Divergence measures and message passing

Tom Minka

Microsoft Research

Cambridge, UK

with thanks to the Machine Learning and Perception Group

Message-Passing Algorithms

Mean-field	MF	[Peterson,Anderson 87]
Loopy belief propagation	BP	[Frey,MacKay 97]
Expectation propagation	EP	[Minka 01]
Tree-reweighted message passing	TRW	[Wainwright,Jaakkola,Willsky 03]
Fractional belief propagation	FBP	[Wiegerinck,Heskes 02]
Power EP	PEP	[Minka 04]

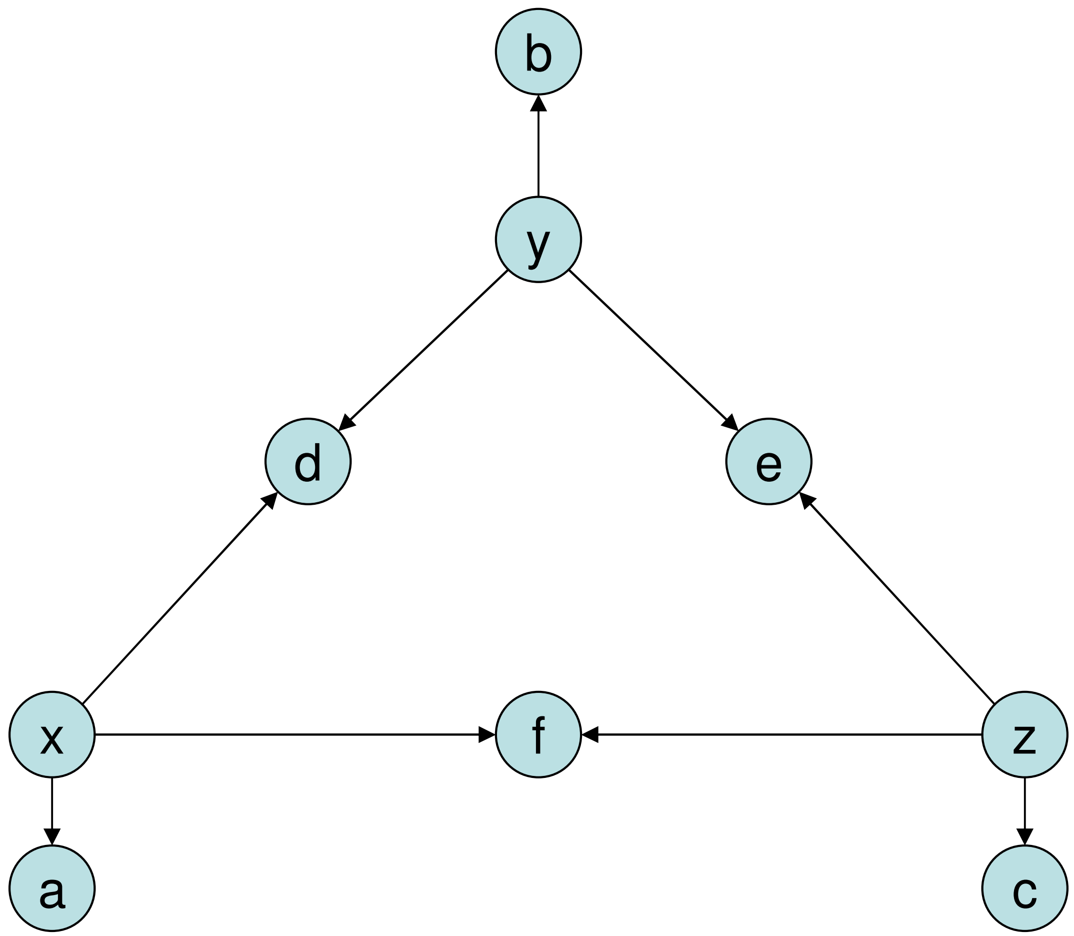
Outline

- Example of message passing
- Interpreting message passing
- Divergence measures
- Message passing from a divergence measure
- Big picture

Outline

- Example of message passing
- Interpreting message passing
- Divergence measures
- Message passing from a divergence measure
- Big picture

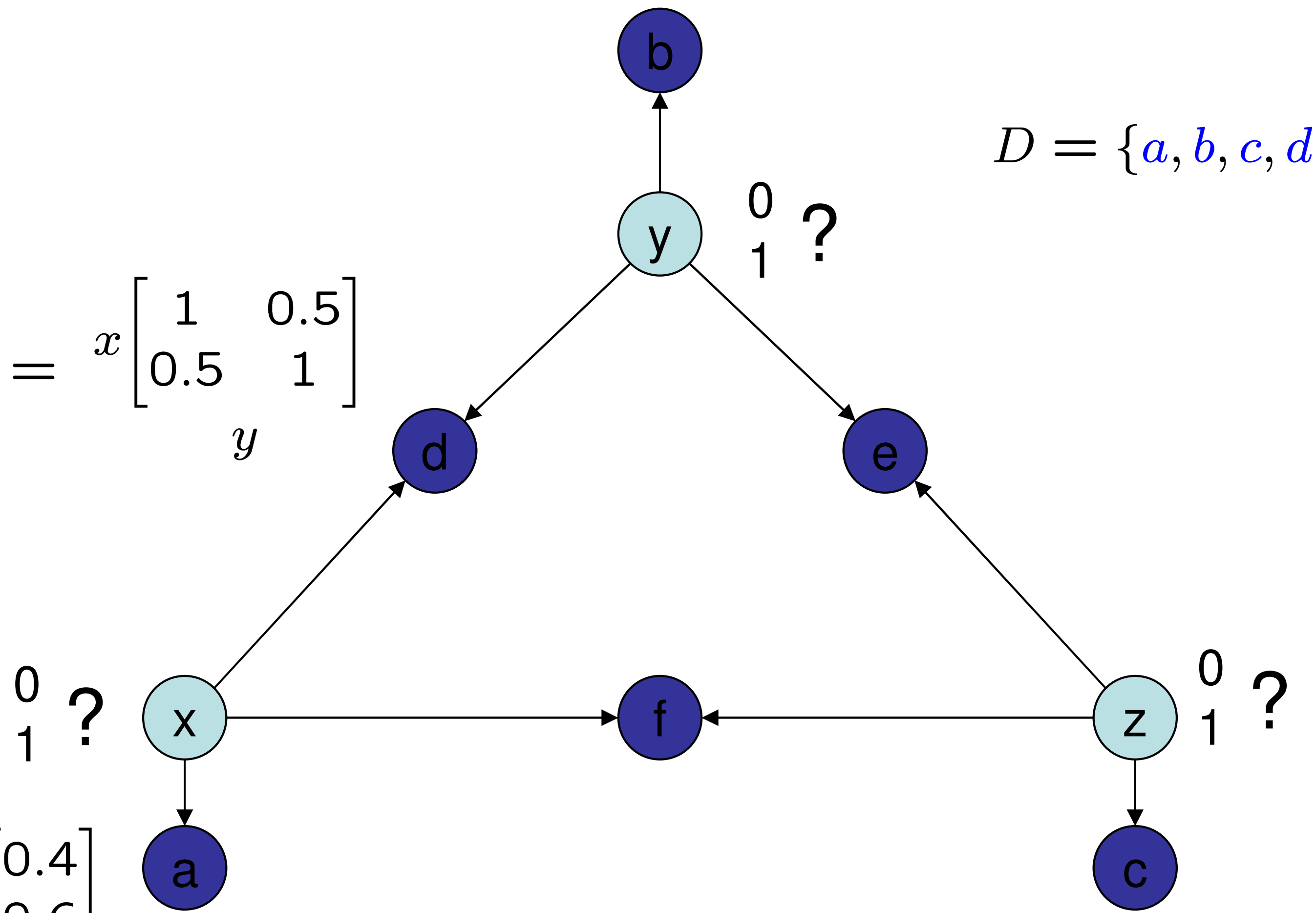
Estimation Problem



Estimation Problem

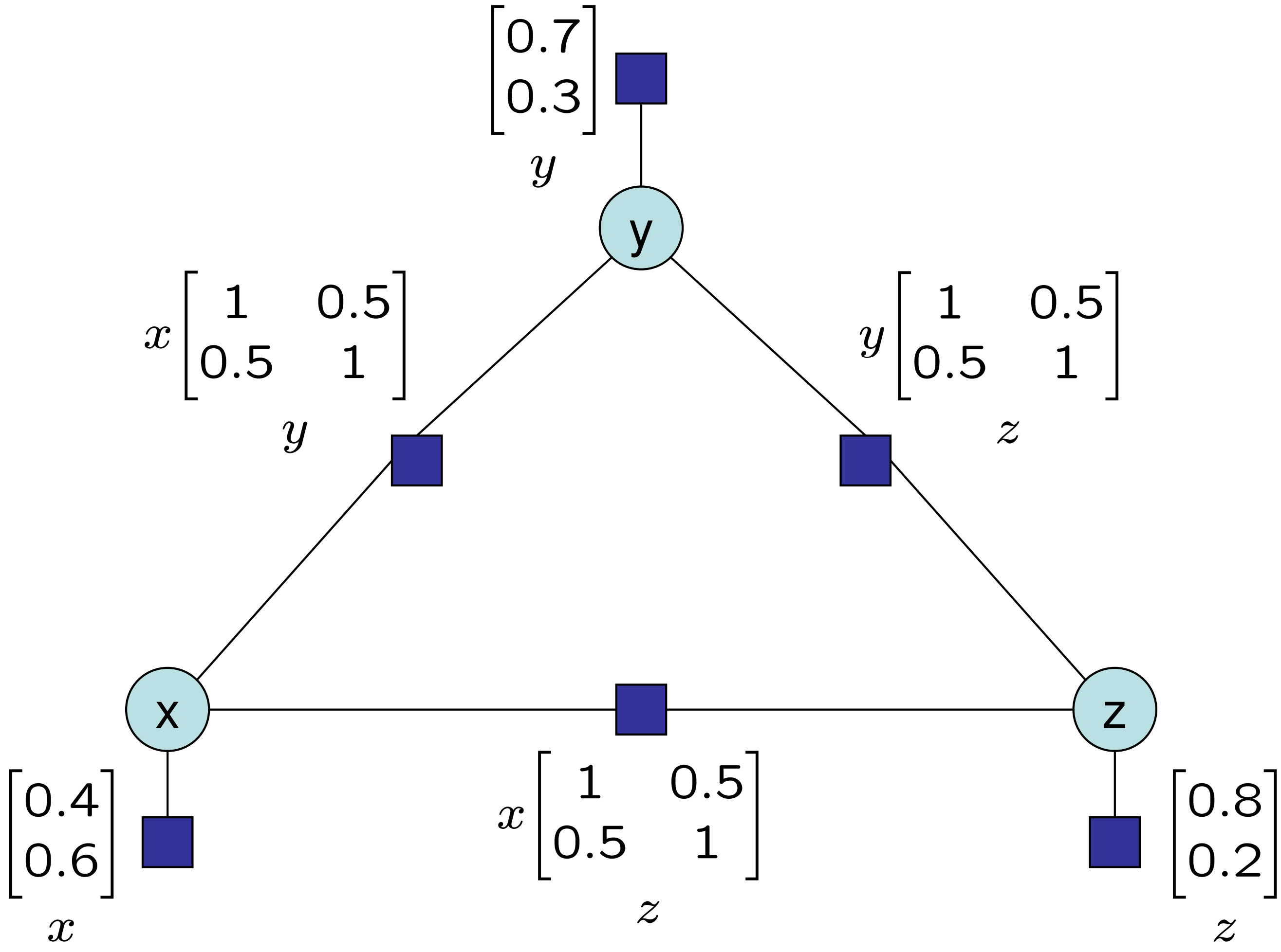
$$D = \{a, b, c, d, e, f\}$$

$$p(d|x,y) = \begin{array}{cc} x & \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \\ & y \end{array}$$



$$p(a|x) = \begin{array}{c} \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \\ x \end{array}$$

Estimation Problem



Estimation Problem

$$p(x,y,z,D) = \begin{array}{cc} x & \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \\ & y \end{array} \begin{array}{cc} y & \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \\ & z \end{array} \begin{array}{cc} x & \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \\ & z \end{array} \begin{array}{c} \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \\ x \end{array} \begin{array}{c} \begin{bmatrix} 0.7 \\ 0.3 \end{bmatrix} \\ y \end{array} \begin{array}{c} \begin{bmatrix} 0.8 & 8 \\ 0.2 & 2 \end{bmatrix} \\ z \end{array}$$

$$p(0,0,0,D) = 0.224$$
$$p(0,0,1,D) = 0.014$$
$$p(0,1,0,D) = 0.024$$
$$p(0,1,1,D) = 0.006$$
$$p(1,0,0,D) = 0.084$$
$$p(1,0,1,D) = 0.021$$
$$p(1,1,0,D) = 0.036$$
$$p(1,1,1,D) = 0.036$$

Queries:

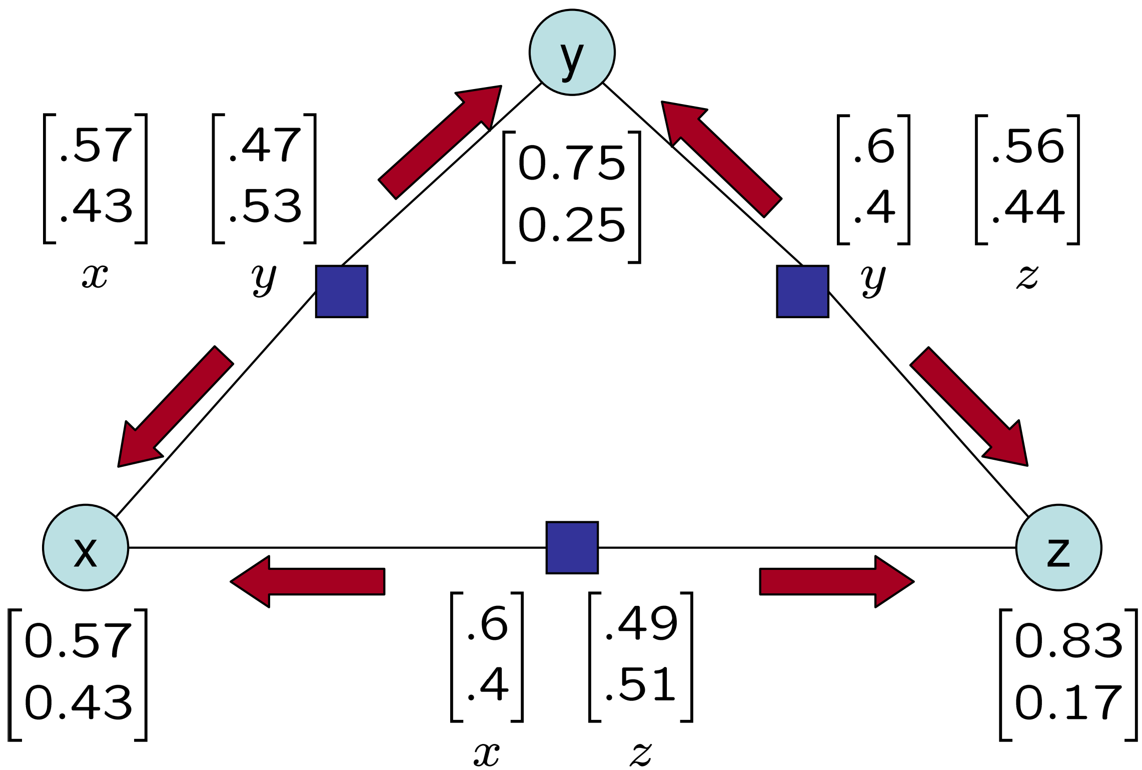
$$p(x,D) = \textstyle\sum_{y,z} p(x,y,z,D)$$

$$p(D) = \textstyle\sum_{x,y,z} p(x,y,z,D)$$

$$(x^\star, y^\star, z^\star) = \mathrm{argmax}\, p(x,y,z,D)$$

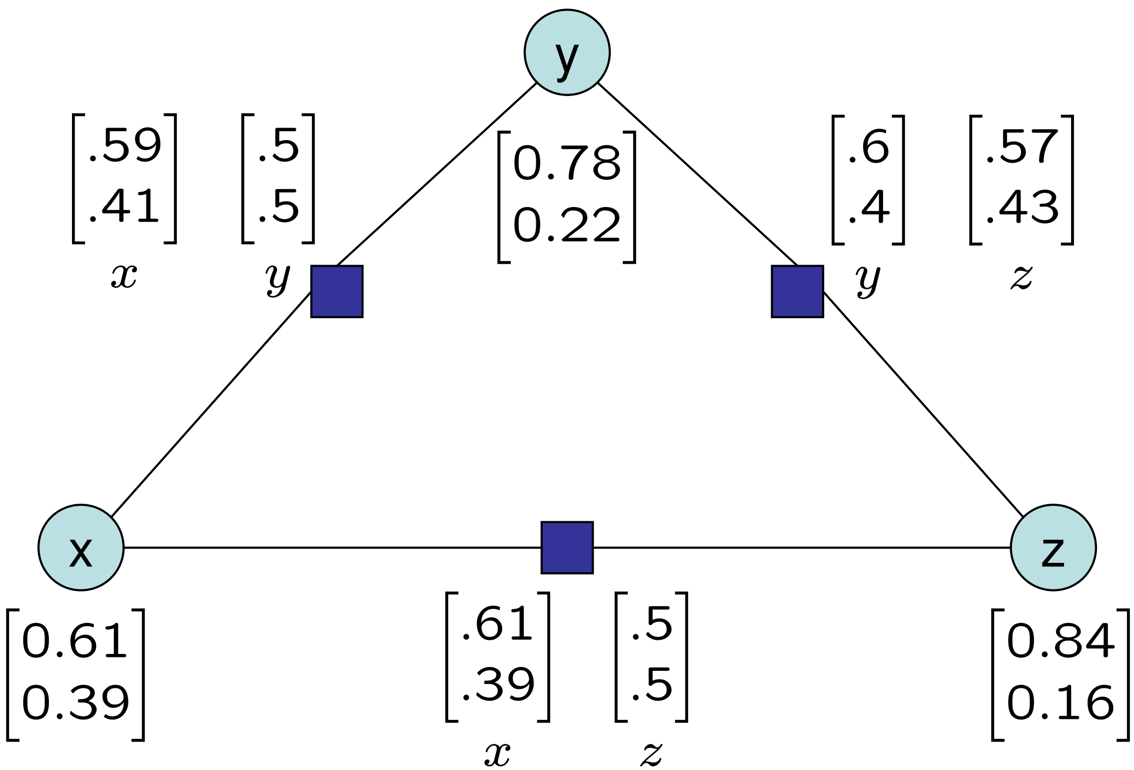
Want to do these *quickly*

Belief Propagation



Belief Propagation

Final



Belief Propagation

Marginals:

$$\underset{x}{\begin{bmatrix}0.60\\0.40\end{bmatrix}} \quad \underset{y}{\begin{bmatrix}0.77\\0.23\end{bmatrix}} \quad \underset{z}{\begin{bmatrix}0.83\\0.17\end{bmatrix}} \quad \text{(Exact)}$$

$$\underset{x}{\begin{bmatrix}0.61\\0.39\end{bmatrix}} \quad \underset{y}{\begin{bmatrix}0.78\\0.22\end{bmatrix}} \quad \underset{z}{\begin{bmatrix}0.84\\0.16\end{bmatrix}} \quad \text{(BP)}$$

Normalizing constant: 0.45 (Exact)
0.44 (BP)

Argmax: (0,0,0) (Exact)
(0,0,0) (BP)

Outline

- Example of message passing
- Interpreting message passing
- Divergence measures
- Message passing from a divergence measure
- Big picture

Message Passing = Distributed Optimization

- Messages represent a simpler distribution $q(x)$ that approximates $p(x)$
  - A *distributed* representation
- Message passing = optimizing q to fit p
  - q stands in for p when answering queries
- Parameters:
  - What type of distribution to construct (approximating family)
  - What cost to minimize (divergence measure)

How to make a message-passing algorithm

1. Pick an approximating family
   - fully-factorized, Gaussian, etc.
2. Pick a divergence measure
3. Construct an optimizer for that measure
   - usually fixed-point iteration
4. Distribute the optimization across factors

Outline

- Example of message passing
- Interpreting message passing
- Divergence measures
- Message passing from a divergence measure
- Big picture

Let p,q be *unnormalized* distributions

Kullback-Leibler (KL) divergence

$$KL(p \mid\mid q) = \int p(x) \log \frac{p(x)}{q(x)} dx + \int (q(x) - p(x)) dx$$

Alpha-divergence (α is any real number)

$$D_\alpha(p \mid\mid q) = \frac{\int_x \alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha} dx}{\alpha(1-\alpha)}$$

Asymmetric, convex

$$D_\alpha(p \mid\mid q) = 0 \qquad \text{if } p = q$$
$$D_\alpha(p \mid\mid q) > 0 \qquad \text{otherwise}$$

Examples of alpha-divergence

$$D_{-1}(p \,||\, q) = \frac{1}{2}\int_x \frac{(q(x) - p(x))^2}{p(x)} dx$$

$$D_0(p \,||\, q) = KL(q \,||\, p)$$

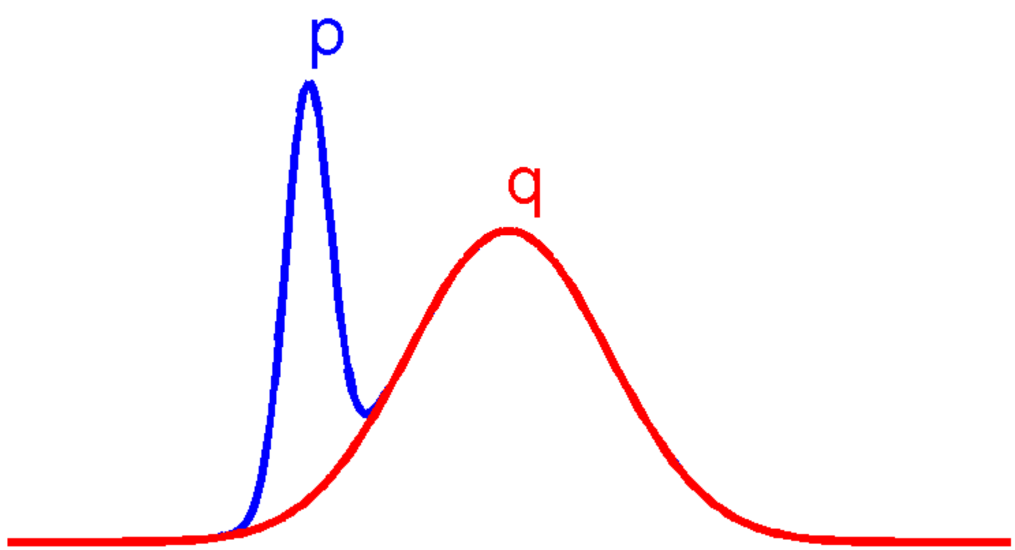
$$D_{\frac{1}{2}}(p \,||\, q) = 2\int_x \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$$

$$D_1(p \,||\, q) = KL(p \,||\, q)$$

$$D_2(p \,||\, q) = \frac{1}{2}\int_x \frac{(p(x) - q(x))^2}{q(x)} dx$$

Minimum alpha-divergence

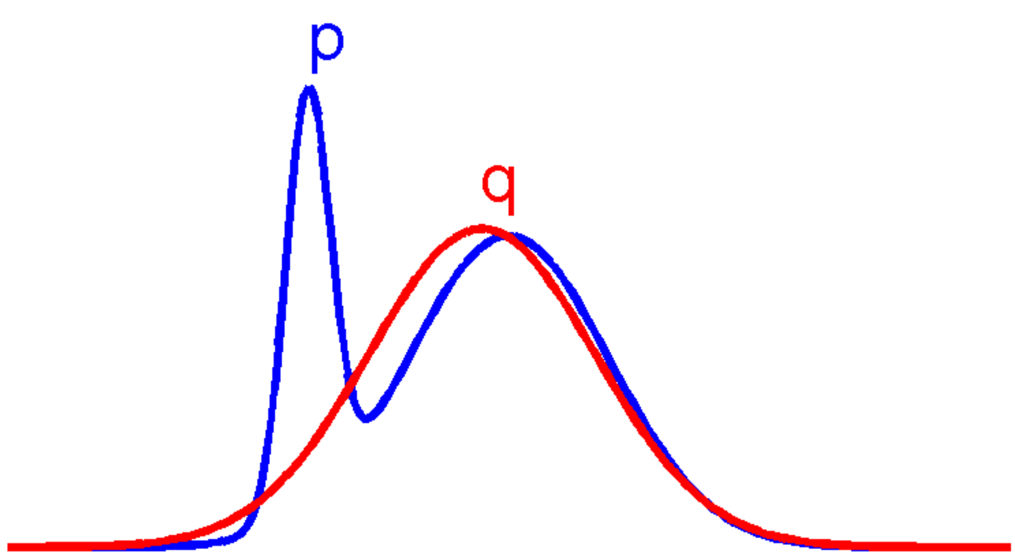
q is Gaussian, minimizes $D_\alpha(p||q)$



$\alpha = -\infty$

Minimum alpha-divergence

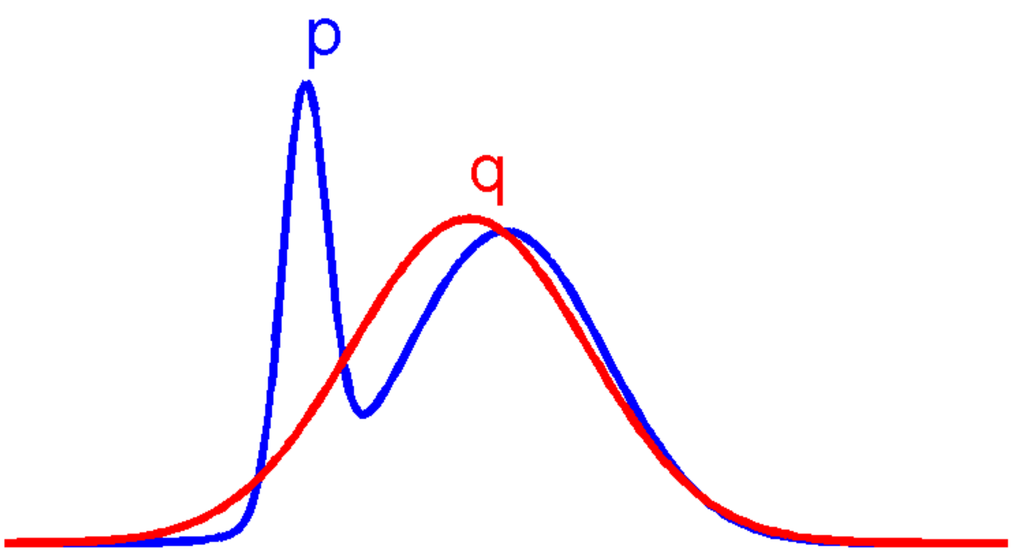
q is Gaussian, minimizes $D_\alpha(p||q)$



$\alpha = 0$

Minimum alpha-divergence

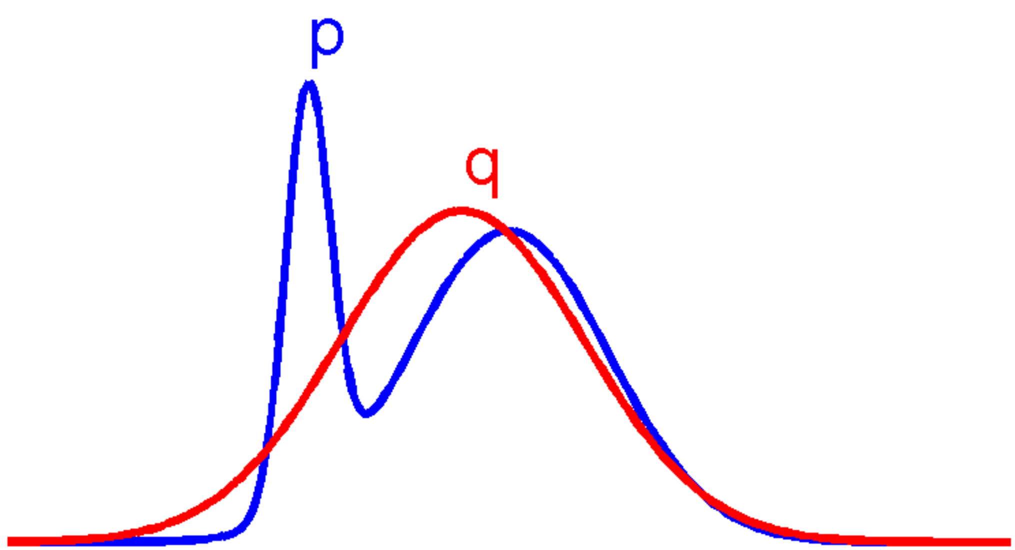
q is Gaussian, minimizes $D_\alpha(p||q)$



$\alpha = 0.5$

Minimum alpha-divergence

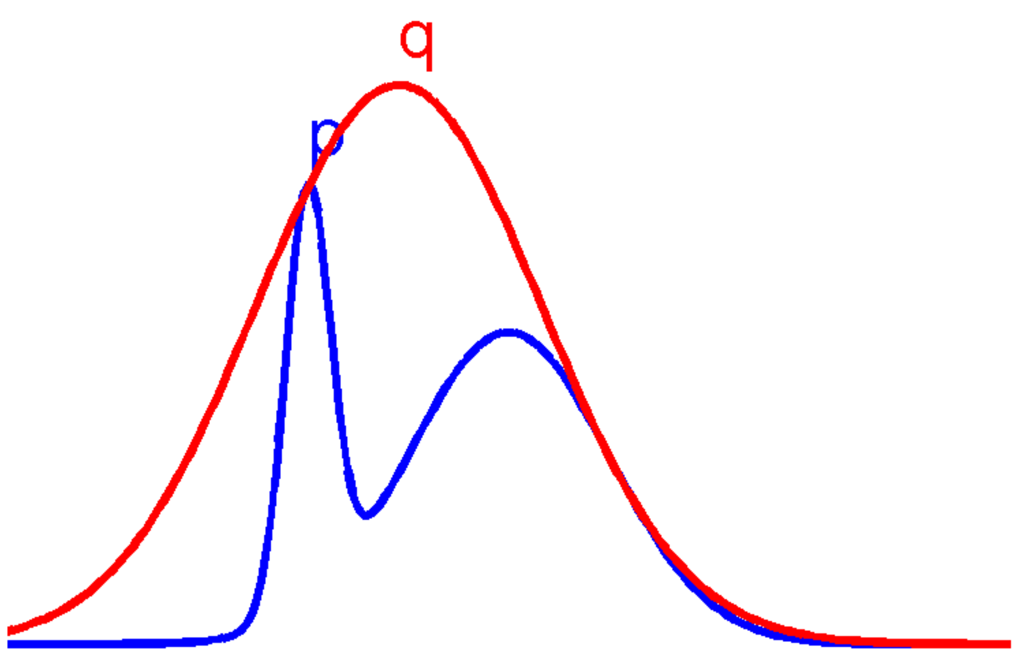
q is Gaussian, minimizes $D_\alpha(p||q)$



$\alpha = 1$

Minimum alpha-divergence

q is Gaussian, minimizes $D_\alpha(p||q)$



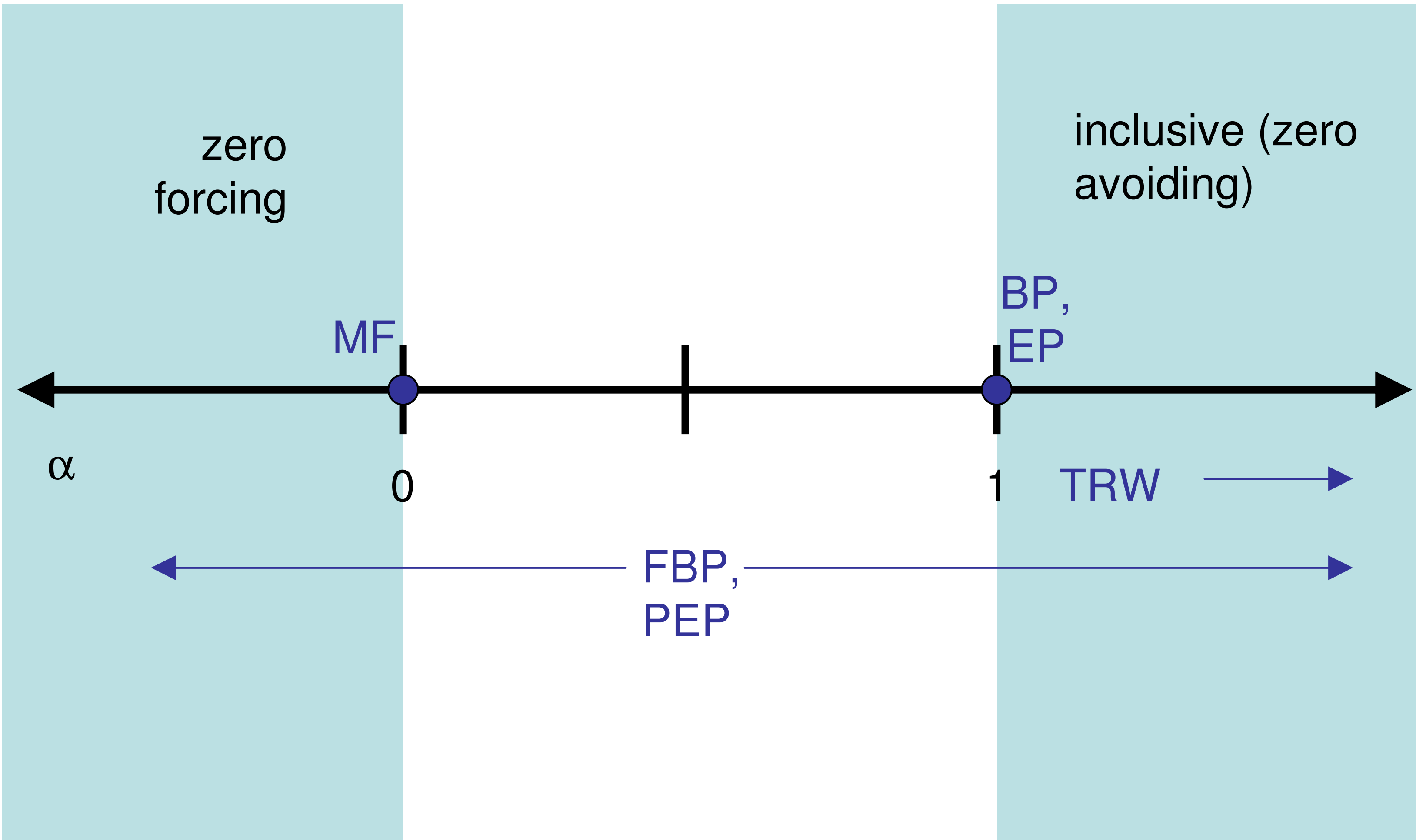
$\alpha = \infty$

Properties of alpha-divergence

- $\alpha \leq 0$ seeks the mode with largest mass (not tallest)
  - *zero-forcing*: p(x)=0 forces q(x)=0
  - underestimates the support of p
- $\alpha \geq 1$ stretches to cover everything
  - *inclusive*: p(x)>0 forces q(x)>0
  - overestimates the support of p

[Frey,Patrascu,Jaakkola,Moran 00]

Structure of alpha space



Other properties

- If q is an exact minimum of alpha-divergence:
- Normalizing constant:

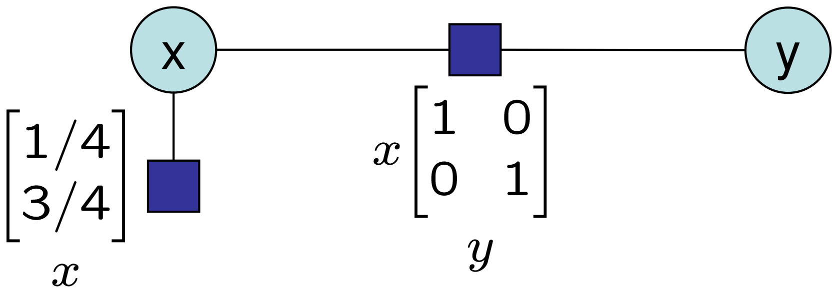
$$\int q(x)dx \leq \int p(x)dx \qquad \text{if } \alpha < 1$$

$$\int q(x)dx = \int p(x)dx \qquad \text{if } \alpha = 1$$

$$\int q(x)dx \geq \int p(x)dx \qquad \text{if } \alpha > 1$$

- If α=1: Gaussian q matches mean,variance of p
  - Fully factorized q matches marginals of p

Two-node example



$$p(x,y) = \underset{x}{\begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}} \quad \begin{array}{cc} x & \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ & y \end{array} \qquad\qquad p(y) = \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}$$

$$q(x,y) = \underset{x}{\begin{bmatrix} a \\ b \end{bmatrix}} \quad \underset{y}{\begin{bmatrix} c \\ d \end{bmatrix}}$$

- q is fully-factorized, minimizes α-divergence to p
- q has correct marginals only for $\alpha = 1$ (BP)

Two-node example

Bimodal distribution

$$p(x,y) = \begin{array}{cc} & y \\ x & \begin{bmatrix} 1/4 & 0 \\ 0 & 3/4 \end{bmatrix} \end{array}$$

$\alpha = 1$ (BP)

$$q(x,y) = \underset{x}{\begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}} \underset{y}{\begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}} = \begin{array}{cc} & y \\ x & \begin{bmatrix} 1/16 & 3/16 \\ 3/16 & 9/16 \end{bmatrix} \end{array}$$

$\alpha = 0$ (MF)

$\alpha \leq 0.5$

$$q(x,y) = \underset{x}{\begin{bmatrix} 0 \\ \sqrt{3}/2 \end{bmatrix}} \underset{y}{\begin{bmatrix} 0 \\ \sqrt{3}/2 \end{bmatrix}} = \begin{array}{cc} & y \\ x & \begin{bmatrix} 0 & 0 \\ 0 & 3/4 \end{bmatrix} \end{array}$$

Good	Bad
•Marginals •Mass	•Zeros •Peak heights
•Zeros •One peak	•Marginals •Mass

Two-node example

Bimodal distribution

$$p(x,y) = \begin{array}{cc} & y \\ x & \begin{bmatrix} 1/4 & 0 \\ 0 & 3/4 \end{bmatrix} \end{array}$$

$\alpha = \infty$

$$q(x,y) = \underset{x}{\begin{bmatrix} 1/2 \\ \sqrt{3}/2 \end{bmatrix}} \underset{y}{\begin{bmatrix} 1/2 \\ \sqrt{3}/2 \end{bmatrix}} = \begin{array}{cc} & y \\ x & \begin{bmatrix} 1/4 & \sqrt{3}/4 \\ \sqrt{3}/4 & 3/4 \end{bmatrix} \end{array}$$

Good	Bad
•Peak heights	•Zeros •Marginals

Lessons

- Neither method is inherently superior – depends on what you care about
- A factorized approx does not imply matching marginals (only for $\alpha=1$)
- Adding y to the problem can change the estimated marginal for x (though true marginal is unchanged)

Outline

- Example of message passing
- Interpreting message passing
- Divergence measures
- Message passing from a divergence measure
- Big picture

Distributed divergence minimization

$$p(x,y,z) = \begin{array}{cc} x & \begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix} \\ & y \end{array} \begin{array}{cc} y & \begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix} \\ & z \end{array} \begin{array}{cc} x & \begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix} \\ & z \end{array} \begin{array}{c} \begin{bmatrix}0.4\\ 0.6\end{bmatrix} \\ x \end{array} \begin{array}{c} \begin{bmatrix}0.7\\ 0.3\end{bmatrix} \\ y \end{array} \begin{array}{c} \begin{bmatrix}0.8\\ 0.2\end{bmatrix} \\ z \end{array}$$

$$qq(x,y,z) = \begin{array}{c} \begin{bmatrix}.59\\ .41\end{bmatrix} \\ x \end{array} \begin{array}{c} \begin{bmatrix}.5\\ .5\end{bmatrix} \\ y \end{array} \begin{array}{c} \begin{bmatrix}.6\\ .4\end{bmatrix} \\ y \end{array} \begin{array}{c} \begin{bmatrix}.57\\ .43\end{bmatrix} \\ z \end{array} \begin{array}{c} \begin{bmatrix}.61\\ .39\end{bmatrix} \\ x \end{array} \begin{array}{c} \begin{bmatrix}.5\\ .5\end{bmatrix} \\ z \end{array} \begin{array}{c} \begin{bmatrix}0.4\\ 0.6\end{bmatrix} \\ x \end{array} \begin{array}{c} \begin{bmatrix}0.7\\ 0.3\end{bmatrix} \\ y \end{array} \begin{array}{c} \begin{bmatrix}0.8\\ 0.2\end{bmatrix} \\ z \end{array}$$

$$q(x,y,z) = \begin{array}{c} \begin{bmatrix}0.61\\ 0.39\end{bmatrix} \\ x \end{array} \begin{array}{c} \begin{bmatrix}0.78\\ 0.22\end{bmatrix} \\ y \end{array} \begin{array}{c} \begin{bmatrix}0.84\\ 0.16\end{bmatrix} \\ z \end{array}$$

Distributed divergence minimization

- Write p as product of factors:

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

- Approximate factors one by one:

$$t_a(x) \to \tilde{t}_a(x)$$

- Multiply to get the approximation:

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

Global divergence to local divergence

- Global divergence:

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

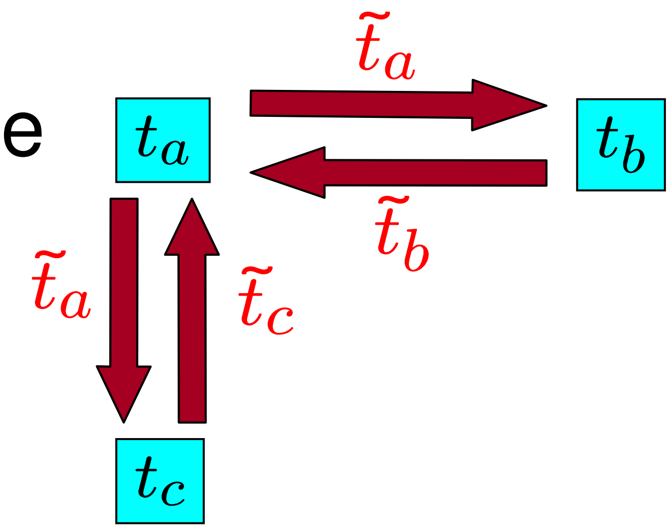
$$D(t_a(x) \prod_{b \neq a} t_b(x) \mid\mid \tilde{t}_a(x) \prod_{b \neq a} \tilde{t}_b(x))$$

- Local divergence:

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

Message passing

- Messages are passed between *factors*
- Messages are factor approximations: $\tilde{t}_a(x)$
- Factor a receives $\tilde{t}_b(x), b \neq a$
  - Minimize local divergence to get $\tilde{t}_a(x)$
  - Send to other factors
  - Repeat until convergence
- Produces all 6 algs



Global divergence vs. local divergence

In general, local ≠ global

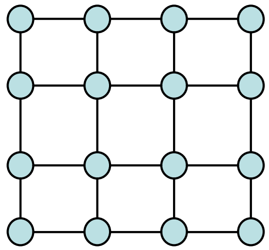
- but results are similar
- BP doesn't minimize global KL, but comes close

Experiment

- Which message passing algorithm is best at minimizing global $D_\alpha(p||q)$?
- Procedure:

1. Run FBP with various α_L
2. Compute global divergence for various α_G
3. Find best α_L (best alg) for each α_G

Results

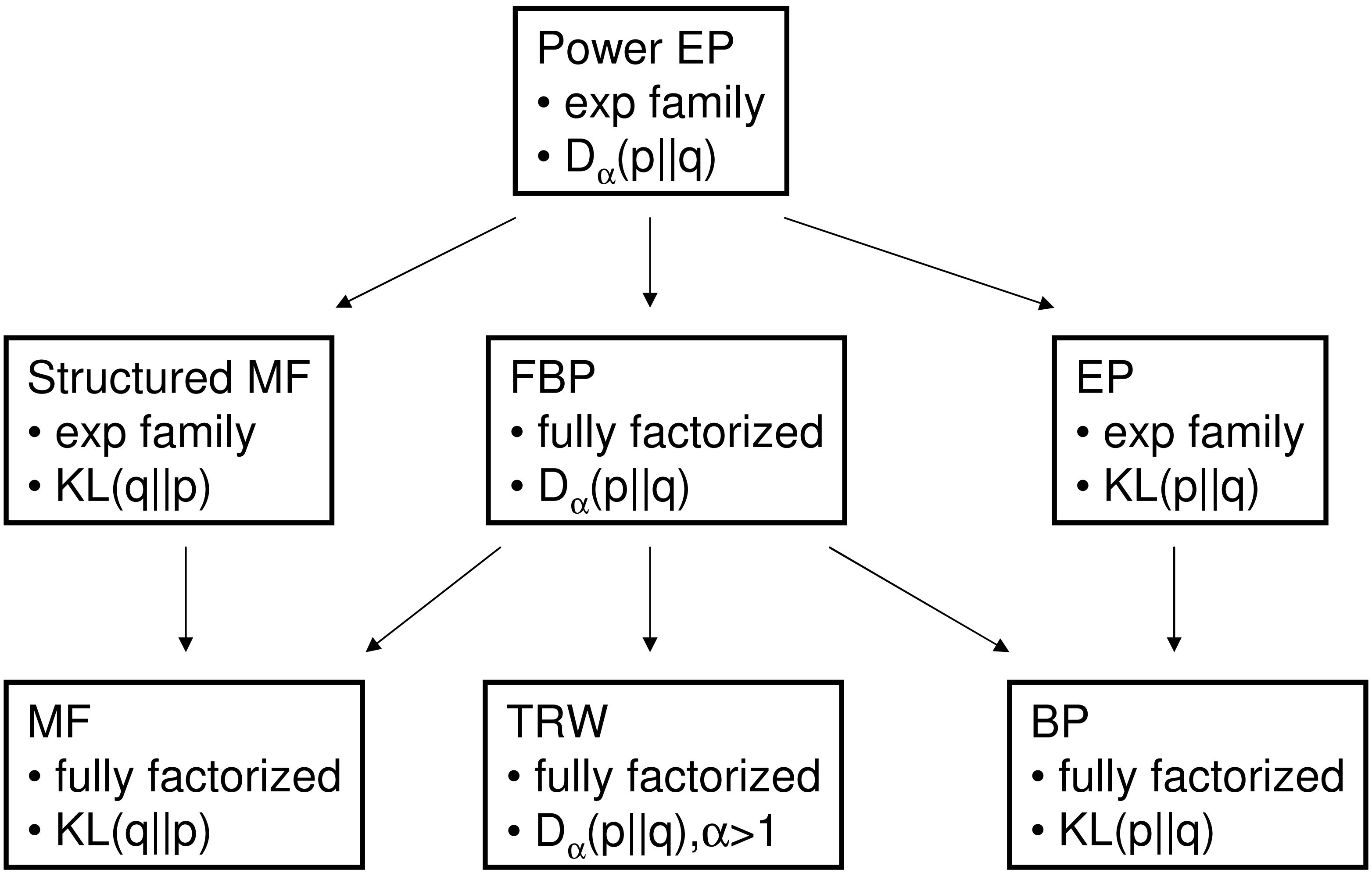


- Average over 20 graphs, random singleton and pairwise potentials: $exp(w_{ij}x_ix_j)$
- Mixed potentials (w ~ U(-1,1)):
  - best $\alpha_L = \alpha_G$ (local should match global)
  - FBP with same α is best at minimizing D_α
    - BP is best at minimizing KL

Outline

- Example of message passing
- Interpreting message passing
- Divergence measures
- Message passing from a divergence measure
- Big picture

Hierarchy of algorithms



Matrix of algorithms

divergence measure

Other divergences?

MF
- fully factorized
- KL(q||p)

Structured MF
- exp family
- KL(q||p)

TRW
- fully factorized
- $D_\alpha(p||q), \alpha>1$

approximation family

Other families? (mixtures)

BP
- fully factorized
- KL(p||q)

EP
- exp family
- KL(p||q)

FBP
- fully factorized
- $D_\alpha(p||q)$

Power EP
- exp family
- $D_\alpha(p||q)$

Other Message Passing Algorithms

Do they correspond to divergence measures?

- Generalized belief propagation [Yedidia,Freeman,Weiss 00]
- Iterated conditional modes [Besag 86]
- Max-product belief revision
- TRW-max-product [Wainwright,Jaakkola,Willsky 02]
- Laplace propagation [Smola,Vishwanathan,Eskin 03]
- Penniless propagation [Cano,Moral,Salmerón 00]
- Bound propagation [Leisink,Kappen 03]

Future work

- Understand existing message passing algorithms
- Understand local vs. global divergence
- New message passing algorithms:
  - Specialized divergence measures
  - Richer approximating families
- Other ways to minimize divergence