Tightness Results for Local Consistency Relaxations in Continuous MRFs

Yoav Wald **Amir Globerson**

School of Computer Science and Engineering
The Hebrew University, Jerusalem 91904, Israel

Abstract

Finding the MAP assignment in graphical models is a challenging task that generally requires approximations. One popular approximation approach is to use linear programming relaxations that enforce local consistency. While these are commonly used for discrete variable models, they are much less understood for models with continuous variables.

Here we define local consistency relaxations of MAP for continuous pairwise Markov Random Fields (MRFs), and analyze their properties. We begin by providing a characterization of models for which this relaxation is tight. These turn out to be models that can be reparameterized as a sum of local convex functions. We also provide a simple formulation of this relaxation for Gaussian MRFs.

Next, we show how the above insights can be used to obtain optimality certificates for loopy belief propagation (LBP) in such models. Specifically, we show that the messages of LBP can be used to calculate upper and lower bounds on the MAP value, and that these bounds coincide at convergence, yielding a natural stopping criterion which was not previously available.

Finally, our results illustrate a close connection between local consistency relaxations of MAP and LBP. They demonstrate that in the continuous case, whenever LBP is provably optimal so is the local consistency relaxation.

1 INTRODUCTION

Graphical models [13] have become a key tool for describing multivariate distributions. For many models of interest, the basic inference task of finding the most likely assignment (also known as the MAP assignment) is computationally hard [26], and one must resort to approximations.

When the model variables are discrete, a popular approximation scheme is linear programming relaxations (LPR). These approximate the MAP problem via minimization of a linear function over locally consistent *pseudo-marginals*. LPRs have several advantages: they provide optimality certificates, they can be optimized via message passing algorithms, they work well in practice, and they are provably exact in some cases (e.g., binary attractive models and trees) [8, 27, 28, 36, 11].

For models with continuous variables, it is less clear how to apply the local consistency perspective of LPRs. For example, the *psuedo-marginals* now become functions rather than a discrete set of variables. Moreover, the standard consistency constraints translate into a continuum of constraints. The goal of the current paper is to study such relaxations and understand when they are tight.

Another commonly used approximate inference algorithm is loopy belief propagation (LBP)[37]. It works by passing messages along the graph, in a manner motivated by variable elimination on tree structured models. Although LBP and LPR are generally distinct algorithms, there are cases where both are exact. For example, both yield the exact MAP for tree models, maximum weight matching [3, 24] and a few other problems (see Section 8). However, there is still no general result linking LPR and LBP. Since LBP in the continuous case is fairly well understood [16, 17], we will want to link our results to known results on LBP.

We begin by defining local consistency MAP relaxations for continuous models. Technically, these will be constructed in an analogous way to the LP relaxations for discrete variables. However, they will typically not correspond to standard linear programs and thus we refer to them as local consistency relaxations (LCRs).

We obtain several surprising results on LCRs. For simplicity of presentation we focus on pairwise MRFs, but extensions to larger cliques are possible (see Section 7). Our first key result is to show that the LCR of a model is tight if the model is "convex decomposable" (CD). A model is CD if it can be expressed as a sum of pairwise convex functions (see [17] for a similar definition).

For continuous MRFs there are several cases where LBP is known to converge to the exact MAP. It is thus of interest to relate these to our LCR results in order to further our understanding of the relation between these two approximation schemes. The relation turns out to be simple and interesting. For Gaussian MRFs LBP is provably optimal when the model is walk-summable [16, 18], a property which turns out to be equivalent to convex decomposability of the model. In other words LCR and LBP are both exact on these models. For general MRFs, LBP is provably optimal for models that are scaled diagonally dominant [17]. These turn out to be a strict subset of CD models, suggesting that either LCR is a stronger approximation or that stronger properties of LBP can be shown.

The above results on LCR and the relation to LBP lead to an important practical implication. It turns out we can use these insights to obtain runtime optimality certificates for loopy belief propagation (LBP) in a wide range of models. Specifically, we show that the messages of LBP can be used to calculate upper and lower bounds on the MAP value, and that these bounds coincide at convergence, yielding a natural stopping criterion for LBP, which was not previously available.

2 MAP IN CONTINUOUS MRFS

We begin by recalling MAP inference in graphical models and how LP relaxations are used to obtain approximate solutions of the problem.

Consider n variables $X_1, \dots, X_n$, and a set of singleton and pairwise functions $f_i(x_i), f_{ij}(x_i, x_j)$ where the pairs ij are a set of edges of a graph $G = (V, E)$. Use these to define a function over $\boldsymbol{x} = x_1, \dots, x_n$:

$$F(\boldsymbol{x}) = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j). \tag{1}$$

For now we do not assume anything about the state spaces of the variables. We refer to $F(\boldsymbol{x})$ as an MRF over $X_1, \dots, X_n$ and consider the MAP problem of minimizing $F(\boldsymbol{x})$.[1]

It will be convenient in what follows to express $F(\boldsymbol{x})$ as a linear function of certain functions of $\boldsymbol{x}$ (our notation follows that of [33]). Assume there exists a vector of functions $\boldsymbol{\phi}_i(x_i)$ of x_i, and a vector $\boldsymbol{\theta}_i$ such that:

$$f_i(x_i) = \langle \boldsymbol{\theta}_i, \boldsymbol{\phi}_i(x_i) \rangle.$$

For example if $\boldsymbol{\phi}_i(x_i) = \left[x_i, x_i^2\right]$ then $f_i(x_i)$ is the quadratic function $f_i(x_i) = \theta_{i,1} x_i + \theta_{i,2} x_i^2$. Similarly assume there exist functions $\boldsymbol{\phi}_{i,j}(x_i, x_j)$ and vectors $\boldsymbol{\theta}_{ij}$ such that:

$$f_{ij}(x_i, x_j) = \langle \boldsymbol{\theta}_{ij}, \boldsymbol{\phi}_{ij}(x_i, x_j) \rangle. \tag{2}$$

Denote the concatenation of all $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{ij}$ by $\boldsymbol{\theta}$ and the concatenation of all $\boldsymbol{\phi}_i, \boldsymbol{\phi}_{ij}$ functions by $\boldsymbol{\phi}$. Furthermore, denote the dimension of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ by m. We can thus write:

$$F(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle, \tag{3}$$

The MAP problem has an equivalent formulation in terms of mean parameters, as shown in [33] and as reviewed next. We define the set of realizable mean parameters:

$$\mathcal{M} = \left\{ \boldsymbol{\mu} \in \mathbb{R}^m \ : \ \exists \hat{p} \in \Delta \text{ s.t } \mathbb{E}_{\hat{p}} \left[\boldsymbol{\phi}(\boldsymbol{x}) \right] = \boldsymbol{\mu} \right\},$$

where Δ is the set of densities over $\boldsymbol{x}$. It can be then shown that the MAP problem corresponds to optimization of a linear function over the set $\mathcal{M}$.

Theorem 2.1. *[33] For the MRF as defined above and the corresponding $\mathcal{M}$ it holds that* [2]*:*

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) = \min_{\boldsymbol{\mu} \in \mathcal{M}} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle. \tag{4}$$

The problem in Eq. (4) has a linear objective over m variables. m is usually not much larger than n, yet the definition of $\mathcal{M}$ involves variables corresponding to densities and thus is generally hard to characterize explicitly. However, there are continuous variable cases where $\mathcal{M}$ does have a compact form. For example, in Gaussian MRFs $\mathcal{M}$ can be expressed via positive semi definiteness constraints (see [33, sec. 3.4.1]).

Finally, we recall the definition of a reparameterization of an MRF.

Definition 1. *We call any set of functions $\bar{f}_i(x_i), \bar{f}_{ij}(x_i, x_j)$ a reparameterization of $F(\boldsymbol{x})$ if it holds that for every $\boldsymbol{x}$:*

$$F(\boldsymbol{x}) = \sum_i \bar{f}_i(x_i) + \sum_{ij} \bar{f}_{ij}(x_i, x_j). \tag{5}$$

3 LOCAL CONSISTENCY RELAXATIONS

Optimizing over the set $\mathcal{M}$ is generally hard. When X are discrete variables $\mathcal{M}$ will involve an exponential number of inequalities. When X are continuous, even in cases where $\mathcal{M}$ is tractable to optimize over, this optimization may be costly (e.g., a semidefinite program). This has prompted considerable research on relaxations of $\mathcal{M}$ and the resulting optimization problem. Most of the work in this context has focused on discrete variables as reviewed next. Our goal is then to extend this framework to the continuous case.

[1] We refer to this as the MAP problem since it corresponds to finding the maximum a posteriori assignment to $\boldsymbol{x}$ in the model $p(\boldsymbol{x}) \propto \exp(-F(\boldsymbol{x}))$.

[2] We note that the right hand side of Eq. (4) should have $\bar{\mathcal{M}}$ instead of $\mathcal{M}$. The closure is omitted for simplicity of presentation.

3.1 LCR FOR DISCRETE VARIABLES

Consider the case where X_i are discrete variables, each with D values, and the functions $\boldsymbol{\phi}$ are defined as follows. $\boldsymbol{\phi}_i(x_i)$ is D dimensional with $\phi_{i,k}(x_i) = \mathcal{I}(x_i = k)$, and $\boldsymbol{\phi}_{ij}(x_i, x_j)$ is $D \times D$ dimensional with $\phi_{ij,kl}(x_i, x_j) = \mathcal{I}(x_i = k, x_j = l)$. The expected values $\mathbb{E}_{\hat{p}}[\boldsymbol{\phi}_i(x_i)]$ and $\mathbb{E}_{\hat{p}}\left[\boldsymbol{\phi}_{ij}(x_i, x_j)\right]$ are simply the singleton marginals $\hat{p}(x_i)$ and $\hat{p}(x_i, x_j)$ respectively. Thus, $\mathcal{M}$ corresponds to the set of all singleton and pairwise marginals that are achieved by some distribution $\hat{p}(\boldsymbol{x})$. This is also known as the marginal polytope [33].

As mentioned earlier, the marginal polytope generally requires an exponential number of inequalities to describe. One natural alternative is to consider a *local consistency relaxation* (LCR) where instead of requiring the marginals to come from a "global" distribution $\hat{p}(\boldsymbol{x})$ we just require the singleton and pairwise marginals to be consistent. In other words, we define the set $\mathcal{M}_L$ as the set of locally consistent pairwise marginals:

$$\mathcal{M}_L = \left\{ \boldsymbol{\mu} \geq 0 : \begin{array}{ll} \sum_{x_i} \mu_i(x_i) = 1 & \\ \sum_{x_j} \mu_{ij}(x_i, x_j) = \mu_i(x_i) & \forall ij, x_i. \end{array} \right\}$$

The local relaxation of the MAP problem is then to minimize $\boldsymbol{\mu} \cdot \boldsymbol{\theta}$ over $\boldsymbol{\mu} \in \mathcal{M}_L$ instead of $\boldsymbol{\mu} \in \mathcal{M}$. We next consider the extension of this relaxation to the continuous variable case.

3.2 LCR FOR CONTINUOUS VARIABLES

For continuous variables, a natural extension of the above is to replace the sum in the local consistency constraints with an integral:

$$\mathcal{M}_L = \left\{ \boldsymbol{\mu} : \begin{array}{ll} \exists \text{ densities } \hat{p}_i, \hat{p}_{ij} \text{ s.t.} & \\ \int \hat{p}_{ij}(x_i, x_j) dx_j = \hat{p}_i(x_i) & \forall ij, x_i \\ \mathbb{E}_{\hat{p}_i}[\boldsymbol{\phi}_i(x_i)] = \boldsymbol{\mu}_i & \forall i \\ \mathbb{E}_{\hat{p}_{ij}}\left[\boldsymbol{\phi}_{ij}(x_i, x_j)\right] = \boldsymbol{\mu}_{ij} & \forall ij. \end{array} \right\} \tag{6}$$

In other words we consider all pairwise consistent densities, and $\mathcal{M}_L$ are all the expected values obtained from such densities. As in the discrete case the local relaxation of MAP is then the problem:

$$\min_{\boldsymbol{\mu} \in \mathcal{M}_L} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle. \tag{7}$$

Note that because $\mathcal{M} \subseteq \mathcal{M}_L$, the above minimum is a lower bound on the true MAP value (as in the case of discrete models [29]).

The key question we ask here is: for which cases is the relaxation in Eq. (7) tight? Before presenting our main result (Thm. 4.1), we will define an even looser relaxation in the next section which will be important for our analysis.

3.3 WEAK LCR AND ITS DUAL

In the constraints of Eq. (6) we require the singleton and pairwise marginals to be completely consistent. Namely, that $\hat{p}_{ij}(x_i)$, the marginal density calculated from $\hat{p}_{ij}(x_i, x_j)$ will equal $\hat{p}_i(x_i)$ for *all* x_i values. A weaker consistency constraint is to enforce that $\hat{p}_{ij}(x_j)$ and $\hat{p}_i(x_i)$ agree only on certain expected values. We next define the set corresponding to this constraint.

Given a vector of functions $\boldsymbol{\psi}_i(x_i)$ for each $i \in V$,[3] we define the set:

$$\mathcal{M}_L^{\psi} = \left\{ \boldsymbol{\mu} : \begin{array}{ll} \exists \text{ densities } \hat{p}_i, \hat{p}_{ij} \text{ s.t.} & \\ \mathbb{E}_{\hat{p}_{ij}}[\boldsymbol{\psi}_i(x_i)] = \mathbb{E}_{\hat{p}_i}[\boldsymbol{\psi}_i(x_i)] & \forall ij \\ \mathbb{E}_{\hat{p}_i}[\boldsymbol{\phi}_i(x_i)] = \boldsymbol{\mu}_i & \forall i \\ \mathbb{E}_{\hat{p}_{ij}}\left[\boldsymbol{\phi}_{ij}(x_i, x_j)\right] = \boldsymbol{\mu}_{ij} & \forall ij. \end{array} \right\}$$

Thus $\mathcal{M}_L^{\boldsymbol{\psi}}$ enforces consistency only in the sense that $\hat{p}_{ij}(x_j)$ and $\hat{p}_i(x_i)$ agree on the expected values of $\boldsymbol{\psi}_i$. Thus, for any choice of $\boldsymbol{\psi}$ we have $\mathcal{M}_L \subseteq \mathcal{M}_L^{\psi}$.

The corresponding relaxation is then defined as:

$$\min_{\boldsymbol{\mu} \in \mathcal{M}_L^{\psi}} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle. \tag{8}$$

Basic relations between the relaxations we consider are thus summarized by the inequalities:

$$\min_{\boldsymbol{\mu} \in \mathcal{M}_L^{\psi}} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \leq \min_{\boldsymbol{\mu} \in \mathcal{M}_L} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \leq \min_{\boldsymbol{\mu} \in \mathcal{M}} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle = \min_{\boldsymbol{x}} F(\boldsymbol{x}).$$

The dual of Eq. (8) will play an important role in subsequent sections:

Lemma 3.1. *Given a pairwise MRF with functions $f_i(x_i), f_{ij}(x_i, x_j)$, the following is a dual problem of Eq. (8):*

$$\max_{\boldsymbol{\delta}} \sum_i \min_{x_i} \left\{ f_i(x_i) + \sum_{j \in N(i)} \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\psi}_i(x_i) \rangle \right\} + \tag{9}$$
$$\sum_{ij} \min_{x_i, x_j} \left\{ f_{ij}(x_i, x_j) - \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\psi}_i(x_i) \rangle - \langle \boldsymbol{\delta}_{ij}, \boldsymbol{\psi}_j(x_j) \rangle \right\}.$$

We note that the bound achieved by this dual, and hence by weak LCR, is dependent on the reparameterization $f_i(x_i), f_{ij}(x_i, x_j)$ being used. The full LCR Eq. (7) may also yield different optimization problems for different reparameterizations, yet it turns out that the bound it achieves is invariant to the reparameterization. The proof of the following lemma can be found in the supplementary material.

Lemma 3.2. *LCR has the same value under any reparameterization $\{f_i(x_i), f_{ij}(x_i, x_j)\}$ of $F(x)$.*

[3] For example, $\boldsymbol{\psi}_i(x_i)$ could be $\left[x_i, x_i^3\right]$.

4 TIGHTNESS OF LCR ON CONVEX DECOMPOSABLE MODELS

We are now ready to provide our main result on the tightness of LCRs. We first recall the definition of Convex Decomposable (CD) models, as introduced by [17][4] in the context of LBP analysis. Next, we show that in these models the local consistency relaxation in Eq. (7) is exact.

Definition 2. *A pairwise MRF is convex decomposable if there exists a reparameterization $\bar{f}_i(x_i), \bar{f}_{ij}(x_i, x_j)$ of $F(\boldsymbol{x})$ such that all the functions in the reparameterization are convex.*

Our result regarding tightness of LCR on CD MRFs is stated in the following theorem:

Theorem 4.1. *For any CD pairwise MRF it holds that*

$$\min_{\boldsymbol{\mu} \in \mathcal{M}_L} \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle = \min_{\boldsymbol{x}} F(\boldsymbol{x}).$$

The proof is provided in the appendix. It relies on the following insights:

- From Lem. 3.2, the dual of weak LCR Eq. (9) taken with respect to any reparameterization of $F(\boldsymbol{x})$ is a lower bound on LCR's value.
- Specifically, we can consider Eq. (9) with $\boldsymbol{\psi}_i = [x_i]$ when the reparameterization $f_i(x_i), f_{ij}(x_i, x_j)$ contains convex functions.
- Assuming $\boldsymbol{x}^*$ is a MAP assignment, the dual assignment $\boldsymbol{\delta}^*$ obtained by setting multipliers according to a subgradient of the pairwise functions $f_{ij}(x_i, x_j)$ taken at $\boldsymbol{x}^*$, achieves a dual objective of $F(\boldsymbol{x}^*)$ in Eq. (9).
- LCR's value cannot be lower than weak LCR's value, which in turn cannot be lower than $F(\boldsymbol{x}^*)$ because we constructed a dual assignment that achieves this objective. LCR is a lower bound on MAP, so that equality must hold.

4.1 RELATION TO LOOPY BP

In [17] it is shown that loopy BP is exact (i.e., guaranteed to converge to the true MAP assignment) on a class of models that satisfy scaled diagonal dominance. These are a strict subset of CD models. In particular, they require that a model is CD as well as its Hessian being scaled diagonally dominant. That is, there must exist a strictly positive vector $\boldsymbol{w} \in \mathbb{R}^n$ and a $0 < \lambda < 1$ such that for any $\boldsymbol{x}$ it holds:

$$\sum_{j \in N(i)} w_j \left| \frac{\partial^2}{\partial x_i \partial x_j} F(\boldsymbol{x}) \right| \le \lambda w_i \frac{\partial^2}{\partial x_i^2} F(\boldsymbol{x}).$$

[4] We note that the definition in [17] poses the additional demands of differentiability and that $f_i(x_i)$ are strictly convex, so the class we define here is a larger one.

Thus, we conclude that given currently known exactness results on LBP, the LCR approximation is stronger in the sense that it is exact whenever LBP is known to be exact. It remains an open question whether it can be shown that LBP is exact on CD models.

5 GAUSSIAN MRFS AND THEIR RELAXATION

Gaussian MRFs (GMRFs) have been studied widely, and are also of practical interest [5, 6, 16, 18, 34]. Here we give a simple characterization of $\mathcal{M}_L$ for GMRFs, and then discuss when the corresponding relaxations are tight. We give a stronger result than Theorem 4.1 by showing that LCR is tight *if and only if* the Gaussian model is CD. Specifically, we show that for non CD models the LCR optimization problem has a value of $-\infty$, and is thus a meaningless relaxation.

Recall that a GMRF $F(\boldsymbol{x})$ is a quadratic form:

$$F(\boldsymbol{x}) = \frac{1}{2} x^\top \Gamma x - h^\top x,$$

where $\Gamma \succeq 0$. The function vectors $\boldsymbol{\phi}_i(x_i), \boldsymbol{\phi}_{ij}(x_i, x_j)$ for this type of MRF are given by

$$\boldsymbol{\phi}_i(x_i) = \{x_i, x_i^2\}, \boldsymbol{\phi}_{ij}(x_i, x_j) = \{x_i x_j\}. \quad (10)$$

As stated in [33], the set of realizable mean parameters $\mathcal{M}$ in this case is the set of all first and second moments that can be realized by a density $\hat{p}$. Namely:

$$\mathcal{M} = \left\{ (\Sigma, \boldsymbol{\eta}) \;:\; \Sigma - \boldsymbol{\eta}\boldsymbol{\eta}^\top \succeq 0 \right\}. \quad (11)$$

Here the elements Σ_{ij} correspond to the expected values of $x_i x_j$, the diagonal elements Σ_{ii} are the expected values of x_i^2 and $\boldsymbol{\eta}$ are the expected values of x_i.

5.1 A CHARACTERIZATION OF $\mathcal{M}_L$ FOR GMRFS

Given this simple form of $\mathcal{M}$, it is interesting to see what $\mathcal{M}_L$ translates into for this case. To characterize $\mathcal{M}_L$ we define:

Definition 3. *Given an $n \times n$ matrix A, and an edge $(i, j) \in E$, define the following sub matrix of A*

$$A_{[ij]} = \begin{bmatrix} A_{ii} & A_{ij} \\ A_{ji} & A_{jj} \end{bmatrix}.$$

Similarly, for an n dimensional vector v, define

$$v_{[ij]} = \begin{bmatrix} v_i \\ v_j \end{bmatrix}.$$

Claim 5.1. *For GMRFs it holds that*

$$\mathcal{M}_L = \{ (\Sigma, \boldsymbol{\eta}) \;:\; \Sigma_{[ij]} - \boldsymbol{\eta}_{[ij]} \boldsymbol{\eta}_{[ij]}^\top \succeq 0 \;\forall ij \in E \}. \quad (12)$$

The proof can be found in the appendix. The characterization is very intuitive: it says that Σ should be such that its pairwise submatrices are valid covariance matrices of two variables with mean given by $\boldsymbol{\eta}$.

5.1.1 LCR is Unbounded for non-CD GMRFs

We now turn to give a full characterization of the GMRFs on which LCR is tight. From Thm. 4.1 we know that whenever a GMRF is convex decomposable then LCR is tight, yet we do not know if tightness holds for any other GMRFs. It turns out that it does not, and in fact the objective of LCR is unbounded for non CD models. The proof relies on the following two claims. First we claim that for GMRFs, under a certain choice of $\boldsymbol{\psi}_i(x_i)$, $\mathcal{M}_L^{\psi}$ as given for the weak LCR in Eq. (8) is equal to $\mathcal{M}_L$:

Claim 5.2. *Let $\mathcal{M}_L$ be the set defined in Eq. (6) with the functions in Eq. (10). For $\boldsymbol{\psi}_i(x_i) = [x_i, x_i^2]$ it holds that:*

$$\mathcal{M}_L = \mathcal{M}_L^{\boldsymbol{\psi}}. \tag{13}$$

From Claim 5.2 we gather that weak LCR achieves the same bound as LCR, and that Eq. (9) is also a dual of LCR. The second claim then states that Eq. (9) is unbounded when the GMRF is not CD.

Claim 5.3. *For any non-CD GMRF, the value of Eq. (9) taken with $\boldsymbol{\psi}_i(x_i) = [x_i, x_i^2]$ is $-\infty$ for any choice of $\boldsymbol{\delta}$.*

The proof of both claims is provided in the appendix. It is easy to check that Slater's conditions hold for LCR as defined by Eq. (12) (see e.g. [4, p. 523] for the conditions). Thus when the optimum of LCR is bounded (i.e. larger than $-\infty$), strong duality holds and Eq. (9) obtains the same optimal value. Claim 5.3 states that this does not happen for non-CD GMRFs. We have thus shown the following corollary:

Corollary 5.4. *LCR is unbounded on any non-CD GMRF.*

5.2 RELATION TO LOOPY BP

Several works have analyzed the properties of LBP when applied to GMRFs [16, 34, 18]. These show that when a model is CD then LBP is exact.[5] Note that the definition of convex decomposability used in [18] is somewhat more restrictive than what we consider here. It requires the pairwise functions $f_{ij}(x_i, x_j)$ in the convex decomposition to be strictly convex.

Given our results above, we conclude that LCR is exact precisely on the models where LBP is known to be exact. That is with the subtle exception of cases where the pairwise functions in the convex decomposition are not strictly convex, and then LBP is not known to be exact.

[5]As stated in Section 1, there are other characterizations of these models, e.g. walk summability, which turn out to be equivalent to convex decomposability.

6 A STOPPING CRITERION FOR LBP

In this section we highlight a practical application resulting from the close LCR and LBP connection. Since LBP does not optimize a clearly defined objective, it is hard to monitor its convergence. Below we show how in many cases, simple upper and lower bounds can be calculated for LBP.

Consider an MRF where f_{ij}, f_i are convex and LBP is known to be exact. We will provide upper and lower bounds on the MAP objective in this case and show that they are tight at convergence. These bounds can thus be used as a reliable, easy to apply stopping criterion for LBP in these cases.[6]

Given our assumption on the MRF, it follows that it is CD. Thus, the LCR is tight, and the maximum of the dual objective Eq. (9) will equal the MAP value. This leads us to the following bounding scheme. At each iteration of BP, calculate an estimate of the MAP and denote it by $\boldsymbol{x}^t$. Clearly $F(\boldsymbol{x}^t)$ is an upper bound on the MAP value.

The procedure for obtaining the lower bound is more involved, although technically simple. It is described (along with the upper bound) in Algorithm 1. First, it is easy to see that d in Algorithm 1 is a lower bound since it is a value of the dual in Eq. (9). Second, this bound is tight when $\boldsymbol{x}^t$ is the MAP assignment, as the following result states:

Lemma 6.1. *Assume $\boldsymbol{x}^t \in \arg\min_{\boldsymbol{x}} F(\boldsymbol{x})$, then the bound returned by the procedure satisfies $p = d$.*

Algorithm 1 Bounding Scheme for Sum of Convex Functions

Input: MAP estimate $\boldsymbol{x}^t$ at time t, accuracy parameter ε, convex reparameterization $\{f_i(x_i), f_{ij}(x_i, x_j)\}$.

For each $ij \in E$, find $g \in \partial f_{ij}(x_i^t, x_j^t)$ and set

$$\hat{\delta}_{ji} = g_i, \ \hat{\delta}_{ij} = g_j.$$

Set for each $i \in V$ and $ij \in E$

$$\bar{f}_i(x_i) = f_i(x_i) + \sum_{j \in N(i)} \hat{\delta}_{ji} x_i$$

$$\bar{f}_{ij}(x_i, x_j) = f_{ij}(x_i, x_j) - \hat{\delta}_{ji} x_i - \hat{\delta}_{ij} x_j$$

Calculate primal and dual objectives [7]

$$p = \sum_i f_i(x_i^t) + \sum_{ij} f_{ij}(x_i^t, x_j^t)$$

$$d = \sum_i \min_{x_i} \bar{f}_i(x_i) + \sum_{ij} \min_{x_i, x_j} \bar{f}_{ij}(x_i, x_j)$$

If $p - d < \varepsilon$ then determine convergence.

---

[6]It is possible to apply these bounds to any iterative algorithm that provides a sequence of approximate MAP assignments. We focus on LBP since the bounds are tight in the case we address.

The proof of this lemma follows exactly the same argument given in the third step for the proof of Thm. 4.1. When the singleton functions f_i are strongly convex and the objective is smooth, then it is also possible to give a guarantee on the size of the bound for $\boldsymbol{x}^t$ that is not a MAP assignment (see Lem. B.1 in the supplementary material).

Thus, we have obtained a scheme that provides tight lower and upper bounds on the iterates of BP in a subclass of the cases when it converges. Figure 1 illustrates these bounds for a Gaussian MRF.

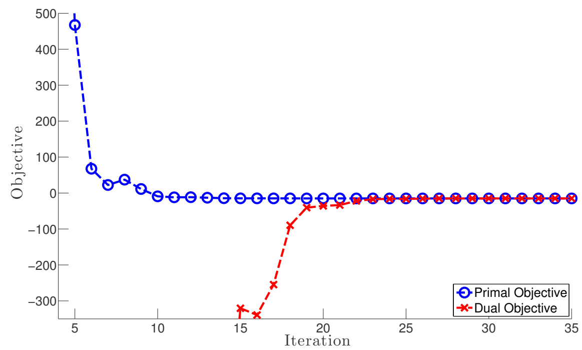

Figure 1: Illustration of the bounds for a 10×10 grid Gaussian MRF. We generate random convex quadratic functions for each node and edge in the model, and use the described bounding scheme to determine LBP's convergence.

7 HIGHER ORDER MODELS

Thus far we only considered pairwise models. In this section we briefly mention the generalization of our results to higher order models. We propose two possible LCR approaches in this context and provide the related tightness results.

A non pairwise MRF may be written as follows:

$$F(\boldsymbol{x}) = \sum_{\alpha \in C} f_\alpha(\boldsymbol{x}_\alpha),$$

where $C \subseteq 2^V$. One option to define $\mathcal{M}_L$ for this case is:

$$\mathcal{M}_L = \left\{ \boldsymbol{\mu} : \begin{array}{ll} \exists \text{ densities } \hat{p}_i, \hat{p}_\alpha \text{ s.t.} & \\ \int \hat{p}_\alpha(\boldsymbol{x}_\alpha) dx_{\alpha \setminus i} = \hat{p}_i(x_i) & \forall \alpha, i \in \alpha, x_i \\ \mathbb{E}_{\hat{p}_\alpha}\left[\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)\right] = \boldsymbol{\mu}_\alpha & \forall \alpha \in C. \end{array} \right\} \tag{14}$$

In other words, $\mathcal{M}_L$ constrains marginals over the α sets to agree on singleton marginals. Another reasonable choice is to define a set $\mathcal{M}_C$ that enforces stronger consistency constraints than $\mathcal{M}_L$:

$$\int \hat{p}_\alpha(\boldsymbol{x}_\alpha) dx_{\alpha \setminus \beta} = \int \hat{p}_\beta(\boldsymbol{x}_\beta) dx_{\beta \setminus \alpha} \quad \forall \alpha, \beta \in C. \tag{15}$$

Instead of consistency over single variables, $\mathcal{M}_C$ enforces consistency on the overlap of pairs $\alpha, \beta \in C$.

Similar derivations to those of the pairwise case lead to tightness characterization for the constraints above. For our tightness results to carry, $F(\boldsymbol{x})$ should be given as a sum of convex functions. The first result states that if the functions $f_\alpha(\boldsymbol{x}_\alpha)$ are all convex, then the $\mathcal{M}_L$ relaxation is tight.

Claim 7.1. *If $F(\boldsymbol{x})$ is given as a sum of convex functions then* $\min_{\boldsymbol{\mu} \in \mathcal{M}_L} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle = \min_{\boldsymbol{x}} F(\boldsymbol{x})$.

In this case we can also use the bounds described in Section 6. An interesting example of such a model is the one underlying the AMP algorithm of Donoho et al. [19].

The relaxation defined by Eq. (15) may be more complicated to solve due to the additional constraints. At the cost of this complication, the obtained relaxation is invariant to reparameterizations and an analogue of Thm. 4.1 holds.[8]

Definition 4. *An MRF is CD w.r.t $C \subseteq 2^V$ if there exists a reparameterization $\{\bar{f}_\alpha(\boldsymbol{x}_\alpha)\}$ for $F(\boldsymbol{x})$ such that all the functions in the reparameterization are convex.*

Claim 7.2. *For any MRF that is CD w.r.t C it holds that* $\min_{\boldsymbol{\mu} \in \mathcal{M}_C} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle = \min_x F(\boldsymbol{x})$.

See supplementary material for proofs of the above results.

8 RELATED WORK

The current paper studies local consistency relaxations as applied to continuous MRFs, and their relation to loopy belief propagation. Below we briefly survey related results on this relation, focusing on discrete variable models where local consistency relaxations are typically considered.

For discrete tree structured MRFs, it can be shown that LBP and LCR are equivalent in the sense that they are both tight, and in fact there is a mapping between dual LCR variables and BP optima (see [33, p. 200]) .

In [35] Weiss *et al.* consider the relation between LCR and convex belief propagation (and *not* standard LBP). They show that if a convex variant of max-product LBP converges, and the sharpened version of its beliefs (where sharpening means to distribute all probability mass evenly between maximizing arguments of the beliefs) are locally consistent, then they are a solution to the LCR.

For maximum weight bipartite matching it is known that LBP is exact [3] and so is the standard LP relaxation of the problem [31] (see also [1, sec. 6] on the relation between

[7] Notice that in order to calculate d it is not required to perform optimization over $\bar{f}_{ij}(x_i, x_j)$. The manner in which we set $\hat{\boldsymbol{\delta}}$ guarantees that the minimum is attained at (x_i^t, x_j^t). The calculation of d will only demand optimizing $\bar{f}_i(x_i)$ for each i.

[8] This result holds under a non-restrictive assumption that $\alpha \cap \beta \in C$ for all $\alpha, \beta \in C$ and also $\{i\} \in C$ for all $i \in V$.

this relaxation and local consistency relaxations). For non bipartite matching a more subtle result is available [24] showing that if the LCR does not have fractional optima then LBP is exact. Otherwise LBP will not be exact. Related results are available for maximum weight independent set [25] and packing and covering problems [7].

For the problem of decoding LDPC codes, both LCR approaches [9] and LBP [21] have been successful. Although some relations between these have been shown [32, 2], a clear link establishing cases where they are both exact has not been provided.

Tarlow *et al.* [30] show that for the min-cut problem, where LCR is known to be tight (e.g., see [31]), a modified version of LBP is exact as well.

Another connection between the LPR and message passing algorithms relies on the notion of graph covers [23, 12]. It is shown that for discrete models, LPR solves the MAP problem on a model that has the lowest objective amongst a family of graphical models who are in some sense isomorphic to the model it tries to approximate. Message passing algorithms are unable to distinguish between these isomorphic models, and thus an intuitive link between LPR and message passing is made.

On GMRFs, both LBP [16, 34, 18] and graph covers [22] were studied. One of the conclusions reached by studying graph covers [22] is that for non-CD GMRFs, dual coordinate ascent algorithms [10, 14, 15] are bound to fail at producing the MAP estimate. Our result for GMRFs provides an arguably simpler explanation of this failure: These algorithms perform dual coordinate ascent on the LCR, and whenever the LCR is inapplicable (as in the case of non-CD GMRFs) they yield a useless bound. See Section 5.2 for further discussion of GMRFs and LBP.

Relaxations for continuous MRFs have been less studied. The most relevant work is [20], who arrived at the dual of Eq. (7) (notice we did not use this dual, rather we used Eq. (9) which is a dual of Eq. (8)) but did not analyze when these are exact. Another recent work [38] suggested a different relaxation for continuous variables, but one that is more involved than standard LPRs and also has no clear connection to BP.

Another work on MAP estimation in continuous MRFs suggested an algorithm called Linear Coordinate Descent [40]. This algorithm was later generalized by the authors [39]. For both versions, the authors do not give an objective over which the updates yield an improvement at each step. In fact, our analysis provides a very simple interpretation of the algorithm in [40]. It essentially performs dual coordinate ascent on the dual of weak LCR (see Eq. (9)). This also leads to a much simpler convergence proof than that given in [40], since coordinate ascent in this case has a unique maximum and therefore converges globally [4].

9 DISCUSSION

We considered the MAP problem for MRFs on continuous variables, and derived a local relaxation for these. For convex decomposable MRFs we showed that these relaxations are in fact exact. For Gaussian MRFs we provided a stronger result showing that convex decomposability is necessary and sufficient for exactness of local relaxations.

Comparing our results to those on exactness of loopy belief propagation we find that local relaxations are exact in a strictly larger class of models. This further strengthens the known relation between LBP and LCR, adding to numerous other models where exactness of local relaxations coincides with exactness of LBP. It also leaves interesting open questions. For example, does BP in fact converge on general CD models or is the scaled diagonal dominance condition of [17] necessary.

We note that it is possible to use the LCR and its dual to derive coordinate ascent algorithms for the dual LCR. We have indeed derived MPLP-*like* algorithms [27] for this case. However, our experiments (not shown here) indicate that they are typically slower than LBP. Thus, LBP remains an attractive option for optimizing such models, and the tight bounds we develop in Section 6 are very useful when using it in practice. An interesting open problem is to devise such bounds for LBP in other models, and to provide general transformations between LCR and LBP solutions.

A Proofs

A.1 Proof of Lem. 3.1

Proof. Let us write $\mathcal{M}_L^{\psi}$ with auxiliary variables $\boldsymbol{\eta}$:

$$\mathcal{M}_L^{\psi} = \left\{ \boldsymbol{\mu}, \boldsymbol{\eta} : \begin{array}{ll} \exists \hat{p}_i, \hat{p}_{ij} \text{ s.t.} & \\ \mathbb{E}_{\hat{p}_i}\left[\boldsymbol{\phi}_i(x_i)\right] = \boldsymbol{\mu}_i & \forall i \\ \mathbb{E}_{\hat{p}_{ij}}\left[\boldsymbol{\phi}_{ij}(x_i, x_j)\right] = \boldsymbol{\mu}_{ij} & \forall ij \\ \mathbb{E}_{\hat{p}_i}\left[\boldsymbol{\psi}_i(x_i)\right] = \boldsymbol{\eta}_i & \forall i \\ \mathbb{E}_{\hat{p}_{ij}}\left[\boldsymbol{\psi}_i(x_i)\right] = \boldsymbol{\eta}_{ji} & \forall i, j \in N(i) \\ \boldsymbol{\eta}_{ji} = \boldsymbol{\eta}_i & \forall i, j \in N(i). \end{array} \right\}$$

Define for each $i \in V, ij \in E$ the sets of realizable mean parameters:

$$\mathcal{M}_i = \left\{ \boldsymbol{\mu}_i, \boldsymbol{\eta}_i : \begin{array}{c} \exists \hat{p}_i \text{ s.t.} \\ \mathbb{E}_{\hat{p}_i}\left[\boldsymbol{\phi}_i(x_i)\right] = \boldsymbol{\mu}_i \\ \mathbb{E}_{\hat{p}_i}\left[\boldsymbol{\psi}_i(x_i)\right] = \boldsymbol{\eta}_i \end{array} \right\},$$

$$\mathcal{M}_{ij} = \left\{ \boldsymbol{\mu}_{ij}, \boldsymbol{\eta}_{ij}, \boldsymbol{\eta}_{ji} : \begin{array}{c} \exists \hat{p}_{ij} \text{ s.t.} \\ \mathbb{E}_{\hat{p}_{ij}}\left[\boldsymbol{\phi}_{ij}(x_i, x_j)\right] = \boldsymbol{\mu}_{ij} \\ \mathbb{E}_{\hat{p}_{ij}}\left[\boldsymbol{\psi}_i(x_i)\right] = \boldsymbol{\eta}_{ji} \\ \mathbb{E}_{\hat{p}_{ij}}\left[\boldsymbol{\psi}_j(x_j)\right] = \boldsymbol{\eta}_{ij}. \end{array} \right\}$$

Then $\mathcal{M}_L^{\psi}$ can now be written compactly as:

$$\mathcal{M}_L^{\psi} = \left\{ \boldsymbol{\mu}, \boldsymbol{\eta} : \begin{array}{ll} (\boldsymbol{\mu}_i, \boldsymbol{\eta}_i) \in \mathcal{M}_i & \forall i \\ (\boldsymbol{\mu}_{ij}, \boldsymbol{\eta}_{ij}, \boldsymbol{\eta}_{ji}) \in \mathcal{M}_{ij} & \forall ij \\ \boldsymbol{\eta}_{ji} = \boldsymbol{\eta}_i & \forall i, j \in N(i). \end{array} \right\}$$

Using the expressions for $f_i(x_i), f_{ij}(x_i, x_j)$ from Eq. (2), our problem is:

$$\min_{(\boldsymbol{\mu},\boldsymbol{\eta}) \in \mathcal{M}_L^{\psi}} \sum_i \langle \boldsymbol{\theta}_i, \boldsymbol{\mu}_i \rangle + \sum_{ij} \langle \boldsymbol{\theta}_{ij}, \boldsymbol{\mu}_{ij} \rangle$$

We now assign a Lagrange multiplier to each consistency constraint

$$\begin{aligned} \boldsymbol{\delta}_{ji} &\leftrightarrow \boldsymbol{\eta}_{ji} = \boldsymbol{\eta}_i \\ \boldsymbol{\delta}_{ij} &\leftrightarrow \boldsymbol{\eta}_{ij} = \boldsymbol{\eta}_j, \end{aligned}$$

and write the resulting Lagrangian:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\mu}) = \sum_i & \left\{ \langle \boldsymbol{\theta}_i, \boldsymbol{\mu}_i \rangle + \sum_{j \in N(i)} \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\eta}_i \rangle \right\} \\ & + \sum_{ij} \left\{ \langle \boldsymbol{\theta}_{ij}, \boldsymbol{\mu}_{ij} \rangle - \langle \boldsymbol{\delta}_{ij}, \boldsymbol{\eta}_{ij} \rangle - \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\eta}_{ji} \rangle \right\}. \end{aligned}$$

To obtain a dual, we first minimize $\mathcal{L}(\boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\eta})$ w.r.t $\boldsymbol{\mu}, \boldsymbol{\eta}$ under the remaining constraints: $(\boldsymbol{\mu}_i, \boldsymbol{\eta}_i) \in \mathcal{M}_i, (\boldsymbol{\mu}_{ij}, \boldsymbol{\eta}_{ij}, \boldsymbol{\eta}_{ji}) \in \mathcal{M}_{ij}$. Since the relevant variables for each constraint all lie in a single summand, we can push the minimization inside the sums:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\delta}) = \sum_i \min_{\boldsymbol{\mu}_i, \boldsymbol{\eta}_i \in \mathcal{M}_i} & \left\{ \langle \boldsymbol{\theta}_i, \boldsymbol{\mu}_i \rangle + \sum_{j \in N(i)} \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\eta}_i \rangle \right\} + \\ \sum_{ij} \min_{\boldsymbol{\mu}_{ij}, \boldsymbol{\eta}_{ij}, \boldsymbol{\eta}_{ji} \in \mathcal{M}_{ij}} & \left\{ \langle \boldsymbol{\theta}_{ij}, \boldsymbol{\mu}_{ij} \rangle - \langle \boldsymbol{\delta}_{ij}, \boldsymbol{\eta}_{ij} \rangle - \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\eta}_{ji} \rangle \right\}. \end{aligned}$$

Each summand includes optimization over the set of realizable mean parameters, thus according to Thm. 2.1 our Lagrangian is given by:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\delta}) = \sum_i \min_{x_i} & \left\{ \langle \boldsymbol{\theta}_i, \boldsymbol{\phi}_i(x_i) \rangle + \sum_{j \in N(i)} \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\psi}_i(x_i) \rangle \right\} + \\ \sum_{ij} \min_{x_i, x_j} & \left\{ \langle \boldsymbol{\theta}_{ij}, \boldsymbol{\phi}_{ij}(x_i, x_j) \rangle \right. \\ & \left. - \langle \boldsymbol{\delta}_{ij}, \boldsymbol{\phi}_{ij}(x_j) \rangle - \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\phi}_{ji}(x_i) \rangle \right\}. \end{aligned}$$

The desired dual is obtained by maximizing over $\boldsymbol{\delta}$. □

A.2 Proof of Thm. 4.1

Proof. Assume $\{f_i(x_i), f_{ij}(x_i, x_j)\}$ is a reparameterization of $F(\boldsymbol{x})$ for which all the functions are convex. According to Lem. 3.2 the minimum of LCR does not depend on the reparameterization, thus it is enough to prove weak LCR's tightness with respect to this decomposition in order to establish LCR's tightness.

Consider weak LCR taken with $\boldsymbol{\psi}_i(x_i) = [x_i]$. In this case the dual problem Eq. (9) is:

$$\begin{aligned} \max_{\boldsymbol{\delta}} \sum_i \min_{x_i} & \left\{ f_i(x_i) + \sum_{j \in N(i)} \delta_{ji} x_i \right\} + \\ \sum_{ij} \min_{x_i, x_j} & \left\{ f_{ij}(x_i, x_j) - \delta_{ji} x_i - \delta_{ij} x_j \right\}. \end{aligned}$$

Let $\boldsymbol{x}^* \in \arg\min_x F(\boldsymbol{x})$, and for each $ij \in E$ take some $g \in \partial f_{ij}(x_i^*, x_j^*)$. Set the multipliers δ_{ji}, δ_{ij} as the components of g:

$$\delta_{ji}^* = g_i, \delta_{ij}^* = g_j,$$

and define the reparameterizaition obtained under $\boldsymbol{\delta}^*$:

$$\begin{aligned} \bar{f}_i(x_i) &= f_i(x_i) + \sum_{j \in N(i)} \delta_{ji}^* x_i \\ \bar{f}_{ij}(x_i, x_j) &= f_{ij}(x_i, x_j) - \delta_{ji}^* x_i - \delta_{ij}^* x_j. \end{aligned}$$

Each of the functions $\bar{f}_i, \bar{f}_{ij}$ is a sum convex functions, hence is convex. From how we set g, each $\bar{f}_{ij}$ is minimized at (x_i^*, x_j^*). We also have:

$$\sum_i \bar{f}_i(x_i^*) = F(\boldsymbol{x}^*) - \sum_{ij} \bar{f}_{ij}(x_i^*, x_j^*).$$

It holds that $0 \in \partial \bar{f}_{ij}(x_i^*, x_j^*)$ and that $0 \in \partial F(\boldsymbol{x}^*)$ (because $\boldsymbol{x}^*$ is a minimizer). From additivity of the subgradient, we now get $0 \in \partial \sum_i \bar{f}_i(x_i^*)$, and since each function in the sum depends on a different variable it also holds that $0 \in \partial \bar{f}_i(x_i^*)$ for each i. We conclude that x^* minimizes each function in the reparameterization $\bar{f}_i, \bar{f}_{ij}$, and the dual objective our assignment $\boldsymbol{\delta}^*$ achieves is:

$$\sum_i \bar{f}_i(x_i^*) + \sum_{ij} \bar{f}_{ij}(x_i^*, x_j^*) = F(\boldsymbol{x}^*).$$

Any objective reached by a feasible dual assignment yields a lower bound on LCR's optimal objective, thus LCR's optimum is bounded below the MAP objective and we may conclude that the relaxation is tight. □

A.3 Proof of Claims on LCR for GMRFs

A.3.1 Proof of Claims 5.1 and 5.2

Recall that for a GMRF we have:

$$\boldsymbol{\phi}_i = \{x_i, x_i^2\}, \boldsymbol{\phi}_{ij} = \{x_i x_j\}.$$

For any feasible element in $\mathcal{M}_L$, let $\hat{p}_i, \hat{p}_{ij}$ be the densities that generated these feasible mean parameters. For any $ij \in$

$E, i \in V$ denote:

$$\Sigma_{ij} = \mathbb{E}_{\hat{p}_{ij}}\left[x_i x_j\right], \tag{16}$$
$$\Sigma_{ii} = \mathbb{E}_{\hat{p}_{ij}}\left[x_i^2\right] = \mathbb{E}_{\hat{p}_i}\left[x_i^2\right],$$
$$\eta_i = \mathbb{E}_{\hat{p}_{ij}}\left[x_i\right] = \mathbb{E}_{\hat{p}_i}\left[x_i\right].$$

The equalities of expectations with respect to $\hat{p}_i$ and $\hat{p}_{ij}$ hold due to local consistency constraints. Since for any $ij \in E$, the parameters $(\Sigma_{[ij]}, \boldsymbol{\eta}_{[ij]})$ are first and second moments of the density $\hat{p}_{ij}$ then it must hold that:

$$\Sigma_{[ij]} - \boldsymbol{\eta}_{[ij]}\boldsymbol{\eta}_{[ij]}^{\top} \geq 0. \tag{17}$$

On the other hand, assume we are given $(\Sigma, \boldsymbol{\eta})$ such that Eq. (17) holds for all $ij \in E$. Take $\tilde{p}_{ij}$ as the bivariate Gaussian density with moments $(\Sigma_{[ij]}, \boldsymbol{\eta}_{[ij]})$ for all $ij \in E$, and $\tilde{p}_i$ as the univariate Gaussian density with moments (Σ_{ii}, η_i) for all $i \in V$. These densities satisfy the local consistency constraints:

$$\int \tilde{p}_{ij}(x_i, x_j) dx_j = \tilde{p}_i(x_i),$$

which means $(\Sigma, \boldsymbol{\eta}) \in \mathcal{M}_L$ and that Claim 5.1 holds.

To see Claim 5.2 holds, we use the same type of argument and prove $\mathcal{M}_L^{\boldsymbol{\psi}} \subseteq \mathcal{M}_L$. Given any feasible element in $\mathcal{M}_L^{\boldsymbol{\psi}}$, consider the densities $\hat{p}_i, \hat{p}_{ij}$ who generated it. The weak local consistency constraints imposed by $\mathcal{M}_L^{\boldsymbol{\psi}}$ assure that the equalities in Eq. (16) hold. Thus Eq. (17) also holds and this element must also be feasible in $\mathcal{M}_L$.

A.3.2 Proof of Claim 5.3

Proof. Let $\boldsymbol{\delta}$ be any assignment to the dual Eq. (9), we will prove it achieves an objective of $-\infty$. It will then follow that this is the maximal value achieved by the dual, and in turn the minimal value achieved by LCR. Let us define the functions:

$$\bar{f}_i(x_i) = f_i(x_i) + \sum_{j \in N(i)} \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\psi}_i(x_i) \rangle$$
$$\bar{f}_{ij}(x_i, x_j) = f_{ij}(x_i, x_j) - \langle \boldsymbol{\delta}_{ji}, \boldsymbol{\psi}_i(x_i) \rangle - \langle \boldsymbol{\delta}_{ij}, \boldsymbol{\psi}_j(x_j) \rangle.$$

The dual objective may then be written as:

$$\sum_i \min_{x_i} \bar{f}_i(x_i) + \sum_{ij} \min_{x_i, x_j} \bar{f}_{ij}(x_i, x_j).$$

Under the choice of $\boldsymbol{\psi}_i(x_i)$ made in Claim 5.2, each function in $\{\bar{f}_i(x_i), \bar{f}_{ij}(x_i, x_j)\}$ is quadratic. $\{\bar{f}_i(x_i), \bar{f}_{ij}(x_i, x_j)\}$ is also a reparameterization of $F(\boldsymbol{x})$, so for non-CD GMRFs, at least one of these functions must be non-convex. The minimum of a non-convex quadratic function is unbounded, so one of the minimizations in the sum of the dual objective must be unbounded, and the objective achieved by $\boldsymbol{\delta}$ is $-\infty$. □

Acknowledgements

This work was supported by the ISF Centers of Excellence grant 1789/11. We would like to thank Nir Rosenfeld for his useful comments on this paper.

References

[1] C. Arora and A. Globerson. Higher order matching for consistent multiple target tracking. In *ICCV*, 2013.

[2] S. Arora, C. Daskalakis, and D. Steurer. Message passing algorithms and improved lp decoding. In *STOC*, pages 3–12, 2009.

[3] M. Bayati, D. Shah, and M. Sharma. Max-product for maximum weight matching: Convergence, correctness, and lp duality. *IEEE Trans. on Inf. Theory*, 54(3):1241–1251, 2008.

[4] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[5] D. Bickson. *Gaussian Belief Propagation: Theory and Aplication*. PhD thesis, Hebrew University of Jeruslaem, 2008.

[6] B. Cseke and T. Heskes. Properties of bethe free energies and message passing in gaussian models. *Journal of Artificial Intelligence Research*, 41(2):1–24, 2011.

[7] G. Even and N. Halabi. Message-passing algorithms for packing and covering linear programs with zero-one matrices. *arXiv preprint arXiv:1302.3518*, 2013.

[8] J. Feldman, T. Malkin, R. A. Servedio, C. Stein, and M. J. Wainwright. LP decoding corrects a constant fraction of errors. *IEEE Trans. on Inf. Theory*, 53(1):82–89, 2007.

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

[10] A. Globerson and T. Jaakkola. Fixing max-product: Convergent message passing algorithms for map lp-relaxations. *Advances in Neural Information Processing Systems*, 21(1.6), 2007.

[11] J. K. Johnson. *Convex relaxation methods for graphical models: Lagrangian and maximum entropy approaches*. PhD thesis, Massachusetts Institute of Technology, 2008.

[12] R. Koetter and P. O. Vontobel. Graph-covers and iterative decoding of finite length codes. In *Proc. Int. Symp. on Turbo Codes and Related Topics*, 2003.

[13] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[14] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on Pattern Anal. and Mach. Int.*, 28(10):1568–1583, 2006.

[15] V. Kovalevsky and V. Koval. A diffusion algorithm for decreasing energy of max-sum labeling problem. *Glushkov Institute of Cybernetics, Kiev, USSR*, 1975.

[16] D. Malioutov, J. Johnson, and A. Willsky. Walk-sums and belief propagation in gaussian graphical models. *The Journal of Machine Learning Research*, 7:2031–2064, 2006.

[17] C. Moallemi and B. Van Roy. Convergence of the min-sum algorithm for convex optimization. In *Proc. of the 45th Allerton Conference on Communication, Control and Computing*, 2007.

[18] C. Moallemi and B. Van Roy. Convergence of min-sum message passing for quadratic optimization. *IEEE Trans. on Inf. Theory*, 55(5):2413–2423, 2009.

[19] A. Montanari. Graphical models concepts in compressed sensing. *Compressed Sensing: Theory and Applications*, pages 394–438, 2012.

[20] J. Peng, T. Hazan, D. McAllester, and R. Urtasun. Convex max-product algorithms for continuous mrfs with applications to protein folding. In *Proc. ICML*, 2011.

[21] T. J. Richardson, M. A. Shokrollahi, and R. L. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *IEEE Trans. on Inf. Theory*, 47(2):619–637, 2001.

[22] N. Ruozzi and S. Tatikonda. Message-passing algorithms for quadratic minimization. *The Journal of Machine Learning Research*, 14(1):2287–2314, 2013.

[23] N. Ruozzi and S. Tatikonda. Message-passing algorithms: Reparameterizations and splittings. *IEEE Trans. on Inf. Theory*, 59(9):5860–5881, 2013.

[24] S. Sanghavi, D. M. Malioutov, and A. S. Willsky. Linear programming analysis of loopy belief propagation for weighted matching. In *NIPS*, 2007.

[25] S. Sanghavi, D. Shah, and A. Willsky. Message passing for maximum weight independent set. *IEEE Trans. on Inf. Theory*, 55(11):4822–4834, 2009.

[26] S. E. Shimony. Finding maps for belief networks is NP-hard. *Artificial Intell.*, 68(2):399–410, 1994.

[27] 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.

[28] D. Sontag, T. Meltzer, A. Globerson, Y. Weiss, and T. Jaakkola. Tightening LP relaxations for MAP using message-passing. In *Uncertainty in Artificial Intelligence*, pages 503–510. AUAI Press, 2008.

[29] D. A. Sontag. *Approximate inference in graphical models using LP relaxations*. PhD thesis, Massachusetts Institute of Technology, 2010.

[30] D. Tarlow, I. E. Givoni, R. S. Zemel, and B. J. Frey. Graph cuts is a max-product algorithm. In *UAI*, pages 671–680, 2011.

[31] B. Taskar, S. Lacoste-Julien, and M. Jordan. Structured prediction, dual extragradient and Bregman projections. *Journal of Machine Learning Research*, pages 1627–1653, 2006.

[32] P. Vontobel and R. Koetter. On the relationship between linear programming decoding and min-sum algorithm decoding. In *ISITA*, pages 991–996, 2004.

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

[34] Y. Weiss and W. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural computation*, 13(10):2173–2200, 2001.

[35] Y. Weiss, C. Yanover, and T. Meltzer. Map estimation, linear programming and belief propagation with convex free energies. In *UAI*, 2007.

[36] C. Yanover, O. Schueler-Furman, and Y. Weiss. Minimizing and learning energy functions for side-chain prediction. *Journal of Computational Biology*, 15(7):899–911, 2008.

[37] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[38] C. Zach and P. Kohli. A convex discrete-continuous approach for markov random fields. *Computer Vision–ECCV 2012*, 2012.

[39] G. Zhang and R. Heusdens. Generalized linear coordinate-descent message-passing for convex optimization. In *ICASSP*, pages 2009–2012. IEEE, 2012.

[40] G. Zhang and R. Heusdens. Linear coordinate-descent message passing for quadratic optimization. *Neural computation*, 24(12):3340–3370, 2012.