Variational Methods ⇏ Approximate

not intrinsically

But naturally leads to Approximate ⇐ Find the best class of function to fit posterior

Probabilistic & Unsupervised Learning

Factored Variational Approximations and Variational Bayes

KEY to variational approximation:
Remove as few arcs as possible from the moral graph s.t. inference becomes tractable.

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

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

Term 1, Autumn 2017

Expectations 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

Bishop.

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

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

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

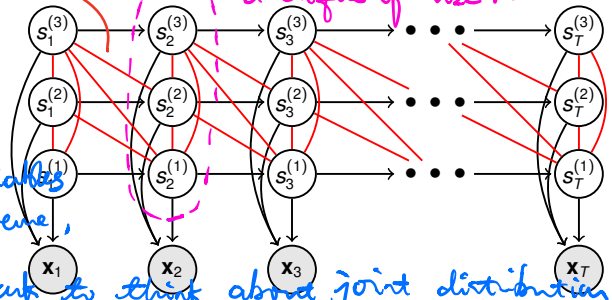
- Computing the conditional probabilities in a very large multiply-connected DAG:

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

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

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

Distributed models



Consider an FHMM with M state variables taking on K values each.

- Moralisation puts simultaneous states $(s_t^{(1)}, s_t^{(2)}, \ldots, s_t^{(M)})$ into a single clique
- Triangulation extends cliques to size $M + 1$
- Each state takes K values $\Rightarrow$ sums over K^{M+1} 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 wrt θ (i.e. ML learning):

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

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:

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

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

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

Recall: maximizing the lower bound of free energy is equivalent to minimizing KL divergence

The log likelihood is bounded below by:

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

EM alternates between:

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

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

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

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

EM Never Decreases the Likelihood

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

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

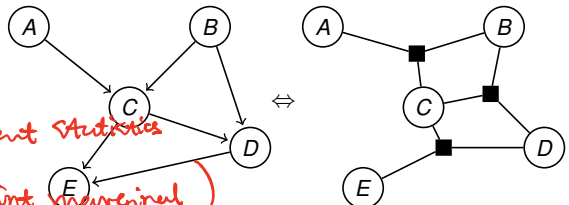
- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

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.

Intractability

The M-step for a graphical model is usually (relatively) easy.

Hard thing:
Too many number of steps in inference
(Finding joint marginals / sufficient statistics of the joint marginal)



Easy thing:
Learning is quite straightforward.

$$P(A,B,C,D,E) = \underbrace{P(A)P(B)P(C|A,B)}_{f_1(A,B,C)}\underbrace{P(D|B,C)}_{f_2(B,C,D)}\underbrace{P(E|C,D)}_{f_3(C,D,E)}$$

- Need expected sufficient stats from marginal posteriors on each factor group.
- Then (at least for a DAG) can optimise each factor parameter vector separately.
- Intractability in EM comes from the difficulty of computing marginal posteriors in graphs with large tree-width or non-linear/non-conjugate conditionals.
- [For non-DAG models, partition function (normalising constant) may also be intractable.]

Free-energy-based variational approximation

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

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

- Each step increases the likelihood.
- A fixed point of the gradient M-step must be at a mode of the expected log-joint.

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 either case, we choose q from within a limited set $\mathcal{Q}$:

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

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})\in\mathcal{Q}\leftarrow\textbf{Constraint}}{\operatorname{argmax}} \mathcal{F}\left(q(\mathcal{Y}), \theta^{(k-1)}\right).$$

M step: unchanged

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

Unlike in GEM, the fixed point may not be at an unconstrained optimum of $\mathcal{F}$.

What do we lose?

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

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

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

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.

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

- This means 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 $P(\mathcal{Y}|\mathcal{X}, \theta^{\mathrm{ML}}) \in \mathcal{Q}$, 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:

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

is equivalent to:

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

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

So, in each E step, the algorithm is trying to find the best approximation to $P(\mathcal{Y}|\mathcal{X})$ in $\mathcal{Q}$ in a KL sense. This is related to ideas in *information geometry*. It also suggests generalisations to other distance measures.

Factored Variational E-step

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

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

Rather than jointly optimise the free energy w.r.t. all of the distributions, we think as iteratively update w.r.t just one of the distributions at a time. E-step itself becomes iterative

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:

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

Iteratively update E step to convergence, then do M-step

Iterate one distribution each time.

Closed form updates for each one of the factors

- q_i updates iterated to convergence to "complete" VE-step.
- In fact, every (VE)$_i$-step separately increases $\mathcal{F}$, so any schedule of (VE)$_i$- and M-steps will converge. Choice can be dictated by practical issues (rarely efficient to fully converge E-step before updating parameters).

As long as we visit all factors

See Ass5. Q1(a).

SEE
Matthew Beal
2003
Chapter 2.2

Factored Variational E-step

The Factored Variational E-step has a general form.

The free energy is: Factored q

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

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

$$\mathcal{H}[q_i] = -\int q_i(\mathcal{Y}_i)\log q_i(\mathcal{Y}_i)\, d\mathcal{Y}_i$$

Look at one of the factors, $\mathcal{Y}_i$

Target:
Find functional form of q_i that maximise the free energy.

Functional Directive

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

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

Expectation w.r.t all other distribution

Constraint of being valid distribution.

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

closed form solution, unique optimum, also global optimum

Function of $\mathcal{Y}_i$ itself

All other factors apart from $\mathcal{Y}_i$ are integrated over

In general, this depends only on the expected sufficient statistics under q_i. Thus, again, we don't actually need the *entire* distributions, just the relevant expectations (now for approximate inference as well as learning).

only thing left is $\mathcal{Y}_i$

* Functional Example : Entropy

$$\mathcal{H}[p] = -\int p(x) \ln p(x) dx \qquad \mathcal{H} : \mathcal{P} \to \mathbb{R}^+$$

* Derivation of ⊛

$$\mathcal{F}(q(\underline{Y}), \theta^{(k-1)}) = \left\langle \log P(\underline{X}, \underline{Y} | \theta^{(k-1)}) \right\rangle_{\prod_j q_j(Y_j)} + \mathcal{H}[q(\underline{Y})]$$

$$\Leftrightarrow \mathcal{F}\left(\prod_j q_j(Y_j), \theta^{(k-1)}\right) = \left\langle \log P(\underline{X}, \underline{Y} | \theta^{(k-1)} \right\rangle_{\prod_j q_j(Y_j)} + \mathcal{H}\left[\prod_j q_j(Y_j)\right]$$

$$= \int q_i(Y_i) \cdot \left\langle \log P(\underline{X}, \underline{Y} | \theta^{(k-1)}) \right\rangle_{\prod_{j \neq i} q_j(Y_i)} dY_i$$

$$+ \mathcal{H}[q_i] + \sum_{j \neq i} \mathcal{H}[q_j]$$

Notice lagrangian constraint : $\int q_i(Y_i) dY_i = 1$.

Also functional derivative : $\frac{\partial}{\partial q_i} \int \mathcal{F}\, q_i(Y_i) dY_i = \mathcal{F}$. And $\frac{\partial}{\partial q_i} \int q_i(Y_i) dY_i = 1$

$$\frac{\partial}{\partial q_i}\left(\mathcal{F} + \lambda\left(\int q_i(Y_i) dY_i - 1\right)\right)$$

not related ↓

$$= \frac{\partial}{\partial q_i}\left(\int q_i(Y_i) \left\langle \log P(\underline{X}, \underline{Y} | \theta^{(k-1)}) \right\rangle_{\prod_{j \neq i} q_j(Y_j)} dY_i - \int q_i(Y_i) \log q_i(Y_i) dY_i - \sum_{j \neq i} \mathcal{H}[q_j]\right.$$

$$\left. + \lambda\left(\int q_i(Y_i) dY_i - 1\right)\right) = 0$$

$$= \left\langle \log P(\underline{X}, \underline{Y} | \theta^{(k-1)}) \right\rangle_{\prod_{j \neq i} q_j(Y_j)} - \log q_i(Y_i) + \lambda = 0.$$

$$\Rightarrow q_i(Y_i) \propto \exp\left\{ \left\langle \log P(\underline{X}, \underline{Y} | \theta^{(k-1)}) \right\rangle_{\prod_{j \neq i} q_j(Y_j)} \right\}$$

Functional Derivative:

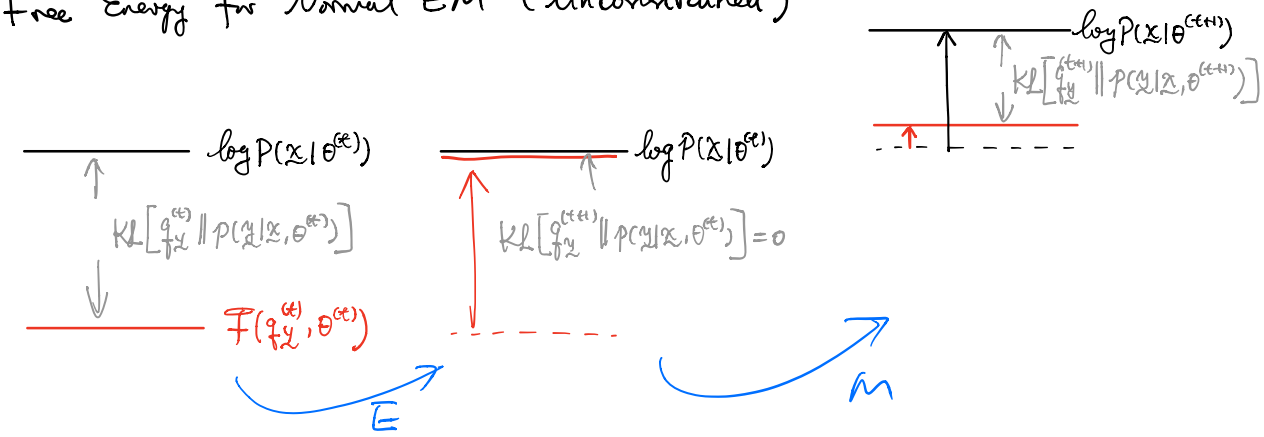
$\mathcal{F}: \mathcal{M} \to \mathbb{R} / \mathbb{C}$ as a functional.

the functional derivative of $\mathcal{F}(f)$, denoted by $\frac{\delta F}{\delta f}$ is defined by:

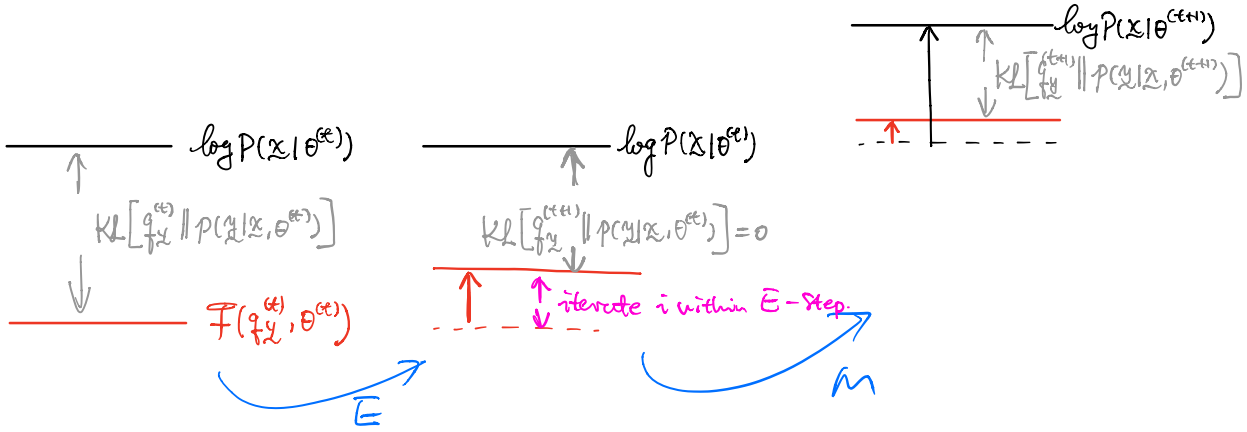
$$\frac{\delta F}{\delta f} = \int \frac{\delta F}{\delta f}(x) \cdot \phi(x)\, dx = \lim_{\varepsilon \to 0} \frac{\mathcal{F}(f+\varepsilon\phi) - \mathcal{F}(f)}{\varepsilon} = \left[\frac{d}{d\varepsilon} F(f+\varepsilon\phi)\right]_{\varepsilon=0}$$

where ϕ is an arbitrary function, and $\varepsilon\phi$ is called the variation of f.

Free Energy for Normal EM (Unconstrained)



Free Energy for constrained EM $Q_Y(Y) = \prod_i Q_{Y_i}(Y_i)$.



Mean-field approximations

Factored to an extreme level that each set has just one latent variable

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.

In Boltzmann Machine:

- Suppose $P(\mathcal{X}, \mathcal{Y})$ has sufficient statistics that are separable in the latent variables: *e.g.* the Boltzmann machine

Sufficient statistics

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

(Iteration of the E-step)

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

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

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

Iteration through this process eventually will find a self consistent point.

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

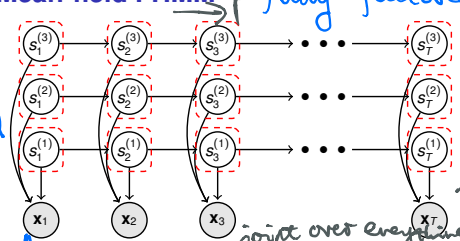
- Thus, we can update each q_i in turn given the means (or, in general, mean sufficient statistics) of the others.
- Each variable sees the mean field imposed by its neighbours, and we update these fields until they all agree.

* Find the mean value of one of the latent variable conditioned on the mean values of everything on its Markov Blanket.

Fully factored case performs better than the structured case.

While factor over T or M brings bias in different direction

Mean-field FHMM



Fully factored to single nodes.

$P(A,B) = P(A|B) \cdot P(B)$.

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

$$P(S_{1:T}^{1:M}, X_{1:T}) = P(X_{1:T} | S_{1:T}^{1:M}) P(S_{1:T}^{1:M}) = \left(\prod_{\tau=1}^{T} P(x_\tau | S_\tau^{1:M})\right) \cdot \left(\prod_{m=1}^{M} \prod_{\tau=1}^{T} P(S_\tau^m | S_{\tau-1}^m)\right)$$

joint over everything

$$q_t^m(s_t^m) \propto \exp\left\langle \log P(\mathbf{s}_{1:T}^{1:M}, \mathbf{x}_{1:T}) \right\rangle_{\prod_{\neg(m,t)} q_{t'}^{m'}(s_{t'}^{m'})}$$

$$= \exp\left\langle \left(\sum_\mu \sum_\tau \log P(s_\tau^\mu | s_{\tau-1}^\mu) + \sum_\tau \log P(\mathbf{x}_\tau | s_\tau^{1:M})\right) \right\rangle_{\prod_{\neg(m,t)} q_{t'}^{m'}}$$

Backward Message

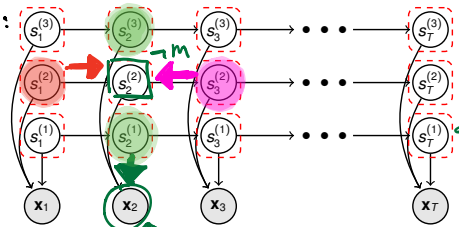
Message from the past

$$\propto \exp\left[\underbrace{\langle \log P(s_t^m | s_{t-1}^m) \rangle_{q_{t-1}^m} + \langle \log P(\mathbf{x}_t | s_t^{1:M}) \rangle_{\prod_{\neg m} q_t^{m'}}}_{\alpha_t^m(i) \propto e^{\sum_j \log \Phi_{ji}^m q_{t-1}^m(j)} \; e^{\langle \log A_i(\mathbf{x}_t) \rangle_{q_t^{\neg m}}}} + \underbrace{\langle \log P(s_{t+1}^m | s_t^m) \rangle_{q_{t+1}^m}}_{\beta_t^m(i) \propto e^{\sum_j \log \Phi_{ij}^m q_{t+1}^m(j)}} \right]$$

Cf. forward-backward: $\alpha_t(i) \propto \sum_j \alpha_{t-1}(j) \Phi_{ji} \cdot A_j(\mathbf{x}_t)$ $\quad \beta_t(i) \propto \sum_j \Phi_{ij} A_j(\mathbf{x}_{t+1}) \beta_{t+1}(j)$

Message from current observation

Mean-field FHMM



Difference between forward-backward ①:

* Since our forward message depends on current estimates of previous time. After doing a forward-backward sweep, the next forward message needs to be different. it's an ITERATIVE process with multiple forward-backward sweep

Difference ②: Evidence term $A_i(x_t)$ appears once in mean-field FHMM, but appear from both forward/backward messages in the normal case

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

The only term that the chain influence with each other

observation

$$q_t^m(s_t^m) \propto \exp\Big[\underbrace{\langle \log P(s_t^m|s_{t-1}^m)\rangle_{q_{t-1}^m} + \langle \log P(\mathbf{x}_t|s_t^{1:M})\rangle_{\prod_{\neg m} q_t^{m'}}}_{\alpha_t^m(i) \propto e^{\sum_j \log \Phi_{ji}^m q_{t-1}^m(j)} \cdot e^{\langle \log A_j(\mathbf{x}_t)\rangle_{q_t^{\neg m}}}} + \underbrace{\langle \log P(s_{t+1}^m|s_t^m)\rangle_{q_{t+1}^m}}_{\beta_t^m(i) \propto e^{\sum_j \log \Phi_{ij}^m q_{t+1}^m(j)}}\Big]$$

Evidence term

Cf. forward-backward: $\alpha_t(i) \propto \sum_j \alpha_{t-1}(j)\Phi_{ji} \cdot A_i(\mathbf{x}_t)$ $\quad \beta_t(i) \propto \sum_j \Phi_{ij} A_j(\mathbf{x}_{t+1})\beta_{t+1}(j)$

But NOT exact forward-backward.

- Yields a message-passing algorithm like forward-backward
- Updates depend only on immediate neighbours in chain
- Chains couple only through joint output
- Multiple passes; messages depend on (approximate) marginals
- Evidence does not appear explicitly in backward message (cf Kalman smoothing)

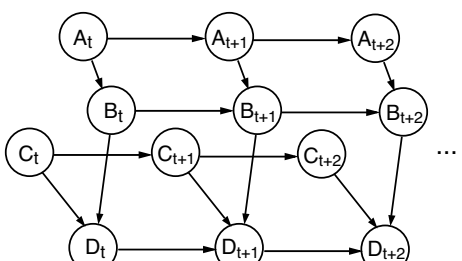
Structured variational approximation

- $q(\mathcal{Y})$ need not be completely factorized.
- For example, suppose $\mathcal{Y}$ can be partitioned into sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ such that computing the expected sufficient statistics under $P(\mathcal{Y}_1|\mathcal{Y}_2, \mathcal{X})$ and $P(\mathcal{Y}_2|\mathcal{Y}_1, \mathcal{X})$ would be tractable.

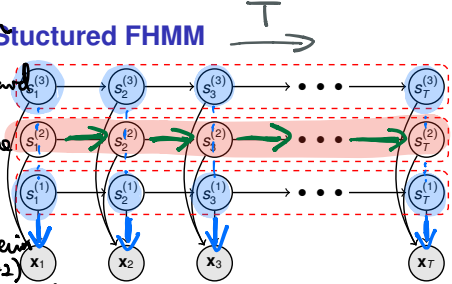
Theorem :

$\Rightarrow$ Then the factored approximation $q(\mathcal{Y}) = q(\mathcal{Y}_1)q(\mathcal{Y}_2)$ is tractable.

- In particular, any factorisation of $q(\mathcal{Y})$ into a product of distributions on trees, yields a tractable approximation.



Compare with the mean field case, we do a forward backward one for the whole chain. Then do again for the next chain (M=1) (M=2). After doing forward-backward for all M=1,2,3, we Repeat the process again until they all agree as in the mean field case.



For the FHMM we can factor the chains:

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

$$q^m(s_{1:T}^m) \propto \exp\left\langle \log P(\mathbf{s}_{1:T}^{1:M}, \mathbf{x}_{1:T}) \right\rangle_{\prod_{\neg m} q^{m'}(s_{1:T}^{m'})}$$

q of a chain

$$= \exp\left\langle \sum_{\mu}\sum_{t} \log P(s_t^{\mu}|s_{t-1}^{\mu}) + \sum_{t} \log P(\mathbf{x}_t|s_t^{1:M}) \right\rangle_{\prod_{\neg m} q^{m'}}$$

Keep the m'th term

$$\propto \exp\left[\sum_{t} \log P(s_t^m|s_{t-1}^m) + \sum_{t} \left\langle \log P(\mathbf{x}_t|s_t^{1:M}) \right\rangle_{\prod_{\neg m} q^{m'}(s_t^{m'})} \right]$$

$\sum_t$: Sum over the chain

Remove M=1

Time factor

$$= \prod_{t} P(s_t^m|s_{t-1}^m) \prod_{t} e^{\left\langle \log P(\mathbf{x}_t|s_t^{1:M}) \right\rangle_{\prod_{\neg m} q^{m'}(s_t^{m'})}}$$

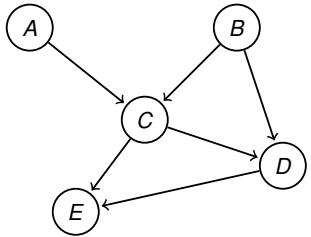
$$\exp\left\{ \left\langle \sum_{\mu=1}^{M}\sum_{t=1}^{T} \log P(S_t^{\mu}|S_{t-1}^{\mu}) \right\rangle_{\prod_{\neg m} q^{m'}} \right\}$$

$$= \exp\left\{ \sum_{t=1}^{T} \log P(S_t^m|S_{t-1}^m) \right\}.$$

This looks like a standard HMM joint, with a modified likelihood term ⇒ cycle through multiple forward-backward passes, updating likelihood terms each time.

Messages on an arbitrary graph

Consider a DAG:



$$P(\mathcal{X}, \mathcal{Y}) = \prod_k P(Z_k | \text{pa}(Z_k))$$

and let $q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i)$ for disjoint sets $\{\mathcal{Y}_i\}$.

We have that the VE update for q_i is given by $q_i^*(\mathcal{Y}_i) \propto \exp \langle \log p(\mathcal{Y}, \mathcal{X}) \rangle_{q_{\neg i}(\mathcal{Y})}$ where $\langle \cdot \rangle_{q_{\neg i}(\mathcal{Y})}$ denotes averaging with respect to $q_j(\mathcal{Y}_j)$ for all $j \neq i$

Then:

$$\log q_i^*(\mathcal{Y}_i) = \left\langle \sum_k \log P(Z_k | \text{pa}(Z_k)) \right\rangle_{q_{\neg i}(\mathcal{Y})} + \text{const}$$

$$= \sum_{j \in \mathcal{Y}_i} \langle \log P(Y_j | \text{pa}(Y_j)) \rangle_{q_{\neg i}(\mathcal{Y})} + \sum_{j \in \text{ch}(\mathcal{Y}_i)} \langle \log P(Z_j | \text{pa}(Z_j)) \rangle_{q_{\neg i}(\mathcal{Y})} + \text{const}$$

Theorem: This defines messages that are passed between nodes in the graph. Each node receives messages from its Markov boundary: parents, children and parents of children (all neighbours in the corresponding factor graph).

Further, if the conditional distribution have conjugate exponential structure, then the variational passing scheme

Co-parent of node j: the other parents of the child nodes apart from j itself

Co-parents.

Parents

Children

child (Y_i)

The update of factors in the variational posterior distribution represents a local calculation on the graph

Non-factored variational methods

The term variational approximation is used whenever a bound on the likelihood (or on another estimation cost function) is optimised, but does not necessarily become tight.

Many further variational approximations have been developed, including:

- parametric forms (e.g. Gaussian) for non-linear models
- non-free-energy-based bounds (both upper and lower) on the likelihood.

Rather than learning datasets/iid observations, it learns a seperate machine takes data

We can also see MAP- or zero-temperature EM and recognition models as parametric forms of variational inference.

produces parameters of variational distribution as output

constrained cases are just delta functions

Guaranteed to converge as it always pushes up free energy

Variational methods can also be used to find an approximate posterior on the parameters.

Previous: use factorisation to find distributions over latent variables (Approximate posterior on latent variables)

Then: use it to find distribution over our parameters. (Approximate posterior on parameters)

Enlarge the graph, often factorisation makes it easy ... ??

Previous: Integrate Free energy (log-joint) over latent.

Now: Integrate evidence/log-likelihood over latent & parameters.

Variational Bayes

Approximate Posterior on PARAMETERS

So far, we have applied Jensen's bound and factorisations to help with integrals over latent variables.

We can do the same for integrals over parameters in order to bound the log marginal likelihood or evidence.

Bound Evidence

Evidence over a model class

$$\log P(\mathcal{X}|\mathcal{M}) = \log \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; P(\mathcal{X},\mathcal{Y}|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M}) = \log \iint p(x,y,\theta|M)\,dy\,d\theta$$

$$= \max_{Q} \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q(\mathcal{Y},\boldsymbol{\theta}) \log \frac{P(\mathcal{X},\mathcal{Y},\boldsymbol{\theta}|\mathcal{M})}{Q(\mathcal{Y},\boldsymbol{\theta})}$$

$$\geq \max_{Q_{\mathcal{Y}},Q_{\boldsymbol{\theta}}} \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y})Q_{\theta}(\boldsymbol{\theta}) \log \frac{P(\mathcal{X},\mathcal{Y},\boldsymbol{\theta}|\mathcal{M})}{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

← Integrate out both the hidden variables and parameters

☆ The constraint that the distribution Q must factor into the product $Q_{\mathcal{Y}}(\mathcal{Y})Q_{\theta}(\boldsymbol{\theta})$ leads to the **variational Bayesian EM algorithm** or just **"Variational Bayes"**.

$Q(y,\theta) = Q_Y(y)Q_\theta(\theta)$

Some call this the "Evidence Lower Bound" (ELBO). I'm not fond of that term

联合调3.

$$\log P(X|M) = \log \iint P(X,Y,\theta|M)\, dy\, d\theta$$

$$= \log \iint Q(Y,\theta) \cdot \frac{P(X,Y,\theta|M)}{Q(Y,\theta)}\, dy\, d\theta$$

Jensen's Inequality.

$$\geq \iint Q(Y,\theta) \cdot \log\left(\frac{P(X,Y,\theta|M)}{Q(Y,\theta)}\right) dy\, d\theta$$

Factorise $Q(Y,\theta) = Q_Y(Y) \cdot Q_\theta(\theta)$.

"My distribution over latent variables are independent of parameters, hence gives a lower bound. Seldom happens... So equality almost never holds"

$$\geq \iint Q_Y(Y) \cdot Q_\theta(\theta) \cdot \log\left(\frac{P(X,Y,\theta|M)}{Q_Y(Y) \cdot Q_\theta(\theta)}\right) dy\, d\theta .$$

$\stackrel{\text{def}}{=} \mathcal{F}(Q_Y(Y), Q_\theta(\theta))$

The VB-Free Energy lower bound.

Model-class Marginal likelihood / Evidence bounded by VB-Free Energy.

Variational Bayesian EM . . .

Coordinate maximization of the VB free-energy lower bound

$$\mathcal{F}(Q_{\mathcal{Y}}, Q_{\boldsymbol{\theta}}) = \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{p(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}|\mathcal{M})}{Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

leads to **EM-like** updates: (Instead of updating single value, we update whole distribution)

$$\begin{cases} Q^*_{\mathcal{Y}}(\mathcal{Y}) \propto \exp \langle \log P(\mathcal{Y}, \mathcal{X}|\boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} & \text{E-like step} \\ Q^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}) \exp \langle \log P(\mathcal{Y}, \mathcal{X}|\boldsymbol{\theta}) \rangle_{Q_{\mathcal{Y}}(\mathcal{Y})} & \text{M-like step} \end{cases}$$

Prior on θ

↑ Instead of optimising single θ, we optimise a distribution of θ.

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

Detailed learning rule

$$\log P(\mathcal{X}) - \mathcal{F}(Q_{\mathcal{Y}}, Q_{\boldsymbol{\theta}}) = \log P(\mathcal{X}) - \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta})}{Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

$$= \iint Q_Y Q_\theta \log \left(\frac{Q_Y Q_\theta \cdot P(X)}{P(X, Y, \theta)} \right) dY d\theta = \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{Q_{\mathcal{Y}}(\mathcal{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{P(\mathcal{Y}, \boldsymbol{\theta}|\mathcal{X})} = KL\left[Q_Y Q_\theta \,\|\, P(Y, \theta|X) \right]$$

$$\log P(X) = KL(Q\|P) + \mathcal{F}(Q_Y, Q_\theta)$$

Beal 2003. Theorem 2.1 : Learning rule of Variational Bayes EM:

Let M be model with parameters θ giving rise to i.i.d observed data $\underline{X} = \{x_1, \cdots, x_n\}$ with latent variables $\underline{Y} = \{y_1, y_2, \cdots, y_n\}$.

We have the log marginal lower bound / Evidence lower bound / VB free energy:

$$\mathcal{F}(Q_Y(Y), Q_\theta(\theta)) = \iint Q_Y(Y)\, Q_\theta(\theta) \cdot \log\left(\frac{P(X, Y, \theta | M)}{Q_Y(Y) Q_\theta(\theta)}\right) dY\, d\theta$$

Which can be optimised by the following step, with t be number of iteration:

VBE - step: $q_{y_i}^{(t+1)}(y_i) = \frac{1}{Z_{y_i}} \exp\left(\int q_\theta^{(t)}(\theta) \cdot \log P(x_i, y_i | \theta, M)\, d\theta\right)$ $\forall i$. ← Iteration in VBE - step.

where $q_Y^{(t+1)}(Y) = \prod_{i=1}^{n} q_{y_i}^{(t+1)}(y_i)$.

VBM - step: $q_\theta^{(t+1)}(\theta) = \frac{1}{Z(\theta)} P(\theta | M) \exp\left(\int q_Y(Y) \log P(X, Y | \theta, M)\, dY\right)$ ← one step in VBM

Prior on θ

Moreover, the learning rule converges to a LOCAL maximum of $\mathcal{F}(Q_Y(Y), Q_\theta(\theta))$

Proof: * VBE:

$$\frac{\delta}{\delta Q_Y(Y)} \mathcal{F}(Q_Y(Y), Q_\theta(\theta)) = \int Q_\theta(\theta) d\theta \cdot \left[\frac{\delta}{\delta Q_Y(Y)} \int Q_Y(Y) \cdot \log\left(\frac{P(X, Y, \theta | M)}{Q_Y(Y) Q_\theta(\theta)}\right) dY\right]$$

$$= \int Q_\theta(\theta) d\theta \cdot \left[\frac{\delta}{\delta Q_Y(Y)} \int Q_Y(Y) \cdot \log\left(\frac{P(X, Y | \theta, M)}{Q_Y(Y)}\right) dY\right]$$

$$= \int Q_\theta(\theta) d\theta \cdot \left[\log\left(\frac{P(X, Y | \theta, M)}{Q_Y(Y)}\right) + \underbrace{Q_Y(Y) \cdot \left(-\frac{1}{Q_Y(Y)}\right)}_{= -1}\right]$$

$$= \int Q_\theta(\theta) d\theta \cdot \left[\log P(X, Y | \theta, M) - \log Q_Y(Y) - 1\right] = 0.$$

$$\Leftrightarrow \log Q_Y^{(t+1)}(Y) = \int \log P(X, Y | \theta, M) \cdot Q_\theta^{(t)}(\theta) \cdot d\theta + \text{const}.$$

t: outer loop

Into factored form:

$$\Leftrightarrow \log Q_{y_i}^{(t+1)}(y_i) = \int \log P(x_i, y_i | \theta, M) \cdot Q_\theta^{(t)}(\theta) d\theta + \text{const}. \quad \forall i.$$

i: iteration within VBE-step

* VBM:

$$\frac{\delta}{\delta Q_\theta(\theta)} \mathcal{F}(Q_Y(Y), Q_\theta(\theta)) = \frac{\delta}{\delta Q_\theta(\theta)} \int d\theta\, Q_\theta(\theta) \left[\int Q_Y(Y) \log \left(\frac{P(X,Y,\theta|M)}{Q_\theta(\theta) Q_Y(Y)} \right) dY \right]$$

$$= \frac{\delta}{\delta Q_\theta(\theta)} \int d\theta\, Q_\theta(\theta) \left[\int Q_Y(Y) \log \left(\frac{P(X,Y|\theta) \cdot P(\theta|M)}{Q_\theta(\theta) \cdot Q_Y(Y)} \right) dY \right]$$

移出 Y

$$= \frac{\delta}{\delta Q_\theta(\theta)} \int d\theta\, Q_\theta(\theta) \left[\int Q_Y(Y) \log P(X,Y|\theta) dY + \int Q_Y(Y) \log \left(\frac{P(\theta|M)}{Q_\theta(\theta) Q_Y(Y)} \right) dY \right]$$

$$= \frac{\delta}{\delta Q_\theta(\theta)} \int d\theta\, Q_\theta(\theta) \left[\int Q_Y(Y) \log P(X,Y|\theta) dY + \log \left(\frac{P(\theta|M)}{Q_\theta(\theta)} \right) \right]$$

$$= \int Q_Y(Y) \log P(X,Y|\theta) dY + \log P(\theta|M) - \log Q_\theta(\theta) + \text{const.} = 0$$

$$\Leftrightarrow \log Q_\theta^{(t+1)}(\theta) = \log P(\theta|M) + \int Q_Y^{(t+1)}(Y) \log P(X,Y|\theta) dY + \text{const.}$$

Prior of θ ; outer loop.

VBM-step is just optimise over whole $Q_\theta(\theta)$ distribution.

So just do one step.

Therefore, the VBEM algorithm is:

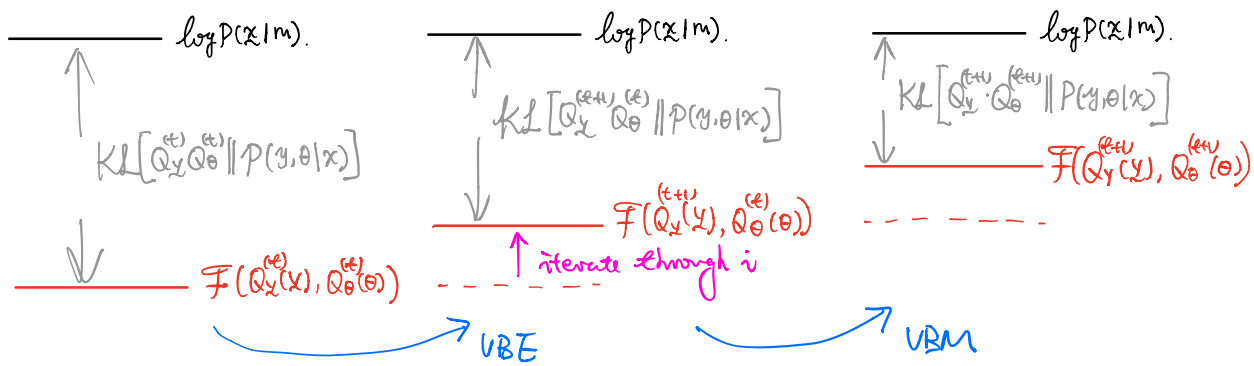
while VB-free energy not converge

VBE [
  for $i = 1:n$
    $\log Q_{Y_i}^{(t+1)}(Y_i) = \int \log P(X_i, Y_i|\theta, M) \cdot Q_\theta^{(t)}(\theta) d\theta + \text{const.}$
  end
]

VBM [$\log Q_\theta^{(t+1)}(\theta) = \log P(\theta|M) + \int Q_Y^{(t+1)}(Y) \log P(X,Y|\theta) dY + \text{const.}$]

end

} gives fixed point iteration.



Conjugate-Exponential models

Let's focus on *conjugate-exponential* (**CE**) latent-variable models:

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

$$P(\mathcal{Y}, \mathcal{X}|\boldsymbol{\theta}) = f(\mathcal{Y}, \mathcal{X})\ g(\boldsymbol{\theta}) \exp\left\{\phi(\boldsymbol{\theta})^{\mathsf{T}} \mathsf{T}(\mathcal{Y}, \mathcal{X})\right\}$$

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

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

$$P(\boldsymbol{\theta}|\nu, \tau) = h(\nu, \tau)\ g(\boldsymbol{\theta})^{\nu} \exp\left\{\phi(\boldsymbol{\theta})^{\mathsf{T}} \tau\right\}$$

  where ν and τ are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- ν: number of pseudo-observations
- τ: values of pseudo-observations

CE :

Exact Inference or Variational Inference.

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

} closed form for variational Bayes

▶ LDA (can do variational LDA).

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

Non-CE :

Variational Inference

Not in the **CE** family: Cannot find normaliser

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

} Non-closed form

Numerical Issues

Note: one can often approximate such models with a suitable choice from the **CE** family.

Conjugate-exponential VB

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

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

$$\begin{aligned} Q_\theta(\boldsymbol{\theta}) &\propto P(\boldsymbol{\theta}) \cdot \exp\left\langle \sum_i \log P(\mathbf{y}_i, \mathbf{x}_i|\boldsymbol{\theta}) \right\rangle_{Q_\mathcal{Y}} \\ &= h(\nu, \boldsymbol{\tau}) g(\boldsymbol{\theta})^\nu e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\mathsf{T} \boldsymbol{\tau}} \quad g(\boldsymbol{\theta})^n e^{\left\langle \log f(\mathcal{Y}, \mathcal{X}) \right\rangle_{Q_\mathcal{Y}}} e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\mathsf{T} \left\langle \sum_i \mathsf{T}(\mathbf{y}_i, \mathbf{x}_i) \right\rangle_{Q_\mathcal{Y}}} \\ &\propto h(\tilde{\nu}, \tilde{\boldsymbol{\tau}}) g(\boldsymbol{\theta})^{\tilde{\nu}} e^{\boldsymbol{\phi}(\boldsymbol{\theta})^\mathsf{T} \tilde{\boldsymbol{\tau}}} \end{aligned}$$

  with $\tilde{\nu} = \nu + n$ and $\tilde{\boldsymbol{\tau}} = \tau + \sum_i \langle \mathsf{T}(\mathbf{y}_i, \mathbf{x}_i') \rangle_{Q_\mathcal{Y}}$ $\Rightarrow$ only need to track $\tilde{\nu}, \tilde{\boldsymbol{\tau}}$.

- $Q_\mathcal{Y}(\mathcal{Y}) = \prod_{i=1}^n Q_{\mathbf{y}_i}(\mathbf{y}_i)$ takes the same form as in the E-step of regular EM

$$\begin{aligned} Q_{\mathbf{y}_i}(\mathbf{y}_i) &\propto \exp \langle \log P(\mathbf{y}_i, \mathbf{x}_i|\boldsymbol{\theta}) \rangle_{Q_\theta} \\ &\propto f(\mathbf{y}_i, \mathbf{x}_i) e^{\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{Q_\theta}^\mathsf{T} \mathsf{T}(\mathbf{y}_i, \mathbf{x}_i)} = P(\mathbf{y}_i|\mathbf{x}_i, \overline{\boldsymbol{\phi}}(\boldsymbol{\theta})) \end{aligned}$$

  with natural parameters $\overline{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{Q_\theta}$ $\Rightarrow$ inference unchanged from regular EM.

* Complete-data likelihood:

$$p(x_i, y_i|\theta) = g(\theta) \cdot f(x_i, y_i) \cdot \exp\left(\phi(\theta)^T T(x_i, y_i)\right) \quad \text{where} \quad g(\theta)^{-1} = \int\int f(x_i, y_i) \exp\left(\phi(\theta)^T T(x_i, y_i)\right) dx_i dy_i$$

* Parameter-prior:

$$p(\theta|\nu, \tau) = h(\nu, \tau) g(\theta)^\nu \exp\left(\phi(\theta)^T \tau\right). \quad \text{where} \quad h(\nu, \tau)^{-1} = \int g(\theta)^\nu \exp\left(\phi(\theta)^T \tau\right) d\theta$$

VBE step:

$$Q_{y_i}^{(t)}(y_i) \propto \exp\left(\left\langle \log p(x_i, y_i|\theta^{(t)}) \right\rangle_{Q_\theta}\right) \qquad \forall i$$

$$\propto f(x_i, y_i) \exp\left(\bar{\phi}^T \cdot T(x_i, y_i)\right), \quad \text{where} \quad \bar{\phi} = \int Q_\theta(\theta)\phi(\theta) d\theta \qquad \forall i$$

$$= p(y_i|x_i, \bar{\phi}(\theta)) \qquad \forall i \qquad = \left\langle \phi(\theta) \right\rangle_{Q_\theta(\theta)}.$$

rewrite as

VBM step:

$$Q_\theta^{(t+1)}(\theta) \propto p(\theta|M) \cdot \exp\left(\left\langle \log p(X, Y|\theta, M) \right\rangle_{Q_X^{(t)}(X)}\right)$$

$$= h(\nu, \tau) \cdot g(\theta)^\nu \exp\left(\phi(\theta)^T \tau\right) \cdot \exp\left(\left\langle \sum_{i=1}^n \log p(x_i, y_i|\theta, M) \right\rangle_{Q_{y_i}^{(t)}(y_i)}\right)$$

$$= h(\nu, \tau) \cdot g(\theta)^\nu \exp\left(\phi(\theta)^T \tau\right) \cdot g(\theta)^n \cdot \exp\left(\left\langle \sum_{i=1}^n f(x_i, y_i) \right\rangle_{Q_{y_i}^{(t)}}\right) \cdot \exp\left(\phi(\theta)^T \cdot \sum_{i=1}^n \left\langle T(x_i, y_i) \right\rangle_{Q_{y_i}^{(t)}}\right)$$

no θ dependees.

$$\propto h(\tilde{\nu}, \tilde{\tau}) \cdot g(\theta)^{\nu+n} \cdot \exp\left(\phi(\theta)^T \left[\tau + \sum_{i=1}^n \left\langle T(x_i, y_i) \right\rangle_{Q_{y_i}^{(t)}}\right]\right)$$

$$\tilde{\nu} = \nu + n \quad \text{and} \quad \tilde{\tau} = \tau + \sum_{i=1}^n \left\langle T(x_i, y_i) \right\rangle_{Q_{y_i}^{(t)}}.$$

Summary of VBEM for CE Models

While VB free energy not converge (loop over t).

VBE
- for $i = 1:n$
  - compute sufficient statistics $\left\langle T(x_i, y_i) \right\rangle_{Q_{y_i}^{(t)}}$
- end
- $\nu \mathrel{+}= n$, $\tau \mathrel{+}= \sum_i \left\langle T(x_i, y_i) \right\rangle_{Q_{y_i}^{(t)}}$

VBM
- compute $\bar{\phi} = \left\langle \phi(\theta) \right\rangle$ under ν and τ

end

The Variational Bayesian EM algorithm

The $Q_Y(\mathcal{Y})$ in EM and VB-EM are exactly the same

EM for MAP estimation (for CE)

Goal: maximize $P(\boldsymbol{\theta}|\mathcal{X}, m)$ wrt $\boldsymbol{\theta}$

E Step: compute

$Q_{\mathcal{Y}}(\mathcal{Y}) \leftarrow p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$ — one parameter value.

M Step:

$\boldsymbol{\theta} \leftarrow \underset{\boldsymbol{\theta}}{\text{argmax}} \int d\mathcal{Y}\, Q_{\mathcal{Y}}(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$ — one point value

Variational Bayesian EM (for CE).

Goal: maximise bound on $P(\mathcal{X}|m)$ wrt $Q_{\boldsymbol{\theta}}$

VB-E Step: compute

$Q_{\mathcal{Y}}(\mathcal{Y}) \leftarrow p(\mathcal{Y}|\mathcal{X}, \bar{\boldsymbol{\phi}})$ — Mean of natural parameters

VB-M Step:

$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \overset{\alpha}{\leftarrow} \exp \int d\mathcal{Y}\, Q_{\mathcal{Y}}(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$ — distribution

$\bar{\phi}(\theta) = \langle \phi(\theta) \rangle_{Q(\theta)}$

Properties:

Pushing up $\mathcal{F}_m$ incorporates model complexity penalty. (Because we're working on evidence not likelihood)

- Reduces to the EM algorithm if $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. (Dirac Delta)
- $\mathcal{F}_m$ 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**, $\bar{\boldsymbol{\phi}}$.

↘ Integrating over parameters incorporating a Occam's Razor effect.

VB and model selection

- Variational Bayesian EM yields an approximate posterior Q_θ over model parameters.
- It also yields an optimised lower bound on the model evidence

$$\max \mathcal{F}_{\mathcal{M}}(Q_\mathcal{Y}, Q_\theta) \leq P(\mathcal{D}|\mathcal{M})$$

- These lower bounds can be compared amongst models to learn the right (structure, connectivity ... of the) model

Model Selection

Free energy depends on $Q_\mathcal{Y}$, Q_θ and Hyperparameter η

* Optimise w.r.t. $Q_\mathcal{Y}$ and Q_θ to obtain a lower bound

Just like mean field, as long as we visit (update) each $q(y_i)$ and $q(\theta)$, pushes up lower bound

- If a continuous domain of models is specified by a hyperparameter η, then the VB free energy depends on that parameter:

$$\mathcal{F}(Q_\mathcal{Y}, Q_\theta, \eta) = \iint d\mathcal{Y}\, d\theta\;\; Q_\mathcal{Y}(\mathcal{Y}) Q_\theta(\theta) \log \frac{P(\mathcal{X}, \mathcal{Y}, \theta|\eta)}{Q_\mathcal{Y}(\mathcal{Y}) Q_\theta(\theta)} \leq P(\mathcal{X}|\eta)$$

A hyper-M step maximises the current bound wrt η:

$$\eta \leftarrow \underset{\eta}{\text{argmax}} \iint d\mathcal{Y}\, d\theta\;\; Q_\mathcal{Y}(\mathcal{Y}) Q_\theta(\theta) \log P(\mathcal{X}, \mathcal{Y}, \theta|\eta)$$

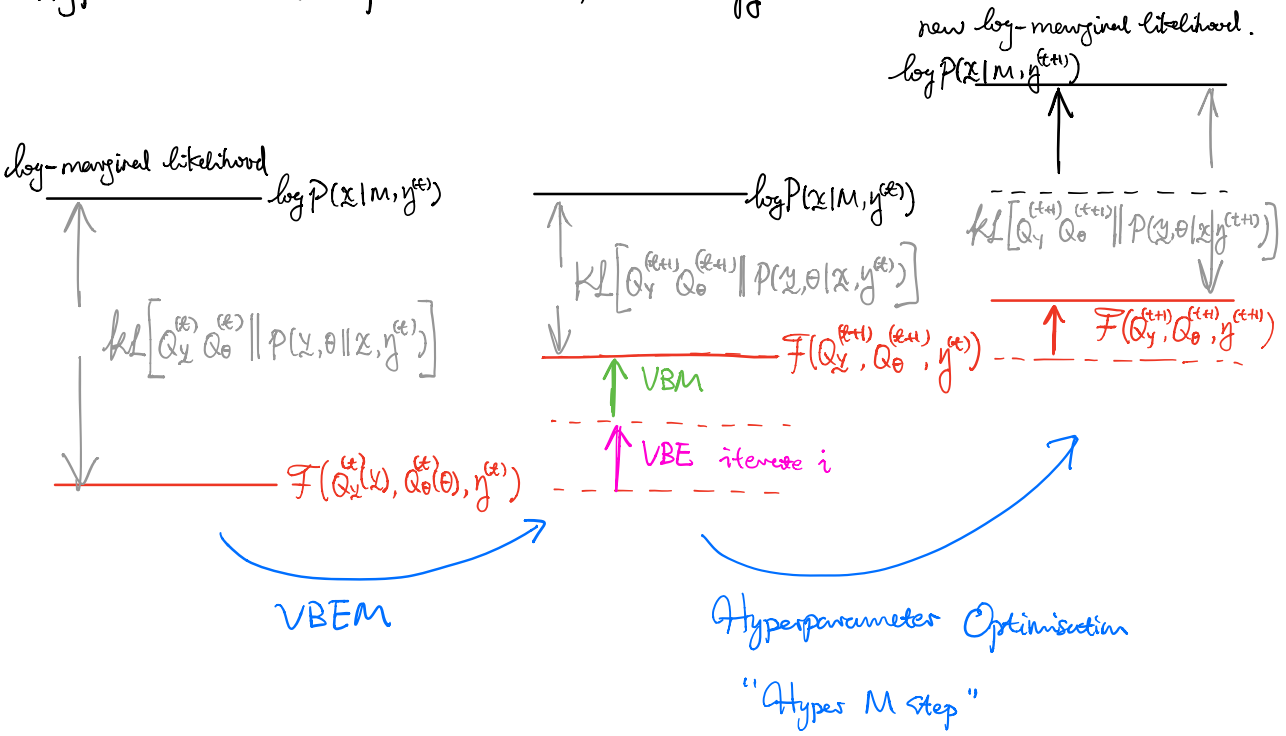
Analogous to standard M step.

Update Hyper parameters to increase Free Energy.

which is approximation to Model Evidence.

Hyper M-step.

Hyper Parameter Optimisation Free Energy



Model Selection:

Compare the log ratio of posterior probabilities for M and M'

$$\log \frac{P(M|\underline{x})}{P(M'|\underline{x})} = \log P(M) + \log P(\underline{x}|M) - \log P(M') - \log P(\underline{x}|M').$$

$$= \log P(M) + \mathcal{F}(Q_{\underline{y}},\theta) + KL\left[Q(\underline{y},\theta) \| P(\underline{y},\theta|\underline{x},M)\right]$$

$$- \log P(M') - \mathcal{F}'(Q_{\underline{y}},\theta) - KL\left[Q'(\underline{y},\theta) \| P(\underline{y},\theta|\underline{x},M')\right]$$

Motivation for ARD:

For a regression problem with too many inputs, some of which are irrelevant to the prediction. So we need a PRIOR (Λ) over the regression parameters embodies the concept of relevance. In this case, $\Lambda_{ii} \sim \mathcal{N}(\underline{0}, \alpha_i^{-1} I)$. We have α_i associated to each input, controls weights from input to output. (variance w.r.t itself, ignore correlation between dimensions)

Under the variational Bayes / Evidence framework, we do a hyperparameter M step, to learn $\underline{\alpha}$

Maximise Evidence w.r.t. α

ARD for unsupervised learning

Automatic Way to Learn dimension

Recall that ARD (automatic relevance determination) was a hyperparameter method to select relevant or useful inputs in regression.

- A similar idea used with variational Bayesian methods can learn a latent dimensionality.
- Consider factor analysis:

  See Ass5, Q1, (h) Select Latent Dimension

  $$\mathbf{x} \sim \mathcal{N}(\Lambda\mathbf{y}, \Psi) \qquad \mathbf{y} \sim \mathcal{N}(0, I)$$ with a column-wise prior $\Lambda_{:i} \sim \mathcal{N}\left(0, \alpha_i^{-1} I\right)$

  (linear)

- The VB free energy is

  Optimise jointly with $\mathcal{Y}$ and Λ

  $$\mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_\Lambda(\Lambda), \Psi, \boldsymbol{\alpha}) = \big\langle \log P(\mathcal{X}, \mathcal{Y}|\Lambda, \Psi) + \log P(\Lambda|\boldsymbol{\alpha}) + \log P(\Psi) \big\rangle_{Q_{\mathcal{Y}} Q_\Lambda} + \ldots$$

  and so hyperparameter optimisation requires

  Prior depending on hyperparameter α

  $$\boldsymbol{\alpha} \leftarrow \text{argmax} \, \langle \log P(\Lambda|\boldsymbol{\alpha}) \rangle_{Q_\Lambda}$$

- Now Q_Λ is Gaussian, with the same form as in linear regression, but with expected moments of $\mathbf{y}$ appearing in place of the inputs.
- Optimisation wrt the distributions, Ψ and $\boldsymbol{\alpha}$ in turn causes some α_i to diverge as in regression ARD.
- In this case, these parameters select "relevant" latent dimensions, effectively learning the dimensionality of $\mathbf{y}$.

Augmented Variational Methods

* Working in a model with observed Gaussian, $p(x|s) = \mathcal{N}(\cdots)$ Then computational complexity of inference scales cubically with the number of data we have (we have to inverse the matrix)

Computational Intractability.

In our examples so far, the approximate variational distribution has been over the "natural" latent variables (and parameters) of the generative model.

* Analytical Intractability with non-conjugate models

Sometimes it may be useful to introduce additional latent variables, solely to achieve computational tractability.

Two examples are GP regression and the GPLVM.

EP will focus on analytic intractability

Make the problem bigger then decompose

Sparse GP approximations

Recall.

GP predictions:

$$y'|X, Y, \mathbf{x}' \sim \mathcal{N}\left(K_{\mathbf{x}'X}(K_{XX}+\sigma^2 I)^{-1}Y, K_{\mathbf{x}'\mathbf{x}'} - K_{\mathbf{x}'X}(K_{XX}^{-1})K_{X\mathbf{x}'} + \sigma^2\right)$$

$O(n^3)$, where problem comes from (Computational Intractability)

Evidence (for learning kernel hyperparameters):

$$\log P(Y|X) = -\frac{1}{2}\log|2\pi(K_{XX}+\sigma^2 I)| - \frac{1}{2}Y(K_{XX}+\sigma^2 I)^{-1}Y^{\mathsf{T}}$$

Computing either form requires inverting the $N \times N$ matrix K_{XX}, in $\mathcal{O}(N^3)$ time.

One proposal to make this more efficient is to find (or select) a smaller set of possibly fictitious measurements U at inputs Z such that

$$P(y'|Z, U, \mathbf{x}') \approx P(y'|X, Y, \mathbf{x}').$$

What values should U and Z take?

Z: Pseudo Input
U: Pseudo Output

Called Sparse Approximation to GP (A reduced matrix for inversion)

X: Actual Input
Z: Sudo Input
Y: Actual Output
U: Sudo Output.
F: Value of Y

Variational Sparse GP approximations

F is still a LATENT variable. We observe Y, but still don't know the exact value of F

Write F for the (smooth) GP function values that underlie Y (so $Y \sim \mathcal{N}(F, \sigma^2 I)$).

Introduce latent measurements U at inputs Z (and integrate over U).

Sudo-Observation Inputs

The likelihood can be written

$$P(Y|X) = \iint dF\, dU\, P(Y, F, U|X, Z) = \iint dF\, dU\, P(Y|F)P(F|U, X, Z)P(U|Z)$$

Now, both U and F are latent, so we introduce a variational distribution $q(F, U)$ to form a free-energy.

Any set of parameters for kernel covariance.

$$\mathcal{F}(q(F, U), \underline{\theta}) = \left\langle \log \frac{P(Y|F)P(F|U, X, Z)P(U|Z)}{q(F, U)} \right\rangle_{q(F,U)}$$

Now, choose the variational form $q(F, U) = P(F|U, X, Z)q(U)$. That is, fix $F|U$ without reference to Y – so information about Y will need to be "compressed" into $q(U)$.

Then

(1) We don't need to consider joint.

$$\mathcal{F}(q(F, U), \theta, Z) = \left\langle \log \frac{P(Y|F)\ \cancel{P(F|U, X, Z)}\ P(U|Z)}{\cancel{P(F|U, X, Z)}\ q(U)} \right\rangle_{P(F|U)q(U)}$$

We have just Gaussian left cut inducing points, not over number of data.

$$= \left\langle \langle \log P(Y|F) \rangle_{P(F|U)} + \log P(U|Z) - \log q(U) \right\rangle_{q(U)}$$

So we can have more data, break data into small sets cut inducing points

Variational Sparse GP approximations

$$\mathcal{F}(q(U),\theta,Z) = \left\langle \langle \log P(Y|F)\rangle_{P(F|U)} + \log P(U|Z) - \log q(U) \right\rangle_{q(U)}$$

Now $P(F|U)$ is fixed by the generative model (rather than being subject to free optimisation).

So we can evaluate that expectation:

$$\begin{aligned}
&\langle \log P(Y|F)\rangle_{P(F|U)} \\
&= \left\langle -\frac{1}{2}\log\left|2\pi\sigma^2 I\right| - \frac{1}{2\sigma^2}\mathrm{Tr}\left[(Y-F)(Y-F)^\mathsf{T}\right] \right\rangle_{P(F|U)} \\
&= -\frac{1}{2}\log\left|2\pi\sigma^2 I\right| - \frac{1}{2\sigma^2}\mathrm{Tr}\left[(Y-\langle F\rangle_{P(F|U)})(Y-\langle F\rangle_{P(F|U)})^\mathsf{T}\right] - \frac{1}{2\sigma^2}\mathrm{Tr}\left[\Sigma_{F|U}\right] \\
&= \log \mathcal{N}\left(Y|K_{XZ}K_{ZZ}^{-1}U,\sigma^2 I\right) - \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX}\right]
\end{aligned}$$

So,

$$\begin{aligned}
\mathcal{F}(q(U),\theta,Z) = \left\langle \log \mathcal{N}\left(Y|K_{XZ}K_{ZZ}^{-1}U,\sigma^2 I\right) + \log P(U|Z) - \log q(U)\right\rangle_{q(U)} \\
- \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX}\right] .
\end{aligned}$$

Variational Sparse GP approximations

* To maximise the free energy, we just set $q(u)$ to the posterior of the PPCA-like model.

A PPCA free energy

$$\mathcal{F}(q(U),\theta,Z) = \left\langle \log \frac{\mathcal{N}\left(Y|K_{XZ}K_{ZZ}^{-1}U,\sigma^2 I\right)P(U|Z)}{q(U)} \right\rangle_{q(U)} - \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX}\right] .$$

difference between PPCA

The expectation is the free energy of a PPCA-like model with normal prior $U \sim \mathcal{N}(0, K_{ZZ})$ and loading matrix $K_{XZ}K_{ZZ}^{-1}$. The maximum of this free energy is the log-likelihood (achieved with q equal to the posterior under the PPCA-like model).

This gives

Gaussian Y with zero mean and $K_{XZ}K_{ZZ}^{-1}K_{ZX} + \sigma^2 I$ covariance

$$\mathcal{F}(q^*(U),\theta,Z) = \log \mathcal{N}\left(Y|0, K_{XZ}K_{ZZ}^{-1}K_{ZX} + \sigma^2 I\right) - \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX} - K_{XZ}K_{ZZ}^{-1}K_{ZX}\right] .$$

Zero mean

Note that we have eliminated all terms in K_{XX}^{-1}.

The only inverse, comes from variational.

We can optimise the free energy numerically with respect to Z and θ to adjust the GP prior and quality of variational approximation.

A similar approach can be used to learn X if they are unobserved (*i.e.* in the GPLVM). Assume $q(X,F,U) = q(X)P(F|X,U)q(U)$. Then $\mathcal{F} = \langle \log P(Y,F,U|X) \log P(X)\rangle_{q(U)q(X)}$ which simplifies into tractable components in much the same way as above.

A few references

- Jordan, Ghahramani, Jaakkola, Saul, 1999. An introduction to variational methods for graphical models. *Machine Learning* **37**:183–233.
- Attias, 2000. A variational Bayesian framework for graphical models. *NIPS 12*. `http://www.gatsby.ucl.ac.uk/publications/papers/03-2000.ps`
- Beal, 2003. Variational algorithms for approximate Bayesian inference. *PhD thesis*, Gatsby Unit, UCL. `http://www.cse.buffalo.edu/faculty/mbeal/thesis/`
- Winn, 2003. Variational message passing and its applications. *PhD thesis*, Cambridge. `http://johnwinn.org/Publications/Thesis.html`; also `VIBES` software for conjugate-exponential graphs.

Some complexities:

- MacKay, 2001. A problem with variational free energy minimization. `http://www.inference.phy.cam.ac.uk/mackay/minima.pdf`
- Turner, MS, 2011. Two problems with variational expectation maximisation for time-series models. In Barber, Cemgil, Chiappa, eds., *Bayesian Time Series Models*. `http://www.gatsby.ucl.ac.uk/~maneesh/papers/turner-sahani-2010-ildn.pdf`
- Berkes, Turner, MS, 2008. On sparsity and overcompleteness in image models. *NIPS 20*. `http://www.gatsby.ucl.ac.uk/~maneesh/papers/berkes-etal-2008-nips.pdf`
- Giordano, R, Broderick, T, and Jordan, MI, 2015. Linear response methods for accurate covariance estimates from mean field variational Bayes. *NIPS*