Lagrangian Relaxation for MAP Estimation in Graphical Models

Jason Johnson*
Dmitry Malioutov
Alan Willsky

45th Allerton Conference on Communication, Control and Computing

September 26, 2007

*jasonj@mit.edu, //ssg.mit.edu/group/jasonj.

Introduction

- MAP estimation on an intractable graph.
  - Reformulate on tractable graph, but with complicating constraints.
  - Relaxing the constraints leads to a tractable, convex dual problem.
- Related Work:
  - Tree-Reweighted Max-Product (Wainwright, Jaakkola, Willsky, IT '05).
  - Convergent TRMP (Kolmogorov, PAMI '06).
  - Max-Sum Diffusion (Werner, PAMI '07).
  - LP Approaches (Feldman, Wainwright, Karger, IT '03; Yanover, Meltzer, Weiss, JMLR '06).
- Our Contribution:
  - Simple, general picture.
  - Smooth optimization method for discrete problems.
  - Similar approach for Gaussian problems.
  - New class of *multi-scale* relaxations (inspired by group renormalization and multigrid).

Graphical Models

A collection of random variables $x = (x_v, v \in V)$ with probability distribution:

$$p(x) = \exp\left\{\frac{1}{\tau}\left(f(x) - \Phi(f,\tau)\right)\right\}$$

with *potential*

$$f(x) = \sum_{E \in \mathcal{G}} f_E(x_E)$$

where $\mathcal{G} \subset 2^V$ defines a (hyper)graph with (hyper)edges $E \in \mathcal{G}$.

The *temperature* $\tau > 0$ controls the level of randomness. Probability concentrates on MAP estimate as τ approaches zero.

The normalization function Φ is the *log-partition function* of statistical mechanics, *cumulant generating function* in statistics.

Boltzmann Machine

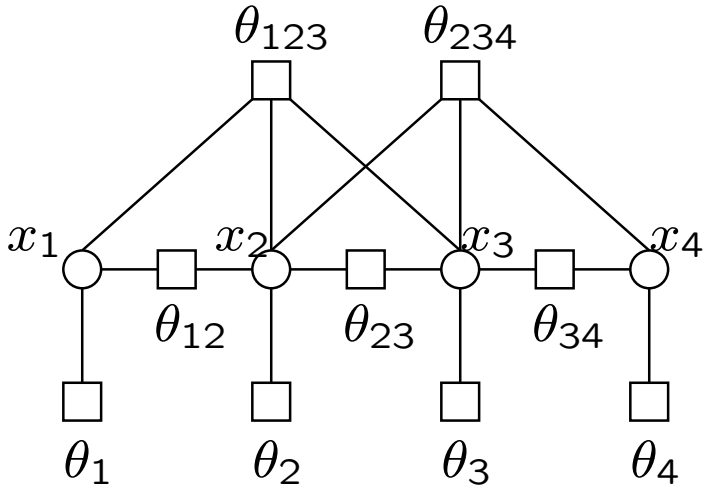
Let $x_v \in \{0, 1\}$ and

$$p(x) \propto \exp f(x) = \exp \sum_{E \in \mathcal{G}} \theta_E \phi_E(x_E)$$

where $\phi_E(x_E) = \prod_{v \in E} x_v$. E.g.,

$$\begin{aligned} V &= \{1, 2, 3, 4\} \\ \mathcal{G} &= \{1, 2, 3, 4, 12, 23, 34, 123, 234\} \\ \vartheta(x) &= \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 \\ &\quad + \theta_{12} x_1 x_2 + \theta_{23} x_2 x_3 + \theta_{34} x_3 x_4 \\ &\quad + \theta_{123} x_1 x_2 x_3 + \theta_{234} x_2 x_3 x_4 \end{aligned}$$

Factor graph representation:

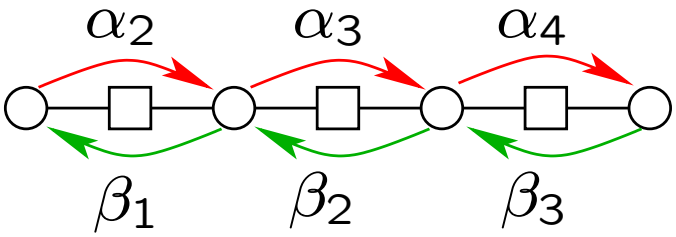


MAP Estimation

Find global configuration $x^* \in \{0,1\}^n$ which maximizes the potential $f(x)$. Consider the Markov chain:

$$f(x_1, \ldots, x_n) = \sum_{t=1}^{n-1} f(x_t, x_{t+1})$$

Max-Sum Algorithm:



$$\begin{aligned} \alpha_t(x_t) &= \max_{x_{t-1}} \{f(x_{t-1}, x_t) + \alpha_{t-1}(x_{t-1})\} \\ \beta_t(x_t) &= \max_{x_{t+1}} \left\{f(x_t, x_{t+1}) + \beta_{t+1}(x_{t+1})\right\} \end{aligned}$$

Computes *max-sum marginal* $\widehat{f}_t = \alpha_t + \beta_t$ of each variable, determines $x_t^* = \arg\max \widehat{f}_t$.

Generalizes to trees and "thin" graphs, complexity *exponential* in width of the graph.

Gaussian Model

Information form of Gaussian density:

$$p(x) \propto \exp\{-\frac{1}{2}x^T Jx + h^T x\}$$

related to mean μ and covariance P by:

$$J = P^{-1}, \;\; h = P^{-1}\mu$$

Graph $\mathcal{G}$ defined by fill-pattern of J.

Max-Sum reduces to *Gaussian elimination*:

$$\begin{aligned} \widehat{J}_A &= J_{AA} - J_{AB}(J_{BB})^{-1}J_{BA} \\ \widehat{h}_A &= h_A - J_{AB}(J_{BB})^{-1}h_B \end{aligned}$$

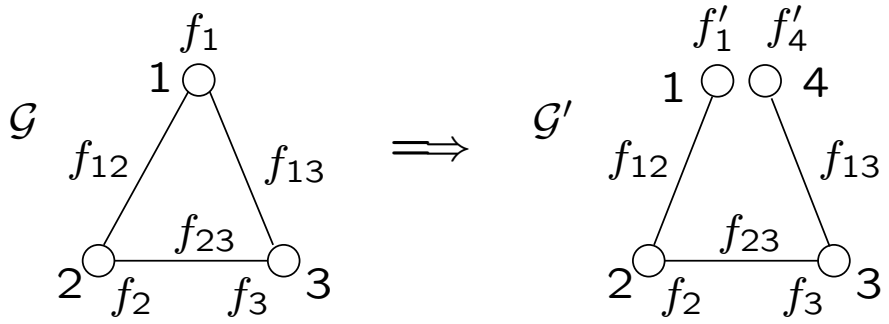
Sparse GE has complexity *cubic* in the width of the graph.

Lagrangian Relaxation: A Toy Example

Consider 3-node cycle, with objective

$$f = f_1 + f_2 + f_3 + f_{12} + f_{23} + f_{13}$$

Let x_4 be a "replica" of x_1, define f' on 4-node chain by $f_1 \rightarrow f'_1 + f'_4$.



Note that

$$f^* \triangleq \max f(x) = \max\{f'(x) | x_1 = x_4\} \leq \max f'(x)$$

Dual Problem:

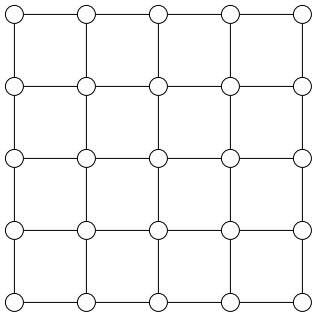
$$g^* \triangleq \min_{\lambda} \max_{x} \{f'(x) + \lambda(x_1 - x_4)\} \geq f^*$$

Minimizes upper bound on f^*. If $x_1^* = x_4^*$, then $g^* = f^*$ and x^* is the correct MAP estimate.

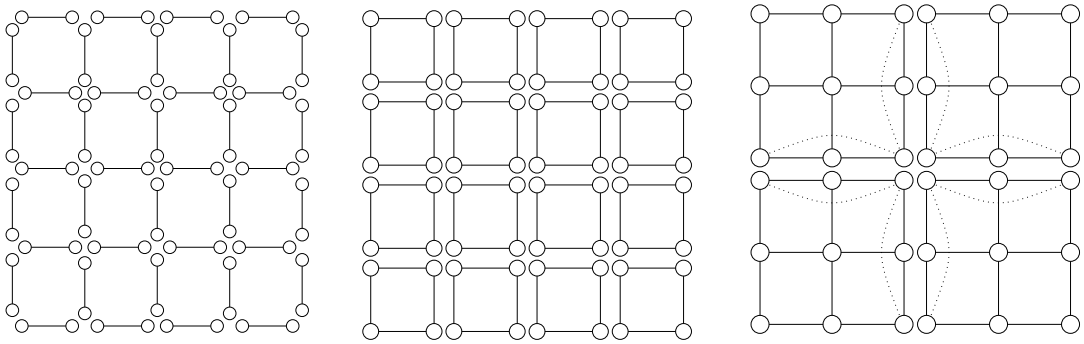
Graphical Decomposition

Replicate nodes/edges of $\mathcal{G}$ to obtain a tractable graph $\mathcal{G}'$ comprised of small/thin components.

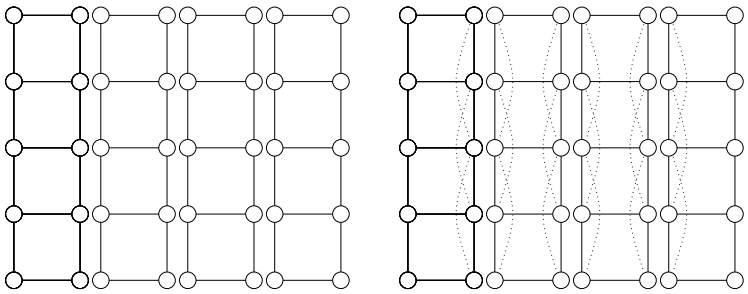
For example, take a 2D grid model:



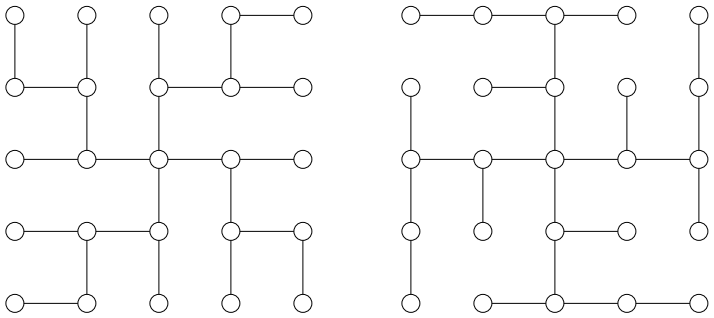
Split into disconnected edges or blocks:



Split into thin strips:



Split into trees:



Tree-decomposition is essentially equivalent to TRMP approach of Martin Wainwright, which considers convex combinations of trees.

Lagrangian Relaxation

MAP Problem:

$$f^* = \max_x f(x), \;\; f(x) = \sum_{E \in \mathcal{G}} f_E(x_E)$$

Define $f'(x')$ on $\mathcal{G}'$ such that $f'(x') = f(x)$ for all *consistent* x' (split f_E among $f'_{E'}$).

MAP equivalent to:

$$\begin{array}{ll} \max & f'(x') \\ \text{s.t.} & \phi_A(x') = \phi_B(x') \text{ for all } A, B \equiv E \end{array}$$

requires that replicated statistics agree.

Relax the replica constraints:

$$f'(x'; \lambda) \triangleq f'(x') + \sum_{A,B} \lambda_{A,B}(\phi_A(x') - \phi_B(x'))$$

Each λ corresponds to a valid *reparameteriza-tion* of f', still consistent with $f(x)$.

The Dual Problem

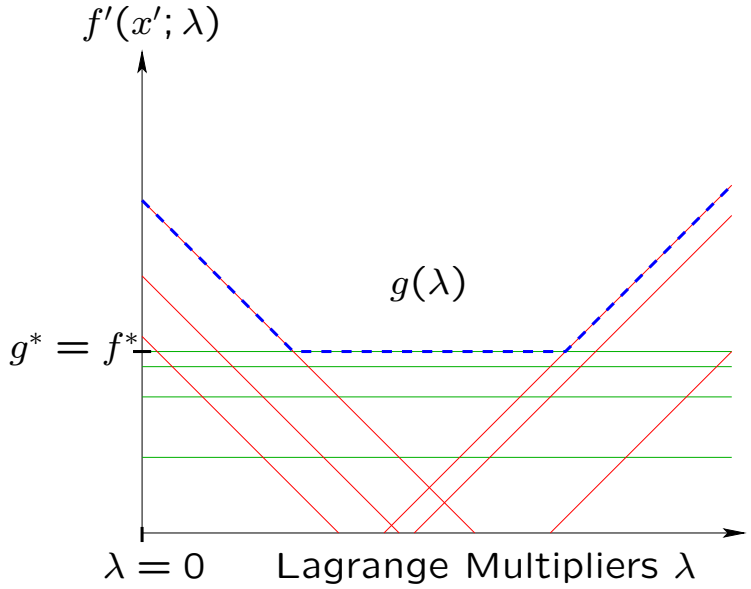
The *dual function* is now tractable to compute:

$$g(\lambda) \triangleq \max_{x'} f'(x'; \lambda)$$

This requires max-sum computations on the tractable graph $\mathcal{G}'$. Note that $g(\lambda) \geq f^*$.

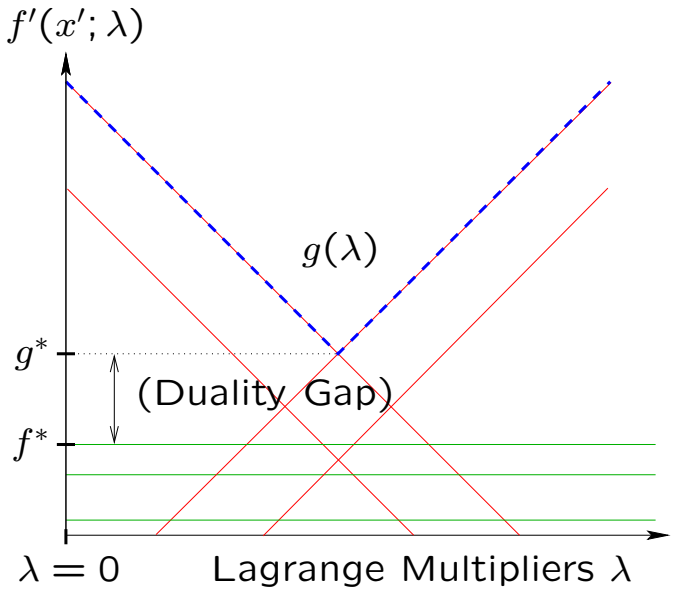
Dual Problem: $\min g(\lambda) \triangleq g^*$ to obtain the tightest bound on f^*. *Strong duality* if $g^* = f^*$.

Geometric picture:



Duality Gap

The invalid x' can "hide" the consistent ones...



Let $\lambda^* \in \arg\min g(\lambda)$. Then, either:

(1) Exists consistent $x' \in \arg\max f'(x';\lambda^*)$. Then, $g^* = f^*$ and MAP estimate is obtained.

(2) There is no consistent $x' \in \arg\max f'(x';\lambda^*)$. Then, $g(\lambda) > f^*$ for *all* λ and we cannot recover MAP estimate.

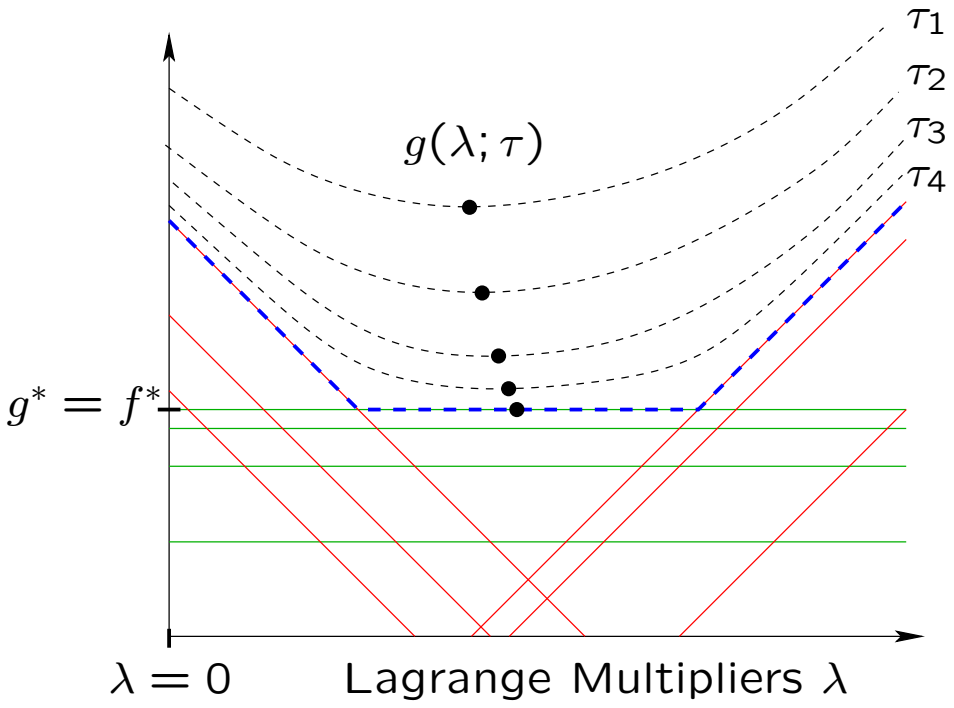
Smooth Dual Problem

Use "log-sum-exp" approximation of "max" function (Boyd, Vandenberghe):

$$g(\lambda;\tau) \triangleq \tau \log \sum_{x'} \exp\left(\frac{f'(x';\lambda)}{\tau}\right) \geq g(\lambda)$$

Uniform convergence to $g(\lambda)$ as $\tau \to 0$.

Idea: minimize $g(\lambda;\tau)$ for decreasing τ. A kind of *interior point method* to minimize $g(\lambda)$.



Convex Optimization Method

Note, $g(\lambda;\tau)$ is the normalization function of the distribution

$$p(x';\lambda,\tau) = \exp\{\tfrac{1}{\tau}(f'(x';\lambda) - g(\lambda;\tau))\}$$

at temperature τ.

Moment-Generating Property:

$$\frac{\partial g(\lambda;\tau)}{\partial \lambda_{A,B}} = p[\phi_A] - p[\phi_B]$$

The minimum over all λ's that couple replicas $E' \in \mathcal{G}'$ of an edge $E \in \mathcal{G}$ obtained when *marginals agree* for all E'.

We develop *iterative scaling* method, adaptation of methods used for *learning* graphical models...

Max-Entropy Duality and Linear-Programming Connection

Smooth Dual Problem:

$$\begin{array}{ll} \min_{f' \sim \mathcal{G}'} & \Phi(f', \tau) \\ \text{s.t.} & f'(x') = f(x) \text{ for all } x' \equiv x. \end{array}$$

Note, $\lim_{\tau \to 0} \Phi(f', \tau) = \max_{x'} f'(x')$.

Maximum Entropy Duality:

$$\begin{array}{ll} \max_{p(x')} & p[f] + \tau H'(p) \\ \text{s.t.} & p(x'_A) = p(x'_B) \text{ for replicas A,B} \end{array}$$

Equivalent to optimization over *local marginal polytope* (Wainwright) of $\mathcal{G}$.

Reduces to LP as $\tau \to 0$, H' serves as *barrier function* for local marginal polytope.

Iterative Scaling: (Csiszar) sequence of information projections (minimum relative entropy), each projection imposes a subset of constraints.

Iterative Marginal Matching

Algorithm: At temperature τ:

Compute marginal potentials:

$$\tilde{f}_k = \tau \log \tilde{p}_k$$

Average:

$$\bar{f}_k = \frac{1}{K} \sum_k \tilde{f}_k$$

Reparameterize:

$$f' \leftarrow f' + (\bar{f} - \tilde{f}_k)$$

Iterate over constraints until convergence.

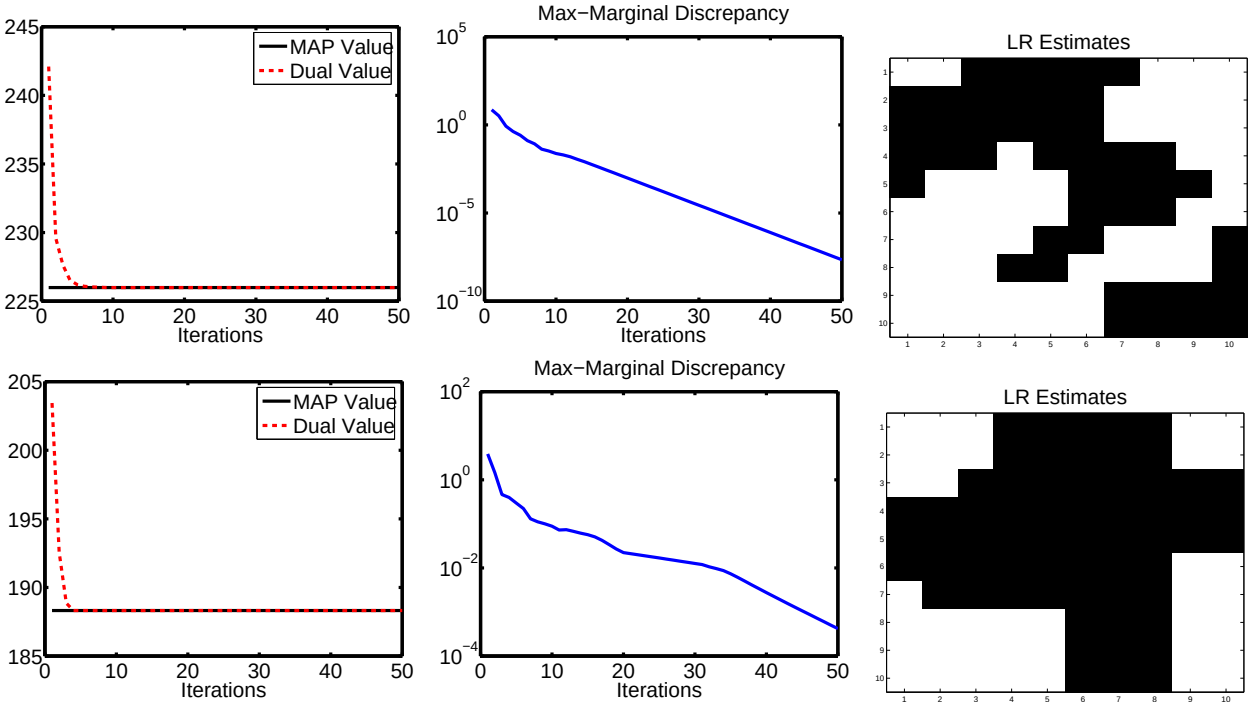
Let $\tau \rightarrow 0$ to minimize $g(\lambda)$. Reduces to *max-sum diffusion*, related to convergent TRMP.

Smoothed approach avoids *non-minimal fixed points* observed by Kolmogorov.

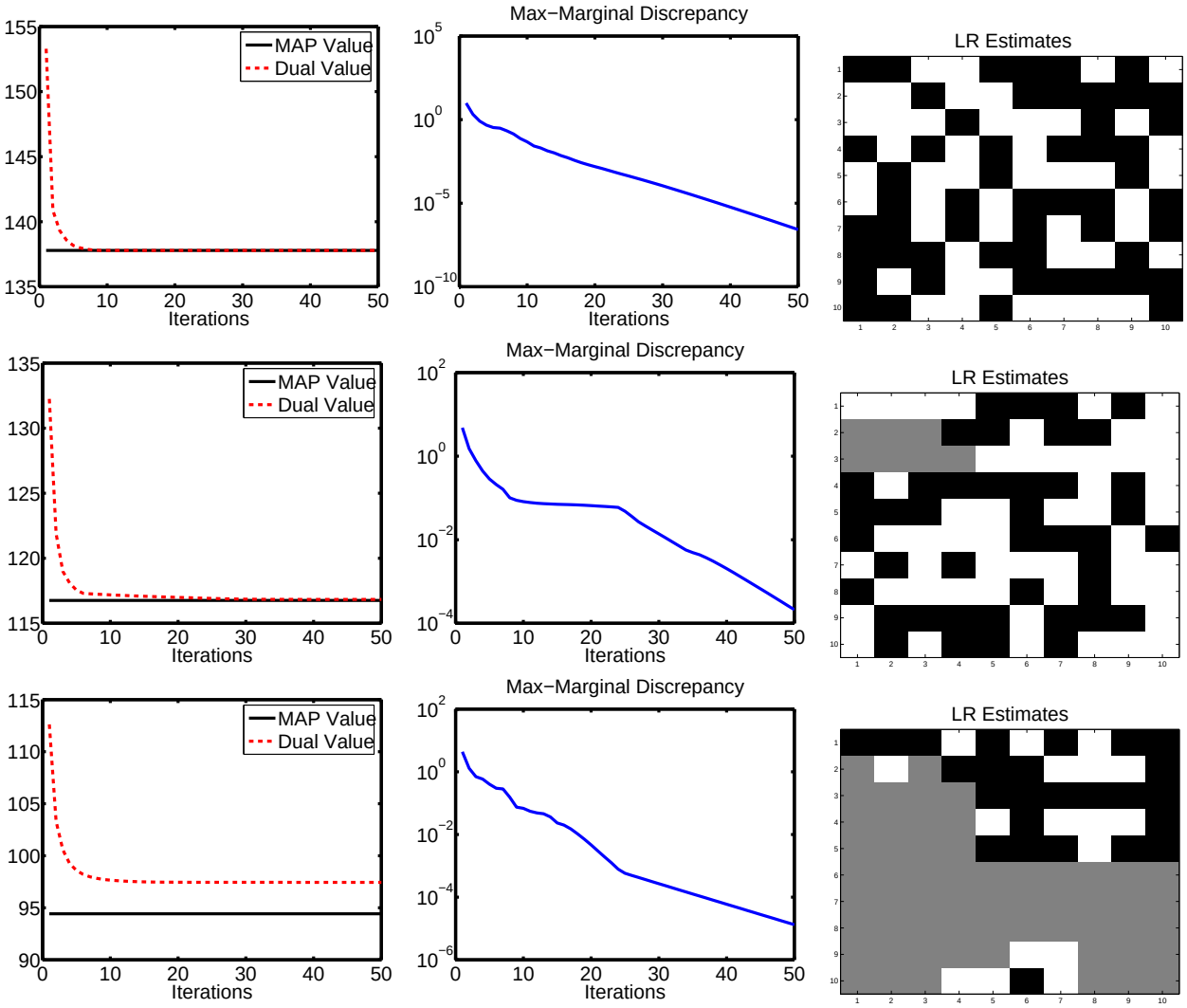
Discrete Example

10×10 Random-field Ising model, vary coupling strength.

"Attractive" Models (no duality gap):



"Frustrated" Model (exhibits duality gap):



Gaussian LR

Let $J = \sum_C [J_C]_V$, $J_C \in \mathbb{R}^{|C|\times|C|}$ and $J_C \succ 0$.

Equivalent model (h', J') on $\mathcal{G}'$:

$$A^T h' = h, \;\; A^T J' A = J, \;\; J' \succ 0$$

where $x' = Ax$ denotes variable replication.

MAP Problem:

$$\begin{array}{ll} \max_{x'} & -\frac{1}{2} x'^T J' x' + h'^T x' \\ \text{s.t.} & x'_a = x'_b, \text{ replicas } a, b \end{array}$$

Linearly-constrained convex QP, no duality gap!

Dual problem:

$$\begin{array}{ll} \min_{h'} & \frac{1}{2} h'^T J'^{-1} h' \\ \text{s.t.} & A^T h' = h \end{array}$$

Regularized Dual Problem:

$$\begin{array}{ll} \min & \frac{1}{2}(h'^T J'^{-1} h' - \log\det J') \\ \text{s.t.} & A^T h' = h,\, A^T J' A = J,\, J' \succ 0 \end{array}$$

Solve by marginal matching algorithm...

Gaussian Moment-Matching

Replica constraints relax to moment constraints:

$\mu_k = p[x_{E_k}]$ and $P_k \triangleq p[(x_{E_k} - \mu_k)(x_{E_k} - \mu_k)']$ equal across all replicas E_k of $E \in \mathcal{G}$.

Algorithm:

Compute marginal potentials:

$$\hat{h}_k = P_k^{-1}\mu_k, \;\; \hat{J}_k = P_k^{-1}$$

Average:

$$\bar{h} = \tfrac{1}{K}\sum_k \hat{h}_k, \;\; \bar{J} = \tfrac{1}{K}\sum_k \hat{J}_k$$

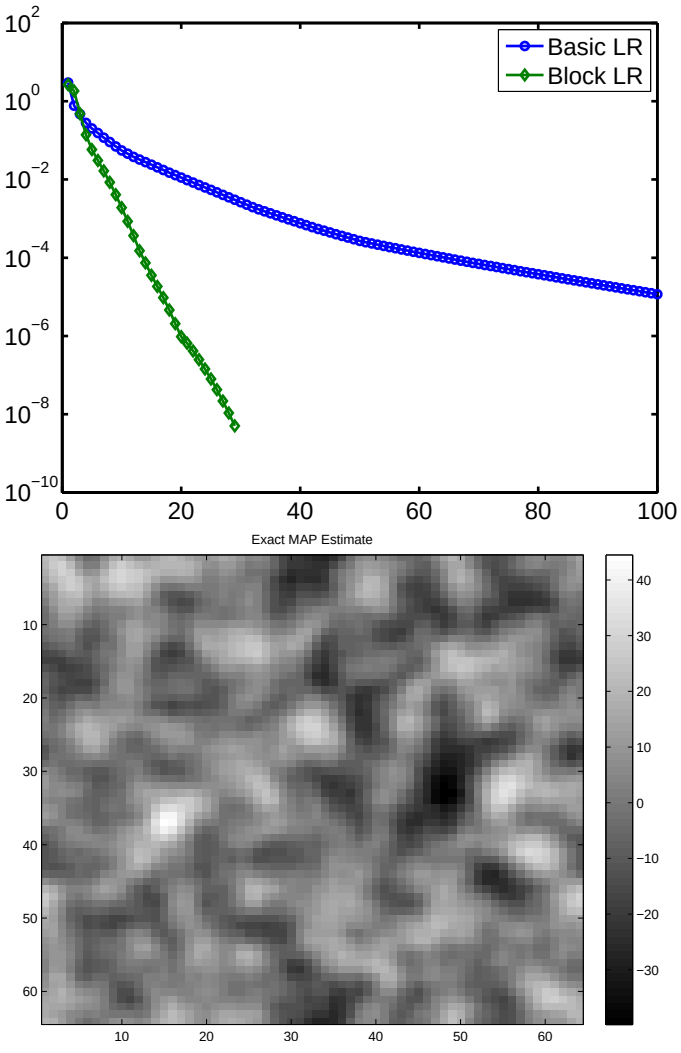
Reparameterize:

$$\begin{aligned} h' &\leftarrow h' + (\bar{h} - \hat{h}_k) \\ J' &\leftarrow J' + (\bar{J} - \hat{J}_k) \end{aligned}$$

Iterate over all constraints until convergence.

Gaussian Example

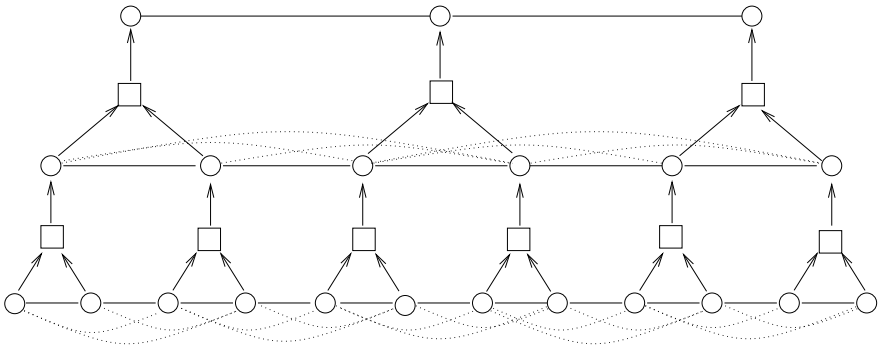
64×64 random-field thin-plate model (penalizes curvature).



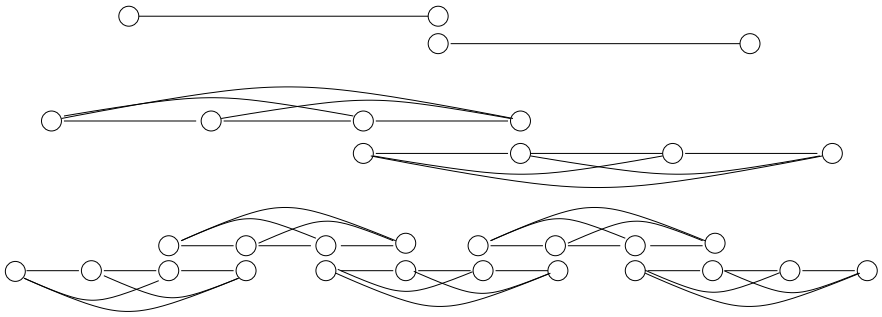
Slower convergence for stronger correlations...

Multi-scale Reparameterization

Define a family of constrained multi-scale models that are *equivalent* to the original MRF.



Relax the cross-scale constraints and break up each level into tractable subgraphs.



Dual Problem minimize the MAP value over all equivalent multi-scale reparameterizations.

Generalized Gaussian LR

Allow general linear constraints on "replicas":

$$\tilde{x}_k \triangleq A_k x_{E_k} \text{ equal for all } E_k \in \mathcal{E}_c$$

Relaxes to moment constraints:

$$\tilde{\mu}_k \triangleq A_k \mu_k \text{ and } \tilde{P}_k \triangleq A_k P_k A_k'$$

must be equal across replicas.

Algorithm:

Compute marginal potentials:

$$\tilde{h}_k = \tilde{P}_k^{-1} \tilde{\mu}_k, \;\; \tilde{J}_k = \tilde{P}_k^{-1}$$

Average:

$$\bar{h} = \frac{1}{K} \sum_k \tilde{h}_k, \;\; \bar{J} = \frac{1}{K} \sum_k \tilde{J}_k$$

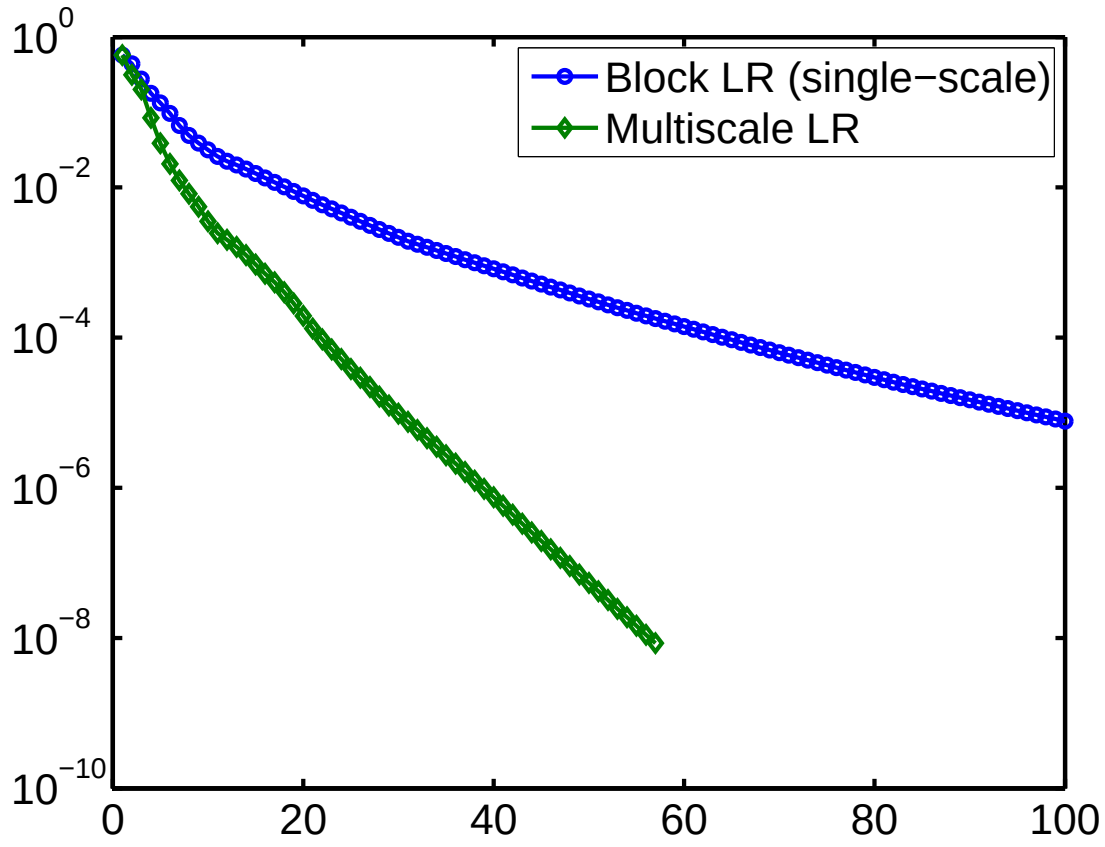
Reparameterize:

$$\begin{aligned} h' &\leftarrow h' + A_k^T (\bar{h} - \tilde{h}_k) \\ J' &\leftarrow J' + A_k^T (\bar{J} - \tilde{J}_k) A_k \end{aligned}$$

Iterate over all constraints until convergence.

Multi-Scale Gaussian Example

128×128 random-field thin membrane model (penalizes image gradient).



Conclusion

Convex approach to MAP estimation using "diffusion" rule to propagate information between tractable subgraphs.

Multi-scale method can accelerate convergence in large graphs with strong correlations.

Challenges:

- ♠ Multi-scale Approach:
  - ♢ *design* coarsening operator
  - ♢ *stochastic* upwards model
- ♠ Discrete Models:
  - ♢ characterize "easy" models in block LR
  - ♢ characterize strong duality in multi-scale approach
  - ♢ *adaptively* enhance formulation to reduce gap
  - ♢ *near-optimal* estimates for hard problems