PyTorch

深度学习：GAN

主讲人：龙良曲

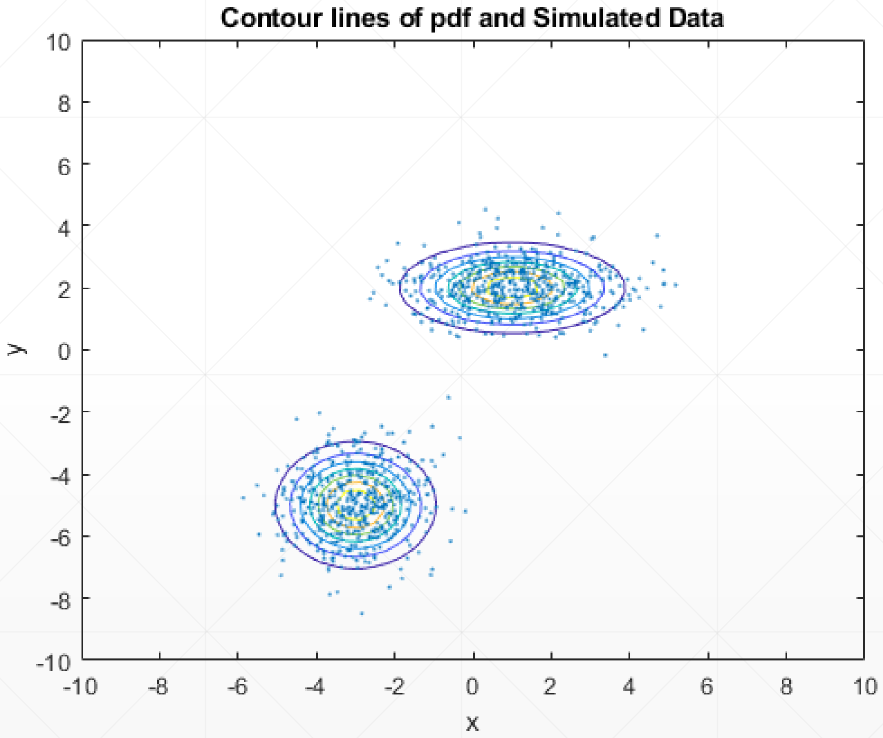
"What I cannot create, I do not understand."

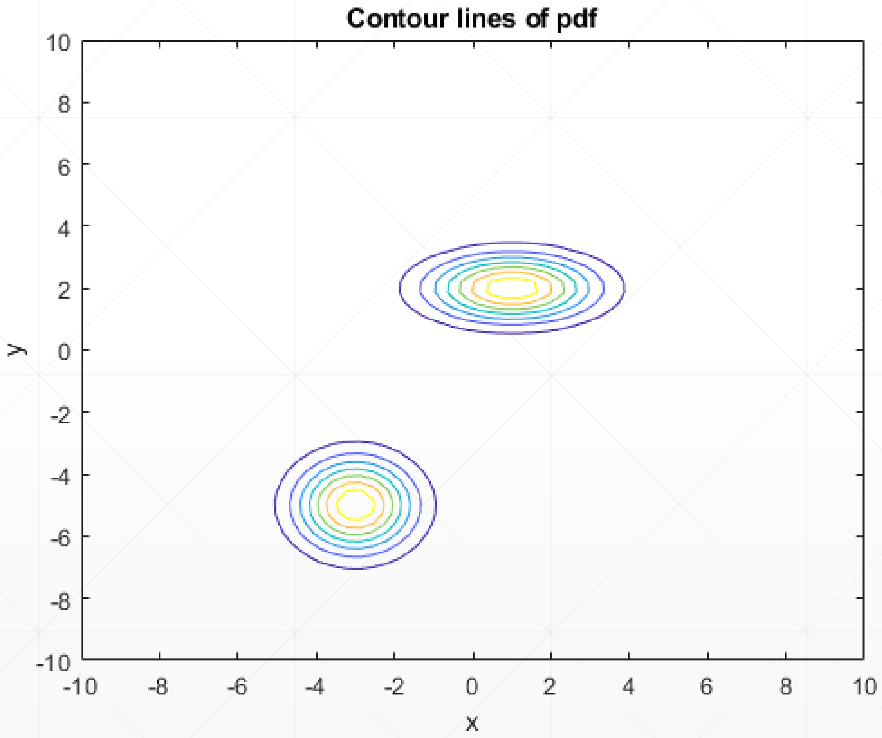
—Richard Feynman

https://blog.openai.com/generative-models/

Our Goal: $p(x)$

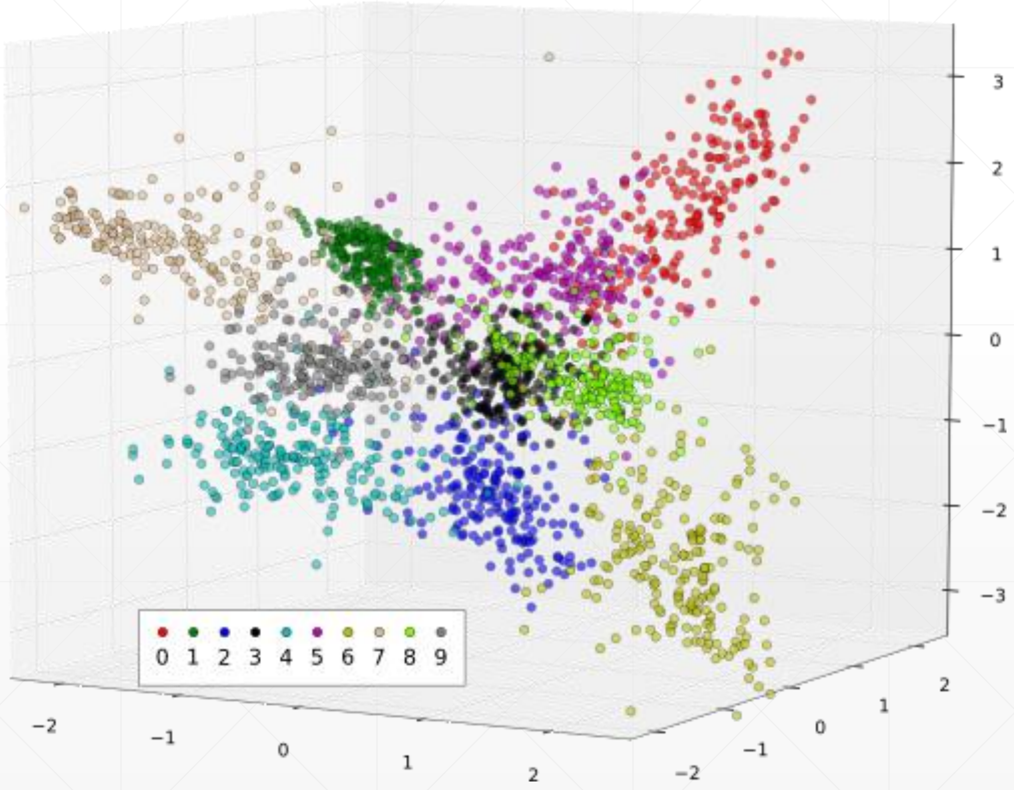
```
mu = [1 2;-3 -5]
sigma = cat(3,[2 0;0 .5],[1 0;0 1])
p = ones(1,2)/2
gm = gmdistribution(mu,sigma,p)
```





https://www.mathworks.com/help/stats/simulate-data-from-a-gaussian-mixture-model.html

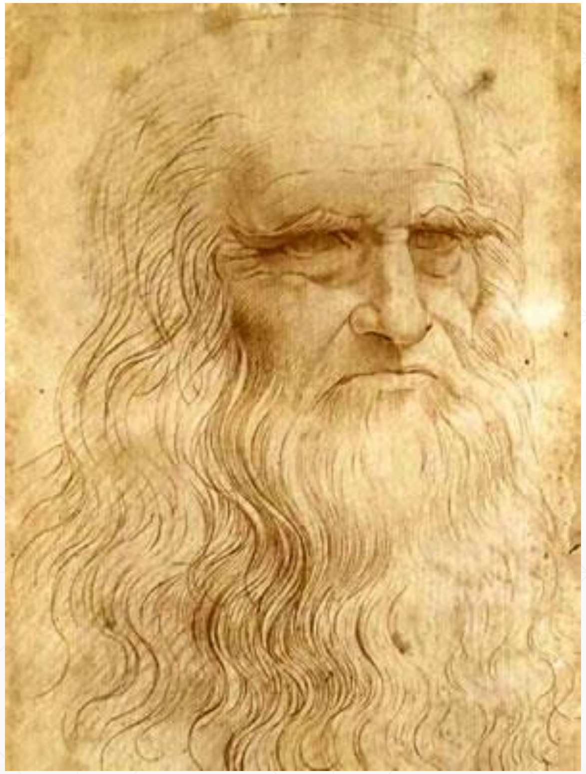
What does $p(x)$ looks like?



http://www.pymvpa.org/examples/mdp_mnist.html

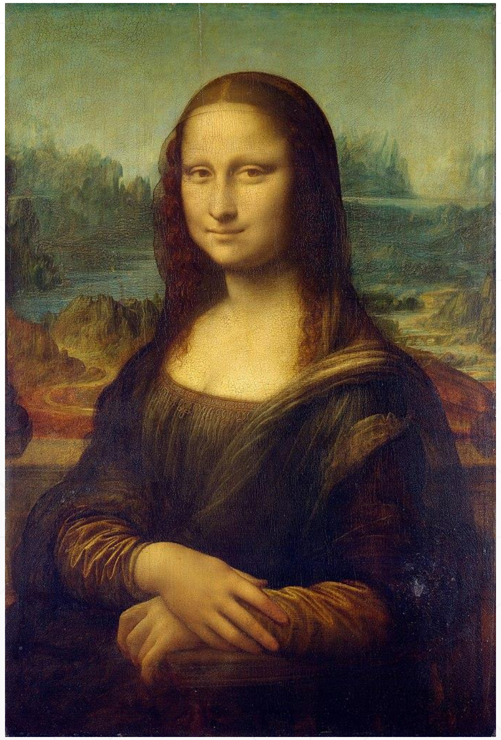
emm, how to learn $p(x)$

- Let's consider the case of growth up of a painter


When firstly began to paint

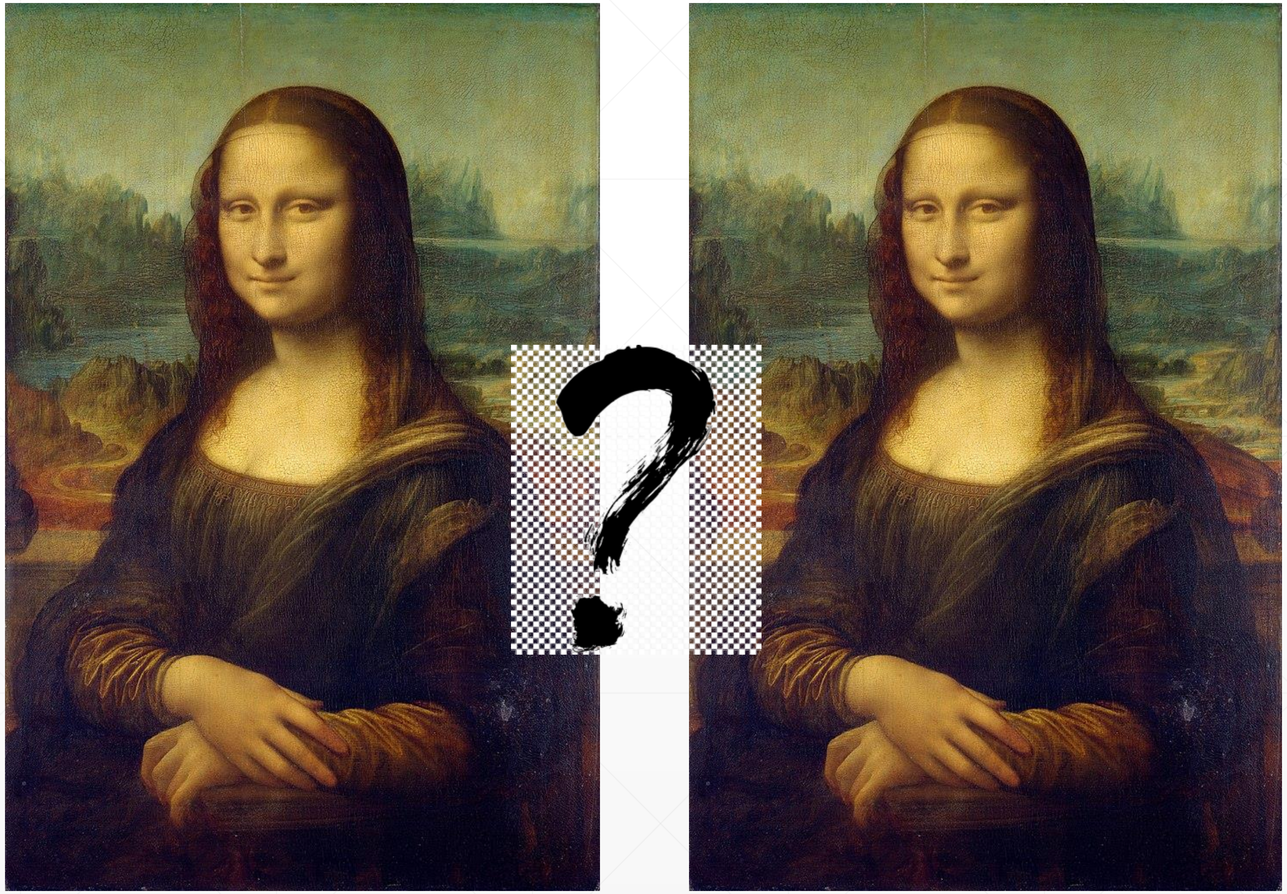
After learned by 5 years

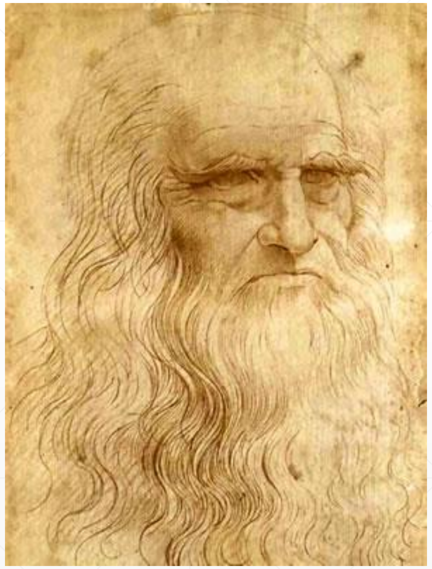


After learned by 10 years




Finally **Nash Equilibrium**

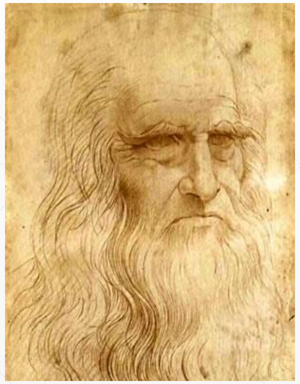


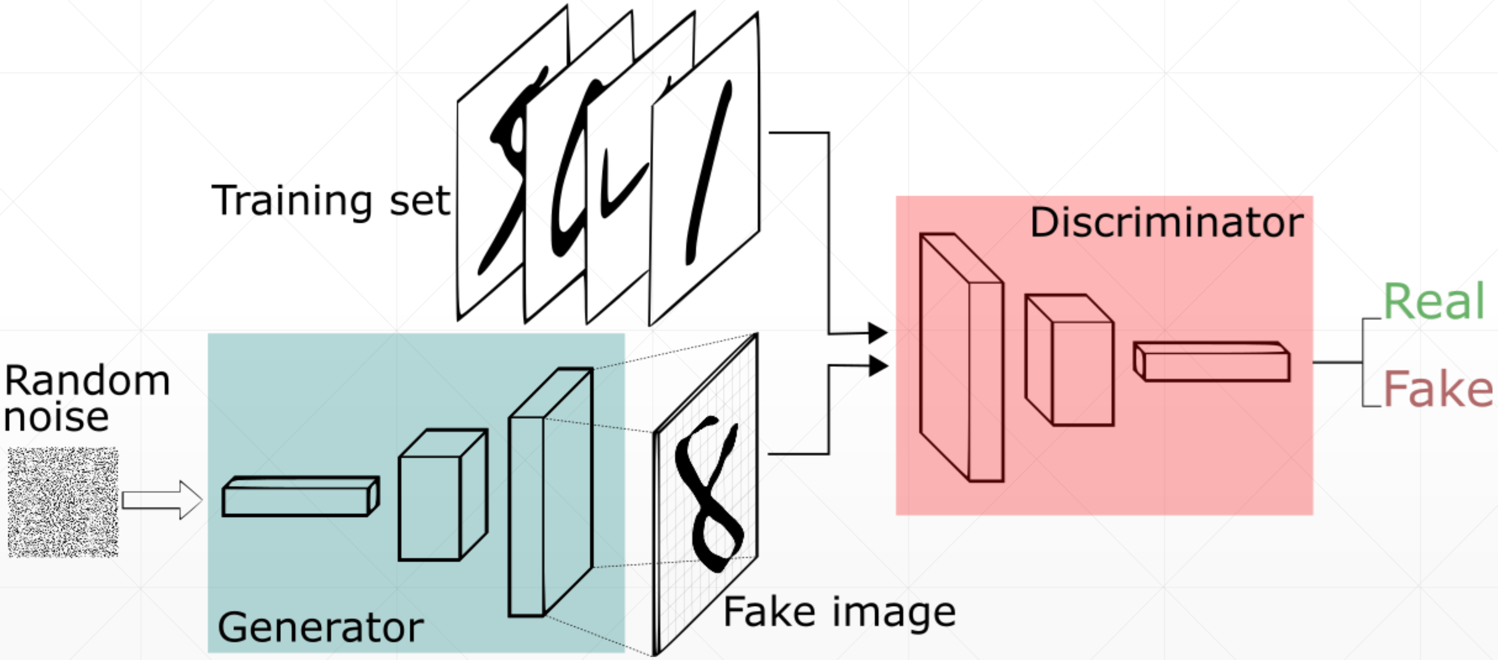


Put it down

- Painter or Generator:

- Critic or Discriminator





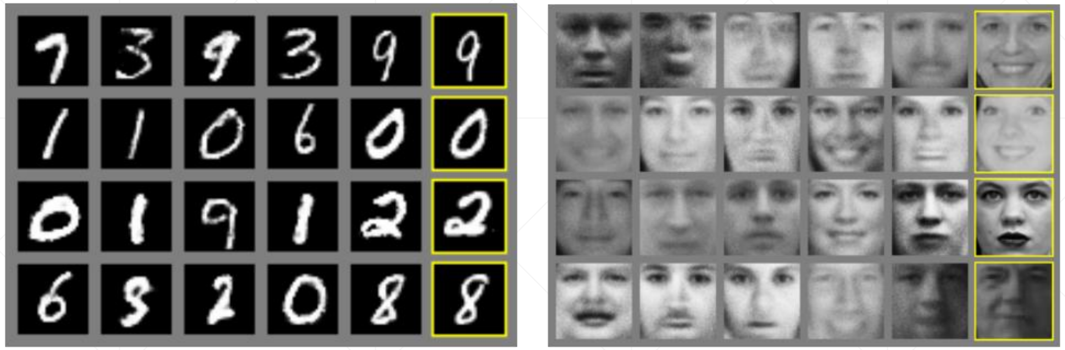
https://towardsdatascience.com/generative-adversarial-networks-explained-34472718707a

How to train?

$$\min_G \max_D L(D, G) = \mathbb{E}_{x \sim p_r(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$
$$= \mathbb{E}_{x \sim p_r(x)}[\log D(x)] + \mathbb{E}_{x \sim p_g(x)}[\log(1 - D(x)]$$

Done!

Figure 1: Class-conditional samples generated by our model.

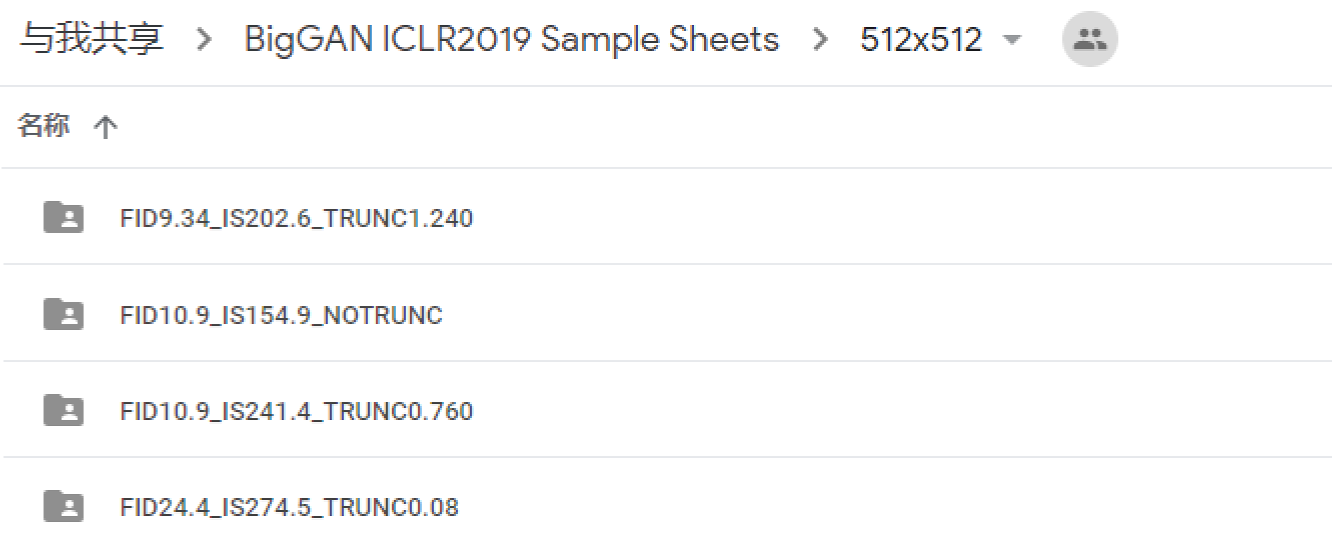


a) b)

c) d)

https://medium.com/syncedreview/biggan-a-new-state-of-the-art-in-image-synthesis-cf2ec5694024

See more realistic samples



https://drive.google.com/drive/folders/1lWC6XEPD0LT5KUnPXeve_kWeY-FxH002

Having Fun

- https://reiinakano.github.io/gan-playground/
- https://affinelayer.com/pixsrv/
- https://www.youtube.com/watch?v=9reHvktowLY&feature=youtu.be
- https://github.com/ajbrock/Neural-Photo-Editor
- https://github.com/nashory/gans-awesome-applications

The End ?

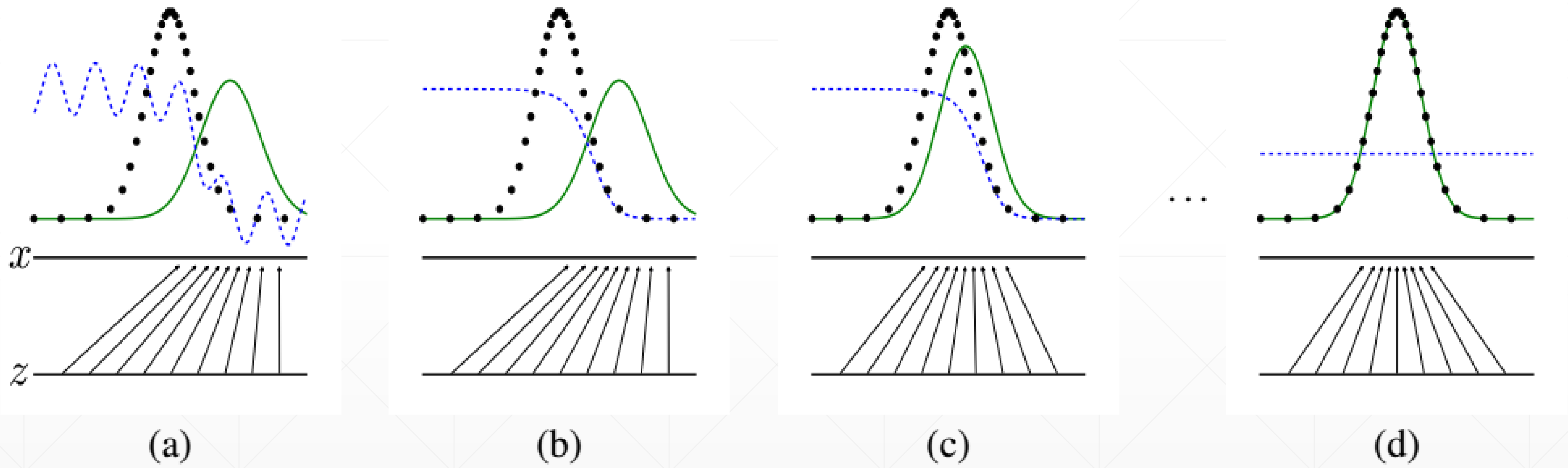
Never end

- Q1. Where will D converge, given fixed G

- Q2. Where will G converge, after optimal D

$$\begin{aligned}\min_G \max_D L(D,G) &= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \\ &= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1-D(x)]\end{aligned}$$

Intuition



Q1. Where will D go (fixed G)

Proposition 1. *For G fixed, the optimal discriminator D is*

$$D_G^*(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{2}$$

Proof. The training criterion for the discriminator D, given any generator G, is to maximize the quantity $V(G, D)$

$$\begin{aligned} V(G, D) &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_z p_{\boldsymbol{z}}(\boldsymbol{z}) \log(1 - D(g(\boldsymbol{z})))dz \\ &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1 - D(\boldsymbol{x}))dx \end{aligned} \tag{3}$$

$$f(\tilde{x}) = Alog\tilde{x} + Blog(1 - \tilde{x})$$

$$\begin{aligned} \frac{df(\tilde{x})}{d\tilde{x}} &= A\frac{1}{ln10}\frac{1}{\tilde{x}} - B\frac{1}{ln10}\frac{1}{1-\tilde{x}} \\ &= \frac{1}{ln10}\left(\frac{A}{\tilde{x}} - \frac{B}{1-\tilde{x}}\right) \\ &= \frac{1}{ln10}\frac{A-(A+B)\tilde{x}}{\tilde{x}(1-\tilde{x})} \end{aligned}$$

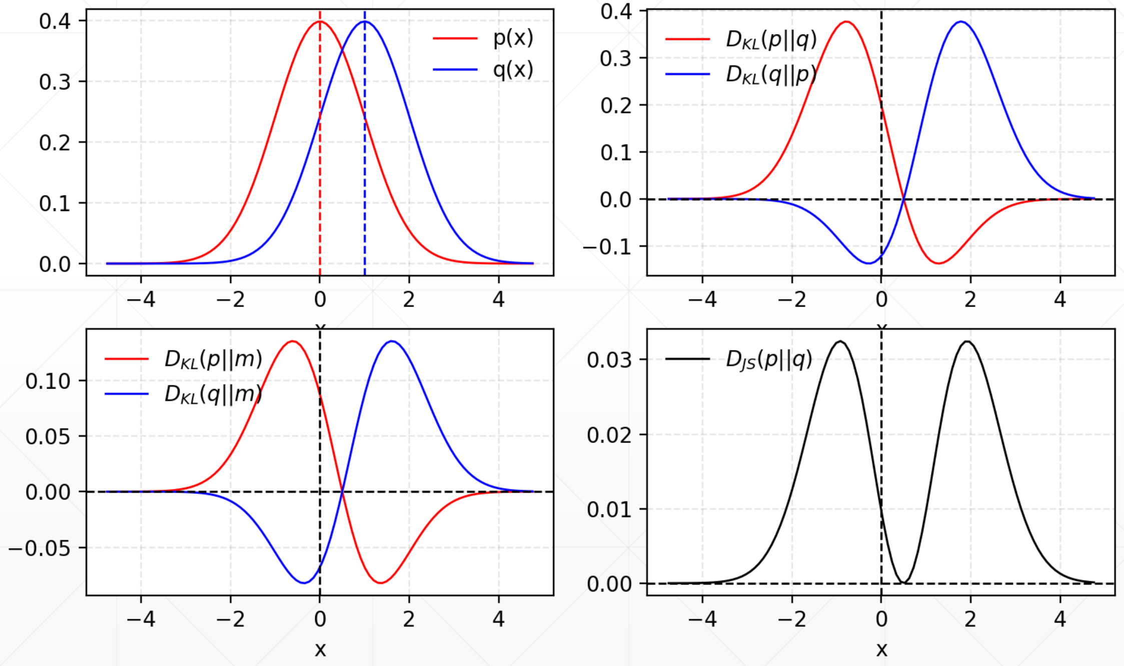
Thus, set $\frac{df(\tilde{x})}{d\tilde{x}} = 0$, we get the best value of the discriminator:

$D^*(x) = \tilde{x}^* = \frac{A}{A+B} = \frac{p_r(x)}{p_r(x)+p_g(x)} \in [0, 1]$

KL Divergence V.S. JS Divergence

$$D_{KL}(p\|q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$

$$D_{JS}(p\|q) = \frac{1}{2} D_{KL}(p\|\frac{p+q}{2}) + \frac{1}{2} D_{KL}(q\|\frac{p+q}{2})$$



Q2. Where will G go (after D*)

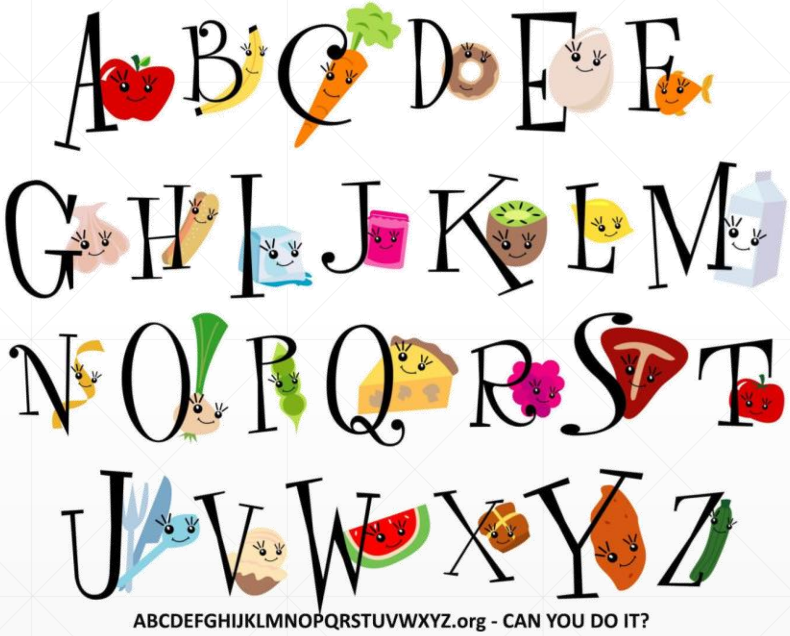
$$\begin{aligned} D_{JS}(p_r \| p_g) =& \frac{1}{2} D_{KL}(p_r || \frac{p_r + p_g}{2}) + \frac{1}{2} D_{KL}(p_g || \frac{p_r + p_g}{2}) \\ =& \frac{1}{2} \left(\log 2 + \int_x p_r(x) \log \frac{p_r(x)}{p_r + p_g(x)} dx \right) + \\ & \frac{1}{2} \left(\log 2 + \int_x p_g(x) \log \frac{p_g(x)}{p_r + p_g(x)} dx \right) \\ =& \frac{1}{2} \left(\log 4 + L(G, D^*) \right) \end{aligned}$$

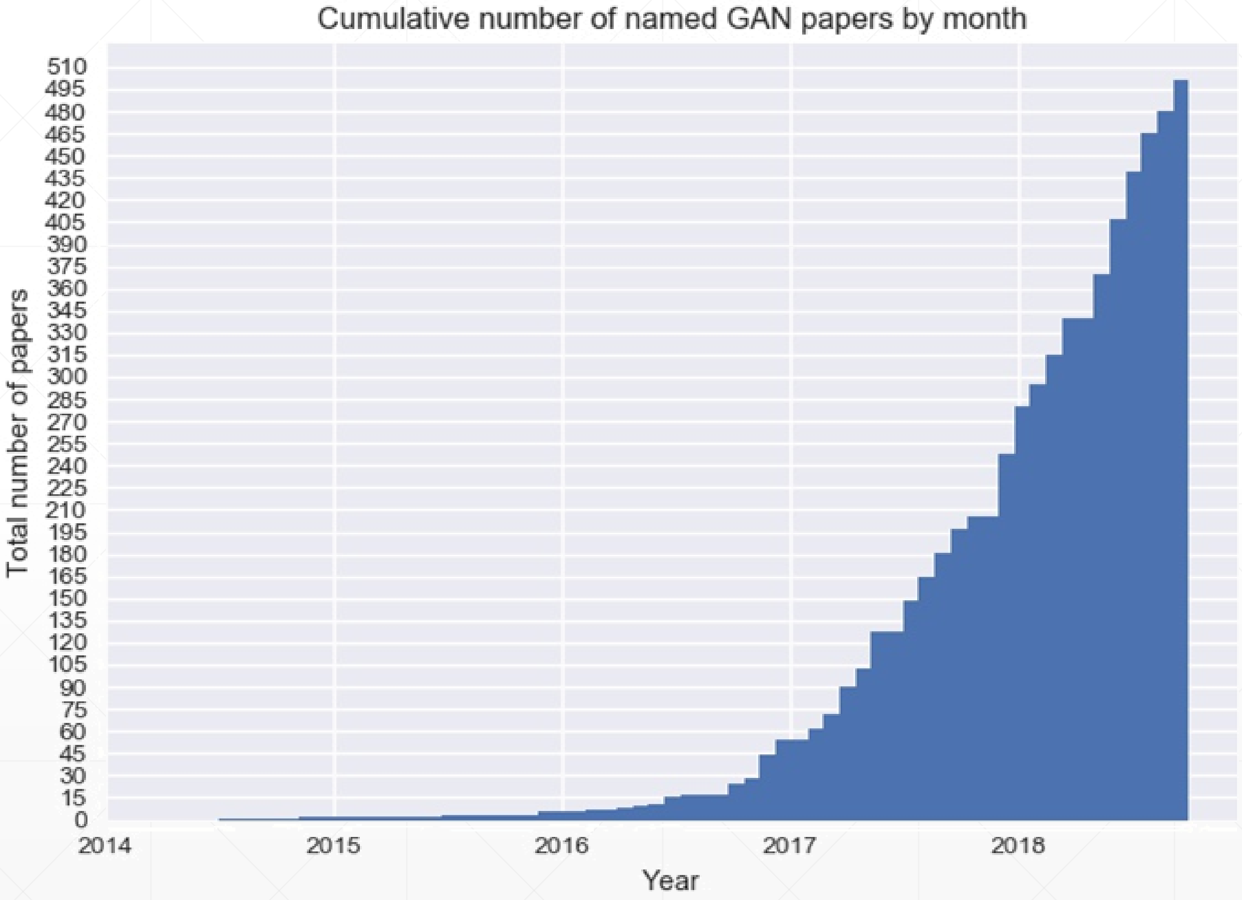
$$D_{JS}(p_r \parallel p_g) \geq 0$$

$$p_r = p_g$$

$$L(G, D^*) = 2D_{JS}(p_r \| p_g) - 2 \log 2$$

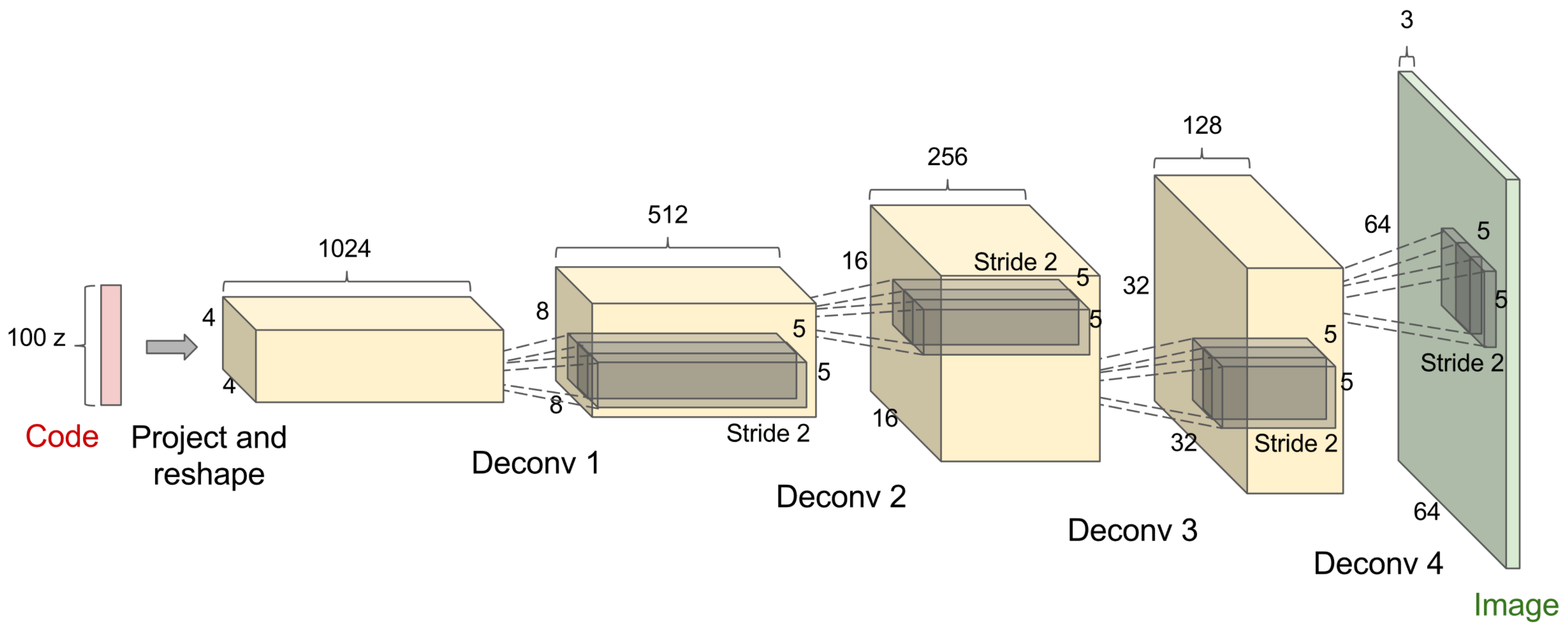
A~Z GAN





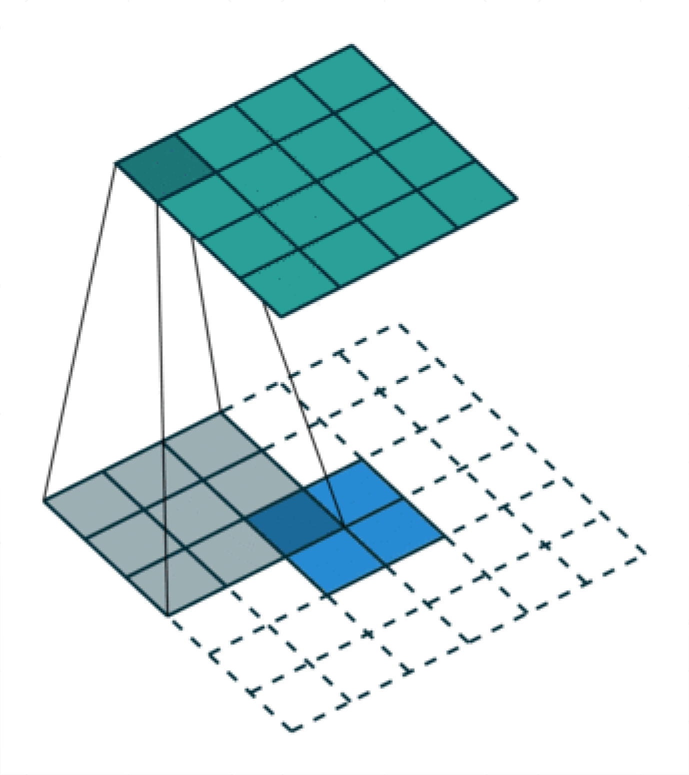
https://github.com/hindupuravinash/the-gan-zoo/blob/master/cumulative_gans.jpg

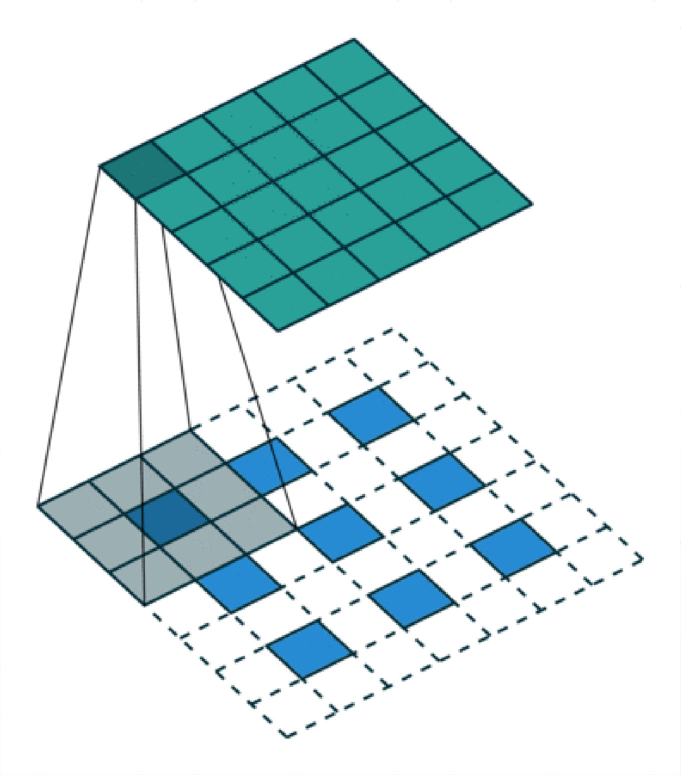
DCGAN



https://blog.openai.com/generative-models/

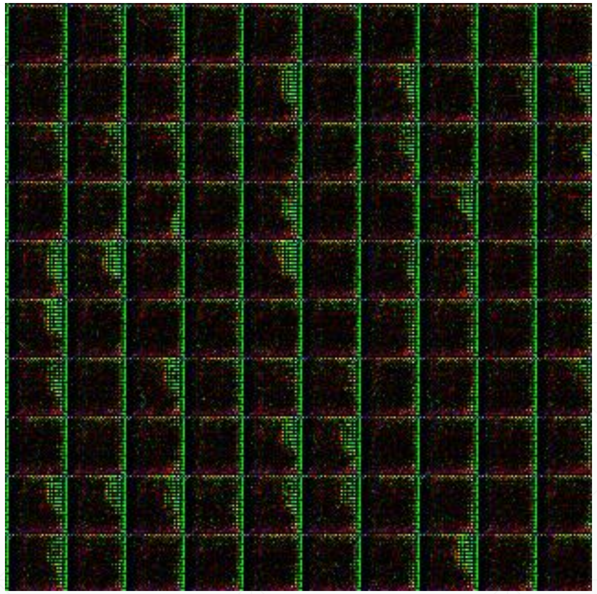
Transposed Convolution



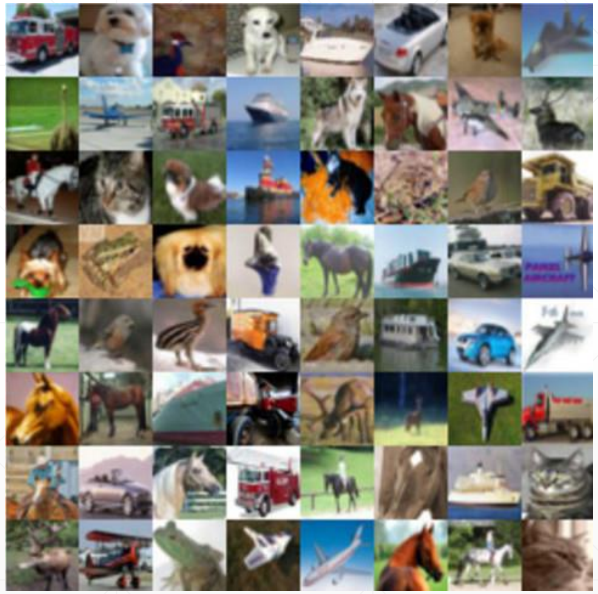


https://datascience.stackexchange.com/questions/6107/what-are-deconvolutional-layers

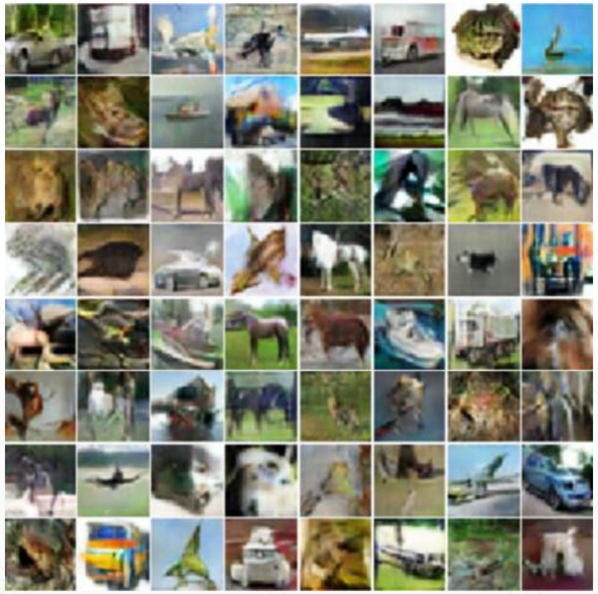
VAE V.S. DCGAN



GT



GAN



https://blog.openai.com/generative-models/

The Last thing?

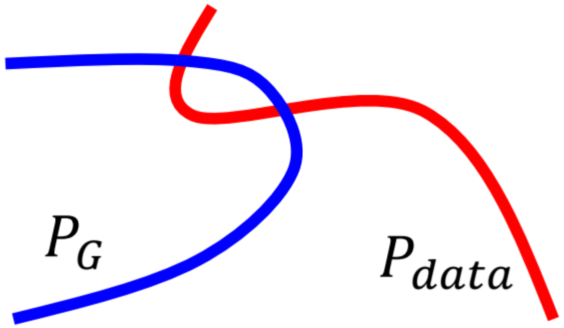
- Training Stability

Why?

- In most cases, P_G and P_{data} are not overlapped.
- 1. The nature of data

Both P_{data} and P_G are low-dim manifold in high-dim space.

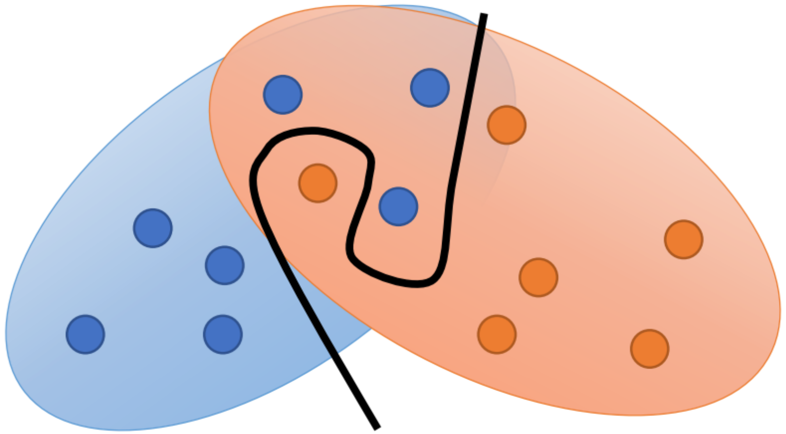
The overlap can be ignored.



- 2. Sampling

Even though P_{data} and P_G have overlap.

If you do not have enough sampling ……



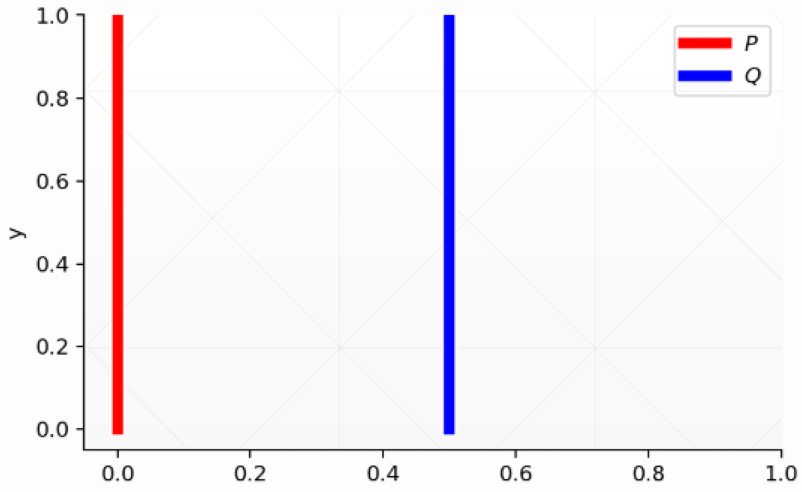
http://speech.ee.ntu.edu.tw/~tlkagk/courses/MLDS_2018/Lecture/WGAN%20(v2).pdf

$$D_{KL}(p\|q) = \int_x p(x)\log\frac{p(x)}{q(x)}dx$$

$$D_{JS}(p\|q) = \frac{1}{2}D_{KL}(p\|\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q\|\frac{p+q}{2})$$

Toy example

$$\forall(x,y)\in P, x=0 \text{ and } y\sim U(0,1)$$
$$\forall(x,y)\in Q, x=\theta, 0\leq\theta\leq 1 \text{ and } y\sim U(0,1)$$



When $\theta \neq 0$:

$$D_{KL}(P\|Q) = \sum_{x=0,y\sim U(0,1)} 1\cdot\log\frac{1}{0} = +\infty$$
$$D_{KL}(Q\|P) = \sum_{x=\theta,y\sim U(0,1)} 1\cdot\log\frac{1}{0} = +\infty$$
$$D_{JS}(P,Q) = \frac{1}{2}\left(\sum_{x=0,y\sim U(0,1)} 1\cdot\log\frac{1}{1/2} + \sum_{x=0,y\sim U(0,1)} 1\cdot\log\frac{1}{1/2}\right) = \log 2$$
$$W(P,Q) = |\theta|$$

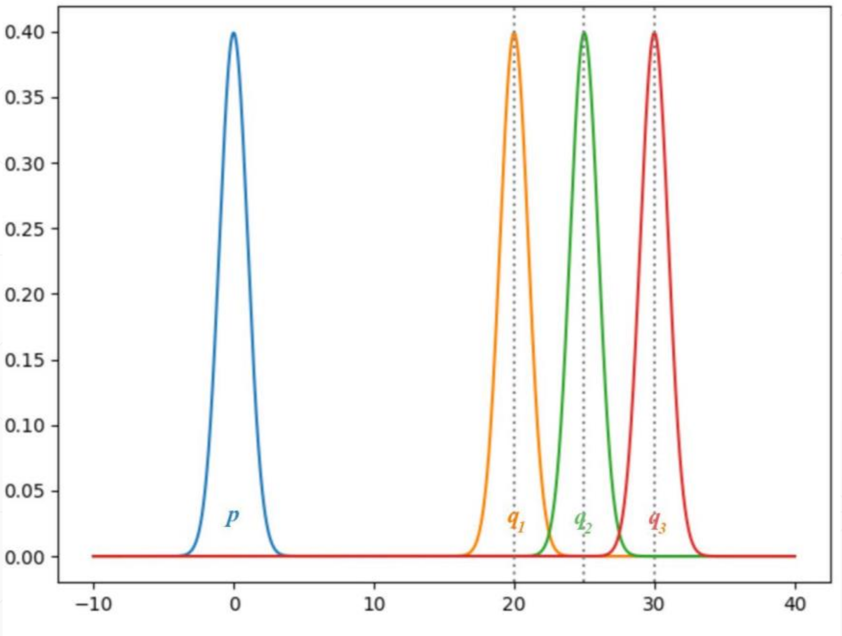
But when $\theta = 0$, two distributions are fully overlapped:

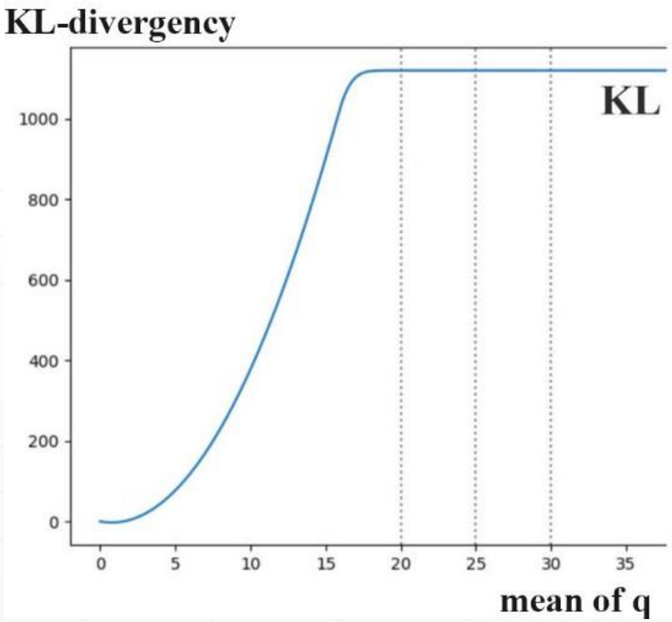
$$D_{KL}(P\|Q) = D_{KL}(Q\|P) = D_{JS}(P,Q) = 0$$
$$W(P,Q) = 0 = |\theta|$$

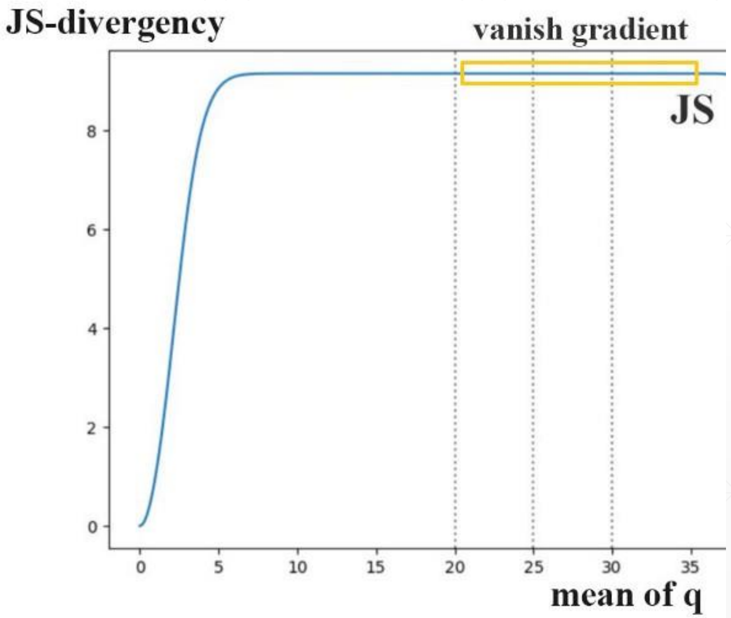
Toy example

MLE is kind of minimize KLD

GAN





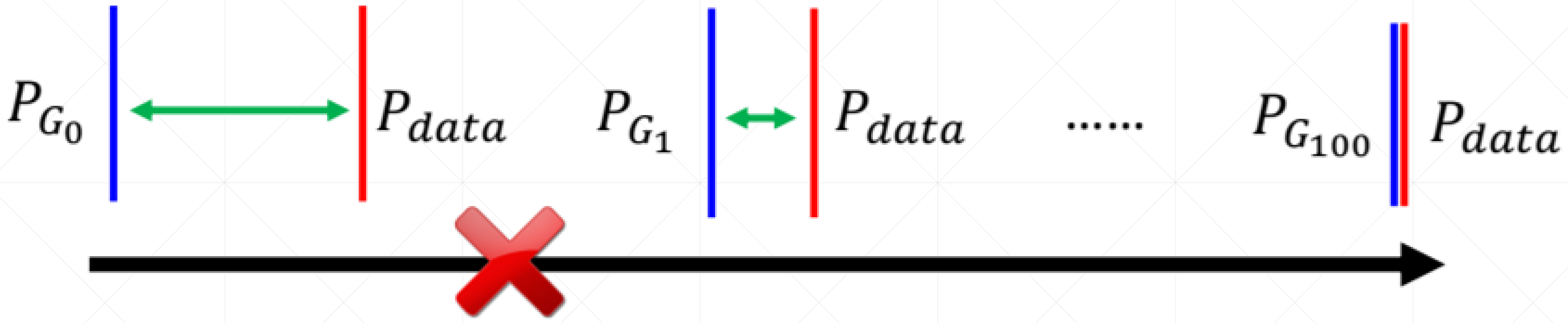


https://medium.com/@jonathan_hui/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490
https://jhui.github.io/2017/01/05/Deep-learning-Information-theory/

$$\min_G \max_D L(D, G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

$$= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x)]$$

JS Divergence



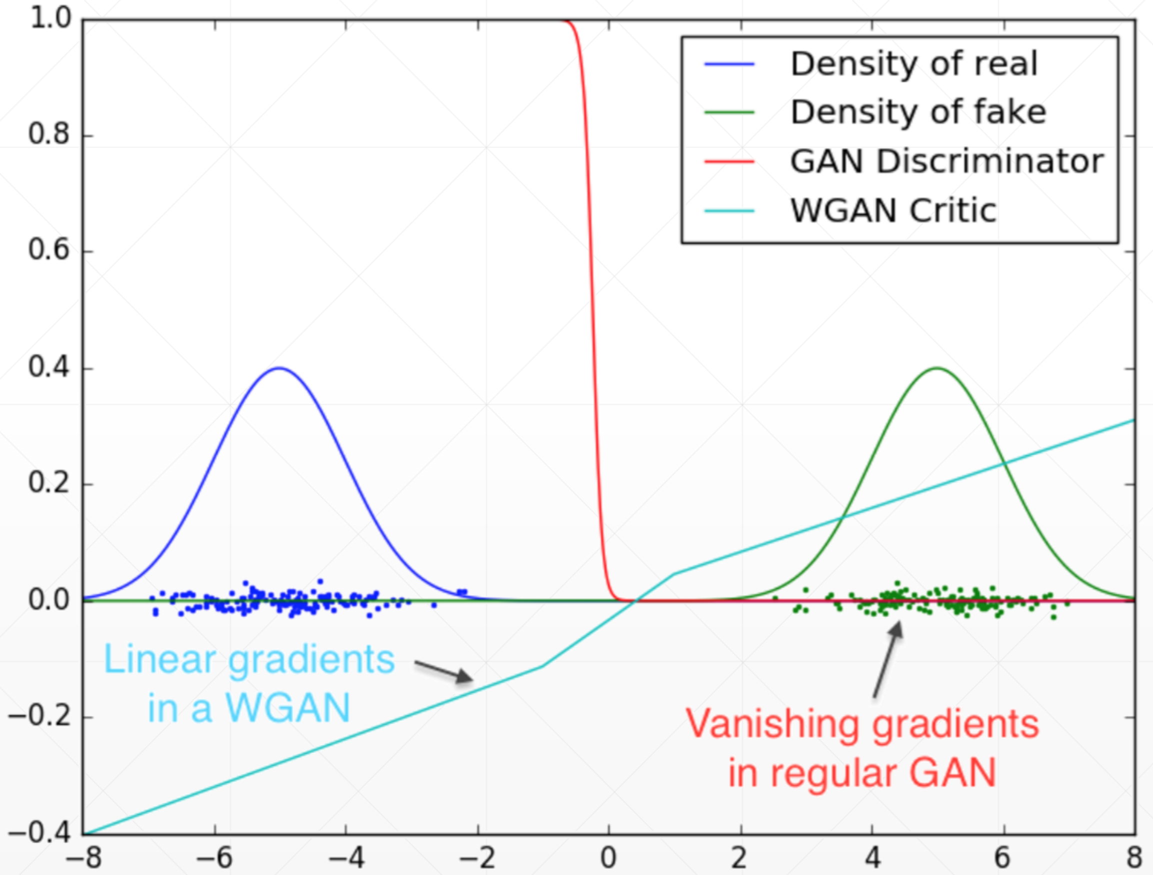
Equally bad

$JS(P_{G_0}, P_{data}) = log2$ $\quad$ $JS(P_{G_1}, P_{data}) = log2$ $\quad$ …… $\quad$ $JS(P_{G_{100}}, P_{data}) = 0$

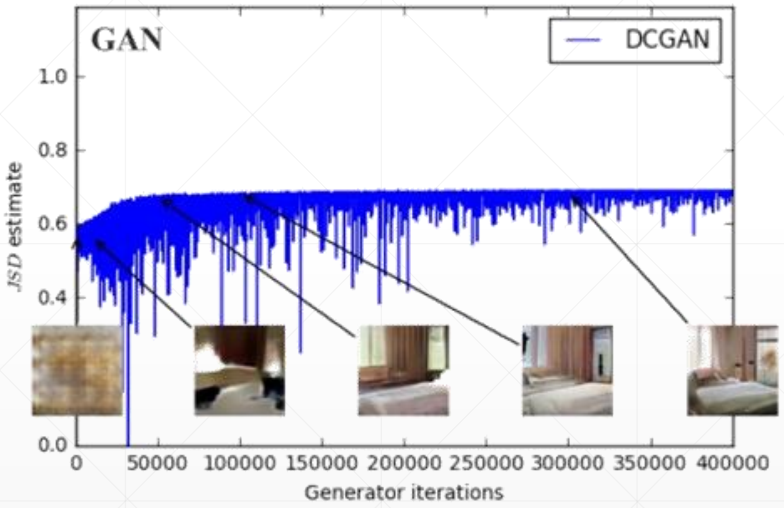
JS divergence is log2 if two distributions do not overlap.

Intuition: If two distributions do not overlap, binary classifier achieves 100% accuracy

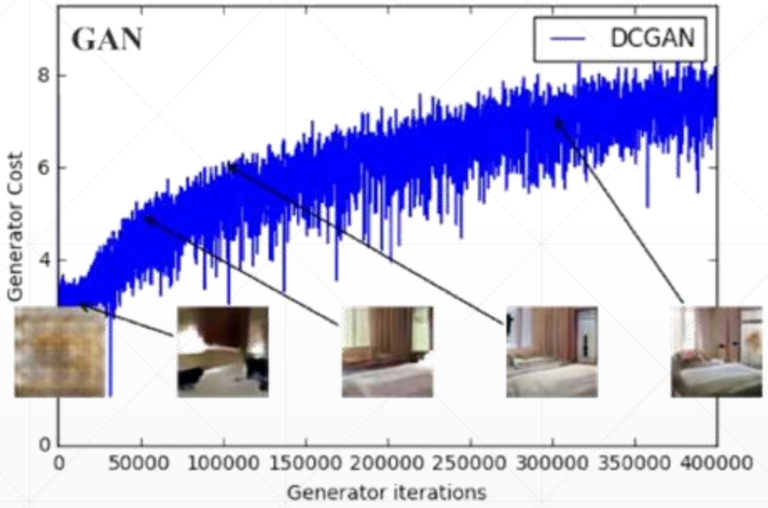
Gradient Vanishing



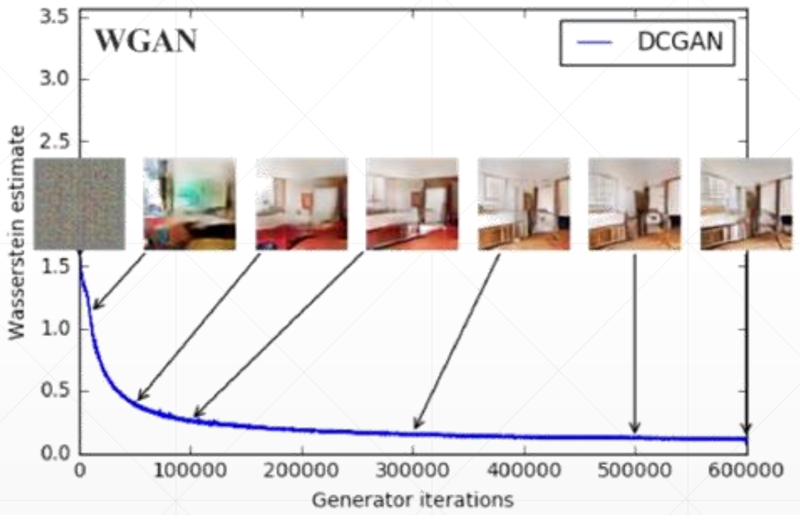
Training Progress Invisible



$$\frac{1}{m}\sum_{i=1}^{m}\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right)$$

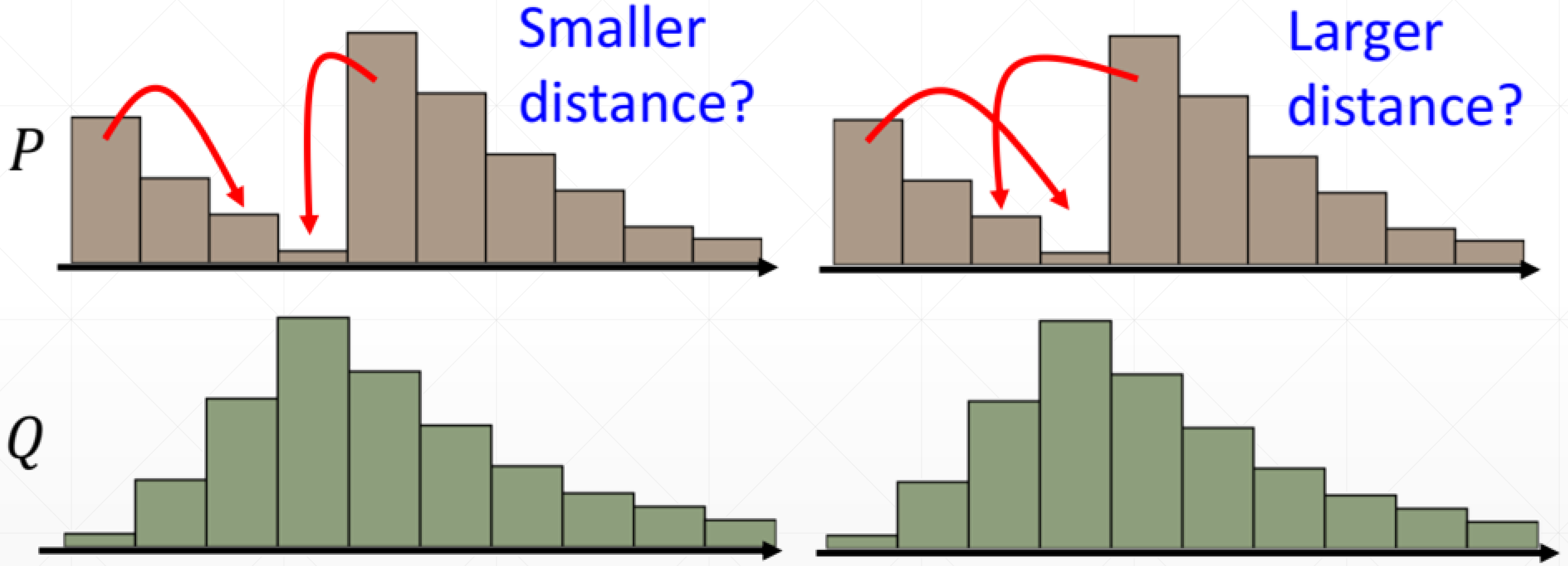


$$\frac{1}{m}\sum_{i=1}^{m}-\log\left(D\left(G\left(z^{(i)}\right)\right)\right)$$



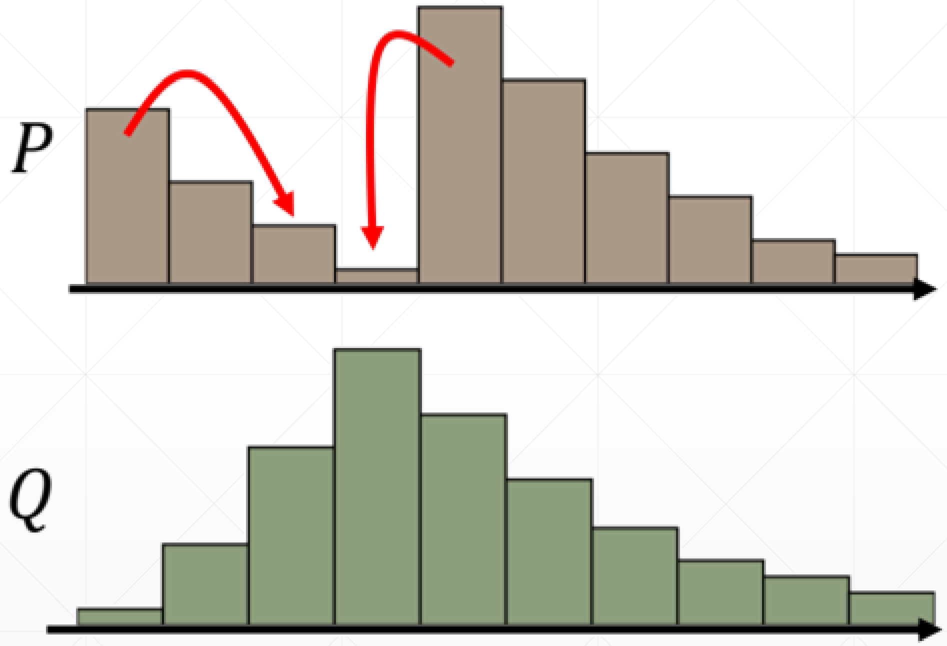
$$\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$$

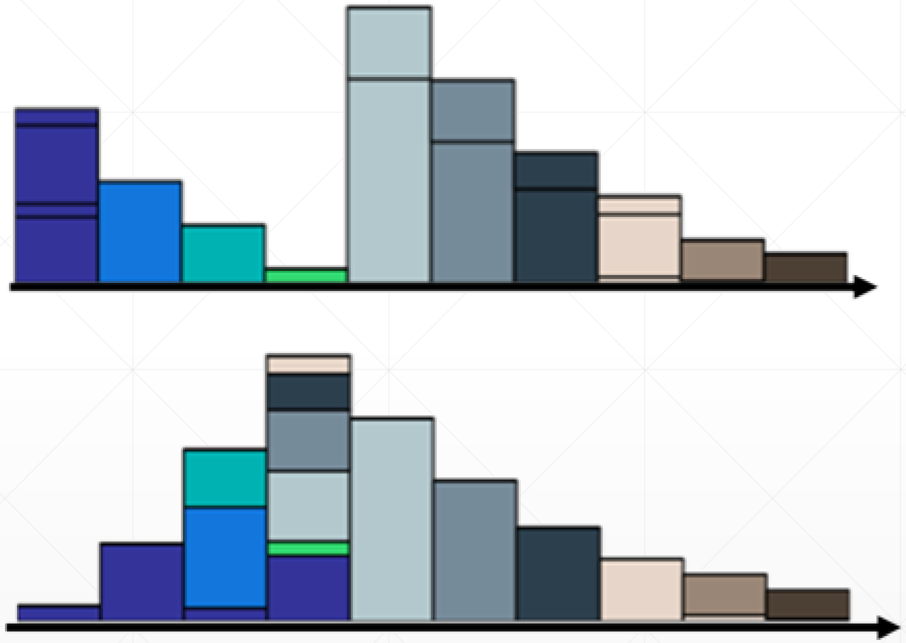
HowTo



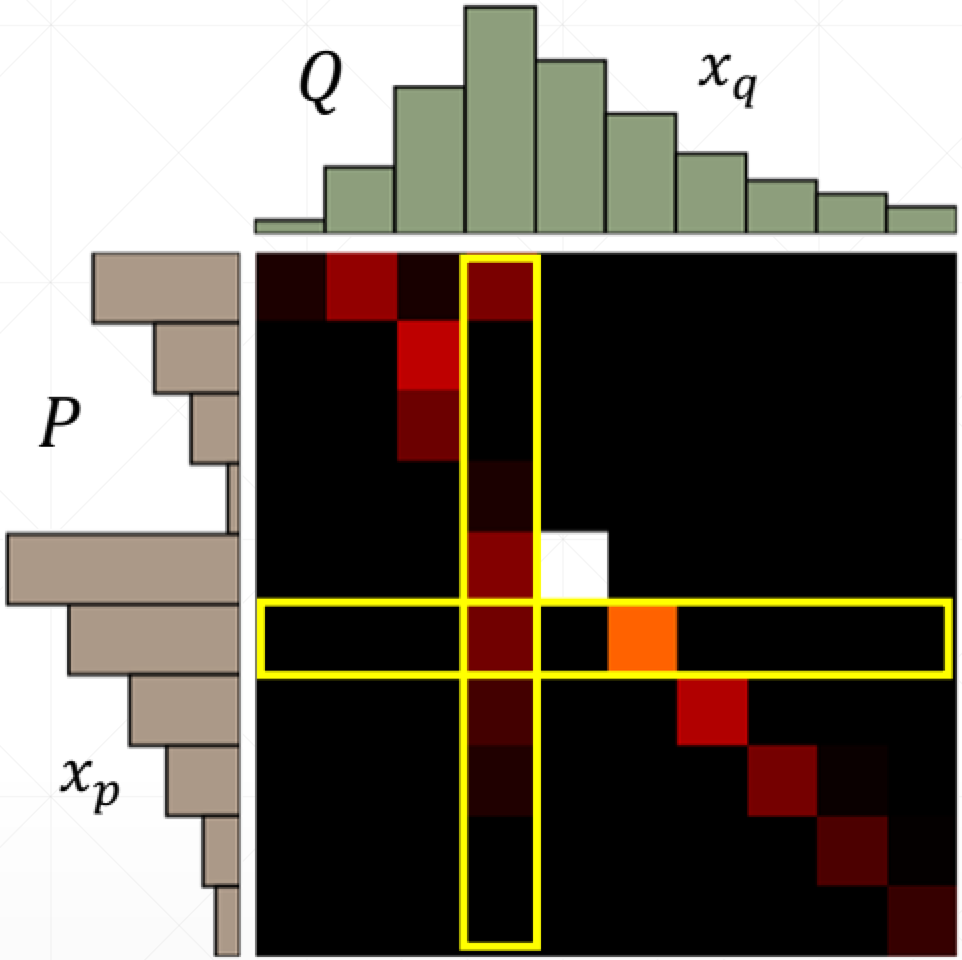
The Least Cost among plans

Best "moving plans" of this example





There many possible "moving plans".



moving plan γ
All possible plan Π

A "moving plan" is a matrix

The value of the element is the amount of earth from one position to another.

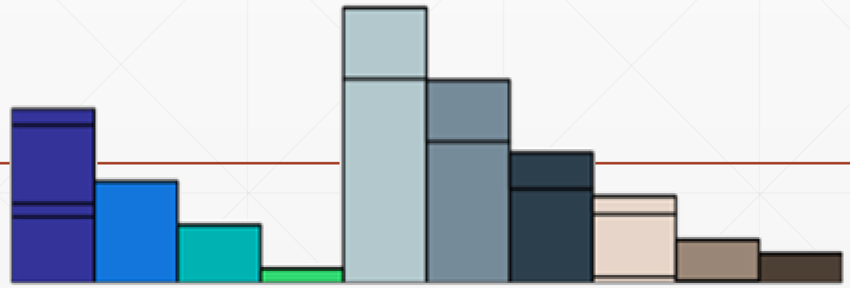
Average distance of a plan γ:

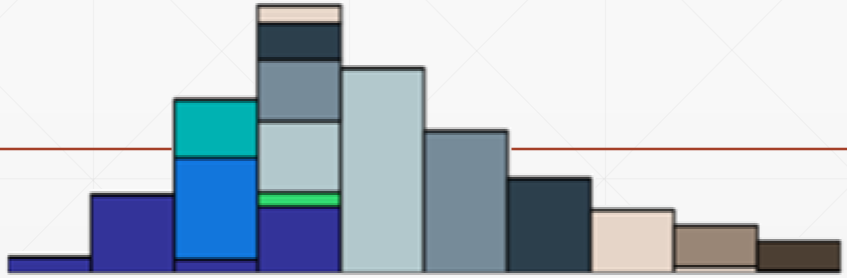
$$B(\gamma) = \sum_{x_p, x_q} \gamma(x_p, x_q) \| x_p - x_q \|$$

Earth Mover's Distance:

$$W(P, Q) = \min_{\gamma \in \Pi} B(\gamma)$$

The best plan





How to compute Wasserstein Distance

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\, \|x - y\| \,\right] ,$$

$$\left[f\left(x^{(i)}\right) - f\left(G\left(z^{(i)}\right)\right)\right]$$

	Discriminator/Critic	Generator
GAN	$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)\right]$	$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} -\log\left(D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right)$
WGAN	$\nabla_w \frac{1}{m} \sum_{i=1}^{m} \left[f\left(x^{(i)}\right) - f\left(G\left(z^{(i)}\right)\right)\right]$	$\nabla_\theta \frac{1}{m} \sum_{i=1}^{m} -f\left(G\left(z^{(i)}\right)\right)$

$$|f(x_1) - f(x_2)| \le |x_1 - x_2| .$$

1-Lipschitz function

WGAN

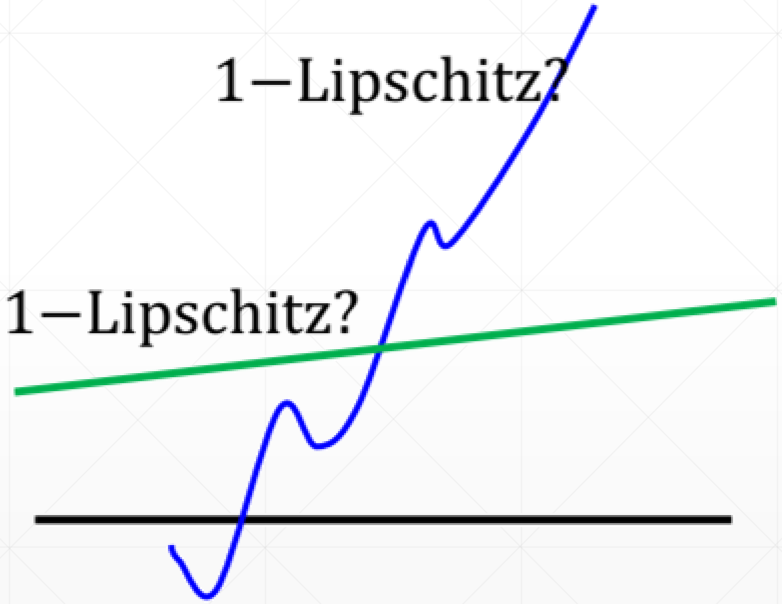
Weight Clipping [Martin Arjovsky, et al., arXiv, 2017]

Force the parameters w between c and -c

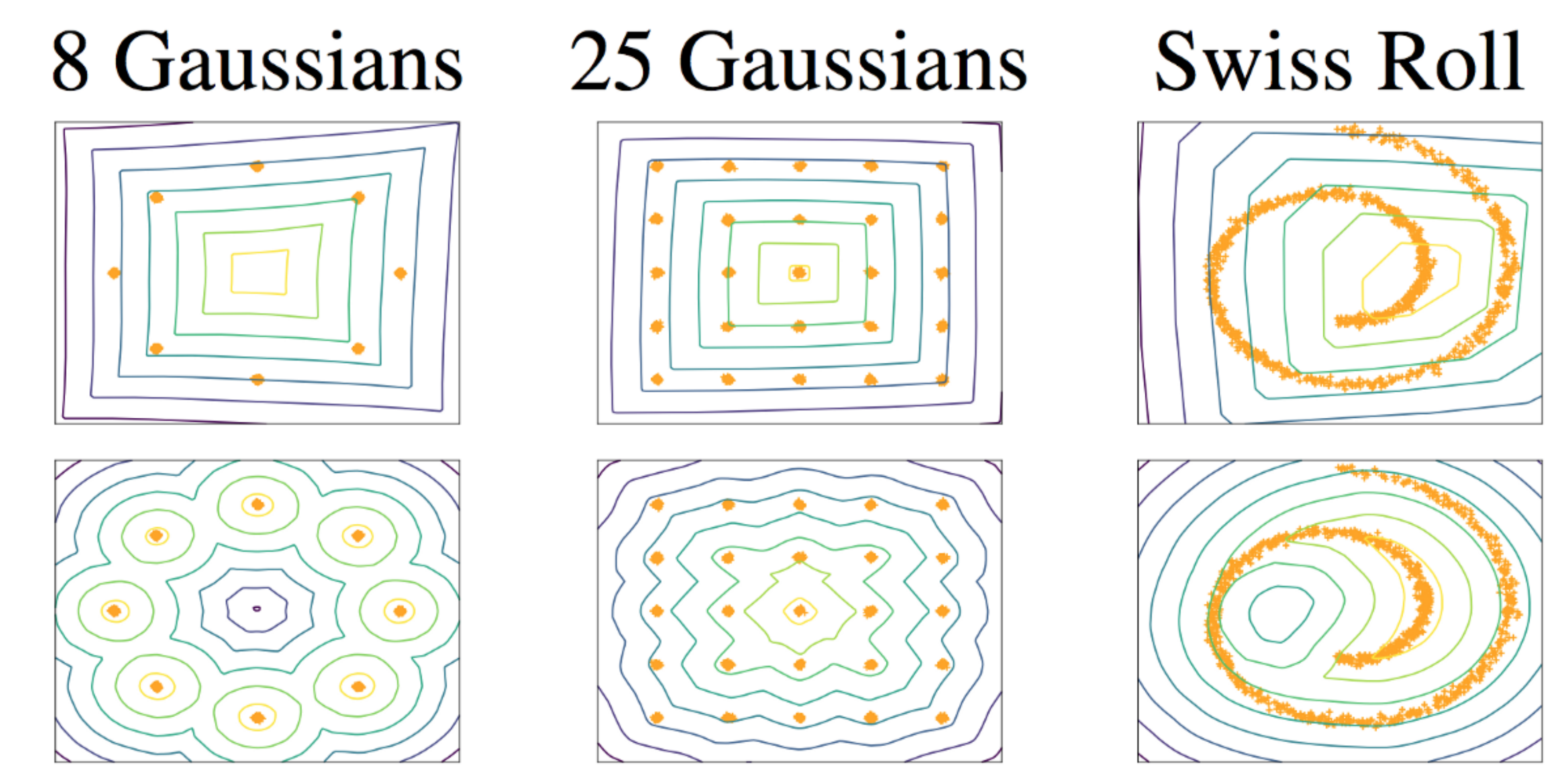
After parameter update, if w > c, w = c;

if w < -c, w = -c

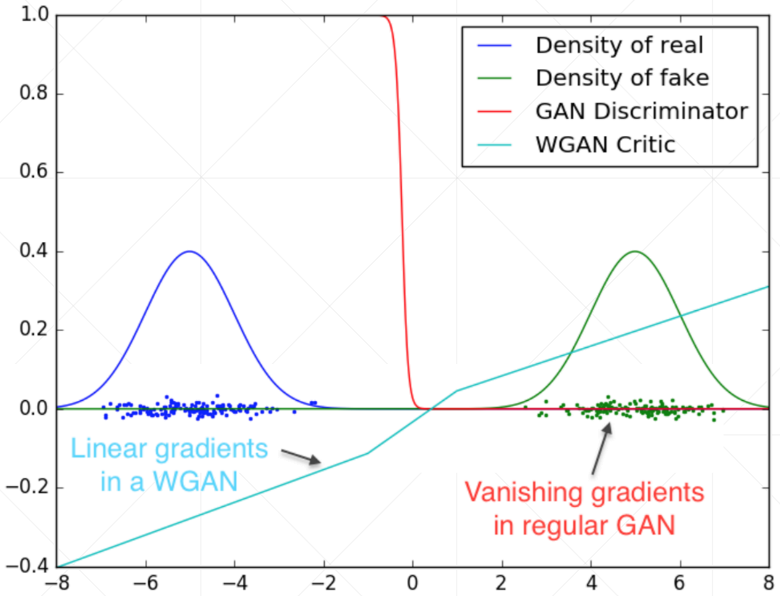
How to fulfill this constraint?



Sort of Regularization



WGAN-Gradient Penalty



$$|f(x_1) - f(x_2)| \le |x_1 - x_2|.$$

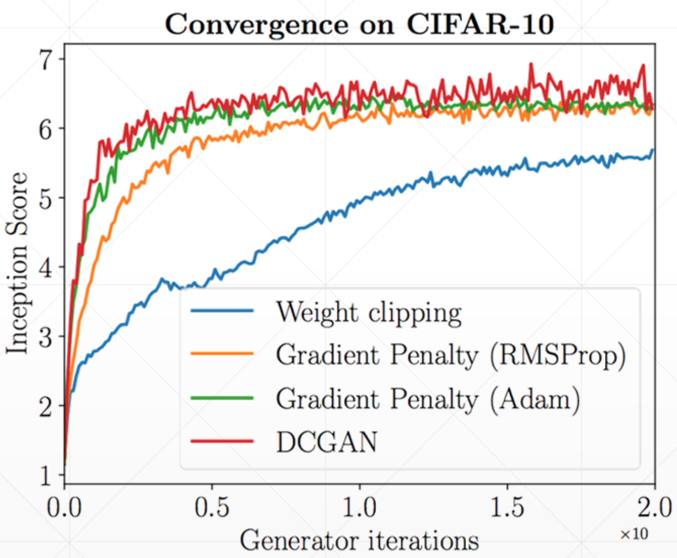
$$L = \underbrace{\underset{\tilde{\boldsymbol{x}} \sim \mathbb{P}_g}{\mathbb{E}}\left[D(\tilde{\boldsymbol{x}})\right] - \underset{\boldsymbol{x} \sim \mathbb{P}_r}{\mathbb{E}}\left[D(\boldsymbol{x})\right]}_{\text{Original critic loss}} + \underbrace{\lambda \underset{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}}{\mathbb{E}}\left[(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1)^2\right]}_{\text{Our gradient penalty}}.$$

where $\hat{\boldsymbol{x}}$ *sampled from* $\tilde{\boldsymbol{x}}$ *and* $\boldsymbol{x}$ *with t uniformly sampled between 0 and 1*

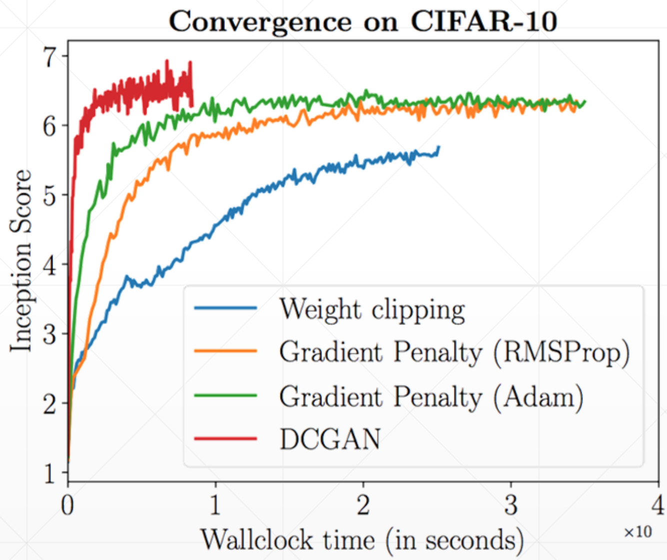
$$\hat{\boldsymbol{x}} = t\tilde{\boldsymbol{x}} + (1-t)\boldsymbol{x} \text{ with } 0 \le t \le 1$$

More stable

DCGAN
LSGAN
WGAN (clipping)
WGAN-GP (ours)
Baseline (G: DCGAN, D: DCGAN)
G: No BN and a constant number of filters, D: DCGAN
G: 4-layer 512-dim ReLU MLP, D: DCGAN
No normalization in either G or D
Gated multiplicative nonlinearities everywhere in G and D
tanh nonlinearities everywhere in G and D
101-layer ResNet G and D



Convergence on CIFAR-10
Inception Score
Generator iterations
×10
Weight clipping
Gradient Penalty (RMSProp)
Gradient Penalty (Adam)
DCGAN



Convergence on CIFAR-10
Inception Score
Wallclock time (in seconds)
×10
Weight clipping
Gradient Penalty (RMSProp)
Gradient Penalty (Adam)
DCGAN

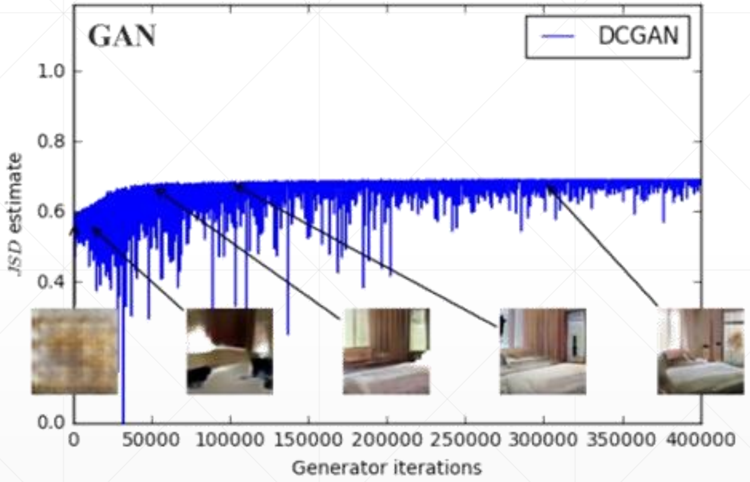
Unsupervised

Method	Score
ALI [8] (in [27])	$5.34 \pm .05$
BEGAN [4]	5.62
DCGAN [22] (in [11])	$6.16 \pm .07$
Improved GAN (-L+HA) [23]	$6.86 \pm .06$
EGAN-Ent-VI [7]	$7.07 \pm .10$
DFM [27]	$7.72 \pm .13$
WGAN-GP ResNet (ours)	$7.86 \pm .07$

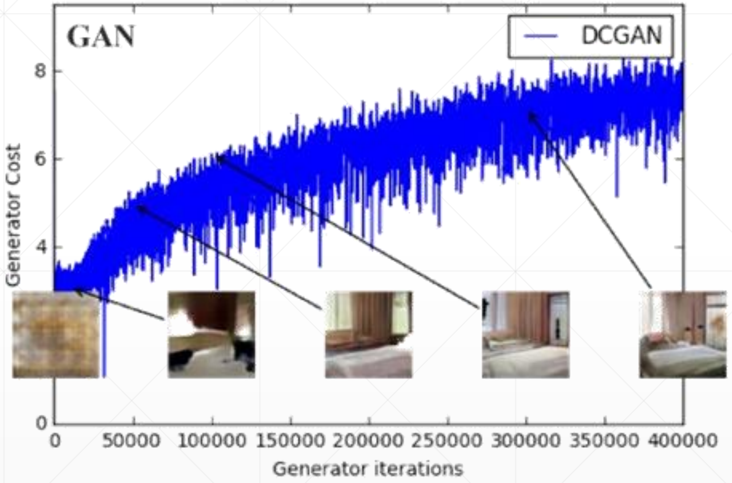
Supervised

Method	Score
SteinGAN [26]	6.35
DCGAN (with labels, in [26])	6.58
Improved GAN [23]	$8.09 \pm .07$
AC-GAN [20]	$8.25 \pm .07$
SGAN-no-joint [11]	$8.37 \pm .08$
WGAN-GP ResNet (ours)	$8.42 \pm .10$
SGAN [11]	$8.59 \pm .12$

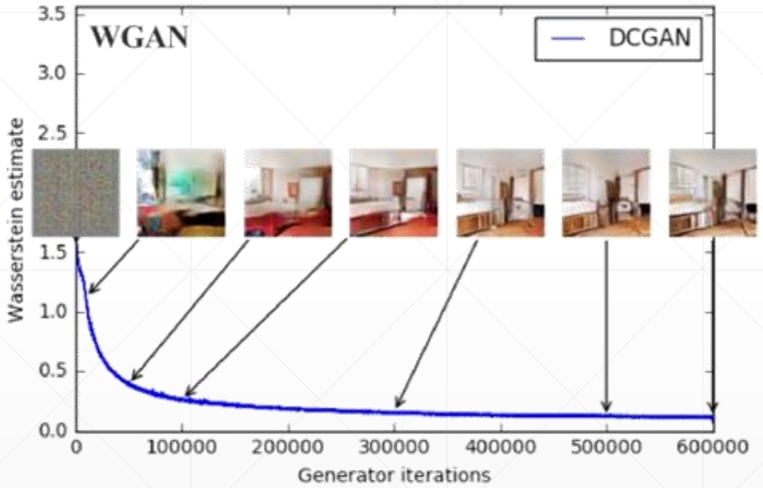
Training Progress Indicator



$$\frac{1}{m}\sum_{i=1}^{m}\log\left(1-D\left(G\left(z^{(i)}\right)\right)\right)$$



$$\frac{1}{m}\sum_{i=1}^{m}-\log\left(D\left(G\left(z^{(i)}\right)\right)\right)$$



$$\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$$

Thank You.