Due Date: April 18th at 23:00

Instructions

- For all questions that are not graded only on the answer, show your work! Any problem without work shown will get no marks regardless of the correctness of the final answer.
- Please try to use a document preparation system such as LaTeX. If you write your answers by hand, note that you risk losing marks if your writing is illegible without any possibility of regrade, at the discretion of the grader.
- Submit your answers electronically via the course GradeScope. Incorrectly assigned answers can be given 0 automatically at the discretion of the grader. To assign answers properly, please:
  - Make sure that the top of the first assigned page is the question being graded.
  - Do not include any part of the answer to any other questions within the assigned pages.
  - Assigned pages need to be placed in order.
  - For questions with multiple parts, the answers should be written in order of the parts within the question.
- Questions requiring written responses should be short and concise when necessary. Unnecessary wordiness will be penalized at the grader's discretion.
- Please sign the agreement below.
- It is your responsibility to follow updates to the assignment after release. All changes will be visible on Overleaf and Piazza.
- Any questions should be directed towards the TAs for this assignment (theoretical part): *Vitória Barin Pacela, Philippe Martin.*

I acknowledge I have read the above instructions and will abide by them throughout this assignment. I further acknowledge that any assignment submitted without the following form completed will result in no marks being given for this portion of the assignment.

Signature: _____

Name: _____

UdeM Student ID: _____

This assignment covers mathematical and algorithmic techniques in the families of deep generative models, such as VAEs, GANs, and contrastive learning.

Question 1 (5-5-6)**.** Consider a latent variable model $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})dz$, where $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ and $\boldsymbol{z} \in \mathbb{R}^K$. The encoder network (aka "recognition model") of a variational autoencoder, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, is used to produce an approximate (variational) posterior distribution over latent variables $\boldsymbol{z}$ for any input data point $\boldsymbol{x}$. [1] This distribution is trained to match the true posterior by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x})||p(\boldsymbol{z}))$$

Let $\mathcal{Q}$ be the family of variational distributions with a feasible set of parameters $\mathcal{P}$; i.e. $\mathcal{Q} = \{q(\boldsymbol{z}; \pi) : \pi \in \mathcal{P}\}$; for example, π can be the mean and standard deviation of a normal distribution. We assume q_ϕ is parameterized by a neural network (with parameters ϕ) that outputs the parameters, $\pi_\phi(\boldsymbol{x})$, of the distribution $q \in \mathcal{Q}$, i.e. $q_\phi(\boldsymbol{z}|\boldsymbol{x}) := q(\boldsymbol{z}; \pi_\phi(\boldsymbol{x}))$.

1.1 Show that maximizing the expected complete data log-likelihood (ECLL)

$$\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})]$$

for a fixed $q(\boldsymbol{z}|\boldsymbol{x})$, w.r.t. the model parameter θ, is equivalent to maximizing

$$\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$$

This means the maximizer of the ECLL coincides with that of the marginal likelihood only if $q(\boldsymbol{z}|\boldsymbol{x})$ perfectly matches $p(\boldsymbol{z}|\boldsymbol{x})$.

1.2 In this and the following task, the goal is to compare amortized variational inference (when q_ϕ is optimized for the whole dataset) with the traditional variational inference (when q_ϕ is optimized individually for each x). Consider a finite training set $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$, n being the size the training data. Let ϕ^* be the maximizer $\arg\max_\phi \sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ with θ fixed. In addition, for each $\boldsymbol{x}_i$ let $q_i \in \mathcal{Q}$ be an "instance-dependent" variational distribution, and denote by q_i^* the maximizer of the corresponding ELBO. Compare $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ and $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. Which one is bigger?

1.3 Following the previous question, compare the two approaches in the second subquestion

(a) in terms of bias of estimating the marginal likelihood via the ELBO, in the best-case scenario (i.e. when both approaches are optimal within the respective families)

(b) from the computational point of view (efficiency)

(c) in terms of memory (storage of parameters)

Question 2 (3-3-7-3-2-2)**.** The β-VAE [1] has been proposed as a follow-up from the VAE, and still is a relevant approach for learning a disentangled latent representation.

[1] Using a recognition model in this way is known as "amortized inference"; this can be contrasted with traditional variational inference approaches (see, e.g., Chapter 10 of Bishop's *Pattern Recognition and Machine Learning*), which fit a variational posterior independently for each new data point.

The Beta-VAE introduces a hyperparameter $\beta \in \mathbb{R}$, $\beta \geq 1$, which gives a higher weight to the regularization term (KL divergence between the real and the estimated posterior of z) in the VAE, as opposed to the reconstruction error. The loss of a β-VAE is given by:

$$\mathcal{L}^{\beta\text{VAE}}_{\boldsymbol{\theta},\boldsymbol{\phi},\beta}(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) + \beta \text{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z})) \tag{1}$$

where $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ is an isotropic Gaussian.

2.1 This question is about the role of the reconstruction error. What would be the optimal solution when $\beta = 0$?

2.2 This question is about the role of the KL divergence term. What would be the optimal solution if the regularization (KL divergence term) is optimized without any reconstruction error?

2.3 In a general autoencoder, let $X \sim p(X)$ be the true generating process for the input data X, $Y = f_\theta(X)$ be the hidden variable, and $p(X, Y)$ their associated joint.

Prove that $\max_\theta \mathbf{I}(X;Y) \geq \max_\theta \mathbf{H}(X) - l(X|Y)$, where $\mathbf{I}(X;Y)$ is the mutual information between X and Y, $\mathbf{H}(X)$ is the entropy of X, and $l(X|Y)$ is the autoencoder loss function for binary inputs.

2.4 Propose a term for the loss that could replace the reconstruction error, considering the tradeoffs from the previous questions. (Hint: SSL-related)

2.5 Let $z_a = f_a(x)$ and $z_b = f_b(x)$ be two different representations of a latent factor under different parameterizations from the same observed data x.

(a) Give an example of z_a, z_b, f_a, f_b that would satisfy $z_a \neq z_b$. Consider the simplest case possible, where $z_a, z_b \in \mathbb{R}^2$ and f_a, f_b are linear transformations.

(b) What does this suggest about the uniqueness of learning a latent variable from data x? How would you mathematically characterize the entire set of $z \in Z$ that could be reached from the same observed data x?

Question 3 (2-2-3)**.** In this question, we are concerned with analyzing the training dynamics of GANs. Consider the following value function

$$V(d, g) = dg \tag{2}$$

with $g \in \mathbb{R}$ and $d \in \mathbb{R}$. We will use this simple example to study the training dynamics of GANs.

1. Consider gradient descent/ascent with learning rate α as the optimization procedure to iteratively minimize $V(d, g)$ w.r.t. g and maximize $V(d, g)$ w.r.t. d. We will apply the gradient descent/ascent to update g and d simultaneously. What is the update rule of g and d? Write your answer in the following form

$$[d_{k+1}, g_{k+1}]^\top = A[d_k, g_k]^\top$$

where A is a 2×2 matrix; i.e. specify the value of A.

2. The optimization procedure you found in 6.1 characterizes a map which has a stationary point [2], what are the coordinates of the stationary points?

3. Analyze the eigenvalues of A and predict what will happen to d and g as you update them jointly. In other word, predict the behaviour of d_k and g_k as $k \to \infty$.

Question 4 (3-5-4-3)**.** Assume that we want to fit a data distribution $p_d(\boldsymbol{x})$ with a parameterized density model defined as

$$p_\theta(\boldsymbol{x}) = \frac{\exp(-E_\theta(\boldsymbol{x}))}{Z_\theta},$$

where $E_\theta(\boldsymbol{x})$ is the *energy* (i.e. an arbitrary function that maps $\boldsymbol{x}$ to a scalar in $\mathbb{R}$) parameterized with parameters θ and Z_θ is the normalizing constant:

$$Z_\theta = \int \exp(-E_\theta(x))dx.$$

The issue with learning $p_\theta(\boldsymbol{x})$ is the estimation of the normalizing constant Z_θ which may involve an intractable integral. However, several methods exist for learning unnormalized probabilistic models, also known as Energy Based Models (EBMs). The questions in this problem will explore some methods related to learning EBMs.

1. Noise Contrastive Estimation (NCE) is a method for learning the parameters of an EBM by contrasting it with another distribution with known density $p_n(\boldsymbol{x})$. $p_n(\boldsymbol{x})$ is usually chosen to be simple with a tractable PDF that we can easily sample from (e.g. a Gaussian distribution). NCE defines the learning objective as a binary classification problem:

   $$\mathcal{L}_{\text{NCE}}(E_\theta, p_d, p_n) := \mathbb{E}_{\boldsymbol{x}\sim p_d}[\log \sigma(E_\theta(\boldsymbol{x}))] + \mathbb{E}_{z\sim p_n}[\log(1 - \sigma(E_\theta(\boldsymbol{z})))],$$

   where σ is the sigmoid function. Assuming that E_θ is expressive enough, give a close form solution of the optimal energy $E_{\theta*}$ as a function of $p_n(\boldsymbol{x})$ and $p_d(\boldsymbol{x})$:

   $$E_{\theta*} = \arg\min_{E_\theta} \mathcal{L}_{\text{NCE}}(E_\theta, p_d, p_n). \tag{3}$$

2. Contrastive learning also learns EBMs via a classification objective. It does so by contrasting *positive* and *negative* pairs of samples. In this question, we will assume that we are given a datasets of tuples representing a positive pair of samples $(\boldsymbol{x}_i, \boldsymbol{x}'_i) \in \mathcal{D}$. The negative pairs of samples are two elements $(\boldsymbol{x}_i, \boldsymbol{x}'_j)$ sampled independently from $\mathcal{D}$ (i.e. $(\boldsymbol{x}_i, \cdot) \sim \mathcal{D}, (\cdot, \boldsymbol{x}'_j) \sim \mathcal{D}$). Let $f_\theta : \mathcal{X} \to S^d$ and $g_\phi : \mathcal{X}' \to S^d$ some mappings to the unit hyper-sphere, with $\boldsymbol{z} = f_\theta(\boldsymbol{x})$ and $\boldsymbol{y} = g_\phi(\boldsymbol{x}')$, we define the contrastive objective as follows:

   $$\mathcal{L}_{\text{contr.}}(f_\theta, g_\phi, \mathcal{D}) := \frac{1}{N}\sum_{i=1}^{N}\left[-\log \frac{e^{\boldsymbol{z}_i^\top \cdot \boldsymbol{y}_i}}{\sum_{j=1}^{M} e^{\boldsymbol{z}_i^\top \cdot \boldsymbol{y}_j}}\right], \tag{4}$$

   where M is the number of negative samples. Show that as as $M \to \infty$, equation 4 can be decomposed into three terms:

   $$\mathcal{L}_{\text{align}}(f_\theta, g_\theta, \mathcal{D}) := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i^\top \cdot \boldsymbol{y}_i,$$

[2] A stationary point is a point on the surface of the graph (of the function) where all its partial derivatives are zero (equivalently, the gradient is zero). Source: https://en.wikipedia.org/wiki/Stationary_point

$$\mathcal{L}_{\text{unif}}(f_\theta, g_\theta, \mathcal{D}) := \frac{1}{N}\sum_{i=1}^{N} \log \frac{1}{M}\sum_{j=1}^{M} e^{\boldsymbol{z}_i^\top \boldsymbol{y}_j}$$

and a constant $\log M$. What does each term represent? What happens as $M \to \infty$? Assume that the dimensionality of $\mathcal{X}$ is greater than d the dimensionality of the hyper-sphere and the underlying density of the distribution generating $\mathcal{D}$ is bounded[3]. Relate your answer to the information maximization principle.

3. The contrastive objective of equation 4 encodes positive and negative pairs of samples independently. Assume that we defined another network that jointly encode $\boldsymbol{x}$ and $\boldsymbol{x}'$: $h_\psi : \mathcal{X} \times \mathcal{X}' \to \mathcal{S}^d$. For this question, assume that $\boldsymbol{z} = h_\psi(\boldsymbol{x}, \boldsymbol{x}')$ and $\boldsymbol{y} = g_\phi(\boldsymbol{x}')$, where g_ϕ is the same mapping as defined in the previous question. Derive an objective function $\mathcal{L}_*(h_\psi, g_\theta, \mathcal{D})$ that have the same property as $\mathcal{L}_{\text{contr.}}$ that we discussed in the previous question.

4. Assume that f_θ, g_ϕ and h_ψ are all deep neural networks. Also, assume that we are training a vision-language encoders like CLIP. We can think of $\boldsymbol{x}$ as image tokens and $\boldsymbol{x}'$ as text tokens. In this case, f_θ would be an image encoder, g_ϕ would be a text encoder and h_ψ would be a joint image-text encoder. Discuss the trade-offs of learning an EBM with $\mathcal{L}_{\text{contr.}}$ and $\mathcal{L}_*$.

References

[1] Higgins, Irina and Matthey, Loïc and Pal, Arka and Burgess, Christopher P. and Glorot, Xavier and Botvinick, Matthew M. and Mohamed, Shakir and Lerchner, Alexander, *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*, ICLR, 2017.

[3] These are details and are not necessary needed for a complete answer.