Wasserstein loss.

more directly minimize distance between data and model distributions

→ make winning a turn in the game correlate with actually reducing distance between generated and true distributions.

Recall Jensen-Shannon divergence is

$$JS(P\|Q) = \frac{1}{2} D_{KL}(P\|M) + \frac{1}{2} D_{KL}(Q\|M) \quad \text{where } M = \frac{1}{2}(P+Q)$$

which was implicit objective in original minimax GAN.

and also $D_{KL}(\cdot\|\cdot)$ was objective in ML-GAN.

Now consider Wasserstein distance instead (comes from optimal transportation theory)

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

where $\Pi(P,Q)$ is set of all joint distributions over x and y such that marginals are P and Q, and $\gamma(x,y)$ can be seen as amount of mass that must be moved from x to y to transform P to Q.

Wasserstein distance is cost of optimal transport plan.

Note that Wasserstein distance is continuous and almost differentiable everywhere, which allows model training to get to optimality.

JS divergence locally saturates as discriminator gets better, so gradients become zero not so with Wasserstein.

One difficulty is the inf in the definition of Wasserstein distance: can't handle it computationally.

Can use the Kantorovich - Rubinstein duality:

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

so $W(P,Q)$ is supremum over all 1-Lipschitz functions.

K-Lipschitz Continuity: given two metric spaces (X, d_X) and (Y, d_Y) the map $f: X \mapsto Y$ is K-Lipschitz continuous if

$$d_Y(f(x_1), f(x_2)) \leq K\, d_X(x_1, x_2) \quad \text{for all } x_1, x_2 \in X.$$

when a function is K-Lipschitz we end up with $K\,W(P,Q)$.

Now if we have a family of parameterized functions $\{f_w\}_{w\in W}$ that are K-Lipschitz we get:

$$W(P,Q) \propto \max_{w\in W} \mathbb{E}_{x\sim P}[f_w(x)] - \mathbb{E}_{z\sim Z}[f_w(g_\theta(z))]$$

and so we can implement using neural networks.

Objective function of generator becomes

$$L_{gen}(w) = \min_\theta \mathbb{E}_{x\sim P}[f_w(x)] - \mathbb{E}_{z\sim Z}[f_w(g_\theta(z))]$$

$$= \min_\theta - \mathbb{E}_{z\sim Z}[f_w(g_\theta(z))]$$

so generator tries to minimize Wasserstein distance rather than fooling discriminator. In a sense, discriminator becomes a 'critic'.