Due Date: Dec 8th (23:00 ET), 2023

Instructions

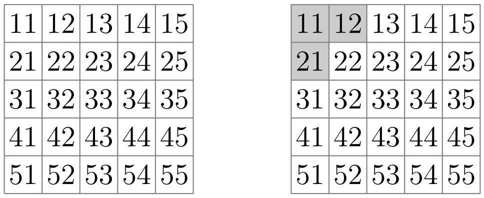
- *For all questions, show your work!*
- *Use LaTeX and the template we provide when writing your answers. You may reuse most of the notation shorthands, equations and/or tables. See the assignment policy on the course website for more details.*
- *The use of AI tools like Chat-GPT to find answers or parts of answers for any question in this assignment is not allowed. However, you can use these tools to improve the quality of your writing, like fixing grammar or making it more understandable. If you do use these tools, you must clearly explain how you used them and which questions or parts of questions you applied them to. Failing to do so or using these tools to find answers or parts of answers may result in your work being completely rejected, which means you'll receive a score of 0 for the entire theory or practical section.*
- *Submit your answers electronically via Gradescope.*
- *TAs for this assignment are* ***Thomas Jiralerspong, Sahar Dastani, and Shuo Zhang.***

Question 1 (5-5-5-5)**.** (**Autoregressive Models**)

One way to enforce autoregressive conditioning is via masking the weight parameters.[1] Consider a two-hidden-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 (Left)) in each of the following 4 cases:



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 1 – (Left) 5 × 5 convolutional feature map. (Right) Template answer.

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.

[1] An example of this is the use of masking in the Transformer architecture.

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.

Your answer should look like Figure 1 (Right).

Answer 1. See Figure 2

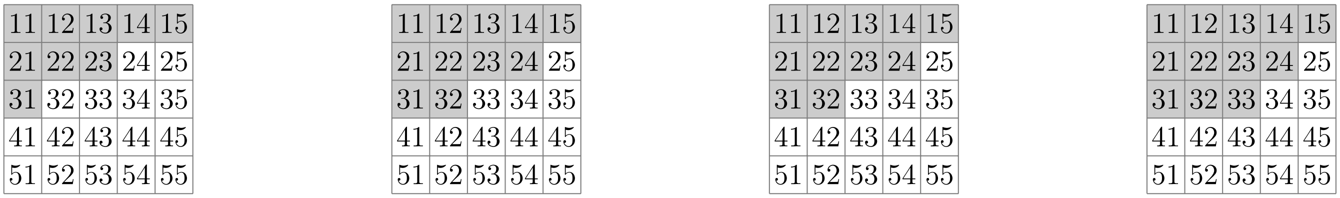


FIGURE 2 – Receptive field under different masking schemes.

Question 2 (5-5). (**Normalizing Flows**)

In this question, we study some properties of normalizing flows. Let $X \sim P_X$ and $U \sim P_U$ be, respectively, the distribution of the data and a base distribution (e.g. an isotropic gaussian). We define a normalizing flow as $F : \mathcal{U} \to \mathcal{X}$ parametrized by $\boldsymbol{\theta}$. Starting with P_U and then applying F will induce a new distribution $P_{F(U)}$ (used to match P_X). Since normalizing flows are invertible, we can also consider the distribution $P_{F^{-1}(X)}$.

However, some flows, like planar flows, are not easily invertible in practice. If we use P_U as the base distribution, we can only sample from the flow but not evaluate the likelihood. Alternatively, if we use P_X as the base distribution, we can evaluate the likelihood, but we will not be able to sample.

2.1 Show that $D_{KL}[P_X||P_{F(U)}] = D_{KL}[P_{F^{-1}(X)}||P_U]$. In other words, the forward KL divergence between the data distribution and its approximation can be expressed as the reverse KL divergence between the base distribution and its approximation.

2.2 Suppose two scenarios: 1) you don't have samples from $p_X(\boldsymbol{x})$, but you can evaluate $p_X(\boldsymbol{x})$, 2) you have samples from $p_X(\boldsymbol{x})$, but you cannot evaluate $p_X(\boldsymbol{x})$. For each scenario, specify if you would use the forward KL divergence $D_{KL}[P_X||P_{F(U)}]$ or the reverse KL divergence $D_{KL}[P_{F(U)}||P_X]$ as the objective to optimize. Justify your answer.

Answer 2. 2.1 We use the probability change of variable formula and perform a change of variable for the integral (and use the inverse function theorem):

$$\begin{aligned}
D_{KL}[p_X(\boldsymbol{x})||p_{F(U)}(\boldsymbol{x})] &= \int_{\mathcal{X}} p_X(\boldsymbol{x}) \log \frac{p_X(\boldsymbol{x})}{p_{F(U)}(\boldsymbol{x})} d\boldsymbol{x} \\
&= \int_{\mathcal{X}} p_X(\boldsymbol{x}) \log \frac{p_X(\boldsymbol{x})}{p_U(F^{-1}(\boldsymbol{x}))|\det J_{F^{-1}}(\boldsymbol{x})|} d\boldsymbol{x} \\
&= \int_{F^{-1}(\mathcal{X})} p_X(F(\boldsymbol{u}))|\det J_F(\boldsymbol{u})| \log \frac{p_X(F(\boldsymbol{u}))|\det J_F(\boldsymbol{u})|}{p_U(\boldsymbol{u})} d\boldsymbol{u} \\
&= \int_{F^{-1}(\mathcal{X})} p_{F^{-1}(X)}(\boldsymbol{u}) \log \frac{p_{F^{-1}(X)}(\boldsymbol{u})}{p_U(\boldsymbol{u})} d\boldsymbol{u} \\
&= D_{KL}[p_{F^{-1}(X)}(\boldsymbol{u})||p_U(\boldsymbol{u})]
\end{aligned}$$

Note that the equality also holds for the reverse: $D_{KL}[P_{F(U)}||P_X] = D_{KL}[P_U||P_{F^{-1}(X)}]$. In general, it is also true that f-divergences (the KL divergence is a particular case) are invariable to invertible transformation, meaning that $D_{KL}[P_X||P_Y] = D_{KL}[f(P_X)||f(P_Y)]$ if f is invertible.

2.2 The forward KL divergence is used when we have samples from the distribution P_X, but we cannot evaluate p_X. On the other hand, the reverse KL divergence is used when we can evaluate p_X. It becomes apparent if we expand the two terms. For the forward KL divergence, we have:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= D_{KL}[p_X(\boldsymbol{x})||p_{F(U)}(\boldsymbol{x})] \\ &= \mathbb{E}_{p_X(\boldsymbol{x})}[-\log p_{F(U)}(\boldsymbol{x})] + cst. \\ &= -\mathbb{E}_{p_X(\boldsymbol{x})}[\log p_U(F^{-1}(\boldsymbol{x})) + \log|J_{F^{-1}}(\boldsymbol{x})|] + cst.\end{aligned}$$

that only require $\boldsymbol{x}$. The second term is denoted as a constant since it does not depend on $\boldsymbol{\theta}$. For the reverse KL divergence, we have:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= D_{KL}[p_{F(U)}(\boldsymbol{x})||p_X(\boldsymbol{x})] \\ &= \mathbb{E}_{p_{F(U)}(\boldsymbol{x})}[\log p_{F(U)}(\boldsymbol{x}) - \log p_X(\boldsymbol{x})] \\ &= \mathbb{E}_{p_U(\boldsymbol{u})}[\log p_U(\boldsymbol{u}) - \log|J_F(\boldsymbol{u})| - \log p_X(F(\boldsymbol{u}))]\end{aligned}$$

that only require to evaluate p_X.

Question 3 (3-8-3-14). (**Variational Autoencoders**)

1. Let $p_x^*(.)$ be the true data distribution and $p_x(.;\theta)$ be the model distribution parametrized over θ, a natural criterion to define if $p_x(.;\theta)$ is accurately portraying $p_x^*(.)$ is the *Maximum Likelihood Estimation* (MLE). Sometimes, knowledge about the data can lead us to adopt a model with hidden intermediate variable z to approximate the data distribution, where only the joint distribution $p_{x,z}(.,.,\theta)$ are explicitly defined. For such models, we need to calculate the marginal likelihood $p_x(.) = \int_z p_{x,z}(.,z,\theta)dz$, however, this proves to be difficult. Why?
   (a) We do not know about $p_(x|z)$ and thus cannot calculate the integral.
   (b) Integration over the hidden variable z can prove to be intractable due to the complexity of $p_(x|z)$ and the curse of dimensionality.
   (c) We don't know and cannot assume what z looks like (i.e. what kind of distribution) and thus cannot calculate the integral.
   (d) The integral over the hidden variable z is intractable because it does not follow a standard distribution like Gaussian or Bernoulli.

2. To avoid the above problem, we can try to avoid $p_x(.)$ and instead aim to establish a lower bound function of it. This involves rewriting the log of the marginal likelihood $\log p_x(.) = \log\int_z p_{x,z}(.,z,\theta)dz$ as a combination of a KL divergence and an *Evidence Lower Bound* (ELBO). This process is facilitated by the introduction of an approximate posterior $q(z|x)$ which approximates the unknown true posterior $p(z|x)$. The choice of q is arbitrary, but we often choose it from simpler classes of distributions such as the Gaussian for practical reasons. Your task is to derive the ELBO function in two ways:

(a) By decomposing the marginal likelihood as the combination of a KL-divergence between variational and true posteriors over z ($D_{KL}(q(z|x)||p(z|x))$) and the ELBO.

(b) By using the Jensen Inequality.

3. What is the significance of the above result ? Select all that apply.

   (a) $p_x(.)$ has a lower bound which is the ELBO.

   (b) Maximizing the ELBO is equivalent to minimizing the distributional difference between the approximation $q(z|x)$ and the true (but intractable) $p(z|x)$.

   (c) The ELBO offers a theoretical bound but is not useful in practice for training models with latent variables.

   (d) The choice of q affects the tightness of the lower bound.

4. This question is about importance weighted autoencoder. When training a variational autoencoder, the standard training objective is to maximize the evidence lower bound (ELBO). Here we consider another lower bound, called the Importance Weighted Lower Bound (IWLB), a tighter bound than ELBO, defined as

$$\mathcal{L}_k = \mathbf{E}_{z_{1:k} \sim q(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{1}{k} \sum_{j=1}^{k} \frac{p(\boldsymbol{x}, z_j)}{q(z_j \mid \boldsymbol{x})} \right]$$

for an observed variable $\boldsymbol{x}$ and a latent variable $\boldsymbol{z}$, k being the number of importance samples. The model we are considering has joint that factorizes as $p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})$, $\boldsymbol{x}$ and $\boldsymbol{z}$ being the observed and latent variables, respectively. In the following questions, one needs to make use of the Jensen's inequality:

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)]$$

for a convex function f.

   (a) Show that IWLB is a lower bound on the log likelihood $\log p(\boldsymbol{x})$.

   (b) Given a special case where $k = 2$, prove that $\mathcal{L}_2$ is a tighter bound than the ELBO (with $k = 1$).

Answer 3.

1. (b)

2. (a)

$$\begin{aligned}
D_{KL}(q_\phi(z|x)||p_\theta(z|x)) &= \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} \\
&= \int_z q_\phi(z|x) \log \frac{q_\phi(z|x) p_\theta(x)}{p_\theta(x,z)} \\
&= \int_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(x,z)} + \int_z q_\phi(z|x) \log p_\theta(x) \\
&= \mathbb{E}_{q_\phi(z|x)}[\log \frac{q_\phi(z|x)}{p_\theta(z,x)}] + \log p_\theta(x) \\
&= \mathbb{E}_{q_\phi(z|x)}[\log q_\phi(z|x) - \log p_\theta(z,x)] + \log p_\theta(x) \\
\log p_\theta(x) &= D_{KL}(q_\phi(z|x)||p_\theta(z|x)) - \mathbb{E}_{q_\phi(z|x)}[\log q_\phi(z|x) - \log p_\theta(z,x)] \\
&= D_{KL}(q_\phi(z|x)||p_\theta(z|x)) + \underbrace{\mathbb{E}_{q_\phi(z|x)}[-\log q_\phi(z|x) + \log p_\theta(z,x)]}_{\text{ELBO}}
\end{aligned}$$

(b)

$$\begin{aligned}
\log p(x) &= \log \int_z p(x,z)\, dz \\
&= \log \int_z q(z|x) \frac{p(x,z)}{q(z|x)}\, dz \quad \text{(Multiply and divide by } q(z|x)\text{)} \\
&= \log \mathbb{E}_{q(z|x)} \left[\frac{p(x,z)}{q(z|x)} \right] \quad \text{(Expectation w.r.t. } q(z|x)\text{)} \\
&\geq \mathbb{E}_{q(z|x)} \left[\log \frac{p(x,z)}{q(z|x)} \right] \quad \text{(Using Jensen's Inequality)} \\
&= ELBO
\end{aligned}$$

3. (a), (b), (d)

4. (a) By Jensen's inequality, we can swap the expectation and logarithm:

$$\begin{aligned}
\mathcal{L}_k &= \mathbf{E}_{z_{1:k} \sim q(z|x)}[\log \frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)}] \\
&\leq \log \mathbf{E}_{z_{1:k} \sim q(z|x)}[\frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)}] \\
&= \log \mathbf{E}_{z \sim q(z|x)}[\frac{p(x,z)}{q(z|x)}] \\
&= \log \int_z q(z|x) \frac{p(x,z)}{q(z|x)} dz \\
&= \log \int_z p(x,z) = \log p(x)
\end{aligned}$$

(b) We make use of the finite form of the Jensen's inequality in the expectation.

$$\begin{aligned}\mathcal{L}_2 &= \mathbf{E}_{z_{1,2}\sim q(z|x)}[\log \frac{1}{2}(\frac{p(x,z_1)}{q(z_1|x)} + \frac{p(x,z_2)}{q(z_2|x)})] \\ &\geq \mathbf{E}_{z_{1,2}\sim q(z|x)}[\frac{1}{2}\log\frac{p(x,z_1)}{q(z_1|x)} + \frac{1}{2}\log\frac{p(x,z_2)}{q(z_2|x)}] \\ &= \mathbf{E}_{z\sim q(z|x)}[\log\frac{p(x,z)}{q(z|x)}] = \mathcal{L}_1\end{aligned}$$

As proven in the first question, $\mathcal{L}_k \leq \log p(x)$, so $\log p(x) \geq \mathcal{L}_2 \geq \mathcal{L}_1$. In other words, $\log p(x) - \mathcal{L}_2 \leq \log p(x) - \mathcal{L}_1$, so $\mathcal{L}_2$ is a tighter bound. One can also prove the general rule $\mathcal{L}_{k+1} \geq \mathcal{L}_k$, the proof of which can be found in the appendix of *Importance weighted autoencoders* (Burda et al. 2016).

Question 4 (2-2-2-3-3-10). (**Generative Adversarial Networks**)

1. Consider a Generative Adversarial Network (GAN) which successfully produces images of apples. Which of the following propositions is false ?
   (a) The generator aims to learn the distribution of apple images.
   (b) The discriminator can be used to classify images as apple vs. non-apple.
   (c) After training the GAN, the discriminator loss eventually reaches a constant value.
   (d) The generator can produce unseen images of apples.
2. Which of the following cost functions is the non-saturating cost function for the generator in GANs (G is the generator and D is the discriminator) ? Note that the cost function will be minimized w.r.t the generator parameters during training.
   (a) $J^{(G)} = \frac{1}{m}\sum_{i=1}^{m}\log(1 - D(G(z^{(i)})))$
   (b) $J^{(G)} = -\frac{1}{m}\sum_{i=1}^{m}\log(D(G(z^{(i)})))$
   (c) $J^{(G)} = \frac{1}{m}\sum_{i=1}^{m}\log(1 - G(D(z^{(i)})))$
   (d) $J^{(G)} = -\frac{1}{m}\sum_{i=1}^{m}\log(G(D(z^{(i)})))$
3. After training a neural network, you observe a large gap between the training accuracy (100%) and the test accuracy (42%). Which of the following methods is commonly used to reduce this gap ?
   (a) Generative Adversarial Networks
   (b) Dropout
   (c) Sigmoid activation
   (d) RMSprop optim
4. Given the two options of (A) saturating cost and (B) non-saturating cost, which cost function would you choose to train a GAN ? Explain your reasoning. (1-2 sentences)
5. You are training a standard GAN, and at the end of the first epoch you take note of the values of the generator and discriminator losses. At the end of epoch 100, the values of the loss functions are approximately the same as they were at the end of the first epoch. Why are the quality of generated images at epoch 1 and epoch 100 not necessarily similar ? (1-2 sentences)

6. Let p_0 and p_1 be two probability distributions with densities f_0 and f_1 (respectively). We want to explore what we can do with a trained GAN discriminator. A trained discriminator is thought to be one which is "close" to the optimal one:

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))].$$

   (a) For the first part of this problem, derive an expression we can use to estimate the Jensen-Shannon divergence (JSD) between p_0 and p_1 using a trained discriminator. We remind that the definition of JSD is $\mathrm{JSD}(p_0, p_1) = \frac{1}{2}\big(KL(p_0\|\mu) + KL(p_1\|\mu)\big)$, where $\mu = \frac{1}{2}(p_0 + p_1)$.

   (b) For the second part, we want to demonstrate 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 $f_0(\boldsymbol{x})$ [2]. 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 $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1 - D^*(\boldsymbol{x}))$.

   Hint: Find the closed form solution for D^*.

Answer 4.

1. (b)
2. (b)
3. (b)
4. Non-saturating cost is generally considered preferable, as the gradienttowards the beginning of training is larger.
5. You should not necessarily expect them to be the same since the losses are with respect to different quality models over time. That is, the loss of the generator at epochs 1 and 100 are with respect to a discriminator which might have significantly improved, and the same follows for the loss of the discriminator.
6. The given function to be maximized can be expressed as a functional of D, $G[D] := \int g(D(\boldsymbol{x}), \boldsymbol{x})d\boldsymbol{x}$ where

$$g(D(\boldsymbol{x}), \boldsymbol{x}) := f_1(\boldsymbol{x})\log D(\boldsymbol{x}) + f_0(\boldsymbol{x})\log(1 - D(\boldsymbol{x}))$$

   Setting the functional derivative to be zero yields

$$\frac{\delta G[D]}{\delta D} = \frac{\partial g(D(\boldsymbol{x}), \boldsymbol{x})}{\partial D} = \frac{f_1(\boldsymbol{x})}{D(\boldsymbol{x})} - \frac{f_0(\boldsymbol{x})}{1 - D(\boldsymbol{x})} = 0$$

   solving which gives us $D^*(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})}$.

   (a) Substituting back D^* we see that $G[D^*] = 2\mathrm{JSD}(p_0, p_1) - 2\log 2$ so: $\mathrm{JSD}(p_0, p_1) \approx \frac{1}{2}G[D] + \log 2$.

   (b) Thus we can use D^* to estimate the density of f_1 by rearranging the terms, which yields $f_1 = f_0D^*/(1 - D^*))$.

[2] You might need to use the "functional derivative" to solve this problem. See "19.4.2 Calculus of Variations" of the Deep Learning book or "Appendix D Calculus of Variations" of Bishop's Pattern Recognition and Machine Learning for more information.

Question 5 (5-5-5-5)**.** (**Self-Supervised Learning: Paper Review**)

In this question, you are going to write a **one page review** of the A Simple Framework for Contrastive Learning of Visual Representations paper.

Your review should have the following four sections: Summary, Strengths, Weaknesses, and Reflections. For each of these sections, below we provide a set of questions you should ask about the paper as you read it. Then, discuss your thoughts about these questions in your review.

(5.1) **Summary:**

(a) What is this paper about ?

(b) What is the main contribution ?

(c) Describe the main approach and results. Just facts, no opinions yet.

(5.2) **Strengths:**

(a) Is there a new theoretical insight ?

(b) Or a significant empirical advance ? Did they solve a standing open problem ?

(c) Or a good formulation for a new problem ?

(d) Any good practical outcome (code, algorithm, etc) ?

(e) Are the experiments well executed ?

(f) Useful for the community in general ?

(5.3) **Weaknesses:**

(a) What can be done better ?

(b) Any missing baselines ? Missing datasets ?

(c) Any odd design choices in the algorithm not explained well ? Quality of writing ?

(d) Is there sufficient novelty in what they propose ? Minor variation of previous work ?

(e) Why should anyone care ? Is the problem interesting and significant ?

(5.4) **Reflections:**

(a) How does this relate to other concepts you have seen in the class ?

(b) What are the next research directions in this line of work ?

(c) What (directly or indirectly related) new ideas did this paper give you ? What would you be curious to try ?

This question is subjective and so we will accept a variety of answers. You are expected to analyze the paper and offer your own perspective and ideas, beyond what the paper itself discusses.