Probabilistic Modelling and Reasoning
Exercises 8 — Notes

Spring 2022
Gutmann

These notes are intended to give a summary of relevant concepts from the lectures which are helpful to complete the exercises. It is not intended to cover the lectures thoroughly. Learning this content is not a replacement for working through the lecture material and the exercises.

KL divergence — The Kullback-Leibler divergence measures the "distance" between p and q:

$$\mathrm{KL}(p||q) = \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \tag{1}$$

It satisfies: $\mathrm{KL}(p||q) = 0 \Leftrightarrow p = q$, $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$, $\mathrm{KL}(p||q) \geq 0$. Optimising with respect to the first argument when the second is fixed leads to mode seeking. Optimising with respect to the second argument when the first is fixed produces global fits (moment-matching).

ELBO — For a joint model $p(\mathbf{x}, \mathbf{y})$, the evidence lower bound (ELBO) is

$$\mathcal{L}_{\mathbf{x}}(q) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{q(\mathbf{y}|\mathbf{x})}\right] \tag{2}$$

where $q(\mathbf{y}|\mathbf{x})$ is the variational distribution. It can be rewritten as

$$\log p(\mathbf{x}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y}|\mathbf{x})) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{y}) - \mathrm{KL}(q(\mathbf{y}|\mathbf{x})||p(\mathbf{y})) = \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} \log p(\mathbf{x}, \mathbf{y}) + \mathcal{H}(q)$$

where $\mathcal{H}(q) = -\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{x})\right]$ is the entropy of q. The ELBO is a lower bound on $\log p(\mathbf{x})$. It is maximised when $q(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})$ which makes the bound tight.

EM algorithm — The expectation maximisation (EM) algorithm can be used to learn the parameters $\boldsymbol{\theta}$ of a statistical model $p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})$ with latent (unobserved) variables $\mathbf{h}$ and visible (observed) variables $\mathbf{v}$ for which we have data $\mathcal{D}$. It updates the parameters $\boldsymbol{\theta}$ by iterating between the expectation (E) and the maximisation (M) step:

$$\text{E-step: compute } J(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_{\text{old}})}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})] \qquad \text{M-step: } \boldsymbol{\theta}_{\text{new}} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, J(\boldsymbol{\theta}) \tag{3}$$

The update rule produces a sequence of parameters for which the log-likelihood is guaranteed to never decrease, i.e. $\ell(\boldsymbol{\theta}_{\text{new}}) \geq \ell(\boldsymbol{\theta}_{\text{old}})$.

Amortisation — For iid data $\mathbf{v}_1, \ldots, \mathbf{v}_n$, the ELBO for the statistical model $p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})$ is

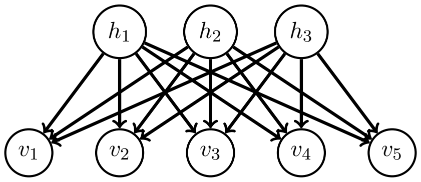
$$\mathcal{L}_{\mathcal{D}} = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}, q) \qquad \mathcal{L}_i(\boldsymbol{\theta}, q) = \mathcal{L}_{\mathbf{v}_i}(\boldsymbol{\theta}, q) = \mathbb{E}_{q(\mathbf{h}|\mathbf{v}_i)}\left[\log \frac{p(\mathbf{v}_i, \mathbf{h}; \boldsymbol{\theta})}{q(\mathbf{h}|\mathbf{v}_i)}\right] \tag{4}$$

Learning n $q(\mathbf{h}|\mathbf{v}_i)$ is too costly when n is large. Amortisation means that the $q(\mathbf{h}|\mathbf{v}_i)$ are parametrised as $q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v}_i) = q(\mathbf{h}; \boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v}_i))$, where $q(\mathbf{h}; \boldsymbol{\lambda})$ is some parametric model and $\boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v})$ and its parameters $\boldsymbol{\phi}$ are shared among all n data points. $\boldsymbol{\phi}$ is learned by maximising $\mathcal{L}_{\mathcal{D}}$.

Reparametrisation We can sample from distributions by deterministically transforming another random variable $\boldsymbol{\epsilon}$ drawn from some base distribution $p(\boldsymbol{\epsilon})$. For example, $h \sim \mathcal{N}(h; \mu, \sigma^2) \Leftrightarrow h = \mu + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(\epsilon, 0, 1)$. Sampling $\mathbf{h} \sim q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v})$ as $\mathbf{h} = \mathbf{t}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{v})$ allows us to write the ELBO in terms of an expectation with respect to $\boldsymbol{\epsilon}$ and hence pull $\nabla_{\boldsymbol{\phi}}$ inside the expectation when maximising the ELBO with gradient ascent.

VAE — The variational autoencoder (VAE) is a deep latent variable model. The model is defined by the DAG on the left with a standard normal distribution $p(\mathbf{h}) = \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I})$ for $\mathbf{h}$ and conditionals $p(v_k|\mathbf{h}; \boldsymbol{\theta}) = p(v_k; \boldsymbol{\eta}_k)$ for the visibles v_k, where $\boldsymbol{\eta}_k = \boldsymbol{\eta}^k_{\boldsymbol{\theta}}(\mathbf{h})$ is a nonlinear mapping called the decoder (network) that maps the latents $\mathbf{h}$ to the parameters $\boldsymbol{\eta}$ of a parametric family $\{p(v; \boldsymbol{\eta})\}_{\boldsymbol{\eta}}$.

DAG:



- Gaussian autoencoder:

$$p(v_k; \boldsymbol{\eta}_k) = \mathcal{N}(v_k; m_k, s_k^2), \quad v_k \in \mathbb{R}$$

- Bernoulli VAE:

$$p(v_k; \boldsymbol{\eta}_k) = p_k^{v_k}(1 - p_k)^{(1-v_k)}, \quad v_k \in \{0, 1\}$$

The model is learned by stochastic gradient ascent on $\mathcal{L}_{\mathcal{D}}$ using amortisation and e.g. reparametrisation. The variational distribution is often assumed to be a factorised Gaussian $q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v}) = \prod_k \mathcal{N}(h_k; \mu_k(\mathbf{v}), \sigma_k^2(\mathbf{v}))$, where the means and variances $(\mu_1, \ldots, \mu_H, \sigma_1^2, \ldots, \sigma_H^2)$ are outputs of the encoder (network) $\boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v})$. The ELBO for a d-dimensional data point $\mathbf{v}_i = (v_{i1}, \ldots, v_{id})$ is

$$\mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underbrace{\sum_{k=1}^{d} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v}_i)} \left[\log p(v_{ik}; \boldsymbol{\eta}^k_{\boldsymbol{\theta}}(\mathbf{h}))\right]}_{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v}_i)} \log p(\mathbf{v}_i|\mathbf{h})} + \underbrace{\frac{1}{2} \sum_{k=1}^{H} \left(1 + \log(\sigma_k^2(\mathbf{v}_i)) - \sigma_k^2(\mathbf{v}_i) - \mu_k^2(\mathbf{v}_i)\right)}_{-\mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v}_i)||p(\mathbf{h}))} \tag{5}$$