# CSCI-GA.2565-001 Machine Learning: Part 3

Chuanyang Jin

Dec 19, 2022

## 1 Variational Inference and Monte Carlo Gradients

In this question, we will review the details of variational inference (VI), in particular, we will implement the gradient estimators that make VI tractable.

We consider the latent variable model $p(\mathbf{z}, \mathbf{x}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)$ where $\mathbf{x}_i, \mathbf{z}_i \in \mathbb{R}^D$. Recall that in VI, we find an approximation $q_\lambda(\mathbf{z})$ to $p(\mathbf{z}|\mathbf{x})$.

(A) Let $V_1(\lambda)$ be the set of variational approximations $\{q_\lambda : q_\lambda(\mathbf{z}) = \prod_{i=1}^{N} q(\mathbf{z}_i; \lambda_i)\}$ where $\lambda_i$ are parameters learned for each datapoint $\mathbf{x}_i$. Now consider $f_\lambda(\mathbf{x})$ as a deep neural network with *fixed* architecture where $\lambda$ parametrizes the network. Let $V_2(\lambda) = \{q_\lambda : q_\lambda(\mathbf{z}) = \prod_{i}^{N} q(\mathbf{z}_i; f_\lambda(\mathbf{x}_i))\}$. Which of the two families ($V_1$ or $V_2$) is more expressive, i.e. approximates a larger set of distributions? **Prove** your answer.

Will your answer change if we let $f_\lambda$ represent *variable* architecture, e.g. if $\lambda$ parametrizes the set of multi-layered perceptrons of all sizes? Why or why not?

*Solution.*

$V_1$ is more expressive, i.e. approximates a larger set of distributions.

Let $q(\mathbf{z}; \theta_1, ..., \theta_N) \in V_2(\lambda)$, where $\theta_i = f_\lambda(x_i)$ for $i = 1, ..., N$. Since it can always be decomposed as $\prod_{i=1}^{N} q(\mathbf{z}_i; \lambda_i)$ where $\lambda_i$ are parameters learned for each datapoint $\mathbf{x}_i$, we have $q(\mathbf{z}; \theta_1, ..., \theta_N) \in V_1(\lambda)$. Therefore, $V_2 \subseteq V_1$. On the other hand, $V_1 \nsubseteq V_2$ in general. An counterexample would be when the *fixed* architecture of the network in $V_2$ only gives one-dimensional outputs, but the parameters in $V_1$ can be multi-dimensional.

Our answer will change if we let $f_\lambda$ represent *variable* architecture. $V_2 \subseteq V_1$ still holds for the same reason. Let $q(\mathbf{z}; \theta_1, ..., \theta_N) \in V_1(\lambda)$. If $\lambda$ parametrizes the set of multi-layered perceptrons of all sizes, since the number of data points is finite, there always exists an MLP to fit all the $\theta_i$'s such that $f_\lambda(x_i) = \theta_i$ for $i = 1, ..., N$. Therefore, $V_2 \subseteq V_1$. We conclude that $V_1$ and $V_2$ are equally expressive. □

(B) For variational inference to work, we need to compute unbiased estimates of the gradient of the ELBO. In class, we learnt two such estimators: score function (REINFORCE) and pathwise (reparametrization) gradients. Let us see this in practice for a simpler inference problem.

Consider the dataset of $N = 100$ one-dimensional data points $\{x_i\}_{i=1}^{N}$ in `data.csv`. Suppose we want to minimize the following expectation with respect to a parameter $\mu$:

$$\min_{\mu} \mathop{\mathbb{E}}_{z \sim \mathcal{N}(\mu,1)} \left[ \sum_{i=1}^{N} (x_i - z)^2 \right] \tag{1}$$

(i) Write down the score function gradient for this problem. Using a suitable reparametrization, write down the reparameterization gradient for this problem.

*Solution.*

Score function gradient:

$$\begin{aligned}
\nabla_\mu \mathbb{E}_{z\sim\mathcal{N}(\mu,1)}[f(z)] &= \int_z \nabla_\mu p_\mu(z) f(z) dz \\
&= \int_z p_\mu(z) \frac{\nabla_\mu p_\mu(z)}{p_\mu(z)} f(z) dz \\
&= \int_z p_\mu(z) \nabla_\mu \log p_\mu(z) f(z) dz \\
&= \mathbb{E}_{z\sim\mathcal{N}(\mu,1)} \left[\nabla_\mu \log p_\mu(z) f(z)\right] \\
&= \mathbb{E}_{z\sim\mathcal{N}(\mu,1)} \left[\sum_{i=1}^{N} (z-\mu)(x_i - z)^2\right]
\end{aligned}$$

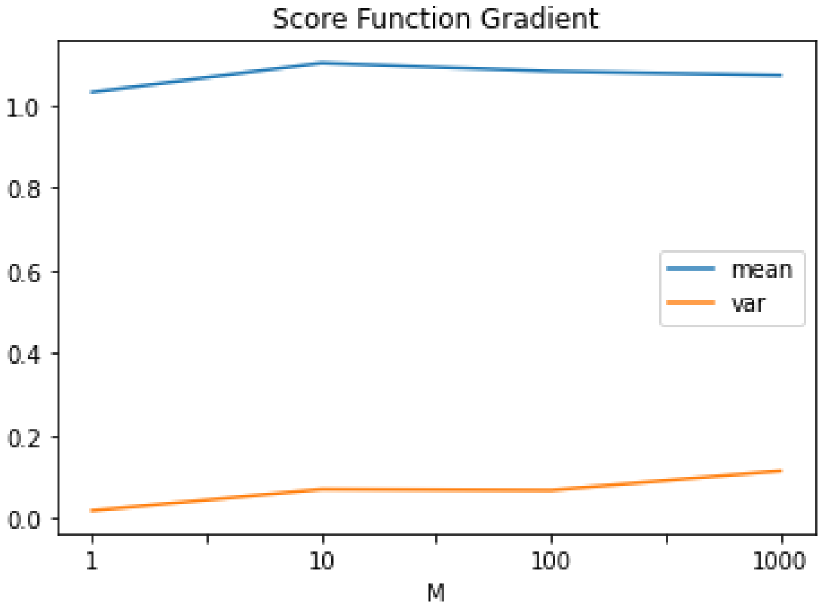
Reparameterization gradient: let $z = t_\mu(\epsilon) = \mu + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$.

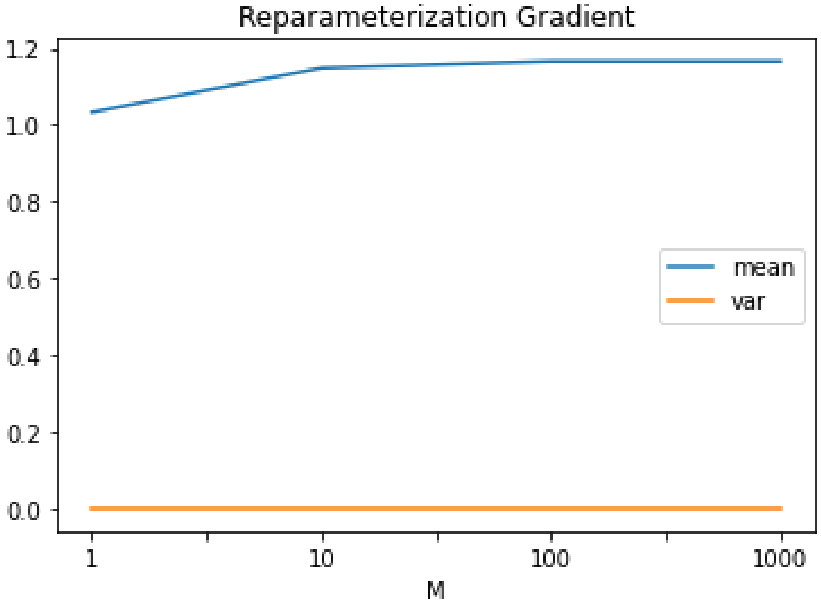
$$\begin{aligned}
\nabla_\mu \mathbb{E}_{z\sim\mathcal{N}(\mu,1)}[f(z)] &= \nabla_\mu \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}[f(t_\mu(\epsilon))] \\
&= \int_\epsilon p(\epsilon) \nabla_\mu f(t_\mu(\epsilon)) d\epsilon \\
&= \int_\epsilon p(\epsilon) \nabla_z f(z) \nabla_\mu t_\mu(\epsilon) d\epsilon \\
&= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)} \left[\nabla_z f(z) \nabla_\mu t_\mu(\epsilon)\right] \\
&= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)} \left[2\sum_{i=1}^{N} (\mu + \epsilon - x_i)\right] \\
&= 2N\mu - 2\sum_{i=1}^{N} x_i
\end{aligned}$$

□

(ii) Using PyTorch and for each of these two gradient estimators, perform gradient descent using $M = \{1, 10, 100, 1000\}$ gradient samples for $T = 10$ trials. Plot the mean and variance of the final estimate for $\mu$ for each value of $M$ across the $T$ trials. *You should have two graphs, one for each gradient estimator. Each of the graph should contain two plots, one for the means and one for the variances. The $x$-axis should be $M$, hence each of these plots will have four points.*

*Solution.*





□

(C) What conditions do you require on $p(z)$ and $f(z)$ ($f(z) = \sum_{i=1}^{N}(x_i - z)^2$ in this case) for each of the two gradient estimators to be valid? Do these apply to both continuous and discrete distributions $p(z)$?

*Solution.*

For the score gradient estimators to be valid, $p(z)$ is required to let $\log p_\theta(z)$ be differentiable with respect to the parameters $\theta$.

For the reparameterization gradient estimator to be valid, $f(z)$ must be differentiable with respect to the random variable $z$. □

## 2 Bayesian Parameters versus Latent Variables

(A) Consider the model $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$ where the inverse-variance is distributed $\lambda = 1/\sigma^2 \sim \text{Gamma}(\alpha, \beta)$. Show that the predictive distribution $y^\star | \mathbf{w}, \mathbf{x}^\star, \alpha, \beta$ for a datapoint $\mathbf{x}^\star$ follows a generalized T distribution

$$T(t; \nu, \mu, \theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\theta\sqrt{\pi\nu}} \Big(1 + \frac{1}{\nu}\Big(\frac{t-\mu}{\theta}\Big)^2\Big)^{-\frac{\nu+1}{2}}$$

with degree $\nu = 2\alpha$, mean $\mu = \mathbf{w}^\top \mathbf{x}^\star$ and scale $\theta = \sqrt{\beta/\alpha}$. You may use the property $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$.

*Solution.*

$$\begin{aligned}
p(y_i|x_i) &= \int_{\lambda=0}^{\infty} p(y_i|x_i, \lambda) p(\lambda) d\lambda \\
&= \int_{\lambda=0}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y-w^T x}{\sigma})^2} \beta e^{-\beta\sqrt{\lambda}} \frac{(\beta\lambda)^{\alpha-1}}{\Gamma(\alpha)} d\lambda \\
&= \int_{\lambda=0}^{\infty} \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2}(y-w^T x)^2} \beta e^{-\beta\sqrt{\lambda}} \frac{(\beta\lambda)^{\alpha-1}}{\Gamma(\alpha)} d\lambda \\
&= \int_{\lambda=0}^{\infty} \sqrt{\frac{\lambda}{2\pi}} e^{-\lambda(\frac{(y-w^T x)^2}{2}+\beta)} \beta^\alpha \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)} d\lambda \\
&= \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)} \int_{\lambda=0}^{\infty} \lambda^{\alpha-\frac{1}{2}} e^{-\lambda(\frac{(y-w^T x)^2}{2}+\beta)} d\lambda \\
&= \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)} \cdot \frac{1}{(\frac{(y-w^T x)^2}{2} + \beta)^{\alpha+\frac{1}{2}}} \cdot \int_{\lambda=0}^{\infty} \left(\lambda(\frac{(y-w^T x)^2}{2} + \beta)\right)^{\alpha-\frac{1}{2}} e^{-\lambda(\frac{(y-w^T x)^2}{2}+\beta)} d\left(\lambda(\frac{(y-w^T x)^2}{2} + \beta)\right) \\
&= \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+\frac{1}{2})}{(\frac{(y-w^T x)^2}{2} + \beta)^{\alpha+\frac{1}{2}}} \\
&= \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)\sqrt{2\pi\beta}} \cdot \beta^{\alpha+\frac{1}{2}} \cdot \Big(\beta + \frac{(y-\mu)^2}{2}\Big)^{-(\alpha+\frac{1}{2})} \\
&= \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)\sqrt{2\pi\beta}} \Big(1 + \frac{(t-\mu)^2}{2\beta}\Big)^{-(\alpha+\frac{1}{2})} \\
&= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\theta\sqrt{\pi\nu}} \Big(1 + \frac{1}{\nu}\Big(\frac{t-\mu}{\theta}\Big)^2\Big)^{-\frac{\nu+1}{2}}
\end{aligned}$$

□

(B) Using your expression in (A), write down the MLE objective for $\mathbf{w}$ on $N$ arbitrary labelled datapoints $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Do not optimize this objective.

*Solution.*

$$\begin{aligned}
\max_{\mathbf{w}} \prod_{i=1}^{N} p(y_i|x_i) &= \max_{\mathbf{w}} \prod_{i=1}^{N} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\theta\sqrt{\pi\nu}} \Big(1 + \frac{1}{\nu}\Big(\frac{y_i - \mathbf{w}^T x_i}{\theta}\Big)^2\Big)^{-\frac{\nu+1}{2}} \\
&= \max_{\mathbf{w}} \prod_{i=1}^{N} \Big(1 + \frac{1}{\nu}\Big(\frac{y_i - \mathbf{w}^T x_i}{\theta}\Big)^2\Big)^{-\frac{\nu+1}{2}} \\
&= \max_{\mathbf{w}} \log \Big[\prod_{i=1}^{N} \Big(1 + \frac{1}{\nu}\Big(\frac{y_i - \mathbf{w}^T x_i}{\theta}\Big)^2\Big)^{-\frac{\nu+1}{2}}\Big] \\
&= \max_{\mathbf{w}} \sum_{i=1}^{N} \log \Big(1 + \frac{1}{\nu}\Big(\frac{y_i - \mathbf{w}^T x_i}{\theta}\Big)^2\Big)^{-\frac{\nu+1}{2}}
\end{aligned}$$

□

(C) Now consider the model $y_i \sim \mathcal{N}\Big(f(\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}), \sigma^2\Big)$ where $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\sigma^2$ is known, and $f$ is a deep neural network parametrized by $\mathbf{w}$.

(i) Write down an expression for the predictive distribution $y^\star|\mathbf{X}, \mathbf{y}, \mathbf{x}^\star$, where $\mathbf{X}, \mathbf{y}$ denote the training datapoints. *(You may leave your answer as an integral.)*

*Solution.*

$$\begin{aligned}
p(y^\star|\mathbf{X}, \mathbf{y}, \mathbf{x}^\star) &= \int p(y^\star|\mathbf{x}^\star, \mathbf{z}^\star) p(\mathbf{z}^\star|\mathbf{X}, \mathbf{y}) d\mathbf{z}^\star \\
&= \int \frac{1}{2\pi\sigma^2} \exp\Big[-\frac{(\mathbf{y}^\star - f(\mathbf{x}^\star, \mathbf{z}^\star; \mathbf{w}))^2}{2\sigma^2}\Big] p(\mathbf{z}^\star|\mathbf{X}, \mathbf{y}) d\mathbf{z}^\star
\end{aligned}$$

□

(ii) Describe how you would approximate this distribution using variational inference and how you can use your approximation to make a prediction for $\mathbf{x}^\star$. Your answer should include the distribution $p(\cdot)$ that you wish to approximate *(which may or may not be the predictive distribution itself)*, the distribution $q(\cdot)$ that is the variational approximation, as well as the variational objective.

*Solution.*

We want to approximate $p(\mathbf{z}|\mathbf{X}, \mathbf{y})$, the posterior distribution of the latent variable $\mathbf{z}^\star$ given the training data $\mathbf{X}$ and $\mathbf{y}$.

The variational objective is to maximize the ELBO, $\mathbb{E}_q[\log p(\mathbf{y}, \mathbf{z}|\mathbf{X})] - \mathbb{E}_q[\log q_\lambda(\mathbf{z})]$. We can use the score function gradient to optimize it.

Then we have the variational approximation $q_{\lambda^\star}(\mathbf{z})$ for $p(\mathbf{z}^\star|\mathbf{X}, \mathbf{y})$, and infer $\mathbf{z}^\star$ by taking $\mathbf{z}^\star = \mathrm{argmax}_z\, q_{\lambda^\star}(z)$.

Finally, we have $y^\star \sim \mathcal{N}(f(\mathbf{x}^\star, \mathbf{z}^\star, \mathbf{w}), \sigma^2)$. We make a prediction for $\mathbf{x}^\star$ by taking

$$y^\star = \operatorname*{argmax}_c \mathcal{N}(c; f(\mathbf{x}^\star, \mathbf{z}^\star, \mathbf{w}), \sigma^2) = f(\mathbf{x}^\star, \mathbf{z}^\star, \mathbf{w})$$

.

□

(D) Finally, consider the model $y \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$ where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, I)$ and $\sigma^2$ is known.

Derive a closed-form expression for the predictive distribution $y^\star|\mathbf{X}, \mathbf{y}, \mathbf{x}^\star$. What are the parameters of this predictive distribution and how do you optimize them?

*Solution.*

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \mathcal{N}((\mathbf{X}^T\mathbf{X}+I)^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2(\mathbf{X}^T\mathbf{X}+I)^{-1})$$

Our predictive distribution is

$$\begin{aligned} p(y^\star|\mathbf{X},\mathbf{y},\mathbf{x}^\star) &= \int_{\mathbf{w}} N(\mathbf{y}^\star|\mathbf{w}^T\mathbf{x}^\star, \sigma^2)\mathcal{N}((\mathbf{X}^T\mathbf{X}+I)^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2(\mathbf{X}^T\mathbf{X}+I)^{-1})d\mathbf{w} \\ &= N(\mathbf{y}^\star|((\mathbf{X}^T\mathbf{X}+I)^{-1}\mathbf{X}^T\mathbf{y})^T\mathbf{x}^\star, \sigma^2(I+\mathbf{x}^\star(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^\star)) \end{aligned}$$

So we don't need to optimize **w** anymore. We only need to optimize the hyper-parameters on the marginal likelihood of the data. □

(E) Of the three models defined in parts (A), (C), and (D) above, which are latent variable models and which are not? Why? *(If any are ambiguous, explain why.)*

*Solution.*

The models defined in parts (C) and (D) are latent variable models, since unobserved random variables **z** and **w** are introduced to indirectly estimate the distribution.

The model in part (A) is not a latent variable model, since no such unobserved random variables are introduced. □

(F) Of the three models defined in parts (A), (C), and (D) above, which are Bayesian models and which are not? Why? *(If any are ambiguous, explain what is Bayesian about it and what is not.)*

*Solution.*

The models defined in parts (C) and (D) are Bayesian models, since they assume prior distributions on some model parameters and compute their posterior distributions.

The model defined in (A) is ambiguous. It assumes prior distribution on the variance $\sigma^2$, which is from the Bayesian perspective, but there are no priors over any model parameters in predicting new data points. □

## 3 Normalizing Flows

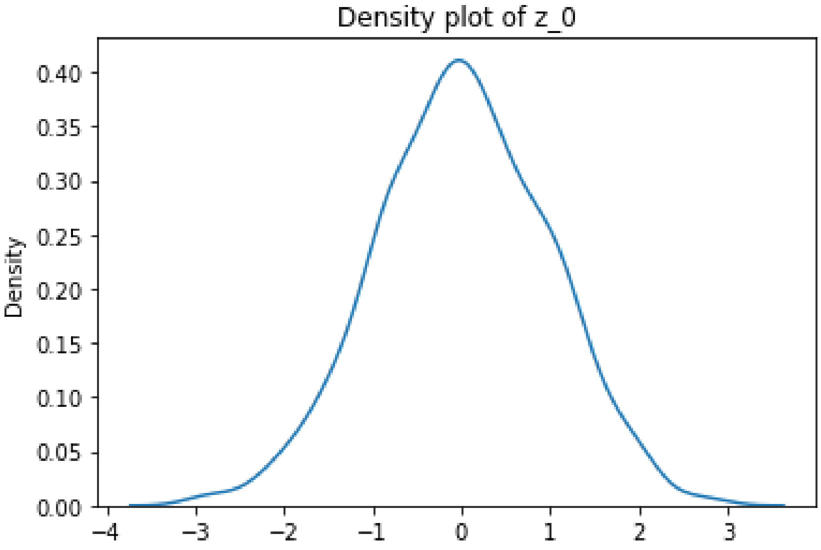
In this question, we will review how we can use invertible transformations and the change-of-variables formula to turn simple distributions into complex ones. Such transformations are known as *normalizing flows*. One reason that flows are useful is that they can map unimodal distributions into multimodal ones, while still allowing for a tractable density.

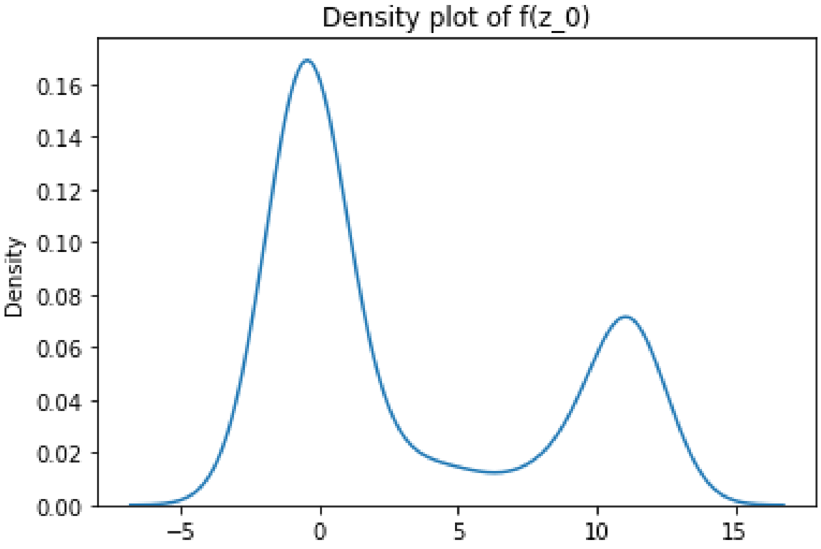
(A) Let $z_0 \sim \mathcal{N}(0, 1)$. Produce a density plot of $z_0$ using $N = 1000$ samples.

Now look up "Planar Flow" (Equation 4) of The Expressive Power of a Class of Normalizing Flow Models. Denote this flow as $f$. Choose an invertible non-linearity $h$ and find values of $w, b, u$ such that $f(z_0)$ is a multimodal distribution. Plot the density plot of $f(z_0)$ using the same $N = 1000$ samples as above.

*Note that* $d = 1$ *in this question. Also, it is fine to choose a* $h$ *that is only invertible on its output range, e.g. the sigmoid function on* $(0, 1)$.

*Solution.*





$f(z_0) = z_0 + uh(w^T z_0 + b)$
where $u = 10$, $w = 10$, $b = 5$, $h(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. □

(B) Use the change-of-variables formula and write down an explicit expression for the density of $f(z_0)$. This depends on your choice of $h$.

*Solution.*

$f(z_0) = z_0 + uh(w^T z_0 + b) = z_0 + \frac{10}{1+e^{-10\cdot z_0+5}}$ □

(C) Let's generalize this to $D$-dimensional variables and to a sequence of invertible functions $f$ (not necessarily planar flows). We can sample $\mathbf{z}^{(0)} \sim \mathcal{N}(\mathbf{0}, I)$ and then transform that sample iteratively using a sequence of invertible functions, $f_1, \ldots, f_K$, to finally obtain a sample $\mathbf{z}^{(K)}$ where

$$\mathbf{z}^{(K)} = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{z}^{(0)}).$$

Using the change-of-variables formula, write down a formula for the log-density of $\mathbf{z}^{(K)}$ using the functions $\{f_k\}_{k=1}^K$, their inverses and Jacobians, as well as the log-density of $\mathbf{z}^{(0)}$.

*Solution.*

The change-of-variables formula gives:
If $b = g(a)$, then $f_b(b) = f_a(a)|\frac{\partial a}{\partial b}|$ where $|\frac{\partial a}{\partial b}|$ is the Jacobian matrix. Using the formula, the density of $\mathbf{z}^{(K)}$ can be expressed as

$$\begin{aligned} f(\mathbf{z}^{(K)}) &= f(\mathbf{z}^{(K-1)})|H^K| \\ &= f(\mathbf{z}^{(K-2)})|H^{K-1}||H^K| \\ &\ldots \\ &= f(\mathbf{z}^0)\prod_{i=1}^K |H^i| \end{aligned}$$

Then

$$\begin{aligned} \log(f(\mathbf{z}^{(K)})) &= \log(f(\mathbf{z}^0)\prod_{i=1}^K |H^i|) \\ &= \log(f(\mathbf{z}^0))\sum_{i=1}^K \log|H^i| \end{aligned}$$

Since $\mathbf{z}^{(K)} = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{z}^{(0)})$, we have $\mathbf{z}^{(0)} = f_1^{-1} \circ \cdots \circ f_{K-1}^{-1} \circ f_K^{-1}(\mathbf{z}^{(K)})$. Then

$$\begin{aligned} \log(f(\mathbf{z}^{(K)})) &= \log(f(\mathbf{z}^0))\sum_{i=1}^K \log|H^i| \\ &= \log(f(f_1^{-1} \circ \cdots \circ f_{K-1}^{-1} \circ f_K^{-1}(\mathbf{z}^{(K)})) + \sum_{i=1}^K \log|H^i| \end{aligned}$$

□

(D) Let us consider how we can improve variational inference using flows.

(i) How can we define an approximation $q_\lambda(\mathbf{z})$ to $p(\mathbf{z}|\mathbf{x})$ using flows?

(ii) How do we train the model, i.e. optimize $\lambda$ to yield a good approximation to $p(\mathbf{z}|\mathbf{x})$?

(iii) In what way is this more flexible than using a Gaussian approximation for $q_\lambda(\mathbf{z})$?

*Solution.*

(i) We can use flows to map $\mathbf{z}$ into another vector space: $q_\lambda(\mathbf{z}) = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{z})$.

(ii) We employ methods similar to variational inference. We can take the score gradient or reparametrization gradient of the KL divergence between $p(\mathbf{z}|\mathbf{x})$ and $q_\lambda(\mathbf{z})$. Note that we can take the derivative of the function level by level from the outer flows to the inner flows.

(iii) Gaussian approximation can only create a single modal. But as shown in (b), we can use this flow-based methods to create a multi-modal distribution.

□

## 4 Causal Inference: Doubly Robust Estimators

Denote unit $i$'s treatment, outcome, and its vector of covariates by $T_i$, $Y_i$, and $X_i$. Let us model propensity scores by $e(x;\hat{\theta})$ and the outcome by $f(x;\hat{\psi})$. Assume strong ignorability and positivity hold. We define the Doubly Robust Estimator (DRE) for the average outcome when treated, $\mathbb{E}[Y^{(1)}]$, by:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_iY_i}{e(X_i;\hat{\theta})} - \frac{T_i - e(X_i;\hat{\theta})}{e(X_i;\hat{\theta})}f(X_i;\hat{\psi})\right] \tag{2}$$

(A) Suppose the propensity model is correctly specified, i.e. $e(x;\hat{\theta}) = P(T_i = 1|X_i = x)$. Given any function $f$, show that the DRE is unbiased.

*Solution.*

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_iY_i}{e(X_i;\hat{\theta})} - \frac{T_i - e(X_i;\hat{\theta})}{e(X_i;\hat{\theta})}f(X_i;\hat{\psi})\right]\right]\\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\left[\frac{T_iY_i}{e(X_i;\hat{\theta})} - \frac{T_i - e(X_i;\hat{\theta})}{e(X_i;\hat{\theta})}f(X_i;\hat{\psi})\right]\right]\\
&= \mathbb{E}_{T,X,Y}\left[\frac{TY}{e(X;\hat{\theta})} - \frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{T,X,Y}\left[\frac{TY}{e(X;\hat{\theta})}\right] - \mathbb{E}_{T,X,Y}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{T,X,Y}\left[\frac{TY^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{T,X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right] \quad \text{since } Y = TY^1 + (1-T)Y^0 \text{ and } T = 0 \text{ when } Y = Y^0\\
&= \mathbb{E}_{X,Y}\mathbb{E}_{T|X,Y}\left[\frac{TY^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{T,X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X,Y]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{T,X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X,Y]Y^1}{P(T=1|X)}\right] - \mathbb{E}_{T,X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{X,Y}\left[Y^1\right] - \mathbb{E}_{T,X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right] \quad \text{since strong ignorability} \rightarrow \mathbb{E}[T|X,Y] = \mathbb{E}[T|X] = P(T=1|X)\\
&= \mathbb{E}_{X,Y}\left[Y^1\right] - \mathbb{E}_{X}\mathbb{E}_{T|X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{X,Y}\left[Y^1\right] - \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X] - e(X;\hat{\theta})}{e(X;\hat{\theta})}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{X,Y}\left[Y^1\right] - \mathbb{E}_{X}\left[\frac{P(T=1|X) - P(T=1|X)}{P(T=1|X)}f(X;\hat{\psi})\right]\\
&= \mathbb{E}_{X,Y}\left[Y^1\right]
\end{aligned}$$

Therefore, the DRE is unbiased. □

(B) Suppose the model $f(x;\hat{\psi})$ is correctly specified, i.e. $f(x;\hat{\psi}) = \mathbb{E}[Y_i^{(1)}|X_i = x] := \mathbb{E}[Y_i|X_i = x, T = 1]$. Given any function $e$ taking values in $(0,1)$, show that the DRE is unbiased.

*Solution.*

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_iY_i}{e(X_i;\hat{\theta})} - \frac{T_i - e(X_i;\hat{\theta})}{e(X_i;\hat{\theta})} f(X_i;\hat{\psi})\right]\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{T,X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})} f(X;\hat{\psi})\right] \quad \text{by (a)} \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{T,X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X]\right] \quad \text{since } f(x;\hat{\psi}) = \mathbb{E}[Y_i^{(1)}|X_i = x] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{X}\mathbb{E}_{T|X}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X]\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X] - e(X;\hat{\theta})}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X]\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X]}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X] - \mathbb{E}[Y^{(1)}|X]\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X]}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X]\right] + \mathbb{E}_{X}\left[\mathbb{E}[Y^{(1)}|X]\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X]}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X]\right] + \mathbb{E}\left[Y^{(1)}\right] \\
&= \mathbb{E}_{X}\mathbb{E}_{Y|X}\left[\frac{\mathbb{E}[T|X]Y^1}{e(X;\hat{\theta})}\right] - \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X]}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X]\right] + \mathbb{E}\left[Y^{(1)}\right] \\
&= \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X]\mathbb{E}[Y^1|X]}{e(X;\hat{\theta})}\right] - \mathbb{E}_{X}\left[\frac{\mathbb{E}[T|X]}{e(X;\hat{\theta})} \mathbb{E}[Y^{(1)}|X]\right] + \mathbb{E}\left[Y^{(1)}\right] \\
&= \mathbb{E}\left[Y^{(1)}\right]
\end{aligned}
$$

Therefore, the DRE is unbiased. □

(C) Recall that control variates improve Monte Carlo estimators by defining a new estimator with the same expectation but lower variance. For some estimator $f$, this is done by taking a function $g$ with $\mathbb{E}[g(x)] = 0$, and defining the new estimator as $\hat{f}(x) = f(x) - ag(x)$ for some $a \in \mathbb{R}$. Here $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$, but the variances are not equal. The value of $a$ that makes the variance of $\hat{f}$ smallest is $a^* = \frac{Cov(f,g)}{Var(g)}$.

When both $e(x;\hat{\theta})$ and $f(x;\hat{\psi})$ are correctly specified, use control variates to justify the use of Doubly Robust Estimators.

*Solution.*

Consider $f(T, X, Y) = \frac{TY}{e(X;\hat{\theta})}$, $g(T, X, Y) = \frac{T_i - e(X_i;\hat{\theta})}{e(X_i;\hat{\theta})}$.

$$\begin{aligned}
\mathbb{E}(f) &= \mathbb{E}_{T,X,Y}\left[\frac{TY}{e(X;\hat{\theta})}\right] \\
&= \mathbb{E}_{T,X,Y}\left[\frac{TY}{P(T=1|X)}\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X,Y]Y}{P(T=1|X)}\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X,Y]Y}{P(T=1|X)}\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{\mathbb{E}[T|X]Y}{P(T=1|X)}\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{P(T=1|X)Y^1}{P(T=1|X)}\right] \\
&= \mathbb{E}_X\mathbb{E}_{Y|X}\left[\frac{P(T=1|X)Y^1}{P(T=1|X)}\right] \\
&= \mathbb{E}_X\left[\frac{P(T=1|X)}{P(T=1|X)}\right]\mathbb{E}\left[Y^1|X\right] \\
&= \mathbb{E}\left[Y^1|X\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(g) &= \mathbb{E}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}\right] \\
&= \mathbb{E}\left[\frac{T}{e(X;\hat{\theta})}\right] - 1 \\
&= \mathbb{E}\left[\frac{T}{P(T=1|X)}\right] - 1 \\
&= \frac{P(T=0|X)\times 0 + P(T=1|X)\times 1}{P(T=1|X)} - 1 \\
&= 1 - 1 \\
&= 0
\end{aligned}$$

$$
\begin{aligned}
Var(g) &= \mathrm{Var}\left[\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}\right] \\
&= \mathbb{E}\left[(\frac{T}{e(X;\hat{\theta})} - 1)^2\right] \\
&= \mathbb{E}\left[(\frac{T}{P(T=1|X)} - 1)^2\right] \\
&= \mathbb{E}\left[\frac{T^2}{P(T=1|X)^2} - 2\frac{T}{P(T=1|X)} + 1\right] \\
&= \mathbb{E}\left[\frac{T^2}{P(T=1|X)^2}\right] - 2\mathbb{E}\left[\frac{T}{P(T=1|X)}\right] + 1 \\
&= \mathbb{E}\left[\frac{T^2}{P(T=1|X)^2}\right] - 2 + 1 \\
&= \mathbb{E}\left[\frac{T^2}{P(T=1|X)^2}\right] - 1
\end{aligned}
$$

$$
\begin{aligned}
Cov(f,g) &= \mathbb{E}\left[(f - \mathbb{E}(f))(g - \mathbb{E}(g)\right] \\
&= \mathbb{E}\left[(\frac{TY}{e(X;\hat{\theta})} - \mathbb{E}[Y^1|X])\frac{T - e(X;\hat{\theta})}{e(X;\hat{\theta})}\right] \\
&= \mathbb{E}\left[(\frac{TY}{P(T=1|X)} - \mathbb{E}[Y^1|X])\frac{T - P(T=1|X)}{P(T=1|X)}\right] \\
&= \mathbb{E}\left[\frac{T^2Y}{P(T=1|X)^2}\right] - \mathbb{E}\left[\frac{TY}{P(T=1|X)}\right] - \mathbb{E}\left[\frac{\mathbb{E}\left[Y^1|X\right]T}{P(T=1|X)^2}\right] + \mathbb{E}\left[Y^1|X\right] \\
&= \mathbb{E}\left[\frac{T^2Y}{P(T=1|X)^2}\right] - \mathbb{E}\left[Y^1|X\right] - \mathbb{E}\left[Y^1|X\right] + \mathbb{E}\left[Y^1|X\right] \\
&= \mathbb{E}_{T,X,Y}\left[\frac{T^2Y}{P(T=1|X)^2}\right] - \mathbb{E}\left[Y^1|X\right] \\
&= \mathbb{E}_{T,X,Y}\left[\frac{T^2Y^1}{P(T=1|X)^2}\right] - \mathbb{E}\left[Y^1|X\right] \\
&= \mathbb{E}_{T,X}\mathbb{E}_{Y|T,X}\left[\frac{T^2Y}{P(T=1|X)^2}\right] - \mathbb{E}\left[Y^1|X\right] \\
&= \mathbb{E}_{T,X}\left[\frac{T^2Y}{P(T=1|X)^2}\right]\mathbb{E}\left[Y^1|X\right] - \mathbb{E}\left[Y^1|X\right]
\end{aligned}
$$

Since $a^* = \frac{Cov(f,g)}{Var(g)} = \mathbb{E}\left[Y^1|X\right]$, our use of Doubly Robust Estimators is justified. □

## 5 Generative Models with $f$-divergences

Given two distributions $P$ and $Q$ with density functions $p$ and $q$, we define the $f$-divergence as:

$$D_f(P\|Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

with respect to a convex function $f$.

(A) How can we estimate $f$-divergences using likelihood-ratio estimators as we learned in class?

*Solution.*

We sample equal number of data from distributions $p$ and $q$. We assign a label $T$ for each data: $T = 0$ for data from $p$ and $T = 1$ for data from $q$. Then we train a distriminative model to predict whether each data come from $p$ or $q$. For a given data $x$, the model outputs $P(T = 1|x)$. Then

$$\begin{aligned}
\frac{p(x)}{q(x)} &= \frac{P(x|T=0)}{P(x|T=1)} \\
&= \frac{\frac{P(x,T=0)}{P(T=0)}}{\frac{P(x,T=1)}{P(T=1)}} \\
&= \frac{P(x,T=0)}{P(x,T=1)} \quad \text{since we sample equal number of data from } p \text{ and } q \rightarrow P(T=0) = P(T=1) \\
&= \frac{P(T=0|x)P(X=x)}{P(T=1|x)P(X=x)} \\
&= \frac{P(T=0|x)}{P(T=1|x)} \\
&= \frac{1-P(T=1|x)}{P(T=1|x)}
\end{aligned}$$

So we can use the model output $P(T = 1|x)$ to calculate $\frac{p(x)}{q(x)}$, and then insert it into the function $f$ to get $f(\frac{p(x)}{q(x)})$.
Then we can use Monte Carlo to estimate $f$-divergence. If we denote each data from $q$ as $x_1, x_2, ...x_n$, we can estimate $\widehat{f(\frac{p(x)}{q(x)})}$ using the above methods, and our estimation of $f$-divergence will be

$$D_f(P\|Q) \approx \frac{1}{n}\sum_{i=1}^{n} \widehat{f(\frac{p(x)}{q(x)})}$$

□

(B) Using the estimate from your solution to (A), show how we can minimize $D_f(P\|Q_\theta)$ with respect to $\theta$.

*Solution.*

$$
\begin{aligned}
\nabla_\theta D_f(P\|Q) &= \nabla_\theta \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx \\
&= \int \nabla_\theta \left(q(x) f\left(\frac{p(x)}{q(x)}\right)\right) dx \\
&= \int \nabla_\theta \left(q(x)\right) f\left(\frac{p(x)}{q(x)}\right) dx + \int q(x) \nabla_\theta \left(f\left(\frac{p(x)}{q(x)}\right)\right) dx \\
&= \mathbb{E}\left[\frac{\nabla_\theta(q(x))}{q(x)} f\left(\frac{p(x)}{q(x)}\right)\right] + \mathbb{E}\left[\nabla_\theta\left(f\left(\frac{p(x)}{q(x)}\right)\right)\right] \\
&= \mathbb{E}\left[\nabla_\theta\left(\log q(x)\right) f\left(\frac{p(x)}{q(x)}\right)\right] + \mathbb{E}\left[\nabla_\theta\left(f\left(\frac{p(x)}{q(x)}\right)\right)\right] \\
&= \mathbb{E}\left[\nabla_\theta\left(\log q(x)\right) f\left(\frac{p(x)}{q(x)}\right)\right] + \mathbb{E}\left[f'\left(\frac{p(x)}{q(x)}\right)\left(-\frac{p(x)}{q^2(x)}\right)\nabla_\theta(q(x))\right] \\
&= \mathbb{E}\left[\nabla_\theta\left(\log q(x)\right) f\left(\frac{p(x)}{q(x)}\right)\right] - \mathbb{E}\left[f'\left(\frac{p(x)}{q(x)}\right)\left(\frac{p(x)}{q(x)}\right)\nabla_\theta(\log q(x))\right]
\end{aligned}
$$

Following part (a), we can estimate $\frac{p(x)}{q(x)}$. We also have $q$ and $f$, so we can derive $\nabla_\theta(\log q(x))$ and $f'$. Then each term in the above expression is known.
We use Monte Carlo to estimate the gradient:

$$
\nabla_\theta D_f(P\|Q) \approx \frac{1}{n}\sum_{i=1}^{n}\left[\nabla_\theta\left(q(x_i)\right) f\left(\frac{p(x_i)}{q(x_i)}\right) - f'\left(\frac{p(x_i)}{q(x_i)}\right)\left(\frac{p(x_i)}{q(x_i)}\right)\nabla_\theta \log q(x_i)\right]
$$

□

(C) Let $\mathcal{Q}$ denote the family of distributions that $Q_\theta$ lives in. Assume that $P \notin \mathcal{Q}$. Compare the KL and Reverse KL divergences. What are properties that each $f$-divergence imposes on its corresponding minimizer $Q_\theta^\star$, i.e.

$$
Q_\theta^\star = \arg\min_{Q_\theta \in \mathcal{Q}} D_f(P\|Q_\theta)
$$

| Name | $D_f(P\|Q)$ | $f(u)$ |
|---|---|---|
| Kullback-Leibler | $\int p(x)\log\frac{p(x)}{q(x)}dx$ | $u\log u$ |
| Reverse KL | $\int q(x)\log\frac{q(x)}{p(x)}dx$ | $-\log u$ |

*Solution.*

For KL divergence,

- When $u$ is near 1, which means $p(x)$ and $q(x)$ are similar, $f(u) = u\log u$ will be close to 0.
- When $u$ is very small, which means $p(x)$ is small and $q(x)$ is large, $f(u) = u\log u$ will be negatively small.
- When $u$ is very large, which means $p(x)$ is large and $q(x)$ is small, $f(u) = u\log u$ will also be large.

For reverse KL divergence,

- When $u$ is near 1, which means $p(x)$ and $q(x)$ are similar, $f(u) = -\log u$ will be close to 0.
- When $u$ is very small, which means $p(x)$ is small and $q(x)$ is large, $f(u) = -\log u$ will be large.

- When $u$ is very large, which means $p(x)$ is large and $q(x)$ is small, $f(u) = -\log u$ will be small.

So minimizing KL divergence can avoid the situation where $p(x)$ is large and $q(x)$ is small, while minimizing reverse KL divergence can avoid the situation where $p(x)$ is small and $q(x)$ is large.
In other words, minimizing KL divergence ensure that $q(x)$ will be nonzero whenever $p(x)$ is nonzero, while minimizing reverse KL divergence ensures that $q(x)$ will be zero whenever $p(x)$ is zero. □

## 6 Reinforcement Learning with Sparse Rewards

Suppose that you have a robot that should learn to move from any of a set of starting positions $s_0 \in \mathcal{S}_0$ to a goal position $G$. The robot can move by performing a sequence of small, low-level continuous actions, such as rotating its joints. However, you do not know how to explicitly program a sequence of actions that will move the robot from any $s_0$ to $G$. You decide to use a reinforcement learning algorithm. Your RL algorithm uses the policy gradient to learn a policy $\pi_\theta(a|s)$ that maps from states $s$ to distributions over actions $a$. You consider learning episodes of a finite number of $T$ steps.

(A) You start by designing the Markov Decision Process (MDP) that defines the robot's learning environment. The robot receives reward $R = 1000$ when it reaches location $G$ and $R = 0$ upon entering any other position. What is the Monte Carlo estimate of the policy gradient when $G$ is not reached in $T$ steps in any of the sample trajectories? What happens as the robot explores in this environment and what will its learning process look like?

*Solution.*

The Monte Carlo estimate of the policy gradient is in the form

$$\frac{1}{S}\sum_{s=1}^{S} \nabla_\theta \log \pi_\theta(\gamma_s) r_{\gamma_s}$$

where $S$ is the number of actions sampled from the current policy, $\pi$ is the current policy, $\gamma_s$ is the exploration action we take, and $r_{\gamma_s}$ is its reward.
When $G$ is not reached in $T$ steps in any of the sample trajectories, $r_{\gamma_s}$ is always 0, so the Monte Carlo estimate of the policy gradient will be 0.
The exploration depends on the Monte Carlo estimate of the policy gradient. If $G$ is not reached in $T$ steps in any of the sample trajectories, the policy gradient will stay 0, and we will never update the policy. Only if $G$ can be reached, we will have nonzero policy gradient and update the policy accordingly. The optimization will favor policies that can reach $G$ in $T$ steps. □

(B) Suppose the robot must start in $s_0$ for all learning trajectories. Name two ways you could alter the MDP environment to improve exploration and gradients.

*Solution.*

First way: Instead of only assigning reward $R = 1000$ to the position $G$, we can assign rewards to positions according to their distance from $G$. Positions close to $G$ will have higher rewards, and positions far from $G$ will have lower rewards. In this way, we will update the policy even when $G$ is not reached in $T$ steps in any of the sample trajectories.
Second way: Based on the first way, we can also let the reward decay as the number of steps $t$ increases. For example, we can set a decay rate of 0.9 for each step. □