Wasserstein GAN

Reading group paper by Bun

1. Motivation

Say we want to learn the probability distribution, $P_r(x)$, of our data.

Why do this?

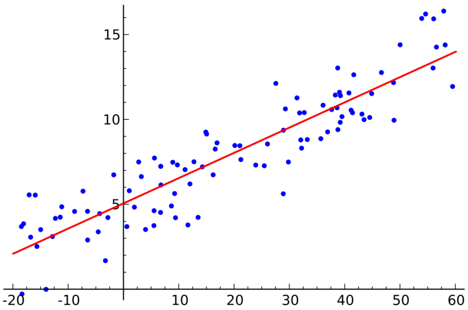
- Can generate new samples from distribution
- Find likelihood of samples (measure uncertainty)
- Learn full conditional distributions (not pointwise estimates)

So lets learn the density via maximum likelihood?

$$\max_{\theta \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$

Where P_θ is the distribution of some parametrized density and $x^{(i)}$ are samples from $P_r(x)$. This can be shown to be equivalent to minimising the $D_{KL}\left(P_r \| P_\theta\right)$.

What issues might we run into when doing this? P_r density might not be continuous (manifold hypothesis, high dimensional data we usually work with might lie on lower dimension manifolds). Or maybe try learn a (smooth parameterised) mapping from a lower dimensional distribution, Z, to P_r. Since both Z and P_r are projections of low dimensional manifolds into a higher dimensional space, they are most likely disjoint. $D_{KL}\left(P_r \| P_\theta\right)$ between disjoint sets is infinite. Usually practitioners add noise to data while training to overcome this (density now continuous).



However, this leads to blurry images as significant amounts of noise need to be added to overcome both issues.

2. Other distances

So we found that D_{KL} is not a great divergence to optimise. What other options are there?

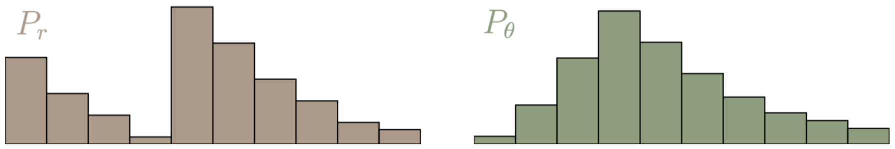
- $D_{JS} = D_{KL}\left(P_r \middle\| \frac{P_\theta+P_r}{2}\right)+D_{KL}\left(P_\theta \middle\| \frac{P_\theta+P_r}{2}\right)$, traditional GAN metric
- Earth-Mover/Wasserstein distance (explained later)

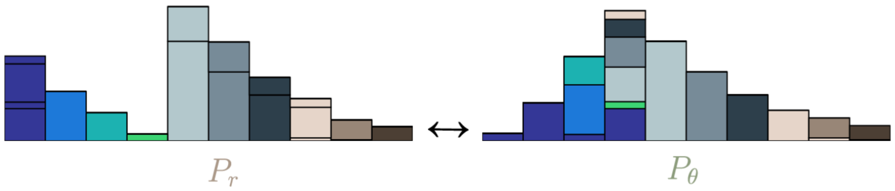
2.1 What is the Wasserstein distance

So what is this distance?

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\, \|x - y\| \,\right]$$

Intuitively, γ defines the path one distribution must take to move its mass from P_g to P_r. Lets look at a discrete example.

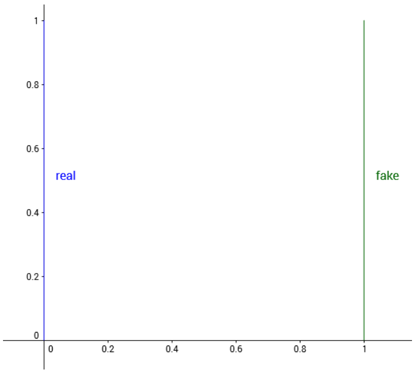




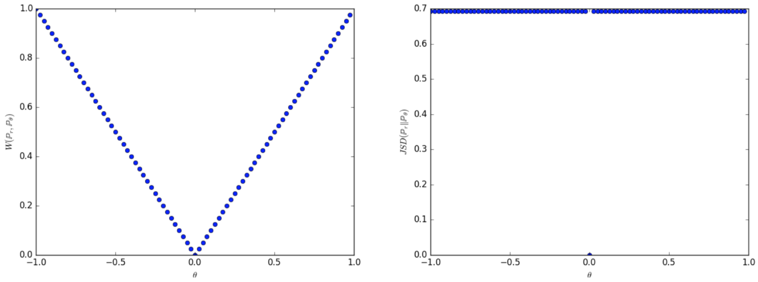
The path cost is calculated by summing the amount of mass, m, and the distance that needs to be travelled, d. The minimum path cost defines the wasserstein distance between both distributions.

2.2 Comparison

How do they perform when trying to minimise the distance/divergence of two disjoint distributions? Let P_0 be the distribution of (0,Z) where $Z \sim U[0, 1]$ and P_θ be (θ,Z) where θ is a single variable parameter.



$$
\begin{aligned}
W(\mathbb{P}_0, \mathbb{P}_\theta) &= |\theta|, \\
JS(\mathbb{P}_0, \mathbb{P}_\theta) &= \begin{cases} \log 2 & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}
\end{aligned}
$$



As can be seen in this example, we can optimise P_θ to converge to P_0 with gradient descent under the Wasserstein distance but not the standard GAN JS divergence. It can be shown that the Wasserstein distance is continuous and diffrentiable almost everywhere if we are using a smooth parameterised function from a lower dimensional distribution, Z, to P_r.

3. How to calculate this distance?

We saw earlier that the distance was:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\big[\, \|x - y\| \,\big]$$

This isn't tractable as finding the minimum path cost grows exponentially with dimensions. The an insight of the paper is that the wasserstein distance has a dual representation.

$$W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]$$

Where f is a critic that maps $\chi \to \mathbb{R}$. Note, it is no longer trying to classify real or generated inputs, just maximising the value. In order to maintain the Lipschitz constraint, the weights of the critic are clipped. Once the critic is trained to convergence, we can backpropogate the wasserstein gradient through the generator. This process of converge critic and update generator is repeated until convergence.

One benefit of this procedure, is the discriminator/critic can be trained to convergence. As we saw with the Uniform example, the JS divergence frequently has zero gradient if the discriminator trains to optimality. This does not occur with WGAN and the gradient estimate improves as the critic is trained even more.

4. Interesting points

- Wasserstein distance is found to be very correlated with image quality but cannot be compared between different critic architectures.
- Hard to mode collapse (no more min-max game or balancing generator and discriminator capacity)