

# Study Material - Youtube

## Document Information

- **Generated:** 2025-08-02 00:43:32
- **Source:** [https://youtu.be/\\_IBfVkrvqAI](https://youtu.be/_IBfVkrvqAI)
- **Platform:** Youtube
- **Word Count:** 2,199 words
- **Estimated Reading Time:** ~10 minutes
- **Number of Chapters:** 3
- **Transcript Available:** Yes (analyzed from video content)

## Table of Contents

1. Wasserstein GANs (WGANs): A Deep Dive
  2. Self-Assessment for This Video
  3. Key Takeaways from This Video
- 

## Video Overview

This lecture provides a detailed mathematical foundation for Wasserstein Generative Adversarial Networks (WGANs). It begins by identifying a critical flaw in the training of standard GANs: the saturation of divergence metrics like Jensen-Shannon (JS) divergence, which leads to vanishing gradients and training instability.

The core of the lecture introduces the **Wasserstein distance**, also known as the Earth Mover's Distance, as a "softer" and more robust alternative. This metric is derived from the theory of **Optimal Transport**, which conceptualizes the distance between two distributions as the minimum "work" required to transform one into the other.

A significant portion of the lecture is dedicated to explaining the mathematical underpinnings of this concept, including the definition of a **transport plan** and the challenge of its direct computation. The instructor then presents the **Kantorovich-Rubinstein duality**, a pivotal theorem that reframes the intractable Wasserstein distance calculation into a more manageable optimization problem. This dual formulation is the key to creating a practical WGAN architecture.

Finally, the lecture details how to construct a WGAN by replacing the standard GAN's discriminator with a "critic" that is constrained to be **1-Lipschitz**. It explains the new objective function and the practical methods for enforcing this constraint, leading to a more stable and reliable training process for generative models.

## Learning Objectives

Upon completing this lecture, students will be able to: - **Understand the limitations of standard GANs**, particularly the issue of vanishing gradients due to divergence metric saturation. - **Define the Wasserstein distance** and explain its intuitive meaning through the "Earth Mover's Distance" analogy. - **Grasp the concept of Optimal Transport** and how a "transport plan" represents the transformation between two distributions. - **Explain the Kantorovich-Rubinstein duality** and its role in making the Wasserstein distance computationally tractable. - **Describe the architecture of a WGAN**, including the function of the generator and the "critic". - **Understand the 1-Lipschitz constraint** and why it is essential for the WGAN critic. - **Formulate the WGAN objective function** and outline the training algorithm.

## Prerequisites

To fully understand the content of this lecture, students should have a foundational knowledge of: - **Generative Adversarial Networks (GANs):** Basic architecture and training principles. - **Probability and Statistics:** Concepts of probability distributions ( $P_x, P_\theta$ ), density functions, and expected value ( $\mathbb{E}$ ). - **Calculus:** Derivatives, gradients, and the concepts of minimization and maximization. - **Linear Algebra:** Understanding of norms (e.g., L2 norm). - **Neural Networks:** Familiarity with network parameters ( $\theta, \omega$ ) and gradient-based optimization.

## Key Concepts Covered

- Divergence Metric Saturation
  - Wasserstein Distance (Earth Mover's Distance)
  - Optimal Transport
  - Transport Plan ( $\gamma$ )
  - Kantorovich-Rubinstein Duality
  - 1-Lipschitz Functions
  - WGAN Architecture (Generator and Critic)
  - WGAN Objective Function
- 

## Wasserstein GANs (WGANs): A Deep Dive

This lecture explores Wasserstein GANs (WGANs), an advancement in generative modeling that addresses the training instability often encountered in standard GANs. We will begin by understanding the problem with traditional GANs and then delve into the mathematical solution provided by the Wasserstein distance.

### The Problem with Standard GANs: Divergence Saturation

The training of standard Generative Adversarial Networks (GANs) often suffers from instability. A primary reason for this is the choice of the divergence metric used to measure the “distance” between the real data distribution ( $P_x$ ) and the generated data distribution ( $P_\theta$ ).

At (00:11), the instructor highlights the core problem: > **Solution:** Use a ‘softer’ divergence metric, that does not saturate when the manifolds of the supports of  $P_x$  &  $P_\theta$  do not align.

In standard GANs, the objective function implicitly minimizes the **Jensen-Shannon (JS) divergence**. The JS divergence, along with similar metrics like the Kullback-Leibler (KL) divergence, has a critical flaw in the context of GANs. When the supports of the two distributions,  $P_x$  and  $P_\theta$ , are disjoint (meaning they do not overlap), the JS divergence saturates to a constant value ( $\log 2$ ).

**Why is this a problem?** When the divergence is constant, its gradient with respect to the generator's parameters ( $\theta$ ) becomes zero. This is known as the **vanishing gradient problem**. If the gradient is zero, the generator receives no information on how to adjust its parameters to produce better samples, and learning stalls. In high-dimensional spaces, it is highly probable that the manifolds on which the real and generated data lie are disjoint, making this a frequent and severe issue.

### The Wasserstein Metric: A “Softer” Alternative

To overcome the saturation problem, we need a metric that provides a meaningful, non-zero gradient even when the distribution supports do not overlap. The **Wasserstein metric**, introduced at (00:15), provides such a solution. It is derived from the concept of **Optimal Transport** (00:31).

## Intuitive Foundation: Optimal Transport and Earth Mover's Distance

The Wasserstein-1 distance is also known as the **Earth Mover's Distance (EMD)**. This name provides a powerful intuition.

Imagine you have a pile of dirt representing one probability distribution ( $P_x$ ) and a hole of the same volume representing another distribution ( $P_\theta$ ). The EMD is the minimum “cost” or “work” required to move the dirt to fill the hole perfectly.

- **Work/Cost:** The work is calculated as the **amount of dirt moved** multiplied by the **distance it is moved**.
- **Optimal Plan:** We want to find the most efficient way to move the dirt—the “transport plan” that minimizes the total work.

This minimum work gives us a true distance metric between the two distributions. Unlike JS divergence, even if the pile of dirt and the hole are far apart (disjoint supports), there is still a well-defined cost to move the dirt, and this cost provides a useful gradient for optimization.

## Mathematical Analysis of Wasserstein Distance

The formal definition of the Wasserstein-1 distance captures the Earth Mover's intuition.

At (02:08), the instructor introduces the formula:

$$W(P_x, P_{\hat{x}}) = \inf_{\gamma \in \Pi(P_x, P_{\hat{x}})} \mathbb{E}_{(x, \hat{x}) \sim \gamma} [\|x - \hat{x}\|_p]$$

Let's break this down: -  $W(P_x, P_{\hat{x}})$  is the Wasserstein distance between the real distribution  $P_x$  and the generated distribution  $P_{\hat{x}}$ . The subscript  $p$  on the norm (e.g.,  $p = 2$  for Euclidean distance) defines the specific Wasserstein- $p$  distance. -  $\Pi(P_x, P_{\hat{x}})$  is the set of all possible **joint distributions**  $\gamma(x, \hat{x})$  whose marginal distributions are  $P_x$  and  $P_{\hat{x}}$ . -  $\gamma(x, \hat{x})$  is a **transport plan**. It specifies how much probability mass should be moved from a point  $x$  to a point  $\hat{x}$  to transform  $P_x$  into  $P_{\hat{x}}$ . -  $\inf_{\gamma \in \Pi}$  signifies that we are searching for the infimum (the minimum cost) over all possible transport plans. -  $\mathbb{E}_{(x, \hat{x}) \sim \gamma} [\|x - \hat{x}\|_p]$  is the expected cost for a given transport plan  $\gamma$ . It's the average distance the mass travels.

The transport plan  $\gamma$  must satisfy the marginal constraints (03:45): 1.  $\int \gamma(x, \hat{x}) d\hat{x} = P_x(x)$  2.  $\int \gamma(x, \hat{x}) dx = P_{\hat{x}}(\hat{x})$

These constraints ensure that the total mass moved from  $P_x$  and the total mass arriving at  $P_{\hat{x}}$  are conserved.

**Problem:** The optimization over the space of all possible transport plans,  $\Pi(P_x, P_{\hat{x}})$ , is computationally intractable.

## Kantorovich-Rubinstein Duality: Making WGANs Practical

The key to making the Wasserstein distance usable in practice is the **Kantorovich-Rubinstein duality** (39:08). This powerful theorem provides an alternative, dual formulation for the Wasserstein-1 distance that is much easier to work with.

The dual formulation is (39:23):

$$W(P_x, P_\theta) = \sup_{\|f\|_L \leq 1} (\mathbb{E}_{x \sim P_x} [f(x)] - \mathbb{E}_{\hat{x} \sim P_\theta} [f(\hat{x})])$$

- **Supremum (sup):** We are now maximizing instead of minimizing.
- **1-Lipschitz Constraint:** The maximization is performed over all **1-Lipschitz functions**, denoted by  $\|f\|_L \leq 1$ .

A function  $f$  is 1-Lipschitz if its gradient has a norm of at most 1 everywhere. This means the function's output cannot change more rapidly than its input.

$$\|\nabla_x f(x)\| \leq 1$$

Or, more generally:

$$\frac{\|f(x_1) - f(x_2)\|}{\|x_1 - x_2\|} \leq 1 \quad \text{for all } x_1 \neq x_2$$

This constraint is crucial. It prevents the function  $f$  from becoming too steep, ensuring that the distance measure remains meaningful.

## **The WGAN Architecture and Training Algorithm**

The Kantorovich-Rubinstein dual form fits perfectly into the adversarial training framework of GANs.

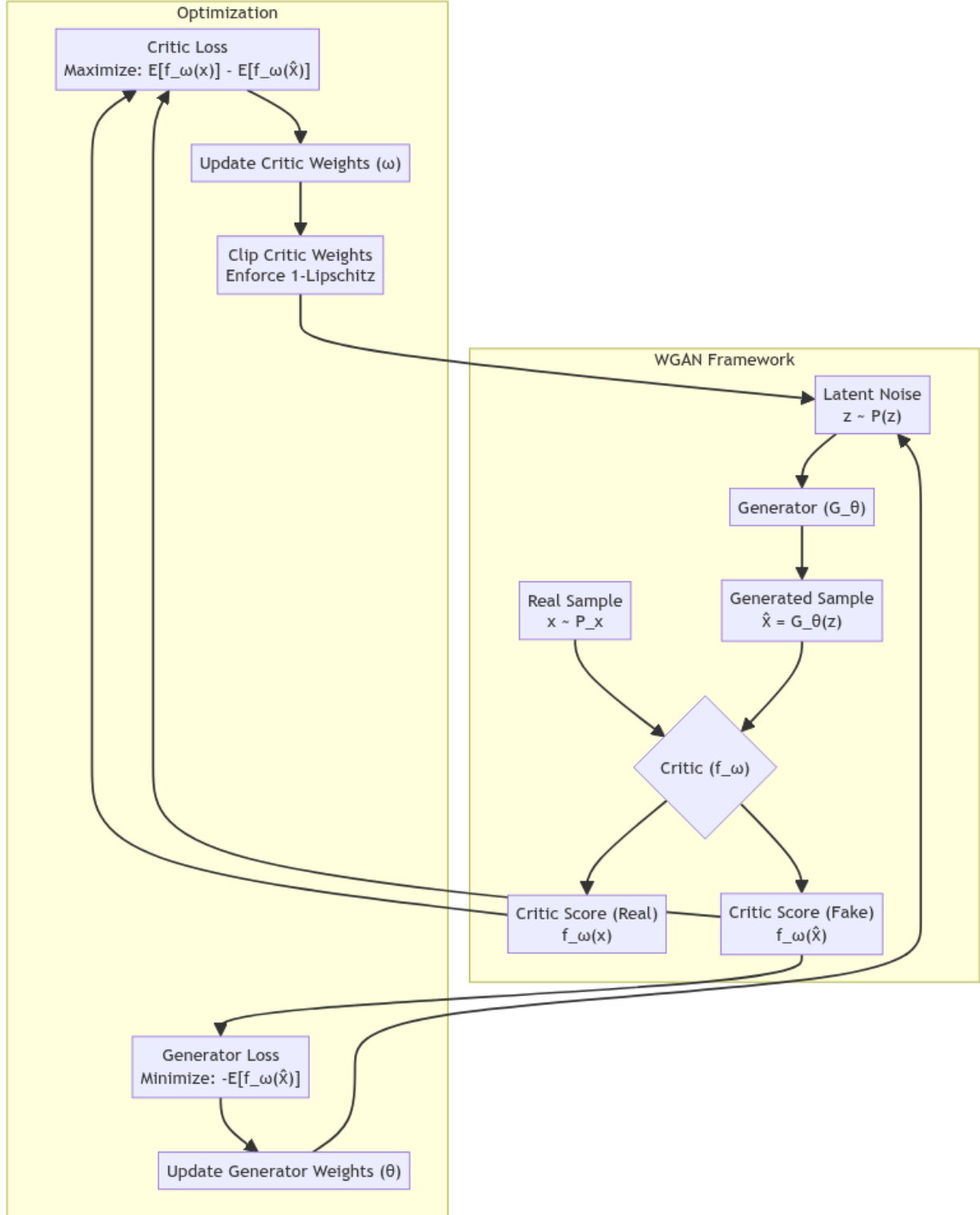


Figure 1: Flowchart illustrating the WGAN training process. The generator tries to produce samples that get high scores from the critic, while the critic tries to maximize the score difference between real and fake samples.

## WGAN Objective Function

The training process becomes a min-max game based on the dual formulation:

$$\min_{\theta} \max_{\omega: \|f_{\omega}\|_L \leq 1} \left( \mathbb{E}_{x \sim P_x} [f_{\omega}(x)] - \mathbb{E}_{z \sim P_z} [f_{\omega}(G_{\theta}(z))] \right)$$

- **The Critic ( $f_{\omega}$ ):** This network (formerly the discriminator) is trained to maximize the objective, effectively estimating the Wasserstein distance. It outputs a scalar score, not a probability.
- **The Generator ( $G_{\theta}$ ):** This network is trained to minimize the objective, which means it tries to generate samples that the critic scores similarly to real samples.

## Enforcing the 1-Lipschitz Constraint

The most critical part of implementing a WGAN is enforcing the 1-Lipschitz constraint on the critic  $f_{\omega}$ . The original paper proposed a simple technique:

- **Weight Clipping (45:31):** After each gradient update for the critic, its weights ( $\omega$ ) are clipped to a small, fixed range, such as  $[-0.01, 0.01]$ . This forces the weights to stay small, which in turn helps to bound the gradients and approximate the Lipschitz constraint.

While simple, weight clipping can lead to issues like capacity underuse or exploding gradients if the clipping value is not chosen carefully. More advanced techniques like Gradient Penalty (WGAN-GP) have since been developed, but weight clipping is the foundational method.

## Conclusion: Why WGANs are More Stable

The conclusion of the lecture, summarized at (49:24), is that training a WGAN is more stable than training a naive GAN.

- **Key Advantage:** The Wasserstein metric does not saturate when the real and generated distributions have disjoint supports. This ensures that the generator always receives a meaningful gradient, preventing the training process from stalling.
- **Practical Implementation:** The Kantorovich-Rubinstein duality allows us to approximate the Wasserstein distance with a critic network.
- **The Critic's Role:** The critic is not a classifier but a regressor that learns a 1-Lipschitz function to measure the distance between distributions.

This shift from a classification-based divergence (JS) to a true distance metric (Wasserstein) is the fundamental reason for the improved stability and performance of WGANs.

---

## Self-Assessment for This Video

### 1. Conceptual Questions:

- What is the primary reason for training instability in standard GANs? Explain the concept of “divergence saturation.”
- In your own words, describe the “Earth Mover’s Distance” analogy for the Wasserstein distance. Why is it considered a “softer” metric than JS divergence?
- What is a “transport plan” in the context of Optimal Transport? What do the marginal constraints on a transport plan ensure?
- Why is the Kantorovich-Rubinstein duality essential for building a practical WGAN?
- What is a 1-Lipschitz function? Why must the critic in a WGAN be 1-Lipschitz?

### 2. Mathematical Questions:

- Write down the primal and dual formulations of the Wasserstein-1 distance. Explain every term in both equations.

- Given two 1D discrete distributions,  $P_x$  with mass at points  $\{x_1, \dots, x_k\}$  and  $P_{\hat{x}}$  with mass at points  $\{\hat{x}_1, \dots, \hat{x}_L\}$ , how would you represent a transport plan between them?
- Explain the WGAN min-max objective function. How does the generator's loss differ from the critic's loss?

### 3. Application and Implementation Questions:

- What is the main architectural difference between a standard GAN's discriminator and a WGAN's critic?
- How did the original WGAN paper propose to enforce the 1-Lipschitz constraint? What are the potential drawbacks of this method?
- Outline the alternating training steps for the generator and the critic in a WGAN.

## Key Takeaways from This Video

- **Problem:** Standard GANs fail when the real and generated data manifolds do not overlap because the JS divergence saturates, providing no gradient for the generator.
- **Solution:** Use the **Wasserstein distance**, a metric from Optimal Transport theory that does not saturate and provides meaningful gradients even for disjoint distributions.
- **Optimal Transport:** The Wasserstein distance is the minimum “work” (mass  $\times$  distance) needed to transform one distribution into another. This is defined by finding the optimal “transport plan.”
- **Kantorovich-Rubinstein Duality:** This theorem converts the intractable primal problem of finding the best transport plan into a tractable dual problem of finding the best 1-Lipschitz function that maximizes the difference in expectations.
- **WGAN Implementation:**
  - The discriminator is replaced by a **critic** that outputs a real-valued score.
  - The objective function is changed to the dual form of the Wasserstein distance.
  - The critic must be constrained to be **1-Lipschitz**, which is practically achieved through methods like weight clipping.
- **Result:** WGANs offer significantly more stable training and a loss that better correlates with the quality of generated samples.

## Visual References

A graph illustrating the core problem with standard GANs: the saturation of the Jensen-Shannon (JS) divergence. It visually shows how the gradient vanishes when the real and generated distributions have little overlap, explaining the need for a new distance metric. (at 04:20):


Wasserstein's Metric (Optimal Transport)

Given two distributions  $p_x$  &  $p_{\tilde{x}}$ ,


$$W(p_x || p_{\tilde{x}}) = \min_{\pi \in \Pi(p_x, p_{\tilde{x}})} \mathbb{E}_{x, \tilde{x} \sim \pi} \|x - \tilde{x}\|_2$$

$\pi$ : a joint distribution between  $p_x$  &  $p_{\tilde{x}}$

$\Pi(p_x, p_{\tilde{x}})$  : All

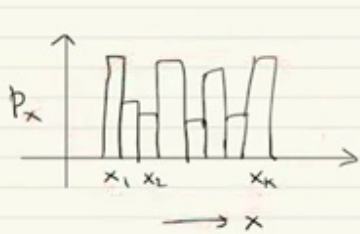


IIT Madras  
B.S. Degree



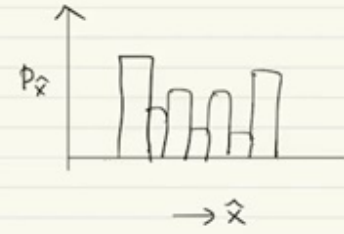
A key diagram explaining the 'Earth Mover's Distance' intuition for the Wasserstein distance. It visually represents two probability distributions as piles of earth and the 'transport plan' as the work required to move one pile to match the other. (at 09:15):

Suppose  $p_x$  &  $p_{\tilde{x}}$  are two 1-d discrete pmf.




$p_x$

$x$




$p_{\tilde{x}}$

$\tilde{x}$



IIT Madras  
B.S. Degree



A slide presenting the Kantorovich-Rubinstein duality theorem. It shows the intractable primal form of the Wasserstein distance (involving transport plans) and its equivalent, computationally tractable dual form (involving 1-Lipschitz functions), which is the mathematical foundation of



as a joint distribution b/w  $p_x$  &  $p_z$

	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	...	$\hat{x}_L$	$p_z \rightarrow$
$x_1$	0.1	0.2	0.6		0.05	} Joint dist. b/w $p_x$ & $p_z$
$x_2$						
$\vdots$						
$\vdots$						
$x_k$						

$\downarrow p_x$

Every re-distribution scheme is a joint distribution and is called a "transport-plan".

WGANs. (at 16:30):

A summary slide showing the final WGAN architecture and its objective function. It contrasts the standard GAN discriminator with the WGAN 'critic' and clearly lays out the loss functions for both the critic and the generator, highlighting the removal of the log-sigmoid function. (at

the mass from  $x$  to  $\hat{x}$ .

Avg. work done in a transport plan:

$$\int_{x, \hat{x}} \pi(x, \hat{x}) \cdot \|x - \hat{x}\| dx d\hat{x}$$

$$= \mathbb{E}_{\pi(x, \hat{x})} \|x - \hat{x}\|$$

23:50):