Unsupervised Learning

Variational Methods
and Other Approximations

Maneesh Sahani

`maneesh@gatsby.ucl.ac.uk`

Gatsby Computational Neuroscience Unit, and
MSc in Intelligent Systems, Dept Computer Science
University College London

Term 1, Autumn 2006

Integrals in Statistical Modelling

- **Parameter estimation**

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \int d\mathcal{Y}\; P(\mathcal{Y}|\theta) P(\mathcal{X}|\mathcal{Y}, \theta)$$

  (or using EM)

$$\theta^{\text{new}} = \underset{\theta}{\text{argmax}} \int d\mathcal{Y}\; P(\mathcal{Y}|\mathcal{X}, \theta^{\text{old}}) \log P(\mathcal{X}, \mathcal{Y}|\theta)$$

- **Prediction**

$$p(x|\mathcal{D}, m) = \int d\theta\; p(\theta|\mathcal{D}, m) p(x|\theta, \mathcal{D}, m)$$

- **Model selection or weighting** (by marginal likelihood)

$$p(\mathcal{D}|m) = \int d\theta\; p(\theta|m) p(\mathcal{D}|\theta, m)$$

These integrals are often intractable:

- **Analytic intractability**: integrals may not have closed form in non-linear, non-Gaussian models $\Rightarrow$ numerical integration.
- **Computational intractability**: Numerical integral (or sum if $\mathcal{Y}$ or θ are discrete) may be exponential in data or model size.

Examples of Intractability

- Bayesian marginal likelihood/model evidence for Mixture of Gaussians: exact computations are exponential in number of data points

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \int d\theta \; p(\theta) \prod_{i=1}^{N} \sum_{s_i} p(\mathbf{x}_i | s_i, \theta) p(s_i|\theta)$$

$$= \sum_{s_1} \sum_{s_2} \ldots \sum_{s_N} \int d\theta \; p(\theta) \prod_{i=1}^{N} p(\mathbf{x}_i | s_i, \theta) p(s_i|\theta)$$

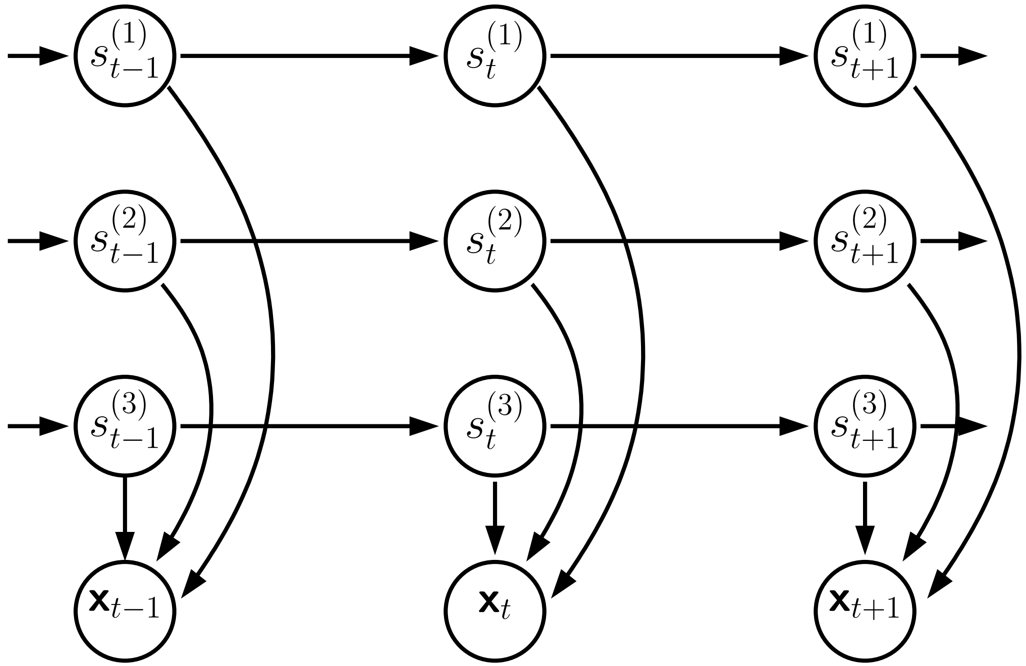
- Computing the conditional probability of a variable in a very large multiply connected directed graphical model:

$$p(x_i | X_j = a) = \sum_{\text{all settings of } \mathbf{y} \backslash \{i,j\}} p(x_i, \mathbf{y}, X_j = a) / p(X_j = a)$$

- Computing the hidden state distribution in a general nonlinear dynamical system

$$p(\mathbf{y}_t | \mathbf{x}_1, \ldots, \mathbf{x}_T) \propto \int p(\mathbf{y}_t | \mathbf{y}_{t-1}) p(\mathbf{x}_t | \mathbf{y}_t) p(\mathbf{y}_{t-1} | \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) p(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_t | \mathbf{y}_t) d\mathbf{y}_{t-1}$$

Distributed models



In the FHMM, moralisation puts simultaneous states $s_t^{(1)}, s_t^{(2)}, s_t^{(3)}$ into a single clique.

- M state variables, K values $\Rightarrow$ sums over K^{2M} terms.
- Factorial *prior* $\not\Rightarrow$ Factorial *posterior* (explaining away).

Variational methods approximate the posterior, often in a factored form. To see how they work, we need to review the free-energy interpretation of EM.

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ.

Goal: Maximize the log likelihood (i.e. ML learning) wrt θ:

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\begin{aligned}\int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})}\, d\mathcal{Y} &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta)\, d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y})\, d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta)\, d\mathcal{Y} + \mathbf{H}[q],\end{aligned}$$

where $\mathbf{H}[q]$ is the entropy of $q(\mathcal{Y})$.
So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + \mathbf{H}[q]$$

The E and M steps of EM

The log likelihood is bounded below (Jensen) by:

$$\mathcal{F}(q,\theta) = \langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{q(\mathcal{Y})} + \textbf{H}[q],$$

EM alternates between:

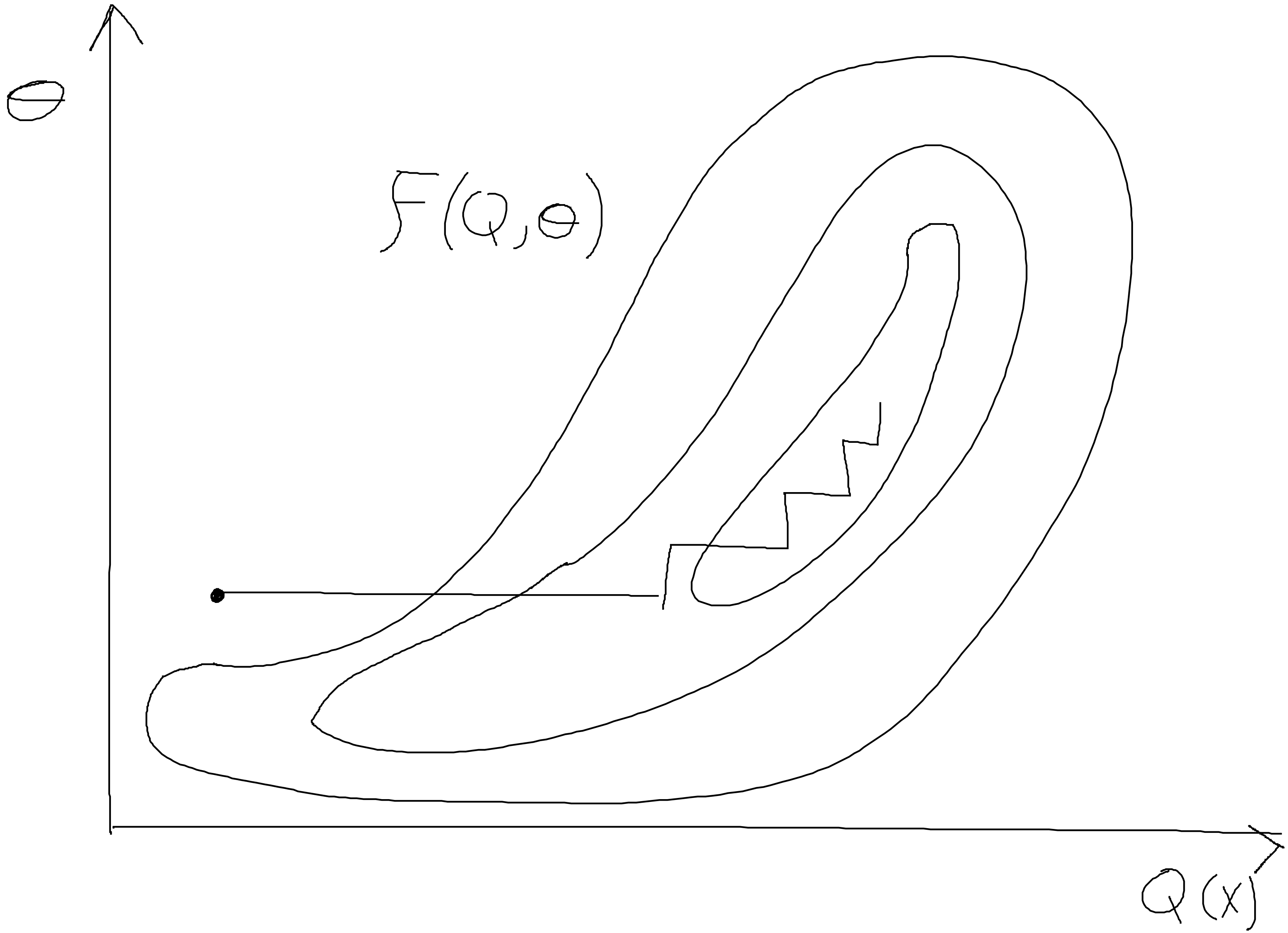
E step: optimise $\mathcal{F}(q,\theta)$ wrt distribution over hidden variables holding parameters fixed:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})}{\text{argmax}}\ \mathcal{F}\big(q(\mathcal{Y}), \theta^{(k-1)}\big) = P(\mathcal{Y}|\mathcal{X}, \theta^{(k-1)})$$

M step: maximise $\mathcal{F}(q,\theta)$ wrt parametersholding hidden distribution fixed:

$$\theta^{(k)} := \underset{\theta}{\text{argmax}}\ \mathcal{F}\big(q^{(k)}(\mathcal{Y}), \theta\big) = \underset{\theta}{\text{argmax}}\ \ \langle \log P(\mathcal{Y},\mathcal{X}|\theta)\rangle_{q^{(k)}(\mathcal{Y})}$$

EM as Coordinate Ascent in $\mathcal{F}$



EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big) \underset{\text{Jensen}}{\leq} \ell\big(\theta^{(k)}\big),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that $\theta^{(k)} \neq \theta^{(k-1)}$ iff $\mathcal{F}$ increases, then the overall EM iteration will step to a new value of θ iff the likelihood increases.

Variational Approximations to the EM algorithm

What if finding expected sufficient stats under $P(\mathcal{Y}|\mathcal{X}, \theta)$ is computationally intractable?

In the **generalised EM** algorithm, we argued that intractable maximisations could be replaced by gradient M-steps. For the E-step we could:

- Parameterise $q = q_\rho(\mathcal{Y})$ and take a gradient step in ρ.
- Assume some simplified form for q, usually factored: $q = \prod_i q_i(\mathcal{Y}_i)$ where $\mathcal{Y}_i$ partition $\mathcal{Y}$, and maximise within this form.

In both cases, we assume $q \in \mathcal{Q}$, and optimise within this class:

VE step: maximise $\mathcal{F}(q, \theta)$ wrt restricted latent distribution given parameters:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y}) \in \mathcal{Q}}{\text{argmax}} \; \mathcal{F}\big(q(\mathcal{Y}), \theta^{(k-1)}\big).$$

M step: unchanged

$$\theta^{(k)} := \underset{\theta}{\text{argmax}} \; \mathcal{F}\big(q^{(k)}(\mathcal{Y}), \theta\big) = \underset{\theta}{\text{argmax}} \int q^{(k)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

This maximises a lower bound on the log likelihood.

What do we lose?

What does restricting q to $\mathcal{Q}$ cost us?

- Recall that the free-energy is bounded above by Jensen:

$$\mathcal{F}(q, \theta) \leq \ell(\theta^{\text{ML}})$$

  Thus, as long as every step increases $\mathcal{F}$, convergence is still guaranteed.

- But, since $P(\mathcal{Y}|\mathcal{X}, \theta^{(k)})$ may not lie in $\mathcal{Q}$, we no longer saturate the bound after the E-step. Thus, the likelihood may not increase on each full EM step.

$$\ell\big(\theta^{(k-1)}\big) \underset{\text{E step}}{\not=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big) \underset{\text{Jensen}}{\leq} \ell\big(\theta^{(k)}\big),$$

- Thus, we may not converge to a maximum of ℓ.

The hope is that by *increasing a lower bound* on ℓ we will find a decent solution.
[Note that if $P(\mathcal{Y}|\mathcal{X}, \theta^{\text{ML}}) \in \mathcal{Q}$, then θ^{ML} is a fixed point of the variational algorithm.]

KL divergence

Recall that

$$\begin{aligned}\mathcal{F}(q,\theta) &= \langle \log P(\mathcal{X},\mathcal{Y}|\theta)\rangle_{q(\mathcal{Y})} + \mathbf{H}[q] \\ &= \langle \log P(\mathcal{X}|\theta) + \log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{q(\mathcal{Y})} - \langle \log q(\mathcal{Y})\rangle_{q(\mathcal{Y})} \\ &= \langle \log P(\mathcal{X}|\theta)\rangle_{q(\mathcal{Y})} - \mathbf{KL}[q\|P(\mathcal{Y}|\mathcal{X},\theta)].\end{aligned}$$

Thus,

E step maximise $\mathcal{F}(q,\theta)$ wrt the distribution over latents, given parameters:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})\in\mathcal{Q}}{\text{argmax}}\ \mathcal{F}\big(q(\mathcal{Y}), \theta^{(k-1)}\big).$$

is equivalent to:

E step minimise $\mathbf{KL}[q\|p(\mathcal{Y}|\mathcal{X},\theta)]$ wrt distribution over latents, given parameters:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})\in\mathcal{Q}}{\text{argmin}} \int q(\mathcal{Y}) \log \frac{q(\mathcal{Y})}{p(\mathcal{Y}|\mathcal{X},\theta^{(k-1)})} d\mathcal{Y}$$

So, in each E step, the algorithm is trying to find the best approximation to $P(\mathcal{Y}|\mathcal{X})$ in $\mathcal{Q}$.

This is related to ideas in *information geometry*.

Factored Variational E-step

The most common form of variational approximation partitions $\mathcal{Y}$ into disjoint sets $\mathcal{Y}_i$ with

$$\mathcal{Q} = \Big\{ q \;\Big|\; q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i) \Big\}.$$

In this case the E-step is itself iterative:

(Factored VE step)$_i$: maximise $\mathcal{F}(q, \theta)$ wrt $q_i(\mathcal{Y}_i)$ given other q_j and parameters:

$$q_i^{(k)}(\mathcal{Y}_i) := \underset{q_i(\mathcal{Y}_i)}{\text{argmax}} \;\; \mathcal{F}\Big(q_i(\mathcal{Y}_i) \prod_{j \neq i} q_j(\mathcal{Y}_j), \theta^{(k-1)}\Big).$$

The q_is can be updated iteratively until convergence before moving on to the M-step. Alternatively, we can make a single pass over all q_i (starting from values at the last step) and then perform an M-step. Each VE step increases $\mathcal{F}$, so convergence is still guaranteed.

Factored Variational E-step

The Factored Variational E-step has a general form.

The free energy is:

$$\mathcal{F}\Big(\prod_j q_j(\mathcal{Y}_j), \theta^{(k-1)}\Big) = \Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)}) \Big\rangle_{\prod_j q_j(\mathcal{Y}_j)} + \mathbf{H}\Big[\prod_j q_j(\mathcal{Y}_j)\Big]$$

$$= \int d\mathcal{Y}_i \; q_i(\mathcal{Y}_i) \Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)}) \Big\rangle_{\prod_{j \neq i} q_j(\mathcal{Y}_j)} + \mathbf{H}[q_i] + \sum_{j \neq i} \mathbf{H}[q_j]$$

Now, taking the variational derivative of the Lagrangian (enforcing normalisation of q_i):

$$\frac{\delta}{\delta q_i}\Big(\mathcal{F} + \lambda\Big(\int q_i - 1\Big)\Big) = \Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)}) \Big\rangle_{\prod_{j \neq i} q_j(\mathcal{Y}_j)} - \log q_i(\mathcal{Y}_i) - 1 + \lambda$$

$$(= 0) \quad \Rightarrow \quad q_i(\mathcal{Y}_i) \propto \exp \Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)}) \Big\rangle_{\prod_{j \neq i} q_j(\mathcal{Y}_j)}$$

In general, this depends only on the expected sufficient statistics under q_j. Thus, once again, we don't actually need the *entire* distributions, just the *relevant* expectations.

Mean-field Approximations

If $\mathcal{Y}_i = y_i$ (*i.e.*, q is factored over all variables) then the variational technique is often called a "mean field" approximation.

Suppose $P(\mathcal{X}, \mathcal{Y})$ is log-linear, *e.g.* the Boltzmann machine:

$$P(\mathcal{X}, \mathcal{Y}) = \frac{1}{Z} \exp\left(\sum_{ij} W_{ij} s_i s_j + \sum_i b_i s_i \right)$$

with some $s_i \in \mathcal{Y}$ and others observed.

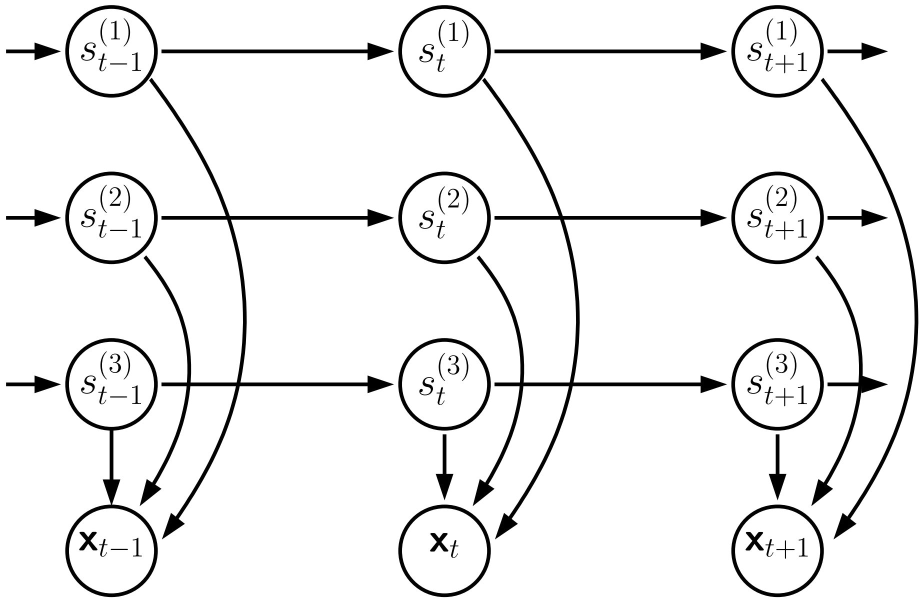
Expectations wrt a fully factored q distribute over all $s_i \in \mathcal{Y}$

$$\langle \log P(\mathcal{X}, \mathcal{Y}) \rangle_{\prod q_i} = \sum_{ij} W_{ij} \langle s_i \rangle_{q_i} \langle s_j \rangle_{q_j} + \sum_i b_i \langle s_i \rangle_{q_i}$$

(where q_i for $s_i \in \mathcal{X}$ is a delta function on observed value).

Thus, we can update each q_i in turn given the means of the others. Each variable is seeing the *mean* field imposed by its neighbours. We update these fields until they all agree.

Mean-field FHMM



The mean-field approach to the FHMM with

$$q(s_{1:T}^{1:M}) = \prod_{m,t} q_t^m(s_t^m)$$

yields a variant of the usual forward-backward algorithm. Coupling between the different chains only takes place through the joint output distribution. Each update depends only on the immediate neighbours.

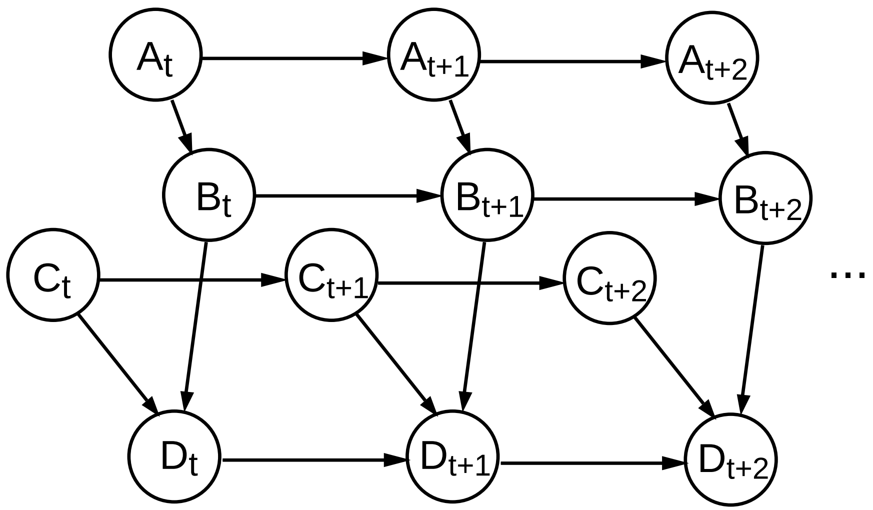
$$\begin{aligned}
q_{t'}^{m'}(s_{t'}^{m'}) &\propto \exp \left\langle \log P(\mathbf{s}_{1:T}^{1:M}, \mathbf{x}_{1:T}) \right\rangle_{\prod_{\neg(m',t')} q_t^m(s_t^m)} \\
&= \exp \left\langle \sum_m \sum_t \log P(s_t^m | s_{t-1}^m) + \sum_t \log P(\mathbf{x}_t | s_t^{1:M}) \right\rangle_{\prod_{\neg(m',t')} q_t^m} \\
&\propto \exp \left[\left\langle \log P(s_{t'}^{m'} | s_{t'-1}^{m'}) \right\rangle_{q_{t'-1}^{m'}} + \left\langle \log P(s_{t+1'}^{m'} | s_{t'}^{m'}) \right\rangle_{q_{t'+1}^{m'}} + \left\langle \log P(\mathbf{x}_{t'} | s_{t'}^{1:M}) \right\rangle_{\prod_{\neg m} q_{t'}^m} \right]
\end{aligned}$$

Structured Variational Approximations

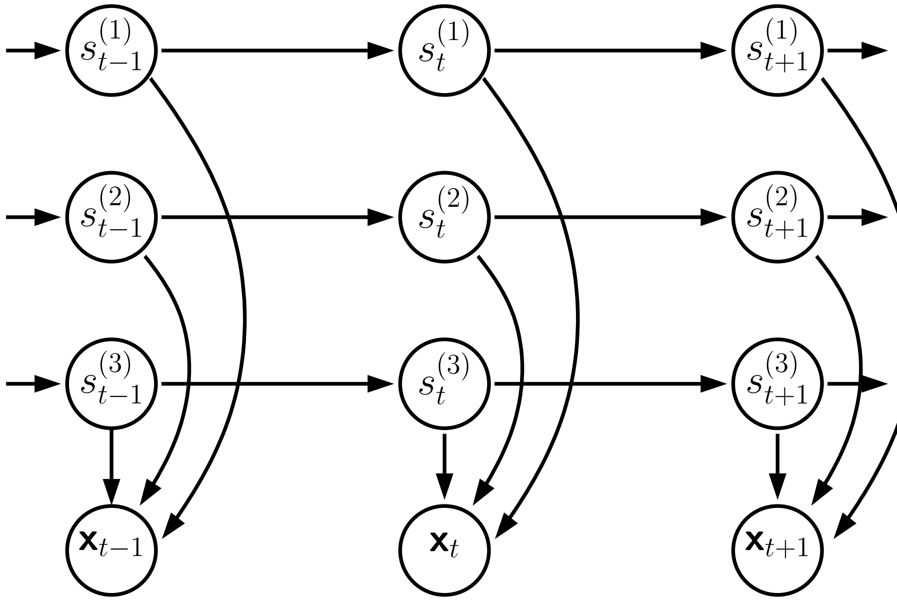
$q(\mathcal{Y})$ need not be completely factorized.

For example, suppose you can partition $\mathcal{Y}$ into sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ such that computing the expected sufficient statistics under $q(\mathcal{Y}_1)$ and $q(\mathcal{Y}_2)$ is tractable.
Then $q(\mathcal{Y}) = q(\mathcal{Y}_1)q(\mathcal{Y}_2)$ is tractable.

If you have a graphical model, you may want to factorize $q(\mathcal{Y})$ into a product of trees, which are tractable distributions.



Stuctured FHMM



The most natural structured approximation in the FHMM is to factor each chain from the others

$$q(s_{1:T}^{1:M}) = \prod_m q^m(s_{1:T}^m)$$

Updates within each chain are then found by a forward-backward algorithm, with a modified "likelihood" term.

$$\begin{aligned}
q^{m'}(s_{1:T}^{m'}) &\propto \exp \left\langle \log P(\mathbf{s}_{1:T}^{1:M}, \mathbf{x}_{1:T}) \right\rangle_{\prod_{\neg m'} q^m(s_{1:T}^m)} \\
&= \exp \left\langle \sum_m \sum_t \log P(s_t^m | s_{t-1}^m) + \sum_t \log P(\mathbf{x}_t | s_t^{1:M}) \right\rangle_{\prod_{\neg m'} q^m} \\
&\propto \exp \left[\sum_t \log P(s_t^{m'} | s_{t-1}^{m'}) + \sum_t \left\langle \log P(\mathbf{x}_{t'} | s_{t'}^{1:M}) \right\rangle_{\prod_{\neg m} q^m s_{t'}^m} \right] \\
&= \prod_t P(s_t^{m'} | s_{t-1}^{m'}) \quad \prod_t e^{\left\langle \log P(\mathbf{x}_{t'} | s_{t'}^{1:M}) \right\rangle_{\prod_{\neg m} q^m s_{t'}^m}}
\end{aligned}$$

Variational Approximations and Graphical Models I

Let $q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i)$.

Variational approximation maximises $\mathcal{F}$:

$$\mathcal{F}(q) = \int q(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}) d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) d\mathcal{Y}$$

Focusing on one term, q_j, we can write this as:

$$\mathcal{F}(q_j) = \int q_j(\mathcal{Y}_j) \left\langle \log p(\mathcal{Y}, \mathcal{X}) \right\rangle_{\neg q_j(\mathcal{Y}_j)} d\mathcal{Y}_j + \int q_j(\mathcal{Y}_j) \log q_j(\mathcal{Y}_j) d\mathcal{Y}_j + \text{const}$$

Where $\langle \cdot \rangle_{\neg q_j(\mathcal{Y}_j)}$ denotes averaging w.r.t. $q_i(\mathcal{Y}_i)$ for all $i \neq j$

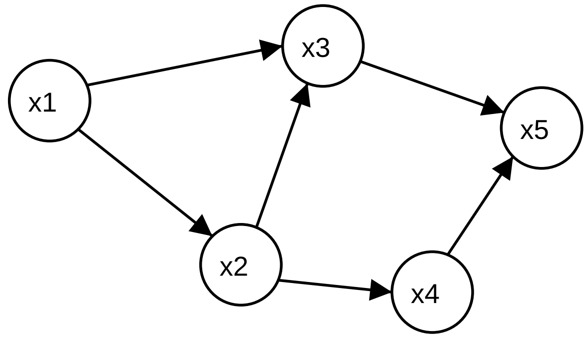
Optimum occurs when:

$$q_j^*(\mathcal{Y}_j) = \frac{1}{Z} \exp \left\langle \log p(\mathcal{Y}, \mathcal{X}) \right\rangle_{\neg q_j(\mathcal{Y}_j)}$$

Variational Approximations and Graphical Models II

Optimum occurs when:

$$q_j^*(\mathcal{Y}_j) = \frac{1}{Z} \exp \langle \log p(\mathcal{Y}, \mathcal{X}) \rangle_{\neg q_j(\mathcal{Y}_j)}$$



Assume graphical model: $p(\mathcal{Y}, \mathcal{X}) = \prod_i p(X_i | \text{pa}_i)$

$$\begin{aligned} \log q_j^*(\mathcal{Y}_j) &= \left\langle \sum_i \log p(X_i | \text{pa}_i) \right\rangle_{\neg q_j(\mathcal{Y}_j)} + \text{const} \\ &= \langle \log p(\mathcal{Y}_j | \text{pa}_j) \rangle_{\neg q_j(\mathcal{Y}_j)} + \sum_{k \in \text{ch}_j} \langle \log p(X_k | \text{pa}_k) \rangle_{\neg q_j(\mathcal{Y}_j)} + \text{const} \end{aligned}$$

This defines messages that get passed between nodes in the graph. Each node receives messages from its Markov boundary: parents, children and parents of children.

Variational Message Passing (Winn and Bishop, 2004)

Variational Approximations to Bayesian Learning

$$\begin{aligned}\log p(\mathcal{X}) &= \log \int \int p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \, d\mathcal{Y} \, d\boldsymbol{\theta} \\ &\geq \int \int q(\mathcal{Y}, \boldsymbol{\theta}) \log \frac{p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})}{q(\mathcal{Y}, \boldsymbol{\theta})} \, d\mathcal{Y} \, d\boldsymbol{\theta}\end{aligned}$$

Constrain $q \in \mathcal{Q}$ s.t. $q(\mathcal{Y}, \boldsymbol{\theta}) = q(\mathcal{Y}) q(\boldsymbol{\theta})$.

This results in the **variational Bayesian EM algorithm**.

More about this later (when we study model selection).

The Other KL

Variational methods find $q = \operatorname{argmin} \textbf{KL}[q\|p(y|x)]$:

- guaranteed convergence;
- maximising lower bound may help ℓ increase;
- (factored approximation) distributes for message-passing.

What about the 'other' KL ($q = \operatorname{argmin} \textbf{KL}[p(y|x)\|q]$)?

Crucially, for a factored approximation the (clique) marginals are correct:

$$
\begin{aligned}
\operatorname*{argmin}_{q_i} \textbf{KL}\Big[P(\mathcal{Y}|\mathcal{X})\Big\|\prod q_j(\mathcal{Y}_j|\mathcal{X})\Big] &= \operatorname*{argmin}_{q_i} -\int d\mathcal{Y}\; P(\mathcal{Y}|\mathcal{X}) \log \prod_j q_j(\mathcal{Y}_j|\mathcal{X}) \\
&= \operatorname*{argmin}_{q_i} -\sum_j \int d\mathcal{Y}\; P(\mathcal{Y}|\mathcal{X}) \log q_j(\mathcal{Y}_j|\mathcal{X}) \\
&= \operatorname*{argmin}_{q_i} -\int d\mathcal{Y}_i\; P(\mathcal{Y}_i|\mathcal{X}) \log q_i(\mathcal{Y}_i|\mathcal{X}) \\
&= P(\mathcal{Y}_i|\mathcal{X})
\end{aligned}
$$

and the marginals are what we need for learning.

But (perversely) this means finding the best q for this KL is intractrable!

Expectation Propagation (EP)

The distribution we need to approximate is often a (normalised) product of factors:

$$P(\mathcal{Y}|\mathcal{X}) = \frac{P(\mathcal{Y},\mathcal{X})}{P(\mathcal{X})} = \frac{1}{Z}\prod_i P(s_i|\mathsf{pa}(s_i)) \propto \prod_{i=1}^{N} f_i(\mathcal{Y}_i)$$

where the $\mathcal{Y}_i$ are not necessarily disjoint.

We wish to approximate this by a product of *simpler* terms: $q(\mathcal{Y}) \stackrel{\text{def}}{=} \prod_{i=1}^{N} \tilde{f}_i(\mathcal{Y}_i)$

$$\min_{q(\mathcal{Y}_i)} \textbf{KL}\left[\prod_{i=1}^{N} f_i(\mathcal{Y}_i) \middle\| \prod_{i=1}^{N} \tilde{f}_i(\mathcal{Y}_i)\right]$$ (intractable)

$$\min_{\tilde{f}_i(\mathcal{Y}_i)} \textbf{KL}\left[f_i(\mathcal{Y}_i) \middle\| \tilde{f}_i(\mathcal{Y}_i)\right]$$ (simple, non-iterative, inaccurate)

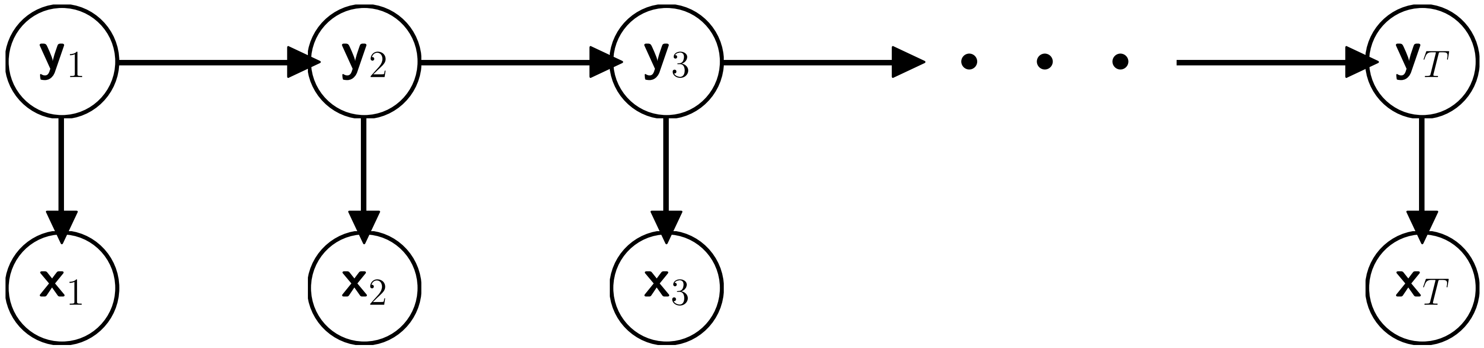
$$\min_{\tilde{f}_i(\mathcal{Y}_i)} \textbf{KL}\left[f_i(\mathcal{Y}_i)\prod_{j\neq i} \tilde{f}_j(\mathcal{Y}_i) \middle\| \tilde{f}_i(\mathcal{Y}_i)\prod_{j\neq i} \tilde{f}_j(\mathcal{Y}_i)\right]$$ (simple, iterative, accurate) $\longleftarrow$ EP

Expectation Propagation II

Input $f_1(\mathcal{Y}_1) \dots f_N(\mathcal{Y}_N)$
Initialize $\tilde{f}_1(\mathcal{Y}_1) = f_1(\mathcal{Y}_1)$, $\tilde{f}_i(\mathcal{Y}_i) = 1$ for $i > 1$, $q(\mathcal{Y}) = \prod_i \tilde{f}_i(\mathcal{Y}_i)$
repeat
  for $i = 1 \dots N$ **do**
    Deletion: $q_{\neg i}(\mathcal{Y}) \leftarrow \dfrac{q(\mathcal{Y})}{\tilde{f}_i(\mathcal{Y}_i)} = \prod_{j \neq i} \tilde{f}_j(\mathcal{Y}_j)$
    Projection: $\tilde{f}_i^{\text{new}}(\mathcal{Y}) \leftarrow \underset{f(\mathcal{Y}_i)}{\text{argmin}}\ \mathbf{KL}[f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}) \| f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y})]$
    Inclusion: $q(\mathcal{Y}) \leftarrow \tilde{f}_i^{\text{new}}(\mathcal{Y}_i)\, q_{\neg i}(\mathcal{Y})$
  end for
until convergence

- KL minimisation (projection) only depends on $q_{\neg i}(\mathcal{Y})$ marginalised to $\mathcal{Y}_i$.
- $\tilde{f}_i(\mathcal{Y})$ in exponential family $\rightarrow$ projection step is moment matching.
- Update order need not be sequential.
- Minimizes the opposite KL to variational methods.
- Loopy belief propagation and assumed density filtering are special cases.
- No convergence guarantee (although convergent forms can be developed).
- The names (deletion, projection, inclusion) are not the same as in (Minka, 2001).

EP for a NLSSM



$$p(\mathbf{y}_t|\mathbf{y}_{t-1}) = \phi_t(\mathbf{y}_t, \mathbf{y}_{t-1}) \qquad \textit{e.g. } \exp(-\|\mathbf{y}_t - h_s(\mathbf{y}_{t-1})\|^2/2\sigma^2)$$

$$p(\mathbf{x}_t|\mathbf{y}_t) = \psi_t(\mathbf{y}_t) \qquad \textit{e.g. } \exp(-\|\mathbf{x}_t - h_o(\mathbf{y}_t)\|^2/2\sigma^2)$$

Then $f_t(\mathbf{y}_t, \mathbf{y}_{t-1}) = \phi_t(\mathbf{y}_t, \mathbf{y}_{t-1})\psi_t(\mathbf{y}_t)$. As ϕ_t and ψ_t are non-linear, EP is not generally tractable. Assume $\tilde{f}_t(\mathbf{y}_t, \mathbf{y}_{t-1})$ is Gaussian. Then,

$$q_{\neg t}(\mathbf{y}_t, \mathbf{y}_{t-1}) = \sum_{\substack{\mathbf{y}_1 \cdots \mathbf{y}_{t-2} \\ \mathbf{y}_{t+1} \cdots \mathbf{y}_T}} \prod_{t' \neq t} \tilde{f}_{t'}(\mathbf{y}_{t'}, \mathbf{y}_{t'-1}) = \underbrace{\sum_{\mathbf{y}_1 \cdots \mathbf{y}_{t-2}} \prod_{t' < t} \tilde{f}_{t'}(\mathbf{y}_{t'}, \mathbf{y}_{t'-1})}_{\alpha_{t-1}(\mathbf{y}_{t-1})} \underbrace{\sum_{\mathbf{y}_{t+1} \cdots \mathbf{y}_T} \prod_{t' > t} \tilde{f}_{t'}(\mathbf{y}_{t'}, \mathbf{y}_{t'-1})}_{\beta_t(\mathbf{y}_t)}$$

with both α and β Gaussian.

$$\tilde{f}_t(\mathbf{y}_t, \mathbf{y}_{t-1}) = \operatorname*{argmin}_{f \in \mathcal{N}} \mathbf{KL}[\phi_t(\mathbf{y}_t, \mathbf{y}_{t-1})\psi_t(\mathbf{y}_t)\alpha_{t-1}(\mathbf{y}_{t-1})\beta_t(\mathbf{y}_t) \| f(\mathbf{y}_t, \mathbf{y}_{t-1})\alpha_{t-1}(\mathbf{y}_{t-1})\beta_t(\mathbf{y}_t)]$$

Moment Matching

Recall that for exponential family $q(x) = \frac{1}{Z(\boldsymbol{\theta})} e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}}$:

$$
\begin{aligned}
\operatorname*{argmin}_q \mathbf{KL}[p(x)\|q(x)] &= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbf{KL}[p(x)\|\frac{1}{Z(\boldsymbol{\theta})}e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}}] \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} -\int dx\; p(x) \log \frac{1}{Z(\boldsymbol{\theta})}e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}} \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} -\int dx\; p(x)\mathbf{S}(x)\cdot\boldsymbol{\theta} + \log Z(\boldsymbol{\theta}) \\
\frac{\partial}{\partial\boldsymbol{\theta}} &= -\int dx\; p(x)\mathbf{S}(x) + \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}}\int dx\; e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}} \\
&= -\langle\mathbf{S}(x)\rangle_p + \frac{1}{Z(\boldsymbol{\theta})}\int dx\; e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}}\mathbf{S}(x) \\
&= -\langle\mathbf{S}(x)\rangle_p + \langle\mathbf{S}(x)\rangle_q
\end{aligned}
$$

So minimum is found by matching sufficient stats. This is usually moment matching.

How do we calculate $\langle\mathbf{S}(x)\rangle_p$? Low dimensional integral $\rightarrow$ Quadrature, Laplace approx ...