Due Date: April 5th 23:59, 2019

Instructions

- *For all questions, show your work!*
- *Starred questions are **hard** questions, not **bonus** questions.*
- *Please use a document preparation system such as LaTeX, unless noted otherwise.*
- *Unless noted that questions are related, assume that notation and defintions for each question are self-contained and independent*
- *Submit your answers electronically via Gradescope.*
- ***TAs for this assignment are Shawn Tan, Samuel Lavoie, and Chin-Wei Huang.***

This assignment covers mathematical and algorithmic techniques underlying the three most popular families of deep generative models, variational autoencoders (VAEs, Questions 1-3), autoregressive models (Question 4), and generative adversarial networks (GANs, Questions 5-7).

Question 1 (8-8)**.** Reparameterization trick is a standard technique that makes the samples of a random variable differentiable. Consider a random vector $Z \in \mathbb{R}^K$ with a density function $q(\boldsymbol{z}; \phi)$. We want to find a deterministic function $\boldsymbol{g} : \mathbb{R}^K \to \mathbb{R}^K$ that depends on ϕ, to transform a random variable Z_0 having a ϕ-independent density function $q(\boldsymbol{z}_0)$, such that $\boldsymbol{g}(Z_0)$ has the same density as Z. Recall the change of density for a bijective, differentiable $\boldsymbol{g}$:

$$q(\boldsymbol{g}(\boldsymbol{z}_0)) = q(\boldsymbol{z}_0) \left| \det \left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} \right) \right|^{-1} \tag{1}$$

1. Assume $q(\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ and $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$, where $\mu \in \mathbb{R}^K$ and $\sigma \in \mathbb{R}^K_{>0}$. Show that $\boldsymbol{g}(\boldsymbol{z}_0)$ is distributed by $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$ using Equation (1).
2. Assume instead $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \boldsymbol{S}\boldsymbol{z}_0$, where $\boldsymbol{S}$ is a non-singular $K \times K$ matrix. Derive the density of $\boldsymbol{g}(\boldsymbol{z}_0)$ using Equation (1).

Answer 1.

1. Define $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$. Let $\boldsymbol{J}$ be the Jacobian matrix $\boldsymbol{J} = \frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0}$. Note that here $\boldsymbol{J}$ is a diagonal matrix with $\boldsymbol{J}_{ii} = \sigma_i$. Since $\boldsymbol{g}$ has inverse $\boldsymbol{g}^{-1}(\boldsymbol{z}) = (\boldsymbol{z} - \mu)/\sigma$ (where the division is elementwise), we can write $\boldsymbol{z}_0 = \boldsymbol{g}^{-1}(\boldsymbol{g}(\boldsymbol{z}_0)) = (\boldsymbol{z} - \mu)/\sigma$ and $||\boldsymbol{z}_0||^2 = \sum_{i=1}^K \frac{(\boldsymbol{z}_i - \mu_i)^2}{\sigma_i^2}$ where $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$:

$$q(\boldsymbol{z}) = (2\pi)^{-\frac{K}{2}} e^{-\frac{||\boldsymbol{z}_0||^2}{2}} \prod_{i=1}^K \sigma_i^{-1} = \prod_{i=1}^K \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\frac{(\boldsymbol{z}_i - \mu_i)^2}{\sigma_i^2}} \right)$$

   The right hand side is the density of $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$.
2. Here $\boldsymbol{J} = \boldsymbol{S}$. Let $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ be the singular decomposition. Let $\Sigma = \boldsymbol{S}\boldsymbol{S}^\top = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top$. We have $\boldsymbol{g}^{-1}(\boldsymbol{z}) = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\top(\boldsymbol{z} - \mu)$ and

$$||\boldsymbol{z}_0||^2 = \boldsymbol{g}^{-1}(\boldsymbol{z})^\top \boldsymbol{g}^{-1}(\boldsymbol{z}) = (\boldsymbol{z} - \mu)^\top \boldsymbol{U}\boldsymbol{D}^{-2}\boldsymbol{U}^\top(\boldsymbol{z} - \mu) = (\boldsymbol{z} - \mu)^\top \Sigma^{-1}(\boldsymbol{z} - \mu)$$

   where $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{z}_0)$. Again, by the change of variable formula,

$$\begin{aligned} q(\boldsymbol{z}) &= (2\pi)^{-\frac{K}{2}} e^{-\frac{||\boldsymbol{z}_0||^2}{2}} |\det \boldsymbol{S}|^{-1} \\ &= (\det 2\pi \boldsymbol{I}_K)^{-\frac{1}{2}} |\det \boldsymbol{S}\boldsymbol{S}^\top|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{z}-\mu)^\top \Sigma^{-1}(\boldsymbol{z}-\mu)} \\ &= (\det 2\pi\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{z}-\mu)^\top \Sigma^{-1}(\boldsymbol{z}-\mu)} \end{aligned}$$

   which is the density of $\mathcal{N}(\mu, \Sigma)$ where $\Sigma = \boldsymbol{S}\boldsymbol{S}^\top$.

Question 2 (5-5-6)**.** Consider a latent variable model $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ where $\boldsymbol{z} \in \mathbb{R}^K$, and $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$. 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):

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

Let $\mathcal{Q}$ be the family of variational distributions with a feasible set of parameters $\mathcal{P}$; i.e. $\mathcal{Q} = \{q(\boldsymbol{z}; \pi) : \pi \in \mathcal{P}\}$; for example π can be mean and standard deviation of a normal distribution. We assume q_ϕ is parameterized by a neural network (with parameters ϕ) that outputs the parameters, $\pi_\phi(\boldsymbol{x})$, of the distribution $q \in \mathcal{Q}$, i.e. $q_\phi(\boldsymbol{z}|\boldsymbol{x}) := q(\boldsymbol{z}; \pi_\phi(\boldsymbol{x}))$.

1. Show that maximizing the expected complete data log likelihood

   $$\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$$

   for a fixed $q(\boldsymbol{z}|\boldsymbol{x})$, wrt the model parameter θ, gives the maximizer of the biased log marginal likelihood: $\arg\max_\theta\{\log p_\theta(\boldsymbol{x}) + B(\theta)\}$, where $B(\theta)$ is non-positive. Find $B(\theta)$.
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 of $\sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ 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 ?
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)
   (b) from the computational point of view (efficiency)
   (c) in terms of memory (storage of parameters) https://www.overleaf.com/13215018tppysmgzxppx

Answer 2.

1. Using the fact that $q(\boldsymbol{z}|\boldsymbol{x})$ and $p(\boldsymbol{z})$ do not depend on θ and the identity $1 = p_\theta(\boldsymbol{z}|\boldsymbol{x})/p_\theta(\boldsymbol{z}|\boldsymbol{x})$, we have

   $$\begin{aligned}\arg\max_\theta \left\{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]\right\} &= \arg\max_\theta \left\{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{p_\theta(\boldsymbol{z}|\boldsymbol{x})}\frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{q(\boldsymbol{z}|\boldsymbol{x})}\right]\right\} \\ &= \arg\max_\theta \left\{\log p_\theta(\boldsymbol{x}) - D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))\right\}\end{aligned}$$

   That it, maximizer of the expected complete data log likelihood aims at maximizing the marginal likelihood and reducing the KL divergence between $q(\boldsymbol{z}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{z}|\boldsymbol{x})$; the latter gap causes the bias.
2. Denote by $\mathcal{L}[q(\boldsymbol{z})]$ the lower bound corresponding to the variational distribution $q(\boldsymbol{z})$. Due to the finiteness of the recognition model, the distributions q_ϕ can represent is a subset of $\mathcal{Q}$. Given q_i^* the maximizer of the ELBO (argmax over the family $\mathcal{Q}$), the inference gap of q_{ϕ^*} can be decomposed as the sum of two positive terms:

   $$\log p_\theta(\boldsymbol{x}_i) - \mathcal{L}[q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)] = (\log p_\theta(\boldsymbol{x}_i) - \mathcal{L}[q_i^*(\boldsymbol{z})]) + (\mathcal{L}[q_i^*(\boldsymbol{z})] - \mathcal{L}[q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)])$$

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.

Using the equality $\log p_\theta(\boldsymbol{x}) - \mathcal{L}[q(\boldsymbol{z}|\boldsymbol{x})] = D_{\mathrm{KL}}(q(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$, we have

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

The last term is non-negative since q_i^* is the maximizer among $\mathcal{Q}$, so the KL on the LHS is no smaller than the KL on the RHS.

3. (a) Since the only term in the ELBO that depends on θ is the expected complete data log likelihood, there is a bias due to the KL divergence. As demonstrated by the last question, the recognition model cannot do better than the optimal q_i^*.

   (b) Computationally, the point of using a recognition model is to amortize the cost of inference, so that one does not need to update q_i for each data point $\boldsymbol{x}_i$ until the approximation is good enough ; instead one can simply use the output of the recognition model as a reasonable approximation, since the encoder, which is normally jointly trained with the decoder, is constantly updated.

   (c) In terms of number of parameters, amortization allows for $\mathcal{O}(1)$ storage (considering a fixed size recognition model) instead of learning a new $q_i(\boldsymbol{z})$ for each $\boldsymbol{x}_i$, which is $\mathcal{O}(n)$.

Question 3 (6-6)**.** Since variational inference provides a lower-bound on the log marginal likelihood of the data, it gives us a biased estimate of the marginal likelihood. Therefore, methods of "tightening" the bound (i.e. finding a higher valid lower bound) may be desirable.

Consider a latent variable model with the joint $p(\boldsymbol{x}, \boldsymbol{h})$ where $\boldsymbol{x}$ and $\boldsymbol{h}$ are the observed and unobserved random variables, respectively. Now let $q(\boldsymbol{h})$ be a variational approximation to $p(\boldsymbol{h}|\boldsymbol{x})$. Define

$$\mathcal{L}_K = \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\log \frac{1}{K}\sum_{j=1}^{K} \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]$$

Note that $\mathcal{L}_1$ is equivalent to the evidence lower bound (ELBO).

1. Show that $\mathcal{L}_K$ is a lower bound of the log marginal likelihood $\log p(\boldsymbol{x})$.
2. Show that $\mathcal{L}_K \geq \mathcal{L}_1$; i.e. $\mathcal{L}_K$ is a family of lower bounds tighter than the ELBO.

Answer 3.

1. Applying Jensen's inequality and linearity of expectation gives

$$\mathbb{E}\left[\log \frac{1}{K}\sum_{j=1}^{K} \frac{p_j}{q_j}\right] \leq \log \frac{1}{K}\sum_{j=1}^{K} \mathbb{E}_{q_j}\left[\frac{p_j}{q_j}\right] = \log \frac{1}{K}\sum_{j=1}^{K} \int p(\boldsymbol{x}, \boldsymbol{h}_j) d\boldsymbol{h}_j = \log p(\boldsymbol{x})$$

2. Note that $a = \frac{1}{K}\sum_{j=1}^{K} a_j$ can be viewed as the expectation over a uniform measure ; i.e. $j \sim \mathrm{Unif}([1, \cdots, K])$ and $a = \mathbb{E}_j[a_j]$. Now we can apply Jensen's Inequality to obtain the lower bound:

$$\mathbb{E}\left[\log \frac{1}{K}\sum_{j=1}^{K} \frac{p_j}{q_j}\right] = \mathbb{E}\left[\log \mathbb{E}_j\left[\frac{p_j}{q_j}\right]\right] \geq \mathbb{E}\left[\mathbb{E}_j\left[\log \frac{p_j}{q_j}\right]\right] = \mathbb{E}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{h})}{q(\boldsymbol{h})}\right]$$

since $\boldsymbol{h}_j$'s are identically distributed.

Question 4 (5-5-5-5)**.** One way to enforce autoregressive conditioning is via masking the weight parameters.[2] Consider a two-layer convolutional neural network without kernel flipping, with kernel size 3×3 and padding size 1 on each border (so that an input feature map of size 5×5 is convolved into a 5×5 output). Define mask of type A and mask of type B as

$$(\boldsymbol{M}^A)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j < 2 \\ 0 & \text{elsewhere} \end{cases} \qquad (\boldsymbol{M}^B)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

where the index starts from 1. Masking is achieved by multiplying the kernel with the binary mask (elementwise). Specify the receptive field of the output pixel that corresponds to the third row and the third column (index 33 of Figure 1) in each of the following 4 cases:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 1 – 5×5 convolutional feature map.

1. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
2. If we use $\boldsymbol{M}^A$ for the first layer and $\boldsymbol{M}^B$ for the second layer.
3. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^A$ for the second layer.
4. If we use $\boldsymbol{M}^B$ for the first layer and $\boldsymbol{M}^B$ for the second layer.

Answer 4. See Figure 2

11	**12**	**13**	**14**	**15**
21	**22**	**23**	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

11	**12**	**13**	**14**	**15**
21	**22**	**23**	**24**	25
31	**32**	33	34	35
41	42	43	44	45
51	52	53	54	55

11	**12**	**13**	**14**	**15**
21	**22**	**23**	**24**	25
31	**32**	33	34	35
41	42	43	44	45
51	52	53	54	55

11	**12**	**13**	**14**	**15**
21	**22**	**23**	**24**	25
31	**32**	**33**	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 2 – Receptive field under different masking schemes.

Question 5 (10)**.** Let P_0 and P_1 be two probability distributions with densities f_0 and f_1 (respectively). This problem demonstrates 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 $f_0(\boldsymbol{x})$.

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

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x} \sim P_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim P_0}[\log(1 - D(\boldsymbol{x}))]$$

2. An example of this is the use of masking in the Transformer architecture (Problem 3 of TP2 practical part).

Answer 5. The given function to be maximized can be expressed as a functional of D, $G[D] := \int g(D(\boldsymbol{x}), \boldsymbol{x}) d\boldsymbol{x}$ where

$$g(D(\boldsymbol{x}), \boldsymbol{x}) := f_1(\boldsymbol{x}) \log D(\boldsymbol{x}) + f_0(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))$$

Setting the functional derivative to be zero yields

$$\frac{\delta G[D]}{\delta D} = \frac{\partial g(D(\boldsymbol{x}), \boldsymbol{x})}{\partial D} = \frac{f_1(\boldsymbol{x})}{D(\boldsymbol{x})} - \frac{f_0(\boldsymbol{x})}{1 - D(\boldsymbol{x})} = 0$$

solving which gives us $D^*(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})}$. Thus we can use D^* to estimate the density of f_1 by rearranging the terms, which yields $f_1 = f_0 D^*/(1 - D^*))$.

Question 6 (5-5-6)**.** While generative adversarial networks were originally formulated as minimizing the Jensen-Shannon (JS)-divergence, the framework can be generalized to use other divergences, such as the Kullback–Leibler (KL)-divergence. In this exercise we see how KL can be approximated (bounded from below) via a function $T : \mathcal{X} \to \mathbb{R}$ (i.e. the discriminator). Let q and p be probability density functions and recall the definition of the KL divergence $D_{\mathrm{KL}}(p||q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$.

*1. Let $R_1[T] := \mathbb{E}_p[T(x)] - E_q[e^{T(x)-1}]$.
   (a) The convex conjugate of a function $f(u)$ is defined as $f^*(t) = \sup_{u \in \mathrm{dom} f} ut - f(u)$. Show that the convex conjugate of $f(u) = u \log u$ is $f^*(t) = e^{t-1}$, and its biconjugate[3], i.e. the convex conjugate of its convex conjugate, is $f^{**}(u) := (f^*)^*(u) = u \log u$.
   (b) Use the fact found above to show that $D_{\mathrm{KL}}(p||q) = \sup_T R_1[T]$, where the supremum is taken over the set of all (measurable) functions $\mathcal{X} \to \mathbb{R}$. Start from the following step

$$\sup_{T(x)} \int p(x)T(x) - q(x)e^{T(x)-1} dx = \int \sup_{t \in \mathbb{R}} p(x)t - q(x)e^{t-1} dx$$

   which you don't need to prove.

*2. Let $r(x) = e^{T(x)}/\mathbb{E}_q[e^{T(x)}]$ and $R_2[T] := \mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}]$.
   (a) Verify that rq is a proper density function, i.e. integrating to 1.
   (b) Show that $D_{\mathrm{KL}}(p||q) \geq R_2[T]$, with equality if and only if $T(x) = \log(p(x)/q(x)) + c$ where c is a constant independent of x.

3. Compare the two representations of the KL divergence. For fixed $T(x)$, $p(x)$ and $q(x)$, which one of $R_1[T]$ and $R_2[T]$ is greater than or equal to the other ?

Answer 6.

1. This problem can be solved using the functional derivative. Alternatively, we use the technique of *convex conjugate*. Since the stationary point of $ut - u \log u$ is $u^* = e^{t-1}$, so the maximizer is $f^*(t) = e^{t-1}$. Likewise, the stationary point of $tu - e^{t-1}$ is $1 + \log u$, so the maximizer is $f^{**}(u) = u \log u$. Now, rewrite $R_1[T]$ as

$$\begin{aligned}
&\sup_{T(x)} \int p(x)T(x) - q(x)e^{T(x)-1} dx \\
&= \int \sup_t p(x)t - q(x)e^{t-1} dx \\
&= \int q(x) \sup_t \left(\frac{p(x)}{q(x)} t - e^{t-1} \right) dx
\end{aligned}$$

3. More generally, the biconjugate of f is equal to itself if f is a lower semi-continuous convex function (this is known as the **Fenchel-Monreau Theorem**).

Thus, setting $u = p(x)/q(x)$ in the above integral yields

$$\int q(x)\frac{p(x)}{q(x)}\log\frac{p(x)}{q(x)}dx = D_{\mathrm{KL}}(p||q)$$

This is known as the **f-divergence** representation of the KL divergence.

2. By definition, $q(x)r(x)$ is a normalized density function, since

$$\int q(x)r(x)dx = \frac{1}{\mathbb{E}_q[e^{T(x)}]}\int q(x)e^{T(x)}dx = \frac{\mathbb{E}_q[e^{T(x)}]}{\mathbb{E}_q[e^{T(x)}]} = 1$$

Now write $R_2[T]$ as

$$\mathbb{E}_p[T(x)] - \log\mathbb{E}_q[e^{T(x)}] = \mathbb{E}_p[\log e^{T(x)}] - \log\mathbb{E}_q[e^{T(x)}] = \mathbb{E}_p[\log e^{T(x)}/\mathbb{E}_q[e^{T(x)}]]$$

which is equal to $\mathbb{E}_p[\log r(x)]$. The gap between the KL and the lower bound is $D_{\mathrm{KL}}(p||q) - \mathbb{E}[\log r(x)] = \mathbb{E}_p[\log p(x) - \log q(x)r(x)]$. Since qr is a normalized density function, the gap is equal to $D_{\mathrm{KL}}(p||qr)$, which is non-negative. Furthermore, the KL divergence is zero if and only if $p = qr$, which means $r(x) = p(x)/q(x)$, and that $e^{T(x)} \propto p(x)/q(x)$. As a result, the optimal $T(x)$ is equal to the log likelihood ratio $\log(p(x)/q(x))$ up to some constant c.

3. We can express the difference as

$$R_1 - R_2 = \log\mathbb{E}_q[e^{T(x)}] - \mathbb{E}_q[e^{T(x)-1}] = \log\mathbb{E}_q[e^{T(x)-1}] + 1 - \mathbb{E}_q[e^{T(x)-1}] \leq 0$$

since the elementary inequality $t - 1 - \log t \geq 0$ holds for any $t > 0$. Thus $R_1 \leq R_2$.

Question 7 (10)**.** Let $q, p : \mathcal{X} \to [0, \infty)$ be probability density functions with disjoint (i.e. non-overlapping) support ; more formally, $\{x \in \mathcal{X} : p(x) > 0 \text{ and } q(x) > 0\} = \varnothing$. What is the Jensen Shannon Divergence (JSD) between p and q ? Recall that JSD is defined as $D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||r) + \frac{1}{2}D_{KL}(q||r)$ where $r(x) = \dfrac{p(x) + q(x)}{2}$.

Answer 7. By definition $D_{\mathrm{JS}}(p||q) = \frac{1}{2}\mathbb{E}_p[\log p - \log r] + \frac{1}{2}\mathbb{E}_q[\log q - \log r]$. The first KL is

$$\mathbb{E}_p[\log p] - \mathbb{E}_p[\log(p+q)/2] = \mathbb{E}_p[\log p] - \mathbb{E}_p[\log(p+q) - \log 2] = \log 2$$

since p and q has non-overlapping support. Likewise the second KL is also $\log 2$, so $D_{\mathrm{JS}}(p||q) = \log 2$.