

CS 446 / ECE 449 — Homework 3

acard6

Version 1.1

Instructions.

- Homework is due **Tuesday, March 02, at noon CDT**.
- Everyone must submit individually at gradescope under **hw3** and **hw3code**.
- The “written” submission at **hw3** **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use L^AT_EX, markdown, google docs, MS word, whatever you like; but it must be typed!
- When submitting at **hw3**, gradescope will ask you to mark out boxes around each of your answers; please do this precisely!
- Please make sure your NetID is clear and large on the first page of the homework.
- Your solution **must** be written in your own words. Please see the course webpage for full academic integrity information. Briefly, you may have high-level discussions with at most 3 classmates, whose NetIDs you should place on the first page of your solutions, and you should cite any external reference you use; despite all this, your solution must be written in your own words.
- We reserve the right to reduce the auto-graded score for **hw3code** if we detect funny business (e.g., your solution lacks any algorithm and hard-codes answers you obtained from someone else, or simply via trial-and-error with the autograder).
- When submitting to **hw3code**, only upload **hw3_vae.py** and **hw3_gan.py**. Additional files will be ignored.

Version History.

- 1.0 Initial Version.
- 1.1 Update code/hw3_vae.py.

1. Variational Auto-Encoders [Written]

We use VAEs to learn the distribution of the data x . Let z denote the unobserved latent variable. We refer to the approximated posterior $q_\phi(z|x)$ as the encoder and to the conditional distribution $p_\theta(x|z)$ as the decoder. Use these names to answer the following questions.

- (a) We are interested in modeling data $x \in \{0, 1\}^G$. Hence, we choose $p_\theta(x|z)$ to follow G independent Bernoulli distributions. Recall, a Bernoulli distribution has a probability density function of

$$P(k) = \begin{cases} 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}.$$

Use \hat{y}_j to denote the $j^{\text{th}} \in [1, G]$ dimension of the decoder's output, and similarly x_j the $j^{\text{th}} \in [1, G]$ dimension of the data x . Write down the explicit form for $p_\theta(x|z)$ in terms of \hat{y}_j and x_j .

- (b) We further assume that $z \in \mathbb{R}^2$ and that $q_\phi(z|x)$ follows a multi-variate Gaussian distribution with an identity covariance matrix. What is the output dimension of the encoder?
- (c) We want to maximize the log-likelihood $\log p_\theta(x)$. To this end we introduce a joint distribution $p_\theta(x, z)$ and reformulate the log-likelihood via

$$\log p_\theta(x) = \log \sum_z q_\phi(z|x) \frac{p_\theta(x, z)}{q_\phi(z|x)}.$$

Use Jensen's inequality to obtain a bound on the log-likelihood and divide the bound into two parts, one of which is the Kullback-Leibler divergence $\text{KL}(q_\phi(z|x), p(z))$.

- (d) State at least two properties of the KL-divergence.
- (e) Recall, the evidence lower bound (ELBO) of the log likelihood, $\log p_\theta(x)$, is

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x), p(z)). \quad (1)$$

We can also write the ELBO as

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z) + \log p(z) - \log(q_\phi(z|x))]. \quad (2)$$

Practically, will training a VAE using the formulation in Eq. 1 be the same as the one in Eq. 2? If not, why use one formulation over another?

- (f) Observe that the ELBO in Eq. 1 works for any q_ϕ distribution. Is it a good idea to choose $q_\phi(z|x) := \mathcal{N}(0, I)$? In other words, why is an encoder necessary?
- (g) Let

$$q_\phi(z|x) = \frac{1}{\sqrt{2\pi\sigma_\phi^2}} \exp\left(-\frac{1}{2\sigma_\phi^2}(z - \mu_\phi)^2\right).$$

What is the value for the KL-divergence $\text{KL}(q_\phi(z|x), q_\phi(z|x))$ and why?

- (h) Further, let

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

Note the difference of the means for $p(z)$ and $q_\phi(z|x)$ while their standard deviations are identical. Assume that $\sigma = \sigma_\phi = \sigma_p$. What is the value for the KL-divergence $\text{KL}(q_\phi(z|x), p(z))$ in terms of μ_p , μ_ϕ and σ ?

- (i) Now, let $q_\phi(z|x)$ and $p(z)$ be arbitrary probability distributions. We want to find that $q_\phi(z|x)$ which maximizes

$$\sum_z q_\phi(z|x) \log p_\theta(x|z) - \text{KL}(q_\phi(z|x), p(z))$$

subject to $\sum_z q_\phi(z|x) = 1$. Ignore the non-negativity constraints. State the Lagrangian and compute its stationary point, i.e., solve for $q_\phi(z|x)$ which depends on $p_\theta(x|z)$ and $p(z)$. Make sure to get rid of the Lagrange multiplier.

- (j) Which of the following terms should $q_\phi(z|x)$ be equal to: (1) $p(z)$; (2) $p_\theta(x|z)$; (3) $p_\theta(z|x)$; (4) $p_\theta(x, z)$.

Solution.

Answer A.) we know that the $p_\theta(x|z) = \prod_j p(x_j|z)$, knowing that $p(x_j = 1) = \hat{y}_j$ and $p(x_j = 0) = 1 - \hat{y}_j$ we can formulate $p_\theta(x|z) = \prod_{j=1}^G (\hat{y}_j)^{x_j} (1 - \hat{y}_j)^{1-x_j}$

Answer B.) if $z \in \mathbb{R}^2$ and $q_\Phi(z|x) = \mathcal{N}(\mu, \sigma^2 I)$, and if I is 2d the the encoder is also 2d since the identity matrix has to match up with the latent variable.

Answer C.) By reformulating the log-likelihood to $\log p_\theta(x) = \log \sum_z q_\Phi(z|x) \frac{p_\theta(x, z)}{p_\theta(z|x)}$ we can then write it as $\log \sum_z q_\Phi(z|x) \frac{p_\theta(x, z)}{q_\Phi(z|x)} \frac{q_\Phi(z|x)}{p_\theta(z|x)}$ (since we can approximate $p_\theta(z|x) \approx q_\Phi(z|x)$) we then use Jensen's inequality on the log to get $\log \sum_z q_\Phi(z|x) \frac{p_\theta(x, z)}{q_\Phi(z|x)} \frac{q_\Phi(z|x)}{p_\theta(z|x)} \leq \sum_z q_\Phi(z|x) \log \left(\frac{p_\theta(x, z)}{q_\Phi(z|x)} \frac{q_\Phi(z|x)}{p_\theta(z|x)} \right)$ $= \sum_z q_\Phi(z|x) \log \frac{p_\theta(x, z)}{q_\Phi(z|x)} + \sum_z q_\Phi(z|x) \log \frac{q_\Phi(z|x)}{p_\theta(z|x)} = \mathcal{L}(p_\theta, q_\Phi) + KL(q_\Phi, p_\theta)$ so $\mathcal{L}(p_\theta, q_\Phi) + KL(q_\Phi, p_\theta) \geq \mathcal{L}(p_\theta, q_\Phi)$

Answer D.) Properties of KL divergence include, relative entropy is always non-negative, no upper bound exist for the general case, and relative entropy remains well-defined for continuous distributions, and furthermore is invariant under parameter transformations.

Answer E.) yes they are the same since we can rewrite the expectation as $\mathbb{E}_{q_\Phi(z|x)}[\log p(x|z) + \log p(z) - \log q(z|x)] = \mathbb{E}_{q_\Phi(z|x)}[\log p(x|z)] + \mathbb{E}_{q_\Phi(z|x)}[\log p(z) - \log q(z|x)]$ using the subtraction properties of 2 logs we can then combine them and expand the definition of expectation to become $\mathbb{E}_{q_\Phi(z|x)}[\log p(x|z)] + \mathbb{E}_{q_\Phi(z|x)}[\log \frac{p(z)}{q(z|x)}] = \mathbb{E}_{q_\Phi(z|x)}[\log p(x|z)] + \sum_z q_\Phi(z|x) \log \frac{p(z)}{q_\Phi(z|x)}$ we can then substitute with the definition of KL divergence of $-KL(q_\Phi, p) = \int_z q_\Phi(z|x) \log \frac{p(z)}{q_\Phi(z|x)}$. By substituting we can reach the conclusion that eq. 1 and eq. 2 are in fact the same

Answer F.) if we use the identity matrix for the prior we assume that each sample of q is i.i.d with a normal distribution, which may not perfectly reflect the reality of the nature of real world data. An encoder is necessary to map out helps us extract the most from an image in the form of data and establish useful correlations between various inputs within the network.

Answer G.) KL-divergence is defined part E can be further simplified as $-KL(q_\Phi, p) = \int_z q_\Phi(z|x) \log \frac{p(z)}{q_\Phi(z|x)} = -\int q \log q + \int q \log p$ if $p=q$ and q is a gaussian distribution then we get that $\int_z q \log q dz = \int_z \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(z-\mu_q)^2}{2\sigma_q^2}\right) * \left(-\frac{1}{2} \log(2\pi\sigma_q^2)\right) * \left(-\frac{(z-\mu_q)^2}{2\sigma_q^2}\right) dz = -\frac{1}{2}(1 + \log 2\pi\sigma_q^2)$, doing the same for $KL(q, p)$ we get $-\int_z q \log p dz = -\int_z \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(z-\mu_p)^2}{2\sigma_p^2}\right) * \left(-\frac{1}{2} \log(2\pi\sigma_p^2)\right) * \left(-\frac{(z-\mu_p)^2}{2\sigma_p^2}\right) dz = \frac{1}{2} \log 2\pi\sigma_p^2 + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2}$ combining the 2 parts we can simplify and get $KL(q, p) = \log \frac{\sigma_q^2}{\sigma_p^2} + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} - \frac{1}{2}$ since they're equal $\log(1)=0$, the means cancel, and the ratio of the variance is 1 which resolves to 1/2

Answer H.) using the equation of KL from the last part of $KL(q, p) = \log \frac{\sigma_p}{\sigma_\Phi} + \frac{\sigma_\Phi^2 + (\mu_\Phi - \mu_p)^2}{2\sigma_p^2} - \frac{1}{2}$ since the variance/std is the same between p and q their ratio is just 1 leaving the equation to be equal to $K(q, p) = \log \frac{\sigma}{\sigma} + \frac{\sigma^2 + (\mu_\Phi - \mu_p)^2}{2\sigma^2} - \frac{1}{2} = \frac{\sigma^2 + (\mu_\Phi - \mu_p)^2}{2\sigma^2} - \frac{1}{2}$

Answer I.) $L(q_\Phi, \lambda) = \sum_z q_\Phi \log p_\theta - KL(q_\Phi, p(z)) + \lambda((\sum_z q_\Phi(z|x)) - 1)$ $\frac{\delta L}{\delta q_\Phi} = \log p_\theta - \log q_\Phi - 1 - \lambda = 0 \rightarrow \log q_\Phi = \log p_\theta - \lambda - 1 \rightarrow$

$\sum_z q_\Phi = \sum_z \exp(\log p_\theta - (\lambda + 1)) = 1$ take the log of 1 and q to get
 $\log \sum_z \exp(\log p_\theta - (\lambda + 1)) = 0 \rightarrow \frac{\delta}{\delta \lambda} \sum_z \exp(\log p_\theta - (\lambda + 1)) = \sum_z \exp(\log p_\theta - \lambda) = 1$
multiply by p(z) on both sides $\sum_z p(z) \exp(\log p_\theta - \lambda) = p(x)$, take the log of both sides
 $\log p(x) = \log \sum_z p(z) \exp(\log p_\theta - \lambda) = \sum_z p(z) \log p_\theta - \lambda$ using jensen ineq we note that $\log p(x) <$
 $\sum_z p(z) \log p_\theta - \lambda =$
 $p(x) \leq \exp \sum_z p(z) \log p_\theta - \lambda$ therefore $q(z|x)$ is maximized by $q_\Phi = p(z) \exp(\log(p_\theta(x|z)))$, when
we get rid of the lagrangian by taking the sum over all q over z is equal to 1

Answer J.) $q_\Phi(z|x)$ should be set equal to $p_\theta(z|x)$ since we can closely approximate probability $p(z|x)$ as $q(z|x)$

2. Variational Auto-Encoders [Coding]

In this assignment, you will implement a Variational Autoencoder and train it on MNIST digits. Each datapoint x in the MNIST dataset is a 28×28 grayscale image (i.e., pixel values are between 0 and 1) of a handwritten digit in $\{0, \dots, 9\}$, and a label indicating which number. The prior over each digit's latent representation z is a multivariate standard normal distribution, i.e., $z \sim \mathcal{N}(0, I)$. For all questions, we set the dimension of the latent space D_z to 2. Given the latent representation z for an image, the distribution over all 784 pixels in the image is given by a product of independent Bernoulli, whose characteristic probabilities are given by the output of a neural network $f_\theta(z)$ (the decoder):

$$p_\theta(x|z) = \prod_{d=1}^{784} \text{Ber}(x_d | f_\theta(z)). \quad (3)$$

Relevant files: `hw3_vae.py`, `hw3_utils.py`.

- (a) **Decoder Architecture.** Given a latent representation z , the decoder produces a 784-dimensional vector representing the Bernoulli distribution characteristic probability, i.e., the probability for every pixel in the image being labeled 1. Define the decoder parameters in the method `__init__` of the `Decoder` class and implement the corresponding `forward` function. The decoder architecture is a multi-layer perceptron (i.e., a fully-connected neural network), with two hidden layers, followed each by a non linearity: `tanh` after the first layer and `sigmoid` after the second layer. The hidden dimension is set to 500 units.
- (b) **Distributions.**
 - i. Implement the method `logpdf_diagonal_gaussian` that, given a latent representation z , a mean μ and the variance σ^2 , outputs the log-likelihood of the normal distribution $\mathcal{N}(\mu, \sigma^2 I)$.
 - ii. Implement a function `logpdf_bernoulli` that, given a sample x and a probability p , outputs the log-likelihood of a Bernoulli distribution.
 - iii. Implement the function `sample_diagonal_gaussian` which uses the reparametrization trick to sample z from Diagonal Gaussian $z \sim \mathcal{N}(\mu, \sigma^2 I)$.
 - iv. Implement the function `sample_Bernoulli` which samples a configuration x from a Bernoulli distribution characterized by a probability p .
- (c) **Variational Objective.** Complete the function `elbo` with the ELBO loss implementation corresponding to Eq. 2.
- (d) **Training.** Train the model for 200 epochs. **Hint:** Run the `main` function and make sure the number of epochs is set-up correctly in `parse_args`.
- (e) **Visualization.**
 - i. **Samples from the generative model.** Complete the method `visualize_data_space` following the instructions:
 - Sample a z from the prior $p(z)$. Use `sample_diagonal_gaussian`.
 - Use the generative model to parameterize a Bernoulli distribution over x given z . Use `self.decoder` and `array_to_image`. Plot this distribution $p(x|z)$.
 - Sample x from the distribution $p(x|z)$. Plot this sample.
 - Repeat the steps above for 10 samples z from the prior. Concatenate all your plots into one 10×2 figure where the first column is the distribution over x and the second column is a sample from this distribution. Each row will be a new sample from the prior. Hint: use the function `concat_images`.
 - Attach the figure to your report.
 - ii. **Latent space visualization.** Produce a scatter plot in the latent space, where each point in the plot represents a different image in the training set. Complete the method `visualize_latent_space` following the instructions:

- Encode each image in the training set. Use *self.encoder*.
 - Plot the mean vector μ of $q_\phi(z|x)$ in the 2D latent space with a scatter plot. Make sure to color each point according to the class label (0 to 9).
 - Attach the scatter plot to your report.
- iii. **Interpolation between two classes.** Complete the method *visualize_inter_class_interpolation* following the instructions:
- Sample 3 pairs of data points (*self.train_images*) with different classes (*self.train_labels*).
 - Encode the data in each pair, and take the mean vectors. Note that the encoder produces a mean vector and a variance one.
 - Interpolate between these mean vectors. We denote the output by z_α , with $\alpha \in [0, 1]$ and the interpolation step being 0.1. Hint: use the function *interpolate_mu*.
 - Along the interpolation, plot the distributions $p(x|z_\alpha)$ in the same figure.
 - Use *concat_images* to concatenate these plots into one figure.
 - Attach the plot to your report.

Solution.

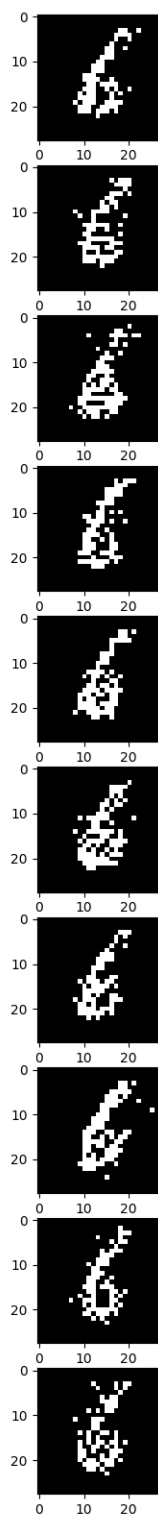
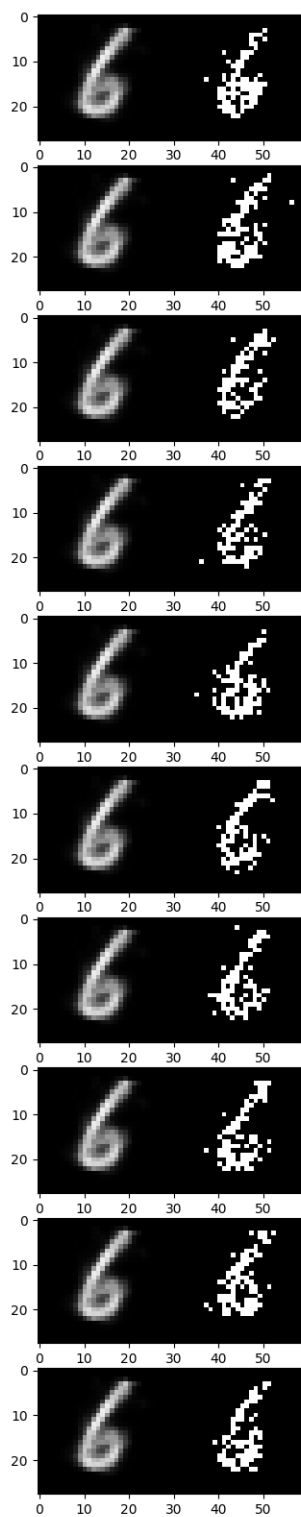


figure 1: epoch = 0

3. Generative Adversarial Networks [Written]

Here we discuss distribution-comparison-related problems in Generative Adversarial Networks (GANs).

- (a) What is the cost function for classical GANs? Use $D_\omega(x)$ as the discriminator and $G_\theta(x)$ as the generator.
- (b) Assume arbitrary capacity for both discriminator and generator. In this case we refer to the discriminator using $D(x)$, and denote the distribution on the data domain induced by the generator via $p_G(x)$. State an equivalent problem to the one asked for in part (a), by using $p_G(x)$.
- (c) Assuming arbitrary capacity, derive the optimal discriminator $D^*(x)$ in terms of $p_{\text{data}}(x)$ and $p_G(x)$.

Hint: you can think of fixing generator $G(\cdot)$ to find the optimal value for discriminator $D(\cdot)$.

- (d) Assume arbitrary capacity and an optimal discriminator $D^*(x)$ from (c), show that the optimal generator $G^*(x)$ generates the distribution $p_G^* = p_{\text{data}}$, where $p_{\text{data}}(x)$ is the data distribution.

Hint: you may need the Jensen-Shannon divergence:

$$\text{JSD}(p_{\text{data}}, p_G) = \frac{1}{2} \text{KL}(p_{\text{data}}, M) + \frac{1}{2} \text{KL}(p_G, M),$$

where $M = \frac{1}{2}(p_{\text{data}} + p_G)$.

- (e) More recently, researchers have proposed to use the Wasserstein distance instead of divergences to train the models since the KL divergence often fails to give meaningful information for training. Consider three distributions, $\mathbb{P}_1 \sim U[0, 1]$, $\mathbb{P}_2 \sim U[0.5, 1.5]$, and $\mathbb{P}_3 \sim U[1, 2]$, where $U[a, b]$ is uniform distribution over $[a, b]$. Calculate $\text{KL}(\mathbb{P}_1, \mathbb{P}_2)$, $\text{KL}(\mathbb{P}_1, \mathbb{P}_3)$, $\mathbb{W}_1(\mathbb{P}_1, \mathbb{P}_2)$, and $\mathbb{W}(\mathbb{P}_1, \mathbb{P}_3)$, where $\mathbb{W}_1(\cdot, \cdot)$ is the Wasserstein-1 distance between two distributions.

Hint: this subproblem requires no *real* mathematical computation. What you need to do is to understand the intuitive meaning of KL-divergence and Wasserstein distance. You may find wiki of *Earth mover's distance* and *Wasserstein metric* useful.

Solution.

Answer A.) the classical cost function for a GAN includes the generators cost function and the discriminators loss. loss of the discriminator is $\text{loss}_d = -\sum_x \log(D_\omega(x)) - \sum_z \log(1 - D_\omega(G_\theta(z)))$. Where as the loss for the generator is $\text{loss}_g = -\sum_z \log(D_\omega(G_\theta(z)))$. So the cost is just $\text{cost} = \text{loss}_d + \text{loss}_g$

Answer B.) using $p_G(x)$ as the distribution on the data domain induced by the generator we get the

$$-\int_x p_{\text{data}}(x) \log(D_\omega(x)) dx - \int_z p_z(z) \log(1 - D_\omega(G_\theta(z))) dz =$$

$$-\int_x p_{\text{data}}(x) \log(D_\omega(x)) + p_G(x) \log(1 - D_\omega(x)) dx = \int_x L(x, D, \dot{D}) dx$$

Answer C.) The optimal value for the discriminator assuming arbitrary capacity and a fixed $G(\cdot)$ and $D(\cdot)$ for our loss L , we find that we derive our w.r.t D and set it equal to 0 to get $\frac{\delta L(x, D, \dot{D})}{\delta D} = -\frac{p_{\text{data}}}{D} + \frac{p_G}{1-D} = 0 \rightarrow D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)}$

Answer D.) To find the optimal generator we begin by substituting $D^*(x)$ for $D(x)$ in the equation from part B, we see that we get

$$-\int_x p_{\text{data}}(x) \log(D^*(x)) + p_G(x) \log(1 - D^*(x)) dx =$$

$$-\int_x p_{\text{data}}(x) \log\left(\frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)}\right) + p_G(x) \log\left(1 - \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_G(x)}\right) dx$$

$$= -2\text{JSD}(p_{\text{data}}, p_G) + \log(4) = \text{JSD}(p_{\text{data}}, p_G) = -\text{KL}(p_{\text{data}}, M) - \text{KL}(p_G, M) + \log(4)$$

which gives us the optimal generator is $p_G(x) = p_{\text{data}}(x)$

Answer E.) the probability of a uniform distribution is $\frac{1}{b-a}$ so to find KL we integrate

$$\text{KL}(\mathbb{P}_1, \mathbb{P}_2) = \int_x p_{\mathbb{P}_1}(x) \log\left(\frac{p_{\mathbb{P}_1}(x)}{p_{\mathbb{P}_2}(x)}\right) = \int_{0.5}^1 \frac{1}{1-0} \log\left(\frac{\frac{1}{1-0}}{\frac{1}{1.5-0.5}}\right) dx = \int_{0.5}^1 \log(1) dx = 0,$$

given that the probability of $(\mathbb{P})_3$ is also 1, when put into the KL divergence it goes to 0. For

the Wasserstein distance we get that

$$W(\mathbb{P}_1, \mathbb{P}_2) = \min_{p_j(x, x') \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \mathbb{E}[|x - x'|] = \int f(x) dp_{\mathbb{P}_1} - \int f(x) dp_{\mathbb{P}_2} \rightarrow$$

$W(\mathbb{P}_1, \mathbb{P}_2) = (\int_0^1 |F(x) - G(x)| dx)$, where F and G are the cdf's of $\mathbb{P}_1, \mathbb{P}_2$ respectively, and since $G = 1/2F$ on the interval 0 to 1 we set $2 * dp_{\mathbb{P}_2} = dp_{\mathbb{P}_1}$ W to be

$$\int_0^1 2f(x) - \frac{1}{2} dx = \frac{3}{4}$$

We repeat the process for $W(\mathbb{P}_3, \mathbb{P}_3)$ and we note that \mathbb{P}_3 is out of the interval 0 to 1 so thus we just get that W is just $\int_0^1 \frac{1}{1} dx = 1$

4. Generative Adversarial Networks [Coding]

In this problem, you need to implement a Generative Adversarial Network and train it on MNIST digits.

Table 1: **Discriminator Architecture**

<i>Layer No.</i>	<i>Layer Type</i>	<i>Kernel Size</i>	<i>Stride</i>	<i>Padding</i>	<i>Output Channels</i>
1	conv2d	3	1	1	2
2	ReLU	-	-	-	2
3	MaxPool	2	2	-	2
4	conv2d	3	1	1	4
5	ReLU	-	-	-	4
6	MaxPool	2	2	-	4
7	conv2d	3	1	0	8
8	ReLU	-	-	-	8
9	Linear	-	-	-	1

Table 2: **Generator Architecture**

<i>Layer No.</i>	<i>Layer Type</i>	<i>Kernel Size</i>	<i>Stride</i>	<i>Padding</i>	<i>Bias</i>	<i>Output Channels</i>
1	Linear	-	-	-	✓	1568
2	LeakyReLU(0.2)	-	-	-	-	1568
3	Upsample(scale=2)	-	-	-	✗	32
4	conv2d	3	1	1	✓	16
5	LeakyReLU(0.2)	-	-	-	-	16
6	Upsample(scale=2)	-	-	-	✗	16
7	conv2d	3	1	1	✓	8
8	LeakyReLU(0.2)	-	-	-	-	8
9	conv2d	3	1	1	✓	1
10	sigmoid	-	-	-	-	1

- (a) Implement a discriminator **DNet** in `hw3_gan.py` using the architecture described in Tab. 1. Layers contain bias if corresponding `torch` function has an option for adding one.

Remark 1: From layer 8 to layer 9, you need to flatten each data entry from a matrix to a vector.

- (b) Implement a generator **GNet** in `hw3_gan.py` using the architecture described in Tab. 2.

Remark 2: From layer 2 to layer 3, you need to reshape each data to size $(32, 7, 7)$ in the format of *CHW*. Note, $1568 = 32 \times 7 \times 7$.

Remark 3: For (a) and (b), please define layers in `__init__` with **exactly the same** order as they appear in Tab. 1 and Tab. 2.

Remark 4: We have listed **all** layers for discriminator and generator. No need to add any extra components.

- (c) Implement the weight initialization function `_weight_init` in **DNet** and **GNet**: use `kaiming_uniform` for weights and 0 for the bias if the layer contains bias.

Hint: to iterate over all layers of an `nn.Module`, you may find `self.children()` useful. See `children()` function explained in <https://pytorch.org/docs/stable/generated/torch.nn.Module.html>.

- (d) Implement the loss function for the discriminator: `_get_loss_d` of GAN class in `hw3_gan.py`.
Hint: you may find `torch.nn.BCEWithLogitsLoss` useful.
- (e) Implement the loss function for the generator: `_get_loss_g` of GAN class in `hw3_gan.py`.
Hint: you may find `torch.nn.BCEWithLogitsLoss` useful.
- (f) Attach generated images after training.
Remark 5: the provided code default saves images during training. You can choose three of the saved ones and indicate the corresponding epochs.
Remark 6: with default training settings, you should obtain reasonable generated images after around 30 epochs.

Solution.

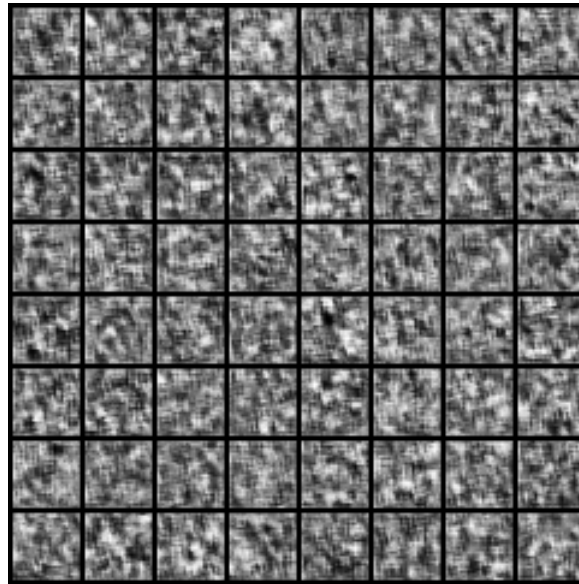


figure 1: epoch = 0



figure 2: epoch = 30



figure 3: epoch = 70