Due Date: April 18th at 23:00

Instructions

- For all questions that are not graded only on the answer, show your work! Any problem without work shown will get no marks regardless of the correctness of the final answer.
- Please try to use a document preparation system such as LaTeX. If you write your answers by hand, note that you risk losing marks if your writing is illegible without any possibility of regrade, at the discretion of the grader.
- Submit your answers electronically via the course GradeScope. Incorrectly assigned answers can be given 0 automatically at the discretion of the grader. To assign answers properly, please:
  - Make sure that the top of the first assigned page is the question being graded.
  - Do not include any part of the answer to any other questions within the assigned pages.
  - Assigned pages need to be placed in order.
  - For questions with multiple parts, the answers should be written in order of the parts within the question.
- Questions requiring written responses should be short and concise when necessary. Unnecessary wordiness will be penalized at the grader's discretion.
- Please sign the agreement below.
- It is your responsibility to follow updates to the assignment after release. All changes will be visible on Overleaf and Piazza.
- Any questions should be directed towards the TAs for this assignment (theoretical part): *Vitória Barin Pacela, Philippe Martin.*

I acknowledge I have read the above instructions and will abide by them throughout this assignment. I further acknowledge that any assignment submitted without the following form completed will result in no marks being given for this portion of the assignment.

Name: **Félix Wilhelmy**

UdeM Student ID: **20333575**

Signature: _____

Question 1

Question 1.1

Let's start from the well known identity of the ELBO equation

$$\log p_\theta(x) = \mathcal{L}(q, p) + \mathrm{KL}\Big(q(z|x) \,\|\, p_\theta(z|x)\Big)$$

The lower bound can be further defined as

$$\mathcal{L}(q, p) = \mathbb{E}_{q(z|x)}\left[\frac{\log p(x, z)}{\log q(z \mid x)}\right]$$

We can further define it as

$$\mathcal{L}(q, p) = \mathbb{E}_{q(z|x)}\left[\log p(x, z)\right] - \mathbb{E}_{q(z|x)}\left[\log q(z \mid x)\right]$$

By integrating this into ELBO, we get

$$\log p_\theta(x) = \mathbb{E}_{q(z|x)}\left[\log p(x, z)\right] - \mathbb{E}_{q(z|x)}\left[\log q(z \mid x)\right] + \mathrm{KL}\Big(q(z|x) \,\|\, p_\theta(z|x)\Big)$$

The term $\mathbb{E}_{q(z|x)}\left[\log p(x, z)\right]$ is the expected complete data log-likelihood (ECLL). By rearranging our equation and isolating the ECLL, we get

$$\mathbb{E}_{q(z|x)}\left[\log p(x, z)\right] = \log p_\theta(x) + \mathrm{KL}\Big(q(z|x) \,\|\, p_\theta(z|x)\Big) - \mathbb{E}_{q(z|x)}\left[\log q(z \mid x)\right]$$

In the problem, the variational distribution $q(z|x)$ is fixed, so $\mathbb{E}_{q(z|x)}\big[\log q(z|x)\big]$ is constant with respect to the model parameters θ.

Therefore, maximizing the expected complete data log-likelihood

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

with respect to θ is equivalent to maximizing

$$\log p_\theta(x) - \mathrm{KL}\Big(q(z|x) \,\|\, p_\theta(z|x)\Big)$$

Question 1.2

Since the instance-dependent variational distribution is optimized solely for the individual data point x_i, it is free to adapt fully to the true posterior $p_\theta(z|x_i)$. In the ideal case where the variational family $\mathcal{Q}$ is rich enough, one could achieve

$$\mathrm{KL}\big(q_i^*(z)\|p_\theta(z|x_i)\big) \approx 0$$

In contrast, the amortized variational inference solution is, via a single neural network, trained over the entire dataset, it must strike a balance among all x_i. Hence, its approximation capability is limited. Optimizing each separately gives

$$\sum_{i=1}^{n} \mathcal{L}_i\big(q_{\phi^*}\big) \;\le\; \sum_{i=1}^{n} \max_q \mathcal{L}_i(q) \;=\; \sum_{i=1}^{n} \mathcal{L}_i\big(q_i^*\big).$$

If we substitute $\mathcal{L}_i(q) = \log p_\theta(x_i) - D_{KL}(q\|p_\theta(z \mid x_i))$, we get

$$\sum_{i=1}^{n} \big[\log p_\theta(x_i) - D_{\mathrm{KL}}\big(q_{\phi^*}(z \mid x_i) \,\|\, p_\theta(z \mid x_i)\big)\big] \;\le\; \sum_{i=1}^{n} \big[\log p_\theta(x_i) - D_{\mathrm{KL}}\big(q_i^*(z) \,\|\, p_\theta(z \mid x_i)\big)\big].$$

The terms $\log p_\theta(x_i)$ cancel, yielding

$$\sum_{i=1}^{n} D_{\mathrm{KL}}\big(q_{\phi^*}(z \mid x_i) \,\|\, p_\theta(z \mid x_i)\big) \;\ge\; \sum_{i=1}^{n} D_{\mathrm{KL}}\big(q_i^*(z) \,\|\, p_\theta(z \mid x_i)\big).$$

Since each KL divergence is nonnegative, this implies for each i

$$D_{\mathrm{KL}}\big(q_{\phi^*}(z \mid x_i) \,\|\, p_\theta(z \mid x_i)\big) \;\ge\; D_{\mathrm{KL}}\big(q_i^*(z) \,\|\, p_\theta(z \mid x_i)\big).$$

Hence the amortized posterior has a KL gap to the true posterior that is at least as large as the instance-wise optimum.

Question 1.3

We compare the two approaches **amortized variational inference** (when q_ϕ is optimized for the whole dataset) and the traditional **instance-dependent approach** (with a per-data point variational distribution $q_i(z)$) along three dimensions.

Question 1.3.a

As discussed in Question 1.2, the optimal instance-dependent solution $q_i^*(z)$ can approach the true posterior arbitrarily well (i.e., with KL ≈ 0) given a sufficiently expressive family $\mathcal{Q}$. However, the amortized solution $q_{\phi^*}(z|x)$ is limited by its shared parameterization. This inherent constraint means that the resulting KL divergence will be larger in the amortized case, thereby introducing additional bias in the marginal likelihood estimate.

Question 1.3.b

Let C_f denote the computational cost of a single forward pass through the recognition network. For amortized inference, the cost of obtaining the variational parameters for a new data point is $\mathcal{O}(C_f)$.

In contrast, with the instance-dependent approach, assume that finding the optimal variational parameters via iterative optimization requires T iterations with per-iteration cost C_{i}. Then, for each data point, the cost is on the order of $\mathcal{O}(T \cdot C_{\text{i}})$.

For large n or if T is large (due to slow convergence), $\mathcal{O}(T \cdot C_{\text{i}})$ can be significantly larger than $\mathcal{O}(C_f)$, indicating that amortized inference not only reduces per-data point cost, but also avoids the repeated iterative optimizations required by the instance-dependent approach.

Question 1.3.c

Let p denote the number of parameters in the latent variable model $q_\phi(z|x)$. In amortized inference, regardless of the dataset size n, the total parameter storage remains $\mathcal{O}(p)$.

Conversely, for the instance-dependent approach, if we store an independent set of variational parameters π_i for each data point, and if each π_i has p' parameters (often $p' \ll p$), then the total memory requirement is $\mathcal{O}(n \cdot p')$.

This scaling can become prohibitive for large datasets, resulting in a memory complexity that grows linearly with n, while amortized inference remains fixed in memory usage irrespective of the data size.

Question 2

Question 2.1

When $\beta = 0$, the KL regularization term is removed from the β-VAE objective and it reduces to

$$\mathcal{L}^{\beta\text{VAE}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]$$

Without the KL term, there is no incentive for the learned posterior $q_{\boldsymbol{\phi}}(z|x)$ to adhere to any prior $p_{\boldsymbol{\theta}}(z)$. As a result, the encoder is not required to learn a meaningful or "organized" latent space z.

The optimal solution is achieved by maximizing the conditional likelihood $p_{\boldsymbol{\theta}}(x|z)$ (or equivalently, minimizing its negative log-likelihood), which forces the network to reconstruct the input x from the latent code z as accurately as possible.

Thus, the model will focus solely on reducing the reconstruction error, potentially in a nearly deterministic fashion akin to that of a conventional autoencoder. This reconstruction-driven objective can lead to a degenerate latent variable model that lacks the desired regularization or disentanglement properties.

Question 2.2

If we consider optimizing only the regularization (KL divergence) term while ignoring the reconstruction error, the β-VAE objective becomes

$$\mathcal{L}^{\beta\text{VAE}}_{\boldsymbol{\theta},\boldsymbol{\phi},\beta}(\boldsymbol{x}) = \beta\text{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,||\, p_{\boldsymbol{\theta}}(\boldsymbol{z}))$$

However, since the reconstruction likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ does not contribute to the objective, the model is not encouraged to use any information from $\boldsymbol{x}$ to shape the distribution of $\boldsymbol{z}$. As a result, the encoder always outputs the prior and the latent representation is independent of the input. Although this trivially minimizes the KL divergence, it leads to a degenerate latent variable model that cannot reconstruct the input data effectively.

Question 2.3

We wish to show

$$\max_{\theta}\Big[\mathbf{I}_{\theta}(X;Y)\Big] \geq \max_{\theta,\theta'}\Big[\mathbf{H}(X) - l_{\theta,\theta'}(X|Y)\Big]$$

In other words, the best achievable mutual information under our model should be at least as large as the best difference between the entropy of X and the autoencoder reconstruction loss.

By definition of mutual information,

$$\mathbf{I}_\theta(X;Y) = \mathbf{H}(X) - \mathbf{H}_\theta(X \mid Y)$$

where $\mathbf{H}_\theta(X \mid Y)$ is the conditional entropy under the true conditional $p_\theta(X \mid Y)$. Substituting this into the maximization gives

$$\max_\theta \Big[\mathbf{H}(X) - \mathbf{H}_\theta(X \mid Y)\Big] \geq \max_{\theta,\theta'} \Big[\mathbf{H}(X) - l_{\theta,\theta'}(X|Y)\Big]$$

Observe that $\mathbf{H}(X)$ is a property of the data distribution alone and does not depend on θ or θ'. Consequently, when comparing the two maxima, $\mathbf{H}(X)$ cancels out, leaving

$$\max_\theta \Big[-\mathbf{H}_\theta(X \mid Y)\Big] \geq \max_{\theta,\theta'} \Big[-l_{\theta,\theta'}(X|Y)\Big]$$

Next, we write both the conditional entropy and the autoencoder loss explicitly as expectations over the joint $p_\theta(X, Y)$

$$\mathbf{H}_\theta(X \mid Y) = -\mathbb{E}_{p_\theta(X,Y)}\big[\log p_\theta(X \mid Y)\big] \quad \text{and} \quad l_{\theta,\theta'}(X|Y) = -\mathbb{E}_{p_\theta(X,Y)}\big[\log q_{\theta'}(X \mid Y)\big]$$

because the decoder $q_{\theta'}(X \mid Y)$ assigns a likelihood to each true (X, Y) pair and we measure its quality by the expectation of the negative log likelihood under that same joint.

Plugging these into the previous inequality immediately yields

$$\max_\theta \Big[\mathbb{E}_{p_\theta(X,Y)}\big[\log p_\theta(X \mid Y)\big]\Big] \geq \max_{\theta,\theta'} \Big[\mathbb{E}_{p_\theta(X,Y)}\big[\log q_{\theta'}(X \mid Y)\big]\Big]$$

This is exactly the variational lower-bound inequality cited in hint (b). Hence, minimizing the expected reconstruction error $l_{\theta,\theta'}$ provides a lower bound on $I(X;Y)$.

Question 2.4

As an alternative to the reconstruction loss, consider a contrastive objective that maximizes mutual information between two views of X. Let

$$Y = f_\theta(X) \quad \text{and} \quad Y^+ = f_\theta(X^+) \quad \text{and} \quad Y_j^- = f_\theta(X_j^-).$$

Here, X^+ is an augmented "positive" view of X, and $\{X_j^-\}_{j=1}^M$ are independently sampled "negative" examples. We then define the contrastive loss

$$\mathcal{L}_{\text{InfoNCE}}(\theta) = \mathbb{E}_X \left[-\log \frac{\exp\big(\langle Y, Y^+ \rangle\big)}{\sum_{j=1}^M \exp\big(\langle Y, Y_j^- \rangle\big)} \right]$$

Replacing the autoencoder reconstruction term $l_{\theta,\theta'}(X \mid Y)$ with $\mathcal{L}_{\text{InfoNCE}}$ encourages the encoder to produce representations that are invariant under the chosen augmentations of X while remaining discriminative across different samples.

Question 2.5

Question 2.5.a

Consider any nonzero vector $x \in \mathbb{R}^2$; for example, $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Define two invertible linear encoders

$$f_a(x) = A\,x = I_{2\times 2}\,x = x \quad \text{and} \quad f_b(x) = B\,x = 2I_{2\times 2}\,x = 2x$$

where $A, B \in \mathbb{R}^{2\times 2}$ and $A \neq B$. Their latent representations are

$$z_a = f_a(x) = x \quad \text{and} \quad z_b = f_b(x) = 2x$$

so $z_a \neq z_b$ despite both mappings being invertible.

Question 2.5.b

More generally, let $z_* = f_*(x)$ be any valid linear encoding of x. Then for any invertible matrix $T \in \mathrm{GL}(2)$,

$$z = T\,z_*$$

is also a valid encoding of x, since T reparameterizes the latent without losing information.

Hence, the full set of possible latent representations is

$$\mathcal{Z}(x) = \{\, T\,z_* : T \in \mathrm{GL}(2)\},$$

showing the latent variable is non-unique up to arbitrary invertible linear transformations.

Question 3

Question 3.1

We are given the value function

$$V(d, g) = d\, g$$

with $d, g \in \mathbb{R}$. We perform gradient ascent on d (to maximize V) and gradient descent on g (to minimize V) simultaneously with a learning rate α.

The gradient of V is

$$\nabla V(d, g) = \begin{pmatrix} g \\ d \end{pmatrix}$$

For gradient ascent in d and gradient descent in g, the update is

$$d_{k+1} = 1\, d_k + \alpha\, g_k \quad \text{and} \quad g_{k+1} = 1\, g_k - \alpha\, d_k$$

Thus, we obtain a 2×2 matrix A that describes the simultaneous update

$$A = \begin{pmatrix} 1 & \alpha \\ -\alpha & 1 \end{pmatrix}$$

Writing these updates together in matrix form, we obtain

$$\begin{pmatrix} d_{k+1} \\ g_{k+1} \end{pmatrix} = A \begin{pmatrix} d_k \\ g_k \end{pmatrix} = \begin{pmatrix} 1 & \alpha \\ -\alpha & 1 \end{pmatrix} \begin{pmatrix} d_k \\ g_k \end{pmatrix}$$

Question 3.2

A stationary point is defined as a point (d, g) that remains invariant under the update,

$$\begin{pmatrix} d_k \\ g_k \end{pmatrix} = A \begin{pmatrix} d_k \\ g_k \end{pmatrix}$$

Therefor, for the stationary point, the updates must satisfy

$$d_{k+1} - d_k = \alpha\, g_k = 0, \quad g_{k+1} - g_k = -\alpha\, d_k = 0$$

Assuming $\alpha > 0$, these equations imply

$$g_k = 0 \quad \text{and} \quad d_k = 0$$

Thus, the unique stationary point is

$$(d_k, g_k) = (0, 0)$$

Question 3.3

To analyze the behavior of the iterates d_k and g_k as $k \to \infty$, we begin by computing the eigenvalues of the update matrix

$$A = \begin{pmatrix} 1 & \alpha \\ -\alpha & 1 \end{pmatrix}$$

The eigenvalues λ satisfy

$$\det\big(A - \lambda I\big) = \det \begin{pmatrix} 1-\lambda & \alpha \\ -\alpha & 1-\lambda \end{pmatrix} = (1-\lambda)^2 + \alpha^2 = 0$$

Solving for λ, we obtain:

$$(1-\lambda)^2 = -\alpha^2 \quad \Longrightarrow \quad 1 - \lambda = \pm i\alpha \quad \Longrightarrow \quad \lambda = 1 \mp i\alpha$$

The magnitude of each eigenvalue is given by

$$|\lambda| = \sqrt{1^2 + \alpha^2} = \sqrt{1 + \alpha^2}$$

This magnitude represents the spectral radius of A, which determines the long-term behavior of the iterates. In particular:

- If $|\lambda| < 1$, the iterates decay to zero.
- If $|\lambda| = 1$, the iterates neither decay nor diverge (they remain constant or oscillate).
- If $|\lambda| > 1$, the iterates grow without bound (diverge).

For any nonzero α, it follows that $\sqrt{1+\alpha^2} > 1$.

Thus the iterates d_k and g_k will diverge as $k \to \infty$. Moreover, because the eigenvalues are complex conjugates, the divergence is accompanied by oscillations. In other words, the joint update causes the system to spiral outward in the (d, g) plane.

Question 4

Question 4.1

Here, we aim to learn the parameters of an energy-based model by contrasting it with a noise distribution with known density $p_n(x)$. To do so, we use Noise Contrastive Estimation (NCE), which formulates the learning problem as a binary classification task that distinguishes samples drawn from the real data distribution $p_d(x)$ from those drawn from the noise distribution $p_n(x)$.

The NCE learning objective is defined as

$$\mathcal{L}_{\text{NCE}}(E_\theta, p_d, p_n) := \mathbb{E}_{\boldsymbol{x}\sim p_d}\big[\log \sigma(E_\theta(\boldsymbol{x}))\big] + \mathbb{E}_{\boldsymbol{z}\sim p_n}\Big[\log\Big(1 - \sigma(E_\theta(\boldsymbol{z}))\Big)\Big]$$

Let $E_\theta(x)$ denote the parameterized energy function of our model, and let $E_{\theta^*}(x)$ denote its optimal version after training. Assuming that the energy function $E_\theta(x)$ is sufficiently expressive, the optimal energy function is defined by

$$E_{\theta^*}(x) = \underset{E_\theta}{\text{argmin}}\ \mathcal{L}_{\text{NCE}}(E_\theta, p_d, p_n)$$

In NCE, we interpret the sigmoid output as the probability that a sample x comes from the data distribution:

$$\sigma\big(E_\theta(x)\big) = p(d|x) = \frac{p_d(x)}{p_d(x) + p_n(x)}$$

Taking the inverse sigmoid (logit function) yields

$$E_\theta(x) = \sigma^{-1}\big(p(d|x)\big) = \ln\left(\frac{p(d|x)}{1 - p(d|x)}\right)$$

Substituting the expression for $p(d|x)$ into this equation, we obtain

$$E_\theta(x) = \ln\left(\frac{\frac{p_d(x)}{p_d(x)+p_n(x)}}{1 - \frac{p_d(x)}{p_d(x)+p_n(x)}}\right) = \ln\left(\frac{p_d(x)}{p_n(x)}\right)$$

Thus, the closed-form solution for the optimal energy function is

$$E_{\theta^*}(x) = \ln\left(\frac{p_d(x)}{p_n(x)}\right)$$

Question 4.2

Let's recall that we are given a dataset D of positive pairs $\{(x_i, x_i')\}_{i=1}^N$ and that the negative pairs are generated by pairing x_i with x_j' sampled independently from D. Let

$$f_\theta : \mathcal{X} \to S_d, \quad g_\phi : \mathcal{X}' \to S_d,$$

be mappings into the unit hypersphere, and define

$$z_i = f_\theta(x_i), \quad y_i = g_\phi(x_i').$$

The contrastive objective is given by

$$\mathcal{L}_{\text{contr.}}(f_\theta, g_\phi, D) := \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\exp\left(z_i^\top y_i\right)}{\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)}\right]$$

where M is the number of negative samples.

Notice that for a given i, we can rewrite the loss as

$$-\log\frac{\exp\left(z_i^\top y_i\right)}{\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)} = -\left(\log\left(\exp\left(z_i^\top y_i\right)\right) - \log\left(\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)\right)\right) = -z_i^\top y_i + \log\left(\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)\right)$$

This gives us the first two terms $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{unif}}$. Dividing and multiplying the second term by M gives

$$\log\left(\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)\right) = \log\left(M \cdot \frac{1}{M}\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)\right) = \log M + \log\left(\frac{1}{M}\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)\right)$$

Thus, the loss for sample i can be decomposed as:

$$-\log\frac{\exp\left(z_i^\top y_i\right)}{\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)} = -z_i^\top y_i + \log\left(\frac{1}{M}\sum_{j=1}^{M}\exp\left(z_i^\top y_j\right)\right) + \log M$$

Averaging over i, we write the overall contrastive loss as

$$\mathcal{L}_{\text{contr.}}(f_\theta, g_\phi, D) = \mathcal{L}_{\text{unif}}(f_\theta, g_\phi, D) - \mathcal{L}_{\text{align}}(f_\theta, g_\phi, D) + \log M$$

where we define the following three components:

Alignment Term $\left(\mathcal{L}_{\text{align}}\right)$ This term directly measures how similar the representations of the positive pairs (z_i, y_i) are. Maximizing $z_i^\top y_i$ promotes similarity between the representations and encourages the pair to be aligned (i.e., to have a large inner product).

$$\mathcal{L}_{\text{align}}(f_\theta, g_\phi, D) := \frac{1}{N}\sum_{i=1}^{N} z_i^\top y_i$$

Uniformity Term $\left(\mathcal{L}_{\mathbf{unif}}\right)$ This term averages the similarity between the anchor z_i and the randomly sampled y_j. It penalizes the concentration of the negative pairs and encourages the overall set of representations (from the negative samples) to be uniformly distributed on the hypersphere. In effect, it prevents the model from collapsing all features into a narrow region.

$$\mathcal{L}_{\text{unif}}(f_\theta, g_\phi, D) := \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{M}\sum_{j=1}^{M} \exp\left(z_i^\top y_j\right)\right)$$

Constant Term This is a normalization constant that emerges naturally from the formulation when considering M negative samples. It does not affect the optimization process.

$$\log M$$

Behavior as $M \to \infty$ When M tends to infinity, the term

$$\frac{1}{M}\sum_{j=1}^{M} \exp\left(z_i^\top y_j\right)$$

converges to the expectation of $\exp\left(z_i^\top y\right)$ over the distribution of negative samples. As a consequence, the uniformity term encourages the representations to be uniformly spread out over the hypersphere. This connects to the information maximization principle in that a representation which is well spread carries more diverse, non-redundant information.

In summary, the overall contrastive objective aims to maximize the mutual information between positive pairs (by aligning z_i and y_i) while simultaneously promoting diversity and/or uniformity of the feature representations (through $\mathcal{L}_{\text{unif}}$). This balance prevents representational collapse and encourages the model to capture informative, discriminative features.

Question 4.3

Assume that we now have a network that jointly encodes the pair (x, x') into a representation on the unit hypersphere, while we retain the previous mapping for the second view:

$$h_\psi : \mathcal{X} \times \mathcal{X}' \to S_d \quad \text{and} \quad g_\phi : \mathcal{X}' \to S_d$$

For each positive pair (x_i, x_i') from the dataset D, we define

$$z_i = h_\psi(x_i, x_i') \quad \text{and} \quad y_i = g_\phi(x_i')$$

We also generate negative pairs by pairing x_i with x_j' for $j \neq i$.

Following the structure of the contrastive objective $\mathcal{L}_{\text{contr.}}$ from the previous question, we derive an analogous objective function $\mathcal{L}_*(h_\psi, g_\phi, D)$ that encourages the joint representation z_i to align with the corresponding y_i while repelling negatives. Thus, we define the loss as

$$\mathcal{L}_*(h_\psi, g_\phi, D) := \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\exp\big(z_i^\top y_i\big)}{\sum_{j=1}^{M}\exp\big(z_i^\top y_j\big)}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\exp\big(h_\psi(x_i, x_i')^\top g_\phi(x_i')\big)}{\sum_{j=1}^{M}\exp\big(h_\psi(x_i, x_i')^\top g_\phi(x_j')\big)}\right]$$

where M denotes the number of negative samples.

Although the mathematical formulation of $\mathcal{L}_*(h_\psi, g_\phi, D)$ is nearly identical to the standard contrastive loss $\mathcal{L}_{\text{contr.}}$, the crucial difference lies in the representation learning. In $\mathcal{L}_*(\cdot)$, the joint encoder h_ψ fuses information from both x and x', potentially capturing richer and more informative interactions between the modalities compared to the single-view encoder used in $\mathcal{L}_{\text{contr.}}$. Therefore, while the loss expressions are similar, they are not equivalent in terms of the learned representations or performance.

Question 4.4

We consider two different approaches for learning an EBM:

1. Using the standard contrastive objective $L_{\text{contr.}}$ (from Question 4.2), which typically learns separate encoders for different modalities and aligns them based on their inner products.
2. Using the joint objective $\mathcal{L}_*$ (from Question 4.3), which leverages a joint encoder h_ψ together with a separate text encoder g_ϕ to capture the combined multimodal correspondence.

Representational Flexibility While $L_{\text{contr.}}$ learns separate encoders whose outputs are aligned by inner product—allowing each modality to be indexed or fine-tuned independently. $\mathcal{L}_*$ fuses modalities via a joint encoder, capturing richer cross-modal interactions at the cost of modular reuse.

Computational Complexity The contrastive approach keeps parameter counts low and optimization per network simple, whereas the joint objective adds a larger fused network, increasing both parameter count and training overhead for simultaneous multimodal fusion.

Robustness & Overfitting Separate embeddings in $L_{\text{contr.}}$ serve as implicit regularizers that limit overfitting in each modality, while the joint encoder in $\mathcal{L}_*$ can over-parameterize cross-modal correlations and thus demands stronger explicit regularization.

Information Maximization $L_{\text{contr.}}$ explicitly decomposes into alignment and uniformity terms to transparently maximize mutual information, whereas $\mathcal{L}_*$ also encourages shared information but does so within a coupled latent space, potentially obscuring modality-specific discriminative features.