Towards Principled Methods for Training GANs

1. Discriminator reaches zero loss, ~~ruins generator~~ :(
2. Ruins generator Gradients :(
3. Adding continuous noise helps :)

0. Preliminaries

- g_θ generator D discriminator
- $g_\theta : Z \to X$
- $z \sim p(z)$
- P_g vs P_r ← real dist.
- ~~$P_{data}(x)$~~ $\mathbb{P}_r$
- $L(D, g_\theta) = \mathbb{E}_{P_r}[\log D(x)] + \mathbb{E}_{P_g}[\log(1 - D(x))]$
- Only need 2nd term for training g_θ

1. Discriminator Reaches 0 loss

Key – P_r and P_g are on low dim manifolds
– telling manifolds apart is easy

Lemma: $g : Z \to X$, g is "nice", $g(Z)$ is countable union of manifolds with dimension $\leq \dim(Z)$

PF: not obvious. Peano curves are counter example

g = Affine maps + Diffeomorphisms ✓

Lemma: w.h.p 2 manifolds intersect in lower dimensional manifolds

Thm: perfect discriminator: force it to be smooth on the manifolds-

$D^*(x) = 0$ or 1 at all important places

$\nabla_x D^*(x) = 0$

2. Sad Generator

– Investigate $\nabla_\theta \mathbb{E}_{z \sim p(z)}[\log(1 - D(g_\theta(z)))]$

$$\nabla_\theta \mathbb{E}_z[\log(1 - D(g_\theta(z))] \leq \mathbb{E}_z\left[\frac{\|\nabla_\theta D(g_\theta(z))\|^2}{|1 - D(g_\theta(z))|^2}\right]$$

$$\leq \mathbb{E}_z\left[\frac{\|\nabla_x D(g_\theta(z))\|^2 \, \|J_\theta g_\theta(z)\|^2}{|1 - D(g_\theta(z))|^2}\right]$$

$$|D(x) - D^*(x)| \leq \epsilon \quad \forall x$$

$$\|\nabla_x (D(x) - D^*(x))\| \leq \epsilon \quad \forall x$$

$$\nabla_\theta \mathbb{E}_z [\log(1 - D(g_\theta(z))] \leq \frac{\epsilon^2}{(1-\epsilon)^2} \mathbb{E}_z [\|J_\theta g_\theta(z)\|^2]$$

$$\leq \frac{\epsilon^2}{(1-\epsilon)^2} M^2 \to 0 \quad \text{as } \epsilon \to 0$$

Thm: as $D \to D^*$, $\nabla_\theta L(D, g_\theta) \to 0$

Not covered: $\Delta\theta = \nabla_\theta \mathbb{E}_z [-\log D(g_\theta(z))]$

tldr; $\to$ cauchy dist. (under generous assumptions)

3. Random Noise

Replace p_g and p_r with $p_{g+\epsilon}$, $p_{r+\epsilon}$

$$p_{x+\epsilon}(x) = \mathbb{E}_{y \sim p_x} [p_\epsilon(x-y)] = \int p_\epsilon(x-y) \, d p_x(y)$$

Recall $D^*(x) = \frac{p_r(x)}{p_g(x) + p_r(x)} \longrightarrow D^*(x) = \frac{p_{r+\epsilon}(x)}{p_{g+\epsilon}(x) + p_{r+\epsilon}(x)}$

Thm: $\mathbb{E}_{z \sim p(z)} [\nabla_\theta \log(1 - D^*(g_\theta(z))]$ $\quad \epsilon, \epsilon' \sim N(0, \sigma^2 I)$

$$= \mathbb{E}_{z, \epsilon'} \Big[a(z) \int p_\epsilon(\tilde{g}_\theta(z) - y) \nabla_\theta \|\tilde{g}_\theta(z) - y\|^2 \, d p_r(y) - b(z) \int p_\epsilon(\tilde{g}_\theta(z) - y) \nabla_\theta \|\tilde{g}_\theta(z) - y\|^2 \, d p_g(y) \Big] = 2 \nabla_\theta JSD(p_r \| p_g)$$

$$a(z) = \frac{1}{2\sigma^2} \frac{1}{p_{r+\epsilon}(g_\theta(z)) + p_{g+\epsilon}(g_\theta(z))}, \quad b(z) = a(z) \frac{p_{r+\epsilon}(g_\theta(z))}{p_{g+\epsilon}(g_\theta(z))}$$