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.

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.

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.

Variational inference (VI) — We compute the posterior $q(\mathbf{y}|\mathbf{x})$ as $\mathrm{argmax}_{q \in \mathcal{Q}} \mathcal{L}_{\mathbf{x}}(q)$ where $\mathcal{Q}$ is the variational family.

Mean-field VI — Assumes that q fully factorises: $q(\mathbf{y}|\mathbf{x}) = \prod_i q(y_i|\mathbf{x})$. In coordinate ascent VI, each q_i is sequentially updated as

$$q_i(y_i|\mathbf{x}) = \frac{1}{Z} \exp\left[\mathbb{E}_{q(\mathbf{y}_{\setminus i}|\mathbf{x})}\left[\log p(\mathbf{x}, \mathbf{y})\right]\right] \tag{3}$$

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:

$$\text{E-step: compute } J(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{h}|\mathcal{D};\boldsymbol{\theta}_{\text{old}})}[\log p(\mathcal{D}, \mathbf{h}; \boldsymbol{\theta})] \qquad \text{M-step: } \boldsymbol{\theta}_{\text{new}} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{4}$$

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