Midterm Exam II
ECE 685D– Introduction to Deep Learning
Fall 2024

Instructor: Prof. Vahid Tarokh
ECE Department, Duke University

Nov 18th 2024
10:05 AM - 11:20 AM

Name: _____
NetID: _____

This exam contains 14 pages and 11 questions. There are 15 points bonus. This is a closed-book exam. No exam aids are permitted except for a one-sided letter-sized cheat sheet. Communication with others is strictly prohibited.

Distribution of Marks

Question	Points	Score
1	4	
2	4	
3	4	
4	5	
5	5	
6	5	
7	15	
8	23	
9	10	
10	20	
11	20	
Total:	115	

For these multiple-choice questions, circle ALL of the correct answers. These questions can have more than one correct answer.

1. (4 points) For a Generative Adversarial Network (GAN), what is the accuracy of the discriminator if the generator is perfectly trained (i.e., at its global optimum) with balanced real-fake data?
   (a) 100%
   (b) 50%
   (c) 33.33%
   (d) 0%

   Answer: (b)

2. (4 points) In generative modeling, why is mode collapse problematic?
   (a) It requires more training data.
   (b) It limits the diversity of generated outputs.
   (c) It makes the discriminator weak.
   (d) None of the above.

   Answer: (b)

3. (4 points) Consider a Vanilla GAN. When the discriminator becomes too powerful in the early stages of training, which of the following can happen?
   (a) The generator converges quickly.
   (b) The model achieves perfect accuracy.
   (c) The generator may struggle to improve.
   (d) None of the above.

   Answer: (c)

For these TRUE-FALSE questions, circle the correct statement.

4. (5 points) Given the following statements,

   (i) Restricted Boltzmann Machines (RBMs) are generative models with 3 types of nodes: visible, hidden, and output nodes.

   (ii) RBMs perform unsupervised learning.

   What is the correct answer?

   (a) Statement (i) is TRUE, Statement (ii) is FALSE.

   (b) Statement (i) is FALSE, Statement (ii) is TRUE.

   (c) Both statements are FALSE.

   (d) Both statements are TRUE.

   Answer: (b)

5. (5 points) Given the following statements,

   (i) Autoencoders perform supervised learning.

   (ii) Denoising Autoencoders train by adding noise to the output.

   What is the correct answer?

   (a) Statement (i) is TRUE, Statement (ii) is FALSE.

   (b) Statement (i) is FALSE, Statement (ii) is TRUE.

   (c) Both statements are FALSE.

   (d) Both statements are TRUE.

   Answer: (c)

6. (5 points) Given the following statements,

   (i) The optimal linear autoencoders (with both encoder and decoder linear transformations) have paired encoding and decoding matrices that can be constructed using Principle Component Analysis (PCA).

   (ii) Sparse Autoencoders use the L_2 norm to sparsify the latent variables.

   What is the correct answer?

   (a) Statement (i) is TRUE, Statement (ii) is FALSE.

   (b) Statement (i) is FALSE, Statement (ii) is TRUE.

   (c) Both statements are FALSE.

   (d) Both statements are TRUE.

   Answer: (a)

7. (15 points) (**Slow Feature Analysis**) Consider the following three-dimensional input signal $\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]$, where:

$$\begin{aligned} x_1(t) &= \cos(6t), \\ x_2(t) &= \sin(7t) - \sin(5t) + \cos(12t), \\ x_3(t) &= \sin(t). \end{aligned}$$

All components vary quickly, but hidden in these signals is a slowly varying feature. Let the slowly varying function be defined as:

$$f(t) = Ax_1(t) + Bx_2(t) + Cx_3(t) + Dx_1(t)x_3(t) + Ex_1^2(t)$$

where A, B, C, D, and E are scalar constants.

Find the values $\theta = [A, B, C, D, E] \in \mathbb{R}^5$ that make $f(t)$ non-zero and minimize the following objective function:

$$\mathcal{L}(f) = \frac{1}{6}\sum_{i=1}^{6}\left[f\left(\frac{\pi}{2}i\right) - f\left(\frac{\pi}{2}(i-1)\right)\right]^2$$

This page is intentionally left blank.

8. (**VAE**) Recall that the VAE optimizes the ELBO given as:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = E_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})],$$

and the Kullback–Leibler (KL) divergence is defined as:

$$D_{KL}\left[P||Q\right] = \int P(\mathbf{x}) \log\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) d\mathbf{x}.$$

In other words if $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ are sampled from the true distribution $p(\mathbf{x})$ then a VAE optimizes the following objective

$$\arg\max_{\theta,\phi} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\theta, \phi; \mathbf{x}_i)$$

Next, as in the lecture notes, we consider a VAE with the true data generating distribution $p(\mathbf{x})$, encoder $q_\phi(\mathbf{z}|\mathbf{x})$, decoder $p_\theta(\mathbf{x}|\mathbf{z})$ and latent variable distribution $p_z(\mathbf{z})$. If this VAE is trained well, then we intuitively expect that the two joint distributions $q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{z})p_z(\mathbf{z})$ are close to each other. To this end, let us consider $D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})||p_\theta(\mathbf{x}|\mathbf{z})p_z(\mathbf{z})\right]$.

(a) (6 points) Show that

$$D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})||p_\theta(\mathbf{x}|\mathbf{z})p_z(\mathbf{z})\right] = -E_{p(\mathbf{x})}\left[\mathcal{L}(\theta, \phi; \mathbf{x})\right].$$

(b) (5 points) Does the VAE trained as above empirically minimize

$$D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})||p_\theta(\mathbf{x}|\mathbf{z})p_z(\mathbf{z})\right]$$

over the parameters ϕ and θ?

(c) (12 points) Consider a one-dimensional data point $x \in \mathbb{R}$ and a one-dimensional latent variable $z \in \mathbb{R}$ with a standard normal, $p_z(z) \sim \mathcal{N}(0, 1)$. Assume that $q_\phi(z|x) \sim \mathcal{N}(\mu(x), x^2 + 4)$, where:

$$\mu(x) = w_1 x + b_1,$$

and $p_\theta(x|z) \sim \mathcal{N}(\nu(z), 1)$, with:

$$\nu(z) = w_2 z + b_2.$$

To train a VAE, a student has drawn a sample $x_1 = 0$ from $p(x)$ and has initialized the parameters as $w_1 = w_2 = 1, b_1 = b_2 = 0$. The student chooses a learning rate $\epsilon = 1$ and uses stochastic gradient ascent to maximize ELBO $\mathcal{L}(w_1, w_2, b_1, b_2; x_1)$. What is the new value of w_1 and w_2 when the student has completed the training? (You do not need to simplify your result).

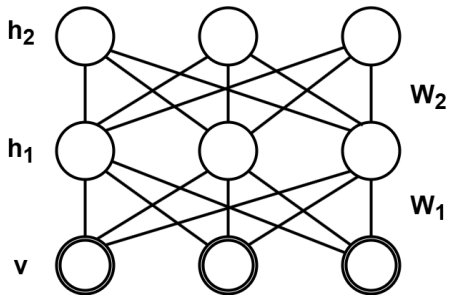
Hint: The KL divergence between $P : X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Q : X \sim \mathcal{N}(\mu_2, \sigma_2^2)$ is given by

$$D_{KL}\left[P||Q\right] = \frac{1}{2}\left[\frac{(\mu_2 - \mu_1)^2}{\sigma_2^2} + \frac{\sigma_1^2}{\sigma_2^2} - \ln\frac{\sigma_1^2}{\sigma_2^2} - 1\right]$$

This page is intentionally left blank.

This page is intentionally left blank.

9. (**Deep Boltzmann Machine**) Consider the Deep Boltzmann Machine in the figure below, with visible units $\mathbf{v} = [v_1, v_2, v_3] \in \{0,1\}^3$, and 2 equal-dimension layers of hidden units $\mathbf{h_1} = [h_{11}, h_{12}, h_{13}] \in \{0,1\}^3$, $\mathbf{h_2} = [h_{21}, h_{22}, h_{23}] \in \{0,1\}^3$.



The energy function $E(\mathbf{v}, \mathbf{h_1}, \mathbf{h_2})$ is defined as:

$$E(\mathbf{v}, \mathbf{h_1}, \mathbf{h_2}) = -\mathbf{v}^T \mathbf{W}_1 \mathbf{h}_1 - \mathbf{h}_1^T \mathbf{W}_2 \mathbf{h}_2,$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{3\times 3}$. In an expression of $\mathbf{a^T W b}$, a specific element of the weight matrix $W_{i,j}$ denotes the connection between node a_i and b_j. The joint probability of $\mathbf{v}, \mathbf{h_1}, \mathbf{h_2}$ is given as:

$$p(\mathbf{v}, \mathbf{h_1}, \mathbf{h_2}) = e^{-E(\mathbf{v}, \mathbf{h_1}, \mathbf{h_2})}/Z.$$

where $Z = \sum_{\mathbf{v}, \mathbf{h_1}, \mathbf{h_2}} e^{-E(\mathbf{v}, \mathbf{h_1}, \mathbf{h_2})}$ is the partition function.

Derive the following conditional probability density functions:

(a) (3 points) $p(v_3 = 1 | \mathbf{h}_1, \mathbf{h}_2)$ given $\mathbf{h}_1 = [1,1,1]$, $\mathbf{h}_2 = [1,0,1]$

(b) (4 points) $p(h_{11} = 1 | \mathbf{v}, \mathbf{h}_2)$ given $\mathbf{v} = [0,1,1]$, $\mathbf{h}_2 = [0,0,1]$

(c) (3 points) $p(h_{22} = 1 | \mathbf{v}, \mathbf{h}_1)$ given $\mathbf{v} = [1,0,0]$, $\mathbf{h}_1 = [1,0,1]$.

Write the answers in the weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$.

This page is intentionally left blank.

10. (**GAN**) Let $f(x)$, $x \in \mathbb{R}$, denote the true probability density function of the data X, defined as:

$$f(x) = \begin{cases} 2xe^{-x^2}, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0. \end{cases}$$

Consider a generator $G(\cdot)$ that takes the random variable $Z \sim \text{Unif}(0, e/2)$ and produces the fake random variable $\widehat{X} = G(Z)$ with $G(z) = z + 3$. Consider the Vanilla GAN's loss given by:

$$L(D, G) = E_{X \sim f(x)}[\log D(x)] + E_Z[\log(1 - D(G(Z)))]. \tag{1}$$

(a) (5 points) Show that the optimum discriminator is given by:

$$D(x) = \begin{cases} 1, & \text{for } x > 3 + e/2 \\ H(x), & \text{otherwise} \end{cases}$$

for some $H(x)$.

(b) (5 points) Give an explicit formula for $H(x)$.

(c) (10 points) Compute the probability of the discriminator declaring the input is fake when the data point is real and $D(x) < 0.5$.

This page is intentionally left blank.

11. **(Exponential Moving Average Regression)** Consider an Exponential Moving Average (EMA) process defined recursively by

$$y_t = \alpha x_t + (1-\alpha) y_{t-1}, \quad t \geq 1,$$

where $0 < \alpha < 1$ is an unknown smoothing parameter, $y_0 = 0$ is given, and $\{x_t\}$ are i.i.d. Gaussian random variables:

$$x_t \sim N(\mu, \sigma^2),$$

with μ and σ^2 known.

(a) (10 points) For observations $\{y_1, y_2\}$, express y_1 and y_2 in terms of α, μ, σ^2, and the underlying x_1, x_2. Write the joint density $p(y_1, y_2)$ in terms of α and the Gaussian density of x_t. Then write the corresponding log-likelihood function $\log p(y_1, y_2)$ as a function of α.

(b) (10 points) Derive the Maximum Likelihood Estimator (MLE) for α. You may leave the answer as the solution to an equation and do not need to compute it explicitly.

This page is intentionally left blank.