CS5787: Exercises 3

https://github.com/mitkrieg/dl-assignment-3

Mitchell Krieger
mak483@cornell.edu

1 Theory: Question 1 [40 pts]

a) $\theta \neq 0$

- KL Divergence:
  When $\theta = 0$ there is no overlap between P and Q. This means that whenever P is greater than zero, Q is zero. Applying this to the KL formula at $x = 0$ we get

$$D_{KL}(P||Q) = P(0) \log \frac{P(0)}{0} dx = \infty \tag{1}$$

  We can see from the above that because the logarithm in the equation is undefined, we get a value for KL Divergence of ∞.

- JS Divergence:
  The JS Divergence considers the KL Divergence between each distribution and the average of the two distributions. The average distribution $\frac{P+Q}{2}$ will split the distribution to be half at $x = 0$ and half at $x = \theta$. Therefore, The KL Divergence between P and the average will be:

$$D_{KL}(P||\frac{P+Q}{2}) = 1 \cdot \log(\frac{1}{0.5}) = \log(2) \tag{2}$$

  $D_{KL}(Q||\frac{P+Q}{2})$ will give the same result. So applying the JS Divergence formula:

$$D_{JS}(P||Q) = \frac{1}{2} \cdot \log(2) + \frac{1}{2} \cdot \log(2) = \log(2) \tag{3}$$

- Wasserstein Distance:
  The Wasserstein Distance is θ because both distributions have the same uniform distribution between [0,1] for y. Because they are identical except for the x-value of P is 0 and the x-value of Q is θ, the cost of moving P to Q would be θ.

b) $\theta = 0$

- KL Divergence:
  Zero because the distributions are the same

$$\begin{aligned} D_{KL}(P||Q) &= \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{P(x)} dx \\ &= \int_{-\infty}^{\infty} P(x) \log 1 \\ &= \int_{-\infty}^{\infty} P(x) 0 \\ &= 0 \end{aligned} \tag{4}$$

- JS Divergence:
  $\frac{P+Q}{2} = P = Q$ because both distributions are uniform between [0,1] for y and

centered at $x = 0$. So the JS Divergence is zero because the distributions are the same

$$\begin{aligned} D_{JS}(P||Q) &= \frac{1}{2}D_{KL}(P||\frac{P+Q}{2}) + \frac{1}{2}D_{KL}(Q||\frac{P+Q}{2}) \\ &= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0 \end{aligned} \tag{5}$$

- Wasserstein Distance: Zero because both distributions are uniform between [0,1] for y and centered at $x = 0$. So there is no cost to transform P into Q.

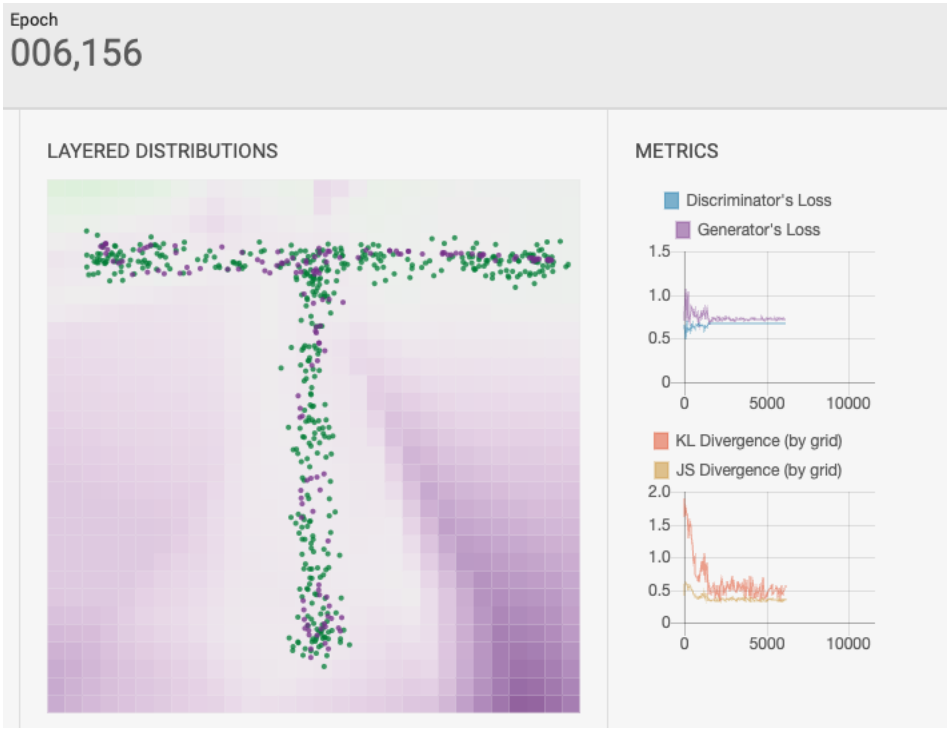
c) The advantage of the Wasserstein Distance is that it is both interpretable and stable when compared to KL and JS Divergence. When $\theta \neq 0$, KL is infinite and JS is $\log 2$, where as Wasserstein is θ. In this context, θ is a much clearer measure of the distances between the two distribution because it is simply the distance between the x-values of the distributions. In addition, both KL and JS have discontinuous jumps from when they overlap to when then don't, KL becomes infinite and JS although still finite jumps from 0 to $\log 2$. Wasserstein distance on the other hand, increases linearly as θ grows. These are big advantages because it would make the gradient in training a model more stable because it is continuous and not super sensitive to small changes (from not overlapping to overlapping).

2 Theory: Question 2 [10 pts]

LSTMs will process inputs sequentially. So at each time step, the LSTM will have to multiply the input by the hidden state, the previous hidden state by the current hidden state, and other smaller operations like the activation functions for the gates and addition for adding bias and cell state updates. The dominant operation here is the multiplication of the hidden states. This operation has a time complexity of $O(d^2)$ where d is the dimension of the hidden state. Because we do this at every time step t for all N, the total complexity is $O(d^2N)$. However, the size of the hidden state is a constant value, and not a function of the size of the input, so the total time complexity is linear $O(N)$.

Transformers on the other hand use self-attention, which requires every value of the input to be compared to all other values in the input. This is an $O(N^2)$ operation. The other components of transformers, such as the feed forward network, encoder/decoder and positional encoding are all generally $O(N)$, so they get dominated by the self-attention mechanism's computational complexity. So the total time complexity is quadratic at $O(N^2)$

3 Practical: Question 3 [5 pts]



4 Practical: Question 4 [20 pts]

5 Practical: Question 5 [25 pts]