Introduction to Optimal Transport

Mikhail Persiianov, Grigoriy Ksenofontov

Moscow, 2026

Overview

Course Organization

Birdseye of the course

Recap: what is generative modeling?

From generative modeling to domain translation

Optimal transport: core ideas

Wasserstein GAN: what it does — and doesn't — learn

Neural Optimal Transport (NOT): learning the map

Conclusion

Course Organization

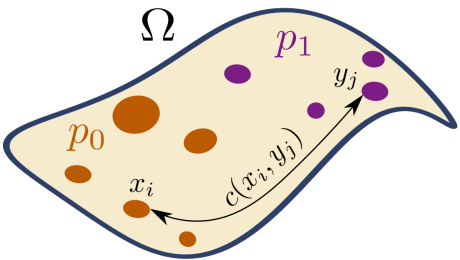
Course Organizers

Mikhail Persiianov

Grigoriy Ksenofontov

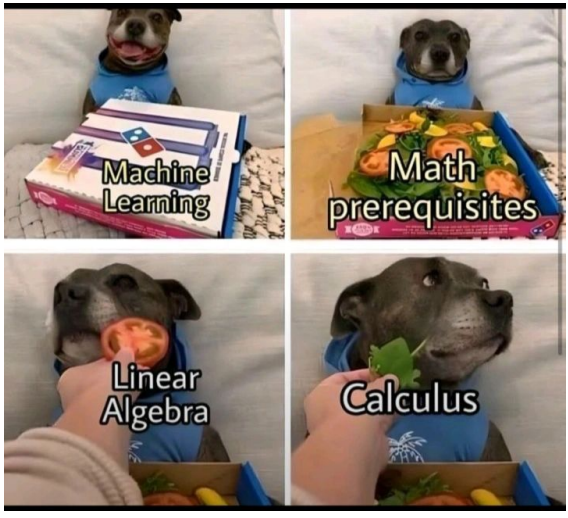
Course Topics

- Optimal Transport (Monge and Kantorovich formulations)
- Entropic Optimal Transport (Sinkhorn)
- Schrödinger Bridge Problem
- Score-based (diffusion) models for Optimal Transport
- Schrödinger Bridge models and bridge matching
- Flow-based models and Optimal Transport



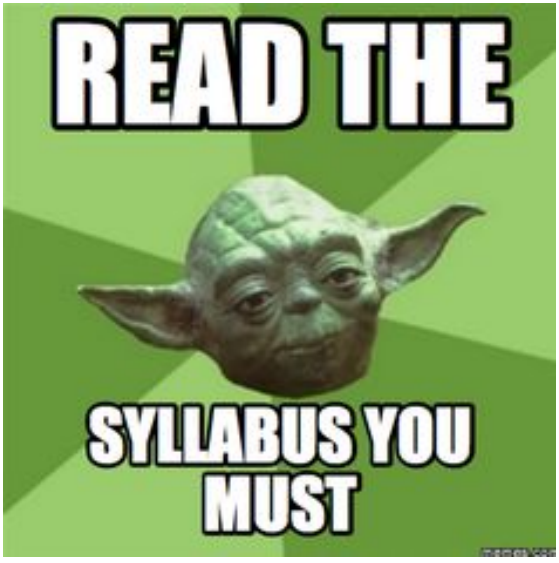
Course Prerequisites

- Calculus
- Probability Theory and Statistics
- Linear Algebra
- Optimization Methods
- Deep Generative Modeling
- Python (PyTorch)



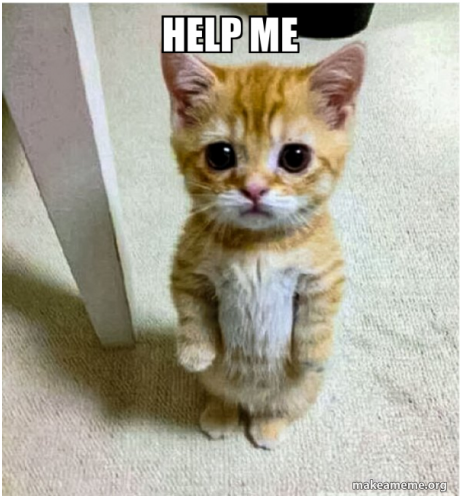
Course Schedule

- **10.02** Lecture 1
- **17.02** Lecture 2
- **24.02** Student Reports 1
- **03.03** Q/A Session HW1
- **10.03** Lecture 3
- **17.03** Lecture 4
- **24.03** Student Reports 2
- **31.03** Q/A Session HW2
- **07.04** Lecture 5
- **14.04** Lecture 6
- **21.04** Student Reports 3
- **28.04** Q/A Session HW3
- **05.05** Project Defence
- **12.05** Exam



Course Assistance and Consultation

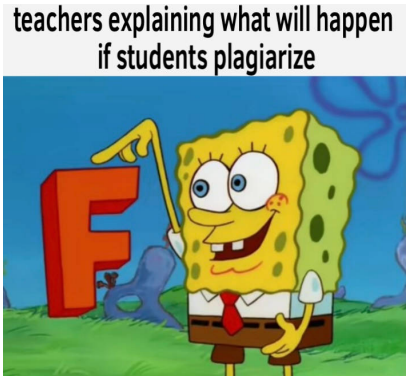
If you run into challenging homework problems, please don't hesitate to ask for help in our course Telegram group chat:



Reminder: Student Academic Integrity

Disciplinary penalties are imposed for:

- cheating, plagiarism, fabrication or falsification of data or results;
- copying, rewriting, paraphrasing, or summarizing of text, discoveries, or insights without acknowledging and/or citing the source;
- allowing other students to copy one's own work, using another student's solutions or code;
- Mindless use of ChatGPT and similar tools is prohibited.



Birdseye of the course

Generative modeling as mass transport

- Real data is described by a distribution p_{data}
- We start from a simple, known distribution

$$z \sim p_0 \quad \text{(Gaussian, Uniform)}$$

- A generator defines a transformation

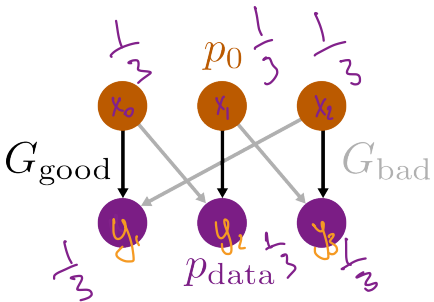
$$x = G(z), \qquad G_\sharp p_0 \approx p_{\text{data}}$$

- Generation is not just sampling — **it is moving probability mass**

Question:

How should mass move from p_0 to p_{data}?

This is a transport problem.



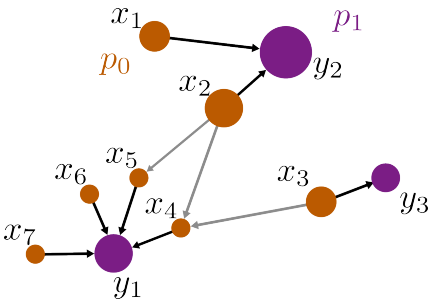
Simple distribution $p_0 \rightarrow$ structured data p_{data}

Why optimal transport?

- Matching distributions is **not enough**
- Many mappings can satisfy

$$T_\sharp p_0 = p_1$$

- But these mappings can be:
  - Arbitrary
  - Semantically wrong
  - Geometrically distorted
- We need a notion of:
  - **Cost** — how expensive is moving mass?
  - **Geometry** — what is "near" or "far"?
  - **Optimality** — which mapping should we prefer?



Source distribution $p_0 \rightarrow$ target distribution p_1

Optimal transport provides this structure.

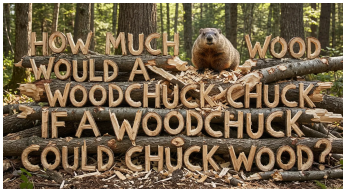
Recap: what is generative modeling?

What is generative modeling?

Generative modeling (GM, Generative AI, GenAI) refers to

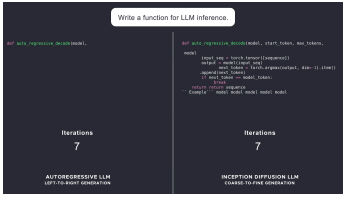
machine learning models that learn a data distribution and generate new, realistic samples.

Images[1]



Video[2]

Text[3]



Informally: neural networks that create new digital content, not just analyze existing data.

[1] Google AI Blog
[2] Kling AI
[3] Inception Labs AI Blog

Unconditional generative modeling: the problem

Data assumption. We observe samples

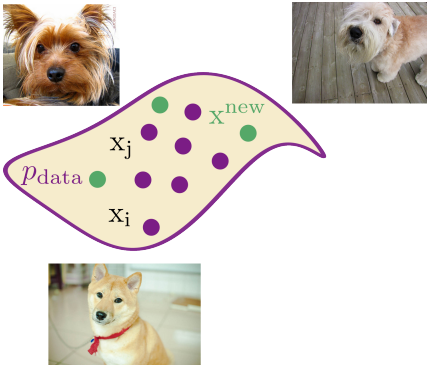
$$X = \{x_1, \ldots, x_N\} \text{ i.i.d. } x_i \sim p_{\text{data}},$$

where the true data distribution p_{data} is **unknown** and accessible only through samples (e.g. p_{data} on $\mathbb{R}^D$).

Generative modeling task. Given the dataset X, learn a model that allows:

- **Sampling:** generate new samples $x^{\text{new}} \sim p_{\text{data}}$,
- **(Optional) Density estimation:** evaluate $p_{\text{data}}(x)$.

In this course, we focus primarily on sampling.



Unconditional generative modeling: setup and learning

Canonical setup.

- Sample from a simple prior:

$$z \sim p_0(z) \quad (\mathcal{N}(0, I),\ \text{Uniform}[0,1]^d)$$

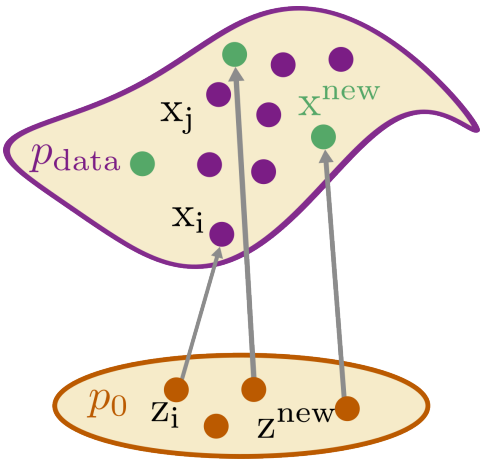
- Transform via a generator:

$$x = G_\theta(z), \qquad \underbrace{G_\theta \sharp p_0}_{p_\theta} \approx p_{\text{data}}$$

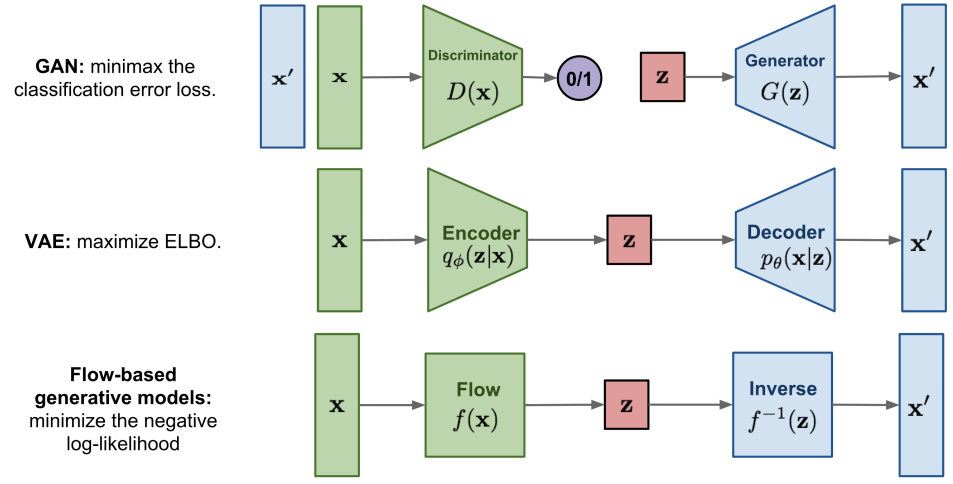
Learning principle.

- Choose a parametric model p_θ
- Define a discrepancy $\mathcal{L}(p_\theta, p_{\text{data}})$
- Optimize parameters: $\min_\theta \mathcal{L}(p_\theta, p_{\text{data}})$

Key perspective: Learning G_θ means learning how to *move probability mass*.



Common paradigms in generative modeling



4

[4] Lilian Weng (2018). **Flow-based Deep Generative Models.** Blog post comparing GAN, VAE, and flow-based models. URL: https://lilianweng.github.io/posts/2018-10-13-flow-models/.

What distribution matching does not specify

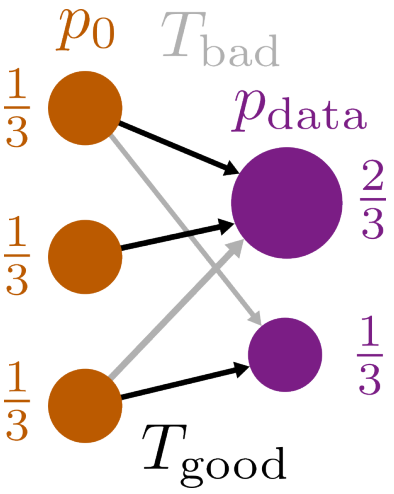
Most generative models answer:

- Do the distributions match?

They do not specify:

- How probability mass should move
- Which sample maps to which
- Whether the transformation is meaningful

This ambiguity becomes painful in domain translation.



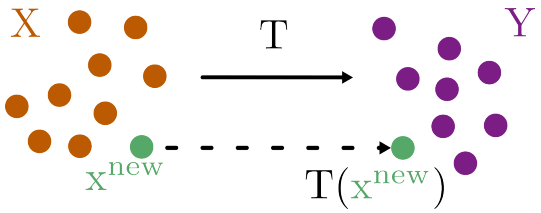
From generative modeling to domain translation

Domain translation: the task

Goal:
learn a *mapping* between two data domains

$$\mathcal{X} \longrightarrow \mathcal{Y}$$

using only samples from each domain.



In this course:

$$\mathcal{X}, \mathcal{Y} \in \left\{\mathbb{R}^D, \mathbb{S}^D\right\}, \qquad \mathbb{S} = \{1, \ldots, S\}$$

- $\mathbb{R}^D$: continuous data (images, signals, embeddings)
- $\mathbb{S}^D$: discrete / categorical data (tokens, labels, text)

Key requirement: the learned map should **generalize** to unseen data (similar, but not identical to the training samples).

Domain translation: applications

Figure 1: Day → Night translation for autonomous driving[7].

Figure 2: Sketch → Image translation for image editing[7].

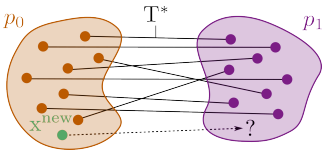
Other examples:

- **Medical imaging:** MRI → CT, image → segmentation[5].
- **Remote sensing:** satellite imagery, land-cover mapping[6].
- **Cross-modality:** LiDAR → RGB, night → day for perception or surveillance[7].

[5]Xiaofeng Liu et al. (2022). **"Act: Semi-supervised domain-adaptive medical image segmentation with asymmetric co-training".** In: *MICCAI*. Springer, pp. 66–76.
[6]Yadang Chen et al. (2022). **"Semi-supervised contrastive learning for few-shot segmentation of remote sensing images".** In: *Remote Sensing* 14.17, p. 4254.
[7]Gaurav Parmar et al. (2024). **"One-Step Image Translation with Text-to-Image Models".** In: *arXiv e-prints*.

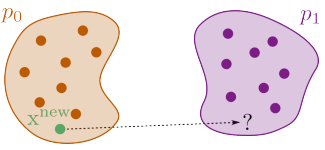
Supervision regimes in domain translation

Supervised



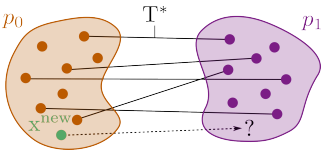
+ Well-posed: explicit input–output pairs.
– Paired data collection is expensive.

Unsupervised (our focus)



+ Unpaired data is easy to collect.
– Multiple valid mappings (solution ambiguity).

Semi-supervised



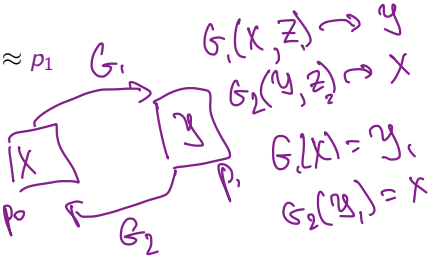
+ Combines weak pairing with abundant unpaired data.
– Limited theory and principled algorithms.

Question: *which of the generative modeling methods we know can be applied in each setting?*

GAN-based domain translation and its limitations

Setup:

$$G_\theta : \mathcal{X} \to \mathcal{Y}, \quad G_\theta \sharp p_0 \approx p_1$$



- Only the **marginal distributions** are matched[8]
- Many generators satisfy $G_\theta \sharp p_0 = p_1$
- GANs do **not enforce**:
  - Sample-wise correspondence
  - Geometric consistency
  - Minimal distortion
- **Typical failure modes:** mode mixing, semantic mismatch, arbitrary mappings[9]

Good discriminator score $\neq$ meaningful transport.

[8] Ian Goodfellow et al. (2020). **"Generative adversarial networks".** In: *Communications of the ACM* 63.11, pp. 139–144.

[9] Jun-Yan Zhu et al. (2017). **"Unpaired image-to-image translation using cycle-consistent adversarial networks".** In: *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232.

The Missing Ingredient: Geometry

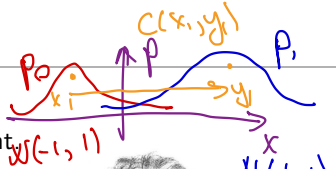
What standard GANs lack:

- A cost for moving probability mass
- A preference for minimal distortion
- A notion of optimality

Optimal transport provides exactly this structure, making the mapping meaningful.

Optimal transport: core ideas

The Monge problem (1781)[10]



Let p_0, p_1 be probability distributions on $\mathbb{R}^D$ with finite first moment.

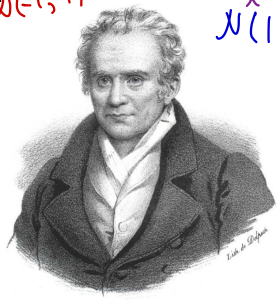
Question (Monge, 1781):

What is the most efficient way to move the mass of p_0 to p_1?

Formally, given a cost function $c : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_+$,

$$\min_T \ \mathbb{E}_{x\sim p_0}\big[c(x, T(x))\big] \quad \text{s.t.} \quad T_\sharp p_0 = p_1.$$

- T is a **deterministic transport map**
- Each point x is sent to exactly one location $T(x)$



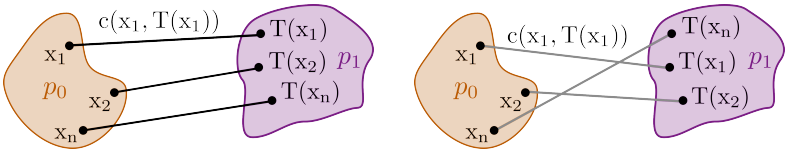
Gaspard Monge (1746–1818)

[10] Gaspard Monge (1781). **"Mémoire sur la théorie des déblais et des remblais".** In: *Mem. Math. Phys. Acad. Royale Sci.*, pp. 666–704.

Wasserstein-1 distance as a Monge problem

Special case (Wasserstein-1). If the cost is chosen as $c(x, y) = \|x - y\|_2$, then the optimal value of the Monge problem defines the **Wasserstein-1 distance**[11]:

$$\mathbb{W}_1(p, q) \overset{\text{def}}{=} \min_{T_\sharp p_0 = p_1} \mathbb{E}_{x \sim p_0} \|x - T(x)\|_2.$$



The minimizer T^* (if it exists) is called the **optimal transport map**.

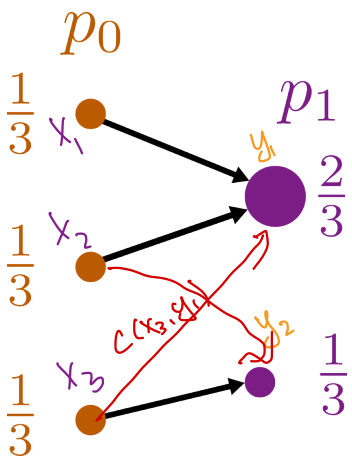
Informally, $\mathbb{W}_1$ measures the minimal effort required to move the mass of p_0 onto p_1.

[11] Cédric Villani et al. (2008). **Optimal transport: old and new.** Vol. 338. Springer.

[11] Here $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_D^2}$ denotes the Euclidean norm.

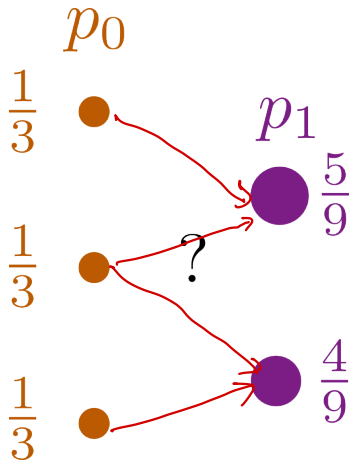
Interpreting the Monge formulation

- A **transport map** $T : \mathbb{R}^D \to \mathbb{R}^D$ assigns *each source point* x to *exactly one target point* $T(x)$
- Transport is **deterministic**:
  - **No splitting:** all mass at x goes to a single location $T(x)$
  - **No merging:** a target point receives mass from at most one source
- This enforces a **one-to-one assignment** of probability mass



Limitations of the Monge formulation

- An optimal transport map T may **not exist**
- Source and target distributions may have:
  - different supports (e.g. disconnected or lower-dimensional)
  - different intrinsic geometries
- The resulting optimization problem is **highly non-convex**



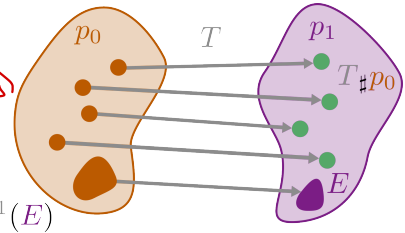
In many situations, optimal transport requires splitting mass, which Monge cannot represent.

Maps and push-forward measures

Map (transport): $T : \mathcal{X} \longrightarrow \mathcal{Y}$

Discrete intuition:

$$T_{\#} : \begin{cases} \delta_x \mapsto \delta_{T(x)} \\ \sum_i \delta_{x_i} \mapsto \sum_i \delta_{T(x_i)} \\ \sum_i a_i \delta_{x_i} \mapsto \sum_i a_i \delta_{T(x_i)} \end{cases}$$



General definition:

$$(T_{\#} p_0)(E) \stackrel{\text{def}}{=} p_0\big(T^{-1}(E)\big)$$

Change of variables:

$$\int_{\mathcal{Y}} g(y)\, d(T_{\#} p_0)(y) = \int_{\mathcal{X}} g(T(x))\, dp_0(x)$$

Density case (smooth T):

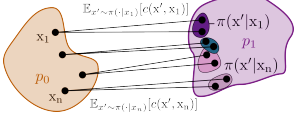
$$p_1(y) = p_0\big(T^{-1}(y)\big) \left|\det \nabla T^{-1}(y)\right|$$

The Kantorovich optimal transport problem[12]

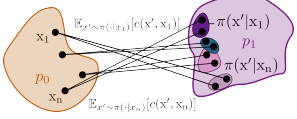
Kantorovich's formulation. The optimal transport problem is

$$\min_{\pi \in \Pi(p,q)} \mathbb{E}_{(x,x') \sim \pi} \big[c(x, x') \big],$$

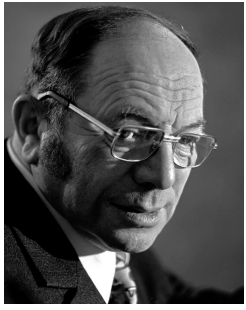
where $\Pi(p, q)$ is the set of **transport plans (couplings)**.
A transport plan $\pi(x, x')$ is a joint distribution on $\mathbb{R}^D \times \mathbb{R}^D$ satisfying the marginal constraints $\int \pi(x, x')\, dx' = p(x)$, $\int \pi(x, x')\, dx = q(x')$.



Optimal plan



Suboptimal plan



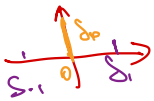
Leonid Kantorovich (1912–1986)

Mass can be split and merged to minimize the total transport cost.

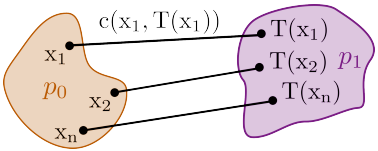
[12] Leonid V Kantorovich (1942). **"On the translocation of masses".** In: *Dokl. Akad. Nauk. USSR (NS)*. vol. 37, pp. 199–201.

Monge vs. Kantorovich Formulations of Optimal Transport

Monge's formulation



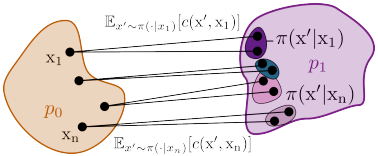
$$\inf_{T_\sharp p_0 = p_1} \mathbb{E}_{x\sim p_0}\left[c(x, T(x))\right]$$



- Asymmetric constraint: $T_\sharp p_0 = p_1$
- No mass splitting $\Rightarrow$ solution may not exist
- Example: $p_0 = \delta_0, \quad p_1 = \frac{1}{2}(\delta_{-1} + \delta_1)$

Kantorovich's formulation

$$\min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{(x,x')\sim\pi}\left[c(x, x')\right]$$



- Symmetric in (p_0, p_1)
- Allows mass splitting
- A minimizer π^* always exists

Kantorovich's formulation is a **relaxation** of Monge's. For many (p_0, p_1), the optimal plan is deterministic: $\pi^*(\cdot|x) = \delta_{T^*(x)}$, and both formulations yield the same $\mathbb{W}_1$.

Properties of the Wasserstein-1 Distance[13]

Let $\mathcal{P}_1(\mathbb{R}^D)$ denote the set of probability measures on $\mathbb{R}^D$ with finite first moment.

Theorem (Existence of optimal transport plans)

Let $p_0, p_1 \in \mathcal{P}_1(\mathbb{R}^D)$. The Kantorovich problem

$$\mathbb{W}_1^{(p_0,p_1)} = \min_{\pi \in \Pi(p_0,p_1)} \mathbb{E}_{(x,x')\sim\pi} \|x - x'\|_2$$

admits at least one minimizer $\pi^* \in \Pi(p_0, p_1)$. In general, the minimizer need not be unique.

Theorem (Wasserstein-1 is a metric)

The function $\mathbb{W}_1(\cdot,\cdot)$ defines a **metric** *on $\mathcal{P}_1(\mathbb{R}^D)$:*

- **(Triangle inequality)** $\mathbb{W}_1(p_1, p_3) \leq \mathbb{W}_1(p_1, p_2) + \mathbb{W}_1(p_2, p_3)$
- **(Identity of indiscernibles)** $\mathbb{W}_1(p, q) = 0 \iff p = q$
- **(Symmetry)** $\mathbb{W}_1(p, q) = \mathbb{W}_1(q, p)$

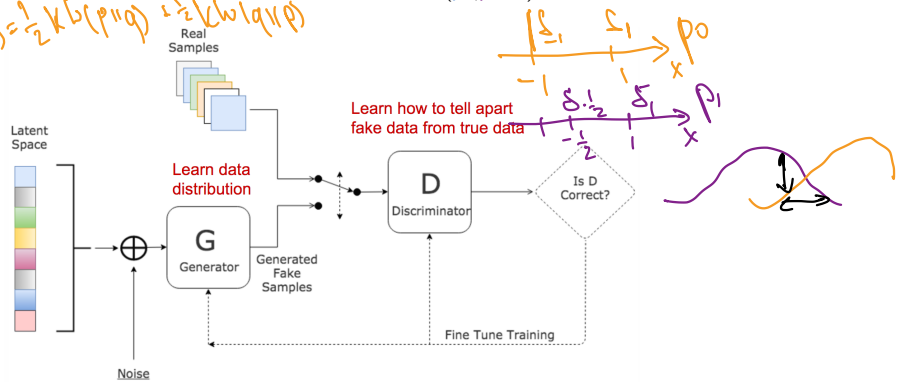
[13] Filippo Santambrogio (2015). **"Optimal transport for applied mathematicians"**. In.

Wasserstein GAN: what it does — and doesn't — learn

Wasserstein Loss for Generative Modeling

Goal. Map a simple latent distribution p_0 to the data distribution p_{data} using a generator G_θ.

$$\min_\theta \mathbb{W}_1(p_\theta, p_{\text{data}}) = \min_\theta \mathbb{W}_1(G_\theta \sharp p_0, p^*) = \min_\theta \Big\{ \min_{\pi \in \Pi(p_\theta, p_{\text{data}})} \mathbb{E}_{(x,x') \sim \pi} \|x - x'\|_2 \Big\}.$$



Question: Why use $\mathbb{W}_1$ instead of KL divergence?

Differentiability of $\mathbb{W}_1$ w.r.t. the generator[14]

Theorem (Differentiability of $\mathbb{W}_1$)

Let $\mathcal{X} \subset \mathbb{R}^D$ be a compact set (e.g., $\mathcal{X} = [-1, 1]^D$), and let $p_{\text{data}} \in \mathcal{P}(\mathcal{X})$ be the data distribution. Let p_0 be a fixed latent distribution on $\mathcal{Z}$, and let $G_\theta : \mathcal{Z} \to \mathcal{X}$ be a generator parameterized by θ. Define the generated distribution $p_\theta \stackrel{\text{def}}{=} G_\theta \sharp p_0$.

Then:

1. *If $G_\theta(z)$ is continuous in θ, then $\mathbb{W}_1(p_\theta, p_{\text{data}})$ is continuous in θ.*
2. *If $G_\theta(z)$ is locally Lipschitz in θ and satisfies mild regularity conditions, then $\mathbb{W}_1(p_\theta, p_{\text{data}})$ is differentiable almost everywhere in θ.*
3. **These statements do not hold** *for the KL or JS divergences.*

Key challenge. Although $\mathbb{W}_1$ has favorable optimization properties, its computation requires solving a **constrained OT problem**.

[14] Martin Arjovsky, Soumith Chintala, and Léon Bottou (2017). **"Wasserstein generative adversarial networks".** In: *International conference on machine learning*. PMLR, pp. 214–223.

Dual form of $\mathbb{W}_1$[15]

Theorem (Kantorovich–Rubinstein duality)

Let $p_\theta, p_{\text{data}} \in \mathcal{P}_1(\mathbb{R}^D)$. Then the Wasserstein–1 distance admits the dual representation

$$\mathbb{W}_1(p_\theta, p_{\text{data}}) = \max_{\|f\|_L \leq 1} \{\mathbb{E}_{x\sim p_\theta}[f(x)] - \mathbb{E}_{x\sim p_{\text{data}}}[f(x)]\},$$

where the maximum is taken over all **1-Lipschitz functions** $f : \mathbb{R}^D \to \mathbb{R}$ *(called gen).*

Recall that the Lipschitz seminorm is defined as

$$\|f\|_L \stackrel{\text{def}}{=} \sup_{x\neq x'} \frac{|f(x) - f(x')|}{\|x - x'\|_2}.$$

The constraint $\|f\|_L \leq 1$ can be interpreted (almost everywhere) as $\|\nabla_x f(x)\|_2 \leq 1$, by Rademacher's theorem.

[15] Cédric Villani et al. (2008). **Optimal transport: old and new.** Vol. 338. Springer.

Revisiting the Wasserstein loss for generative modeling

Theoretical objective.

Map a fixed latent distribution p_0 to the data distribution p_{data} using a generator G_θ.

$$\min_\theta \mathbb{W}_1(p_\theta, p_{\text{data}}) = \min_\theta \mathbb{W}_1(G_\theta \sharp p_0, p_{\text{data}}) = \min_\theta \sup_{\|f\|_L \leq 1} \{\mathbb{E}_{x\sim p_\theta}[f(x)] - \mathbb{E}_{x\sim p_{\text{data}}}[f(x)]\}.$$

Practical reality of Wasserstein GANs.

1. Restrict f to a neural network f_ω (the *critic*) and optimize over ω;
2. Use the change of variables $x = G_\theta(z)$ with $z \sim p_0$ (reparameterization trick);
3. Replace expectations over p_{data} with empirical averages over the dataset.

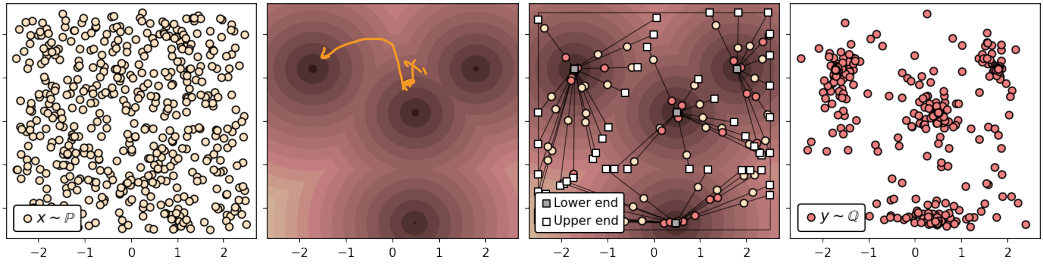
$$\min_\theta \max_\omega \left\{ \mathbb{E}_{z\sim p_0}\left[f_\omega(G_\theta(z))\right] - \frac{1}{M}\sum_{m=1}^{M} f_\omega(y_m) \right\}.$$

Problem: how to enforce the constraint $\|f_\omega\|_L \leq 1$?

Issues of Wasserstein GANs[16]

The paper introduces **controlled OT benchmarks**: pairs (p_0, p_1) with known optimal transport maps, exact transport cost $\mathbb{W}_1$, and ground-truth OT gradients.

Key empirical finding: most existing WGAN-based OT solvers poorly approximate the true $\mathbb{W}_1$ distance and its geometry.



[16]Alexander Korotin, Alexander Kolesov, and Evgeny Burnaev (2022). **"Kantorovich strikes back! Wasserstein GANs are not optimal transport?"** In: *Advances in Neural Information Processing Systems* 35, pp. 13933–13946.

Neural Optimal Transport (NOT): learning the map

Dual formulation of optimal transport

Kantorovich dual problem[17]

Let p_0, p_1 be probability distributions on $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^D$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a cost function. The Kantorovich OT cost admits the dual representation

$$\min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{(x,x') \sim \pi}[c(x, x')] \;=\; \sup_{g \oplus f \leq c} \left\{ \mathbb{E}_{x \sim p_0}[g(x)] + \mathbb{E}_{x' \sim p_1}[f(x')] \right\}.$$

The supremum is taken over functions[18] $g \in L^1(p_0), f \in L^1(p_1)$, such that for all $(x, x') \in \mathcal{X} \times \mathcal{Y}$,

$$g(x) + f(x') \;\leq\; c(x, x').$$

Interpretation. The functions g and f are called **Kantorovich (dual) potentials** and provide a lower bound on the transport cost.

[17] Cédric Villani et al. (2008). **Optimal transport: old and new.** Vol. 338. Springer.

[9] Here $L^1(p_0)$ denotes the space of functions integrable w.r.t. p_0, i.e., $g \in L^1(p_0) \leftrightarrow \int_{\mathcal{X}} |g(x)|\, p_0(x) \mathrm{d}x < \infty$.

Intuition Behind Dual Potentials

- Dual potentials act as **prices / value functions** assigned to locations:

$$g(x) \text{ for source } x \sim p_0, \qquad f(y) \text{ for target } y \sim p_1.$$

- The constraint

$$g(x) + f(y) \leq c(x, y)$$

  enforces **no arbitrage**: the value gained by moving mass from x to y cannot exceed the transport cost.
- Maximizing the dual objective finds the **tightest lower bound** on the true transport cost.
- At optimality, the inequality is tight on transported pairs:

$$g(x) + f(y) = c(x, y) \quad \text{(complementary slackness)}.$$

Wasserstein-1 case. The dual potential is a **1-Lipschitz function** whose gradient indicates the *direction and geometry* of optimal transport.

Dual potentials describe optimal transport geometry without explicitly moving mass.

c-transform and reduced dual formulation

Starting point (Kantorovich dual). For a cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$,

$$\min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{(x,x') \sim \pi}[c(x, x')] = \sup_{g \oplus f \leq c} \left\{ \mathbb{E}_{x \sim p_0}[g(x)] + \mathbb{E}_{x' \sim p_1}[f(x')] \right\}.$$

Fix $f \in L^1(p_1)$. The constraint $g(x) + f(x') \leq c(x, x')$ implies

$$g(x) \;\leq\; c(x, x') - f(x') \quad \forall x' \in \mathcal{Y}.$$

c-transform. The largest admissible g is the *c-transform* of f:

$$f^c(x) \stackrel{\text{def}}{=} \inf_{x' \in \mathcal{Y}} \left\{ c(x, x') - f(x') \right\}.$$

Reduced dual (single potential). Substituting $g = f^c$ yields

$$\min_{\pi \in \Pi(p_0, p_1)} \mathbb{E}_{(x,x') \sim \pi}[c(x, x')] = \sup_{f \in L^1(p_1)} \left\{ \mathbb{E}_{x \sim p_0}[f^c(x)] + \mathbb{E}_{x' \sim p_1}[f(x')] \right\}.$$

This formulation involves **one potential only** and has **no explicit constraints**.

Semi-dual max–min formulation

Starting point (reduced dual). From the c-transform formulation,

$$\min_{\pi\in\Pi(p_0,p_1)} \mathbb{E}_{(x,x')\sim\pi}[c(x,x')] = \sup_{f\in L^1(p_1)} \{\mathbb{E}_{x\sim p_0}[f^c(x)] + \mathbb{E}_{x'\sim p_1}[f(x')]\}\,.$$

Using the definition of the c-transform,

$$f^c(x) = \inf_{x'\in\mathcal{Y}} \{c(x,x') - f(x')\},$$

we obtain

$$\sup_f \left\{ \int_{\mathcal{X}} \inf_{x'\in\mathcal{Y}} \big(c(x,x') - f(x')\big)\, dp_0(x) + \int_{\mathcal{Y}} f(x')\, dp_1(x') \right\}.$$

Interchanging infimum and integral[18] yields

$$\sup_f \inf_{T:\mathcal{X}\to\mathcal{Y}} \left\{ \int_{\mathcal{X}} c(x, T(x))\, dp_0(x) - \int_{\mathcal{X}} f(T(x))\, dp_0(x) + \int_{\mathcal{Y}} f(x')\, dp_1(x') \right\}.$$

Semi-dual form.

$$\min_{\pi\in\Pi(p_0,p_1)} \mathbb{E}_{(x,x')\sim\pi}[c(x,x')] = \sup_f \inf_T \mathcal{L}(f,T),$$

where f is a dual potential and T acts as a transport map.

[18] Omar Anza Hafsa and Jean-Philippe Mandallena (2003). **"Interchange of infimum and integral".** In: *Calculus of Variations and Partial Differential Equations* 18.4, pp. 433–449.

Practical considerations: empirical semi-dual

Theoretical objective (semi-dual). From the semi-dual formulation,

$$\sup_{f} \inf_{T} \mathcal{L}(f, T),\ \mathcal{L}(f, T) = \int_{\mathcal{X}} c(x, T(x))\, dp_0(x) - \int_{\mathcal{X}} f(T(x))\, dp_0(x) + \int_{\mathcal{Y}} f(x')\, dp_1(x').$$

Practical setting. We only observe empirical samples $\{x_n\}_{n=1}^N \sim p_0,\ \{y_m\}_{m=1}^M \sim p_1$, and restrict T and f to parametric families.

Empirical semi-dual objective. Let $T_\theta : \mathcal{X} \to \mathcal{Y}$ (transport map) and $f_\omega : \mathcal{Y} \to \mathbb{R}$ (dual potential) be neural networks. We optimize[19]

$$\sup_{\omega} \inf_{\theta} \widehat{\mathcal{L}}(\omega, \theta) = \sup_{\omega} \inf_{\theta} \left\{ \frac{1}{N} \sum_{n=1}^{N} \left[c(x_n, T_\theta(x_n)) - f_\omega(T_\theta(x_n)) \right] + \frac{1}{M} \sum_{m=1}^{M} f_\omega(y_m) \right\}.$$

Optimization is performed via stochastic gradient ascent in ω and descent in θ.

[19] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2023). **"Neural Optimal Transport".** In: *The Eleventh International Conference on Learning Representations*.

Qualitative Examples of NOT[20]

Deterministic (one-to-one)

Celeba (female) $\rightarrow$ anime
128×128 images

Outdoor $\rightarrow$ church
128×128 images

[20] Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev (2023). **"Neural Optimal Transport".** In: *The Eleventh International Conference on Learning Representations*.

Conclusion

Conclusion: Generative Modeling as Mass Transport

Big picture.

- Generative modeling can be viewed as **moving probability mass** from a simple prior to a complex data distribution.
- Learning a generator means learning *how mass moves*, not just how distributions match.
- **Distribution matching alone is insufficient**:
  - mappings may be arbitrary,
  - geometry may be distorted,
  - semantics may be lost.
- Meaningful generation and translation require **cost**, **geometry**, and **optimality**.

Optimal transport provides exactly this structure.

Conclusion: From OT to Neural Optimal Transport

Methodological lessons.

- **Optimal transport** formalizes mass movement via:
  - transport plans and maps,
  - dual potentials encoding geometry.
- **Wasserstein GANs** improve optimization stability, but do *not* reliably recover OT geometry or transport maps.
- **Neural Optimal Transport (NOT)** directly optimizes the semi-dual OT formulation:
  - learning transport maps T,
  - learning dual potentials f,
  - producing geometrically meaningful domain translations.

Generative modeling, domain translation, and OT are unified by one question:
How should probability mass move—optimally and meaningfully?