How does the GAN make $\boldsymbol{p}_{\textbf{data}} \approx \boldsymbol{p}_{\textbf{model}}$?

The Generator Loss function we constructed is a proxy to achieve the goal

$$p_{\text{model}} \approx p_{\text{data}}$$

That is: the distribution of samples produced by the Generator is (approximately) the same as the "true" distribution

- we note that we don't know the "true" p_{data}
    - we only have available a sample and those the training set defines an *empirical* distribution

There are several ways to quantify

$$p_{\mathrm{model}} \approx p_{\mathrm{data}}$$

One choice would be the minimization of KL Divergence

- $\mathbf{KL}(p_{\mathrm{data}} || p_{\mathrm{model}})$

Review: definition of the KL Divergence

Mathematically:

$$
\begin{array}{lll}
\text{KL}(p||q) & = & - \sum_x p(x) \log q(x) + \sum_x p(x) \log p(x) \\
& = & \sum_x p(x) * (\log p(x) - \log q(x)) \\
& = & \mathbb{E}_{\mathbf{x} \sim p} (\log p(x) - \log q(x))
\end{array}
$$

You can see that it is

- the point-wise difference between the (log) probability of $\mathbf{x}$ in distributions p and q
- averaged over the distribution of $\mathbf{x} \sim p$

and thus is a point-wise measure of the dis-similarity of the two distributions.

As a reminder:

We now show that using this as a loss function

- results in a estimation of the model distribution p_{model}
- that is the Maximum Likelihood estimator of the training examples (represented by p_{data})

Choose p_{model} to Minimize

$$
\begin{array}{lll}
\mathbf{KL}(p_{\text{data}} || p_{\text{model}}) & = & \int_{\mathbf{x}} p_{\text{data}}(\mathbf{x}) \left(\log \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{model}}\mathbf{x})} \right) d\mathbf{x} & \text{Definition of KL} \\
& = & \mathbb{E}_{\mathbf{x} \in p_{\text{data}}} \log(p_{\text{data}}(\mathbf{x})) - \log(p_{\text{model}}(\mathbf{x})) & \text{Definiton of log} \\
\text{minimizing KL} & & & \\
& \approx & \mathbb{E}_{\mathbf{x} \in p_{\text{data}}} - \log(p_{\text{model}}(\mathbf{x})) & \text{Since } \log(p_{\text{data}}(
\end{array}
$$

So minimizing $\mathbf{KL}$ is equivalent to

- minimizing the Negative Log Likelihood
- in other words: *maximizing* the Log Likelihood

Notice that the expectation is over the "true" distribution p_{data}.

The expectation is certainly reasonable for training put perhaps not best for the purposes of generating synthetic data

- Measures fidelity to training data
- NOT how "realistic" the synthetic data is
- the penalty for p_{model} placing large probability mass around a particular $\hat{\mathbf{x}}'$ is small when $p_{\text{data}}(\hat{\mathbf{x}}') \approx 0$
    - so Generator may create large quantity of synthetic data that is improbable given the training set

If we knew the true p_{data}, a better objective to minimize for the purpose of generating synthetic data would be the similar

$$\mathbf{KL}(p_{\text{model}}||p_{\text{data}})$$

which is equivalent to maximizing

$$\mathbb{E}_{\mathbf{x} \in p_{\text{model}}} - \log(p_{\text{data}}(\mathbf{x}))$$

The expectation is over the synthetic data, not the true data

- $\log(p_{\text{data}}(\mathbf{x}))$ is defined as log of Perplexity
    - an element of "surprise" in seeing $\mathbf{x}$
- So the expectation asks: for each synthetic datum generated, how likely is it to occur in the true distribution ?

This is merely a theoretical argument

- In practical terms: we only have empirical p_{data}

So can't evaluate

- log Perplexity
$$p_{\text{data}}(\hat{\mathbf{x}})$$
for $\hat{\mathbf{x}} \in p_{\text{model}}$
- unless synthetic $\hat{\mathbf{x}}$ replicates a sample in the training data

Jensen-Shannon Divergence

We have observed that the KL divergence is *not* symmetric

$$\mathbf{KL}(P||Q) \neq \mathbf{KL}(Q||P)$$

because the expectations are taken over different distributions.

An alternative measure of similarity of two distributions is the Jensen-Shannon Divergence (JSD)

$$\begin{aligned} \mathrm{JSD}(P||Q) &= \mathrm{JSD}(Q||P) \\ &= \frac{1}{2}\,\mathrm{KL}\left(P\,||\,\frac{P+Q}{2}\right) + \\ &\quad \frac{1}{2}\,\mathrm{KL}\left(Q\,||\,\frac{P+Q}{2}\right) \end{aligned}$$

This measure is

- symmetric
- is a kind of mixture of $\mathbf{KL}(P||Q)$ and $\mathbf{KL}(Q||P)$.

[Huszar (https://arxiv.org/pdf/1511.05101.pdf)](https://arxiv.org/pdf/1511.05101.pdf) has a Generalized JSD which interpolates between the two terms

$$\begin{aligned} \mathrm{JSD}_\pi(P||Q) &= \mathrm{JSD}(Q||P) \\ &= \pi \, \mathrm{KL} \left(\, P \, || \, \pi P + (1-\pi)Q \, \right) + \\ & \quad (1-\pi) \, \mathrm{KL} \left(\, Q \, || \, \pi P + (1-\pi)Q \, \right) \end{aligned}$$

The Generalized JSD

- **Not** symmetric although
  $\mathrm{JSD}_\pi(P||Q) = \mathrm{JSD}_{1-\pi}(Q||P)$

Huszar shows that, for small values of π

$$\frac{\mathrm{JSD}_{\pi}(P||Q)}{\pi} \approx \mathrm{KL}\left(\, P \,||\, Q\right)$$

and

$$\frac{\mathrm{JSD}_{1-\pi}(P||Q)}{1-\pi} \approx \mathrm{KL}\left(\, Q \,||\, P\right)$$

In the first case

- $\mathrm{JSD}_{\pi}(P||Q)$ is proportional to Maximum Likelihood

In the second case

- $\mathrm{JSD}_{1-\pi}(P||Q)$ is proportional to $\mathrm{KL}\left(\, Q \,||\, P\right)$

In implementing Generalized JSD

- The Discriminator is trained (as usual) on a mix of real and fake examples
    - But *not* in equal numbers
    - π is fraction of samples from Q
    - $(1 - \pi)$ is fraction of samples from P
    - $\pi < \frac{1}{2}$: real samples over represented
    - $\pi > \frac{1}{2}$: biased toward Q
- Explains why we often see training with Generator updated twice for each update of Discriminator ?

Adversarial Training and the Jensen-Shannon Divergence

The Discriminator Loss $\mathcal{L}_D$

- summed over all examples
  - (ignoring the $\frac{1}{2}$ from the previous presentation where we assumed equal number of Real and Fake)

is

$$\mathcal{L}_D \quad = \quad - \left(\mathbb{E}_{\mathbf{x^{(i)}} \in p_{\text{data}}} \log D(\mathbf{x^{(i)}}) + \mathbb{E}_{\mathbf{x^{(i)}} \in p_{\text{model}}} \log \left(1 - D(\mathbf{x^{(i)}})\right) \right) \quad D(G(\mathbf{z})) = \mathbf{x}^{(}$$

We also showed that the optimal Discriminator results in

$$D^*(\mathbf{x}) = \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{model}}(\mathbf{x}) + p_{\text{data}}(\mathbf{x})}$$

Plugging $D^*(\mathbf{x})$ into $\mathcal{L}_D$ (Goodfellow Equation):

$$
\begin{aligned}
\mathcal{L}_D &= -\left(\mathbb{E}_{\mathbf{x^{(i)}}\in p_{\text{data}}} \log \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{model}}(\mathbf{x})+p_{\text{data}}(\mathbf{x})} + \mathbb{E}_{\mathbf{x^{(i)}}\in p_{\text{model}}} \log \frac{p_{\text{model}}(\mathbf{x})}{p_{\text{model}}(\mathbf{x})+p_{\text{data}}(\mathbf{x})}\right) \\
&= -\left(\mathbf{KL}(p_{\text{data}}||p_{\text{model}}(\mathbf{x}) + p_{\text{data}}(\mathbf{x})) + \mathbf{KL}(p_{\text{model}}||p_{\text{model}}(\mathbf{x}) + p_{\text{data}}(\mathbf{x})\right) \\
&= -\left(\log 4 + \mathbf{KL}(p_{\text{data}}||\frac{p_{\text{model}}(\mathbf{x})+p_{\text{data}}(\mathbf{x})}{2}) + \mathbf{KL}(p_{\text{model}}||\frac{p_{\text{model}}(\mathbf{x})+p_{\text{data}}(\mathbf{x})}{2}\right)
\end{aligned}
$$

$$
= -\left(\log 4 + 2 * \mathrm{JSD}(p_{\text{data}}||p_{\text{model}})\right)
$$

The above equations shows that

- minimizing KL Divergence (second line above)
- under the assumption that the Discriminator can train to be the **optimal** adversary

results in $\mathcal{L}_D$ becoming equivalent to Jensen-Shannon Distance (last line above)

So solving the minimax optimally minimizes the JSD divergence between p_{data} and p_{model}.

In [2]:

```
print("Done")
```

```
Done
```