Minimax estimation of optimal transport maps

Vincent Divol

Journée SMAI-SIGMA

CEREMADE
Université Paris Dauphine - PSL

vincent.divol@psl.eu

vincentdivol.github.io

Jon Niles-Weed

Aram Pooladian

Regression: $Y_i = T_0(X_i) + \varepsilon_i$

Observe: $(X_1, Y_1), \ldots, (X_n, Y_n)$ **paired data**

Goal: estimate T_0

Non-parametric least squares:

$\hat{T} \in \arg\min_{T \in \mathcal{T}} \sum_{i=1}^{n} \|T(X_i) - Y_i\|^2$

$\mathbb{E}\|\hat{T} - T_0\|^2 \leq \inf_{T \in \mathcal{T}} \|T - T_0\|^2 + \delta_{n,\mathcal{T}}$

model error estimation error

Regression: $Y_i = T_0(X_i) + \varepsilon_i$

Observe: $(X_1, Y_1), \ldots, (X_n, Y_n)$ **paired data**

Goal: estimate T_0

Non-parametric least squares:

$$\hat{T} \in \arg\min_{T \in \mathcal{T}} \sum_{i=1}^{n} \|T(X_i) - Y_i\|^2$$

$$\mathbb{E}\|\hat{T} - T_0\|^2 \leq \inf_{T \in \mathcal{T}} \|T - T_0\|^2 + \delta_{n,\mathcal{T}}$$

model error estimation error

$\to$ What if we only have access to $\{X_i\}$ and $\{Y_j\}$? (uncoupled regression)

$\to$ What if we only have access to $\{X_i\} \sim \mu$ and $\{Y_j = T_0(X'_j)\} \sim (T_0)_\sharp\mu$, with $X_i \perp X'_j$?

Application: computational biology

population of stem cells evolve through time
$\to$ observing a cell destroys it

[Schiebinger & al. 19, Moriel & al. 21, Demetci & al. 21]

$X_1, \ldots, X_n \sim \mu_t$

time t

$T_0(X'_1), \ldots, T_0(X'_n)$
$= Y_1, \ldots, Y_n \sim \mu_{t+1}$

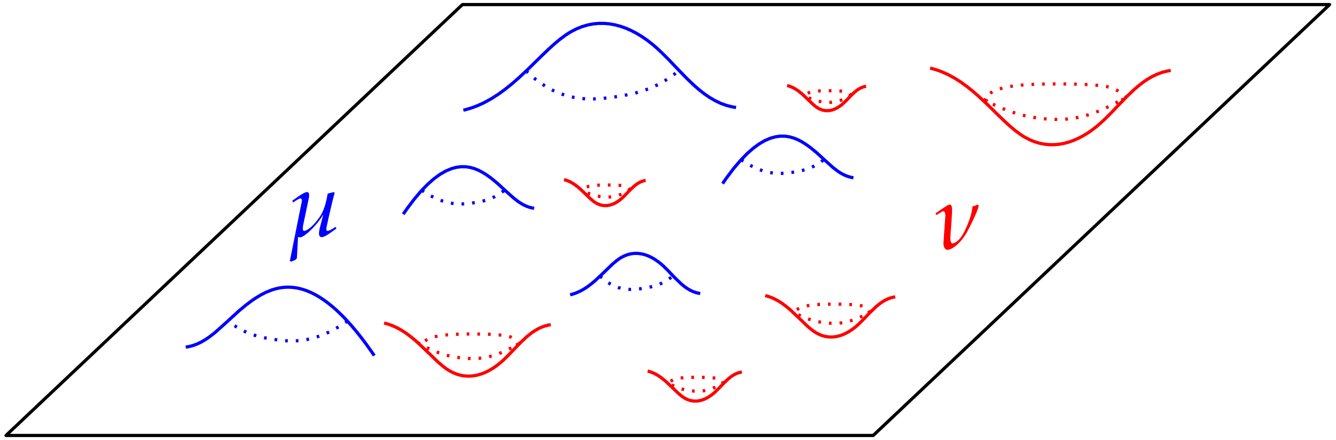
time $t+1$

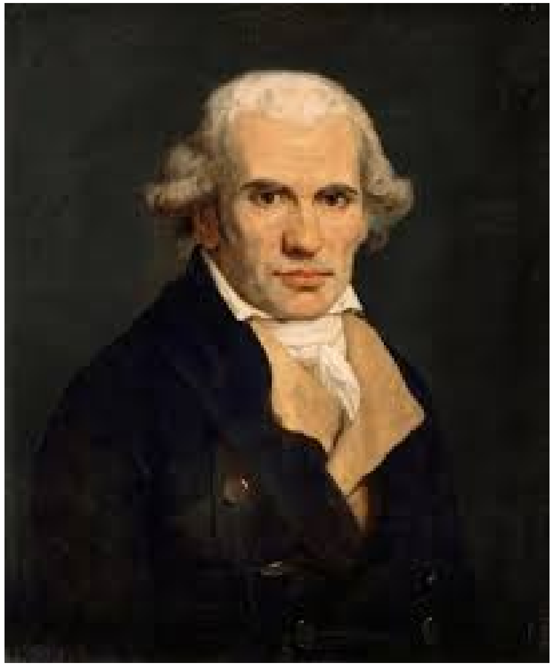
What is the transformation T_0?

Problem: Lack of identifiability
$\to$ Given μ, ν, there are **many** Ts with $T_\sharp \mu = \nu$

minimize $\int \|x - T(x)\|^2 \mathrm{d}\mu(x)$ (Monge)

under the constraint $T_\sharp \mu = \nu$





$\rightarrow$ Existence?

minimize $\int \|x - y\|^2 \mathrm{d}\pi(x, y)$ (Kantorovitch)

under the constraint $\pi \in \Pi(\mu, \nu)$

$\pi(A \times \mathbb{R}^d) = \mu(A) \quad \pi(\mathbb{R}^d \times B) = \nu(B)$

$\rightarrow$ Linear problem under linear constraints!

minimize $\int \|x - y\|^2 \mathrm{d}\pi(x, y)$ (Kantorovitch)

under the constraint $\pi \in \Pi(\mu, \nu)$

$\pi(A \times \mathbb{R}^d) = \mu(A) \quad \pi(\mathbb{R}^d \times B) = \nu(B)$

→ Linear problem under linear constraints!

Discrete setting: $\mu = \sum_{i=1}^{n} \mu_i \delta_{x_i}$

$\nu = \sum_{j=1}^{m} \nu_j \delta_{y_j}$

$C_{ij} = \|x_i - y_j\|^2$

minimize $\langle C, \pi \rangle$

under the constraints $\quad \forall i, \sum_j \pi_{ij} = \mu_i \quad \forall j, \sum_i \pi_{ij} = \nu_j$

$\rightarrow$ nm variables, $n + m$ constraints minimize $S(\phi) = \mu(\phi) + \nu(\phi^*)$ (Dual problem)

where $\phi^*(y) = \sup_x \langle x, y \rangle - \phi(x)$

(Lagrange)

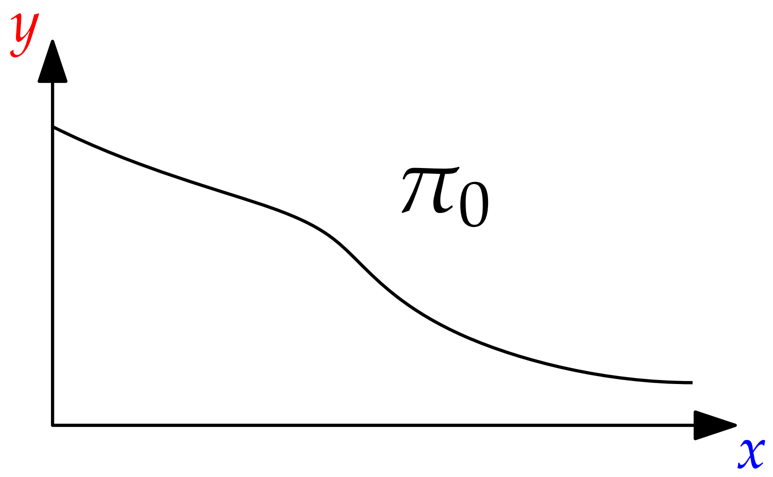
$\rightarrow$ complexity $O(nm(n+m)) = O(n^3)$ if $n = m$

minimize $S(\phi) = \mu(\phi) + \nu(\phi^*)$ (Dual problem)

where $\phi^*(y) = \sup_x \langle x, y \rangle - \phi(x)$

$\rightarrow$ complexity $O(nm(n+m)) = O(n^3)$ if $n = m$

(Brenier)



(Lagrange)

Theorem: if μ has a density on $\mathbb{R}^d$, then (Monge) has a unique solution T_0, equal to $\nabla \phi_0$ where ϕ_0 is the (convex) Brenier potential.

So far...

- $X_1, \ldots, X_n \sim \mu \qquad\qquad Y_1, \ldots, Y_n \sim \nu = (T_0)_\sharp \mu$

- $T_0 = \nabla \phi_0$ where $\phi_0 = \arg\min_\phi S(\phi) = \int \phi d\mu + \int \phi^* d\nu$

- $\hat{T} = \nabla \hat{\phi}$ where

$$\hat{\phi} = \arg\min_{\phi \in \mathcal{F}} S_n(\phi) = \frac{1}{n} \sum_{i=1}^n \phi(X_i) + \frac{1}{n} \sum_{i=1}^n \phi^*(Y_i)$$

$\mathcal{F}$ = family of candidate potentials

- The "size" of $\mathcal{F}$ is measured by its metric entropy:

$N(h) =$ smallest number of L_∞ balls of radius h needed to cover $\mathcal{F}$

Theorem: if μ satisfies a Poincaré inequality, if ϕ_0 and all potentials in $\mathcal{F}$ are (uniformly) smooth, and if

$$\log N(h) \lesssim_{\log(1/h)} h^{-\gamma} \qquad \gamma \geq 0$$

Then

$$\mathbb{E}[\|\nabla\hat{\phi} - \nabla\phi_0\|^2_{L_2(\mu)}] \lesssim \inf_{\phi\in\mathcal{F}}(S(\phi) - S(\phi_0)) + n^{-\left(\frac{2}{2+\gamma}\wedge\frac{1}{\gamma}\right)}$$

→ Generalizes previous theoretical and applied works [Hütter Rigollet 21], [Makkuva & al. 20], [Bunne & al. 22], [Vacher Vialard 21]

→ New (near) minimax results in many different situations: approximation with NNs, Barron spaces, RKHS, spiked model, etc.

An example: the spiked transport model

[Niles-Weed Rigollet 21]

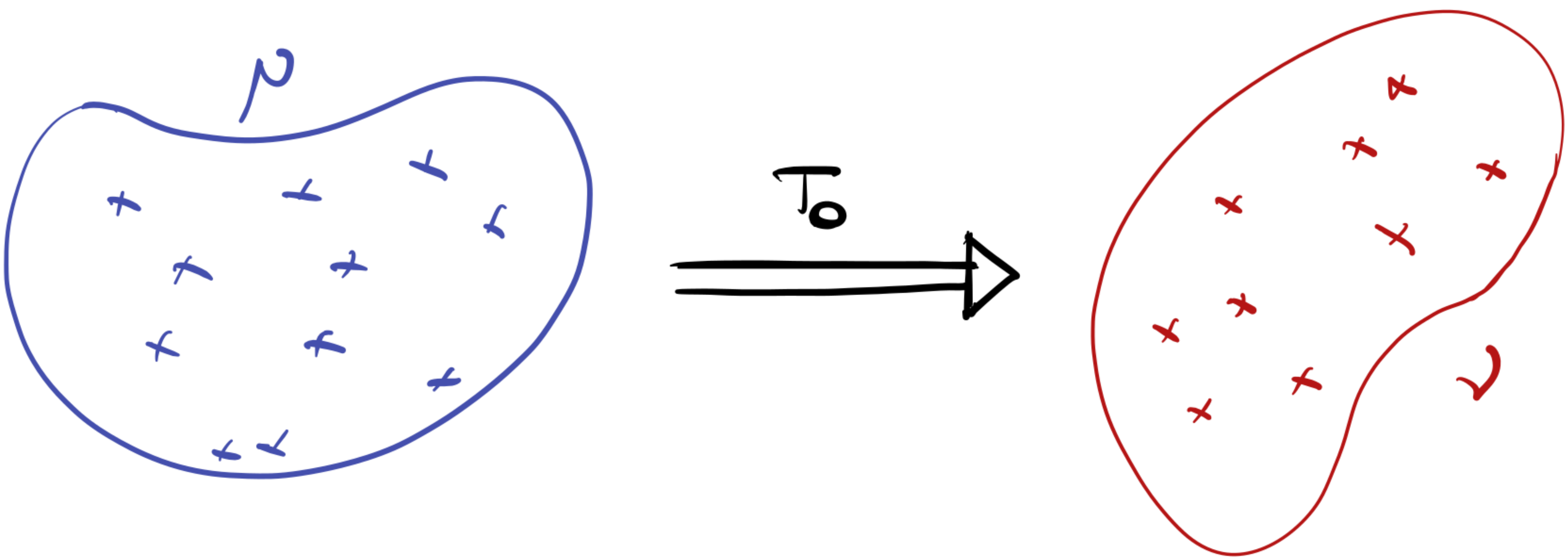
$U \subset \mathbb{R}^d$ unknown k-dimensional subspace, $k \ll d$ (the spike)

$$T_0(x) = T_0'(\pi_U(x)) + \pi_U^\top(x)$$

where $T_0' : U \to U$ is of regularity $\alpha \geq 1$

$\to$ design $\mathcal{F}$ adapted to the model with

$$\mathbb{E}[\|\nabla\hat{\phi}_{\mathcal{F}} - T_0\|^2_{L_2(\mu)}] \lesssim n^{-\frac{2(\alpha+1)}{2\alpha+k+2}}$$



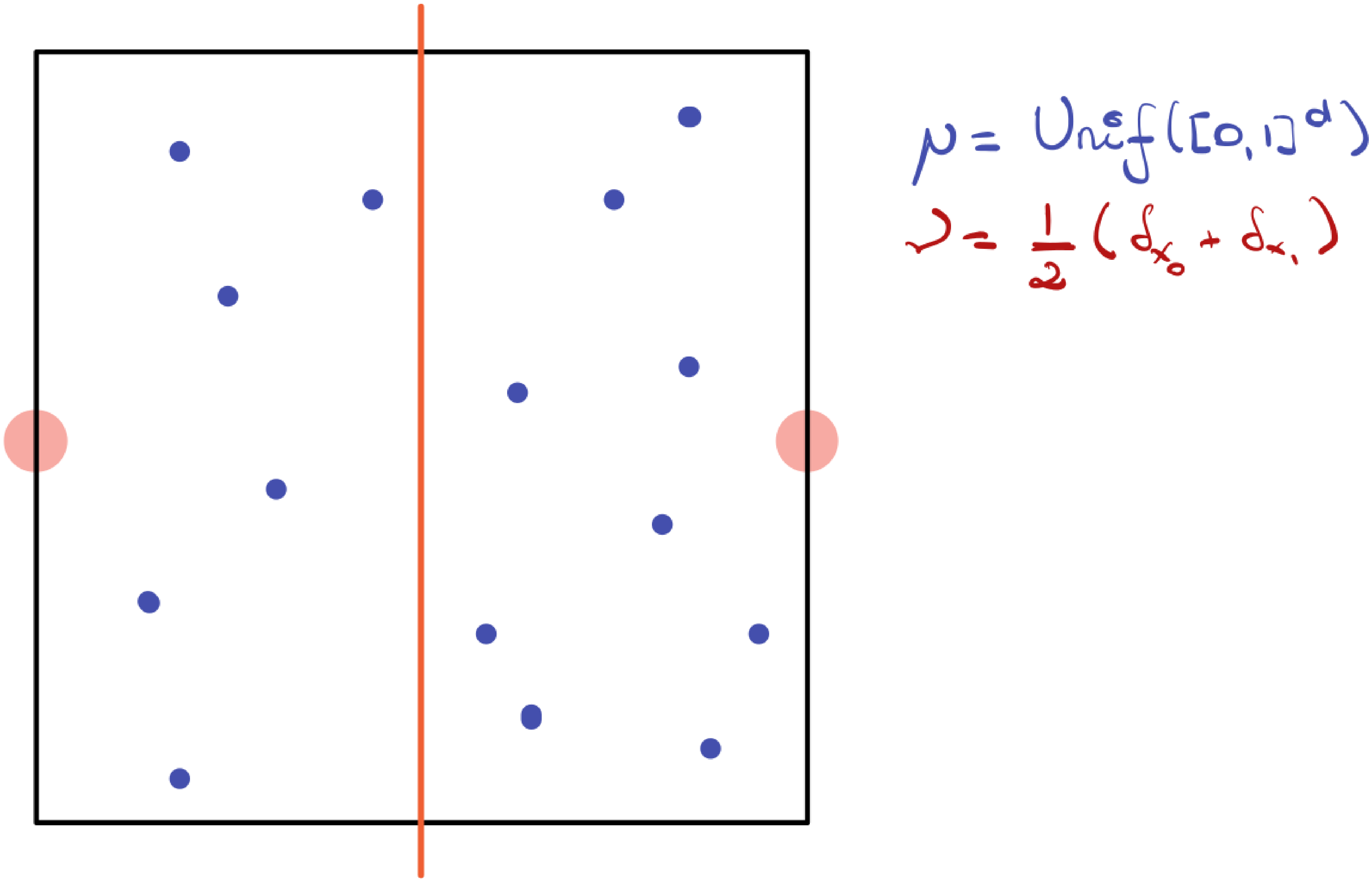
$X_1, \ldots, X_n \sim \mu$ $\qquad$ $Y_1, \ldots, Y_n \sim \nu$

Theorem: If T_0 is bi-Lipschitz, μ is almost uniform on a nice domain in $\mathbb{R}^d$. Then,

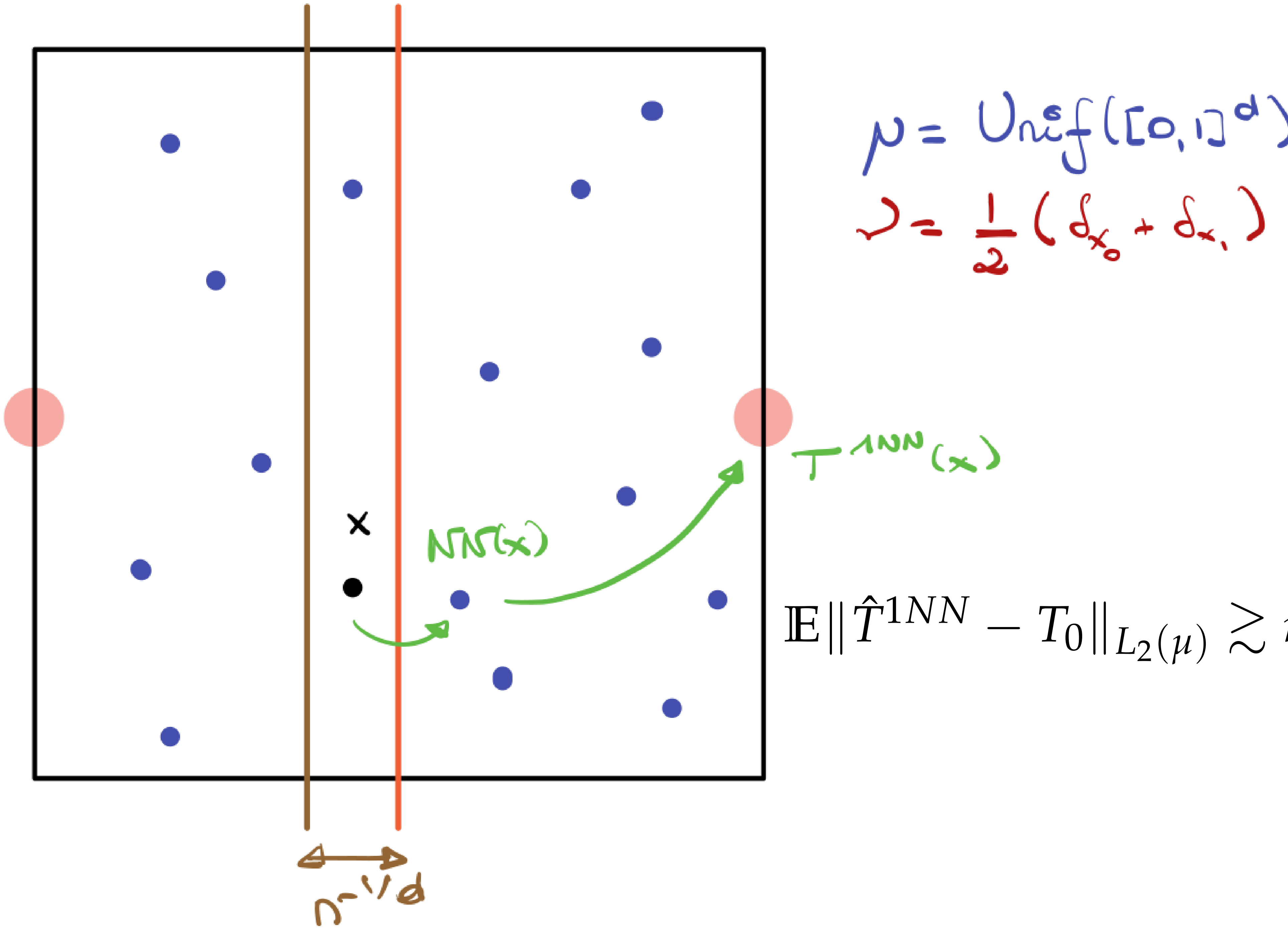
$$\mathbb{E}\|\hat{T}^{1NN} - T_0\|_{L_2(\mu)} \lesssim n^{-\frac{1}{d}}$$

[Manole & al. 21]

$\Rightarrow$ What if T_0 is not even continuous?

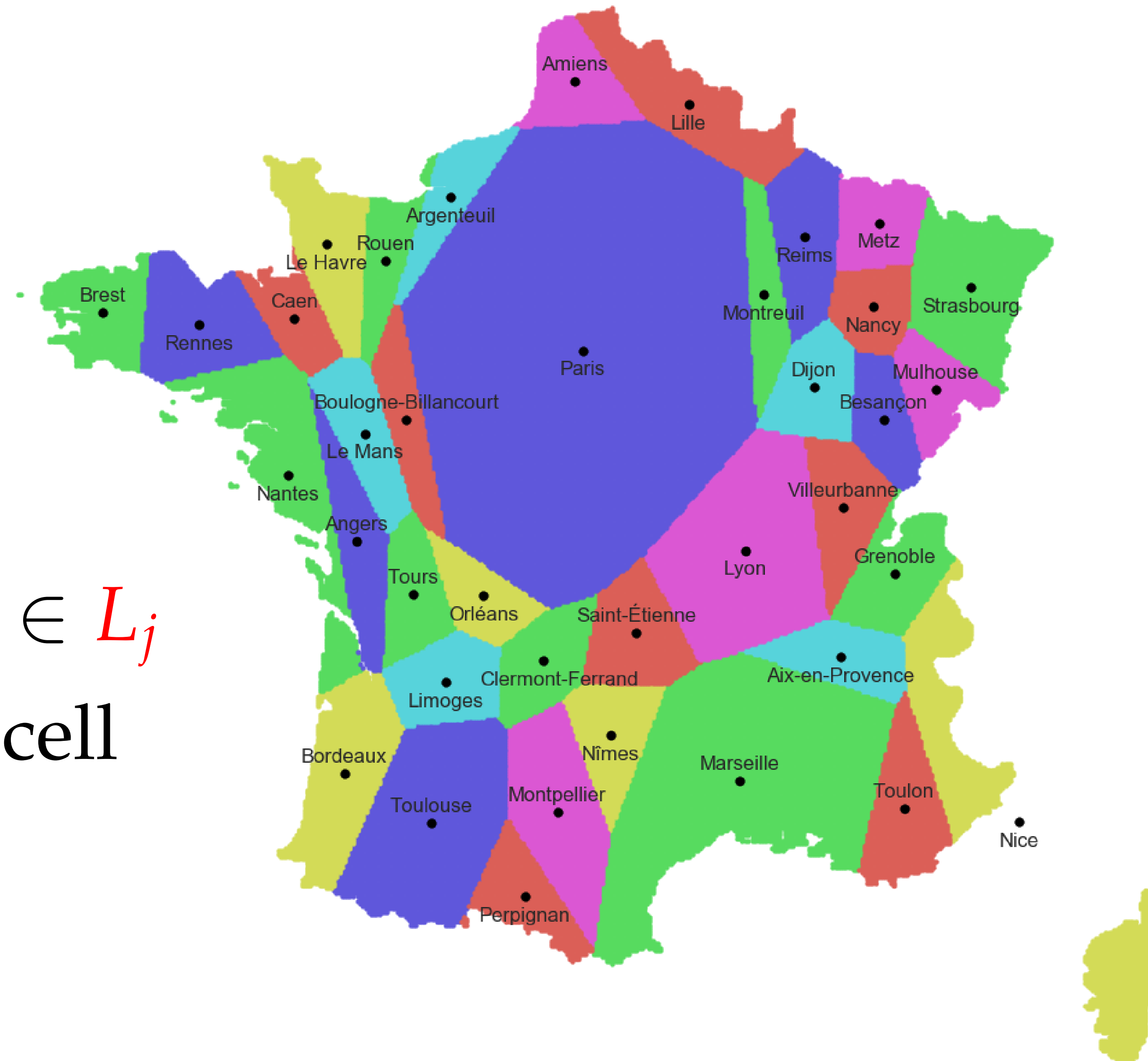


⇒ What if T_0 is not even continuous?



$$\mathbb{E}\|\hat{T}^{1NN} - T_0\|_{L_2(\mu)} \gtrsim n^{-\frac{1}{2d}}$$

The semi-discrete case: μ has a density, $\nu = \sum_{j=1}^{J} q_j \delta_{y_j}$



$$T_0(x) = y_j \text{ if } x \in L_j$$

L_j = Laguerre cell

$$\text{minimize} \int \|x - y\|^2 \mathrm{d}\pi(x, y) + \varepsilon \, \mathrm{KL}(\pi \| \mu \otimes \nu)$$

under the constraint $\pi \in \Pi(\mu, \nu)$ (Schrödinger)

Imagine that you observe a system of diffusing particles which is in thermal equilibrium. Suppose that at a given time t_0 you see that their repartition is almost uniform and that at $t_1 > t_0$ you find a spontaneous and significant deviation from this uniformity. You are asked to explain how this deviation occurred. What is its most likely behaviour?

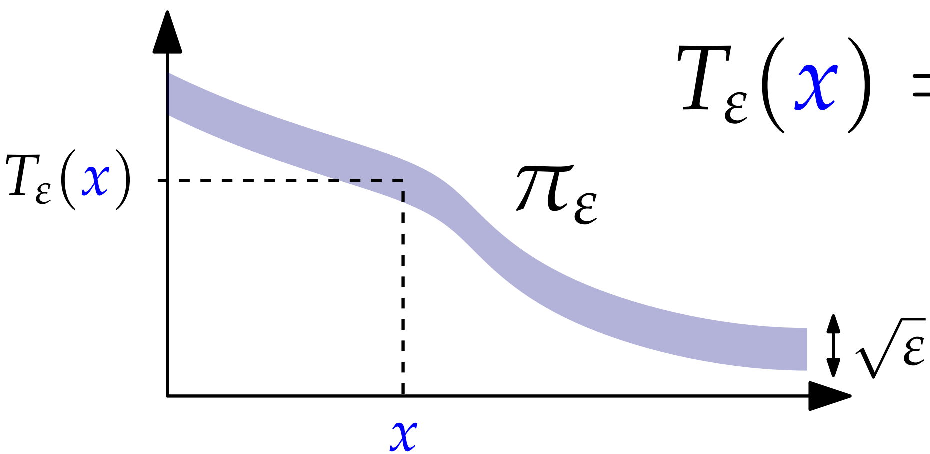
$$\text{minimize} \int \|x - y\|^2 \mathrm{d}\pi(x,y) + \varepsilon\, \mathrm{KL}(\pi \| \mu \otimes \nu)$$

$$\text{under the constraint } \pi \in \Pi(\mu, \nu) \quad \text{(Schrödinger)}$$

$$\pi_\varepsilon(x,y) = e^{\frac{xy - \phi_\varepsilon(x) - \psi_\varepsilon(y)}{\varepsilon}} \mathrm{d}\mu(x)\mathrm{d}\nu(y)$$

→ entropic Brenier potential

$$T_\varepsilon(x) = \int y \,\mathrm{d}\pi_\varepsilon^x(y)$$



→ complexity $O(n^2/\varepsilon^2)$ through Sinkhorn's algorithm [Cuturi, 13]

Theorem:

· μ almost uniform on compact convex support

· $\nu = \sum_{j=1}^{J} q_j \delta_{y_j}$, $q_j \geq q_{\min}$

· $\hat{T}_\varepsilon = T_\varepsilon^{\mu_n \to \nu_n}$

Then, for $\varepsilon \simeq n^{-1/2}$,

$$\mathbb{E}\|\hat{T}_\varepsilon - T_0\|_{L_2(\mu)} \leq \underbrace{\|T_\varepsilon - T_0\|_{L_2(\mu)}}_{\text{bias}} + \underbrace{\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon\|_{L_2(\mu)}}_{\text{fluctuations}}$$

$$\lesssim \sqrt{\varepsilon} + \frac{1}{\sqrt{n\varepsilon}} \simeq n^{-\frac{1}{4}}$$

[Pooladian, D., Niles-Weed, ICML 23]

$\rightarrow$ This rate is minimax optimal!

Take-home messages

• Discontinuous transport maps arise naturally in OT

• Semi-discrete OT: toy model to understand relevant phenomena

• Entropic smoothing $\rightarrow$ fast computations + improved statistical rates

$\rightarrow$ Can we prove similar phenomena in other discontinuous settings (manifolds?)

$\rightarrow$ Can we design a selection procedure for ε?