Midterm Exam II
ECE 685D– Introduction to Deep Learning
Fall 2023

Instructor: Prof. Vahid Tarokh
ECE Department, Duke University

Nov 20th 2023
10:05 AM - 11:20 AM
(Exam duration: 75 minutes)

Name: _____
Duke ID (NetID): _____

This exam contains 10 pages and 5 questions. This exam has 115 points of which 15 are bonus points. This is a closed-book exam. No exam aids are allowed. You are not allowed to communicate with others.

Distribution of Marks

Question	Points	Score
1	20	
2	35	
3	15	
4	25	
5	20	
Total:	115	

1. (**Ridge and Lasso Regularization**) Consider the Elastic Network, which utilizes both l_1 and l_2 regularization. The objective function is given as follows:

$$F(w) = \frac{1}{2}\sum_{j=1}^{n}(y^{(j)} - \sum_{i=1}^{d} w_i x_i^{(j)})^2 + \alpha \sum_{i=1}^{d} |w_i| + \frac{\lambda}{2}\sum_{i=1}^{d} w_i^2 \tag{1}$$

Here $(x^{(j)}, y^{(j)})$ is the j$^{\text{th}}$ example in the training data, w is a d-dimensional weight vector, λ is a regularization parameter for the l_2 norm of w, and α is a regularization parameter for the l_1 norm of w. This network is a generalization of Ridge and Lasso regression: It reverts to Lasso when $\lambda = 0$, and it reverts to Ridge when $\alpha = 0$.

Let g, h, c, α be real constants $(h, \alpha > 0)$, and consider the following piece-wise function:

$$f(x) = c + gx + 0.5hx^2 + \alpha|x| \tag{2}$$

(a) (5 points) What is the x^* that minimizes $f(x)$? Write x^* as a piece-wise function of g.

(b) (15 points) Fixing all parameters in the objective function $F(w)$ except w_k, for $k \in \mathbb{R}$. This function is in the form of $f(x)$, where $x = w_k$. Write g and h in $F(w)$. Write the update rule for the Elastic Net.

Solution:

(a) The piecewise function $f(x)$ can be written as:

$$f^+(x) = c + gx + 0.5hx^2 + \alpha x \tag{3}$$

$$f^-(x) = c + gx + 0.5hx^2 - \alpha x \tag{4}$$

We have $x^+ = -\frac{g+\alpha}{h}$ and $x^- = -\frac{g-\alpha}{h}$

$$x^* = \begin{cases} x^+ \text{ if } x^+ > 0 \\ x^- \text{ if } x^- < 0 \\ 0 \text{ if } x^+ < 0, x^- > 0 \end{cases}$$

Or

$$x^* = \begin{cases} -\dfrac{g+\alpha}{h} \text{ if } g < -\alpha \\ 0 \text{ if } g \in [-\alpha, \alpha] \\ -\dfrac{g-\alpha}{h} \text{ if } g > \alpha \end{cases}$$

(b)

$$g = \sum_{j=1}^{n} x_k^{(j)}(\sum_{i\neq k} w_i x_i^{(j)} - y_i) \tag{5}$$

$$h = \lambda + \sum_{j=1}^{n}(x_k^{(j)})^2 \tag{6}$$

The update rule can be written as follows:

for $k \in \{1, 2, \ldots, d\}$ **do**

$g = \sum_{j=1}^{n} x_k^{(j)} (\sum_{i \neq k} w_i x_i^{(j)} - y_i)$

$h = \lambda + \sum_{j=1}^{n} (x_k^{(j)})^2$

$$\mathrm{w}_k = \begin{cases} -\dfrac{g+\alpha}{h} \text{ if } g < -\alpha \\ 0 \text{ if } g \in [-\alpha, \alpha] \\ -\dfrac{g-\alpha}{h} \text{ if } g > \alpha \end{cases}$$

end for=0

Rubric: a) 1 point for each of the regular cases of the piecewise function, 3 points for the 0 case. b) 5 points for g, 5 points for h, 5 points for update rule 2 points for update rule if they condition on w instead of g Error carried forward is allowed for the update rule if the student understands that gradient descent is no longer needed, since we have found the minimizer in part (a)

2. (**GAN**) Let $p^*(x), x \in \mathbb{R}$ denote the true, data-generating probability density function (PDF) given by Gaussian distribution with $(\mu, \sigma) = (1, 1)$ (i.e., $\mathcal{N}(1, 1)$). Consider a GAN model consisting of the generator G and the discriminator D. The minimax loss is given as follows:

$$L(D, G) = E_{x \sim real} log D_{\theta_d}(x) + E_z log(1 - D(G_{\theta_g}(z))) \tag{7}$$

The generator $g(\cdot)$ takes a standard uniform random variable $z \sim Uniform(0, 1)$ and produces a fake random variable $x = g(z)$ from the Gaussian distribution $\mathcal{N}(\mu_g, 1)$.

(a) (10 points) Write down the optimal discriminator $D(x)$. Hint: Write the solution in the form of an integral (No need to simplify the solution)

(b) (10 points) Instead of the optimal discriminator, consider the following discriminator below:

$$D(x) = \begin{cases} \text{"x is real", if } x \geq 1 - \gamma \\ \text{"x is fake", otherwise} \end{cases}$$

Compute the probability of a real data point being declared as fake.

(c) (15 points) If we fix this discriminator, determine the threshold $\gamma > 0$ such that maximize $L^*(D, G)$. You do not have to compute γ explicitly.

Solution:

(a) The minimax objective needs to be maximized with respect to the discriminator D, which results in the 2nd term of the objective with an optimal discriminator becoming

$$E_z log(1 - D(G_{\theta_g}(z))) = \int_z p(z) log(1 - D(G_{\theta_g}(z))) dz, \tag{8}$$

as an optimal discriminator has $D(x) = 1$ for all real data x. This could also be written via the change of variables $x = G(z)$ in terms of the distribution $p_G(x)$ of samples generated by G given sampled z, as

$$\int_x p_G(x) log(1 - D(x)) dx. \tag{9}$$

Hence, the relevant part of the objective for the discriminator value at a particular x become:

$$\max_{D(x)} p(x) \log(D(x)) + p_G(x) \log(1 - D(x))$$

To optimize this, we can take derivatives with respect to the discriminator probability D(x) and solve for a stationary point. Taking derivatives, we have:

$$0 = p(x) \frac{1}{D(x)} - p_G(x) \frac{1}{1 - D(x)}$$

Thus,

$$D^*(x) = \frac{p(x)}{p(x) + p_G(x)}$$

(b) Given some $x \sim p^*(x)$,

$$P(\text{“}x \text{ is fake”}) = P(x < 1-\gamma) = \int_{x=-\infty}^{1-\gamma} p^*(x)dx = \Phi\left(\frac{(1-\gamma)-1}{1}\right) = \Phi\left(-\gamma\right), \quad (10)$$

where Φ is the standard normal CDF / z-score function.

(c) This will be true if the D's decision threshold $1-\gamma$ is halfway between the distributions $p^*(x) = \mathcal{N}(1,1)$ and $p_G(\mu_g, 1)$ (since they have the same variance). In other words, $1-\gamma = \frac{1+\mu_g}{2}$, i.e. $\gamma = \frac{1-\mu_g}{2}$.

3. (15 points) (**Slow Feature Analysis**) Consider the two-dimensional input signals:

$$x_1(t) = cos(t) + cos(22t)$$

and

$$x_2(t) = cos(11t)$$

Both components are quickly varying but hidden in the signals is the slowly varying function:

$$f(t) = A + Bx_1(t) + Cx_2(t) + Dx_1(t)x_2(t) + Ex_1^2(t) + Fx_2^2(t)$$

Find $\theta = [A, B, C, D, E, F]$ such that it minimizes:

$$\mathcal{L}(f) = \frac{1}{4}\sum_{i=1}^{4}[f(\frac{\pi}{4}i) - f(\frac{\pi}{4}(i-1))]^2$$

You do not need to simplify the answer.

Solution:

Note that due to an error in the loss function, $cos(t)$ does not actually minimize the loss. Therefore two solutions are accepted.

1) $\theta = [0, 0, 0, 0, 0, 0]$

2) $\theta = [1, 1, 0, 0, 0, -2]$

Technically any value for A will not affect the loss, so all answers for A are accepted.

4. (**LVM**) Consider a three-dimensional data point $x \in \mathbb{R}^3$ and $h = [h_1, h_2]$ where $h_1, h_2 \in \{0, 1\}$. The joint probability density function (PDF) of x and h is defined as follows:

$$p(x, h) = \exp(-E(x, h))/Z$$

where Z is the normalization constant and $E(x, h)$ is the energy function given by:

$$E(x, h) = -hW^T x - c^T x + bh + 0.5x^T x \tag{11}$$

with

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, c = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, b = 0 \tag{12}$$

(a) (10 points) Write down the conditional PDFs $p(h|x)$ and $p(x|h)$.

$$\begin{aligned} p(x) &= \sum_{h_1 \in \{0,1\}} \sum_{h_2 \in \{0,1\}} p(x, h) \\ &= \frac{1}{Z} \exp(-\frac{1}{2} x^T x)(1 + \exp(x_1 + x_3) + \exp(x_2 + x_3) + \exp(x_1 + x_2 + 2x_3)) \end{aligned}$$

$$\begin{aligned} p(h|x) &= \frac{p(x, h)}{p(x)} \\ &= \frac{Z^{-1} \exp(hw^T x) \exp(-0.5x^T x)}{Z^{-1} \exp(-0.5x^T x)(1 + \exp(x_1 + x_3) + \exp(x_2 + x_3) + \exp(x_1 + x_2 + 2x_3))} \\ &= \frac{\exp(hw^T x)}{1 + \exp(x_1 + x_3) + \exp(x_2 + x_3) + \exp(x_1 + x_2 + 2x_3)} \end{aligned}$$

Taking the equation from part (b), we have

$$\begin{aligned} p(x|h) &= \frac{p(x, h)}{p(h)} \\ &= \frac{Z^{-1} \exp(hW^T x) \exp(-0.5x^T x)}{cZ^{-1} \exp(\frac{1}{2}(hW^T)^T (hW^T))} \\ &= \frac{1}{c} \exp(hW^T x - 0.5x^T x - 0.5(hW^T)^T (hW^T)) \\ &= \frac{1}{c} \exp(-\frac{1}{2}(x - hW^T)^T I^{-1} (x - hW^T)) \\ &= N(x \in \mathbb{R}^3; hW^T, I) \end{aligned}$$

(b) (15 points) Compute the marginal PDF $p(h)$ (i.e., 4 components)

Recall that

$$N(x \in \mathbb{R}^3; \mu, \Sigma) = (2\pi)^{-3/2} \det(\Sigma)^{-1/2} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu))$$

and we have

$$\begin{aligned} p(h) &= \frac{1}{Z}\int_x \exp(hW^T x - \frac{1}{2}x^T x)dx \\ &= \frac{1}{Z}\exp(\frac{1}{2}(hW^T)^T(hW^T))\int_x \exp(hW^T x - \frac{1}{2}x^T x - \frac{1}{2}(hW^T)^T(hW^T))dx \\ &= \frac{1}{Z}\exp(\frac{1}{2}(hW^T)^T(hW^T))\int_x \exp(-\frac{1}{2}(x - hW^T)^T I^{-1}(x - hW^T))dx \end{aligned}$$

We hence find

$$p(h) = \frac{c}{Z}\exp(\frac{1}{2}(hW^T)^T(hW^T)), c \in \mathbb{R}$$

Here, $\frac{c}{Z}$ is a normalizing factor whose value is

$$\frac{c}{Z} = \frac{1}{\sum_{h_1\in\{0,1\},h_2\in\{0,1\}}\exp(\frac{1}{2}(hW^T)^T(hW^T))}$$

Common reasons for point deduction:
If you wrote $p(h|x)$ correctly, either in the form of the key or $p(h_1|x) * p(h_1|x)$, you got 4 points. If you consider the model to be a Bernoulli-Bernoulli RBM, you may be taken off 2 points. If you wrote the formula correctly yet did not simplify the integral for $p(x|h)$ and $p(h)$, you will be deducted 4 points and 8 points respectively. If your formula is wrong for $p(h)$, you will be deducted 12+ points for part (b).

5. (**VAE**) Consider a three-dimensional data point $x \in \mathbb{R}^3$ and two-dimensional latent variable $z \in \mathbb{R}^2$ with a standard normal prior $p(z) = \mathcal{N}(0, I)$. Let $\sigma(\cdot)$ denote the logistic sigmoid function.

   Let the variational posterior distribution $q(z|x)$ be the Gaussian distribution with mean vector $\sigma(Ax+a)$, and the diagonal covariance matrix whose diagonal elements are given by $\sigma(Bx+b)$ where:

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & -1 & -2 \end{bmatrix}, a = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, B = \begin{bmatrix} -1 & 1 & 1 \\ 1 & w_1 & -2 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}. \tag{13}$$

   Similarly, let the generator distribution $p(x|z)$ be Gaussian with mean vector $\sigma(Cz + c)$ and diagonal covariance matrix whose diagonal elements are given by $\sigma(Dz + d)$ where:

$$C = \begin{bmatrix} 2 & 0 \\ -1 & w_2 \\ 2 & 1 \end{bmatrix}, c = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}, D = \begin{bmatrix} 1 & -2 \\ -1 & 2 \\ -1 & -2 \end{bmatrix}, d = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}. \tag{14}$$

   Recall the Evidence Lower Bound (ELBO) given below:

$$\mathcal{L}(\theta) = \mathbb{E}_{z\sim q(z|x)}[\log p(x|z)] - D_{KL}[(q(z|x)||p(z)] \tag{15}$$

   where $\theta = [w_1, w_2]$.

   (a) (5 points) Compute $\mathcal{L}(\theta)$. Given the multivariate Gaussian $p(x) \sim N(\mu_1, \Sigma_1), q(x) \sim N(\mu_2, \Sigma_2)$

$$\begin{aligned} KL(p(x)||q(x)) &= \mathbb{E}_{x\sim p(x)}[\log p(x) - \log q(x)] \\ &= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} + \frac{1}{2}\mathbb{E}_{x\sim p(x)}[-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)] \\ &= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} + \frac{1}{2}\mathbb{E}_{x\sim p(x)}[-tr\{\Sigma_1^{-1}(x-\mu_1)(x-\mu_1)^T\} + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)] \\ &= \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} + \frac{1}{2}\mathbb{E}_{x\sim p(x)}[-tr\{\Sigma_1^{-1}\Sigma_1\} + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)] \\ &= \frac{1}{2}(\log\frac{\det\Sigma_2}{\det\Sigma_1} - d) + \frac{1}{2}[(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) + tr\{\Sigma_2^{-1}\Sigma_1\}] \end{aligned}$$

   Students will full points (2 pts) if they expand the equation to the expectation or integral form (like in the 2nd step.) Then, we have

$$\mu_1 = \sigma(Ax+a), \Sigma_1 = diag\{\sigma(Bx+b)\}, \mu_2 = 0, \Sigma_2 = \mathbf{I}_d$$

   Students are supposed to explicitly compute the value of the four parameters above (2pts)

   The $\mathbb{E}_{z\sim q(z|x)}[\log p(x|z)]$ term can be kept with either integral or expectation form. However, the student is supposed to compute the parameter value of

$$N(\mu_3 = \sigma(Cx+c), \Sigma_3 = diag\{\sigma(Dx+d)\}$$

   and plug it into the term (1pt).

(b) (15 points) Compute the gradient $\nabla_\theta \mathcal{L}(\theta)$.Do not need to simplify the solution.
***This is different from the example exam because z remains a random variable but not a deterministic sample/samples. Hence, the solution is slightly different. Part (a) is graded leniently if students use the deterministic setting. However, doing so might lead to an incorrect equation to start with in part (b)

$$\text{Given } x,\ \mathbb{E}_{z\sim q(z|x)}[\log p(x|z)] = \int_z q(z|x)\log p(x|z)\ dz$$
$$= \int_z N(\sigma(Ax+a), diag\{\sigma(Bx+b)\})\log N(x;\sigma(Cz+c), diag\{\sigma(Dz+d)\})\ dz$$

$$\text{Then, recall } KL(q(z|x)||p(z)) = \frac{1}{2}(\log\frac{\det\Sigma_2}{\det\Sigma_1} - d) + [(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) + tr\{\Sigma_2^{-1}\Sigma_1\}]$$

The student should show that the KL divergence term contributes only to the gradient of w_1 by equations (6 pts) in integral or expectation form.

Likewise, the student should show the likelihood term contributes to the gradient of both w_1 *and* w_2 with equations in integral form (7 pts).

Students are expected to plug the actual numerical value into the equations and expand them (but are not required to simplify the equations). If students did not write any non-trivial equations or leave the answer blank, deduct an additional 1-2 pts.

1. Solving this particular question

We're given:

- Prior: $p(z) = \mathcal{N}(0, I_2)$.
- Approx posterior:
  $q(z \mid x) = \mathcal{N}(\mu_1, \Sigma_1)$ with
  $\mu_1 = \sigma(Ax + a)$, $\Sigma_1 = \mathrm{diag}(\sigma(Bx + b))$.
- Likelihood (generator):
  $p(x \mid z) = \mathcal{N}(\mu_3, \Sigma_3)$ with
  $\mu_3 = \sigma(Cz + c)$, $\Sigma_3 = \mathrm{diag}(\sigma(Dz + d))$.
- ELBO:

$$\mathcal{L}(\theta) = \mathbb{E}_{z \sim q(z|x)}[\log p(x \mid z)] - D_{KL}\big(q(z \mid x) \,\|\, p(z)\big),$$

  where $\theta = [w_1, w_2]$ appear inside A, B, C, D.

1.1 KL term $D_{KL}(q(z \mid x) \| p(z))$

They remind you of the KL between Gaussians:

> For $p(x) \sim \mathcal{N}(\mu_1, \Sigma_1)$, $q(x) \sim \mathcal{N}(\mu_2, \Sigma_2)$,
>
> $$KL(p\|q) = \frac{1}{2}\log\frac{\det\Sigma_2}{\det\Sigma_1} + \frac{1}{2}\mathbb{E}_p[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2$$

Simplified form (what they derive below) is the standard:

$$KL(p\|q) = \frac{1}{2}\left[\log\frac{\det\Sigma_2}{\det\Sigma_1} - d + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right],$$

for dimension d.

In our ELBO, we need $D_{KL}(q(z \mid x)\|p(z))$.
So **match p and q to the formula**:

- $p = q(z \mid x)$:
  $\mu_1 = \sigma(Ax + a)$, $\Sigma_1 = \mathrm{diag}(\sigma(Bx + b))$.
- $q = p(z)$:
  $\mu_2 = 0$, $\Sigma_2 = I_2$ (since prior is standard normal).

Then plug into the formula with $d = 2$:

- $\det\Sigma_2 = 1,\ \Sigma_2^{-1} = I$.

So

$$\begin{aligned} D_{KL}(q(z \mid x)\|p(z)) &= \frac{1}{2}\left[\log\frac{1}{\det\Sigma_1} - 2 + \mathrm{tr}(\Sigma_1) + (\mu_2-\mu_1)^T(\mu_2-\mu_1)\right] \\ &= \frac{1}{2}\left[-\log\det\Sigma_1 - 2 + \mathrm{tr}(\Sigma_1) + \|\mu_1\|^2\right], \end{aligned}$$

where

- $\mu_1 = \sigma(Ax + a)$,
- $\Sigma_1 = \mathrm{diag}(\sigma(Bx + b))$.

That's the KL part they expect you to write explicitly.

1.2 Reconstruction term $\mathbb{E}_{q(z|x)}[\log p(x \mid z)]$

Given

$$p(x \mid z) = \mathcal{N}(\mu_3, \Sigma_3), \quad \mu_3 = \sigma(Cz + c), \quad \Sigma_3 = \mathrm{diag}(\sigma(Dz + d)).$$

Log Gaussian density in 3D:

$$\log p(x \mid z) = -\frac{1}{2}\left[3\log(2\pi) + \log\det\Sigma_3 + (x - \mu_3)^T\Sigma_3^{-1}(x - \mu_3)\right].$$

So the reconstruction term is

$$\mathbb{E}_{q(z|x)}[\log p(x \mid z)] = -\frac{1}{2}\,\mathbb{E}_{q(z|x)}\Big[3\log(2\pi) + \log\det\Sigma_3 + (x - \mu_3)^T\Sigma_3^{-1}(x - \mu_3)\Big],$$

with μ_3, Σ_3 as above.

Because of the nonlinearity $\sigma(Cz + c)$, this expectation usually has **no closed form**, so it's perfectly fine (and standard) to keep it as an expectation in the final expression, just like the instructions say.

1.3 Put them together: final expression for $\mathcal{L}(\theta)$

So for a given x:

$$\boxed{\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}_{z\sim q(z|x)}[\log p(x \mid z)] - D_{KL}\big(q(z \mid x)\,\|\,p(z)\big) \\ &= -\frac{1}{2}\,\mathbb{E}_{q(z|x)}\Big[3\log(2\pi) + \log\det\Sigma_3 + (x - \mu_3)^T\Sigma_3^{-1}(x - \mu_3)\Big] \\ &\quad -\frac{1}{2}\left[-\log\det\Sigma_1 - 2 + \mathrm{tr}(\Sigma_1) + \|\mu_1\|^2\right],\end{aligned}}$$

with

$$\mu_1 = \sigma(Ax + a), \quad \Sigma_1 = \mathrm{diag}(\sigma(Bx + b)), \quad \mu_3 = \sigma(Cz + c), \quad \Sigma_3 = \mathrm{diag}(\sigma(Dz + d)).$$

That's exactly what they're aiming for in part (a):
KL in closed form, reconstruction term in expectation form, and all the Gaussian parameters expressed from the given matrices.

---

2. "Universal template" you can reuse

For **any Gaussian VAE** with:

- prior $p(z) = \mathcal{N}(\mu_p, \Sigma_p)$,
- variational posterior $q_\phi(z \mid x) = \mathcal{N}(\mu_q(x), \Sigma_q(x))$,
- likelihood $p_\theta(x \mid z) = \mathcal{N}(\mu_x(z), \Sigma_x(z))$,

the **ELBO for a single datapoint** is:

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

KL between Gaussians (universal formula):

For dimension d,

$$\boxed{D_{KL}\big(\mathcal{N}(\mu_q,\Sigma_q)\,\|\,\mathcal{N}(\mu_p,\Sigma_p)\big)=\frac{1}{2}\Big[\log\frac{\det\Sigma_p}{\det\Sigma_q}-d+\operatorname{tr}(\Sigma_p^{-1}\Sigma_q)+(\mu_p-\mu_q)^T\Sigma_p^{-1}(\mu_p-\mu_q)\Big].}$$

This is the **template you always plug into.**

Log Gaussian (for reconstruction term):

If $x\in\mathbb{R}^d$,

$$\boxed{\log\mathcal{N}(x;\mu_x,\Sigma_x)=-\frac{1}{2}\Big[d\log(2\pi)+\log\det\Sigma_x+(x-\mu_x)^T\Sigma_x^{-1}(x-\mu_x)\Big].}$$

So the reconstruction term is *always*:

$$\boxed{\mathbb{E}_{q(z|x)}[\log p(x\mid z)]=-\frac{1}{2}\mathbb{E}_{q(z|x)}\Big[d\log(2\pi)+\log\det\Sigma_x(z)+(x-\mu_x(z))^T\Sigma_x(z)^{-1}(x-\mu_x(z))\Big].}$$

Depending on how $\mu_x(z),\Sigma_x(z)$ depend on z, you either:

- simplify the expectation analytically (if linear), or
- leave it as an expectation (or approximate by Monte Carlo).

TL;DR template

For VAE with Gaussian $q(z|x)$, $p(z)$, $p(x|z)$:

1. Identify $\mu_q(x),\Sigma_q(x)$ and μ_p,Σ_p; plug into Gaussian KL formula.
2. Identify $\mu_x(z),\Sigma_x(z)$; plug into log-Gaussian formula inside an expectation.
3. ELBO = reconstruction expectation – KL.

Follow these 3 steps and you can solve essentially any exam/homework question in this style.

Nice, this is the “gradient of ELBO w.r.t. θ=[w₁,w₂]” part.

The key is:

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

with

- $q_\theta(z|x) = \mathcal{N}(\mu_1, \Sigma_1)$,
  $\mu_1 = \sigma(Ax + a)$ (no w_1, w_2 here),
  $\Sigma_1 = \mathrm{diag}(\sigma(Bx + b))$ (depends on w_1 through B, not on w_2).
- $p_\theta(x|z) = \mathcal{N}(\mu_3, \Sigma_3)$,
  $\mu_3 = \sigma(Cz + c)$ (depends on w_2 via C),
  $\Sigma_3 = \mathrm{diag}(\sigma(Dz + d))$ (no w_1, w_2).

So:

- **KL term** depends on w_1 only (through Σ_1).
- **Likelihood term** depends on **both**:
  - on w_1 via the *distribution* $q(z|x)$ inside the expectation,
  - on w_2 via $\log p(x|z)$.

Below is exactly what they want: gradient in integral / expectation form, not fully simplified.

1. Write the ELBO for a fixed x

From the statement (they already wrote the first line):

$$\mathbb{E}_{z\sim q(z|x)}[\log p(x|z)] = \int q(z|x)\ \log p(x|z)\, dz = \int \mathcal{N}\big(z;\mu_1,\Sigma_1\big)\ \log \mathcal{N}\big(x;\mu_3,\Sigma_3\big)\, dz,$$

with
$\mu_1 = \sigma(Ax + a),\ \Sigma_1 = \mathrm{diag}(\sigma(Bx + b))$
$\mu_3 = \sigma(Cz + c),\ \Sigma_3 = \mathrm{diag}(\sigma(Dz + d))$.

And the KL between Gaussians (here $q(z|x)$ vs prior $p(z) = \mathcal{N}(0, I)$):

$$D_{KL}(q(z|x)\|p(z)) = \frac{1}{2}\Big(\log\frac{\det\Sigma_2}{\det\Sigma_1} - d + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1)\Big),$$

where for us $\mu_2 = 0,\ \Sigma_2 = I_2,\ d = 2$.

So:

$$\mathcal{L}(\theta) = \int q(z|x)\ \log p(x|z)\, dz - D_{KL}(q(z|x)\|p(z)).$$

2. Gradient w.r.t. w_1

Here w_1 only appears in $B \to \Sigma_1 \to$ in both:

- the distribution $q(z|x)$ used in the expectation, and
- the KL term (through Σ_1).

So:

$$\frac{\partial \mathcal{L}}{\partial w_1} = \underbrace{\frac{\partial}{\partial w_1} \int q(z|x) \log p(x|z)\, dz}_{\text{from likelihood term}} - \underbrace{\frac{\partial}{\partial w_1} D_{KL}(q(z|x)\|p(z))}_{\text{from KL term}}.$$

Do each part:

2.1 Likelihood part

Because $p(x|z)$ does **not** depend on w_1, only q does:

$$\frac{\partial}{\partial w_1} \int q(z|x) \log p(x|z)\, dz = \int \frac{\partial q(z|x)}{\partial w_1} \log p(x|z)\, dz.$$

And explicitly:

$$q(z|x) = \mathcal{N}\big(z; \mu_1, \Sigma_1(w_1)\big),$$

so you can write

$$\frac{\partial q(z|x)}{\partial w_1} = \frac{\partial}{\partial w_1} \mathcal{N}\big(z; \mu_1, \Sigma_1(w_1)\big)$$

and plug that into the integral (you don't have to simplify further).

2.2 KL part

Using the Gaussian KL formula and remembering that only Σ_1 depends on w_1:

$$D_{KL}(q\|p) = \frac{1}{2}\Big(-\log\det\Sigma_1 - d + \operatorname{tr}(\Sigma_1) + \|\mu_1\|^2 \Big),$$

so

$$\frac{\partial}{\partial w_1} D_{KL}(q\|p) = \frac{1}{2}\left[-\operatorname{tr}\left(\Sigma_1^{-1}\frac{\partial \Sigma_1}{\partial w_1}\right) + \operatorname{tr}\left(\frac{\partial \Sigma_1}{\partial w_1}\right)\right],$$

with $\frac{\partial \Sigma_1}{\partial w_1}$ diagonal and only the $(2,2)$ entry non-zero (because the second row of B contains w_1). If you want to be explicit:

- let $u = Bx + b$, $s = \sigma(u)$, $\Sigma_1 = \operatorname{diag}(s)$;
- $u_2 = x_1 + w_1 x_2 - 2x_3 + b_2$;
- $\frac{\partial s_2}{\partial w_1} = \sigma(u_2)\big(1 - \sigma(u_2)\big) x_2$;
- so $\frac{\partial \Sigma_1}{\partial w_1} = \operatorname{diag}(0, \frac{\partial s_2}{\partial w_1})$.

You do **not** need to simplify further—that matches the rubric.

So we can summarise:

$$\boxed{\frac{\partial \mathcal{L}}{\partial w_1} = \int \frac{\partial q(z|x)}{\partial w_1} \log p(x|z)\, dz \;-\; \frac{1}{2}\left[-\operatorname{tr}\left(\Sigma_1^{-1}\frac{\partial \Sigma_1}{\partial w_1}\right) + \operatorname{tr}\left(\frac{\partial \Sigma_1}{\partial w_1}\right)\right].}$$

This shows clearly that **only** w_1 appears in the KL term.

3. Gradient w.r.t. w_2

Here w_2 appears only in $C \to \mu_3(z) \to$ only in the **likelihood term**, not in q or the KL term.

So:

$$\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial}{\partial w_2} \int q(z|x) \log p(x|z)\, dz - \frac{\partial}{\partial w_2} D_{KL}(q\|p).$$

But KL doesn't depend on w_2, so its derivative is 0 and:

$$\frac{\partial \mathcal{L}}{\partial w_2} = \int q(z|x) \frac{\partial}{\partial w_2} \log p(x|z)\, dz.$$

Use the Gaussian log-density:

$$\log p(x|z) = -\frac{1}{2}\left[3\log(2\pi) + \log\det\Sigma_3 + (x-\mu_3(z))^T\Sigma_3^{-1}(x-\mu_3(z))\right],$$

and only $\mu_3(z) = \sigma(Cz+c)$ depends on w_2 (via the entry $C_{2,2} = w_2$). So

$$\frac{\partial}{\partial w_2}\log p(x|z) = -\frac{1}{2}\frac{\partial}{\partial w_2}\left[(x-\mu_3(z))^T\Sigma_3^{-1}(x-\mu_3(z))\right],$$

with $\frac{\partial \mu_3}{\partial w_2}$ coming through the chain rule:

$$\frac{\partial \mu_3}{\partial w_2} = \sigma'(Cz+c) \odot \frac{\partial (Cz)}{\partial w_2},$$

and only the second component is non-zero because that row of C has $w_2 z_2$.

Thus, in integral form:

$$\boxed{\frac{\partial \mathcal{L}}{\partial w_2} = \int q(z|x)\left[-\frac{1}{2}\frac{\partial}{\partial w_2}\left((x-\mu_3(z))^T\Sigma_3^{-1}(x-\mu_3(z))\right)\right] dz.}$$

This shows the likelihood term contributes to gradie ↓ of **both** w_1 **and** w_2, while the KL term only involves w_1—exactly what the question and rubric are asking you to show.