial time is *claw-free graphs*, see Figure A.1. A *claw* is a graph isomorphic to $K_{1,3}$ (i.e. a star configuration with one vertex adjacent to each of 3 other vertices, no two of which are adjacent to each other), and a graph is claw-free if it does not contain a claw as an induced subgraph (see 3.3 for basic definitions). Historically, claw-free perfect graphs have attracted attention, being an early class on which the strong perfect graph conjecture was shown to hold, and a class that could be recognized efficiently using decomposition methods (Parthasarathy and Ravindra, 1976; Chvátal and Sbihi, 1988). Interestingly, the MWSS problem on a claw-free graph can be considered a generalization of the weighted matching problem for a general graph in the following way:

- Consider a graph G and its line graph L. A stable set of L is naturally in 1-1 correspondence with a matching of G. Hence, a MWSS of L is equivalent to finding a maximum weight matching of G. This was famously solved by the blossom algorithm (Edmonds, 1965).
- In any line graph L, it is clear that $\forall v \in V(L)$, the neighbors of v can be expressed as the union of two cliques (possibly with edges between the cliques). Any graph satisfying this property is called a *quasi-line graph*.

It is easily seen that any line graph is a quasi-line graph, and any quasi-line graph is claw-free. In each case, the converse is not true. See Figure A.1 for examples.

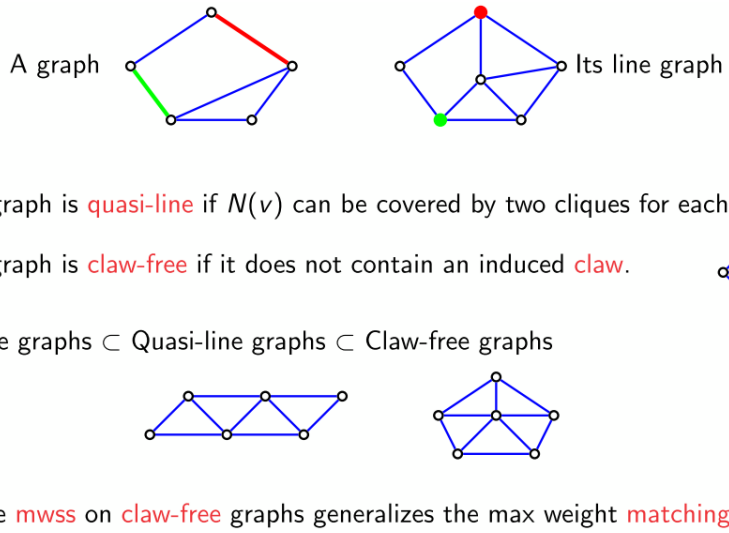


Figure A.1: Illustrations of line, quasi-line and claw-free graphs , from (Faenza, 2011)

The first polynomial time algorithm for finding a MWSS in a general claw-free graph was provided by Minty (1980) using matching-like arguments. Over time, several approaches yielded improvements, but for many years the best runtime was $O(n^6)$. In a recent breakthrough, Faenza et al. (2011) derived an $O(n(m + n \log n))$ algorithm using decomposition methods. This was in part inspired by the success of this approach in analyzing perfect graphs (Chudnovsky et al., 2006) and in a grand characterization of all claw-free graphs (Chudnovsky and Seymour, 2005, 2007, 2008a,b,c,d, 2010, 2012). The decomposition used by Faenza et al. (2011) is weaker but admits efficient algorithms to produce the decomposition and deal with all cases; key ideas are summarized below.

A.2.1 Strips and their composition

To facilitate their decomposition result, Chudnovsky and Seymour (2005) introduced the notion of *strips* and their composition. The version here is slightly modified as used by Faenza et al. (2011).

A *strip* (G, A) is a graph G (not necessarily connected) with a multi-family A of either one *(1-strip)* or two *(2-strip)* designated cliques, called the *extremities* of the strip. Given a family F of k vertex-disjoint strips, $F = \{(G^j, A^j), j \in \{1, \ldots, k\}\}$, and a partition $\mathcal{P} = \{P_1, \ldots, P_m\}$ of

the multi-set of all the extremities $\cup_{j\in\{1,\ldots,k\}} A^j$, the *composition* w.r.t. $\mathcal{P}$ is the graph G defined by: $V(G) = \cup_{j=1}^{k} V(G^j)$; $u, v \in V(G)$ are adjacent iff either they were adjacent in some G^j, or they come from extremities A^i, A^j which are in the same class of $\mathcal{P}$. For this graph G, $(F, \mathcal{P})$ defines a *strip decomposition* of G. Note $\forall P \in \mathcal{P}$, the vertices $\cup_{A\in P} A$ form a clique of G, called a *partition-clique*.

A helpful previous result is from Oriolo et al. (2008): If G is the composition of strips $(G^i, A^i), i = 1, \ldots, k$ and a MWSS can be found for each strip in time $O(p_i(n_i))$, then a MWSS of G can be found in time $O(\sum_{i=1}^{k} p_i(n_i) + n^2 \log n)$.[1]

A.2.2 Summary of the Faenza et al. (2011) algorithm for MWSS of a claw-free graph in $O(n(m + n \log n))$ time

Let G be a claw-free graph with stability number α.

Algorithm 1 runs in $O(mn)$ time and determines either: (i) $\alpha \leq 3$; or (ii) G is *net-free*[2]; or (iii) G is the composition of (*5-wheel*[3] or *distance simplicial*) strips, breaking apart at *articulation cliques*, in which case the algorithm returns the decomposition.

In case (i), find a MWSS by enumeration in $O(mn)$. In case (iii), solve MWSS on distance simplicial strips in $O(m_i)$ by the method of Pulleyblank and Shepherd (1993); solve MWSS on 5-wheel strips in $O(m_i n_i)$ by enumeration; then use the algorithm of Oriolo et al. (2008) to combine to a MWSS of G, total time $O(n(m + n \log n))$. In case (ii), G has no articulation clique, and hence no net; in this case use results of Brandstädt and Dragan (2003) and Hempel and Kratsch (2002), and find a MWSS in $O(n(m + n \log n))$.

---

[1] The composition of strips (G^j, A^j) with each G^j claw-free/quasi-line/line may not lead to a graph with the same property. However, this does hold provided we require that, for each strip, the property must hold on an auxiliary graph where an additional vertex is added for each extremity, with the neighbors of each additional vertex equal to the members of its associated extremity.

[2] A *net* is the graph with six vertices $\{v_1, v_2, v_3, s_1, s_2, s_3\}$ and edges v_1v_2, v_2v_3, v_1v_3, and v_is_i for $i = 1, 2, 3$. In a quasi-line graph, a net clique is always an articulation clique, hence a quasi-line graph with no articulation clique is net-free.

[3] A *5-wheel* is an induced 5-cycle together with an extra vertex called its *center* which is complete to the 5-cycle.

A.2.3 Description of the structure of quasi-line graphs

The following is due to Chudnovsky and Seymour (2005, 2012). A *homogeneous pair of cliques* in G is a pair (A, B) such that:

- A, B are cliques in G with $A \cap B = \emptyset$;
- no vertex in $G \setminus (A \cup B)$ has both a neighbor and a non-neighbor in A; and the same for B; and
- at least one of A or B has cardinality at least 2.

It is shown that every quasi-line graph that does not contain a homogeneous pair of cliques is either a circular interval graph, or a composition of linear interval strips. This is then refined by demonstrating that the only kind of homogeneous pairs of cliques required are those corresponding to the inverse mappings of ends of a fuzzy interval, where at least one of the ends has inverse map with at least two members, leading to the characterization: Every connected quasi-line graph is either a fuzzy circular interval graph, or the composition of fuzzy linear interval strips.

Appendix B

Appendix for the NMRF Approach to MAP Inference

In this Appendix, we consider only binary pairwise MRFs and relax the assumption of Section 4.1. That is, we shall consider reparameterizations where more than one enode from an edge clique group is present in the pruned NMRF. While this can only make it harder to prove perfection of the resulting pruned NMRF, the possible benefit is that there may be less (or no) incident singleton snodes present. This is typically effected by an edge potential 'absorbing' one or both incident singleton variable potentials. Although one might expect that edge nodes can cause more trouble than singleton nodes, in the sense of making odd holes or antiholes more likely (since they typically interact with more nodes in an NMRF), we shall show that in some cases, the net effect is helpful.

In particular, we show in Section B.2 that this technique expands the range of 2-connected models that can be efficiently mapped to perfect pruned NMRFs beyond those identified in Section 4.5 to include:

(i) A general multi-triangle, where the topology is any number of triangles on a common base, allowing all edges to take any weight (attractive or repulsive);

(ii) A frustrated cycle of any size; and

(iii) Some topologies on a complete graph on 4 variables (K_4 topology, which has treewidth 3) containing a frustrated cycle; specifically, the topology must contain at least one non-frustrated triangle.

It is known that the LP relaxation on the triplet-consistent polytope, which we shall call LP+TRI, is tight for any model without a frustrated cycle, and also for any model with treewidth ≤ 2 (Wainwright and Jordan, 2004). This clearly includes the $B_R, T_{m,n}$ and U_n structures of Section 4.5, and also the structures described in (i) and (ii) above. Sontag (2010) has shown that, for models with binary variables, the triplet-consistent polytope TRI is equal to the cycle polytope. Barahona and Mahjoub (1986) showed that the cycle polytope is equal to the marginal polytope for *symmetric* (i.e. no singleton potentials) *planar* binary pairwise models, hence for these models, LP+TRI is always tight. Further, David Sontag has demonstrated a result similar to our Theorem 4.2.1 decomposition result for perfect NMRFs that applies to the LP+TRI approach (unpublished private correspondence). In particular, if LP+TRI is tight on each of two models, which are then pasted together on one common variable, then LP+TRI is tight on the combined model. Hence, LP+TRI can handle all the models described in Section 4.5, and in addition the models described in (i) and (ii) above. From extensive experiments, we believe that LP+TRI might also be tight on models in category (iii), even though these models have treewidth 3 and contain a frustrated cycle.[1]

We leave open the following interesting questions for future work: (1) Can we precisely characterize the set of binary pairwise models that can be handled by the NMRF approach, where we now allow any reparameterization including absorbing singleton nodes? (2) Is this set a subset of the models that can be handled by the LP+TRI approach?

In considering the first question (1), we caution that, while the decomposition result of Theorem 4.2.1 always allows blocks with the $B_R, T_{m,n}$ and U_n structures of Section 4.5 to be pasted together in any fashion, more work is required to check conditions under which the new structures (i), (ii) and (iii) described in this Appendix, may be pasted together while preserving the ability to map efficiently to a perfect pruned NMRF.[2]

[1]For *symmetric* models in category (iii), because the complete graph on 4 variables is planar, the earlier results immediately imply that LP+TRI is tight, but we believe it might also be tight for models of category (iii) with arbitrary singleton potentials, though we have not proved this.

[2]This is because, as stated in Theorem 4.2.1, when pasting blocks together on a variable, the sub-NMRFs must have the same snodes for that variable, but the constructions in this chapter explicitly remove some snodes.

B.1 Absorbing Singleton Potentials, Breaks, Surrogate snodes and Phantom Edges

As described in Section 3.8, singleton transformations can always be applied to leave just one enode per edge in a pruned NMRF, though this will typically leave snodes for each singleton potential. Here we consider edges that *absorb* both incident singleton potentials (one can also consider edges that absorb the singleton potential at just one end, but we leave that for future work), i.e. we consider the reparameterization $\psi'_{ij}(x_i, x_j) = \psi_i(x_i) + \psi_{ij}(x_i, x_j) + \psi_j(x_j), \psi'(x_i) = 0, \psi'(x_j) = 0$. This removes all snodes at i and j, though now we must recognize that any of the 4 possible enodes for such an absorbing edge could be present in the pruned NMRF. In fact, at most 3 enodes will be present after pruning since at least one will be the minimum and can be pruned, though we cannot know in advance which enode this will be. Henceforth we make the conservative (harder) assumption that all 4 enodes are present and show that nevertheless, certain additional MRF structures can be guaranteed to yield a perfect pruned NMRF.

Definition B.1.1. Given a particular reparameterization of a sub-MRF, a *break* at a variable vertex is a missing snode in the pruned NMRF (typically because it has been absorbed by an incident edge) such that it is not possible for enodes from some incident edges to connect through the vertex. A particular reparameterization of a sub-MRF signed topology is *unbroken* if a reparameterization is used such that it contains no breaks.

A break can be very helpful but typically introduces new difficulties. We illustrate the idea and introduce related new terms with the following example. Consider a frustrated 5-cycle $v_1, \ldots, v_5$ that has been reparameterized as in Section 3.8 to have one enode per edge. For a range of singleton potentials, this will be unbroken and hence will form an odd hole in the pruned NMRF, as described in Section 4.5. However, if we arrange that a particular vertex, say v_3, is broken with respect to the incident edges $v_2 - v_3$ and $v_3 - v_4$, then the odd hole could be avoided. This might be achieved as follows: (i) first reparameterize as in Section 3.8 to get one non-zero enode per edge, choosing a reparameterization such that both the $v_2 - v_3$ enode and the $v_3 - v_4$ enode have setting $v_3 = 0$, hence they only connect at v_3 via the snode $(v_3 = 1)$; (ii) now add a *phantom edge* $v_1 - v_3$, which did not exist in the original MRF; this initially has $\psi_{13}(x_1, x_3) = 0 \; \forall x_1, x_3$ but then is reparameterized to absorb the singleton potentials $\psi_1(x_1)$ and $\psi_3(x_3)$, i.e. $\psi'_{13}(x_1, x_3) = \psi_1(x_1) + \psi_3(x_3), \psi'_1 =$

$0, \psi'_3 = 0$. This now prevents the original odd hole from connecting at v_3, apparently solving the problem. However, at least one new odd hole has been introduced, formed by NMRF nodes from the $v_3 - v_4 - v_5 - v_1$ section of the original MRF together with either one or two enodes from the new phantom $v_1 - v_3$ edge. To see this, recall that we have chosen the $v_3 = 0$ setting and suppose that the $v_5 - v_1$ enode has setting $v_1 = x \in \{0, 1\}$. Let $\bar{x} = 1 - x$. We assume that the phantom $v_1 - v_3$ edge has all 4 enodes, so in particular it has $(v_1 = \bar{x}, v_3 = 1)$ which would connect the ends (i.e. the enode for $v_3 - v_4$ and that for $v_5 - v_1$) with one enode, and it has $\{(v_1 = \bar{x}, v_3 = 0), (v_1 = x, v_3 = 1)\}$ which would connect the ends with two enodes. Thus, there is a new odd hole - we solved one problem but introduced another, for no net benefit.

If the 5-cycle were part of a larger MRF, including say v_6 that is unconnected to any of the 5-cycle, one might think that we could form the break at v_3 yet avoid the problem above by introducing a phantom edge $v_3 - v_6$. However, this does not work: either the $(v_3 = 1, v_6 = 0)$ or $(v_3 = 1, v_6 = 1)$ enode could play the role of what we call a *surrogate snode*, i.e. it would play the same role as the original $(v_3 = 1)$ snode did in connecting the original odd hole formed by $v_1, \ldots, v_5$.

Despite these difficulties, we shall show, perhaps surprisingly, that the break idea can allow us to extend the range of signed MRF structures that may be handled efficiently with the NMRF method, for any valid potentials.

B.2 Additional Tractable Models

By reparameterizing to use edges that absorb singleton potentials, we show that all the following structures may be added to the list of those described in Chapter 4. Unless otherwise shown, all singleton and edge potentials may take any value (edges may be attractive or repulsive). Recall Theorem 3.5.5 (Strong Perfect Graph Theorem) and observe that here, to show that a perfect pruned NMRF may be efficiently constructed for these examples, since we cannot rely on the results of Section 4.4, we must check for possible odd holes and antiholes (we need only check for odd antiholes of size ≥ 7 since an antihole of size 5 is isomorphic to a hole of the same size).

B.2.1 General multi-triangles

A general multi-triangle, shown in Figure B.1, is a generalized version of the $T_{m,n}$ and U_n structures described in Section 3.4 and illustrated in Figures 3.2 and 3.3. It consists of n triangles $\{s, v_i, t\}$ on a common base $s - t$. Whereas in the $T_{m,n}$ and U_n structures, the edges are restricted to be either attractive or repulsive in particular configurations, here any edge types are allowed throughout the structure.

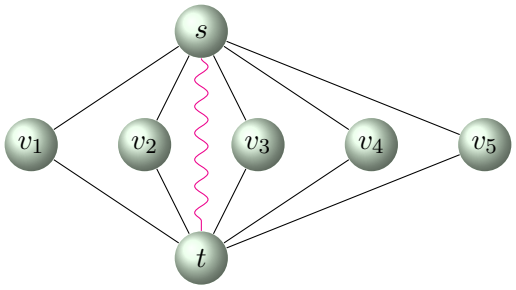


Figure B.1: An example multi-triangle structure with $n = 5$. All edges shown may be attractive or repulsive. The black solid edges are reparameterized to have one enode per edge. The purple wavy edge shows an edge that has been reparameterized so as to absorb its incident snodes.

Reparameterization. For all edges $s - v_i$ and $v_i - t$, reparameterize such that we have just one enode per edge. Always choose the reparameterization such that $s - v_i$ has setting $s = 0$ (if the edge is attractive, use $(s = 0, v_i = 0)$; if repulsive, use $(s = 0, v_i = 1)$). Next $\forall i$ choose the reparameterization of $v_i - t$ such that $s - v_i$ and $v_i - t$ have opposite settings for v_i and hence the enodes connect 'directly' in the NMRF without using an snode at v_i. Let edge $s - t$ be reparameterized so as to absorb the snodes at s and t.

Lemma B.2.1. *With the reparameterization above, a general multi-triangle has a perfect pruned NMRF.*

Proof. (i) *No odd holes.* If a $v_i - t$ and $v_j - t$ enode connect directly at t, then they have different t settings and there's no way to use an $s - t$ enode to connect the $s - v_i$ and $s - v_j$ enodes at s without it also being adjacent to one of the $v_i - t$ or $v_j - t$ enodes, forming a chord. If $v_i - t$ and $v_j - t$ do not connect directly at t then in order to connect to a hole, 2 extra nodes are needed: one

at s and one at t (one can't connect both with one node without forming a chord). Hence, such a hole cannot be odd.

(ii) *No odd antiholes of size* ≥ 7. Suppose an antihole A of size ≥ 7 exists in the pruned NMRF. Let x be a node in A and N_x be those members of A which are adjacent to x, then $|N_x| \geq 4$ and $\forall y \in N_x, y$ is adjacent to at least one other member of N_x. We consider possible candidate members of A. No snode can be in A by Lemma 4.3.1.

An $s - v_i$ enode is adjacent to no other $s - v_j, i \neq j$ by construction. It might be adjacent to a v_i snode, but this can't be in A as just noted, and up to 2 $s - t$ snodes. This is not enough, hence no $s - v_i$ enode can be in A. The only remaining candidates are $v_i - t$ enodes and $s - t$ enodes. Suppose x is a $v_i - t$ enode with setting $t = 0$ (a similar argument applies with setting $t = 1$). N_x can consist only of $v_j - t$ or $s - t$ enodes, all of which must have setting $t = 1$ to be adjacent to x. But each of N_x must be adjacent to ≥ 1 other member of N_x hence none of N_x can be of the form $v_j - t$. So all of N_x must be made up of $s - t$ enodes, but there are at most 2 of these with $t = 1$, contradiction. This leaves $s - t$ nodes as the only possible members, but there are at most 4 of these, contradiction.

□

B.2.2 Frustrated cycles of any size

In Chapter 4, it was shown how to handle a frustrated cycle on 3 vertices. One special case of the multi-triangle structure of Section B.2.1 with $n = 2$ is a 4-cycle $\{s, v_1, t, v_2\}$ with any types of edges, together with an $s-t$ edge that goes across (the purple wavy edge shown in Figure B.1). Note that the 4-cycle could be frustrated, and the $s-t$ edge could be arbitrarily weak. Indeed, it could be non-existent and then be introduced as a phantom edge, as described in Section B.1. Hence, we have shown that a frustrated 4-cycle may be mapped to a perfect pruned NMRF by adding a phantom edge that cuts across the cycle. We shall show how this idea may be modified slightly and extended to allow a frustrated cycle of any size ≥ 5. See Figure B.2 for an example with 6 variables.

Reparameterization. There is a cycle of n variables $x_1, \ldots, x_n$ with $n \geq 5$. Pick one 'star' vertex to which all the phantom edges will be incident. Here we assume the star is x_1, as illustrated in Figure B.2. For all edges of the original cycle, reparameterize such that we obtain just one enode

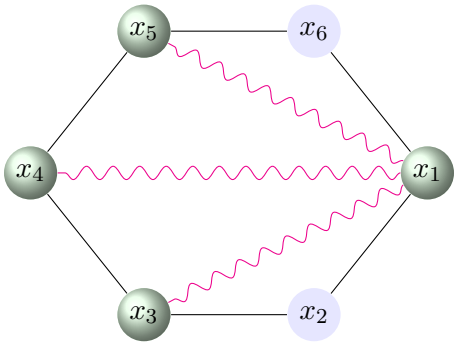


Figure B.2: An example frustrated cycle on 6 variables. All edges shown may be attractive or repulsive. The black solid edges are reparameterized to have one enode per edge. The purple wavy edges indicate phantom edges that were added and reparameterized so as to absorb all incident snodes. x_1 is the 'star' variable, to which all phantom edges are adjacent. The enodes of the original cycle connect directly at x_2 and x_6 (shown in blue, unshaded) but do not connect directly at x_3, x_4 or x_5 (they might or might not connect at x_1).

per edge in the following way. Pick an orientation to take around the cycle starting with the star, and an initial value (0 or 1) for the star variable setting of the first enode. Here we shall assume a clockwise orientation and an initial value of 0 but a similar argument applies for other choices. With these choices, the first enode will be $(x_1 = 0, x_2 = 0)$ if the edge is attractive, or $(x_1 = 0, x_2 = 1)$ if repulsive. Now continue around the cycle in the chosen direction, selecting the enode as follows: at the second and last variables (here x_2 and x_n), ensure that the incident enodes connect directly at the variable, i.e. they have different settings for the variable; at all other variables, ensure that the incident enodes do not connect directly, i.e. they have the same setting for the variable, hence at these variables there will breaks once snodes are absorbed. (This means that for general edge types, after dealing with all the edges and returning to the star at the beginning of the cycle, the edges may either connect or not at the star.) Introduce phantom edges from the star to all non-adjacent variables on the cycle and reparameterize them so as to absorb the incident snodes (the star snodes can be absorbed among them in any way).

Lemma B.2.2. *With the reparameterization above for a cycle C_n of size $n \geq 5$, a perfect pruned NMRF is obtained.*

Proof. Let $s = x_1$ be the star. Let $u - v$ be an edge of C_n where neither u nor v is adjacent to s in the original cycle. By construction, there is one $u - v$ enode, say $(u = x, v = y)$ and its neighbors are all either $s - u$ or $s - v$ enodes.

(i) *No odd holes.* Suppose an odd hole H exists and consider candidate members. Clearly the snodes at x_2 and x_n could not be in H. Consider an edge $u - v$. To form an odd hole, we must have $(s = a, u = \bar{x})$ and $(v = \bar{y}, s = a)$ for some $a \in \{0, 1\}$ else we have a triangle. To continue past $(v = \bar{y}, s = a)$, we cannot have anything with setting $s = \bar{a}$ else it would form a chord, so we must have $(v = y, s = a)$; but then there is no way to continue without forming a chord. Hence there is no such $u - v$ enode in H.

We next consider if enodes from the remaining 4 edges of the original cycle could be in H. The possibilities are: (1) 2 enodes ($\{x_1 - x_2, x_2 - x_3\}$, or $\{x_{n-1} - x_n, x_n - x_1\}$) together with enodes from the connecting phantom edge - but this cannot work since there could be at most 2 enodes from any one edge (else a triangle is formed), so there are insufficient nodes; or (2) all 4 of the enodes under consideration (if they connect at s) together with additional enodes from phantom edges - but this cannot work since, if the 4 enodes connect at s then they have different settings there, and any additional enode from a phantom edge will form a chord. Hence, the only remaining candidates for members of H are the enodes from phantom edges, where more than one phantom edge must be involved to have sufficient nodes. To connect across different phantom edges, there must be enodes with different settings for s but there is no way to do this without at least one enode being adjacent to ≥ 3 others, contradiction.

(ii) *No odd antiholes of size* ≥ 7. We proceed as in the Proof of Lemma B.2.1. Consider if an enode from a $u - v$ edge could be in A. It has 4 neighbors (2 from each incident phantom edge), each of which is adjacent to 2 of the other 4. But to be in A, there must be 2 that are adjacent to 2 of the other 4, and 2 that are adjacent to 1 of the other 4, contradiction. Neither an $x_2 - x_3$ nor $x_5 - x_6$ enode can be in A (each has up to 4 neighbors including the snode, but this cannot be in A by Lemma 4.3.1).

Next consider the enode for $s - x_2$ (the same argument applies to $x_n - s$). Its possible neighbors in A are the enodes from phantom edges at s (the possible neighbor of the enode from $x_n - s$ could not work since it would not be adjacent to any of the other possible neighbors). It can be adjacent to at most 2 enodes from each phantom edge clique group (the 2 which have different setting for s).

Neighbor enodes from different phantom edge groups are not adjacent to each other (they overlap only on s where they have the same setting). The only possibility is 2 enodes from each of 2 phantom edge clique groups, but then each of them is adjacent to just one other neighbor (the one in its same clique group), contradiction.

The only remaining possible nodes of A are enodes from phantom edge clique groups. If one of them, say p, is in A, then there must be a set of 4 neighbors of p, say q_1, q_2, q_3 and q_4, all in order going around A, with p and q_1 both adjacent to q_3 and q_4, and q_3 not adjacent to q_4. Hence, q_3 and q_4 must be in different clique groups with the same setting for s, say $s = a$. To be adjacent to both, p and q_1 must each have setting $s = \bar{a}$. Now q_2 is not adjacent to q_1, so has setting $s = \bar{a}$ but then q_2 is adjacent to q_3, contradiction. □

Remarks. This construction for a single cycle is superficially similar to *triangulation* of the MRF topology, where all added edges are incident to one variable, and suggests a link to the LP+TRI approach.

Note that all the edges introduced as phantom edges could equally be edges of any type that were already in the original MRF.

B.2.3 Some frustrated K_4 structures (treewidth 3)

We require that there is at least one non-frustrated triangle (which could contain 0 or 2 repulsive edges), while all other edges and all singleton potentials are unrestricted. See Figure B.3 for the reparameterization. The edges shown in solid blue must form a non-frustrated triangle, which allows their enodes to be chosen such that none of them connect directly at any of the triangle's vertices x_2, x_3, x_4.

Lemma B.2.3. *With the reparameterization of a K_4 structure shown in Figure B.3, a perfect pruned NMRF is obtained.*

Proof. The argument is similar to that in the proof for frustrated cycles in Section B.2.2. For each of odd holes and odd antiholes: first consider the solid blue edges, which are similar to the $u - v$ edges of a frustrated cycle; next consider the purple wavy edges with 4 enodes, which are similar to the phantom edges of a frustrated cycle. □

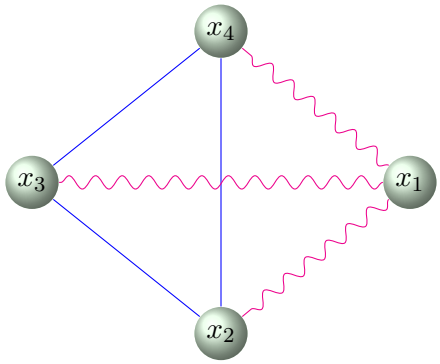


Figure B.3: An example of a frustrated yet tractable K_4. All edges shown may be attractive or repulsive, except that the triangle shown with solid blue edges must be non-frustrated (i.e. contain either 0 or 2 repulsive edges). The solid blue edges are reparameterized to have one enode per edge such that none of them connects directly to another. The purple wavy edges are reparameterized so as to absorb all snodes.

B.3 Discussion

We have provided examples to show that the idea of reparameterizing edges so as to absorb singleton potentials can significantly increase the range of binary pairwise models on which MAP inference is tractable with the NMRF method. We hope to be able to characterize the enlarged set exactly in future work, and to determine if this is a subset of those models that may be handled in polynomial time using the LP+TRI method.

Appendix C

Appendix for Discrete Methods to Approximate the Partition Function

Here we provide details and proofs of several of the results in Chapter 6. To establish these, we also derive additional preliminary results where required.

Lemma C.0.1. *Unless q_i or $q_j \in \{0, 1\}$, all entries of the pseudo-marginal μ_{ij} are strictly > 0, whether (i, j) is associative or repulsive.*[1]

Proof. First assume $\alpha_{ij} > 0$. Considering (6.2) and using Lemmas 6.4.1 and 6.4.3, we have that element-wise

$$\mu_{ij} \geq \begin{pmatrix} (1-q_i)(1-q_j) & q_j(1-q_i)/(1+\alpha_{ij}) \\ q_i(1-q_j)/(1+\alpha_{ij}) & q_i q_j \end{pmatrix} \tag{C.1}$$

which proves the result for this case. If $\alpha_{ij} < 0$ then flip either q_i or q_j. As in the proof of Lemma 6.3.1, pseudo-marginal entries change position but not value. □

In order to prove Theorem 6.4.2, we first show the following result.

Theorem C.0.2. *If all edges incident to X_i are associative then at any stationary point of the Bethe free energy, $\sigma(\theta_i) \leq q_i \leq \sigma(\theta_i + W_i)$. Remark the same sandwich result holds for the true marginal p_i.*

[1] Here we assume α_{ij} is finite, see footnote 1 in Chapter 6.

Proof. We first prove the left inequality. Consider (6.9). Using $\alpha_{ij} > 0 \ \forall j \in \mathcal{N}(i)$ and Lemma 6.4.1 we have

$$\begin{aligned} Q_i &= \frac{\prod_{j\in\mathcal{N}(i)}(q_i - \xi_{ij})}{q_i^{d_i-1}} \frac{(1-q_i)^{d_i-1}}{\prod_{j\in\mathcal{N}(i)}(1+\xi_{ij}-q_i-q_j)} \\ &\leq \frac{\prod_{j\in\mathcal{N}(i)} q_i(1-q_j)}{q_i^{d_i-1}} \frac{(1-q_i)^{d_i-1}}{\prod_{j\in\mathcal{N}(i)}(1-q_i)(1-q_j)} \\ &= \frac{q_i}{1-q_i} \text{ which gives the result.} \end{aligned}$$

To obtain the right inequality, flip all variables as in Section 6.3.1. Using the first inequality, (6.7) and Lemma 6.3.1 yields $1 - q_i \geq \sigma(-\theta_i - W_i) \Leftrightarrow q_i \leq \sigma(\theta_i + W_i)$ since $1 - \sigma(-x) = \sigma(x)$. To show the result for the true marginal, let $m_{i=a} = \sum_{x:x_i=a} \exp(\sum_{i\in V} \theta_i x_i + \sum_{(i,j)\in E} W_{ij} x_i x_j)$ then using (6.1), $p_i = \frac{m_{i=1}}{m_{i=1}+m_{i=0}}$. Since all $W_{ij} > 0$ the result follows. □

Using the technique of flipping variables (6.8), we obtain the more powerful Theorem 6.4.2 as a corollary.

Theorem 6.4.2 For general edge types (associative or repulsive), let $W_i = \sum_{j\in\mathcal{N}(i):W_{ij}>0} W_{ij}$, $V_i = -\sum_{j\in\mathcal{N}(i):W_{ij}<0} W_{ij}$. At any stationary point of the Bethe free energy, $\sigma(\theta_i - V_i) \leq q_i \leq \sigma(\theta_i + W_i)$. The same result holds for the true marginal p_i.

Proof. Using (6.8), flip all variables adjacent to X_i with a repulsive edge, i.e. set $\mathcal{R} = \{j \in \mathcal{N}(i) : W_{ij} < 0\}$. The resulting new model is fully associative around X_i so we may apply Theorem C.0.2 to yield the result. □

A lemma which shall soon prove helpful.

Lemma C.0.3. *For* $q_i, q_j \in [0,1], 0 \leq q_i + q_j - 2q_iq_j \leq 1$.

Proof. Let $f = q_i + q_j - 2q_iq_j$. To show the left inequality, consider $m = \min(q_i, q_j)$ and $M = \max(q_i, q_j)$, then $f \geq 2m(1-M) \geq 0$. For the right inequality observe $1 - f = (1 - q_i)(1 - q_j) + q_iq_j \geq 0$. □

Using Lemma C.0.3, we show the following result.

Lemma 6.4.3 (Upper bound for ξ_{ij} for an attractive edge) If $\alpha_{ij} > 0$, then
$q_j - \xi_{ij} \geq \frac{q_j(1-q_i)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \geq \frac{q_j(1-q_i)}{1+\alpha_{ij}} = \frac{q_j(1-q_i)}{K_{ij}}$,

$q_i - \xi_{ij} \geq \frac{q_i(1-q_j)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \geq \frac{q_i(1-q_j)}{1+\alpha_{ij}} = \frac{q_i(1-q_j)}{K_{ij}}$.
Also $\xi_{ij} \leq m(\alpha_{ij}+M)/(1+\alpha_{ij}) \Rightarrow \xi_{ij} - q_iq_j \leq \frac{\alpha_{ij}m(1-M)}{1+\alpha_{ij}}$, where $m = \min(q_i, q_j)$ and $M = \max(q_i, q_j)$.

Proof. We prove the first inequality. The second follows by Lemma C.0.3 and those for $q_i - \xi_{ij}$ follow by symmetry. The final inequality follows by combining the earlier ones. Let $\xi_{ij} = q_j + y$ and substitute into (6.4),

$$\alpha_{ij}y^2 + y[\alpha_{ij}(q_j - q_i) - 1] + q_j(q_i - 1) = 0.$$

The function is a convex parabola which at $y = 0$ is at $q_j(q_i - 1) \leq 0$.[2] From Lemma 6.4.1 we know that the left root is at $\xi_{ij} \geq q_iq_j$ so we may take the derivative there, i.e. at $q_j + y = q_iq_j \Leftrightarrow y = q_j(q_i - 1)$ and by convexity establish a lower bound for $q_j - \xi_{ij}$. □

We now turn to results on second and third derivatives and thence submodularity.

Theorem 6.5.1. (Second derivatives for each edge) For any edge (i, j), for any α_{ij}, writing $f = f_{ij}$ and $\mu_{ab} = \mu_{ij}(a, b)$ from (6.2),

$$\frac{\partial^2 f}{\partial q_i^2} = \frac{1}{T_{ij}}q_j(1 - q_j)$$
$$\frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} = \frac{1}{T_{ij}}(\mu_{01}\mu_{10} - \mu_{00}\mu_{11})$$
$$\frac{\partial^2 f}{\partial q_j^2} = \frac{1}{T_{ij}}q_i(1 - q_i)$$

where $T_{ij} = q_iq_j(1 - q_i)(1 - q_j) - (\xi_{ij} - q_iq_j)^2 \geq 0$ with equality only for q_i or $q_j \in \{0, 1\}$. Further $\mu_{01}\mu_{10} - \mu_{00}\mu_{11} = q_iq_j - \xi_{ij}$ and has the sign of $-\alpha_{ij}$.

Proof. We begin with the same approach as Korč et al. (2012) but extend the analysis and derive stronger results.

For notational convenience add a third pseudo-dimension restricted to the value 1. Let $\mathbf{y} = (y_1, y_2, y_3)$ be the vector with components $y_1 = x_i$, $y_2 = x_j$ and $y_3 = 1$ where $x_i, x_j \in \mathbb{B}$. Define $\pi(\mathbf{y}) = \mu_{ij}(x_i, x_j)$, and $\phi(\mathbf{y}) = W_{ij}$ if $\mathbf{y} = (1, 1, 1)$ or $\phi(\mathbf{y}) = 0$ otherwise. Let $\mathbf{r} = (q_i, q_j, 1)$. Define function h used in entropy calculations as $h(z) = -z \log z$.

[2] This confirms neatly that we must take the left root else $y > 0 \Rightarrow \mu_{01} < 0$ (a contradiction).

Consider (6.5) but instead of solving for ξ_{ij} explicitly, express f as an optimization problem, minimizing free energy subject to local consistency and normalization constraints in order to use techniques from convex optimization. We have $f(q_i, q_j) = g(\mathbf{r})$ where

$$g(\mathbf{r}) = \min_{\pi} \sum_{\mathbf{y}} \big(-\phi(\mathbf{y})\pi(\mathbf{y}) - h(\pi(\mathbf{y}))\big)$$
$$\text{s.t.} \sum_{\mathbf{y}:y_k=1} \pi(\mathbf{y}) = r_k \;\; k = 1, 2, 3. \tag{C.2}$$

The Lagrangian can be written as

$$L_{\mathbf{r}}(\pi, \boldsymbol{\lambda}) = \sum_{\mathbf{y}} [(-\phi(\mathbf{y}) - \langle \mathbf{y}, \boldsymbol{\lambda}\rangle)\pi(\mathbf{y}) - h(\pi(\mathbf{y}))] + \langle \mathbf{r}, \boldsymbol{\lambda}\rangle$$

and its derivative is

$$\frac{\partial L_{\mathbf{r}}(\pi, \boldsymbol{\lambda})}{\partial \pi} = -\phi(\mathbf{y}) - \langle \mathbf{y}, \boldsymbol{\lambda}\rangle + 1 + \log \pi$$

which yields a minimum at

$$\pi_{\boldsymbol{\lambda}}(\mathbf{y}) = \exp(\phi(\mathbf{y}) + \langle \mathbf{y}, \boldsymbol{\lambda}\rangle - 1). \tag{C.3}$$

Since the minimization problem in (14) is convex and satisfies the weak Slater's condition (the constraints are affine), strong duality applies and $g(\mathbf{r}) = \max_{\boldsymbol{\lambda}} G(\mathbf{r}, \boldsymbol{\lambda}) = G(\mathbf{r}, \boldsymbol{\lambda}^*(\mathbf{r}))$ where the dual is simply

$$G(\mathbf{r}, \boldsymbol{\lambda}) = \min_{\pi} L_{\mathbf{r}}(\pi, \boldsymbol{\lambda}) = -\sum_{\mathbf{y}} \pi_{\boldsymbol{\lambda}}(\mathbf{y}) + \langle \mathbf{r}, \boldsymbol{\lambda}\rangle. \tag{C.4}$$

Let $D_k(\mathbf{r}, \boldsymbol{\lambda}) = \frac{\partial G(\mathbf{r}, \boldsymbol{\lambda})}{\partial \lambda_k}$ then $D_k(\mathbf{r}, \boldsymbol{\lambda}^*) = 0, \; k = 1, 2, 3.$

Hence $\frac{\partial g}{\partial r_k} = \frac{\partial G}{\partial r_k} = \lambda_k$ using (C.4). Focusing on our goal of obtaining second derivatives, we consider $\frac{\partial^2 g}{\partial r_l \partial r_k} = \frac{\partial \lambda_k}{\partial r_l}$ which we shall express in terms of $C_{kl} := \frac{\partial^2 G}{\partial \lambda_l \partial \lambda_k} = \frac{\partial D_k}{\partial \lambda_l}$.

Differentiating $D_k(\mathbf{r}, \boldsymbol{\lambda}^*) = 0$ with respect to r_l,

$$0 = \frac{\partial D_k(\mathbf{r}, \boldsymbol{\lambda}^*)}{\partial r_l} = \frac{\partial D_k}{\partial r_l} + \sum_{p=1}^{3} \frac{\partial D_k}{\partial \lambda_p} \frac{\partial \lambda_p}{\partial r_l} \quad k, l = 1, 2, 3.$$

Considering (C.4), $\frac{\partial D_k}{\partial r_l} = \frac{\partial^2 G}{\partial r_l \partial \lambda_k} = \delta_{kl}$ hence $0 = \delta_{kl} + \sum_p C_{kp} \frac{\partial^2 g}{\partial r_l \partial r_p}$. Thus $\frac{\partial^2 g}{\partial r_l \partial r_k} = -[C^{-1}]_{kl}$. Using its definition and (C.4), we have

$$C_{kl} = \frac{\partial^2 G}{\partial \lambda_l \partial \lambda_k} = \frac{\partial}{\partial \lambda_l} \Big(-\sum_{\mathbf{y}} y_k \pi_{\boldsymbol{\lambda}}(\mathbf{y}) + r_k \Big)$$
$$= -\sum_{\mathbf{y}} y_k y_l \pi_{\boldsymbol{\lambda}}(\mathbf{y}) = -\sum_{\mathbf{y}:y_k=y_l=1} \pi_{\boldsymbol{\lambda}}(\mathbf{y}).$$

Earlier work Korč et al. (2012) stopped here, recognizing that $\det C \leq 0$. We more precisely characterize this matrix

$$C = -\begin{pmatrix} \mu_{10}+\mu_{11} & \mu_{11} & \mu_{10}+\mu_{11} \\ \mu_{11} & \mu_{01}+\mu_{11} & \mu_{01}+\mu_{11} \\ \mu_{10}+\mu_{11} & \mu_{01}+\mu_{11} & 1 \end{pmatrix} \tag{C.5}$$

Recall constraints $\mu_{00}+\mu_{01}+\mu_{10}+\mu_{11} = 1$, $\mu_{01}+\mu_{11} = q_j$, $\mu_{10}+\mu_{11} = q_i$. Note C is symmetric.

Applying the result above and using Cramer's rule,

$$\frac{\partial^2 f}{\partial q_i^2} = \frac{\partial^2 g}{\partial r_1^2} = -\frac{1}{\det C}(\mu_{01}+\mu_{11})(\mu_{00}+\mu_{10}) = \frac{q_j(1-q_j)}{-\det C}$$

$$\frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} = \frac{\partial^2 g}{\partial r_1 \partial r_2} = \frac{(\mu_{01}\mu_{10}-\mu_{00}\mu_{11})}{-\det C}$$

$$\frac{\partial^2 f}{\partial q_j^2} = \frac{\partial^2 g}{\partial r_2^2} = -\frac{1}{\det C}(\mu_{10}+\mu_{11})(\mu_{00}+\mu_{01}) = \frac{q_i(1-q_i)}{-\det C}.$$

Using (C.5) and simplifying, we obtain $-\det C = \mu_{00}\mu_{10}\mu_{11} + \mu_{10}\mu_{11}\mu_{01} + \mu_{11}\mu_{10}\mu_{00} + \mu_{01}\mu_{00}\mu_{10}$. By Lemma C.0.1 this is strictly > 0 unless q_i or $q_j \in \{0,1\}$. Substituting in terms from (6.2) and simplifying establishes $-\det C = T_{ij}$ from the statement of the theorem, and $\mu_{01}\mu_{10} - \mu_{00}\mu_{11} = q_i q_j - \xi_{ij}$. The sign follows from Lemma 6.4.1 or observing from (C.3) that $\frac{\mu_{00}\mu_{11}}{\mu_{01}\mu_{10}} = e^{W_{ij}} = \alpha_{ij} + 1$. □

Lemma C.0.4 (Finite 3rd derivatives)**.** *For any edge (i,j) with $\alpha_{ij} > 0$, if $q_i, q_j \in (0,1)$ then all third derivatives exist and are finite.*

Proof. Using Theorem 6.5.1 noting $T_{ij} > 0$ strictly and considering (6.2), it is sufficient to show $\frac{\partial \xi_{ij}}{\partial q_k}$ is finite. We may assume $k \in \{i,j\}$ else the derivative is 0 and by symmetry need only check $\frac{\partial \xi_{ij}}{\partial q_i}$. Differentiating (6.4),

$$\frac{\partial \xi_{ij}}{\partial q_i} = \frac{\alpha_{ij}(q_j - \xi_{ij}) + q_j}{1 + \alpha_{ij}(q_i - \xi_{ij} + q_j - \xi_{ij})},$$

clearly finite for $\alpha_{ij} > 0$ since recalling (6.2), $q_i - \xi_{ij}$ and $q_j - \xi_{ij}$ are elements of the pseudo-marginal and hence are non-negative (or use Lemma 6.4.3). □

Theorem 6.5.2. If a binary pairwise MRF is submodular on an edge (i,j), i.e. $\alpha_{ij} > 0$, then the multi-label discretized MRF for any discretization $\mathcal{M}$ is submodular for that edge. In particular, if

the MRF is fully associative/submodular, i.e. $\alpha_{ij} > 0 \; \forall (i,j) \in \mathcal{E}$, then the multi-label discretized MRF is fully submodular for any discretization.

Proof. For any edge (i,j), let f be the pairwise function f_{ij} from (6.5) and note the submodularity requirement (6.6). Let $x = (x_1, x_2)$, $y = (y_1, y_2)$ be any points in $[0,1]^2$. Define $s(x,y) = (s_1, s_2) = (\min(x_1, y_1), \min(x_2, y_2))$, and $t(x,y) = (t_1, t_2) = (\max(x_1, y_1), \max(x_2, y_2))$. Let $g(x,y) = f(s_1, s_2) + f(t_1, t_2) - f(s_1, t_2) - f(s_2, t_1)$, call this the submodularity of the rectangle defined by x, y. We must show $g(x,y) \leq 0$. Note f is continuous in $[0,1]^2$ hence so also is g. We shall show that $\forall (x,y) \in (0,1)^2$, $g(x,y) < 0$ then the result follows by continuity.

Assume $x, y \in (0,1)^2$. Consider derivatives of f in the compact set $R = [s_1, t_1] \times [s_2, t_2]$. Using (6.9) and Lemma C.0.1, first derivatives exist and are bounded. By Theorem 6.5.1 and Lemma C.0.4 the same holds for second and third derivatives. Further, Theorem 6.5.1 and Lemma C.1.1 show that $\frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} < 0$.

If a rectangle is sliced fully along each dimension so as to be subdivided into sub-rectangles then summing the submodularities of all the sub-rectangles, internal terms cancel and we obtain the submodularity of the original rectangle.

Hence there exists an ϵ such that if we subdivide the rectangle defined by x, y into sufficiently small sub-rectangles with sides $< \epsilon$ and apply Taylor's theorem up to second order with the remainder expressed in terms of the third derivative evaluated in the interval, then the second order terms dominate and the submodularity of each small sub-rectangle < 0. Summing over all sub-rectangles provides the result. □

C.1 Bethe Bound Propagation (BBP)

In order to derive our approach of Bethe bound propagation (BBP), we extend the analysis of bounds on ξ_{ij} from Lemmas 6.4.1 and 6.4.3.

Lemma C.1.1 (Better lower bound for ξ_{ij})**.** *If $\alpha_{ij} > 0$, then $\xi_{ij} \geq q_i q_j + \alpha_{ij} q_i q_j (1 - q_i)(1 - q_j) / [1 + \alpha_{ij}(q_i + q_j - 2 q_i q_j)]$, equality only possible at an edge, i.e. one or both of $q_i, q_j \in \{0, 1\}$.*

Proof. Write $\xi_{ij} = q_i q_j + y$ and substitute into (6.4),

$$\alpha_{ij} y^2 - y[1 + \alpha_{ij}(q_i + q_j - 2 q_i q_j)] + \alpha_{ij} q_i q_j (1 - q_i)(1 - q_j) = 0.$$

We have a convex parabola which at $y = 0$ is above the abscissa (unless q_i or $q_j \in \{0, 1\}$) and has negative gradient by Lemma C.0.3. Hence all roots are at $y \geq 0$ and given convexity we can bound below using the tangent at $y = 0$ which yields the result. □

Bounds were already derived on stationary points in Theorems C.0.2 and 6.4.2. Here we show for variables with only associative edges how we can iteratively improve these bounds, sometimes with striking results. Note that a fully associative model is not required, and as in Section 6.3.2, any model may be selectively flipped to yield local associativity around a particular node.

We first assume all $\alpha_{ij} \geq 0$ and adopt the approach of Theorem C.0.2, now using the better bound from Lemma C.1.1 to obtain

$$q_i - \xi_{ij} \leq q_i - q_i q_j - \frac{\alpha_{ij} q_i q_j (1-q_i)(1-q_j)}{1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)} = q_i(1-q_j)\Big[1 - \frac{\alpha_{ij} q_j (1-q_i)}{1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)}\Big],$$

$$1 + \xi_{ij} - q_i - q_j \geq 1 + q_i q_j - q_i - q_j + \frac{\alpha_{ij} q_i q_j (1-q_i)(1-q_j)}{1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)} = (1-q_i)(1-q_j)\Big[1 + \frac{\alpha_{ij} q_i q_j}{1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)}\Big].$$

Hence $Q_i \leq \frac{q_i}{1-q_i} \prod_{j \in \mathcal{N}(i)} R_{ij}^{-1}$ where

$$R_{ij} = \frac{1 + \frac{\alpha_{ij} q_i q_j}{1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)}}{1 - \frac{\alpha_{ij} q_j (1-q_i)}{1 + \alpha_{ij}(q_i + q_j - 2q_i q_j)}} = 1 + \frac{\alpha_{ij} q_j}{1 + \alpha_{ij} q_i (1 - q_j)},$$

monotonically increasing with q_j and decreasing with q_i. Hence

$$e^{W_{ij}} = 1 + \alpha_{ij} \geq R_{ij} \geq L_{ij} := 1 + \frac{\alpha_{ij} A_j}{1 + \alpha_{ij}(1 - B_i)(1 - A_j)} \tag{C.6}$$

Using Theorem C.0.2, we initialize $A_i = \sigma(\theta_i)$ and $B_i = 1 - \sigma(\theta_i + W_i)$.

Using (6.9), at any stationary point we must have

$$q_i \geq 1/[1 + \exp(-\theta_i)/L_i]$$

where $L_i = \prod_{j \in \mathcal{N}(i)} L_{ij}$. Intuitively, in an associative model, if variable i has neighbors j which are likely to be 1 (i.e. high A_j) then this pulls up the probability that i will be 1 (i.e. raises A_i).

Flipping all variables,

$$1 - q_i \geq 1/[1 + \exp(\theta_i + W_i)/U_i]$$

where $U_i = \prod_{j \in \mathcal{N}(i)} U_{ij}$ with

$$e^{-W_{ij}} \geq U_{ij} := 1 + \frac{\alpha_{ij} B_j}{1 + \alpha_{ij}(1 - A_i)(1 - B_j)}.$$

It is also possible to write this as

$$\sigma(\theta_i + \log L_i) \leq q_i \leq \sigma(\theta_i + W_i - \log U_i).$$

This establishes a message passing type of algorithm for iteratively improving the bounds $\{A_i, B_i\}$. Repeat until convergence:

$$\text{new } A_i \leftarrow (1 + \exp(-\theta_i)/L_i)^{-1}$$

$$\text{new } B_i \leftarrow (1 + \exp(\theta_i + W_i)/U_i)^{-1}$$

$$\text{recompute } L_i, U_i \text{ using new } A_i, B_i.$$

Lemma C.1.2. *At every iteration, all of A_i, B_i, L_{ij}, U_{ij} monotonically increase.*

Proof. All of the dependencies are monotonically increasing on all inputs. The first iteration yields an increase since each $L_{ij}, U_{ij} > 1$. □

Since $A_i + B_i \leq 1$, each is bounded above and we achieve monotonic convergence. Combining this with the main global optimization approach can dramatically reduce the range of values that need be considered, leading to significant time savings. Convergence is rapid even for large, densely connected graphs. Each iteration takes $O(|\mathcal{E}|)$ time; a good heuristic is to run for up to 50 iterations, terminating early if all parameters improve by less than a threshold value. This adds negligible time to the global optimization.

This procedure alone can produce impressive results. For example, running on a 100-node graph with independent random edge probability 0.04 (hence average degree 4), each W_{ij} and θ_i drawn randomly from Uniform $[0, 1]$ and then adjusting $\theta_i \leftarrow \theta_i - \sum_{j \in \mathcal{N}(i)} W_{ij}/2$ in order to be unbiased, convergence takes about 11 iterations yielding final average bracket width of 0.05 after starting with average bracket width of 0.40. Greater connectivity, higher edge strengths and smaller individual node potentials make the problem more challenging and may widen the returned final brackets significantly.

C.1.1 BBP for general models

A repulsive edge (i, j) may always be flipped to associative by flipping variable j, which flips its Bethe bounds $A_j \leftrightarrow B_j$. Using Theorem 6.4.2 we can extend the analysis above to run BBP on

any model, see Algorithm 2. Performance in terms of convergence speed and final bracket width is similar for associative and non-associative models.

C.2 Extending *curvMesh* to a General Model

Here we extend the analysis of Section 6.7 by considering repulsive edges to show that for a general binary pairwise model, we can still compute useful bounds (which turn out to be very similar to the earlier bounds for attractive models) for a sufficient mesh width.

Our main tool for dealing with a repulsive edge is to flip the variable at one end (see Section 6.3) to yield an attractive edge, then we can apply earlier results. We denote the flipped model parameters with a $'$. For example, if just variable X_j is flipped, then $q'_j = q(X'_j = 1) = q(1 - X_j = 1) = 1 - q_j$. Since $\alpha_{ij} = e^{W_{ij}} - 1$ and here $W'_{ij} = -W_{ij}$, the following relationship holds if one end of an edge is flipped,

$$\frac{\alpha'_{ij}}{1 + \alpha'_{ij}} = \frac{e^{-W_{ij}} - 1}{e^{-W_{ij}}} = 1 - e^{W_{ij}} = -\alpha_{ij}. \tag{C.7}$$

Note that, for an attractive edge, $\frac{\alpha'_{ij}}{1+\alpha'_{ij}} \in (0, 1)$, as is $-\alpha_{ij}$ for a repulsive edge. Recall that when we flip some set of variables, by construction $\mathcal{F}' = \mathcal{F} + constant$ (see Section 6.3).

The Hessian terms from Theorem 6.5.3 still apply. Our goal is to bound the magnitude of each entry H_{ij} for a general binary pairwise model, then the earlier analysis will provide the result. Whereas for a fully attractive model, we assumed a maximum edge weight W with $0 \leq W_{ij} \leq W$, now we assume $|W_{ij}| \leq W$.

C.2.1 Edge terms

First consider H_{ij} for an edge $(i, j) \in \mathcal{E}$. If the edge is attractive, then the earlier analysis holds (it makes no difference if other edges are attractive or repulsive). If it is repulsive, then $H_{ij} > 0$. Consider a model where just X_j is flipped. $H_{ij} = \frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j} = -\frac{\partial^2 \mathcal{F}'}{\partial q'_i \partial q'_j} = -H'_{ij}$. Hence using (6.19) and (C.7), in practice an upper bound may be computed from Lemma 6.7.1 using $k = -\alpha_{ij}$ and $A'_j = B_j, B'_j = A_j$. The theoretical bound for an attractive edge from (6.20) becomes $H_{ij} \leq \frac{-\alpha_{ij}}{\bar{\eta}(1-\alpha_{ij}^2)}$. As we should expect from the attractive case, the following result holds.

Lemma C.2.1. *For a repulsive edge,* $\frac{1}{1-\alpha_{ij}^2} = O(e^{-W_{ij}})$.

Algorithm 2 BBP for a general binary pairwise model

{Initialize}

for all $i \in \mathcal{V}$ **do**

$W_i = \sum_{j \in \mathcal{N}(i): W_{ij} > 0} W_{ij}$,

$V_i = -\sum_{j \in \mathcal{N}(i): W_{ij} < 0} W_{ij}$,

$A_i = \sigma(\theta_i - V_i)$, $B_i = 1 - \sigma(\theta_i + W_i)$

end for

for all $(i, j) \in \mathcal{E}$ **do**

$\alpha_{ij} = \exp(|W_{ij}|) - 1$

end for

{Main loop}

repeat

for all $i \in \mathcal{V}$ **do**

$L_i = 1, U_i = 1$ {Initialize for this pass}

for all $j \in \mathcal{N}(i)$ **do**

if $W_{ij} > 0$ **then**

{Associative edge}

$L_i* = 1 + \frac{\alpha_{ij} A_j}{1 + \alpha_{ij}(1 - B_i)(1 - A_j)}$

$U_i* = 1 + \frac{\alpha_{ij} B_j}{1 + \alpha_{ij}(1 - A_i)(1 - B_j)}$

else

{Repulsive edge}

$L_i* = 1 + \frac{\alpha_{ij} B_j}{1 + \alpha_{ij}(1 - B_i)(1 - B_j)}$

$U_i* = 1 + \frac{\alpha_{ij} A_j}{1 + \alpha_{ij}(1 - A_i)(1 - A_j)}$

end if

end for

$A_i = 1/(1 + \exp(-\theta_i + V_i)/L_i)$

$B_i = 1/(1 + \exp(\theta_i + W_i)/U_i)$

end for

until All A_i,B_i changed by $<$ THRESH **or** run MAXITER times

Proof. Let $u = -W_{ij}$, then $\alpha_{ij} = e^{-u} - 1$ and $\frac{1}{1-\alpha_{ij}^2} = \frac{1}{(1-\alpha_{ij})(1+\alpha_{ij})} = \frac{1}{e^{-u}(2-e^{-u})} = O(e^u)$. □

Hence, noting that we may flip any neighbors j of i which are adjacent via repulsive edges to obtain $\frac{1}{\eta_i(1-\eta_i)} = O(e^{T+\Delta W/2})$ as before, where now $W = \max_{(i,j)\in\mathcal{E}} |W_{ij}|$, we see that for our new second derivative method, just as in the fully attractive case, $a = O(e^{W(1+\Delta/2)+T})$.

We provide a further interesting result, deriving a lower bound for ξ_{ij} for a repulsive edge.

Lemma C.2.2 (Lower bound for ξ_{ij} for a repulsive edge, analogue of Lemma 6.4.3)**.** *For any repulsive edge* (i, j), $q_iq_j - \xi_{ij} \leq -\alpha_{ij}p_{ij}$ *where* $p_{ij} = \min\{q_iq_j, (1-q_i)(1-q_j)\}$.

Proof. Consider a model where just variable X_j is flipped, and let all new quantities be designated by the symbol $'$. Consider the joint pseudo-marginal (6.2). In the new model the columns are switched since $\mu'_{ij}(a, b) = q(X'_i = a, X'_j = b) = q(X_i = a, X_j = 1 - b) = \mu_{ij}(a, 1 - b)$, hence

$$\mu'_{ij} = \begin{pmatrix} 1 + \xi'_{ij} - q'_i - q'_j & q'_j - \xi'_{ij} \\ q'_i - \xi'_{ij} & \xi'_{ij} \end{pmatrix} = \begin{pmatrix} q_j - \xi_{ij} & 1 + \xi_{ij} - q_i - q_j \\ \xi_{ij} & q_i - \xi_{ij} \end{pmatrix}. \tag{C.8}$$

Applying Lemma 6.4.3 to the new model, $\xi'_{ij} - q'_iq'_j \leq \frac{\alpha'_{ij}}{1+\alpha'_{ij}} m'(1 - M')$. Substituting in $\xi'_{ij} = q_i - \xi_{ij}$ from (C.8) and using (C.7), we have $(q_i - \xi_{ij}) - q_i(1 - q_j) \leq -\alpha_{ij}m'(1 - M')$. Since $m' = \min\{q_i, 1 - q_j\}$ and $M' = \max\{q_i, 1 - q_j\}$, noting $q_i \leq 1 - q_j \Leftrightarrow q_i + q_j \leq 1 \Leftrightarrow q_iq_j \leq (1 - q_i)(1 - q_j)$, the result follows. □

Hence for a repulsive edge (i, j), using (6.10), we have

$$T_{ij} = q_iq_j(1 - q_i)(1 - q_j) - (\xi_{ij} - q_iq_j)^2 \geq p_{ij}P_{ij} - \alpha_{ij}^2 p_{ij}^2,$$

where $P_{ij} = \max\{q_iq_j, (1 - q_i)(1 - q_j)\}$.

C.2.2 Diagonal terms

Consider the H_{ii} terms from Theorem 6.5.3, which is true for a general model. If all neighbors of X_i are adjacent via attractive edges, then, as in Section 6.7.2, $H_{ii} \leq \frac{1}{\eta_i(1-\eta_i)} \left(1 - d_i + \sum_{j\in\mathcal{N}(i)} \frac{1}{1-\left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2}\right)$.

If any neighbors are connected to X_i by a repulsive edge, then consider a new model where those neighbors are flipped, so now all edges incident to X_i are attractive, and designate the new model

parameters with a $'$. As before, observe $\mathcal{F} = \mathcal{F}' + constant$, hence $H_{ii} = \frac{\partial^2 \mathcal{F}}{\partial q_i^2} = \frac{\partial^2 \mathcal{F}'}{\partial q_i'^2} = H'_{ii}$. Using (C.7) we obtain that for a general model,

$$H_{ii} \leq \frac{1}{\eta_i(1-\eta_i)} \left(1 - d_i + \sum_{j \in \mathcal{N}(i): W_{ij} > 0} \frac{1}{1 - \left(\frac{\alpha_{ij}}{1+\alpha_{ij}} \right)^2} + \sum_{j \in \mathcal{N}(i): W_{ij} < 0} \frac{1}{1-\alpha_{ij}^2} \right). \tag{C.9}$$

Similarly to the analysis in Section 6.7.3.1, using Lemma C.2.1 gives that for a general model, $b = \max_{i \in \mathcal{V}} H_{ii} = O(\Delta e^{W(1+\Delta/2)+T})$, just as for a fully attractive model, where now $W = \max |W_{ij}|$.

C.3 Power Network

The simulated sub-network used in the experiment is shown in Figure C.1.

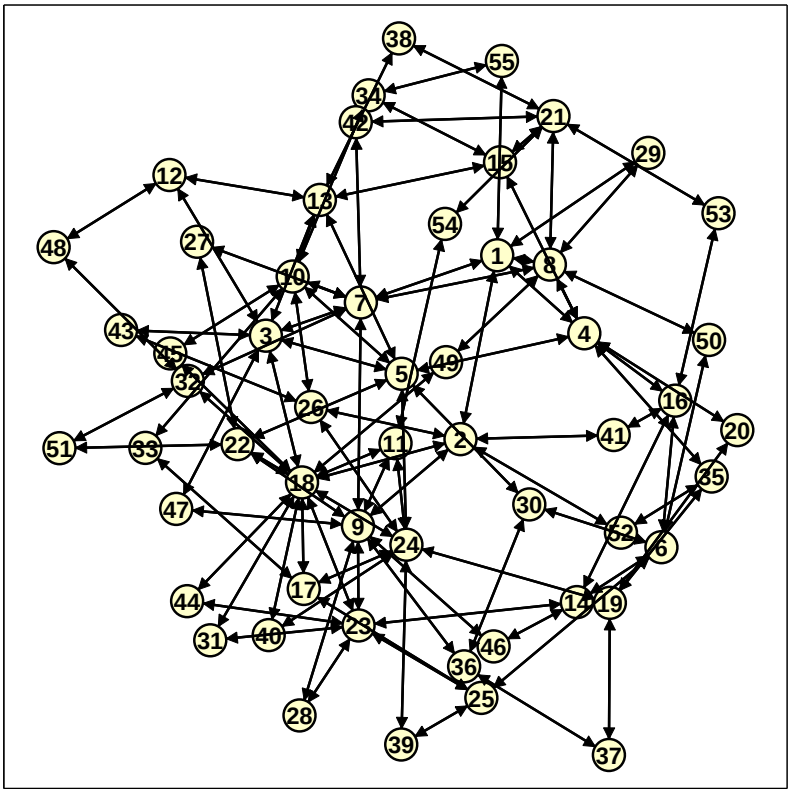


Figure C.1: Sub-network used for the power experiment described in the main text, Section 6.9.2.

C.4 Replacing W with GAP for Attractive Models

In this Section, we show how for an attractive model, the theoretical mesh size bound $N = O(\frac{nmW}{\epsilon})$ may be improved to $N = O\left(\frac{nm}{\epsilon} \min\left[W, \frac{n}{G}\right]\right)$, where G is the GAP for the model, which we define as the difference in energy between the lowest and next lowest energy states (i.e. it is the difference in score between the MAP and the 2nd best configuration, where the score of a state is defined as minus the energy), *normalized to some level of potential strengths*. For this, we scale all potentials s.t. $\max_{(i,j)\in\mathcal{E}} W_{ij} = 1$. Clearly GAP ≥ 0 and GAP$= 0$ is possible, for example in a symmetric model. As far as we are aware, the only work relating the size of the expected GAP to the distribution of weights in a model is in (Aldous, 2001; Wästlund, 2009), which addresses bipartite matching.

The idea is as follows: In the mesh approach, we seek a point with maximum score plus Bethe entropy. The Bethe entropy is $S_B = \sum_{i\in\mathcal{V}} S_i - \sum_{(i,j)\in\mathcal{E}} I_{ij}$, i.e. the sum of the singleton entropies minus the sum of the edge mutual informations, where $I_{ij} \geq 0$. Hence, $S_B \leq S^* = n \log 2$, which is independent of the potential functions. Indeed, a better upper bound may be available but this is sufficient for our needs.[3] We shall show that the maximum score obtainable at a particular value of q_i is upper bounded by the linear interpolation of the maximum values at $q_i = 0$ and $q_i = 1$. The difference in these two score values scales proportionately with the overall potential strengths (or equivalently, it scales inversely with the temperature). For sufficiently high potential strengths, we will not need to check more than a certain distance away from whichever end (0 or 1) has higher score. This distance declines inversely to the potential strengths, and this gives the result.

We begin with results on how the score varies if it is maximized over $n - 1$ variables while the other one variable is varied.

Notation. Define the *score* of a point in marginal distribution space given by $q = \{q_i : i \in \mathcal{V}; q_{ij} : (i,j) \in \mathcal{E}\}$ to be $f(q) = \sum_{i\in\mathcal{V}} \theta_i q_i + \sum_{(i,j)\in\mathcal{E}} W_{ij} q_{ij}$, i.e. the negative of the energy. For an attractive model, we have $W_{ij} > 0\ \forall (i,j) \in \mathcal{E}$.

Let $M_i(x) = \max_{q\in\mathcal{L}:q_i=x} f(q)$, i.e. the maximum score over the local polytope subject to $q_i = x$.

[3] Indeed, when edge weights are strongly positive, this can lower dramatically the maximum entropy at the optimum of the Bethe free energy, as described in Section 7.5. Hence, it may be possible to improve our result.

Lemma C.4.1. *For any $x \in [0,1]$, $M_i(x)$ is achieved by a configuration with $q_j \in \{0, x, 1\}\ \forall j \in \mathcal{V}$.*

Proof. The problem is an LP with variables $\{q_j : j \in \mathcal{V} \setminus i\}$ and $\{q_{jk} : (j,k) \in \mathcal{E}\}$. The constraints of the local polytope may be written:

(i) $q_j \geq 0\ \forall j \in \mathcal{V} \setminus i$

(ii) $q_j \leq 1\ \forall j \in \mathcal{V} \setminus i$

(iii) $q_{jk} \geq 0\ \forall (j,k) \in \mathcal{E}$

(iv) $q_{jk} \leq q_j\ \forall (j,k) \in \mathcal{E}$

(v) $q_{jk} \leq q_k\ \forall (j,k) \in \mathcal{E}$

(vi) $1 + q_{jk} - q_j - q_k \geq 0\ \forall (j,k) \in \mathcal{E}$

Observe that if all $W_{jk} > 0$ then $q_{jk} = \min(q_j, q_k)$ and (vi) will never be binding. Hence the same solution will be obtained by considering the modified LP' without it. The optimum of LP' is achieved at a vertex. We have $q_i = x$ and it is easily seen that a vertex must satisfy $q_j \in \{0, x, 1\}\ \forall j \in \mathcal{V}$. □

Lemma C.4.2. *$M_i(y)$ is a convex function of y.*

Proof. For any $x \in [0,1]$, take an $\arg\max$ of $M_i(x)$ as in Lemma C.4.1. Divide the variables $j \neq i$ into 3 sets: $A_x = \{j \in \mathcal{V} \setminus i : q_j = 0\}, B_x = \{j \in \mathcal{V} \setminus i : q_j = x\}$ and $C_x = \{j \in \mathcal{V} \setminus i : q_j = 1\}$. Define the function $f_x : [0,1] \to \mathbb{R}$ given by $f_x(y) = f(q(y;x))$ where $q(y;x)$ is defined by:

$$\forall j \in \mathcal{V}, q_j(y;x) = \begin{cases} 0 & j \in A_x \\ y & j \in B_x \cup \{i\}\ ; \\ 1 & j \in C_x \end{cases} \quad \forall (j,k) \in \mathcal{E}, q_{jk}(y;x) = \min(q_j(y;x), q_k(y;x)).$$

Note that $f_x(y)$ is linear. Observe that $M_i(y) = \sup_{x \in [0,1]} f_x(y)$, i.e. is the pointwise $\sup$ of a set of linear functions, hence is convex. □

With these results, we turn to mesh construction and consider the points required in dimension i. Since the model is attractive, $M_i(0)$ and $M_i(1)$ may be computed efficiently (and will have $\arg\max$ at an integer vertex). Let the higher be achieved at $q_i = H_i \in \{0,1\}$, and the lower at $q_i = L_i = 1 - H_i$, with difference $D_i = M_i(H_i) - M_i(L_i)$.

In the mesh approach, we are seeking a point with maximum score plus Bethe entropy. By Lemma C.4.2, as we move in dimension i from H_i to L_i, the maximum possible score declines at least as quickly as the linear interpolation. Hence there is no need to check more than t_i^* from H_i where $t_i^* D_i = S^*$. If all potentials are scaled by a factor k, then so too will all the D_i terms. Hence $D_i = G_i \max W_{ij}$ where G_i is the gap in dimension i when the model is scaled s.t. $\max W_{ij} = 1$. Note that GAP as defined above, i.e. the difference in score between the MAP and second best configuration when the model is normalized, satisfies that the GAP, $G = \min_{i \in \mathcal{V}} G_i$.

Combining this with the earlier result means that the total number of points $N = O\left(\frac{nmW}{\epsilon} \frac{S^*}{GW}\right) = O\left(\frac{n^2 m}{G\epsilon}\right)$.

C.4.1 Holding singleton potentials fixed while edge potentials scale

In the analysis above, we assumed that *all* potentials scale together, in order to conclude that the differences $D_i = M_i(H_i) - M_i(L_i)$ would scale with them. An interesting question is to ask how the D_i vary if the distribution of singleton potentials is held fixed, while only the edge potentials scale. We studied this empirically by considering an attractive model with 10 variables, all pairwise adjacent. We generated 200 models, each with $\theta_i \sim [-T_{max}, T_{max}]$ and $W_{ij} \sim [0, W]$, then observed the empirical difference in scores between the MAP configuration and the second best configuration as T_{max} and W were varied. See Figure C.2 for results.

Observe that the average difference in scores between MAP and 2nd best seems to asymptote to a constant value (approximately $\frac{3}{2} T_{max}$). Hence, it appears that if just edge weights are varied, while singleton potentials are kept fixed, then the benefit described above will not be seen.

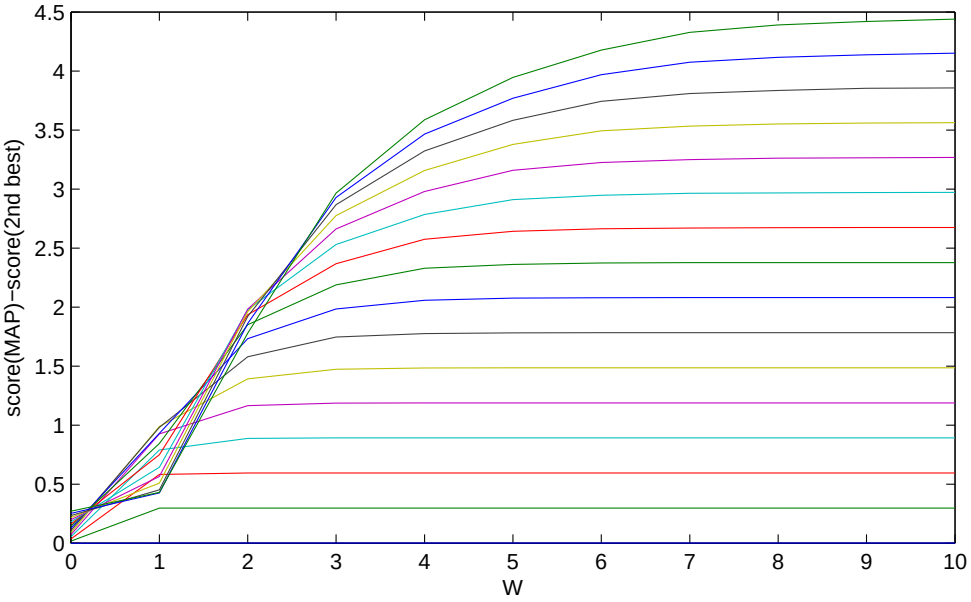


Figure C.2: Difference in score between MAP and 2nd best for various values of T_{max}. Average over 200 runs, complete graph on 10 variables, $\theta_i \sim [-T_{max}, T_{max}]$ and $W_{ij} \sim [0, W]$. Each curve is for a different value of T_{max} as it was varied from 0 to 3 in steps of 0.2.

Appendix D

Appendix for Understanding the Bethe Approximation

Here we provide further details and derivations of several of the results in Chapter 7.

7.3 Homogeneous Cycles

Tree-Reweighted Approximation

The tree-reweighted approximation (TRW) of Wainwright et al. (2005) provides a family of upper bounds on the true entropy and partition function, based on selecting a convex combination of spanning trees of the MRF graph.

Lemma D.0.3. *In the homogeneous case for n connected variables with topology $G(\mathcal{V}, \mathcal{E})$ (e.g. C_n or K_n) with edge weights W and no singleton potentials, the minimum TRW partition function Z_T*

is achieved with uniform edge appearance probability r and marginals satisfying

$$\mu_{ij} = \begin{pmatrix} x_T & \frac{1}{2} - x_T \\ \frac{1}{2} - x_T & x_T \end{pmatrix} = \mu_T \; \forall (i,j) \in \mathcal{E}$$

$$\begin{aligned} \log Z_T &= -E_T + S_T \\ &= mWx_T \quad + (n-1)S(\mu_T) + (2-n)\log 2 \\ \textit{where } x_T &= \frac{e^{W/2r}}{2(1+e^{W/2r})} = \frac{1}{2}\sigma(W/2r), \\ r &= \frac{n-1}{m}, m = |\mathcal{E}| \end{aligned}$$

In particular, if $G = C_n$ then $r = \frac{n-1}{n}$, or if $G = K_n$ then $r = \frac{2}{n}$.
Further, if the TRW optimization is performed over the cycle polytope, then the same result applies except (similar to the Bethe case) $x_{TC} = \max(x_T, 1/2g)$, where g is the size of the smallest odd chordless cycle in G (if none exists then $x_{TC} = x_T$).

Proof. Let $\mathcal{L}$ be the local polytope and R the spanning tree polytope. For the optimal TRW bound, we seek

$$\log Z_T = \min_{\rho \in R} \max_{\mu \in \mathcal{L}} \left(-E(\mu) + \sum_{t \in S} \rho_t S(\mu_t) \right) \tag{D.1}$$

where here

$$-E(\mu) = \frac{W}{2} \sum_{(i,j) \in \mathcal{E}} \mu_{ij}(0,0) + \mu_{ij}(1,1)$$

and μ_t is the projection of the global μ distribution onto the spanning tree $t \in S$, hence, as is standard,

$$S(\mu_t) = \sum_{i \in \mathcal{V}} S(\mu_i) + \sum_{(i,j) \in \mathcal{E}(t)} S(\mu_{ij}) - S(\mu_i) - S(\mu_j).$$

Considering (D.1), the outer optimization is minimizing with respect to ρ a pointwise max of a linear function of ρ, hence is minimizing a convex function of ρ. Given the symmetry of the problem, this implies that the best TRW bound is achieved when each edge has equal weight $r = \frac{n-1}{m}$, and

$$\log Z_T = \max_{\mu \in \mathcal{L}} \sum_{(i,j) \in \mathcal{E}} \left[\frac{W}{2} \left(\mu_{ij}(0,0) + \mu_{ij}(1,1) \right) + r S(\mu_{ij}) \right] + \sum_{i \in \mathcal{V}} S(\mu_i)(1 - r \cdot \text{degree}(i)).$$

Observe that if $r = 1$ (a tree) then this is exactly the Bethe optimization problem.

It is easy to check that $\mu_i = \frac{1}{2}\ \forall i$ is a stationary point. The remaining results follow from Lemma 7.3.1 and differentiating what must be maximized with respect to x_T to obtain a maximum at $x_T = \frac{e^{W/2r}}{2(1+e^{W/2r})}$, cf Lemma 7.3.2.

□

7.5 General Homogeneous Graphs

7.5 Threshold result for attractive models

Lemma 7.5.1. Consider a symmetric homogeneous MRF on n vertices with $d-$regular topology and edge weights W. $q = (\frac{1}{2}, \ldots, \frac{1}{2})$ is a stationary point of the Bethe free energy but for W above a critical value, this is not a minimum. Specifically, let H be the Hessian of the Bethe free energy at q, x_B be the value from Lemma 7.3.2 and $\mathbf{1}$ be the vector of length n with 1 in each dimension; then $\mathbf{1}^T H \mathbf{1} = n[d - 4x_B(d-1)]/x_B < 0$ if $x_B > \frac{1}{4}\frac{d}{d-1} \Leftrightarrow W > 2\log\frac{d}{d-2}$.

Proof. We use Lemma 7.3.2 and the following expressions for the Hessian $H_{jk} = \frac{\partial^2 \mathcal{F}}{\partial q_j \partial q_k}$ from (Weller and Jebara, 2013a):

$$H_{jk} = \begin{cases} \frac{q_j q_k - \xi_{jk}}{T_{jk}} & \text{if } (j,k) \in \mathcal{E} \\ 0 & \text{if } (j,k) \notin \mathcal{E} \end{cases}, \quad H_{jj} = -\frac{d_j - 1}{q_j(1-q_j)} + \sum_{k \in \mathcal{N}(j)} \frac{q_k(1-q_k)}{T_{jk}},$$

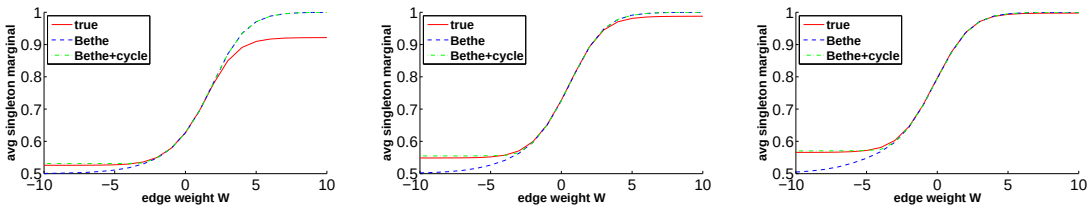
where $d_j = |\mathcal{N}(j)|$ is the degree of j, and $T_{jk} = q_j q_k (1-q_j)(1-q_k) - (\xi_{jk} - q_j q_k)^2$. Taking these together with (7.4), and using symmetry, we have $x_B = \frac{1}{2}\sigma(W/2), T_{jk} = T = x_B(\frac{1}{2} - x_B)$ and

$$\begin{aligned} \mathbf{1}^T H \mathbf{1} &= n\left[-4(d-1) + \frac{d}{4T} + \frac{d}{T}\left(\frac{1}{4} - x_B\right)\right] \\ &= n\left[d - 4x_B(d-1)\right]/x_B. \end{aligned}$$

□

7.5.1 Further results on entropy and polytope

We have shown that in an attractive model, the Bethe entropy approximation can lead to singleton marginals being pulled toward the extreme values of 0 or 1. When repulsive edges are present and we have a frustrated cycle, there is also an effect that can go the other way, pushing singleton



(a) $T_{max} = 1$ (b) $T_{max} = 2$ (c) $T_{max} = 3$

Figure D.1: Average singleton marginal vs. uniform edge weight W for true, Bethe, Bethe+cycle. C_5 topology with $\theta_i \sim [0, T_{max}]$, all edge weights set to W. Bethe and Bethe+cycle overlap for positive W. Average shown over 20 runs for each set of parameters.

marginals toward $\frac{1}{2}$. This effect is due to the polytope relaxation. One way to see this is to observe that the minimum energy configuration on the local polytope for a symmetric frustrated cycle has all singleton marginals of $\frac{1}{2}$, whereas on the marginal polytope it is integral (Wainwright and Jordan, 2008, §8.4.1).

To examine these effects, we ran experiments on a model with 5 nodes arranged in a cycle. Each $\theta_i \sim [0, T_{max}]$ and all edge weights were set to uniform W. T_{max} and W were varied to observe their effect. Singleton marginals were computed using Bethe (on local), Bethe+cycle (which in this context is the same as Bethe+marginal) and with the true distribution. See Figure D.1 for results.

Observe that for strongly positive W, the Bethe entropy approximation pulls the marginals toward 1. This behavior is the same for Bethe and Bethe+cycle, demonstrating that it is an effect due to the entropy approximation. Note we are observing this effect on a model which clearly has just one cycle. As singleton potential strengths are raised, the relative effect diminishes. On the other hand, for strongly negative W (which causes a highly frustrated cycle since the cycle is odd), the curve for Bethe is pulled toward 0.5, but the Bethe+cycle curve is not, indicating that this is a polytope effect.

7.6 Experiments

7.6.1 Implementation and validation

7.6.1.1 Optimizing over the cycle polytope

We provide details of our dual decomposition algorithm to optimize over the cycle polytope, see Algorithm 3. This relies on the ϵ-approximation mesh method of Weller and Jebara (2013a), as improved in (Weller and Jebara, 2014a) to handle general (non-attractive) binary pairwise models. Even if the initial model is attractive, as the dual variables update, the modified potential parameters may become repulsive. Note that a lower bound on the Bethe free energy $\mathcal{F}$ is equivalent to a lower bound on $-\log Z_B$ or an upper bound on $\log Z_B$, the Bethe $\log$ partition function, see §7.2.2.

Our goal is to minimize $\mathcal{F}$ subject to the cycle constraints (7.6) to yield what we define as $-\log Z_{BC}$. Introduce Lagrangian multipliers $\lambda = \{\lambda_{C,F}\}$ for each such constraint on C and F, and consider

$$\mathcal{L}(\mu,\lambda) = \mathbb{E}_\mu(E) - S_B(\mu) + \sum_{C,F} \lambda_{C,F} \left(1 - \sum_{(i,j)\in F} (\mu_{ij}(0,0) + \mu_{ij}(1,1)) - \sum_{(i,j)\in C\setminus F} (\mu_{ij}(1,0) + \mu_{ij}(0,1)) \right) \tag{D.2}$$

$$= \mathcal{F}(\mu) + \lambda^\perp g(\mu) \quad \text{defining } g \text{ appropriately from the line above.}$$

Let $\mathcal{G}$ be the dual function, i.e. $\mathcal{G}(\lambda) := \inf_\mu \mathcal{L}(\mu,\lambda)$. For any $\lambda \succcurlyeq 0$, this is a lower bound for $\mathcal{F}(\mu^*)$ where μ^* is the optimum feasible (i.e. in the cycle polytope) primal point. For any feasible μ, $\mathcal{F}(\mu)$ provides an upper bound on $\mathcal{F}(\mu^*)$.

We shall identify $\sup_\lambda \mathcal{G}(\lambda) = \sup_\lambda \inf_\mu \mathcal{L}(\mu,\lambda)$ subject to $\lambda \succcurlyeq 0$, which is be the best lower bound we can obtain. We do this as follows: given λ, absorb the constraint terms from (D.2) into the energy, reparameterize appropriately and minimize using the approach of Weller and Jebara (2014a). Then update λ using projected sub-gradient descent with g and repeat to convergence.

Note that for a complete graph K_n, the set of all chordless cycles is the set of all $\binom{n}{3}$ triplets. This provides a polynomial upper bound on the number of chordless cycles for a graph on n vertices, since for any graph that is not complete, adding a missing edge can only increase the number.

Following the methods of (Boyd and Mutapcic, 2007, §3.2) with a typical step size schedule, it is easy to see that we converge in the dual, and that as a consequence of the ϵ-approximate inner

Algorithm 3 Dual decomposition algorithm to compute lower bound for $-\log Z_B$ on the cycle polytope

{Initialize. Take inputs $\epsilon, n, \{\theta_i, \theta_{ij} = \frac{W_{ij}}{2} I\}$ with all $|W_{ij}| \leq W$, $|\theta_i| \leq T$; $\lambda^0, \{s_k\}$ step sizes}

$E_{const} \leftarrow 0$ {keeps track of Energy constant through reparameterizations}

for all $i \in \mathcal{V}$ **do**

$\theta_i \leftarrow \theta_i - \sum_{j \in \mathcal{N}(i)} W_{ij}/2$

$E_{const} += \sum_{j \in \mathcal{N}(i)} -W_{ij}/2$

end for

save all base $\{\theta_i, W_{ij}\}, E_{const}$ parameters

$\{\lambda_{C,F}\} \leftarrow$ some initial values λ^0, all ≥ 0; typically initialize all to 0

$t \leftarrow 0$ {iteration number}

{Main loop}

repeat

{First absorb the constraint terms into the energy parameters}

load all base $\{\theta_i, W_{ij}\}, E_{const}$ parameters

for all chordless cycles C **do**

for all odd $F \subseteq C$ **do**

for all edge $(i,j) \in F$ **do**

$W_{ij} \leftarrow W_{ij} + 2\lambda^t_{C,F}$

$\theta_i \leftarrow \theta_i - \lambda^t_{C,F}, \quad \theta_j \leftarrow \theta_j - \lambda^t_{C,F}$

end for

for all edge $(i,j) \in C \setminus F$ **do**

$W_{ij} \leftarrow W_{ij} - 2\lambda^t_{C,F}$

$\theta_i \leftarrow \theta_i + \lambda^t_{C,F}, \quad \theta_j \leftarrow \theta_j + \lambda^t_{C,F}$

$E_{const} += \lambda^t_{C,F}$

end for

end for

end for

{Now solve the ϵ-approx $\log Z_B$ problem on the local polytope}

run the algorithm from Weller and Jebara (2014a) using $\epsilon, \{\theta_i, W_{ij}\}$ to return $-\log Z_t$ at $\mu^t = \{q^*_i, \xi^*_{ij}\}$ using (Welling and Teh, 2001) $\xi^*_{ij}(q^*_i, q^*_j, W_{ij})$

$\mathcal{G}(\lambda^t) \leftarrow -\log Z_t + E_{const}$

{Update the $\{\lambda_{C,F}\}$ with subgradient g; Increment t}

for all $(i,j) \in \mathcal{E}$ **do**

$\text{mainDiag}_{ij} \leftarrow 1 + 2\xi^*_{ij} - q^*_i - q^*_j, \quad \text{offDiag}_{ij} \leftarrow q^*_i + q^*_j - 2\xi^*_{ij}$

end for

for all chordless cycles C **do**

for all odd $F \subseteq C$ **do**

$g_{C,F} = 1 - \sum_{(i,j) \in F} \text{mainDiag}_{ij} - \sum_{(i,j) \in C \setminus F} \text{offDiag}_{ij}$

end for

end for

$\lambda^{t+1} \leftarrow \max(\lambda^t + s_t g, 0)$ {This projects onto the feasible set, i.e. projected subgradient descent}

$t \leftarrow t + 1$

until convergence

output final $\mathcal{G}(\lambda^{t-1})$ as best lower bound on $-\log Z_{BC}$

solver, the final dual solution is also accurate to within the same ϵ, i.e. the final dual value less ϵ provides a lower bound on $-\log Z_B$ for the cycle polytope.

Rounding to yield a primal feasible solution was achieved by taking a minimum convex combination with the uniform distribution, which has pairwise marginal of $\begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}$ for each edge, so as just to satisfy all cycle inequalities.

In practice, solving over the cycle polytope was often *faster* than solving over the standard local polytope due to the following approach. We employed a schedule of declining ϵ values (typically $1, 0.5, 0.25, 0.1$). For each ϵ value, we would run the dual decomposition method *using an LP relaxation* for the multi-label discrete problem. We would then move to the next, smaller ϵ value, warm starting with the last set of dual λ values. At the very end, if an integer solution was not returned, then we would restart with the final ϵ value at the latest λ values, henceforth using an integer solver (we used gurobi for both the LP and ILP solvers). Computationally difficult problems tend to have strongly negative edge weights. The approach described tends to lead to dual variables λ s.t. in the inner solver, models with significantly less strongly negative edge weights are being used. Since the inner solver takes much more time for higher $|W_{ij}|$ values, this was often very helpful.

7.6.1.2 Optimizing over the marginal polytope

We present our dual decomposition approach to optimize over the marginal polytope. We impose $4\binom{n}{2}$ constraints: each $\delta_{ij}(x_i, x_j)$ dual variable enforces consistency for an edge (i, j) at settings $X_i = x_i, X_j = x_j$ (singleton consistency and summing to 1 follow from constraints of the local polytope).

$$\begin{aligned}
\min_{\mu\in\mathcal{M}} \mathcal{F}_\theta(\mu) &= \min_{\mu^K\in\mathcal{M}} \min_{\mu\in\mathcal{L}} \max_{\delta} \mathcal{F}_\theta(\mu) + \sum_{(i,j)\in\mathcal{E};x_i,x_j\in\{0,1\}} \delta_{ij}(x_i,x_j)\left[\mu^K_{ij}(x_i,x_j) - \mu_{ij}(x_i,x_j)\right] \\
&\geq \max_{\delta} \min_{\mu^K\in\mathcal{M}} \min_{\mu\in\mathcal{L}} \mathcal{F}_\theta(\mu) + \sum_{(i,j)\in\mathcal{E};x_i,x_j\in\{0,1\}} \delta_{ij}(x_i,x_j)\left[\mu^K_{ij}(x_i,x_j) - \mu_{ij}(x_i,x_j)\right] \\
&= \max_{\delta} \min_{\mu^K\in\mathcal{M}} \min_{\mu\in\mathcal{L}} \mathcal{F}_{\theta'}(\mu) + \sum_{(i,j)\in\mathcal{E};x_i,x_j\in\{0,1\}} \delta_{ij}(x_i,x_j)\mu^K_{ij}(x_i,x_j) \\
&= \max_{\delta} \left[\min_{\mu\in\mathcal{L}} \mathcal{F}_{\theta'}(\mu) + \min_{\mu^K\in\mathcal{M}} \sum_{(i,j)\in\mathcal{E};x_i,x_j\in\{0,1\}} \delta_{ij}(x_i,x_j)\mu^K_{ij}(x_i,x_j)\right] \qquad \text{(D.3)}
\end{aligned}$$

θ' is given by $\theta'_{ij}(x_i,x_j) = \theta_{ij}(x_i,x_j) + \delta_{ij}(x_i,x_j)\ \forall(i,j) \in \mathcal{E}; x_i,x_j \in \{0,1\}$ [since $\mathcal{F} = E - S, E = -\theta\cdot\mu$].

We use subgradient descent (as before with the cycle polytope) to attain a lower bound. Each iteration, we use the new μ^K when computing the subgradient. When minimizing over $\mu^K \in \mathcal{M}$, the optimum will be achieved at a vertex so we solve by enumeration over all 2^n vertices.

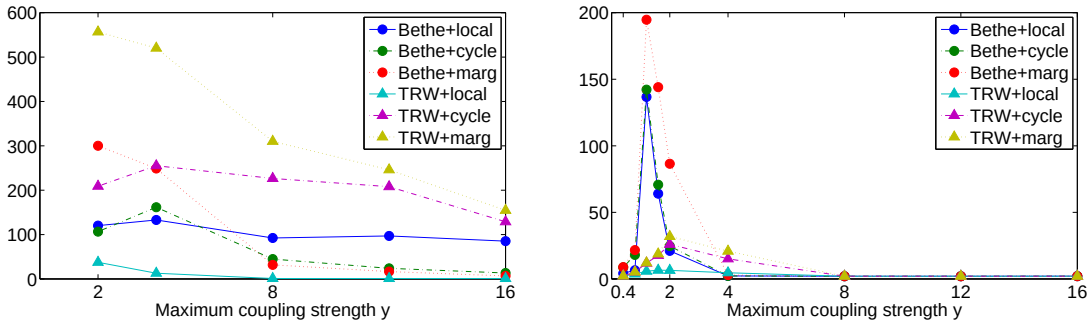
The term in square brackets in (D.3) is concave in δ, hence if $\epsilon = 0$ we converge to the optimum lower bound. Note strong duality is not guaranteed.

Rounding to achieve a primal feasible solution was achieved by solving an LP to find the closest point in the marginal polytope.

7.6.1.3 Further details on Frank-Wolfe

FW provides no runtime guarantee when applied to a non-convex surface such as the Bethe free energy. In Figure D.2 we show the empirical average number of iterations required to reach within 0.01 of the returned best value, comparing Bethe and TRW across local, cycle and marginal polytopes, for different parameter settings. Note that different convergence criteria were used for Bethe and TRW, with the duality gap examined for TRW, which is why we report this number of iterations, which provides a better basis for comparison than the total number of iterations.

At each iteration, to compute the optimal vertex of the appropriate polytope to move toward: for local and cycle polytopes, we solve the respective LP; for the marginal polytope, this is impractical, so we enumerate over all 2^n configurations, which clearly scales poorly. For the LP for the cycle



(a) General models

(b) Attractive models

Figure D.2: Average number of iterations of FW required to reach within 0.01 of the returned best value

polytope, the number of chordless cycles in a graph with n vertices is upper bounded by the number in a complete graph with n vertices, hence is $O(n^3)$, though it is typically not efficient to identify them.

Appendix E

Appendix for Clamping Variables and Approximate Inference

Here we provide the following additional information relating to Chapter 8 on Clamping Variables and Approximate Inference:

- Figure E.1 showing examples of the $f_c(x)$ function introduced in Lemma 8.4.2;
- In Section E.1, theoretical results on the Hessian leading to proofs of Theorem 8.4.4 and (a stronger version of) Theorem 8.4.5 from §8.4.1, and Lemma 8.6.1 from §8.6; and
- In Section E.2, additional illustrative experimental results with details on the Mpower selection heuristic.

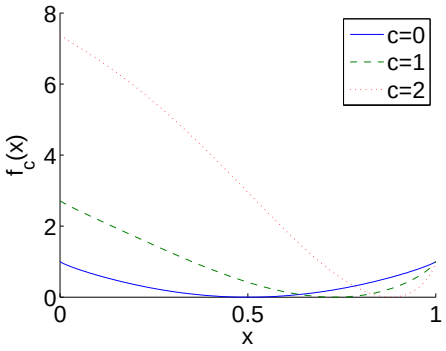


Figure E.1: Plots of upper bound $f_c(x)$ against x for various values of c

E.1 The Hessian and Proofs of Earlier Results

In this Section, we first discuss properties of the Hessian in §E.1.1, then use these in §E.1.2 to prove Theorems 8.4.4 and 8.4.5, and Lemma 8.6.1. Define the *interior* to be all points $q \in (0,1)^n$. Recall that $r^*(x) = (r_1^*(q_i), \ldots, r_{i-1}^*(q_i), r_{i+1}^*(q_i), \ldots, r_n^*(q_i))$ with corresponding pairwise terms $\{\xi_{ij}^*\}$, is an $\arg\max$ of $\mathcal{G}(q) = -\mathcal{F}(q)$ where q_i is held fixed at a particular value. For notational convenience, define $r_i^* = q_i$.

E.1.1 Properties of the Hessian

From (Weller and Jebara, 2013a), we have all terms of the Hessian matrix $H_{jk} = \frac{\partial^2 \mathcal{F}}{\partial q_j \partial q_k}$:

$$H_{jk} = \begin{cases} \frac{q_j q_k - \xi_{jk}}{T_{jk}} & \text{if } (j,k) \in \mathcal{E} \\ 0 & \text{if } (j,k) \notin \mathcal{E} \end{cases}, \quad H_{jj} = -\frac{d_j - 1}{q_j(1-q_j)} + \sum_{k \in \mathcal{N}(j)} \frac{q_k(1-q_k)}{T_{jk}}, \tag{E.1}$$

where $d_j = |\mathcal{N}(j)|$ is the degree of j, and $T_{jk} = q_j q_k (1-q_j)(1-q_k) - (\xi_{jk} - q_j q_k)^2 \geq 0$, with equality only at an edge (i.e. q_j or $q_k \in \{0,1\}$). For an attractive edge (j,k), in the interior, as shown in (Weller and Jebara, 2013a, Lemma 14 in Supplement), $\xi_{jk} - q_j q_k > 0$ and hence $H_{jk} < 0$.

Now write

$$H_{jj} = \frac{1}{q_j(1-q_j)} + \sum_{k \in \mathcal{N}(j)} \left(\frac{q_k(1-q_k)}{T_{jk}} - \frac{1}{q_j(1-q_j)} \right). \tag{E.2}$$

Consider the term in large parentheses for some $k \in \mathcal{N}(j)$. First observe that the term is ≥ 0, strictly > 0 in the interior, whether the edge is attractive or repulsive. Since $H_{jj} > 0$, on the surface $\frac{\partial \mathcal{F}}{\partial q_j}\Big|_{r^*} = 0$, we have

$$\frac{\partial r_j^*}{\partial r_k^*} = -\frac{H_{jk}}{H_{jj}}\Big|_{r^*}, \tag{E.3}$$

which also holds for $k = i$ where we define $r_i^* = q_i$.

Further, we may incorporate the term for k to obtain

$$H_{jj} \geq \frac{1}{q_j(1-q_j)} + \frac{q_k(1-q_k)}{T_{jk}} - \frac{1}{q_j(1-q_j)} = \frac{q_k(1-q_k)}{T_{jk}},$$

with equality iff j has no neighbor other than k (again allowing $k = i$), in which case,

$$\frac{\partial r_j^*}{\partial r_k^*} = \frac{\xi_{jk}^* - r_j^* r_k^*}{r_k^*(1-r_k^*)}. \tag{E.4}$$

We also show the following results, though the remainder of this Section §E.1.1 is not used until later when we prove Theorem 8.4.5 in §E.1.2.1.

Considering the term in large parentheses from (E.2), using the definition of T_{jk}, we may write

$$\left(\frac{q_k(1-q_k)}{T_{jk}} - \frac{1}{q_j(1-q_j)}\right) = \left(\frac{\xi_{jk}-q_jq_k}{T_{jk}}\right)\left(\frac{\xi_{jk}-q_jq_k}{q_j(1-q_j)}\right) = -H_{jk}\beta_{j\to k}, \qquad \text{(E.5)}$$

where we define $\beta_{j\to k} = \frac{\xi_{jk}-q_jq_k}{q_j(1-q_j)}$, which as mentioned in the main paper after Theorem 8.4.4, is equal to $\frac{\mathrm{Cov}_q(X_j,X_k)}{\mathrm{Var}_q(X_j)}$, called in finance the beta of X_k with respect to X_j. This is clearly positive for an attractive edge. We next show that the range of $\beta_{j\to k}$ is bounded, as would be expected for beta.

Lemma E.1.1. *In the interior, for an edge (j,k): if attractive, $0 < \beta_{j\to k} \le \frac{\alpha_{jk}}{\alpha_{jk}+1} = 1 - e^{-W_{jk}} < 1$; if repulsive, $-1 < e^{W_{jk}} - 1 = \alpha_{jk} \le \beta_{j\to k} < 0$. In either case, $|\beta_{j\to k}| = \left|\frac{\xi_{jk}-q_jq_k}{q_j(1-q_j)}\right| \le 1 - e^{-|W_{jk}|} < 1$.*

Proof. This follows from (Weller and Jebara, 2013a, Lemma 6) and the corresponding flipped result (Weller and Jebara, 2014a, Lemma 10 in Supplement; consider each of the 2 cases for p_{jk} therein). □

Define $\beta^*_{j\to k} = \beta_{j\to k}\big|_{r^*}$. Regarding (E.4), note that $\beta^*_{j\to k} \ge \frac{\partial r^*_k}{\partial r^*_j}$ with equality iff $\mathcal{N}(k) = \{j\}$. This notation will become clear when we use it in §E.1.2.1 to prove Theorem 8.4.5.

E.1.2 Derivation of earlier results

Using the results of §E.1.1, we first provide a general Theorem from which Lemma 8.6.1 follows as an immediate corollary.

Theorem E.1.2. *For any binary pairwise MRF where the Bethe free energy is convex, adding further variables to the model and holding them at fixed singleton marginal values (optimum pairwise marginals are computed using the formula of Welling and Teh, 2001), leaves the Bethe free energy over the original variables convex.*

Proof. The Bethe free energy is convex ⇔ the Hessian is everywhere positive semi-definite. When new variables are added to the system, considering (E.1) and (E.2), the only effect on the sub-Hessian restricted to the original variables is potentially to increase the diagonal terms H_{jj} for any original variable j which is adjacent to a new variable. By Weyl's inequality, this can only increase the minimum eigenvalue of the sub-Hessian, and the result follows. □

Since the Bethe free energy is convex for any model whose entire topology contains at most one cycle (Pakzad and Anantharam, 2002), Lemma 8.6.1 follows.

We next turn to Theorem 8.4.4, then use this to prove a stronger version of Theorem 8.4.5. Keep in mind that, as shown in (Weller and Jebara, 2013a), each stationary point lies in an open region in the interior $q \in (0,1)^n$. Further, as discussed in §8.4.1, we assume that at any $\arg\max$ point $r^*(q_i)$, the reduced Hessian $H_{\setminus i}$ is non-singular. Hence, writing $\nabla_{n-1}\mathcal{F}\big|_{q_i}$ for the $(n-1)$-vector of partial derivatives $\frac{\partial \mathcal{F}(q)}{\partial q_j}\Big|_{q_i}$ $\forall j \neq i$, there is an open region around any $(q_i, r^*(q_i))$ where the function $\nabla_{n-1}\mathcal{F}\big|_{q_i} = 0$ may be well approximated by an invertible linear function, allowing us to solve (as in the implicit function theorem) for the total derivatives $\frac{dr_j^*}{dq_i}$ as the unique solutions to the linear system $\frac{dr_j^*}{dq_i} = \frac{\partial r_j^*}{\partial q_i} + \sum_{k \notin \{i,j\}} \frac{\partial r_j^*}{\partial r_k^*}\frac{dr_k^*}{dq_i}$ $\forall j \neq i$, where here $\frac{\partial r_j^*}{\partial r_k^*}$ always means on the surface $\nabla_{n-1}\mathcal{F}\big|_{q_i} = 0$. In addition, since $H_{\setminus i}$ is real, symmetric, positive definite, with all main diagonal ≥ 0 and all off-diagonal ≤ 0, it is an M-matrix (indeed a Stieltjes matrix), which we shall use in §E.1.2.1. We assume these points for the rest of this Section.

Notation: Let $D_j = \frac{dr_j^*}{dq_i}$, and $\partial_{jk} = \frac{\partial r_j^*}{\partial r_k^*}$, so $D_j = \sum_{k \notin \{i,j\}} \partial_{jk} D_k + \partial_{ji}$ $\forall j \neq i$. For notational convenience, define $r_i^* = q_i$ and take $D_i = 1$. Let $[n] = \{1, \ldots, n\}$ and $[n] \setminus i = \{1, \ldots, n\} \setminus \{i\}$. Note that $\partial_{jk} = \frac{\partial r_j^*}{\partial r_k^*} \leq \beta^*_{k \to j}$ (equality iff j has no neighbor other than k), as defined above. We shall write Hessian terms such as H_{jk} to mean $H_{jk}\big|_{r^*}$ where this is implied by the context.

We first need the following Lemma.

Lemma E.1.3. *Consider a MRF with n variables, where then one more variable X_{n+1} is added with singleton marginal r^*_{n+1}, adjacent to exactly one of the original n variables, say X_a with $a \in [n]$ (note we allow $a = i$), then: $D_1, \ldots, D_n$ are unaffected, and $D_{n+1} = \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{r^*_a(1-r^*_a)} D_a$.*

Proof. We have the linear system $D_j = \sum_{k \notin \{i,j\}} \partial_{jk} D_k + \partial_{ji}$ $\forall j \in [n] \setminus i$. When X_{n+1} is added, this yields a new equation for D_{n+1}, which as shown in (E.4), is $D_{n+1} = \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{r^*_a(1-r^*_a)} D_a$, and the only other equation that changes is the one for D_a, where we write ∂'_{ak} and ∂'_{ai} for the new coefficients. Hence, it is sufficient to show that the earlier solutions for $D_1, \ldots, D_n$ satisfy the new equation for D_a, i.e. if $D_a = \sum_{k \in [n+1] \setminus \{i,a\}} \partial'_{ak} D_k + \partial'_{ai}$.

Observe from (E.3) that $\partial'_{ak} = \partial_{ak} H_{aa} / H'_{aa}$ $\forall k \in [n]$, where H'_{aa} incorporates the new X_{n+1}

variable. Hence,

$$\begin{aligned}
\sum_{k\in[n+1]\setminus\{i,a\}} \partial'_{ak} D_k + \partial'_{ai} &= \frac{H_{aa}}{H'_{aa}} \left(\sum_{k\notin\{i,j\}} \partial_{ak} D_k + \partial_{ai} \right) + \partial'_{a,n+1} D_{n+1} \\
&= \frac{H_{aa}}{H'_{aa}} D_a + \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{T_{a,n+1} H'_{aa}} \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{r^*_a (1 - r^*_a)} D_a \quad \text{by (E.3), (E.1) and just above} \\
&= \frac{D_a}{H'_{aa}} \left[H_{aa} + \frac{\left(\xi^*_{a,n+1} - r^*_a r^*_{n+1}\right)^2}{T_{a,n+1} r^*_a (1 - r^*_a)} \right] \\
&= \frac{D_a}{H'_{aa}} \left[H_{aa} + \left(\frac{r^*_{n+1}(1 - r^*_{n+1})}{T_{a,n+1}} - \frac{1}{r^*_a (1 - r^*_a)} \right) \right] \quad \text{(definition of } T_{a,n+1}\text{)} \\
&= \frac{D_a}{H'_{aa}} \left[H_{aa} + (H'_{aa} - H_{aa}) \right] = D_a \qquad \square
\end{aligned}$$

Theorem 8.4.4 may now be proved by induction on $|C_k|$. The base case $|C_k| = 1$ follows from (E.4). The inductive step follows from Lemma E.1.3 by considering a leaf.

E.1.2.1 Proof of (stronger version of) Theorem 8.4.5:

As above, we have the linear system given by the following equations:

$$D_j = \sum_{k\notin\{i,j\}} \partial_{jk} D_k + \partial_{ji} \quad \forall j \neq i \qquad \Leftrightarrow -\partial_{ji} = \sum_{k\neq i} [\partial_{jk} - \delta_{jk}] D_k \tag{E.6}$$

$$\text{with } \partial_{jk} = \frac{\partial r^*_j}{\partial r_k *} = -\frac{H_{jk}}{H_{jj}} \;\; k \notin \{i,j\}, \;\; \partial_{jj} := 0, \qquad \partial_{ji} = \frac{\partial r^*_j}{\partial q_i} = -\frac{H_{ji}}{H_{jj}}, \;\; \delta_{jk} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}.$$

Hence we may rewrite (E.6), multiplying by $-H_{jj}$, to give the equivalent system

$$\sum_{k\neq i} H_{jk} D_k = -H_{ji} \quad \forall j \neq i \tag{E.7}$$

Note equation (E.7) makes intuitive sense: for each variable X_j, we have $\mathcal{F}_j = 0$ at a stationary point, then taking the total derivative with respect to q_i gives $H_{ji} + \sum_{k\neq i} H_{jk} D_k = 0$.

By Theorem 8.4.4, we have the complete solution vector $D_k \; \forall k \neq i$ provided the topology is acyclic. In this setting, we rewrite the result of Theorem 8.4.4 using the β^* notation from above: $D_k = \prod_{(s\to t)\in P(i\rightsquigarrow k)} \beta^*_{s\to t}$, where here $P(i \rightsquigarrow k)$ is the *unique* path from i to k.

For a general graph, there may be many paths from i to k. Let $\Pi(i \rightsquigarrow k)$ be the set of all such directed paths. For any r^*, for any particular path $P(i \rightsquigarrow k) \in \Pi(i \rightsquigarrow k)$, define its *weight* to be

$W[P(i \rightsquigarrow k)] = \prod_{(s \to t) \in P(i \rightsquigarrow k)} \beta^*_{s \to t}$. We shall prove the following result:

$$D_k \geq \max_{P(i \rightsquigarrow k) \in \Pi(i \rightsquigarrow k)} W[P(i \rightsquigarrow k)]. \tag{E.8}$$

Note this is clearly stronger than Theorem 8.4.5 since $\forall j \in \mathcal{N}(i)$, the path going directly $i \to j$ is one member of $\Pi(i \rightsquigarrow j)$, though in general there may be many others.

For any particular r^*, let G' be the weighted directed graph formed from the topology of the MRF by replacing each undirected edge $s - t$ by two directed edges: $s \to t$ with weight $\beta^*_{s \to t}$ and $t \to s$ with weight $\beta^*_{t \to s}$. Note that in an attractive model, all $\beta^*_{s \to t} \in (0, 1)$, see Lemma E.1.1.

It is a simple application of Dijkstra's algorithm to construct from G' a tree of all maximum weight directed paths from i to each vertex $j \neq i$, which we call $\mathcal{T}$.[1] (For our purpose we just need to know that such a tree $\mathcal{T}$ exists.)

We want to solve (E.7), which we write as $H_{\setminus i} D = -H_i$, where we want to solve for D, which is the vector of $D_k \; \forall k \neq i$, and H_i is the ith column of H without its ith element. Let $H^{\mathcal{T}}_{\setminus i}$ be the reduced Hessian for the model on $\mathcal{T}$ (which is missing some edges), and $H^{\mathcal{T}}_i$ be the ith column of the Hessian for the model on $\mathcal{T}$ without its ith element. In the sub-model with only the edges of $\mathcal{T}$, by construction and Theorem 8.4.4, $D^{\mathcal{T}}_k = \max_{P(i \rightsquigarrow k) \in \Pi(i \rightsquigarrow k)} W[P(i \rightsquigarrow k)]$. Hence, it is sufficient to show that adding the extra edges from $\mathcal{T}$ to G cannot decrease any D_k. This forms the remainder of the proof, where we shall require the following nonsingular M-matrix property of $H_{\setminus i}$: its inverse is elementwise non-negative (Fan, 1958, Theorem 5').

Let $\Delta = H_{\setminus i} - H^{\mathcal{T}}_{\setminus i}$ (this accounts for edges in $E(G) \setminus E(\mathcal{T})$ not incident to i), $\eta = H_i - H^{\mathcal{T}}_i$ (this accounts for edges in $E(G) \setminus E(\mathcal{T})$ incident to i) and $\delta = D - D^{\mathcal{T}}$. We must show that $\delta \geq 0$ elementwise. We have $H^{\mathcal{T}}_{\setminus i} D^{\mathcal{T}} = -H^{\mathcal{T}}_i$ and $H_{\setminus i} D = -H_i$, hence $H^{\mathcal{T}}_{\setminus i} D^{\mathcal{T}} - \eta = -H^{\mathcal{T}}_i - \eta = -H_i = H_{\setminus i} D = (H^{\mathcal{T}}_{\setminus i} + \Delta)(D^{\mathcal{T}} + \delta)$, hence $-\eta = (H^{\mathcal{T}}_{\setminus i} + \Delta)\delta + \Delta D^{\mathcal{T}} \Leftrightarrow \delta = (H_{\setminus i})^{-1}(-\eta - \Delta D^{\mathcal{T}})$. Thus, it is sufficient to show that the $(n-1)$ vector $-\eta - \Delta D^{\mathcal{T}}$ is elementwise non-negative.

Recall (E.1) and (E.5). $-\eta - \Delta D^{\mathcal{T}}$ may be written as the sum of $-\eta_e - \Delta_e D^{\mathcal{T}}$, with one η_e and Δ_e for each edge $e = (s, t)$ in $E(G) \setminus E(\mathcal{T})$. For each such edge e, we have 2 cases:

Case 1, $i \notin \{s, t\}$: $\eta_e = 0$; Δ_e has only 4 non-zero elements, at locations $(s, s), (s, t), (t, s), (t, t)$.

[1] We want the max of the prod of edge weights $\Leftrightarrow$ max of the log of the prod of edge weights $\Leftrightarrow$ max of the sum of the log of edge weights (all negative) $\Leftrightarrow$ min of the sum of - log of the edge weights (all positive); so really we construct the usual shortest directed paths tree using - log of the edge weights, which are all positive.

Showing only these elements,

$$\Delta_e = \begin{array}{c} \\ s \\ t \end{array} \begin{array}{c} \begin{array}{cc} s & t \end{array} \\ \begin{pmatrix} -H_{st}\beta^*_{s\to t} & H_{st} \\ H_{st} & -H_{st}\beta^*_{t\to s} \end{pmatrix} \end{array} = -H_{st} \begin{array}{c} \\ s \\ t \end{array} \begin{array}{c} \begin{array}{cc} s & t \end{array} \\ \begin{pmatrix} \beta^*_{s\to t} & -1 \\ -1 & \beta^*_{t\to s} \end{pmatrix} \end{array}, \text{ where } -H_{st} > 0 \text{ for an attractive edge.}$$

Hence, $-\eta_e - \Delta_e D^{\mathcal{T}}$ is 0 everwhere except element s which is $-H_{st}(D^{\mathcal{T}}_t - D^{\mathcal{T}}_s \beta^*_{s\to t})$, and element t which is $-H_{st}(D^{\mathcal{T}}_s - D^{\mathcal{T}}_t \beta^*_{t\to s})$. Observe that both expressions are ≥ 0 by construction of $\mathcal{T}$ (for example, considering the first bracketed term, observe that $D^{\mathcal{T}}_t$ is the maximum weight of a path from i to t, whereas $D^{\mathcal{T}}_s \beta^*_{s\to t}$ is the weight of a path to t going through s).

Case 2, $i \in \{s, t\}$: WLOG suppose the edge is (i, s). $-\eta_e$ is zero everywhere except element s which is $-H_{is}$ (positive). Δ_e has just one non-zero element at (s, s) which is $-H_{is}\beta^*_{s\to i}$. Hence, $-\eta_e - \Delta_e D^{\mathcal{T}}$ is 0 everwhere except element s which is $-H_{is}(1 - D^{\mathcal{T}}_s \beta^*_{s\to i}) > 0$ by Lemma E.1.1.

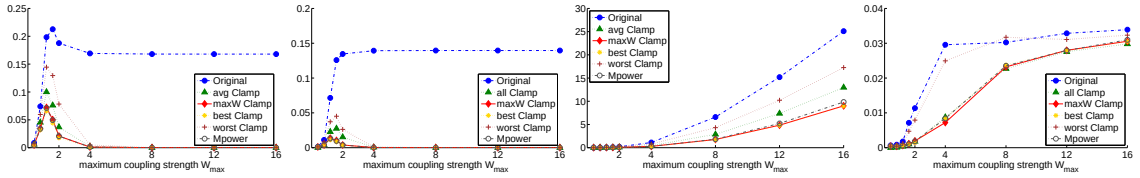
This completes the proof.

E.2 Additional Experiments

All of the experiments reported in §8.5 were also run at other settings. In particular, the earlier results show the poor performance of the standard Bethe approximation in estimating singleton marginals for attractive models with low singleton potentials, and indicate how clamping repairs this. Here, in Figures E.2-E.4, we show results for the same topologies using the higher singleton potentials $T_{max} = 2$ for attractive models, and also show results with low singleton potentials $T_{max} = 0.1$ for general (non-attractive) models.

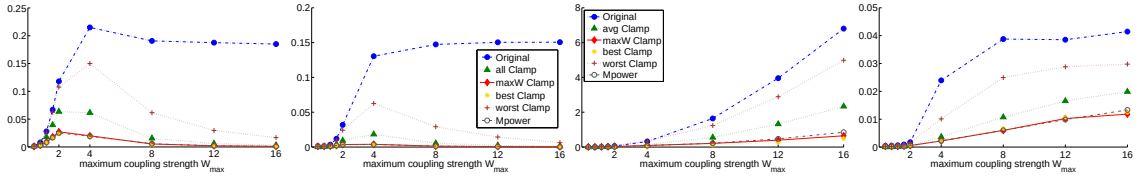
Note that in some examples of attractive models, when the 'worst clamp' variable was clamped, the resulting Bethe approximation to $\log Z$ appears to worsen (see Figure E.4a), which seems to conflict with Theorem 8.4.1. The explanation is that in these examples, Frank-Wolfe is failing to find the global Bethe optimum, as was confirmed by spot checking.

Next we show results for a particular fixed topology we call a 'lamp', see Figure E.5, which illustrates how maxW can sometimes select a poor variable to clamp. We explain the Mpower selection heuristic and demonstrate that it performs much better on this topology.



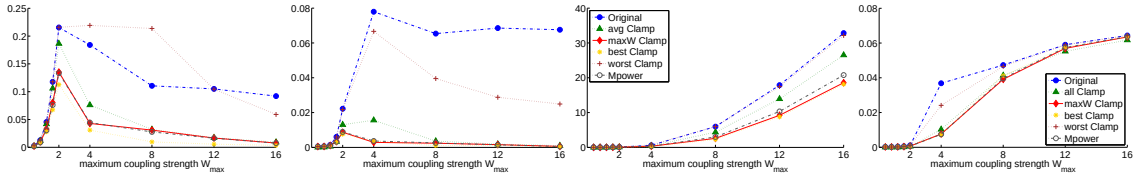
(a) attractive $\log Z, T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z, T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure E.2: Average errors vs true, **complete graph on $n = 10$**. Consistent legend throughout.



(a) attractive $\log Z, T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z, T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure E.3: Average errors vs true, **random graph on $n = 10, p = 0.5$**. Consistent legend throughout.



(a) attractive $\log Z, T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z, T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure E.4: Average errors vs true, **random graph on $n = 50, p = 0.1$**. Consistent legend throughout.

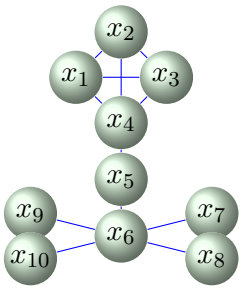


Figure E.5: 'Lamp' topology.
maxW is likely to choose x_6 since it has the highest degree, but x_4 is typically a better choice since it lies on cycles. Mpower can recognize this and make a better choice.

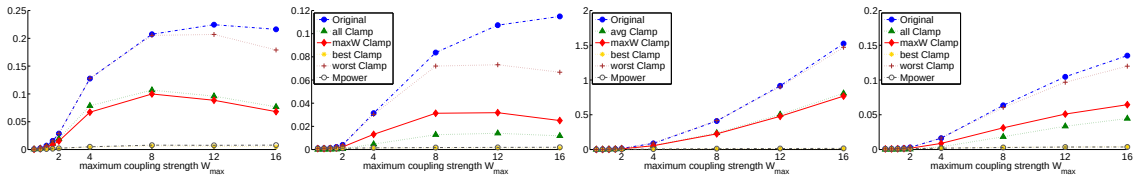
E.2.1 Mpower heuristic

We would like an efficient way to select a variable to clamp which lies on many heavy simple cycles. One problem is how to define heavy. Even with a good definition, it is still NP-hard to search over all simple cycles. The idea for Mpower is as follows: assign each edge (i, j) a weight based on $|W_{ij}|$ and create a matrix M of these weights. If M is raised to the kth power, then the ith diagonal element in M^k is the sum over all paths of length k from i to i of the product of the edge weights along the path. Ideally, we might consider the sum $\sum_{k=1}^{\infty} M^k$ and use the diagonal elements to rank the vertices, choosing the one with highest total score. Recalling (E.8), it is sensible to assign edge weights M_{ij} based on possible $\beta^*_{i\to j}$ values. Given Lemma E.1.1, a first idea is to use $1 - e^{-|W_{ij}|}$.

However, we'd like to be sure that the matrix series $\sum_{k=1}^{\infty} M^k$ is convergent, allowing it to be computed as $(I - M)^{-1} - I$ (since we shall be interested only in ranking the diagonal terms, in fact there is no need to subtract I at the end). Thus, we need the spectral radius $\rho(M) < 1$. A sufficient condition is that all row sums are < 1. Since each term $1 - e^{-|W_{ij}|} < 1$ and there at most $n - 1$ such elements in any row, our first heuristic was to set $M_{ij} = \frac{1}{n-1}(1 - e^{-|W_{ij}|})$. We then made two adjustments.

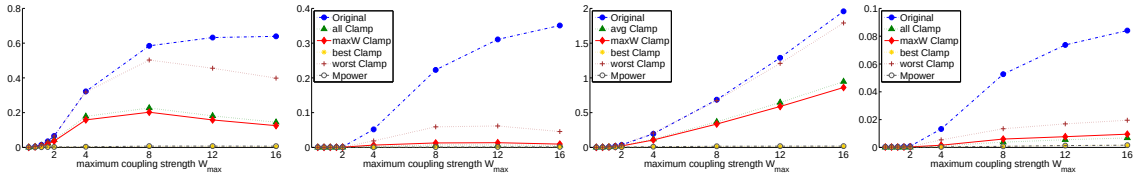
First, note that the series $\sum_{k=1}^{\infty} M^k$ overcounts all cycles, though at an exponentially decaying rate. It is hard to repair this. However, it also includes relatively high value terms coming from paths from i to any neighbor j and straight back again, along with all powers of these. We should like to discard all of these, hence from each ith diagonal term of $(I-M)^{-1}$, we subtract $s_i/(1-s_i)$, where s_i is the ith diagonal term of M^2. This is very similar to the final version we used, and gives only very marginally worse results on the examples we considered.

For our final version, we observe that $1 - e^{-|W_{ij}|}$ decays rapidly, and $\approx \tanh \frac{|W_{ij}|}{2}$. Given the form of the loop series expansion for a single cycle, which contains $\tanh \frac{W_{ij}}{4}$ terms (Weller et al., 2014, Lemma 5), we tried instead using $M_{ij} = \frac{1}{n-1} \tanh \frac{|W_{ij}|}{4}$, and it is for this heuristic that results are shown in Figures E.6 (for $T_{max} = 2$) and E.7 (for $T_{max} = 0.1$). Observe that for this topology, Mpower performs close to optimally (almost the same results as for best Clamp), significantly outperforming maxW in most settings. Note, however, that in the experiments on random graphs reported in §8.5, Mpower did not outperform the simpler maxW heuristic. In future work, we hope to improve the selection methods.



(a) attractive $\log Z$, $T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z$, $T_{max} = 2$ (d) general margs, $T_{max} = 2$

Figure E.6: Average errors vs true, **'lamp' topology $T_{max} = 2$**. Consistent legend throughout. Mpower performs well, significantly better than maxW.



(a) attractive $\log Z$, $T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z$, $T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure E.7: Average errors vs true, **'lamp' topology $T_{max} = 0.1$**. Consistent legend throughout. Mpower performs well, significantly better than maxW for $\log Z$.