1 Exercise 1

To generate a sample of such a discrete random variable X, we use the inverse c.d.f theorem. If F is the c.d.f of X, then:

$$U \sim \mathcal{U}[0,1] \implies F^{(-1)}(U) \sim X$$

where $F^{(-1)}$ is the inverse c.d.f function.
In the discrete case we have

$$\begin{aligned} F^{(-1)}(u) &= \inf\{x \in \mathbb{R}, F(x) \geq u\} \\ &= \inf\{x \in \mathbb{R}, \sum_{k=1}^{n} p_k \mathbf{1}_{\{X_k \leq x\}} \geq u\} \\ &= \{x_k : s_{k-1} < u \leq s_k\} \end{aligned}$$

with $s_k = \sum_{i=1}^{k} p_i$ if $k \geq 1$; 0 if $k = 0$.
To simulate one sample x of X, we sample one point u from U. If $\sum_{i=1}^{k-1} p_i < u \leq \sum_{i=1}^{k} p_i$ for one $k \geq 1$, we choose $x = x_k$.

2 Exercise 2

2.1

The parameters are:

- $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, p$, with $\sum_{j=1}^{p} \alpha_j = 1$.
- $\mu_j \in \mathbb{R}^d$, $j = 1, \ldots, p$.
- $\Sigma_j \in S_d^{++}(\mathbb{R})$, $j = 1, \ldots, p$.

For any $\theta = (\alpha_1, \ldots, \alpha_p, \mu_1, \ldots, \mu_p, \Sigma_1, \ldots, \Sigma_p)$, its corresponding density function $p_\theta(x)$ for any observation x is given by

$$p_\theta(x) = \sum_{j=1}^{p} \alpha_j f_{\mu_j, \Sigma_j}(x)$$

where $f_{\mu_j,\Sigma_j} : x \mapsto \frac{1}{\sqrt{2\pi|\Sigma_j|}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}$ is the p.d.f of $\mathcal{N}(\mu_j, \Sigma_j)$.
Thus the likelihood of θ given the outcomes $(x_i)_{1 \leq i \leq n}$ of i.i.d n-sample $(X_i)_{1 \leq i \leq n}$ is

$$\mathcal{L}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \sum_{j=1}^{p} \alpha_j f_{\mu_j, \Sigma_j}(x).$$

2.2

We choose θ randomly with fixed $d = 2$.

2.3

We aim to maximise $Q(\theta|\theta_t) = \mathbb{E}_{Z \sim p_{\theta_t}(Z|x)}[\log p(Z, x|\theta^t)]$.

- Initialization (k-means or empirical): set up $\hat{\alpha}^0, \hat{\mu}^0, \hat{\Sigma}^0$.
- E step: Compute $\tau_{i,j}(\theta_t) = \tau_{i,j}^t = \frac{\hat{\alpha}_j^t f_{\hat{\mu}_j^t, \hat{\Sigma}_j^t}(x_i)}{\sum_{j=1}^{p} \hat{\alpha}_j^t f_{\hat{\mu}_j^t, \hat{\Sigma}_j^t}}(x_i)$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$.
- M step: Update the parameters

1. $\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} \tau_{i,j}^t$.
2. $\hat{\mu}_j = \frac{\sum_{i=1}^{n} \tau_{i,j}^t x_i}{\sum_{i=1}^{n} \tau_{i,j}^t}$.
3. $\hat{\Sigma}_j = \frac{1}{\sum_{i=1}^{n} \tau_{i,j}^t} \tau_{i,j}^t (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T$.

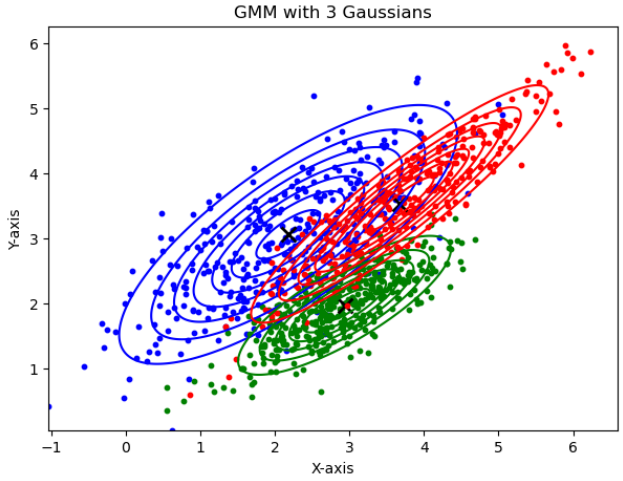
where these quantities are derived from the minimization problem

$$\min_{\theta \in K} -Q(\theta|\theta^t) = \min_{\theta \in K} -\sum_{i,j} [\log \alpha_j + \log f_{\mu_j,\Sigma_j}(x_i)] \tau_{i,j}(\theta^t)$$

where $K = \{\sum_{j=1}^{p} \alpha_j = 1, \alpha_j \geq 0, \Sigma_j \in S_d^{++}(\mathbb{R}) \; \forall 1 \leq j \leq p\}$ is a convex set and the objective function is a convex function. So it's a convex problem. The solutions verify $\nabla_\theta Q(\theta|\theta^t) = 0$.

Go back to E step if $t < N$ where N is fixed or $|\log \mathcal{L}(x_1, \ldots, x_n; \theta^t) - \log \mathcal{L}(x_1, \ldots, x_n; \theta^{t-1})| < \epsilon$ where ϵ is a tolerance.

2.4



The plot above shows us that the estimated parameters are close to the true ones. Given a number of clusters, the algorithm manages to approximate relatively well the distribution of the data.

2.5

The shape of the data shows us that using a gaussian mixture model on it might be a good idea. The data seem to have (overlapping) clusters, which GMMs are particularly good at modeling since each Gaussian component can represent one cluster. Moreover the overall distribution does not appear to be completely unimodal. That's why it could fit with a combination of several Gaussian distributions.