Wasserstein GAN

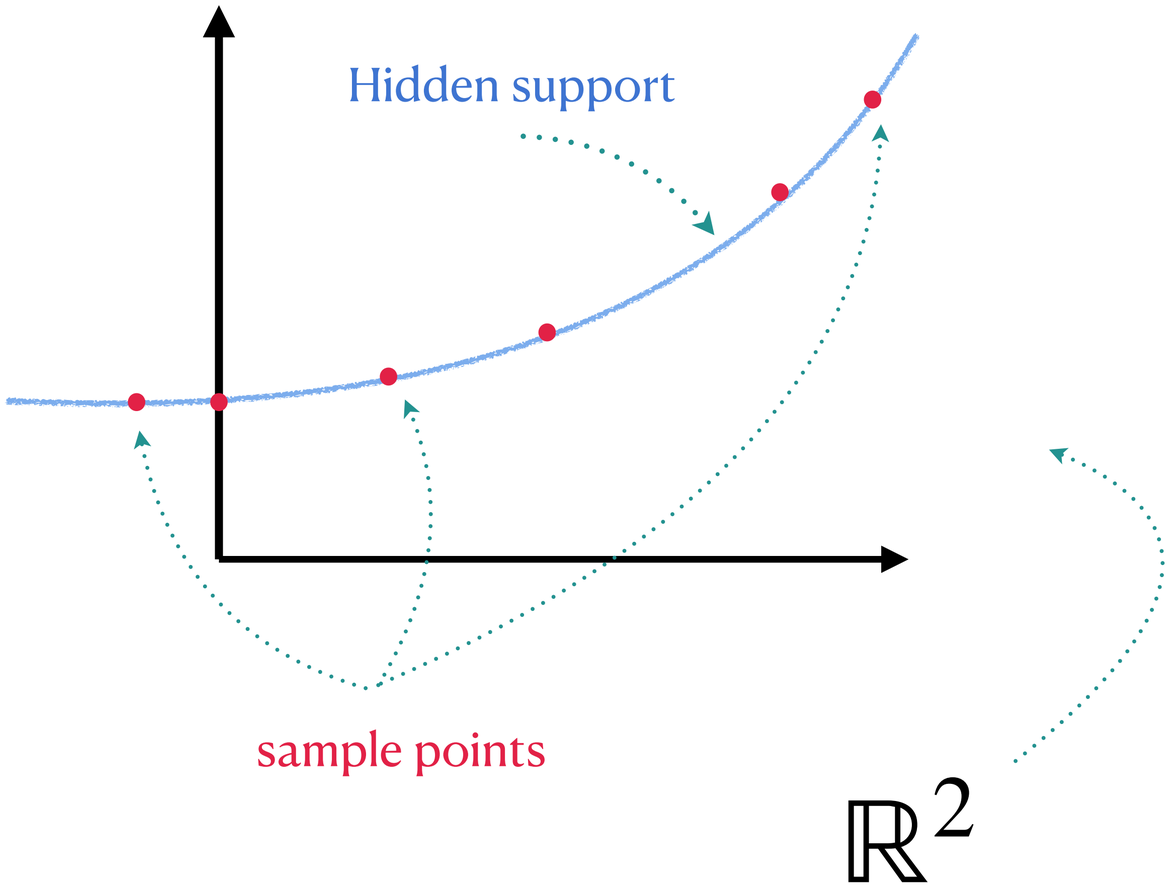
Martin Arjovsky , Soumith Chintala , and Leon Bottou

Presented by: Aditya Kumar Akash and Peyman Morteza

Part I: Motivation and theoretical aspects of WGAN

Unsupervised Learning Problem

- A sample data points from a hidden support is given.
- How to develop an algorithm to “learn” the hidden structure?



Unsupervised Learning Problem

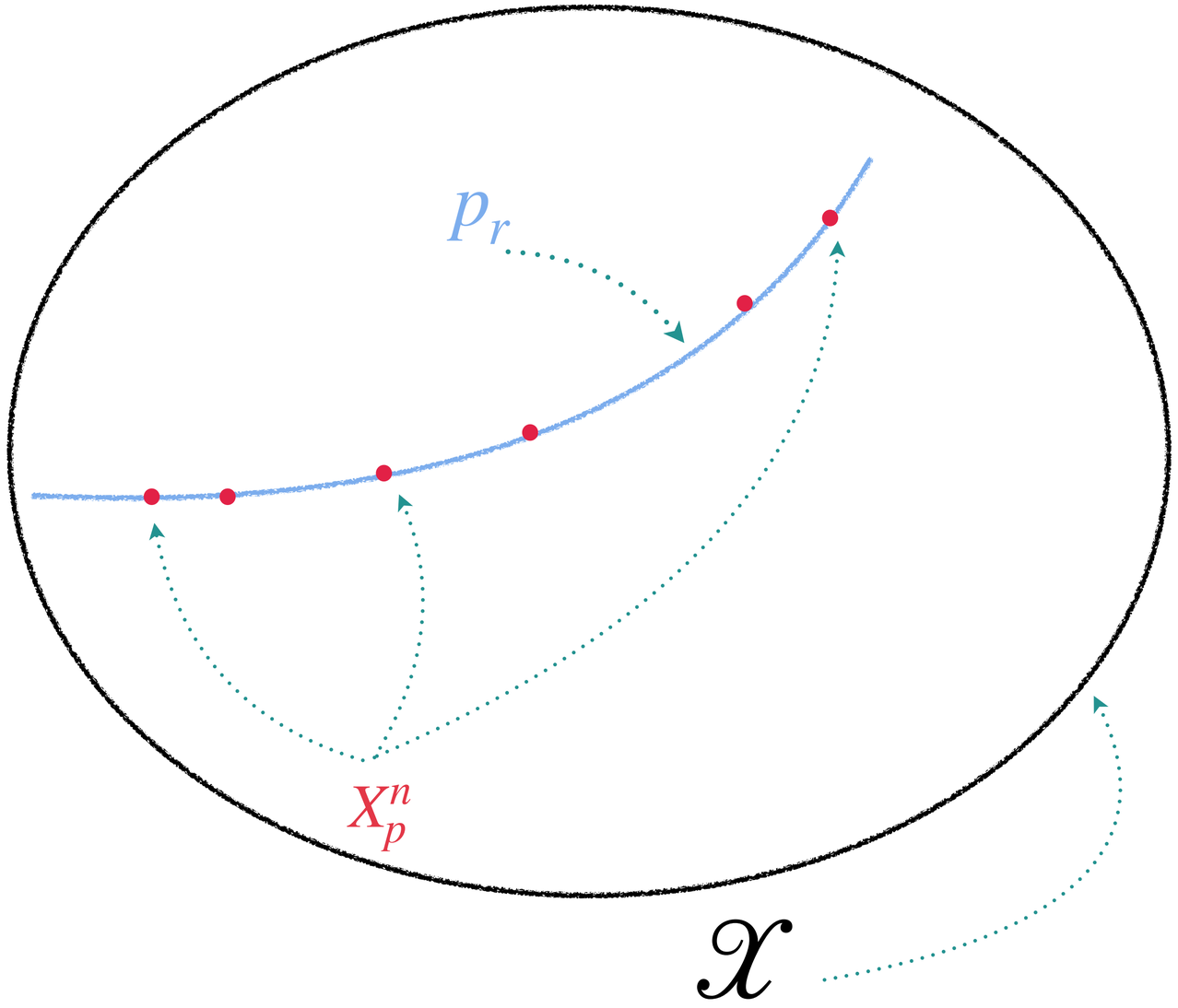
- What if the underlying data points are in high dimension and the underlying structure has a complicated support?
- Why to study complexified version of the problem?

Image Data

Music Data

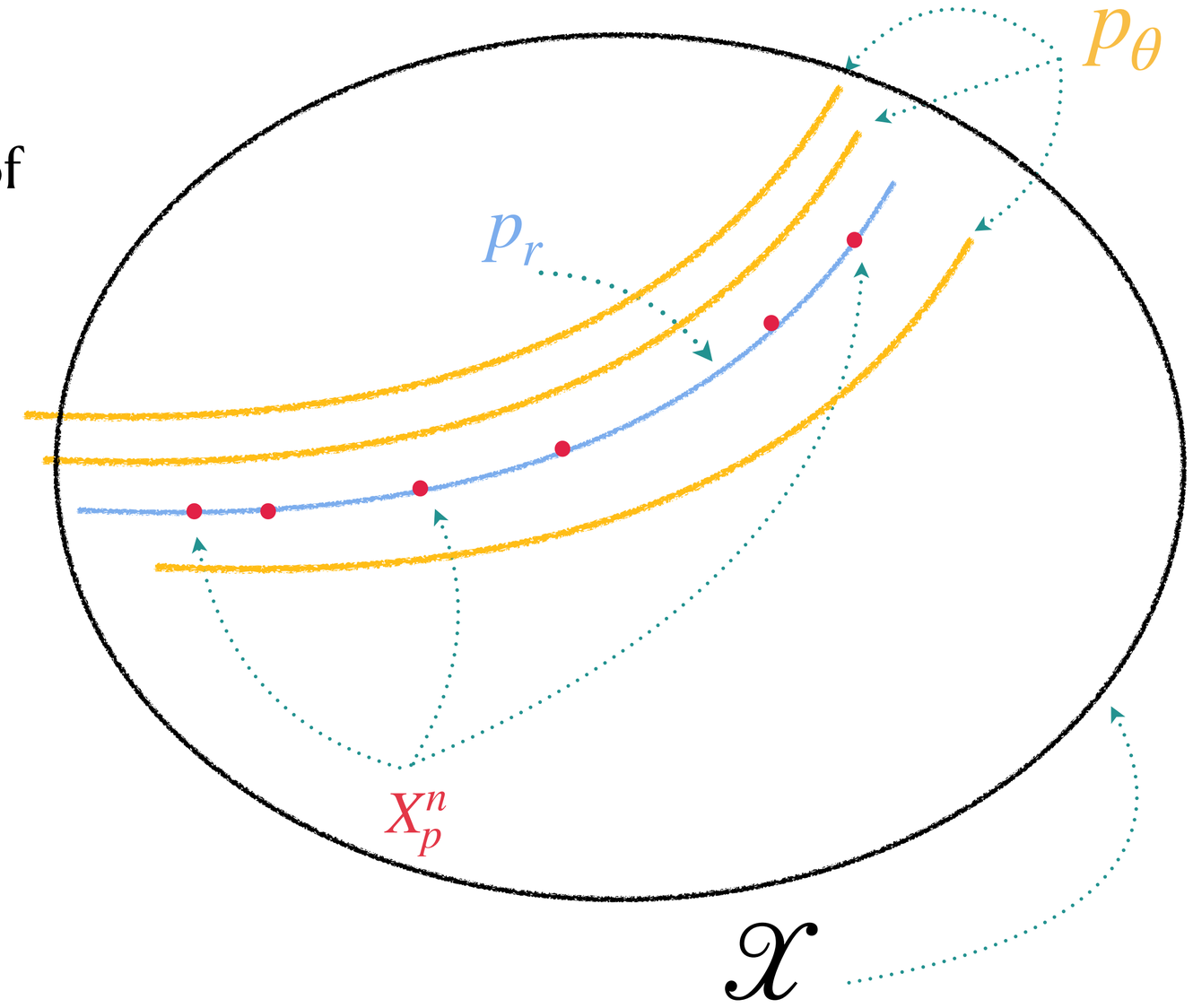
Problem formulation

- Let $\mathcal{X}$ be a compact metric space and let Σ denotes the Borel sets of $\mathcal{X}$ and P denotes the set of all probability measures on $(\mathcal{X}, \Sigma)$.
- Let $p_r \in P$ be a hidden probability measure that generates the real data.
- Let $X_p^n = \{x_1^p, \ldots, x_n^p\}$ be a given i.i.d sample generated by p_r.



Problem formulation

- $p_\theta, \theta \in \Theta$: a parametrized family of probability distributions.
- $\theta^* := argmin_{\theta \in \Theta}\big(d(p_r, p_\theta)\big)$
- What is a good candidate for d?



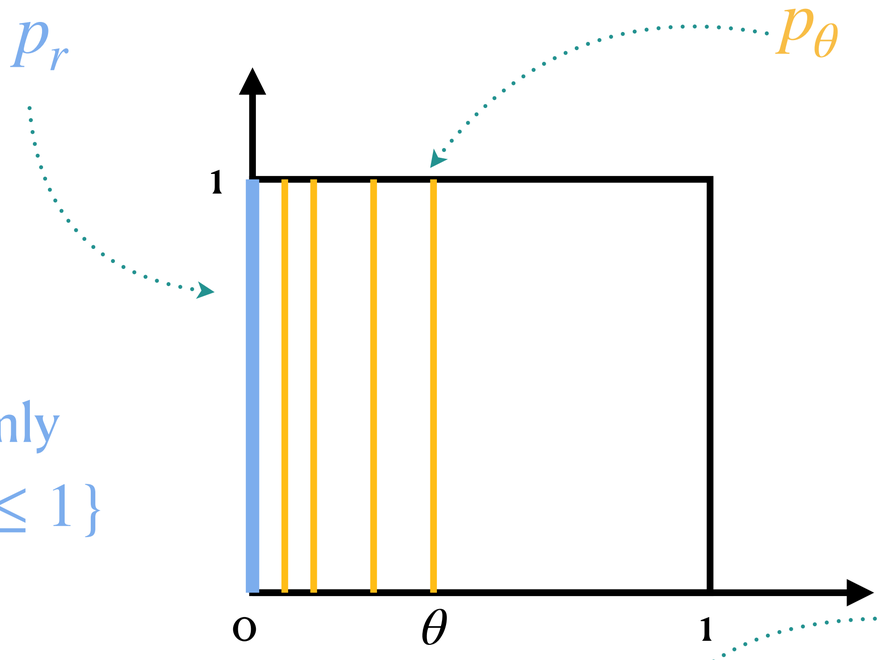
Motivating Example

Looking for "metric" d such that

$d(p_\theta, p_r) \to 0$ as $\theta \to 0$

p_r : measure on $[0,1]^2$ that is uniformly distributed on $\{(0,y) \in [0,1]^2 \mid 0 \le y \le 1\}$

p_θ : measure on $[0,1]^2$ that is uniformly distributed on $\{(\theta, y) \in [0,1]^2 \mid 0 \le y \le 1\}$



Equip with uniform measure μ

$\mathcal{X} = [0,1] \times [0,1]$

$\Theta = [0,1]$

$g_\theta : [0,1] \to [0,1] \times [0,1]$

$g_\theta(x) = (\theta, x)$

$p_\theta := (g_\theta)_* \mu$

“Distance” functions on P

$$KL(p||q) := \int_{\mathcal{X}} \log(\frac{f_p(x)}{f_q(x)}) f_p(x) dx$$

Kullback–Leibler divergence:
Shows up as asymptotic of MLE

$$JS(p,q) := KL(p, \frac{p+q}{2}) + KL(q, \frac{p+q}{2})$$

Jensen–Shannon divergence:
Symmetrized version of KL

$$\delta(p,q) := \sup_{K \in \Sigma} |p(K) - q(K)|$$

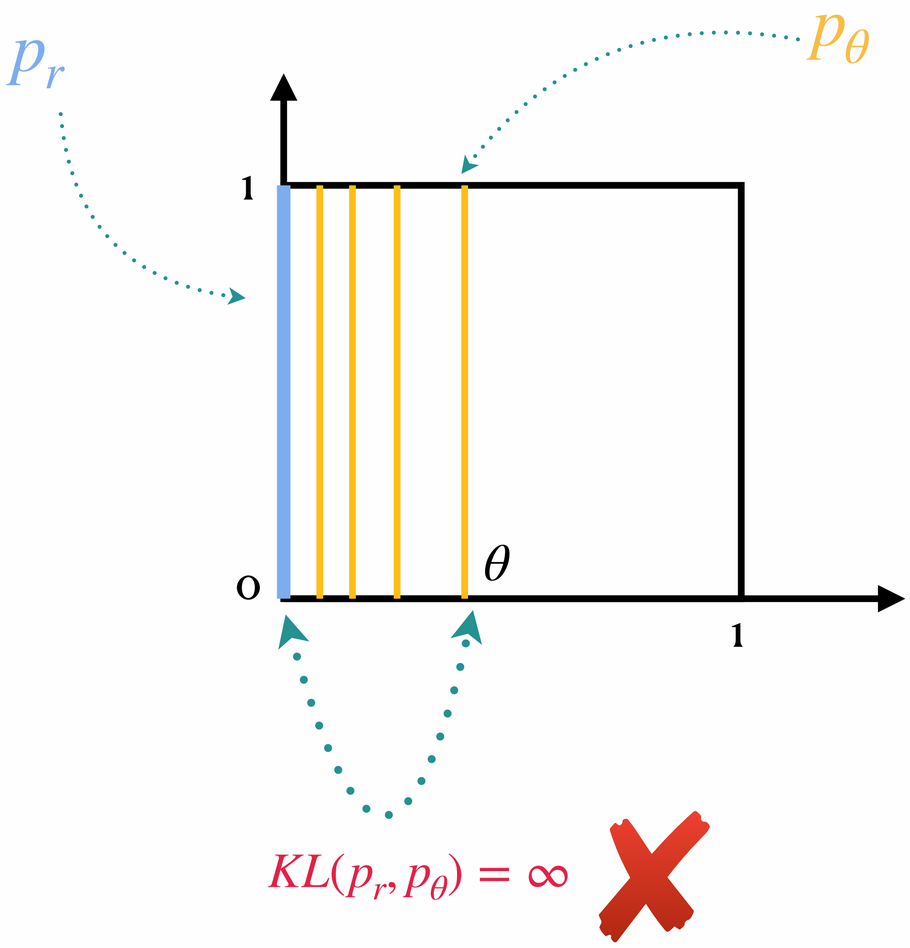
Total variation between two measures

KL Divergence

Looking for a "metric" , d, such that $d(p_\theta, p_r) \to 0$ as $\theta \to 0$

If $\theta > 0 \implies p_r$ and p_θ have disjoint support

$\implies KL(p_r, p_\theta) = \infty \implies KL$ is not even defined in this particular example.



p_r : measure on $[0,1]^2$ that is uniformly distributed on $\{(0,y) \in [0,1]^2 \mid 0 \le y \le 1\}$

p_θ : measure on $[0,1]^2$ that is uniformly distributed on $\{(\theta, y) \in [0,1]^2 \mid 0 \le y \le 1\}$

JS Divergence

Looking for a "metric" , d, such that $d(p_\theta, p_r) \to 0$ as $\theta \to 0$

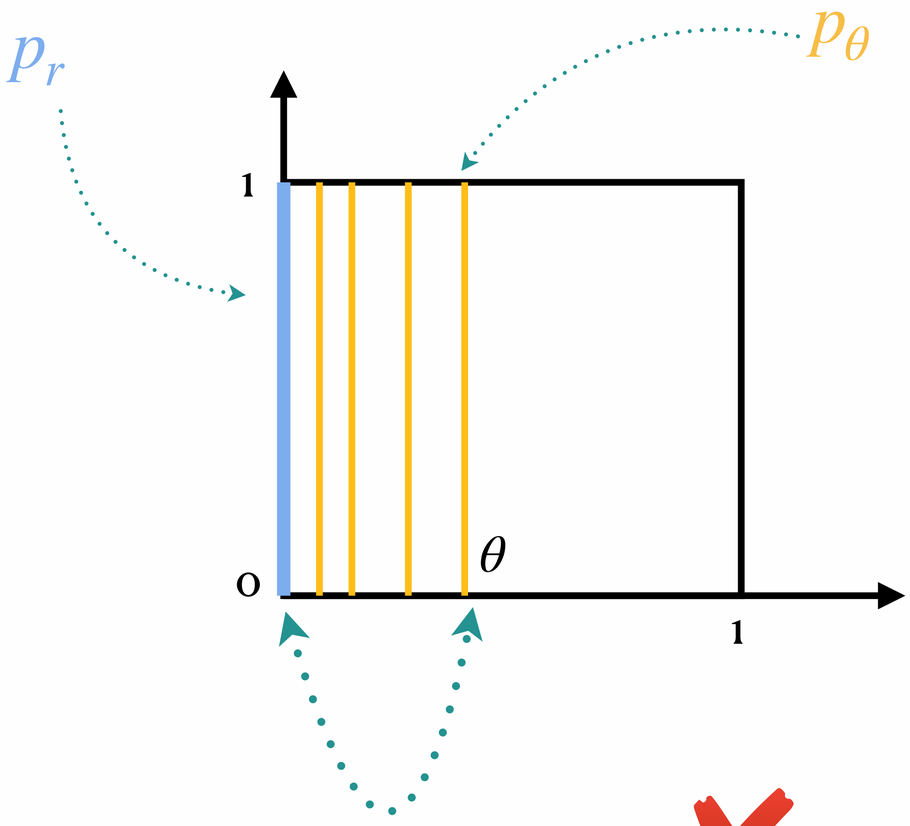
If $\theta > 0 \implies \frac{p_r + p_\theta}{2}$ has mass $\frac{1}{2}$ on $y = 0, y = \theta$

$$\implies JS(p_r, p_\theta) = \frac{1}{2}\log(2) + \frac{1}{2}\log(2) = \log(2)$$

$\implies$ JS is constant and does not change by θ.

p_r : measure on $[0,1]^2$ that is uniformly distributed on $\{(0,y) \in [0,1]^2 \mid 0 \le y \le 1\}$

p_θ : measure on $[0,1]^2$ that is uniformly distributed on $\{(\theta, y) \in [0,1]^2 \mid 0 \le y \le 1\}$



JS does not have an informative gradient!

Total Variation

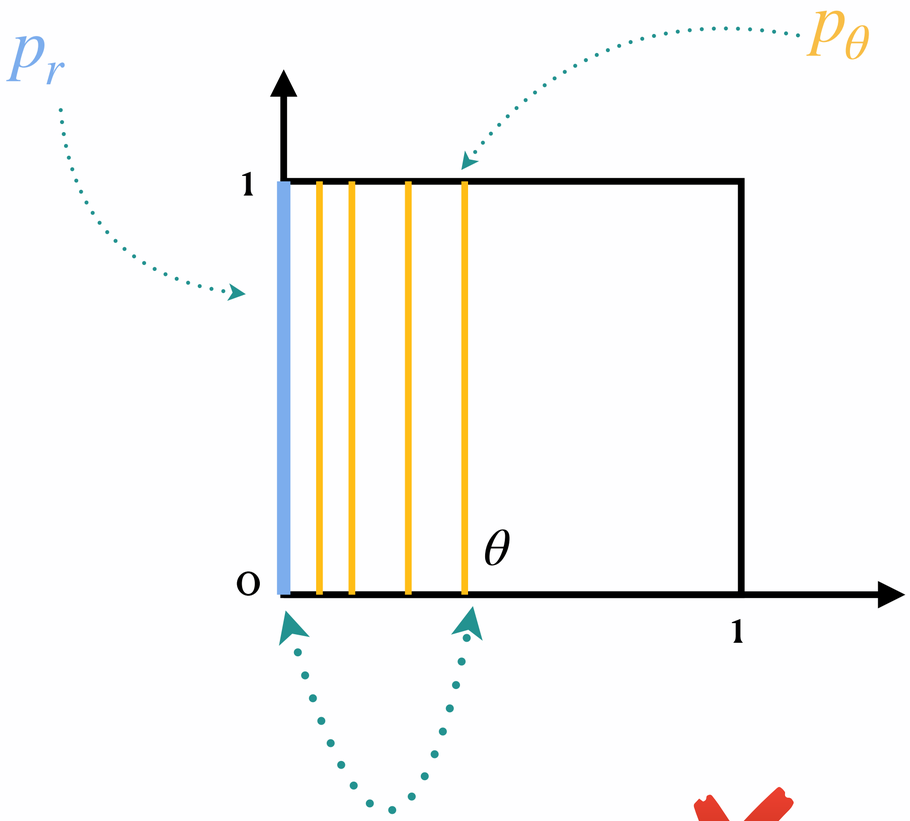
Looking for a "metric" , d, such that $d(p_\theta, p_r) \to 0$ as $\theta \to 0$

If $\theta > 0 \implies \delta(p_r, p_\theta) = 1$ is constant and does not change by θ.

If $\theta = 0 \implies \delta(p_r, p_\theta) = 0$.

p_r : measure on $[0,1]^2$ that is uniformly distributed on $\{(0,y) \in [0,1]^2 \mid 0 \le y \le 1\}$

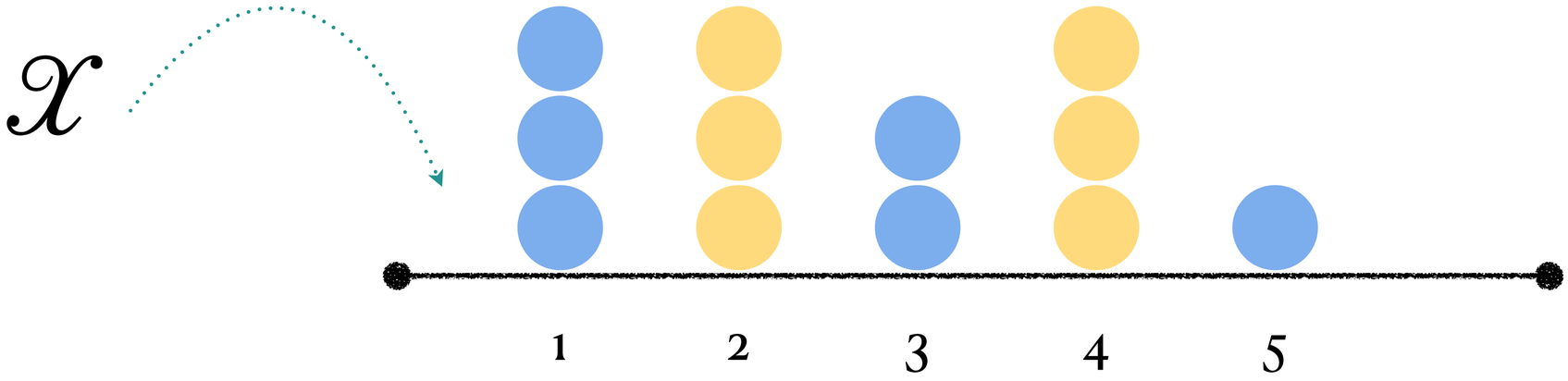
p_θ : measure on $[0,1]^2$ that is uniformly distributed on $\{(\theta, y) \in [0,1]^2 \mid 0 \le y \le 1\}$



$\delta(p_r, p_\theta) \equiv 1$

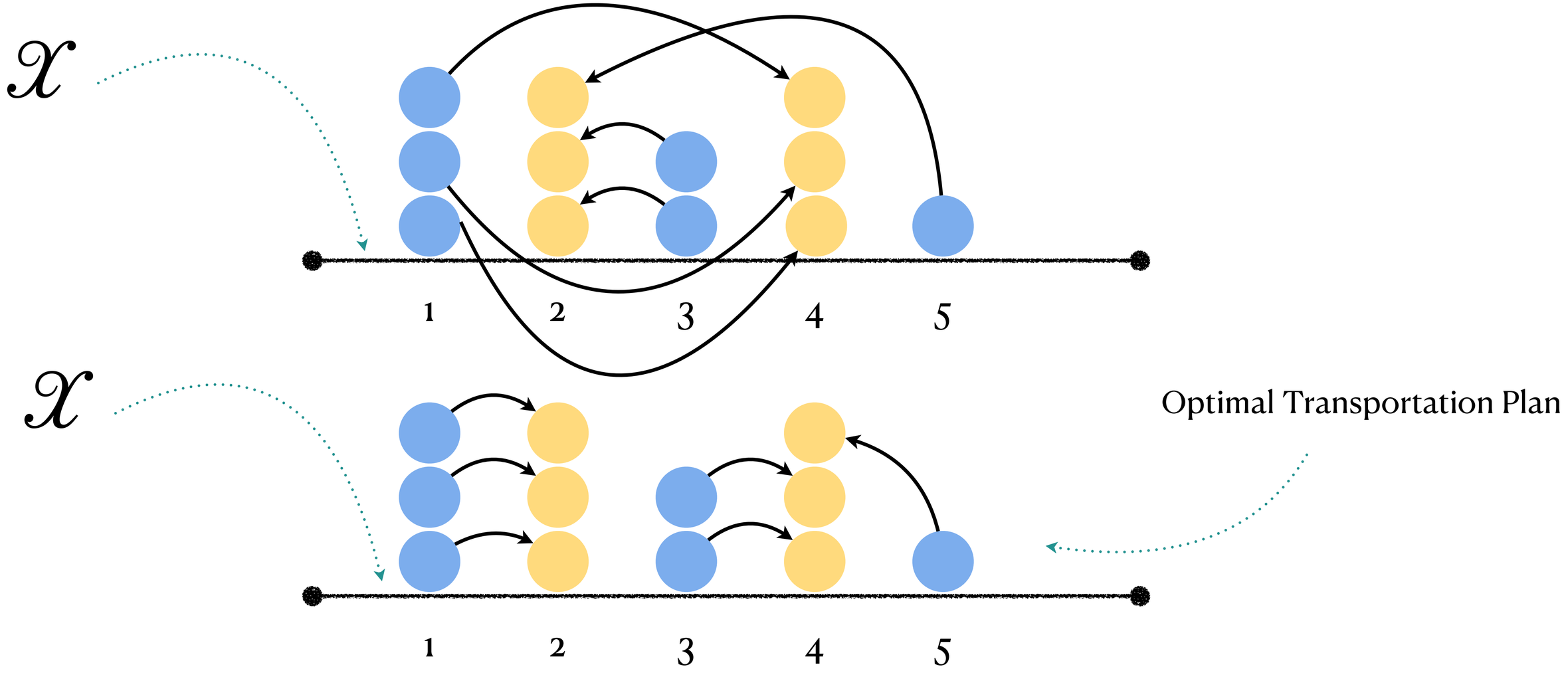
Total variation does not have an informative gradient!

Wasserstein Distance

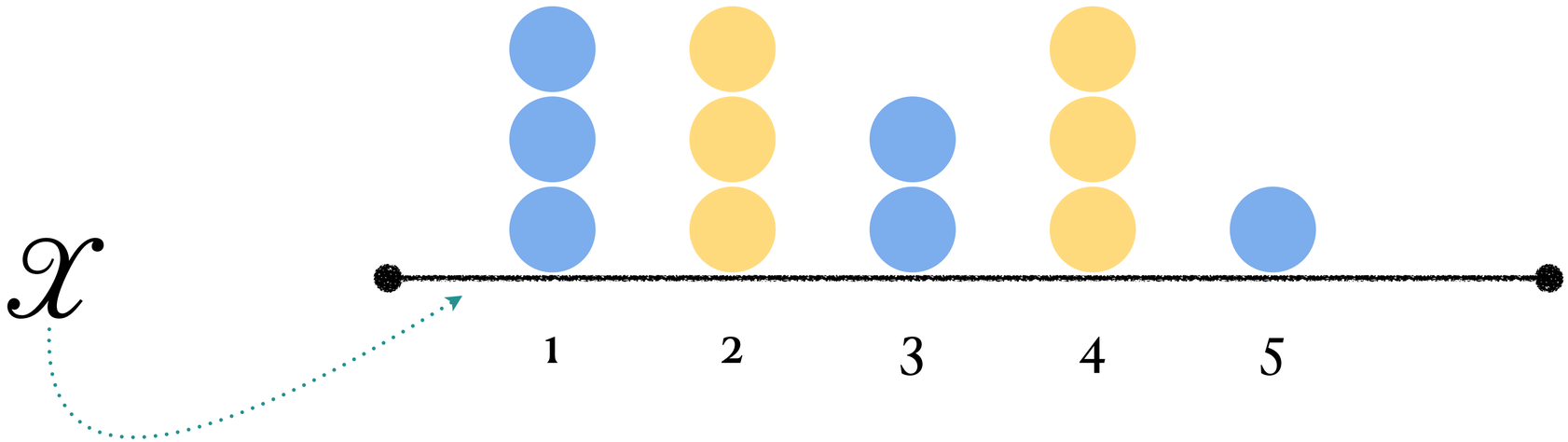


- Transport each blue ball and match it to a yellow ball.
- $c(x, y) = |x - y|$ cost of moving a ball located at position x to position y.
- Find a transportation plan with minimal cost.

Wasserstein Distance



Wasserstein Distance

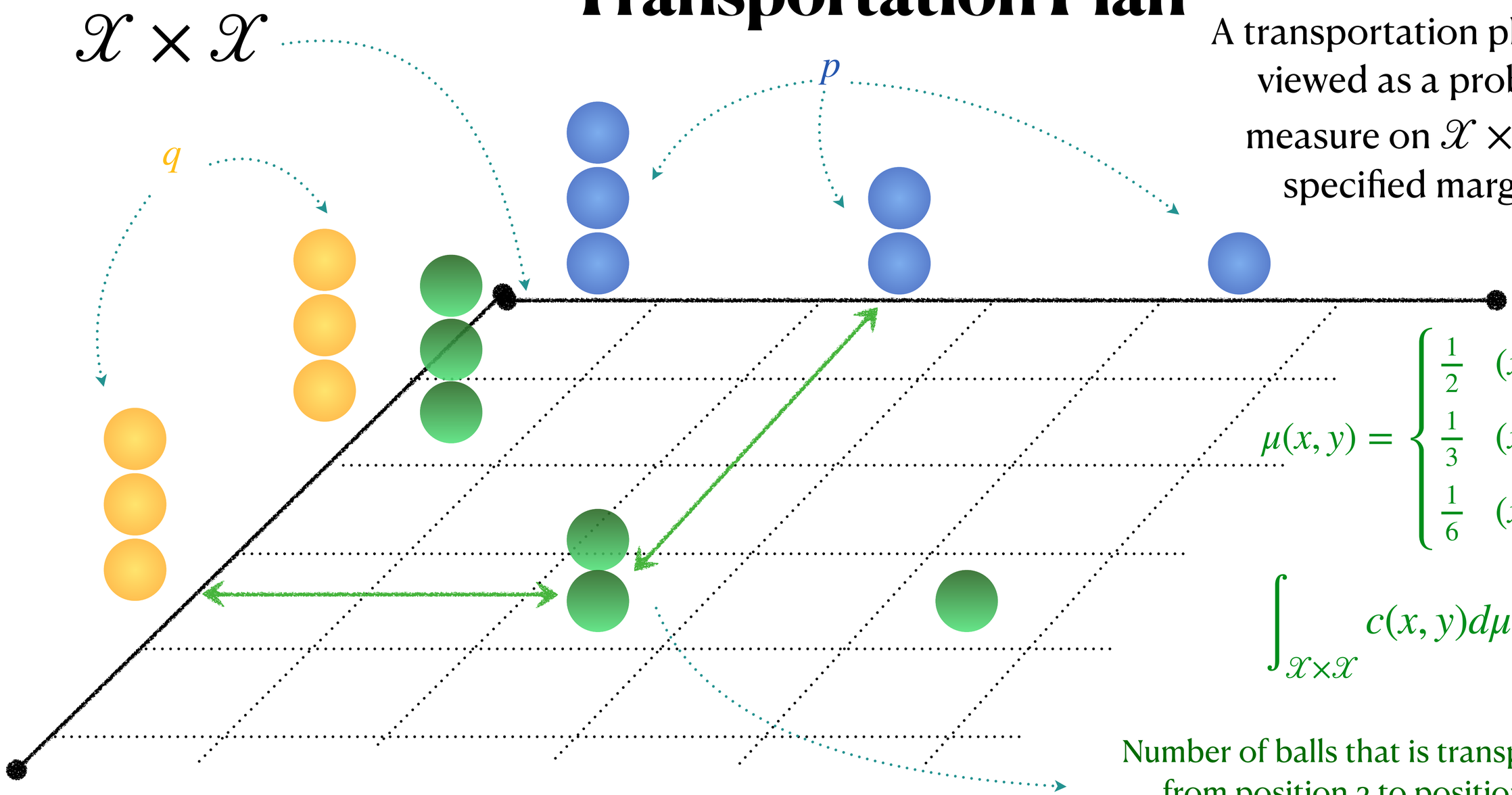


$$p(x) = \begin{cases} \frac{1}{2} & x = 1 \\ \frac{1}{3} & x = 3, \\ \frac{1}{6} & x = 5 \end{cases}$$

$$q(x) = \begin{cases} \frac{1}{2} & x = 2 \\ \frac{1}{2} & x = 4 \end{cases},$$

Each distribution of balls can be viewed as a probability measure on $\mathcal{X}$

Transportation Plan



A transportation plan can be viewed as a probability measure on $\mathcal{X} \times \mathcal{X}$ with specified marginals.

$$\mu(x,y) = \begin{cases} \frac{1}{2} & (x,y) = (1,2) \\ \frac{1}{3} & (x,y) = (3,4), \\ \frac{1}{6} & (x,y) = (5,4) \end{cases}$$

$$\int_{\mathcal{X}\times\mathcal{X}} c(x,y)d\mu(x,y) = 1$$

Number of balls that is transported from position 3 to position 4.

Wasserstein Distance

$$\Pi(p,q) := \{\mu \in Prob(\mathcal{X} \times \mathcal{X}) \,|\, (\pi_1)_*(\mu) = p, (\pi_2)_*(\mu) = q\}\,.$$

The set of all transportation plans between p and q

$$c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

$$c(x,y) = d_{\mathcal{X}}(x,y)$$

$$W(p,q) := \inf_{\mu \in \Pi(p,q)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)d\mu$$

Wasserstein distance is the infimum cost among all transportation plans

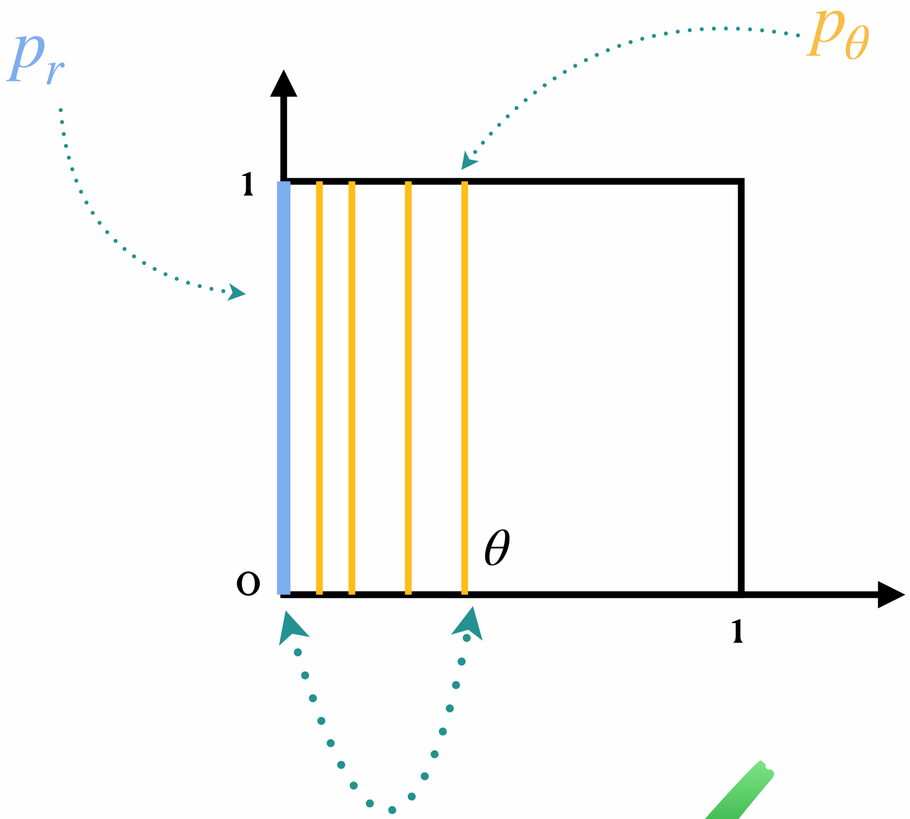
Wasserstein Distance

Looking for a "metric" , d, such that $d(p_\theta, p_r) \to 0$ as $\theta \to 0$

$$W(p_r, p_\theta) = \theta \to 0.$$

p_r : measure on $[0,1]^2$ that is uniformly distributed on $\{(0,y) \in [0,1]^2 \mid 0 \le y \le 1\}$

p_θ : measure on $[0,1]^2$ that is uniformly distributed on $\{(\theta, y) \in [0,1]^2 \mid 0 \le y \le 1 \ , \theta \in (0,1)\}$

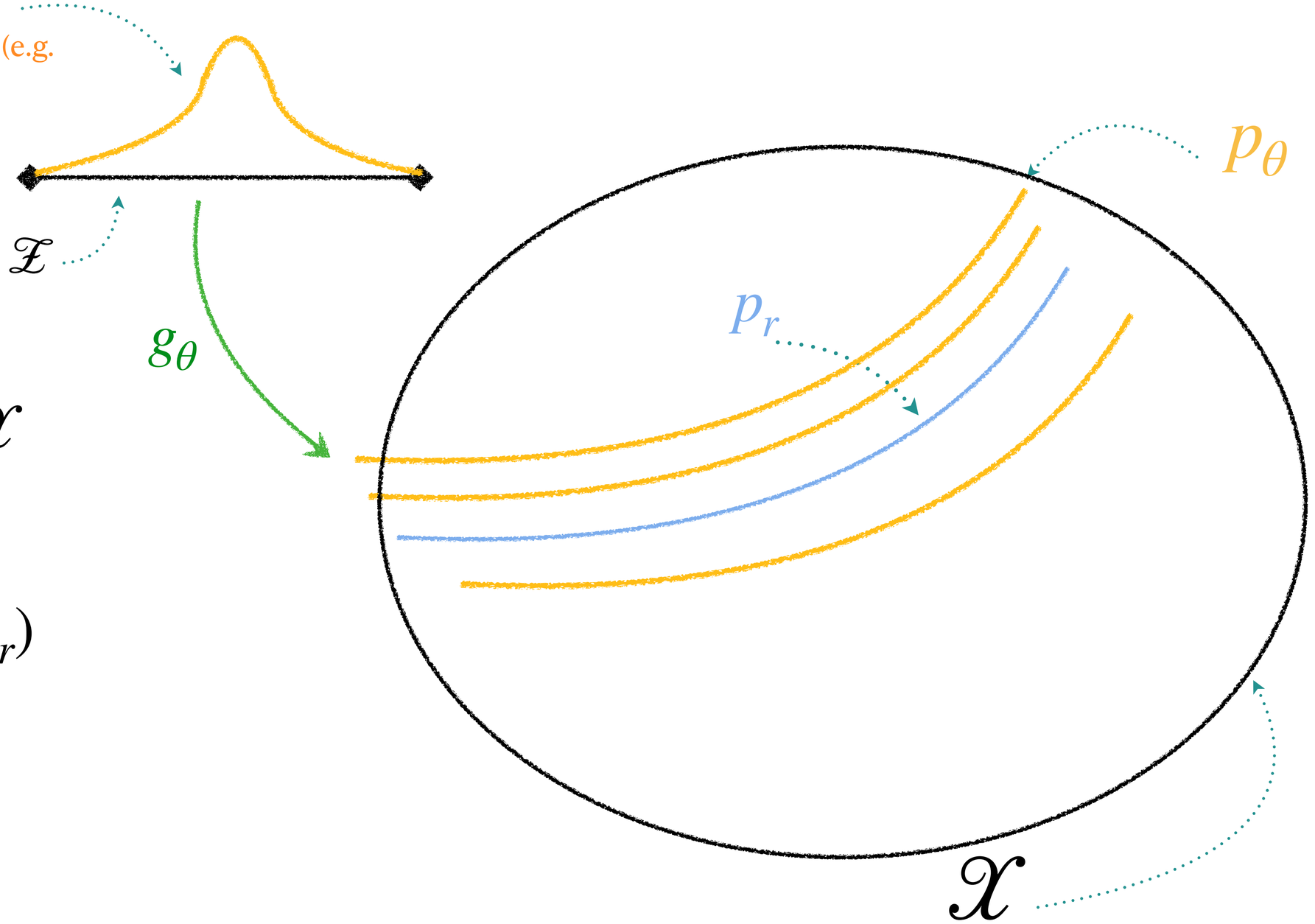


$W(p_r, p_\theta) = \theta$

Wasserstein distance looks to be a good candidate!

General case formulation

μ: your favorite measure on $\mathcal{Z}$ (e.g. gaussian, uniform, etc)



$$g : \mathcal{Z} \times \Theta \to \mathcal{X}$$

$$p_\theta := (g_\theta)_* \mu$$

$$\phi(\theta) := W(p_\theta, p_r)$$

Formal Theorem

Theorem.

- *If g is continuous in θ then ϕ is continuous.*
- *If g is locally Lipschitz and the local constants $L(\theta, z)$ satisfies,*

$$\int_{\mathcal{Z}} L(\theta, z) d\mu(z) < \infty$$

then ϕ is differentiable almost everywhere.

Theorem 1 in the paper.

Corollary. *If g_θ is a neural net and $\int_{\mathcal{Z}} \|z\| \, d\mu(z) < \infty$ then $\phi(\theta)$ is continuous and almost everywhere differentiable.*

Corollary 1 in the paper

Optimizing ϕ

$$\phi(\theta) := W(p_\theta, p_r)$$

$$W(p_\theta, p_r) := \inf_{\mu \in \Pi(p_\theta, p_r)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)d\mu$$

Not obvious from this representation how to compute ϕ in general or differentiate it.

$$\Pi(p_r, p_\theta) := \{\mu \in Prob(\mathcal{X} \times \mathcal{X}) \,|\, (\pi_1)_*(\mu) = p_r, (\pi_2)_*(\mu) = p_\theta\} \,.$$

Kantorovich-Rubinstein duality

No intractable inf anymore!

$$\phi(\theta) := W(p_r, p_\theta) = \sup_{\psi \in Lip_1} \left(\int_{\mathcal{X}} \psi(x) dp_r(x) - \int_{\mathcal{X}} \psi(x) dp_\theta(x) \right)$$

Theorem 3 in WGAN paper

$$f^* = argmax_{\psi \in Lip_1} \left(\int_{\mathcal{X}} \psi(x) dp_r(x) - \int_{\mathcal{X}} \psi(x) dp_\theta(x) \right)$$

$$\frac{\partial \phi}{\partial \theta} = - \int_{\mathcal{Z}} \frac{\partial \left(f^* \circ g_\theta \right)(z)}{\partial \theta} d\mu(z)$$

When both sides are defined

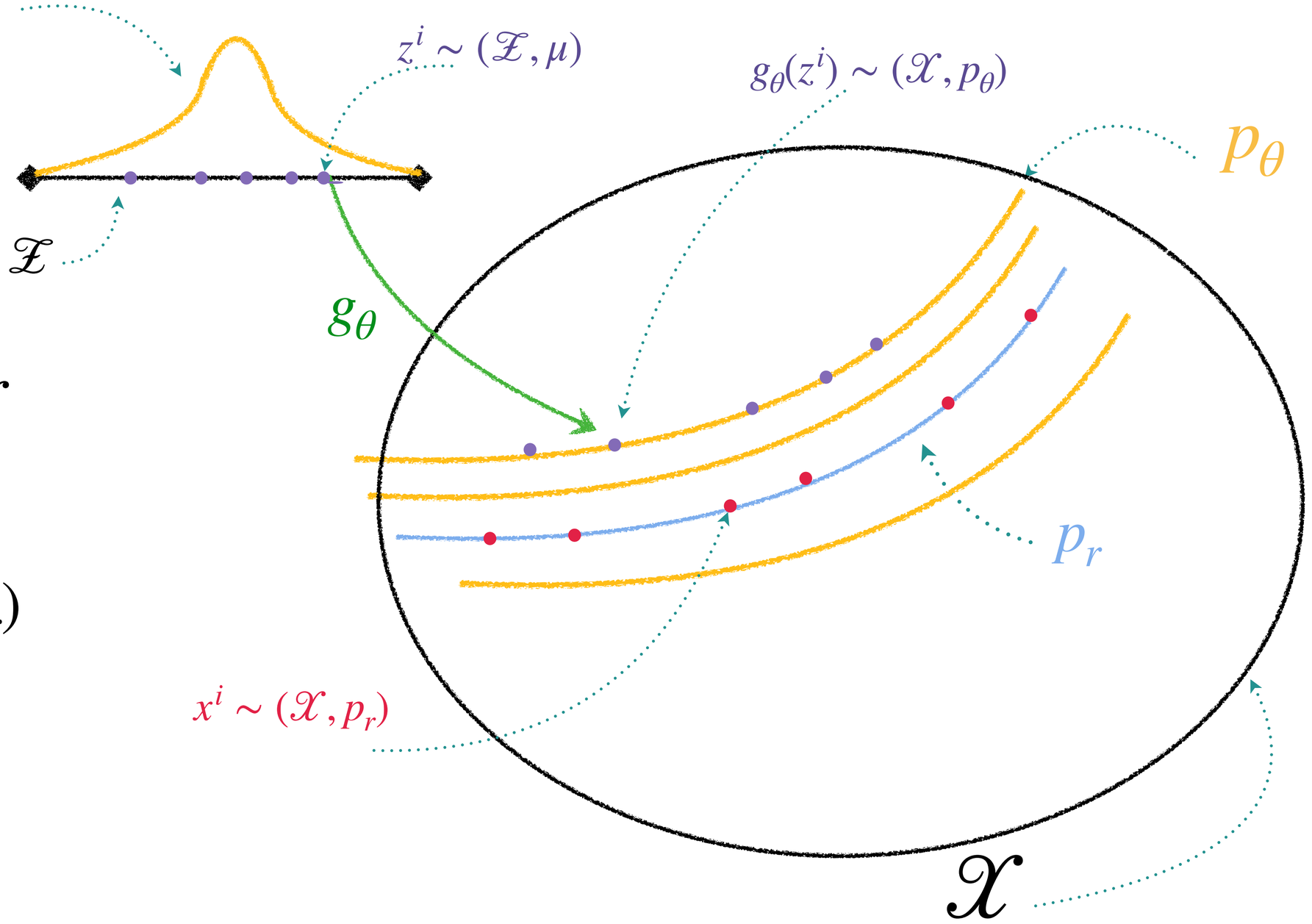
WGAN Algorithm

μ: your favorite measure on $\mathcal{Z}$ (e.g. gaussian, uniform, etc)

$$g : \mathcal{Z} \times \Theta \to \mathcal{X}$$

$$p_\theta := (g_\theta)_* \mu$$

$$\phi(\theta) := W(p_\theta, p_r)$$



WGAN Algorithm

$$x_i \sim (\mathcal{X}, p_r) \qquad z_i \sim (\mathcal{Z}, \mu)$$

$$g_\theta : \mathcal{Z} \to \mathcal{X}, \theta \in \Theta \qquad \phi(\theta) := W(p_r, p_\theta) \qquad \frac{\partial \phi}{\partial \theta} = -\int_{\mathcal{Z}} \frac{\partial \left(f^* \circ g_\theta\right)(z)}{\partial \theta} d\mu(z) \qquad \frac{\partial}{\partial \theta}\left(\frac{1}{n}\sum_{i=1}^{n} f^* \circ g_\theta\right)(z^i)\Big)$$

A neural net that generates samples

Want to minimize with gradient descent

Want to estimate

Need to estimate f^*

Needs to be compact

$$f_w : \mathcal{X} \to \mathbb{R}, w \in \mathcal{W} \qquad \Big(\int_{\mathcal{X}} f_w(x) dp_r(x) - \int_{\mathcal{X}} f_w(x) dp_\theta(x)\Big) \qquad \frac{\partial}{\partial w}\Big(\sum_{i=1}^{n} \frac{1}{n} f_w(x^i) - \frac{1}{n}\sum_{i=1}^{n} f_w(g_\theta(z^i))\Big)$$

Second neural net for estimating f^*

Want to maximize with gradient ascent

We can compute gradient of the estimation with respect to w

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

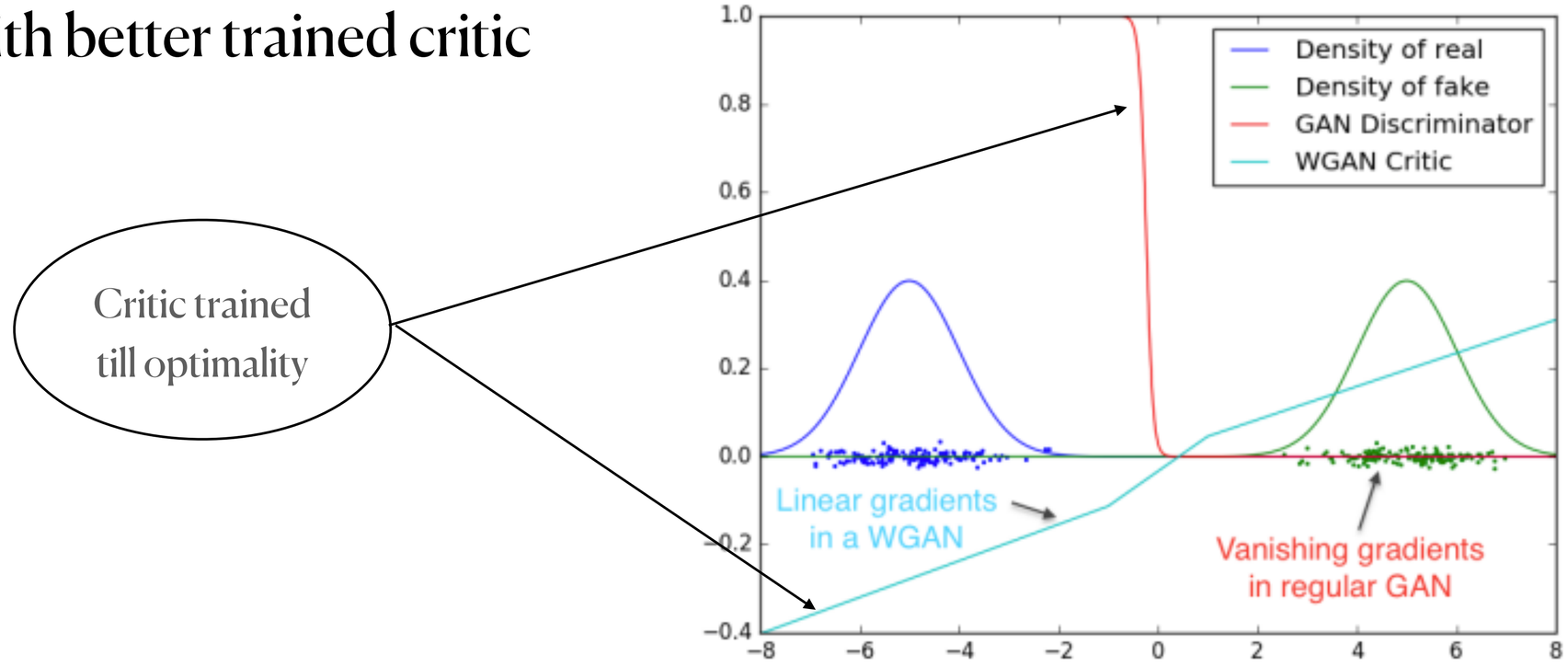
The algorithm from WGAN paper

Part II: Training and empirical aspects of WGAN

Training WGAN

A better GAN loss function

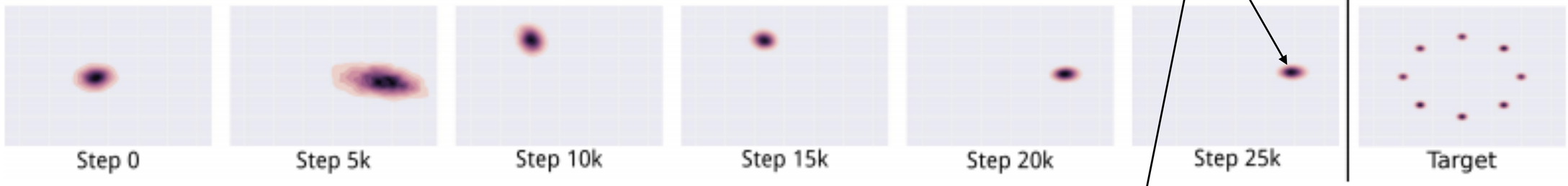
- EM distance is continuous and differentiable a.e
- Reliable gradients with better trained critic



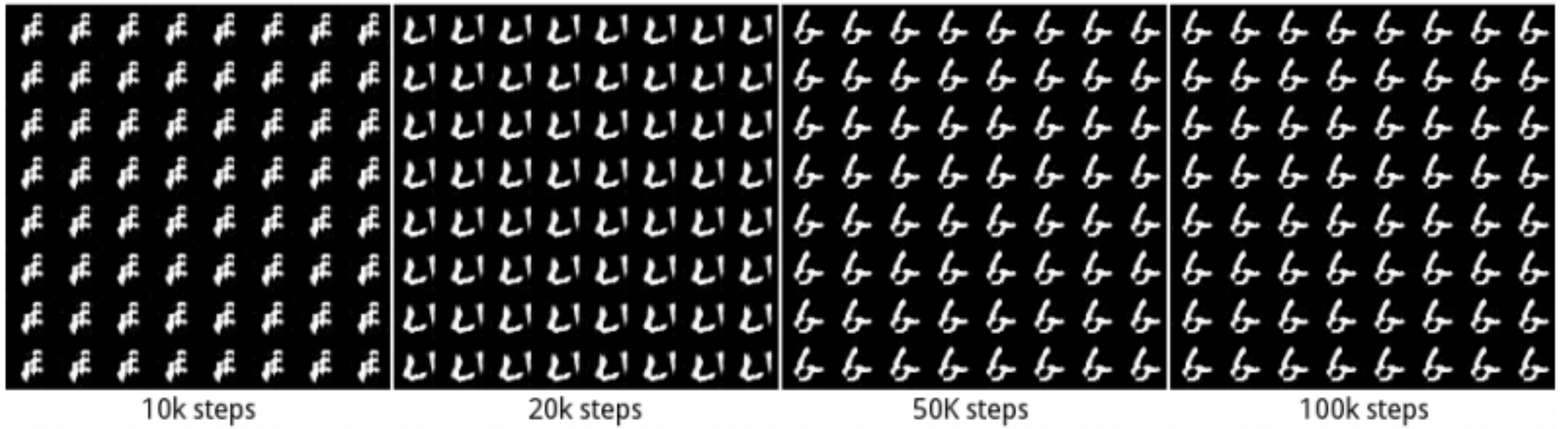
Training WGAN

No mode collapse for WGAN

- Mode collapse for Standard GAN



Single mode learnt



Training WGAN

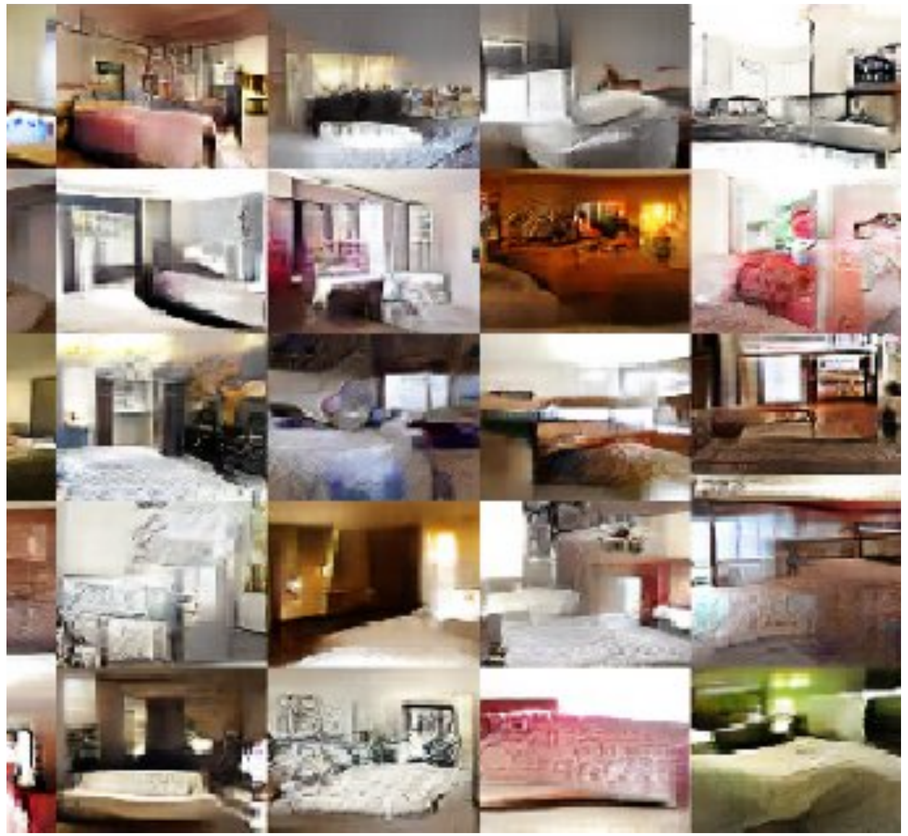
Explanations for mode collapse ?

- Standard GAN
  - For a fixed discriminator, generator uses few modes to fool it
  - Training critic til optimality provides less meaningful gradients
- WGAN
  - For critic trained till optimality, we get more reliable gradients

Empirical Results

Datasets & Baselines

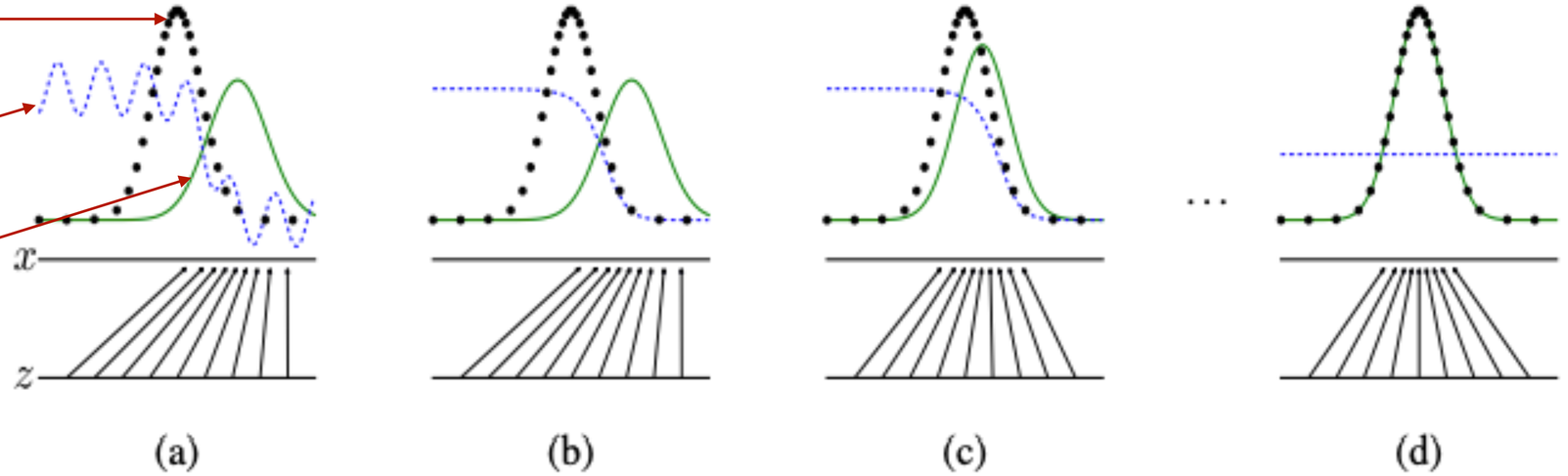
- Image generation Datasets
  - LSUN-Bedroom dataset



- Baselines
  - DCGAN - GAN with conv architectures

Standard GAN Training

- Data distribution
- Discriminator D
- Generator G

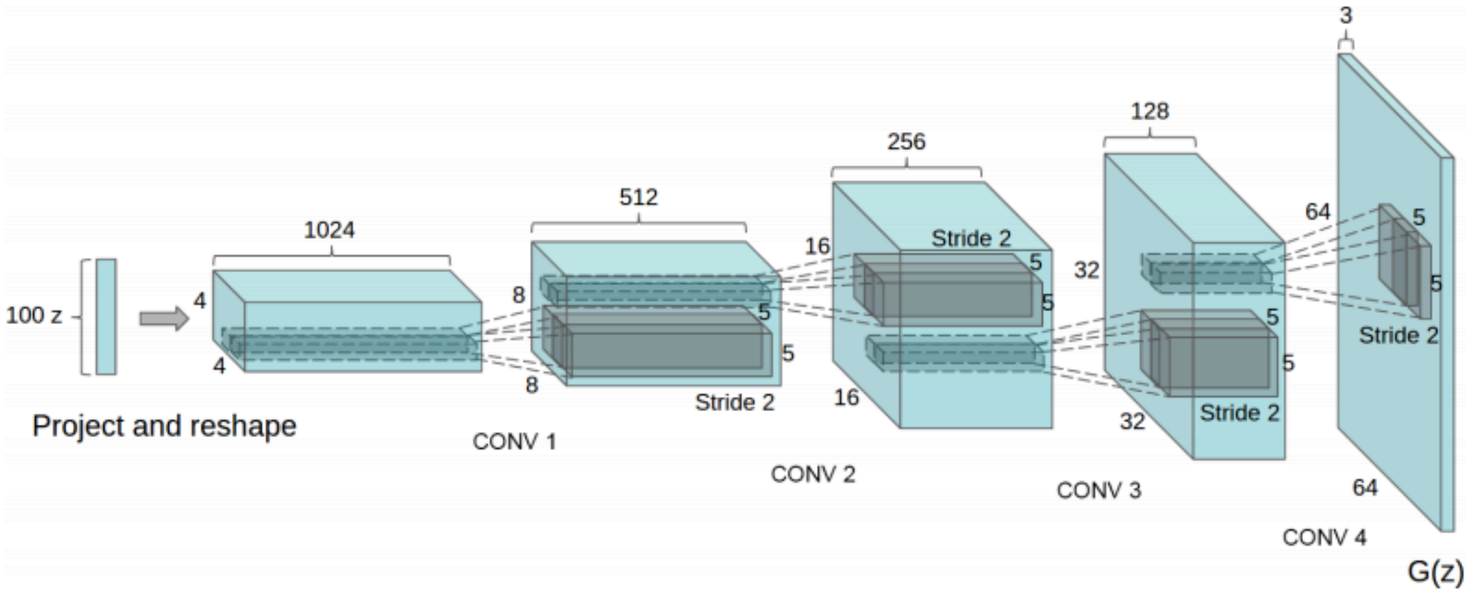


Objective

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

DCGAN Generator

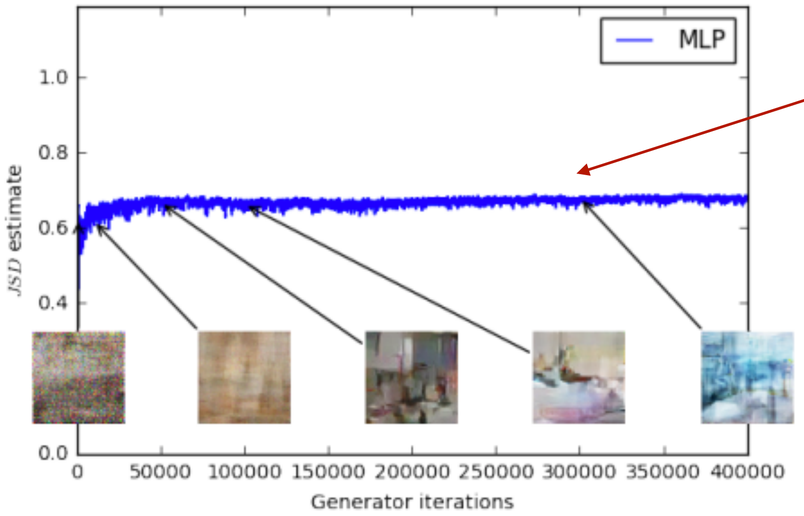
- Replace pooling layers with convolutions
- Use batchnorm
- Remove FC hidden layers
- Use ReLU for G
- Use Leaky ReLU for D

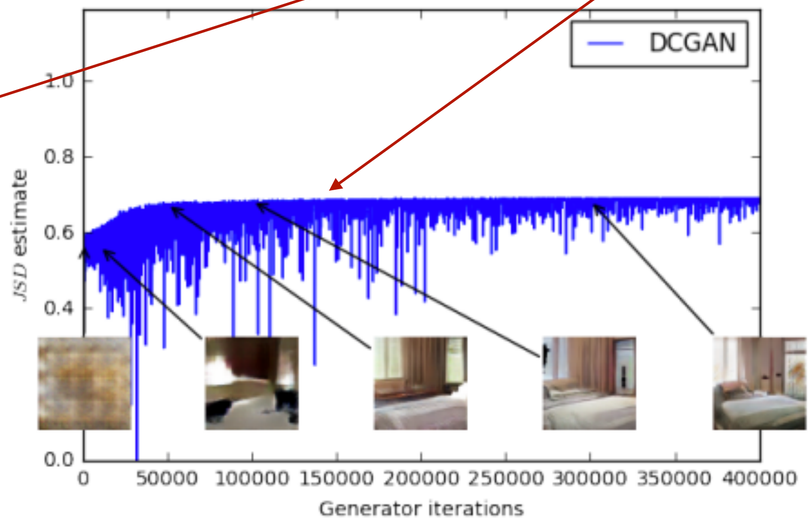


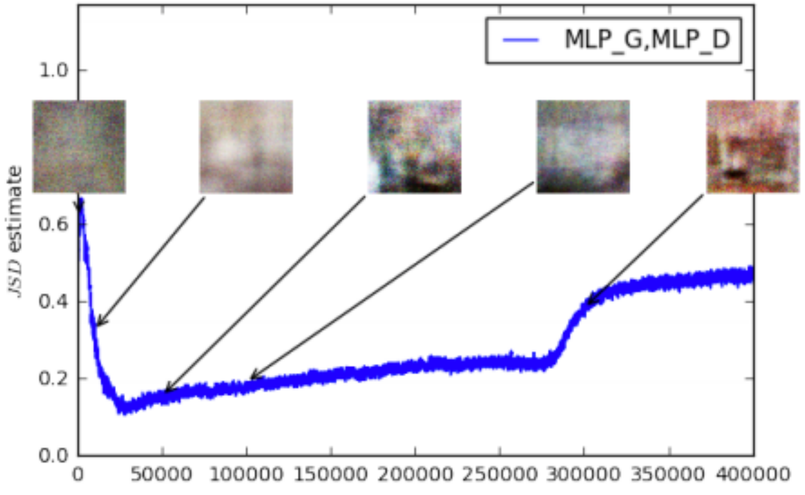
WGAN as Meaningful loss

JS estimates with standard GAN training

Loss close to log(2)=0.69







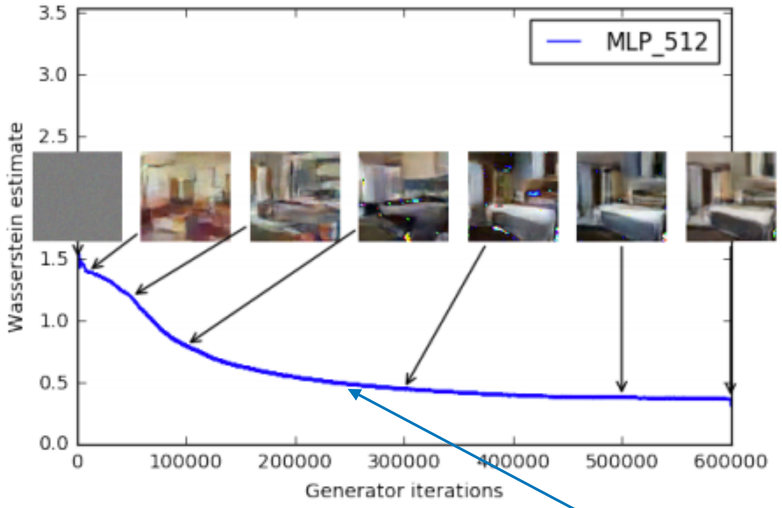
Increasing error (JS Estimate)
for MLP generator (FC Network)

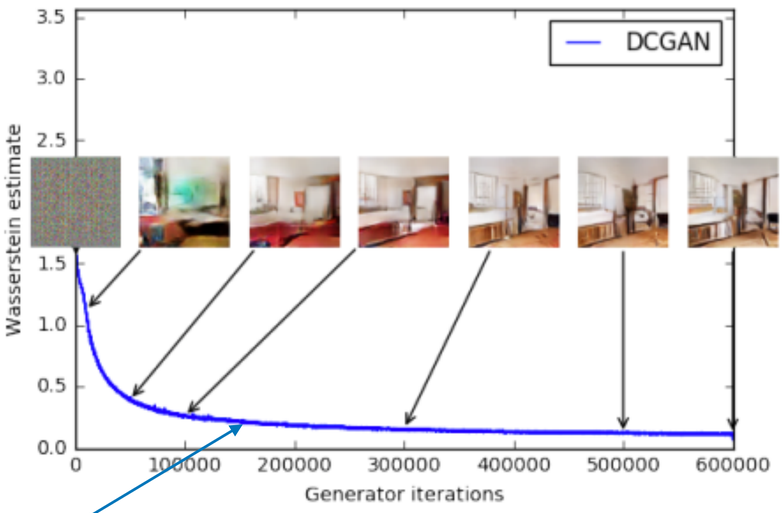
For DCGAN generator
sample quality better,
But no correlation to JS estimates

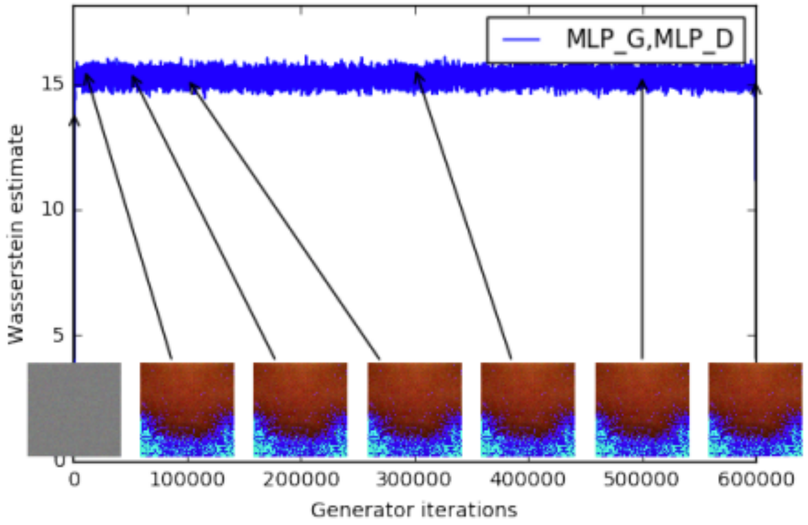
For MLP gen and disc,
Curve fluctuates regardless
of quality

WGAN as Meaningful loss

Wasserstein estimates with WGAN training







Wasserstein loss decreases,
as image quality increases

High Learning Rate used,
Loss and quality are constant

WGAN Training

Negative Result

- Training unstable with
  - Momentum based optimizer eg ADAM
  - High learning rates
- Since critic loss is non-stationary
  - momentum a potential cause
  - Works with RMSProp

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta)\left(\frac{\delta C}{\delta w}\right)^2$$

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{E[g^2]_t}} \frac{\delta C}{\delta w}$$

Adaptive LR

Improved Stability - WGAN

WGAN vs Standard GAN

WGAN training

Standard GAN training

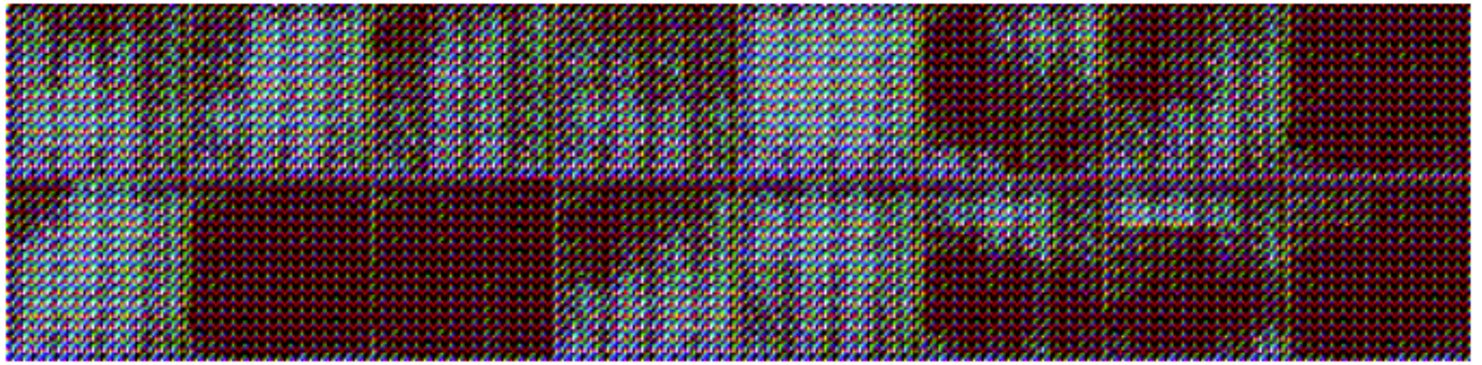
High quality samples for both

Improved Stability - WGAN

WGAN vs Standard GAN

WGAN training
- Able to produce samples

Standard GAN training
- fails



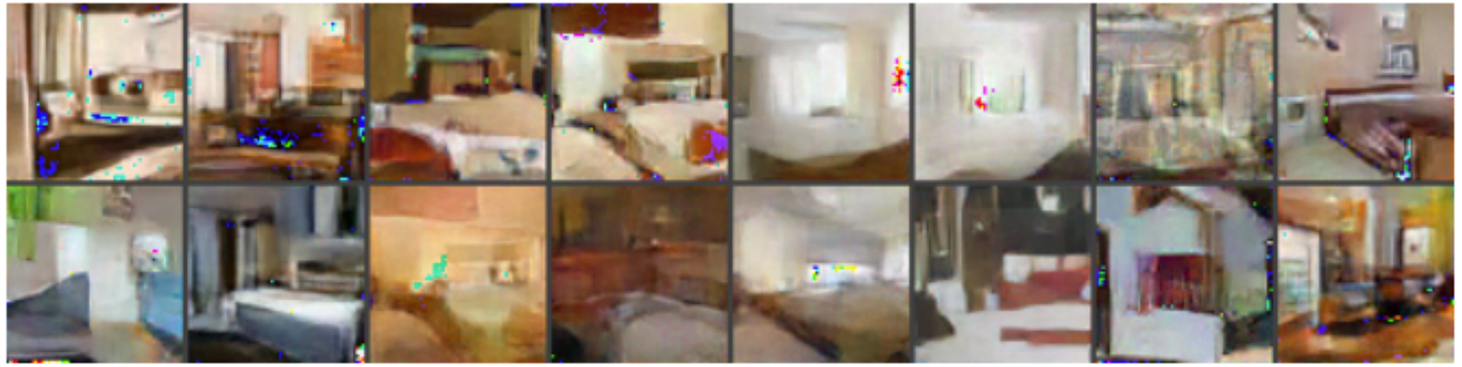
DCGAN Generator without batch normalization
and constant number of filter each layer

Improved Stability - WGAN

WGAN vs Standard GAN

MLP Generator

WGAN training
- Able to produce good samples



Standard GAN training
- Lower quality images

Notice the mode collapse for standard GAN samples

Related Works

Works with Integral Probability Metrics (IPM)

- IPM with class $\mathcal{F}$: $d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_\theta}[f(x)]$
- $\mathcal{F}$: 1-Lipschitz gives $W(\mathbb{P}_r, \mathbb{P}_\theta)$
- $\mathcal{F}$: $\{ f : f(x) \in [-1,1] \forall x \}$ gives $d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_\theta) = \delta(\mathbb{P}_r, \mathbb{P}_\theta)$ (Total variation distance)
- $\mathcal{F}$: $\{ f : f(x) \in [0,m] \forall x \}$ for Energy-based GANs (EBGANs)
    - Same behavior as Total variation distance

Works with Integral Probability Metrics (IPM)

- $\mathcal{F} = \{f \sim \mathcal{H} : ||f||_{\mathcal{H}} \leq 1\}$, for some RKHS $\mathcal{H}$ associated with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ gives Maximum Mean Discrepancy (MMD)
  - MMD is a proper metric & no need to maximize the inner supremum
  - Disadvantage: quadratic computational cost
  - Linear approximations have worse sample complexity
- Generative Moment Matching Networks (GMMs): Generative networks using MMD
  - Quadratic cost as a function of samples
  - Vanishing gradients for low-bandwidth kernels

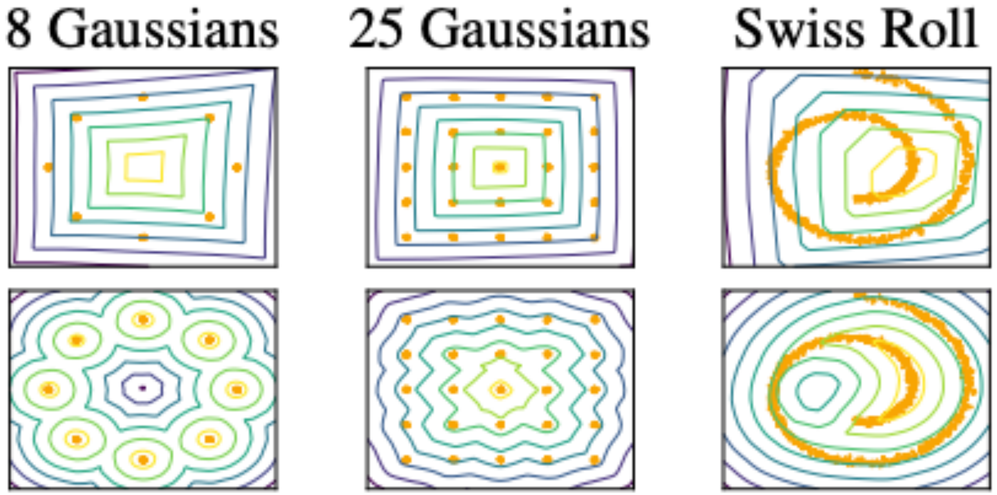
Follow-up Works

WGAN with Gradient Penalty

- Difficulties with weight constraints:
  - Optimization difficulties - Resulting critic can have a pathological value surface

Weight clipping critic fail to capture higher moments

Critic with Gradient Penalty



Value surface of WGAN critics

WGAN with Gradient Penalty

- A differentiable function is 1-Lipschitz iff it has gradients with norm at most 1 everywhere.

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

Soft constraint

- Stable training

Sliced WGAN

- Wasserstein computation in dual involves solving saddle point problem
- Another variant of Wasserstein uses random projections and is a single minimization
- Wasserstein: $$W_2^2(D, F) = \frac{1}{|F|} \min_{\sigma \in \Sigma_{|F|}} \sum_{i=1}^{|F|} ||D_{\sigma(i)} - F||_2^2$$

Can be solved by sorting for 1D case

Sliced Wasserstein: $$\min_{\theta} \frac{1}{|\hat{\Omega}|} \sum_{\omega \in \hat{\Omega}} W_2^2(D^\omega, F^\omega)$$

Random projection directions

Summary

- Introduced Wasserstein distances for GANs
    - Theoretical analysis of EM distance behavior in comparison to popular divergences like KL, JS
- Define a form of GAN using duality : Wasserstein GAN
    - Meaningful loss metric for distributions
    - Stable training of GAN
    - Gets rid of issues like mode collapse
- Experiments show stable learning of image generators

Thank You!

Questions ?

Quiz Problems

According to the paper, which of the following gives a more sensible cost function for learning probability distribution that are supported in low dimensional submanifolds?

- KL divergence
- JS divergence
- Total variation distance(TV)
- Wasserstein distance (Earth-Mover) 

Quiz Problems

Which of the following statements is true:

- In WGAN, we can train the critic till optimality.
- WGAN solves the mode collapse issue in GAN.
- WGAN training is not stable when one uses momentum based optimizer or high learning rate.
- All of the above. 

Quiz Problems

Which of the following found to be true in empirical evaluation of WGANs?

- Experiments illustrate that Wasserstein estimates correlates well with the quality of the generated samples. ✓
- WGAN training is reported to be stable with Adam optimizer.
- It is observed that WGANs are much more stable than GANs when one varies the generator architecture. ✓
- JS estimates follow the same correlation with image quality as the Wasserstein estimates.

References

- The slides have materials from following works:
- Wasserstein GANs - https://arxiv.org/abs/1701.07875
- Generative Adversarial Networks - https://arxiv.org/abs/1406.2661
- Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks - https://arxiv.org/abs/1511.06434
- Unrolled Generative Adversarial Networks - https://arxiv.org/abs/1611.02163
- Generative Modelling using the Sliced Wasserstein Distance - https://arxiv.org/abs/1803.11188
- Improved Training of Wasserstein GANs - https://arxiv.org/abs/1704.00028
- Understanding RMSprop - https://towardsdatascience.com/understanding-rmsprop-faster-neural-network-learning-62e116fcf29a
- LSUN Dataset: https://www.yf.io/p/lsun
- Image Sources at slide 4: https://www.pinterest.com/pin/382383824583177634/ and https://www.pinterest.com/pin/161777811601964865/