Lý Thuyết GAN: Hàm Mục Tiêu và Jensen-Shannon Divergence

1 Hàm Mục Tiêu GAN và Mối Liên Hệ với Binary Cross-Entropy (BCE)

1.1 Hàm Mục Tiêu GAN

Hàm mục tiêu của GANs được định nghĩa như sau:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))]$$

Ý nghĩa:

- $x \sim p_{\text{data}}$: Dữ liệu thật từ phân phối p_{data}.
- $z \sim p(z)$: Nhiễu ngẫu nhiên (thường là Gaussian hoặc uniform).
- $G(z)$: Dữ liệu giả được tạo bởi Generator.
- $D(x)$: Xác suất mà Discriminator gán cho x là thật ($D(x) \approx 1$).
- $\mathbb{E}$: Kỳ vọng (trung bình trên tất cả mẫu).
- Discriminator (D) cố gắng tối đa hóa $V(G, D)$.
- Generator (G) cố gắng tối thiểu hóa $V(G, D)$.

1.2 Liên Hệ với Binary Cross-Entropy (BCE)

Binary Cross-Entropy (BCE) là hàm mất mát dùng để đo lường hiệu suất của một bộ phân loại nhị phân. Công thức BCE cho một mẫu dữ liệu là:

$$L_{\text{BCE}} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

- y: Nhãn thật (1 cho dữ liệu thật, 0 cho dữ liệu giả).
- $\hat{y}$: Xác suất dự đoán (tức là $D(x)$ hoặc $D(G(z))$).

Trong GANs, Discriminator được huấn luyện để phân biệt dữ liệu thật ($y = 1$) và dữ liệu giả ($y = 0$). Hàm mất mát của D có thể được viết lại như sau:
Đối với dữ liệu thật ($x \sim p_{\text{data}}, y = 1$):

$$L_{\text{BCE, real}} = -\log D(x)$$

Đối với dữ liệu giả $(G(z), y = 0)$:

$$L_{\text{BCE, fake}} = -\log(1 - D(G(z)))$$

Khi lấy kỳ vọng trên toàn bộ dữ liệu thật và giả, hàm mất mát tổng của D là:

$$\begin{aligned} L_D &= \mathbb{E}_{x\sim p_{\text{data}}}[-\log D(x)] + \mathbb{E}_{z\sim p(z)}[-\log(1 - D(G(z)))] \\ &= -\left(\mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]\right) \end{aligned}$$

Nhận thấy rằng hàm mục tiêu $V(G, D)$ của GANs chính là nghịch đảo của hàm mất mát BCE:

$$V(G, D) = -L_D$$

Kết luận: Tối đa hóa $V(G, D)$ tương đương với tối thiểu hóa hàm mất mát BCE của Discriminator. Điều này cho thấy GANs sử dụng BCE để huấn luyện D, với mục tiêu làm D phân biệt tốt nhất giữa dữ liệu thật và giả.

2 Suy Ra Jensen-Shannon (JS) Divergence

2.1 Discriminator Tối Ưu

Để suy ra JS Divergence, chúng ta cần tìm Discriminator tối ưu D^* khi Generator G cố định. Hàm mục tiêu được viết lại dưới dạng tích phân (thay vì kỳ vọng trên mẫu):

$$V(G, D) = \int_x p_{\text{data}}(x) \log D(x)\, dx + \int_z p(z) \log(1 - D(G(z)))\, dz$$

Vì $z \sim p(z)$ và $G(z) \sim p_g$, phần thứ hai có thể được viết lại theo $x \sim p_g(x)$:

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

Để tối đa hóa $V(G, D)$ theo D, ta lấy đạo hàm của biểu thức bên trong tích phân theo $D(x)$:

$$\frac{\partial}{\partial D(x)} [p_{\text{data}}(x) \log D(x) + p_g(x) \log(1 - D(x))] = \frac{p_{\text{data}}(x)}{D(x)} - \frac{p_g(x)}{1 - D(x)}$$

Đặt đạo hàm bằng 0 để tìm cực đại:

$$\frac{p_{\text{data}}(x)}{D(x)} = \frac{p_g(x)}{1 - D(x)}$$

$$p_{\text{data}}(x)(1 - D(x)) = p_g(x)D(x)$$

$$p_{\text{data}}(x) - p_{\text{data}}(x)D(x) = p_g(x)D(x)$$

$$p_{\text{data}}(x) = D(x)(p_{\text{data}}(x) + p_g(x))$$

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

Ý nghĩa: Discriminator tối ưu $D^*(x)$ phân biệt thật và giả dựa trên tỷ lệ giữa phân phối thật (p_{data}) và phân phối giả (p_g).

2.2 Thay D^* vào Hàm Mục Tiêu

Khi $D = D^*$, ta thay $D^*(x)$ vào $V(G, D)$:

$$V(G, D^*) = \int_x \left[p_{\text{data}}(x) \log \left(\frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} \right) + p_g(x) \log \left(1 - \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} \right) \right] dx$$
$$= \int_x \left[p_{\text{data}}(x) \log \left(\frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} \right) + p_g(x) \log \left(\frac{p_g(x)}{p_{\text{data}}(x) + p_g(x)} \right) \right] dx$$

Định nghĩa $m(x) = \frac{p_{\text{data}}(x) + p_g(x)}{2}$. Ta có thể viết lại:

$$V(G, D^*) = \int_x \left[p_{\text{data}}(x) \log \left(\frac{p_{\text{data}}(x)/m(x)}{(p_{\text{data}}(x) + p_g(x))/(2m(x))} \right) + p_g(x) \log \left(\frac{p_g(x)/m(x)}{(p_{\text{data}}(x) + p_g(x))/(2m(x))} \right) \right] dx$$
$$= \int_x \left[p_{\text{data}}(x) \log \left(\frac{p_{\text{data}}(x)}{m(x)} \right) + p_g(x) \log \left(\frac{p_g(x)}{m(x)} \right) \right] dx - \int_x [p_{\text{data}}(x) + p_g(x)] \log 2 \, dx$$

Vì $\int_x p_{\text{data}}(x) \, dx = 1$ và $\int_x p_g(x) \, dx = 1$, ta có:

$$\int_x [p_{\text{data}}(x) + p_g(x)] \log 2 \, dx = 2 \log 2$$

Phần còn lại chính là KL divergence:

$$V(G, D^*) = KL(p_{\text{data}}||m) + KL(p_g||m) - 2 \log 2$$
$$= 2 \cdot \frac{1}{2} \left(KL(p_{\text{data}}||m) + KL(p_g||m) \right) - 2 \log 2$$
$$= 2JS(p_{\text{data}}||p_g) - 2 \log 2$$

Jensen-Shannon Divergence được định nghĩa là:

$$JS(p_{\text{data}}||p_g) = \frac{1}{2} KL(p_{\text{data}}||m) + \frac{1}{2} KL(p_g||m), \quad m = \frac{p_{\text{data}} + p_g}{2}$$

Vậy:

$$V(G, D^*) = 2JS(p_{\text{data}}||p_g) - 2 \log 2$$

2.3 Ý Nghĩa

- Khi D tối ưu, hàm mục tiêu của GANs tương đương với việc tối thiểu hóa JS Divergence giữa p_{data} và p_g.
- JS Divergence đo lường sự khác biệt giữa hai phân phối. Khi G cải thiện, p_g tiến gần p_{data}, làm giảm JS Divergence.
- Hằng số $-2 \log 2$ không ảnh hưởng đến tối ưu hóa, vì nó không phụ thuộc vào G.

3 Chứng Minh Bài Toán Cực Tiểu–Cực Đại Đạt Giá Trị Cực Đại Khi $p_g = p_{\text{data}}$

3.1 Mục Tiêu

Chúng ta cần chứng minh rằng điểm cân bằng của bài toán $\min_G \max_D V(G, D)$ đạt được khi $p_g = p_{\text{data}}$, và tại đó $V(G, D) = -2 \log 2$.

3.2 Khi $p_g = p_{\textbf{data}}$

Thay $p_g = p_{\text{data}}$ vào $D^*(x)$:

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{\text{data}}(x)} = \frac{1}{2}$$

Điều này có nghĩa Discriminator không thể phân biệt thật và giả, vì chúng có cùng phân phối. Thay $D^*(x) = \frac{1}{2}$ vào $V(G, D)$:

$$\begin{aligned} V(G, D^*) &= \mathbb{E}_{x\sim p_{\text{data}}}\left[\log \frac{1}{2}\right] + \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - \frac{1}{2}\right)\right] \\ &= \int_x p_{\text{data}}(x) \log \frac{1}{2}\, dx + \int_z p(z) \log \frac{1}{2}\, dz \\ &= \log \frac{1}{2} \int_x p_{\text{data}}(x)\, dx + \log \frac{1}{2} \int_z p(z)\, dz \\ &= \log \frac{1}{2} \cdot 1 + \log \frac{1}{2} \cdot 1 = -\log 2 - \log 2 = -2 \log 2 \end{aligned}$$

3.3 Kiểm Tra Giá Trị Cực Đại

Khi $p_g = p_{\text{data}}$, JS Divergence đạt giá trị nhỏ nhất:

$$JS(p_{\text{data}}||p_{\text{data}}) = 0$$

Thay vào:

$$V(G, D^*) = 2 \cdot 0 - 2 \log 2 = -2 \log 2$$

- Đây là giá trị nhỏ nhất mà G có thể đạt được, vì JS Divergence không âm ($JS \geq 0$).
- Đối với D, $D^*(x) = \frac{1}{2}$ là điểm tối ưu, vì bất kỳ $D(x) \neq \frac{1}{2}$ sẽ làm giảm $V(G, D)$ (do đạo hàm đã được kiểm tra ở bước 2.1).

3.4 Ý Nghĩa

- Khi $p_g = p_{\text{data}}$, Generator tạo ra dữ liệu giống hệt dữ liệu thật, và Discriminator không thể phân biệt (gán xác suất $\frac{1}{2}$ cho mọi mẫu).
- Đây là điểm cân bằng Nash của trò chơi, nơi không bên nào có thể cải thiện thêm mà không làm đối thủ tốt hơn.