Due Date: April 5th 23:59, 2019

Instructions

- *For all questions, show your work!*
- *Starred questions are **hard** questions, not **bonus** questions.*
- *Please use a document preparation system such as LaTeX, unless noted otherwise.*
- *Unless noted that questions are related, assume that notation and defintions for each question are self-contained and independent*
- *Submit your answers electronically via Gradescope.*
- ***TAs for this assignment are Shawn Tan, Samuel Lavoie, and Chin-Wei Huang.***

This assignment covers mathematical and algorithmic techniques underlying the three most popular families of deep generative models, variational autoencoders (VAEs, Questions 1-3), autoregressive models (Question 4), and generative adversarial networks (GANs, Questions 5-7).

Question 1 (8-8)**.** Reparameterization trick is a standard technique that makes the samples of a random variable differentiable. Consider a random vector $Z \in \mathbb{R}^K$ with a density function $q(\boldsymbol{z}; \phi)$. We want to find a deterministic function $\boldsymbol{g} : \mathbb{R}^K \to \mathbb{R}^K$ that depends on ϕ, to transform a random variable Z_0 having a ϕ-independent density function $q(\boldsymbol{z}_0)$, such that $\boldsymbol{g}(Z_0)$ has the same density as Z. Recall the change of density for a bijective, differentiable $\boldsymbol{g}$:

$$q(\boldsymbol{g}(\boldsymbol{z}_0)) = q(\boldsymbol{z}_0) \left| \det \left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} \right) \right|^{-1} \tag{1}$$

1. Assume $q(\boldsymbol{z}_0) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ and $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$, where $\mu \in \mathbb{R}^K$ and $\sigma \in \mathbb{R}^K_{>0}$. Show that $\boldsymbol{g}(\boldsymbol{z}_0)$ is distributed by $\mathcal{N}(\mu, \text{diag}(\sigma^2))$ using Equation (1).
2. Assume instead $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \boldsymbol{S}\boldsymbol{z}_0$, where $\boldsymbol{S}$ is a non-singular $K \times K$ matrix. Derive the density of $\boldsymbol{g}(\boldsymbol{z}_0)$ using Equation (1).

Answer 1.

1.

Question 2 (5-5-6)**.** Consider a latent variable model $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ where $\boldsymbol{z} \in \mathbb{R}^K$, and $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$. The encoder network (aka "recognition model") of 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 datapoint $\boldsymbol{x}$.[1] This distribution is trained to match the true posterior by maximizing the evidence lower bound (ELBO):

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

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 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] 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 an Machine Learning*), which fit a variational posterior independently for each new datapoint.

1. Show that maximizing the expected complete data log likelihood

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

for a fixed $q(\boldsymbol{z}|\boldsymbol{x})$, wrt the model parameter θ, gives the maximizer of the biased log marginal likelihood: $\arg\max_\theta\{\log p_\theta(\boldsymbol{x}) + B(\theta)\}$, where $B(\theta)$ is non-positive. Find $B(\theta)$.

2. Consider a finite training set $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$, n being the size the training data. Let ϕ^* be the maximizer of $\sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ 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 ?

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)

Answer 2.

1.

Question 3 (6-6)**.** Since variational inference provides a lower-bound on the log marginal likelihood of the data, it gives us a biased estimate of the marginal likelihood. Therefore, methods of "tightening" the bound (i.e. finding a higher valid lower bound) may be desirable.

Consider a latent variable model with the joint $p(\boldsymbol{x}, \boldsymbol{h})$ where $\boldsymbol{x}$ and $\boldsymbol{h}$ are the observed and unobserved random variables, respectively. Now let $q(\boldsymbol{h})$ be a variational approximation to $p(\boldsymbol{h}|\boldsymbol{x})$. Define

$$\mathcal{L}_K = \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\log \frac{1}{K}\sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]$$

Note that $\mathcal{L}_1$ is equivalent to the evidence lower bound (ELBO).

1. Show that $\mathcal{L}_K$ is a lower bound of the log marginal likelihood $\log p(\boldsymbol{x})$.
2. Show that $\mathcal{L}_K \geq \mathcal{L}_1$; i.e. $\mathcal{L}_K$ is a family of lower bounds tighter than the ELBO.

Answer 3.

1.

Question 4 (5-5-5-5)**.** One way to enforce autoregressive conditioning is via masking the weight parameters.[2] Consider a two-layer convolutional neural network without kernel flipping, with kernel size 3×3 and padding size 1 on each border (so that an input feature map of size 5×5 is convolved into a 5×5 output). Define mask of type A and mask of type B as

$$(\boldsymbol{M}^A)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j < 2 \\ 0 & \text{elsewhere} \end{cases} \qquad (\boldsymbol{M}^B)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

where the index starts from 1. Masking is achieved by multiplying the kernel with the binary mask (elementwise). Specify the receptive field of the output pixel that corresponds to the third row and the third column (index 33 of Figure 1) in each of the following 4 cases:

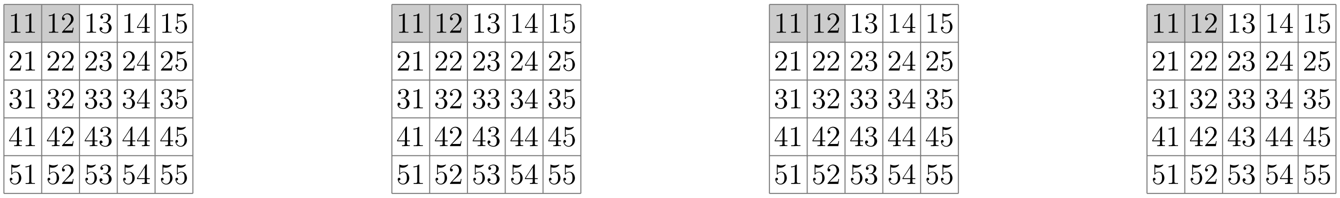
[2] An example of this is the use of masking in the Transformer architecture (Problem 3 of TP2 practical part).

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 1 – 5×5 convolutional feature map.

1. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
2. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer.
3. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
4. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer.

Answer 4. See Figure 2 for an example answer



▨ 11	▨ 12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

▨ 11	▨ 12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

▨ 11	▨ 12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

▨ 11	▨ 12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 2 – Receptive field under different masking schemes.

Question 5 (10)**.** Let P_0 and P_1 be two probability distributions with densities f_0 and f_1 (respectively). This problem demonstrates that a optimal GAN Discriminator (i.e. one which is able to distinguish between examples from P_0 and P_1 with minimal NLL loss) can be used to express the probability density of a datapoint $\boldsymbol{x}$ under f_1, $f_1(\boldsymbol{x})$ in terms of $f_0(\boldsymbol{x})$.

Assume f_0 and f_1 have the same support. Show that $f_1(\boldsymbol{x})$ can be estimated by $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1 - D(\boldsymbol{x}))$ by establishing the identity $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1 - D^*(\boldsymbol{x}))$, where

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim P_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_0}[\log(1 - D(\boldsymbol{x}))]$$

Answer 5.

Question 6 (5-5-6)**.** While generative adversarial networks were originally formulated as minimizing the Jensen-Shannon (JS)-divergence, the framework can be generalized to use other divergences, such as the Kullback–Leibler (KL)-divergence. In this exercise we see how KL can be approximated (bounded from below) via a function $T : \mathcal{X} \to \mathbb{R}$ (i.e. the discriminator). Let q and p be probability density functions and recall the definition of the KL divergence $D_{\mathrm{KL}}(p||q) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$.

*1. Let $R_1[T] := \mathbb{E}_p[T(x)] - E_q[e^{T(x)-1}]$.

   (a) The convex conjugate of a function $f(u)$ is defined as $f^*(t) = \sup_{u\in\mathrm{dom}f} ut - f(u)$. Show that the convex conjugate of $f(u) = u\log u$ is $f^*(t) = e^{t-1}$, and its biconjugate[3], i.e. the convex conjugate of its convex conjugate, is $f^{**}(u) := (f^*)^*(u) = u\log u$.

[3] More generally, the biconjugate of f is equal to itself if f is a lower semi-continuous convex function (this is known as the **Fenchel-Monreau Theorem**).

(b) Use the fact found above to show that $D_{\mathrm{KL}}(p||q) = \sup_T R_1[T]$, where the supremum is taken over the set of all (measurable) functions $\mathcal{X} \to \mathbb{R}$. Start from the following step

$$\sup_{T(x)} \int p(x)T(x) - q(x)e^{T(x)-1}dx = \int \sup_{t\in\mathbb{R}} p(x)t - q(x)e^{t-1}dx$$

which you don't need to prove.

*2. Let $r(x) = e^{T(x)}/\mathbb{E}_q[e^{T(x)}]$ and $R_2[T] := \mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}]$.

(a) Verify that rq is a proper density function, i.e. integrating to 1.

(b) Show that $D_{\mathrm{KL}}(p||q) \geq R_2[T]$, with equality if and only if $T(x) = \log(p(x)/q(x)) + c$ where c is a constant independent of x.

3. Compare the two representations of the KL divergence. For fixed $T(x)$, $p(x)$ and $q(x)$, which one of $R_1[T]$ and $R_2[T]$ is greater than or equal to the other ?

Answer 6.

Question 7 (10)**.** Let $q, p : \mathcal{X} \to [0, \infty)$ be probability density functions with disjoint (i.e. non-overlapping) support ; more formally, $\{x \in \mathcal{X} : p(x) > 0 \text{ and } q(x) > 0\} = \varnothing$. What is the Jensen Shannon Divergence (JSD) between p and q ? Recall that JSD is defined as $D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||r) + \frac{1}{2}D_{KL}(q||r)$ where $r(x) = \dfrac{p(x)+q(x)}{2}$.

Answer 7.