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:

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:

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:

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:

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:

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:

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:

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

Nice — this is a classic SFA exam trick: sample the inputs at the given times, expand the nonlinear basis (here the monomials given), then force the feature values to be constant across the sample times so the discrete-time derivative is zero (hence the SFA loss is minimized). The lecture explains exactly this "expand to nonlinear basis then solve a linear problem" approach. Lecture Part 11 LFM

Below is a compact, worked solution you can copy into an exam.

---

1) Sample the signals at the required time points

We need values at

$t_n = \frac{\pi}{2}n,\ n = 0, 1, \ldots, 6.$

Evaluate the basis functions at these t_n:

- $x_1(t) = \cos(6t) \Rightarrow x_{1,n} = \cos(6t_n) = \cos(3\pi n) = (-1)^n.$
- $x_3(t) = \sin t \Rightarrow x_{3,n} = \{0, 1, 0, -1, 0, 1, 0\}$ for $n = 0..6.$
- $\cos(12t) = \cos(6\pi n) = 1$ at all sample points.
- $\sin(7t)$ and $\sin(5t)$ reduce mod 2π; computing them gives the sequence below.

Make a small table (only distinct rows matter). the result:

n	x1 (=(-1)^n)	sin7t	sin5t	x2 = sin7t − sin5t + cos12t	x3	x1·x3	x1²
0	1	0	0	1	0	0	1
1	−1	−1	1	−1	1	−1	1
2	1	0	0	1	0	0	1
3	−1	1	−1	3	−1	1	1
4	1	0	0	1	0	0	1
5	−1	−1	1	−1	1	−1	1
6	1	0	0	1	0	0	1

(You can verify the sin(7t) and sin(5t) residues by reducing $7n$ and $5n$ modulo 4, since $\sin(k\pi/2)$ cycles every 4.)

2) Write f at the sampled points

Given

$$f(t) = Ax_1 + Bx_2 + Cx_3 + Dx_1x_3 + Ex_1^2,$$

at sample n denote f_n. Using the table:

$$f_0 = A + B + E.$$

$$f_1 = -A - B + C - D + E.$$

$$f_3 = -A + 3B - C + D + E.$$

(Indices 2,4,6 repeat the same values as index 0; index 5 repeats index 1.)

3) Require the discrete differences to be zero (so $L = 0$)

SFA discrete loss is zero iff all f_n are equal across the sampled times. So require $f_1 = f_0$ and $f_3 = f_0$. Subtracting:

From $f_1 - f_0 = 0$:

$$(-A - B + C - D + E) - (A + B + E) = -2A - 2B + C - D = 0.$$

From $f_3 - f_0 = 0$:

$$(-A + 3B - C + D + E) - (A + B + E) = -2A + 2B - C + D = 0.$$

4) Solve the linear system

Add the two equations:

$$(-2A - 2B + C - D) + (-2A + 2B - C + D) = -4A = 0 \Rightarrow A = 0.$$

Plugging $A = 0$ into either equation gives the same constraint:

$$-2B + C - D = 0 \quad \Longrightarrow \quad C = 2B + D.$$

So the solution set is

$$A = 0, \qquad C = 2B + D, \qquad B, D, E \text{ free},$$

with the only requirement that f is not identically zero (so not all B, D, E chosen to make constant zero).

5) Pick a simple non-zero solution (exam-style)

Take the simplest nontrivial choice, e.g.

$$B = 1,\ D = 0,\ E = 0 \quad \Rightarrow \quad C = 2.$$

So one valid answer is

$$\boxed{\theta = [A, B, C, D, E] = [0,\ 1,\ 2,\ 0,\ 0].}$$

This gives $f_n = Ax_1 + Bx_2 + Cx_3 + Dx_1x_3 + E = 0 + 1 \cdot x_2 + 2 \cdot x_3 + 0 + 0$. Plugging any sampled n from the table yields $f_n = 1$ for every n, so all differences vanish and $L(f) = 0$. That satisfies "non-zero" and minimizes the loss (zero is the global minimum here).

(There are infinitely many valid θ because B, D, E are free as long as $C = 2B + D$ and the overall feature isn't identically 0.)

Remark / Exam tip

- The key trick: evaluate the basis at the discrete sample times and convert the SFA objective to linear constraints on coefficients (expand nonlinear terms as additional basis functions). This is exactly the SFA method described in lecture: expand to a finite nonlinear function space and solve a linear problem for the slowest output. Lecture Part 11 LFM
- In exams, show the sampling table and then write the small linear system — graders like to see the explicit equations and one simple non-zero solution.

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.

(a) Derivation

We want to compute the KL between the two **joint** distributions

$$P(x,z) = q_\phi(z \mid x)\, p(x) \quad \text{and} \quad Q(x,z) = p_\theta(x \mid z)\, p_z(z),$$

i.e.

$$D_{KL}\big[\, q_\phi(z \mid x)p(x) \,\|\, p_\theta(x \mid z)p_z(z)\big] = \iint p(x)\, q_\phi(z \mid x)\, \log \frac{q_\phi(z \mid x)p(x)}{p_\theta(x \mid z)p_z(z)}\, dz\, dx.$$

Now expand the log and split the integrals:

$$\begin{aligned} D &= \iint p(x)q_\phi(z \mid x)\big(\log q_\phi(z \mid x) - \log p_\theta(x \mid z) - \log p_z(z) + \log p(x)\big)\, dz\, dx \\ &= \iint p(x)q_\phi(z \mid x) \log q_\phi(z \mid x)\, dz\, dx - \iint p(x)q_\phi(z \mid x) \log p_\theta(x \mid z)\, dz\, dx \\ &\quad - \iint p(x)q_\phi(z \mid x) \log p_z(z)\, dz\, dx + \iint p(x)q_\phi(z \mid x) \log p(x)\, dz\, dx. \end{aligned}$$

Since $\int q_\phi(z \mid x)\, dz = 1$, the last term reduces to

$$\iint p(x)q_\phi(z \mid x) \log p(x)\, dz\, dx = \int p(x) \log p(x)\, dx = \mathbb{E}_{p(x)}[\log p(x)],$$

which depends only on the data distribution $p(x)$ (a constant w.r.t. θ, ϕ).

Group the remaining terms: note

$$\int q_\phi(z \mid x) \log q_\phi(z \mid x)\, dz - \int q_\phi(z \mid x) \log p_z(z)\, dz = D_{KL}\big[q_\phi(z \mid x) \,\|\, p_z(z)\big].$$

Recognize the bracket as the ELBO

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

Thus we obtain the common identity

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

Important note (exam point): $\mathbb{E}_{p(x)}[\log p(x)]$ is a constant **independent** of θ, ϕ because it depends only on the true data distribution $p(x)$. So up to an additive constant we have

$$D_{KL}[q_\phi p] \;=\; -\mathbb{E}_{p(x)}[\mathcal{L}(\theta, \phi; x)] \;+\; \text{const}.$$

If your course/sheet writes the identity without the $\mathbb{E}_{p(x)}[\log p(x)]$ term, they are implicitly dropping that constant (because it doesn't affect optimization over θ, ϕ). Always mention this in an exam.

This page is intentionally left blank.

(b) Does training the VAE empirically minimize that KL over ϕ, θ?

Short answer (exam-ready):

Yes, up to an additive constant and up to approximation by the empirical data distribution. Maximizing the ELBO (or equivalently minimizing −ELBO) on the training set is equivalent to minimizing $D_{KL}[q_\phi(z|x)p(x) \,||\, p_\theta(x|z)p_z(z)]$ with respect to θ, ϕ, except for the additive constant $\mathbb{E}_{p(x)}[\log p(x)]$ (which does not depend on the parameters). In practice we replace the expectation over the true $p(x)$ by the empirical average over the training samples, so we minimize the empirical estimate of that KL.

Caveats you should state in an exam:

1. **Empirical vs true expectation:** We optimize the empirical average $\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\theta, \phi; x_i)$. That converges to $\mathbb{E}_{p(x)}[\mathcal{L}]$ only as $N \to \infty$ under usual conditions. So training minimizes the empirical surrogate of the true expected KL.
2. **Model/approximation limitations:** The equality above assumes exact optimization and flexible families. In practice:
   - $q_\phi(z|x)$ (encoder) is an approximate posterior family (gaussians with diagonal covariances usually) — this creates an **approximation gap**.
   - $p_\theta(x|z)$ (decoder) may not be expressive enough.
   - Optimization is nonconvex → we find local minima and use SGD; global optimum not guaranteed.
3. **Amortization gap:** Because we use a single encoder network shared across data points (amortized inference), there is an additional gap compared to optimizing per-data variational parameters.
4. **Optimization issues:** Vanishing KL (posterior collapse) or poor local optima can prevent achieving the theoretical minimum.

Final exam-summary (one-liner you can write on paper)

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

so maximizing expected ELBO (or maximizing empirical ELBO) **is equivalent** to minimizing that KL up to the constant $\mathbb{E}_{p(x)}[\log p(x)]$ and up to approximation/empirical sampling errors.

Setup (one-dim, $x = 0$)

Given:

- Prior $p_z(z) = \mathcal{N}(0, 1)$.
- Approximate posterior $q_\phi(z \mid x) = \mathcal{N}(\mu(x),\ \sigma^2(x))$ with $\mu(x) = w_1 x + b_1,\ \sigma^2(x) = x^2 + 4.$
- Likelihood $p_\theta(x \mid z) = \mathcal{N}(\nu(z), 1)$ with $\nu(z) = w_2 z + b_2$.
- Drawn sample $x_1 = 0$.
- Initial parameters: $w_1 = w_2 = 1,\ b_1 = b_2 = 0.$
- Learning rate $\epsilon = 1$.
- We run *one* SGD ascent step maximizing ELBO for that single sample.

For $x = 0$:

$$\mu = \mu(0) = w_1 \cdot 0 + b_1 = b_1, \qquad \sigma^2 = \sigma^2(0) = 0^2 + 4 = 4.$$

At the initialization $b_1 = 0$ so $\mu = 0$. Under q: $E[z] = \mu = 0,\ E[z^2] = \sigma^2 + \mu^2 = 4.$

ELBO for single x

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] \ - \ D_{KL}(q_\phi(z|x) \| p_z(z)).$$

We only need gradients of $\mathcal{L}$ w.r.t. w_2 and w_1.

Gradient w.r.t. w_2

$$\mathbb{E}_q[\log p_\theta(x|z)] = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\mathbb{E}_q[(x - w_2 z - b_2)^2].$$

Differentiate the expected quadratic term w.r.t. w_2:

$$\frac{\partial}{\partial w_2}\Big(-\tfrac{1}{2}\mathbb{E}[(x - w_2 z - b_2)^2]\Big) = \mathbb{E}\big[(x - w_2 z - b_2)\, z\big].$$

The KL term does **not** depend on w_2 (it depends on μ, σ^2 only), so

$$\frac{\partial \mathcal{L}}{\partial w_2} = \mathbb{E}_q\big[(x - w_2 z - b_2) z\big].$$

Now plug the numbers at initialization: $x = 0,\ b_2 = 0,\ w_2 = 1$, and $E[z^2] = 4$:

$$\left.\frac{\partial \mathcal{L}}{\partial w_2}\right|_{\text{init}} = \mathbb{E}[(0 - 1 \cdot z - 0)\, z] = \mathbb{E}[-z^2] = -4.$$

Gradient-ascent update with $\epsilon = 1$:

$$w_2^{\text{new}} = w_2^{\text{old}} + \epsilon \cdot \frac{\partial \mathcal{L}}{\partial w_2} = 1 + 1 \cdot (-4) = -3.$$

Gradient w.r.t. w_1

w_1 enters only through $\mu(x) = w_1 x + b_1$. For our sample $x = 0$,

$$\frac{\partial \mu}{\partial w_1} = x = 0.$$

Hence the chain rule gives

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial \mu} \cdot \frac{\partial \mu}{\partial w_1} = (\cdots) \cdot 0 = 0.$$

So w_1 does not change in this update:

$$w_1^{\text{new}} = w_1^{\text{old}} = 1.$$

(If you had asked about b_1, it would change because $\partial \mu / \partial b_1 = 1$.)

Final answer

After the SGD ascent step (learning rate 1) on the single sample $x = 0$:

$$\boxed{w_1 = 1, \qquad w_2 = -3.}$$

✅ Where the term $-\frac{1}{2}\log(2\pi)$ comes from

For a Gaussian with **unit variance**:

$$p_\theta(x \mid z) = \mathcal{N}(x; \nu(z), 1)$$

the probability density is:

$$p(x \mid z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \nu(z))^2\right).$$

Taking logarithm:

$$\log p(x \mid z) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(x - \nu(z))^2.$$

This is a **standard identity**:

$$\log \mathcal{N}(x; \mu, 1) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(x - \mu)^2.$$

When computing the **reconstruction term**:

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

if $p_\theta(x|z) = \mathcal{N}(\nu(z), 1)$, then:

$$\mathbb{E}_{q(z|x)}[\log p_\theta(x|z)] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_q[(x - \nu(z))^2].$$

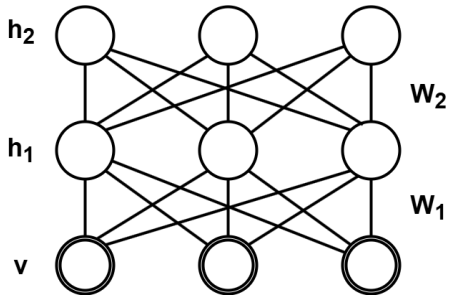
The first part:

$$-\frac{1}{2}\log(2\pi)$$

is a constant; it does **not depend on** w_1, w_2, b_1, b_2.

So we usually ignore it when taking gradients (but it is still part of the complete formula).

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 $\mathbf{W}_2$.

Quick derivation (one-line)

Energy:

$$E(\mathbf{v}, \mathbf{h}_1, \mathbf{h}_2) = -\mathbf{v}^\top W_1 \mathbf{h}_1 - \mathbf{h}_1^\top W_2 \mathbf{h}_2.$$

For a binary unit $x \in \{0, 1\}$ with Markov blanket inputs giving total input s, the conditional is

$$P(x = 1 \mid \text{neighbors}) = \sigma(s) \quad \text{with } \sigma(s) = \frac{1}{1 + e^{-s}}.$$

The total input s is the sum of weights from neighbors multiplied by their states (this is standard for Boltzmann machines / RBMs). Lecture Part 14 RBM

Notation / indexing reminder

- $W_1 \in \mathbb{R}^{3\times 3}$ with entries $W_{1_{ij}}$ connecting visible v_i to h_{1j}.
  The energy term $-\mathbf{v}^\top W_1 \mathbf{h}_1$ expands as $-\sum_{i,j} v_i W_{1_{ij}} h_{1j}$.
- $W_2 \in \mathbb{R}^{3\times 3}$ with entries $W_{2_{jk}}$ connecting h_{1j} to h_{2k}.
  The energy term $-\mathbf{h}_1^\top W_2 \mathbf{h}_2 = -\sum_{j,k} h_{1j} W_{2_{jk}} h_{2k}$.

So:

- input to visible v_i from $\mathbf{h}_1$ is $s_{v_i} = \sum_j W_{1_{ij}} h_{1j}$.
- input to hidden h_{1j} from $\mathbf{v}$ and $\mathbf{h}_2$ is $s_{h_{1j}} = \sum_i v_i W_{1_{ij}} + \sum_k W_{2_{jk}} h_{2k}$.
- input to hidden h_{2k} from $\mathbf{h}_1$ is $s_{h_{2k}} = \sum_j h_{1j} W_{2_{jk}}$.

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

Visible unit v_3 gets input from all h_{1j} via row $i = 3$ of W_1:

$$s_{v_3} = \sum_{j=1}^{3} W_{1_{3j}}\, h_{1j} = W_{1_{31}} + W_{1_{32}} + W_{1_{33}} \quad (\text{since } h_{1j} = 1\ \forall j).$$

Hence

$$\boxed{p(v_3 = 1 \mid \mathbf{h}_1, \mathbf{h}_2) = \sigma\left(W_{1_{31}} + W_{1_{32}} + W_{1_{33}}\right).}$$

(Note: v_3 does **not** directly depend on h_2 in the energy — only via h_1.)

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

Here h_{11} is h_1 with index $j = 1$. Its total input is

$$s_{h_{11}} = \sum_{i=1}^{3} v_i \, W_{1_{i1}} \ + \ \sum_{k=1}^{3} W_{2_{1k}} \, h_{2k}.$$

Plugging the given values:

$$\sum_i v_i W_{1_{i1}} = 0 \cdot W_{1_{11}} + 1 \cdot W_{1_{21}} + 1 \cdot W_{1_{31}} = W_{1_{21}} + W_{1_{31}},$$

and

$$\sum_k W_{2_{1k}} h_{2k} = W_{2_{11}} \cdot 0 + W_{2_{12}} \cdot 0 + W_{2_{13}} \cdot 1 = W_{2_{13}}.$$

So total

$$s_{h_{11}} = W_{1_{21}} + W_{1_{31}} + W_{2_{13}}.$$

Thus

$$\boxed{p(h_{11} = 1 \mid \mathbf{v}, \mathbf{h}_2) = \sigma\left(W_{1_{21}} + W_{1_{31}} + W_{2_{13}}\right).}$$

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

Unit h_{22} is h_2 with index $k = 2$. Its conditional depends only on $\mathbf{h}_1$ (not directly on $\mathbf{v}$):

$$s_{h_{22}} = \sum_{j=1}^{3} h_{1j} \, W_{2_{j2}}.$$

Given $\mathbf{h}_1 = [1, 0, 1]$:

$$s_{h_{22}} = 1 \cdot W_{2_{12}} + 0 \cdot W_{2_{22}} + 1 \cdot W_{2_{32}} = W_{2_{12}} + W_{2_{32}}.$$

Therefore

$$\boxed{p(h_{22} = 1 \mid \mathbf{v}, \mathbf{h}_1) = \sigma\left(W_{2_{12}} + W_{2_{32}}\right).}$$

Final remarks

- Each conditional is the sigmoid of the sum of the connecting-weight entries multiplied by neighbor states (Markov blanket). This is precisely the RBM / binary Boltzmann conditional form; see the RBM lecture slides for the conditional logistic formula and Markov-blanket argument. Lecture Part 14 RBM

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 $D(x) < 0.5$.

Quick facts (from the problem)

- Real data density
$$f(x) = \begin{cases} 2xe^{-x^2}, & x \geq 0 \\ 0, & x < 0 \end{cases}.$$
- Generator: $Z \sim \text{Unif}(0, \frac{e}{2})$, $G(z) = z + 3$.
So the *model* (fake) random variable $\tilde{X} = G(Z)$ is uniform on $[3,\ 3 + \frac{e}{2}]$.
- Therefore the model (fake) density is

$$p_{\text{model}}(x) = \begin{cases} \dfrac{2}{e}, & x \in \left[3,\ 3 + \frac{e}{2}\right], \\ 0, & \text{otherwise.} \end{cases}$$

(a) Optimum discriminator $D^*(x)$

For vanilla GAN (fixed G), the optimal discriminator is the well-known ratio

$$D^*(x) = \frac{p_{\text{real}}(x)}{p_{\text{real}}(x) + p_{\text{model}}(x)}.$$

(Reference: lecture slides / Goodfellow et al.). Lecture Part 16 GAN

Now observe that whenever $p_{\text{model}}(x) = 0$ we get $D^*(x) = 1$. The model support is exactly $[3, 3 + \frac{e}{2}]$, so outside that interval $p_{\text{model}}(x) = 0$ and $D^*(x) = 1$. Inside the overlap $[3, 3 + \frac{e}{2}]$ both densities are positive and the formula above applies.

So one can write

$$D^*(x) = \begin{cases} 1, & x \notin [3,\ 3 + \frac{e}{2}], \\ \dfrac{f(x)}{f(x) + \frac{2}{e}}, & x \in [3,\ 3 + \frac{e}{2}]. \end{cases}$$

(Which matches the exam's statement that $D(x) = 1$ where the model density is zero, and $H(x)$ elsewhere.)

(b) Explicit formula for $H(x)$

From the general formula, for $x \in [3, 3 + \frac{e}{2}]$,

$$H(x) \equiv D^*(x) = \frac{f(x)}{f(x) + p_{\text{model}}(x)} = \frac{2xe^{-x^2}}{2xe^{-x^2} + \frac{2}{e}} = \frac{xe^{-x^2}}{xe^{-x^2} + \frac{1}{e}}.$$

You can also write it as the equivalent form

$$H(x) = \frac{2xe^{-x^2}}{2xe^{-x^2} + \frac{2}{e}} = \frac{xe^{-x^2}}{xe^{-x^2} + e^{-1}}.$$

(And, recall: for x outside the model support the correct value is $D^*(x) = 1$.)

(c) Probability that the discriminator declares *fake* when the data point is real and $D(x) < 0.5$

We want

$$\Pr_{X\sim f}\left(D^*(X) < 0.5\right) = \int_{\{x:\, D^*(x)<0.5\}} f(x)\,dx.$$

Inside the overlap region $x \in [3, 3+\frac{e}{2}]$,

$$D^*(x) < 0.5 \iff \frac{f(x)}{f(x)+\frac{2}{e}} < \frac{1}{2} \iff f(x) < \frac{2}{e}.$$

But note $f(x) = 2xe^{-x^2}$. For any $x \geq 3$ we have $2xe^{-x^2}$ extremely small compared with $2/e$ (numerically $2/e \approx 0.736$, while $f(3) = 6e^{-9} \approx 7.4 \times 10^{-4}$). Thus for the whole overlap interval $[3, 3+\frac{e}{2}]$ the inequality $f(x) < 2/e$ holds — hence $D^*(x) < 0.5$ for every x in the model support.

Outside the overlap $p_{\text{model}}(x) = 0$ so $D^*(x) = 1$ and therefore those x do **not** satisfy $D^*(x) < 0.5$.

So the set $\{x : D^*(x) < 0.5\}$ is exactly the overlap interval $[3,\ 3+\frac{e}{2}]$. Therefore

$$\Pr_{X\sim f}\left(D^*(X) < 0.5\right) = \int_3^{3+\frac{e}{2}} 2xe^{-x^2}\,dx.$$

This integral is straightforward:

$$\int 2xe^{-x^2}\,dx = -e^{-x^2} + C.$$

Hence

$$\Pr\left(D^*(X) < 0.5\right) = \Big[-e^{-x^2}\Big]_{x=3}^{x=3+\frac{e}{2}} = e^{-9} - e^{-\left(3+\frac{e}{2}\right)^2}.$$

Numeric approximation

- $e^{-9} \approx 1.234 \times 10^{-4}$.
- $\left(3+\frac{e}{2}\right) \approx 4.35914$, so $\exp\left(-(3+\frac{e}{2})^2\right)$ is about 5.6×10^{-9}, negligible here.

Therefore

$$\Pr\left(D^*(X) < 0.5\right) \approx 1.234 \times 10^{-4}\ (\approx 0.0001234).$$

Short summary you can write in the exam

1. State the general optimum discriminator: $D^*(x) = \dfrac{p_{\text{real}}(x)}{p_{\text{real}}(x)+p_{\text{model}}(x)}$. Lecture Part 16 GAN
2. Note model support $[3, 3+\frac{e}{2}]$ so $p_{\text{model}}(x) = 2/e$ on that interval, zero elsewhere. Conclude $D^*(x) = 1$ where $p_{\text{model}}(x) = 0$, and $H(x) = \dfrac{xe^{-x^2}}{xe^{-x^2}+e^{-1}}$ for $x \in [3, 3+\frac{e}{2}]$.
3. Compute $\Pr(D^*(X) < 0.5) = \int_3^{3+e/2} 2xe^{-x^2}dx = e^{-9} - e^{-(3+e/2)^2} \approx 1.234 \times 10^{-4}$.

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.

(a) Express y_1, y_2, then $p(y_1, y_2)$ and $\log p(y_1, y_2)$

Step 1: Unroll the recursion

For $t = 1$:

$$y_1 = \alpha x_1 + (1-\alpha) y_0 = \alpha x_1$$

For $t = 2$:

$$\begin{aligned} y_2 &= \alpha x_2 + (1-\alpha) y_1 \\ &= \alpha x_2 + (1-\alpha)(\alpha x_1) \\ &= \alpha x_2 + \alpha(1-\alpha) x_1 \end{aligned}$$

So

$$\boxed{y_1 = \alpha x_1, \qquad y_2 = \alpha x_2 + \alpha(1-\alpha) x_1}$$

Step 2: Invert to get x_1, x_2 in terms of y_1, y_2, α

From $y_1 = \alpha x_1$:

$$x_1 = \frac{y_1}{\alpha}$$

From $y_2 = \alpha x_2 + \alpha(1-\alpha) x_1$:

$$\begin{aligned} y_2 &= \alpha x_2 + \alpha(1-\alpha) x_1 \\ \frac{y_2}{\alpha} &= x_2 + (1-\alpha) x_1 \\ x_2 &= \frac{y_2}{\alpha} - (1-\alpha) x_1 = \frac{y_2}{\alpha} - (1-\alpha)\frac{y_1}{\alpha} = y_1 + \frac{y_2 - y_1}{\alpha} \end{aligned}$$

So

$$\boxed{x_1 = \frac{y_1}{\alpha}, \qquad x_2 = y_1 + \frac{y_2 - y_1}{\alpha}}$$

⭐ Step 3 Goal

We want **$p(y_1, y_2)$**.

But we only know the density of **x_1, x_2** (Gaussian).

So we convert between them:

$$(y_1, y_2) \longleftrightarrow (x_1, x_2)$$

Because:

- We know **x_1, x_2 are Gaussian**
- y's are **deterministic functions of x's**

So the density transformation follows:

$$p_Y(y_1, y_2) = p_X(x_1, x_2) \cdot |\det(J)|$$

This is the **change-of-variables formula**.

⭐ Why do we need the Jacobian?

Because when you change variables, you need to account for how the space stretches.

But—good news—**in this problem, the transformation is linear**, so the Jacobian is very simple.

⭐ Step 3 explained in 4 simple steps

Step 3.1: Express x_1 and x_2 in terms of y_1 and y_2

We already found:

$$x_1 = \frac{y_1}{\alpha}$$

$$x_2 = y_1 + \frac{y_2 - y_1}{\alpha}$$

This makes (x_1, x_2) a function of (y_1, y_2).

Step 3.2: Write the joint density of x_1, x_2

Since x_1, x_2 are independent Gaussian:

$$p_X(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2}{2\sigma^2}\right)$$

Nothing fancy here.

Step 3.3: Compute the Jacobian determinant

The transformation is:

$$(x_1, x_2) = f(y_1, y_2, \alpha)$$

The Jacobian matrix is just the matrix of partial derivatives:

$$J = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix}$$

Now compute each small derivative:

🔷 **For x_1:**

$$x_1 = \frac{y_1}{\alpha}$$

So:

$$\frac{\partial x_1}{\partial y_1} = \frac{1}{\alpha}, \qquad \frac{\partial x_1}{\partial y_2} = 0$$

🔷 **For x_2:**

$$x_2 = y_1 + \frac{y_2 - y_1}{\alpha}$$

Take derivatives:

$$\frac{\partial x_2}{\partial y_1} = 1 - \frac{1}{\alpha} = -\frac{1-\alpha}{\alpha}$$

$$\frac{\partial x_2}{\partial y_2} = \frac{1}{\alpha}$$

🔷 **So the Jacobian is:**

$$J = \begin{pmatrix} \frac{1}{\alpha} & 0 \\ -\frac{1-\alpha}{\alpha} & \frac{1}{\alpha} \end{pmatrix}$$

🔷 **Determinant of J**

$$\det J = \frac{1}{\alpha} \cdot \frac{1}{\alpha} - 0 = \frac{1}{\alpha^2}$$

That's it!

Step 3.4: Apply change of variables

$$\boxed{p(y_1, y_2) = p_X(x_1, x_2) \left| \frac{1}{\alpha^2} \right|}$$

Substitute the Gaussian density:

$$p(y_1, y_2) = \frac{1}{2\pi\sigma^2} \frac{1}{\alpha^2} \exp\left[-\frac{1}{2\sigma^2} \left(\left(\frac{y_1}{\alpha} - \mu \right)^2 + \left(y_1 + \frac{y_2 - y_1}{\alpha} - \mu \right)^2 \right) \right]$$

And that is the final joint density.

Step 4: Log-likelihood as a function of α

Ignoring constants in α:

$$\log p(y_1, y_2 \mid \alpha) = -2\log\alpha - \frac{1}{2\sigma^2}\left[\left(\frac{y_1}{\alpha} - \mu\right)^2 + \left(y_1 + \frac{y_2 - y_1}{\alpha} - \mu\right)^2\right] + C$$

So

$$\boxed{\ell(\alpha) := \log p(y_1, y_2 \mid \alpha) = -2\log\alpha - \frac{1}{2\sigma^2}\left[\left(\frac{y_1}{\alpha} - \mu\right)^2 + \left(y_1 + \frac{y_2 - y_1}{\alpha} - \mu\right)^2\right] + C}$$

(b) MLE for α

Differentiate $\ell(\alpha)$ w.r.t. α and set to zero.

Using the expression above, after simplification you get:

$$\frac{d\ell}{d\alpha} = 0 \quad \Longleftrightarrow \quad 2\sigma^2\alpha^2 + \alpha\left(\mu y_2 + y_1^2 - y_1 y_2\right) - \left(2y_1^2 - 2y_1 y_2 + y_2^2\right) = 0$$

This is a **quadratic equation in** α:

$$\boxed{2\sigma^2\alpha^2 + \alpha\left(\mu y_2 + y_1^2 - y_1 y_2\right) - \left(2y_1^2 - 2y_1 y_2 + y_2^2\right) = 0}$$

The **MLE** $\hat{\alpha}$ is the root of this quadratic that lies in $(0, 1)$.
You can solve it explicitly with the quadratic formula if you want, but the exam says it's enough to leave it as "the solution of this equation," so this is the final required form.