Problem 17.1

How many parameters are needed to create a 1D mixture of Gaussians with $n = 5$ components (equation 17.4)? State the possible range of values that each parameter could take.

$$P(x) = \sum_{i=1}^{n} \lambda_n \cdot \text{Norm}_x[\mu_n, \sigma_n^2] \tag{17.4}$$

In a 1D mixture of Gaussians, the latent variable z is discrete, and the prior $P(z)$ is a categorical distribution with one probability λ_n for each possible valye of z. The likelihood $P(x|z = n)$ of the data x given that the latent variable takes value n is normally distributed with mean μ_n and variance σ_n^2. The marginal distribution of the data $P(x)$ is obtained by summing over all possible values of the latent variable.

Therefore, since the mixing coefficients must sum to 1, $\lambda_5 = 1 - \sum_{i=1}^{4} \lambda_i$, meaning that there 4 independent mixing coefficients. Each Gaussian component has two parameters, the mean and the variance, so there are $5 \times 2 = 10$ parameters in total. This gives a total of $4 + 10 = 14$ parameters.

For the ranges of these values, the mixing coefficients must be non-negative and sum to 1, so $0 \leq \lambda_i \leq 1$ and $\sum_{i=1}^{n} \lambda_i = 1$. The means μ_i can take any real value, and the variances σ_i^2 must be positive, so $\sigma_i^2 > 0$.

Problem 17.2

A function is concave if its second derivative is less than or equal to zero everywhere. Show that this is true for the function $g(x) = \log(x)$.

The first derivative of $g(x) = \log(x)$ is:

$$\frac{d}{dx}\log(x) = \frac{1}{x}$$

The second derivative is:

$$\frac{d^2}{dx^2}\log(x) = -\frac{1}{x^2}$$

Since the second derivative is negative for all $x > 0$, the function $g(x) = \log(x)$ is concave for $x > 0$.

Problem 17.3

For convex functions, Jensen's inequality works the other way around.

$$g(\mathbb{E}[x]) \leq \mathbb{E}[g(x)] \tag{17.31}$$

A function is convex if its second derivative is greater than or equal to zero everywhere. Show that the function $g(x) = x^{2n}$ is convex for arbitrary $n \in [1, 2, \ldots]$. Use this result with Jensen's inequality to show that the square of the mean $\mathbb{E}[x]$ of a distribution $P(x)$ must be less than or equal to its second moment $\mathbb{E}[x^2]$.

The first derivative of $g(x) = x^{2n}$ is:

$$\frac{d}{dx}x^{2n} = 2nx^{2n-1}$$

The second derivative is:

$$\frac{d^2}{dx^2}x^{2n} = 2n(2n-1)x^{2n-2}$$

Where $n \geq 1$, the second derivative is positive for all $x \in \mathbb{R}$ since the prefactor $2n(2n-1)$ is positive. Therefore, the function $g(x) = x^{2n}$ is convex for $n \in [1, 2, \ldots]$.

Applying Jensen's inequality to the convex function $g(x) = x^2$, we have:

$$\mathbb{E}[x^2] \geq (\mathbb{E}[x])^2$$

This states that the mean of a distribution $P(x)$ squared is less than or equal to the second moment of the distribution.

Problem 17.4

Show that the ELBO, as expressed in equation 17.18, can alternatively be derived from the KL divergence between the variational distribution $q(\mathbf{z}|\mathbf{x})$ and the true posterior distribution $P(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})$:

$$D_{KL}[q(\mathbf{z}|\mathbf{x})||P(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})] = \int q(\mathbf{z}|\mathbf{x}) \log \left[\frac{q(\mathbf{z}|\mathbf{x})}{P(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})}\right] d\mathbf{z} \tag{17.32}$$

Start by using Bayes' rule (equation 17.19).

$$D_{KL}[q(\mathbf{z}|\mathbf{x})||P(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})] = \int q(\mathbf{z}|\mathbf{x}) \log\left[\frac{q(\mathbf{z}|\mathbf{x})}{P(\mathbf{z}|\mathbf{x},\boldsymbol{\phi})}\right] d\mathbf{z}$$

Problem 17.5

The reparameterization trick computes the derivative of an expression of a function $f(x)$:

$$\frac{\partial}{\partial \boldsymbol{\phi}} \mathbb{E}_{P(x|\boldsymbol{\phi})}[f(x)]$$

with respect to the parameters $\boldsymbol{\phi}$ of the distribution $P(x|\boldsymbol{\phi})$. Show that this derivative can also be computed as:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\phi}} \mathbb{E}_{P(x|\boldsymbol{\phi})}[f(x)] &= \mathbb{E}_{P(x|\boldsymbol{\phi})}\left[f(x)\frac{\partial}{\partial \boldsymbol{\phi}}\log(P(x|\boldsymbol{\phi}))\right] \\ &\approx \frac{1}{I}\sum_{i=1}^{I} f(x_i)\frac{\partial}{\partial \boldsymbol{\phi}}\log(P(x|\boldsymbol{\phi})).\end{aligned}$$

This method is known as the REINFORCE algorithm or score function estimator.

Problem 17.6

Why is it better to use spherical linear interpolation rather than regular linear interpolation when moving between points in the latent space? Hint: consider figure 8.13

Problem 17.7

Derive the EM algorithm for the 1D mixture of Gaussians algorithm with N components. To do this, you need to (i) find an expression for the posterior distribution $P(z|x)$ over the latent variable $z \in \{1, 2, \ldots, N\}$ for a data point x and (ii) find an expression that updates the evidence lower bound given the posterior distributions for all of the data points. You will need to use Lagrange multipliers to ensure that the weights $\lambda_1, \ldots, \lambda_N$ of the Gaussians sum to one.