SC4/SM8 Advanced Topics in Statistical Machine Learning
Chapter 5/6: Variational Methods

Yee Whye Teh
Department of Statistics
Oxford

https://github.com/ywteh/advml2020

ELBO

The main idea of variational Bayes is to turn posterior inference in intractable Bayesian models into optimization.

The key quantity is ELBO (Evidence Lower BOund):

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

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

It equals log-evidence iff $q(\mathbf{z}, \theta) = p(\mathbf{z}, \theta | \mathbf{X})$.

Variational families

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

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

$$q_{\mathbf{Z}}(\mathbf{z}) \propto \exp\left(\int \log p(\mathbf{X}, \mathbf{z}, \theta) q_{\Theta}(\theta)\, d\theta\right),$$

$$q_{\Theta}(\theta) \propto \exp\left(\int \log p(\mathbf{X}, \mathbf{z}, \theta) q_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}\right).$$

The distinction between parameters θ and latent variables $\mathbf{z}$ disappears in Bayesian modelling, so we will drop θ from the notation and collect all unobserved quantities into $\mathbf{z}$.

Mean-field variational family

In **mean-field variational family** $\mathcal{Q}$, variational distribution fully factorizes

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

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.

CAVI

Doing sequential updates for each individual factor z_j, we obtain **Coordinate Ascent Variational Inference (CAVI)** algorithm

Input: a model $p(\mathbf{z}, \mathbf{x})$, dataset $\mathbf{x}$
Output: a variational posterior $q(\mathbf{z})$

while the ELBO has not converged **do**

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

return $q\left(\mathbf{z}\right) = \prod_{j=1}^{m} q_j\left(z_j\right)$

CAVI in exponential families

When the complete conditionals $p\left(z_j|\mathbf{z}_{-j},\mathbf{x}\right)$ belong to an exponential family

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

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

$$\begin{aligned}
q_j(z_j) &\propto \exp\left[\mathbb{E}_{-j}\log p\left(z_j|\mathbf{z}_{-j},\mathbf{x}\right)\right] \\
&= \exp\left[\log h\left(z_j\right) + \left\{\mathbb{E}_{-j}\eta_j\left(\mathbf{z}_{-j},\mathbf{x}\right)\right\}^\top z_j - \mathbb{E}_{-j}A\left(\eta_j\left(\mathbf{z}_{-j},\mathbf{x}\right)\right)\right] \\
&\propto h\left(z_j\right)\exp\left[\left\{\mathbb{E}_{-j}\eta_j\left(\mathbf{z}_{-j},\mathbf{x}\right)\right\}^\top z_j\right]
\end{aligned}$$

Example: 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 $z_{dn} \sim \mathrm{Discrete}(\theta_d)$, i.e. $p(z_{dn} = k|\theta_d) = \theta_{dk}$
      2. Draw a word $w_{dn} \sim \mathrm{Discrete}(\beta_{z_{dn}})$, i.e. $p(w_{dn} = v|\beta, z) = \beta_{z_{dn}v}$

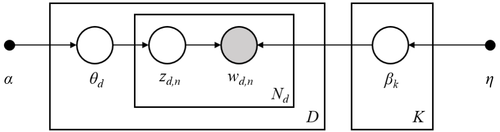

Figure: Graphical model representation of LDA. Plates represent replication, for example there are D documents each having a topic proportion vector θ_d

Latent Dirichlet Allocation

Mean-field family:

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

1. Complete conditional on the topic assignment is a multinomial

$$p(z_{dn}=k|\theta_d,\beta,w_d) \propto \theta_{dk}\beta_{k,w_{dn}} = \exp\left(\log\theta_{dk} + \log\beta_{k,w_{dn}}\right).$$

2. Complete conditional on the topic proportions is a Dirichlet

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

3. Complete conditional on the topics is another Dirichlet

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

Variational Autoencoder (VAE)

- A **probabilistic deep generative model**: a pair of neural networks jointly trained to approximately copy inputs at the outputs while passing them through a lower-dimensional representation.
  - An encoder / recognition model $q_\phi(z|x)$, of **latent codes** $z \in \mathbb{R}^{d_z}$, given inputs $x \in \mathbb{R}^{d_x}$, $d_z \ll d_x$, parametrized by a neural network with weights ϕ,
  - A decoder / generative model $p_\theta(x|z)$, of outputs $x \in \mathbb{R}^{d_x}$, given codes $z \in \mathbb{R}^{d_z}$, parametrized by a neural network with weights θ.

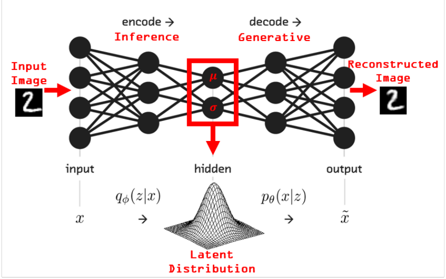

Figure: Figure from Kaggle tutorial on VAEs for MNIST

VAE ELBO

The decoder specifies the likelihood and the encoder is a variational approximation to the intractable posterior of latent codes.
ELBO for a single observation x:

$$\begin{aligned}
\mathcal{L}(x,\theta,\phi) &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z)\right] + H\left(q_\phi\left(\cdot|x\right)\right) \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi\left(z|x\right)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(z)}{q_\phi\left(z|x\right)}\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right] \\
&= -KL\left(q_\phi\left(z|x\right)||p(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right]. \qquad (1)
\end{aligned}$$

The common choice is $q_\phi(z|x) = \mathcal{N}\left(z|\mu_\phi\left(x\right), \Sigma_\phi\left(x\right)\right)$, where $\mu_\phi(x)$ and $\Sigma_\phi(x)$ are the outputs of a neural network. The prior is typically $p(z) = \mathcal{N}(0, I)$, so the KL term is tractable.

$$KL\left(q_\phi\left(z|x\right)||p(z)\right) = \frac{1}{2}\left[\mu_\phi\left(x\right)^\top \mu_\phi\left(x\right) + \mathrm{tr}\left(\Sigma_\phi\left(x\right)\right) - \log\det\left(\Sigma_\phi\left(x\right)\right) - d_z\right].$$

VAE ELBO

ELBO on the whole set of observations $\{x_i\}_{i=1}^n$, average over individual terms in (1):

$$\mathcal{L}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}_{q_\phi(z|x_i)} \left[\log p_\theta \left(x_i | z \right) \right] - KL \left(q_\phi \left(z | x_i \right) || p(z) \right) \right\}. \quad (2)$$

- Lower bound on the (scaled) model evidence $\frac{1}{n} \log p_\theta \left(\{x_i\}_{i=1}^n \right) = \frac{1}{n} \sum_{i=1}^n \log p_\theta \left(x_i \right)$, since $\mathcal{L}(x_i, \theta, \phi) \leq \log p_\theta \left(x_i \right)$, for all i.
- Use Stochastic gradient descent to jointly maximize (2) with respect to θ and ϕ using minibatches of observations x_i at the time in order to compute unbiased estimators of the gradients of ELBO.

Reparametrization trick

- The terms $\mathbb{E}_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)]$ are generally not tractable.
- A simple idea: obtain an unbiased estimator with drawing a single $z_i \sim q_\phi(z|x_i)$ and estimating

$$\hat{\mathbb{E}}_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] = \log p_\theta(x_i|z_i).$$

- Problem: cannot compute the gradients of this estimator with respect to ϕ as explicit dependence on the variational parameters ϕ has been lost.
- Solution is the so called "Reparametrization trick": a draw $z_i \sim \mathcal{N}(z|\mu_\phi(x), \Sigma_\phi(x))$ can be written as $z_i = \mu_\phi(x) + \Sigma_\phi^{1/2}(x)\epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, I)$, so can rewrite

$$\mathbb{E}_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] = \mathbb{E}_\epsilon\left[\log p_\theta\left(x_i|\mu_\phi(x) + \Sigma_\phi^{1/2}(x)\epsilon\right)\right],$$

and use an estimator of the form

$$\log p_\theta\left(x_i|\mu_\phi(x) + \Sigma_\phi^{1/2}(x)\epsilon_i\right),$$

based on a single draw $\epsilon_i \sim \mathcal{N}(0, I)$, with gradients w.r.t. ϕ and θ both available.

Other criteria

Lower bounds other than ELBO are possible. If have access to to some stricly positive unbiased estimator $\hat{p}_\theta(x)$ of $p_\theta(x)$, with

$$\int \hat{p}_\theta(x)\, q_{\theta,\phi}(u|x)\, du = p_\theta(x),$$

where $u \sim q_{\theta,\phi}(\cdot|x)$ denotes all random variables used to compute the estimator and ϕ parametrizes the sampling distribution of u.
By Jensen's inequality:

$$\int \log \hat{p}_\theta(x)\, q_{\theta,\phi}(u|x)\, du \quad \leq \quad \log \int \hat{p}_\theta(x)\, q_{\theta,\phi}(u|x)\, du \leq \log p_\theta(x).$$

- In the standard VAE ELBO, $u = z$ and $\hat{p}_\theta(x) = p_\theta(x, z) / q_\phi(z|x)$
- Other options include Importance Weighted Autoencoder (IWAE) using s importance samples $u = \{z_j\}_{j=1}^{s}$, with $z_j \sim q_\phi(\cdot|x)$

$$\hat{p}_\theta(x) = \frac{1}{s} \sum_{j=1}^{s} \frac{p_\theta(x, z_j)}{q_\phi(z_j|x)}.$$