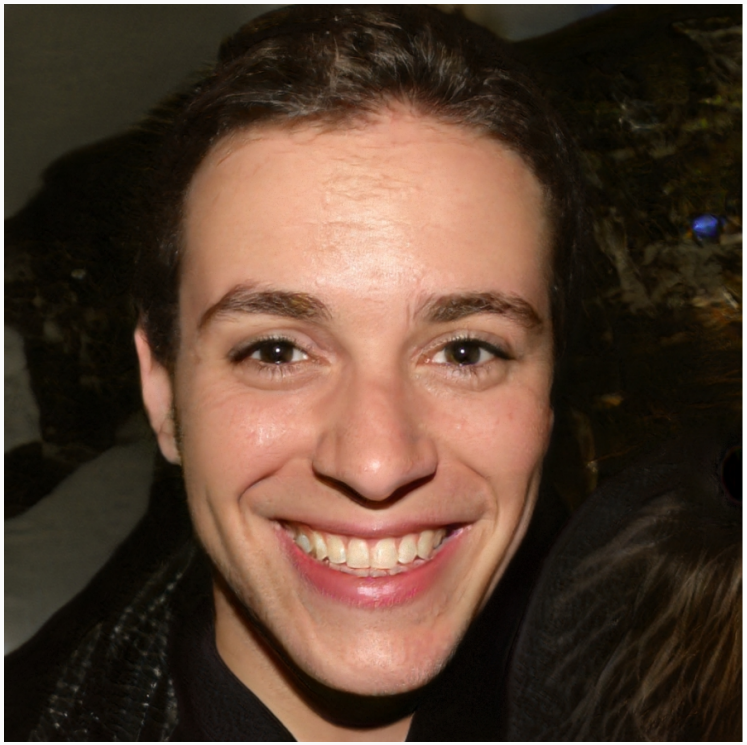
Generative Adversarial Networks: theory and practice.

Ugo Tanielian

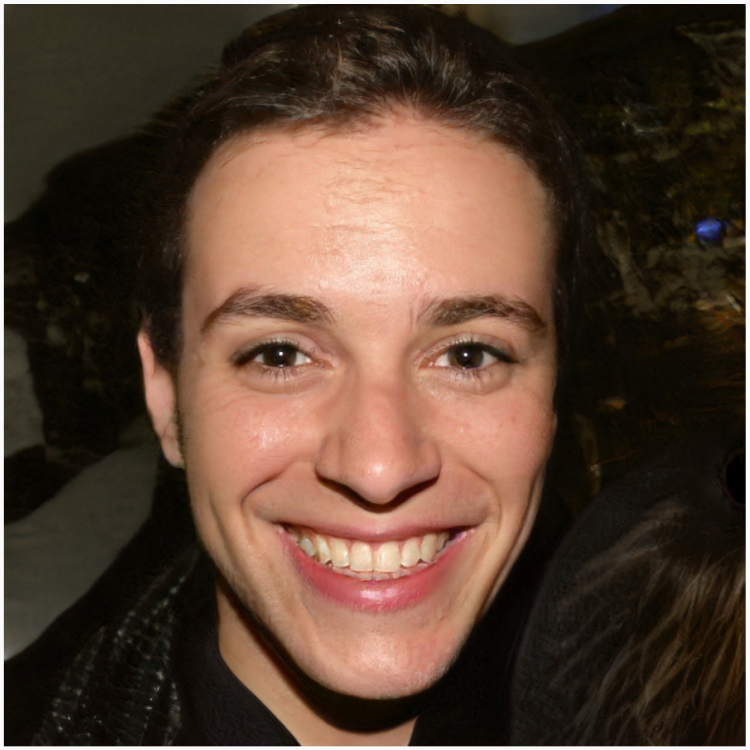
Criteo, Sorbonne Universite, Rennes Universite

Let's play a game! (1): Generative Adversarial Networks



Which face is real.

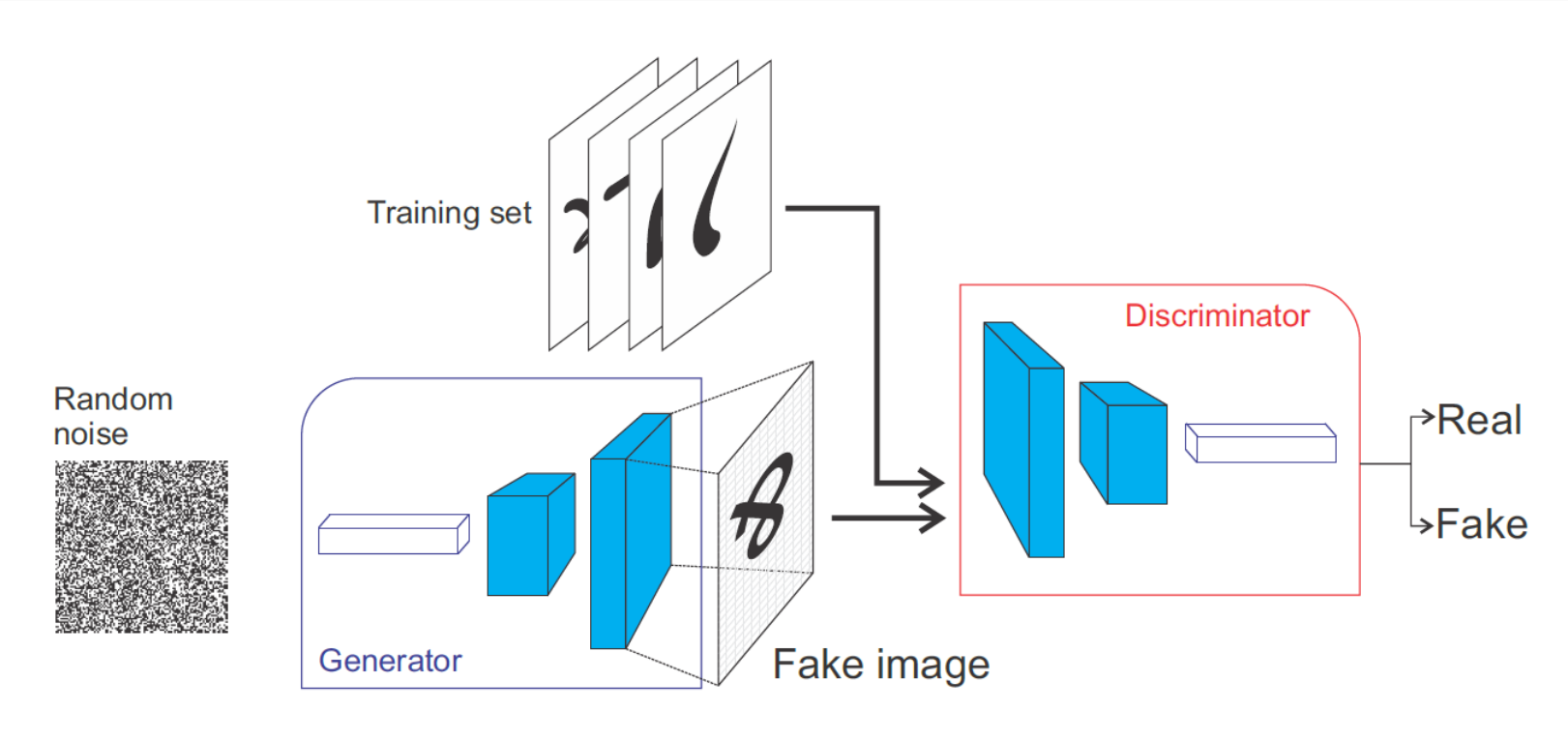
Let's play a game (2): Generative Adversarial Networks



Which face is real.

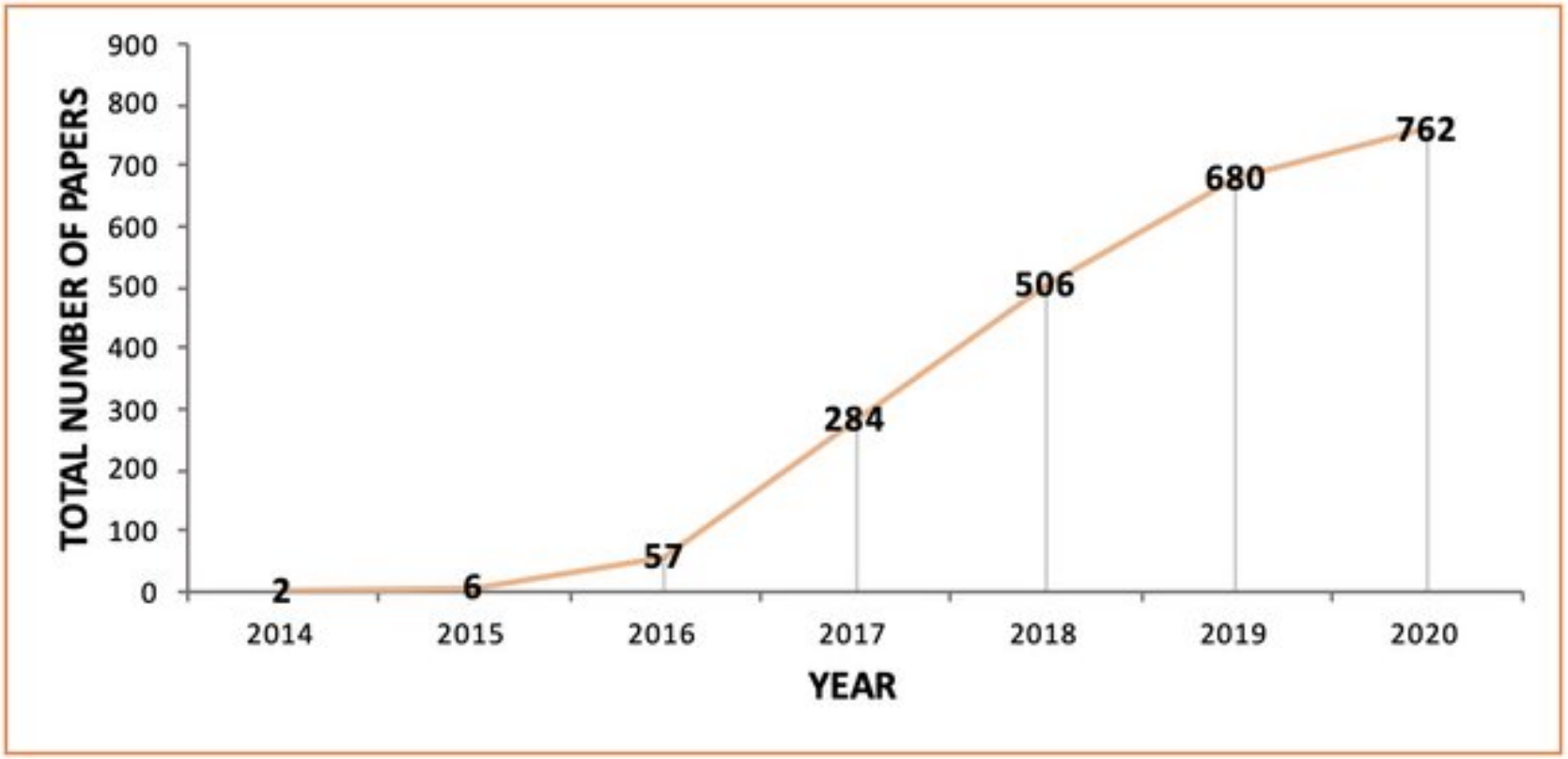
Quick introduction to GANs

Presentation: Generative Adversarial Networks



Source: medium.

The GAN Zoo



Source: researchgate.

Motivation

Generative models aim at generating artificial contents (with randomness).

Pros

- Simple generation.
- Work extremely well with high-dimensional data.
- Allow manifold discovering: image interpolation.

[Abdal et al., 2019].

Cons

- Unknown probability density function: we cannot easily check low density areas.
- Tricky training.

Outstanding image generation: human faces

This person does not exist.

Merchandising: virtual try on problem.

vue.ai.

Art: Edmond de Belamy.

https://en.wikipedia.org/wiki/Edmond_de_Belamy

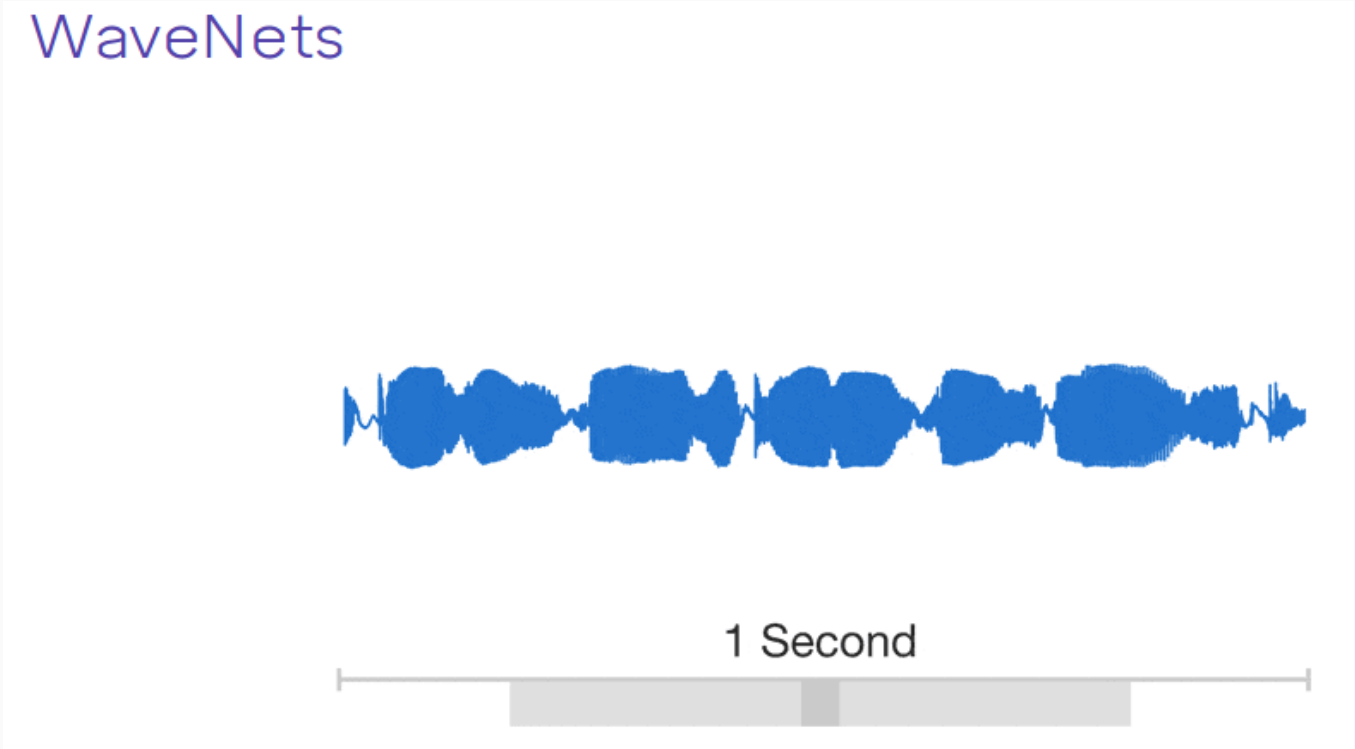
Interactive image generation.

GAN paint studio, [Bau et al., 2020].

Other solutions:

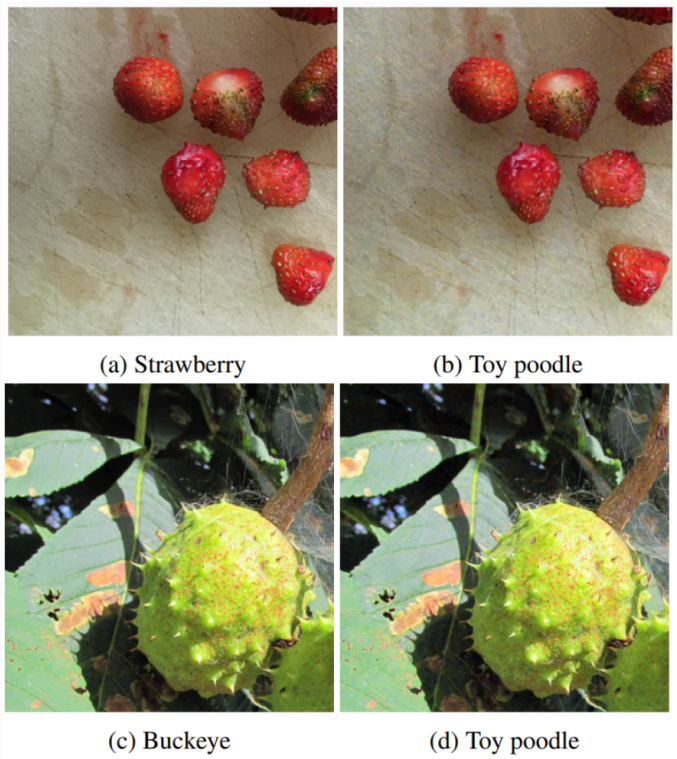
- Interactive GAN [Zhu et al., 2016],
- GauGANs by NVIDIA [Park et al., 2019].

Speech synthesis

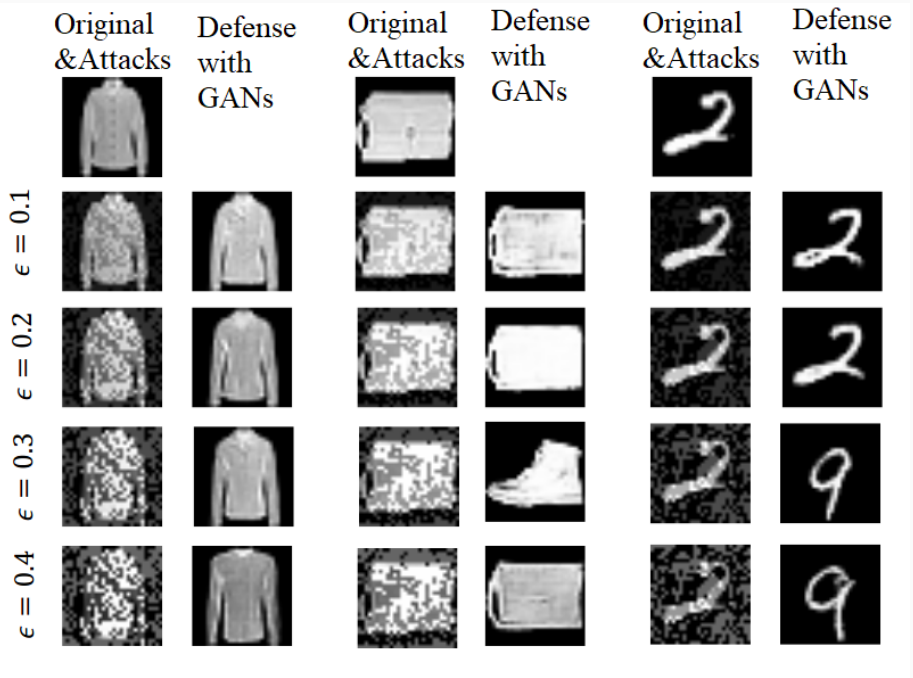


WaveNet by DeepMind.

GANs for robustness



(a) Attacking deep nets with GANs: [Xiao et al., 2018].



(b) Defending deep nets with GANs: [Samangouei et al., 2018].

Last but not least: GANs for physics

- Using GANs to solve SDEs [Yang et al., 2018].
- Synthetic data generation [Takahashi et al., 2019] and Monte Carlo simulation of SDEs using GANs [van Rhijn et al., 2021].

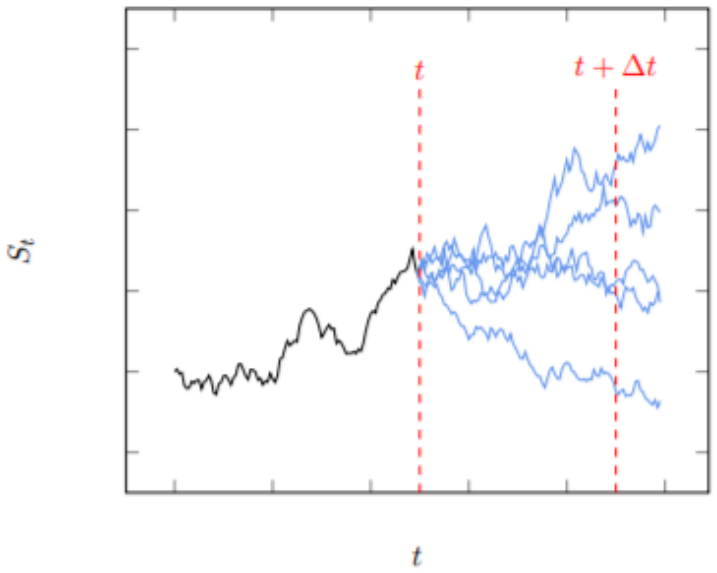


Figure 1: Illustration of the problem setting: given the process $\{S_t\}$ up to time t, obtain samples from the process at time $t + \Delta t$ using a GAN.

(a) Attacking deep nets with GANs: [Xiao et al., 2018].

- Market prediction [Xingyu et al., 2018]: a model that learns the properties of data without explicit assumptions or mathematical formulations; stochastic process cannot do without non-trivial assumptions.

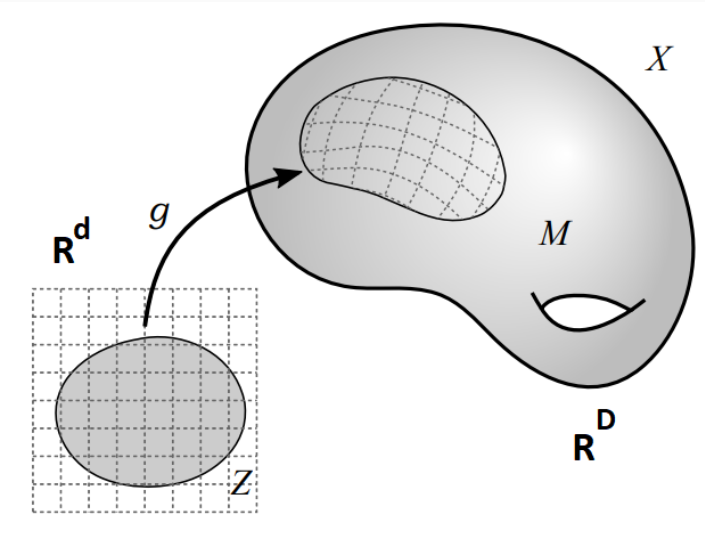
Outline

Mathematical context

The data

1. Data:
   - ▷ Target distribution: probability measure $\mu^\star$ on $\mathbb{R}^D$.
   - ▷ Finite-samples: $X_1, \ldots, X_n$ i.i.d. as $\mu^\star$. μ_n: empirical measure.
   - ▷ Objective: how can we sample from $\mu^\star$?

2. Latent variable:
   - ▷ Z defined on $\mathbb{R}^d$.
   - ▷ Z is typically uniform or Gaussian.
   - ▷ $d \ll D$: the manifold hypothesis.



Source: [Shao et al., 2018].

The Generator

Generator: a parametric family of functions from $\mathbb{R}^d$ to $\mathbb{R}^D$.

- ▷ Each G_θ is a neural network.
- ▷ Definition: $G_\theta(Z) \overset{\mathscr{L}}{\sim} \mu_\theta$.
- ▷ Notation: $\mathscr{G} = \{G_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^P$.
- ▷ Associated family of distributions: $\mathscr{P} = \{\mu_\theta : \theta \in \Theta\}$.
- ▷ Each μ_θ is a candidate to represent $\mu^\star$.

The discriminator

- Discriminator: a parametric family of functions from $\mathbb{R}^D$ to $\mathbb{R}$.
- Notation: $\mathscr{D} = \{D_\alpha : \alpha \in \Lambda\}$, $\Lambda \subseteq \mathbb{R}^Q$.
- In GANs algorithms, each D_α is a neural network.
- D_α is trained to distinguish between real and fake samples.

Source: https://www.wikihow.com.

Adversarial principle

- Objective: solve

$$\inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda}\Big[\mathbb{E}\log(D_\alpha(X))+\mathbb{E}\log(1-D_\alpha(G_\theta(Z)))\Big].$$

  - ▷ The higher $D(x)$, the higher the probability that x is drawn from $\mu^\star$.
  - ▷ The generator and the discriminator have opposite objectives.
  - ▷ Forget: estimation by maximum likelihood.
  - ▷ Forget: a strategy based on nonparametric density estimation.

- Empirical version:

$$\inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda}\Big[\frac{1}{n}\sum_{i=1}^{n}\log(D_\alpha(X_i))+\mathbb{E}\log(1-D_\alpha(G_\theta(Z)))\Big].$$

- The min-max optimum is found by alternative stochastic gradient descent.
- Generative principle: $\hat{\theta}_n \to G_{\hat{\theta}_n} \to G_{\hat{\theta}_n}(Z_1), G_{\hat{\theta}_n}(Z_2)\ldots \to$ new images.

- **Reminder**: for μ and ν probability measures on $\mathbb{R}^D$,

$$D_{\mathrm{JS}}(\mu, \nu) = \frac{1}{2} D_{\mathrm{KL}}\left(\mu \,\middle\|\, \frac{\mu+\nu}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(\nu \,\middle\|\, \frac{\mu+\nu}{2}\right).$$

- **Reminder**: for μ and ν probability measures on $\mathbb{R}^D$,

$$D_{\mathrm{JS}}(\mu,\nu) = \frac{1}{2} D_{\mathrm{KL}}\left(\mu \,\middle\|\, \frac{\mu+\nu}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(\nu \,\middle\|\, \frac{\mu+\nu}{2}\right).$$

- Idealization: $\mathscr{D} = \mathscr{D}_\infty$, the set of all functions from $\mathbb{R}^D$ to $[0,1]$.

$$\sup_{D \in \mathscr{D}_\infty} \Big[\mathbb{E}\log(D(X)) + \mathbb{E}\log(1 - D(G_\theta(Z)))\Big] = 2D_{\mathrm{JS}}(\mu^\star, \mu_\theta) - \ln 4.$$

- Consequence:

$$\inf_{\theta\in\Theta} \sup_{D \in \mathscr{D}_\infty} \Big[\mathbb{E}\log(D(X)) + \mathbb{E}\log(1 - D(G_\theta(Z)))\Big] = 2\inf_{\theta\in\Theta} D_{\mathrm{JS}}(\mu^\star, \mu_\theta) - \ln 4.$$

The role of the discriminator

In practice, one has always $\mathscr{D} = \{D_\alpha : \alpha \in \Lambda\}$

$$\sup_{\alpha \in \Lambda} \Big[\mathbb{E} \log(D_\alpha(X)) + \mathbb{E} \log(1 - D_\alpha(G_\theta(Z))) \Big]$$

acts like a divergence between the distributions μ_θ and the empirical distribution μ_n.

- Neural net divergence [Arora et al., 2017]
- Adversarial divergence [Liu et al., 2017]

Different variants of the discriminator's objective

1. Least squares GANs [Mao et al., 2017]: related to the Pearson-ξ^2 div.

$$\sup_{\alpha\in\Lambda}\sum_{i=1}^{n}(D_\alpha(X_i)-1)^2+\sum_{i=1}^{n}D_\alpha(G_\theta(Z_i))^2, \qquad \inf_{\theta\in\Theta}\sum_{i=1}^{n}(D_\alpha(G_\theta(Z_i))-1)^2.$$

2. [Nowozin et al., 2016] proposed f-GANs and showed that any f-divergence can be used for training GANs:

$$\inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda}\mathbb{E}D_\alpha(X)-\mathbb{E}(f^\star\circ D_\alpha)(G_\theta(Z)), \quad f^\star \text{ convex conjugate.}$$

3. When approximating other probability metrics
   - ▷ Wasserstein GANs [Arjovsky et al., 2017]:

$$\inf_{\theta\in\Theta}\sup_{\alpha\in\Lambda}\mathbb{E}_{\mu^\star}\ D_\alpha-\mathbb{E}_{\mu_\theta}\ D_\alpha.$$

   - ▷ MMD-GANs [Dziugaite et al., 2015, Li et al., 2015], Energy-based GANs [Zhao et al., 2017], Fisher GANs, Sobolev GANs...
   - ▷ No need to be absolutely continuous.
   - ▷ D_α is now a critic.

Wasserstein GANs

From GANs to WGANs

- Analysis of GANs [Goodfellow et al., 2014] made in [Biau, Cadre, Sangnier, and T., 2018] (Chapter 2).
- Drawbacks of orignal GANs formulation...
  - ▷ The training process of GANs is unstable.
  - ▷ Mode collapse phenomenon.
  - ▷ WGANs have become a standard in machine learning.
  - ▷ In the present study: both $\mathscr{G}$ and $\mathscr{D}$ are feed-forward neural networks.

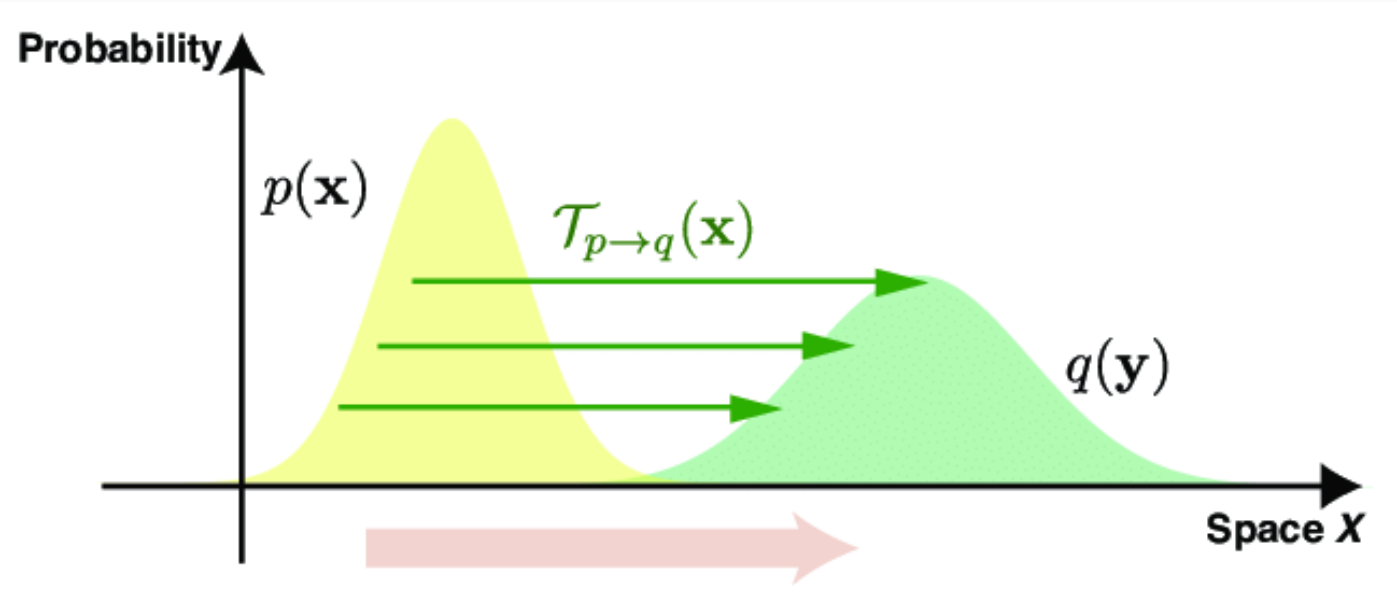
Reminder on the Wasserstein distance

- **Reminder**: for μ and ν probability measures in $P_1(E)$,

$$W_1(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{E\times E} \|x-y\|\pi(\mathrm{dx},\mathrm{dy}).$$

- **Dual form**:

$$W_1(\mu,\nu) = \sup_{f\in \mathrm{Lip}_1} |\mathbb{E}_\mu f - \mathbb{E}_\nu f|.$$



Source: https://www.wikihow.com.

General principle of WGANs

- Theoretical WGANs (T-WGANs):

$$\inf_{\theta\in\Theta} \sup_{f\in \mathrm{Lip}_1} |\mathbb{E}_{\mu^\star} f - \mathbb{E}_{\mu_\theta} f| = \inf_{\theta\in\Theta} W_1(\mu^\star, \mu_\theta).$$

- WGANs: in practice, one always has a parametric $\mathscr{D} = \{D_\alpha : \alpha \in \Lambda\}$:

$$\inf_{\theta\in\Theta} \sup_{\alpha\in\Lambda} |\mathbb{E}_{\mu^\star} D_\alpha - \mathbb{E}_{\mu_\theta} D_\alpha| = ??$$

- Empirical WGANs:

$$\inf_{\theta\in\Theta} \sup_{\alpha\in\Lambda} \Big[\frac{1}{n}\sum_{i=1}^{n} D_\alpha(X_i) - \mathbb{E} D_\alpha(G_\theta(Z))\Big] = ??$$

Notation & Objective of the present section

- For $\mathscr{D} \subseteq \text{Lip}_1$, the Integral Probability Metric $d_{\mathscr{D}}$ is

$$d_{\mathscr{D}}(\mu, \nu) = \sup_{f \in \mathscr{D}} |\mathbb{E}_\mu f - \mathbb{E}_\nu f|.$$

- Unified notation:

$$\text{T-WGANs: } \inf_{\theta \in \Theta} d_{\text{Lip}_1}(\mu^\star, \mu_\theta) \quad \text{and} \quad \Theta^\star = \underset{\theta \in \Theta}{arg\,min}\, d_{\text{Lip}_1}(\mu^\star, \mu_\theta),$$

$$\text{WGANs: } \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu^\star, \mu_\theta) \quad \text{and} \quad \bar{\Theta} = \underset{\theta \in \Theta}{arg\,min}\, d_{\mathscr{D}}(\mu^\star, \mu_\theta),$$

$$\text{Empirical WGANs: } \inf_{\theta \in \Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta) \quad \text{and} \quad \hat{\Theta}_n = \underset{\theta \in \Theta}{arg\,min}\, d_{\mathscr{D}}(\mu_n, \mu_\theta).$$

Final goal !

The present section aims at studying those three sets and we want to compare them with respect to the d_{Lip_1} distance.

$$\inf_{\theta\in\Theta} \underbrace{d_{\text{Lip}_1}(\mu^\star, \mu_\theta)}_{\text{T-WGANs:}} \overset{??}{\ll} \sup_{\bar{\theta}\in\bar{\Theta}} \underbrace{d_{\text{Lip}_1}(\mu^\star, \mu_{\bar{\theta}})}_{\text{WGANs:}} \overset{??}{\ll} \sup_{\hat{\theta}_n\in\hat{\Theta}_n} \underbrace{d_{\text{Lip}_1}(\mu^\star, \mu_\theta)}_{\text{Empirical WGANs:}},$$

d_{Lip_1} is the evaluation metric !

First objective: what are the properties of $d_{\mathscr{D}}$?

- It is clear that the properties of d_{Lip_1} are well known: studied by [Villani, 2008].
- We parameterize $\mathscr{D}$ with the newly defined GroupSort activation.

$$\tilde{\sigma}(x_1, x_2, \ldots, x_{2n-1}, x_{2n}) = (\max(x_1, x_2), \min(x_1, x_2), \ldots, \max(x_{2n-1}, x_{2n}), \min(x_{2n-1}, x_{2n}))$$

Theorem 1 ([Anil et al., 2018])

Assume that $E \subset \mathbb{R}^D$ is compact. Then, for any $f \in Lip_1(E)$ and any $\varepsilon > 0$, there exists a GroupSort neural network D such that $\|f - D\|_\infty \leqslant \varepsilon$.

First objective: what are the properties of $d_{\mathscr{D}}$?

- It is clear that the properties of d_{Lip_1} are well known: studied by [Villani, 2008].
- We parameterize $\mathscr{D}$ with the newly defined GroupSort activation.

$$\tilde{\sigma}(x_1, x_2, \ldots, x_{2n-1}, x_{2n}) = (\max(x_1, x_2), \min(x_1, x_2), \ldots, \max(x_{2n-1}, x_{2n}), \min(x_{2n-1}, x_{2n}))$$

Theorem 1 ([Anil et al., 2018])

Assume that $E \subset \mathbb{R}^D$ is compact. Then, for any $f \in Lip_1(E)$ and any $\varepsilon > 0$, there exists a GroupSort neural network D such that $\|f - D\|_\infty \leqslant \varepsilon$.

Consequences

There exists a discriminator $\mathscr{D}$ with weight constraints such that:

1. Each $D_\alpha \in \mathscr{D}$ is 1-Lipschitz.
2. The neural IPM $d_{\mathscr{D}}$ is a metric on $\mathscr{P} \cup \{\mu^\star\}$.
3. The neural IPM $d_{\mathscr{D}}$ metrizes weak convergence in $\mathscr{P} \cup \{\mu^\star\}$.
4. GroupSort networks studied in [Biau, Sangnier, and T., 2021].

Studying:

$$\Theta^\star = \underset{\theta\in\Theta}{arg\,min}\; d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \quad \text{and} \quad \bar{\Theta} = \underset{\theta\in\Theta}{arg\,min}\; d_{\mathscr{D}}(\mu^\star, \mu_\theta).$$

and their differences...

Theorem 2 (From [Biau, Sangnier, and T., 2020])

The functions $\theta \mapsto d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ and $\theta \mapsto d_{\mathscr{D}}(\mu^\star, \mu_\theta)$ are Lipschitz continuous, and the Lipschitz constant of $d_{\mathscr{D}}$ is independent of $\mathscr{D}$.

Consequently,

Corollary 1

The sets $\Theta^\star$ and $\bar{\Theta}$ are non empty.

Third objective: understanding the optimization error

- Error when minimizing $d_{\mathscr{D}}(\mu^\star, \mu_\theta)$ instead of $d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$:

Optimization error

$$0 \leqslant \varepsilon_{\text{optim}} = \sup_{\bar{\theta} \in \bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta).$$

$$\varepsilon_{\text{optim}} \leqslant \sup_{\theta \in \Theta} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) - d_{\mathscr{D}}(\mu^\star, \mu_\theta) \right] = T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}).$$

- Controlling $T_{\mathscr{P}}$ means controlling the gap between d_{Lip_1} and $d_{\mathscr{D}}$.
- Note: $\mathscr{D} \subset \mathscr{D}' \Rightarrow T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}) \searrow$ $\quad \mathscr{P} \subset \mathscr{P}' \Rightarrow T_{\mathscr{P}}(\mathrm{Lip}_1, \mathscr{D}) \nearrow$

Theorem 3 ([Biau et al., 2020])

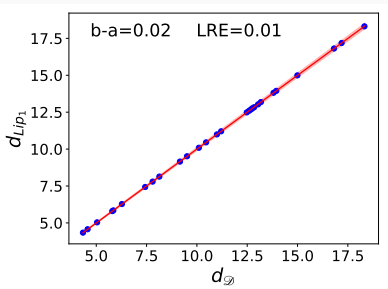
For all $\varepsilon > 0$, there exists a class of discriminators $\mathscr{D}$ such that

$$0 \leqslant \varepsilon_{optim} \leqslant T_{\mathscr{P}}(Lip_1, \mathscr{D}) \leqslant c\varepsilon.$$

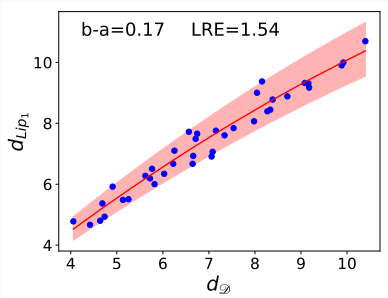
- Message: For any generative model $\mathscr{P}$ and any ε, one can find a discriminator such that the loss in performance is of the order of ε.

Empirical comparisons of the two distances

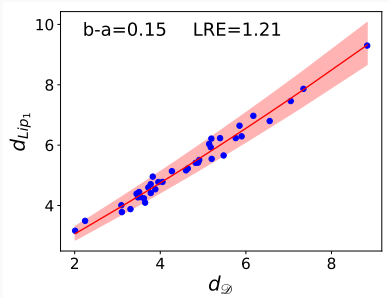
- Setting: μ_1 and μ_2 are mixtures of bivariate Gaussian densities.
- Note: when $K \nearrow$ we have $(b - a) \nearrow$.



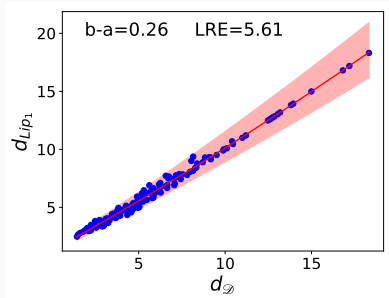
(a) 1 component.



(b) 2 components.

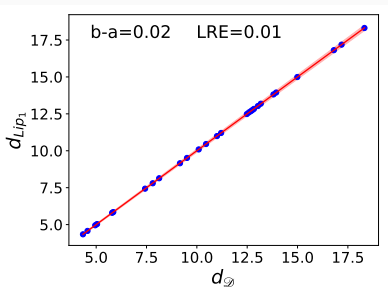


(c) 9 components.

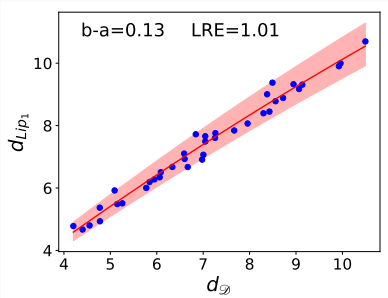


(d) 25 components.

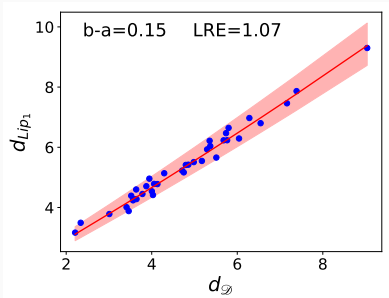
Figure 3: Discriminator $\mathscr{D}$ with depth $q = 2$.



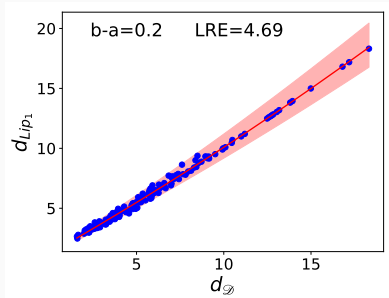
(a) 1 components.



(b) 2 components.



(c) 9 components.



(d) 25 components.

Figure 4: Discriminator $\mathscr{D}$ with depth $q = 5$.

Fourth objective: analyzing asymptotic properties

- Data: $X_1, \ldots, X_n$ i.i.d. as $\mu^\star$.
- Optimization problem:

$$\inf_{\theta\in\Theta} \sup_{\alpha\in\Lambda} \Big[\frac{1}{n} \sum_{i=1}^{n} D_\alpha(X_i) - \mathbb{E} D_\alpha(G_\theta(Z)) \Big] = \inf_{\theta\in\Theta} d_{\mathscr{D}}(\mu_n, \mu_\theta).$$

- Recall $\hat{\Theta}_n = \underset{\theta\in\Theta}{arg\,min}\, d_{\mathscr{D}}(\mu_n, \mu_\theta)$.
- Upper bounds on the performance of WGANs

$$\begin{aligned} 0 &\leqslant d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n}) - \inf_{\theta\in\Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta) \\ &\leqslant \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}}, \end{aligned}$$

Understanding $\varepsilon_{\text{estim}}$

Lemma 1

$\varepsilon_{estim} \to 0$ *almost surely as* $n \to \infty$.

Key inequality

$$0 \leqslant \varepsilon_{\text{estim}} + \varepsilon_{\text{optim}} \leqslant 2T_{\mathscr{P}}(\text{Lip}_1, \mathscr{D}) + 2d_{\mathscr{D}}(\mu^\star, \mu_n).$$

- $T_{\mathscr{P}}(\text{Lip}_1, \mathscr{D}) \nearrow$ when the capacity of $\mathscr{P} \nearrow$.
- The discriminator plays a more ambivalent role.
- Next step: bounds on $d_{\mathscr{D}}(\mu^\star, \mu_n)$.

Proposition 1 (From [Biau, Sangnier, and T., 2020])

More generally, if $\mu^\star$ is γ sub-Gaussian, then with probability at least $1-\eta$,

$$d_{\mathscr{D}}(\mu^\star, \mu_n) \leqslant \frac{c}{\sqrt{n}} + 8\gamma\sqrt{eD}\sqrt{\frac{\log(1/\eta)}{n}}.$$

Remark: c is $O(qQ^{3/2}(D^{1/2}+q))$.

Theorem 4 (From [Biau, Sangnier, and T., 2020])

More generally, if $\mu^\star$ is γ sub-Gaussian, then, for all $\varepsilon > 0$, there exists a discriminator $\mathscr{D}$ such that, with probability at least $1-\eta$,

$$0 \leqslant \varepsilon_{estim} + \varepsilon_{optim} \leqslant 2\varepsilon + \frac{2c}{\sqrt{n}} + 16\gamma\sqrt{eD}\sqrt{\frac{\log(1/\eta)}{n}}.$$

Warning: c is a function of ε.

Fifth objective: understanding the overall performance of WGANs

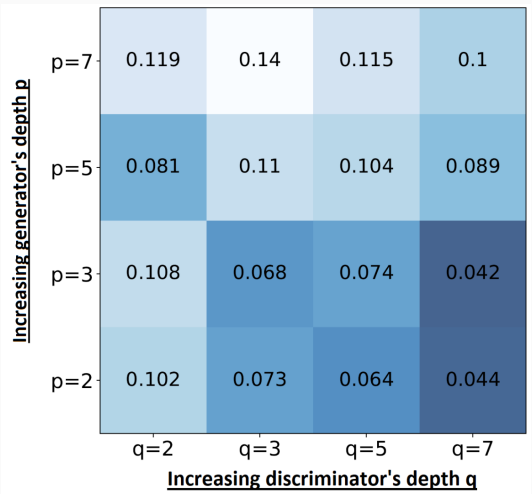
$$d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n}) \leqslant \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} + \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$$

$$= \varepsilon_{\mathrm{estim}} + \varepsilon_{\mathrm{optim}} + \varepsilon_{\mathrm{approx}}$$

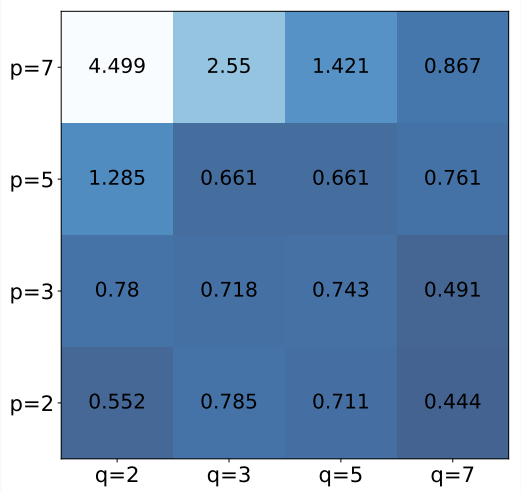
- $\varepsilon_{\mathrm{estim}} = \sup_{\theta_n \in \hat{\Theta}_n} \left[d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n}) - d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}_n}) \right]$ (data)
- $\varepsilon_{\mathrm{optim}} = \sup_{\bar{\theta} \in \bar{\Theta}} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\bar{\theta}}) - \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ (metric discrepancy)
- $\varepsilon_{\mathrm{approx}} = \inf_{\theta \in \Theta} d_{\mathrm{Lip}_1}(\mu^\star, \mu_\theta)$ (model)

Synthetic experiments

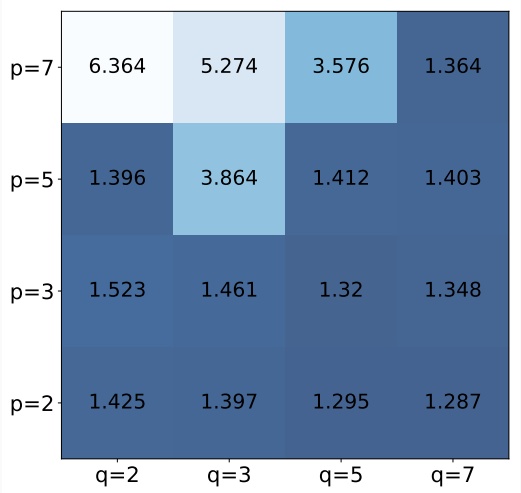
- Setting: $\mu^\star$ is a mixture of Gaussian densities with 2, 4 or 9 components.
- A family of generators: $\{\mathscr{G}_p : p = 2, 3, 5, 7\}$.
- A family of discriminators: $\{\mathscr{D}_q : q = 2, 3, 5, 7\}$.
- We draw $X_1, \ldots, X_n$ drawn from $\mu^\star$ with $n = 5000$.
- We plot the performance: $\sup_{\theta_n \in \hat{\Theta}_n} d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\hat{\theta}_n}) \leqslant \varepsilon_{\text{estim}} + \varepsilon_{\text{optim}} + \varepsilon_{\text{approx}}$.



(a) $K = 2$.



(b) $K = 4$.



(c) $K = 9$.

Figure 5: $d_{\mathrm{Lip}_1}(\mu^\star, \mu_{\theta_n})$ for different generator's and discriminator's capacity.

A too small discriminator facilitate instability and mode collapse

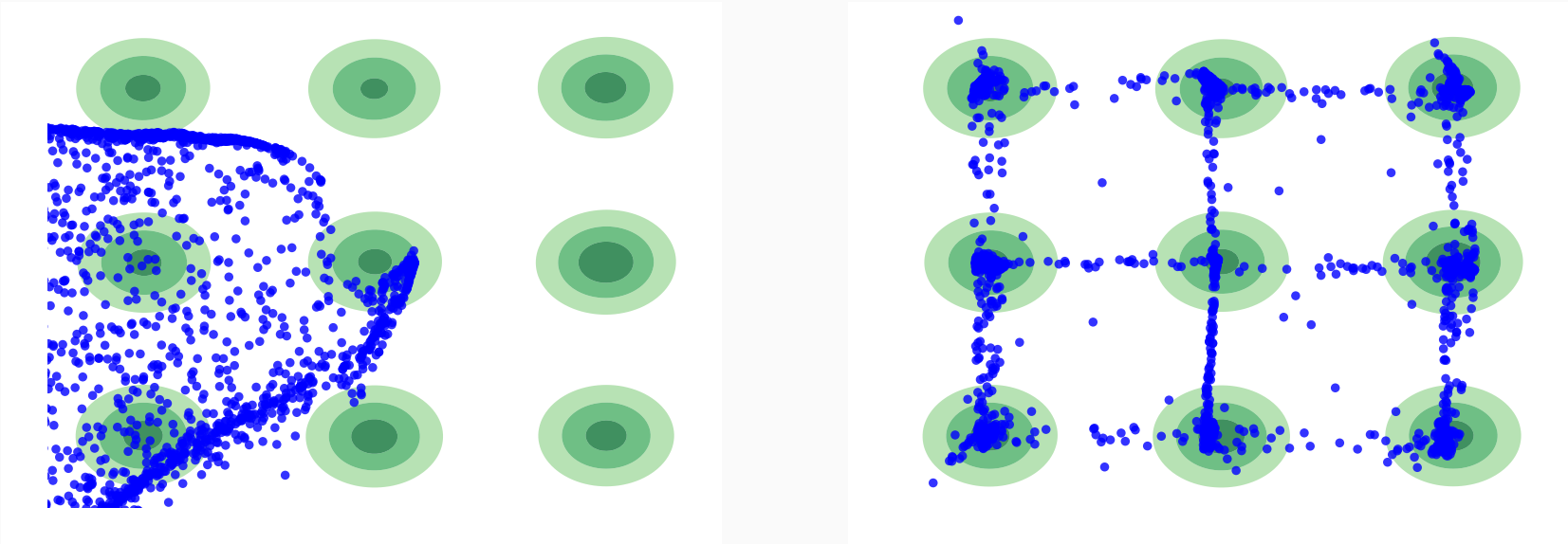


Figure 6: Left: Discriminator's depth=2, Generator's=4. Right: Discriminator's depth=5, Generator's=4

Transition

So far, we have:

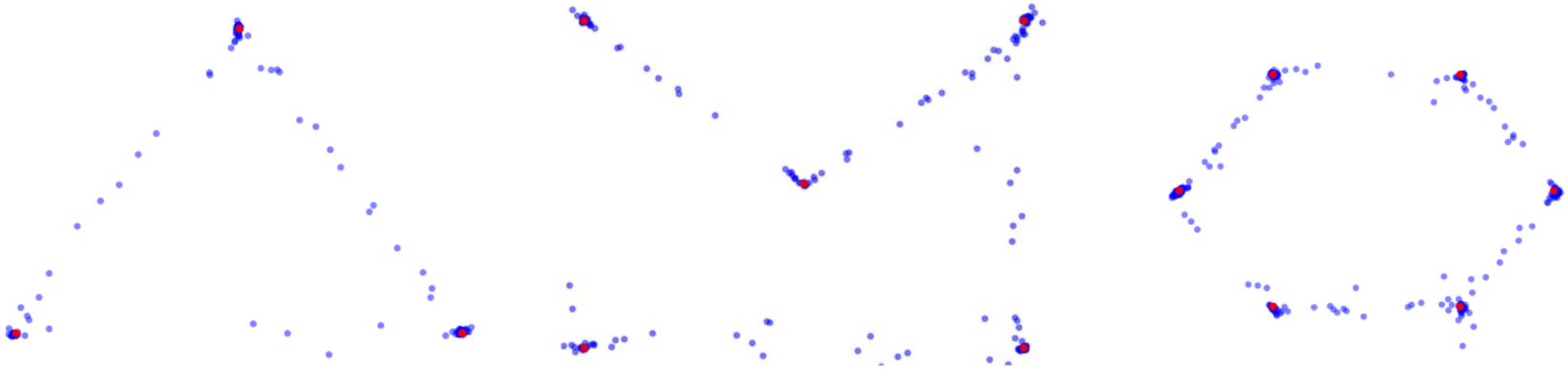
- Focused on understanding the discrepancy between $d_{\mathscr{D}}$ and d_{Lip_1} (W_1).
- The consequences on the optimum found in WGANs.

Let's forget about neural discriminators...

- Moving back to T-WGANs.
- Could we find their optimality properties in a simplified setting ?
- Since $d_{\mathscr{D}}$ approximates d_{Lip_1} understanding one could help explaining the other...

Optimal WGANs

A new question arises...



(a) $n = 3$. (b) $n = 5$. (c) $n = 6$.

Figure 4: Illustration of WGANs on few-shot learning ($n = 8, 16, 32$).

Setting

- Now, let U be a uniform random variable on $[0,1]^p$ and.
- For $K > 0$, let $\mathrm{Lip}_K(E, E')$ be the set of K-Lipschitz continuous functions.
- For $G \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)$, $G_\sharp U$ denotes the pushforward distribution of U by G.

New goal:

- Finding an optimal $\widehat{G}_K \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)$:

$$W_1(\widehat{G}_{K\sharp}U, \mu_n) = \inf_{G \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)} W_1(G_\sharp U, \mu_n). \tag{1}$$

Motivation: understanding what is the underlying objective of GANs.

Contributions

1. We start with an asymptotic analysis of $W_1(\widehat{G}_{K\sharp U}, \mu)$ as the sample size n tends to infinity (both univariate and multivariate).
2. Then, we provide a thorough finite sample analysis of the case $d = 1$. We explicitly describe the (two) functions achieving the infimum in (1).
3. Finally, we move to the setting where $d > 1$ and derive a finite sample bound on the infimum in (1).

The multivariate case is much more complicated...

Theorem 5 (From [Stephanovitch, T, Biau, Cadre, and Klutchnikoff 2022])

Let $\widehat{G}_K \in \widehat{\mathscr{G}}_K$. Assume that μ is of order 1, and let F^{-1} be the generalized inverse of the distribution function F of μ, i.e., for all $u \in (0,1)$, $F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geqslant u\}$.

1. *Assume that $S(\mu)$ is bounded.*
   - (*i*) *If $F^{-1} \in \mathrm{Lip}_{K_0}([0,1], \mathbb{R})$ for some $K_0 > 0$, then, for all $K \geqslant K_0$,*
   $$\lim_{n\to\infty} W_1(\widehat{G}_{K\sharp U}, \mu) = 0 \text{ a.s.}$$
   - (*ii*) *If $F \in \mathrm{Lip}_{K_1}(\mathbb{R}, [0,1])$ for some $K_1 > 0$, then, for all $K < 1/K_1$,*
   $$\liminf_{n\to\infty} W_1(\widehat{G}_{K\sharp U}, \mu) > 0 \text{ a.s.}$$
2. *Assume that $S(\mu)$ is unbounded. Then, for all $K > 0$,*
   $$\liminf_{n\to\infty} W_1(\widehat{G}_{K\sharp U}, \mu) > 0 \text{ a.s.}$$

Asymptotic analysis: case $d > 1$

Theorem 6 (From [Stephanovitch, T, Biau, Cadre, and Klutchnikoff 2022])

Let $\widehat{G}_K \in \widehat{\mathscr{G}}_K$. Assume that μ is of order 1 and that $\lambda_d(S(\mu)) > 0$, where λ_d denotes the Lebesgue measure on $\mathbb{R}^d$. Then, for all $K > 0$,

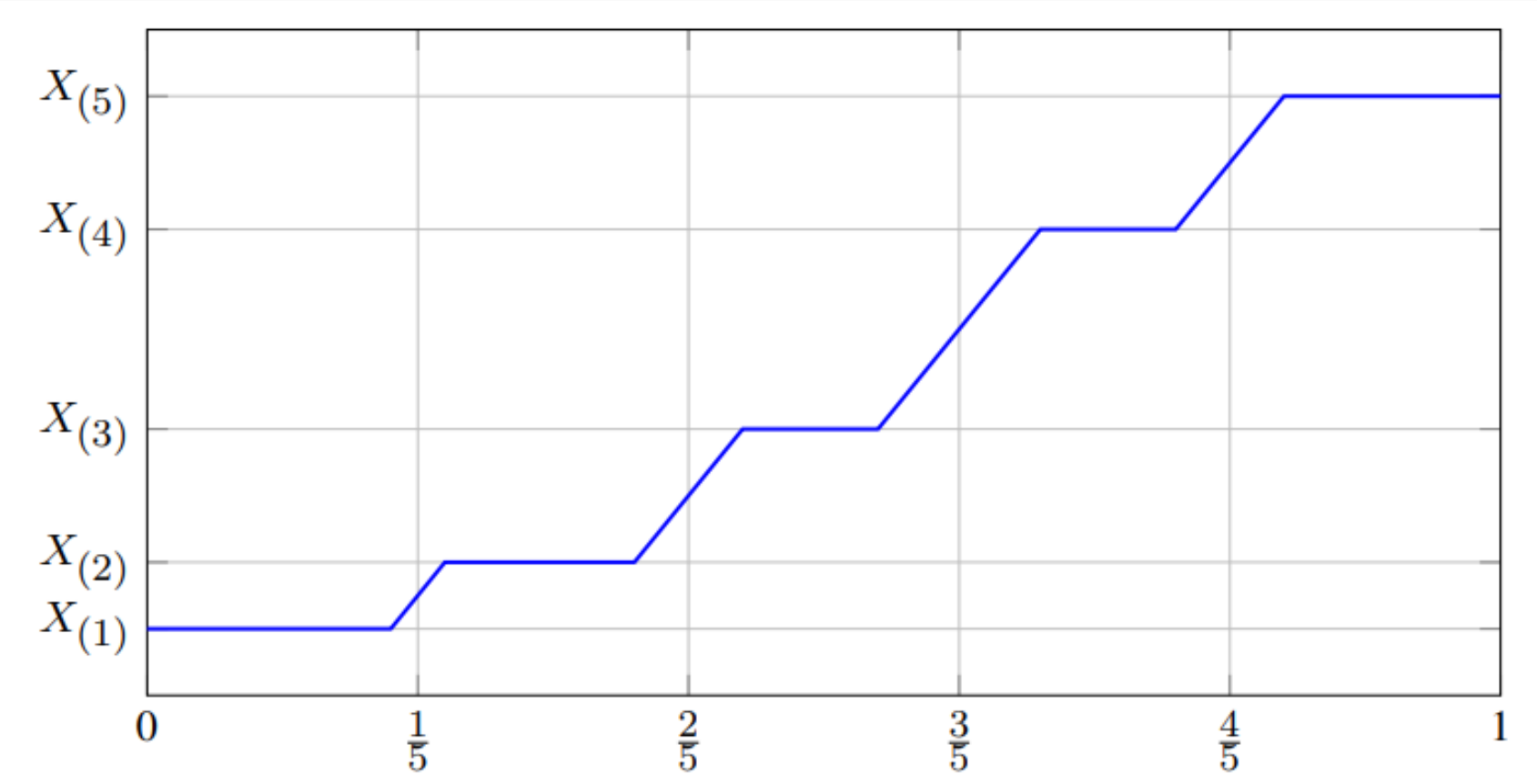
$$\liminf_{n \to \infty} \; W_1(\widehat{G}_{K\sharp U}, \mu) > 0 \text{ a.s.}$$

Important remark: K is fixed here !

We introduce the following function $\widehat{G}_K^\star : [0,1] \to \mathbb{R}$, and will show that it plays a key role in solving Problem (1).

$$\widehat{G}_K^\star(u) = \begin{cases} X_{(1)} & \text{if } u \in \left[0, \frac{1}{n} - \frac{X_{(2)} - X_{(1)}}{2K}\right] \\ X_{(i)} + K\left(u - \left(\frac{i}{n} - \frac{X_{(i+1)} - X_{(i)}}{2K}\right)\right) & \text{if } u \in \left[\frac{i}{n} - \frac{X_{(i+1)} - X_{(i)}}{2K}, \frac{i}{n} + \frac{X_{(i+1)} - X_{(i)}}{2K}\right] \\ & \text{for } 1 \leqslant i \leqslant n-1 \\ X_{(i+1)} & \text{if } u \in \left[\frac{i}{n} + \frac{X_{(i+1)} - X_{(i)}}{2K}, \frac{i+1}{n} - \frac{X_{(i+2)} - X_{(i+1)}}{2K}\right] \\ & \text{for } 1 \leqslant i \leqslant n-2 \\ X_{(n)} & \text{if } u \in \left[\frac{n-1}{n} + \frac{X_{(n)} - X_{(n-1)}}{2K}, 1\right]. \end{cases}$$

Illustration in 1D



[Stéphanovitch et al., 2022]

Theorem 7 (From [Stephanovitch, T, Biau, Cadre, and Klutchnikoff 2022])

Assume that $K \geqslant n \max_{i=1,\dots,n-1}(X_{(i+1)} - X_{(i)})$

$$W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \inf_{G \in \mathrm{Lip}_K([0,1],\mathbb{R})} W_1(G_\sharp U, \mu_n) = \frac{1}{4K}\sum_{i=1}^{n-1}(X_{(i+1)} - X_{(i)})^2.$$

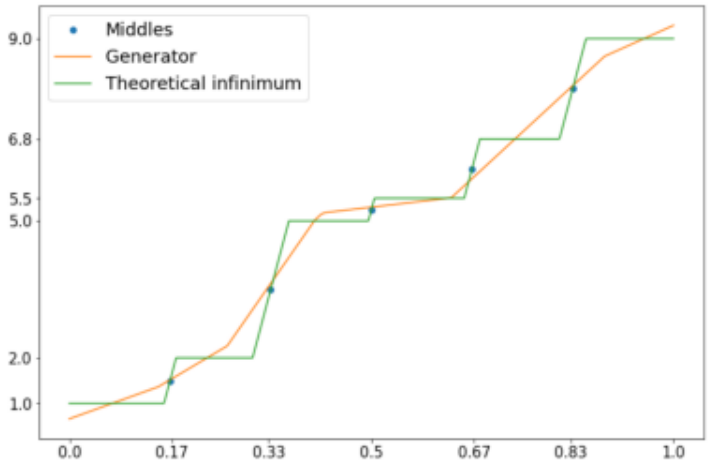
Moreover, $\widehat{\mathscr{G}}_K = \{\widehat{G}^\star_K, \widehat{G}^\star_K \circ S\}$, *where* $S(u) = 1 - u$, $u \in [0,1]$.

K is not fixed anymore !

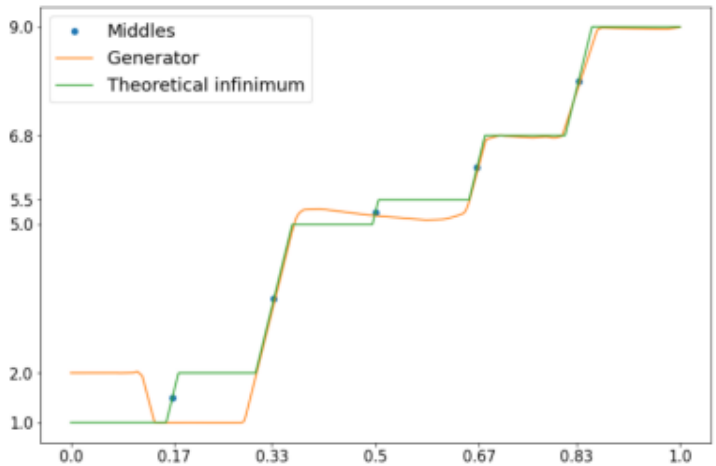
$\widehat{G}^\star_{K\sharp U}$ has atoms at the X_i's, of respective sizes

$$\frac{1}{n} - \frac{X_{(2)} - X_{(1)}}{2K} \quad \text{for } X_{(1)}, \qquad \frac{1}{n} - \frac{X_{(n)} - X_{(n-1)}}{2K} \quad \text{for } X_{(n)},$$
$$\frac{1}{n} - \frac{X_{(i+1)} - X_{(i-1)}}{2K} \quad \text{for } X_{(i)},\ i = 2, \dots, n-1,$$

Experience in 1D



(a) Fitting $n = 5$ data points with a generator depth equal to 3. $W_1(\widehat{G}^{\star}_{K\sharp U}, \mu_n) = 0.080$ and $W_1(G^{\theta}_{\sharp U}, \mu_n) = 0.501$.



(b) Fitting $n = 5$ data points with a generator depth equal to 5. $W_1(\widehat{G}^{\star}_{K\sharp U}, \mu_n) = 0.080$ and $W_1(G^{\theta}_{\sharp U}, \mu_n) = 0.165$.

[Stéphanovitch et al., 2022]

The shortest path plays a key role.

The set of paths connecting all data points $X_1, \ldots, X_n$, while minimizing the sum of the squared Euclidean distances, is defined as follows:

$$(k, \sigma) \in \arg\min \left\{ \sum_{i=1}^{n+k'-1} \|X_{\sigma'(i+1)} - X_{\sigma'(i)}\|^2 : k' \in \mathbb{N}, \sigma' \in \mathscr{S}_{k'} \right\}, \qquad (2)$$

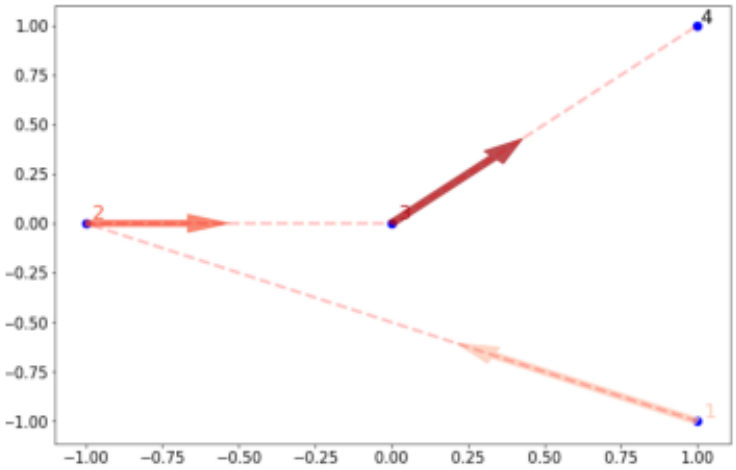
with

$$\sigma'(\{1, \ldots, n+k'\}) = \{1, \ldots, n\} \quad \text{and} \quad \sigma'(j) \neq \sigma'(j+1)$$

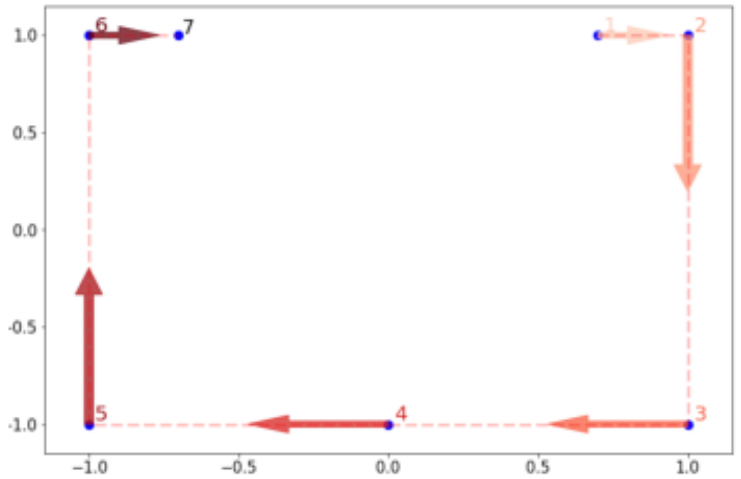
Note that

- k may be strictly positive (repetition).
- (k, σ) may not be unique.
- σ depends on k.

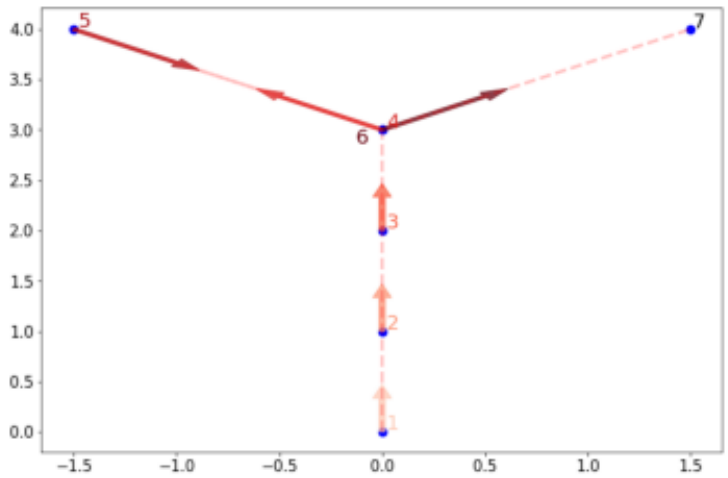
Some examples of shortest paths in 2D



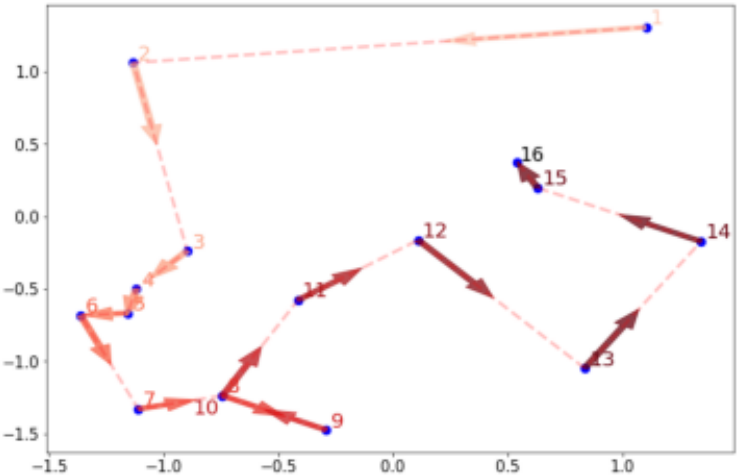
(a) Shortest path with $n = 4$, $k = 0$ in (5.1).



(b) Shortest path with $n = 7$, $k = 0$ in (5.1).



(c) Shortest path with $n = 6$, $k = 1$ in (5.1).



(d) Shortest path with $n = 15$, $k = 1$ in (5.1).

[Stéphanovitch et al., 2022]

Defining $\hat{G}$

Let us now provide some intuition on $\widehat{G}_K^\star : [0,1] \to \mathbb{R}^d$ is obtained.

1. The function strictly follows σ, one of the optimal paths in (2).
2. There exists $0 \leqslant t_1 < \cdots < t_{n+k} \leqslant 1$, $\widehat{G}_K^\star(t_j) = X_{\sigma(j)}$, $j \in \{1, \ldots, n+k\}$.
3. $\varphi(i)$ is the length of time $\widehat{G}_K^\star$ stays constant at X_i.
4. Now, we note V_j the time steps where the function $\widehat{G}_K^\star$ has arrived on a sample point $X_{\sigma(j)}$ and will pause for a time equal to $\varphi(\sigma(j))$.

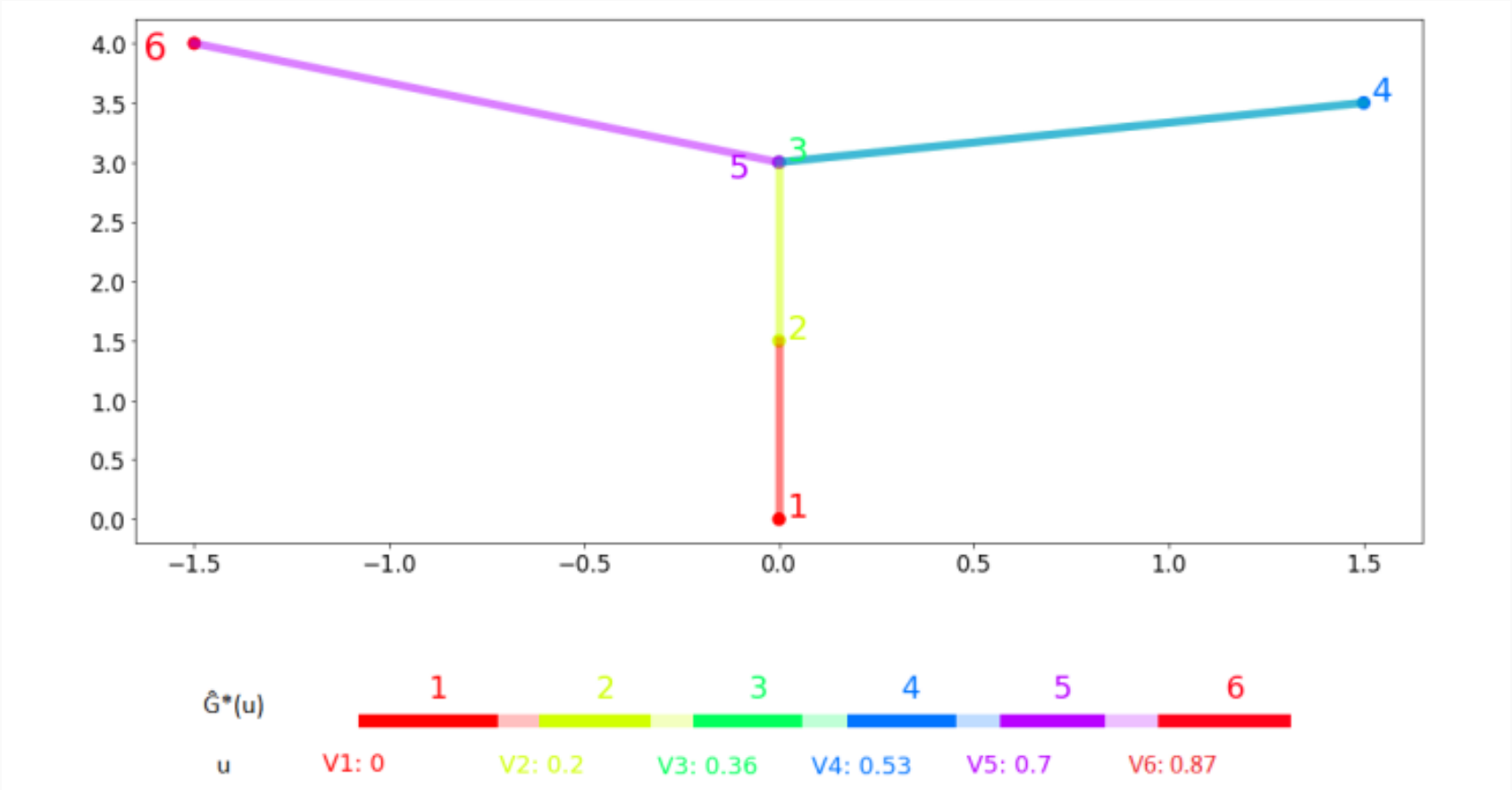
$$V_j = V_{j-1} + \varphi(\sigma(j-1)) + \frac{\|X_{\sigma(j)} - X_{\sigma(j-1)}\|}{K}.$$

As in the univariate case, outliers are (slightly) forgotten !

$\widehat{G}_K^\star : [0,1] \to \mathbb{R}^d$ is defined as follows:

$$\widehat{G}_K^\star(u) = \begin{cases} X_{\sigma(j)} & \text{if } u \in [V_j, V_j + \varphi(\sigma(j))] \\ & \text{for } 1 \leqslant j \leqslant n+k \\ X_{\sigma(j)} + \left(u - (V_j + \varphi(\sigma(j)))\right) K \frac{X_{\sigma(j+1)} - X_{\sigma(j)}}{\|X_{\sigma(j+1)} - X_{\sigma(j)}\|} & \text{if } u \in [V_j + \varphi(\sigma(j)), V_{j+1}] \\ & \text{for } 1 \leqslant j \leqslant n+k-1. \end{cases}$$

Understanding $\hat{G}$: an example in 2D



[Stéphanovitch et al., 2022]

Proposition 1 (From [Stephanovitch, T, Biau, Cadre, and Klutchnikoff 2022])

Assume that

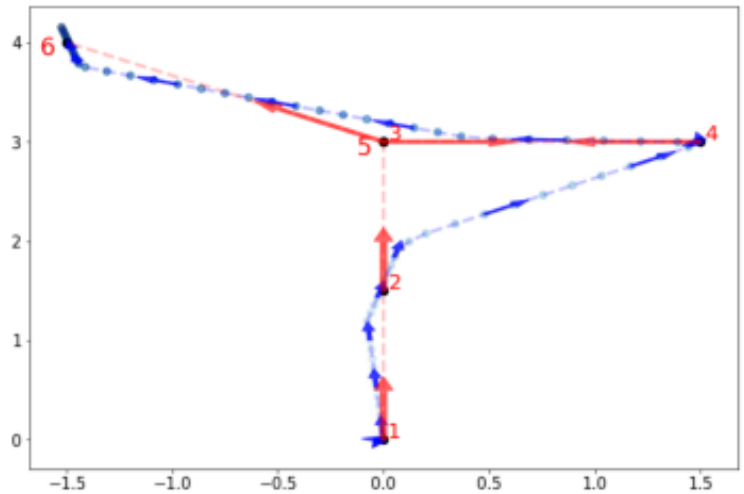
$$K \geqslant n \max_{i=1,\dots,n} \sum_{j \in \sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|),$$

and let $\widehat{G}^{\star}_K \in \mathrm{Lip}_K([0,1], \mathbb{R}^d)$ *(defined previously). Then*

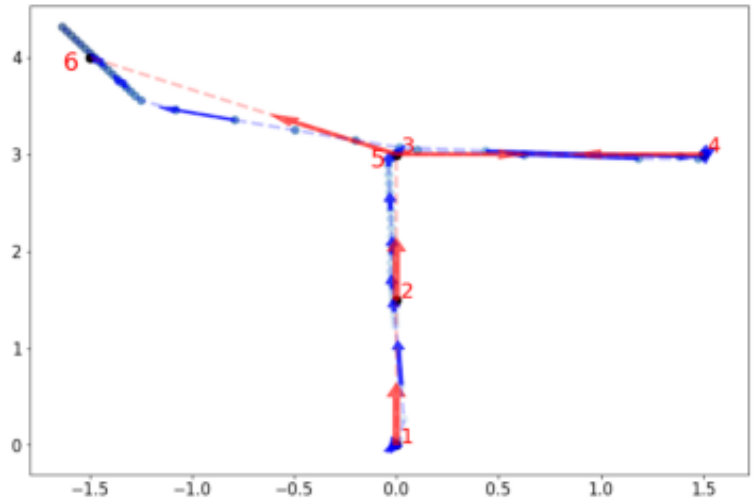
$$W_1(\widehat{G}^{\star}_{K\sharp U}, \mu_n) = \frac{1}{4K} \sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2.$$

- Quite a constraint on K !
- Is this the optimal generator ?

Experiences in 2D



(a) The sample size is $n = 5$ and the depth of the generator is equal to 3. The WGAN misses the shortest path leading to a deteriorated 1-Wasserstein distance: $W_1(\widehat{G}^{\star}_{K\sharp U}, \mu_n) = 0.030$ and $W_1(G^{\theta}_{\sharp U}, \mu_n) = 0.286$.



(b) The sample size is $n = 5$ and the depth of the generator is equal to 6. The WGAN is closer to the shortest path: $W_1(\widehat{G}^{\star}_{K\sharp U}, \mu_n) = 0.018$ and $W_1(G^{\theta}_{\sharp U}, \mu_n) = 0.174$.

[Stéphanovitch et al., 2022]

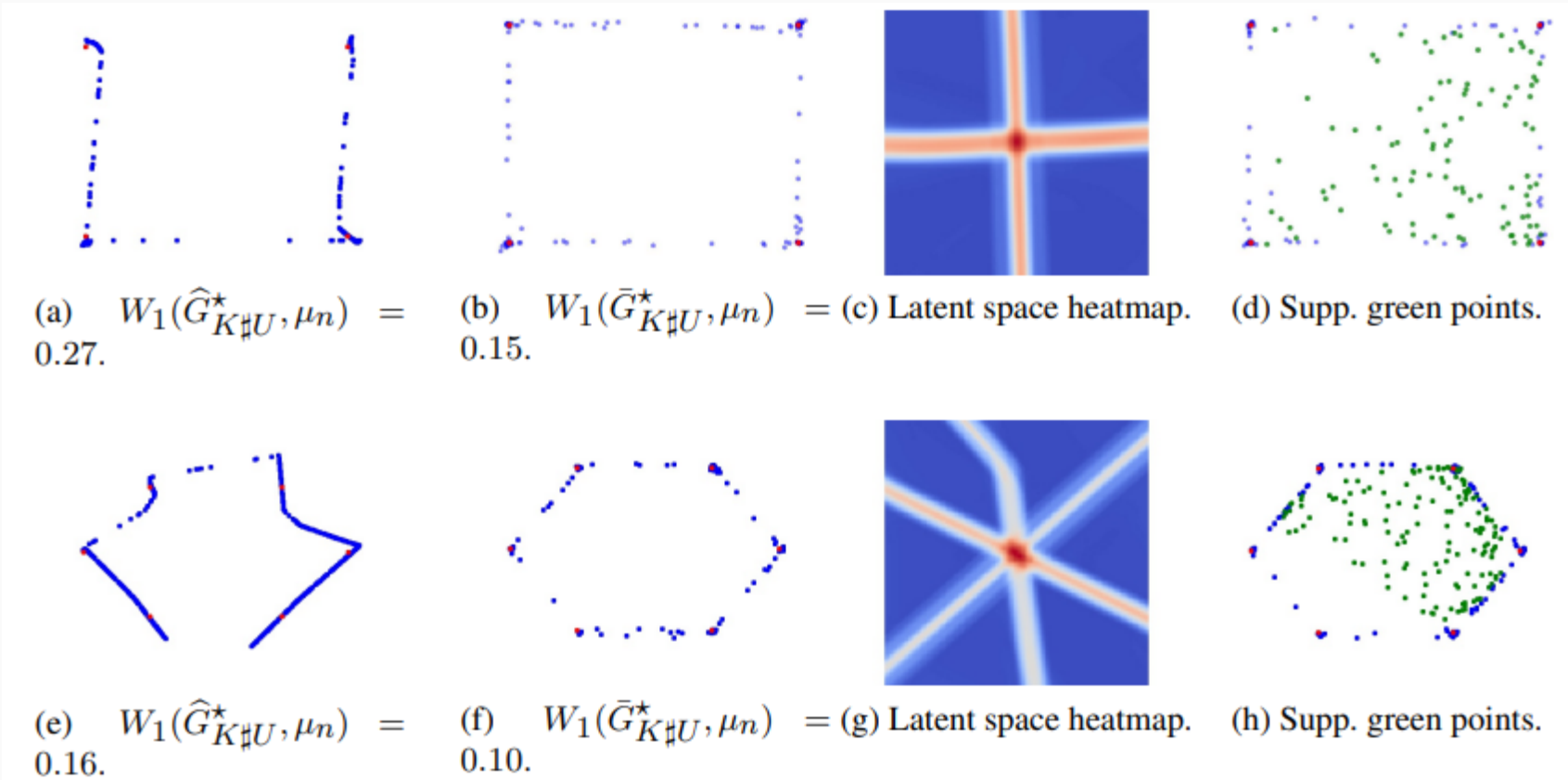
What happens when you put a stronger constraint on K ?

Assume that

$$K \leqslant n \max_{i=1,\dots,n} \sum_{j \in \sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|),$$

In this setting, one cannot construct $\hat{G}$.

What happens when you increase the dimension of the latent space ?



(a) $W_1(\widehat{G}^{\star}_{K\sharp U}, \mu_n) = 0.27.$ (b) $W_1(\bar{G}^{\star}_{K\sharp U}, \mu_n) = 0.15.$ (c) Latent space heatmap. (d) Supp. green points.

(e) $W_1(\widehat{G}^{\star}_{K\sharp U}, \mu_n) = 0.16.$ (f) $W_1(\bar{G}^{\star}_{K\sharp U}, \mu_n) = 0.10.$ (g) Latent space heatmap. (h) Supp. green points.

[Stéphanovitch et al., 2022]

Conclusion

General questions (1): Are GANs memorizing the dataset?

- Few shot learning regime: memorization doable...
- Huge dataset. $K \to \infty \implies$ underfitting.

Interesting case: the regular simplex...

$$K = n \max_{i=1,\dots,n} \sum_{j \in \sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|) = \sum_{i \in [1,n-1]} \|X_{i+1} - X_i\|$$

We have $\varphi(i) = 0$ for all $i \in [1, n-1] \implies$ no memorization. Consequently, how are interpolation done ?

General questions (2): WGANs work because they fail ?

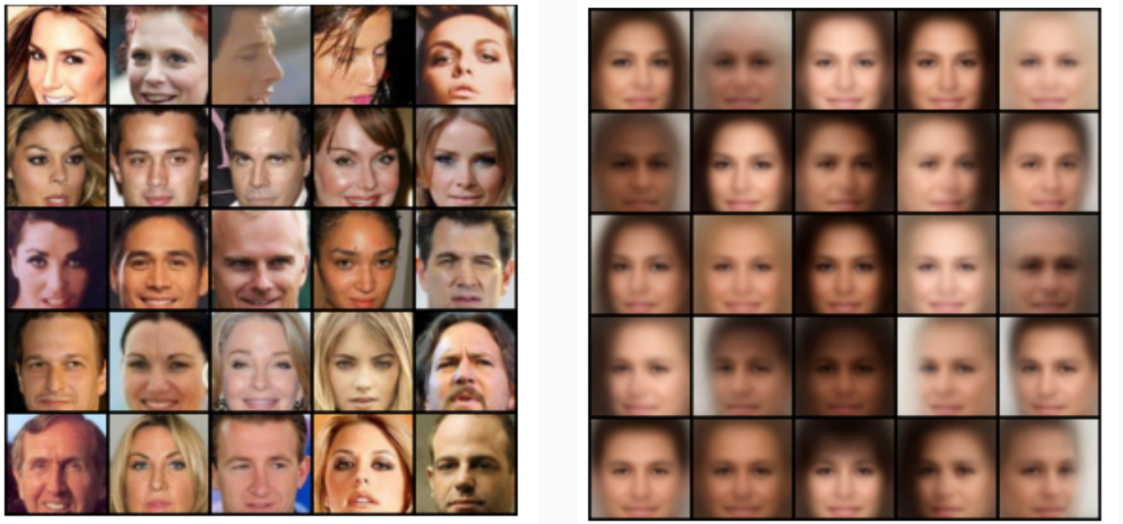


Figure 7: Left: $W(\mu_n, \tilde{\mu}_n) = 51.40$, Right: $W(\mu_n, \mu_n^k) = 40.15$ (k-means) [Stanczuk et al., 2021]

- Interesting properties of convolutional networks ??

$$\underset{\theta \in \Theta}{arg\,min}\, d_{\mathscr{D}}(\mu_n, \mu_\theta) \neq \underset{\theta \in \Theta}{arg\,min}\, d_{\mathrm{Lip}_1}(\mu_n, \mu_\theta).$$

- The discriminator punishes more samples out of the target manifold...
- Failure of the L_2 distance as a perceptual distance.

References i


Abdal, R., Qin, Y., and Wonka, P. (2019).
Image2stylegan: How to embed images into the stylegan latent space?
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441.


Anil, C., Lucas, J., and Grosse, R. B. (2018).
Sorting out lipschitz function approximation.
In *ICML*.


Arjovsky, M., Chintala, S., and Bottou, L. (2017).
Wasserstein Generative Adversarial Networks.
In *International Conference on Machine Learning*.


Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. (2017).
Generalization and equilibrium in generative adversarial nets (gans).
CoRR, abs/1703.00573.


Bau, D., Strobelt, H., Peebles, W., Zhou, B., Zhu, J.-Y., Torralba, A., et al. (2020).
Semantic photo manipulation with a generative image prior.
arXiv preprint arXiv:2005.07727.


Belomestny, D., Moulines, E., Naumov, A., Puchkin, N., and Samsonov, S. (2021).
Rates of convergence for density estimation with gans.
arXiv preprint arXiv:2102.00199.


Biau, G., T., U., and Sangnier, M. (2020).
Some theoretical insights into wasserstein gans.
arXiv preprint arXiv:2006.02682.


Dziugaite, G., Roy, D., and Ghahramani, Z. (2015).
Training generative neural networks via Maximum Mean Discrepancy optimization.
In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*.


Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative Adversarial Nets.
In *Advances in Neural Information Processing Systems*.


Li, Y., Swersky, K., and Zemel, R. (2015).
Generative Moment Matching Networks.
In *International Conference on Machine Learning*.


Liang, T. (2018).
On how well generative adversarial networks learn densities: Nonparametric and parametric results.
arXiv:1811.03179.


Liu, S., Bousquet, O., and Chaudhuri, K. (2017).
Approximation and convergence properties of generative adversarial learning.
In Guyon, I., Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5551–5559. Curran Associates, Inc., Red Hook.


Luise, G., Pontil, M., and Ciliberto, C. (2020).
Generalization properties of optimal transport gans with latent distribution learning.
arXiv preprint arXiv:2007.14641.


Mao, X., Li, Q., Xie, H., Lau, R., Wang, Z., and Smolley, S. (2017).
Least Squares Generative Adversarial Networks.
In *IEEE International Conference on Computer Vision*.


Nowozin, S., Cseke, B., and Tomioka, R. (2016).
f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization.
In *Neural Information Processing Systems*.


Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019).
Semantic image synthesis with spatially-adaptive normalization.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346.


Samangouei, P., Kabkab, M., and Chellappa, R. (2018).
Defense-gan: Protecting classifiers against adversarial attacks using generative models.
In *International Conference on Learning Representations*.


Schreuder, N., Brunel, V.-E., and Dalalyan, A. (2021).
Statistical guarantees for generative models without domination.
In *Algorithmic Learning Theory*, pages 1051–1071. PMLR.


Shao, H., Kumar, A., and Thomas Fletcher, P. (2018).
The riemannian geometry of deep generative models.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323.


Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., and Póczos, B. (2018).
Nonparametric density estimation with adversarial losses.
In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10246–10257. Curran Associates Inc.


Stanczuk, J., Etmann, C., Kreusser, L. M., and Schönlieb, C.-B. (2021).
Wasserstein gans work because they fail (to approximate the wasserstein distance).
arXiv preprint arXiv:2103.01678.


Stéphanovitch, A., T., U., Cadre, B., Klutchnikoff, N., and Biau, G. (2022).
Optimal 1-wasserstein distance for wgans.
arXiv preprint arXiv:2201.02824.


Takahashi, S., Chen, Y., and Tanaka-Ishii, K. (2019).
Modeling financial time-series with generative adversarial networks.
Physica A: Statistical Mechanics and its Applications, 527:121261.


Uppal, A., Singh, S., and Póczos, B. (2019).
Nonparametric density estimation & convergence rates for gans under besov ipm losses.
arXiv preprint arXiv:1902.03511.


van Rhijn, J., Oosterlee, C. W., Grzelak, L. A., and Liu, S. (2021).
Monte carlo simulation of sdes using gans.
arXiv preprint arXiv:2104.01437.


Villani, C. (2008).
***Optimal transport: old and new*, volume 338.**
Springer Science & Business Media.


Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., and Song, D. (2018).
Generating adversarial examples with adversarial networks.
arXiv preprint arXiv:1801.02610.


Xingyu, Z., Zhisong, P., Guyu, H., Siqi, T., and Cheng, Z. (2018).
Stock market prediction on high-frequency data using generative adversarial nets.
Mathematical Problems in Engineering, 2018:11.


Yang, L., Zhang, D., and Karniadakis, G. E. (2018).
Physics-informed generative adversarial networks for stochastic differential equations.
arXiv preprint arXiv:1811.02033.


Zhao, J., Mathieu, M., and LeCun, Y. (2017).
Energy-based Generative Adversarial Network.
In *International Conference on Learning Representations*.


Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. (2016).
Generative Visual Manipulation on the Natural Image Manifold.
In *European Conference on Computer Vision*.

Complementary work: analysis of convergence rates for adversarial divergences

Analysis of the following risk:

$$d_{\mathscr{D}}(\mu_{\hat{\theta}_n}, \mu^\star) - \min_{\theta \in \Theta} d_{\mathscr{D}}(\mu_\theta, \mu^\star).$$

1. $d_{\mathscr{D}}$ is an IPM:
   - ▷ Assumptions: both $\mu^\star$ and $\mathscr{D}$ corresponds to a non-parametric class of Sobolev spaces [Liang, 2018] and [Singh et al., 2018].
   - ▷ Assumptions: both $\mu^\star$ and $\mathscr{D}$ corresponds to a non-parametric class of Besov spaces [Uppal et al., 2019].
   - ▷ Assumptions: $\mu^\star$ is the pushforward distribution of a Lipschitz generator and $\mathbb{D}$ corresponds to the class of α-smooth functions [Schreuder et al., 2021].
2. $d_{\mathscr{D}}$ is a Sinkhorn divergence: [Luise et al., 2020].
3. $d_{\mathscr{D}}$ approximates the Jensen-Shannon divergence: [Biau, Cadre, Sangnier, and T., 2018] and [Belomestny et al., 2021].