

A decorative border at the top of the slide featuring a repeating geometric pattern of interlocking diamonds in dark green and light green.

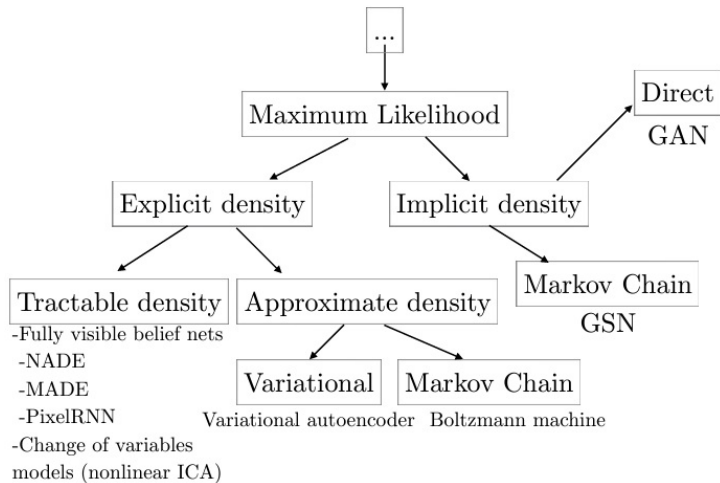
## X. Optimal Transport-based Method

Young-geun Kim

Department of Statistics and Probability

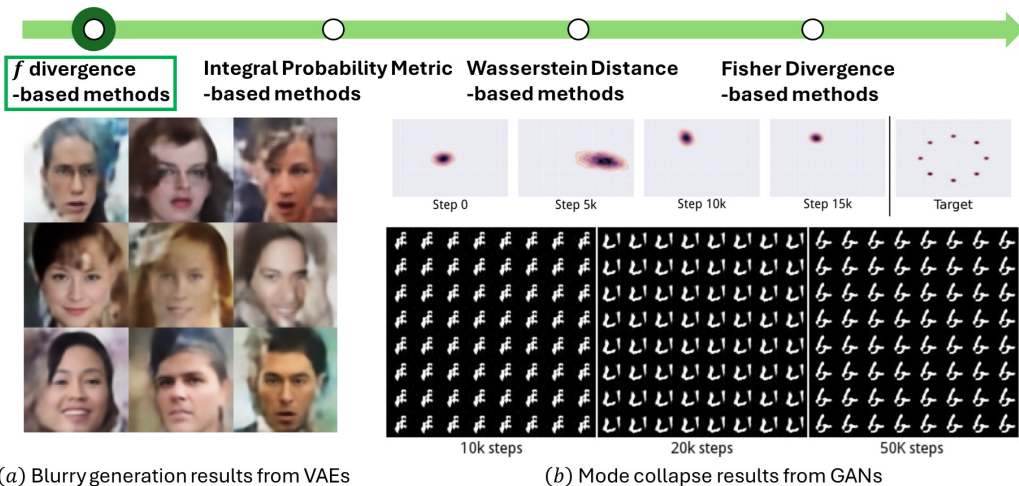
STT 997 (SS 2025)

# Introduction



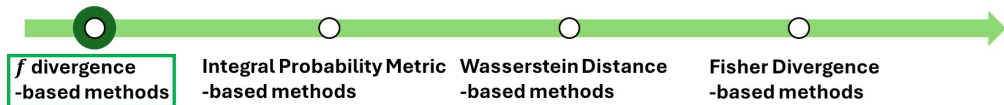
The figure is from Goodfellow (2016).

# Introduction



Images are edited from Tolstikhin et al. (2018) and Metz et al. (2017).

# Introduction

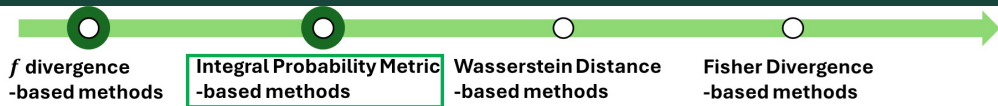


- The mode collapse phenomenon in GANs demonstrates that the  $p_\theta$  fails in capturing the support of  $p_n$ .
- Several works have criticized  $f$ -divergence,

$$\mathcal{D}_f(p_n \| p_\theta) = \int f\left(\frac{p_n(\vec{x})}{p_\theta(\vec{x})}\right) p_\theta(\vec{x}) d\vec{x},$$

pointing out that it is based on the density ratio  $p_n/p_\theta$ , and this dependency may be a reason for the observed failures in  $f$ -divergence-based methods.

# Introduction



- As an alternative to density ratios, a line of work has proposed focusing on discrepancy measures that are effective regardless of the differences between the supports of  $p_n$  and  $p_\theta$ .
- For example, Generative Moment Matching Networks (Li et al., 2015) aim to minimize:

$$\left\| \int \varphi(\vec{x}) d\mathbb{P}_n(\vec{x}) - \int \varphi(\vec{x}) d\mathbb{P}_\theta(\vec{x}) \right\|^2 \quad (1)$$

where the integrated terms  $\varphi(\vec{x})$  represent vectors of finite moments, e.g.,  $\varphi(x) = (c, \sqrt{2cx}, x^2)^T$  in the univariate case with second-order moments.

- This loss function is a special case of integral probability metrics,  $\gamma_{\mathcal{F}}(p_n, p_\theta)$ , where  $\mathcal{F}$  denotes a set of summary statistics functions, such as moments.

# Integral Probability Metric

- The IPMs can be expressed as:

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \int f(\vec{x}) d\mathbb{P}(\vec{x}) - \int f(\vec{x}) d\mathbb{Q}(\vec{x}) \right|$$

where  $\mathcal{F}$  is a class of real-valued functions, and  $\mathbb{P}$  and  $\mathbb{Q}$  are probability measures.

# Integral Probability Metric

- **Total Variation Distance:** The total variation distance,  $\delta(p, q) = \frac{1}{2} \int |p(\vec{x}) - q(\vec{x})| d\vec{x}$ , has an alternative expression:

$$\sup_{A \in \mathcal{A}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \sup_{A \in \mathcal{A}} \left| \int I(\vec{x} \in A) d\mathbb{P}(\vec{x}) - \int I(\vec{x} \in A) d\mathbb{Q}(\vec{x}) \right|$$

where  $\mathcal{A}$  is the corresponding  $\sigma$ -algebra. Thus, the total variation is the IPM using the set of indicator functions for all events.

- **Earth Mover's Distance:** When  $\mathcal{F}$  consists of all 1-Lipschitz continuous functions,  $\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q})$  corresponds to the Earth mover's distance (or 1-Wasserstein distance), a special case of Wasserstein distances. Further details will be discussed in the subsequent subsection on Wasserstein distances.

# Integral Probability Metric

- **Maximum Mean Discrepancy (MMD):** We denote the kernel mean by  $\mu_{\mathbb{P}}(\vec{x}) := \int k(\vec{x}', \vec{x}) d\mathbb{P}(\vec{x}')$ . Then, the MMD is defined as the difference between kernel means in  $\mathcal{H}$ , the RKHS specified by  $k$ :

$$\text{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

MMD builds a kernel-based test statistic for a two-sample test:

$$H_0 : \mathbb{P} = \mathbb{Q} \text{ vs. } H_1 : \mathbb{P} \neq \mathbb{Q}.$$

- The MMD has important alternative representations:
  - ① **IPM:**  $\text{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) - \int f(\vec{x}) d\mathbb{Q}(\vec{x}) \right).$
  - ② **Kernel function form:**

$$\begin{aligned} & \text{MMD}_k^2(\mathbb{P}, \mathbb{Q}) \\ &= \int k(\vec{x}, \vec{x}') d\mathbb{P}(\vec{x}) d\mathbb{P}(\vec{x}') - 2 \int k(\vec{x}, \vec{y}) d\mathbb{P}(\vec{x}) d\mathbb{Q}(\vec{y}) + \int k(\vec{y}, \vec{y}') d\mathbb{Q}(\vec{y}) d\mathbb{Q}(\vec{y}'). \end{aligned} \tag{2}$$



# MMD: Generative Moment Matching Network

1. **Generative Moment Matching Network (GMMN):** GMMNs (Li et al., 2015) propose to use empirical estimators as loss functions to train generative models rather than introducing adversarial networks as in GANs.
  - Given  $(\vec{x}_i)_{i=1}^B$  and  $(G_\theta(\vec{z}_i))_{i=1}^B$ , minibatch samples of size  $B$  from  $\mathbb{P}_n$  and  $\mathbb{P}_\theta$  respectively, the minibatch-based empirical estimators for  $\text{MMD}_k^2(\mathbb{P}_n, \mathbb{P}_\theta)$  can be expressed as

$$\begin{aligned} & \frac{1}{B(B-1)} \sum_{i=1}^B \sum_{j \neq i}^B k(\vec{x}_i, \vec{x}_j) - \frac{2}{B^2} \sum_{i=1}^B \sum_{j=1}^B k(\vec{x}_i, G_\theta(\vec{z}_j)) \\ & + \frac{1}{B(B-1)} \sum_{i=1}^B \sum_{j \neq i}^B k(G_\theta(\vec{z}_i), G_\theta(\vec{z}_j)). \end{aligned} \tag{3}$$

- GMMNs used a mixture of multiple Gaussian kernels with various bandwidth parameters.

# MMD: Generative Moment Matching Network

- Minimizing  $\text{MMD}_k(\mathbb{P}_n, \mathbb{P}_\theta)$  can be interpreted as matching moments between  $\mathbb{P}_n$  and  $\mathbb{P}_\theta$ .
- Let  $k$  be the kernel that defines the MMD, and let  $\varphi(\vec{x})^1$  represent the corresponding kernel feature mapping, i.e.,

$$k(\vec{x}, \vec{x}') = \varphi(\vec{x})^\top \varphi(\vec{x}') \quad (4)$$

- For a univariate example, consider  $k(x, x') = (xx' + c)^2$  for some  $c > 0$ . The feature mapping  $\varphi(x) = (c, \sqrt{2cx}, x^2)^\top$  satisfies Equation (4). Kernels with higher degrees allow for covering higher-order moments.
- The loss of GMMNs, minibatch-based empirical estimators for (squared) MMD, can be expressed as

$$\|B^{-1} \sum_{i=1}^B \varphi(\vec{x}_i) - B^{-1} \sum_{i=1}^B \varphi(G_\theta(\vec{z}_i))\|^2. \quad (5)$$

---

<sup>1</sup>The symbol  $\phi$  is more commonly used, but we use  $\varphi$  here to avoid confusion with parameters for auxiliary networks, e.g., the discriminator in GANs.

## 2. MMD GAN:

- GMMNs face challenges in selecting effective kernels. MMD GANs (Li et al., 2017) overcome this limitation by introducing adversarial kernel learning.
- MMD GANs aim to target  $\max_{k \in \mathcal{K}} \text{MMD}_k(\mathbb{P}_n, \mathbb{P}_\theta)$ , where  $\mathcal{K}$  is a class of kernel functions.
- To model an expressive class  $\mathcal{K}$ , MMD GANs employ a neural network  $E_\phi$  to define  $(k \circ E_\phi)(\vec{x}, \vec{x}') := k(E_\phi(\vec{x}), E_\phi(\vec{x}'))$ , targeting:

$$\max_{\phi} \text{MMD}_{k \circ E_\phi}(\mathbb{P}_n, \mathbb{P}_\theta). \quad (6)$$

- The injectivity of  $E_\phi$  is crucial to retain the important properties of MMDs with usual kernels. MMD-GANs incorporate an encoder architecture for  $E_\phi$ , add a decoder, and introduce a reconstruction error-based penalty term to enforce the injectivity.

## 3. Methods using Other IPMs:

- One of the main challenges in using IPMs,

$$\gamma_{\mathcal{F}}(\mathbb{P}_n, \mathbb{P}_{\theta}) := \sup_{f \in \mathcal{F}} \left| \int f(\vec{x}) d\mathbb{P}_n(\vec{x}) - \int f(\vec{x}) d\mathbb{P}_{\theta}(\vec{x}) \right|,$$

lies in approximating the supremum over the function class  $\mathcal{F}$ .

- While MMD has a tractable representation that allows for the direct use of its empirical estimators, this is not the case for more general IPMs.
- Most methods targeting other IPMs employ neural networks to model elements within  $\mathcal{F}$ . Notably, Wasserstein GANs (Arjovsky et al., 2017) have become one of the most popular methods targeting the 1-Wasserstein distance.

# Outline

- 1 OPTIMAL TRANSPORT
- 2 WASSERSTEIN GENERATIVE MODELS
- 3 APPLICATION

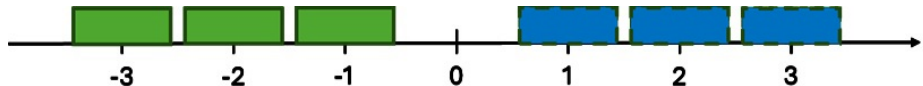
# Optimal Transport



- Another line of work has targeted Wasserstein distances from an optimal transport perspective.
- Useful materials include:
  - 1 *Optimal transport: old and new* (Villani et al., 2009)
  - 2 *Optimal transport for applied mathematicians* (Santambrogio, 2015)

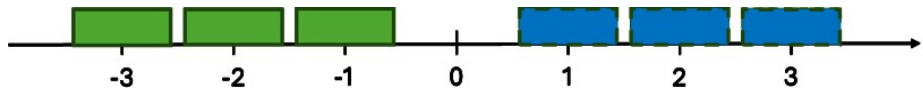
Images are edited from Santambrogio (2015).

# Motivating Example: Monge Problem



- *Optimal Transport* is a mathematical problem focused on identifying maps that most efficiently move one distribution of mass to another and investigating their properties.
- Let's consider a toy example of moving green bold boxes on the left to blue dashed-box regions on the right.
- Which way is more efficient?
  - 1  $T_1$ :  $T_1(-3) = 1$ ,  $T_1(-2) = 2$ , and  $T_1(-1) = 3$
  - 2  $T_2$ :  $T_2(-3) = 3$ ,  $T_2(-2) = 2$ , and  $T_2(-1) = 1$

# Motivating Example: Monge Problem



- The answer depends on criteria. When we consider the squared  $L_2$ -distance between the original green boxes and their transportation results:

$$\begin{aligned} & \left( (-3 - T_1(-3))^2 + (-2 - T_1(-2))^2 + (-1 - T_1(-1))^2 \right) / 3 \\ &= \left( (-3 - 1)^2 + (-2 - 2)^2 + (-1 - 3)^2 \right) / 3 = 16 \end{aligned} \quad (7)$$

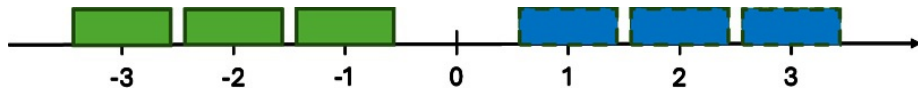
$$\begin{aligned} & \left( (-3 - T_2(-3))^2 + (-2 - T_2(-2))^2 + (-1 - T_2(-1))^2 \right) / 3 \\ &= \left( (-3 - 3)^2 + (-2 - 2)^2 + (-1 - 1)^2 \right) / 3 = 56/3 \end{aligned} \quad (8)$$

Q: How many (transportation) maps are there?

Q: What happens when we consider the  $L_1$ -distance?



# Motivating Example: Monge Problem















- The *Monge Problem* is a classical formulation in Optimal Transport: Given two probability measures  $\mathbb{P}$  and  $\mathbb{Q}$ , and a cost function  $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , the optimal transportation cost is defined as:

$$\min_{T: \mathbb{P}(T(\vec{x})) = \mathbb{Q}(\vec{x})} \int c(\vec{x}, T(\vec{x})) d\mathbb{P}(\vec{x}) \quad (9)$$

The constraint  $\mathbb{P}(T(\vec{x})) = \mathbb{Q}(\vec{x})$  represents the push-forward operation  $T\#\mathbb{P} = \mathbb{Q}$ . Here, the optimal solutions are referred to as *optimal transport maps*.

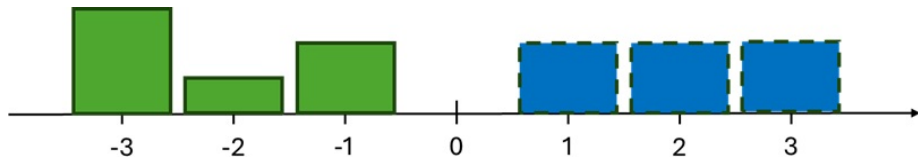
- When the cost function is the squared  $L_2$ -distance,  $c(\vec{x}, \vec{x}') = \|\vec{x} - \vec{x}'\|^2$ ,  $T_1$  is the unique optimal transport map. In the univariate case, the transport map that moves percentiles to corresponding percentiles is optimal.

# Motivating Example: Optimal Transportation Costs as Measures of Generation Quality

								
	1/3	0	0			0.10	0.30	0.40
	0	1/3	0			0.40	0.05	0.25
	0	0	1/3			0.25	0.30	0.10
Joint Distribution					Transportation Cost			

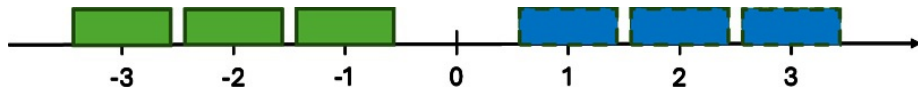
$$\begin{aligned}
 & (0.10 \times (1/3) + 0.30 \times 0 + 0.40 \times 0) \\
 & + (0.40 \times 0 + 0.05 \times (1/3) + 0.25 \times 0) \\
 & + (0.25 \times 0 + 0.30 \times 0 + 0.10 \times (1/3)) = 0.08
 \end{aligned}$$

# Motivating Example: Kantorovich Problem



- The constraint on the optimal transport map,  $\mathbb{P}(T(\vec{x})) = \mathbb{Q}(\vec{x})$ , generally poses challenges in defining the optimal transportation cost.
- The *Kantorovich Problem* is a generalization of the Monge Problem to alleviate these challenges.

# Motivating Example: Kantorovich Problem

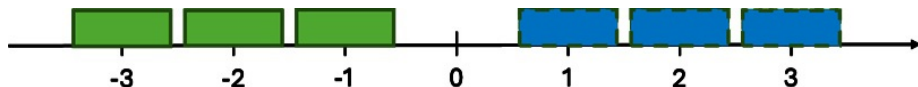


- In the first example, the joint distribution of  $(x, T_1(x))$  can be expressed as:

(Source, Target)	1	2	3	Total
-3	1/3	0	0	1/3
-2	0	1/3	0	1/3
-1	0	0	1/3	1/3
Total	1/3	1/3	1/3	1

- That is, each (deterministic) transport map identifies the corresponding joint distribution of the source data and their transported results.

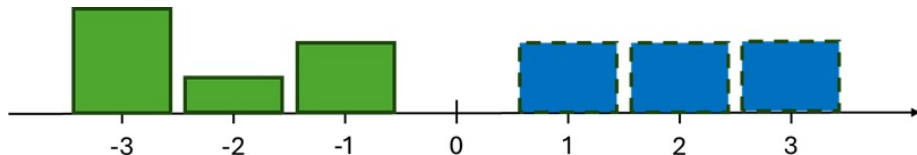
# Motivating Example: Kantorovich Problem



- We denote the joint probability measure by  $\pi(\cdot, \cdot)$  s.t.  $\pi(\cdot, \Omega_{\vec{x}'}) = \mathbb{P}(\cdot)$  and  $\pi(\Omega_{\vec{x}'}, \cdot) = \mathbb{Q}(\cdot)$ . Then, the transportation cost can be expressed as  $\int c(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}')$ .
- The Kantorovich Problem involves minimizing transportation costs among all joint distributions whose marginals are the source and target data distributions:

(Source, Target)	1	2	3	Total
-3	$p_{(-3,1)}$	$p_{(-3,2)}$	$p_{(-3,3)}$	1/3
-2	$p_{(-2,1)}$	$p_{(-2,2)}$	$p_{(-2,3)}$	1/3
-1	$p_{(-1,1)}$	$p_{(-1,2)}$	$p_{(-1,3)}$	1/3
Total	1/3	1/3	1/3	1

# Motivating Example: Kantorovich Problem



(Source, Target)	1	2	3	Total
-3	$p_{(-3,1)}$	$p_{(-3,2)}$	$p_{(-3,3)}$	$1/2$
-2	$p_{(-2,1)}$	$p_{(-2,2)}$	$p_{(-2,3)}$	$1/6$
-1	$p_{(-1,1)}$	$p_{(-1,2)}$	$p_{(-1,3)}$	$1/3$
Total	$1/3$	$1/3$	$1/3$	1

# Motivating Example: Kantorovich Problem

$$\min_{p_{(-3,1)}, \dots, p_{(-1,3)}} \left( c(-3, 1)p_{(-3,1)} + \dots + c(-1, 3)p_{(-1,3)} \right)$$

subject to

$$p_{(-3,1)} + p_{(-3,2)} + p_{(-3,3)} = 1/2,$$

$$p_{(-2,1)} + p_{(-2,2)} + p_{(-2,3)} = 1/6,$$

$$p_{(-1,1)} + p_{(-1,2)} + p_{(-1,3)} = 1/3,$$

$$p_{(-3,1)} + p_{(-2,1)} + p_{(-1,1)} = 1/3,$$

$$p_{(-3,2)} + p_{(-2,2)} + p_{(-1,2)} = 1/3,$$

$$p_{(-3,3)} + p_{(-2,3)} + p_{(-1,3)} = 1/3,$$

$$p_{(-3,1)} \geq 0, \dots, p_{(-1,3)} \geq 0.$$

(10)

# Wasserstein Distance

- The  $p$ -**Wasserstein distance** is defined as:

$$W_p(\mathbb{P}, \mathbb{Q}; d) := \left( \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int d^p(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}') \right)^{1/p}$$

where  $p \in [1, \infty)$ ,  $d$  is a metric defined on  $\mathcal{X}$ ,<sup>2</sup> and  $\Pi(\mathbb{P}, \mathbb{Q})$  is the set of all joint probability measures whose marginals are  $\mathbb{P}$  and  $\mathbb{Q}$ .

- Under some conditions, there exists an optimal transport map  $T^*$  that satisfies:

- ①  $W_p(\mathbb{P}, \mathbb{Q}; d) = \left( \int d^p(\vec{x}, T^*(\vec{x})) d\mathbb{P}(\vec{x}) \right)^{1/p}$
- ②  $\mathbb{P}(T^*(\vec{x})) = \mathbb{Q}(\vec{x})$

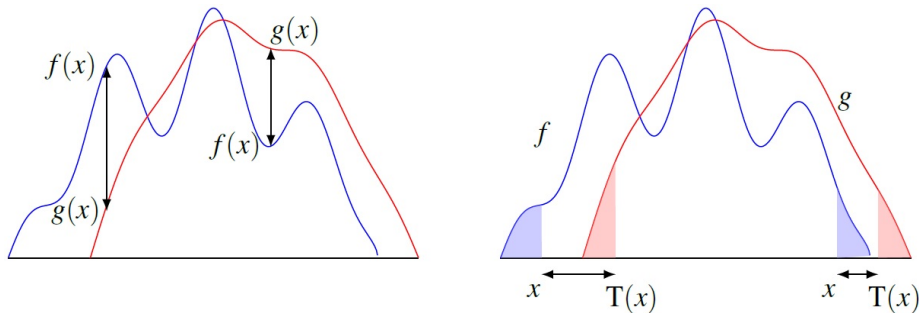
That is, Monge Problem can be equivalent to the Kantorovich Problem, and together they are simply referred to as the Monge-Kantorovich Problem.

---

<sup>2</sup>The Kantorovich Problem is generally defined with the cost function  $c$ .



# Wasserstein Distance



- The  $f$ -divergence utilizes density-ratios to quantify discrepancies between distributions.
- In contrast, Wasserstein distances utilize transportation costs.

The figure is from Santambrogio (2015).

# Wasserstein Distance

- For example, when  $d$  is the Euclidean norm,  $W_p$  becomes the Mallows metric (Mallows, 1972), and has played an important role in deriving asymptotic properties of bootstrap estimators (Bickel and Freedman, 1981; Freedman, 1981).
- Wasserstein distances effectively quantify differences between high-dimensional distributions when their supports are in low-dimensional manifolds.

## Example

(Example 1 in Arjovsky et al., 2017) Let  $Z \sim U[0, 1]$ ,  $\vec{X} = (0, Z)^T$ , and  $G_\theta(Z) = (\theta, Z)^T$ .

- Intuitively,  $\mathcal{D}(\mathbb{P}_{n=\infty}, \mathbb{P}_\theta)$  should decrease as  $\theta$  vanishes.
  - $W_p(\mathbb{P}_{n=\infty}, \mathbb{P}_\theta; |\cdot|) = |\theta|$
  - $\text{JS}(\mathbb{P}_{n=\infty} \parallel \mathbb{P}_\theta) = \log 2$  if  $\theta \neq 0$  and 0 if  $\theta = 0$
  - $\text{KL}(\mathbb{P}_{n=\infty} \parallel \mathbb{P}_\theta) = \infty$  if  $\theta \neq 0$  and 0 if  $\theta = 0$
  - $\delta(\mathbb{P}_{n=\infty}, \mathbb{P}_\theta) = 1$  if  $\theta \neq 0$  and 0 if  $\theta = 0$

# 1–Wasserstein Distance: Duality

- When  $p = 1$  (called *Earth Mover's Distance*), duality holds (Villani et al., 2009; Villani, 2021):

$$W_1(\mathbb{P}, \mathbb{Q}; d) = \sup_{f, g: f(\vec{x}) + g(\vec{x}') \leq d(\vec{x}, \vec{x}')} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') \right). \quad (11)$$

- When we further assume that  $d(\vec{x}, \vec{x}') = \|\vec{x} - \vec{x}'\|$  for some norm  $\|\cdot\|$ , the dual form can be re-expressed as:

$$W_1(\mathbb{P}, \mathbb{Q}; d) = \sup_{\text{Lip}(f) \leq 1} \int f(\vec{x}) d\mathbb{P}(\vec{x}) - \int f(\vec{x}) d\mathbb{Q}(\vec{x}) \quad (12)$$

where  $\text{Lip}(f) := \max\{C \mid |f(\vec{x}) - f(\vec{x}')| \leq C\|\vec{x} - \vec{x}'\|\}$  represents the Lipschitz constant of  $f$ . Note that this dual form implies that  $W_1$  is an IPM.

# Duality: Sketch of Proof

$\sup_{f,g} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') - \int (f(\vec{x}) + g(\vec{x}')) \pi(\vec{x}, \vec{x}') \right) = 0$  when  $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$  and  $= \infty$  otherwise, which implies:

$$\begin{aligned}
 & \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int d(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}') \\
 &= \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \left( \int d(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}') \right. \\
 & \quad \left. + \sup_{f,g} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') - \int (f(\vec{x}) + g(\vec{x}')) d\pi(\vec{x}, \vec{x}') \right) \right) \\
 &= \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \sup_{f,g} \left( \int d(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}') + \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') \right. \\
 & \quad \left. - \int (f(\vec{x}) + g(\vec{x}')) d\pi(\vec{x}, \vec{x}') \right).
 \end{aligned}$$

# Duality: Sketch of Proof

Now, by exchanging the infimum and supremum:<sup>3</sup>

$$\begin{aligned}
 & \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \sup_{f, g} \left( \int d(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}') + \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') \right. \\
 & \quad \left. - \int (f(\vec{x}) + g(\vec{x}')) d\pi(\vec{x}, \vec{x}') \right) \\
 &= \sup_{f, g} \inf_{\pi \geq 0} \left( \int d(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}') + \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') \right. \\
 & \quad \left. - \int (f(\vec{x}) + g(\vec{x}')) d\pi(\vec{x}, \vec{x}') \right) \\
 &= \sup_{f, g} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') + \inf_{\pi \geq 0} \int (d(\vec{x}, \vec{x}') - f(\vec{x}) - g(\vec{x}')) d\pi(\vec{x}, \vec{x}') \right).
 \end{aligned}$$

<sup>3</sup>See Section 1.2. in Santambrogio (2015) for details on conditions to exchange them.

# Duality: Sketch of Proof

Note that  $\inf_{\pi \geq 0} \int (d(\vec{x}, \vec{x}') - f(\vec{x}) - g(\vec{x}')) d\pi(\vec{x}, \vec{x}') = 0$  if  $f(\vec{x}) + g(\vec{x}') \leq d(\vec{x}, \vec{x}')$  for all  $(\vec{x}, \vec{x}')$  and  $= -\infty$  otherwise. This implies:

$$\begin{aligned} & \sup_{f, g} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') + \inf_{\pi \geq 0} \int (d(\vec{x}, \vec{x}') - f(\vec{x}) - g(\vec{x}')) d\pi(\vec{x}, \vec{x}') \right) \\ &= \sup_{f, g: f(\vec{x}) + g(\vec{x}') \leq d(\vec{x}, \vec{x}')} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') \right). \end{aligned}$$

# Duality: Sketch of Proof

Next, we show that:

$$\begin{aligned} W_1(\mathbb{P}, \mathbb{Q}; d) &= \sup_{f, g: f(\vec{x}) + g(\vec{x}') \leq d(\vec{x}, \vec{x}')} \left( \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') \right) \\ &= \sup_{\text{Lip}(f) \leq 1} \int f(\vec{x}) d\mathbb{P}(\vec{x}) - \int f(\vec{x}) d\mathbb{Q}(\vec{x}). \end{aligned} \quad (13)$$

when we further assume that  $d(\vec{x}, \vec{x}') = \|\vec{x} - \vec{x}'\|$  for some norm  $\|\cdot\|$ .

**(i) LHS  $\leq$  RHS:** For any  $f$  and  $g$  s.t.  $f(\vec{x}) + g(\vec{x}') \leq d(\vec{x}, \vec{x}')$  for all  $(\vec{x}, \vec{x}')$ ,

$$f(\vec{x}) \leq f^d(\vec{x}) := \inf_{\vec{x}'} \left( d(\vec{x}, \vec{x}') - g(\vec{x}') \right) \leq d(\vec{x}, \vec{x}) - g(\vec{x}) = -g(\vec{x}). \quad (14)$$

This implies  $\int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}') \leq \int f^d(\vec{x}) d\mathbb{P}(\vec{x}) - \int f^d(\vec{x}) d\mathbb{Q}(\vec{x})$ .

# Duality: Sketch of Proof

Here,  $f^d$  is 1-Lipschitz continuous because:

$$f^d(\vec{x}_1) \leq \inf_{\vec{x}'} \left( d(\vec{x}_1, \vec{x}_2) + d(\vec{x}_2, \vec{x}') - g(\vec{x}') \right) = d(\vec{x}_1, \vec{x}_2) + f^d(\vec{x}_2) \quad (15)$$

and  $d(\vec{x}_1, \vec{x}_2) = d(\vec{x}_2, \vec{x}_1)$ . This implies  $\int f(\vec{x})d\mathbb{P}(\vec{x}) + \int g(\vec{x}')d\mathbb{Q}(\vec{x}') \leq \text{RHS}$ , which concludes the proof for  $\text{LHS} \leq \text{RHS}$ .

**(ii) LHS  $\geq$  RHS:** For any 1-Lipschitz continuous function  $f$  and  $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$ ,

$$\begin{aligned} \int d(\vec{x}, \vec{x}')d\pi(\vec{x}, \vec{x}') &\geq \int \left( f(\vec{x}) - f(\vec{x}') \right) d\pi(\vec{x}, \vec{x}') \\ &= \int f(\vec{x})d\mathbb{P}(\vec{x}) - \int f(\vec{x}')d\mathbb{Q}(\vec{x}'). \end{aligned} \quad (16)$$

Now, sequentially taking the infimum over  $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$  and the supremum over  $f : \text{Lip}(f) \leq 1$  yields  $\text{LHS} \geq \text{RHS}$ .



# Primal vs. Dual Problems

- For general cases with cost function  $c$ :

- 1 (Primal)  $\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int c(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}')$

- 2 (Dual)  $\sup_{f, g: f(\vec{x}) + g(\vec{x}') \leq c(\vec{x}, \vec{x}')} \int f(\vec{x}) d\mathbb{P}(\vec{x}) + \int g(\vec{x}') d\mathbb{Q}(\vec{x}')$

- When  $d(\vec{x}, \vec{x}') = \|\vec{x} - \vec{x}'\|$ :

- 1 (Primal)  $W_1(\mathbb{P}, \mathbb{Q}; d) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int d(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}')$

- 2 (Dual)  $W_1(\mathbb{P}, \mathbb{Q}; d) = \sup_{\text{Lip}(f) \leq 1} \int f(\vec{x}) d\mathbb{P}(\vec{x}) - \int f(\vec{x}) d\mathbb{Q}(\vec{x})$

# Multi-marginal Transport Problem

- The optimal transportation cost is generally defined for multiple distributions as follows:

① (Primal)

$$\inf_{\pi \in \Pi(\mathbb{P}^{(1)}, \dots, \mathbb{P}^{(M)})} \int c(\vec{x}^{(1)}, \dots, \vec{x}^{(M)}) d\pi(\vec{x}^{(1)}, \dots, \vec{x}^{(M)}) \quad (17)$$

where  $\Pi(\mathbb{P}^{(1)}, \dots, \mathbb{P}^{(M)})$  represents the set of all joint probability measures whose marginals are  $\mathbb{P}^{(1)}, \dots, \mathbb{P}^{(M)}$ .

② (Dual)

$$\sup_{f_1, \dots, f_M: f_1(\vec{x}^{(1)}) + \dots + f_M(\vec{x}^{(M)}) \leq c(\vec{x}^{(1)}, \dots, \vec{x}^{(M)})} \sum_{m=1}^M \int f_m(\vec{x}^{(m)}) d\mathbb{P}^{(m)}(\vec{x}^{(m)}) \quad (18)$$

# Outline

- 1 OPTIMAL TRANSPORT
- 2 WASSERSTEIN GENERATIVE MODELS
- 3 APPLICATION

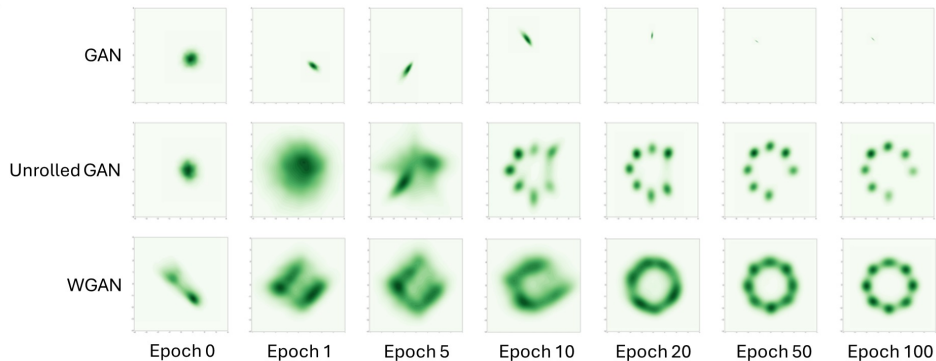
# 1-Wasserstein Distance: Wasserstein GAN

1. **Wasserstein GAN (WGAN):** WGANs model the class of 1-Lipschitz continuous functions using neural networks, denoted by  $f_\phi$ , with the goal of

$$\min_{\theta} \max_{f_\phi} \left( \int f_\phi(\vec{x}) d\mathbb{P}_n(\vec{x}) - \int f_\phi(\vec{x}) d\mathbb{P}_\theta(\vec{x}) \right). \quad (19)$$

- When the set  $\{f_\phi \mid \phi \in \Phi\}$  perfectly approximates the set  $\{f \mid \|f\|_L \leq 1\}$ , Equation (19) equals to  $\min_{\theta} W_1(\mathbb{P}_n, \mathbb{P}_\theta; d)$ .
- The 1-Lipschitz continuity condition can be relaxed to  $C$ -Lipschitz continuity for an arbitrary constant  $C$ : In this case, Equation (19) equals to  $C \min_{\theta} W_1(\mathbb{P}_n, \mathbb{P}_\theta; d)$ . To enforce this, WGANs clip weights and biases in neural network layers during training.

# 1-Wasserstein Distance: Wasserstein GAN



- WGANs performed better than GAN baselines in learning the distribution of Gaussian mixtures with 8 modes.

Images are edited from Arjovsky et al. (2017).

# 1-Wasserstein Distance: WGAN with Gradient Penalty

2. **WGAN with Gradient Penalty (WGAN-GP)**: Gulrajani et al. (2017) proposed a gradient-based penalty term to enforce the Lipschitz constraint in the dual form.
- For any  $\pi \in \Pi(\mathbb{P}_n, \mathbb{P}_\theta)$  and  $0 \leq t \leq 1$ , let  $\vec{X}^{(t)} := t\vec{X} + (1-t)\vec{X}'$  where  $(\vec{X}, \vec{X}') \sim \pi$ . Authors showed that the optimal 1-Lipschitz continuous function  $f^*$  satisfies:

$$\nabla_{\vec{X}^{(t)}} f^*(\vec{X}^{(t)}) = \frac{\vec{X}' - \vec{X}^{(t)}}{\|\vec{X}' - \vec{X}^{(t)}\|} \quad (20)$$

for any norm  $\|\cdot\|$ .

- Motivated by this, authors proposed to target the following:

$$\int f_\phi(\vec{x}) d\mathbb{P}_n(\vec{x}) - \int f_\phi(\vec{x}) d\mathbb{P}_\theta(\vec{x}) + \lambda \left( \int \left( \|\nabla_{\vec{x}} f_\phi(\vec{x})\|_2 - 1 \right)^2 \sum_t (t d\mathbb{P}_n(\vec{x}) + (1-t) d\mathbb{P}_\theta(\vec{x})) \right) \quad (21)$$

# $p$ -Wasserstein Distance: Wasserstein Autoencoder

**3. Wasserstein Autoencoder (WAE):** Tolstikhin et al. (2018) derived an alternative representation of the  $p$ -Wasserstein distance, and proposed WAEs based on the result.

- Let  $\Pi^\dagger(\mathbb{P}_n, \mathbb{P}_\theta; \mathbb{P}_Z)$  be the set of  $\pi^\dagger(\vec{x}, \vec{x}', \vec{z})$  s.t.  $\pi^\dagger(\vec{x}) = \mathbb{P}(\vec{x})$  and  $\pi^\dagger(\vec{x}', \vec{z}) = \mathbb{P}_\theta(\vec{x}'|\vec{z})\mathbb{P}(\vec{z})$ . Then,

$$\inf_{\pi(\vec{x}, \vec{x}') \in \Pi(\mathbb{P}_n, \mathbb{P}_\theta)} \int d^p(\vec{x}, \vec{x}') d\pi(\vec{x}, \vec{x}') \leq \inf_{\pi^\dagger(\vec{x}, \vec{x}', \vec{z}) \in \Pi^\dagger(\mathbb{P}_n, \mathbb{P}_\theta; \mathbb{P}_Z)} \int d^p(\vec{x}, \vec{x}') d\pi^\dagger(\vec{x}, \vec{x}') \quad (22)$$

- When the generator (or decoder) is deterministic, i.e.,  $\mathbb{P}_\theta(\vec{x}'|\vec{z})$  is Dirac whose mass is at  $G_\theta(\vec{z})$ , the equality holds and the RHS can be (re-)expressed as:

$$\inf_{\mathbb{Q}(\vec{z}|\vec{x}): \int q(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}) = p(\vec{z})} \int d^p(\vec{x}, G_\theta(\vec{z})) d\mathbb{Q}(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}). \quad (23)$$

# $p$ -Wasserstein Distance: Wasserstein Autoencoder

- These imply the following alternative representation of  $p$ -Wasserstein distances:

$$W_p(\mathbb{P}_n, \mathbb{P}_\theta; d) = \left( \inf_{\mathbb{Q}(\vec{z}|\vec{x}): \int q(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}) = p(\vec{z})} \int d^p(\vec{x}, G_\theta(\vec{z})) d\mathbb{Q}(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}) \right)^{1/p}. \quad (24)$$

Further details of the proof can be found in Section B of the Supplementary Material in the WAE paper.

- Based on this relation, WAEs introduce encoders  $q_\phi(\vec{z}|\vec{x})$  and target

$$\theta^* \in \arg \min_{\theta} \left( \inf_{\phi \in \Phi(\mathbb{P}_n)} \int d^p(\vec{x}, G_\theta(\vec{z})) d\mathbb{Q}_\phi(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}) \right)^{1/p} \quad (25)$$

where  $\Phi(\mathbb{P}_n) := \{\phi \mid \int q_\phi(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}) = p(\vec{z})\}$ .

- On the RHS,  $q_\phi(\vec{z}|\vec{x})$  can be viewed as an encoder. The constraint in the infimum ensures that the marginal distribution of the posterior distributions matches the prior distributions.



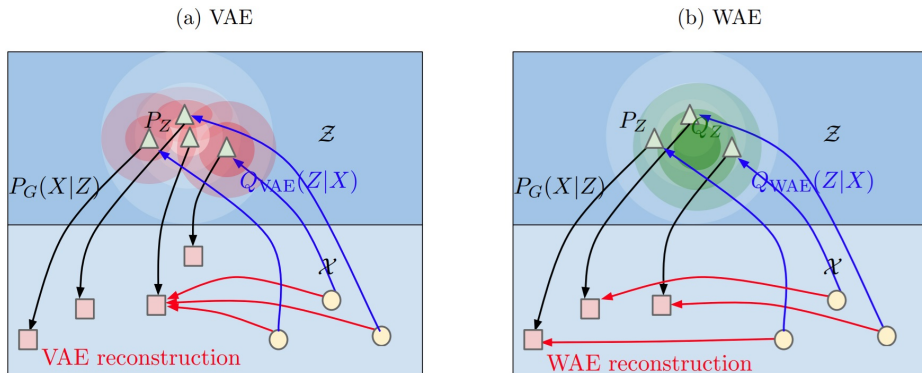
# $p$ -Wasserstein Distance: Wasserstein Autoencoder

- In implementation, WAEs introduce a penalty term to enforce the constraint on  $\phi$ . The loss can be expressed as:

$$\int d^p(\vec{x}, G_\theta(\vec{z})) d\mathbb{Q}_\phi(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}) + \lambda \mathcal{D}_{\vec{Z}} \left( \int q_\phi(\vec{z}|\vec{x}) d\mathbb{P}_n(\vec{x}), p(\vec{z}) \right) \quad (26)$$

where  $\mathcal{D}_{\vec{Z}}$  indicates the statistical distance applied to the distributions of  $\vec{Z}$ . WAEs typically use JS divergence and MMD (Maximum Mean Discrepancy) as measures for  $\mathcal{D}_{\vec{Z}}$ .

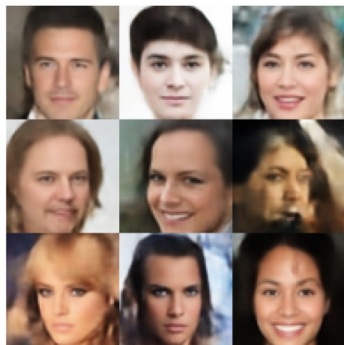
# $p$ -Wasserstein Distance: Wasserstein Autoencoder



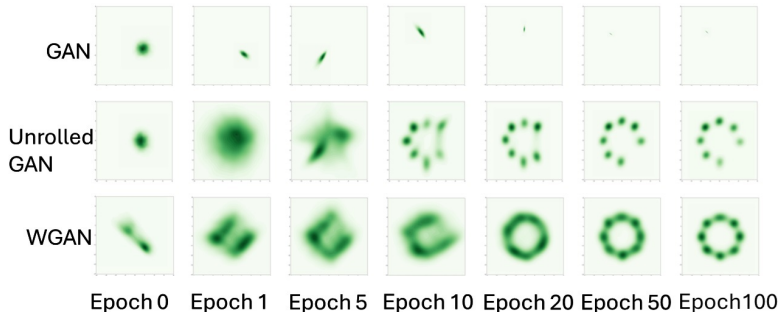
- Compared with the loss of VAEs, the negative ELBO, the penalty term changes from matching  $q_\phi(\vec{z}|\vec{x})$  directly with  $p(\vec{z})$  to matching  $\int q_\phi(\vec{z}|\vec{x})d\mathbb{P}_n(\vec{x})$  with  $p(\vec{z})$ .

The figure is from Tolstikhin et al. (2018).

# Generation Results: WAEs and WGANs



(a) Sharp generation results from WAEs



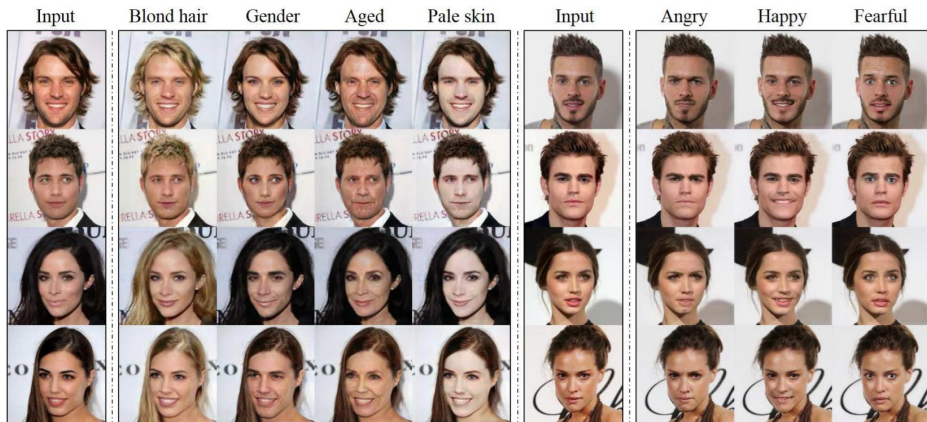
(b) Preventing mode collapse with WGANs

Images are edited from Arjovsky et al. (2017) and Tolstikhin et al. (2018).

# Outline

- 1 OPTIMAL TRANSPORT
- 2 WASSERSTEIN GENERATIVE MODELS
- 3 APPLICATION

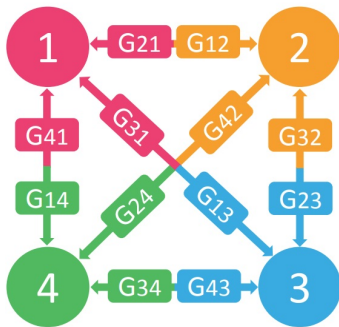
# StarGAN



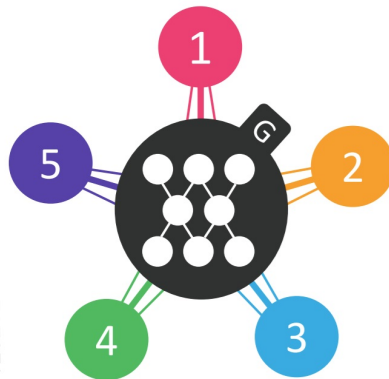
The figure is from Choi et al. (2018).

# StarGAN

(a) Cross-domain models

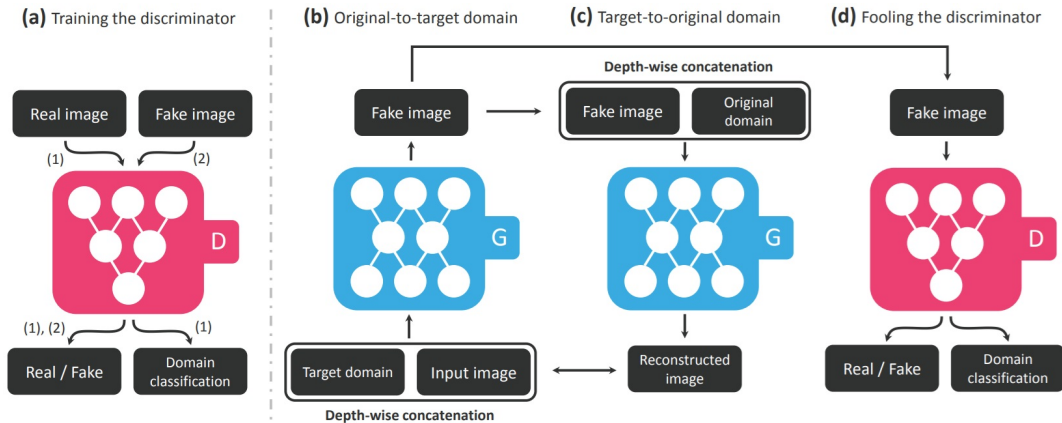


(b) StarGAN



The figure is from Choi et al. (2018).

# StarGAN



The figure is from Choi et al. (2018).

# StarGAN

- Let  $G_\theta(\cdot, \vec{c})$  be a translator that moves data from an arbitrary source domain to the target domain described by  $\vec{c}$ .
- For the adversarial loss component, StarGAN targets:

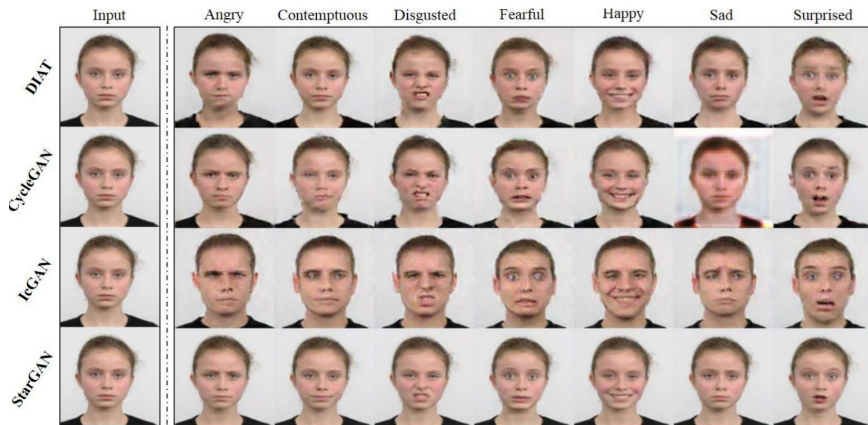
$$\int f_\phi(\vec{x}) d\mathbb{P}_n(\vec{x}) - \int f_\phi(G_\theta(\vec{x}, \vec{c})) d\mathbb{P}_n(\vec{x}, \vec{c}) \quad (27)$$

with an added gradient-penalty term.

- With the optimal  $\theta^*$ , the distribution of  $G_{\theta^*}(\vec{X}, \vec{C})$  matches the real distribution, which is a mixture of sub-populations from each data domain.
- The final objective includes classification and cycle-consistency losses to further enforce that, for a given  $\vec{c}$ , the distribution of  $G_{\theta^*}(\vec{X}, \vec{c})$  aligns with the sub-population for corresponding data domain.



# StarGAN



The figure is from Choi et al. (2018).

# References I

- Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.
- Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, 9(6):1196–1217.
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Freedman, D. A. (1981). Bootstrapping regression models. *The annals of statistics*, 9(6):1218–1228.
- Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

## References II

- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.
- Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30.
- Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *International conference on machine learning*, pages 1718–1727. PMLR.
- Mallows, C. L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, pages 508–515.
- Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. (2017). Unrolled generative adversarial networks. In *International Conference on Learning Representations*.
- Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkhäuser, NY*, 55(58-63):94.

# References III

- Tolstikhin, I., Bousquet, O., Gelly, S., and Schölkopf, B. (2018). Wasserstein auto-encoders. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Villani, C. (2021). *Topics in optimal transportation*, volume 58. American Mathematical Soc.
- Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.