Cost of Moving Mountains: Intro to Optimal Transport

Tony Huang, Layne Malek, Peiyao Sun, Max Wang
Mentor: Fang Rui Lim, Jiamin Jian
Lab of Geometry, University of Michigan

LOG(M)

Project description

Many modern problems compare *distributions* rather than individual points: e.g. two clouds of data, two images treated as mass densities, etc. **But how can we measure the distance between probability measures in a way that respects the geometry of the underlying space?**

In this project, we use **Optimal transport (OT)** to answer the question by viewing one distribution as a pile of "mass" that must be moved to match another distribution with minimal total cost.

Toy example: Bread and Bakery problem Transport bread from bakery x_i to house y_j. A map $T : \mathbb{R} \to \mathbb{R}$ tells us $y = T(x)$, subject to the mass-balance constraint: for any house y_j requiring b_j units,

$$b_j = \sum_{\substack{i=1,\dots,n \\ T(x_i)=y_j}} a_i. \tag{1}$$

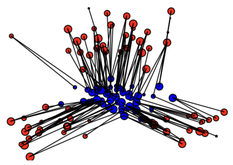
If this constraint is fulfilled, the total cost is $\sum_{i=1}^{n} c\big(x_i, T(x_i)\big)\, a_i$ where c is the cost function.

From Kantorovich to Wasserstein Distance

Historically, OT began with Monge's map-based model, and Kantorovich's relaxation to couplings made the problem well-posed and convex. The Kantorovich formulation seeks

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, \mathrm{d}\pi(x,y) \tag{2}$$

where $\Pi(\mu,\nu)$ is the set of all couplings with marginals μ and ν.



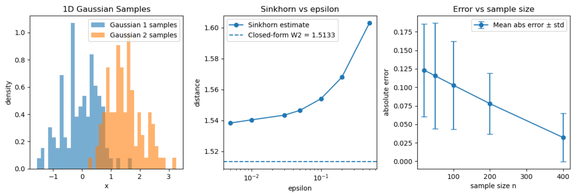
Visualization of Kantorovich OT plan [3]

Sinkhorn Algorithm

To solve the Optimal Transport problem efficiently, we add an **entropic regularization** term $\epsilon H(P)$ to the objective function:

$$\min_{P \in U(a,b)} \langle C, P\rangle - \varepsilon H(P), \qquad H(P) = -\sum_{i,j} P_{ij}\big(\log P_{ij} - 1\big)$$

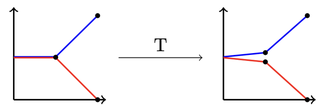
The Core Idea: Instead of heavy linear programming, we use **Sinkhorn Iteration**: $u \leftarrow a \oslash (Kv), v \leftarrow b \oslash (K^\top u)$. By repeatedly scaling the rows and columns of a kernel matrix $K = \exp(-C/\epsilon)$, we quickly converge to the optimal solution.



Sinkhorn vs Closed-form (1D Gaussian)

Causal Optimal Transport

When we apply the logic of classical OT to stochastic processes or time series, a significant problem appears: the flow of information. Take a look at these two models, **left** and **right**. [2]



Standard Wasserstein distance only measures spatial similarity, falsely treating these two as "nearly identical." Yet, they differ fundamentally in the timing of information reveal (filtration). **Left model:** Late information reveal. **Right model:** Early information reveal (allowing you to "anticipate" the future). And this is why we need Causal Optimal Transport (COT).

Causal Transport Maps

The causal transport map can be represented in $\mathbb{R}^3$ as:

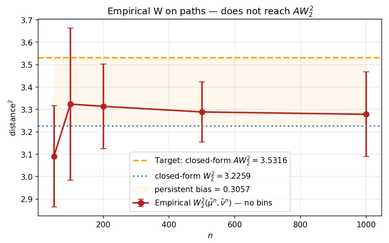
$$(x_0, x_1, x_2) \to T(x_0, x_1, x_2) = \begin{pmatrix} T_0(x_0) \\ T_1(x_0, x_1) \\ T_2(x_0, x_1, x_2) \end{pmatrix} \in \mathbb{R}^3 \tag{3}$$

These causal maps respect the temporal flow of information since each T_i is only allowed to depend on $x_0, \dots, x_i$.
And in terms of the **Adapted Wasserstein Distance**, we can switch couplings $\Pi(\mu,\nu)$ into causal couplings $\Pi_c(\mu,\nu)$.

Unbinned Empirical Failure

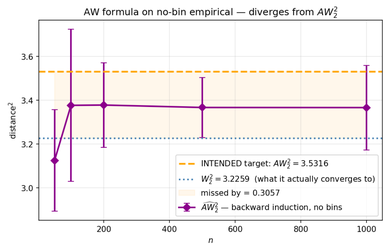
Ordinary empirical path measures do not preserve the temporal information structure required by adapted transport. Without binning or an adapted finite-state approximation, the estimator misses the adapted Wasserstein target AW_2^2 and instead approaches the classical Wasserstein distance W_2^2.

Empirical W_2^2 on Raw Paths



We compute the classical Wasserstein distance directly from unbinned empirical path measures. As the sample size n increases, the estimate moves toward the closed-form W_2^2 benchmark rather than the adapted target AW_2^2, showing that raw empirical paths ignore the filtration structure needed for adapted transport.

Backward Induction Without Binning



Applying the adapted Wasserstein backward induction directly to raw empirical paths does not recover AW_2^2. Without binning or a finite-state scenario tree, the recursion is built on the wrong empirical object, so the estimate stabilizes away from the adapted benchmark and retains a non-vanishing bias.

Gaussian Markov Process

In this experiment, we model μ and ν as the path laws of two one-dimensional Ornstein–Uhlenbeck (OU) processes, simulated in discrete time by the Euler scheme

$$X_{t+1} = (1 - \theta\Delta t)X_t + \theta\Delta t\, m + \sigma\sqrt{\Delta t}\,\xi_{t+1}, \qquad \xi_{t+1} \sim \mathcal{N}(0,1).$$

For μ, we use $m = 0,\ \sigma = 0.3,\ \theta = 1.0,\ X_0 = 0$, and for ν, we use $m = 0.6,\ \sigma = 0.4,\ \theta = 0.85,\ X_0 = 0.5$. These choices induce two Gaussian Markov path measures μ and ν.

Our benchmark studies convergence of an empirical adapted Wasserstein-2 approximation to the exact Gaussian target:

$$AW_2(\hat{\mu}^N, \hat{\nu}^N) \longrightarrow AW_2(\mu,\nu) \quad \text{as } N \to \infty.$$

Pipeline.

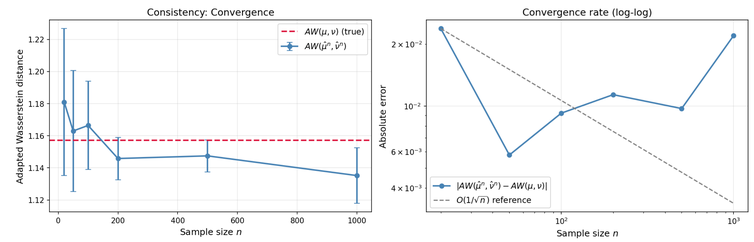
1. Sample N paths from each Euler–OU law.
2. Project whole trajectories onto a shared path-space partition with uniform binning $\hat{\varphi}^N$ [1], yielding finite-support empirical process measures
$$\hat{\mu}^N = \frac{1}{N}\sum_{n=1}^{N} \delta_{\hat{\varphi}^N(X^{(n)})}, \qquad \hat{\nu}^N = \frac{1}{N}\sum_{n=1}^{N} \delta_{\hat{\varphi}^N(Y^{(n)})}.$$
3. Build the induced scenario trees and compute the discrete adapted Wasserstein-2 distance by backward induction on path prefixes.
4. Compare the empirical value with the exact Gaussian reference and plot
$$\big|AW_2(\hat{\mu}^N, \hat{\nu}^N) - AW_2(\mu,\nu)\big|.$$

Closed form for true $AW_2^2(\mu,\nu)$. Given that $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$, $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$, and $X = m_\mu + L_\mu Z$, $Y = m_\nu + L_\nu Z$, where $Z \sim \mathcal{N}(0, I)$, m_μ, m_ν are the mean path vector, and Σ_μ, Σ_ν are the covariance factors, then

$$AW_2^2(\mu,\nu) = \|m_\mu - m_\nu\|^2 + \|L_\mu - L_\nu\|_F^2.$$

Interpretation. Any remaining gap between the empirical curve and the closed-form target comes from two sources: finite-sample error and path-space discretization bias. Thus, convergence requires both increasing sample size and a sufficiently fine partition.

Consistency - Gaussian Markov (OU) Processes, AW_2



References

[1] Beatrice Acciaio and Songyan Hou. Convergence of adapted empirical measures on R^d. *The Annals of Applied Probability*, 34(5):4799–4835, 2024.

[2] Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Manu Eder. Adapted wasserstein distances and stability in mathematical finance. *arXiv preprint arXiv:1901.07450*, 2020.

[3] Gabriel Peyré. Optimal transport with linear programming. `https://nbviewer.org/github/gpeyre/numerical-tours/blob/master/python/optimaltransp_1_linprog.ipynb`, 2024. Numerical Tours of Data Science.