CS 446 / ECE 449 — Homework 5

mukaiyu2

Version 1.0

Instructions.

- Homework is due **Tuesday, April 12, at noon CDT**; no late homework accepted.
- Everyone must submit individually at gradescope under `hw5` and `hw5code`.
- The "written" submission at `hw5` **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use LaTeX, markdown, google docs, MS word, whatever you like; but it must be typed!
- When submitting at `hw5`, 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 `hw5code` 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 `hw5code`, only upload `hw5_vae.py` and `hw5_gan.py`. Additional files will be ignored.

Version History.

1.0 Initial Version.

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

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

We can also write the ELBO as

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

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.

(a)

$$p_\theta(x|z) = \prod_{j=1}^{G} p_\theta(x_j|z) = \prod_{j=1}^{G} \hat{y_j}^{x_j}(1-\hat{y_j})^{1-x_j}$$

(b) The output dimension of the encoder is **2**.

(c) Jensen's inequality: For concave function f:

$$f\left(\sum_z q(z)g(z)\right) \geqslant \sum_z q(z)f(g(z))$$

Since $\big(\log(x)\big)'' = -\frac{1}{x^2} < 0$, $\log(x)$ is concave, then we have:

$$\begin{aligned}
\log p_\theta(x) &= \log \sum_z q_\phi(z|x)\frac{p_\theta(x,z)}{q_\phi(z|x)} \\
&\geqslant \sum_z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)} \\
&= \sum_z q_\phi(z|x)\log\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)} \\
&= \sum_z q_\phi(z|x)\log p_\theta(x|z) - \sum_z q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z)} \\
&= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x), p(z))
\end{aligned}$$

(d) Properties of KL-divergence:

- Lower bounded by 0, equals to 0 iff 2 distributions are the same.
- Asymmetry, meaning that $KL(q,p) \neq KL(p,q)$.

(e) No, they are not the same, in practice we often use Eq. 2. Because the $q_\phi(z|x)$ of $\mathbb{E}_{q_\phi(z|x)}$ is in practice approximated by sampling on a standard Gaussian distribution and shift according to the $\mu(x)$ and $\sigma(x)$ from the encoder.

(f) No, it's not a good idea, otherwise $q_\phi(z|x)$ would be the same as $q_\phi(z)$, or more seriously, $p(z)$. Just as discussed in class, if so, we are not learning anything.
Encoder is necessary because it learns the conditional distribution of z w.r.t. x, the data influences this conditional distribution.
In practice $q_\phi(z|x)$ is approximated by sampling z from $\mathcal{N}(0,I)$ and shift according to $\mu(x)$ and $\sigma(x)$ (here in this question, I).

(g)

$$KL(q_\phi(z|x), q_\phi(z|x)) = 0$$

Doesn't matter what distribution $q_\phi(z|x)$ is, as long as we feed in 2 same distribution to KL-divergence, the result will always be 0 because $\log\frac{p}{p} = 0$.

(h) For 1-dimensional z:

$$\begin{aligned}
KL(q_\phi(z|X), p(z)) &= q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu_\phi)^2\right) \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu_\phi)^2\right)}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu_p)^2\right)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu_\phi)^2\right) \frac{1}{2\sigma^2} \left[(z-\mu_p)^2 - (z-\mu_\phi)^2\right] \\
&= \frac{1}{2\sigma^2\sqrt{2\pi\sigma^2}} \left[(z-\mu_p)^2 - (z-\mu_\phi)^2\right] \exp\left(-\frac{1}{2\sigma^2}(z-\mu_\phi)^2\right)
\end{aligned}$$

(i) Let the Lagrangian multiplier be λ, then:

$$\begin{aligned}
\mathcal{L}(q_\phi(z|x), \lambda) &= \sum_z q_\phi(z|x) \log p_\theta(x|z) - KL(q_\phi(z|x), p(z)) + \lambda \left(\sum_z q_\phi(z|x) - 1\right) \\
&= \sum_z q_\phi(z|x) \log p_\theta(x|z) - \sum_z q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z)} + \lambda \left(\sum_z q_\phi(z|x) - 1\right)
\end{aligned}$$

So taking gradient we have:

$$\frac{\partial \mathcal{L}(q_\phi(z|x), \lambda)}{\partial q_\phi(z|x)} = \log p_\theta(x|z) + \log p(z) - (\log q_\phi(z|x) + 1) + \lambda = 0$$

$$\Downarrow$$

$$\log q_\phi(z|x) = \log p_\theta(x|z) + \log p(z) + \lambda - 1 = \lambda - 1 + \log p_\theta(x, z)$$

$$\Downarrow$$

$$q_\phi(z|x) \propto p_\theta(x, z)$$

Incorporating the constraint $\sum_z q_\phi(z|x) = 1$ by regularization, we have:

$$q_\phi(z|x) = \frac{p_\theta(x, z)}{\sum_{z'} p_\theta(x, z')} = \frac{p_\theta(x, z)}{p_\theta(x)} = p_\theta(z|x)$$

(j) $q_\phi(z|x)$ should be equal to (3) $p_\theta(z|x)$, which is encoding as wished by prior.

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 $\{0, \ldots, 9\}$, 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):

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

Relevant files: *HW5_vae.py*, *HW5_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.

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.

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**

Layer No.	*Layer Type*	*Kernel Size*	*Stride*	*Padding*	*Output Channels*
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**

Layer No.	*Layer Type*	*Kernel Size*	*Stride*	*Padding*	*Bias*	*Output Channels*
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 `hw5_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 `hw5_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 `hw5_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 `hw5_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.