Due Date: April 12, 2020

Instructions

- *For all questions, show your work!*
- *Please use a document preparation system such as LaTeX.*
- *Unless noted that questions are related, assume that notation and definitions for each question are self-contained and independent.*
- *Submit your answers electronically via Gradescope.*
- *TAs for this assignment are* ***Christos Tsirigotis*** *and* ***Philippe Brouillard****.*

This assignment covers mathematical and algorithmic techniques underlying regularization and popular families of deep generative models. Thus, we explore regularization (Question 1), variational autoencoders (VAEs, Questions 2), normalizing flows (Question 3), and generative adversarial networks (GANs, Question 4-5).

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. To ease notation, we write $\boldsymbol{M} = \boldsymbol{X} \odot \boldsymbol{R}$. We begin by developing the squared loss:

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{w}\|^2 &= (\boldsymbol{y} - \boldsymbol{M}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{M}\boldsymbol{w}) \\
&= (\boldsymbol{y}^\top - \boldsymbol{w}^\top \boldsymbol{M}^\top)(\boldsymbol{y} - \boldsymbol{M}\boldsymbol{w}) \\
&= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{w}^\top \boldsymbol{M}^\top \boldsymbol{y} + \boldsymbol{w}^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{w}
\end{aligned}$$

Taking the expected value over $\boldsymbol{R}$ of the above,

$$\begin{aligned}
\mathbb{E}[L(\boldsymbol{w})] &= \mathbb{E}[\boldsymbol{y}^\top \boldsymbol{y}] - \mathbb{E}[2\boldsymbol{w}^\top \boldsymbol{M}^\top \boldsymbol{y}] + \mathbb{E}[\boldsymbol{w}^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{w}] \\
&= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{w}^\top \mathbb{E}[\boldsymbol{M}^\top]\boldsymbol{y} + \boldsymbol{w}^\top \mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}]\boldsymbol{w}
\end{aligned}$$

For some matrix $\boldsymbol{A}$, we have that $(\mathbb{E}[\boldsymbol{A}])_{ij} = \mathbb{E}[\boldsymbol{A}_{ij}]$. For the second term, this implies

$$\begin{aligned}
\mathbb{E}[\boldsymbol{M}^\top]_{ij} &= \mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})^\top]_{ij} \\
&= \mathbb{E}[(\boldsymbol{X} \odot \boldsymbol{R})_{ji}] \\
&= \mathbb{E}[X_{ji} R_{ji}] \\
&= pX_{ji}
\end{aligned}$$

since $R_{ij} \sim \text{Bern}(p)$. We develop the product inside the third term:

$$\begin{aligned}(\boldsymbol{M}^\top \boldsymbol{M})_{ij} &= \sum_{k=1}^{n} M_{ik} M_{kj} \\ &= \sum_{k=1}^{n} X_{ik} R_{ik} X_{kj} R_{kj}\end{aligned}$$

The expected value of this product is then

$$\begin{aligned}\mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}]_{ij} &= \mathbb{E}[(\boldsymbol{M}^\top \boldsymbol{M})_{ij}] \\ &= \sum_{k=1}^{n} X_{ik} X_{kj} \mathbb{E}[R_{ik} R_{kj}]\end{aligned}$$

For the off-diagonal elements, the random variable elements of $\boldsymbol{R}$ are independent:

$$\mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}]_{ij} = p^2 (\boldsymbol{X}^\top \boldsymbol{X})_{ij}, \qquad i \neq j$$

On the diagonal however, we have the same realizations of random variables, implying

$$\mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}]_{ii} = p (\boldsymbol{X}^\top \boldsymbol{X})_{ii}$$

We also make use of the following relation:

$$\begin{aligned}\|\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}\|^2 &= (\boldsymbol{y}^\top - p\boldsymbol{w}^\top \boldsymbol{X}^\top)(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) \\ &= \boldsymbol{y}^\top \boldsymbol{y} - 2p\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{y} + p^2 \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w}\end{aligned}$$

Returning to our expression for the expected value of the loss,

$$\begin{aligned}\mathbb{E}[L(\boldsymbol{w})] &= \boldsymbol{y}^\top \boldsymbol{y} - 2p\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{w}\mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}]\boldsymbol{w} \\ &= \|\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}\|^2 - p^2 \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^\top \mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}]\boldsymbol{w} \\ &= \|\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}\|^2 + \boldsymbol{w}^\top (\mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}] - p^2 \boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{w}\end{aligned}$$

In this last term, every off-diagonal element would be 0. We can then write

$$\mathbb{E}[\boldsymbol{M}^\top \boldsymbol{M}] - p^2 \boldsymbol{X}^\top \boldsymbol{X} = p(1-p)\text{diag}(\boldsymbol{X}^\top \boldsymbol{X}) = p(1-p)\Gamma^\top \Gamma$$

leading to

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

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 are looking for

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

which correspond to the $\boldsymbol{w}$ for which the gradient of the expected loss is zero:

$$\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}[L(\boldsymbol{w})] = \frac{\partial}{\partial \boldsymbol{w}}\|\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}\|^2 + p(1-p)\frac{\partial}{\partial \boldsymbol{w}}\|\Gamma\boldsymbol{w}\|^2 = 0$$

We develop the first term:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{w}}\|\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}\|^2 &= \frac{\partial}{\partial \boldsymbol{w}}\left(\boldsymbol{y}^\top\boldsymbol{y} - 2p\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y} + p^2\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}\right)\\ &= -2p\frac{\partial}{\partial \boldsymbol{w}}\left(\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y}\right) + p^2\frac{\partial}{\partial \boldsymbol{w}}\left(\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}\right)\\ &= -2p\boldsymbol{X}^\top\boldsymbol{y} + p^2(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{w}\\ &= -2p\boldsymbol{X}^\top\boldsymbol{y} + 2p^2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}\end{aligned}$$

and the second:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{w}}\|\Gamma\boldsymbol{w}\|^2 &= \frac{\partial}{\partial \boldsymbol{w}}\left(\boldsymbol{w}^\top\Gamma^\top\Gamma\boldsymbol{w}\right)\\ &= \left(\Gamma^\top\Gamma + \Gamma^\top\Gamma\right)\boldsymbol{w}\\ &= 2\Gamma^\top\Gamma\boldsymbol{w}\end{aligned}$$

The derivative of the expected loss is then

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}[L(\boldsymbol{w})] &= -2p\boldsymbol{X}^\top\boldsymbol{y} + 2p^2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + 2p(1-p)\Gamma^2\boldsymbol{w} = 0\\ \boldsymbol{X}^\top\boldsymbol{y} &= p\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + (1-p)\Gamma^2\boldsymbol{w}\\ \boldsymbol{X}^\top\boldsymbol{y} &= \left(\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + \frac{1-p}{p}\Gamma^2\right)p\boldsymbol{w}\\ \Rightarrow p\boldsymbol{w}^{\text{dropout}} &= \left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{1-p}{p}\Gamma^2\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \qquad (1)\end{aligned}$$

We have then identified that $\lambda^{\text{dropout}} = (1-p)/p$. The probability of dropping an unit is $1-p$. If this is zero, i.e. we keep every unit in the input, $\lambda^{\text{dropout}} = 0$ and our network has no regularization. Adversely, as $1-p$ draws nearer to 1, more and more units in the input can be dropped and $\lambda^{\text{dropout}} \to \infty$. This corresponds to an ever increasing regularization effect, driving the weights that minimize the expected loss to zero.

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. The loss function for such a problem would be

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

Similarly to before, the solution is found where the gradient is zero:

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

For the first term, we have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{w}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 &= \frac{\partial}{\partial \boldsymbol{w}}\left[(\boldsymbol{y}^\top - \boldsymbol{w}^\top\boldsymbol{X})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\right] \\ &= \frac{\partial}{\partial \boldsymbol{w}}\left[\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{w}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}\right] \\ &= -2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w}\end{aligned}$$

and for the second,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{w}}\|\boldsymbol{w}\|^2 &= \frac{\partial}{\partial \boldsymbol{w}}\left[\boldsymbol{w}^\top\boldsymbol{w}\right] \\ &= 2\boldsymbol{I}\boldsymbol{w}\end{aligned}$$

The gradient of the loss is then

$$\begin{aligned}\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} &= -2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{w} + \frac{\lambda^{L_2}}{2}\cdot 2\boldsymbol{I}\boldsymbol{w} = 0 \\ \boldsymbol{X}^\top\boldsymbol{y} &= \left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\lambda^{L_2}}{2}\boldsymbol{I}\right)\boldsymbol{w} \\ \Rightarrow \boldsymbol{w}^{L_2} &= \left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\lambda^{L_2}}{2}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \qquad (2)\end{aligned}$$

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. Comparing solutions (1) and (2), we can first note that the weights $\boldsymbol{w}^{\text{dropout}}$ are all scaled by an additional factor of $1/p$, which is greater than 1. The regularization term in (1) depends on the inputs through Γ, in addition to the regularization parameter λ^{dropout}. This could mean that inputs $\boldsymbol{X}$ of greater amplitude would lead to stronger regularization in dropout. On the other hand, the $\boldsymbol{w}^{L_2}$ solution has no additional scaling, and its regularization term depends entirely on the amplitude of λ^{L_2}. Its tendency to push weight towards 0 does not depend on the magnitude of the inputs.

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 $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ 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

[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.

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. The expression above is another form seen in class for the ELBO, which we note $\mathcal{L}'$:

$$\mathcal{L}'(\theta,\phi;\boldsymbol{x}) = \log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})) \tag{3}$$

This can be set equal to the other expression for the ELBO introduced earlier. Maximizing $\mathcal{L}'$ is then equivalent to maximizing $\mathcal{L}$, which we develop:

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

The maximization procedure at hand is w.r.t. θ and for a fixed $q(\boldsymbol{z} \mid \boldsymbol{x})$, so only the first term above would be maximized. This term represents the ECLL. By the equivalence between the two forms of the ELBO, we have shown that maximizing the ECLL w.r.t. θ is equivalent to maximizing $\mathcal{L}'(\theta,\phi;\boldsymbol{x}) = \log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$.

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. Using expression (3) as the objective ELBO, both maximizing distributions correspond to:

$$\begin{aligned} q_{\phi^*} &= \arg\max_{q_\phi} \sum_{i=1}^{n} \mathcal{L}(\theta, \phi; \boldsymbol{x}_i) \\ &= \arg\min_{q_\phi} \sum_{i=1}^{n} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i) \| p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \\ &= \arg\min_{q_\phi} \sum_{i=1}^{n} \mathbb{E}_{q_\phi} \left[\log q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i) - \log p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)\right] \end{aligned}$$

$$\begin{aligned} q_i^* &= \arg\max_{q_i} \mathcal{L}(\theta, \phi; \boldsymbol{x}_i) \\ &= \arg\min_{q_i} D_{\mathrm{KL}}(q_i(\boldsymbol{z}) \| p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)) \\ &= \arg\min_{q_i} \mathbb{E}_{q_i} \left[\log q_i(\boldsymbol{z}) - \log p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)\right] \end{aligned}$$

It can be seen by comparing the two forms above that q_{ϕ^*} must minimize the KL divergence for the entire training set simultaneously. This distribution then learns to attribute weight to a greater range of values of $\boldsymbol{z}$ in the latent space than those values of $\boldsymbol{z}$ for which the true posterior $p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)$ of an individual training example $\boldsymbol{x}_i$ would bear the most importance. Comparatively, according to the universal approximation theorem, an adequate optimization procedure would push an instance-dependant variational distribution $q_i^*(\boldsymbol{z})$ to be as close as possible to the true posterior $p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$, fine-tuning it to $\boldsymbol{x}_i$ without the hindrance of taking into account other members of the training set. Looking at the KL divergences of both distributions q_{ϕ^*} and q_i^* evaluated at data point $\boldsymbol{x}_i$, it is evident that $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}) \| p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i))$ would be the smallest of the two, as $q_i^*(\boldsymbol{z})$ was trained to reproduce $p_\theta(\boldsymbol{z} \mid \boldsymbol{x}_i)$ exactly, while $q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x}_i)$ was not. The bigger one would then be $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$.

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. With optimal distributions, the marginal likelihood can be estimated via importance sampling of the ELBO. The approach where we have an optimal approximation q_i^* of the posterior for every example in the training set would have very little bias, since we have approximate distributions that are very close of the true density. The other approach where we maximized q_ϕ over the complete training set would have a greater bias, as its family of functions is not as expressive as individually tuned approximate posteriors. However, the first approach would have high variance, being very sensible to variations in the data, while the second would be more robust to model such variations and therefore have lower variance.

(b) from the computational point of view (efficiency)

Answer. Computationally, it would be more efficient to optimize individual posteriors

q_i as the loss landscape would be much less multimodal, with the global minima being easier to find for individual training examples. Adequately minimizing the average loss of the entire set is a much more involved optimization task, with the loss landscape bearing many local minima.

(c) in terms of memory (storage of parameters)

Answer. For models of the same size, learning optimal posteriors for every training example requires n times more storage than the alternative, which learns only one optimal distribution for a given dataset of size n.

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. The Kullback-Leibler divergence between two probability distributions P_X and $P_{F(U)}$ is expressed as:

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

We make use of the relation for a change of variable:

$$\begin{aligned} P_{F(U)}(F(\boldsymbol{u})) &= P_U(F^{-1}(\boldsymbol{x}))|\det J_{F^{-1}}(\boldsymbol{x})| \\ &= P_U(\boldsymbol{u})|\det J_F(\boldsymbol{u})|^{-1} \end{aligned}$$

where $J_{F^{-1}}(X)$ is the Jacobian of the transformation F^{-1} w.r.t. X. We develop the forward KL divergence:

$$\begin{aligned} D_{\mathrm{KL}}[P_X\|P_{F(U)}] &= \mathbb{E}_{P_X}\left[\log P_X(\boldsymbol{x}) - \log P_{F(U)}(F(\boldsymbol{u}))\right] \\ &= \mathbb{E}_{P_X}\left[\log P_X(\boldsymbol{x}) - \log|\det J_{F^{-1}}(\boldsymbol{x})| - \log P_U(F^{-1}(\boldsymbol{x}))\right] \\ &= \mathbb{E}_{P_{F^{-1}(X)}}\left[\log P_X(F(\boldsymbol{u})) + \log|\det J_F(\boldsymbol{u})| - \log P_U(\boldsymbol{u})\right] \end{aligned}$$

We make use of another change of variable:

$$P_{F^{-1}(X)}(F^{-1}(\boldsymbol{x})) = P_X(F(\boldsymbol{u}))|\det J_F(\boldsymbol{u})|$$

Which leads to

$$\begin{aligned} D_{\mathrm{KL}}[P_X\|P_{F(U)}] &= \mathbb{E}_{P_{F^{-1}(X)}}\left[P_{F^{-1}(X)}(F^{-1}(\boldsymbol{x})) - \log P_U(\boldsymbol{u})\right] \\ &= D_{\mathrm{KL}}[P_{F^{-1}(X)}||P_U] \end{aligned}$$

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.

(1) The objective to optimize if we have no samples but can evaluate the distribution would be the reverse KL divergence. We wish to optimize $\boldsymbol{\theta}$, which parameterizes the normalizing flow F. We inspect the reverse KL:

$$\begin{aligned} D_{\text{KL}}[P_{F^{-1}(X)}||P_U] &= \mathbb{E}_{P_{F^{-1}(X)}}[\log P_{F^{-1}(X)}(\boldsymbol{u};\boldsymbol{\theta}) - \log P_U(\boldsymbol{u})] \\ &= \mathbb{E}_{P_{F^{-1}(X)}}[\log P_X(F(\boldsymbol{u};\boldsymbol{\theta})) + \log|\det J_F(\boldsymbol{u};\boldsymbol{\theta})| - \log P_U(\boldsymbol{u})] \end{aligned}$$

We can then apply the transform on samples from P_U to evaluate P_X, and compute the Jacobian of F w.r.t. the samples and conduct the optimization procedure.

(2) In this case, the objective to optimize should be the forward KL divergence:

$$\begin{aligned} D_{\text{KL}}[P_X||P_{F(U)}] &= \mathbb{E}_{P_X}[\log P_X(\boldsymbol{x}) - \log P_{F(U)}(\boldsymbol{x};\boldsymbol{\theta})] \\ &= -\mathbb{E}_{P_X}[P_{F(U)}(\boldsymbol{x};\boldsymbol{\theta})] + \underbrace{\mathbb{E}_{P(X)}[\log P_X(\boldsymbol{x})]}_{C} \end{aligned}$$

where C is a constant in the context of optimization over $\boldsymbol{\theta}$. We perform a change of variable:

$$D_{\text{KL}}[P_X||P_{F(U)}] = -\mathbb{E}_{P_X}[\log P_U(F^{-1}(\boldsymbol{x};\boldsymbol{\theta})) + \log|\det J_{F^{-1}}(\boldsymbol{x};\boldsymbol{\theta})|] + C$$

With samples from $P_X(\boldsymbol{x})$, we can apply the inverse transform on them to evaluate P_U, as well as compute its Jacobian w.r.t. the samples and conduct the optimization procedure.

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. We begin by finding the closed form solution for D^*. We write

$$\begin{aligned} G(D) &= \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))] \\ &= \int_{\boldsymbol{x}} f_1(\boldsymbol{x}) \log D(\boldsymbol{x}) d\boldsymbol{x} + \int_{\boldsymbol{x}} f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x})) d\boldsymbol{x} \\ &= \int_x f_1(\boldsymbol{x}) \log D(\boldsymbol{x}) + f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x})) d\boldsymbol{x} \end{aligned}$$

This is maximal where its functional derivative w.r.t. $D(\boldsymbol{x})$ is null. Writing the integrand as a functional $g(y, \boldsymbol{x}) = f_1(\boldsymbol{x}) \log D(\boldsymbol{x}) + f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))$, we have:

$$\begin{aligned}
\frac{\partial G(D)}{\partial D(\boldsymbol{x})} &= \frac{\partial}{\partial D(\boldsymbol{x})} \int_x g(D(\boldsymbol{x}), \boldsymbol{x}) d\boldsymbol{x} \\
&= \frac{\partial}{\partial y} g(D(\boldsymbol{x}), x) = 0
\end{aligned}$$

The integrand can be rewritten in the simple algebraic form:

$$g(y) = a \log(y) + b \log(1 - y)$$

with $a = p_1(\boldsymbol{x})$, $b = p_0(\boldsymbol{x})$ and $y = D(\boldsymbol{x})$. We are looking for the y that maximizes $g(y)$:

$$\begin{aligned}
g'(y) &= \frac{a}{y} - \frac{b}{1-y} = 0 \\
\frac{a}{y} &= \frac{b}{1-y} \\
y^* &= \frac{a}{a+b}
\end{aligned}$$

We then find that this corresponds to an optimal discriminator of $D^* = \frac{f_1}{f_1 + f_0}$. This can be used to define an estimate for the Jensen-Shannon divergence of the distributions p_0 and p_1:

$$\begin{aligned}
\text{JSD}(p_0, p_1) &= \frac{1}{2} \text{KL}\left(p_1 \| (p_0 + p_1)/2\right) + \frac{1}{2} \text{KL}\left(p_0 \| (p_0 + p_1)/2\right) && (4) \\
&= \frac{1}{2} \mathbb{E}_{x \sim p_1} \left[\log \frac{2p_1}{p_0 + p_1}\right] + \frac{1}{2} \mathbb{E}_{x \sim p_0} \left[\log \frac{2p_0}{p_0 + p_1}\right] && (5) \\
&= \frac{1}{2} \mathbb{E}_{x \sim p_1} \left[\log \frac{p_1}{p_0 + p_1}\right] + \frac{1}{2} \mathbb{E}_{x \sim p_0} \left[\log \left(1 - \frac{p_1}{p_0 + p_1}\right)\right] + \frac{1}{2} \log 4 && (6) \\
& && (7)
\end{aligned}$$

Drawing N samples from the distributions p_0 and p_1, we can approximate these to correspond to their respective probability densities:

$$\begin{aligned}
\text{JSD}(p_0, p_1) &\approx \log 2 + \frac{1}{2N} \sum_{i=1}^{N} \log \frac{f_1(x_i)}{f_0(x_i) + f_1(x_i)} + \log \left(1 - \frac{f_1(x_i)}{f_0(x_i) + f_1(x_i)}\right) \\
&= \log 2 + \frac{1}{2N} \sum_{i=1}^{N} \log D^* + \log\left(1 - D^*\right)
\end{aligned}$$

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

[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.

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. Using the result from 4.1 for the optimal discriminator, it can be trivially shown that

$$\begin{aligned} f_1(\boldsymbol{x}) &= f_0(\boldsymbol{x})\frac{D^*(\boldsymbol{x})}{1-D^*(\boldsymbol{x})} \\ &= f_0(\boldsymbol{x}) \cdot \frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x})+f_0(\boldsymbol{x})} \cdot \frac{f_1(\boldsymbol{x})+f_0(\boldsymbol{x})}{f_0(\boldsymbol{x})} \\ &= f_1(\boldsymbol{x}) \end{aligned}$$

An optimal GAN discriminator is one that is reasonably "close" to D^*, so it can used to estimate $f_1(\boldsymbol{x})$ through:

$$f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})\frac{D(\boldsymbol{x})}{1-D(\boldsymbol{x})}$$

Hint: Find the closed form solution for D^*.

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:

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

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. Bearing in mind the expression for the logistic function $\sigma(x) = (1+e^{-x})^{-1}$, we develop

[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.

the objective function:

$$\begin{aligned}
\mathcal{L}(\psi,\theta) &= \mathbb{E}_{p_D}\log(\sigma(\psi_0 x+\psi_1)) + \mathbb{E}_{p_0}\log(\sigma(-\psi_0 x-\psi_1)) \\
&= \int_x p_D(x)\log(\sigma(\psi_0 x+\psi_1))dx + \int_x p_\theta(x)\log(\sigma(-\psi_0 x-\psi_1))dx \\
&= \int_x \delta_0(x)\log(\sigma(\psi_0 x+\psi_1)) + \delta_\theta\log(\sigma(-\psi_0 x-\psi_1))dx \\
&= \log(\sigma(\psi_1)) + \log(\sigma(-\psi_0\theta-\psi_1)) \\
&= -\log\big(1+e^{-\psi_1}\big) - \log\big(1+e^{\psi_0\theta+\psi_1}\big)
\end{aligned}$$

We can then obtain the velocity field. For the parameters of the critic, the components are

$$\begin{aligned}
v_\psi &= \nabla_\psi\mathcal{L}(\psi,\theta) = [\partial_{\psi_0}\mathcal{L}(\psi_0,\psi_1,\theta),\ \partial_{\psi_1}\mathcal{L}(\psi_0,\psi_1,\theta)] \\
\Rightarrow v_{\psi_0} &= -\frac{\partial}{\partial\psi_0}\log\big(1+e^{\psi_0 x+\psi_1}\big) \\
&= -\frac{\theta e^{\psi_0\theta+\psi_1}}{1+e^{\psi_0\theta+\psi_1}} \\
&= -\theta\cdot\sigma(\psi_0\theta+\psi_1) \\
\Rightarrow v_{\psi_1} &= -\frac{\partial}{\partial\psi_1}\log\big(1+e^{-\psi_1}\big) - \frac{\partial}{\partial\psi_1}\log\big(1+e^{\psi_0\theta+\psi_1}\big) \\
&= \frac{e^{-\psi_1}}{1+e^{-\psi_1}} - \frac{e^{\psi_0\theta+\psi_1}}{1+e^{\psi_0\theta+\psi_1}} \\
&= 1-\sigma(\psi_1)-\sigma(\psi_0\theta+\psi_1)
\end{aligned}$$

And for those of the generator,

$$v_\theta = -\partial_\theta\mathcal{L}(\psi_0,\psi_1,\theta) \tag{8}$$

$$= \frac{\partial}{\partial\theta}\log\big(1+e^{\psi_0\theta+\psi_1}\big) \tag{9}$$

$$= \frac{\psi_0 e^{\psi_0\theta+\psi_1}}{1+e^{\psi_0\theta+\psi_1}} \tag{10}$$

$$= \psi_0\cdot\sigma(\psi_0\theta+\psi_1) \tag{11}$$

We find the stationary points of this field by setting each individual component to 0:

$$\begin{aligned}
v_{\psi_0} &= -\theta\cdot\sigma(\psi_0\theta+\psi_1) = 0 \\
\Rightarrow \theta^* &= 0 \\
v_\theta &= \psi_0\cdot\sigma(\psi_0\theta+\psi_1) = 0 \\
\Rightarrow \psi_0^* &= 0 \\
v_{\psi_1} &= 1-\sigma(\psi_1)-\sigma(\psi_0\theta+\psi_1) = 0 \\
\sigma(\psi_1^*) &= 1-\sigma(\psi_0^*\theta^*+\psi_1^*) \\
\sigma(\psi_1^*) &= 1-\sigma(\psi_1^*) \\
\sigma(\psi_1^*) &= \sigma(-\psi_1^*) \\
\Rightarrow \psi_1^* &= 0
\end{aligned}$$

The stationary points of the velocity field are then $(\psi_0^*, \psi_1^*, \theta) = (0, 0, 0)$.

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

Answer. The Jacobian J of v at any point is obtained through:

$$J = \begin{bmatrix} \partial_{\psi_0} v_{\psi_0} & \partial_{\psi_1} v_{\psi_0} & \partial_\theta v_{\psi_0} \\ \partial_{\psi_0} v_{\psi_1} & \partial_{\psi_1} v_{\psi_1} & \partial_\theta v_{\psi_1} \\ \partial_{\psi_0} v_\theta & \partial_{\psi_1} v_\theta & \partial_\theta v_\theta \end{bmatrix}$$

We will make use of the derivative of the sigmoid function: $\partial_x \sigma(x) = \sigma(x)(1 - \sigma(x)) = \sigma(x)\sigma(-x)$. We write $y = \psi_0\theta + \psi_1$, and compute the derivatives:

$$\begin{aligned} \partial_{\psi_0} v_{\psi_0} &= -\theta \frac{\partial \sigma(y)}{\partial \psi_0} \\ &= -\theta \frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \psi_0} \\ &= -\theta^2 \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_{\psi_1} v_{\psi_0} &= -\theta \frac{\partial \sigma(y)}{\partial \psi_1} \\ &= -\theta \frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \psi_1} \\ &= -\theta \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_\theta v_\theta &= -\sigma(y) - \theta \frac{\partial \sigma(y)}{\partial \theta} \\ &= -\sigma(y) - \theta \frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \theta} \\ &= -\sigma(y) - \psi_0\theta \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_{\psi_0} v_{\psi_1} &= -\frac{\partial \sigma(y)}{\partial \psi_0} \\ &= -\frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \psi_0} \\ &= -\theta \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_{\psi_1} v_{\psi_1} &= -\frac{\partial \sigma(\psi_1)}{\partial \psi_1} - \frac{\partial \sigma(y)}{\partial \psi_1} \\ &= -\frac{\partial \sigma(\psi_1)}{\partial \psi_1} - \frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \psi_1} \\ &= -\sigma(\psi_1)\sigma(-\psi_1) - \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_\theta v_{\psi_1} &= -\frac{\partial \sigma(y)}{\partial \theta} \\ &= -\frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \theta} \\ &= -\psi_0 \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_{\psi_0} v_\theta &= \sigma(y) + \psi_0 \frac{\partial \sigma(y)}{\partial \psi_0} \\ &= \sigma(y) + \psi_0 \frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \psi_0} \\ &= \sigma(y) + \psi_0\theta \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_{\psi_1} v_\theta &= \psi_0 \frac{\partial \sigma(y)}{\partial \psi_0} \\ &= \psi_0 \frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \psi_1} \\ &= \psi_0 \sigma(y)\sigma(-y) \end{aligned}$$

$$\begin{aligned} \partial_\theta v_\theta &= \psi_0 \frac{\partial \sigma(y)}{\partial \theta} \\ &= \psi_0 \frac{\partial \sigma(y)}{\partial y} \frac{\partial y}{\partial \theta} \\ &= \psi_0^2 \sigma(y)\sigma(-y) \end{aligned}$$

Compiling this into the Jacobian,

$$J = \begin{bmatrix} -\theta^2\sigma(y)\sigma(-y) & -\theta\sigma(y)\sigma(-y) & -\sigma(y) - \psi_0\theta\sigma(y)\sigma(-y) \\ -\theta\sigma(y)\sigma(-y) & -\sigma(\psi_1)\sigma(-\psi_1) - \sigma(y)\sigma(-y) & -\psi_0\sigma(y)\sigma(-y) \\ \sigma(y) + \psi_0\theta\sigma(y)\sigma(-y) & \psi_0\sigma(y)\sigma(-y) & \psi_0^2\sigma(y)\sigma(-y) \end{bmatrix}$$

Evaluating the Jacobian at $(\psi_0^*, \psi_1^*, \theta) = (0, 0, 0)$, we obtain:

$$J^* = \begin{bmatrix} 0 & 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}$$

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. The eigenvalues of J^* have been found to be (using Wolfram):

$$\lambda_1 = -\frac{1}{2}, \qquad \lambda_2 = \frac{i}{2}, \qquad \lambda_3 = -\frac{i}{2}$$

As can be seen, the criteria that all eigenvalues of J^* must have a negative real part is only met for one of the three, λ_1. We thus cannot conclude on the local stability around the stationary points in this case from this criterion alone.

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:

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

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. We begin by finding the expression for the regularization term:

$$\begin{aligned} \|\nabla_x C_\psi(x)\|^2 &= \|\frac{\partial}{\partial x}(\psi_0 x + \psi_1)\|^2 \\ &= \|\psi_0\|^2 \\ &= \psi_0^2 \end{aligned}$$

$$\begin{aligned} \mathcal{R}_1(\psi) &= \mathbb{E}_{p_D}\|\nabla_x C_\psi(x)\|^2 \\ &= \int_x p_D(x)\psi_0^2 dx \\ &= \int_x \delta_0(x)\psi_0^2 dx \\ &= \psi_0^2 \end{aligned}$$

$$\begin{aligned} \nabla_\psi \mathcal{R}_1(x) &= [\partial_{\psi_0}\mathcal{R}_1(\psi),\ \partial_{\psi_1}\mathcal{R}_1(\psi)] \\ &= [2\psi_0,\ 0] \end{aligned}$$

The velocity field components are then easily obtained from the previous case:

$$\begin{aligned} \bar{v}_{\psi_0} &= \partial_{\psi_0}\mathcal{L}(\psi, \theta) - \frac{\gamma}{2}\partial_{\psi_0}\mathcal{R}_1(\psi) \\ &= -\theta \cdot \sigma(\psi_0\theta + \psi_1) - \gamma\psi_0 \\ \bar{v}_{\psi_1} &= \partial_{\psi_1}\mathcal{L}(\psi, \theta) - \frac{\gamma}{2}\partial_{\psi_1}\mathcal{R}_1(\psi) \\ &= 1 - \sigma(\psi_1) - \sigma(\psi_0\theta + \psi_1) \\ \bar{v}_\theta &= -\partial_\theta \mathcal{L}(\psi, \theta) \\ &= \psi_0 \cdot \sigma(\psi_0\theta + \psi_1) \end{aligned}$$

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

We find the stationary points of this field by setting each individual component to 0:

$$\begin{aligned}
\bar{v}_\theta &= \psi_0 \cdot \sigma(\psi_0\theta + \psi_1) = 0 \\
\Rightarrow \psi_0^* &= 0 \\
\bar{v}_{\psi_0} &= -\theta \cdot \sigma(\psi_0\theta + \psi_1) - \gamma\psi_0 = 0 \\
\Rightarrow \theta^* &= 0 \\
\bar{v}_{\psi_1} &= 1 - \sigma(\psi_1) - \sigma(\psi_0\theta + \psi_1) = 0 \\
\sigma(\psi_1^*) &= 1 - \sigma(\psi_0^*\theta^* + \psi_1^*) \\
\sigma(\psi_1^*) &= 1 - \sigma(\psi_1^*) \\
\sigma(\psi_1^*) &= \sigma(-\psi_1^*) \\
\Rightarrow \psi_1^* &= 0
\end{aligned}$$

The stationary points of the velocity field are then $(\psi_0^*, \psi_1^*, \theta) = (0, 0, 0)$ once again.

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

Answer. The Jacobian $\bar{J}$ of $\bar{v}$ at any point is obtained through:

$$\bar{J} = \begin{bmatrix} \partial_{\psi_0}\bar{v}_{\psi_0} & \partial_{\psi_1}\bar{v}_{\psi_0} & \partial_{\theta}\bar{v}_{\psi_0} \\ \partial_{\psi_0}\bar{v}_{\psi_1} & \partial_{\psi_1}\bar{v}_{\psi_1} & \partial_{\theta}\bar{v}_{\psi_1} \\ \partial_{\psi_0}\bar{v}_{\theta} & \partial_{\psi_1}\bar{v}_{\theta} & \partial_{\theta}\bar{v}_{\theta} \end{bmatrix}$$

Since $\bar{v}_{\psi_1} = v_{\psi_1}$ and $\bar{v}_\theta = v_\theta$, only the first row of this Jacobian can differ from the one obtained in 5.2. We compute these components:

$$\begin{aligned}
\partial_{\psi_0}\bar{v}_{\psi_0} &= -\theta\frac{\partial\sigma(y)}{\partial\psi_0} - \gamma\frac{\partial\psi_0}{\partial\psi_0} & \partial_{\psi_1}\bar{v}_{\psi_0} &= -\theta\frac{\partial\sigma(y)}{\partial\psi_1} & \partial_{\theta}\bar{v}_{\psi_0} &= -\sigma(y) - \theta\frac{\partial\sigma(y)}{\partial\theta} \\
&= -\theta\frac{\partial\sigma(y)}{\partial y}\frac{\partial y}{\partial\psi_0} - \gamma & &= -\theta\frac{\partial\sigma(y)}{\partial y}\frac{\partial y}{\partial\psi_1} & &= -\sigma(y) - \theta\frac{\partial\sigma(y)}{\partial y}\frac{\partial y}{\partial\theta} \\
&= -\theta^2\sigma(y)\sigma(-y) - \gamma & &= -\theta\sigma(y)\sigma(-y) & &= -\sigma(y) - \psi_0\theta\sigma(y)\sigma(-y)
\end{aligned}$$

The Jacobian of $\bar{v}$ is then

$$\bar{J} = \begin{bmatrix} -\theta^2\sigma(y)\sigma(-y) - \gamma & -\theta\sigma(y)\sigma(-y) & -\sigma(y) - \psi_0\theta\sigma(y)\sigma(-y) \\ -\theta\sigma(y)\sigma(-y) & -\sigma(\psi_1)\sigma(-\psi_1) - \sigma(y)\sigma(-y) & -\psi_0\sigma(y)\sigma(-y) \\ \sigma(y) + \psi_0\theta\sigma(y)\sigma(-y) & \psi_0\sigma(y)\sigma(-y) & \psi_0^2\sigma(y)\sigma(-y) \end{bmatrix}$$

Evaluating the Jacobian at $(\psi_0^*, \psi_1^*, \theta) = (0, 0, 0)$, we obtain:

$$J^* = \begin{bmatrix} -\gamma & 0 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix}$$

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

Answer. The eigenvalues of $\bar{J}^*$ have been found to be (using Wolfram):

$$\lambda_1 = -\frac{1}{2}, \qquad \lambda_2 = \frac{1}{2}\left(-\sqrt{\gamma^2 - 1} - \gamma\right), \qquad \lambda_3 = \frac{1}{2}\left(\sqrt{\gamma^2 - 1} - \gamma\right)$$

The criterion for local asymptotic stability is that all the eigenvalues of $\bar{J}^*$ have a negative real part. We inspect this condition for λ_3, given $\gamma > 1$:

$$\begin{aligned} \gamma &> \sqrt{\gamma^2 - 1} \\ \gamma^2 &> \gamma^2 - 1 \\ 0 &> -1 \end{aligned}$$

which is always true. If $0 < \gamma < 1$, the square roots in λ_2 and λ_3 will produce complex values, but the presence of the $-\gamma$ term in both cases will ensure that these eigenvalues still have a negative real part. We can then conclude that this model will be locally stable in the neighbourhood of the stationary points.

In Problem 2 of the programming assignment, you will verify empirically your claims.