Problem Set 4 - Machine Learning Theories

June 9, 2025

Problem 1
Consider the following two graphical models for continuous random variables $X_1, X_2, X_3 \in \mathbb{R}$ as shown in Figure 1.

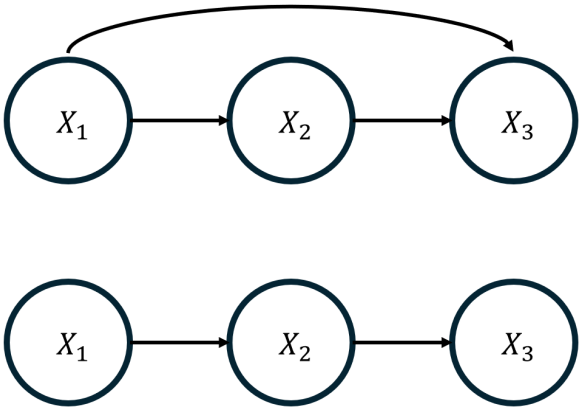

Figure 1: Two graphical models.

(1) Explain the difference between two directed graphical models in terms of the decomposed probability density functions.

(2) Can you find any (conditional) independence in each case? If you can, find the independence and explain what it means.

(3) Assuming X_1, X_2, X_3 are Gaussian random variables, find the minimum number of parameters needed to model the joint probability density function $p(X_1, X_2, X_3)$ for each graphical model.

Problem 2
Show that the conditional independence $x_1 \perp\!\!\!\perp x_2|x_3$ holds but not the independence $x_1 \not\perp\!\!\!\perp x_2$ (in general) for the following joint probability density functions:

(1) $p(x_1, x_2, x_3) = p(x_3)p(x_1|x_3)p(x_2|x_3)$

(2) $p(x_1, x_2, x_3) = p(x_1)p(x_3|x_1)p(x_2|x_3)$

Problem 3

Consider a state-space model where the latent state $x_t \in \mathbb{R}$ evolves according to a first-order linear dynamical system, but the observation model is *misspecified* due to an unknown additive bias. The objective is to reformulate the problem to estimate both the latent state and the bias using an augmented Kalman filter.

Consider the following scalar-valued state-space model:

$$x_t = \alpha x_{t-1} + \beta + \epsilon_{x,t}, \quad \epsilon_{x,t} \sim N(0, s_x^2),$$
$$y_t = \lambda x_t + \delta + \epsilon_{y,t}, \quad \epsilon_{y,t} \sim N(0, s_y^2),$$

where $\alpha, \beta, \lambda \in \mathbb{R}$ are known constants, $\delta \in \mathbb{R}$ is an *unknown constant bias* in the observation model, $\epsilon_{x,t}$ and $\epsilon_{y,t}$ are independent Gaussian noises, and the initial state $x_0 \sim N(\mu_0, \sigma_0^2)$ is known.

(1) Explain why the standard Kalman filter, which assumes an unbiased observation model $y_t = \lambda x_t + \epsilon_{y,t}$, will yield biased estimates for x_t. Using the observation update equations, discuss the consequence of this model misspecification on the posterior mean and variance for x_t, conditioned on the observations $\{y_0, \ldots, y_t\}$.

(2) To overcome the bias, we now reformulate the problem by introducing an augmented state vector:

$$\mathbf{x}_t = \begin{bmatrix} x_t \\ \delta_t \end{bmatrix},$$

where δ_t is the unknown bias, assumed to be a constant over time. We can use the following model: $\delta_t = \delta_{t-1} + \epsilon_{\delta,t}$, where $\epsilon_{\delta,t} \sim N(0, s_\delta^2)$ is a Gaussian noise independent to other noises. Write down the new state transition equation $\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{b} + \epsilon_t$ for $\mathbf{x}_t \in \mathbb{R}^2$. Clearly define the transition matrix $A \in \mathbb{R}^{2\times 2}$, the vector $\mathbf{b} \in \mathbb{R}^2$, and process noise covariance for $\epsilon_t \in \mathbb{R}^2$.

(3) Rewrite the observation equation y_t as a linear function of the augmented state $\mathbf{x}_t$, i.e., $y_t = C\mathbf{x}_t + \epsilon_{y,t}$. Identify the new observation matrix $C \in \mathbb{R}^{1\times 2}$ and the observation noise model.

(4) Assuming the initial augmented state $\mathbf{x}_0 \sim N(\mathbf{m}_0, P_0)$ with known mean $\mathbf{m}_0$ and covariance P_0, derive the Kalman filter update equations (i.e., time update and observation update) for this two-dimensional system.

Problem 4
Let P and Q be two probability distributions over a finite sample space $\mathcal{X}$, such that $P(x) > 0 \Rightarrow Q(x) > 0$ for all $x \in \mathcal{X}$. The Kullback–Leibler (KL) divergence from P to Q is defined as

$$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

Prove that $D_{\mathrm{KL}}(P \,\|\, Q) \geq 0$, with equality if and only if $P = Q$, using the following form of **Jensen's inequality**:

> Let ϕ be a convex function and X a random variable. Then
>
> $$\phi\left(\mathbb{E}[X]\right) \leq \mathbb{E}[\phi(X)].$$
>
> Equality holds if and only if X is almost surely constant or ϕ is linear on the support of X.

(1) Show that $\phi(x) = -\log x$ is convex on $(0, \infty)$.

(2) Apply Jensen's inequality to the random variable $X(x) = \frac{Q(x)}{P(x)}$ under the distribution P, and use this to prove that $D_{\mathrm{KL}}(P \,\|\, Q) \geq 0$.

(3) Explain when equality holds, and conclude that the KL divergence is zero if and only if $P = Q$.

Problem 5

Derive the closed-form expression for the following Kullback-Leibler divergence term between $p(z) = N(z|0, I)$ and $q(z) = N(z|\mu, \text{diag}(\sigma_1^2, \ldots, \sigma_D^2))$ for $z, \mu \in \mathbb{R}^D$:

$$KL(q(z)||p(z)) = \int q(z) \log \left(\frac{q(z)}{p(z)} \right) \, dz.$$

Problem 6
Derive the following expressions, which appear in the formulation of the diffusion models.

(1) Using the forward process equation $z_t = \sqrt{1-\beta_t} z_{t-1} + \sqrt{\beta_t}\epsilon_t$ for $t = 1, \ldots, T$ and the fact that $\epsilon_t, \epsilon_{t'} \sim N(0, I)$ are independent for $t \neq t'$, show that $q(z_t|x) = N(z_t|\sqrt{\alpha_t}x, (1-\alpha_t)I)$, where $\alpha_t = \prod_{\tau=1}^{t}(1-\beta_\tau)$ and $z_0 = x$.

(2) Show that, if z_{t-1} has zero mean and unit variance, the distribution of $z_t = \sqrt{1-\beta_t} z_{t-1} + \sqrt{\beta_t}\epsilon_t$ for $\epsilon_t \sim N(0, I)$ will also have zero mean and unit variance, irrespective of the value of β_t.

(3) Using the fact that $q(z_t|z_{t-1}) = N(z_t|\sqrt{1-\beta_t}z_{t-1}, \beta_t I)$ and $q(z_t|x) = N(z_t|\sqrt{\alpha_t}x, (1-\alpha_t)I)$ and the Bayes rule

$$q(z_{t-1}|z_t, x) = \frac{q(z_t|z_{t-1}, x)q(z_{t-1}|x)}{q(z_t|x)}, \tag{1}$$

show that $q(z_{t-1}|z_t, x) = N(z_{t-1}|m_t(x, z_t), \sigma_t^2 I)$, where $m_t(x, z_t) = \frac{(1-\alpha_{t-1})\sqrt{1-\beta_t}z_t + \sqrt{\alpha_{t-1}}\beta_t x}{1-\alpha_t}$ and $\sigma_t^2 = \frac{\beta_t(1-\alpha_{t-1})}{1-\alpha_t}$.
Furthermore, by using the expressions $z_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon_t$ and $\alpha_t = \prod_{\tau=1}^{t}(1-\beta_\tau)$, show that $m_t(x, z_t)$ can be reformulated as $m_t(x, z_t) = \frac{1}{\sqrt{1-\beta_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_t\right)$.
(Hint: use the Markov property $q(z_t|z_{t-1}, x) = q(z_t|z_{t-1})$ and complete the square with respect to z_{t-1} for the numerator $q(z_t|z_{t-1}, x)q(z_{t-1}|x))$ in (1).)

(4) Derive the following expression that appears during the derivation of the evidence lower bound (ELBO):

$$\begin{aligned}
&\mathbb{E}_q\left[\sum_{t=2}^{T}\log\frac{p(z_{t-1}|z_t, w)}{q(z_{t-1}|z_t, x)} + \log p(x|z_1, w)\right] \\
&= \int q(z_1|x)\log p(x|z_1, w)\ dz_1 - \sum_{t=2}^{T}\int KL(q(z_{t-1}|z_t, x)||p(z_{t-1}|z_t, w))q(z_t|x)\ dz_t,
\end{aligned}$$

where $\mathbb{E}_q[\cdot]$ is the expectation with respect to $q(z_1, \ldots, z_T|x)$.

(5) By using the definition of the KL-divergence $KL(q||p) = \int q(z)\log\left(\frac{q(z)}{p(z)}\right)\ dz$ and the facts that $q(z_{t-1}|z_t, x) = N(z_{t-1}|m_t(x, z_t), \sigma_t^2 I)$ and $p(z_{t-1}|z_t, w) = N(z_{t-1}|\mu(z_t, w, t), \beta_t I)$, show that

$$KL(q(z_{t-1}|z_t, x)||p(z_{t-1}|z_t, w)) = \frac{1}{2\beta_t}\|m_t(x, z_t) - \mu(z_t, w, t)\|^2 + const.$$