Unifying Local Consistency and MAX SAT Relaxations for Scalable Inference with Rounding Guarantees

Stephen H. Bach
University of Maryland

Bert Huang
Virginia Tech

Lise Getoor
University of California, Santa Cruz

Abstract

We prove the equivalence of first-order local consistency relaxations and the MAX SAT relaxation of Goemans and Williamson (1994) for a class of MRFs we refer to as logical MRFs. This allows us to combine the advantages of each into a single MAP inference technique: solving the local consistency relaxation with any of a number of highly scalable message-passing algorithms, and then obtaining a high-quality discrete solution via a guaranteed rounding procedure when the relaxation is not tight. Logical MRFs are a general class of models that can incorporate many common dependencies, such as logical implications and mixtures of supermodular and submodular potentials. They can be used for many structured prediction tasks, including natural language processing, computer vision, and computational social science. We show that our new inference technique can improve solution quality by as much as 20% without sacrificing speed on problems with over one million dependencies.

1 INTRODUCTION

One of the canonical problems for probabilistic modeling is finding the most probable variable assignment, i.e., maximum a posteriori (MAP) inference. For Markov random fields (MRFs), MAP inference is NP-hard in general (Shimony, 1994), so approximations are required in practice. In this paper, we provide a new analysis of approximate MAP inference for a particularly flexible and broad class of MRFs we refer to as *logical MRFs*. In these models, potentials are defined by truth tables of disjunctive logical clauses with nonnegative weights. This class includes many common types of models, such as mixtures of supermodular and submodular potentials, and many of the models that can be defined using the language Markov logic (Richardson and Domingos, 2006).[1] Such models are useful for the many domains that require expressive dependencies, such as natural language processing, computer vision, and computational social science. MAP inference for logical MRFs is still NP-hard (Garey et al., 1976), so we consider two main approaches for approximate inference, each with distinct advantages.

The first approach uses *local consistency relaxations* (Wainwright and Jordan, 2008). Instead of solving a combinatorial optimization over discrete variables, MAP inference is first viewed equivalently as the optimization of marginal distributions over variable and potential states. The marginals are then relaxed to *pseudomarginals*, which are only consistent among local variables and potentials. The primary advantage of local consistency relaxations is that they lead to highly scalable message-passing algorithms, such as dual decomposition (Sontag et al., 2011). However—except for a few special cases—local consistency relaxations produce fractional solutions, which require some rounding or decoding procedure to find discrete solutions. For most MRFs, including logical MRFs, there are no previously known guarantees on the quality of these solutions.

The second approach to tractable MAP inference for logical MRFs is *weighted maximum satisfiability (MAX SAT) relaxation*, in which one views MAP inference as the classic MAX SAT problem and relaxes it to a convex program from that perspective. Given a set of disjunctive logical clauses with associated nonnegative weights, MAX SAT is the problem of finding a Boolean variable assignment that maximizes the sum of the weights of the satisfied clauses. Convex programming relaxations for MAX SAT also produce

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

[1] Markov logic can also include clauses with conjunctions and with negative or infinite weights.

fractional solutions, but unlike local consistency relaxations, they offer theoretically guaranteed rounding procedures (Goemans and Williamson, 1994). However, even though these relaxations are tractable in principle, general-purpose convex program solvers do not scale well to inference for large graphical models (Yanover et al., 2006).

In this paper, we unite these two approaches. Our first contribution is the following theoretical result: for logical MRFs, the first-order local consistency relaxation and the MAX SAT relaxation of Goemans and Williamson (1994) are equivalent. We prove this equivalence by analyzing the local consistency relaxation as a hierarchical optimization and reasoning about KKT conditions of the optimizations at the lower level of the hierarchy. Replacing the optimizations at the lower level with compact solutions shows the equivalence.

This proof of equivalence is important because it reveals that one can combine the advantages of both approaches into a single algorithm by using the message-passing algorithms developed in the graphical models community and the guaranteed rounding procedure of the MAX SAT relaxation. We show that our technique can improve solution quality by as much as 20% on problems with over one million clauses.

2 PRELIMINARIES

In this section, we review MRFs, local consistency relaxations, and MAX SAT relaxations; and we define logical MRFs.

2.1 Markov Random Fields

MRFs are probabilistic graphical models that factor according to the structure of an undirected graph. For the purposes of this paper, we consider MRFs with discrete domains.

Definition 1. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a vector of n random variables, where each variable x_i has discrete domain $\mathcal{X}_i = \{0, 1, \ldots, K_i - 1\}$. Then, let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$ be a vector of m potentials, where each potential $\phi_j(\boldsymbol{x})$ maps states of a subset of the variables $\boldsymbol{x}_j$ to real numbers. Finally, let $\boldsymbol{w} = (w_1, \ldots, w_m)$ be a vector of m real-valued parameters. Then, a **Markov random field** over $\boldsymbol{x}$ is a probability distribution of the form

$$P(\boldsymbol{x}) \propto \exp\left(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})\right) \ .$$

MAP inference, i.e., finding the solution of $\arg\max_{\boldsymbol{x}} \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$, is NP-hard (Shimony, 1994). Thus, various approaches have been developed to approximate MAP inference efficiently.

2.2 Local Consistency Relaxations

A popular approach for tractable inference in MRFs is local consistency relaxation (Wainwright and Jordan, 2008). This approach starts by viewing MAP inference as an equivalent optimization over marginal probabilities. For each $\phi_j \in \boldsymbol{\phi}$, let $\boldsymbol{\theta}_j$ be a marginal distribution over joint assignments $\boldsymbol{x}_j$. For example, $\theta_j(\boldsymbol{x}_j)$ is the probability that the subset of variables associated with potential ϕ_j is in a particular joint state $\boldsymbol{x}_j$. Also, let $x_j(i)$ denote the setting of the variable with index i in the state $\boldsymbol{x}_j$.

With this variational formulation, inference can be relaxed to an optimization over the *first-order local polytope* $\mathbb{L}$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ be a vector of probability distributions, where $\mu_i(k)$ is the marginal probability that x_i is in state k. The first-order local polytope is

$$\mathbb{L} \triangleq \left\{ \boldsymbol{\theta}, \boldsymbol{\mu} \;\middle|\; \begin{array}{ll} \sum_{\boldsymbol{x}_j | x_j(i)=k} \theta_j(\boldsymbol{x}_j) = \mu_i(k) & \forall i, j, k \\ \sum_{\boldsymbol{x}_j} \theta_j(\boldsymbol{x}_j) = 1 & \forall j \\ \sum_{k=0}^{K_i - 1} \mu_i(k) = 1 & \forall i \\ \theta_j(\boldsymbol{x}_j) \geq 0 & \forall j, \boldsymbol{x}_j \\ \mu_i(k) \geq 0 & \forall i, k \end{array} \right\},$$

which constrains each marginal distribution $\boldsymbol{\theta}_j$ over joint states $\boldsymbol{x}_j$ to be consistent only with the marginal distributions $\boldsymbol{\mu}$ over individual variables that participate in the potential ϕ_j.

MAP inference can then be approximated with the *first-order local consistency relaxation*:

$$\arg\max_{(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \mathbb{L}} \ \sum_{j=1}^{m} w_j \sum_{\boldsymbol{x}_j} \theta_j(\boldsymbol{x}_j) \ \phi_j(\boldsymbol{x}_j), \tag{1}$$

which is an upper bound on the true MAP objective. The first-order local consistency relaxation is a much more tractable linear program than exact inference, and it can be applied to any MRF. Much work has focused on solving the first-order local consistency relaxation for large-scale MRFs, which we discuss further in Sections 4 and 6. However, in general, the solutions are fractional, and there are no guarantees on the approximation quality of a tractable discretization of these fractional solutions.

2.3 Logical Markov Random Fields

We now turn to the focus of this paper: logical MRFs, which are MRFs whose potentials $\boldsymbol{\phi}$ are defined by disjunctive Boolean clauses with associated nonnegative weights, formally defined as follows.

Definition 2. Let $\boldsymbol{C} = (C_1, \ldots, C_m)$ be a vector of logical clauses, where each clause $C_j \in \boldsymbol{C}$ is a disjunction of literals and each literal is a variable x or its

negation $\neg x$ such that each variable $x_i \in \boldsymbol{x}$ appears at most once in C_j. Let I_j^+ (resp. I_j^-) $\subseteq \{1, \ldots, n\}$ be the set of indices of the variables that are not negated (resp. negated) in C_j. Then C_j can be written as

$$\left(\bigvee_{i \in I_j^+} x_i \right) \quad \vee \quad \left(\bigvee_{i \in I_j^-} \neg x_i \right) .$$

A **logical Markov random field** is an MRF in which each variable x_i has Boolean domain $\{0, 1\}$, i.e., $K_i = 2$, each potential $\phi_j(\boldsymbol{x}) = 1$ if $\boldsymbol{x}$ satisfies C_j and 0 otherwise, and each parameter $w_j \geq 0$.

Logical MRFs are very expressive. A clause C_j can be viewed equivalently as an implication from conditions to consequences:

$$\bigwedge_{i \in I_j^-} x_i \implies \bigvee_{i \in I_j^+} x_i .$$

Multiple clauses can together express dependencies that cannot be expressed in a single clause, such as multiple sets of conditions implying one set of possible consequences, or one set of conditions implying multiple sets of possible consequences.

The generality of logical MRFs can be stated more broadly: MAP inference for any discrete distribution of bounded factor size can be converted to a MAX SAT problem—and therefore MAP inference for a logical MRF—of size polynomial in the variables and clauses (Park, 2002a). One way to convert MAP inference for a general MRF to MAP inference for a logical MRF is to encode each entry of the original potential table as its own potential defined by a disjunctive clause. The weight of that clause is the negated value of the original potential energy. Weights can be shifted to be all nonnegative by adding a constant to each, but this shift loosens the rounding guarantees introduced in the next subsection. However, in practice, logical MRFs can model many important problems without the need for such conversions.

2.4 MAX SAT Relaxations

The MAP problem for a logical MRF can also be viewed as an instance of MAX SAT and approximately solved from this perspective. The MAX SAT problem is to find a Boolean assignment to the variables $\boldsymbol{x}$ that maximizes the sum of the weights of the satisfied clauses in $\boldsymbol{C}$. A solution to MAX SAT is therefore also the MAP state of the logical MRF defined via $\boldsymbol{C}$. Since MAX SAT is NP-hard, a large body of work has focused on constructing convex programming relaxations for this problem.

Goemans and Williamson (1994) introduced a linear programming relaxation that provides rounding guarantees for the solution. We review their technique and the results of their analysis here. For each variable x_i, associate with it a continuous variable $y_i \in [0, 1]$. Then, let $\boldsymbol{y}^\star$ be the solution to the linear program

$$\operatorname*{arg\,max}_{\boldsymbol{y} \in [0,1]^n} \sum_{C_j \in \boldsymbol{C}} w_j \min \left\{ \sum_{i \in I_j^+} y_i + \sum_{i \in I_j^-} (1 - y_i), 1 \right\} \quad (2)$$

and let each variable x_i be independently set to 1 according to a rounding probability function f, i.e., $p_i = f(y_i^\star)$. Many functions can be chosen for f, but a simple one Goemans and Williamson (1994) analyze is the linear function

$$f(y_i^\star) = \frac{1}{2} y_i^\star + \frac{1}{4} .$$

Let $\hat{W}$ be the expected total weight of the satisfied clauses from using this randomized rounding procedure. More precisely,

$$\hat{W} = \sum_{C_j \in \boldsymbol{C}} w_j \left(1 - \prod_{i \in I_j^+} (1 - p_i) \prod_{i \in I_j^-} p_i \right) . \quad (3)$$

Let $W^\star$ be the maximum total weight of the satisfied clauses over all assignments to $\boldsymbol{x}$, i.e., the weight of the MAX SAT solution. Goemans and Williamson (1994) showed that

$$\hat{W} \geq \frac{3}{4} W^\star .$$

The method of conditional probabilities (Alon and Spencer, 2008) can deterministically find an assignment to $\boldsymbol{x}$ that achieves a total weight of at least $\hat{W}$. Each variable x_i is greedily set to the value that maximizes the expected weight over the unassigned variables, conditioned on either possible value of x_i and the previously assigned variables. This greedy maximization can be applied quickly because, in many models, variables only participate in a small fraction of the clauses, making the change in expectation quick to compute for each variable. Specifically, referring to the definition of $\hat{W}$ (3), the assignment to x_i only needs to maximize over the clauses C_j in which x_i participates, i.e., $i \in I_j^+ \cup I_j^-$, which is usually a small set.

3 EQUIVALENCE ANALYSIS

In this section, we prove the equivalence of the first-order local consistency relaxation and the MAX SAT relaxation of Goemans and Williamson (1994) for logical MRFs (Theorem 6). Our proof analyzes the local consistency relaxation to derive an equivalent, more

compact optimization over only the variable pseudomarginals $\boldsymbol{\mu}$ that is identical to the MAX SAT relaxation. Since the variables are Boolean, we refer to each pseudomarginal $\mu_i(1)$ as simply μ_i. Let $\boldsymbol{x}_j^F$ denote the unique setting such that $\phi_j(\boldsymbol{x}_j^F) = 0$. (I.e., $\boldsymbol{x}_j^F$ is the setting in which each literal in the clause C_j is false.)

We begin by reformulating the local consistency relaxation as a hierarchical optimization, first over the variable pseudomarginals $\boldsymbol{\mu}$ and then over the factor pseudomarginals $\boldsymbol{\theta}$. Due to the structure of local polytope $\mathbb{L}$, the pseudomarginals $\boldsymbol{\mu}$ parameterize inner linear programs that decompose over the structure of the MRF, such that—given fixed $\boldsymbol{\mu}$—there is an independent linear program $\hat{\phi}_j(\boldsymbol{\mu})$ over $\boldsymbol{\theta}_j$ for each clause C_j. We rewrite objective (1) as

$$\arg\max_{\boldsymbol{\mu}\in[0,1]^n} \sum_{C_j \in \boldsymbol{C}} \hat{\phi}_j(\boldsymbol{\mu}), \tag{4}$$

where

$$\hat{\phi}_j(\boldsymbol{\mu}) = \max_{\boldsymbol{\theta}_j} \; w_j \sum_{\boldsymbol{x}_j | \boldsymbol{x}_j \neq \boldsymbol{x}_j^F} \theta_j(\boldsymbol{x}_j) \tag{5}$$

$$\text{s.t.} \sum_{\boldsymbol{x}_j | x_j(i)=1} \theta_j(\boldsymbol{x}_j) = \mu_i \qquad \forall i \in I_j^+ \tag{6}$$

$$\sum_{\boldsymbol{x}_j | x_j(i)=0} \theta_j(\boldsymbol{x}_j) = 1 - \mu_i \quad \forall i \in I_j^- \tag{7}$$

$$\sum_{\boldsymbol{x}_j} \theta_j(\boldsymbol{x}_j) = 1 \tag{8}$$

$$\theta_j(\boldsymbol{x}_j) \geq 0 \qquad \forall \boldsymbol{x}_j \; . \tag{9}$$

It is straightforward to verify that objectives (1) and (4) are equivalent for logical MRFs. All constraints defining $\mathbb{L}$ can be derived from the constraint $\boldsymbol{\mu} \in [0,1]^n$ and the constraints in the definition of $\hat{\phi}_j(\boldsymbol{\mu})$. We have omitted redundant constraints to simplify analysis.

To make this optimization more compact, we replace each inner linear program $\hat{\phi}_j(\boldsymbol{\mu})$ with an expression that gives its optimal value for any setting of $\boldsymbol{\mu}$. Deriving this expression requires reasoning about any maximizer $\boldsymbol{\theta}_j^\star$ of $\hat{\phi}_j(\boldsymbol{\mu})$, which is guaranteed to exist because problem (5) is bounded and feasible[2] for any parameters $\boldsymbol{\mu} \in [0,1]^n$ and w_j.

We first derive a sufficient condition for the linear program to not be fully satisfiable, in the sense that it cannot achieve a value of w_j, the maximum value of the weighted potential $w_j\phi_j(\boldsymbol{x})$. Observe that, by the objective (5) and the simplex constraint (8), showing that $\hat{\phi}_j(\boldsymbol{\mu})$ is not fully satisfiable is equivalent to showing that $\theta_j^\star(\boldsymbol{x}_j^F) > 0$.

[2]Setting $\theta_j(\boldsymbol{x}_j)$ to the probability defined by $\boldsymbol{\mu}$ under the assumption that the elements of $\boldsymbol{x}_j$ are independent, i.e., the product of the pseudomarginals, is always feasible.

Lemma 1. *If*

$$\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) < 1 \; ,$$

then $\theta_j^\star(\boldsymbol{x}_j^F) > 0$.

Proof. By the simplex constraint (8),

$$\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) < \sum_{\boldsymbol{x}_j} \theta_j^\star(\boldsymbol{x}_j) \; .$$

Also, by summing all the constraints (6) and (7),

$$\sum_{\boldsymbol{x}_j | \boldsymbol{x}_j \neq \boldsymbol{x}_j^F} \theta_j^\star(\boldsymbol{x}_j) \leq \sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) \; ,$$

because all the components of $\boldsymbol{\theta}^\star$ are nonnegative, and—except for $\theta_j^\star(\boldsymbol{x}_j^F)$—they all appear at least once in constraints (6) and (7). These bounds imply

$$\sum_{\boldsymbol{x}_j | \boldsymbol{x}_j \neq \boldsymbol{x}_j^F} \theta_j^\star(\boldsymbol{x}_j) < \sum_{\boldsymbol{x}_j} \theta_j^\star(\boldsymbol{x}_j) \; ,$$

which means $\theta_j^\star(\boldsymbol{x}_j^F) > 0$, completing the proof. □

We next show that if $\hat{\phi}_j(\boldsymbol{\mu})$ is parameterized such that it is not fully satisfiable, as in Lemma 1, then its optimum always takes a particular value defined by $\boldsymbol{\mu}$.

Lemma 2. *If* $w_j > 0$ *and* $\theta_j^\star(\boldsymbol{x}_j^F) > 0$*, then*

$$\sum_{\boldsymbol{x}_j | \boldsymbol{x}_j \neq \boldsymbol{x}_j^F} \theta_j^\star(\boldsymbol{x}_j) = \sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) \; .$$

Proof. We prove the lemma via the Karush-Kuhn-Tucker (KKT) conditions (Karush, 1939; Kuhn and Tucker, 1951). Since problem (5) is a maximization of a linear function subject to linear constraints, the KKT conditions are necessary and sufficient for any optimum $\boldsymbol{\theta}_j^\star$.

Before writing the relevant KKT conditions, we introduce some necessary notation. For a state $\boldsymbol{x}_j$, we need to reason about the variables that disagree with the unsatisfied state $\boldsymbol{x}_j^F$. Let

$$d(\boldsymbol{x}_j) \triangleq \left\{ i \in I_j^+ \cup I_j^- | \boldsymbol{x}_j(i) \neq \boldsymbol{x}_j^F(i) \right\}$$

be the set of indices for the variables that do not have the same value in the two states $\boldsymbol{x}_j$ and $\boldsymbol{x}_j^F$.

We now write the relevant KKT conditions for $\boldsymbol{\theta}_j^\star$. Let $\boldsymbol{\lambda}, \boldsymbol{\alpha}$ be real-valued vectors where $|\boldsymbol{\lambda}| = |I_j^+| + |I_j^-| + 1$ and $|\boldsymbol{\alpha}| = |\boldsymbol{\theta}_j|$. Let each λ_i correspond to a constraint (6) or (7) for $i \in I_j^+ \cup I_j^-$, and let λ_Δ correspond to the simplex constraint (8). Also, let each

$\alpha_{\boldsymbol{x}_j}$ correspond to a constraint (9) for each $\boldsymbol{x}_j$. Then, the following KKT conditions hold:

$$\alpha_{\boldsymbol{x}_j} \geq 0 \qquad \forall \boldsymbol{x}_j \quad (10)$$

$$\alpha_{\boldsymbol{x}_j} \theta^\star_j(\boldsymbol{x}_j) = 0 \qquad \forall \boldsymbol{x}_j \quad (11)$$

$$\lambda_\Delta + \alpha_{\boldsymbol{x}_j^F} = 0 \quad (12)$$

$$w_j + \sum_{i \in d(\boldsymbol{x}_j)} \lambda_i + \lambda_\Delta + \alpha_{\boldsymbol{x}_j} = 0 \quad \forall \boldsymbol{x}_j \neq \boldsymbol{x}_j^F \ . \quad (13)$$

Since $\theta^\star_j(\boldsymbol{x}_j^F) > 0$, by condition (11), $\alpha_{\boldsymbol{x}_j^F} = 0$. By condition (12), then $\lambda_\Delta = 0$. From here we can bound the other elements of $\boldsymbol{\lambda}$. Observe that for every $i \in I_j^+ \cup I_j^-$, there exists a state $\boldsymbol{x}_j$ such that $d(\boldsymbol{x}_j) = \{i\}$. Then, it follows from condition (13) that there exists $\boldsymbol{x}_j$ such that, for every $i \in I_j^+ \cup I_j^-$,

$$w_j + \lambda_i + \lambda_\Delta + \alpha_{\boldsymbol{x}_j} = 0 \ .$$

Since $\alpha_{\boldsymbol{x}_j} \geq 0$ by condition (10) and $\lambda_\Delta = 0$, it follows that $\lambda_i \leq -w_j$. With these bounds, we show that, for any state $\boldsymbol{x}_j$, if $|d(\boldsymbol{x}_j)| \geq 2$, then $\theta^\star_j(\boldsymbol{x}_j) = 0$. Assume that for some state $\boldsymbol{x}_j$, $|d(\boldsymbol{x}_j)| \geq 2$. By condition (13) and the derived constraints on $\boldsymbol{\lambda}$,

$$\alpha_{\boldsymbol{x}_j} \geq (|d(\boldsymbol{x}_j)| - 1) w_j > 0 \ .$$

With condition (11), $\theta^\star_j(\boldsymbol{x}_j) = 0$. Next, observe that for all $i \in I_j^+$ (resp. $i \in I_j^-$) and for any state $\boldsymbol{x}_j$, if $d(\boldsymbol{x}_j) = \{i\}$, then $x_j(i) = 1$ (resp. $x_j(i) = 0$), and for any other state $\boldsymbol{x}'_j$ such that $x'_j(i) = 1$ (resp. $x'_j(i) = 0$), $d(\boldsymbol{x}'_j) \geq 2$. By constraint (6) (resp. constraint (7)), $\theta^\star(\boldsymbol{x}_j) = \mu_i$ (resp. $\theta^\star(\boldsymbol{x}_j) = 1 - \mu_i$).

We have shown that if $\theta^\star_j(\boldsymbol{x}_j^F) > 0$, then for all states $\boldsymbol{x}_j$, if $d(\boldsymbol{x}_j) = \{i\}$ and $i \in I_j^+$ (resp. $i \in I_j^-$), then $\theta^\star_j(\boldsymbol{x}_j) = \mu_i$ (resp. $\theta^\star_j(\boldsymbol{x}_j) = 1 - \mu_i$), and if $|d(\boldsymbol{x}_j)| \geq 2$, then $\theta^\star_j(\boldsymbol{x}_j) = 0$. This completes the proof. □

Lemma 1 says if $\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) < 1$, then $\hat{\phi}_j(\boldsymbol{\mu})$ is not fully satisfiable, and Lemma 2 provides its optimal value. We now reason about the other case, when $\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) \geq 1$, and we show that it is sufficient to ensure that $\hat{\phi}_j(\boldsymbol{\mu})$ is fully satisfiable.

Lemma 3. *If $w_j > 0$ and*

$$\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) \geq 1 \ ,$$

then $\theta^\star_j(\boldsymbol{x}_j^F) = 0$.

Proof. We prove the lemma by contradiction. Assume that $w_j > 0$, $\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) \geq 1$, and that the lemma is false, $\theta^\star_j(\boldsymbol{x}_j^F) > 0$. Then, by Lemma 2,

$$\sum_{\boldsymbol{x}_j | \boldsymbol{x}_j \neq \boldsymbol{x}_j^F} \theta^\star_j(\boldsymbol{x}_j) \geq 1 \ .$$

The assumption that $\theta^\star_j(\boldsymbol{x}_j^F) > 0$ implies

$$\sum_{\boldsymbol{x}_j} \theta^\star_j(\boldsymbol{x}_j) > 1,$$

which is a contradiction, since it violates the simplex constraint (8). The possibility that $\theta^\star_j(\boldsymbol{x}_j^F) < 0$ is excluded by the nonnegativity constraints (9). □

For completeness and later convenience, we also state the value of $\hat{\phi}_j(\boldsymbol{\mu})$ when it is fully satisfiable, which follows from the simplex constraint (8).

Lemma 4. *If $\theta^\star_j(\boldsymbol{x}_j^F) = 0$, then*

$$\sum_{\boldsymbol{x}_j | \boldsymbol{x}_j \neq \boldsymbol{x}_j^F} \theta^\star_j(\boldsymbol{x}_j) = 1 \ .$$

We can now combine the previous lemmas into a single expression for the value of $\hat{\phi}_j(\boldsymbol{\mu})$.

Lemma 5.

$$\hat{\phi}_j(\boldsymbol{\mu}) = w_j \min \left\{ \sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i), 1 \right\} \ .$$

Proof. The lemma is trivially true if $w_j = 0$ since any assignment will yield zero value. If $w_j > 0$, then we consider two cases. In the first case, if $\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) < 1$, then, by Lemmas 1 and 2,

$$\hat{\phi}_j(\boldsymbol{\mu}) = w_j \left(\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) \right) \ .$$

In the second case, if $\sum_{i \in I_j^+} \mu_i + \sum_{i \in I_j^-} (1 - \mu_i) \geq 1$, then, by Lemmas 3 and 4,

$$\hat{\phi}_j(\boldsymbol{\mu}) = w_j \ .$$

Factoring out w_j completes the proof. □

This leads to our final equivalence result.

Theorem 6. *For a logical MRF, the first-order local consistency relaxation of MAP inference is equivalent to the MAX SAT relaxation of Goemans and Williamson (1994). Specifically, any partial optimum $\boldsymbol{\mu}^\star$ of objective (1) is an optimum $\boldsymbol{y}^\star$ of objective (2), and vice versa.*

Proof. Substituting the solution of the inner optimization from Lemma 5 into the local consistency relaxation objective (4) gives a projected optimization over only $\boldsymbol{\mu}$ which is identical to the MAX SAT relaxation objective (2). □

We discuss the practical implications of this proof of equivalence in the next section.

4 SOLVING THE RELAXATION

A large body of work has focused on solving local consistency relaxations of MAP inference quickly. Typically, off-the-shelf convex optimization methods do not scale well for large graphical models (Yanover et al., 2006), so a large, important branch of research has investigated highly scalable message-passing algorithms. In this section, we examine how such algorithms can be augmented with the rounding guarantees of the MAX SAT relaxation, since they optimize the same objective. We propose using a message-passing algorithm to optimize the local consistency relaxation and then applying the rounding procedure of Goemans and Williamson (1994) to obtain a high-quality discrete solution. Any algorithm that can find the optimal variable pseudomarginals $\boldsymbol{\mu}^\star$ is applicable. We consider three families of message-passing algorithms from graphical models literature.

The first approach is dual decomposition (DD) (Sontag et al., 2011), which solves a dual problem of (1):

$$\begin{aligned}\min_{\boldsymbol{\delta}} \sum_{i=1}^{n} \max_{x_i} \left(\sum_{j|x_i \in \boldsymbol{x}_j} \delta_{ij}(x_i) \right) \\ + \sum_{j=1}^{m} \max_{\boldsymbol{x}_j} \left(w_j \phi_j(\boldsymbol{x}_j) - \sum_{i|x_i \in \boldsymbol{x}_j} \delta_{ij}(x_i) \right)\end{aligned} \tag{14}$$

where $\boldsymbol{\delta}$ is a vector of dual variables. Since the pseudomarginals $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ do not actually appear in objective (14), only some DD algorithms can be used to find the optimum $\boldsymbol{\mu}^\star$ in order to compute rounding probabilities. Subgradient methods for DD (e.g., Jojic et al. (2010), Komodakis et al. (2011), and Schwing et al. (2012)) can find $\boldsymbol{\mu}^\star$ in many ways, including those described by Anstreicher and Wolsey (2009), Nedić and Ozdaglar (2009), and Shor (1985). Other DD algorithms, such as TRW-S (Kolmogorov, 2006), MSD (Werner, 2007), MPLP (Globerson and Jaakkola, 2007), and ADLP (Meshi and Globerson, 2011), use coordinate descent to solve the dual objective. In general, there is no known way to find the primal solution $\boldsymbol{\mu}^\star$ with coordinate descent DD.

The second approach uses message-passing algorithms to solve objective (1) directly in its primal form and therefore always finds $\boldsymbol{\mu}^\star$. One well-known algorithm is that of Ravikumar et al. (2010), which uses proximal optimization, a general approach that iteratively improves the solution by searching for nearby improvements. The authors also provide rounding guarantees for when the relaxed solution is integral, i.e., the relaxation is tight, allowing the algorithm to converge faster. Such guarantees are complementary to ours, since we consider the case when the relaxation is not tight. Another message-passing algorithm that solves the primal objective is AD^3 (Martins et al., 2011), which uses the alternating direction method of multipliers (ADMM) (Boyd et al., 2011). AD^3 optimizes objective (1) for binary, pairwise MRFs and supports the addition of certain deterministic constraints on the variables. A third example of a primal message-passing algorithm is APLP (Meshi and Globerson, 2011), which is the primal analog of ADLP. Like AD^3, it uses ADMM to optimize the objective.

A third approach is a new one that we identify and deserves special emphasis. The compact form of relaxed MAP inference for logical MRFs derived in Section 3 is subsumed by exact MAP inference for hinge-loss Markov random fields (HL-MRFs) (Bach et al., 2013), which are MRFs defined over continuous variables with hinge-loss functions for potentials. A MAP state of a HL-MRF minimizes the sum of the weighted hinge-loss potentials. Instead of solving the local consistency relaxation in the form of objective (1), objective (2) can be rewritten as a minimization of hinge-loss potentials by subtracting each weighted clause $w_j \phi(\boldsymbol{x})$ from its maximum possible value w_j. HL-MRFs support scalable MAP inference via ADMM, so the same inference algorithm can be applied to relaxed MAP inference for logical MRFs. This approach is appealing because it solves the compact form of the local consistency relaxation directly, optimizing over only $\boldsymbol{\mu}$, thus avoiding the expensive explicit representation of high-dimensional factor marginals.

Any of these three approaches can be used to perform relaxed MAP inference and find $\boldsymbol{\mu}^\star$, the optimal variable pseudomarginals. If $\boldsymbol{\mu}^\star$ is not integral, i.e., the relaxation is not tight, then we apply the rounding procedure of Goemans and Williamson (1994), as discussed in Section 2.4. We let $p_i = f(\mu_i^\star)$ and apply the method of conditional probabilities. This technique combines the fast inference of message-passing algorithms with the rounding guarantees of MAX SAT relaxations for high solution quality.

5 EVALUATION

In this section, we compare our proposed approach to approximate MAP inference with coordinate descent DD, a popular approach to which rounding procedures cannot be applied (because it does not find a primal solution $\boldsymbol{\mu}^\star$). We show that our proposed technique of combining the rounding procedure with message-passing algorithms can significantly improve the quality of approximate inference. We refer to our technique as rounded linear programming or rounded LP. We compare rounded LP with MPLP (Globerson and Jaakkola, 2007), which is a state-of-the-art coordinate

Table 1: Average sizes for each group of MAP tasks.

Group	Target Users	Variables	Clauses
1	10,000	10,019	214,163
2	20,000	20,037	446,109
3	30,000	30,055	685,415
4	40,000	40,073	924,082
5	50,000	50,091	1,156,125

descent DD algorithm. Recent work, e.g., Jojic et al. (2010) and Meshi and Globerson (2011), notes that MPLP often finds the best discrete, primal solutions.

We evaluate rounded LP and MPLP on randomly generated social network analysis problems, in which the task is to predict whether users share the same political ideology, e.g., liberal or conservative. The networks are composed of upvote and downvote edges, representing whether each user liked or disliked something another user said. We assume that we have some attribute information about each user, summarized in an ideology score uniformly distributed in the $[-1, 1]$ interval. This score could be the output of a classifier or an aggregate over features. It represents *local* information about each user, which the model considers in conjunction with the structure of the interactions.

We generate networks based on a procedure of Broecheler et al. (2010). For a target number of users, in-degrees and out-degrees d for each edge type are sampled from the power-law distribution $P(d) \propto 3 \cdot d^{2.5}$. Edges are created by randomly matching nodes until no more can be added. The number of users is initially the target number plus the expected number of users with zero edges, and then users without any edges are removed. We generated 25 such networks in five groups. Table 1 lists the groups, target numbers of users, and the average numbers of variables and clauses in the corresponding MAP tasks, which is determined by the networks' structures.

We construct a logical MRF for each network to model user ideology. We describe the MRFs in terms of their sets of weighted clauses $\boldsymbol{C}$. Associate with each user X in the network a Boolean variable and arbitrarily associate the true state with a liberal ideology and the false state with a conservative ideology. We refer to each such variable as $\textsc{Liberal}(X)$. If the sign of the ideology score is positive then we add to $\boldsymbol{C}$ the clause

$$w_{\sim[0,1]} \quad : \quad \textsc{Liberal}(X)$$

and if its sign is negative we add the clause

$$w_{\sim[0,1]} \quad : \quad \neg\textsc{Liberal}(X) \ .$$

In either case, each clause is weighted with the magnitude of the ideology score. For each upvote edge from user X_1 to X_2 we add the clauses

$$1.0 \quad : \quad \neg\textsc{Liberal}(X_1) \vee \textsc{Liberal}(X_2)$$
$$1.0 \quad : \quad \textsc{Liberal}(X_1) \vee \neg\textsc{Liberal}(X_2) \ ,$$

enforcing a preference for agreeing ideology, and for each downvote edge we add the clauses

$$1.0 \quad : \quad \textsc{Liberal}(X_1) \vee \textsc{Liberal}(X_2)$$
$$1.0 \quad : \quad \neg\textsc{Liberal}(X_1) \vee \neg\textsc{Liberal}(X_2) \ ,$$

enforcing a preference for disagreeing ideology. While these models are motivated by social network analysis, they are of a similar form to many other problems and domains involving collective classification with attractive and repulsive dependencies.

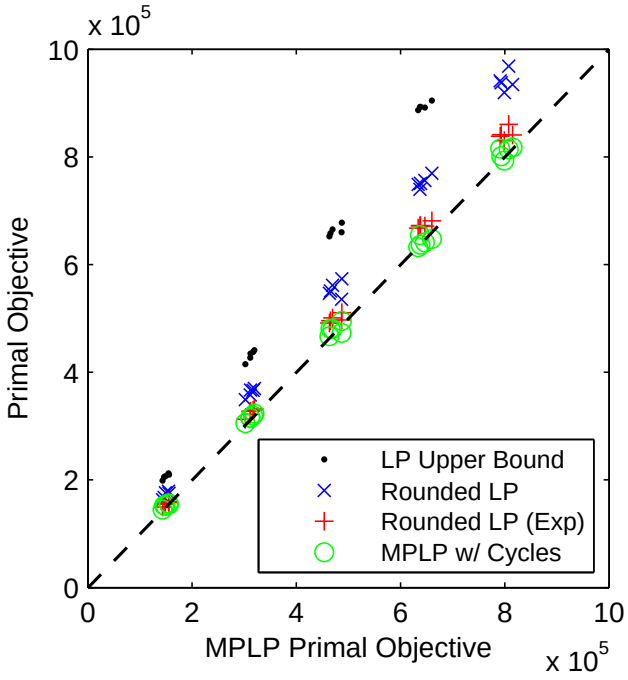

Figure 1: Primal objective scores relative to MPLP.

For each of the 25 logical MRFs, we performed approximate MAP inference using rounded LP and MPLP. For rounded LP, we solved the local consistency relaxation as a HL-MRF, as described in Section 4. We measured the initial linear program objective score ("LP Upper Bound"), which is an upper bound on any discrete primal solution, the expected score $\hat{W}$ (3) using $p_i = f(\mu_i^\star)$ ("Rounded LP (Exp)"), and the final score after rounding using the method of conditional probabilities ("Rounded LP"), as described in Section 2.4. For MPLP, we used the implementation of Globerson et al. (2012) with default settings. We measured the results of both the first-order local consistency relaxation ("MPLP") and iterative cycle tightening ("MPLP w/ Cycles") (Sontag et al., 2008, 2012),

which searches for tighter relaxations to use. The results are summarized in Figure 1, and provided in the supplementary material. All differences in scores between the ten pairs of methods, e.g., "Rounded LP (Exp)" and "MPLP w/ Cycles," are statistically significant using a paired t-test with rejection threshold $p < 0.001$, except "MPLP" and "MPLP w/ Cycles."

On these problems, rounded LP always outperforms MPLP. It finds solutions that are better in expectation than MPLP's solutions, and those solutions are improved further after rounding. What makes these problems difficult is that each pair of clauses for either an upvote or downvote edge is a supermodular potential or submodular potential, respectively. The first-order local consistency relaxation would be tight for a completely supermodular problem (Wainwright and Jordan, 2008), but this mix of potentials makes the problems hard to solve. We found (in experiments not shown) that MPLP's relative performance improves on problems that have many more supermodular potentials than submodular ones, presumably because they are very close to polynomial-time solvable problems. Cycle tightening improves the performance of MPLP, but its impact is limited because there are so many frustrated cycles in these problems. Rounded LP is highly scalable, taking only a minute to solve problems with over one million clauses. Our experiments demonstrate tangible consequences of the approximation guarantee for rounded LP.

6 RELATED WORK

In addition to the various approaches discussed in Section 4, other approaches to approximating MAP inference include tighter linear programming relaxations (Sontag et al., 2008, 2012). These tighter relaxations enforce local consistency on variable subsets that are larger than individual variables, which makes them *higher-order local consistency relaxations*. Mezuman et al. (2013) developed techniques for special cases of higher-order relaxations, such as when the MRF contains cardinality potentials, in which the probability of a configuration depends on the number of variables in a particular state. Some papers have also explored nonlinear convex programming relaxations, e.g., Ravikumar and Lafferty (2006) and Kumar et al. (2006).

Previous analyses have identified particular subclasses whose local consistency relaxations are tight, i.e., the maximum of the relaxed program is exactly the maximum of the original problem. These special classes include graphical models with tree-structured dependencies, binary pairwise models with supermodular potential functions, models encoding bipartite matching problems, and those with *nand* potentials and perfect graph structures (Wainwright and Jordan, 2008; Schrijver, 2003; Jebara, 2009; Foulds et al., 2011). These tightness guarantees are powerful but require more restrictive conditions than our analysis.

While MAP inference is hard to approximate in general (Abdelbar and Hedetniemi, 1998; Park, 2002b), researchers have studied performance guarantees of the first-order local consistency relaxation for special cases. Kleinberg and Tardos (2002) and Chekuri et al. (2005) considered the metric labeling problem. Feldman et al. (2005) used the local consistency relaxation to decode binary linear codes. Our work provides performance guarantees for approximate MAP inference for a new class of models, logical MRFs.

Finally, Huynh and Mooney (2009) introduced a linear programming relaxation for Markov logic (Richardson and Domingos, 2006) inspired by MAX SAT relaxations. Markov logic subsumes logical MRFs, but the relaxation of general Markov logic comes with no known guarantees on the quality of solutions.

7 CONCLUSION

In this paper, we proved that the first-order local consistency relaxation and the MAX SAT relaxation of Goemans and Williamson (1994) are equivalent for logical MRFs. This result is important because the local consistency relaxation can first be solved with any of a number of highly scalable message-passing algorithms, and the output can then be rounded to a discrete solution of guaranteed high quality. We demonstrated this technique by comparing it with coordinate descent DD, showing that applying the guaranteed rounding procedure leads to higher solution quality. Directions for future work include applying our hierarchical optimization analysis to other cases and examining whether rounding guarantees can be developed for higher-order relaxations.

Acknowledgements

The authors would like to thank the anonymous reviewers and several readers, including Kevin Murphy and Ofer Meshi, who helped improve this manuscript. This work was supported by NSF grant IIS1218488, and IARPA via DoI/NBC contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

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

N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley-Interscience, third edition, 2008.

K. M. Anstreicher and L. A. Wolsey. Two "well-known" properties of subgradient optimization. *Mathematical Programming*, 120(1):213–220, 2009.

S. H. Bach, B. Huang, B. London, and L. Getoor. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers*. Now Publishers, 2011.

M. Broecheler, P. Shakarian, and V. S. Subrahmanian. A scalable framework for modeling competitive diffusion in social networks. In *Social Computing (SocialCom)*, 2010.

C. Chekuri, S. Khanna, J. Naor, and L. Zosin. A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM J. Discrete Math.*, 18(3):608–625, 2005.

J. Feldman, M. J. Wainwright, and D. R. Karger. Using linear programming to decode binary linear codes. *Information Theory, IEEE Trans. on*, 51(3): 954–972, 2005.

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

M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete graph problems. *Theoretical Computer Science*, 1(3):237–267, 1976.

A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

A. Globerson, D. Sontag, D. K. Choe, and Y. Li. MPLP, 2012. URL `http://cs.nyu.edu/~dsontag/code/mplp_ver2.tgz`.

M. X. Goemans and D. P. Williamson. New 3/4-approximation algorithms for the maximum satisfiability problem. *SIAM J. Discrete Math.*, 7(4): 656–666, 1994.

T. Huynh and R. Mooney. Max-margin weight learning for Markov logic networks. In *European Conference on Machine Learning (ECML)*, 2009.

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

V. Jojic, S. Gould, and D. Koller. Accelerated dual decomposition for MAP inference. In *International Conference on Machine Learning (ICML)*, 2010.

W. Karush. Minima of Functions of Several Variables with Inequalities as Side Constraints. Master's thesis, University of Chicago, 1939.

J. Kleinberg and É. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. *J. ACM*, 49(5):616–639, 2002.

V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 28(10): 1568–1583, 2006.

N. Komodakis, N. Paragios, and G. Tziritas. MRF energy minimization and beyond via dual decomposition. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 33(3):531–552, 2011.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Berkeley Symp. on Math. Statist. and Prob.*, 1951.

M. P. Kumar, P. H. S. Torr, and A. Zisserman. Solving Markov random fields using second order cone programming relaxations. In *Computer Vision and Pattern Recognition (CVPR)*, 2006.

A. Martins, M. Figueiredo, P. Aguiar, N. Smith, and E. Xing. An augmented Lagrangian approach to constrained MAP inference. In *International Conference on Machine Learning (ICML)*, 2011.

O. Meshi and A. Globerson. An alternating direction method for dual MAP LP relaxation. In *European Conference on Machine learning (ECML)*, 2011.

E. Mezuman, D. Tarlow, A. Globerson, and Y. Weiss. Tighter linear program relaxations for high order graphical models. In *Uncertainty in Aritificial Intelligence (UAI)*, 2013.

A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM J. Optimization*, 19(4):1757–1780, 2009.

J. D. Park. Using weighted MAX-SAT engines to solve MPE. In *National Conference on Artificial Intelligence (AAAI)*, 2002a.

J. D. Park. MAP complexity results and approximation methods. In *Uncertainty in Artificial Intelligence (UAI)*, 2002b.

P. Ravikumar and J. Lafferty. Quadratic programming relaxations for metric labeling and Markov random field MAP estimation. In *International Conference on Machine Learning (ICML)*, 2006.

P. Ravikumar, A. Agarwal, and M. J. Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *J. Mach. Learn. Res.*, 11:1043–1080, 2010.

M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer-Verlag, 2003.

A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Globally convergent dual MAP LP relaxation solvers using Fenchel-Young margins. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

S. E. Shimony. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence*, 68(2):399–410, 1994.

N. Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer-Verlag, 1985.

D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *Uncertainty in Aritificial Intelligence (UAI)*, 2008.

D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, pages 219–254. MIT Press, 2011.

D. Sontag, D. K. Choe, and Y. Li. Efficiently searching for frustrated cycles in MAP inference. In *Uncertainty in Aritificial Intelligence (UAI)*, 2012.

M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, 2008.

T. Werner. A linear programming approach to max-sum problem: A review. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 29(7):1165–1179, 2007.

C. Yanover, T. Meltzer, and Y. Weiss. Linear programming relaxations and belief propagation – An empirical study. *J. Mach. Learn. Res.*, 7:1887–1907, 2006.