Please submit your solutions via Ilias. The submission is not a formal requirement for passing the exam but doing the exercises will be very helpful to do so. Submissions should be a single PDF document (note that Jupyter notebooks can and should also be downloaded as PDFs, and not only submitted as .ipynb files).

1. **Intuition Question** *(Affine transformations)*

   Consider an m-dimensional normally-distributed random vector $\mathbf{u} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Let $\mathbf{x} = f(\mathbf{u}|\mathbf{A}, \mathbf{b}) := \mathbf{A}\mathbf{u} + \mathbf{b}$ be an affine transformation, where $\mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{b} \in \mathbb{R}^n$.

   (a) Using the formula for the transformation of random variables from the lecture, write the conditions for $f(\mathbf{u}|\mathbf{A}, \mathbf{b})$ to be a bijection. What is $f^{-1}(\mathbf{x})$?

   (b) Using $f^{-1}(\mathbf{x})$ from the previous part, derive $p_{\mathbf{x}}(\mathbf{x})$ in terms of $\mathbf{A}, \mathbf{b}, \mu, \mathbf{\Sigma}$.

   (c) For $m = n$, find a parameterization of $\mathbf{A}, \mathbf{b}$ such that $f(\mathbf{u}|\mathbf{A}, \mathbf{b})$ is always invertible. (*Hint*: enforce invertibility of $\mathbf{A}$ using Cholesky factorization or other methods.)

2. **Intuition Question** *(Conditional Densities)*

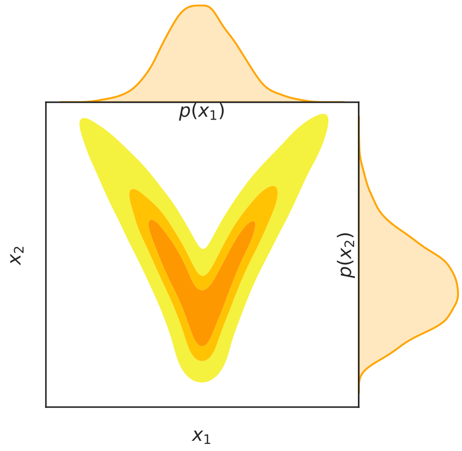
   Consider a 2-dimensional random variable $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \in \mathbb{R}^2$ drawn from a bivariate standard normal distribution $\mathbf{u} \sim \mathcal{N}(0, I_{2\times 2})$. A transformation $f : \mathbb{R}^2 \to \mathbb{R}^2$ is applied to produce $\mathbf{x} \in \mathbb{R}^2$.

   You are given THREE such transforms and TWO resulting density plots. For each transform, which of the density plots (zero or more) are possible for any set of the transformation parameters? Here, $a, c > 0,\ b, d \in \mathbb{R}$, and $g(\cdot), h(\cdot)$ are arbitrary functions with $g > 0$ everywhere. Assume sufficient flexibility for g, h.

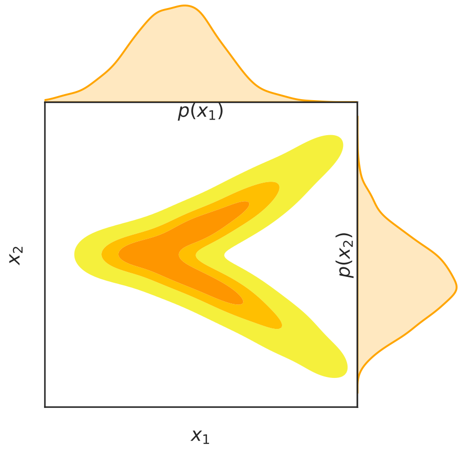
   (a) $f(\mathbf{u}|a, b, c, d) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a \cdot u_1 + b \\ c \cdot u_2 + d \end{bmatrix}$

   (b) $f(\mathbf{u}|a, b, g(\cdot), h(\cdot)) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a \cdot u_1 + b \\ g(u_1) \cdot u_2 + h(u_1) \end{bmatrix}$

   (c) $f(\mathbf{u}|a, b, g(\cdot), h(\cdot)) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} g(u_2) \cdot u_1 + h(u_2) \\ a \cdot u_2 + b \end{bmatrix}$



(a)



(b)

3. **Theory Question** *(Density estimation)* You have familiarized yourselves with the **Kullback-Liebler (KL) divergence** in the previous tutorial. This question deals with certain properties of the KL divergence, and its use in density estimation using normalizing flows. The KL divergence, which measures the discrepancy between two distributions $p(x)$ and $q(x)$, is defined as

$$D_{\mathrm{KL}}\left[p(x) \| q(x)\right] = \mathbb{E}_{p(x)}\left[\log \frac{p(x)}{q(x)}\right].$$

   (a) Is $D_{\mathrm{KL}}[p(x)\|q(x)] = D_{\mathrm{KL}}[q(x)\|p(x)]\ \forall p, q$? If not, provide a counterexample.

   (b) You are given an i.i.d. dataset $\{x_n\}_{n=1}^N$. x_n's are samples from the true data distribution $p_{\text{true}}(x)$. Given an approximating data distribution $q_\phi(x)$ parameterized with ϕ, the goal is to estimate the maximum likelihood estimate (MLE) of ϕ over the dataset.

   Prove that as the dataset size $N \to \infty$,

$$\theta_{\text{MLE}} = \underset{\phi}{\text{argmax}}\ q_\phi(\{x_n\}_{n=1}^N) \to \underset{\theta}{\text{argmin}}\ D_{\mathrm{KL}}[p_{\text{true}}(x)\|q_\phi(x)].$$

   In other words, for large number of datapoints, the MLE estimate converges to the parameter that minimizes the KL divergence between the true data distribution and the approximating data distribution. If needed, use the *uniform law of large numbers*[1].

   (c) Let $x_n \in \mathbb{R}^d$. Our goal is to approximate the density $p_{\text{true}}(x)$ by $q_\phi(x)$, such that we can generate new datapoints $x' \sim q_\phi(x)$. This can be done using a normalizing flow, which transforms a random vector which can be sampled *easily* into data from the *desired distribution.*

   Consider a general bijective transformation $x = f(u|\phi) : \mathbb{R}^d \to \mathbb{R}^d$. Assuming $u \sim \mathcal{N}(0, I)$, density estimation over x is performed by learning the parameters of the transformation f. The distribution over x induced by f can be learnt by minimizing $D_{\mathrm{KL}}[p_{\text{true}}(x)\|q_\phi(x)]$. Using the result of the previous part, write a loss function for finding optimal parameters of the normalizing flow.

   (d) (*Optional*) $D_{\mathrm{KL}}[q_\phi(x)\|p_{\text{true}}(x)]$ is usually referred to as the reverse KL divergence, as opposed to the previously mentioned $D_{\mathrm{KL}}[p_{\text{true}}(x)\|p(x|\theta)]$, which is called the forward KL divergence. Minimizing both of these takes the approximating distribution closer to the true distribution. In the context of learning normalizing flows, how should you choose between the two objectives? (*Hint*: note that the true distribution may be either hard to sample from, or hard to evaluate.)

4. **Practical Question** In this exercise you will train normalizing flows for density estimation using `Pyro`, a probabilistic programming framework based on `Pytorch`. See `Exercise_10.ipynb`.

[1] Law of large numbers Section 4.4 – Wikipedia, the Free Encyclopedia.