Towards Principled Methods For Training Generative Adversarial Networks

Introduction

> Always, g_Θ is a neural network parameterized by Θ, and the main difference is how g_Θ is trained

$$KL(P_r||P_g) = \int P_r(x)log$$

> This cost function is not symetrical between P_r and P_g:
> If $P_r(x) > P_g(x)$, then x is a point with higher probability of coming from the data than being a generated sample. This is the core of the phenomenon commonly described as **mode dropping**. It is important to note that when $P_r(x) > 0$ but $P_g(x) \to 0$, the integrand inside the KL grows quickly to infinity, meaning that this cost function assigns an extremely high cost to a generator's distribution not covering parts of the data.

> If $P_r(x) < P_g(x)$, then x has low probability of being a data point, but high probability of being generated by our model. This is the case when we see our generator outputting an image that doesn't look real. In this case, when $P_r(x) \to 0$ and $P_g(x) > 0$, we see that the value inside the KL goes to 0, meaning that this cost function will **pay extremely low cost for generating fake looking samples**.

$$JSD(P_r \parallel P_g) =$$

$$L(D^*, g_\Theta) = 2JSD(P_r \parallel P_g) - 2\log 2$$

> We will prove that as the approximation gets better, either we see vanishing gradients or the massively unstable behaviour we ses in practice, depending on which cost function we use.
> **In practice, as the discriminator gets better, the updates to the generator get consistently worse.**

Sources Of Instability

> The theory tells us that the trained discriminator will have cost at most $2\log 2 - 2JSD(P_r \parallel P_g)$. However, in practice, if we just train D till convergence, its error will go to 0. **The only way this can happen is if the distributions are not continuous, or they have disjoint supports.**

> **One possible cause for the distributions not to be continuous is if their supports lie on low dimensional manifolds.**

> **Lemma 1. Let $g : Z \to X$ be a function composed by affine transformations and pointwise nonlinearities, which can either be rectifiers, leaky rectifiers, or smooth strictly increasing functions (such as the sigmoid, tanh, softplus, etc). Then, $g(Z)$ is contained in a countable union of manifolds of dimension at most dim Z. Therefore, if the dimension of Z is less than the one of X, $g(Z)$ will be a set of measure 0 in X.**

> If two manifolds don't perfectly align, their intersection $L = M \bigcap P$ will be a finite union of manifolds with dimensions strictly lower than both the dimension of M and the one of P.

> These two theorems tell us that there are perfect discriminators which are smooth and constant almost everywhere in M and P. The fact that the discriminator is constant in both manifolds points to the fact that we won't really be able to learn anything by backproping through it.