Homework 5

6.7960 Deep Learning
Fall 2025

Instructions: Complete the following questions. There are a total of 29 points for this homework. Each question is marked with its corresp onding points. We will be using this notebook. Please include any code you write in your report or simply convert your notebook to a PDF and submit that alongside your report. **Unless otherwise stated, please justify your answers.**

Collaboration: If you collaborate with other students on the homework, list the names of all your collaborators.

Submission: Upload a **PDF** of your response through **Canvas** by **11/12 at 11:59pm**.

Notation: We will use this set of math notation specified on course website, whose LaTeX source is available on Canvas. For example, c is a scalar, $\mathbf{b}$ is a vector and $\mathbf{W}$ is a matrix. You are encouraged (although not enforced) to follow this notation.

Variational Autoencoders (14pt)

Answer: Solution notebook

Variational autoencoders (VAEs), unlike standard autoencoders, are generative models. We would like to sample a vector from some standard normal distribution $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and have that sample map to some element in the data distribution $p(\mathbf{x})$.

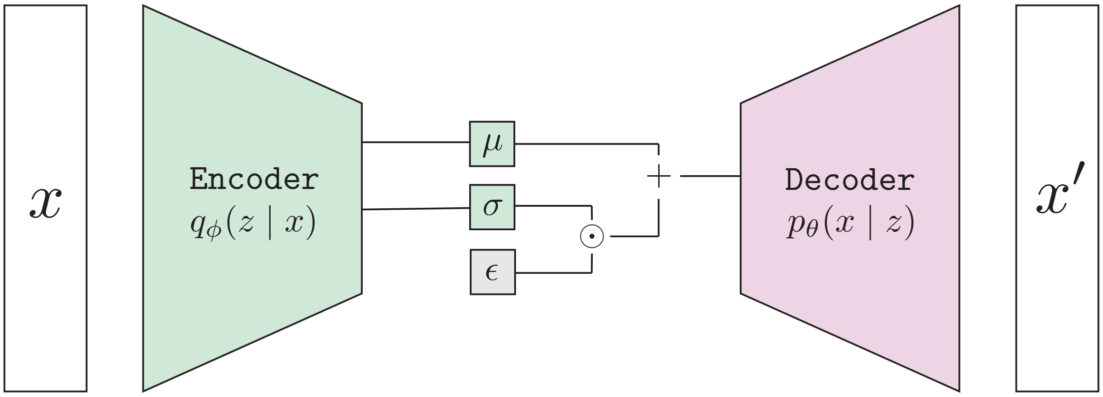

Figure 1: A variational autoencoder (VAE) diagram

Instead of directly maximizing $p(\mathbf{x})$, we instead aim to maximize a lower bound on $\log p(\mathbf{x})$

called the evidence lower bound (ELBO) Kingma and Welling [2019]:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{first term}} - \underbrace{D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))}_{\text{second term}}, \tag{1}$$

where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a fixed prior distribution over the latent $\mathbf{z}$.

1. **(1 pt)** Explain what each of each of the two terms in the above loss function (i.e., the first term and the second term) is doing. Please answer in a few sentences. Note that here we aim to *maximize* the objective function.

   Answer: The first term is reconstruction loss, which encourages reconstructing the original image. The second term is the KL divergence between the learned distribution of the latent $\mathbf{z}$ and the prior, and it encourages the encoder to generate standard normal $\mathbf{z}$.

2. **(1 pt)** In a VAE, each input datapoint x is mapped to a distribution in the latent space by encoding each x as both a mean μ and variance σ^2:

$$\mu_z = f^\mu_\phi(x),$$
$$\sigma^2_z = f^\sigma_\phi(x).$$

   In this question, we will think about how rich the structure of the means in the embedding space (i.e., the quantities $f^\mu_\phi(x)$ for all input datapoints) alone must be in a fully-trained VAE. Suppose you successfully train a VAE such that the marginal $q_\phi(\mathbf{z})$ becomes a unit Gaussian, and $p_\theta(x)$ is equal to the data distribution p_{data}. In other words, assume you achieved *perfect* encoder and decoder after training a VAE.

   (a) **(0.5 pt)** Does this imply that decoding only the means of the embeddings recovers the data distribution? In other words, are the sets of means of embedding sufficient for data distribution recovery, or are the variances of embeddings also necessary? Explain your answer in a few sentences.

   Answer: No, the means alone may not cover the full data distribution, such that the decoding only covers the full data distribution when combined with the variances.

   (b) **(0.5 pt)** Does this imply that the means of the embeddings (i.e., $f^\mu_\phi(x)$ above) are Gaussian distributed? In other words, is the random variable $f^\mu_\phi(x)$ Gaussian when $x \sim p_{\text{data}}$? Explain your answer in a few sentences.

   Hint: Note that the random variable $\mathbf{z} \sim \mathcal{N}(f^\mu_\phi(x), f^\sigma_\phi(x))$ is unit Gaussian by assumption, when $x \sim p_{\text{data}}$. Does this imply that its mean is also Gaussian?

   Answer: No. Imagine two subsets of x. Within each, $f^\sigma_\phi(x)$ is constant, and $f^\mu_\phi(x) \sim \mathcal{N}(0, 1 - f^\sigma_\phi(x)$, so the marginal $z \sim \mathcal{N}(0, 1)$ within that subset. However, the mixture of the two distributions of means (two gaussians with different standard deviations) will not itself be gaussian.

Grader comments: Please clarify this problem for the next year!

3. **(1 pt)** Suppose you successfully train an autoencoder (not VAE) such that $g(f(\mathbf{x})) = \mathbf{x}$ where f is the deterministic encoder and g is the deterministic decoder. Does this imply that $g(\mathbf{z}) \sim p_{\texttt{data}}$ when $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$? Explain your answer in a few sentences.

   Answer: No, the latents are distributed according to an arbitrary distribution, so sampling from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ will likely not return the original data distribution

4. **(6 pts)** Implement the VAE architecture in the provided colab.

5. **(1 pt)** Train the VAE model on the provided FashionMNIST dataset.

6. **(4 pt)** Using your trained model, complete the `plot_latents` function such that for a given pair of latent dimensions (i, j) the `plot_latents` plots a 10×10 grid of images sampled from different pairs of latent values in dimensions i, j (Figure 2). Latent dimensions not equal to i or j should be set to zero.

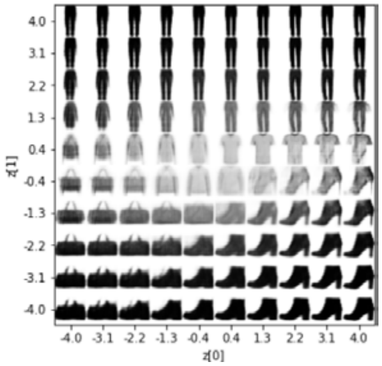

Figure 2: Example latent visualization.

Grader comments: Commented out question 7 here about β-VAE. Code cells were deleted in the collab file.

Diffusion Models (17pt)

A diffusion process works by sequentially adding noise to a given input x_0 for T timesteps, and our goal is to learn a function that removes the added noise (i.e., *denoising*). While there are many variations around this central concept, we will be looking closely at one particular approach from [Ho et al., 2020].

As T approaches infinity, this process results in noise drawn from a zero-mean isotropic Gaussian at $\mathbf{x}_T$. This process is depicted in Figure 3. The random variable $\mathbf{x}_t$ conditioned on $\mathbf{x}_{t-1}$ is distributed according to $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ with the following form:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\,\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad \text{where} \;\; \alpha_t = 1 - \beta_t. \tag{2}$$

Note: the distribution q refers to the joint distribution over all $\mathbf{x}_0, \dots, \mathbf{x}_T$. Conditionals of this distribution are denoted as $q(\cdot|\cdot)$ for subsets of the variables in the joint configuration.

Let's first try to understand the properties of the forward diffusion process in terms of $q(\mathbf{x}_t|\mathbf{x}_0)$. We will use these properties later to derive a loss for the reverse diffusion process.

7. **(2pt)** Adding two independent Gaussian random variables results in another Gaussian random variable. Using this property, one can derive a closed-form representation for the Gaussian distribution $q(\mathbf{x}_t|\mathbf{x}_0)$ in terms of α_t and $\mathbf{x}_0$. Prove that

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{3}$$

where $\bar{\alpha}_t$ refers to $\prod_{s=1}^{t} \alpha_s$. Provide a detailed explanation of how you arrived at your answer.

Answer: Using $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\,\mathbf{x}_{t-1}, \beta_t\mathbf{I})$, we conclude that $\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}Z_t$ where $Z_t = \mathcal{N}(0, \mathbf{I})$. Similarly, we get

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}Z_t \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\beta_{t-1}}Z_{t-1}) + \sqrt{\beta_t}Z_t \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{\alpha_t\beta_{t-1}}Z_{t-1} + \sqrt{\beta_t}Z_t \\
&= \cdots \\
&= \sqrt{\alpha_t\alpha_{t-1}\dots\alpha_1}\mathbf{x}_0 \\
&\quad + \sqrt{\beta_t}Z_t + \sqrt{\alpha_t\beta_{t-1}}Z_{t-1} + \dots + \sqrt{\alpha_t\alpha_{t-1}\dots\alpha_2\beta_1}Z_1,
\end{aligned}$$

where $Z_1, Z_2 \dots, Z_t = \mathcal{N}(0, \mathbf{I})$ are independent random variables. Therefore, we have

$$\mathbb{E}[\mathbf{x}_t|\mathbf{x}_0] = \sqrt{\bar{\alpha}_t}\mathbf{x}_0,$$

and we can show the covariance by induction.

$$\begin{aligned}
\mathrm{cov}(\mathbf{x}_t|\mathbf{x}_0) &= (\beta_t + \alpha_t\beta_{t-1} + \dots + \alpha_t\alpha_{t-1}\dots\alpha_2\beta_1)\mathbf{I} \\
&= (1 - \alpha_t + \alpha_t(\beta_{t-1} + \dots + \alpha_{t-1}\dots\alpha_2\beta_1))\mathbf{I} \\
&= (1 - \alpha_t + \alpha_t\ \mathrm{cov}(\mathbf{x}_{t-1}|\mathbf{x}_0))\mathbf{I} \\
&= (1 - \alpha_t + \alpha_t(1 - \bar{\alpha}_{t-1}))\mathbf{I} \\
&= (1 - \alpha_t\bar{\alpha}_{t-1})\mathbf{I} \\
&= (1 - \bar{\alpha}_t)\mathbf{I}.
\end{aligned}$$

Therefore,

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}).$$

Grader comments: The final answer is $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$.

8. **(1pt)** Let us calculate how far $\mathbf{x}_T$ is from being $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in KL divergence. For any d-dimensional Gaussian distribution $\mathcal{N}(\mathbf{a}, \Sigma)$, we have

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mathbf{a}, \Sigma) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})\right) = \frac{1}{2}\left(\|\mathbf{a}\|_2^2 + \mathrm{tr}(\Sigma) - \log(\det(\Sigma)) - d\right)$$

where $\mathrm{tr}(\Sigma)$ is the sum of diagonal entries of the covariance matrix Σ. Use the above formula and compute $D_{\mathrm{KL}}\left(q(\mathbf{x}_T|\mathbf{x}_0) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})\right)$. Conclude that if T is large, then $q(\mathbf{x}_T|\mathbf{x}_0)$ is close to being unit Gaussian.

Answer:

$$D_{\mathrm{KL}}\left(q(\mathbf{x}_T|\mathbf{x}_0) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})\right) = \frac{1}{2}\left(\bar{\alpha}_T\|\mathbf{x}_0\|_2^2 + d(1-\bar{\alpha}_T) - d\log(1-\bar{\alpha}_T) - d\right).$$

Thus, since for large T we know that $\bar{\alpha}_T$ goes to zero, the above KL divergence also goes to zero.

Now, our real goal is to learn a function $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_t)$ that removes the noise added (i.e., *denoising*). Indeed, learning such a function enables us to remove noise from $\mathbf{x}_t$ to recover $\mathbf{x}_{t-1}$. By starting with a completely noisy image (i.e., a unit Gaussian), we can perform sequential denoising to generate an image with this approach.

One strategy to learning $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is to minimize the KL-divergence between this distribution and the true reverse process $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$:

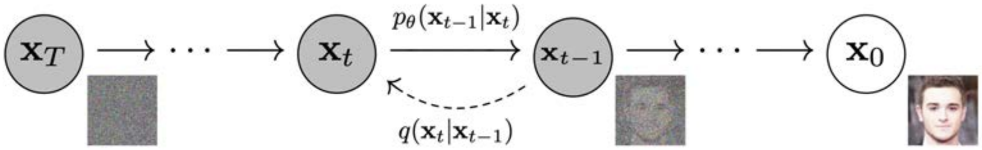
$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

Figure 3: The directed graphical model for the diffusion process.

But what is the form of the reverse diffusion process $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$? We will derive this in the following questions.

To derive this expression, we use the closed form for the expression $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$, which works out to be (you may refer to Ho et al. [2020] for more details):

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \tag{4}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \tag{5}$$

$$\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \cdot \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \cdot \mathbf{x}_0 \tag{6}$$

Also, note from Eqn. 3, we have that

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon_t, \tag{7}$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a Gaussian random variable with unit variance.

Using the above expressions:

9. **(1pt)** Derive an expression for mean and the variance of $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \epsilon_t)$ in terms of the above quantities.

Answer: According to the given formula, we have

$$\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \cdot \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \cdot \mathbf{x}_0.$$

Also, we have $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon_t$. Therefore, we can solve this for $\mathbf{x}_0$ to find

$$\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{x}_t - \frac{\sqrt{(1-\bar{\alpha}_t)}}{\sqrt{\bar{\alpha}_t}}\epsilon_t.$$

Now, to obtain the final result, we replace the above formula into the formula for $\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0)$ and obtain

$$\tilde{\mu}(\mathbf{x}_t, \epsilon_t) = \tilde{\mu} \quad \mathbf{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{x}_t - \frac{\sqrt{(1-\bar{\alpha}_t)}}{\sqrt{\bar{\alpha}_t}}\epsilon_t \Big)\Big(\tag{8}$$

$$= \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \cdot \mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \quad \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{x}_t - \frac{\sqrt{(1-\bar{\alpha}_t)}}{\sqrt{\bar{\alpha}_t}}\epsilon_t \Big)\Big(\tag{9}$$

$$= \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t). \tag{10}$$

Moreover, the covariance matrix is also given as $\tilde{\beta}_t\mathbf{I}$.

Grader comments: Final answer is:

$$\tilde{\mu}(\mathbf{x}_t, \epsilon_t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t).$$

10. **(1pt)** Derive an expression for $D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\epsilon_t)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))$. Assume that we parameterize $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)=\mathcal{N}(\mu_\theta(\mathbf{x}_t,t),\Sigma_t)$, where $\Sigma_t=\sigma_t^2\mathbf{I}$.

    Hint: You can use the following formula for the KL divergence between two d-dimensional Gaussian vectors:

$$D_{\mathrm{KL}}\left(\mathcal{N}\left(\mathbf{a},\Sigma_a\right)\parallel\mathcal{N}\left(\mathbf{b},\Sigma_b\right)\right)=\frac{1}{2}\Big(\log(\det(\Sigma_b))-\log(\det(\Sigma_a))-d \\ +(\mathbf{a}-\mathbf{b})^T\Sigma_b^{-1}(\mathbf{a}-\mathbf{b})+\mathrm{tr}(\Sigma_b^{-1}\Sigma_a)\Big).\Bigg($$

    Answer: The final answer is derived after replacing the mean and covariance matrices derived in the previous question in the provided hint.

Congrats! You now have an objective function for denoising, which you can use to train a diffusion model.

Grader comments: **The old answer:** In practice, however, people often reparameterize the optimization problem in terms of $\epsilon_t=\mathbf{x}_t-\mathbf{x}_{t-1}$ and predict ϵ_t with a model $\epsilon_\theta(\mathbf{x}_t,t)$. Let's formalize this idea!

In practice, people often parameterize the optimization problem in terms of the total noise $\epsilon_t=\mathrm{x}_t-\mathrm{x}_0$ and predict ϵ_t with a model $\epsilon_\theta(\mathrm{x}_t,t)$. Let's formalize this idea!

11. **(2pt)** Let us use the ϵ-parameterization to express $\mu_\theta(\mathbf{x}_t,t)$ as

$$\mu_\theta(\mathbf{x}_t,t)=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right)\Bigg(\tag{11}$$

    Show that if we fix $\sigma_t^2=\tilde{\beta}_t$, then

$$\arg\min_\theta D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\epsilon_t)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))=\arg\min_\theta\|\epsilon_t-\epsilon_\theta(\mathbf{x}_t,t)\|_2^2.$$

    Answer: The answer is derived after replacing the mean and covariance matrices in the result of the previous part.

    Grader comments: Answer to the previous question is something like half of

$$\frac{1}{\sigma_t^2}\|\mu_\theta(x_t,t)-\tilde{\mu}(x_t,t)\|+d\left[\frac{\tilde{\beta}}{\sigma_t^2}-\log\frac{\tilde{\beta}}{\sigma_t^2}-1\right]\Bigg($$

    We can substitute in our new values. From 9 we have

$$\tilde{\mu}(x_t,\epsilon_t)=\frac{1}{\sqrt{\alpha_t}}\left[x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_t\right]\Bigg($$

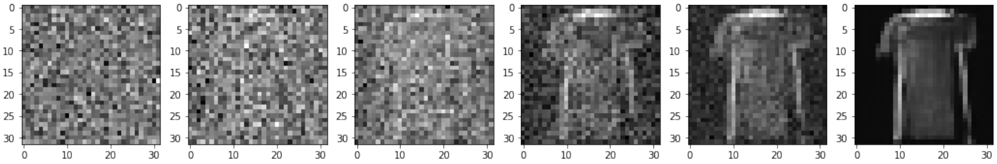

Figure 4: Diffusion inference figure example

So the difference of the two μ is $\frac{1-\alpha_t}{\sqrt{\alpha_t}\sqrt{1-\bar{\alpha}_t}}\left[\epsilon_\theta(x_t, t) - \epsilon_t\right]$

They should then note that the α_t is a positive constant with respect to θ, so it doesn't affect the argmin.

Also $1 - \alpha_t = \beta_t$ so they may write this differently.

Thus, we obtain the following diffusion modeling loss, which is used in the coding section of this problem set:

$$\mathcal{L}_{\texttt{Diffusion}} = \|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2. \tag{12}$$

12. **(6pt)** In the ′ colab, complete the `FIXME`s and train the diffusion model.

13. **(2pt)** Create a plot with six horizontally-aligned subfigures showing the diffusion model inference $\mathbf{x}_t$ for $t \in \{200, 100, 50, 20, 10, 0\}$. An example can be seen in Figure 4. Your plots will look different.

14. **(2pt)** There are many similarities between VAEs and diffusion models. Answer each question with a sentence or two.

    (a) **(1pt)** What is the equivalent of the encoder, decoder, and latent variable in the diffusion model?

    Answer: The encoder is the product of all $q(x_t|x_{t-1})$, the decoder is the product of all $p_\theta(x_{t-1}|x_t)$. The latent is $\mathbf{x}_T$.

    (b) **(1pt)** What is the difference between the encoder in a diffusion model and VAE?
    Answer: The encoder is not learned in the diffusion model. Alternatively, the latent is smaller than the input in a VAE.

References

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL https://arxiv.org/abs/2006.11239.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *CoRR*, abs/1906.02691, 2019. URL http://arxiv.org/abs/1906.02691.