Banach Wasserstein GAN

Jonas Adler
KTH – Royal Institute of Technology
Elekta

Sebastian Lunz
University of Cambridge

The Wasserstein metric

- The Wasserstein distance between probability distributions of images on a space B is defined as

$$\mathrm{Wass}(\mathbb{P}_G, \mathbb{P}_r) := \inf_{\pi \in \Pi(\mathbb{P}_G, \mathbb{P}_r)} \mathbb{E}_{(X_1, X_2) \sim \pi} d_B(X_1, X_2).$$

- The Kantorovich-Rubinstein duality provides a way of computing the Wasserstein distance more efficiently

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

- The dependence of f on the metric is encoded in the Lipschitz condition

$$\mathrm{Lip}(f) \leq 1 \quad \Leftrightarrow \quad |f(x) - f(y)| \leq d_B(x, y).$$

- In Wasserstein GANs, we approximate the function f in the Kantorovich duality with a neural network D.
- We train neural network G as Generator, using Wasserstein distance between ground truth and generated image distribution as loss.
- The theory holds in any Polish (e.g. separable completely metrizable) space, but in practice everyone uses $B = L^2$.
- To generalize from L^2, we need to enforce 1-Lipschitz constraint on D in a more general setting.

Banach Spaces

- Banach spaces can be used to model images.
- Banach space B consists of a vector space and a norm $\|\cdot\|$ that defines a notion of length on B.
- The *dual* space B^* is the the space of all bounded linear functionals $B \to \mathbb{R}$, equipped with the norm

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

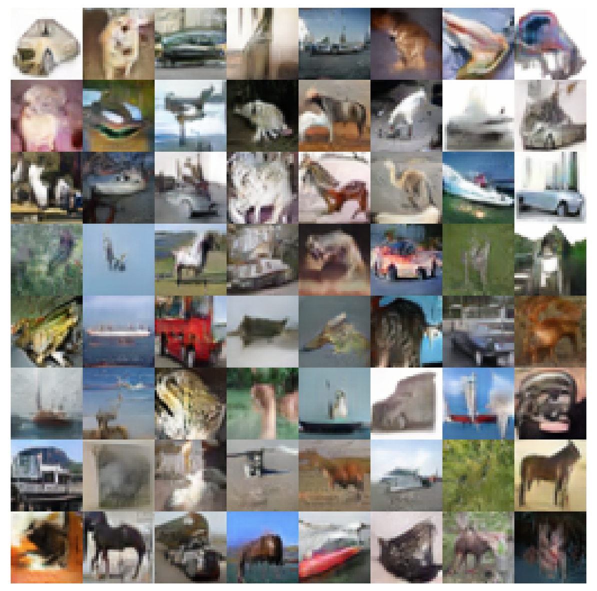
- Classical Banach spaces include *Sobolev spaces* $W^{s,p}$.

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

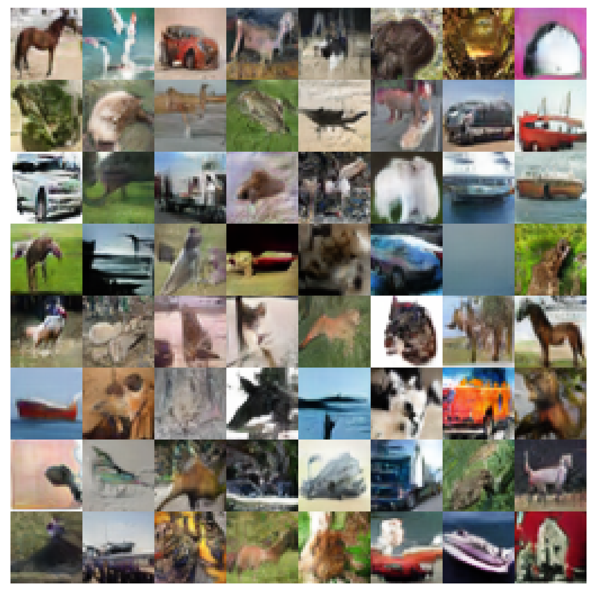
- For any $s, p \geq 1$, define

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

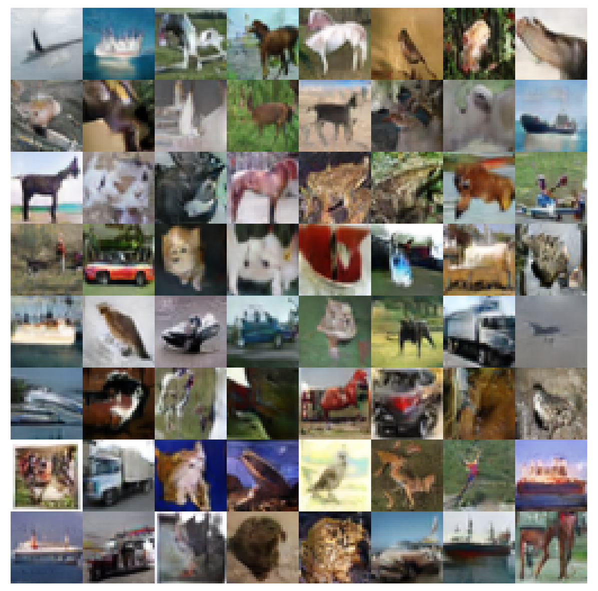
- The parameter p controls the emphasis on outliers, with higher values corresponding to a stronger focus on outliers.
- A negative value of s corresponds to amplifying low frequencies, prioritizing the global structure of the image. High values of s amplify high frequencies, putting emphasis on sharp local structures, like the edges or ridges.



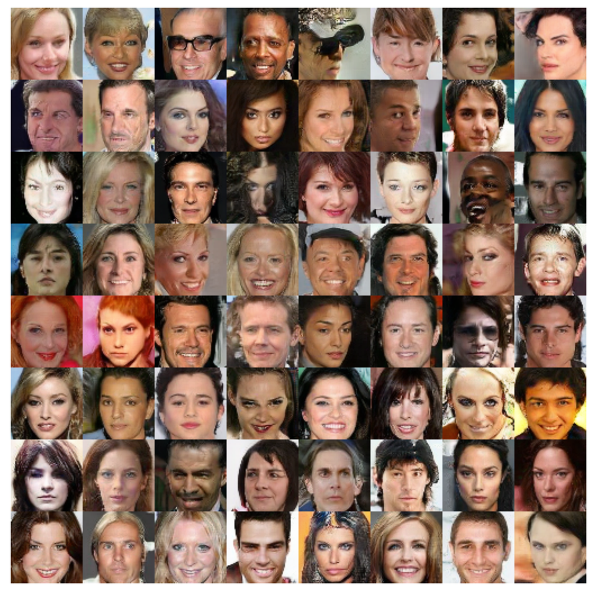
$p = 1.3$



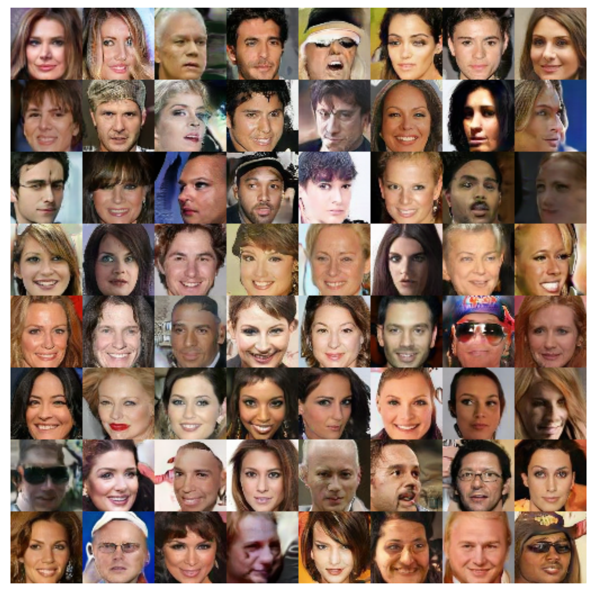
$p = 2.0$



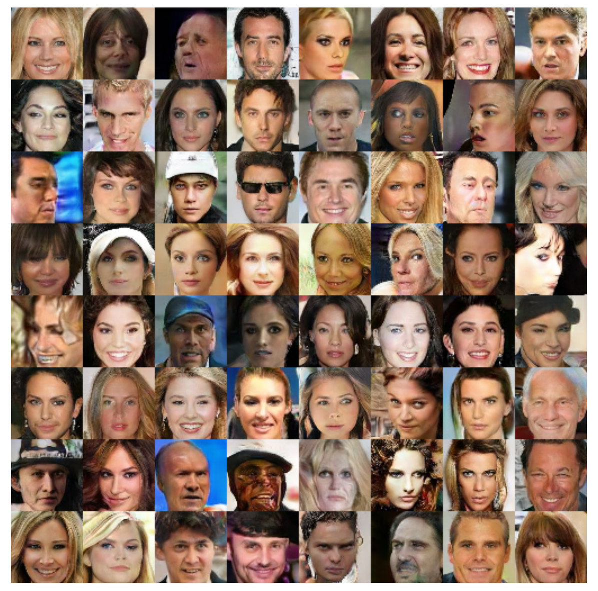
$p = 10.0$



$s = -2$



$s = 0$



$s = 2$

Lipschitz constraint in Banach spaces

A function f is called Fréchet differentiable at $x \in B$ if there exists $\partial f(x) \in B^*$ such that

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

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

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

Implementation

- The loss used to train the critic D in Banach Wasserstein GANs

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

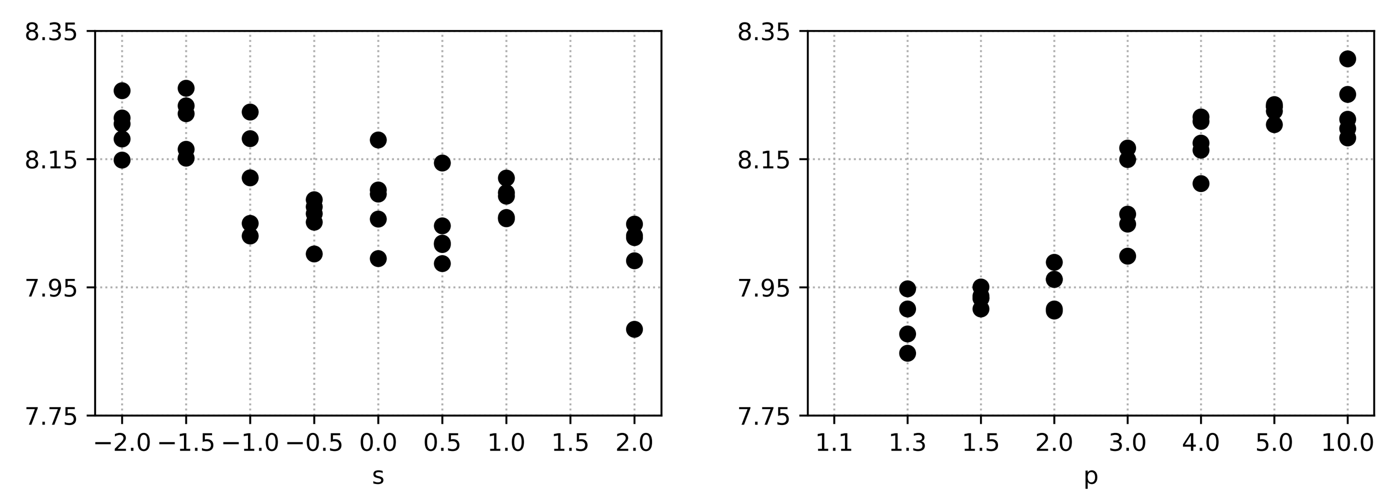
- If a closed form for the dual norm is available, $\|\partial D(\hat{X})\|_{B^*}$ can be computed using readily available automatic differentiation software at no performance loss.
- Heuristics for parameter choices can be built on the assumption that D is scale preserving on the deepest point $x \to \partial D(x)$, leading to

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

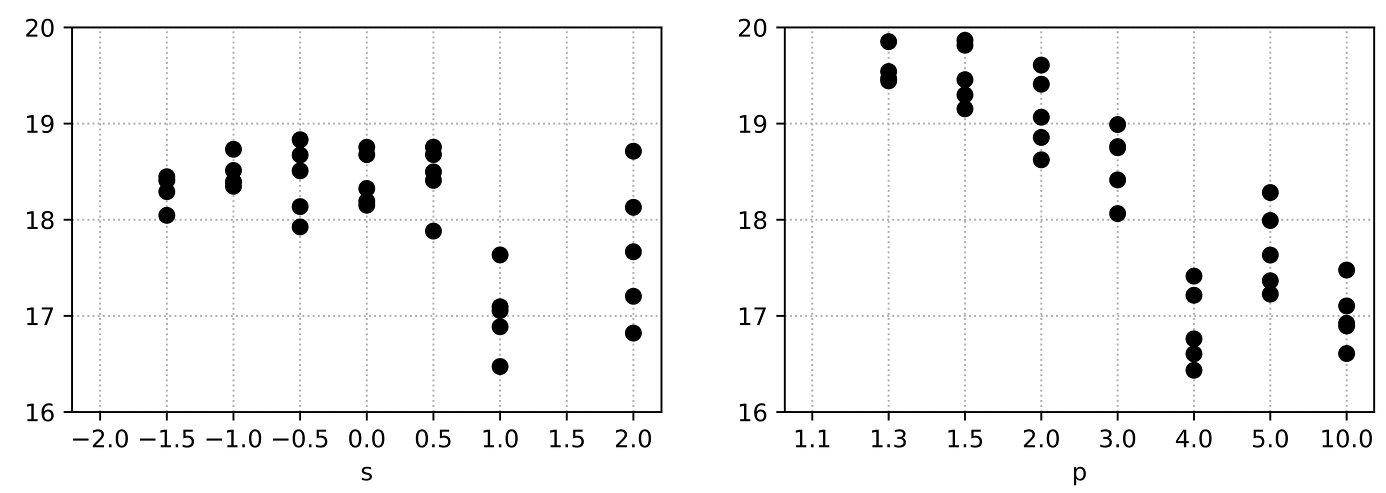
Evaluation

- We evaluate FID and Inception Scores using a range of norms.
- We observe an improvement for high p (focusing on outliers) and for low s (focusing on large scales).
- Similar results were observed on CelebA.

Inception scores on CIFAR-10



FID scores on CIFAR-10



- In order to visually assess the impact of the choice of norm, we plot the Fréchet derivatives ∂D of the discriminator.