Probabilistic Graphical Models

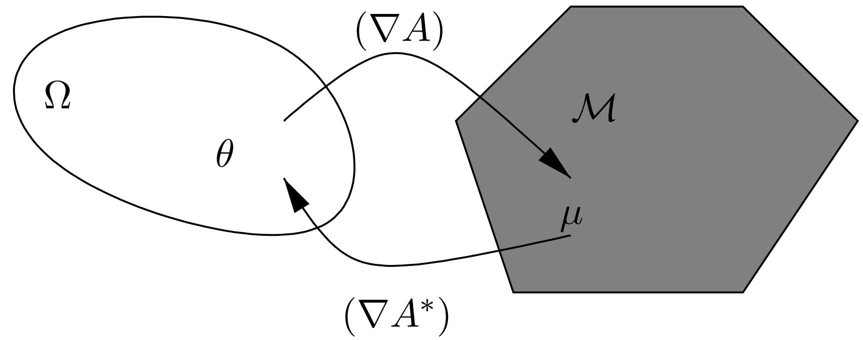
Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 23:
Reweighted Sum-Product Belief Propagation,
Convex Surrogates for Variational Learning

Some figures and examples courtesy M. Wainwright & M. Jordan,
Graphical Models, Exponential Families, & Variational Inference, 2008.

Inference as Optimization

$$p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\}$$
$$A(\theta) = \log \sum_{x \in \mathcal{X}} \exp\{\theta^T \phi(x)\}$$
$$\mathcal{M} = \mathrm{conv}\{\phi(x) \mid x \in \mathcal{X}\}$$



- Express log-partition as optimization over all distributions $\mathcal{Q}$

$$A(\theta) = \sup_{q \in \mathcal{Q}} \left\{ \sum_{x \in \mathcal{X}} \theta^T \phi(x) q(x) - \sum_{x \in \mathcal{X}} q(x) \log q(x) \right\}$$

Jensen's inequality gives arg max: $q(x) = p(x \mid \theta)$

- More compact to optimize over relevant *sufficient statistics*:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu + H(p(x \mid \theta(\mu)) \right\}$$

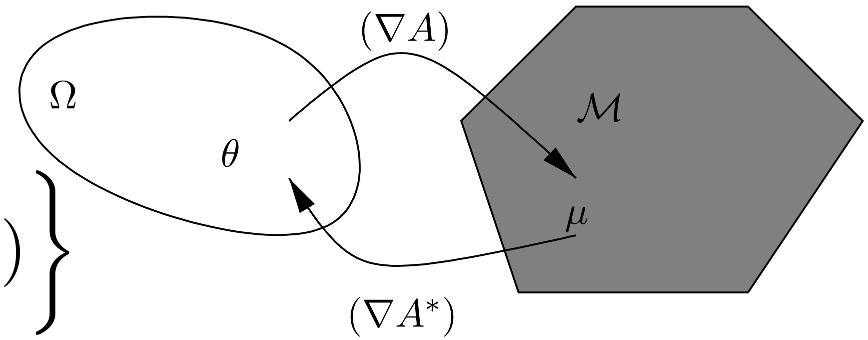
concave function (linear plus entropy) over a convex set

$$\mu = \sum_{x \in \mathcal{X}} \phi(x) q(x) = \sum_{x \in \mathcal{X}} \phi(x) p(x \mid \theta(\mu))$$

Variational Inference Approximations

$$p(x \mid \theta) = \exp\{\theta^T \phi(x) - A(\theta)\}$$

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \theta^T \mu + H(p(x \mid \theta(\mu)) \right\}$$



Mean Field: Lower bound log-partition function

- Restrict optimization to some simpler subset $\mathcal{M}_- \subset \mathcal{M}$
- Imposing conditional independencies makes entropy tractable

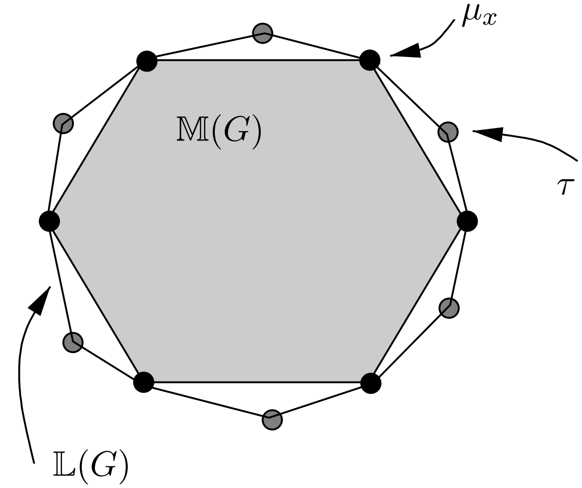
Bethe & Loopy BP: Approximate log-partition function

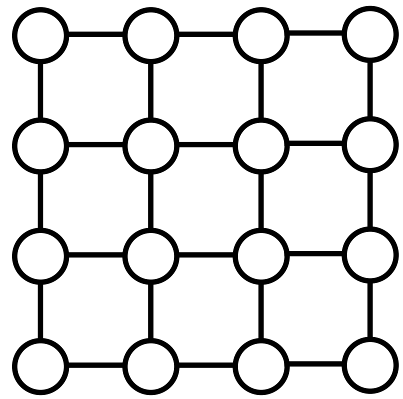
- Define tractable outer bound on constraints $\mathcal{M}_+ \supset \mathcal{M}$
- Tree-based models give approximation to true entropy

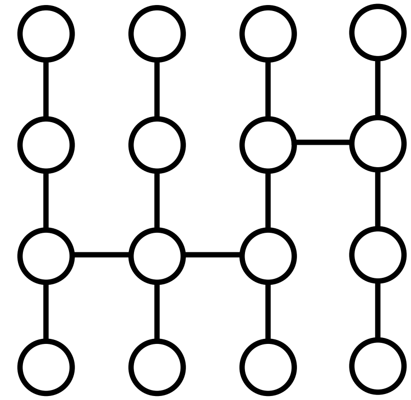
Reweighted BP: Upper bound log-partition function

- Define tractable outer bound on constraints $\mathcal{M}_+ \supset \mathcal{M}$
- Tree-based models give tractable upper bound on true entropy

Tree-Based Outer Approximations







- For some graph G, denote true marginal polytope by $\mathbb{M}(G)$
- Associate marginals with nodes and edges, and impose the following *local consistency* constraints $\mathbb{L}(G)$

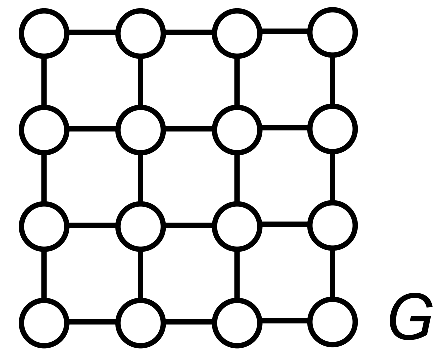
$$\sum_{x_s} \mu_s(x_s) = 1, \quad s \in \mathcal{V} \qquad \mu_s(x_s) \geq 0, \mu_{st}(x_s, x_t) \geq 0$$

$$\sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s), \quad (s,t) \in \mathcal{E}, x_s \in \mathcal{X}_s$$

- For any graph, this is a *convex* outer bound: $\mathbb{M}(G) \subseteq \mathbb{L}(G)$
- For any tree-structured graph *T*, we have $\mathbb{M}(T) = \mathbb{L}(T)$

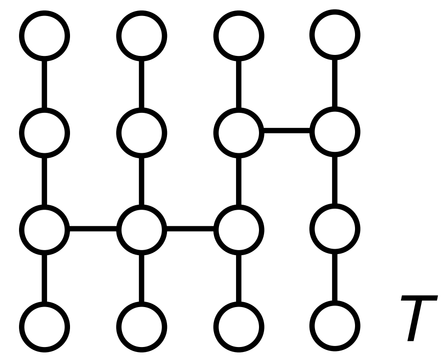
Tree-Based Entropy Bounds

$$p(x) = \frac{1}{Z} \exp\left\{ -\sum_{(s,t)\in\mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s\in\mathcal{V}} \phi_s(x_s) \right\}$$



$$H(\mu(T)) = \sum_{s\in\mathcal{V}} H_s(\mu_s) - \sum_{(s,t)\in\mathcal{E}(T)} I_{st}(\mu_{st})$$

$$H(\mu) \le H(\mu(T))$$ *for any tree T*



Maximum entropy property of exponential families:

- Original distribution maximizes entropy subject to constraints

$$\mathbb{E}_p[\phi_{st}(x_s, x_t)] = \mu(x_s, x_t), \qquad (s,t) \in \mathcal{E}$$

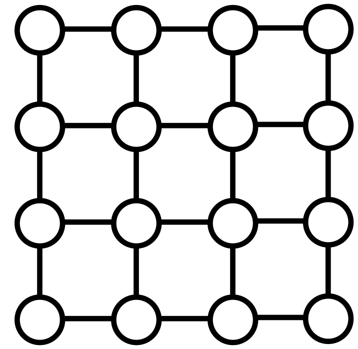
- Tree-structured distribution maximizes subject to a *subset* of the full constraints (those corresponding to edges in tree):

$$\mathbb{E}_p[\phi_{st}(x_s, x_t)] = \mu(x_s, x_t), \qquad (s,t) \in \mathcal{E}(T)$$

Tree-Based Entropy Bounds

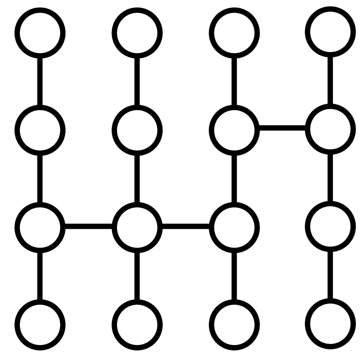
$$p(x) = \frac{1}{Z} \exp\left\{ - \sum_{(s,t)\in\mathcal{E}} \phi_{st}(x_s, x_t) - \sum_{s\in\mathcal{V}} \phi_s(x_s) \right\}$$

$$H(\mu(T)) = \sum_{s\in\mathcal{V}} H_s(\mu_s) - \sum_{(s,t)\in\mathcal{E}(T)} I_{st}(\mu_{st})$$



$$H(\mu) \leq H(\mu(T))$$ *for any tree T*

$$H(\mu) \leq \sum_{s\in\mathcal{V}} H_s(\mu_s) - \sum_{(s,t)\in\mathcal{E}} \rho_{st} I_{st}(\mu_{st})$$

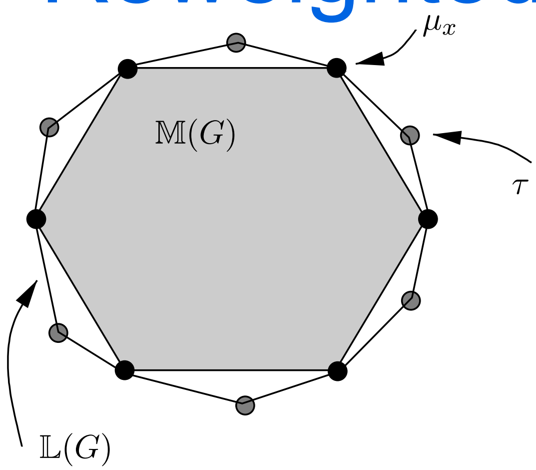


- Family of bounds depends on edge appearance probabilities from some distribution over subtrees in the original graph:

$$H(\mu) \leq \sum_{T} \rho(T) H(\mu(T)) \qquad \rho_{st} = \mathbb{E}_\rho\left[\mathbb{I}\left[(s,t) \in E(T)\right]\right]$$

Must only specify a single scalar parameter per edge

Reweighted Bethe Variational Methods



$$A(\theta) \leq \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_\rho(\tau) \right\}$$

$$H_\rho(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} \rho_{st} I_{st}(\tau_{st})$$

- Local consistency constraints are convex, but allow globally inconsistent *pseudo-marginals* on graphs with cycles
- Assuming we pick weights corresponding to some distribution on acyclic sub-graphs, have *upper bound* on true entropy
- This defines a *convex surrogate* to true variational problem

Issues to resolve:

- Given edge weights, how can we efficiently find the best pseudo-marginals? A message-passing algorithm?
- There are many distributions over spanning trees. How can we find the best edge appearance probabilities?

Reweighted Belief Propagation

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

Standard Loopy BP:

$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)}$$

$$q_t(x_t) \propto \psi_t(x_t) \prod_{u\in\Gamma(t)} m_{ut}(x_t)$$

Lagrangian derivation generalizes to reweighted Bethe objective

Reweighted BP:

$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t)^{1/\rho_{st}} \frac{q_t(x_t)}{m_{st}(x_t)}$$

$$q_t(x_t) \propto \psi_t(x_t) \prod_{u\in\Gamma(t)} m_{ut}(x_t)^{\rho_{ut}}$$

For loopy graphs, "down-weights" messages to be more uniform

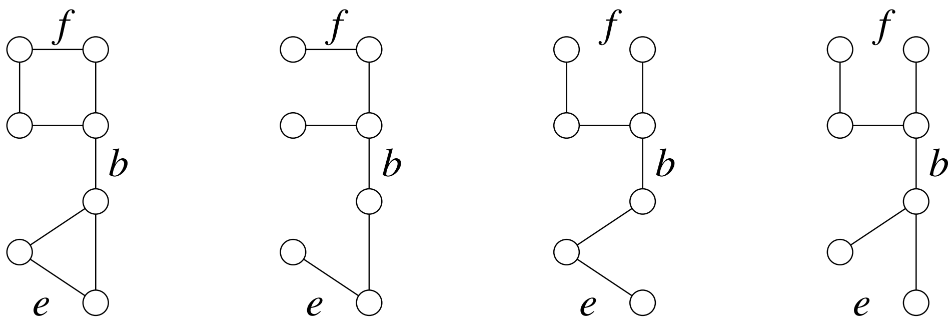
$$m_{ts}(x_s) \propto \left[\sum_{x_t} \psi_{st}(x_s, x_t)^{1/\rho_{st}} \frac{q_t(x_t)}{m_{st}(x_t)^{1/\rho_{st}}} \right]^{\rho_{st}}$$

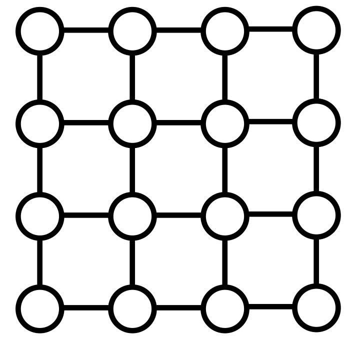
$$q_t(x_t) \propto \psi_t(x_t) \prod_{u\in\Gamma(t)} m_{ut}(x_t)$$

Applying a change of variables:

$$m_{ut}(x_t) \leftarrow m_{ut}(x_t)^{\rho_{ut}}$$

Spanning Tree Polytope

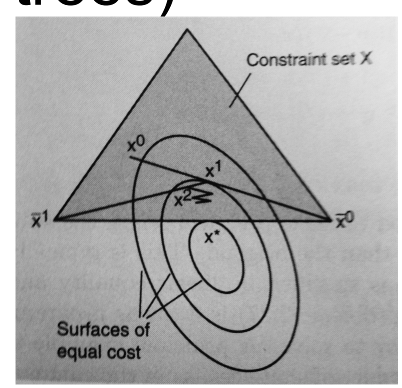




$$\rho_b = 1, \rho_e = \frac{2}{3}, \rho_f = \frac{1}{3} \qquad\qquad \rho_{st} = \frac{1}{2}$$

$$A(\theta) \leq \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_\rho(\tau) \right\} \qquad H_\rho(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} \rho_{st} I_{st}(\tau_{st})$$

- Bound holds assuming edge weights lie in the *spanning tree polytope* (generated by some valid distribution on trees)
- Optimize via *conditional gradient* method:
  - ➢ Find descent direction by maximizing linear function (gradient) over constraint set
  - ➢ For spanning tree polytope, this reduces to a maximum weight spanning tree problem
  - ➢ Iteratively tightens bound on partition function



Bertsekas 1999

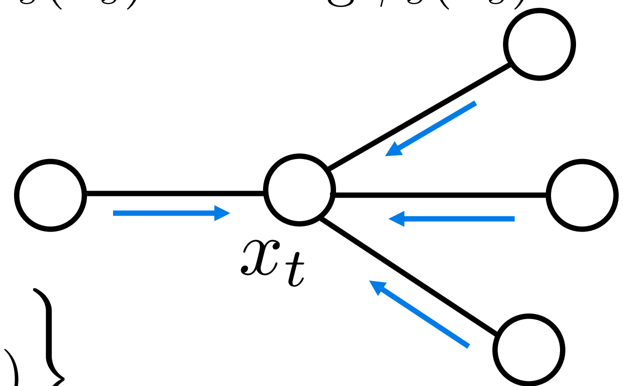
MF & Reweighted BP: Message Passing

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

$$\phi_{st}(x_s, x_t) = -\log \psi_{st}(x_s, x_t)$$
$$\phi_s(x_s) = -\log \psi_s(x_s)$$

Beliefs: *pseudo-marginals*

$$q_t(x_t) = \frac{1}{Z_t} \psi_t(x_t) \prod_{u\in\Gamma(t)} m_{ut}(x_t)$$



Mean Field

$$m_{ts}(x_s) \propto \exp\left\{ -\sum_{x_t} \phi_{st}(x_s, x_t) q_t(x_t) \right\}$$

Loopy BP

$$m_{ts}(x_s) \propto \sum_{x_t} \psi_{st}(x_s, x_t) \frac{q_t(x_t)}{m_{st}(x_t)}$$

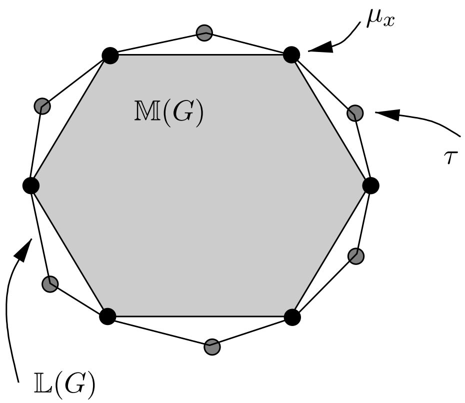
Reweight BP

$$m_{ts}(x_s) \propto \left[\sum_{x_t} \psi_{st}(x_s, x_t)^{1/\rho_{st}} \frac{q_t(x_t)}{m_{st}(x_t)^{1/\rho_{st}}} \right]^{\rho_{st}}$$

- Reweighted BP becomes loopy BP when $\rho_{st} = 1$
- Reweighted BP approaches mean field as $\rho_{st} \to \infty$

  Geometric mean is limit of power mean

MF & Reweighted BP: Variational Objective



$$A(\theta) \approx \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_\rho(\tau) \right\}$$

$$H_\rho(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} \rho_{st} I_{st}(\tau_{st})$$

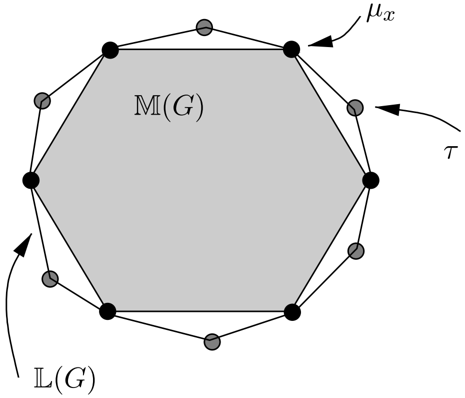
- View edge weights as positive, tunable parameters
- In the limit where they become very large:

$$\tau_{st} \to \infty \quad \Rightarrow \quad \textit{optimum sets} \;\; I_{st}(\tau_{st}) = 0 \quad \Rightarrow \quad \tau_{st}(x_s, x_t) = \tau_s(x_s)\tau_t(x_t)$$

Mean Field: For acyclic edge set $\rho_{st} = 1$, otherwise $\rho_{st} \to \infty$

- *Objective:* Lower bounds true $A(\theta)$, but non-convex
- *Message-passing:* Guaranteed convergent, but local optima

MF & Reweighted BP: Variational Objective



$$A(\theta) \approx \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_\rho(\tau) \right\}$$

$$H_\rho(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} \rho_{st} I_{st}(\tau_{st})$$

Loopy BP: For all edges, set $\rho_{st} = 1$

- *Objective:* Approximation, possibly poor, generally non-convex
- *Message-passing:* Multiple optima, may not convergent
- *But*, for some models gives most accurate marginal estimates

Reweighted BP: Respect spanning tree polytope, $0 < \rho_{st} \leq 1$

- *Objective:* Upper bounds true $A(\theta)$, convex
- *Message-passing:* Single global optimum, typically convergent

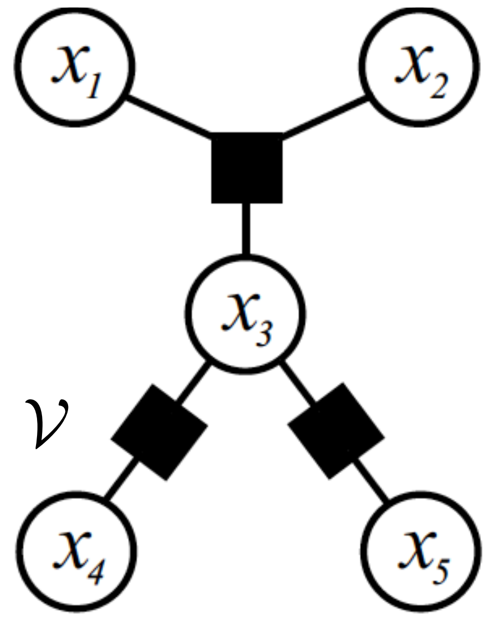
Mean Field: For acyclic edge set $\rho_{st} = 1$, otherwise $\rho_{st} \to \infty$

- *Objective:* Lower bounds true $A(\theta)$, but non-convex
- *Message-passing:* Guaranteed convergent, but local optima

Undirected Graphical Models

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \prod_{f \in \mathcal{F}} \psi_f(x_f \mid \theta_f)$$

$$Z(\theta) = \sum_x \prod_{f \in \mathcal{F}} \psi_f(x_f \mid \theta_f)$$



$\mathcal{F} \longrightarrow$ set of hyperedges linking subsets of nodes $f \subseteq \mathcal{V}$

$\mathcal{V} \longrightarrow$ set of N nodes or vertices, $\{1, 2, \ldots, N\}$

- Assume an exponential family representation of each factor:

$$p(x \mid \theta) = \exp\left\{ \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_f) - A(\theta) \right\}$$

$$\psi_f(x_f \mid \theta_f) = \exp\{\theta_f^T \phi_f(x_f)\} \qquad A(\theta) = \log Z(\theta)$$

- Partition function *globally* couples the local factor parameters

Learning for Undirected Models

- Undirected graph encodes dependencies within a single training example:

$$p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} \frac{1}{Z(\theta)} \prod_{f \in \mathcal{F}} \psi_f(x_{f,n} \mid \theta_f) \qquad \mathcal{D} = \{x_{\mathcal{V},1}, \ldots, x_{\mathcal{V},N}\}$$

- Given N independent, identically distributed, completely observed samples:

$$\log p(\mathcal{D} \mid \theta) = \left[\sum_{n=1}^{N} \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_{f,n}) \right] - N A(\theta)$$

$$p(x \mid \theta) = \exp\left\{ \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_f) - A(\theta) \right\}$$

Learning for Undirected Models

- Undirected graph encodes dependencies within a single training example:

$$p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} \frac{1}{Z(\theta)} \prod_{f \in \mathcal{F}} \psi_f(x_{f,n} \mid \theta_f) \qquad \mathcal{D} = \{x_{\mathcal{V},1}, \ldots, x_{\mathcal{V},N}\}$$

- Given N independent, identically distributed, completely observed samples:

$$\log p(\mathcal{D} \mid \theta) = \left[\sum_{n=1}^{N} \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_{f,n}) \right] - N A(\theta)$$

- Take gradient with respect to parameters for a single factor:

$$\nabla_{\theta_f} \log p(\mathcal{D} \mid \theta) = \left[\sum_{n=1}^{N} \phi_f(x_{f,n}) \right] - N \mathbb{E}_\theta[\phi_f(x_f)]$$

- Must be able to compute *marginal distributions* for factors in current model:
  - Tractable for tree-structured factor graphs via sum-product
  - What about general factor graphs or undirected graphs?

Convex Likelihood Surrogates

$$\log p(\mathcal{D} \mid \theta) = \left[\sum_{n=1}^{N} \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_{f,n}) \right] - NA(\theta)$$

$$\log p(\mathcal{D} \mid \theta) \geq \left[\sum_{n=1}^{N} \sum_{f \in \mathcal{F}} \theta_f^T \phi_f(x_{f,n}) \right] - NB(\theta) \triangleq L_B(\theta)$$

where we pick a bound satisfying $A(\theta) \leq B(\theta), B(\theta)$ convex

- Apply reweighted Bethe (generalizes to higher-order factors):

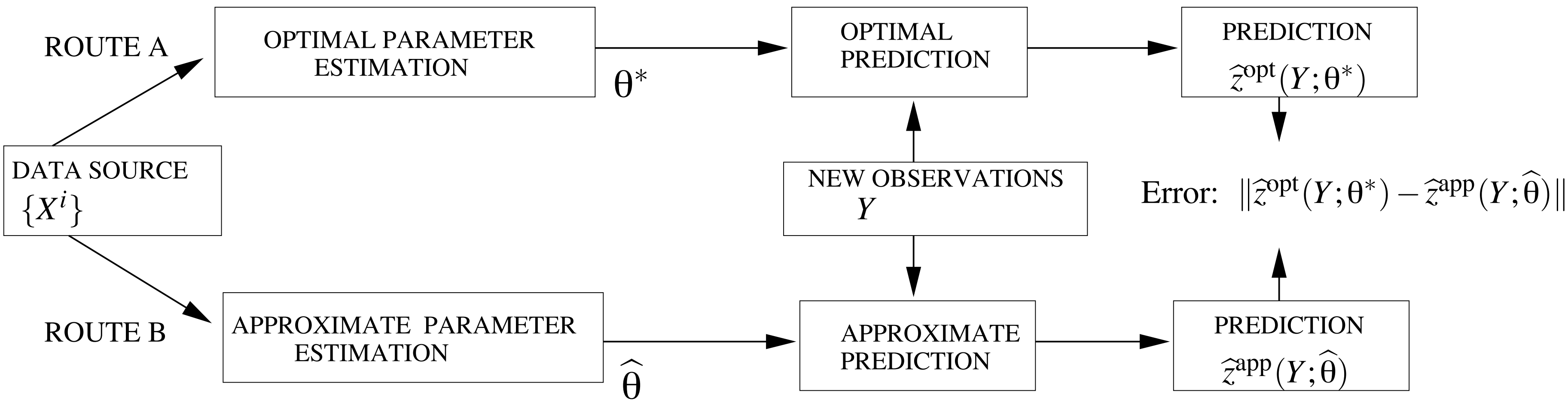
$$B(\theta) = \sup_{\tau \in \mathbb{L}(G)} \left\{ \theta^T \tau + H_\rho(\tau) \right\} \qquad H_\rho(\tau) = \sum_{s \in \mathcal{V}} H_s(\tau_s) - \sum_{(s,t) \in \mathcal{E}} \rho_{st} I_{st}(\tau_{st})$$

$$\nabla_{\theta_f} L_B(\theta) = \left[\sum_{n=1}^{N} \phi_f(x_{f,n}) \right] - N\mathbb{E}_\tau[\phi_f(x_f)]$$

- Gradients depend on expectations of *pseudo-marginals* produced by applying reweighted BP to current model

Approximate Learning & Inference:

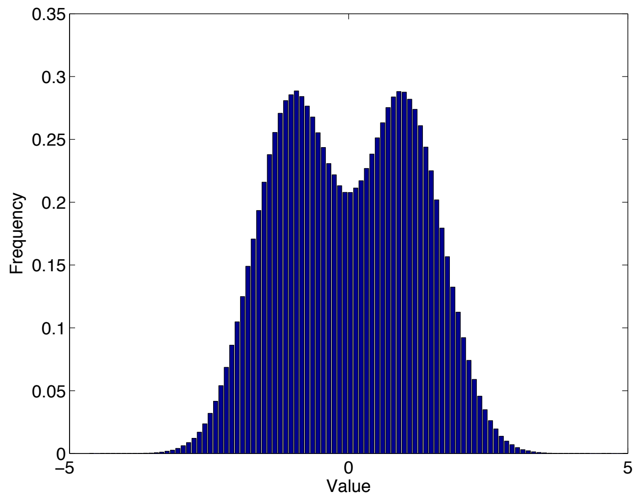
Two Wrongs Make a Right

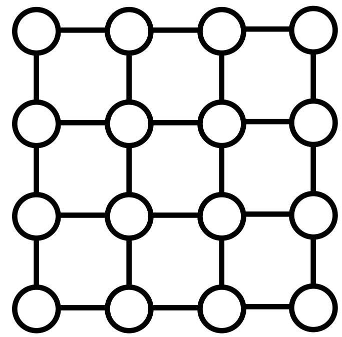


- Empirical Folk Theorem: Performance is better if the inference approximations used to learn parameters from training data are “matched” to those used for test examples
- Actual Theorem roughly shows: If learn based on *convex* upper bound to true partition function, can bound error on predictions for test examples which are “close” to training data
- Non-convexity & local optima bad in theory & practice

Wainwright 2006

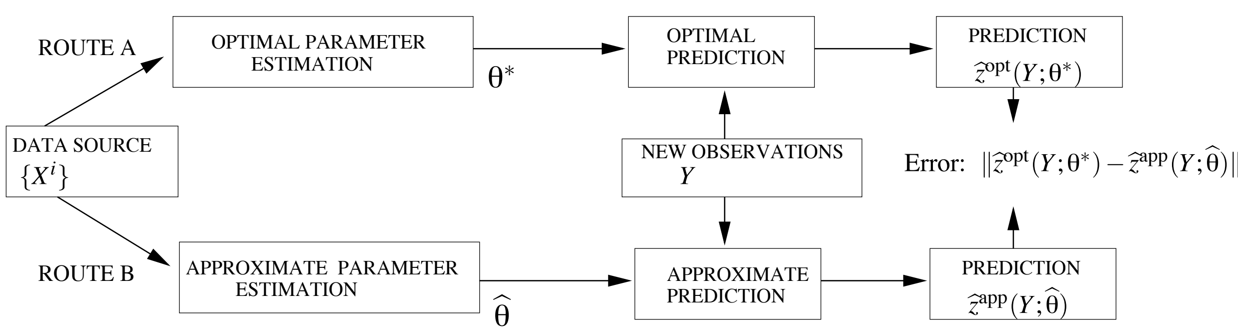
Example: Spatially Coupled Mixtures

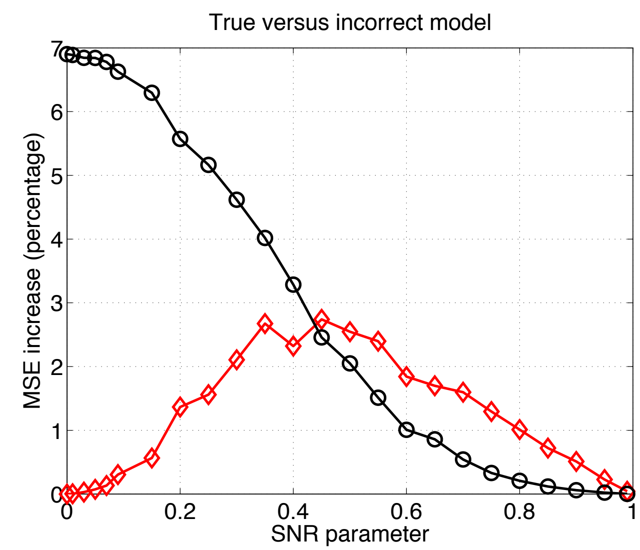




8x8 grid

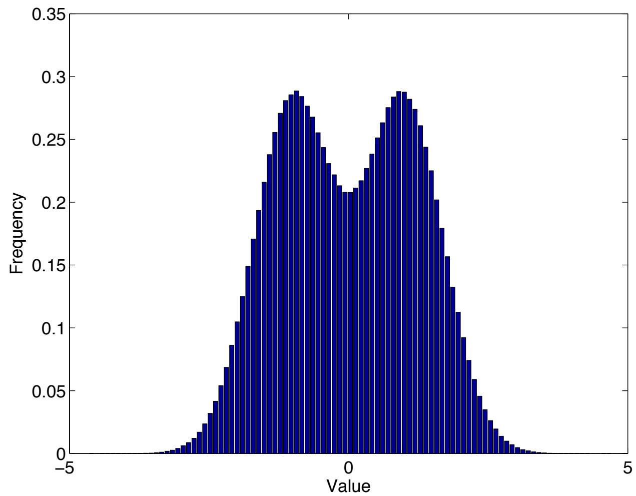
Real-valued spatial fields from mixture of two Gaussians, with positive spatial correlation in mixture component selection

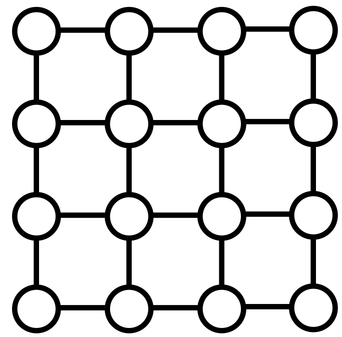




Wainwright 2006

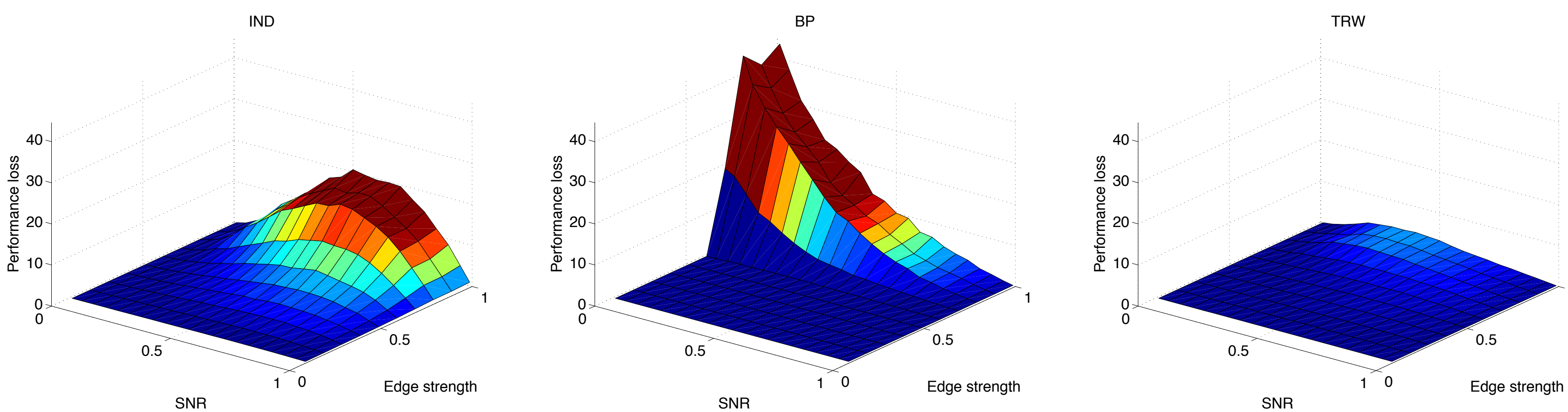
Example: Spatially Coupled Mixtures





8x8 grid

Real-valued spatial fields from mixture of two Gaussians, with positive spatial correlation in mixture component selection



Wainwright 2006