Inverse optimal transport and related problems

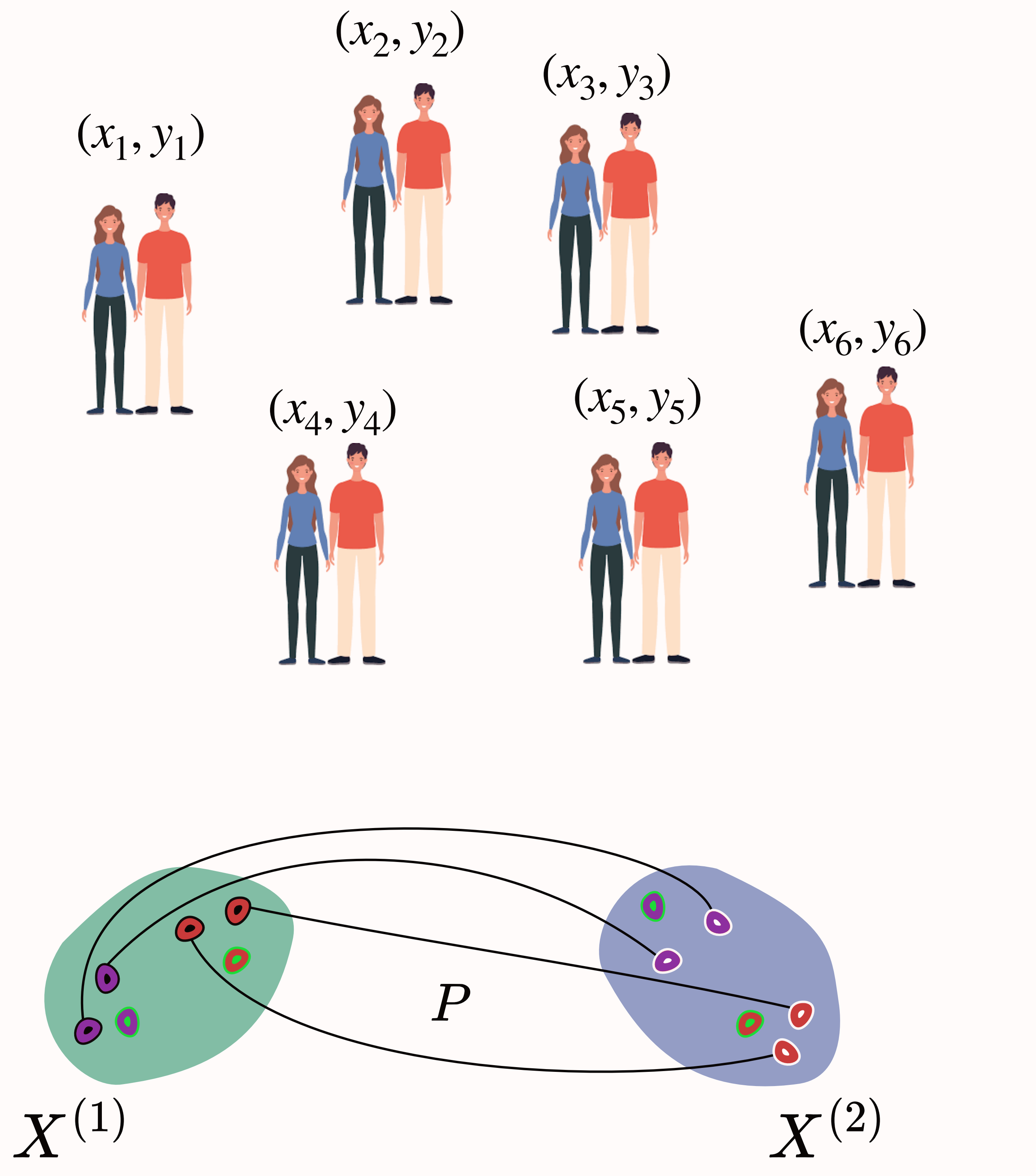
Clarice Poon (University of Warwick)

Joint work with: Francisco Andrade and Gabriel Peyré (ENS Paris)

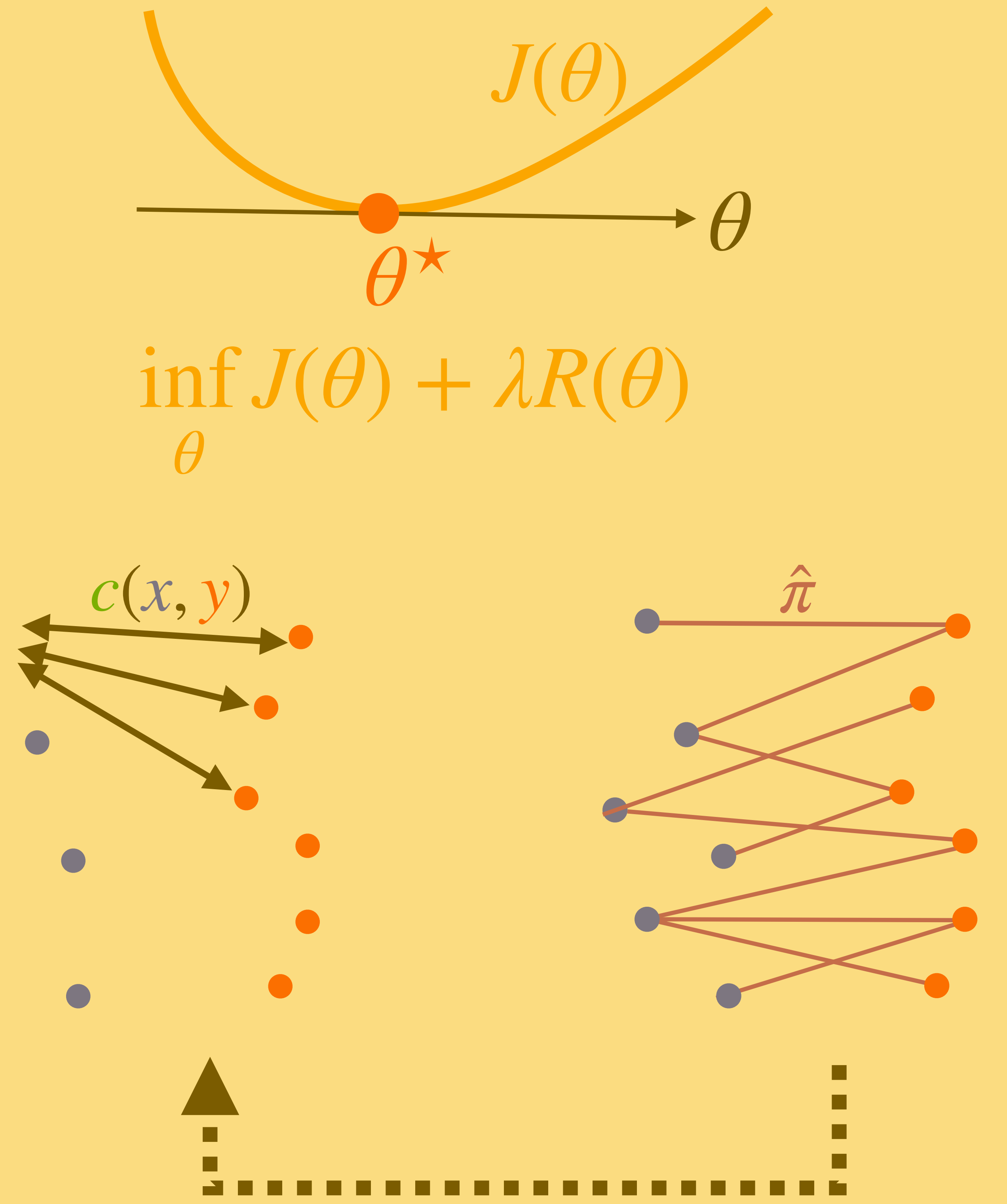
Papers:
Sparsistency for inverse optimal transport, ICLR 2024
Learning from Samples: Inverse Problems over measures via Sharpened Fenchel-Young Losses https://arxiv.org/abs/2505.07124

Outline

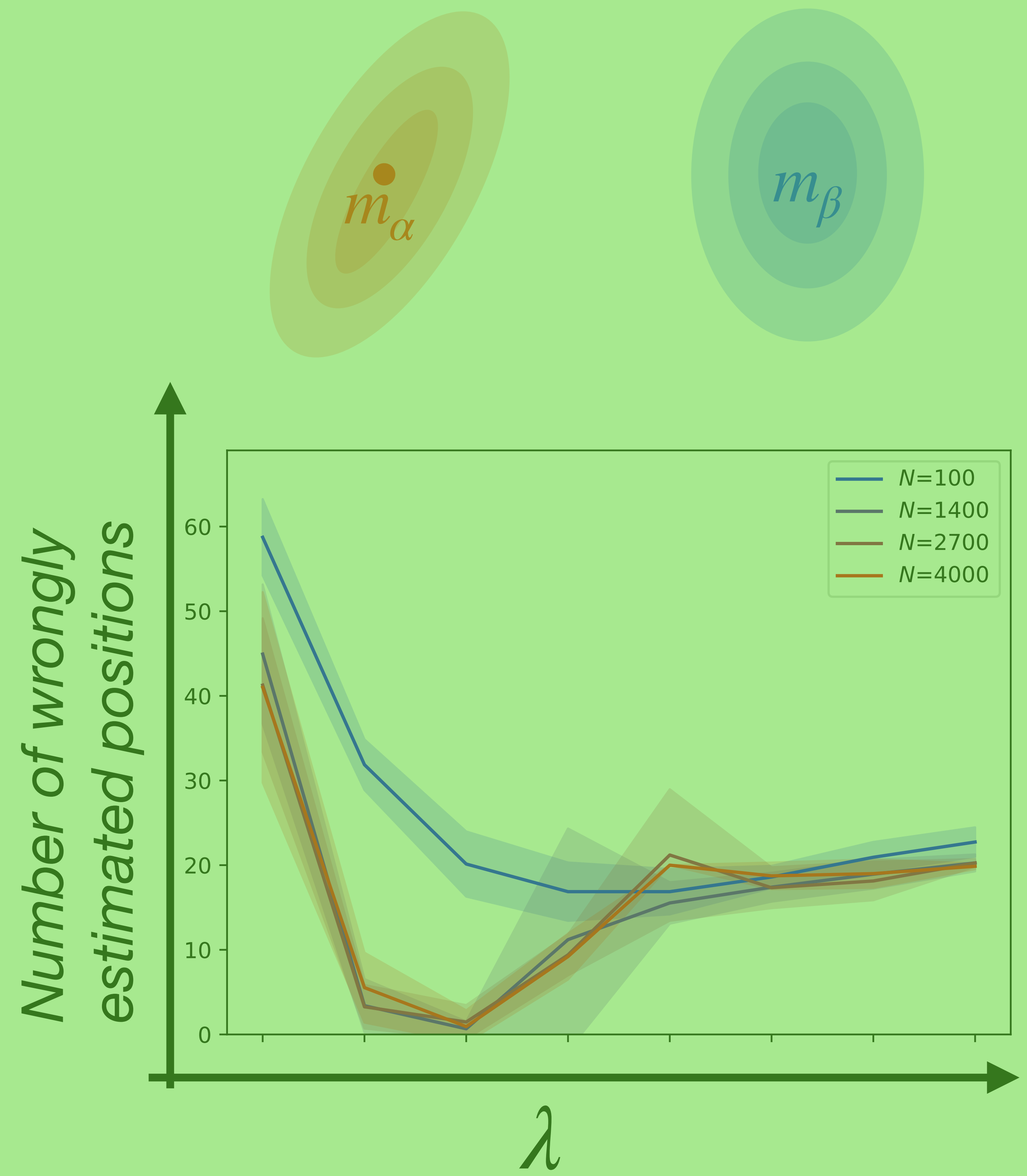
Inverse problems in OT



Learning framework



Recovery guarantees

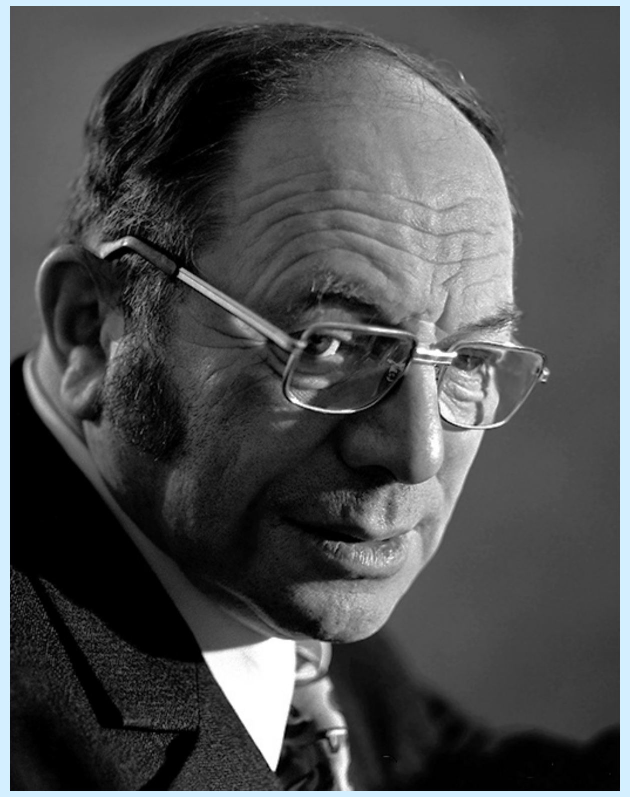


Optimal transport

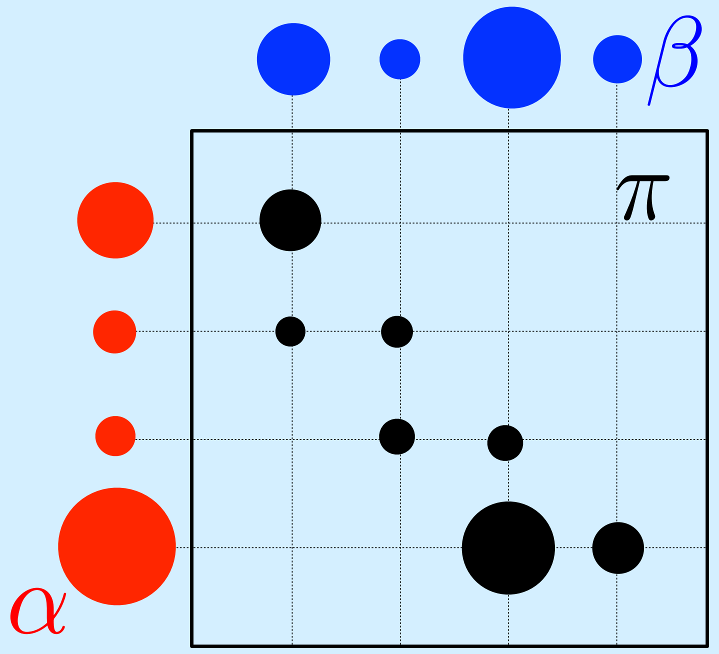
$c(x, y)$ is the **cost** of moving from position x to y

Monge 1781: Given probability measures $\alpha \in \mathscr{P}(\mathscr{X})$ and $\beta \in \mathscr{P}(\mathscr{Y})$, find the optimal way of transporting α to β.

$$\mathrm{OT}(\alpha, \beta) = \inf_{T_\sharp \alpha = \beta} \int c(x, T(x)) d\alpha(x)$$



Kantorovich 1942



$$\mathrm{OT}(\alpha, \beta) = \inf_{\pi_1 = \alpha, \pi_2 = \beta} \int c(x, y) d\pi(x, y)$$

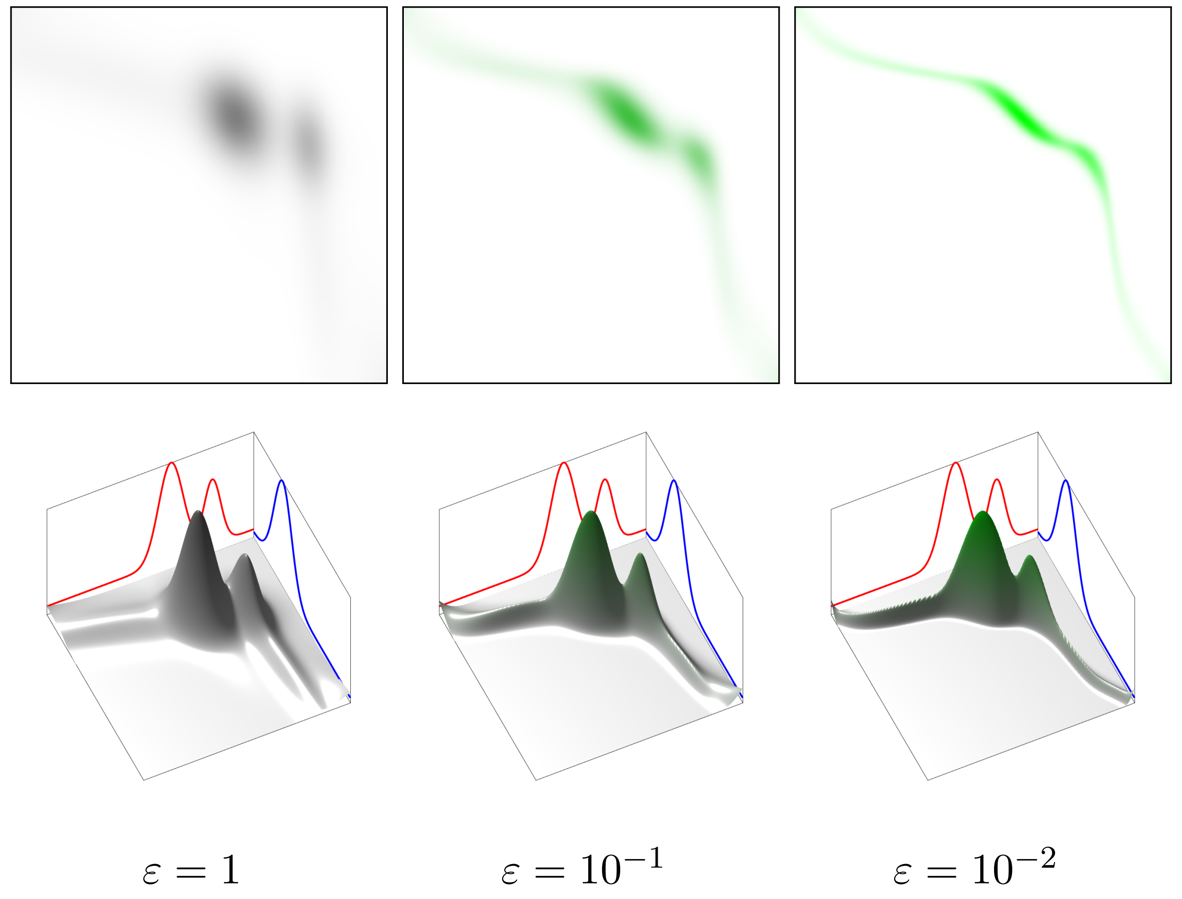
$$\longleftrightarrow \quad \sup_{f \oplus g \leq c} \int f d\alpha + \int g d\beta$$

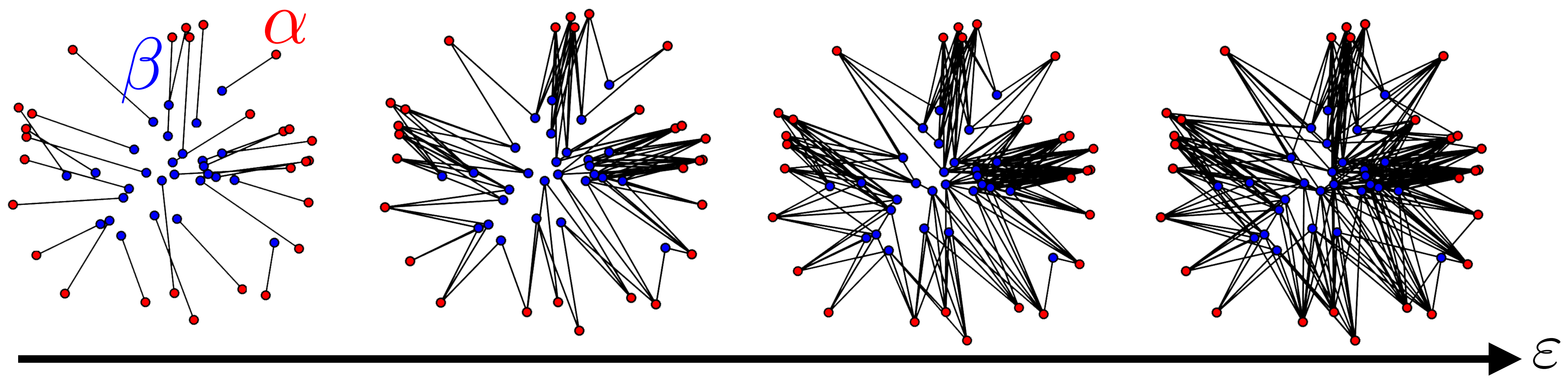
Convex

Entropic optimal transport

Regularize with $\mathrm{KL}(\pi\,|\,\mu) := \int \log(d\pi/d\mu)d\pi$ and $\epsilon > 0$

$$\mathrm{eOT}(\alpha, \beta) = \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int c(x,y)d\pi(x,y) + \epsilon \mathrm{KL}(\pi\,|\,\alpha \otimes \beta)$$





Images credit [Peyré & Cuturi '19]

Entropic optimal transport

Regularize with $\mathrm{KL}(\pi | \mu) := \int \log(d\pi/d\mu) d\pi$ and $\epsilon > 0$

$$\mathrm{eOT}(\alpha, \beta) = \inf_{\pi \in \mathcal{U}(\alpha,\beta)} \int c(x, y) d\pi(x, y) + \epsilon \mathrm{KL}(\pi | \alpha \otimes \beta)$$

- **Natural** modelling assumption
- Alleviates the **curse of dimensionality** [Genevay et al '19, Mena & Weed '19]

$$\mathrm{eOT}(\alpha, \beta) - \mathrm{eOT}(\alpha_n, \beta_n) = \mathcal{O}(n^{-1/2})$$

- **Fast algorithms** available [Sinkhorn '64, Cuturi '13]

$$\sup_{f,g} \int f d\alpha + \int g d\beta - \epsilon \int \exp\left(\frac{f(x) + g(y) - c(x, y)}{\epsilon}\right) d\alpha(x) d\beta(y)$$

Sinkhorn algorithm

- **Fast algorithms** available [Sinkhorn '64, Cuturi '13]

$$\sup_{f,g} \int f d\alpha + \int g d\beta - \epsilon \int \exp\left(\frac{f(x) + g(y) - c(x,y)}{\epsilon}\right) d\alpha(x) d\beta(y)$$

First order optimality:

$$\exp(-f/\epsilon) = \int \exp((g(y) - c(x,y))/\epsilon) d\beta(y)$$

$$\exp(-g/\epsilon) = \int \exp((f(x) - c(x,y))/\epsilon) d\alpha(y)$$

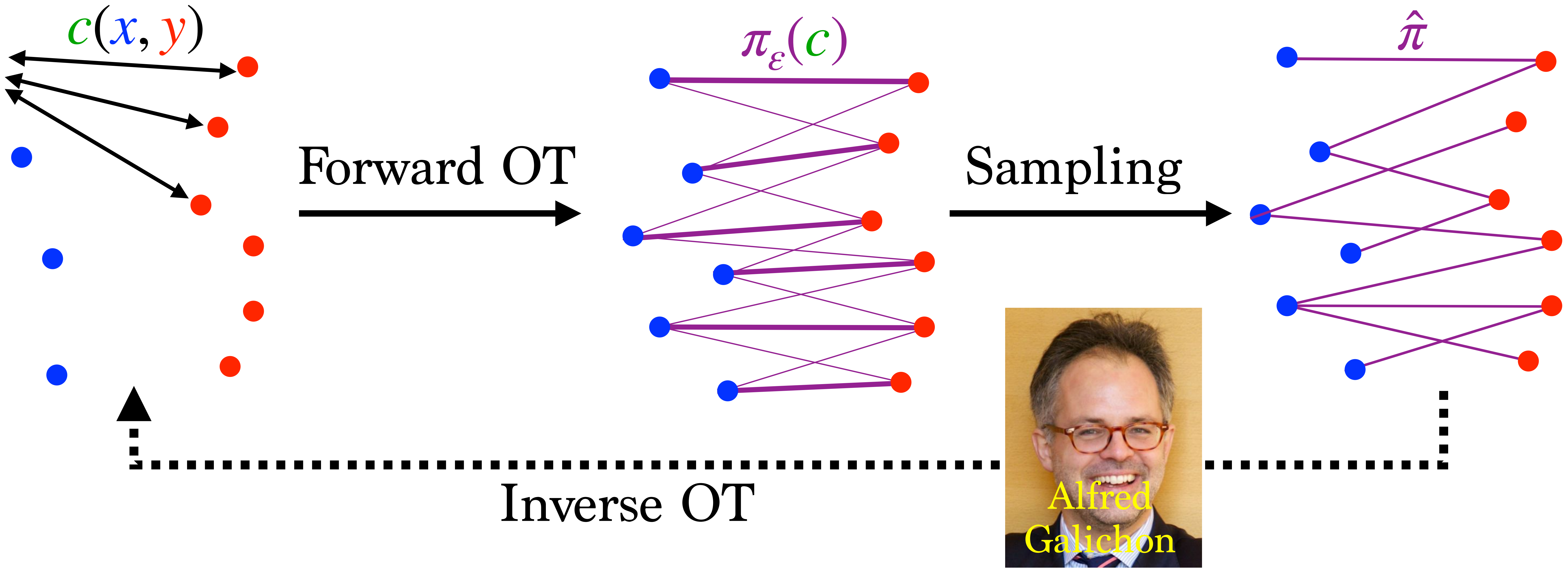
Sinkhorn is alternating minimisation:

$$f_{k+1} = -\epsilon \log\left(\int \exp((g_k(y) - c(x,y))/\epsilon) d\beta(y)\right)$$

$$g_{k+1} = -\epsilon \log\left(\int \exp((f_{k+1}(x) - c(x,y))/\epsilon) d\alpha(y)\right)$$

Inverse optimal transport

Given probability measures α, β and a ground cost $c(x, y)$, compute the optimal coupling.

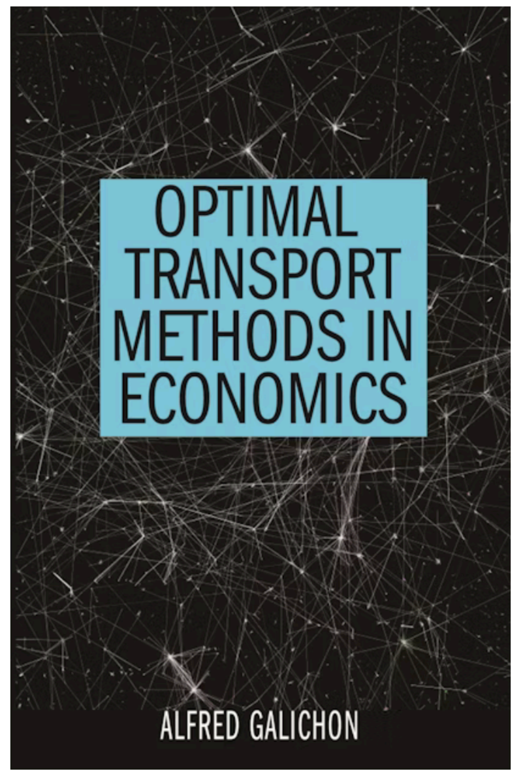


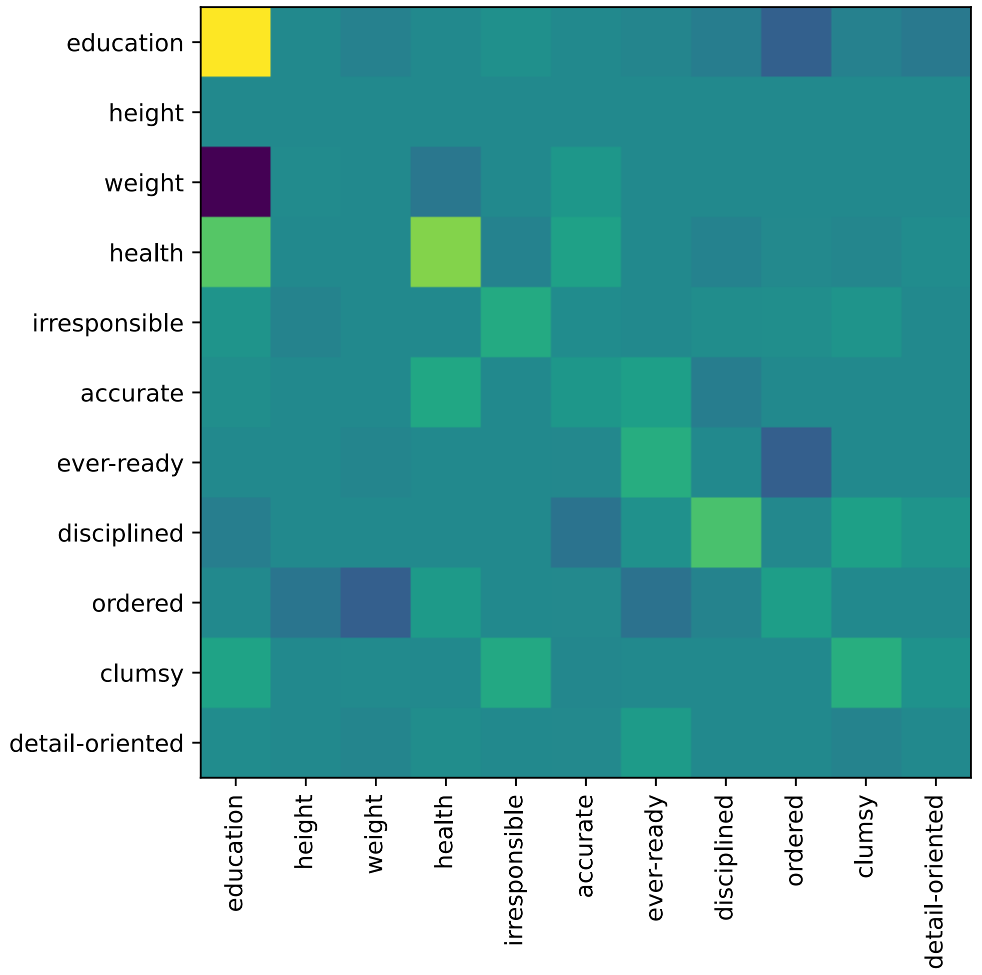
Suppose you observe how two populations are coupled. How can we infer the 'cost' that led to this coupling?

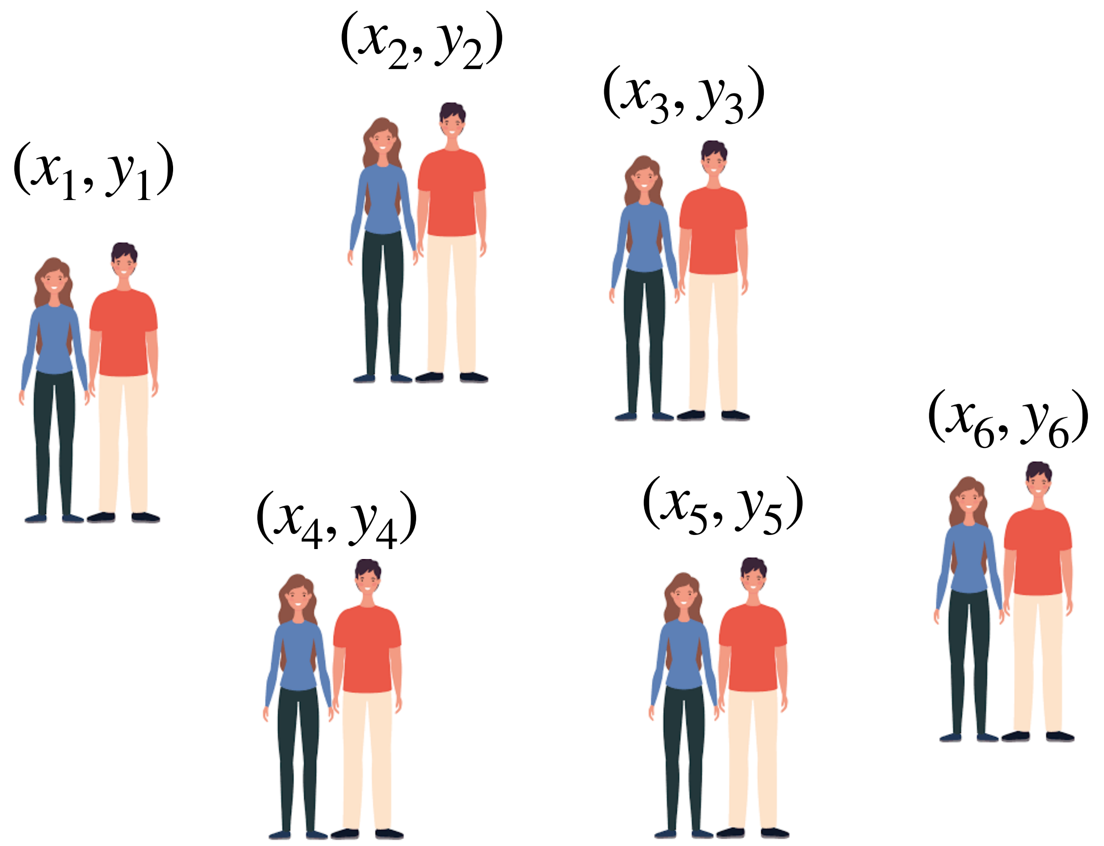
Galichon, Alfred, and Bernard Salanié. "Cupid's Invisible Hand: Social Surplus and Identification in Matching Models." (2015).
Galichon, Alfred. *Optimal transport methods in economics*. Princeton University Press, 2016.
Dupuy, Arnaud, Alfred Galichon, and Yifei Sun. "Estimating matching affinity matrices under low-rank constraints." *Information and Inference: A Journal of the IMA* 8.4 (2019): 677-689.

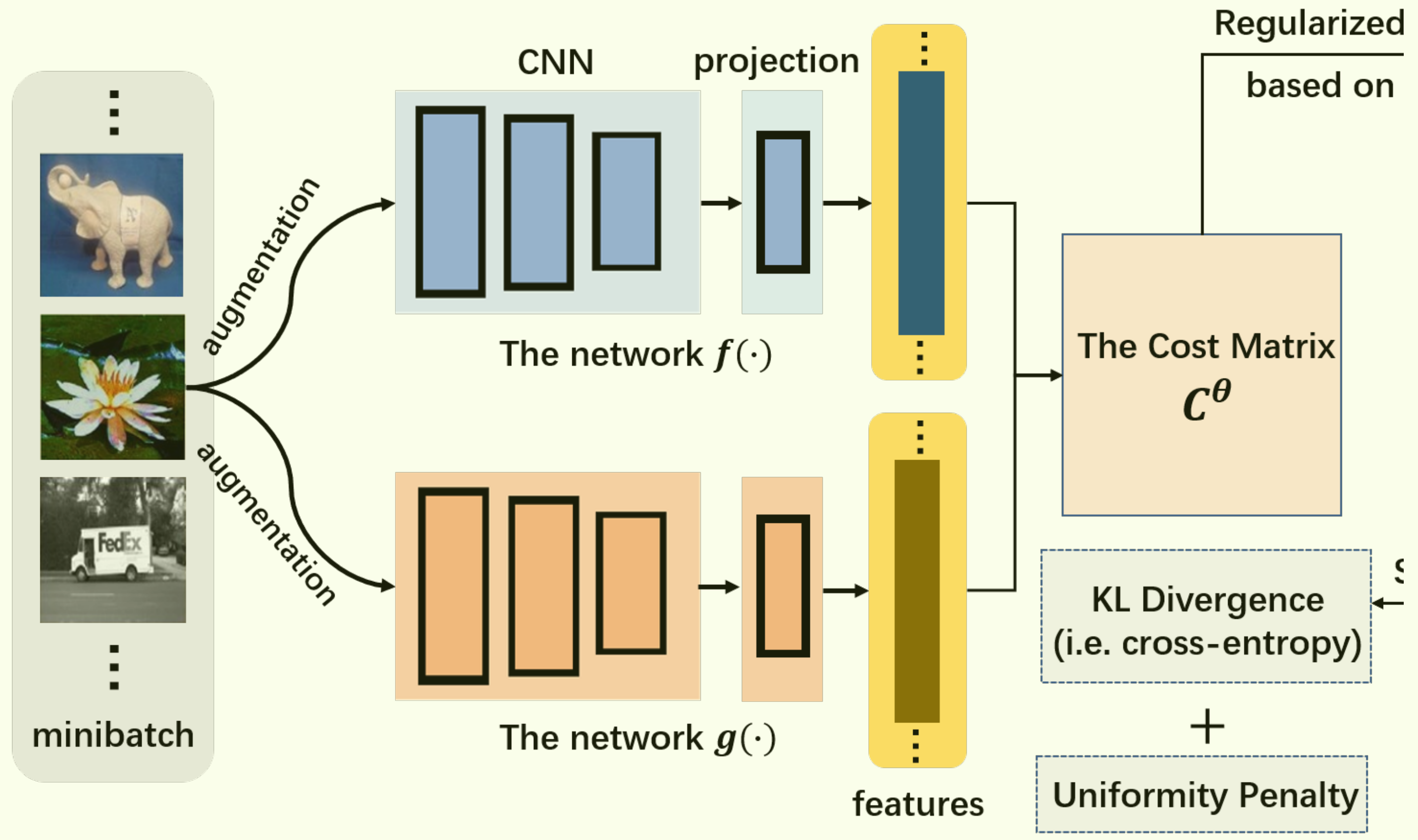
Understanding matching

Dupuy & Galichon 2014. "Personality traits and the marriage market."

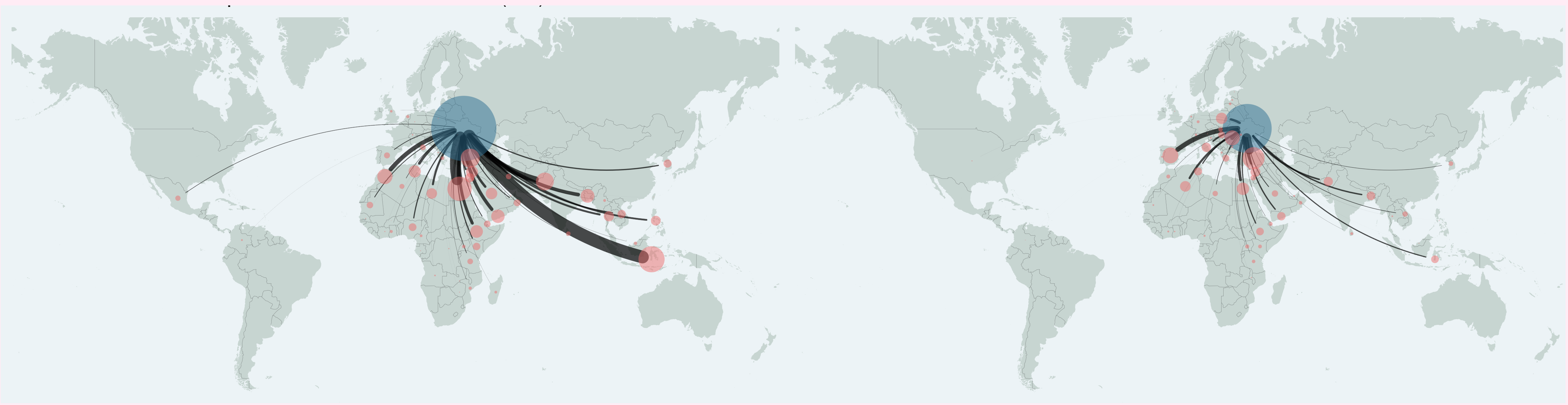








Shi et al. "Understanding and generalizing contrastive learning from the inverse optimal transport perspective." *ICLR*2023.

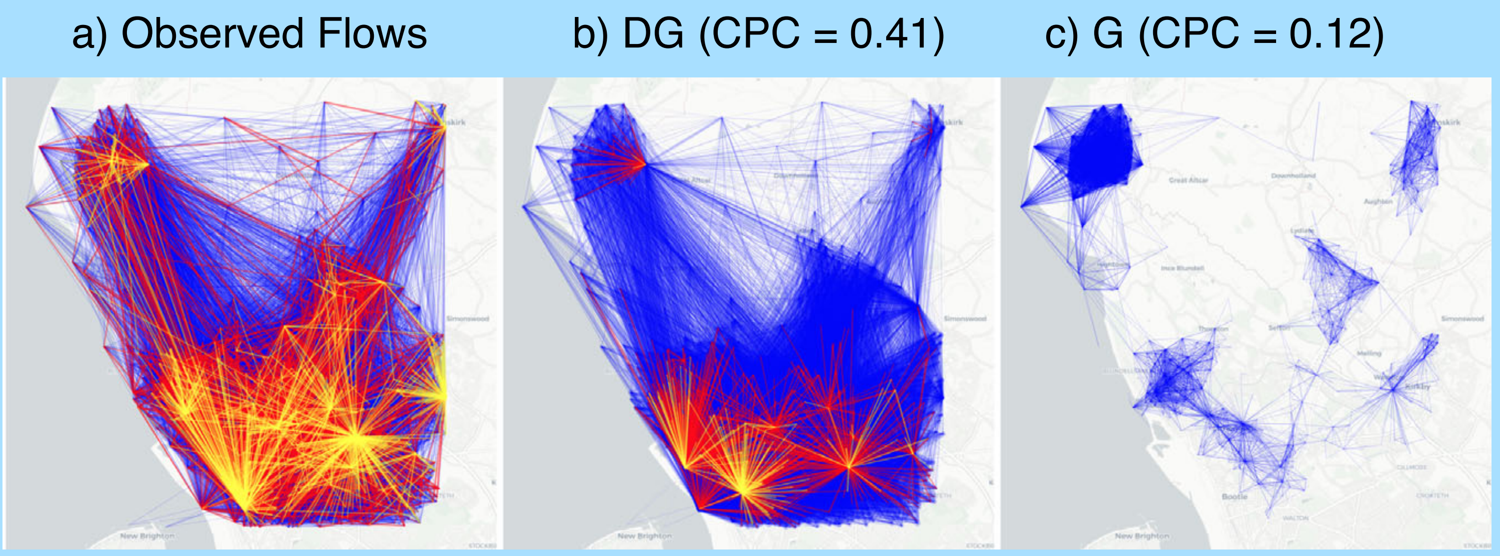


Modelling Global Trade with Optimal Transport

Thomas Gaskin, Marie-Therese Wolfram, Andrew Duncan, Guven Demirel

A Deep Gravity model for mobility flows generation

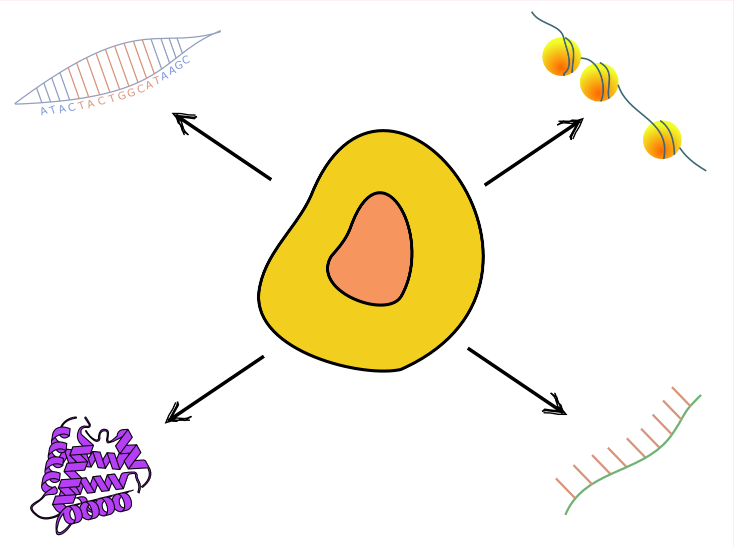
Filippo Simini[1,2,3], Gianni Barlacchi[4], Massimiliano Luca[5,6] & Luca Pappalardo[7]

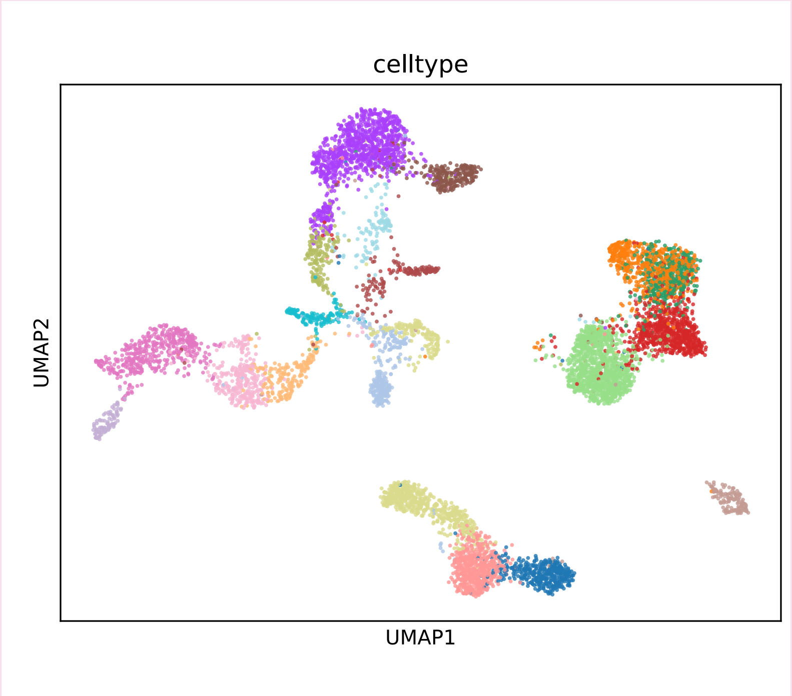


Cell genomics

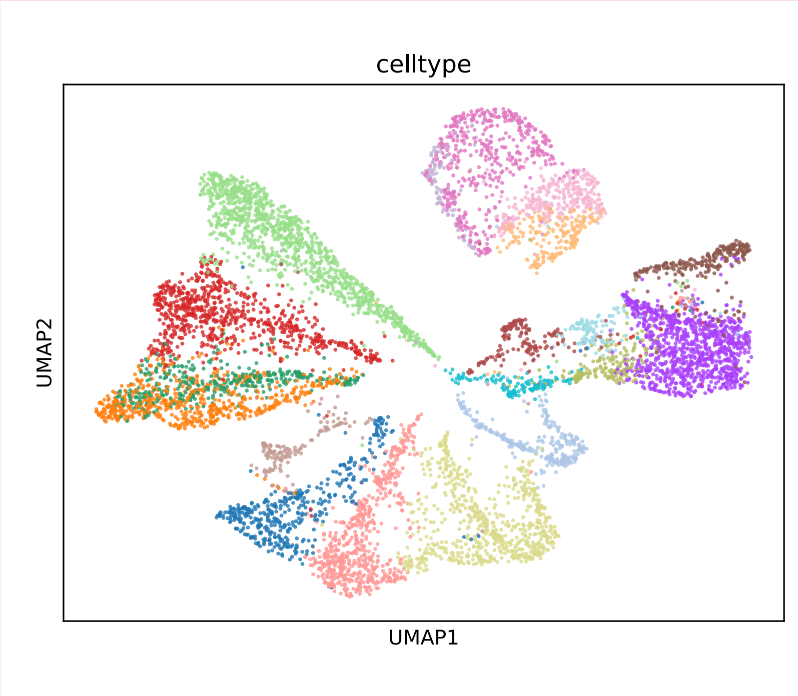
Jules Samaran, Gabriel Peyré, Laura Cantini

Cellular data is multimodal



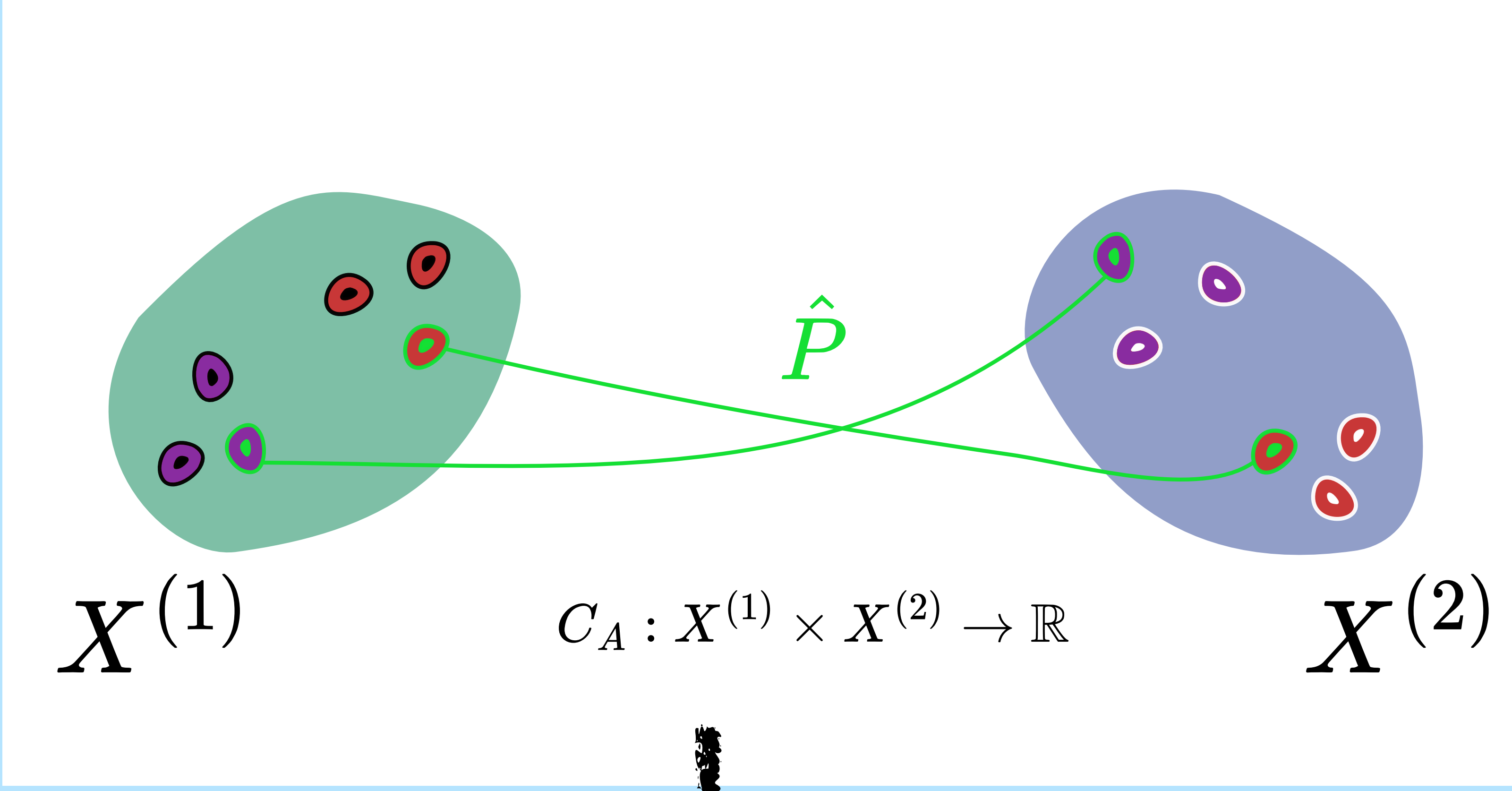


Gene expression

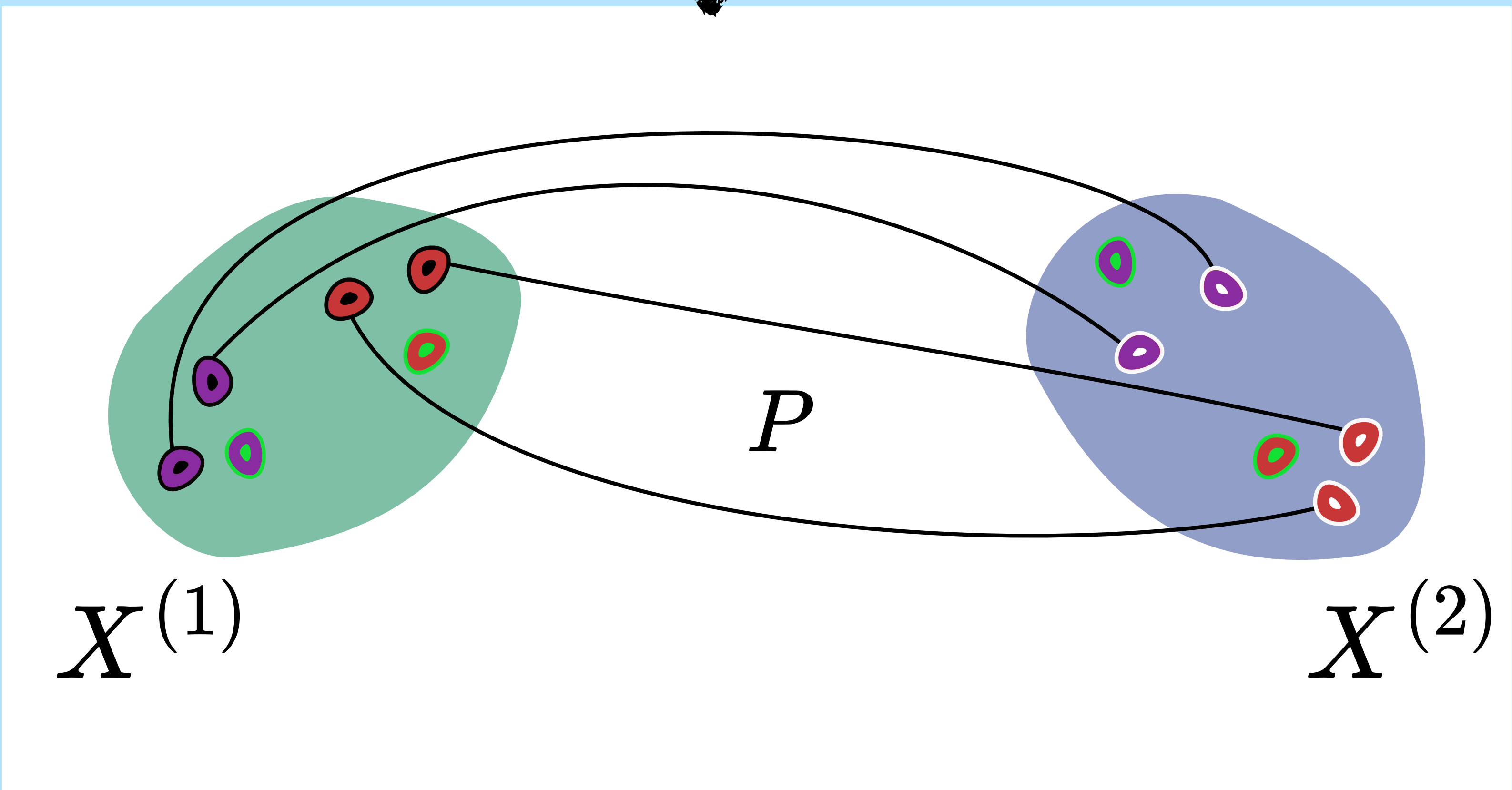


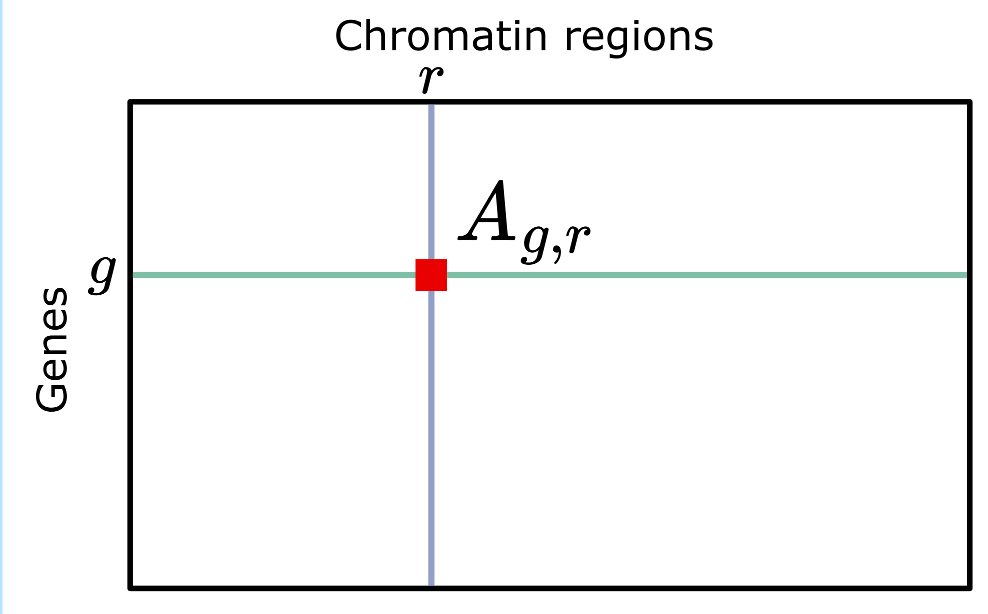
Chromatin accessibility

- B1 B
- CD4+ T activated
- CD4+ T naive
- CD8+ T
- CD14+ Mono
- CD16+ Mono
- Erythroblast
- G/M prog
- HSC
- Lymph prog
- MK/E prog
- NK
- Naive CD20+ B
- Normoblast
- Plasma cell
- Proerythroblast
- Transitional B
- cDC2
- pDC

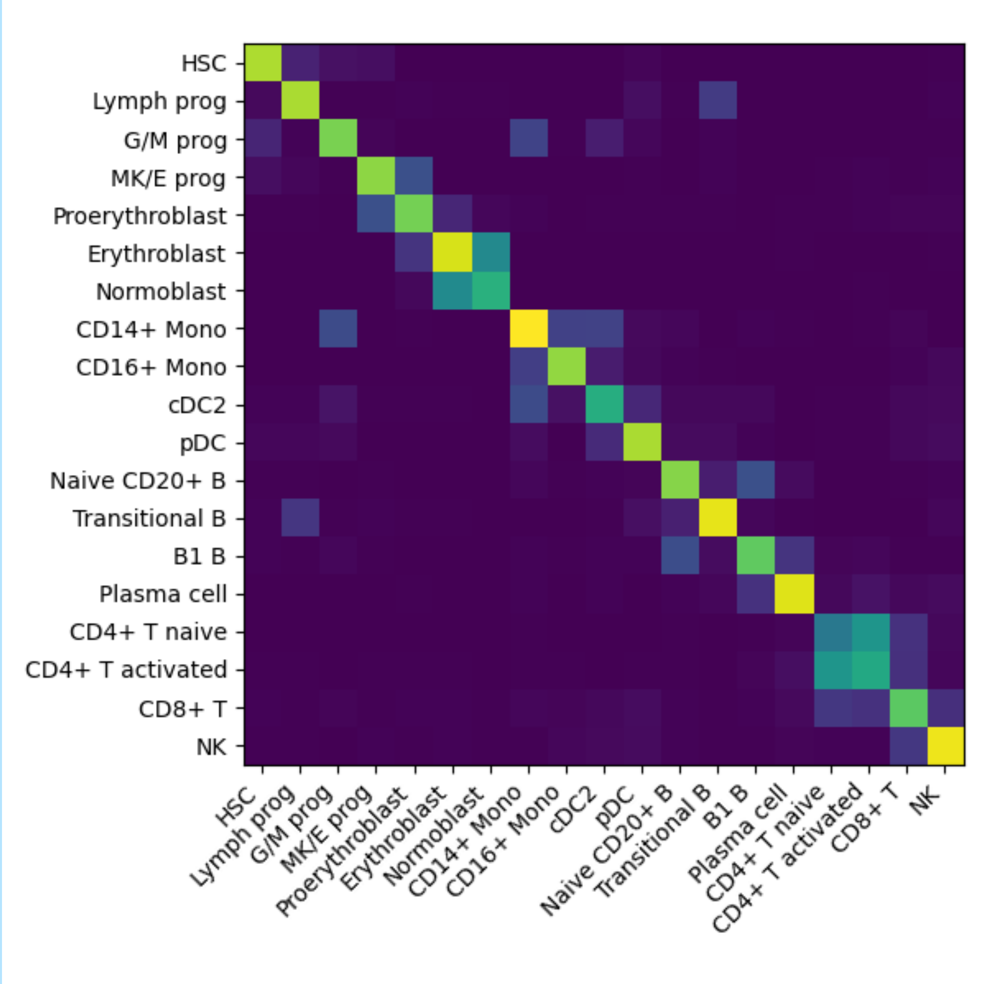


Learn the missing links





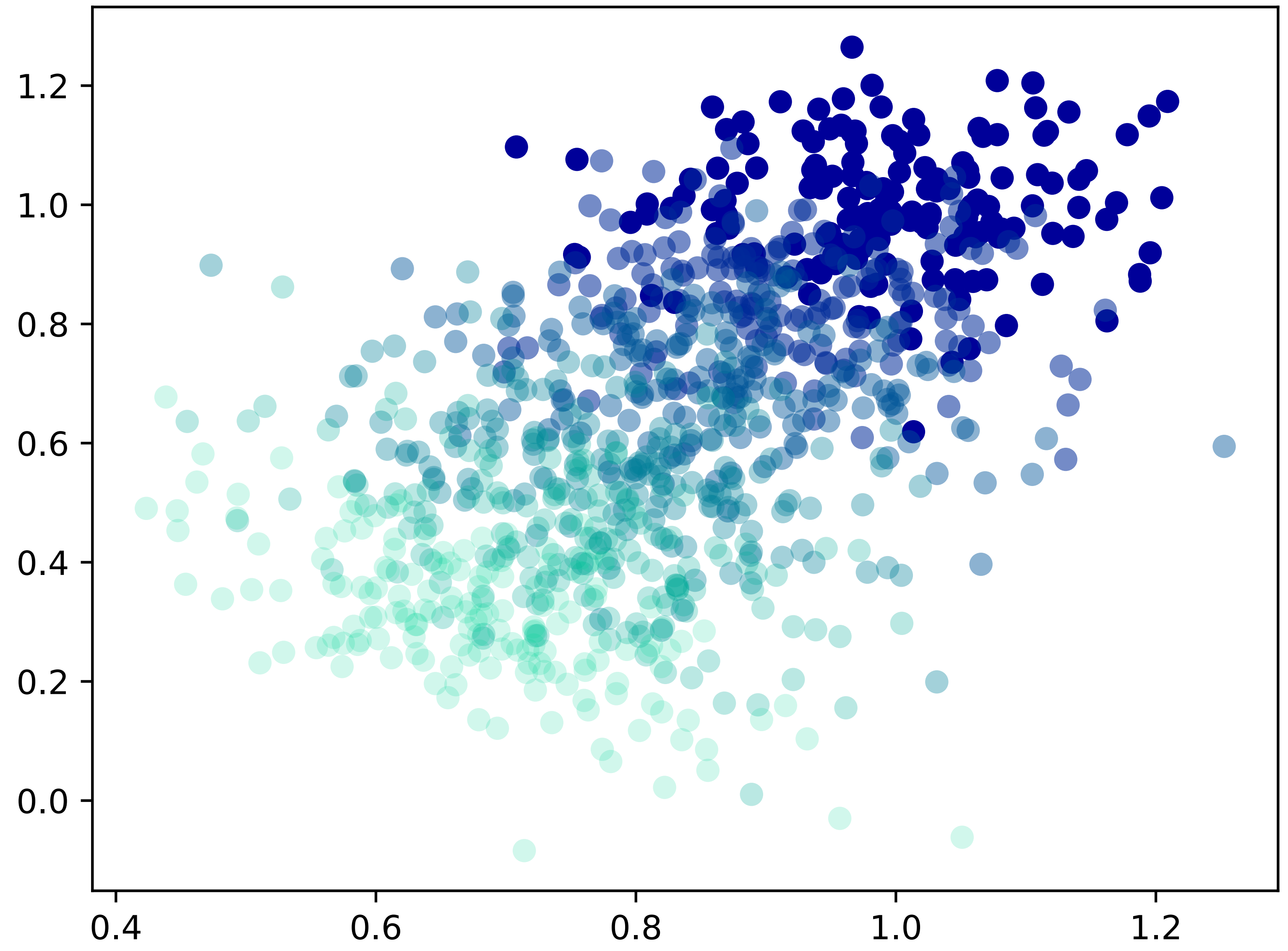
$$C_A(x_1, x_2) = x_1^T A x_2$$



89% accuracy

How do populations evolve?

$$\frac{dX_i}{dt} = \mathbf{v}_t(X_i) \quad \overset{X_i \overset{iid}{\sim} \rho_t}{\longleftrightarrow} \quad \partial_t \rho_t + \nabla \cdot (\rho_t \mathbf{v}_t) = 0$$



Goal: Given iid samples of $\rho_{t_0}, \rho_{t_1}, \rho_{t_2}, \ldots \rho_{t_T}$, find $\mathbf{v}_t$.

Examples:
$\mathbf{v} = \nabla V$ or $\mathbf{v} = \int \nabla_1 W(\cdot, x) d\rho(x)$

Modelling using Wasserstein gradient flows:

Suppose $\mathbf{v}_t = -\nabla \delta \mathscr{F}(\rho_t)$ for some $\mathscr{F} : \mathscr{P}(X) \to \mathbb{R}$

Model using discretised dynamics: $\alpha_{k+1} = \underset{\alpha \in \mathscr{P}(X)}{\operatorname{argmin}} \mathscr{F}(\alpha) + \frac{1}{2\tau} W_2^2(\alpha, \alpha_k)$

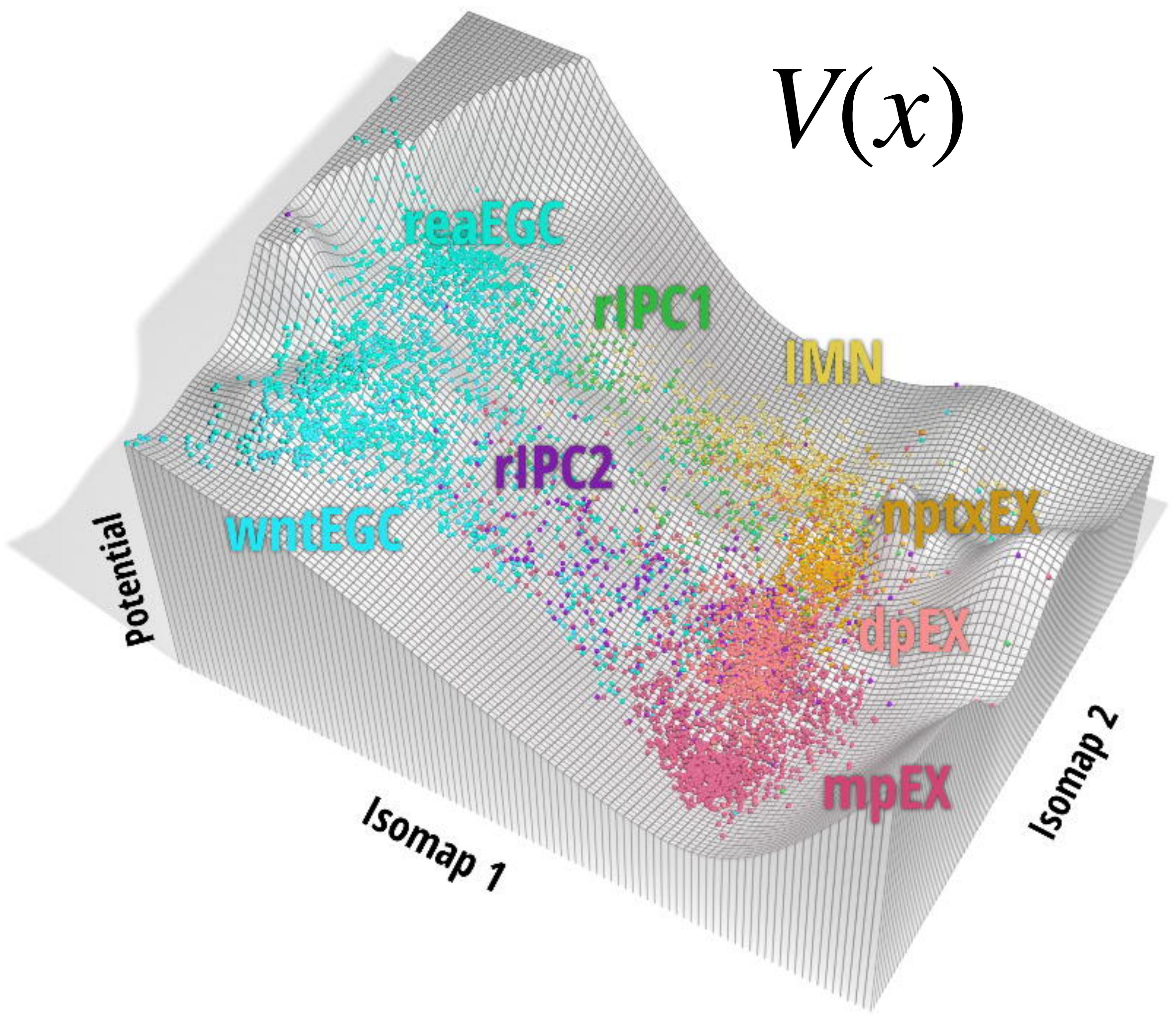
Jordon Kinderlehrer Otto scheme (JKO)

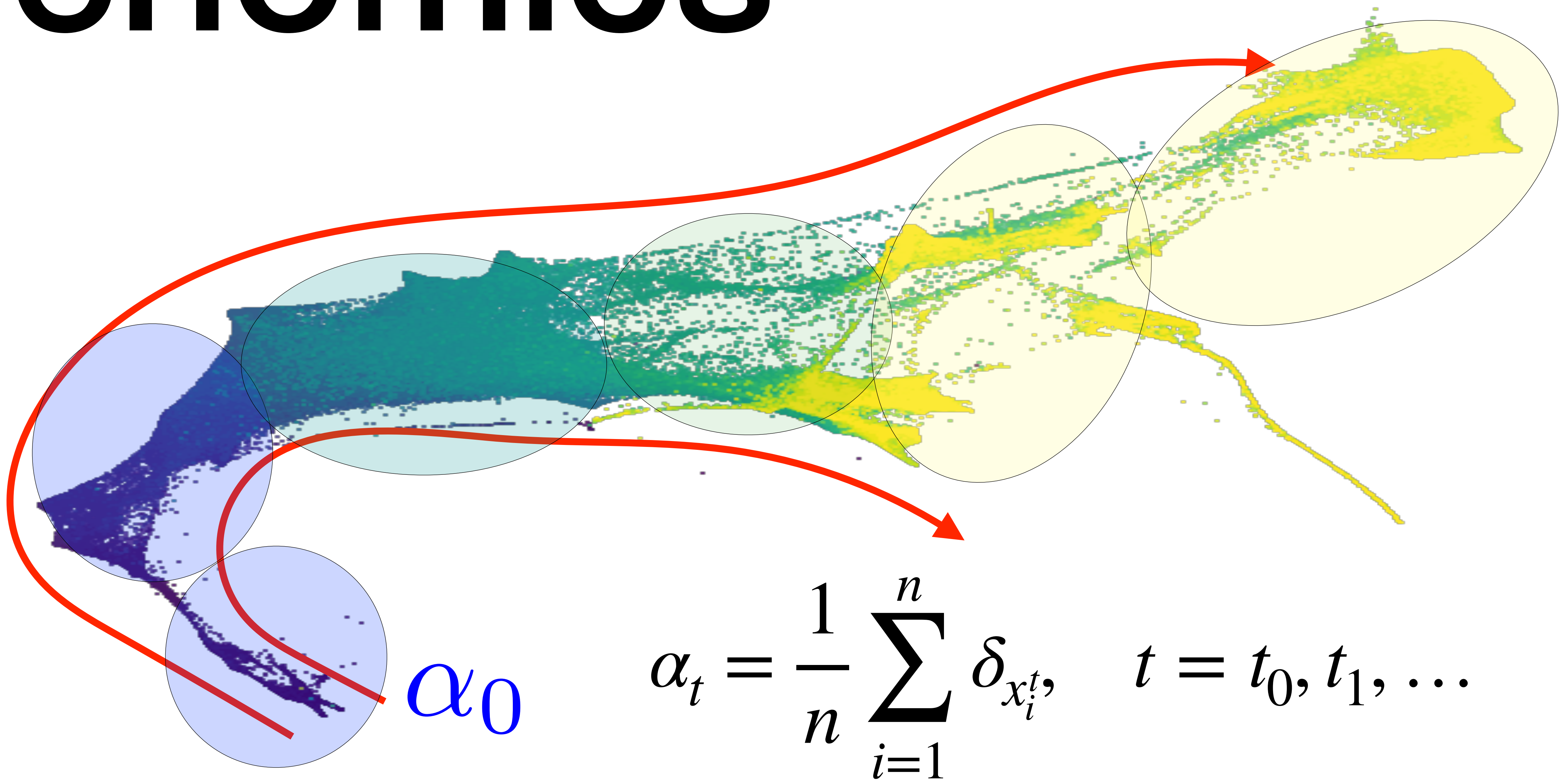
↔ Discretisation of curve of probability measures $\alpha_k \approx \rho_{k\tau}$

Example: $\mathscr{F}(\rho) = \int V(x) d\rho(x)$ so $\delta \mathscr{F} = V$ and $\mathbf{v} = \nabla V$

Example: $\mathscr{F}(\rho) = \frac{1}{2} \int W(x, y) d\rho(x) d\rho(y)$, $\delta \mathscr{F}(\rho) = \int W(\cdot, x) d\rho(x)$ and $\mathbf{v} = \int \nabla_1 W(\cdot, x) d\rho(x)$

Cell Genomics





$$\alpha_t = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i^t}, \quad t = t_0, t_1, \ldots$$

$x_i^t =$ vector of genes

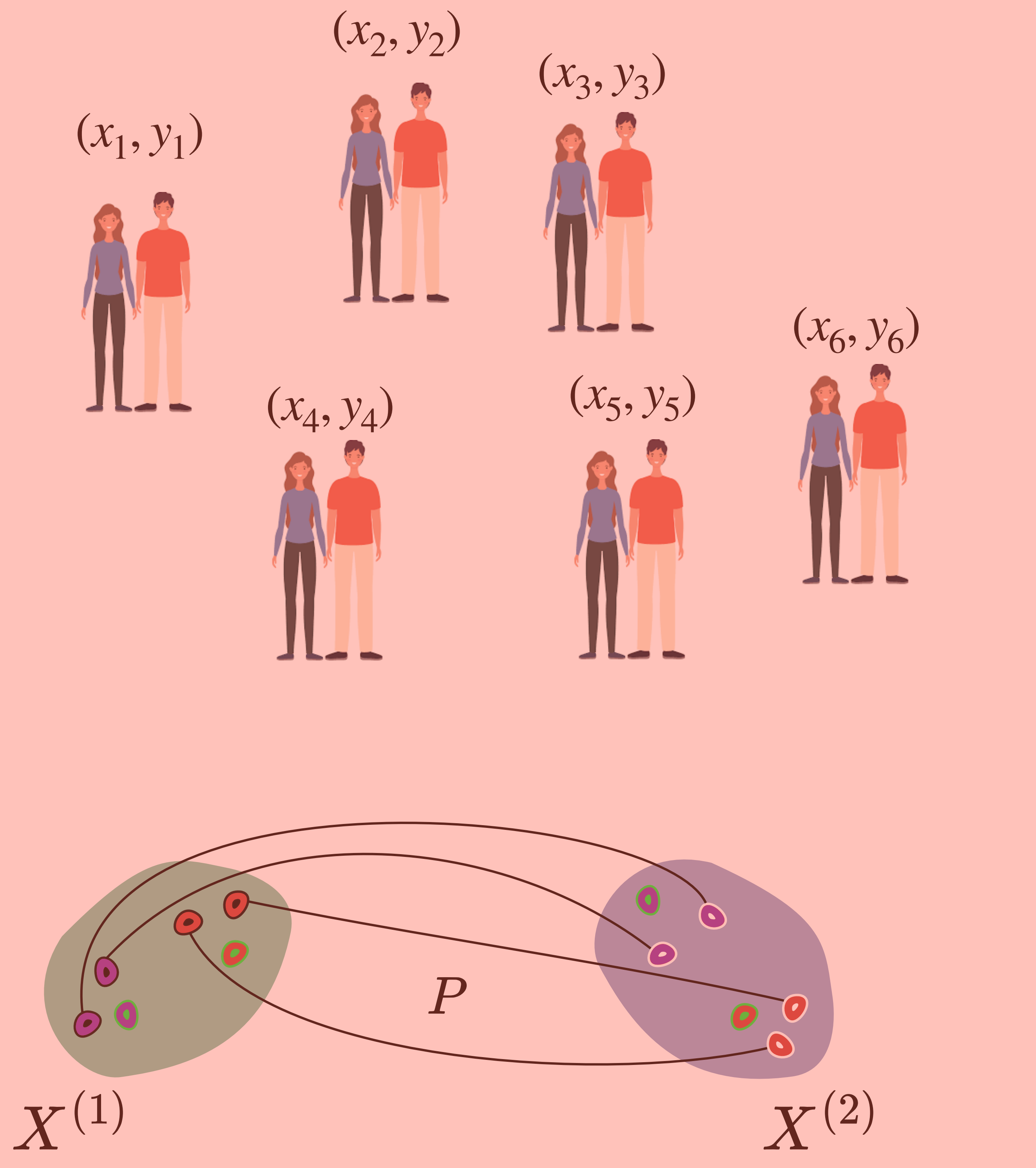
Model as JKO flow

$$\alpha_{t+1} = \operatorname{argmin}_\rho \int V(x) d\rho(x) + W_2^2(\rho, \alpha_t)$$

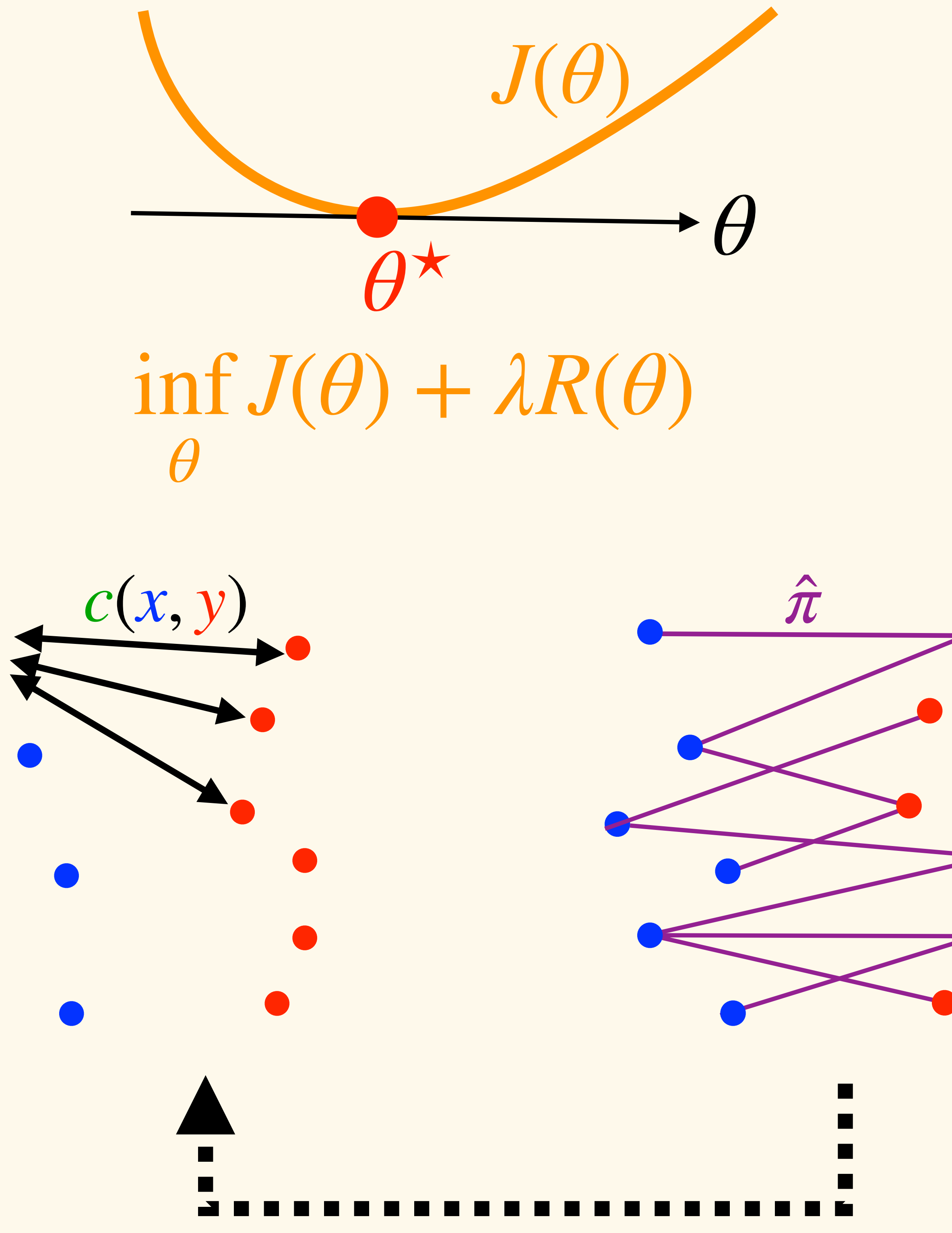
Reference: Learning cell fate landscapes from spatial transcriptomics using Fused Gromov-Wasserstein.
Geert-Jan Huizing, Gabriel Peyré, Laura Cantini

Outline

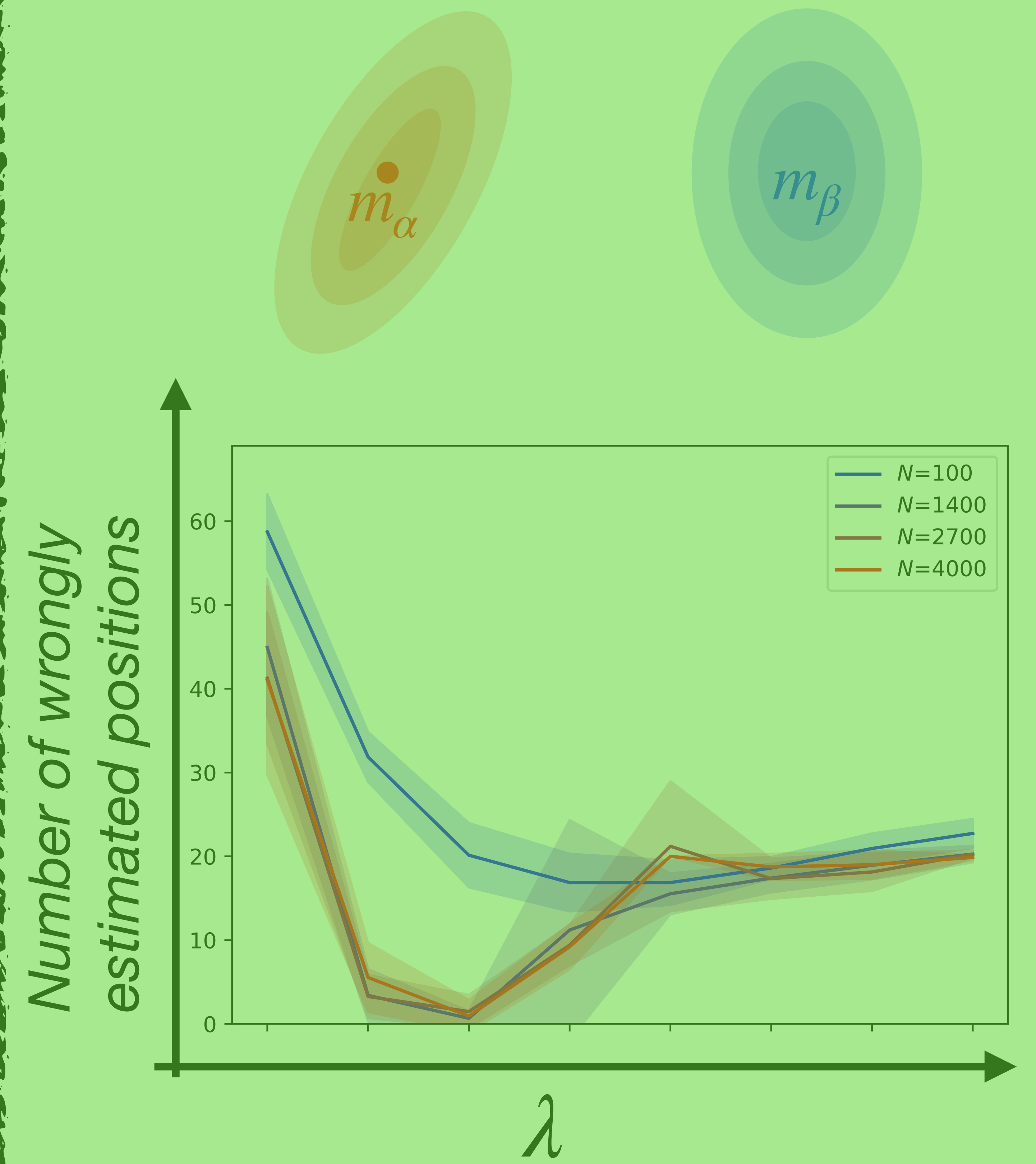
Inverse problems in OT



Learning framework

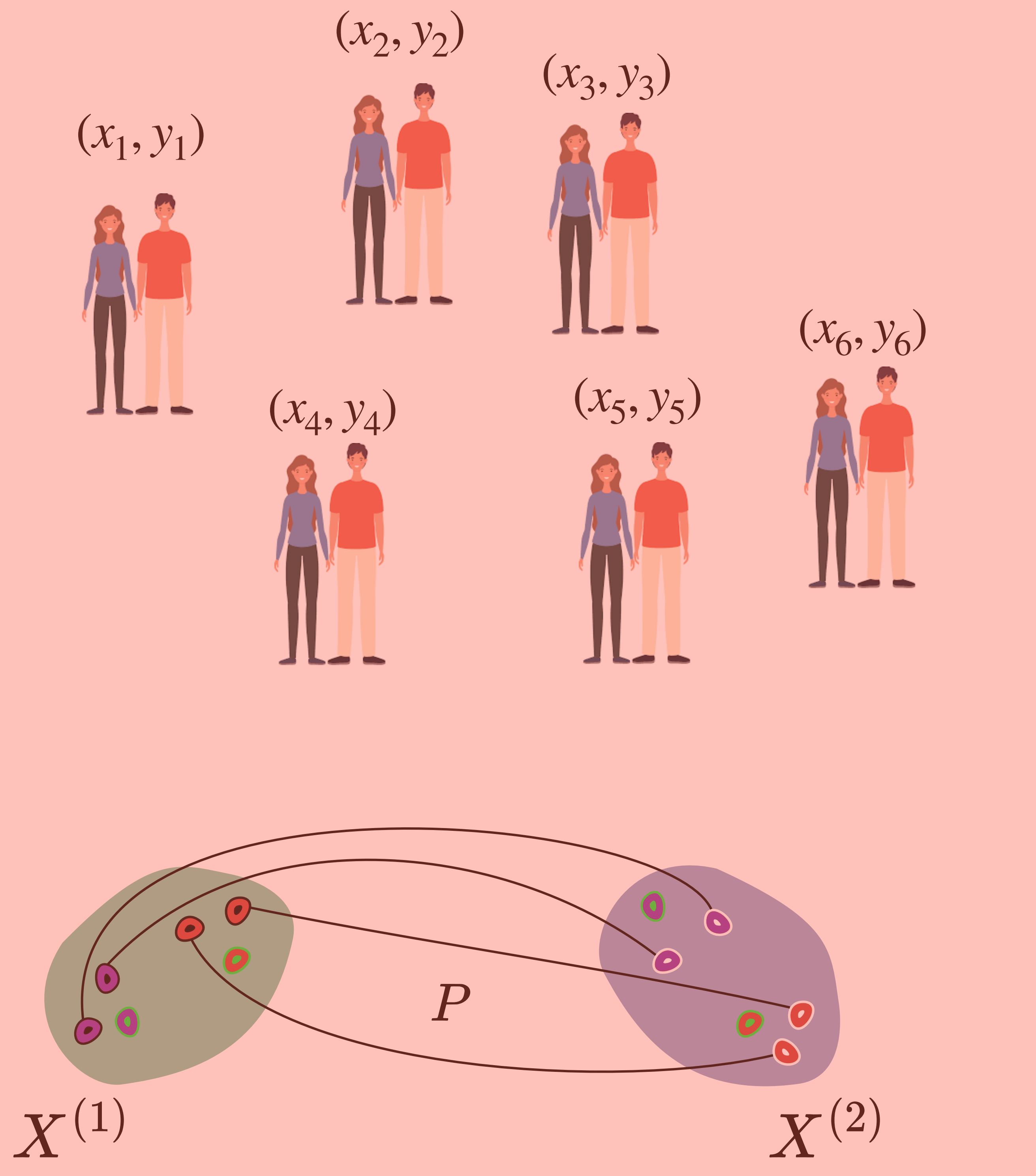


Recovery guarantees

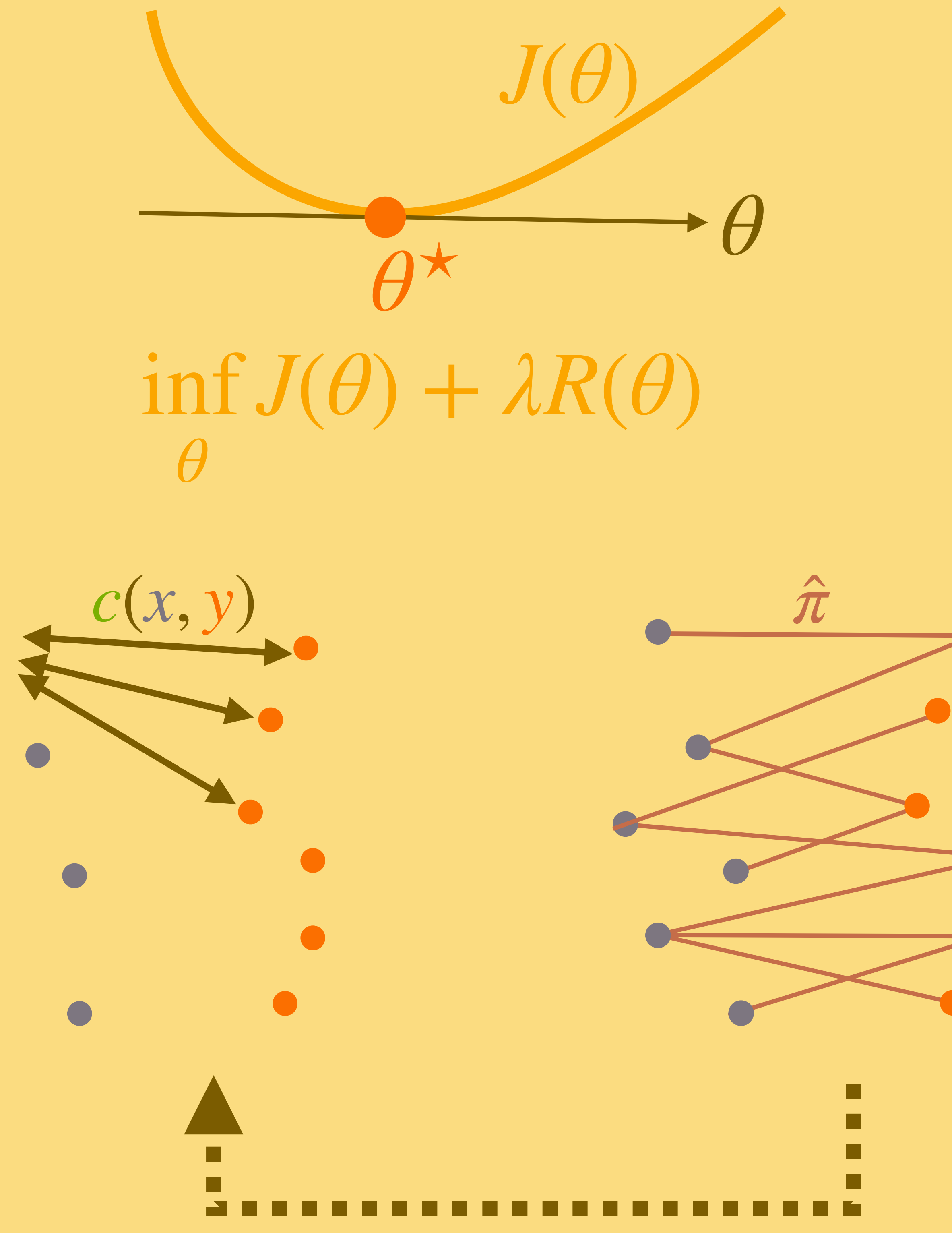


Outline

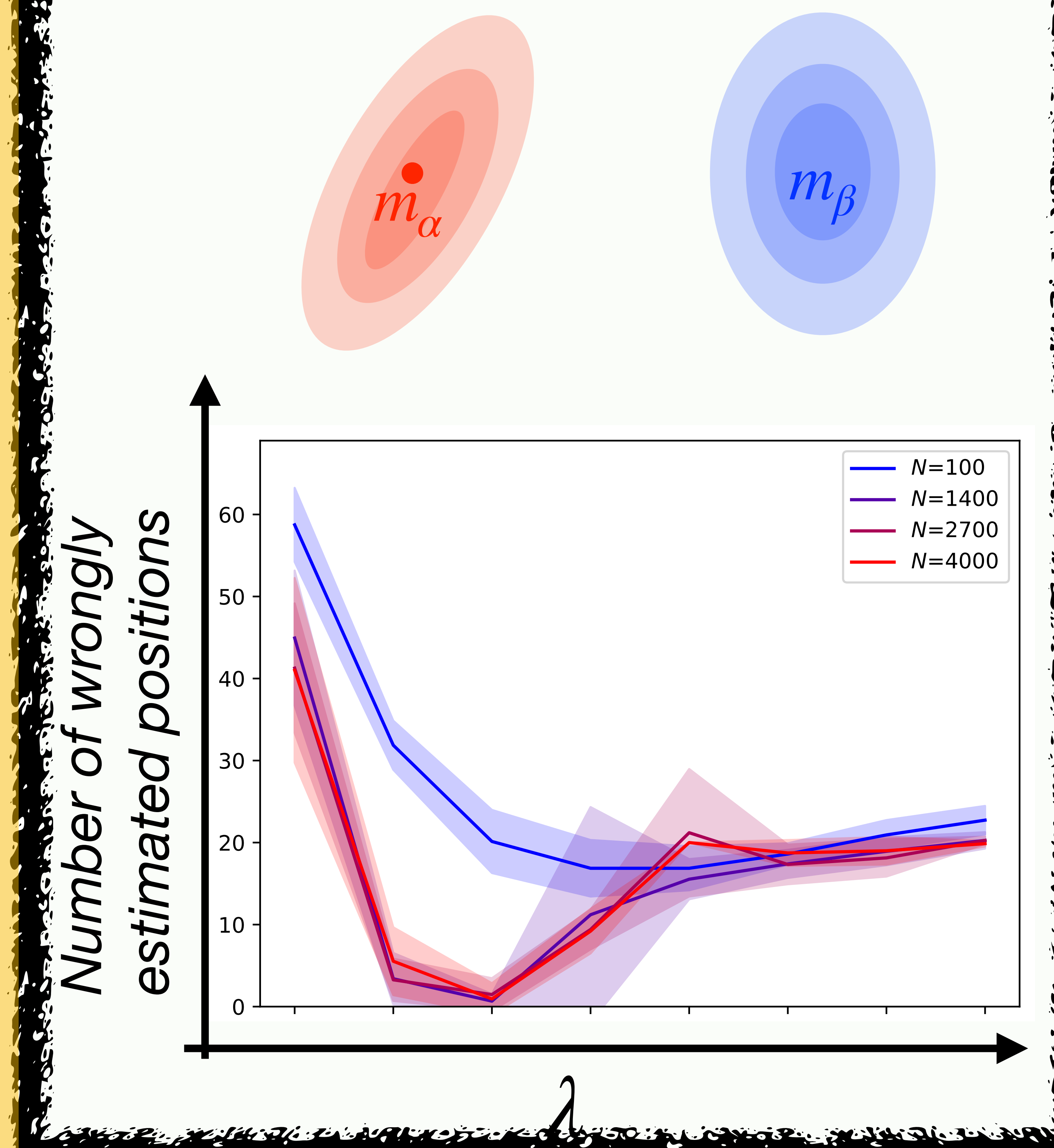
Inverse problems in OT



Learning framework



Recovery guarantees



Stability guarantee

Suppose that $\pi^\star = P_\Omega(c_\theta^\star) := \text{argmin}_c \langle c_{\theta^\star}, \pi \rangle + \Omega(\pi)$. Assume that $\hat{\pi}$ is noisy version of $\pi^\star$. Solve:

$$\inf_{\theta \in \mathbb{R}^p} \mathscr{L}(c_\theta, \hat{\pi}, \hat{\Omega}) + \lambda R(\theta)$$

Theorem: Let $\gamma > 0$ be the 'noise' level. If there is:

- Measurement stability, $|\langle \psi, \hat{\pi} - \pi^\star \rangle| \leq \gamma$ for all basis elements ψ
- Forward stability: $|\langle \psi, P_\Omega(c_\theta^\star) - P_{\hat{\Omega}}(c_\theta^\star) \rangle| \leq \gamma$ for all basis elements ψ
- Local curvature: $J(\theta) := \mathscr{L}(c_\theta | \hat{\pi}, \hat{\Omega})$ is locally strongly convex and Lipschitz smooth.

Then, the minimizer θ satisfies $\|\theta - \theta^\star\| = \mathcal{O}(\lambda + \gamma)$.

Assumptions

Loss to minimize in the case of **iOT**:

$$\inf_{f,g,c} \langle c - (f \oplus g), \hat{\pi} \rangle + \epsilon \int \exp\left(\frac{f(x) + g(y) - c(x,y)}{\epsilon}\right) d\hat{\alpha}(x) d\hat{\beta}(y)$$

Non-uniqueness for iOT

1. If f, g minimizers $\iff$ $f + a, g - a$ are minimisers for all constants a.
2. Replace c with $c - u \oplus v \iff f + u, g + v$ are minimisers.

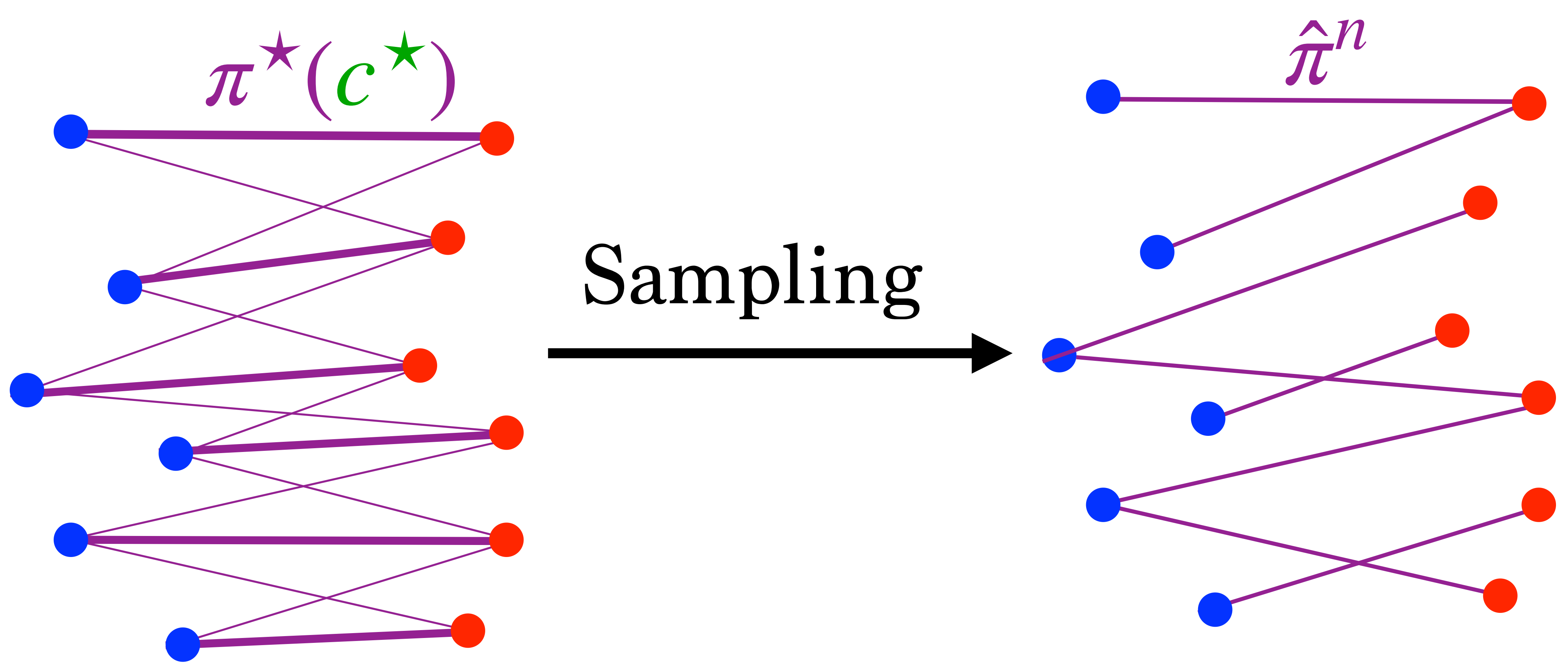
Assume $c_\theta(x,y) = \sum_k \theta_k c_k(x,y)$ are centred:

$$\int c_k(x,y) d\alpha(x) = 0, \quad \int c_k(x,y) d\beta(y) = 0$$

and linearly independent.

Assume that α, β have compact supports.

Sample complexity for iUOT



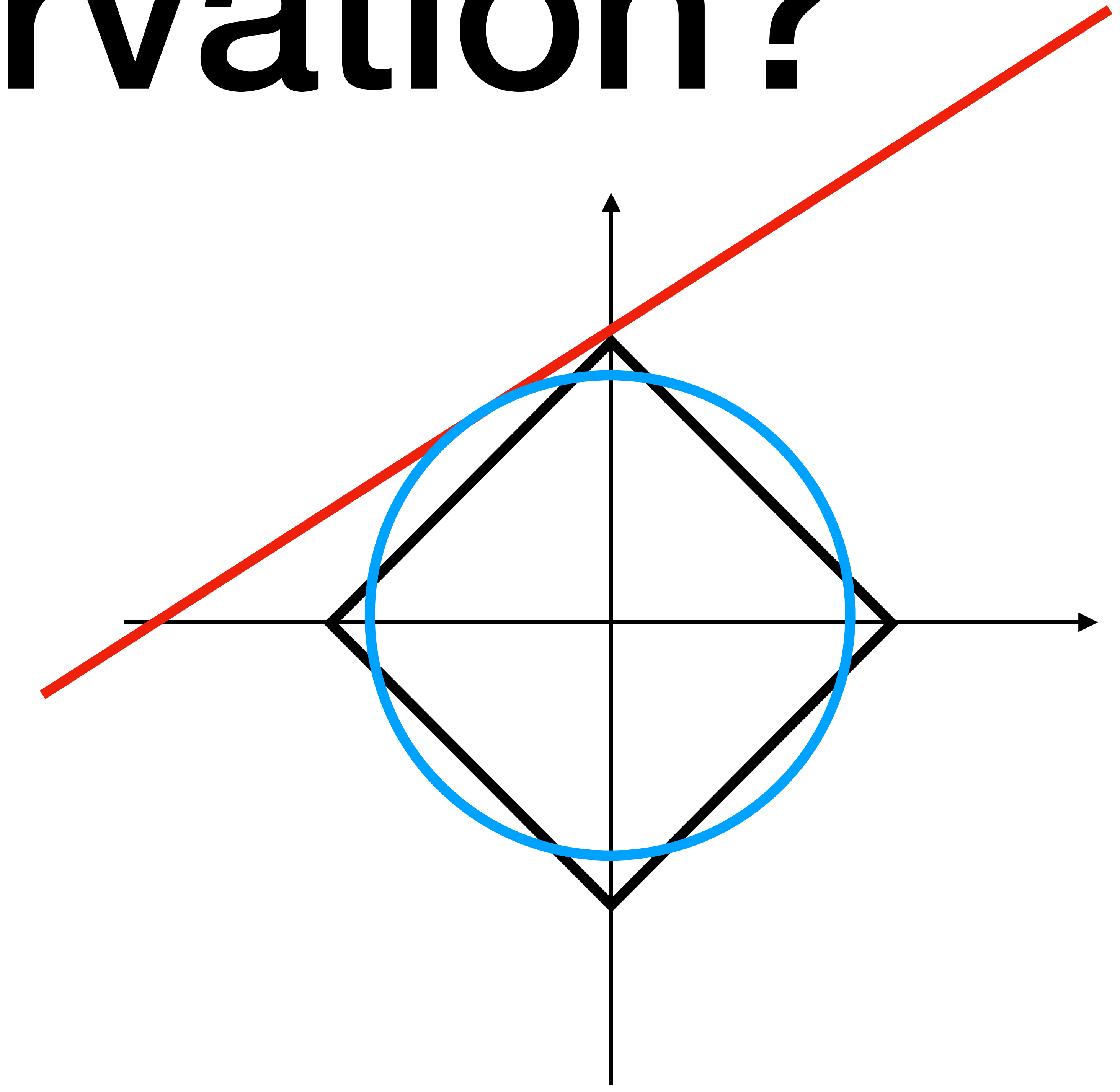
How accurate is $c_{\theta^{n,\lambda}}$ constructed from $\hat{\pi}^n$?

Theorem:
Let $c^\star = \Phi\theta^\star$ be the cost that gave rise to $\pi^\star, \alpha^\star, \beta^\star$. Let R be a convex, l.s.c. regularizer. Given $(x_i, y_i) \sim \pi^\star$ iid with $i = 1,\ldots,n$, the solution to $\theta = \operatorname{argmin}_\theta L(\theta; \hat{\pi}_n, \hat{\alpha}_n, \hat{\beta}_n) + \lambda R(\theta)$ is unique and with probability at least $1 - e^{-t}$,

$$\|\theta - \theta^\star\| = \mathcal{O}\left(\lambda + \frac{\sqrt{\log(n) + t}}{\sqrt{n}}\right)$$

Structure preservation?

Typical regularisers:



Sparsity. $R(\theta) = \sum_{i=1}^{n} |\theta_i|$

Low rank $R(\theta) = \sum_{i=1}^{n} \sigma_i(\theta)$ where $\sigma_i(\theta)$ = singular values of θ.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta; \hat{\pi}_n, \hat{\alpha}_n, \hat{\beta}_n) + \lambda R(\theta)$$

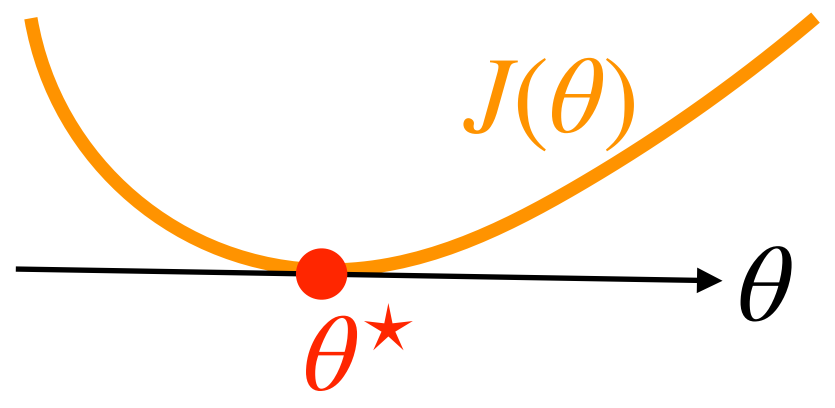
***Question**:*

If $\theta^\star$ that generated $\pi^\star, \alpha^\star, \beta^\star$ is sparse/low rank, is the solution $\hat{\theta}$ also of the same sparsity/rank when n is large enough and λ is small enough?

Structure preservation

Certificate: Let $M := \nabla^2 J(\theta^\star)$, and define

- $\hat{z}_1 = M_{(:,I)} M_{(I,I)}^{-1} \mathrm{Sign}(\theta_I^\star)$ in the case of l1 where I is the support of $\theta^\star$
- $\hat{z}_* = M P_T (P_T M P_T)^{-1} \mathrm{Sign}(\theta_I^\star)$ where P_T is the projection onto the row/column space spanned by $\theta^\star$.

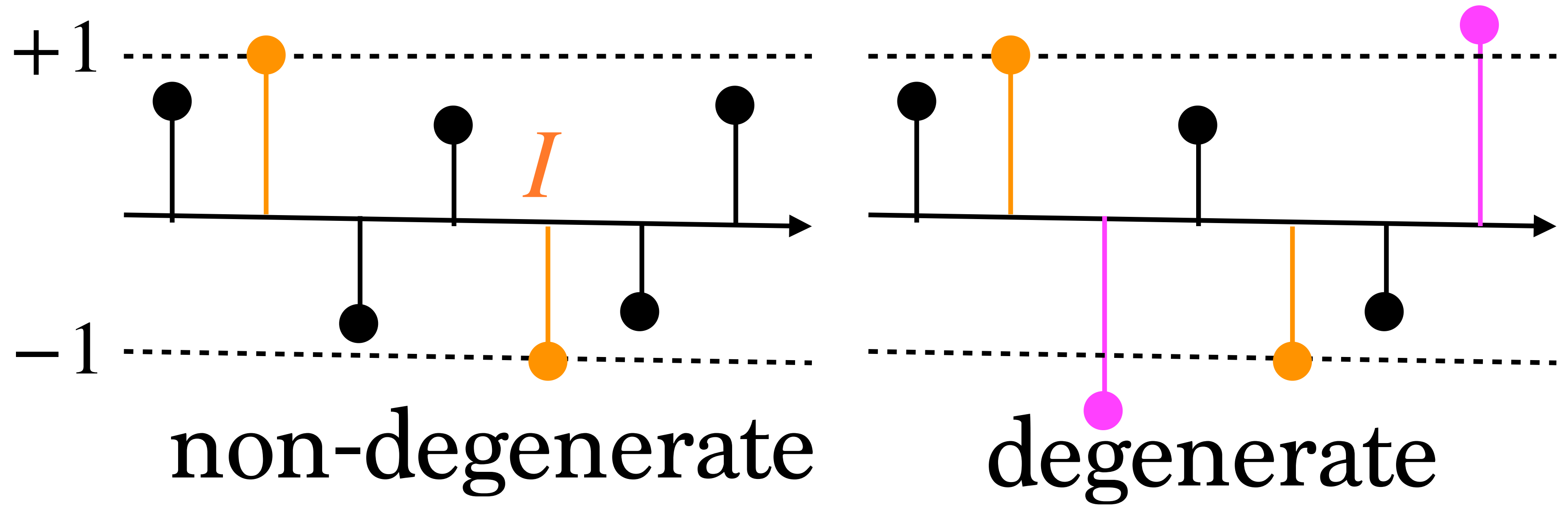


$$J(\theta) := L(c_\theta \,|\, \pi(c_{\theta^\star}))$$

We say that

$\hat{z}_1$ is *non degenerate* if $\|\hat{z}^1_{I^c}\|_\infty < 1$

$\hat{z}^*$ is non degenerate if $\|P_T^\perp \hat{z}^*\|_2 < 1$.



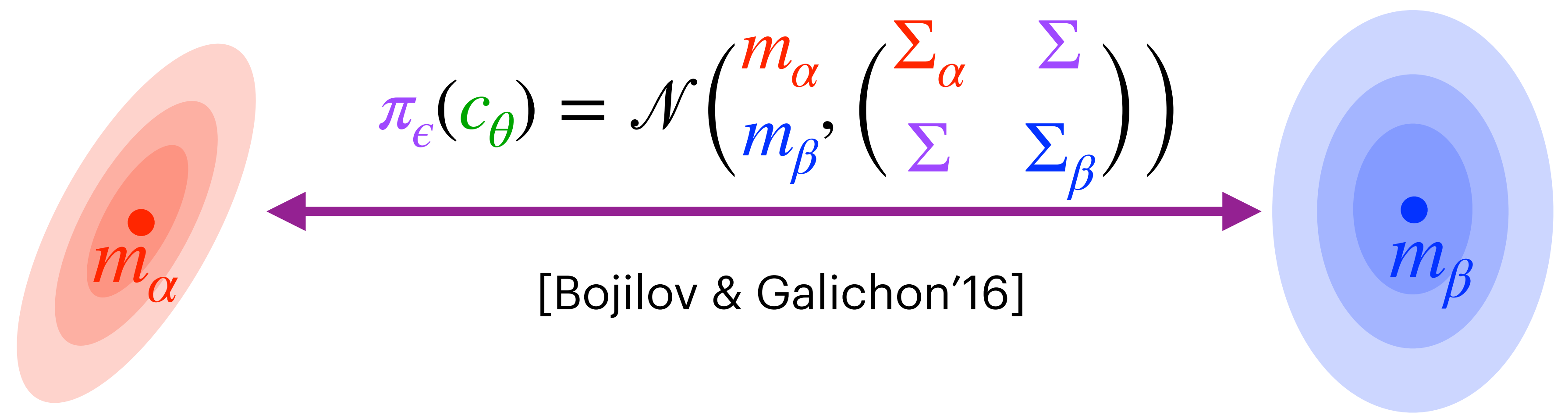
Theorem: Suppose that $\hat{z}$ is nondegenerate.

Then, provided that $1 \gtrsim \lambda \gtrsim \sqrt{\frac{t + \log(n)}{n}}$, with probability at least $1 - e^{-t}$,

- $\mathrm{Supp}(\theta_{n,\lambda}) = \mathrm{Supp}(\theta^\star)$ in the case of $R(\theta) = \|\theta\|_1$
- $\mathrm{Rank}(\theta_{n,\lambda}) = \mathrm{Rank}(\theta^\star)$ in the case of $R(\theta) = \|\theta\|_*$.

The iOT loss in the Gaussian setting

$\alpha = \mathcal{N}(m_\alpha, \Sigma_\alpha)$ $\qquad$ $\beta = \mathcal{N}(m_\beta, \Sigma_\beta)$



Proposition:

$$\partial^2 J(\theta^\star) = 2\varepsilon\left[4\varepsilon^2(\Sigma_\beta - \Sigma^T\Sigma_\alpha\Sigma)^{-1} \otimes (\Sigma_\alpha - \Sigma\Sigma_\beta{}^{-1}\Sigma^\top)^{-1} + (\theta^{\star\top} \otimes \theta^\star)\right]^{-1}$$

→ Numerically check when $\eta^\star$ is non-degenerated.

$$\min_\theta L(c_\theta \,|\, \hat{\pi}) + \lambda\|\theta\|_1$$

$\lambda = \lambda_0 \varepsilon$, $\varepsilon \to 0$

$$\min_{\theta \succ 0} \frac{1}{2}\log\det(\theta) + \frac{1}{2}\langle\theta, \hat{\theta}^{-1}\rangle + \lambda_0\|\theta\|_1$$

Graphical-Lasso

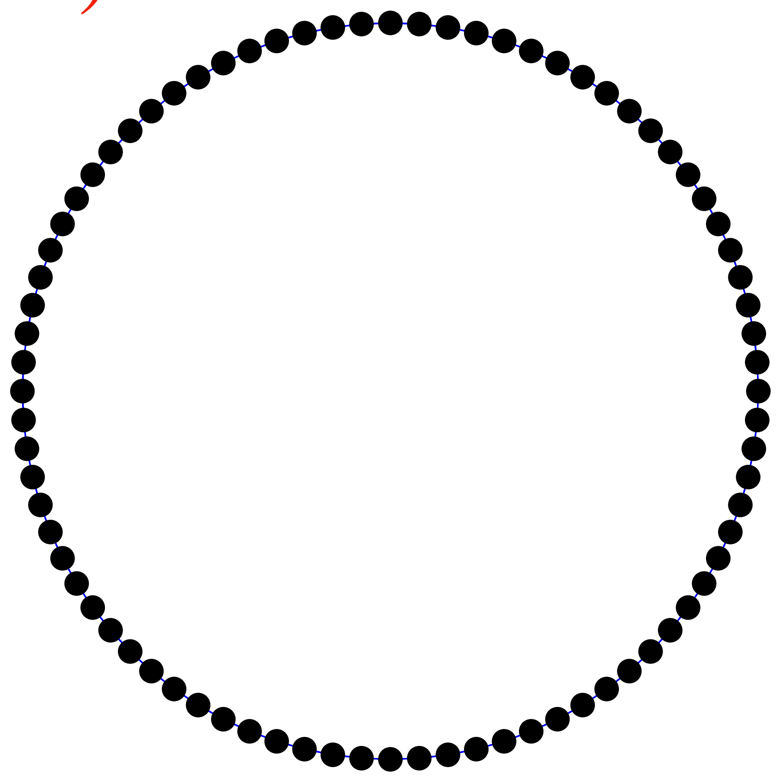
$\lambda = \lambda_0/\varepsilon$, $\varepsilon \to +\infty$

$$\min_\theta \frac{1}{2}\|(\Sigma_\beta^{\frac{1}{2}} \otimes \Sigma_\alpha^{\frac{1}{2}})(\theta - \hat{\theta})\|_F^2 + \lambda_0\|\theta\|_1$$

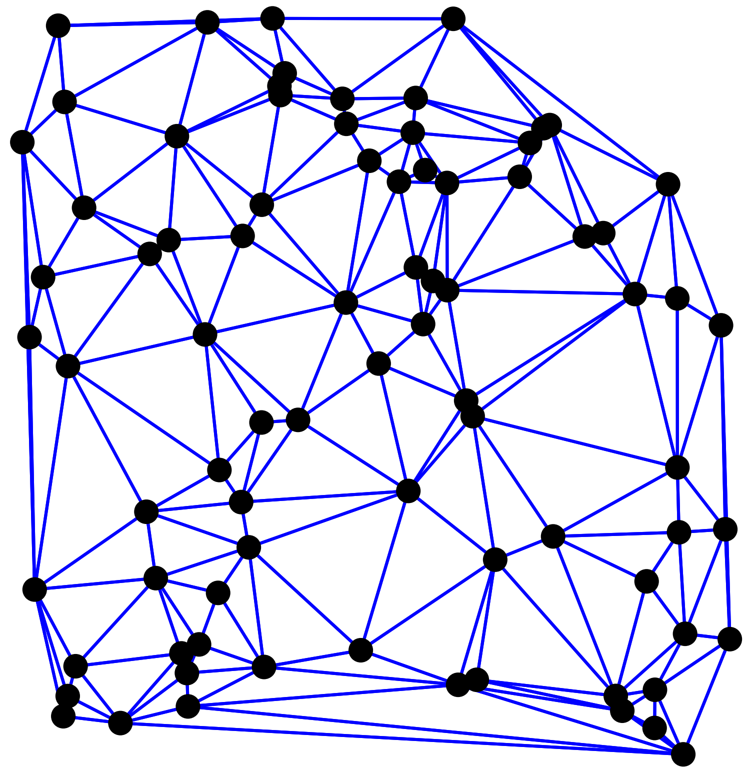
Lasso

Numerical illustrations of ℓ_1 certificates

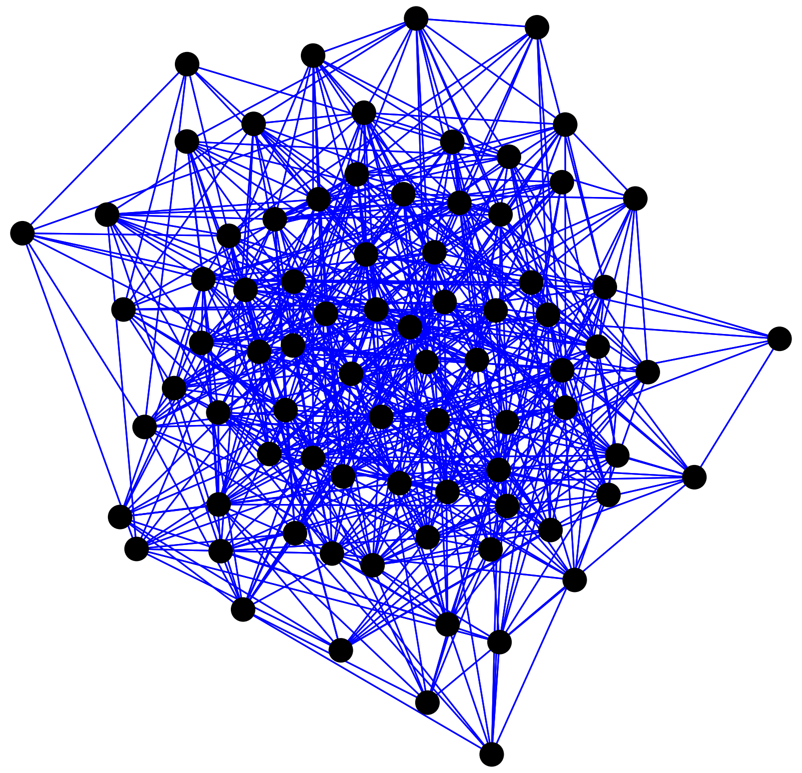
$\theta^\star = \delta I + \mathrm{diag}(G1) - G$



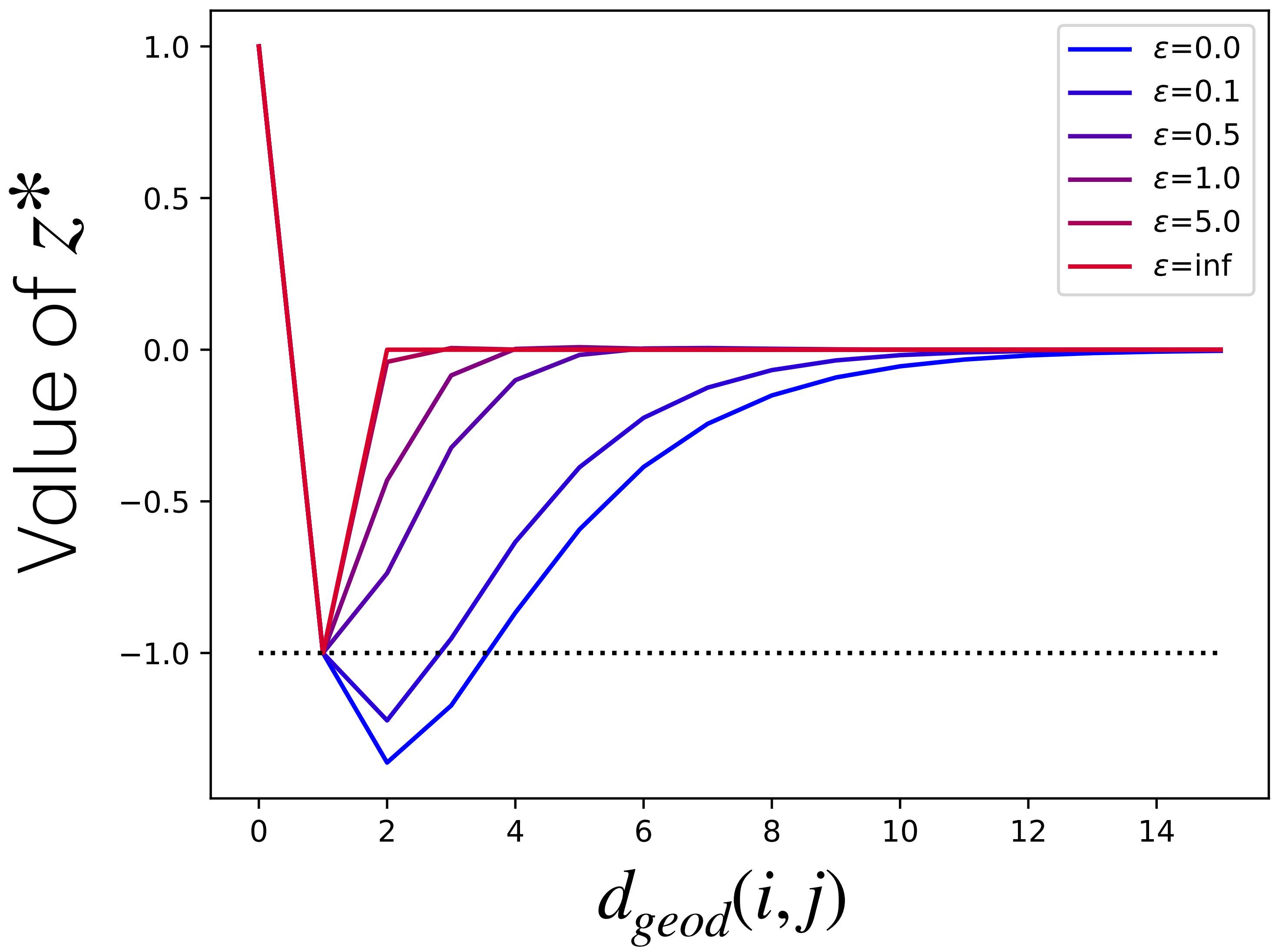
Circular

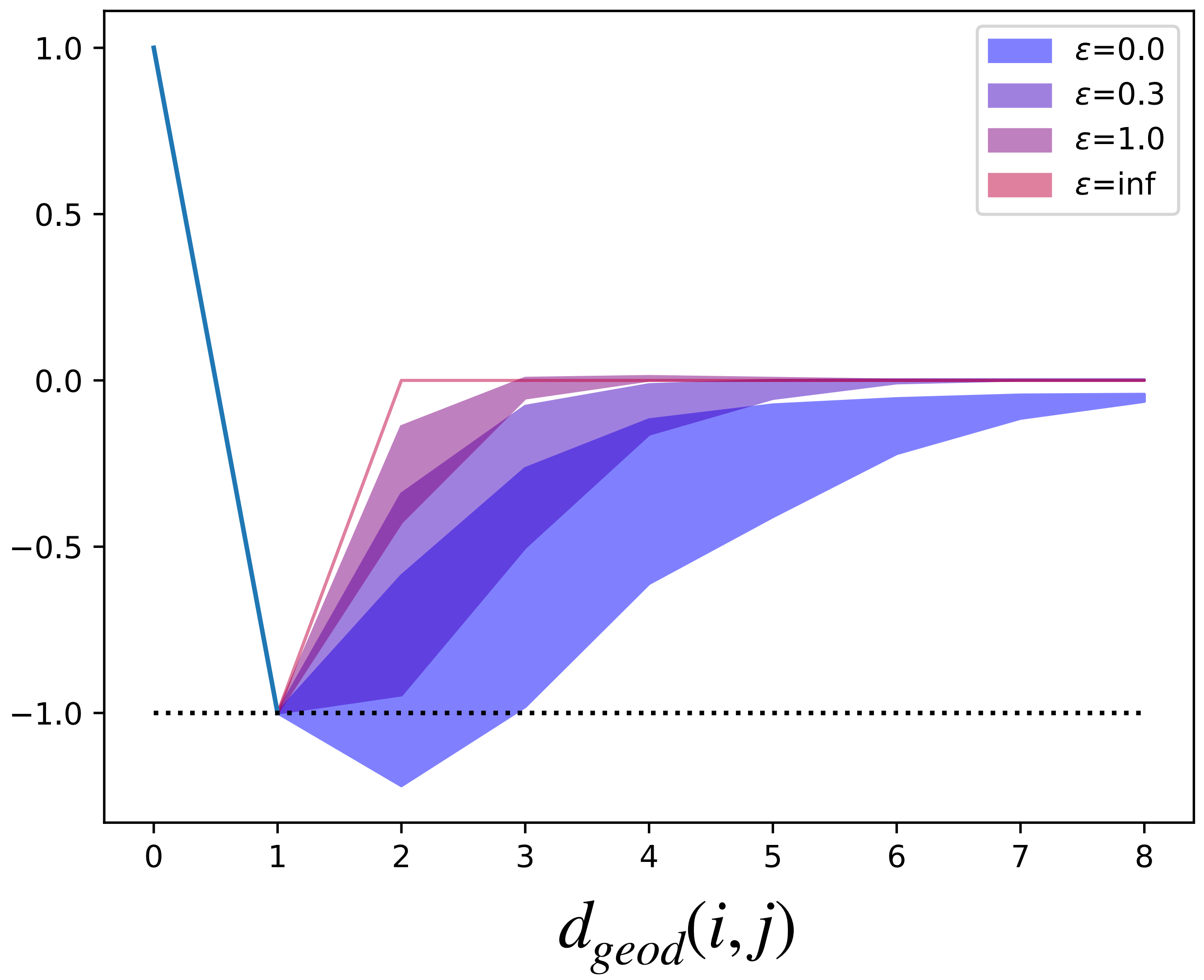


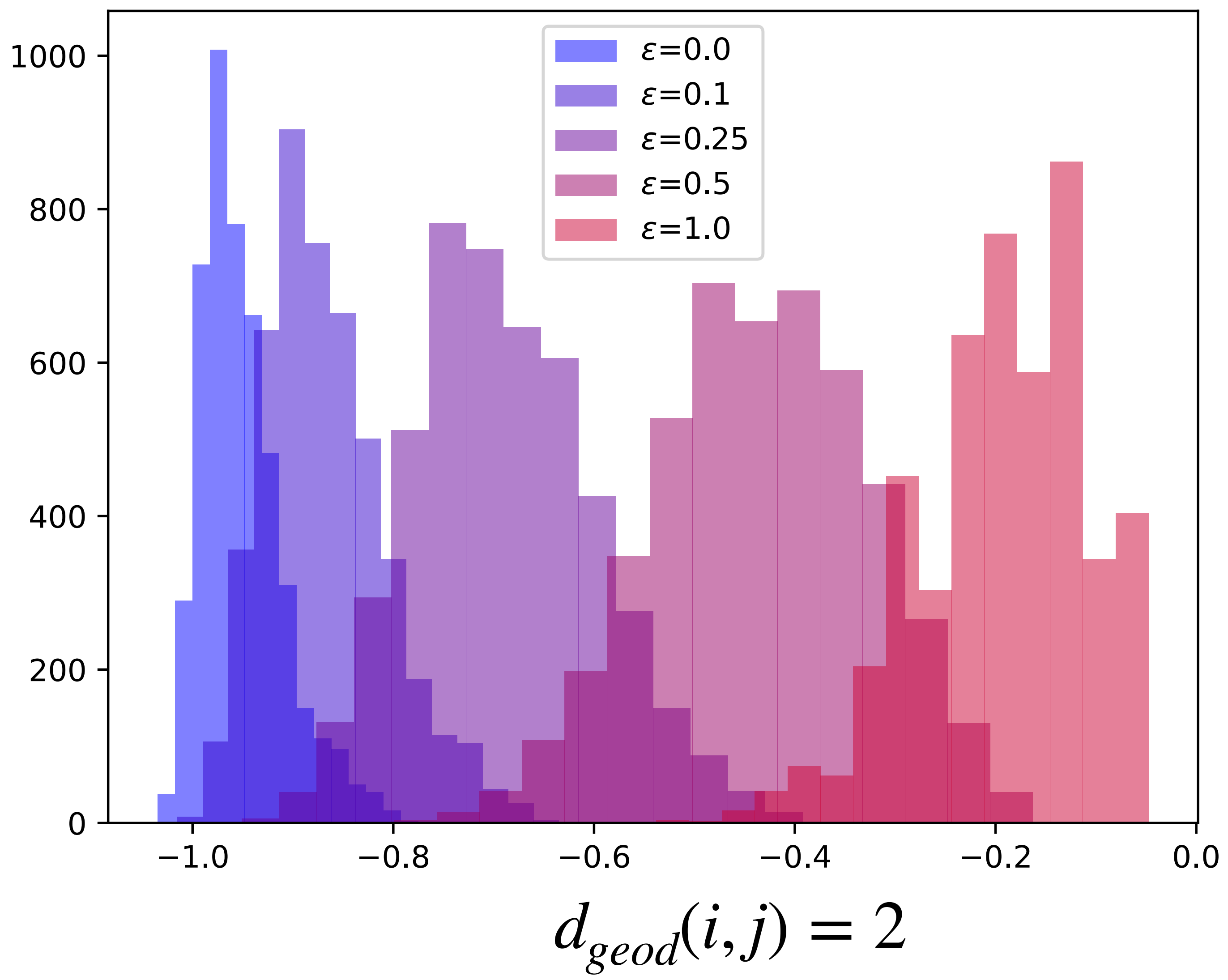
Planar



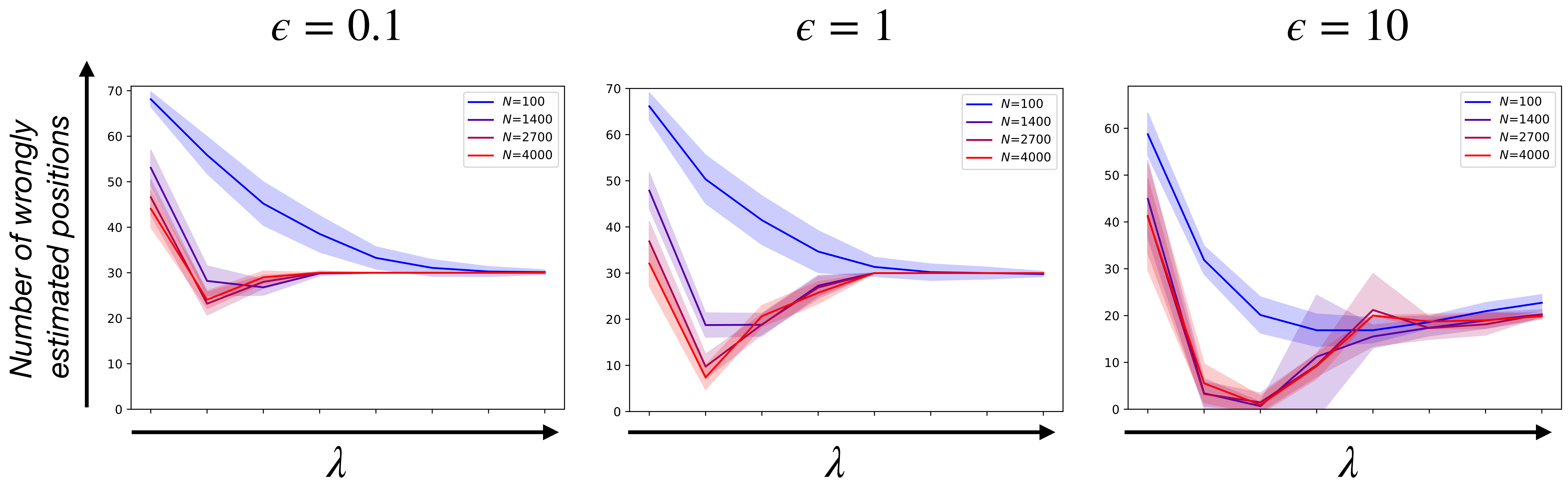
Erdös-Rényi







Numerical illustrations



Recovery performance of ℓ_1-iOT for a circular graph.

The iJKO loss

$$L_r(\theta) := \langle V_\theta, \alpha^{k+1} \rangle - \inf_{\alpha \in \mathscr{P}(\mathscr{X})} \langle V_\theta, \alpha \rangle + \frac{1}{\tau} W^2_{2,\epsilon}(\alpha, \alpha^k \,|\, \alpha \otimes \alpha^k) + r\mathrm{KL}(\alpha \,|\, \alpha^{k+1}) \,.$$

$$\theta^s = \operatorname*{argmin}_\theta \frac{\lambda}{r} R(\theta) + L_r(\theta) \quad \xrightarrow{r \to \infty} \quad \operatorname*{argmin}_\theta \lambda R(\theta) + \mathrm{Var}_{\alpha^{k+1}} \left[V_\theta + \tau^{-1} f^*(\alpha^{k+1}, \alpha^k) \right]$$

Gaussian experiment

Consider $\frac{d}{dt} X_t = -\nabla V(X_t)$ where $X_0 \sim \mathcal{N}(\mathbf{m}^\star, \Sigma^\star)$ and $V(x) = x^\top \theta^\star x$.

Then, $\alpha_t := \mathrm{law}(X_t) = \mathcal{N}(\mathbf{m}_t, \Sigma_t)$ with $\mathbf{m}_t = e^{-2t\theta^\star} \mathbf{m}^\star$ and $\Sigma_t = e^{-2t\theta^\star} \Sigma^\star e^{-2t\theta^\star}$.

Suppose we observe samples of α_t at discrete time point.

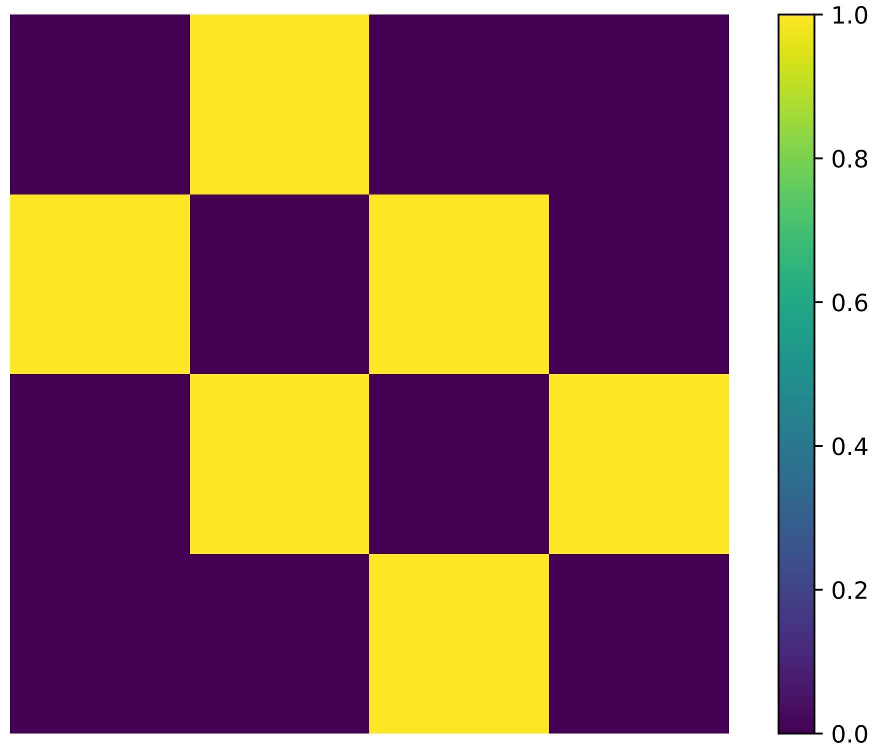
c.f. loss $\|\nabla V_\theta + \tau^{-1} \nabla f^*(\alpha^{k+1}, \alpha^k)\|^2_{L^2(\alpha^{k+1})}$ of Terpin, Lanzetti, Gadea, and Dörfler. *Learning diffusion at lightspeed*

Question: what kind of $\theta^\star$ are easy to recover?

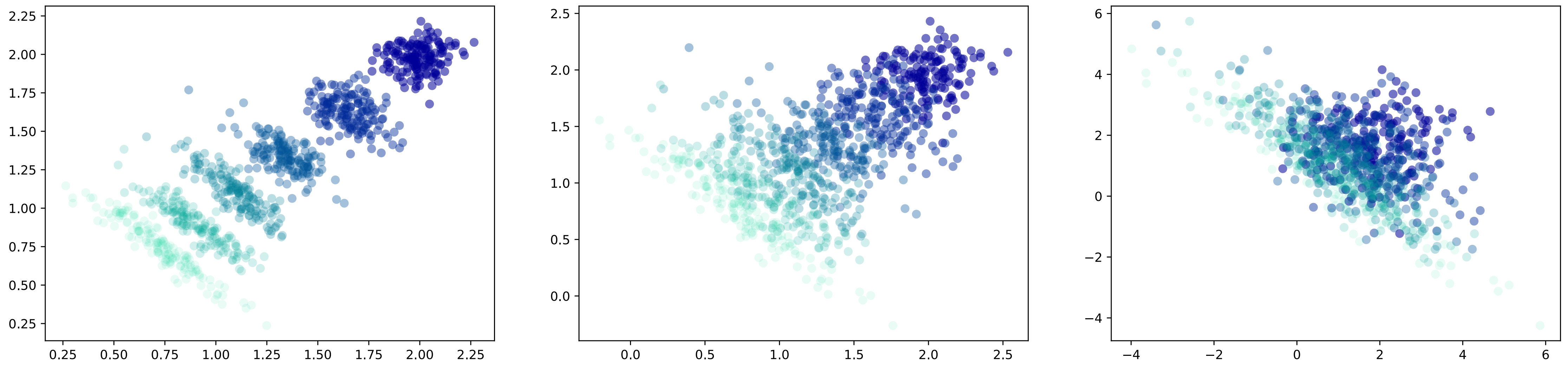
Investigate using the closed form certificate at $r = \infty$

The sparse setting

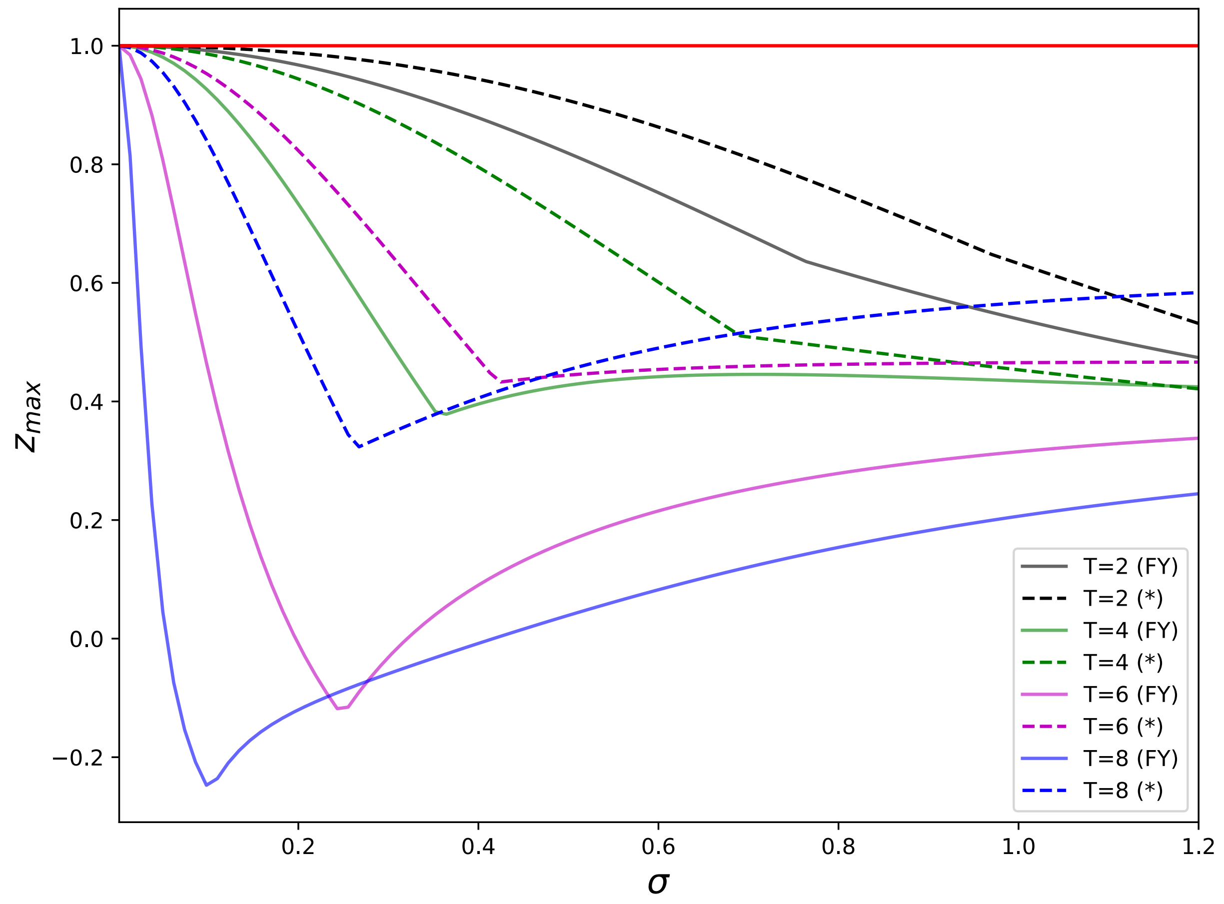
$\theta^\star =$



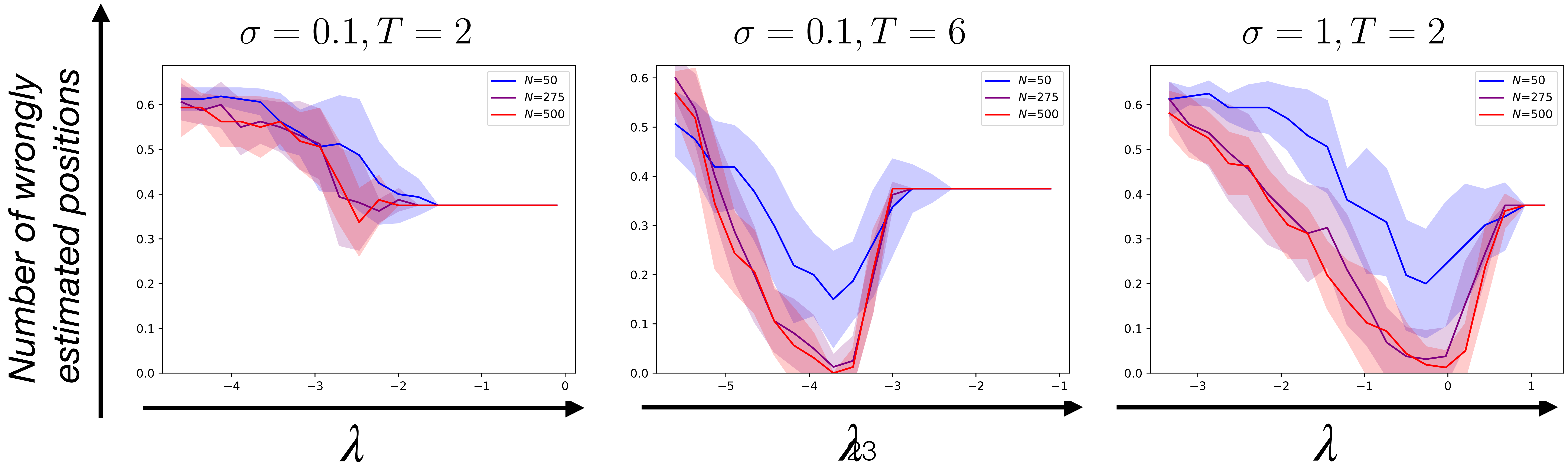
Evolution across 6 time points with $\tau = 0.1$, $m^\star = 2 \cdot \mathbf{1}$ and $\Sigma^\star = \sigma^2 \mathrm{Id}$.



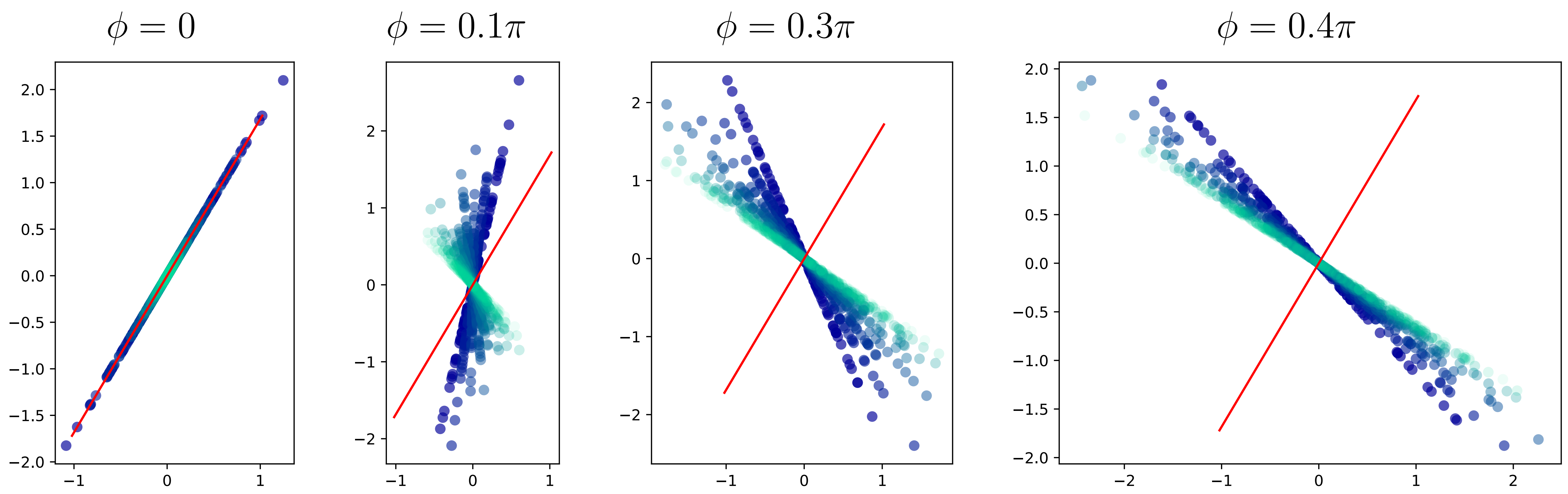
$\sigma = 0.1$ $\sigma = 0.2$ $\sigma = 1$

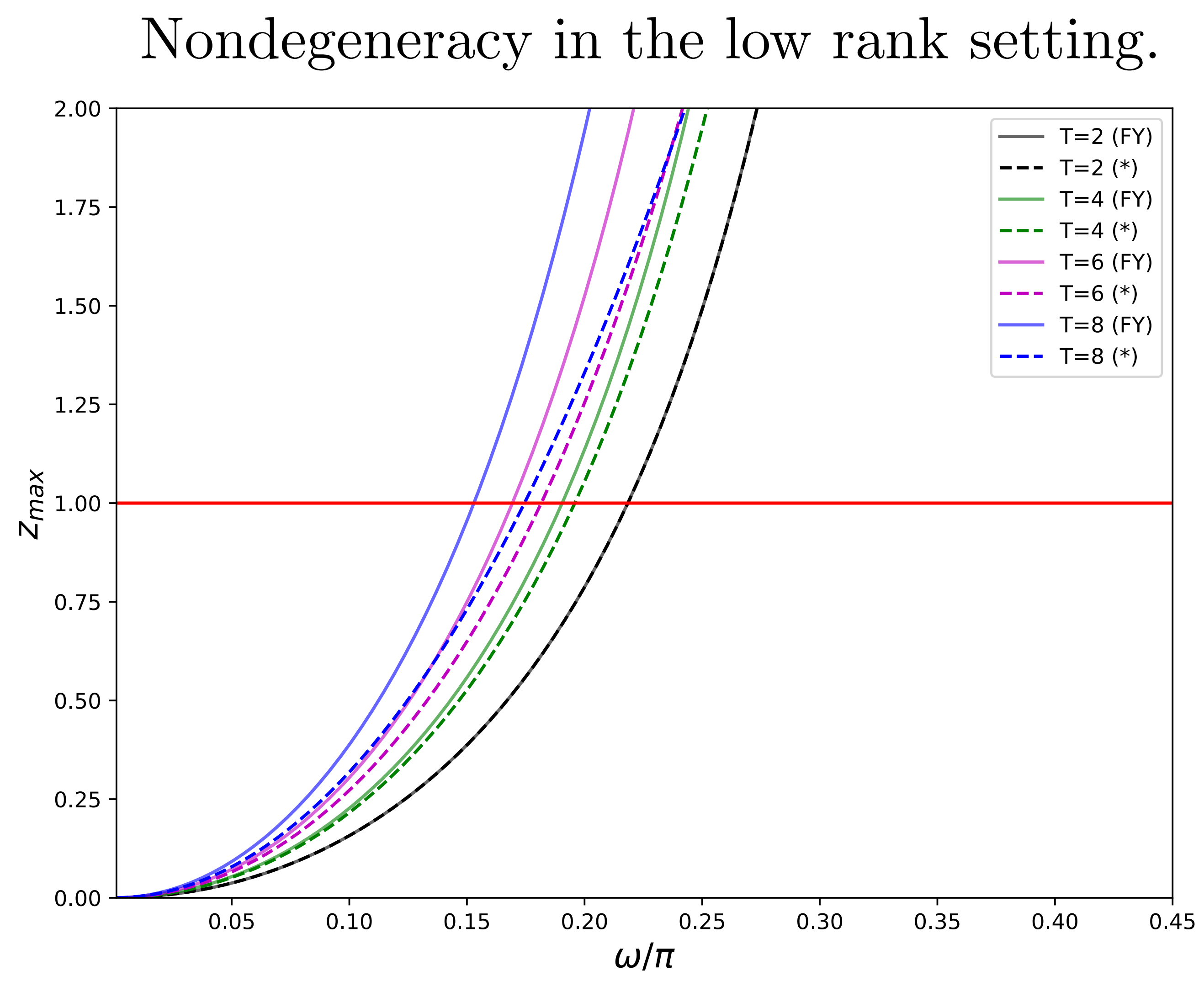


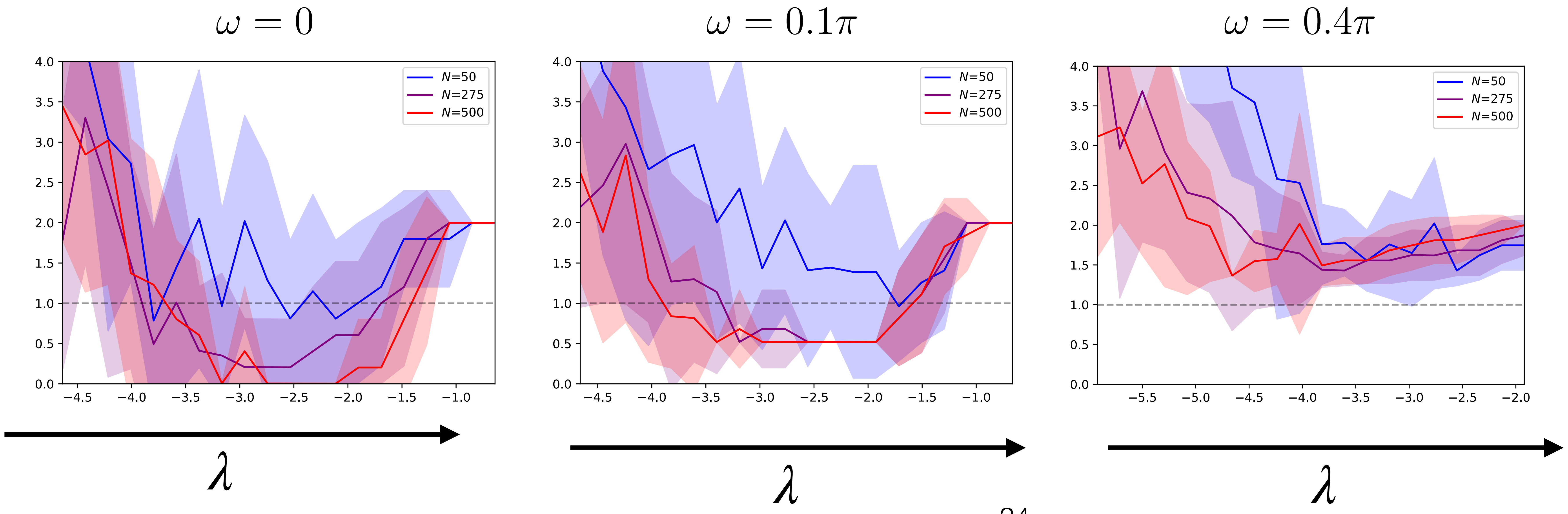
$\Sigma^\star = \sigma^2 I, \quad m^\star = 2 \cdot \mathbf{1}$



The low rank setting







$$\theta^\star = uu^\top$$

$$\Sigma^\star = \delta I + u_\phi u_\phi^\top \text{ with}$$

$$u_\phi = \cos(\phi)u + \sin(\phi)u^\perp$$

Summary

- Optimal transport computes a coupling given two distributions and a cost metric.
- In some applications, the metric is unknown; or we might be interested in recovering certain dynamics given observations of probability distributions.
- Fenchel-Young losses allow us to construct convex losses to handle these inverse problems with probability measures.
- We derived a theoretical analysis of the sample complexity, and structural properties of regularization.

Sparsistency for inverse optimal transport, ICLR 2024

Learning from Samples: Inverse Problems over measures via Sharpened Fenchel-Young Losses