Variational Inference and Learning I

Basic Properties and Use

Michael U. Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, The University of Edinburgh

Spring Semester 2022

Recap

- Learning and inference often involves intractable integrals
- For example:
  - Marginalisation: $p(\mathbf{x}) = \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$
  - Likelihood in case of unobserved variables:
    $L(\boldsymbol{\theta}) = p(\mathcal{D}; \boldsymbol{\theta}) = \int_{\mathbf{u}} p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u}$
- We can use Monte Carlo integration and sampling to approximate the integrals.
- Alternative: variational approach to (approximate) inference and learning.

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 and 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 and learning

$\log(u)$ is a concave function

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

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

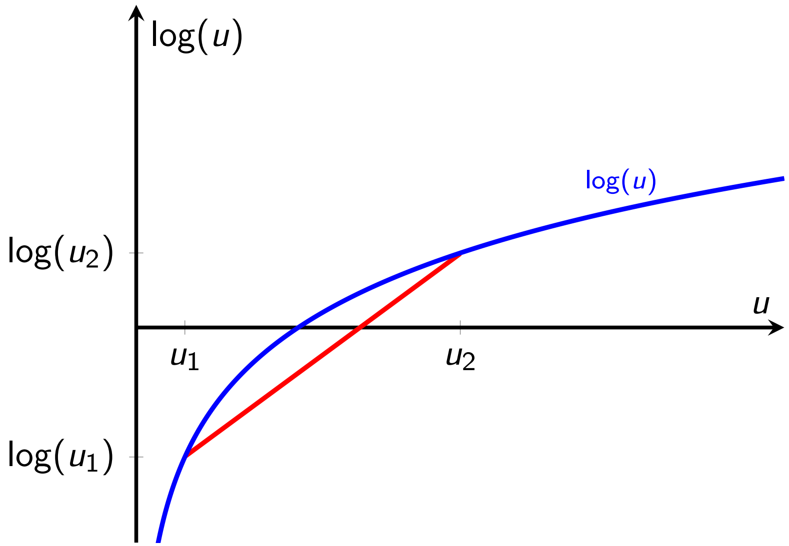
  $(1-a)x + ay$ with $a \in [0,1]$ linearly interpolates between x and y.

- 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 $\text{KL}(p||q)$

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

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

Non-negativity of the KL divergence

Non-negativity follows from the concavity of the logarithm.

$$
\begin{aligned}
-\mathrm{KL}(p||q) &= -\mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] && (2)\\
&= \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})}\right] && (3)\\
&\leq \log \underbrace{\mathbb{E}_{p(\mathbf{x})}\left[\frac{q(\mathbf{x})}{p(\mathbf{x})}\right]}_{\int p(\mathbf{x})q(\mathbf{x})/p(\mathbf{x})d\mathbf{x}=1} && (4)
\end{aligned}
$$

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

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

KL divergence and maximum likelihood-estimation

- Assume your data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is sampled iid from $p_*(\mathbf{x})$.
- Your model is $p(\mathbf{x}; \boldsymbol{\theta})$. Consider KL div $\mathrm{KL}(p_*(\mathbf{x})||p(\mathbf{x}; \boldsymbol{\theta}))$

$$\mathrm{KL}(p_*(\mathbf{x})||p(\mathbf{x}; \boldsymbol{\theta})) = \mathbb{E}_{p_*(\mathbf{x})}\left[\log \frac{p_*(\mathbf{x})}{p(\mathbf{x}; \boldsymbol{\theta})}\right] \tag{6}$$

$$= \mathbb{E}_{p_*(\mathbf{x})} \log p_*(\mathbf{x}) - \mathbb{E}_{p_*(\mathbf{x})} \log p(\mathbf{x}; \boldsymbol{\theta}) \tag{7}$$

- $\operatorname{argmin}_{\boldsymbol{\theta}} \mathrm{KL}(p_*(\mathbf{x})||p(\mathbf{x}; \boldsymbol{\theta})) = \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{p_*(\mathbf{x})} \log p(\mathbf{x}; \boldsymbol{\theta})$
- Approximating the expectation $\mathbb{E}_{p_*(\mathbf{x})}$ with a sample average gives log-likelihood (scaled by $1/n$)

$$\frac{1}{n}\ell(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta}) \tag{8}$$

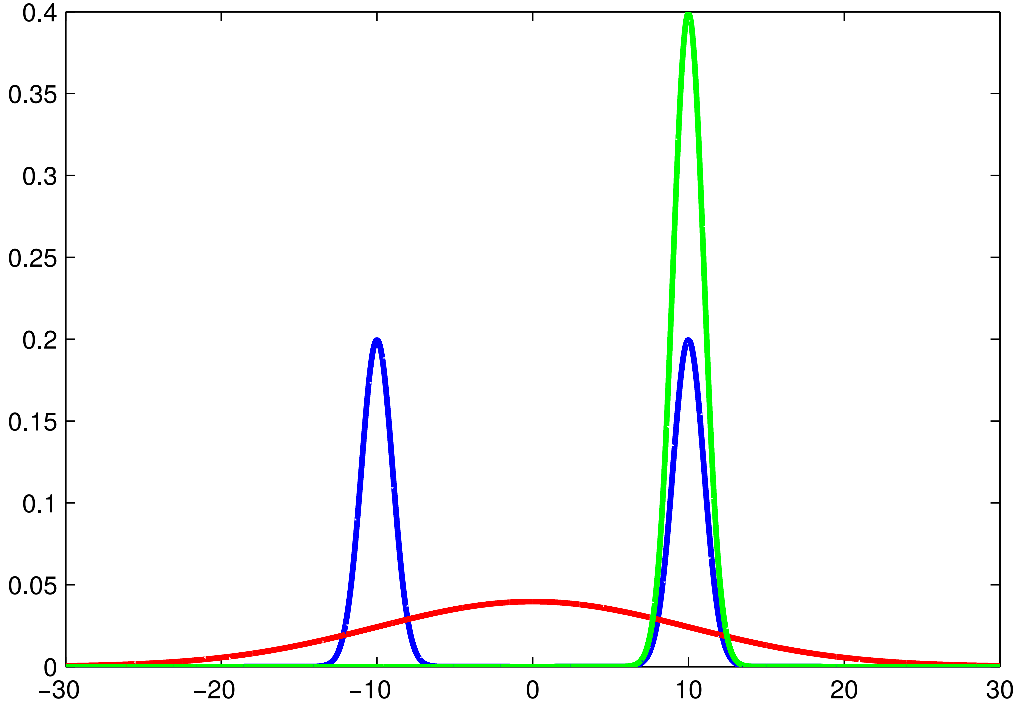
- Hence: $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \approx \operatorname{argmin}_{\boldsymbol{\theta}} \mathrm{KL}(p_*(\mathbf{x})||p(\mathbf{x}; \boldsymbol{\theta}))$

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

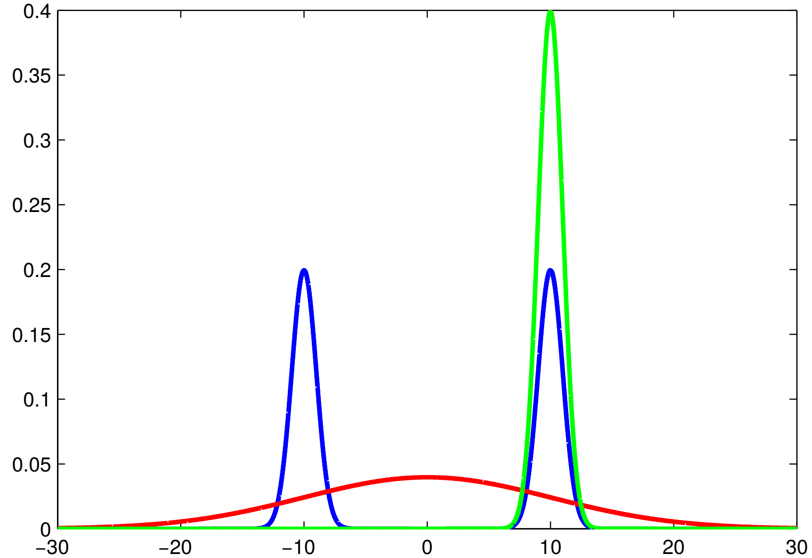
Asymmetry of the KL divergence

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

- Optimal q avoids regions where p is small.
  (but can be small where p is large)
- Produces good local fit, "mode seeking"

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

- Optimal q is nonzero where p is nonzero
  (and does not care about regions where p is small)
- Corresponds to MLE; produces global fit/moment matching

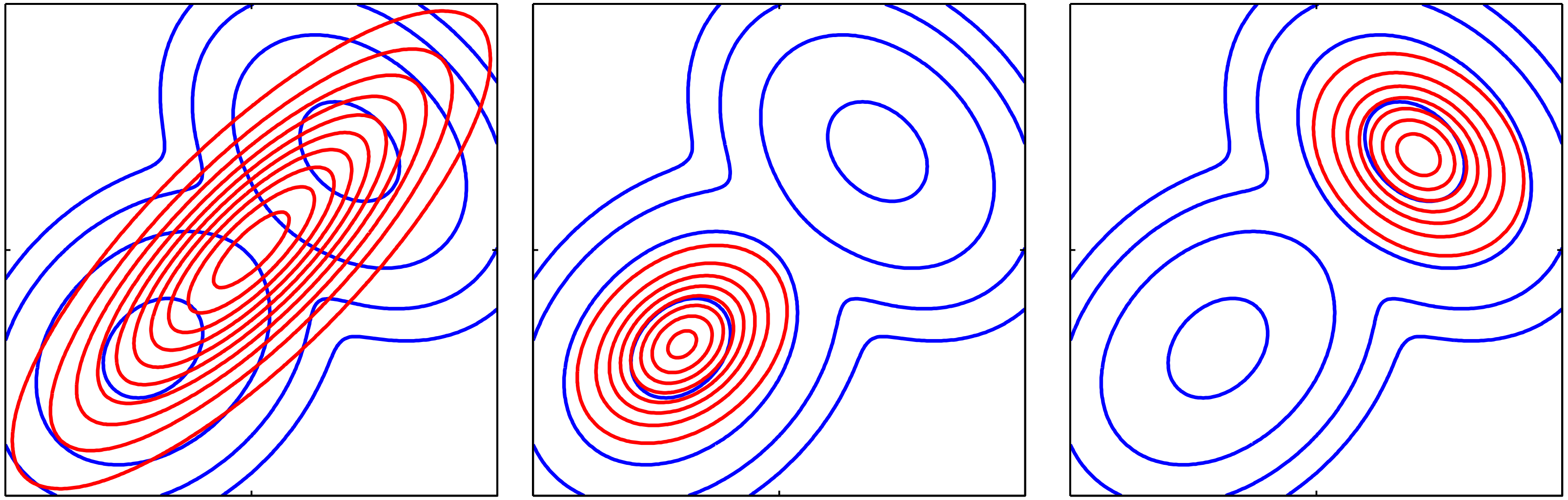


Asymmetry of the KL divergence

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

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

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



$\min_q$ KL(p || q) $\quad$ $\min_q$ KL(q || p) $\quad$ $\min_q$ 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 and 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 and learning

Variational lower bound: auxiliary distribution

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

- Like in importance sampling, we write

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

  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

$$\log p(\mathbf{x}) = \log \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \left[\frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})} \right] \tag{10}$$

- Instead of approximating the expectation with a sample average, use now the concavity of the logarithm.

Variational lower bound: concavity of the logarithm

- Concavity of the log gives

$$\log p(\mathbf{x}) = \log \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \geq \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \quad (11)$$

  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) $\mathcal{L}_{\mathbf{x}}(q)$

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x},\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \quad (12)$$

- 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

$$\begin{aligned}\mathcal{L}_{\mathbf{x}}(q) &= \log p(\mathbf{x}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) & (13)\\ &= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{y}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y})) & (14)\\ &= \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x},\mathbf{y}) + \mathcal{H}(q) & (15)\end{aligned}$$

  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

$$\mathcal{H}(q) = -\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{x})\right] \quad (16)$$

  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 because 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

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y} \tag{17}$$

  using importance sampling and Jensen's inequality.
- 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

$$\mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) = \log p(\mathbf{x}) - \mathcal{L}_{\mathbf{x}}(q) \tag{18}$$

  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: Inference becomes optimisation.

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:
  - optimisation with respect to pdf/pmf $q(\mathbf{y}|\mathbf{x})$
  - computation of the expectation
- Restrict search space to family of variational distributions $q(\mathbf{y}|\mathbf{x})$ for which $\mathcal{L}_{\mathbf{x}}(q)$ is computable.
- 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. $q(y_i|\mathbf{x}) = \mathcal{N}(y_i; \mu_i(\mathbf{x}), \sigma_i^2(\mathbf{x}))$
- Discussed in more detail later.
- $\mathcal{L}_{\mathbf{x}}(q)$ can be computed analytically in closed form only in special cases.

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 and learning

Program

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

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

$$\hat{p}(\mathbf{y}|\mathbf{x}_o) = \underset{q \in \mathcal{Q}}{\operatorname{argmax}} \, \mathcal{L}_{\mathbf{x}_o}(q) \tag{19}$$

  $\mathcal{Q}$ is the set of pdfs/pmfs in which we search for the solution
- The decomposition of the log marginal gives

$$\log p(\mathbf{x}_o) = \mathrm{KL}(q(\mathbf{y}|\mathbf{x}_o)||p(\mathbf{y}|\mathbf{x}_o)) + \mathcal{L}_{\mathbf{x}_o}(q) = \text{const} \tag{20}$$

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

$$\underset{q \in \mathcal{Q}}{\operatorname{argmax}} \, \mathcal{L}_{\mathbf{x}_o}(q) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}(q(\mathbf{y}|\mathbf{x}_o)||p(\mathbf{y}|\mathbf{x}_o)) \tag{21}$$

Posterior as compromise between prior and fit

- Equivalent forms of the ELBO:

$$\mathcal{L}_{\mathbf{x}_o}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x}_o)} \log p(\mathbf{x}_o|\mathbf{y}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x}_o)||p(\mathbf{y})) \qquad (22)$$

- 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:

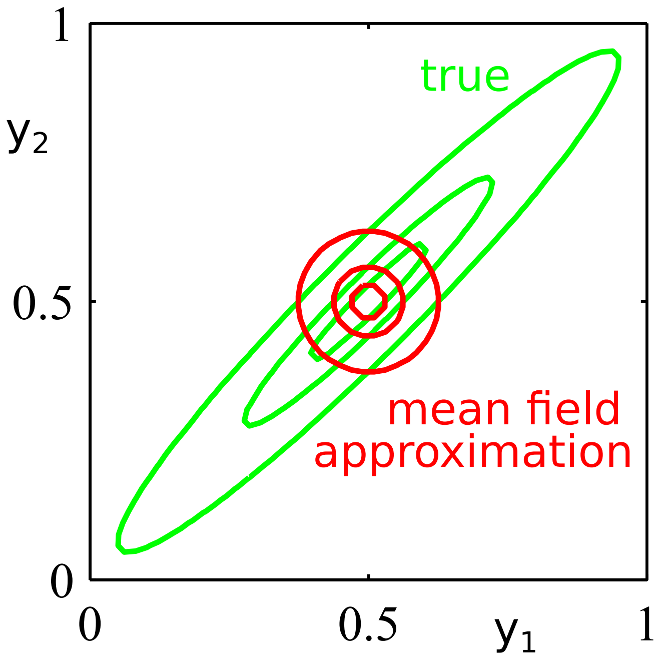
$$\mathcal{L}_{\mathbf{x}_o}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x}_o)} \log p(\mathbf{x}_o, \mathbf{y}) + \mathcal{H}(q) \tag{23}$$

- By maximising $\mathcal{L}_{\mathbf{x}_o}(q)$ we find a q that
  - produces likely imputations $\mathbf{y}$
  - 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

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

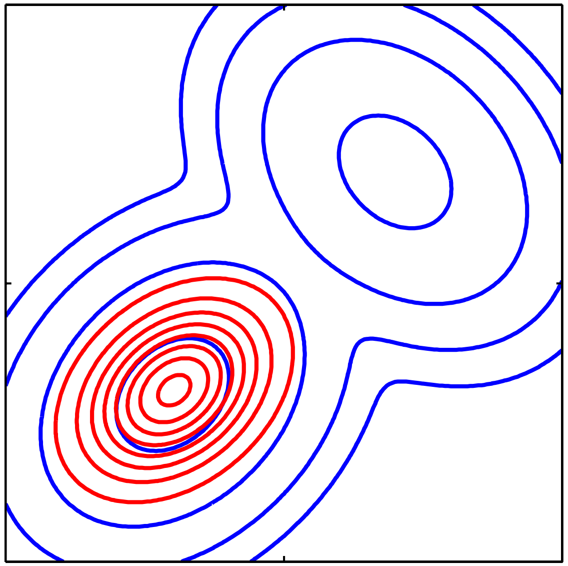
- When minimising $\mathrm{KL}(q||p)$ with respect to q, q will try very hard to be zero where p is small.
- Assume true posterior is correlated bivariate Gaussian and we work with $\mathcal{Q} = \{q(\mathbf{y}|\mathbf{x}_o) : q(\mathbf{y}|\mathbf{x}_o) = q(y_1|\mathbf{x}_o)q(y_2|\mathbf{x}_o)\}$ (independence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



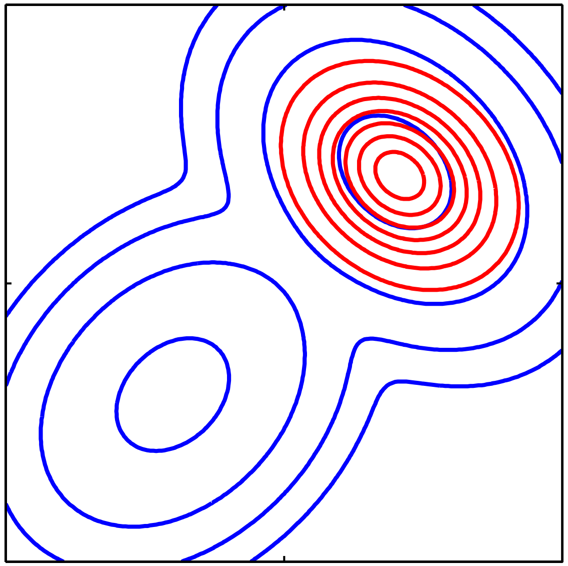
(Bishop, Figure 10.2)

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)

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).
- See slides on *Intractable Likelihood Functions*: the log likelihood function $\ell(\boldsymbol{\theta})$ is implicitly defined by the integral

$$\ell(\boldsymbol{\theta}) = \log p(\mathcal{D}; \boldsymbol{\theta}) = \log \int_{\mathbf{h}} p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}, \tag{24}$$

  which is generally intractable.
- We could approximate $\ell(\boldsymbol{\theta})$ and its gradient using Monte Carlo integration.
- Here: use the variational approach.

Parameter estimation in presence of unobserved variables

- We had

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \tag{25}$$

$$= \log p(\mathbf{x}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) \tag{26}$$

- Substitute

$$\mathbf{x} \to \mathcal{D}, \qquad \mathbf{y} \to \mathbf{h}, \qquad p(\mathbf{x}, \mathbf{y}) \to p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta}) \tag{27}$$

- We then have

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \mathbb{E}_{q(\mathbf{h}|\mathcal{D})}\left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h}|\mathcal{D})}\right] \tag{28}$$

$$= \log p(\mathcal{D}; \boldsymbol{\theta}) - \mathrm{KL}(q(\mathbf{h}|\mathcal{D})||p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})) \tag{29}$$

- Notation $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q)$ highlights dependency on $\boldsymbol{\theta}$ and q.

MLE by maximising the ELBO

- Using $\ell(\boldsymbol{\theta})$ for the log-likelihood $\log p(\mathcal{D}; \boldsymbol{\theta})$, we have

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \ell(\boldsymbol{\theta}) - \mathrm{KL}(q(\mathbf{h}|\mathcal{D})||p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})) \tag{30}$$

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

$$\max_q \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \ell(\boldsymbol{\theta}) \tag{31}$$

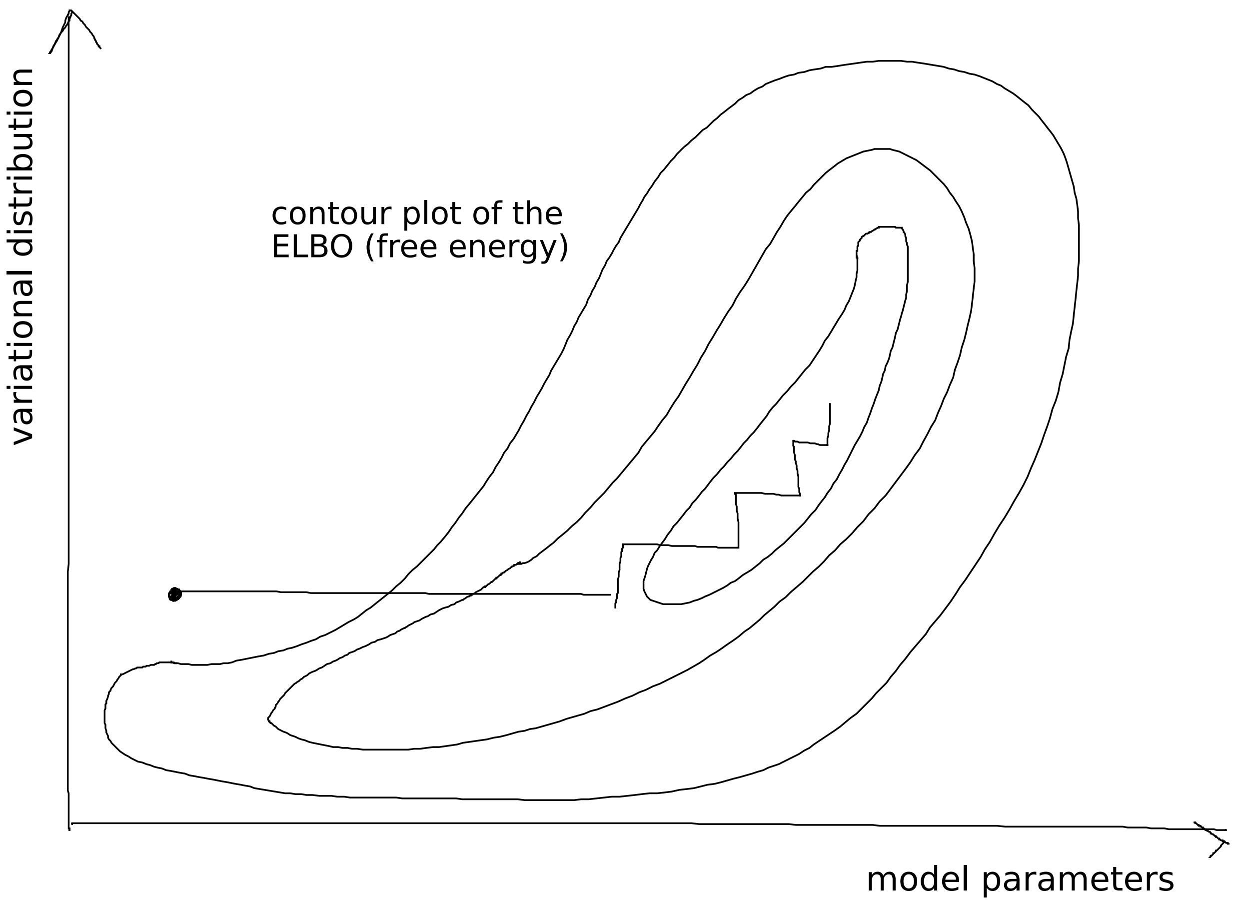
- Maximum likelihood estimation (MLE)

$$\max_{\boldsymbol{\theta}, q} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \tag{32}$$

  MLE = maximise the ELBO $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q)$ with respect to $\boldsymbol{\theta}$ and q
- Restricted search space $\mathcal{Q}$ leads to approximate estimate of $\boldsymbol{\theta}$ 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`)

Where is the "expectation"?

- The optimisation with respect to q is called the "expectation step"

$$\max_{q \in \mathcal{Q}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{h}|\mathcal{D})} \left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h}|\mathcal{D})} \right] \tag{33}$$

- Denote the best q by q^* so that

$$\max_{q \in \mathcal{Q}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q^*) = \mathbb{E}_{q^*(\mathbf{h}|\mathcal{D})} \left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q^*(\mathbf{h}|\mathcal{D})} \right] \tag{34}$$

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

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

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

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

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

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

$$\max_q \mathcal{L}_\mathcal{D}(\boldsymbol{\theta}_k, q) \tag{35}$$

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

$$\ell(\boldsymbol{\theta}_k) = \mathcal{L}_\mathcal{D}(\boldsymbol{\theta}_k, q^*_{k+1}) \tag{36}$$

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

$$\max_{\boldsymbol{\theta}} \mathcal{L}_\mathcal{D}(\boldsymbol{\theta}, q^*_{k+1}) \tag{37}$$

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

$$\mathcal{L}_\mathcal{D}(\boldsymbol{\theta}_{k+1}, q^*_{k+1}) \geq \mathcal{L}_\mathcal{D}(\boldsymbol{\theta}_k, q^*_{k+1}) = \ell(\boldsymbol{\theta}_k) \tag{38}$$

- 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 and learning
   - Inference: approximating posteriors
   - Learning with Bayesian models
   - Learning with statistical models and unobserved variables
   - (Variational) EM algorithm