Understanding the Dynamics of GANs

Objective of GAN training : Given true distribution P, set of generators $\mathcal{G} = \{G_u ; u \in \mathcal{U}\}$ and set of discriminators $\mathcal{D} = \{D_v ; v \in \mathcal{V}\}$ for monotone $m : \mathbb{R} \to \mathbb{R}$ (log, identity map, ...)

$$\arg\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} \mathbb{E}_{x \sim P}[m(D_v(x))] + \mathbb{E}_{x \sim G_u}[m(1 - D_v(x))]$$

how it is trained :- Simultaneous (stochastic) gradient descent
- learn both generator + discriminator

issues - (a) Mode collapse :- Generator only learns subset of "variations / features" of true distribution

(b) Vanishing Gradients : Gradient updates for generators → 0

Open question : Can we understand convergence behaviour of GANs?

- Model (in this paper) for GAN dynamics
  - (refer paper for discussion on why Gaussian is useless)
  - true distribution, generators are bimodal gaussians with variance 1.

generator :

$$\mathcal{G} = \left\{ \tfrac{1}{2} N(\mu_1, 1) + \tfrac{1}{2} N(\mu_2, 1) ; \mu_1, \mu_2 \in \mathbb{R} \right\}$$

loss function:

$$d_{TV}(P,Q) = \frac{1}{2} \int_{\Omega} |P(x) - Q(x)|\, dx$$

≥

$$= \max_{A} P(A) - Q(A)$$ abs. val missing?

for GMM,

$$d_{TV}(G_{\mu_1}, G_{\mu_2}) = \max_{E = I_1 \cup I_2} G_{\mu_1}(E) - G_{\mu_2}(E), \quad (4)$$

where $I_1 \cap I_2 = \phi$, $I_1, I_2 \subseteq \mathbb{R}$

Proof of (4):

Thm A.1: let f be any analytic function with at most n-zeros. then $f \circ N(0,\sigma^2)$ has at most n-zeros.

Thm A.2: Any linear combination $F(x)$ of p.d.f of k Gaussians with same variance has at most $k-1$ zeros, provided at least two Gaussians have different means. In particular, for any $\mu \neq \nu$ $f(x) = D_\mu(x) - D_\nu(x)$ has at most three zeros.

Proof simple geometric argument for 2-gaussian, sufficient mean separation case. for close-by-gaussians, more involved.

-

$F(x) = \sum C_i f_i(x)$, $f_i(x) = N(\mu_i, 1)$

$|\mu_i - \mu_j|$ $i \neq j$ is small

this part not perfectly clear ↘

consider $g_i(x) = N(\mu_i, 1/D^2)$, $D \gg 1$

then apply geometric arg. to $G(x) = \sum C_i g_i(x)$

convolve with some fat Gaussian to get back F from G. then apply Thm A.1 to show. # zeroes is unchanged.

Discriminators:

$$\mathcal{D} = \{ \mathbb{1}_{[l_1,r_1]} + \mathbb{1}_{[l_2,r_2]} \mid l, r \in \mathbb{R}^2, \; l_1 \le r_1 \le l_2, r_2 \}$$

then, finding best fit in T.V to unknown G_{μ^*} is equivalent to finding

$$\hat{\mu} = \arg\min_{\mu} \max_{l,r} L(\mu, l, r), \text{ where}$$

$$L(\mu, l, r) = \mathbb{E}_{x \sim G_{\mu^*}}[D(x)] + \mathbb{E}_{x \sim G_{\mu}}[1 - D(x)]$$

A.2: By plugging in the definitions for the discriminators as above, using the model of two Gaussians, it can be shown that the function L is smooth in $l, r, \hat{\mu}$

Dynamics: A Optimal discriminator dynamics

$$l^{(t)}, r^{(t)} = \arg\max_{l,r} L(\hat{\mu}^{(t)}, l, r)$$

$$\hat{\mu}^{(t+1)} = \hat{\mu}^{(t)} - \eta_g \nabla_{\mu} L(\hat{\mu}^{(t)}, l^{(t)}, r^{(t)})$$

B First order dynamics

$$\hat{\mu}^{(t+1)} = \hat{\mu}^{(t)} - \eta_g \nabla_{\mu} L(\hat{\mu}^{(t)}, l^{(t)}, r^{(t)})$$

$$r^{(t+1)} = r^{(t)} + \eta_d \nabla_r L(\hat{\mu}^{(t)}, l^{(t)}, r^{(t)})$$

$$l^{(t+1)} = l^{(t)} + \eta_d \nabla_l L(\hat{\mu}^{(t)}, l^{(t)}, r^{(t)})$$

Main result

Thm 3.1 Fix $\delta > 0$ sufficiently small and $C > 0$. Let $\mu^* \in \mathbb{R}^2$ so that $|\mu_i^*| \leq C$ and $|\mu_1^* - \mu_2^*| \geq \delta$. Then for all initial points $\hat{\mu}^{(0)}$ s.t $|\hat{\mu}_i^{(0)}| \leq C$, $|\hat{\mu}_1^{(0)} - \hat{\mu}_2^{(0)}| \geq \delta$, if we let $\eta = \text{poly}(1/\delta, e^{-C^2})$ and $T = \text{poly}(1/\delta, e^{-C^2})$, then if $\hat{\mu}^{(T)}$ is specified by the optimal discriminator dynamics, $d_{TV}(G_{\mu^*}, G_{\hat{\mu}^{(T)}}) \leq$

Proof: convention: $\mu_1^* \leq \mu_2^*$, $\hat{\mu}_1 \leq \hat{\mu}_2$.

$$f(\hat{u}) \triangleq f_{u^*}(\hat{u}) = d_{TV}(G_{\hat{u}}, G_{\mu^*})$$

$$F(\hat{\mu}, x) = G_{\mu^*}(x) - G_{\hat{\mu}}(x)$$

for any $\delta > 0$

$\text{Rect}(\delta) = \text{Rect}(\mu^*, \delta) = \{\hat{u} \mid |\mu_i^* - \hat{\mu}_j| < \delta, \text{ some } i, j\}$

$\text{opt}(\delta) = \text{opt}(\mu^*, \delta) = \{\hat{u} : |\hat{\mu}_i - \mu_i^*| < \delta \ \forall i\}$

$B(C)$ - box with length of side C around origin.

$\text{sep}(\gamma) = \{v \in \mathbb{R}^2 : |v_1 - v_2| > \gamma\}$

$\delta' = D\delta$, $D \ll 1$

set

$\eta = \text{poly}(1/\delta, e^{-C^2})$ $\quad k_1 : (e^{-C^2})^{k_1} \ll \frac{1}{D}$

$= (e^{-C^2})^{k_1} \cdot (1/\delta)^{k_2}$

$\ll D \left(\frac{D}{\delta'}\right)^{k_2}$ $\quad k_2 > 1 \Rightarrow \ll \frac{D^2}{\delta'} = \frac{D^2}{D\delta}$

$\left(e^{-c^2}\right)^{k_1} \cdot \left(\frac{1}{\delta}\right)^{k_2}$ $\quad k_2 \geq 1, \quad k_1 : \left(e^{-c^2}\right)^{k_1} < \left(\frac{\delta'}{\delta}\right)$

$\left(\frac{\delta'}{\delta}\right)$.

$$n = \left(e^{-c^2}\right)^{k_1} \left(\frac{1}{\delta}\right)^{k_2} \qquad k \geq 1$$

$$= \left(e^{-c^2}\right)^{k_1} \left(\frac{D}{\delta'}\right)$$

set k_1 : $\left(e^{-c^2}\right)^{k_1} << \left(\frac{\delta'^2}{D}\right) = \left(\delta'\delta\right)$

$\Rightarrow n < \delta'$. Now, if $\hat{\mu} \in opt(\delta')$,

we have $\hat{\mu}^{(t+1)} \in opt(2\delta')$. for sufficiently small constant, by B.2, $d_{TV}(G_{\mu^*}(x), G_{\hat{\mu}^{(t+1)}}(x)) < \delta'$.

B.2: If two univariate Gaussians with unit variance have means within distance at most Δ, then their T.V distance is at most $O(1) \cdot \Delta$.

contra positively,

$d_{TV} > \delta \Rightarrow \hat{\mu} \notin opt(\delta')$

choose $n = \frac{1}{poly(c, \epsilon^{c^2}, 1/\delta)}$