Unsupervised Learning

Variational Methods
and Other Approximations

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, and
MSc in Intelligent Systems, Dept Computer Science
University College London

Term 1, Autumn 2006
Integrals in Statistical Modelling

• Parameter estimation

\[ \hat{\theta} = \arg\max_{\theta} \int dY P(Y|\theta)P(X|Y, \theta) \]

(or using EM)

\[ \theta^{\text{new}} = \arg\max_{\theta} \int dY P(Y|X, \theta^{\text{old}}) \log P(X, Y|\theta) \]

• Prediction

\[ p(x|D, m) = \int d\theta p(\theta|D, m)p(x|\theta, D, m) \]

• Model selection or weighting (by marginal likelihood)

\[ p(D|m) = \int d\theta p(\theta|m)p(D|\theta, m) \]

These integrals are often intractable:

• **Analytic intractability**: integrals may not have closed form in non-linear, non-Gaussian models \(\Rightarrow\) numerical integration.

• **Computational intractability**: Numerical integral (or sum if \(Y\) or \(\theta\) are discrete) may be exponential in data or model size.
Examples of Intractability

• Bayesian marginal likelihood/model evidence for Mixture of Gaussians: exact computations are exponential in number of data points

\[ p(x_1, \ldots, x_N) = \int d\theta \ p(\theta) \prod_{i=1}^{N} \sum_{s_i} p(x_i|s_i, \theta)p(s_i|\theta) \]

\[ = \sum_{s_1} \sum_{s_2} \ldots \sum_{s_N} \int d\theta \ p(\theta) \prod_{i=1}^{N} p(x_i|s_i, \theta)p(s_i|\theta) \]

• Computing the conditional probability of a variable in a very large multiply connected directed graphical model:

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

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

\[ p(y_t|x_1, \ldots, x_T) \propto \int p(y_t|y_{t-1})p(x_t|y_t)p(y_{t-1}|x_1, \ldots, x_{t-1})p(x_{t+1}, \ldots, x_t|y_t)dy_{t-1} \]
In the FHMM, moralisation puts simultaneous states $s_{t}^{(1)}, s_{t}^{(2)}, s_{t}^{(3)}$ into a single clique.

- $M$ state variables, $K$ values $\Rightarrow$ sums over $K^{2M}$ terms.
- Factorial prior $\not\Rightarrow$ Factorial posterior (explaining away).

Variational methods approximate the posterior, often in a factored form. To see how they work, we need to review the free-energy interpretation of EM.
The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{x_i\}$; Latent variables $\mathcal{Y} = \{y_i\}$; Parameters $\theta$.

Goal: Maximize the log likelihood (i.e. ML learning) wrt $\theta$:

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

Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen’s inequality:

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

Now,

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

$$= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y} + H[q],$$

where $H[q]$ is the entropy of $q(\mathcal{Y})$.

So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + H[q]$$
The log likelihood is bounded below (Jensen) by:

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

EM alternates between:

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

$$q^{(k)}(Y) := \arg\max_{q(Y)} \mathcal{F}(q(Y), \theta^{(k-1)}) = P(Y|X, \theta^{(k-1)})$$

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

$$\theta^{(k)} := \arg\max_{\theta} \mathcal{F}(q^{(k)}(Y), \theta) = \arg\max_{\theta} \langle \log P(Y, X|\theta) \rangle_{q^{(k)}(Y)}$$
EM as Coordinate Ascent in $\mathcal{F}$

$F(Q, \theta)$

$Q(x)$

$\theta$
EM Never Decreases the Likelihood

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

\[ \ell(\theta^{(k-1)}) = E \text{ step } \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \leq M \text{ step } \mathcal{F}(q^{(k)}, \theta^{(k)}) \leq \ell(\theta^{(k)}), \]

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

If the M-step is executed so that \( \theta^{(k)} \neq \theta^{(k-1)} \) iff \( \mathcal{F} \) increases, then the overall EM iteration will step to a new value of \( \theta \) iff the likelihood increases.
Variational Approximations to the EM algorithm

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

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

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

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

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

$$ q^{(k)}(\mathcal{Y}) := \arg\max_{q(\mathcal{Y}) \in \mathcal{Q}} \mathcal{F}(q(\mathcal{Y}), \theta^{(k-1)}) . $$

**M step**: unchanged

$$ \theta^{(k)} := \arg\max_{\theta} \mathcal{F}(q^{(k)}(\mathcal{Y}), \theta) = \arg\max_{\theta} \int q^{(k)}(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y} , $$

This maximises a lower bound on the log likelihood.
What do we lose?

What does restricting \( q \) to \( Q \) cost us?

- Recall that the free-energy is bounded above by Jensen:
  \[
  \mathcal{F}(q, \theta) \leq \ell(\theta^{\text{ML}})
  \]
  Thus, as long as every step increases \( \mathcal{F} \), convergence is still guaranteed.

- But, since \( P(Y|X, \theta^{(k)}) \) may not lie in \( Q \), we no longer saturate the bound after the E-step. Thus, the likelihood may not increase on each full EM step.
  \[
  \frac{\ell(\theta^{(k-1)})}{\mathcal{F}(q^{(k)}, \theta^{(k-1)})} \leq \frac{\mathcal{F}(q^{(k)}, \theta^{(k)})}{\ell(\theta^{(k)})},
  \]
  Thus, we may not converge to a maximum of \( \ell \).

The hope is that by increasing a lower bound on \( \ell \) we will find a decent solution. [Note that if \( P(Y|X, \theta^{\text{ML}}) \in Q \), then \( \theta^{\text{ML}} \) is a fixed point of the variational algorithm.]
KL divergence

Recall that

\[ F(q, \theta) = \langle \log P(X, Y|\theta) \rangle_{q(Y)} + H[q] \]
\[ = \langle \log P(X|\theta) + \log P(Y|X, \theta) \rangle_{q(Y)} - \langle \log q(Y) \rangle_{q(Y)} \]
\[ = \langle \log P(X|\theta) \rangle_{q(Y)} - \text{KL}[q||P(Y|X, \theta)]. \]

Thus,

**E step** maximise \( F(q, \theta) \) wrt the distribution over latents, given parameters:

\[ q^{(k)}(Y) := \arg\max_{q(Y) \in Q} F(q(Y), \theta^{(k-1)}). \]

is equivalent to:

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

\[ q^{(k)}(Y) := \arg\min_{q(Y) \in Q} \int q(Y) \log \frac{q(Y)}{p(Y|X, \theta^{(k-1)})} dY \]

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

This is related to ideas in *information geometry*. 
The most common form of variational approximation partitions $\mathcal{Y}$ into disjoint sets $\mathcal{Y}_i$ with

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

In this case the E-step is itself iterative:

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

$$q_i^{(k)}(\mathcal{Y}_i) := \arg\max_{q_i(\mathcal{Y}_i)} \mathcal{F}(q_i(\mathcal{Y}_i) \prod_{j \neq i} q_j(\mathcal{Y}_j), \theta^{(k-1)}).$$

The $q_i$s 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:

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

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

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

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

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

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 $Y_i = y_i$ (i.e., $q$ is factored over all variables) then the variational technique is often called a “mean field” approximation.

Suppose $P(\mathcal{X}, \mathcal{Y})$ is log-linear, e.g. the Boltzmann machine:

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

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

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

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

(where $q_i$ for $s_i \in \mathcal{X}$ 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.
Mean-field FHMM

The mean-field approach to the FHMM with

$$q(s_{1:T}^1) = \prod_{m,t} q^m_t (s^m_t)$$

yields a variant of the usual forward-backward algorithm. Coupling between the different chains only takes place through the joint output distribution. Each update depends only on the immediate neighbours.

$${q'}^m_t (s^m_t') \propto \exp \langle \log P(s_{1:T}^1, x_{1:T}) \rangle \prod_{(m',t')} q^{m'}_t (s^{m'}_t)$$

$$= \exp \left\langle \sum_m \sum_t \log P(s^m_t | s^m_{t-1}) + \sum_t \log P(x_t | s_{1:T}^1) \right\rangle \prod_{(m',t')} q^{m'}_t$$

$$\propto \exp \left[ \langle \log P(s^m_{t'} | s^m_{t'-1}) \rangle_{q^m_{t'-1}} + \langle \log P(s^m_{t+1} | s^m_{t'}) \rangle_{q^{m'}_{t+1}} + \langle \log P(x_{t'} | s_{1:T}^1) \rangle_{\prod_{m} q^{m'}_{t'}} \right]$$
Structured Variational Approximations

$q(\mathcal{Y})$ need not be completely factorized.

For example, suppose you can partition $\mathcal{Y}$ into sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ such that computing the expected sufficient statistics under $q(\mathcal{Y}_1)$ and $q(\mathcal{Y}_2)$ is tractable. Then $q(\mathcal{Y}) = q(\mathcal{Y}_1)q(\mathcal{Y}_2)$ is tractable.

If you have a graphical model, you may want to factorize $q(\mathcal{Y})$ into a product of trees, which are tractable distributions.
The most natural structured approximation in the FHMM is to factor each chain from the others

\[ q(s_{1:T}^1) = \prod_m q^m(s_{1:T}^m) \]

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

\[
q^m(s_{1:T}^m) \propto \exp \left\langle \log P(s_{1:T}^1, x_{1:T}) \right\rangle \prod_{-m'} q^{m'}(s_{1:T}^{m'}) \\
= \exp \left\langle \sum_m \sum_t \log P(s_t^m|s_{t-1}^m) + \sum_t \log P(x_t|s_{1:T}^1) \right\rangle \prod_{-m'} q^{m'} \\
\propto \exp \left[ \sum_t \log P(s_t^m'|s_{t-1}^m) + \sum_t \langle \log P(x_t'|s_{1:T}^1) \rangle \prod_{-m} q^m s_{t'}^m \right] \\
= \prod_t P(s_t^m'|s_{t-1}^m) \prod_t e^{\langle \log P(x_t'|s_{1:T}^1) \rangle \prod_{-m} q^m s_{t'}^m} 
\]
Let \( q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i) \).

Variational approximation maximises \( F \):

\[
F(q) = \int q(\mathcal{Y}) \log p(\mathcal{Y}, \mathcal{X}) d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) d\mathcal{Y}
\]

Focusing on one term, \( q_j \), we can write this as:

\[
F(q_j) = \int q_j(\mathcal{Y}_j) \langle \log p(\mathcal{Y}, \mathcal{X}) \rangle_{-q_j(\mathcal{Y}_j)} d\mathcal{Y}_j + \int q_j(\mathcal{Y}_j) \log q_j(\mathcal{Y}_j) d\mathcal{Y}_j + \text{const}
\]

Where \( \langle \cdot \rangle_{-q_j(\mathcal{Y}_j)} \) denotes averaging w.r.t. \( q_i(\mathcal{Y}_i) \) for all \( i \neq j \)

Optimum occurs when:

\[
q_j^*(\mathcal{Y}_j) = \frac{1}{Z} \exp \langle \log p(\mathcal{Y}, \mathcal{X}) \rangle_{-q_j(\mathcal{Y}_j)}
\]
Optimum occurs when:

\[ q^*_j(Y_j) = \frac{1}{Z} \exp \langle \log p(Y, X) \rangle_{-q_j(Y_j)} \]

Assume graphical model: \( p(Y, X) = \prod_i p(X_i|\text{pa}_i) \)

\[
\log q^*_j(Y_j) = \left\langle \sum_i \log p(X_i|\text{pa}_i) \right\rangle_{-q_j(Y_j)} + \text{const} \\
= \langle \log p(Y_j|\text{pa}_j) \rangle_{-q_j(Y_j)} + \sum_{k \in \text{ch}_j} \langle \log p(X_k|\text{pa}_k) \rangle_{-q_j(Y_j)} + \text{const}
\]

This defines messages that get passed between nodes in the graph. Each node receives messages from its **Markov boundary**: parents, children and parents of children.

Variational Message Passing (Winn and Bishop, 2004)
Variational Approximations to Bayesian Learning

\[
\log p(\mathcal{X}) = \log \int \int p(\mathcal{X}, \mathcal{Y}|\theta)p(\theta) \, d\mathcal{Y} \, d\theta \\
\geq \int \int q(\mathcal{Y}, \theta) \log \frac{p(\mathcal{X}, \mathcal{Y}, \theta)}{q(\mathcal{Y}, \theta)} \, d\mathcal{Y} \, d\theta
\]

Constrain \( q \in Q \) s.t. \( q(\mathcal{Y}, \theta) = q(\mathcal{Y})q(\theta) \).

This results in the \textbf{variational Bayesian EM algorithm}.

More about this later (when we study model selection).
The Other KL

Variational methods find $q = \text{argmin} \ KL[q \| p(y|x)]$:

- guaranteed convergence;
- maximising lower bound may help $\ell$ increase;
- (factored approximation) distributes for message-passing.

What about the ‘other’ KL ($q = \text{argmin} \ KL[p(y|x) \| q]$)?

Crucially, for a factored approximation the (clique) marginals are correct:

$$\arg\min_{q_i} KL \left[ P(Y|X) \bigg\| \prod q_j(Y_j|X) \right] = \arg\min_{q_i} - \int dY \ P(Y|X) \log \prod q_j(Y_j|X)$$

$$= \arg\min_{q_i} - \sum_j \int dY_j \ P(Y|X) \log q_j(Y_j|X)$$

$$= \arg\min_{q_i} - \int dY_i \ P(Y_i|X) \log q_i(Y_i|X)$$

$$= P(Y_i|X)$$

and the marginals are what we need for learning.

But (perversely) this means finding the best $q$ for this KL is intractable!
Expectation Propagation (EP)

The distribution we need to approximate is often a (normalised) product of factors:

\[ P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{1}{Z} \prod_i P(s_i|\text{pa}(s_i)) \propto \prod_{i=1}^N f_i(Y_i) \]

where the \( Y_i \) are not necessarily disjoint.

We wish to approximate this by a product of *simpler* terms:

\[ q(Y) \overset{\text{def}}{=} \prod_{i=1}^N \tilde{f}_i(Y_i) \]

\[
\begin{align*}
\min_{q(Y_i)} \text{KL} \left[ \prod_{i=1}^N f_i(Y_i) \parallel \prod_{i=1}^N \tilde{f}_i(Y_i) \right] & \quad \text{(intractable)} \\
\min_{\tilde{f}_i(Y_i)} \text{KL} \left[ f_i(Y_i) \parallel \tilde{f}_i(Y_i) \right] & \quad \text{(simple, non-iterative, inaccurate)} \\
\min_{\tilde{f}_i(Y_i)} \text{KL} \left[ f_i(Y_i) \prod_{j \neq i} \tilde{f}_j(Y_i) \parallel \tilde{f}_i(Y_i) \prod_{j \neq i} \tilde{f}_j(Y_i) \right] & \quad \text{(simple, iterative, accurate)} \leftarrow \text{EP}
\end{align*}
\]
Expectation Propagation II

Input $f_1(\mathcal{Y}_1) \ldots f_N(\mathcal{Y}_N)$
Initialize $\tilde{f}_1(\mathcal{Y}_1) = f_1(\mathcal{Y}_1)$, $\tilde{f}_i(\mathcal{Y}_i) = 1$ for $i > 1$, $q(\mathcal{Y}) = \prod_i \tilde{f}_i(\mathcal{Y}_i)$
repeat
  for $i = 1 \ldots N$ do
    Deletion: $q_{-i}(\mathcal{Y}) \leftarrow \frac{q(\mathcal{Y})}{\tilde{f}_i(\mathcal{Y}_i)} = \prod_{j \neq i} \tilde{f}_j(\mathcal{Y}_j)$
    Projection: $\tilde{f}_i^{\text{new}}(\mathcal{Y}) \leftarrow \arg\min_{f(\mathcal{Y}_i)} \text{KL}[f(\mathcal{Y}_i)q_{-i}(\mathcal{Y}) || f(\mathcal{Y}_i)q_{-i}(\mathcal{Y})]$
    Inclusion: $q(\mathcal{Y}) \leftarrow \tilde{f}_i^{\text{new}}(\mathcal{Y}_i) q_{-i}(\mathcal{Y})$
  end for
until convergence

- KL minimisation (projection) only depends on $q_{-i}(\mathcal{Y})$ marginalised to $\mathcal{Y}_i$.
- $\tilde{f}_i(\mathcal{Y})$ in exponential family $\rightarrow$ projection step is moment matching.
- Update order need not be sequential.
- Minimizes the opposite KL to variational methods.
- Loopy belief propagation and assumed density filtering are special cases.
- No convergence guarantee (although convergent forms can be developed).
- The names (deletion, projection, inclusion) are not the same as in (Minka, 2001).
EP for a NLSSM

\[
p(y_t | y_{t-1}) = \phi_t(y_t, y_{t-1})
\]
\[
p(x_t | y_t) = \psi_t(y_t)
\]

Then \( f_t(y_t, y_{t-1}) = \phi_t(y_t, y_{t-1})\psi_t(y_t) \). As \( \phi_t \) and \( \psi_t \) are non-linear, EP is not generally tractable. Assume \( \tilde{f}_t(y_t, y_{t-1}) \) is Gaussian. Then,

\[
q_{-t}(y_t, y_{t-1}) = \sum_{y_1\cdots y_{t-2}} \prod_{t' \neq t} \tilde{f}_{t'}(y_{t'}, y_{t'-1}) = \sum_{y_1\cdots y_{t-2}} \prod_{t' < t} \tilde{f}_{t'}(y_{t'}, y_{t'-1}) \sum_{y_{t+1}\cdots y_T} \prod_{t' > t} \tilde{f}_{t'}(y_{t'}, y_{t'-1})
\]

with both \( \alpha \) and \( \beta \) Gaussian.

\[
\tilde{f}_t(y_t, y_{t-1}) = \arg\min_{f \in \mathcal{N}} \text{KL}[\phi_t(y_t, y_{t-1})\psi_t(y_t)\alpha_{t-1}(y_{t-1})\beta_t(y_t) || f(y_t, y_{t-1})\alpha_{t-1}(y_{t-1})\beta_t(y_t)]
\]
Moment Matching

Recall that for exponential family $q(x) = \frac{1}{Z(\theta)} e^{S(x) \cdot \theta}$:

$$\arg\min_q \text{KL}[p(x) \| q(x)] = \arg\min_\theta \text{KL}[p(x) \| \frac{1}{Z(\theta)} e^{S(x) \cdot \theta}]$$

$$= \arg\min_\theta - \int dx\ p(x) \log \frac{1}{Z(\theta)} e^{S(x) \cdot \theta}$$

$$= \arg\min_\theta - \int dx\ p(x) S(x) \cdot \theta + \log Z(\theta)$$

$$\frac{\partial}{\partial \theta} = - \int dx\ p(x) S(x) + \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int dx\ e^{S(x) \cdot \theta}$$

$$= - \langle S(x) \rangle_p + \frac{1}{Z(\theta)} \int dx\ e^{S(x) \cdot \theta} S(x)$$

$$= - \langle S(x) \rangle_p + \langle S(x) \rangle_q$$

So minimum is found by matching sufficient stats. This is usually moment matching.

How do we calculate $\langle S(x) \rangle_p$? Low dimensional integral $\rightarrow$ Quadrature, Laplace approx . . .