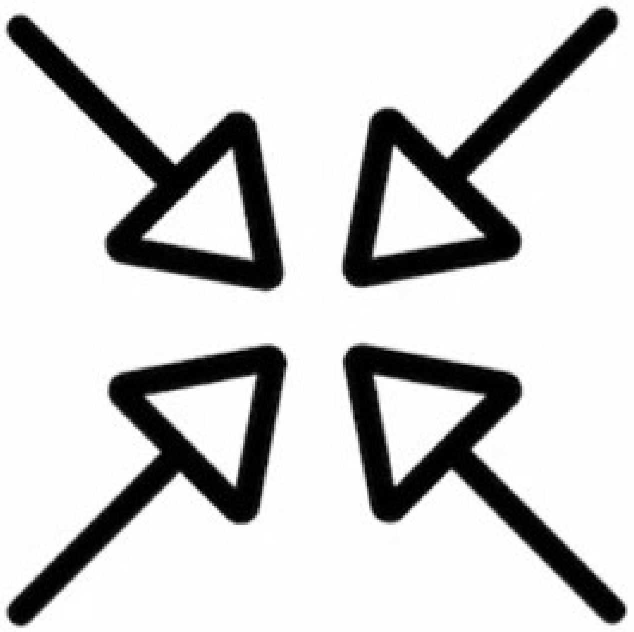
deeplearning.ai

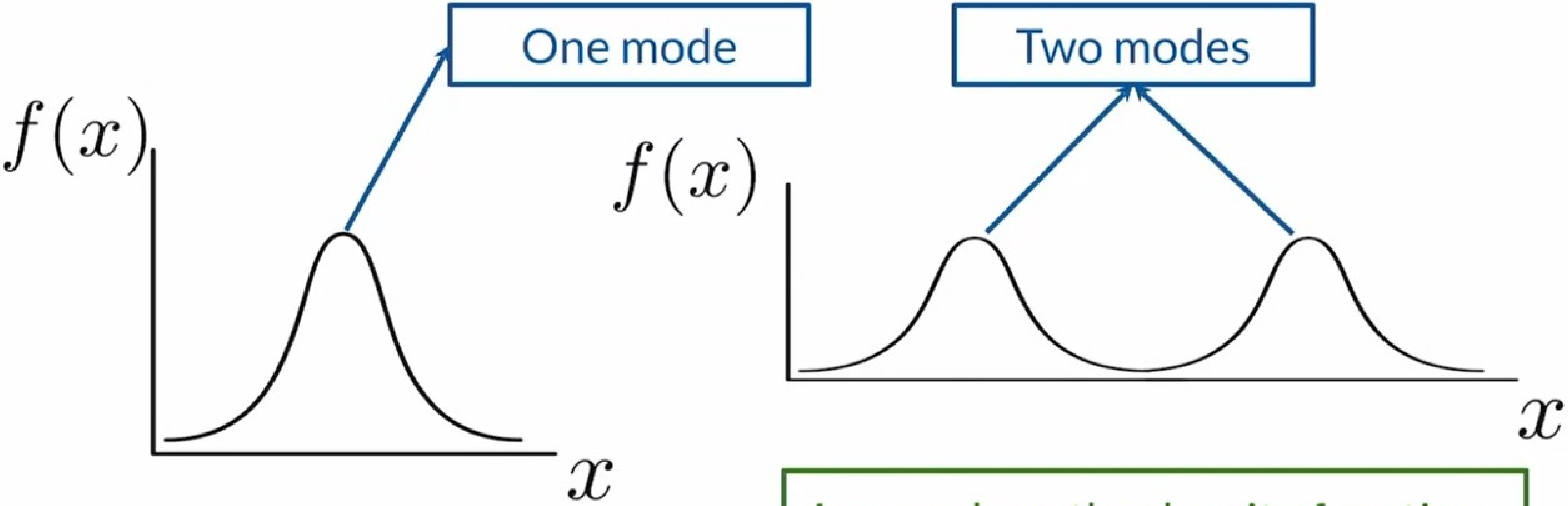
Mode Collapse

Outline

- Modes in distributions
- Mode collapse in GANs
- Intuition behind it during training



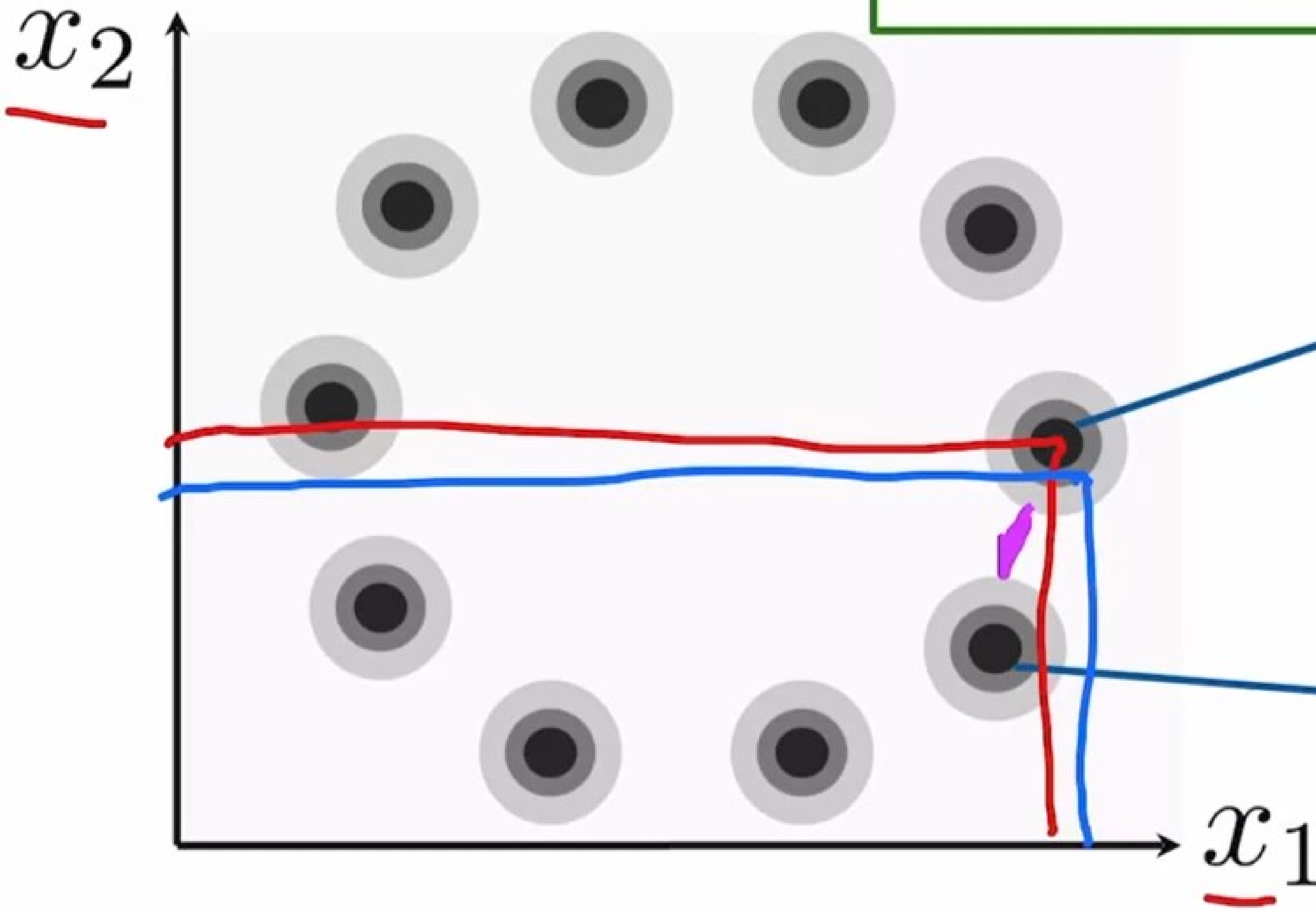
Mode Collapse

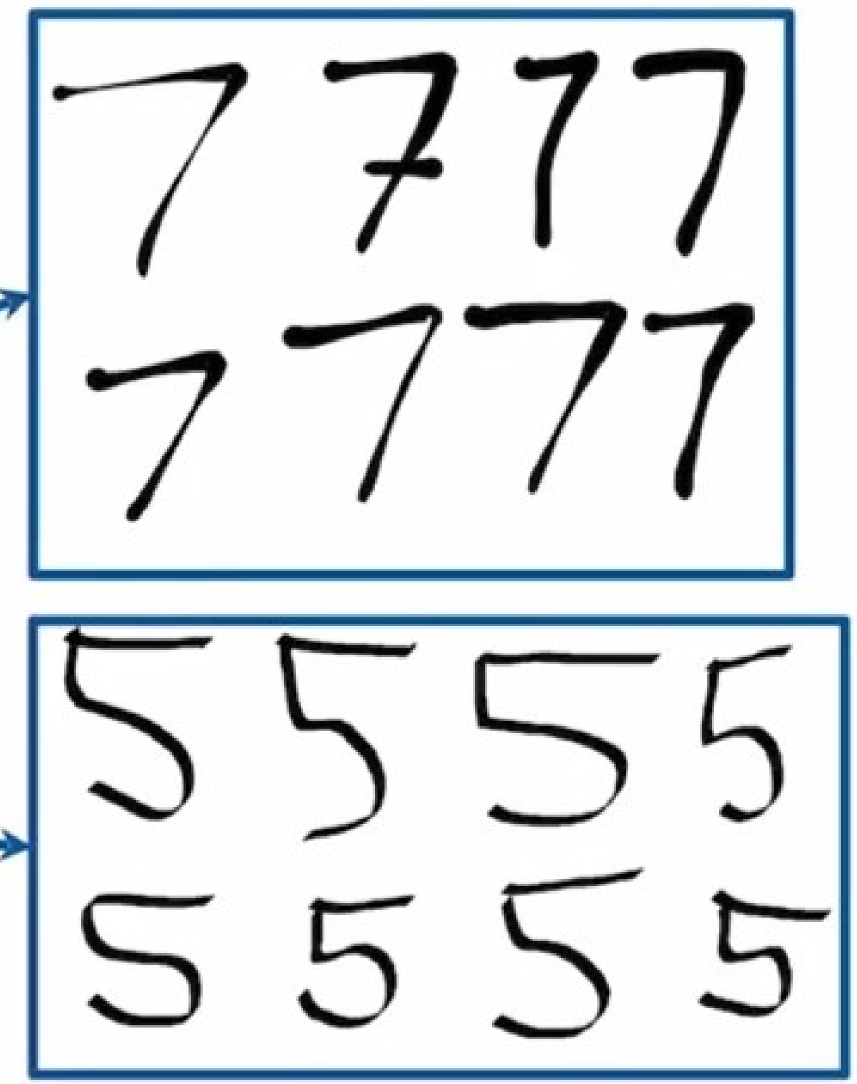


Any peak on the density function is a mode!

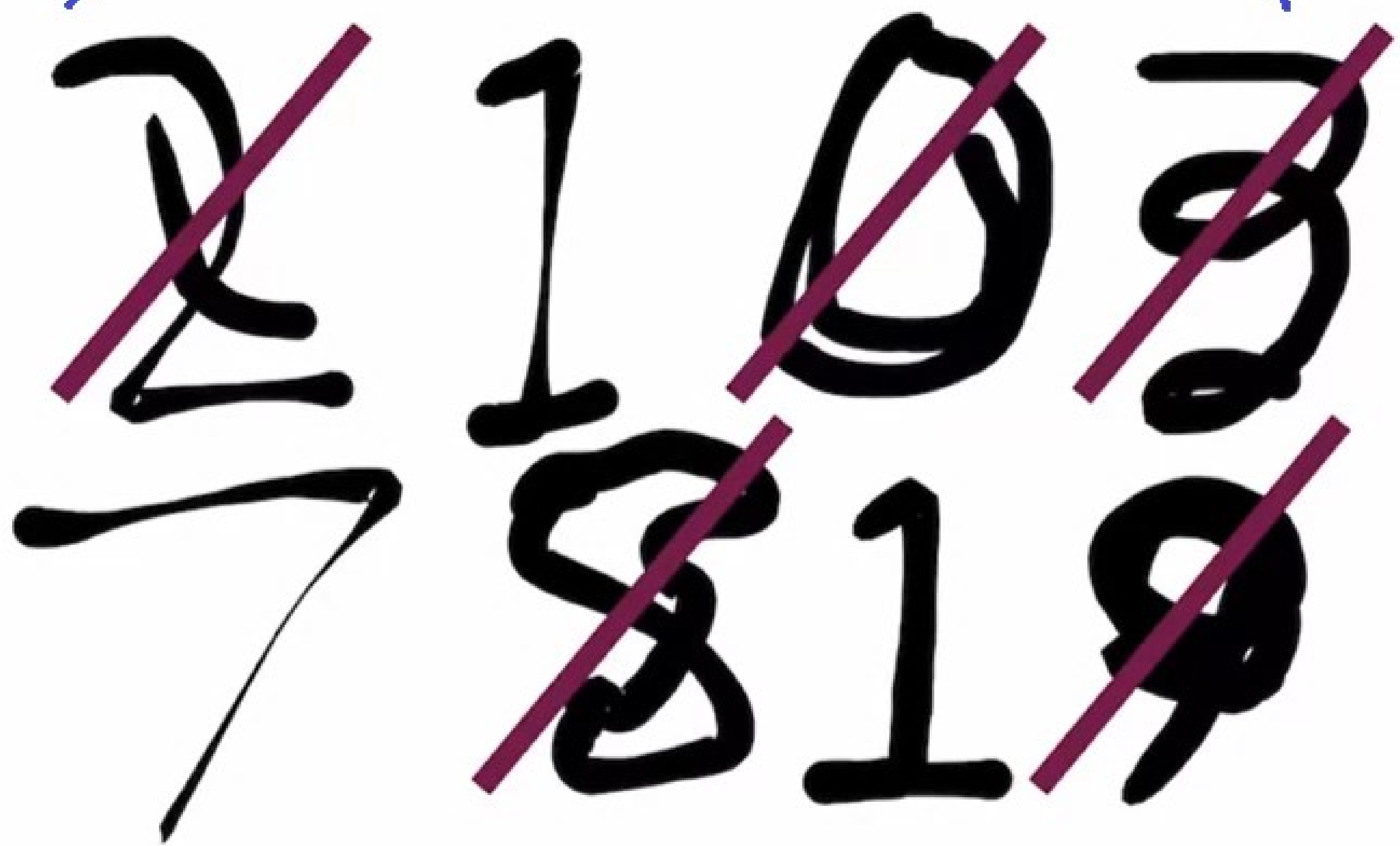
Mode Collapse

10 different modes, 1 per digit





Mode Collapse → Lesser than 10 modes → 8

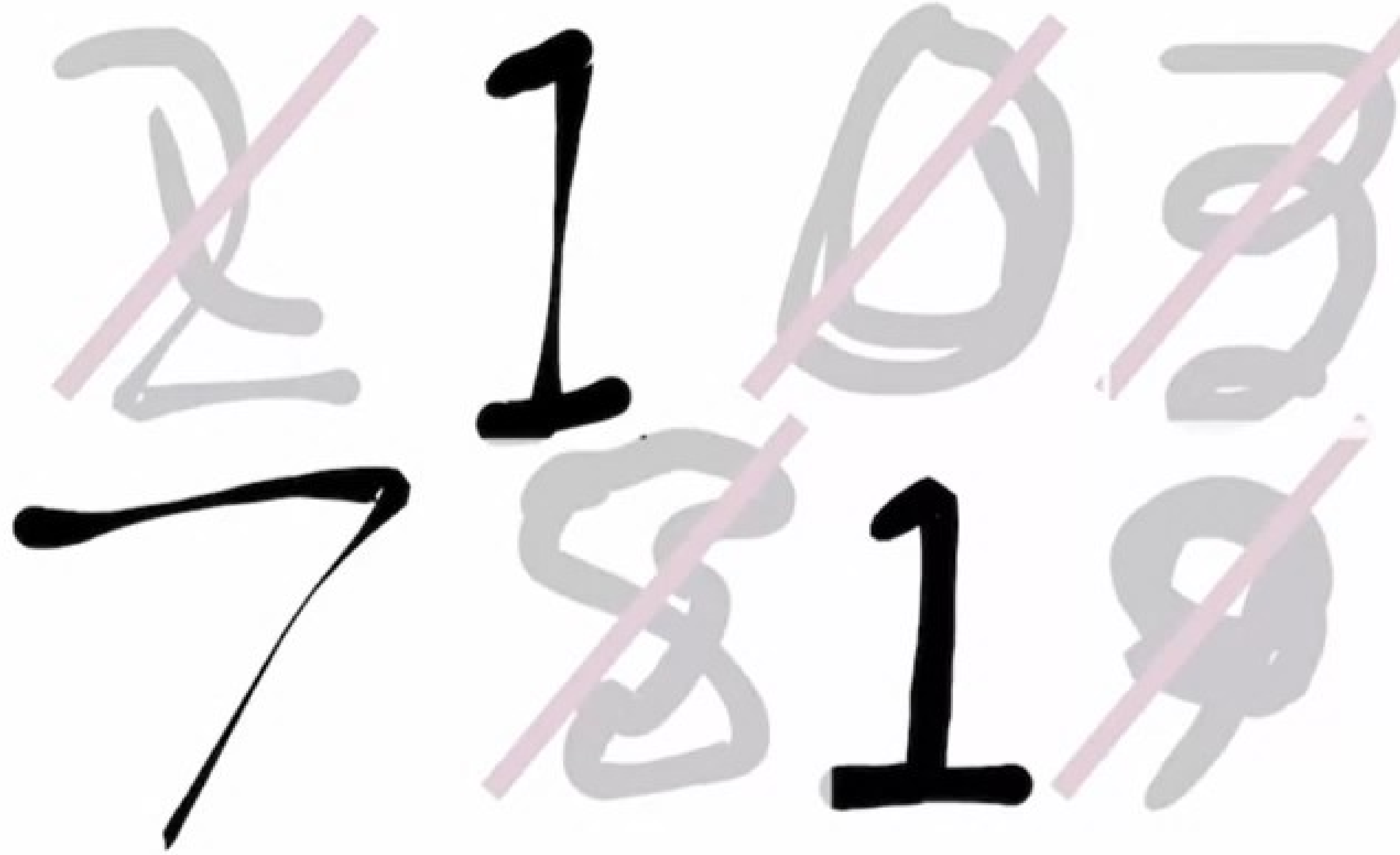


Fakes

Discriminator

The discriminator misclassifies fake handwritten digits 1 and 7. Thus generator will produce more of 1s and 7s to fool the discriminator

Mode Collapse



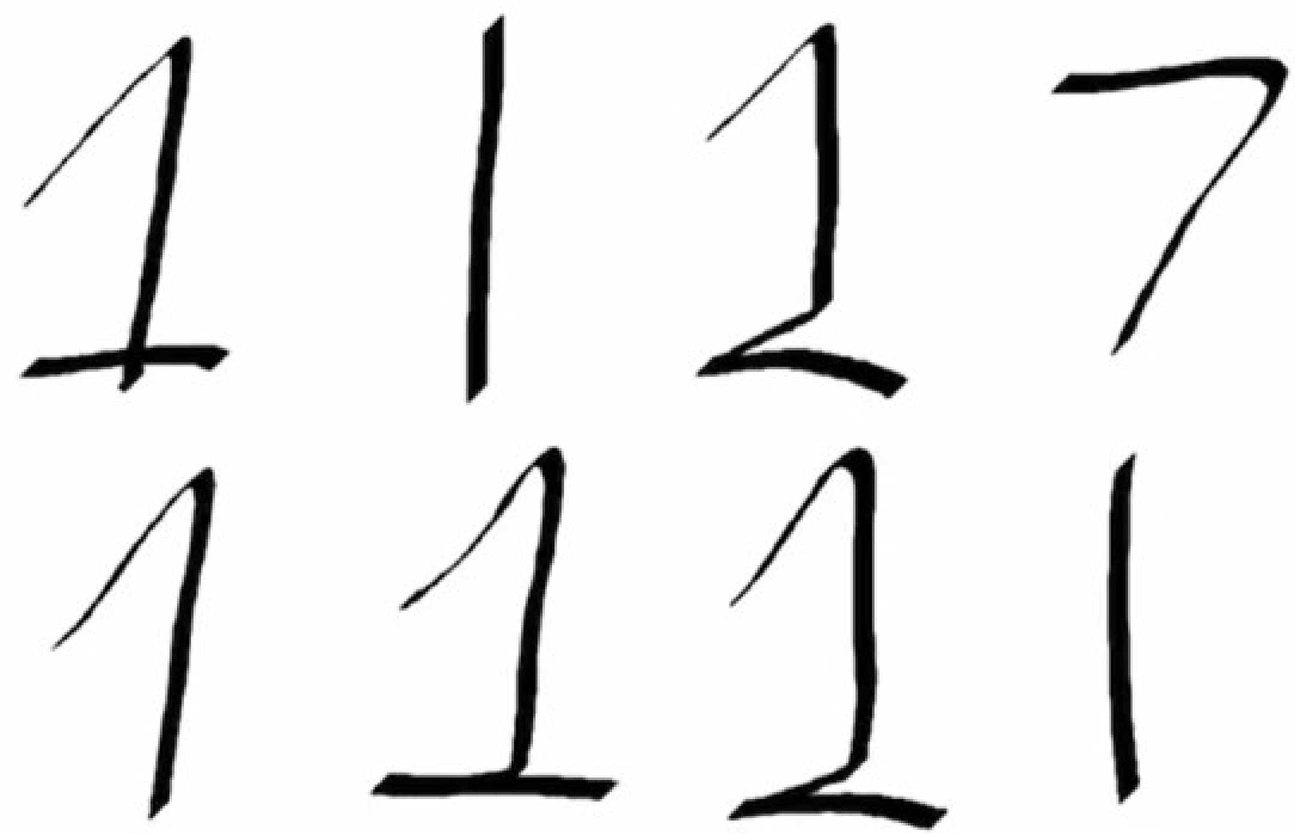
Fakes that fooled the discriminator

Generator will create more of such fakes (1s and 7s) that can fool the discriminator.

Mode Collapse

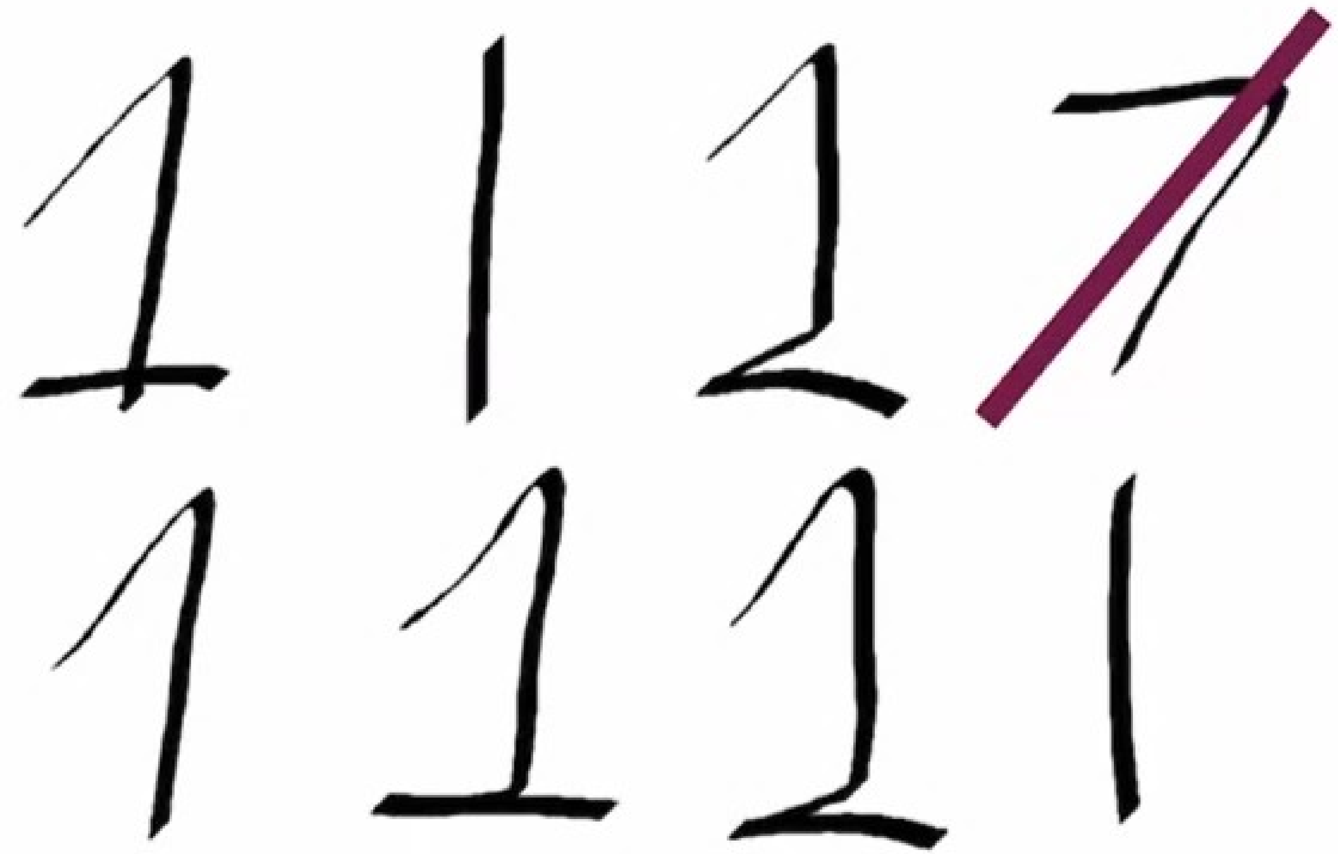
Mode collapsing into 2 modes due to cost function minima

Generator



Mode Collapse

Discriminator

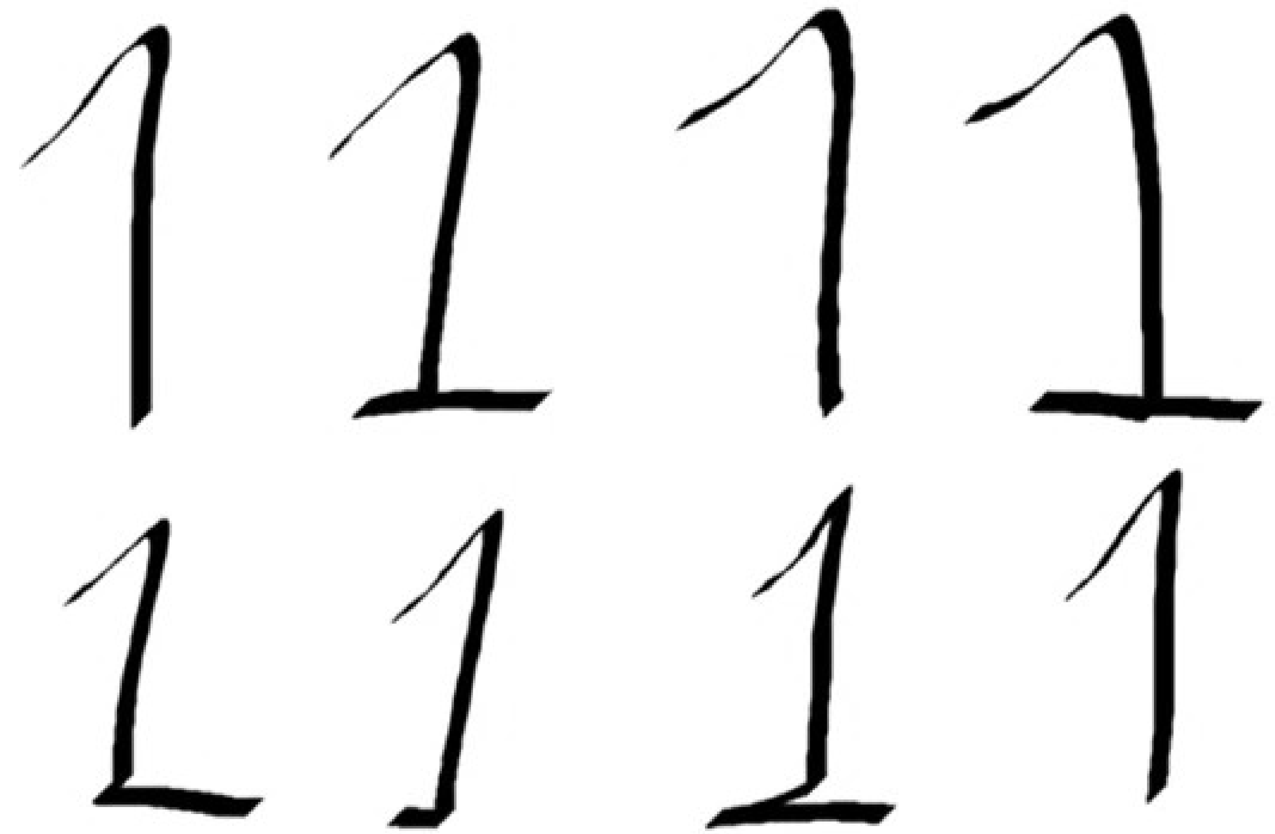


Fakes

Discriminator learns to identify fake handwritten digit 7. Thus generator will produce more of 1s now

Mode Collapse

Generator



Hence the mode now collapses to single mode. Now the generator would either learn to get out of this, otherwise it will fail to do so. In other words the mode will get out of this local minima and may get into some other cost function minima

Summary

- Modes are peaks in the distribution of features
- Typical with real-world datasets
- Mode collapse happens when the generator gets stuck in one mode

deeplearning.ai

Problem with BCE Loss

Outline

- BCE Loss and the end objective in GANs
- Problem with BCE Loss

BCE Loss in GANs

$$\underline{J(\theta)} = -\frac{1}{m}\sum_{i=1}^{m}\Big[\overbrace{y^{(i)} \cancel{\log h(x^{(i)},\theta)}}^{\text{Real}} + \overbrace{\boxed{(1-y^{(i)})\log(1-h(x^{(i)},\theta))}}^{\text{Fake}}\Big]$$

Prediction

Label

Features

Parameters

minimax

Generator → Maximize cost

↓

Fake cost

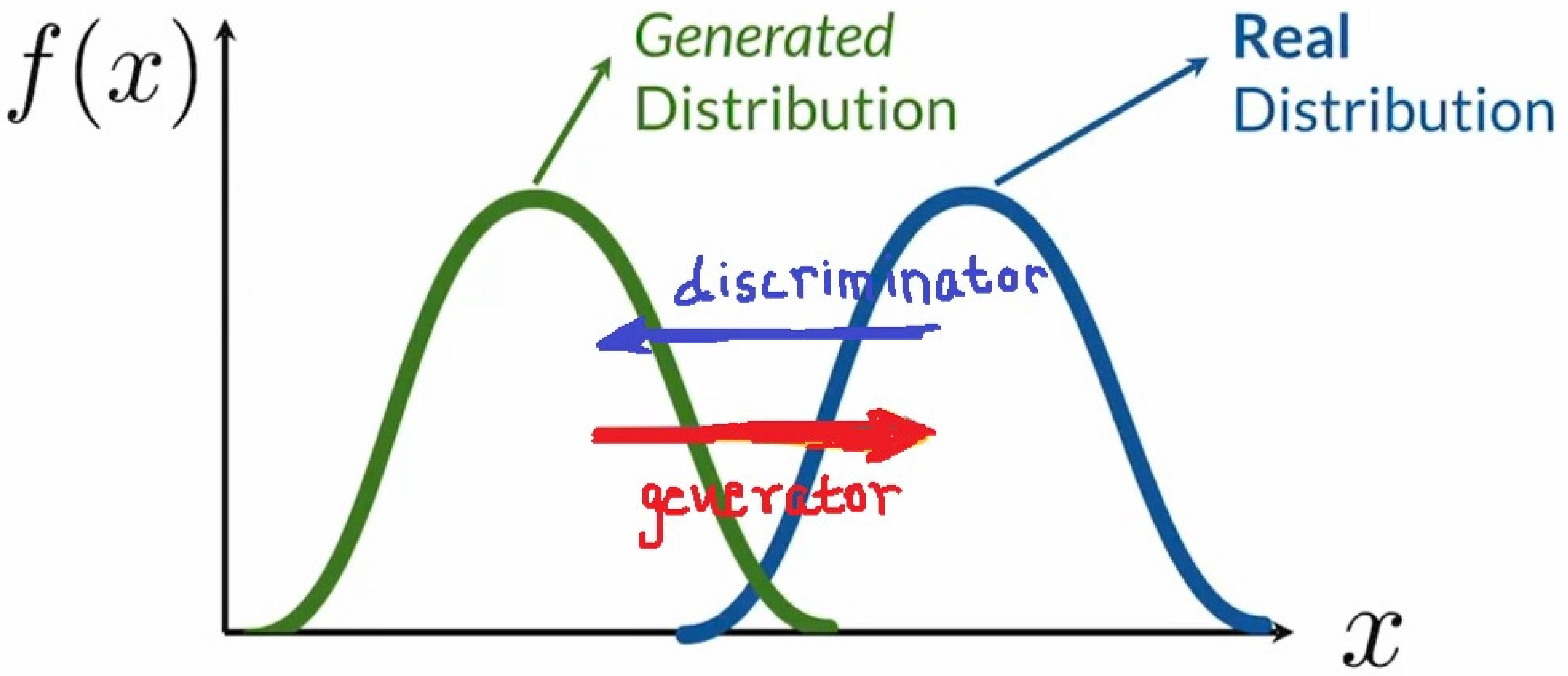
Discriminator → Minimize cost

Real

Fake

Objective in GANs

Make the generated and real distributions look similar



BCE Loss in GANs

Criticizing is more straightforward

Single output

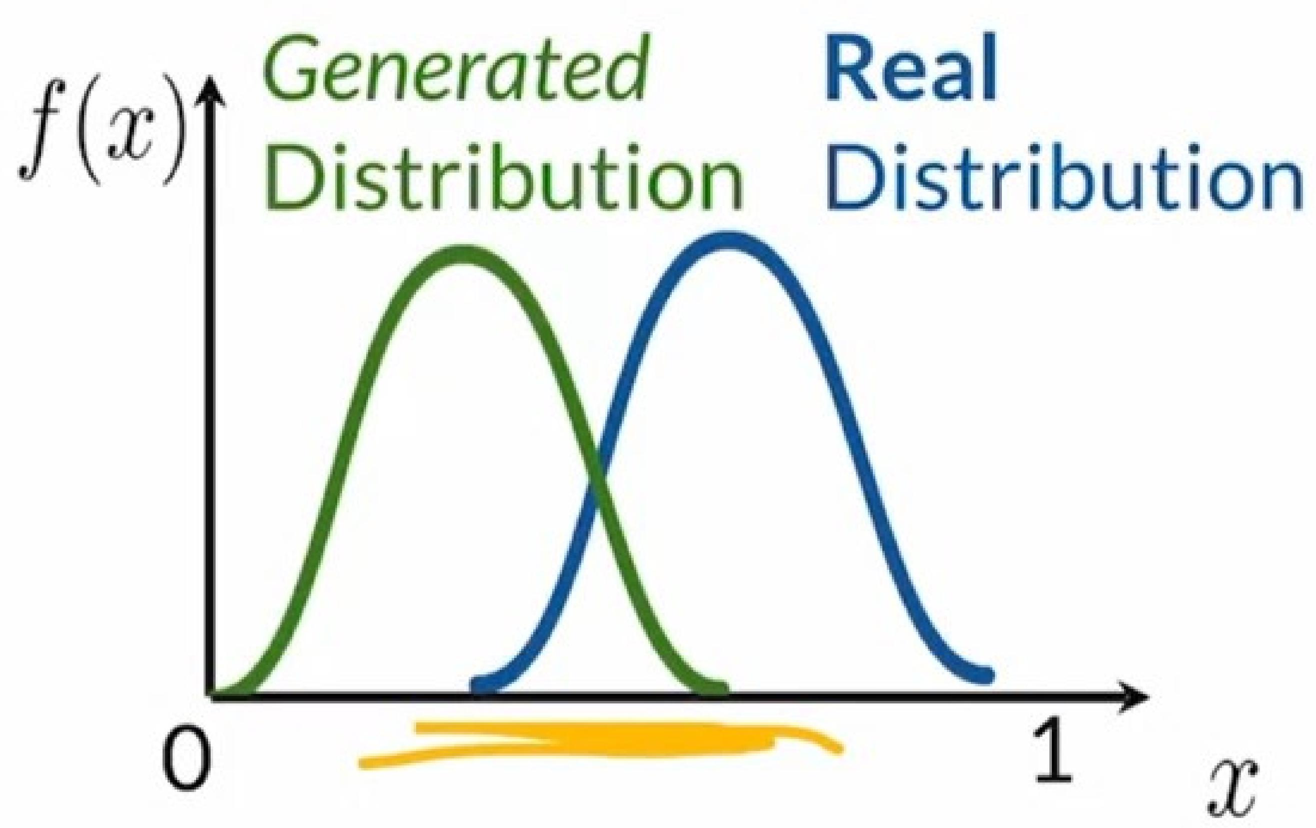
Easier to train than the generator

Complex output

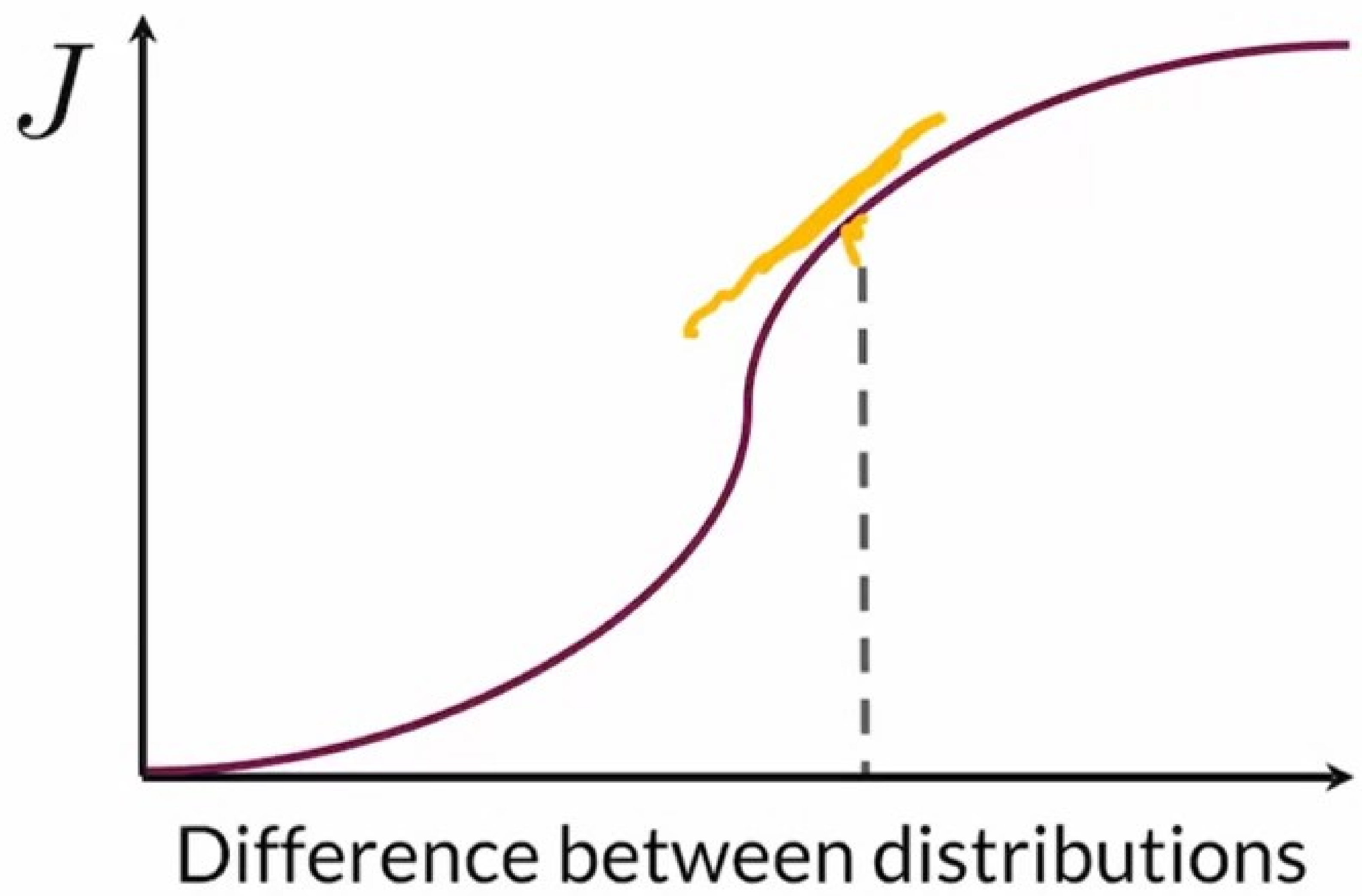
Difficult to train

Often, the discriminator gets better than the generator

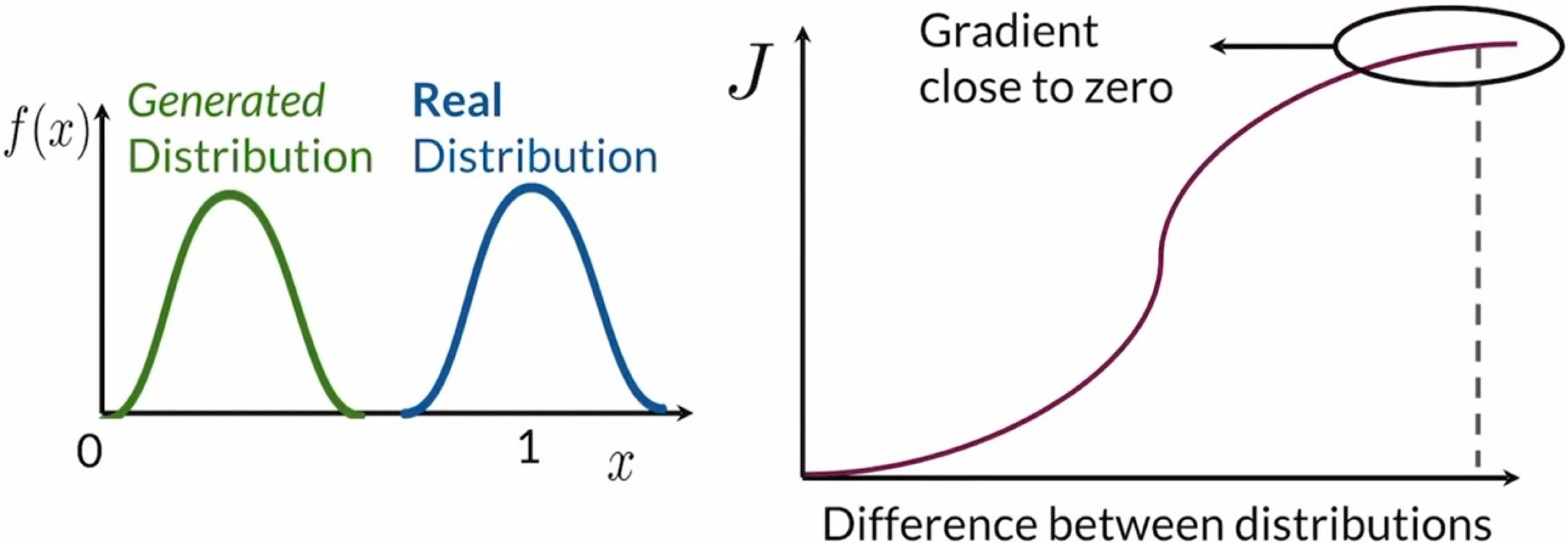
Problems with BCE Loss



Initially disc is also poor in classification



Problems with BCE Loss



Later disc will outperform generator i.e. gradient approaches zero at this minima. Hence generator will not learn anything now. This is k/a problem of vanishing gradients in GANs

Summary

- GANs try to make the real and generated distributions look similar
- When the discriminator improves too much, the function approximated by BCE Loss will contain flat regions
- Flat regions on the cost function = **vanishing gradients**

deeplearning.ai

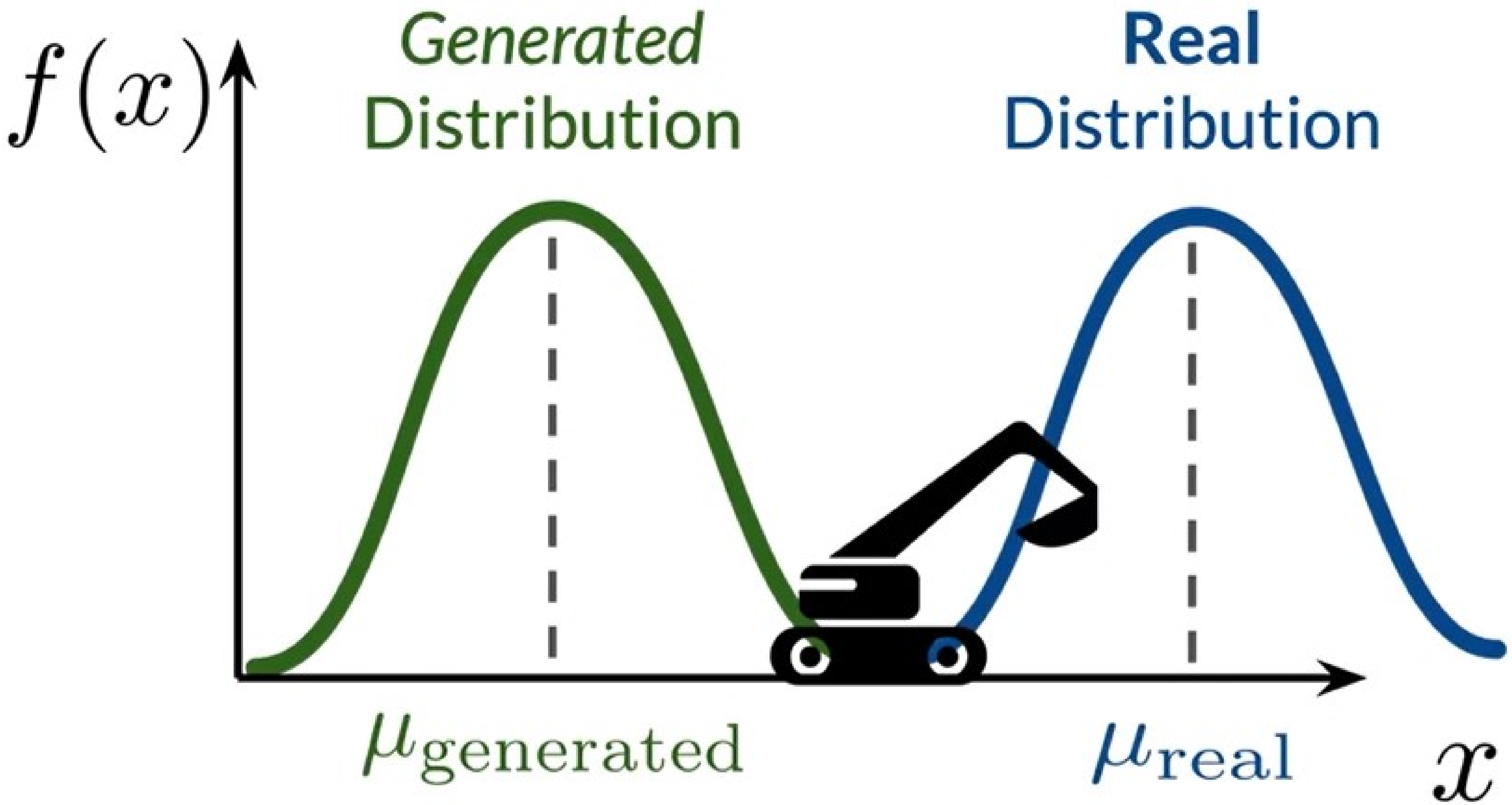
Earth Mover's Distance

Outline

new cost function

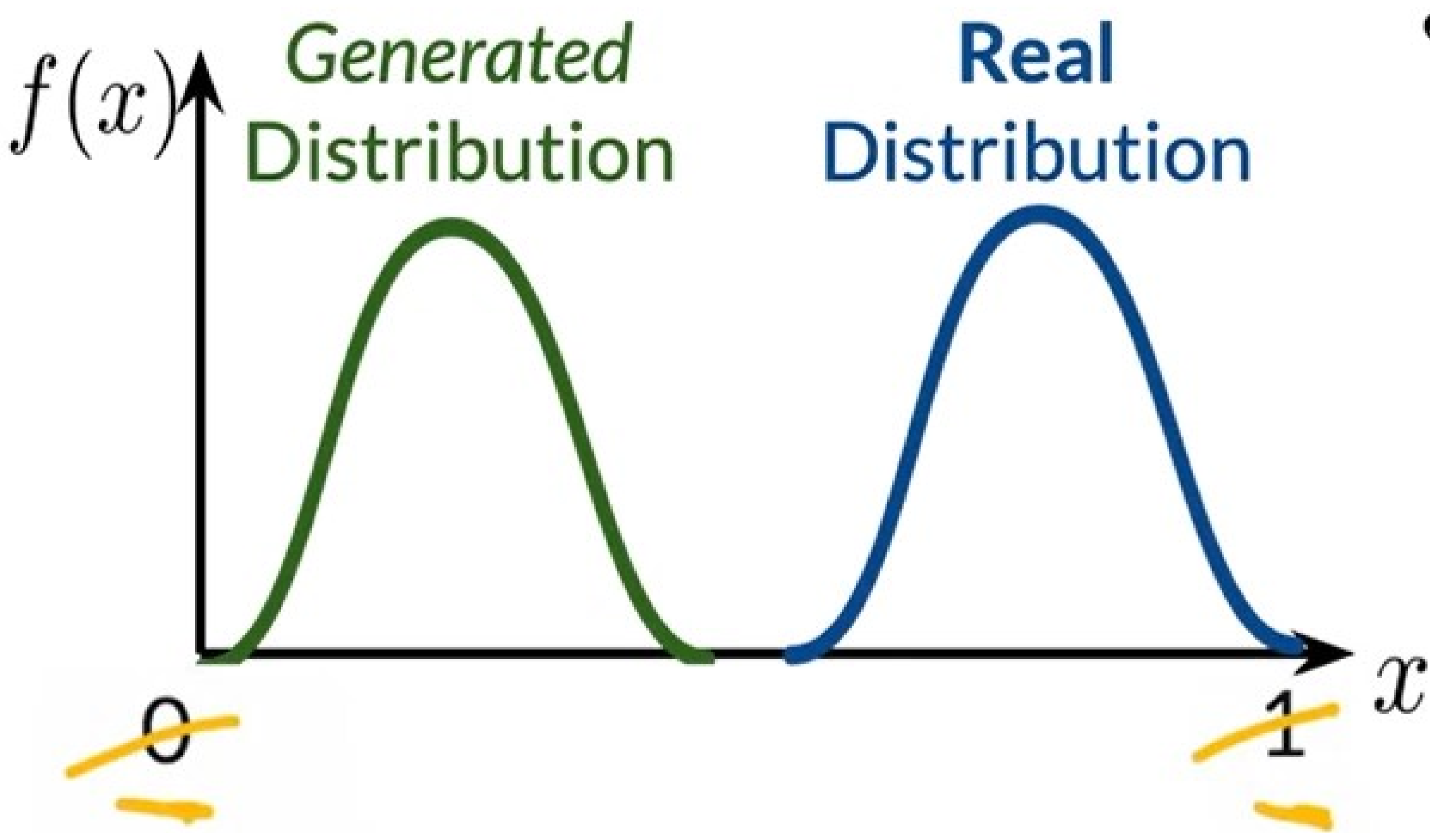
- Earth Mover's Distance (EMD)
- Why it solves the vanishing gradient problem of BCE Loss

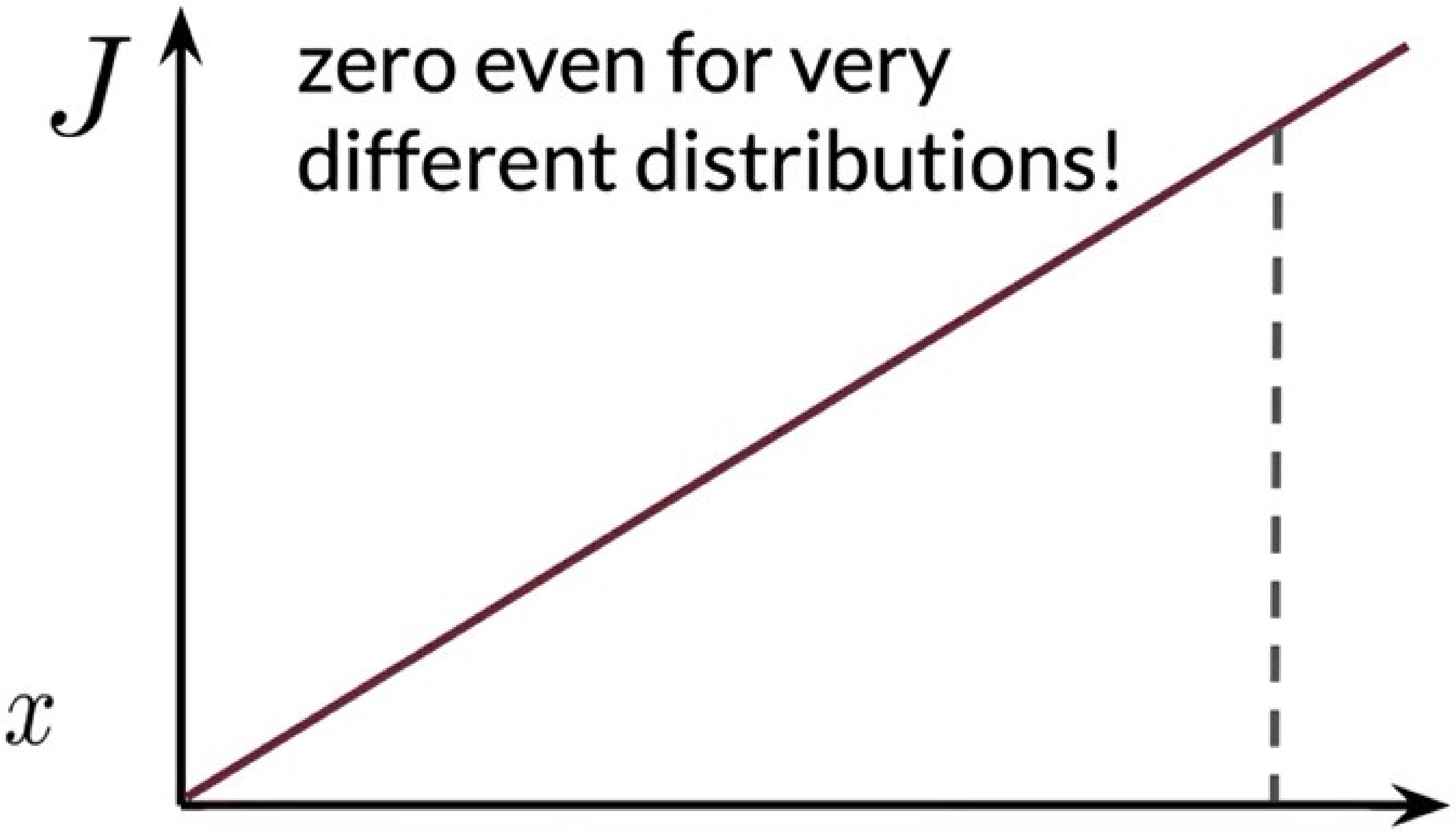
Earth Mover's Distance



Effort to make the *generated* distribution equal to the **real** distribution

Earth Mover's Distance





Summary

- Earth mover's distance (EMD) is a function of amount and distance
- Doesn't have flat regions when the distributions are very different
- Approximating EMD solves the problems associated with BCE

deeplearning.ai

Wasserstein Loss

Outline

- BCE Loss Simplified
- W-Loss and its comparison with BCE Loss

BCE Loss Simplified

$$J(\theta) = \boxed{-\frac{1}{m}\sum_{i=1}^{m}}\left[\boxed{y^{(i)}\log h(x^{(i)},\theta)} + \boxed{(1-y^{(i)})\log(1-h(x^{(i)},\theta))}\right]$$

$$\min_{d}\max_{g} -[\mathbb{E}(\log(d(x))) + \mathbb{E}(1-\log(d(g(z))))]$$

Discriminator

Minimize cost

Generator

Maximize cost

W-Loss

No log in W-Loss function and hence it is not bounded between 0 and 1

W-Loss approximates the Earth Mover's Distance

$$\min_{g} \max_{c} \; \mathbb{E}(c(x)) - \mathbb{E}(c(\overbrace{g(z)}^{\hat{x}}))$$

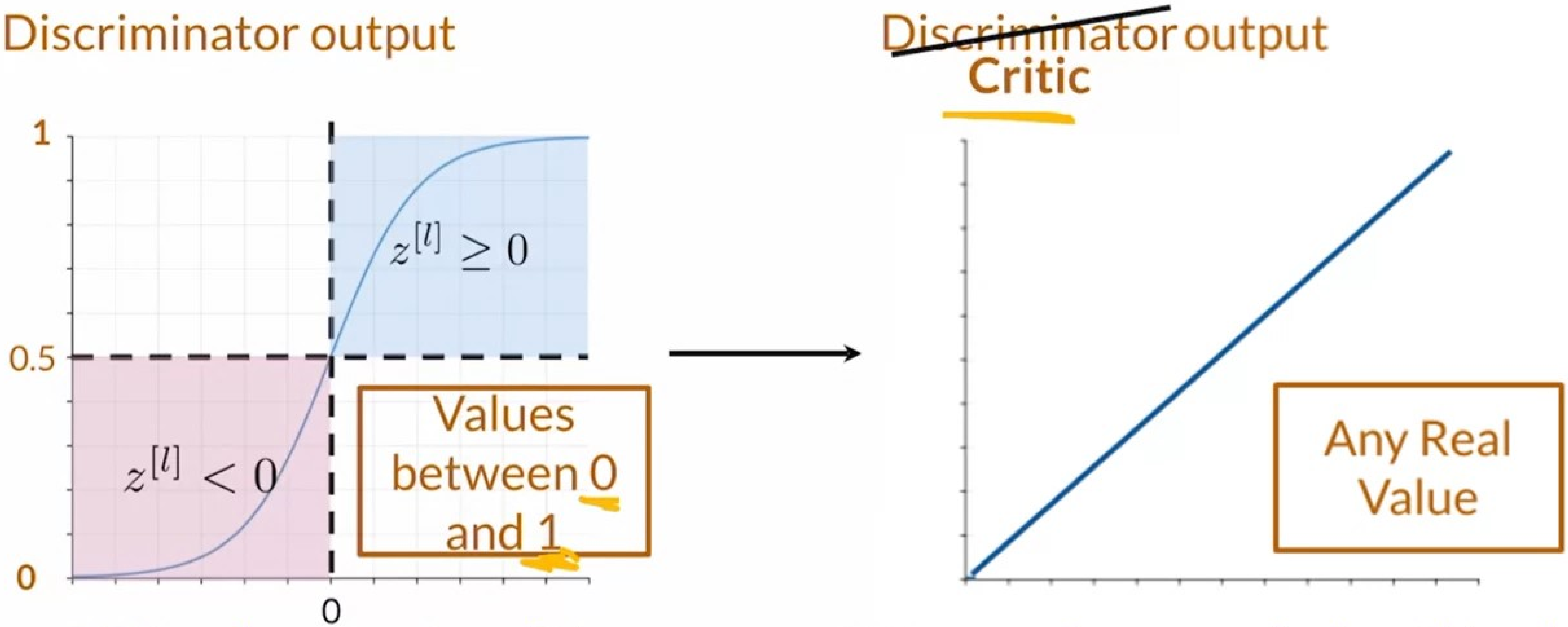
Generator

Minimize the distance

Critic

Maximize the distance

Discriminator Output



BCE Loss between 0 and 1 i.e discrimates class into 0 and 1

W Loss can be any real value. This solves the problem of vanishing gradients

W-Loss vs BCE Loss

BCE Loss	W-Loss
Discriminator outputs between 0 and 1	Critic outputs any number
$-[\mathbb{E}(\log(d(x))) + \mathbb{E}(1 - \log(d(g(z))))]$	$\mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))$

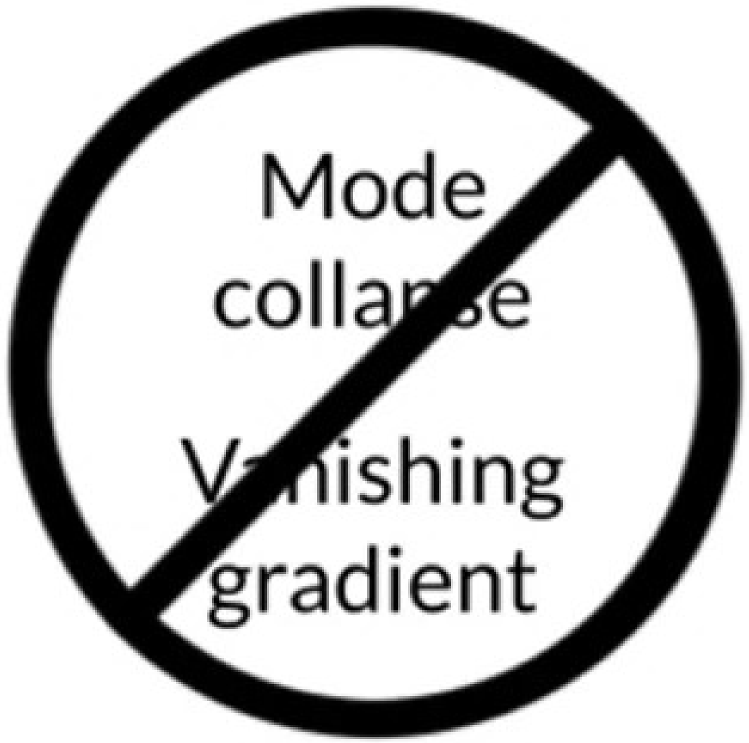
distance b/e real and fake from ground truth

distance b/w real and fake distributions

W-Loss helps with mode collapse and vanishing gradient problems

Summary

- W-Loss looks very similar to BCE Loss
- W-Loss prevents mode collapse and vanishing gradient problems



Condition on Wasserstein Critic

deeplearning.ai

Outline

- Continuity condition on the critic's neural network
- Why this condition matters

Condition on W-Loss

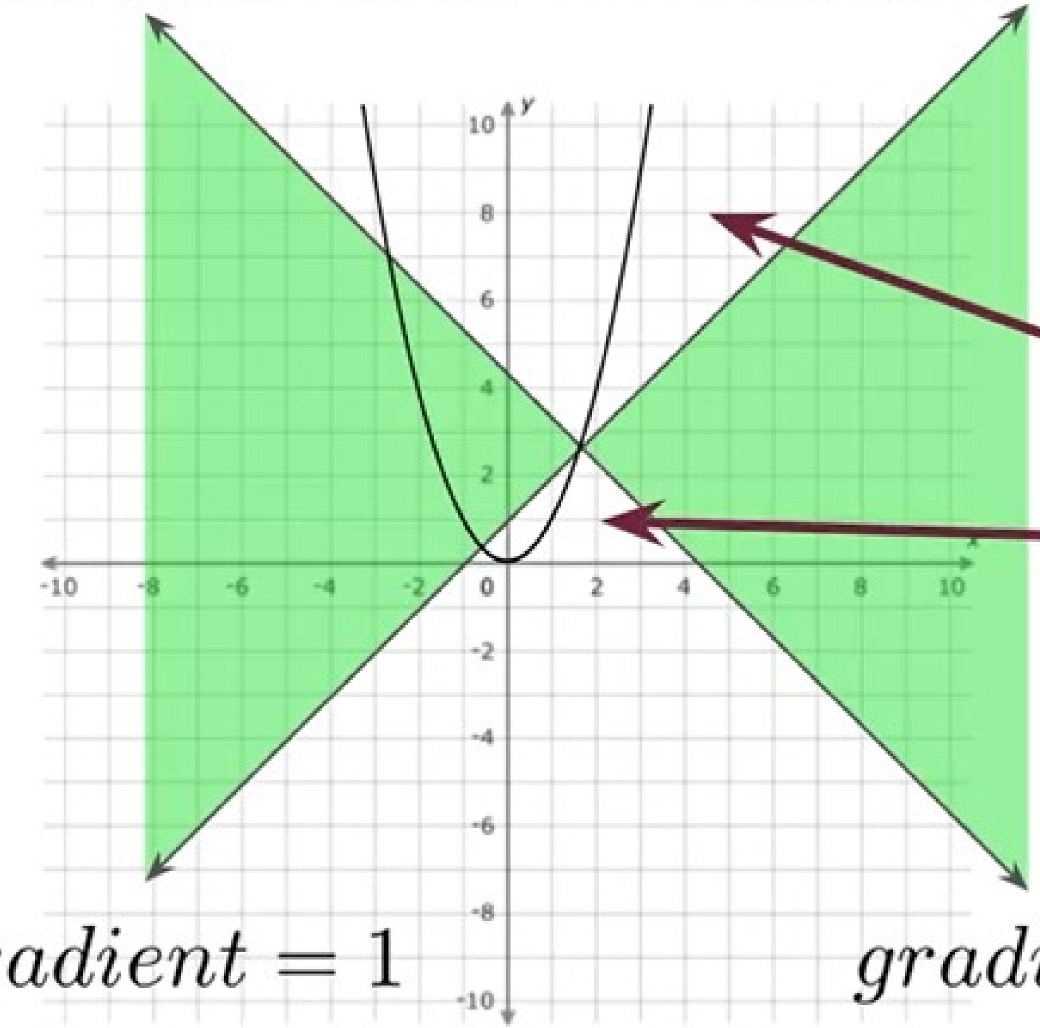
$$\min_{g} \max_{\boxed{c}} \mathbb{E}(\boxed{c}(x)) - \mathbb{E}(\boxed{c}(g(z)))$$

Needs to be 1-Lipschitz Continuous

1-L

Condition on W-Loss

Critic needs to be **1**-L Continuous



$gradient = 1$ $gradient = -1$

The norm of the gradient should be at most **1** for *every point*

In other words, the function must grow within the green region.

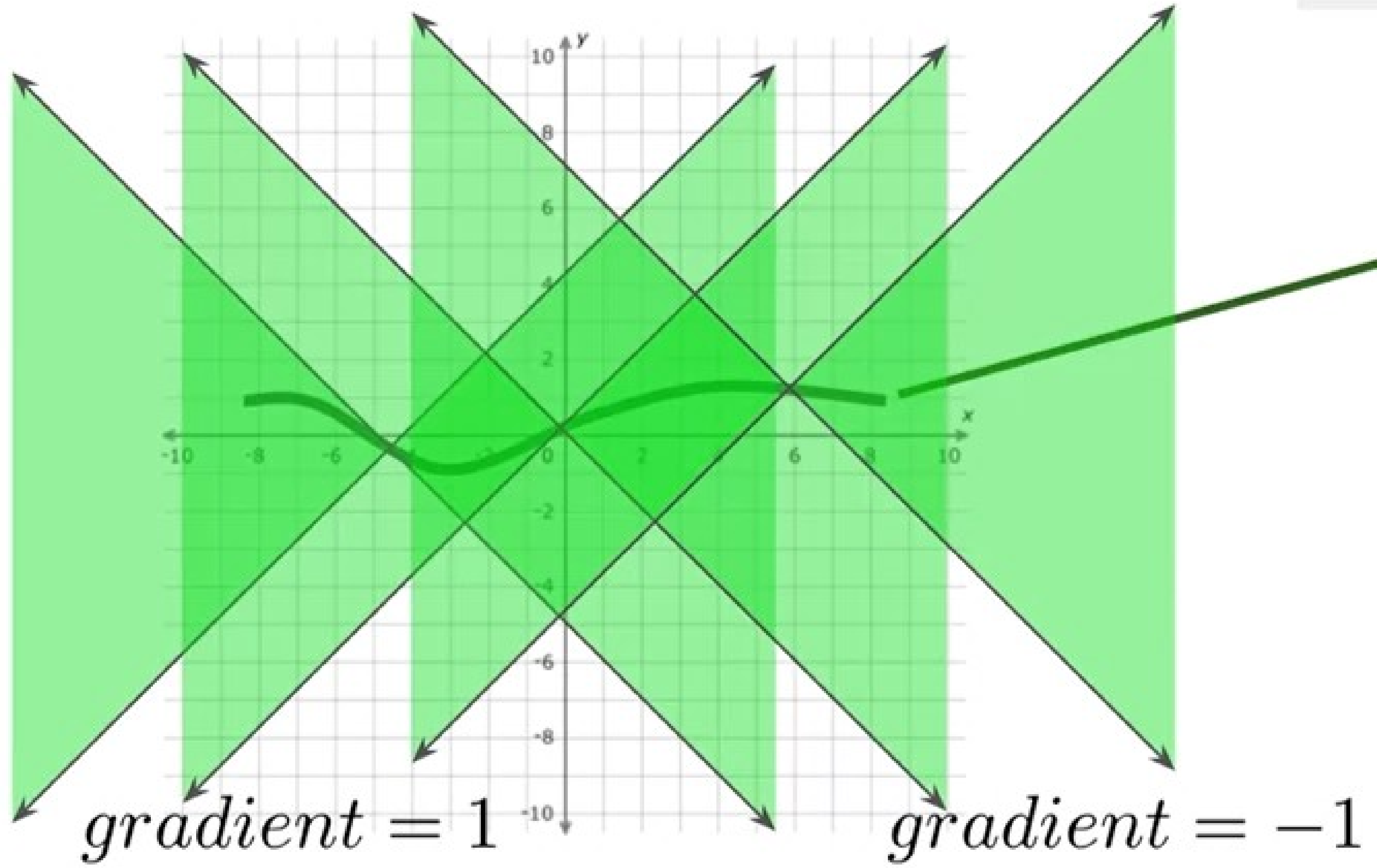
Not 1-L Continuous

The function growth should be less than linear growth. This is crucial to avoid excess growth of W-Loss. This also ensures the loss to be in a valid range.

Condition on W-Loss

Critic needs to be **1**-L Continuous

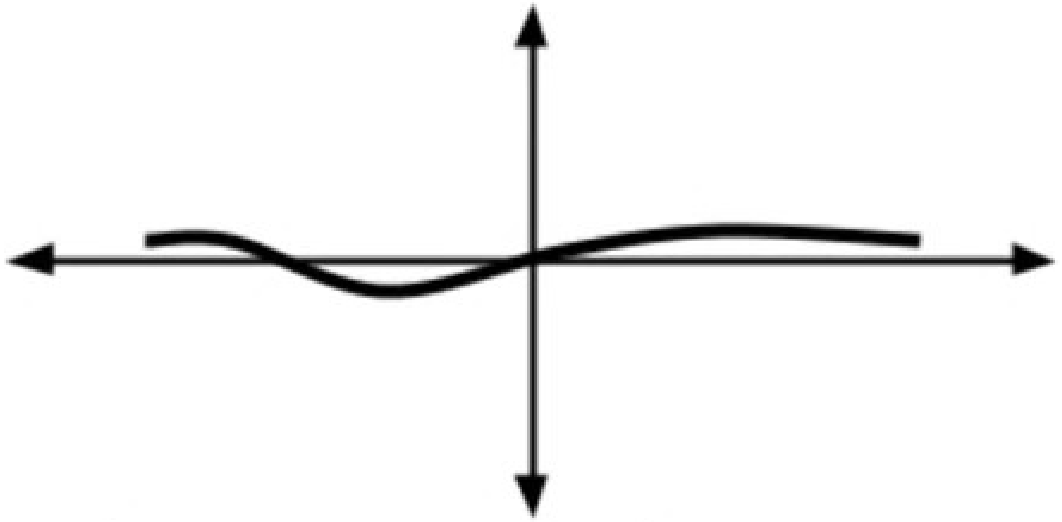
The norm of the gradient should be at most **1** for *every point*



Needed for training stable neural networks with W-Loss

Summary

- Critic's neural network needs to be 1-L Continuous when using W-Loss
- This condition ensures that W-Loss is validly approximating Earth Mover's Distance



Next up, we will look some methods to ensure that this condition is satisfied

deeplearning.ai

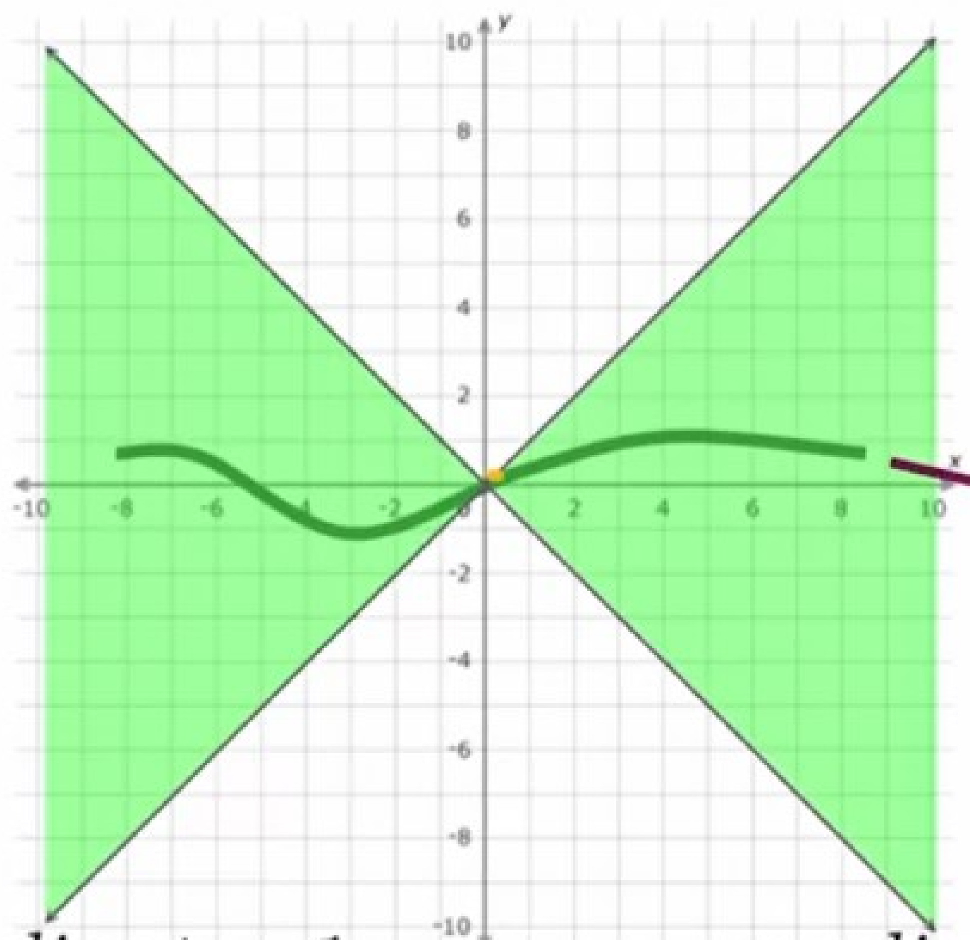
1-Lipschitz Continuity Enforcement

Outline

- Weight clipping and gradient penalty
- Advantages of gradient penalty

1-L Enforcement

Critic needs to be 1-L Continuous



$gradient = 1$ $\qquad\qquad$ $gradient = -1$

Norm of the gradient at most 1

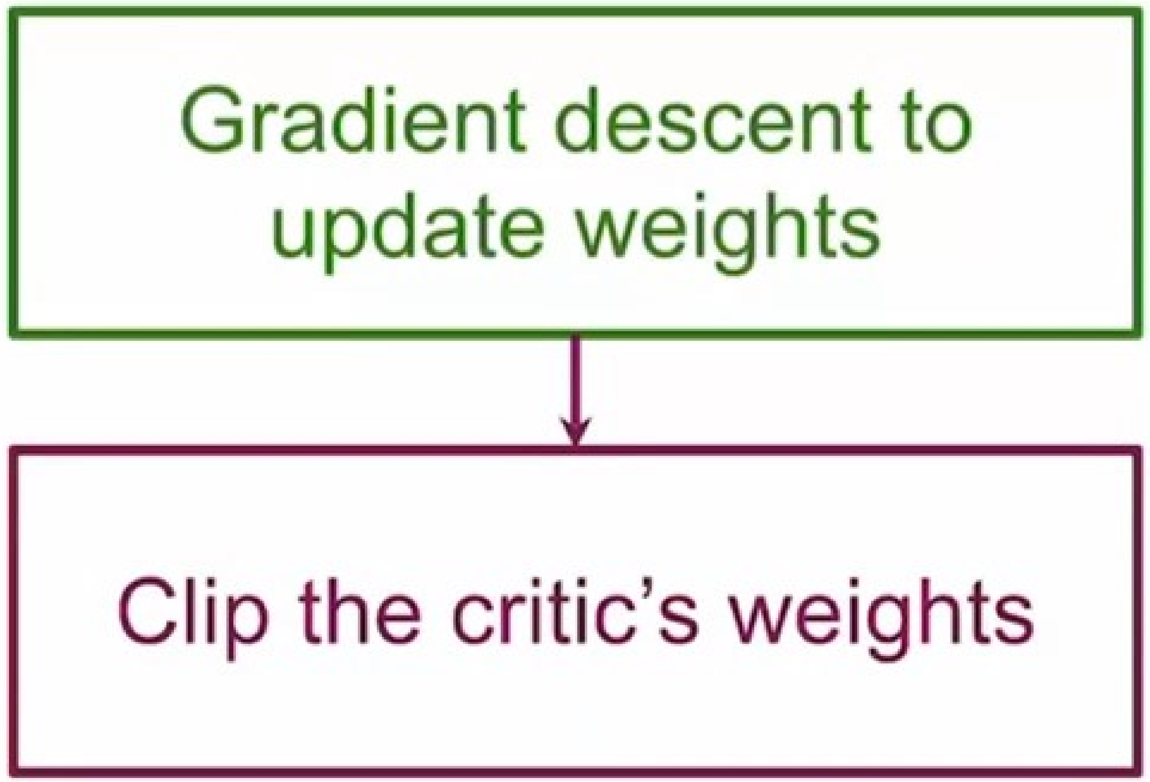
Δ

$$||\nabla f(x)||_2 \leq 1$$

Slope of the function
at most 1

1-L Enforcement: Weight Clipping

Weight clipping forces the weights of the critic to a fixed interval



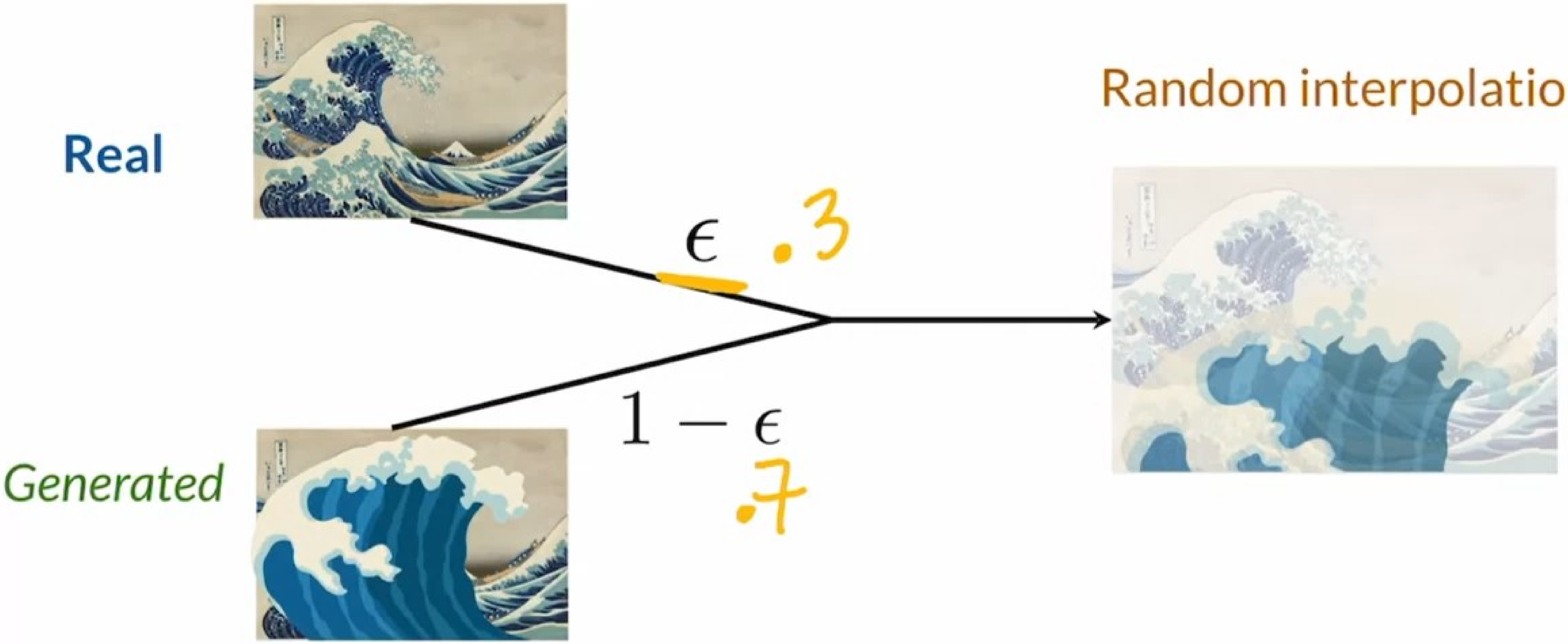
Limits the learning ability of the critic

1-L Enforcement: Gradient Penalty

$$\min_{g} \max_{c} \; \mathbb{E}(c(x)) - \mathbb{E}(c(g(z))) \boxed{+ \lambda \text{reg}}$$

Regularization of the critic's gradient

1-L Enforcement: Gradient Penalty



1-L Enforcement: Gradient Penalty

$$(||\nabla c(\boxed{\hat{x}})||_2 - 1)^2$$

Regularization term

$$\downarrow$$

$$\epsilon \boxed{x} + (1 - \epsilon) \boxed{g(z)}$$

Interpolation

Real

Generated

Putting It All Together

$$\min_{g} \max_{c} \boxed{\mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))} \boxed{+\lambda\mathbb{E}(||\nabla c(\hat{x})||_2 - 1)^2}$$

Makes the GAN less prone to **mode collapse** and **vanishing gradient**

Tries to make the critic be 1-L Continuous, for the loss function to be **continuous and differentiable**

Summary

- Weight clipping and gradient penalty are ways to enforce 1-L continuity
- Gradient penalty tends to work better