Quadratic Programming Relaxations for Metric Labeling and Markov Random Field MAP Estimation

Pradeep Ravikumar PRADEEPR@CS.CMU.EDU
John Lafferty LAFFERTY@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

Quadratic program relaxations are proposed as an alternative to linear program relaxations and tree reweighted belief propagation for the metric labeling or MAP estimation problem. An additional convex relaxation of the quadratic approximation is shown to have additive approximation guarantees that apply even when the graph weights have mixed sign or do not come from a metric. The approximations are extended in a manner that allows tight variational relaxations of the MAP problem, although they generally involve non-convex optimization. Experiments carried out on synthetic data show that the quadratic approximations can be more accurate and computationally efficient than the linear programming and propagation based alternatives.

1. Introduction

Undirected graphical models, or Markov random fields (MRFs), are natural tools in many domains, from image processing to social network modeling. A key inference problem for MRFs is to compute the maximum a posteriori (MAP) configuration—the most probable labeling—which is used in multiple applications such as image denoising, protein folding and error control coding. For arbitrary graphs and parameter settings this problem is NP-hard, but various approximate techniques have been proposed that have enabled the application of MRFs to a range of practical problems.

For tree-structured distributions, the MAP estimate for random fields can be computed efficiently by dynamic programming. It can also be computed in polynomial time using graph cuts (Greig et al., 1989) when the parameter settings yield a submodular energy function. In the general setting, a widely used approximation technique is max-product belief propagation (Pearl, 1988). The algorithm is convergent on trees, and its fixed point configuration upon convergence can be shown to be locally optimal with respect to a large set of moves (Weiss & Freeman, 2001). A similar message passing algorithm, tree-reweighted max product (Wainwright et al., 2005), has stronger correctness and convergence guarantees. Boykov et al. (2001) have proposed graph-cut based algorithms that efficiently find a local energy minimum with respect to two types of large moves. A different direction has been taken in recent work on linear program relaxations for the MAP problem in the specific setting of metric labeling. In the metric labeling formulation, a weighted graph and a metric on labels specifies the energy or cost of different labelings of a set of objects, and the goal is to find a minimum cost labeling. Casting this as an integer linear program, Kleinberg and Tardos (1999) proposed linear relaxations for specific metrics. Chekuri et al. (2005) recently extended these techniques using the natural linear relaxation of the metric labeling task, and obtained stronger approximation guarantees.

In this paper, we propose a quadratic programming (QP) relaxation to the MAP or metric labeling problem. While the linear relaxations have $O(|E|k^2)$ variables, where $|E|$ is the number of edges in the graph and k is the number of labels, in our QP formulation there are kn variables, and yet we show that the quadratic objective function more accurately represents the energy in the graphical model. In particular, we show that the QP formulation computes the MAP solution exactly. Under certain conditions the relaxation results in a non-convex problem however, which requires an intractable search over local minima. This motivates an additional convex approximation to the relaxation, which we show satisfies an additive approximation guarantee. We also extend the relaxation to general variational "inner polytope" relaxations which we also show to compute the MAP exactly. Experiments indicate that our quadratic relaxation with the convex approximation outperforms or is comparable to existing methods under many

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006.

settings.

In the following section, we establish some notation and recall the relevant background. In Section 2.1 we review linear relaxations for MAP estimation. In Section 3, we describe the quadratic relaxation, prove that it is tight, and detail its convex approximation. In Section 4, we then extend the above relaxation to show the tightness of various variational inner polytope relaxations. Finally, in Section 5 and Section 6, we present our experimental results and conclusions.

2. Notation and Background

Consider a graph $G = (V, E)$, where V denotes the set of nodes and E denotes the set of edges. Let X_s be a random variable associated with node s, for $s \in V$, yielding a random vector $X = \{X_1, \ldots, X_n\}$, and let $\phi = \{\phi_\alpha, \alpha \in I\}$ denote the set of potential functions (or sufficient statistics) for a set I of cliques in G. Associated with ϕ is a vector of parameters $\theta = \{\theta_\alpha, \alpha \in I\}$. With this notation, the exponential family of distributions of X, associated with ϕ and G is given by

$$p(x;\theta) \; = \; \exp\left(\sum_\alpha \theta_\alpha \phi_\alpha - \Psi(\theta)\right).$$

As discussed in (Yedidia et al., 2001), at the expense of increasing the state space one can assume without loss of generality that the graphical model is a pairwise Markov random field, *i.e.*, the set of cliques I is the set of edges $\{(s,t) \in E\}$, so that

$$p(x;\theta) \propto \exp\left(\sum_{s \in V} \theta_s \phi_s(x_s) + \sum_{(s,t)\in E} \theta_{st}\phi_{st}(x_s, x_t)\right).$$

If each X_s takes values in a discrete set $\mathcal{X}_s$, we can represent any potential function as a linear combination of indicator functions, $\phi_s(x_s) = \sum_j \phi_s(j)\, \mathcal{I}_j(x_s)$ and $\phi_{st}(x_s, x_t) = \sum_{jk} \phi_{st}(j,k)\, \mathcal{I}_{j,k}(x_s, x_t)$ where

$$\mathcal{I}_j(x_s) \quad = \quad \begin{cases} 1 & x_s = j \\ 0 & \text{otherwise} \end{cases}$$

and

$$\mathcal{I}_{j,k}(x_s, x_t) \quad = \quad \begin{cases} 1 & x_s = j \text{ and } x_t = k \\ 0 & \text{otherwise}. \end{cases}$$

We thus consider pairwise MRFs with indicator potential functions as

$$p(x|\theta) \propto \exp\left(\sum_{s,j} \theta_{s;j}\mathcal{I}_j(x_s) + \sum_{s,t;j,k} \theta_{s,j;t,k}\mathcal{I}_{j,k}(x_s, x_t)\right).$$

The MAP problem is then given by

$$x^* = \underset{x}{\operatorname{argmax}} \sum_{s,j} \theta_{s;j}\mathcal{I}_j(x_s) + \sum_{s,t;j,k} \theta_{s,j;t,k}\mathcal{I}_{j,k}(x_s, x_t). \tag{1}$$

2.1. Linear Relaxations

MAP estimation in the discrete case is essentially a combinatorial optimization problem, and it can be cast as an integer program. Recent work has studied approximate MAP estimation using linear relaxations (Bertsimas & Tsitsiklis, 1997). Letting variables $\mu(s;j)$ and $\mu(s,j;t,k)$ correspond to the indicator variables $\mathcal{I}_j(x_s)$ and $\mathcal{I}_{j,k}(x_s, x_t)$, we obtain the following integer linear program (ILP),

$$\begin{aligned} \max \quad & \sum_{s;j} \theta_{s;j}\, \mu_1(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu_2(s,j;t,k) \\ \text{such that} \quad & \sum_k \mu_2(s,j;t,k) = \mu_1(s;j) \\ & \sum_j \mu_1(s;j) = 1 \\ & \mu_1(s;j) \in \{0,1\} \\ & \mu_2(s,j;t,k) \in \{0,1\}. \end{aligned}$$

This ILP can then be relaxed to the following linear program (LP),

$$\begin{aligned} \max \quad & \sum_{s;j} \theta_{s;j}\, \mu_1(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu_2(s,j;t,k) \qquad (2) \\ \text{such that} \quad & \sum_k \mu_2(s,j;t,k) = \mu_1(s;j) \\ & \sum_j \mu_1(s;j) = 1 \\ & 0 \le \mu_1(s;j) \le 1 \\ & 0 \le \mu_2(s,j;t,k) \le 1. \end{aligned}$$

Chekuri et al. (2005) propose the above LP relaxation as an approximation algorithm for the metric labeling task, which is the MAP problem with spatially homogeneous MRF parameters; thus, $\theta_{s,j;t,k} = w_{st}\, d(j,k)$, where w_{st} is a non-negative edge weight and d is a metric that is the same for all the edges. Kleinberg and Tardos (1999) proposed related linear relaxations for specific metrics. The above LP relaxation was also proposed for the general pairwise graphical model setting by Wainwright and Jordan (2003). Letting θ and $\phi(x)$ denote the vectors of parameters and potential functions, respectively, and letting $\langle \theta, \phi(x) \rangle$ denote the inner product

$$\langle \theta, \phi(x) \rangle = \sum_{s;j} \theta_{s;j}\mathcal{I}_j(x_s) + \sum_{(s,t)\in E;\, j,k} \theta_{s,j;t,k}\mathcal{I}_{j,k}(x_s, x_t)$$

the MAP problem is then given by

$$x^* = \underset{x}{\operatorname{argmax}} \langle \theta, \phi(x) \rangle = \sup_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle$$

where $\mathcal{M}$ is the set of moment parameters

$$\mathcal{M} = \left\{ \mu \,:\, \sum_x p(x)\phi(x) = \mu \text{ for some distribution } p \right\}.$$

The polytope $\mathcal{M}$ can be seen to be upper-bounded by the set LOCAL(G) of all single and pairwise vectors μ_1 and μ_2 that satisfy the local consistency constraints

$$\begin{aligned} \textstyle\sum_k \mu_2(s,j;t,k) &= \mu_1(s;j) \\ \textstyle\sum_j \mu_1(s,j) &= 1 \\ 0 \le \mu_1(s;j) &\le 1 \\ 0 \le \mu_2(s,j;t,k) &\le 1. \end{aligned}$$

Wainwright and Jordan (2003) thus proposed the upper-bounding relaxation of using $LOCAL(G)$ as an outer bound for the polytope $\mathcal{M}$,

$$\mu^* = \sup_{\mu \in LOCAL(G)} \langle \theta, \mu \rangle, \tag{3}$$

which is the same LP formulation as in equation (2). Furthermore, Wainwright et al. (2005) show that under certain conditions, the tree-reweighted belief propagation updates solve the dual of the LP in equation (3); since strong duality holds, the tree updates also give the optimal primal value for the LP.

3. Quadratic Relaxation

In the linear relaxation of equation (2), the variables $\mu_2(s,j;t,k)$ are relaxations of the indicator variables $\mathcal{I}_{j,k}(x_s,x_t)$, with a value of one indicating that for edge $(s,t) \in E$, variable x_s is labeled j and variable x_t is labeled k. These pairwise variables are constrained by demanding that they be consistent with the corresponding "marginal" variables $\mu_1(s,j)$. Note, however, that the binary indicator variables satisfy the additional "independence" constraint

$$\mathcal{I}_{j,k}(x_s,x_t) = \mathcal{I}_j(x_s)\,\mathcal{I}_k(x_t).$$

This then suggests that constraining the relaxation variables in a similar manner, $\mu_2(s,j;t,k) = \mu_1(s;j)\,\mu(t;k)$, might yield a tighter relaxation. This leads to the following quadratic program

$$\begin{aligned} \max \quad & \sum_{s;j} \theta_{s;j}\mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\,\mu(s;j)\,\mu(t;k) \\ \text{subject to} \quad & \sum_j \mu(s;j) = 1 \\ & 0 \le \mu(s;j) \le 1 \end{aligned} \tag{4}$$

The following result shows that the relaxation is in fact tight; the proof uses the probabilistic method.

Theorem 3.1. *The optimal value of problem* (4) *is equal to the optimal value of the MAP problem* (1).

Proof. Let the optimal MAP energy be

$$e_{\text{MAP}} = \max_x \sum_{s;j} \theta_{s;j}\mathcal{I}_j(x_s) + \sum_{s,t;j,k} \theta_{s,j;t,k}\,\mathcal{I}_{j,k}(x_s,x_t)$$

and let the optimal value of the relaxed problem be

$$e^* = \max_\mu \sum_{s;j} \theta_{s;j}\mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\,\mu(s;j)\,\mu(t;k)$$

where $\sum_j \mu(s;j) = 1$ and $\mu(s;j) \in [0,1]$. Clearly, $e^* \ge e_{\text{MAP}}$ since problem (4) is a relaxation of problem (1). We now show that $e_{\text{MAP}} \ge e^*$.

Let μ^* be an optimal solution of problem (4), and consider the following randomized rounding scheme. For each node s, assign it value j with probability $\mu(s;j)$. The expected energy of such a rounding is

$$\begin{aligned} e_R &= \sum_{s;j} \theta_{s;j}\,\mu(s;j)) + \sum_{s,j;t,k} \theta_{s,j;t,k}\,\mu(s;j)\,\mu(t;k) \\ &= e^* \end{aligned}$$

But there has to exist some discrete assignment y whose energy is greater than the expected energy; $e(y) \ge e^*$. Since e_{MAP} is the energy of the optimal configuration $e_{\text{MAP}} \ge e(y)$, which thus gives $e_{\text{MAP}} \ge e^*$. □

Note that the randomization in the proof just shows the existence of a discrete solution with the same energy as that of the optimal real relaxation. The problem of obtaining such a discrete solution efficiently is considered next.

Theorem 3.2. *Any solution of the MAP problem* (1) *efficiently yields a solution of the relaxation* (4) *and vice versa. Thus the relaxation* (4) *is equivalent to the MAP problem* (1).

Proof. From theorem 3.1, the optimal values of problems (1) and (4) are equal; let e^* denote this maximum energy. Let $\hat{x}$ be an optimal solution of the MAP problem (1). As problem (4) is a relaxation of the MAP problem, $\mu(s;j) = \mathcal{I}(\hat{x};j)$ is also a feasible and optimal solution for (4).

For the converse, let μ^* be an optimal solution of problem (4). Its energy is given by

$$e^* = \sum_{s;j} \theta_{s;j}\mu^*(s;j) + \sum_{(s,t)\in E;j,k} \theta_{s,j;t,k}\mu^*(s;j)\mu^*(t;k) \tag{5}$$

If each $\mu^*(s;j)$ is integer valued, that is, in $\{0,1\}$, then we can use μ^* itself as the feasible optimal solution for the MAP problem (1). Otherwise, consider μ^* to be real valued; we (efficiently) construct a labeling y with the maximum energy e^*.

Consider an unlabeled node s. Assign it label $y_s = \operatorname{argmax}_j \theta_{s;j} + \sum_{t:(s,t)\in E;k} \theta_{s,j;t,k}\mu^*(t;k)$. Now, set $\mu^*(s;y_s) = 1$ and $\mu^*(s;k) = 0$; $k \neq y_s$. Continue with this labeling process until all nodes are labeled. It can be shown that the energy of this assignment y is equal to the energy e^* of the optimal MAP assignment. In particular, each time we take up an unlabeled node t, we select a labeling that does not decrease the expected energy of the unlabeled nodes given the labelings of the labeled nodes. Given that the initial expected energy of all unlabeled nodes was e, the energy at the end of the process, that is, of the assignment y, is thus at least e^*. $\square$

3.1. Convex Approximation

The previous section showed that the relaxation in equation (4), while a simple extension of the LP in equation (2), is actually equivalent to the MAP problem. This yields the interesting result that the MAP problem is solvable in polynomial time if the edge parameter matrix $\Theta = [\theta_{s,j;t,k}]$ is negative definite, since in this case the QP (4) is a convex program. Note also that the quadratic program has a simple set of constraints (only linear and box constraints), which are also small in number, and is thus a simple problem instance of general convex optimization. It should also be stressed that for an n node graph, the QP has only kn variables while the LP has $O(k^2|E|)$ variables.

The case where the edge parameter matrix Θ is not negative definite yields a non-convex program. While one could carry out an iterative search procedure up to a local maximum as in the max-product algorithm, we now describe a convex approximation that provides a polynomial time solution with additive bound guarantees.

Consider the quadratic integer program (QIP) corresponding to the QP, given by

$$\begin{aligned}
\max_{\mu} \quad & \sum_{s;j} \theta_{s;j}\,\mu(s;j) + \sum_{s,t;j,k} \theta_{s,j;t,k}\,\mu(s;j)\,\mu(t;k) \\
\text{subject to} \quad & \sum_j \mu(s;j) = 1 \qquad\qquad (6)\\
& \mu(s;j) \in \{0,1\} \\
& \mu(s,j;t,k) \in \{0,1\}.
\end{aligned}$$

This is clearly equivalent to the MAP problem in equation (1). Let $\Theta = [\theta_{s,j;t,k}]$ be a parameter matrix that is not negative semi-definite. Let $d(s,i)$ be the (positive) diagonal terms that need to be subtracted from the matrix to make it negative semi-definite. An upper bound for d is $d(s,i) \leq \sum_{(t,k)} |\theta_{s,j;t,k}|$, since the negative of a diagonally dominant matrix is negative semi-definite. Let $\Theta' = \Theta - \operatorname{diag}\{d(s;i)\}$ be the negative semi-definite matrix obtained by subtracting off diagonal elements $d(s;i)$. Also, let

$$\theta'_{s;j} = \theta_{s;j} + d(s;j). \qquad (7)$$

Now, for binary $\mu(s;i) \in \{0,1\}$, we have that $\mu(s;i)^2 = \mu(s;i)$; in particular, $d(s;i)\mu(s;j) - d(s;i)\mu(s;j)^2 = 0$. We thus get that the following QIP is equivalent to the QIP (6),

$$\begin{aligned}
\max_{\mu} \quad & \sum_{s;j} \theta'_{s;j}\,\mu(s;j) + \sum_{s,t;j,k} \theta'_{s,j;t,k}\,\mu(s;j)\,\mu(t;k) \\
\text{such that} \quad & \sum_j \mu(s;j) = 1 \\
& \mu(s;j) \in \{0,1\}
\end{aligned}$$

Relaxing this QIP as before, we obtain the following optimization problem.

$$\begin{aligned}
\max_{\mu} \quad & \sum_{s;j} \theta'_{s;j}\,\mu(s;j) + \sum_{s,t;j,k} \theta'_{s,j;t,k}\mu(s;j)\mu(t;k) \\
\text{such that} \quad & \sum_j \mu(s;j) = 1 \\
& \mu(s;j) \in [0,1]
\end{aligned}$$

This is a convex program solvable in polynomial time. The optimality results of the previous section do not follow, however, and the relaxation (8) is not always tight. But as shown next, we can get an additive approximation bound for the discrete solution obtained using the rounding procedure described in the previous section.

Theorem 3.3. *Let μ^* be the optimal solution for the convex QP* (8)*, and let e^* be the optimal MAP energy. Then there is a discrete configuration y (from μ^*) with energy $E(y)$ satisfying*

$$\begin{aligned}
E(y) \quad &\geq \quad e^* - \sum_{s,i} d(s;i)\mu^*(s;i)(1-\mu^*(s;i)) \\
&\geq \quad e^* - \frac{1}{4}\sum_{s,i} d(s;i).
\end{aligned}$$

This result shows that if either Θ is close to negative definite, so that $\sum_{s,i} d(s;i)$ is small, or if the solution is close to integral, so that $\mu^*(s;i)$ is close to zero or one, then the convex relaxation achieves a solution that is close to the optimal MAP solution.

Proof. Just as in the proof of Theorem 3.2, given μ^*, the optimal solution to the convex relaxation, we can efficiently obtain a discrete solution y whose energy is

$$E(y) = \sum_{s;j} \theta_{s;j} \mu^*(s;j) + \sum_{(s,t)\in E;j,k} \theta_{s,j;t,k} \mu^*(s;j) \mu^*(t;k).$$

On the other hand, the optimal value of the convex QP is given by

$$\begin{aligned} e^*_{CQP} &= \sum_{s;j} (\theta_{s;j} + d(s;i))\, \mu^*(s;j) \\ &\quad + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu^*(s;j)\, \mu^*(t;k) \\ &\quad - \sum_{(s,i)} d(s;i) \mu^*(s;i)^2. \end{aligned}$$

We then have that

$$\begin{aligned} E(y) &= e^*_{CQP} - \sum_{(s,i)} d(s;i) \mu^*(s;i)[1 - \mu^*(s;i)] \\ &\geq e^* - \sum_{s,i} d(s;i)\, \mu^*(s;i)\, (1 - \mu^*(s;i)) \\ &\geq e^* - \sum_{s,i} \frac{d(s;i)}{4} \end{aligned}$$

and the result follows. $\square$

3.2. Iterative Update Procedure

Just as tree-reweighted max product gives a set of iterative updates for solving the LP in equation (3), we might ask if there is an iterative update counterpart for the QP. Max-product is a co-ordinate ascent algorithm in the dual (Lagrangian) space for the LP; however, since the dual space of the QP (4) is more complicated, we look at a set of fixed point co-ordinate ascent updates in its primal space.

The QP is given by

$$\begin{aligned} \mu^* &= \max_{\mu} \sum_{s;j} \theta_{s;j} \mu(s;j) \\ &\quad + \sum_{s,t;j,k} \theta_{s,j;t,k}\, \mu(s;j)\, \mu(t;k) \end{aligned} \tag{8}$$

subject to

$$\begin{aligned} \sum_j \mu(s;j) &= 1 \\ \mu(s;j) &\in [0,1]. \end{aligned}$$

Consider node s, and suppose that values for $\mu(t;.)$ are fixed for other nodes $t \neq s$. Then the optimal parameter values $\mu(s;.)$ for node s are given by

$$\begin{aligned} \mu(s;.) &= \max_{\mu(s;.)} \sum_j \theta_{s;j} \mu(s;j) \\ &\quad + \sum_{t;j,k} \theta_{s,j;t,k}\, \mu(s;j)\, \mu(t;k) \end{aligned}$$

subject to $\sum_j \mu(s;j) = 1$. This is easily seen to be solved by taking

$$j^*(s) = \operatorname*{argmax}_j \theta_{s;j} + \sum_{t;j,k} \theta_{s,j;t,k} \mu(t;k)$$

and setting $\mu(s,j) = \mathcal{I}_{j^*(s)}(j)$. This is essentially the iterative conditional modes algorithm (Besag, 1986), which iteratively updates each node with a labeling that most increases the energy, holding fixed the labels of the other nodes.

A better iterative procedure, with stronger and faster convergence properties, albeit for convex programs, is projected conjugate gradient ascent (Axelsson & Barker, 2001). Thus, another advantage of our convex approximation is that we can use conjugate gradient ascent as a simple iterative procedure that is guaranteed to converge (unlike co-ordinate ascent for max product). This makes the convex approximation to the QP applicable to large scale problems.

4. Inner Polytope Relaxations

In the previous section, we obtained a quadratic relaxation by imposing an "independence" constraint on the parameters $\mu(s,j;t,k)$ in equation (4). We also showed that this relaxation is actually tight, and is equivalent to the MAP problem. In this section, we show how one can think of this relaxation as the counterpart of mean-field for MAP, and how any of the corresponding relaxation counterparts of structured mean-field are also tight.

Consider Wainwright and Jordan (2003)'s polytope formulation of MAP in equation (3), given by

$$\mu^* = \max_x \langle \theta, \phi \rangle = \sup_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle$$

where $\mathcal{M}$ is the convex hull of all configuration potentials $\phi(x)$. The second equality follows from the fact that in a linear program, the optimum occurs at an extremal point $\phi(x^*)$. Thus, if $M_I \subset \mathcal{M}$ is any subset of the marginal polytope that includes all of the vertices, then the equations

$$\mu^* = \max_x \langle \theta, \phi \rangle = \sup_{\mu \in M_I} \langle \theta, \mu \rangle$$

still hold. In other words, any relaxation of the indicator variables to $\mu(s,j;t,k) \in \mathcal{M}_I$ would lead to a tight relaxation, as long as $\mathcal{M}_I$ contains all vertices. In contrast, tree-reweighted max product is not tight, since the domain set for its relaxed parameters is $LOCAL(G) \supseteq \mathcal{M}$; see Section 2.1.

As described in (Wainwright & Jordan, 2003), one can think of structured mean field methods as inner polytope approximations. For the given graph G and a subgraph H, let

$$\mathcal{E}(H) = \{\theta' \mid \theta'_{st} = \theta_{st}\, 1_{(s,t)\in H}\}$$

where θ_{st} is the vector of natural parameters associated with edge (s,t). Thus, the parameters not included in the subgraph are set to zero. Now for the subgraph H, we can define the following set of moment parameters:

$$\mathcal{M}(G;H) = \{\mu \mid \mu = E_\theta[\phi(x)] \text{ for some } \theta \in \mathcal{E}(H)\}\,.$$

In essence, the moment parameters in $\mathcal{M}(G;H)$ must be realizable by a distribution that respects the structure of H. For any $H \subseteq G$, the relation $\mathcal{M}(G;H) \subseteq \mathcal{M}(G)$ thus always holds, and $\mathcal{M}(G;H)$ is an inner polytope approximation to $\mathcal{M}$. In particular, taking H to be the completely disconnected graph (*i.e.* no edges) H_0, we have,

$$\begin{aligned}\mathcal{M}(G;H_0) \;\;=\;\; & \{\mu(s;j), \mu(s,j;t,k) \mid \\ & \;\; 0 \le \mu(s;j) \le 1 \\ & \;\; \mu(s,j;t,k) = \mu(s;j)\mu(t;k)\}\end{aligned}$$

which can be seen to be equal to the feasible set of the QP relaxation (4). For this subgraph $H = H_0$, the mean field relaxation thus becomes

$$\begin{aligned}&\sup_{\mu\in\mathcal{M}(G;H_0)} \langle\theta,\mu\rangle \\ &= \sup_{\mu\in\mathcal{M}(G;H_0)} \sum_{s;j}\theta_{s;j}\mu(s;j) + \sum_{st;jk}\theta_{s,j;t,k}\mu(s,j;t,k) \\ &= \sup_{\mu\in\mathcal{M}(G;H_0)} \sum_{s;j}\theta_{s;j}\mu(s;j) + \sum_{st;jk}\theta_{s,j;t,k}\mu(s;j)\mu(t;k)\end{aligned}$$

which is equivalent to the quadratic relaxation in equation (4). Thus, we can, in principle, use any "structured mean-field" relaxation of the form $\sup_{\mu\in M(G;H)}\langle\theta,\mu\rangle$ to solve the MAP *exactly*. The caveat is that this problem, like structured mean field, is a non-convex problem. However, while structured mean field only solves for an approximate value of the log-partition function, the results from Section 3 show that its counterpart for the MAP problem is exact, if the global optimum can be found.

5. Experiments

The quadratic relaxation with the convex approximation was evaluated by comparing it against three competing methods: the linear programming relaxation (Chekuri et al., 2005), the tree-reweighted max product algorithm (Wainwright et al., 2005), and iterative conditional modes (ICM) (Besag, 1986). For tree-reweighted max product, we use the sequential update variant detailed in (Kolmogorov, 2005), which has better convergence properties than the originally proposed algorithm.

The approximate MAP algorithms were compared on different potential functions and coupling types for 2D nearest neighbor grid graphs with 100 nodes and a label set of size four. The node potentials were generated uniformly $\mathcal{U}(-1,1)$, while the edge potentials were generated as a product of an edge weight and a distance function on labels. For different settings of an edge coupling-strength parameter, d_{coup}, the edge weight was selected from $\mathcal{U}(-d_{coup}, d_{coup})$ for the mixed coupling, from $\mathcal{U}(0, 2\,d_{coup})$ for the positive coupling, and from $\mathcal{U}(-2\,d_{coup}, 0)$ for the negative coupling. The following four commonly used distances were used for the distance function: Ising, $\phi(l,m) = lm$; uniform, or Potts, $\phi(l,m) = \mathbb{I}(l=m)$; quadratic, $\phi(l,m) = (l-m)^2$; linear $\phi(l,m) = |l-m|$.

Figures 1 and 2 show plots of the value (energy) of the MAP estimates using the different algorithms for a range of model types. For any given setting of parameters and potential functions, a higher value is closer to the MAP estimate and is thus better. As the plots show, the quadratic relaxation slightly outperforms tree-reweighted max product for mixed and positive couplings, and is comparable or slightly worse for negative coupling. The quadratic approximation typically beats both ICM and the linear relaxation.

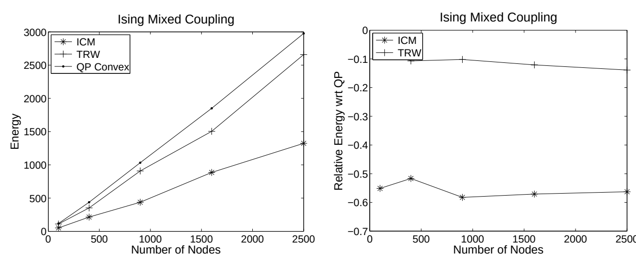

Figure 3. Comparison of ICM and TRW on larger graphs, using Ising potentials with mixed coupling. The right plot shows $(e_{\text{ICM}} - e_{\text{QP}})/e_{\text{ICM}}$ and $(e_{\text{TRW}} - e_{\text{QP}})/e_{\text{TRW}}$.

In Figure 3 we compare the MAP estimates from different algorithms on larger graphs, using the Ising potential function with mixed coupling. The quadratic relaxation is seen to outperform ICM and tree-reweighted max product, even as the number of nodes increases.

We note that since the convex approximation to the QP is a convex program, it can be solved (in polynomial time) using standard QP solvers for small problems, and for larger-scale problems one can use iterative projected conjugate gradient, which provides a fast iterative method that is guaranteed to converge. In our experiments, the computation time for the QP method was comparable to that required by tree-reweighted max product, which in turn required much less time to solve than the linear programming relaxation. This is due primarily to the fact that the linear program has $|E|k^2$ variables, while the convex quadratic

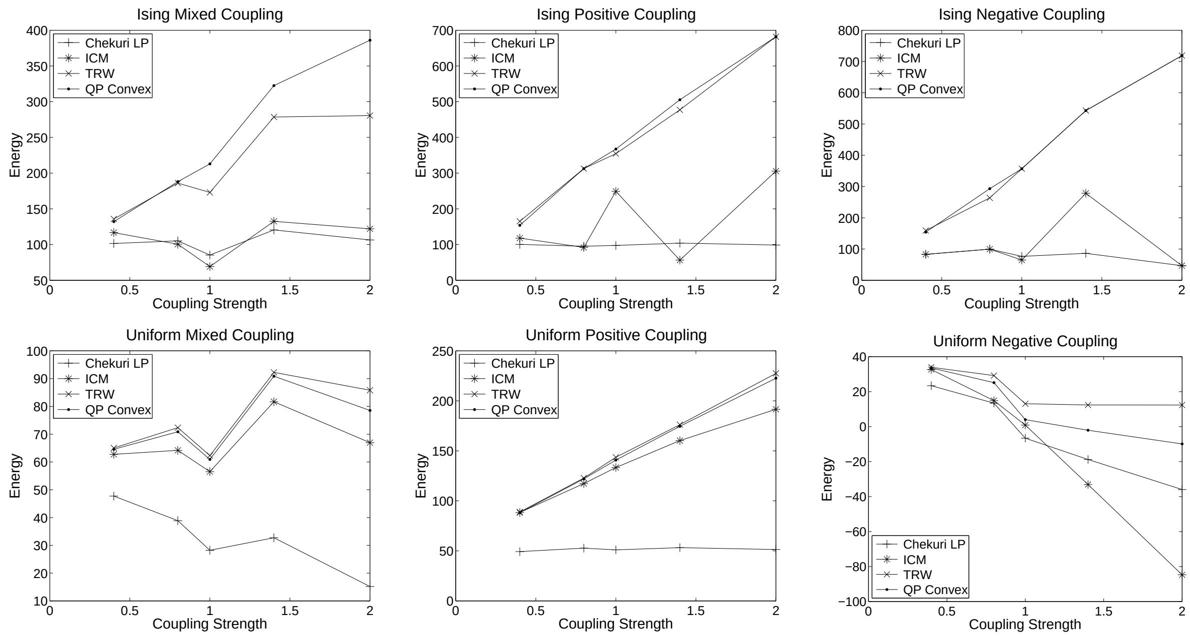

Figure 1. Comparison of linear relaxation (LP), iterative conditional modes (ICM), tree-reweighted max product (TRW), and quadratic programming relaxation (QP) on 10×10 grid graphs using Ising potentials (top row) and uniform potentials (bottom) with mixed (left), positive (center) and negative (right) couplings. A better MAP estimate has a higher value.

relaxation has only nk variables, where n is the number of nodes in the graph, $|E|$ is the number of edges, and k is the number of labels.

6. Conclusions

This paper has proposed a quadratic programming relaxation to the MAP problem for random fields, or the metric labeling problem. The quadratic objective function more accurately represents the energy in the graphical model while using fewer variables than the linear relaxation. It was shown that the QP relaxation is tight and computes the MAP solution exactly. However, under certain conditions the relaxation results in a non-convex problem, which requires an intractable search over local minima. This led to an additional convex approximation to the relaxation, for which there is an additive approximation guarantee. The quadratic programming approximation was also extended to general variational "inner polytope" relaxations that also compute the MAP exactly. Experiments demonstrated that the quadratic relaxation, with the convex approximation, can outperform existing methods under many settings, while also being computationally attractive.

Acknowledgments

This research was supported in part by NSF grants IIS-0312814 and IIS-0427206. The authors thank Ramin Zabih and Vladimir Kolmogorov for helpful comments.

References

Axelsson, O., & Barker, V. A. (2001). *Finite element solution of boundary value problems: theory and computation*. Society for Industrial and Applied Mathematics.

Bertsimas, D., & Tsitsiklis, J. (1997). *Introduction to linear optimization*. Athena Scientific.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, *23*.

Chekuri, C., Khanna, S., Naor, J., & Zosin, L. (2005). A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM Journal on Discrete Mathematics*, *18*, 608–625.

Greig, D., Porteous, B., & Seheult, A. (1989). Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society, Series B*, *51*.

Kleinberg, J., & Tardos, E. (1999). Approximation algorithms for classification problems with pairwise relationships: Metric partitioning and markov random fields.

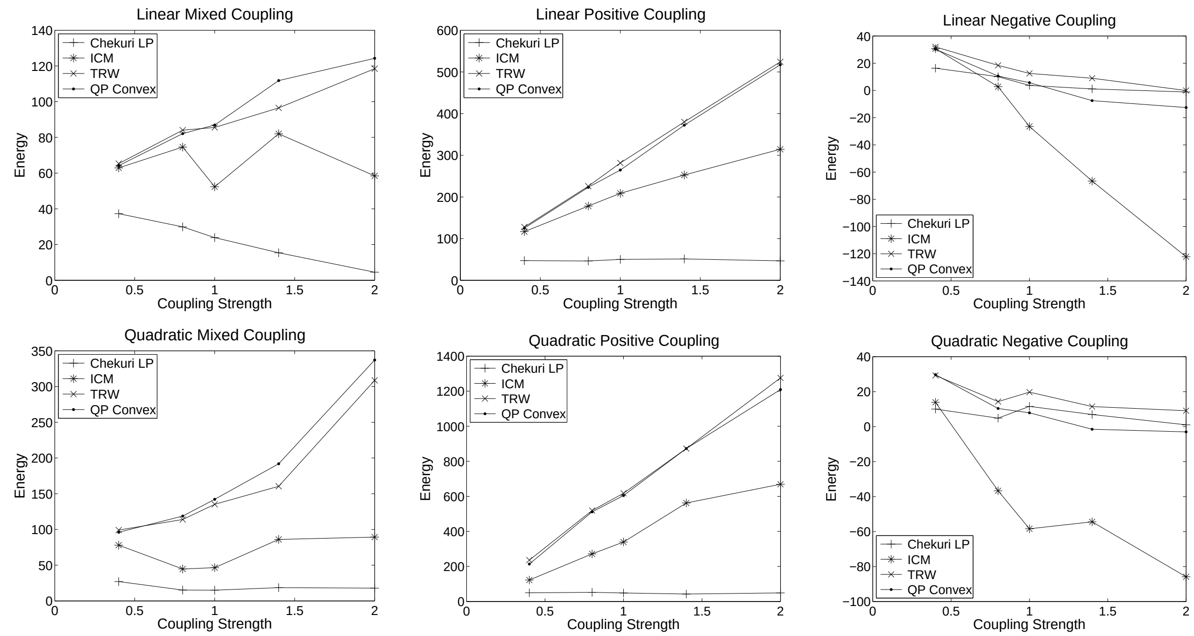

Figure 2. Comparison of linear relaxation (LP), iterative conditional modes (ICM), tree-reweighted max product (TRW), and quadratic programming relaxation (QP) on 10×10 grid graphs using linear potentials (top row) and quadratic potentials (bottom) with mixed (left), positive (center) and negative (right) couplings. A better MAP estimate has a higher value.

IEEE Symposium on the Foundations of Computer Science.

Kolmogorov, V. (2005). Convergent tree-reweighted message passing for energy minimization. *AISTATS*.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc.

Wainwright, M., Jaakkola, T., & Willsky, A. (2005). Map estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. *IEEE Transactions on Information Theory*, *51*, 3697–3717.

Wainwright, M. J., & Jordan, M. I. (2003). Variational inference in graphical models: The view from the marginal polytope. *Allerton Conference on Communication, Control, and Computing*.

Weiss, Y., & Freeman, W. T. (2001). On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, *47*.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2001). Understanding belief propagation and its generalizations. *IJCAI 2001 Distinguished Lecture track*.