Variational Inference and Learning I

Fundamentals, mean-field VI, and the EM algorithm

Michael U. Gutmann

Probabilistic Machine Learning (INFR11298)
School of Informatics, The University of Edinburgh

Autumn Semester 2026

Recap

- Learning and inference often involve integrals that are hard to compute.
- For example:
  - Marginalisation/inference: $p(\mathbf{x}) = \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$
  - Log-likelihood for iid data $\mathbf{v}_1, \ldots, \mathbf{v}_n$ in case of unobserved variables: $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int_{\mathbf{h}} p(\mathbf{h}, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}$
  - The gradient of $\ell(\boldsymbol{\theta})$ requires expectations with respect to the posteriors $p(\mathbf{h}|\mathbf{v}_i; \boldsymbol{\theta})$, which are generally intractable too.
- Here: variational approach to inference and learning to deal with some of the intractabilities that we encountered.

History

Variational methods have a long history, in particular in physics. For example:

- Fermat's principle (1650) to explain the path of light: "light travels between two given points along the path of shortest time" (see e.g. `http://www.feynmanlectures.caltech.edu/I_26.html`)
- Principle of least action in classical mechanics and beyond (see e.g. `http://www.feynmanlectures.caltech.edu/II_19.html`)
- Finite elements methods to solve problems in fluid dynamics or civil engineering.

Loosely speaking: the general idea is to frame the original problem in terms of an optimisation problem.

Program

1. Preparations
2. The variational principle
3. Application to inference
4. Application to learning

Program

1. Preparations
   - Concavity of the logarithm and Jensen's inequality
   - Kullback-Leibler divergence and its properties
2. The variational principle
3. Application to inference
4. Application to learning

$\log(u)$ is a concave function

- $\log(u)$ is a concave function

  $$\log((1-a)u_1 + au_2) \geq (1-a)\log(u_1) + a\log(u_2) \qquad a \in [0,1]$$

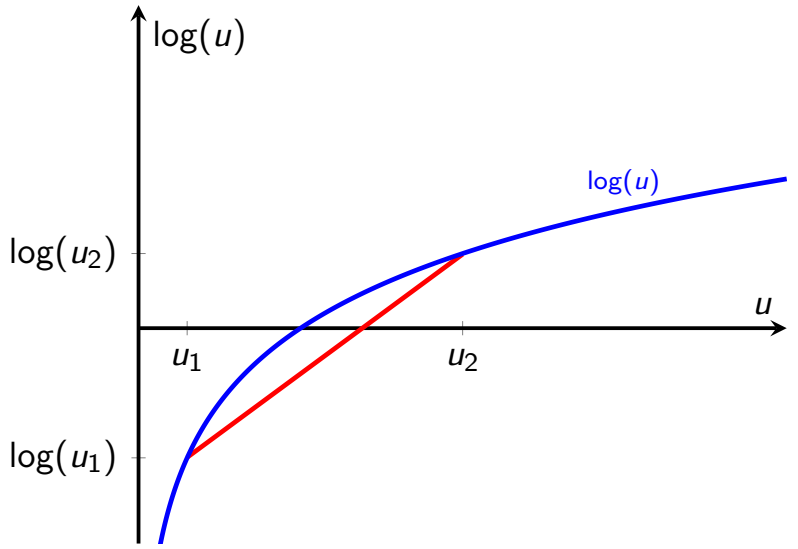
  $(1-a)x + ay$ with $a \in [0,1]$ is a weighted average of x and y.

- log(average) $\geq$ average (log)

- Generalisation

  $$\log \mathbb{E}[g(\mathbf{x})] \geq \mathbb{E}[\log g(\mathbf{x})]$$

  with $g(\mathbf{x}) > 0$



- Called Jensen's inequality for concave functions.

Kullback-Leibler divergence

- Kullback Leibler divergence $\mathrm{KL}(p||q)$

$$\mathrm{KL}(p||q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} = \mathbb{E}_{p(\mathbf{x})} \left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})} \right] \quad (1)$$

- Properties
  - $\mathrm{KL}(p||q) = 0$ if and only if (iff) $p = q$
    (they may be different on sets of probability zero under p)
  - $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$
  - $\mathrm{KL}(p||q) \geq 0$
- Non-negativity follows from the concavity of the logarithm.

Non-negativity of the KL divergence

Non-negativity follows from the concavity of the logarithm.

$$-\mathrm{KL}(p||q) = -\mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \tag{2}$$

$$= \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})}\right] \tag{3}$$

$$\leq \log \underbrace{\mathbb{E}_{p(\mathbf{x})}\left[\frac{q(\mathbf{x})}{p(\mathbf{x})}\right]}_{\int p(\mathbf{x})\frac{q(\mathbf{x})}{p(\mathbf{x})}d\mathbf{x}=1} \tag{4}$$

Hence $-\mathrm{KL}(p||q) \leq \log(1) = 0$ and thus

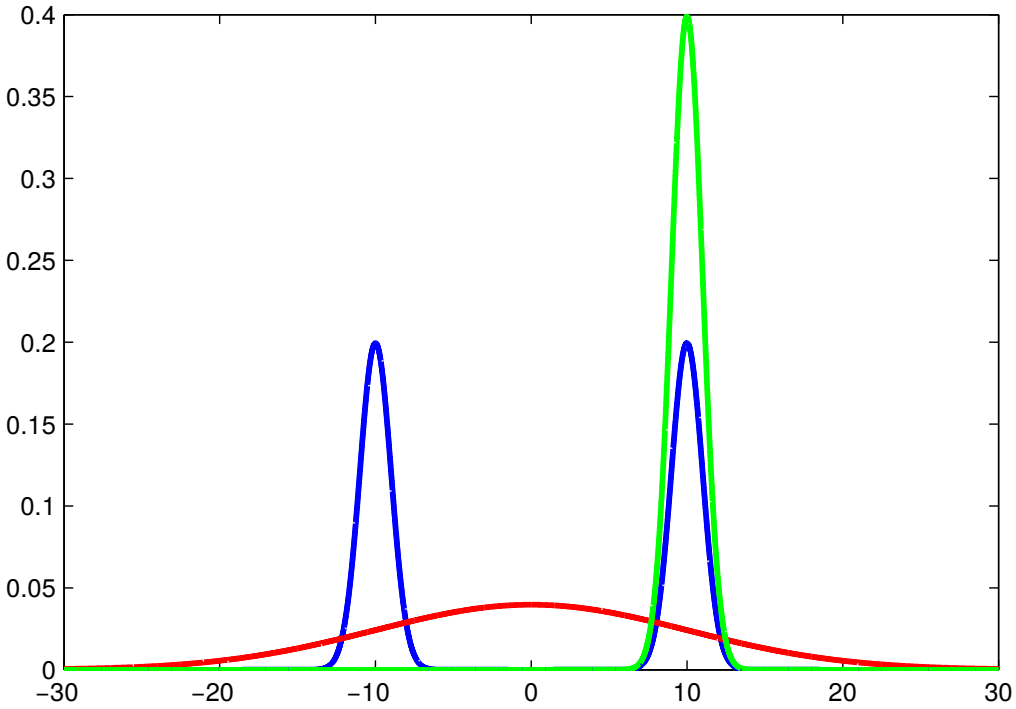
$$\mathrm{KL}(p||q) \geq 0 \tag{5}$$

Asymmetry of the KL divergence

Blue: mixture of Gaussians $p(x)$ (fixed)

Green: (unimodal) Gaussian q that minimises $\mathrm{KL}(q||p)$

Red: (unimodal) Gaussian q that minimises $\mathrm{KL}(p||q)$

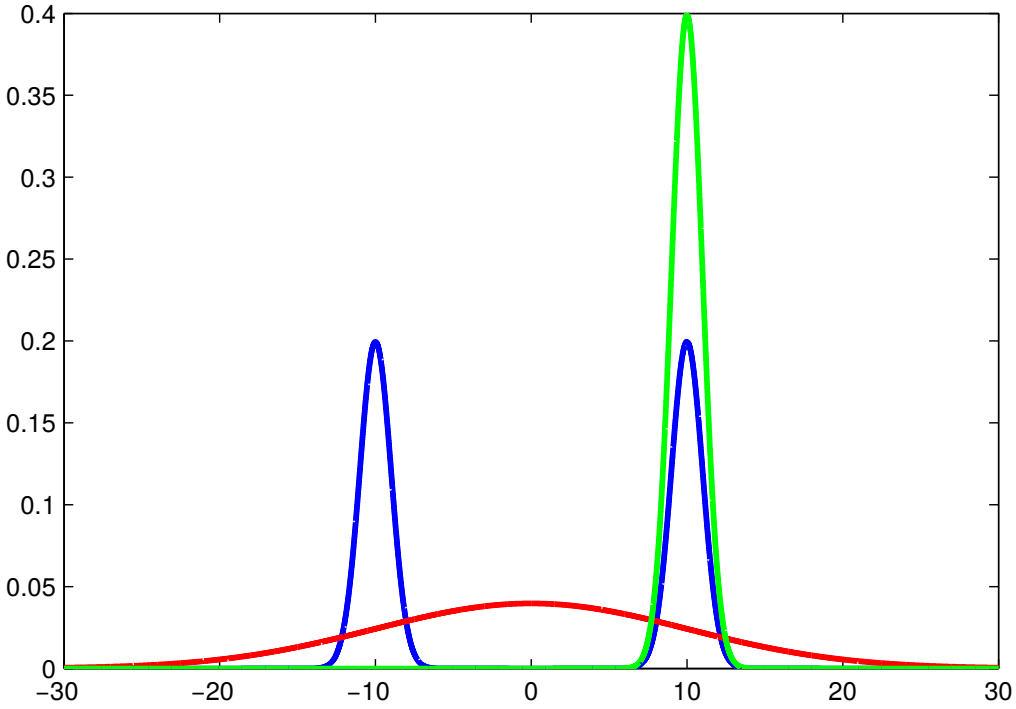


Barber Figure 28.1, Section 28.3.4

Asymmetry of the KL divergence

$$\operatorname{argmin}_q \mathrm{KL}(q||p) = \operatorname{argmin}_q \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x}$$

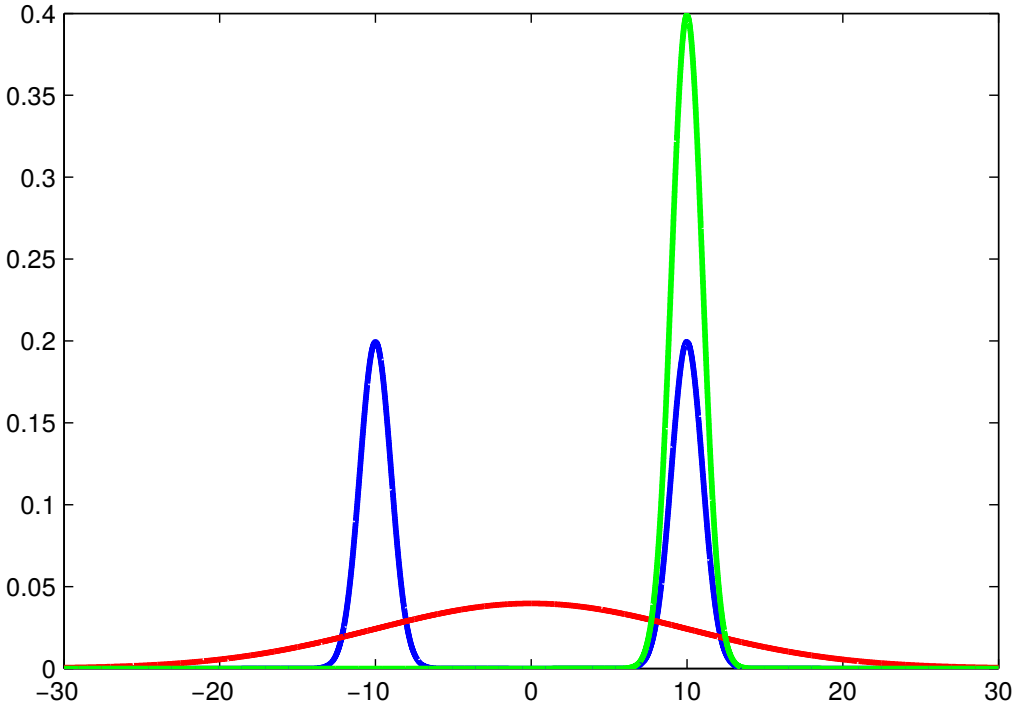
- Large penalty when $q(\mathbf{x})$ is large but $p(\mathbf{x})$ is small.
- No penalty for $q(\mathbf{x}) \neq p(\mathbf{x})$ when $q(\mathbf{x})$ is small.
- Encourages q to stay inside the high-density regions of p, even if it misses some. Produces good local fit, "mode seeking".



Asymmetry of the KL divergence

$$\operatorname{argmin}_q \mathrm{KL}(p||q) = \operatorname{argmin}_q \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x}$$

- Large penalty when $q(\mathbf{x})$ is small but $p(\mathbf{x})$ is large.
- No penalty for $q(\mathbf{x}) \neq p(\mathbf{x})$ when $p(\mathbf{x})$ is small.
- Encourages q to cover all high-density regions of p, even if it spreads beyond them. Produces global fit, corresponds to MLE.

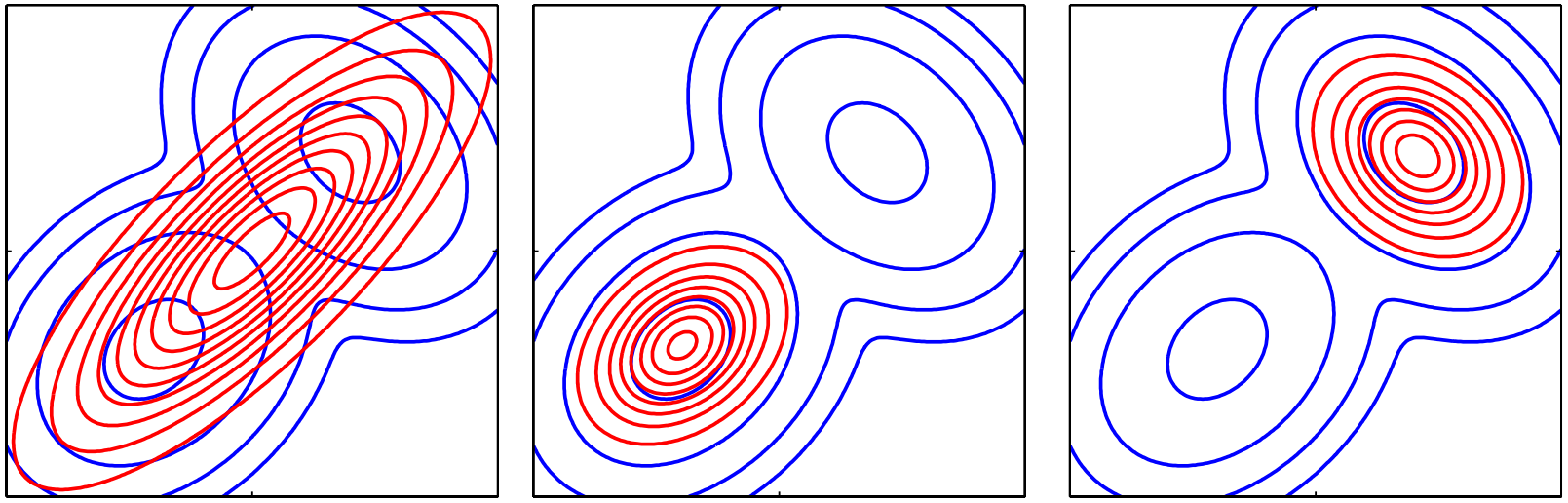


Asymmetry of the KL divergence

Blue: mixture of Gaussians $p(\mathbf{x})$ (fixed)

Red: optimal (unimodal) Gaussians $q(\mathbf{x})$

Global fit (left) versus mode seeking (middle and right). (two local minima are shown)



$\min_q$ KL(p || q) $\min_q$ KL(q || p) $\min_q$ KL(q || p)

Bishop Figure 10.3

Program

1. Preparations
   - Concavity of the logarithm and Jensen's inequality
   - Kullback-Leibler divergence and its properties
2. The variational principle
3. Application to inference
4. Application to learning

Program

1. Preparations

2. The variational principle
   - Variational lower bound
   - Maximising the ELBO to compute the marginal and conditional from the joint

3. Application to inference

4. Application to learning

Variational lower bound: auxiliary distribution

Consider joint pdf /pmf $p(\mathbf{x}, \mathbf{y})$ with marginal $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$

- We can write $p(\mathbf{x})$ as

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \frac{q(\mathbf{y}|\mathbf{x})}{q(\mathbf{y}|\mathbf{x})} \mathrm{d}\mathbf{y} = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \left[\frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})} \right] \tag{6}$$

  where $q(\mathbf{y}|\mathbf{x})$ is an auxiliary distribution (called the variational distribution in the context of variational inference/learning) for a given $\mathbf{x}$.

- Log marginal is

$$\log p(\mathbf{x}) = \log \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \left[\frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})} \right] \tag{7}$$

- Approximating the expectation with a sample average leads to importance sampling. Another approach is to work with the concavity of the logarithm instead.

Variational lower bound: concavity of the logarithm

- Concavity of the log gives

$$\log p(\mathbf{x}) = \log \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \geq \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \tag{8}$$

  This is the variational lower bound for $\log p(\mathbf{x})$.

- Right-hand side is called the (variational) free energy $\mathcal{F}_{\mathbf{x}}(q)$ or the evidence lower bound (ELBO) $\mathcal{L}_{\mathbf{x}}(q)$

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \tag{9}$$

- Since q is a function, the ELBO is a functional, which is a mapping that depends on a function.

Properties of the ELBO

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right]$$

- By manipulating the definition of the ELBO, we obtain the following equivalent forms

$$\begin{aligned}
\mathcal{L}_{\mathbf{x}}(q) &= \log p(\mathbf{x}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) && (10)\\
&= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{y}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y})) && (11)\\
&= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x},\mathbf{y}) + \mathcal{H}(q) && (12)
\end{aligned}$$

  where $p(\mathbf{y})$ is the marginal of $p(\mathbf{x},\mathbf{y})$ and $\mathcal{H}(q)$ is the entropy of q.

- Entropy is a measure of randomness/variability of a variable

$$\mathcal{H}(q) = -\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{x})\right] \quad (13)$$

  Larger entropy means more variability.

Properties of the ELBO (proof)

- First expression:

$$\begin{aligned}
\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] &= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{q(\mathbf{y}|\mathbf{x})}\right] \\
&= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{y}|\mathbf{x})}{q(\mathbf{y}|\mathbf{x})} + \log p(\mathbf{x})\right] \\
&= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{y}|\mathbf{x})}{q(\mathbf{y}|\mathbf{x})}\right] + \log p(\mathbf{x}) \\
&= -\mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) + \log p(\mathbf{x})
\end{aligned}$$

- Second expression is obtained similarly but using $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$ instead of $p(\mathbf{x},\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ above.
- Third expression from the definition of the entropy.

Tightness of the ELBO

- From $\mathcal{L}_{\mathbf{x}}(q) = \log p(\mathbf{x}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x}))$ and non-negativity of the KL divergence, we have
  1. $\log p(\mathbf{x}) \geq \mathcal{L}_{\mathbf{x}}(q)$ (as before)
  2. $\log p(\mathbf{x}) = \mathcal{L}_{\mathbf{x}}(q) \Leftrightarrow q(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$
- Maximising $\mathcal{L}_{\mathbf{x}}(q)$ with respect to q yields both $\log p(\mathbf{x})$ and the conditional $p(\mathbf{y}|\mathbf{x})$ at the same time.
- Makes sense: if we know $p(\mathbf{x}, \mathbf{y})$ and $p(\mathbf{x})$, we know $p(\mathbf{y}|\mathbf{x})$, and vice versa, since $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{x})$.

Alternative approach

- We started from the task of approximating the marginal

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y} \tag{14}$$

- Alternative starting point is the task of approximating the conditional $p(\mathbf{y}|\mathbf{x})$ for some given $\mathbf{x}$ by a distribution $q(\mathbf{y}|\mathbf{x})$.
- Measuring the quality of the approximation $q(\mathbf{y}|\mathbf{x})$ by $\mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x}))$ gives

$$\mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) = \log p(\mathbf{x}) - \mathcal{L}_{\mathbf{x}}(q) \tag{15}$$

Same key result as before.

Variational principle

- By maximising the ELBO

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right]$$

  we can split the joint $p(\mathbf{x},\mathbf{y})$ into $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$

$$\log p(\mathbf{x}) = \max_q \mathcal{L}_{\mathbf{x}}(q)$$

$$p(\mathbf{y}|\mathbf{x}) = \operatorname*{argmax}_q \mathcal{L}_{\mathbf{x}}(q)$$

- Highlights the variational principle: The inference problem is expressed in terms of an optimisation problem.

Solving the optimisation problem

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right]$$

- Difficulties when maximising the ELBO over all q:
  - Optimisation is over functions (pdfs/pmfs), not a finite set of parameters
  - $\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}$ depends on q, so the gradient cannot be moved inside it
- Restrict search space to a family $\mathcal{Q}$ of tractable distributions $q(\mathbf{y}|\mathbf{x})$.
- Family $\mathcal{Q}$ specified by
  - independence assumptions, e.g. $q(\mathbf{y}|\mathbf{x}) = \prod_i q(y_i|\mathbf{x})$, which corresponds to "mean-field" variational inference
  - parametric assumptions, e.g. $q(y_i|\mathbf{x}) = \mathcal{N}(y_i; \mu_i(\mathbf{x}), \sigma_i^2(\mathbf{x}))$
- Price: unless $p(\mathbf{y}|\mathbf{x}) \in \mathcal{Q}$, $\max_{q\in\mathcal{Q}} \mathcal{L}_{\mathbf{x}}(q)$ is only a lower bound on $\log p(\mathbf{x})$, and the optimal q only approximates $p(\mathbf{y}|\mathbf{x})$.
- $\mathcal{L}_{\mathbf{x}}(q)$ has a closed form only in special cases; otherwise it can be estimated, e.g. by sampling from q. (details later)

Program

1. Preparations

2. The variational principle
   - Variational lower bound
   - Maximising the ELBO to compute the marginal and conditional from the joint

3. Application to inference

4. Application to learning

Program

1. Preparations
2. The variational principle
3. Application to inference
   - The mechanics
   - Interpretation
   - Nature of the approximation
   - Mean-field variational inference
4. Application to learning

Approximate posterior inference

- Inference task: given value $\mathbf{x} = \mathbf{x}_o$ and joint pdf/pmf $p(\mathbf{x}, \mathbf{y})$, compute $p(\mathbf{y}|\mathbf{x}_o)$.
- Variational approach: estimate the posterior by solving an optimisation problem

$$\hat{p}(\mathbf{y}|\mathbf{x}_o) = \operatorname*{argmax}_{q \in \mathcal{Q}} \mathcal{L}_{\mathbf{x}_o}(q) \tag{16}$$

  $\mathcal{Q}$ is the set of pdfs/pmfs in which we search for the solution
- From the basic property of the ELBO in Equation (10)

$$\log p(\mathbf{x}_o) = \mathrm{KL}(q(\mathbf{y}|\mathbf{x}_o)||p(\mathbf{y}|\mathbf{x}_o)) + \mathcal{L}_{\mathbf{x}_o}(q) = \text{const} \tag{17}$$

- Because the sum of the KL and ELBO is constant, we have

$$\operatorname*{argmax}_{q \in \mathcal{Q}} \mathcal{L}_{\mathbf{x}_o}(q) = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(q(\mathbf{y}|\mathbf{x}_o)||p(\mathbf{y}|\mathbf{x}_o)) \tag{18}$$

Posterior as compromise between prior and fit

- Equivalent forms of the ELBO:

$$\mathcal{L}_{\mathbf{x}_o}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x}_o)} \log p(\mathbf{x}_o|\mathbf{y}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x}_o)||p(\mathbf{y})) \tag{19}$$

- By maximising $\mathcal{L}_{\mathbf{x}_o}(q)$ we find a q that
  - produces $\mathbf{y}$ which are likely explanations of $\mathbf{x}_o$
  - stays close to the prior $p(\mathbf{y})$
- If included in the search space $\mathcal{Q}$, $p(\mathbf{y}|\mathbf{x}_o)$ is the optimal q, which means that the posterior fulfils the two desiderata best.

As compromise between variable and likely imputations

- Equivalent forms of the ELBO:

$$\mathcal{L}_{\mathbf{x}_o}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x}_o)} \log p(\mathbf{x}_o, \mathbf{y}) + \mathcal{H}(q) \tag{20}$$

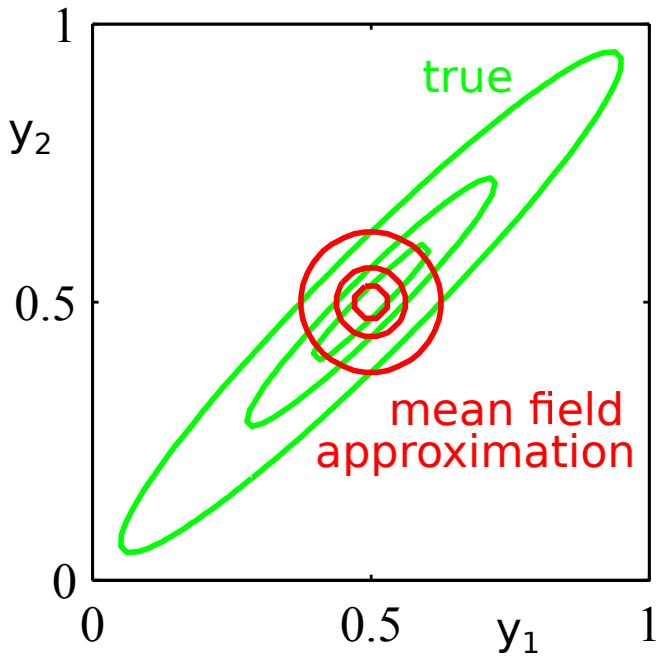
- By maximising $\mathcal{L}_{\mathbf{x}_o}(q)$ we find a q that
  - produces likely imputations (filled-in data) $\mathbf{y}$
  - is maximally variable
- If included in the search space $\mathcal{Q}$, $p(\mathbf{y}|\mathbf{x}_o)$ is the optimal q, which means that the posterior fulfils the two desiderata best.

Nature of the approximation

$$\text{argmax}_{q \in \mathcal{Q}} \mathcal{L}_{\mathbf{x}_o}(q) = \text{argmin}_{q \in \mathcal{Q}} \text{KL}(q(\mathbf{y}|\mathbf{x}_o)||p(\mathbf{y}|\mathbf{x}_o))$$

- When minimising $\text{KL}(q||p)$ with respect to q, q will try very hard to be zero where p is small.
- Assume true posterior is correlated bivariate Gaussian and we work with $\mathcal{Q} = \{q(\mathbf{y}|\mathbf{x}_o) : q(\mathbf{y}|\mathbf{x}_o) = q(y_1|\mathbf{x}_o)q(y_2|\mathbf{x}_o)\}$
  (independence but no parametric assumptions, called mean-field VI)

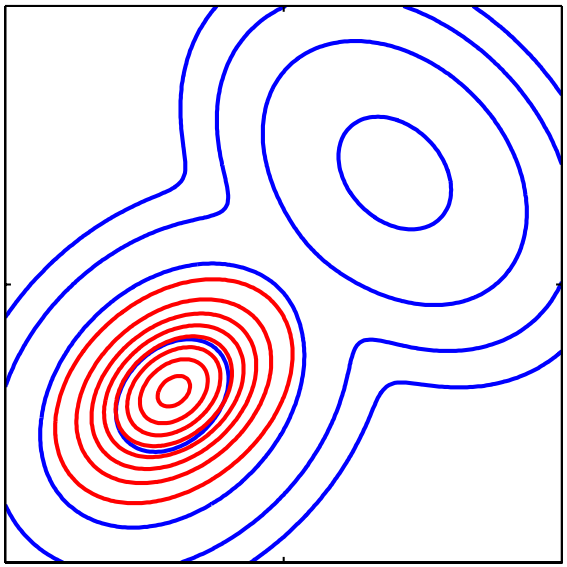
- Optimal q is Gaussian.
- Mean is correct but variances dictated by the variances of $p(\mathbf{y}|\mathbf{x}_o)$ along the y_1 and y_2 axes.
- Posterior variance is underestimated.



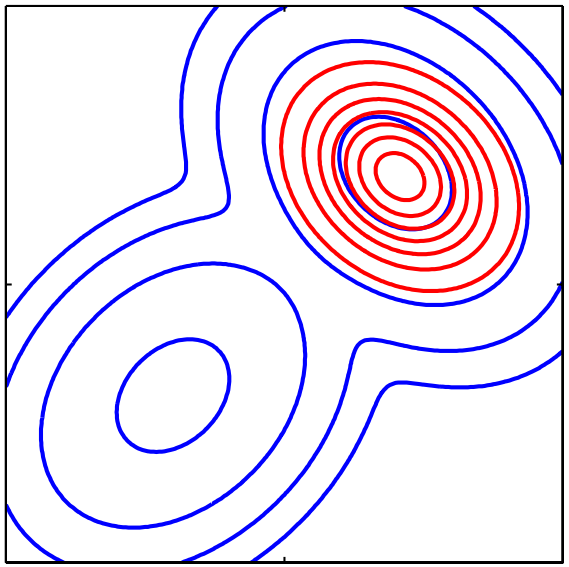
(Bishop, Figure 10.2)

Nature of the approximation

- Assume that true posterior is multimodal, but that the family of variational distributions $\mathcal{Q}$ only includes unimodal distributions.
- The optimal $q(\mathbf{y}|\mathbf{x}_o)$ only covers one mode: "mode-seeking behaviour".



local optimum



local optimum

Blue: true posterior
Red: approximation

Bishop Figure 10.3 (adapted)

Mean-field variational inference

- In mean field variational inference, we assume that the variational distribution $q(\mathbf{y}|\mathbf{x})$ fully factorises, i.e.

$$q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{d} q_i(y_i|\mathbf{x}) \tag{21}$$

  when $\mathbf{y}$ is d-dimensional.
- Independence assumption but no parametric assumption
- An approach to learning the q_i is to update one at a time while keeping the others fixed, called coordinate ascent variational inference (CAVI).
- We next derive the corresponding update equations.

Mean-field VI: ELBO

With $q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{d} q_i(y_i|\mathbf{x})$, we have

$$\begin{aligned}\mathcal{L}_{\mathbf{x}}(q) =& \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log q(\mathbf{y}|\mathbf{x}) \\ =& \mathbb{E}_{q_1(y_1|\mathbf{x})} \cdots \mathbb{E}_{q_d(y_d|\mathbf{x})} \left[\log p(\mathbf{x}, \mathbf{y})\right] - \\ & \mathbb{E}_{q_1(y_1|\mathbf{x})} \cdots \mathbb{E}_{q_d(y_d|\mathbf{x})} \left[\sum_{i=1}^{d} \log q_i(y_i|\mathbf{x})\right]\end{aligned}$$

Second term simplifies:

$$\begin{aligned}\mathbb{E}_{q_1} \cdots \mathbb{E}_{q_d} \left[\sum_{i=1}^{d} \log q_i(y_i|\mathbf{x})\right] &= \sum_{i=1}^{d} \mathbb{E}_{q_1(y_1|\mathbf{x})} \cdots \mathbb{E}_{q_d(y_d|\mathbf{x})} \left[\log q_i(y_i|\mathbf{x})\right] \\ &= \sum_{i=1}^{d} \mathbb{E}_{q_i(y_i|\mathbf{x})} \left[\log q_i(y_i|\mathbf{x})\right]\end{aligned}$$

Mean-field VI: ELBO

Define $q(\mathbf{y}_{\setminus 1}|\mathbf{x}) = \prod_{i=2}^{d} q_i(y_i|\mathbf{x})$ so that

$$q(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{d} q_i(y_i|\mathbf{x}) = q_1(y_1|\mathbf{x}) q(\mathbf{y}_{\setminus 1}|\mathbf{x}) \tag{22}$$

Hence

$$\mathbb{E}_{q_1(y_1|\mathbf{x})} \cdots \mathbb{E}_{q_d(y_d|\mathbf{x})} \left[\log p(\mathbf{x}, \mathbf{y})\right] = \mathbb{E}_{q_1(y_1|\mathbf{x})} \mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})} \left[\log p(\mathbf{x}, \mathbf{y})\right]$$

ELBO becomes

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q_1(y_1|\mathbf{x})} \mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})} \left[\log p(\mathbf{x}, \mathbf{y})\right] - \sum_{i=1}^{d} \mathbb{E}_{q_i(y_i|\mathbf{x})} \left[\log q_i(y_i|\mathbf{x})\right] \tag{23}$$

We next maximise $\mathcal{L}_{\mathbf{x}}(q)$ with respect to q_1 while keeping the other q_i fixed.

Mean-field VI: Optimisation

As we optimise with respect to q_1 we can drop additive terms from the ELBO that do not depend on q_1. This gives the objective

$$J(q_1) = \mathbb{E}_{q_1(y_1|\mathbf{x})}\mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})}\left[\log p(\mathbf{x}, \mathbf{y})\right] - \mathbb{E}_{q_1(y_1|\mathbf{x})}\left[\log q_1(y_1|\mathbf{x})\right] \quad (24)$$

Define $\bar{p}(y_1|\mathbf{x})$

$$\bar{p}(y_1|\mathbf{x}) = \frac{1}{Z}\exp\left[\mathbb{E}_{q(\mathbf{y}_{\setminus 1}|\mathbf{x})}\left[\log p(\mathbf{x}, \mathbf{y})\right]\right] \quad (25)$$

Then

$$J(q_1) = \mathbb{E}_{q_1(y_1|\mathbf{x})}\left[\log \bar{p}(y_1|\mathbf{x}) - \log Z\right] - \mathbb{E}_{q_1(y_1|\mathbf{x})}\left[\log q_1(y_1|\mathbf{x})\right] \quad (26)$$

$$= -\mathbb{E}_{q_1(y_1|\mathbf{x})}\left[\log \frac{q_1(y_1|\mathbf{x})}{\bar{p}(y_1|\mathbf{x})}\right] - \log Z \quad (27)$$

$$= -\mathrm{KL}(q_1(y_1|\mathbf{x})||\bar{p}(y_1|\mathbf{x})) - \log Z \quad (28)$$

Thus

$$\operatorname*{argmax}_{q_1} J(q_1) = \operatorname*{argmin}_{q_1} \mathrm{KL}(q_1(y_1|\mathbf{x})||\bar{p}(y_1|\mathbf{x})) = \bar{p}(y_1|\mathbf{x}) \quad (29)$$

Mean-field VI: CAVI updates

- The same calculation for any i gives the optimal q_i when the others are fixed:

$$q_i(y_i|\mathbf{x}) \leftarrow \bar{p}(y_i|\mathbf{x}) \propto \exp\left[\mathbb{E}_{q(\mathbf{y}_{\setminus i}|\mathbf{x})}\left[\log p(\mathbf{x},\mathbf{y})\right]\right]$$

  where $q(\mathbf{y}_{\setminus i}|\mathbf{x}) = \prod_{j\neq i} q_j(y_j|\mathbf{x})$.
- If $p(\mathbf{x},\mathbf{y})$ factorises into factors $\phi_c(\mathbf{x}_c,\mathbf{y}_c)$, those without y_i can be dropped, as they get absorbed into the normalising constant:

$$q_i(y_i|\mathbf{x}) \propto \exp\Big[\sum_{c:\, y_i\in\mathbf{y}_c} \mathbb{E}_{q(\mathbf{y}_{\setminus i}|\mathbf{x})}\left[\log \phi_c(\mathbf{x}_c,\mathbf{y}_c)\right]\Big]$$

- Hence q_i only depends on the q_j of variables y_j that share a factor with y_i, i.e. its Markov blanket.

Mean-field VI: CAVI algorithm

- CAVI algorithm:
  1. Initialise $q_1, \ldots, q_d$.
  2. For $i = 1, \ldots, d$ (fixed or random order): update q_i, using the newest values of the other q_j.
  3. Repeat step 2 until the ELBO stops changing.
- Several sweeps are generally needed, as each update depends on the other q_j.
- Coordinate ascent in function space: the ELBO never decreases and is bounded by $\log p(\mathbf{x})$, so it converges, typically to a local optimum that depends on the initialisation.

Program

1. Preparations
2. The variational principle
3. Application to inference
   - The mechanics
   - Interpretation
   - Nature of the approximation
   - Mean-field variational inference
4. Application to learning

Program

1. Preparations
2. The variational principle
3. Application to inference
4. Application to learning
   - Learning with Bayesian models
   - Learning with statistical models and unobserved variables
   - (Variational) EM algorithm

Learning by Bayesian inference

- Task 1: For a Bayesian model $p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{x}, \boldsymbol{\theta})$, compute the posterior $p(\boldsymbol{\theta}|\mathcal{D})$
- Formally the same problem as before: $\mathcal{D} = \mathbf{x}_o$ and $\boldsymbol{\theta} \equiv \mathbf{y}$.
- Task 2: For a Bayesian model $p(\mathbf{v}, \mathbf{h}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{v}, \mathbf{h}, \boldsymbol{\theta})$, compute the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ where the data $\mathcal{D}$ are for the visibles $\mathbf{v}$ only.
- With the equivalence $\mathcal{D} = \mathbf{x}_o$ and $(\mathbf{h}, \boldsymbol{\theta}) \equiv \mathbf{y}$, we are formally back to the problem just studied.

Parameter estimation in presence of unobserved variables

- Task: For the model $p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})$, estimate the parameters $\boldsymbol{\theta}$ from data $\mathcal{D}$ on the visibles $\mathbf{v}$ only ($\mathbf{h}$ is unobserved).
- To evaluate the log likelihood function $\ell(\boldsymbol{\theta})$, we need to evaluate the integral

$$\ell(\boldsymbol{\theta}) = \log p(\mathcal{D}; \boldsymbol{\theta}) = \log \int_{\mathbf{h}} p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}, \tag{30}$$

  which is generally intractable.
- The integral has the same form as the marginal $p(\mathbf{x}) = \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$ considered before, so we can use the ELBO again with

$$\mathbf{x} \to \mathcal{D}, \qquad \mathbf{y} \to \mathbf{h}, \qquad p(\mathbf{x}, \mathbf{y}) \to p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta}) \tag{31}$$

MLE by maximising the ELBO

- This gives

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \mathbb{E}_{q(\mathbf{h}|\mathcal{D})}\left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h}|\mathcal{D})}\right] \tag{32}$$

$$= \log p(\mathcal{D}; \boldsymbol{\theta}) - \mathrm{KL}(q(\mathbf{h}|\mathcal{D})||p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})) \tag{33}$$

$$= \ell(\boldsymbol{\theta}) - \mathrm{KL}(q(\mathbf{h}|\mathcal{D})||p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})) \tag{34}$$

- If the search space $\mathcal{Q}$ is unrestricted or includes $p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})$

$$\max_{q} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \ell(\boldsymbol{\theta}) \tag{35}$$

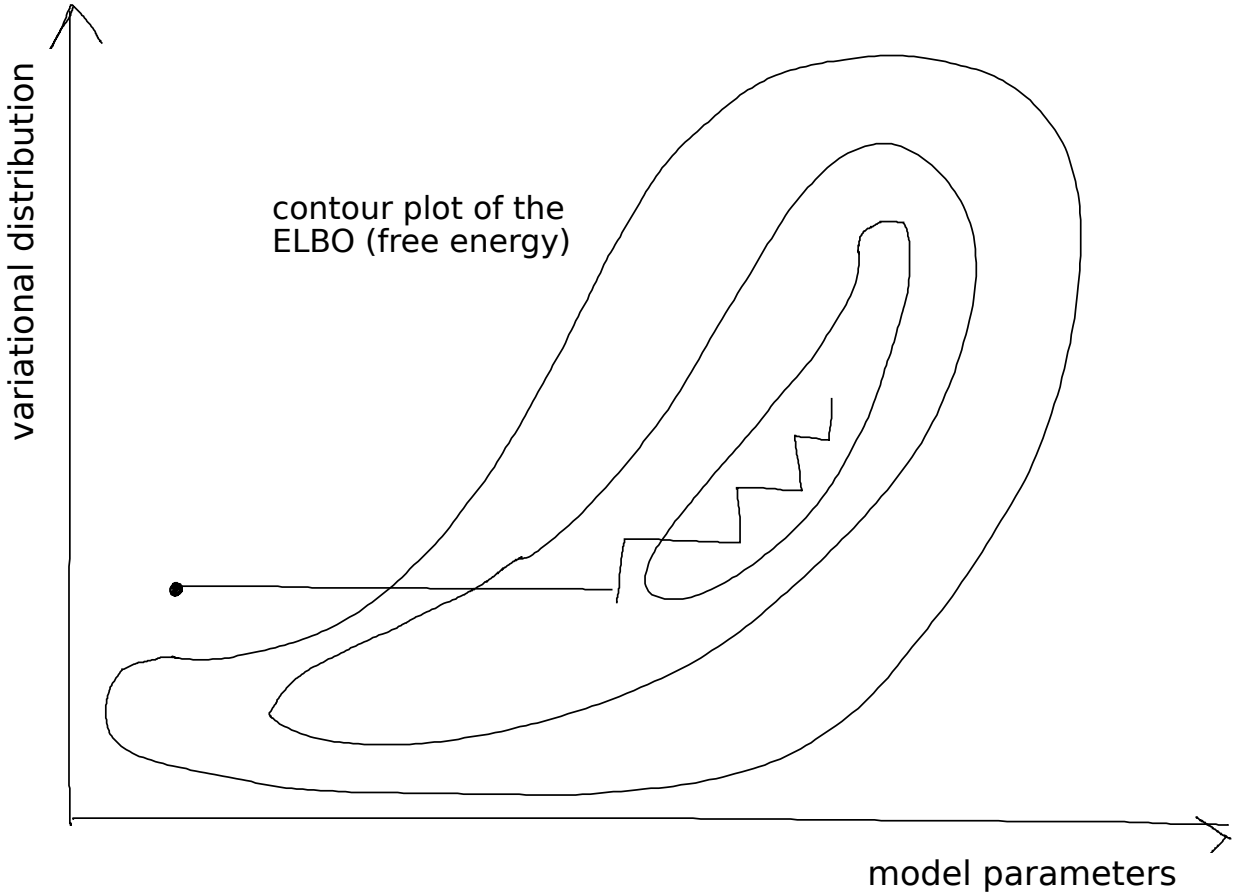
- Maximum likelihood estimation (MLE)

$$\max_{\boldsymbol{\theta}, q} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \tag{36}$$

  MLE = maximise the ELBO $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q)$ with respect to $\boldsymbol{\theta}$ and q
- Restricting the search space $\mathcal{Q}$ leads to an approximation when estimating $\boldsymbol{\theta}$ and $p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})$.

Variational EM algorithm

Variational expectation maximisation (EM): maximise $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q)$ by iterating between maximisation with respect to $\boldsymbol{\theta}$ and maximisation with respect to q (coordinate ascent).



(Adapted from `http://www.cs.cmu.edu/~tom/10-702/Zoubin-702.pdf`)

Maximisation with respect to q (E-step)

- The optimisation with respect to q is called the "expectation step".
- Denote the best $q \in \mathcal{Q}$ by q^* so that

$$\max_{q \in \mathcal{Q}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q^*) = \mathbb{E}_{q^*(\mathbf{h}|\mathcal{D})}\left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q^*(\mathbf{h}|\mathcal{D})}\right] \quad (37)$$

which is defined in terms of an expectation and the reason for the name "expectation step".

Maximisation with respect to $\boldsymbol{\theta}$ (M-step)

- With q fixed, only the first term of the ELBO depends on $\boldsymbol{\theta}$:

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \mathbb{E}_{q(\mathbf{h}|\mathcal{D})}\left[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})\right] - \underbrace{\mathbb{E}_{q(\mathbf{h}|\mathcal{D})}\left[\log q(\mathbf{h}|\mathcal{D})\right]}_{\text{does not depend on } \boldsymbol{\theta}}$$

- M-step: $\boldsymbol{\theta}_{k+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q(\mathbf{h}|\mathcal{D})}\left[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})\right]$
- For maximisation by gradient ascent:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \mathbb{E}_{q(\mathbf{h}|\mathcal{D})}\left[\nabla_{\boldsymbol{\theta}} \log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})\right]$$

- For iid data $\mathcal{D} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$ and $q(\mathbf{h}|\mathcal{D}) = \prod_i q_i(\mathbf{h}_i|\mathbf{v}_i)$:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \sum_{i=1}^{n} \mathbb{E}_{q_i(\mathbf{h}_i|\mathbf{v}_i)}\left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta})\right]$$

- Same as $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ in slides *Intractable Likelihood Functions*, case of unobserved variables, but with the posterior $p(\mathbf{h}_i|\mathbf{v}_i; \boldsymbol{\theta})$ replaced by q_i. Key difference: q_i stays fixed throughout the M-step, whereas $p(\mathbf{h}_i|\mathbf{v}_i; \boldsymbol{\theta})$ changes with every update of $\boldsymbol{\theta}$.

Classical EM algorithm

- Denote the parameters at iteration k by $\boldsymbol{\theta}_k$.
- We know that the optimal q for the expectation step is $q^*(\mathbf{h}|\mathcal{D}) = p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta}_k)$
- If we can compute the posterior $p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta}_k)$, we obtain the (classical) EM algorithm that iterates between:

  E-step: compute the expectation

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q^*) = \underbrace{\mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})]}_{\substack{\text{interpretation: expected} \\ \text{completed log-likelihood of } \boldsymbol{\theta}}} - \underbrace{\mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)} \log p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta}_k)}_{\substack{\text{does not depend on } \boldsymbol{\theta} \text{ and} \\ \text{does not need to be computed}}}$$

  M-step: maximise with respect to $\boldsymbol{\theta}$

$$\boldsymbol{\theta}_{k+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q^*) = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})]$$

- Gradient: same expression as for $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, but with the posterior fixed at $\boldsymbol{\theta}_k$. The two coincide at $\boldsymbol{\theta} = \boldsymbol{\theta}_k$.

Classical EM algorithm never decreases the log likelihood

- Assume you have updated the parameters and start iteration $k+1$ with optimisation with respect to q

$$\max_q \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_k, q) \tag{38}$$

- Optimal solution q^*_{k+1} is the posterior $p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta}_k)$ so that

$$\ell(\boldsymbol{\theta}_k) = \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_k, q^*_{k+1}) \tag{39}$$

- Optimise with respect to the $\boldsymbol{\theta}$ while keeping q fixed at q^*_{k+1}

$$\max_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q^*_{k+1}) \tag{40}$$

- Due to maximisation, updated parameter $\boldsymbol{\theta}_{k+1}$ is such that

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{k+1}, q^*_{k+1}) \geq \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_k, q^*_{k+1}) = \ell(\boldsymbol{\theta}_k) \tag{41}$$

- From variational lower bound: $\ell(\boldsymbol{\theta}) \geq \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q)$. Hence:

$$\ell(\boldsymbol{\theta}_{k+1}) \geq \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{k+1}, q^*_{k+1}) \geq \ell(\boldsymbol{\theta}_k)$$

$\Rightarrow$ EM yields non-decreasing sequence $\ell(\boldsymbol{\theta}_1), \ell(\boldsymbol{\theta}_2), \ldots$.

Program recap

1. Preparations
   - Concavity of the logarithm and Jensen's inequality
   - Kullback-Leibler divergence and its properties
2. The variational principle
   - Variational lower bound
   - Maximising the ELBO to compute the marginal and conditional from the joint
3. Application to inference
   - The mechanics
   - Interpretation
   - Nature of the approximation
   - Mean-field variational inference
4. Application to learning
   - Learning with Bayesian models
   - Learning with statistical models and unobserved variables
   - (Variational) EM algorithm