Variational Approximations

10-702: Statistical Machine Learning

Zoubin Ghahramani

Machine Learning Department, CMU

April 16th, 2008
Motivation

Many statistical inference problems result in intractable computations...

- Bayesian posterior over model parameters:

\[
P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}
\]

- Computing posterior over hidden variables (e.g. for E step of EM):

\[
P(H|V, \theta) = \frac{P(V|H, \theta)P(H|\theta)}{P(V|\theta)}
\]

- Computing marginals in a multiply-connected graphical models:

\[
P(x_i|x_j = e) = \sum_{x \setminus \{x_i, x_j\}} P(x|x_j = e)
\]

**Solutions:** Markov chain Monte Carlo, variational approximations
Example: Binary latent factor model

Model with $K$ binary latent variables, $s_i \in \{0, 1\}$, organised into a vector $s = (s_1, \ldots, s_K)$ real-valued observation vector $y$
parameters $\theta = \{[\mu_i, \pi_i]_{i=1}^K, \sigma^2\}$

$s \sim$ Bernoulli
$y|s \sim$ Gaussian

$$p(s|\pi) = p(s_1, \ldots, s_K|\pi) = \prod_{i=1}^K p(s_i|\pi_i) = \prod_{i=1}^K \pi_i^{s_i} (1 - \pi_i)^{(1-s_i)}$$

$$p(y|s_1, \ldots, s_K, \mu, \sigma^2) = \mathcal{N} \left( \sum_{i=1}^K s_i \mu_i, \sigma^2 I \right)$$

EM optimizes bound on likelihood:

$$\mathcal{F}(q, \theta) = \langle \log p(s, y|\theta) \rangle_{q(s)} - \langle \log q(s) \rangle_{q(s)}$$

where $\langle \rangle_q$ is expectation under $q$:

$$\langle f(s) \rangle_q \overset{\text{def}}{=} \sum_s f(s)q(s)$$

**Exact E step:** $q(s) = p(s|y, \theta)$ distribution over $2^K$ states: intractable for large $K$
Example: Binary latent factor model

Model with $K$ binary latent variables, $s_i \in \{0, 1\}$, organised into a vector $s = (s_1, \ldots, s_K)$

real-valued observation vector $y$

parameters $\theta = \{\{\mu_i, \pi_i\}_{i=1}^K, \sigma^2\}$

$s \sim \text{Bernoulli}$

$y|s \sim \text{Gaussian}$

Figure 2: **Left panel**: Original source images used to generate data. **Middle panel**: Observed images resulting from mixture of sources. **Right panel**: Recovered sources from Lu et al (2004)
Given a set of observed (visible) variables $V$, a set of unobserved (hidden / latent / missing) variables $H$, and model parameters $\theta$, optimize the log likelihood:

$$\mathcal{L}(\theta) = \log p(V|\theta) = \log \int p(H, V|\theta) dH,$$

Using Jensen’s inequality, for any distribution of hidden variables $q(H)$ we have:

$$\mathcal{L}(\theta) = \log \int q(H) \frac{p(H, V|\theta)}{q(H)} dH \geq \int q(H) \log \frac{p(H, V|\theta)}{q(H)} dH = \mathcal{F}(q, \theta),$$

defining the $\mathcal{F}(q, \theta)$ functional, which is a lower bound on the log likelihood.

In the EM algorithm, we alternately optimize $\mathcal{F}(q, \theta)$ wrt $q$ and $\theta$, and we can prove that this will never decrease $\mathcal{L}$. 
The E and M steps of EM

The lower bound on the log likelihood:

$$F(q, \theta) = \int q(H) \log \frac{p(H, V|\theta)}{q(H)} dH = \int q(H) \log p(H, V|\theta) dH + \mathcal{H}(q),$$

where $\mathcal{H}(q) = -\int q(H) \log q(H) dH$ is the entropy of $q$. We iteratively alternate:

**E step:** maximize $F(q, \theta)$ wrt the distribution over hidden variables given the parameters:

$$q^{[k]}(H) := \arg\max_{q(H)} F(q(H), \theta^{[k-1]}) = p(H|V, \theta^{[k-1]}).$$

**M step:** maximize $F(q, \theta)$ wrt the parameters given the hidden distribution:

$$\theta^{[k]} := \arg\max_{\theta} F(q^{[k]}(H), \theta) = \arg\max_{\theta} \int q^{[k]}(H) \log p(H, V|\theta) dH,$$

which is equivalent to optimizing the expected complete log likelihood $\log p(H, V|\theta)$, since the entropy of $q(H)$ does not depend on $\theta$. 
Often $p(H|V, \theta)$ is computationally intractable, so an exact E step is out of the question.

**Assume some simpler form for $q(H)$**, e.g. $q \in \mathcal{Q}$, the set of fully-factorized distributions over the hidden variables: $q(H) = \prod_i q(H_i)$

**E step** (approximate): maximize $\mathcal{F}(q, \theta)$ wrt the distribution over hidden variables given the parameters:

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

**M step**: maximize $\mathcal{F}(q, \theta)$ wrt the parameters given the hidden distribution:

$$\theta^{[k]} := \arg\max_{\theta} \mathcal{F}(q^{[k]}(H), \theta) = \arg\max_{\theta} \int q^{[k]}(H) \log p(H, V|\theta) dH,$$

This maximizes a lower bound on the log likelihood.

Using the fully-factorized $q$ is sometimes called a **mean-field approximation**.
Example: Binary latent factor model

Model with $K$ binary latent variables, $s_i \in \{0, 1\}$, organised into a vector $s = (s_1, \ldots, s_K)$.

Real-valued observation vector $y$.

Parameters $\theta = \{\{\mu_i, \pi_i\}_{i=1}^K, \sigma^2\}$.

$s \sim \text{Bernoulli}$

$y|s \sim \text{Gaussian}$

$$p(s|\pi) = p(s_1, \ldots, s_K|\pi) = \prod_{i=1}^K p(s_i|\pi_i) = \prod_{i=1}^K \pi_i^{s_i}(1-\pi_i)^{1-s_i}$$

$$p(y|s_1, \ldots, s_K, \mu, \sigma^2) = \mathcal{N}\left(\sum_{i=1}^K s_i \mu_i, \sigma^2 I\right)$$

EM optimizes bound on likelihood:

$$\mathcal{F}(q, \theta) = \langle \log p(s, y|\theta) \rangle_q(s) - \langle \log q(s) \rangle_q(s)$$

where $\langle \rangle_q$ is expectation under $q$:

$$\langle f(s) \rangle_q \overset{\text{def}}{=} \sum_s f(s) q(s)$$

**Exact E step:** $q(s) = p(s|y, \theta)$ distribution over $2^K$ states: intractable for large $K$.
Example: Binary latent factors model (cont.)

\[ \mathcal{F}(\mathbf{q}, \theta) = \langle \log p(\mathbf{s}, \mathbf{y}|\theta) \rangle_{q(\mathbf{s})} - \langle \log q(\mathbf{s}) \rangle_{q(\mathbf{s})} \]

\[
\log p(\mathbf{s}, \mathbf{y}|\theta) + c \\
= \sum_{i=1}^{K} s_i \log \pi_i + (1 - s_i) \log (1 - \pi_i) - D \log \sigma - \frac{1}{2\sigma^2} (\mathbf{y} - \sum_i s_i \mu_i)^\top (\mathbf{y} - \sum_i s_i \mu_i) \\
= \sum_{i=1}^{K} s_i \log \pi_i + (1 - s_i) \log (1 - \pi_i) - D \log \sigma \\
- \frac{1}{2\sigma^2} \left( \mathbf{y}^\top \mathbf{y} - 2 \sum_i s_i \mu_i^\top \mathbf{y} + \sum_i \sum_j s_i s_j \mu_i^\top \mu_j \right) \\
\]

we therefore need \( \langle s_i \rangle \) and \( \langle s_i s_j \rangle \) to compute \( \mathcal{F} \).

These are the expected *sufficient statistics* of the hidden variables.
Example: Binary latent factors model (cont.)

Variational approximation:

\[
q(s) = \prod_i q_i(s_i) = \prod_{i=1}^{K} \lambda_i^{s_i} (1 - \lambda_i)^{(1-s_i)}
\]

where \(\lambda_i\) is a parameter of the variational approximation modelling the posterior mean of \(s_i\) (compare to \(\pi_i\) which models the prior mean of \(s_i\)).

Under this approximation we know \(\langle s_i \rangle = \lambda_i\) and \(\langle s_i s_j \rangle = \lambda_i \lambda_j + \delta_{ij} (\lambda_i - \lambda_i^2)\).

\[
\mathcal{F}(\lambda, \theta) = \sum_i \lambda_i \log \frac{\pi_i}{\lambda_i} + (1 - \lambda_i) \log \frac{1 - \pi_i}{(1 - \lambda_i)}
\]

\[
- D \log \sigma - \frac{1}{2\sigma^2} (y - \sum_i \lambda_i \mu_i)^\top (y - \sum_i \lambda_i \mu_i)
\]

\[
- \frac{1}{2\sigma^2} \sum_i (\lambda_i - \lambda_i^2) \mu_i \mu_i^\top - \frac{D}{2} \log(2\pi)
\]
Fixed point equations for the binary latent factors model

Taking derivatives w.r.t. $\lambda_i$:

$$\frac{\partial F}{\partial \lambda_i} = \log \frac{\pi_i}{1 - \pi_i} - \log \frac{\lambda_i}{1 - \lambda_i} + \frac{1}{\sigma^2} (y - \sum_{j \neq i} \lambda_j \mu_j)^\top \mu_i - \frac{1}{2\sigma^2} \mu_i^\top \mu_i$$

Setting to zero we get fixed point equations:

$$\lambda_i = f \left( \log \frac{\pi_i}{1 - \pi_i} + \frac{1}{\sigma^2} (y - \sum_{j \neq i} \lambda_j \mu_j)^\top \mu_i - \frac{1}{2\sigma^2} \mu_i^\top \mu_i \right)$$

where $f(x) = 1/(1 + \exp(-x))$ is the logistic (sigmoid) function.

**Learning algorithm:**

**E step:** run fixed point equations until convergence of $\lambda$ for each data point.

**M step:** re-estimate $\theta$ given $\lambda$s.
KL divergence

Note that

**E step** maximize $\mathcal{F}(q, \theta)$ wrt the distribution over hidden variables, given the parameters:

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

is equivalent to:

**E step** minimize $\mathcal{KL}(q \parallel p(H|V, \theta))$ wrt the distribution over hidden variables, given the parameters:

$$q^{[k]}(H) := \arg\min_{q(H) \in \mathcal{Q}} \int q(H) \log \frac{q(H)}{p(H|V, \theta^{[k-1]})} dH$$

So, in each E step, the algorithm tries to find the best approximation to $p$ in $\mathcal{Q}$. This is related to ideas in *information geometry*.
\[
\log p(V) = \log \int \int p(V, H|\theta)p(\theta) \, dH \, d\theta \\
\geq \int \int q(H, \theta) \log \frac{p(V, H, \theta)}{q(H, \theta)} \, dH \, d\theta
\]

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

This results in the variational Bayesian EM algorithm.
Goal:

- Infer number of clusters
- Infer intrinsic dimensionality of each cluster

Under the assumption that each cluster is Gaussian
Mixture of Factor Analysers

**True data**: 6 Gaussian clusters with dimensions: (1 7 4 3 2 2) embedded in 10-D

**Inferred structure:**

<table>
<thead>
<tr>
<th>number of points per cluster</th>
<th>intrinsic dimensionalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1 7 4 3 2 2</td>
</tr>
<tr>
<td>8</td>
<td>1 1 2</td>
</tr>
<tr>
<td>16</td>
<td>1 1 4 2</td>
</tr>
<tr>
<td>32</td>
<td>1 6 3 3 2 2</td>
</tr>
<tr>
<td>64</td>
<td>1 7 4 3 2 2</td>
</tr>
<tr>
<td>128</td>
<td>1 7 4 3 2 2</td>
</tr>
</tbody>
</table>

- Finds the clusters and dimensionalities efficiently.
- The model complexity reduces in line with the lack of data support.

demos: run_simple and ueda_demo
Discrete hidden states, $s_t$.
Observations $y_t$.

How many hidden states?
What structure state-transition matrix?

demo: vbhmm_demo
Let $q(H) = \prod_i q_i(H_i)$.  

Variational approximation maximises $\mathcal{F}$:  
\[
\mathcal{F}(q) = \int q(H) \log p(H, V) dH - \int q(H) \log q(H) dH
\]

Focusing on one term, $q_j$, we can write this as:  
\[
\mathcal{F}(q_j) = \int q_j(H_j) \langle \log p(H, V) \rangle_{q_j(H_j)} dH_j + \int q_j(H_j) \log q_j(H_j) dH_j + \text{const}
\]

Where $\langle \cdot \rangle_{q_j(H_j)}$ denotes averaging w.r.t. $q_i(H_i)$ for all $i \neq j$  

Optimum occurs when:  
\[
q_j^*(H_j) = \frac{1}{Z} \exp \langle \log p(H, V) \rangle_{q_j(H_j)}
\]
Optimum occurs when:

\[ q_j^*(H_j) = \frac{1}{Z} \exp \langle \log p(H, V) \rangle_{\sim q_j(H_j)} \]

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

\[
\log q_j^*(H_j) = \left\langle \sum_i \log p(X_i|\text{pa}_i) \right\rangle_{\sim q_j(H_j)} + \text{const}
\]

\[
= \left\langle \log p(H_j|\text{pa}_j) \right\rangle_{\sim q_j(H_j)} + \sum_{k \in \text{ch}_j} \left\langle \log p(X_k|\text{pa}_k) \right\rangle_{\sim q_j(H_j)} + \text{const}
\]

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)
Expectation Propagation (EP)

Data (iid) $\mathcal{D} = \{x^{(1)} \ldots, x^{(N)}\}$, model $p(x|\theta)$, with parameter prior $p(\theta)$.

The parameter posterior is:

$$p(\theta|\mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\theta) \prod_{i=1}^{N} p(x^{(i)}|\theta)$$

We can write this as product of factors over $\theta$:

$$p(\theta) \prod_{i=1}^{N} p(x^{(i)}|\theta) = \prod_{i=0}^{N} f_{i}(\theta)$$

where $f_{0}(\theta) \overset{\text{def}}{=} p(\theta)$ and $f_{i}(\theta) \overset{\text{def}}{=} p(x^{(i)}|\theta)$ and we will ignore the constants.

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

$$q(\theta) \overset{\text{def}}{=} \prod_{i=0}^{N} \tilde{f}_{i}(\theta)$$

$$\min_{q(\theta)} \mathcal{KL} \left( \prod_{i=0}^{N} f_{i}(\theta) \left\| \prod_{i=0}^{N} \tilde{f}_{i}(\theta) \right\| \right)$$  \hspace{1cm} \text{(intractable)}

$$\min_{\tilde{f}_{i}(\theta)} \mathcal{KL} \left( f_{i}(\theta) \left\| \tilde{f}_{i}(\theta) \right\| \right)$$  \hspace{1cm} \text{(simple, non-iterative, inaccurate)}

$$\min_{\tilde{f}_{i}(\theta)} \mathcal{KL} \left( f_{i}(\theta) \prod_{j \neq i} \tilde{f}_{j}(\theta) \left\| \tilde{f}_{i}(\theta) \prod_{j \neq i} \tilde{f}_{j}(\theta) \right\| \right)$$  \hspace{1cm} \text{(simple, iterative, accurate) ← EP}
**Expectation Propagation II**

Input $f_0(\theta) \ldots f_N(\theta)$

Initialize $\tilde{f}_0(\theta) = f_0(\theta)$, $\tilde{f}_i(\theta) = 1$ for $i > 0$, $q(\theta) = \prod_i \tilde{f}_i(\theta)$

repeat
  for $i = 0 \ldots N$ do
    Deletion: $q_{\neg i}(\theta) \leftarrow q(\theta) \prod_{j \neq i} \tilde{f}_j(\theta)$
    Projection: $\tilde{f}_i^{\text{new}}(\theta) \leftarrow \arg\min_{f(\theta)} \mathcal{KL}(f_i(\theta) q_{\neg i}(\theta) \| f(\theta) q_{\neg i}(\theta))$
    Inclusion: $q(\theta) \leftarrow \tilde{f}_i^{\text{new}}(\theta) q_{\neg i}(\theta)$
  end for
until convergence

**The EP algorithm.** Some variations are possible: here we assumed that $f_0$ is in the exponential family, and we updated sequentially over $i$.

- Tries to minimize the opposite KL to variational methods
- $\tilde{f}_i(\theta)$ in exponential family $\rightarrow$ projection step is moment matching
- No convergence guarantee (although convergent forms can be developed)
Some Further Readings

Appendix: The binary latent factors model for an i.i.d. data set

Assume data set \( \mathcal{D} = \{y^{(1)} \ldots , y^{(N)}\} \) of \( N \) points and params \( \theta = \{[\mu_i, \pi_i]_{i=1}^K, \sigma^2 \} \)

Use a factorised distribution:
\[
q(s) = \prod_{n=1}^N q_n(s^{(n)}) = \prod_{n=1}^N \prod_{i=1}^K q_n(s_i^{(n)}) = \prod_{n} \prod_{i} (\lambda_i^{(n)})^{s_i^{(n)}} (1 - \lambda_i^{(n)})^{1-s_i^{(n)}}
\]

\[
p(\mathcal{D}|\theta) = \prod_{n=1}^N p(y^{(n)}|\theta)
\]

\[
p(y^{(n)}|\theta) = \sum_{s} p(y^{(n)}|s, \mu, \sigma)p(s|\pi)
\]

\[
\mathcal{F}(q(s), \theta) = \sum_n \mathcal{F}_n(q_n(s^{(n)}), \theta) \leq \log p(\mathcal{D}|\theta)
\]

\[
\mathcal{F}_n(q_n(s^{(n)}), \theta) = \langle \log p(s^{(n)}, y^{(n)}|\theta) \rangle_{q_n(s^{(n)})} - \langle \log q_n(s^{(n)}) \rangle_{q_n(s^{(n)})}
\]

We need to optimise w.r.t. \( q_n(s^{(n)}) \) for each data point, so

**E step:** optimize \( q_n(s^{(n)}) \) (i.e. \( \lambda^{(n)} \)) for each \( n \).

**M step:** re-estimate \( \theta \) given \( q_n(s^{(n)}) \)'s.
Appendix: How tight is the lower bound?

It is hard to compute a nontrivial general upper bound.

To determine how tight the bound is, one can approximate the true likelihood by a variety of other methods.

One approach is to use the variational approximation as a proposal distribution for importance sampling.

But this will generally not work well. See exercise 33.6 in David MacKay’s textbook.