# Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport?

Korotin Alexander [*,1,2], Kolesov Alexander [*,2] & Burnaev Evgeny [1,2]

[1] *Skolkovo Institute of Science and Technology*, [2] *Artificial Intelligence Research Institute (AIRI)*,

[*] *Equal contribution*

**Question :** How precisely do WGANs solve the OT problem?

## Background on Optimal Transport .

For distributions $\mathbb{P},\mathbb{Q}$ on $\mathbb{R}^D$ the Monge's form of the Wasserstein-1 distance is

$$\mathbb{W}_1(\mathbb{P},\mathbb{Q}) = \min_{T\sharp\mathbb{P}=\mathbb{Q}} \int \|x - T(x)\|_2 d\mathbb{P}(x).$$

The Kantorovich's relaxation is given by

$$\mathbb{W}_1(\mathbb{P},\mathbb{Q}) = \min_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \int_{\mathbb{R}^D\times\mathbb{R}^D} \|x - y\|_2 d\pi(x,y).$$

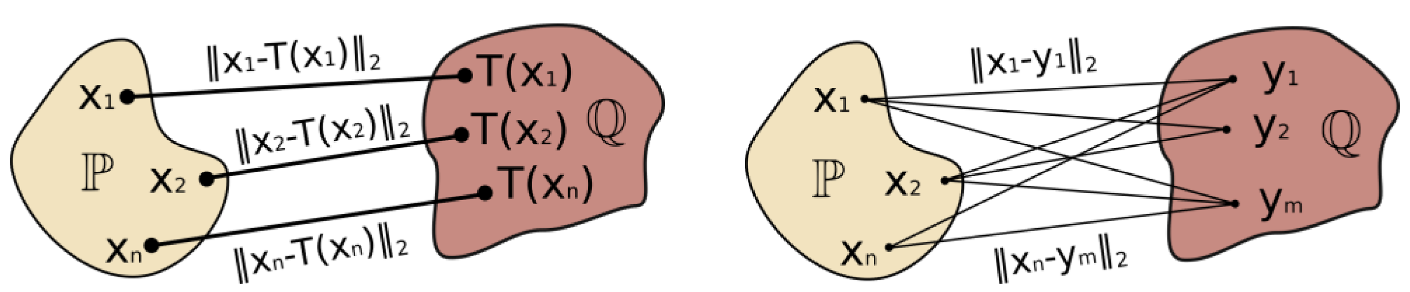



Figure 1. Monge's and Kantorovich's OT formulations.

## Dual formulation of $\mathbb{W}_1$ .

For distributions $\mathbb{P},\mathbb{Q}$, the dual formulation of $\mathbb{W}_1$ is given by

- Conventional (⌊**LS**⌉):

$$\mathbb{W}_1(\mathbb{P},\mathbb{Q}) = \max_{f\oplus g\leq\|\cdot\|_2} \int f(x)d\mathbb{P}(x) + \int g(y)d\mathbb{Q}(y);$$

- c-transform (⌊**MM:B**⌉, ⌊**MM:Bv2**⌉, ⌊**MM**⌉, ⌊**MM:R**⌉):

$$\mathbb{W}_1(\mathbb{P},\mathbb{Q}) = \max_f \int f(x)d\mathbb{P}(x) + \int \min_{x\in\mathbb{R}^D}\{\|x-y\|_2 - f(x)\}d\mathbb{Q}(y);$$

- 1-Lipschitz constraint (⌊**WC**⌉, ⌊**GP**⌉, ⌊**LP**⌉, ⌊**SN**⌉, ⌊**SO**⌉):

$$\mathbb{W}_1(\mathbb{P},\mathbb{Q}) = \max_{\|f\|_L\leq 1} \int f(x)d\mathbb{P}(x) - \int f(y)d\mathbb{Q}(y).$$

Typically, the dual potential is recovered with neural nets which are trained with SGD.

## Optimal Transport in GANs .

Let $\mathbb{P} = \mathbb{P}_\alpha$ is a parametric distribution and $\mathbb{Q}$ is the data distribution. Typically, $\mathbb{P}_\alpha = G_\alpha\sharp\mathbb{S}$ obtained by a generator network $G_\alpha$ from a fixed latent $\mathbb{S}$.

The **loss** function for the generator $G_\alpha$ is

$$\mathbb{W}_1(\mathbb{P}_\alpha,\mathbb{Q}) = \int_z f^*(G_\alpha(z))d\mathbb{S}(z) - \int f^*(y)d\mathbb{Q}(y).$$

The derivative of the loss is

$$\frac{\partial\mathbb{W}_1(\mathbb{P}_\alpha,\mathbb{Q})}{\partial\alpha} = \int_z J_\alpha G_\alpha(z)^T \nabla f^*(G_\alpha(z))d\mathbb{S}(z).$$

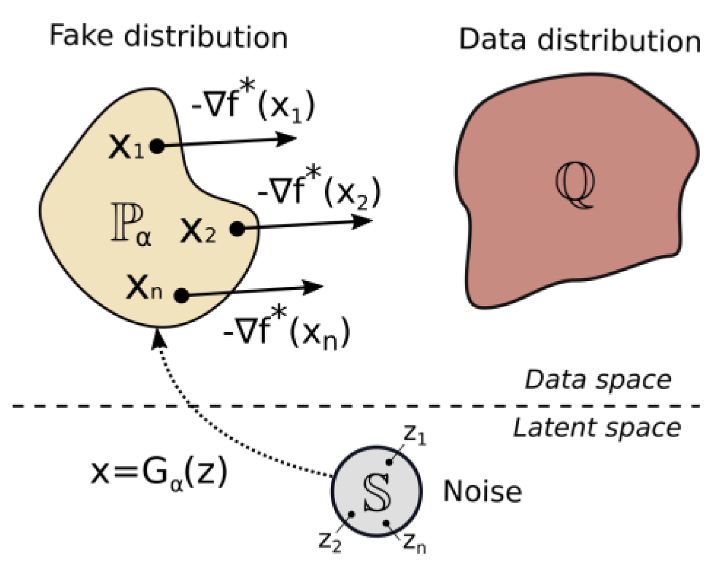



Figure 2. The anti-gradient $-\nabla f^*(x)$ shows where to move the mass of each $x = G_\alpha(z)$ to make the generated $\mathbb{P}_\alpha$ closer to $\mathbb{Q}$ in $\mathbb{W}_1$.

**Problem:** There are no non-trivial pairs $\mathbb{P},\mathbb{Q}$ with known OT gradient $\nabla f^*$.

## Constructing benchmark pairs.

We propose a way to construct pairs of $\mathbb{P},\mathbb{Q}$ with a known OT map, cost, gradient.

**1: Definition**

For a 1-Lipschitz function $u:\mathbb{R}^D\to\mathbb{R}$, we say that a transport plan $\pi\in\Pi(\mathbb{P},\mathbb{Q})$ is $u$-*ray monotone* (decreasing) if $u(x)-u(y) = \|x-y\|_2$ holds $\pi$-almost surely for $x,y\in\mathbb{R}^D$.

**2: Proposition**

Let $\pi\in\Pi(\mathbb{P},\mathbb{Q})$ be a $u$-ray monotone transport plan for a 1-Lipschitz function $u$. Then it is an optimal plan between $\mathbb{P},\mathbb{Q}$.

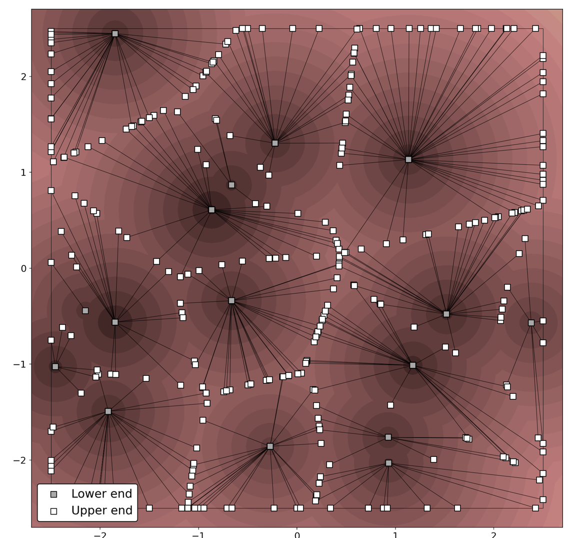
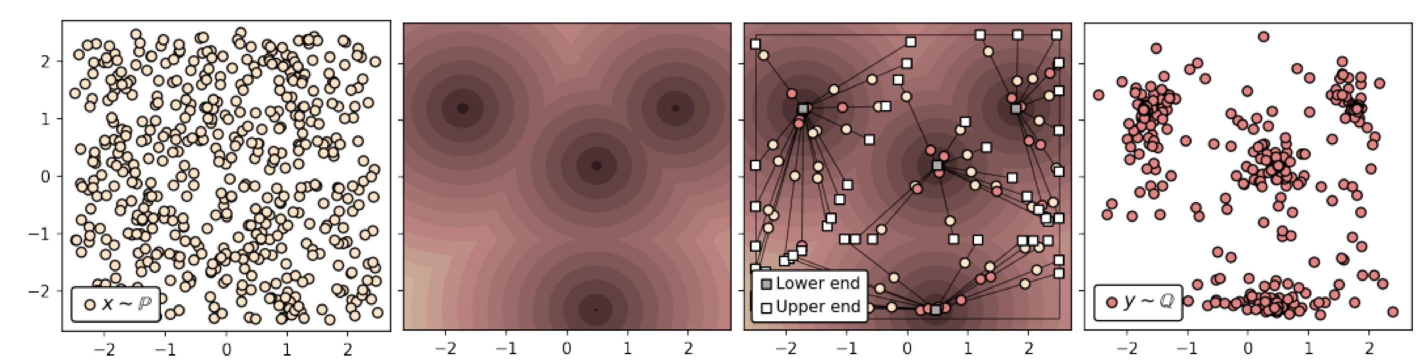


Figure 3. Truncated transport rays of a random MinFunnel $N = 16, D = 2$.

**Parameterizing 1-Lipschitz functions:**

$u:\mathbb{R}^D\to\mathbb{R}$, we employ the following 1-Lipschitz *MinFunnel* functions

$$u(x) = \min_n\{\|x-a_n\|_2 + b_n\}.$$



(a) Samples $x\sim\mathbb{P}$. (b) 1-Lipshitz potential $u:\mathbb{R}^D\to\mathbb{R}$. (c) Transport rays for $x$ and map $y = T(x)$. (d) Pushforward samples $y = T(x)\sim\mathbb{Q}$.

Figure 4. The pipeline of constructing of benchmark pairs.

**The construction pipeline:**

- Pick absolutely continuous distribution $\mathbb{P}$.
- Take 1-Lipschitz MinFunnel function $u(x)$.
- Find $u$'s transport rays, $u$-forward map.
- Move samples from $\mathbb{P}$ to $\mathbb{Q}$ by the map.

## High-dimensional and images benchmark pairs.

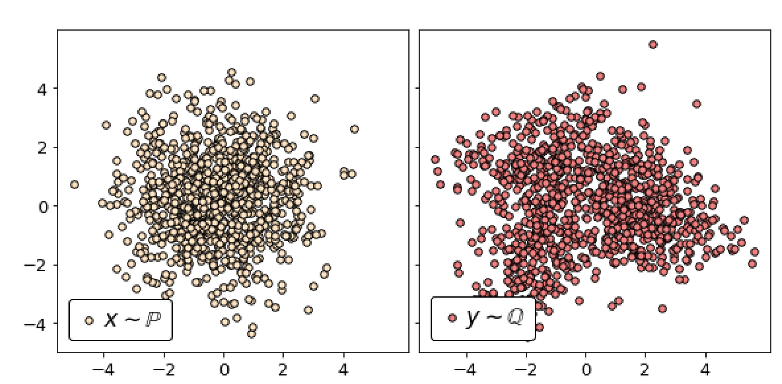


Figure 5. The pair with $N = 4, D = 32$.

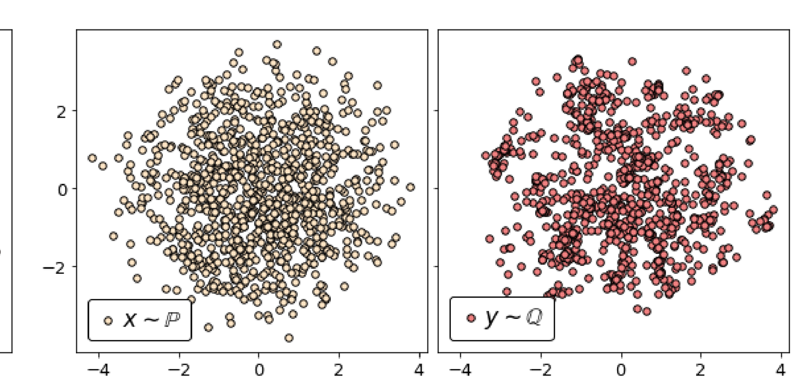


Figure 6. The pair with $N = 64, D = 4$.

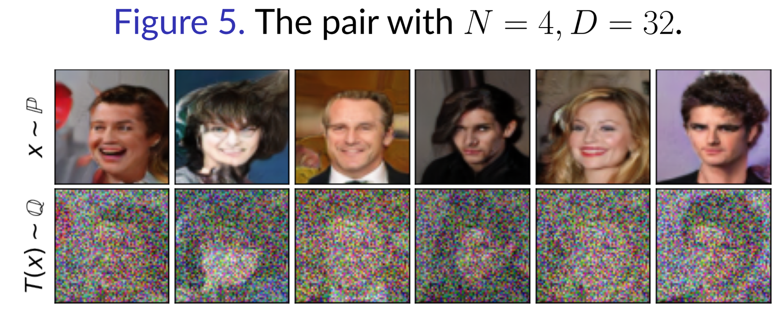


Figure 7. Samples of CelebA from the pair with $(N,p) = (1,10)$.

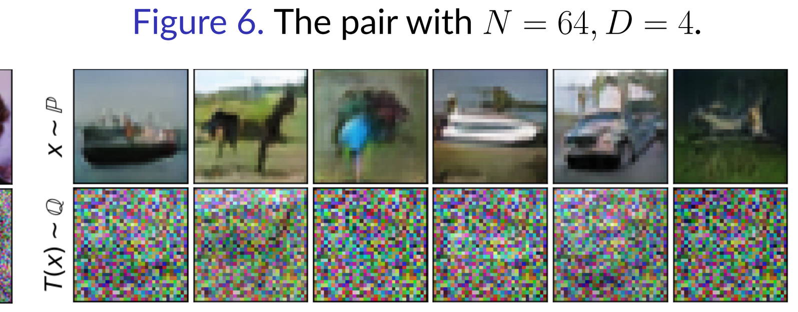


Figure 8. Samples of CIFRA-10 from the pair with $(N,p) = (1,10)$.

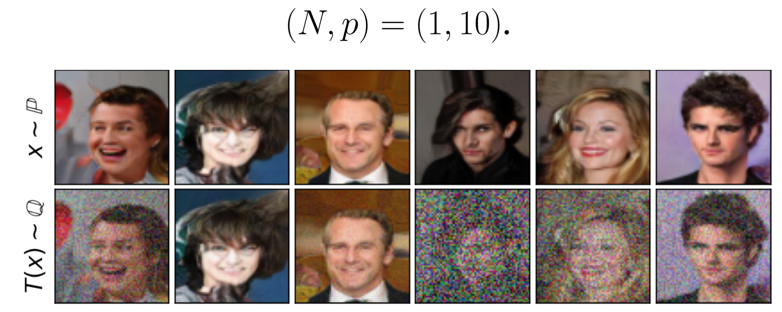


Figure 9. Samples of CelebA from the pair with $(N,p) = (16,100)$.

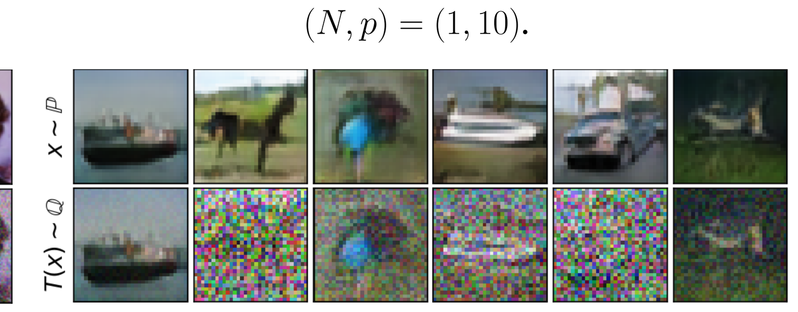
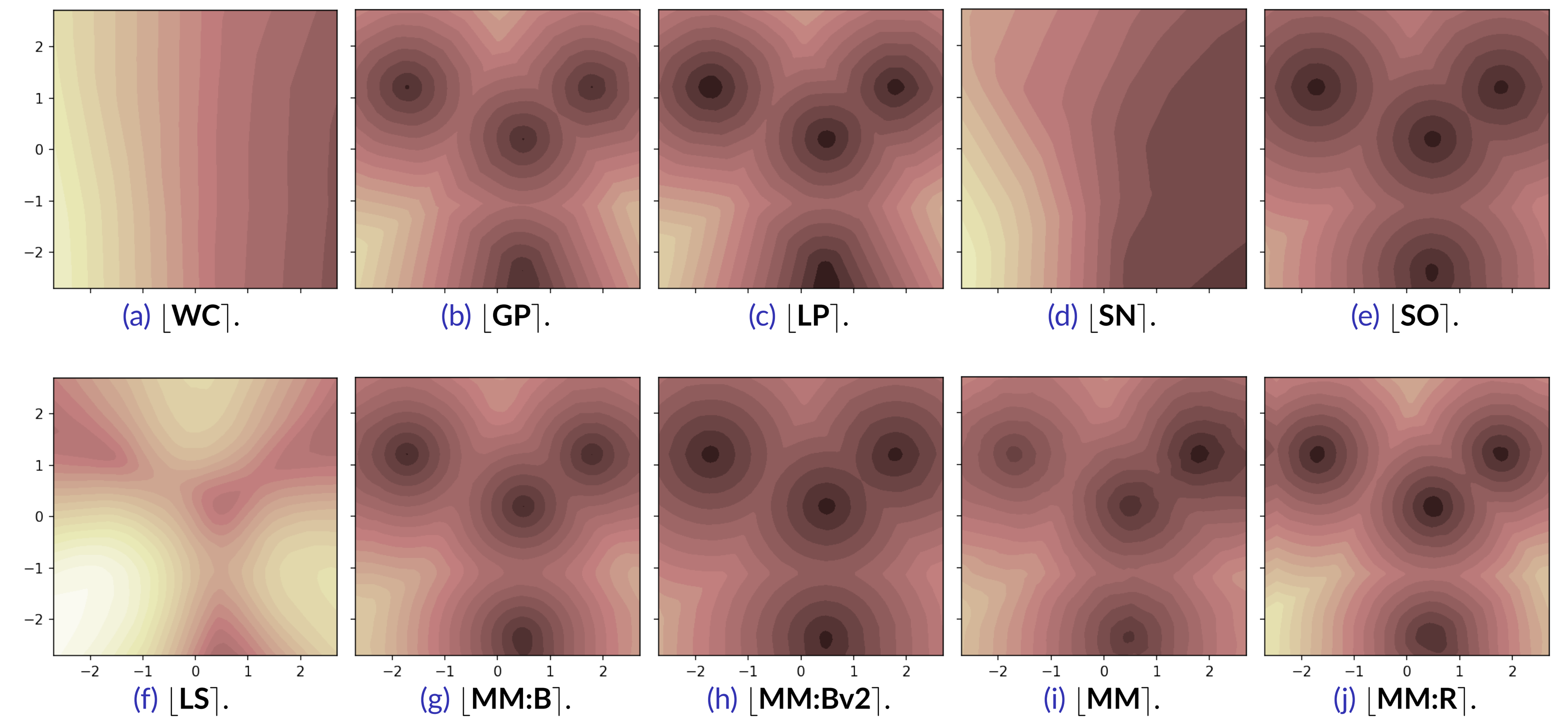


Figure 10. Samples of CIFAR-10 from the pair with $(N,p) = (16,100)$.

## Evaluating dual surfaces in 2D.



(a) ⌊WC⌉. (b) ⌊GP⌉. (c) ⌊LP⌉. (d) ⌊SN⌉. (e) ⌊SO⌉. (f) ⌊LS⌉. (g) ⌊MM:B⌉. (h) ⌊MM:Bv2⌉. (i) ⌊MM⌉. (j) ⌊MM:R⌉.

Figure 11. Surfaces of potentials $f_\theta$ learned by OT solvers on the pair with $D = 2, N = 4$. The ground truth optimal potential is in Figure 4.

## Qualitative comparison of dual solvers.

- $\mathcal{L}^2$ - metric:

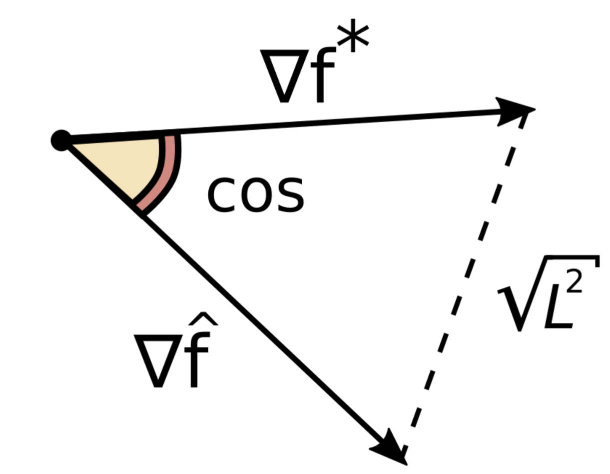
$$\mathcal{L}^2(\nabla\hat{f},\nabla f^*) = \|\nabla\hat{f} - \nabla f^*\|_2^2;$$

- Cosine similarity:

$$\cos(\nabla\hat{f},\nabla f^*) = \frac{\langle\nabla\hat{f},\nabla f^*\rangle}{\|\nabla\hat{f}\|_2\|\nabla f^*\|_2};$$

- Deviation of $\hat{\mathbb{W}}_1$ from $\mathbb{W}_1$:

$$\text{dev} = 100\% \cdot \frac{|\mathbb{W}_1 - \widehat{\mathbb{W}}_1|}{\mathbb{W}_1}.$$




## Results.

| | ⌊WC⌉ | ⌊GP⌉ | ⌊LP⌉ | ⌊SN⌉ | ⌊SO⌉ | ⌊LS⌉ | ⌊MM:B⌉ | ⌊MM:Bv2⌉ | ⌊MM⌉ | ⌊MM:R⌉ |
|---|---|---|---|---|---|---|---|---|---|---|
| HD | 0.11 | 0.56 | 0.56 | 0.09 | 0.61 | -0.51 | -0.54 | 0.45 | 0.53 | 0.31 |
| Img | 0.33 | 0.92 | 0.92 | 0.35 | - | 0.36 | 0.38 | 0.09 | 0.46 | 0.91 |

Table 1. The top row demonstrates $cos$ for dual solvers for pairs with $N = 256, D = 128$, while the bottom row shows $cos$ for image pairs $N = 16, p = 100$

| | ⌊WC⌉ | ⌊GP⌉ | ⌊LP⌉ | ⌊SN⌉ | ⌊SO⌉ | ⌊LS⌉ | ⌊MM:B⌉ | ⌊MM:Bv2⌉ | ⌊MM⌉ | ⌊MM:R⌉ |
|---|---|---|---|---|---|---|---|---|---|---|
| HD | $\gg$10 | 0.89 | 0.9 | 1.57 | 0.77 | 2.07 | 1.98 | $\gg$10 | 0.73 | 1.38 |
| Img | $\gg$10 | 1.34 | 1.12 | 3.32 | - | 1.87 | 1.28 | $\gg$10 | 0.16 | 0.95 |

Table 2. The top row demonstrates $L^2$ metric for dual solvers for pairs with $N = 256, D = 128$, while the bottom row shows $\mathcal{L}^2$ metric for image pairs $N = 16, p = 100$

| | ⌊WC⌉ | ⌊GP⌉ | ⌊LP⌉ | ⌊SN⌉ | ⌊SO⌉ | ⌊LS⌉ | ⌊MM:B⌉ | ⌊MM:Bv2⌉ | ⌊MM⌉ | ⌊MM:R⌉ |
|---|---|---|---|---|---|---|---|---|---|---|
| HD | 0.36 | 0.67 | 0.69 | 0.12 | 0.8 | -0.32 | -0.32 | 2.61 | 0.79 | 1.02 |
| Img | $\gg$100 | 36.16 | 36.75 | 13.26 | - | 5.24 | 4.21 | 24.36 | $\gg$100 | 23.46 |

Table 3. The top row demonstrates $dev$ metric for dual solvers for pairs with $N = 64, D = 128$, while the bottom row shows $dev$ for image pairs $N = 16, p = 100$

## Conclusion.

- All the solvers should **not** be considered as **meaningful estimators of** $\mathbb{W}_1$.
- The gradient $\nabla\hat{f}$ recovered by some solvers shows **positive** $cos$ with the ground-truth gradient $\nabla f^*$, but mostly they provide **poor** $\mathcal{L}^2$ metric.

More details and code are represented in GitHub page.

GitHub

https://github.com/justkolesov/Wasserstein1Benchmark