Probabilistic and Bayesian Machine Learning

Day 4: Variational Approximations

Yee Whye Teh
ywteh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit
University College London

http://www.gatsby.ucl.ac.uk/∼ywteh/teaching/probmodels

The E and M steps of EM

The lower bound on the log likelihood is given by:

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

EM alternates between:

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

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

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

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

Variational Approximations to the EM algorithm

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

Generalised EM algorithm replaces intractable maximisations with gradient M-steps. For the E-step we could:

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

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

VE step: Find $q^{(k)}$ within restricted class $\mathcal{Q}$ with

$$\mathcal{F}(q^{(k)}(\mathbf{Y}), \theta^{(k-1)}) \geq \mathcal{F}(q^{(k-1)}(\mathbf{Y}), \theta^{(k-1)})$$

M step: Find $\theta^{(k)}$ with

$$\mathcal{F}(q^{(k)}(\mathbf{Y}), \theta^{(k)}) \geq \mathcal{F}(q^{(k)}(\mathbf{Y}), \theta^{(k-1)})$$

This increases a lower bound on the log likelihood (but not necessarily the log likelihood itself...).

KL divergence

Recall that

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

Thus,

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

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

is equivalent to:

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

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

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

This is related to ideas in information geometry.

Factored Variational E-step

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

$$\mathcal{Q} = \left\{ q \;\middle|\; q(\mathbf{Y}) = \prod_i q_i(\mathbf{Y}_i) \right\}.$$

In this case the E-step is itself iterative:

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

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

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:

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

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

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

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

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

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

Suppose $P(\mathbf{X}, \mathbf{Y})$ is an exponential family distribution, *e.g.* the Boltzmann machine:

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

with some $\mathbf{Y} = \{s_i\}$ unobserved while others are observed.

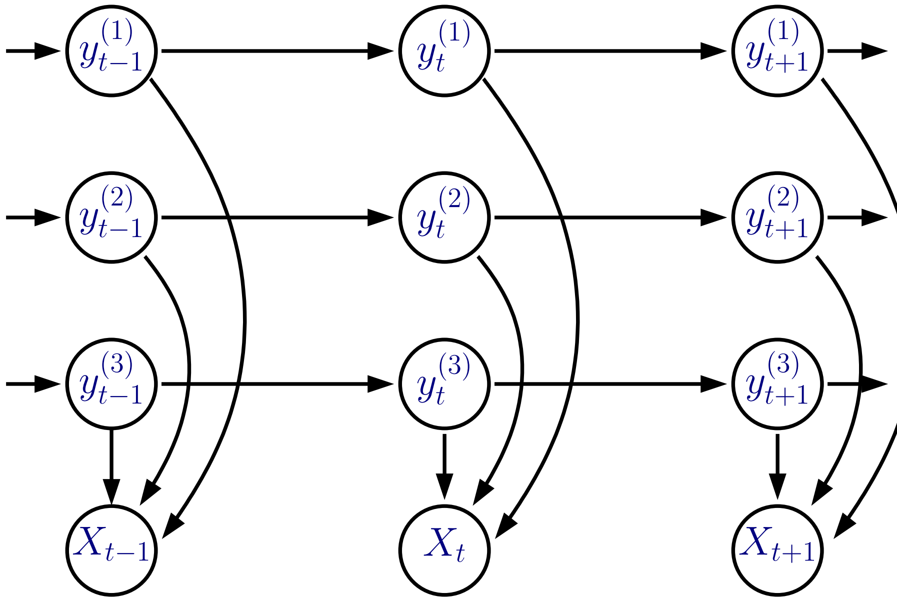
Expectations wrt a fully factored q distribute over all $s_i \in \mathbf{Y}$

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

(where q_i for $s_i \in \mathbf{X}$ 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.

Factorial HMMs



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

$$q(y_{1:\tau}^{1:M}) = \prod_m q^m(y_{1:\tau}^m)$$

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

$$\begin{aligned}
q^{m'}(y_{1:\tau}^{m'}) &\propto \exp \left\langle \log P(y_{1:\tau}^{1:M}, x_{1:\tau}) \right\rangle_{\prod_{\neg m'} q^m(y_{1:\tau}^m)} \\
&= \exp \left\langle \sum_m \sum_t \log P(y_t^m | y_{t-1}^m) + \sum_t \log P(x_t | y_t^{1:M}) \right\rangle_{\prod_{\neg m'} q^m(y_{1:\tau}^m)} \\
&\propto \exp \left[\sum_t \log P(y_t^{m'} | y_{t-1}^{m'}) + \sum_t \left\langle \log P(x_{t'} | y_{t'}^{1:M}) \right\rangle_{\prod_{\neg m} q^m(y_{1:\tau}^m)} \right] \\
&= \prod_t P(y_t^{m'} | y_{t-1}^{m'}) \quad \prod_t \exp \left\langle \log P(x_{t'} | y_{t'}^{1:M}) \right\rangle_{\prod_{\neg m} q^m(y_{t'}^m)}
\end{aligned}$$

Variational Bayesian Learning

Let the hidden latent variables be $\mathbf{Y}$, data $\mathbf{X}$ and the parameters $\boldsymbol{\theta}$.

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

$$\begin{aligned}
\log P(\mathbf{X}) &= \log \int d\mathbf{Y}\, d\boldsymbol{\theta}\; P(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta}) \\
&= \log \int d\mathbf{Y}\, d\boldsymbol{\theta}\; Q(\mathbf{Y}, \boldsymbol{\theta}) \frac{P(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta})}{Q(\mathbf{Y}, \boldsymbol{\theta})} \\
&\geq \int d\mathbf{Y}\, d\boldsymbol{\theta}\; Q(\mathbf{Y}, \boldsymbol{\theta}) \log \frac{P(\mathbf{X}, \mathbf{Y}, \boldsymbol{\theta})}{Q(\mathbf{Y}, \boldsymbol{\theta})}.
\end{aligned}$$

The saturating $Q(\mathbf{Y}, \boldsymbol{\theta}) = P(\mathbf{Y}, \boldsymbol{\theta}|\mathbf{X})$ is almost always intractable.
Use a simpler, factorised approximation $Q(\mathbf{Y}, \boldsymbol{\theta}) = Q_{\mathbf{Y}}(\mathbf{Y}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

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

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

Variational Bayesian Learning

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

$$Q^{(k)}_{\mathbf{Y}}(\mathbf{Y}) \propto \exp \langle \log P(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta}) \rangle_{Q^{(k-1)}_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \qquad E-like\ step$$

$$Q^{(k)}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}) \exp \langle \log P(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta}) \rangle_{Q^{(k)}_{\mathbf{Y}}(\mathbf{Y})} \qquad M-like\ step$$

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

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

Conjugate-Exponential Families

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

- The joint probability over variables is in the exponential family:

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

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

- The prior over parameters is conjugate to this joint probability:

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

  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

Variational Bayes for Conjugate-Exponential Families

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

(a) $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is also **conjugate**:

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\bar{\eta}, \bar{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\bar{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \bar{\boldsymbol{\nu}}\right\}$$

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

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

$$Q_{\mathbf{Y}_i}(\mathbf{Y}_i) \propto f(\mathbf{Y}_i, \mathbf{X}_i) \exp\left\{\boldsymbol{\phi}(\bar{\boldsymbol{\theta}})^\top \mathbf{T}(\mathbf{Y}_i, \mathbf{X}_i)\right\} = P(\mathbf{Y}_i | \mathbf{X}_i, \bar{\boldsymbol{\theta}})$$

where $\phi(\bar{\boldsymbol{\theta}}) = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$.

Key points:

- The approximate parameter posterior is of the same form as the prior, so it is **easily summarized** in terms of two sets of hyperparameters, $\bar{\eta}$ and $\bar{\boldsymbol{\nu}}$;
- The approximate latent 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

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

E Step: compute

$$Q_{\mathbf{Y}}^{(k)}(\mathbf{Y}) = P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}^{(k-1)})$$

M Step:

$$\boldsymbol{\theta}^{(k)} = \operatorname*{argmax}_{\boldsymbol{\theta}} \left\langle \log P(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta}) \right\rangle_{Q_{\mathbf{Y}}^{(k)}(\mathbf{Y})}$$

Variational Bayesian EM

Goal: lower bound $p(\mathbf{X})$

VB-E Step: compute

$$Q_{\mathbf{Y}}^{(k)}(\mathbf{Y}) = P(\mathbf{Y}|\mathbf{X}, \bar{\boldsymbol{\theta}}^{(k-1)})$$

VB-M Step:

$$Q_{\boldsymbol{\theta}}^{(k)}(\boldsymbol{\theta}) \propto \exp \left\langle \log P(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \right\rangle_{Q_{\mathbf{Y}}^{(k)}(\mathbf{Y})}$$

Properties:

- Reduces to the EM algorithm if $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.
- Free energy increases monotonically.
- Analytical parameter distributions.
- 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, but **using expected natural parameters**, $\bar{\boldsymbol{\theta}}$.

End Notes

Theses on Variational Bayes:

Matthew Beal (2003).
Variational Algorithms for Approximate Bayesian Inference. Gatsby Unit, UCL.

John Winn (2003).
Variational Message Passing and its Applications. Physics, Cambridge.

Alternative view point:

M. J. Wainwright and M. I. Jordan (2008).
Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1:1-305.

End Notes