Variational Bayes

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/atsml/
The main idea of variational Bayes is to turn posterior inference in intractable Bayesian models into optimization.

The key quantity is ELBO

$$\mathcal{F}(q) = \mathbb{E}_q [\log p(X, z, \theta)] + H(q)$$

which is a lower bound on log-evidence $\log p(X)$.

It equals log-evidence iff $q(z, \theta) = p(z, \theta|X)$. 
Variational families

VB minimises the divergence $\text{KL} \left( q(z, \theta) \| p(z, \theta | X) \right)$ over some variational family $Q$ or, equivalently, maximises the ELBO, i.e., finds the tightest lower bound on the log-evidence.

If $Q$ consists of variational distributions which factorise across the latents and the parameters: $q(z, \theta) = q_Z(z) q_\Theta(\theta)$, we obtain the alternating Bayesian EM updates

$$q_Z(z) \propto \exp \left( \int \log p(X, z, \theta) q_\Theta(\theta) \ d\theta \right),$$

$$q_\Theta(\theta) \propto \exp \left( \int \log p(X, z, \theta) q_Z(z) \ dz \right).$$

The distinction between parameters $\theta$ and latent variables $z$ disappears in Bayesian modelling, so we will drop $\theta$ from the notation and collect all unobserved quantities into $z$. 
In **mean-field variational family** $\mathcal{Q}$, variational distribution fully factorizes:

$$q(z) = \prod_{j=1}^{m} q_j(z_j),$$

Unable to capture posterior correlations between the latent variables $z_j$ and $z_{j'}$ for $j \neq j'$; the best we can hope for is a rich representations of the posterior marginals.
Doing sequential updates for each individual factor $z_j$, we obtain **Coordinate Ascent Variational Inference (CAVI)** algorithm

**Input**: a model $p(z, x)$, dataset $x$

**Output**: a variational posterior $q(z)$

while the ELBO has not converged do

- for $j = 1, \ldots, m$
  - $q_j(z_j) \propto \exp \left[ \mathbb{E}_{z_{-j} \sim q} \log p(z_j | z_{-j}, x) \right]$
  - $\text{ELBO}(q) = \mathbb{E}_{z \sim q} \left[ \log p(x, z) \right] + H(q)$

return $q(z) = \prod_{j=1}^{m} q_j(z_j)$
CAVI in exponential families

When the complete conditionals \( p(z_j|z_{-j}, x) \) belong to an exponential family

\[
p(z_j|z_{-j}, x) = h(z_j) \exp \left[ \eta_j (z_{-j}, x)^\top z_j - A(\eta_j (z_{-j}, x)) \right],
\]

\( q_j \) belongs to the same family and CAVI simplifies to updating natural parameters

\[
q_j(z_j) \propto \exp \left[ \mathbb{E}_{-j} \log p(z_j|z_{-j}, x) \right]
= \exp \left[ \log h(z_j) + \left\{ \mathbb{E}_{-j} \eta_j (z_{-j}, x) \right\}^\top z_j - \mathbb{E}_{-j} A(\eta_j (z_{-j}, x)) \right]
\propto h(z_j) \exp \left[ \left\{ \mathbb{E}_{-j} \eta_j (z_{-j}, x) \right\}^\top z_j \right]
\]
Latent Dirichlet Allocation

Used for topic modelling in a collection of documents: each text document typically blends multiple topics.
- each document is a probability distribution over topics
- each topic is a probability distribution over words

Goal is to find the posterior

\[ p(\text{topics, proportions, assignments} | \text{observed words}) \]
Latent Dirichlet Allocation

$D$: the number of documents, $K$: the number of topics, $V$: the size of the vocabulary.

1. For each topic in $k = 1, \ldots, K$,
   1. Draw a distribution over $V$ words $\beta_k \sim \text{Dir}_V(\eta)$
2. For each document in $d = 1, \ldots, D$,
   1. Draw a vector of topic proportions $\theta_d \sim \text{Dir}_K(\alpha)$
   2. For each word in $n = 1, \ldots, N_d$,
      1. Draw a topic assignment $z_{dn} \sim \text{Discrete}(\theta_d)$, i.e. $p(z_{dn} = k | \theta_d) = \theta_{dk}$
      2. Draw a word $w_{dn} \sim \text{Discrete}(\beta_{z_{dn}})$, i.e. $p(w_{dn} = v | \beta, z) = \beta_{z_{dn} v}$

Figure: Graphical model representation of LDA. Plates represent replication, for example there are $D$ documents each having a topic proportion vector $\theta_d$
Latent Dirichlet Allocation

Mean-field family:

\[ q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k; \lambda_k) \prod_{d=1}^{D} \left\{ q(\theta_d; \gamma_d) \prod_{n=1}^{N_d} q(z_{dn}; \phi_{dn}) \right\}. \]

1. Complete conditional on the topic assignment is a multinomial

\[ p(z_{dn} = k|\theta_d, \beta, w_d) \propto \theta_d^k \beta_{k,w_{dn}} = \exp(\log \theta_d^k + \log \beta_{k,w_{dn}}). \]

2. Complete conditional on the topic proportions is a Dirichlet

\[ p(\theta_d|z_d) = \text{Dir}_K \left( \theta_d; \alpha + \sum_{n=1}^{N_d} z_{dn} [\cdot] \right). \]

3. Complete conditional on the topics is another Dirichlet

\[ p(\beta_k|z, w) = \text{Dir}_V \left( \beta_k; \eta + \sum_{d=1}^{D} \sum_{n=1}^{N_d} z_{dn} [k] w_{dn} [\cdot] \right). \]