

Copyright Notice

These slides are distributed under the Creative Commons License.

[DeepLearning.AI](#) makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite [DeepLearning.AI](#) as the source of the slides.

For the rest of the details of the license, see

<https://creativecommons.org/licenses/by-sa/2.0/legalcode>

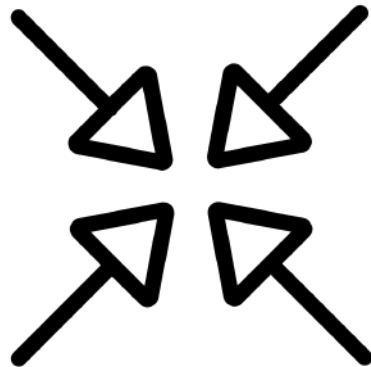


deeplearning.ai

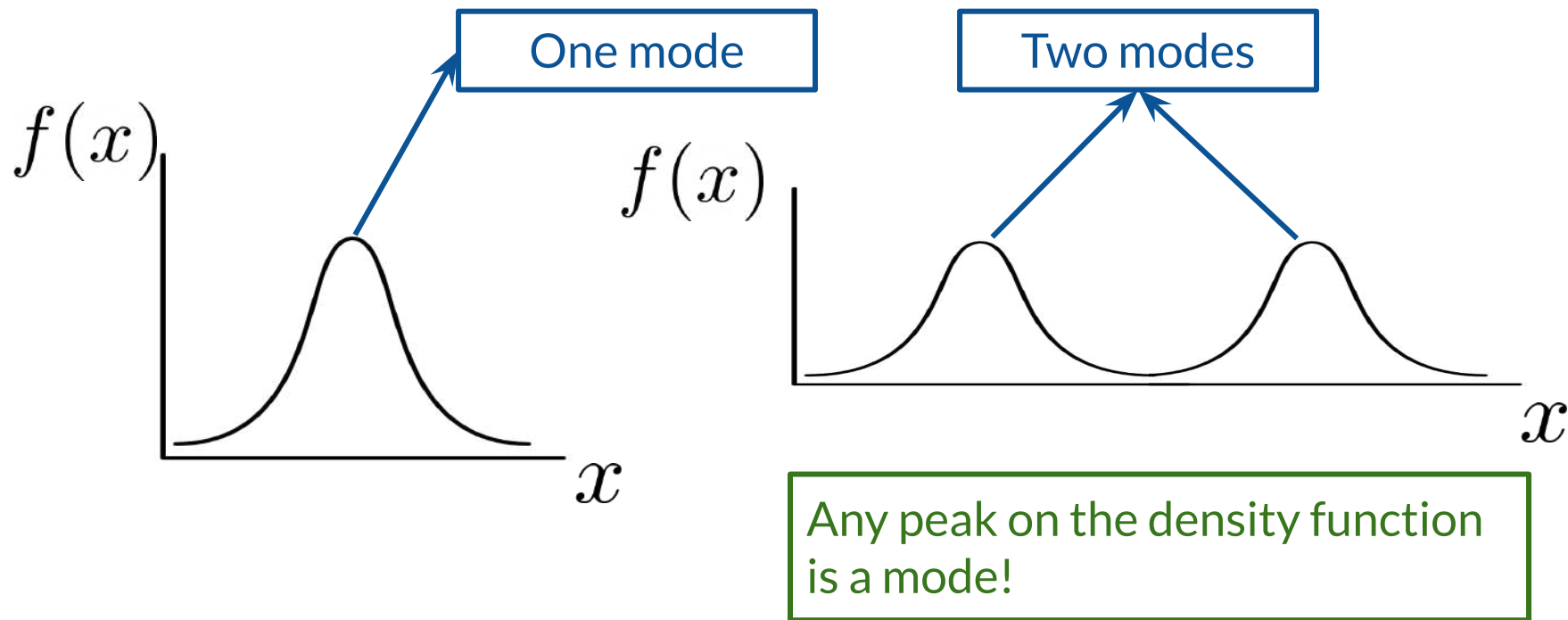
Mode Collapse

Outline

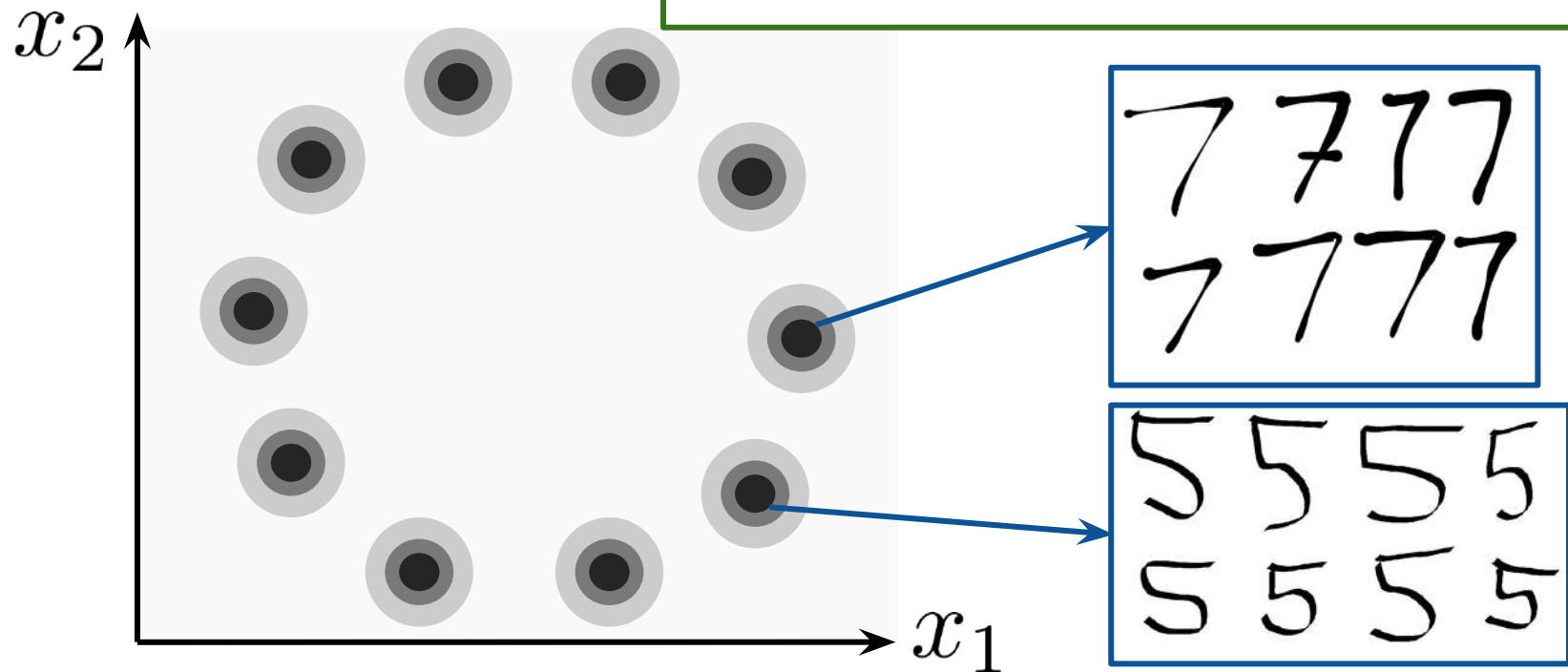
- Modes in distributions
- Mode collapse in GANs
- Intuition behind it during training



Mode Collapse



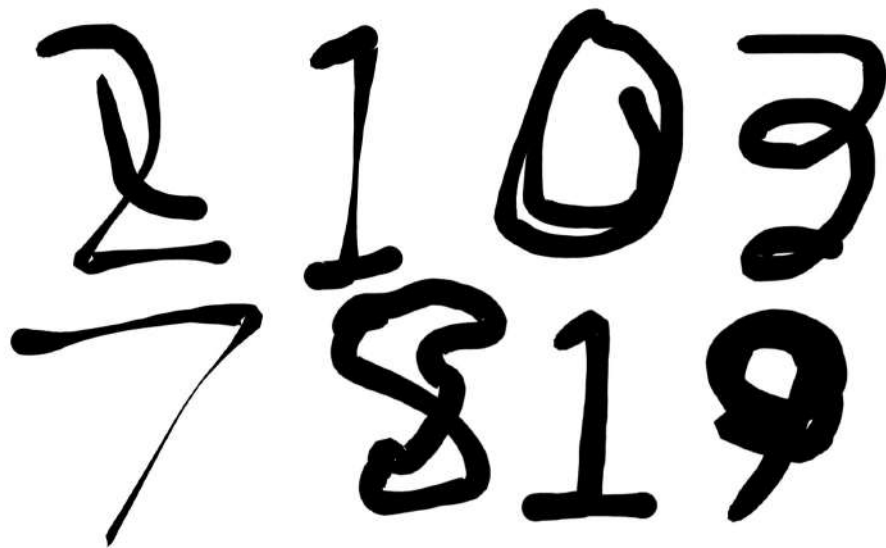
Mode Collapse



Mode Collapse



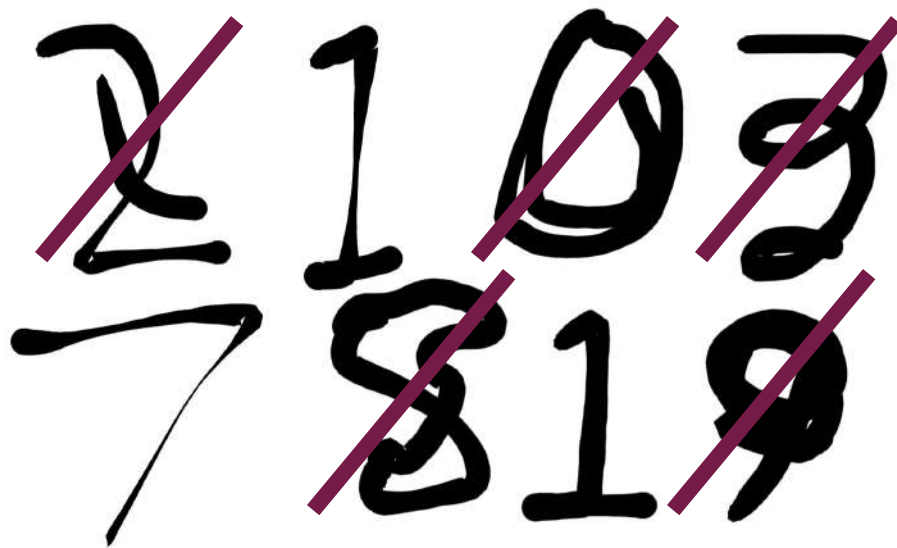
Discriminator



Mode Collapse



Discriminator

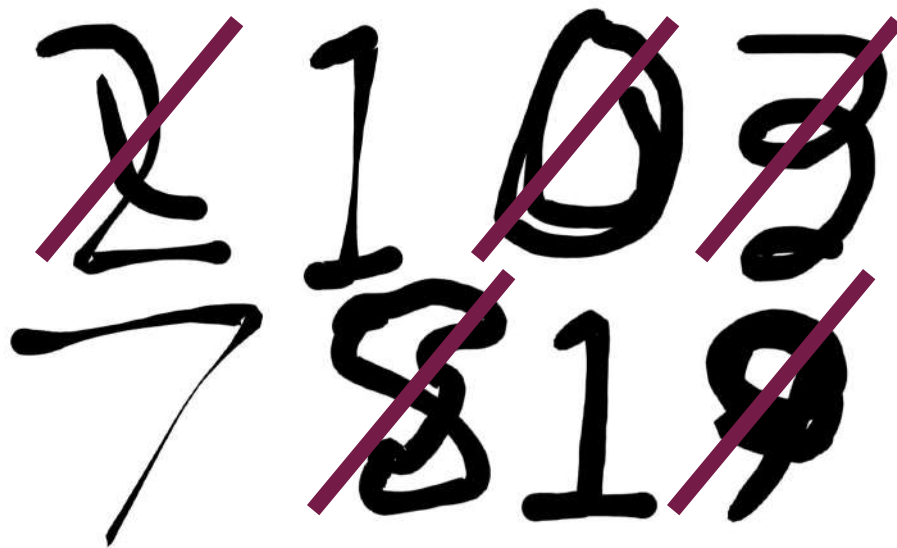


Fakes

Mode Collapse



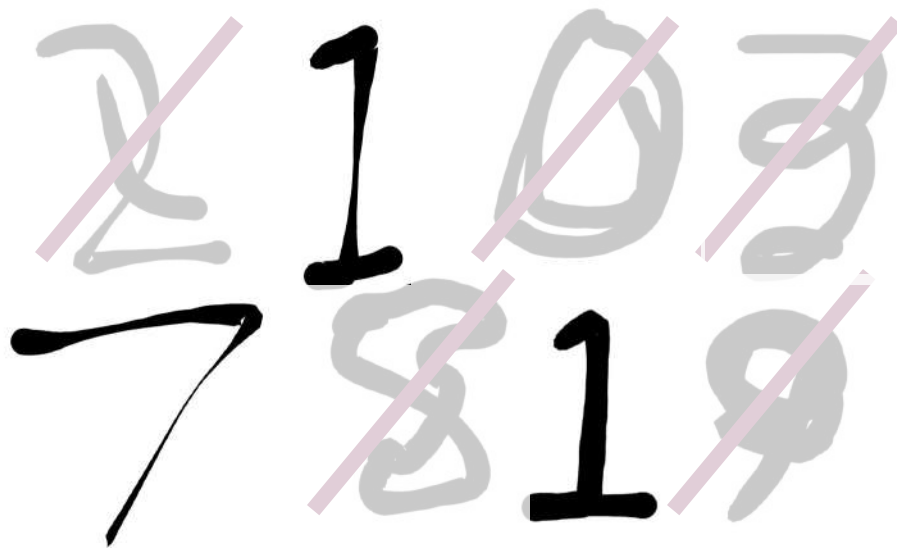
Generator



Mode Collapse



Generator

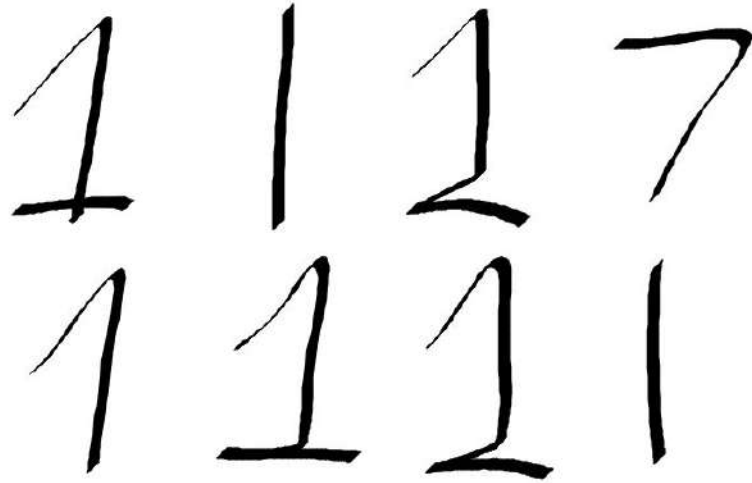


Fakes that
fooled the
discriminator

Mode Collapse



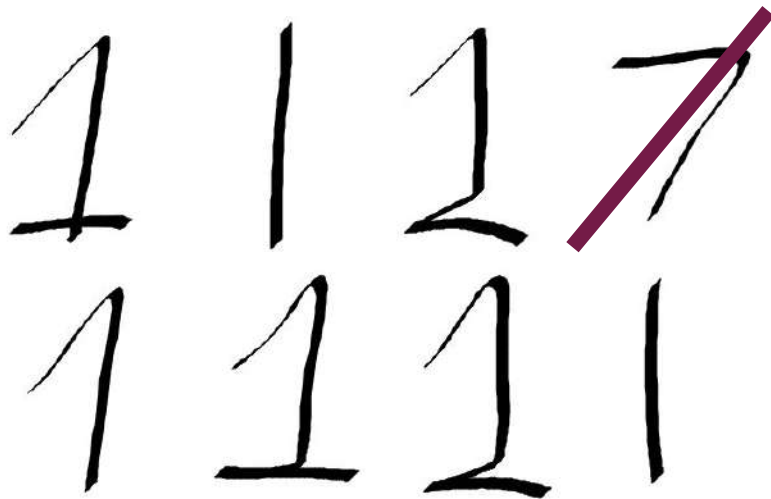
Generator



Mode Collapse



Discriminator

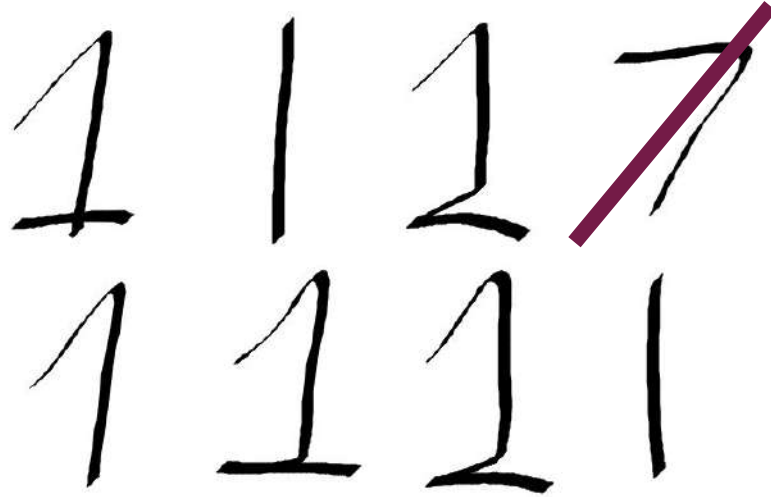


Fakes

Mode Collapse



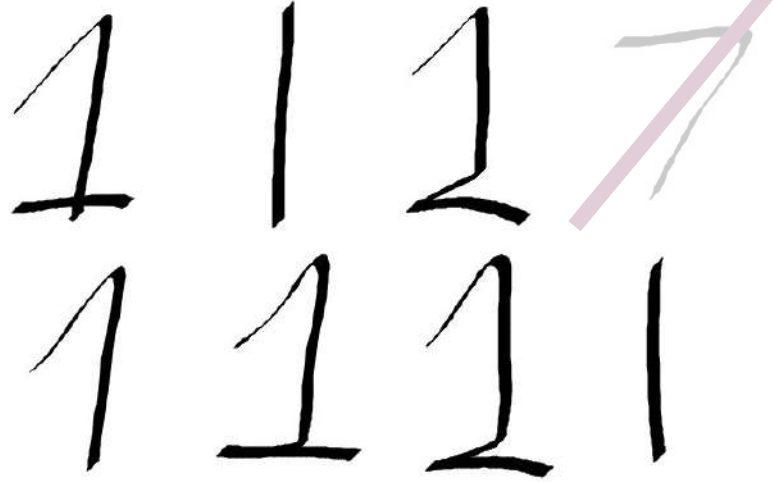
Generator



Mode Collapse



Generator

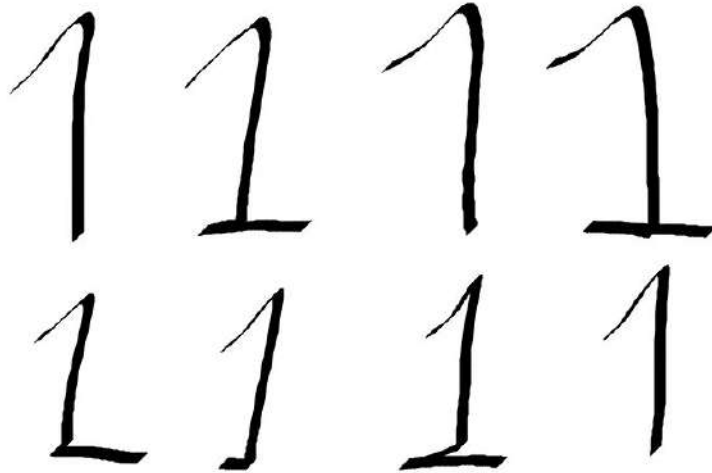


Fakes that
fooled the
discriminator

Mode Collapse



Generator



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



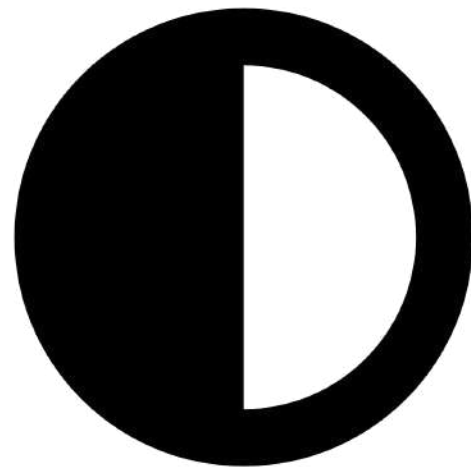


deeplearning.ai

Problem with BCE Loss

Outline

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



BCE Loss in GANs

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

Prediction

Label

Features

Parameters



Generator

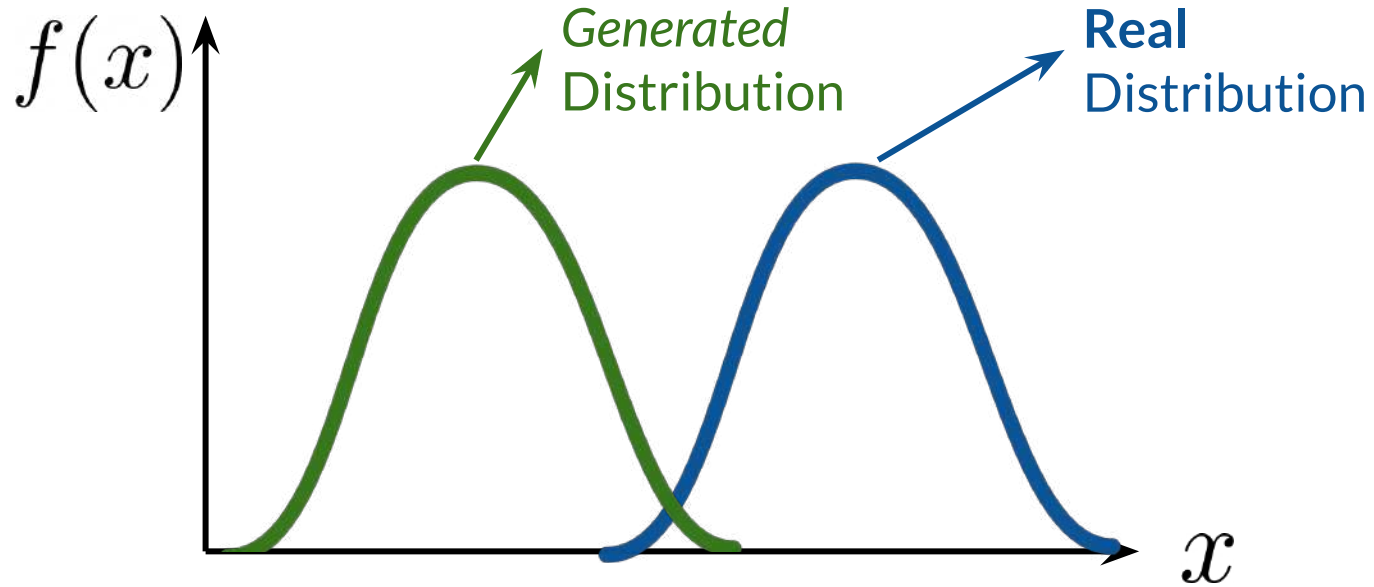
Maximize
cost



Discriminator

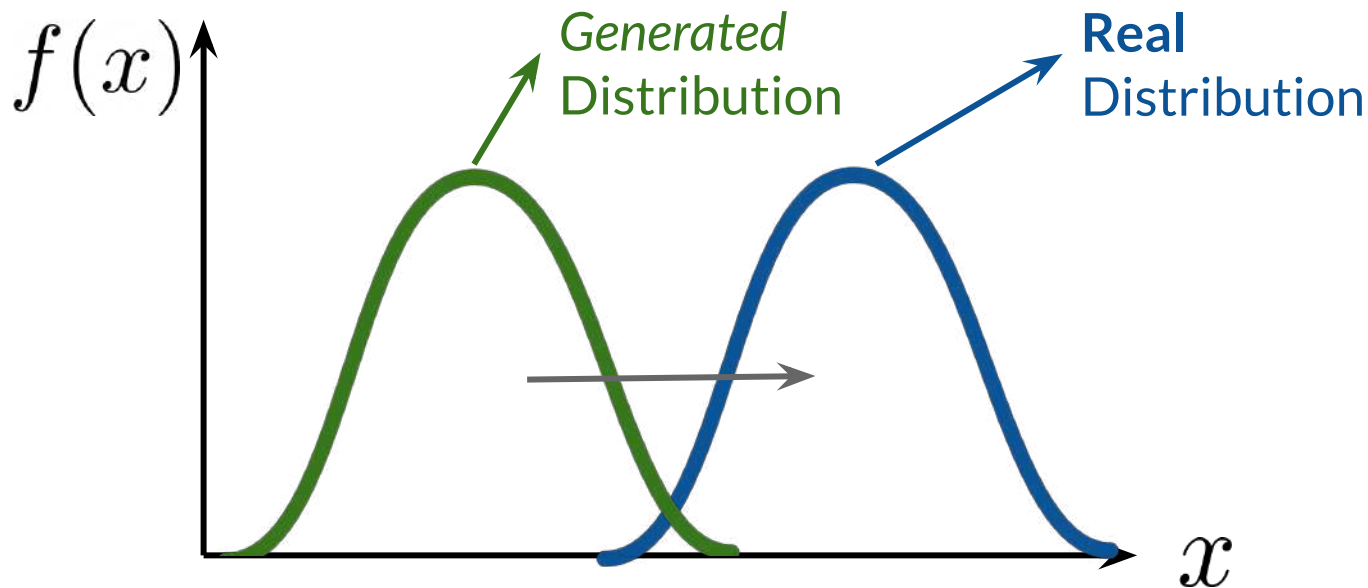
Minimize
cost

Objective in GANs



Objective in GANs

Make the generated and real distributions look similar



BCE Loss in GANs

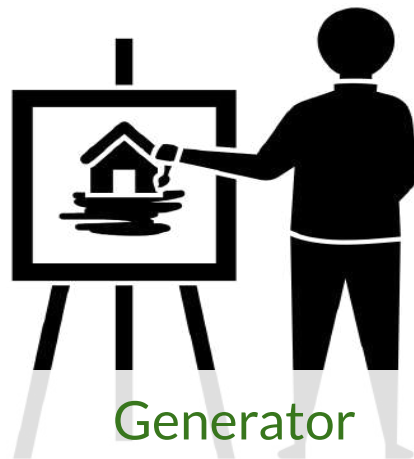
Criticizing is more straightforward



Discriminator

Single output

Easier to train
than the
generator



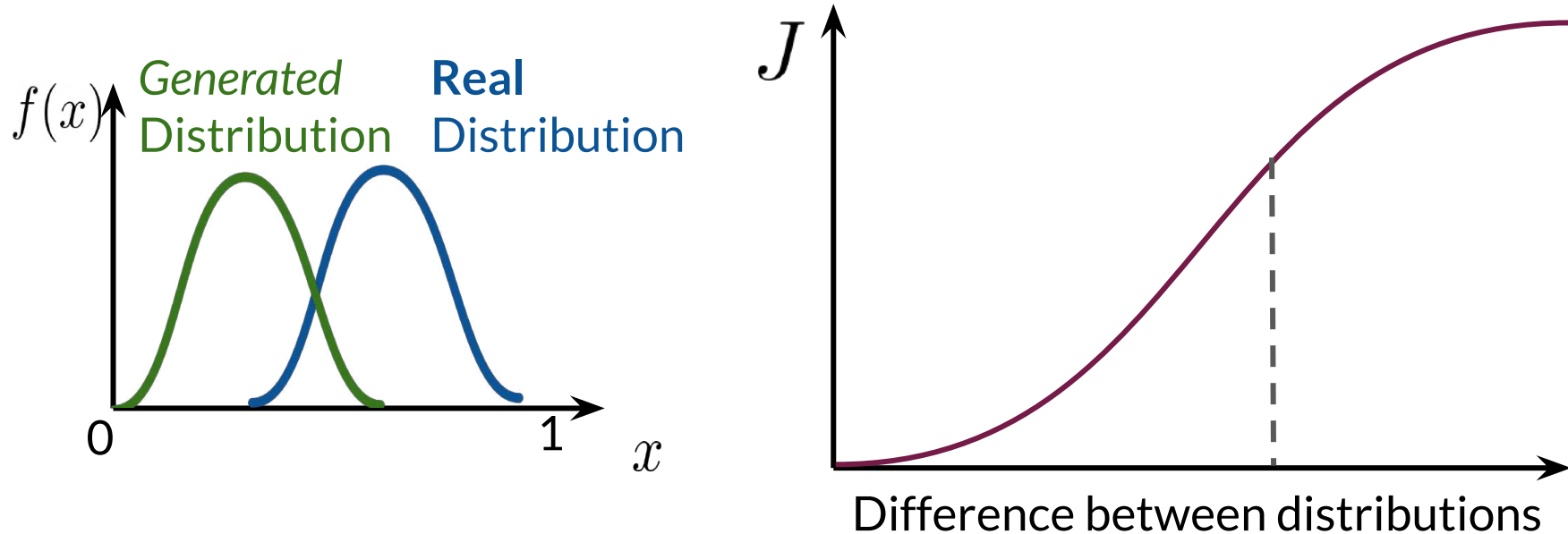
Generator

Complex
output

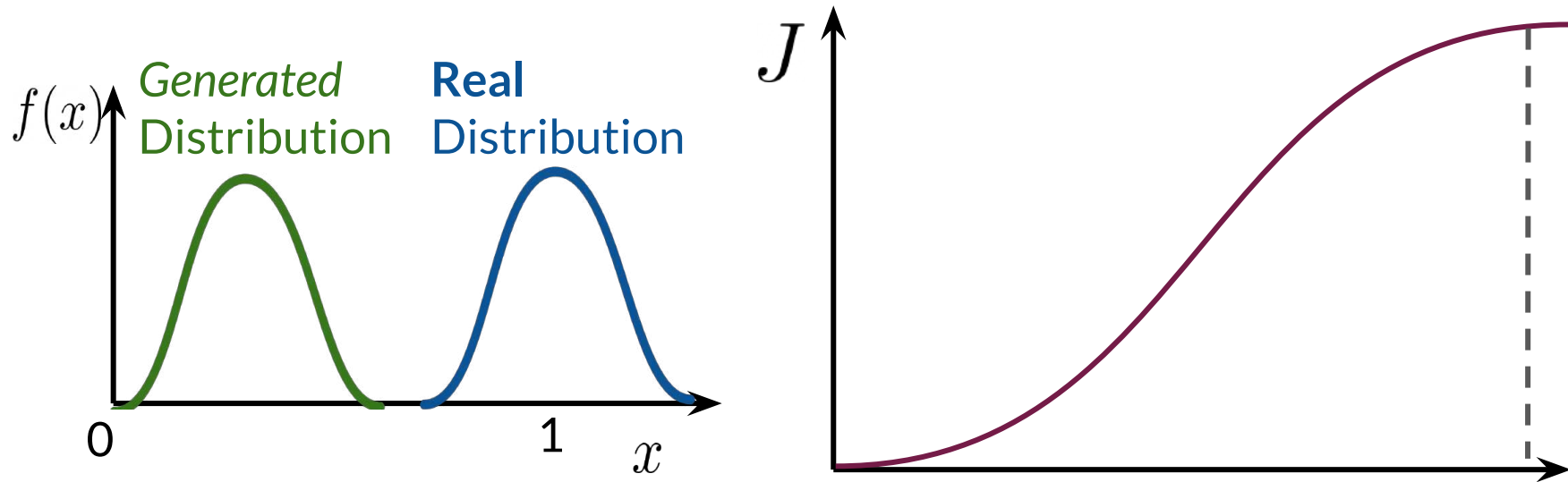
Difficult to
train

Often, the discriminator gets better than the generator

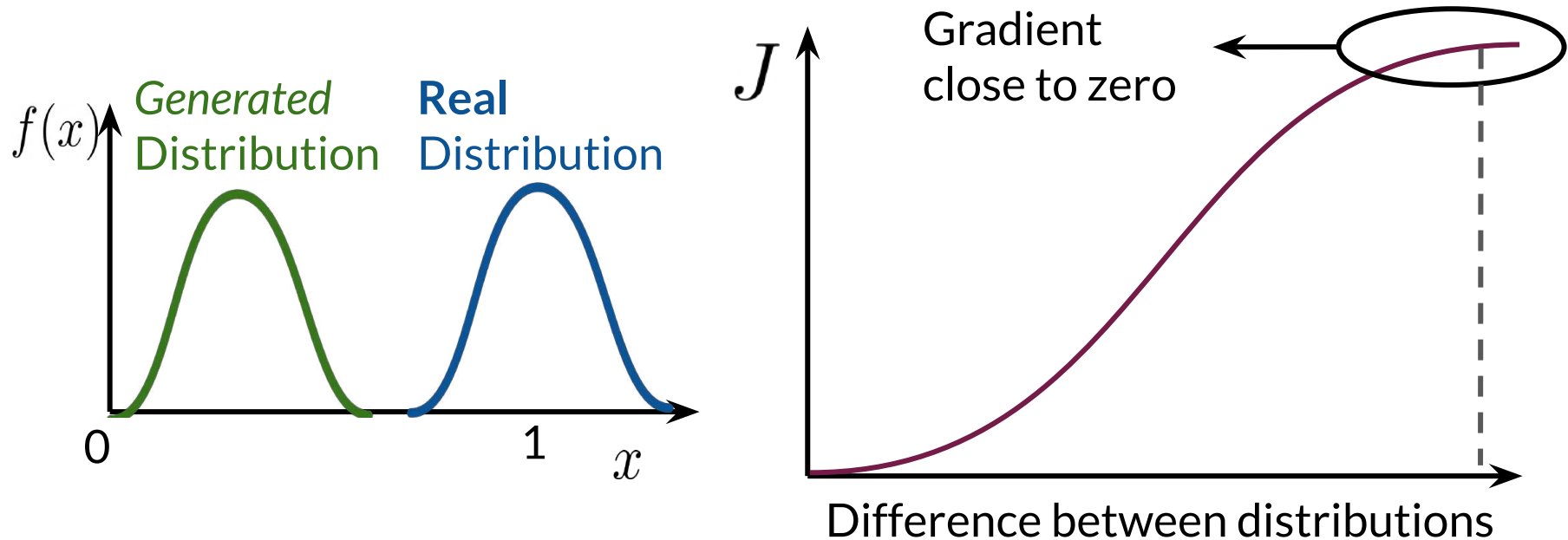
Problems with BCE Loss



Problems with BCE Loss

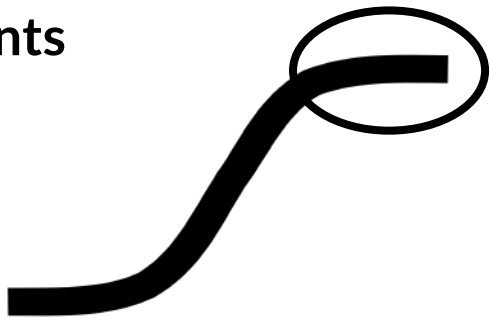


Problems with BCE Loss



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**



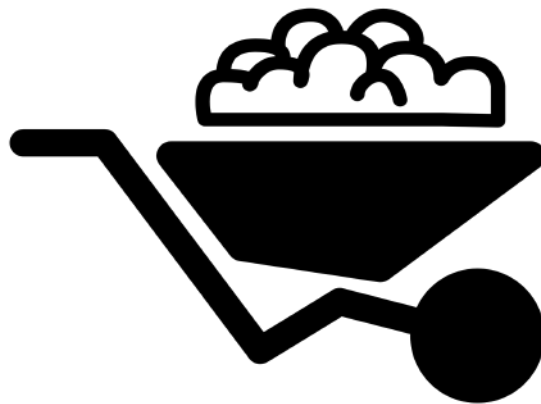


deeplearning.ai

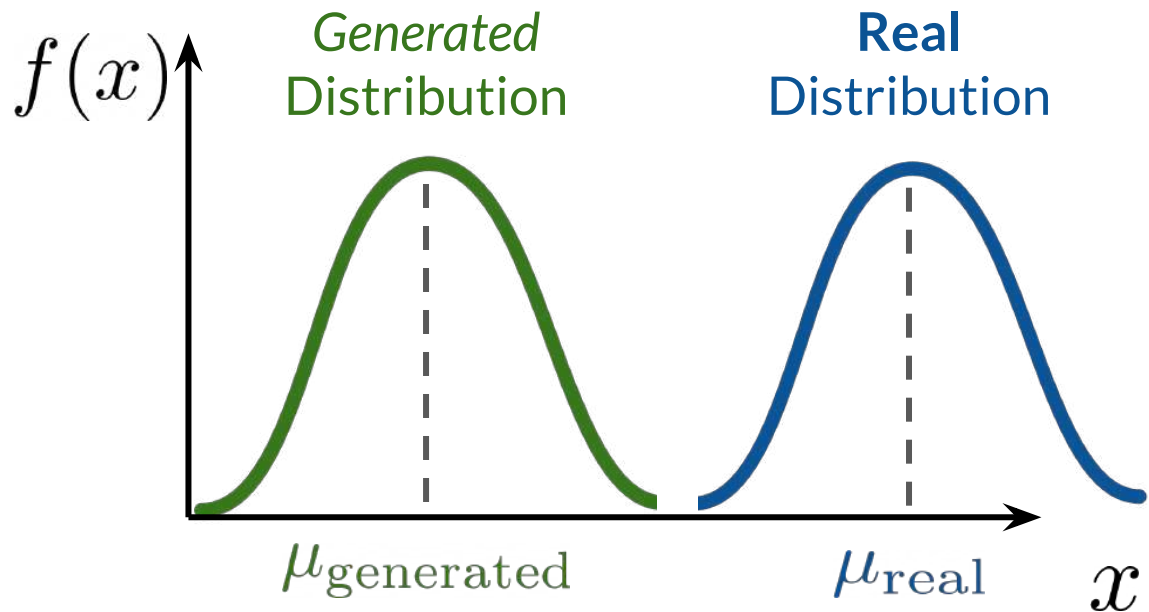
Earth Mover's Distance

Outline

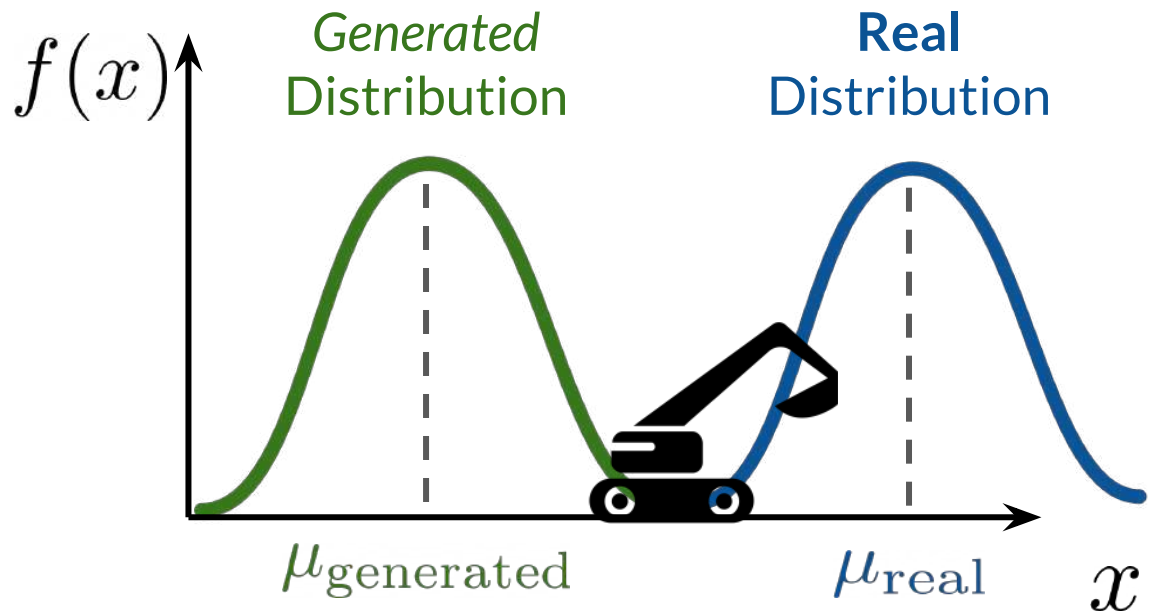
- Earth Mover's Distance (EMD)
- Why it solves the vanishing gradient problem of BCE Loss



Earth Mover's Distance



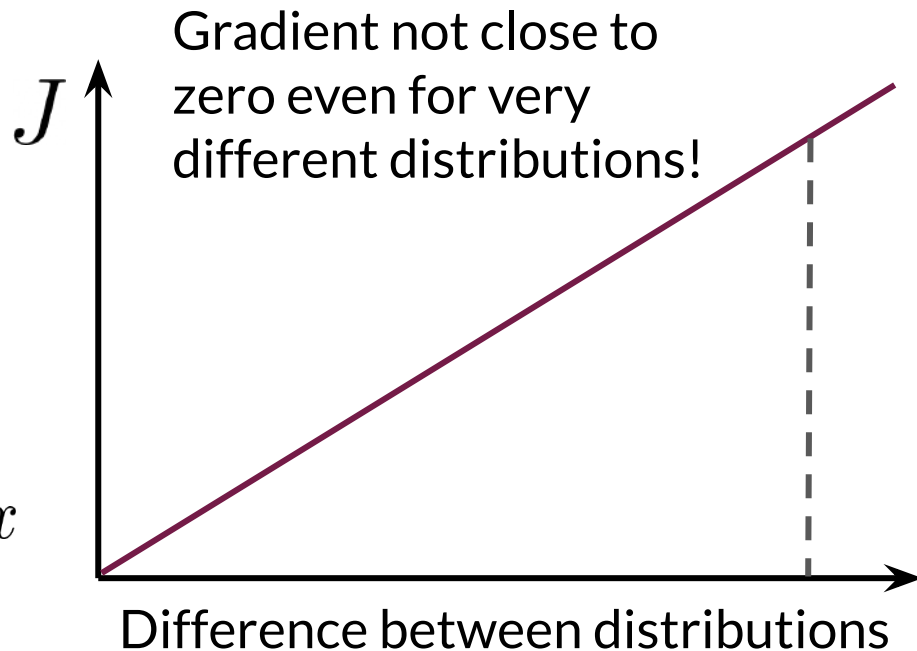
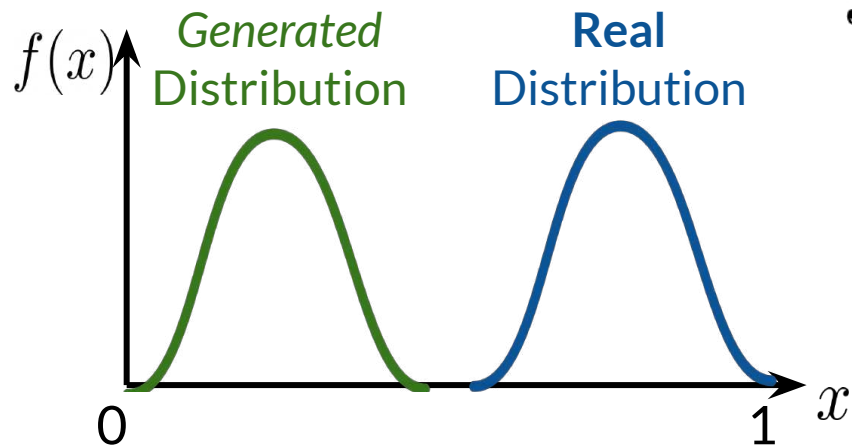
Earth Mover's Distance



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

Depends on the distance and amount moved

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

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

$\min_d \max_g$



Discriminator

Minimize
cost



Generator

Maximize
cost

BCE Loss Simplified

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

$$\min_d \max_g -[\mathbb{E}(\log(d(x))) + \mathbb{E}(\log(g(x)))]$$



Minimize
cost



Maximize
cost

BCE Loss Simplified

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

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



Minimize
cost



Maximize
cost

W-Loss

W-Loss approximates the Earth Mover's Distance

W-Loss

W-Loss approximates the Earth Mover's Distance

$$\min_g \max_c \mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))$$

W-Loss

W-Loss approximates the Earth Mover's Distance

$$\min_g \max_c \mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))$$



Maximize
the
distance

W-Loss

W-Loss approximates the Earth Mover's Distance

$$\min_g \max_c \mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))$$



Generator

Minimize
the
distance

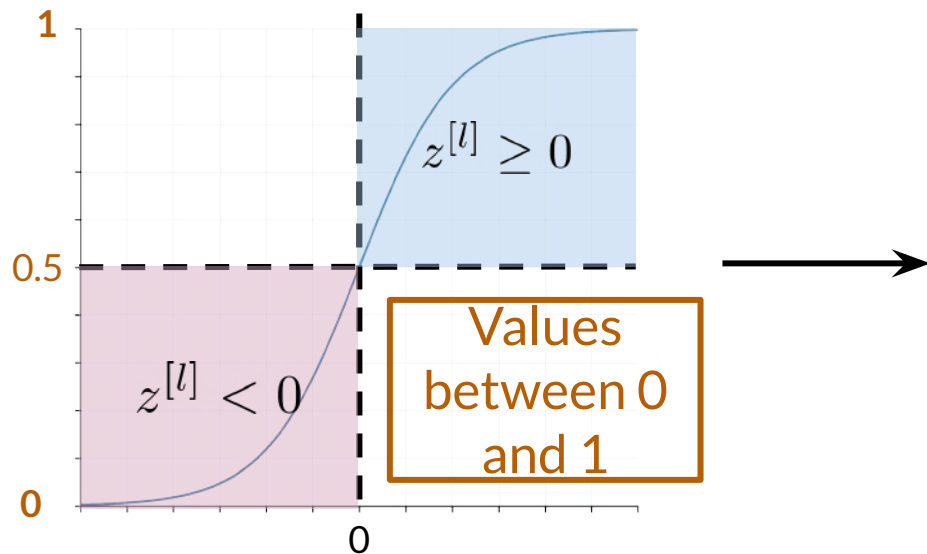


Critic

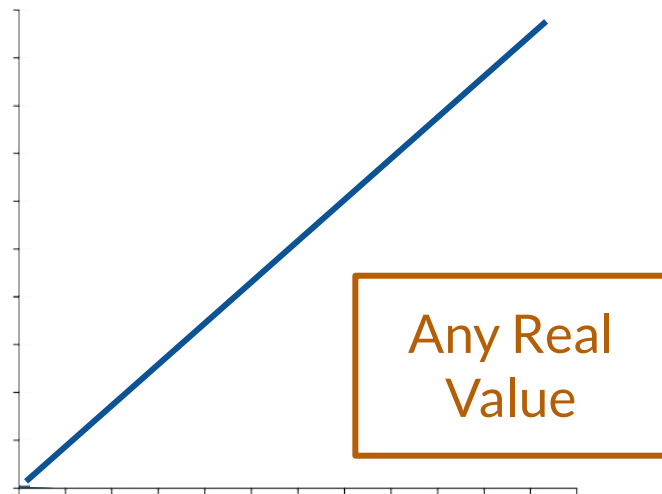
Maximize
the
distance

Discriminator Output

Discriminator output



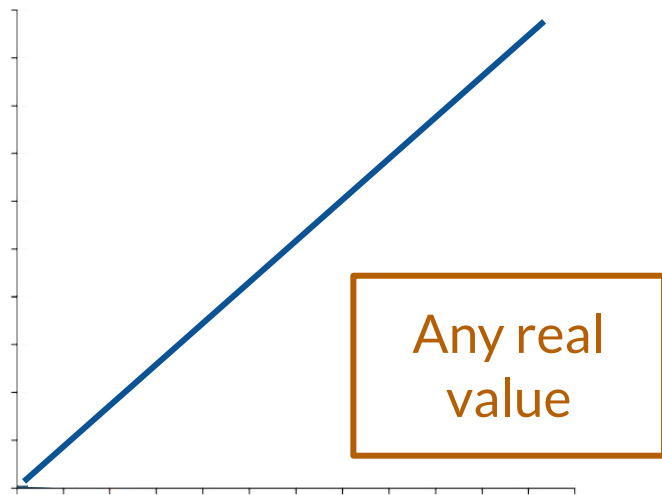
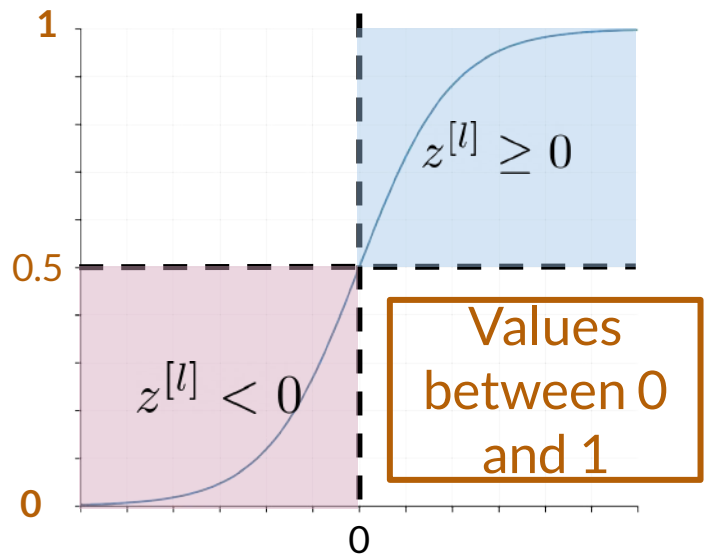
Discriminator output



Discriminator Output

Discriminator output

~~Discriminator output~~
Critic



W-Loss vs BCE Loss

BCE Loss

W-Loss

Discriminator outputs between 0 and 1

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

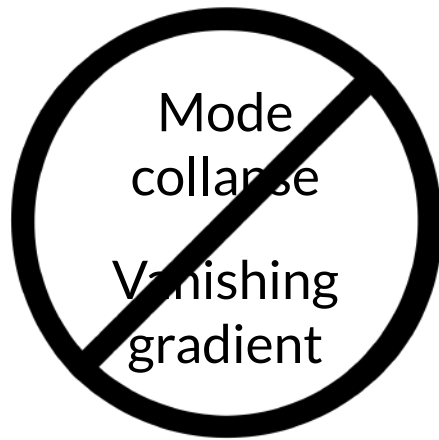
Critic outputs any number

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

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





deeplearning.ai

Condition on Wasserstein Critic

Outline

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



Condition on W-Loss

$$\min_g \max_c \mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))$$

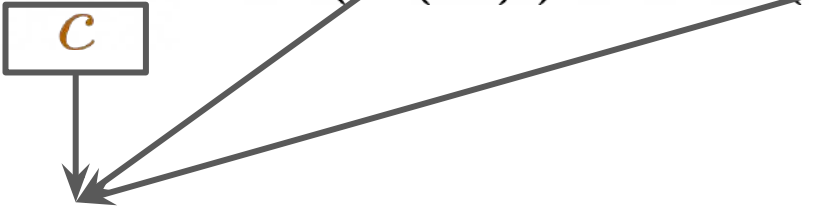
Condition on W-Loss

$$\min_g \max_c \mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))$$

Condition on W-Loss

$$\min_g \max_c \mathbb{E}(c(x)) - \mathbb{E}(c(g(z)))$$

Condition on W-Loss

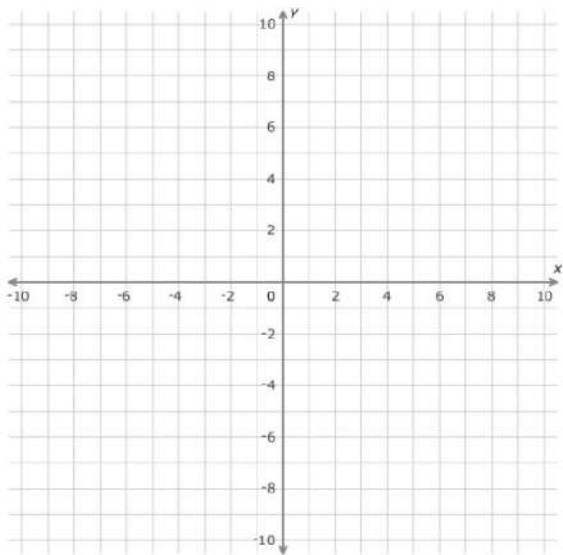
$$\min_g \max_c \mathbb{E}(\boxed{c}(x)) - \mathbb{E}(\boxed{c}(g(z)))$$
A diagram with two arrows pointing from the boxed 'c' in the equation to the text 'Needs to be 1-Lipschitz Continuous'. One arrow starts from the bottom of the box and points down. The other starts from the top-right corner of the box and points down and to the right.

Needs to be 1-Lipschitz Continuous

Condition on W-Loss

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

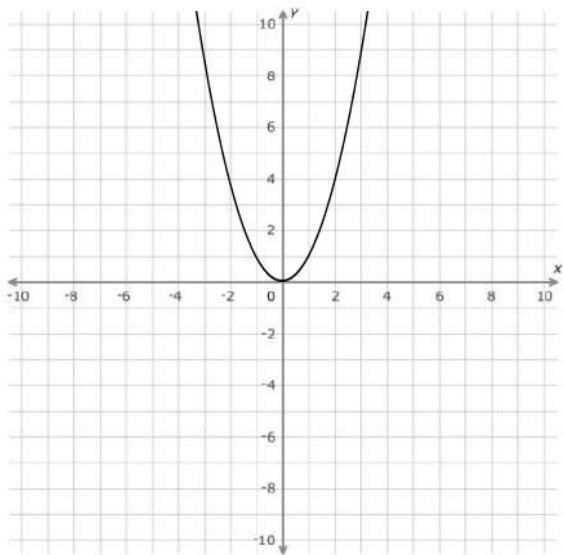
The norm of the gradient should be at most **1** for *every point*



Condition on W-Loss

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

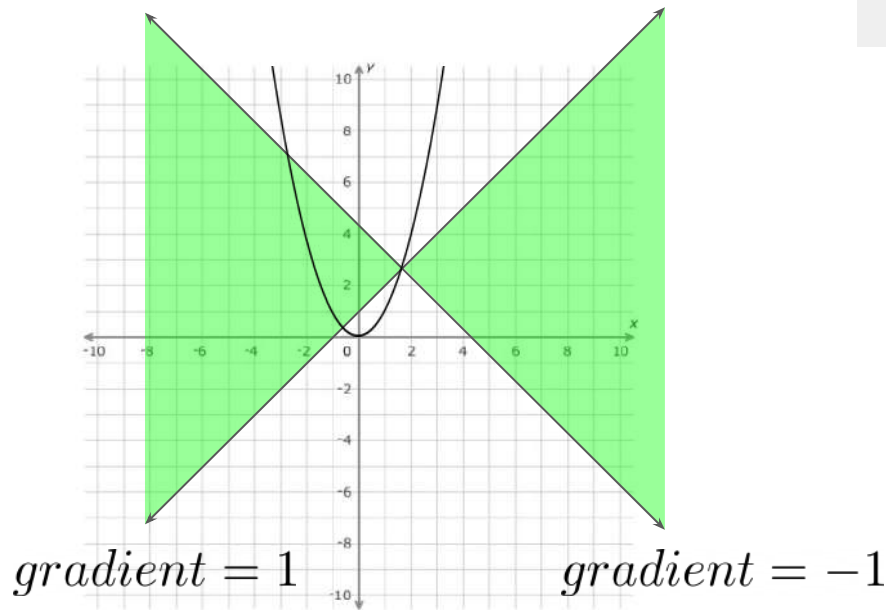
The norm of the gradient should be at most **1** for every point



Condition on W-Loss

Critic needs to be 1-L Continuous

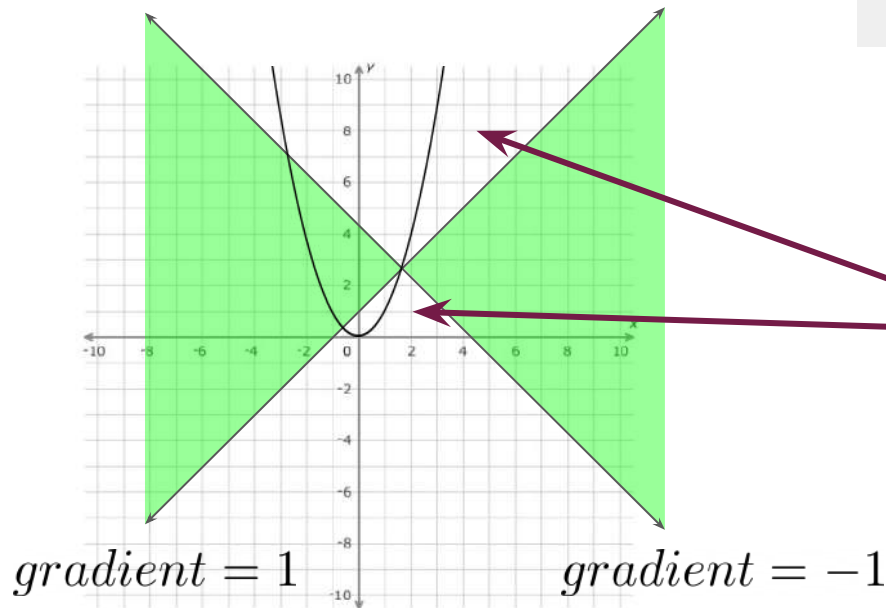
The norm of the gradient should be at most **1** for every point



Condition on W-Loss

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

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

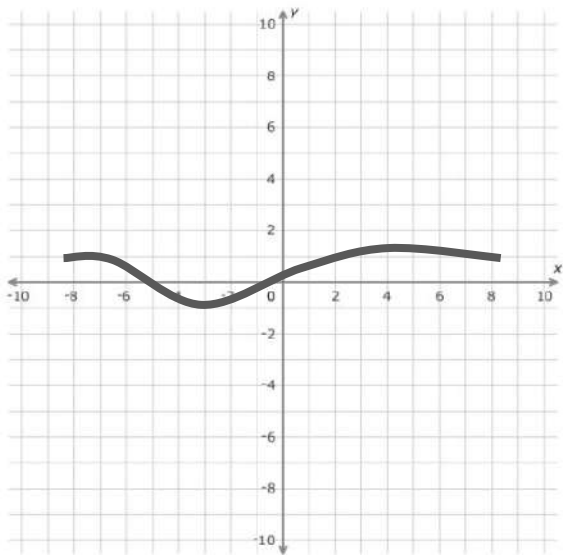


Not 1-L Continuous

Condition on W-Loss

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

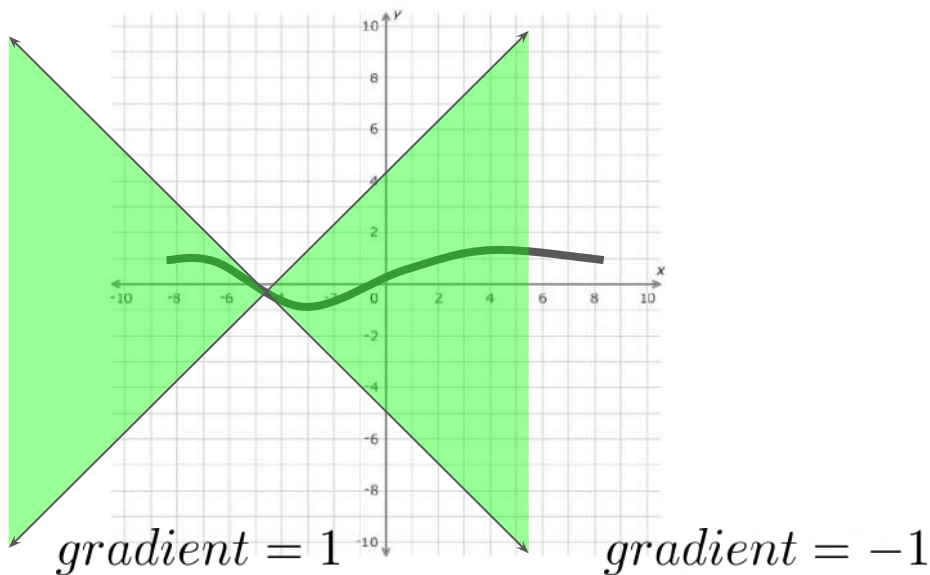
The norm of the gradient should be at most **1** for *every point*



Condition on W-Loss

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

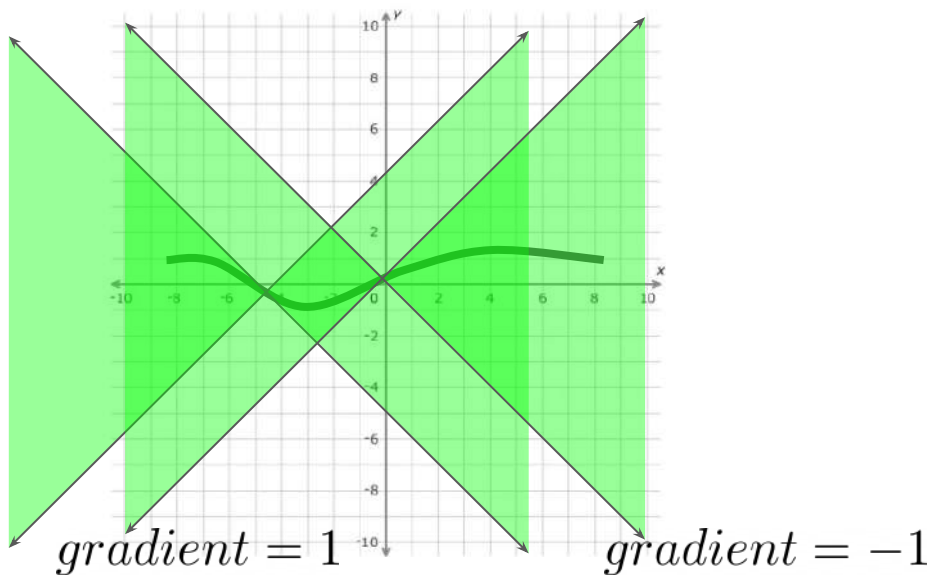
The norm of the gradient should be at most **1** for every point



Condition on W-Loss

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

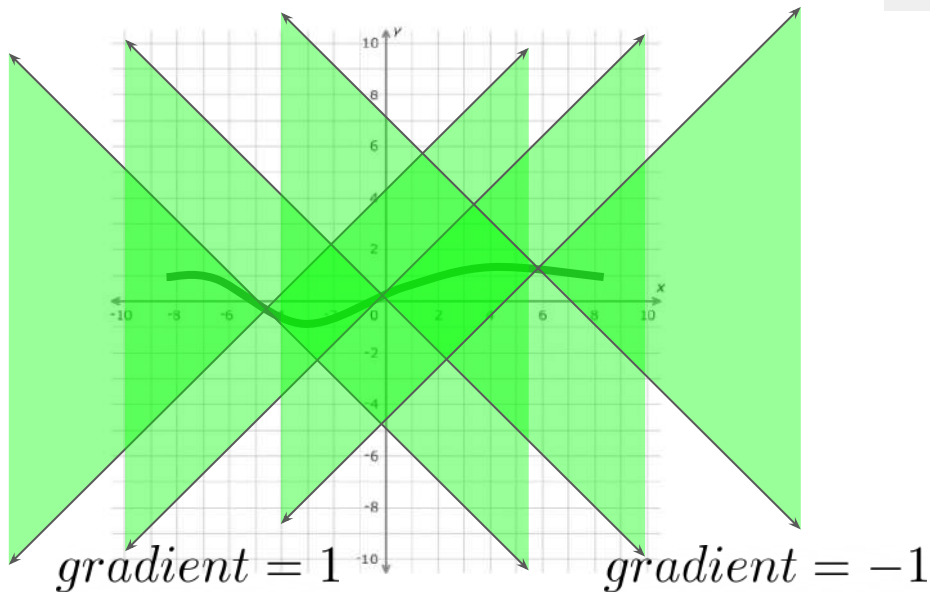
The norm of the gradient should be at most **1** for every point



Condition on W-Loss

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

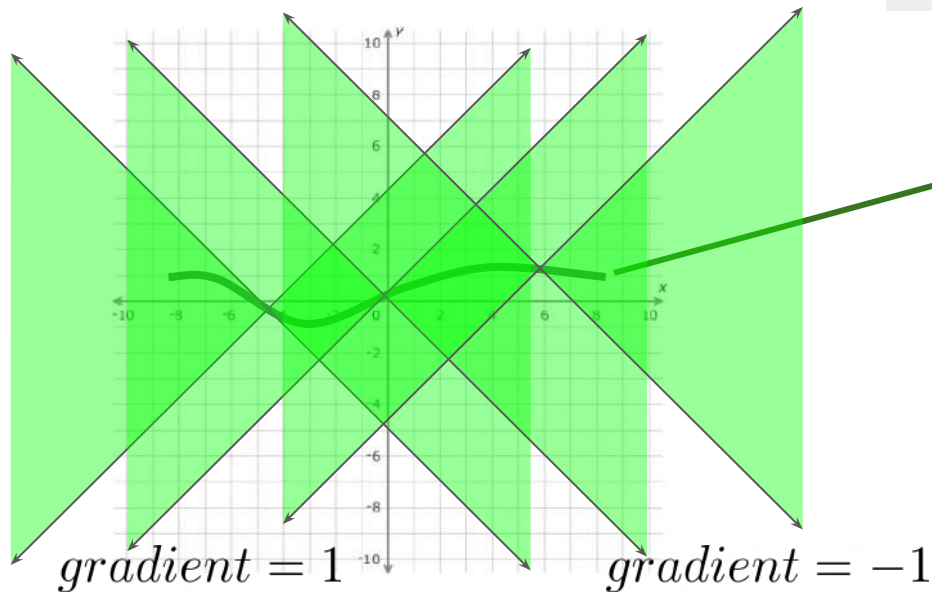
The norm of the gradient should be at most **1** for every point



Condition on W-Loss

Critic needs to be 1-L Continuous

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



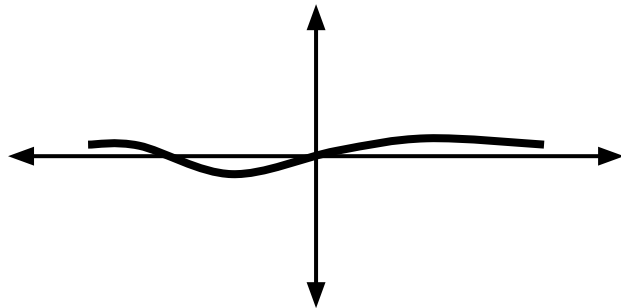
1-L Continuous

W-Loss is valid

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





deeplearning.ai

1-Lipschitz Continuity Enforcement

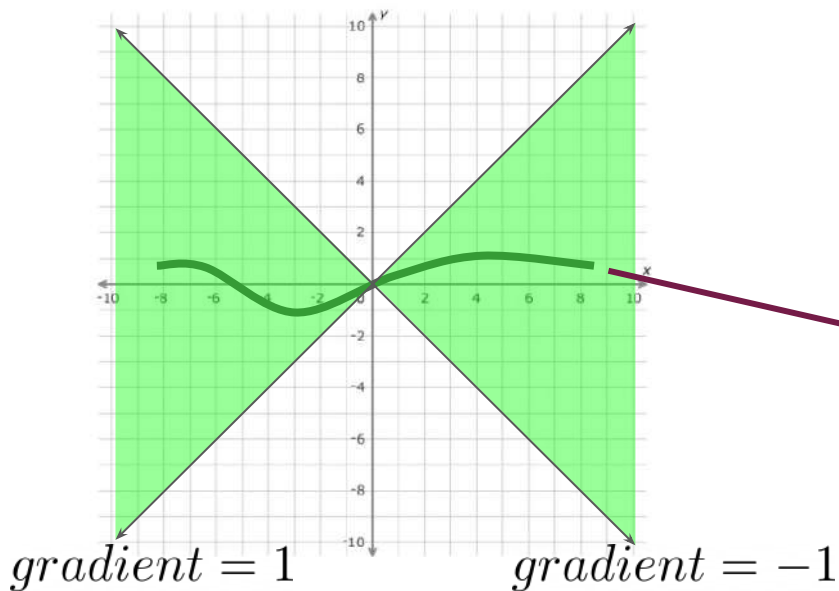
Outline

- Weight clipping and gradient penalty
- Advantages of gradient penalty



1-L Enforcement

Critic needs to be 1-L Continuous



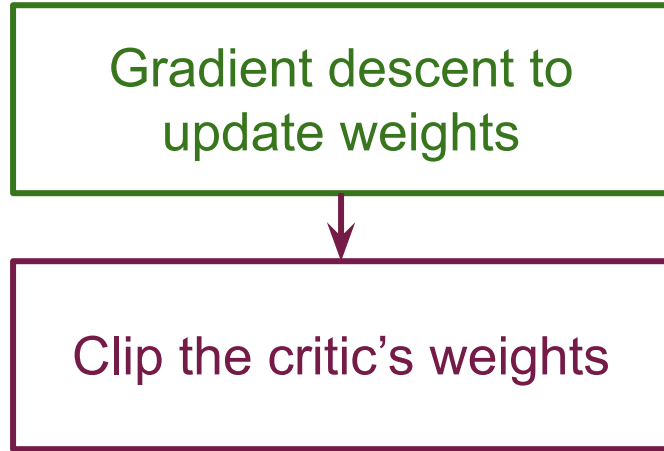
Norm of the gradient at most 1

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

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

$$\min_g \max_c \mathbb{E}(c(x)) - \mathbb{E}(c(g(z))) + \lambda \text{reg}$$



Regularization of the
critic's gradient

1-L Enforcement: Gradient Penalty

Real



ϵ

Random interpolation



1-L Enforcement: Gradient Penalty

Real



Random interpolation



ϵ

$1 - \epsilon$

\hat{x}

Generated



1-L Enforcement: Gradient Penalty

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

Regularization term

1-L Enforcement: Gradient Penalty

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

Regularization term

1-L Enforcement: Gradient Penalty

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

Regularization term




$$\epsilon x + (1 - \epsilon)g(z)$$

Interpolation

1-L Enforcement: Gradient Penalty

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

Regularization term


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

Real

Interpolation

1-L Enforcement: Gradient Penalty

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

Regularization term



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

Interpolation

Real

Generated

Putting It All Together

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

Putting It All Together

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

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

Putting It All Together

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

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

