Problem 1 (5 marks)

(MLE as KL Minimization) Suppose you are given N observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ from some true underlying data distribution $p_{data}(\boldsymbol{x})$ (may assume N to be very large, e.g., infinity). To learn it, you assume a parametrized distribution $p(\boldsymbol{x}|\theta)$ and estimate the parameters θ using MLE. Show that doing MLE is equivalent to finding θ that minimizes the KL divergence between the true distribution $p_{data}(\boldsymbol{x})$ and the assumed distribution $p(\boldsymbol{x}|\theta)$. Note that KL divergence between two probability distributions p and q is asymmetric and can be defined in two different ways: $KL(p||q)$ or $KL(q||p)$. For this problem, minimizing only one of these two will be equivalent to MLE. Why not the other one?

Problem 2 (5 marks)

(Distribution of Empirical Mean of Gaussian Observations) Consider N scalar-valued observations $x_1, \ldots, x_N$ drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Consider their empirical mean $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$. Representing the empirical mean as a linear transformation of a random variable, derive the probability distribution of $\bar{x}$.

Problem 3 (15 marks)

(Benefits of Hierarchical Modeling?) Consider a dataset of test-scores of students from M schools in a district: $\boldsymbol{x} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M} = \{x_1^{m}, \ldots, x_{N_m}^{(m)}\}_{m=1}^{M}$, where N_m denotes the number of students in school m. Assume the scores of students in school m are drawn independently as $x_n^{(m)} \sim \mathcal{N}(\mu_m, \sigma^2)$ where the Gaussian's mean μ_m is unknown and the variance σ^2 is same for all schools and known (for simplicity). Assume the means $\mu_1, \ldots, \mu_M$ of the M Gaussians to also be Gaussian distributed $\mu_m \sim \mathcal{N}(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are hyperparameters.

1. Assume the hyperparameters μ_0 and σ_0^2 to be known. Derive the posterior distribution of μ_m and write down the mean and variance of this posterior distribution. **Note:** While you can derive it the usual way, the derivation will be much more compact if you use the result of Problem 2 and think of each school's data as a *single* observation (the empirical mean of observations) having the distribution derived in Problem 2.

2. Assume the hyperparameter μ_0 to be unknown (but still keep σ_0^2 as fixed for simplicity). Derive the marginal likelihood $p(\boldsymbol{x}|\mu_0, \sigma^2, \sigma_0^2)$ and use MLE-II to estimate μ_0 (note again that σ^2 and σ_0^2 are known here). Note: Looking at the form/expression of the marginal likelihood, if the MLE-II result looks obvious to you, you may skip the derivation and directly write the result.

3. Consider using this MLE-II estimate of μ_0 from part (2) in the posteriors of each μ_m you derived in part (1). Do you see any benefit in using the MLE-II estimate of μ_0 as opposed to using a known value of μ_0?

Problem 4 (20 marks)

Binary Latent Matrices Consider modeling an $N \times K$ binary matrix $\mathbf{Z}$ with its entries assumed to be generated independently as follows

$$\begin{aligned} Z_{nk}|\pi_k &\sim \text{Bernoulli}(\pi_k) \qquad n = 1, \ldots, N, k = 1, \ldots, K \\ \pi_k &\sim \text{Beta}(\alpha/K, 1) \qquad k = 1, \ldots, K \end{aligned}$$

- Integrate out $\{\pi_k\}_{k=1}^{K}$ and derive the expression for the marginal prior $p(\mathbf{Z}|\alpha)$ and show that it can be written in form of a product of ratios of Beta functions.

- Derive the distribution $p(Z_{nk}|Z_{-nk})$ where Z_{-nk} denotes all the entries in k-th column of $\mathbf{Z}$, except Z_{nk}. Since Z_{nk} is binary, it suffices to compute $p(Z_{nk} = 1|Z_{-nk})$ (hint: Use Bayes rule). Explain why the form of the result makes intuitive sense.

- As a function of α, what will be the expected number of ones in each column of $\mathbf{Z}$, and in all of $\mathbf{Z}$?

Problem 5 (30 marks)

(Spike-and-Slab Model for Sparsity) Suppose w is a real-valued r.v. that can either be close to zero with probability π, or take a wide range of real values with probability $(1-\pi)$. An example of this could be in a regression problem where w is the weight of some feature. The feature could be irrelevant for predicting the output (in which case we would expect w to be close to zero) or be useful (in which case we would expect w to be non-zero with a wide range of possible values). We want to infer w from data taking a Bayesian approach. Note that, in practice, $\boldsymbol{w}$ is a vector (with each entry modeled this way) but here we will consider the scalar w case.

A popular approach to solve such problems is to impose a *spike and slab prior* on w. Let $b \in \{0, 1\}$ be a binary random variable and define the following *conditional* prior on w:

$$p(w|b, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}) = \begin{cases} \mathcal{N}(w|0, \sigma^2_{\text{spike}}) & b = 0 \\ \mathcal{N}(w|0, \sigma^2_{\text{slab}}) & b = 1, \end{cases}$$

Depending on the value of b (which itself is unknown), w is assumed drawn from one of the two distributions: a "peaky" one $\mathcal{N}(w|0, \sigma^2_{\text{spike}})$ with variance σ^2_{spike} being very small, and a "flat" one $\mathcal{N}(w|0, \sigma^2_{\text{slab}})$, with $\sigma_{\text{slab}} \gg \sigma_{\text{spike}}$. So, basically, the value of the binary "mask" b decides whether the feature is relevant or not.

We usually don't know b, so we must either infer it with w, or marginalize it if we care about the value of w.

- Assume a prior $p(b = 1) = \pi = 1/2$, which means both Gaussians are equally likely for w. What is the *marginal* prior $p(w|\sigma^2_{\text{spike}}, \sigma^2_{\text{slab}})$, i.e., the prior over w after integrating out b?

- Plot this marginal prior distribution for $(\sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}) = (1, 100)$. Briefly comment on how the shape of this distribution compares with that of a typical Gaussian distribution?

- Suppose someone gave us a "noisy" version of w defined as $x = w + \epsilon$ where $\epsilon \sim \mathcal{N}(\epsilon|0, \rho^2)$. This is equivalent to writing $p(x|w, \rho^2) = \mathcal{N}(x|w, \rho^2)$. Assume the variance ρ^2 to be known. Given x, what is the posterior distribution of b, $p(b = 1|x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \rho^2)$? Note that w must NOT appear in this expression (has to be integrated out first). Plot the resulting posterior $p(b = 1|x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \rho^2)$ as a function of x.

- Given the noisy observation $x = w + \epsilon$ as defined above, what is the posterior distribution of w, i.e., $p(w|x, \sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}, \rho^2)$? Note that b must NOT appear in this expression (has to be integrated out or summed over since b is discrete).

- Assume $(\sigma^2_{\text{spike}}, \sigma^2_{\text{slab}}) = (1, 100)$, the noise variance $\rho^2 = 0.01$. For these settings of the hyperparameters, plot the posterior distribution of w given a noisy observation $x = 3$.

Do not submit the code for this part. All of the answers/derivations for this part (including the plots) should be in the PDF writeup.