Unsupervised Learning

Variational Approximations

Zoubin Ghahramani
zoubin@gatsby.ucl.ac.uk

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

Autumn 2003

Review: The EM algorithm

Given a set of observed (visible) variables V, a set of unobserved (hidden / latent / missing) variables H, and model parameters θ, 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 θ, 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:

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

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

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

$$q^{[k]}(H) := \underset{q(H)}{\operatorname{argmax}} \; \mathcal{F}\big(q(H), \theta^{[k-1]}\big).$$

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

$$\theta^{[k]} := \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}\big(q^{[k]}(H), \theta\big) = \underset{\theta}{\operatorname{argmax}} \; \int q^{[k]}(H) \log p(H,V|\theta) dH,$$

which is equivalent to optimizing the expected complete-data likelihood $p(H,V|\theta)$, since the entropy of $q(H)$ does not depend on θ.

Variational Approximations to the EM algorithm

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:

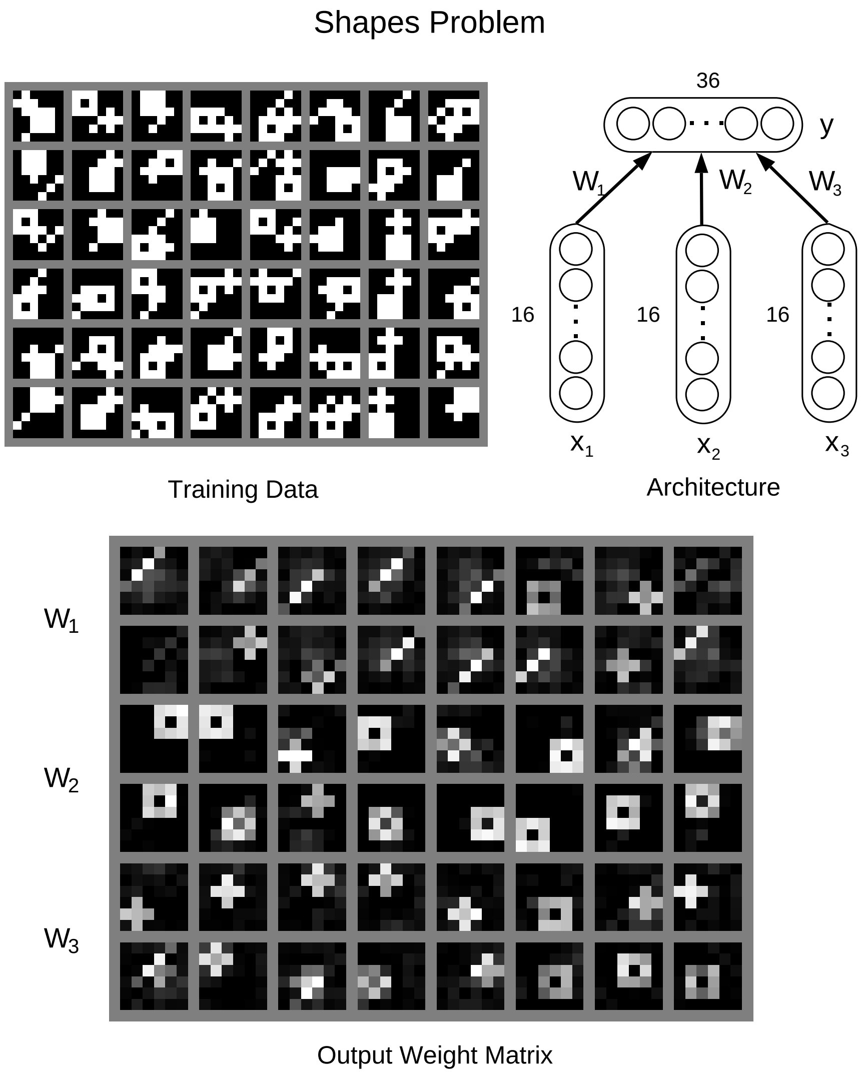
$$q^{[k]}(H) := \underset{q(H)\in\mathcal{Q}}{\operatorname{argmax}} \; \mathcal{F}\big(q(H), \theta^{[k-1]}\big).$$

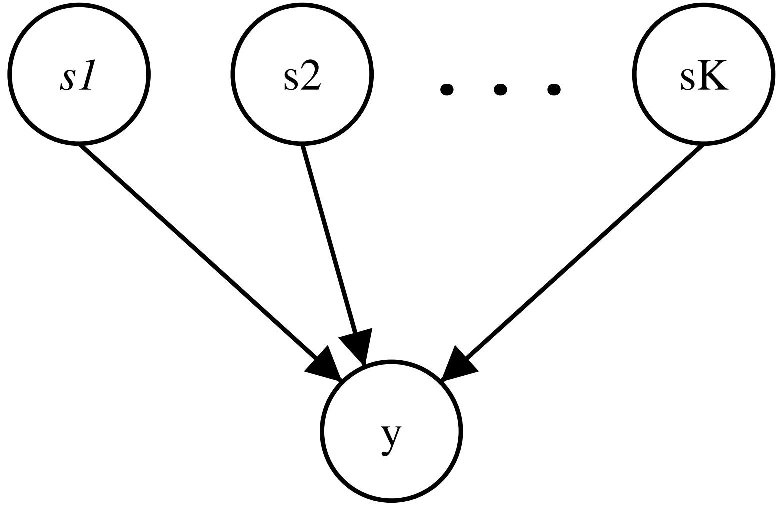
M step : maximize $\mathcal{F}(q,\theta)$ wrt the parameters given the hidden distribution:

$$\theta^{[k]} := \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}\big(q^{[k]}(H), \theta\big) = \underset{\theta}{\operatorname{argmax}} \; \int q^{[k]}(H) \log p(H, V|\theta) dH,$$

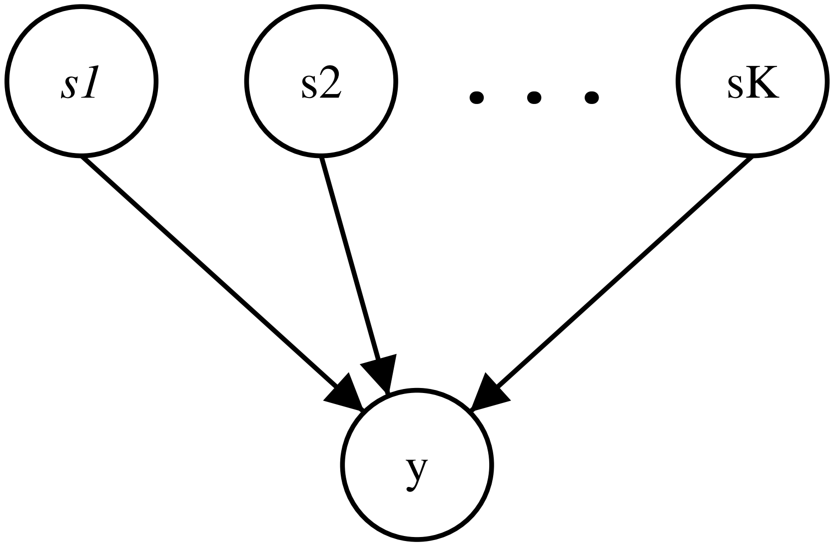
This maximizes a lower bound on the log likelihood.
Using the fully-factorized form of q is sometimes called a **mean-field approximation**.

Example: A binary latent factors model





Example: Binary latent factors model



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

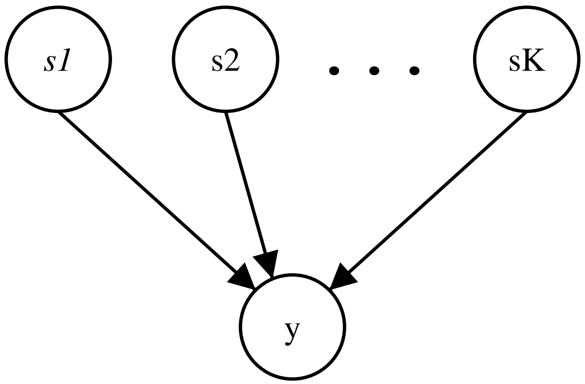
$$p(\mathbf{s}|\boldsymbol{\pi}) = p(s_1, \ldots, s_K|\boldsymbol{\pi}) = \prod_{i=1}^K p(s_i|\pi_i) = \prod_{i=1}^K \pi_i^{s_i}(1-\pi_i)^{(1-s_i)}$$

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

EM optimizes lower bound on likelihood: $\mathcal{F}(q, \boldsymbol{\theta}) = \langle \log p(\mathbf{s}, \mathbf{y}|\boldsymbol{\theta}) \rangle_{q(\mathbf{s})} - \langle \log q(\mathbf{s}) \rangle_{q(\mathbf{s})}$
where $\langle\rangle_q$ is defined expectation under q: $\langle f(\mathbf{s}) \rangle_q \equiv \sum_{\mathbf{s}} f(\mathbf{s}) q(\mathbf{s})$

Exact E step: $q(\mathbf{s}) = p(\mathbf{s}|\mathbf{y}, \boldsymbol{\theta})$ is a distribution over 2^K states, **intractable** for large K

Example: Binary latent factors model (cont)



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

$$\begin{aligned}
\log \quad & p(\mathbf{s}, \mathbf{y}|\boldsymbol{\theta}) + c \\
= \quad & \textstyle\sum_{i=1}^{K} s_i \log \pi_i \quad + (1 - s_i) \log(1 - \pi_i) - D \log \sigma - \frac{1}{2\sigma^2} (\mathbf{y} - \sum_i s_i \boldsymbol{\mu}_i)^\top (\mathbf{y} - \sum_i s_i \boldsymbol{\mu}_i) \\
= \quad & \textstyle\sum_{i=1}^{K} s_i \log \pi_i \quad + (1 - s_i) \log(1 - \pi_i) - D \log \sigma \\
& - \frac{1}{2\sigma^2} \left(\mathbf{y}^\top \mathbf{y} - 2 \sum_i s_i \boldsymbol{\mu}_i^\top \mathbf{y} + \sum_i \sum_j s_i s_j \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_j \right)
\end{aligned}$$

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:

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

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

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = & \sum_i \lambda_i \log \frac{\pi_i}{\lambda_i} + (1-\lambda_i) \log \frac{(1-\pi_i)}{(1-\lambda_i)} \\
& - D \log \sigma - \frac{1}{2\sigma^2}(\mathbf{y} - \sum_i \lambda_i \boldsymbol{\mu}_i)^\top (\mathbf{y} - \sum_i \lambda_i \boldsymbol{\mu}_i) \\
& - \frac{1}{2\sigma^2} \sum_i (\lambda_i - \lambda_i^2) {\boldsymbol{\mu}_i}^\top \boldsymbol{\mu}_i - \frac{D}{2} \log(2\pi)
\end{aligned}$$

Fixed point equations for the binary latent factors model

Taking derivatives w.r.t. λ_i:

$$\frac{\partial \mathcal{F}}{\partial \lambda_i} = \log \frac{\pi_i}{1-\pi_i} - \log \frac{\lambda_i}{1-\lambda_i} + \frac{1}{\sigma^2}(\mathbf{y} - \sum_{j \neq i} \lambda_j \boldsymbol{\mu}_j)^\top \boldsymbol{\mu}_i - \frac{1}{2\sigma^2} \boldsymbol{\mu}_i{}^\top \boldsymbol{\mu}_i$$

Setting to zero we get fixed point equations:

$$\lambda_i = f\left(\log \frac{\pi_i}{1-\pi_i} + \frac{1}{\sigma^2}(\mathbf{y} - \sum_{j \neq i} \lambda_j \boldsymbol{\mu}_j)^\top \boldsymbol{\mu}_i - \frac{1}{2\sigma^2} \boldsymbol{\mu}_i{}^\top \boldsymbol{\mu}_i \right)$$

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

Learning algorithm:

E step: run fixed point equations until convergence of $\boldsymbol{\lambda}$ for each data point.
M step: re-estimate $\boldsymbol{\theta}$ given $\boldsymbol{\lambda}$s.

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

Assume a data set $\mathcal{D} = \{\mathbf{y}^{(1)} \ldots, \mathbf{y}^{(N)}\}$ of N points. Parameters $\boldsymbol{\theta} = \{\{\boldsymbol{\mu}_i, \pi_i\}_{i=1}^{K}, \sigma^2\}$
Use a factorised distribution: $q(\mathbf{s}) = \prod_{n=1}^{N} q_n(\mathbf{s}^{(n)}) = \prod_{n=1}^{N} \prod_{i=1}^{K} q_n(s_i^{(n)})$

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{n=1}^{N} p(\mathbf{y}^{(n)}|\boldsymbol{\theta}) \\
p(\mathbf{y}^{(n)}|\boldsymbol{\theta}) &= \sum_{\mathbf{s}} p(\mathbf{y}^{(n)}|\mathbf{s}, \boldsymbol{\mu}, \sigma) p(\mathbf{s}|\boldsymbol{\pi}) \\
\mathcal{F}(q(\mathbf{s}), \boldsymbol{\theta}) &= \sum_{n} \mathcal{F}_n(q_n(\mathbf{s}^{(n)}), \boldsymbol{\theta}) \leq \log p(\mathcal{D}|\boldsymbol{\theta}) \\
\mathcal{F}_n(q_n(\mathbf{s}^{(n)}), \boldsymbol{\theta}) &= \left\langle \log p(\mathbf{s}^{(n)}, \mathbf{y}^{(n)}|\boldsymbol{\theta}) \right\rangle_{q_n(\mathbf{s}^{(n)})} - \left\langle \log q_n(\mathbf{s}^{(n)}) \right\rangle_{q_n(\mathbf{s}^{(n)})}
\end{aligned}
$$

We need to optimise w.r.t. the distribution over latent variables for *each data point*, so

E step: optimize $q_n(\mathbf{s}^{(n)})$ (i.e. $\boldsymbol{\lambda}^{(n)}$) for each n.
M step: re-estimate $\boldsymbol{\theta}$ given $q_n(\mathbf{s}^{(n)})$'s.

KL divergence

Note that

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

$$q^{[k]}(H) := \underset{q(H)\in\mathcal{Q}}{\operatorname{argmax}}\ \mathcal{F}\big(q(H), \theta^{[k-1]}\big).$$

is equivalent to:

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

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

So, in each E step, the algorithm is trying to find the best approximation to p in $\mathcal{Q}$.

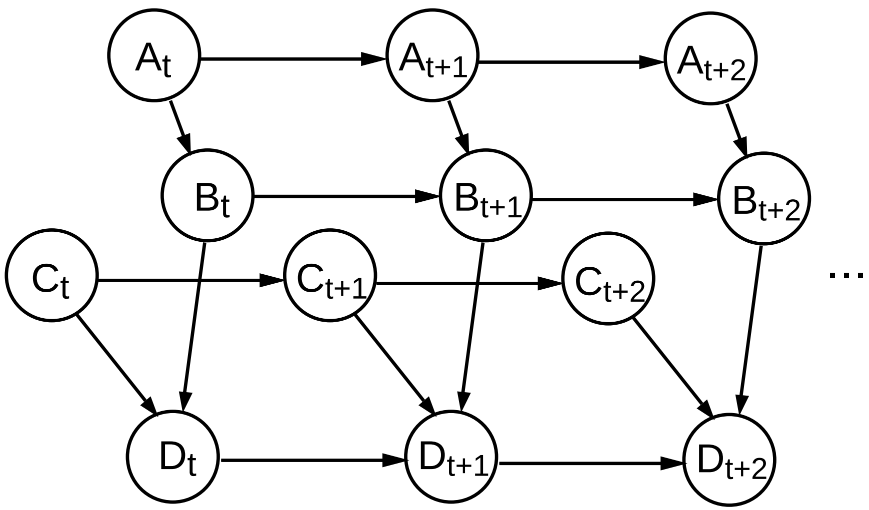
This is related to ideas in *information geometry*.

Structured Variational Approximations

$q(H)$ need not be completely factorized.

For example, suppose you can partition H into sets H_1 and H_2 such that computing the expected sufficient statistics under $q(H_1)$ and $q(H_2)$ is tractable.
Then $q(H) = q(H_1)q(H_2)$ is tractable.

If you have a graphical model, you may want to factorize $q(H)$ into a product of trees, which are tractable distributions.



Variational Approximations to Bayesian Learning

$$
\begin{aligned}
\log p(V) &= \log \int \int p(V, H|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \; dH \; d\boldsymbol{\theta} \\
&\geq \int \int q(H, \boldsymbol{\theta}) \log \frac{p(V, H, \boldsymbol{\theta})}{q(H, \boldsymbol{\theta})} \; dH \; d\boldsymbol{\theta}
\end{aligned}
$$

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

This results in the **variational Bayesian EM algorithm**.

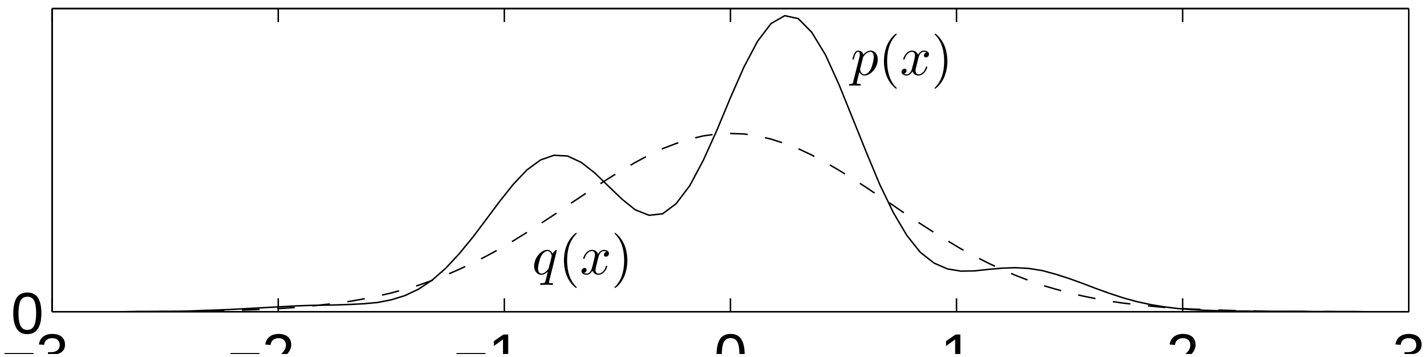
More about this later (when we study model selection).

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 as a proposal distribution for **importance sampling**.



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

Readings

- MacKay, D.J.C. (2003) Information Theory, Inference, and Learning Algorithms. Chapter 33.

- Ghahramani, Z. (1995) Factorial learning and the EM algorithm. In Adv Neur Info Proc Syst 7. Available at: www.gatsby.ucl.ac.uk/∼zoubin/

- Ghahramani, Z. and Beal, M.J. (2000) Graphical models and variational methods. In Saad & Opper (eds) Advanced Mean Field Method—Theory and Practice. MIT Press. Available at: www.gatsby.ucl.ac.uk/∼zoubin/papers/advmf.ps.gz

- Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. and Saul, L.K. (1999) An Introduction to Variational Methods for Graphical Models. Machine Learning 37:183-233. Available at: www.gatsby.ucl.ac.uk/∼zoubin/papers/varintro.ps.gz