Approximate Inference

Part 2 of 2

Tom Minka

Microsoft Research, Cambridge, UK

Machine Learning Summer School 2009

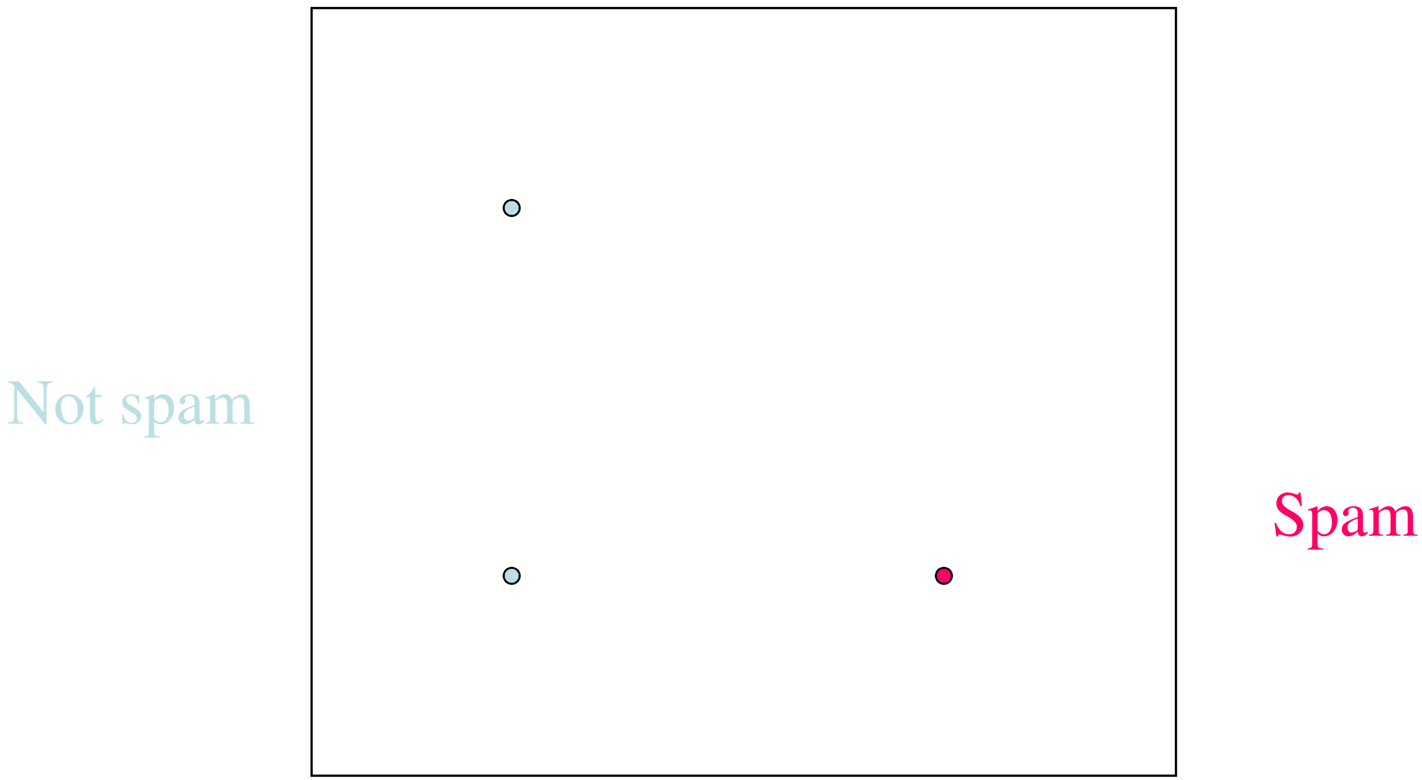
http://mlg.eng.cam.ac.uk/mlss09/

Expectation Propagation

- Fits an exponential-family approximation to the posterior
- Belief propagation is a special case
- Kalman filtering is a special case
- Does not always converge
  - May get stuck due to improper distributions (negative variances)
  - May oscillate due to loopy graph

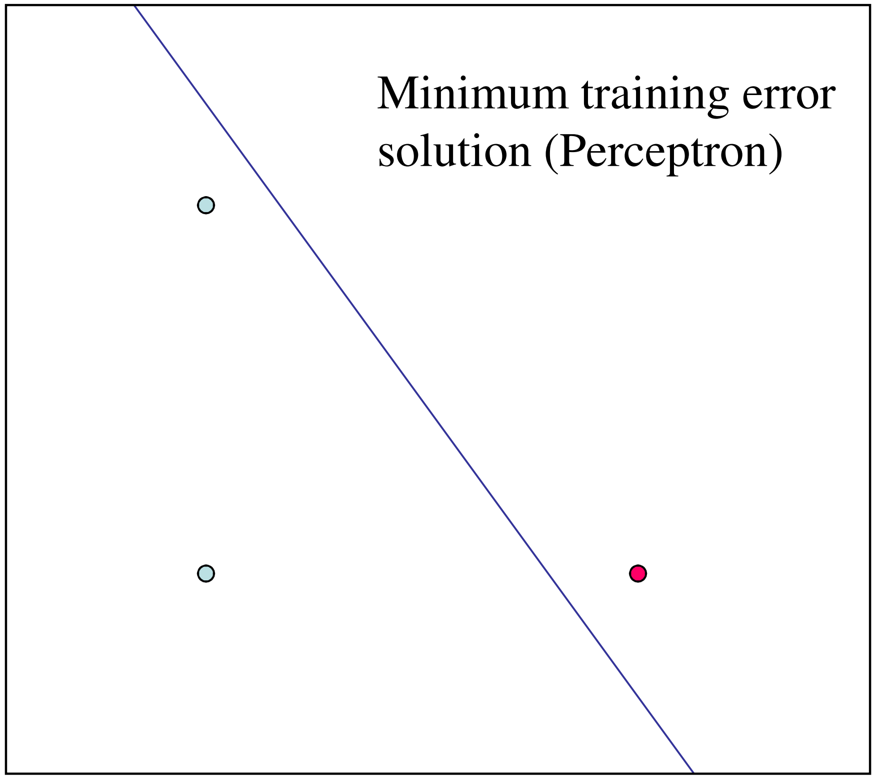
Classification problems

Spam filtering by linear separation



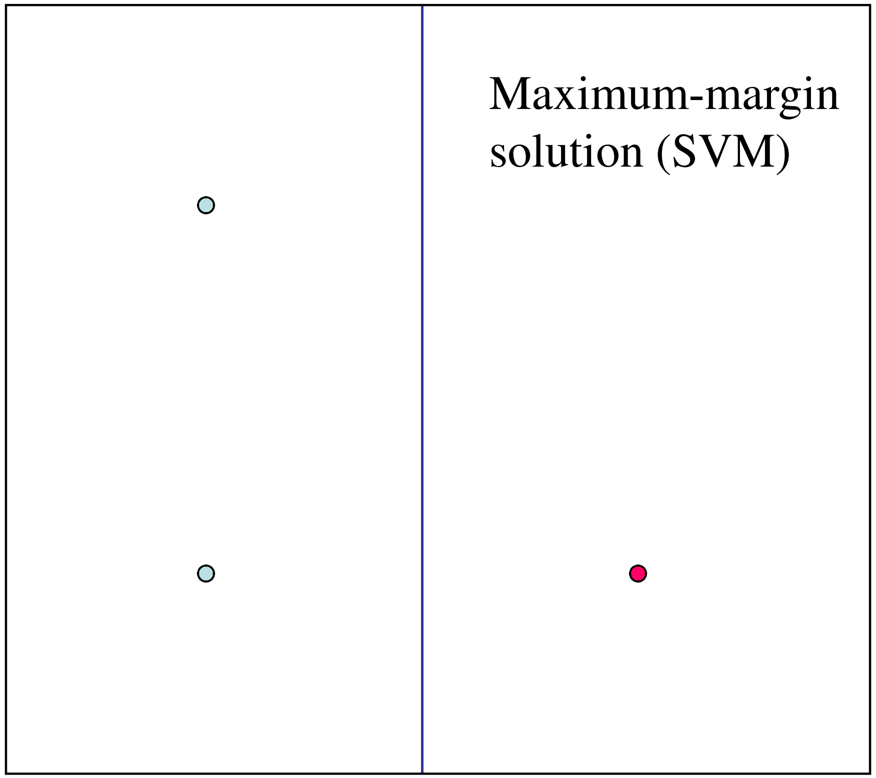
Choose a boundary that will generalize to new data

Linear separation



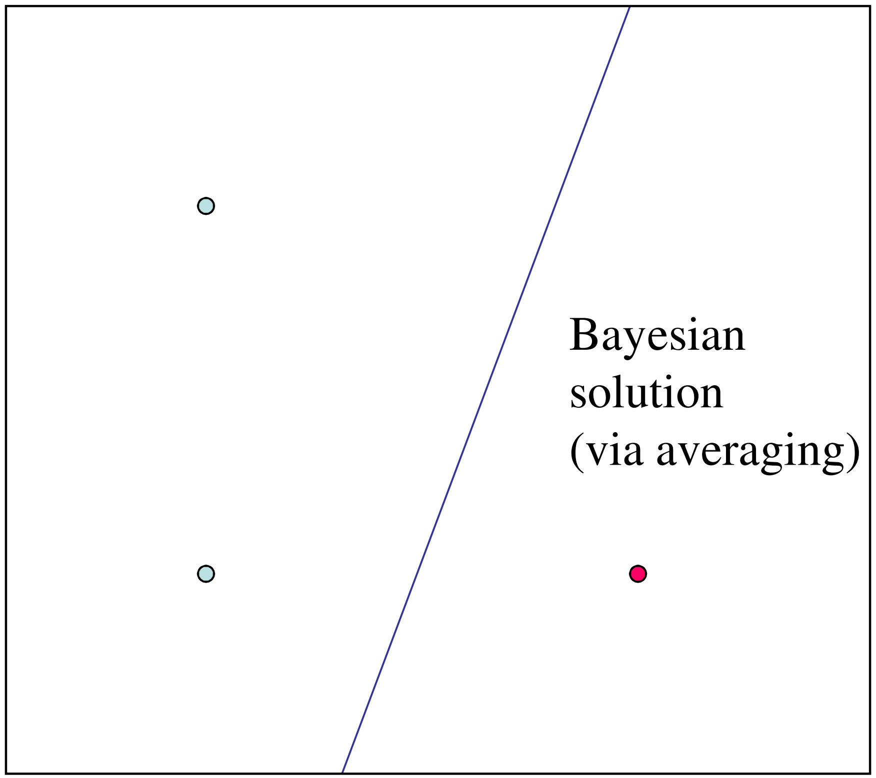
Too arbitrary – won't generalize well

Linear separation



Ignores information in the vertical direction

Linear separation



Has a margin, and uses information in all dimensions

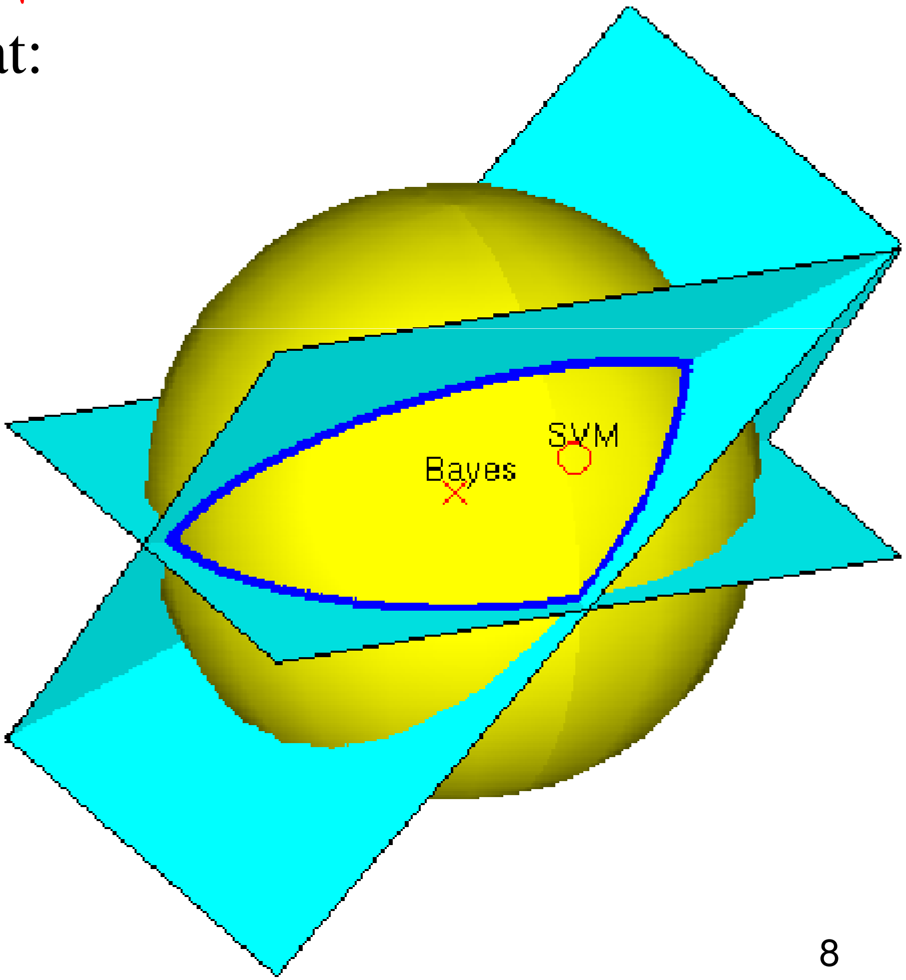
Geometry of linear separation

$p(w) \prod_i p(y_i \mid w, x_i) \propto p(w \mid y_1 \ldots y_n)$

Separator is any vector w such that:

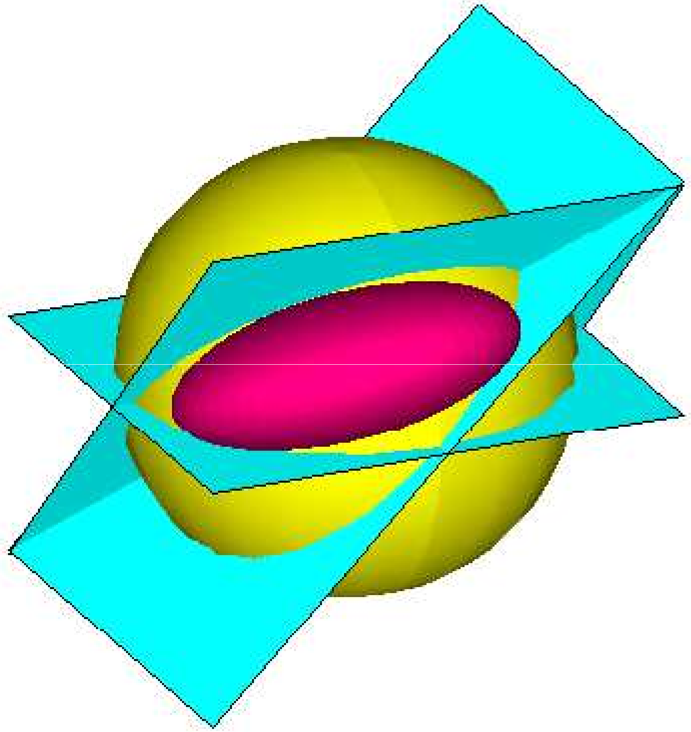
$$\mathbf{w}^T \mathbf{x}_i > 0 \quad \text{(class 1)}$$
$$\mathbf{w}^T \mathbf{x}_i < 0 \quad \text{(class 2)}$$
$$\|\mathbf{w}\| = 1 \quad \text{(sphere)}$$

$\int_w p(y \mid w, x)\, p(w \mid D)\, dw$



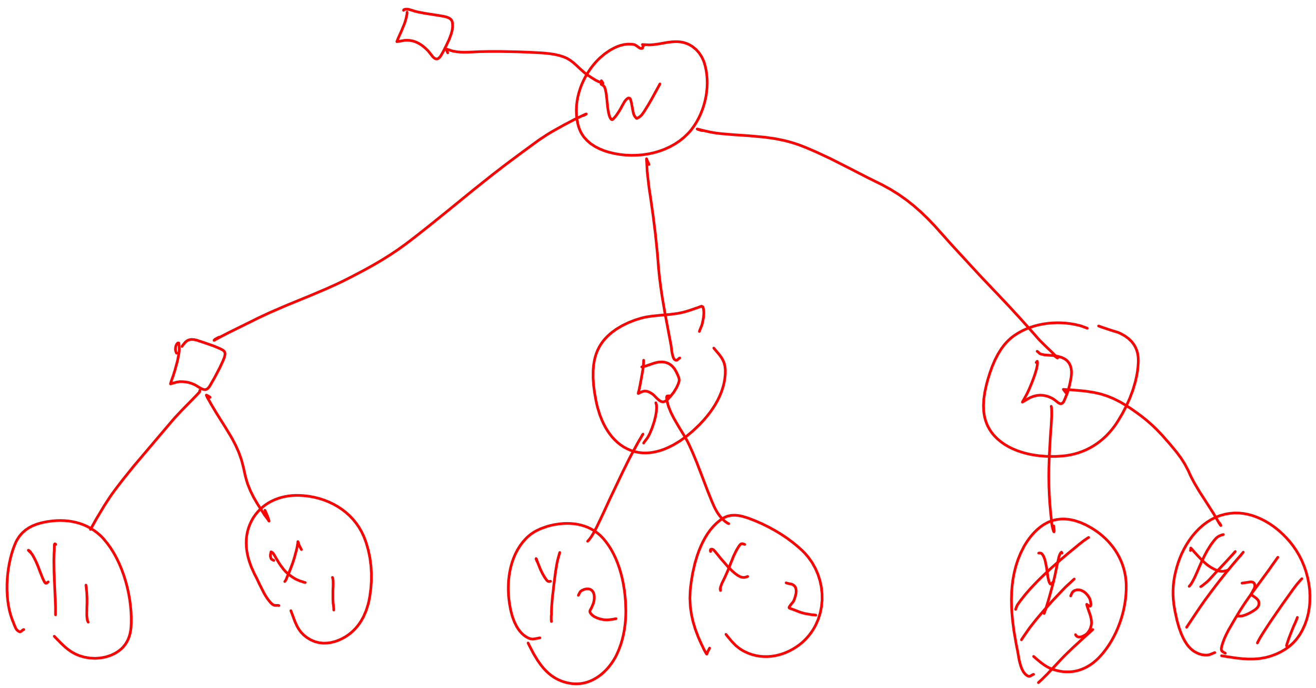
This set has an unusual shape
SVM: Optimize over it
Bayes: Average over it

Performance on linear separation



EP Gaussian approximation to posterior

Factor graph



$$p(y_i = \pm 1 \mid \mathbf{x}_i, \mathbf{w}) = I(y_i \mathbf{x}_i^\top \mathbf{w} > 0)$$

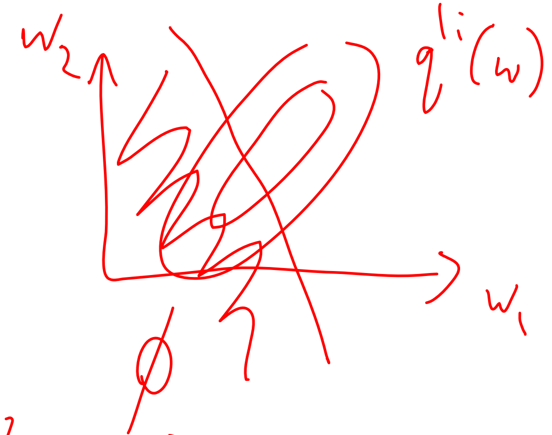
$$p(\mathbf{w}) = N(\mathbf{w}; \mathbf{0}, \mathbf{I})$$

Computing moments

$$p(y_i = \pm 1 \mid \mathbf{x}_i, \mathbf{w}) = I(y_i \mathbf{x}_i^\top \mathbf{w} > 0) = f_i(w)$$

$$q^{\setminus i}(\mathbf{w}) = N(\mathbf{w}; \mathbf{m}^{\setminus i}, \mathbf{V}^{\setminus i})$$

$$\tilde{f}_i(w) = \frac{\text{proj}[f_i(w)\, q^{\setminus i}(w)]}{q^{\setminus i}(w)}$$



$$Z(m^{\setminus i}, V^{\setminus i}) = \int_w f(w)\, q^{\setminus i}(w)\, dw = p(u > 0)$$

$$= \int_w I(\underbrace{y_i x_i^T w} > 0)\, N(w; m^{\setminus i}, V^{\setminus i})\, dw$$

$$u \sim N(y_i x_i^T m^{\setminus i}, x_i^T V^{\setminus i} x_i)$$

$$\phi\left(\frac{y_i x_i^T m^{\setminus i}}{}\right.$$

Computing moments

$$p(u>0) = \Phi\left(\frac{E[u]}{\sqrt{var(u)}}\right) = \Phi\left(\frac{y_i x_i^T m^{\backslash i}}{\sqrt{x_i^T V^{\backslash i} x_i}}\right)$$

Time vs. accuracy

A typical run on the 3-point problem

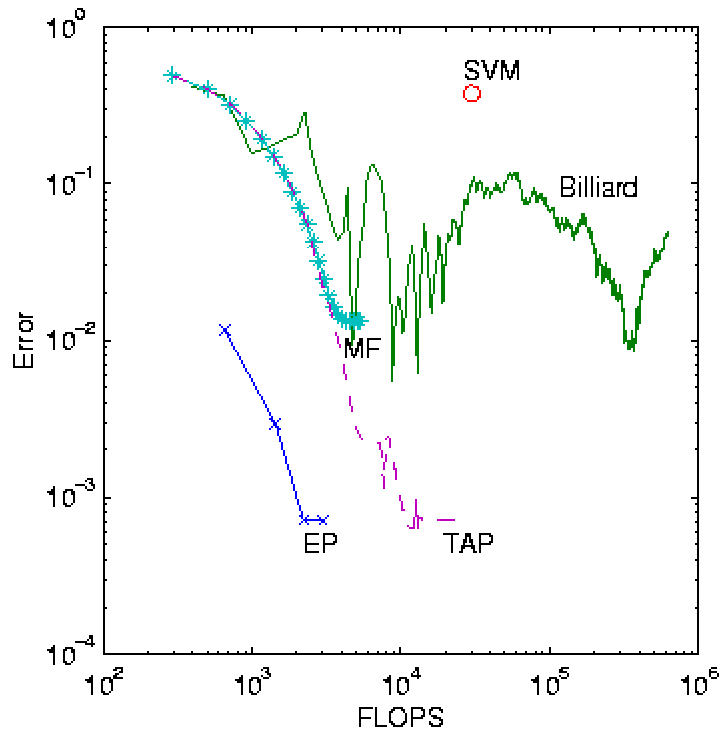
Error = distance to true mean of w

Billiard = Monte Carlo sampling
(Herbrich et al, 2001)

Opper&Winther's algorithms:

MF = mean-field theory

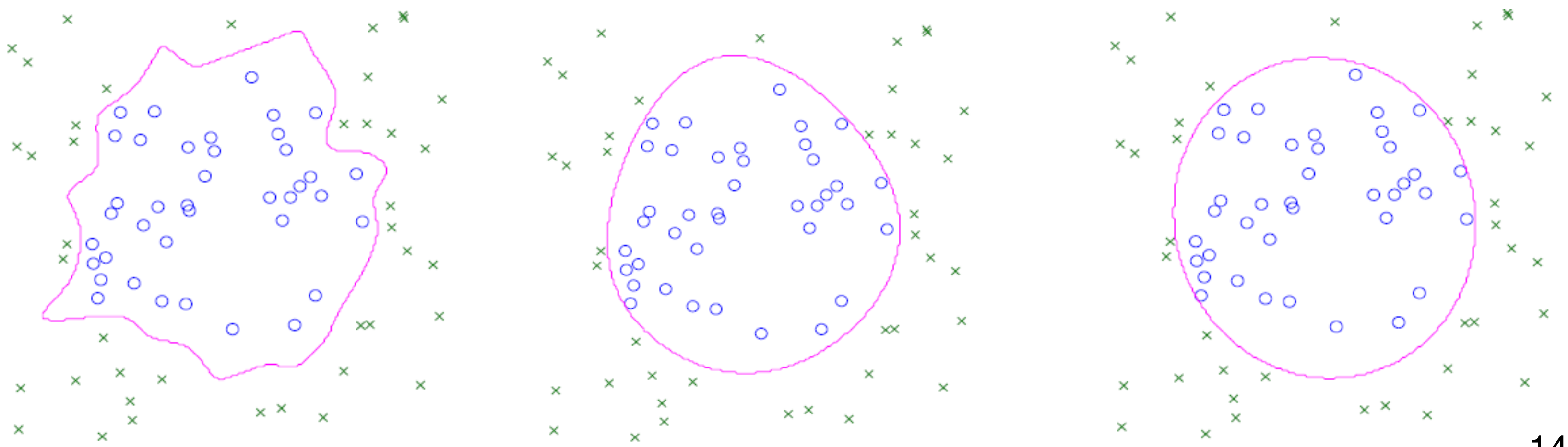
TAP = cavity method (equiv to Gaussian EP for this problem)



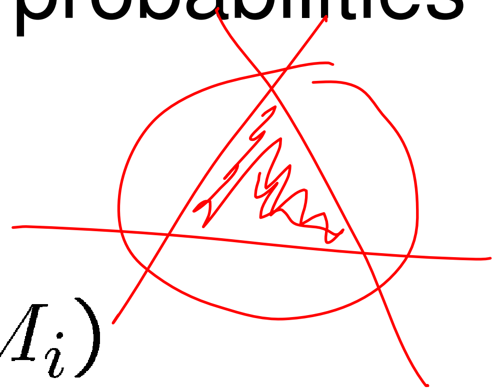
Gaussian kernels

- Map data into high-dimensional space so that

$$\phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$$



Bayesian model comparison



- Multiple models M_i with prior probabilities $p(M_i)$
- Posterior probabilities:

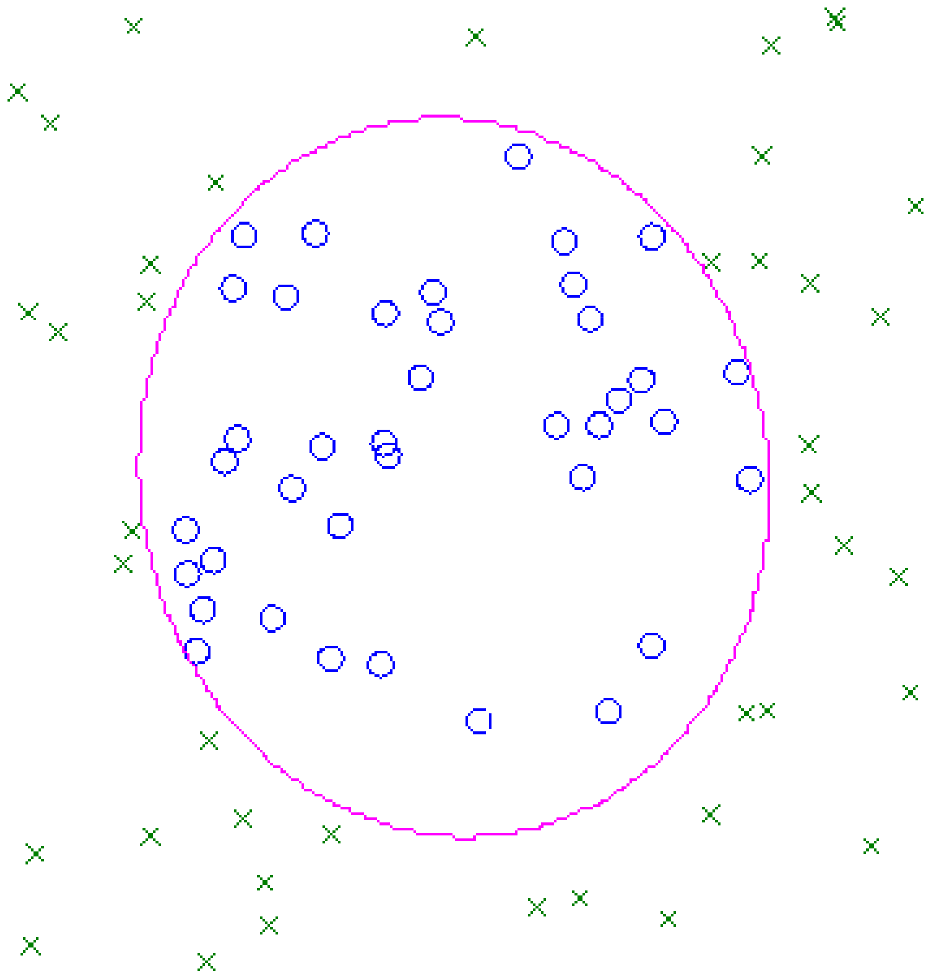
$$p(M_i|D) \propto p(D|M_i)p(M_i)$$

- For equal priors, models are compared using model evidence:

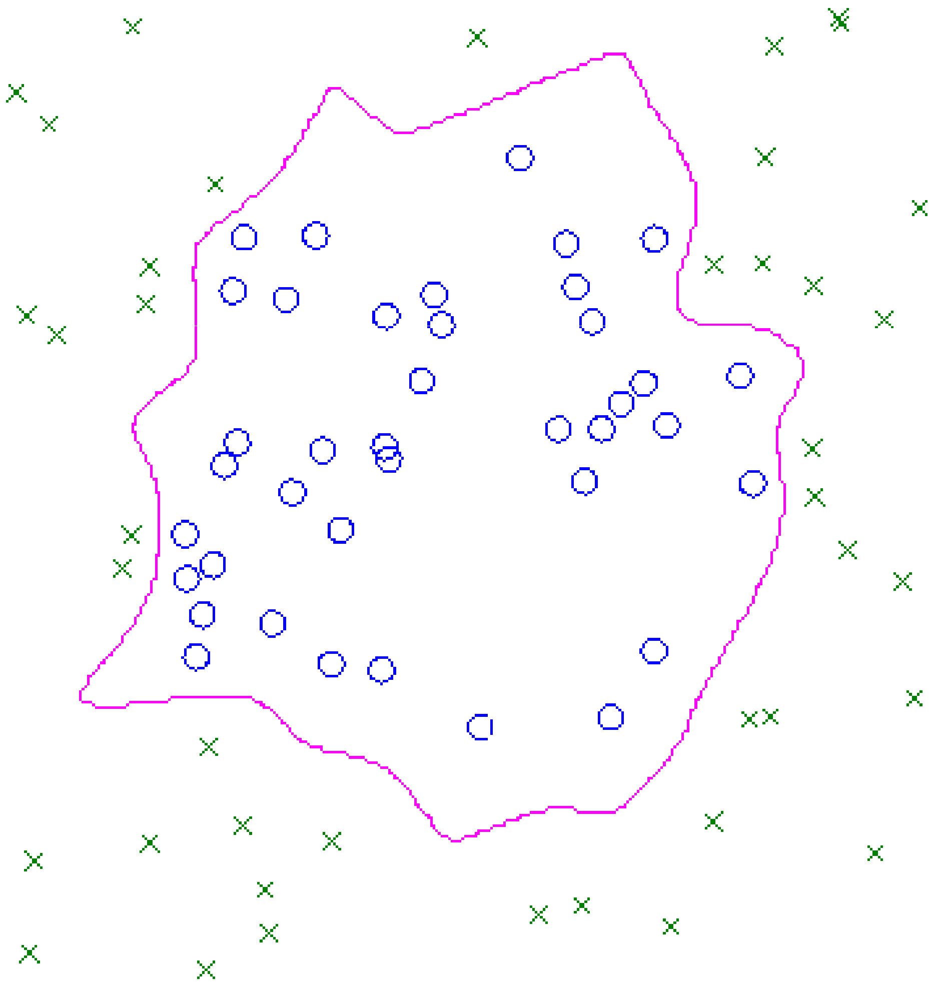
$$p(D|M_i) = \int_\theta p(D, \theta|M_i)d\theta$$

$$p(D) = \int_w p(w) \prod_i I(y_i x_i^T w > 0)\, dw$$

Highest-probability kernel



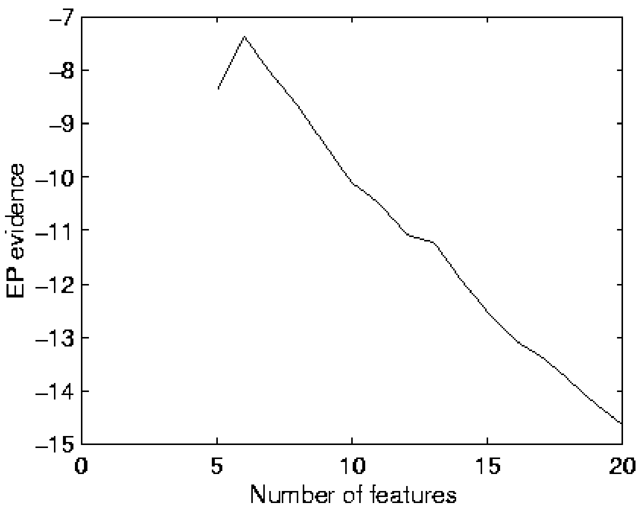
Margin-maximizing kernel



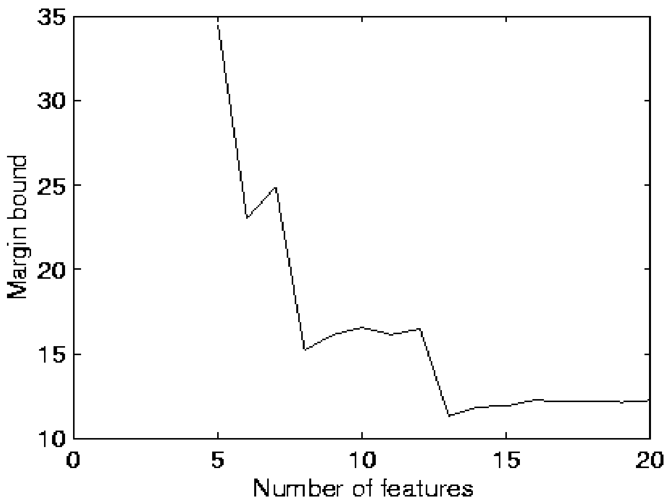
Bayesian feature selection

Synthetic data where 6 features are relevant (out of 20)

Bayes picks 6



Margin picks 13



Further reading

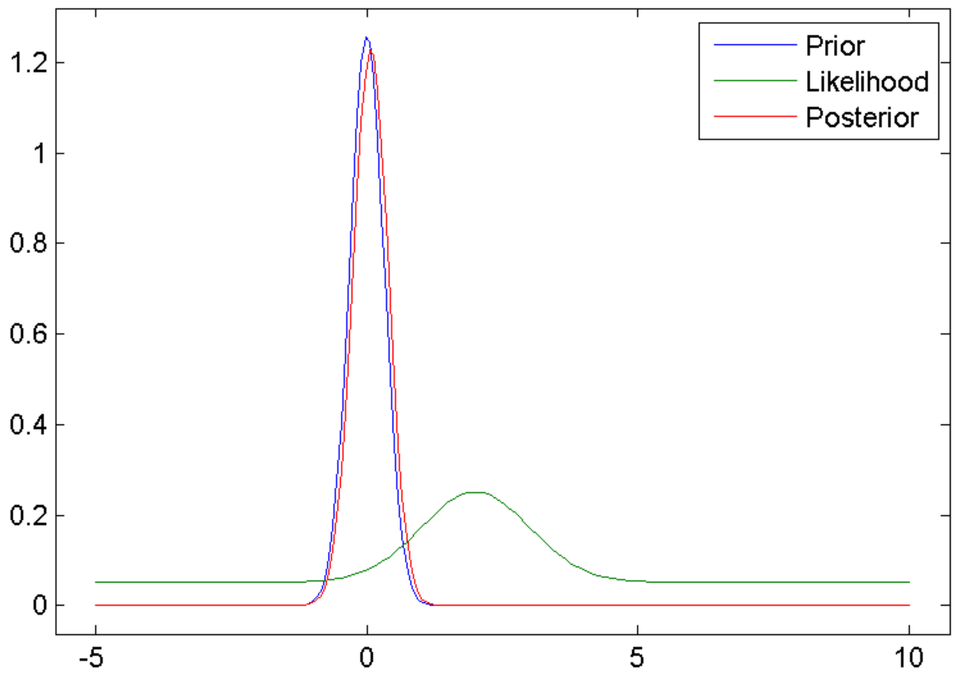
- "Divergence measures and message passing" http://research.microsoft.com/~minka/papers/
- EP bibliography
  http://research.microsoft.com/~minka/papers/ep/roadmap.html
- EP quick reference
  http://research.microsoft.com/~minka/papers/ep/minka-ep-quickref.pdf

How negative variances arise in EP

$$p(x) = N(x; 0, 0.1) \qquad f(x) = p(y = 2 \mid x)$$

$$p(y \mid x) = 0.5N(y; x, 1) + 0.5N(y; 0, 10)$$

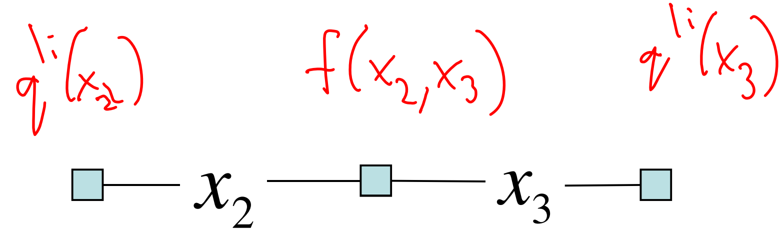
$$\text{proj}[f(x)p(x)] = N(x; 0.068, 0.104)$$



$$\tilde{f}(x) = \frac{\text{proj}[f(x)p(x)]}{\cancel{f(x)}\, p(x)}$$

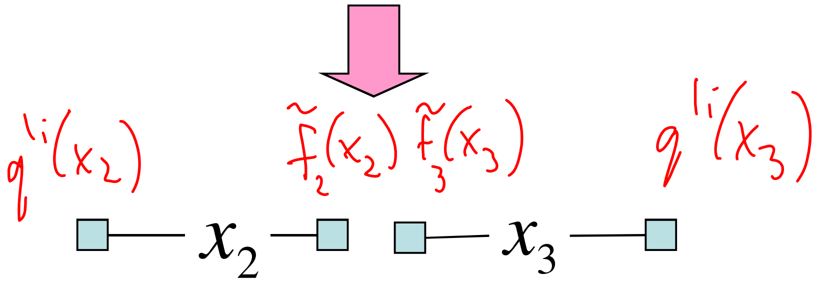
$\rightarrow$ negative variance

Belief propagation



Special case when marginal distribution already has the correct form (no projection)

only approx is factorization of posterior



$$\tilde{f}_2(x_2) = \frac{\mathrm{proj}\left[q^{\setminus i}(x_2) \int q^{\setminus i}(x_3) f(x_2, x_3)\, dx_3 \right]}{q^{\setminus i}(x_2)}$$

$$\tilde{f}_2(x_2) = \int q^{\setminus i}(x_3) f(x_2, x_3)\, dx_3$$

Divergence minimization

- For exponential families, moment matching step can be interpreted as minimizing KL divergence

$$q(x) = \text{proj}[p(x)] \longleftrightarrow q(x) = \arg\min KL(p(x) \,\|\, q(x))$$

$$\min KL(p \,\|\, q) = \int p(x) \log \frac{p(x)}{q(x)} dx = \min -\int p(x) \log q(x)\, dx$$

$$q(x) = \frac{\exp(\theta^T \phi(x))}{\int_X \exp(\theta^T \phi(x)) dx} \qquad = -\int p(x)\, \theta^T \phi(x)\, dx + \log \int_X \exp(\theta^T \phi(x)) dx$$

$$\frac{d}{d\theta} = -\int p(x) \phi(x) dx + \int_X \phi(x) q(x) dx = 0$$

Global divergence to local divergence

$p(x) = \prod_a f_a(x)$ $q(x) = \prod_a \tilde{f}_a(x)$

- Global divergence:

$$D(p(x) \mid\mid q(x)) =$$

$$D(f_a(x) \prod_{b \neq a} f_b(x) \mid\mid \tilde{f}_a(x) \prod_{b \neq a} \tilde{f}_b(x))$$

- Local divergence:

$$D(f_a(x) \prod_{b \neq a} \tilde{f}_b(x) \mid\mid \tilde{f}_a(x) \prod_{b \neq a} \tilde{f}_b(x))$$

Fixed points of EP

- Fixed points of EP are the stationary points of the EP model evidence:

$$\underline{\tilde{Z}(\tilde{f}_1,\ldots,\tilde{f}_n)} = \underline{\left(\int_{\mathbf{x}} q(\mathbf{x})d\mathbf{x}\right)^{1-n}} \underline{\prod_{i=1}^{n}\int_{\mathbf{x}} \frac{f_i(\mathbf{x})}{\tilde{f}_i(\mathbf{x})} q(\mathbf{x})d\mathbf{x}}$$

$$\times \int f_i \, q^{\setminus i}(x) \, dx$$

where $\underline{q(\mathbf{x}) = \prod_{i=1}^{n} \tilde{f}_i(\mathbf{x})}$

("Divergence measures and message passing", Minka, 2005)

Other divergences

- Same recipe can be used to minimize other divergence measures
- Minimizing KL(q||p) leads to mean-field approximation
- Minimizing alpha-divergence leads to tree-reweighted belief propagation and power EP

Special property of KL(q||p)

- Minimizing local divergence is equivalent to minimizing global divergence

$q(x) = \prod_a \tilde{f}_a(x)$

global $KL(q(x) \| p(x)) = \int_x q(x) \log \frac{q(x)}{p(x)} dx$

$p(x) = \prod_a f_a(x)$

local $KL(\tilde{f}_a(x) \prod_{b \neq a} \tilde{f}_b(x) \| f_a(x) \prod_{b \neq a} \tilde{f}_b(x)) =$

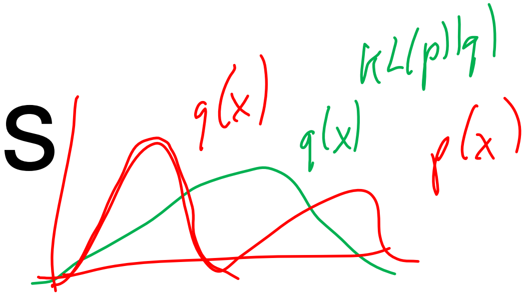
$\int_x q(x) \log \frac{\tilde{f}_a(x) \prod_{b \neq a} \tilde{f}_b(x)}{f_a(x) \prod_{b \neq a} \tilde{f}_b(x)} dx = \int_x q(x) \log \frac{\tilde{f}_a(x)}{f_a(x)}$

$\sum_i KL(q(x) \| f_a(x) \prod_{b \neq a} \tilde{f}_b(x)) \simeq KL(q(x) \| p(x))$

$$\int q(x) \log \frac{q(x)}{p(x)}$$

KL(p||q)

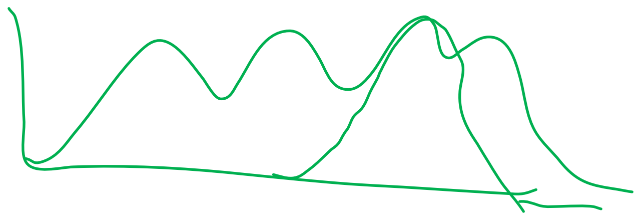
Other properties



- KL(q||p) is *mode-seeking*
  - Distant modes of posterior are ignored, rather than averaged together
- KL(q||p) is *zero-forcing*

  $p(x)=0 \Rightarrow q(x)=0$

  - Causes under-estimate of variance
- When scaled by model evidence, the optimal q is a pointwise lower bound on p (in conjugate-exponential case)



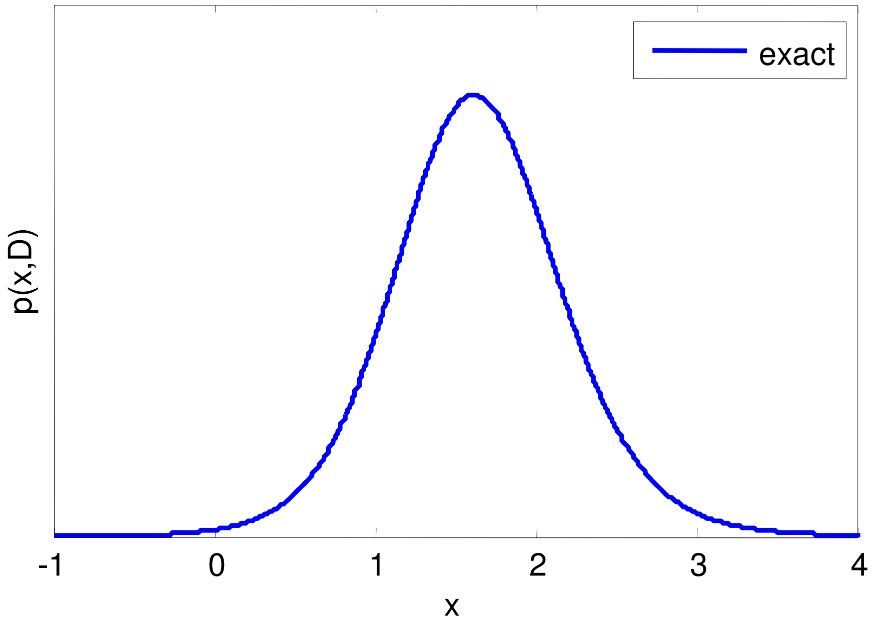
Clutter problem

- Want to estimate x given multiple y’s

$$p(x) \sim N(0,100)$$

$$p(y_i \mid x) = (0.5)N(y_i; x, 1) + (0.5)N(y_i; 0, 10)$$

Exact posterior

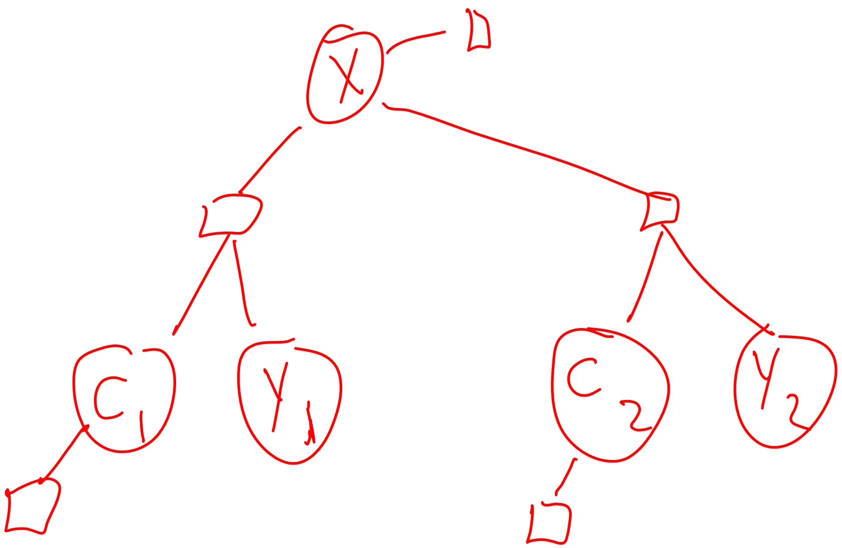


Clutter problem with latent variables

$$p(x) \sim N(0,100)$$

$$p(c_i = 1) = 0.5$$

$$p(y_i \mid c_i, x) = N(y_i; x, 1)^{c_i} \, N(y_i; 0, 10)^{1-c_i}$$



$q(x) \prod_i q(c_i)$

Strategy

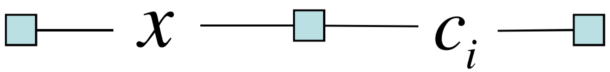
- Approximate *each* factor by a Gaussian in x and Bernoulli in c

$$p(y_i \mid c_i, x) = N(y_i; x, 1)^{c_i} N(y_i; 0, 10)^{1-c_i}$$

$$\approx N(x; m_i, v_i) p_i^{c_i} (1 - p_i)^{1-c_i}$$

Approximating a single factor

$q^{\backslash i}(x)$ $\quad f(x, c_i)$ $\quad q^{\backslash i}(c_i)$

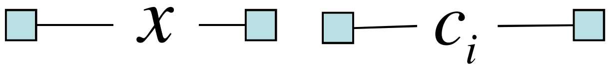


$$KL\left(q(x)\, q(c_i) \,\|\, f(x, c_i)\, q^{\backslash i}(x)\, q^{\backslash i}(c_i) \right)$$

$$q(x) = q^{\backslash i}(x)\, \tilde{f}(x)$$

$$q(c_i) = \tilde{f}(c_i)\, q^{\backslash i}(c_i)$$

$q^{\backslash i}(x)$ $\quad \tilde{f}(x)\, \tilde{f}(c_i)$ $\quad q^{\backslash i}(c_i)$

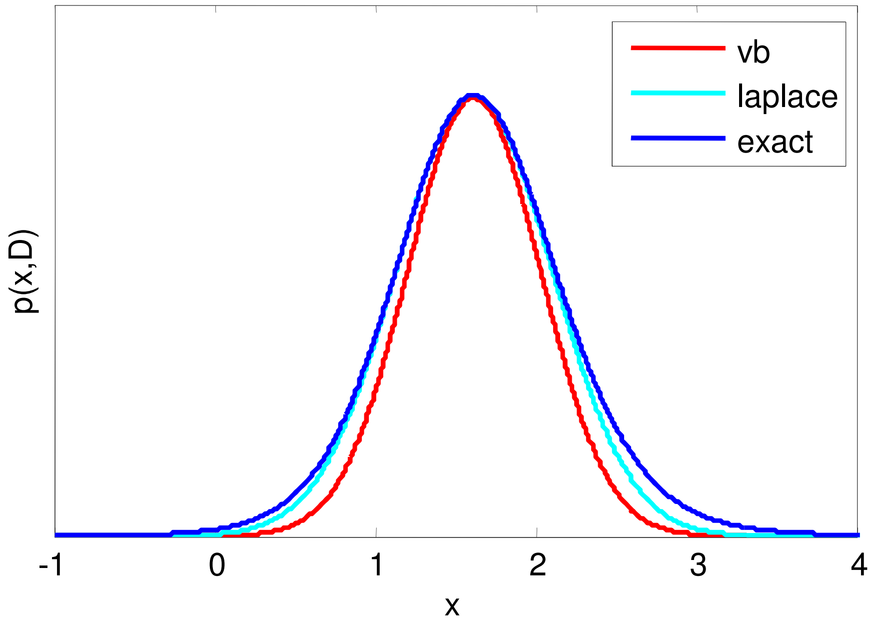


$$\tilde{f}(x) \propto \exp\left(\sum_c q(c) \log f(x, c_i = c) \right)$$

$$\tilde{f}(c) \propto \exp\left(\int_x q(x) \log f(x, c)\, dx \right)$$

Two factors

KL-minimizing Gaussian (vb)



Accuracy

Posterior mean:

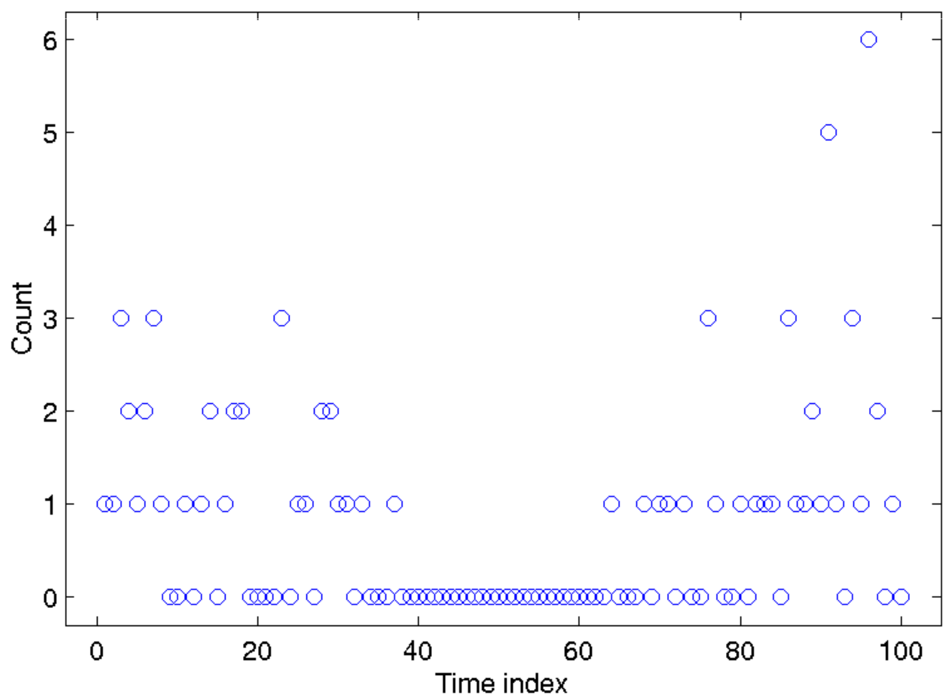
exact = 1.64864
ep = 1.64514
laplace = 1.61946
vb = 1.61834

Posterior variance:

exact = 0.359673
ep = 0.311474
laplace = 0.234616
vb = 0.171155

Example: Poisson tracking

- y_t is a Poisson-distributed integer with mean $\exp(x_t)$



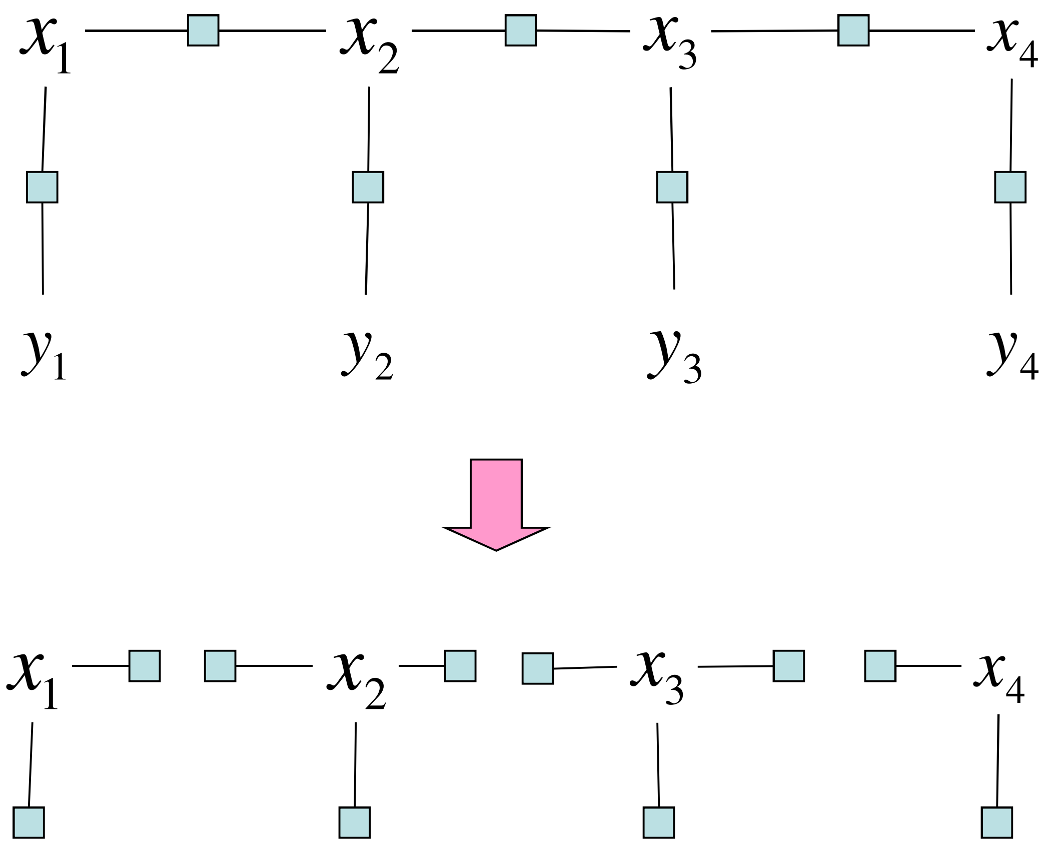
Poisson tracking model

$$p(x_1) \sim N(0,100)$$

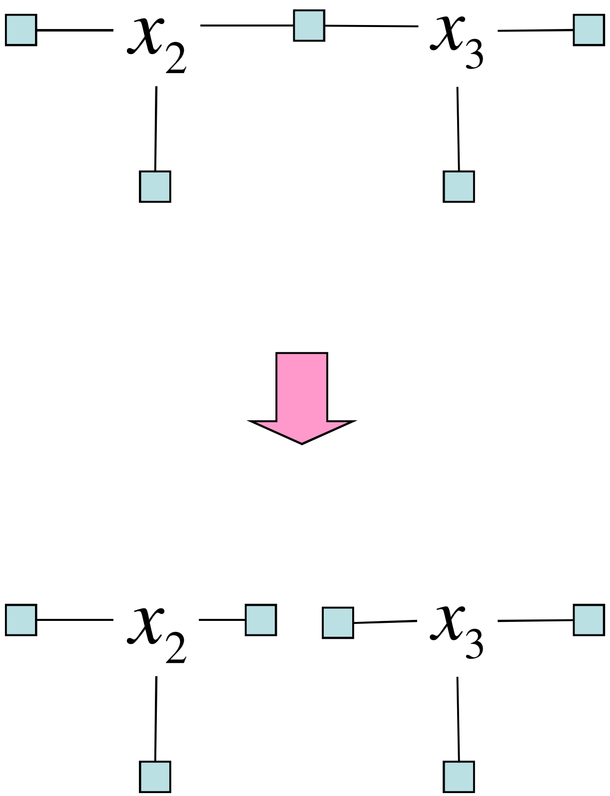
$$p(x_t \mid x_{t-1}) \sim N(x_{t-1}, 0.01)$$

$$p(y_t \mid x_t) = \exp(y_t x_t - e^{x_t}) / y_t!$$

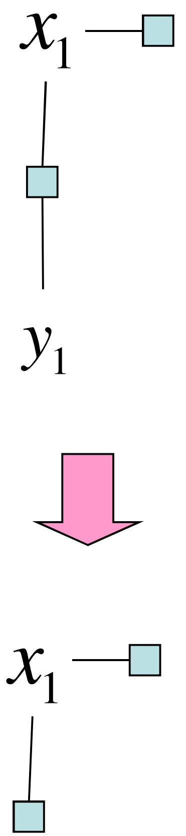
Factor graph

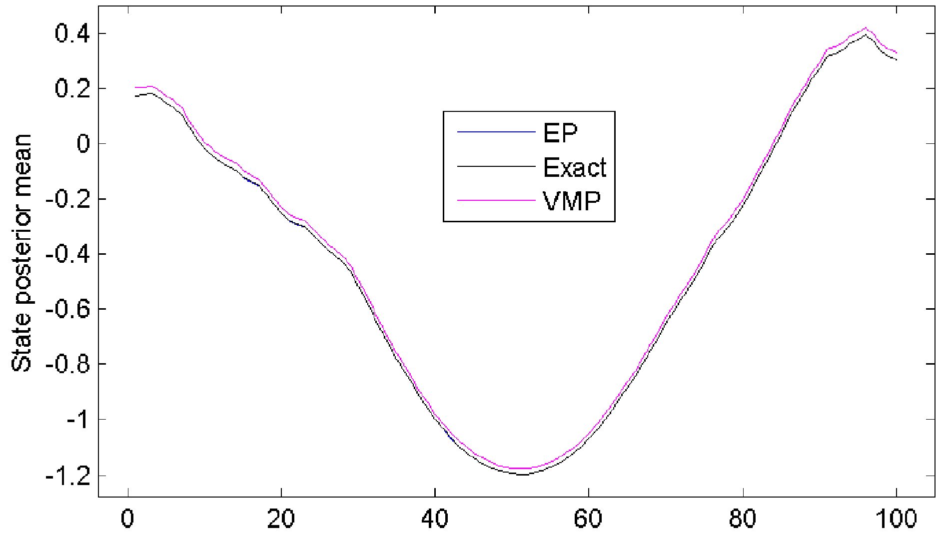


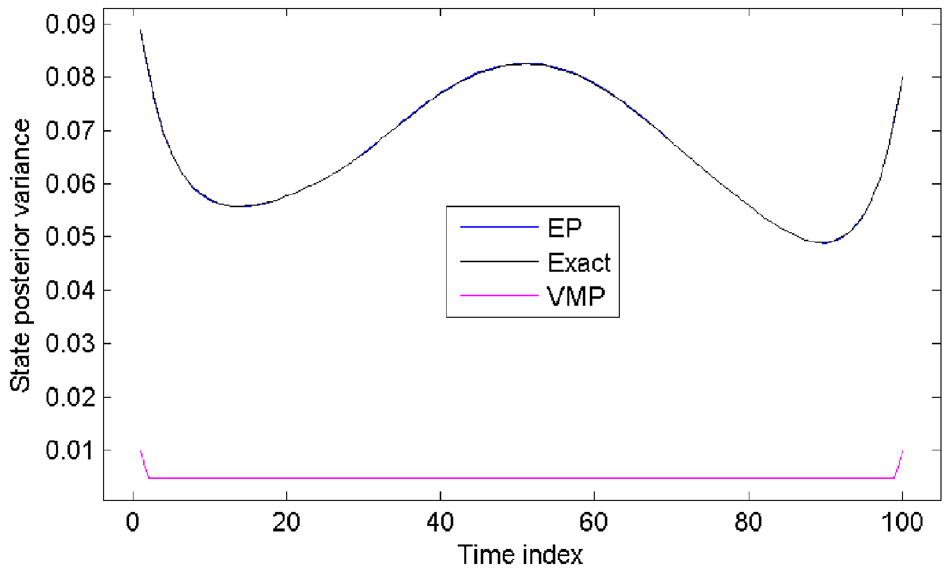
Splitting in context



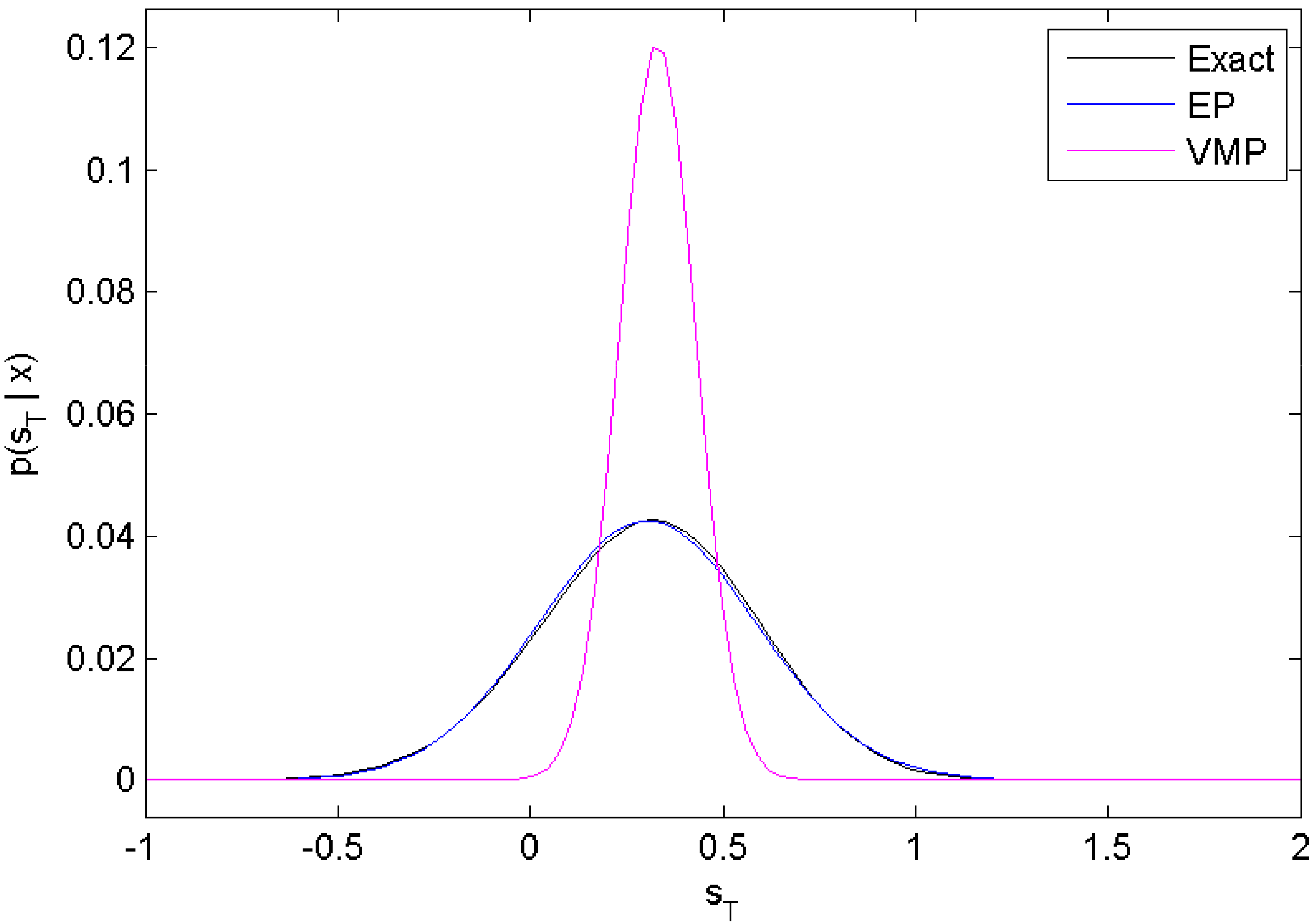
Approximating a measurement factor




State posterior mean
EP
Exact
VMP
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
0
20
40
60
80
100


State posterior variance
EP
Exact
VMP
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
20
40
60
80
100
Time index

Posterior for the last state



Why is VMP so certain?

- Uncertainty does not propagate correctly though the Markov chain

$$p(x_1) \sim N(0,100)$$

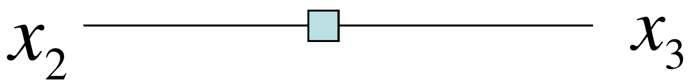
$$p(x_t \mid x_{t-1}) \sim N(x_{t-1}, 0.01)$$

$$\mathrm{var}(x_1) \approx 0.01$$

$$\mathrm{var}(x_2) \approx 0.005$$

VMP messages for Gaussian factor

$q(x_2)$ $\quad N(x_3; x_2, v) \quad$ $q(x_3)$



$N(x_2; E[x_3], v)$ $\quad$ $N(x_3; E[x_2], v)$

Simple example

$$p(x_1) \sim N(0, 100) \qquad p(x_2 \mid x_1) \sim N(x_1, 0.01)$$

$p(x_2) \sim N(0, 100.01)$

$N(x_1; 0, 0.01)$

$\square \longrightarrow x_1 \longrightarrow \square \longleftarrow x_2$

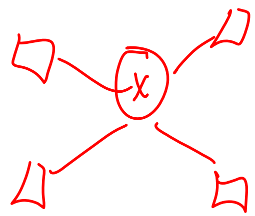
$N(x_2; 0, 0.01)$

$q_{vmp}(x_2) = N(0, 0.01)$

$q_{vmp}(x_1) = N(0, 0.00999)$

Conclusions

- Variational message passing does not handle chains correctly
  - But it is not as troubled by loops



- VMP works best on tight cliques or star graphs
- Make factor graph as compact as possible
  - Remove missing data from factor graph