# Expectation propagation for approximate Bayesian inference

**Thomas P. Minka**

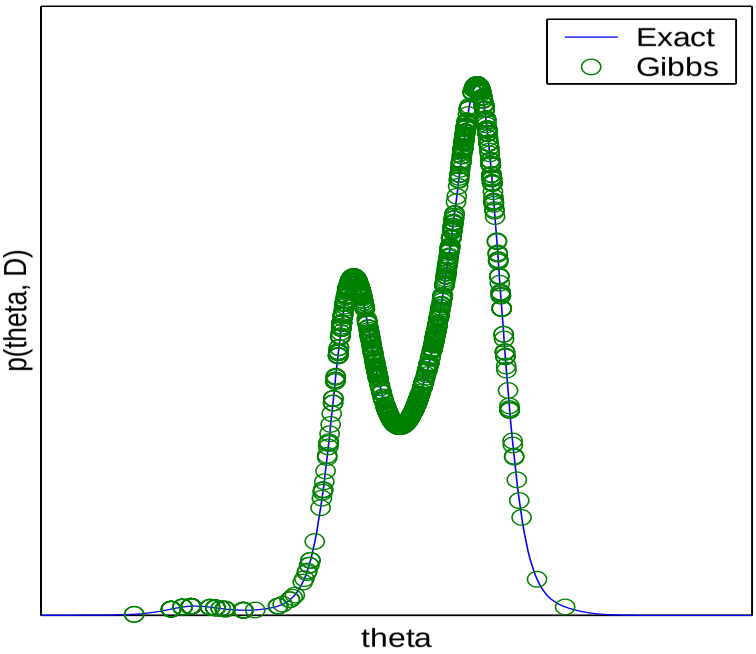
MIT Media Lab

tpminka@media.mit.edu

www.media.mit.edu/~tpminka/

# Approximating posterior distributions

## Sampling

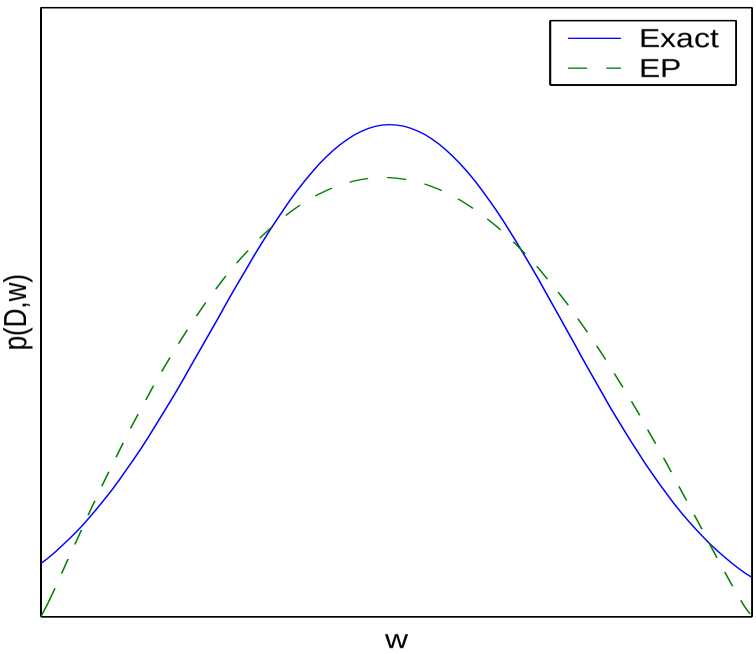



good for complex,

multimodal posteriors

slow, predictable

## Deterministic approximation



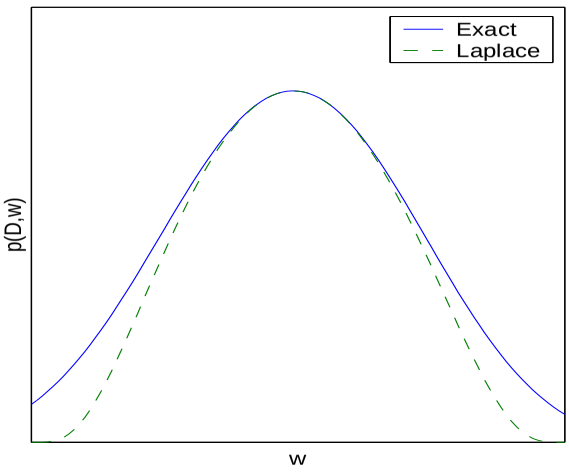

good for simple,

smooth posteriors

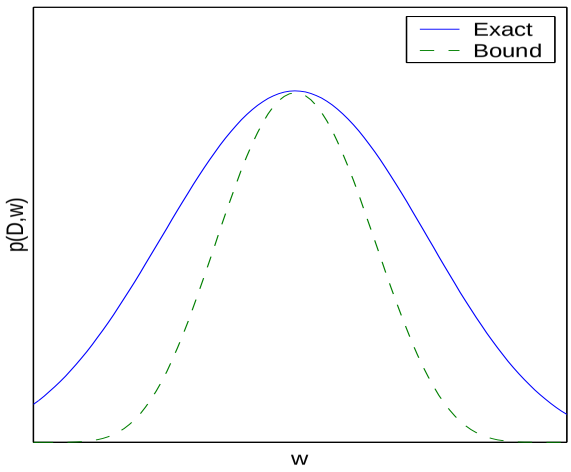
fast, unpredictable

# Deterministic approximation

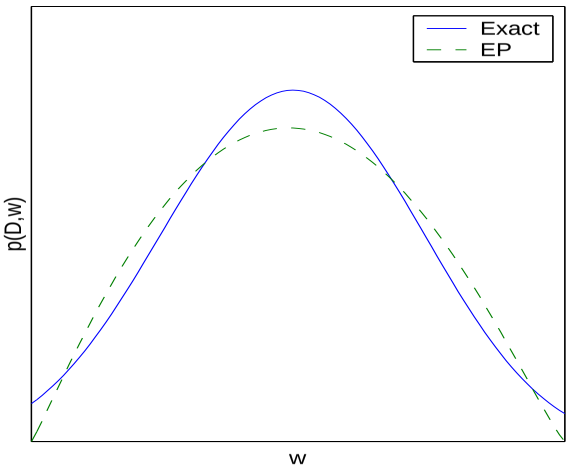
Laplace's method




Variational bound




Minimize
KL-divergence




- Assumed-density filtering
- Belief propagation
- Expectation propagation

# Outline

- Assumed-density filtering (ADF)

  = Sequential KL-projection

  (Boyen&Koller, Opper&Winther, Barber&Sollich, Lauritzen, ...)

- Expectation propagation (EP)

  = ADF + iterative refinement

- Belief propagation

  = EP for factorized posterior

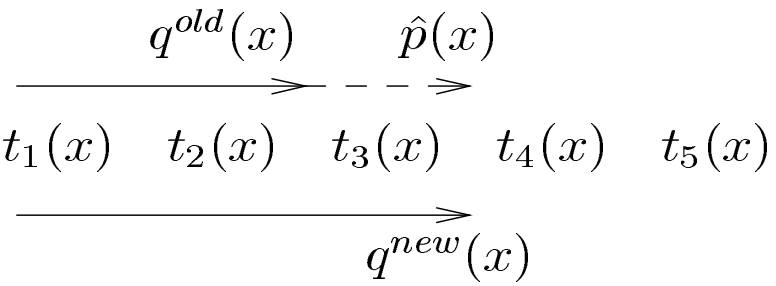
- Application of EP: Bayes point machine
  - Voting all linear classifiers
  - Choosing feature space (kernel)

# Assumed-density filtering

Break joint into product of terms: $p(D, x) = \prod_j t_j(x)$

Want $p(x|D)$

Approximate each one-step posterior:




Example:

$$p(D, x) = p(x) \prod_i p(y_i|x)$$

1. Init $q(x) = p(x)$
2. Loop $i$:

$$\begin{aligned} \hat{p}(x) &= \frac{p(y_i|x)q^{old}(x)}{\int p(y_i|x)q^{old}(x)dx} \\ q^{new}(x) &= \text{argmin}_q \; D(\hat{p} \;||\; q) \end{aligned}$$

# Gaussian filtering

At each step, assume posterior is Gaussian:

$$\text{argmin}_q D(\hat{p} \,||\, q) \qquad\qquad q(x) \sim \mathcal{N}(x; m, v)$$

$$\Downarrow$$

$$\begin{aligned} E_q[x] &= E_{\hat{p}}[x] \\ m &= \int x\hat{p}(x)dx \\ E_q[x^2] &= E_{\hat{p}}[x^2] \\ v + m^2 &= \int x^2\hat{p}(x)dx \end{aligned}$$

$$\mathcal{N}(x; m^{old}, v^{old})p(y_i|x) \Rightarrow \mathcal{N}(x; m^{new}, v^{new})$$

q preserves certain expectations of p

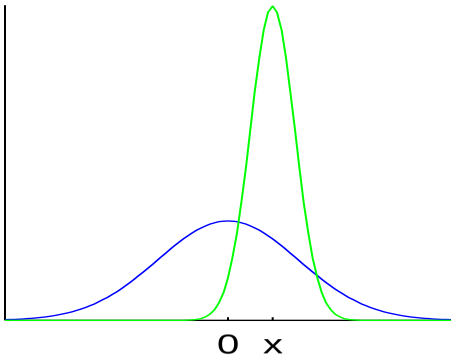
$\longrightarrow$ true for any assumed density in exponential family

(Gamma, Dirichlet, multinomial, ...)

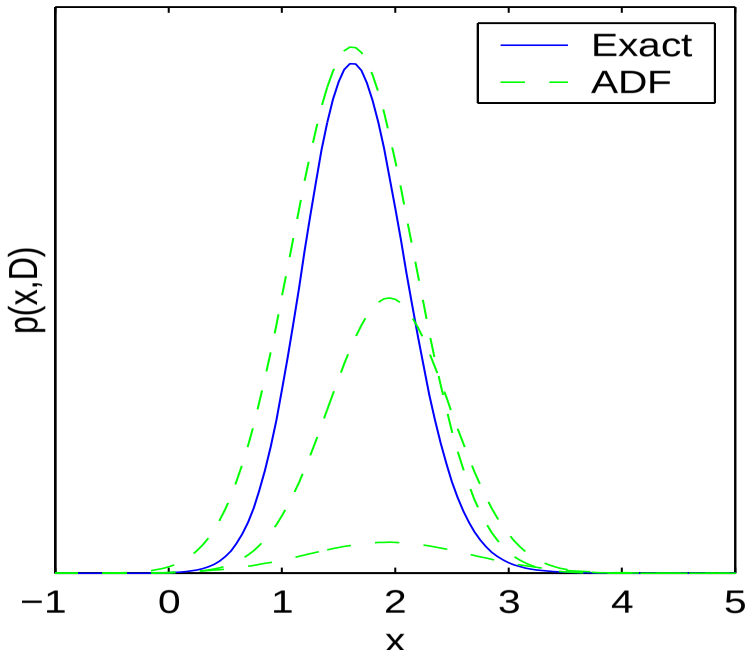
# Example

Data model

$$p(y|x) = \frac{1}{2}\mathcal{N}(y; x, 1) + \frac{1}{2}\mathcal{N}(y; 0, 10)$$




ADF posterior for three orderings of same data:




x = 2

n = 20

ADF is too sensitive to ordering

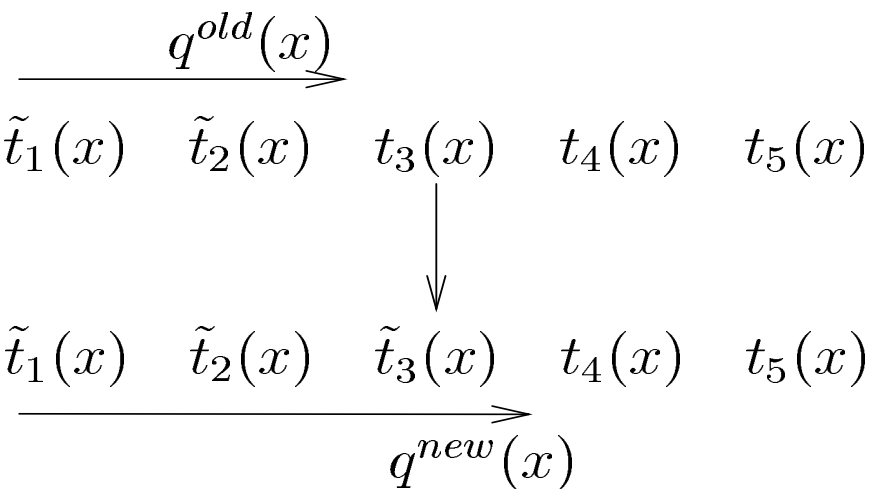
Can we make it independent of ordering?

# Another view of ADF

Approximate posterior for exact terms

=

Exact posterior for approximate terms




$$\tilde{t}_3(x) = \frac{q^{new}(x)}{q^{old}(x)}$$

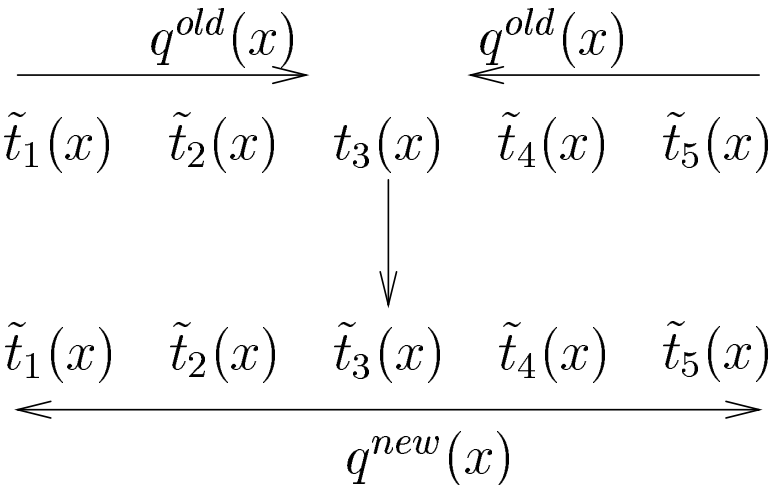
$q$ in exp family $\Rightarrow$ $\tilde{t}$ in exp family

Approximate each term as Gaussian, then multiply

What if we continue to refine the Gaussians?

# Expectation Propagation

Refine each term on the basis of all other terms:



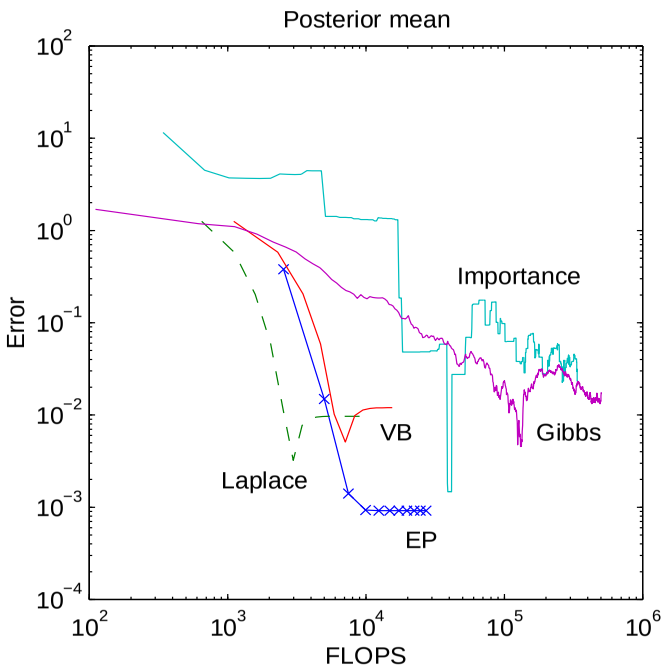

1. Remove $\tilde{t}_i$:

$$q^{old}(x) \propto \frac{q^{new}(x)}{\tilde{t}_i(x)}$$

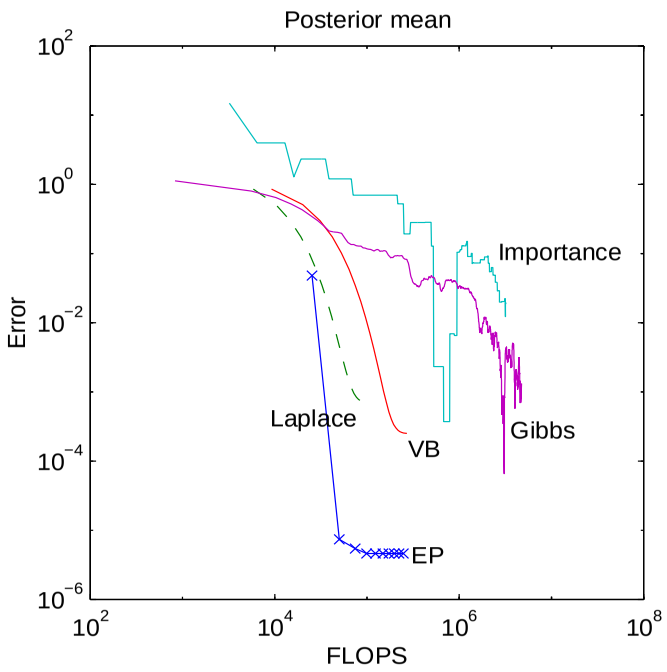
2. Compute $q^{new}(x)$ from $q^{old}(x)$ and $t_i(x)$ (ADF)

3. $\tilde{t}_i(x) = \frac{q^{new}(x)}{q^{old}(x)}$

# Performance




n=20




n=200

ADF = first ‘x’ of EP

VB = variational bound

Deterministic methods improve with more data

(posterior is more Gaussian)

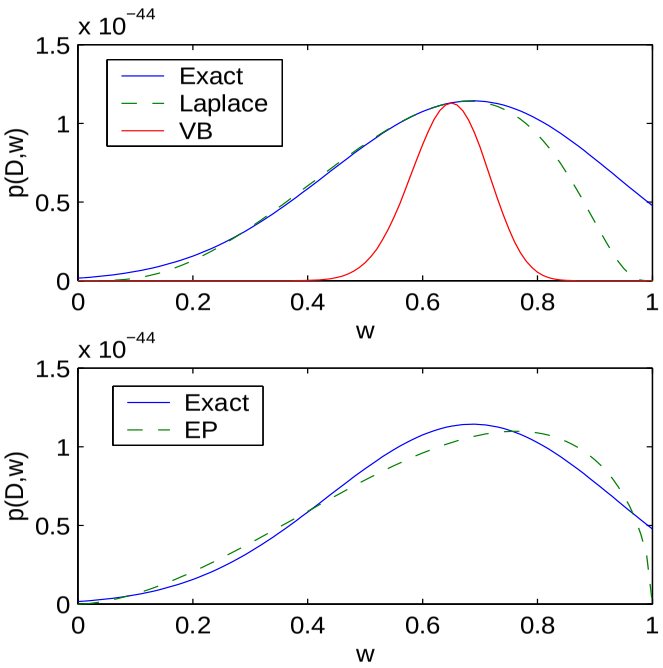
Sampling methods do not care

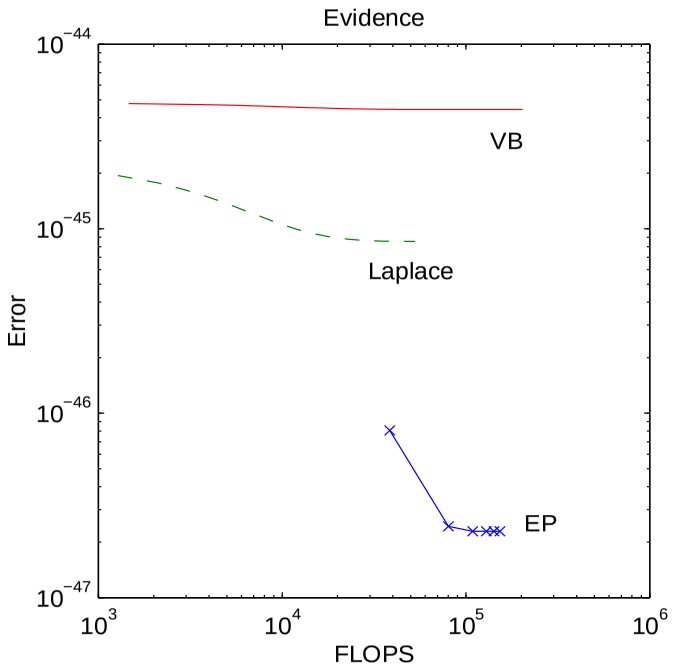
# Mixture weights example

$$p(y|w) = w\mathcal{N}(y;0,3) + (1-w)\mathcal{N}(y;1,3)$$
$$q(w) \sim \text{Dirichlet}(a_1, a_2)$$

Expectations are $E[\log(w)]$, $E[\log(1-w)]$

Typical result



# Factorized approximation

$$\text{argmin}_q D(\hat{p} \,||\, q) \qquad\qquad q(x) = \prod_k q_k(x_k)$$

$$\Rightarrow q_k(x_k) = \hat{p}(x_k) \qquad\qquad (\text{marginals})$$

Factorized ADF (Boyen&Koller)

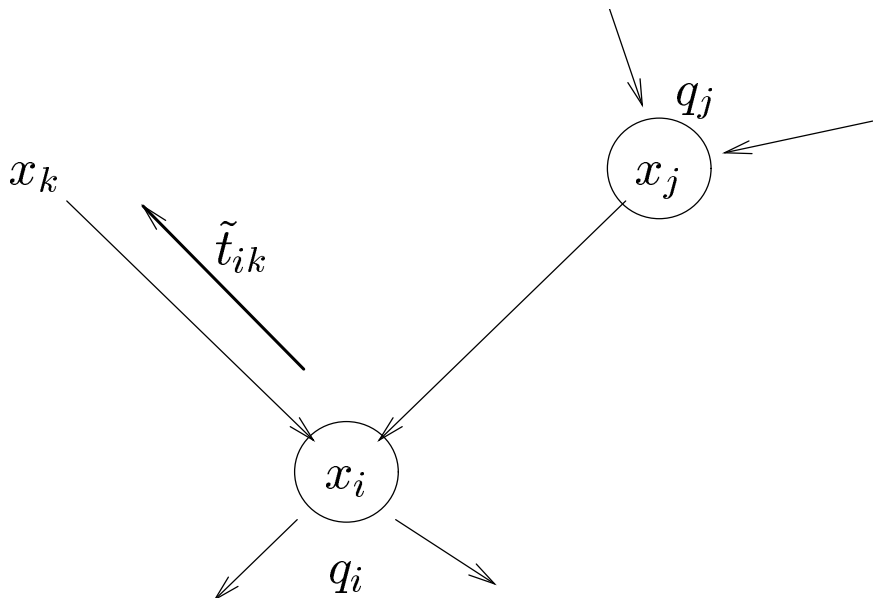
$$p(\mathbf{x}) \;=\; \prod_i p(x_i|\text{pa}(x_i))$$

1. Init $q(\mathbf{x}) = 1$
2. Loop $i$:

$$\begin{aligned} \hat{p}(\mathbf{x}) &\propto p(x_i|\text{pa}(x_i))q^{old}(\mathbf{x}) \\ q_k^{new}(x_k) &= \hat{p}(x_k) \end{aligned}$$

# Factorized EP

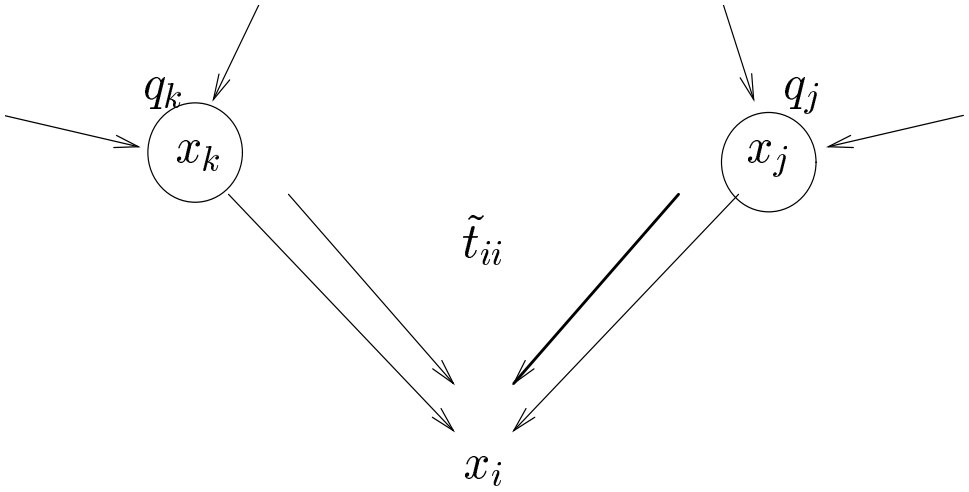
$$
\begin{aligned}
\tilde{t}_i(\mathbf{x}) &= \prod_k \tilde{t}_{ik}(x_k) \qquad \text{(messages)} \\
q^{new}(\mathbf{x}) &\propto \prod_j \tilde{t}_j(\mathbf{x}) \qquad \text{(belief state)} \\
q^{old}(\mathbf{x}) &\propto \prod_{j \neq i} \tilde{t}_j(\mathbf{x}) \qquad \text{(partial belief state)} \\
\tilde{t}_{ik}(x_k) &= \frac{\hat{p}(x_k)}{q_k^{old}(x_k)} \\
&= \int \delta_{x_k} p(x_i|\text{pa}(x_i)) \prod_{j \neq k} q_j^{old}(x_j) d\mathbf{x} \\
&\quad \text{(message } i \to k\text{)}
\end{aligned}
$$




$$\tilde{t}_{ik}(x_k) = \int_{x_i, x_j} p(x_i|x_k, x_j) q_i(x_i) q_j(x_j)$$

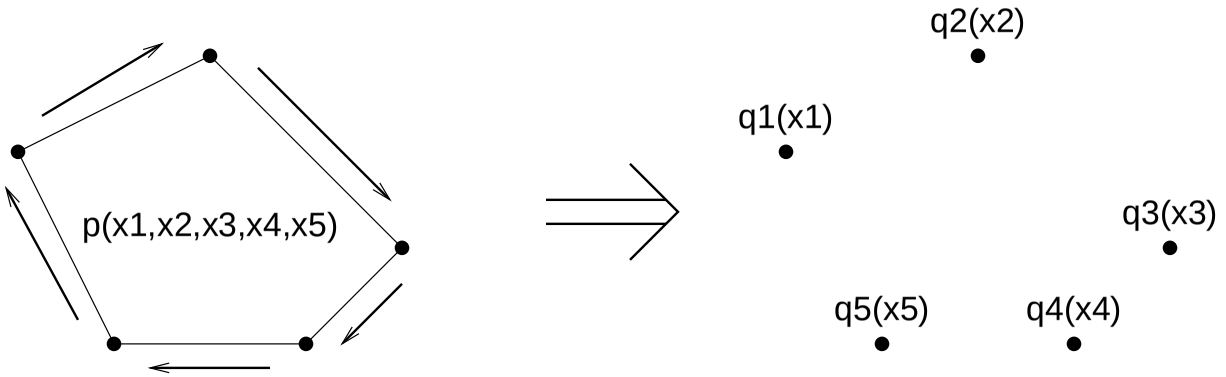
# Factorized EP cont'd

Parents to child:




$$\tilde{t}_{ii}(x_i) = \int_{x_k, x_j} p(x_i|x_k, x_j) q_k(x_k) q_j(x_j)$$

Loopy belief propagation is factorized EP




Can use other approximations within EP, e.g. Markov chain

# Bayes point machine

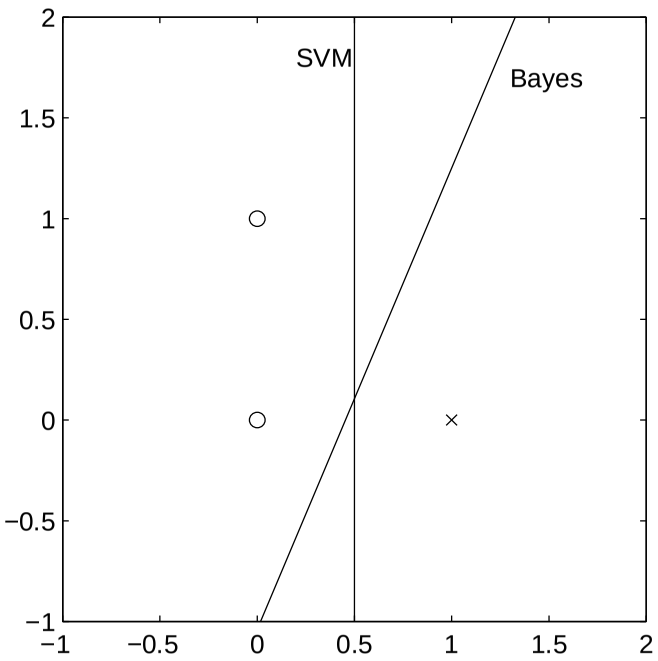
Use w to classify y

$$\mathbf{w}^{\mathrm{T}}\mathbf{y}_i > 0 \qquad (\text{class } 1)$$

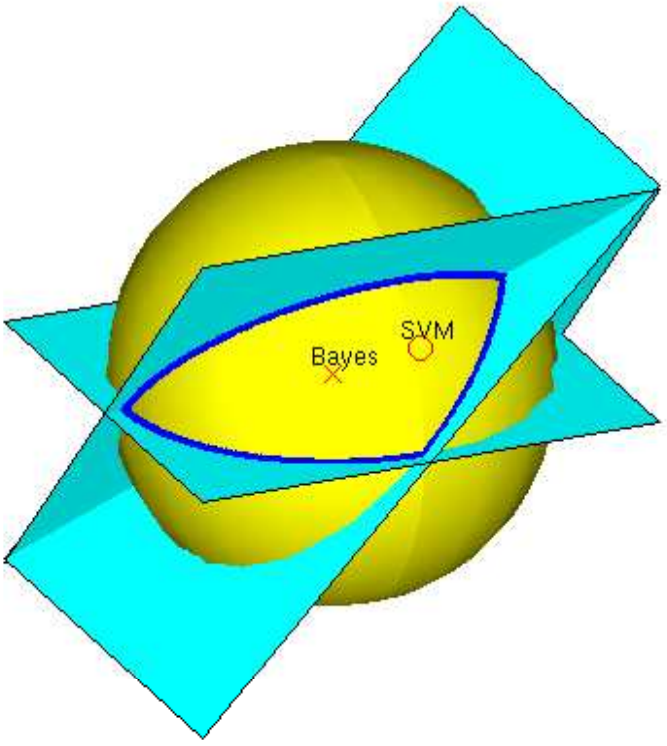
$$\mathbf{w}^{\mathrm{T}}\mathbf{y}_i < 0 \qquad (\text{class } 2)$$

SVM $\longrightarrow$ Maximize margin

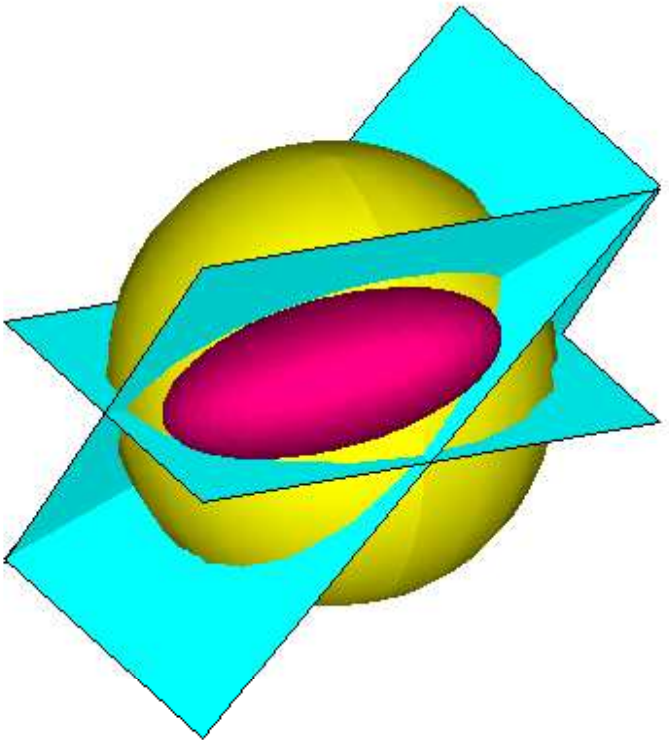
Bayes $\longrightarrow$ Vote all separators




Version space




EP Gaussian posterior



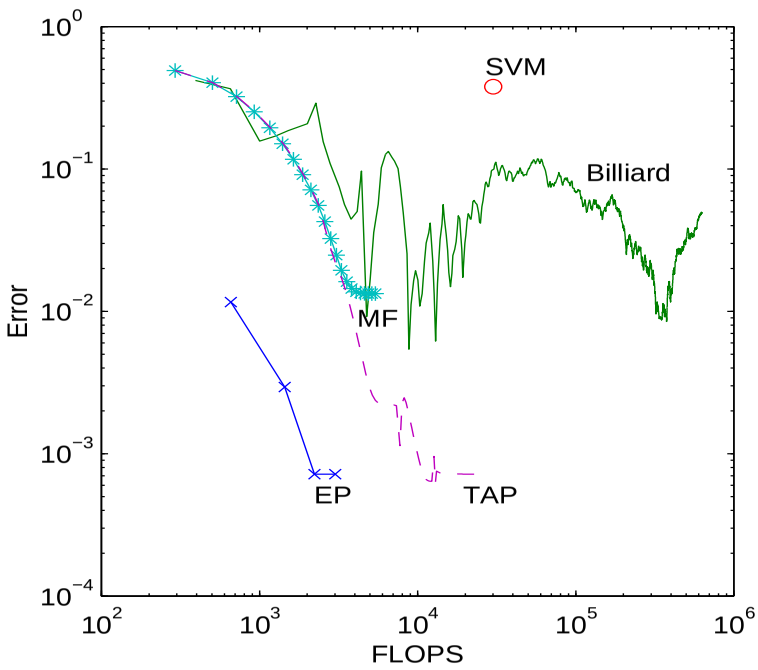
# Performance

Billiard = Monte Carlo

Opper&Winther's algs:

MF = mean-field theory

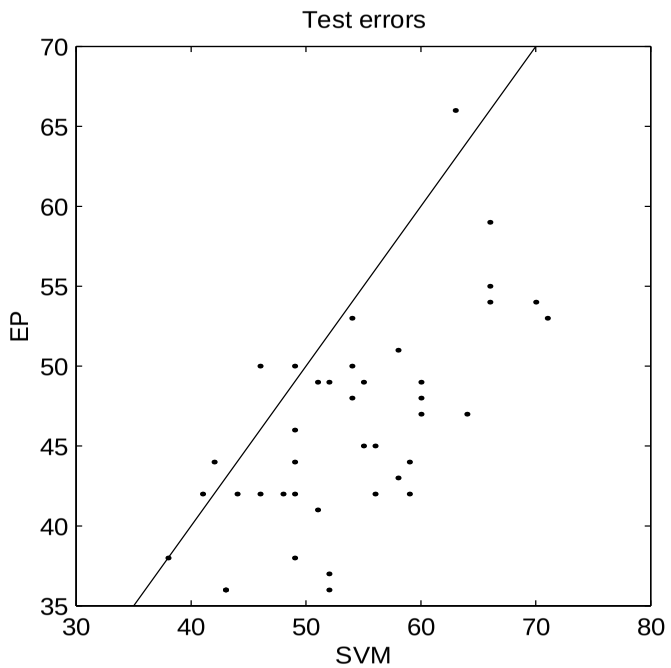
TAP = cavity method
(equiv to Gaussian EP)

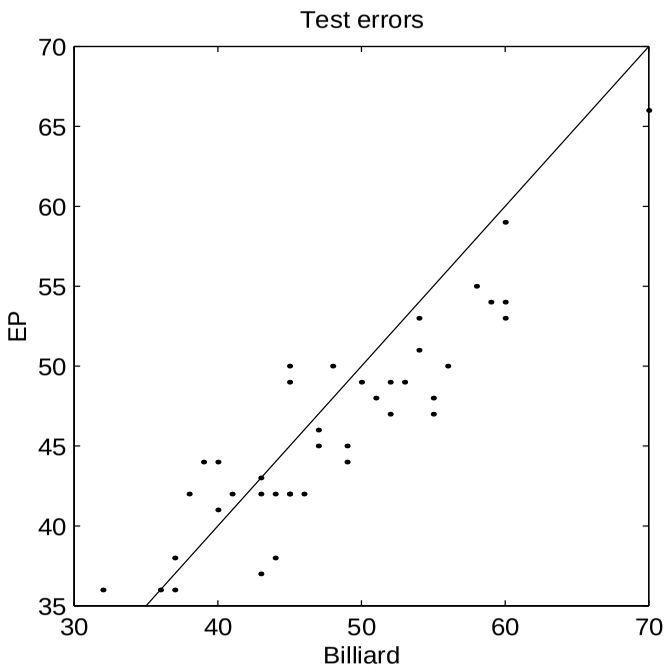



Digit recognition ('3' from '5')

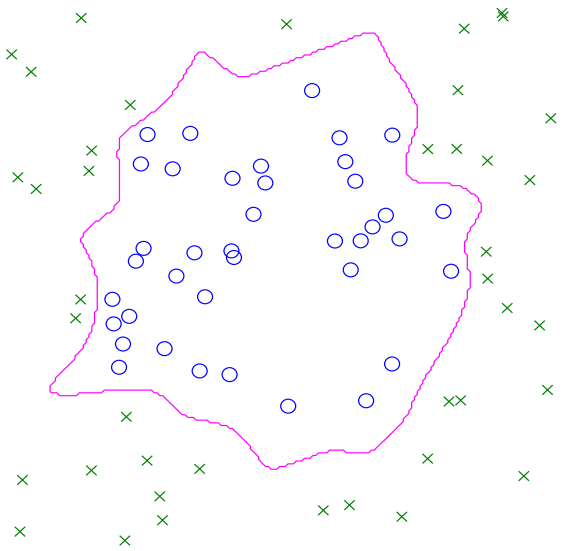
40 random train/test splits

64 dimensions, linear

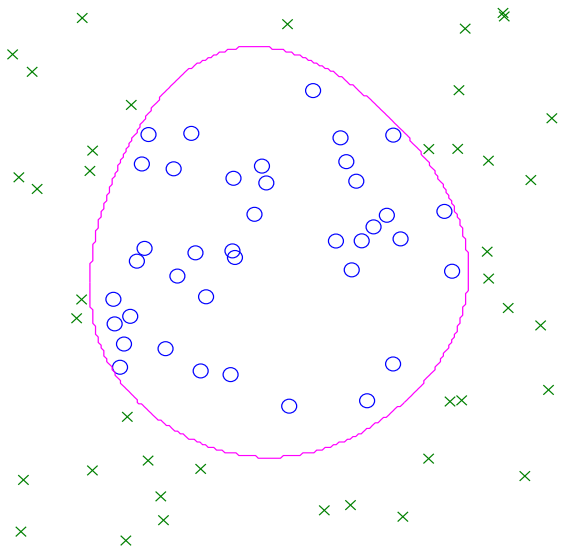


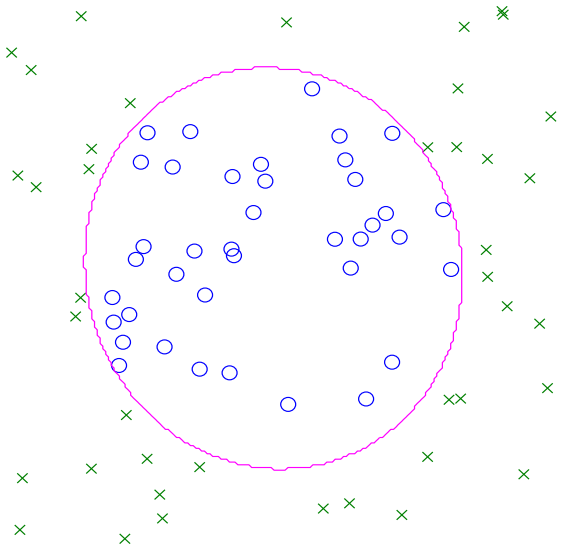
# Kernel selection



Gaussian kernel, width 0.08
(SVM choice)



Gaussian kernel, width 0.6
(Bayes choice among Gaussians)



Quadratic kernel
(Bayes choice)

| Kernel | $R^2/\rho^2$ | $\log(p(D))$ |
|---|---|---|
| $\sigma = 0.08$ | 18 | -39 |
| $\sigma = 0.6$ | 108 | -19 |
| quadratic | 656 | -16 |

SVM and EP have similar boundaries, but prefer different kernels

# Summary

- Expectation propagation = assumed-density filtering plus iterative refinement
- Order-free, more accurate
- Generalizes belief propagation to hybrid nets and non-factorized approximations
- Generalizes TAP method for Bayes point machine
- Like belief propagation, may not converge
- No error estimate available