Banach Wessenstian GAN *

NeurIPS 2018

Sheng-Je Huang

Institute of Communications Engineering
National Chiao Tung University, Hsinchu

Feburary 20 , 2019

---

*Jonas Adler, KTH-Royal institute of Technology Research and Physics
Sebastian Lunz, University of Cambridge

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Outline

- **Introduction**
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Introduction

- Extend WGAN implemented via a **gradient penalty** (GP) term to any separable complete normed space.
- Efficiently implemented BWGAN by replacing the ℓ^2 norm into **dual norm**.
- Give theoretically grounded heuristics for the choice of regularization parameters.

Outline

- Introduction
- **Background**
  - **Generative adversarial networks**
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Generative adversarial networks

- GANs perform generative modeling by learning a map $G : Z \to B$ from a low-dimensional **latent space** Z to **image space** B, mapping a fixed noise distribution $\mathbb{P}_Z$ to a distribution of generated images $\mathbb{P}_G$.

- The famous **minimax** game between generator G and critic D

$$\min_G \max_D \mathbb{E}_{X\sim\mathbb{P}_r}[\log(D(x))] + \mathbb{E}_{Z\sim\mathbb{P}_Z}[\log(1 - D(G_\Theta(Z)))]. \qquad (1)$$

- Using **Jensen-Shannon divergence** as distance measure between the dirstirbutions $\mathbb{P}_G$ and $\mathbb{P}_r$.

Outline

- Introduction
- Background
  - Generative adversarial networks
  - **Wasserstein metrics**
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Wasserstein metrics (1/2)

- To overcome undesirable behavior of the JSD in the presence of **singular measures**, using the Wasserstein metric to quantify the distance between the distributions $\mathbb{P}_G$ and $\mathbb{P}_r$.

- The Wasserstein distance provide **meaningful gradients** to the gernerator even when the measures are mutually singular.

- The Wasserstein-p, $p \geqq 1$, distance is defined as

$$\mathrm{Wass}_p(\mathbb{P}_G, \mathbb{P}_r) := \left(\inf_{\pi \in \Pi(\mathbb{P}_G, \mathbb{P}_r)} \mathbb{E}_{(X_1, X_2) \sim \pi} d_B(X_1, X_2)^p \right)^{1/p} \tag{2}$$

Wasserstein metrics (2/2)

- The infimum is highly intractable.
- The **Kantorovich-Rubinstein duality** provides a way of more efficiently computing the Wasserstein-1 distance.

$$\mathrm{Wass}_p(\mathbb{P}_G, \mathbb{P}_r) = \sup_{\mathrm{Lip}(f) \leq 1} \mathbb{E}_{X \sim \mathbb{P}_G} f(X) - \mathbb{E}_{X \sim \mathbb{P}_r} f(X) \tag{3}$$

- The supremum is taken over all Lipschitz continuous functions $f : B \to \mathbb{R}$ with Lipschitz constant equal or less than one.
- If we consider γ-Lipschitz with a function $f : B \to \mathbb{R}$, we can get

$$|f(x) - f(y)| \leq \gamma d_B(x, y).$$

Wasserstein metrics

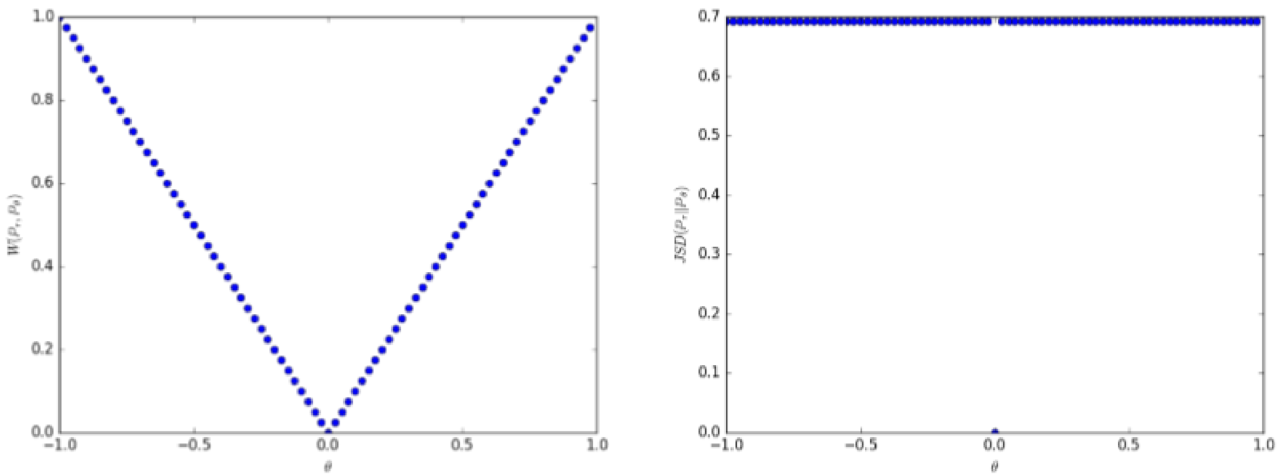


Figure 1: These plots show $\rho(\mathbb{P}_\theta, \mathbb{P}_0)$ as a function of θ when ρ is the EM distance (left plot) or the JS divergence (right plot). The EM plot is continuous and provides a usable gradient everywhere. The JS plot is not continuous and does not provide a usable gradient.

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - **Wasserstein GAN**
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Wasserstein GAN

- Implementing GANs with the Wasserstein metric requires to approximate the supremum in (3) with a neural network.

- Original WGAN [1] using **weight clipping** to statisfy the Lipschitz constraint.

- However weight clipping in WGAN leads to optimization difficulties, and that even when optimization succeeds the resulting critic can have a pathological value surface.

[1] Martin Arjovsky, Soumith Chintala, and Leon Boeou. Wasserstein Generative Adversarial Networks. *International Conference on Machine Learning, ICML*, 2017

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - **Improved Wasserstein GAN**
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Improved Wasserstein GAN [2]

- **Gradient penalty** as an uncontrollable additional constraint becomes another characterization of 1-Lipschitz functions.

- In particular, they prove that if $B = \mathbb{R}^n, d(x,y)_B = \|x-y\|_2$ we have the gradient characterization

$$f \text{ is 1-Lipschitz} \iff \|\nabla f(x)\|_2 \leq 1 \text{ for all } x \in \mathbb{R}^n.$$

- Penalty term to the **loss function** of D that takes the form

$$\mathbb{E}_{\widehat{X}}\left(\|\nabla D(\widehat{X})\|_2 - 1\right)^2 \tag{4}$$

[2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in Neural information Processing Systems (NIPS)*, 2017.

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - **Banach spaces**
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Banach spaces (1/3)

- We often choose the ℓ^2 norm as underlying distance measure on **image space**, but many other distance notions are possible that account for more specific image features.

- If a vector space B is equipped with a notion of **length**, a norm $\|\cdot\|_B : B \to \mathbb{R}$, we call it a **normed space**.

- A normed space is called a **Banach space** if it is complete, that is, Cauchy sequences converge.

- All separable Banach spaces are Polish spaces and we can define Wasserstein metrics on them using the induced metric $d_B(x, y) = \|x - y\|_B$.

Banach spaces (2/3)

- For any Banach space B, we can consider the space of all bounded linear functionals $B \to \mathbb{R}$, which we will denote B^* and call the **topological dual of** B.

- Banach space with norm $\| \cdot \|_{B^*} : B^* \to \mathbb{R}$ given by

$$\|x^*\|_{B^*} = \sup_{x \in B} \frac{x^*(x)}{\|x\|_B}. \tag{5}$$

Banach spaces (3/3)

The set of functions $x : \Omega \to \mathbb{R}$ with **norm**

- L^p**-spaces:**

$$\|x\|_{L^p} = \left(\int_\Omega x(t)^p dt \right)^{1/p} \tag{6}$$

is a Banach with dual $[L^p]^* = L^q$ where $1/p + 1/q = 1$.

- **Sobolev spaces:**

$$\|x\|_{W^{1,2}} = \left(\int_\Omega x(t)^2 + |\nabla x(t)|^2 dt \right)^{1/2} \tag{7}$$

It can rewrite the equation by multiplying with ξ in the Fourier space

$$\|x\|_{W^{s,p}} = \left(\int_\Omega \left(\mathcal{F}^{-1} \left[(1 + |\xi|^2)^{s/2} \mathcal{F} x \right] (t) \right)^p dt \right)^{1/p}. \tag{8}$$

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- **Banach Wasserstein GANs**
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- Conclusion

Banach Wasserstein GANs

- In particular, for any Banach space B with norm $\|\cdot\|_B$, we will derive the loss function

$$L = \frac{1}{\gamma}(\mathbb{E}_{X\sim\mathbb{P}_\theta} D(X) - \mathbb{E}_{X\sim\mathbb{P}_r} D(X)) + \lambda \mathbb{E}_{\hat{X}} \left(\frac{1}{\gamma} \|\partial D(\hat{X})\|_{B^*} - 1 \right)^2 \tag{9}$$

where $\lambda, \gamma \in \mathbb{R}$ are **regularization parameters**.

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - **Enforcing the Lipschitz constraint**
  - Regularization parameter
- Computational results
- Conclusion

Enforcing the Lipschitz constraint (1/3)

- We require a more **general notion of gradient**:
  The function f is call *Fréchet differentiable* at $x \in B$ if there is a bounded linear map $\partial f(x) : B \to \mathbb{R}$ such that

$$\lim_{\|h\|_B \to 0} \frac{1}{\|h\|_B} |f(x+h) - f(x) - [\partial f(x)](h)| = 0. \tag{10}$$

- The gradient $\nabla f(x)$ in $\mathbb{R}^n$ with the standard inner product is connected to the Fréchet derivative via $[\partial f(x)](h) = \nabla f(x) \cdot h$.

Enforcing the Lipschitz constraint (2/3)

- **Lemma 1** *Assume $f : B \to \mathbb{R}$ is Fréchet differentiable. Then f is γ-Lipschitz if and only if*

$$\|\partial f(x)\|_{B^*} \leq \gamma \quad \forall x \in B. \tag{11}$$

- According to the **Lemma 1**, we can get the γ-*Lipschitz contraints*

$$|f(x) - f(y)| \leq \gamma \|x - y\|_B$$

Enforcing the Lipschitz constraint (3/3)

- Gradient norm penalization requires characterizing the dual B^* of B. In a **finite dimension**, there is an linear continuous bijection $\iota : \mathbb{R}^n \to B$ given by

$$\iota(x)_i = x_i. \tag{12}$$

- We can write $f = g \circ \iota$ where $g : \mathbb{R}^n \to \mathbb{R}$ and we can get $\partial f(x) = \iota^*(\partial g(\iota(x)))$ by the **chain rule**. ($\iota^* : \mathbb{R}^n \to B^*$ is the adjoint of ι.)

- The derivative in **finite dimensional Banach spaces** can be done using standard automatic differentiation libraries.

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - **Regularization parameter**
- Computational results
- Conclusion

Regularization parameter (1/2)

- Regularization term:

$$\lambda \mathbb{E}_{\hat{X}} \left(\frac{1}{\gamma} \|\partial D(\hat{X})\|_{B^*} - 1 \right)^2 .$$

- In order to avoid having to hand-tune parameters for every choice of norm, author derive some **heuristic parameter** choice rules.

- Assuming that G is the zero-generator and symmetry of $\mathbb{P}_r$, the D will be decided by a single constant $f(x) = c\|x\|_B$.
  We can form the optimization problem

$$\min_{c \in \mathbb{R}} \mathbb{E}_{X \sim \mathbb{P}_r} \left[-\frac{c\|X\|_B}{\gamma} + \frac{\lambda (c-\gamma)^2}{\gamma^2} \right].$$

Regularization parameter (2/2)

- By solving optimization problem, we can obtain

$$c = \gamma \left(1 + \frac{\mathbb{E}_{X \sim \mathbb{P}_r} \|X\|_B}{2\lambda}\right).$$

- Since the norm has Lipschitz constant 1, we want $c \approx \gamma$. To has a small relative error, we get the heuristic rule

$$\lambda \approx \mathbb{E}_{X \sim \mathbb{P}_r} \|X\|_B.$$

- Assuming λ was appropriately chosen, we find in general (by lemma 1) $\|\partial D(x)\|_{B^*} \approx \gamma$. We want to enforce $\|x\|_{B^*} \approx \|\partial D(x)\|_{B^*}$, hence $\gamma \approx \|x\|_{B^*}$.

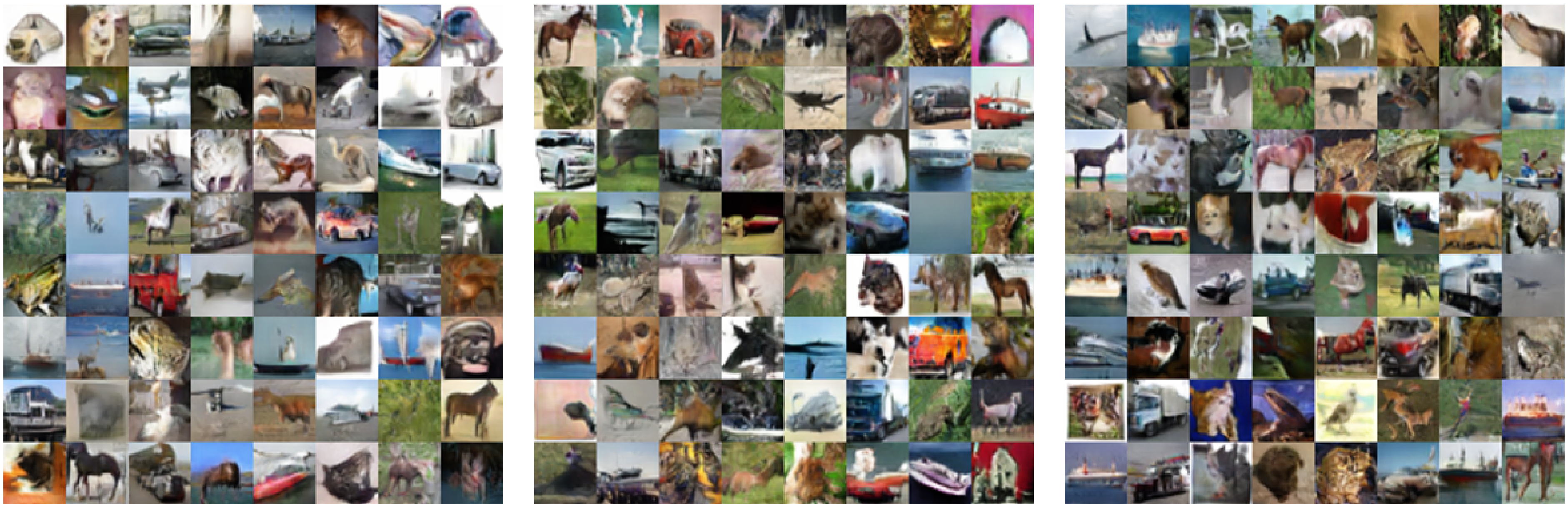
- We pick the expected values as a represnetative, we can finally obtain the heuristic

$$\gamma \approx \mathbb{E}_{X \sim \mathbb{P}_r} \|X\|_{B^*}.$$

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- **Computational results**
- Conclusion

Computational results

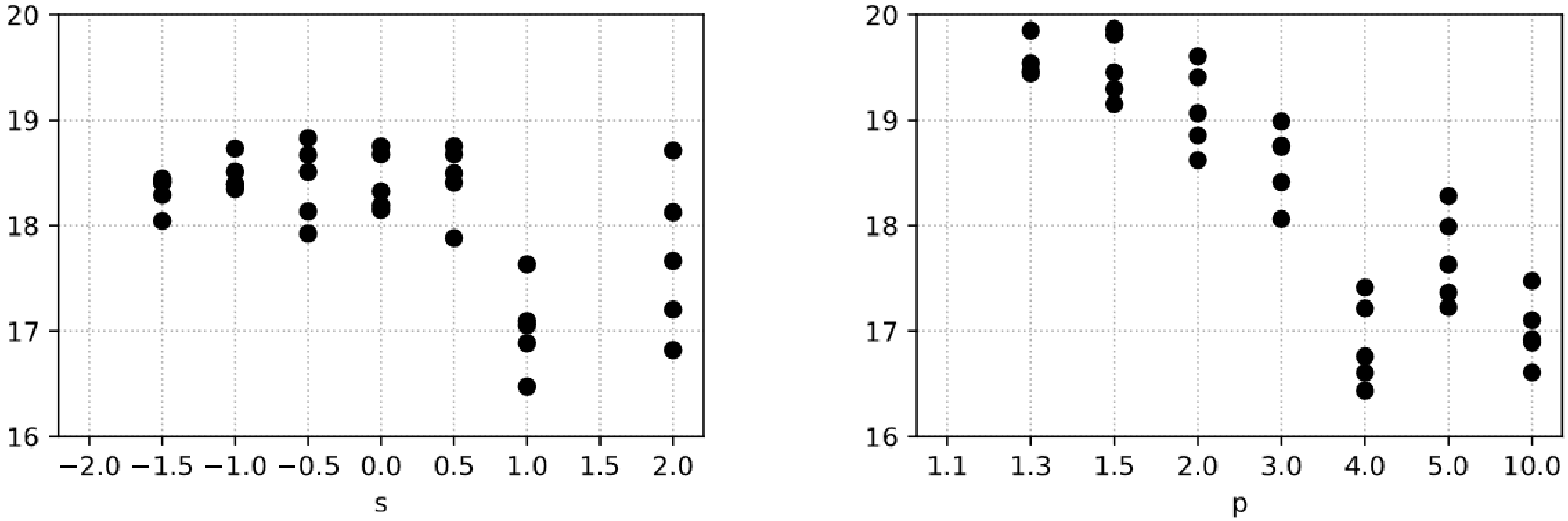


(a) $p = 1.3$ (b) $p = 2.0$ (c) $p = 10.0$

Figure 1: Generated CIFAR-10 samples for some L^p spaces.

Computational results

- A high image quality corresponds to **low** FID scores.



(a) $W^{s,2}$ (b) L^p

Figure 2: FID scores for BWGAN on CIFAR-10.

Computational results

- A high image quality corresponds to **high** Inception scores.

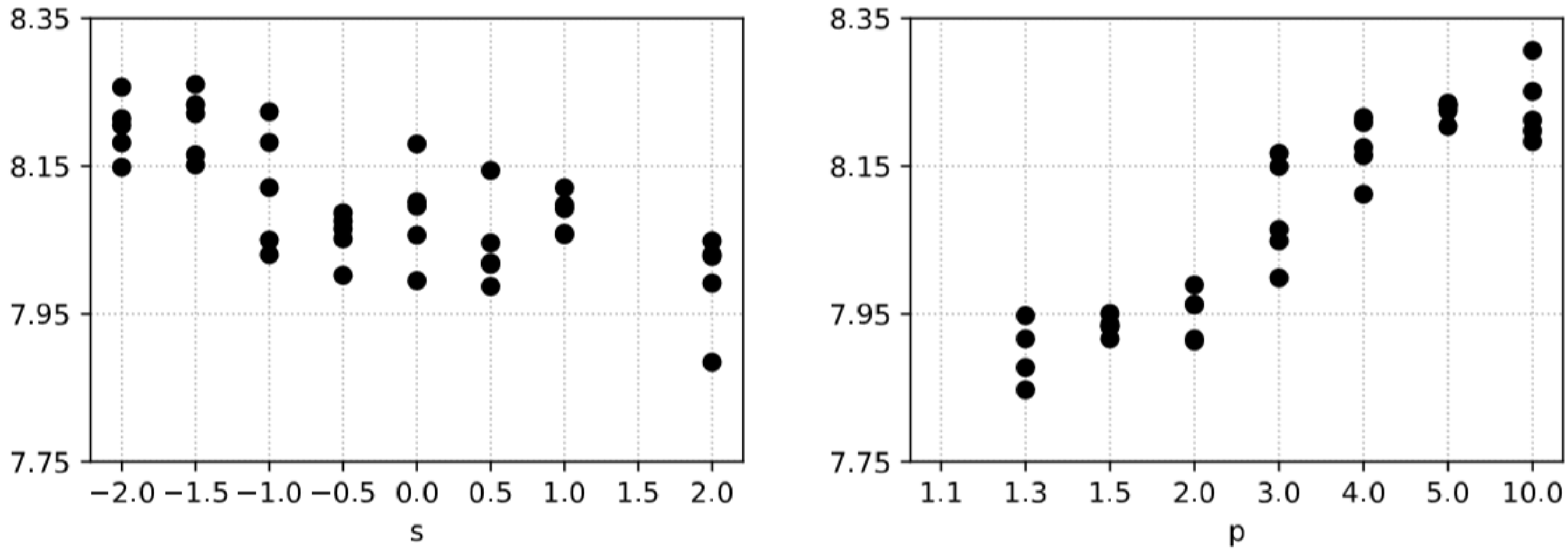


Figure 3: Inception scores for BWGAN on CIFAR-10.

Computational results

Method	Inception Score
DCGAN [16]	$6.16 \pm .07$
EBGAN [21]	$7.07 \pm .10$
WGAN-GP [7]	$7.86 \pm .07$
CT GAN [20]	$8.12 \pm .12$
SNGAN [14]	$8.22 \pm .05$
$W^{-\frac{3}{2},2}$-BWGAN	$8.26 \pm .07$
L^{10}-BWGAN	$8.31 \pm .07$
Progressive GAN [9]	$8.80 \pm .05$

Figure 4: Inception scores on CIFAR-10.

Computational results

- Different norm are suitable for different dataset.

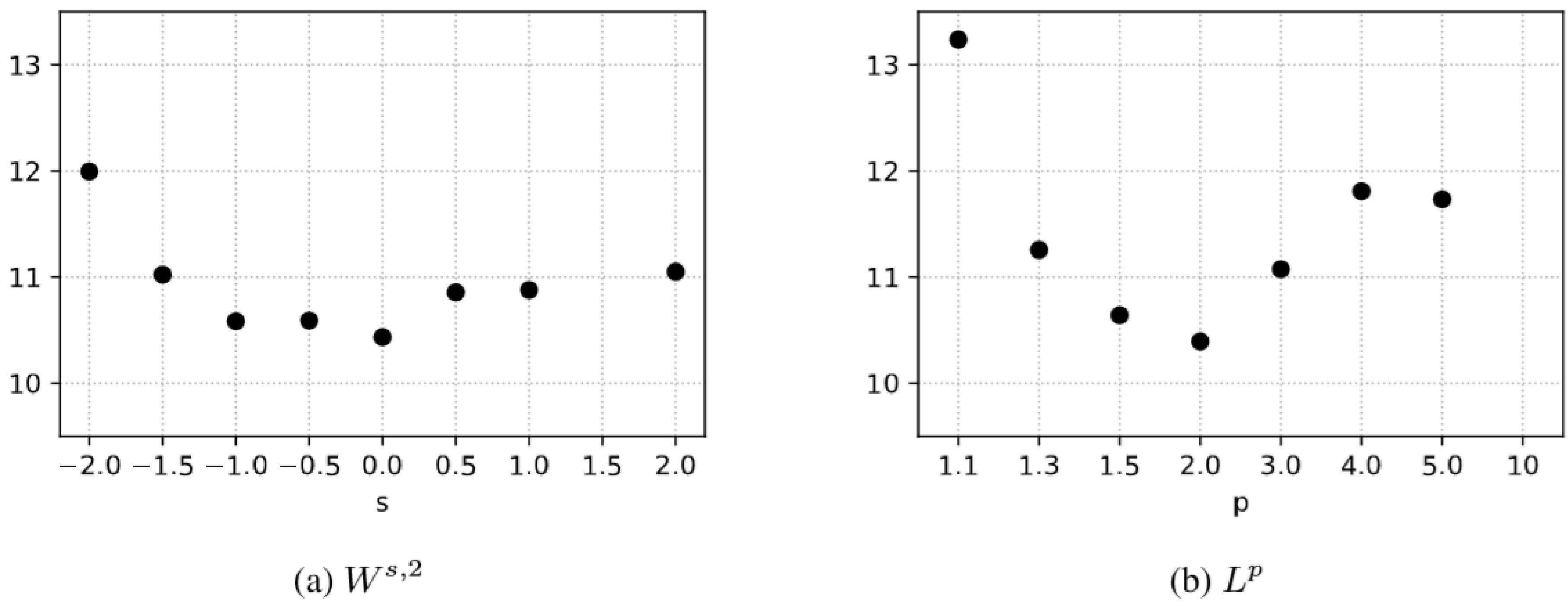


Figure 5: FID scores for BWGAN on CelebA.

Outline

- Introduction
- Background
  - Generative adversarial networks
  - Wasserstein metrics
  - Wasserstein GAN
  - Improved Wasserstein GAN
  - Banach spaces
- Banach Wasserstein GANs
  - Enforcing the Lipschitz constraint
  - Regularization parameter
- Computational results
- **Conclusion**

Conclusion

- This paper analyzed the dependence of WGANs on the notion of distance between **images**.
- Showed how choosing distances other than the ℓ^2 metric can be used to make WGANs focus on **particular image features** of interest.
- Generalize of WGANs with gradient norm penalization to **Banach spaces**, allowing to easily implement WGANs for a wide range of underlying norms on images.
- This work was motivated by images, the theory is general and can be applied to data in **any normed space**.