# Deep Generative Models

## Lecture 10

Roman Isachenko

Moscow Institute of Physics and Technology

Autumn, 2021

# Recap of previous lecture

## Likelihood-free learning

- Likelihood is not a perfect measure quality measure for generative model.
- Likelihood could be intractable.

Imagine we have two sets of samples

- $\mathcal{S}_1 = \{\mathbf{x}_i\}_{i=1}^{n_1} \sim \pi(\mathbf{x})$ – real samples;
- $\mathcal{S}_2 = \{\mathbf{x}_i\}_{i=1}^{n_2} \sim p(\mathbf{x}|\boldsymbol{\theta})$ – generated (or fake) samples.

## Two sample test

$$H_0 : \pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}), \quad H_1 : \pi(\mathbf{x}) \neq p(\mathbf{x}|\boldsymbol{\theta})$$

If test statistic $T(\mathcal{S}_1, \mathcal{S}_2) < \alpha$, then accept $H_0$, else reject it.

- $p(\mathbf{x}|\boldsymbol{\theta})$ minimizes the value of test statistic $T(\mathcal{S}_1, \mathcal{S}_2)$.
- It is hard to find an appropriate test statistic in high dimensions. $T(\mathcal{S}_1, \mathcal{S}_2)$ could be learnable.

# Recap of previous lecture

- **Generator:** generative model $\mathbf{x} = G(\mathbf{z})$, which makes generated sample more realistic.
- **Discriminator:** a classifier $D(\mathbf{x}) \in [0, 1]$, which distinguishes real samples from generated samples.

## GAN optimality theorem

The minimax game

$$\min_G \max_D V(G, D) = \min_G \max_D \left[\mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(G(\mathbf{z})))\right]$$

has the global optimum $\pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})$, in this case $D^*(\mathbf{x}) = 0.5$.

$$\min_G V(G, D^*) = \min_G \left[2JSD(\pi||p) - \log 4\right] = -\log 4, \quad \pi(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta}).$$

If the generator could be any function and the discriminator is optimal at every step, then the generator is guaranteed to converge to the data distribution.

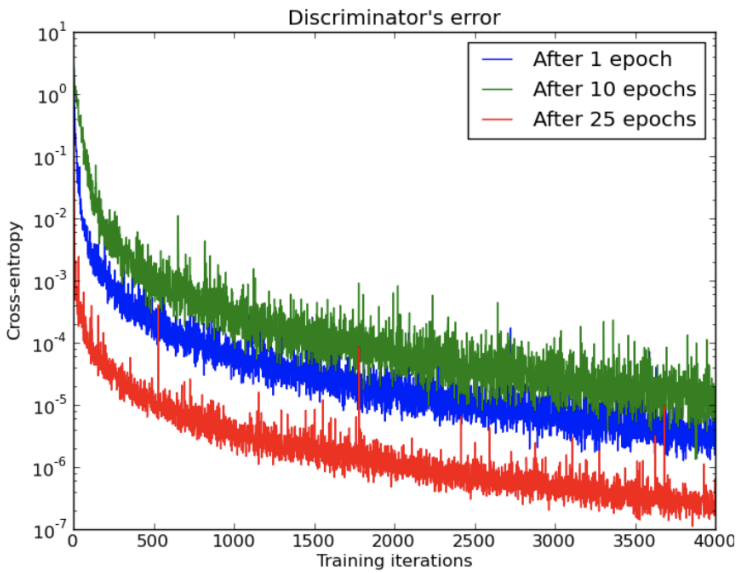
*Goodfellow I. J. et al. Generative Adversarial Networks, 2014*

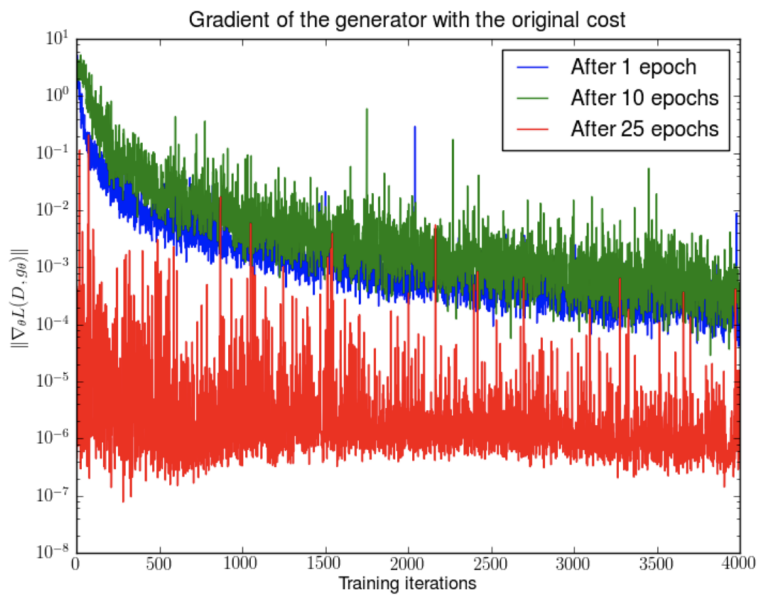
# Vanishing gradients

## Objective

$$\min_G \max_D V(G, D) = \min_G \max_D \left[ \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(G(\mathbf{z}))) \right]$$

Early in learning, $G$ is poor, $D$ can reject samples with high confidence. In this case, $\log(1 - D(G(\mathbf{z})))$ saturates.




*Arjovsky M., Bottou L. Towards Principled Methods for Training Generative Adversarial Networks, 2017*

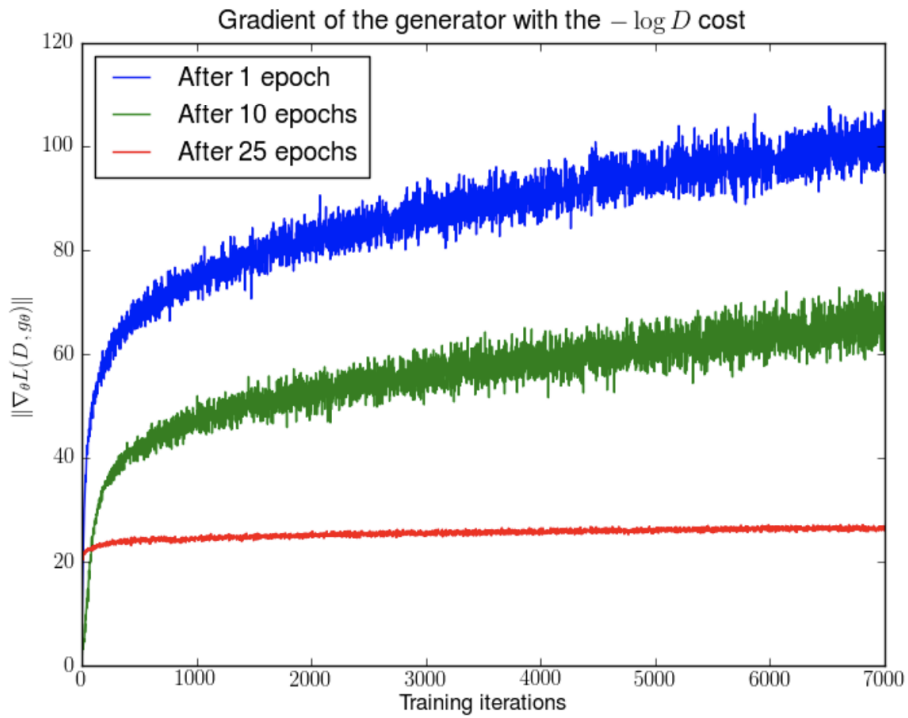
# Vanishing gradients

## Objective

$$\min_G \max_D V(G, D) = \min_G \max_D \left[ \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(G(\mathbf{z}))) \right]$$

## Non-saturating GAN

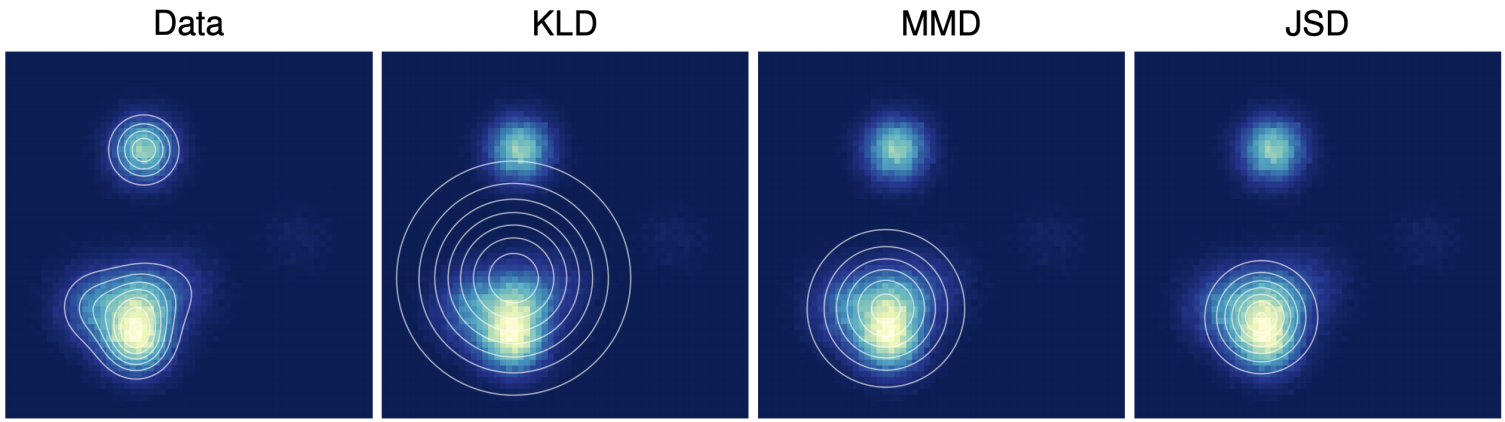
- Maximize $\log D(G(z))$ instead of minimizing $\log(1 - D(G(\mathbf{z})))$.
- Gradients are getting much stronger, but the training is unstable (with increasing mean and variance).




*Arjovsky M., Bottou L. Towards Principled Methods for Training Generative Adversarial Networks, 2017*

# Mode collapse

The phenomena where the generator of a GAN collapses to one or few distribution modes.




Alternate architectures, adding regularization terms, injecting small noise perturbations and other millions bags and tricks are used to avoid the mode collapse.

---

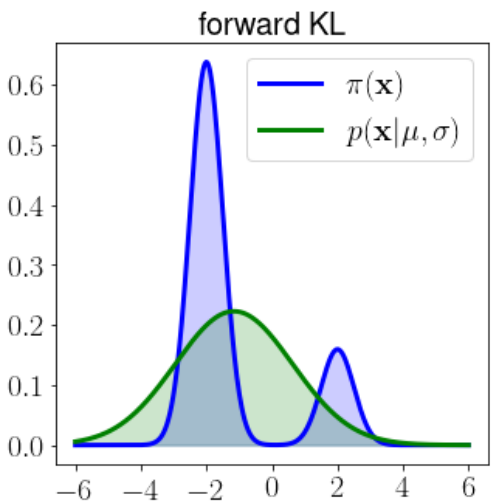
*Goodfellow I. J. et al. Generative Adversarial Networks, 2014*
*Metz L. et al. Unrolled Generative Adversarial Networks, 2016*

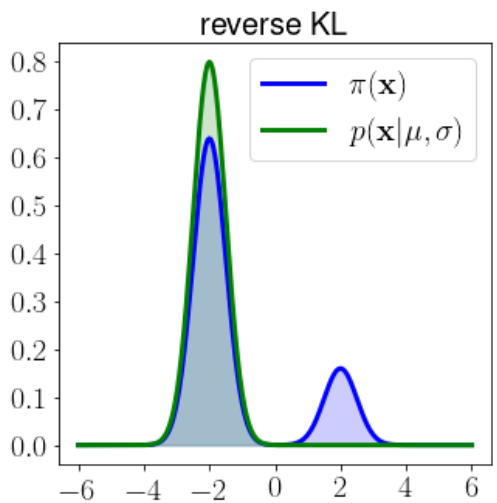
# Jensen-Shannon vs Kullback-Leibler

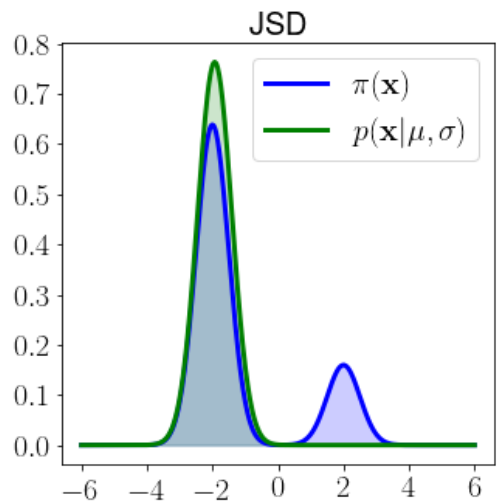
## Mode covering vs mode seeking

$$KL(\pi||p) = \int \pi(\mathbf{x}) \log \frac{\pi(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}, \quad KL(p||\pi) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\pi(\mathbf{x})} d\mathbf{x}$$

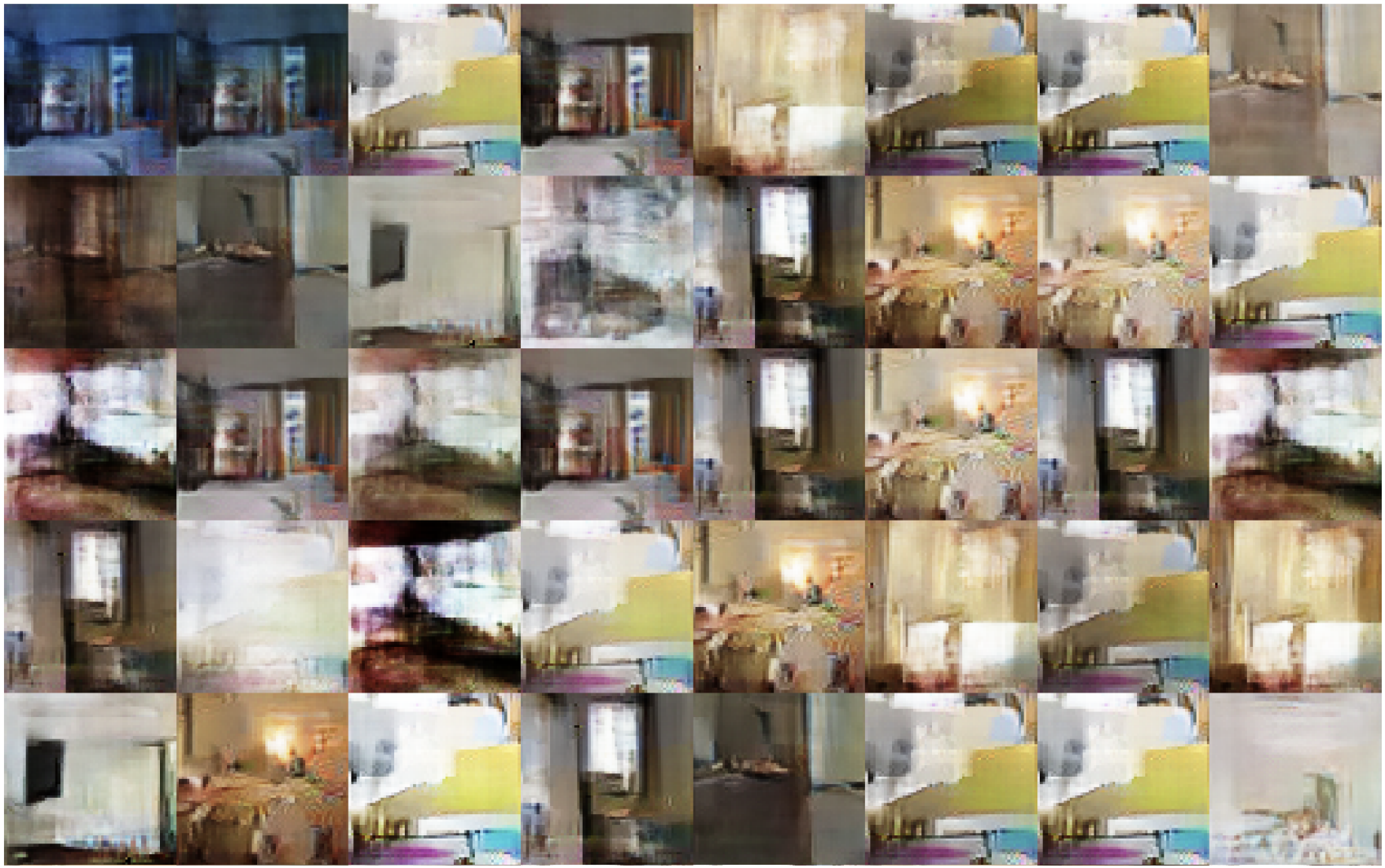
$$JSD(\pi||p) = \frac{1}{2}\left[KL\left(\pi(\mathbf{x})||\frac{\pi(\mathbf{x})+p(\mathbf{x})}{2}\right) + KL\left(p(\mathbf{x})||\frac{\pi(\mathbf{x})+p(\mathbf{x})}{2}\right)\right]$$



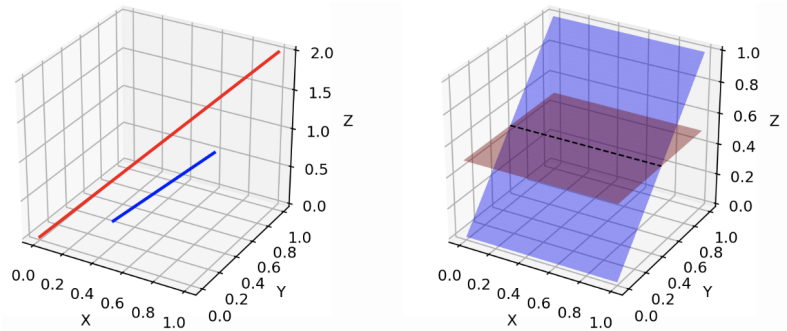
# Mode collapse: Deep Convolutional GAN



*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

# Informal theoretical results

- Since $\mathbf{z}$ usually has lower dimensionality compared to $\mathbf{x}$, manifold $G(\mathbf{z})$ has a measure 0 in $\mathbf{x}$ space. Hence, support of $p(\mathbf{x}|\boldsymbol{\theta})$ lies on low-dimensional manifold.
- Distribution of real images $\pi(\mathbf{x})$ is also concentrated on a low dimensional manifold.




- If $\pi(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\theta})$ have disjoint supports, then there is a smooth optimal discrimator. We are not able to learn anything by backproping through it.
- For such low-dimensional disjoint manifolds

$$KL(\pi||p) = KL(p||\pi) = \infty, \quad JSD(\pi||p) = \log 2$$

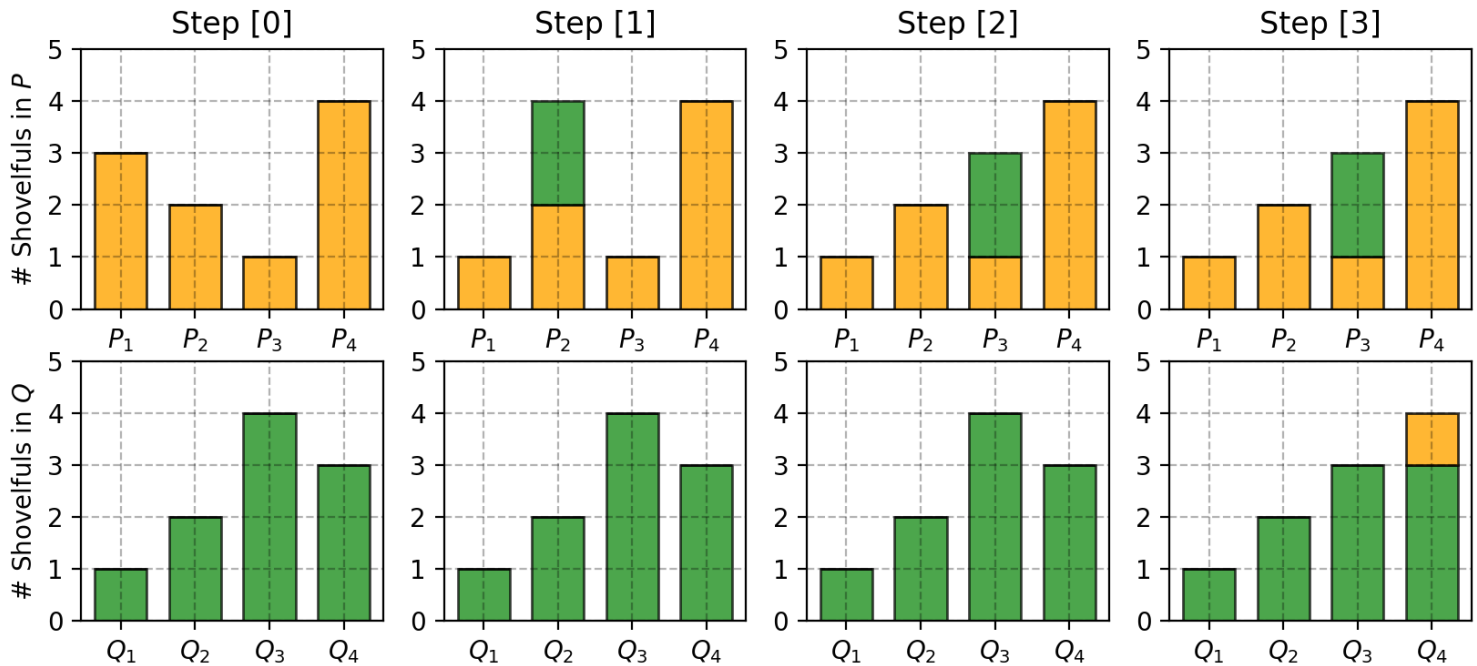
- Adding continuous noise to the inputs of the discriminator smoothes the distributions of the probability mass.

---

*Weng L. From GAN to WGAN, 2019*
*Arjovsky M., Bottou L. Towards Principled Methods for Training Generative Adversarial Networks, 2017*

# Wasserstein distance (discrete)

Also called Earth Mover's distance. The minimum cost of moving and transforming a pile of dirt in the shape of one probability distribution to the shape of the other distribution.




$$W(P, Q) = 2(\text{step } 1) + 2(\text{step } 2) + 1(\text{step } 3) = 5$$

*Weng L. From GAN to WGAN, 2019*

# Wasserstein distance (continuous)

$$W(\pi, p) = \inf_{\gamma \in \Gamma(\pi, p)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma} \|\mathbf{x}-\mathbf{y}\| = \inf_{\gamma \in \Gamma(\pi, p)} \int \|\mathbf{x}-\mathbf{y}\| \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$$

- $\gamma(\mathbf{x}, \mathbf{y})$ – transportation plan (the amount of "dirt" that should be transported from point $\mathbf{x}$ to point $\mathbf{y}$)

$$\int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} = p(\mathbf{y}); \quad \int \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \pi(\mathbf{x}).$$

- $\gamma(\mathbf{x}, \mathbf{y})$ – the amount, $\|\mathbf{x} - \mathbf{y}\|$– the distance.
- $\Gamma(\pi, p)$ – the set of all joint distributions $\Gamma(\mathbf{x}, \mathbf{y})$ with marginals $\pi$ and $p$.

For better understanding of transportation plan function $\gamma$, try to write down the plan for previous discrete case.

---

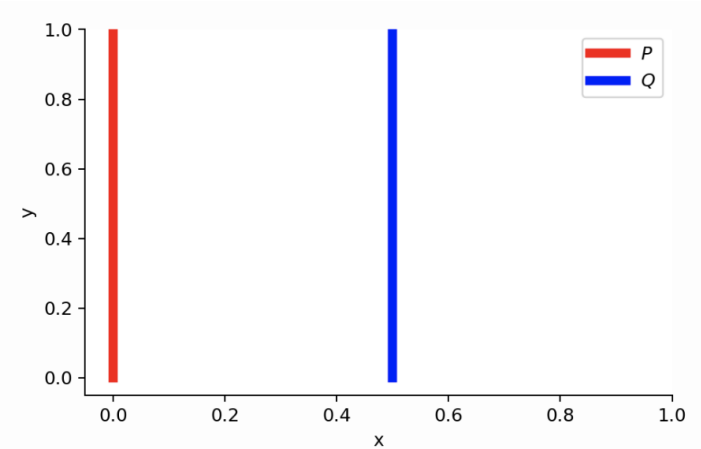
*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

# Wasserstein distance vs KL vs JSD

Consider 2d distributions

$$\pi(x, y) = (0, U[0, 1])$$

$$p(x, y|\theta) = (\theta, U[0, 1])$$




- $\theta = 0$. Distributions are the same

$$KL(\pi||p) = KL(p||\pi) = JSD(p||\pi) = W(\pi, p) = 0$$

- $\theta \neq 0$

$$KL(\pi||p) = \int_{U[0,1]} 1 \log \frac{1}{0} dy = \infty = KL(p||\pi)$$

$$JSD(\pi||p) = \frac{1}{2}\left(\int_{U[0,1]} 1 \log \frac{1}{1/2} dy + \int_{U[0,1]} 1 \log \frac{1}{1/2} dy\right) = \log 2$$

$$W(\pi, p) = |\theta|$$

---

*Weng L. From GAN to WGAN, 2019*
*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

# Wasserstein distance vs KL vs JSD

### Theorem 1

Let $G(\mathbf{z}, \boldsymbol{\theta})$ be (almost) any feedforward neural network, and $p(\mathbf{z})$ a prior over $\mathbf{z}$ such that $\mathbb{E}_{p(\mathbf{z})}\|\mathbf{z}\| < \infty$. Then therefore $W(\pi, p)$ is continuous everywhere and differentiable almost everywhere.

### Theorem 2

Let $\pi$ be a distribution on a compact space $\mathcal{X}$ and $\{p_t\}_{t=1}^{\infty}$ be a sequence of distributions on $\mathcal{X}$.

$$KL(\pi||p_t) \to 0 \,(\text{or } KL(p_t||\pi) \to 0) \tag{1}$$

$$JSD(\pi||p_t) \to 0 \tag{2}$$

$$W(\pi||p_t) \to 0 \tag{3}$$

Then, considering limits as $t \to \infty$, (1) implies (2), (2) implies (3).

---

*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

# Wasserstein GAN

## Wasserstein distance

$$W(\pi||p) = \inf_{\gamma \in \Gamma(\pi,p)} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\|\mathbf{x}-\mathbf{y}\| = \inf_{\gamma \in \Gamma(\pi,p)} \int \|\mathbf{x}-\mathbf{y}\|\gamma(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y}$$

The infimum across all possible joint distributions in $\Gamma(\pi, p)$ is intractable.

## Theorem (Kantorovich-Rubinstein duality)

$$W(\pi||p) = \frac{1}{K} \max_{\|f\|_L \leq K} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})\right],$$

where $\|f\|_L \leq K$ are $K-$Lipschitz continuous functions $(f : \mathcal{X} \to \mathbb{R})$

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq K\|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \text{for all } \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}.$$

---

*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

# Wasserstein GAN

## Theorem (Kantorovich-Rubinstein duality)

$$W(\pi||p) = \frac{1}{K} \max_{\|f\|_L \leq K} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})\right],$$

- Now we have to ensure that $f$ is $K$-Lipschitz continuous.
- Let $f(\mathbf{x}, \phi)$ be a feedforward neural network parametrized by $\phi$.
- If parameters $\phi$ lie in a compact set $\mathbf{\Phi}$ then $f(\mathbf{x}, \phi)$ will be $K$-Lipschitz continuous function.
- Let the parameters be clamped to a fixed box $\mathbf{\Phi} \in [-0.01, 0.01]^d$ after each gradient update.

$$K \cdot W(\pi||p) = \max_{\|f\|_L \leq K} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})\right] \geq$$

$$\geq \max_{\phi \in \mathbf{\Phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}, \phi) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x}, \phi)\right]$$

---

*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

# Wassestein GAN

## Vanilla GAN objective

$$\min_G \max_D \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(G(\mathbf{z})))$$

## WGAN objective

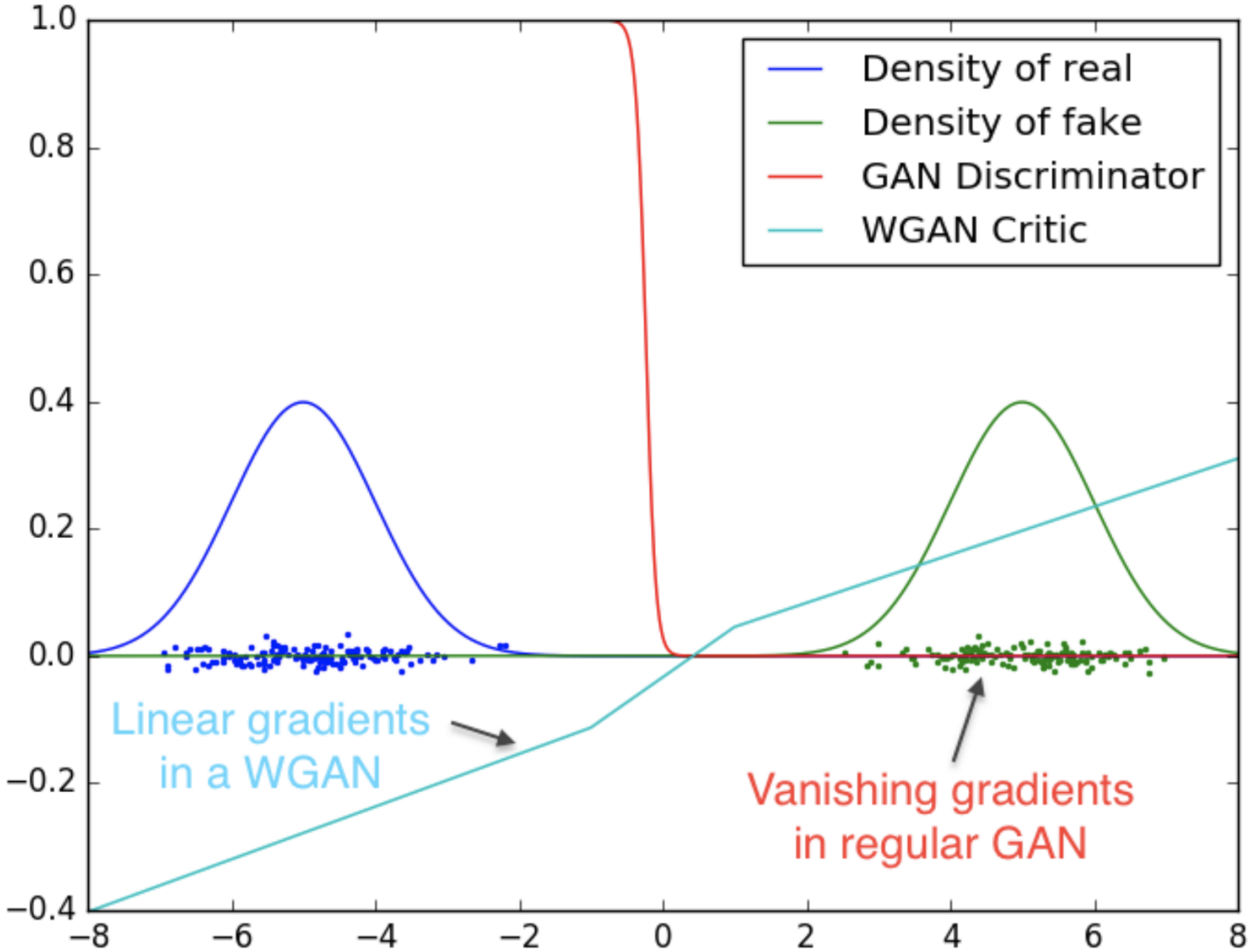
$$\min_G W(\pi||p) = \min_G \max_{\phi \in \mathbf{\Phi}} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}, \phi) - \mathbb{E}_{p(\mathbf{z})} f(G(\mathbf{z}), \phi)\right].$$

- Discriminator $D$ is similar to the function $f$, but not the same (it is not a classifier anymore). In the WGAN model, function $f$ is usually called *critic*.
- "Weight clipping is a clearly terrible way to enforce a Lipschitz constraint". If the clipping parameter is large, it is hard to train the critic till optimality. If the clipping parameter is too small, it could lead to vanishing gradients.

---

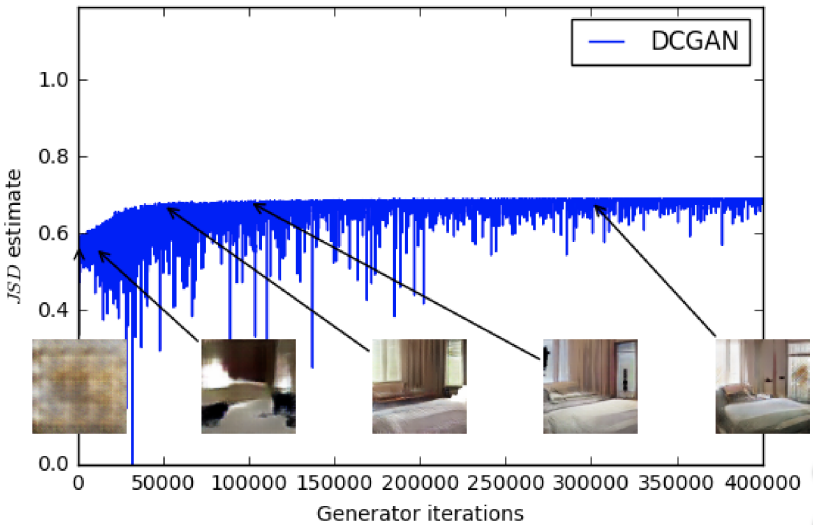
*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

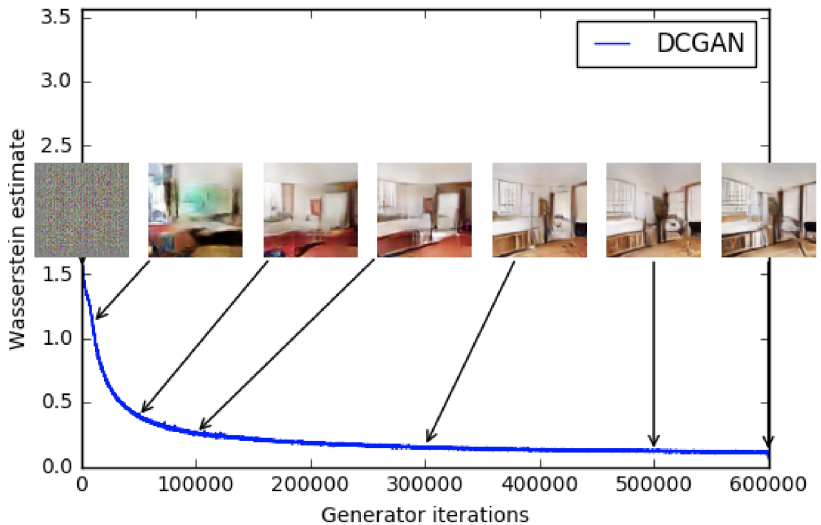
# Wasserstein GAN



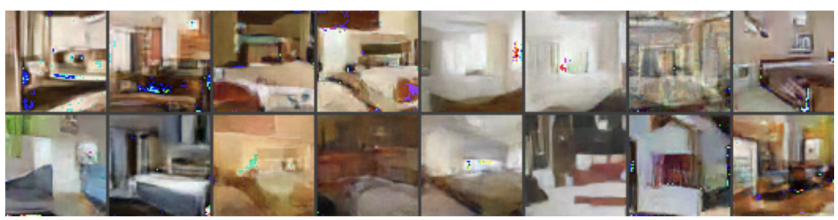

*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

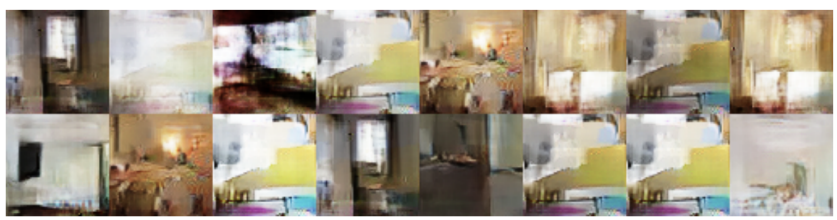
# Wasserstein GAN




- *JSD* correlates poorly with the sample quality. Stays constast nearly maximum value $\log 2 \approx 0.69$.
- *W* is highly correlated with the sample quality.





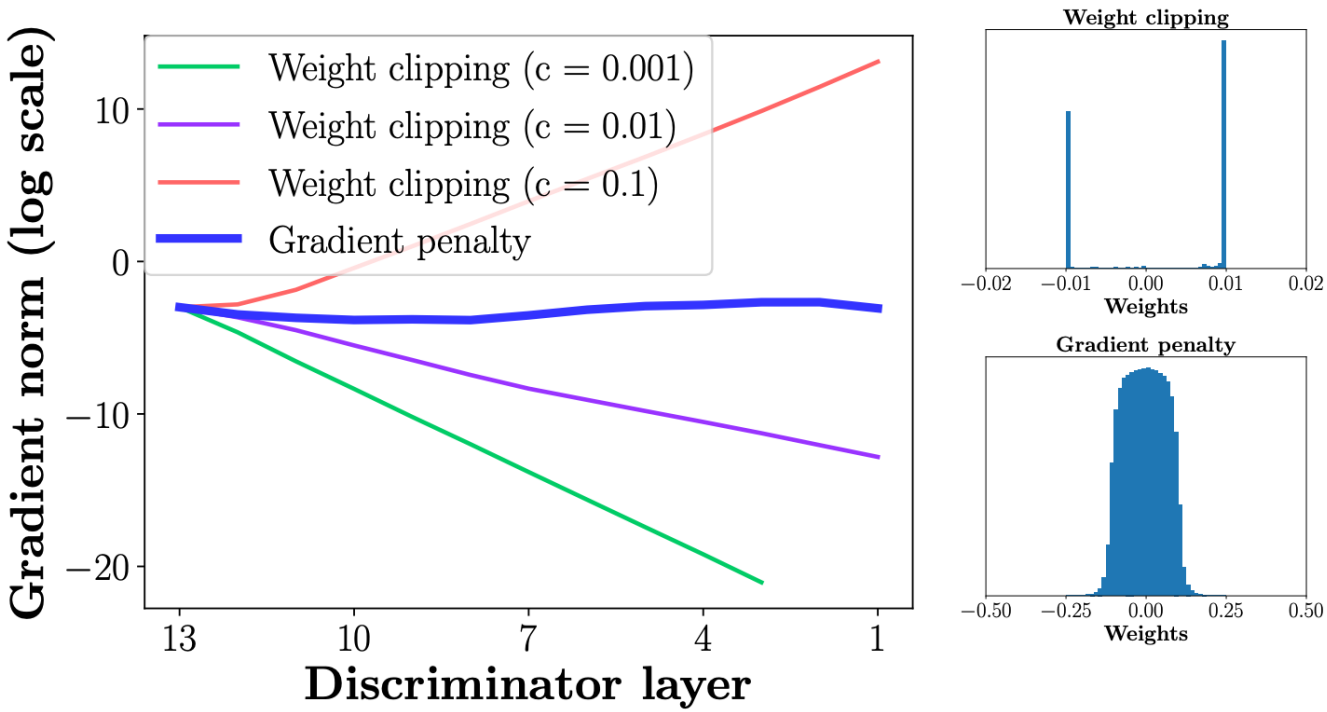
"In no experiment did we see evidence of mode collapse for the WGAN algorithm."

*Arjovsky M., Chintala S., Bottou L. Wasserstein GAN, 2017*

# Wasserstein GAN with Gradient Penalty

## Weight clipping analysis

- The critic ignores higher moments of the data distribution.
- The gradients either grow or decay exponentially.




Gradient penalty makes the gradients more stable.

---

*Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017*

# Wasserstein GAN with Gradient Penalty

## Theorem

Let $\pi(\mathbf{x})$ and $p(\mathbf{x})$ be two distribution in $\mathcal{X}$, a compact metric space. Then, there is 1-Lipschitz function $f^*$ which is the optimal solution of

$$\max_{\|f\|_L \leq 1} \left[\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})\right].$$

Let $\gamma$ be the optimal transportation plan between $\pi(\mathbf{x})$ and $p(\mathbf{x})$. Then, if $f^*$ is differentiable, $\gamma(\mathbf{x} = \mathbf{y}) = 0$ and $\hat{\mathbf{x}}_t = t\mathbf{x} + (1-t)\mathbf{y}$ with $\mathbf{x} \sim \pi(\mathbf{x})$, $\mathbf{y} \sim p(\mathbf{x}|\boldsymbol{\theta})$, $t \in [0, 1]$ it holds that

$$\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[\nabla f^*(\hat{\mathbf{x}}_t) = \frac{\mathbf{y} - \hat{\mathbf{x}}_t}{\|\mathbf{y} - \hat{\mathbf{x}}_t\|}\right] = 1.$$

## Corollary

$f^*$ has gradient norm 1 almost everywhere under $\pi(\mathbf{x})$ and $p(\mathbf{x})$.

---

*Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017*

# Wasserstein GAN with Gradient Penalty

A differentiable function is 1-Lipschtiz if and only if it has gradients with norm at most 1 everywhere.

## Gradient penalty

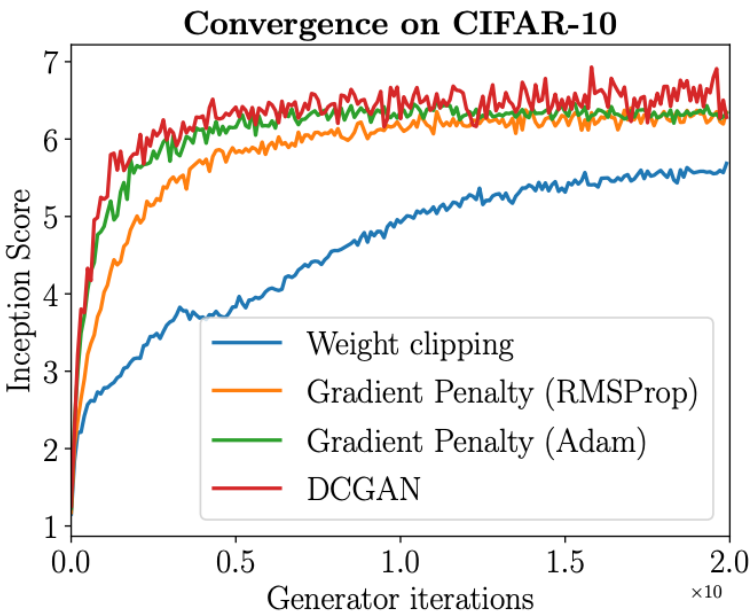
$$W(\pi||p) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})} f(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} f(\mathbf{x})}_{\text{original critic loss}} + \lambda \underbrace{\mathbb{E}_{U[0,1]} \left[ (\|\nabla_{\hat{\mathbf{x}}} f(\hat{\mathbf{x}})\|_2 - 1)^2 \right]}_{\text{gradient penalty}},$$

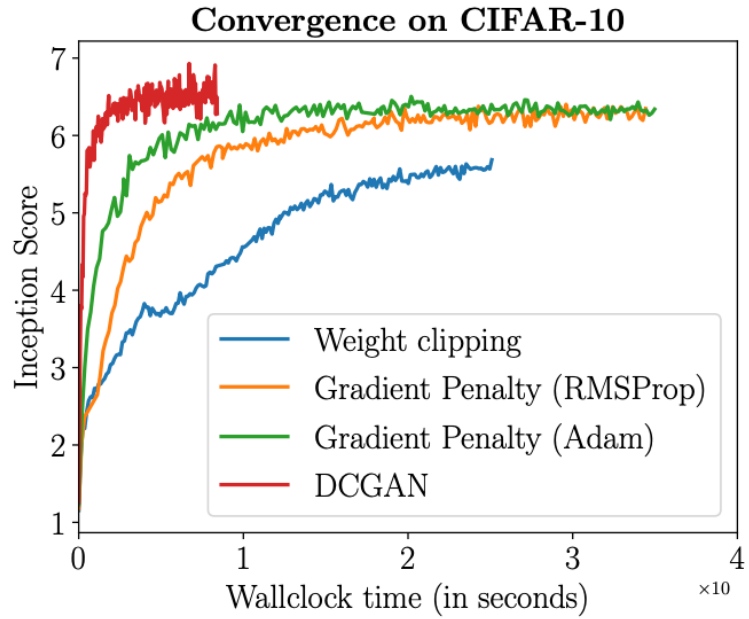
- Samples $\hat{\mathbf{x}}_t = t\mathbf{x} + (1-t)\mathbf{y}$ with $t \in [0, 1]$ are uniformly sampled along straight lines between pairs of points: $\mathbf{x}$ from the data distribution $\pi(\mathbf{x})$ and $\mathbf{y}$ from the generator distribution $p(\mathbf{x}|\boldsymbol{\theta})$.
- Enforcing the unit gradient norm constraint everywhere is intractable, it turns out to be sifficient to enforce it only along these straight lines.

---

*Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017*

# Wasserstein GAN with Gradient Penalty




## WGAN-GP convergence

| Min. score | Only GAN | Only WGAN-GP | Both succeeded | Both failed |
|---|---|---|---|---|
| 1.0 | 0 | 8 | 192 | 0 |
| 3.0 | 1 | 88 | 110 | 1 |
| 5.0 | 0 | 147 | 42 | 11 |
| 7.0 | 1 | 104 | 5 | 90 |
| 9.0 | 0 | 0 | 0 | 200 |

*Gulrajani I. et al. Improved Training of Wasserstein GANs, 2017*

# Summary

- Mode collapse and vanishing gradients are the two main problems of vanilla GAN. Lots of tips and tricks has to be used to make the GAN training is stable and scalable.
- DCGAN is the first GAN with deep convolutional architecture.
- KL and JS divergences work poorly as model objective in the case of disjoint supports.
- Earth-Mover distance is a more appropriate objective function for distribution matching problem.
- Wasserstein GAN uses Kantorovich-Rubinstein duality to obtain EM distance.
- Weight clipping is a terrible way to enforce Lipschitzness. Gradient Penalty works better.