Week 2

Problem 1: Expectation and variance

For the following distributions write down their density function and their support. Express $\mathbb{E}[X]$ and $\mathrm{Var}[X]$ as integrals and evaluate them.

a) $X \sim \mathrm{Uniform}[-3,5]$ *b)* $X \sim \mathcal{N}(m, s^2)$

Part (a):

$$\begin{aligned} p(x) &= \frac{1}{8}\mathbb{1}_{[-3,5]}(x), \\ \mathrm{support}[p] &= [-3,5], \\ \mathbb{E}[X] &= \int_{-3}^{5} \frac{x}{8}\,\mathrm{d}x = 1, \\ \mathrm{Var}[X] &= \int_{-3}^{5} \frac{(x-1)^2}{8}\,\mathrm{d}x = \frac{64}{12}. \end{aligned}$$

Both integrals are elementary to compute.

Part (b):

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right), \\ \mathrm{support}[p] &= \mathbb{R}, \\ \mathbb{E}[X] &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi s^2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)\mathrm{d}x = m, \\ \mathrm{Var}[X] &= \int_{-\infty}^{\infty} \frac{(x-m)^2}{\sqrt{2\pi s^2}} \exp\left(-\frac{(x-m)^2}{2s^2}\right)\mathrm{d}x = s^2. \end{aligned}$$

For these two integrals, first make a change of variables $y = \frac{x-m}{s}$, $\mathrm{d}y = \frac{\mathrm{d}x}{s}$. Together with the fact that $\int_{-\infty}^{\infty} p(x)\,dx = 1$, this allows to reduce to the integrals $\int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\mathrm{d}y$ and $\int_{-\infty}^{\infty} \frac{y^2}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\mathrm{d}y$, respectively. The first integral is zero by symmetry, and the second can be computed using integration by parts, giving the result of 1.

Problem 2: Gradients and score function trick

Given a family of densities $p_\theta(x)$, depending on a parameter θ, and functions $f(x,\varphi)$ and $g(x,\theta)$:

1. Write as an expectation over p_θ: $\frac{\partial}{\partial\varphi}\mathbb{E}_{x\sim p_\theta(x)}[f(x,\varphi)] = \mathbb{E}_{x\sim p_\theta(x)}[\quad\quad]$.
2. Show that $\frac{\partial}{\partial\theta}\mathbb{E}_{x\sim p(x,\theta)}[f(x,\varphi)] = \mathbb{E}_{x\sim p_\theta(x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x) f(x,\varphi)\right]$ (hint: you will need the fact that $\frac{\partial}{\partial\theta}\log h(\theta) = \frac{1}{h(\theta)}\frac{\partial}{\partial\theta}h(\theta)$). This is known as the **score function trick**.
3. Write as an expectation over p_θ: $\frac{\partial}{\partial\theta}\mathbb{E}_{x\sim p_\theta(x)}[g(x,\theta)] = \mathbb{E}_{x\sim p_\theta(x)}[\quad\quad]$.

How can each of these gradients be approximated using the Monte Carlo method?

1. $\frac{\partial}{\partial\varphi}\mathbb{E}_{x\sim p_\theta(x)}[f(x,\varphi)] = \mathbb{E}_{x\sim p_\theta(x)}\left[\frac{\partial}{\partial\varphi}f(x,\varphi)\right]$. This can be approximated by sampling $x_1,\ldots,x_n$ from p_θ and computing $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\varphi}f(x_i,\varphi)$.

2. Derivation:

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathbb{E}_{x\sim p(x,\theta)}\left[f(x,\varphi)\right] &= \frac{\partial}{\partial\theta}\int p_\theta(x)f(x,\varphi)\,\mathrm{d}x \\ &= \int \frac{\partial}{\partial\theta}p_\theta(x)f(x,\varphi)\,\mathrm{d}x \\ &= \int p_\theta(x)\frac{\partial}{\partial\theta}\log p_\theta(x)f(x,\varphi)\,\mathrm{d}x \\ &= \mathbb{E}_{x\sim p_\theta(x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(x)f(x,\varphi)\right].\end{aligned}$$

This can be approximated by sampling $x_1,\ldots,x_n$ from p_θ and computing $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log p_\theta(x_i)f(x_i,\varphi)$.

3. $\frac{\partial}{\partial\theta}\mathbb{E}_{x\sim p_\theta(x)}\left[g(x,\theta)\right] = \mathbb{E}_{x\sim p_\theta(x)}\left[\frac{\partial}{\partial\theta}g(x,\theta) + \frac{\partial}{\partial\theta}\log p_\theta(x)g(x,\theta)\right]$ (product rule, combining the previous two parts). This can be approximated by sampling $x_1,\ldots,x_n$ from p_θ and computing $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}g(x_i,\theta) + \frac{\partial}{\partial\theta}\log p_\theta(x_i)g(x_i,\theta)$.

Problem 3: Reparametrisation trick

1. If $\epsilon\sim\mathcal{N}(0,1)$, what is the distribution of $X = a + b\epsilon$, where $a,b\in\mathbb{R}$?
2. Use Part 1 to rewrite as an expectation over $\mathcal{N}(\mu,\sigma^2)$ as an expectation over $\mathcal{N}(0,1)$:

$$\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[f(X)\right] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\qquad\qquad\right].$$

3. Rewrite as an expectation over $\mathcal{N}(0,1)$: $\frac{\partial}{\partial\mu}\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[f(X)\right] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\qquad\qquad\right]$. This is known as the **reparametrisation trick**.
4. Use Problem 2.2 to rewrite the expression in Part 3 as an expectation over $\mathcal{N}(\mu,\sigma^2)$ instead, then rewrite it as an expectation over $\mathcal{N}(0,1)$ using Part 2:

$$\frac{\partial}{\partial\mu}\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[f(X)\right] = \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[\qquad\qquad\qquad\qquad\right] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\qquad\qquad\qquad\qquad\right]$$

Which expression – this one or the one in Part 3 – is better suited for Monte Carlo estimation?

1. X is distributed according to $\mathcal{N}(a,b^2)$.
2. $\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[f(X)\right] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[f(\mu+\sigma\epsilon)\right]$.
3. $\frac{\partial}{\partial\mu}\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[f(X)\right] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\frac{\partial}{\partial\mu}f(\mu+\sigma\epsilon)\right]$.
4. Using the score function trick,

$$\begin{aligned}\frac{\partial}{\partial\mu}\mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[f(X)\right] &= \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[\frac{\partial}{\partial\mu}\log p(X)f(X)\right] \\ &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\left(\frac{\partial}{\partial\mu}\log p(\mu+\sigma\epsilon)\right)f(\mu+\sigma\epsilon)\right] \\ &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\frac{\epsilon}{\sigma}f(\mu+\sigma\epsilon)\right]\end{aligned}$$

The expression in Part 3 often has lower variance when p is an approximate posterior. However, the expression in Part 4 does not require differentiating f and could be more efficient.

Problem 4: KL divergence

Recall that the KL divergence between two distributions over $\mathbb{R}^d$ with densities p and q is defined as follows: $\mathrm{KL}\,(p \,\|\, q) = \int_{\mathbb{R}^d} \log\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x$.

1. Compute the divergence between two Gaussian distributions, $\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma_1^2) \,\|\, \mathcal{N}(\mu_2, \sigma_2^2)\right)$.
2. Show that for three full-support distributions with densities p_1, p_2, q,

$$\mathrm{KL}\left(\frac{p_1 + p_2}{2} \,\|\, q\right) \leq \frac{1}{2}\mathrm{KL}(p_1 \,\|\, q) + \frac{1}{2}\mathrm{KL}(p_2 \,\|\, q),$$
$$\mathrm{KL}\left(q \,\|\, \frac{p_1 + p_2}{2}\right) \leq \frac{1}{2}\mathrm{KL}(q \,\|\, p_1) + \frac{1}{2}\mathrm{KL}(q \,\|\, p_2).$$

This property is called **convexity** of the KL divergence.

1. We write the KL as an integral and evaluate:

$$\begin{aligned}
&\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma_1^2) \,\|\, \mathcal{N}(\mu_2, \sigma_2^2)\right) \\
&= \int_{-\infty}^{\infty} \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)}\right) \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \left(\log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{((x-\mu_1)-(\mu_2-\mu_1))^2}{2\sigma_2^2}\right) \frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)\mathrm{d}x \\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{\mathrm{Var}\left[\mathcal{N}(\mu_1, \sigma_1^2)\right]}{2\sigma_1^2} + \frac{\mathrm{Var}\left[\mathcal{N}(\mu_1, \sigma_1^2)\right] + (\mu_2 - \mu_1)^2}{2\sigma_2^2} \\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.
\end{aligned}$$

2. One way to show this is to introduce a parameter α, set $p_\alpha = \alpha p_1 + (1-\alpha)p_2$. We will show that $\mathrm{KL}(p_\alpha \,\|\, q)$ and $\mathrm{KL}(q \,\|\, p_\alpha)$ are convex functions of α, which implies the desired inequalities by setting $\alpha = \frac{1}{2}$.

   To show convexity, we show the second derivative with respect to α is non-negative. For the first KL, we have

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\mathrm{KL}(p_\alpha \,\|\, q) &= \frac{\partial}{\partial\alpha}\int p_\alpha(x)\log\left(\frac{p_\alpha(x)}{q(x)}\right)\mathrm{d}x \\
&= \int (p_1(x) - p_2(x))\log\left(\frac{p_\alpha(x)}{q(x)}\right) + (p_1(x) - p_2(x))\,\mathrm{d}x \\
&= \int (p_1(x) - p_2(x))\log\left(\frac{p_\alpha(x)}{q(x)}\right)\mathrm{d}x,
\end{aligned}$$

   and

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\frac{\partial}{\partial\alpha}\mathrm{KL}(p_\alpha \,\|\, q) &= \frac{\partial}{\partial\alpha}\int (p_1(x) - p_2(x))\log\left(\frac{p_\alpha(x)}{q(x)}\right)\mathrm{d}x \\
&= \int \frac{(p_1(x) - p_2(x))^2}{p_\alpha(x)}\,\mathrm{d}x \\
&\geq 0.
\end{aligned}$$

For the second KL,

$$\begin{aligned}\frac{\partial}{\partial\alpha}\mathrm{KL}(q \,\|\, p_\alpha) &= \frac{\partial}{\partial\alpha}\int q(x)\log\left(\frac{q(x)}{p_\alpha(x)}\right)\mathrm{d}x \\ &= -\int q(x)\frac{p_1(x)-p_2(x)}{p_\alpha(x)}\,\mathrm{d}x,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\alpha}\frac{\partial}{\partial\alpha}\mathrm{KL}(q \,\|\, p_\alpha) &= \frac{\partial}{\partial\alpha}\int -q(x)\frac{p_1(x)-p_2(x)}{p_\alpha(x)}\,\mathrm{d}x \\ &= \int q(x)\frac{(p_1(x)-p_2(x))^2}{p_\alpha(x)^2}\,\mathrm{d}x \\ &\geq 0.\end{aligned}$$

Week 3

Problem 1: Importance-weighted ELBO

In lecture, we briefly mentioned that instead of the ELBO

$$\log p_\theta(x) \geq \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)}\right]$$

one can use a tighter bound on the likelihood in a VAE (or any latent variable model):

$$\log p_\theta(x) \geq \mathbb{E}_{z_1,\ldots,z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x \mid z_k)p(z_k)}{q_\phi(z_k \mid x)}\right].$$

1. Show that as $K \to \infty$, the new bound converges to $\log p_\theta(x)$, so the IW bound can be made arbitrarily tight (hint: what does the $\frac{1}{K}\sum_{k=1}^{K}$ turn into as $K \to \infty$?).
2. (Harder:) Show that the new bound is tighter than the ELBO, i.e., that for all $K \geq 1$,

$$\mathbb{E}_{z_1,\ldots,z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x \mid z_k)p(z_k)}{q_\phi(z_k \mid x)}\right] \geq \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)}\right]$$

(hint: this uses the fact that for any $c_1,\ldots,c_k$, we have $\log\left(\frac{1}{K}\sum_{k=1}^{K} c_k\right) \geq \frac{1}{K}\sum_{k=1}^{K}\log c_k$), which is a form of Gibbs' inequality).

1. Using the weak convergence of the Monte Carlo estimator and the continuous mapping theorem, the estimate approaches

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x \mid z_k)p(z_k)}{q_\phi(z_k \mid x)}\right] &\xrightarrow{K\to\infty} \log \mathbb{E}_{z\sim q_\phi(z|x)} \frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)} \\ &= \int_z p_\theta(x \mid z)p(z)\,\mathrm{d}z \\ &= p_\theta(x).\end{aligned}$$

2. Using the hint:

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x \mid z_k)p(z_k)}{q_\phi(z_k \mid x)}\right] &\geq \mathbb{E}_{z_1,\ldots,z_K\sim q_\phi(z|x)}\left[\frac{1}{K}\sum_{k=1}^{K}\log \frac{p_\theta(x \mid z_k)p(z_k)}{q_\phi(z_k \mid x)}\right] \\ &= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x \mid z)p(z)}{q_\phi(z \mid x)}\right],\end{aligned}$$

where the equality expresses the unbiasedness of the Monte Carlo estimator.

Problem 2: Denoising autoencoder

This problem considers VAEs in which the data and latent spaces are of the same dimension d, which is connected to diffusion models (lectures 8-10).

1. (a) Suppose that the encoder $q(z \mid x)$ in a VAE is fixed to be Gaussian: $q(z \mid x) = \mathcal{N}(z; x, \sigma^2 I_d)$, where $\sigma > 0$ is a constant. What is the VAE objective for the parameters of the decoder $p_\theta(x \mid z)$ in this case? Assume the decoder is Gaussian with fixed variance: $p_\theta(x \mid z) = \mathcal{N}(x; \mu_\theta(z), \tau^2 I_d)$.

   (b) Suppose we optimise the VAE over $x \sim \pi_{\text{data}}$. Show that the optimal $\mu_\theta(z)$ equals the conditional expectation $\mathbb{E}[X \mid Z = z]$, where (X, Z) is distributed according to $\pi_{\text{data}}(x)q(z \mid x)$ (hint: use the fact that optimising $\mathbb{E}_{Y \sim p(Y)}[\|Y - c\|^2]$ over c gives $c^* = \mathbb{E}[Y]$).

2. Now consider instead a decoder $p(x \mid z)$ that is fixed to be Gaussian ($p(x \mid z) = \mathcal{N}(x; z, \tau^2 I_d)$). Show that the true posterior $p(z \mid x) \propto p(z)p_\theta(x \mid z)$ is also Gaussian and give its parameters (hint: the product of two Gaussian densities is proportional to a Gaussian density). What is the optimal encoder $q_\phi(z \mid x)$ in this case?

1. (a) The VAE objective in this case is:

$$\begin{aligned}
&\mathbb{E}_{x \sim \pi_{\text{data}}} \left[\mathbb{E}_{z \sim q(z|x)} \left[\log p_\theta(x \mid z) \right] - \mathrm{KL}(q(z \mid x) \,\|\, p(z)) \right] \\
&= \mathbb{E}_{x \sim \pi_{\text{data}}} \left[-\frac{d}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \mathbb{E}_{z \sim q(z|x)} [\|x - \mu_\theta(z)\|^2] - \mathrm{KL}(q(z \mid x) \,\|\, p(z)) \right],
\end{aligned}$$

   which, treating the terms that do not depend on θ as constant, is proportional simply to

$$\mathbb{E}_{x \sim \pi_{\text{data}}} \mathbb{E}_{z \sim q(z|x)} [\|x - \mu_\theta(z)\|^2].$$

   (b) The optimal $\mu_\theta(z)$ is the one that minimises $\mathbb{E}_{x \sim \pi_{\text{data}}} \mathbb{E}_{z \sim q(z|x)} [\|x - \mu_\theta(z)\|^2]$. For a fixed z, by the hint, this recovers the mean of the conditional of $\pi_{\text{data}}(x)q(z \mid x)$ on z.

2. The true posterior is given by

$$\begin{aligned}
p(z \mid x) &\propto p(z)p_\theta(x \mid z) \\
&\propto \exp\left(-\frac{\|z\|^2}{2} \right) \exp\left(-\frac{\|x - z\|^2}{2\tau^2} \right) \\
&\propto \exp\left(-\frac{1 + \tau^2}{2\tau^2} \left\| z - \frac{x}{1 + \tau^2} \right\|^2 \right),
\end{aligned}$$

   where the proportionality treats x as constant, so $p(z \mid x) = \mathcal{N}\left(z; \frac{x}{1+\tau^2}, \frac{\tau^2}{1+\tau^2} I_d\right)$. The optimal encoder is $q_\phi(z \mid x) = p(z \mid x)$. Notice that as $\tau \to 0$, this approaches δ_x, while as $\tau \to +\infty$, it approaches $\mathcal{N}(0, I_d)$, which is consistent with what we should expect.

Week 4

Problem 1: Simple pushforward

The Jacobian matrix $\mathrm{J}_f(x)$ of a map $f : \mathbb{R}^2 \to \mathbb{R}^2$ at a point x is defined as the 2×2 matrix $\left(\mathrm{J}_f(x)\right)_{i,j} = \frac{\partial f_i}{\partial x_j}$. For the given distributions $\mathbb{P}$ and $\mathbb{Q}$, find a map f and its inverse f^{-1} such that $\mathbb{Q} = f_{\#}\mathbb{P}$. Additionally, compute $\mathrm{J}_f(x)$ and $\det \mathrm{J}_f(x)$.

1. **Shift:** $\mathbb{P}$ is $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$, $\mathbb{Q}$ is $\mathcal{N}\left(\begin{bmatrix}1\\-2\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$;
2. **Scaling:** $\mathbb{P}$ is $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$, $\mathbb{Q}$ is $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}3 & 0\\0 & 0.5\end{bmatrix}\right)$;
3. **Rotation:** $\mathbb{P}$ is $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}3 & 0\\0 & 0.5\end{bmatrix}\right)$, $\mathbb{Q}$ is $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1.75 & 1.25\\1.25 & 1.75\end{bmatrix}\right)$;

(For this problem, you may wish to review how multivariate Gaussians transform under linear changes of variables.)

1. $f(x) = x + \begin{bmatrix}1\\-2\end{bmatrix}, f^{-1}(y) = y - \begin{bmatrix}1\\-2\end{bmatrix}, \mathrm{J}_f(x) = I_2, \det \mathrm{J}_f(x) = 1.$
2. $f(x) = \begin{bmatrix}\sqrt{3} & 0\\0 & \sqrt{0.5}\end{bmatrix} x, f^{-1}(y) = \begin{bmatrix}\frac{1}{\sqrt{3}} & 0\\0 & \sqrt{2}\end{bmatrix} y, \mathrm{J}_f(x) = \begin{bmatrix}\sqrt{3} & 0\\0 & \sqrt{0.5}\end{bmatrix}, \det \mathrm{J}_f(x) = \sqrt{1.5}.$
3. $f(x) = Ax, f^{-1}(y) = A^{-1}y, \mathrm{J}_f(x) = A, \det \mathrm{J}_f(x) = \det A = 1$, where A is a $\frac{\pi}{4}$ rotation matrix (among other choices):
$$A = \begin{bmatrix}\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2}\\\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\end{bmatrix}.$$

Problem 2: Infinitesimal change of variables

This problem considers a map $f : \mathbb{R}^d \to \mathbb{R}^d$ of the form $f(x) = x + u(x)\Delta$, where $u : \mathbb{R}^d \to \mathbb{R}^d$ is a vector field and $\Delta > 0$. This map represents transporting points for a time Δ along the direction given by the vector field. Such maps are important in dynamics-based generative models, which will be discussed at the end of the course.

1. Convince yourself **informally** that for smooth enough u and small enough Δ, f is invertible.
2. Show using the variable change formula that the following identity holds for a distribution with density p and a diffeomorphism $f : \mathbb{R}^d \to \mathbb{R}^d$:
$$\log(f_{\#}p)(f(x)) - \log p(x) = -\log\det \mathrm{J}_f(x).$$
3. Define $p_\Delta := f_{\#}p$, that is, the distribution that sampled by drawing samples x from p and transporting them for time Δ in the direction $u(x)$.
   (a) Compute $\mathrm{J}_f(x)$ in terms of $\mathrm{J}_u(x)$.
   (b) Show that $\log\det \mathrm{J}_f(x) = \mathrm{Tr}\left[\mathrm{J}_u(x)\right]\Delta + O(\Delta^2)$.
   (For this problem, you will need to have some knowledge of the matrix exponential. If this is unfamiliar to you, solve this exercise *assuming that $J_u(x)$ is a diagonal matrix*, using the Taylor expansion $\log(1+t) = t + O(t^2)$. Then read about the matrix exponential and use the fact that $\log\det A = \mathrm{Tr}\left[\log A\right]$ to solve the problem in the general case.)
   (c) Combine Part 2 of this problem with Parts 3(a,b) to show the following identity:
$$\log p_\Delta(f(x)) - \log p(x) = -\mathrm{Tr}\left[\mathrm{J}_u(x)\right]\Delta + O(\Delta^2).$$

Observe that $\mathrm{Tr}\left[\mathrm{J}_u(x)\right]$ is the divergence of the vector field u at x. We have thus related the change of density when transporting points along u to the divergence of u, commonly given as a physical interpretation for the divergence and an intuition for the divergence theorem in vector calculus.

1. If u is L-Lipschitz, then for $\Delta < \frac{1}{L}$, f is invertible, since
$$\|f(x) - f(y)\| = \|x - y + (u(x) - u(y))\Delta\| \geq \|x - y\| - \|(u(x) - u(y))\Delta\| \geq (1 - L\Delta)\|x - y\|,$$
which is positive if $x \neq y$.
2. This is simply the logarithm of the formula $(f_{\#}p)(f(x)) = p(f(x))/\det \mathrm{J}_f(x)$.
3. (a) $\mathrm{J}_f(x) = I + \mathrm{J}_u(x)\Delta$.
   (b) Using the Taylor expansion of the logarithm, $\log\det \mathrm{J}_f(x) = \log\det(I + \mathrm{J}_u(x)\Delta) = \mathrm{Tr}\left[\mathrm{J}_u(x)\right]\Delta + O(\Delta^2)$. Notice that convergence relies on positivity of this determinant, which holds by Part 1 of this problem for small enough Δ.
   (c) This follows directly from the previous parts by substituting.

Week 5

Problem 1: Optimal GAN discriminator

Assume that the generator is defined as $G_\theta : \mathbb{R}^{d_{\text{latent}}} \to \mathbb{R}^{d_{\text{data}}}$, and the discriminator $D : \mathbb{R}^{d_{\text{data}}} \to [0,1]$ outputs a probability $p(\text{real} \mid x)$. Let p_{data} be the density of the real data distribution and $p_\theta = (G_\theta)_{\#} p_{\text{latent}}$ be the density of the distribution of generated samples.
Show that for a fixed generator G_θ the optimal discriminator D^* for the following optimisation problem

$$\max_D \left\{ \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_{\text{latent}}} [\log(1 - D(G_\theta(z)))] \right\}$$

has the following form: $D^*(\text{real} \mid x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)}$ (hint: for $a, b > 0$ find the maximiser of $a \log(x) + b \log(1-x)$ with respect to x, then use this result to find the optimal D^*).

First, in the hint, the optimiser is shown by basic calculus to be $\frac{a}{a+b}$.
Now, we rewrite the objective as a single expectation over x:

$$2\mathbb{E}_{x \sim \frac{p_{\text{data}} + p_\theta}{2}} \left[\frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)} \log D(x) + \frac{p_\theta(x)}{p_{\text{data}}(x) + p_\theta(x)} \log(1 - D(x)) \right].$$

Now apply the hint with $a \leftarrow \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)}$, $b \leftarrow \frac{p_\theta(x)}{p_{\text{data}}(x) + p_\theta(x)}$, and $x \leftarrow D(x)$ to get the optimal discriminator $D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)}$.

Problem 2: Alternative GAN losses

1. Alternatively, losses for training discriminator and generator can be written as follows:

$$\mathcal{L}_{\text{MSE}}(\varphi) = \mathbb{E}_{x \sim p_{\text{data}}} \left[(1 - D_\varphi(x))^2 \right] + \mathbb{E}_{z \sim p_{\text{latent}}} \left[(D_\varphi(G_\theta(z)))^2 \right],$$
$$\mathcal{L}_{\text{MSE}}(\theta) = \mathbb{E}_{z \sim p_{\text{latent}}} \left[(1 - D_\varphi(G_\theta(z)))^2 \right],$$

   which are minimised with respect to φ and θ respectively. This contrasts with the vanilla objective

$$\mathcal{L}_{\text{vanilla}}(\theta, \varphi) = \mathbb{E}_{x \sim p_{\text{data}}} \left[\log D_\varphi(x) \right] + \mathbb{E}_{z \sim p_{\text{latent}}} \left[\log(1 - D_\varphi(G_\theta(z))) \right]$$

   Show that the gradients $\nabla_\varphi \mathcal{L}_{\text{MSE}}(\varphi)$ and $\nabla_\theta \mathcal{L}_{\text{MSE}}(\theta)$ are proportional to $\nabla_\varphi \mathcal{L}_{\text{vanilla}}(\varphi, \theta)$ and $\nabla_\theta \mathcal{L}_{\text{vanilla}}(\varphi, \theta)$ respectively.

2. In order to improve the training stability one can use 'non-saturating' losses for training the generator: $-\log D(G_\theta(z))$ in place of $\log(1 - D(G_\theta(z)))$:

$$\mathcal{L}_{\text{non-sat}}(\theta) = \mathbb{E}_{z \sim p_{\text{latent}}} \left[-\log D(G_\theta(z)) \right].$$

   Are the non-saturating GAN gradients proportional to the vanilla ones?

1. The problem is to be understood as: **for a given** x, the expected gradients of the MSE losses are proportional to the expected gradients of the vanilla losses.

   Rewriting the objectives as expectations over $x \sim \frac{p_{\text{data}} + p_\theta}{2}$ as in the previous problem, we have

$$\mathcal{L}_{\text{vanilla}}(\theta, \varphi) \propto \mathbb{E}_{x \sim \frac{p_{\text{data}} + p_\theta}{2}} \left[\frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)} \log D_\varphi(x) + \frac{p_\theta(x)}{p_{\text{data}}(x) + p_\theta(x)} \log(1 - D_\varphi(x)) \right],$$
$$\mathcal{L}_{\text{MSE}}(\theta, \varphi) \propto \mathbb{E}_{x \sim \frac{p_{\text{data}} + p_\theta}{2}} \left[\frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_\theta(x)} (1 - D_\varphi(x))^2 + \frac{p_\theta(x)}{p_{\text{data}}(x) + p_\theta(x)} (D_\varphi(x))^2 \right].$$

Let us know compute the gradient $\nabla_\varphi \mathcal{L}_{\text{vanilla}}(\theta, \varphi)$ and $\nabla_\varphi \mathcal{L}_{\text{MSE}}(\theta, \varphi)$ for a given x. For simplicity let's set $A = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_\theta(x)}$, we also note that the expections in the following equations is taken with respect to $\frac{p_{\text{data}}(x)+p_\theta(x)}{2}$:

$$\begin{aligned}
\nabla_\varphi \mathcal{L}_{\text{vanilla}}(\theta, \varphi) &= \nabla_\varphi \mathbb{E}\left[A \log D_\varphi(x) + (1-A)\log(1 - D_\varphi(x))\right] \\
&= \mathbb{E}\left[\left(\frac{A}{D_\varphi(x)} - \frac{1-A}{1-D_\varphi(x)}\right)\nabla_\varphi D_\varphi(x)\right] \\
&= \mathbb{E}\left[\frac{D_\varphi(x) - A}{D_\varphi(x)(D_\varphi(x)-1)}\nabla_\varphi D_\varphi(x)\right], \\
\nabla_\varphi \mathcal{L}_{\text{MSE}}(\theta, \varphi) &= \nabla_\varphi \mathbb{E}\left[A(1 - D_\varphi(x))^2 + (1-A)(D_\varphi(x))^2\right] \\
&= \mathbb{E}\left[2\Big(-A(1 - D_\varphi(x)) + (1-A)D_\varphi(x)\Big)\nabla_\varphi D_\varphi(x)\right] \\
&= \mathbb{E}\left[2(D_\varphi(x) - A)\nabla_\varphi D_\varphi(x)\right].
\end{aligned}$$

One key observation for that exercise is that in $\nabla_\varphi \mathcal{L}_{\text{vanilla}}$ the coefficient $\frac{1}{D_\varphi(x)(D_\varphi(x)-1)}$ can lead to the vanishing gradients, whereas there is not such issuse in $\nabla_\varphi \mathcal{L}_{\text{MSE}}$.

The gradients with respect to θ can be derived in the same manner.

2. Let us compute the gradiensts with respect to θ for $\mathcal{L}_{\text{vanilla}}$ and $\mathcal{L}_{\text{non-sat}}$:

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\text{vanilla}}(\theta, \varphi) &= \nabla_\theta \mathbb{E}_{z\sim p_{\text{latent}}}\left[\log(1 - D_\varphi(G_\theta(z)))\right] \\
&= \mathbb{E}_{z\sim p_{\text{latent}}}\left[\frac{1}{1 - D_\varphi(G_\theta(z))}\nabla_\theta D_\varphi(G_\theta(z))\right] \\
\nabla_\theta \mathcal{L}_{\text{non-sat}}(\theta, \varphi) &= \nabla_\theta \mathbb{E}_{z\sim p_{\text{latent}}}\left[-\log D_\varphi(G_\theta(z))\right] \\
&= \mathbb{E}_{z\sim p_{\text{latent}}}\left[-\frac{1}{D_\varphi(G_\theta(z))}\nabla_\theta D_\varphi(G_\theta(z))\right]
\end{aligned}$$

Notice how for the $\nabla_\theta \mathcal{L}_{\text{vanilla}}$ the coefficient $\frac{1}{1-D_\varphi(G_\theta(z))}$ can lead to the gradients exploding when $D_\varphi(G_\theta(z)) \approx 1$.

Problem 3: Wasserstein GAN

Read **Wasserstein GAN paper**. Let $\mathbb{P}_r$ represent the distribution of real data and $\mathbb{P}_g$ the distribution of generated data. Answer the following:

1. Explain (informally) how using $\text{KL}(\mathbb{P}_r \,\|\, \mathbb{P}_g)$ as a measure of distance between two distributions differs from $\mathbb{W}_1(\mathbb{P}_r, \mathbb{P}_g)$ and why the latter is more sensible.

2. (**Hard**) Explain (informally) why:

$$\inf_{\pi\in\Pi(\mathbb{P}_r,\mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\pi}\left[\|x - y\|\right] = \sup_{\|f\|_L\leq 1}\left\{\mathbb{E}_{x\sim\mathbb{P}_r}\left[f(x)\right] - \mathbb{E}_{y\sim\mathbb{P}_g}\left[f(y)\right]\right\}$$

   Answer the following questions:

   (a) Give the definition of a 1-Lipschitz function and explain (informally) why f should be 1-Lipschitz.

   (b) If f is parameterised by a neural network, what are the possible ways to make f 1-Lipschitz?

1. The minimisation of KL divergence is problematic when the distributions have non-intersecting supports. In such cases, KL divergence becomes infinite, preventing us from using it to meaningfully define the convergence of one distribution toward the other. The problem of non-intersecting supports is particularly acute, as many data distributions are supported on low-dimensional compact manifolds. The more detailed answer is provided in Section 2 of **Wasserstein GAN paper**. For a more in-depth analysis of the manifold hypothesis in relation to generative modelling please refer to **this paper**.

2. OT is a highly non-trivial topic. Some intuilitve explanations for the problem can be found in the blog posts **[1]** and **[2]**. For the rigorous treatment of the topic please refer to the books by F. Santambrogio and C. Villani.

   (a) A function $f : U \to \mathbb{R}$ for open $U \subset \mathbb{R}^d$ is 1-Lipshitz if $\forall x_1, x_2 \; |f(x_1) - f(x_2)| \leq \|x_1 - x_2\|$. Observe that 1-Lipschitzness implies that the function is continuous, and the norm of its gradient is bounded. The latter observation can be used to enforce the Lipschitz property. Some intuition on why f should be 1-Lipschitz can be found in **Solution 1** to the exercise 3 of the homework for **Week 6**.

   (b) As was established in 2.a, enforcing the Lipschitz property is equivalent to enforcing the bound on the norm of the gradient. They ways to achieve this, among others, include weight clipping, gradient clipping, or spectral normalization.

Week 6

Problem 1: Fréchet distance

Given $p_1(x) = \mathcal{N}(x; \mu_1, \Sigma_1)$ and $p_2(x) = \mathcal{N}(x; \mu_2, \Sigma_2)$, the Fréchet distance between p_1 and p_2 can be written in closed form as

$$\mathcal{W}_2^2(p_1, p_2) = \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right)$$

1. Find simpler expressions in the cases (1) $\Sigma_1 = \Sigma_2$, (2) p is standard normal ($\mathcal{N}(0, I)$).
2. Show that $\mathcal{W}_2^2(p_1, p_2) = \mathcal{W}_2^2(p_2, p_1)$.
3. Fréchet inception distance (FID) is computed using Gaussians fit to features of an InceptionNet pretrained on ImageNet. Explain the motivation for using InceptionNet features instead of using Gaussians fit to raw pixel values.

1. Closed-form expressions for $\mathcal{W}_2^2(p_1, p_2)$ given different p_1 and p_2:

   (1) $\mathcal{W}_2^2(\mathcal{N}(\mu_1, \Sigma), \mathcal{N}(\mu_2, \Sigma)) = \|\mu_1 - \mu_2\|_2^2$;

   (2) $\mathcal{W}_2^2(\mathcal{N}(\mu, \Sigma), \mathcal{N}(0, I)) = \|\mu\|_2^2 + \mathrm{Tr}\left(I + \Sigma - 2\Sigma^{1/2}\right) = \left\{\Sigma = U\Lambda U^T;\ UU^T = I\right\}$
   $= \|\mu\|_2^2 + \mathrm{Tr}\left(I + \Lambda - 2\Lambda^{1/2}\right) = \|\mu\|_2^2 + \sum_1^d \left(1 - \sqrt{\lambda_i}\right)^2$;

   (3) assuming $\Sigma_i = D_i$ - diagonal matrix:
   $\mathcal{W}_2^2(\mathcal{N}(\mu_1, D_1), \mathcal{N}(\mu_2, D_2)) = \|\mu_1 - \mu_2\|_2^2 + \mathrm{Tr}\left(D_1 + D_2 - 2(D_1 D_2)^{1/2}\right)$
   $= \|\mu_1 - \mu_2\|_2^2 + \sum_1^d \left(\sqrt{d_{1,i}} - \sqrt{d_{2,i}}\right)^2$.

2. In order to show that Fréchet distance is symmetric, it is enough to prove that

$$\mathrm{Tr}\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right) = \mathrm{Tr}\left((\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}\right).$$

   (The remaining terms are obviously invariant to exchanging p_1 and p_2.)

   Proof. We prove the conjecture in two steps:

   (1) let us show that $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$ and $\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}$ have the same eigenvalues. Let λ be the eigenvalue of $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$ with the corresponding eigenvector v, then

   $$\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}v = \lambda v.$$

   Using the fact that $\Sigma_2 = \Sigma_2^{1/2}\Sigma_2^{1/2}$, multiply the expressions by $\Sigma_2^{1/2}\Sigma_1^{1/2}$ from the left:

   $$\underbrace{\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}}\underbrace{\Sigma_2^{1/2}\Sigma_1^{1/2}v}_{u} = \lambda\underbrace{\Sigma_2^{1/2}\Sigma_1^{1/2}v}_{u} \implies \Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}u = \lambda u.$$

   thus, if (λ, v) is an eigenvalue-eigenvector pair of $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$, then $(\lambda, \Sigma_2^{1/2}\Sigma_1^{1/2}v)$ is an eigenvalue-eigenvector pair of $\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}$.

   (2) Let $A = \Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$ and $B = \Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}$; so we must show $\mathrm{Tr}(A^{1/2}) = \mathrm{Tr}(B^{1/2})$. Given that A and B have the same eigenvalues, they admit eigendecompositions $A = U\Lambda U^T$ and $B = V\Lambda V^T$. Finally,

   $$\mathrm{Tr}(A^{1/2}) = \mathrm{Tr}(U\Lambda^{1/2}U^T) = \mathrm{Tr}(\Lambda^{1/2}) = \mathrm{Tr}(V^T B^{1/2} V) = \mathrm{Tr}(B^{1/2}),$$

   completing the proof.

□

3. Pointwise distance in pixel space is not necessarily a good measure of similarity between images, because pixel intensities might have quite a significant variability that does not meaningfully correspond to visual variability. Ideally, we want to avoid having such variablility. In theory, classifiers like InceptionNet learn meaningful features that can allow to distinguish between images, since the features required for distinguishing images are similar to those required to classify them. That is why using features of a learnt classifier might be actually advantageous. It is possible to actually see how such features look like.

Problem 2: InfoNCE

For this problem, please read this paper. Given a dataset of i.i.d. samples $\{x_i\}_{i=1}^N$, we want learn a representation of x_i given by $c_i = \text{Enc}(x_i)$. This problem can be solved by using (V)AE, i.e., $x \xrightarrow{\text{Enc}} c \xrightarrow{\text{Dec}} \hat{x} \approx x$. However, such a method requires training a reconstruction network Dec. Instead, one can try to maximise some measure of similarity between the distributions of x_i and c_i. This is what InfoNCE loss allows us to do.

1. Following the paper, what are the disadvantages of training a reconstruction model Dec?
2. The paper proposes to use mutual information $I(x,c) = \text{KL}\Big(p(x,c)\|p(x)\otimes p(c)\Big)$ as a measure of similarity between distributions of x_i and c_i.
   (a) Show that $I(x,c) = \mathbb{E}_{p(x,c)} \log \frac{p(x|c)}{p(x)}$
   (b) The InfoNCE loss for a tuple of (x, x_{neg}, c) can be written as follows.

$$\mathcal{L}(f) = -\mathbb{E}_{(x,x_{\text{neg}},c)} \log \frac{f(x,c)}{f(x,c)+f(x_{\text{neg}},c)}.$$

   Assuming that the optimum $f(x,c)$ is given by $\frac{p(x|c)}{p(x)}$ and $f(x_{\text{neg}},c) = \frac{p(x_{\text{neg}}|c)}{p(x_{\text{neg}})} \approx 1$, show

$$\mathcal{L}(f) \geq -\mathbb{E}_{(x,c)} \log \frac{p(x \mid c)}{p(x)} = -I(x,c).$$

   In other words, miminising $\mathcal{L}(f)$ allows to maximise mutual information between x and c.

1. We want to avoid spending compute on training the model to reconstruct the data objects, as that does not help us to learn a representation of the data that is useful for downstream uses (such as classification), but forces the models to concentrate on high-frequency features.

2. (a) $$I(x,c) = \int p(x,c) \log \frac{p(x,c)}{p(x)p(c)}\, dx\, dc = \underbrace{\int p(x,c) \log \frac{p(x \mid c)\cancel{p(c)}}{p(x)\cancel{p(c)}}\, dx\, dc}_{=\mathbb{E}_{p(x,c)} \log \frac{p(x|c)}{p(x)}}$$

$$= \int p(c) \left[\int p(x \mid c) \log \frac{p(x \mid c)}{p(x)}\, dx \right] dc = \mathbb{E}_{p(c)} \left[\text{KL}\left(p(x \mid c)\|p(x)\right) \right].$$

   (b) $$\mathcal{L}(f) = \mathbb{E}_{(x,x_{\text{neg}},c)} \log \left(1 + \frac{f(x_{\text{neg}},c)}{f(x,c)}\right) = \mathbb{E}_{(x,x_{\text{neg}},c)} \log \left(1 + \frac{p(x)}{p(x \mid c)}\right)$$ where the inequality uses that $\log(1+a)$

$$\geq \mathbb{E}_{(x,c)} \log \frac{p(x)}{p(x \mid c)} = -I(x,c),$$

Problem 3: Simple OT

Let $p(x) = \frac{1}{2}\mathbb{1}_{[0,2]}(x)$ and $q(x) = \frac{1}{2}\mathbb{1}_{[1,3]}(x)$, where $\mathbb{1}_A$ is the indicator function of A, i.e., $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise.

1. Show that $f(x) = x + 1$ is the optimal transport map for the problem $\min_{f:f_{\#p}=q} \int_{\mathbb{R}} |x - f(x)|\, p(x)\, dx$. Notice that this problem is 1-OT (the absolute value of the distance is not squared), not 2-OT as used in FID.
2. Show that $f(x) = \begin{cases} x+2 & \text{if } x \in [0,1] \\ x & \text{if } x \in (1,2] \end{cases}$ is another OT map with the same cost . What does this tell about the uniqueness of 1-OT maps?

1. For this problem we consider two approaches to the solution, which will be presented quite informally (because the full formalisation would require us to detail the concepts that are beyond the scope of this course). Nevertheless, the presented solutions should provide some intuition into how such problems can be solved; for a mathematically rigourous treatment of the problem refer to F. Santambrogio (Chapter 2, Sections 1-2), and F. Maggi (Part 1).

Solution 1: OT via duality

We can provide the following lower bound for the problem using some 1-Lipschitz function $\varphi : \mathbb{R} \to \mathbb{R}$.

$$\int \varphi(x)p(x)\,dx - \int \varphi(y)q(y)\,dy$$

given the push-forward $f_{\#p} = q$, we substitute $y = f(x)$

$$= \int [\varphi(x) - \varphi(f(x))]\, p(x)\, dx$$

applying the 1-Lipschitz property, $|\varphi(x) - \varphi(x')| \leq |x - x'|$

$$\leq \int |x - f(x)| p(x)\, dx$$

Since the previous inequality holds for any 1-Lipschitz function φ and any transport map f such that $f_{\#}p = q$, we can take the supremum over all possible 1-Lipschitz functions and infimum over all possible transport maps:

$$\sup_{\varphi} \left\{ \int \varphi(x)p(x)\,dx - \int \varphi(y)q(y)\,dy \right\} \leq \inf_{f:f_{\#p}=q} \left\{ \int |x - f(x)|p(x)\,dx \right\}$$

Should the pair (φ, f) be one for which the equality holds, this would immediately imply that f is an optimal transport map. It is easy to check that $\varphi(x) = -x$ and $f(x) = x + 1$ constitute such a pair:

$$\begin{aligned} \int \varphi(x)p(x)dx - \int \varphi(y)q(y)dy &= \int_0^2 -x\frac{1}{2}dx - \int_1^3 -y\frac{1}{2}dy \\ &= -\left[\frac{x^2}{4}\right]_0^2 + \left[\frac{y^2}{4}\right]_1^3 = -1 + \frac{9-1}{4} = 1 \\ \int_{\mathbb{R}} |x - f(x)|p(x)dx &= \int_0^2 |x - (x+1)| \cdot \frac{1}{2}dx \\ &= \frac{1}{2}\int_0^2 |-1|dx = \frac{1}{2} \cdot 2 = 1. \end{aligned}$$

Solution 2: OT via monotonicity

To show that $f(x) = x + 1$ is the optimal transport map, we construct a monotone map using the cumulative distribution functions (CDFs) and justify its optimality based on the cost function properties.

Constructing the monotone map using CDFs Let F_p and F_q be the CDFs of p and q, respectively. A fundamental property of CDFs is that they push their own distribution to a Uniform distribution on $[0, 1]$:

$$(F_p)_{\#}p = \text{Uniform}[0, 1] \quad \text{and} \quad (F_q)_{\#}q = \text{Uniform}[0, 1.$$

We can see that the transport map f happens to be equal to

$$f(x) = F_q^{-1}(F_p(x))$$

, since, given $p(x) = \frac{1}{2}\mathbb{1}_{[0,2]}(x)$ and $q(x) = \frac{1}{2}\mathbb{1}_{[1,3]}(x)$, we can calculate:

$$F_p(x) = \int_0^x \frac{1}{2}\, dt = \frac{x}{2} \quad \text{for } x \in [0, 2],$$
$$F_q(x) = \int_1^x \frac{1}{2}\, dt = \frac{x-1}{2} \quad \text{for } x \in [1, 3].$$

Inverting F_q gives $F_q^{-1}(y) = 2y + 1$. Substituting $F_p(x)$ into this inverse:

$$f(x) = F_q^{-1}\left(\frac{x}{2}\right) = 2\left(\frac{x}{2}\right) + 1 = x + 1$$

This map is strictly monotone on the support of p.

Optimality of monotone maps for one-dimensional 1-OT. For the cost function $c(x, y) = |x - y|$, the OT map between p and q is always given by $f(x) = F_q^{-1}(F_p(x))$. This is because for any $x < x'$ and $y < y'$, the following inequality holds:

$$|x - y| + |x' - y'| \leq |x - y'| + |x' - y|$$

Informally, this means that"crossing" paths (sending x to y' and x' to y) is more costly than maintaining the order by sending x to y and x' to y'. Since the OT map constructed is monotone, it ensures that no mass 'crosses' and is therefore an optimal transport map.

2. As found above, the cost of the optimal transport map in Part 1 is 1. For the map in Part 2, it is not hard to see that for $f(x) = \begin{cases} x + 2 & \text{if } x \in [0, 1] \\ x & \text{if } x \in (1, 2] \end{cases}$ the cost is

$$\int_{\mathbb{R}} |x - f(x)| p(x)\, dx = 2 \cdot \frac{1}{2} \int_{\mathbb{R}} \mathbb{1}_{[0,1]}(x)\, dx = 1.$$

Since the cost for this function is the same as in part 1 this function is also an optimal transport map. Given that we found that we found two optimal transform maps for the same problem, this tells us that in general the solutions to the OT problems are not guaranteed to be unique.

You might think that this example contradicts the argument for **Solution 2**, which claims that monotone maps should be the optimal maps. But note that this argument establishes the implication only in one way, namely, that 'if a map is monotone, it shall be optimal'. Yet, it does not exclude the possibility that there exists such non-monotone map that is also optimal. Indeed, for this example such a map exists.