Week 7: Adversarial Models or Integral Probability Measures

Models so Far

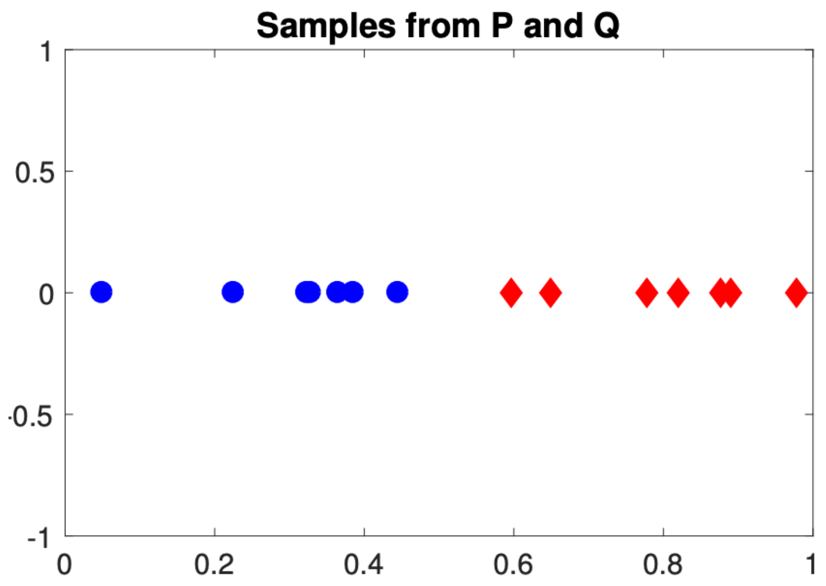
- All models so far used maximum likelihood estimation
- Two main directions:
- Precise likelihood estimation:
    - AR and flow-models
    - Fewer approximations but worse scaling behaviour
- Approximate likelihood estimation:
    - VI and diffusion models

This Week: Model without Point Estimation

- This week – consider a next type of model
- Will not require implicit likelihood estimation
- Potentially allow for more expressive models

Two-Sample Test

- We are given N samples from two distributions p(x) and q(y)
- Several questions we can ask:
  - Are the p and q distributions different?
  - What is the distance between these distributions?
- Statistical test to answer the first question – “two sample test”



Probability Divergence

- A way to measure the difference between the distributions
- Does not have to be symmetric
- Several famous divergences:

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x,$$

Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$	$(\sqrt{u}-1)^2$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \mathrm{d}x$	$-(u+1) \log \frac{1+u}{2} + u \log u$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \mathrm{d}x - \log(4)$	$u \log u - (u+1) \log(u+1)$

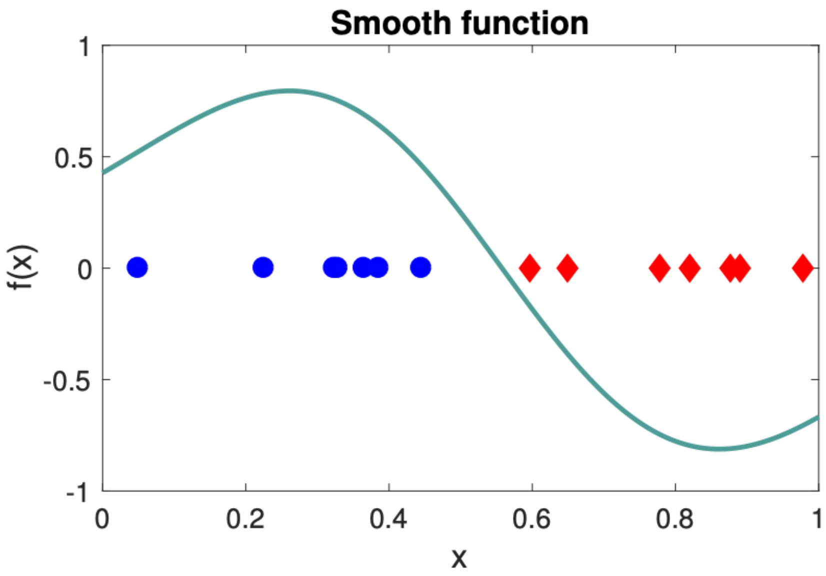
Calculating Divergence Between Samples

- How do we compute the divergence between two samples?
- Naïve Idea: Estimate p and q, then compute divergence
- Very hard as estimation is something we wish to avoid

Integral Probability Metrics

- Want to compute two sample distance without estimating p,q
- Idea - compute a function f(x) which maximizes:

$$E_P[f(x)] - E_Q[f(x)]$$

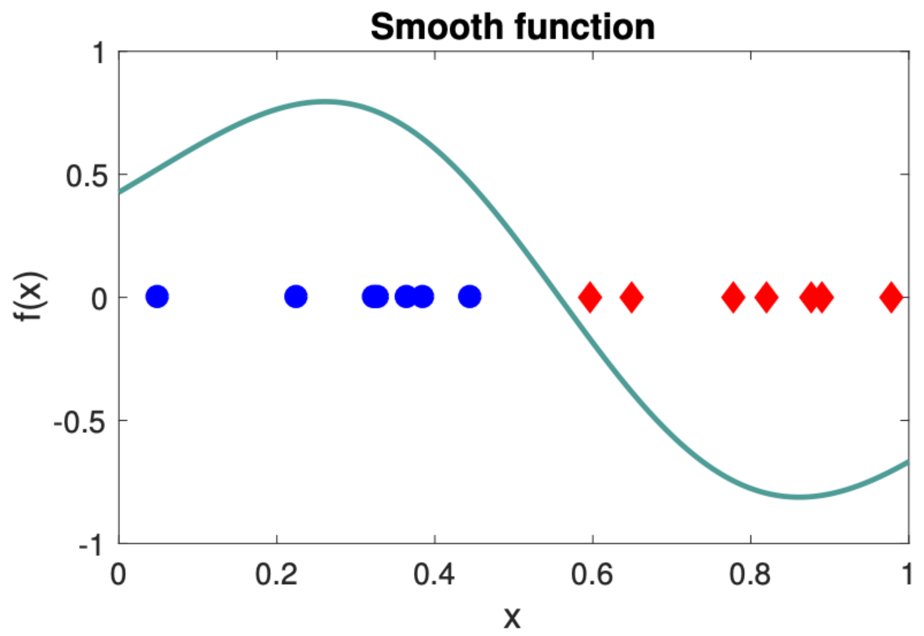


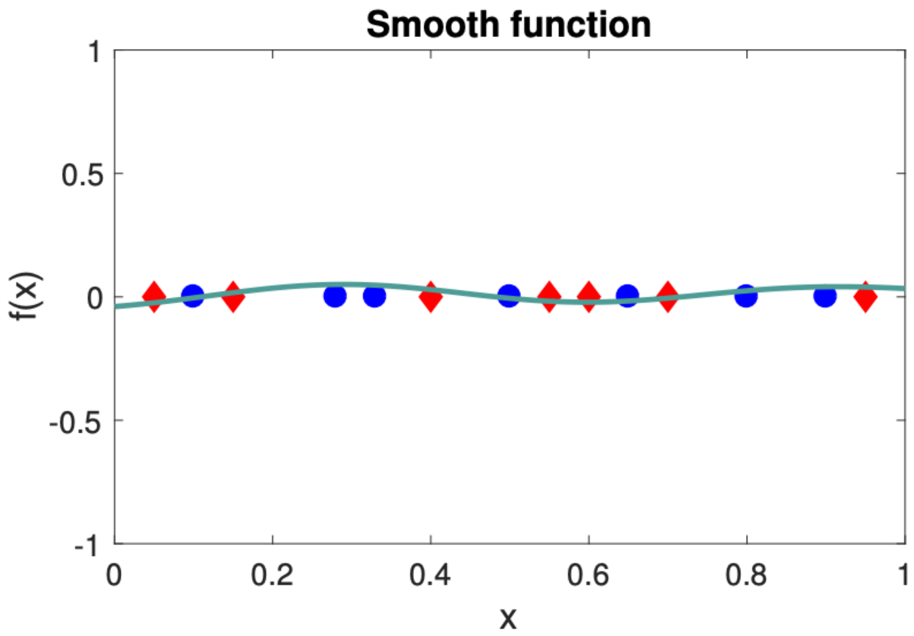
Such IPM Capture Near and Far Samples

- The difference is large when P,Q are different and small when near

$$E_P[f(x)] - E_Q[f(x)]$$

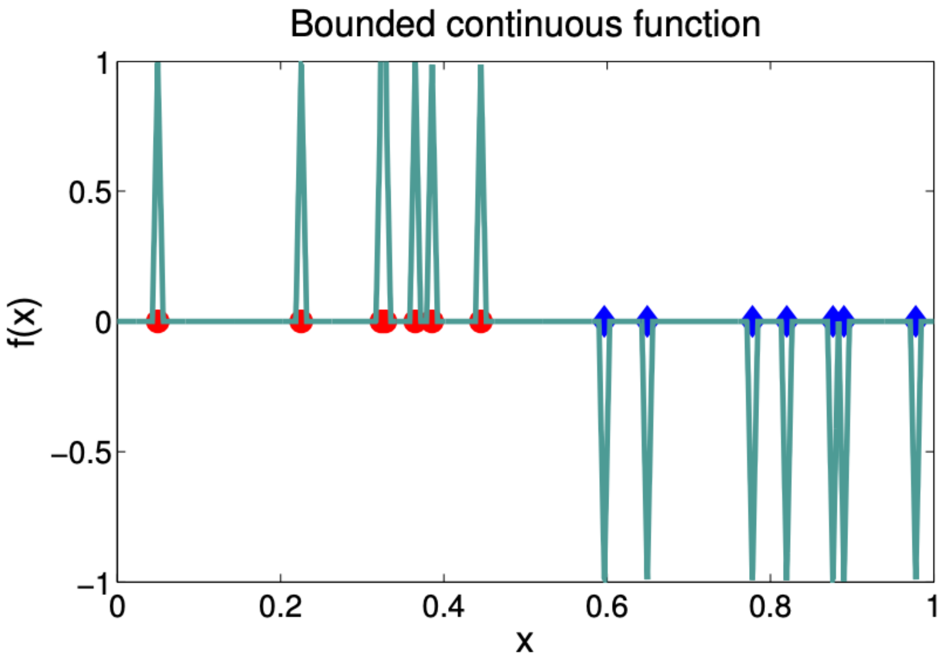
- Wonderful, so we have our two sample test!

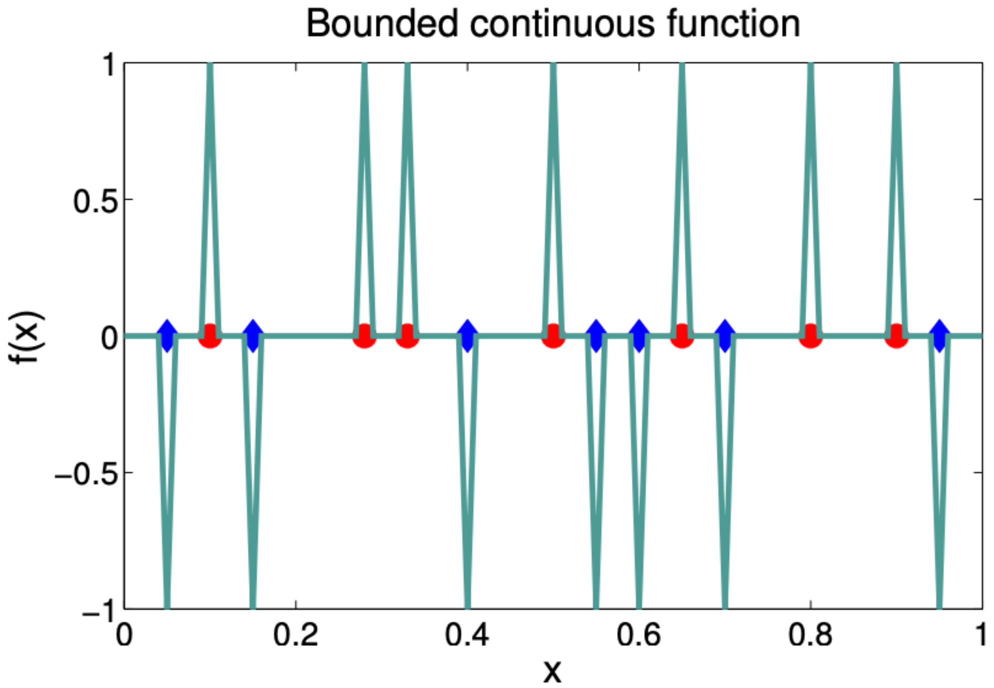




No So Fast

- What happens when f(x) is non-smooth?
- IPM can have arbitrarily small values – even for very different P,Q





Generative Adversarial Models

- Let’s forget the last issue for now
- Assume we are given a set of sample $x_1, x_2 \ldots x_N$ from P
- We also have a generator function G, mapping noise to samples Q
- E.g. if we sample noise vectors $z_1, z_2 .. z_N$ -> $G(z_1), G(z_2) .. G(z_N)$
- How similar are is the distribution of generated and true samples?

Measuring Generative Models 2-Sample Stats

- Idea: measure distance between generated distribution Q(=G(z)) and true distribution P using the 2 sample statistic

$$E_P[d(x)] - E_Z[d(G(z))]$$

Standard GAN Loss

- In the standard GAN loss, the IPM is given by:

$$E_P[\sigma(f(x))] + E_Z[1 - \sigma(f(G(z)))]$$

- The function f is called the discriminator
- Should have high value for real samples, low value for generated
- Lower value of IPM -> harder it is to distringuish generated from real

GAN Training

- Now that we can measure the distance between P and G(Z)
- Can we use this to improve G?
- Can we find a function G such that distribution G(Z) identical to P?
- Optimize G to minimize the IPM!

$$min_G max_f E_P[\sigma(f(x))] + E_Z[1 - \sigma(f(G(z)))]$$

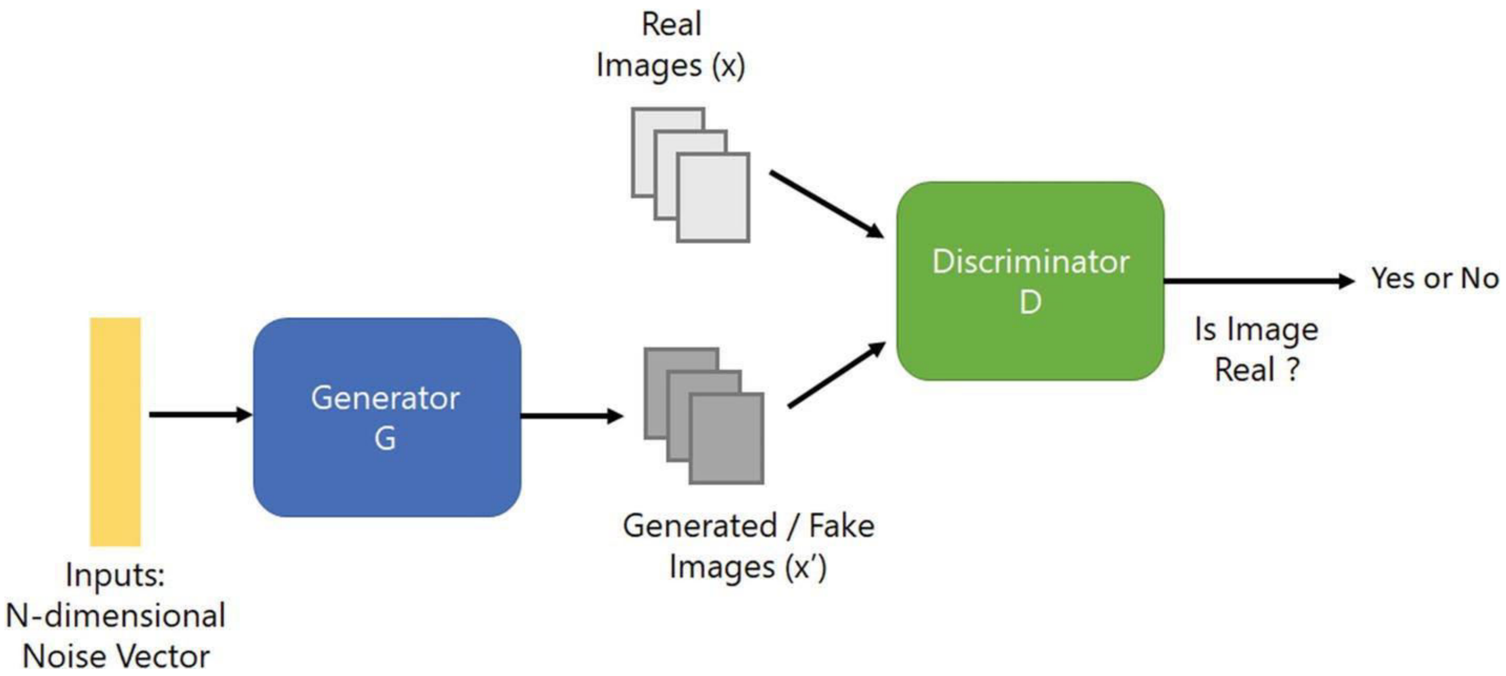
Adversarial Training

- This type of training is called adversarial
- f is trained to maximally discriminate against G
- G is trained to maximally fool f – make it hard to distinguish

$$min_G max_f E_P[\sigma(f(x))] + E_Z[1 - \sigma(f(G(z)))]$$

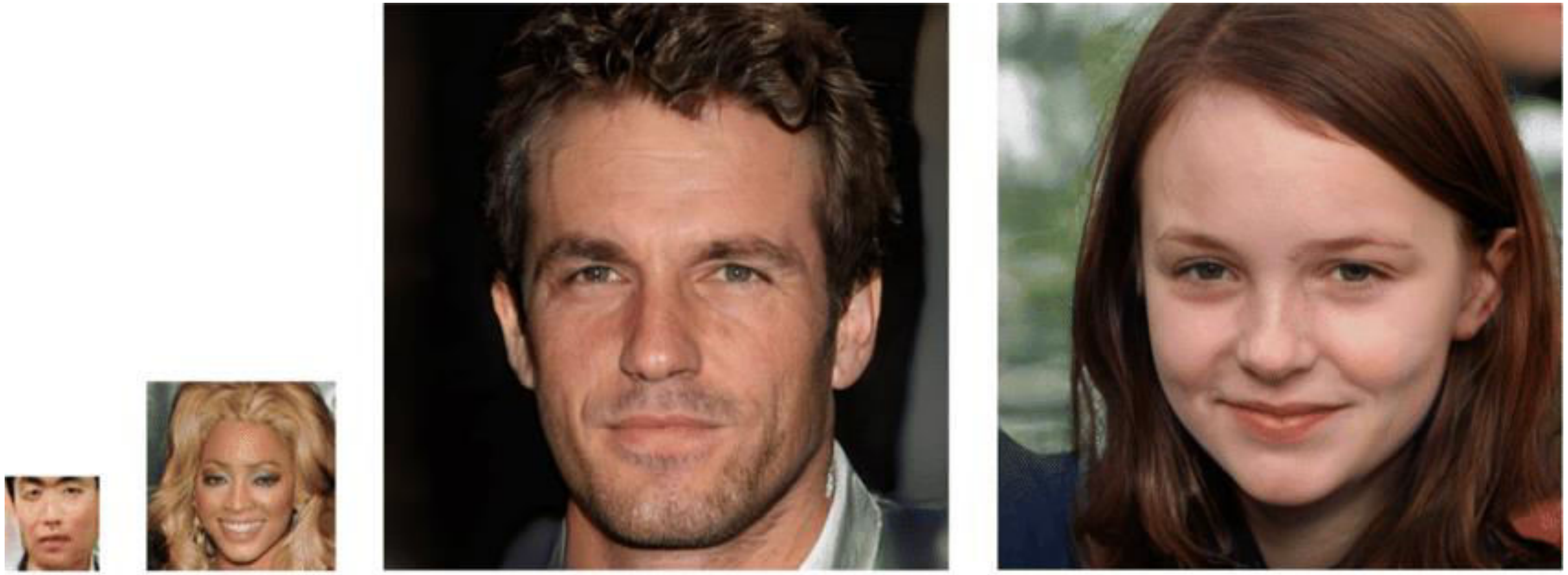
GAN Training for Images

- Generator network – maps noise vector to images
- Discriminator network – maps image to number
  - High for real, low ffor generated



Sampling Using GAN Models

- Sampling is very easy
- Sample random noise vector from Gaussian distribution
- Map to image using generator G



(a) (b) (c) (d)

The Optimal Solution to GAN = JS-Divergence

- The GAN loss is:

$$L(G, f) = \sum_x p(x) \log(f(x)) + q(x) \log(1 - f(x))$$

- Solving for the optimal f at every x, we obtain:

$$f^*(x) = \frac{p(x)}{p(x)+q(x)}$$

- Subtituing back:

$$L(G, f^*) = \sum_X p \log(\frac{p}{p+q}) + q \log(\frac{q}{p+q}) = \sum_X p \log(\frac{2p}{p+q}) + q \log(\frac{2q}{p+q}) - 2log(2)$$

- The JS divergence is given by:

The Optimal Solution to GAN = JS-Divergence

- The JS-divergence is given by:

$$\frac{1}{2}\int p(x)\log\frac{2p(x)}{p(x)+q(x)} + q(x)\log\frac{2q(x)}{p(x)+q(x)}\,\mathrm{d}x$$

- Our expression for the GAN loss:

$$L(G,f^*) = \sum_X p\log(\frac{p}{p+q}) + q\log(\frac{q}{p+q}) = \sum_X p\log(\frac{2p}{p+q}) + q\log(\frac{2q}{p+q}) - 2log(2)$$

- Can simply be written as:

$$L(G,f^*) = 2JS(p||q) - 2\log(2)$$

f-GAN: IPM Lower Bound the f-Divergence

- It was shown that for divergence

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x,$$

- Is lower bounded by the IPM:

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}\left[g_f(V_\omega(x))\right] + \mathbb{E}_{x \sim Q_\theta}\left[-f^*(g_f(V_\omega(x)))\right],$$

- Where f* is the Fenchel conjugate of f (don’t worry what this means)

Example of f-GANs

- Here are the GAN equivalent of some famous divergences
  - Standard GAN is also one of them

$$F(\theta, \omega) = \mathbb{E}_{x \sim P}\left[g_f(V_\omega(x))\right] + \mathbb{E}_{x \sim Q_\theta}\left[-f^*(g_f(V_\omega(x)))\right],$$

Name	$D_f(P\|Q)$	Generator $f(u)$
Kullback-Leibler	$\int p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$	$u \log u$
Reverse KL	$\int q(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$	$-\log u$
Pearson χ^2	$\int \frac{(q(x)-p(x))^2}{p(x)} \, \mathrm{d}x$	$(u-1)^2$
Squared Hellinger	$\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x$	$(\sqrt{u}-1)^2$
Jensen-Shannon	$\frac{1}{2}\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x$	$-(u+1)\log \frac{1+u}{2} + u \log u$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} \, \mathrm{d}x - \log(4)$	$u \log u - (u+1)\log(u+1)$

Output activation g_f	dom_{f^*}	Conjugate $f^*(t)$
v	$\mathbb{R}$	$\exp(t-1)$
$-\exp(-v)$	$\mathbb{R}_-$	$-1-\log(-t)$
v	$\mathbb{R}$	$\frac{1}{4}t^2 + t$
$1-\exp(-v)$	$t < 1$	$\frac{t}{1-t}$
$\log(2) - \log(1+\exp(-v))$	$t < \log(2)$	$-\log(2-\exp(t))$
$-\log(1+\exp(-v))$	$\mathbb{R}_-$	$-\log(1-\exp(t))$

Probability Distance Measures

- The *Total Variation* (TV) distance

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)| \ .$$

- The *Kullback-Leibler* (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log\left(\frac{P_r(x)}{P_g(x)}\right) P_r(x) d\mu(x) \ ,$$

- The *Jensen-Shannon* (JS) divergence

$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m) \ ,$$

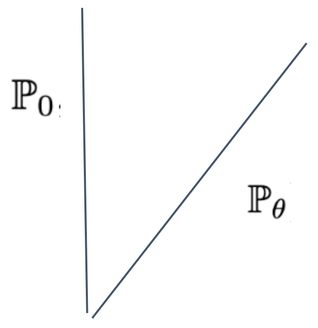
where $\mathbb{P}_m$ is the mixture $(\mathbb{P}_r + \mathbb{P}_g)/2$. This divergence is symmetrical and always defined because we can choose $\mu = \mathbb{P}_m$.

- The *Earth-Mover* (EM) distance or Wasserstein-1

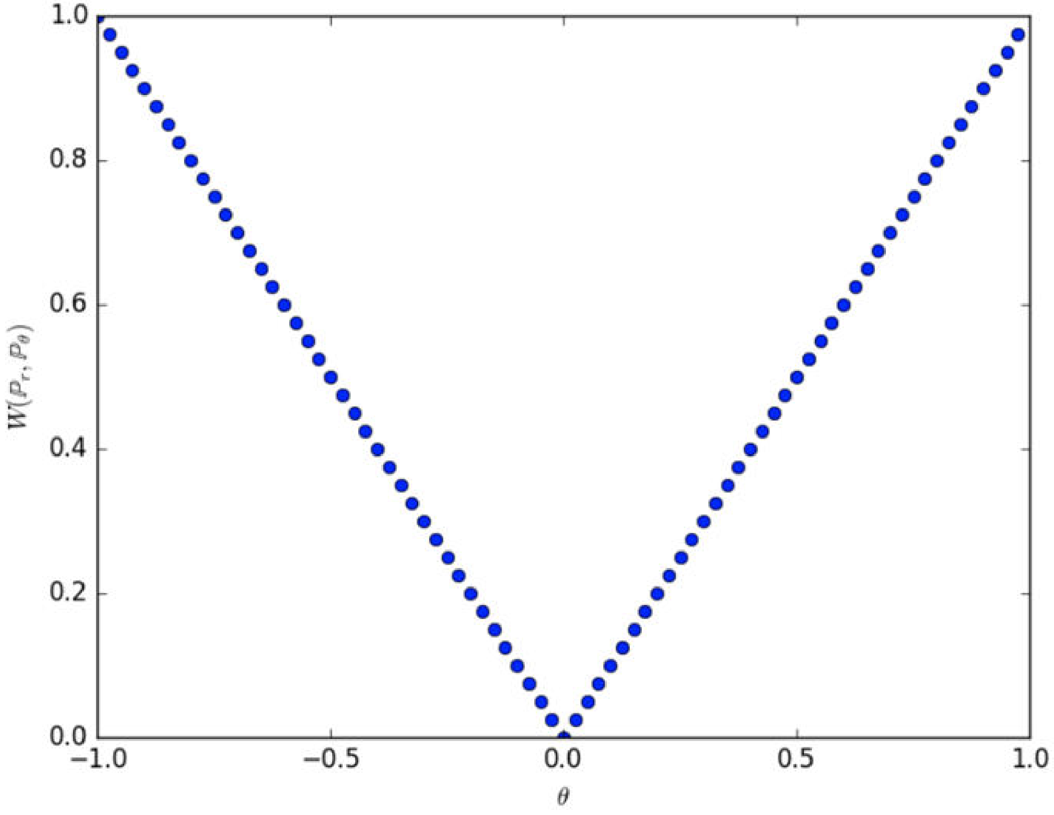
$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} \big[\|x - y\| \big] \ , \tag{1}$$

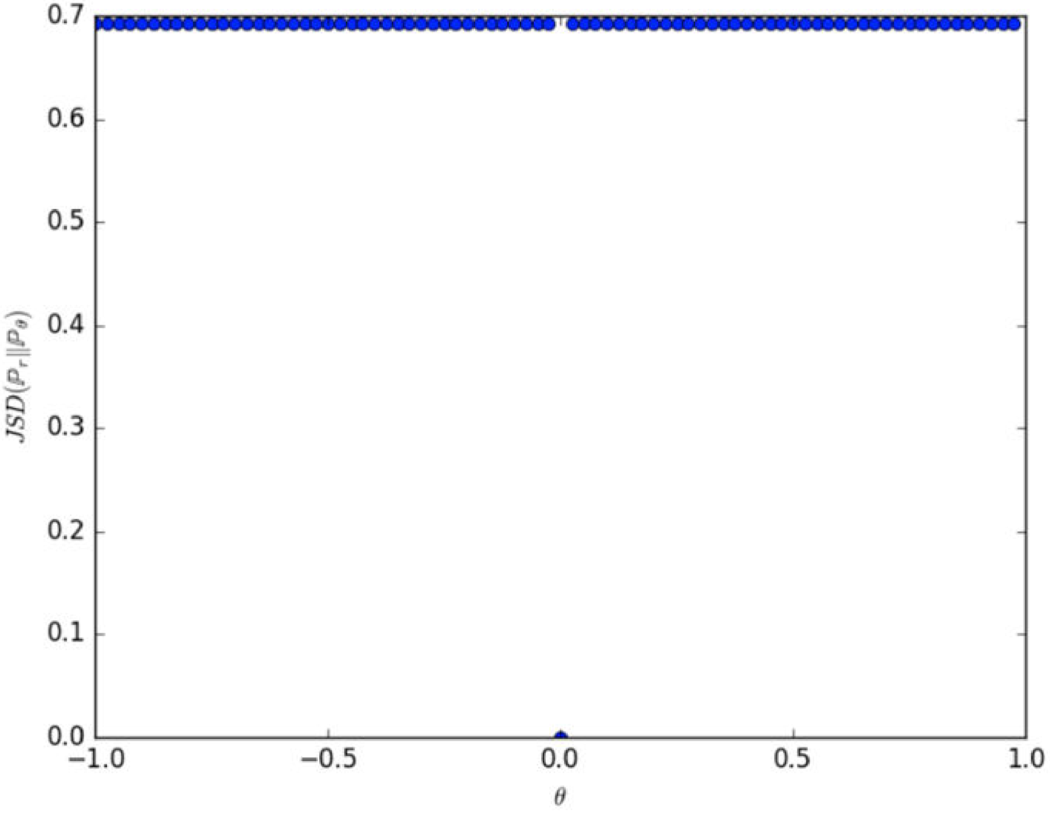
Different distance measures on toy example

- $W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|,$
- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ . \end{cases}$



Different distance measures on toy example (2)





What Does All This Mean?

- The short story – GANs diverge

$$min_G max_f E_P[\sigma(f(x))] + E_Z[1 - \sigma(f(G(z)))]$$

- If P=Q, then no f can distinguish between them
- But this only happens at infinity
- For all case where P!=Q:
  - Choose f(x)=1 for all samples for P (which are finite)
  - Choose f(x)=0 for all other samples
- The above loss will have constant values for many f, gradient 0

Other Divergences

- The *Total Variation* (TV) distance

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)| \ .$$

- The *Kullback-Leibler* (KL) divergence

$$KL(\mathbb{P}_r \| \mathbb{P}_g) = \int \log \left(\frac{P_r(x)}{P_g(x)} \right) P_r(x) d\mu(x) \ ,$$

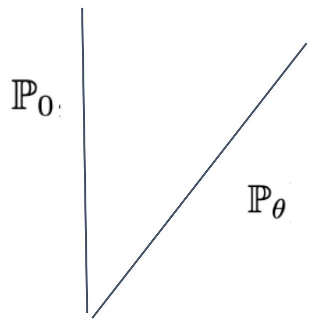
- The *Jensen-Shannon* (JS) divergence

$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \| \mathbb{P}_m) + KL(\mathbb{P}_g \| \mathbb{P}_m) \ ,$$

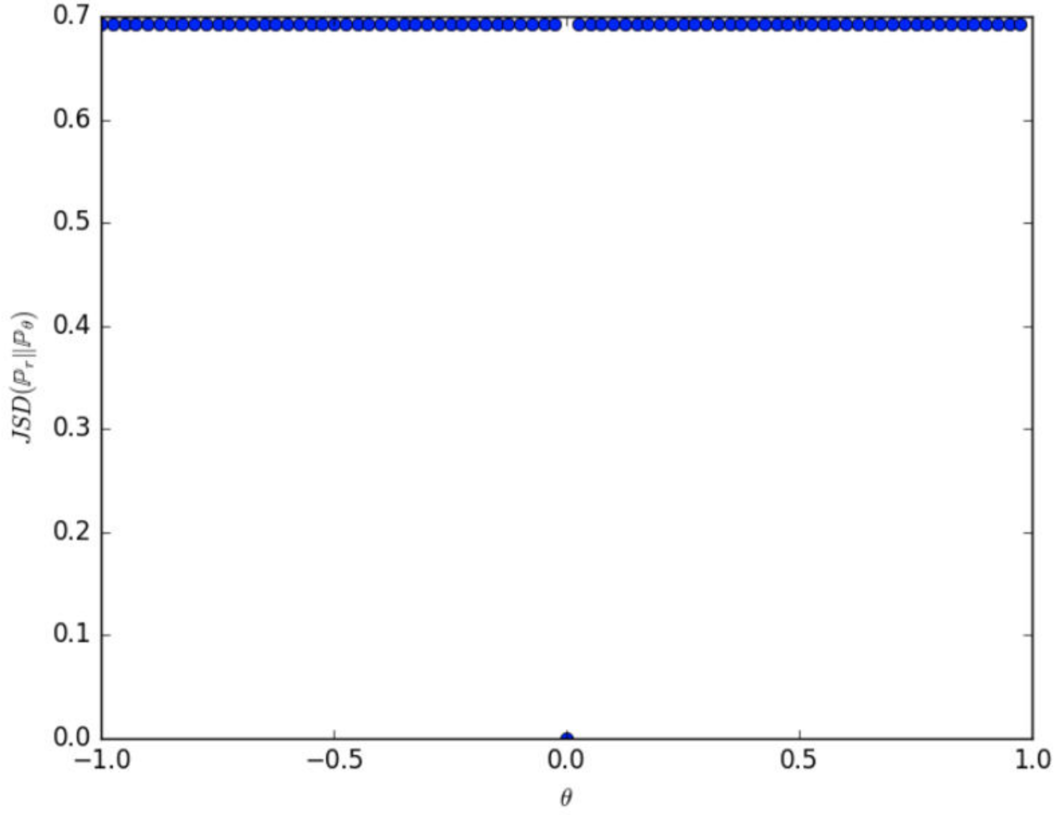
where $\mathbb{P}_m$ is the mixture $(\mathbb{P}_r + \mathbb{P}_g)/2$. This divergence is symmetrical and always defined because we can choose $\mu = \mathbb{P}_m$.

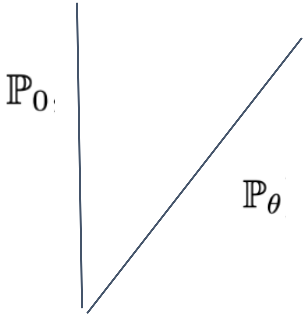
Different divergences on toy example

- $JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- $KL(\mathbb{P}_\theta \| \mathbb{P}_0) = KL(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ , \end{cases}$
- and $\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \ , \\ 0 & \text{if } \theta = 0 \ . \end{cases}$



Different distance measures on toy example (2)





We Need a More Expressive Measure

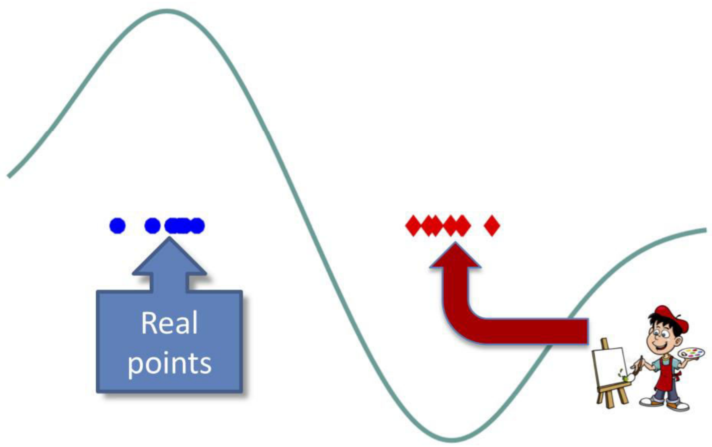
- The above shows that exisiting diverences are not expressive enough
- They fail when Q and P have different supports
- They do not vary smoothly
- Need better ideas!

Idea: Enforce Smoothly Varying Function

- The key idea: enforce the discriminator function is smooth
- This way it will provide gradient at all points
- Mathematically: Lipschitzness is a measure of smoothness

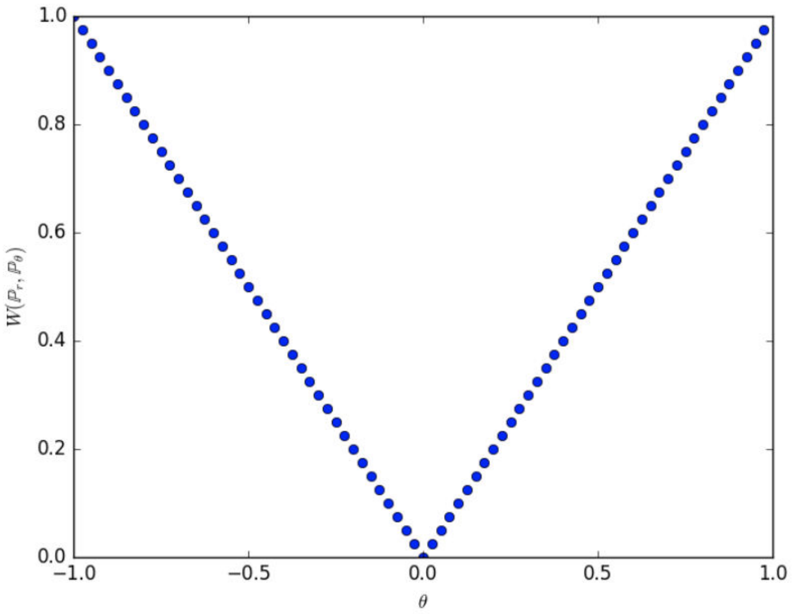
$$W_1(P, Q) = \sup_{||f||_L \le 1} E_P f(X) - E_Q f(Y).$$

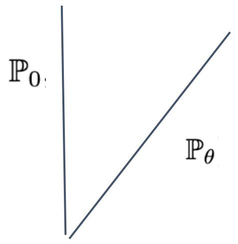
$$||f||_L := \sup_{x \neq y} |f(x) - f(y)| / ||x - y||$$



Back to the Toy Example

- The W-1 distance provides a smooth gradient





How to enforce Lipschitz-1 function f

WGAN: weight clipping $\mathrm{clip}(w, -c, c)$

WGAN-GP: $\mathrm{E}_{p_{\mathcal{D}}(x)}\left[|D_\psi(x)|^2 + \|\nabla_x D_\psi(x)\|^2\right]$

Mescheder et al.: $R_1(\psi) := \frac{\gamma}{2}\mathrm{E}_{p_{\mathcal{D}}(x)}\left[\|\nabla D_\psi(x)\|^2\right]$

Zhang et al.: $$\min_D L_{cr} = \min_D \sum_{j=m}^{n} \lambda_j \left\|D_j(x) - D_j(T(x))\right\|^2,$$

Spectral Normalization

Ensure that every layer in the network is Lip-1

Fast method for speeding up eigenvalue computation based on power method

$$\sigma(A) := \max_{\boldsymbol{h}:\boldsymbol{h}\neq 0} \frac{\|A\boldsymbol{h}\|_2}{\|\boldsymbol{h}\|_2} = \max_{\|\boldsymbol{h}\|_2 \leq 1} \|A\boldsymbol{h}\|_2,$$

$$\bar{W}_{\mathrm{SN}}(W) := W/\sigma(W).$$

The Wasserstein GAN

Finally this can be seen as an adversarial task:

$$\max_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]$$

$$\nabla_\theta W(\mathbb{P}_r, \mathbb{P}_\theta) = -\mathbb{E}_{z\sim p(z)}[\nabla_\theta f(g_\theta(z))]$$

WGAN algorithm

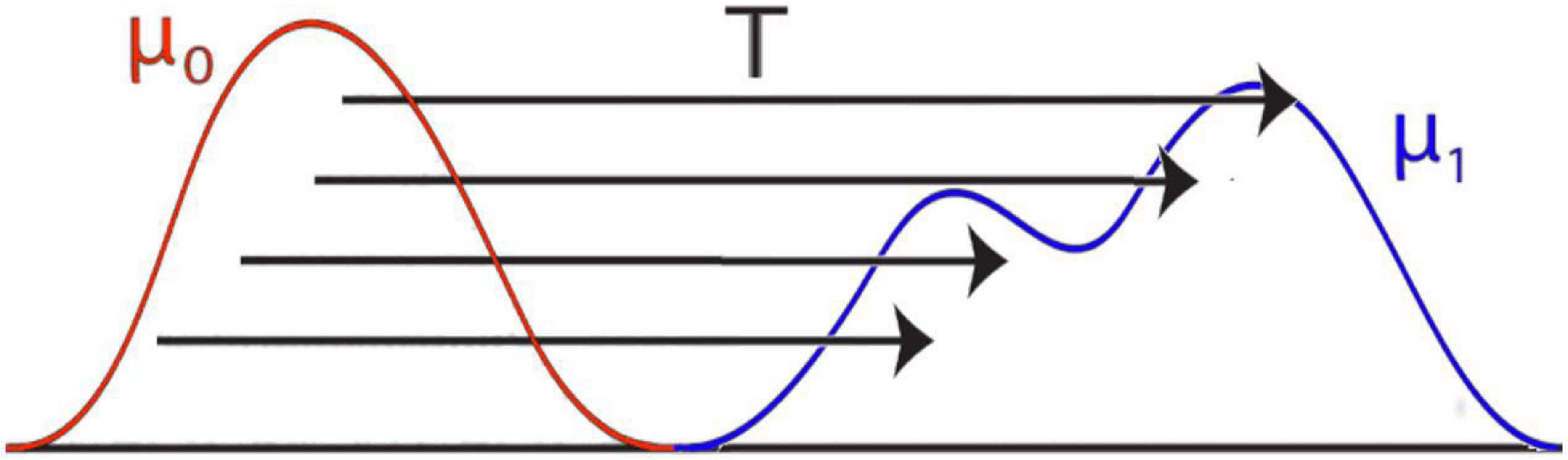
Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
    for $t = 0, ..., n_{\text{critic}}$ **do**
        Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
        Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
        $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
        $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
        $w \leftarrow \text{clip}(w, -c, c)$
    end for
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
    $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

Optimal Transport

- The L1-IPM is the dual of the Wasserstein distance
- The Wasserstein distance provides optimal matching between samples
- In 1-D is is called the Earth Mover's Distance

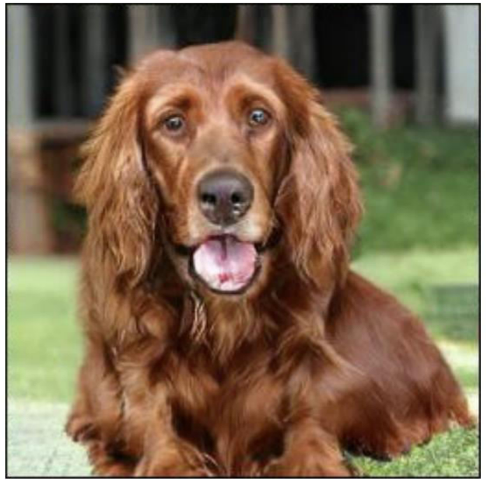


N. Papadakis, Optimal Transport for Image Processing, habilitation à diriger des recherches, Université de Bordeaux, Dec. 2015

BigGAN

Large Scale GAN Training for High Fidelity Natural Image Synthesis
Andrew Brock, Jeff Donahue, Karen Simonyan - DeepMind

- Objective: GAN image generation at the ImageNet scale
- A tour-de-force of GAN tricks – not one idea




Visual Results

(a) 128×128

(b) 256×256

(c) 512×512

(d)

StyleGAN

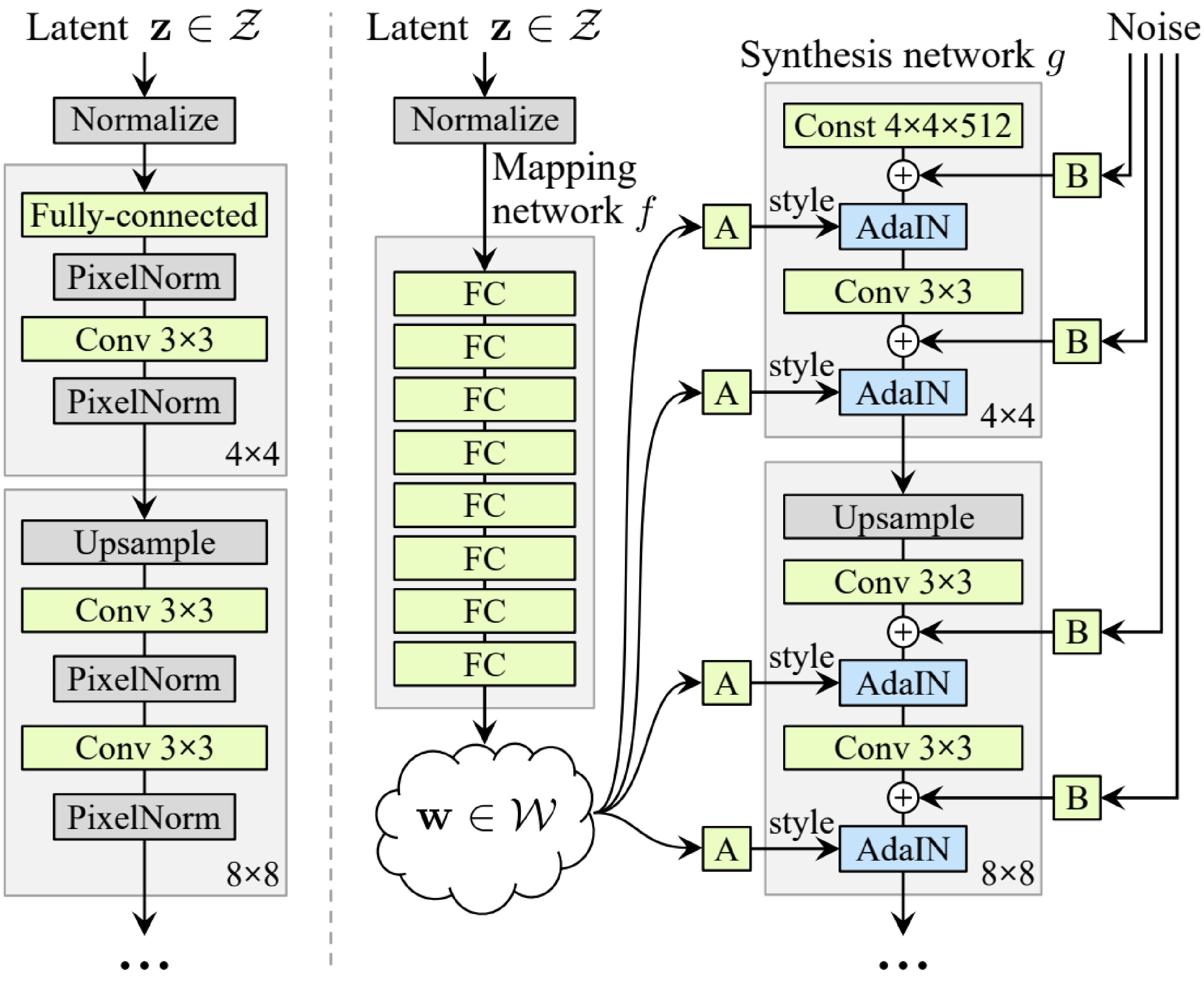
A Style-Based Generator Architecture for Generative Adversarial Networks

Tero Karras, Samuli Laine, Timo Aila - NVIDIA

- Objective: model the distribution of a **single** class
- Main idea: inject *style* code at every level of the generator

Architecture

- Noise to per-layer style code
- Contrast with single code in DCGAN
- Code injected via AdaIN



(a) Traditional (b) Style-based generator

Effective Disentanglement Between Layers

- Codes of different layers are responsible for different scales of attributes