Estimating the "Wrong" Graphical Model: Benefits in the Computation-Limited Setting
Martin J. Wainwright

Daniel L. Klein

APMA 2950P

March 24, 2010

Outline

1. Introduction
   - Scope
   - Wainwright's contributions
   - Review of MRFs and variational approximation
2. Method
   - Convex surrogates to the cumulant-generating function
   - Approximate parameter estimation
   - Joint estimation and prediction
3. Results
   - Analysis
   - Performance bounds
   - Experimental results
4. Conclusion

Problem domain

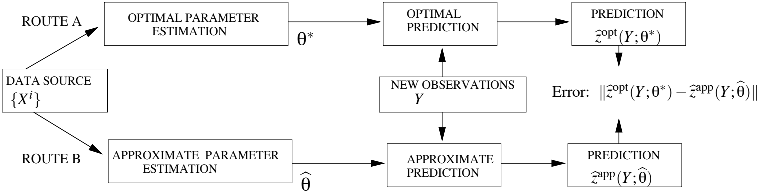
- Problem: joint parameter estimation and prediction in Markov random field.
- Tasks: smoothing, denoising, interpolation, missing data, etc.
- Applications: signal processing (denoising), machine learning (smoothing, interpolation), natural language processing (missing data), etc.

Approach

- Problem (detailed): given samples $\{X_1, \ldots, X_n\}$ from some unknown underlying model $p(\cdot;\theta^*)$, the first step is to form an estimate of the model parameters. Now suppose that we are given a noisy observation of a new sample $Z \sim p(\cdot;\theta^*)$, and that we wish to form a (near-)optimal estimate of Z using the fitted model, and the noisy observation (denoted Y).
- Principled route to obtaining approximations: relax the original optimization problem and take the optimal solutions to the relaxed problem as approximations to the exact values.

Two routes to a solution

Top route is optimal.



Bottom route introduces two approximations. Can we make these two errors (estimation and prediction) cancel out?
The bottom route is used in tree-reweighted sum-product, reweighted GBP, semidefinite relaxations, "convexified" expectation propagation, etc.

Markov random field: setup

- Undirected graph: $G = (V, E)$.
- Discrete state space: $\{0, 1, \ldots, m-1\}$.
- Singleton potentials:

$$\theta_s(x_s) \triangleq \sum_{j=1}^{m-1} \theta_{s;j} \mathbb{I}_j[x_s]$$

with $j = 0$ excluded to guarantee affine independence.

- Pairwise potentials:

$$\theta_{st}(x_s, x_t) \triangleq \sum_{j=1}^{m-1} \sum_{k=1}^{m-1} \theta_{st;jk} \mathbb{I}_j[x_s] \mathbb{I}_k[x_t]$$

similarly excluding $j = 0$ and $k = 0$.

Markov random field: global probability

Probability mass function

$$p(x;\theta) = \exp\left\{ \sum_{s\in V} \theta_s(x_s) + \sum_{(s,t)\in E} \theta_{st}(x_s, x_t) - A(\theta) \right\}$$

with normalizing term

$$A(\theta) \triangleq \log\left[\sum_{x\in X^n} \exp\left\{ \sum_{s\in V} \theta_s(x_s) + \sum_{(s,t)\in E} \theta_{st}(x_s, x_t) \right\} \right].$$

Markov random field: exponential family

The collection of distributions is a regular and minimal exponential family.

- Exponential parameter (vector) θ.
- Sufficient statistics (vector) ϕ.

Compactly, $p(x;\theta) = \exp\{\langle\theta(x),\phi\rangle - A(\theta)\}$, where $\theta \in \mathbb{R}^d$ with $d = N(m-1) + |E|(m-1)^2$.

Dimensionality of θ assumed not to be a problem.

Markov random field: properties of normalization term

It is clear that the normalization term is the log-partition function. We have the following properties (Lemma 1):

(a) A is a convex function of the parameters; strictly so when the sufficient statistics are affinely independent.

(b) A is infinitely differentiable, with

$$\frac{\partial A}{\partial \theta_\alpha} = \mathbb{E}_\theta[\phi_\alpha(X)] \quad \text{and} \quad \frac{\partial^2 A}{\partial \theta_\alpha \partial \theta_\beta} = \mathrm{cov}_\theta\{\phi_\alpha(X), \phi_\beta(X)\}.$$

Mean parameters correspond to marginal probabilities, e.g.,

$$\mu_{s;j} = \mathbb{E}_\theta[\mathbb{I}_j[x_s]] = p(X_s = j; \theta).$$

Background: cumulant-generating functions

Given a random variable $x \sim X = P(x)$, if there exists an $h > 0$ such that

$$M(t) \triangleq \left\langle e^{tx} \right\rangle$$

is defined for $|t| < h$, then we say that $M(t)$ is the moment-generating function for X.
We define the cumulant-generating function by

$$R(t) \triangleq \log M(t)$$

and we have the simple properties

$$\mu_X = R'(0) \quad \text{and} \quad \sigma_X^2 = R''(0).$$

Exact variational principle: conjugate dual function

Convexity and continuity gaurantee existence of variational representation, given in terms of conjugate dual function A^*, of the form

$$A(\theta) = \sup_{\mu \in \mathrm{MARG}_\phi(G)} \{\theta^T \mu - A^*(\mu)\}.$$

But what is A^*? Solving the constrained entropy maximization problem gives us

$$A^*(\mu) = \begin{cases} -H(p(\cdot; \theta(\mu))) & \text{if } \mu \in \mathrm{MARG}_\phi(G) \\ +\infty & \text{otherwise.} \end{cases}$$

Unfortunately, the complexity of the polytope $\mathrm{MARG}_\phi(G)$ grows non-polynomially in the size of G (notable exception: trees!).

Relaxed problem

We work with the relaxed optimization problem

$$B(\theta) \triangleq \max_{\tau \in \mathrm{REL}_\phi(G)} \{\theta^T \tau - B^*(\tau)\}$$

where:

- we must assume that B^* is strictly convex and twice-differentiable,
- $\mathrm{REL}_\phi(G)$ is a convex and compact set that acts as an outer bound to $\mathrm{MARG}_\phi(G)$, and,
- τ can be understood as pseudomarginals,

Relaxed problem: properties of convex surrogate

Our surrogate has the following properties:

- for each θ, $B(\theta)$ obtains a unique optimum $\tau(\theta)$,
- the function B is convex, and,
- the function B is differentiable with $\nabla B(\theta) = \tau(\theta)$.

These properties resemble the properties of A, so naming it the "convex surrogate" is justified.

Danskin's theorem

Properties follow from noting that the hypotheses are satisfied.

Theorem

(Danskin, 1966) Suppose $\phi(x, z)$ *is a continous function such that* $\phi : \mathbb{R}^n \times Z \to \mathbb{R}$ *with* $Z \subset \mathbb{R}^m$ *compact and assume that* ϕ *is convex in* x *for every* z*. Define the set of maximizing points*

$$Z_0(x) = \left\{ \bar{z} : \phi(x, \bar{z}) = \max_{z \in Z} \phi(x, z) \right\}.$$

Then, letting $f(x) = \max_{z \in Z} \phi(x, z)$*, we conclude:*

(i) $f(x)$ *is convex, and,*

(ii) $f(x)$ *is differentiable where* $Z_0(x)$ *consists of a single point, and at such points,*

$$\nabla f(x) = \frac{\partial}{\partial x} \phi(x, \bar{z}).$$

Danskin's theorem: intuition

Example: consider the special case of a single coin-flip with parameter $z = \theta$ the probability of getting heads and x the outcome of the flip (1 if heads). Then we have

$$\phi(x, z) = P(X = x|\theta)$$

which satisfies the conditions so

$$f(x) = \max_{\theta} P(X = x|\theta)$$

is convex, differentiable, and has a single point $Z_0(x)$ with the gradient condition.
In fact, this is completely uninteresting since our data are not able to vary continuous.

Danskin's theorem: intuition

Example: consider the special case of a nondegenerate set of i.i.d. draws $x_1, \ldots, x_n$ ($n > 1$) from a normal distribution with parameters $z = (\mu, \sigma)$. Then we have

$$\phi(x, z) = \log P(X_1 = x_1, \ldots, X_n = x_n | \mu, \sigma)$$

is convex and continuous in the data for any fixed parameters. Then, letting

$$f(x) = \max_{\mu, \sigma} P(X_1 = x_1, \ldots, X_n = x_n | \mu, \sigma),$$

we have that $f(x)$ is convex, differentiable, and has a single point set $Z_0(x)$ at which

$$\nabla f(x_1, \ldots, x_n) = \frac{\partial}{\partial x} P(x_1, \ldots, x_n | \hat{\mu}, \hat{\sigma}).$$

Example: convexified Bethe surrogate

Introduce standing example, an approximation exact for tree-structured MRF.

Relaxed polytope: local consistency of singleton and pairwise pseudomarginals.

Entropy approximation: associate collection $\mathcal{T}$ of spanning trees. Then define strictly convex function

$$B_\rho^*(\tau) \triangleq \sum_{T \in \mathcal{T}} \rho(T) \left\{ \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E(T)} I_{st}(\tau_{st}) \right\} .$$

Bethe surrogate and reweighted sum-product: use messages

$$M_{ts}(x_s) \leftarrow \sum_{x_t} \exp\left\{ \theta_t(x_t) \frac{\theta_{st}(x_s, x_t)}{\rho_{st}} \right\} \frac{\prod_{u \in \Gamma(t) \setminus s} [M_{ut}(x_t)]^{\rho_{ut}}}{[M_{st}(x_t)]^{1-\rho_{st}}} .$$

Joint estimation and prediction: setup

We want to find the posterior (predictive) distribution using

$$p(z|y;\theta) \propto p(z;\theta)p(y|z).$$

In the exponential family setting, the posterior can be given the form $\theta + \gamma(y)$ where determining the function γ can take some work.

Joint estimation and prediction: procedure

1. Form parameter estimate $\hat{\theta}^n$ from initial data $\{x^1, \ldots, x^n\}$ by maximizing the surrogate likelihood ℓ_B.
2. Given new noisy observation y specified by the factorized conditional distribution

$$p(y|z) = \prod_{s=1}^{N} p(y_s|z_s),$$

   incorporate it into the model by forming the new exponential parameter

$$\hat{\theta}_s^n(\cdot) + \gamma_s(y).$$

3. Use message-passing algorithm to compute approximate marginals $\tau(\hat{\theta} + \gamma)$, and use these marginals to compute prediction $\hat{z}(y; \tau)$.

Estimator: asymptotic results

[Regularizer sneakily introduced: note that this shouldn't have any asymptotic effect.]

Under sane conditions (non-negative, convex regularizer with parameter $\lambda^n = o(1/\sqrt{n})$), we have:

(a) $\hat{\theta}^n \xrightarrow{P} \hat{\theta}$, where $\hat{\theta}$ may be distinct from the true parameter θ^*, and,

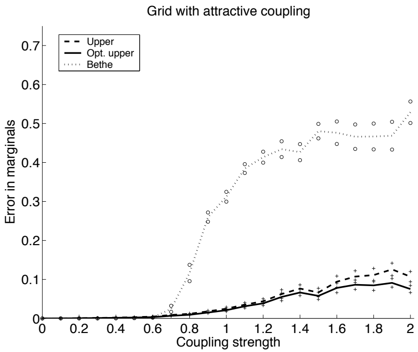
(b) the estimator is asymptotically normal.

Proof: clever use of the gradient and unique optimum properties of the convex surrogate.

Note that this estimator is inconsistent: the estimated model differs from the true model in in the limit of large data (even with the weak regularizer!?).

Estimator: global stability

Note that standard sum-product message-passing is not stable with respect to its inputs for tightly coupled MRFs due to the existence of multiple optima.



Some convex relaxation methods are provably globally stable.

Performance: problem setup

- Measure performance (mean-squared error) loss against Bayes optimum.
- Focus on the infinite data limit.
- Assume the multinomial random vector $X = \{X_s, s \in V\}$ is a label vector for the components in a finite mixture of Gaussians.
- Introduce, for each node $s \in V$, r.v.s Z_s and Y_s with

$$p(Z_s = z_s | X_s = j) \sim N(\nu_j, \sigma_j^2)$$

and

$$Y_s = \alpha Z_s + \sqrt{1 - \alpha^2} W_s.$$

Performance: Bayes least square estimator

Optimal BLSE (minimal MSE) takes the form

$$\hat{z}_s^{opt}(Y;\theta^*) \triangleq \sum_{j=0}^{m-1} \mu_{s;j}(\theta^* + \gamma(Y)) \left[\omega_j(\alpha)(Y_s - \alpha\nu_j) + \nu_j\right]$$

where

$$\omega_j(\alpha) \triangleq \frac{\alpha\sigma_j^2}{\alpha^2\sigma_j^2 + (1-\alpha^2)}.$$

To calculate this, we need θ^* (unknown) and marginals (impractical to compute).

Performance: approximate prediction

Instead, use the surrogate-based predictor

$$\hat{z}_s^{app}(Y;\hat{\theta}) \triangleq \sum_{j=0}^{m-1} \tau_{s;j}(\hat{\theta}+\gamma(Y))\left[\omega_j(\alpha)(Y_s-\alpha\nu_j)+\nu_j\right].$$

Can we bound the (difference in) MSE

$$\Delta R(\alpha,\theta^*,\hat{\theta}) \triangleq R^{app}(\alpha,\hat{\theta} - R^{opt}(\alpha,\theta^*))$$

from above?

Performance: role of stability

In passing, at $\alpha \approx 1$ limit, marginals don't really matter; at $\alpha \approx 0$ limit, inconsistency errors cancel variational errors.
Introduce Lipschitz stability

$$L(\theta^*; \hat{\theta}) \triangleq \sup_{\delta \in \mathbb{R}^d} \sigma_{max}(\nabla^2 A(\theta^* + \delta) - \nabla^2 B(\hat{\theta} + \delta)).$$

Then we have (Theorem 7)

$$\Delta R(\alpha, \theta^*, \hat{\theta}) \leq$$
$$\mathbb{E}\left\{ \min\left(1, L(\theta^*; \hat{\theta}) \frac{||\gamma(Y; \alpha)||_2}{\sqrt{N}}\right) \sqrt{\frac{\sum_{s=1}^N |g_1(Y_s) - g_0(Y_s)|^4}{N}} \right\}.$$

Taking various limits, we get asymptotic optimality.

Tree-reweighted sum-product

Specified by collection of edge weights ρ_{st}, one for each edge (s,t) of the graph, where the vector of edge weights belongs to the spanning tree polytope.

Fix ρ. The procedure is

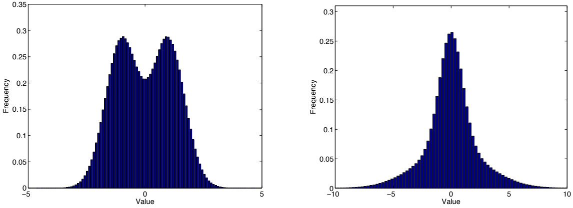
1. Compute empirical marginal distributions $\hat{\mu}_{s;j}$ and $\hat{\mu}_{st;jk}$ and hence approximate parameters

$$\hat{\theta}^n_{s;j} \triangleq \log \hat{\mu}_{s;j} \quad \text{and} \quad \hat{\theta}^n_{st;jk} \triangleq \rho_{st} \log \frac{\hat{\mu}_{st;jk}}{\hat{\mu}_{s;j}\hat{\mu}_{t;k}}.$$

2. Form new exponential parameter $\hat{\theta}_a s^n + \gamma_s(Y)$, where γ_s is appropriate to Gaussian mixture model.
3. Compute approximate marginals $\tau(\hat{\theta} + \gamma)$ by running tree-reweighted sum-product with edge weights ρ_{st} on model with parameters $\hat{\theta} + \gamma$. These give $\hat{z}^{app}$.

Experimental setup: mixtures

We have a mixture of $m = 2$ Gaussians.



(a) (b)

Figure 3: Histograms of different Gaussian mixture ensembles. (a) Ensemble A: a bimodal ensemble with $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$. (b) Ensemble B: mimics a heavy-tailed distribution, with $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_1, \sigma_1^2) = (0, 9)$.

Out graph is a 2D grid with $N = 64$ nodes, where $x \in \{-1, +1\}^N$ are spins. Consider attractive and mixed coupling.

Comparison: true model versus approximate model

Attractive couping, equal variances.

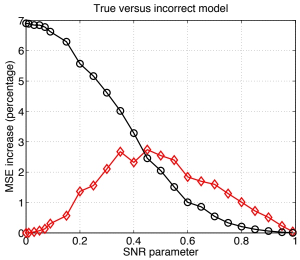
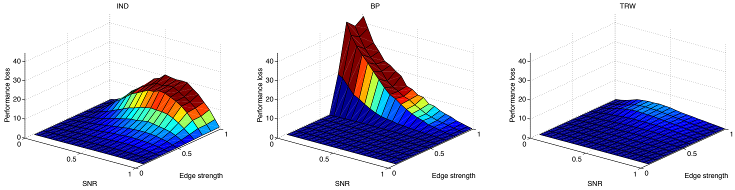


Figure 4: Line plots of percentage increase in MSE relative to Bayes optimum for the TRW method applied to the true model (black circles) versus the approximate model (red diamonds) as a function of observation SNR for grids with $N = 64$ nodes, and attractive coupling $\beta = 0.70$. As predicted by theory, using the "incorrect" model leads to superior performance, when prediction is performed using the approximate TRW method, for a range of SNR.

Comparison: tree-reweighted and ordinary sum-product

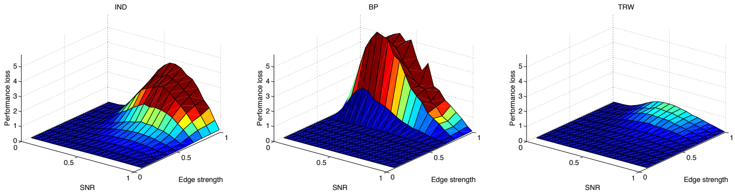
Attractive coupling, equal means.



Left to right: independence, ordinary BP, tree-reweighted

Comparison: tree-reweighted and ordinary sum-product

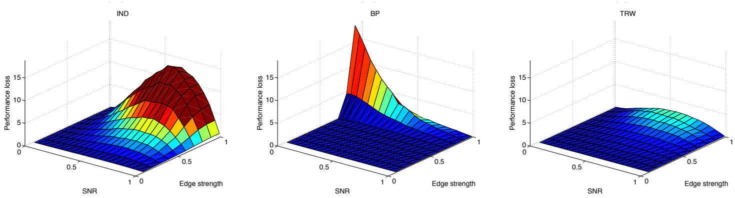
Mixed coupling, equal variances.



Left to right: independence, ordinary BP, tree-reweighted

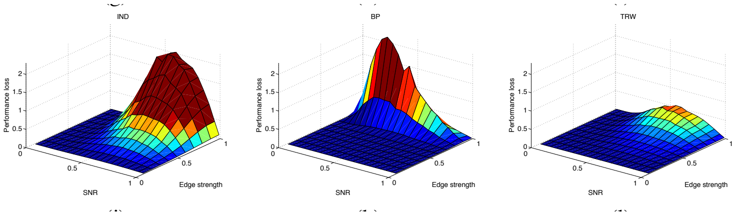
Comparison: tree-reweighted and ordinary sum-product

Mixed coupling, equal means.



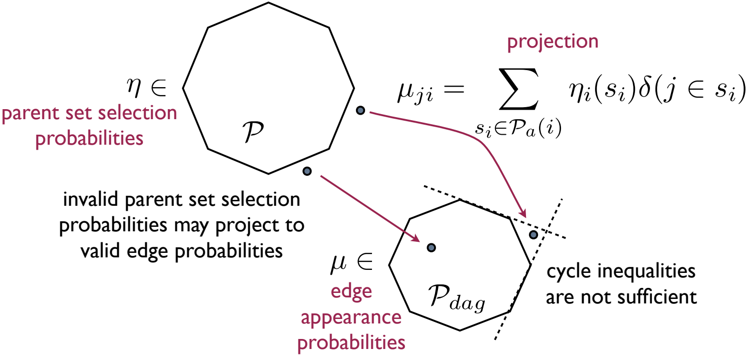
Left to right: independence, ordinary BP, tree-reweighted

Comparison: tree-reweighted and ordinary sum-product



Left to right: independence, ordinary BP, tree-reweighted

Connections to Tommi Jaakkola's PTG talk



Summary

Punch line: in computation-limited setting, using an inconsistent parameter estimator is provably and empirically beneficial.