Please submit your solutions via Ilias. The submission is not a formal requirement for passing the exam but doing the exercises will be very helpful to do so. Submissions should be a single PDF document (note that Jupyter notebooks can and should also be downloaded as PDFs, and not only submitted as .ipynb files).

## Expectation Maximization and Gaussian Mixture Models

1. **Short Theory Questions — Expectation Maximization (EM)**

   Let's say we want to find the maximum likelihood estimate for a model: $\arg\max_{\theta} \log p(x|\theta)$. We can use EM, if introducing a latent variable $z$, makes maximizing the complete-data log-likelihood $\log p(x, z|\theta)$ possible. EM allows us to maximize the marginal log-likelihood $\log p(x|\theta)$ by iteratively computing:

   1. E-step: $p(z|x, \theta_{old})$ to derive $\mathbb{E}_{p(z|x,\theta_{old})} \log p(x, z|\theta)$
   2. M-step: $\arg\max_{\theta} \mathbb{E}_{p(z|x,\theta_{old})} \log p(x, z|\theta)$.

   It is, however, non-obvious how this scheme maximizes the marginal log-likelihood. Here, you will show that repeatedly running these two steps guarantees that one monotonically increases $\log p(x|\theta)$. To make the argument in part (c), you will derive the required tools in parts (a) and (b).

   (a) Show that for any distribution $q(z)$ over latent variables, $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(x,z|\theta)}{q(z)}\right]$ lower bounds the marginal log-likelihood.
   *Hint: You can use Jensen's inequality for that.*
   ***Simplified form of Jensen's inequality****: Let $X$ be a real-valued random variable, and $\phi$ be a concave function (e.g.* $\log$*). Then*

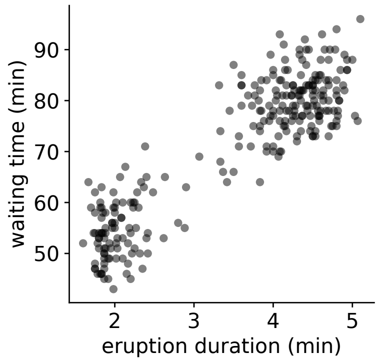
   $$\phi(\mathbb{E}[X]) \geq \mathbb{E}[\phi(X)].$$

   (b) We can write the marginal log-likelihood as the sum of the previously derived lower bound and an additional term: $\log p(x|\theta) = \mathcal{L}(q) + ....$ Derive the additional term.
   *Hint: Start by splitting $\mathcal{L}(q)$ into the terms depending on $z$ and the terms not depending on $z$. Note that the answer will take the form of the KL-divergence $D_{KL}$.*

   (c) Now argue that EM monotonically increases the marginal log-likelihood. Think about what happens to $\mathcal{L}$, as well as to the $D_{KL}$, during the E and M steps. Your answer should only require a few sentences.
   *Hint: The KL divergence satisfies $D_{KL}(q(z)||p(z)) = 0 \iff q = p$.*

2. **Theory Question — EM for Gaussian Mixtures**

   The plot below shows a classic dataset that you already saw in the lecture: the interval between eruptions of the *Old Faithful* geyser in Yellowstone National Park in the US, plotted against the duration of the eruption following the waiting time. Datasets like this, which show a "cluster" structure, are often modeled with *Gaussian mixture models* (GMMs): each datum $\boldsymbol{x}_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$ is a real vector. The data are assumed to come from $k$ separate Gaussian distributions (in this concrete case, $n = 272, d = 2, k = 2$), according to the following generative process:




For each datum $i \in [1, \dots, n]$:

- draw a discrete cluster identity $\boldsymbol{z}_i \in \{0,1\}^k, \sum_j z_{ij} = 1$ ("one-hot") with

$$p(\boldsymbol{z}_i \mid \pi) = \prod_{j=1}^{k} \pi_j^{z_{ij}}$$

- draw the datum $x_i \in \mathbb{R}^d$ from one of $k$ Gaussian distributions, selected by $\boldsymbol{z}_i$ with probability

$$p(\boldsymbol{x}_i \mid \boldsymbol{z}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{j=1}^{k} \mathcal{N}(\boldsymbol{x}_i; \mu_j, \Sigma_j)^{z_{ij}}$$

This process involves the following parameters: a probability $\pi \in \mathbb{R}^k_{+,0}, \sum_j^k \pi_j = 1$, and cluster parameters $\mu_j \in \mathbb{R}^d, \Sigma_j \in \mathbb{R}^{d \times d}$.

Write down an EM algorithm for the GMM above that finds a maximum likelihood assignment for the parameters $\pi_j, \mu_j, \Sigma_j$ defined above. To do so, proceed according to the example you have seen in the lecture, including all derivations, and not just the final result. Write down each step incl. the assumptions, rules etc. that you make/apply:

1. For the E-step, compute the expectation of the complete-data log-likelihood $\log p(\boldsymbol{z}, \boldsymbol{x} \mid \mu, \Sigma, \pi)$ under the posterior $p(\boldsymbol{z} \mid \mu, \Sigma, \boldsymbol{x})$.
2. For the M-step, analytically maximize this expression in $\pi, \mu, \Sigma$.

*Hints:*
***Constrained optimization***: *You may need to enforce constraints within the M-step of $\pi$, an easy technique to do so is to use Lagrange multipliers.*

***Matrix calculus***: *These identities may be helpful:*

- *Let $f : \mathbb{R}^{n \times n} \to \mathbb{R},\ f(A) = x^T A^{-1} x$, then $\nabla_A f(A) = -A^{-T} x x^T A^{-T}$.*
- *Let $f : \mathbb{R}^{n \times n} \to \mathbb{R},\ f(A) = det(A)$, then $\nabla_A f(A) = det(A)(A^{-1})^T$.*

3. **Practical Question** In this week's practical question, we will implement such an EM algorithm for a very common dataset that also shows a "cluster" structure, namely the iris dataset. You will find the required information in notebook `Exercise_06.ipynb`, which can be solved without solving the questions above.