Tutorial on variational approximation methods

Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

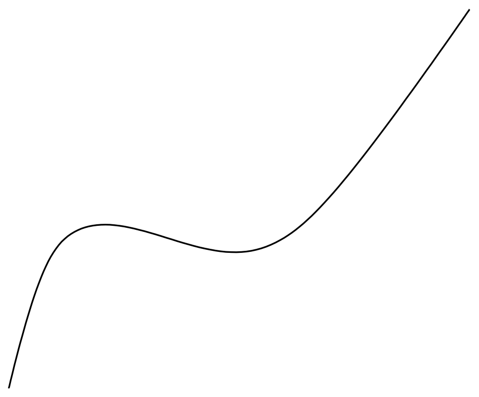
Tutorial topics

- A bit of history
- Examples of variational methods
- A brief intro to graphical models
- Variational mean field theory
  - Accuracy of variational mean field
  - Structured mean field theory
- Variational methods in Bayesian estimation
- Convex duality and variational factorization methods
  - Example: variational inference and the QMR-DT

Variational methods

- Classical setting: “finding the extremum of an integral involving a function and its derivatives”

  Example: finding the trajectory of a particle under external field



- The key idea here is that the problem of interest is formulated as an *optimization problem*

Variational methods cont'd

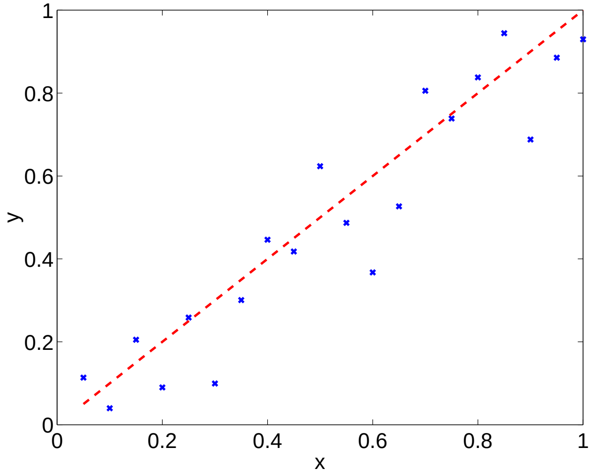
- Variational methods have a long history in physics, statistics, control theory as well as economics.
  - calculus of variations (physics)
  - linear/non-linear moments problems (statistics)
  - dynamic programming (control theory)
- Variational formulations appear naturally also in machine learning contexts:
  - regularization theory
  - maximum entropy estimation
- Recently variational methods been used and further developed in the context of approximate inference and estimation

Examples of variational methods

- In classical examples the formulation itself is given but for us this is one of the key problems
- We provide here a few examples that highlight
  1. how to cast problems as optimization problems
  2. how to find an approximate solution when the exact solution is not feasible
- The examples we use involve
  a) finite element methods for solving differential equations
  b) large deviation methods (Chernoff bound)

Linear regression example

- n examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$



Find the parameter vector $\mathbf{w}$ that minimizes the squared error

$$\min \sum_{t=1}^{n} \left(y_t - \mathbf{w}^T \mathbf{x}_t \right)^2 \;\Rightarrow\; \mathbf{w}^* = \left(\sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^T \right)^{-1} \left(\sum_{t=1}^{n} \mathbf{x}_i y_i \right)$$

Optimization problem desired solution

- We wish to find a (variational) optimization problem given the solution

Example: finite element method

- Consider the following simple 1-D Laplace (or Poisson) differential equation

$$-u''(x) = f(x), \quad x \in (a, b)$$

  with homogeneous boundary conditions $u(a) = u(b) = 0$.

- If $u^*(x)$ denotes the solution, we can recover it by minimizing the following "silly" functional

$$J_{silly}(u) = \int_a^b (u^{*\prime}(x) - u'(x))^2 dx$$

  subject to the boundary conditions.

Finite element method cont'd

- The "silly" functional is useful since we can evaluate the relevant part of it

$$\begin{aligned} J_{silly}(u) &= \int_a^b (u^{*\prime}(x) - u'(x))^2 dx \\ & \cdots \\ &= \text{const.} + \underbrace{\int_a^b u'(x)^2 dx - 2\int_a^b u(x) f(x) dx}_{J(u)} \end{aligned}$$

  where $J(u)$ only depends on $u(x)$ and the constant is defined in terms of the unknown solution $u^*(x)$.

- We typically cannot find the solution exactly but $J(u)$ still permits us to evaluate and rank approximate solutions

Finite element method cont'd

- To find an approximate solution, we specify a set of basis functions $\{\phi_1(x), \ldots, \phi_n(x)\}$ consistent with the boundary conditions and find the "best" solution of the form

$$u(x) = c_1\phi_1(x) + \cdots + c_n\phi_n(x)$$

- Substituting this tentative solution back into our (quadratic) objective $J(u)$ gives a quadratic objective for the coefficients

$$J(c) = \sum_{ij} c_i c_j \underbrace{\left[\int_a^b \phi_i'(x)\phi_j'(x)dx\right]}_{A_{ij}} - 2\sum_i c_i \underbrace{\left[\int_a^b \phi_i(x)f(x)dx\right]}_{r_i}$$

- Minimizing with respect to the coefficients c finally leads to the following "fixed point equations" chracterizing the best approximate solution:

$$Ac = r$$

Summary of the example

- The key steps were:
  1. We defined a natural objective with reference to the optimal solution
  2. We identified the relevant portion of this objective so that it could be evaluated without reference to the unknown solution
  3. The resulting objective permitted us to search for the best approximate solution within a parametric family
  4. The best approximate solution were characterized by a set of fixed point equations

Large deviation example

- Let $x_1, \ldots, x_n$ be a set of zero mean i.i.d. random variables distributed according to a known distribution P.
- In large deviation theory, we are interested in evaluating the probability that the sum $x_1 + \ldots + x_n$ deviates substantially from its mean

$$\text{Prob}\left(\sum_i x_i \geq n\epsilon\right) = E\{\,\underbrace{\text{step}(\sum_i x_i - n\epsilon)}_{\text{event indicator}}\,\}$$

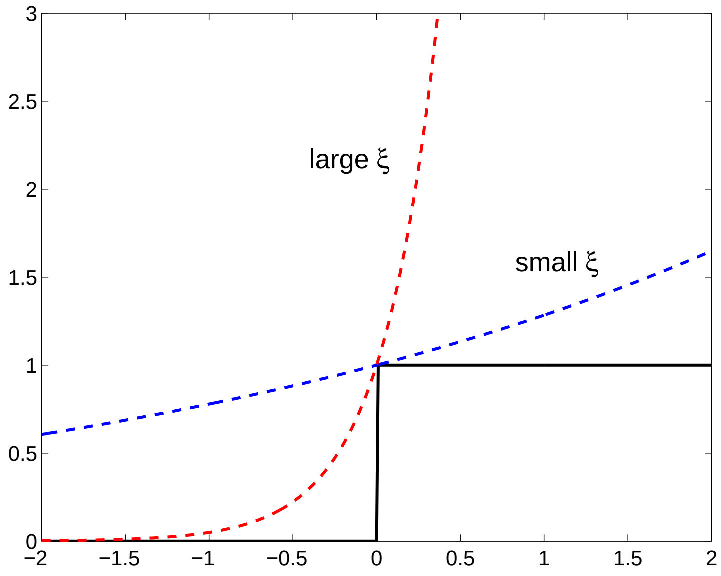
where step$(z) = 1$ for $z \geq 0$ and zero otherwise.

Large deviation example cont'd

$$\text{Prob}\Big(\sum_i x_i \geq n\epsilon\Big) = E\Big\{\underbrace{\text{step}\Big(\sum_i x_i - n\epsilon\Big)}_{\text{event indicator}}\Big\}$$

- To actually evaluate this probability, we make use of the following transformation

$$\text{step}(z) = \min_{\xi \geq 0} e^{\xi z}$$



Large deviation example cont'd

$$\begin{aligned}
\text{Prob}(\sum_i x_i \geq n\epsilon) &= E\{\,\text{step}(\sum_i x_i - n\epsilon)\,\} \\
&= E\{\,\min_\xi\; e^{\xi\,(\sum_i x_i - n\epsilon)}\} \quad \text{(introduce the transform)} \\
&\leq \min_\xi E\{\,e^{\xi\,(\sum_i x_i - n\epsilon)}\} \quad \text{(exchange the operations)}
\end{aligned}$$

where we can finally evaluate the expectation since the argument factors across the independent variables

- "min" and "$E\{\cdot\}$" do not commute and thus we get an upper bound.

Summary of the example

1. Even simple functions (such as the step function) can be cast as optimization problems
2. We can simplify computations by exchanging non-commutative operations ("min" and "$E\{\cdot\}$") and obtain a bound on the quantity of interest
3. The resulting bound can be optimized for accuracy

Note: we could have cast the large deviation example in a form similar to the previous example but this would have required a bit longer explanation...

Probabilistic inference

- Let $P(x_1, \ldots, x_n)$ be the distribution of interest over n variables

  We divide the set of variables into

  1. "visible" variables x_v whose marginal distribution $P(x_v)$ we are interested in computing
  2. "hidden" variables x_h whose posterior distribution $P(x_h|x_v)$ we want

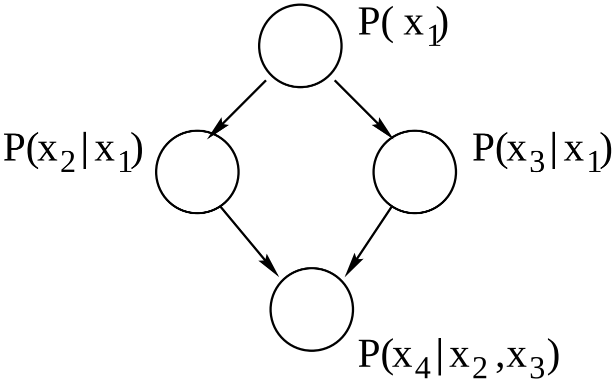
  Evaluating the marginal or posterior involves summing over all configurations of the hidden variables x_h

  $$P(x_v) = \sum_{x_h} P(x_h, x_v)$$

- The complexity of this operation depends on the structure or *factorization* of the joint distribution P
- We try to capture the factorization explicitly in terms of graphs

Graph models: examples

- Bayesian networks:



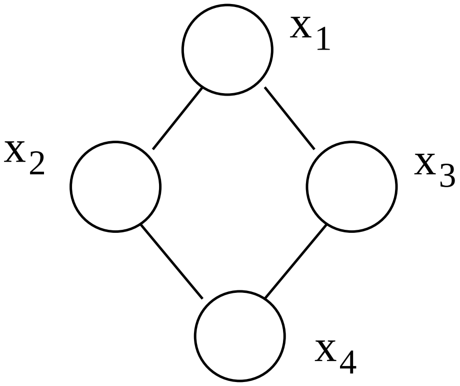
The joint distribution factors according to nodes given their "parents" in the graph

$$\begin{aligned} P(x_1,\ldots,x_4) &= P(x_1)\,P(x_2|x_1)\,P(x_3|x_1)\,P(x_4|x_2,x_3) \\ &= \phi_1(x_1)\,\psi_{12}(x_1,x_2)\,\psi_{13}(x_1,x_3)\,\psi_{234}(x_2,x_3,x_4) \end{aligned}$$

- Normalization is local, i.e., each component in the product is properly normalized

Graph models: examples

- Markov random fields (undirected graph models):



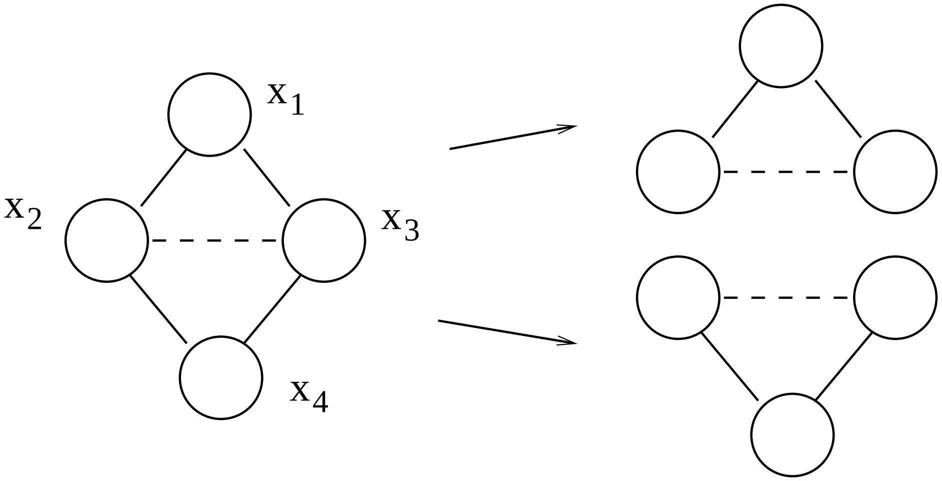
The joint distribution factors across the "cliques" in the graph

$$P(x_1, \ldots, x_4) = \frac{1}{Z} \psi_{12}(x_1, x_2)\, \psi_{13}(x_1, x_3)\, \psi_{24}(x_2, x_4)\, \psi_{34}(x_3, x_4)$$

- Normalization is global

Graph models: complexity

- The complexity of exact inference in the graph model depends on the clique sizes of the *triangulated* graph

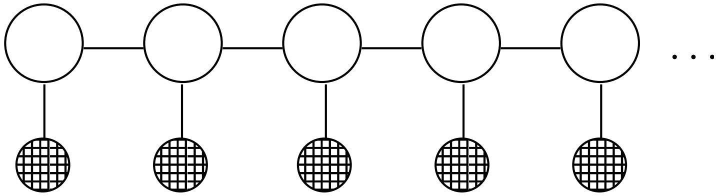


$$P(x_1, \ldots, x_4) = \psi_{123}(x_1, x_2, x_3)\, \psi_{234}(x_2, x_3, x_4)$$

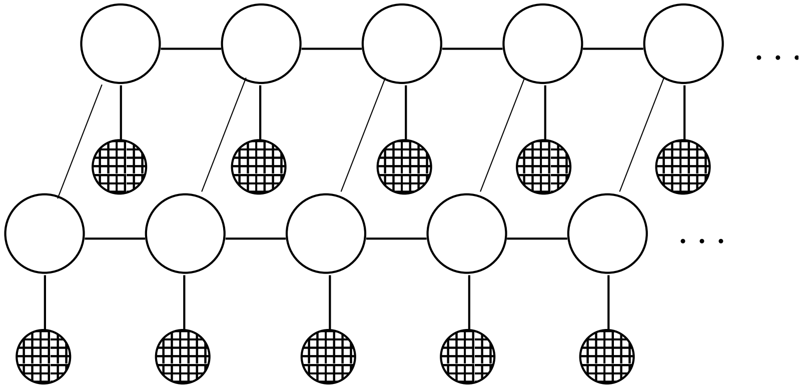
(the clique size here is three)

Specific examples

HMMs and coupled HMMs



The clique size here is 2 and the graph is already triangulated



The clique size is 2; three after triangulation

- the shaded nodes denote “visible” variables
- the unshaded nodes are “hidden” variables

Exact methods for inference

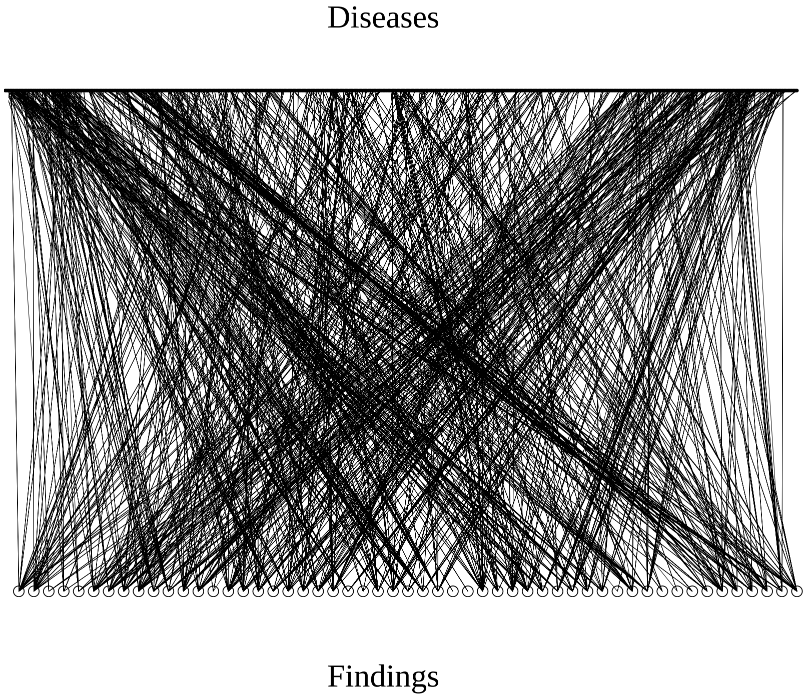
- Forward-backward (chains)
- Belief propagation (trees)
- Junction tree algorithm (clique trees)

All the technigues heavily exploit the graph structure

BUT...

Example: medical diagnosis

- The QMR database (Middleton et al.)



- Probabilistic inference is in general NP-hard
- Even approximate inference is in general NP-hard

Approximate methods

- Belief propagation (even for graphs with cycles)
- Search algorithms
- Sampling methods
- Variational methods

Our focus is on variational methods...

Approximate methods and graph structure

- Approximate inference relies on additional factorization or structure in the probability model

  1. small clique size in the original, not triangulated graph (e.g., in coupled HMMs)
  2. parametric conditional probabilities as in the noisy-OR model

$$\begin{aligned} P(x_4 = 0|x_2, x_3) &= e^{-(\theta_0+\theta_1 x_2+\theta_2 x_3)} \\ P(x_4 = 1|x_2, x_3) &= 1 - e^{-(\theta_0+\theta_1 x_2+\theta_2 x_3)} \end{aligned}$$

  (written here in an exponentiated form)

- There's little one can do (in general) if the model has large (unstructured) probability tables associated with the cliques

Probabilistic inference problem again

- $P(x_1, \ldots, x_n)$ is the distribution of interest and the variables are divided into sets of

  1. "visible" variables x_v whose marginal distribution $P(x_v)$ we are interested evaluating and
  2. "hidden" variables x_h whose posterior distribution $P(x_h|x_v)$ we want

  Evaluating the marginal or posterior involves summing over all configurations of the hidden variables x_h

$$P(x_v) = \sum_{x_h} P(x_h, x_v)$$

- To find a variational solution we need to cast these computations as optimization problems

Variational probabilistic inference

- We start with a "silly" objective function:

$$J(Q) = -KL(Q_{x_h} \| P_{x_h|x_v}) + \log P(x_v)$$

  where Q is a *variational distribution* over the hidden variables x_h.

1) If $Q(x_h) = P(x_h|x_v)$, the posterior probability over the hidden variables x_h, we recover the desired log-marginal $\log P(x_v)$.

2) $\log P(x_v) \geq J(Q)$ for all Q, i.e., the objective function is a *lower bound* on the desired quantity.

3) The slack in the bound is given by the KL-divergence between Q and the true posterior.

Variational probabilistic inference cont'd

- This objective function can be rewritten without reference to the posterior or the marginal

$$\begin{aligned} J(Q) &= -KL(Q_{x_h} \| P_{x_h|x_v}) + \log P(x_v) \\ & \quad \ldots \\ &= \sum_{x_h} Q(x_h) \log P(x_h, x_v) + H(Q) \end{aligned}$$

  where $H(Q)$ is the (Shannon) entropy of Q.

- Taking the logarithm of the joint distribution is attractive because of the factorization

$$\log P(x_h, x_v) = \log \prod_{c \in C} \psi_c(x_c) = \sum_{c \in C} \log \psi_c(x_c)$$

Variational mean field method

- The variational formulation is exact and thus not (yet) manageable
- We can, however, restrict the maximization of

  $$J(Q) = \sum_{x_h} Q(x_h) \log P(x_h, x_v) + H(Q)$$

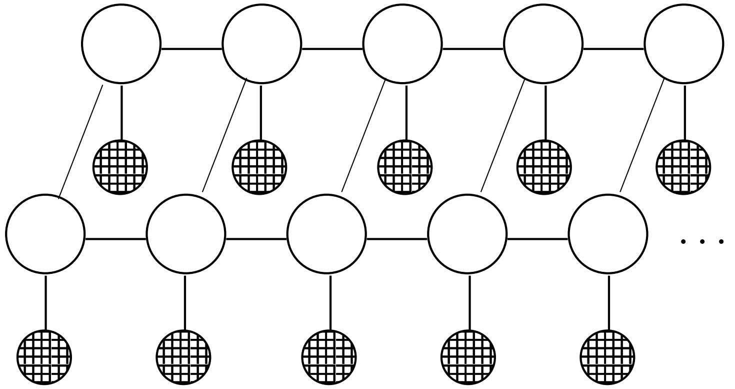
  to some manageable class of distributions Q (cf. finite element methods)

- The simplest choice is the class of completely factored or *mean field* distributions

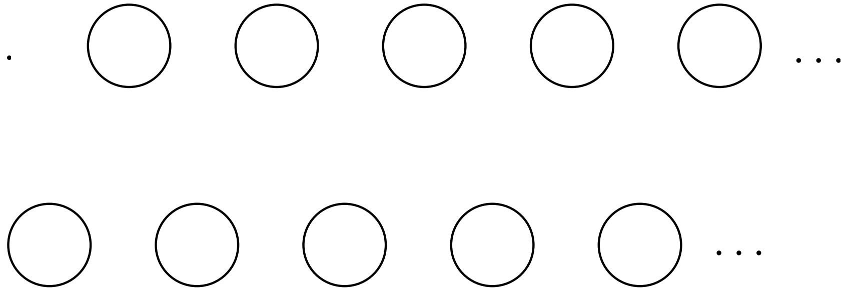
  $$Q(x_h) = \prod_{i \in h} Q_i(x_i)$$

Variational mean field method cont'd

- We want to approximate a complicated graph model with a simple factored model



Coupled HMMs



factored model

- The measure of distance between the two models is the KL-divergence

$$KL(Q_{x_h} \| P_{x_h|x_v})$$

Variational mean field method: updates

- How do we optimize the simple variational distribution?
- We derive mean field equations (updates) by maximizing

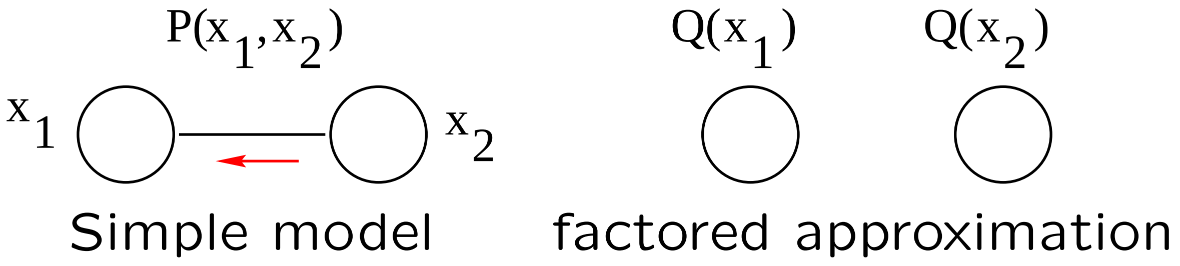
$$J(Q) = \sum_{x_h} Q(x_h) \log P(x_h, x_v) + H(Q)$$

with respect to each of the marginals $Q_i(x_i)$ in

$$Q(x_h) = \prod_{i \in h} Q_i(x_i)$$

while keeping the remaining marginals fixed.

Variational mean field updates cont'd



- The updates are made "locally" by geometrically averaging the ignored dependencies with respect to the other component distributions

$$Q_1(x_1) \;\leftarrow\; \frac{1}{Z_1} \times \exp\big(\sum_{x_2} Q_2(x_2) \log P(x_1, x_2) \big)$$

  where Z_1 normalizes the right hand side across x_1.

  In other words, each x_i only sees the "mean effect" of the dependencies

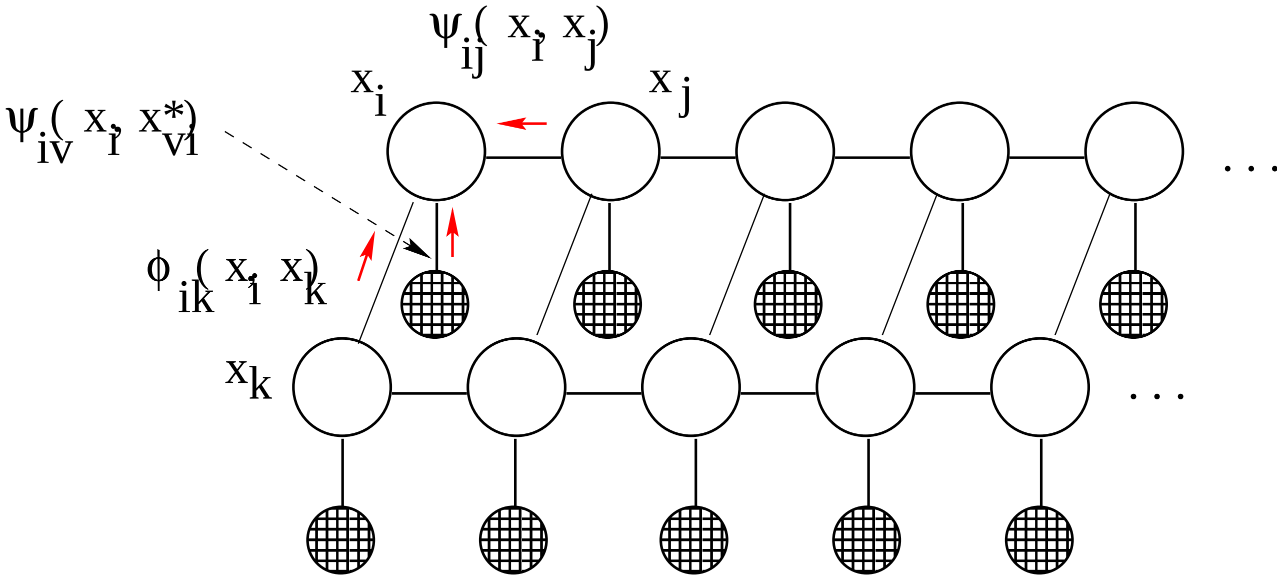
- These update equations are the "fixed point equations"

Variational mean field updates cont'd

- Example: coupled HMMs

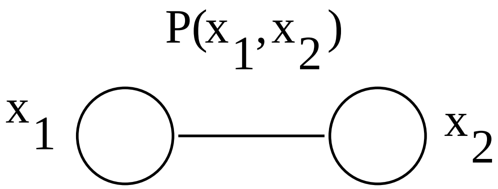
$$Q_i(x_i) \;\leftarrow\; \frac{1}{Z_i} \times \overbrace{\psi_{iv}(x_i, x^*_{vi})}^{\text{evidence at } i} \times \overbrace{\exp\Big(\sum_{x_j} Q_j(x_j) \log \psi_{ij}(x_i, x_j)\Big)}^{\text{within chain dependencies}}$$

$$\times \underbrace{\exp\Big(\sum_{x_k} Q_k(x_k) \log \phi_{ik}(x_i, x_k)\Big)}_{\text{between chains dependencies}}$$



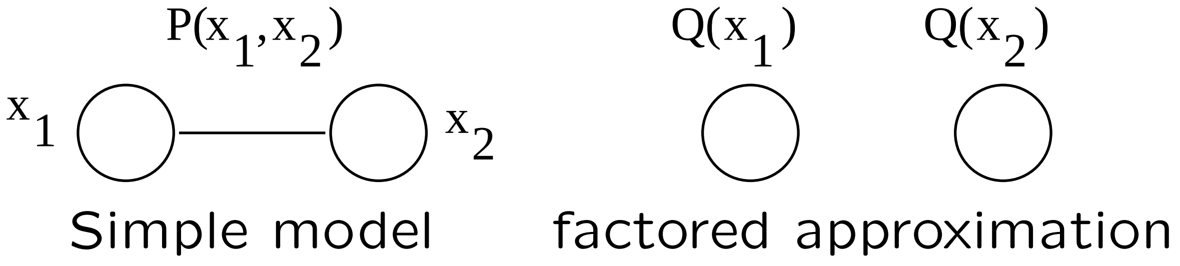
Accuracy of variational mean field

- General thoughts:
  - weak dependence (near independence) ⇒ good accuracy
  - accurate even for large systems with weak interactions (law of large numbers)
  - strong dependencies ⇒ poor approximation
- We'll analyze the accuracy a bit in the context of the following simple example



(no visible nodes)

Simple example



- Since there are no visible variables the log-marginal we are trying to approximate is

$$\log \sum_{\mathbf{x}_h} P(x_1, x_2) = \log \sum_{x_1, x_2} P(x_1, x_2) = \log 1 = 0$$

- The variational mean field approximation does not achieve this value, however, but rather

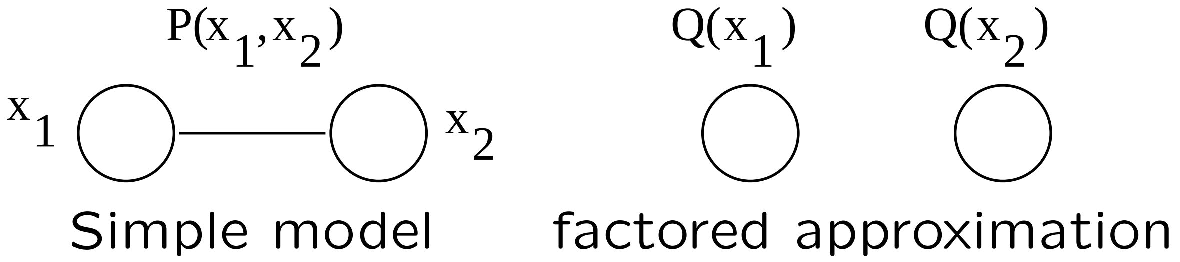
$$J(Q) = -KL\left(Q_{x_1,x_2} \,\|\, P_{x_1,x_2} \right) + 0$$

unless the factored approximation

$$Q(x_1, x_2) = Q_1(x_1)\, Q_2(x_2)$$

equals $P(x_1, x_2)$

Simple example cont'd

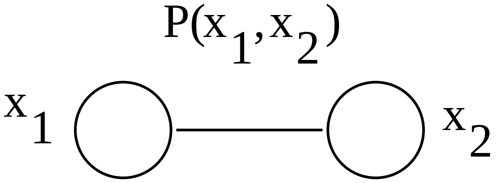


- There are two notions of accuracy:

  1. How close the lower bound $J(Q)$ is to zero (or more generally to the log-marginal $\log P(x_v)$)
  2. How close $Q_1(x_1)$ and $Q_2(x_2)$ are to the true marginals $P_1(x_1)$ and $P_2(x_2)$

- The two notions of accuracy need not be strongly correlated (and indeed they are not)

Simple example cont'd



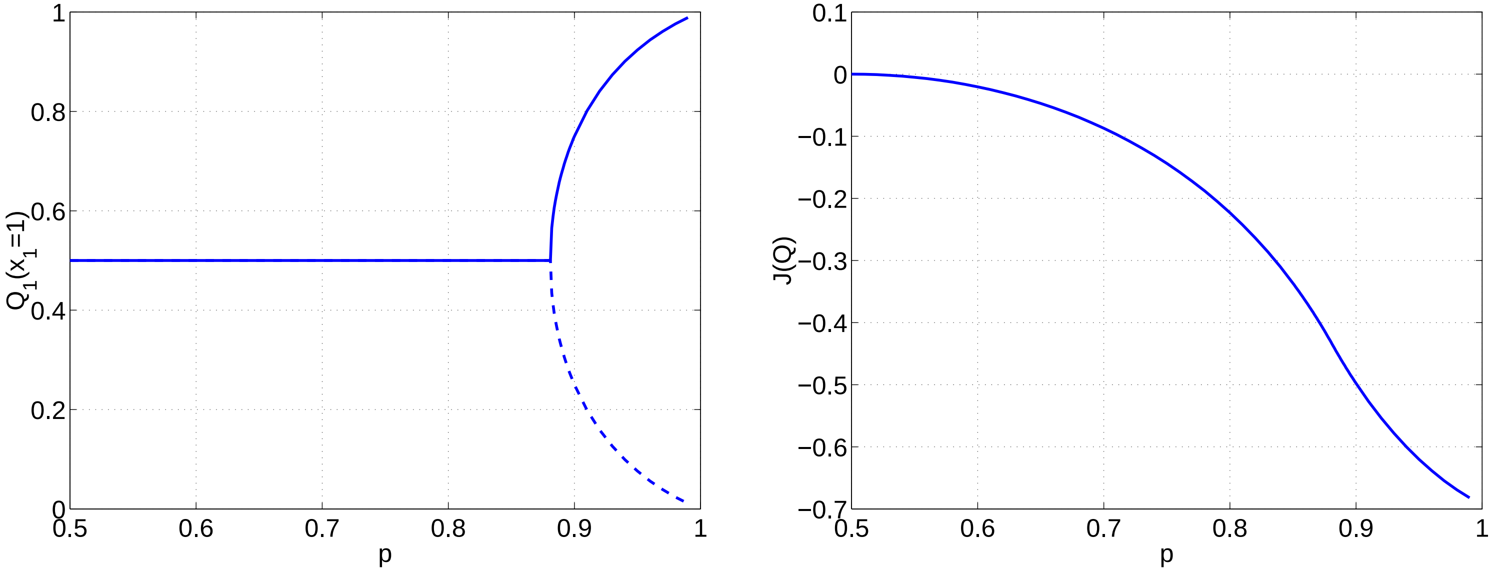
- We introduce a single parameter p that characterizes the dependence between x_1 and x_2 in a simple symmetric model:

$P(0,0) = (1-p)/2$	$P(0,1) = p/2$
$P(1,0) = p/2$	$P(1,1) = (1-p)/2$

  Note that the marginals $P_1(x_1)$ and $P_2(x_2)$ are uniform (equal to 0.5) regardless of the value of p.

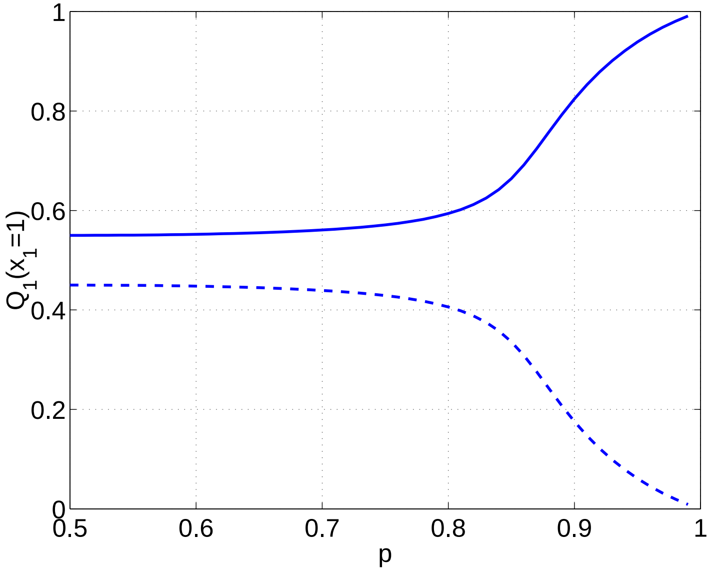
- We might expect that increasing $p \geq 0.5$ gradually degrades the accuracy of the variational approximation

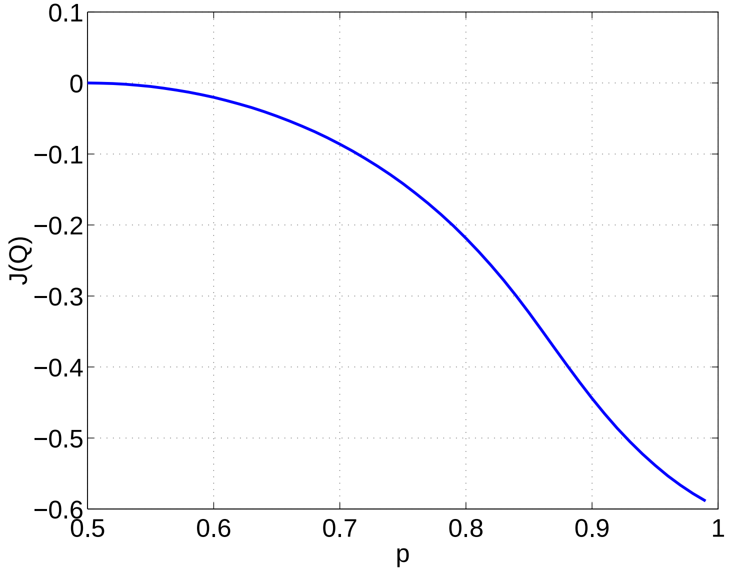
Simple example: results



- The variational marginals undergo spontaneous symmetry breaking as the dependence increases.
- The rate at which the lower bound $J(Q)$ deteriorates slows down when the symmetry is broken

Simple example: results without symmetry





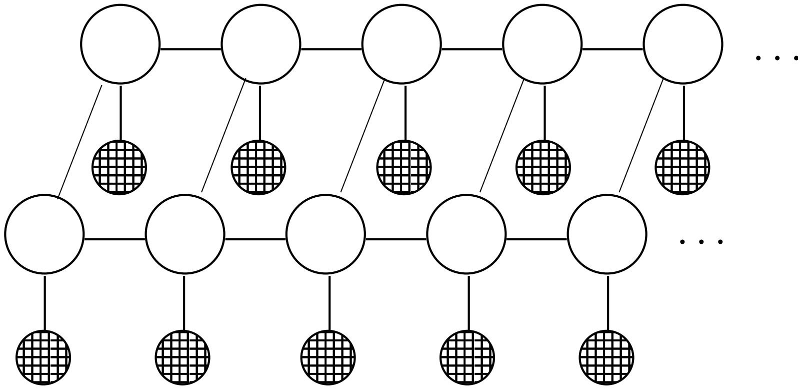
- The phase transition disappears when the problem is no longer exactly symmetrical (e.g., $P_1(x_1 = 1) > 0.5$)
- The variational marginals still deteriorate rapidly after some critical point p^*

Summary of the simple example

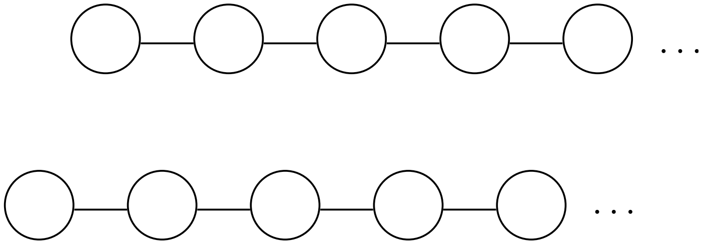
- What can we conclude?
  - The variational marginals remain accurate when the dependencies are relatively weak
  - Larger models offer more opportunities for degradation (symmetries, frustration, explaining away etc.)
  - The variational lower bound can remain fairly accurate even in larger systems
- We should be able improve both forms of accuracy if we dispense with the completely factored (mean field) assumption

Structured mean field approach

- In a structured mean field approach, we want to approximate a complicated graph model with a *tractable graph model*



Coupled HMMs



Structured approximation

- The measure of distance between the two models is the KL-divergence as before

$$KL(Q_{x_h} \| P_{x_h|x_v})$$

Structured mean field approach cont'd

- We can restrict the maximization of

$$J(Q) = \sum_{x_h} Q(x_h) \log P(x_h, x_v) + H(Q)$$

to *any* manageable class of distributions Q.

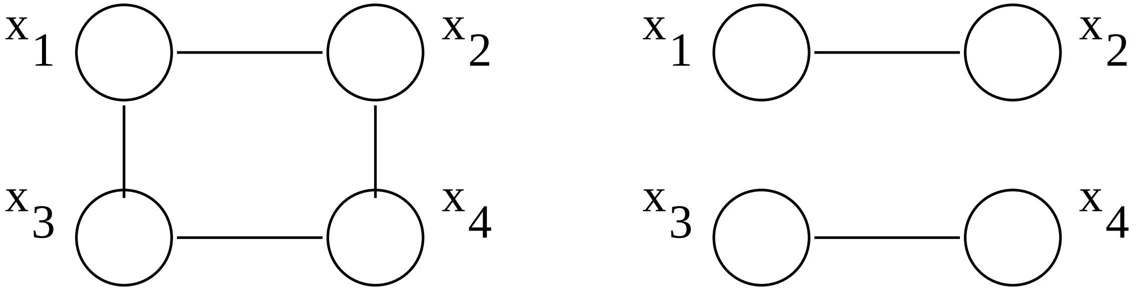
- In a structured approximation, we partition the hidden variables x_h into disjoint sets of variables $x_{h_1}, \ldots, x_{h_m}$
- The variational distribution now factors across these sets but no constraints are imposed within the sets

$$Q(x_h) = \prod_{i=1}^{m} Q_{h_i}(x_{h_i})$$

As a result we incorporate any interactions within each partition exactly while performing a mean field approximation across the partitions

Structured mean field method

- Simple example (all hidden variables):

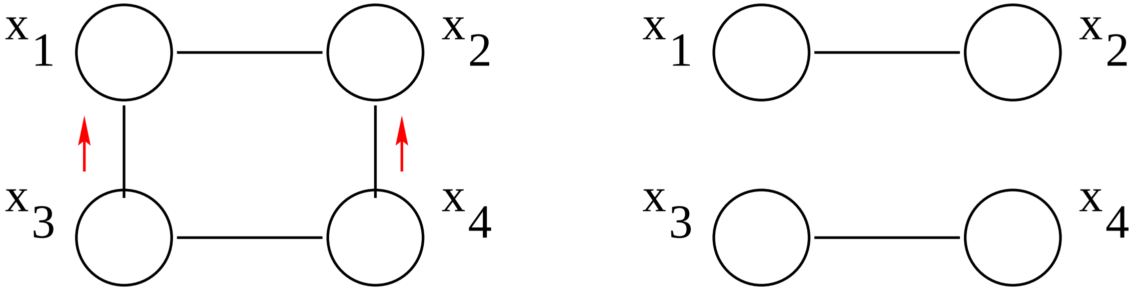


Original graph structured approximation

$$
\begin{aligned}
P(x) &= \frac{1}{Z}\psi_{12}(x_1,x_2)\psi_{23}(x_1,x_3)\psi_{34}(x_3,x_4)\psi_{42}(x_4,x_2) \\
Q(x) &= Q_{12}(x_1,x_2)Q_{34}(x_3,x_4)
\end{aligned}
$$

Example cont'd

- Simple example (only hidden variables):



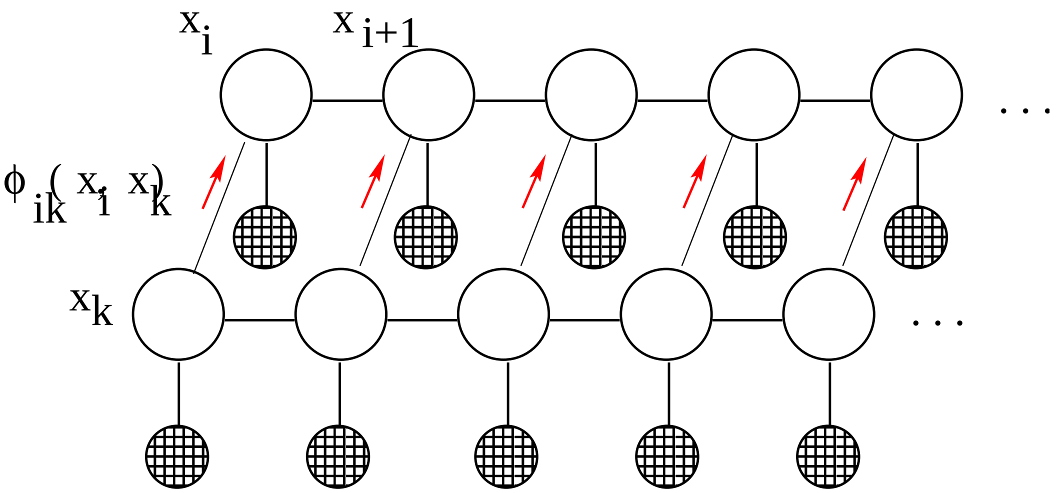
Original graph structured approximation

$$
\begin{aligned}
Q_{12}(x_1,x_2) \leftarrow \frac{1}{Z_{12}} \overbrace{\psi_{12}(x_1,x_2)}^{\text{exact potential}} \quad & \times \overbrace{\exp(\sum_{x_3} Q_{34}(x_3) \log \psi_{13}(x_1,x_3))}^{\text{mean 1-3 dependence}} \\
& \times \underbrace{\exp(\sum_{x_4} Q_{34}(x_4) \log \psi_{42}(x_4,x_2))}_{\text{mean 2-4 dependence}}
\end{aligned}
$$

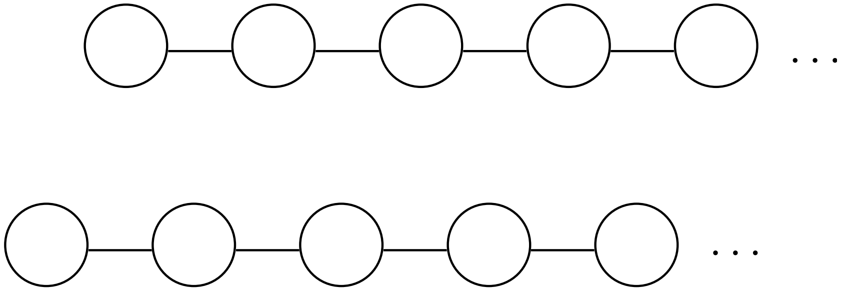
Example: coupled HMMs

- Factored approximation across chains: $Q(x_h) = Q_{h_1}(x_{h_1})\, Q_{h_2}(x_{h_2})$

$$\begin{aligned} Q_{h_1}(x_{h_1}) \leftarrow \frac{1}{Z_{h_1}} \quad & \times \quad (\text{ first HMM model }) \\ & \times \quad \underbrace{\prod_i \exp(\sum_{x_k} Q_{h_2}(x_k) \log \phi_{ik}(x_i, x_k)\,)}_{\text{mean dependencies between the chains}} \end{aligned}$$



Coupled HMMs



Structured approximation

Extensions

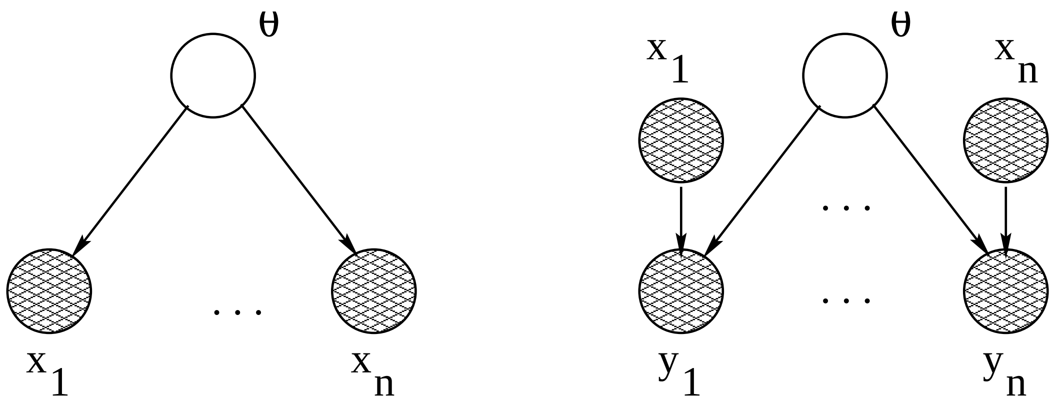
1. The components need not be disjoint (e.g., spanning trees)
2. We can use directed graph models as variational approximating distributions
3. etc.

- These extensions require additional search for the structure of the approximating distribution (difficult)

Bayesian estimation

- Bayesian estimation is *inference*
- In principle, the variational methods we have discussed so far should be applicable to such inference problems
- We distinguish here two cases of Bayesian estimation
  1. no hidden variables
  2. with hidden variables
- The case with hidden variables is often considerably more involved (the posterior has multiple modes)

Bayesian estimation without hidden variables



Density estimation   Classification/regression

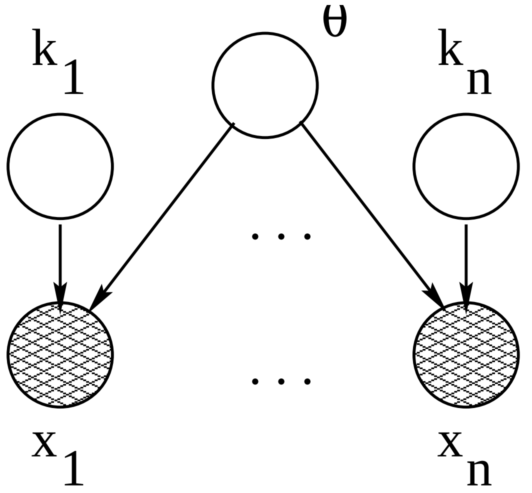
- The mean field/structured mean field approach does not directly help us here since there's no additional (explicit) factorization of θ

- We could, alternatively, impose parametric constraints on the variational posterior distribution $Q(\theta)$ (e.g., Gaussian).

- We may also use variational methods to impose additional factorization (discussed later)

Bayesian estimation with hidden variables

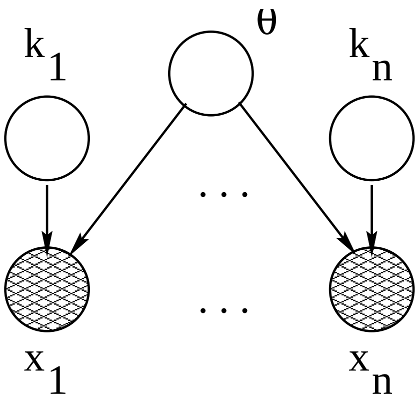
- a Bayesian mixture model



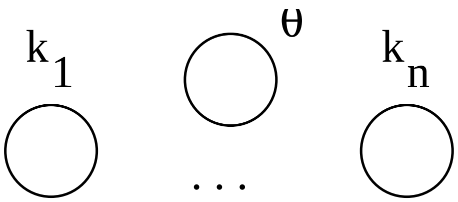
$$P(x, k, \theta) = \left[\prod_{i=1}^{n} P(k_i)\, P(x_i|k_i, \theta) \right] P(\theta)$$

where we assume for simplicity that the parameters θ only influence the conditionals $P(x_i|k_i, \theta)$ and not the mixing proportions $P(k_i)$.

Bayesian mixture model cont'd



mixture model



factored approximation

- Our factored approximation is

$$Q(k, \theta) = Q(k)Q(\theta) = \left[\prod_{i=1}^{n} Q_i(k_i)\right] Q(\theta)$$

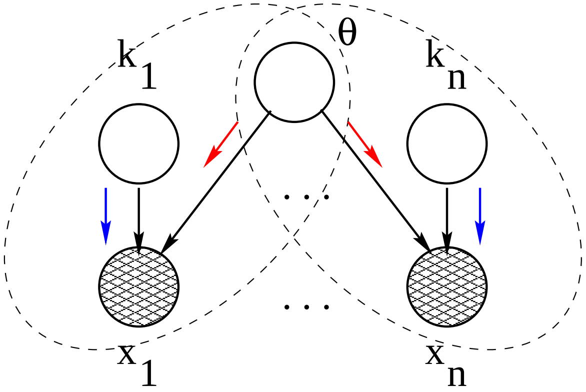
Bayesian mixture model: updates

1. Variational posterior over mixing variables

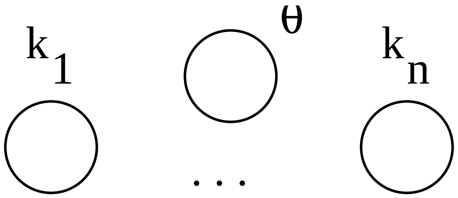
$$Q_i(k_i) \leftarrow \frac{1}{Z_i} \times P(k_i) \times \overbrace{\exp(\int Q(\theta) \log P(x_i|k_i,\theta) d\theta \,)}^{\text{mean } \theta\text{–}k_i \text{ dependence}}$$

2. Variational posterior over parameters

$$Q(\theta) \leftarrow \frac{1}{Z} \times P(\theta) \times \prod_{i=1}^{n} \overbrace{\exp(\sum_{k_i} Q_i(k_i) \log P(x_i|k_i,\theta) \,)}^{\text{mean } \theta\text{–}k_i \text{ dependence}}$$



mixture model



factored approximation

Bayesian mixture model: discussion

- Imposing any factorization of the form

  $$Q(k,\theta) = Q(k)Q(\theta)$$

  makes the solution liable to sudden deterioration of the component marginals (as in the context of the simple example presented earlier)

- This may happen if the mixing variables and the parameters are strongly dependent in the true posterior

- Since in Bayesian estimation we really want the marginals (or at least $Q(\theta)$), we need to be a bit more careful...

Additional factorization

- Success of approximate inference/estimation depends in part on additional structure in the probability model beyond what is visible in the graph (pairwise potentials, parametric conditional probabilities, etc.)
- We can rely on such additional structure to impose further factorization of the joint distribution
- Useful factorization "decouples" dependent variables as in

$$P(x_i|x_{pa_i}) \approx \prod_{j \in pa_i} \psi_{ij}(x_i, x_j)$$

  where x_{pa_i} denotes the "parents" of x_i in a Bayesian network.
- When can we impose such factorization? in a principled way?

Imposing additional factorization cont'd

Theorem (factorization theorem)

Assuming x_i and all its parents x_{pa_i} are discrete variables, then we can **always** *find a variational transform of the form*

$$P(x_i|x_{pa_i}) = \min_{\xi} \left\{ \prod_{j \in pa_i} \psi_{ij}(x_i, x_j; \xi) \right\}$$

(analogous formulation holds for "max")

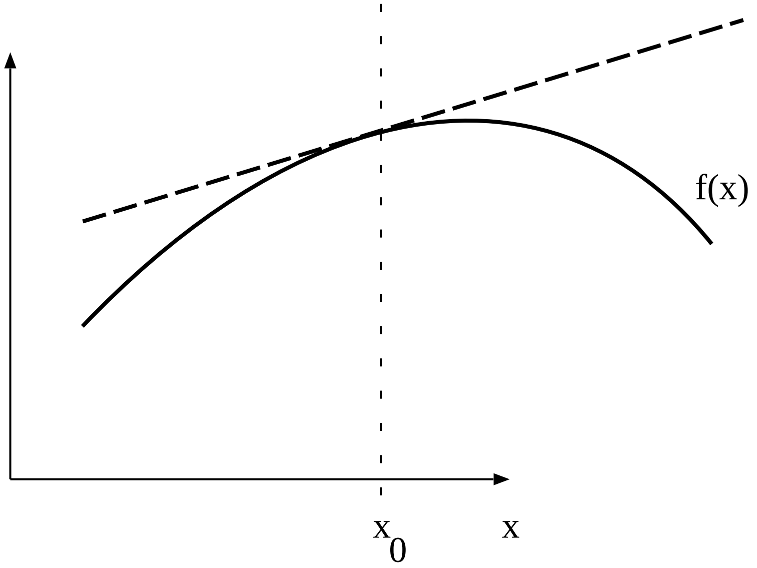
- Unfortunately the mere existence of such transforms is not sufficient; we need transforms that lead to accurate approximations and they are much harder to find

Factorization via convex duality

- Basic idea: for a concave function $f(x)$

  any tangent plane serves as a bound

  $$f(x) \leq \lambda\, x - f^*(\lambda)$$



- The bound is exact for one of the tangents

  $$f(x) = \min_{\lambda} \{\, \lambda\, x - f^*(\lambda)\, \}$$

  (the conjugate function $f^*(\lambda)$, also concave, can be similarly expressed in terms of $f(x)$, hence the duality)

Factorization via convex duality: example

- Factorization transform for the noise-OR model

$$
\begin{aligned}
P(f=1|d,\theta) &= \left[1 - e^{-(\,\theta_1 d_1 + \ldots + \theta_n d_n\,)}\right] \\
&= e^{\log\left[1 - e^{-(\,\theta_1 d_1 + \ldots + \theta_n d_n\,)}\right]} \\
&\leq e^{\lambda(\,\theta_1 d_1 + \ldots + \theta_n d_n\,) - f^*(\lambda)}
\end{aligned}
$$

  since $f(x) = \log\left[1 - e^{-x}\right]$ is a concave (convex down) function.

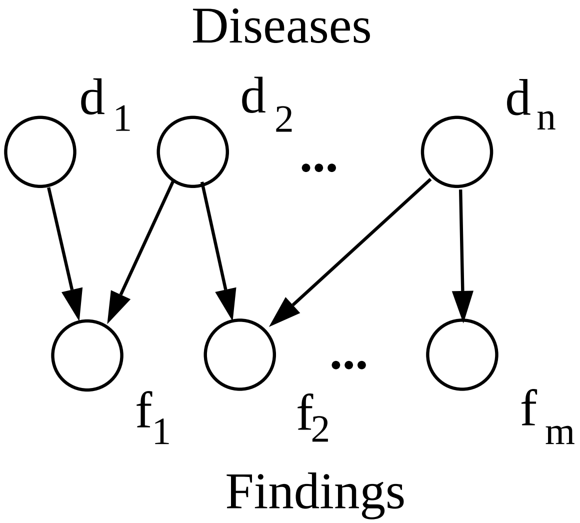
- This is exactly the type of factorization we have discussed

$$
P(f=1|d,\theta) = \min_{\lambda} \left\{ \prod_{j=1}^{n} e^{\lambda\,\theta_j d_j - f^*(\lambda)/n} \right\}
$$

- Similar transforms can be derived for logistic and other models

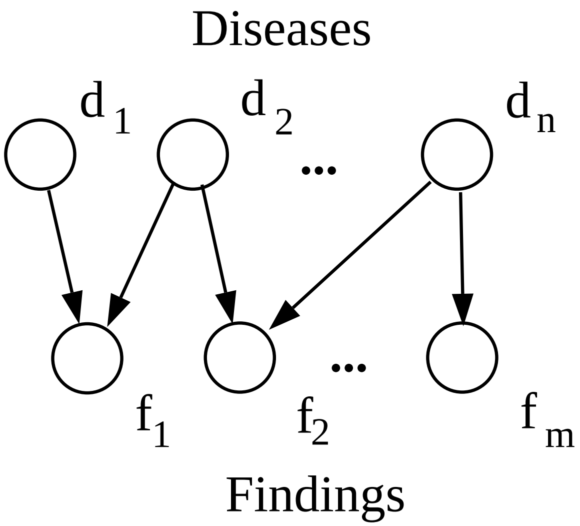
The QMR belief network

(Shwe et al. 1991)



- Contains over 600 significant diseases and about 4000 associated findings in internal medicine.
- Embodies extensive expert (and statistical) knowledge.

QMR: Statistical assumptions



- The assumptions:

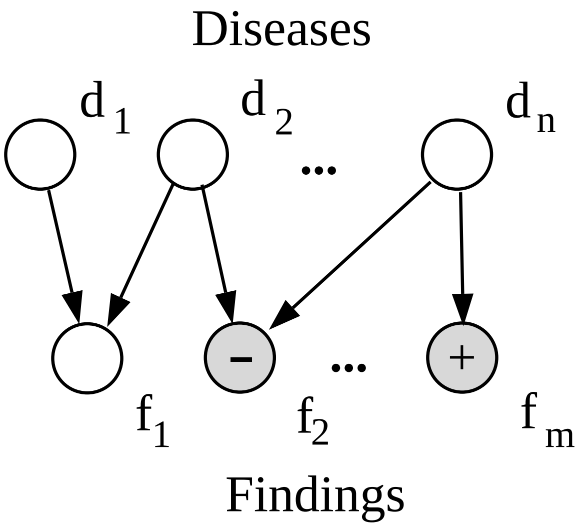
  (1) The diseases are marginally independent.

  (2) The findings are conditionally independent.

  (3) "Causal" independence (noisy-OR)

$$
\begin{aligned}
P(f_k^-|d) &= e^{-(\,\theta_{k1}\,d_1 + \ldots + \theta_{kn}\,d_n\,)} \\
P(f_k^+|d) &= 1 - e^{-(\,\theta_{k1}\,d_1 + \ldots + \theta_{kn}\,d_n\,)}
\end{aligned}
$$

Diagnostic inference

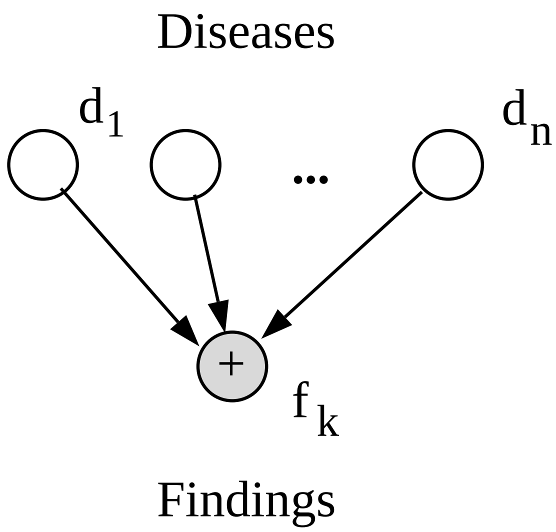


- The inference problem:

  Compute the posterior probabilities for the diseases given the instantiated findings.

  the order in which the evidence is absorbed from the findings is immaterial.

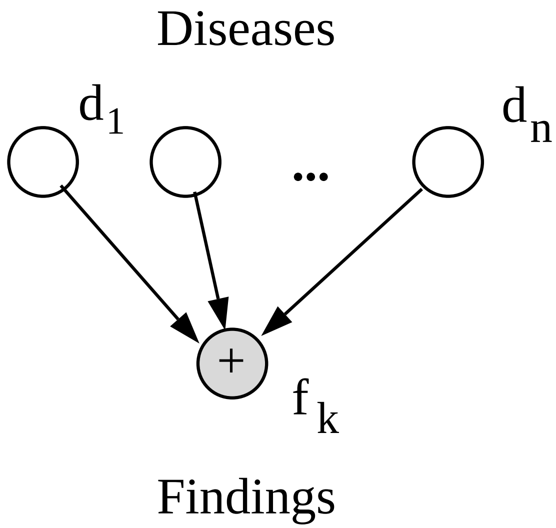
Evidence from positive findings



$$
\begin{aligned}
P(d|f_k^+) &= c\,P(f_k^+|d)P(d) \\
&= c\left[1 - e^{-(\,\theta_{k1}\,d_1 + \ldots + \theta_{kn}\,d_n\,)}\right] P(d)
\end{aligned}
$$

- Marginal independence no longer holds.
- Absorbing the evidence from positive findings is (typically) exponentially costly in the number of positive findings.

Evidence from positive findings



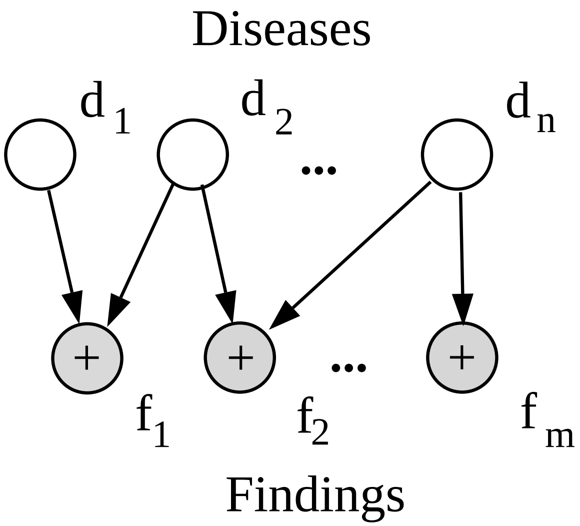
Variational evidence

- We can transform the positive evidence according to

$$\begin{aligned} P(f_k^+|d) &= \left[1 - e^{-(\,\theta_{k1}\,d_1 + \ldots + \theta_{kn}\,d_n\,)}\right] \\ &= e^{\log\left[1 - e^{-(\,\theta_{k1}\,d_1 + \ldots + \theta_{kn}\,d_n\,)}\right]} \\ &\leq e^{\lambda\,(\,\theta_{k1}\,d_1 + \ldots + \theta_{kn}\,d_n\,) - f^*(\lambda)} \end{aligned}$$

since $f(x) = \log\left[1 - e^{-x}\right]$ is a convex down function.

Evidence from positive findings

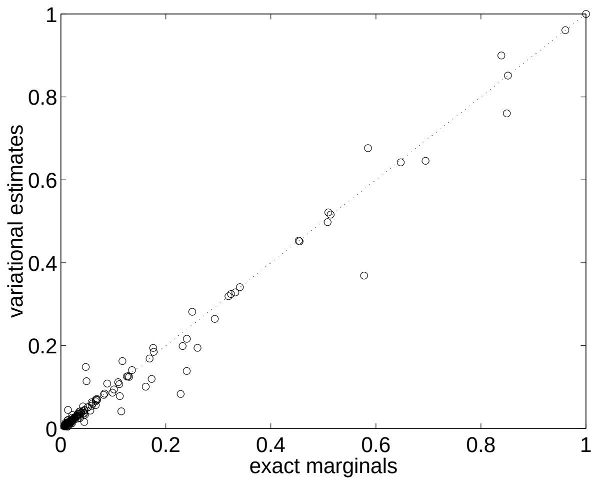


With every variational transformation

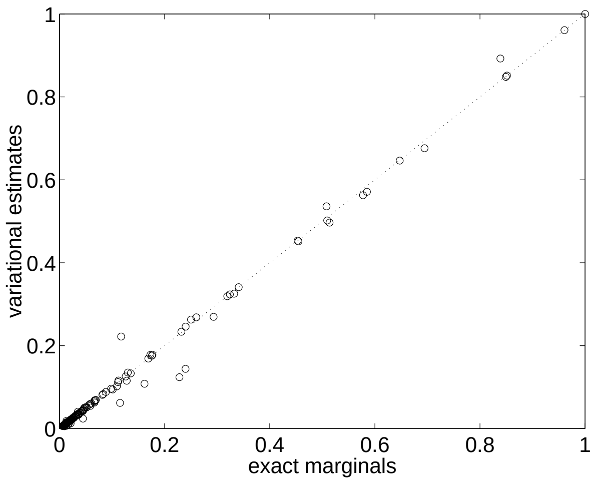
- we lose some accuracy
- we reduce the computation time by a factor of two.

⇒ We can balance the accuracy of inference with the available computational resources

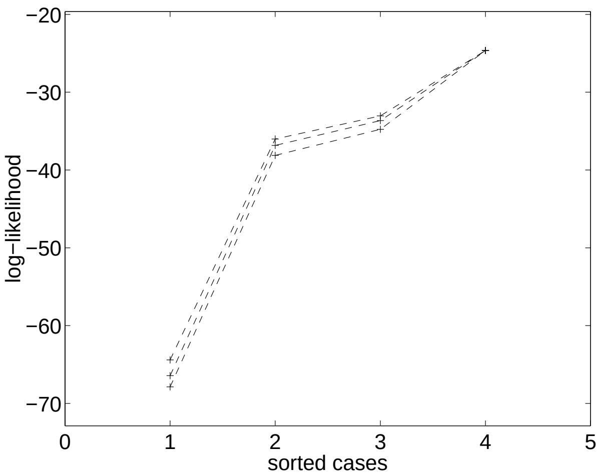
Performance on actual medical cases



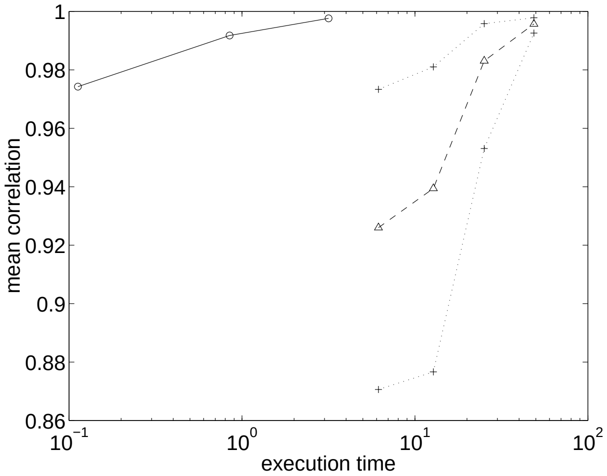
12 findings treated exactly



16 findings treated exactly



bounds on log-likelihood



comparison to sampling

The QMR belief network: conclusions

- The approximation results are *closed form expressions* and can be analyzed for reliability.
- We may also use more sophisticated variational methods to obtain rigorous upper and lower bounds on the desired disease marginals (practical?)

Discussion

- There are a number of variational methods we did not discuss
  - recursive variational algorithms
  - on-line variational methods for Bayesian estimation
  - variational methods for structured Bayesian estimation (with hyperparameters)
  - etc.
- Current and future directions:
  - smooth combination of variational methods with other methods (e.g., junction tree, sampling)
  - better characterization of the accuracy of variational approximation methods
  - new formulations

References

(please see the references in the associated tutorial paper)