

Approximate inference in graphical models: Some progress and open questions

Martin Wainwright

Departments of Statistics and Electrical Engineering/Computer Science
UC Berkeley, USA

Alekh Agarwal, UC Berkeley

Michael Jordan, UC Berkeley

Based on some joint works with:

John Lafferty, CMU

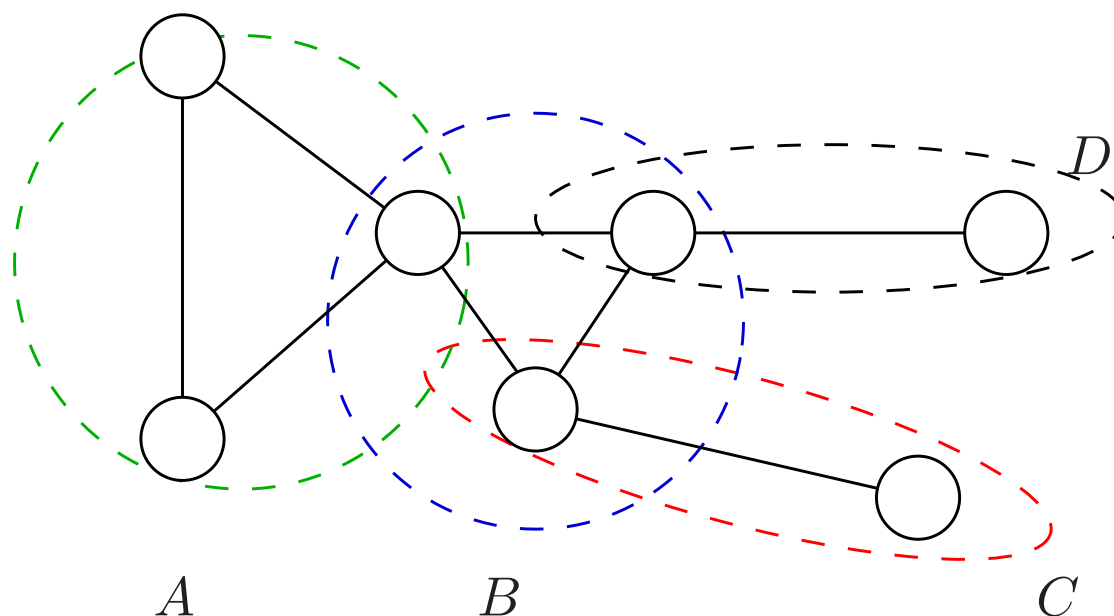
Pradeep Ravikumar, UC Berkeley

NIPS Workshops

December 2008

Graphical models

- probability plus graph theory
- Markov random field: random vector (X_1, \dots, X_N) with distribution factoring according to a graph $G = (V, E)$:



- distribution factorizes over graph cliques:

$$p(x_1, \dots, x_N; \theta) \propto \exp(\theta_A(x_A)) \exp(\theta_B(x_B)) \exp(\theta_C(x_C)) \exp(\theta_D(x_D)).$$

Some computational challenges

For a given class of graphical models:

$$p(x_1, \dots, x_N; \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathfrak{C}} \exp(\theta_C(x_C)).$$

Canonical problems:

1. Computing most probable configurations:

$$\text{MAP problem: } x^* \in \arg \max_{x \in \mathcal{X}^N} \left\{ \sum_{C \in \mathfrak{C}} \theta_C(x_C) \right\}$$

2. Summation/integration problems:

$$\text{Marginalization: } p(x_s) = \sum_{x_t, t \neq s} \frac{\prod_{C \in \mathfrak{C}} \exp(\theta_C(x_C))}{Z(\theta)}$$

$$\text{Likelihood: } Z(\theta) = \int_{\mathcal{X}^N} \prod_{C \in \mathfrak{C}} \exp(\theta_C(x_C)) dx.$$

3. Learning model parameters or graph structure

Outline: Some progress and questions

A. LP relaxations for MAP

- exact linear programs over the marginal polytope
- LP relaxations
- finite convergence via rounding schemes

B. Interactions between approximate inference and learning

- surrogate likelihoods based on convex variational relaxations
- when does it help to use the “wrong” model?

A. Any MAP problem \equiv Linear program

Key steps:

1. Cost function is additive over graph structure:

$$F^* = \max_{x \in \mathcal{X}^N} F(x_1, \dots, x_N) = \max_{x \in \mathcal{X}^N} \left[\sum_{C \in \mathfrak{C}} \theta_C(x_C) \right]$$

2. Equivalent to optimize over distributions supported on \mathcal{X}^N :

$$F^* = \max_{p \in \mathcal{P}} \sum_{x \in \mathcal{X}^N} p(x_1, \dots, x_N) \left\{ \sum_{C \in \mathfrak{C}} \theta_C(x_C) \right\} = \max_{p \in \mathcal{P}} \mathbb{E}_p \left[\sum_{C \in \mathfrak{C}} \theta_C(x_C) \right].$$

3. Linear program over globally consistent marginals:

$$F^* = \max_{\{\mu_C, C \in \mathfrak{C}\}} \sum_{C \in \mathfrak{C}} \left[\sum_{x_C} \mu_C(x_C) \theta_C(x_C) \right] = \max_{\mu_C} \sum_{C \in \mathfrak{C}} \mathbb{E}_{\mu_C} [\theta_C(x_C)].$$

Marginal polytope of a graphical model

$$\mathbb{M}(G) = \{ \mu_C, C \in \mathfrak{C} \mid \exists p \text{ such that } \mu_C \text{ is marginal over } x_C \}.$$

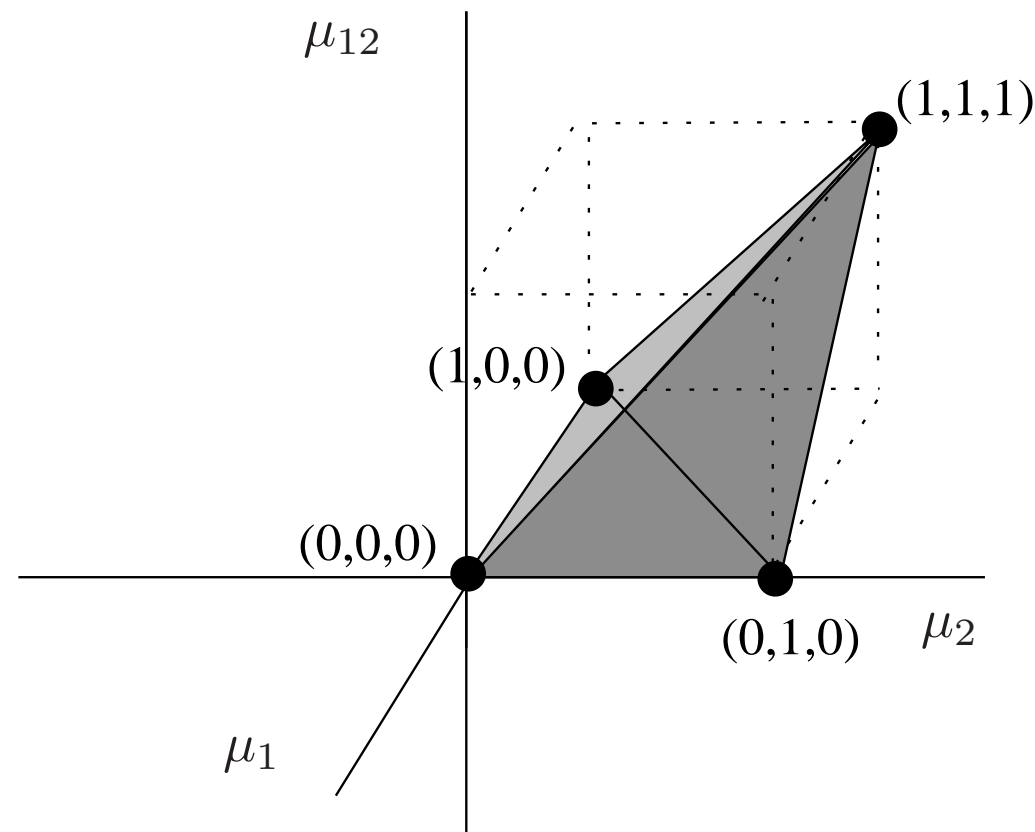
(WaiJor 2003, 2008)

Example:

- graph over two binary random variables:

$$F(x; \theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2.$$

- relevant marginal probs.:
singletons $\mu_i = p(x_i = 1)$ and
joint $\mu_{12} = p(x_1 = 1, x_2 = 1)$.



Equivalently, convex hull $\{ (x_1, x_2, x_1 x_2) \mid (x_1, x_2) \in \{0, 1\}^2 \}$.

First-order (tree-based) LP relaxation

- applies to any graphical model in factor graph form:

$$p(x_1, \dots, x_N; \theta) \propto \prod_{s \in V} \exp(\theta_s(x_s)) \prod_{C \in \mathfrak{C}} \exp(\theta_C(x_C))$$

- maintain pseudomarginals at each site and over each clique:

$$\mu_s(x_s) = \begin{bmatrix} \mu_s(0) & \mu_s(1) & \cdots & \mu_s(m-1) \end{bmatrix}$$

$$\mu_{st}(x_s, x_t) = \begin{bmatrix} \mu_{st}(0,0) & \mu_{st}(0,1) & \cdots & \mu_{st}(0,m-1) \\ \mu_{st}(1,0) & \mu_{st}(1,1) & \cdots & \mu_{st}(1,m-1) \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{st}(m-1,0) & \mu_{st}(m-1,1) & \cdots & \mu_{st}(m-1,m-1) \end{bmatrix}$$

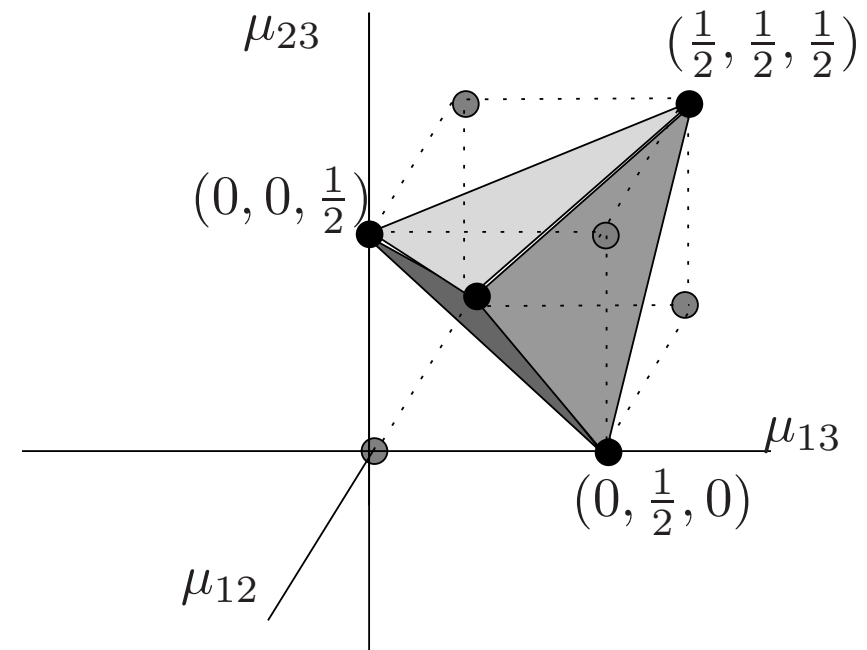
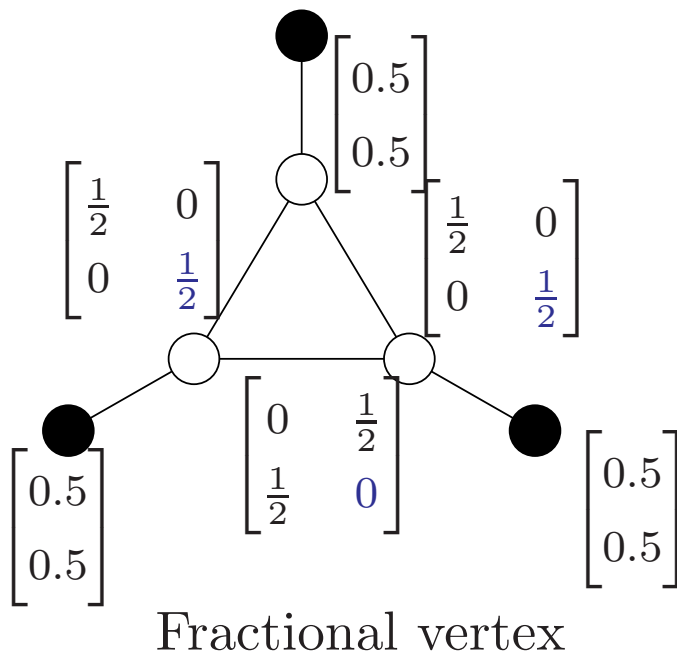
- enforce non-negativity, *normalization constraints* $\sum_{x_s} \mu_s(x_s) = 1$, and *marginalization constraints*

$$\mu_s(x_s) = \sum_{x_t, t \in \mathfrak{C}, t \neq s} \mu_C(x_C)$$

Fractional versus integral vertices

Two possible outcomes to solving first-order LP relaxation:

- integral vertex obtained: LP relaxation recovers MAP configuration
- fractional vertex obtained: LP relaxation is loose



Projection shows $M(G) \subsetneq [0, 1]^3$

Past and on-going work: A partial list....

- first-order LP and (reweighted) max-product: Wainwright et al., 2002
- various methods for tackling LP:
 - TRW max-product can be trapped at non-dual-optima: Kolmogorov, 2005
 - TRW never trapped for binary QPs: Kolmogorov & Wainwright, 2006
 - dual-ascent methods (can be trapped): Globerson & Jaakkola, 2007
 - limits of convex free energies and LP: Weiss et al., 2006, Johnson et al., 2007
 - subgradient methods (convergent LP-solvers): Feldman et al, 2003; Komodakis et al., 2007
 - efficient variants of interior-point methods: Vontobel, 2008
 - proximal point algorithms: Ravikumar et al., 2008
- some performance guarantees
 - worst-case guarantees: Kleinberg & Tardos, 2001; Chekuri et al., 2005
 - average-case analysis via dual-witness methods: Daskalakis et al., 2008
- higher-order LP relaxations
 - efficient algorithms: Sontag et al., 2008; Hazan & Shashua, 2008
 - unified framework with moment matrices: Wainwright & Jordan, 2005
- ordinary max-product and matching: Bayati et al, 2005; Huang & Jebara, 2007

Rounding and its uses

- rounding methods convert fractional LP solution to an integer configuration (Raghavan & Thompson, 1987; Vazirani, 2003)
- classical rounding theory: rounded solution x^{rou} has log. probability within some factor $\alpha(N; \theta) \in (0, 1]$:

$$F(x^*; \theta) \geq F(x^{\text{rou}}; \theta) \geq \alpha(N; \theta) F(x^*; \theta).$$

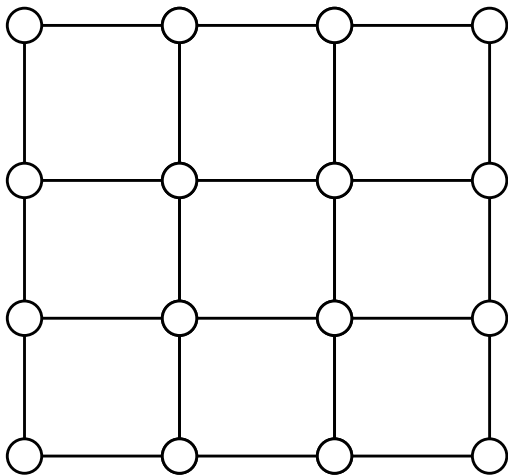
- for tree-based first-order relaxation, such “classical” results exist in special cases (Kleinberg & Tardos, 2001, Chekuri et al., 2005)
- an alternative use of rounding
 - apply a rounding to an intermediate solution μ^n of an LP-solver
 - if successful, yields finite termination for LP-solving

Forest-inducing subsets

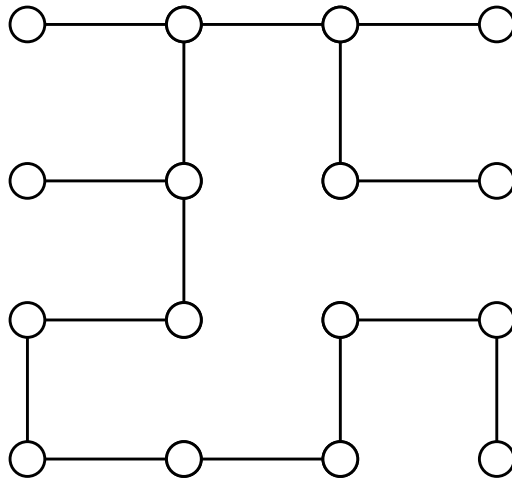
An edge set $E' \subseteq E$ such that $G' = (V, E')$ is a union of trees $\{T_1, T_2, \dots, T_M\}$ with disjoint vertex sets:

$$E' = E(T_1) \cup E(T_2) \cup \dots \cup E(T_M)$$

$$V = V_1 \cup V_2 \cup \dots \cup V_M$$

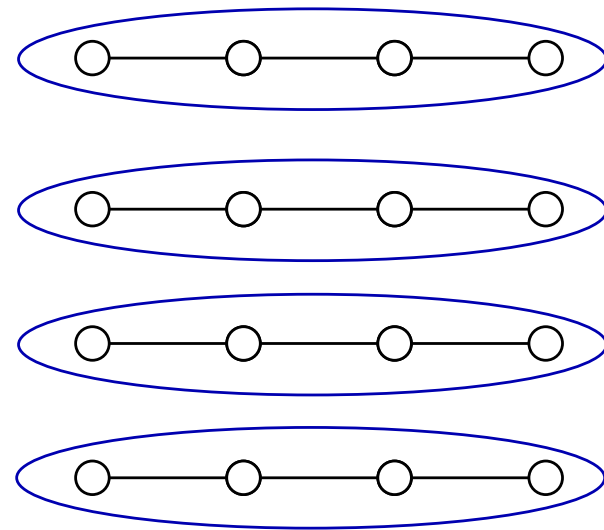


(a) Graph G



(b) $M = 1$

T_1



(c) $M = 4$

Forest-based randomized rounding

- produces a random integral solution $X = (X_1, \dots, X_N) \in \mathcal{X}^N$ with desirable properties
- applied to any interior-based LP-solving method that produces a sequence $\{\mu^n\} \subseteq \mathbb{L}(G)$
- based on a given forest-inducing collection of subtrees $\{T_1, T_2, \dots, T_M\}$:

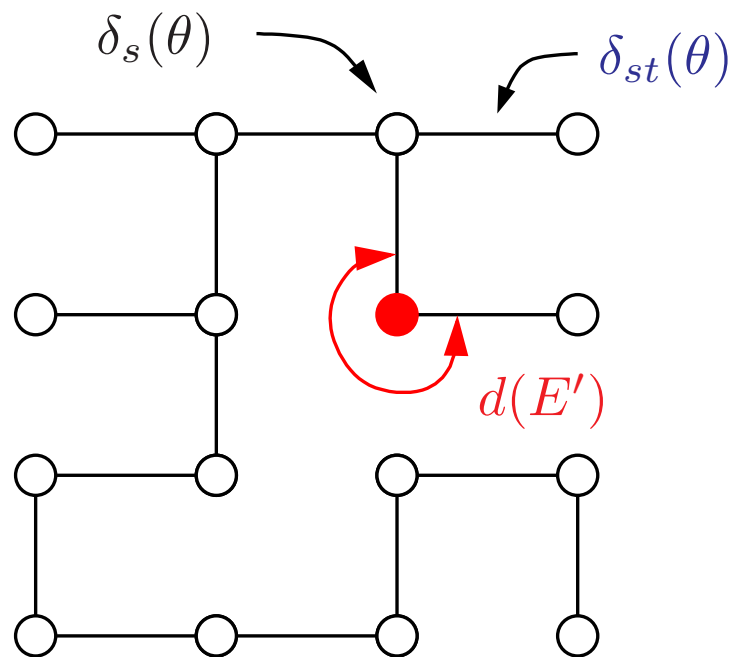
1. For subtree indices $i = 1, \dots, M$, sample

$$X_{V_i} \sim p(x_{V_i}; \mu^n(T_i)) = \prod_{s \in V_i} \mu^n(x_s) \prod_{(s,t) \in E_i} \frac{\mu^n(x_s, x_t)}{\mu^n(x_s) \mu(x_t)}.$$

2. Form the global configuration $X \in \mathcal{X}^N$ by concatenating all the local random samples:

$$X := (X_{V_1}, \dots, X_{V_M}).$$

Theoretical guarantees on forest-rounding



$$\delta_{st}(\theta) := \max_{x_s, x_t} [\theta_{st}(x_s, x_t)] - \min_{x_s, x_t} [\theta_{st}(x_s, x_t)]$$

$$\delta_G(\theta) := \max \left\{ \max_{(s,t) \in E} \delta_{st}(\theta), \max_{s \in V} \delta_s(\theta) \right\}.$$

$$d(E') := \max_{s \in V} |\{t \in V \mid (s, t) \notin E'\}|.$$

Theorem: Suppose that:

(RavAgaWai08)

- MAP solution is unique ($\arg \max_x F(x; \theta) = \{x^*\}$),
- the first-order LP relaxation is tight, achieved at a vertex μ^* of $\mathbb{L}(G)$.

Then forest-rounding succeeds with probability at least $1 - \epsilon$ once

$$\|\mu^n - \mu^*\|_1 \leq \frac{\min_{x \neq x^*} [F(x^*; \theta) - F(x; \theta)]}{\delta_G(\theta)} \frac{\epsilon}{1 + d(E')}.$$

Relevant factors

- **problem difficulty:** how well-separated is the MAP optimum relative to clique parameters θ_s, θ_{st} ?

$$\Delta(\theta; G) := \frac{\min_{x \neq x^*} [F(x^*; \theta) - F(x; \theta)]}{\delta_G(\theta)}$$

- **choice of algorithm:** how fast does $\|\mu^n - \mu^*\|_1$ decay?

Corollary: Given an algorithm that is linearly convergent in ℓ_2 -norm — that is

$$\|\mu^n - \mu^*\|_2 \leq \gamma^n \|\mu^0 - \mu^*\|_2.$$

Then forest-rounding succeeds with probability greater than $1 - \epsilon$ for all iterations $n \geq \lceil n^* \rceil$, where

$$n^* := \frac{\log \sqrt{Nm + N^2 m^2} + \log \|\mu^0 - \mu^*\|_2 + \log \left(\frac{1+d(E')}{\Delta(\theta; G)} \right) + \log(1/\epsilon)}{\log(1/\gamma)}$$

Proximal algorithms

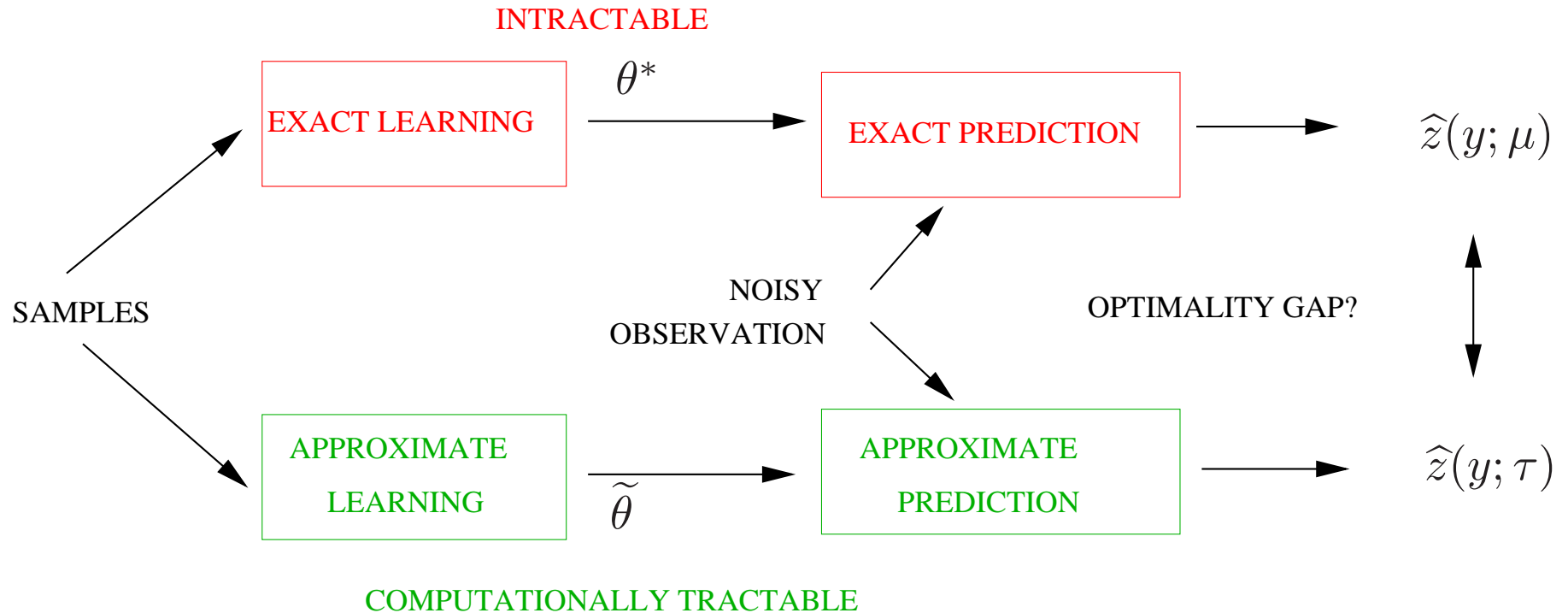
- standard class of methods in optimization theory (TseBer93; IueTeb95)
- can be fruitfully applied to solving first-order LP relaxation, as can randomized rounding for finite termination (PraAgaWai08)
- instead of solving LP directly, generate sequence of iterates:

$$\mu^{n+1} = \arg \min_{\mu \in \mathbb{L}(G)} \{ \mu^T \theta + D_f(\mu \parallel \mu^n) \}.$$

- examples of proximal function

$$\begin{array}{ll} \text{Quadratic} & D_f(\mu \parallel \mu^n) = \|\mu - \mu^n\|_2^2 \\ \text{Kullback-Leibler} & D_f(\mu \parallel \mu^n) = \sum_i \mu_i \log \frac{\mu_i}{\mu_i^n} \end{array}$$

B. Approximate learning and prediction



Key question:

- in computation-limited setting, errors arise in both learning and prediction steps
- is it possible to systematically learn the “wrong” model so that errors largely cancel?

Intractability of likelihood calculations

- graphical model as an exponential family:

$$p(x_1, x_2, \dots, x_N; \theta) = \left\{ \sum_{C \in \mathfrak{C}} \theta_C(x_C) - A(\theta) \right\}$$

where $A(\theta) = \log \sum_x \exp \left(\sum_C \theta_C(x_C) \right)$.

- given i.i.d. $X^{(1)}, \dots, X^{(n)}$ samples, might consider methods based on likelihood

$$\ell_A(\theta; X^{(i)}) = \frac{1}{n} \sum_{i=1}^n \log \mathbb{P}(X^{(i)}; \theta) = \sum_{C \in \mathfrak{C}} \langle \theta_C, \hat{\mu}_C \rangle - A(\theta)$$

where the *empirical marginals* take the form:

$$\hat{\mu}_C(\bar{x}_C) = \frac{\text{card} \{i = 1, \dots, n \mid X_C^{(i)} = \bar{x}_C\}}{n}.$$

Surrogate likelihoods via variational methods

Obstacle: Computing A is difficult, but has the variational representation

$$\log \int_{\mathcal{X}^N} \exp \left(\sum_{C \in \mathfrak{c}} \theta_C(x_C) \right) dx = A(\theta) = \sup_{\mu \in \mathcal{M}(G)} \{ \mu^T \theta - A^*(\mu) \}$$

where $\mathcal{M}(G)$: globally realizable mean parameters
 $A^*(\mu)$: negative entropy functional

Strategy:

(a) Construct a “surrogate” B for A via:

$$B(\theta) := \max_{\tau \in \mathcal{L}(G)} \{ \mu^T \theta - B^*(\tau) \}$$

where $\mathcal{L}(G)$: convex outer bound on $\mathcal{M}(G)$
 $B^*(\mu)$: convex approximation to A^*

(b) Compute surrogate maximum likelihood estimate (MLE):

$$\tilde{\theta} = \arg \max_{\theta} \{ \mu^T \theta - B(\theta) \}.$$

Some examples

1. Convexified Bethe/Kikuchi approximations (WaiJaaWil03, HesWie3, Wei07)

$$\begin{aligned}\mathcal{L}(G) &= \left\{ \tau_s, \tau_{st} \mid \sum_{x_s} \tau_s(x_s) = 1, \quad \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s) \right\} \\ B^*(\tau) &= - \sum_{s \in V} H_s(\tau_s) + \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}).\end{aligned}$$

2. Log-determinant approximation (WaiJor05)

$$\begin{aligned}\mathcal{L}(G) &= \left\{ \tau_s, \tau_{st} \mid \text{associated covariance matrix } M_1[\tau] \succeq 0 \right\}. \\ B^*(\tau) &= - \log \det M_1[\tau]\end{aligned}$$

3. Convexified forms of expectation propagation
4. Many others waiting to be developed.....

Asymptotics of surrogate MLEs

- $\tilde{\theta}$ obtained by maximizing surrogate likelihood

$$\tilde{\theta} := \max_{\theta} \left\{ \underbrace{\theta^T \hat{\mu}^n - B(\theta)}_{\ell_B(\theta; \hat{\mu}^n)} \right\}$$

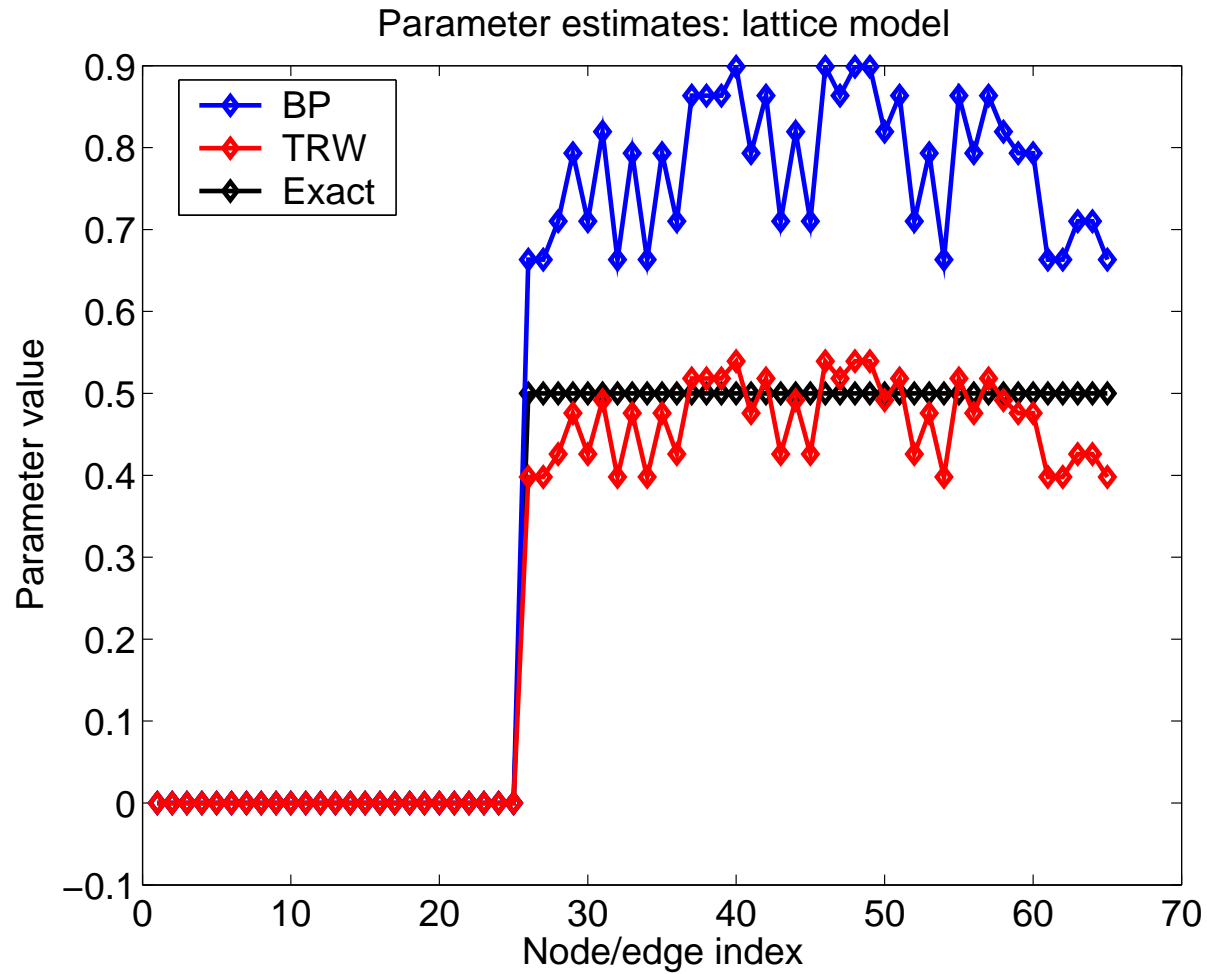
- population criterion based on minimizing criterion with infinite sample size (i.e., $\hat{\mu}^\infty = \mu^*$)
- define limit estimator: $\tilde{\theta} = \arg \max_{\theta} \ell_B(\theta; \mu^*)$.
- as the number of samples $n \rightarrow +\infty$, the estimator converges $\hat{\theta}^n \xrightarrow{a.s.} \tilde{\theta}$, and **asymptotic normality** holds:

$$\sqrt{n}[\hat{\theta}^n - \tilde{\theta}] \xrightarrow{d} \mathcal{N}(0, [\nabla^2 B(\tilde{\theta})]^{-1} \nabla^2 A(\theta^*) [\nabla^2 B(\tilde{\theta})]^{-1})$$

- however, estimator is **asymptotically inconsistent**:

$$\hat{\theta}^n \rightarrow \tilde{\theta} \neq \theta^*$$

Asymptotic inconsistency

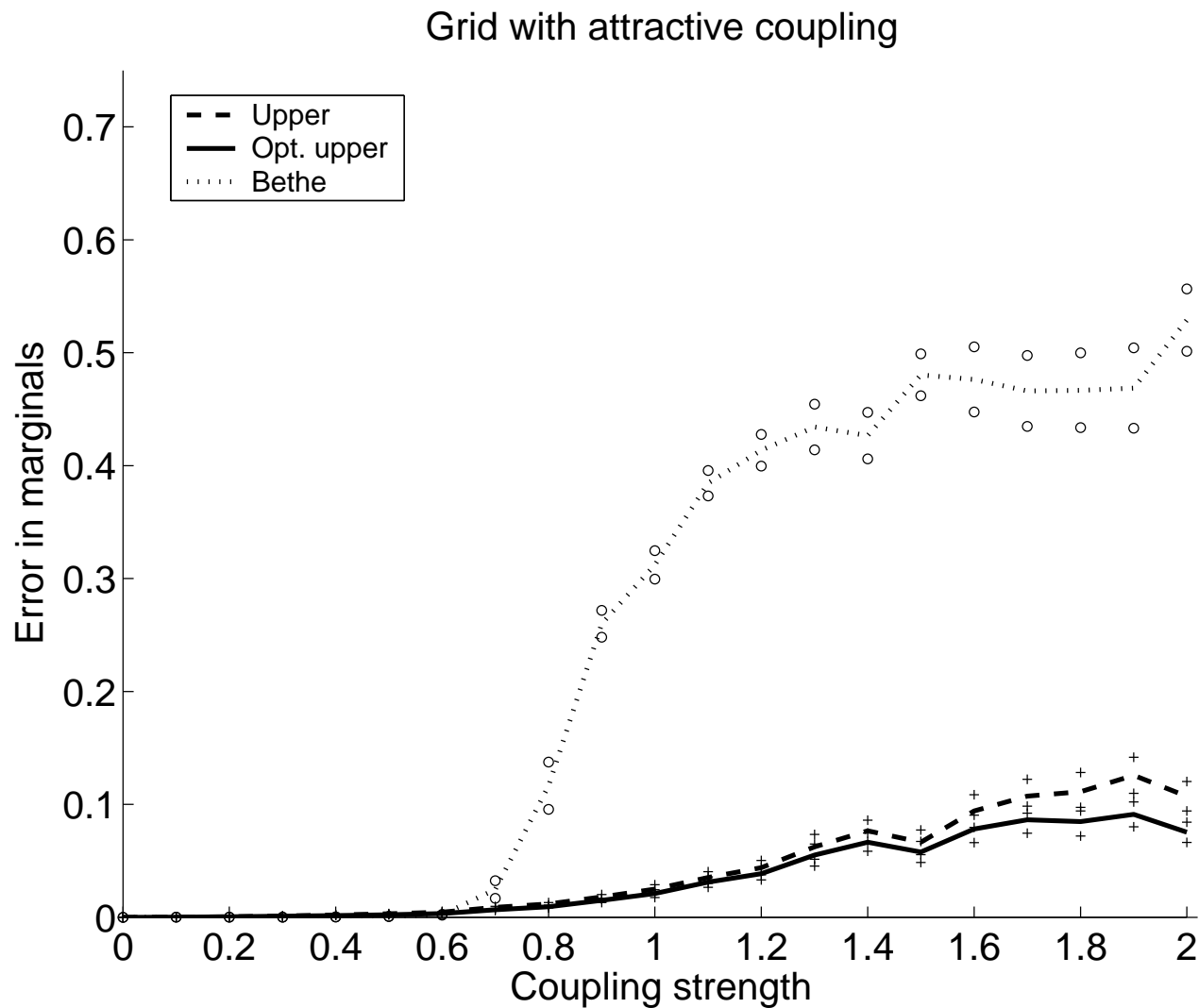


Even with infinite data, surrogate likelihood returns *wrong model!*

Why are convex surrogates still useful?

- conjugate dual pair (A, A^*) generate a bijection $\theta \leftrightarrow \mu$ between:
 - models p_θ indexed by parameters θ , and
 - marginal/mean parameters μ associated with model
- under suitable technical conditions, surrogate dual pair (B, B^*) also generates a bijection $\theta \leftrightarrow \tau$ between:
 - models p_θ indexed by parameters θ , and
 - approximate marginal/mean parameters τ
- good properties of (B, B^*) induce corresponding good properties of inference and learning algorithms

Instability of non-convex variational methods

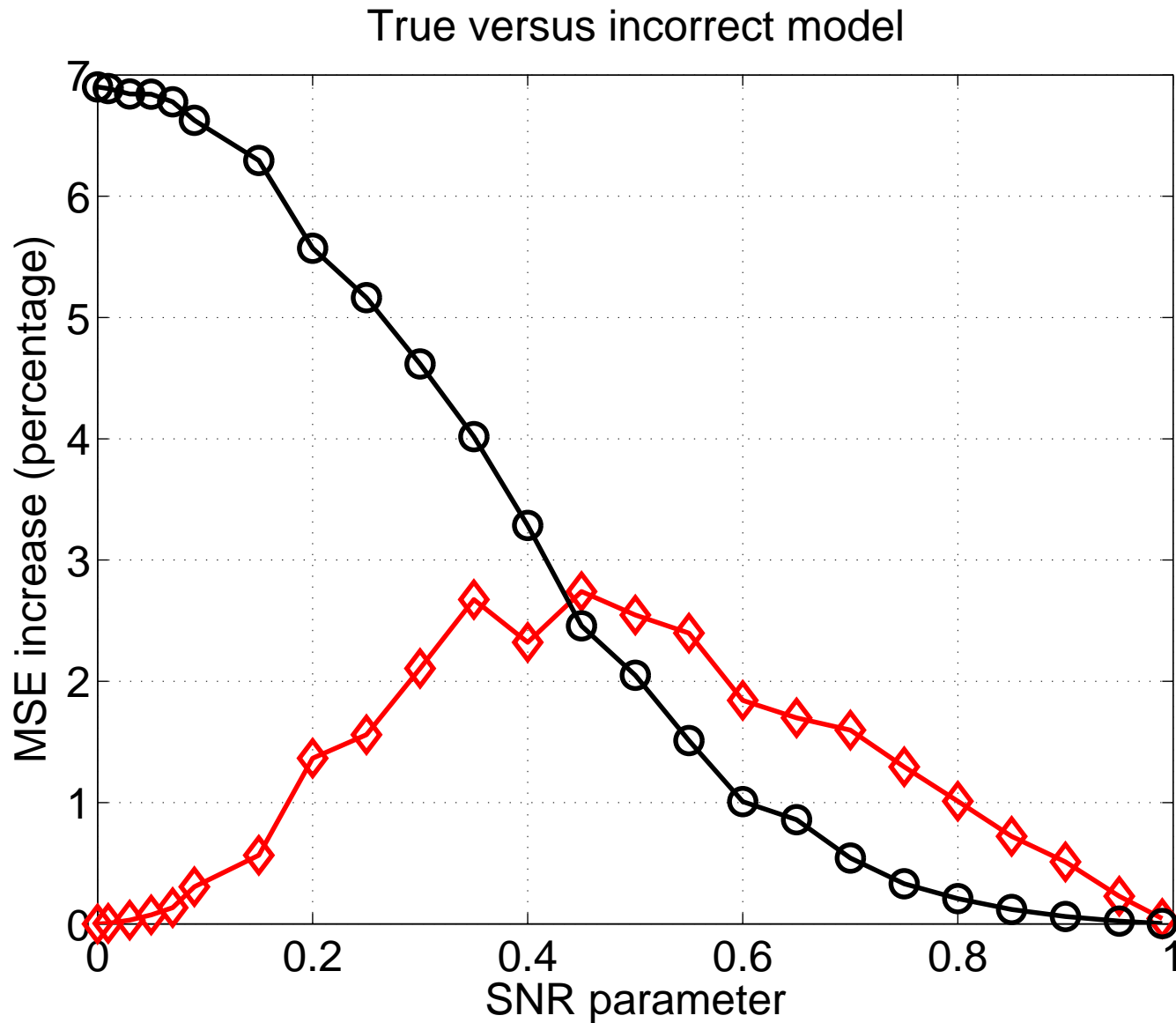


- algorithms applied to perturbed models or with numerical errors.
- stability is essential in such settings

Interactions in approx. learning/prediction

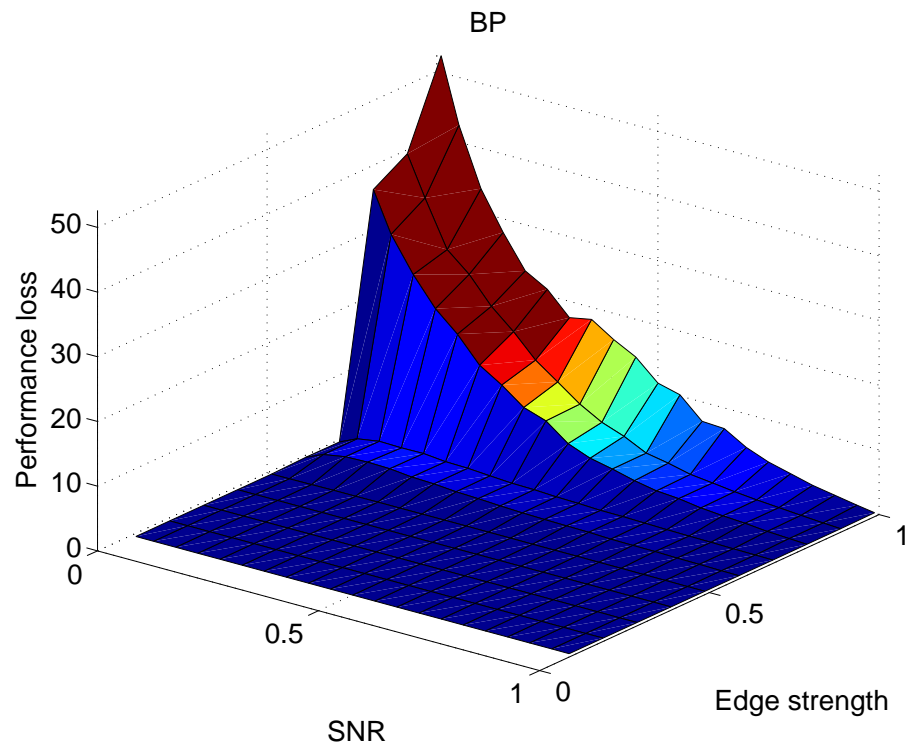
- standard prediction/classification problem
 - initial set of training data $\{(X^{(1)}, Y^{(n)}), \dots, (X^{(n)}, Y^{(n)})\}$
 - subsequent test data to be predicted
- such problems often involve graphical model structure:
 - image denoising/deblurring in computer vision
 - sentence classification in language processing
- three possible methods to consider
 - **Oracle/Oracle:** True model given by oracle, and prediction performed optimally via oracle (gold-standard).
 - **Oracle/Surrogate:** True model given by oracle, and prediction performed based on (B, B^*) .
 - **Surrogate/Surrogate:** Incorrect model estimated by surrogate ℓ_B , and prediction performed based on (B, B^*) .

Predicting using “wrong” model superior

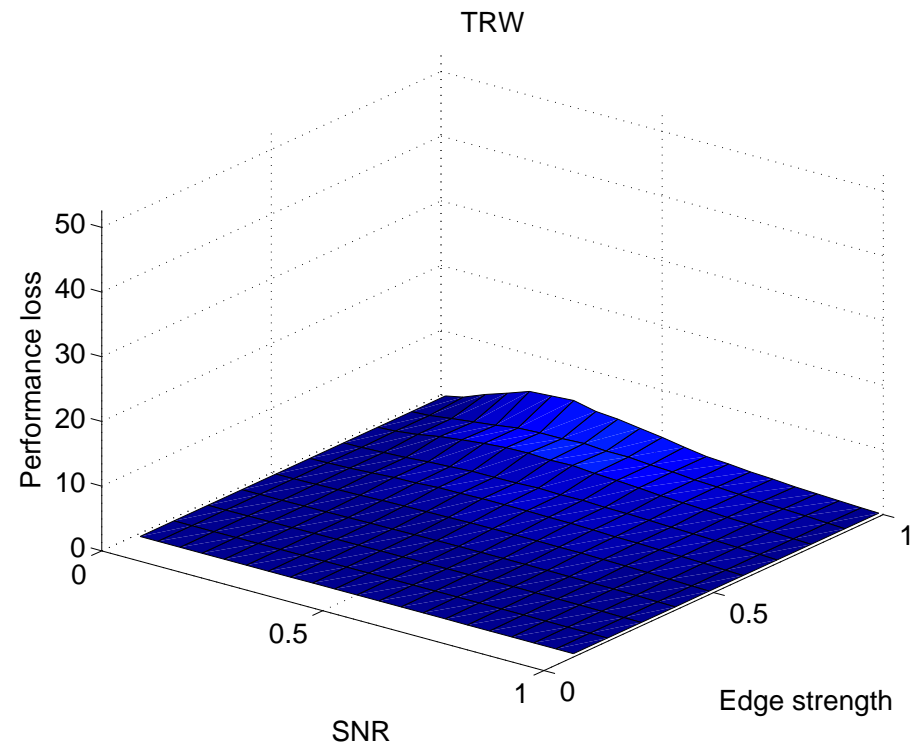


(Wainwright, 2006; JMLR)

(Non)-convexity in joint learning/prediction



Ordinary BP



Reweighted BP (convex)

Summary

- some facets of approximate inference in graphical models:
 - LP relaxations: graph-based algorithms and rounding schemes
 - interactions between approximate learning and prediction
 - high-dimensional scaling in graphical model selection
- many open questions
 - hierarchies of LP relaxations (based on Kikuchi clustering):
 - * fast and/or adaptive algorithms (Sontag et al., 2008)
 - * guarantees for specific problem classes, and average-case analysis?
 - other hierarchies of moment-based relaxations (SOCP, SDP):
 - * fast algorithms?
 - * analysis and performance guarantees? (Kumar et al., 2007)
 - convex surrogates and approximate learning:
 - * other surrogates and their properties?
 - * other settings in which using “wrong” model is beneficial?
 - * extensions to continuous variables?

Forthcoming survey paper

Graphical models, exponential families and variational methods, M. J. Wainwright and Michael I. Jordan. *Foundations and Trends in Statistical Machine Learning*, Vol 1., Numbers 1–2.

- Chapter 7: Convex Relaxations and Upper bounds
- Chapter 8: Integer programming, max-product, and LP relaxations
- Chapter 9: Moment matrices, semidefinite constraints and conic programming