Variational Mean Field for Graphical Models

CS/CNS/EE 155

Baback Moghaddam

Machine Learning Group

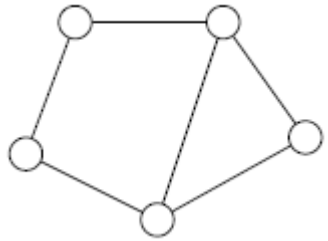
baback@jpl.nasa.gov

JPL

Approximate Inference

- Consider general UGs (*i.e.*, not tree-structured)

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathbf{C}} \exp\{\theta_C(x_C)\}$$



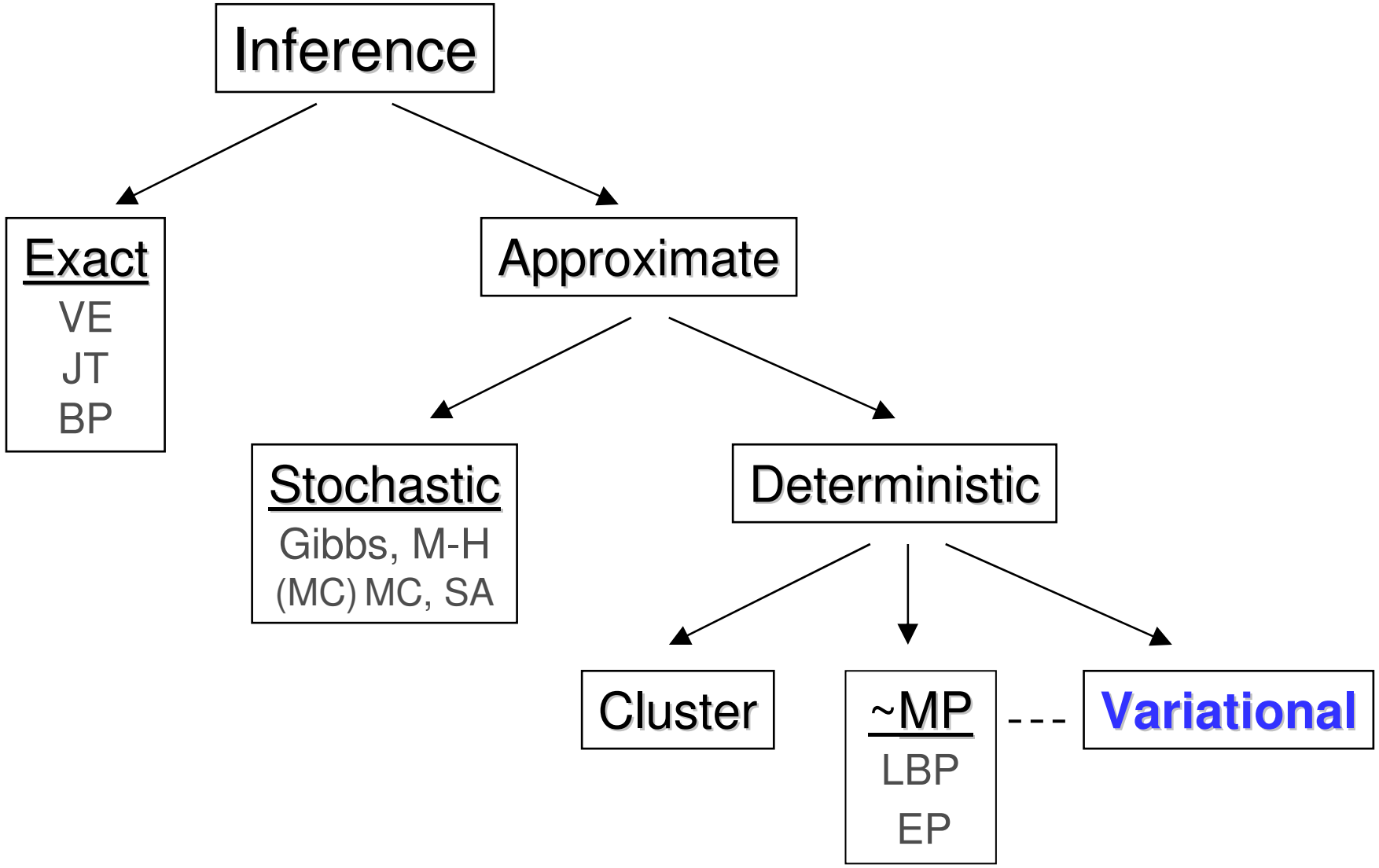
- All basic computations are intractable (for large G)

  - likelihoods & partition function $Z = \sum_{x \in \mathcal{X}^N} \prod_{C \in \mathcal{C}} \exp\{\theta_C(x_C)\}$

  - marginals & conditionals $p(X_s = x_s) = \sum_{x_t,\, t \neq s} \prod_{C \in \mathcal{C}} \exp\{\theta_C(x_C)\}$

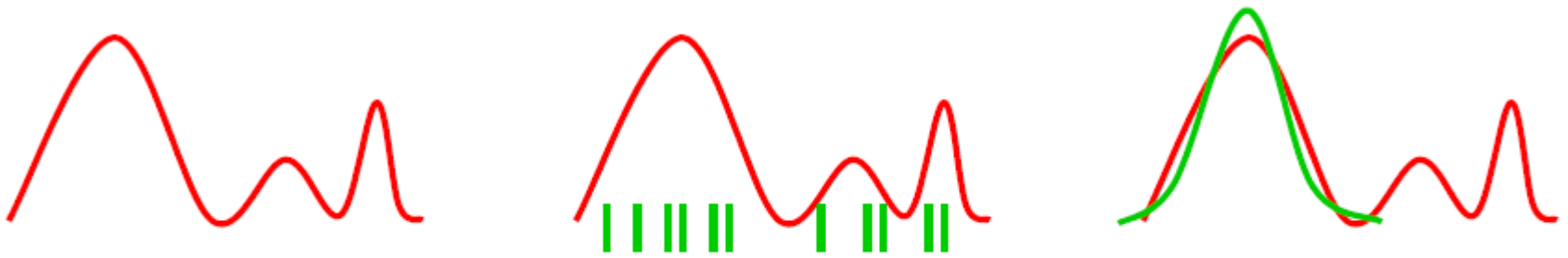
  - finding modes $\widehat{\mathbf{x}} = \arg\max_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}) = \arg\max_{\mathbf{x} \in \mathcal{X}^N} \prod_{C \in \mathcal{C}} \exp\{\theta_C(x_C)\}$

Taxonomy of Inference Methods



Approximate Inference

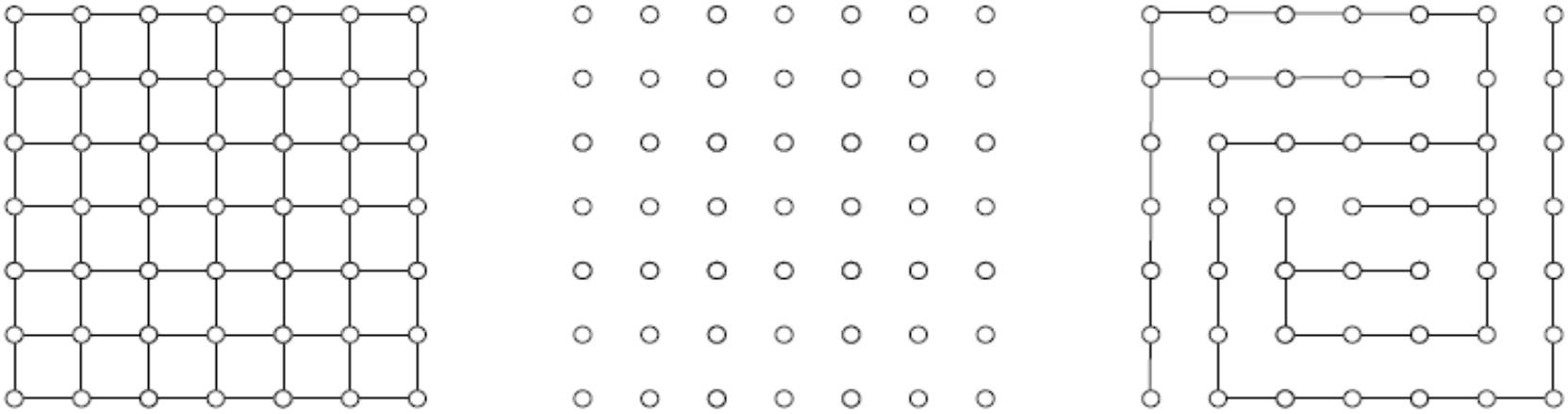
- Stochastic (Sampling)
  - Metropolis-Hastings, Gibbs, (Markov Chain) Monte Carlo, *etc*
  - Computationally *expensive*, but is "exact" (in the limit)
- Deterministic (Optimization)
  - Mean Field (MF), Loopy Belief Propagation (LBP)
  - Variational Bayes (VB), Expectation Propagation (EP)
  - Computationally *cheaper*, but is not exact (gives bounds)



Mean Field : Overview

- General idea
  - approximate $p(x)$ by a simpler factored distribution $q(x)$
  - minimize "distance" $D(p||q)$ - *e.g.,* Kullback-Liebler

original G | (Naïve) MF H_0 | structured MF H_s

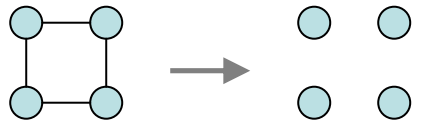


$$p(x) \propto \prod_c \phi_c(x_c) \qquad q(x) \propto \prod_i q_i(x_i) \qquad q(x) \propto q_A(x_A)\, q_B(x_B)$$

Mean Field : Overview

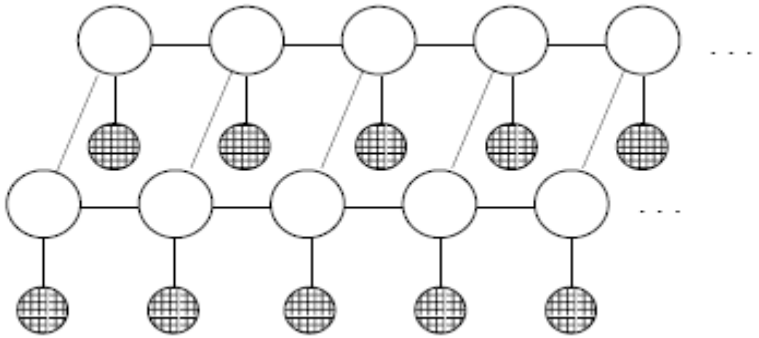
- Naïve MF has roots in *Statistical Mechanics* (1890s)
  - physics of spin glasses (Ising), ferromagnetism, *etc*
  - why is it called "Mean Field" ?

    with full factorization : $E[x_i x_j] = E[x_i]\ E[x_j]$



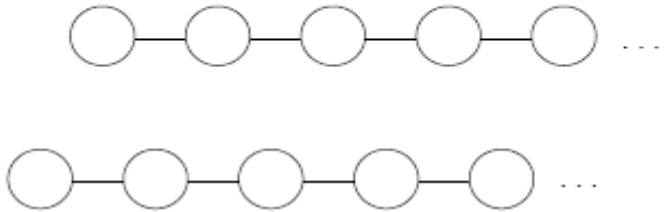
- Structured MF is more "modern"

Coupled HMM



Structured MF approximation

(with tractable chains)



KL Projection $D(Q||P)$

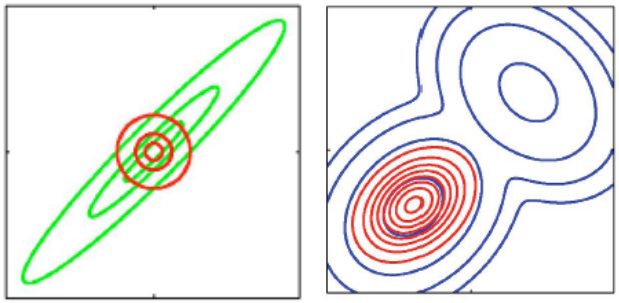
- Infer hidden h given visible v (clamp v nodes with δ's)

$$P(h|v) = \prod_{c \in C(G)} f_c(h_c), \quad Q(h) = \prod_{c \in C(G')} q_c(h_c)$$

- Variational : optimize KL globally

$$\min_Q D(Q||P) = \sum_h Q(h) \ln \frac{Q(h)}{P(h|v)}$$

$P = 0$ forces $Q = 0$



the right density form for Q "falls out"

KL is *easier* since we're taking $E[.]$ wrt simpler Q

Q seeks mode with the largest mass (not height)

so it will tend to *underestimate* the support of P

KL Projection $D(P||Q)$

- Infer hidden h given visible v (clamp v nodes with δ's)

$$P(h|v) = \prod_{c \in C(G)} f_c(h_c), \quad Q(h) = \prod_{c \in C(G')} q_c(h_c)$$

- Expectation Propagation (EP) : optimize KL locally

$$\min_{q_c} D(P||Q) = \sum_h P(h|v) \ln \frac{P(h|v)}{Q(h)}$$

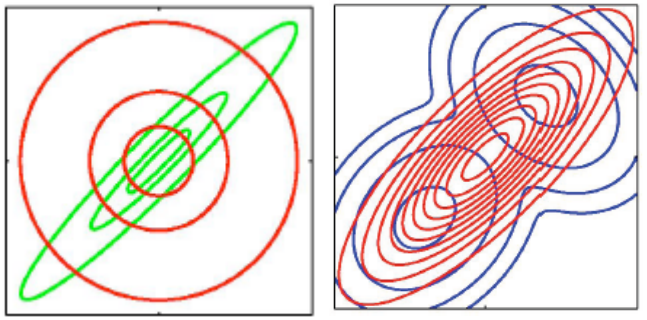
this KL is *harder* since we're taking $E[.]$ wrt P

no nice global solution for Q "falls out"

must sequentially tweak each q_c (match moments)

Q covers *all* modes so it *overestimates* support

$P > 0$ forces $Q > 0$



α - divergences

- The 2 basic KL divergences are *special cases* of

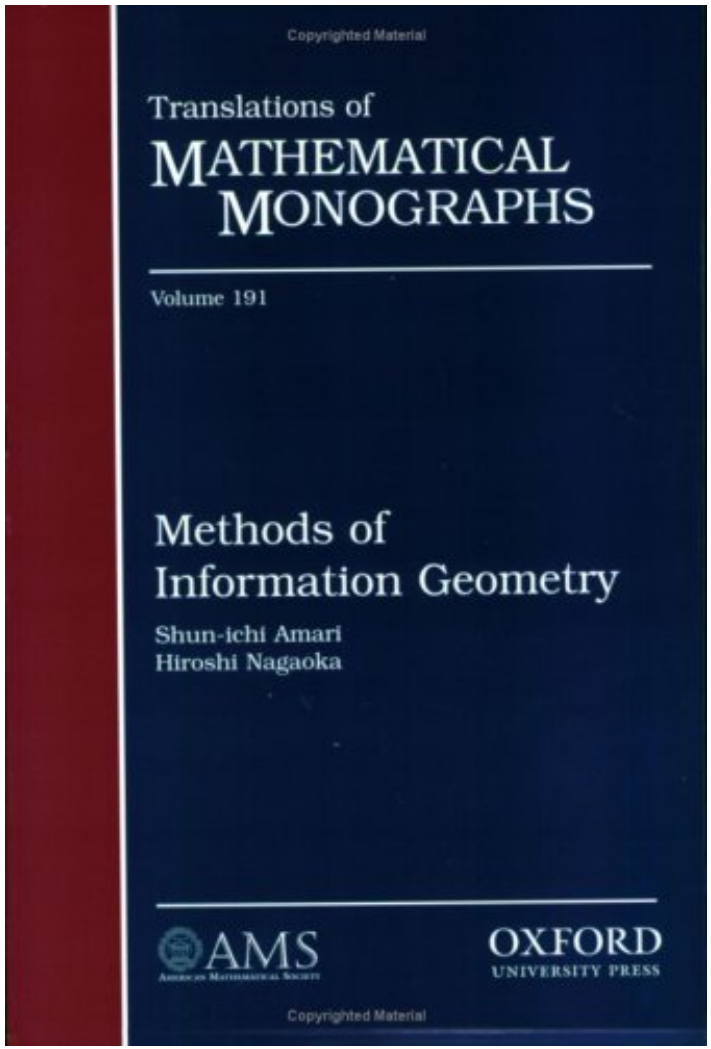
$$D_\alpha(p\,\|\,q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2}\,dx\right)$$

- $D_\alpha(p||q)$ is non-negative and 0 iff $p = q$
  - when $\alpha \to -1$ we get $KL(P||Q)$
  - when $\alpha \to +1$ we get $KL(Q||P)$
  - when $\alpha = 0$ $D_0(P||Q)$ is proportional to Hellinger's distance (metric)

$$D_H(p\,\|\,q) = \int (p(x)^{1/2} - q(x)^{1/2})^2\,dx$$

So many variational approximations must exist, one for each α!

for more on α - divergences



Shun-ichi Amari

for specific examples of $\alpha = \pm 1$

See Chapter 10

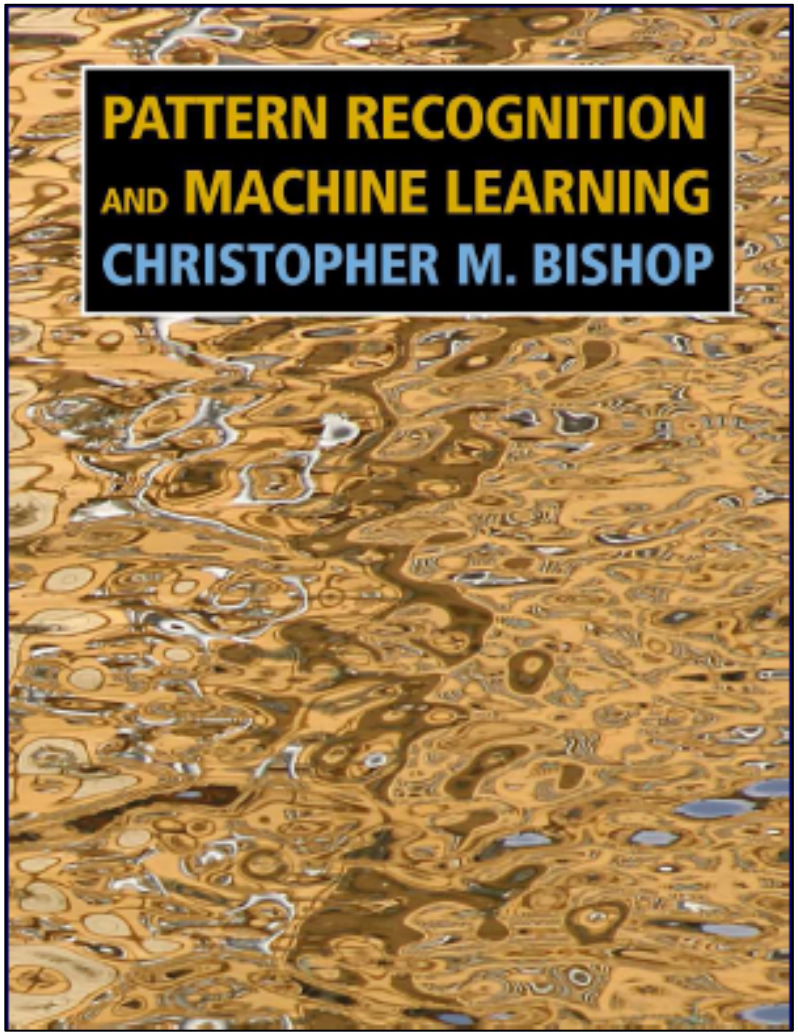
Variational Single Gaussian

Variational Linear Regression

Variational Mixture of Gaussians

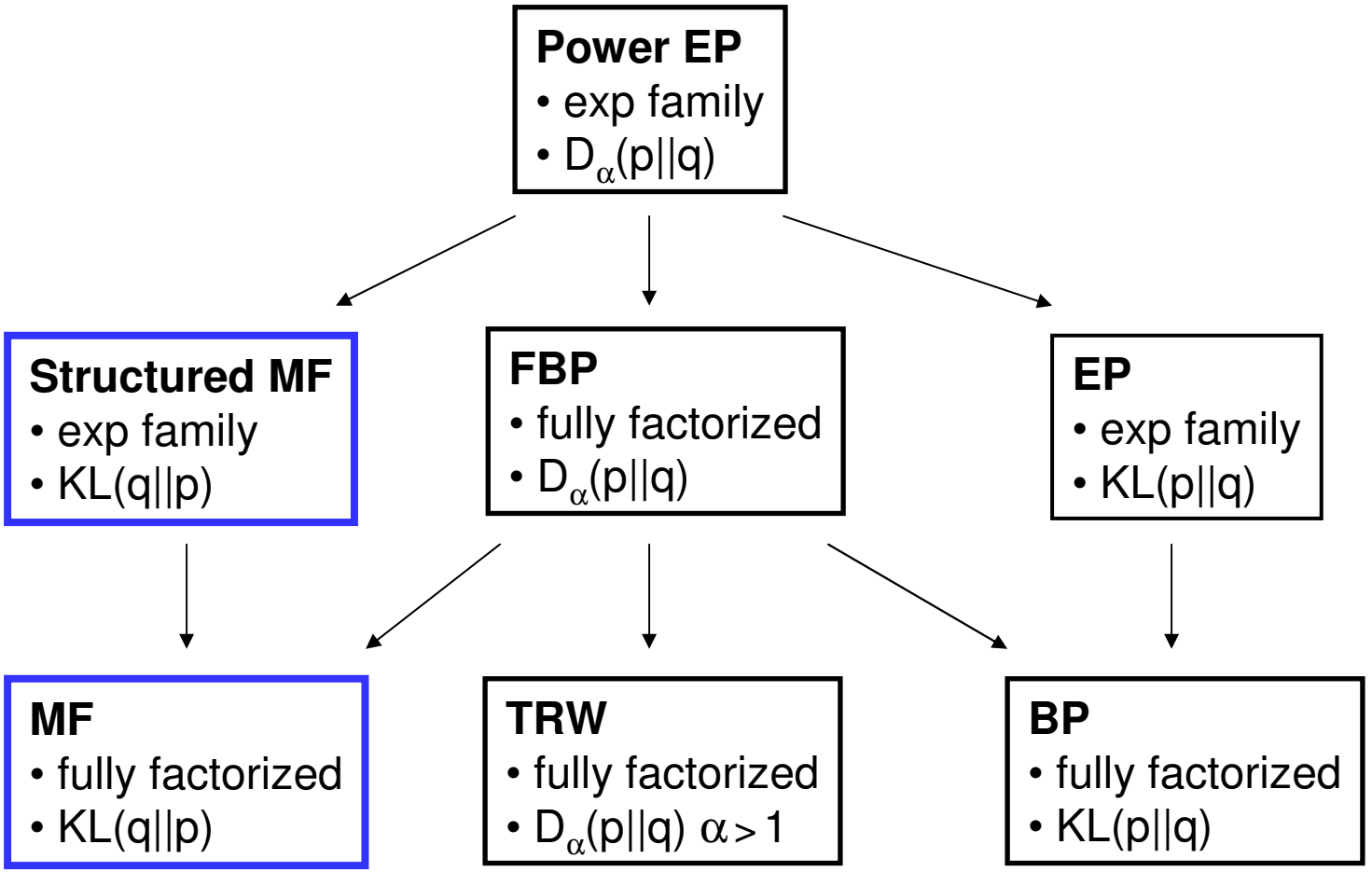
Variational Logistic Regression

Expectation Propagation ($\alpha = -1$)



Hierarchy of Algorithms

(based on α and structuring)



by *Tom Minka*

Variational MF

$$p(x) = \frac{1}{Z}\prod_c \gamma_c(x_c) = \frac{1}{Z} e^{\psi(x)} \qquad \psi(x) = \sum_c \log(\gamma_c(x_c))$$

$$\log Z = \log \int e^{\psi(x)}\, dx$$

$$= \log \int Q(x) \frac{e^{\psi(x)}}{Q(x)} dx \quad \geq \quad E_Q \log[e^{\psi(x)} / Q(x)]$$

Jensen's

$$= \sup_Q E_Q \log[e^{\psi(x)} / Q(x)]$$

$$= \sup_Q \{ E_Q[\psi(x)] + H[Q(x)] \}$$

Variational MF

$$\log Z \quad \geq \quad \sup_Q \{ E_Q[\psi(x)] + H[Q(x)] \}$$

Equality is obtained for $Q(x) = P(x)$ (all Q admissible)

Using *any* other Q yields a lower bound on $\log Z$

The slack in this bound is KL-divergence $D(Q||P)$

Goal: restrict Q to a *tractable subclass* $\mathbf{Q}$
optimize with $\sup_Q$ to tighten this bound

note we're (also) maximizing entropy $H[Q]$

Variational MF

$$\log Z \quad \geq \quad \sup_Q \{ E_Q[\psi(x)] + H[Q(x)] \}$$

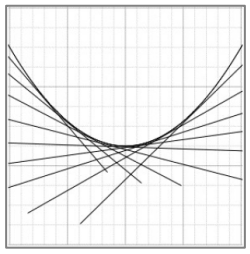
Most common specialized family :

"log-linear models" $\psi(x) = \sum_c \theta_c \phi_c(x_c) = \theta^T \phi(x)$

linear in parameters θ (natural parameters of EFs)

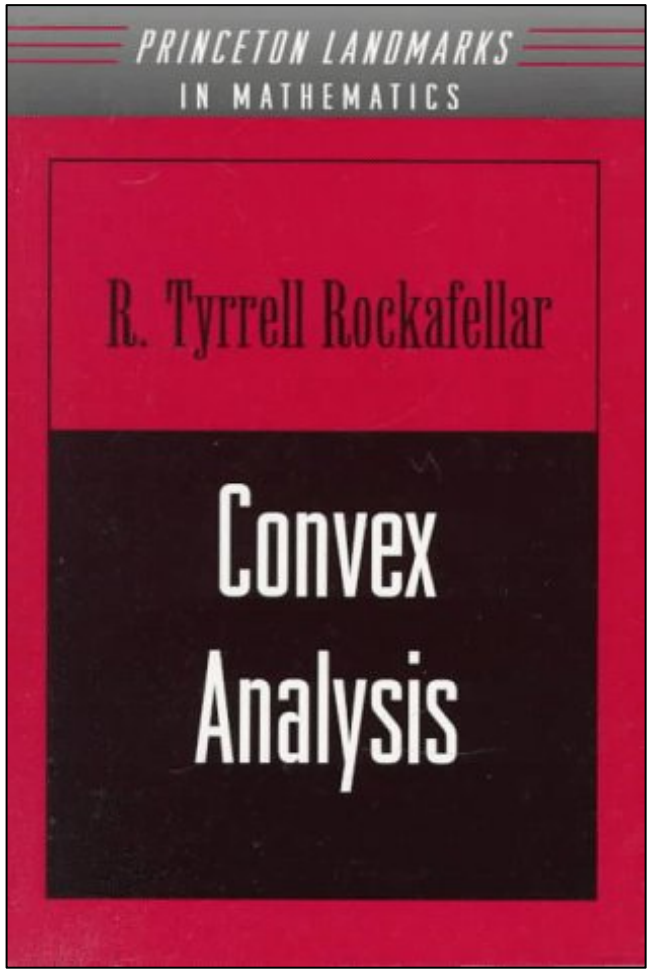
clique potentials $\phi(x)$ (sufficient statistics of EFs)

Fertile ground for plowing Convex Analysis

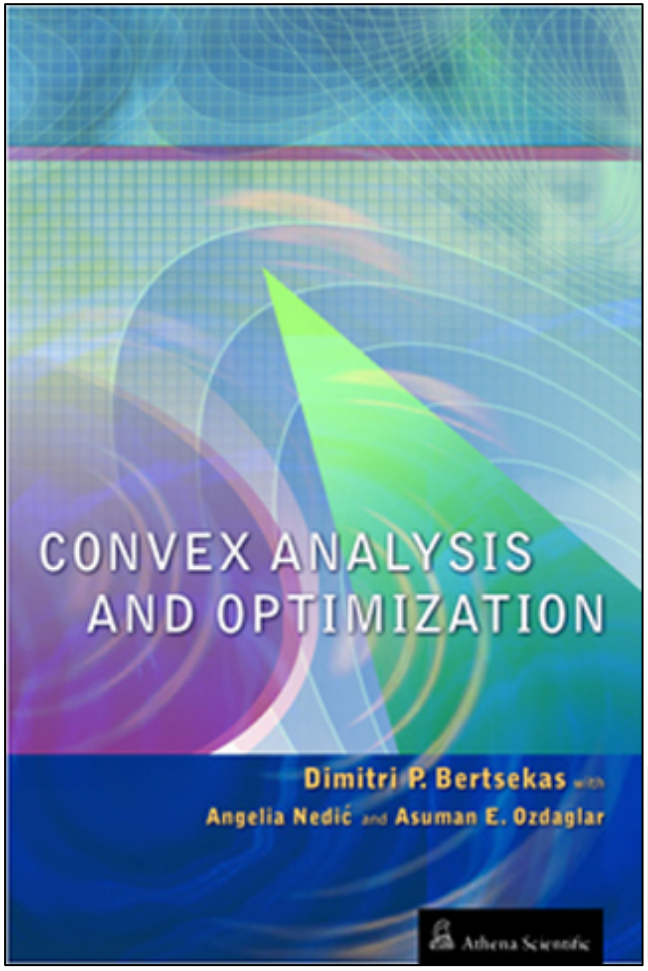


Convex Analysis

The Old Testament



The New Testament



Variational MF for EF

$$\log Z \geq \sup_Q \{ E_Q[\psi(x)] + H[Q(x)] \}$$

$$\log Z \geq \sup_Q \{ E_Q[\theta^T \phi(x)] + H[Q(x)] \}$$

$$\log Z \geq \sup_Q \{ \theta^T E_Q[\phi(x)] + H[Q(x)] \}$$

$$A(\theta) \geq \sup_{\mu \in M} \{ \theta^T \mu - A^*(\mu) \}$$

EF notation

$\boldsymbol{M}$ = set of all *moment* parameters realizable under subclass $\mathbf{Q}$

Variational MF for EF

So it looks like we are just optimizing a concave function (linear term + negative-entropy) over a convex set

Wait … that doesn't sound so hard! Yet it is **hard** **...** Why?

1. graph probability (being a *measure*) requires a very large number of **marginalization** constraints for *consistency* (leads to a typically beastly marginal polytope M in the discrete case)

e.g., a complete 7-node graph's polytope has over 10^8 facets !

In fact, optimizing just the linear term alone can be hard

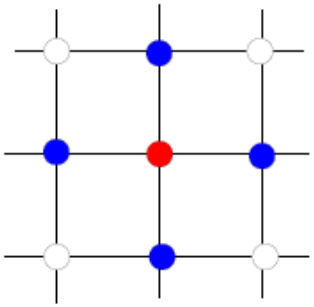
2. exact computation of **entropy** $-A^*(\mu)$ is highly non-trivial

(hence the famed Bethe & Kikuchi approximations)

Gibbs Sampling for Ising

- Binary MRF $G = (V,E)$ with pairwise clique potentials

$$p(\mathbf{x};\theta) \propto \exp\Big\{\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t\Big\}$$



1. pick a node s at random
2. sample u ~ Uniform(0,1)
3. update node s :

$$x_s^{(m+1)} = \begin{cases} 1 & \text{if } u \le \{1+\exp[-(\theta_s + \sum_{t\in\mathcal{N}(s)}\theta_{st}x_t^{(m)})]\}^{-1} \\ 0 & \text{otherwise} \end{cases}$$

4. goto step 1

a slower stochastic version of ICM

Naive MF for Ising

- use a variational mean parameter at each site $\nu_s \in (0, 1)$

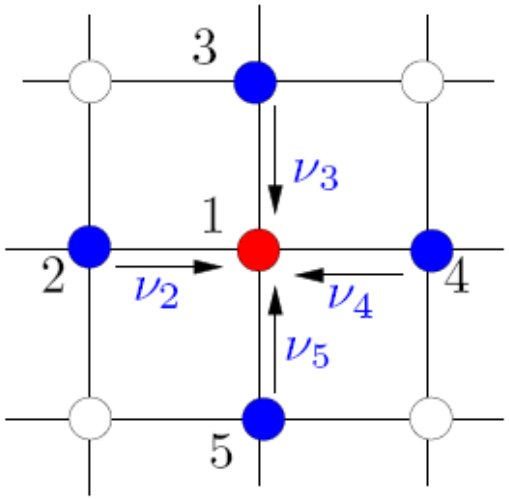
1. pick a node *s* at random
2. update its parameter :

$$\nu_s \longleftarrow \left\{1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st}\nu_t)]\right\}^{-1}$$

3. goto step 1

- deterministic "loopy" message-passing

- how well does it work? depends on θ



Graphical Models as EF

- *G*(*V*,*E*) with nodes $X_s \in \{0, 1, \ldots, m_s - 1\}$
- sufficient stats : $\mathbb{I}_j(x_s) \quad \text{for} \quad s = 1, \ldots n, \quad j \in \mathcal{X}_s$

  $\mathbb{I}_{jk}(x_s, x_t) \quad \text{for} \quad (s,t) \in E, \quad (j,k) \in \mathcal{X}_s \times \mathcal{X}_t$
- clique potentials $\theta_s(x_s) := \sum_{j \in \mathcal{X}_s} \theta_{s;j} \mathbb{I}_j(x_s)$ likewise for θ_{st}
- probability $p(\mathbf{x}; \theta) \propto \exp\{\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t)\}$
- log-partition $A(\theta) = \log \sum_{\mathbf{x} \in \mathcal{X}^n} \exp\{\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t)\}$

- mean parameters $\mu_s(x_s) := \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}_j(x_s)$,

  $\mu_{st}(x_s, x_t) := \sum_{(j,k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk} \mathbb{I}_{jk}(x_s, x_t)$

Variational Theorem for EF

- For any mean parameter μ where $\theta(\mu)$ is the corresponding natural parameter

$$A^*(\mu) = \begin{cases} -H(p_{\theta(\mu)}) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}}. \end{cases}$$

in relative interior of M

not in the closure of M

- the log-partition function has this variational representation

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \theta, \mu \rangle - A^*(\mu) \right\}$$

- this supremum is achieved at the moment-matching value of μ

$$\mu = \int_{\mathcal{X}^m} \phi(x) p_\theta(x) \nu(dx) = \mathbb{E}_\theta[\phi(X)] = \nabla A(\theta(\mu))$$

Legendre-Fenchel Duality

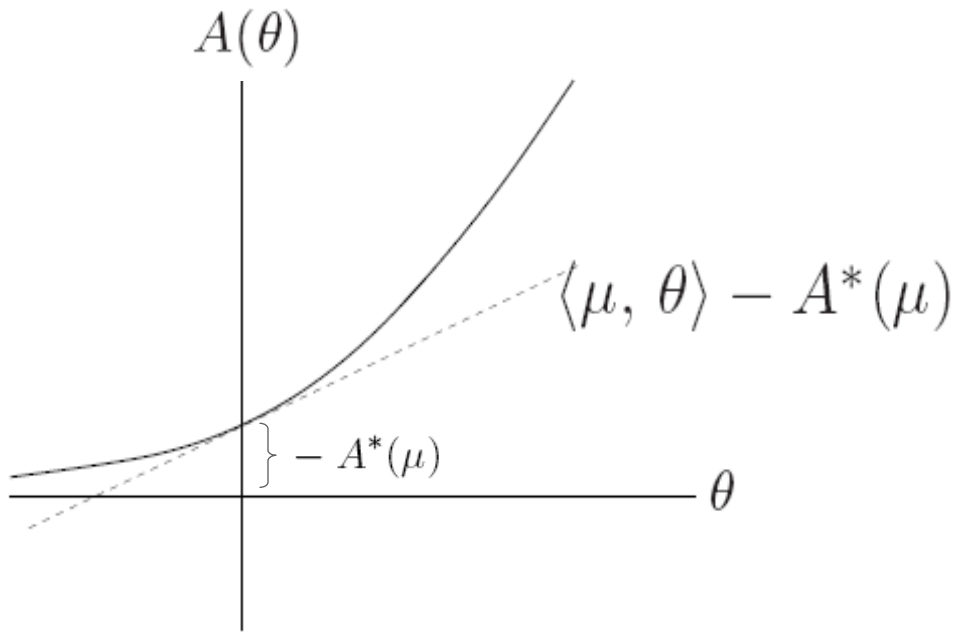
• **Main Idea**: (convex) functions can be "supported" (lower-bounded) by a continuum of lines (hyperplanes) whose intercepts create a *conjugate dual* of the original function (and vice versa)

conjugate dual of A

$$A^*(\mu) := \sup_{\theta \in \Omega} \{\langle \mu, \theta \rangle - A(\theta)\}$$

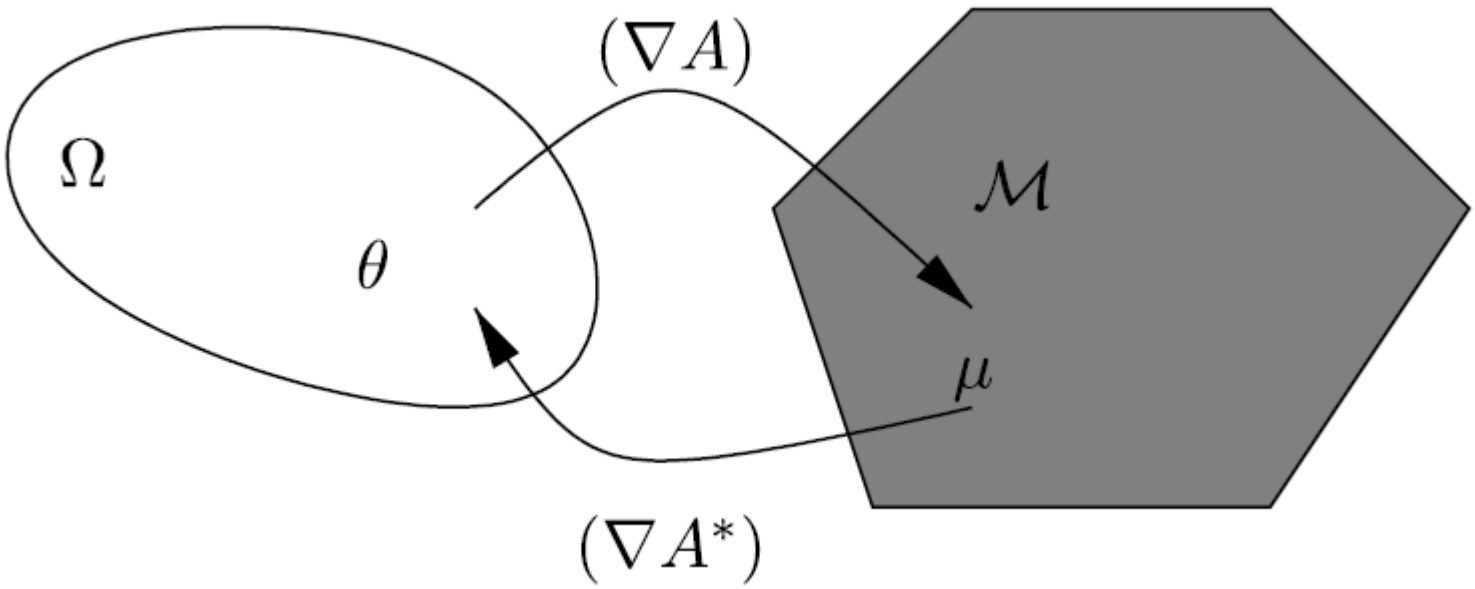
conjugate dual of $A*$

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$$



Note that $A^{**} = A$ (iff A is convex)

Dual Map for EF



Two equivalent parameterizations of the EF

Bijective mapping between Ω and the *interior* of M

Mapping is defined by the gradients of A and its dual $A*$

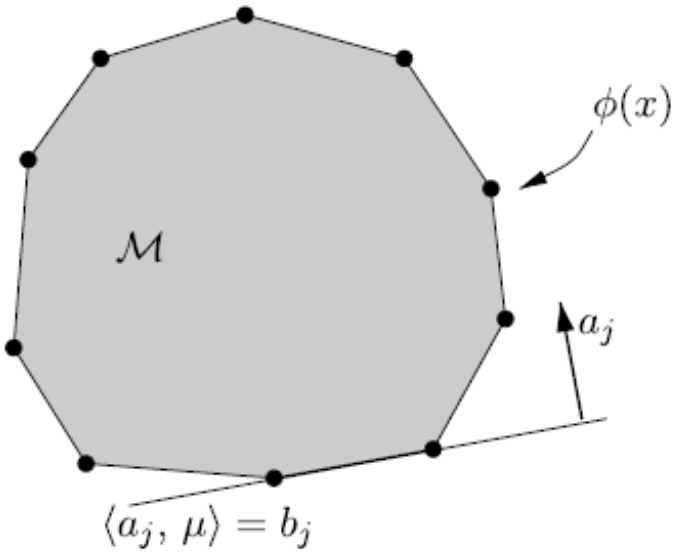
Shape & complexity of M depends on X and size and structure of G

Marginal Polytope

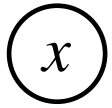
- $G(V,E)$ = graph with discrete nodes
- Then M = convex hull of all $\phi(x)$

$$\mathcal{M} := \{\mu \in \mathbb{R}^d \mid \exists\, p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$$

- equivalent to intersecting half-spaces $a^T \mu > b$
- difficult to characterize for large G
- hence difficult to optimize over
- interior of M is 1-to-1 with Ω



The Simplest Graph



- $G(V,E)$ = a single Bernoulli node $\phi(x) = x$
- density $p(x;\theta) \propto \exp\{\theta\, x\}$
- log-partition $A(\theta) = \log\left[1 + \exp(\theta)\right]$ (of course we knew this)
- we know A^* too, but let's solve for it variationally

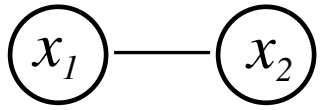
$$A^*(\mu) := \sup_{\theta \in \mathbb{R}} \{\mu\theta - \log[1 + \exp(\theta)]\}$$

- differentiate à stationary point $\mu = \exp(\theta)/[1 + \exp(\theta)]$
- rearrange to $\theta(\mu) := \log[\mu/(1-\mu)]$, substitute into A^*

$$\begin{aligned} A^*(\mu) &= \mu \log[\mu/(1-\mu)] - \log\left[1 + \frac{\mu}{1-\mu}\right] \\ &= \mu \log \mu + (1-\mu)\log(1-\mu), \end{aligned}$$

Note: we found *both* the mean parameter and the lower bound using the variational method

The 2nd Simplest Graph



- $G(V,E)$ = 2 connected Bernoulli nodes $\quad \phi(x) = \{x_1, x_2, x_1 x_2\}$
- $p(x;\theta) \propto \exp\{\theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2\}$
- moments

$$\begin{array}{rcccl} \mu_1 & = & \mathbb{E}[X_1] & = & p(x_1 = 1) \\ \mu_2 & = & \mathbb{E}[X_2] & = & p(x_2 = 1) \\ \mu_{12} & = & \mathbb{E}[X_1 X_2] & = & p(x_1 = 1, x_2 = 1) \end{array}$$

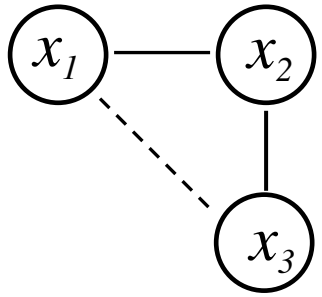
moment constraints

$$\begin{array}{rcl} \mu_1 & \geq & \mu_{12} \\ \mu_2 & \geq & \mu_{12} \\ \mu_{12} & \geq & 0 \\ 1 + \mu_{12} & \geq & \mu_1 + \mu_2 \end{array}$$

- $$\begin{aligned} A^*(\mu) &= -H\left(p(x;\mu)\right) \\ &= \sum_{x_1,x_2} p(x;\mu) \log p(x;\mu) \\ &= \mu_{12} \log \mu_{12} + (\mu_1 - \mu_{12}) \log (\mu_1 - \mu_{12}) + (\mu_2 - \mu_{12}) \log (\mu_2 - \mu_{12}) \\ &\quad + (1 + \mu_{12} - \mu_1 - \mu_2) \log (1 + \mu_{12} - \mu_1 - \mu_2) \end{aligned}$$

- variational problem $\quad A(\theta) = \max\{\theta_1 \mu_1 + \theta_2 \mu_2 + \theta_{12} \mu_{12} - A^*(\mu)\}$
- solve (it's still easy!) $\quad \hat{\mu}_1(\theta) = \dfrac{\exp\{\theta_1\} + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}{1 + \exp\{\theta_1\} + \exp\{\theta_2\} + \exp\{\theta_1 + \theta_2 + \theta_{12}\}}$

The 3rd Simplest Graph



3 nodes → 16 constraints

# of constraints blows up real fast:

7 nodes → 200,000,000+ constraints

hard to keep track of valid μ's

(*i.e.*, the full shape and extent of M)

no more checking our results against closed-forms expressions that we already knew in advance!

unless G remains a tree, entropy $A*$ will not decompose nicely, *etc*

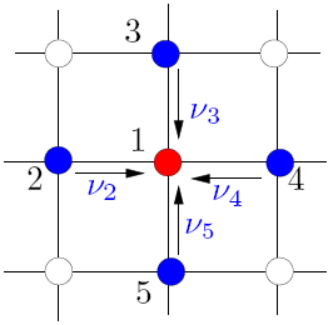
Variational MF for Ising

- tractable subgraph $H = (V, \emptyset)$
- fully-factored distribution $p(\mathbf{x};\theta) = \prod_{s \in V} p(x_s;\theta_s)$
- moment space $\mathcal{M}_{tr}(G;H) = \{\mu \mid \mu_{st} = \mu_s \mu_t, \ \mu_s \in [0,1]\}$
- entropy is *additive* : $-\sum_{s \in V} \mu_s \log \mu_s + (1-\mu_s)\log(1-\mu_s)]$
- variational problem for $A(\theta)$

$$\max_{\mu_s \in [0,1]} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - [\sum_{s \in V} \mu_s \log \mu_s + (1-\mu_s)\log(1-\mu_s)] \right\}$$

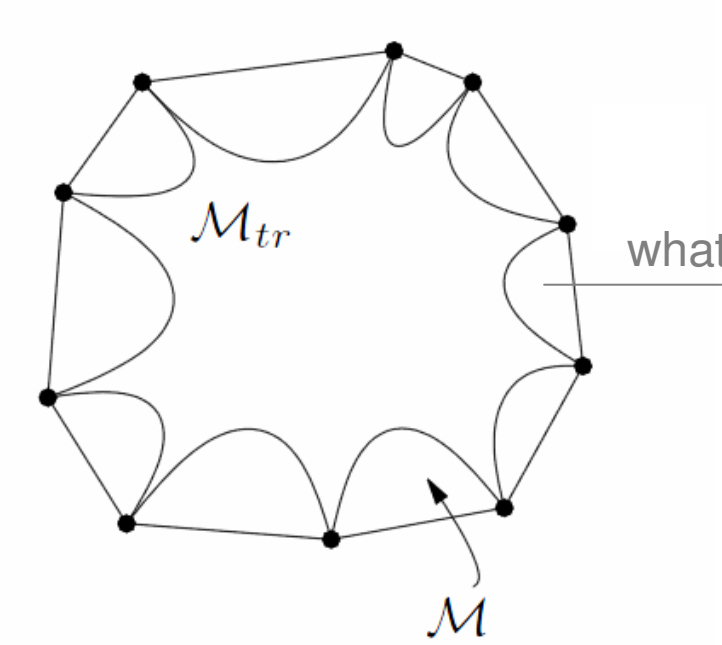
- using coordinate ascent :

$$\mu_s \longleftarrow \{1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t)]\}^{-1}$$



Variational MF for Ising

- M_{tr} is a *non-convex* inner approximation $M_{tr} \subset M$



what causes this funky curvature?

$$\mathcal{M}_{tr}(G;H) = \{\mu \mid \boxed{\mu_{st} = \mu_s \mu_t,} \;\; \mu_s \in [0,1] \}$$

- optimizing over M_{tr} must then yield a *lower bound*

$$A(\theta) \;\geq\; \sup_{\tilde{\mu} \in \mathcal{M}_{tr}} \left\{ \langle \theta, \tilde{\mu} \rangle - A^*(\tilde{\mu}) \right\}$$

Factorization with Trees

- suppose we have a tree $G = (V,T)$
- useful factorization for trees

$$p(\mathbf{x};\theta) \quad = \quad \prod_{s\in V} \mu_s(x_s) \quad \prod_{(s,t)\in E} \frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

- entropy becomes

$$-A^*(\mu) = \mathbb{E}_\mu[-\log p_\mu(X)] \quad = \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st})$$

- singleton terms $\quad H_s(\mu_s) := -\sum_{x_s\in\mathcal{X}_s} \mu_s(x_s)\log\mu_s(x_s)$

- pairwise terms $\quad I_{st}(\mu_{st}) := \sum_{(x_s,x_t)\in\mathcal{X}_s\times\mathcal{X}_t} \mu_{st}(x_s,x_t)\log\frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\,\mu_t(x_t)}$

  Mutual Information

Variational MF for Loopy Graphs

- pretend entropy factorizes like a tree (Bethe approximation)
- define *pseudo* marginals $\{\tau_s, s \in V\} \quad \{\tau_{st}, (s,t) \in E\}$

must impose these normalization and marginalization constraints

$$\begin{cases} \sum_{x_s} \tau_s(x_s) = 1 \\ \sum_{x'_t} \tau_{st}(x_s, x'_t) = \tau_s(x_s), \quad \forall\, x_s \in \mathcal{X}_s, \\ \sum_{x'_s} \tau_{st}(x'_s, x_t) = \tau_t(x_t), \quad \forall\, x_t \in \mathcal{X}_t. \end{cases}$$

- define local polytope $L(G)$ obeying these constraints
- note that $M(G) \subseteq L(G)$ for *any* G

with equality *only* for trees : $M(G) = L(G)$

Variational MF for Loopy Graphs

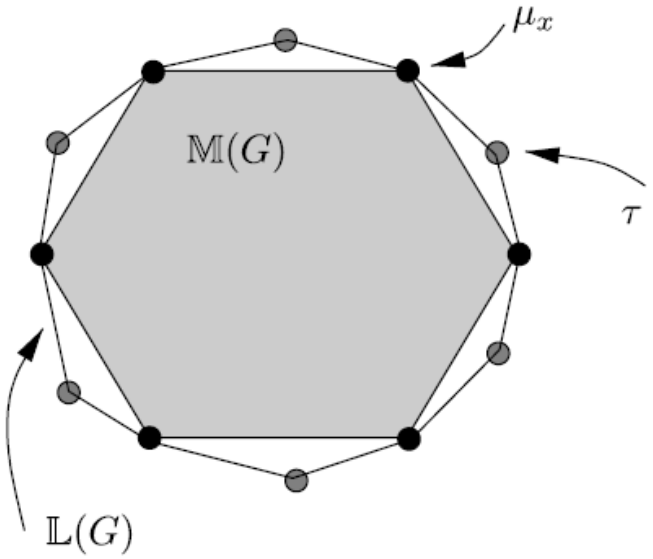
$L(G)$ is an *outer* polyhedral approximation

solving this Bethe Variational Problem we get the LBP eqs !

$$\max_{\tau \in \mathrm{LOCAL}(G)} \left\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \right\}$$

so fixed points of LBP are the
stationary points of the BVP

$$-A^*_{Bethe}(\mu) \quad = \quad \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$$



this not only illuminates what was originally an educated “hack” (LBP)
but suggests new convergence conditions and improved algorithms (TRW)

see ICML'2008 Tutorial

Graphical models and variational methods:
Message-passing, convex relaxations, and all that

Martin Wainwright

Department of Statistics, and

Department of Electrical Engineering and Computer Science,

UC Berkeley, Berkeley, CA USA

Email: wainwrig@{stat,eecs}.berkeley.edu

For further information (tutorial slides, papers, course lectures), see:

www.eecs.berkeley.edu/~wainwrig/

Summary

- SMF can also be cast in terms of “Free Energy” *etc*
- Tightening the var bound = min KL divergence
- Other schemes (*e.g*, “Variational Bayes”) = SMF
  - with additional conditioning (hidden, visible, parameter)
- Solving variational problem gives both μ and $A(\theta)$
- Helps to see problems through lens of Var Analysis

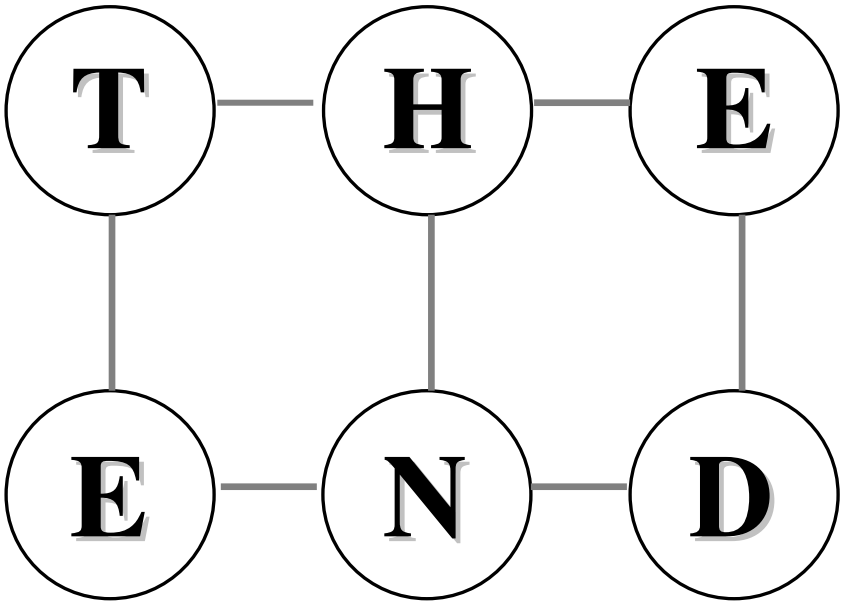
Matrix of Inference Methods

Exact Deterministic approximation Stochastic approximation

	Chain (online)	Low treewidth	High treewidth
Discrete	BP = forwards Boyen-Koller (ADF), beam search	VarElim, Jtree, recursive conditioning	Loopy BP, mean field, structured variational, EP, graph-cuts Gibbs
Gaussian	BP = Kalman filter	Jtree = sparse linear algebra	Loopy BP Gibbs
Other	EKF, UKF, moment matching (ADF) Particle filter	EP, EM, VB, NBP, Gibbs	EP, variational EM, VB, NBP, Gibbs

BP = Belief Propagation, EP = Expectation Propagation, ADF = Assumed Density Filtering, EKF = Extended Kalman Filter, UKF = unscented Kalman filter, VarElim = Variable Elimination, Jtree= Junction Tree, EM = Expectation Maximization, VB = Variational Bayes, NBP = Non-parametric BP

by *Kevin Murphy*


T
H
E
E
N
D