These notes summarise selected lecture concepts and are not a substitute for working through the lecture slides, tutorials, and self-study exercises. Feel free to personalise and develop them into your own summary sheet.

ELBO for iid data — For a model $p(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})$ and iid data $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, the overall ELBO is a sum of ELBOs for each data point

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

Can be expressed as $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}, q) = \ell(\boldsymbol{\theta}) - \sum_{i=1}^{n} \mathrm{KL}(q(\mathbf{h}_i|\mathbf{v}_i)||p(\mathbf{h}_i|\mathbf{v}_i; \boldsymbol{\theta}))$.

Amortisation — We represent $q(\mathbf{h}|\mathbf{v})$ as a member of a parametric family $q(\mathbf{h}; \boldsymbol{\lambda})$ whose parameter $\boldsymbol{\lambda}$ is the output of a learnable function $\boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v})$. Instead of optimising $\mathcal{L}_i$ for each data point separately, we learn the parameters $\boldsymbol{\phi}$ of $\boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v})$ by maximising $\mathcal{L}_{\mathcal{D}}$. While more efficient this introduces an amortisation gap: $\mathcal{L}_i(\boldsymbol{\theta}, q)$ for $q(\mathbf{h}_i; \boldsymbol{\lambda}_{\hat{\boldsymbol{\phi}}}(\mathbf{v}_i))$ may not attain $\max_q \mathcal{L}_i(\boldsymbol{\theta}, q)$.

Reparameterisation — Reparameterisation is the idea of expressing a random variable as a deterministic function $\mathbf{t}_{\boldsymbol{\phi}}$ of noise $\boldsymbol{\epsilon}$ and parameters $\boldsymbol{\phi}$. For conditional random variables this becomes: $\mathbf{h} \sim q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{v}) \iff \mathbf{h} = \mathbf{t}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{v}), \quad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$. This removes the $\boldsymbol{\phi}$ dependency from the expectation in the ELBO and makes optimisation easier:

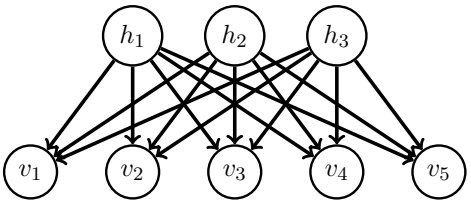
$$\mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{L}_i(\boldsymbol{\theta}, q_{\boldsymbol{\phi}}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{h}_i|\mathbf{v}_i)}\left[\log \frac{p(\mathbf{v}_i, \mathbf{h}_i; \boldsymbol{\theta})}{q_{\boldsymbol{\phi}}(\mathbf{h}_i|\mathbf{v}_i)}\right] = \mathbb{E}_{p(\boldsymbol{\epsilon}_i)}\left[\log \frac{p(\mathbf{v}_i, \mathbf{t}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_i, \mathbf{v}_i); \boldsymbol{\theta})}{q_{\boldsymbol{\phi}}(\mathbf{t}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}_i, \mathbf{v}_i)|\mathbf{v}_i)}\right] \tag{2}$$

Variational autoencoder — A deep latent variable model with H independent latents h_i and $d > H$ conditionally independent visibles v_i, typically generated nonlinearly from the latents.

$$p(\mathbf{h}) = \prod_{i=1}^{H} p(h_i) \tag{3}$$

$$p(\mathbf{v}|\mathbf{h}; \boldsymbol{\theta}) = \prod_{k=1}^{d} p(v_k|\mathbf{h}; \boldsymbol{\theta}) \tag{4}$$

$$p(v_k|\mathbf{h}; \boldsymbol{\theta}) = p(v_k; \boldsymbol{\eta}_{\boldsymbol{\theta}}^k(\mathbf{h})) \tag{5}$$



The decoder $\boldsymbol{\eta}_{\boldsymbol{\theta}}^k(\mathbf{h})$ maps $\mathbf{h}$ to the parameters $\boldsymbol{\eta}$ of the parametric model $p(v_k; \boldsymbol{\eta})$. Its parameters $\boldsymbol{\theta}$ are learned by maximising the ELBO using amortised variational distributions $q(\mathbf{h}|\mathbf{v}) = q(\mathbf{h}; \boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v}))$ and reparameterisation. The mapping $\boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v})$ is called the encoder. If $p(\mathbf{h}) = \mathcal{N}(\mathbf{h}; \mathbf{0}, \mathbf{I})$ and $q(\mathbf{h}; \boldsymbol{\lambda}_{\boldsymbol{\phi}}(\mathbf{v}))$ is a factorised Gaussian with means $\mu_k(\mathbf{v})$ and variances $\sigma_k^2(\mathbf{v})$

$$\mathcal{L}_i(\boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{k=1}^{d} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{h}_i|\mathbf{v}_i)}\left[\log p(v_{ik}; \boldsymbol{\eta}_{\boldsymbol{\theta}}^k(\mathbf{h}_i))\right] + \frac{1}{2}\sum_{k=1}^{H}\left(1 + \log(\sigma_k^2(\mathbf{v}_i)) - \sigma_k^2(\mathbf{v}_i) - \mu_k^2(\mathbf{v}_i)\right) \tag{6}$$

Gaussian VAEs — $p(v_k; \boldsymbol{\eta}_k)$ is Gaussian with $\boldsymbol{\eta}_k = (\mu_k, s_k^2)$ equal to the mean and variance.

Bernoulli VAE — $p(v_k; \eta_k)$ is Bernoulli with $\eta_k = p_k$ equal to the success probability.