Due Date: April 12, 2020

Question 1 (7-5-5-3)**.** The point of this question is to understand and compare the effects of different regularizers (specifically dropout and weight decay) on the weights of a network. Consider a linear regression problem with input data $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, weights $\boldsymbol{w} \in \mathbb{R}^{d\times 1}$ and targets $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$. Suppose that dropout is applied to the input (with probability $1-p$ of dropping the unit i.e. setting it to 0). Let $\boldsymbol{R} \in \mathbb{R}^{n\times d}$ be the dropout mask such that $\boldsymbol{R}_{ij} \sim \text{Bern}(p)$ is sampled i.i.d. from the Bernoulli distribution.

For a squared error loss function with dropout, we then have:

$$L(\boldsymbol{w}) = ||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2$$

1.1 Let Γ be a diagonal matrix with $\Gamma_{ii} = (\boldsymbol{X}^\top \boldsymbol{X})_{ii}^{1/2}$. Show that the *expectation (over* $\boldsymbol{R}$*)* of the loss function can be rewritten as $\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$. *Hint: Note we are trying to find the expectation over a squared term and use* $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$.
Answer:
We have:

$$\mathbb{E}[L(\boldsymbol{w})] = \mathbb{E}[||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2]$$

develop the expression, we get:

$$\mathbb{E}[L(\boldsymbol{w})] = \mathbb{E}[(\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w})^T(\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}))]$$
$$\mathbb{E}[L(\boldsymbol{w})] = \mathbb{E}[(\boldsymbol{y}^T - (\boldsymbol{X} \odot \boldsymbol{R})^T\boldsymbol{w}^T)(\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}))]$$
$$\mathbb{E}[L(\boldsymbol{w})] = \mathbb{E}[\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w} - \boldsymbol{y}(\boldsymbol{X} \odot \boldsymbol{R})^T\boldsymbol{w}^T + (\boldsymbol{X} \odot \boldsymbol{R})^T\boldsymbol{w}^T(\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}]$$

because the expectation is over $\boldsymbol{R}$:

$$\mathbb{E}[L(\boldsymbol{w})] = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\mathbb{E}[\boldsymbol{X} \odot \boldsymbol{R}]\boldsymbol{w} + \boldsymbol{w}^T\mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})^T(\boldsymbol{X} \odot \boldsymbol{R})]\boldsymbol{w}$$

Compute $\mathbb{E}[\boldsymbol{X} \odot \boldsymbol{R}]$:
Since the expectation is over $\boldsymbol{R}$ and $\boldsymbol{R}_{ij} \sim \text{Bern}(p)$, we get:

$$\mathbb{E}[\boldsymbol{X} \odot \boldsymbol{R}] = p\boldsymbol{X}_{ij}$$

Compute $\mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})^T(\boldsymbol{X} \odot \boldsymbol{R})]$:

— $i \neq j$: $\boldsymbol{R}_{k,i}$ and $\boldsymbol{R}_{k,j}$ are independent. Therefore,$\mathbb{E}[\boldsymbol{R}_{k,i}\boldsymbol{R}_{k,j}] = \mathbb{E}[\boldsymbol{R}_{k,i}]\mathbb{E}[\boldsymbol{R}_{k,j}] = p^2$:

$$\mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})^T(\boldsymbol{X} \odot \boldsymbol{R})]_{i,j} = \sum_{k=1}^{n}(\boldsymbol{X}_{k,i}\mathbb{E}[\boldsymbol{R}_{k,i}]\mathbb{E}[\boldsymbol{R}_{k,j}]\boldsymbol{X}_{k,j}) = p^2(\boldsymbol{X}^T\boldsymbol{X})_{i,j}$$

— $i = j$: $\boldsymbol{R}_{k,i}$ and $\boldsymbol{R}_{k,j}$ are dependent. Therefore,$\mathbb{E}[\boldsymbol{R}_{k,i}\boldsymbol{R}_{k,j}] = p$:

$$\mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})^T(\boldsymbol{X} \odot \boldsymbol{R})]_{i,j} = \sum_{k=1}^{n}(\boldsymbol{X}_{k,i}\mathbb{E}[\boldsymbol{R}_{k,i}]\mathbb{E}[\boldsymbol{R}_{k,j}]\boldsymbol{X}_{k,j}) = p(\boldsymbol{X}^T\boldsymbol{X})_{i,j}$$

The two cases give us:

$$\mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})^T(\boldsymbol{X} \odot \boldsymbol{R})] = p^2(\boldsymbol{X}^T\boldsymbol{X}) + p(1-p)diag(\boldsymbol{X}^T\boldsymbol{X})$$

Return to the first equation:

$$\mathbb{E}[||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2] = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\mathbb{E}[\boldsymbol{X} \odot \boldsymbol{R}]\boldsymbol{w} + \boldsymbol{w}^T\mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})^T(\boldsymbol{X} \odot \boldsymbol{R})]\boldsymbol{w}$$

$$\mathbb{E}[||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2] = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^Tp\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^Tp^2(\boldsymbol{X}^T\boldsymbol{X}) + p(1-p)diag(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{w}$$

$$\mathbb{E}[||\boldsymbol{y} - (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}||^2] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||diag(\boldsymbol{X}^T\boldsymbol{X})^{1/2}\boldsymbol{w}||^2$$

Therefore,

$$\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$$

1.2 Show that the solution $\boldsymbol{w}^{\text{dropout}}$ that minimizes the expected loss from question 1.1 satisfies

$$p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

where λ^{dropout} is a regularization coefficient depending on p. How does the value of p affect the regularization coefficient, λ^{dropout} ?

Answer:

We have:

$$\mathbb{E}[L(\boldsymbol{w})] = ||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2$$

$$\frac{\partial\mathbb{E}[L(\boldsymbol{w})]}{\partial\boldsymbol{w}} = \frac{\partial}{\partial\boldsymbol{w}}(||\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}||^2 + p(1-p)||\Gamma\boldsymbol{w}||^2)$$

$$\frac{\partial\mathbb{E}[L(\boldsymbol{w})]}{\partial\boldsymbol{w}} = 2p\boldsymbol{X}^T(p\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2p(1-p)(\Gamma\boldsymbol{w})\Gamma$$

$$\frac{\partial\mathbb{E}[L(\boldsymbol{w})]}{\partial\boldsymbol{w}} = 2p(p\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}) + 2p(1-p)\Gamma^2\boldsymbol{w}$$

Setting $\frac{\partial\mathbb{E}[L(\boldsymbol{w})]}{\partial\boldsymbol{w}} = 0$

$$2p(p\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}) + 2p(1-p)\Gamma^2\boldsymbol{w} = 0$$

$$p\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y} + (1-p)\Gamma^2\boldsymbol{w} = 0$$

$$p\boldsymbol{w}(\boldsymbol{X}^T\boldsymbol{X} + \frac{(1-p)}{p}\Gamma^2) - \boldsymbol{X}^T\boldsymbol{y} = 0$$

$$p\boldsymbol{w}(\boldsymbol{X}^T\boldsymbol{X} + \frac{(1-p)}{p}\Gamma^2) = \boldsymbol{X}^T\boldsymbol{y}$$

$$p\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \frac{(1-p)}{p}\Gamma^2)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Therefore,

$$p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

where $\lambda^{\text{dropout}} = \frac{(1-p)}{p}$.

How does the value of p affect the regularization coefficient, λ^{dropout} ?

Since $\lambda^{\text{dropout}} = \frac{(1-p)}{p}$, p is a probability so it is between 0 and 1, therefore, for p close to 1, λ^{dropout} is small and by decreasing p, λ^{dropout} grows larger.

1.3 Express the loss function for a linear regression problem without dropout and with L^2 regularization, with regularization coefficient λ^{L_2}. Derive its closed form solution $\boldsymbol{w}^{L_2}$.
Answer:
For the loss function for a linear regression problem with L^2 regularization, , we then have:

$$L_2(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 + \lambda^{L_2}||\boldsymbol{w}||^2$$

The closed form solution:

$$\frac{\partial L_2(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{\partial}{\partial \boldsymbol{w}}(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||^2 + \lambda^{L_2}||\boldsymbol{w}||^2)$$

$$\frac{\partial L_2(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{\partial}{\partial \boldsymbol{w}}((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda^{L_2}(\boldsymbol{w})^T(\boldsymbol{w}))$$

$$\frac{\partial L_2(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{X}^T\boldsymbol{y} + 2\lambda^{L_2}\boldsymbol{w}$$

Setting $\frac{\partial L_2(\boldsymbol{w})]}{\partial \boldsymbol{w}} = 0$

$$\frac{\partial L_2(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{X}^T\boldsymbol{y} + 2\lambda^{L_2}\boldsymbol{w} = 0$$

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda^{L_2}\boldsymbol{I})\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

Therefore,

$$\boldsymbol{w}^{L_2} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda^{L_2}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

1.4 Compare the results of 1.2 and 1.3: identify specific differences in the equations you arrive at, and discuss qualitatively what the equations tell you about the similarities and differences in the effects of weight decay and dropout (1-3 sentences).

Answer:

From the results of 1.2 and 1.3, we see that weight decay penalty grows polynomially in the depth of the network, however, dropout penalty grows exponentially.

Question 2 (5-5-6)**.** Consider a latent variable model $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})dz$, where $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ and $\boldsymbol{z} \in \mathbb{R}^K$. The encoder network (aka "recognition model") of variational autoencoder, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, is used to produce an approximate (variational) posterior distribution over latent variables $\boldsymbol{z}$ for any input datapoint $\boldsymbol{x}$.[1] This distribution is trained to match the true posterior by maximizing the evidence lower bound (ELBO):

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

Let $\mathcal{Q}$ be the family of variational distributions with a feasible set of parameters $\mathcal{P}$; i.e. $\mathcal{Q} = \{q(\boldsymbol{z}; \pi) : \pi \in \mathcal{P}\}$; for example π can be mean and standard deviation of a normal distribution. We assume q_ϕ is parameterized by a neural network (with parameters ϕ) that outputs the parameters, $\pi_\phi(\boldsymbol{x})$, of the distribution $q \in \mathcal{Q}$, i.e. $q_\phi(\boldsymbol{z}|\boldsymbol{x}) := q(\boldsymbol{z}; \pi_\phi(\boldsymbol{x}))$.

2.1 Show that maximizing the expected complete data log likelihood (ECLL)

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

for a fixed $q(\boldsymbol{z}|\boldsymbol{x})$, wrt the model parameter θ, is equivalent to maximizing

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

This means the maximizer of the ECLL coincides with that of the marginal likelihood only if $q(\boldsymbol{z}|\boldsymbol{x})$ perfectly matches $p(\boldsymbol{z}|\boldsymbol{x})$.
Answer:

We have:

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})dz$$

$$\log p_\theta(\boldsymbol{x}) = \log \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})dz$$

$$\max_\theta \mathbb{E}[\log p_\theta(\boldsymbol{x})] = \max_\theta \mathbb{E}[\log \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})dz]$$

However, marginalization of the latent variable is often intractable.
From the lecture, we know that :

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}(\theta, \phi; \boldsymbol{x})$$

So the ELBO is a lower bound to the log marginal likelihood. Therefore, maximizing it with respect to the model parameters θ approximately maximizes the log marginal likelihood.

We know also that:

$$\log p_\theta(\boldsymbol{x}) = \mathcal{L}(\theta, \phi; \boldsymbol{x}) + D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$$

[1] Using a recognition model in this way is known as "amortized inference" ; this can be contrasted with traditional variational inference approaches (see, e.g., Chapter 10 of Bishop's *Pattern Recognition an Machine Learning*), which fit a variational posterior independently for each new datapoint.

So,

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$$

Maximizing ELBO can be shown to minimize $D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$.
Therefore,

$$\max_\theta \mathbb{E}[\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))] = \max_\theta \mathbb{E}[\max_q \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})]]$$

The KL divergence determines the tightness of the lower bound, where we have equality if the KL divergence is zero, which happens if $q(\boldsymbol{z}|\boldsymbol{x})$ perfectly matches $p(\boldsymbol{z}|\boldsymbol{x})$.

2.2 Consider a finite training set $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$, n being the size the training data. Let ϕ^* be the maximizer $\arg\max_\phi \sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ with θ fixed. In addition, for each $\boldsymbol{x}_i$ let $q_i \in \mathcal{Q}$ be an "instance-dependent" variational distribution, and denote by q_i^* the maximizer of the corresponding ELBO. Compare $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ and $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. Which one is bigger ?

Answer:

the variational gap $\mathbb{G}$ is equal to:

$$\mathbb{G} = \log p_\theta(\boldsymbol{x}) - \mathcal{L}(q_i^*)$$

The inference gap decomposes as the sum of approximation and amortization gaps:

$$\mathbb{G} = \log p_\theta(\boldsymbol{x}) - \mathcal{L}(q_{\phi^*}) + \mathcal{L}(q_{\phi^*}) - \mathcal{L}(q_i^*)$$

$\log p_\theta(\boldsymbol{x}) - \mathcal{L}(q_{\phi^*})$ is the approximation gap and $\mathcal{L}(q_{\phi^*}) - \mathcal{L}(q_i^*)$ is the amortization gap.

$$\log p_\theta(\boldsymbol{x}) - \mathcal{L}(q_{\phi^*}) < \log p_\theta(\boldsymbol{x}) - \mathcal{L}(q_i^*)$$
$$\mathcal{L}(q_{\phi^*}) > \mathcal{L}(q_i^*)$$

Knowing that $\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$:

$$\mathcal{L}(q_{\phi^*}) > \mathcal{L}(q_i^*)$$
$$\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) > \log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$$

$$D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) < D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$$

2.3 Following the previous question, compare the two approaches in the second subquestion

(a) in terms of bias of estimating the marginal likelihood via the ELBO, in the best case scenario (i.e. when both approaches are optimal within the respective families)
Answer:
Amortized case is optimal on complex datasets however instance dependent case is optimal at local distribution.

(b) from the computational point of view (efficiency)
Answer:
amortized case is more efficient than instance dependent case.

(c) in terms of memory (storage of parameters)
Answer:
Amortized case needs more memory than the instance dependent case because the global parameters are shared across all datapoints.

Question 3 (6-4)**.** In this question, we study some properties of normalizing flows. Let $X \sim P_X$ and $U \sim P_U$ be, respectively, the distribution of the data and a base distribution (e.g. an isotropic gaussian). We define a normalizing flow as $F : \mathcal{U} \rightarrow \mathcal{X}$ parametrized by $\boldsymbol{\theta}$. Starting with P_U and then applying F will induce a new distribution $P_{F(U)}$ (used to match P_X). Since normalizing flows are invertible, we can also consider the distribution $P_{F^{-1}(X)}$.

However, some flows, like planar flows, are not easily invertible in practice. If we use P_U as the base distribution, we can only sample from the flow but not evaluate the likelihood. Alternatively, if we use P_X as the base distribution, we can evaluate the likelihood, but we will not be able to sample.

3.1 Show that $D_{KL}[P_X||P_{F(U)}] = D_{KL}[P_{F^{-1}(X)}||P_U]$. In other words, the forward KL divergence between the data distribution and its approximation can be expressed as the reverse KL divergence between the base distribution and its approximation.
Answer:
Using the formula for the density of a flow-based model and a change of variables, we have the following:

$$D_{KL}[P_X||P_{F(U)}] = \mathbb{E}_{P_X(X)}[\log P_X(X) - \log P_{F(U)}(X)]$$

$$D_{KL}[P_X||P_{F(U)}] = \mathbb{E}_{P_X(X)}[\log P_X(X) - \log P_U(F^{-1}(X)) - \log |det\frac{\partial F^{-1}(X)}{\partial X}|]$$

$$D_{KL}[P_X||P_{F(U)}] = \mathbb{E}_{P_X(X)}[\log P_X(X) - \log |det\frac{\partial F^{-1}(X)}{\partial X}| - \log P_U(F^{-1}(X))]$$

$$D_{KL}[P_X||P_{F(U)}] = \mathbb{E}_{P_U(U)}[\log P_X(F(U)) + \log |det\frac{\partial F(U)}{\partial U}| - \log P_U(U)]$$

$$D_{KL}[P_X||P_{F(U)}] = \mathbb{E}_{P_U(U)}[\log P_{F^{-1}}(X) - \log P_U(U)]$$

$$D_{KL}[P_X||P_{F(U)}] = D_{KL}[P_{F^{-1}(X)}||P_U]$$

3.2 Suppose two scenario: 1) you don't have samples from $p_X(\boldsymbol{x})$, but you can evaluate $p_X(\boldsymbol{x})$, 2) you have samples from $p_X(\boldsymbol{x})$, but you cannot evaluate $p_X(\boldsymbol{x})$. For each scenario, specify if you would use the forward KL divergence $D_{KL}[P_X||P_{F(U)}]$ or the reverse KL divergence $D_{KL}[P_{F(U)}||P_X]$ as the objective to optimize. Justify your answer.
Answer:

— **We don't have samples from $p_X(\boldsymbol{x})$, but we can evaluate $p_X(\boldsymbol{x})$**: The reverse KL divergence is suitable for this situation because we can minimize $D_{KL}[P_{F(U)}||P_X]$ even if we can only evaluate $p_X(\boldsymbol{x})$ by a multiplicative normalizing constant and stochastic gradient-based methods to obtain an unbiased estimate of the gradient of D_{KL} where we need to be able to sample from the base distribution as well as compute and differentiate through the transformation F and its Jacobian determinant. Therefore, we can fit a flow-based model by minimizing the reverse KL divergence even if we cannot evaluate $p_X(\boldsymbol{x})$.

— **We have samples from $p_X(\boldsymbol{x})$, but we cannot evaluate $p_X(\boldsymbol{x})$**: The forward KL divergence is suitable for this situation because having a set of samples, we can estimate the expectation using Monte Carlo technique and minimizing Monte Carlo approximation is equivalent to fitting the flow-based model to the samples by maximum likelihood estimation, where d, we need to compute F^{-1}. Therefore, we can train a flow model with maximum likelihood even if we cannot evaluate $p_X(\boldsymbol{x})$.

Question 4 (3-7)**.** Let p_0 and p_1 be two probability distributions with densities f_0 and f_1 (respectively). We want to explore what we can do with a trained GAN discriminator. A trained discriminator is thought to be one which is "close" to the optimal one:

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))].$$

4.1 For the first part of this problem, derive an expression we can use to estimate the Jensen-Shannon divergence using a trained discriminator. We remind that the definition of JSD is $\text{JSD}(p_0, p_1) = \frac{1}{2}\big(KL(p_0\|\mu) + KL(p_1\|\mu)\big)$, where $\mu = \frac{1}{2}(p_0 + p_1)$.
Answer:
Find the closed form solution for D^*:

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))]$$

We need to maximize the quantity:

$$\int_{\boldsymbol{x}\sim p_1} p_1(\boldsymbol{x}) \log D(\boldsymbol{x}) d\boldsymbol{x} + \int_{\boldsymbol{x}\sim p_0} p_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x})) d\boldsymbol{x} = \int_{\boldsymbol{x}\sim p_1} p_1(\boldsymbol{x}) \log D(\boldsymbol{x}) + p_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x})) d\boldsymbol{x}$$

So, it can be written in a simpler form:

$$\boldsymbol{y} = a \log \boldsymbol{y} + b \log(1 - \boldsymbol{y})$$

$$\boldsymbol{y}' = \frac{a}{\boldsymbol{y}} + \frac{b}{1 - \boldsymbol{y}}$$

Find optimal $\boldsymbol{y}*$ by setting $\boldsymbol{y}' = 0$

$$\boldsymbol{y}* = \frac{a}{a+b}$$

The discriminator does not need to be defined outside of $Supp(p_1) \cup Supp(p_1)$. Therefore:

$$D^* := \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})}$$

Putting the optimal descriminator into the equation:

$$\mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))] = \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})}] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})})]$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})}] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(\frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})})]$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log \frac{p_1(\boldsymbol{x})}{2 * \frac{1}{2}(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x}))}] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(\frac{p_0(\boldsymbol{x})}{2 * \frac{1}{2}(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x}))})]$$

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log \frac{p_1(\boldsymbol{x})}{\frac{1}{2}(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x}))}] - \log 2 + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(\frac{p_0(\boldsymbol{x})}{\frac{1}{2}(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x}))})] - \log 2 \\
&= \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log \frac{p_1(\boldsymbol{x})}{\frac{1}{2}(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x}))}] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(\frac{p_0(\boldsymbol{x})}{\frac{1}{2}(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x}))})] - 2\log 2 \\
&= 2\mathrm{JSD}(p_0, p_1) - 2\log 2 \\
&= 2\mathrm{JSD}(p_0, p_1) - \log 4
\end{aligned}
$$

Therefore:

$$
\mathrm{JSD}(p_0, p_1) = \frac{1}{2}(\arg\max_D \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))] + \log 4)
$$

4.2 For the second part, we want to demonstrate that a optimal GAN Discriminator (i.e. one which is able to distinguish between examples from p_0 and p_1 with minimal NLL loss) can be used to express the probability density of a datapoint $\boldsymbol{x}$ under f_1, $f_1(\boldsymbol{x})$ in terms of $f_0(\boldsymbol{x})$[2]. Assume f_0 and f_1 have the same support. Show that $f_1(\boldsymbol{x})$ can be estimated by $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1 - D(\boldsymbol{x}))$ by establishing the identity $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1 - D^*(\boldsymbol{x}))$.

Answer:

We have:

$$D = \int_{\boldsymbol{x}} p_1(\boldsymbol{x}) \log D(\boldsymbol{x}) + p_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x})) d\boldsymbol{x}$$

From the previous question, we get :

$$D^* := \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})}$$

Since the optimal discriminator and the trained discriminator are “close”, we can replace D by D^*:

$$= \int_{\boldsymbol{x}} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})} + p_0(\boldsymbol{x}) \log(1 - \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})}) d\boldsymbol{x}$$

$$= \int_{\boldsymbol{x}} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})} + p_0(\boldsymbol{x}) \log \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})} d\boldsymbol{x}$$

$$= \int_{\boldsymbol{x}} p_1(\boldsymbol{x})[\log p_1(\boldsymbol{x}) - \log(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})] + p_0(\boldsymbol{x})[\log p_0(\boldsymbol{x}) - \log(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})] d\boldsymbol{x}$$

$$= \int_{\boldsymbol{x}} p_1(\boldsymbol{x}) \log p_1(\boldsymbol{x}) d\boldsymbol{x} - \int_{\boldsymbol{x}} p_1(\boldsymbol{x}) \log(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})) d\boldsymbol{x} + \int_{\boldsymbol{x}} p_0(\boldsymbol{x}) \log p_0(\boldsymbol{x}) d\boldsymbol{x} - \int_{\boldsymbol{x}} p_0(\boldsymbol{x}) \log(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})) d\boldsymbol{x}$$

The functional derivative of $\int_{\boldsymbol{x}} p_1(\boldsymbol{x}) \log p_1(\boldsymbol{x}) d\boldsymbol{x}$:

$$\frac{\partial}{\partial p_1(\boldsymbol{x})} \int_{\boldsymbol{x}} p_1(\boldsymbol{x}) \log p_1(\boldsymbol{x}) d\boldsymbol{x} = \frac{\partial}{\partial x} \int_{\boldsymbol{x}} ((p_1(\boldsymbol{x}) + \gamma p_1(\boldsymbol{x})) \log(p_1(\boldsymbol{x}) + \gamma p_1(\boldsymbol{x})) - p_1(\boldsymbol{x}) \log p_1(\boldsymbol{x})) d\boldsymbol{x} = \log p_1(\boldsymbol{x}) +$$

On the other hand, we can make the derivative of D^*

$$D^* = \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})}$$

$$\frac{\partial D^*}{\partial p_1(\boldsymbol{x})} \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})}$$

$$\frac{\partial D^*}{\partial \boldsymbol{x}} \frac{\frac{p_1(\boldsymbol{x})}{\partial \boldsymbol{x}}(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x})) - (\frac{p_1(\boldsymbol{x})}{\partial \boldsymbol{x}} + \frac{p_0(\boldsymbol{x})}{\partial \boldsymbol{x}}) p_1(\boldsymbol{x})}{(p_1(\boldsymbol{x}) + p_0(\boldsymbol{x}))^2}$$

Knowing that $\frac{p_1(\boldsymbol{x})}{\partial \boldsymbol{x}} = f_1(\boldsymbol{x})$ and $\frac{p_0(\boldsymbol{x})}{\partial \boldsymbol{x}} = f_0(\boldsymbol{x})$

[2] You might need to use the “functional derivative” to solve this problem. See “19.4.2 Calculus of Variations” of the Deep Learning book or “Appendix D Calculus of Variations” of Bishop’s Pattern Recognition and Machine Learning for more information.

Therefore,

$$f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1 - D^*(\boldsymbol{x}))$$

Question 5 (1-2-1-1-2-3)**.** In this question, we are concerned with analyzing the training dynamics of GANs under gradient ascent-descent. We denote the parameters of the critic and the generator by ψ and θ respectively. The objective function under consideration is the Jensen-Shannon (standard) GAN one:

$$\mathcal{L}(\psi,\theta) = \mathbb{E}_{p_D}\log(\sigma(C_\psi(x))) + \mathbb{E}_{p_\theta}\log(\sigma(-C_\psi(x)))$$

where σ is the logistic function. For ease of exposition, we will study the continuous-time system which results from the (alternating) discrete-time system when learning rate, $\eta > 0$, approaches zero:

$$\begin{aligned}\psi^{(k+1)} &= \psi^{(k)} + \eta v_\psi(\psi^{(k)},\theta^{(k)}) \\ \theta^{(k+1)} &= \theta^{(k)} + \eta v_\theta(\psi^{(k+1)},\theta^{(k)})\end{aligned} \quad \xRightarrow{\eta\to 0^+} \quad \begin{aligned}\dot{\psi} &= v_\psi(\psi,\theta) \\ \dot{\theta} &= v_\theta(\psi,\theta)\end{aligned} \qquad \begin{aligned}v_\psi(\psi,\theta) &:= \nabla_\psi\mathcal{L}(\psi,\theta) \\ v_\theta(\psi,\theta) &:= -\nabla_\theta\mathcal{L}(\psi,\theta)\end{aligned}$$

The purpose is to initiate a study on the stability of the training algorithm. For this reason, we will utilize the following simple setting: Both training and generated data have support on $\mathbb{R}$. In addition, $p_D = \delta_0$ and $p_\theta = \delta_\theta$. This means that both of them are Dirac distributions[3] which are centered at $x = 0$, for the real data, and at $x = \theta$, for the generated. The critic, $C_\psi : \mathbb{R} \to \mathbb{R}$, is $C_\psi(x) = \psi_0 x + \psi_1$.

5.1 Derive the expressions for the "velocity" field, v, of the dynamical system in the joint parameter space (ψ_0,ψ_1,θ), and find its stationary points $(\psi_0^*,\psi_1^*,\theta^*)$.[4]
Answer:

$$v = \begin{pmatrix} v_{\psi_0}(\psi,\theta) \\ v_{\psi_1}(\psi,\theta) \\ v_\theta(\psi,\theta) \end{pmatrix} = \begin{pmatrix} \nabla_{\psi_0}\mathcal{L}(\psi,\theta) \\ \nabla_{\psi_1}\mathcal{L}(\psi,\theta) \\ -\nabla_\theta\mathcal{L}(\psi,\theta) \end{pmatrix}$$

We know that $p_D = \delta_0$ and $p_\theta = \delta_\theta$, Dirac distributions, which are centered at $x = 0$, for the real data, and at $x = \theta$, for the generated.
Also,$(g \circ f)' = (g' \circ f)f'$, $log'(u) = \frac{u'}{u}$ and $\sigma'(x) = \sigma(x)(1-\sigma(x))$

Calculate $v_{\psi_0}(\psi,\theta)$:

$$v_{\psi_0}(\psi,\theta) = \frac{\partial}{\partial\psi_0}(\log(\sigma(C_\psi(x))) + \log(\sigma(-C_\psi(x))))$$

$$v_{\psi_0}(\psi,\theta) = \frac{\partial}{\partial\psi_0}(\log(\sigma(\psi_1)) + \log(\sigma(-\psi_0\theta-\psi_1)))$$

$$v_{\psi_0}(\psi,\theta) = 0 + \frac{\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))\theta^2}{\sigma(-\psi_0\theta-\psi_1)}$$

$$v_{\psi_0}(\psi,\theta) = (1-\sigma(-\psi_0\theta-\psi_1))\theta^2$$

Calculate $v_{\psi_0}(\psi,\theta)$:

$$v_{\psi_1}(\psi,\theta) = \frac{\partial}{\partial\psi_1}(\log(\sigma(C_\psi(x))) + \log(\sigma(-C_\psi(x))))$$

[3] If $p_X = \delta_z$, then $p(X = z) = 1$.

[4] To find the stationary points, set $v = 0$ and solve for each of the parameters.

$$v_{\psi_1}(\psi,\theta) = \frac{\partial}{\partial\psi_1}(\log(\sigma(\psi_1)) + \log(\sigma(-\psi_0\theta - \psi_1)))$$

$$v_{\psi_0}(\psi,\theta) = \frac{\sigma(\psi_1)(1-\sigma(\psi_1))}{\sigma(\psi_1)} + \frac{\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))}{\sigma(-\psi_0\theta-\psi_1)}$$

$$v_{\psi_0}(\psi,\theta) = (1-\sigma(\psi_1)) + (1-\sigma(-\psi_0\theta-\psi_1))$$

Calculate $v_\theta(\psi,\theta)$:

$$v_\theta(\psi,\theta) = \frac{\partial}{\partial\theta}(\log(\sigma(C_\psi(x))) + \log(\sigma(-C_\psi(x))))$$

$$v_\theta(\psi,\theta) = \frac{\partial}{\partial\theta}(\log(\sigma(\psi_1)) + \log(\sigma(-\psi_0\theta - \psi_1)))$$

$$v_\theta(\psi,\theta) = 0 + \frac{\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2}{\sigma(-\psi_0\theta-\psi_1)}$$

$$v_\theta(\psi,\theta) = (1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2$$

Therefor,

$$v = \begin{pmatrix} v_{\psi_0}(\psi,\theta) \\ v_{\psi_1}(\psi,\theta) \\ v_\theta(\psi,\theta) \end{pmatrix} = \begin{pmatrix} \nabla_{\psi_0}\mathcal{L}(\psi,\theta) \\ \nabla_{\psi_1}\mathcal{L}(\psi,\theta) \\ -\nabla_\theta\mathcal{L}(\psi,\theta) \end{pmatrix} = \begin{pmatrix} (1-\sigma(-\psi_0\theta-\psi_1))\theta^2 \\ (1-\sigma(\psi_1)) + (1-\sigma(-\psi_0\theta-\psi_1)) \\ -(1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2 \end{pmatrix}$$

To find the stationary points, we set $v = 0$ and solve for each of the parameters

$$v = \begin{pmatrix} v_{\psi_0}(\psi,\theta) \\ v_{\psi_1}(\psi,\theta) \\ v_\theta(\psi,\theta) \end{pmatrix} = \begin{pmatrix} \nabla_{\psi_0}\mathcal{L}(\psi,\theta) \\ \nabla_{\psi_1}\mathcal{L}(\psi,\theta) \\ -\nabla_\theta\mathcal{L}(\psi,\theta) \end{pmatrix} = \begin{pmatrix} (1-\sigma(-\psi_0\theta-\psi_1))\theta^2 \\ (1-\sigma(\psi_1)) + (1-\sigma(-\psi_0\theta-\psi_1)) \\ -(1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$(1-\sigma(-\psi_0\theta-\psi_1))\theta^2 = 0 \rightarrow \theta = 0$(because we can't set $\sigma = 0$ for the exponential part.
$(1-\sigma(\psi_1)) + (1-\sigma(-\psi_0\theta-\psi_1)) = 0 \rightarrow 2\psi_1 = -\psi_0\theta$
$-(1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2 = 0 \rightarrow \psi_0 = 0$
stationary points $(\psi_0^*, \psi_1^*, \theta^*) = (0,0,0)$

5.2 Derive J^*, the (3×3) Jacobian of v at $(\psi_0^*, \psi_1^*, \theta^*)$.

Answer:

$$J = \begin{pmatrix} \nabla^2_{\psi_0}\mathcal{L}(\psi,\theta) & \nabla_{\psi_1}\nabla_{\psi_0}\mathcal{L}(\psi,\theta) & \nabla_{\theta}\nabla_{\psi_0}\mathcal{L}(\psi,\theta) \\ \nabla_{\psi_0}\nabla_{\psi_1}\mathcal{L}(\psi,\theta) & \nabla^2_{\psi_1}\mathcal{L}(\psi,\theta) & \nabla_{\theta}\nabla_{\psi_1}\mathcal{L}(\psi,\theta) \\ \nabla_{\psi_0}\nabla_{\theta}\mathcal{L}(\psi,\theta) & \nabla_{\psi_1}\nabla_{\theta}\mathcal{L}(\psi,\theta) & \nabla^2_{\theta}\mathcal{L}(\psi,\theta) \end{pmatrix}$$

Calculate $\nabla^2_{\psi_0}\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\psi_0}(1-\sigma(-\psi_0\theta-\psi_1))\theta^2 = \theta\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))\psi_0$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla^2_{\psi_0}\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla_{\psi_1}\nabla_{\psi_0}\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\psi_1}(1-\sigma(-\psi_0\theta-\psi_1))\theta^2 = \theta\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla_{\psi_1}\nabla_{\psi_0}\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla_{\theta}\nabla_{\psi_0}\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\theta}(1-\sigma(-\psi_0\theta-\psi_1))\theta^2 = 2\theta + \theta^2\psi_0\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1)) - 2\theta\sigma(-\psi_0\theta-\psi_1)$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla_{\theta}\nabla_{\psi_0}\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla_{\psi_0}\nabla_{\psi_1}\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\psi_0}((1-\sigma(\psi_1)) + (1-\sigma(-\psi_0\theta-\psi_1))) = \theta\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla_{\theta}\nabla_{\psi_0}\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla^2_{\psi_1}\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\psi_1}((1-\sigma(\psi_1)) + (1-\sigma(-\psi_0\theta-\psi_1))) = \sigma(\psi_1)(1-\sigma(\psi_1)) + \sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla^2_{\psi_1}\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla_{\theta}\nabla_{\psi_1}\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\theta}((1-\sigma(\psi_1)) + (1-\sigma(-\psi_0\theta-\psi_1))) = 0 + \sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))\psi_0$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla_{\theta}\nabla_{\psi_1}\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla_{\psi_0}\nabla_\theta\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\psi_0}(1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2 = 2\psi_0(1-\sigma(-\psi_0\theta-\psi_1)) - \psi_0^2\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla_{\psi_1}\nabla_{\psi_0}\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla_{\psi_1}\nabla_\theta\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\psi_1}(1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2 = \psi_0^2\theta\sigma(-\psi_0\theta-\psi_1)(1-\sigma(-\psi_0\theta-\psi_1))$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla_{\psi_1}\nabla_\theta\mathcal{L}(\psi,\theta) = 0$

Calculate $\nabla_\theta^2\mathcal{L}(\psi,\theta)$:

$$= \frac{\partial}{\partial\theta}(1-\sigma(-\psi_0\theta-\psi_1))\psi_0^2 = \psi_0^2\theta\sigma(-\psi_0\theta-\psi_1)$$

at $(\psi_0^*, \psi_1^*, \theta^*)$, $\nabla_\theta^2\mathcal{L}(\psi,\theta) = 0$

Therefore,

$$J^* = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

For a continuous-time system to be locally asymptotically stable it suffices that all eigenvalues of J^* have negative real part. Otherwise, further study is needed to conclude. However, this case is not great news since the fastest achievable convergence is sublinear.

5.3 Find the eigenvalues of J^* and comment on the system's local stability around the stationary points.

Answer:

$$J^* = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\det |J^* - \lambda I| = 0$$

$$\det \begin{pmatrix} -\lambda & 0 & 0 \\ 0 & -\lambda & 0 \\ 0 & 0 & -\lambda \end{pmatrix} = 0$$

$$-\lambda(\lambda^2 0) + 0 + 0 = 0 \rightarrow \lambda = 0$$

The zero matrix has only zero as its eigenvalues [source], therefore, we can not say anything about the local stability however we need more studies.

Now we will include a gradient penalty, $\mathcal{R}_1(\psi) = \mathbb{E}_{p_D} \|\nabla_x C_\psi(x)\|^2$, to the critic's loss, so the regularized system becomes:

$$\begin{aligned} \dot{\psi} &= \bar{v}_\psi(\psi, \theta) & \bar{v}_\psi(\psi, \theta) &:= \nabla_\psi \mathcal{L}(\psi, \theta) - \tfrac{\gamma}{2} \nabla_\psi \mathcal{R}_1(\psi) \\ \dot{\theta} &= \bar{v}_\theta(\psi, \theta) & \bar{v}_\theta(\psi, \theta) &:= -\nabla_\theta \mathcal{L}(\psi, \theta) \end{aligned}$$

for $\gamma > 0$. Repeat 1-2-3 for the modified system and compare the stability of the two.

5.4 Derive the expressions for the "velocity" field, $\bar{v}$, of the dynamical system in the joint parameter space (ψ_0, ψ_1, θ), and find its stationary points $(\psi_0^*, \psi_1^*, \theta^*)$. [5]
Answer:

$$\bar{v} = \begin{pmatrix} \nabla_{\psi_0} \mathcal{L}(\psi, \theta) - \frac{\gamma}{2} \nabla_{\psi 0} \mathcal{R}_1(\psi) \\ \nabla_{\psi_1} \mathcal{L}(\psi, \theta) - \frac{\gamma}{2} \nabla_{\psi 1} \mathcal{R}_1(\psi) \\ -\nabla_\theta \mathcal{L}(\psi, \theta) \end{pmatrix}$$

$\mathcal{R}_1(\psi) = \mathbb{E}_{p_D} \|\nabla_x C_\psi(x)\|^2 = \mathbb{E}_{p_D} \|\psi_0\|^2$
$\nabla_{\psi 0} \mathcal{R}_1(\psi) = 2\psi_0$
$\nabla_{\psi 1} \mathcal{R}_1(\psi) = 0$
Adding the results to the first part of the question, we get:

$$\bar{v} = \begin{pmatrix} (1 - \sigma(-\psi_0\theta - \psi_1))\theta^2 - \gamma\psi_0 \\ (1 - \sigma(\psi_1)) + (1 - \sigma(-\psi_0\theta - \psi_1)) \\ -(1 - \sigma(-\psi_0\theta - \psi_1))\psi_0^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

[5] To find the stationary points, set $v = 0$ and solve for each of the parameters.

5.5 Derive $\bar{J}^*$, the (3×3) Jacobian of $\bar{v}$ at $(\psi_0^*, \psi_1^*, \theta^*)$.
Answer:

5.6 Find the eigenvalues of $\bar{J}^*$ and comment on the system's local stability around the stationary points.