Polynomial-Time Exact Inference in NP-Hard Binary MRFs via Reweighted Perfect Matching

Nicol N. Schraudolph
jmlr@schraudolph.org
adaptive tools AG
Canberra ACT 2602, Australia

Abstract

We develop a new form of reweighting (Wainwright et al., 2005b) to leverage the relationship between Ising spin glasses and perfect matchings into a novel technique for the exact computation of MAP states in hitherto intractable binary Markov random fields. Our method solves an $n \times n$ lattice with external field and random couplings much faster, and for larger n, than the best competing algorithms. It empirically scales as $O(n^3)$ even though this problem is NP-hard and non-approximable in polynomial time. We discuss limitations of our current implementation and propose ways to overcome them.

1 INTRODUCTION

It is well known that inference in Markov random fields (MRFs) is NP-hard in general; this includes tasks such as calculating the partition function, finding an optimal (*maximum a posteriori*, MAP, or ground) state, conditioning on a subgraph, computing marginal probabilities, and so forth. Much work in graphical models therefore proceeds via the following route:

1. identify a tractable class of graphs (distributions);
2. develop efficient, exact inference methods for it;
3. leverage these methods into techniques applicable (at a cost) to a wider class of graphs/distributions.

The prime example for this is belief propagation a.k.a. message passing (Pearl, 1988), which at its heart is an efficient inference mechanism for trees, *i.e.,* graphs without cycles. By aggregating clusters of nodes into supernodes, the junction tree algorithm (Lauritzen and Spiegelhalter, 1988) leverages this into an exact inference method for any graph, at a cost exponential in its *treewidth*, a measure of its structural complexity. Loopy belief propagation (Weiss, 1997; Frey and MacKay, 1998; Yedidia et al., 2001) provides a heuristic alternative for graphs with cycles that can yield good results but may fail to converge.

Another example is the graph cut approach, in which binary MRF ground states are computed in polynomial time via a duality with maximum-weight network flow. In its original form (Greig et al., 1989) this construction only applies to MRFs whose edge potentials obey a submodularity constraint (Kolmogorov and Zabih, 2004). The QPBO algorithm (Kolmogorov and Rother, 2007) leverages this into a method that provides a partial labeling of the ground state for MRFs with some nonsubmodular edges; this can be improved further by solving a series of related QPBO problems (Rother et al., 2007). These methods tend to work well as long as there are not too many nonsubmodular edges.

The restriction to binary node states can be overcome by employing constructions that reduce a non-binary MRF to either a sequence of binary MRFs on the same graph (Boykov et al., 2001), or a single binary MRF on a more complex graph (Ishikawa, 2003).

Globerson and Jaakkola (2007) have noted that binary MRFs are closely related to Ising spin glasses, which have long been studied in statistical physics. When defined over a *planar* graph, partition function (Kasteleyn, 1961; Fisher, 1961) and ground state (Bieche et al., 1980; Barahona, 1982) of an Ising model can be computed in polynomial time by establishing a correspondence with perfect matchings in a related graph. In this approach to inference in graphical models the graph's *genus* (as opposed to its treewidth, or submodularity) thus determines its tractability.

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

It is possible to leverage the polynomial-time computation of the partition function for planar Ising models into a general method for nonplanar ones, at a cost exponential in the graph's genus. A proposal of Kasteleyn (1961) to this effect was fleshed out by Galluccio and Loebl (1999), though implemented only for toroidal (*i.e.,* genus one) lattices (Galluccio et al., 2000). Here we present and implement a generic and practical algorithm to lift the exact ground state computation from planar to nonplanar Ising models.

The remainder of this paper is organized as follows: In Section 2 we briefly review the method of Schraudolph and Kamenetsky (2008, 2009) for computing a ground state of a planar Ising model, and show how it falls short on nonplanar graphs. Section 3 addresses this shortfall by using reweighting (Wainwright et al., 2005b) in a novel way, employing convex mixtures of graph embeddings. In Section 4 we find that a straightforward implementation of this idea empirically yields polynomial-time convergence on an NP-hard and non-approximable problem of great practical interest. Sections 5 and 6 describe important refinements of our algorithm, before we conclude with a brief discussion (Section 7).

2 ISING MODELS

Schraudolph and Kamenetsky (2008, 2009) define an Ising model to be a pairwise binary MRF defined over a graph $G(\mathcal{V},\mathcal{E})$ with an energy function of the form

$$E(\boldsymbol{y}) \;:=\; \sum_{(i,j)\in\mathcal{E}} [\![y_i \neq y_j]\!]\,\theta_{ij}, \qquad (1)$$

where $\boldsymbol{y} \in \{0,1\}^{|\mathcal{V}|}$ is the model's binary state vector, $[\![\cdot]\!]$ denotes the indicator function, and θ_{ij} the *disagreement cost* of edge (i,j). Any pairwise binary MRF can be expressed in this form (1) by replacing its unary (node) potentials with edges to an additional field (or bias) node whose value is fixed (Globerson and Jaakkola, 2007; Schraudolph and Kamenetsky, 2008, 2009). Note that this makes the Ising model topologically more complex than the corresponding MRF; planar Ising models correspond to MRFs that are outerplanar w.r.t. nodes with non-zero unary potential.

2.1 PLANAR ISING MODELS

A ground state of a planar Ising model can be computed as follows:[1]

1. Embed G (*i.e.,* draw it without edge crossings) on a plane or sphere. This can be done in linear time (Boyer and Myrvold, 2004).

[1] We only give a brief overview here; see Schraudolph and Kamenetsky (2008) for a detailed treatment.

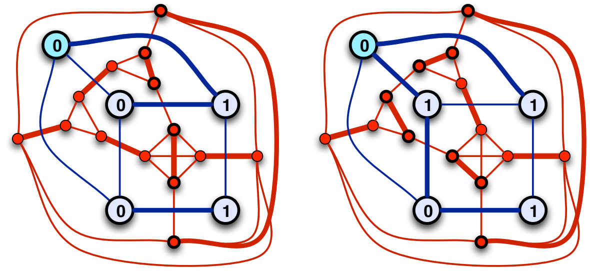


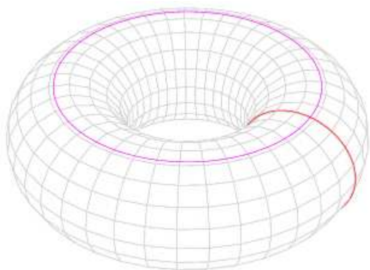
Figure 1: A planar Ising model with four nodes (large, blue) plus bias node (cyan, top left) and its expanded dual (small nodes, red), in two different states (left & right), with the corresponding cut of the graph and perfect matching of its expanded dual (both in bold).

2. Construct the dual graph G^*, which has a node for each face of G's embedding. The weight of an edge in G^* is the disagreement cost of the edge of G that it crosses.
3. Expand G^* by replacing each of its nodes with a k-clique, where k is the degree of the node. The new clique-internal edges are given zero weight.
4. Compute a minimum-weight perfect matching $\mathcal{M}^*(-\boldsymbol{\theta})$ in the expanded dual with negated edge costs. A perfect matching is a subset of the edges in which each node has degree one; minimum-weight perfect matchings can be computed in $O(|\mathcal{V}|^3|\mathcal{E}|)$ by the Blossom algorithm (Edmonds, 1965a,b; Kolmogorov, 2009).[2]
5. The complement $\mathcal{C}^*(\boldsymbol{\theta}) := \mathcal{E}\backslash\mathcal{M}^*(-\boldsymbol{\theta})$ is a minimum-cost cut of G, hence induces a ground state $\boldsymbol{y}^* := \operatorname{argmin}_{\boldsymbol{y}} E(\boldsymbol{y})$ of G.

The key to making this construction work is the complementarity between perfect matchings of the expanded dual and cuts of the original graph, used in the final step. Consider Figure 1: By definition, every perfect matching $\mathcal{M}$ of the expanded dual matches an even number of nodes in each clique via clique-internal edges. Thus the perimeter of every face of the original model is crossed an even number of times by dual edges that are *not* part of the matching.

Let us call two sets $\mathcal{A}, \mathcal{B}$ *consistent* w.r.t. each other, and write $\mathcal{A} \triangleq \mathcal{B}$, iff their intersection contains an even number of elements. Above we have seen that by construction, $\mathcal{E}\backslash\mathcal{M} \triangleq \mathcal{F}$ for every face perimeter $\mathcal{F} \subseteq \mathcal{E}$ of G. The set $\boldsymbol{\mathcal{F}} \subseteq 2^{\mathcal{E}}$ of face perimeters of a plane embedded graph is a *cycle basis* for the graph, *i.e.,* every cycle $\mathcal{O} \subseteq \mathcal{E}$ can be composed from face perimeters via symmetric set differences. It is easy to show that

[2] Gabow (1990) established an $O(|\mathcal{V}|(|\mathcal{E}| + |\mathcal{V}|\log|\mathcal{V}|))$ time complexity, but this has never been implemented.



courtesy of
Wikipedia

Figure 2: A lattice graph embedded on the surface of a torus; the two cycles indicated form a homology basis.

symmetric set differences preserve consistency, *i.e.,*

$$\mathcal{A} \mathrel{\hat{=}} \mathcal{B} \wedge \mathcal{A} \mathrel{\hat{=}} \mathcal{C} \implies \mathcal{A} \mathrel{\hat{=}} (\mathcal{B} \triangle \mathcal{C}) . \tag{2}$$

Thus $\mathcal{E} \backslash \mathcal{M} \mathrel{\hat{=}} \mathcal{O}$ for every cycle $\mathcal{O} \subseteq \mathcal{E}$ — which means that $\mathcal{E} \backslash \mathcal{M}$ is a cut. Since $\mathcal{M}^*(-\boldsymbol{\theta})$ has minimum *negated* weight, it is in fact a maximum-weight perfect matching w.r.t. $\boldsymbol{\theta}$. Its complement $\mathcal{C}^*(\boldsymbol{\theta})$ must therefore be a cut of minimum cost, hence induces a state that minimizes the energy (1) of the model.

2.2 NONPLANAR ISING MODELS

Although nonplanar graphs cannot be drawn on a plane or sphere without edge crossings, they can be embedded in topologically more complex surfaces. The genus (*i.e.,* number of holes) of the topologically simplest orientable surface in which a graph can be embedded is its genus g. A toroidal grid for instance has genus one since it can be drawn on a torus (Figure 2).

The dual graph G^* is well-defined for any embedding of G regardless of its genus, and so Steps 1–4 of the method for planar graphs (Section 2.1) can be applied here just as well. The only caveat is that finding an optimal (*i.e.,* minimum-genus) embedding is NP-complete (Thomassen, 1989) and non-approximable (Chen et al., 1997) in general. The method we will introduce in Section 3, however, does not require optimal embeddings. We have developed heuristics which yield embeddings of sufficient quality for our purposes in reasonable time (in preparation).

What breaks down for nonplanar Ising models is the complementarity between perfect matchings in the dual and cuts in the model graph: nonplanar embedding surfaces have holes, and graph cycles that thread or encircle a hole (*e.g.,* those shown in Figure 2) cannot be generated from symmetric differences of face perimeters. In short, $\boldsymbol{\mathcal{F}}$ is not a cycle basis for nonplanar embeddings, so $\mathcal{E} \backslash \mathcal{M}$ may no longer be a cut.

When applied to a nonplanar embedding, Step 5 in Section 2.1 yields an *extended ground state* (Thomas and Middleton, 2007): a minimum-weight edge set that is consistent w.r.t. the face perimeters of the embedding, *i.e.,* all elements of $\boldsymbol{\mathcal{F}}$. Attempts to label the graph's nodes on this basis, however, may lead to contradictions along cycles that do not lie in the cycle space of $\boldsymbol{\mathcal{F}}$, such as those shown in Figure 2.

More formally, let $\boldsymbol{\mathcal{X}} \subseteq 2^{\mathcal{E}}$ denote the set of *extended states* of our Ising model $G(\mathcal{V}, \mathcal{E})$, *i.e.,* all edge sets consistent with the faces $\boldsymbol{\mathcal{F}}$ of its embedding:

$$\boldsymbol{\mathcal{X}} \;:=\; \{\mathcal{X} \subseteq \mathcal{E} : (\forall \mathcal{F} \in \boldsymbol{\mathcal{F}})\ \mathcal{X} \mathrel{\hat{=}} \mathcal{F}\} . \tag{3}$$

The set $\boldsymbol{\mathcal{C}} \subseteq 2^{\mathcal{E}}$ of *cuts* of G is defined analogously, replacing $\boldsymbol{\mathcal{F}}$ in (3) with any cycle basis of G. Cuts are by definition consistent with all cycles of G, so $\boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{X}}$. $\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{X}}$ iff $\boldsymbol{\mathcal{F}}$ is a cycle basis of G, *i.e.,* comprises the faces of a plane (that is, genus zero) embedding of G.

The set of extended ground states is given by

$$\boldsymbol{\mathcal{X}}^*(\boldsymbol{\theta}) \;:=\; \operatorname*{argmin}_{\mathcal{X} \in \boldsymbol{\mathcal{X}}} E_{\mathcal{X}}(\boldsymbol{\theta}) , \tag{4}$$

where $E_{\mathcal{S}}(\boldsymbol{\theta}) := \sum_{(i,j) \in \mathcal{S}} \theta_{ij}$ for any edge set $\mathcal{S} \subseteq \mathcal{E}$, and argmin is deemed to return the location of *all* minima. Note that the method of Section 2.1 only returns one (arbitrary) extended ground state $\mathcal{X}^* \in \boldsymbol{\mathcal{X}}^*(\boldsymbol{\theta})$.

The set $\boldsymbol{\mathcal{C}}^*(\boldsymbol{\theta})$ of minimum-cost cuts is defined analogously by replacing $\boldsymbol{\mathcal{X}}$ in (4) with $\boldsymbol{\mathcal{C}}$. There is a 1:2 correspondence between cuts and induced node states in Ising models, which turns into a bijection when the label of the field node is held fixed (Schraudolph and Kamenetsky, 2008). In either case, $\boldsymbol{\mathcal{C}}^*(\boldsymbol{\theta})$ exactly induces the ground states of the Ising model. Since $\boldsymbol{\mathcal{C}} \subseteq \boldsymbol{\mathcal{X}}$, we know that $E_{\mathcal{X}^*}(\boldsymbol{\theta}) \leq E_{\mathcal{C}^*}(\boldsymbol{\theta})$, *i.e.,* the extended ground state energy is a lower bound on the energy of true ground states.

3 REWEIGHTING

Reweighting approaches the problem of calculating the partition function (Wainwright et al., 2005a) *resp.* a MAP state (Wainwright et al., 2005b) of an intractable distribution via a convex mixture of tractable ones. We will now briefly introduce this technique, then show that it can also be used to leverage the efficient method we have to calculate extended ground states of nonplanar Ising models (Section 2.2) into a practical algorithm for obtaining true ground states.

3.1 CONVENTIONAL REWEIGHTING

Following Wainwright et al. (2005b), we replace our single Ising model $G(\mathcal{V}, \mathcal{E})$ with a collection $G_k(\mathcal{V}_k, \mathcal{E}_k)$ of such models ($k = 1, 2, \ldots$), where $(\forall k)\ \mathcal{V}_k \subseteq \mathcal{V}$ and $\mathcal{E}_k \subseteq \mathcal{E}$, *i.e.,* the G_k are all subgraphs of G. We also require that $\bigcup_k \mathcal{E}_k = \mathcal{E}$, *i.e.,* every edge occurs at least once in the collection. Each G_k has its own disagreement cost vector $\boldsymbol{\theta}_k$, subject to the constraints $(i,j) \notin \mathcal{E}_k \Rightarrow [\boldsymbol{\theta}_k]_{ij} = 0$ and $\sum_k \boldsymbol{\theta}_k = \boldsymbol{\theta}$.[3] We shall indicate by a subscript k whenever we refer to a property of G_k, as opposed to G — thus $\boldsymbol{\mathcal{F}}_k$ refers to the face perimeters of G_k's embedding, $\boldsymbol{y}_k^*(\boldsymbol{\theta}_k)$ to one of its ground states, and so forth.

By its definition, $E_{\mathcal{C}^*}(\boldsymbol{\theta})$ is a pointwise minimum of linear functions, and hence concave in $\boldsymbol{\theta}$. We can thus invoke Jensen's inequality to obtain the lower bound

$$\sum_k E_{\mathcal{C}_k^*}(\boldsymbol{\theta}_k) \;\leq\; E_{\mathcal{C}^*}(\boldsymbol{\theta}) \tag{5}$$

on the ground state energy. The key idea now is to make this bound as tight as possible by maximizing the left-hand side of (5) w.r.t. the $\boldsymbol{\theta}_k$.

Wainwright et al. (2005b) proved that (5) becomes tight iff there is *agreement* between the ground states of all G_k, *i.e.,* there are ground states $\boldsymbol{y}_k^*$ that agree wherever they overlap:

$$(\forall k_1, k_2)\;\; v \in \mathcal{V}_{k_1} \cap \mathcal{V}_{k_2} \;\Rightarrow\; [\boldsymbol{y}_{k_1}^*]_v = [\boldsymbol{y}_{k_2}^*]_v\,. \tag{6}$$

A ground state of G can then be obtained by simply combining the agreeing ground states $\boldsymbol{y}_k^*$ of the G_k.

This strategy of course only makes sense if the G_k are computationally more tractable than G. Thus Wainwright et al. (2005a,b) decompose a graph with cycles into a collection of spanning trees; Globerson and Jaakkola (2007) approximate the partition function of a nonplanar Ising model by decomposing it into a collection of spanning planar graphs. By contrast, we will now introduce a form of reweighting that does not rely on decomposition into simpler subgraphs.

3.2 A NEW FORM OF REWEIGHTING

When reweighting is applied to *extended* ground states, we obtain a lower bound analogous to (5) which we can maximize. Since we no longer have node states, the agreement condition (6) must be reformulated in terms of edges:

$$\begin{aligned} &\forall k\; \exists \mathcal{X}_k^* \in \boldsymbol{\mathcal{X}}_k^*(\boldsymbol{\theta}_k) : (\forall k_1, k_2) \\ &(i,j) \in \mathcal{E}_{k_1} \cap \mathcal{E}_{k_2} \;\Rightarrow\; [(i,j) \in \mathcal{X}_{k_1}^* \Leftrightarrow (i,j) \in \mathcal{X}_{k_2}^*]\,. \end{aligned} \tag{7}$$

When agreement holds, the resulting extended ground state $\mathcal{X}^* := \bigcup_k \mathcal{X}_k^*$ will be consistent with the cycle space of $\bigcup_k \boldsymbol{\mathcal{F}}_k$, *i.e.,* all cycles that can be composed from the face perimeters of the entire collection. This offers an exciting prospect: if we design our collection of G_k such that $\bigcup_k \boldsymbol{\mathcal{F}}_k$ is a cycle basis of G, then (7) implies an extended ground state $\mathcal{X}^*$ that is consistent with *every* cycle of G — in other words: a true ground state. We call such a collection of G_k *consistent.*

[3] For simplicity, we omit the weighting coefficients ρ used by Wainwright et al. (2005b).

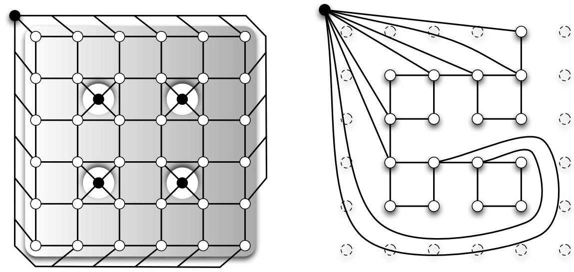


Figure 3: Left: Regular embedding of a 6×6 grid with external field. Connections to the field node (black disks) from the interior of the grid are routed through four holes in the embedding surface (shaded). Right: Plane embedding of a homology basis for the embedding on the left. Together, the faces of these two embeddings constitute a cycle basis for the full graph.

It is not difficult to find a cycle basis for a given embedded graph: Augmenting the faces $\boldsymbol{\mathcal{F}}$ of the embedding with a *homology basis* $\boldsymbol{\mathcal{H}}$ for instance creates a cycle basis $\boldsymbol{\mathcal{F}} \cup \boldsymbol{\mathcal{H}}$. Loosely speaking, a homology basis for an orientable surface of genus g comprises $2g$ cycles that thread *resp.* encircle each hole in the surface; Figure 2 gives a homology basis for the torus ($g = 1$), Figure 3 (right) one for the embedding ($g = 4$) shown in Figure 3 (left). A homology basis for an embedded graph $G(\mathcal{V}, \mathcal{E})$ can be computed in $O(|\mathcal{E}| + |\mathcal{V}| \log |\mathcal{V}|)$ time (Erickson and Whittlesey, 2005).

A consistent collection of G_k can thus be constructed recursively as follows:

1. $G_1 := G$, $k := 1$
2. find embedding of G_k
3. find homology basis $\boldsymbol{\mathcal{H}}_k$ of G_k
4. IF $\boldsymbol{\mathcal{H}}_k = \epsilon$ THEN EXIT
5. build G_{k+1} : $\mathcal{E}_{k+1} := \cup \boldsymbol{\mathcal{H}}_k$
6. $k := k + 1$; GOTO 2

That is, in the k^{th} iteration we add the (homology basis of the)^k graph G to our collection, unless and until $\boldsymbol{\mathcal{H}}_k$ is empty, *i.e.,* G_k is plane embedded.

If G is planar to begin with, the resulting collection will only contain G itself, and reweighting reduces to the algorithm of Section 2.1 for planar Ising models. If G derives from a planar MRF (*i.e.,* without the field node it would be planar) then it has a planar homology basis, and the collection will consist of that and G. In general, the minimum size $\varrho(G)$ of a consistent collection for G is an interesting and to our knowledge new graph invariant: while $\varrho(G) = 1$ comprises just the planar graphs, for $k > 1$ the class $G : \varrho(G) = k$ contains graphs of unbounded genus that are nonetheless, in a deep and precise sense, topologically simpler than

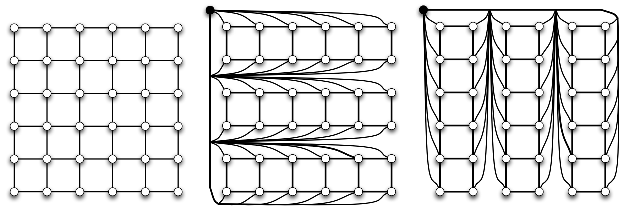


Figure 4: The three plane subgraphs with which we augment the nonplanar grid of Figure 3 (left): the full grid without external field (left), and horizontal (center) *resp.* vertical (right) strips with external field.

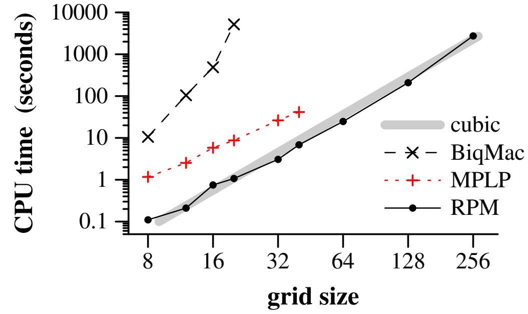


Figure 5: CPU time taken by our implementation (RPM) *resp.* MPLP (Sontag et al., 2008) on a MacBook laptop (2.2 GHz Intel Core 2 Duo) and a state-of-the-art SDP-based branch & bound algorithm (Rendl et al., 2010) on the BiqMac server (3.0 GHz Intel Xeon 5160) to compute a ground state of an $n \times n$ square lattice Ising model with random coupling strength and external field, against grid size n.

graphs of the class $G : \varrho(G) = k + 1$. Note that determining a ground state has been shown to be NP-hard already for graphs with $\varrho(G) = 2$ (Barahona, 1982).

In practice the above recursive method for constructing a consistent collection may be too parsimonious, in that we find it difficult to achieve agreement (7) with it. This is a key problem with conventional reweighting as well. It is often found that enriching the collection (*i.e.,* using more and larger subgraphs than strictly necessary) improves the probility of reaching agreement. Unlike conventional reweighting, in doing so we are not limited by tractability concerns: we can efficiently compute extended ground states for any graph. Thus there is nothing to stop us from employing, say, a collection comprising many copies of the entire graph G, diversely embedded so that $\bigcup_k \boldsymbol{\mathcal{F}}_k$ is a cycle basis.

To summarize, the key difference here to conventional reweighting is that instead of making the computation tractable by restricting the modeled distribution to a subgraph, we do so by relaxing the ground state to the extended ground state. In both cases, the missing constraints are enforced by requiring agreement across the collection.

4 EMPIRICAL RESULTS

Noting that the lower bound (5) on extended ground states is nonsmooth, concave, and piecewise linear, we use a Fortran implementation of the BT (bundle trust) optimizer (Schramm and Zowe, 1992) to maximize it. Since we always use $G_1 := G$, we can set

$$\boldsymbol{\theta}_1 \;:=\; \boldsymbol{\theta} - \sum_{k=2,3,\ldots} \boldsymbol{\theta}_k \tag{8}$$

to obtain an unconstrained optimization problem over the disagreement costs of edges in $\mathcal{E}_2, \mathcal{E}_3, \ldots$ A subgradient w.r.t. $[\boldsymbol{\theta}_k]_{ij}$ (needed by BT) is then given by

$$[\![(i,j) \in \mathcal{X}_k^*]\!] \;-\; [\![(i,j) \in \mathcal{X}_1^*]\!]\,. \tag{9}$$

The Blossom V C++ code (Kolmogorov, 2009) is used to incrementally compute maximum-weight perfect matchings in the inner loop of the optimization.

We choose binary MRFs over rectangular grids as our benchmark problem, which translate into rectangular Ising lattices with external field. Many computer vision problems have been cast into this form. Barahona (1982) has proven that determining a ground state, or its energy, in such an Ising model is NP-hard, even when disagreement costs are restricted to $\{-1, 0, +1\}$. There is also an applicable proof of polynomial-time non-approximability (Bertoni et al., 1997).

An $n \times m$ rectangular grid with external field has a regular (not necessarily optimal) embedding of genus $\lceil n/2-1 \rceil \lceil m/2-1 \rceil$; the case $n = m = 6$ is shown in Figure 3 (left). We employ square grids ($n = m$) of varying size with uniformly random disagreement costs $\boldsymbol{\theta}$. For reweighting, we have found it advantageous to augment the regular embedding with the 3 planar subgraphs illustrated in Figure 4; the 4 graphs form a consistent collection, so we obtain a true ground state.

Figure 5 shows the CPU time taken by our reweighted perfect matching (RPM) method to calculate a ground state for grid sizes ranging from $n = 8$ to $n = 256$. We find cubic scaling (broad gray line) over the entire range of grid sizes, from 0.1 CPU seconds for an 8×8 grid to about 45 CPU minutes for 256×256.

For comparison, we submitted the same problems to the BiqMac solver (`http://biqmac.uni-klu.ac.at/`), which runs a state-of-the-art branch & bound algorithm based on a semidefinite programming (SDP) relaxation (Rendl et al., 2010). As Figure 5 (dashed line) shows, this took 2–4 orders of magnitude longer than our method and exhibits worse scaling with grid size. For grid sizes up to 20×20, where BiqMac was able to compute an answer within the 3 CPU hour max-

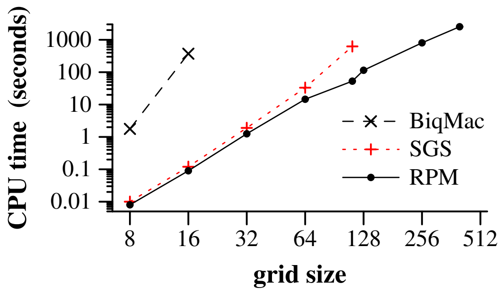


Figure 6: As Figure 5, but also comparing our RPM against the method of Liers et al. (2004) at the Spin Glass Server (SGS, 1.86 GHz Intel Celeron M) on $n \times n$ toroidal lattices with random coupling strength but no external field. MPLP did not work here.

imum time available on the server, we verified that both methods return the same result. Rendl et al. (2010) found their algorithm to be faster than all other generic methods they considered; thus these results suggest that our novel form of reweighting constitutes a substantial advance in combinatorial optimization.

We also compared performance to MPLP (Sontag et al., 2008), which represents the current state of the art in message-passing algorithms. MPLP is an order of magnitude slower than RPM on the 8×8 grid but scales even better than RPM, so that at $n = 40$ it only takes 6 times as long as our method. Whenever it converged it returned the same result as the other methods; however, we could not get it to converge for grids any larger than 40×40, presumably due to numerical limitations.

For grid sizes above $n = 100$, our implementation of RPM sometimes fails to converge; above $n = 200$ these failures become frequent. We suspect numerical problems: reweighting tends to create degeneracy in the collection and may well exceed the capabilities of standard IEEE-754 floating-point arithmetic for problems of such size. While Blossom V is templated C++ code, the Fortran code of the BT optimizer unfortunately chains us to IEEE-754 at present. In the following two sections we describe improvements that should not only further increase the speed of RPM, but also facilitate its implementation in arbitrary-precision integer arithmetic. This will allow us to probe the limits of the observed polynomial-time regime, which we must encounter at some point — unless P=NP.

Rendl et al. (2010) found the method of Liers et al. (2004), which is specialized to 2D and 3D lattices with periodic boundary conditions but no external field, to be faster than their algorithm on such grids. We therefore performed a direct comparison of our approach with that method (running on the Spin Glass Server, `http://www.informatik.uni-koeln.de/ls_juenger/research/spinglass/`) on 2D toroidal grids. For reweighting we augmented the full graph ($g = 1$) with two planar subgraphs obtained by deleting a row of vertical *resp.* column of horizontal edges. Note that this problem is *not* NP-hard.

Figure 6 shows that while both methods perform similarly on smaller grids, our approach is more than twice as fast at 64×64, and almost twelve times as fast for $n = 112$, the largest grid size for which the Spin Glass Server returned an en exact answer. Our code scales well beyond that (up to 400×400) before we run into numerical problems. Again we verified that all methods produce identical results. Given that we are comparing our first implementation (with known inefficiencies) of a new, generic approach against an established, specialized solver here, we find these results very encouraging.

We tried but could not get MPLP to work on toroidal grids of any size. Like all message-passing algorithms, MPLP relies on the node potentials to break the initial symmetry; without such an external field it simply does not converge. We tried addressing this problem by promoting an arbitrary node in the toroidal grid to field node, but the resulting bias was too weak to guide MPLP to the ground state: it converged very slowly, and only to a local optimum.

5 SLACK SHARING

Maximizing a lower bound like (5) amounts to solving a linear program (LP). In each iteration of this process we call Blossom, which is itself an LP, to compute extended ground states. Running two nested loops of linear programs, unaware of each other, is clearly inefficient. We now show how to combine both levels into a single modified LP. This amounts to a single run of the Blossom algorithm, extended by a technique we call *slack sharing*, on the entire reweighting collection.

The Blossom algorithm (Edmonds, 1965a,b; Kolmogorov, 2009) solves the following LP very efficiently:

$$\underset{\boldsymbol{x} \geq \mathbf{0}}{\text{minimize}}\ \boldsymbol{\theta}^{\top}\boldsymbol{x} \quad \text{s.t.:} \quad |\{v\}|_{\boldsymbol{x}} = 1 \ \ \forall\, v \in \mathcal{V}, \tag{10}$$

$$|\mathcal{B}|_{\boldsymbol{x}} \geq 1 \ \ \forall\, \mathcal{B} \in \boldsymbol{\mathcal{B}}, \tag{11}$$

where $\boldsymbol{\mathcal{B}}$ contains all subsets of $\mathcal{V}$ with an odd number of nodes greater than one, and for any subset of nodes, $|\cdot|_{\boldsymbol{x}}$ sums the values of $\boldsymbol{x}$ along the boundary:

$$(\forall\, \mathcal{S} \subseteq \mathcal{V}) \qquad |\mathcal{S}|_{\boldsymbol{x}} := \sum_{\substack{(i,j)\in\mathcal{E}:\\ i\in\mathcal{S}, j\notin\mathcal{S}}} x_{ij}\,. \tag{12}$$

At the solution the primal vector $\boldsymbol{x}$ becomes binary, and indicates the minimum-weight perfect matching:

$x^*_{ij} = [\![(i,j) \in \mathcal{M}^*(\boldsymbol{\theta})]\!]$. The dual program is

$$\underset{\boldsymbol{y}}{\text{maximize}} \quad \sum_{v\in\mathcal{V}} y_v + \sum_{\mathcal{B}\in\boldsymbol{\mathcal{B}}} y_{\mathcal{B}} \tag{13}$$

$$\text{subject to:} \quad s_{ij} \geq 0 \quad \forall (i,j) \in \mathcal{E}, \tag{14}$$

$$y_{\mathcal{B}} \geq 0 \quad \forall \mathcal{B} \in \boldsymbol{\mathcal{B}}, \tag{15}$$

where $\boldsymbol{s}$ is a vector of *slacks*, *i.e.,* reduced edge costs:

$$s_{ij} \;:=\; \theta_{ij} - y_i - y_j - \sum_{\substack{\mathcal{B}\in\boldsymbol{\mathcal{B}}:\\ i\in\mathcal{B}, j\notin\mathcal{B}}} y_{\mathcal{B}}. \tag{16}$$

When we integrate reweighting into Blossom's LP, the primal optimization becomes

$$\underset{\boldsymbol{\theta}_k}{\text{maximize}} \;\; \underset{\boldsymbol{x}_k \geq \boldsymbol{0}}{\text{minimize}} \;\; \boldsymbol{\theta}_k^\top \boldsymbol{x}_k \tag{17}$$

subject to the additional constraint $\sum_k \boldsymbol{\theta}_k = \boldsymbol{\theta}$. Surprisingly, the dual program remains unchanged, except that (14) becomes

$$\sum_{k:(i,j)\in\mathcal{E}_k} [\boldsymbol{s}_k]_{ij} \;\geq\; 0 \quad \forall (i,j) \in \mathcal{E}, \tag{18}$$

i.e., corresponding edges across the collection *share* their slack. To integrate reweighting into Blossom, we thus merely have to implement a mechanism for slack sharing. Blossom is quite an intricate algorithm (see Kolmogorov, 2009), so this is by no means a trivial task. However, given that edges which share slacks always reside in distinct connected components (there are no edges *between* graphs in a reweighting collection), we foresee no undue complications here.

6 DEALING WITH DEGENERACY

For degenerate Ising models in our collection, Blossom only returns one of their multiple extended ground states. Because these samples from the sets $\boldsymbol{\mathcal{X}}^*_k(\boldsymbol{\theta}_k)$ are drawn independently, they may fail to agree even when (7) holds. (We can observe this empirically by drawing edge costs $\boldsymbol{\theta}$ randomly from $\{-1, 1\}$, which produces highly degenerate Ising models.) At present we have no way to distinguish this state of affairs from a genuine failure to reach agreement (7). Launching a search over the sets of extended ground states for samples that are in agreement with each other looks unpromising: the problem appears quite similar to weighted set cover, which is NP-hard.

Instead we deal with degeneracy as follows: Assume that the edge costs $\boldsymbol{\theta}$ of a degenerate Ising model are integers.[4] Now add to $\boldsymbol{\theta}$ the fractional vector $\boldsymbol{\rho} : \rho_i = 2^{-i}, i = 1, 2, \ldots |\mathcal{E}|$. Since $\boldsymbol{\rho}$ gives each extended state its unique fractional energy level which is not affected by $\boldsymbol{\theta}$, the Ising model with edge costs $\boldsymbol{\theta}+\boldsymbol{\rho}$ is non-degenerate. Since $\boldsymbol{\rho}$ raises the energy of any given state by at most $\sum_{i=1}^{|\mathcal{E}|} 2^{-i} < 1$, the unique extended ground state for $\boldsymbol{\theta}+\boldsymbol{\rho}$ is one of the extended ground states of the original, degenerate model with integer costs.

This trick requires us to lengthen the representation of our integer costs by $|\mathcal{E}|$ bits. Since Blossom only performs linear operations on the cost vector, however, this increases the complexity of our algorithm only by a factor of $|\mathcal{E}|/l$, where $l \geq \lceil \log_2(\theta_{\max} - \theta_{\min}) \rceil$ is the length of integer representation needed for the original, degenerate problem.

In practice we can do better still by first solving the degenerate problem, then adding fractional energies only to those edges that do not agree at the optimum (if any). We then incrementally re-optimize the modified problem, and repeat as necessary until we have either reached agreement or added fractional energies to all edges without doing so.

[4] Given the finite precision of floating-point computer arithmetic, this can be assumed without loss of generality.

7 DISCUSSION

Above we have given a polynomial-time algorithm for an NP-hard and non-approximable problem. The catch is that a given collection may not reach agreement at the optimum, and thus fail to provide a useful answer. Empirically we find that expanding the collection helps in such an event; unless P=NP the required input expansion must be exponential in the worst case.

However, that worst case appears to be far more elusive than one would think: In our experiments RPM is able to reliably solve larger MAP problems than any other method, and much faster, even though we only ever expand the input graph by a fixed factor of < 4.

At present our reweighting collections are hand-crafted *a priori*. Better understanding of the cause(s) of disagreement at the optimum should yield heuristics to automate and optimize the collection design process.

The Ising ground state problem is equivalent to MAX CUT, which has very direct transformations from many other NP-hard problems. Our work thus has implications beyond machine learning and statistical physics.

An implementation of RPM will be available as part of our isinf code at `http://nic.schraudolph.org/isinf/`.

References

F. Barahona. On the computational complexity of Ising spin glass models. *Journal of Physics A: Mathematical, Nuclear and General*, 15(10):3241–3253, 1982.

A. Bertoni, P. Campadelli, G. Gangai, and R. Posenato.

Approximability of the ground state problem for Ising spin glasses. *Journal of Complexity*, 13:326–339, 1997.

I. Bieche, R. Maynard, R. Rammal, and J. P. Uhry. On the ground states of the frustration model of a spin glass by a matching method of graph theory. *Journal of Physics A: Mathematical and General*, 13:2553–2576, 1980.

J. M. Boyer and W. J. Myrvold. On the cutting edge: Simplified $O(n)$ planarity by edge addition. *Journal of Graph Algorithms and Applications*, 8(3):241–273, 2004. Reference implementation (C source code): http://jgaa.info/accepted/2004/BoyerMyrvold2004.8.3/planarity.zip.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11): 1222–1239, 2001.

J. Chen, S. P. Kanchi, and A. Kanevsky. A note on approximating graph genus. *Information Processing Letters*, 61 (6):317–322, 1997.

J. Edmonds. Maximum matching and a polyhedron with 0,1-vertices. *Journal of Research of the National Bureau of Standards*, 69B:125–130, 1965a.

J. Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965b.

J. Erickson and K. Whittlesey. Greedy optimal homotopy and homology generators. In *Proc. 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1038–1046, Vancouver, BC, 2005. SIAM. ISBN 0-89871-585-7.

M. E. Fisher. Statistical mechanics of dimers on a plane lattice. *Physical Review*, 124(6):1664–1672, 1961.

B. J. Frey and D. J. C. MacKay. A revolution: Belief propagation in graphs with cycles. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, volume 10, pages 479–485. MIT Press, 1998.

H. N. Gabow. Data structures for weighted matching and nearest common ancestors with linking. In *Proc. 1st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 434–443, San Francisco, CA, 1990. SIAM. ISBN 0-89871-251-3.

A. Galluccio and M. Loebl. On the theory of Pfaffian orientations. i. perfect matchings and permanents. *Journal of Combinatorics*, 6:81–100, 1999.

A. Galluccio, M. Loebl, and J. Vondrák. New algorithm for the Ising problem: Partition function for finite lattice graphs. *Phys Rev Letters*, 84(26):5924–5927, June 2000.

A. Globerson and T. Jaakkola. Approximate inference using planar graph decomposition. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, pages 473–480, Cambridge, MA, 2007. MIT Press.

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

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

P. W. Kasteleyn. The statistics of dimers on a lattice: I. the number of dimer arrangements on a quadratic lattice. *Physica*, 27(12):1209–1225, 1961.

V. Kolmogorov. Blossom V: A new implementation of a minimum cost perfect matching algorithm. *Mathematical Programming Computation*, 1 (1):43–67, 2009. http://www.cs.ucl.ac.uk/staff/V.Kolmogorov/papers/BLOSSOM5.html.

V. Kolmogorov and C. Rother. Minimizing nonsubmodular functions with graph cuts — a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(7):1274–1279, July 2007.

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

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, B 50(2):157–224, 1988.

F. Liers, M. Jünger, G. Reinelt, and G. Rinaldi. Computing exact ground states of hard Ising spin glass problems by branch-and-cut. In A. Hartmann and H. Rieger, editors, *New Optimization Algorithms in Physics*, pages 47–68. Wiley, 2004.

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

F. Rendl, G. Rinaldi, and A. Wiegele. Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. *Mathematical Programming*, 121(2): 307, 2010.

C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality. In *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, June 2007.

H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM J. Optimization*, 2:121–152, 1992.

N. N. Schraudolph and D. Kamenetsky. Efficient exact inference in planar Ising models. Tech. Rep. arXiv:0810.4401, Oct. 2008. http://aps.arxiv.org/abs/0810.4401

N. N. Schraudolph and D. Kamenetsky. Efficient exact inference in planar Ising models. In D. Koller, Y. Bengio, D. Schuurmans, L. Bottou, and A. Culotta, editors, *Advances in Neural Information Processing Systems 21*, Cambridge, MA, 2009. MIT Press.

D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In D. A. McAllester and P. Myllymäki, editors, *Proc. 24th Conf. Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2008.

C. K. Thomas and A. A. Middleton. Matching Kasteleyn cities for spin glass ground states. *Physical Review B*, 76(22):22406, 2007.

C. Thomassen. The graph genus problem is NP-complete. *Journal of Algorithms*, 10(4):568–576, 1989.

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

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. MAP estimation via agreement on trees: Message-passing and linear programming. *IEEE Transactions on Information Theory*, 51(11):3697–3717, 2005b.

Y. Weiss. Belief propagation and revision in networks with loops. AI memo 1616, CBCL paper 155, MIT, Department of Brain and Cognitive Sciences, November 1997.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 689–695. MIT Press, 2001.