Class 9

Shikhar Saxena

February 16, 2023

Contents

Recap

$$\begin{aligned} D_f(p\|q) &= \int q(x) \sup_t \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\} dx \\ &= \sup_t \int q(x) \left\{ T_1(x)\frac{p(x)}{q(x)} - f^*(T_1(x)) \right\} \\ &\geq \int q(x)T_1(x)\frac{p(x)}{q(x)} - f^*(T_1(x)) \end{aligned}$$

$$\therefore D_f(p\|q) = \sup_{T:X\to\mathbb{R}} \int_X (T(x)p(x) - f^*(T(x))q(x))\, dx$$

Using parameterized family of function T_ϕ to approximate the optimal function T, can minimize an f-divergence between p and p_θ (**GAN problem**)

$$\begin{aligned} \theta_s &= arg\ \min_\theta \sup_\phi \left[E_{x\sim p} T_\phi(x) - E_{x\sim p_\theta} f^*(T_\phi(x)) \right] \\ &= arg\ \min_\theta \sup_\phi \left[E_{x\sim p} T_\phi(x) - E f^*(T_\phi(g_\theta(z))) \right] \end{aligned}$$

where $z \to g_\theta(z) = x$

In 2014 paper, I. Goodfellow et al, Generative Adversarial Networks

Parameterizing: $T_\phi(x) = \log(d_\phi(x))$

$$\theta_f = arg\ \min_\theta \sup_\phi \left[E_{x\sim p} \log(d_\phi(x)) + E_{z\sim q} \log(1 - d_\phi(g_\theta(z))) \right]$$

Algorithm 1: GAN

for *training iterations* **do**
  for k *steps* **do**
    Sample minibatch of m noise sample $\{z^1, z^2, \ldots, z^m\}$ from priors $p_g(z)$
    Sample minibatch of m examples $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ from data generating distribution $p_{data}(x)$
    Update the discriminator by ascending its stochastic gradient

$$u^i \leftarrow \nabla_\phi \left(1/m \sum_{i=1}^{m} \log d_\phi(x^i) + \log(1 - d_\phi(g_\theta(z^i))) \right) \quad (\phi^{i+1} \leftarrow \phi^i + \eta u^i)$$

  end
  Sample minibatch of m noise samples $\{z^1, \ldots, z^m\}$
  Update the generator by descending

$$v_i = \nabla_\theta 1/m \log(1 - d_\phi(g_\theta(z^i))) \quad (\theta^{i+1} \leftarrow \theta^i - \eta v^i)$$

end

Algorithm for GAN

Proposition 1. *For G fixed, optimal discriminator D is*

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

Proof. Recall for Two-player game

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} \log D(x) + E_{z \sim p_Z(z)} [\log(1 - D(G(z)))]$$

Therefore, for a given G, discriminator D wants to maximize $V(G, D)$

$$V(G, D) = \int_x p_{data}(x) \log(D(x)) dx + p_g(x) \log(1 - D(x)) dx \tag{1}$$

For any $(a, b) \in \mathbb{R}^2$ the function $f(y) : y \to a \log y + b \log(1 - y)$ achieves max in $[0, 1]$ at $\frac{a}{a+b}$.

$f'(y) = \frac{a}{y} + \frac{b(-1)}{1-y} = 0 \implies y = \frac{a}{a+b}$

Therefore (1), proves our proposition.

□

We reformulate our loss function as $\min_G C(G)$ where $C(G) = \max_D V(G, D) = V(G, D_G^*)$.

Theorem 1. *The global minimum of the virtual training criterion $C(G)$ is achieved if and only if $p_g = p_{data}$. At that point, $C(G)$ achieves the value $-\log 4$.*

Proof. <u>if part</u>: If $p_g = p_{data}$, $D_G^* = 1/2$. Thus, $C(G) = -\log 4$.

<u>only if part</u>:

$$
\begin{aligned}
C(G) &= V(G, D_G^*) \\
&= E_{x \sim p_{data}}\left[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}\right] + E_{x \sim p_g}\left[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}\right] \\
&= -\log 4 + KL\left(p_{data} \middle\| \frac{p_{data} + p_g}{2}\right) + KL\left(p_g \middle\| \frac{p_{data} + p_g}{2}\right) \\
&= -\log 4 + 2\ JSD(p_{data} \| p_g)
\end{aligned}
$$

Minimum value of JSD is 0 which is achieved when $p_g = p_{data}$.

Thus, $-\log 4$ is the best value since, $JSD \geq 0$ with equality holding only when $p_g = p_{data}$. $\square$