Due Date: April 12th 2021

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* $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$.
We'll start from the fact that for the $l2$ norm of any vector $\boldsymbol{v}$ we have $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^\top \boldsymbol{v}$. Writing the squared error loss using this relation and manipulating:

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

Now taking the expectation w.r.t $\boldsymbol{R}$ and holding all terms that do not depend on it out of the expectation (i.e. $\boldsymbol{y}, \boldsymbol{w}$), we get:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] &= \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{y}^\top \boldsymbol{y}] - \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{y}^\top (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}] - \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{w}^\top (\boldsymbol{X} \odot \boldsymbol{R})^\top \boldsymbol{y}] + \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{w}^\top (\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})\boldsymbol{w}] \\ &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \mathbb{E}_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})]\boldsymbol{w} - \boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top]\boldsymbol{y} + \boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})]\boldsymbol{w} \end{aligned}$$

We can simplify further by noting that since $\boldsymbol{X} \odot \boldsymbol{R}$ is an element-wise product, we can take the expectation element-wise as well, so we can push the expectation over to $\boldsymbol{R}$ and get $\mathbb{E}_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})] = \boldsymbol{X} \odot \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}] = p\boldsymbol{X}$, because the expectation for all elements is the same and is equal to p, so it's like all elements of $\boldsymbol{X}$ have been multiplied by p, so when all elements of a matrix in an element-wise product are the same, it becomes equivalent to a scalar multiplication. Same goes for $\mathbb{E}_{\boldsymbol{R}}[(\boldsymbol{X} \odot \boldsymbol{R})^\top] = (\boldsymbol{X} \odot \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}])^\top = p\boldsymbol{X}^\top$. Plugging these results back to our equation for $\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})]$, and defining $\boldsymbol{M} = (\boldsymbol{X} \odot \boldsymbol{R})^\top (\boldsymbol{X} \odot \boldsymbol{R})$, we arrive at:

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

So we're only left with computing $\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{M}]$. According to the dimensions of $\boldsymbol{X}, \boldsymbol{R}$, we know that $\boldsymbol{M}$ is $d \times d$ dimensional. Let's compute $\boldsymbol{M}_{ij}$. It's the dot product of the $i-$th row of $(\boldsymbol{X} \odot \boldsymbol{R})^\top$, and the $j-$th column of $(\boldsymbol{X} \odot \boldsymbol{R})$:

$$\boldsymbol{M}_{ij} = \langle (\boldsymbol{X} \odot \boldsymbol{R})^\top_{[i,:]}, (\boldsymbol{X} \odot \boldsymbol{R})_{[:,j]} \rangle = \langle (\boldsymbol{X} \odot \boldsymbol{R})_{[:,i]}, (\boldsymbol{X} \odot \boldsymbol{R})_{[:,j]} \rangle$$

Where the last equation comes from replacing the $i-$th row of $(\boldsymbol{X} \odot \boldsymbol{R})^\top$ by the $i-$th column of $(\boldsymbol{X} \odot \boldsymbol{R})$. Now we can rewrite this dot product as a sum:

$$\boldsymbol{M}_{ij} = \langle (\boldsymbol{X} \odot \boldsymbol{R})_{[:,i]}, (\boldsymbol{X} \odot \boldsymbol{R})_{[:,j]} \rangle = \sum_{k=1}^{n} \boldsymbol{X}_{ki} \boldsymbol{R}_{ki} \boldsymbol{X}_{kj} \boldsymbol{R}_{kj}$$

To take the expected of the matrix $\boldsymbol{M}$ we can take the expected of each element, so computing $\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{M}_{ij}]$, and noting that $\boldsymbol{X}$ and its elements are independent of $\boldsymbol{R}$ (thus coming out of the expectations):

$$\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{M}_{ij}] = \mathbb{E}_{\boldsymbol{R}}[\sum_{k=1}^{n} \boldsymbol{X}_{ki} \boldsymbol{R}_{ki} \boldsymbol{X}_{kj} \boldsymbol{R}_{kj}] = \sum_{k=1}^{n} \boldsymbol{X}_{ki} \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{ki} \boldsymbol{R}_{kj}] \boldsymbol{X}_{kj}$$

We know that any element $\boldsymbol{R}_{mn}$ of the matrix $\boldsymbol{R}$ is identically and independently distributed according to a Bern(p). So we have two cases for $\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{ki} \boldsymbol{R}_{kj}]$:

(a) $i \neq j$: Then $\boldsymbol{R}_{ki}, \boldsymbol{R}_{kj}$ would be two different entries of $\boldsymbol{R}$ and thus independent (but identically distributed). In that case $\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{ki} \boldsymbol{R}_{kj}] = \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{ki}] \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{kj}] = p^2$

(b) $i = j$: Then $\boldsymbol{R}_{ki}, \boldsymbol{R}_{kj}$ would be the same entry in $\boldsymbol{R}$ and thus are an identical random variable. In that case $\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{ki} \boldsymbol{R}_{kj}] = \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{ki} \boldsymbol{R}_{ki}]$. Now we can use the hint and write $\mathrm{Var}(Z) + \mathbb{E}[Z]^2 = \mathbb{E}[Z^2]$, with $Z = \boldsymbol{R}_{ki}$. We know that the variance for a Bernoulli random variable is $p(1-p)$, and its mean is $\mathbb{E}[\boldsymbol{R}_{ki}] = p$, thus $\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{R}_{ki} \boldsymbol{R}_{ki}] = p(1-p) + p^2$.

So we have the expectation of $\boldsymbol{M}$. For diagonal entries $(i = j)$:

$$\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{M}_{ii}] = (p^2 + p(1-p)) \sum_{k=1}^{n} \boldsymbol{X}_{ki} \boldsymbol{X}_{ki} = (p^2 + p(1-p))(\boldsymbol{X}^\top \boldsymbol{X})_{ii} = (p^2 + p(1-p))\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})$$

For off-diagonal entries $(i \neq j)$:

$$\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{M}_{ij}] = p^2 \sum_{k=1}^{n} \boldsymbol{X}_{ki} \boldsymbol{X}_{kj} = p^2 (\boldsymbol{X}^\top \boldsymbol{X})_{ij}$$

Finally adding these terms we have:

$$\mathbb{E}_{\boldsymbol{R}}[\boldsymbol{M}] = (p^2 + p(1-p))\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X}) + p^2 (\boldsymbol{X}^\top \boldsymbol{X})_{ij} = p^2 (\boldsymbol{X}^\top \boldsymbol{X}) + p(1-p)\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})$$

Where last equality comes from the fact that off-diagonal terms have zeros on the diagonal and adding $p^2 \mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})$ from the contribution of diagonal entries to the expectation gives $p^2 (\boldsymbol{X}^\top \boldsymbol{X})$. Now we have everything to compute $\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})]$:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})] &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top p \boldsymbol{X} \boldsymbol{w} - \boldsymbol{w}^\top p \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{R}}[\boldsymbol{M}] \boldsymbol{w} \\
&= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top p \boldsymbol{X} \boldsymbol{w} - \boldsymbol{w}^\top p \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{w}^\top (p^2 \boldsymbol{X}^\top \boldsymbol{X} + p(1-p)\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})) \boldsymbol{w}
\end{aligned}$$

For a diagonal matrix D we have $D = D^{1/2} D^{1/2}$, and also $D^{1/2} = (D^{1/2})^\top$, therefore $D = (D^{1/2})^\top D^{1/2}$. Now applying this relation to $\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})$ yields the following for the last term $\boldsymbol{w}^\top (\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})) \boldsymbol{w}$ in the equation above:

$$\boldsymbol{w}^\top (\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})) \boldsymbol{w} = \boldsymbol{w}^\top (\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})^{1/2})^\top (\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})^{1/2}) \boldsymbol{w} = \|\Gamma \boldsymbol{w}\|^2$$

Plugging back:

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

This concludes the proof.

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} ?

Let's take the derivative w.r.t $\boldsymbol{w}$ to obtain the stationary points. We also note that the derivative of the squared norm of a vector-valued function is $\frac{\partial\|\boldsymbol{f}(\boldsymbol{x})\|^2}{\partial\boldsymbol{x}} = \frac{\partial\boldsymbol{f}(\boldsymbol{x})^\top\boldsymbol{f}(\boldsymbol{x})}{\partial\boldsymbol{x}} = 2\frac{\partial\boldsymbol{f}(\boldsymbol{x})^\top}{\partial\boldsymbol{x}}\boldsymbol{f}(\boldsymbol{x})$.

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

Setting that equal to zero to get to stationary points:

$$\begin{aligned}&\frac{\partial\mathbb{E}_{\boldsymbol{R}}[L(\boldsymbol{w})]}{\partial\boldsymbol{w}} = -2p\boldsymbol{X}^\top(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) + 2p(1-p)\Gamma^\top\Gamma\boldsymbol{w} = 0 \\ &\implies p\boldsymbol{X}^\top(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) = (1-p)\Gamma^\top\Gamma p\boldsymbol{w} \\ &\overset{\Gamma \text{ is diagonal}}{\implies} p\boldsymbol{X}^\top\boldsymbol{y} - p\boldsymbol{X}^\top\boldsymbol{X}p\boldsymbol{w} = (1-p)\Gamma^2 p\boldsymbol{w} \\ &\implies p\boldsymbol{w} = (p\boldsymbol{X}^\top\boldsymbol{X} + (1-p)\Gamma^2)^{-1}p\boldsymbol{X}^\top\boldsymbol{y} = (\boldsymbol{X}^\top\boldsymbol{X} + \frac{(1-p)}{p}\Gamma^2)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\ &\implies p\boldsymbol{w}^{\text{dropout}} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top\boldsymbol{y}\end{aligned}$$

Why is this a minimizer of the expected loss ? Because the expected loss is convex in $\boldsymbol{w}$ (sum of two convex terms is also convex), and the stationary point of a convex function is a minimum. $\lambda^{\text{dropout}} = \frac{1-p}{p}$ is a monotonically decreasing function of p vanishing at $p = 1$ and diverging to ∞ at $p = 0$. In this simple linear regression case, as $p \to 1$ and as result $\lambda \to 0$, it means that no input will be dropped out and the problem reduces to a normal linear regression, and this is confirmed by the fact that the normal equation for linear regression is recovered in this case:

$$\begin{aligned}\lim_{p\to 1} p\boldsymbol{w} &= \lim_{p\to 1}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\ \boldsymbol{w} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}\end{aligned}$$

On the other hand, as $p \to 0$, and as a result $\lambda \to \infty$, it means that almost no input will be kept, which results in output being degenerate and only yielding zero outputs. In this case, the regularization term gets much higher weight and since it's in an inverse form, it'll push $p\boldsymbol{w}$ towards zero to keep this degenerate state. So we can see the tradeoffs and based on this intuition, a moderate value should be used for p in more complex settings other than linear regression.

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 $\boldsymbol{w}^{L_2}$.

For a squared error loss function with L2 penalty, we then have:

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

Note that again the loss is convex in $\boldsymbol{w}$ so its stationary point will be a minimizer of the loss. Also here we don't have any expectation as there is no stochastic element present. Taking the derivative and setting it to zero:

$$\begin{aligned}
\frac{\partial L(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{\partial}{\partial \boldsymbol{w}}\|(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w})\|^2 + \frac{\partial}{\partial \boldsymbol{w}}\lambda^{L_2}\|\boldsymbol{w}\|^2 \\
&= 2\frac{\partial(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w})^\top}{\partial \boldsymbol{w}}(\boldsymbol{y} - p\boldsymbol{X}\boldsymbol{w}) + 2\lambda^{L_2}\frac{\partial(\boldsymbol{w})^\top}{\partial \boldsymbol{w}}\boldsymbol{w} \\
&= -2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + 2\lambda^{L_2}\boldsymbol{w} = 0 \\
&\implies \boldsymbol{w}^{L_2} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}$$

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).

Let's look at them at the same time:

$$\begin{aligned}
p\boldsymbol{w}^{\text{dropout}} &= (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{\text{dropout}}\Gamma^2)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
\boldsymbol{w}^{L_2} &= (\boldsymbol{X}^\top\boldsymbol{X} + \lambda^{L_2}\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}
\end{aligned}$$

Both add the regularization term in the inverse term besides $\boldsymbol{X}^\top\boldsymbol{X}$, and as in each case $\lambda \to 0$, we recover the normal equation, the solution of unaltered linear regression $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. In the other limit ($\lambda \to \infty$), both of them push all the weights (or $p\boldsymbol{w}$) to 0 (because of the large term in the inverse), but with different mechanisms. One important difference is that L2 penalty regularizes all weights in a balanced way, i.e. doesn't impose stronger regularization on any set of weights, whereas dropout does. Since the denominator depends on $\lambda\Gamma^2$, and since Γ^2 contains the variance in every direction, regularization is affected by the variance in each direction and thus is not the same for all directions/weights.

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 the evidence lower bound (ELBO):

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

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. $q_\phi(\boldsymbol{z}|\boldsymbol{x}) := q(\boldsymbol{z}; \pi_\phi(\boldsymbol{x}))$.

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.

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 $p(\boldsymbol{z}|\boldsymbol{x})$.

$$\begin{aligned}
\max_\theta \text{ECLL} &= \max_\theta \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})] = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x},\boldsymbol{z})] \\
&= \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\Big(\log p_\theta(\boldsymbol{x},\boldsymbol{z}) - \log p_\theta(\boldsymbol{x}) + \log p_\theta(\boldsymbol{x})\Big)d\boldsymbol{z} \\
&= \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\Big(\log p_\theta(\boldsymbol{z}\mid\boldsymbol{x}) + \log p_\theta(\boldsymbol{x})\Big)d\boldsymbol{z} \\
&= \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\Big(\log p_\theta(\boldsymbol{z}\mid\boldsymbol{x}) - \log q(\boldsymbol{z}\mid\boldsymbol{x}) + \log q(\boldsymbol{z}\mid\boldsymbol{x}) + \log p_\theta(\boldsymbol{x})\Big)d\boldsymbol{z} \\
&= \max_\theta -\int q(\boldsymbol{z}\mid\boldsymbol{x})\Big(\log q(\boldsymbol{z}\mid\boldsymbol{x}) - \log p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Big)d\boldsymbol{z} + \int q(\boldsymbol{z}\mid\boldsymbol{x})\Big(\log q(\boldsymbol{z}\mid\boldsymbol{x}) + \log p_\theta(\boldsymbol{x})\Big)d\boldsymbol{z} \\
&= \max_\theta -D_{\mathrm{KL}}\Big(q(\boldsymbol{z}\mid\boldsymbol{x})\|p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Big) + \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\Big(\log q(\boldsymbol{z}\mid\boldsymbol{x}) + \log p_\theta(\boldsymbol{x})\Big)d\boldsymbol{z} \\
&= \max_\theta -D_{\mathrm{KL}}\Big(q(\boldsymbol{z}\mid\boldsymbol{x})\|p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Big) + \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\log q(\boldsymbol{z}\mid\boldsymbol{x})d\boldsymbol{z} + \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\log p_\theta(\boldsymbol{x})d\boldsymbol{z}
\end{aligned}$$

But $\max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\log q(\boldsymbol{z}\mid\boldsymbol{x})d\boldsymbol{z}$ has no dependence on θ, thus we can discard it from the maximization as the choice of θ doesn't affect it.

$$\max_\theta \text{ECLL} = \max_\theta -D_{\mathrm{KL}}\Big(q(\boldsymbol{z}\mid\boldsymbol{x})\|p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Big) + \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\log p_\theta(\boldsymbol{x})d\boldsymbol{z}$$

As $p_\theta(\boldsymbol{x})$ is independent of $\boldsymbol{z}, q(\boldsymbol{z}\mid\boldsymbol{x})$ (because q is fixed by assumption), then we can pull it out from the integral, and the integral reduces to 1 (integral of density over its domain is 1) $\int q(\boldsymbol{z}\mid\boldsymbol{x})\log p_\theta(\boldsymbol{x})d\boldsymbol{z} = \log p_\theta(\boldsymbol{x})\int q(\boldsymbol{z}\mid\boldsymbol{x})d\boldsymbol{z} = \log p_\theta(\boldsymbol{x})$. Pluggin this result back:

$$\begin{aligned}
\max_\theta \text{ECLL} &= \max_\theta -D_{\mathrm{KL}}\Big(q(\boldsymbol{z}\mid\boldsymbol{x})\|p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Big) + \max_\theta \int q(\boldsymbol{z}\mid\boldsymbol{x})\log p_\theta(\boldsymbol{x})d\boldsymbol{z} \\
&= \max_\theta -D_{\mathrm{KL}}\Big(q(\boldsymbol{z}\mid\boldsymbol{x})\|p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Big) + \max_\theta \log p_\theta(\boldsymbol{x})
\end{aligned}$$

Thus we have shown that under the stated assumptions, maximizing ECLL yields the same result as maximizing $\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}\Big(q(\boldsymbol{z}\mid\boldsymbol{x})\|p_\theta(\boldsymbol{z}\mid\boldsymbol{x})\Big)$

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 ?

We know that the log-likelihood is given by:

$$\log p(x) = \mathcal{L}(p, q) + D_{\mathrm{KL}}(q(\boldsymbol{z} \mid \boldsymbol{x})\|p(\boldsymbol{z} \mid \boldsymbol{x})$$

But since θ is fixed, then the likelihood of the data under the generator ($\log p_\theta(x)$ is fixed, therefore, we can compare ELBO to get a comparison for KL divergences:

$$\begin{aligned}
&\mathcal{L}(p_\theta, q_{\phi^*}) + D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) = \mathcal{L}(p_\theta, q_i^*) + D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) \\
\Longrightarrow &\mathcal{L}(p_\theta, q_{\phi^*}) - \mathcal{L}(p_\theta, q_i^*) = D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) - D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))
\end{aligned}$$

For comparing ELBO, note that necessarily $\mathcal{L}(p_\theta, q_{\phi^*}) \leq \mathcal{L}(p_\theta, q_i^*)$, because otherwise it should be the case that $\mathcal{L}(p_\theta, q_{\phi^*}) > \mathcal{L}(p_\theta, q_i^*)$, and this results in contradiction, because q_i^* supposed to be $\arg\max_q \mathcal{L}(\theta, q; \boldsymbol{x}_i)$, but now we have found another solution ϕ^* that is makes instance-dependent ELBO even larger which contradicts the fact that q_i^* resulted in the maximum ELBO. So $\mathcal{L}(p_\theta, q_{\phi^*}) \leq \mathcal{L}(p_\theta, q_i^*)$, and since we had:

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

, then $\mathcal{L}(p_\theta, q_{\phi^*}) \geq \mathcal{L}(p_\theta, q_i^*)$ is equivalent to the following:

$$\begin{aligned}
&D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) - D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) \leq 0 \\
\Longrightarrow &D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) \leq D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))
\end{aligned}$$

So KL divergence for ϕ^* is bigger (in fact greater or equal than KL divergence for q_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)

The bias of estimation is just the following:

$$\begin{aligned}
\mathrm{bias}_{\phi*} &= \log p_\theta(x) - \mathcal{L}(p_\theta, q_{\phi^*}) = D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) \\
\mathrm{bias}_{q_i*} &= \log p_\theta(x) - \mathcal{L}(p_\theta, q_i^*) = D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))
\end{aligned}$$

So we need to compare KL divergences. From previous section we know that KL for ϕ^* is *not* strictly greater than that of q_i^*, it's *greater or equal* than that. So in the best case scenario that is mentioned, we have the equality and thus $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x})) = D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z} \mid \boldsymbol{x})\|p_\theta(\boldsymbol{z} \mid \boldsymbol{x}))$. Therefore both approaches to estimating the marginal likelihood have the same bias.

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

If we're to estimate the marginal log-likelihood for all n samples in the training dataset, then using ϕ^*, we only do $\arg\max_\phi \sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ once for a summation, and then use it n times to obtain the log-likelihood of each sample. But using q_i^*, we need to do $\arg\max_q \mathcal{L}(\theta, q; \boldsymbol{x}_i)$ for every sample, and then use the results to get the log-likelihood estimate of each sample by ELBO. So we do $\arg\max_q \mathcal{L}(\theta, q; \boldsymbol{x}_i)$ for $n-1$ times more which renders this approach compute inefficient.

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

As was mentioned above, for ϕ^* we only need one neural net that gives $\arg\max_\phi \sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ so we need to store only parameters for this network. Whereas in the other approach we need to store the parameters of n networks, one per instance-dependent variational approximator, so it's much more storage ($n-1$ times the former case). Therefore, again, using ϕ^* is more efficient, this time in terms of memory. We note that these benefits in memory and compute are at the price of a larger bias when we're not in the best case scenario.

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 $F : \mathcal{U} \to \mathcal{X}$ 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.

We know from class slides that if f is an invertible function, then:

$$X \sim p_X, Y = f(X) \longrightarrow p_Y(y) = p_X(x)|\det \frac{\partial f^{-1}(y)}{\partial y}|$$

Applying this identity to $P_{F(U)}, P_{F^{-1}(X)}$, we get:

$$U \sim p_U, X = F(U) \longrightarrow p_{F(U)}(x) = p_U(u)|\det \frac{\partial F^{-1}(x)}{\partial x}| = p_U(F^{-1}(x))|\det \frac{\partial F^{-1}(x)}{\partial x}|$$
$$X \sim p_X, U = F^{-1}(X) \longrightarrow p_{F^{-1}(X)}(u) = p_X(x)|\det \frac{\partial F(u)}{\partial u}| = p_X(F(u))|\det \frac{\partial F(u)}{\partial u}|$$

Plugging these results in KL divergences:

$$\begin{aligned} D_{KL}[P_X||P_{F(U)}] &= \int p_X(x) \log \frac{p_X(x)}{p_{F(U)}(x)} dx = \int p_X(x) \log \frac{p_X(x)}{p_U(F^{-1}(x))|\det \frac{\partial F^{-1}(x)}{\partial x}|} dx \\ &= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) - \log|\det \frac{\partial F^{-1}(x)}{\partial x}|] \\ D_{KL}[P_{F^{-1}(X)}||P_U] &= \int p_{F^{-1}(X)}(u) \log \frac{p_{F^{-1}(X)}(u)}{p_U(u)} du = \int p_{F^{-1}(X)}(u) \log \frac{p_X(F(u))|\det \frac{\partial F(u)}{\partial u}|}{p_U(u)} du \\ &= \mathbb{E}_{u\sim P_{F^{-1}(X)}}[\log p_X(F(u)) - \log p_U(u) + \log|\det \frac{\partial F(u)}{\partial u}|] \end{aligned}$$

Now they look very similar. Using the identity that $|\det J(g)| = \frac{1}{|\det J(g^{-1})|}$, where J denotes the Jacobian, we have that:

$$|\det \frac{\partial F^{-1}(x)}{\partial x}| = \frac{1}{|\det \frac{\partial F(x)}{\partial x}|}$$

Substituting this result back:

$$\begin{aligned}
D_{KL}[P_X||P_{F(U)}] &= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) - \log|\det \frac{\partial F^{-1}(x)}{\partial x}|] \\
&= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) - \log \frac{1}{|\det \frac{\partial F(x)}{\partial x}|}] \\
&= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) + \log|\det \frac{\partial F(x)}{\partial x}|] \\
D_{KL}[P_{F^{-1}(X)}||P_U] &= \mathbb{E}_{u\sim P_{F^{-1}(X)}}[\log p_X(F(u)) - \log p_U(u) + \log|\det \frac{\partial F(u)}{\partial u}|]
\end{aligned}$$

We have to recall that if $Y = g(X)$ and g is some invertible function, then the expectation of any function $h(Y)$ w.r.t. the distribution of Y is:

$$\mathbb{E}_{y\sim P_Y}[h(Y)] = \int h(y)p_Y(y)dy = \int h(g(x))p_X(x)dx = \mathbb{E}_{x\sim P_X}[h(g(X))]$$

Now since $U = F^{-1}(X)$, similar to the above we have:

$$\mathbb{E}_{u\sim P_{F^{-1}(X)}}[h(U)] = \int h(u)p_{F^{-1}(X)}(u)du = \int h(F^{-1}(x))p_X(x)dx = \mathbb{E}_{x\sim P_X}[h(F^{-1}(X))]$$

Now if we define $h(U) = \log p_X(F(u)) - \log p_U(u) + \log|\det \frac{\partial F(u)}{\partial u}|$, replacing every u by $F^{-1}(x)$ will give the expectation w.r.t. P_X:

$$\begin{aligned}
D_{KL}[P_{F^{-1}(X)}||P_U] &= \mathbb{E}_{u\sim P_{F^{-1}(X)}}[h(U)] = \mathbb{E}_{x\sim P_X}[h(F^{-1}(X))] \\
&= \mathbb{E}_{x\sim P_X}[\log p_X(F(F^{-1}(x))) - \log p_U(F^{-1}(x)) + \log|\det \frac{\partial F(F^{-1}(x))}{\partial F^{-1}(x)}|] \\
&= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) + \log|\det \frac{\partial x}{\partial F^{-1}(x)}|] \\
&= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) - \log|\det \frac{\partial F^{-1}(x)}{\partial x}|] \\
&= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) - \log|\frac{1}{\det \frac{\partial F(x)}{\partial x}}|] \\
&= \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) + \log|\det \frac{\partial F(x)}{\partial x}|] \\
&= D_{KL}[P_X||P_{F(U)}]
\end{aligned}$$

Thus 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.

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.

Let's take a look at the form of KL divergence written as an expectation:

$$D_{KL}[P_X||P_{F(U)}] = \mathbb{E}_{x\sim P_X}[\log \frac{p_X(x)}{p_{F(U)}(x)}] = \mathbb{E}_{x\sim P_X}[\log p_X(x) - \log p_U(F^{-1}(x)) + \log|\det \frac{\partial F(x)}{\partial x}|]$$

$$D_{KL}[P_{F^{-1}(X)}||P_U] = \mathbb{E}_{u\sim P_{F^{-1}(X)}}[\log \frac{p_{F^{-1}(X)}(u)}{p_U(u)}] = \mathbb{E}_{u\sim P_{F^{-1}(X)}}[\log p_X(F(u)) - \log p_U(u) + \log|\det \frac{\partial F(u)}{\partial u}|]$$

It's now clear that if we choose forward KL as our objective, then its gradients w.r.t. flow parameters (F is parameterized by a neural network) don't depend on $\log p_X(x)$, therefore we don't need to evaluate it, so we'll use scenario **2** with **forward KL**. On the other hand, if we choose reverse KL as our objective, then its gradients w.r.t. flow parameters F do depend on $\log p_X(F(u))$, therefore we do need to be able to evaluate $p_X(x)$ at any point, but since we sample from $P_{F^{-1}(X)}$, we aren't sampling from $P_X(x)$ so we'll use scenario **1** with **reverse KL**.

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 $\mu = \frac{1}{2}(p_0 + p_1)$.

We'll first derive the closed form solution for D^* using calculus of variations. From sec. 19.4.2 we know that for differentiable functions $f(\boldsymbol{x})$ and differentiable functions $g(y, \boldsymbol{x})$ with continuous derivatives, that ($\frac{\delta}{\delta f(\boldsymbol{x})}$ denotes the functional derivative)

$$\frac{\delta}{\delta f(\boldsymbol{x})} \int g(f(\boldsymbol{x}), \boldsymbol{x}) d\boldsymbol{x} = \frac{\partial}{\partial y} g(f(\boldsymbol{x}), \boldsymbol{x})$$

To find D^*, we need to compute the following:

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

Rewriting it in terms of integral (instead of expectations):

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

Taking the functional derivative w.r.t. D and setting to zero to find D^*, we get (using the corresponding densities f_0, f_1 for distributions p_0, p_1)

$$\frac{\delta}{\delta D}\int f_1(\boldsymbol{x})\log D(\boldsymbol{x})d\boldsymbol{x} + \frac{\delta}{\delta D}\int f_0(\boldsymbol{x})\log(1-D(\boldsymbol{x}))d\boldsymbol{x} = 0$$

Now using the functional derivative property from the Deep Learning book, by defining $y = D(\boldsymbol{x}), g_1(y,\boldsymbol{x}) = f_1(\boldsymbol{x})\log y, g_0(y,\boldsymbol{x}) = f_0(\boldsymbol{x})\log y$ we arrive at:

$$\begin{aligned}
\frac{\delta}{\delta D}\text{log-likelihood} &= \frac{\delta}{\delta D}\int f_1(\boldsymbol{x})\log D(\boldsymbol{x})d\boldsymbol{x} + \frac{\delta}{\delta D}\int f_0(\boldsymbol{x})\log(1-D(\boldsymbol{x}))d\boldsymbol{x} \\
&= \frac{\partial}{\partial y}(f_1(\boldsymbol{x})\log y) + \frac{\partial}{\partial y}(f_0(\boldsymbol{x})\log 1-y) \\
&= \frac{f_1(\boldsymbol{x})}{y} - \frac{f_0(\boldsymbol{x})}{1-y} \\
&= \frac{f_1(\boldsymbol{x})}{D(\boldsymbol{x})} - \frac{f_0(\boldsymbol{x})}{1-D(\boldsymbol{x})} \qquad \text{(Replacing } y = D(\boldsymbol{x})\text{)} \\
&= 0 \qquad \text{(setting } \frac{\delta}{\delta D}\text{log-likelihood} = 0 \text{ to obtain } D^*\text{)}
\end{aligned}$$

From the last equality above we can find D^*:

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

Now for estimating the JSD, let's write and manipulate it:

$$\text{JSD}(p_0,p_1) = \frac{1}{2}\big(KL(p_0\|\mu) + KL(p_1\|\mu)\big) = \frac{1}{2}\int f_0(\boldsymbol{x})\log\frac{f_0(\boldsymbol{x})}{\frac{f_0(\boldsymbol{x})+f_1(\boldsymbol{x})}{2}}d\boldsymbol{x} + \frac{1}{2}\int f_1(\boldsymbol{x})\log\frac{f_1(\boldsymbol{x})}{\frac{f_0(\boldsymbol{x})+f_1(\boldsymbol{x})}{2}}d\boldsymbol{x}$$

But note that from the equation for $D^*(\boldsymbol{x})$ we can easily get:

$$\begin{aligned}
\frac{f_1(\boldsymbol{x})}{\frac{f_0(\boldsymbol{x})+f_1(\boldsymbol{x})}{2}} &= 2D^*(\boldsymbol{x}) \\
\frac{f_0(\boldsymbol{x})}{\frac{f_0(\boldsymbol{x})+f_1(\boldsymbol{x})}{2}} = \frac{f_0(\boldsymbol{x})}{f_1(\boldsymbol{x})}\frac{f_1(\boldsymbol{x})}{\frac{f_0(\boldsymbol{x})+f_1(\boldsymbol{x})}{2}} &= (\frac{1}{D^*(\boldsymbol{x})} - 1)2D^*(\boldsymbol{x}) = 2 - 2D^*(\boldsymbol{x})
\end{aligned}$$

Plugging these results back into the equation for $\text{JSD}(p_0,p_1)$:

$$\begin{aligned}
\text{JSD}(p_0,p_1) &= \frac{1}{2}\int f_0(\boldsymbol{x})\log\frac{f_0(\boldsymbol{x})}{\frac{f_0(\boldsymbol{x})+f_1(\boldsymbol{x})}{2}}d\boldsymbol{x} + \frac{1}{2}\int f_1(\boldsymbol{x})\log\frac{f_1(\boldsymbol{x})}{\frac{f_0(\boldsymbol{x})+f_1(\boldsymbol{x})}{2}}d\boldsymbol{x} \\
&= \frac{1}{2}\int f_0(\boldsymbol{x})\log 2(1-D^*(\boldsymbol{x}))d\boldsymbol{x} + \frac{1}{2}\int f_1(\boldsymbol{x})\log 2D^*(\boldsymbol{x})d\boldsymbol{x} \\
&= \frac{1}{2}\int f_0(\boldsymbol{x})\log 2d\boldsymbol{x} + \frac{1}{2}\int f_0(\boldsymbol{x})\log(1-D^*(\boldsymbol{x}))d\boldsymbol{x} \\
&+ \frac{1}{2}\int f_1(\boldsymbol{x})\log 2d\boldsymbol{x} + \frac{1}{2}\int f_1(\boldsymbol{x})\log D^*(\boldsymbol{x})d\boldsymbol{x}
\end{aligned}$$

But $\int f_0(\boldsymbol{x}) \log 2 d\boldsymbol{x} = \log 2 \int f_0(\boldsymbol{x}) d\boldsymbol{x} = \log 2$ (same for the integral over f_1), therefore:

$$\begin{aligned}
\text{JSD}(p_0, p_1) &= \frac{1}{2}\int f_0(\boldsymbol{x}) \log 2 d\boldsymbol{x} + \frac{1}{2}\int f_0(\boldsymbol{x}) \log(1 - D^*(\boldsymbol{x})) d\boldsymbol{x} \\
&+ \frac{1}{2}\int f_1(\boldsymbol{x}) \log 2 d\boldsymbol{x} + \frac{1}{2}\int f_1(\boldsymbol{x}) \log D^*(\boldsymbol{x}) d\boldsymbol{x} \\
&= \log 2 + \frac{1}{2}\Big(\mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D^*(\boldsymbol{x}))]\Big) \\
&= \log 2 + \frac{1}{2}\text{log-likelihood}(D^*, G)
\end{aligned}$$

We can use either of the last two equations above to estimate the JSD (G is the generator with distribution p_0) log-likelihood is the discriminator loss.

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

From the equation of the optimal discriminator we have that:

$$D^*(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})}$$

With a bit of manipulation we get:

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

This concludes the derivation.

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:

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

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.

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]

The velocity field is $\boldsymbol{v} = (\boldsymbol{v}_\psi, \boldsymbol{v}_\theta)^\top = (\boldsymbol{v}_{\psi_0}, \boldsymbol{v}_{\psi_1}, \boldsymbol{v}_\theta)^\top = (\frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_0}, \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_1}, -\frac{\partial \mathcal{L}(\psi,\theta)}{\partial \theta})^\top$. To lighten the notation with derivatives of the sigmoid function, let's define $\delta(x)$ according to the following:

$$\frac{d}{dx}\sigma(f(x)) = \frac{d}{df(x)}\sigma(f(x))\frac{df(x)}{dx} = \sigma(f(x))(1-\sigma(f(x)))\frac{df(x)}{dx} = \delta(f(x))\frac{df(x)}{dx}$$

Please note that this δ is *not* the Dirac function, it's just a temporary notation to lighten the equations. I'll use δ as Dirac function later. Now by noting that we can push partials into the expectation, we may compute the velocity field:

$$\begin{aligned}
\boldsymbol{v} &= \begin{pmatrix} \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_0} \\ \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_1} \\ -\frac{\partial \mathcal{L}(\psi,\theta)}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{p_D}[\delta(C_\psi(x))x/\sigma(C_\psi(x))] + \mathbb{E}_{p_\theta}[-\delta(-C_\psi(x))x/\sigma(-C_\psi(x))] \\ \mathbb{E}_{p_D}[\delta(C_\psi(x))/\sigma(C_\psi(x))] + \mathbb{E}_{p_\theta}[-\delta(-C_\psi(x))/\sigma(-C_\psi(x))] \\ -\nabla_\theta[\mathbb{E}_{p_D}\log(\sigma(C_\psi(x))) + \mathbb{E}_{p_\theta}\log(\sigma(-C_\psi(x)))] \end{pmatrix} \\
&= \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ 0 - \nabla_\theta[\mathbb{E}_{p_\theta}\log(\sigma(-C_\psi(x)))] \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ \mathbb{E}_{p_\theta}[\psi_0\delta(-C_\psi(x))/\sigma(-C_\psi(x))] \end{pmatrix} \\
&= \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ \mathbb{E}_{p_\theta}[\psi_0\sigma(C_\psi(x))] \end{pmatrix}
\end{aligned}$$

Now for computing the expectations, we notice the following property for any function f with Dirac function focused at x_0 ($\delta(x - x_0)$):

$$\int_{-\infty}^{+\infty} f(x)\delta(x - x_0)dx = f(x_0)$$

Now if $g_X(x) = \delta(x - x_0)$ is a density function we can equivalently write:

$$\int_{-\infty}^{+\infty} f(x)\delta(x - x_0)dx = \mathbb{E}_{x\sim g_X(x)}[f(x)] = f(x_0)$$

Now we can replace the expectations by the function inside them evaluated at $x = 0$ for $\mathbb{E}_{p_D}$, and at $x = \theta$ for $\mathbb{E}_{p_\theta}$:

$$\boldsymbol{v} = \begin{pmatrix} \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_0} \\ \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_1} \\ -\frac{\partial \mathcal{L}(\psi,\theta)}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ \mathbb{E}_{p_\theta}[\psi_0\sigma(C_\psi(x))] \end{pmatrix} = \begin{pmatrix} -\theta\sigma(C_\psi(\theta)) \\ \sigma(-\psi_1) - \sigma(C_\psi(\theta)) \\ \psi_0\sigma(C_\psi(\theta)) \end{pmatrix}$$

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.

Setting it to zero:

$$\boldsymbol{v} = \begin{pmatrix} -\theta\sigma(C_\psi(\theta)) \\ \sigma(-\psi_1) - \sigma(C_\psi(\theta)) \\ \psi_0\sigma(C_\psi(\theta)) \end{pmatrix} = \boldsymbol{O}$$

All entries in $\boldsymbol{v}$ should be zero, and since sigmoid can't be zero, therefore $\theta = \psi_0 = 0$ for $\boldsymbol{v}_1 = \boldsymbol{v}_3 = 0$. Then for $\boldsymbol{v}_2$ to be 0, we need $\sigma(-\psi_1) = \sigma(C_\psi(\theta))$. For two sigmoids to be equal, their input should be equal (because sigmoid is a bijection), therefore we need $-\psi_1 = C_\psi(\theta) = \psi_0\theta + \psi_1 = \psi_1$ (because we found that $\psi_0 = \theta = 0$), hence $\psi_1 = 0$ as well. Thus the velocity field has a single stationary point $(\psi_0^*, \psi_1^*, \theta^*) = (0, 0, 0)$.

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

Since $\boldsymbol{v}$ is a vector-valued function, a Jacobian is defined for it as follows (again let $g(x) = \sigma(x)(1 - \sigma(x))$:

$$J(\boldsymbol{v}) = \begin{pmatrix} \frac{\partial \boldsymbol{v}_{\psi_0}}{\partial \psi_0} & \frac{\partial \boldsymbol{v}_{\psi_0}}{\partial \psi_1} & \frac{\partial \boldsymbol{v}_{\psi_0}}{\partial \theta} \\ \frac{\partial \boldsymbol{v}_{\psi_1}}{\partial \psi_0} & \frac{\partial \boldsymbol{v}_{\psi_1}}{\partial \psi_1} & \frac{\partial \boldsymbol{v}_{\psi_1}}{\partial \theta} \\ \frac{\partial \boldsymbol{v}_{\theta}}{\partial \psi_0} & \frac{\partial \boldsymbol{v}_{\theta}}{\partial \psi_1} & \frac{\partial \boldsymbol{v}_{\theta}}{\partial \theta} \end{pmatrix} = \begin{pmatrix} -\theta^2 g(C_\psi(\theta)) & -\theta g(C_\psi(\theta)) & -\sigma(C_\psi(\theta)) \\ -\theta g(C_\psi(\theta)) & -g(-\psi_1) - g(C_\psi(\theta)) & -\psi_0 g(C_\psi(\theta)) \\ \sigma(C_\psi(\theta)) + \psi_0\theta g(C_\psi(\theta)) & \psi_0 g(C_\psi(\theta)) & \psi_0^2 g(C_\psi(\theta)) \end{pmatrix}$$

Computing $J^* = J(\boldsymbol{v}^*)$ by plugging $(\psi_0^*, \psi_1^*, \theta^*) = (0, 0, 0)$ into the Jacobian above:

$$\begin{aligned} J(\boldsymbol{v}^*) &= \begin{pmatrix} -\theta^2 g(C_\psi(\theta)) & -\theta g(C_\psi(\theta)) & -\sigma(C_\psi(\theta)) \\ -\theta g(C_\psi(\theta)) & -g(-\psi_1) - g(C_\psi(\theta)) & -\psi_0 g(C_\psi(\theta)) \\ \sigma(C_\psi(\theta)) + \psi_0\theta g(C_\psi(\theta)) & \psi_0 g(C_\psi(\theta)) & \psi_0^2 g(C_\psi(\theta)) \end{pmatrix}_{|\boldsymbol{v}=\boldsymbol{v}*} \\ &= \begin{pmatrix} 0 & 0 & -\sigma(C_\psi(0)) \\ 0 & -g(-\psi_1^*) - g(C_\psi(0)) & 0 \\ \sigma(C_\psi(0)) & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -\sigma(\psi_1^*) \\ 0 & -g(-\psi_1^*) - g(\psi_1^*) & 0 \\ \sigma(\psi_1^*) & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & -\sigma(0) \\ 0 & -g(0) - g(0) & 0 \\ \sigma(0) & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -1/2 \\ 0 & -1/4 - 1/4 & 0 \\ 1/2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & -1/2 \\ 0 & -1/2 & 0 \\ 1/2 & 0 & 0 \end{pmatrix} \\ \implies J^* &= \begin{pmatrix} 0 & 0 & -1/2 \\ 0 & -1/2 & 0 \\ 1/2 & 0 & 0 \end{pmatrix} \end{aligned}$$

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.

Following the definition of eigenvalues, λ is an eigenvalue of J^* if there exists some $\boldsymbol{u}$ for which we have $J^*\boldsymbol{u} = \lambda\boldsymbol{u}$. This is equivalent to $(J^* - \lambda\boldsymbol{I})\boldsymbol{u} = \boldsymbol{O}$. In order to have a non-trivial solution for $\boldsymbol{u}$, $(J^* - \lambda\boldsymbol{I})$ should *not* be invertible or equivalently, it should be the case that $|\det(J^* - \lambda\boldsymbol{I})| = 0$. Under such case, $\boldsymbol{u}$ will have a non-trivial solution. Now we'll compute the

$|\det(J^* - \lambda \boldsymbol{I})|$, set it to zero and find the eigenvalues and eigenvectors:

$$\det(J^* - \lambda \boldsymbol{I}) = \det \begin{pmatrix} -\lambda & 0 & -1/2 \\ 0 & -1/2 - \lambda & 0 \\ 1/2 & 0 & -\lambda \end{pmatrix} = -\lambda^2(\lambda + 1/2) - 1/4(1/2 + \lambda)$$
$$= -(\lambda + 1/2)(\lambda^2 + 1/4) = 0$$

So eigenvalues would be $\{-1/2, i/2, -i/2\}$ (where i denotes the complex values). It's clear that not all eigenvalues have negative real parts, hence the dynamical system is not stable in two directions (around the stationary point). It doesn't diverge as there's no positive real part to any eigenvalue, but rather there are complex eigenvalues that result in oscillation. We can find which directions oscillate and which direction damps by finding the eigenvectors. It's not asked here but after carrying out the computations for eigenvectors, they're as follows:

$$\lambda = -1/2 \to \boldsymbol{u} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \lambda = i/2 \to \boldsymbol{u} = \begin{pmatrix} i \\ 0 \\ 1 \end{pmatrix} \qquad \lambda = -i/2 \to \boldsymbol{u} = \begin{pmatrix} -i \\ 0 \\ 1 \end{pmatrix}$$

So $\lambda = -1/2$ corresponds to ψ_1, and around the stationary point, we'll have damping (stability) in this direction, whereas we'll have oscillations (instability) in directions ψ_0, θ.

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:

$$\begin{array}{ll} \dot{\psi} = \bar{v}_\psi(\psi, \theta) & \bar{v}_\psi(\psi, \theta) \coloneqq \nabla_\psi \mathcal{L}(\psi, \theta) - \frac{\gamma}{2}\nabla_\psi \mathcal{R}_1(\psi) \\ \dot{\theta} = \bar{v}_\theta(\psi, \theta) & \bar{v}_\theta(\psi, \theta) \coloneqq -\nabla_\theta \mathcal{L}(\psi, \theta) \end{array}$$

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]

The velocity field is $\bar{\boldsymbol{v}} = (\bar{\boldsymbol{v}_\psi}, \bar{\boldsymbol{v}_\theta})^\top = (\bar{\boldsymbol{v}_{\psi_0}}, \bar{\boldsymbol{v}_{\psi_1}}, \bar{\boldsymbol{v}_\theta})^\top$. To lighten the notation with derivatives of the sigmoid function, again we define $g(x)$ according to the following:

$$\frac{d}{dx}\sigma(f(x)) = \frac{d}{df(x)}\sigma(f(x))\frac{df(x)}{dx} = \sigma(f(x))(1 - \sigma(f(x)))\frac{df(x)}{dx} = g(f(x))\frac{df(x)}{dx}$$

Now by noting that we can push partials into the expectation, we may compute the velocity field:

$$\bar{\boldsymbol{v}} = \begin{pmatrix} \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_0} - \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_0} \\ \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_1} - \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_1} \\ -\frac{\partial \mathcal{L}(\psi,\theta)}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{p_D}[\delta(C_\psi(x))x/\sigma(C_\psi(x))] + \mathbb{E}_{p_\theta}[-\delta(-C_\psi(x))x/\sigma(-C_\psi(x))] \\ \mathbb{E}_{p_D}[\delta(C_\psi(x))/\sigma(C_\psi(x))] + \mathbb{E}_{p_\theta}[-\delta(-C_\psi(x))/\sigma(-C_\psi(x))] \\ -\nabla_\theta[\mathbb{E}_{p_D}\log(\sigma(C_\psi(x))) + \mathbb{E}_{p_\theta}\log(\sigma(-C_\psi(x)))] \end{pmatrix} - \begin{pmatrix} \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_0} \\ \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_1} \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ 0 - \nabla_\theta[\mathbb{E}_{p_\theta}\log(\sigma(-C_\psi(x)))] \end{pmatrix} - \begin{pmatrix} \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_0} \\ \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_1} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ \mathbb{E}_{p_\theta}[\psi_0\delta(-C_\psi(x))/\sigma(-C_\psi(x))] \end{pmatrix}$$

$$- \begin{pmatrix} \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_0} \\ \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_1} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ \mathbb{E}_{p_\theta}[\psi_0\sigma(C_\psi(x))] \end{pmatrix} - \begin{pmatrix} \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_0} \\ \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_1} \\ 0 \end{pmatrix}$$

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

Now for computing the expectations, we use the same property we used with Dirac function focused at x_0 ($\delta(x - x_0)$) ; If $g_X(x) = \delta(x - x_0)$ is a density function we can write:

$$\int_{-\infty}^{+\infty} f(x)\delta(x - x_0)dx = \mathbb{E}_{x\sim g_X(x)}[f(x)] = f(x_0)$$

Then we can replace the expectations by the function inside them evaluated at $x = 0$ for $\mathbb{E}_{p_D}$, and at $x = \theta$ for $\mathbb{E}_{p_\theta}$. Computing the first term of $\bar{\boldsymbol{v}}$ (just the same as before):

$$\bar{\boldsymbol{v}} = \begin{pmatrix} \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_0} \\ \frac{\partial \mathcal{L}(\psi,\theta)}{\partial \psi_1} \\ -\frac{\partial \mathcal{L}(\psi,\theta)}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \mathbb{E}_{p_D}[x\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[x\sigma(C_\psi(x))] \\ \mathbb{E}_{p_D}[\sigma(-C_\psi(x))] - \mathbb{E}_{p_\theta}[\sigma(C_\psi(x))] \\ \mathbb{E}_{p_\theta}[\psi_0\sigma(C_\psi(x))] \end{pmatrix} = \begin{pmatrix} -\theta\sigma(C_\psi(\theta)) \\ \sigma(-\psi_1) - \sigma(C_\psi(\theta)) \\ \psi_0\sigma(C_\psi(\theta)) \end{pmatrix}$$

Now computing the other term of $\bar{\boldsymbol{v}}$, we need to compute $\nabla_\psi \mathcal{R}_1(\psi)$:

$$\mathcal{R}_1(\psi) = \mathbb{E}_{p_D}\|\nabla_x C_\psi(x)\|^2$$

Since x is 1-D, $\nabla_x C_\psi(x)$ amounts to $\frac{d}{dx}C_\psi(x) = \frac{d}{dx}(\psi_0 x + \psi_1) = \psi_0$. Therefore for $\mathcal{R}_1(\psi)$ we have (noting that the expected value of a constant is itself, and that $\|a\|^2 = a^2$ for scalar a):

$$\mathcal{R}_1(\psi) = \mathbb{E}_{p_D}\|\psi_0\|^2 = \|\psi_0\|^2 = \psi_0^2$$

Now we can compute the second term in $\bar{\boldsymbol{v}}$:

$$-\begin{pmatrix} \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_0} \\ \frac{\gamma}{2}\frac{\partial \mathcal{R}_1(\psi)}{\partial \psi_1} \\ 0 \end{pmatrix} = -\begin{pmatrix} \frac{\gamma}{2}(2\psi_0) \\ 0 \\ 0 \end{pmatrix} = -\begin{pmatrix} \gamma\psi_0 \\ 0 \\ 0 \end{pmatrix}$$

Setting $\bar{\boldsymbol{v}}$ to zero:

$$\bar{\boldsymbol{v}} = \begin{pmatrix} -\theta\sigma(C_\psi(\theta)) - \gamma\psi_0 \\ \sigma(-\psi_1) - \sigma(C_\psi(\theta)) \\ \psi_0\sigma(C_\psi(\theta)) \end{pmatrix} = \boldsymbol{O}$$

All entries in $\boldsymbol{v}$ should be zero, and since sigmoid can't be zero, therefore $\psi_0 = 0$ for $\bar{\boldsymbol{v}}_3 = 0$. Then for $\bar{\boldsymbol{v}}_1$ to be 0, we need $-\theta\sigma(C_\psi(\theta)) - \gamma\psi_0 = -\theta\sigma(C_\psi(\theta)) - 0 = 0$, thus $\theta = 0$. For $\bar{\boldsymbol{v}}_2$ to be 0, we need $\sigma(-\psi_1) = \sigma(C_\psi(\theta))$. For two sigmoids to be equal, their input should be equal (because sigmoid is a bijection), therefore we need $-\psi_1 = C_\psi(\theta) = \psi_0\theta + \psi_1 = \psi_1$ (because we found that $\psi_0 = \theta = 0$), hence $\psi_1 = 0$ as well. Thus again the velocity field has a single stationary point $(\psi_0^*, \psi_1^*, \theta^*) = (0, 0, 0)$.

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

We observed that the only difference between $\boldsymbol{v}, \bar{\boldsymbol{v}}$ was the subtraction of $\gamma\psi_0$ in $\boldsymbol{v}_{\psi_0}$. So all entries of the Jacobian remain the same and we need to recompute only the first row. So we'll reuse most of the computations from 5.2:

$$\bar{J}(\bar{\boldsymbol{v}}) = \begin{pmatrix} \frac{\partial \bar{\boldsymbol{v}}_{\psi_0}}{\partial \psi_0} & \frac{\partial \bar{\boldsymbol{v}}_{\psi_0}}{\partial \psi_1} & \frac{\partial \bar{\boldsymbol{v}}_{\psi_0}}{\partial \theta} \\ \frac{\partial \bar{\boldsymbol{v}}_{\psi_1}}{\partial \psi_0} & \frac{\partial \bar{\boldsymbol{v}}_{\psi_1}}{\partial \psi_1} & \frac{\partial \bar{\boldsymbol{v}}_{\psi_1}}{\partial \theta} \\ \frac{\partial \bar{\boldsymbol{v}}_{\theta}}{\partial \psi_0} & \frac{\partial \bar{\boldsymbol{v}}_{\theta}}{\partial \psi_1} & \frac{\partial \bar{\boldsymbol{v}}_{\theta}}{\partial \theta} \end{pmatrix} = \begin{pmatrix} -\theta^2 g(C_\psi(\theta)) - \frac{\partial \gamma\psi_0}{\partial \psi_0} & -\theta g(C_\psi(\theta)) - \frac{\partial \gamma\psi_0}{\partial \psi_1} & -\sigma(C_\psi(\theta)) - \frac{\partial \gamma\psi_0}{\partial \theta} \\ -\theta g(C_\psi(\theta)) & -g(-\psi_1) - g(C_\psi(\theta)) & -\psi_0 g(C_\psi(\theta)) \\ \sigma(C_\psi(\theta)) + \psi_0\theta g(C_\psi(\theta)) & \psi_0 g(C_\psi(\theta)) & \psi_0^2 g(C_\psi(\theta)) \end{pmatrix}$$

$$= \begin{pmatrix} -\theta^2 g(C_\psi(\theta)) - \gamma & -\theta g(C_\psi(\theta)) - 0 & -\sigma(C_\psi(\theta)) - 0 \\ -\theta g(C_\psi(\theta)) & -g(-\psi_1) - g(C_\psi(\theta)) & -\psi_0 g(C_\psi(\theta)) \\ \sigma(C_\psi(\theta)) + \psi_0\theta g(C_\psi(\theta)) & \psi_0 g(C_\psi(\theta)) & \psi_0^2 g(C_\psi(\theta)) \end{pmatrix}$$

Computing $\bar{J}^* = \bar{J}(\boldsymbol{v}^*)$ by plugging $(\psi_0^*, \psi_1^*, \theta^*) = (0, 0, 0)$ into the Jacobian above:

$$\bar{J}(\boldsymbol{v}^*) = \begin{pmatrix} -\theta^2 g(C_\psi(\theta)) - \gamma & -\theta g(C_\psi(\theta)) & -\sigma(C_\psi(\theta)) \\ -\theta g(C_\psi(\theta)) & -g(-\psi_1) - g(C_\psi(\theta)) & -\psi_0 g(C_\psi(\theta)) \\ \sigma(C_\psi(\theta)) + \psi_0 \theta g(C_\psi(\theta)) & \psi_0 g(C_\psi(\theta)) & \psi_0^2 g(C_\psi(\theta)) \end{pmatrix}_{|\boldsymbol{v}=\boldsymbol{v}*}$$

$$= \begin{pmatrix} -\gamma & 0 & -\sigma(C_\psi(0)) \\ 0 & -g(-\psi_1^*) - g(C_\psi(0)) & 0 \\ \sigma(C_\psi(0)) & 0 & 0 \end{pmatrix} = \begin{pmatrix} -\gamma & 0 & -\sigma(\psi_1^*) \\ 0 & -g(-\psi_1^*) - g(\psi_1^*) & 0 \\ \sigma(\psi_1^*) & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} -\gamma & 0 & -\sigma(0) \\ 0 & -g(0) - g(0) & 0 \\ \sigma(0) & 0 & 0 \end{pmatrix} = \begin{pmatrix} -\gamma & 0 & -1/2 \\ 0 & -1/4 - 1/4 & 0 \\ 1/2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} -\gamma & 0 & -1/2 \\ 0 & -1/2 & 0 \\ 1/2 & 0 & 0 \end{pmatrix}$$

$$\implies \bar{J}^* = \begin{pmatrix} -\gamma & 0 & -1/2 \\ 0 & -1/2 & 0 \\ 1/2 & 0 & 0 \end{pmatrix}$$

So the only difference is the addition of $-\gamma$ to J_{11}^* compared to $\bar{J}^*$, other elements are the same. But we'll see that this difference has a significant impact on stability.

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

Following the definition of eigenvalues, λ is an eigenvalue of $\bar{J}^*$ if there exists some $\boldsymbol{u}$ for which we have $\bar{J}^*\boldsymbol{u} = \lambda\boldsymbol{u}$. This is equivalent to $(\bar{J}^* - \lambda\boldsymbol{I})\boldsymbol{u} = \boldsymbol{O}$. In order to have a non-trivial solution for $\boldsymbol{u}$, $(\bar{J}^* - \lambda\boldsymbol{I})$ should *not* be invertible or equivalently, it should be the case that $|\det(\bar{J}^* - \lambda\boldsymbol{I})| = 0$. Under such case, $\boldsymbol{u}$ will have a non-trivial solution. Now we'll compute the $|\det(\bar{J}^* - \lambda\boldsymbol{I})|$, set it to zero and find the eigenvalues and eigenvectors:

$$\det(\bar{J}^* - \lambda\boldsymbol{I}) = \det \begin{pmatrix} -\gamma - \lambda & 0 & -1/2 \\ 0 & -1/2 - \lambda & 0 \\ 1/2 & 0 & -\lambda \end{pmatrix} = -\lambda(\lambda + 1/2)(\lambda + \gamma) - 1/4(1/2 + \lambda)$$

$$= -(\lambda + 1/2)(\lambda^2 + \gamma\lambda + 1/4) = -(\lambda + 1/2)(\lambda - \frac{-\gamma + \sqrt{\gamma^2 - 4\frac{1}{4}}}{2})(\lambda - \frac{-\gamma - \sqrt{\gamma^2 - 4\frac{1}{4}}}{2})$$

$$= -(\lambda + 1/2)(\lambda - \frac{-\gamma + \sqrt{\gamma^2 - 1}}{2})(\lambda - \frac{-\gamma - \sqrt{\gamma^2 - 1}}{2})$$

So eigenvalues would be $\{-\frac{1}{2}, \frac{1}{2}(-\gamma + \sqrt{\gamma^2 - 1}), \frac{1}{2}(-\gamma - \sqrt{\gamma^2 - 1})\}$. We know that $\gamma > 0$, if $0 < \gamma < 1$, then $\sqrt{\gamma^2 - 1} = i\sqrt{1 - \gamma^2}$ (where i denotes the complex values), and thus the real part of all eigenvalues would be negative $\text{Re}(\lambda) = \{-\frac{1}{2}, -\frac{1}{2}\gamma, -\frac{1}{2}\gamma\}$. If $\gamma \geq 1$, then $\sqrt{\gamma^2 - 1} \geq 0$, and thus the third eigenvalue is real and negative $\frac{1}{2}(-\gamma - \sqrt{\gamma^2 - 1}) < 0$ as its the sum of two negative real numbers. For the second eigenvalue $\frac{1}{2}(-\gamma + \sqrt{\gamma^2 - 1})$, note that $\gamma^2 > \gamma^2 - 1$, thus since $\gamma > 0$, taking the square root we get $\gamma > \sqrt{\gamma^2 - 1}$, and finally $-\gamma + \sqrt{\gamma^2 - 1} < 0$, hence this eigenvalue will also be real and negative. Now because all eigenvalues have negative real parts for any $\gamma > 0$, hence the dynamical system around the stationary point is stable in all directions. It's not asked here but after carrying out the computations for eigenvectors, they're

as follows:

$$\lambda = -1/2 \rightarrow \boldsymbol{u} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \lambda = i/2 \rightarrow \boldsymbol{u} = \begin{pmatrix} \frac{1}{\sqrt{\gamma^2-1}-\gamma} \\ 0 \\ 1 \end{pmatrix} \qquad \lambda = -i/2 \rightarrow \boldsymbol{u} = \begin{pmatrix} -\frac{1}{\sqrt{\gamma^2-1}+\gamma} \\ 0 \\ 1 \end{pmatrix}$$

So $\lambda = -1/2$ corresponds to ψ_1, and around the stationary point, we'll have damping (stability) in this direction regardless of the value of γ, and we'll have damping oscillations (stability) in directions ψ_0, θ if $0 < \gamma < 1$, and if $\gamma > 1$ there won't be any oscillation and it'll exponentially damp. We observe that adding the gradient penalty significantly improves the stability problem around the stationary point of this simple setup. Probably exploiting this intuition in more complex settings will yield similar results and would have stabilizing effect on training GANs. That's why gradient penalty has become popular.

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