Lagrangian Relaxation Methods for Intractable Graphical Models

Jason K. Johnson

(Joint work with Dmitry Malioutov and Alan Willsky)

September 21, 2005

Outline

- Graphical Models
- MAP Estimation
- Lagrangian Relaxation
- Toy Examples

Graphical Models

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

$$\mu_\tau(x) = \exp\left\{\frac{1}{\tau}(\vartheta(x) - \Phi_\tau(\vartheta))\right\}$$

with *potential*

$$\vartheta(x) = \sum_{E \in \mathcal{G}} \vartheta_E(x_E)$$

where $(V, \mathcal{G})$ defines a (hyper)graph with (hyper)edges $\mathcal{G} \subset 2^V$. The *temperature*, $\tau > 0$, controls the level of randomness.

The *Helmholtz free energy** normalizes the distribution:

$$\Phi_\tau(\vartheta) = \tau \log \sum_x \exp\left\{\frac{\vartheta(x)}{\tau}\right\}$$

and plays an important role in analysis of families of models in this form.

*Convex conjugate of Gibbs free energy. Also known as the *log-partition function* or *cumulant generating function*.

Boltzmann Machine

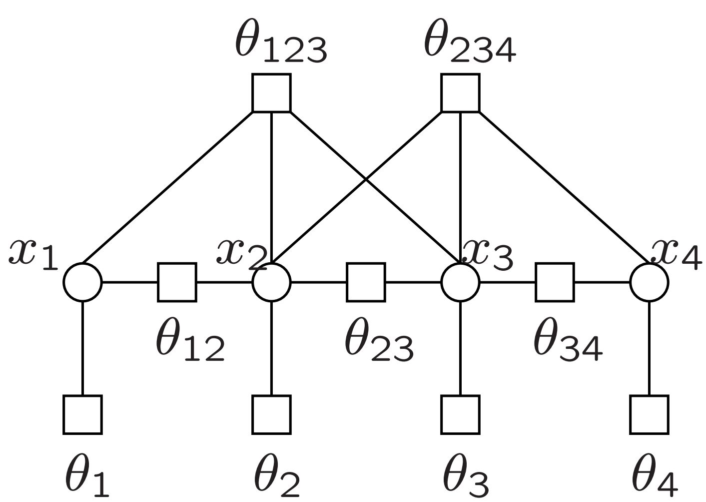
Let $x_v \in \{0,1\}$ and

$$P_\tau(x) \propto \exp \vartheta(x) = \exp \sum_{E \in \mathcal{G}} \theta_E \phi_E(x_E)$$

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:

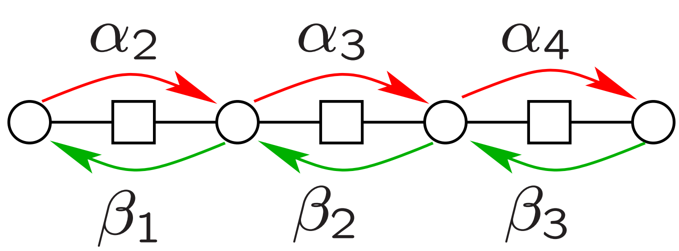


MAP Estimation

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

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

Forwards-Backwards (Viterbi) Algorithm:



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

Compute *max-marginal* of each variable:

$$\begin{aligned}
\widehat{\vartheta}_t(x_t) &\equiv \max_{x_{\setminus t}} \vartheta(x_t, x_{\setminus t}) \\
&= \alpha_t(x_t) + \beta_t(x_t)
\end{aligned}$$

which determines $x_t^* = \arg\max \widehat{\vartheta}_t$.

Useful Properties of $\Phi_\tau(\vartheta)$

Convexity $\Phi_\tau(\sum_k \rho_k \vartheta_k) \leq \sum_k \rho_k \Phi_\tau(\vartheta_k)$

Sub-Additivity $\Phi_\tau(\sum_k \vartheta_k) \leq \sum_k \Phi_\tau(\vartheta_k)$

Moment-Generating Property

$$\frac{\partial \Phi_\tau(\vartheta)}{\partial \vartheta_E(x_E)} = \mu_\tau(x_E) \equiv \sum_{x_{\setminus E}} \mu_\tau(x_E, x_{\setminus E})$$

Zero-Temperature Limit

$$\Phi_\tau(\vartheta) \to \Phi_0(\vartheta) \equiv \max_x \vartheta(x) \quad (\tau \to 0^+)$$

Marginal-Potentials $\rightarrow$ *Max-Marginals*

$$\begin{aligned} \widehat{\vartheta}_\tau(x_E) &\equiv \tau \log \sum_{x_{\setminus E}} \exp\left\{\frac{\vartheta(x_E, x_{\setminus E})}{\tau}\right\} \\ &\rightarrow \max_{x_{\setminus E}} \vartheta(x_E, x_{\setminus E}) \quad (\tau \to 0^+) \end{aligned}$$

Lagrangian Relaxation

Decompose the graphical model into M tractable subgraph models $\{\mathcal{G}_k, \vartheta_k\}$ such that

$$\begin{aligned}
\mathcal{G} &= \cup_k \mathcal{G}_k \\
\vartheta(x) &= \sum_k \vartheta_k(x), \quad \forall x \\
\vartheta_k(x) &= \sum_{E \in \mathcal{G}_k} \vartheta_E^{(k)}(x_E)
\end{aligned}$$

We define the *Lagrangian* ($M = 2$) by

$$\begin{aligned}
L(u, v; \lambda) &= \vartheta_1(u) + \vartheta_2(v) + \langle \lambda, \phi(u) - \phi(v) \rangle \\
&= \{\vartheta_1(u) + \lambda(u)\} + \{\vartheta_2(v) - \lambda(v)\} \\
&= \vartheta_1(x; \lambda) + \vartheta_2(x; \lambda)
\end{aligned}$$

where $\lambda(x) \equiv \langle \lambda, \phi(x) \rangle$ corresponds to constraints $\phi(u) = \phi(v)$; e.g., in discrete models

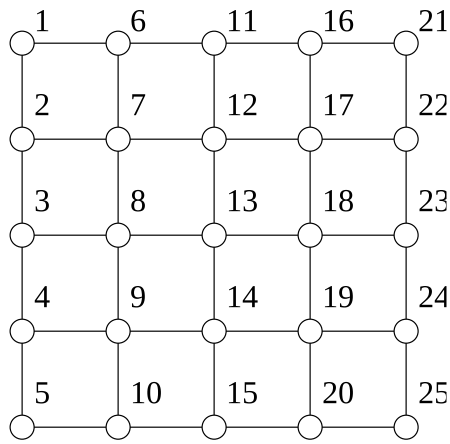
$$\delta_{x_E}(u) = \delta_{x_E}(v)$$

for all $E \in \mathcal{G}_1 \cap \mathcal{G}_2$ and $x_E \in \mathcal{X}_E$, which gives

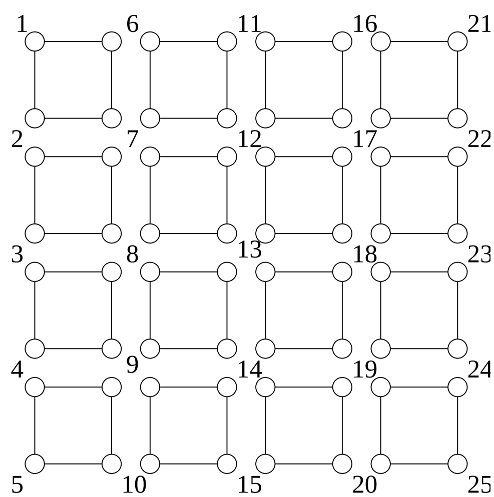
$$\lambda(x) = \sum_{E \in \mathcal{G}_1 \cap \mathcal{G}_2} \lambda_E(x_E)$$

Graphical Decompositions

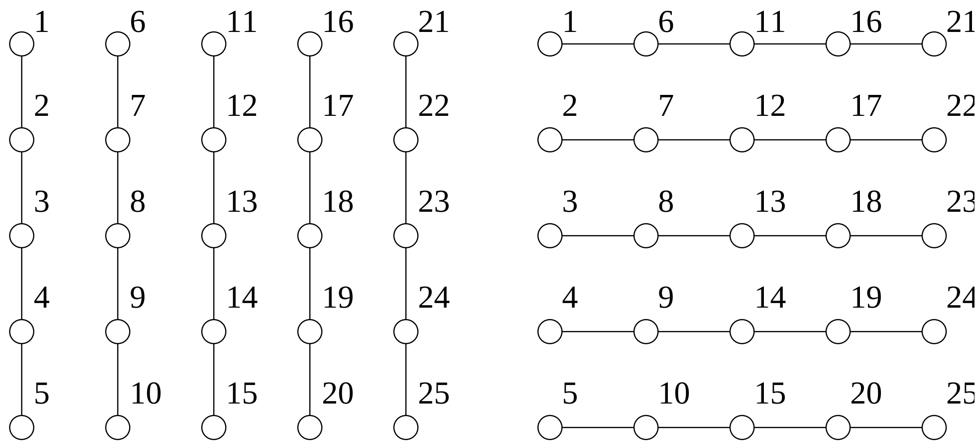
Original Graph



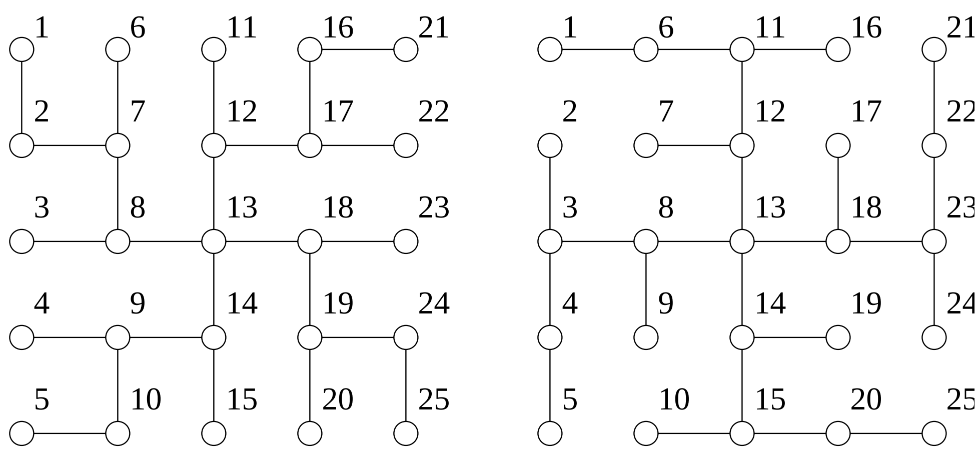
Decomposition into 2×2 subcells



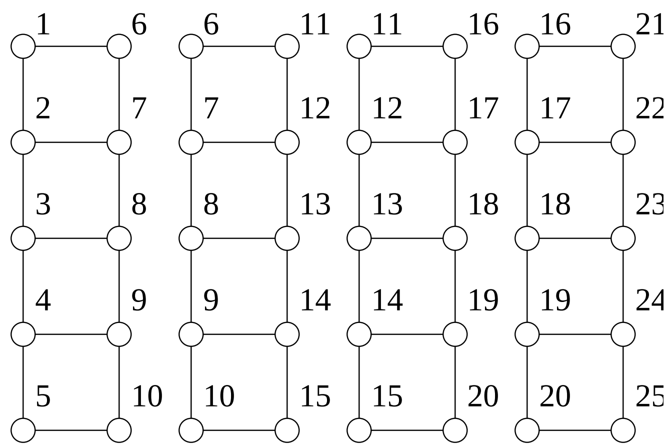
Embedded Trees



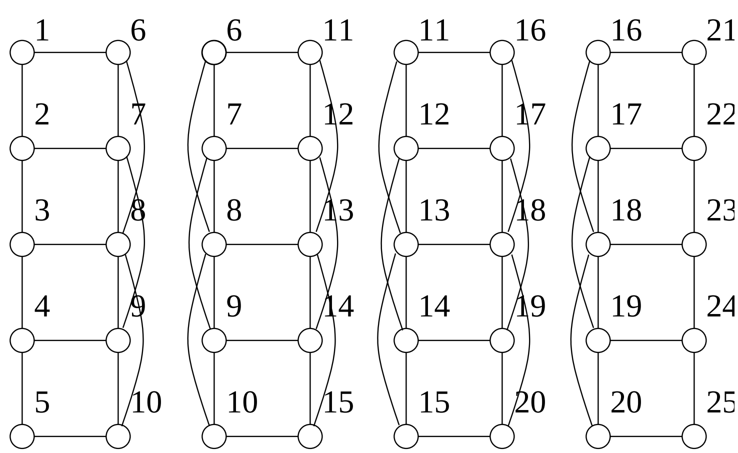
Spanning Trees



Narrow Subgraphs



Enhanced Formulations



The Dual Problem

Dual Function

$$\begin{aligned}\Psi(\lambda) &= \max_{u,v} L(u,v;\lambda)\\ &= \max_{u} \vartheta_1(u;\lambda) + \max_{v} \vartheta_2(v;\lambda)\end{aligned}$$

which is convex and piecewise linear. Evaluation of $\Psi(\lambda) \equiv$ Max-Product on subgraphs.

Weak Duality For *all* λ it holds that

$$\Psi(\lambda) = \max_{u,v} L(u,v;\lambda) \geq \max_{x} L(x,x;\lambda) = \vartheta^*$$

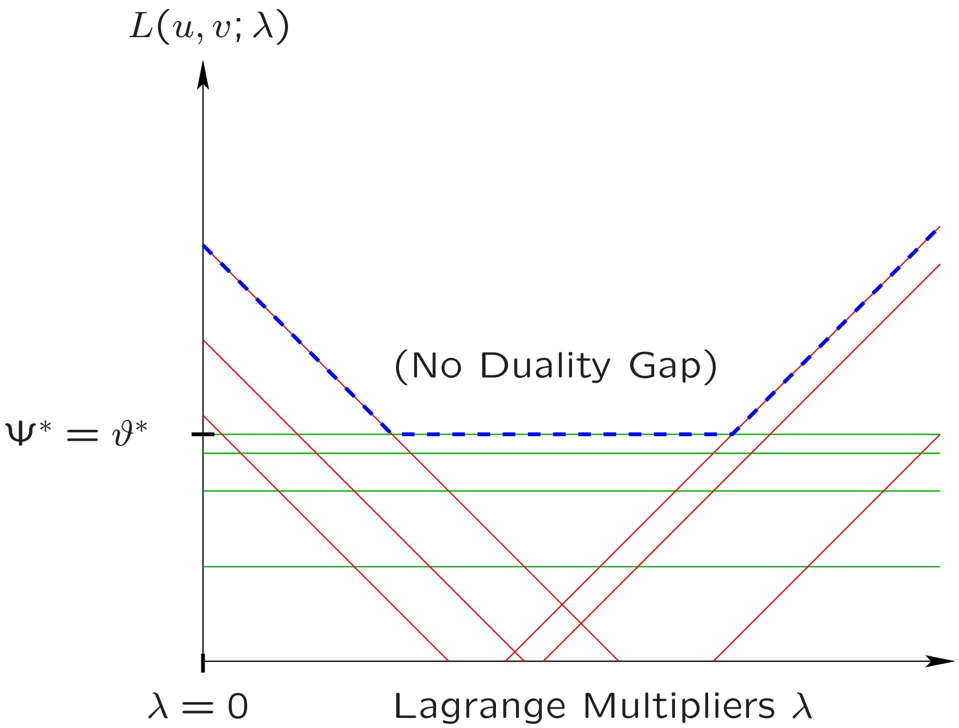
Dual Problem minimize the upper bound

$$\Psi^* = \min_{\lambda} \Psi(\lambda) \geq \vartheta^*$$

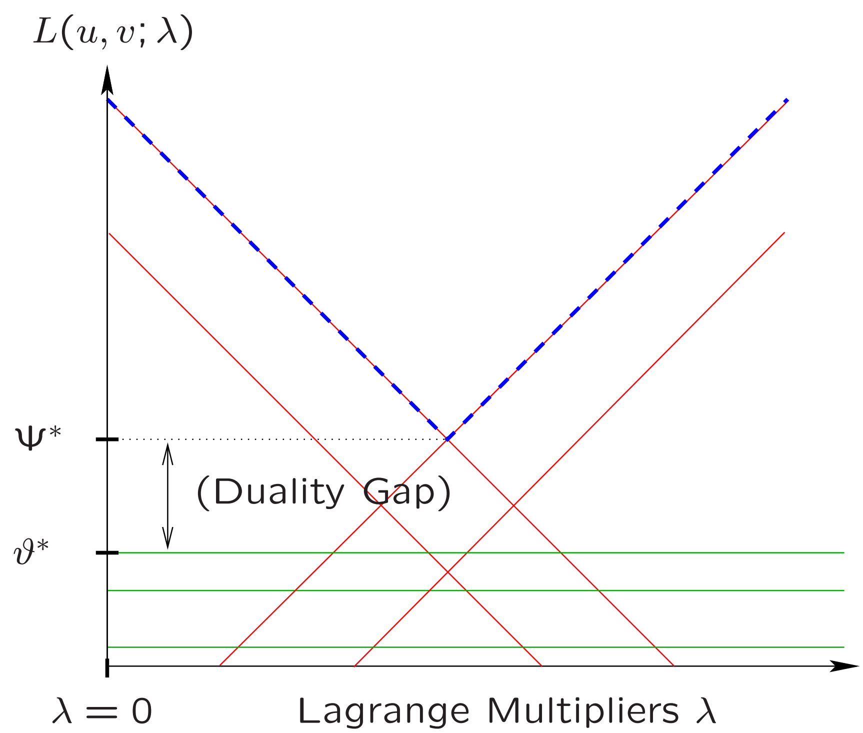
Condition for Strong Duality $\Psi^* = \vartheta^*$ if and only if $\exists(x,\lambda)$ s.t.

$$x \in \arg\max \vartheta_1(\cdot;\lambda) \cap \arg\max \vartheta_2(\cdot;\lambda)$$

Then, $\lambda \in \arg\min \Psi$ and $x \in \arg\max \vartheta$.



$L(u, v; \lambda)$
(No Duality Gap)
$\Psi^* = \vartheta^*$
$\lambda = 0$
Lagrange Multipliers λ



$L(u, v; \lambda)$
Ψ^*
(Duality Gap)
ϑ^*
$\lambda = 0$
Lagrange Multipliers λ

Low-Temperature Bounds

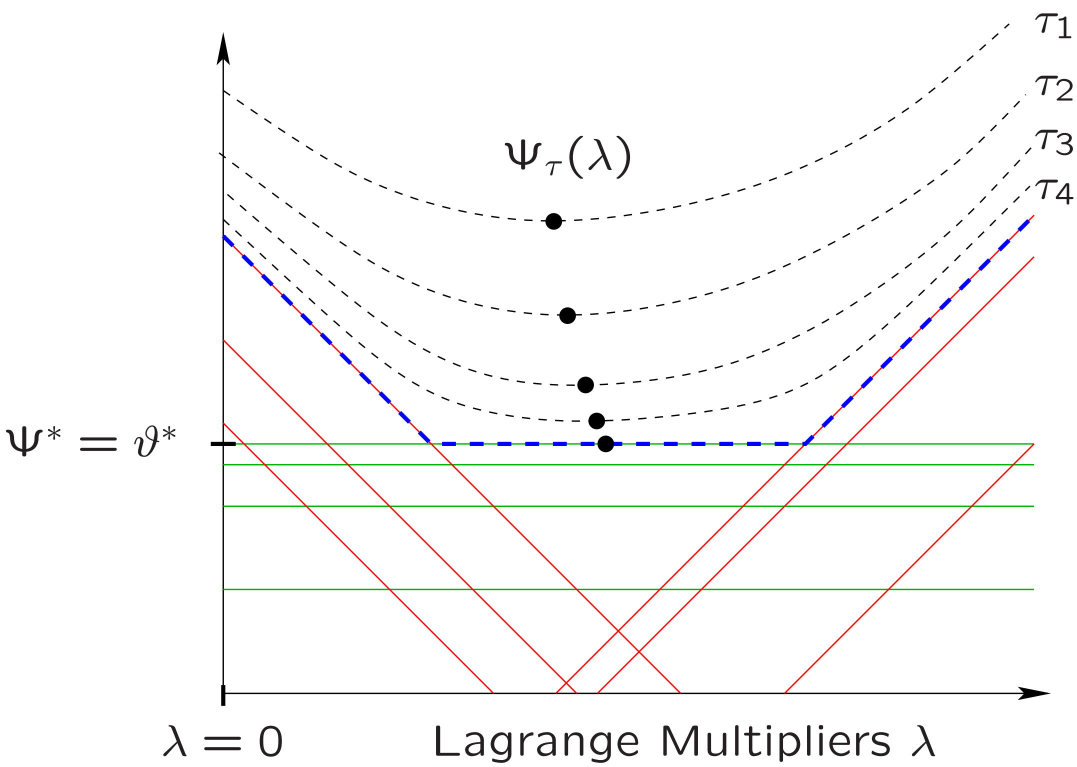
Sub-additive bound on Helmholtz free energy:

$$\Phi_\tau(\vartheta) \leq \Phi_\tau(\vartheta_1 + \lambda) + \Phi_\tau(\vartheta_2 - \lambda) \equiv \Psi_\tau(\lambda)$$

For all λ it holds that:

$$\Psi_\tau(\lambda) \geq \Psi(\lambda)$$

$$\lim_{\tau \to 0^+} \Psi_\tau(\lambda) = \Psi(\lambda)$$



Marginal Equalization

Minimize $\Psi_\tau(\lambda)$ w.r.t $\lambda_E(x_E)$:

$$\frac{\partial \Psi_\tau(\lambda)}{\partial \lambda_E(x_E)} = \mu_1(x_E;\tau,\lambda) - \mu_2(x_E;\tau,\lambda)$$

Hence, λ is optimal if and only if τ-marginals agree on all edges $E \in \mathcal{G}_1 \cap \mathcal{G}_2$.

For each $E \in \mathcal{G}$ do the following:

(1) Calculate marginal on edge E for each subgraph with $E \in \mathcal{G}_k$.

$$\widehat{\vartheta}^{(k)}_{\tau,E}(x_E) = \tau \log \sum_{x_{\backslash E}} \exp\left\{\frac{\vartheta_k(x_E, x_{\backslash E})}{\tau}\right\}$$

(2) Calculate the mean marginal potential.

$$\bar{\vartheta}_{\tau,E}(x_E) = \frac{1}{M}\sum_{k=1}^{M} \widehat{\vartheta}^{(k)}_{\tau,E}(x_E)$$

(3) Update potentials to equalize marginals.

$$\vartheta^{(k)}_E(x_E) \to \vartheta^{(k)}_E(x_E) + \left(\bar{\vartheta}_{\tau,E}(x_E) - \widehat{\vartheta}^{(k)}_{\tau,E}(x_E)\right)$$

Max-Marginal Equalization

In the zero-temperature limit, marginal equalization reduces max-marginal equalization.

For each $E \in \mathcal{G}$ do the following:

(1) Calculate max-marginal on edge E for each subgraph with $E \in \mathcal{G}_k$.

$$\widehat{\vartheta}_E^{(k)}(x_E) = \max_{x_{\setminus E}} \vartheta_k(x_E, x_{\setminus E})$$

(2) Calculate the mean max-marginal potential.

$$\bar{\vartheta}_E(x_E) = \frac{1}{M} \sum_{k=1}^{M} \widehat{\vartheta}_E^{(k)}(x_E)$$

(3) Update potentials to equalize max-marginals.

$$\vartheta_E^{(k)}(x_E) \to \vartheta_E^{(k)}(x_E) + \left(\bar{\vartheta}_E(x_E) - \widehat{\vartheta}_E^{(k)}(x_E) \right)$$

Converges to a global minimum of $\Psi(\lambda)$.

Relation to TRMP

M. Wainwright developed the *Tree-Reweighted Max-Product Algorithm* (TRMP) which tries to minimize the convexity bound of $\Phi_0(\vartheta) = \max \vartheta$:

$$\begin{array}{rl} \min & \sum_k \rho_k \Phi_0(\vartheta_k) \\ \text{s.t.} & \sum_k \rho_k \vartheta_k = \vartheta \end{array}$$

where ϑ_k are defined on spanning trees and $\{\rho_k\}$ are a set on positive weights which sum to one.

Because λ parameterizes all decompositions of ϑ between the subgraphs, LR is equivalent to minimizing the sub-additive bound:

$$\begin{array}{rl} \min & \sum_k \Phi_0(\vartheta_k) \\ \text{s.t.} & \sum_k \vartheta_k = \vartheta \end{array}$$

However, since $\rho_k \Phi_0(\theta_k) = \Phi_0(\rho_k \theta_k)$ for $\rho_k > 0$, the value of both problems are equal.

Toy Examples

10×10 grid with binary-valued variables.

Random Self-Potentials: $\vartheta_i(x_i) \sim N(0, \sigma^2)$.

Attractive Model: $\vartheta_{i,j}(x_i, x_j) = \delta(x_i = x_j)$.

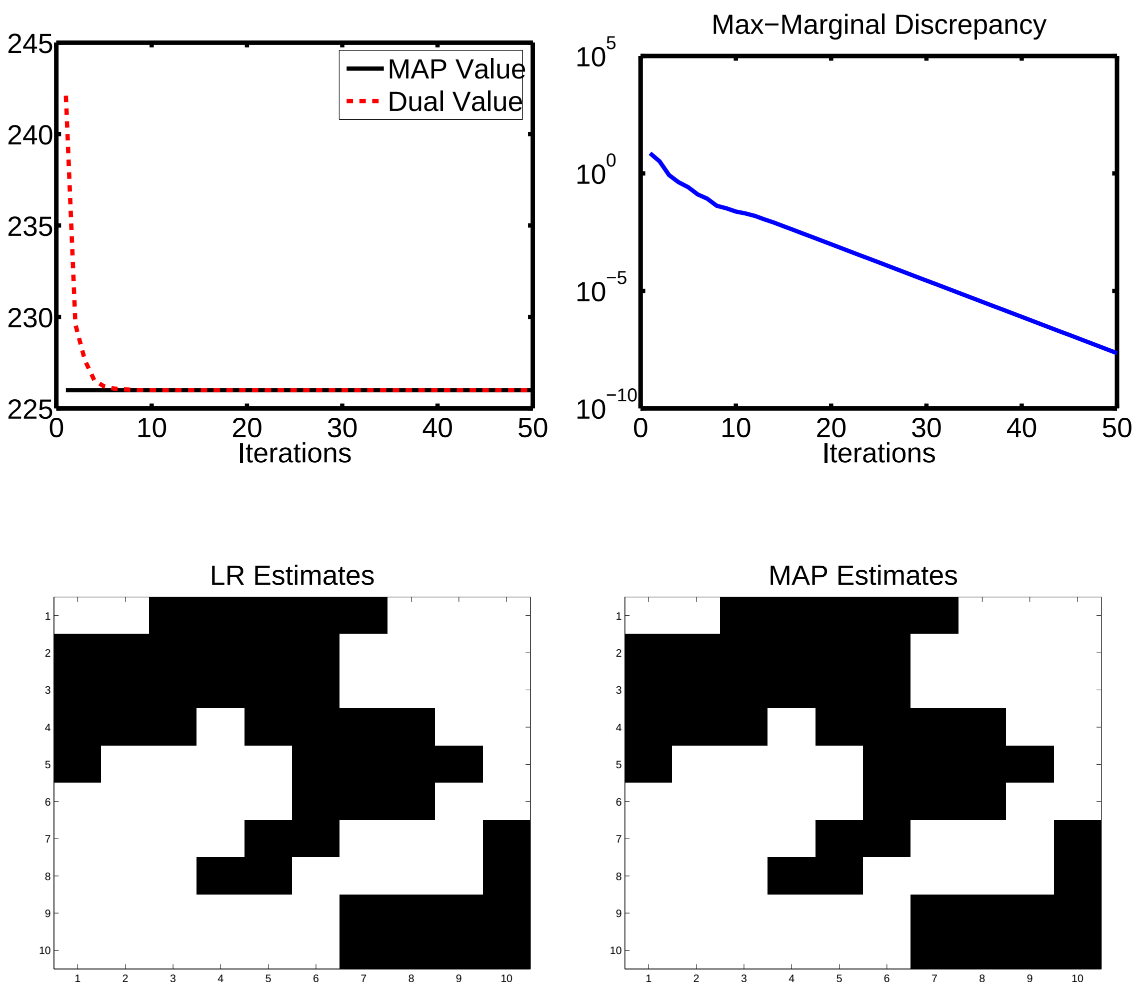
Frustrated Model: $\vartheta_{i,j}(x_i, x_j) = \pm\delta(x_i = x_j)$.

LR Based on vertical and horizontal chains.

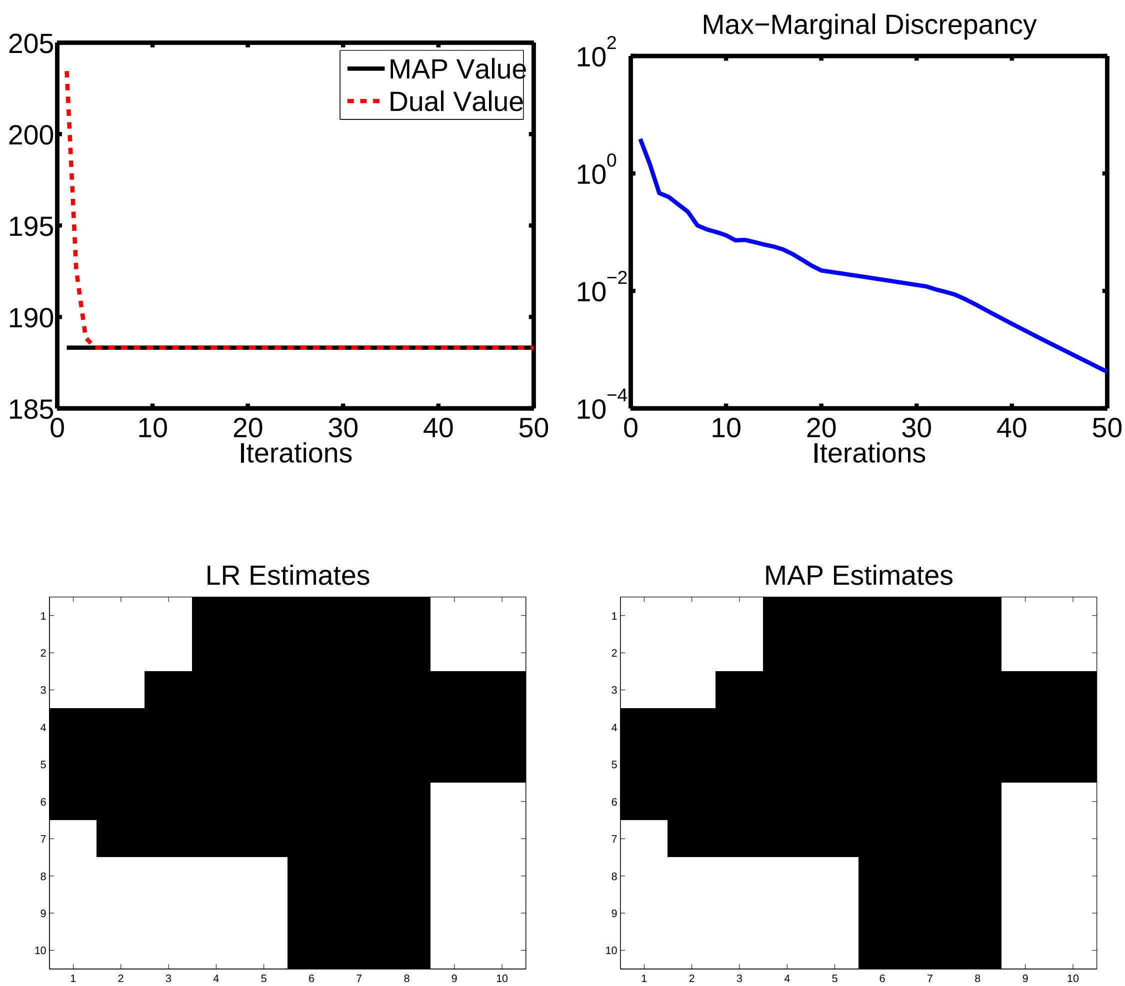
Solved by Max-Marginal Equalization.

Compare to exact MAP estimate (Junction tree algorithm).

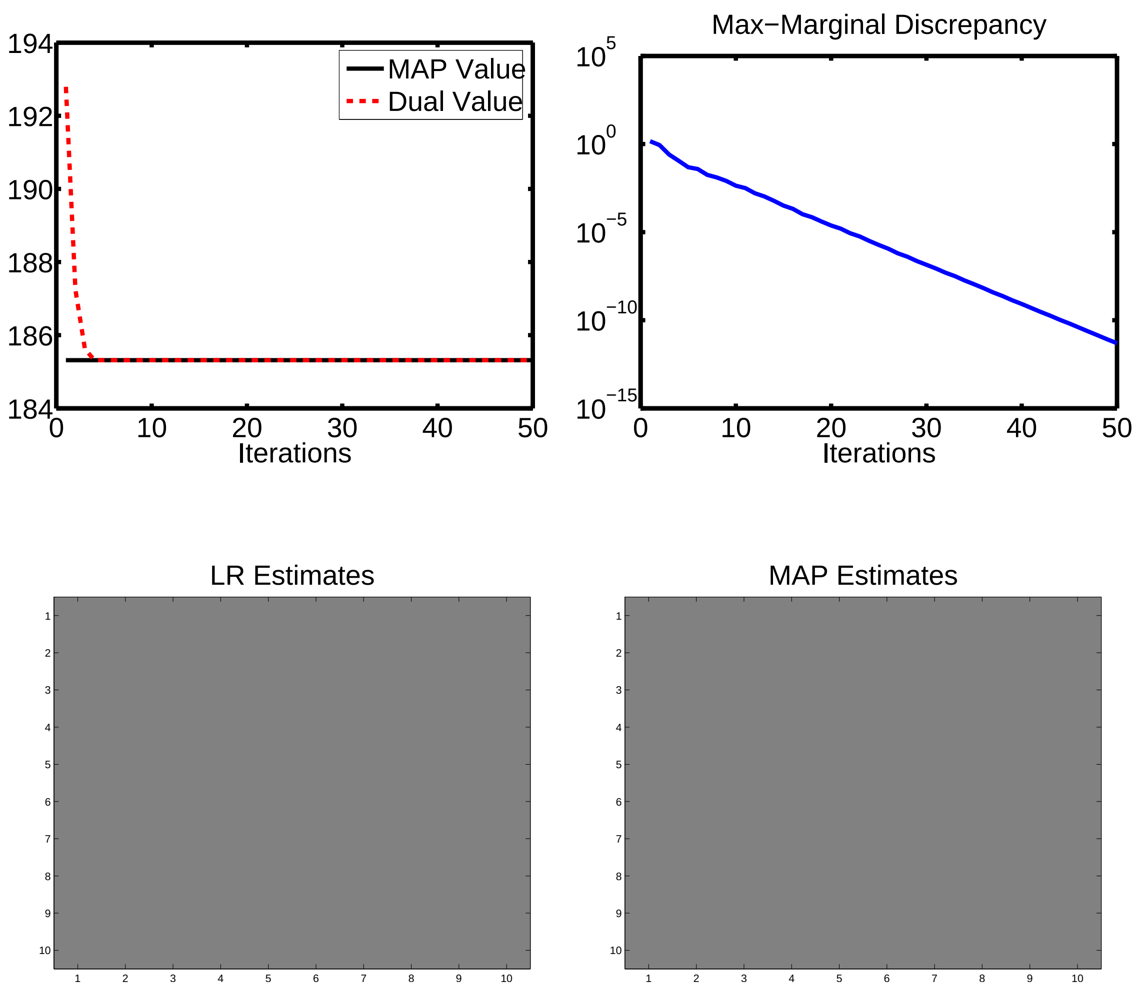
Attractive Potentials ($\sigma = 2.0$)



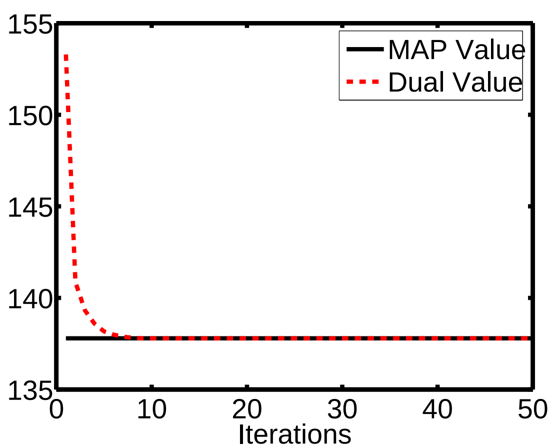
Attractive Potentials ($\sigma = 1.0$)

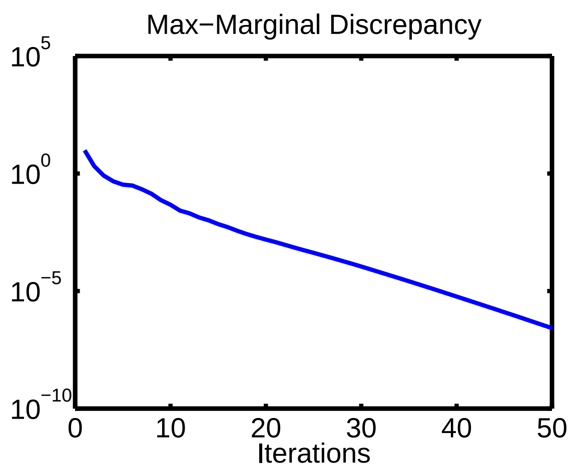


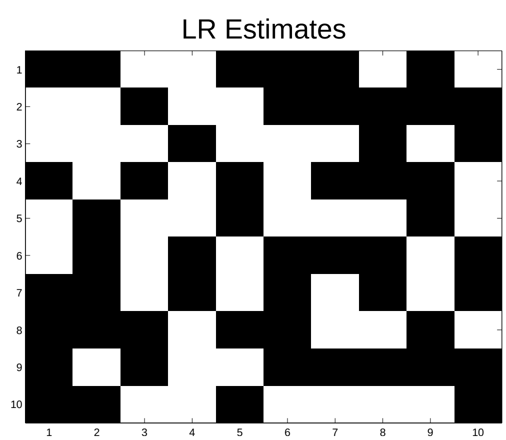
Attractive Potentials ($\sigma = 0.5$)

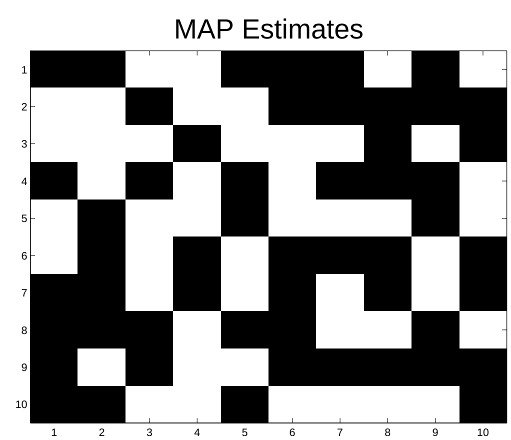


Frustrated Potentials ($\sigma = 2.0$)

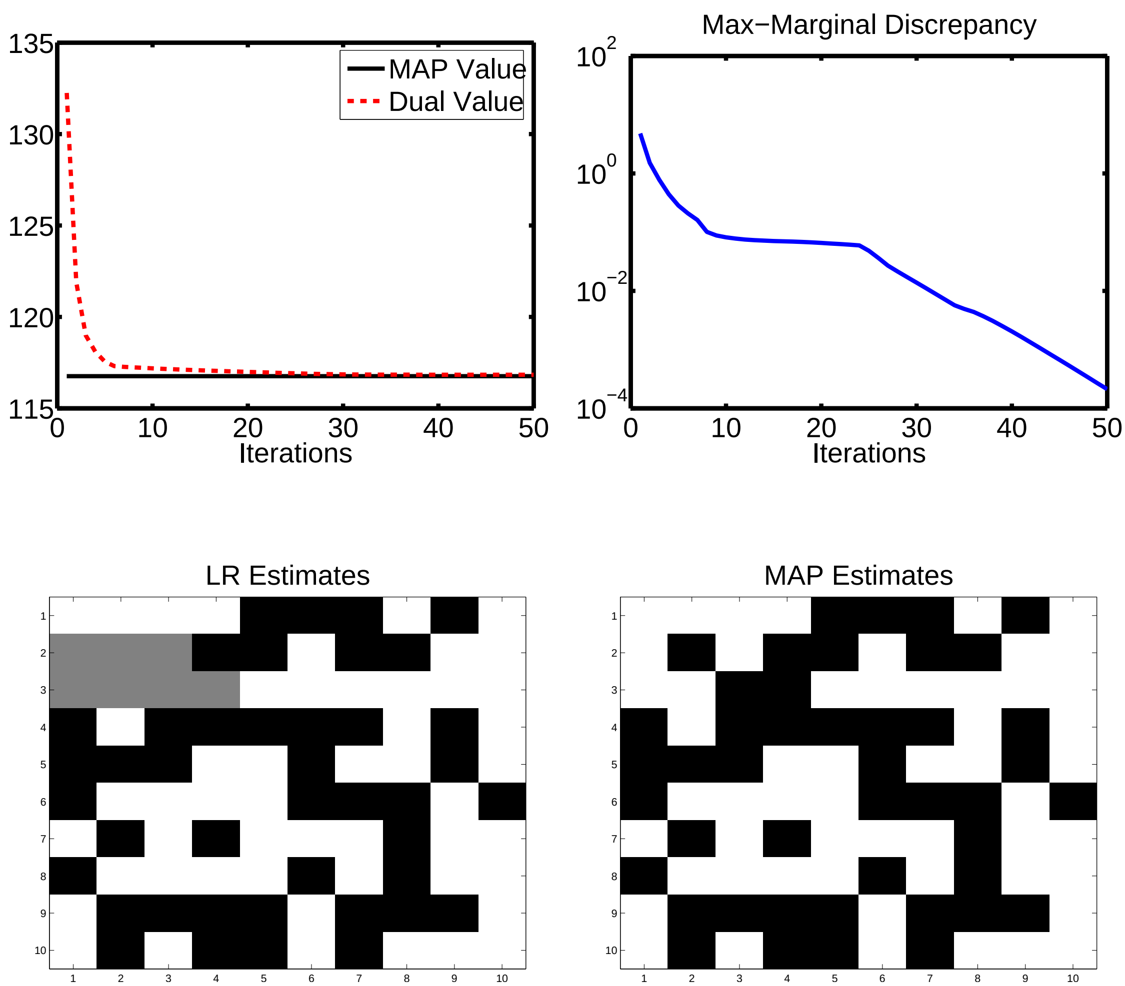




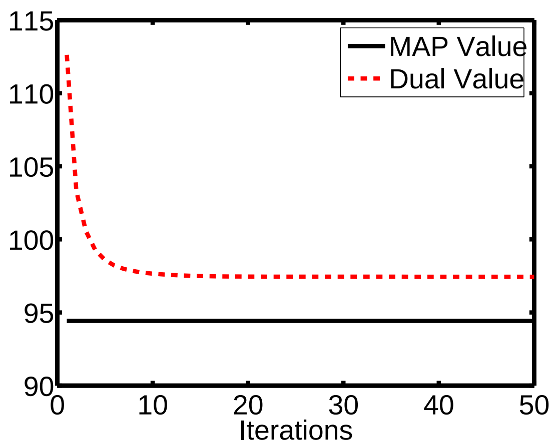


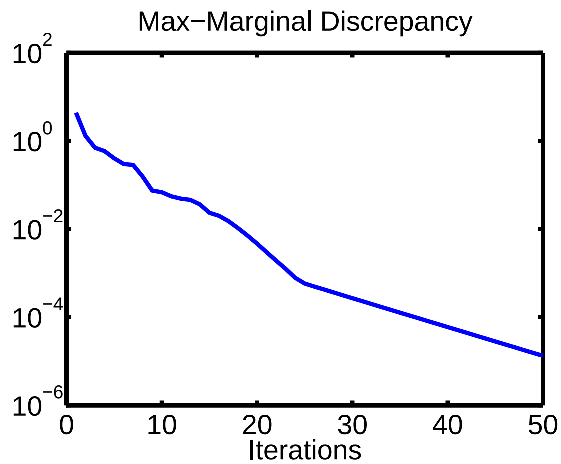


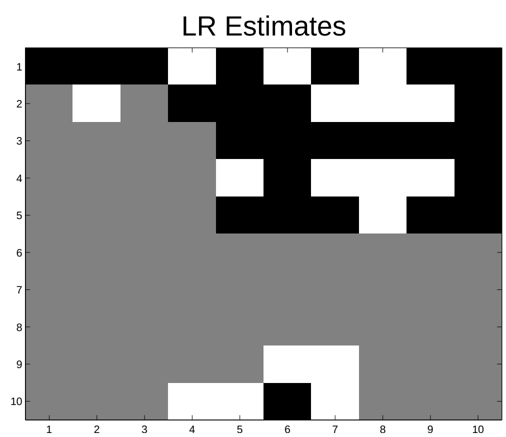
Frustrated Potentials ($\sigma = 1.5$)

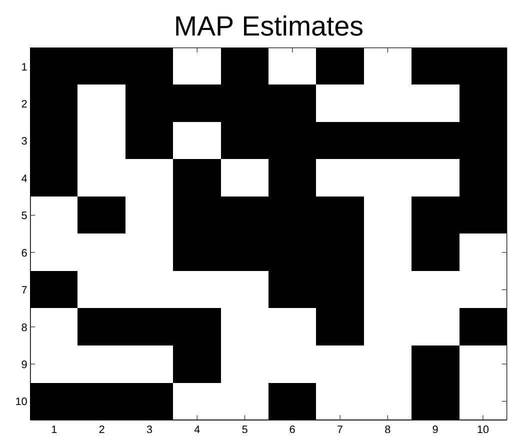


Frustrated Potentials ($\sigma = 1.0$)

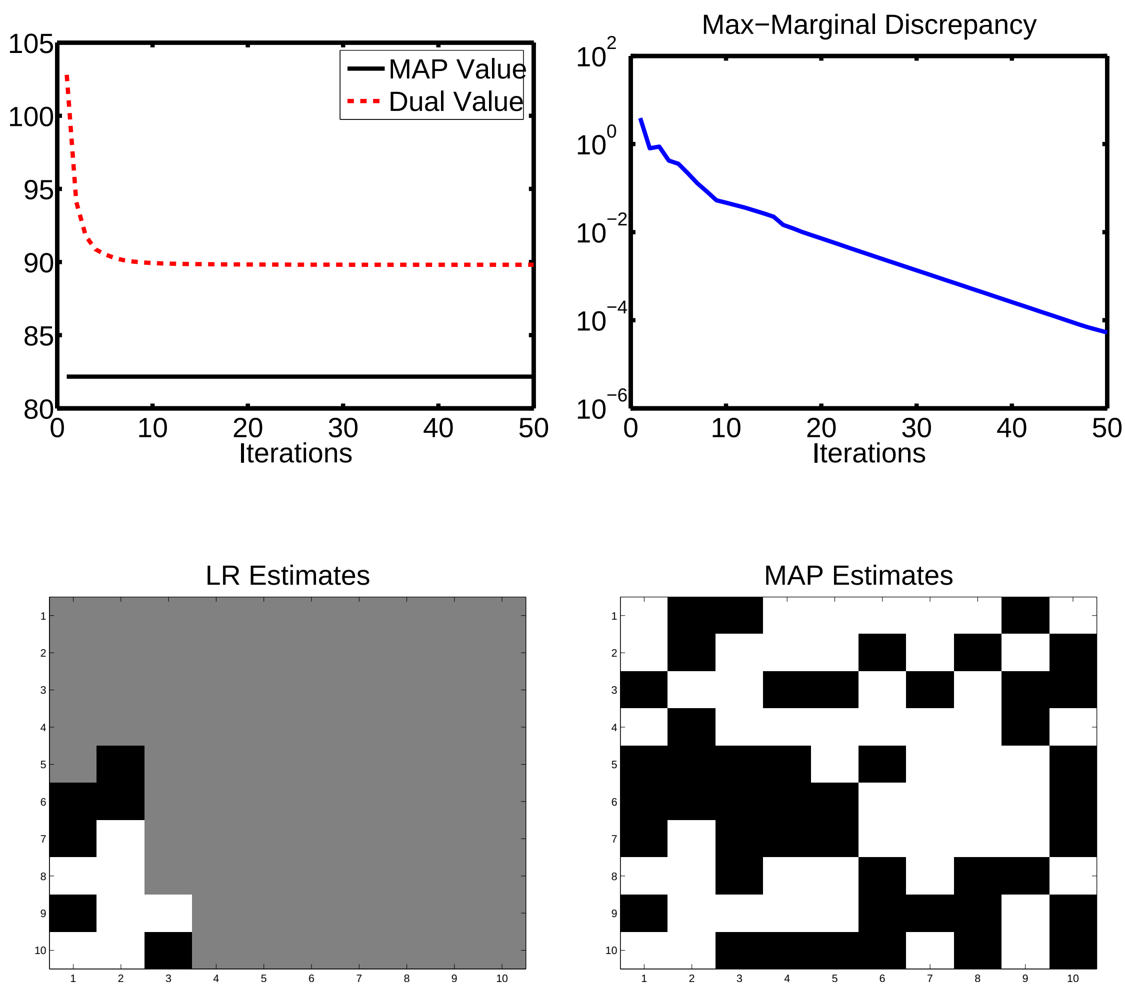








Frustrated Potentials ($\sigma = 0.7$)



Summary

- LR convex optimization problem, solution bounds value of MAP estimate.

- Convergent algorithms based on max-marginal equalization.

- Under favorable circumstances, provides (partial or complete) MAP estimate.

- Future Work:

  – Enhance LR formulation by adding edges.

  – Gaussian Models: strong duality, correctness of means, upper-bounds on variances.

  – Connections to Recursive Cavity Modeling and Expectation Propagation.