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.

Intractability — The cost of integrals (sums) grows exponentially with the dimension (curse of dimensionality), making likelihoods defined via them intractable.

Key identity — For a non-negative function $g(\mathbf{x};\boldsymbol{\theta})$, with $\pi(\mathbf{x};\boldsymbol{\theta}) = g(\mathbf{x};\boldsymbol{\theta})/\int g(\mathbf{x}';\boldsymbol{\theta})\mathrm{d}\mathbf{x}'$

$$\nabla_{\boldsymbol{\theta}} \log \int g(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x} = \mathbb{E}_{\pi(\mathbf{x};\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}} \log g(\mathbf{x};\boldsymbol{\theta})\right]. \tag{1}$$

Unobserved variables — For a model $p(\mathbf{h},\mathbf{v};\boldsymbol{\theta})$ with unobserved (latent or missing) variables $\mathbf{h}$ and iid data $\mathbf{v}_1,\dots,\mathbf{v}_n$, each with its own $\mathbf{h}_i$, the key identity with $\mathbf{x} = \mathbf{h}_i$ gives

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\mathbf{h}_i,\mathbf{v}_i;\boldsymbol{\theta})\mathrm{d}\mathbf{h}_i \tag{2}$$

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{h}_i|\mathbf{v}_i;\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{h}_i,\mathbf{v}_i;\boldsymbol{\theta})\right]. \tag{3}$$

The gradient averages the gradients for "filled-in" data $(\mathbf{h}_i,\mathbf{v}_i)$ over the posterior: one inference problem per data point.

Unnormalised models — For an unnormalised model $\tilde{p}(\mathbf{v};\boldsymbol{\theta})$ with partition function $Z(\boldsymbol{\theta}) = \int \tilde{p}(\mathbf{v};\boldsymbol{\theta})\mathrm{d}\mathbf{v}$, so that $p(\mathbf{v};\boldsymbol{\theta}) = \tilde{p}(\mathbf{v};\boldsymbol{\theta})/Z(\boldsymbol{\theta})$, and features $\mathbf{m}(\mathbf{v};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{v};\boldsymbol{\theta})$, the key identity with $\mathbf{x} = \mathbf{v}$ gives

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \tilde{p}(\mathbf{v}_i;\boldsymbol{\theta}) - n \log Z(\boldsymbol{\theta}), \qquad \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{m}(\mathbf{v}_i;\boldsymbol{\theta}) - n\,\mathbb{E}_{p(\mathbf{v};\boldsymbol{\theta})}\left[\mathbf{m}(\mathbf{v};\boldsymbol{\theta})\right]. \tag{4}$$

The gradient compares the features on the data with their expectation under the model. $Z(\boldsymbol{\theta})$ cannot be ignored: the two terms in the log-likelihood balance each other.

Both — For $p(\mathbf{h},\mathbf{v};\boldsymbol{\theta}) = \tilde{p}(\mathbf{h},\mathbf{v};\boldsymbol{\theta})/Z(\boldsymbol{\theta})$ with $\mathbf{m}(\mathbf{h},\mathbf{v};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{h},\mathbf{v};\boldsymbol{\theta})$, the key identity with $\mathbf{x} = \mathbf{h}$ and $\mathbf{x} = (\mathbf{h},\mathbf{v})$ gives

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int \tilde{p}(\mathbf{h},\mathbf{v}_i;\boldsymbol{\theta})\mathrm{d}\mathbf{h} - n \log Z(\boldsymbol{\theta}) \tag{5}$$

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{h}|\mathbf{v}_i;\boldsymbol{\theta})}\left[\mathbf{m}(\mathbf{h},\mathbf{v}_i;\boldsymbol{\theta})\right] - n\,\mathbb{E}_{p(\mathbf{h},\mathbf{v};\boldsymbol{\theta})}\left[\mathbf{m}(\mathbf{h},\mathbf{v};\boldsymbol{\theta})\right]. \tag{6}$$