In Proceedings of *The 45th Allerton Conference on Communication, Control and Computing*, September, 2007.

Lagrangian Relaxation for MAP Estimation in Graphical Models

Jason K. Johnson, Dmitry M. Malioutov and Alan S. Willsky

***Abstract*— We develop a general framework for MAP estimation in discrete and Gaussian graphical models using Lagrangian relaxation techniques. The key idea is to reformulate an intractable estimation problem as one defined on a more tractable graph, but subject to additional constraints. Relaxing these constraints gives a tractable dual problem, one defined by a thin graph, which is then optimized by an iterative procedure. When this iterative optimization leads to a consistent estimate, one which also satisfies the constraints, then it corresponds to an optimal MAP estimate of the original model. Otherwise there is a "duality gap", and we obtain a bound on the optimal solution. Thus, our approach combines convex optimization with dynamic programming techniques applicable for thin graphs. The popular tree-reweighted max-product (TRMP) method may be seen as solving a particular class of such relaxations, where the intractable graph is relaxed to a set of spanning trees. We also consider relaxations to a set of small induced subgraphs, thin subgraphs (e.g. loops), and a connected tree obtained by "unwinding" cycles. In addition, we propose a new class of multiscale relaxations that introduce "summary" variables. The potential benefits of such generalizations include: reducing or eliminating the "duality gap" in hard problems, reducing the number of Lagrange multipliers in the dual problem, and accelerating convergence of the iterative optimization procedure.**

I. Introduction

Graphical models are probability models for a collection of random variables on a graph: the nodes of the graph represent random variables and the graph structure encodes conditional independence relations among the variables. Such models provide compact representations of probability distributions, and have found many practical applications in physics, statistical signal and image processing, error-correcting coding and machine learning. However, performing optimal estimation in such models using standard junction tree approaches generally is intractable in large-scale estimation scenarios. This motivates the development of variational techniques to perform approximate inference, and, in some cases, recover the optimal estimate.

We consider a general Lagrangian relaxation (LR) approach to *maximum a posteriori* (MAP) estimation in graphical models. The general idea is to reformulate the estimation problem on an intractable graph as a constrained estimation over an augmented model defined on a larger, but more tractable graph. Then, using Lagrange multipliers to relax the constraints, we obtain a tractable estimation problem that gives an upper-bound on the original problem. This leads to a convex optimization problem of minimizing the upper-bound as a function of Lagrange multipliers.

The authors are with the Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. {jasonj,dmm,willsky}@mit.com.

We consider a variety of strategies to augment the original graph. The simplest approach breaks the graph into many small, overlapping subgraphs, which involves replicating some variables. Similarly, the graph can be broken into a set of thin subgraphs, as in the TRMP approach, or "unrolled" to obtain a larger, but connected, thin graph. We show that all of these approaches are essentially equivalent, being characterized by the set of maximal cliques of the augmented graph. More generally, we also consider the introduction of "summary" variables, which leads naturally to multiscale algorithms. We develop a general optimization approach based on marginal and max-marginal matching procedures, which enforce consistency between replicas of a node or edge, and moment-matching in the multiscale relaxation. We show that the resulting bound is tight if and only if there exists an optimal assignment in the augmented model that satisfies the constraints. In that case, we obtain the desired MAP estimate of the original model. When there is a duality gap, this is evidenced by the occurrence of "ties" in the resulting set of max-marginals, which requires further augmentation of the model to reduce and ultimately eliminate the duality gap. We focus primarily on discrete graphical models with binary variables, but also consider the extension to Gaussian graphical models. In the Gaussian model, we find that, whenever LR is "well-posed", so that the augmented model is valid, it leads to a tight bound and the optimal MAP estimate, and also gives *upper-bounds* on variances that provide a measure of confidence in the MAP estimate.

II. Background

We consider probabilistic graphical models [1], [2], [3], which are probability distributions of the form

$$p(x_1,\ldots,x_n) = \frac{1}{Z}\exp\{f(x)\} = \frac{1}{Z}\exp\left\{\sum_{C\in\mathscr{G}} f_C(x_C)\right\} \quad (1)$$

where each function f_C only depends on a subset of variables $x_C = (x_v, v \in C)$ and Z is a normalization constant of the model, called the *partition function* in statistical physics. If the sum ranges over all *cliques* of the graph, which are the fully connected subsets of variables, this representation is sufficient to realize any Markov model on $\mathscr{G}$ [1]. However, it is also common to consider restricted Markov models where only singleton and pairwise interactions are specified. In general, we specify the set of interactions by a hypergraph $\mathscr{G} \subset 2^V$, where 2^V represents the set of all subsets of V. The elements of $\mathscr{G}$ are its *hyperedges*, which generalizes the usual concept of a graph with pairwise edges.

Discrete Models. While our approach is applicable for general discrete models, we focus on models with binary

variables. One may use either the Boltzmann machine representation $x_v \in \{0,1\}$, or that of the Ising model $x_v \in \{-1,+1\}$. These models can be represented as in (1) with

$$f(x;\theta) = \sum_{E\in\mathscr{G}} \theta_E \phi_E(x_E),\ \ \phi_E(x_E) = \prod_{v\in E} x_v \qquad (2)$$

This defines an *exponential family* [4] of probability distributions based on model features ϕ and parameterized by θ. $\Phi(\theta) \triangleq \log Z(\theta)$ is the *log-partition function* and has the *moment-generating property*: $\frac{\partial\Phi(\theta)}{\partial\theta_E} = \mathbb{E}_\theta\{\phi_E(x)\} \triangleq \eta_E$. Here, η are the *moments* of the distribution, which serve both as an alternate parameterization of the exponential family and, in graphical models, to specify the marginal distributions on cliques of the model. Inference in discrete models using junction tree methods, either to compute the mode or the marginals, is generally linear in the number of variables *n* but grows exponentially in the *width* of the graph [2], which is determined by the size of the maximal cliques in a junction tree representation of the graph. Hence, exact inference is only tractable for *thin* graphs, that is, where one can build an equivalent junction tree with small cliques.

Gaussian Models. We also consider Gaussian graphical models [5], [6] represented in *information form*:

$$p(x) = \exp\{-\tfrac{1}{2}x^T Jx + h^T x - \Phi(h,J)\} \qquad (3)$$

where J is the *information matrix*, h a potential vector and $\Phi(h,J) = \frac{1}{2}\{h^T J^{-1}h - \log\det J + n\log 2\pi\}$. This corresponds to the standard form of the Gaussian model specified by the covariance matrix $P = J^{-1}$ and mean vector $\hat{x} = J^{-1}h$. This translates into an exponential family where we identify (h,J) with the parameters θ and $(\hat{x},P)$ with the moments η. In general, the complexity of inference in Gaussian models is $\mathscr{O}(n^3)$. The fill pattern of J determines the Markov structure of the Gaussian model: $(i,j) \in \mathscr{G}$ if $J_{i,j} \neq 0$. Using more efficient recursive inference methods that exploit sparsity, such as junction trees or sparse Gaussian elimination, the complexity is linear in *n* but cubic in the width of the graph, which is still impractical for many large-scale estimation problems.

III. Discrete Lagrangian Relaxation

To begin with, consider the problem of maximizing the following objective function, defined over a hypergraph $\mathscr{G} \subset 2^V$ based on a vertex set $V = \{1,\ldots,n\}$ corresponding to discrete variables $x = (x_1,\ldots,x_n)$.

$$f(x) = \sum_{E\in\mathscr{G}} f_E(x_E) \qquad (4)$$

For instance, this may be defined as $f(x) = \langle\theta,\phi(x)\rangle$ in an exponential family graphical model, such that each term corresponds to a feature $f_E(x_E) = \theta_E\phi_E(x_E)$. Then, we seek x^* to maximize $f(x)$ to obtain the MAP estimate of (1).

An Illustrative Example. To briefly convey the basic concept, we consider a simple pairwise model defined on a 3-node cycle $\mathscr{G}$ represented in Fig. 1. Here, the augmented graph $\mathscr{G}'$ is a 4-node chain, where node 4 is a replica of node 1. We copy all the potentials on the nodes and edges

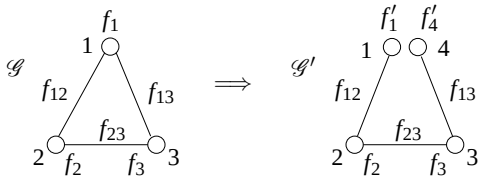

Fig. 1. A simple illustrative example of Lagrangian relaxation.

from $\mathscr{G}$ to $\mathscr{G}'$. For the replicated variables, x_1' and x_4', we split f_1 between f_1' and f_4' such that $f_1(y) = f_1'(y) + f_4'(y)$ for $y \in \{0,1\}$. Now the problem $\max_x f(x)$ is equivalent to maximizing $f'(x')$ subject to the constraint $x_1' = x_4'$. To solve the latter we relax the constraint using Lagrange multipliers: $L(x',\lambda) = f'(x') + \lambda(x_1' - x_4')$. The additional term $\lambda(x_1' - x_4')$ modifies the self-potentials: $f_1' \leftarrow f_1'(x_1') + \lambda x_1'$ and $f_4' \leftarrow f_4'(x_4') - \lambda x_4'$, parameterizing a family of models on $\mathscr{G}'$ all of which are equivalent to f under the constraint $x_1' = x_4'$. For a fixed λ, solving $\max_x L(x,\lambda) \triangleq g(\lambda)$ gives an upper bound on $f^* = \max_x f(x)$, so by optimizing λ to minimize $g(\lambda)$, we find the tightest bound $g^* = \min_\lambda g(\lambda)$. If the constraint $x_1' = x_4'$ is satisfied in the final solution, then there is strong duality $g^* = f^*$ and we obtain the correct MAP assignment for $f(x)$.

We now discuss the general procedure and develop our approach to optimize $g(\lambda)$ in more difficult cases.

A. Obtaining a Tractable Graph by Vertex Replication

In this section, we consider approaches that involve *replicating* variables to define the augmented model. The basic constraints in designing $\mathscr{G}'$ are as follows: $\mathscr{G}'$ is comprised of replicas of nodes and edges of $\mathscr{G}$. Every node and edge of $\mathscr{G}$ must be represented at least once in $\mathscr{G}'$. Finally, $\mathscr{G}'$ should be a thin graph, which relates to the complexity of our method.

To help illustrate the various strategies, we consider a pairwise model $f(x)$ defined on 5×5 grid, as seen in Fig. 2(a). A natural approach is to break the model up into small subgraphs. The simplest method is to break the graph up into its composite interactions. For pairwise models, this means that we split the graph into a set of disjoint edges as shown in (b). Here, each internal node of the graph is replicated four times. To reduce the number of replicated nodes, and hence the number of constraints, it is also useful to merge many of these smaller subgraphs into larger thin graphs. One approach is to group edges into *spanning trees* of the graph as seen in (c). Here, each edge must be including in at least one tree, and some edges are replicated in multiple trees. The TRMP approach is based on this idea. One could also allow multiple replicas of a node in the same connected component of $\mathscr{G}'$. For instance, by taking a spanning tree of the graph and then adding an extra leaf node for each missing edge we obtain the graph seen in (d).

It is also tractable to use small subgraphs that are not trees. We can break the graph into a set of short loops as in (e) or a set of induced subgraphs as in (f) where we select a set of 3×3 subgraphs that overlap on their

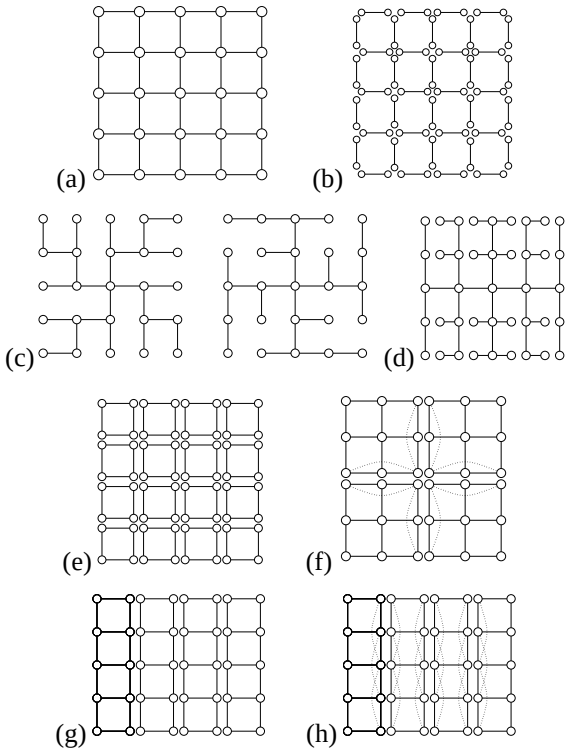

Fig. 2. Illustrations of a variety of possible ways to obtain a tractable graph structure from a 5×5 grid by replicating some vertices of the graph.

boundary. In such cases, including additional edges in the overlap of these subgraphs, such as the dotted edges in (f), can enhance the relaxations that we consider. Finally, we reduce the number of constraints in these formulations by again grouping subgraphs to form larger subgraphs that are still thin, as shown in (g). This will also lead to tractability in our methods. Again, it can be useful to include extra edges in the overlap of these subgraphs as in (h), although this increases the width of the subgraph and affects the computational complexity of our methods.

Notation. Let $\mathscr{G}'$ denote the augmented graph (or collection of subgraphs), which is based on an extended vertex set V', comprised of replicas of nodes in V. We assume that all edges of this graph are also replicas of edges of the original graph $\mathscr{G}$.[1] Thus, there is a well-defined surjective map $\Gamma : \mathscr{G}' \to \mathscr{G}$, each edge $E' \in \mathscr{G}'$ is a replica an edge $E = \Gamma(E') \in \mathscr{G}$, and every edge of $\mathscr{G}$ has at least one such replica. This notation is overloaded for nodes by treating them as singleton edges of $\mathscr{G}$. We also denote the set-valued inverse of Γ by $\mathscr{R}(E) \triangleq \Gamma^{-1}(E)$, which is the set of replicas of E, and let $r_E \triangleq |\mathscr{R}(E)|$ denote the number of replicas. This defines an equivalence relation on $\mathscr{G}'$: $A, B \in \mathscr{G}'$ are equivalent $A \equiv B$ if $\Gamma(A) = \Gamma(B)$, that is, if $A, B \in \mathscr{R}(E)$ are replicas of the same edge $E \in \mathscr{G}$.

B. Equivalent Constrained Estimation Problem

We now define a corresponding objective function $f'(x')$, where $x' = (x'_v)_{v \in V'}$ are the variables of the augmented model. For each hyperedge $E \in \mathscr{G}$ (including individual nodes), we split the function $f_E(x_E)$ among a set of replica functions $\{f'_{E'}, E' \in \mathscr{R}(E)\}$, requiring that these are *consistent*,

$$f_E(x_E) = \sum_{E' \in \mathscr{R}(E)} f'_{E'}(x_E) \text{ for all } x_E. \tag{5}$$

Using the parametric representation $f(x) = \langle \theta, \phi(x) \rangle$, this consistency condition is equivalent to requiring $\theta_E = \Sigma_{E'} \theta'_{E'}$. We will see that the LR approach to follow may be viewed as an optimization over all such possible consistent splittings. Next, we define the augmented objective function over the graph $\mathscr{G}'$ as

$$f'(x') \triangleq \sum_{E \in \mathscr{G}'} f'_E(x'_E). \tag{6}$$

This insures that $f(x) = f'(x')$ where $x' = \zeta(x)$ is the replicated version of x, defined by $x'_{v'} = x_v$ for all $v' \in \mathscr{R}(v)$. This equivalence holds for all *consistent* configurations $x' \in \zeta(\mathbb{X})$, where x' is self-consistent over various replicas of the same node. Thus, we are led to an equivalent optimization problem in the augmented model subject to consistency constraints:

$$f^* \triangleq \max_{x \in \mathbb{X}} f(x) = \max_{x' \in \zeta(\mathbb{X})} f'(x') \tag{7}$$

Expressing the consistency constraint as a set of linear constraints on the model features ϕ, we obtain:

$$\begin{array}{ll} \text{maximize} & f'(x') \\ \text{subject to} & \phi_A(x'_A) = \phi_B(x'_B) \text{ for all } A \equiv B. \end{array} \tag{8}$$

Recall that, in the discrete binary model, these features are defined $\phi_E(x_E) = \Pi_{v \in E} x_v$. Clearly, there is some redundancy in these constraints: $x_a = x_b$ for all replicated nodes $a \equiv b$ would insure that the edges agree. However, these redundant edge-wise feature constraints do enhance the following relaxation.

C. Lagrangian Relaxation

We have now defined an equivalent model on a tractable graph. However, the equivalent *constrained* optimization is still intractable, because the constraints couple some variables of $\mathscr{G}'$, spoiling its tractable structure. This suggests the use of Lagrangian duality to relax those complicating constraints. Introducing Lagrange multipliers $\lambda_{A,B}$ for each constraint, we define the *Lagrangian*, which is a modified version of the objective function:

$$L(x', \lambda) = f'(x') + \sum_{A \equiv B} \lambda_{A,B} (\phi_A(x'_A) - \phi_B(x'_B)) \tag{9}$$

Grouping terms by edges $E \in \mathscr{G}'$, and using $f'_E(x_E) = \theta'_E \phi_E(x_E)$, this is represented

$$\begin{aligned} L(x', \lambda) &= \sum_{E \in \mathscr{G}'} f'_E(x'_E; \lambda) \\ f'_E(x'_E; \lambda) &= \theta'_E(\lambda) \phi_E(x'_E) \\ \theta'_E(\lambda) &= \theta'_E + \Sigma_B \lambda_{E,B} - \Sigma_A \lambda_{A,E} \end{aligned} \tag{10}$$

Note that the Lagrange multipliers may be interpreted as parameterizing all consistent splittings, $\theta'(\lambda)$ spans the subspace of all consistent θ' parameters.[2]

[1] In the case that we introduce extra edges in $\mathscr{G}'$, as in (f) and (h), we also add corresponding edges to $\mathscr{G}$ to maintain this convention.

[2] We obtain a minimal λ parameterization by only using a subset of constraints in (9), such that $\{(\phi_A(x') - \phi_B(x'))\}$ are linearly independent.

It is tractable to maximize the Lagrangian, as it is defined over the thin graph $\mathscr{G}'$. The value of this maximization defines the *dual function*:

$$g(\lambda) = \max_{x'} L(x', \lambda) \tag{11}$$

Note that this is an *unconstrained* optimization over $\mathbb{X}'$, and its solution need not lead to a consistent $x' \in \zeta(\mathbb{X})$. However, if this x' is consistent then it is an optimal solution of the constrained optimization problem (7), and hence $x = \zeta^{-1}(x')$ (which is well-defined for consistent x') is also an optimal solution of the original problem. This is the goal of our approach, to find tractable relaxations of the MAP estimation problem which lead to the correct MAP estimate. This motivates solution of the *dual problem*:

$$\min_{\lambda} g(\lambda) \triangleq g^* \tag{12}$$

Appealing to well-known results [7], [8], we conclude:

Proposition 1 (Lagrangian duality): We have $g(\lambda) \geq f^*$ for all λ. Hence $g^* \geq f^*$. If $g(\lambda^*) = g^*$, then one of the following holds:

(i) There exists a consistent solution:

$$x' \in \arg\max_{x' \in \mathbb{X}'} L(x'; \lambda^*) \cap \zeta(\mathbb{X}).$$

Then, we have *strong duality* $g^* = f^*$ and the set of *all* MAP estimates is obtained as:

$$\arg\max_{x' \in \zeta(\mathbb{X})} f'(x') = \arg\max_{x' \in \mathbb{X}} L(x', \lambda^*) \cap \zeta(\mathbb{X}).$$

(ii) There are no consistent solutions:

$$\arg\max_{x' \in \mathbb{X}'} L(x'; \lambda^*) \cap \zeta(\mathbb{X}) = \emptyset.$$

Then, there is a *duality gap* $g^* > f^*$ and *no* choice of λ will provide a consistent solution.

Also, condition (i) holds only if $g(\lambda^*) = g^*$.

This result generalizes the analogous *strong tree-agreement* optimality condition for TRMP, and clarifies its connection to standard Lagrangian duality results for integer programs. To provide some intuition, we present the following geometric interpretation illustrated in Fig. 3. The dual function is the maximum over a finite set of linear functions in λ indexed by x'. For each $x' \in \mathbb{X}'$, there is a linear function $g(\lambda; x') = \langle a(x'), \lambda \rangle + b(x')$, with $a(x') = (\phi_A(x') - \phi_B(x'))_{A \equiv B}$, which is the gradient, and $b(x') = f'(x')$. The graph of each of these functions defines a hyperplane in $\mathbb{R}^{d+1}$, where d is the number of constraints. The flat hyperplanes, with $a = 0$, correspond to consistent assignments $x' \in \zeta(\mathbb{X})$. The remaining sloped hyperplanes represent inconsistent assignments. Hence, the highest flat hyperplane corresponds to the optimal MAP estimate, with height equal to f^*. The dual function $g(\lambda)$ is defined by the maximum height over this set of hyperplanes for each λ, and is therefore convex, piece-wise linear and greater than or equal to f^* for all λ. In the case of a duality gap, the inconsistent hyperplanes hide the consistent ones, as depicted in (a), so that the minimum of the dual function is defined

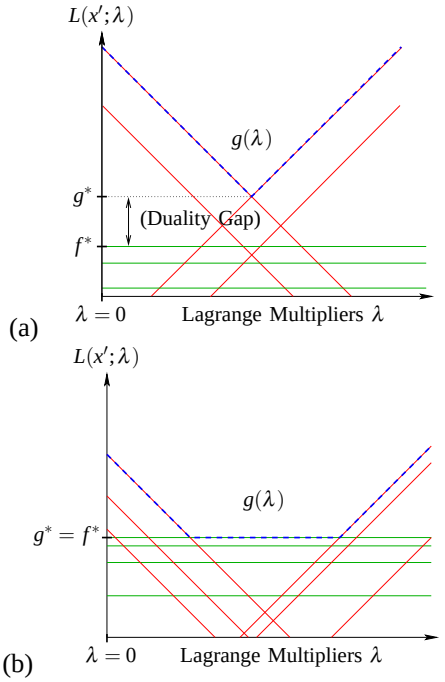

Fig. 3. Illustration of the Lagrangian duality in the cases that (a) there is a duality gap and (b) there is no duality gap (strong duality holds).

by an intersection of slanted hyperplanes corresponding to inconsistent assignments of x'. If there is no duality gap, as depicted in (b), then the minimum is defined by the flat hyperplane corresponding to a consistent assignment. Its intersection with slanted hyperplanes defines the polytope of optimal Lagrange multipliers over which the maximum flat hyperplane is exposed.

D. Linear Programming Formulations

We briefly consider a connection between this LR picture and TRMP [9], [10] and related linear programming approaches [11], [12], [13]. This analysis also serves to understand when different relaxations of the MAP estimation problem will be equivalent.

The *epigraph* of the dual function is defined as the set of all points $(\lambda, h) = \mathbb{R}^{d+1}$ where $g(\lambda) \leq h$, that is, where $a(x')\lambda + b(x') \leq h$ for all x'. Thus, the minimum of the dual function is equal to the lowest point of the epigraph, which defines a linear program (LP) over $(\lambda, h) \in \mathbb{R}^{d+1}$:

$$\begin{array}{ll} \text{minimize} & h \\ \text{subject to} & \langle a(x'), \lambda \rangle + b(x') \leq h \text{ for all } x'. \end{array} \tag{13}$$

Note that there are exponentially many constraints in this formulation, so it is intractable. However, recalling that it *is* tractable to compute the dual function for a given λ, using the max-product algorithm applied to the thin graph $\mathscr{G}'$, we seek a more tractable representation of this LP. To achieve this, we consider the LP dual problem obtained by dualizing the constraints, which is always tight [8]. This LP dual should be distinguished from our Lagrangian dual (12) that is the subject of our paper.

Introducing non-negative Lagrange multipliers $\mu(x') \geq 0$ for each inequality constraint, indexed by $x' \in \mathbb{X}'$, we obtain

the LP Lagrangian:

$$\begin{aligned} M(h,\lambda;\mu) &= h+\mu\left[\langle a(x'),\lambda\rangle + b(x') - h\right] \\ &= \langle \mu[a],\lambda\rangle + \mu[b] + (1-\mu[1])h, \end{aligned} \quad (14)$$

where μ denotes μ-weighted summation, e.g., $\mu[a] = \sum_{x'}\mu(x')a(x')$. The LP dual function is then:

$$M^*(\mu) \triangleq \min_{h,\lambda} M(h,\lambda;\mu) = \begin{cases} \mu[b], & \mu[1]=1 \text{ and } \mu[a]=0 \\ -\infty, & \text{otherwise.} \end{cases} \quad (15)$$

Note that $\mu > 0$ and $\mu[1] = 1$ imply that μ is a probability distribution and $\mu[\cdot]$ an expectation operator. Recalling $a(x') \triangleq (\phi_A(x') - \phi_B(x'), A \equiv B)$ and $b(x') \triangleq f'(x')$, we obtain the dual LP:

$$\max_{\mu\geq 0} M^*(\mu) = \begin{cases} \text{maximize} & \mu[f'] \\ \text{subject to} & \mu[\phi_A] = \mu[\phi_B] \text{ for } A \equiv B \end{cases} \quad (16)$$

We seek a probability distribution over all configurations of the augmented model that maximizes the expected value of $f'(x')$ subject to constraints that the moments specifying marginal distributions are consistent for replicated nodes and edges of the graph. This is a convex relaxation of the constrained version of problem (4), where the objective and constraint functions have been replaced by their expected values under μ. Note that only marginals $\mu_{E'}$ over hyperedges $E \in \mathscr{G}'$ are needed to evaluate both the objective and the constraints of this LP. Hence, it reduces to one defined over the *marginal polytope* $\mathscr{M}(\mathscr{G}')$ [9], defined as the set of all *realizable* collections of marginals over the hyperedges of $\mathscr{G}'$. Moreover, if the graph $\mathscr{G}'$ is *chordal* [2], then its marginal polytope has a simple characterization. Let $\mathscr{M}_{\text{local}}(\mathscr{G}')$ denote the *local marginal polytope* defined as the set of all edge-wise marginal specifications that are consistent on intersections of edges. In general, $\mathscr{M}(\mathscr{G}') \subset \mathscr{M}_{\text{local}}(\mathscr{G}')$. However, in chordal graphs it holds that $\mathscr{M}(\mathscr{G}') = \mathscr{M}_{\text{local}}(\mathscr{G}')$. Thus, if $\mathscr{G}'$ is a thin chordal graph, we obtain a tractable LP whose value is equivalent to g^* in our framework.[3]

One last step shows the connection to LP approaches [9], [11], [12]. The key observation is that, roughly speaking,

$$\mathscr{M}_{\text{local}}(\mathscr{G}') \cap \{\mu | \mu(x_A) = \mu(x_B), A \equiv B\} \equiv \mathscr{M}_{\text{local}}(\mathscr{G}). \quad (17)$$

This is seen by replicating marginals from $\mathscr{G}$ to $\mathscr{G}'$, or by copying (consistent) replicated marginals back to $\mathscr{G}$. For such consistent μ, we have $\mu[f'] = \mu[f]$, which gives:

$$g^* = \max_{\mu\in\mathscr{M}_{\text{local}}(\mathscr{G})} \mu[f] \geq \max_{\mu\in\mathscr{M}(\mathscr{G})} \mu[f] = f^*. \quad (18)$$

The maximum over $\mathscr{M}_{\text{local}}(\mathscr{G})$ gives an upper-bound on the maximum over $\mathscr{M}(\mathscr{G}) \subset \mathscr{M}_{\text{local}}(\mathscr{G})$. The latter is equivalent to exact MAP estimation and the bound becomes tight if $\mathscr{G}$ is the set of maximal cliques of a chordal graph. This discussion leads to the following characterization of LR:

Proposition 2 (LR Hierarchy): Equivalence: Let $\mathscr{G}'_1$ and $\mathscr{G}'_2$ be the set of maximal cliques of two chordal augmented graphs. If $\Gamma^{-1}(\mathscr{G}_1) = \Gamma^{-1}(\mathscr{G}_2)$ then $g^*_1 = g^*_2$ for the respective dual problems. Let $g^*(\mathscr{G})$ denote the common dual value of all such chordal relaxations where $\Gamma^{-1}(\mathscr{G}') = \mathscr{G}$. *Monotonicity:* If $\mathscr{G}_1 \subset \mathscr{G}_2$ then $g^*(\mathscr{G}_1) \geq g^*(\mathscr{G}_2)$. *Strong Duality:* If $\mathscr{G}$ is the set of maximal cliques of a chordal graph, then $g^*(\mathscr{G}) = f^*$.

[3]Some graphs shown in Fig. 2 are not chordal, but they can be extended to a thin chordal graph by adding a few edges. If no two of these new edges are equivalent when mapped into $\mathscr{G}$, then this does not change g^*.

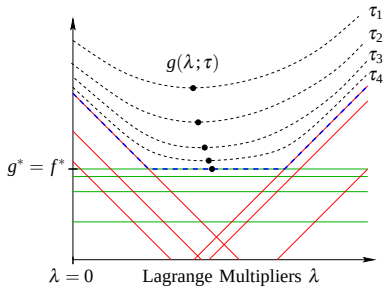

Fig. 4. Illustration of the "log-sum-exp" smooth approximation of the dual function, as a function of "temperature" τ, and of an optimization procedure for minimizing the non-smooth dual function through a sequence of smooth minimizations.

E. Smooth Relaxation of the Dual Problem

In this section, we develop an approach to solve the dual problem. One approach to minimize $g(\lambda)$ is to use non-smooth optimization methods, such as the subgradient method [8]. Here, we consider an alternative, based on the following smooth approximation of $g(\lambda)$:

$$g(\lambda;\tau) \triangleq \tau \log \sum_{x'\in\mathbb{X}} \exp\left(\frac{L(x';\lambda)}{\tau}\right) \quad (19)$$

As illustrated if Fig. 4, the parameter $\tau > 0$ controls the trade-off between smoothness of $g(\lambda;\tau)$ and how well it approximates $g(\lambda)$. This is known as the "log-sum-exp" approximation to the "max function" [14]:

$$g(\lambda) \leq g(\lambda;\tau) \leq g(\lambda) + \tau \log|\mathbb{X}| \text{ for all } \tau > 0. \quad (20)$$

Hence, $g(\lambda;\tau) \to g(\lambda)$ *uniformly* as $\tau \to 0$ and, hence, $g^*(\tau) \triangleq \min_\lambda g(\lambda;\tau)$ converges to g^*.

The function $g(\lambda;\tau)$ has another useful interpretation. Consider the Gibbs distribution defined by

$$p_{\lambda,\tau}(x') = \exp\left(\frac{L(x',\lambda) - g(\lambda;\tau)}{\tau}\right) \quad (21)$$

Here, $\tau > 0$ is the "temperature" and $g(\lambda;\tau)$ normalizes the distribution for each choice of λ and τ, and is equal to the Helmholtz free energy $\mathscr{F}_H(\theta') = \tau\Phi_\tau(\theta')$, where $\Phi_\tau(\theta') = \log\Sigma\exp(\tau^{-1}\langle\theta',\phi'(x')\rangle)$ is the usual log-partition function. Thus, $g(\lambda;\tau)$ is a strictly convex, analytic function. Using the moment-generating property of $\Phi_\tau(\theta')$, the gradient of $g(\lambda;\tau)$ is computed as:

$$\begin{aligned} \frac{\partial g(\lambda;\tau)}{\partial\lambda_{A,B}} &= \frac{\partial\Phi_\tau}{\partial\theta'_A}\frac{\partial\theta'_A}{\partial\lambda_{A,B}} + \frac{\partial\Phi_\tau}{\partial\theta'_B}\frac{\partial\theta'_B}{\partial\lambda_{A,B}} \\ &= p_{\lambda,\tau}[\phi_A] - p_{\lambda,\tau}[\phi_B] \end{aligned} \quad (22)$$

where we use $p[\cdot]$ to denote expectation under p. Thus, appealing to strict convexity, there is a unique $\lambda^*(\tau)$ that

ALGORITHM 1 (Discrete LR)
Iterate until convergence:
For $E \in \mathscr{G}$ where $r_E > 1$
  For $E' \in \mathscr{R}(E)$
    $\hat{f}_{\tau,E'}(x'_{E'}) = \tau \log p_{\tau,\lambda}(x'_{E'})$
  end
  $\bar{f}_{\tau,E}(x_E) = r_E^{-1} \sum_{E'} \hat{f}_{\tau,E'}(x_E)$
  For $E' \in \mathscr{R}(E)$
    $f_{E'}(x_E) \leftarrow f_{E'}(x_E) + \big(\bar{f}_{\tau,E'}(x_E) - \hat{f}_{\tau,E'}(x_E)\big)$
  end
end

minimizes $g(\lambda;\tau)$ and it is also the unique solution of the set of moment-matching conditions:

$$p_{\lambda,\tau}[\phi_A] = p_{\lambda,\tau}[\phi_B], \text{ for all } A \equiv B.$$

These moment-matching conditions are equivalent to requiring that the marginal distributions $p_{\lambda,\tau}(x_A)$ and $p_{\lambda,\tau}(x_B)$ are equal for $x_A = x_B$. We also note that $\frac{\partial g(\lambda;\tau)}{\partial \tau} = p_{\lambda,\tau}[-\log p_{\lambda,\tau}]$, which is the *entropy* of $p_{\lambda,\tau}$ and is positive for all λ. Hence, for a decreasing sequence $\tau_k > 0$ converging to zero, $g(\lambda;\tau)$ converges *monotonically* to $g(\lambda)$. Likewise, $g^*(\tau_k)$ converges *monotonically* to g^*.

Rather than directly optimizing $g(\lambda)$, we instead perform a sequence of minimizations with respect to the functions $g(\lambda;\tau_k)$. At each step, the previous estimate of $\lambda_k^* = \arg\min g(\lambda;\tau_k)$ is used to initialize an iterative method to minimize $g(\lambda;\tau_{k+1})$. This is illustrated in Fig. 4. At each step, we use the following optimization procedure based on the marginal agreement condition.

1) Iterative Log-Marginal Averaging: To minimize $g(\lambda;\tau)$ for a specified τ, starting from an initial guess for λ (or, equivalently, an initial splitting of f), we develop a block coordinate-descent method. Our approach is in the same spirit as the iterative proportional fitting procedure [15].

We begin with the case that the augmented model is defined so that no two replicas of a node are contained in the same connected component of $\mathscr{G}'$. Then, at each step, we minimize over the set of all Lagrange multipliers associated with features defined within any replica of E. This is equivalent to solving the condition that the corresponding marginal distributions $p_{\lambda,\tau}(x'_{E'})$ are consistent for all $E' \in \mathscr{R}(E)$. Algorithm 1 summarizes the method, which involves computing the log-marginal of each replica edge, and then updates the functions $f'_{E'}$ according to the rule:

$$f'_{E'}(x_E) \leftarrow f'_{E'}(x_E) + (\bar{f}_{\tau,E}(x_E) - \hat{f}_{\tau,E'}(x_E)) \qquad (23)$$

where

$$\hat{f}_{\tau,E'}(x'_{E'}) = \tau \log p_{\lambda,\tau}(x'_{E'}), \;\; \bar{f}_{\tau,E}(x_E) = r_E^{-1} \sum_{E' \in \mathscr{R}(E)} \hat{f}_{\tau,E'}(x_E).$$

After the update, the new log-marginals of all replicas E' are equal to $\bar{f}_{\tau,E}$. Also, these updates maintain a consistent representation: $\sum_{E'}(\bar{f}_{\tau,E} - \hat{f}_{\tau,E'}) = 0$. To handle augmented models with multiple replicas of E in the same connected subgraph, we only update a *subset* of replicas at each step, where no two replicas are in the same subgraph. In some cases, this requires including an extra replica of E to act as an intermediary in the update step.

Each step of the procedure requires that we compute the marginal distributions of each replica E' in their respective subgraphs. In the graphs are thin, these marginals can be computed efficiently, with computation linear in the size of each subgraph, using standard belief propagations methods and their junction tree variants. Moreover, if we take some care to store the messages computed by belief propagation, it is possible to amortize the cost of this inference, by only updating a few "messages" at each step. In fact, it is only necessary to update those messages along the directed path from the last updated node or edge to the location in the tree (or junction tree) of the node or edge currently being updated. We find that this generally allows a complete set of updates to be computed with complexity linear in n. Similar ideas are discussed in [16].

Using Algorithm 1, together with a rule to gradually reduce τ, we obtain a simple algorithm which generates a sequence λ_k such that $g(\lambda_k)$ converges to g^* and λ_k converge to a point in the set of optimal Lagrange multipliers.

2) Iterative Max-Marginal Averaging: We now consider what happens as τ approaches zero. The main insight is that the (non-normalized) log-marginals converge to *max-marginals* in the limit as τ approaches zero:

$$\hat{f}_{\tau,E'}(x'_{E'}) + g(\lambda,\tau) \rightarrow \hat{f}_{E'}(x'_{E'}) \triangleq \max_{x'_{\setminus E'}} f'(x'_{E'}, x'_{\setminus E'};\lambda) \qquad (24)$$

Hence, as τ becomes small, the marginal agreement conditions are similar to a set of *max-marginal agreement* conditions among all replicas of an edge or node. One could consider a "zero-temperature" version of Algorithm 1 aimed at solving these max-marginal conditions directly:

$$\begin{aligned} f'_{E'}(x_E) &\leftarrow f'_{E'}(x_E) + \big(\bar{f}_E(x_E) - \hat{f}_{E'}(x_E)\big) \\ \bar{f}_E(x_E) &= r_E^{-1} \sum_{E'} \hat{f}_{E'}(x_E) \end{aligned} \qquad (25)$$

Here, $\bar{f}_E$ is the averaged max-marginal over all replicas of E. Note that $\hat{f}_{E'}(x_E) \geq \hat{f}_E(x_E) \triangleq \max_{x_{\setminus E}} f(x)$ for all x_E and $E' \in \mathscr{R}(E)$, which implies $\bar{f}_E(x_E) \geq \hat{f}_E(x_E)$. This "zero-temperature" approach has close ties to max-sum diffusion (see [13] and reference therein) and Kolmogorov's serial approach to TRMP [16].

In our framework, one can show that $\lambda^* \triangleq \lim_{\tau \rightarrow 0} \lambda^*(\tau)$ is well-defined and minimizes $g(\lambda)$. This point λ^* also satisfies the max-marginal agreement condition and is therefore a fixed point of max-marginal averaging. However, the max-marginal agreement condition by itself does not uniquely determine λ^* and, in fact, is not sufficient to insure that $g(\lambda)$ is minimized (this is related to the existence of non-minimal fixed-points observed by Kolmogorov). Hence, our approach to minimize $g(\lambda;\tau)$ while gradually reducing the temperature has the advantage that it cannot get stuck in such spurious fixed-points. It also helps to accelerate convergence, because the initial optimization at higher temperatures serves to smooth over irregularities of the dual function.

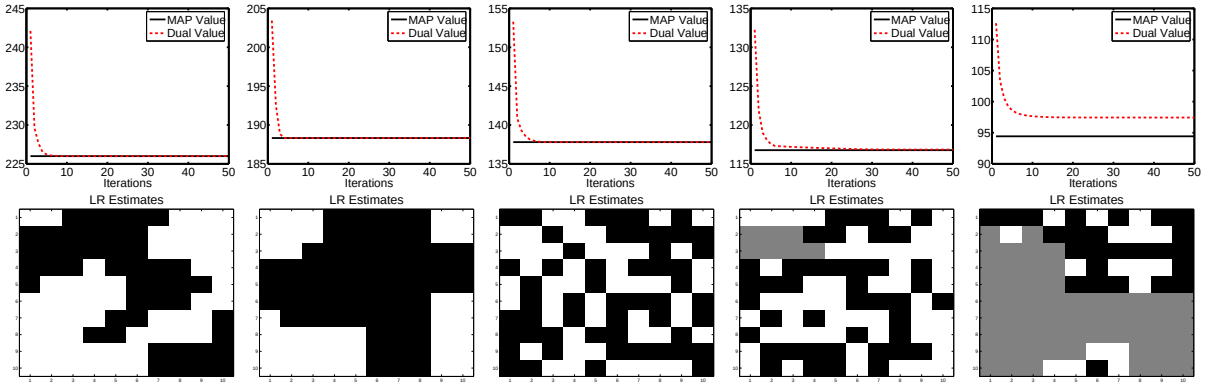

Fig. 5. Five examples for discrete LR showing: (top row) convergence of $g(\lambda)$ to g^* compared to f^* (horizontal line); (bottom row) the resulting estimates generated by relaxed max-marginals (grey areas denote non-unique maximum). The first two columns are examples of attractive models with $\sigma = 2$ and 1. The last three columns are frustrated models with $\sigma = 1.5, 1,$ and .7.

Computational Examples. In this section we provide some preliminary results using our approach to solve binary MRFs. These examples are for a binary model $x_v \in \{-1,+1\}$ defined on a 10×10 grid similar to the one seen in Fig. 2(a). For each node, we include a node potential $f_v(x_v) = \theta_v x_v$ with $\theta_v \sim N(0,\sigma^2)$. For each edge, we include an edge potential $f_{u,v}(x_u,x_v) = \theta_{uv} x_u x_v$ with $\theta_{uv} = 1$ in the "attractive" model and random $\theta_{uv} = \pm 1$ in the "frustrated" model. Hence, σ controls the strength of node potentials relative to edge potentials. As seen in Fig. 5, we obtain strong duality $g^* = f^*$ and recover the correct MAP estimates in attractive models. This is consistent with a result on optimality of TRMP in attractive models [10]. In the frustrated model, the same holds with strong node potentials, but as σ is decreased the frustration of the edge potentials cause a duality gap. However, even in these cases, we have observed that some nodes have a unique maximum in their re-summed max-marginals, and these nodes provide a partial MAP estimate that agrees with the correct global MAP estimate. This is apparently related to the *weak tree agreement* condition for partial optimality in TRMP [10].

IV. Gaussian Lagrangian Relaxation

In this section we apply the LR approach to the problem of MAP estimation in Gaussian graphical models, which is equivalent to maximizing a quadratic objective function

$$f(x;h,J) = -\frac{1}{2}x^T Jx + h^T x, \tag{26}$$

where $J \succ 0$ is sparse with respect to $\mathscr{G}$. Again, we construct an augmented model, which is now specified by an information form (h',J'), defined by a larger graph $\mathscr{G}'$. For consistency, we also require $f'(\zeta(x);h',J') = f(x;h,J)$ for all x. Denoting variable replication by $\zeta(x) = Ax$, this is equivalent to $A^T J'A = J$ and $A^T h' = h$. In order for the dual function to be well-defined, we also require that $J' \succ 0$. For general $J \succ 0$, it is possible that, for a given augmented graph $\mathscr{G}'$, there do not exist any $J' \succ 0$ defined on $\mathscr{G}'$ such that $A^T J'A = J$. To avoid this issue, we will focus on models that are of the form:

$$f(x) = \sum_{E \in \mathscr{F}} f_E(x_E) \tag{27}$$

where $\mathscr{F}$ is a hyper-graph, composed of cliques of $\mathscr{G}$, and each term $f_E(x_E)$ is itself a quadratic form $f_E(x_E) = -\frac{1}{2}x_E^T J_E x_E + h_E^T x_E$ based on $J_E \succ 0$. Then, $J = \sum_E [J_E]_V$ is the sum of these (zero-padded) submatrices. Then, it is simple to obtain a valid augmented model. We split each J_E between its replicas as $J_{E'} = r_E^{-1} J_E$ to obtain $J' = \sum_{E' \in \mathscr{F}'} [J_{E'}]_{V'} \succ 0$.

If there exists a representation of J in terms of 2×2 *pairwise* interactions $J_E \succ 0$, it is said to be *pairwise normalizable*. This condition is equivalent to the walk-summability condition considered in [17], which is related to the convergence (and correctness) of a variety of approximate inference methods [17], [18]. Here, we show that for the more general class of models of the form (27), we obtain a convergent iterative method for solving the dual problem that is tractable provided the cliques are not too large. Moreover, for this class of Gaussian models, we show that there is *no duality gap* and we always converge to the unique MAP estimate of the model. As an additional bonus, we also find that, by solving marginal agreement conditions in the augmented Gaussian model, we obtain a set of upper-bounds on the variances of each variable, although these bounds are often rather loose.

A. Gaussian LR with Linear Constraints

We begin by considering the Lagrangian dual of the following linearly-constrained quadratic program:

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2}x'^T J'x' + h'^T x' \\ \text{subject to} & x'_a = x'_b \text{ for all } a \equiv b. \end{array} \tag{28}$$

We may express the linear constraints on x' as $Hx' = 0$. Relaxing these constraints leads to the following dual function:

$$\begin{aligned} g(\lambda) &= \max_{x'} \{-\tfrac{1}{2}x'^T J'x' + (h' + H^T\lambda)x'\} \\ &= \tfrac{1}{2}(h' + H^T\lambda)^T J'^{-1}(h' + H^T\lambda) \end{aligned} \tag{29}$$

Moreover, by strong duality of quadratic programming [7], it holds that $g^* = f^*$. We also note the following equivalent representation of the dual problem:

$$g^* = \left\{ \begin{array}{ll} \text{minimize} & \frac{1}{2}h'^T J'^{-1}h' \\ \text{subject to} & A^T h' = h \end{array} \right. \tag{30}$$

ALGORITHM 2 (Gaussian LR)

Iterate until convergence:
For $E \in \mathscr{G}$ where $r_E > 1$
  For $E' \in \mathscr{R}(E)$
    Compute moments $(\hat{x}_{E'}, P_{E'})$ in (h', J').
    $\hat{J}_{E'} = P_{E'}^{-1}$, $\hat{h}_{E'} = P_{E'}^{-1} h_{E'}$
  end
  $\bar{J}_E = r_E^{-1} \sum_{E'} \hat{J}_{E'}$, $\bar{h}_E = r_E^{-1} \sum_{E'} \hat{h}_{E'}$
  For $E' \in \mathscr{R}(E)$
    $J'_{E',E'} \leftarrow J'_{E',E'} + (\bar{J}_E - \hat{J}_{E'})$
    $h'_{E'} \leftarrow h'_{E'} + (\bar{h}_E - \hat{h}_{E'})$
  end
end

Here, h' is the problem variable, and we consider all possible choices of h' that are consistent with h under the constraint $x' = Ax$. The optimal choice of h' in this problem is the one which leads to consistency in the estimate $\hat{x}' = J'^{-1} h'$.

B. Quadratic Constraints and Log-Det Regularization

Although, in Gaussian models, it is sufficient to include only linear constraints (there is no duality gap), our method can also accommodate quadratic constraints, and this results in faster convergence and tighter bounds on variances. Consider the constrained optimization problem:

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2} x'^T J' x' + h'^T x' \\ \text{subject to} & x_a = x_b, x_a^2 = x_b^2 \text{ for all } a \equiv b, \\ & x_{a_1} x_{a_2} = x_{b_1} x_{b_2} \text{ for all } (a_1, a_2) \equiv (b_1, b_2). \end{array} \tag{31}$$

This leads to the following equivalent version of the dual problem with problem variables (h', J'):

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} h'^T J'^{-1} h' \\ \text{subject to} & A^T h' = h,\ A^T J' A = J,\ J' \succ 0. \end{array} \tag{32}$$

Any solution of the linearly-constrained relaxation provides a feasible point for this problem, so the value of (32) is less than or equal to that of (30). However, since there is no duality gap in (30), the value of the two problem are equal, both achieve $g^* = f^*$ and obtain the MAP estimate.

While the choice of J' does not affect the value of the dual problem, it does effect variance estimates and convergence of iterative methods. Hence, we regularize the choice of J' by adding a penalty $-\frac{1}{2} \log\det J'$ to the objective of (32), which also serves as a barrier function enforcing $J' \succ 0$. The resulting objective function is then equivalent to $\Phi(h, J)$, which shows a parallel to our earlier approach for "smoothing" the dual function in discrete problems.

C. Gaussian Moment-Matching

We develop an approach in the same spirit as the Gaussian iterative scaling method [6]. We minimize the log-partition function with respect to the information parameters over all replicas of a node or edge, subject to consistency and positive definite constraints. The optimality condition for this minimization is that the marginal moments (means and variances) of all replicas are equalized. It can be shown that the following information-form updates achieve this objective. First, for all replicas E' of E, we compute the marginal information parameters given by sparse Gaussian elimination of $C = V' \setminus E'$ in (J', h'):

$$\begin{aligned} \hat{J}_{E'} &= J'_{E',E'} - J'_{E',C} (J'_{C,C})^{-1} J'_{C,E'} \\ \hat{h}_{E'} &= h'_{E'} - J'_{E',C} (J'_{C,C})^{-1} h'_C \end{aligned} \tag{33}$$

This is equivalent to $\hat{J}_{E'} = P_{E'}^{-1}$ and $\hat{h}_{E'} = P_{E'}^{-1} \hat{x}_{E'}$. Next, we average these marginal information forms over all replicas:

$$\bar{J}_E = r_E^{-1} \sum_{E'} \hat{J}_{E'},\ \bar{h}_E = r_E^{-1} \sum_{E'} \hat{h}_{E'} \tag{34}$$

Finally, we update the information form according to:

$$\begin{aligned} J'_{E',E'} &\leftarrow J'_{E',E'} + (\bar{J}_E - \hat{J}_{E'}) \\ h'_{E'} &\leftarrow h'_{E'} + (\bar{h}_E - \hat{h}_{E'}) \end{aligned} \tag{35}$$

Using the characterization of positive-definiteness of a block matrix in terms of a principle submatrix and its Schur complement, it can be shown that this update preserves positive definiteness of J'. It also preserves consistency, e.g., $\sum_{E'} (\bar{J}_E - \hat{J}_{E'}) = 0$. After the update, the new marginal information parameters for all replicas of E are equal to $(\bar{h}_E, \bar{J}_E)$. Algorithm 2 summarizes this iterative approach for solving the Gaussian LR problem.

Lastly, using the fact that $\bar{f}_E(x_E) \geq \hat{f}_E(x_E)$ for all x_E and that there is no duality gap upon convergence, we conclude that the final equalized marginal information must satisfy $\bar{J}_E \preceq \hat{J}_E \triangleq J_{E,E} - J_{E,\setminus E} (J_{\setminus E,\setminus E})^{-1} J_{\setminus E,E}$. Hence, LR gives an upper-bound on the true variance: $P_E = (\hat{J}_E)^{-1} \preceq (\bar{J}_E)^{-1}$. If each replica of E is contained in a separate connected component of $\mathscr{G}'$, then a tighter bound holds: $P_E \preceq (r_E \bar{J}_E)^{-1}$.

Computational Examples. We apply LR for two Gaussian models defined on a 50×50 2D grid with correlation lengths comparable to the size of the field. First, we use the *thin membrane* model, which encourages neighboring nodes to be similar by having potentials $f_{ij} = (x_i - x_j)^2$ for each edge $\{i, j\} \in \mathscr{G}$. We split the 2D model into vertical strips of narrow width K, which have overlap L (we vary K and set $L = 2$). We impose marginal agreement conditions in $K \times L$ blocks in these overlaps. The updates are done consecutively, from top to bottom blocks, from the left to the right strip. A full update of all the blocks constitutes one iteration. We compare LR to loopy belief propagation (LBP). The LBP variances are underestimates by 21.5 percent (averaged over all nodes), while LR variances for $K = 8$ are overestimates by 16.1 percent. In Figure 6 (top) we show convergence of LR for several values of K, and compare it to LBP. The convergence of variances is similar to LBP, while for the means LR converges considerably faster. In addition, the means in LR converge faster than using block Gauss-Seidel on the same set of overlapping $K \times 50$ vertical strips.

Next, we use the *thin plate model*, which enforces that each node v is close to the average of its nearest neighbors $N(v)$ in the grid, and penalizes curvature. At each node there is a potential: $f_i(x_i, x_{N(i)}) = (x_i - \frac{1}{|N(i)|} \sum_{j \in N(i)} x_j)^2$. LBP does not converge for this model. LR gives rather loose variance

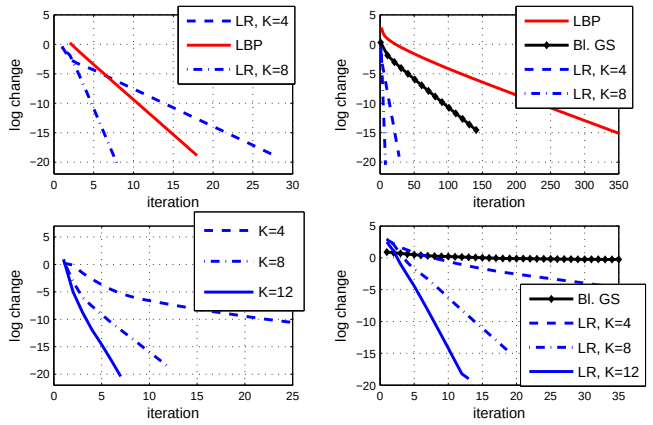

Fig. 6. Convergence plots for variances (left) and means (right), in the thin-membrane model (top) and thin-plate model (bottom).

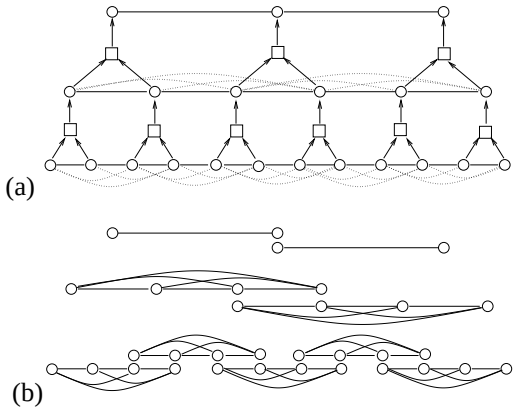

Fig. 7. Illustration of multiscale LR method. (a) First, we define an equivalent multiscale model subject to cross-scale constraints. Relaxing these constraints leads to a set of single-scale models. (b) Next, each single scale is relaxed to a set of tractable subgraphs.

bounds for this more difficult model: for $K = 12$, it overestimates the variances by 75.4 percent. More importantly, it accelerate convergence of the means. In Figure 6 (bottom) we show convergence plots for means and variances, for several values of K. As K increases, the agreement is achieved faster, and for $K = 12$ agreement is achieved in under 13 iterations for both means and variances. We note that LR with $K = 4$ converges much faster for the means than block Gauss-Seidel.

V. Multi-Scale Lagrangian Relaxation

In this section, we propose an extension of the LR method considered thus far. Previously, we have considered relaxations based on augmented models where $x' = \zeta(x)$ involves replication of variables. Here, we consider more general definition of ζ to allow the augmented model to include *summary variables*, such as a sum over a subset of variables, or any linear combination of these. In discrete models, summary variables can also be non-linear functions of x. For example, "parity bits" are used in coding applications and the "majority rule" is used to define coarse-scale binary variables in the renormalization group approach [19].

Using this idea, we develop a *multiscale* Lagrangian relaxation approach for MRFs defined on grids. The purpose of this relaxation is similar to that of the multigrid and renormalization group methods [20], [19]. Iterative methods generally involve simple rules that propagate information locally within the graph. Using a multiscale representation of the model allows information to propagate through coarse scales, which improves the rate of convergence to global equilibrium. Also, in discrete problems, such multiscale representations can help to avoid local minima. In the context of our convex LR approach, we expect this to translate into a reduction of the duality gap to obtain the optimal MAP estimate in a larger class of problems.

A. An Equivalent Multiscale Model

We illustrate the general idea with a simple example based on a 1D Markov chain. While this case is actually tractable by exact methods, it serves to illustrate our approach, which generalizes to 2D grids and 3D lattices. In Fig. 7, we show how to construct the augmented model $f'(x')$ defined on a graph $\mathscr{G}'$. This is done in two stages.

First, as illustrated in Fig. 7(a), we introduce coarse-scale representations of the fine scale variables by recursively defining summary variables at coarser scales to be functions of variables at the next level down. This defines a set of cross-scale constraints, denoted by the square nodes. To allow interactions between coarse-scale variables, while maintaining consistency with the original single-scale model, we introduce extra edges (the dotted ones in Fig. 7(a)) between blocks of nodes that have a (solid) edge between their summary nodes at the next coarser scale. This representation allows us to define a family of constrained multiscale models that are all equivalent to the original single-scale model. For 2D and 3D lattices, this model is still intractable even after relaxing the cross-scale constraints because each scale is itself intractable.

Next, to obtain a tractable dual problem, we break up the graph into smaller subgraphs, introducing additional constraints to enforce consistency among replicated variables. In the example, we break the augmented graph at each scale into its maximal cliques, shown in Fig. 7(b). This defines the final augmented model and the corresponding graph. In a 2D graph, the same idea applies, but we obtain a set of maximal cliques consisting of overlapping 2×4 and 4×2 blocks of the grid. Alternatively, we could break up the 2D grid into a set of width 2 vertical strips, as discussed previously.

Now, the procedure is essentially the same as before. We start with the equivalent constrained optimization problem defined on the augmented graph, now subject to both in-scale and cross-scale constraints. We obtain a tractable problem by introducing Lagrange multipliers to relax these constraints. Then we iteratively adjust the Lagrange multipliers to minimize the dual function, with the aim of eliminating constraint violations to obtain the desired MAP estimate. This is equivalent to adjusting the augmented model $f'(x')$ on $\mathscr{G}'$, subject to the constraint that it remains equivalent to $f(x)$ for all $x' = Ax$.

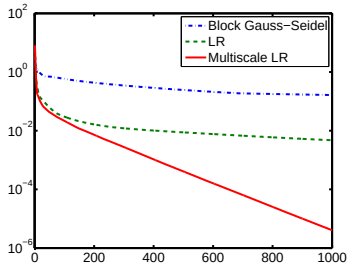

Fig. 8. Convergence of single- and multi-scale LR and block Gauss-Seidel.

B. Gaussian Multiscale Moment-Matching

We demonstrate this approach in the Gaussian model. To carry out the minimization, we again use a block coordinate-descent method that finds an exact minimum over a subset of Lagrange multipliers at each step. The replica constraints are handled the same as before. Here, we briefly summarize our approach to handle the cross-scale summary constraints. Let x_1 and x_2 denote two random vectors at consecutive scales coupled by the constraint $x_2 = Ax_1$. Let $(\hat{h}_1, \hat{J}_1)$ and $(\hat{h}_2, \hat{J}_2)$ denote their corresponding *marginal* information parameters. Relaxing the constraints $x_2 = Ax_1$ and $x_2 x_2^T = Ax_1 x_1^T A^T$, with Lagrange multipliers $(\lambda, -\frac{1}{2}\Lambda)$, leads to the following optimality conditions:

$$\begin{aligned} (\hat{J}_2 + \Lambda)^{-1} &= A(\hat{J}_1 - A^T \Lambda A)^{-1} A^T \qquad (36) \\ (\hat{J}_2 + \Lambda)^{-1}(\hat{h}_2 + \lambda) &= A(\hat{J}_1 - A^T \Lambda A)^{-1}(\hat{h}_1 - A^T \lambda) \end{aligned}$$

We find that the solution is:[4]

$$\begin{aligned} \Lambda &= \tfrac{1}{2}\{(A\hat{J}_1^{-1}A^T)^{-1} - \hat{J}_2\} \\ \lambda &= \tfrac{1}{2}\{(A\hat{J}_1^{-1}A^T)^{-1} A\hat{J}_1^{-1}\hat{h}_1 - \hat{h}_2\} \qquad (37) \end{aligned}$$

The model (h', J') is then updated by adding (λ, Λ) to the coarse-scale and subtracting $(A^T\lambda, A^T\Lambda A)$ from the fine scale. This update enforces the moment conditions $\hat{x}_2 = A\hat{x}_1$ and $P_2 = AP_1A^T$ while maintaining consistency of the model (h', J'). Similar updates can be derived when there are multiple replicas of x_1 and x_2. These methods, together with those described previously, are used to minimize the dual function in the Gaussian multiscale relaxation.

Multiscale Example. We provide a preliminary result involving a 1D thin-membrane model with 1024 nodes. It is defined to have a long correlation length comparable to the length of the field. Using a random h-vector, we solve for the MAP estimates using three methods: a standard block Gauss-Seidel iteration using overlapping blocks of size 4; the (single-scale) Gaussian LR method with the same choice of blocks; and the multiscale LR method. The convergence of all three methods are shown in Fig. 8. We see that the single-scale LR approach is moderately faster than block Gauss-Seidel, but introducing coarser-scales into the method leads to a significant speed-up in the rate of convergence.

[4]The formula (37) corresponds to a generalization of Algorithm 2, in which the moments $(\hat{x}_1, P_1)$ of fine-scale variables x_1 are replaced by the corresponding moments $(A\hat{x}_1, A\hat{P}_1A^T)$ of the summary statistic $\tilde{x}_1 = Ax_1$.

VI. Discussion

We have introduced a general Lagrangian relaxation framework for MAP estimation in both discrete and Gaussian graphical models. This provides a new interpretation of some existing methods, provides deeper insights into those methods, and leads to new generalizations, such as the multiscale relaxation introduced here. There are many promising directions for further work. While we have considered discrete and Gaussian models separately, the basic approach should extend to the richer class of conditionally Gaussian models [1] including both discrete and continuous variables. In discrete models, designing augmented models that capture more structure of the original problem leads to reduced duality gaps and optimal MAP estimates in larger classes of models. It would be of great interest to finds ways to *adaptively* search this hierarchy of relaxations to efficiently reduce and eventually eliminate the duality gap with minimal computation. It is also of interest to consider approaches to identity provably *near-optimal* estimates, perhaps using the relaxed max-marginal estimates, in cases where it is not tractable to completely eliminate the duality gap.

References

[1] S. Lauritzen. *Graphical Models*. Oxford Univ. Press, 1996.

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

[3] B. Frey, editor. *Graphical models for machine learning and digital communication*. MIT Press, 1998.

[4] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Dept. of Stat., UC Berkeley, 2003.

[5] A. Dempster. Covariance selection. *Biometrics*, 28(1), March 1972.

[6] T. Speed and H. Kiiveri. Gaussian Markov distributions over finite graphs. *Annals Stat.*, 14(1), March 1986.

[7] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

[8] D. Bertsimas and J. Tsitsiklis. *Linear Opitimization*. Athena Scientific, 1997.

[9] M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on (hyper)trees: Message-passing and linear-programming approaches. *IEEE Trans. Information Theory*, November 2005.

[10] V. Kolmogorov and M. Wainwright. On the optimality of tree-reweighted max-product message passing. In *Proc. Conf. Uncertainty in AI*, July 2005.

[11] J. Feldman, M. Wainwright, and D. Karger. Using linear programming to decode binary linear codes. *IEEE Trans. Information Theory*, 51, March 2005.

[12] C. Yanover, T. Meltzer, and Y. Weiss. Linear programming relaxations and belief propagation. *J. Machine Learning Research*, 7, 2006.

[13] T. Werner. A linear programming approach to max-sum, a review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, July 2007.

[14] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[15] L. Ruschendorf. Convergence of the iterative proportional fitting procedure. *Annals Stat.*, 23, 1995.

[16] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, January 2005.

[17] D. Malioutov, J. Johnson, and A. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *J. Machine Learning Research*, 7, October 2006.

[18] V. Chandrasekaran, J. Johnson, and A. Willsky. Estimation in Gaussian graphical models using tractable subgraphs: a walk-sum analysis. *IEEE Trans. Signal Processing*, to appear.

[19] B. Gidas. A renormalization group approach to image processing problems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11, February 1989.

[20] U. Trottenberg, C. Oosterlee, and A. Schuller. *Multigrid*. Academic Press, 2001.