WGAN is a type of Generative Adversarial Network (GAN) that improves training stability by using the *Wasserstein-1 distance* (Earth Mover's distance) instead of the Jensen-Shannon divergence. In the following, a detailed mathematical derivation will be done.

1 GAN Objective

The original *GAN* objective is:

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

where,

- D: Discriminator (critic in $WGAN$).
- G: Generator.
- p_{data}: Data distribution.
- p_z: Latent space distribution (e.g., Gaussian or uniform).

2 Problems with Original GAN

The *Jensen-Shannon divergence* used in the original GAN can cause vanishing gradients when p_G and p_{data} have disjoint supports. $WassersteinGAN - WGAN$ proposes using the $EarthMover's$ distance to mitigate this.

3 Earth Mover's Distance (Wasserstein-1 Distance)

The *Earth Mover's* distance between p_r (real data distribution) and p_g (generated data distribution) is defined as:

$$W(p_r, p_g) = \inf_{\gamma\in\Pi(p_r,p_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] \tag{2}$$

where,

- $\Pi(p_r, p_g)$ is the set of all joint distributions with marginals p_r and p_g.

4 Kantorovich-Rubinstein Duality

The *Wasserstein-1 distance* can be reformulated using *Kantorovich-Rubinstein duality*:

$$W(p_r, p_g) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim p_r}[f(x)] - \mathbb{E}_{x \sim p_g}[f(x)] \tag{3}$$

where,

- f: A *Lipschitz-continuous* function with a *Lipschitz* constant of at most $1\|f\|_L \leq 1$

In *WGAN*, f corresponds to the discriminator D, which is now called the *critic*.

5 WGAN Objective

The generator G and critic D are optimized using the following loss functions:

- Critic's loss:
$$L_D = -\mathbb{E}_{x \sim p_r}[D(x)] + \mathbb{E}_{z \sim p_z}[D(G(z))] \tag{4}$$
- Generator's loss:
$$L_G = -\mathbb{E}_{z \sim p_z}[D(G(z))] \tag{5}$$

The optimization is carried out iteratively:

1 Optimize D to **approximate** the *Wasserstein distance*.

2 Optimize G to **minimize** the *Wasserstein distance*.

6 Enforcing the Lipschitz Constraint

To ensure D is *Lipschitz-continuous*, the original *WGAN* clipped $D's$ weights to a small range $(-c, c)$. However, weight clipping can cause capacity issues. *WGAN-GP* introduces a gradient penalty:

- Gradient penalty term:
$$\lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}} \left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \right] \tag{6}$$
- Critic loss with gradient penalty:
$$L_D = -\mathbb{E}_{x \sim p_r}[D(x)] + \mathbb{E}_{z \sim p_z}[D(G(z))] + \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}} \left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \right] \tag{7}$$

where,

- $p_{\hat{x}}$: Uniform sampling along straight lines between real and generated data points.