Homework 3

Deep Learning 2025 Spring

Due on 2025/3/24

1 True or False

Problem 1. If we optimize q_θ w.r.t. a multi-modal distribution p using KL-divergence $\mathrm{KL}(q_\theta \,\|\, p)$, we will get a distribution that uniformly covers all the modes.

Problem 2. The reparameterization trick applied in VAE helps passing gradient back to the encoder.

Problem 3. It is easy to compute the exact posterior $p(\mathsf{z}|\mathsf{x})$ using VAE.

Problem 4. In β-VAE, large β enforces latent variables to be correlated with each other.

2 Q&A

Problem 5. (EM Algorithm) In statistics, expectation–maximization (EM) algorithm is an iterative method to find (local) maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables.[1] Consider a latent variable model with parameter θ

$$p_\theta(\mathsf{x}, \mathsf{z}) = p_\theta(\mathsf{x}|\mathsf{z})p(\mathsf{z})$$

and we want to find the MLE of θ, i.e.,

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_\theta \log p_\theta(\mathsf{x}) = \arg\max_\theta \log \sum_z p_\theta(\mathsf{x}, \mathsf{z})$$

The E-step (expectation) of EM algorithm is given by

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{z\sim p_{\theta^{(t)}}(\mathsf{z}|\mathsf{x})}\left[\log p_\theta(\mathsf{x}, \mathsf{z})\right]$$

and the M-step (maximization) is given by

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)}).$$

Prove that the following optimization process is equivalent to EM algorithm. Define $F(\theta, q) = \mathbb{E}_{z\sim q}\left[\log p_\theta(\mathsf{x}, \mathsf{z})\right] + H(q)$, where $H(\cdot)$ is Shanon entropy.

[1] https://en.wikipedia.org/wiki/Expectation-maximization_algorithm

(E-step)

$$q^{(t)} = \arg\max_q F(\theta^{(t)}, q)$$

(M-step)

$$\theta^{(t)} = \arg\max_\theta F(\theta, q^{(t)})$$

Problem 6. (KL-Divergence)

1. (Gaussian) Prove that the KL-divergence between two d-dimensional Gaussian distributions $\mathcal{N}_0(\mu_0, \Sigma_0)$ and $\mathcal{N}_1(\mu_1, \Sigma_1)$ has the following form:

$$\mathrm{KL}(\mathcal{N}_0 \,\|\, \mathcal{N}_1) = \frac{1}{2}\left\{\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - d + \log\frac{|\Sigma_1|}{|\Sigma_0|}\right\}.$$

2. (Convexity) Let $\lambda \in [0, 1] \subset \mathbb{R}$. p_1, p_2, q_1 and q_2 are discrete distributions over alphabet $\mathcal{Y} = \{1, 2, \ldots, n\}$ with nonzero probabilities. Prove

$$\mathrm{KL}(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2) \le \lambda \mathrm{KL}(p_1 \,\|\, q_1) + (1-\lambda)\mathrm{KL}(p_2 \,\|\, q_2).$$

3. (Inclusive/Exclusive) We can recognize the difference of inclusive and exclusive KL via a simple example. Consider the target distribution

$$p(\mathsf{x}) = \frac{1}{3}\mathcal{N}(-3, 1) + \frac{2}{3}\mathcal{N}(3, 1)$$

which is a multi-modal Gaussian mixture. We model the variational distribution $q(\mathsf{x})$ as a Gaussian distribution with mean μ and variance σ^2, where μ and σ are unknown parameters. Write a program to find the optimal μ and σ w.r.t. inclusive and exclusive KL respectively. You need to submit a figure demonstrating the original distribution and two devrived variational distributions.

4. (Variational Inference) While we used *reverse KL-divergence* $\mathrm{KL}(q_\psi(z|x) \| p(z|x))$ to conduct variational inference (i.e., optimize the first term $q_\psi(z|x)$ with parameter ψ to approximate the second term $p(z|x)$) in lecture, Bob proposes to use the *forward KL-divergence* $\mathrm{KL}(p(z|x) \| q_\psi(z|x))$. In this case, what would be the objective for $q_\psi(z|x)$? What are the pros and cons if we use this objective for $q_\psi(z|x)$ in VAE?

   Hint: The objective should be in the form of expectation.

Problem 7. (GM-VAE) In standard VAEs, the prior of the latent variables is assumed to be an isotropic Gaussian. In this problem, we use a mixture of Gaussian distributions as the prior to allow more complicated latent representations, named Gaussian Mixture Variational Auto-Encoder (GM-VAE).

Consider a latent variable model $p_{\mu,\sigma,\theta}(\mathsf{x}, \mathsf{w}, \mathsf{z}) = p(\mathsf{z})p_{\mu,\sigma}(\mathsf{w}|\mathsf{z})p_\theta(\mathsf{x}|\mathsf{w})$, where an observable sample x is generated from latent variable w and z:

$$\mathsf{z} \sim \text{Categorical}(\pi),\ \mathbb{P}(\mathsf{z}=k)=\pi_k \text{ for } 1 \le k \le K \text{ and } \sum_{k=1}^{K} \pi_k = 1$$

$$\mathsf{w}|\mathsf{z} \sim \prod_{k=1}^{K} \mathcal{N}(\mu_k, \sigma_k^2 I)^{\mathbb{I}(\mathsf{z}=k)}$$

$$\mathsf{x}|\mathsf{w} \sim \mathcal{N}(\mu_\theta(\mathsf{w}), \sigma_\theta^2(\mathsf{w}))$$

where $\mu = [\mu_1, \ldots, \mu_K]$, $\sigma = [\sigma_1, \ldots, \sigma_K]$, and θ are trainable parameters. The prior distribution over z is uniform over alphabet $\{1, \ldots, K\}$. Define a variational model $q_{\psi,\phi}(\mathsf{w}, \mathsf{z}|\mathsf{x}) = q_\psi(\mathsf{w}|\mathsf{x}) q_\phi(\mathsf{z}|\mathsf{w}, \mathsf{x})$, where ψ and ϕ are trainable parameters.

1. Derive ELBO for $\log p_{\mu,\sigma,\theta}(\mathsf{x})$. Your answer should include 3 terms containing $p(\mathsf{z})$, $p_{\mu,\sigma}(\mathsf{w}|\mathsf{z})$ and $p_\theta(\mathsf{x}|\mathsf{w})$ respectively.
2. Design a training procedure for GM-VAE.