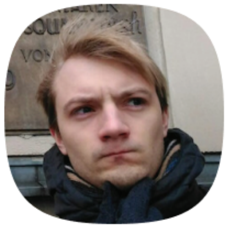
Vladislav Belavin, Maxim Borisyak



Introduction to distances: Wasserstein

2021

NATIONAL RESEARCH UNIVERSITY

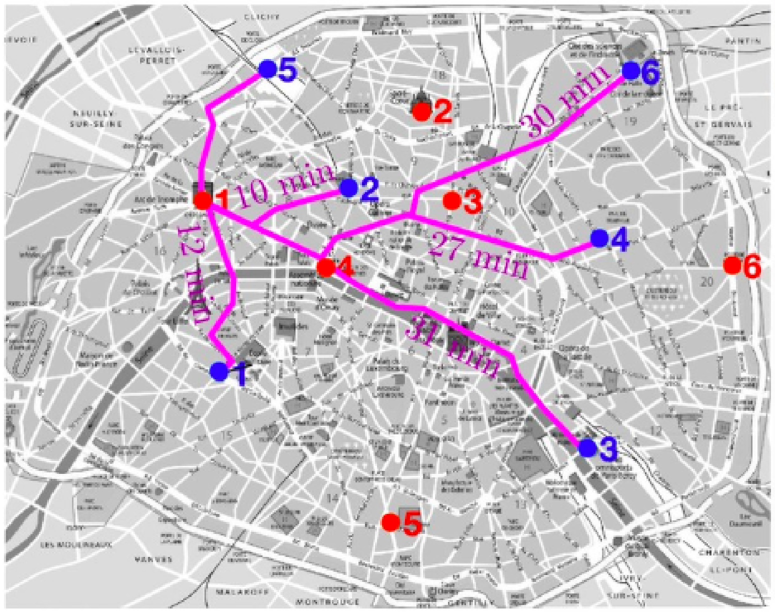
LAMBDA • HSE

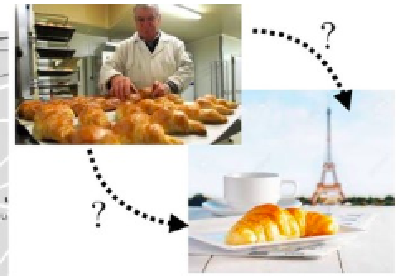
SCHOOL OF DATA ANALYSIS

Schaffhausen Institute of Technology

Motivation

French Bakeries



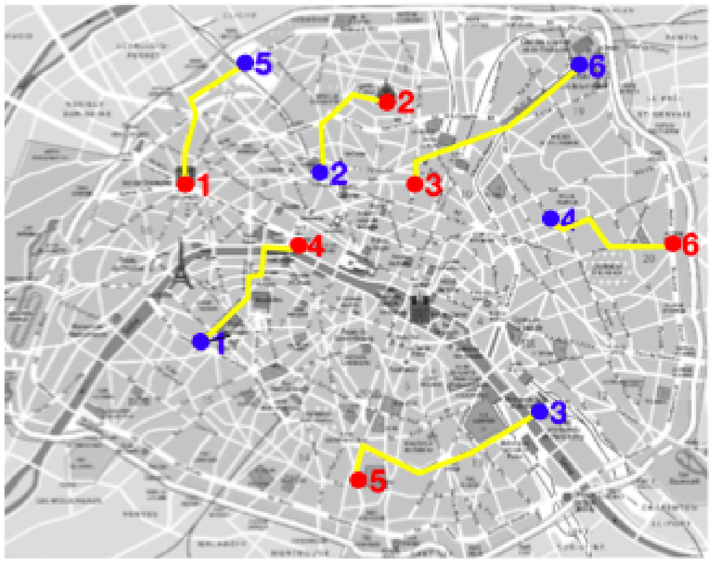


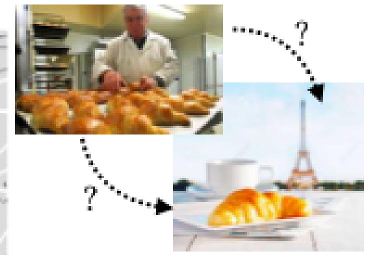
c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15

Given a set of N bakeries and M cafes, what is the optimal way to transport loaves of bread between them?

http://www.gpeyre.com/

French Bakeries



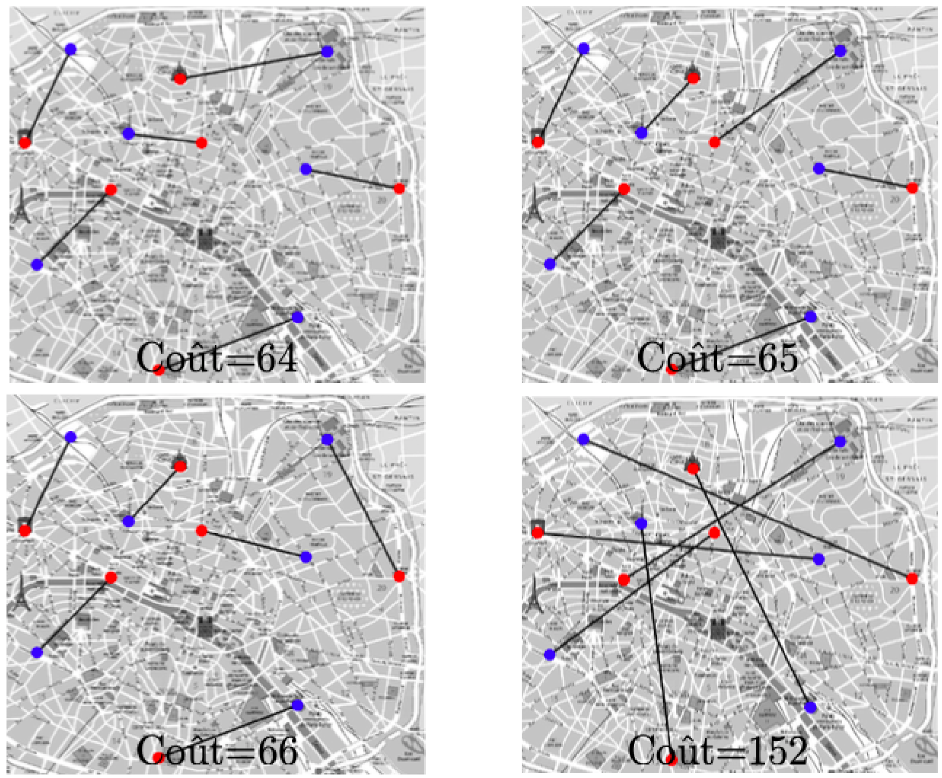


c_{ij}	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	9	34	15

Price = 10+7+15+10+14+9 = 65 min

http://www.gpeyre.com/

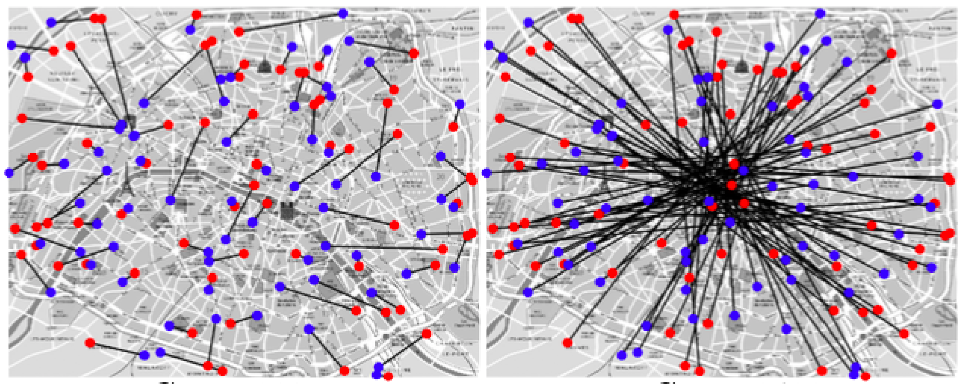
More Bakeries



We can estimate different possibilities, using the same matrix of costs.

Optimal Transport: Monge

The number of calculations rises as factorial.



Cout=200 Cout=1493

We thus need to solve a problem:

$$\min_{\sigma \in \mathrm{Perm}_n} \sum_{i=1}^{n} C_{i,\sigma(i)}$$

Earth Mover's (EM) distance

Formulate the Bakery problem: Kantorovich

What if bakeries can produce different mass of breads?

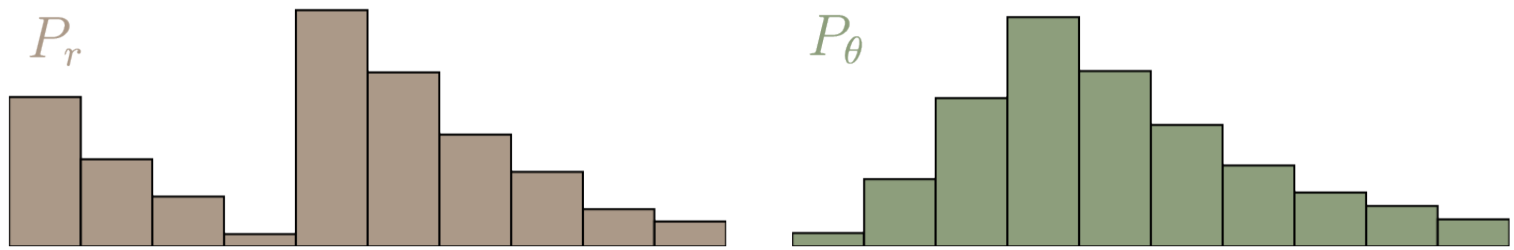
- Let:
  - $p_i, \ \ i \in 1...N$ the mass of bread held by each bakery;
  - $q_j, \ \ j \in 1...M$ the mass of bread desired by each cafe;
  - x_i, y_j the positions of bakeries and cafes;
  - $\sum_i p_i = \sum_j q_j = 1$, and cost is proportional to work (mass$\times$distance).

Find an optimal coupling $\gamma_{i,j}$ – a quantity of how much bread is delivered from bakery i to cafe j – for the mass of bread moved from p_i to q_j. This defines the Earth Mover's (EM) distance:

$$\text{EMD} = \inf_{\gamma \in \Pi} \sum_{x,y} ||x - y|| \gamma(x, y) = \inf_{\gamma \in \Pi} \mathop{\mathbb{E}}_{(x,y) \sim \gamma} ||x - y||.$$

Why EMD?

Imagine that we want to move the events from P_r to P_θ. We also want to save effort, that is, not to move large pieces over long distances.



This is a problem that is solved by many construction workers every day. In fact, this is the optimal transport problem from P_r to P_θ.

Figure: `https://vincentherrmann.github.io/blog/wasserstein/`

Why EMD?

Imagine that we want to move the events from P_r to P_θ. We also want to save effort, that is, not to move large pieces over long distances.

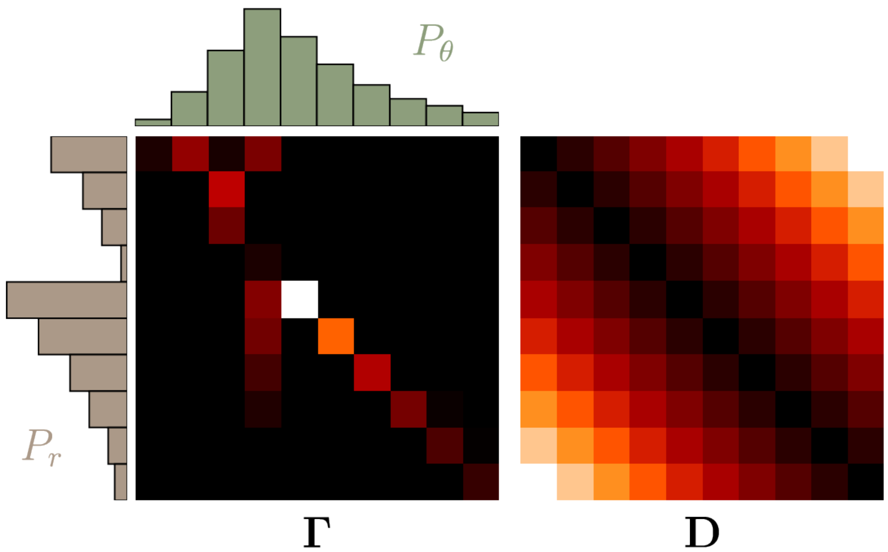
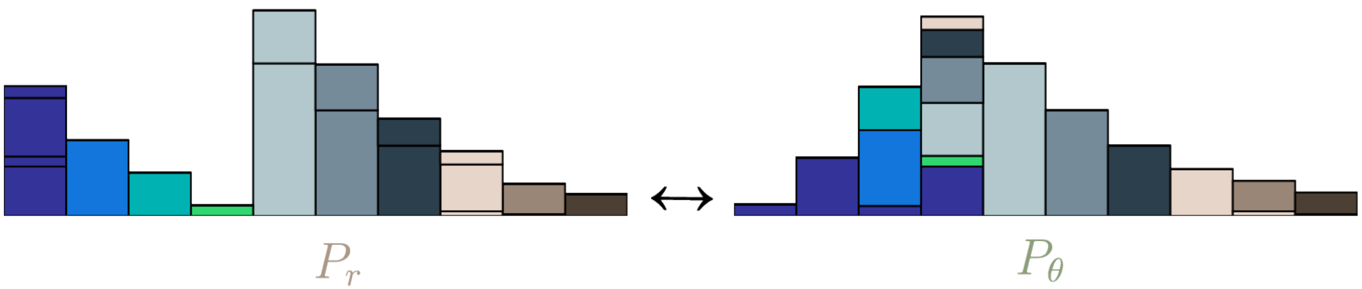


Figure: `https://vincentherrmann.github.io/blog/wasserstein/`

Why EMD?

Imagine that we want to move the events from P_r to P_θ. We also want to save effort, that is, not to move large pieces over long distances.



$$\mathrm{EMD}(P_r, P_\theta) = \inf_{\gamma \in \Pi} \sum_{x,y} ||x - y|| \gamma(x, y) = \inf_{\gamma \in \Pi} \mathop{\mathbb{E}}_{(x,y) \sim \gamma} ||x - y||,$$

Figure: https://vincentherrmann.github.io/blog/wasserstein/

Wasserstein distance

Wasserstein Distance

For continuous case, there are a set of p-Wasserstein distances, with $W_p(P_x, Q_y)$ defined with $x \in M$, $y \in M$ and a distance D on x, y:

$$W_p(p_x, q_y) = \inf_{\gamma \in \Pi(x,y)} \int D(x, y)^p \gamma(x, y) dxdy,$$

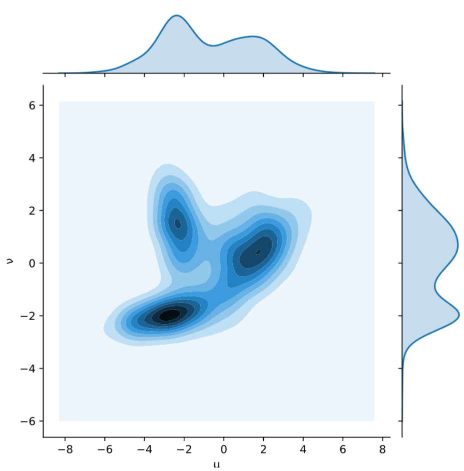
where $\Pi(x, y)$ is a set of all joint distributions having P_x, Q_y as their marginals. And $\gamma(x, y)$ denotes the amount of “mass” to move from x to y.

W_1 Distance

W_1 distance with Euclidean norm is:

$$W(p_x, q_y) = \inf_{\gamma \in \Pi(x,y)} \int D(x,y)\gamma(x,y)dxdy = \inf_{\gamma \in \Pi(x,y)} \mathbb{E}(||x - y||)$$

Which brings an evident connection to EMD.

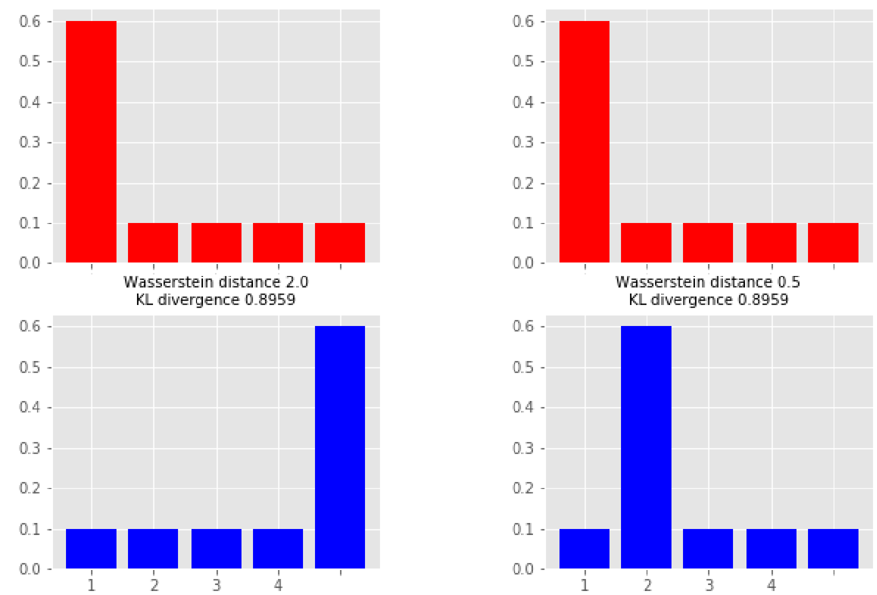


Two dimensional representation of the transport plan between horizontal μ and vertical ν pdfs. Note, that this is not unique plan. The inf must be taken over all possible plans.

Picture: `https://en.wikipedia.org/wiki/Wasserstein_metric`

W vs KL

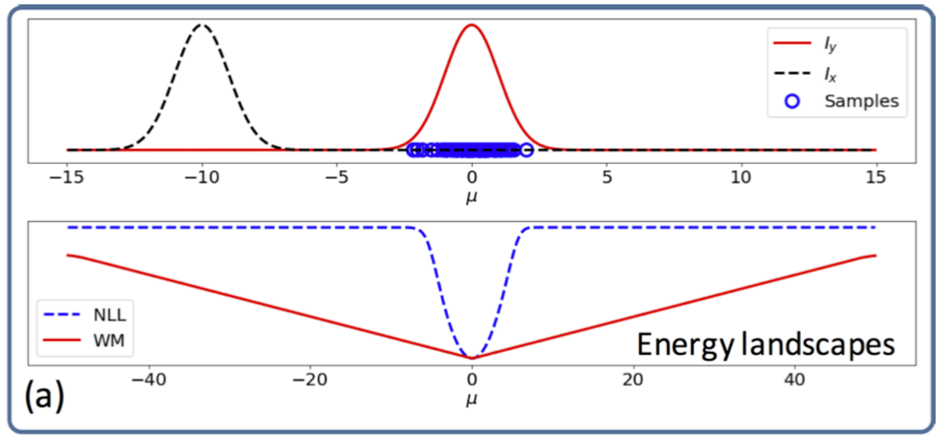
EMD also takes into account the distance at which the differences in the distributions are located.
This is exactly what we need!



From: `https://bit.ly/38BR24M`

W vs KL

Wasserstein loss landscape is less sensitive to the initial point, hence a gradient descent approach would easily converge to the optimal point regardless of the starting point.
Cons: W loss depends on the distance measure, i.e. one more thing to tune.

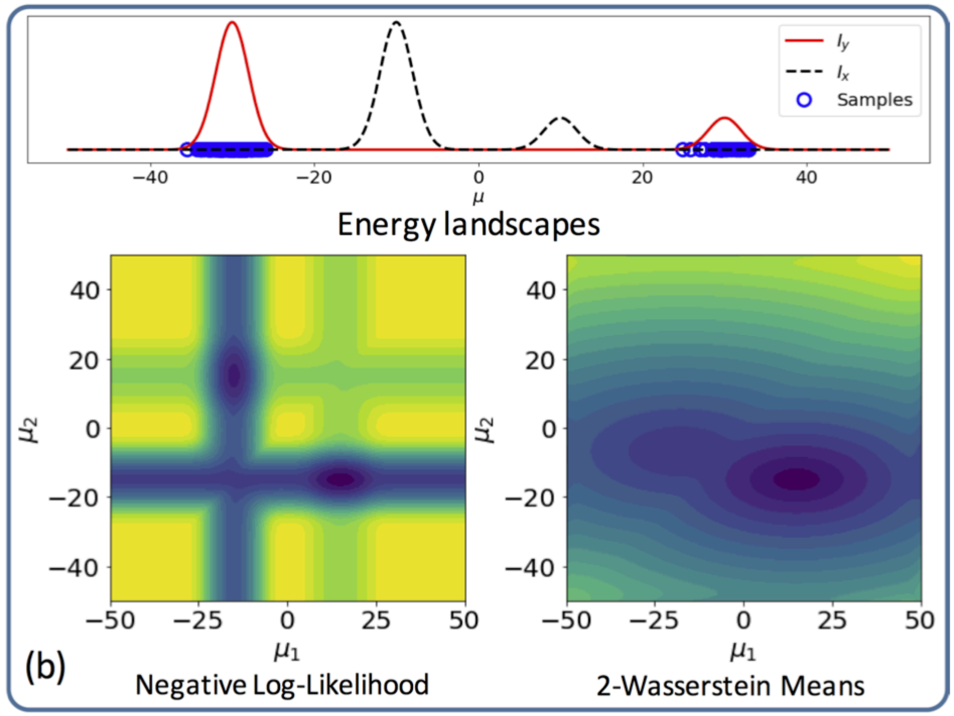


From: `https://arxiv.org/pdf/1711.05376.pdf`

W vs KL

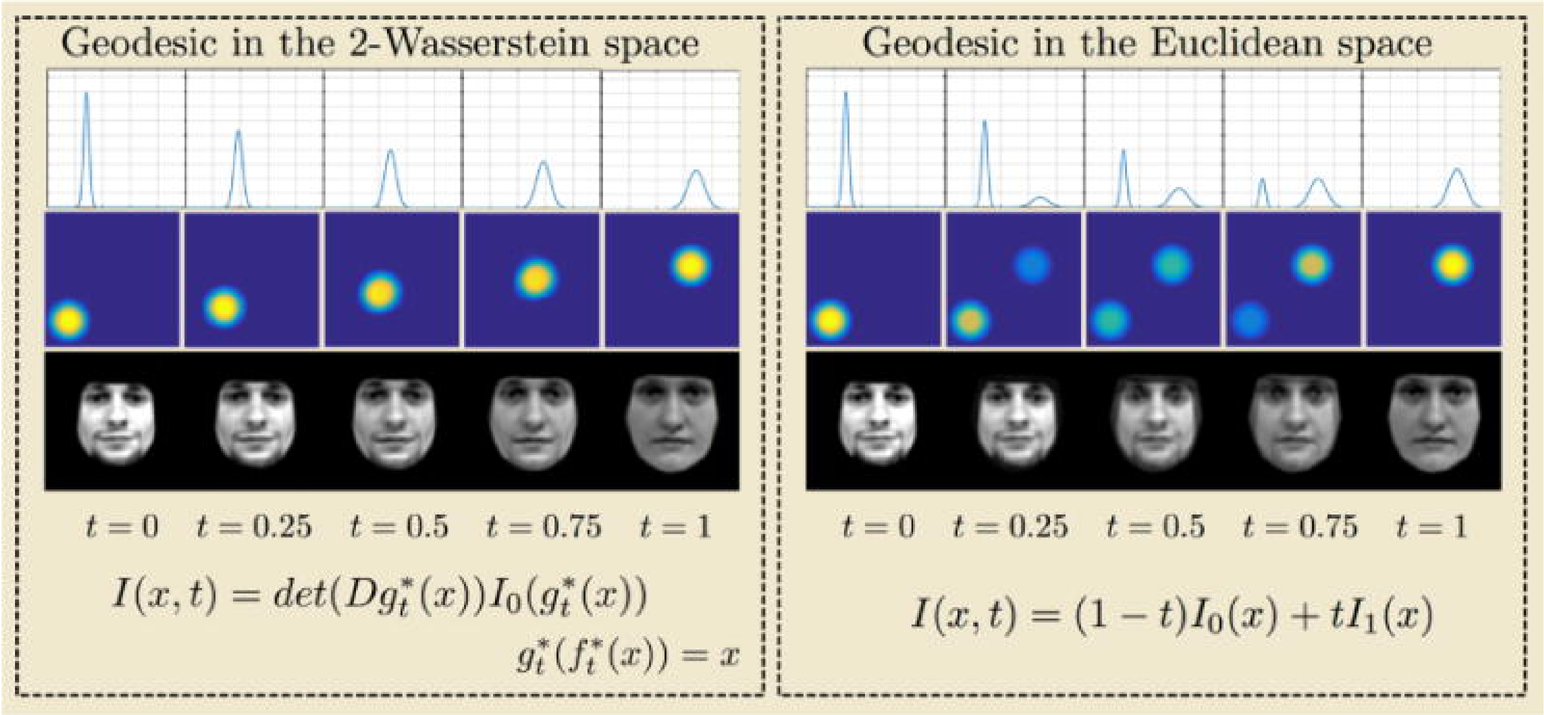
Also, the Wasserstein metric suffers less from the local minima in multimodal cases and is much smoother.

Claim without proof: it is connected to the fact that Wasserstein takes into account the distance at which the differences in the distributions are located.



From: `https://arxiv.org/pdf/1711.05376.pdf`

W_2 for interpolation



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6024256/

Conclusion

Key point (one more time :) There is no single good way to evaluate a generative model. Most likely, the quality metrics should depend on the further use.

- Wasserstein distance does not have zero-gradient problem like KL/JS divergence;
- more robust in multidimensional multimodal cases than KL;
- quite difficult to optimize (dozens of papers are devoted to the optimization problem of Wasserstein distance).

Thank you for your attention!

Vladislav Belavin
vbelavin@hse.ru
SchattenGenie
hse_lambda