Optimal Transportation and Applications

May 16, 2025

Outline

Concepts

Denote the set of probability measures on $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$, and similarly for $\mathcal{P}(\mathcal{Y})$.
For $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, we denote by $\Pi(\mu, \nu)$ the set of all transport plans, i.e., probability measures π on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν:

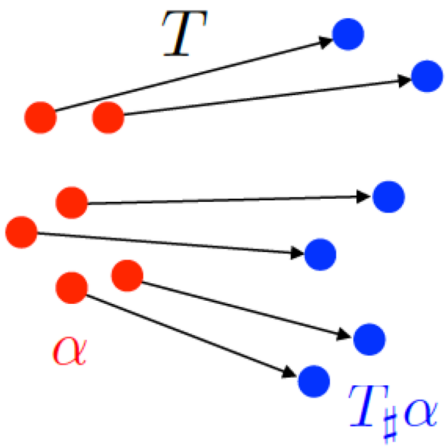
$$\Pi(\mu, \nu) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \pi(A \times \mathcal{Y}) = \mu(A), \pi(\mathcal{X} \times B) = \nu(B)\} \tag{1}$$

For a measurable map $T : \mathcal{X} \to \mathcal{Y}$ and a measure $\mu \in \mathcal{P}(\mathcal{X})$, the pushforward measure $T_{\#}\mu \in \mathcal{P}(\mathcal{Y})$ is defined by:

$$T_{\#}\mu(B) = \mu(T^{-1}(B)) \tag{2}$$

for all measurable sets $B \subset \mathcal{Y}$.

Monge problem



Let $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ be probability measures, and let $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ be a cost function. The Monge problem is to find a measurable map $T : \mathcal{X} \to \mathcal{Y}$ that solves:

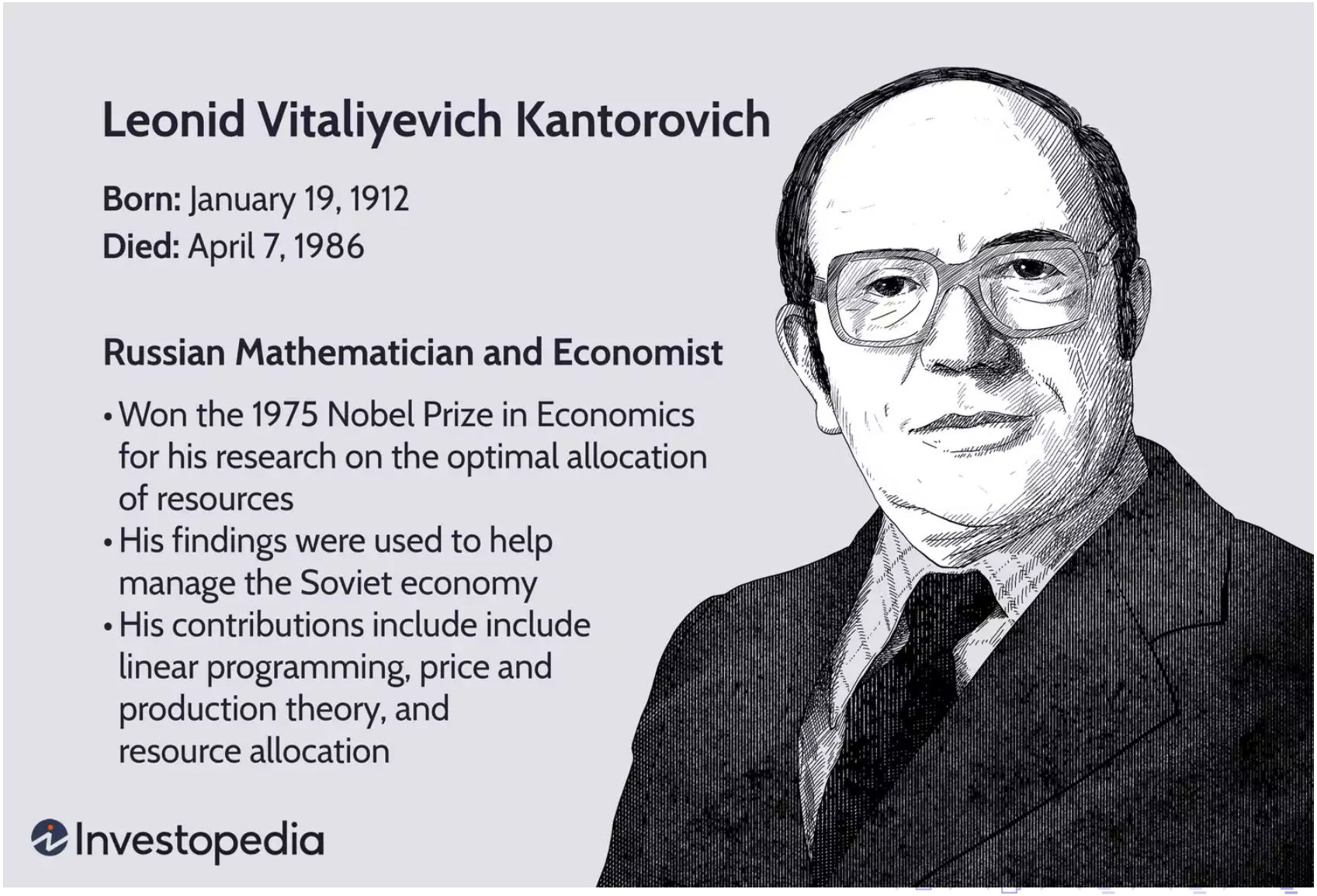
$$\inf_{T : T_{\#}\mu = \nu} \int_{\mathcal{X}} c(x, T(x))\, d\mu(x) \qquad (3)$$

Hardness of Monge problem

The Monge problem is challenging both theoretically and computationally for several reasons:

1. The constraint $T_{\#}\mu = \nu$ is highly non-linear in T: given distribution function $f(x)dx = \mu, g(x)dy = \nu$, with simple change of variable we have $f(x) = g(T(x))\det(\nabla T)$, which lacks tools for analysis.
2. The problem may not have a solution, e.g., if μ has atoms (point masses) while ν does not.
3. The formulation does not allow mass splitting, which can be problematic for discrete measures with different numbers of support points.

Relaxation of Monge problem: the Kantorovich problem



The Kantorovich problem

The Kantorovich problem is to find a transport plan $\pi \in \Pi(\mu, \nu)$ that solves:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y) \tag{4}$$

This is a linear programming problem in an infinite-dimensional space.

Unlike the Monge problem, the Kantorovich formulation:

1. Always has a solution under mild conditions on c (e.g., lower semicontinuity)
2. Allows mass splitting, making it applicable to a broader range of scenarios
3. Has a well-defined dual problem, which is helpful for computation.

Discrete formulation

In the discrete setting, when $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, the Kantorovich problem becomes a finite-dim linear program:

$$\min_{\pi \in \mathbb{R}_+^{n \times m}} \sum_{i=1}^{n} \sum_{j=1}^{m} c(x_i, y_j) \pi_{ij} \tag{5}$$

subject to:

$$\sum_{j=1}^{m} \pi_{ij} = a_i \quad \forall i \in \{1, \ldots, n\} \tag{6}$$

$$\sum_{i=1}^{n} \pi_{ij} = b_j \quad \forall j \in \{1, \ldots, m\} \tag{7}$$

Here, π_{ij} represents the amount of mass transported from location x_i to location y_j.

Wasserstein distance

When the cost function is a distance raised to a power, the optimal transport cost defines a distance between probability distributions, known as the Wasserstein distance. For $p \geq 1$, the p-Wasserstein distance between $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p \, d\pi(x, y) \right)^{1/p} \tag{8}$$

The 2-Wasserstein distance, also known as the Earth Mover's Distance, has various applications in machine learning such as image processing.

Kantorovich-Rubinstein duality for OT in $\mathbb{R}^n$

The dual formulation of the Kantorovich problem is:

$$\sup_{(\phi,\psi)\in\Phi_c} \left\{ \int_{\mathcal{X}} \phi(x)\, d\mu(x) + \int_{\mathcal{Y}} \psi(y)\, d\nu(y) \right\} \tag{9}$$

where Φ_c is the set of pairs of functions $\phi : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ and $\psi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ that satisfy:

$$\phi(x) + \psi(y) \leq c(x,y) \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \tag{10}$$

Application of duality in Neural OT[1]

Eq. (9) is proved to be equal to

$$W_2^2(\nu,\mu) = C_{\nu,\mu} - \inf_{f \in \mathbf{CVX}} \{\mathbb{E}_\nu[f(X)] + \mathbb{E}_\mu[f^*(Y)]\} \tag{11}$$

where **CVX** stands for the set of convex functions, $C_{\nu,\mu} := \{\mathbb{E}_\nu[\|X\|^2] + \mathbb{E}_\mu[\|Y\|^2]\}$, and the $f^*(x) = \min_{x'} < x', x > -f(x')$ is the conjugate function of f, which is always convex. The **CVX** condition restricts the search space for f which becomes handy for design of optimization algorithms.

[1] Makkuva, Ashok, et al. "Optimal transport mapping via input convex neural networks." International Conference on Machine Learning. PMLR, 2020.

Computation of f^*

Convex conjugate function f^* is not available explicitly in most of applications, but can be characterized as

$$f^*(y) = \sup_{g \in \mathbf{CVX}} \langle y, \nabla g(y) \rangle - f(\nabla g(y)) \tag{12}$$

with the maximum being achieved at $g = f^*$, the semi-dual formulation (11) can be rewritten as

$$W_2^2(\nu, \mu) = \sup_{f \in \mathbf{CVX}} \inf_{g \in \mathbf{CVX}} \mathcal{V}_{\nu,\mu}(f, g) + C_{\nu,\mu}, \tag{13}$$

where $\mathcal{V}_{\nu,\mu}(f, g)$ is a functional of f and g defined as

$$\mathcal{V}_{\nu,\mu}(f,g) = -\mathbb{E}_\nu[f(X)] - \mathbb{E}_\mu[\langle Y, \nabla g(Y) \rangle - f(\nabla g(Y))]. \tag{14}$$

Continued

Since now the only variables are convex functions f, g with no constraints, they can be learned by neural network. Check for example [2][3].

[2] Korotin, Alexander, Daniil Selikhanovych, and Evgeny Burnaev. "Neural Optimal Transport." The Eleventh International Conference on Learning Representations.

[3] Amos, Brandon, Lei Xu, and J. Zico Kolter. "Input convex neural networks." International conference on machine learning. PMLR, 2017.

The Breiner Potential

One of the most celebrated results in optimal transport theory is Brenier's theorem, which characterizes the solution to the Monge problem when the cost function is the squared Euclidean distance.

Theorem (Brenier)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be probability measures where μ is absolutely continuous with respect to Lebesgue measure. Then:

1. *There exists a unique (up to μ-null sets) optimal transport map T that solves the Monge problem with cost $c(x, y) = \|x - y\|^2$.*
2. *This optimal map is given by $T = \nabla\varphi$ for some convex function $\varphi : \mathbb{R}^d \to \mathbb{R}$.*
3. *The optimal transport plan in the Kantorovich problem is unique (a.e.) and is given by $\pi = (id \times \nabla\varphi)_{\#}\mu$.*

Why is Brenier theorem important?

Brenier's theorem establishes a profound connection between optimal transport and convex analysis. It states that the optimal transport map is the gradient of a convex function, which has significant implications for computational methods and applications.

Remark

Noticing that the Brenier theorem holds with constraint *almost everywhere* (a.e.), which means that exceptions occur in a null set (a set with measure 0). **This indicates the fundamental difference between the fixed support (with finite support size) and the free support case: the former cannot be treated as a Monge problem, so you would have to study the coupling instead of the transport map.**

Dynamic optimal transport

The Benamou-Brenier formulation [4] recasts the Wasserstein-2 distance as a fluid-dynamics problem:

$$W_2^2(\mu_0, \mu_1) = \inf_{(\rho, v)} \int_0^1 \int_{\mathbb{R}^d} \rho_t(x) \|v_t(x)\|^2 \, dx \, dt \tag{15}$$

subject to:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \quad \rho_0 = \mu_0, \quad \rho_1 = \mu_1 \tag{16}$$

where ρ_t is the density at time t and v_t is the velocity field. Eq. (16) is called the continuity equation.

[4] Benamou, Jean-David, and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem." Numerische Mathematik 84.3 (2000): 375-393.

(Continued) Dynamic OT

This formulation has several advantages:

1. It provides a continuous **interpolation** between the source and target measures (Which can serve as training data for generation model)
2. It has connections to fluid dynamics and PDEs
3. It can be generalized to handle additional constraints or costs

McCann's interpolation

The solutions to the Benamou-Brenier problem define geodesics in the Wasserstein space. Specifically, for $t \in [0,1]$, the measures ρ_t trace the unique constant-speed geodesic connecting μ_0 and μ_1 in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.
When μ_0 is absolutely continuous (i.e., having density function a.e.), the geodesic can be written explicitly in terms of the optimal transport map T from μ_0 to μ_1:

$$\rho_t = ((1-t)\cdot id + t \cdot T)_{\#}\mu_0 \tag{17}$$

This expression, known as McCann's interpolation, provides a simple way to visualize the optimal transport between two measures.

Connections with Fokker-Planck equation

The prestigious diffusion model[5] is based on the Fokker-Planck function, which describes the diffusion process:

$$\mathrm{d}X = f(X,t)\mathrm{d}t + \sigma(X,t)\,\mathrm{d}W, \tag{18}$$

$f(X,t) : \mathbb{R}^n \times [0,1] \to \mathbb{R}^n$ is the divergence vector field at the diffusion time t, which is learned with neural network, $\sigma(X,t) \in \mathbb{R}^{n\times n}$ is some coefficient matrix, $\mathrm{d}W \in \mathbb{R}^n$ is the Brown motion in $\mathbb{R}^n$.

[5] A good tutorial is https://diffusion.csail.mit.edu/

The Fokker-Planck equation

If $x_t \sim p(x,t)$, then $p(x,t)$ satisfies the Fokker-Planck equation:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot \big(f(x,t)p(x,t)\big) + \nabla \cdot \big(D(x,t)\nabla p(x,t)\big),$$

where $D(x,t) = \frac{1}{2}\sigma(x,t)\sigma^\top(x,t)$. Recall the continuity equation (16) in the formula of dynamic OT, we have the following observation:

Remark

The solution v_t in the dynamic OT (15) can be seen as the evolution dynamics of $dX_t = v_t(X_t)dt$ from distribution μ_0 to μ_1, such that the motion energy is minimized.

Applications in generative model

We will spend several pages to illustrate the generative model, and how is the ideas in OT applied to it. In particular, we introduce OT's application in [Lipman22][6] and [Liu22][7]. The rectified flow, proposed in [Liu22], is the foundation of the SOTA generative models, such as Stable Diffusion 3, FLUX and JanusFlow.

[6] Lipman, Yaron, et al. "Flow Matching for Generative Modeling." The Eleventh International Conference on Learning Representations.

[7] Liu, Xingchao, Chengyue Gong, and Qiang Liu. "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow." The Eleventh International Conference on Learning Representations.

Generative model: problem setting

The problem of generating can be reduced to the following:

Problem

Given empirical observations of two distributions π_0, π_1 on $\mathbb{R}^d$, find a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$, which, in the infinite data limit, yields $Z_1 := T(Z_0) \sim \pi_1$ when $Z_0 \sim \pi_0$, that is, (Z_0, Z_1) is a coupling (a.k.a transport plan) of π_0 and π_1.

Generative model: applications

1. **Generative modeling**: This is the case when π_1 is an **empirically observed unknown distribution** (of e.g., images), and π_0 an elementary distribution, such as the standard Gaussian distribution. We are interested in finding a nonlinear transform that turns a point drawn from π_0 to point that follows the data distribution π_1.
2. **Transfer modeling**: This is the case when **both π_0 and π_1 are empirically observed unknown distributions**, and we want to build a procedure to transfer a data point from π_0 to a point that follows π_1, or vice versa. This task admits enormous applications, such as domain adaption in transfer learning, image editing, and sim2real in robotics.

Generative model: example

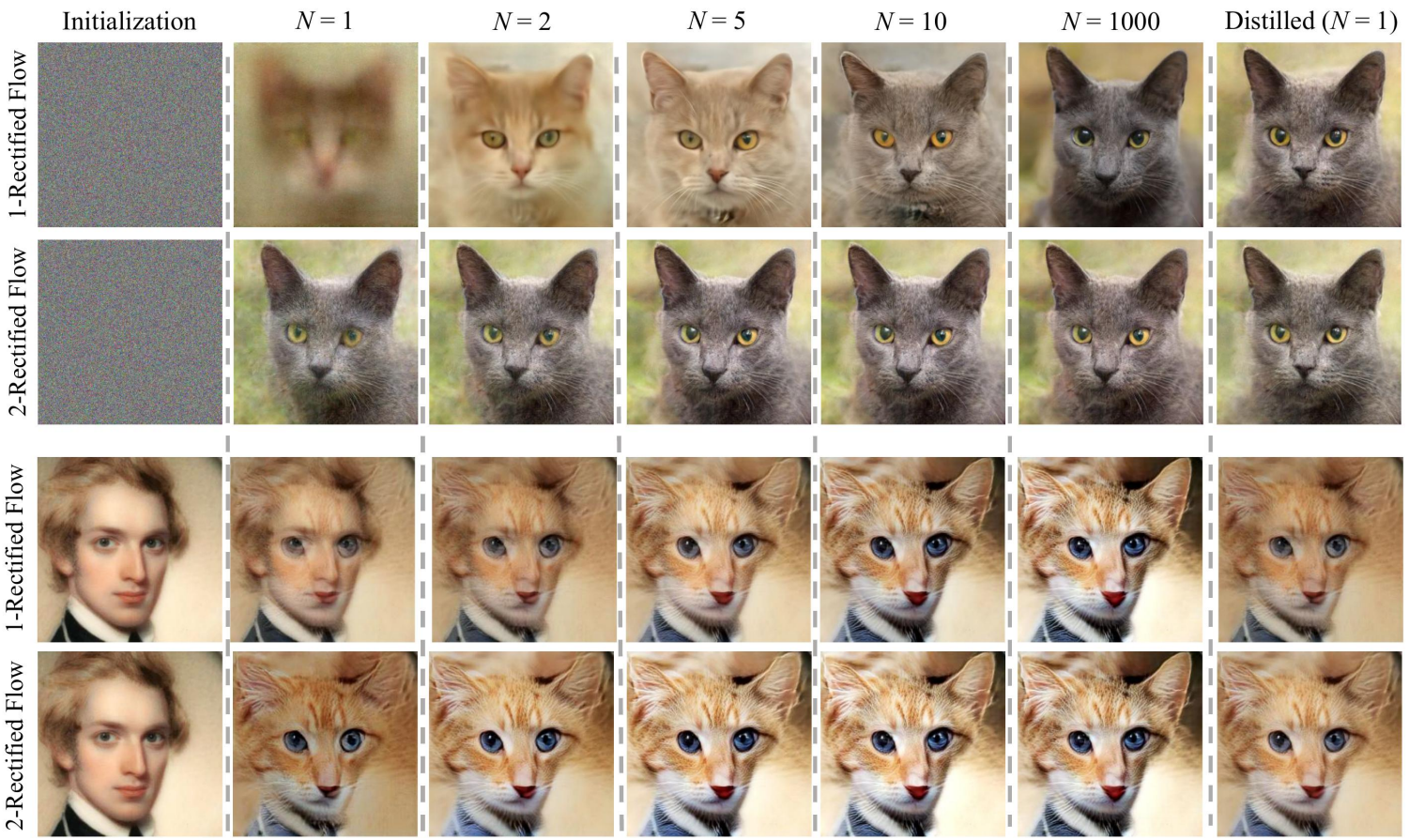


Figure: Generating (upper 2 rows) and transferring (lower 2 rows) of rectified flow.

Generative model: trained to simulate MaCann's interpolation

Recall that in the diffusion model, the neural network simulates the divergence vector field $f(X, t)$ in Eq. (18). We will use $v(X, t)$ to denote $f(X, t)$ for consistency with [Lipman22].
In order to "causalize" the interpolation process X_t, by "projecting" it to the space of causally simulatable ODEs of form $\mathrm{d}Z_t = v(Z_t, t)\mathrm{d}t$. A natural way to the L2 projection on the velocity field, find v by minimizing the least squares loss with the line directions $X_1 - X_0$:

$$\min_v \int_0^1 \mathbb{E}\left[\|(X_1 - X_0) - v(X_t, t)\|^2\right] \mathrm{d}t. \tag{19}$$

continued

Theoretically, the solution can be represented using conditional expectation:

$$v(z,t) = \mathbb{E}[X_1 - X_0 \mid X_t = z],$$

which is the average of the directions of the lines passing through point z at time t.

Rectified flow

For the solution Z_t obtained from samples (X_0, X_1), denote the rectified flow $\boldsymbol{Z} = \{Z_t : t \in [0,1]\}$ induced from (X_0, X_1) by $\boldsymbol{Z} = \text{Rectflow}((X_0, X_1))$. Then the following properties hold:

Theorem

1. The ODE trajectories Z_t and the interpolation X_t have the same marginal distributions, that is,

$$\text{Law}(Z_t) = \text{Law}(X_t), \quad \forall t \in [0,1].$$

Hence, (Z_0, Z_1) forms a coupling of π_0 and π_1.
2. (Z_0, Z_1) guarantees to yield no larger transport cost than (X_0, X_1) simultaneously for 'all convex cost functions' c, that is,

$$\mathbb{E}[c(Z_1 - Z_0)] \leq \mathbb{E}[c(X_1 - X_0)], \quad \forall \text{ convex } c\colon \mathbb{R}^d \to \mathbb{R}.$$

Why rectified flow?

The property 2 of rectified flow indicates that, applying the Rectflow($\cdot$) operator recursively yields a sequence of rectified flows, whose transport cost descends. Recall the formular of dynamic optimal transport, the rectified flow is actually performing gradient descent for dynamic OT[8].
Since the McCann's interpolation gives a gradient field that is constant (invariant to time t), we do not have to use too many steps to simulate the stochastic differential equation Eq. (18). This explains why the Rectified flow model only requires a very short number of steps to generate high-quality images, see Fig. **??**.

[8] Liu, Qiang. "Rectified flow: A marginal preserving approach to optimal transport." arXiv preprint arXiv:2209.14577 (2022).

JKO flow

Jordan et al. study diffusion processes under the lens of the OT metric and introduce a scheme that is now known as the JKO flow: Starting with ρ_0, and given a real-valued energy function $J : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ driving the evolution of the system, they define iteratively for $t \geq 0$, :

$$\rho_{t+1} = \arg \min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} J(\rho) + \frac{1}{2\tau} W^2(\rho, \rho_t)\,, \tag{20}$$

where τ is a time step parameter.

Analogy with Euclidean space

In Euclidean space, the gradient flow is written:

$$\begin{cases} \dot{X}(t) = -\nabla f(X(t)), \quad t > 0 \\ X(0) = \boldsymbol{x}_0 \end{cases} \tag{21}$$

1. forward scheme: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \nabla f(\boldsymbol{x}_k)$ (gradient descent)
2. backward scheme:

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \left(f(\boldsymbol{x}) + \frac{1}{2\alpha_k} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \right)$$

$$\Updownarrow$$

$$\mathbf{0} = \nabla f(\boldsymbol{x}_{k+1}) + \frac{1}{\alpha_k}(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$$

We got the *proximal point method*.

JKO flow

Recall the SDE form of the diffusion process, it is natural to regard JKO flow as the proximal point method for functionals in Wasserstein space. For illustrations on the theory, see chapter 15 of [Vil08][9].

[9]Villani, Cédric. Optimal transport: old and new. Vol. 338. Berlin: springer, 2008.

Tracking Cell Dynamics with scRNA-seq and Optimal Transport

Challenge in Time-Series scRNA-seq:

1. scRNA-seq captures gene expression at single-cell resolution across time.
2. Cells are destroyed during sequencing $\rightarrow$ no paired measurements over time.
3. Only unpaired snapshots are available at different time points.

This breaks the link between cells across time, which leads to:

$\Diamond$ Cannot track individual cell trajectories

$\Diamond$ Cannot directly observe gene expression dynamics

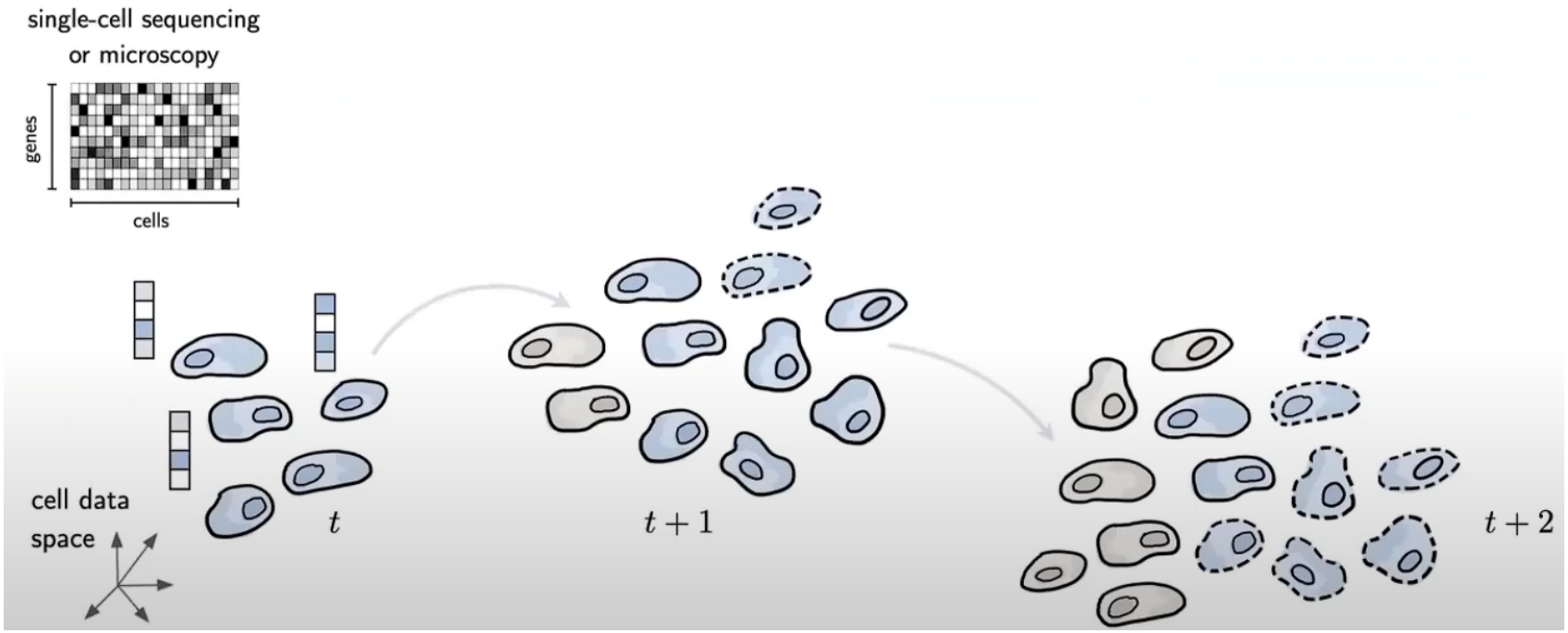
Continued

OT is a powerful tool to infer relationships between cells across time points.

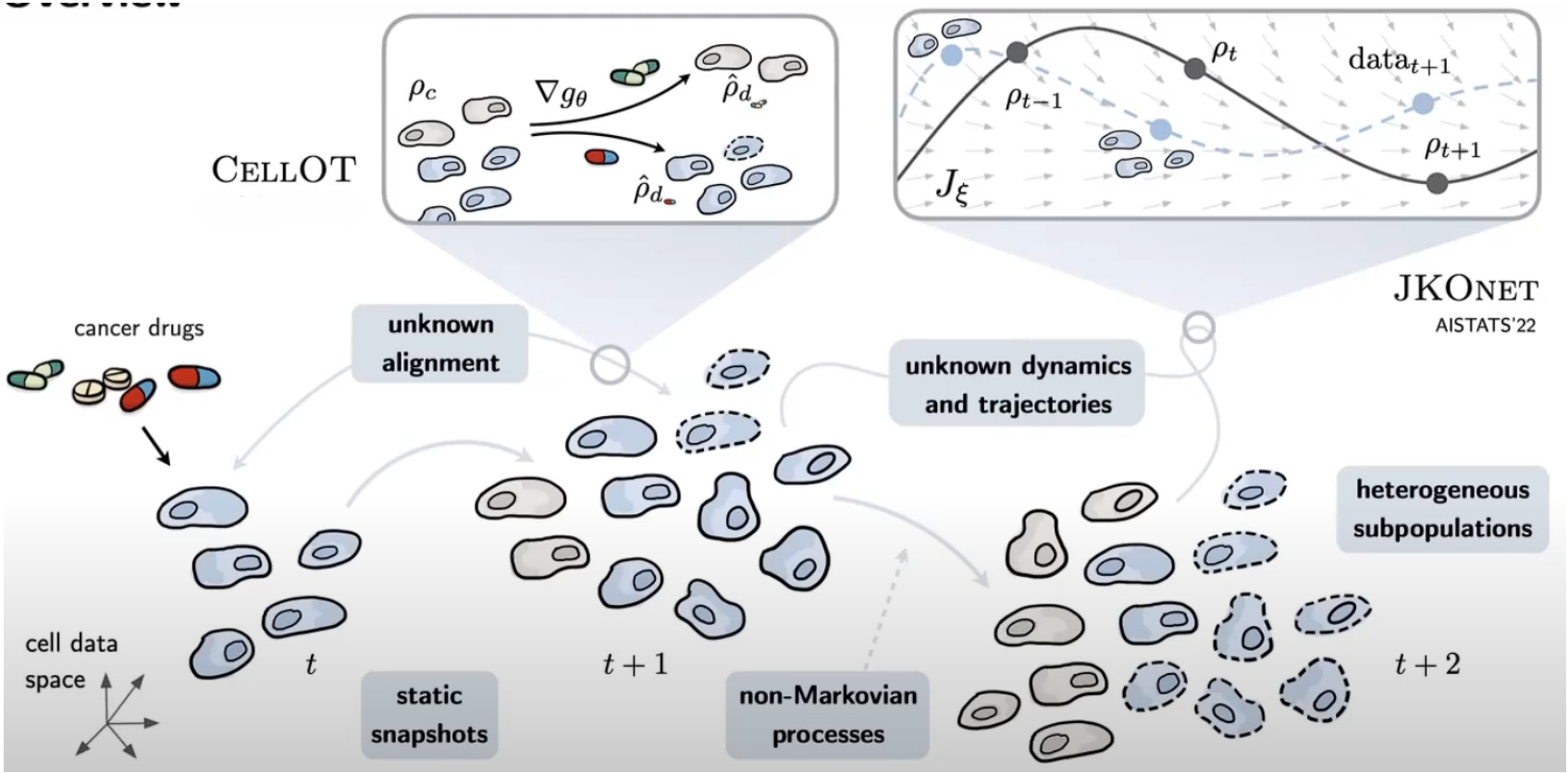
1. Helps reconstruct cell age relationships and developmental trajectories.
2. Enables dynamic modeling of gene expression changes.[10][11][12][13][14]

[10] Schiebinger, G. et al. Cell 176, 928–943. e22 (2019)
[11] Tong, A. et al. Proc. 37th ICML 9526–9536 (PMLR, 2020).
[12] Huguet, G. et al. Adv. Neur. Inf. Process. Syst. 35, 29705–29718 (2022)
[13] Bunne, Charlotte, et al. Nature Reviews Methods Primers 4.1 (2024): 58.
[14] Sha, Y., Qiu, Y., Zhou, P. et al. Nat Mach Intell 6, 25–39 (2024).

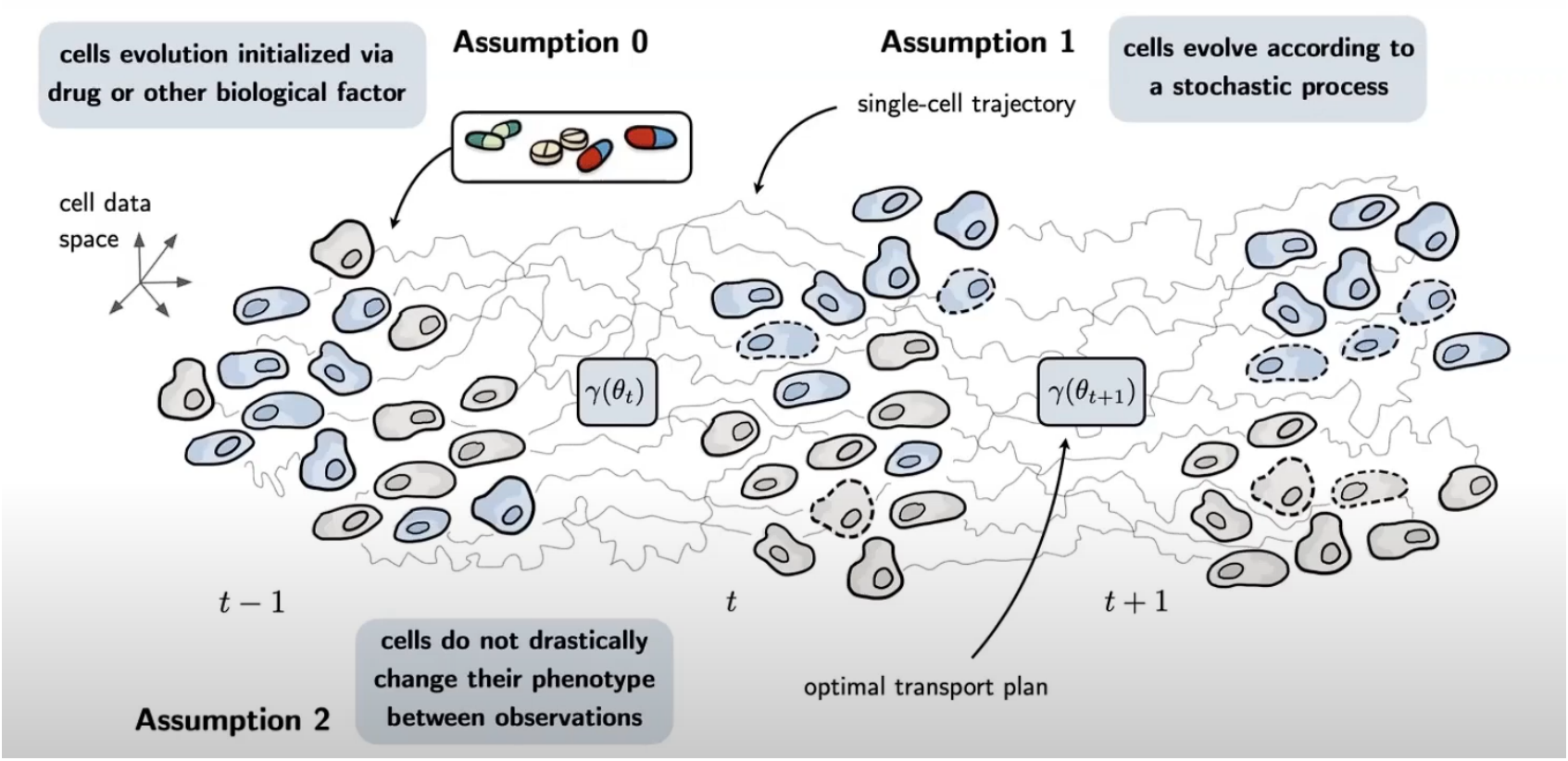
Workflow of stem cell research



Cell dynamics



Assumptions in stem cell evolution



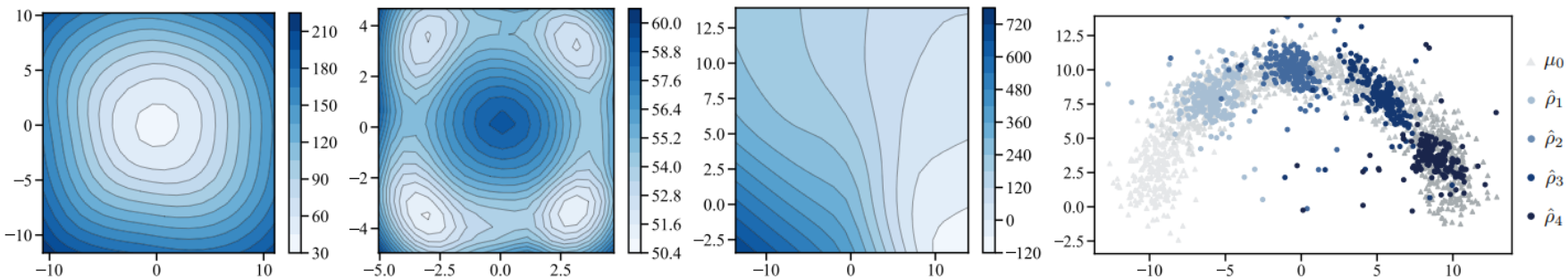
JKO flow in cell dynamics[15]

The biologists believe that the lineage and fate determination of cell follows the *Waddington landscape*: The evolution path tend to descend on some energy function.
Recall the energy function $J(\rho)$ in Eq. (20), with snapshot observation data, we can learn the energy with the matching between cells in the same lineage, so that the future evolution can be predicted.

[15] Bunne, Charlotte, et al. "Proximal optimal transport modeling of population dynamics." International Conference on Artificial Intelligence and Statistics. PMLR, 2022.

Energy-Driven and arbitrary Population Dynamics

Figure 3: Overview on different tasks including trajectory- and potential-based dynamics.



(a) Quadratic Potential. (b) Styblinski Potential. (c) Semicircle Trajectory *with* teacher forcing. (d) Predicted Population Evolution on Semicircle Trajectory.

Figure 4: **Results of JKOnet on Potential- and Trajectory-based Dyanamics.** (a)-(c) Contour plots of the energy functionals J_ξ of JKONET on potential- and trajectory-based population dynamics in different training settings (i.e., trained with or without teacher forcing § 3.2), color gradients depict the magnitude of J_ξ. (d) Predicted population snapshots $(\hat{\rho}_1, \ldots, \hat{\rho}_4)$ (blue) and data trajectory $(\mu_0, \ldots, \mu_4)$ (gray).

Variations of OT

1. Unbalanced OT
2. Gromov-Wasserstein distance
3. . . .

Unbalanced Optimal Transport (UOT)

- **Motivation:** Handle situations where the total mass of the source and target distributions may differ ($\mu(X) \neq \nu(Y)$).
- Introduces a **marginal relaxation** or a **penalty** for adding/removing mass.
- Several formulations exist. One common approach using entropic regularization and KL divergence for marginal penalties is:

$$\min_{\gamma \in \mathcal{M}_+(X \times Y)} \int_{X \times Y} c(x,y) d\gamma(x,y) + \lambda_1 \mathrm{KL}(\pi_X \# \gamma \| \mu) + \lambda_2 \mathrm{KL}(\pi_Y \# \gamma \| \nu)$$

where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters, and $\pi_X \# \gamma$ and $\pi_Y \# \gamma$ are the marginals of γ.

Advantages: More flexible for real-world data. Applications in domain adaptation, generative modeling, etc.

Gromov-Wasserstein (GW) Distance

- **Goal:** Compare probability distributions μ on (X, d_X) and ν on (Y, d_Y), where X and Y can be different metric spaces.
- Compares the **internal structures** (pairwise distances) of the two spaces. **Gromov-Wasserstein Problem:**

$$GW_p^p(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int_{X\times Y} \int_{X\times Y} |d_X(x,x') - d_Y(y,y')|^p d\gamma(x,y) \right.$$

- **Intuition:** Finds a coupling γ that aligns points such that their relative distances are as similar as possible.
- **Applications:** Shape matching, graph comparison, multi-modal learning.

Gromov-Wasserstein distance

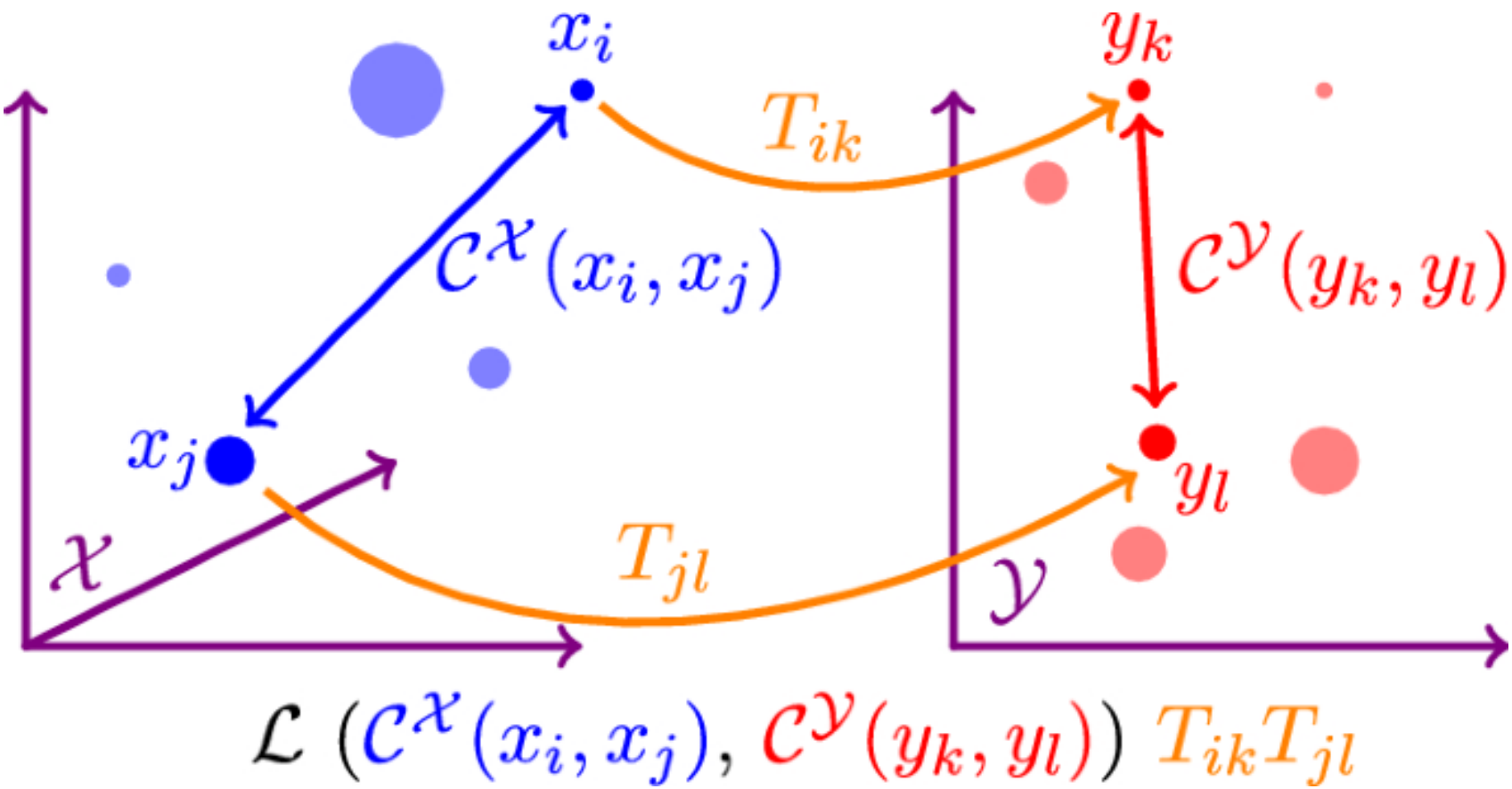


Figure: $C^{\mathcal{X}}$ is the cost within space $\mathcal{X}$, such as a distance.

Wasserstein Barycenter

Given probability measures $\{\mu_1, \ldots, \mu_N\}$ and weights $\{\lambda_1, \ldots, \lambda_N\}$, the Wasserstein barycenter μ^* minimizes:

$$\mu^* = \operatorname*{argmin}_{\mu} \sum_{i=1}^{N} \lambda_i W_p^p(\mu, \mu_i)$$

where W_p is the p-Wasserstein distance:

$$W_p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int \|x - y\|^p d\gamma(x, y) \right)^{1/p}$$

Key Idea

A "mean distribution" that balances mass transport costs.

Wasserstein Barycenter for Neuroimaging data average

Computing an average of brain imaging data is import:

1. Reduce Noise and Increase Signal Reliability
2. Capture Generalizable Patterns
3. Support Statistical Testing
4. Facilitate Visualization and Interpretation (so that it is more smooth visually)

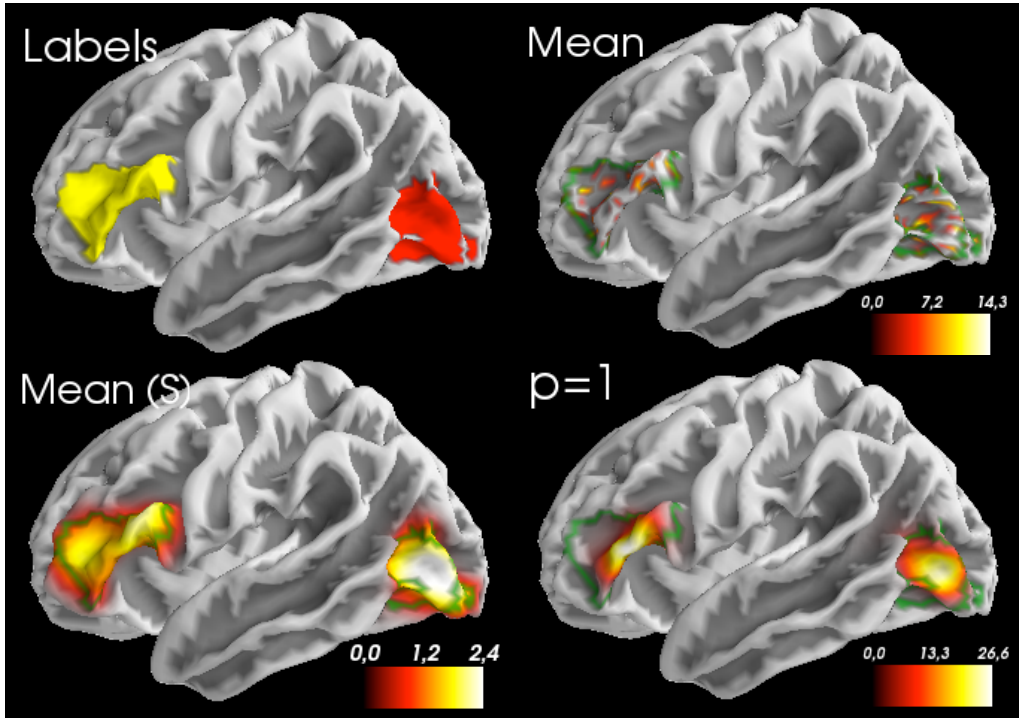


Figure: The standard averaging referred to as *Mean*, the averaging after Gaussian smoothing is referred to as *Mean (S)*, and the Kantorovich mean (p=1). The Mean shows signal spots in random places, marked in green. The *Kantorovich mean clearly shows strong signal spots in specific brain areas, without blurring them like Gaussian smoothing does.* In contrast, Gaussian smoothing not only blurs the signals but also makes them weaker.

Wasserstein barycenter for model ensembling

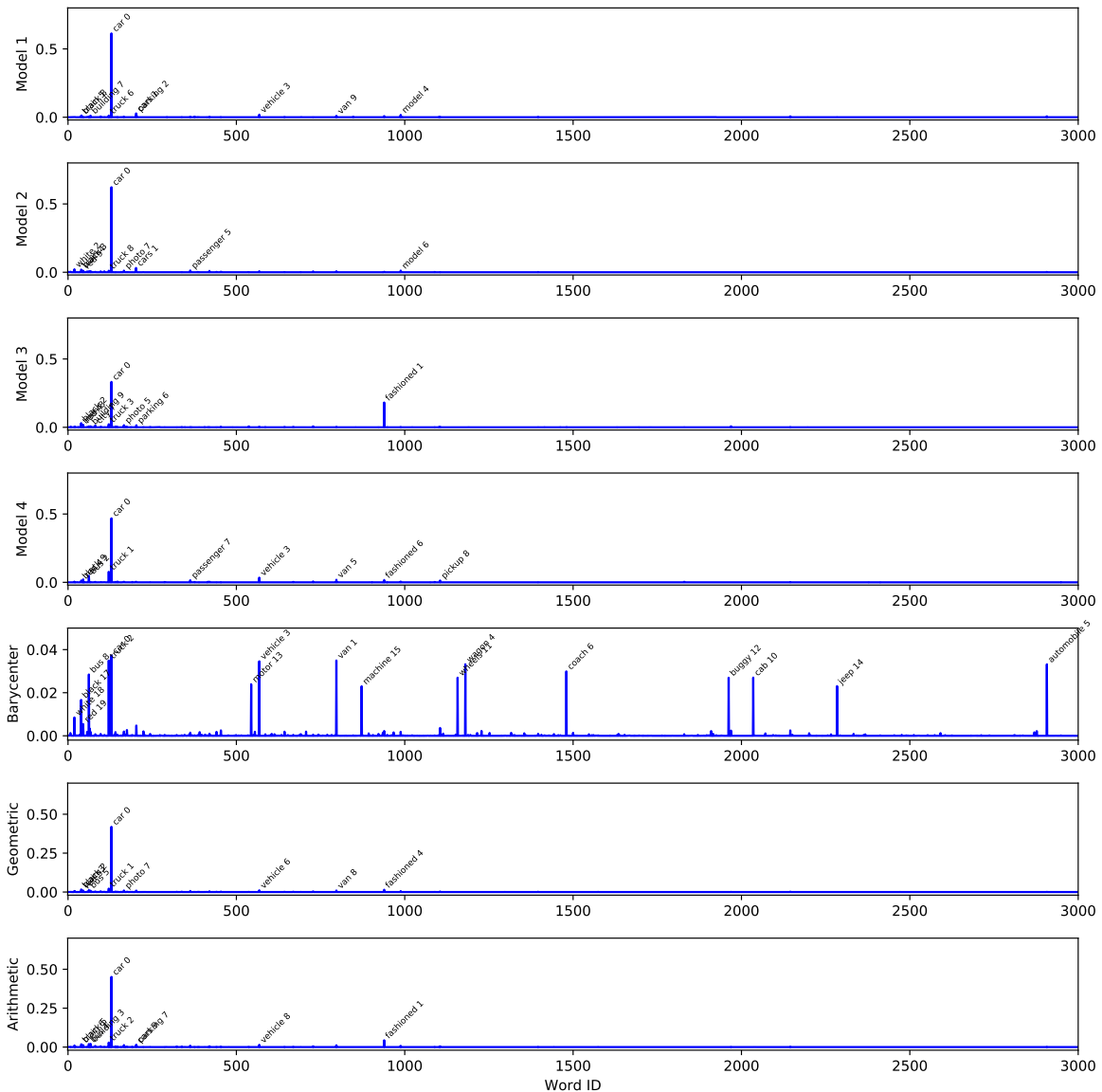
When multiple prediction models are available, model ensembling can be performed by computing the Wasserstein barycenter of their probability outputs, providing a common reference distribution for all models.

Accuracy	resnet18 alone	resnet34 alone	Arithmetic	Geometric	W. Barycenter
Validation	0.7771	0.8280	0.8129	0.8123	**0.8803**
Test	0.7714	0.8171	0.8071	0.8060	**0.8680**

Table: Attribute-based classification. The W. barycenter ensembling achieves better accuracy by exploiting the cross-domain similarity matrix K, compared to a simple linear-transform of probability mass from one domain to another as for the original models or their simple averages.

Wasserstein barycenter for ensembling caption models

Figure: Visualization of the word distributions of W. barycenter, arithmetic and geometric means based on four captioning models. Wasserstein ensembling has better diversity.



Domain Adaption

The goal of domain adaption is to train a model on a source domain and make it perform well on a different but related target domain, where labeled data may be limited or unavailable.

Domain Adaption with OT

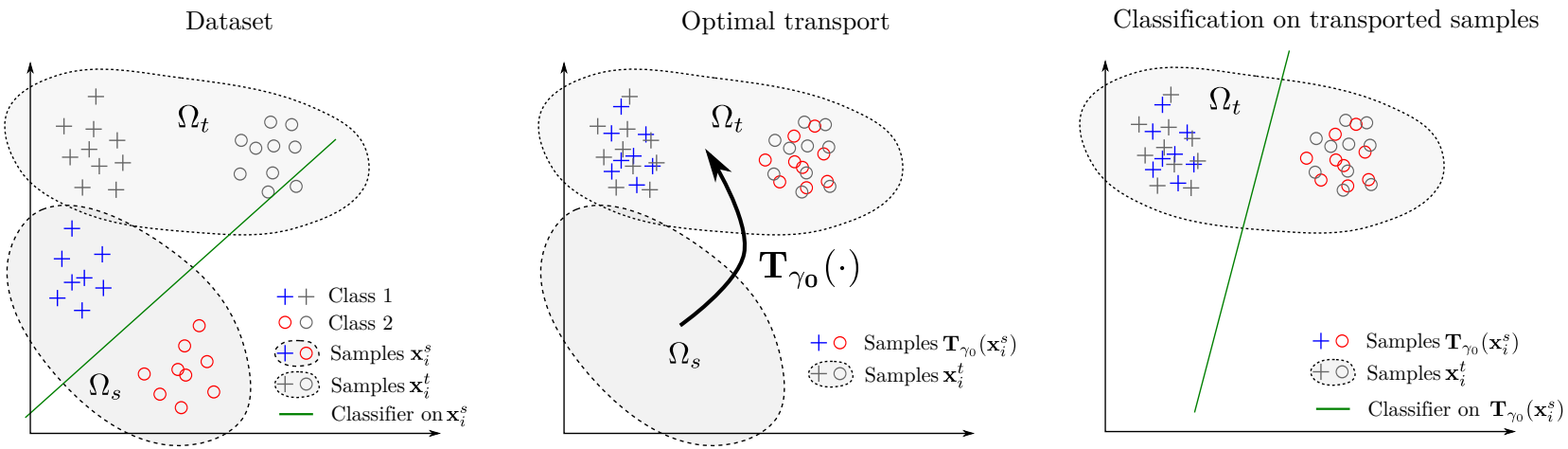


Figure: Illustration of OT for domain adaptation. (left) dataset for training, *i.e.* source domain, and testing, *i.e.* target domain. Note that a classifier estimated on the training examples clearly does not fit the target data. (middle) a data dependent transportation map $\mathbb{T}_{\gamma 0}$ is estimated and used to transport the training samples onto the target domain. Note that this transformation is usually not linear. (right) the transported labeled samples are used for estimating a classifier in the target domain.

Continued

To solve the adaptation problem:

1. Estimate μ_s and μ_t from indicators of samplings in source domain X_s and target domain X_t
2. Find a transport map $\mathbf{T}$ from μ_s to μ_t
3. Use $\mathbf{T}$ to transport labeled samples X_s and train a classifier from them.

The problem of finding such a transportation of minimal cost has already been investigated in the literature. For instance, the optimal transportation problem as defined by Monge is the solution of the following minimization problem:

$$\mathbb{T}_0 = \underset{\mathbf{T}}{\operatorname{argmin}} \int_{\Omega_s} c(\mathbf{x}, \mathbf{T}(\mathbf{x}))d\mu(\mathbf{x}), \quad \text{s.t.} \quad \mathbf{T}\#s = \mu_t \tag{22}$$

Color Transfer

- Goal: Transfer the color palette of one image (source) to another (target).
- Applications:
  - Artistic style editing
  - Image harmonization
  - Film color grading
- Challenge: Achieve natural-looking results while preserving image content.



Why Use Optimal Transport?

- Color distributions in source and target images can differ significantly.
- OT finds a mapping that minimally "moves" one distribution into another.
- It ensures:
  - Global color structure is preserved
  - Mapping is theoretically grounded
- Especially effective when histogram matching fails.

OT for Color Transfer: Pipeline

1. Sample color pixels from source and target images.
2. Estimate color histograms or empirical distributions.
3. Solve the OT problem (usually Kantorovich).
4. Apply the computed transport plan to recolor source pixels.

Wasserstein Distributionally Robust Optimization

- Standard machine learning minimizes risk over an empirical distribution:

$$\min_{\theta} \mathbb{E}_{\hat{P}}[\ell(x, \theta)]$$

- But what if $\hat{P}$ is not a good estimate of the true distribution P?
- DRO protects against distributional shifts by optimizing over a set of plausible distributions:

$$\min_{\theta} \sup_{Q \in \mathcal{U}(\hat{P})} \mathbb{E}_{Q}[\ell(x, \theta)]$$

- $\mathcal{U}(\hat{P})$: uncertainty set around empirical distribution

Why Use Wasserstein Distance?

- Wasserstein distance provides a meaningful way to measure differences between distributions:

$$W_c(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \int c(x, y)\, d\gamma(x, y)$$

- Captures geometry of the data space (unlike KL or TV divergence).
- Can define the uncertainty set as a *Wasserstein ball*:

$$\mathcal{U}(\hat{P}) = \{Q : W_c(Q, \hat{P}) \leq \rho\}$$

- Leads to robustness against perturbations in the input distribution.

Wasserstein DRO Objective

Given loss $\ell(x, \theta)$, the W-DRO objective becomes:

$$\min_{\theta} \sup_{Q: W(Q, \hat{P}) \leq \rho} \mathbb{E}_Q[\ell(x, \theta)]$$

Can be reformulated using strong duality (under convexity assumptions):

$$\min_{\theta} \mathbb{E}_{\hat{P}}[\ell(x, \theta)] + \rho \cdot \Omega(\theta)$$

Where $\Omega(\theta)$ acts as a robustness regularizer.

Interpretation and Benefits

- W-DRO learns models that are robust to small changes in data distribution.
- Especially useful in:
  - Domain shift
  - Noisy data
  - Fairness-aware learning
- Provides a theoretical robustness guarantee.
- Leads to better generalization in out-of-distribution (OOD) settings.

Why Use Wasserstein Distance?

- Captures geometric structure between distributions.
- Suitable for complex, non-Gaussian temporal dynamics.
- Goes beyond pointwise matching — compares entire sample distributions.

OT FOR TIME-SERIES IMPUTATION

- Learn an imputation model (e.g., Transformer) that fills missing values.
- View completed sequences as empirical distributions.
- Use Wasserstein discrepancy to align predicted and real distributions.

Computing Wasserstein Discrepancy

Definition (Proximal Spectral Wasserstein Discrepancy)

The PSW discrepancy seeks a transport plan $\mathbf{T} \in \mathbb{R}_+^{n \times m}$ *that transports the distribution* α *to* β *at minimal cost, defined as:*

$$\mathcal{P}^{\kappa}(\alpha, \beta) := \min_{\mathbf{T} \geq 0} \left\langle \mathbf{D}^{(\mathrm{F})}, \mathbf{T} \right\rangle + \kappa \left(D_{\mathrm{KL}}(\mathbf{T}\mathbf{1}_m \| \Delta_n) + D_{\mathrm{KL}}(\mathbf{T}^{\top}\mathbf{1}_n \| \Delta_m) \right)$$

where $\mathbf{D}^{(\mathrm{F})}$ *is the pairwise distance matrix computed using Pairwise Spectral Distance;* κ *is the matching strength;* $\Delta_n = \mathbf{1}_n/n$ *and* $\Delta_m = \mathbf{1}_m/m$ *are uniform simplex vectors;* $\mathcal{P}^{\kappa}$ *denotes the PSW discrepancy.*

Objective Function

- Total loss combines reconstruction and distributional alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda W(P, Q)$$

- $\mathcal{L}_{\text{recon}}$: e.g., MSE on observed entries.
- $W(P, Q)$: Wasserstein discrepancy.
- λ: Weighting factor.

Results

- Wasserstein discrepancy enforces distributional consistency.
- It regularizes imputation models toward realistic global structure.
- Enhances quality of imputed values beyond just fitting observed data.