# Variational Inference and Learning

Michael Gutmann

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

Spring Semester 2020

# Recap

- Learning and inference often involves intractable integrals
- For example: marginalisation

$$p(\mathbf{x}) = \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$$

- For example: 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 of interest 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 is concave

- $\log(u)$ is concave

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

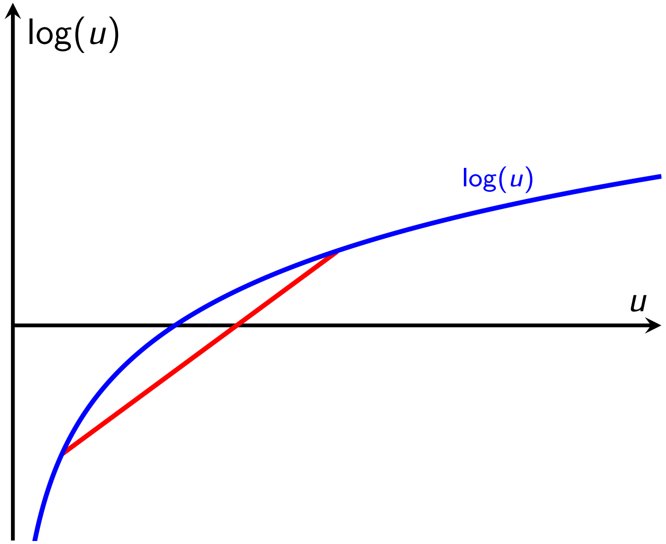
  $(1-a)x + ay$ with $a \in [0,1]$ linearly interpolates between $x$ and $y$.

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

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

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

# Non-negativity of the KL divergence

Non-negativity follows from the concavity of the logarithm.

$$
\begin{aligned}
\mathbb{E}_{p(\mathbf{x})}\left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})}\right] &\leq \log \mathbb{E}_{p(\mathbf{x})}\left[\frac{q(\mathbf{x})}{p(\mathbf{x})}\right] \\
&= \log \int p(\mathbf{x}) \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} \\
&= \log \int q(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \log 1 = 0.
\end{aligned}
$$

From

$$
\mathbb{E}_{p(\mathbf{x})}\left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})}\right] \leq 0
$$

it follows that

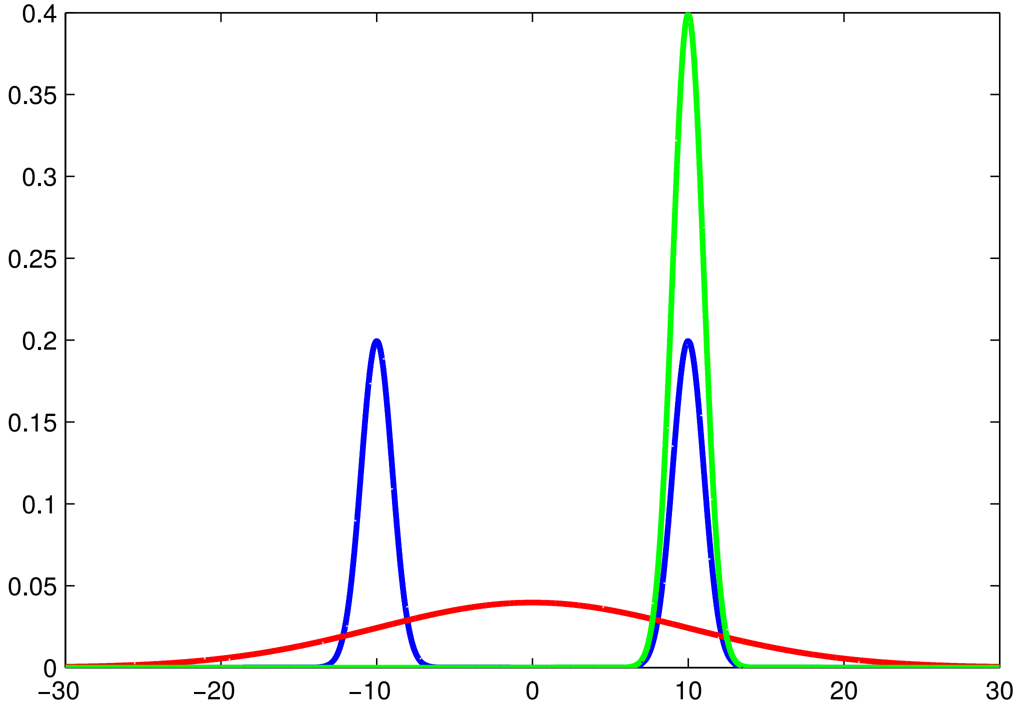
$$
\mathrm{KL}(p||q) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] = -\mathbb{E}_{p(\mathbf{x})}\left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})}\right] \geq 0
$$

# Asymmetry of the KL divergence

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

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

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




Barber Figure 28.1, Section 28.3.4

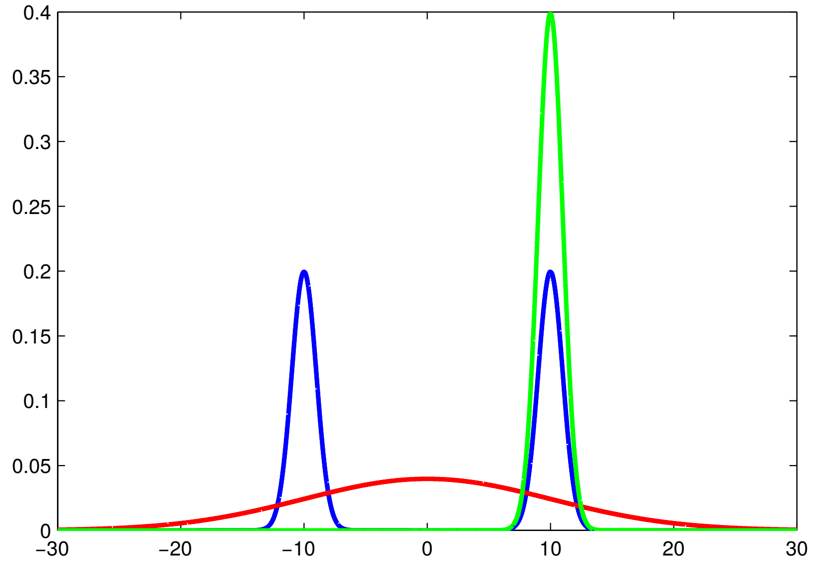
# Asymmetry of the KL divergence

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

- 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

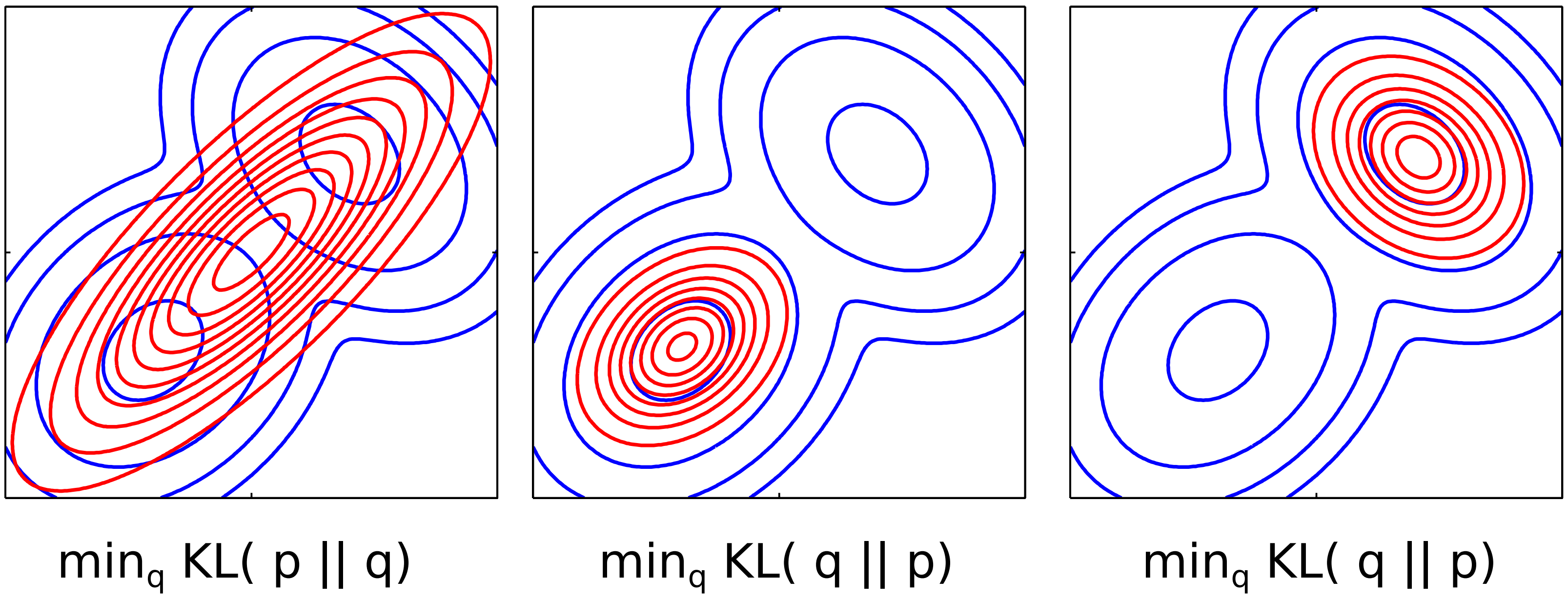


# Asymmetry of the KL divergence

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)




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
    - Free energy and the decomposition of the log marginal
    - Free energy maximisation 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

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

  where $q(\mathbf{y})$ is an auxiliary distribution (called the variational distribution in the context of variational inference/learning)

- Log marginal is

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

- 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

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

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

- Right-hand side is called the (variational) free energy

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

  It depends on $\mathbf{x}$ through the joint $p(\mathbf{x},\mathbf{y})$, and on the auxiliary distribution $q(\mathbf{y})$

  (since $q$ is a function, the free energy is called a functional, which is a mapping that depends on a function)

# Decomposition of the log marginal

- We can re-write the free energy as

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

- Hence: $\log p(\mathbf{x}) = \mathrm{KL}(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x})) + \mathcal{F}(\mathbf{x}, q)$
- $\mathrm{KL} \geq 0$ implies the bound $\log p(\mathbf{x}) \geq \mathcal{F}(\mathbf{x}, q)$ that we have derived on the previous slide.
- $\mathrm{KL}(q||p) = 0$ iff $q = p$ implies that for $q(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})$, the free energy is maximised and equals $\log p(\mathbf{x})$ .

# Alternative approach

- We started from the task of approximating the marginal

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

  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 $q(\mathbf{y})$.
- Measuring the quality of the approximation $q(\mathbf{y})$ by $\mathrm{KL}(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x}))$ gives the same decomposition:

$$\log p(\mathbf{x}) = \mathrm{KL}(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x})) + \mathcal{F}(\mathbf{x}, q)$$

# Variational principle

- By maximising the free energy

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

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

$$\log p(\mathbf{x}) = \max_{q(\mathbf{y})} \mathcal{F}(\mathbf{x}, q)$$

$$p(\mathbf{y}|\mathbf{x}) = \operatorname{argmax}_{q(\mathbf{y})} \mathcal{F}(\mathbf{x}, q)$$

- You can think of free energy maximisation as a "function" that takes as input a joint $p(\mathbf{x}, \mathbf{y})$ and returns as output the (log) marginal and the conditional.

# Variational principle

- Given $p(\mathbf{x}, \mathbf{y})$, consider the inference tasks
  1. compute $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$
  2. compute $p(\mathbf{y}|\mathbf{x})$
- Variational principle: we can formulate the inference problems as an optimisation problem.
- Maximising the free energy

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

  gives
  1. $\log p(\mathbf{x}) = \max_{q(\mathbf{y})} \mathcal{F}(\mathbf{x}, q)$
  2. $p(\mathbf{y}|\mathbf{x}) = \operatorname{argmax}_{q(\mathbf{y})} \mathcal{F}(\mathbf{x}, q)$
- Inference becomes optimisation.
- The (optimal) variational distribution $q(\mathbf{y})$ depends on the value of $\mathbf{x}$. Notation to highlight the dependency: $q(\mathbf{y}|\mathbf{x})$.

# Solving the optimisation problem

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

- Difficulties when maximising the free energy:
  - optimisation with respect to pdf/pmf $q(\mathbf{y})$
  - computation of the expectation
- Restrict search space to family of variational distributions $q(\mathbf{y})$ for which $\mathcal{F}(\mathbf{x}, q)$ is computable.
- Family $\mathcal{Q}$ specified by
  - independence assumptions, e.g. $q(\mathbf{y}) = \prod_i q(y_i)$, which corresponds to "mean-field" variational inference
  - parametric assumptions, e.g. $q(y_i) = \mathcal{N}(y_i; \mu_i(\mathbf{x}), \sigma_i^2(\mathbf{x}))$
- Optimisation is generally challenging: lots of research on how to do it (keywords: stochastic variational inference, black-box variational inference)

# Program

1. Preparations

2. The variational principle
   - Variational lower bound
   - Free energy and the decomposition of the log marginal
   - Free energy maximisation 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

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

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

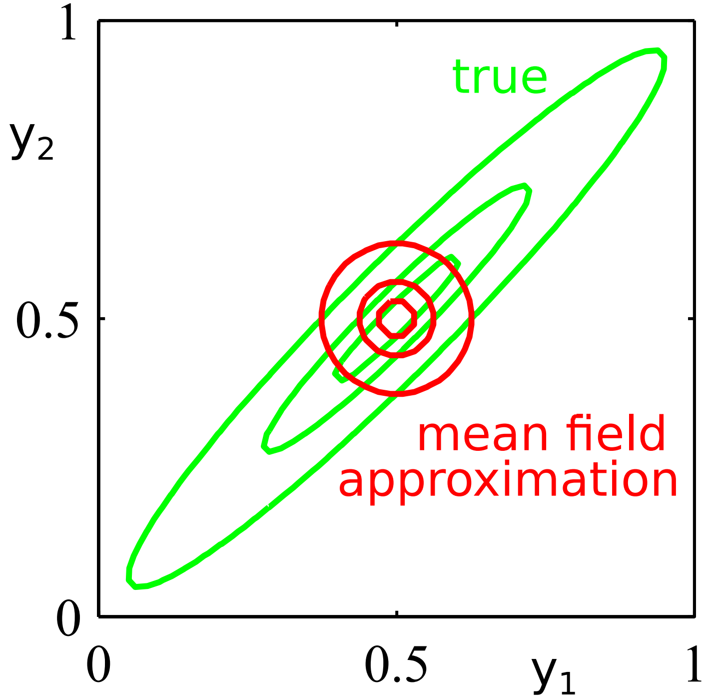
$$\log p(\mathbf{x}_o) = \mathrm{KL}(q(\mathbf{y})||p(\mathbf{y}|\mathbf{x}_o)) + \mathcal{F}(\mathbf{x}_o, q) = \text{const}$$

- Because the sum of the KL and free energy term is constant we have

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

# Nature of the approximation

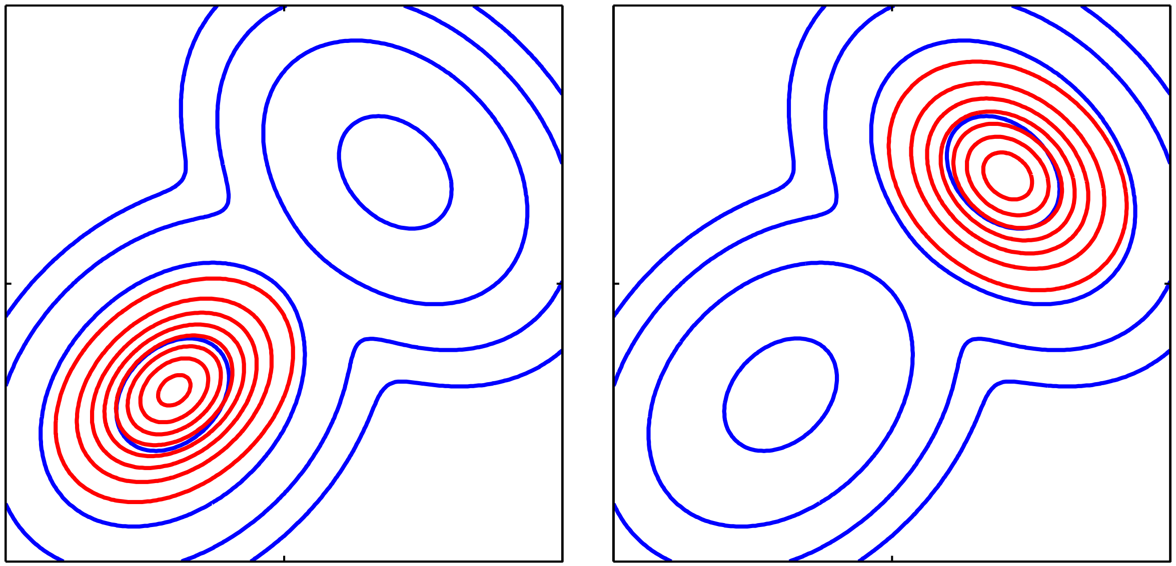
- When minimising $\mathrm{KL}(q||p)$ with respect to $q$, $q$ will try very hard to be zero where $p$ is small.
- Assume true posterior is correlated bivariate Gaussian and we work with $\mathcal{Q} = \{q(\mathbf{y}) : q(\mathbf{y}) = q(y_1)q(y_2)\}$
  (independence but no parametric assumptions)
- $\hat{p}(\mathbf{y}|\mathbf{x}_o)$, i.e. $q(\mathbf{y})$ that minimises $\mathrm{KL}(q||p)$, is Gaussian.
- Mean is correct but variances dictated by the marginal variances of $p(\mathbf{y})$ 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 learned approximate posterior $\hat{p}(\mathbf{y}|\mathbf{x}_o)$ 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.
- But the variational distribution $q(\mathbf{y})$ becomes $q(\mathbf{h}, \boldsymbol{\theta})$.
- Often: assume $q(\mathbf{h}, \boldsymbol{\theta})$ factorises as $q(\mathbf{h})q(\boldsymbol{\theta})$ (see Barber Section 11.5)

# 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

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

  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

- Foundational result that we have derived

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

$$\log p(\mathbf{x}) = \max_{q(\mathbf{y})} \mathcal{F}(\mathbf{x}, q) \qquad p(\mathbf{y}|\mathbf{x}) = \operatorname*{argmax}_{q(\mathbf{y})} \mathcal{F}(\mathbf{x}, q)$$

- Correspondences:

$$\mathbf{v} \equiv \mathbf{x} \qquad \mathbf{h} \equiv \mathbf{y} \qquad p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) \equiv p(\mathbf{x}, \mathbf{y})$$

- Foundational result becomes

$$\log p(\mathbf{v}; \boldsymbol{\theta}) = \mathrm{KL}(q(\mathbf{h})||p(\mathbf{h}|\mathbf{v}; \boldsymbol{\theta})) + \mathcal{F}(\mathbf{v}, q; \boldsymbol{\theta}) \quad \mathcal{F}(\mathbf{v}, q; \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{h})}\left[\log \frac{p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h})}\right]$$

$$\log p(\mathbf{v}; \boldsymbol{\theta}) = \max_{q(\mathbf{h})} \mathcal{F}(\mathbf{v}, q; \boldsymbol{\theta}) \qquad p(\mathbf{h}|\mathbf{v}; \boldsymbol{\theta}) = \operatorname*{argmax}_{q(\mathbf{h})} \mathcal{F}(\mathbf{v}, q; \boldsymbol{\theta})$$

- Plug in $\mathcal{D}$ for $\mathbf{v}$: $\log p(\mathbf{v}; \boldsymbol{\theta})$ becomes $\log p(\mathcal{D}; \boldsymbol{\theta})$, which is $\ell(\boldsymbol{\theta})$

# Approximate MLE by free energy maximisation

- With $\mathbf{v} = \mathcal{D}$ and $\ell(\boldsymbol{\theta}) = \log p(\mathcal{D}; \boldsymbol{\theta})$, the equations become

$$\ell(\boldsymbol{\theta}) = \mathrm{KL}(q(\mathbf{h})||p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})) + \overbrace{\mathcal{F}(\mathcal{D}, q; \boldsymbol{\theta})}^{J_{\mathcal{F}}(q,\boldsymbol{\theta})} \qquad J_{\mathcal{F}}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{h})}\left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h})}\right]$$

$$\ell(\boldsymbol{\theta}) = \max_{q(\mathbf{h})} J_{\mathcal{F}}(q, \boldsymbol{\theta}) \qquad p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta}) = \operatorname*{argmax}_{q(\mathbf{h})} J_{\mathcal{F}}(q, \boldsymbol{\theta})$$

  Write $J_{\mathcal{F}}(q, \boldsymbol{\theta})$ for $\mathcal{F}(\mathcal{D}, q; \boldsymbol{\theta})$ when data $\mathcal{D}$ are fixed.

- Maximum likelihood estimation (MLE)

$$\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \max_{q(\mathbf{h})} J_{\mathcal{F}}(q, \boldsymbol{\theta})$$

  MLE = maximise the free energy with respect to $\boldsymbol{\theta}$ and $q(\mathbf{h})$

- Restricting the search space $\mathcal{Q}$ for the variational distribution $q(\mathbf{h})$ for computational reasons leads to an approximation.

# Free energy as sum of completed log likelihood and entropy

- We can write the free energy as

$$J_{\mathcal{F}}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{h})}\left[\log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h})}\right] = \mathbb{E}_{q(\mathbf{h})}\left[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})\right] - \mathbb{E}_{q(\mathbf{h})}\left[\log q(\mathbf{h})\right]$$

- $-\mathbb{E}_{q(\mathbf{h})}[\log q(\mathbf{h})]$ is the entropy of $q(\mathbf{h})$
  (entropy is a measure of randomness or variability, see e.g. Barber Section 8.2)
- $\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})$ is the log-likelihood for the filled-in data $(\mathcal{D}, \mathbf{h})$
- $\mathbb{E}_{q(\mathbf{h})}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})]$ is the weighted average of these "completed" log-likelihoods, with the weighting given by $q(\mathbf{h})$.

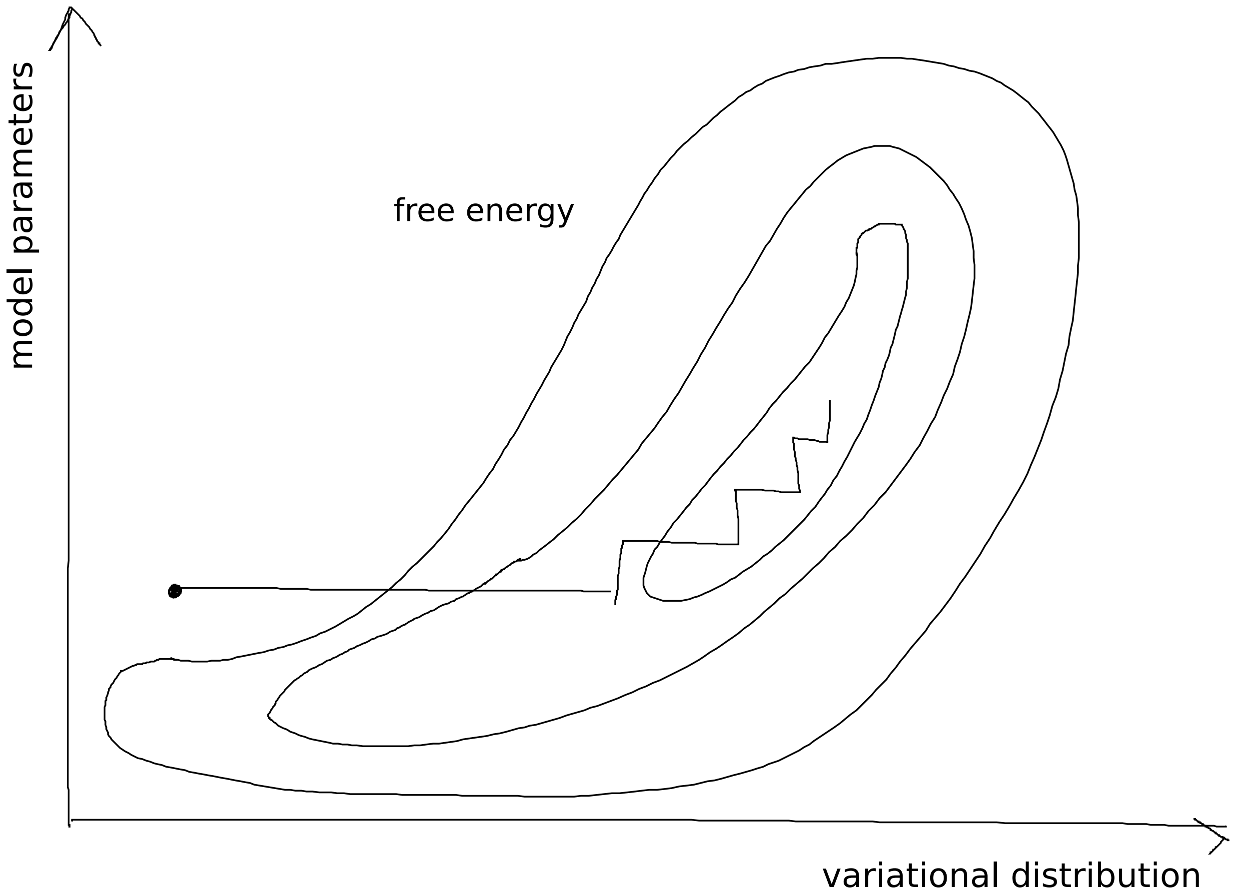
# Free energy as sum of completed log likelihood and entropy

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

- When maximising $J_{\mathcal{F}}(q, \boldsymbol{\theta})$ with respect to $q$ we look for random variables $\mathbf{h}$ (filled-in data) that
  - are maximally variable (large entropy)
  - are maximally compatible with the observed data (according to the model $p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})$)
- If included in the search space $\mathcal{Q}$, $p(\mathbf{h}|\mathcal{D}; \boldsymbol{\theta})$ is the optimal $q$, which means that the posterior fulfils the two desiderata best.

# Variational EM algorithm

Variational expectation maximisation (EM): maximise $J_{\mathcal{F}}(q, \boldsymbol{\theta})$ by iterating between maximisation with respect to $q$ and maximisation with respect to $\boldsymbol{\theta}$ (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"

$$\max_{q \in \mathcal{Q}} J_{\mathcal{F}}(q, \boldsymbol{\theta}) = \max_{q \in \mathcal{Q}} \mathbb{E}_q \left[ \log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h})} \right]$$

- Denote the best $q$ by $q^*$ so that $\max_{q \in \mathcal{Q}} J_{\mathcal{F}}(q, \boldsymbol{\theta}) = J_{\mathcal{F}}(q^*, \boldsymbol{\theta})$
- By definition of $J_{\mathcal{F}}(q, \boldsymbol{\theta})$, we have

$$J_{\mathcal{F}}(q^*, \boldsymbol{\theta}) = \mathbb{E}_{q^*} \left[ \log \frac{p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})}{q^*(\mathbf{h})} \right]$$

- $J_{\mathcal{F}}(q^*, \boldsymbol{\theta})$ is defined in terms of an expectation and the reason for the name "expectation step".

# Classical EM algorithm

- From

$$\ell(\boldsymbol{\theta}_k) = \mathrm{KL}(q(\mathbf{h})||p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)) + J_{\mathcal{F}}(q,\boldsymbol{\theta}_k)$$

  we know that the optimal $q(\mathbf{h})$ is $q^*(\mathbf{h}) = p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)$
- If we can compute the posterior $p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_k)$, we obtain the (classical) EM algorithm that iterates between:

  Expectation step

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

  Maximisation step

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

# Classical EM algorithm never decreases the log likelihood

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

$$\max_q J_{\mathcal{F}}(q, \boldsymbol{\theta}_k)$$

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

$$\ell(\boldsymbol{\theta}_k) = J_{\mathcal{F}}(q^*_{k+1}, \boldsymbol{\theta}_k)$$

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

$$\max_{\boldsymbol{\theta}} J_{\mathcal{F}}(q^*_{k+1}, \boldsymbol{\theta})$$

- Because of maximisation, optimiser $\boldsymbol{\theta}_{k+1}$ is such that

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

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

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

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

# Examples

- Work through the examples in Barber Section 11.2 for the classical EM algorithm.
- Example 11.4 treats the cancer-asbestos-smoking example that we had in an earlier lecture.

# 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
   - Free energy and the decomposition of the log marginal
   - Free energy maximisation 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