STCS 6701: Probabilistic Machine Learning
Homework 1

Due: Fri, Oct 17, 2025 at 11:59 pm ET

Instructions

- All homework should be typeset using LaTeX. Box your answers whenever appropriate.
- Standard late policy applies. Everyone has a total five late days throughout the semester. You are free to use them for whatever reason, no need to inform course staff.
- The homework should be turned in via Gradescope before the deadline (more details will be announced closer to the deadline).
- Turn in the code as well as the writeup.
- You can use any programming language you like.

Learning Objectives

By the end of this assignment, you will be able to:

- Derive and use **blocked** and **collapsed** Gibbs samplers for finite mixtures (Normal–Normal and Normal–Gamma).
- Integrate out parameters to obtain **cluster–factorized marginals** and **posterior predictives**, and explain their impact on label updates.
- Implement and analyze **Gaussian** mixture models.
- Write a concise, **reproducible** mini–report with diagnostics (e.g., traces/ACF).
- Brainstorm ideas for your final project.

Facts you may use without proof

1. **Normal–Normal marginal likelihood (integrate out μ).** If $x_1, \dots, x_n \mid \mu \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with known σ^2, and $\mu \sim \mathcal{N}(m_0, \tau^2)$, then

$$p(x_{1:n}) = (2\pi\sigma^2)^{-n/2} \left(\frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} \right)^{1/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \bar{x})^2 - \frac{n(\bar{x} - m_0)^2}{2(\sigma^2 + n\tau^2)} \right), \qquad \bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j.$$

2. **Single Normal–Gamma marginal is Student-t.** Let the prior be $\lambda \sim \text{Gamma}(a_0, b_0)$ (shape–rate) and $\mu \mid \lambda \sim \mathcal{N}\big(m_0, (\kappa_0\lambda)^{-1}\big)$. For a single observation,

$$x \mid \mu, \lambda \sim \mathcal{N}(\mu, \lambda^{-1}) \quad \Longrightarrow \quad p(x) \;=\; \iint \mathcal{N}(x \mid \mu, \lambda^{-1})\, \mathcal{N}(\mu \mid m_0, (\kappa_0\lambda)^{-1})\, \text{Gamma}(\lambda \mid a_0, b_0)\, d\mu\, d\lambda,$$

and the marginal is Student-t:

$$x \sim t_\nu\big(m_0,\; s^2\big), \qquad \nu = 2a_0, \quad s^2 = \frac{b_0}{a_0} \cdot \frac{\kappa_0 + 1}{\kappa_0}.$$

Gamma parameterization (shape–rate). We use $\text{Gamma}(a, b)$ with density $p(\lambda) = \dfrac{b^a}{\Gamma(a)}\, \lambda^{a-1} e^{-b\lambda}$ (shape a, rate b).

Warning: many libraries use shape–scale (scale = 1/rate). Verify the parametrization of the Gamma distribution when implementing later.

Question 1: A simple Finite Gaussian Mixture

Motivation. In this homework we will iterate through Box's Loop on the Gaussian mixture model (GMM): start from a very simple model, derive inference, check assumptions, and then add complexity. We begin with a minimal mixture for real-valued data: unknown component means, fixed known variances, and fixed equal mixing weights.

Comment. We will consider two inference strategies: a **blocked Gibbs sampler** (the "regular" Gibbs seen in class), which alternates sampling parameters and labels; and a **collapsed Gibbs sampler**, which analytically integrates out some parameters and samples labels directly. Collapsing reduces variance in conditionals and often yields faster mixing and better convergence; this is an instance of *Rao–Blackwellization*.

Model. Let $x_i \in \mathbb{R}$ for $i = 1, \cdots, N$ and fix the number of components $K \geq 2$.

$$z_i \mid K \sim \mathrm{Cat}\Big(\tfrac{1}{K}, \ldots, \tfrac{1}{K}\Big), \qquad i = 1, \ldots, N, \tag{1}$$

$$\mu_k \sim \mathcal{N}(m_0,\ \tau_0^2), \qquad k = 1, \ldots, K, \tag{2}$$

$$x_i \mid (z_i = k, \mu_k) \sim \mathcal{N}(\mu_k,\ \sigma_k^2), \qquad \sigma_k^2 > 0, \tag{3}$$

with fixed prior parameters $m_0 \in \mathbb{R}$ and $\tau_0^2 > 0$.

Notation. For each $k \in \{1, \ldots, K\}$ define

$$I_k = \{i : z_i = k\}, \qquad N_k = |I_k|, \qquad S_k^{(1)} = \sum_{i \in I_k} x_i, \qquad \bar{x}_k = \frac{S_k^{(1)}}{N_k} \ \text{(defined if } N_k \geq 1).$$

For single–site updates of z_i, use leave–one–out versions $N_{k,-i}$ and $S_{k,-i}^{(1)}$ computed on $I_k \setminus \{i\}$, with $\bar{x}_{k,-i} = S_{k,-i}^{(1)}/N_{k,-i}$ when $N_{k,-i} \geq 1$.

Part A: Derivations

(a) **Label update.** Derive the complete conditional $p(z_i = k \mid x_i, \mu_{1:K})$ and write it as a fully *normalized* categorical distribution over $k \in \{1, \ldots, K\}$.

(b) **Mean update.** Derive the complete conditional $p(\mu_k \mid x_{1:N}, z_{1:N})$. Your answer must be a univariate Normal with its mean and variance expressed in terms of $m_0, \tau_0^2, \sigma_k^2$ and the cluster summaries $(N_k, S_k^{(1)})$.

Part B: Collapsed labels (integrate out the means)

(c) **Cluster–factorized marginal.** By integrating out $\mu_{1:K}$ from the joint, derive $p(x_{1:N} \mid z_{1:N})$. Your final expression must factor across clusters, i.e., $p(x_{1:N} \mid z_{1:N}) = \prod_{k=1}^{K} p(X_k)$ with $X_k = \{x_i : z_i = k\}$, and be written in closed form using only cluster summaries and the prior hyperparameters.

(d) **Collapsed label update.** Using your result from (c), derive the single–site *collapsed* conditional $p(z_i = k \mid x_{1:N}, z_{-i})$. Write it as a normalized categorical over $k = 1, \ldots, K$ and treat both cases explicitly:

- assigning i to a *nonempty* cluster ($N_{k,-i} \geq 1$);
- assigning i to an *empty* cluster ($N_{k,-i} = 0$), in which case use the prior predictive induced by (2) and (3).

Part D: A graphical depiction

(e) Draw a graphical model using plate notation of this model using plate notation.

Question 2: Learning Variances and Weights (Theory only)

Motivation. We add complexity to the mixture: learn each component's variance and the mixing weights. You will (A) derive the blocked Gibbs conditionals and (B) integrate out component parameters to obtain collapsed label updates (keeping the weights explicit), then (C) reflect on what changed relative to Question 1.

Model. Let $x_{1:N} \in \mathbb{R}$ and fix the number of components $K \geq 2$.

$$\pi \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K), \tag{4}$$

$$z_i \mid \pi \sim \mathrm{Cat}(\pi_1, \ldots, \pi_K), \qquad i = 1, \ldots, N, \tag{5}$$

$$x_i \mid (z_i = k, \mu_k, \lambda_k) \sim \mathcal{N}(\mu_k,\ \lambda_k^{-1}), \tag{6}$$

$$\lambda_k \sim \mathrm{Gamma}(a_0, b_0), \tag{7}$$

$$\mu_k \mid \lambda_k \sim \mathcal{N}\big(m_0,\ (\kappa_0 \lambda_k)^{-1}\big), \qquad k = 1, \ldots, K, \tag{8}$$

with fixed prior parameters $\alpha_k > 0$, $a_0, b_0, \kappa_0 > 0$, and $m_0 \in \mathbb{R}$. We use the *shape–rate* parameterization for the Gamma: $\mathrm{Gamma}(a, b)$ has density $p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$.

Notation. For each k, let $I_k = \{i : z_i = k\}$, $N_k = |I_k|$,

$$S_k^{(1)} = \sum_{i \in I_k} x_i, \qquad S_k^{(2)} = \sum_{i \in I_k} x_i^2, \qquad \bar{x}_k = S_k^{(1)}/N_k \ \text{ (defined if } N_k \geq 1).$$

For single–site label updates, use leave–one–out summaries $N_{k,-i}, S_{k,-i}^{(1)}, S_{k,-i}^{(2)}$ computed on $I_k \setminus \{i\}$, and $\bar{x}_{k,-i} = S_{k,-i}^{(1)}/N_{k,-i}$ when $N_{k,-i} \geq 1$.

Part A: Blocked Gibbs (derive complete conditionals; normalized)

(a) **Weights.** Derive $p(\pi \mid z_{1:N})$ and write it in closed form.

(b) **Means.** Derive $p(\mu_k \mid \lambda_k,\ x_{1:N},\ z_{1:N})$. Your answer must be a univariate Normal with explicit mean and variance in terms of (m_0, κ_0) and $(N_k, \bar{x}_k)$.

(c) **Precisions.** Derive $p(\lambda_k \mid \mu_k,\ x_{1:N},\ z_{1:N})$. Your answer must be a Gamma distribution with explicit shape and rate in terms of $(a_0, b_0, \kappa_0, m_0)$ and $(N_k, S_k^{(1)}, S_k^{(2)}, \mu_k)$. *(Alternatively, you may first integrate out μ_k and derive $p(\lambda_k \mid x_{1:N}, z_{1:N})$; state clearly which route you chose.)*

(d) **Labels.** Derive the single–site conditional

$$p(z_i = k \mid x_i,\ \mu_{1:K},\ \lambda_{1:K},\ \pi)$$

and write it as a *normalized* categorical distribution over $k = 1, \ldots, K$.

Part B: Component–collapsed labels (integrate μ_k, λ_k; keep π explicit)

When updating z_i, define the current contents of cluster k excluding i by $D_{k,-i} = \{x_j : z_j = k,\ j \neq i\}$ with summaries $N_{k,-i}, S_{k,-i}^{(1)}, S_{k,-i}^{(2)}$ and $\bar{x}_{k,-i} = S_{k,-i}^{(1)}/N_{k,-i}$ when $N_{k,-i} \geq 1$.

(e) **Posterior for component k given $D_{k,-i}$ (up to proportionality).** Write $p(\mu_k, \lambda_k \mid D_{k,-i})$ up to a normalizing constant by combining (6)–(8) over $x \in D_{k,-i}$. Identify the updated constants (often denoted $(m_{k,-i}, \kappa_{k,-i}, a_{k,-i}, b_{k,-i})$) *in terms of* $D_{k,-i}$ and $(m_0, \kappa_0, a_0, b_0)$. Explicitly state how you handle the case $N_{k,-i} = 0$.

(f) **Posterior predictive under component k.** Define

$$p_{\text{pred}}(x_i \mid D_{k,-i}) \;=\; \iint \mathcal{N}(x_i \mid \mu_k, \lambda_k^{-1})\, p(\mu_k, \lambda_k \mid D_{k,-i})\, d\mu_k\, d\lambda_k.$$

Evaluate this integral in closed form. State explicitly how you treat $N_{k,-i} = 0$.

(g) **Component–collapsed label conditional.** Derive the complete conditional for the single site update that the single–site update is $p(z_i = k \mid x_{1:N},\ z_{-i},\ \pi)$.

Part C: What changed relative to Question 1? (short discussion)

In *3–6 sentences*, state precisely which formulas/conditionals differ from Q1 and give an intuitive explanation. At minimum, address:

- the appearance of π in the blocked label update and its effect
- the role of λ_k in the mean/label updates (uncertainty in variance);
- why the *collapsed* label update now uses a non-normal predictive (tail–heaviness when clusters are small), and how this differs from the Normal predictive in Q1.

Part D: A graphical depiction

(h) Draw a graphical model using plate notation of this model using plate notation.

Question 3: Implementation — Collapsed vs. Blocked Mixture Inference

Setup (choose one). Select a single modeling setting for implementation:

- **S1 (Fixed variances).** The Q1 model: $K \geq 2$ components, unknown means, *known* σ_k^2, and fixed known mixing weights.
- **S2 (Learned weights & variances).** The Q2 model: Dirichlet weights and Normal–Gamma components.

Task. For your chosen setting (S1 or S2), implement *both*:

(a) a **blocked** Gibbs sampler as specified by the complete conditionals from earlier questions;

(b) a **collapsed** label sampler obtained by integrating out the appropriate component parameters (cf. Question 2).

Data. Construct a synthetic dataset consistent with your chosen setting. State N, K, and the true parameter values used to generate the data.

What to hand in.

- A brief description of the model and hyperparameters used for the run(s).
- One figure visualizing the fitted clustering for your dataset.
- At least one diagnostic figure comparing the two samplers on your run(s).
- A discussion on the comparison of the behavior of the *collapsed* vs. *blocked* samplers for your setting.

Notes. Keep your writeup clear and self-contained.

Question 3.1: GMM Implementation & Mini-Report

Goal. Implement a Gaussian mixture on real data and write a short, academic-style report (as described below). The inference may use either a blocked or collapsed Gibbs sampler. You may work with an iid Gaussian mixture, as in Question 2, or extend the framework to a non-Gaussian mixture of your choice. The exercise is as much about clear exposition as it is about implementation: motivate your modeling decisions, specify the full generative model precisely, and present your inference approach and results in a concise mini academic report.

Choose one modeling track

- **Track A (iid Normal mixture)** on a real tabular dataset of your choice with learned variances and mixture proportions.
- **Track B (mixture for image segmentation with learned variances)** on an image (e.g., a favorite artwork).
- **Track C (your variant)**: any other finite non-Gaussian mixture model (e.g., using the senate voting data, one could try to fit a Bernoulli Mixture).

Your writeup should include:

i. **Introduction.** What problem are you solving? Why does a mixture solves this problem?

ii. **Model.** Write the full joint distribution for the generative model you use. State all distributions and parameterizations (e.g., Gamma shape–rate). Draw the graphical model.

iii. **Inference.** In the main text, summarize your inference approach and main implementation choices (details may go in an Appendix).

iv. **Data & setup.** Describe your dataset/image, preprocessing, choice of K, priors, etc. Compare the performance of your choices.

v. **Results.** Present the main outcomes of the mixture model: number of clusters found, posterior summaries of component parameters, and representative visualizations (e.g., scatter plots with cluster labels, segmentation maps). Include at least one simple quantitative check of model fit (such as held-out log predictive).

vi. **Diagnostics.** Provide at least one convergence diagnostic (e.g., traces of N_k and one mean/precision element) and briefly comment on mixing.

vii. **Discussion (short).** Interpret the results of the mixture model, how many clusters are there and why? Do you see any meaningful patterns among the clusters? etc,. Reflect on assumptions and trade-offs. One or two sentences on "what you would try next."

Formatting

2–4 pages total (not including figures and the Appendix, if applicable). Every figure needs a caption. This is a *soft limit* — but please don't go way over 4 pages.

Question 4: Final Project Ideas

This problem is intended to help you brainstorm ideas for the project. Consider some dataset of your choosing. If you have a data set in mind for your final project then we encourage you to use it for this exercise as well.

1. **Variables in the data.** What are the variables in the data and what are some of their relationships to each other? Do you expect some of the variables to be correlated? Do you expect others to be (conditionally) independent?

2. **Latent Variables:** What are some latent variables you could introduce to capture the correlations between model variables? What are some latent variables that could summarize aspects of the data? What other latent variables could be hidden in the data?

3. **Research Questions:** Formulate several questions you might be able to answer with the data. Write down the three most interesting ones.