Generative Adversarial Networks (GANs) are a class of deep learning models designed to generate new data samples that resemble a given dataset. They consist of two main components, a generator and a discriminator, that are trained simultaneously in a game-theoretic setup. Let's dive into the mathematics of GANs, their training process, and optimization details.

1 GANs Setup

GANs are defined by two neural networks:

- **Generator** G: This network takes random noise z as input and generates a data sample $G(z)$ that should resemble the data in the original dataset.
- **Discriminator** D: This network takes a data sample (either real or generated) as input and outputs the probability that it is real (i.e., from the actual dataset) rather than generated.

Let $x \sim p_{data}(x)$ be real data sampled from the data distribution, and $z \sim p_z(z)$ be random noise sampled from a simple distribution (e.g., Gaussian).

2 The Objective Function (Minimax Game)

The *GAN* framework sets up a $mini - max$ game between G and D. The discriminator's job is to correctly classify real vs. generated data, while the generator's job is to "fool" the discriminator into classifying generated data as real.

The optimization problem for *GANs* can be formulated as:

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

- **Discriminator's Objective**: Maximize the probability of correctly classifying real data ($D(x)$) as real (outputting 1 for real samples) and generated data ($D(G(z))$) as fake (outputting 0 for generated samples).
- **Generator's Objective**: Minimize the probability that the discriminator correctly classifies generated data as fake, i.e., minimizing $\mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$, effectively maximizing $\log D(G(z))$.

In each training step, D tries to maximize $V(D,G)$ while G tries to minimize it.

3 Training GANs (Alternating Optimization)

GANs are trained through an alternating optimization process, where we update D and G iteratively.

1. Update the Discriminator. To optimize the discriminator, we maximize $V(D, G)$ with respect to D while keeping G fixed. The discriminator's loss function is:

$$L_D = -\left(\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]\right) \tag{2}$$

   This updates D to better classify real samples as real and fake samples as fake.

2. Update the Generator. To optimize the generator, we minimize $V(D, G)$ with respect to G while keeping D fixed. The generator's loss function is:

$$L_G = -\mathbb{E}_{z\sim p_z(z)}[\log D(G(z))] \tag{3}$$

   This formulation of the generator's loss encourages it to generate samples that maximize the discriminator's output, effectively "fooling" it into classifying fake samples as real.

4 Gradient Descent Update

Using gradient descent, the parameters of D and G are updated iteratively.

For the discriminator:

$$\theta_D \leftarrow \theta_D + \eta\nabla_{\theta_D} L_D \tag{4}$$

For the generator:

$$\theta_G \leftarrow \theta_G - \eta\nabla_{\theta_G} L_G \tag{5}$$

5 Optimal Discriminator Derivation

The objective function of $GANs$ described in eq.(1) can be written as:

$$V(D, G) = \int_x p_{\text{data}}(x)\log D(x)\,dx + \int_x p_g(x)\log(1 - D(x))\,dx \tag{6}$$

where $p_{data}(x)$ is the true data distribution, and $p_g(x)$ is the distribution of generated data when $z \sim p_z(z)$ is passed through G, i.e., $p_g(x) = G(z)$.

To derive the optimal discriminator, we assume G is fixed and maximize $V(D, G)$ with respect to D. The goal is to solve:

$$\max_D \int_x p_{\text{data}}(x)\log D(x)\,dx + \int_x p_g(x)\log(1 - D(x))\,dx \tag{7}$$

Step-by-Step Maximization

Let $f(D) = p_{\text{data}}(x) \log D(x) + p_g(x) \log(1 - D(x))$. To find the optimal D that maximizes f, we take the derivative with respect to $D(x)$ and set it to zero.

$$\frac{\partial f}{\partial D(x)} = \frac{p_{\text{data}}(x)}{D(x)} - \frac{p_g(x)}{1 - D(x)} = 0 \tag{8}$$

Rearranging terms,

$$\frac{p_{\text{data}}(x)}{D(x)} = \frac{p_g(x)}{1 - D(x)} \tag{9}$$

Cross-multiplying gives:

$$p_{\text{data}}(x)(1 - D(x)) = p_g(x)D(x) \tag{10}$$

Expanding and solving for $D(x)$:

$$p_{\text{data}}(x) - p_{\text{data}}(x)D(x) = p_g(x)D(x) \tag{11}$$

$$p_{\text{data}}(x) = D(x)(p_{\text{data}}(x) + p_g(x)) \tag{12}$$

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} \tag{13}$$

Thus, the optimal discriminator $D^*(x)$ for a fixed generator G is:

$$D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} \tag{14}$$

6 Jensen-Shannon Divergence Derivation

Now that we have D^*, let's substitute it back into the objective function to see how it relates to the *Jensen-Shannon Divergence.*

The *GAN* objective function at the optimal discriminator D^* is:

$$V(D^*, G) = \mathbb{E}_{x\sim p_{\text{data}}(x)} [\log D^*(x)] + \mathbb{E}_{x\sim p_g(x)} [\log D^*(x)] \tag{15}$$

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

$$V(D^*, G) = \mathbb{E}_{x\sim p_{\text{data}}(x)} \left[\log \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}\right] + \mathbb{E}_{x\sim p_g(x)} \left[\log \frac{p_g(x)}{p_{\text{data}}(x) + p_g(x)}\right] \tag{16}$$

The two terms in this expression represent the $\log -likelihoods$ of $D^*(x)$ predicting real data for real samples, and predicting generated (fake) data for generated samples, respectively.

To get to the *JSD*, we rewrite each term separately. Notice that the *JSD* between two distributions p and q is defined as:

$$\mathrm{JSD}(p||q) = \frac{1}{2}\mathbb{E}_{x\sim p}\left[\log\frac{p(x)}{\frac{p(x)+q(x)}{2}}\right] + \frac{1}{2}\mathbb{E}_{x\sim q}\left[\log\frac{q(x)}{\frac{p(x)+q(x)}{2}}\right] \tag{17}$$

This expression shows that the *JSD* is the average of the *Kullback-Leibler (KL)* divergence from each distribution to the "*mixture distribution*" $m(x) = \frac{p(x)+q(x)}{2}$

Let's work toward matching this form by rewriting $V(D^*, G)$ in terms of expectations of ratios involving the mixture distribution.

The first term in $V(D^*, G)$ is:

$$\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}\right] = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\frac{p_{\text{data}}(x)}{\frac{p_{\text{data}}(x)+p_g(x)}{2}}\right] - \log 2 \tag{18}$$

Similarly, the second term can be written as:

$$\mathbb{E}_{x\sim p_g(x)}\left[\log\frac{p_g(x)}{p_{\text{data}}(x)+p_g(x)}\right] = \mathbb{E}_{x\sim p_g(x)}\left[\log\frac{p_g(x)}{\frac{p_{\text{data}}(x)+p_g(x)}{2}}\right] - \log 2 \tag{19}$$

Now, we substitute these rewritten terms back into the expression for $V(D^*, G)$:

$$V(D^*, G) = \left(\mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\frac{p_{\text{data}}(x)}{\frac{p_{\text{data}}(x)+p_g(x)}{2}}\right] + \mathbb{E}_{x\sim p_g(x)}\left[\log\frac{p_g(x)}{\frac{p_{\text{data}}(x)+p_g(x)}{2}}\right]\right) - 2\log 2 \tag{20}$$

This form is now exactly twice the *Jensen-Shannon Divergence* between p_{data} and p_g, minus a constant:

$$V(D^*, G) = 2 \cdot \mathrm{JSD}(p_{\text{data}}||p_g) - \log 4 \tag{21}$$

In other words, this result in eq.(21) shows that maximizing $V(D, G)$ with respect to D corresponds to maximizing the *JSD* between the real and generated data distributions. In other words, the discriminator's role is to maximize the separation between p_{data} and p_g by making the *JSD* as large as possible.