

Probabilistic & Unsupervised Learning

Variational EM and Variational Bayesian Learning

Maneesh Sahani

`maneesh@gatsby.ucl.ac.uk`

**Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London**

Term 1, Autumn 2009

Integrals in Statistical Modelling

- **Parameter estimation**

$$\hat{\theta} = \operatorname{argmax}_{\theta} \int d\mathcal{Y} P(\mathcal{Y}|\theta) P(\mathcal{X}|\mathcal{Y}, \theta)$$

(or using EM)

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

- **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

$$\begin{aligned} p(\mathbf{x}_1, \dots, \mathbf{x}_N) &= \int d\theta \, p(\theta) \prod_{i=1}^N \sum_{s_i} p(\mathbf{x}_i | s_i, \theta) p(s_i | \theta) \\ &= \sum_{s_1} \sum_{s_2} \dots \sum_{s_N} \int d\theta \, p(\theta) \prod_{i=1}^N p(\mathbf{x}_i | s_i, \theta) p(s_i | \theta) \end{aligned}$$

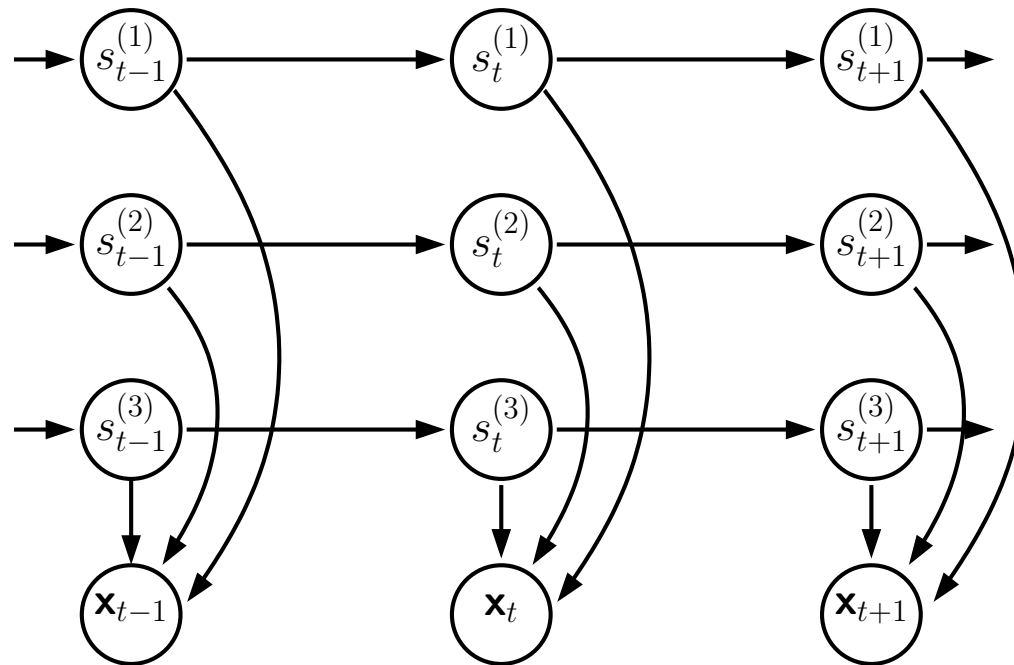
- Computing the conditional probability of a variable in a very large multiply connected directed graphical model:

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

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

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

Distributed models



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

- M state variables, K values \Rightarrow sums over K^{2M} terms.
- Factorial *prior* \nRightarrow 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:

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

EM alternates between:

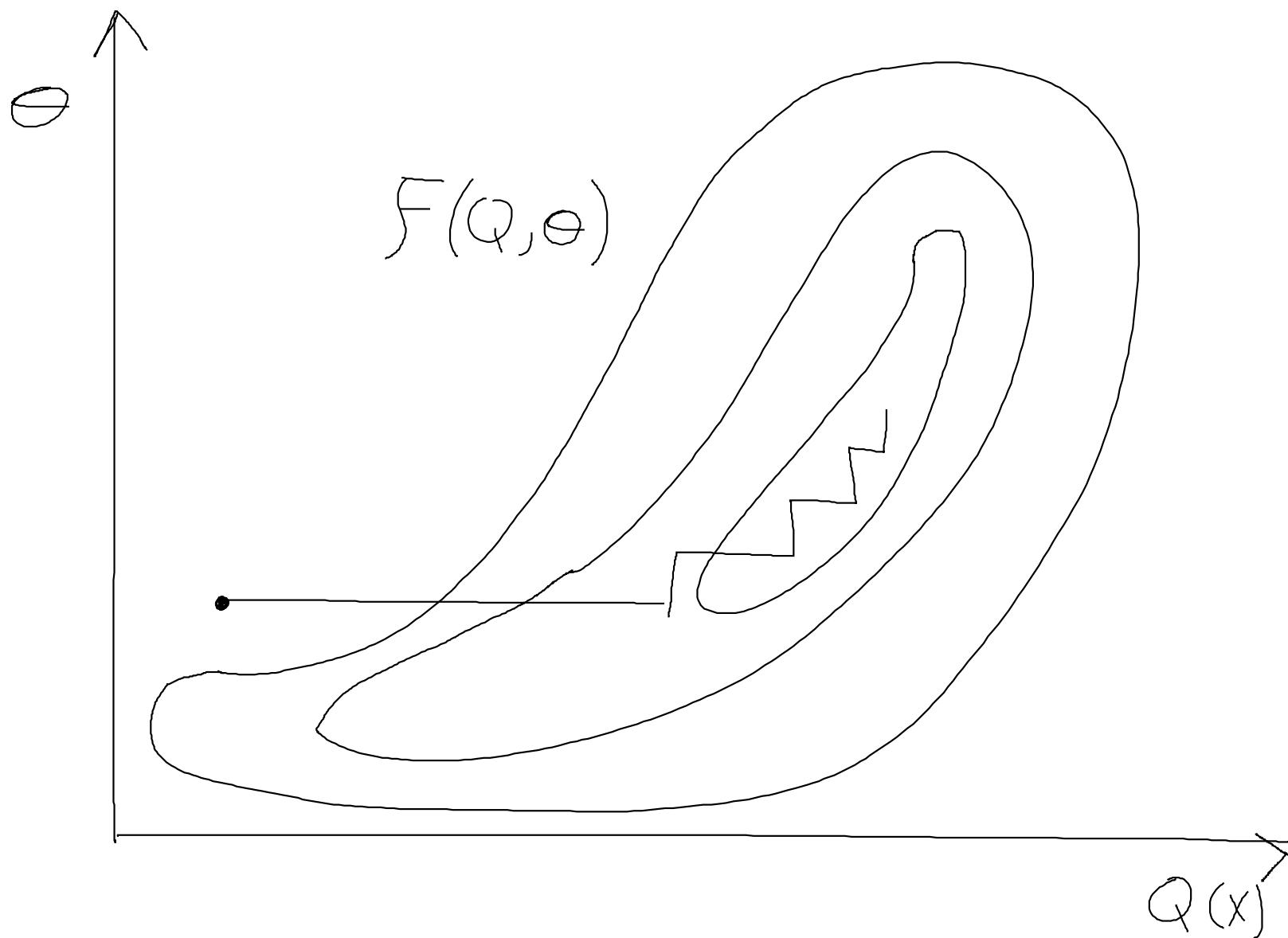
E step: optimise $\mathcal{F}(q, \theta)$ wrt distribution over hidden variables holding parameters fixed:

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

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

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

EM as Coordinate Ascent in \mathcal{F}



EM Never Decreases the Likelihood

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

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

- 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:

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

M step: unchanged

$$\theta^{(k)} := \operatorname{argmax}_{\theta} \mathcal{F}(q^{(k)}(\mathcal{Y}), \theta) = \operatorname{argmax}_{\theta} \int q^{(k)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y},$$

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.

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

- 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:

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

is equivalent to:

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

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

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

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

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:

$$q_i^{(k)}(\mathcal{Y}_i) := \operatorname{argmax}_{q_i(\mathcal{Y}_i)} \mathcal{F}\left(q_i(\mathcal{Y}_i) \prod_{j \neq i} q_j(\mathcal{Y}_j), \theta^{(k-1)}\right).$$

The q_i s 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:

$$\begin{aligned}\mathcal{F}\left(\prod_j q_j(\mathcal{Y}_j), \theta^{(k-1)}\right) &= \left\langle \log P(\mathcal{X}, \mathcal{Y} | \theta^{(k-1)}) \right\rangle_{\prod_j q_j(\mathcal{Y}_j)} + \mathbf{H}\left[\prod_j q_j(\mathcal{Y}_j)\right] \\ &= \int d\mathcal{Y}_i q_i(\mathcal{Y}_i) \left\langle \log P(\mathcal{X}, \mathcal{Y} | \theta^{(k-1)}) \right\rangle_{\prod_{j \neq i} q_j(\mathcal{Y}_j)} + \mathbf{H}[q_i] + \sum_{j \neq i} \mathbf{H}[q_j]\end{aligned}$$

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

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

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

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.

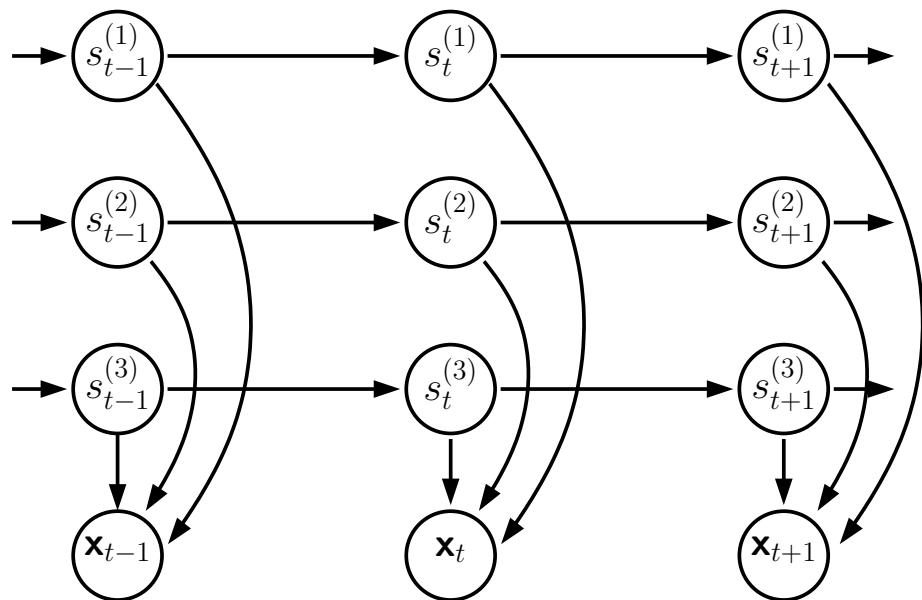
Expectations wrt a fully factored q distribute over all $s_i \in \mathcal{Y}$

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

(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.

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

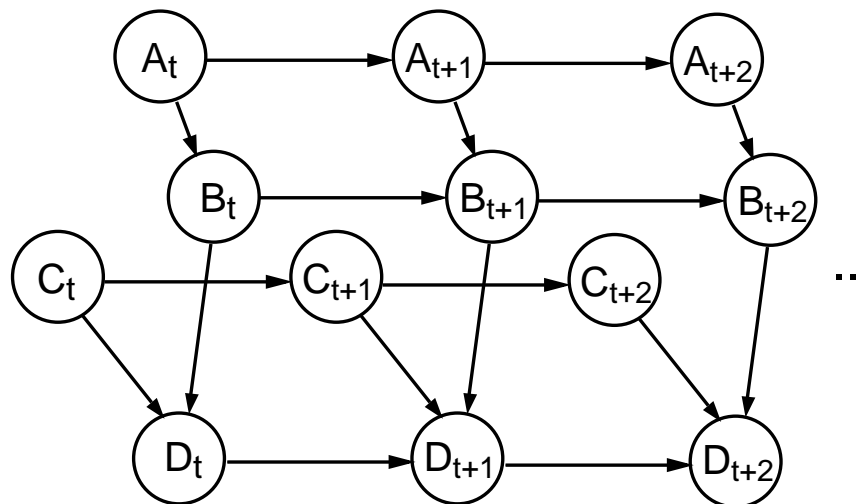
Structured Variational Approximations

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

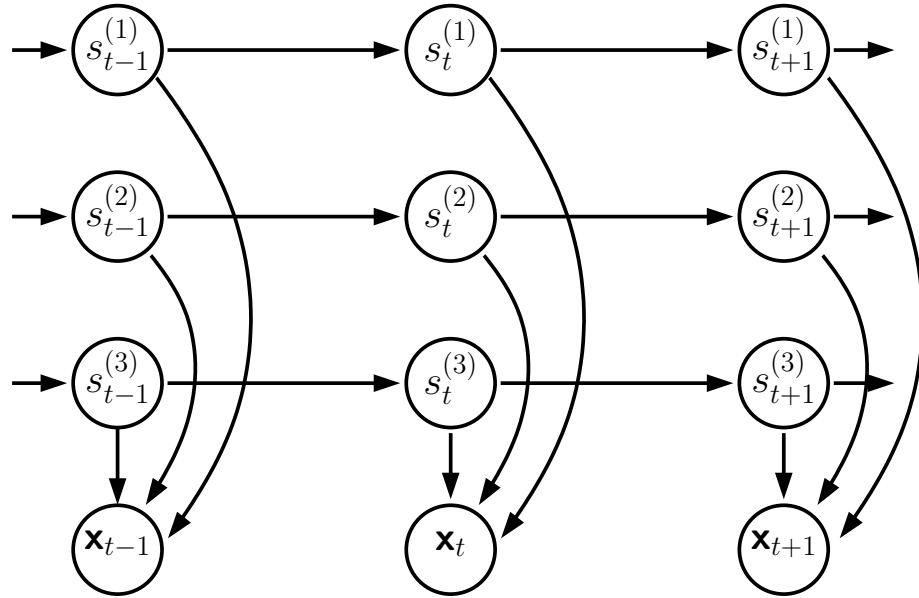
For example, suppose you can partition \mathcal{Y} into sets \mathcal{Y}_1 and \mathcal{Y}_2 such that computing the expected sufficient statistics under $q(\mathcal{Y}_1)$ and $q(\mathcal{Y}_2)$ is tractable.

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.



Structured 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.

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

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:

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

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:

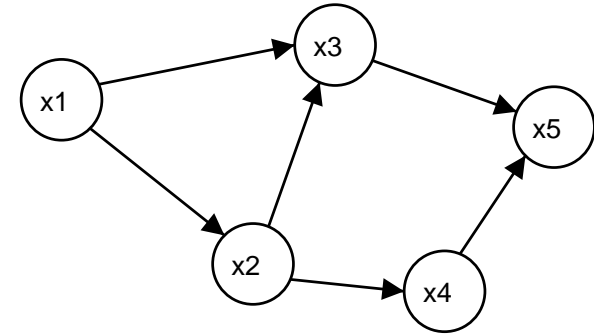
$$q_j^*(\mathcal{Y}_j) = \frac{1}{Z} \exp \langle \log p(\mathcal{Y}, \mathcal{X}) \rangle_{\neg q_j(\mathcal{Y}_j)}$$

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 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**.

Variational Bayesian Learning

Lower Bounding the Marginal Likelihood

Let the hidden latent variables be \mathcal{Y} , data \mathcal{X} and the parameters θ .

Lower bound the marginal likelihood (Bayesian model evidence) using Jensen's inequality:

$$\begin{aligned}\log P(\mathcal{X}) &= \log \int d\mathcal{Y} d\theta P(\mathcal{X}, \mathcal{Y}, \theta) && ||m \\ &= \log \int d\mathcal{Y} d\theta Q(\mathcal{Y}, \theta) \frac{P(\mathcal{X}, \mathcal{Y}, \theta)}{Q(\mathcal{Y}, \theta)} \\ &\geq \int d\mathcal{Y} d\theta Q(\mathcal{Y}, \theta) \log \frac{P(\mathcal{X}, \mathcal{Y}, \theta)}{Q(\mathcal{Y}, \theta)}.\end{aligned}$$

The saturating $Q(\mathcal{Y}, \theta) = P(\mathcal{Y}, \theta | \mathcal{X})$ is almost always intractable.

Use a simpler, factorised approximation $Q(\mathcal{Y}, \theta) = Q_{\mathcal{Y}}(\mathcal{Y})Q_{\theta}(\theta)$:

$$\begin{aligned}\log P(\mathcal{X}) &\geq \int d\mathcal{Y} d\theta Q_{\mathcal{Y}}(\mathcal{Y})Q_{\theta}(\theta) \log \frac{P(\mathcal{X}, \mathcal{Y}, \theta)}{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\theta}(\theta)} \\ &= \mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_{\theta}(\theta), \mathcal{X}).\end{aligned}$$

Maximize this lower bound. The resulting value is the approximation to the evidence.

Variational Bayesian Learning ...

Maximizing this **lower bound**, \mathcal{F} , leads to **EM-like** updates:

$$Q_{\mathcal{Y}}^*(\mathcal{Y}) \propto \exp \langle \log P(\mathcal{Y}, \mathcal{X} | \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \quad E\text{-like step}$$

$$Q_{\boldsymbol{\theta}}^*(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}) \exp \langle \log P(\mathcal{Y}, \mathcal{X} | \boldsymbol{\theta}) \rangle_{Q_{\mathcal{Y}}(\mathcal{Y})} \quad M\text{-like step}$$

Maximizing \mathcal{F} is equivalent to minimizing KL-divergence between the *approximate posterior*, $Q(\boldsymbol{\theta})Q(\mathcal{Y})$ and the *true posterior*, $P(\boldsymbol{\theta}, \mathcal{Y} | \mathcal{X})$.

$$\begin{aligned} \log P(\mathcal{X}) - \mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathcal{X}) &= \\ \log P(\mathcal{X}) - \int d\mathcal{Y} d\boldsymbol{\theta} Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})}{Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} &= \\ \int d\mathcal{Y} d\boldsymbol{\theta} Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{P(\mathcal{Y}, \boldsymbol{\theta} | \mathcal{X})} &= KL(Q || P) \end{aligned}$$

Conjugate-Exponential models

Let's focus on *conjugate-exponential* (CE) models, which satisfy (1) and (2):

- **Condition (1).** The joint probability over variables is in the exponential family:

$$P(\mathcal{Y}, \mathcal{X} | \boldsymbol{\theta}) = f(\mathcal{Y}, \mathcal{X}) g(\boldsymbol{\theta}) \exp \{ \boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathcal{Y}, \mathcal{X}) \}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of *natural parameters*, \mathbf{u} are *sufficient statistics*

- **Condition (2).** The prior over parameters is conjugate to this joint probability:

$$P(\boldsymbol{\theta} | \eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu}) g(\boldsymbol{\theta})^\eta \exp \{ \boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu} \}$$

where η and $\boldsymbol{\nu}$ are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- η : number of pseudo-observations
- $\boldsymbol{\nu}$: values of pseudo-observations

Conjugate-Exponential examples

In the **CE** family:

- Gaussian mixtures
- factor analysis, probabilistic PCA
- hidden Markov models and factorial HMMs
- linear dynamical systems and switching models
- discrete-variable belief networks

Other as yet undreamt-of models can combine Gaussian, Gamma, Poisson, Dirichlet, Wishart, Multinomial and others.

Not in the **CE** family:

- Boltzmann machines, MRFs (no simple conjugacy)
- logistic regression (no simple conjugacy)
- sigmoid belief networks (not exponential)
- independent components analysis (not exponential)

Note: one can often approximate these models with models in the **CE** family.

A Useful Result

Given an iid data set $\mathcal{X} = (\mathcal{X}_1, \dots, \mathcal{X}_n)$, if the model is **CE** then:

(a) $Q_{\theta}(\theta)$ is also **conjugate**, i.e.

$$Q_{\theta}(\theta) = h(\tilde{\eta}, \tilde{\nu}) g(\theta)^{\tilde{\eta}} \exp \{ \phi(\theta)^{\top} \tilde{\nu} \}$$

where $\tilde{\eta} = \eta + n$ and $\tilde{\nu} = \nu + \sum_i \bar{\mathbf{u}}(\mathcal{Y}_i, \mathcal{X}_i)$.

(b) $Q_{\mathcal{Y}}(\mathcal{Y}) = \prod_{i=1}^n Q_{\mathcal{Y}_i}(\mathcal{Y}_i)$ is of the **same form** as in the E step of regular EM, but using **pseudo parameters** computed by averaging over $Q_{\theta}(\theta)$

$$Q_{\mathcal{Y}_i}(\mathcal{Y}_i) \propto f(\mathcal{Y}_i, \mathcal{X}_i) \exp \{ \bar{\phi}(\theta)^{\top} \mathbf{u}(\mathcal{Y}_i, \mathcal{X}_i) \} = P(\mathcal{Y}_i | \mathcal{X}_i, \bar{\phi}(\theta))$$

KEY points:

(a) the approximate parameter posterior is of the same form as the prior, so it is **easily summarized** in terms of two sets of hyperparameters, $\tilde{\eta}$ and $\tilde{\nu}$;

(b) the approximate hidden variable posterior, *averaging over all parameters*, is of the same form as the hidden variable posterior for a *single setting of the parameters*, so again, it is **easily computed** using the usual methods.

The Variational Bayesian EM algorithm

EM for MAP estimation

Goal: maximize $p(\boldsymbol{\theta}|\mathcal{X}, m)$ w.r.t. $\boldsymbol{\theta}$

E Step: compute

$$q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) = p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}^{(t)})$$

M Step:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) d\mathcal{Y}$$

Variational Bayesian EM

Goal: lower bound $p(\mathcal{X}|m)$

VB-E Step: compute

$$q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) = p(\mathcal{Y}|\mathcal{X}, \bar{\boldsymbol{\phi}}^{(t)})$$

VB-M Step:

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) = \exp \left[\int q_{\mathcal{Y}}^{(t+1)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta}) d\mathcal{Y} \right]$$

Properties:

- Reduces to the EM algorithm if $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.
- \mathcal{F}_m increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions (but not constrained to be Gaussian).
- VB-E step has same complexity as corresponding E step.
- We can use the junction tree, belief propagation, Kalman filter, etc, algorithms in the VB-E step of VB-EM, but **using expected natural parameters**, $\bar{\boldsymbol{\phi}}$.

Variational Bayes: History of Models Treated

- multilayer perceptrons (Hinton & van Camp, 1993)
- mixture of experts (Waterhouse, MacKay & Robinson, 1996)
- hidden Markov models (MacKay, 1995)
- other work by Jaakkola, Jordan, Barber, Bishop, Tipping, etc

Examples of Variational Learning of Model Structure

- mixtures of factor analysers (Ghahramani & Beal, 1999)
- mixtures of Gaussians (Attias, 1999)
- independent components analysis (Attias, 1999; Miskin & MacKay, 2000; Valpola 2000)
- principal components analysis (Bishop, 1999)
- linear dynamical systems (Ghahramani & Beal, 2000)
- mixture of experts (Ueda & Ghahramani, 2000)
- discrete graphical models (Beal & Ghahramani, 2002)
- **VIBES software** for conjugate-exponential graphs (Winn, 2003)

ARD for unsupervised learning

A idea similar to supervised ARD can be used with **Variational Bayesian** methods to learn the dimensionality of a latent space. Consider factor analysis:

$$\mathbf{x} \sim \mathcal{N}(\Lambda \mathbf{y}, \Psi) \quad \mathbf{y} \sim \mathcal{N}(0, I)$$

with a prior

$$\Lambda_i \sim \mathcal{N}(0, \alpha_i^{-1} I)$$

The VB free energy is a function of the data, $Q_{\mathcal{Y}}(\mathcal{Y})$, $Q_{\Lambda}(\Lambda)$ **and** **α** :

$$\mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_{\Lambda}(\Lambda), \mathcal{X}, \alpha) = \langle \log P(\mathcal{X}, \mathcal{Y} | \Lambda, \Psi) + \log P(\Lambda | \alpha) + \log P(\Psi) \rangle_{Q_{\mathcal{Y}} Q_{\Lambda}} + \mathbf{H}[Q_{\mathcal{Y}}] + \mathbf{H}[Q_{\Lambda}]$$

Optimising this wrt the distributions and α in turn (like EM) causes some α_i to diverge, restricting the effective dimensionality of \mathbf{y} .