

Regularized and Neural Approaches to Solve Optimal Transport

Marco Cuturi

Viña del Mar | Apple Inc. | 8 Jan. 2024

Outline of the Tutorial

Prelude **Warm-Up: Starting with Optimal Matchings**

- OT problems as extensions / generalizations of CS.101 optimal matchings

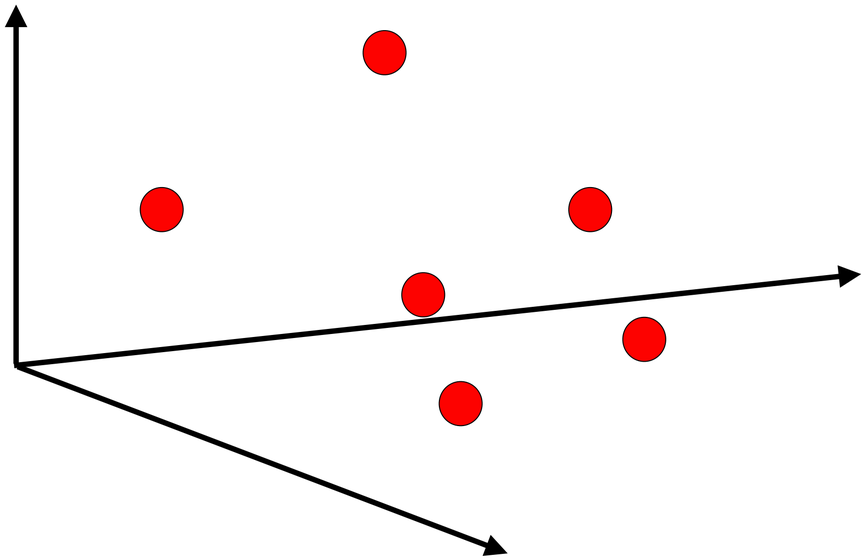
Part 1 Kantorovich Formulation of OT, Discrete Computations

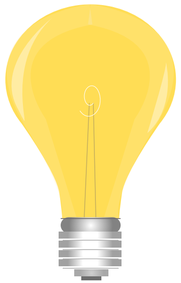
Part 2 Monge Formulation, Duality, Gangbo-McCann-Brenier Theorems

A Puzzle

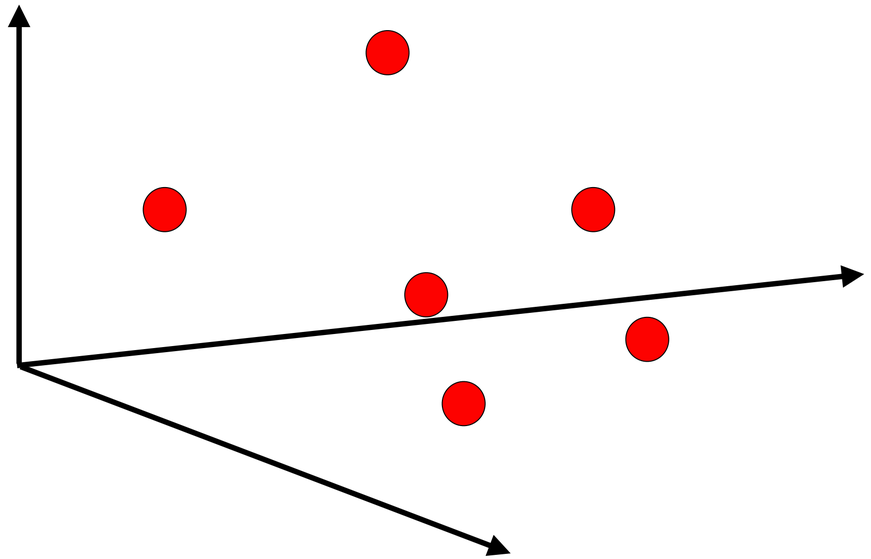


A Puzzle





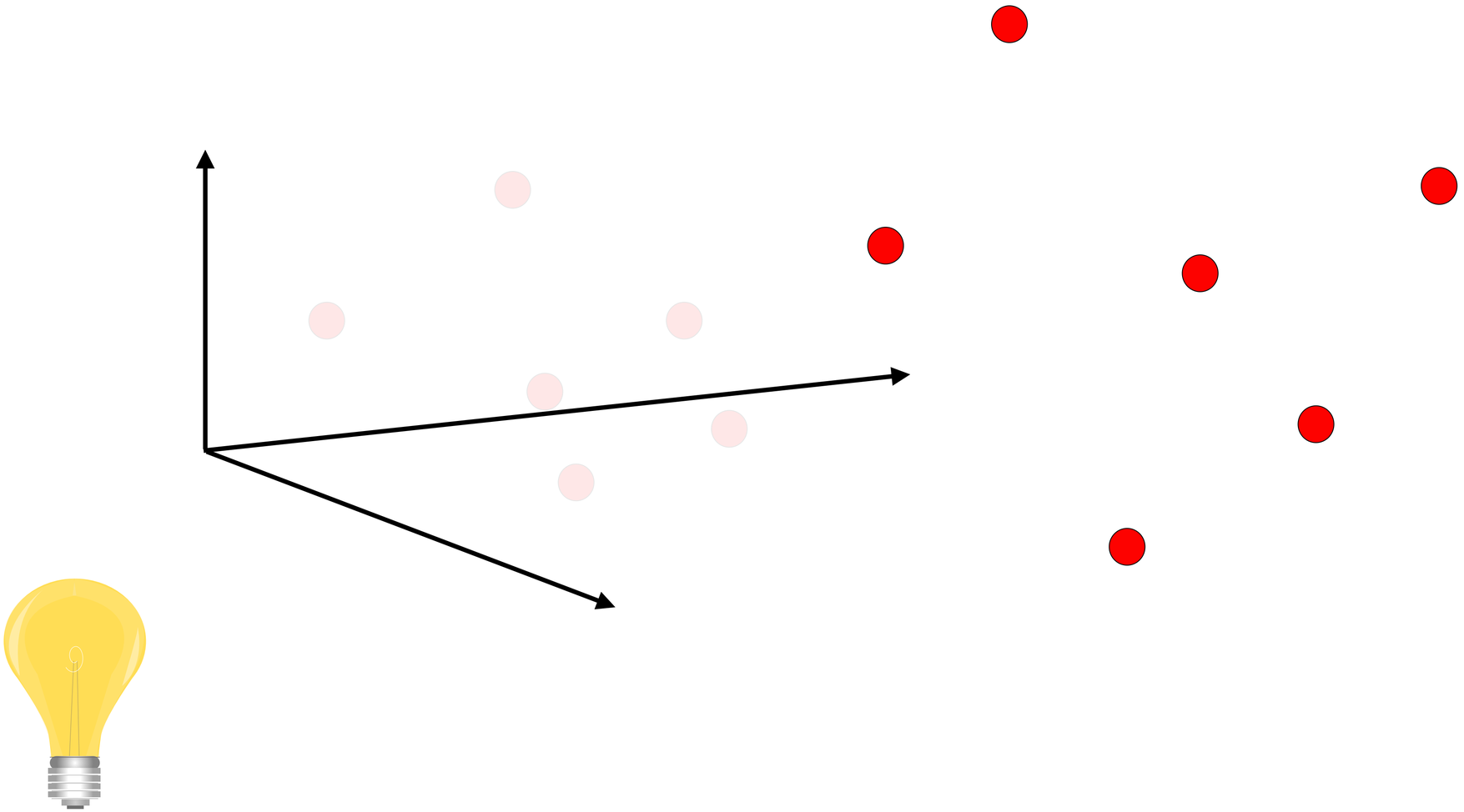
A Puzzle



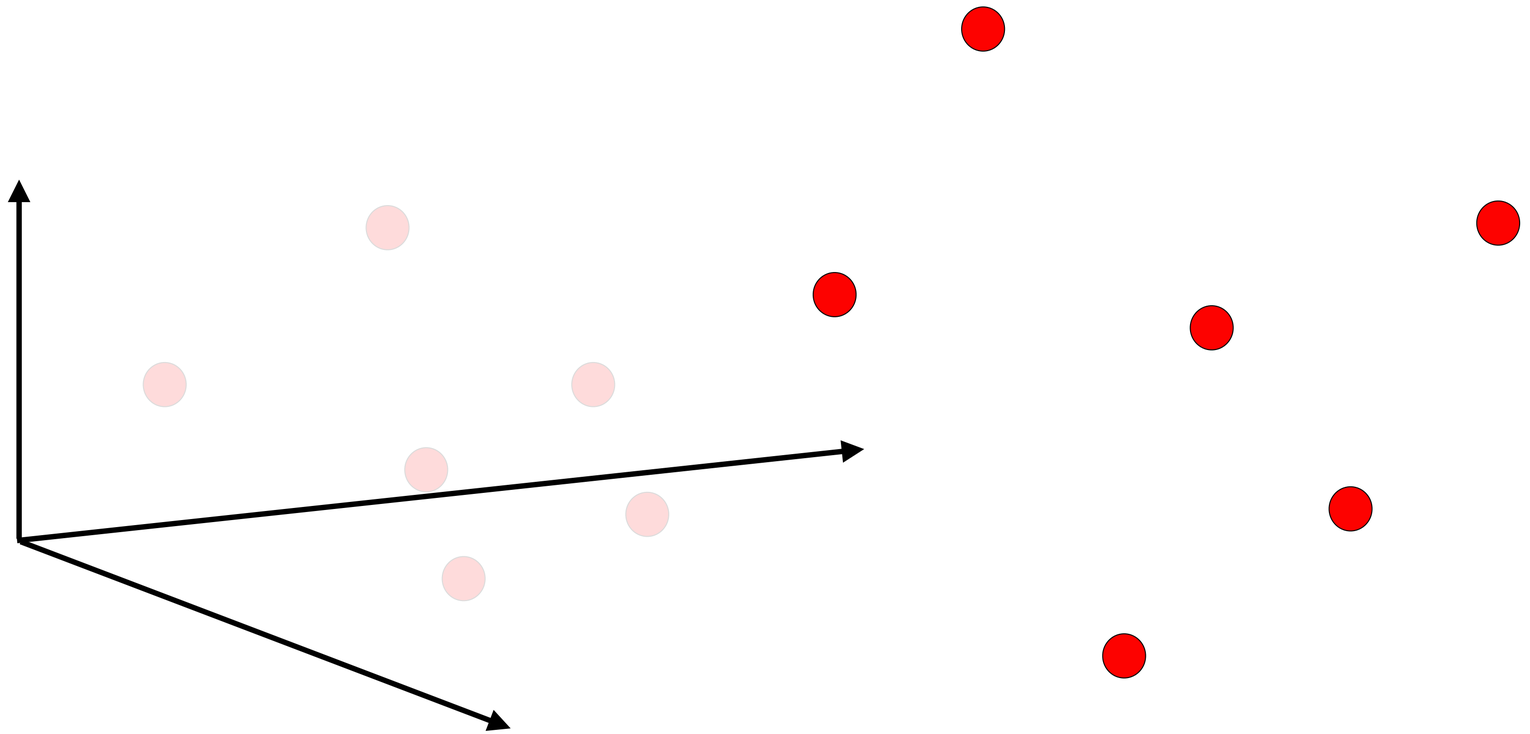
A Puzzle



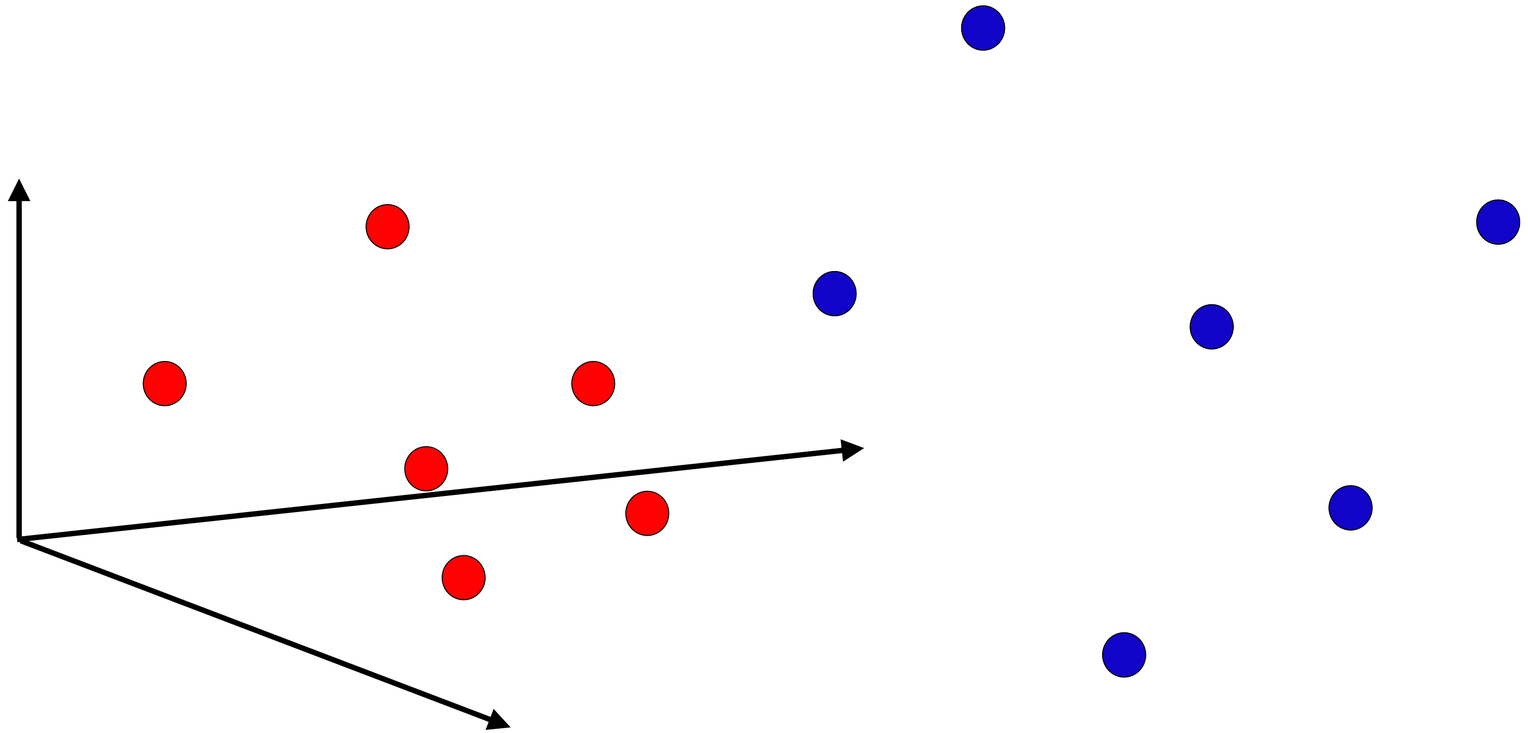
A Puzzle



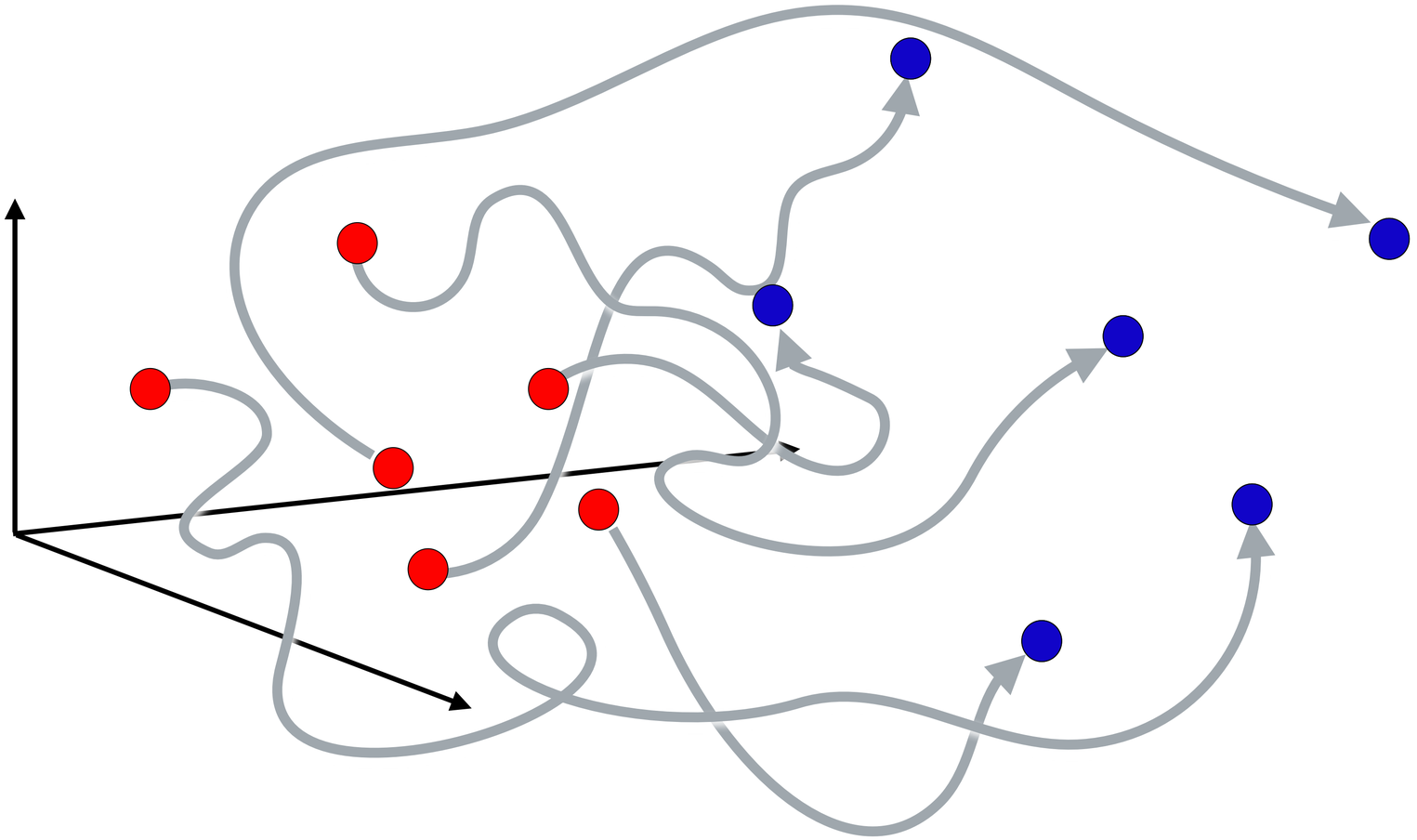
The Puzzle is: Who Went Were?



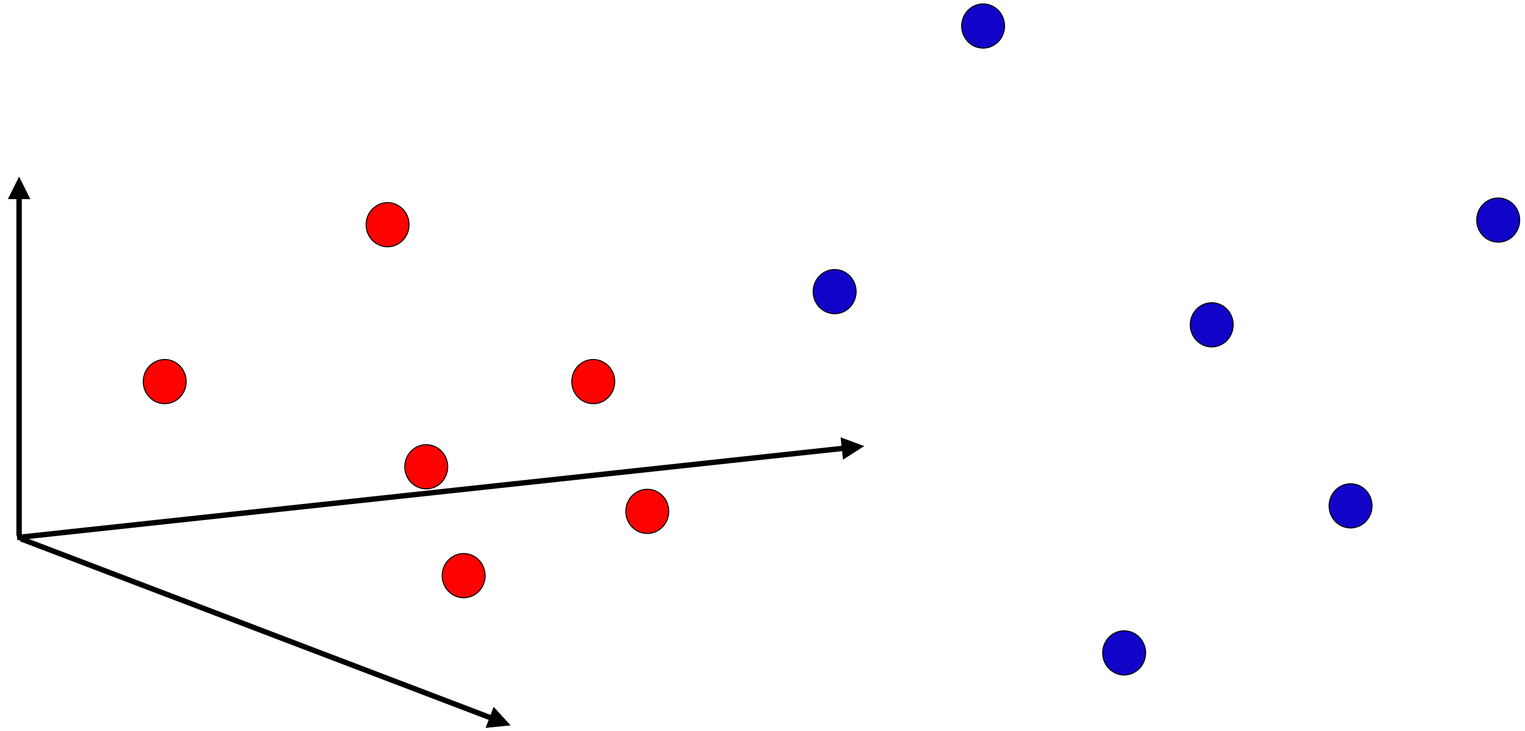
The Puzzle is: Who Went Were?



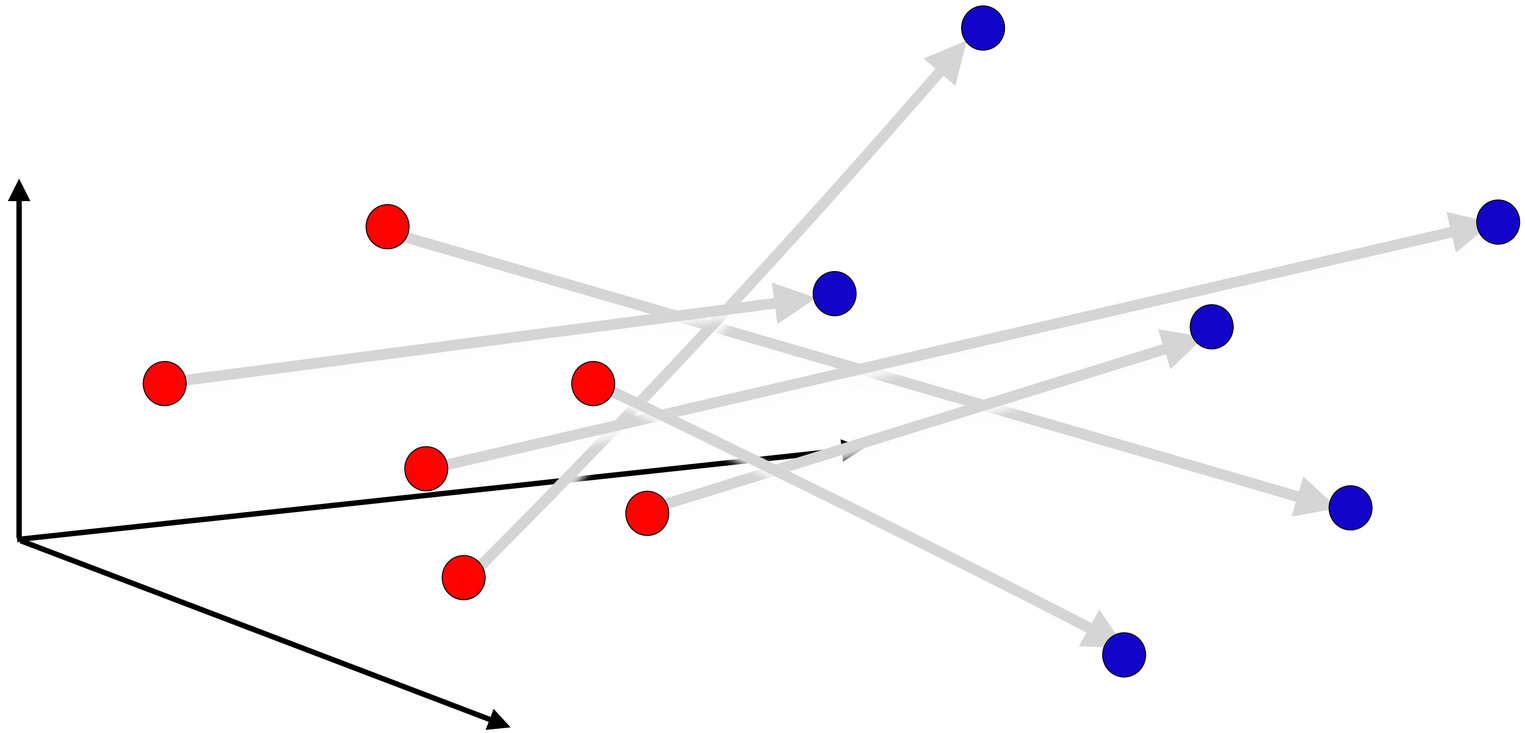
The Puzzle is: Who Went Were?



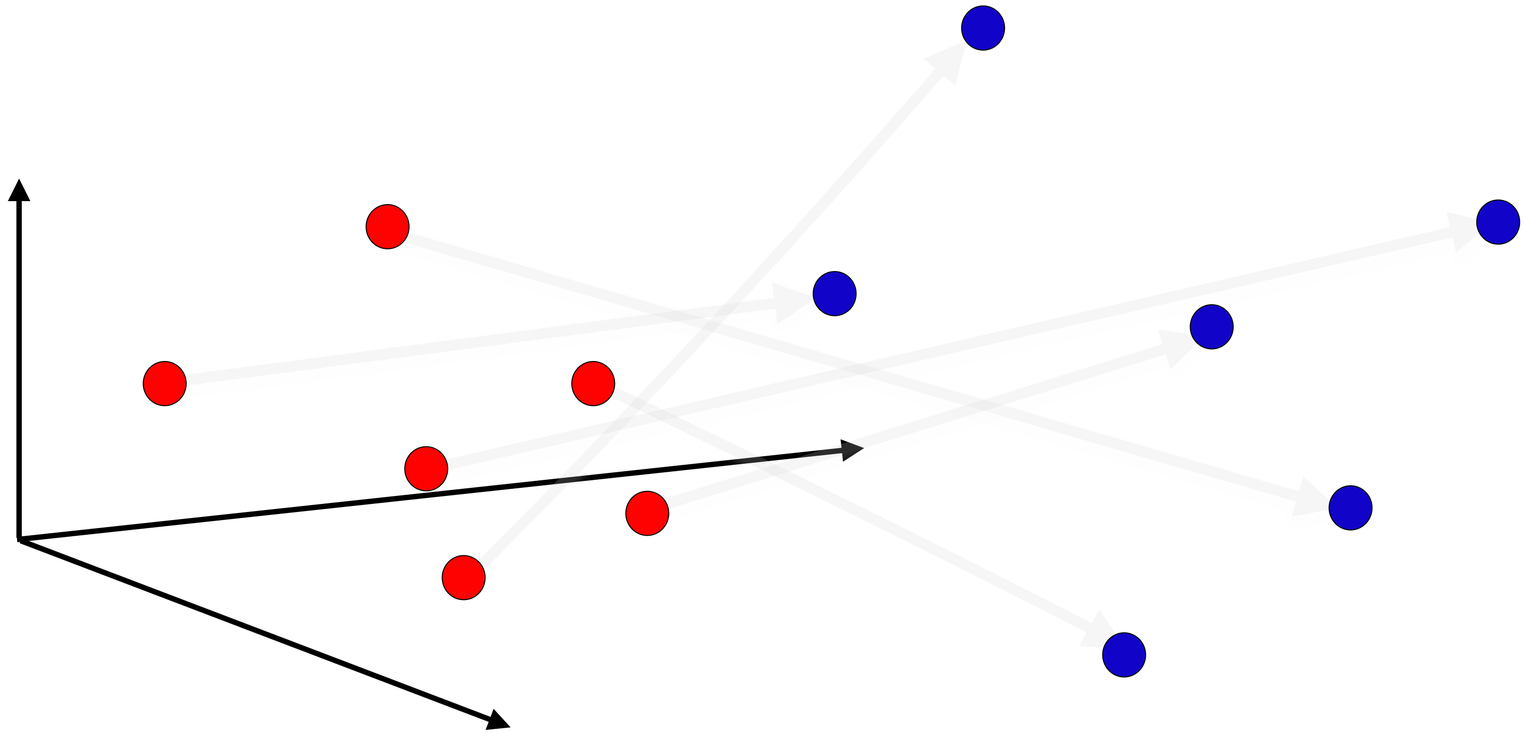
For Now, Assume Particles Moved Along Straight Lines



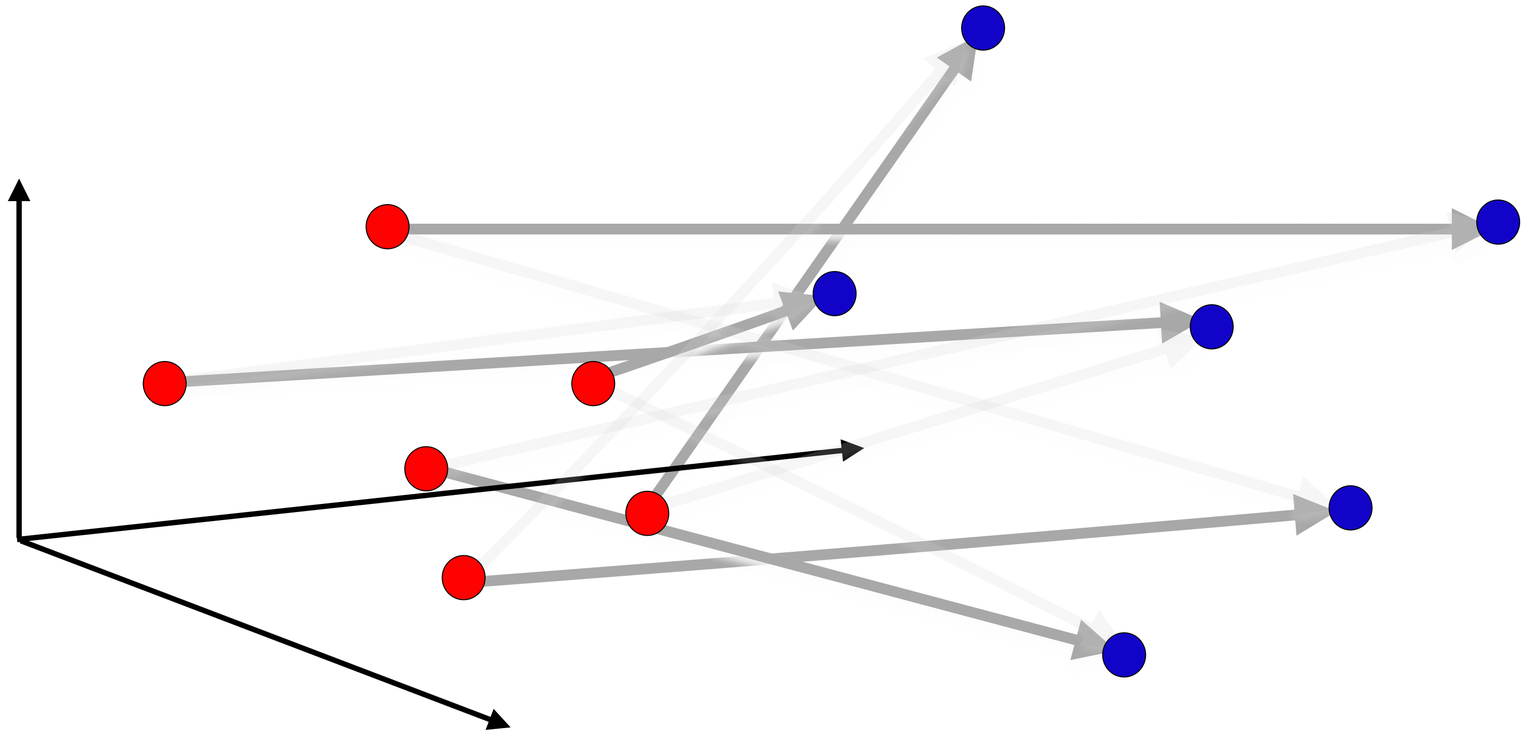
For Now, Assume Particles Moved Along Straight Lines



For Now, Assume Particles Moved Along Straight Lines

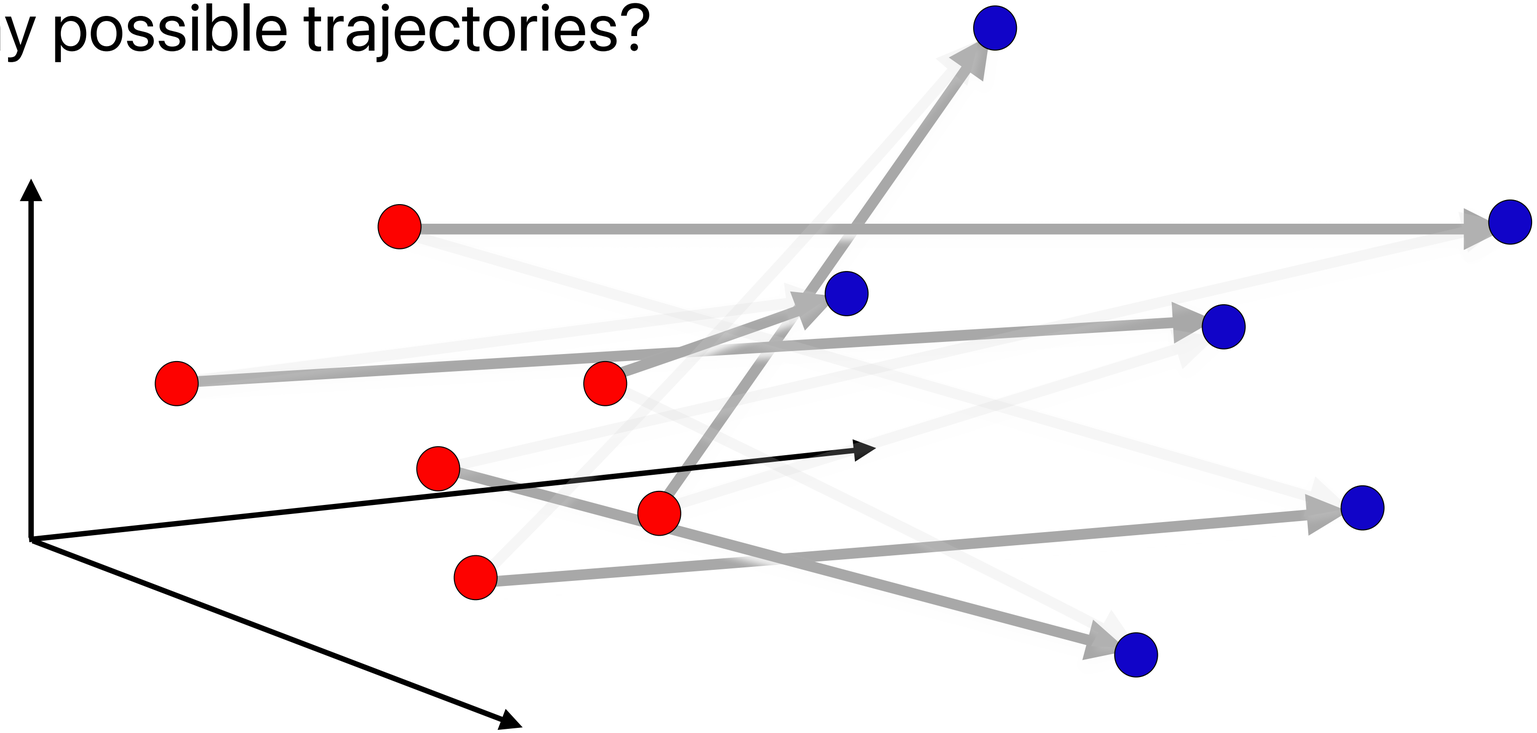


For Now, Assume Particles Moved Along Straight Lines



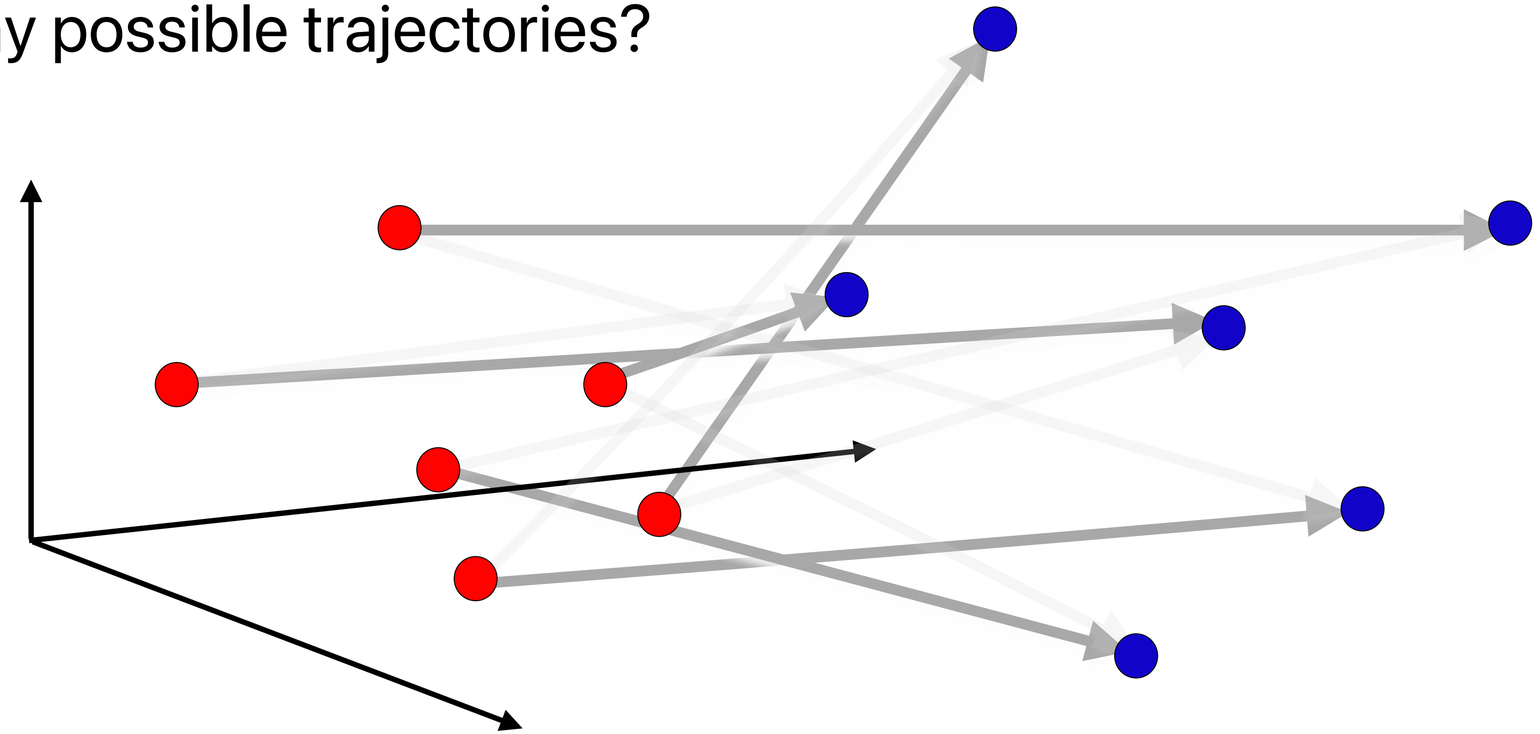
For Now, Assume Particles Moved Along Straight Lines

How many possible trajectories?



For Now, Assume Particles Moved Along Straight Lines

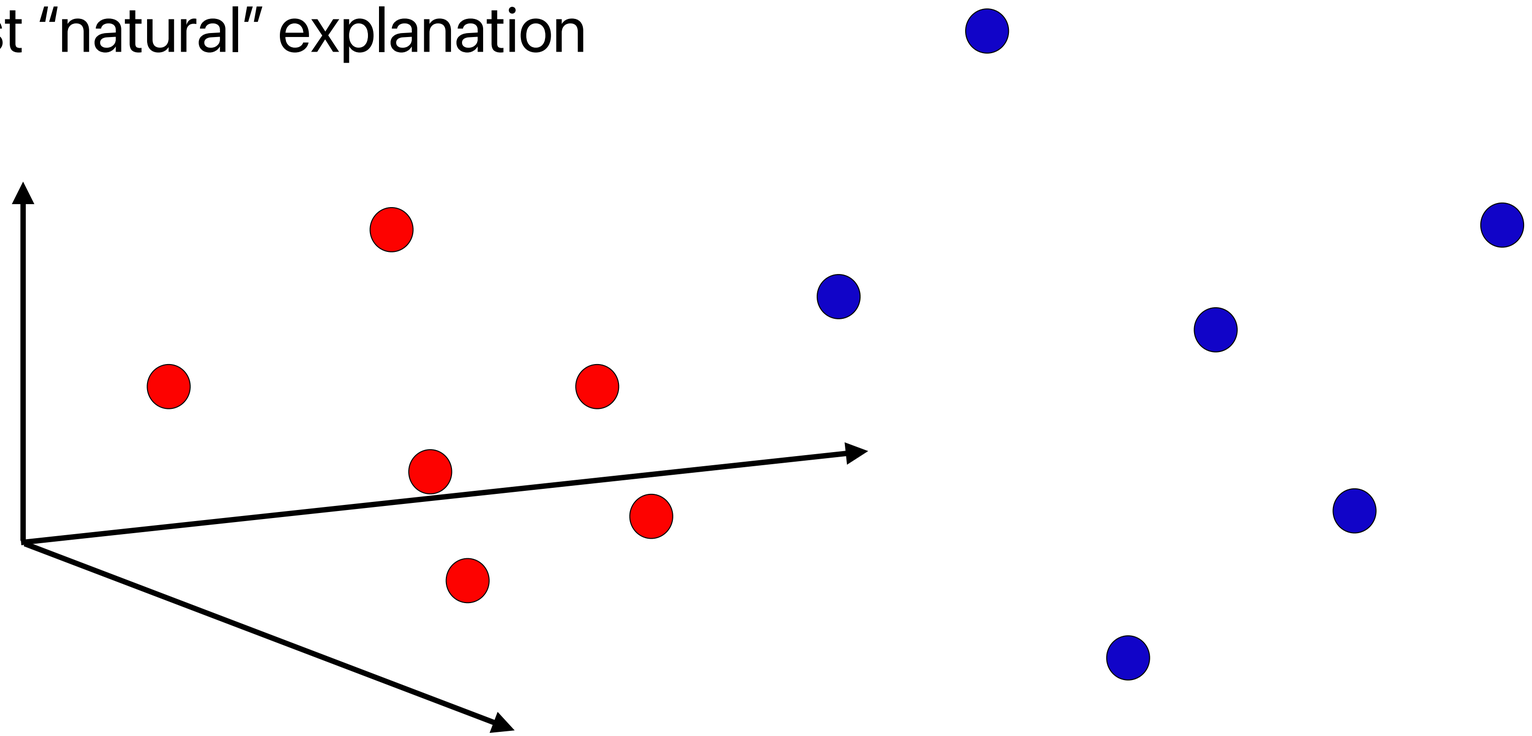
How many possible trajectories?



There are 6! = 6 x 5 x 4 x 3 x 2 = 720 permutations

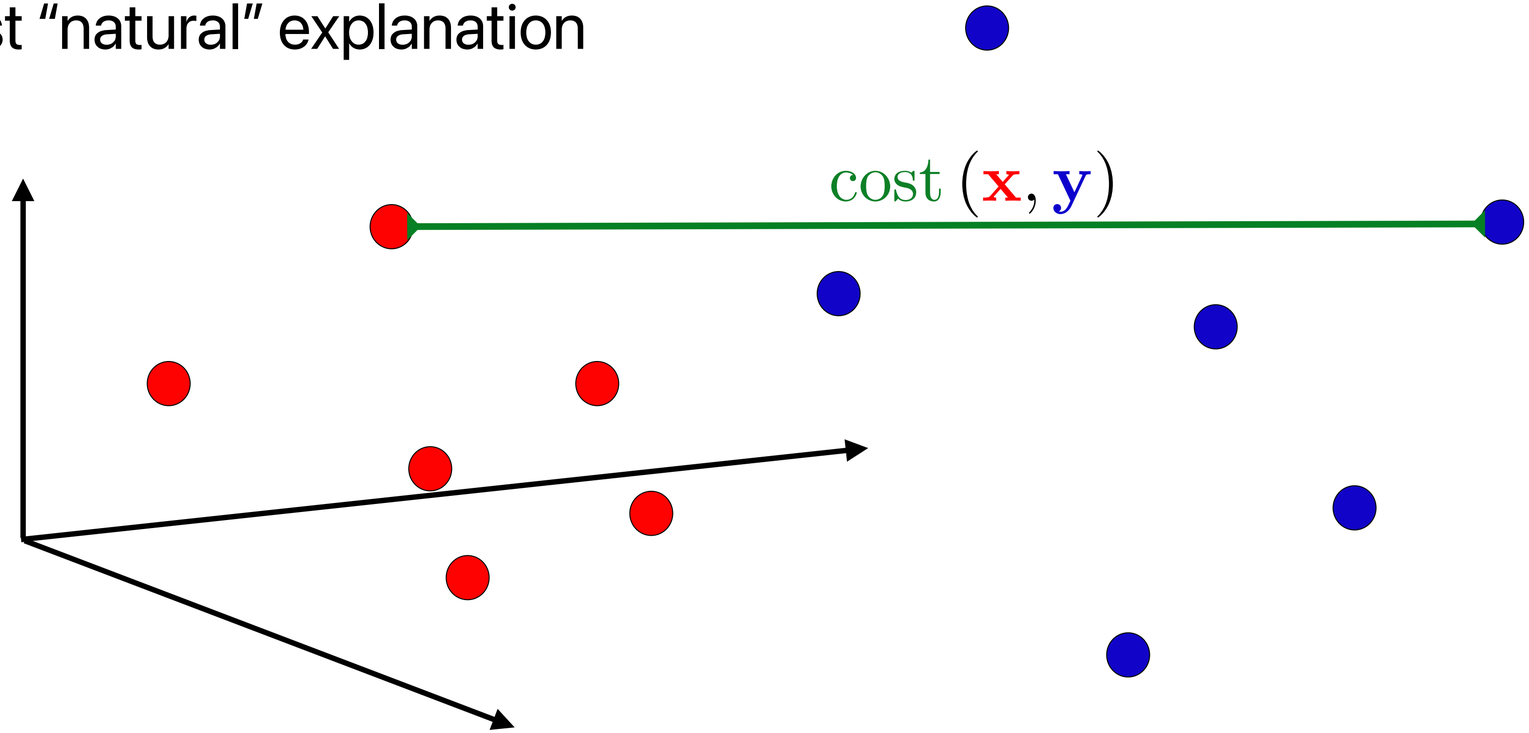
Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem

Goal: Find most “natural” explanation



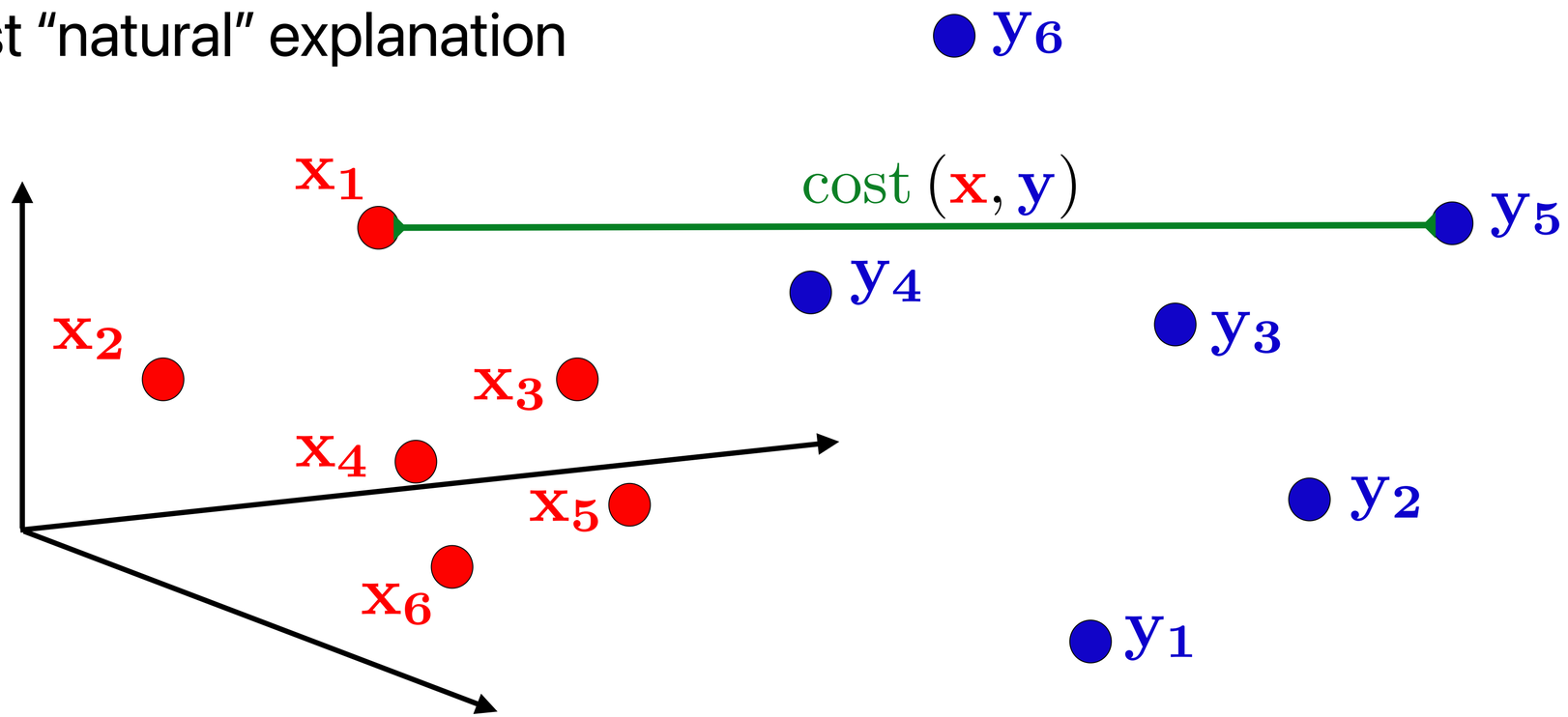
Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem

Goal: Find most “natural” explanation



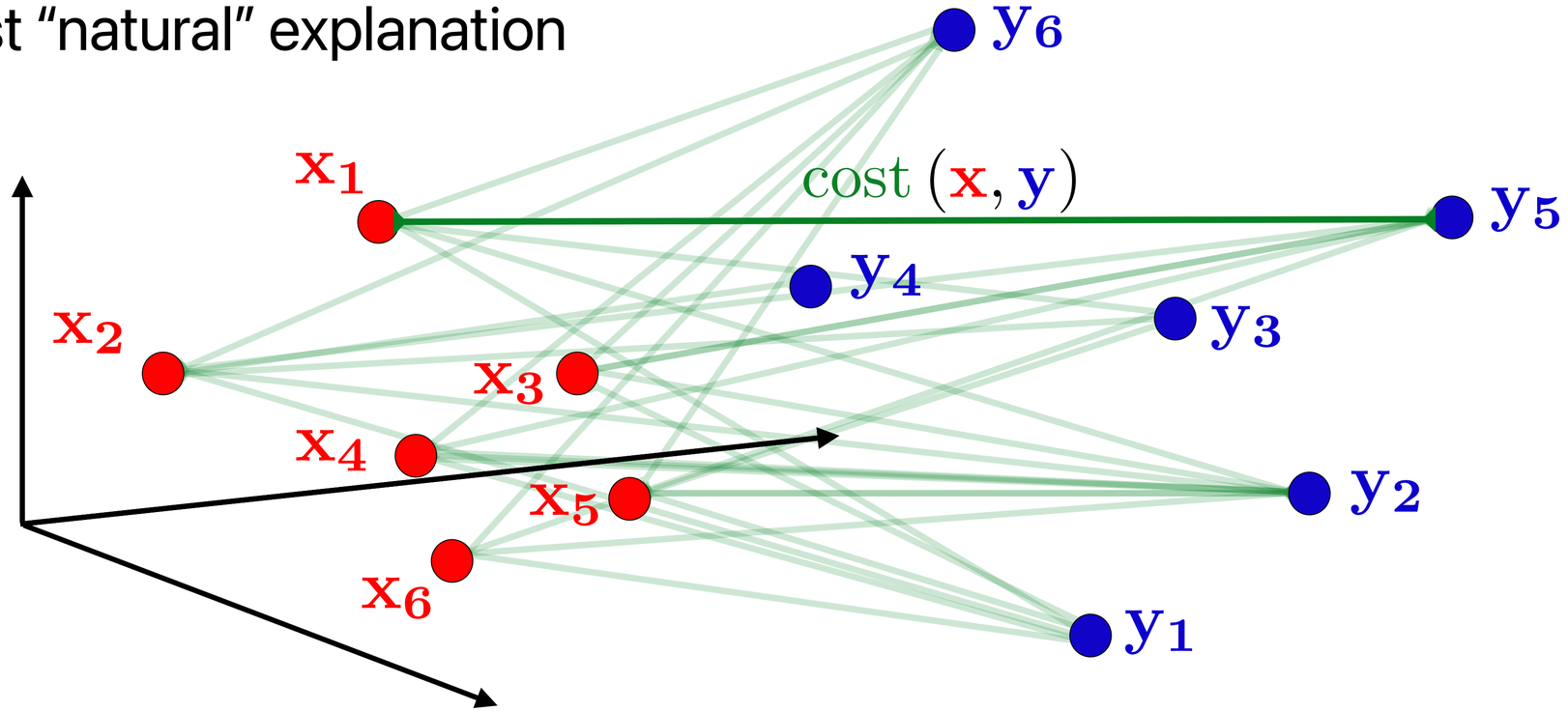
Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem

Goal: Find most “natural” explanation



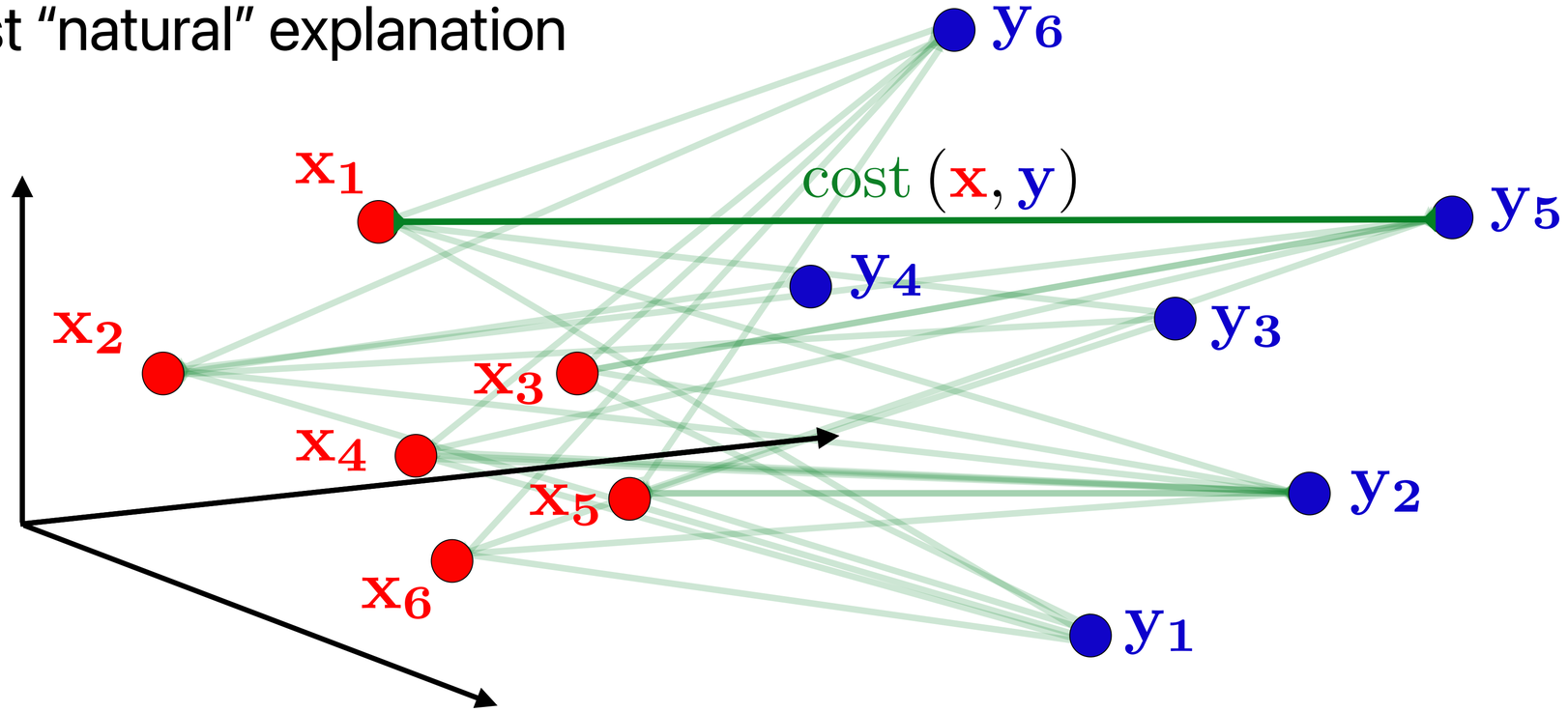
Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem

Goal: Find most “natural” explanation



Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem

Goal: Find most “natural” explanation



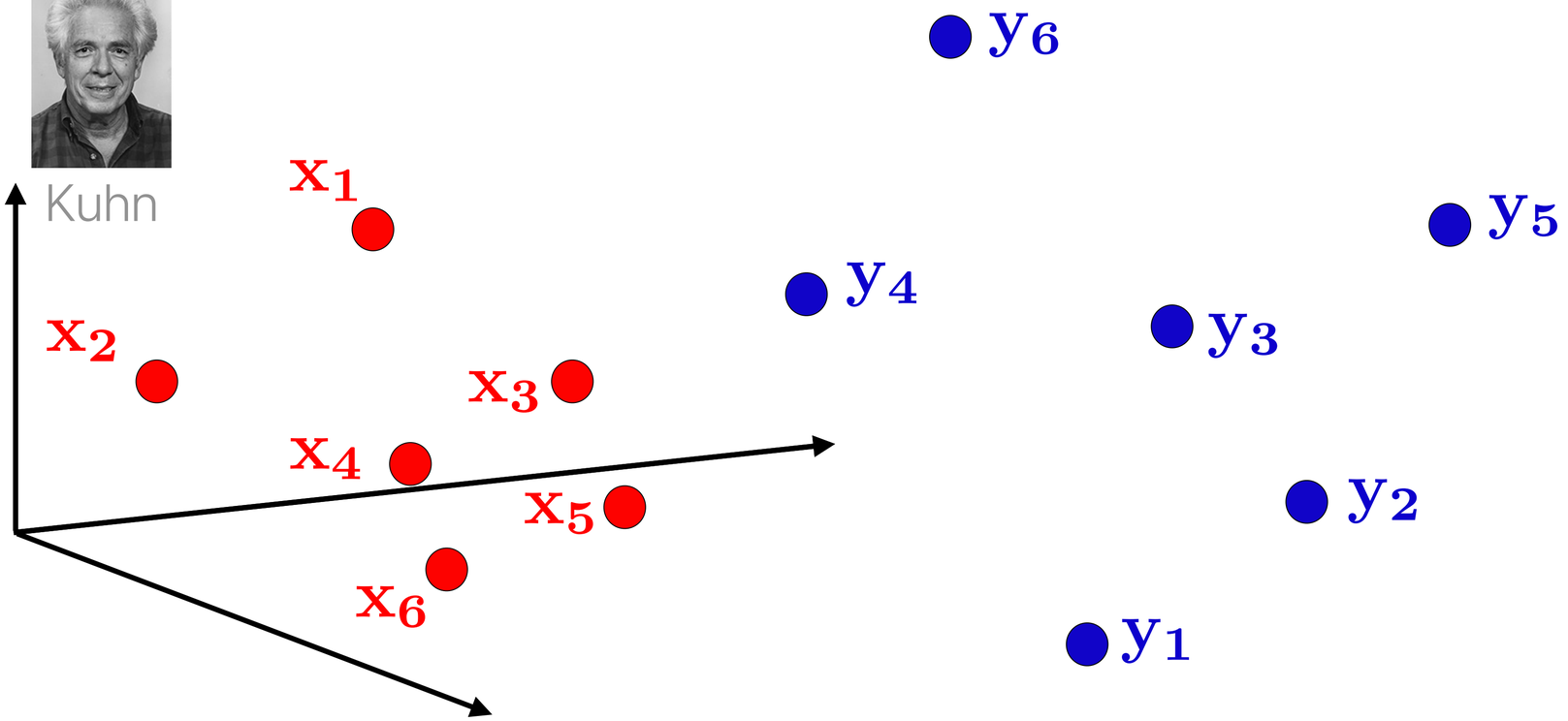
“natural” will be encoded with a cost between pairs of points, to find best permutation:

$$\min_{\sigma \in \mathcal{S}_n} \sum_{i=1}^{n} \boldsymbol{c}\left(\mathbf{x}_i, \mathbf{y}_{\sigma_i}\right)$$

Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem



Egerváry König Kuhn

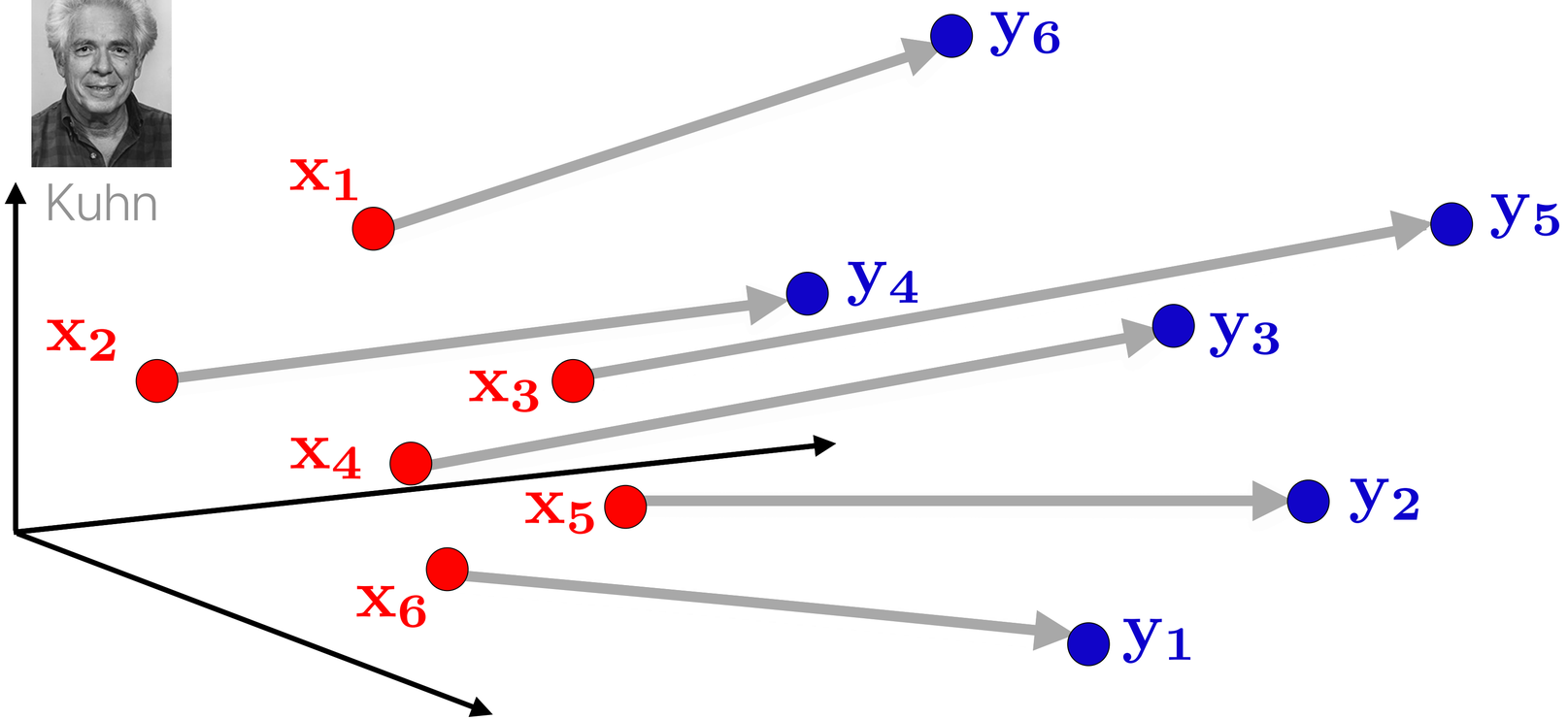


To get the best permutation $(6,4,5,3,2,1)$, run *Hungarian* algorithm, or any of its variants.

Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem



Egerváry König Kuhn

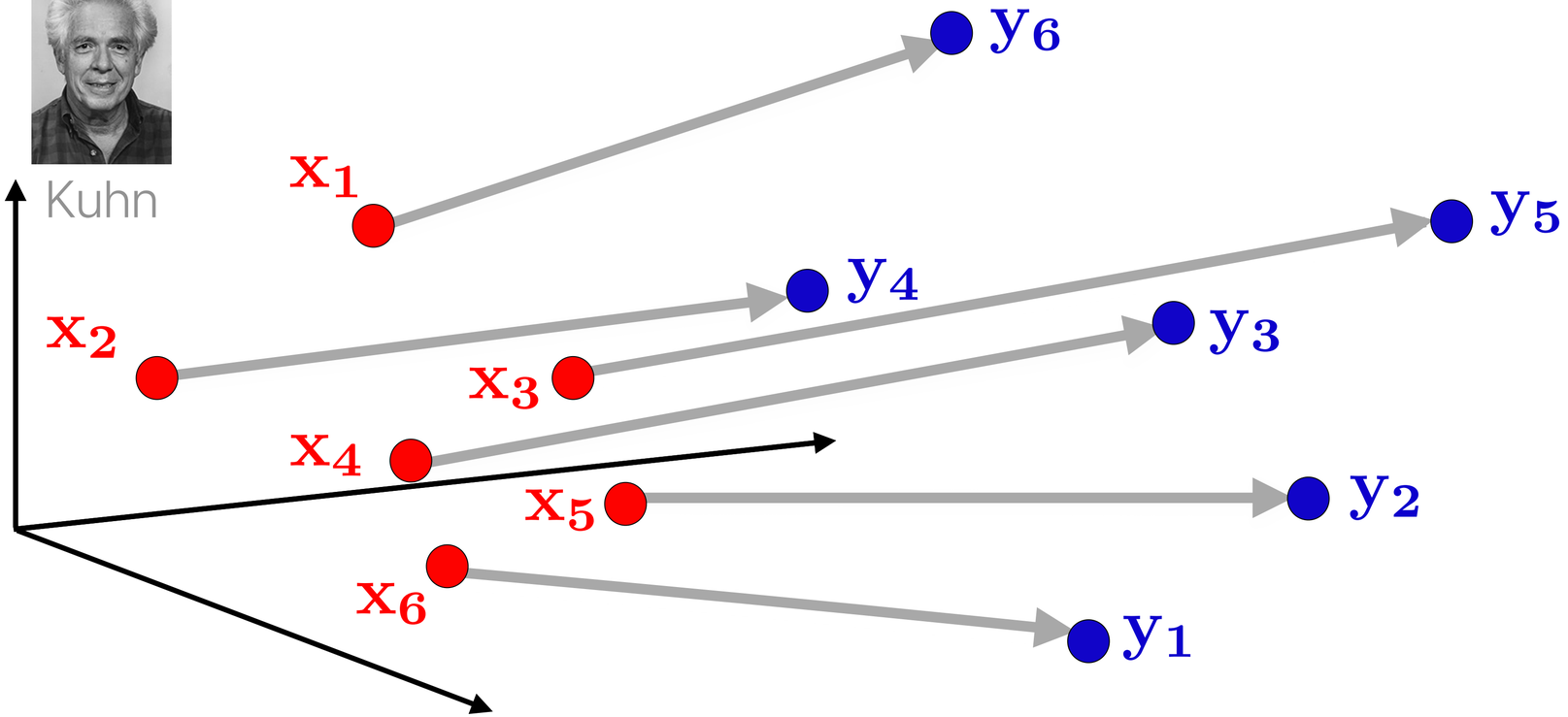


To get the best permutation $(6,4,5,3,2,1)$, run *Hungarian* algorithm, or any of its variants.

Optimal Transport v.0.0, *a.k.a.* Linear Assignment Problem



Egerváry König Kuhn



Sum of lengths = {*Wasserstein, OT, Kantorovich, EMD*} distance between **x**'s & **y**'s.

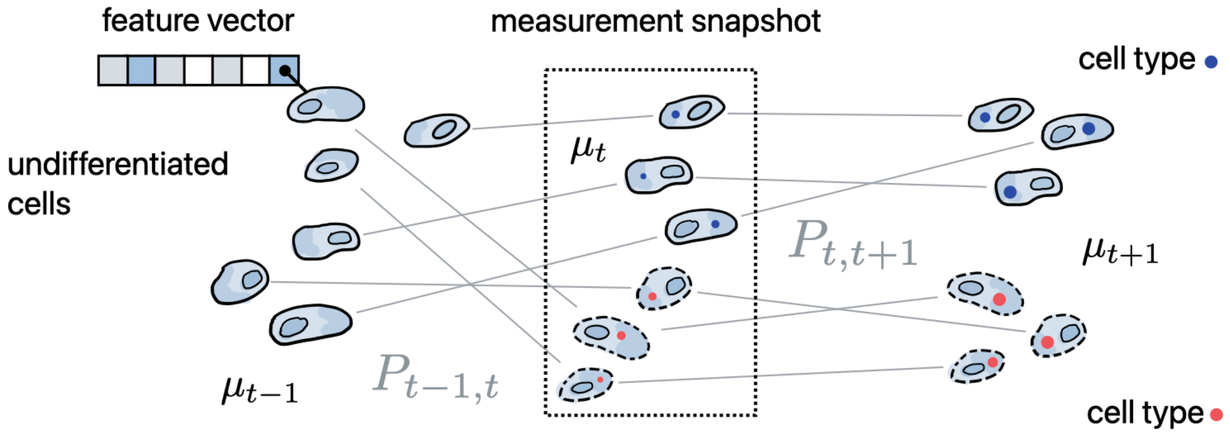


Linear Assignment Problems are Useful

Cell

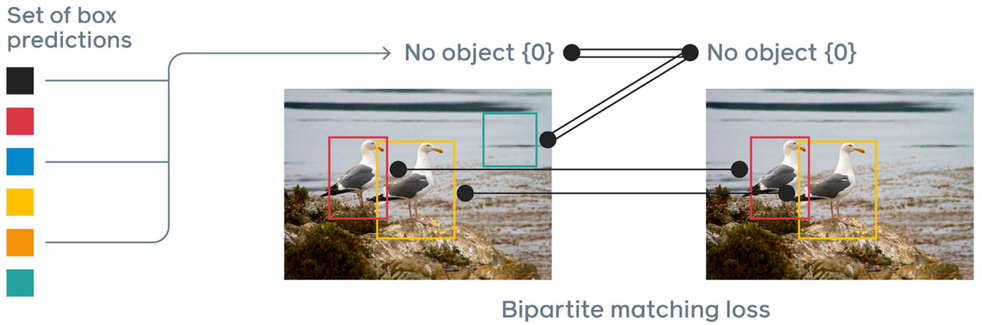
Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

[Schiebinger+19]



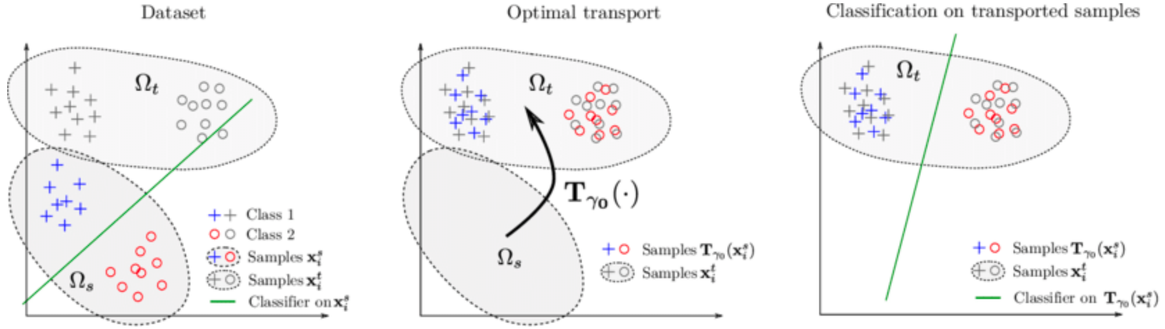
End-to-end object detection with Transformers

[Carion+20]



Optimal Transport for Domain Adaptation

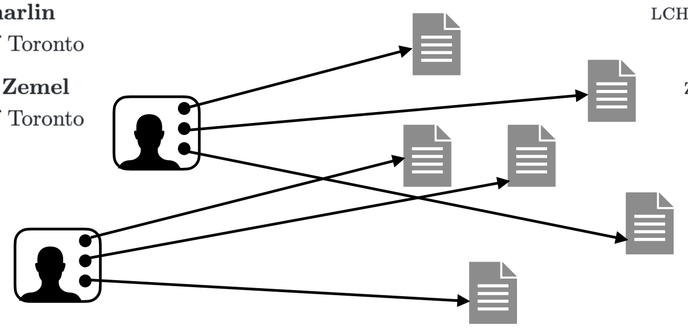
[Courty+15]



The Toronto Paper Matching System: An automated paper-reviewer assignment system

Laurent Charlin LCHARLIN@CS.TORONTO.EDU
University of Toronto

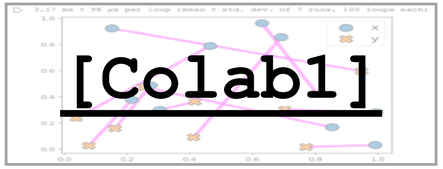
Richard S. Zemel ZEMEL@CS.TORONTO.EDU
University of Toronto

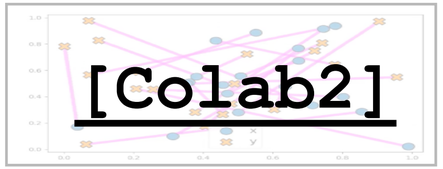


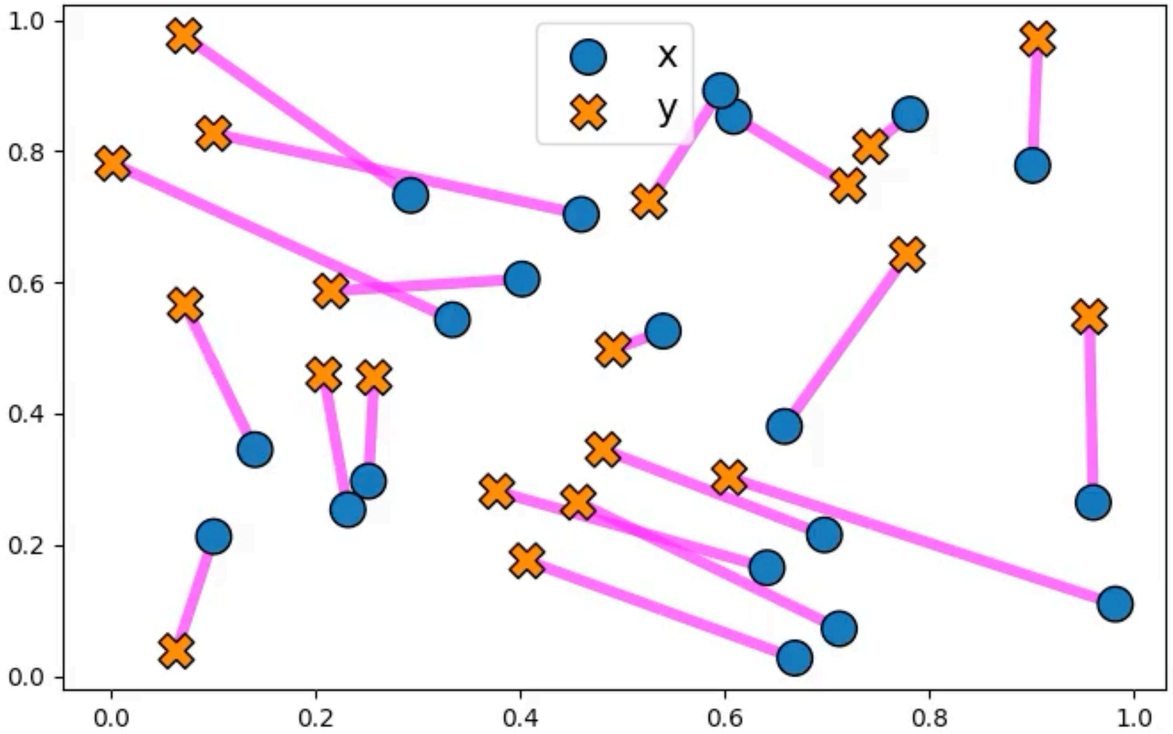
Unfortunately, Optimal Matchings are not Convenient for ML

Computing optimal matchings runs into multiple issues:

- Bad scaling, typically $O(n^3)$ complexity for n points
- Optimal matchings are unstable, and hard to parallelize efficiently on GPU



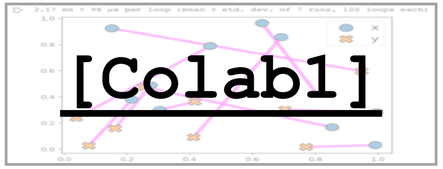


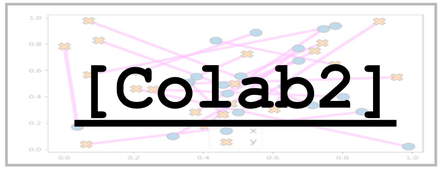


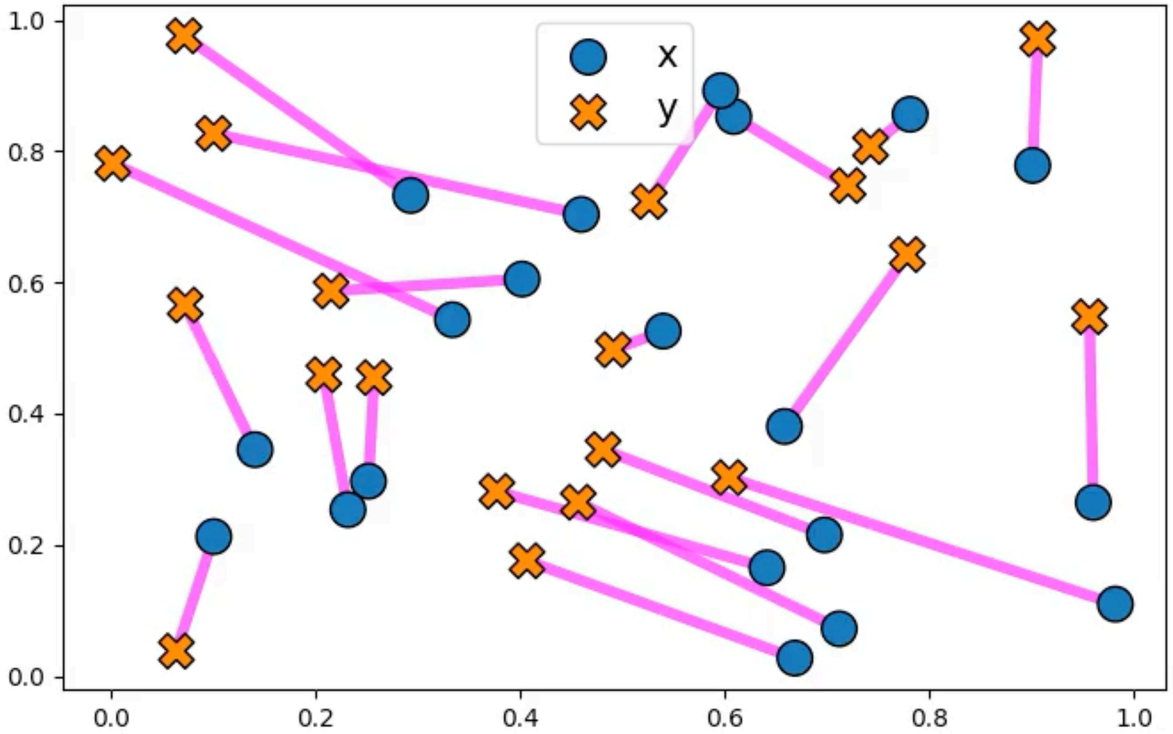
Unfortunately, Optimal Matchings are not Convenient for ML

Computing optimal matchings runs into multiple issues:

- Bad scaling, typically $O(n^3)$ complexity for n points
- Optimal matchings are unstable, and hard to parallelize efficiently on GPU



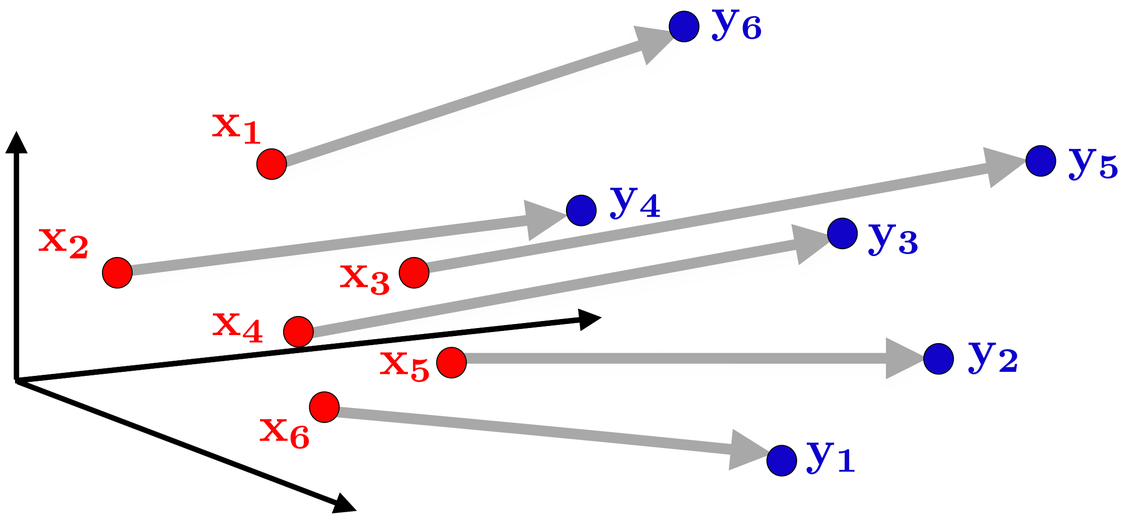




Additionally, Optimal Matchings Raise More Questions than Answers

More problematically, OM are brittle, rigid, and lack generalization:

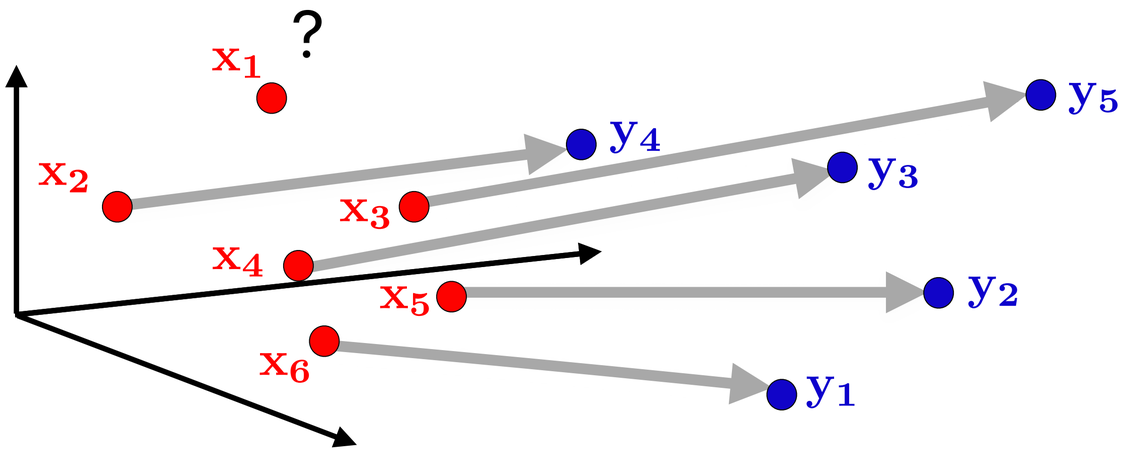
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is no obvious cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new out-of-sample points?
- What if we want stochasticity? From straight-lines to random stochastic paths?



Additionally, Optimal Matchings Raise More Questions than Answers

More problematically, OM are brittle, rigid, and lack generalization:

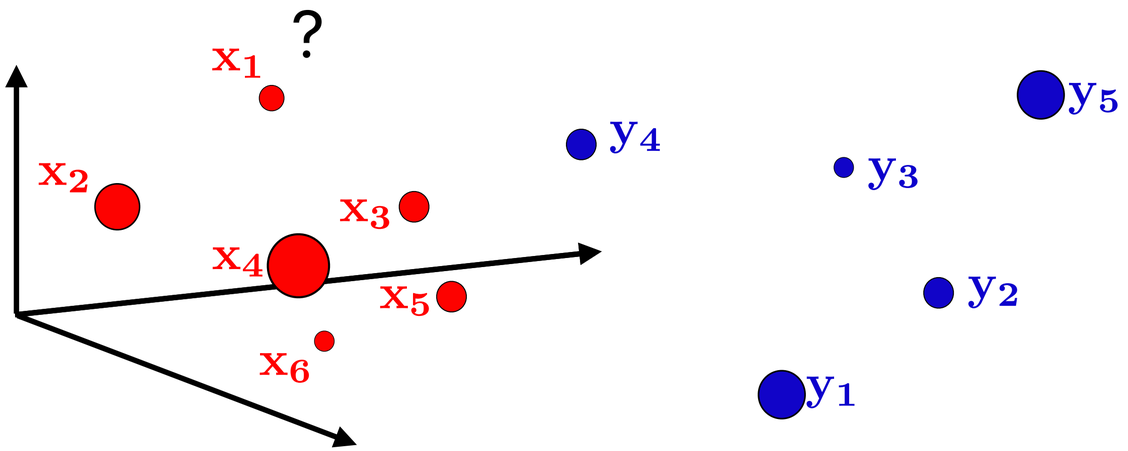
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is no obvious cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new out-of-sample points?
- What if we want stochasticity? From straight-lines to random stochastic paths?



Additionally, Optimal Matchings Raise More Questions than Answers

More problematically, OM are brittle, rigid, and lack generalization:

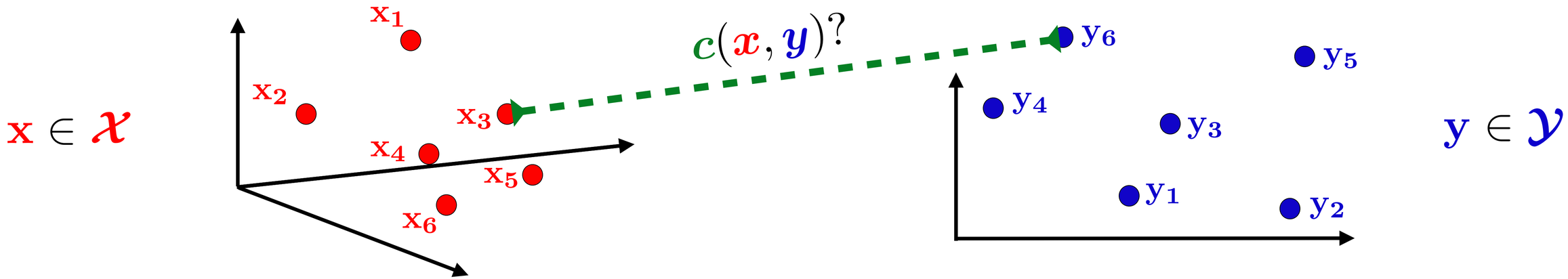
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is no obvious cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new out-of-sample points?
- What if we want stochasticity? From straight-lines to random stochastic paths?



Optimal Matchings Raise More Questions than Answers

Even more problematically, they are brittle, rigid, and lack generalization:

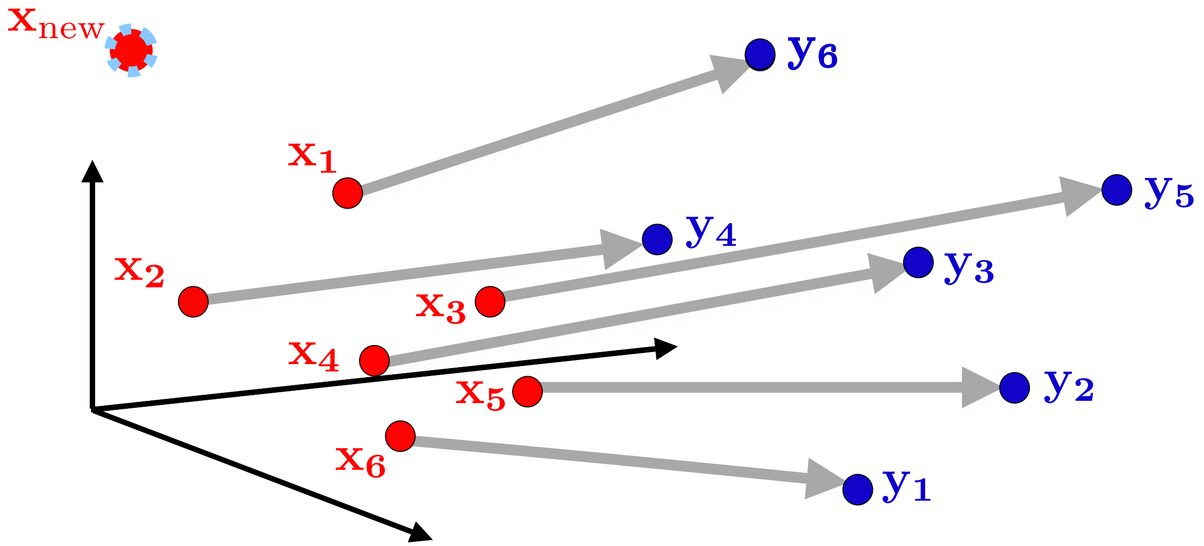
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is **no obvious** cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new out-of-sample points?
- What if we want stochasticity? From straight-lines to random stochastic paths?



Optimal Matchings Raise More Questions than Answers

Even more problematically, they are brittle, rigid, and lack generalization:

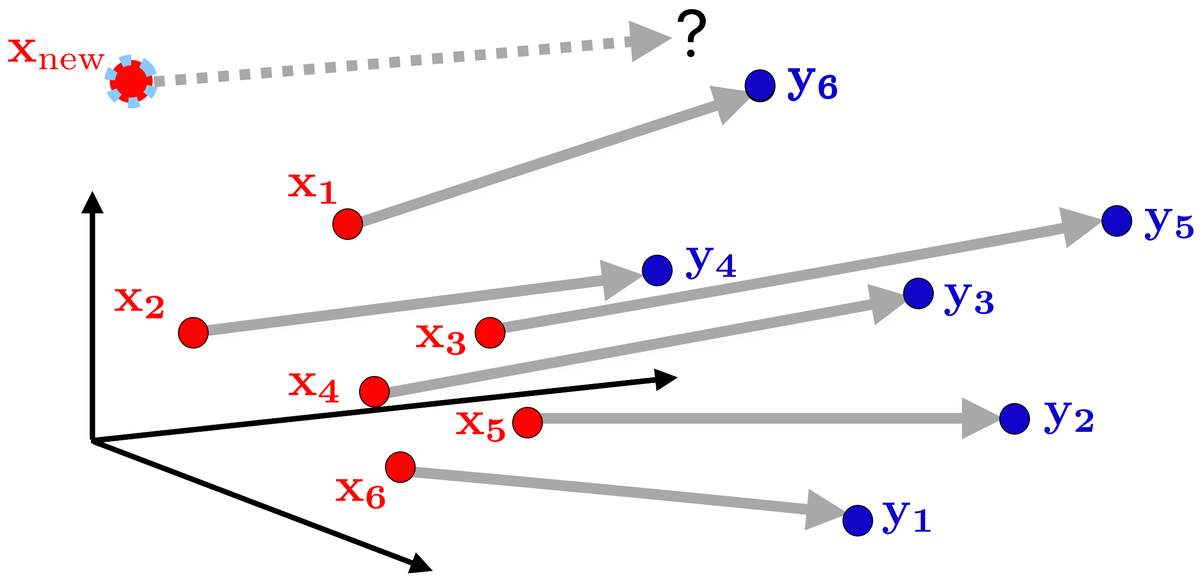
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is no obvious cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new **out-of-sample** points?
- What if we want stochasticity? From straight-lines to random stochastic paths?



Optimal Matchings Raise More Questions than Answers

Even more problematically, they are brittle, rigid, and lack generalization:

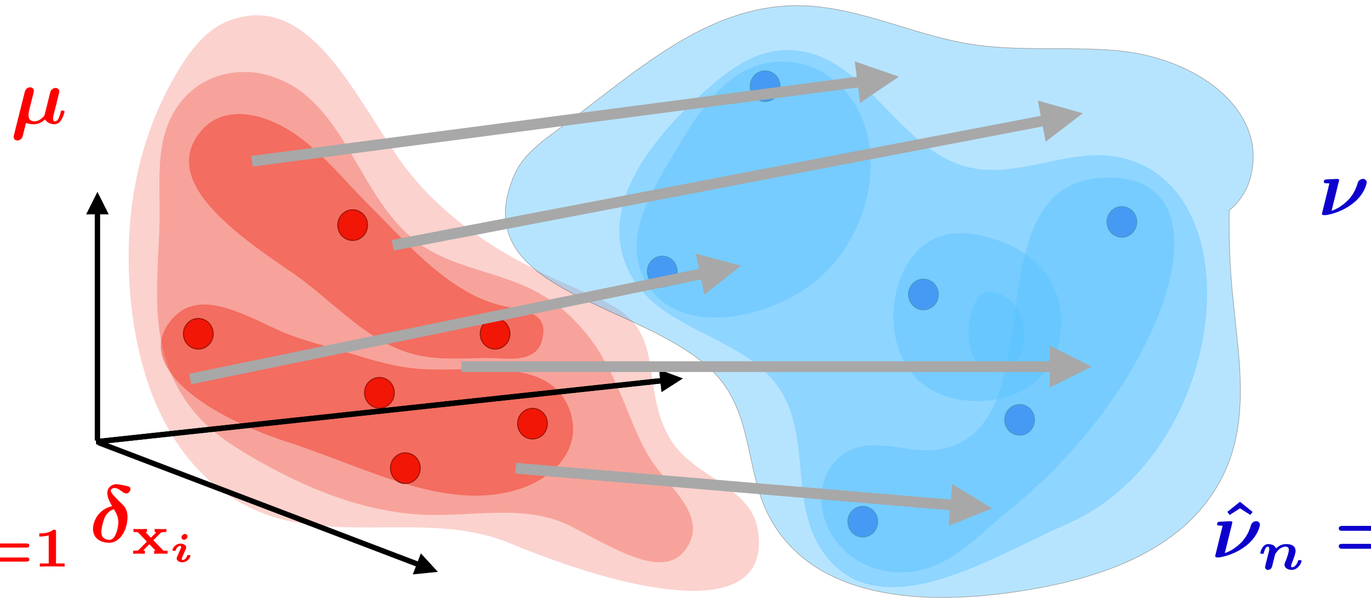
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is no obvious cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new **out-of-sample** points?
- What if we want stochasticity? From straight-lines to random stochastic paths?



Optimal Matchings Raise More Questions than Answers

Even more problematically, they are brittle, rigid, and lack generalization:

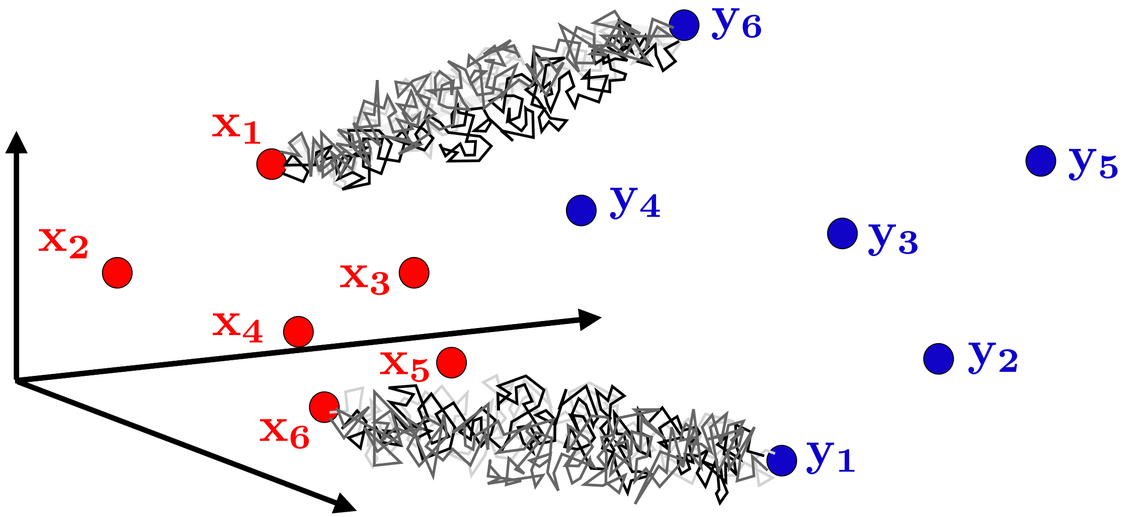
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is no obvious cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new **out-of-sample** points?
- What if we want stochasticity? From straight-lines to random stochastic paths?



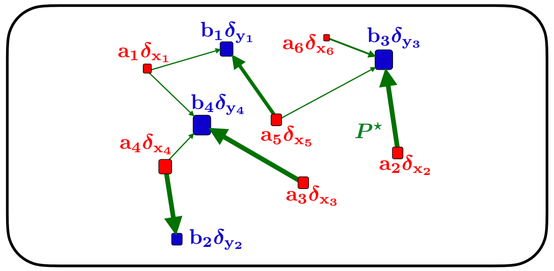
Optimal Matchings Raise More Questions than Answers

Even more problematically, they are brittle, rigid, and lack generalization:

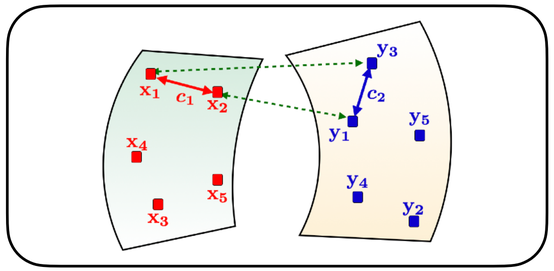
- What if the numbers of **x** and **y** points differ, $n \neq m$? weighted points?
- What if there is no obvious cost function? e.g. **x** and **y** points are in ≠ spaces?
- What if we need ways to generalize, and move new out-of-sample points?
- What if we need **stochasticity**? From straight-lines to stochastic processes?



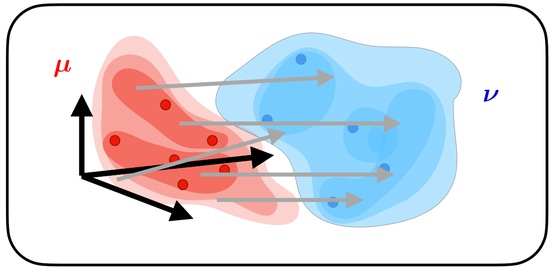
The Jargon of Optimal Transport



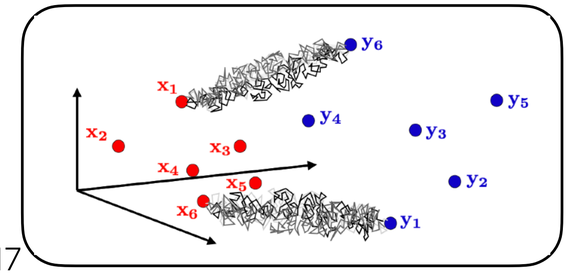
- LAP, Kantorovich (Rubinstein), duality and potential.
- Optimal couplings, transport **plans**, EMD, Wasserstein



- QAP, Gromov-Wasserstein (GW) distance
- Fused GW, graph isomorphism problems



- Monge, transport **map**, Brenier & Gangbo-McCann theorems
- Monge-Ampère PDE



- Dynamic OT Flow-matching approaches, Benamou-Brenier
- Flow matching, ridge matching, rectified flows

Outline of the Tutorial

Prelude Warm-Up: Starting with Optimal Matchings

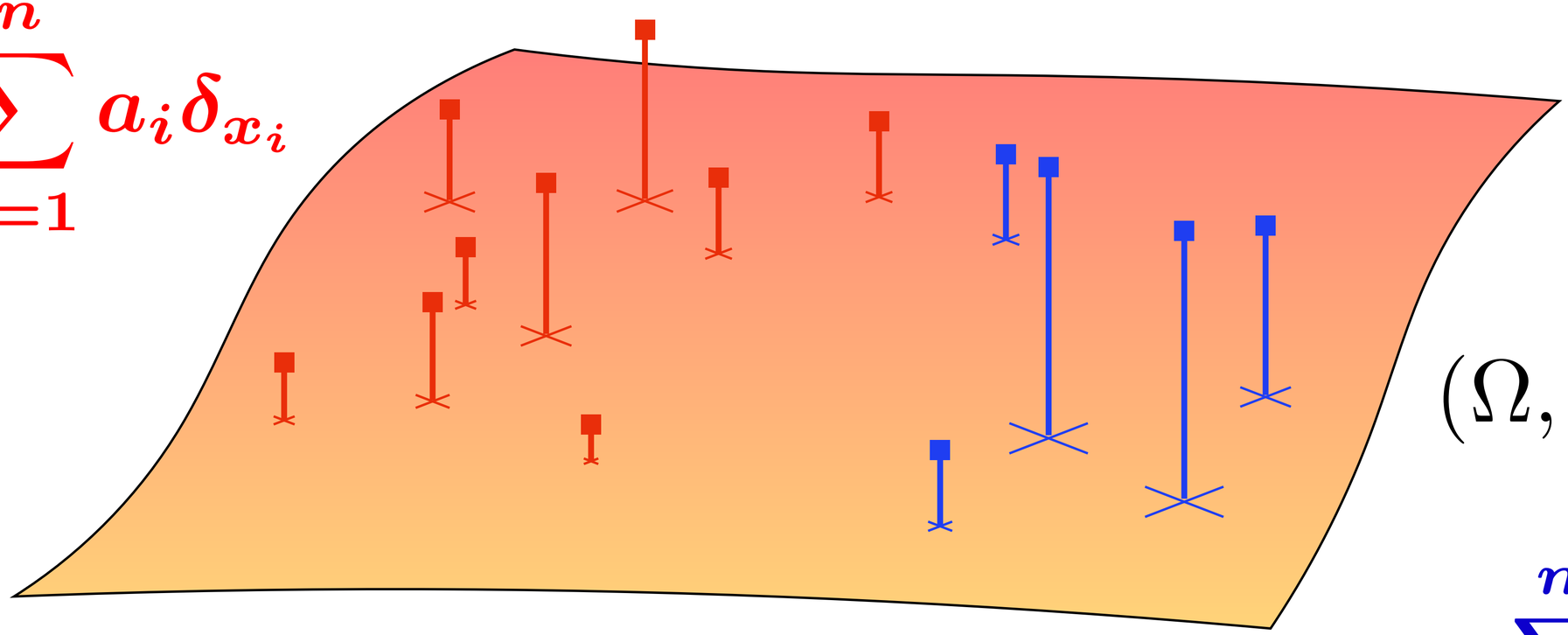
Part 1 **Kantorovich Formulation of OT, Discrete Computations**

- Optimizing over space of coupling matrices, linear programs
- Regularized OT, entropic and rank constrained, unbalanced, quadratic
- Differentiability of OT plans

Part 2 Monge Formulation, Duality, Gangbo-McCann-Brenier Theorems

Optimal Transport between Two Empirical Measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



$$(\Omega, c)$$

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

Mines and Factories Metaphor for Optimal Transport:

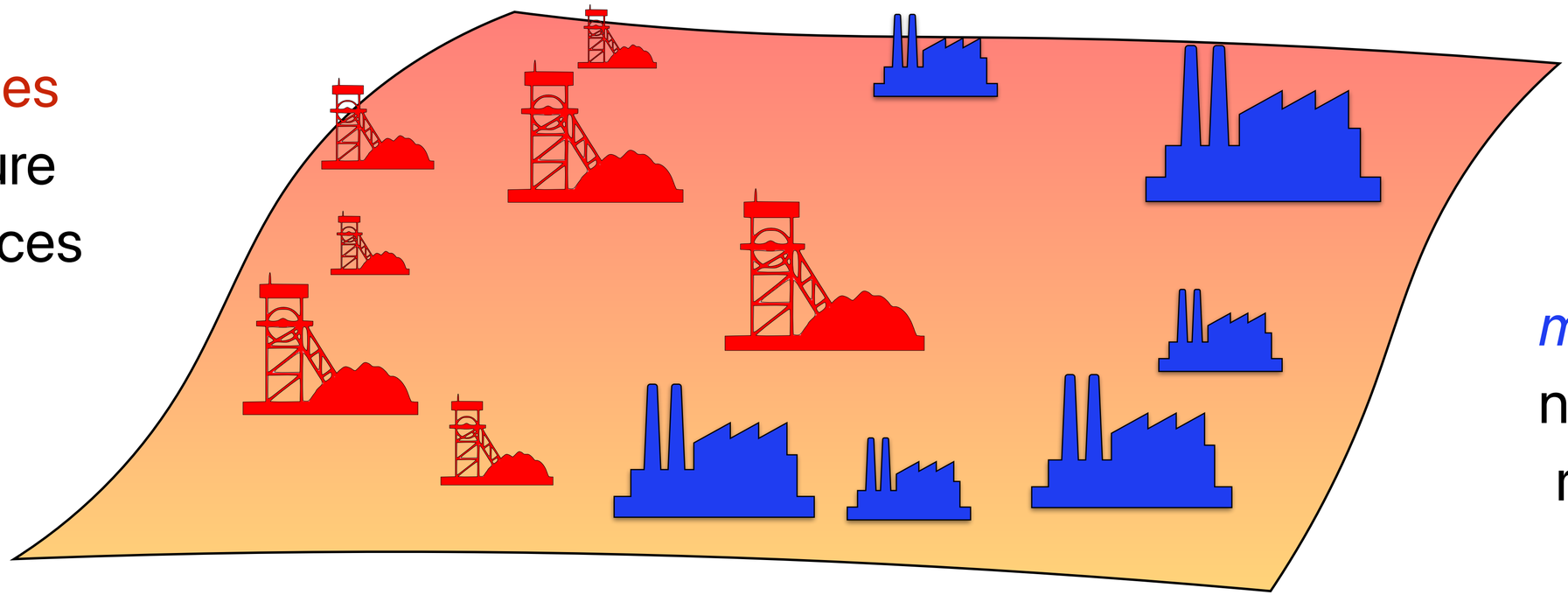

Hitchcock


Kantorovich


Dantzig


Koopmans

n mines
capture
resources



m factories
need those
resources

Mines and Factories Metaphor for Optimal Transport:

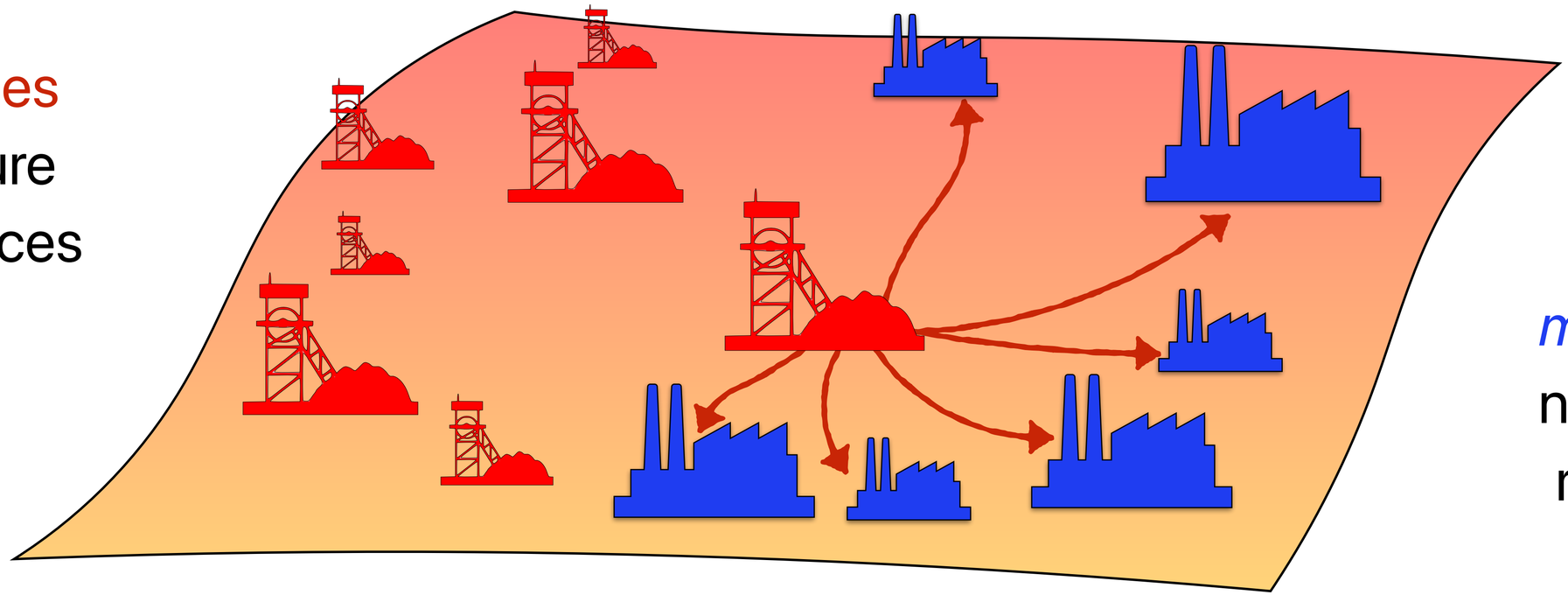





Hitchcock Kantorovich Dantzig Koopmans

Each of the *n* mines must dispatch its production

n mines
capture
resources



m factories
need those
resources

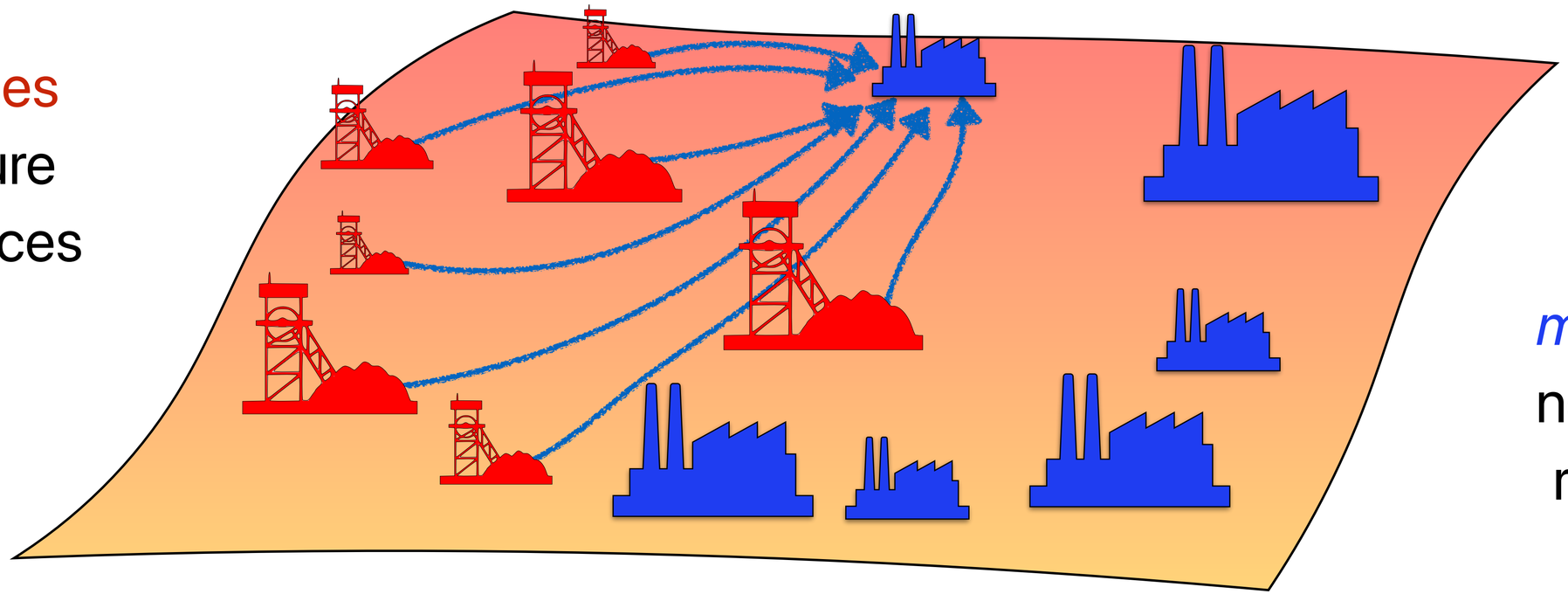
Mines and Factories Metaphor for Optimal Transport:






Hitchcock Kantorovich Dantzig Koopmans

n mines
capture
resources

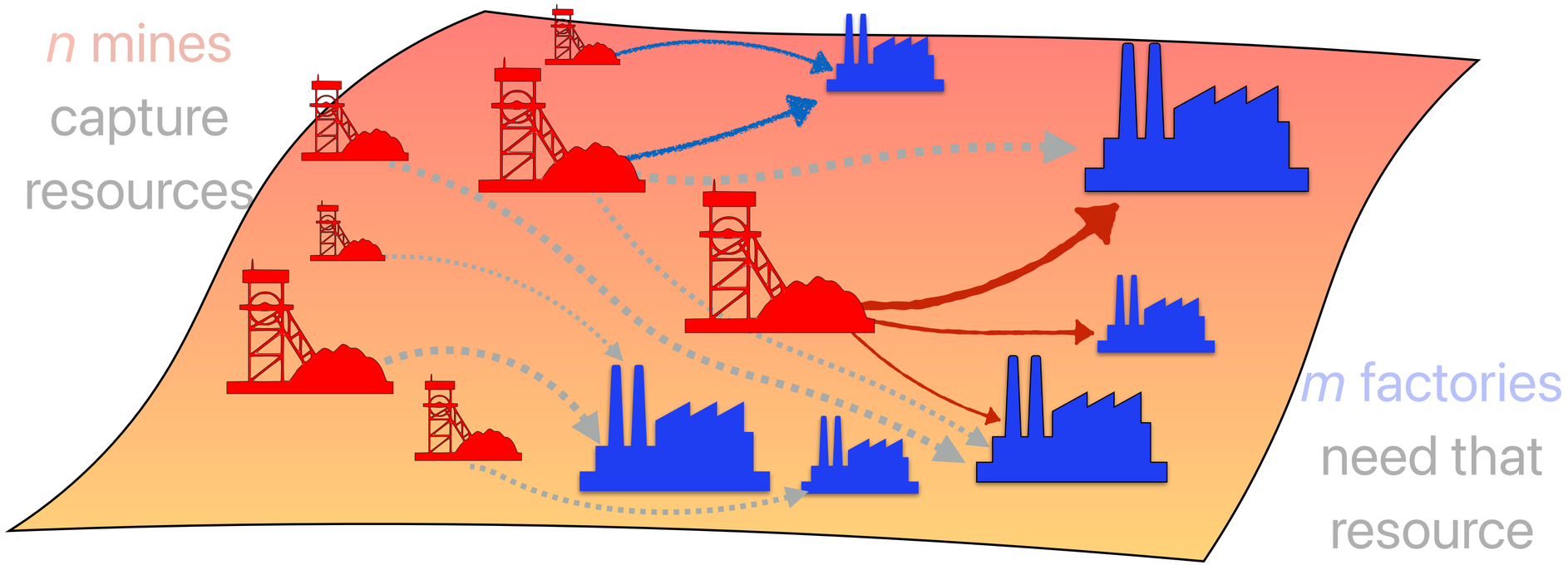


m factories
need those
resources

Each of the *m* factories must get resource it needs

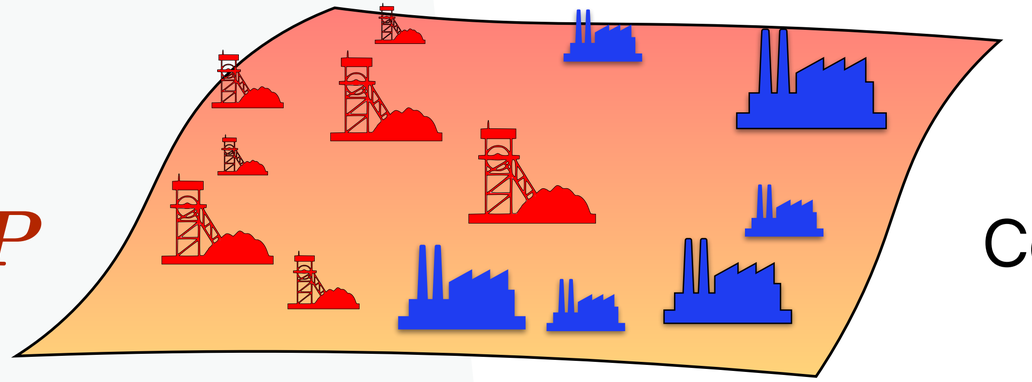
Mines and Factories Metaphor for Optimal Transport:

Optimal Transport computes the **least-costly** transport matrix

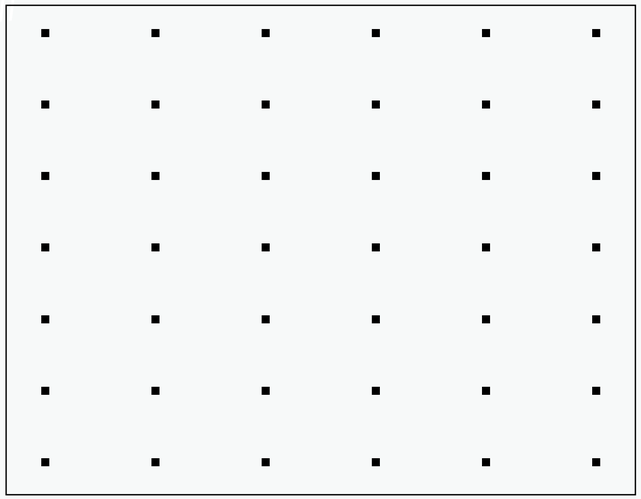


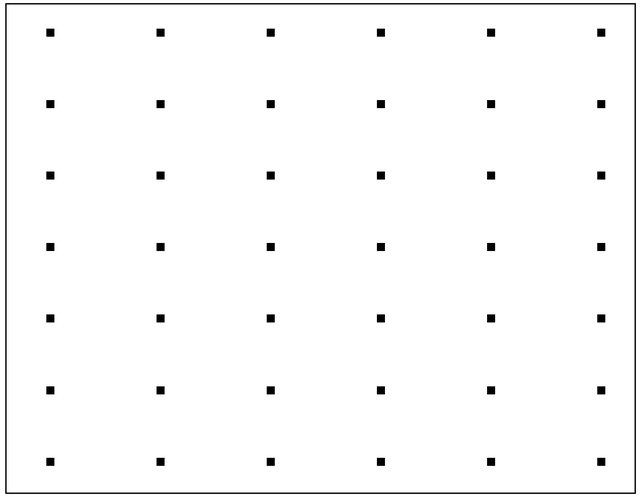
Mines and Factories Metaphor for Optimal Transport:

Couplings P



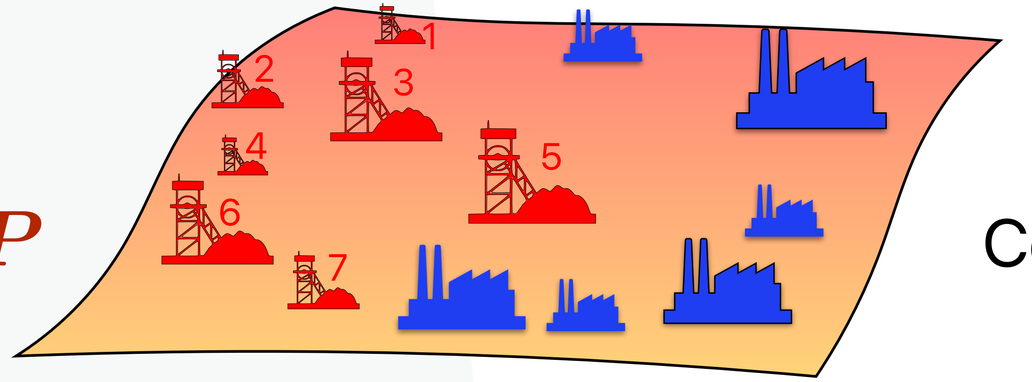
Cost matrix M_{XY}



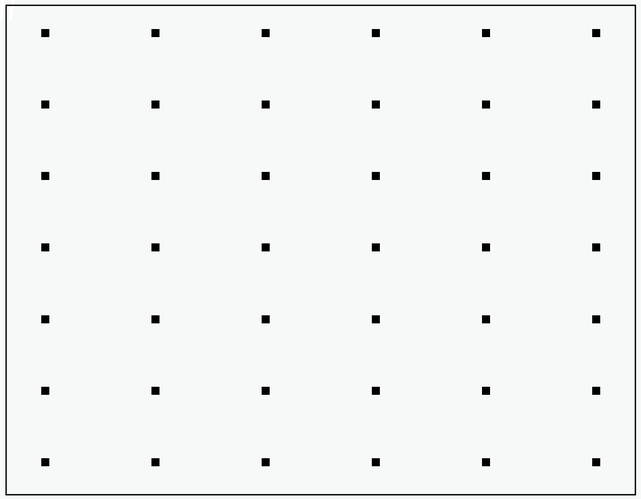


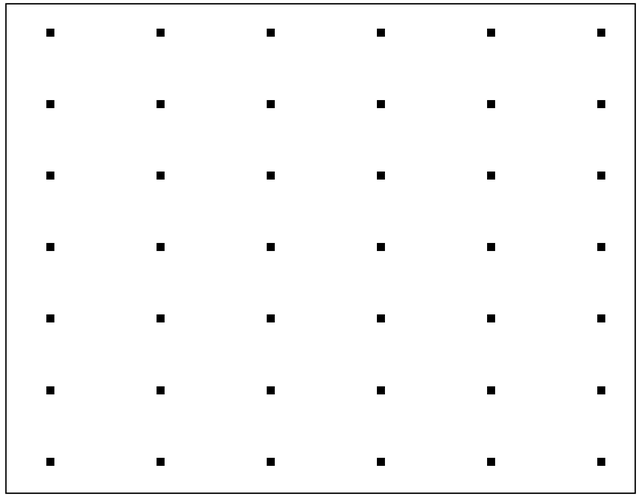
Mines and Factories Metaphor for Optimal Transport:

Couplings P



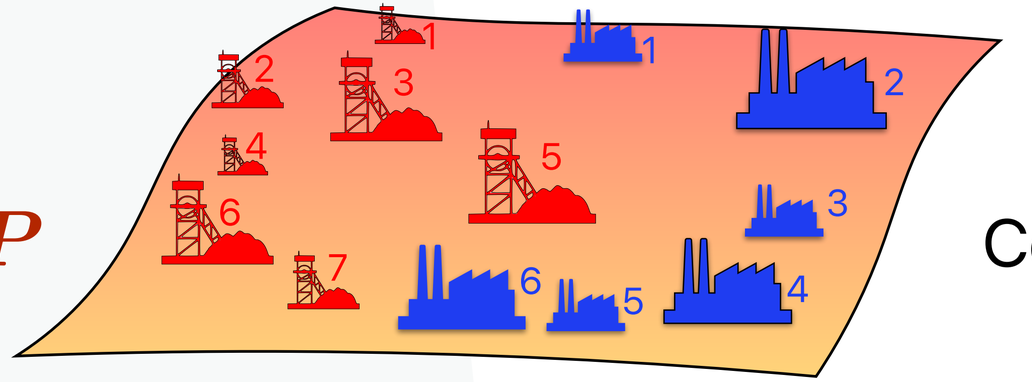
Cost matrix M_{XY}



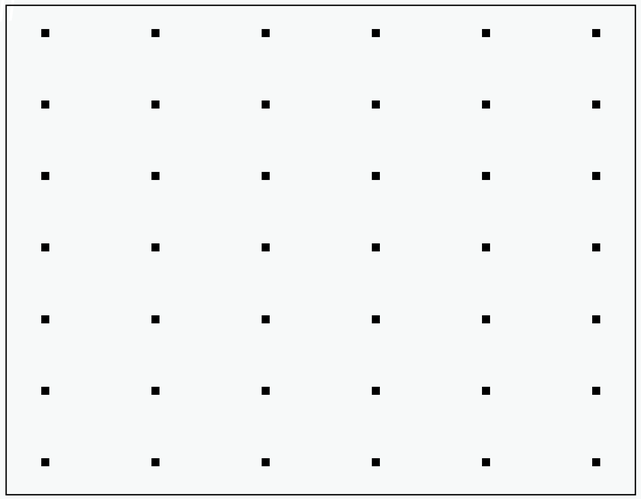


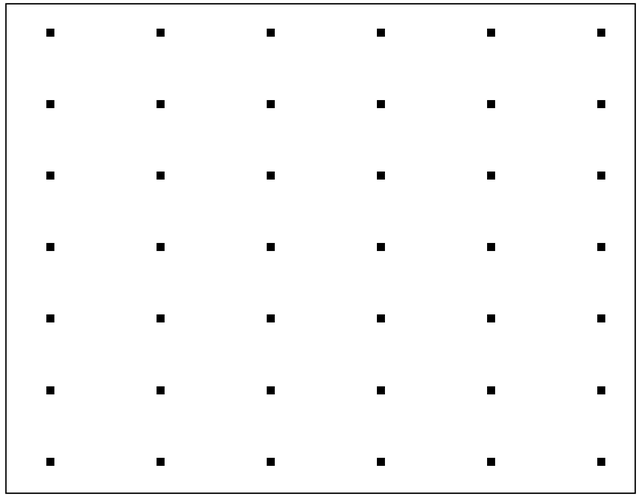
Mines and Factories Metaphor for Optimal Transport:

Couplings P



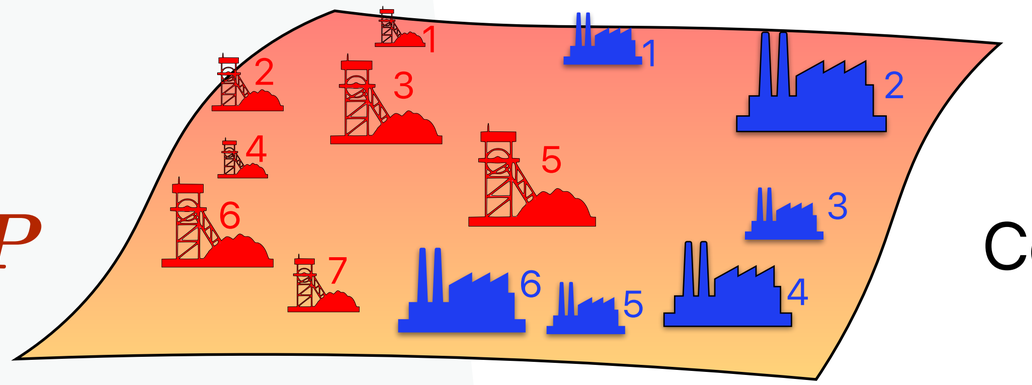
Cost matrix M_{XY}



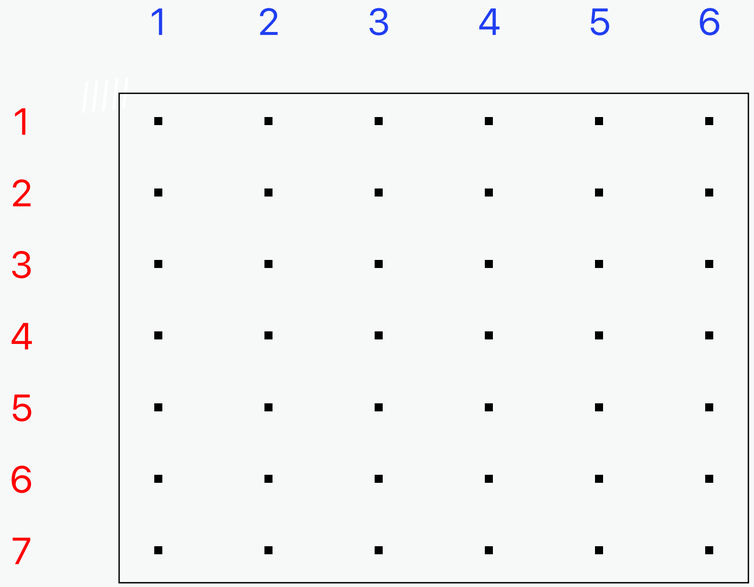


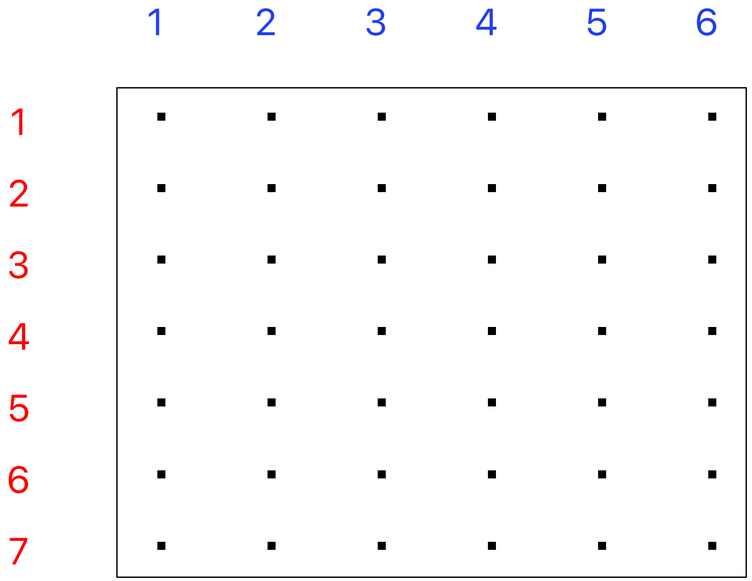
Mines and Factories Metaphor for Optimal Transport:

Couplings P



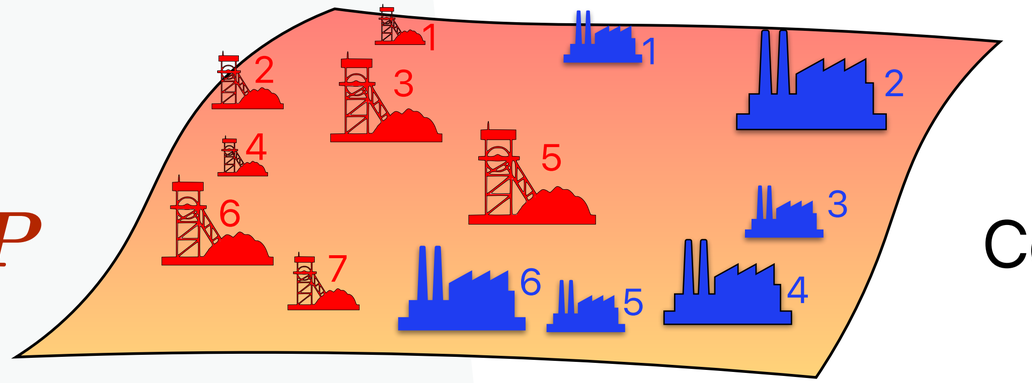
Cost matrix M_{XY}



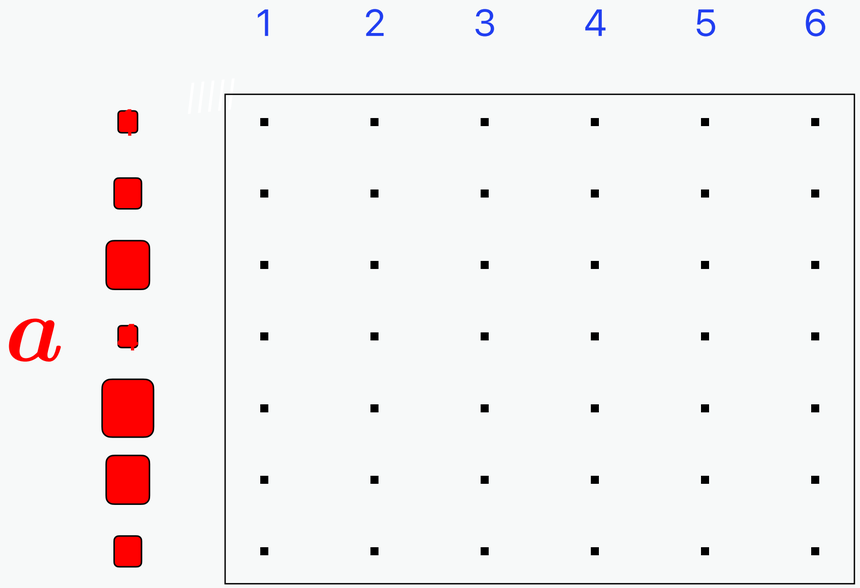


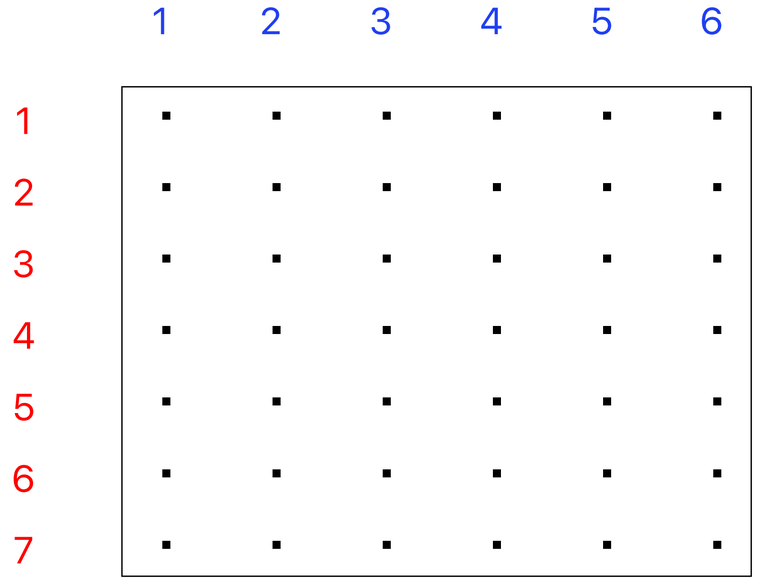
Mines and Factories Metaphor for Optimal Transport:

Couplings P



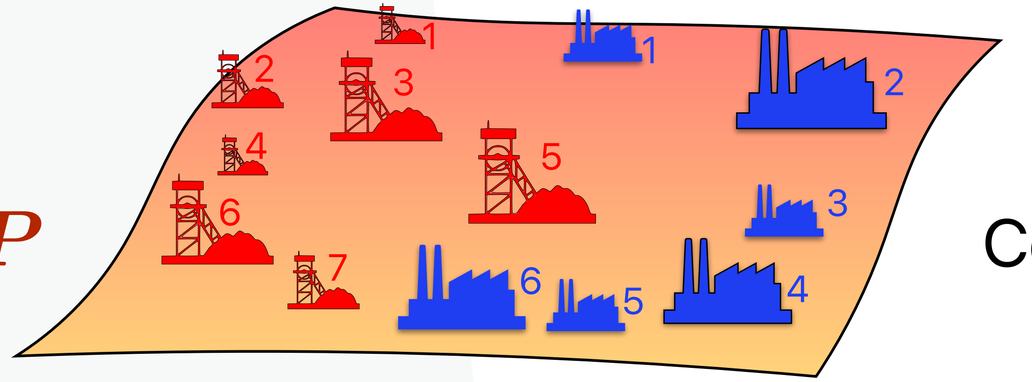
Cost matrix M_{XY}



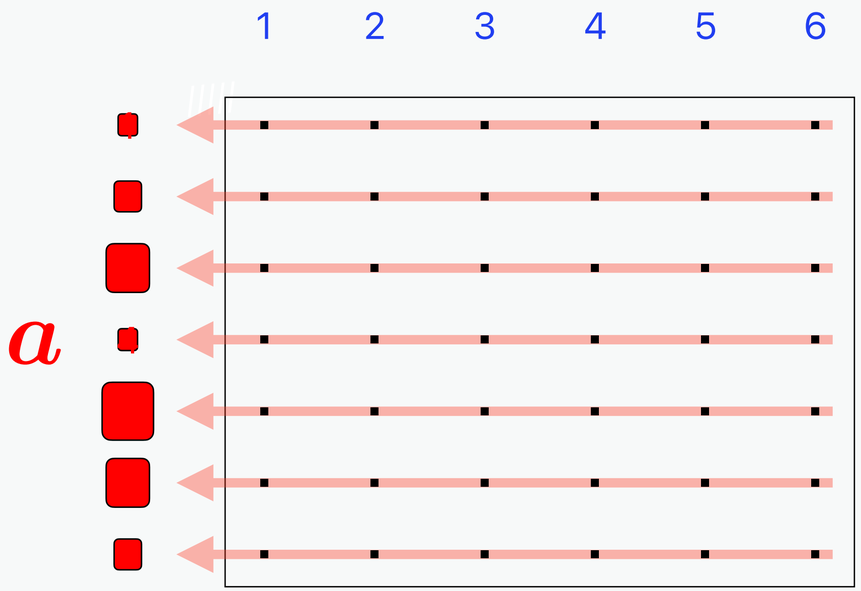


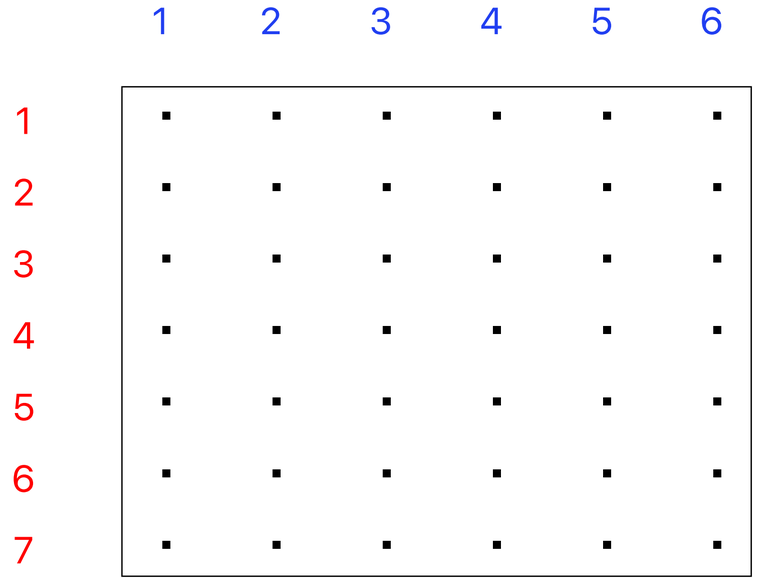
Mines and Factories Metaphor for Optimal Transport:

Couplings P



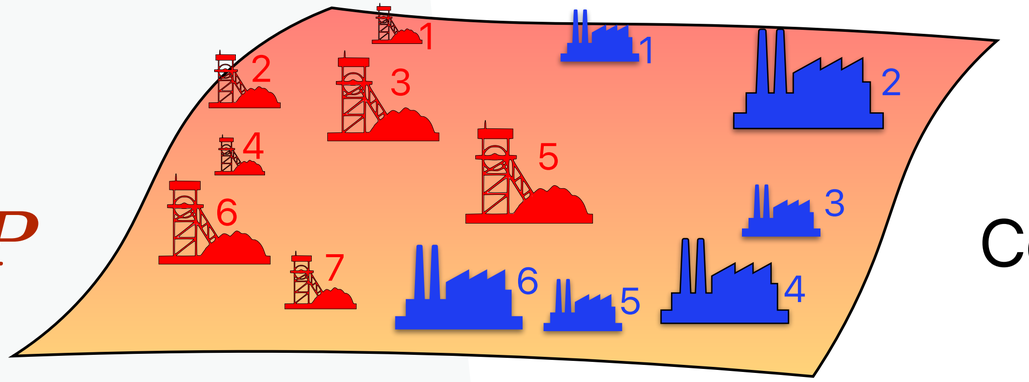
Cost matrix M_{XY}

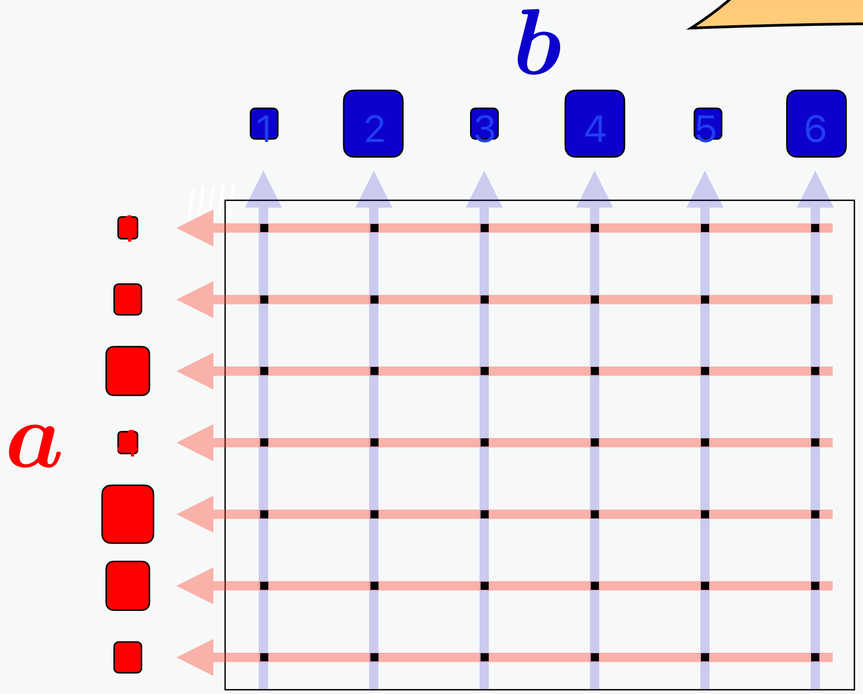




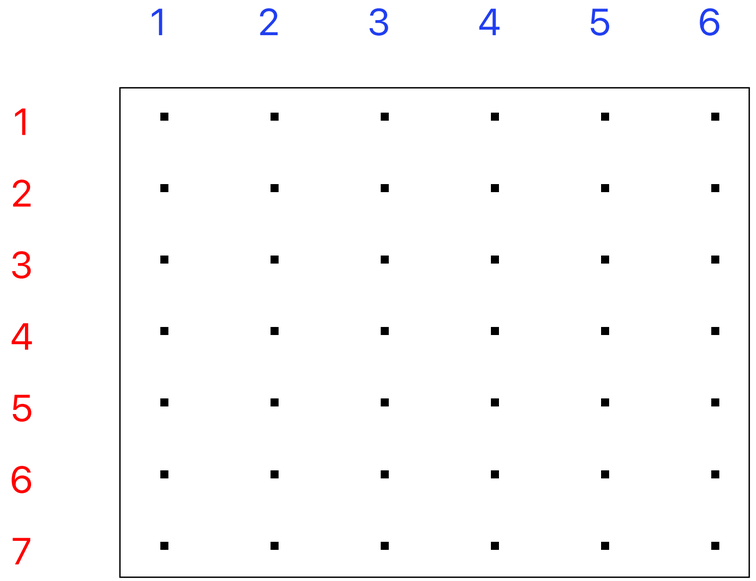
Mines and Factories Metaphor for Optimal Transport:

Couplings P



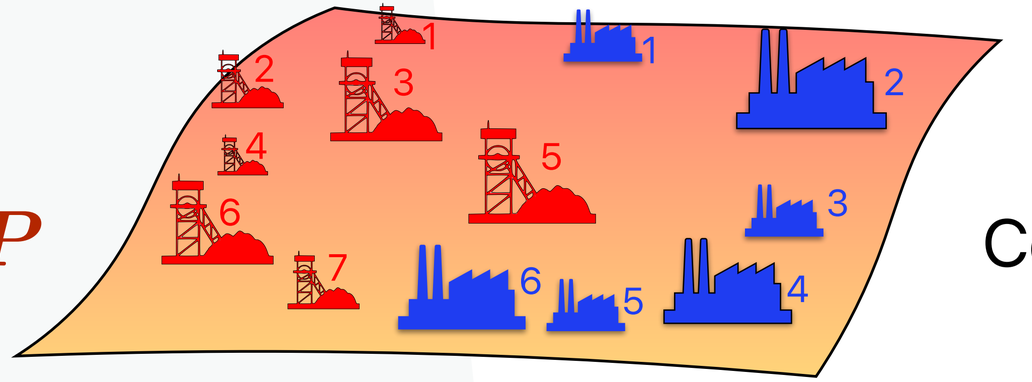


Cost matrix M_{XY}

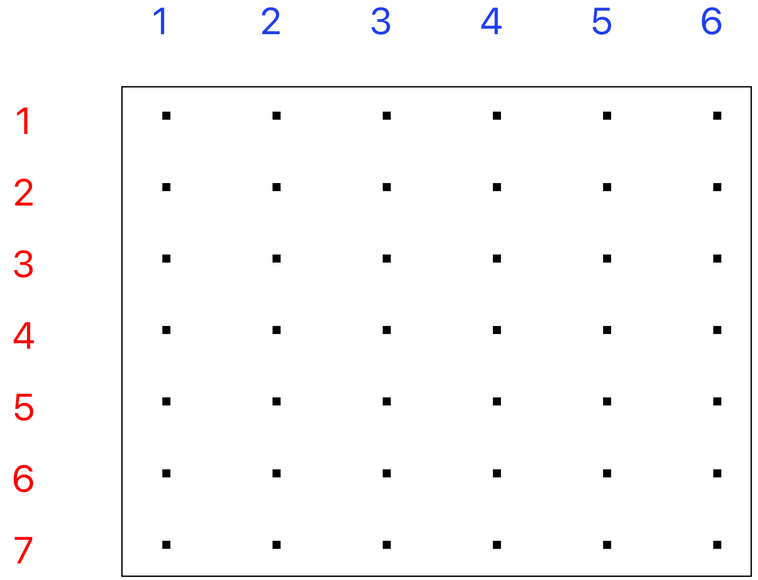


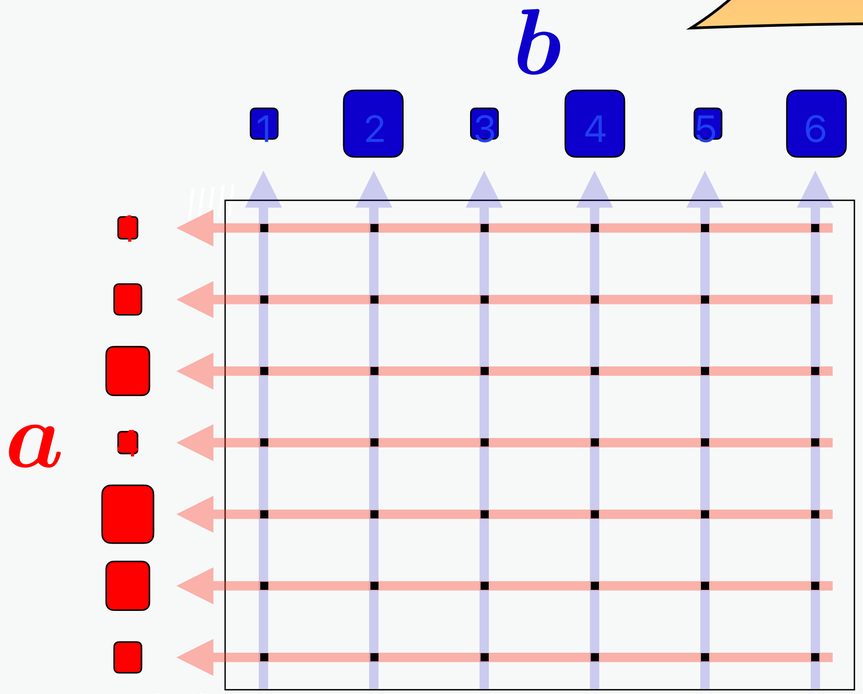
Mines and Factories Metaphor for Optimal Transport:

Couplings $\boldsymbol{P}$



Cost matrix $M_{\boldsymbol{XY}}$

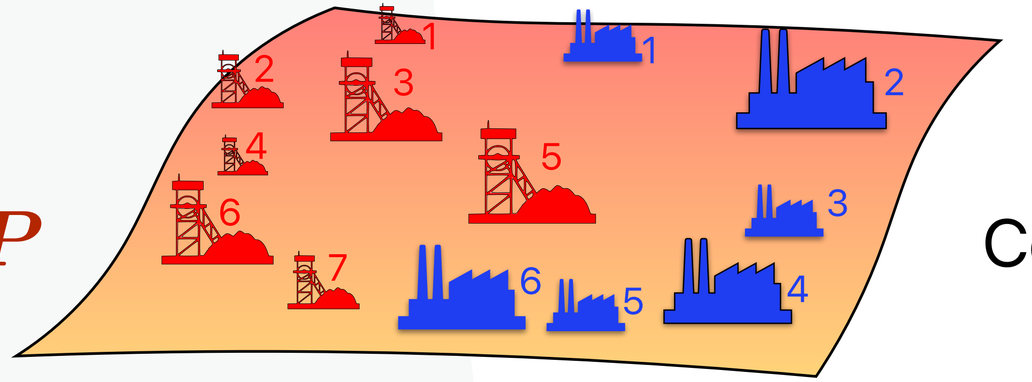




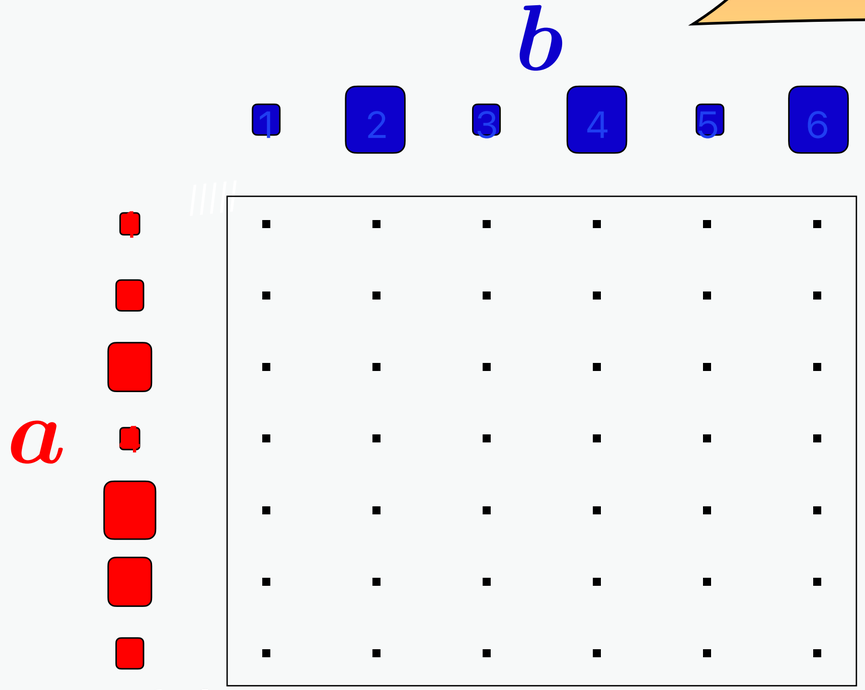
$$U(\boldsymbol{a},\boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

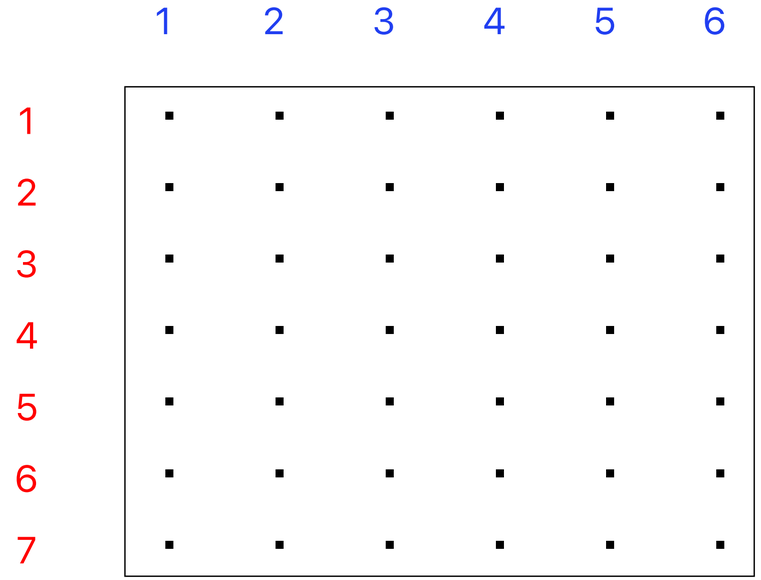
Mines and Factories Metaphor for Optimal Transport:

Couplings $\boldsymbol{P}$



Cost matrix $M_{\boldsymbol{XY}}$

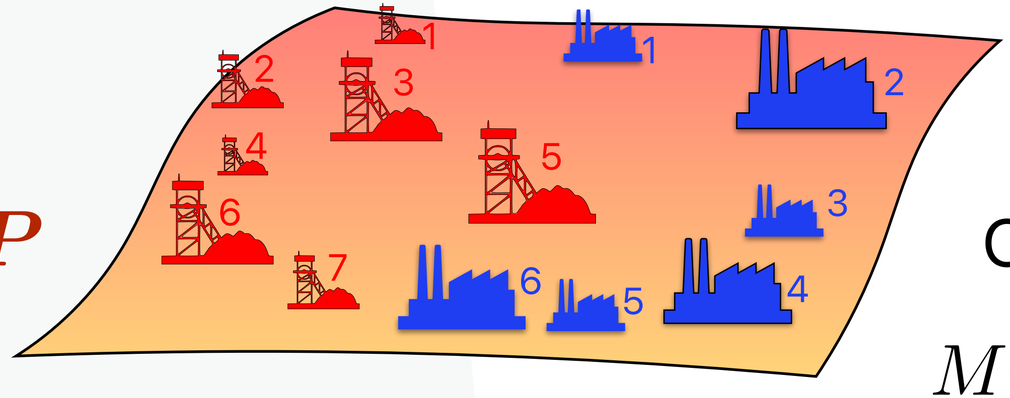




$$U(\boldsymbol{a},\boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

Mines and Factories Metaphor for Optimal Transport:

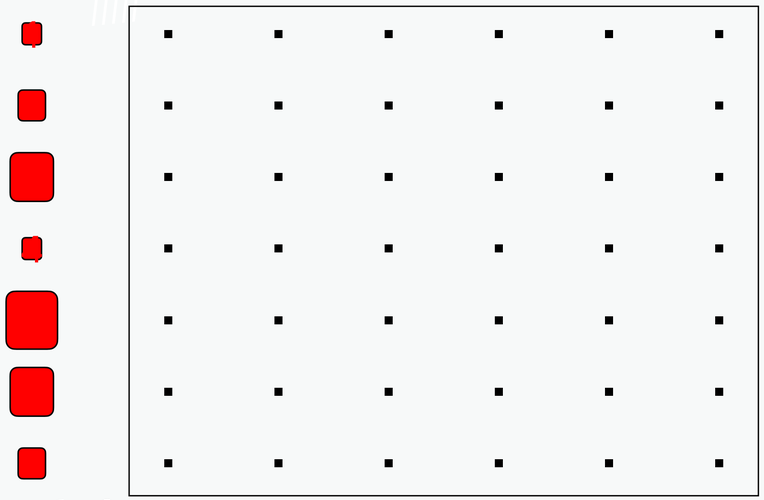
Couplings $\boldsymbol{P}$



$\boldsymbol{b}$



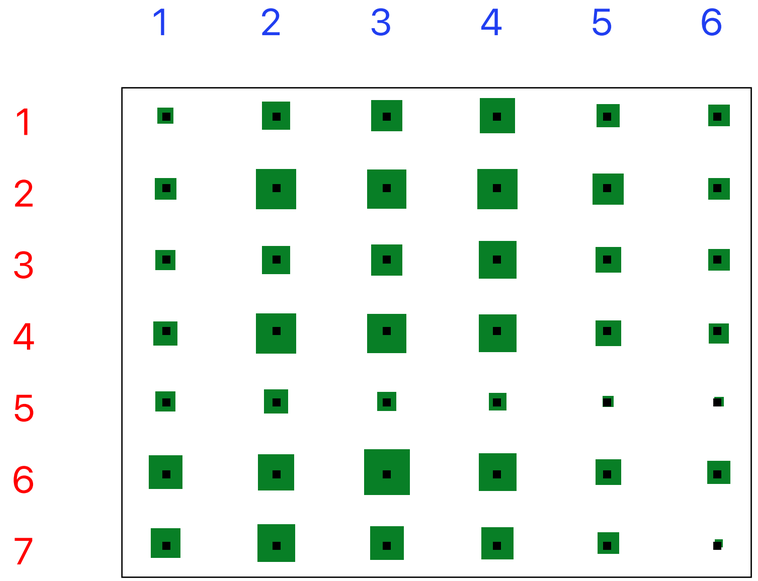
$\boldsymbol{a}$



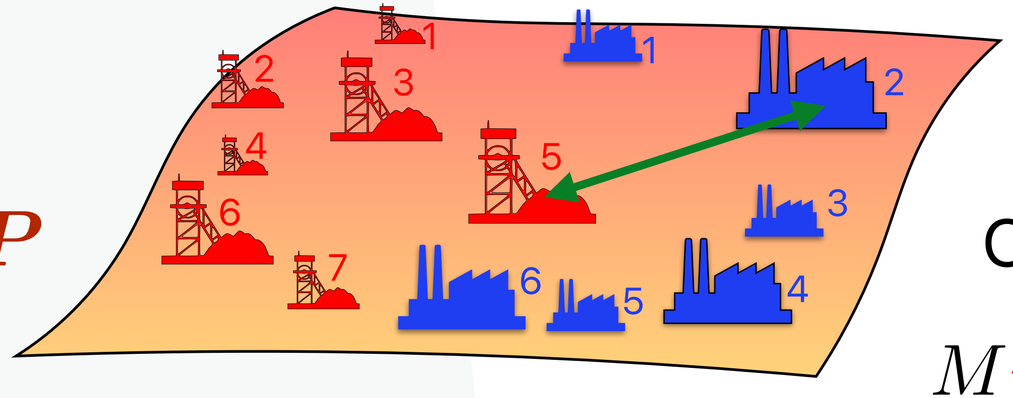
$$U(\boldsymbol{a},\boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

Cost matrix $M_{\boldsymbol{X}\boldsymbol{Y}}$

$$M_{\boldsymbol{X}\boldsymbol{Y}} \stackrel{\text{def}}{=} [\boldsymbol{c}(\boldsymbol{x}_i, \boldsymbol{y}_j)]_{ij}$$

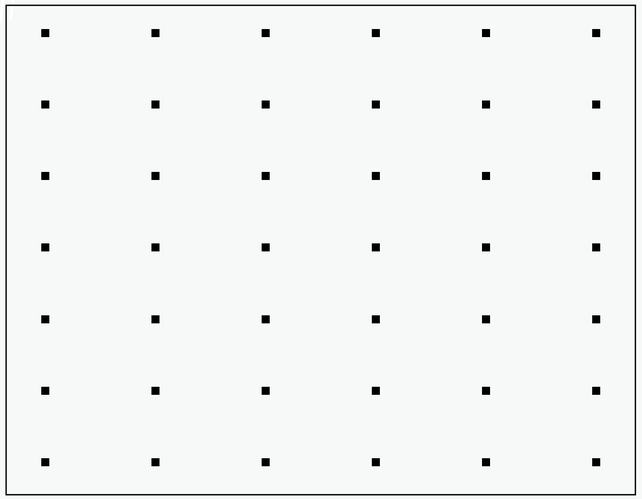


Mines and Factories Metaphor for Optimal Transport:



Couplings $\boldsymbol{P}$

$\boldsymbol{b}$

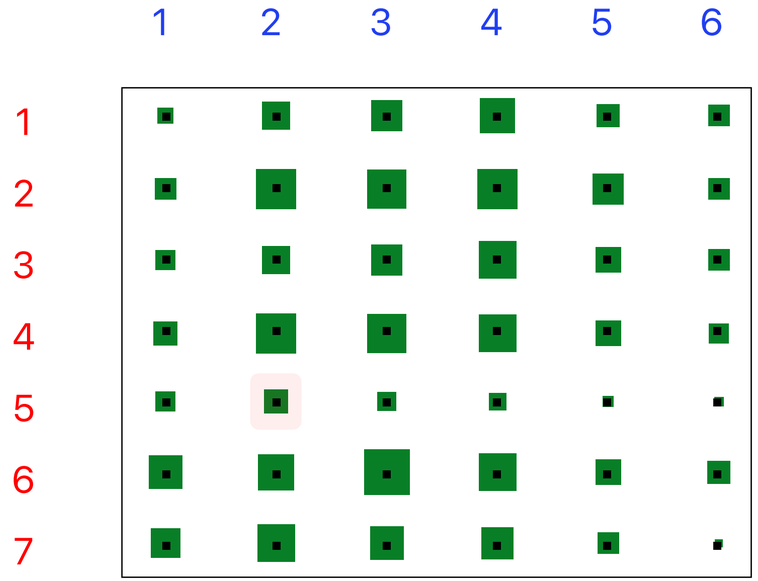


$\boldsymbol{a}$

$$U(\boldsymbol{a},\boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

Cost matrix $M_{\boldsymbol{X}\boldsymbol{Y}}$

$$M_{\boldsymbol{X}\boldsymbol{Y}} \stackrel{\text{def}}{=} [\boldsymbol{c}(\boldsymbol{x}_i, \boldsymbol{y}_j)]_{ij}$$

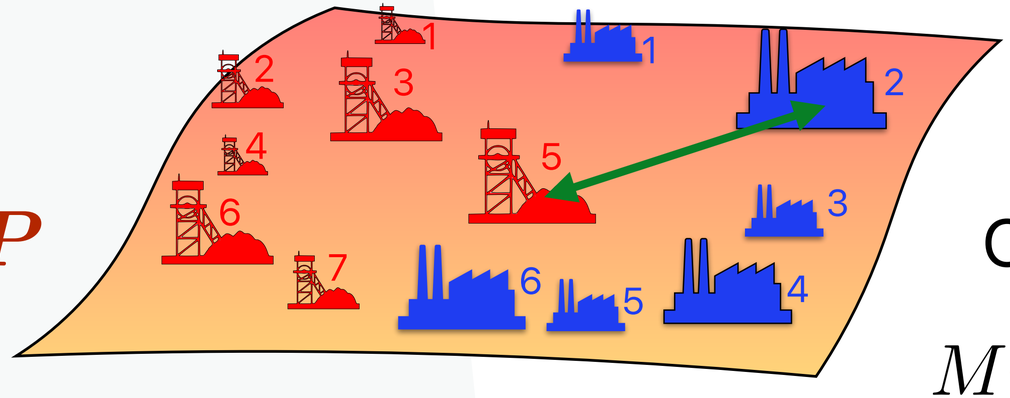


$\boldsymbol{x}_i$

$\boldsymbol{y}_j$

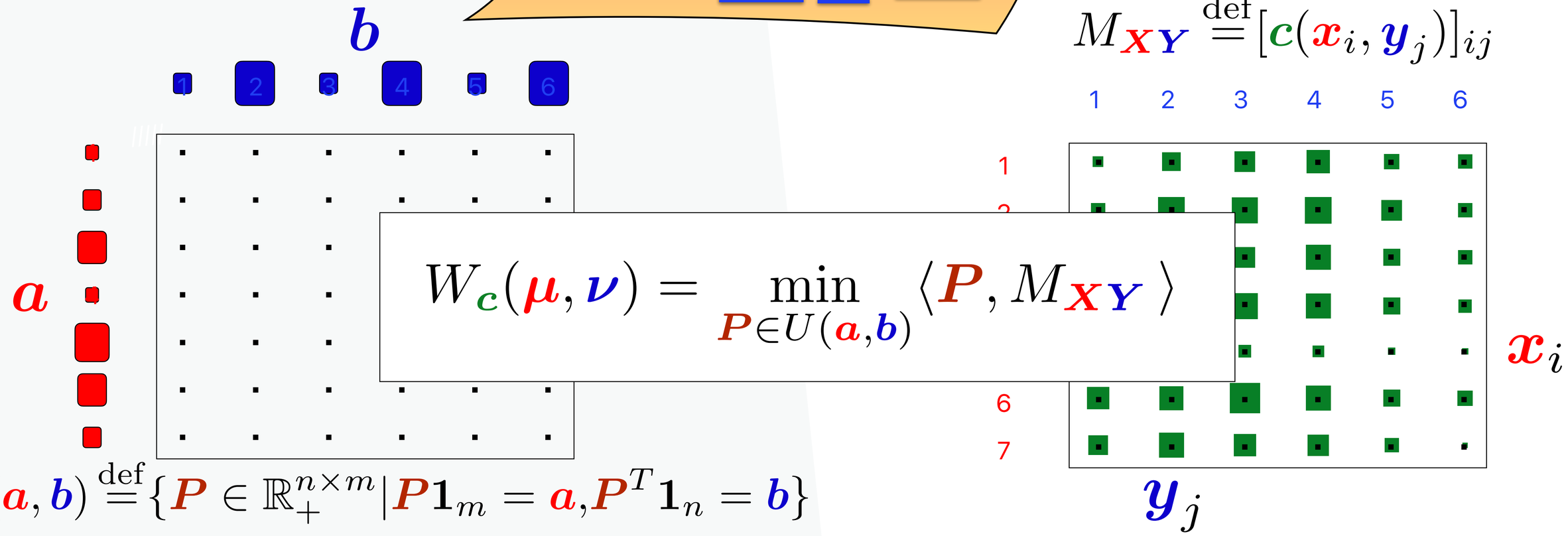
Mines and Factories Metaphor for Optimal Transport:

Couplings P



Cost matrix M_{XY}

$$M_{XY} \stackrel{\text{def}}{=} [c(x_i, y_j)]_{ij}$$



$$W_c(\mu, \nu) = \min_{P \in U(a,b)} \langle P, M_{XY} \rangle$$

$$U(a, b) \stackrel{\text{def}}{=} \{P \in \mathbb{R}_+^{n \times m} | P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\}$$

From Couplings to Monge through Duality

$$\min_{\boldsymbol{P} \in \mathbb{R}_{+}^{n \times m}} \quad \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$
$$\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b}$$

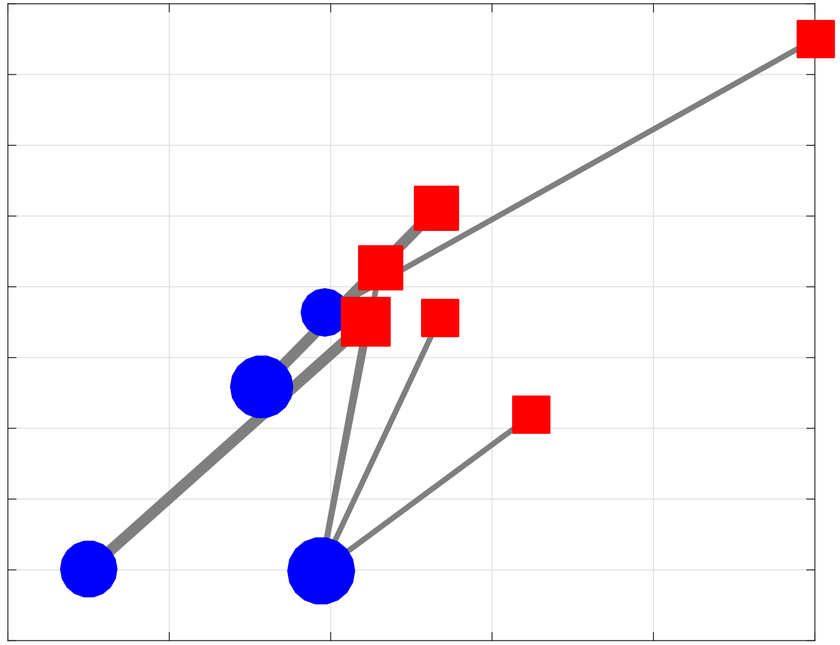
From Couplings to Monge through Duality

$$\min_{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}} \quad \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$
$$\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}$$

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{\boldsymbol{n}},\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{m}}} \quad \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b}$$
$$\boldsymbol{\alpha}_i + \boldsymbol{\beta}_j \leq c(\boldsymbol{x}_i, \boldsymbol{y}_j)$$

From Couplings to Monge through Duality

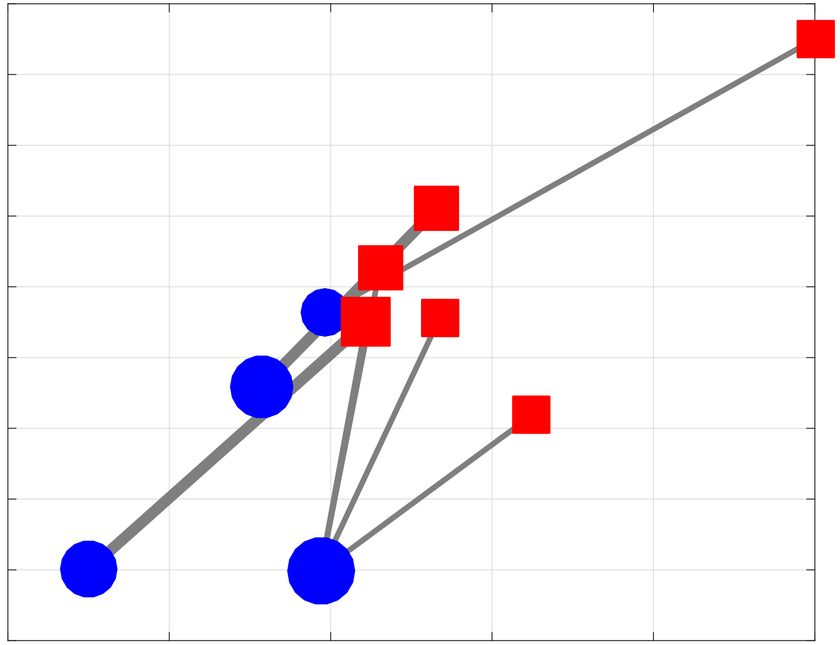
$$\min_{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}} \quad \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$
$$\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}$$



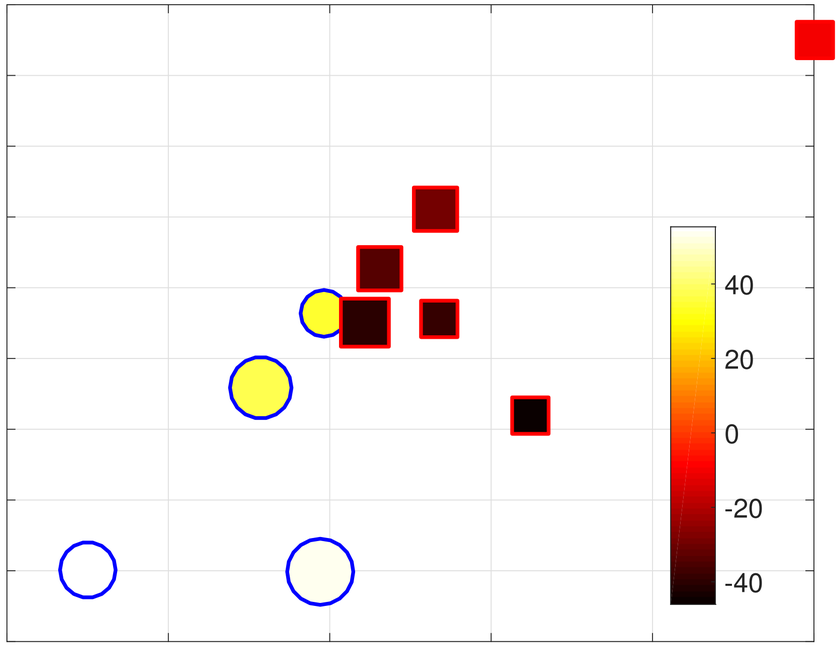
$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{\boldsymbol{n}},\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{m}}} \quad \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b}$$
$$\boldsymbol{\alpha}_i + \boldsymbol{\beta}_j \leq c(\boldsymbol{x}_i, \boldsymbol{y}_j)$$

From Couplings to Monge through Duality

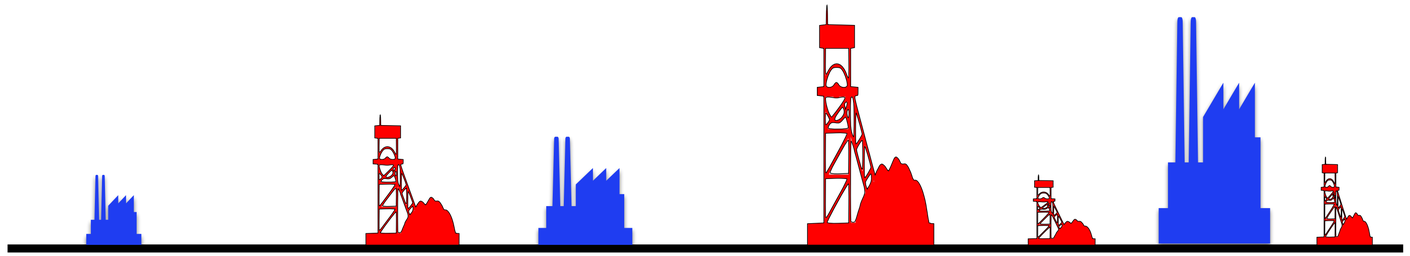
$$\min_{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}} \quad \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$
$$\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}$$



$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^{\boldsymbol{n}},\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{m}}} \quad \boldsymbol{\alpha}^T\boldsymbol{a} + \boldsymbol{\beta}^T\boldsymbol{b}$$
$$\boldsymbol{\alpha}_i + \boldsymbol{\beta}_j \leq c(\boldsymbol{x}_i, \boldsymbol{y}_j)$$



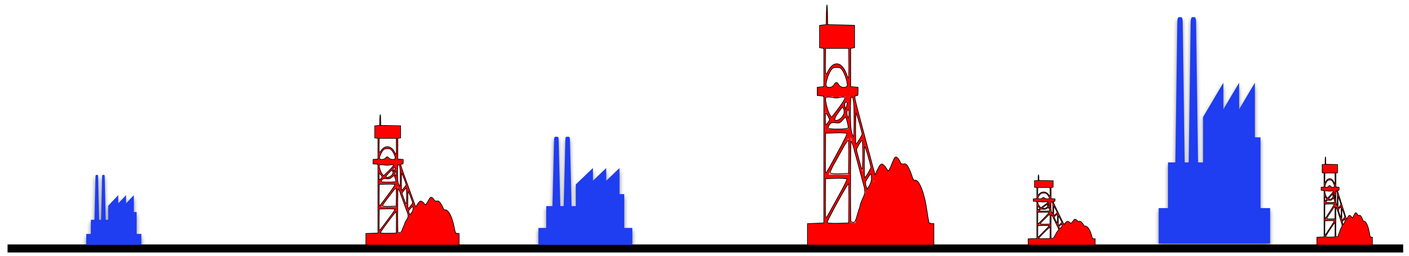
Easy Case: Mines and Factories live in a 1D World



Easy Case: Mines and Factories live in a 1D World

Optimal Transport is **solved in 1D** using **sorting:**
"Repeat until termination: the leftmost mine transfers
as much as it can to the leftmost factory still in need of resource."

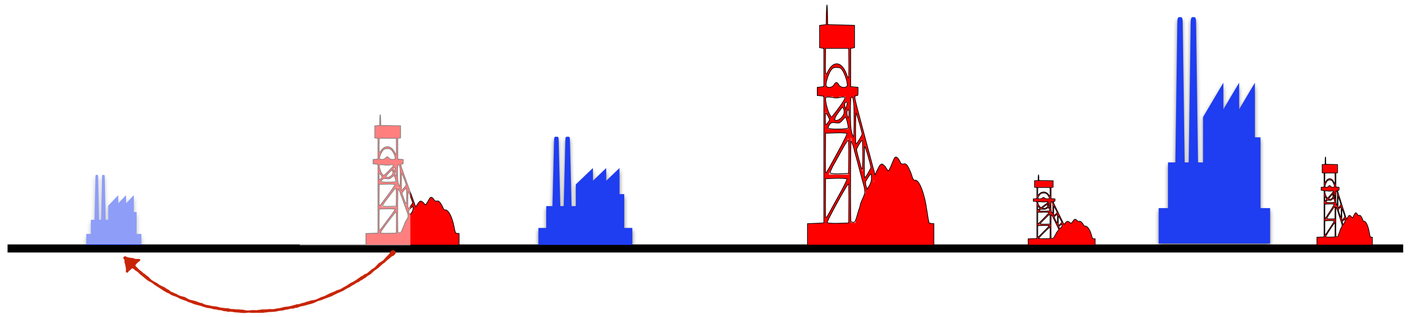
Computation: *O(n log n + m log m)*



Easy Case: Mines and Factories live in a 1D World

Optimal Transport is **solved in 1D** using **sorting:**
"Repeat until termination: the leftmost mine transfers
as much as it can to the leftmost factory still in need of resource."

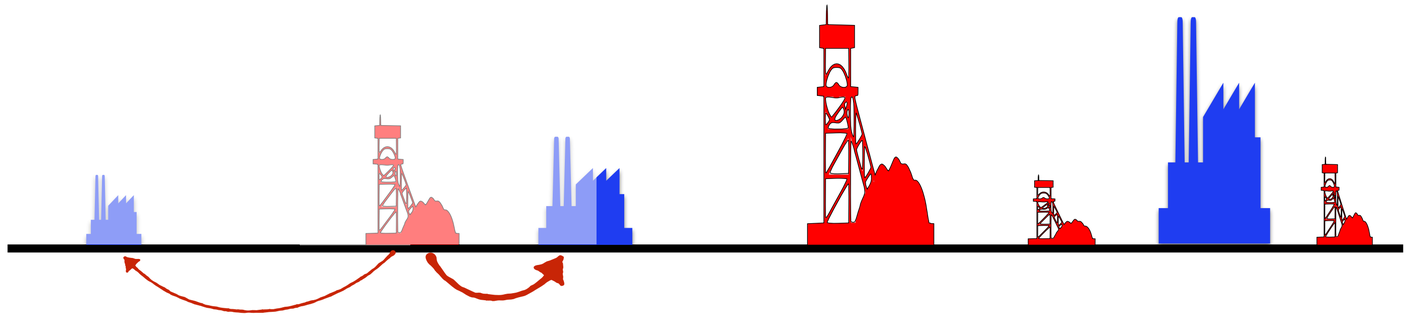
Computation: $O(n \log n + m \log m)$



Easy Case: Mines and Factories live in a 1D World

Optimal Transport is **solved in 1D** using **sorting:**
"Repeat until termination: the leftmost mine transfers
as much as it can to the leftmost factory still in need of resource."

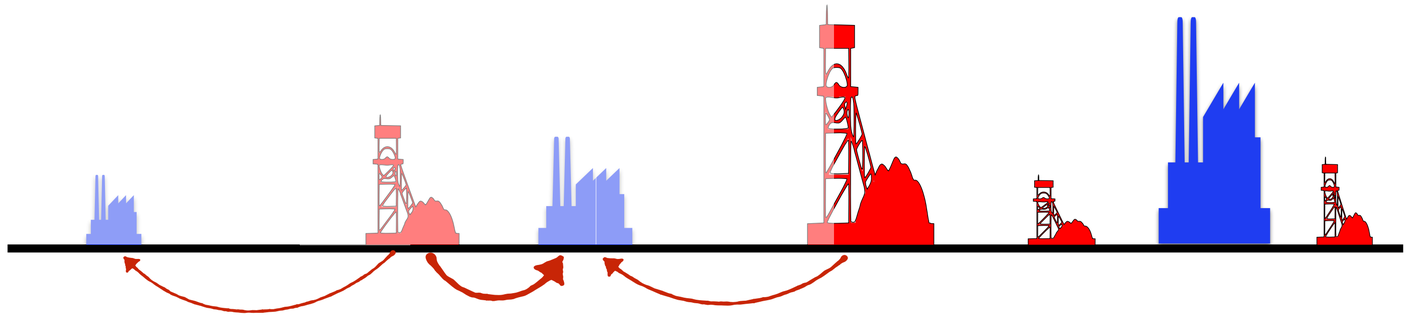
Computation: $O(n \log n + m \log m)$



Easy Case: Mines and Factories live in a 1D World

Optimal Transport is **solved in 1D** using **sorting:**
"Repeat until termination: the leftmost mine transfers
as much as it can to the leftmost factory still in need of resource."

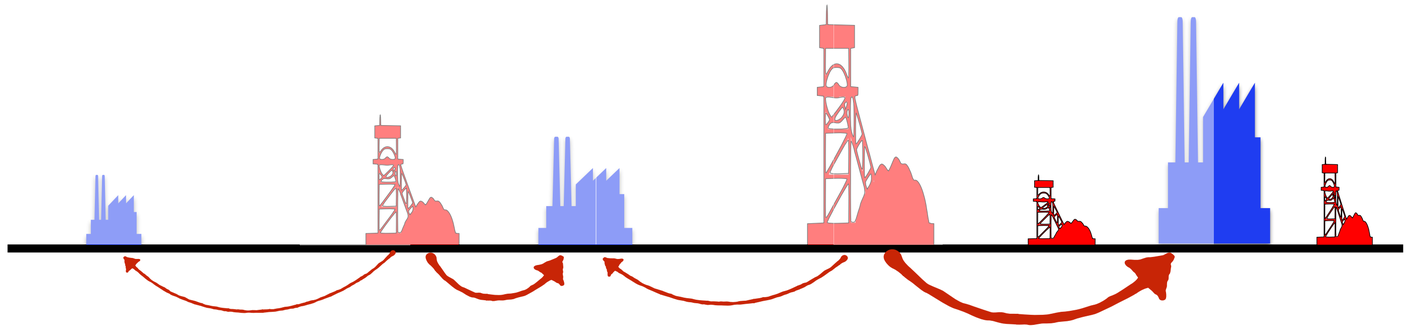
Computation: $O(n \log n + m \log m)$



Easy Case: Mines and Factories live in a 1D World

Optimal Transport is **solved in 1D** using **sorting:**
"Repeat until termination: the leftmost mine transfers
as much as it can to the leftmost factory still in need of resource."

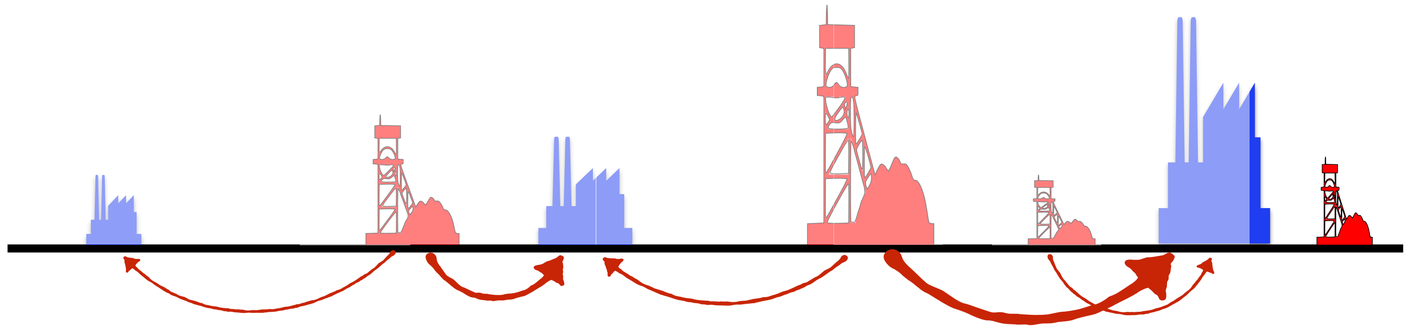
Computation: $O(n \log n + m \log m)$



Easy Case: Mines and Factories live in a 1D World

Optimal Transport is **solved in 1D** using **sorting:**
"Repeat until termination: the leftmost mine transfers
as much as it can to the leftmost factory still in need of resource."

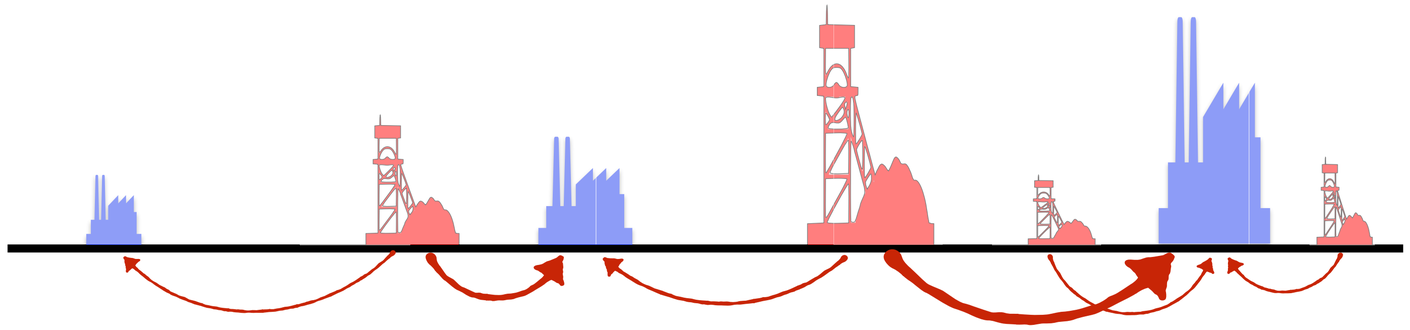
Computation: $O(n \log n + m \log m)$



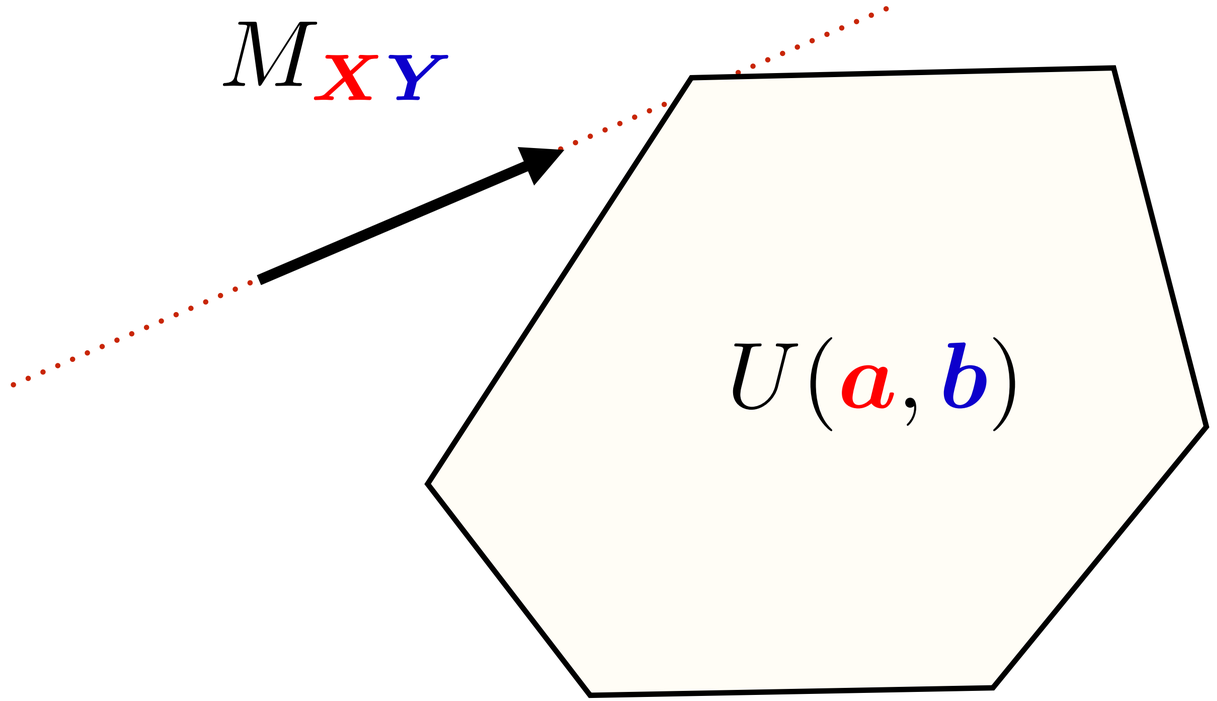
Easy Case: Mines and Factories live in a 1D World

Optimal Transport is **solved in 1D** using **sorting:**
"Repeat until termination: the leftmost mine transfers
as much as it can to the leftmost factory still in need of resource."

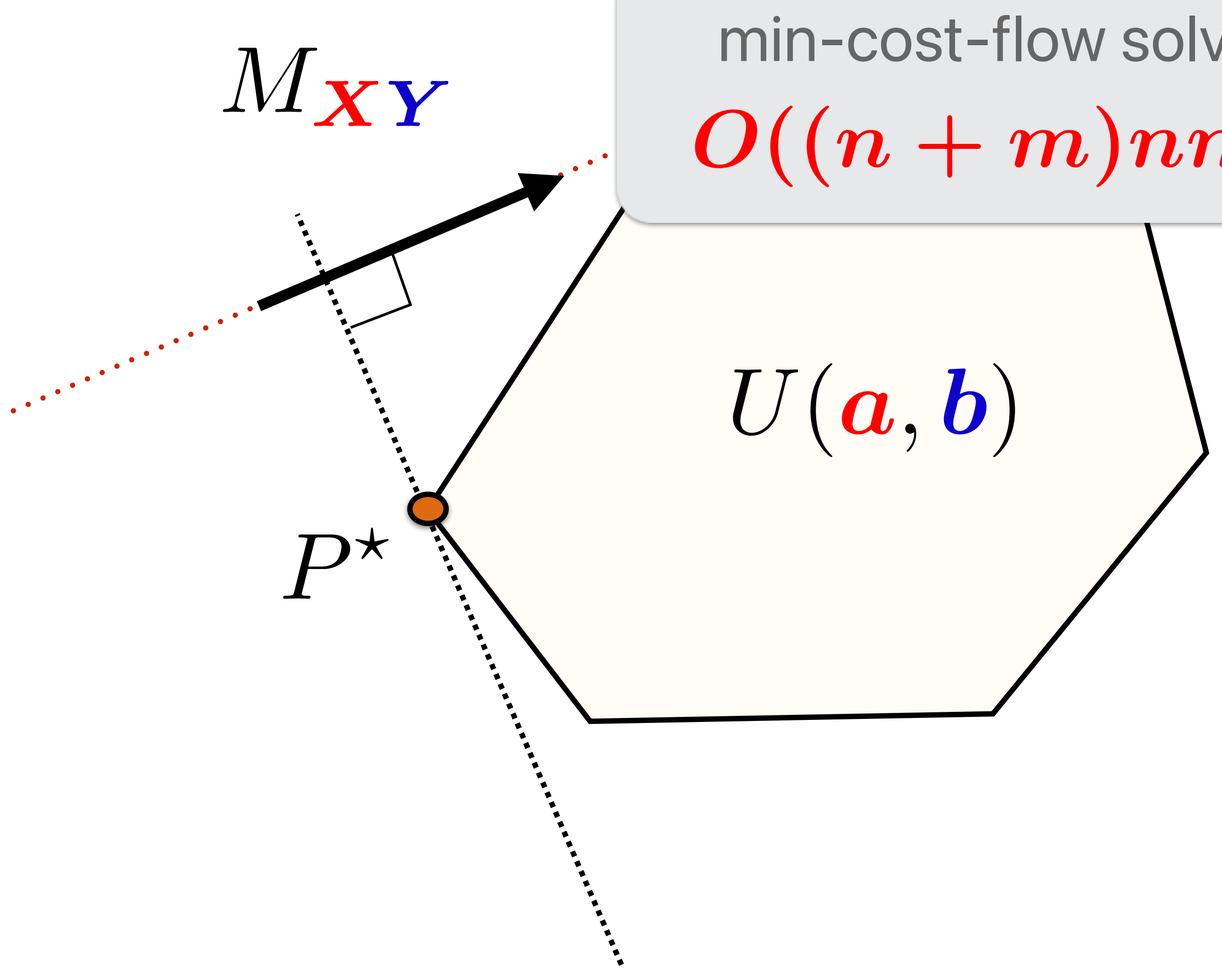
Computation: $O(n \log n + m \log m)$



Solving the Discrete OT Problem in Full Generality



Solving the Discrete OT Problem in Full Generality

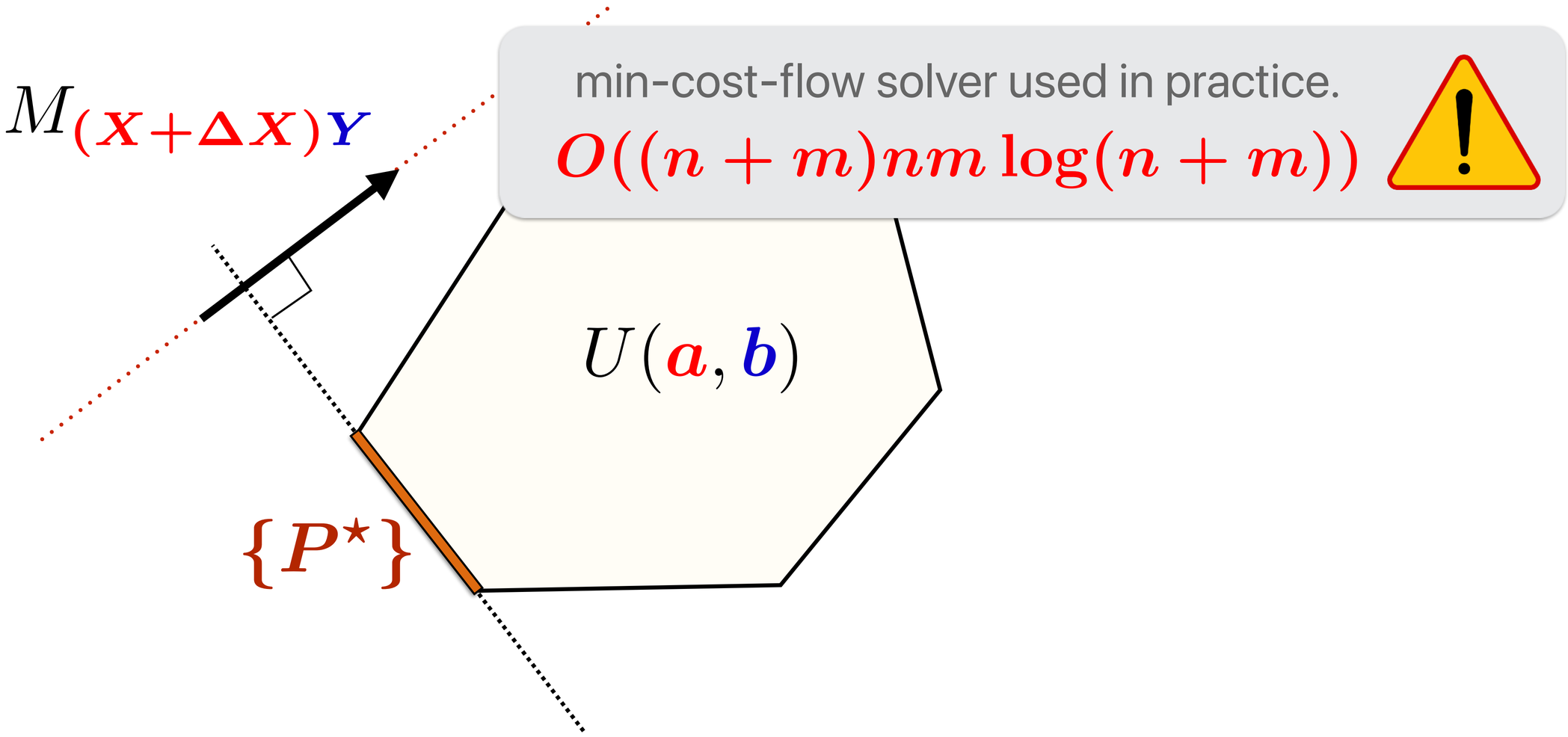


min-cost-flow solver used in practice.

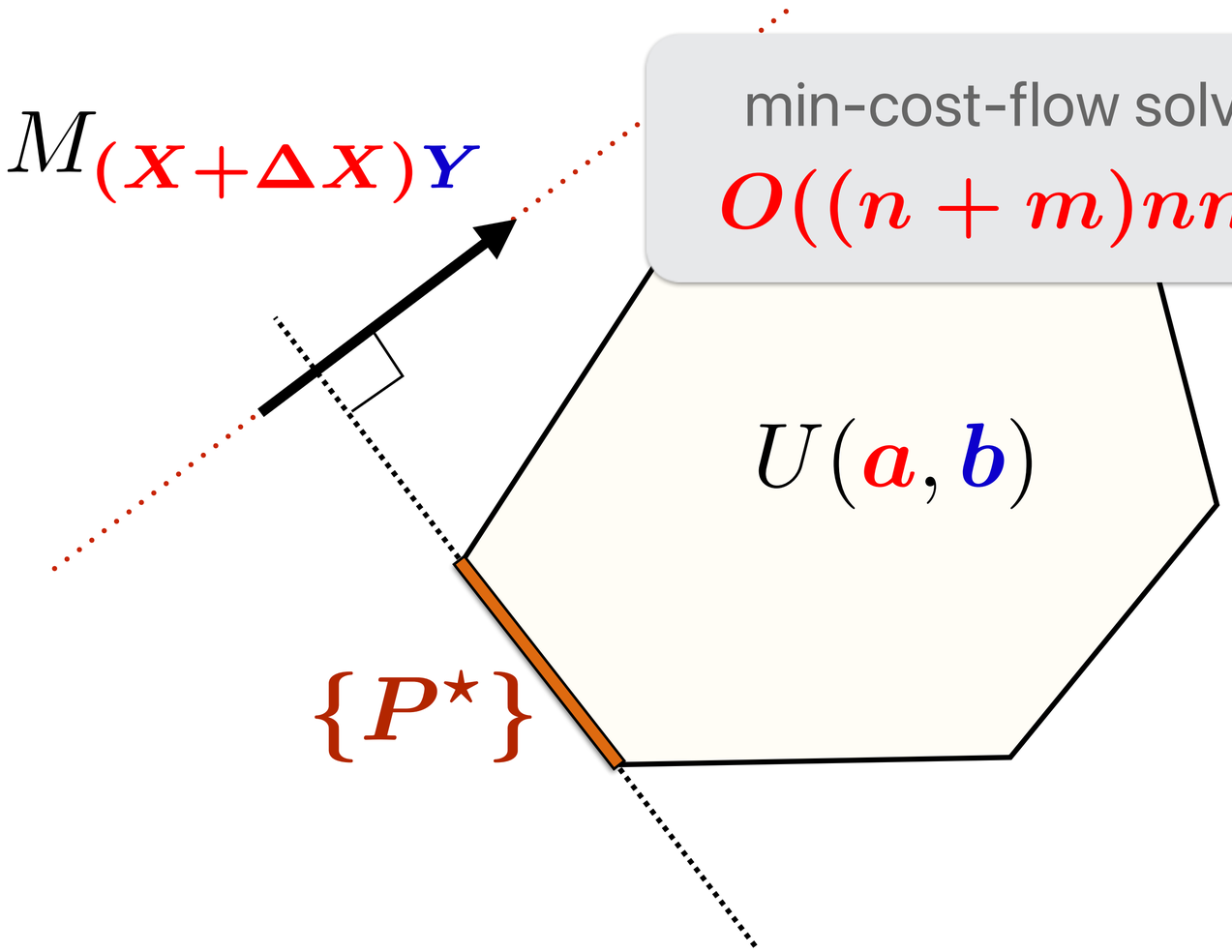
$$O((n+m)nm\log(n+m))$$



Solving the Discrete OT Problem in Full Generality



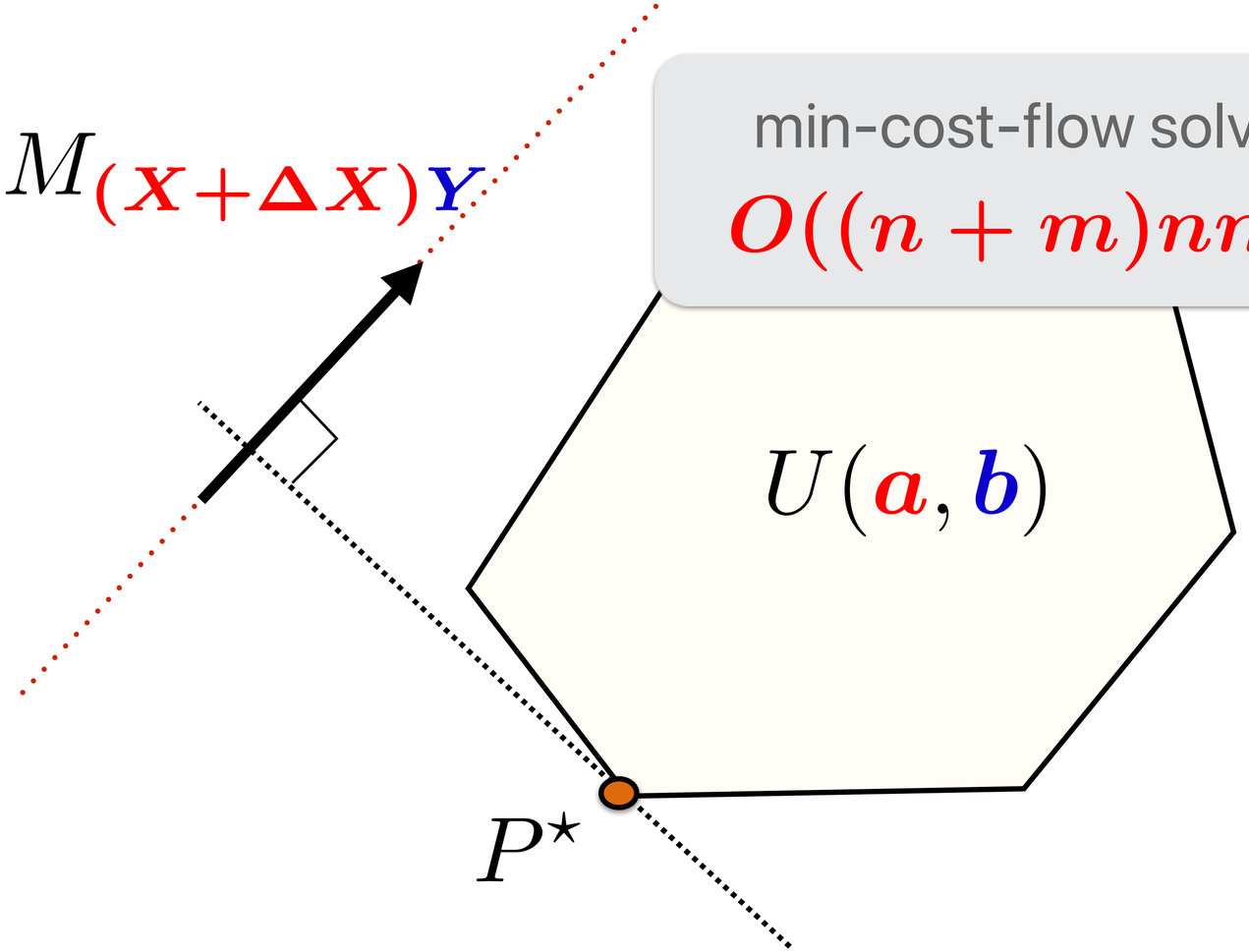
Solving the Discrete OT Problem in Full Generality



min-cost-flow solver used in practice.

$$O((n+m)nm\log(n+m))$$

Solving the Discrete OT Problem in Full Generality

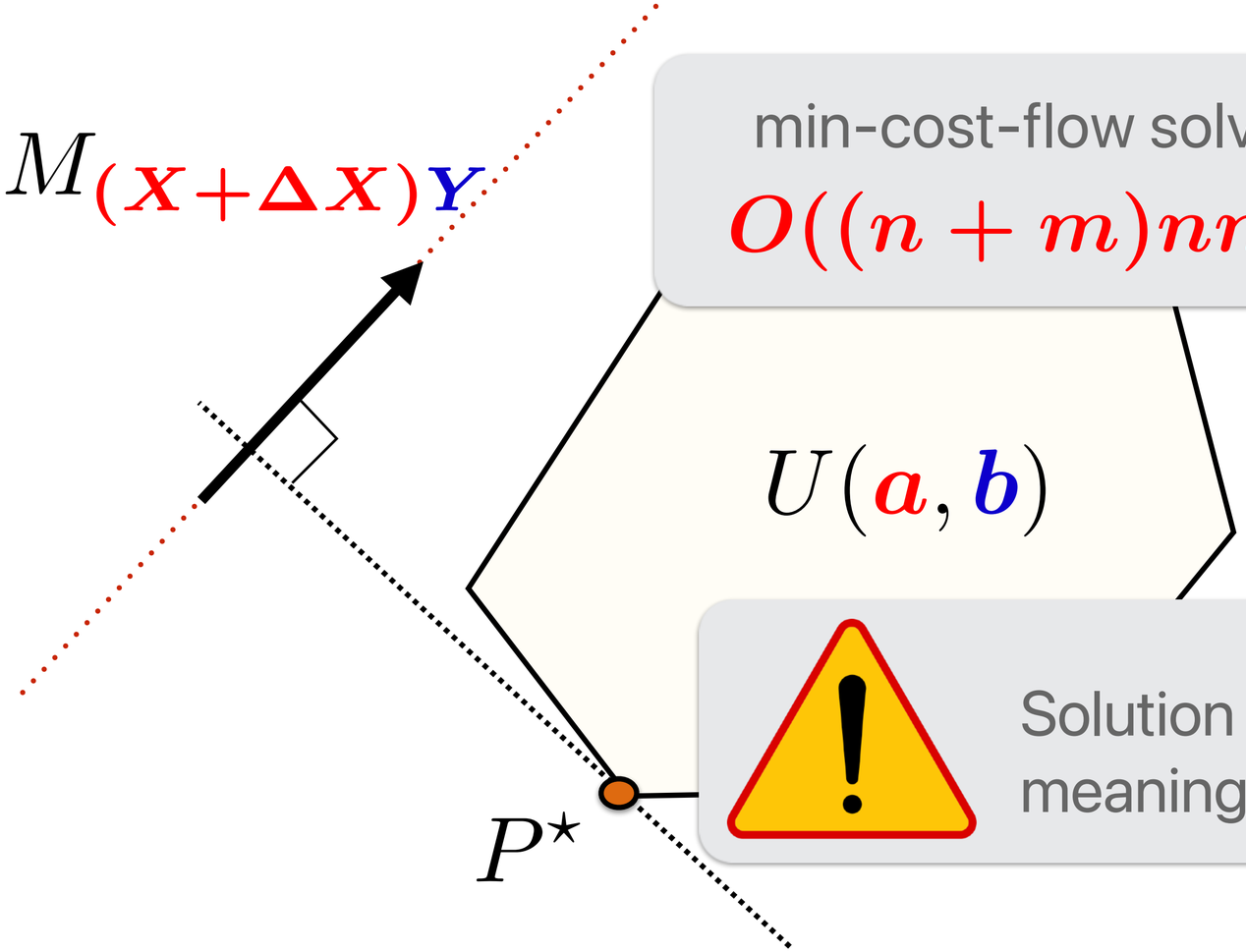


min-cost-flow solver used in practice.

$$O((n+m)nm\log(n+m))$$



Solving the Discrete OT Problem in Full Generality

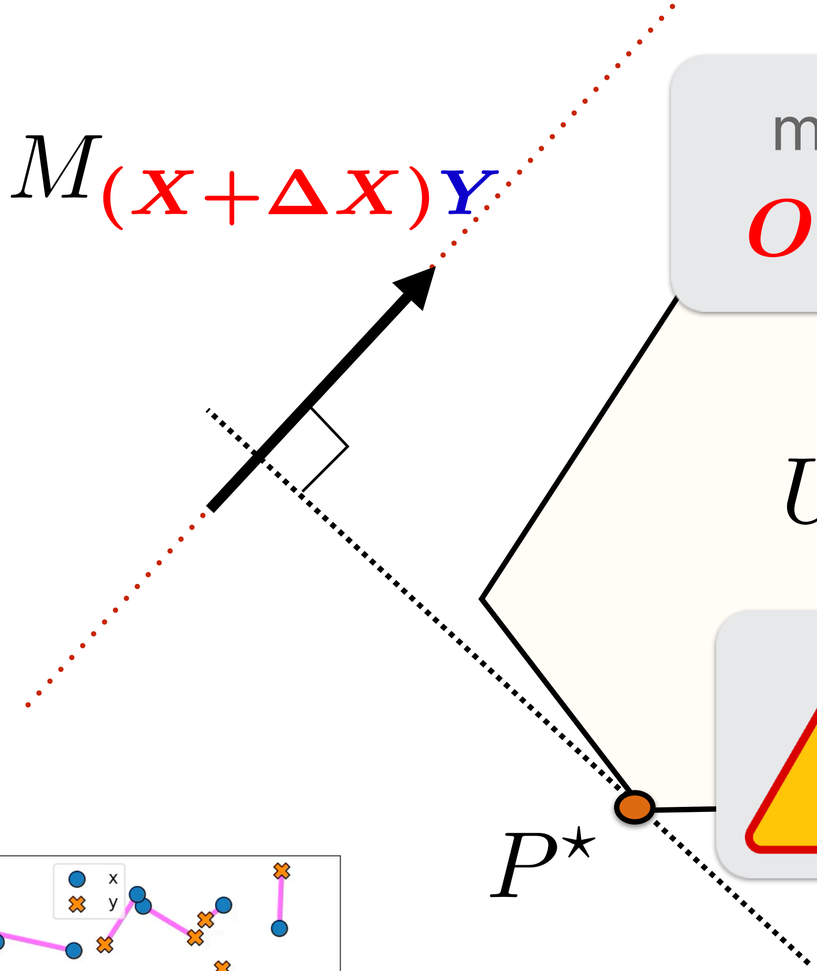


min-cost-flow solver used in practice.

$$O((n+m)nm\log(n+m))$$

Solution $P^\star$ is not unique, and has no meaningful Jacobian w.r.t input *X* or *Y*.

Solving the Discrete OT Problem in Full Generality



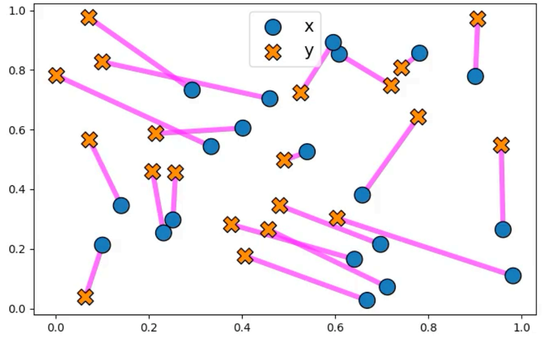
min-cost-flow solver used in practice.

$$O((n+m)nm\log(n+m))$$

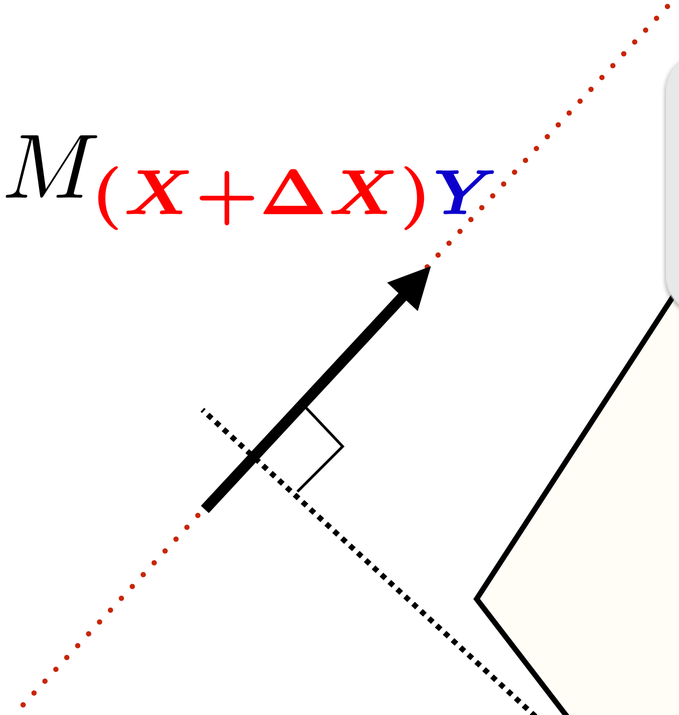




Solution $P^\star$ is not unique, and has no meaningful Jacobian w.r.t input *X* or *Y*.



Solving the Discrete OT Problem in Full Generality



min-cost-flow solver used in practice.

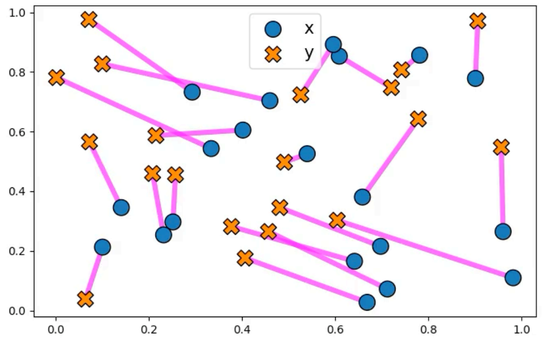
$$O((n+m)nm\log(n+m))$$





Solution $P^\star$ is not unique, and has no meaningful Jacobian w.r.t input *X* or *Y*.

$W_c(\mu, \nu)$ not differentiable



In Higher Dimensions... What matters for practitioners?

$$\text{i.i.d samples } \boldsymbol{x_1, \dots, x_n \sim \mu},\ \boldsymbol{y_1, \dots, y_m \sim \nu},$$
$$\boldsymbol{\hat{\mu}_n} \overset{\text{def}}{=} \frac{1}{n}\sum_i \delta_{\boldsymbol{x_i}},\ \boldsymbol{\hat{\nu}_m} \overset{\text{def}}{=} \frac{1}{m}\sum_j \delta_{\boldsymbol{y_j}}$$

Computational properties

Compute / approximate $\boldsymbol{W_c(\hat{\mu}_n, \hat{\nu}_m)}$

Statistical properties

$$\mathbb{E}\left[\left|\boldsymbol{W_c(\mu, \nu)} - \boldsymbol{W_c(\hat{\mu}_n, \hat{\nu}_m)}\right|\right] \leq f(n, m)?$$

What matters for practitioners?

$$\text{i.i.d samples } \boldsymbol{x_1, \dots, x_n \sim \mu},\ \boldsymbol{y_1, \dots, y_m \sim \nu},$$
$$\boldsymbol{\hat{\mu}_n} \overset{\text{def}}{=} \frac{1}{n} \sum_i \delta_{\boldsymbol{x_i}}, \boldsymbol{\hat{\nu}_m} \overset{\text{def}}{=} \frac{1}{m} \sum_j \delta_{\boldsymbol{y_j}}$$



Computational properties

$$\boldsymbol{O((n+m)nm\log(n+m))}$$

Statistical properties

$$\mathbb{E}\left[\left|\boldsymbol{W_p(\mu,\nu)} - \boldsymbol{W_p(\hat{\mu}_n, \hat{\nu}_m)}\right|\right] \leq f(n,m)?$$

What matters for practitioners?

$$\text{i.i.d samples } \boldsymbol{x_1, \dots, x_n \sim \mu},\ \boldsymbol{y_1, \dots, y_m \sim \nu},$$
$$\boldsymbol{\hat{\mu}_n} \overset{\text{def}}{=} \frac{1}{n}\sum_i \delta_{\boldsymbol{x_i}},\ \boldsymbol{\hat{\nu}_m} \overset{\text{def}}{=} \frac{1}{m}\sum_j \delta_{\boldsymbol{y_j}}$$

Computational properties

$$\boldsymbol{O((n+m)nm\log(n+m))}$$



Statistical properties

$$\mathbb{E}\left[\left|\boldsymbol{W_c(\mu,\nu)} - \boldsymbol{W_c(\hat{\mu}_n,\hat{\nu}_n)}\right|\right] = O(n^{-1/d})$$



What matters for practitioners?

Computational properties

$$O((n+m)nm\log(n+m))$$



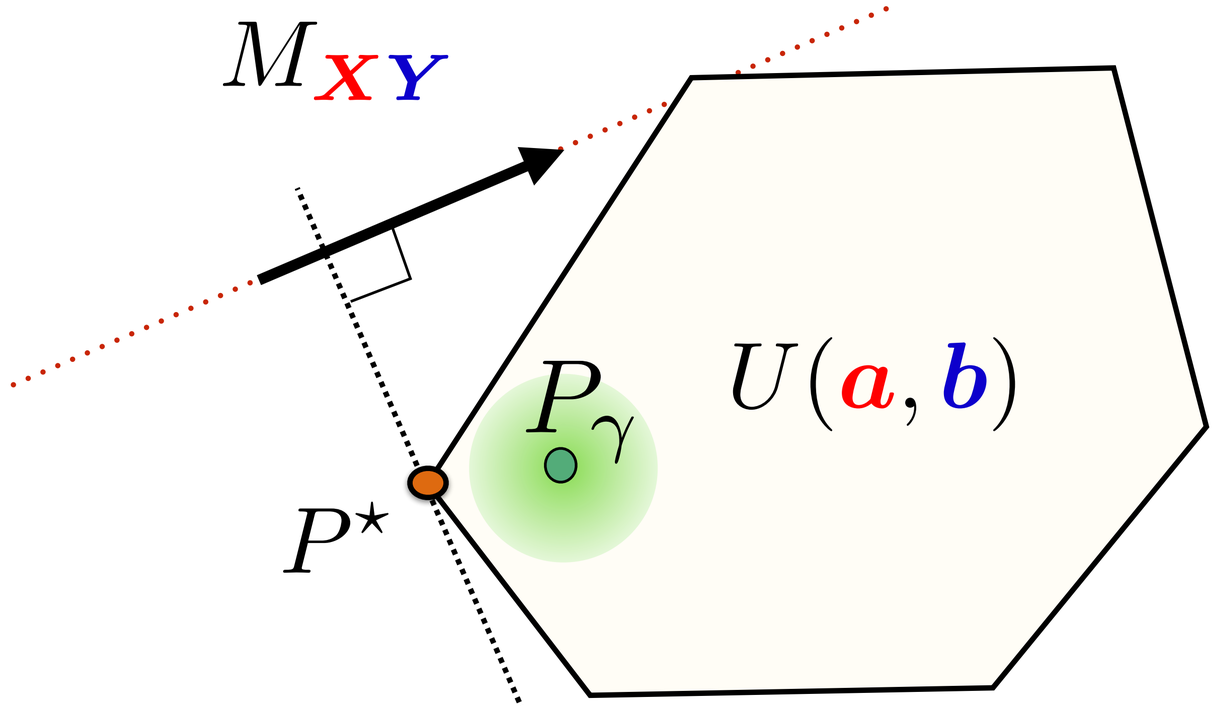
Statistical properties

$$\mathbb{E}\left[|\boldsymbol{W_c}(\boldsymbol{\mu},\boldsymbol{\nu}) - \boldsymbol{W_c}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}},\hat{\boldsymbol{\nu}}_{\boldsymbol{n}})|\right] = O(n^{-1/d})$$



In data sciences, we must regularize the matching problem (and OT problems more generally) to improve on either or both aspects.

Solution: Regularize Linear Program

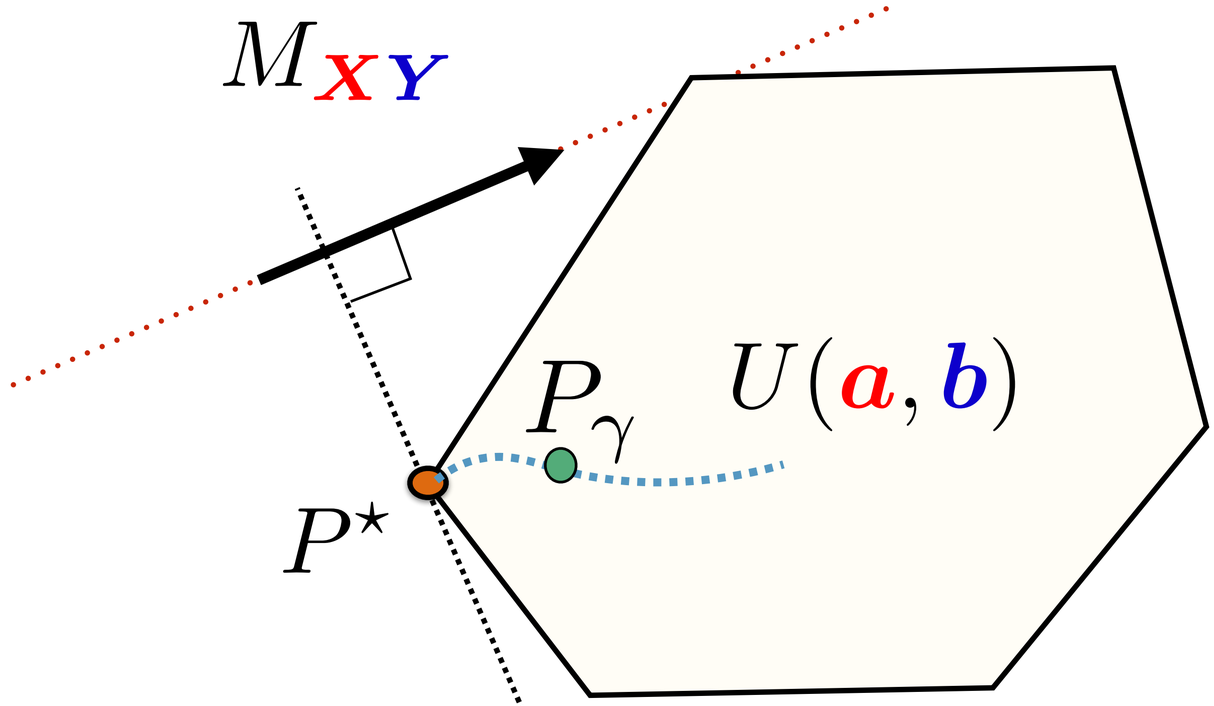


Why regularize?

- Computational efficiency
- Statistical regularity
- Robustness
- Differentiability
- Structure

$$P_\gamma \stackrel{\text{def}}{=} \underset{P \in U(a,b)}{\operatorname{argmin}} \langle P, M_{XY} \rangle + \gamma \operatorname{Reg}(P)$$

Solution: Regularize Linear Program



Why regularize?

- Computational efficiency
- Statistical regularity
- Robustness
- Differentiability
- Structure

Other: [Dessein+16] [Blondel+18]

$$P_\gamma \stackrel{\text{def}}{=} \underset{P \in U(a,b)}{\operatorname{argmin}} \langle P, M_{XY} \rangle - \gamma E(P)$$

Solving Regularised Optimal Transport

Def. Regularized Optimal Transport, $\gamma \geq 0$

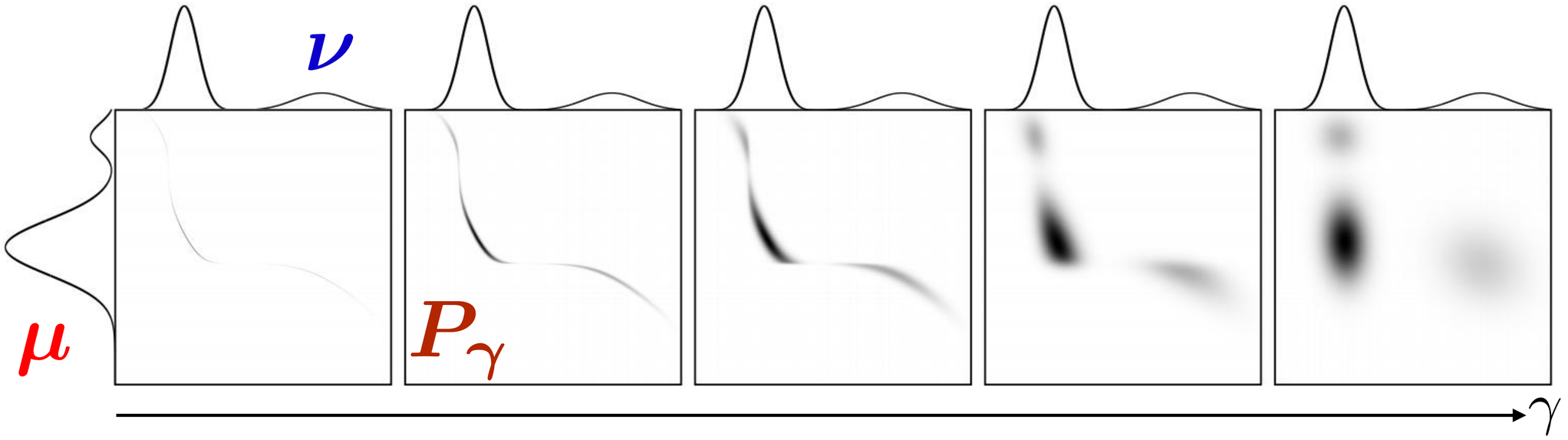
$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

$$E(P) \stackrel{\text{def}}{=} -\sum_{i,j=1}^{nm} P_{ij}(\log P_{ij} - 1)$$

Solving Regularised Optimal Transport

Def. Regularized Optimal Transport, $\gamma \geq 0$

$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$



Solving Regularised Optimal Transport

$$\textbf{Prop. If } P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

$$\text{then } \exists! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m, \text{ such that}$$

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

Solving Regularised Optimal Transport

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

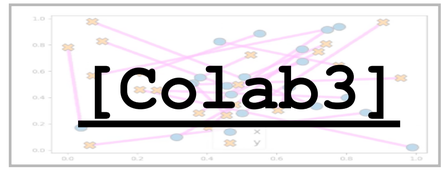
$$L(P, \alpha, \beta) = \sum_{ij} P_{ij} M_{ij} + \gamma P_{ij}(\log P_{ij} - 1) - \alpha^T (P\mathbf{1} - \boldsymbol{a}) - \beta^T (P^T \mathbf{1} - \boldsymbol{b})$$

$$\partial L / \partial P_{ij} = M_{ij} + \gamma \log P_{ij} - \alpha_i - \beta_j$$

$$(\boldsymbol{\partial L / \partial P_{ij}} = 0) \Rightarrow P_{ij} = \boldsymbol{e}^{\frac{\boldsymbol{\alpha_i}}{\boldsymbol{\gamma}}} \; e^{-\frac{M_{ij}}{\gamma}} \; \boldsymbol{e}^{\frac{\boldsymbol{\beta_j}}{\boldsymbol{\gamma}}} = \boldsymbol{u_i} \; K_{ij} \boldsymbol{v_j}$$

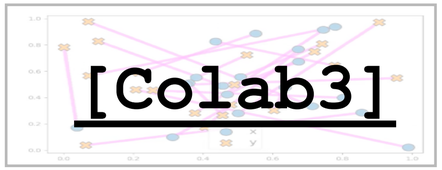
Solving Entropy Regularized OT

$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

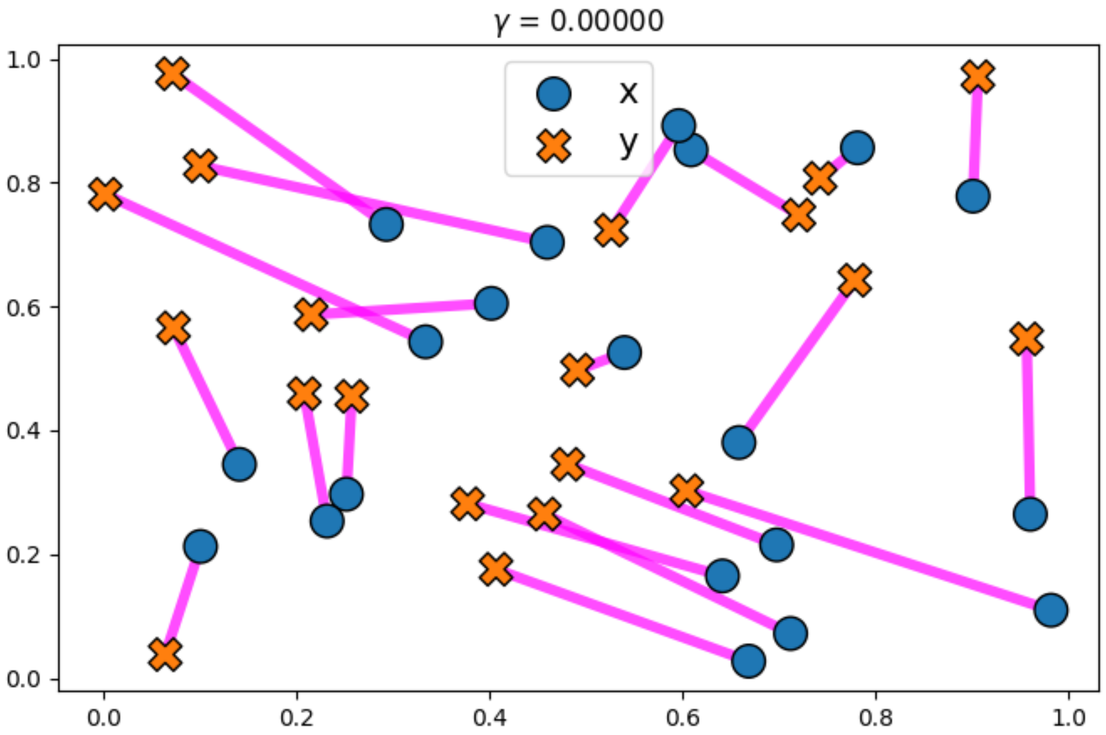


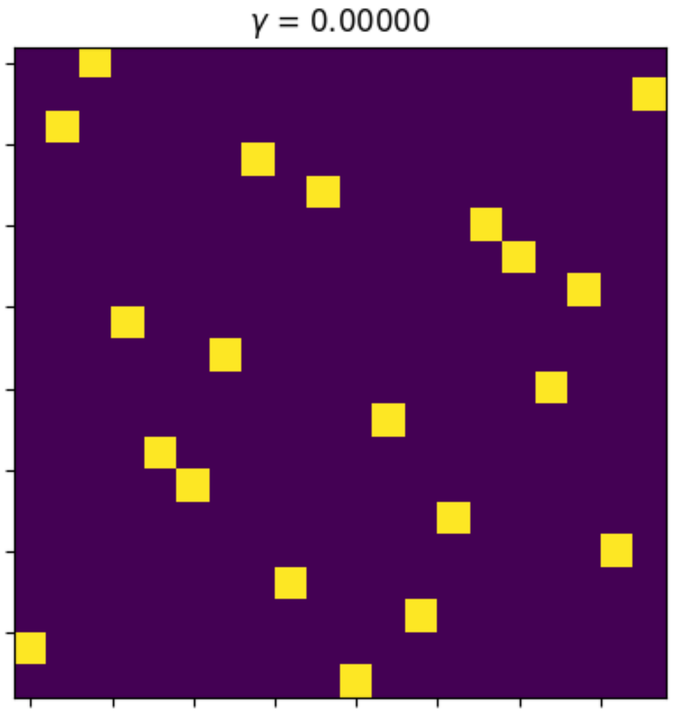
Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$



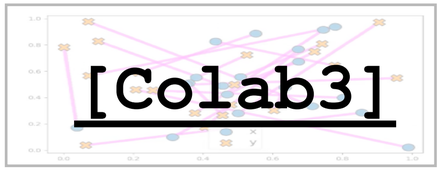
[Colab3]

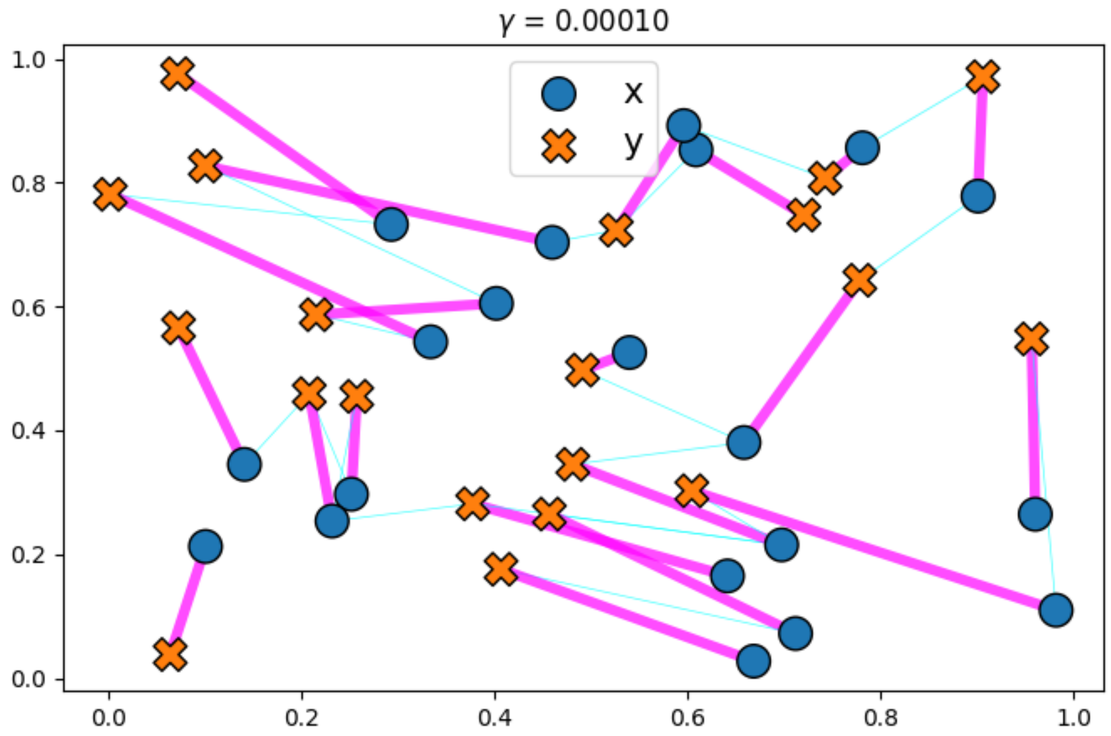


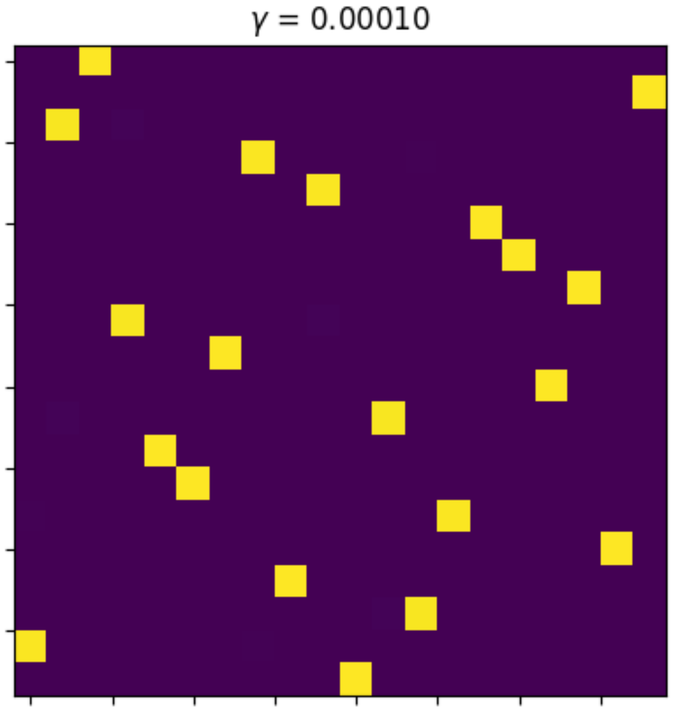


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

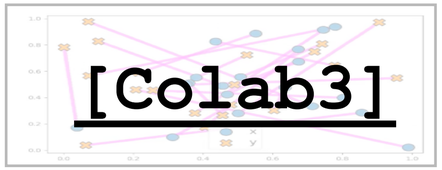


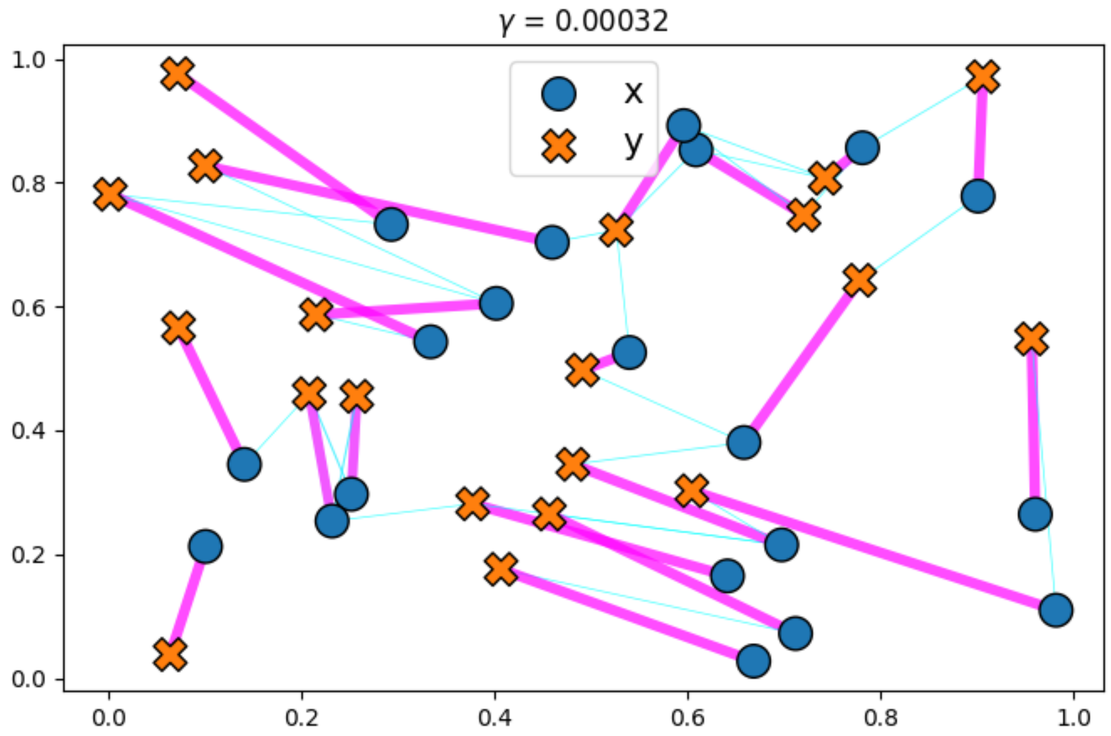


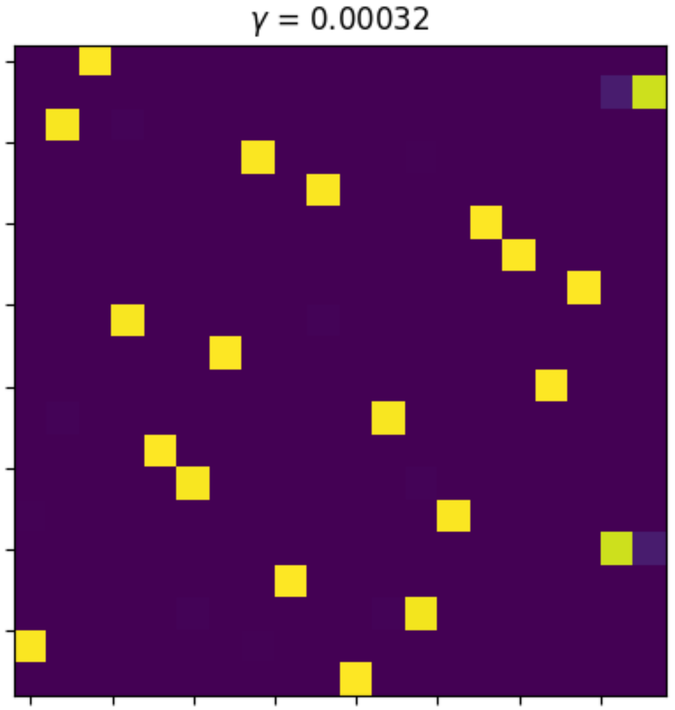


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

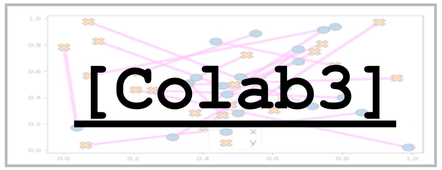


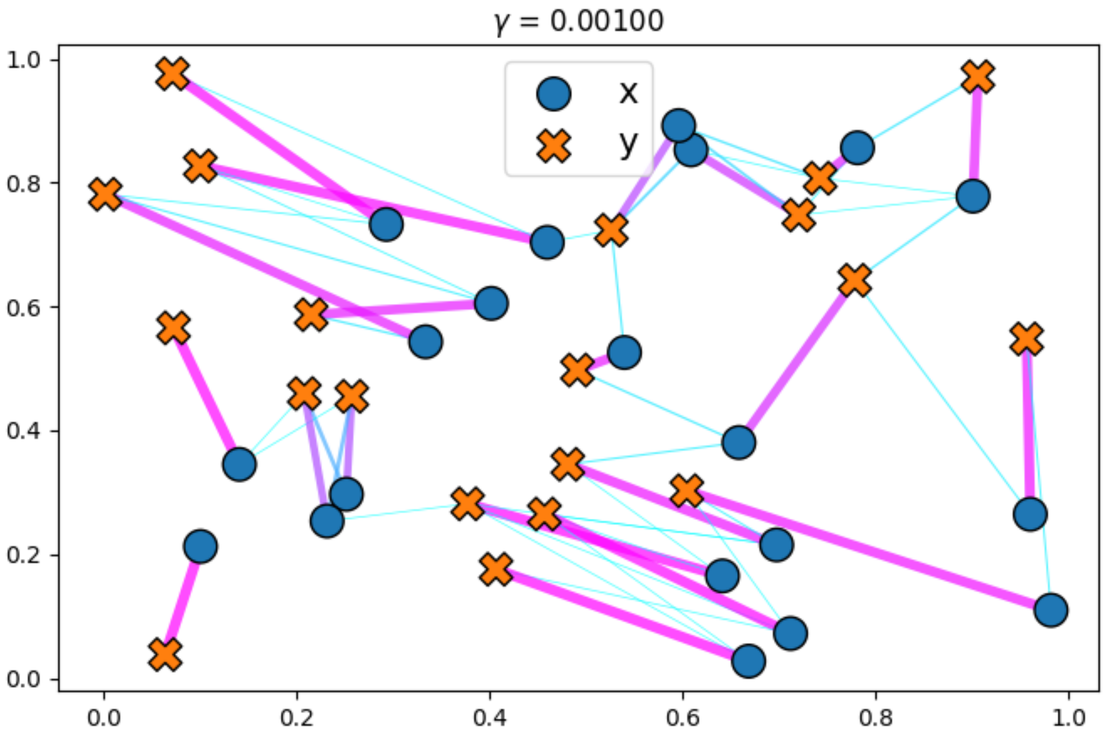


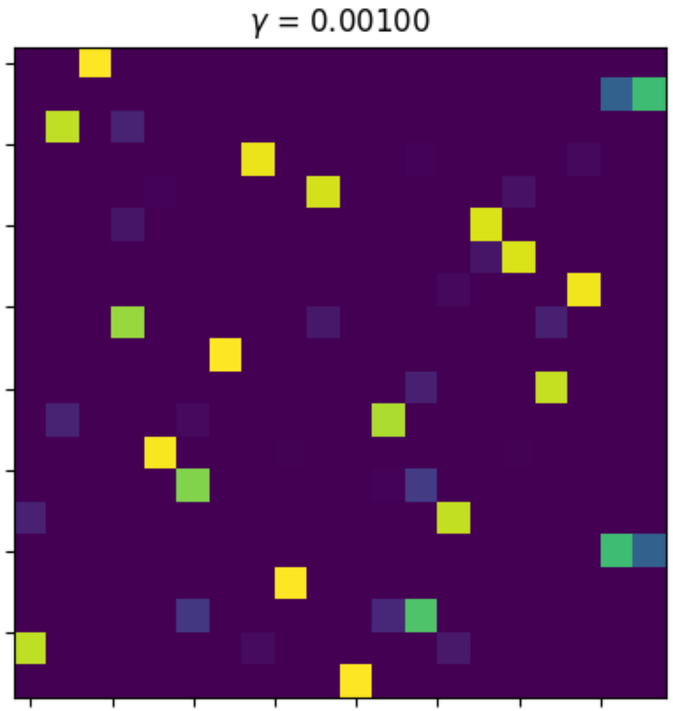


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

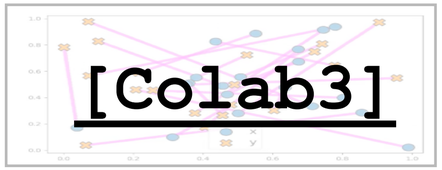




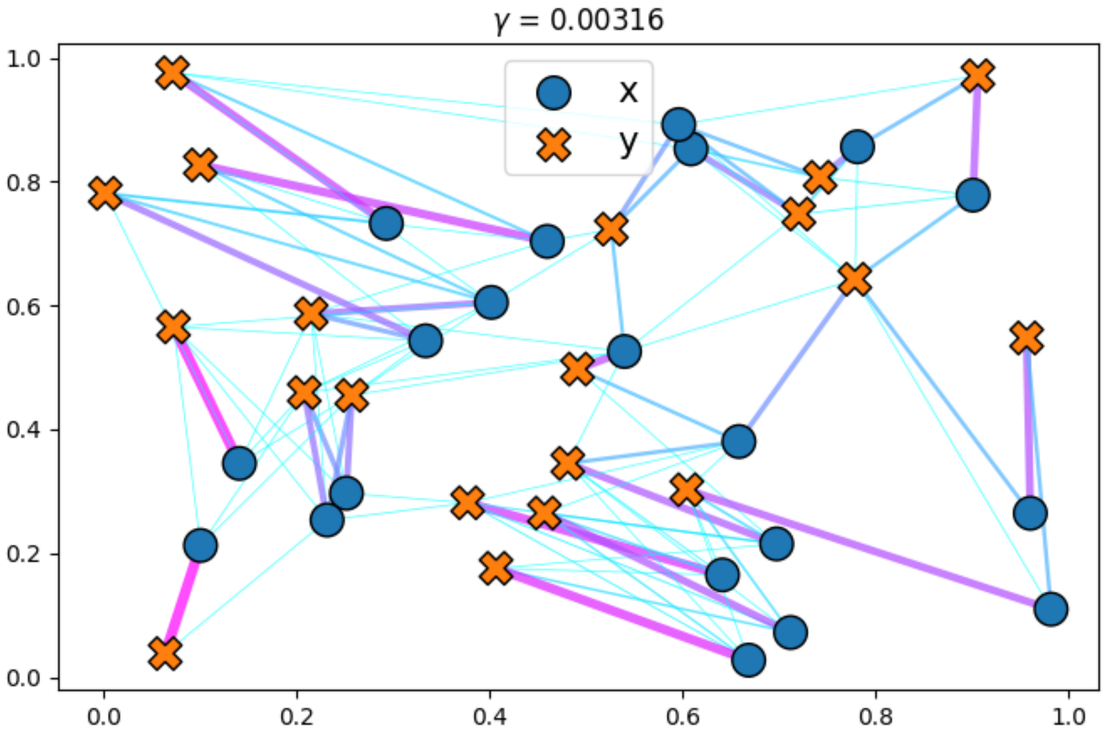


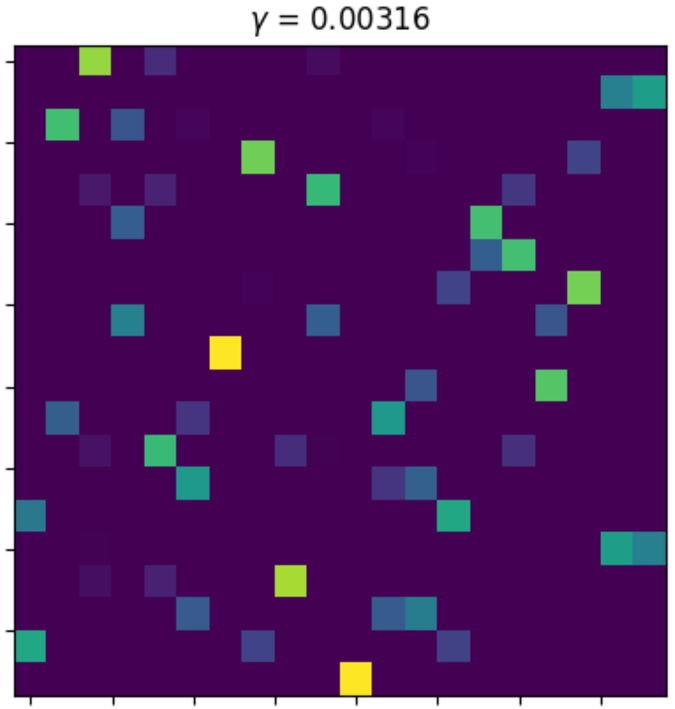
Solving Entropy Regularized OT

$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$



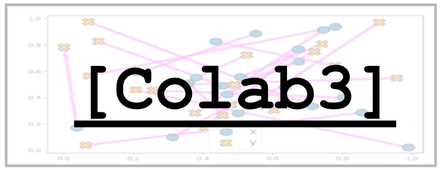
[Colab3]

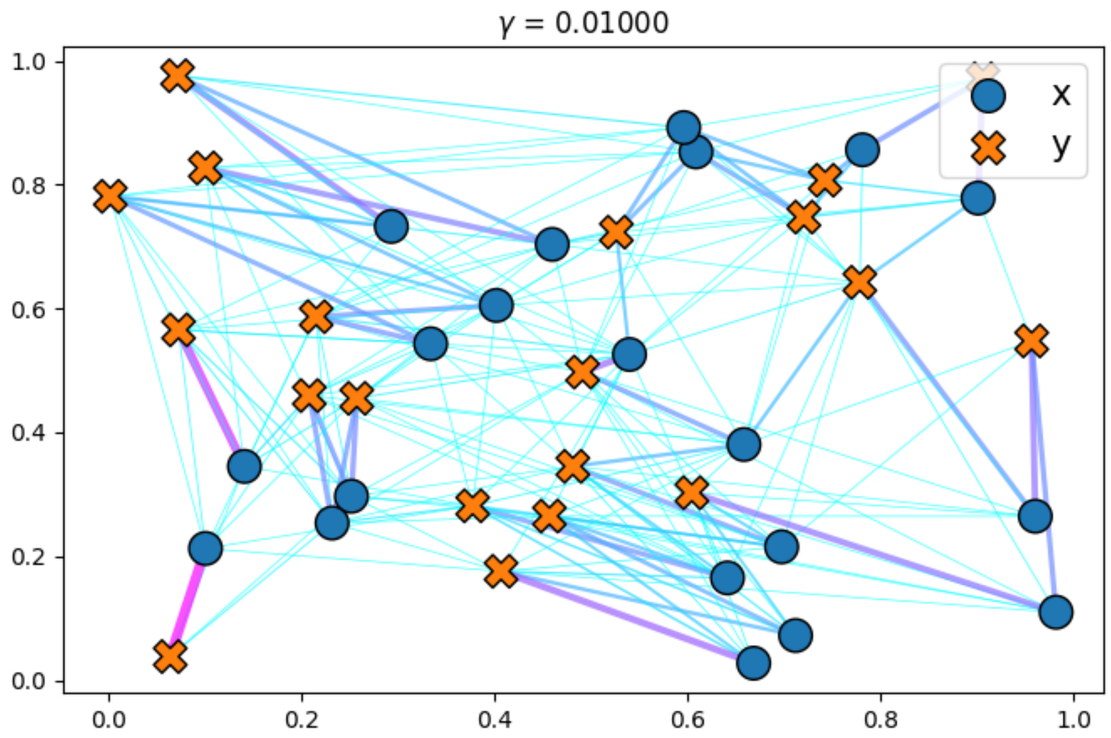


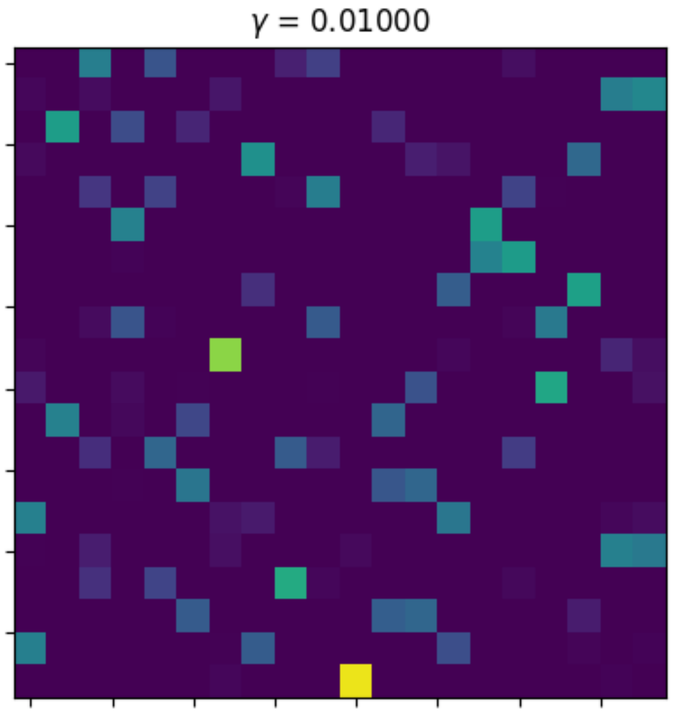


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

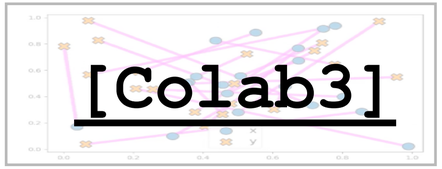


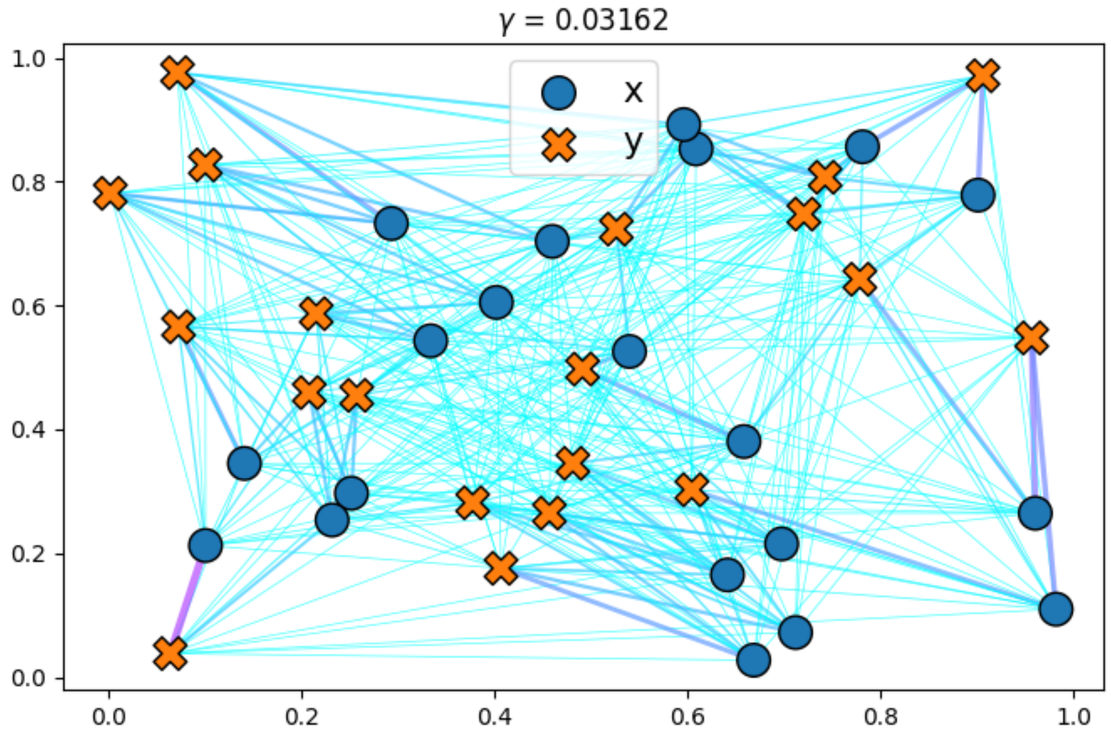


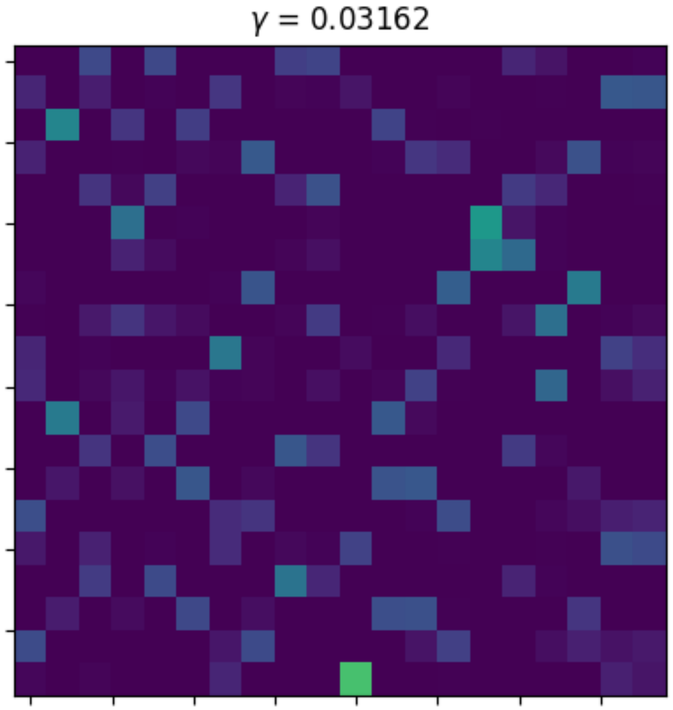


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

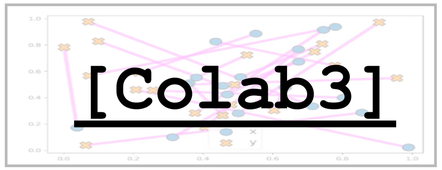


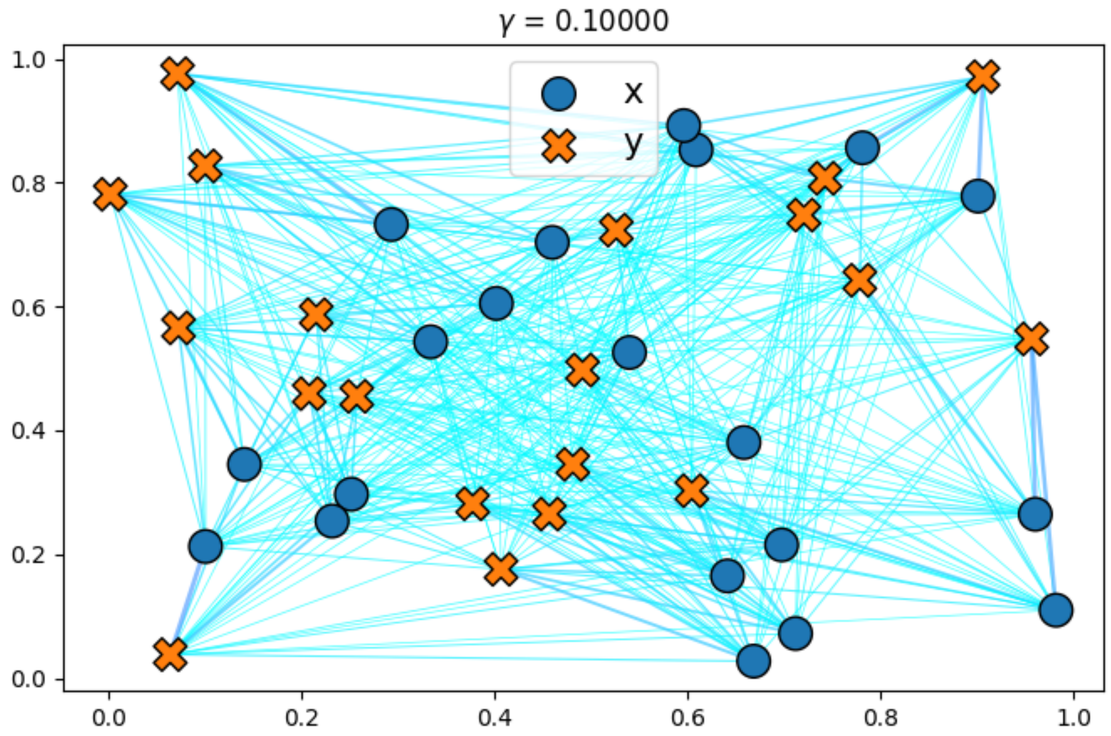


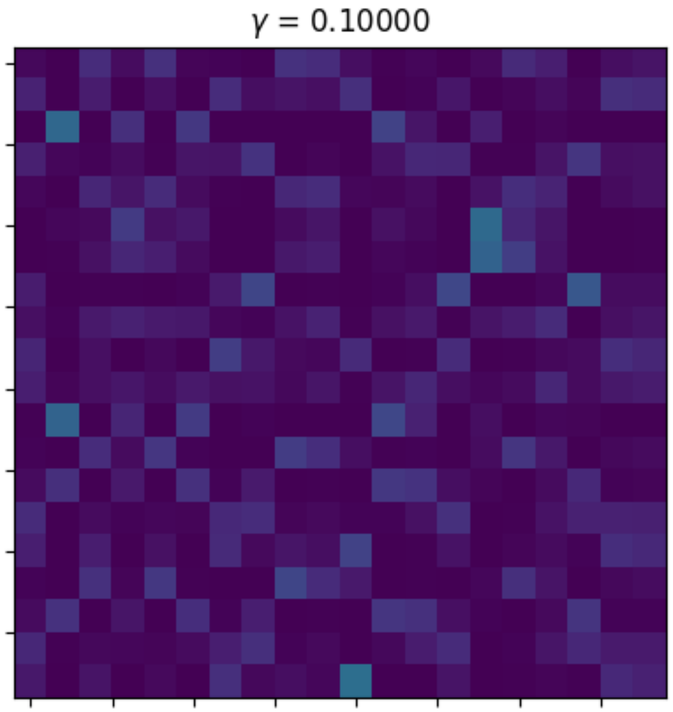


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

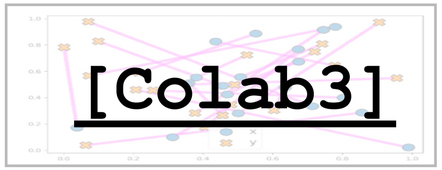


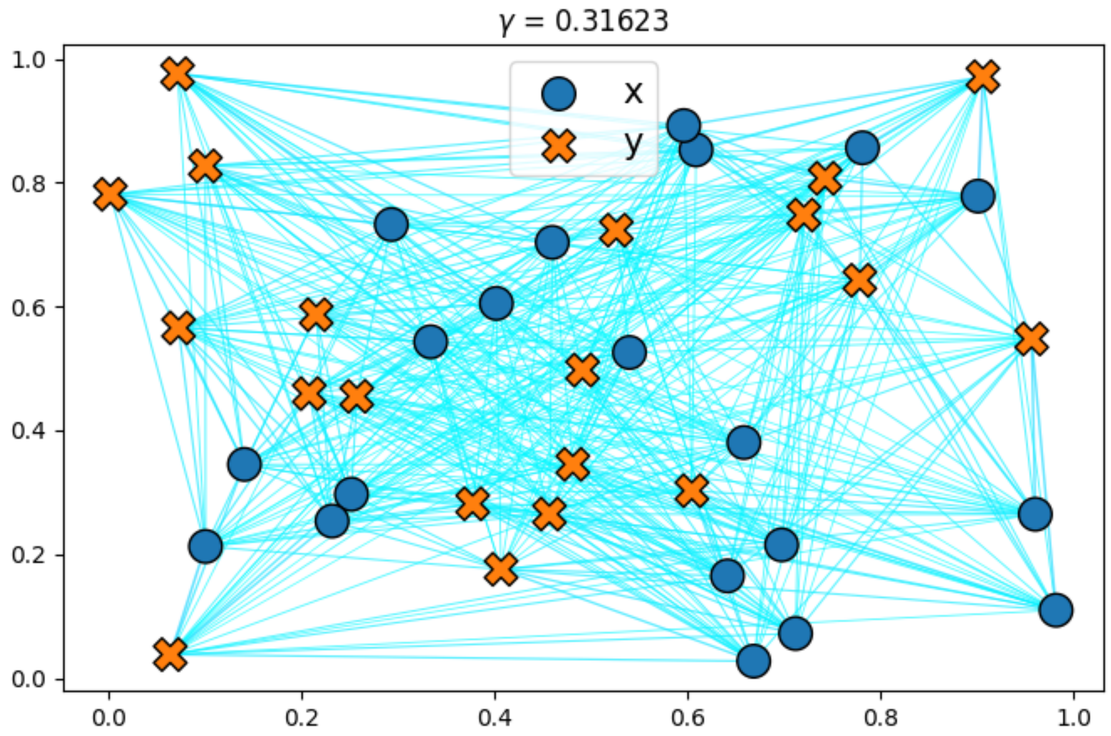


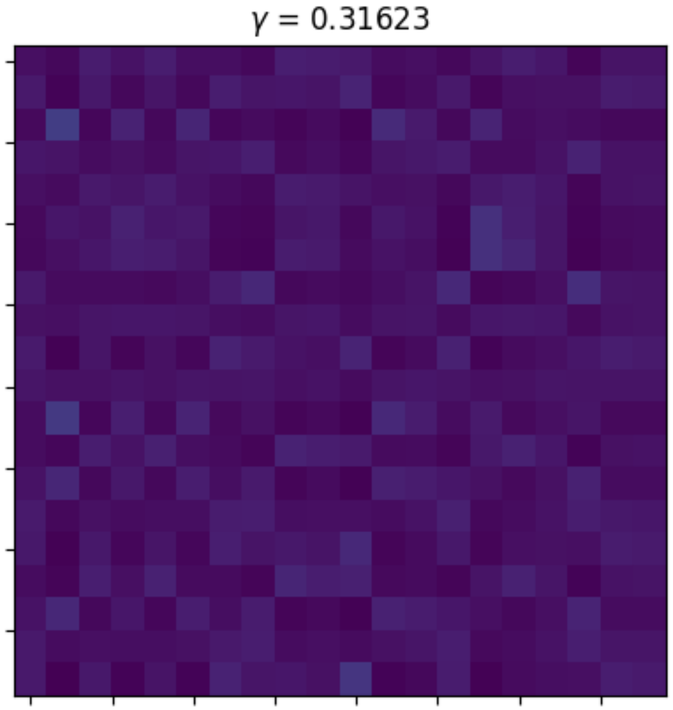


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

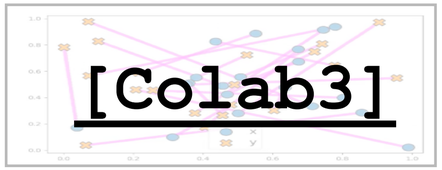


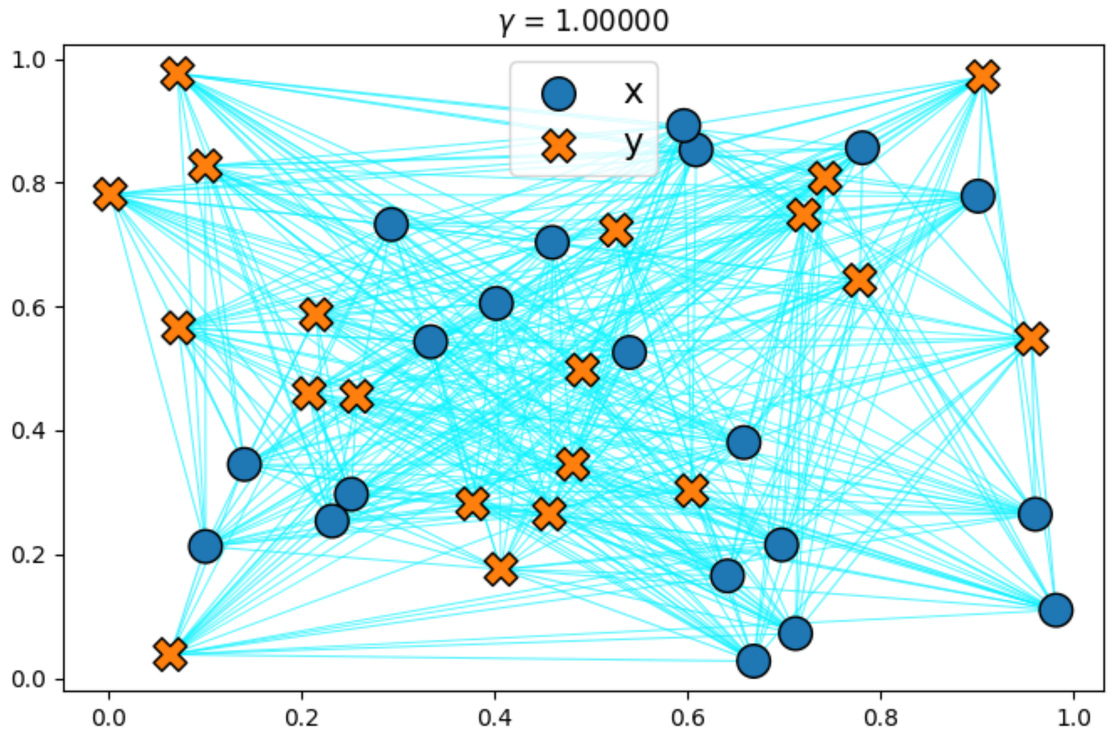


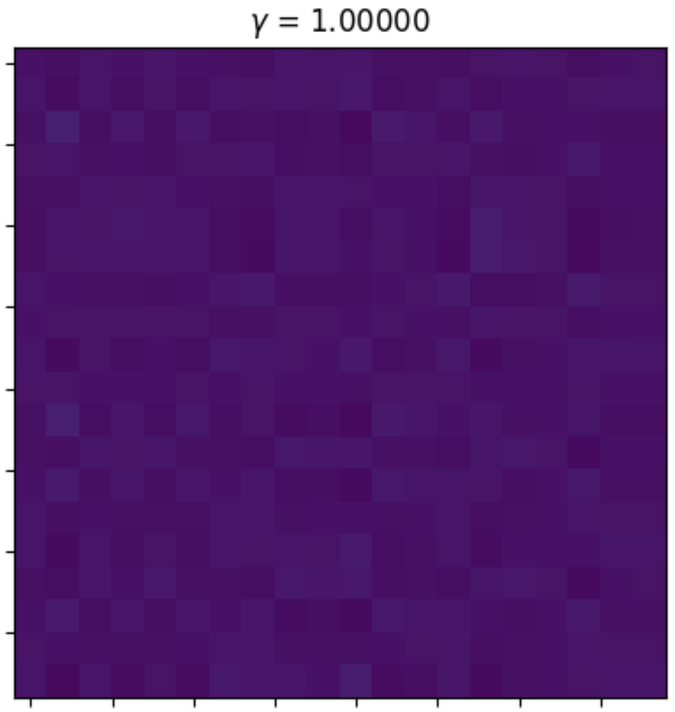


Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$

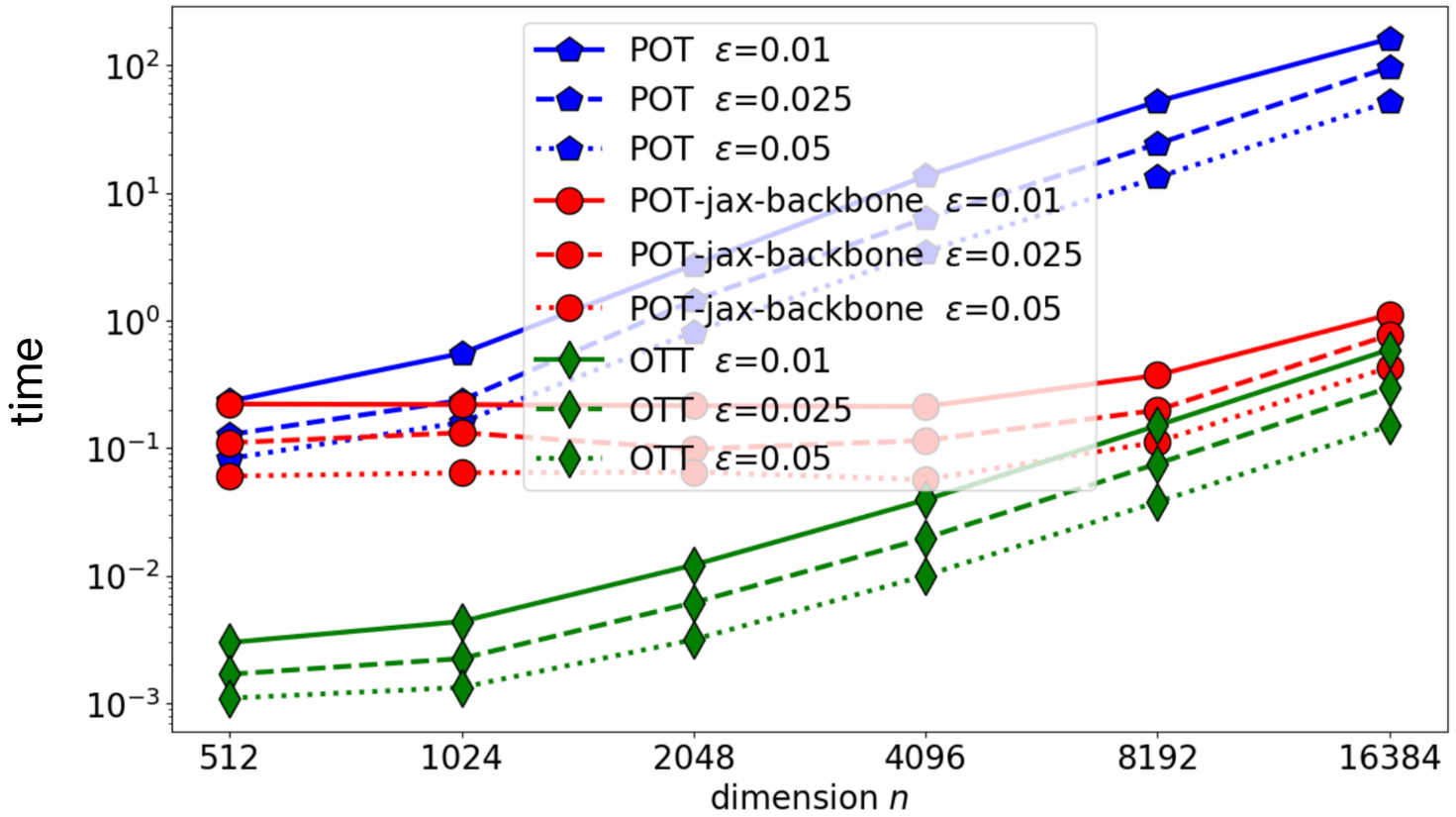






Solving Entropy Regularized OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle - \gamma E(P)$$



Solving Entropy Regularized OT

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$
then $\exists! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}) \mathbf{1}_m & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \mathbf{diag}(\boldsymbol{u}) \mathbf{1}_n & = \boldsymbol{b} \end{cases}$$

Solving Entropy Regularized OT

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$
then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}) \mathbf{1}_m & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \underbrace{\mathbf{diag}(\boldsymbol{u}) \mathbf{1}_n}_{\boldsymbol{u}} & = \boldsymbol{b} \end{cases}$$

Solving Entropy Regularized OT

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \overbrace{\mathbf{diag}(\boldsymbol{v}) \mathbf{1}_m}^{\boldsymbol{v}} & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \underbrace{\mathbf{diag}(\boldsymbol{u}) \mathbf{1}_n}_{\boldsymbol{u}} & = \boldsymbol{b} \end{cases}$$

Solving Entropy Regularized OT

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$
then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \mathbf{diag}(\boldsymbol{u}) K \, \boldsymbol{v} & = \boldsymbol{a} \\ \mathbf{diag}(\boldsymbol{v}) K^T \boldsymbol{u} & = \boldsymbol{b} \end{cases}$$

Solving Entropy Regularized OT

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$
then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

$$P_\gamma \in U(\boldsymbol{a}, \boldsymbol{b}) \Leftrightarrow \begin{cases} \boldsymbol{u} \odot K \boldsymbol{v} = \boldsymbol{a} \\ \boldsymbol{v} \odot K^T \boldsymbol{u} = \boldsymbol{b} \end{cases}$$

Solving Entropy Regularized OT

Prop. If $P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$
then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \overset{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

1. $\boldsymbol{u} = \boldsymbol{a} / K\boldsymbol{v}$
2. $\boldsymbol{v} = \boldsymbol{b} / K^T \boldsymbol{u}$

Solving Entropy Regularized OT

Prop. If $P_\gamma \stackrel{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$

then $\exists ! \boldsymbol{u} \in \mathbb{R}_+^n, \boldsymbol{v} \in \mathbb{R}_+^m$, such that

$$P_\gamma = \mathbf{diag}(\boldsymbol{u}) K \mathbf{diag}(\boldsymbol{v}), \quad K \stackrel{\text{def}}{=} e^{-M_{\boldsymbol{XY}}/\gamma}$$

Sinkhorn's Algorithm (a.k.a. IPFP, RAS): Initialize $\boldsymbol{v}$, then repeat

$$\boldsymbol{u} \leftarrow \frac{\boldsymbol{a}}{K\boldsymbol{v}} \qquad \boldsymbol{v} \leftarrow \frac{\boldsymbol{b}}{K^T \boldsymbol{u}}$$

[Deming+40, Wilson62, Sinkhorn64, Erlander92, Galichon+10, Cuturi13]

Sinkhorn as a Dual Algorithm

Def. Regularized Wasserstein, $\gamma \geq 0$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle - \gamma E(\boldsymbol{P})$$

REGULARIZED DISCRETE PRIMAL

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma (e^{\boldsymbol{\alpha}/\gamma})^T \, K (e^{\boldsymbol{\beta}/\gamma})$$

$$\text{where} \quad K = \left[e^{-\frac{\boldsymbol{D}^p(\boldsymbol{x_i}, \boldsymbol{y_j})}{\gamma}} \right]_{ij}$$

REGULARIZED DISCRETE DUAL

Solving Entropy Regularized OT

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma (e^{\boldsymbol{\alpha}/\gamma})^T \; K (e^{\boldsymbol{\beta}/\gamma})$$

$$\text{where} \quad K = \left[e^{-\frac{\boldsymbol{D}^p(\boldsymbol{x_i}, \boldsymbol{y_j})}{\gamma}} \right]_{ij}$$

REGULARIZED DISCRETE DUAL

Sinkhorn's Algorithm, a.ka. block coordinate ascent

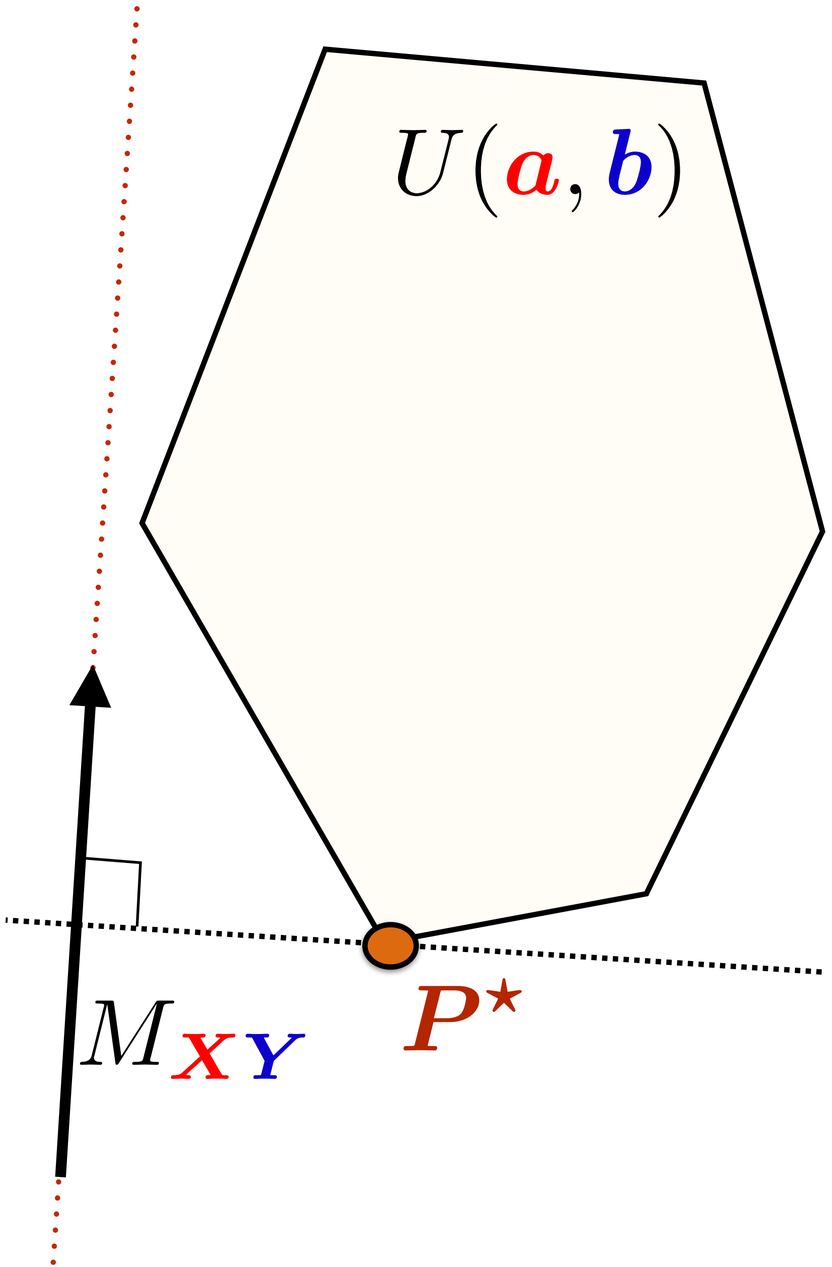
$$(\boldsymbol{u}, \boldsymbol{v}) \stackrel{\text{def}}{=} (e^{\boldsymbol{\alpha}/\gamma}, e^{\boldsymbol{\beta}/\gamma})$$

$$\boldsymbol{\alpha} \leftarrow \gamma \left(\log \boldsymbol{a} - \log \; K (e^{\boldsymbol{\beta}/\gamma}) \right) \qquad \boldsymbol{\beta} \leftarrow \gamma \left(\log \boldsymbol{b} - \log \; K^T (e^{\boldsymbol{\alpha}/\gamma}) \right)$$

Sinkhorn, Wasserstein and Energy Distance

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \qquad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

$$W^p(\mu, \nu) = \langle P^\star, M_{XY} \rangle$$

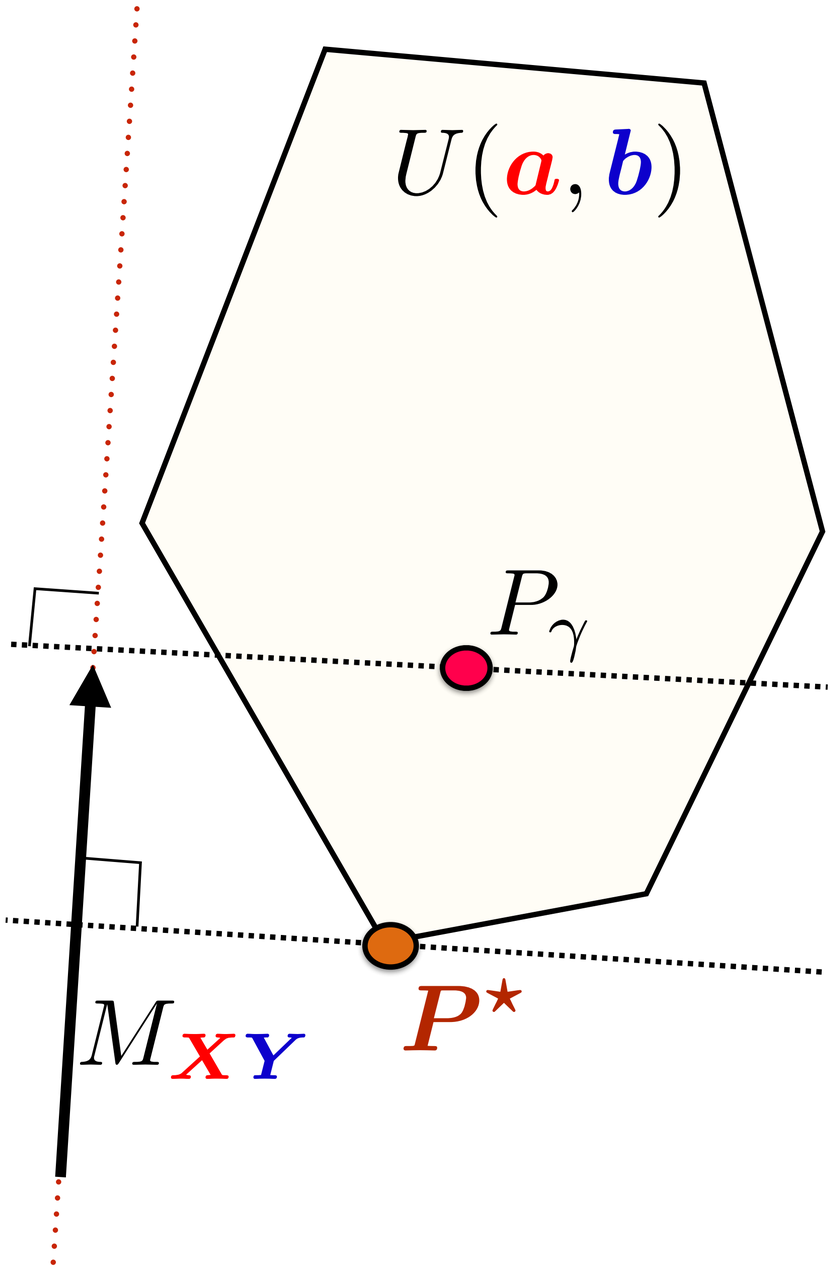


Sinkhorn, Wasserstein and Energy Distance

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \qquad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

$$W_\gamma(\mu, \nu) = \langle P_\gamma, M_{XY} \rangle$$

$$W^p(\mu, \nu) = \langle P^\star, M_{XY} \rangle$$



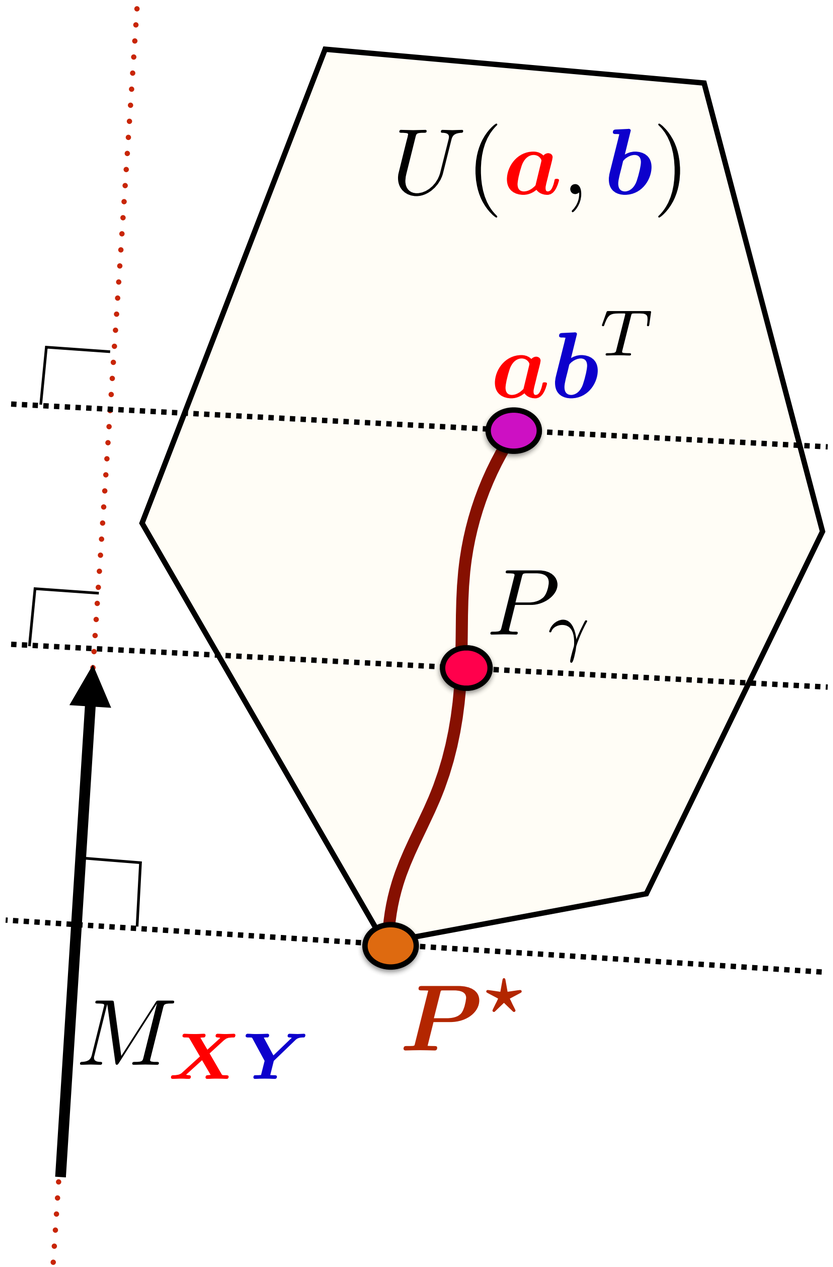
Sinkhorn, Wasserstein and Energy Distance

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \qquad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

$$\mathcal{E}(\mu, \nu) = \langle ab^T, M_{XY} \rangle$$

$$W_\gamma(\mu, \nu) = \langle P_\gamma, M_{XY} \rangle$$

$$W^p(\mu, \nu) = \langle P^\star, M_{XY} \rangle$$



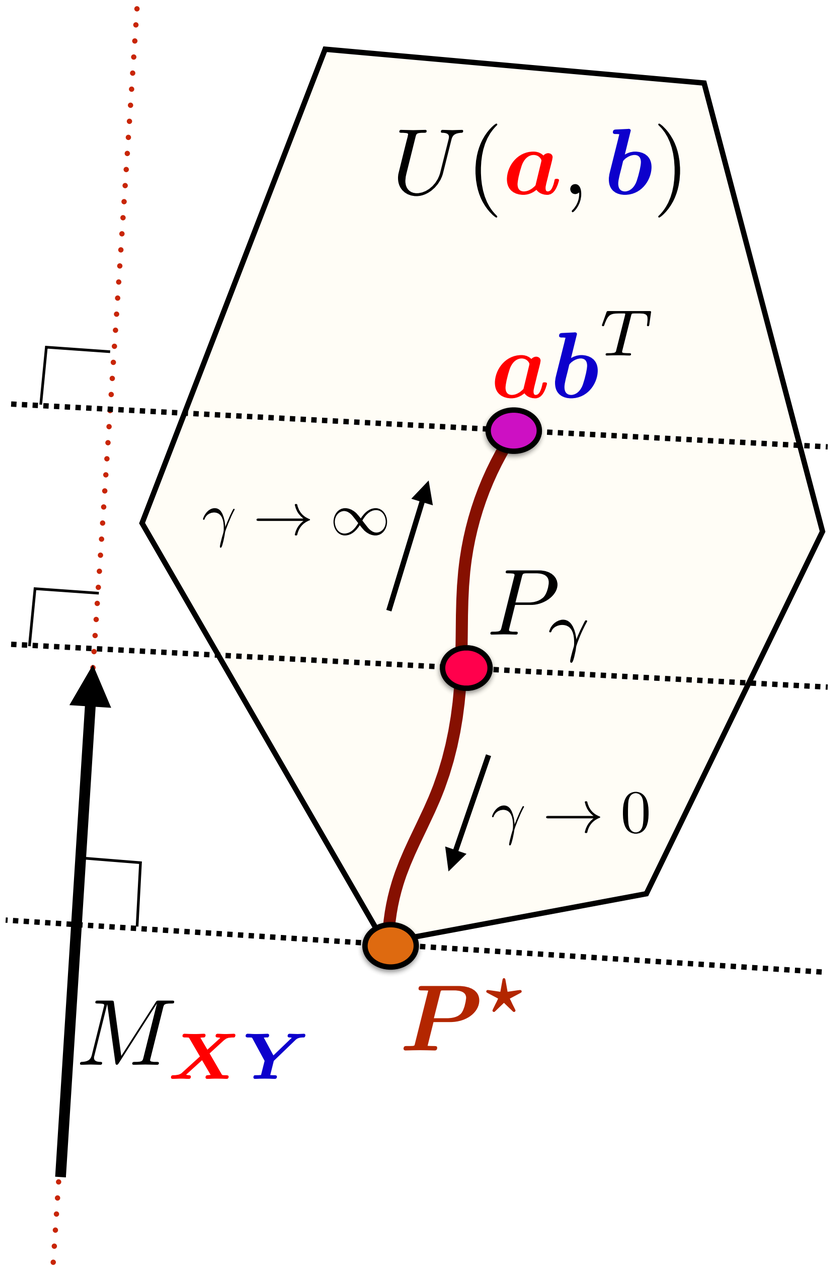
Sinkhorn, Wasserstein and Energy Distance

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \qquad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

$$\mathcal{E}(\mu, \nu) = \langle ab^T, M_{XY} \rangle$$

$$W_\gamma(\mu, \nu) = \langle P_\gamma, M_{XY} \rangle$$

$$W^p(\mu, \nu) = \langle P^\star, M_{XY} \rangle$$



Sinkhorn, Wasserstein and Energy Distance

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{a}\boldsymbol{b}^T, M_{\boldsymbol{XY}} \rangle$$

$$\mathrm{ED}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\nu}, \boldsymbol{\nu}))$$

$$W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle P_\gamma, M_{\boldsymbol{XY}} \rangle - \gamma E(P_\gamma)$$

$$\bar{W}_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(W_\gamma(\boldsymbol{\mu}, \boldsymbol{\mu}) + W_\gamma(\boldsymbol{\nu}, \boldsymbol{\nu}))$$

$$W^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\boldsymbol{XY}} \rangle$$

Sinkhorn, Wasserstein and Energy Distance

$$\mathcal{MMD}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\nu}, \boldsymbol{\nu}))$$

$$\gamma \to \infty \uparrow$$

$$\bar{W}_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(W_\gamma(\boldsymbol{\mu}, \boldsymbol{\mu}) + W_\gamma(\boldsymbol{\nu}, \boldsymbol{\nu}))$$

$$\gamma \to 0 \downarrow$$

$$W^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\boldsymbol{XY}} \rangle$$

How to compare them?

i.i.d samples $\boldsymbol{x_1, \dots, x_n \sim \mu}$, $\boldsymbol{y_1, \dots, y_m \sim \nu}$,

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{n}} \stackrel{\text{def}}{=} \frac{1}{n} \sum_i \delta_{\boldsymbol{x_i}}, \hat{\boldsymbol{\nu}}_{\boldsymbol{m}} \stackrel{\text{def}}{=} \frac{1}{m} \sum_j \delta_{\boldsymbol{y_j}}$$

Computational properties

Effort to compute/approximate $\boldsymbol{\Delta}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}, \hat{\boldsymbol{\nu}}_{\boldsymbol{m}})$?

Statistical properties

$|\boldsymbol{\Delta}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \boldsymbol{\Delta}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}, \hat{\boldsymbol{\nu}}_{\boldsymbol{n}})| \leq f(n)$?

Sinkhorn in between *W* and *MMD*

$$\mathcal{MMD}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\nu}, \boldsymbol{\nu}))$$

$(n+m)^2$

$O(1/\sqrt{n})$

$$W^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\boldsymbol{XY}} \rangle$$

$O((n+m)nm\log(n+m))$

$O(1/n^{1/d})$

Sinkhorn in between *W* and *MMD*

$$\mathcal{MMD}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\mu}) + \mathcal{E}(\boldsymbol{\nu}, \boldsymbol{\nu}))$$

$(n+m)^2$ $\qquad$ $O(1/\sqrt{n})$

$$\bar{W}_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = W_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(W_\gamma(\boldsymbol{\mu}, \boldsymbol{\mu}) + W_\gamma(\boldsymbol{\nu}, \boldsymbol{\nu}))$$

$O((n+m)^2)$ $\qquad$ $O\left(\frac{1}{\gamma^{d/2}\sqrt{n}}\right)$ $\qquad$ **[GCBCP'18]** **[FSVATP'18]**

$$W^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \langle \boldsymbol{P}^\star, M_{\boldsymbol{XY}} \rangle$$

$O((n+m)nm\log(n+m))$ $\qquad$ $O(1/n^{1/d})$

Unbalanced Generalizations for OT

$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

[Frogner+15, Chizat+16]

Unbalanced Generalizations for OT

$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

[Frogner+15, Chizat+16]

$$\min_{\boldsymbol{P} \in \mathbb{R}_+^{\boldsymbol{n} \times \boldsymbol{m}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle + \rho_{\boldsymbol{a}} \mathrm{KL}(\boldsymbol{P} \mathbf{1}_{\boldsymbol{m}} \| \boldsymbol{a}) + \rho_{\boldsymbol{b}} \mathrm{KL}(\boldsymbol{P}^T \mathbf{1}_{\boldsymbol{n}} \| \boldsymbol{b})$$

Unbalanced Generalizations for OT

$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

[Frogner+15, Chizat+16]

$$\min_{\boldsymbol{P} \in \mathbb{R}_+^{\boldsymbol{n} \times \boldsymbol{m}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle + \rho_{\boldsymbol{a}} \mathrm{KL}(\boldsymbol{P}\mathbf{1}_{\boldsymbol{m}} \| \boldsymbol{a}) + \rho_{\boldsymbol{b}} \mathrm{KL}(\boldsymbol{P}^T \mathbf{1}_{\boldsymbol{n}} \| \boldsymbol{b}) - \gamma E(\boldsymbol{P})$$

Still a Sinkhorn-esque Algorithm!

$$\boldsymbol{u} = (\boldsymbol{a} / K\boldsymbol{v})^{\frac{\rho_{\boldsymbol{a}}}{\rho_{\boldsymbol{a}} + \gamma}} \qquad \boldsymbol{v} = \left(\boldsymbol{b} / K^T \boldsymbol{u}\right)^{\frac{\rho_{\boldsymbol{b}}}{\rho_{\boldsymbol{b}} + \gamma}}$$

Unbalanced Generalizations for OT

$$\min_{P \in U(a,b)} \langle P, M_{XY} \rangle$$

[Frogner+15, Chizat+16]

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

Still a Sinkhorn-esque Algorithm!

$$u = (a / Kv)^{\frac{\rho_a}{\rho_a + \gamma}} \qquad v = \left(b / K^T u\right)^{\frac{\rho_b}{\rho_b + \gamma}}$$

$$0 \le \tau_a \le 1 \qquad 0 \le \tau_b \le 1$$

Semi-Balanced Sinkhorn

$$\min_{\boldsymbol{P}\in\mathbb{R}_{+}^{\boldsymbol{n}\times\boldsymbol{m}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle + \rho_{\boldsymbol{a}}\mathrm{KL}(\boldsymbol{P}\mathbf{1}_{\boldsymbol{m}}\|\boldsymbol{a}) + \rho_{\boldsymbol{b}}\mathrm{KL}(\boldsymbol{P}^{T}\mathbf{1}_{\boldsymbol{n}}\|\boldsymbol{b}) - \gamma E(\boldsymbol{P})$$

Semi-Balanced Sinkhorn

$$\min_{\boldsymbol{P}\in\mathbb{R}_+^{\boldsymbol{n}\times\boldsymbol{m}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle + \rho_{\boldsymbol{a}}\mathrm{KL}(\boldsymbol{P}\mathbf{1}_{\boldsymbol{m}}\|\boldsymbol{a}) + \rho_{\boldsymbol{b}}\mathrm{KL}(\boldsymbol{P}^T\mathbf{1}_{\boldsymbol{n}}\|\boldsymbol{b}) - \gamma E(\boldsymbol{P})$$

$$\rho_{\boldsymbol{b}} = \infty$$

Semi-Balanced Sinkhorn

$$\min_{\substack{\boldsymbol{P}\in\mathbb{R}_+^{\boldsymbol{n}\times\boldsymbol{m}} \\ \boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle + \rho_{\boldsymbol{a}}\mathrm{KL}(\boldsymbol{P}\mathbf{1}_{\boldsymbol{m}}\|\boldsymbol{a}) - \gamma E(\boldsymbol{P})$$

Semi-Balanced Sinkhorn

$$\min_{\substack{\boldsymbol{P}\in\mathbb{R}_+^{\boldsymbol{n}\times\boldsymbol{m}} \\ \boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle + \rho_{\boldsymbol{a}}\mathrm{KL}(\boldsymbol{P}\mathbf{1}_{\boldsymbol{m}}\|\boldsymbol{a}) - \gamma E(\boldsymbol{P})$$

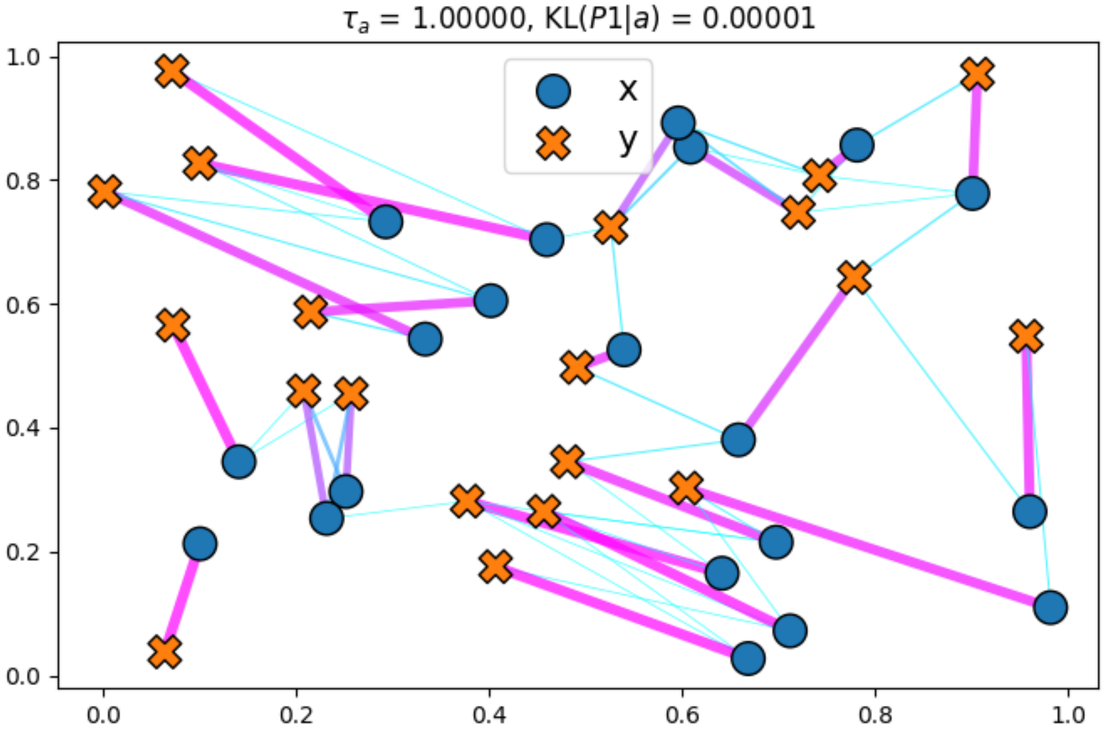
$$\tau_{\boldsymbol{a}} := \frac{\rho_{\boldsymbol{a}}}{\rho_{\boldsymbol{a}}+\gamma}$$

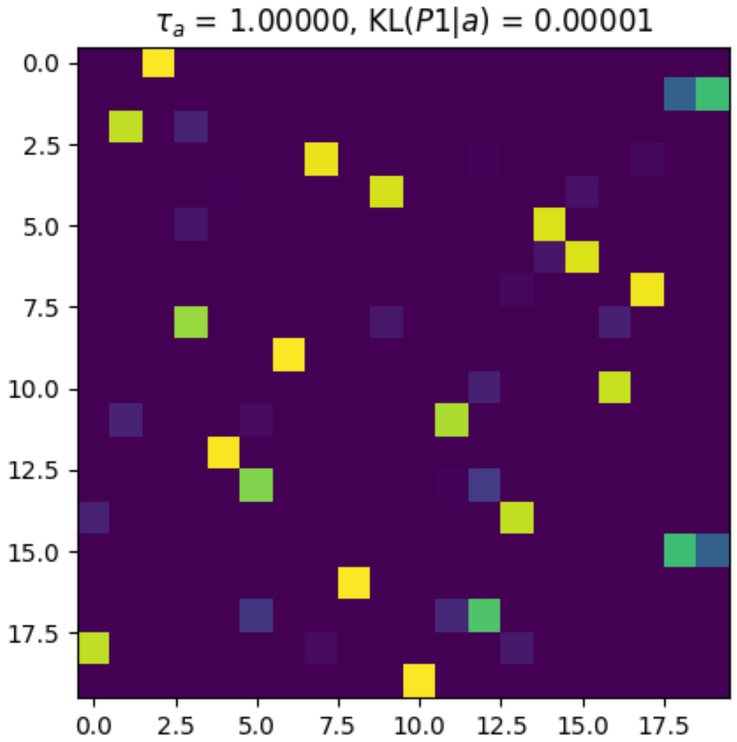
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



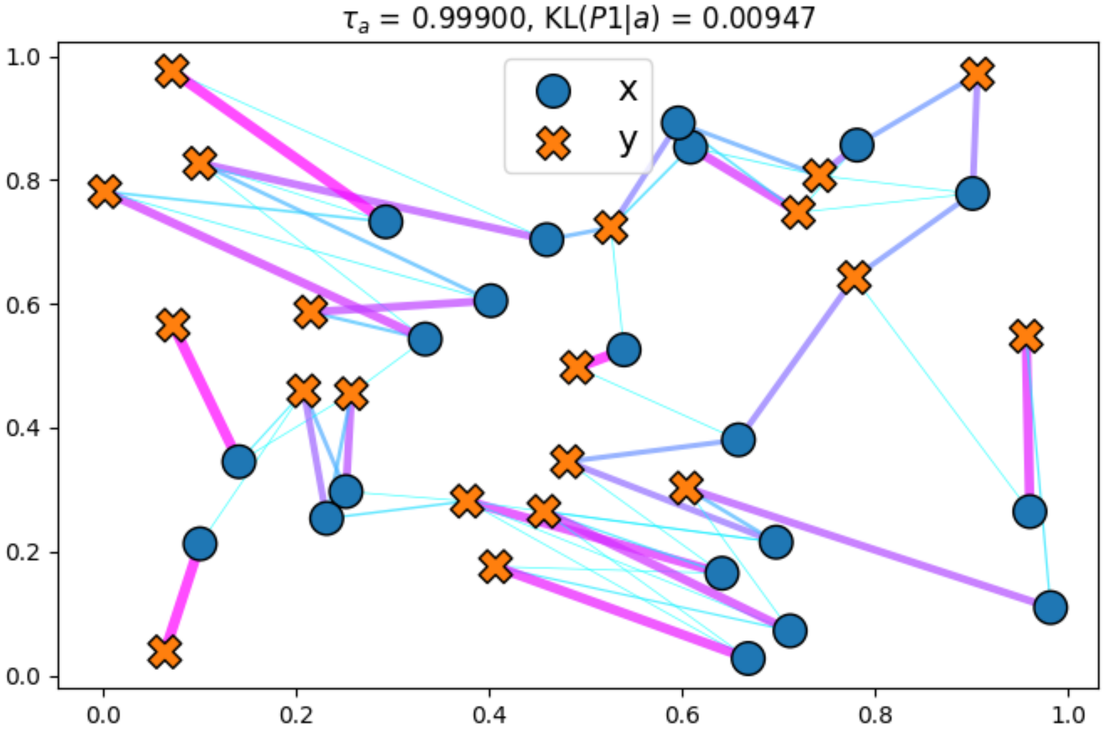


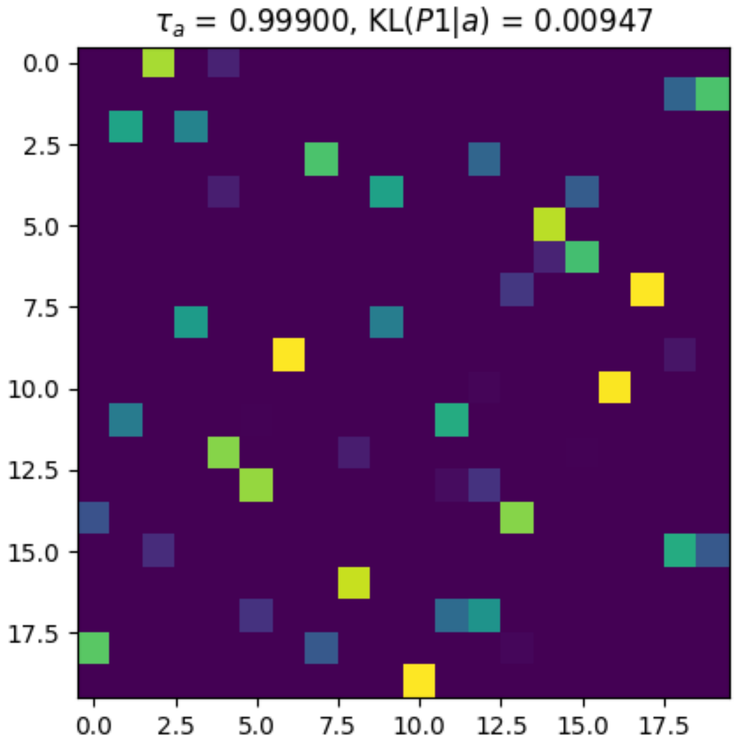
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



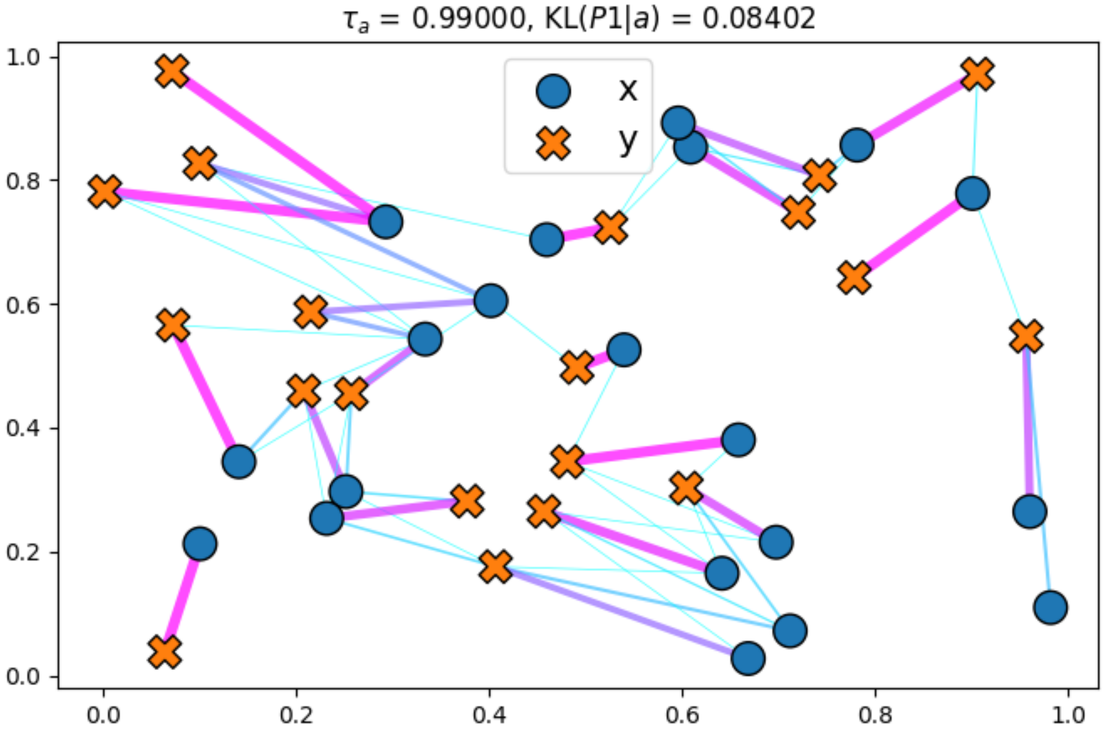


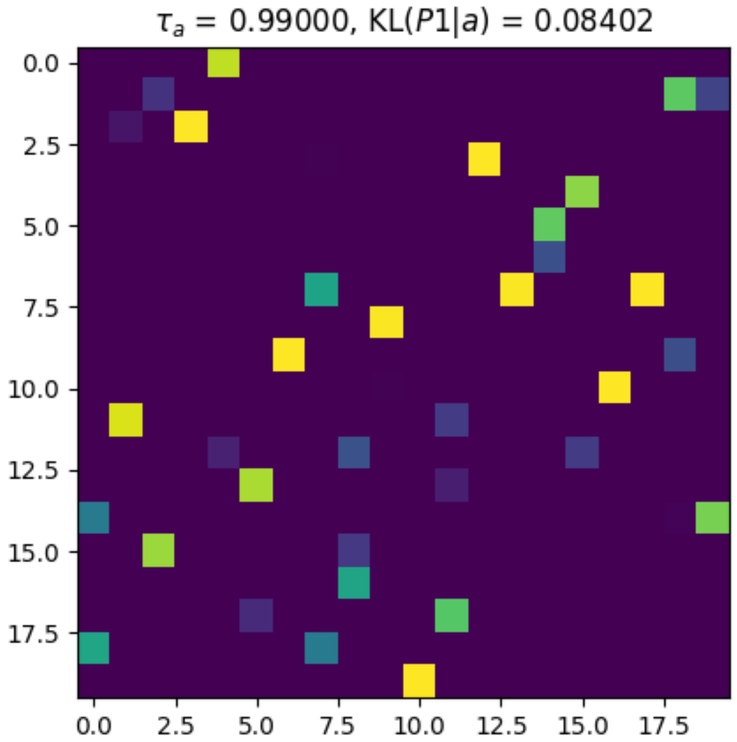
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



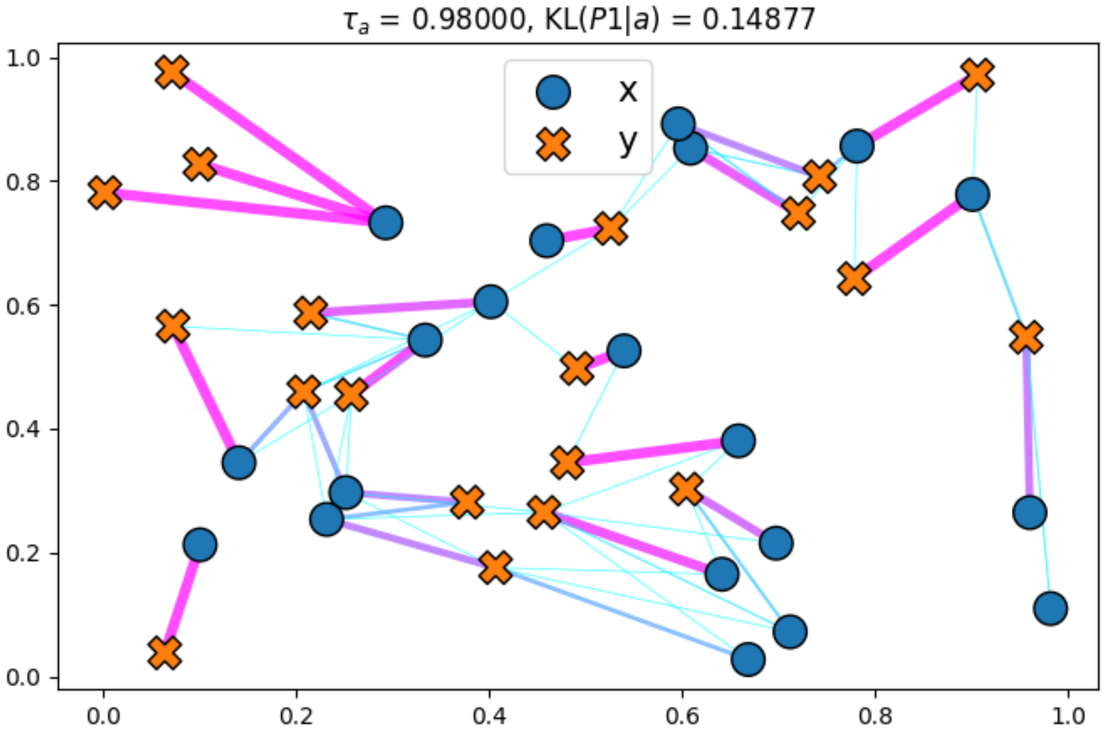


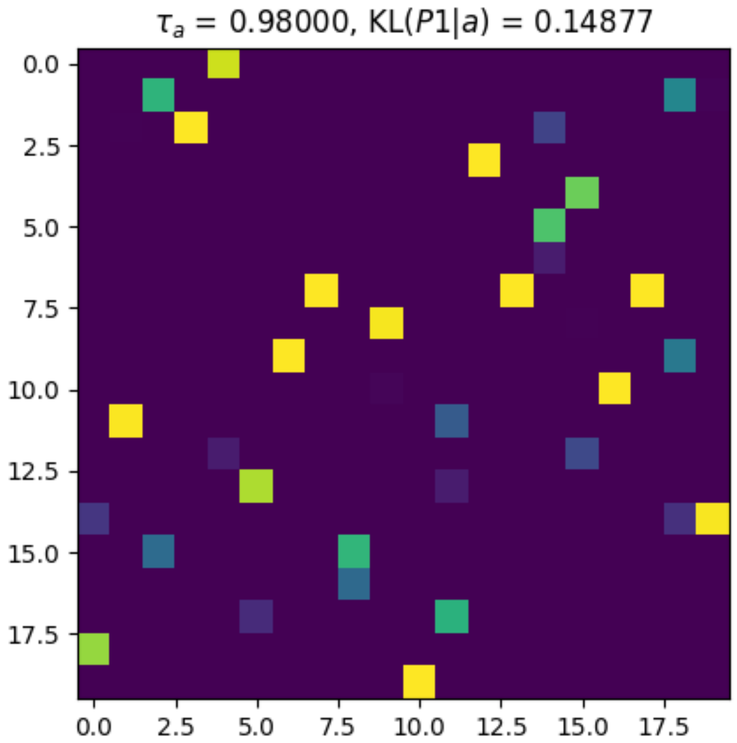
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



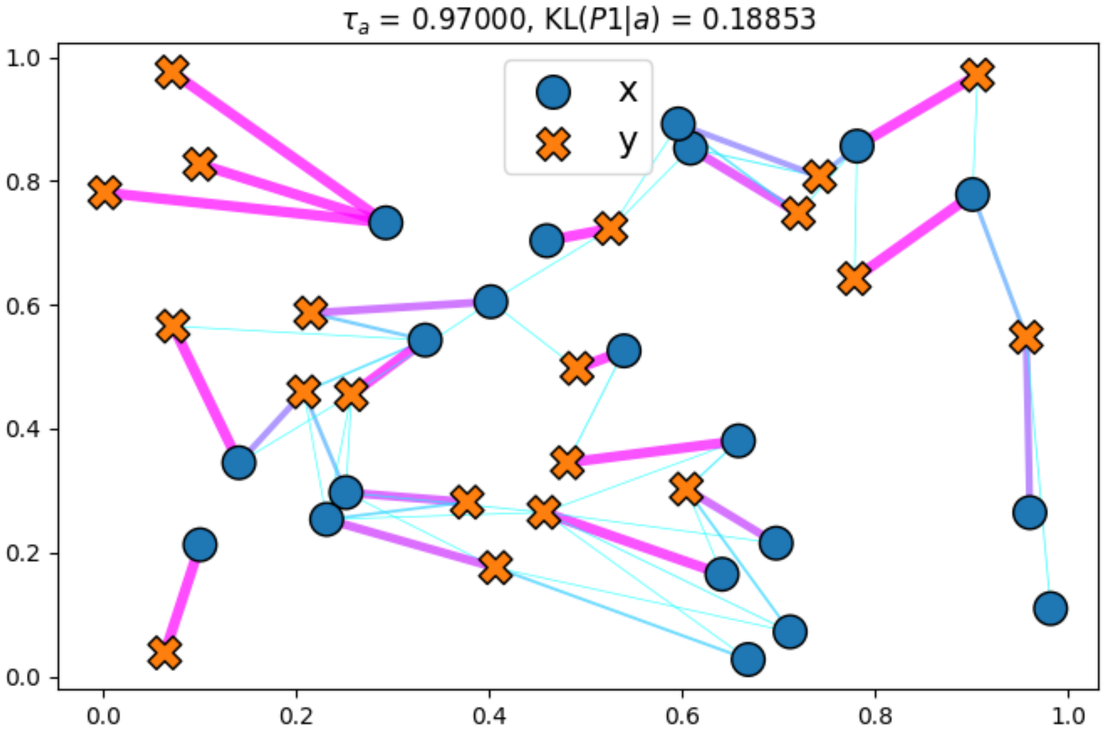


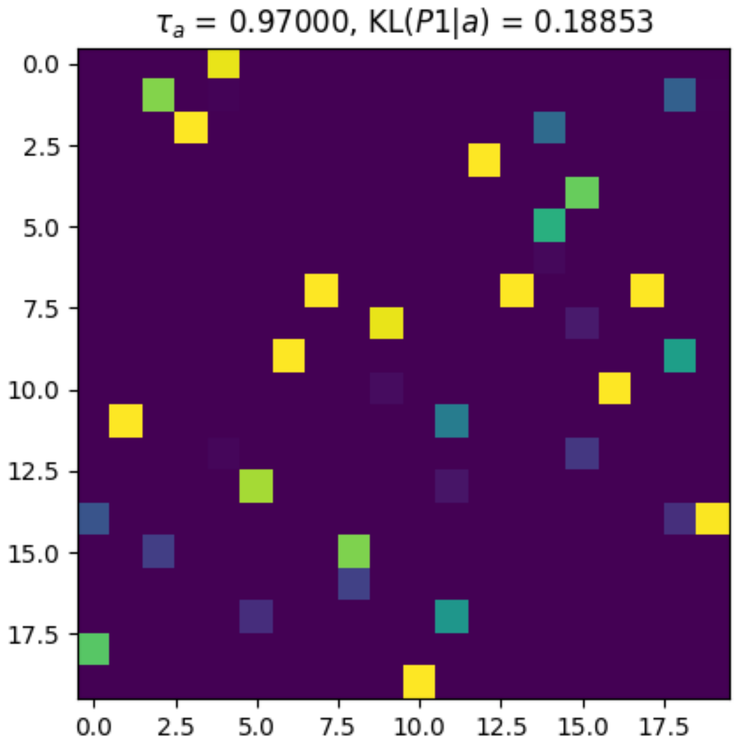
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



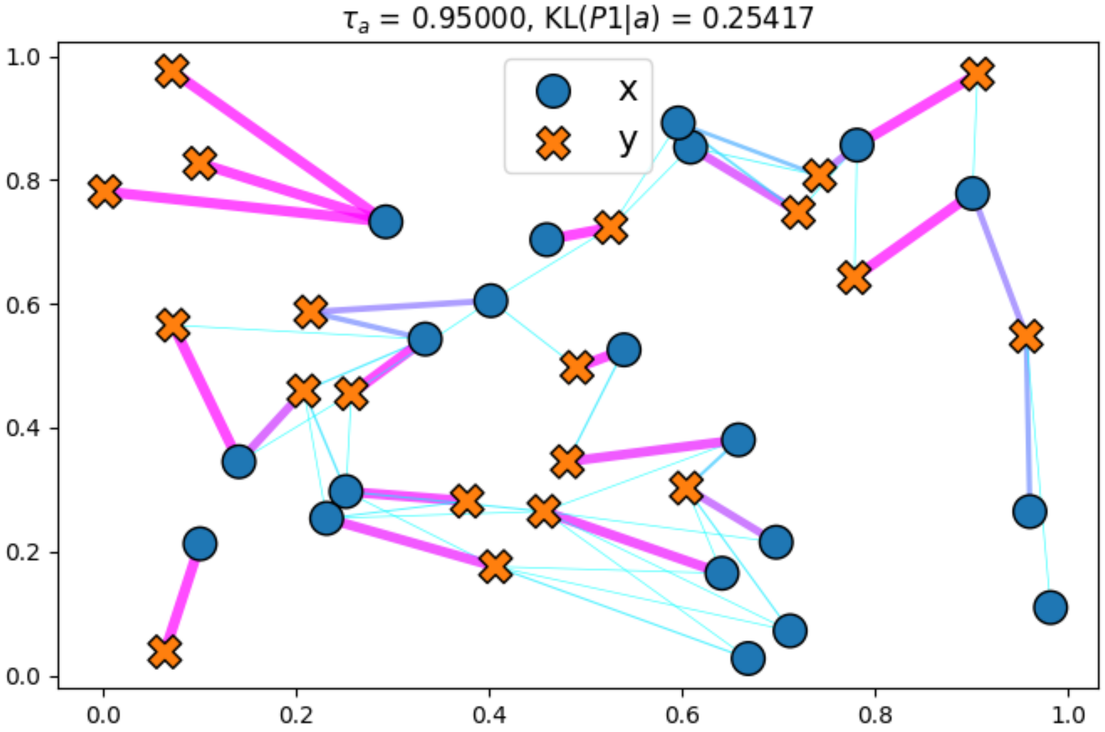


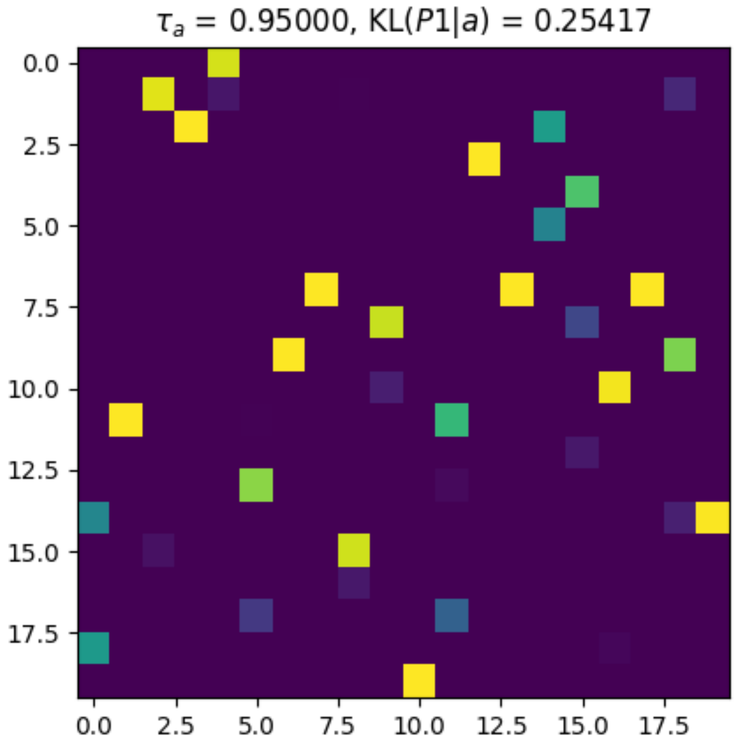
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



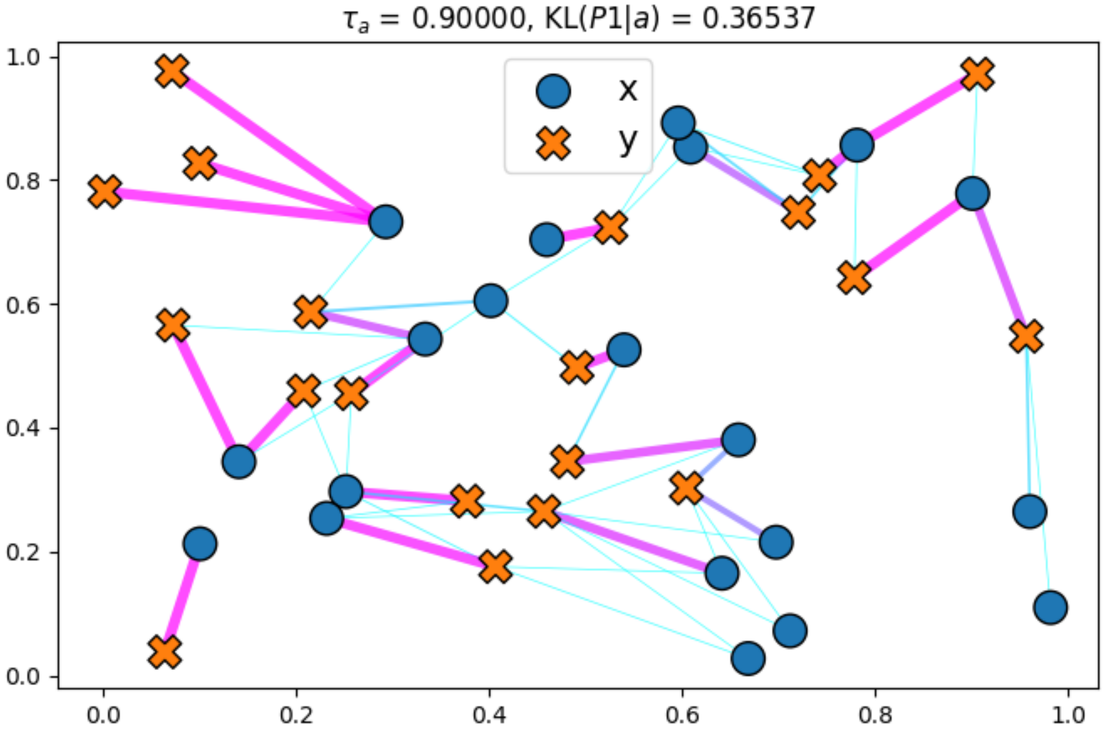


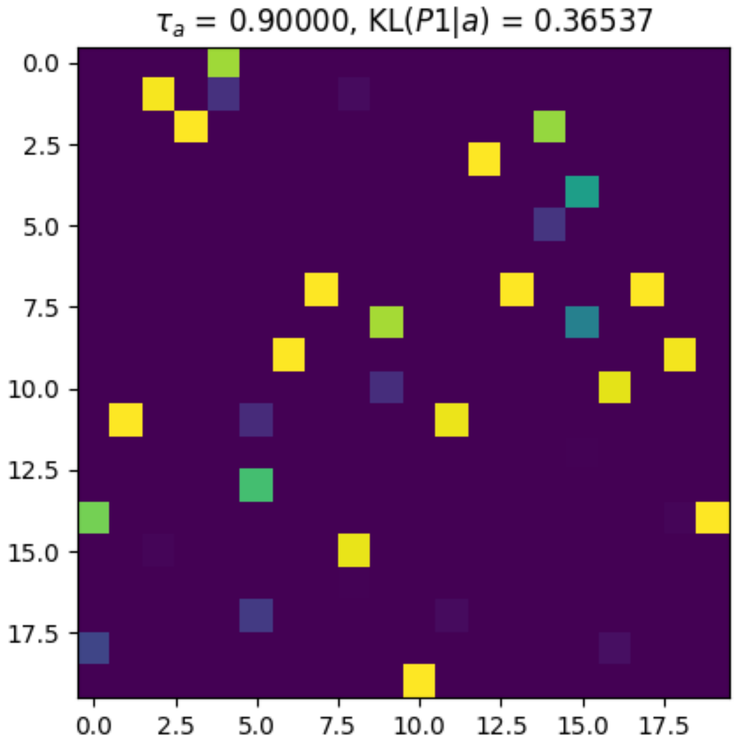
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



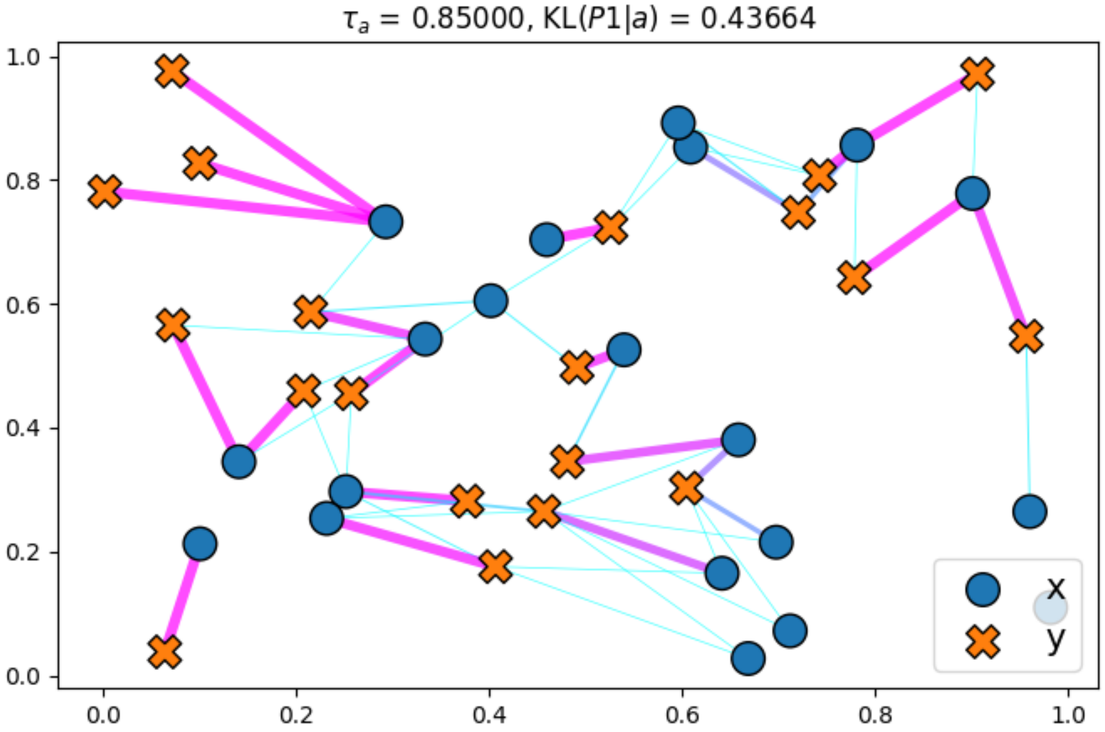


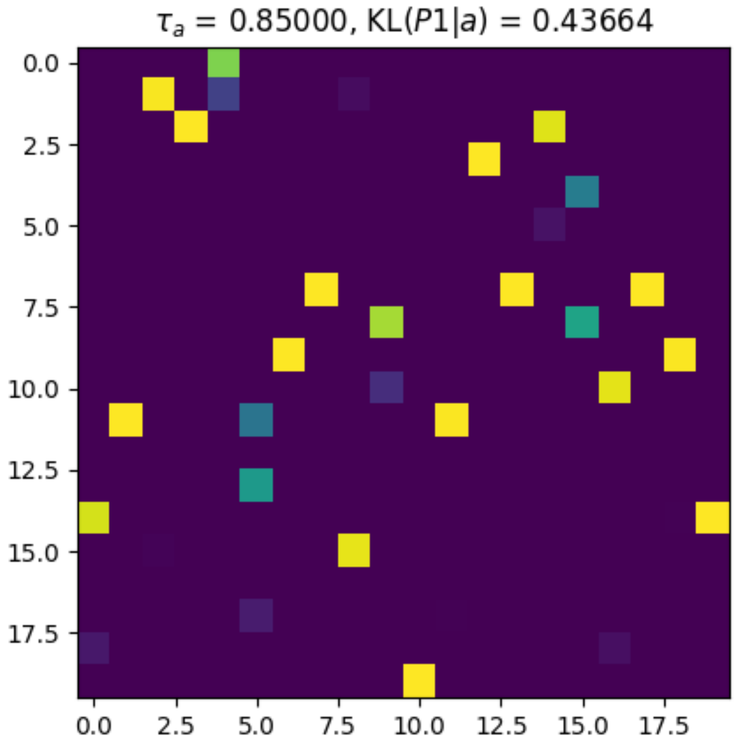
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



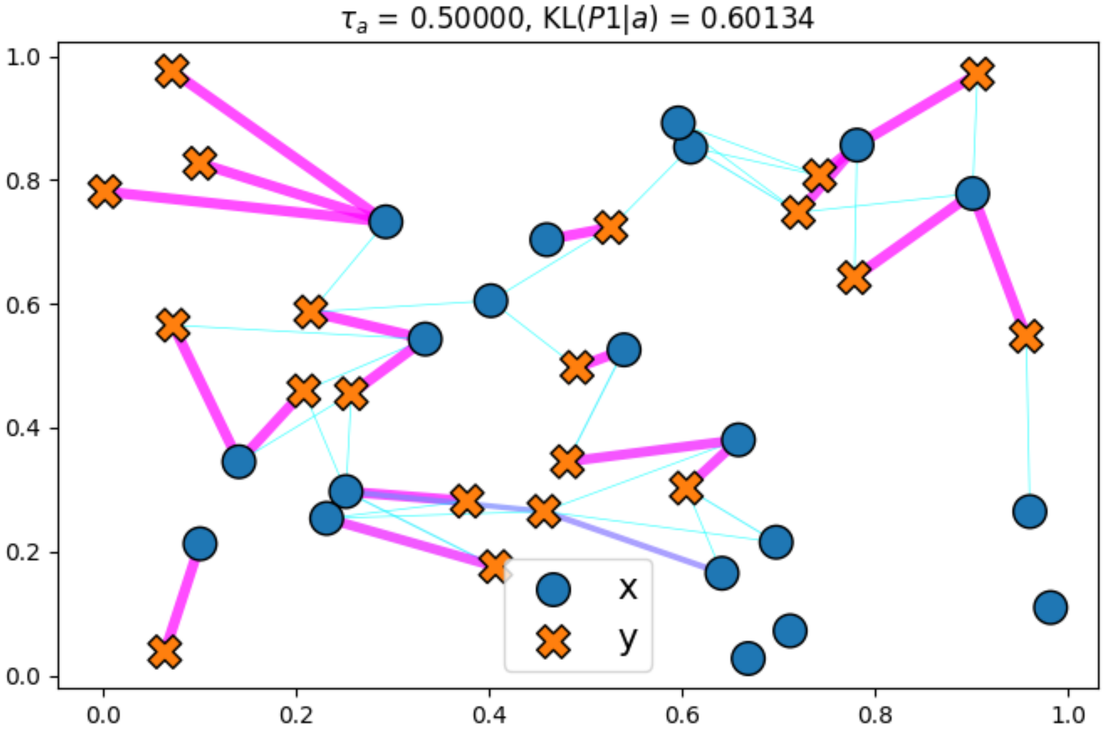


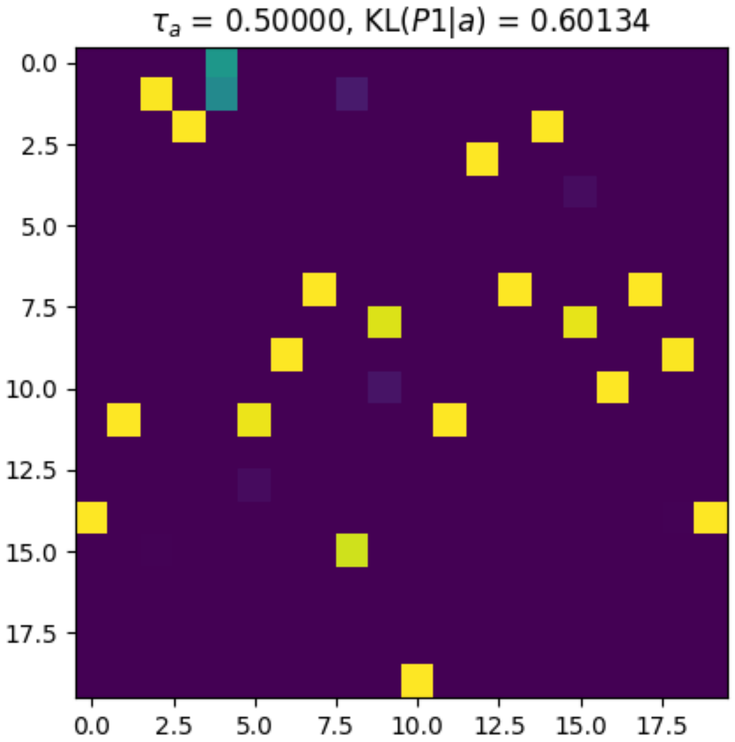
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



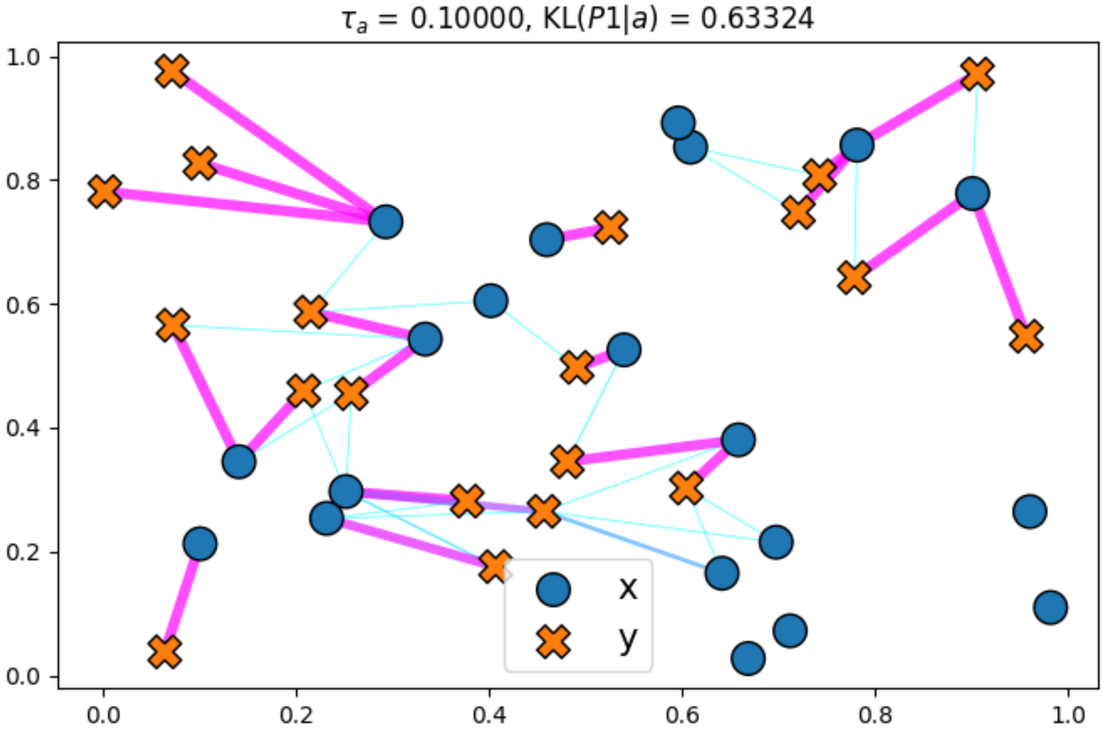


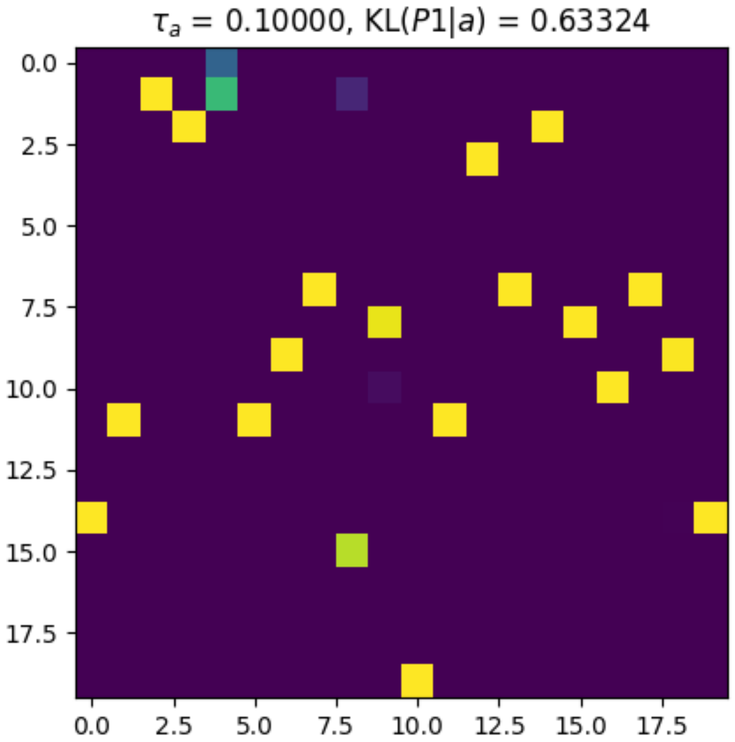
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]



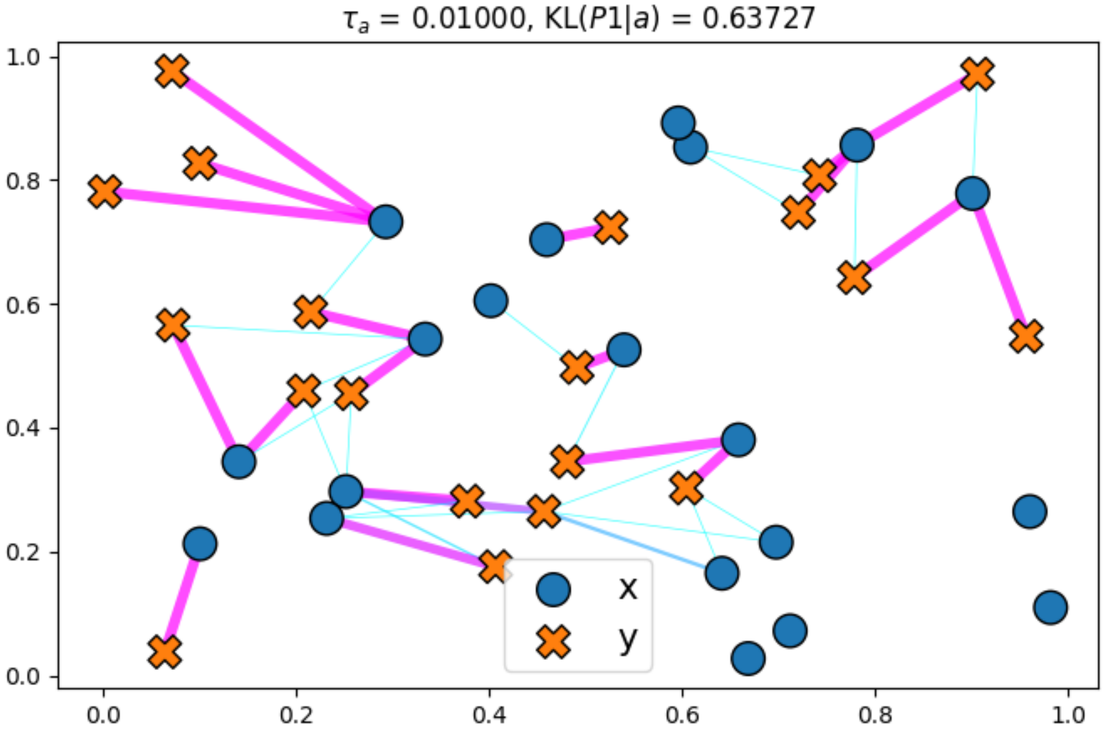


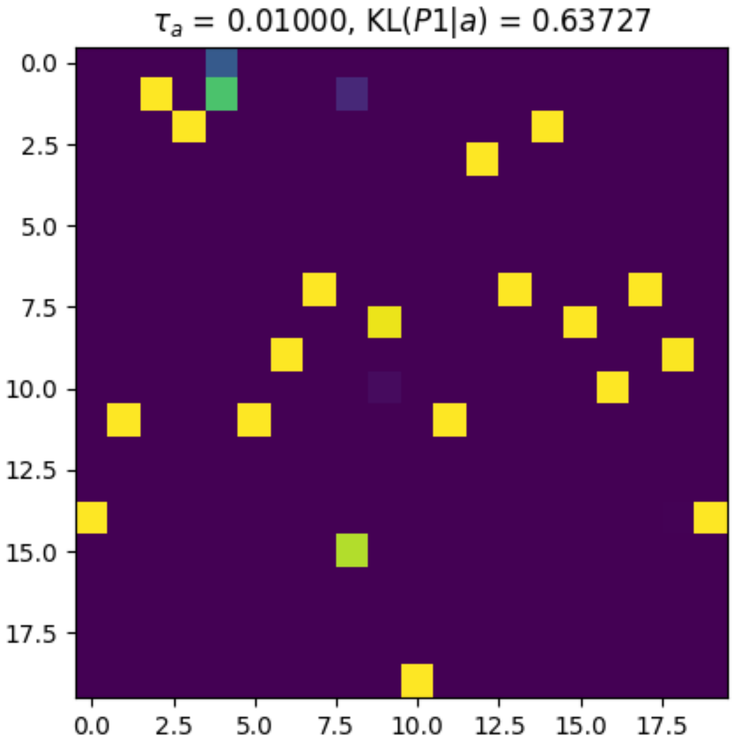
Semi-Balanced Sinkhorn

$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P^T 1_n = b}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P 1_m \| a) - \gamma E(P)$$

$$\tau_a := \frac{\rho_a}{\rho_a + \gamma}$$

[Colab4]





Unbalanced Sinkhorn

$$\min_{\boldsymbol{P}\in\mathbb{R}_{+}^{\boldsymbol{n}\times\boldsymbol{m}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle + \rho_{\boldsymbol{a}}\mathrm{KL}(\boldsymbol{P}\mathbf{1}_{\boldsymbol{m}}\|\boldsymbol{a}) + \rho_{\boldsymbol{b}}\mathrm{KL}(\boldsymbol{P}^{T}\mathbf{1}_{\boldsymbol{n}}\|\boldsymbol{b}) - \gamma E(\boldsymbol{P})$$



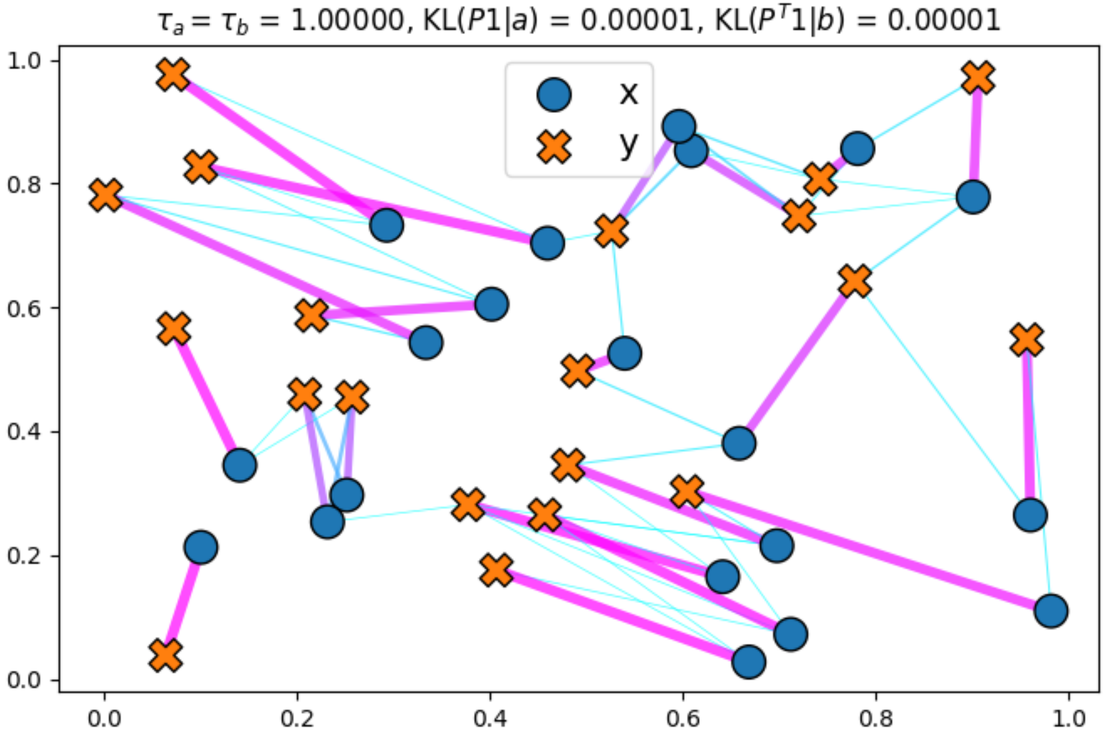
[Colab4]

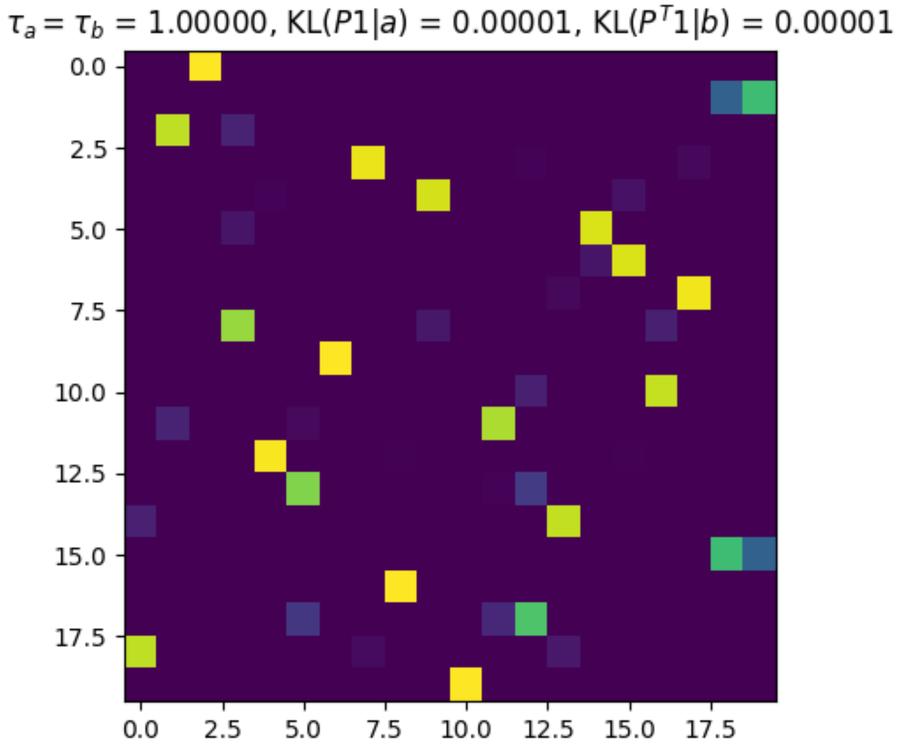
Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$



[Colab4]

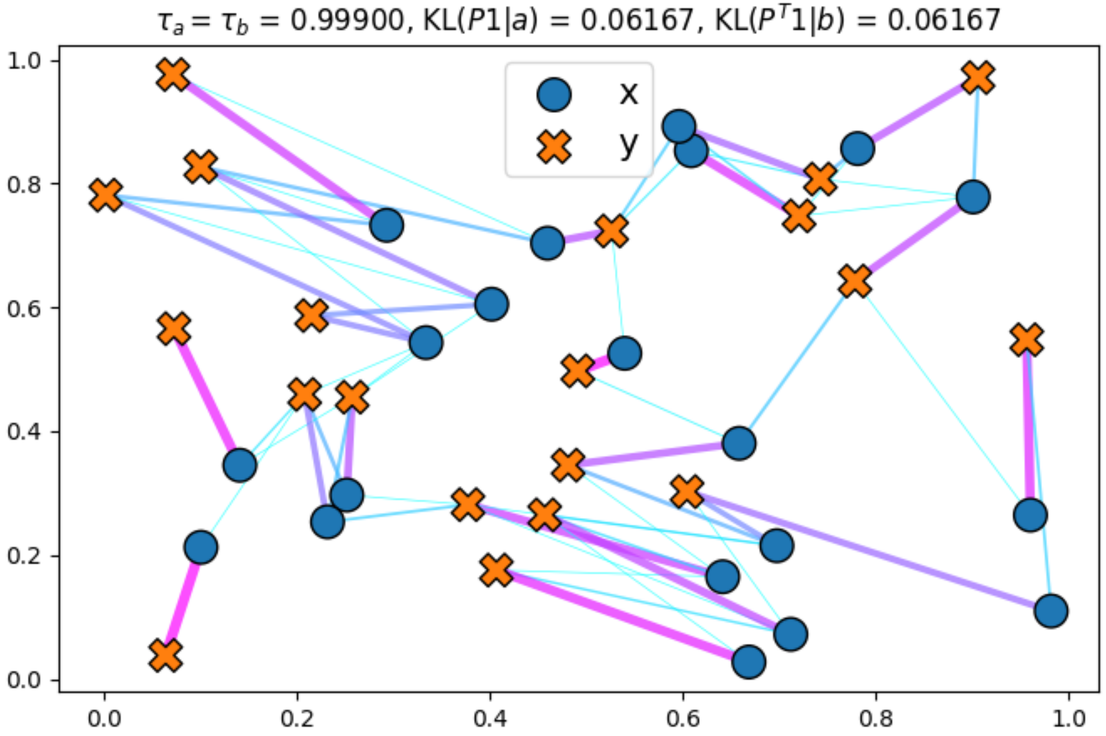


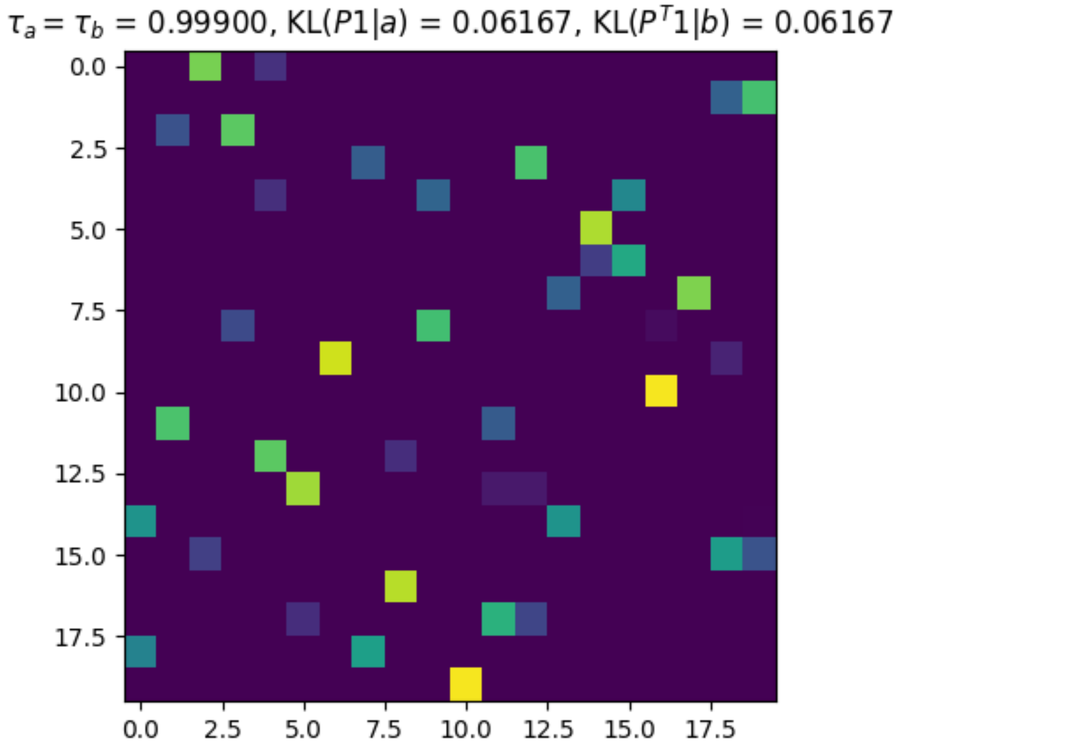


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]



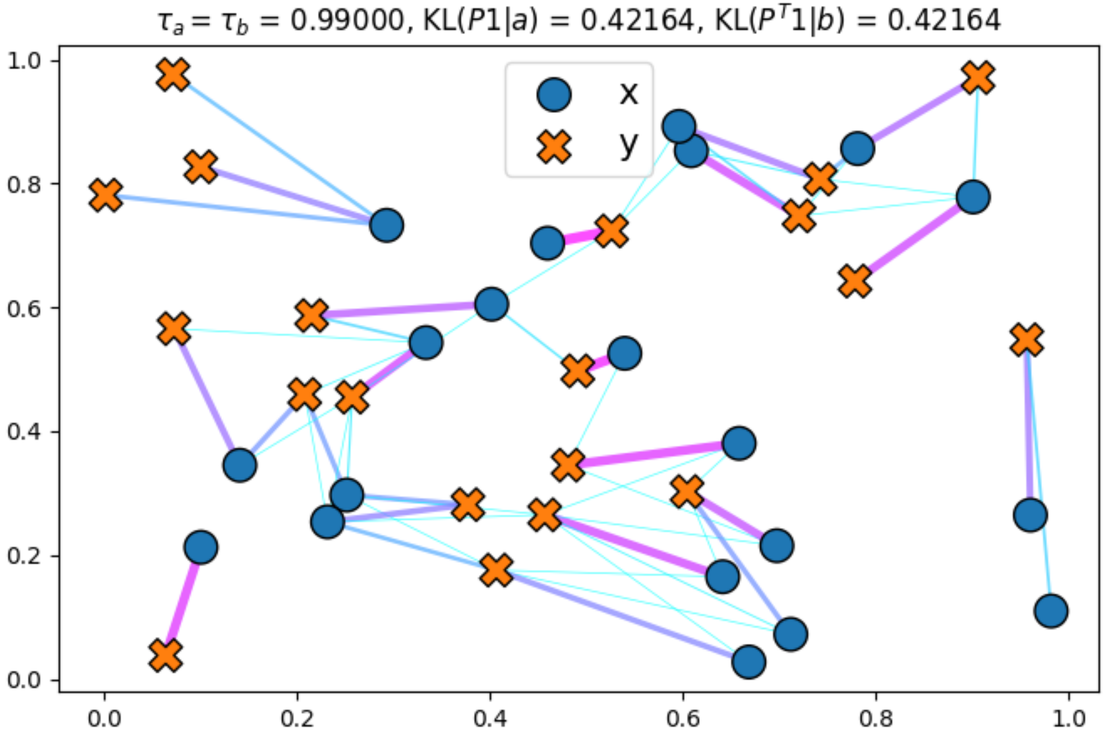


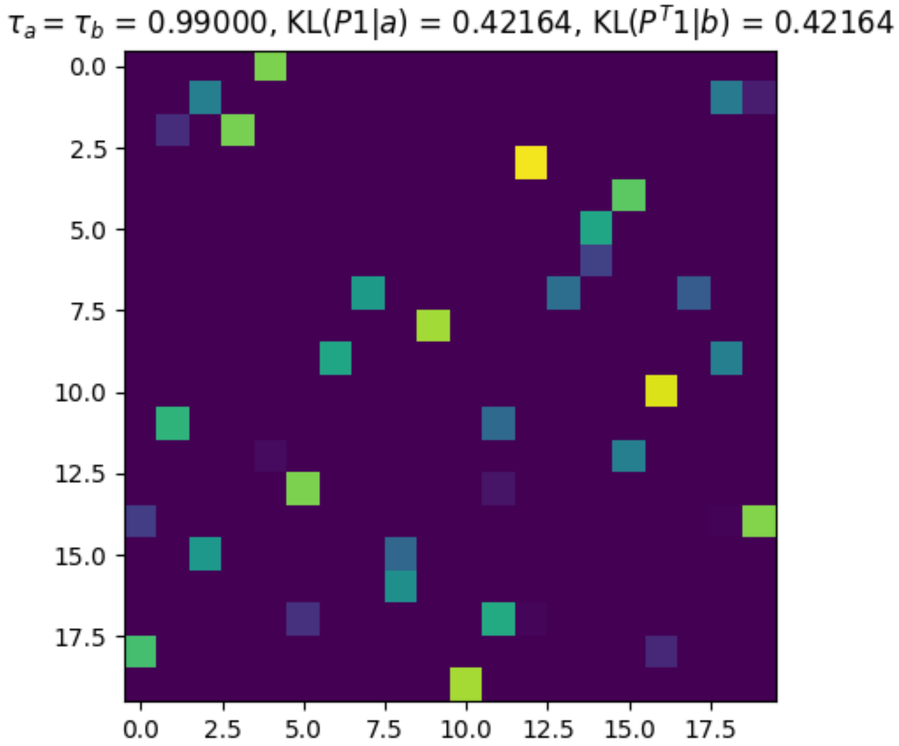
Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$



[Colab4]



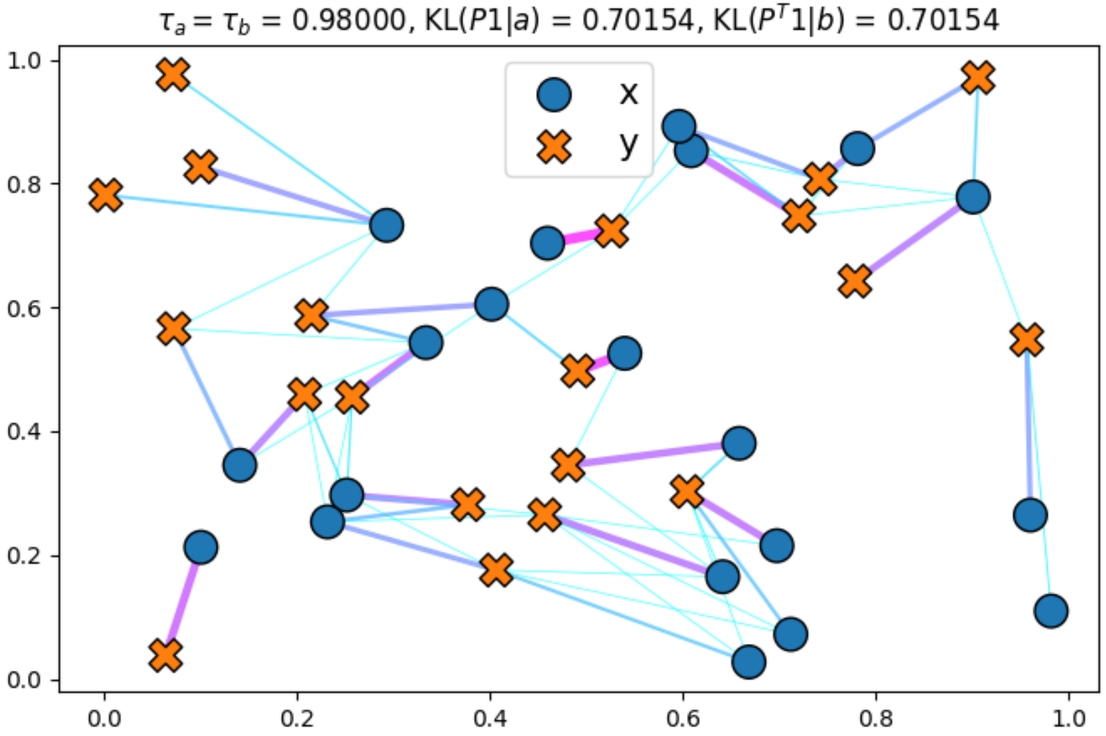


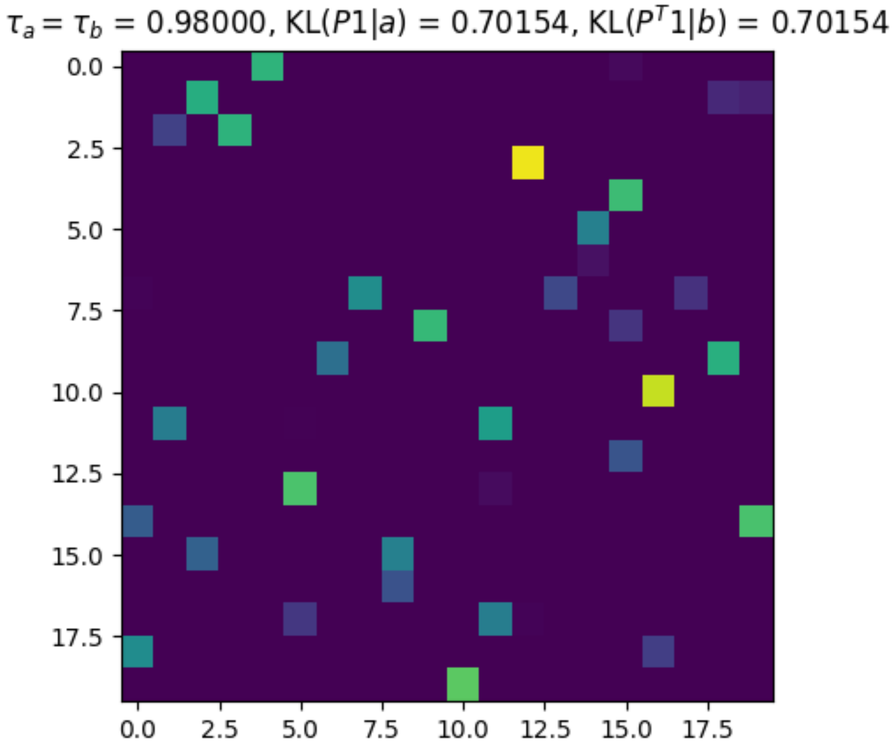
Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$



[Colab4]

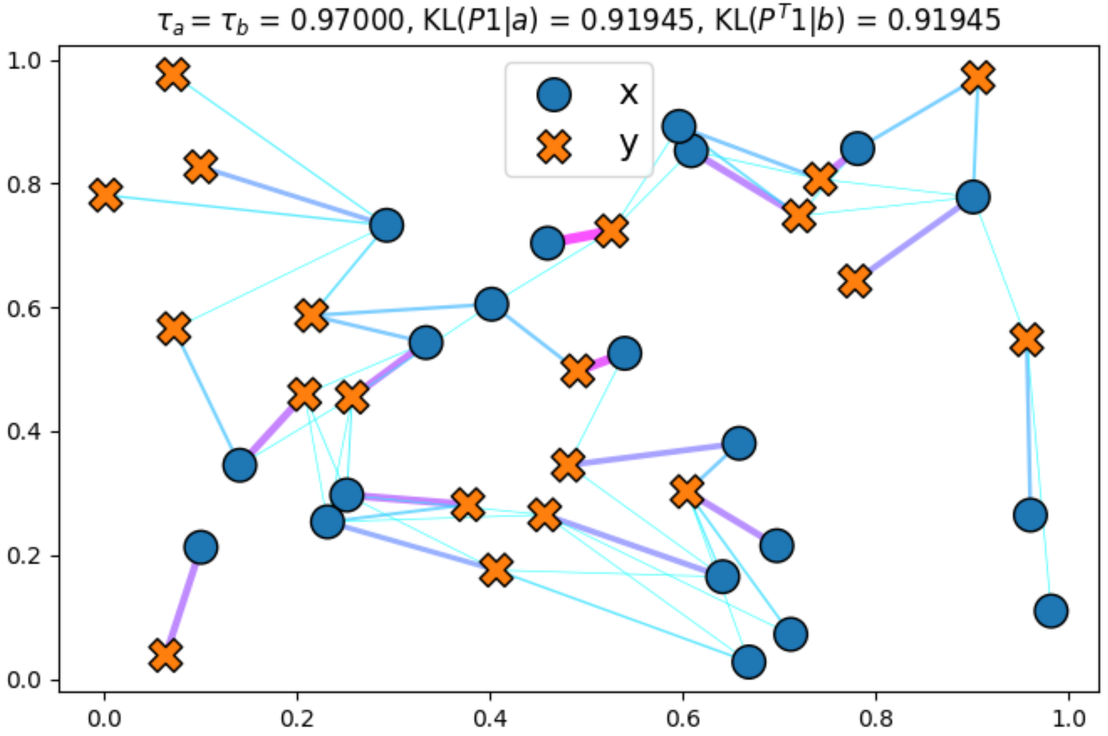


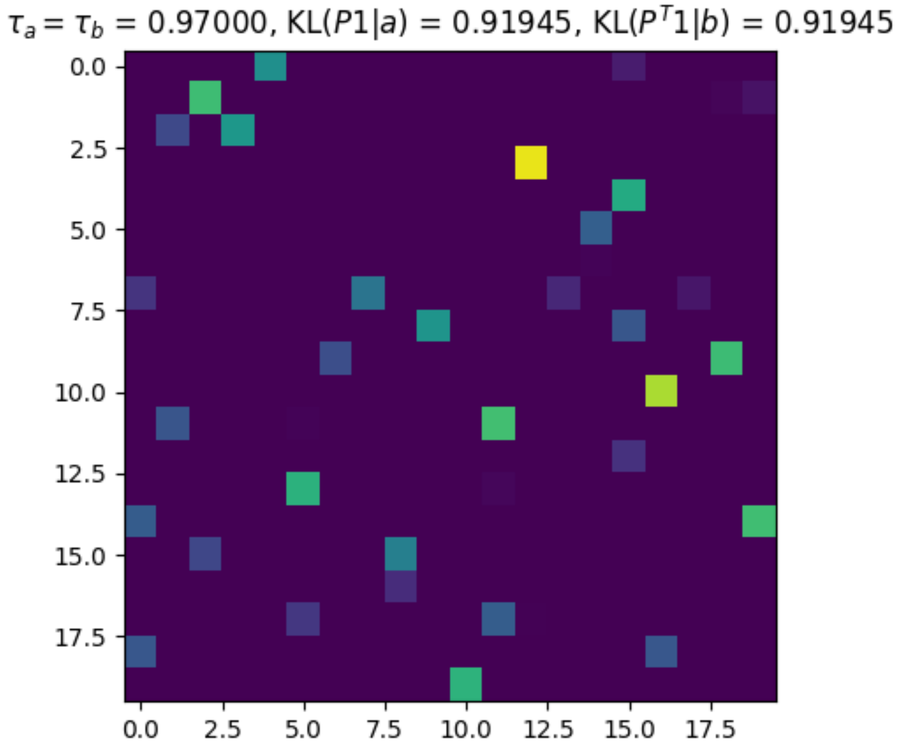


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]

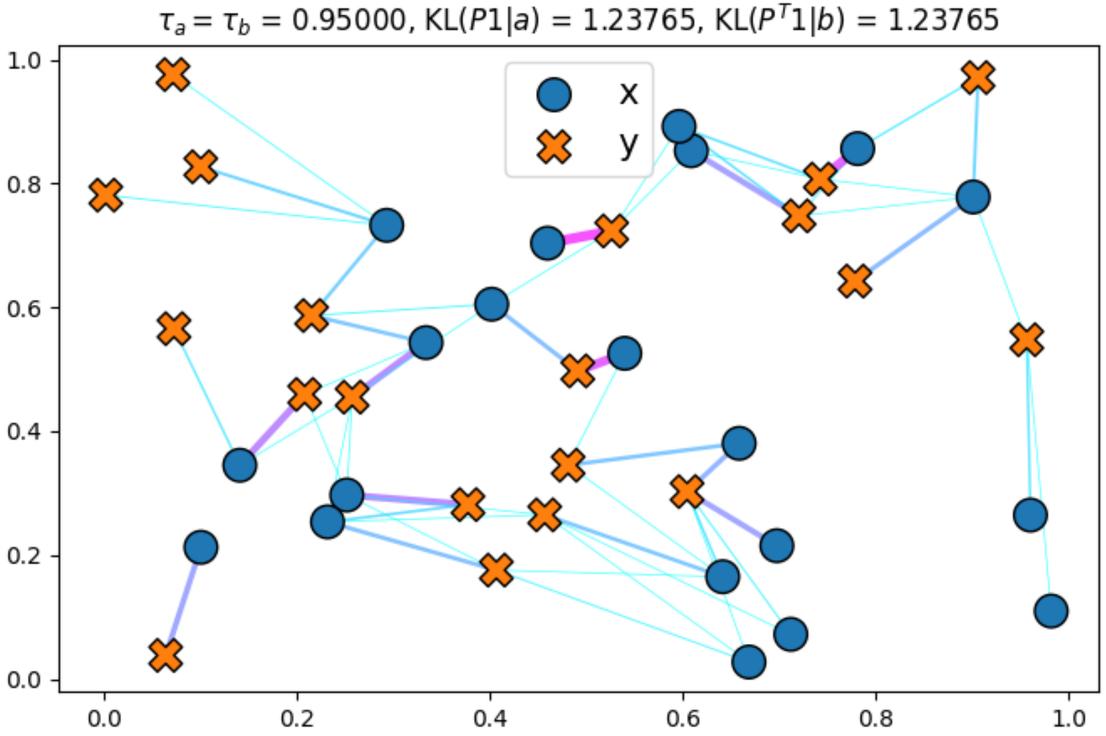


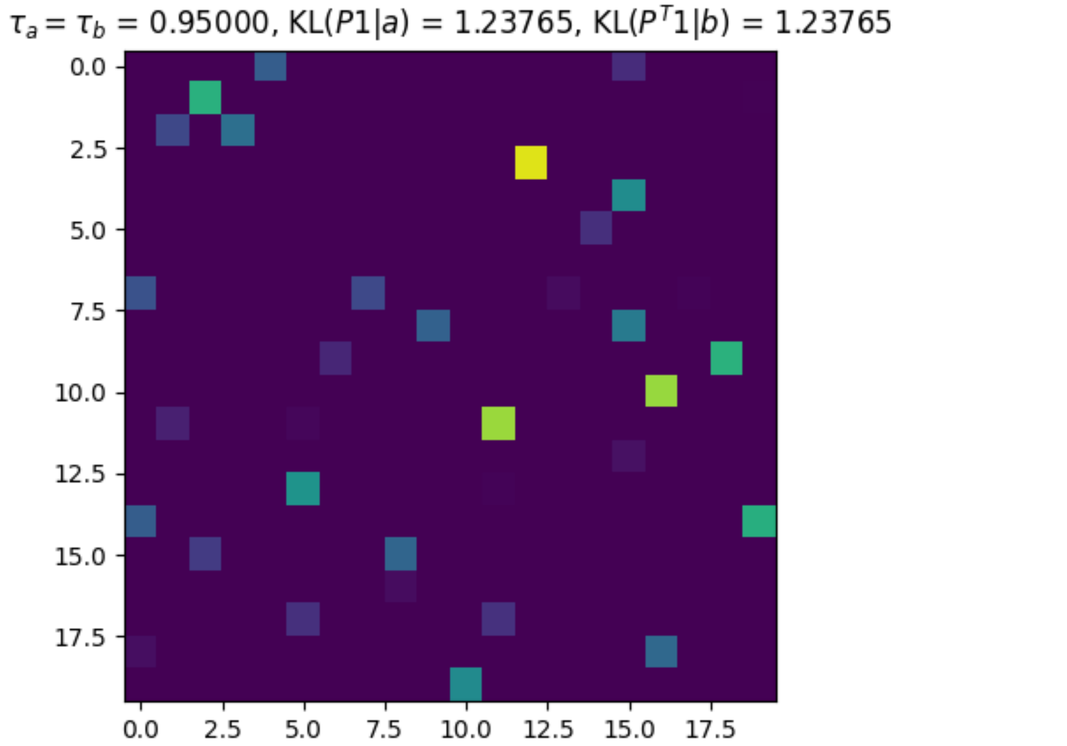


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]

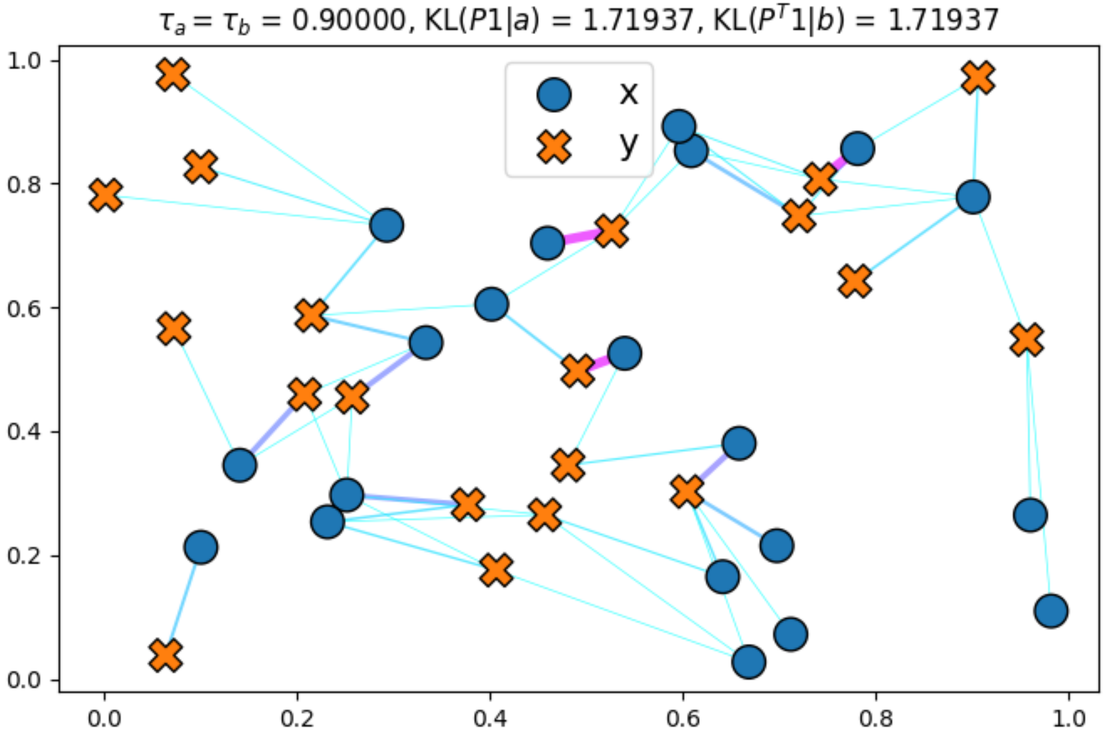


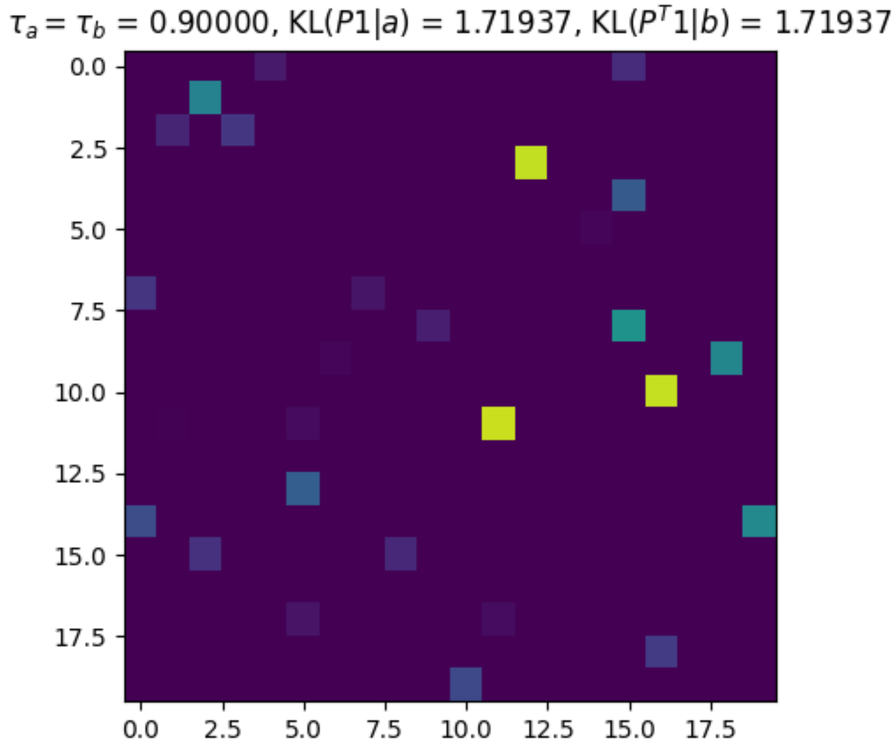


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]

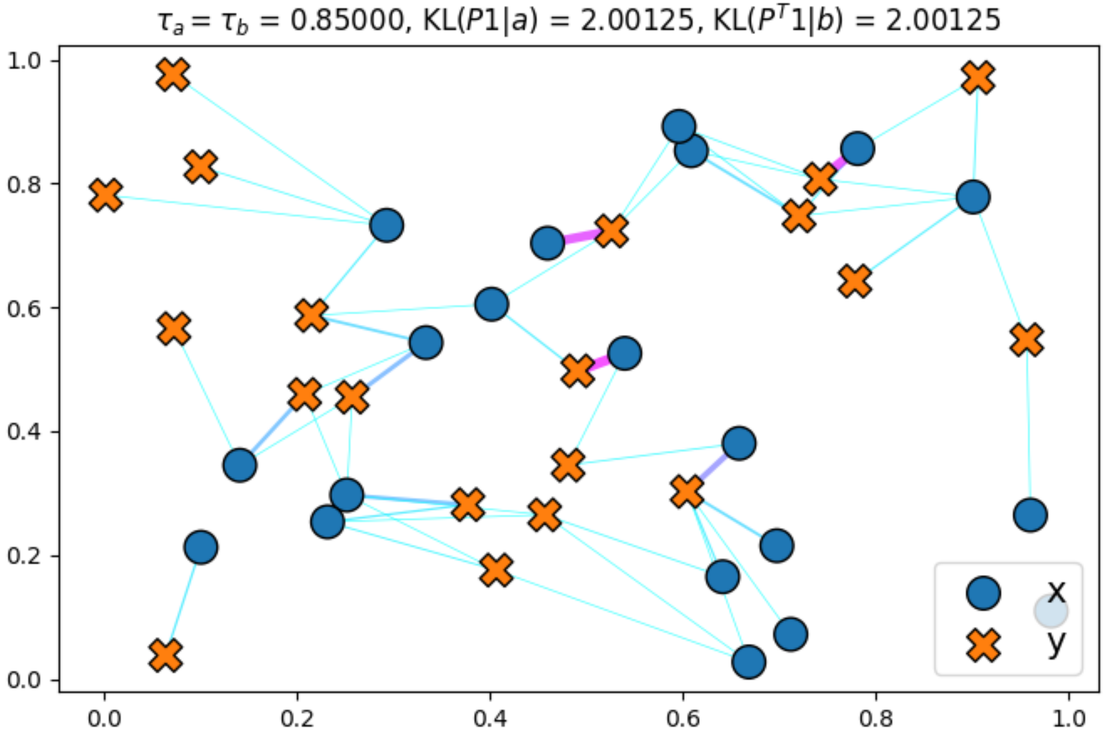


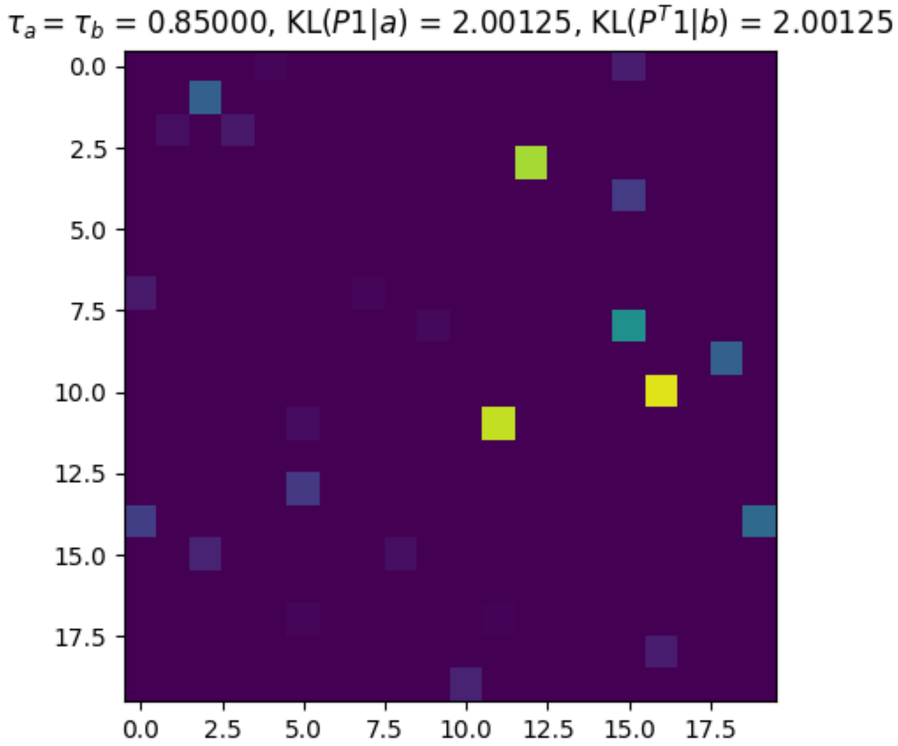


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]

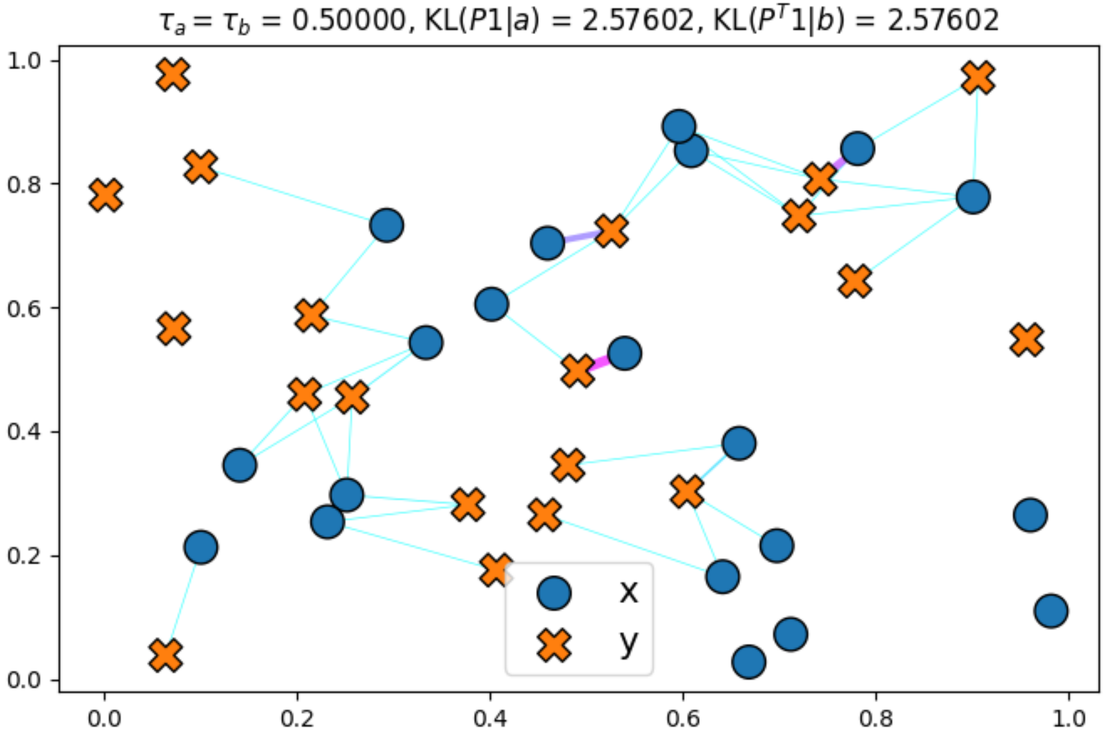


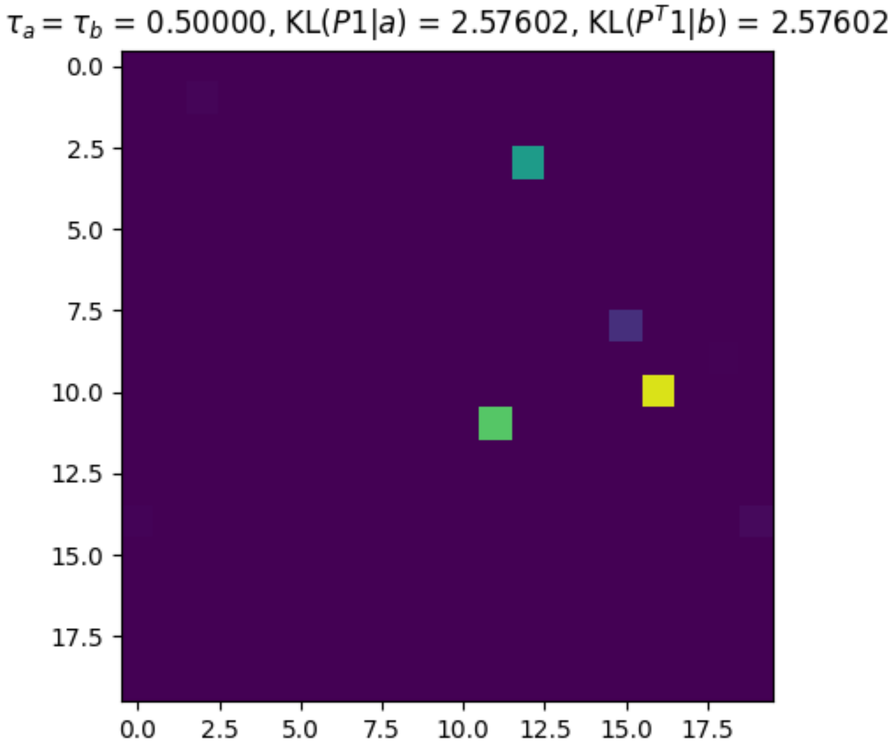


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]

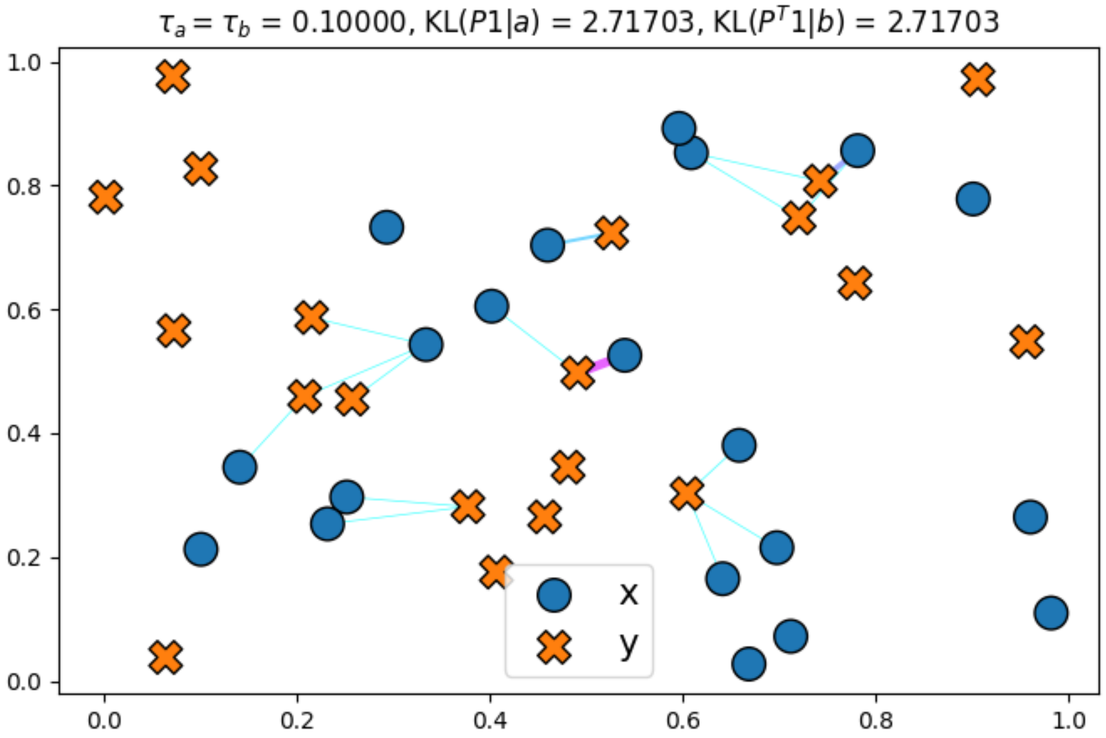


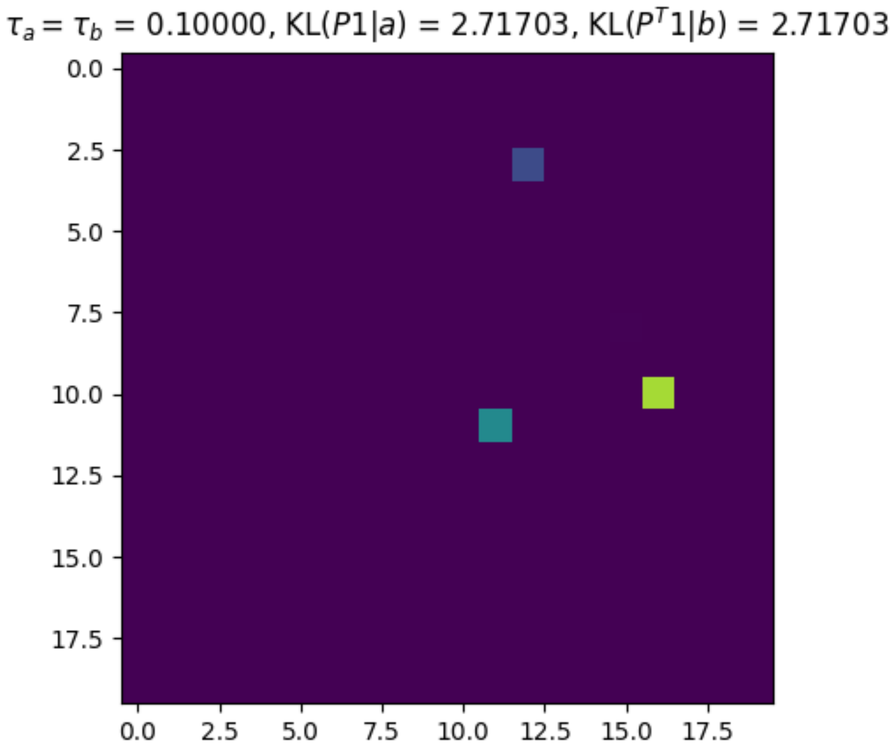


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]

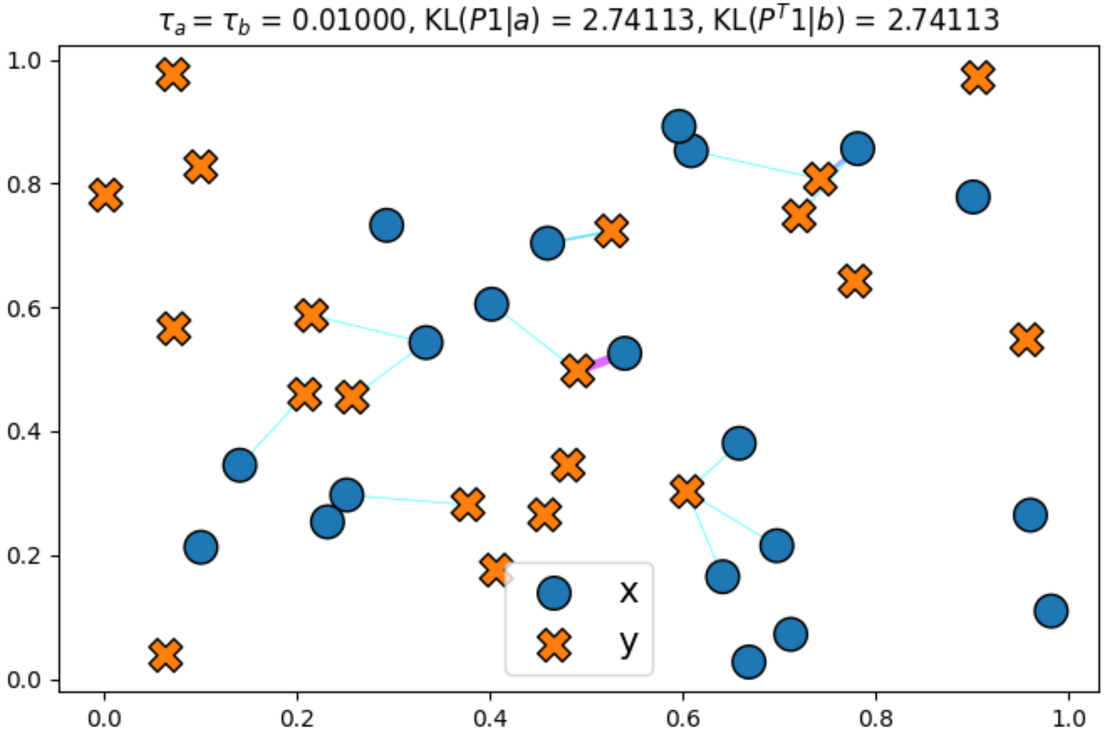


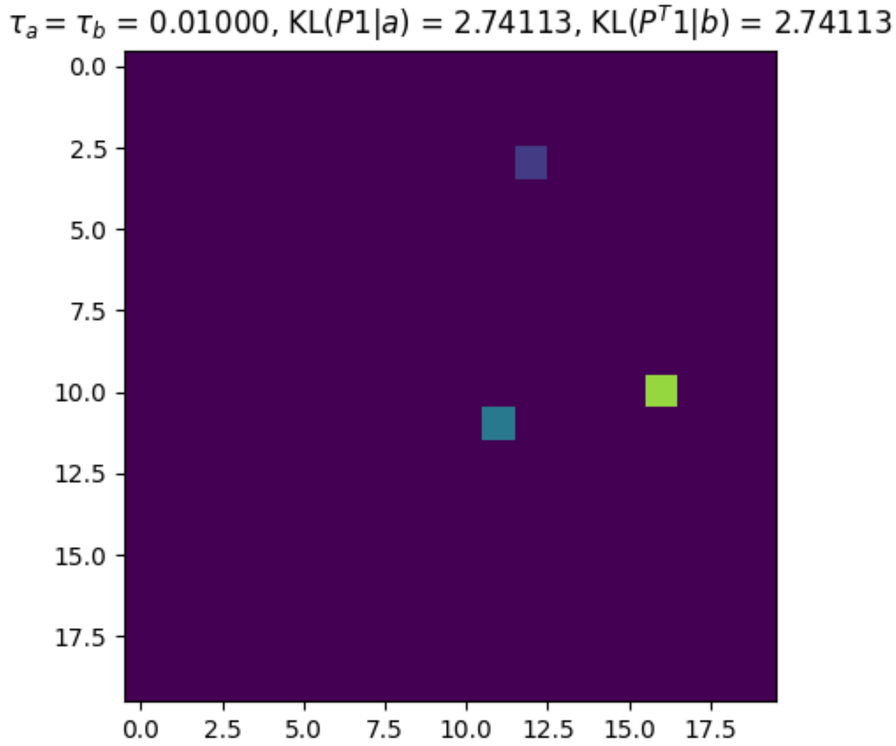


Unbalanced Sinkhorn

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, M_{XY} \rangle + \rho_a \mathrm{KL}(P\mathbf{1}_m \| a) + \rho_b \mathrm{KL}(P^T \mathbf{1}_n \| b) - \gamma E(P)$$

[Colab4]

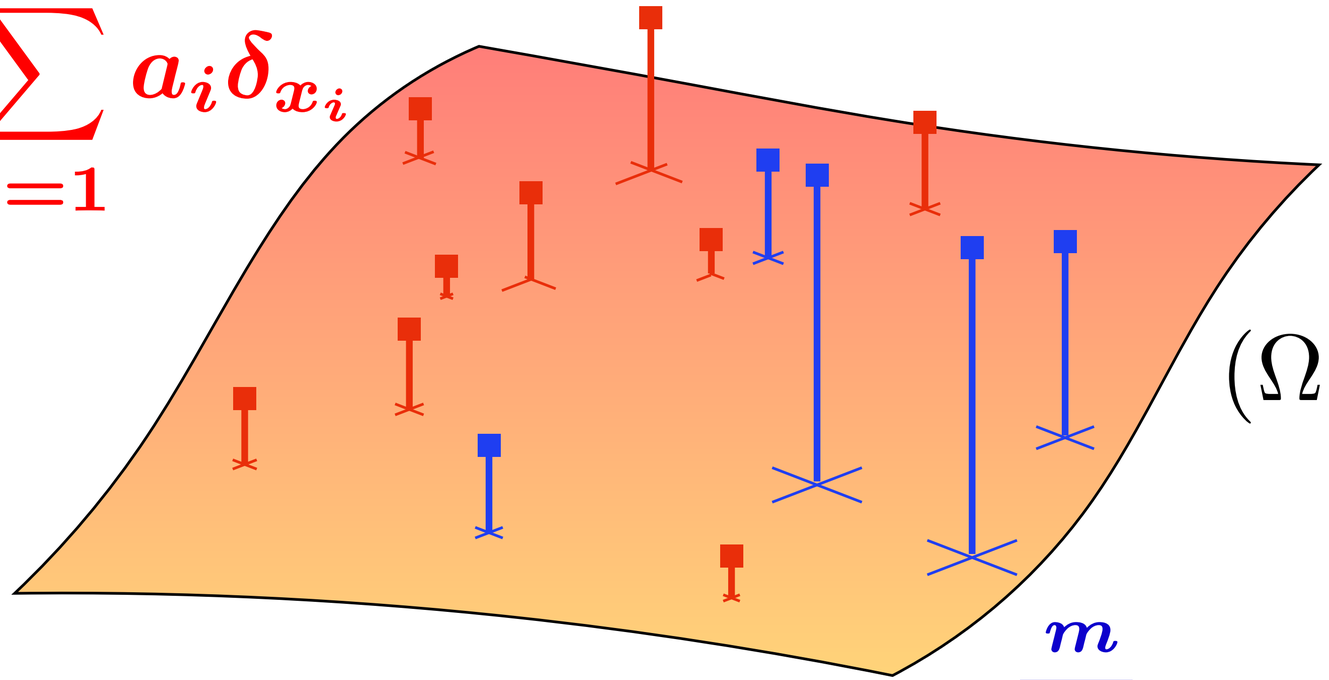




Differentiating the OT Cost (a.k.a. Wasserstein)

$$W((a, X), (b, Y))$$

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



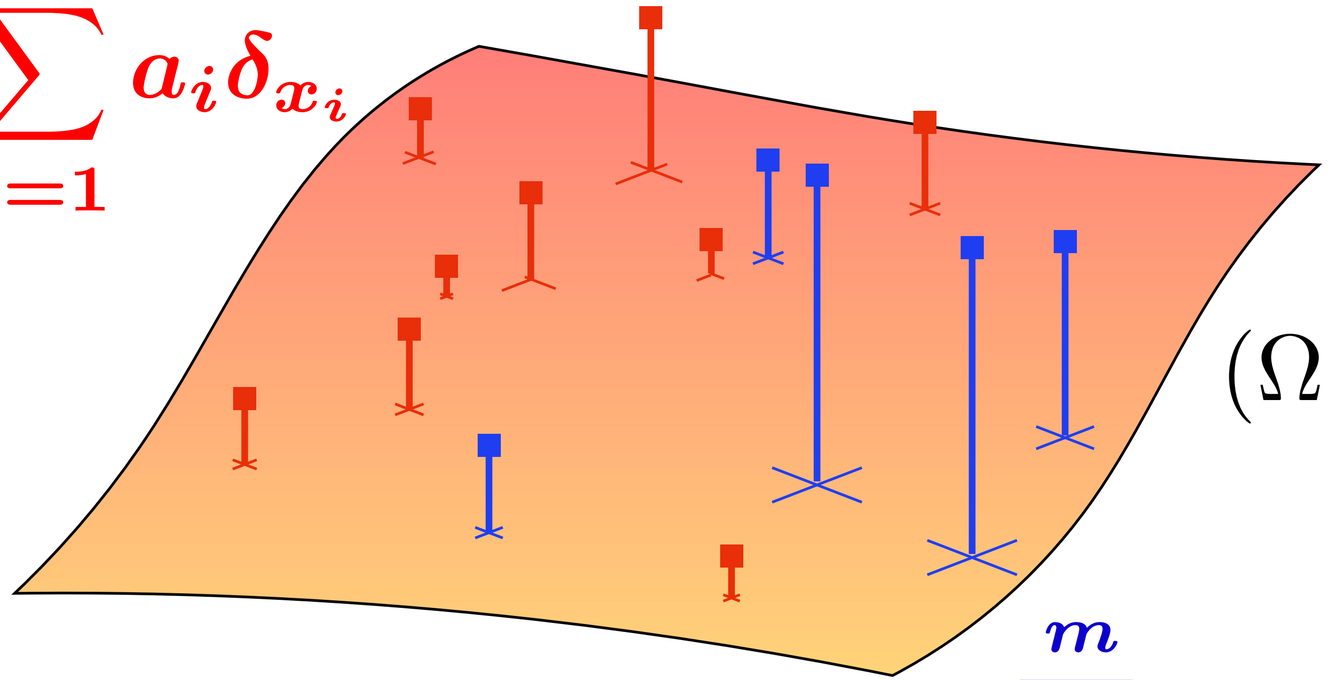
$$(\Omega, D)$$

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

Differentiating the OT Cost (a.k.a. Wasserstein)

$$W((a + \Delta a, X), (b, Y)) = W((a, X), (b, Y)) + \nabla_a W^T \Delta a$$

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



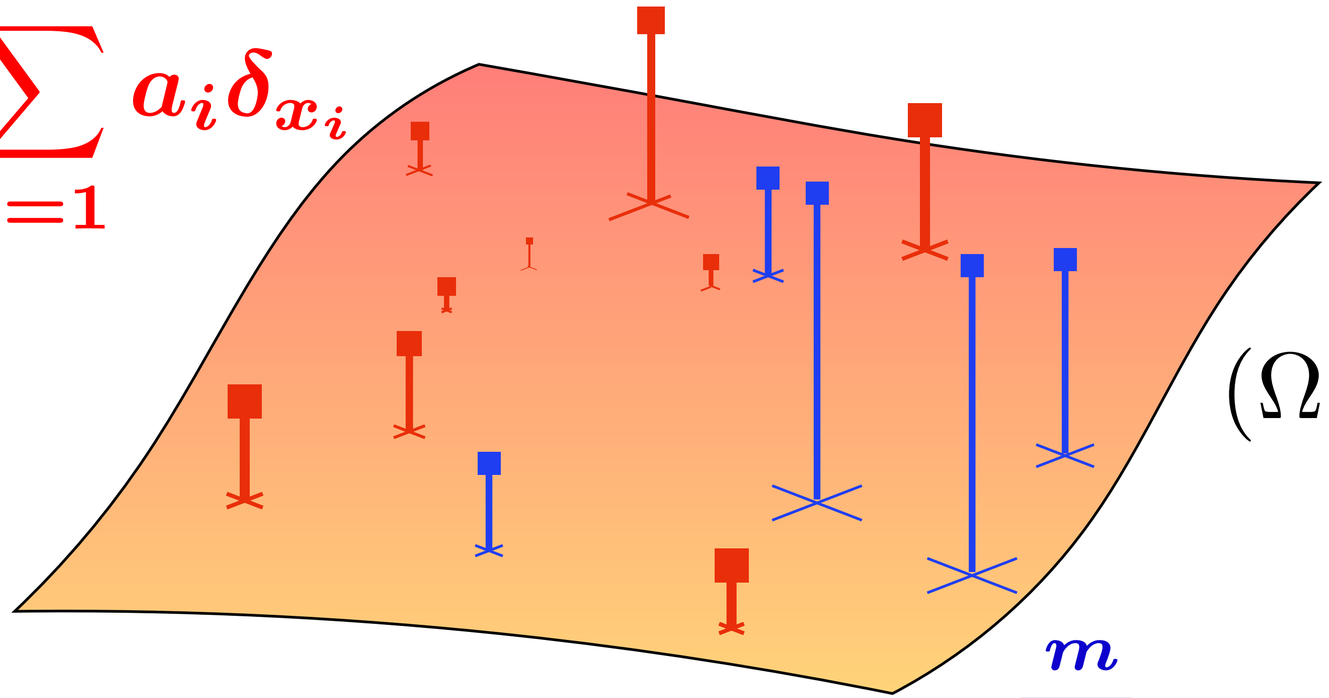
(Ω, D)

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

Differentiating the OT Cost (a.k.a. Wasserstein)

$$W((a + \Delta a, X), (b, Y)) = W((a, X), (b, Y)) + \nabla_a W^T \Delta a$$

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



(Ω, D)

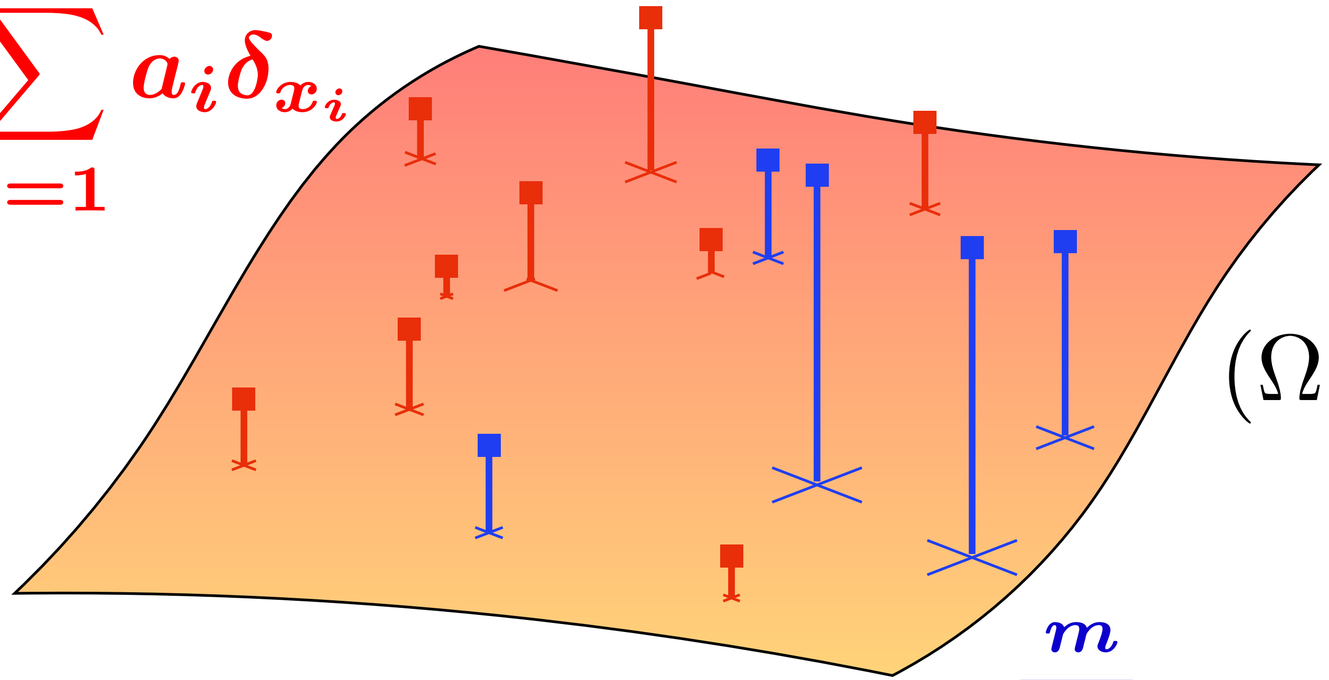
$$a \leftarrow a + \Delta a$$

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

Differentiating the OT Cost (a.k.a. Wasserstein)

$$W((a, X + \Delta X), (b, Y)) = W((a, X), (b, Y)) + \langle \nabla_X W, \Delta X \rangle$$

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



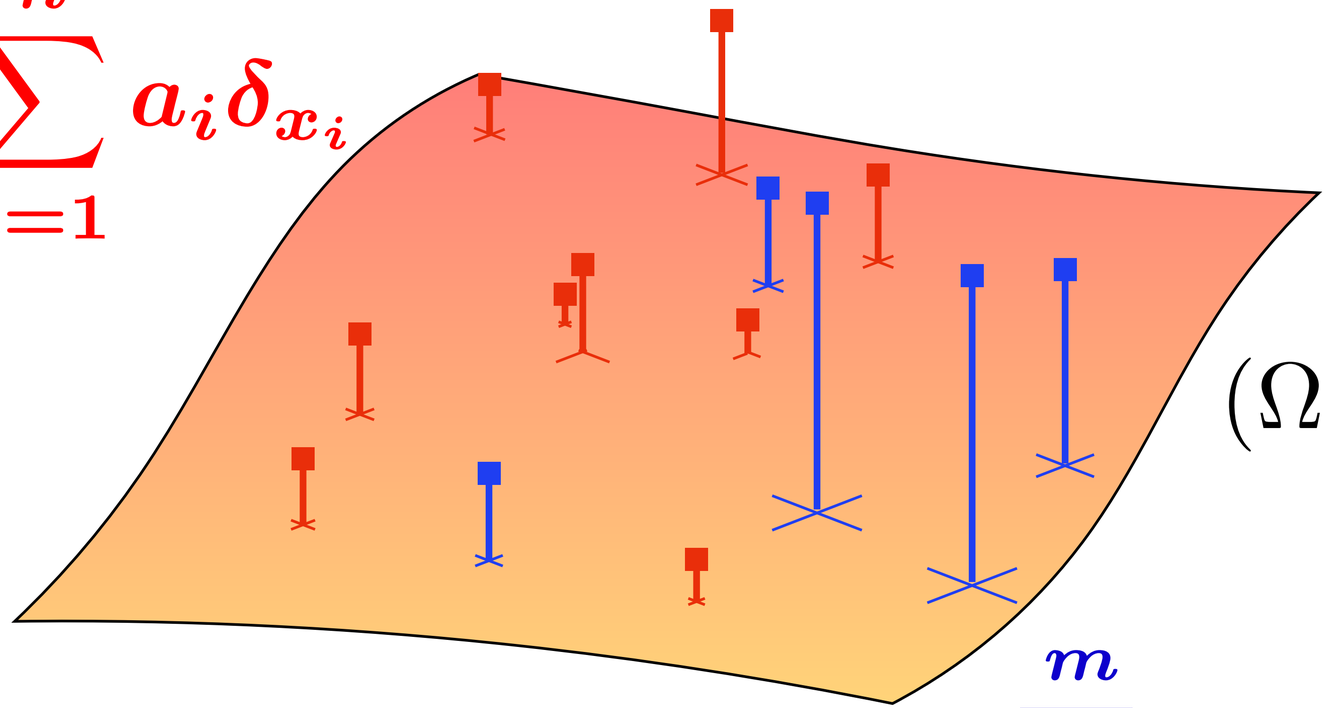
(Ω, D)

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

Differentiating the OT Cost (a.k.a. Wasserstein)

$$W((a, X + \Delta X), (b, Y)) = W((a, X), (b, Y)) + \langle \nabla_X W, \Delta X \rangle$$

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



(Ω, D)

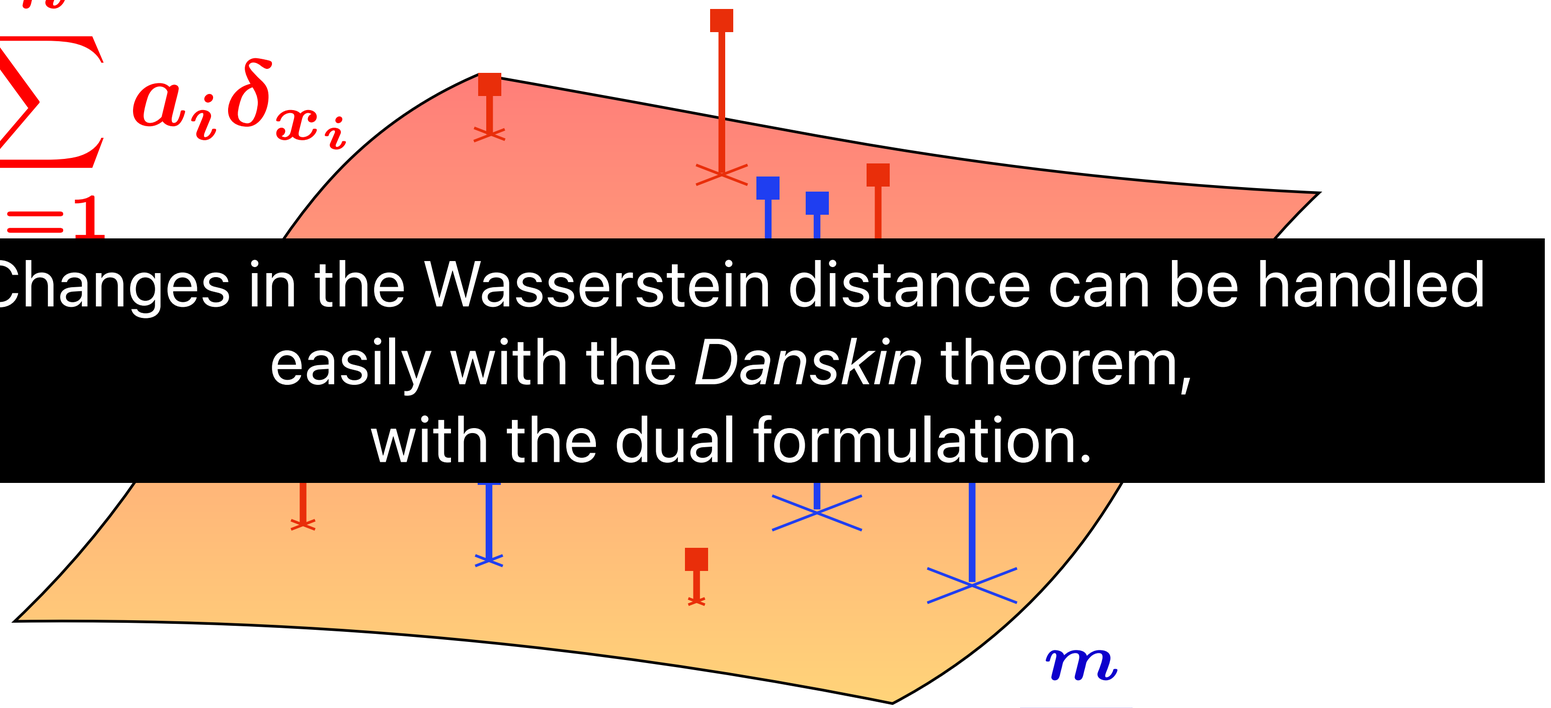
$$X \leftarrow X + \Delta X$$

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

Differentiating the OT Cost (a.k.a. Wasserstein)

$$W((a, X + \Delta X), (b, Y)) = W((a, X), (b, Y)) + \langle \nabla_X W, \Delta X \rangle$$

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



Changes in the Wasserstein distance can be handled easily with the *Danskin* theorem, with the dual formulation.

$$X \leftarrow X + \Delta X$$

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

Reminders on the Danskin Theorem

Use $\boldsymbol{x}$ for either $\boldsymbol{X}$ or $\boldsymbol{a}$, and $\boldsymbol{z}$ for dual variables.

$$f(\boldsymbol{x}) = \min_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z})$$

Reminders on the Danskin Theorem

Use $\boldsymbol{x}$ for either $\boldsymbol{X}$ or $\boldsymbol{a}$, and $\boldsymbol{z}$ for dual variables.

$$f(\boldsymbol{x}) = \min_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z})$$

$$\boldsymbol{z}^\star(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z})$$

Reminders on the Danskin Theorem

Use $\boldsymbol{x}$ for either $\boldsymbol{X}$ or $\boldsymbol{a}$, and $\boldsymbol{z}$ for dual variables.

$$f(\boldsymbol{x}) = \min_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

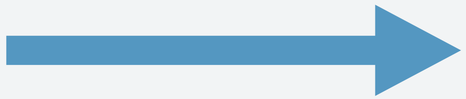
$$\boldsymbol{z}^\star(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z})$$

Reminders on the Danskin Theorem

Use $\boldsymbol{x}$ for either $\boldsymbol{X}$ or $\boldsymbol{a}$, and $\boldsymbol{z}$ for dual variables.

$$f(\boldsymbol{x}) = \min_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

$$\boldsymbol{z}^\star(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z})$$



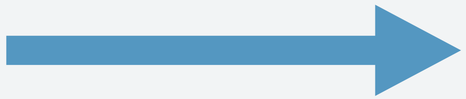
$$\nabla f(\boldsymbol{x}) = \nabla_1 g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

Reminders on the Danskin Theorem

Use $\boldsymbol{x}$ for either $\boldsymbol{X}$ or $\boldsymbol{a}$, and $\boldsymbol{z}$ for dual variables.

$$f(\boldsymbol{x}) = \min_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

$$\boldsymbol{z}^\star(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z})$$



$$\nabla f(\boldsymbol{x}) = \nabla_1 g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

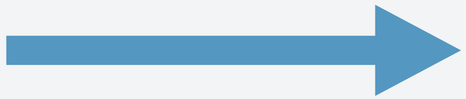
$$\nabla f(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x})) = \nabla_1 g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x})) + J_{\boldsymbol{z}^\star}(\boldsymbol{x})^T \nabla_2 g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

Reminders on the Danskin Theorem

Use $\boldsymbol{x}$ for either $\boldsymbol{X}$ or $\boldsymbol{a}$, and $\boldsymbol{z}$ for dual variables.

$$f(\boldsymbol{x}) = \min_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

$$\boldsymbol{z}^\star(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{z}} g(\boldsymbol{x}, \boldsymbol{z})$$

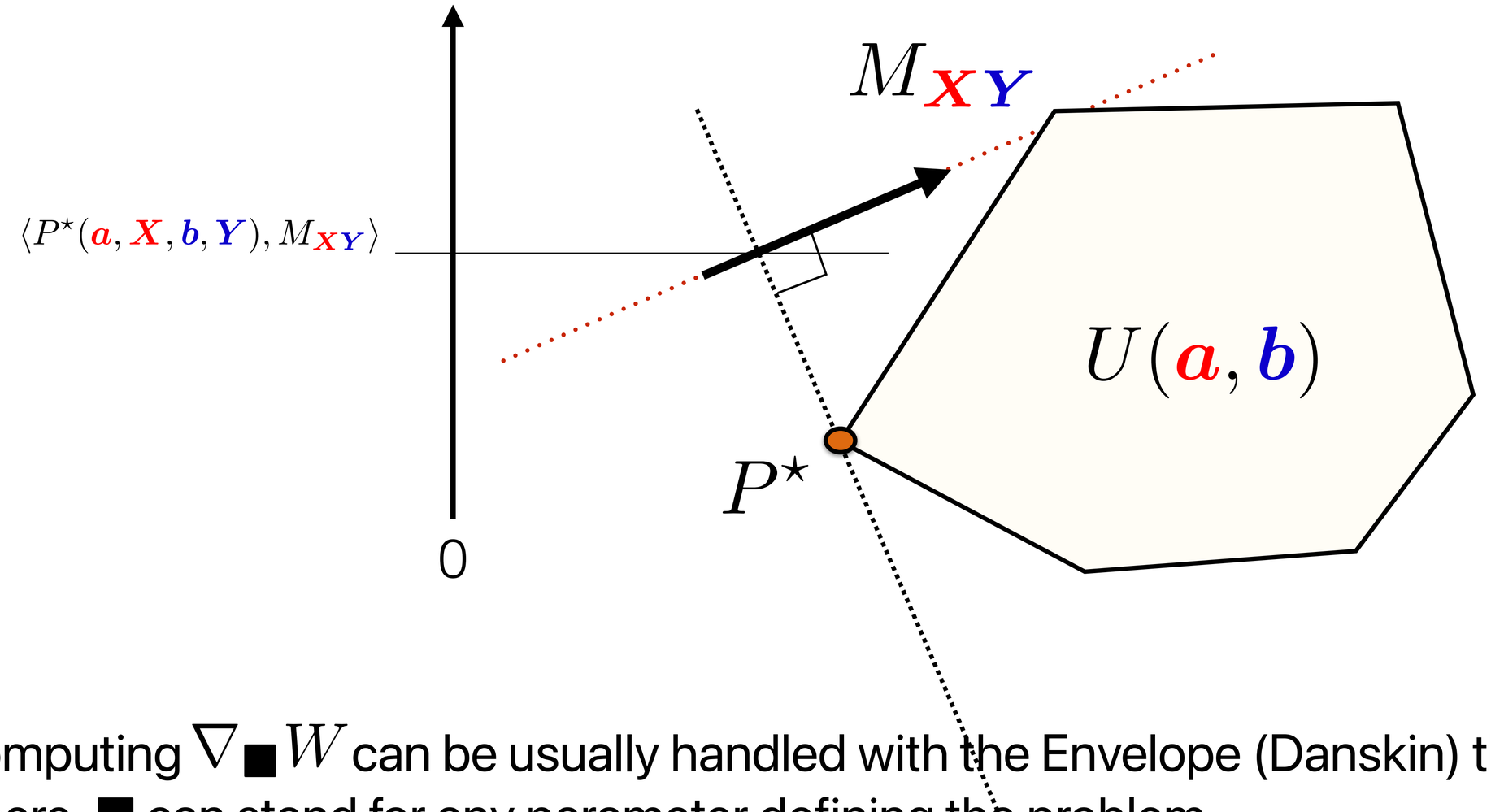


$$\nabla f(\boldsymbol{x}) = \nabla_1 g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

$$\nabla f(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x})) = \nabla_1 g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x})) + J_{\boldsymbol{z}^\star}(\boldsymbol{x})^T \nabla_2 g(\boldsymbol{x}, \boldsymbol{z}^\star(\boldsymbol{x}))$$

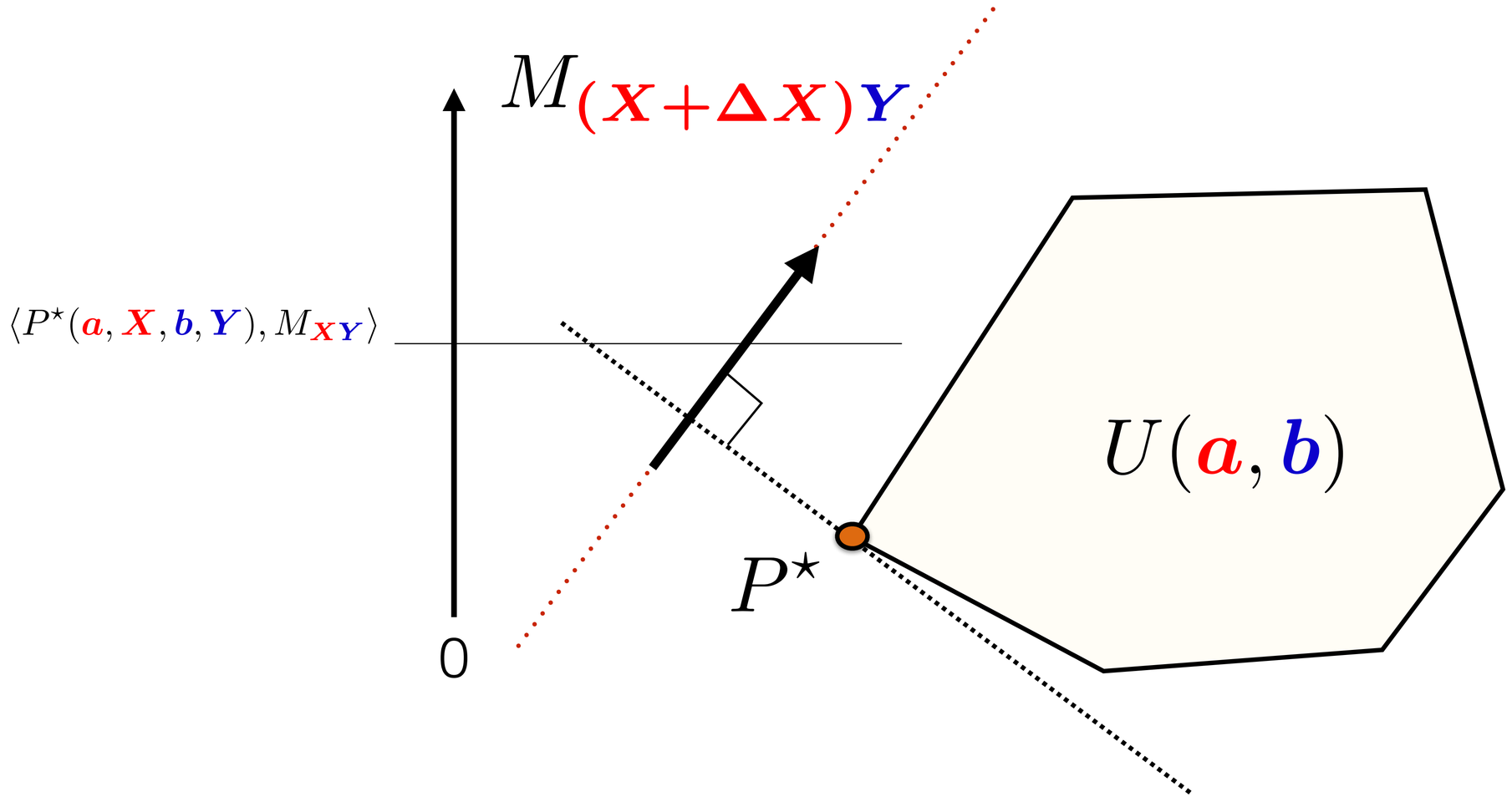
$\boldsymbol{0}$

Differentiating the OT Cost (a.k.a. Wasserstein)



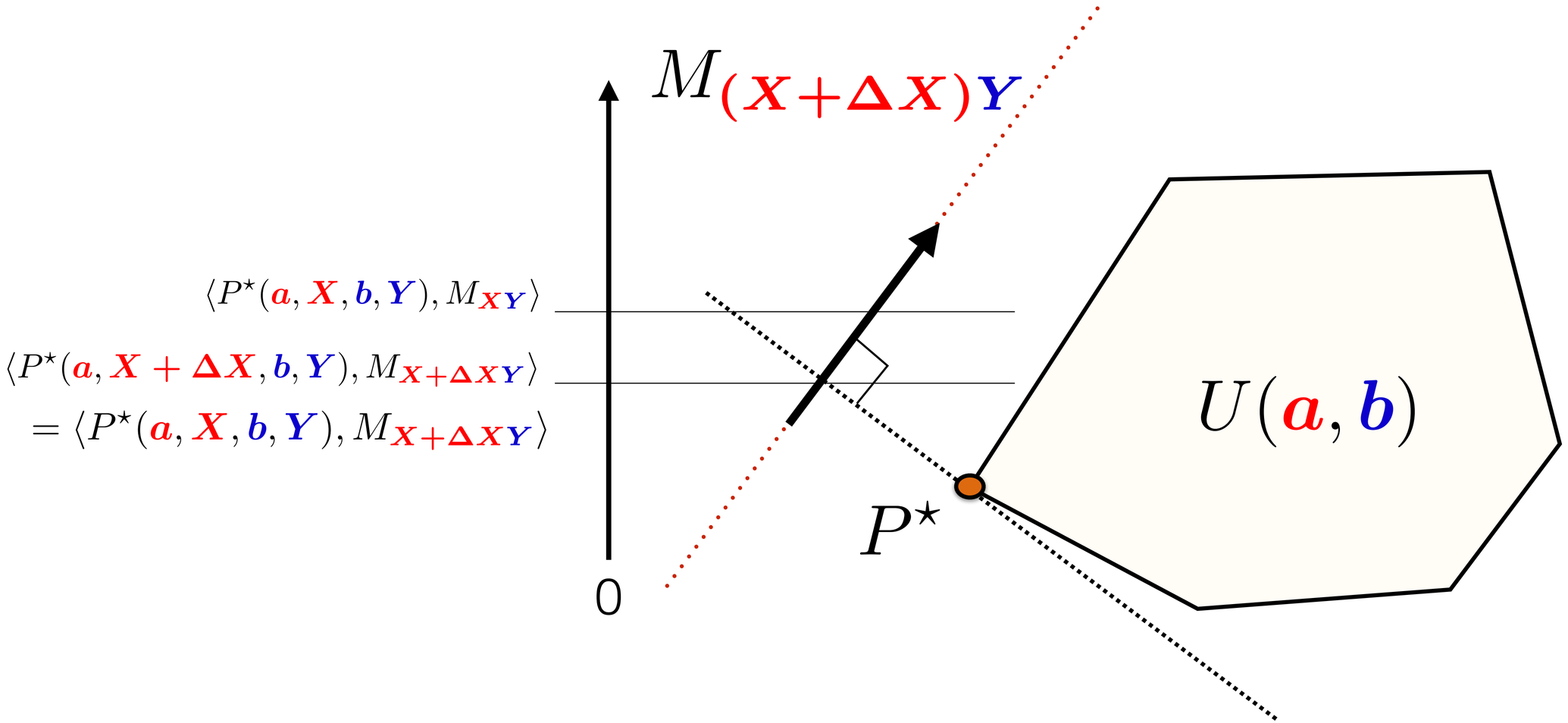
Computing $\nabla_{\blacksquare} W$ can be usually handled with the Envelope (Danskin) theorem, where $\blacksquare$ can stand for any parameter defining the problem.

Differentiating the OT Cost (a.k.a. Wasserstein)



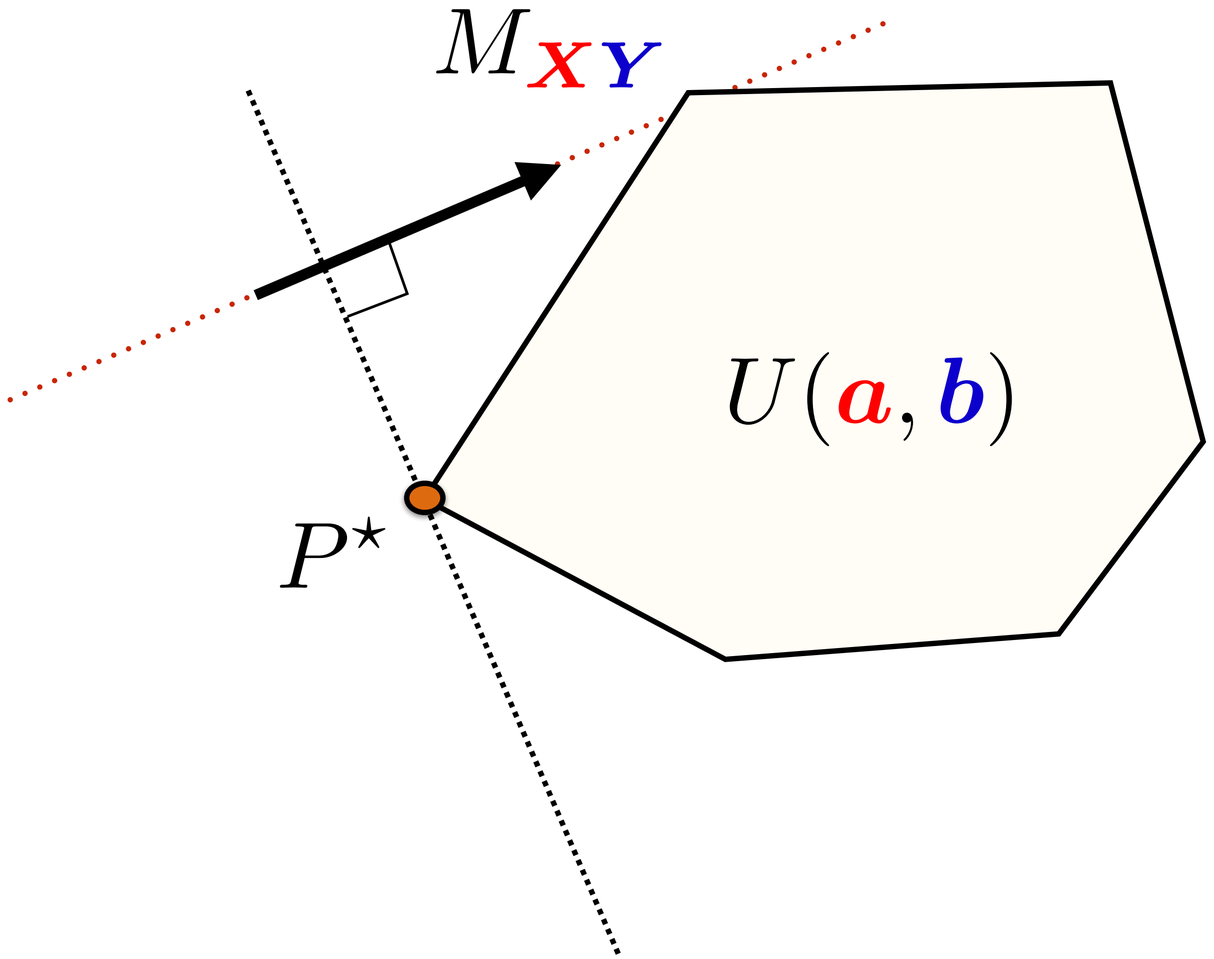
Computing $\nabla_{\blacksquare} W$ can be usually handled with the Envelope (Danskin) theorem, where $\blacksquare$ can stand for any parameter defining the problem.

Differentiating the OT Cost (a.k.a. Wasserstein)

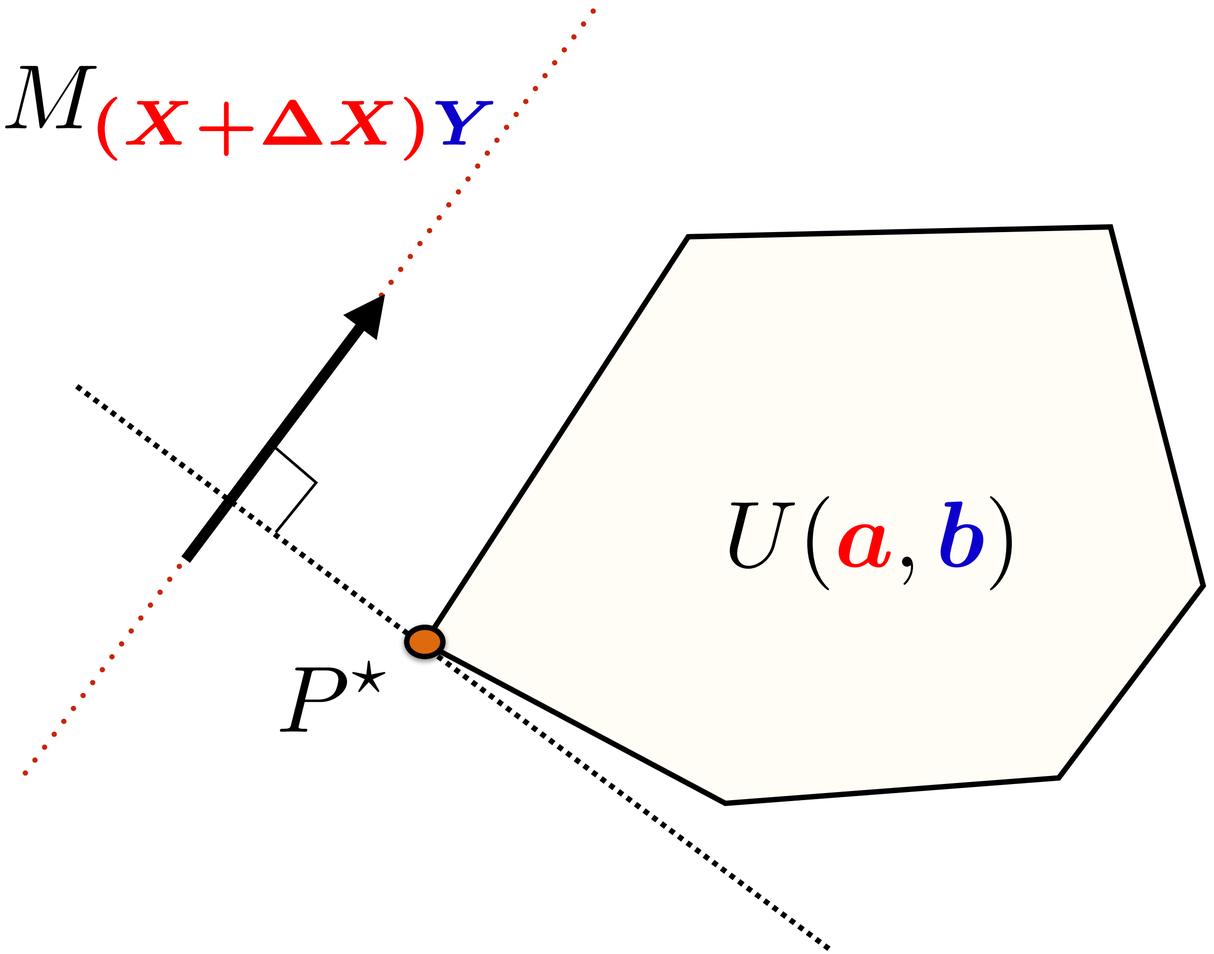


Computing $\nabla_{\blacksquare} W$ can be usually handled with the Envelope (Danskin) theorem, where $\blacksquare$ can stand for any parameter defining the problem.

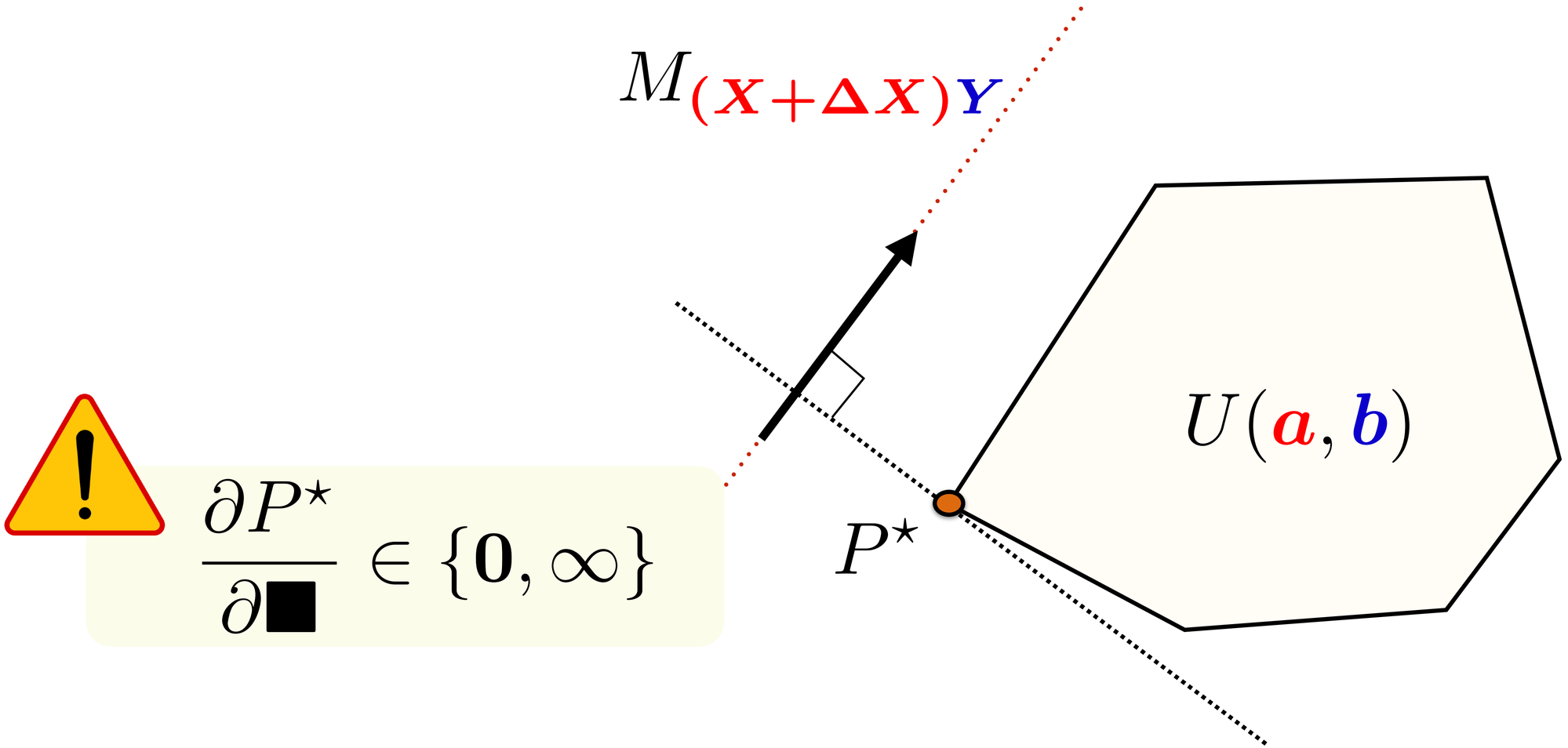
Differentiating the OT Argmin Solution



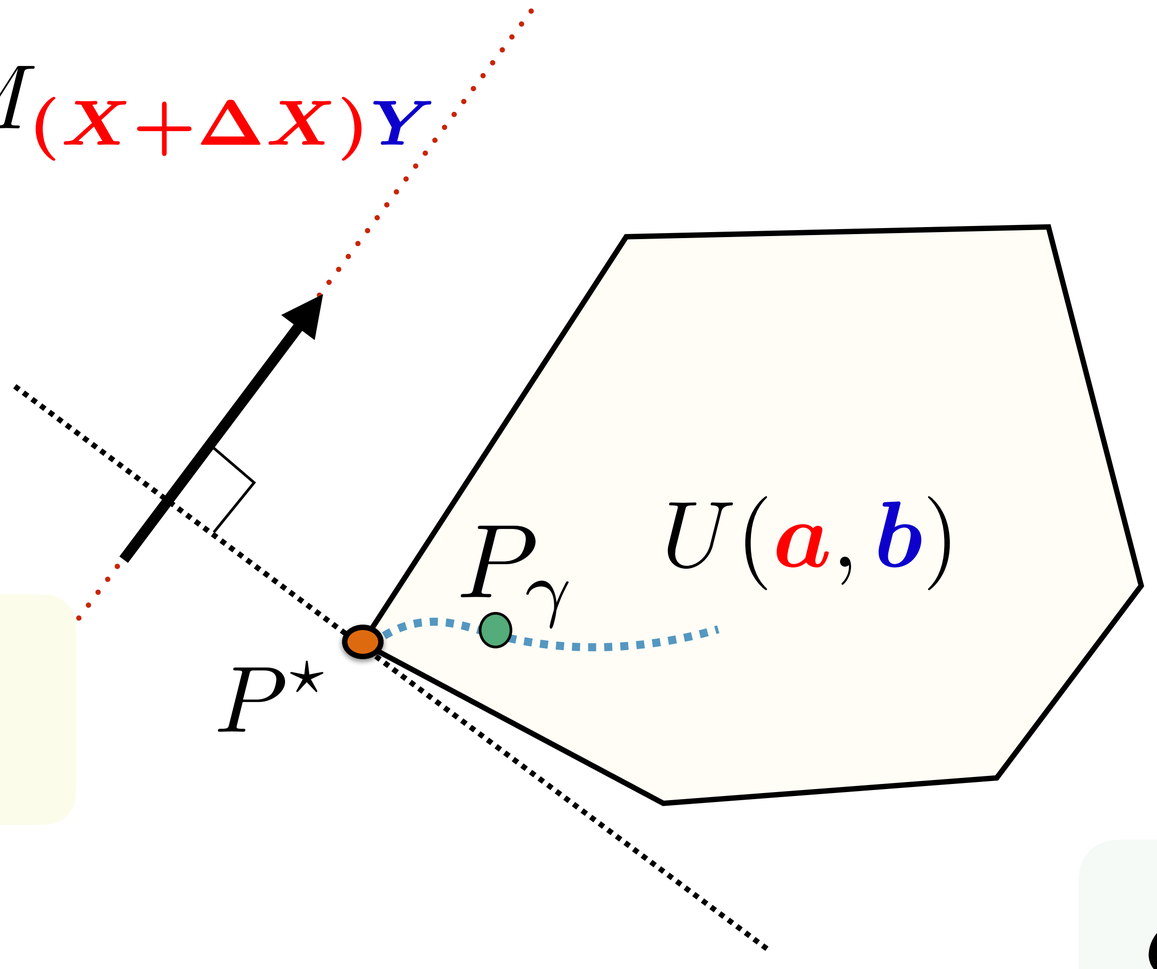
Differentiating the OT Argmin Solution



Differentiating the OT Argmin Solution



Differentiating the OT Argmin Solution

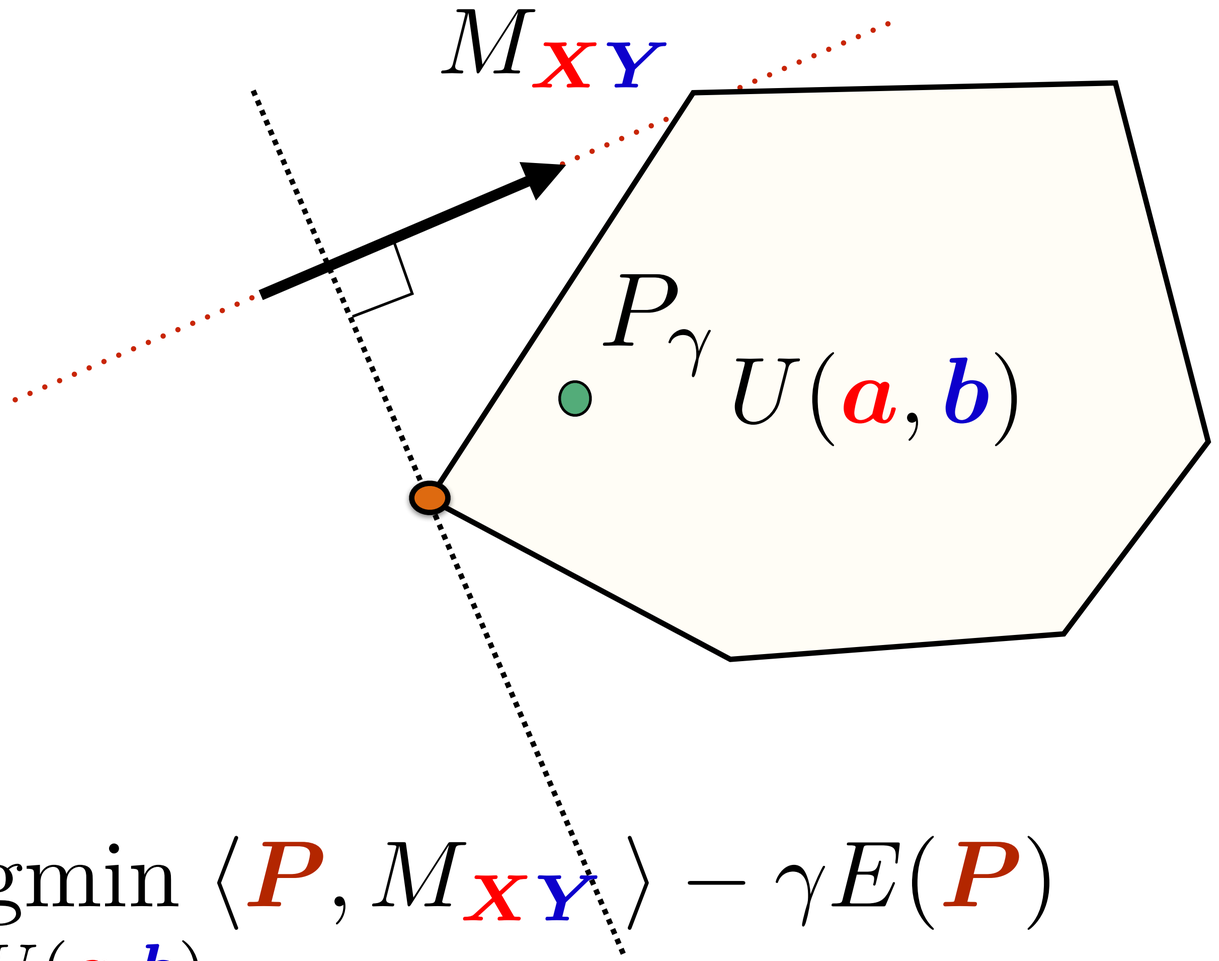


$$\frac{\partial P^\star}{\partial \blacksquare} \in \{0, \infty\}$$

$$P_\gamma \overset{\text{def}}{=} \underset{P \in U(a,b)}{\operatorname{argmin}} \langle P, M_{XY} \rangle - \gamma E(P)$$

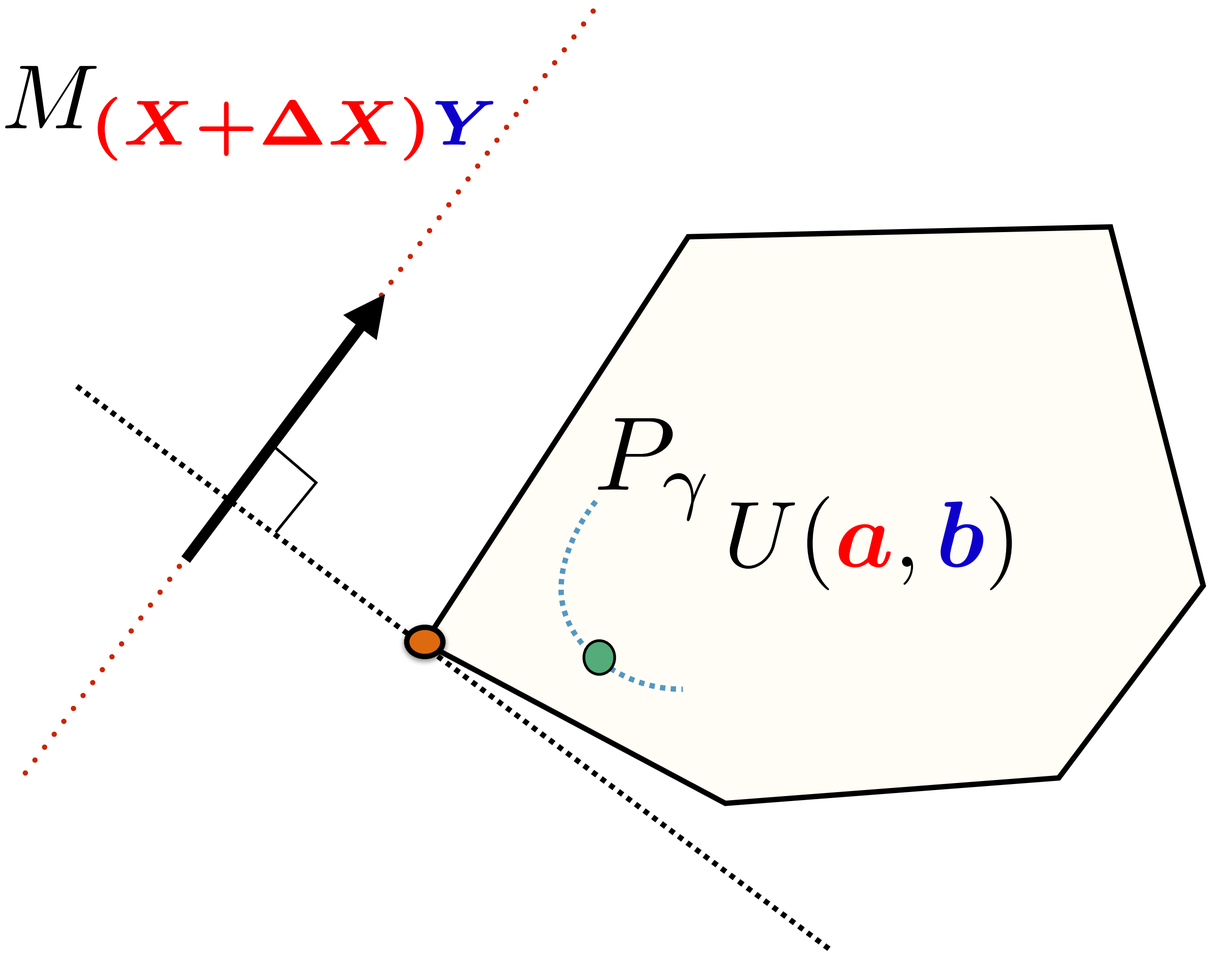
$$\frac{\partial P_\gamma}{\partial \blacksquare} ?$$

Differentiating the OT Argmin Solution



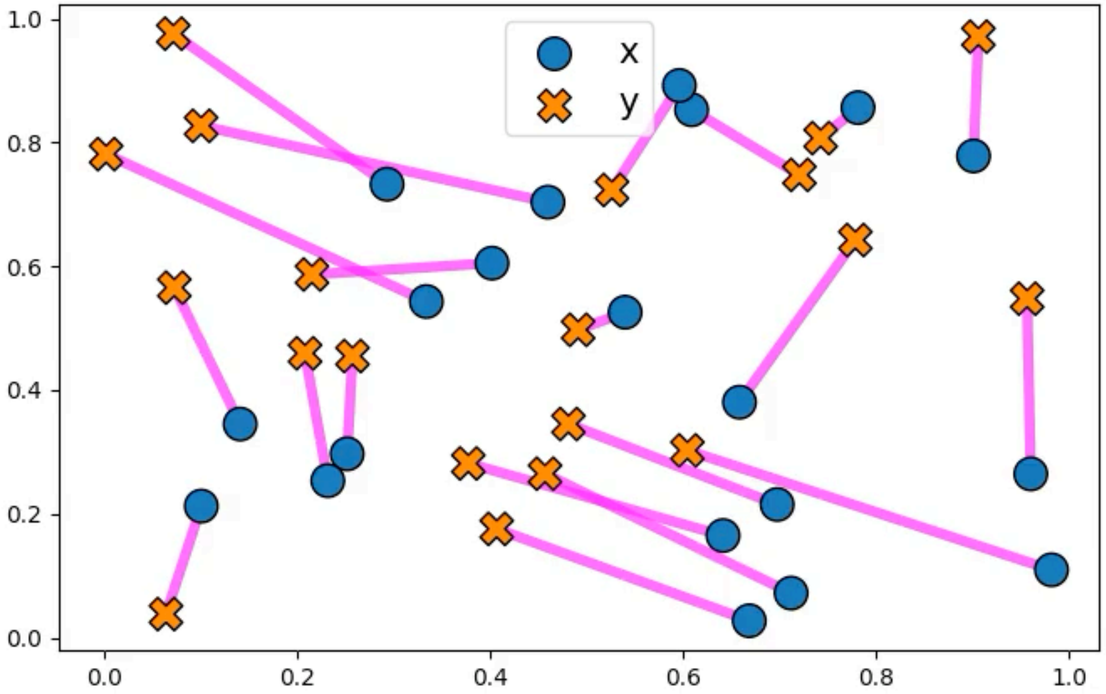
$$P_\gamma \overset{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle - \gamma E(\boldsymbol{P})$$

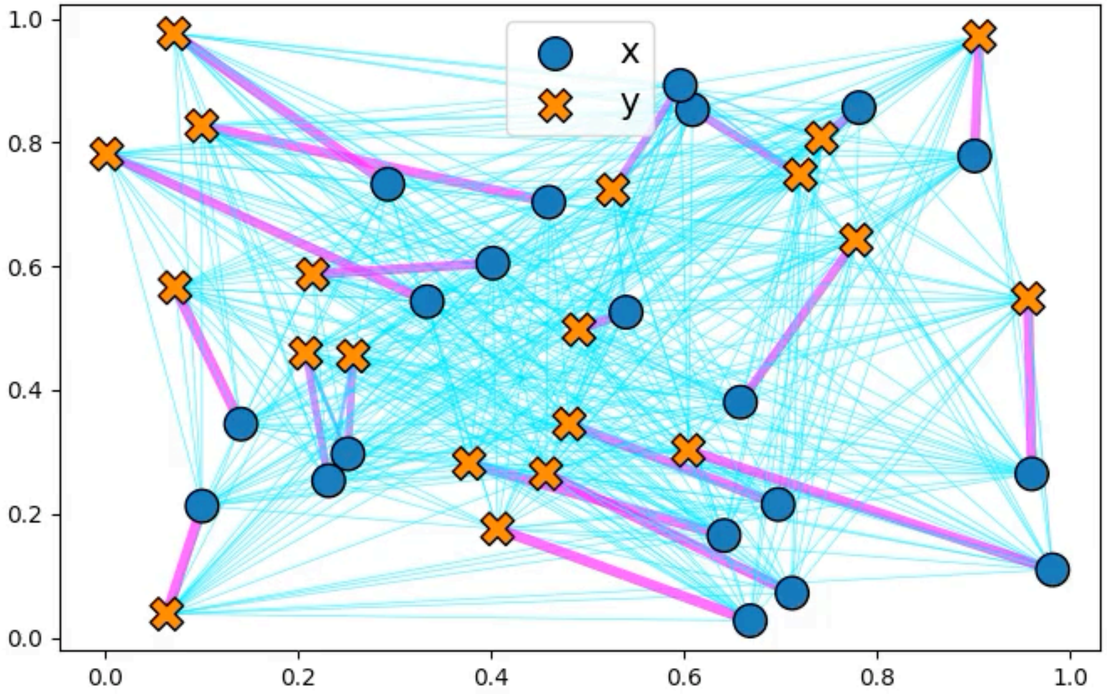
Differentiating the OT Argmin Solution



$$P_\gamma \stackrel{\text{def}}{=} \underset{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})}{\operatorname{argmin}} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle - \gamma E(\boldsymbol{P})$$

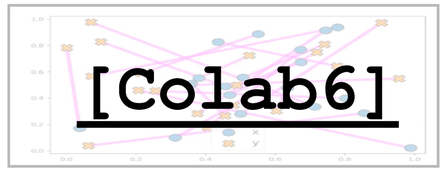
Differentiating the OT Argmin Solution





$$\left(\frac{\partial P_\gamma}{\partial \blacksquare}\right)^T$$ We can see that this seems smoother, how can we get access to that operator?

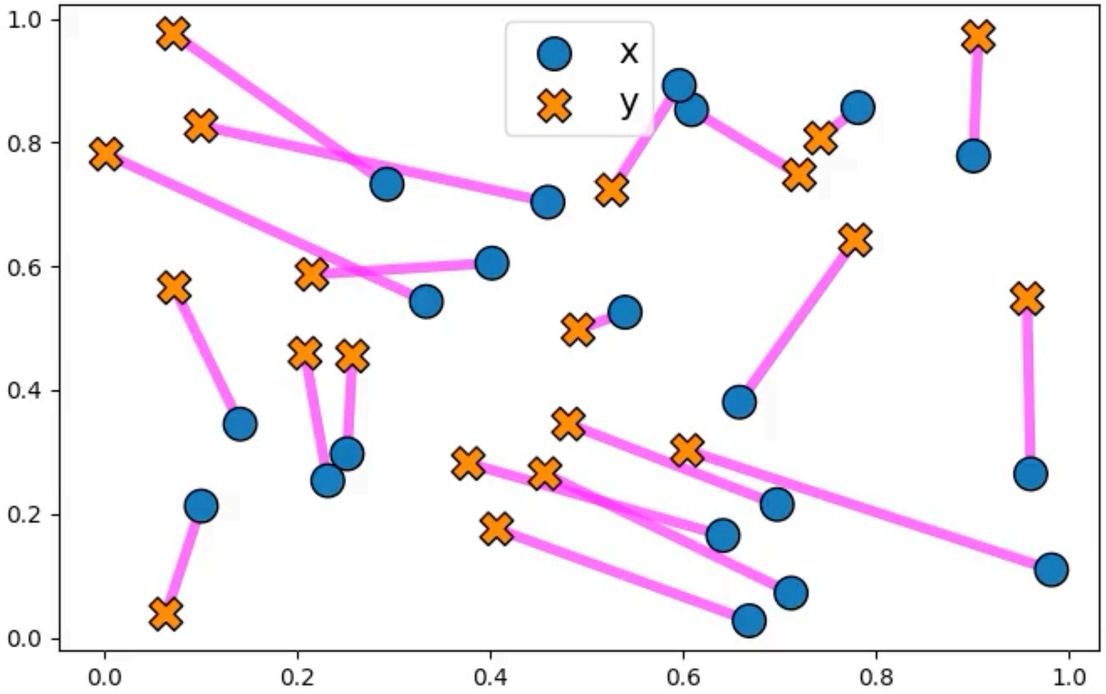
[Bonneel+16,Luise+17, Cuturi+20]

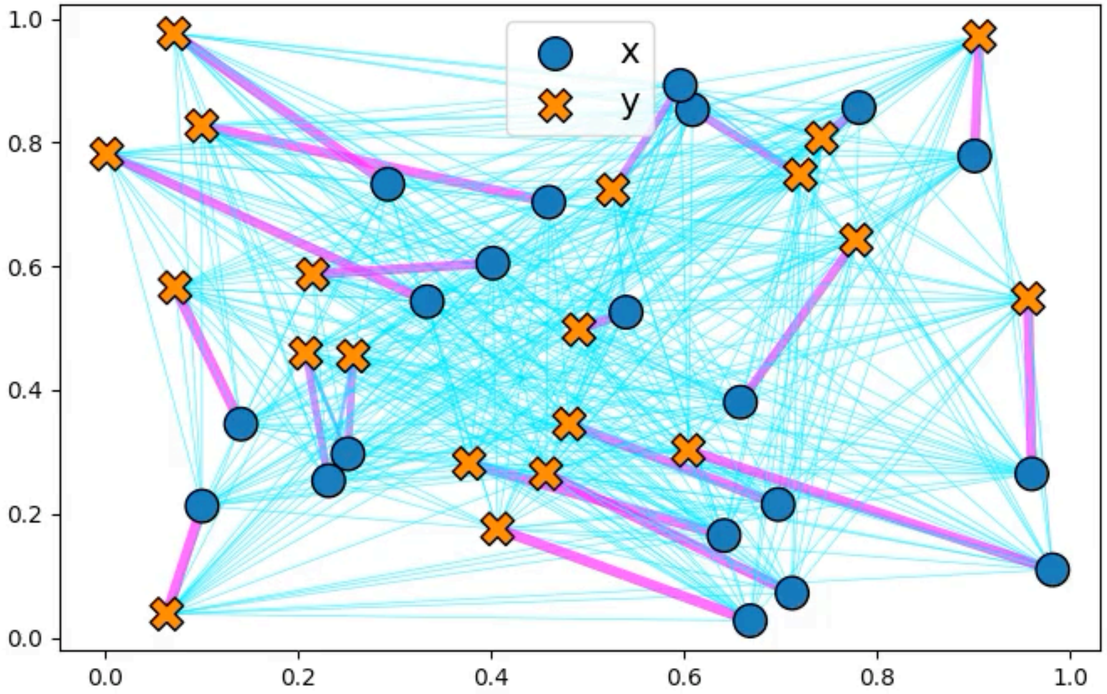


[Colab6]

Sinkhorn Hessians

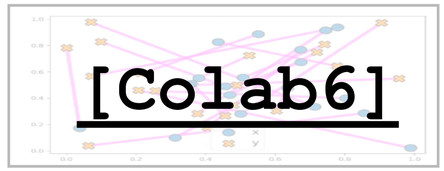
Differentiating the OT Argmin Solution





$$\left(\frac{\partial P_\gamma}{\partial \blacksquare}\right)^T$$ We can see that this seems smoother, how can we get access to that operator?

[Bonneel+16,Luise+17, Cuturi+20]



[Colab6]

Sinkhorn Hessians

Sinkhorn: A Programmer View

Def. For $L \geq 1$, define

$$\boldsymbol{P_L} \stackrel{\text{def}}{=} \mathbf{diag}(\boldsymbol{u_L}) K \mathbf{diag}(\boldsymbol{v_L}),$$

where

$$\boldsymbol{v_0} = \mathbf{1}_m; l \geq 0, \boldsymbol{u_l} \stackrel{\text{def}}{=} \boldsymbol{a} / K \boldsymbol{v_l}, \boldsymbol{v_{l+1}} \stackrel{\text{def}}{=} \boldsymbol{b} / K^T \boldsymbol{u_l}.$$

Prop. $\frac{\partial \boldsymbol{P_L}}{\partial \boldsymbol{X}}, \frac{\partial \boldsymbol{P_L}}{\partial \boldsymbol{a}}$ can be computed recursively, in $O(L)$ kernel $K \times$vector products.

Sinkhorn: A Mathematician View

$$\mathrm{OT}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma (e^{\boldsymbol{\alpha}/\gamma})^T \ K (e^{\boldsymbol{\beta}/\gamma})$$

REGULARIZED DISCRETE DUAL

An unconstrained convex optimization problem!

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma (e^{\boldsymbol{\alpha}/\gamma})^T \ K e^{\boldsymbol{\beta}/\gamma}$$

Gradient w.r.t. each variable is

$$\nabla_{\boldsymbol{\alpha}} \mathcal{E} = \boldsymbol{a} - e^{\boldsymbol{\alpha}/\gamma} \odot \ K e^{\boldsymbol{\beta}/\gamma}$$

$$\nabla_{\boldsymbol{\beta}} \mathcal{E} = \boldsymbol{b} - e^{\boldsymbol{\beta}/\gamma} \odot \ K^T e^{\boldsymbol{\alpha}/\gamma}$$

Sinkhorn: A Mathematician View

$$\mathrm{OT}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma (e^{\boldsymbol{\alpha}/\gamma})^T \; K (e^{\boldsymbol{\beta}/\gamma})$$

REGULARIZED DISCRETE DUAL

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T \boldsymbol{a} + \boldsymbol{\beta}^T \boldsymbol{b} - \gamma (e^{\boldsymbol{\alpha}/\gamma})^T \; K e^{\boldsymbol{\beta}/\gamma}$$

$$H = \begin{bmatrix} \nabla_{\boldsymbol{\alpha}} \mathcal{E} \\ \nabla_{\boldsymbol{\beta}} \mathcal{E} \end{bmatrix} \text{ valued in } \mathbb{R}^{\boldsymbol{n}+\boldsymbol{m}} := \begin{bmatrix} e^{\frac{\boldsymbol{\alpha} \oplus \boldsymbol{\beta} - \boldsymbol{C}}{\gamma}} \mathbf{1}_m - \boldsymbol{a} \\ e^{\frac{\boldsymbol{\beta} \oplus \boldsymbol{\alpha} - \boldsymbol{C}^T}{\gamma}} \mathbf{1}_n - \boldsymbol{b} \end{bmatrix}$$

First order condition: $H(\cdots, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$

Sinkhorn: A Mathematician View

$$F : \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma \to (\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$$

$$H(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma; \boldsymbol{\alpha}, \boldsymbol{\beta}) := \begin{bmatrix} e^{\frac{\boldsymbol{\alpha} \oplus \boldsymbol{\beta} - \boldsymbol{C}}{\gamma}} \mathbf{1}_m - \boldsymbol{a} \\ e^{\frac{\boldsymbol{\beta} \oplus \boldsymbol{\alpha} - \boldsymbol{C}^T}{\gamma}} \mathbf{1}_n - \boldsymbol{b} \end{bmatrix}$$

At the optimum,

$$H(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma; F(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma)) = \mathbf{0}_{\boldsymbol{n}+\boldsymbol{m}}$$

Goal: understand variations of F w.r.t. any input.

Sinkhorn: A Mathematician View

$$F : \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma \to (\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$$

$$H(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{c}, \gamma; \boldsymbol{\alpha}, \boldsymbol{\beta}) := \begin{bmatrix} e^{\frac{\boldsymbol{\alpha} \oplus \boldsymbol{\beta} - \boldsymbol{C}}{\gamma}} \mathbf{1}_m - \boldsymbol{a} \\ e^{\frac{\boldsymbol{\beta} \oplus \boldsymbol{\alpha} - \boldsymbol{C}^T}{\gamma}} \mathbf{1}_n - \boldsymbol{b} \end{bmatrix}$$

implicit function theorem

$$J_{F,\blacksquare} = -\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))\right)^{-1} J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))$$

Using an OT Block in Reverse Mode Differentiation

$J_{F,\blacksquare}^T\,\mathbf{z}$ a.k.a. transpose of $(\mathbf{z}^T J_{F,\blacksquare})$ vjp

See M. Blondel's tutorial

Using an OT Block in Reverse Mode Differentiation

$J_{F,\blacksquare}^T \, \mathbf{z}$ a.k.a. transpose of $(\mathbf{z}^T J_{F,\blacksquare})$ vjp

See M. Blondel's tutorial

$$J_{F,\blacksquare} = -\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right)^{-1} J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))$$

Using an OT Block in Reverse Mode Differentiation

$$J_{F,\blacksquare}^T \, \mathbf{z}$$

a.k.a. transpose of $(\mathbf{z}^T J_{F,\blacksquare})$ vjp

See M. Blondel's tutorial

$$J_{F,\blacksquare} = -\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))\right)^{-1} J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))$$

$$-J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^T \left(\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))\right)^T\right)^{-1} \mathbf{z}$$

Using an OT Block in Reverse Mode Differentiation

$$J_{F,\blacksquare}^T \, \mathbf{z}$$

a.k.a. transpose of $(\mathbf{z}^T J_{F,\blacksquare})$ vjp

See M. Blondel's tutorial

$$J_{F,\blacksquare} = -\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))\right)^{-1} J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))$$

$$-J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^T \left(\underbrace{\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))\right)^T}_{\texttt{jax.vjp}}\right)^{-1} \mathbf{z}$$

Using an OT Block in Reverse Mode Differentiation

$$J_{F,\blacksquare}^T \, \mathbf{z}$$

a.k.a. transpose of $(\mathbf{z}^T J_{F,\blacksquare})$ vjp

See M. Blondel's tutorial

$$J_{F,\blacksquare} = -\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))\right)^{-1} J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))$$

$$-J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^T \left(\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))^T\right)^{-1} \mathbf{z}\right)$$

`jax.vjp`

`jax.linalg.sparse.cg`

Using an OT Block in Reverse Mode Differentiation

$J_{F,\blacksquare}^T \mathbf{z}$ a.k.a. transpose of $(\mathbf{z}^T J_{F,\blacksquare})$ vjp

See M. Blondel's tutorial

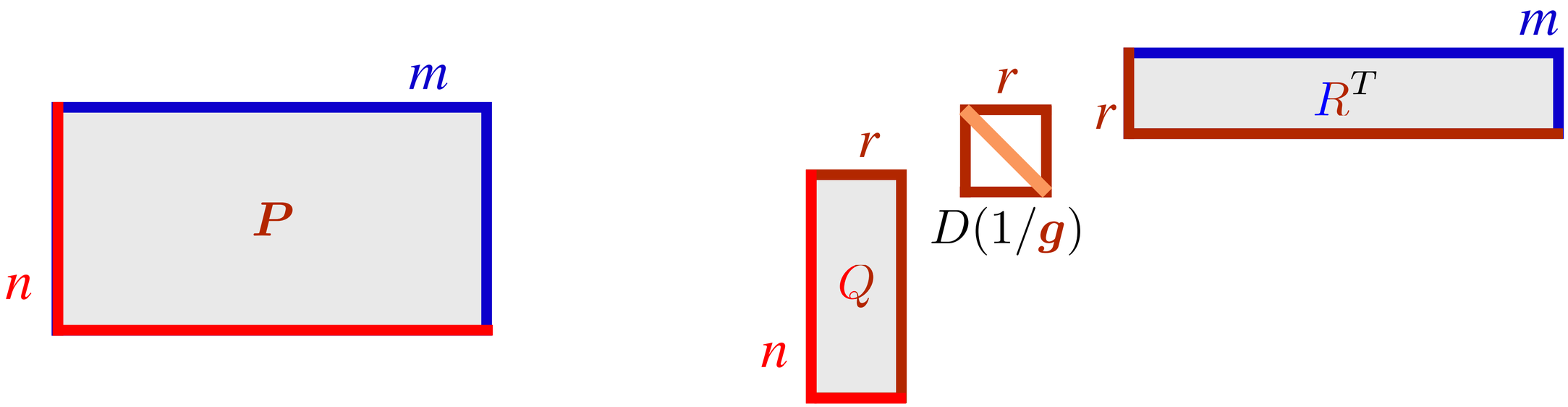
$$J_{F,\blacksquare} = -\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))\right)^{-1} J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))$$

$$-J_{H,\blacksquare}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)^T \left(\left(J_{H,(\boldsymbol{\alpha},\boldsymbol{\beta})}(\blacksquare, (\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*))^T\right)^{-1} \mathbf{z}$$

`jax.vjp` `jax.vjp` `jax.linalg.sparse.cg`

Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$



$$Q \in U(\boldsymbol{a}, \boldsymbol{g}),\ R \in U(\boldsymbol{b}, \boldsymbol{g}), \boldsymbol{g} \in \Sigma_r \Rightarrow Q D(1/\boldsymbol{g})\ R^T \in U(\boldsymbol{a}, \boldsymbol{b})$$

Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\substack{Q \in U(\boldsymbol{a}, \boldsymbol{g}), \\ R \in U(\boldsymbol{b}, \boldsymbol{g}) \\ \boldsymbol{g} \in \Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}} \rangle.$$

[Forrow+19, Scetbon+21]

Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle \longrightarrow$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q, M_{\boldsymbol{XY}} R D(1/\boldsymbol{g})\rangle$$

$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle R, M_{\boldsymbol{XY}}^T Q D(1/\boldsymbol{g})\rangle$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle D(1/\boldsymbol{g}),\, Q^T M_{\boldsymbol{XY}} R\rangle$$

Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle \longrightarrow$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q, \underbrace{M_{\boldsymbol{XY}} R D(1/\boldsymbol{g})}_{C_Q}\rangle$$

$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle R, \underbrace{M_{\boldsymbol{XY}}^T Q D(1/\boldsymbol{g})}_{C_R}\rangle$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle D(1/\boldsymbol{g}), \underbrace{Q^T M_{\boldsymbol{XY}} R}_{c_{\boldsymbol{g}}}\rangle$$

Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle \longrightarrow$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle$$

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q, \underbrace{M_{\boldsymbol{XY}} R D(1/\boldsymbol{g})}_{C_Q}\rangle \quad -\gamma E(Q,\boldsymbol{g},R)$$

$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle R, \underbrace{M_{\boldsymbol{XY}}^T Q D(1/\boldsymbol{g})}_{C_R}\rangle \quad -\gamma E(Q,\boldsymbol{g},R)$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle D(1/\boldsymbol{g}), \underbrace{Q^T M_{\boldsymbol{XY}} R}_{c_{\boldsymbol{g}}}\rangle \quad -\gamma E(Q,\boldsymbol{g},R)$$

Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]

$O(\boldsymbol{n}+\boldsymbol{m})$ if $M_{\boldsymbol{XY}} = \boldsymbol{A}\boldsymbol{B}^T$

$(\boldsymbol{n},\boldsymbol{d})\cdot(\boldsymbol{d},\boldsymbol{m})\cdot(\boldsymbol{m},\boldsymbol{r})\cdot\boldsymbol{r} = O\left((\boldsymbol{n}+\boldsymbol{m})(\boldsymbol{d}+\boldsymbol{r})\right)$

$$\min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle \quad \min_{Q\in U(\boldsymbol{a},\boldsymbol{g})} \langle Q, \underbrace{\boldsymbol{A}\boldsymbol{B}^T R D(1/\boldsymbol{g})}_{C_Q}\rangle \quad -\gamma E(Q,\boldsymbol{g},R)$$

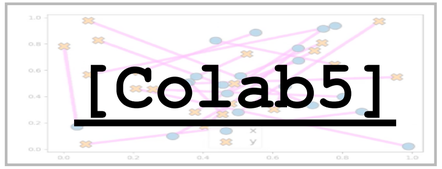
$$\min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle \longrightarrow \min_{R\in U(\boldsymbol{b},\boldsymbol{g})} \langle R, \underbrace{M_{\boldsymbol{XY}}^T Q D(1/\boldsymbol{g})}_{C_R}\rangle \quad -\gamma E(Q,\boldsymbol{g},R)$$

$$\min_{\boldsymbol{g}\in\Sigma_r} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle \quad \min_{\boldsymbol{g}\in\Sigma_r} \langle D(1/\boldsymbol{g}), \underbrace{Q^T M_{\boldsymbol{XY}} R}_{c_{\boldsymbol{g}}}\rangle \quad -\gamma E(Q,\boldsymbol{g},R)$$

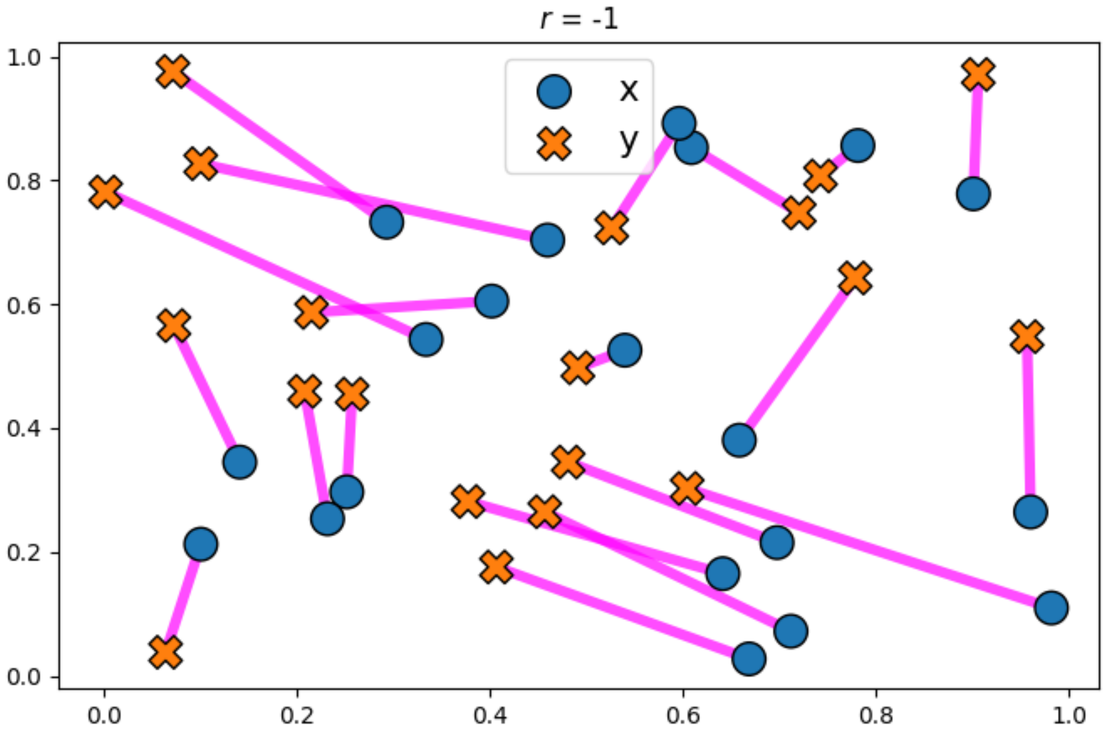
Low-Rank Constrained Discrete Optimal Transport Problem

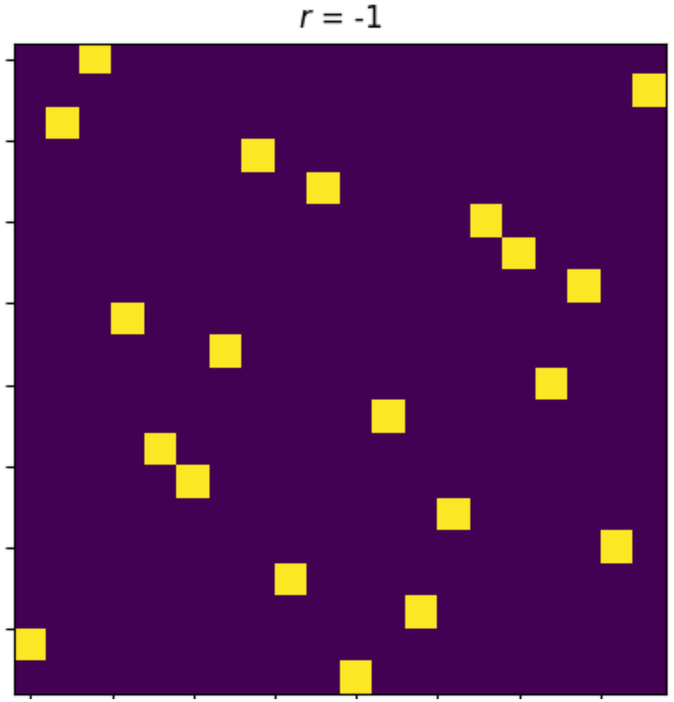
$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]

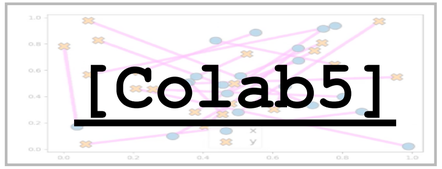




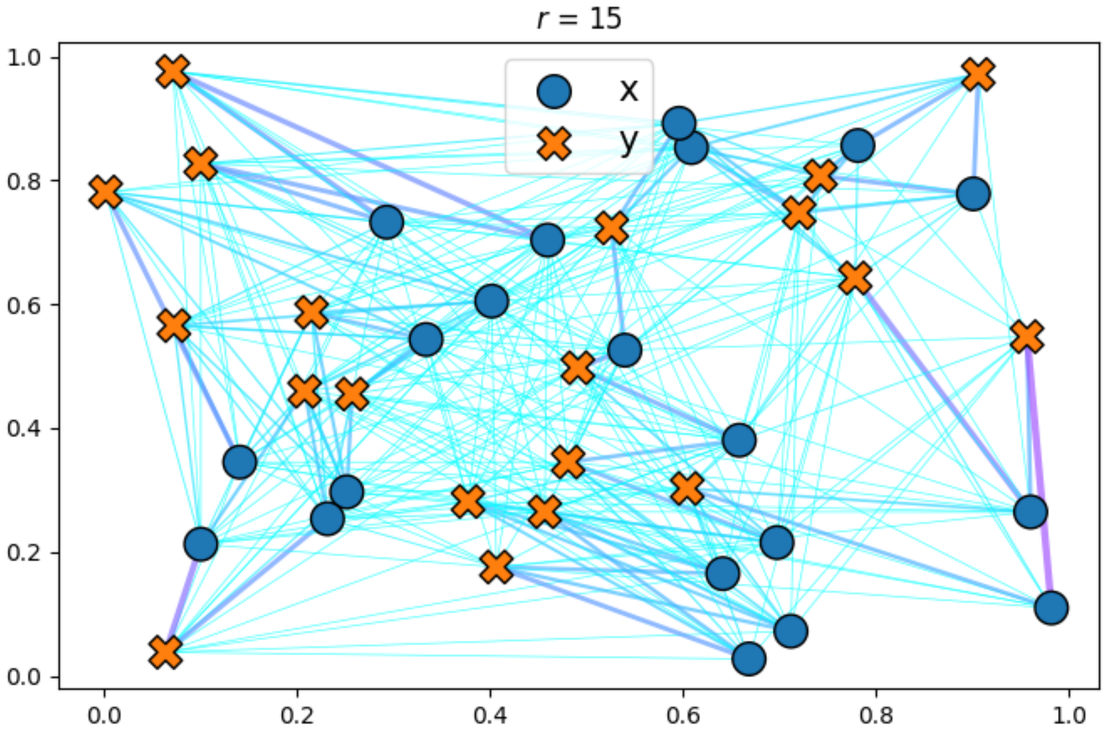
Low-Rank Constrained Discrete Optimal Transport Problem

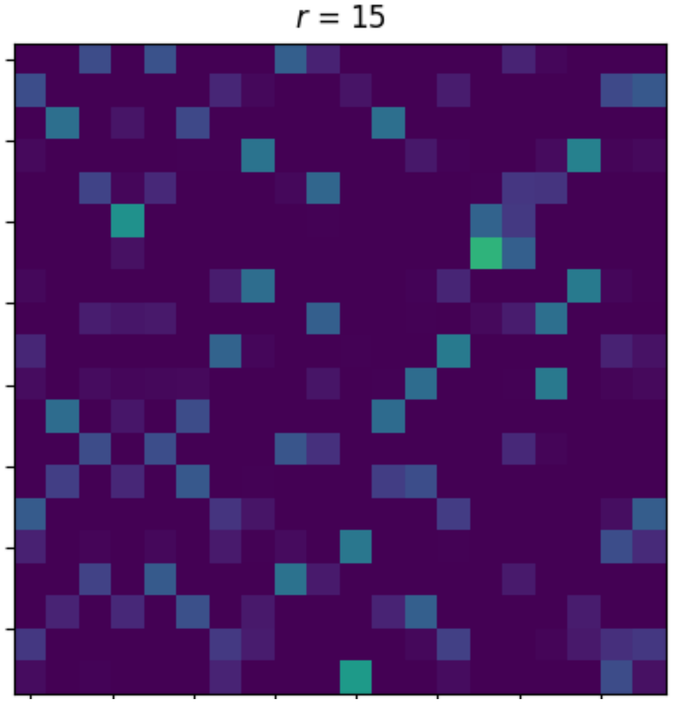
$$\min_{\substack{Q \in U(\boldsymbol{a}, \boldsymbol{g}), \\ R \in U(\boldsymbol{b}, \boldsymbol{g}) \\ \boldsymbol{g} \in \Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}} \rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]

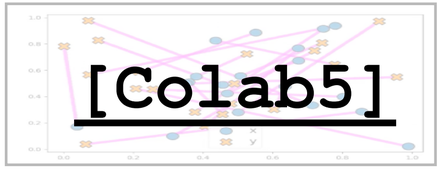




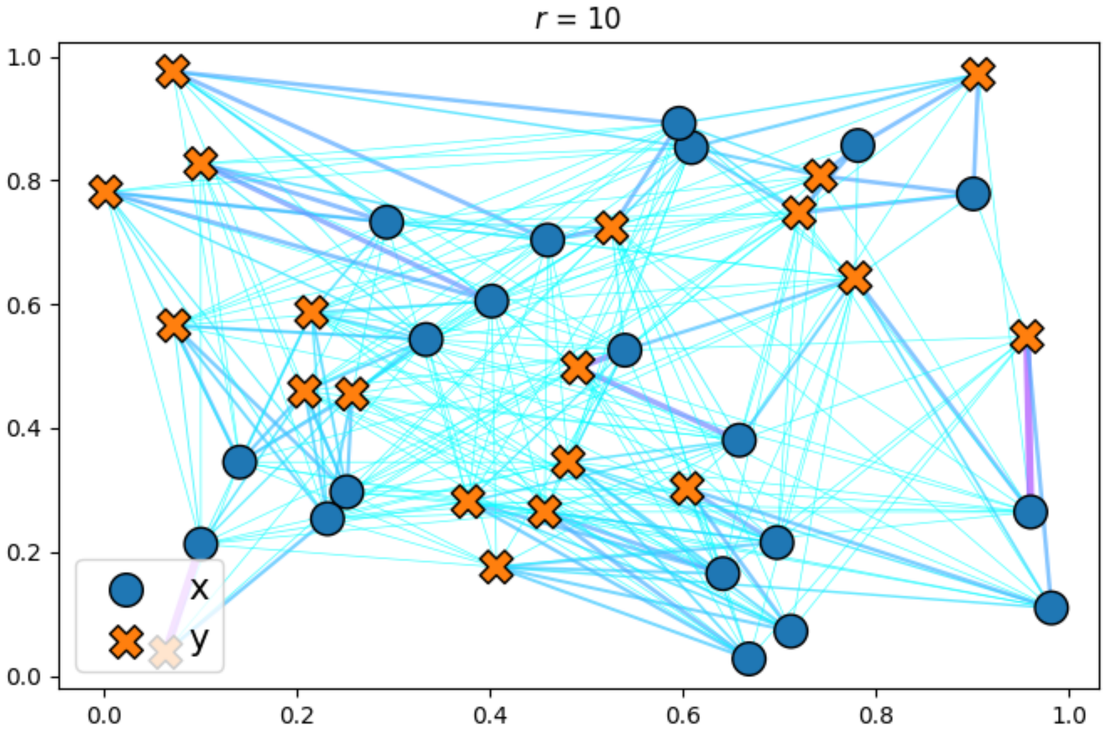
Low-Rank Constrained Discrete Optimal Transport Problem

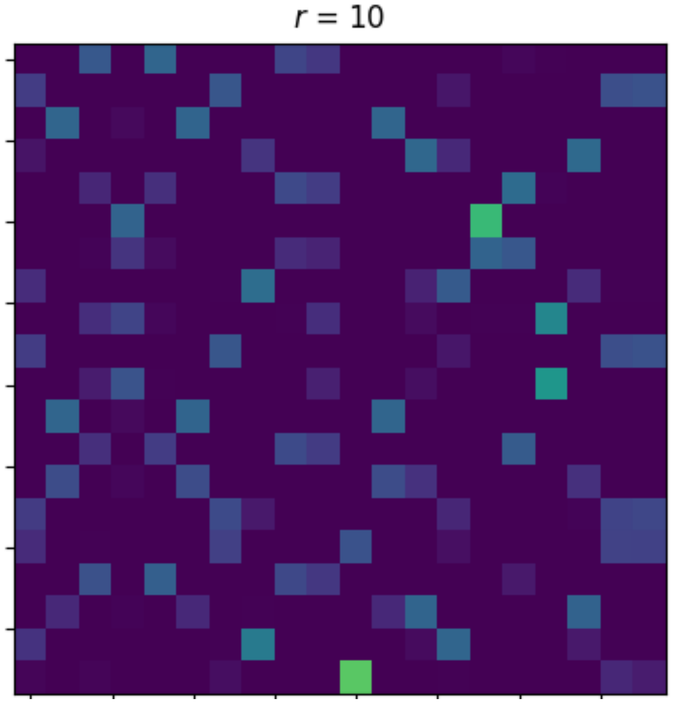
$$\min_{\substack{Q \in U(\boldsymbol{a}, \boldsymbol{g}), \\ R \in U(\boldsymbol{b}, \boldsymbol{g}) \\ \boldsymbol{g} \in \Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}} \rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]

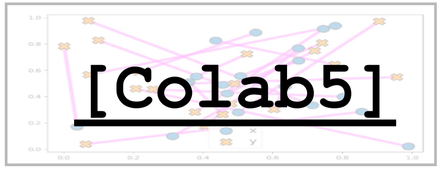




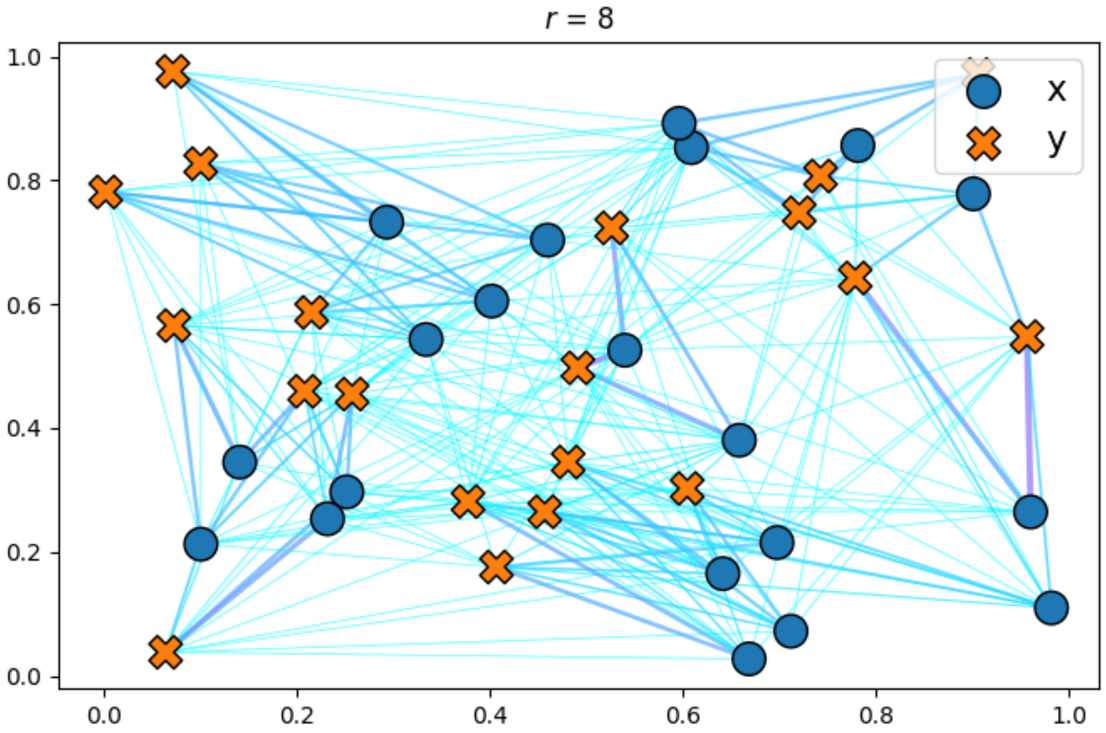
Low-Rank Constrained Discrete Optimal Transport Problem

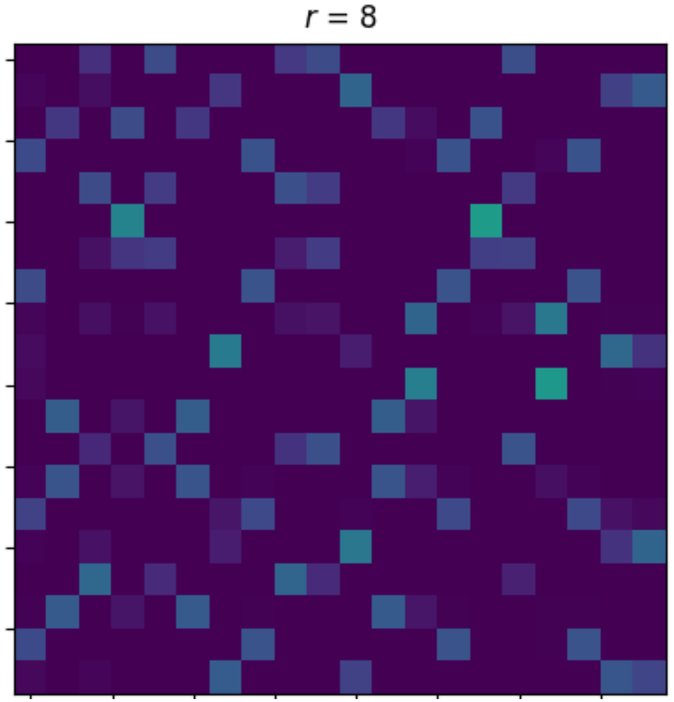
$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]

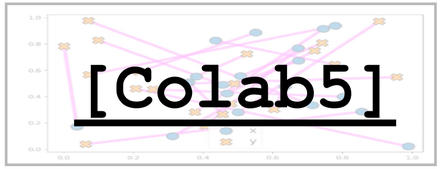




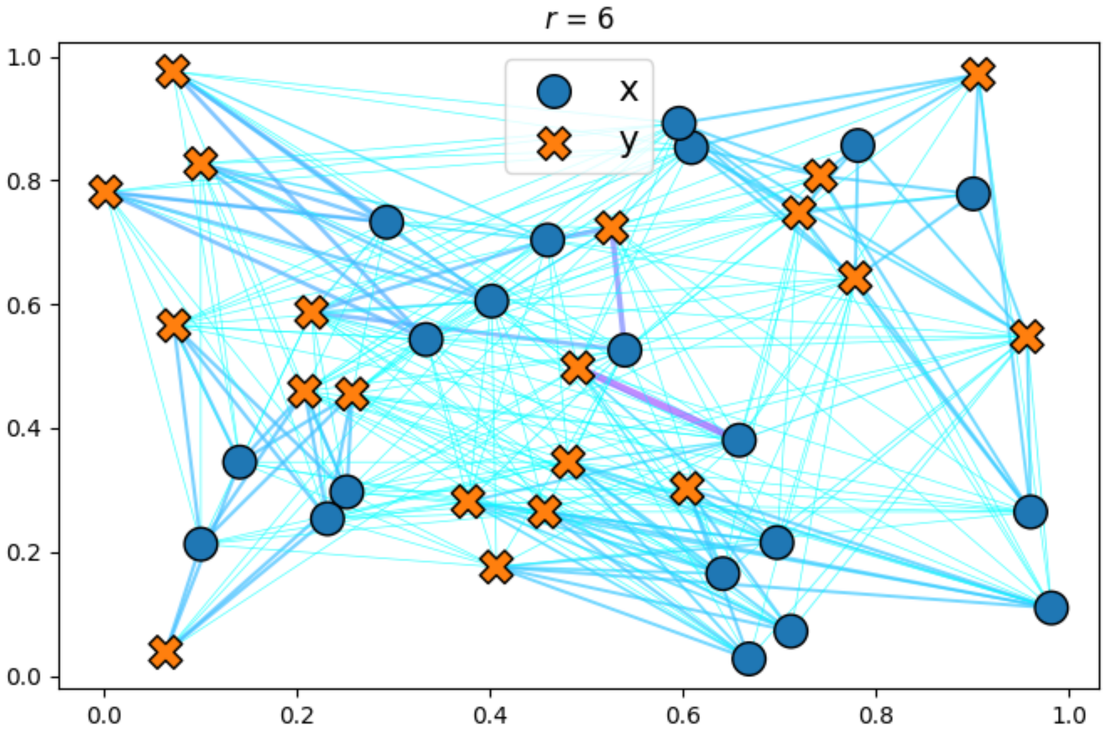
Low-Rank Constrained Discrete Optimal Transport Problem

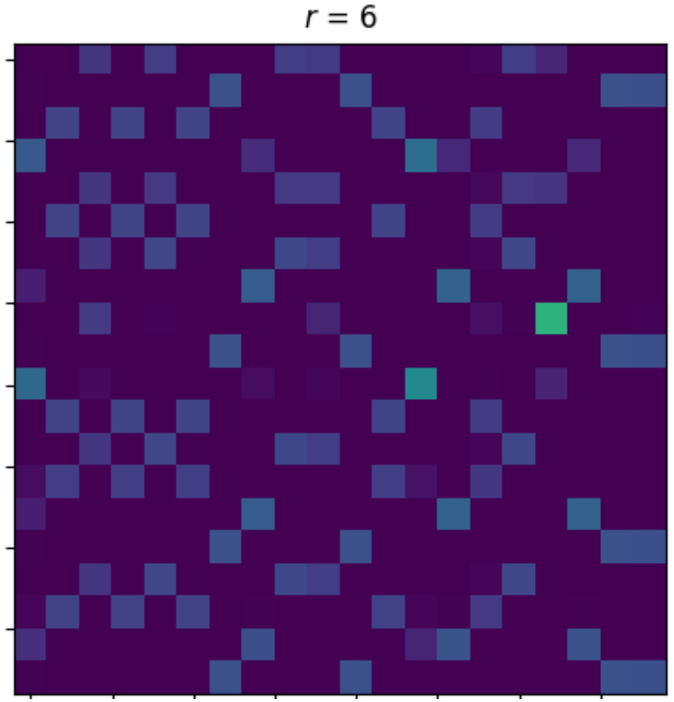
$$\min_{\substack{Q \in U(\boldsymbol{a}, \boldsymbol{g}), \\ R \in U(\boldsymbol{b}, \boldsymbol{g}) \\ \boldsymbol{g} \in \Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}} \rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]

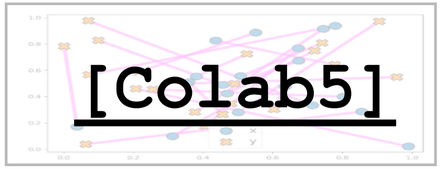




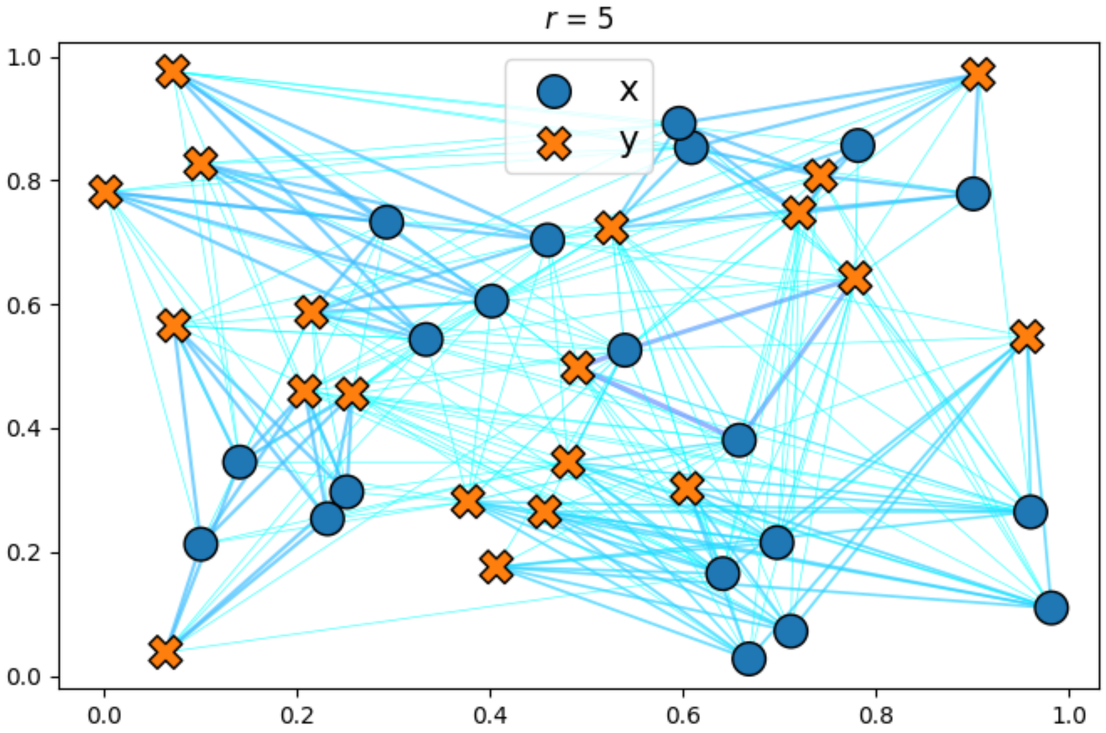
Low-Rank Constrained Discrete Optimal Transport Problem

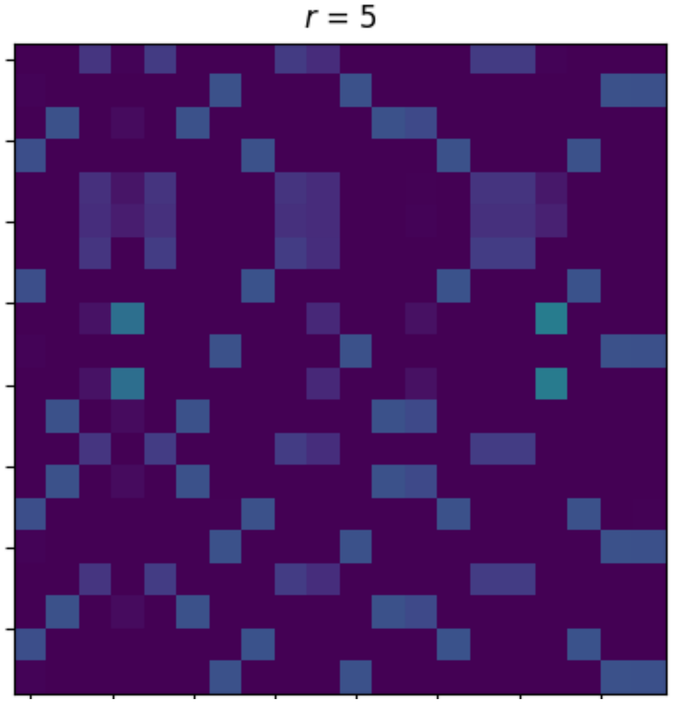
$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]

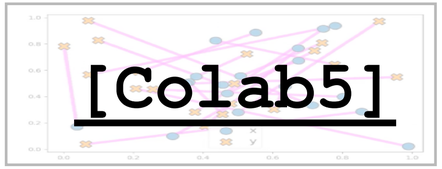




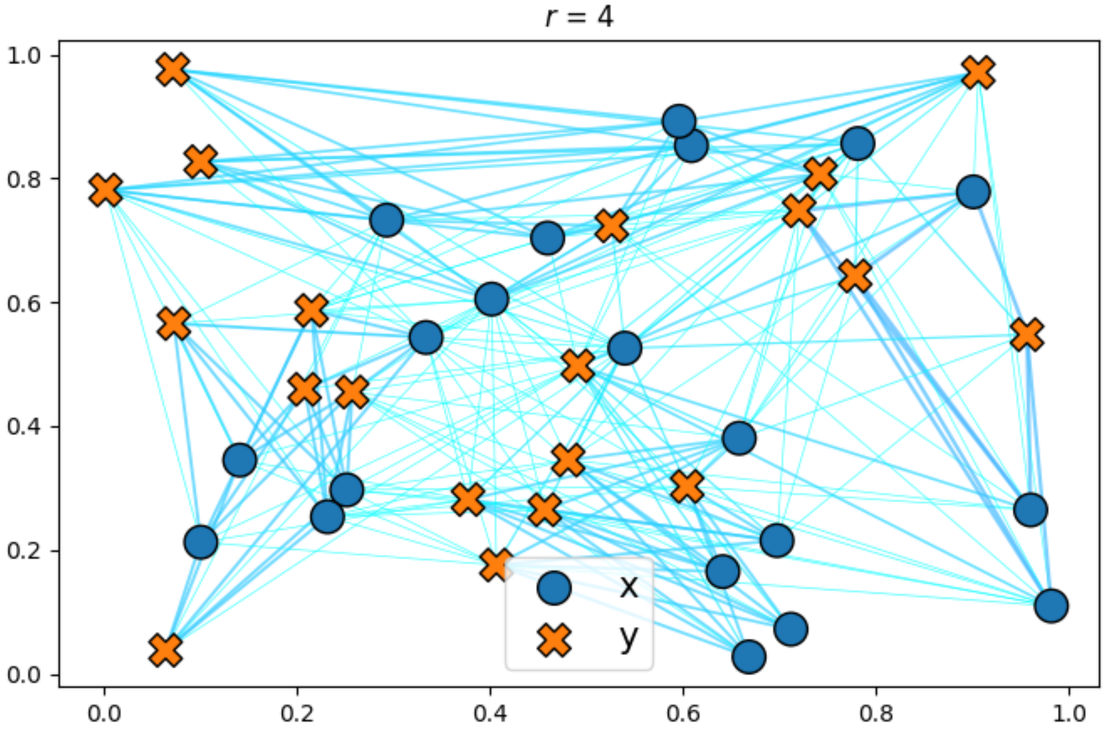
Low-Rank Constrained Discrete Optimal Transport Problem

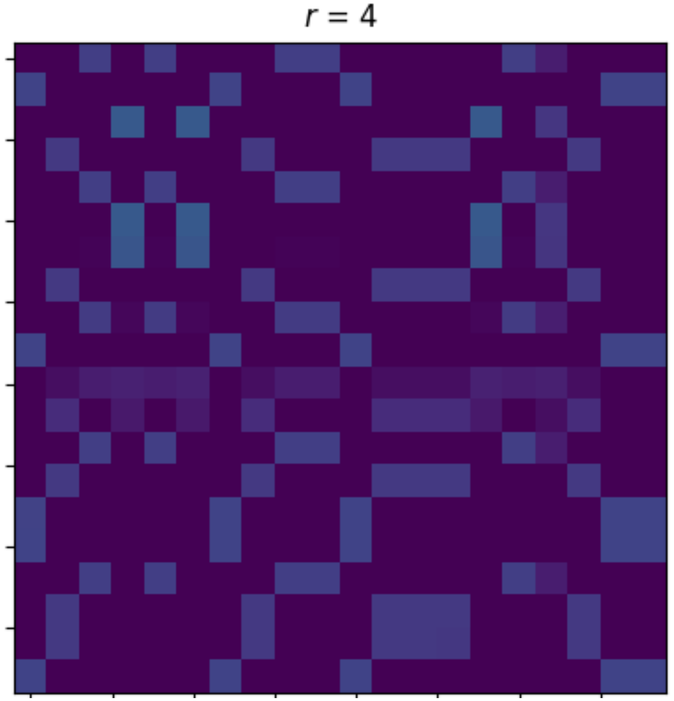
$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]

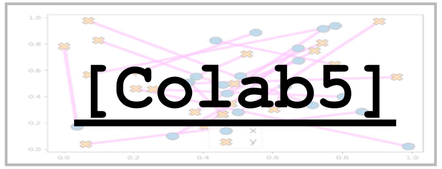




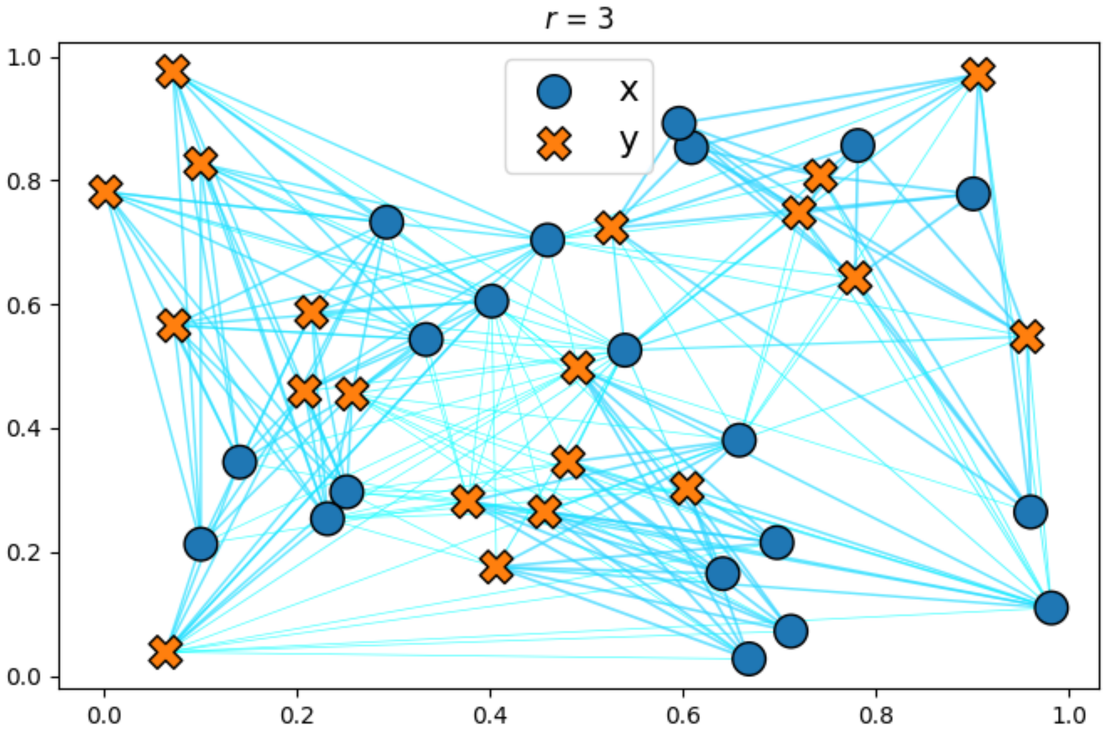
Low-Rank Constrained Discrete Optimal Transport Problem

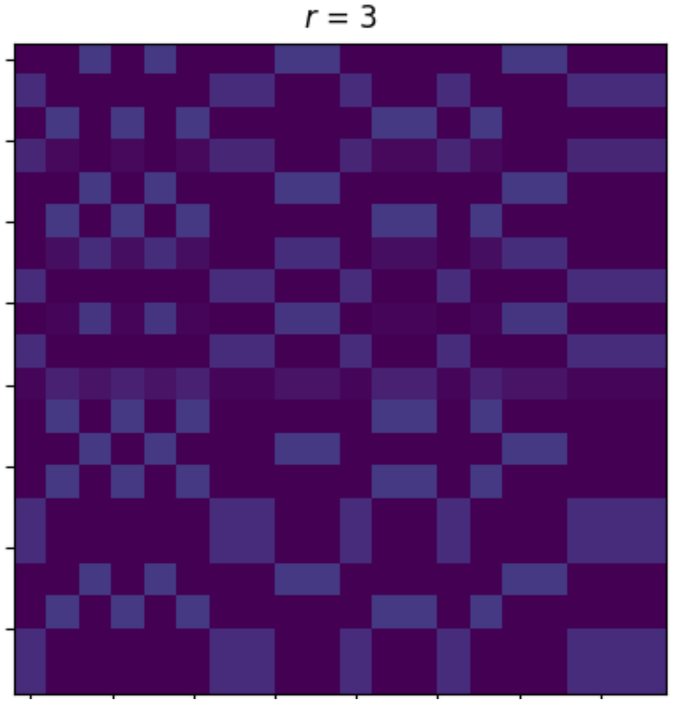
$$\min_{\substack{Q \in U(\boldsymbol{a}, \boldsymbol{g}), \\ R \in U(\boldsymbol{b}, \boldsymbol{g}) \\ \boldsymbol{g} \in \Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}} \rangle.$$

[Forrow+19, Scetbon+21]



[Colab5]



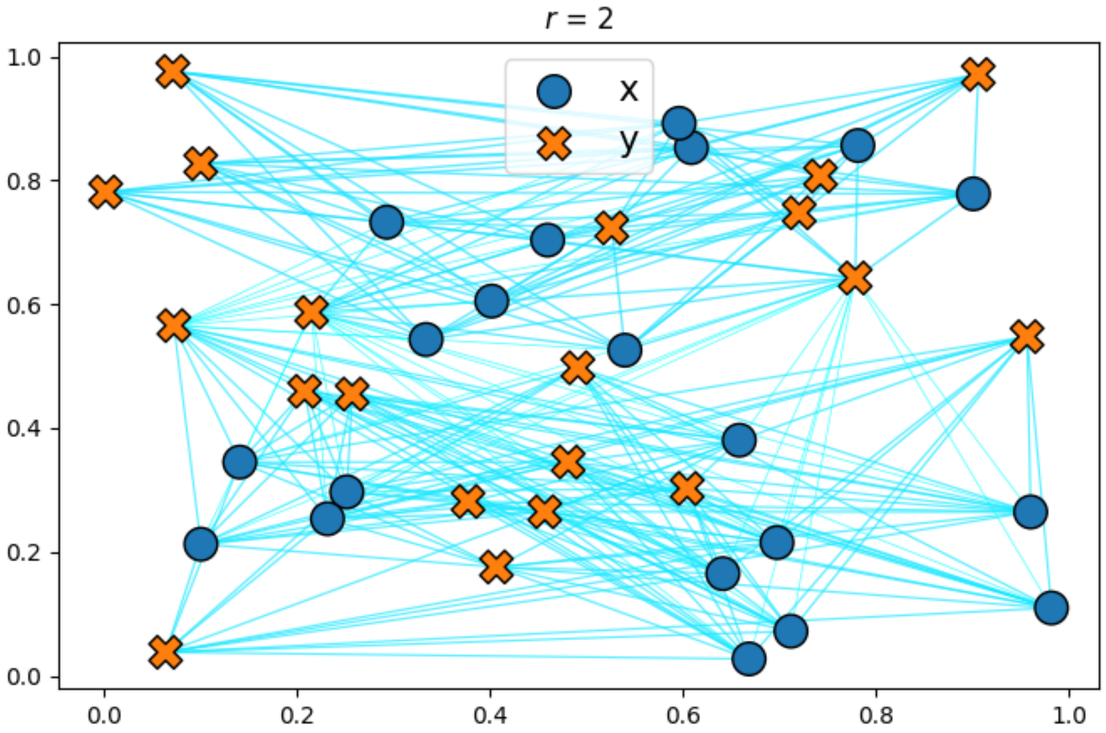


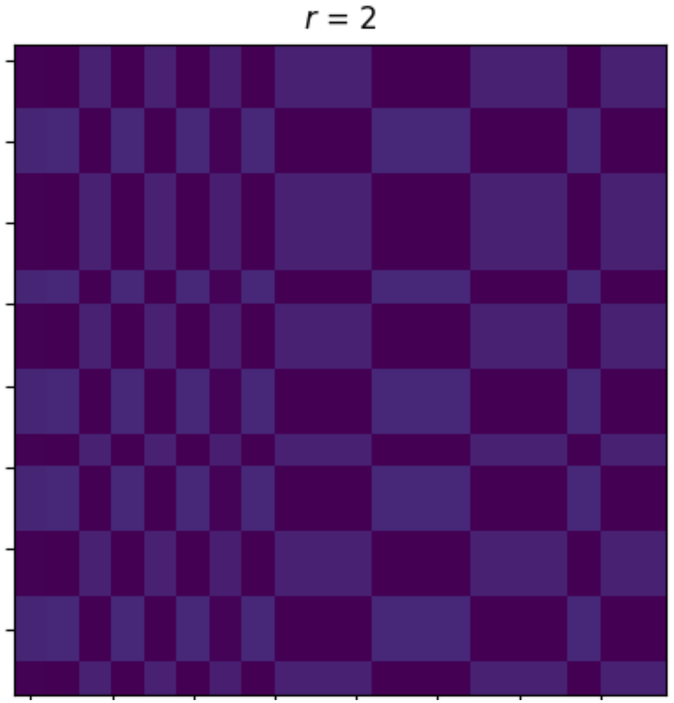
Low-Rank Constrained Discrete Optimal Transport Problem

$$\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g})\\ \boldsymbol{g}\in\Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}}\rangle.$$

[Forrow+19, Scetbon+21]

[Colab5]

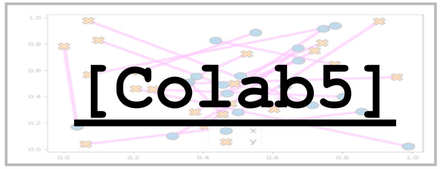




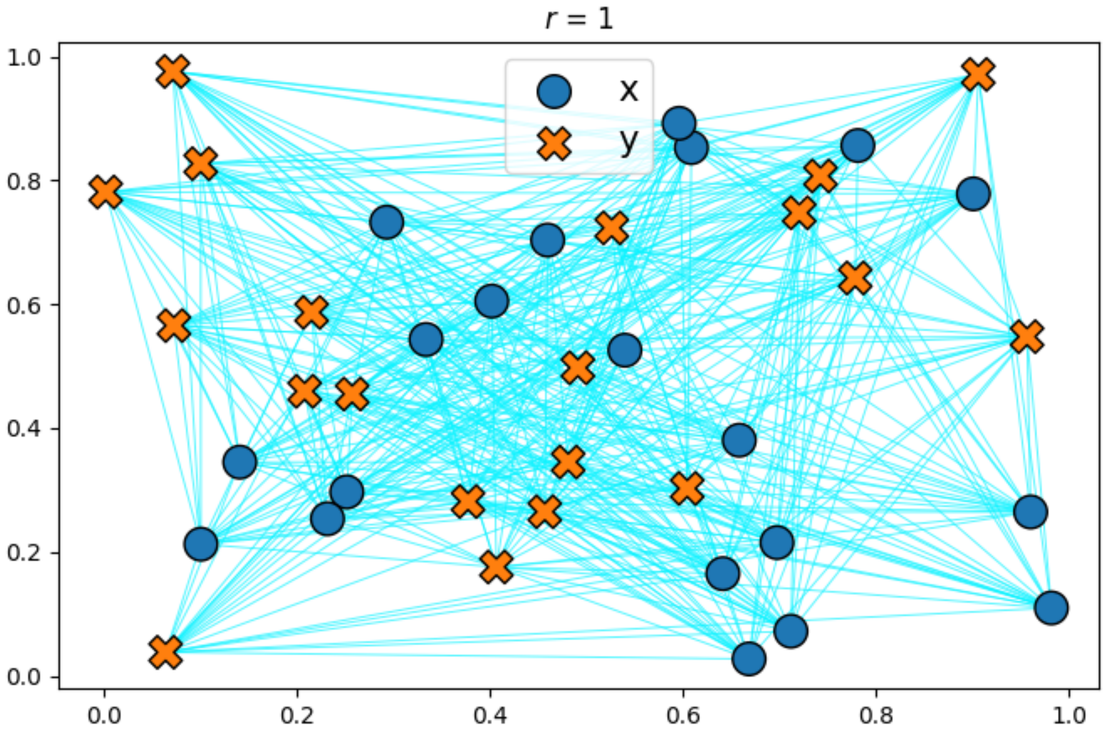
Low-Rank Constrained Discrete Optimal Transport Problem

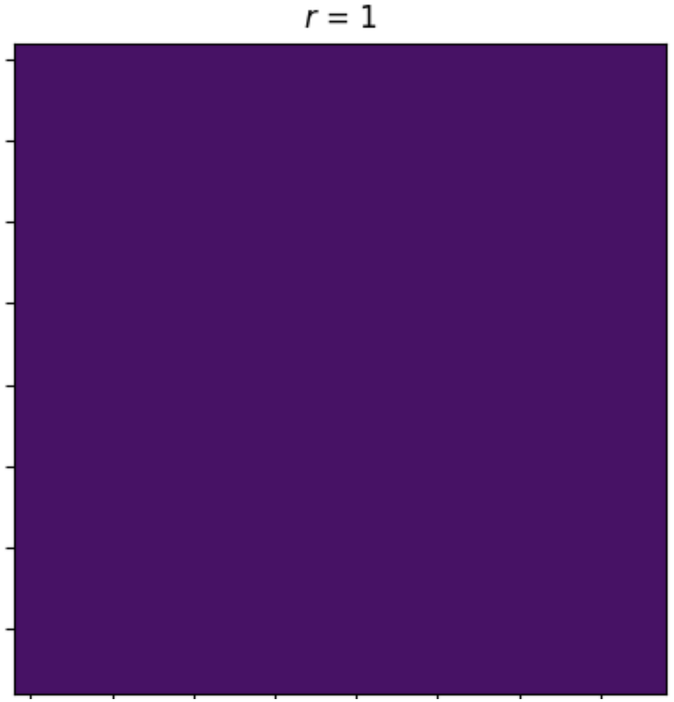
$$\min_{\substack{Q \in U(\boldsymbol{a}, \boldsymbol{g}), \\ R \in U(\boldsymbol{b}, \boldsymbol{g}) \\ \boldsymbol{g} \in \Sigma_r}} \langle Q D(1/\boldsymbol{g})\, R^T, M_{\boldsymbol{XY}} \rangle.$$

[Forrow+19, Scetbon+21]

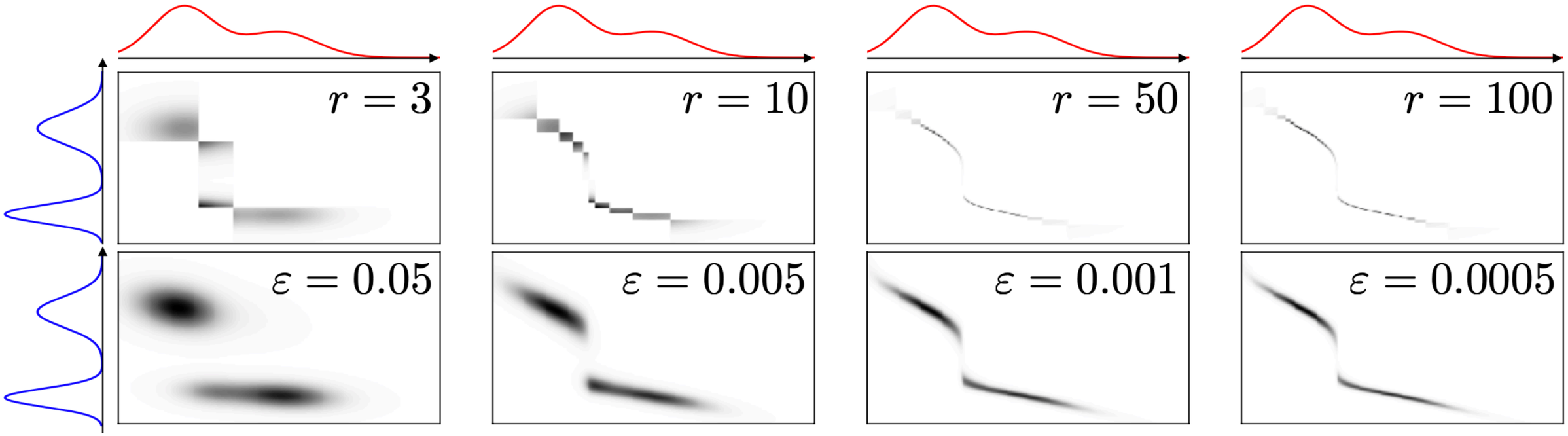


[Colab5]



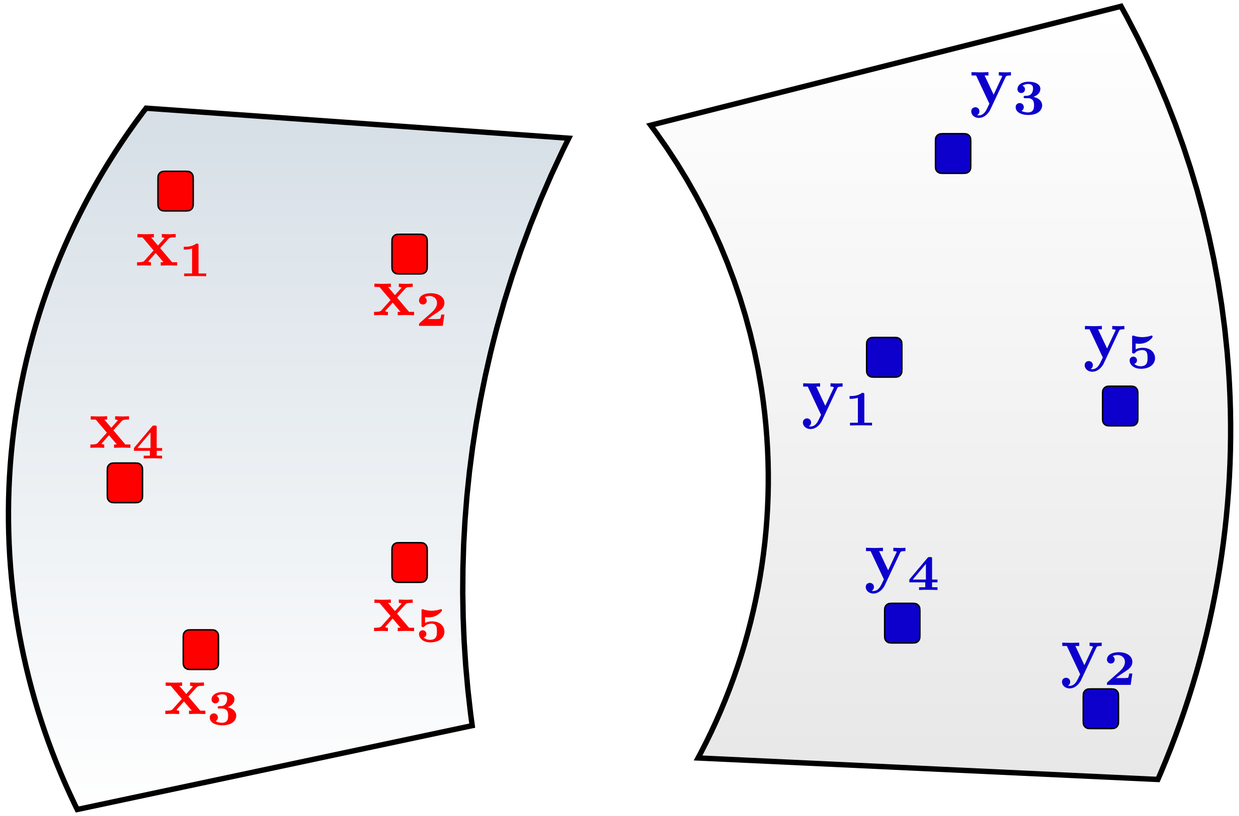


On a Simpler 1D Problem



The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$




Beckman


Koopmans

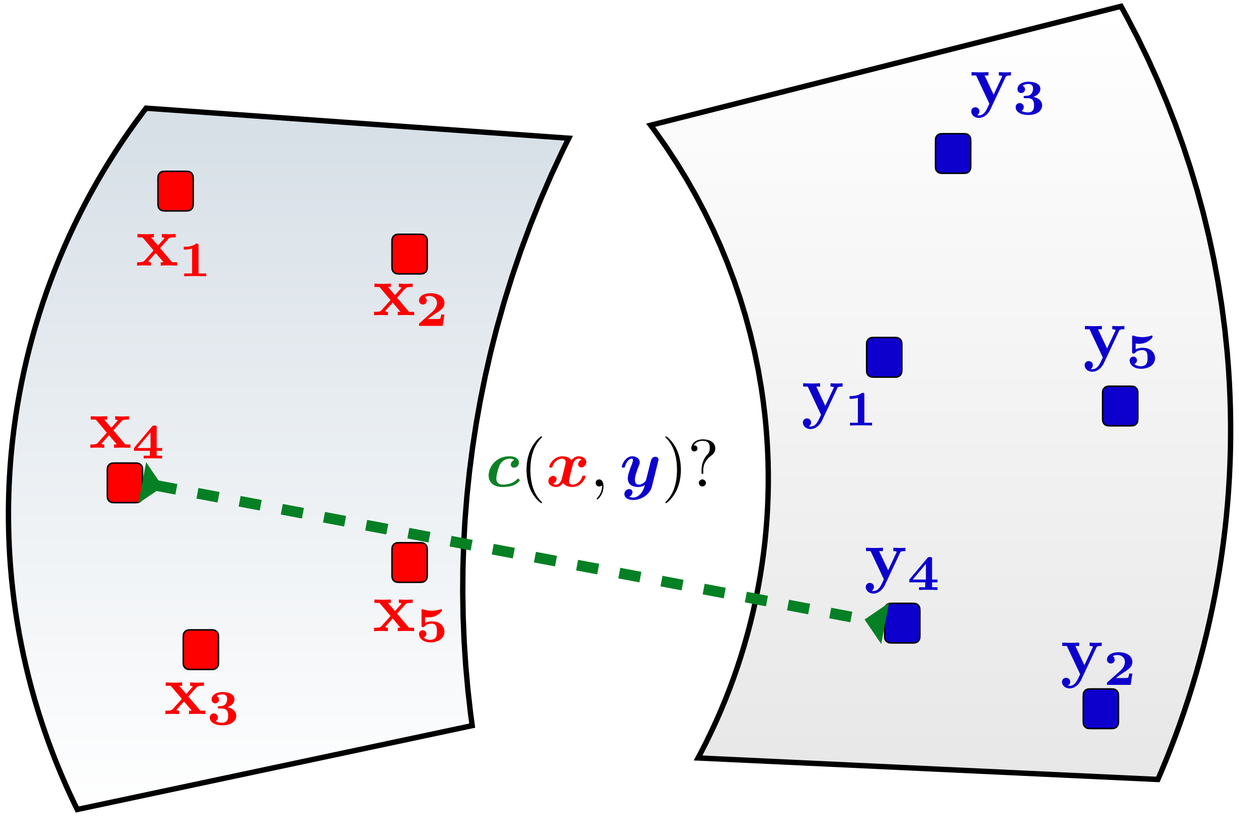

Memoli


Sturm

$$P^\star = \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\operatorname{argmin}} \quad \mathcal{E}(P)$$

The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$




Beckman


Koopmans

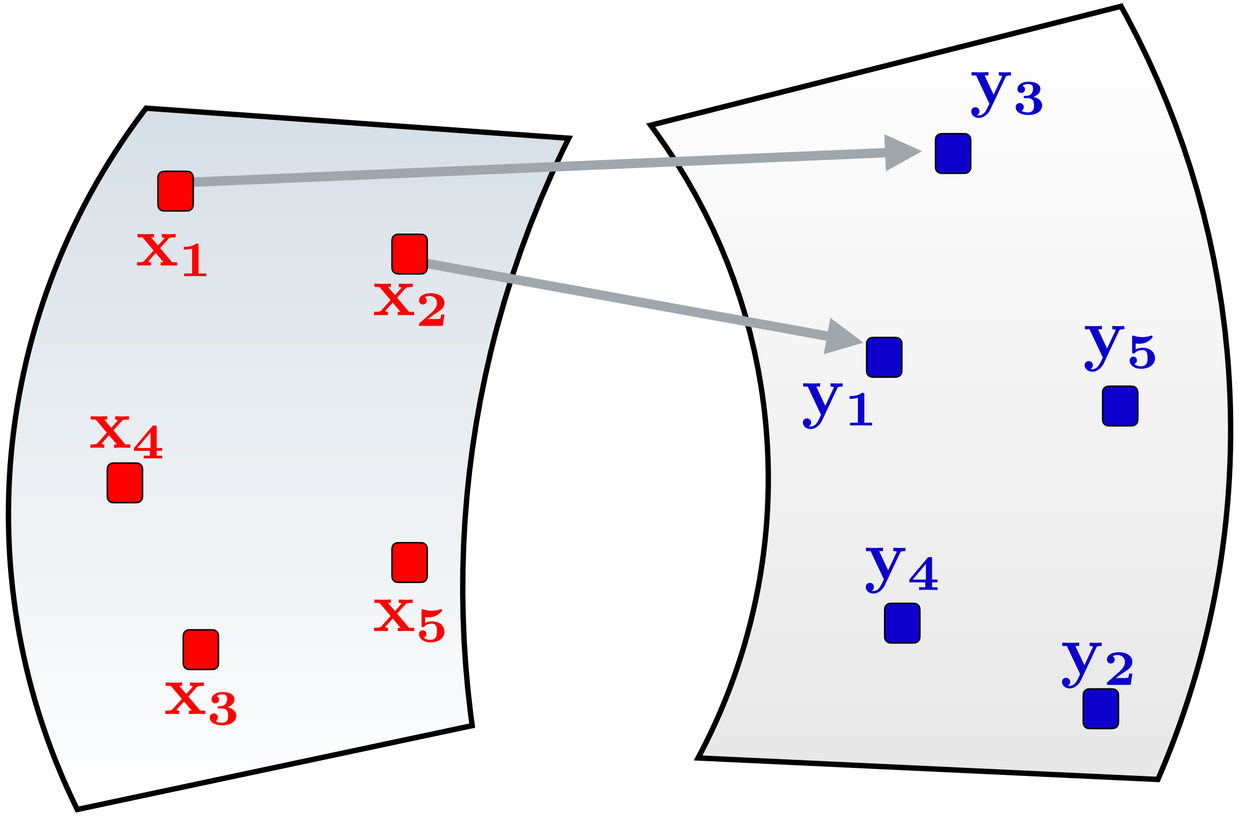

Memoli


Sturm

$$P^\star = \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\operatorname{argmin}} \quad \mathcal{E}(P)$$

The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$




Beckman


Koopmans

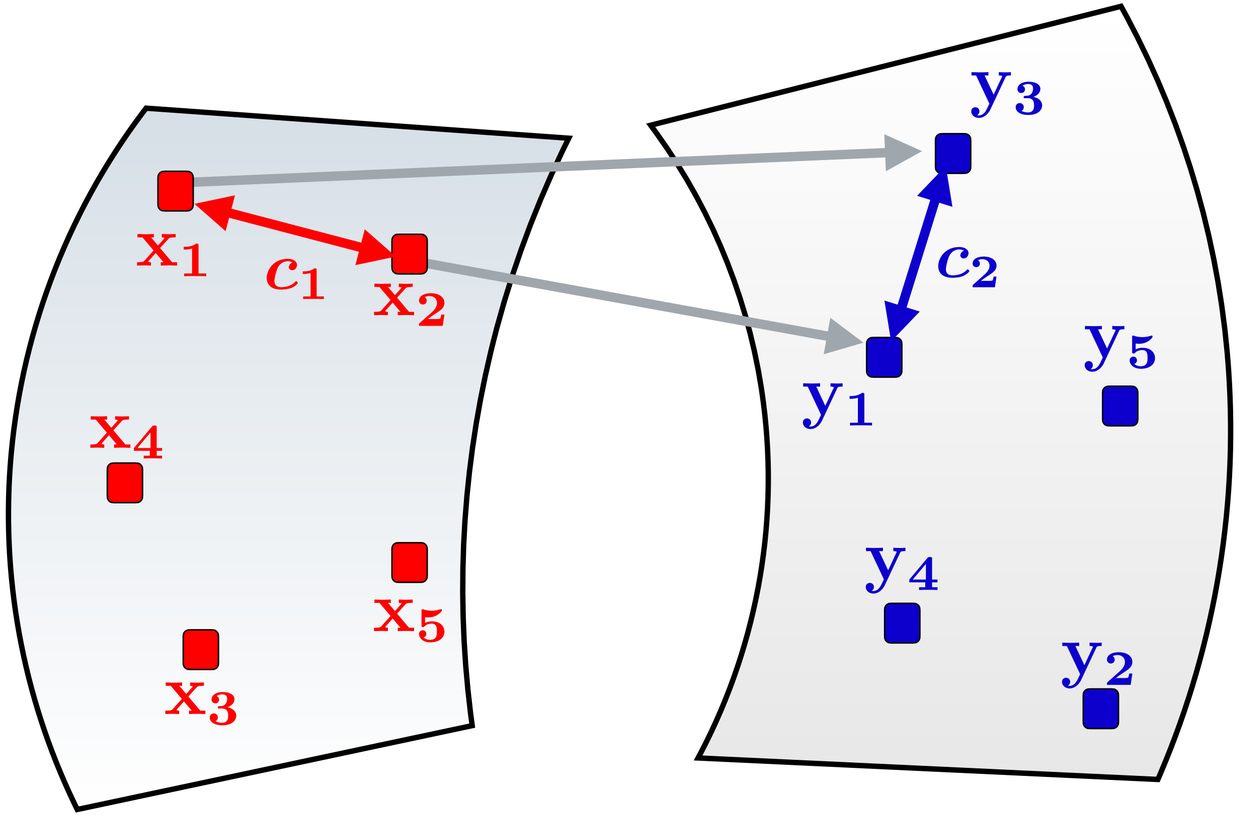

Memoli


Sturm

$$P^\star = \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\operatorname{argmin}} \quad \mathcal{E}(P)$$

The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$




Beckman


Koopmans


Memoli

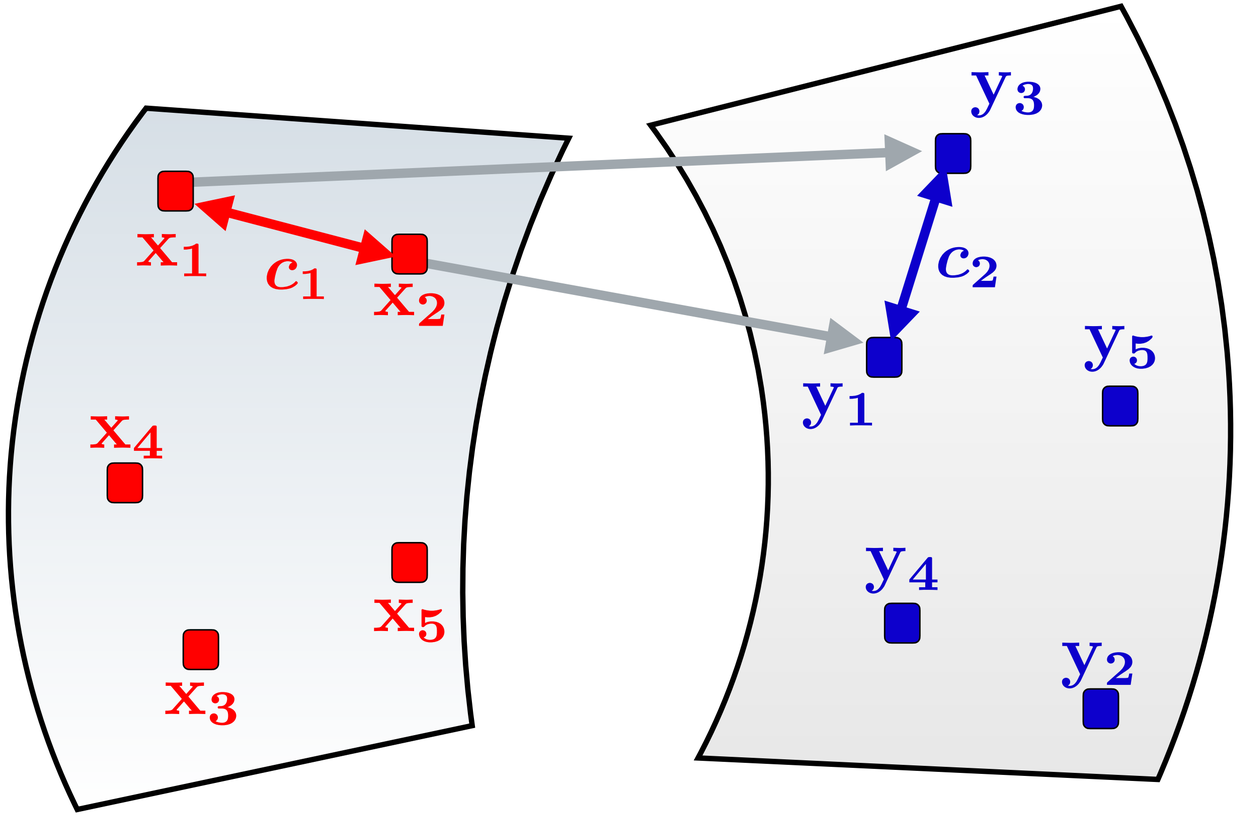

Sturm

$$P^\star = \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\operatorname{argmin}} \quad \mathcal{E}(P)$$

The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$

$$\nabla \mathcal{E}(P) \propto -C_1 P C_2$$




Beckman


Koopmans


Memoli

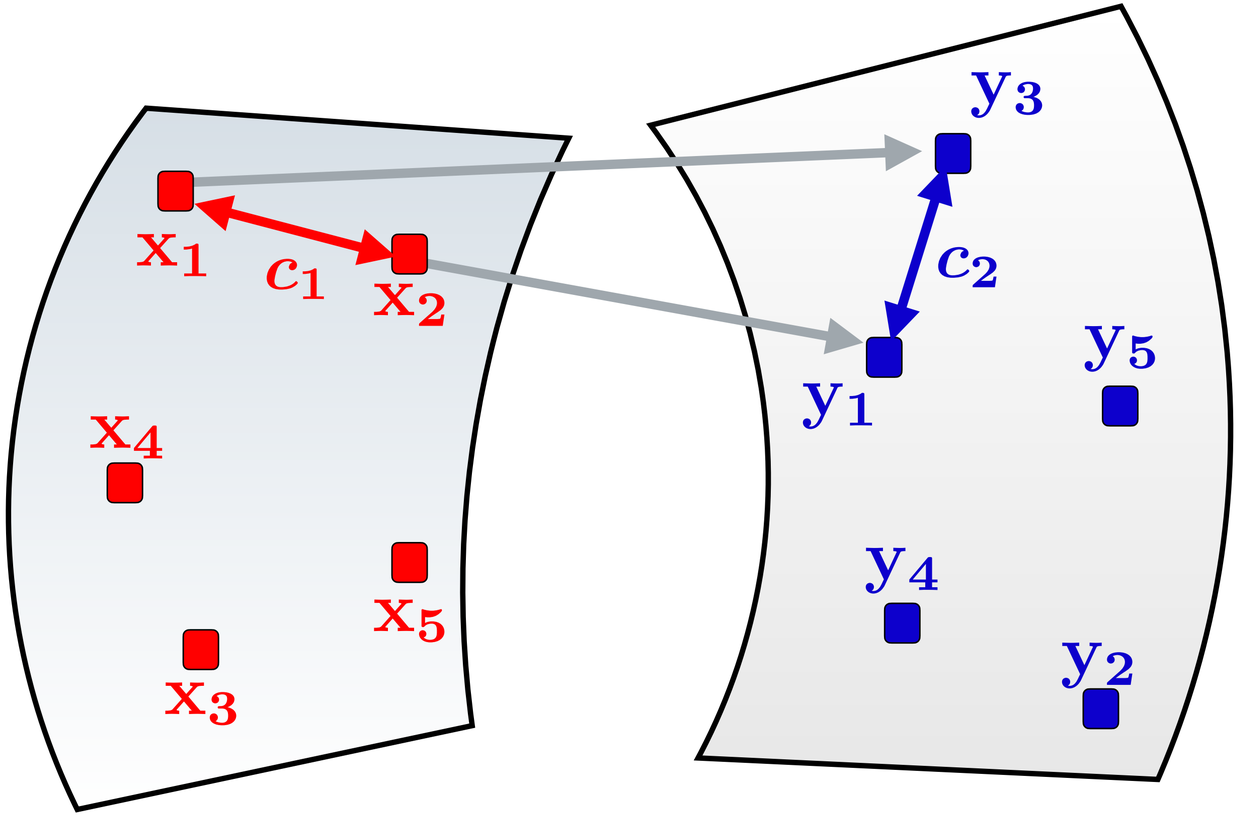

Sturm

$$P_{t+1} \leftarrow \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\operatorname{argmin}} \langle \nabla \mathcal{E}(P_t), P \rangle$$

The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$

$$\nabla \mathcal{E}(P) \propto -C_1 P C_2$$




Beckman


Koopmans


Memoli


Sturm

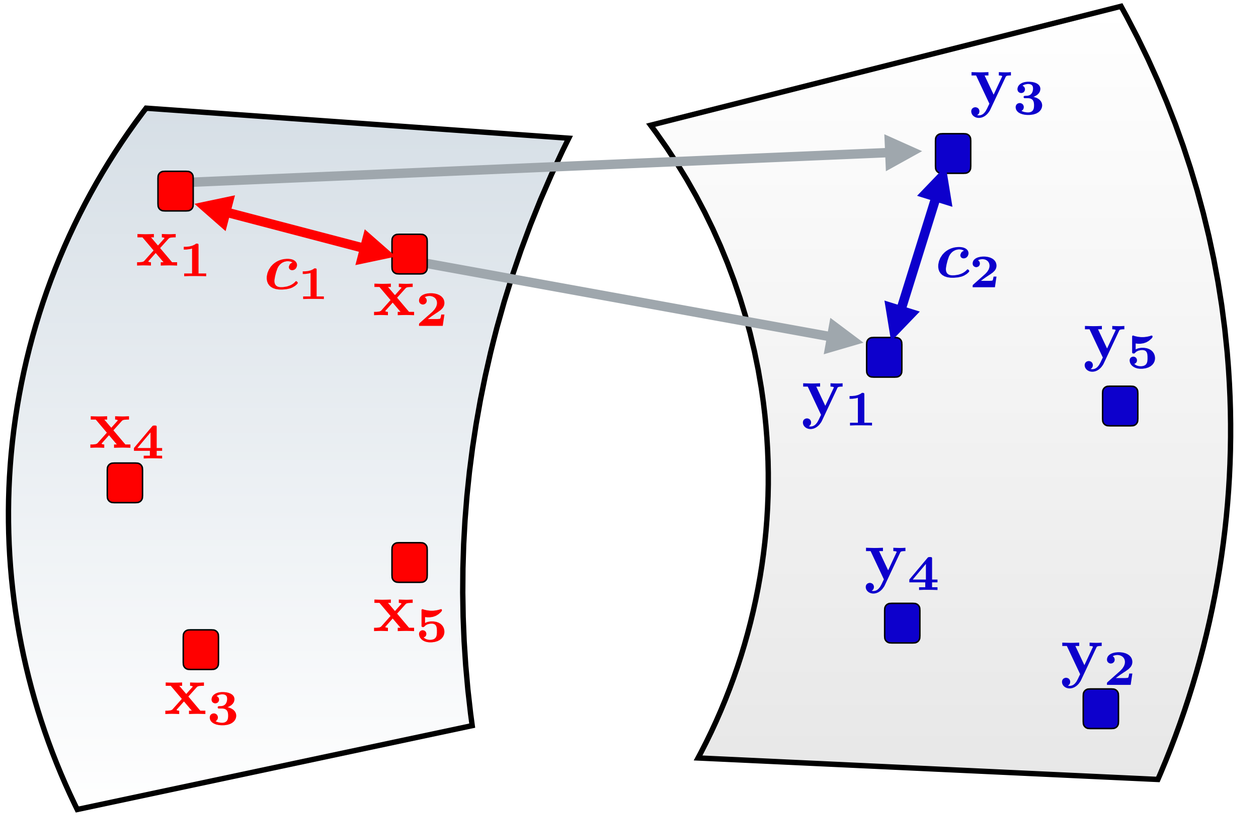
$$P_{t+1} \leftarrow \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\operatorname{argmin}} \langle \nabla \mathcal{E}(P_t), P \rangle$$

Any linear solver can be used

The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$

$$\nabla \mathcal{E}(P) \propto -C_1 P C_2$$




Beckman


Koopmans


Memoli


Sturm

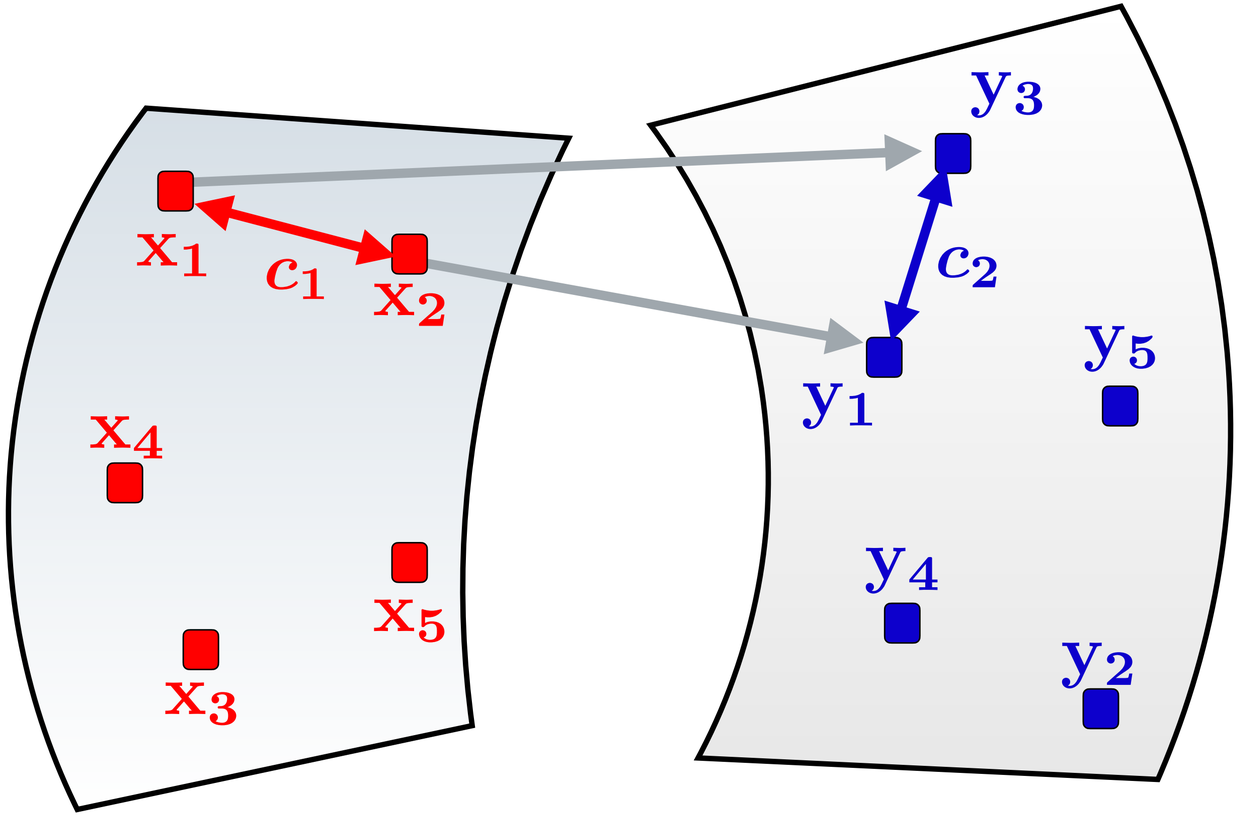
$$P_{t+1} \leftarrow \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\operatorname{argmin}} \langle \nabla \mathcal{E}(P_t), P \rangle - \gamma E(P)$$

Any linear solver can be used

The Quadratic (Gromov-Wasserstein) Optimal Tranport Problem

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} (c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'}))^2$$

$$\nabla \mathcal{E}(P) \propto -C_1 P C_2$$




Beckman


Koopmans


Memoli


Sturm

$$P_{t+1} \leftarrow \underset{\substack{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b \\ \mathrm{rank}(P) = r}}{\mathrm{argmin}} \langle \nabla \mathcal{E}(P_t), P \rangle$$

Any linear solver can be used

GW

Low-Rank and GW are Compatible, yielding Linear Time Algorithm

$$\mathcal{E}(\boldsymbol{P}) := \sum_{i,i',j,j'} \boldsymbol{P}_{ij}\boldsymbol{P}_{i'j'}(\boldsymbol{c_1}(\mathbf{x}_i,\mathbf{x}_{i'}) - \boldsymbol{c_2}(\mathbf{y}_j,\mathbf{y}_{j'}))^2$$

[Scetbon+22]

$$\nabla\mathcal{E}(\boldsymbol{P}) \propto -\boldsymbol{C_1}\boldsymbol{P}\boldsymbol{C_2} \qquad \boldsymbol{C_1} = \boldsymbol{A}\boldsymbol{A}^T \qquad \boldsymbol{C_2} = \boldsymbol{B}\boldsymbol{B}^T$$

$$\boldsymbol{P_{t+1}} \leftarrow \underset{\substack{\boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b} \\ \mathrm{rank}(\boldsymbol{P})=r}}{\mathrm{argmin}} \langle\nabla\mathcal{E}(\boldsymbol{P_t}), \boldsymbol{P}\rangle$$

Low-Rank and GW are Compatible, yielding Linear Time Algorithm

$$\mathcal{E}(\boldsymbol{P}) := \sum_{i,i',j,j'} \boldsymbol{P}_{ij}\boldsymbol{P}_{i'j'}(\boldsymbol{c}_1(\mathbf{x}_i,\mathbf{x}_{i'}) - \boldsymbol{c}_2(\mathbf{y}_j,\mathbf{y}_{j'}))^2$$

[Scetbon+22]

$$\nabla\mathcal{E}(\boldsymbol{P}) \propto -\boldsymbol{C}_1\boldsymbol{P}\boldsymbol{C}_2 \qquad \boldsymbol{C}_1 = \boldsymbol{A}\boldsymbol{A}^T \qquad \boldsymbol{C}_2 = \boldsymbol{B}\boldsymbol{B}^T$$

$$\boldsymbol{P}_{t+1} \leftarrow \underset{\substack{\boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}\\ \mathrm{rank}(\boldsymbol{P})=r}}{\mathrm{argmin}} \langle\nabla\mathcal{E}(\boldsymbol{P}_t), \boldsymbol{P}\rangle$$

$$Q_{t+1}, \boldsymbol{g}_{t+1}, R_{t+1} \leftarrow \arg\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g}),\\ \boldsymbol{g}\in\Sigma_r}} -\langle\boldsymbol{C}_1 \overbrace{Q_t D(\boldsymbol{g}_t)^{-1} R_t^T}^{\boldsymbol{P}_t}\boldsymbol{C}_2, \overbrace{Q D(1/\boldsymbol{g}) R^T}^{\boldsymbol{P}}\rangle.$$

Low-Rank and GW are Compatible, yielding Linear Time Algorithm

$$\mathcal{E}(\boldsymbol{P}) := \sum_{i,i',j,j'} \boldsymbol{P}_{ij}\boldsymbol{P}_{i'j'}(\boldsymbol{c}_1(\mathbf{x}_i,\mathbf{x}_{i'}) - \boldsymbol{c}_2(\mathbf{y}_j,\mathbf{y}_{j'}))^2$$

[Scetbon+22]

$$\nabla\mathcal{E}(\boldsymbol{P}) \propto -\boldsymbol{C_1}\boldsymbol{P}\boldsymbol{C_2} \qquad \boldsymbol{C_1} = \boldsymbol{A}\boldsymbol{A}^T \qquad \boldsymbol{C_2} = \boldsymbol{B}\boldsymbol{B}^T$$

$$\boldsymbol{P_{t+1}} \leftarrow \underset{\substack{\boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}\\ \mathrm{rank}(\boldsymbol{P})=r}}{\mathrm{argmin}} \langle\nabla\mathcal{E}(\boldsymbol{P_t}),\boldsymbol{P}\rangle$$

$$Q_{t+1}, \boldsymbol{g}_{t+1}, R_{t+1} \leftarrow \arg\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g}),\\ \boldsymbol{g}\in\Sigma_r}} -\langle \underbrace{\boldsymbol{C_1}\overbrace{Q_t D(\boldsymbol{g}_t)^{-1} R_t^T}^{\boldsymbol{P_t}}\boldsymbol{C_2}}_{\boldsymbol{P}}, \underbrace{Q D(1/\boldsymbol{g}) R^T}_{\boldsymbol{P}}\rangle.$$

Low-Rank and GW are Compatible, yielding Linear Time Algorithm

$$\mathcal{E}(\boldsymbol{P}) := \sum_{i,i',j,j'} \boldsymbol{P}_{ij}\boldsymbol{P}_{i'j'}(\boldsymbol{c_1}(\mathbf{x}_i,\mathbf{x}_{i'}) - \boldsymbol{c_2}(\mathbf{y}_j,\mathbf{y}_{j'}))^2$$

[Scetbon+22]

$$\nabla\mathcal{E}(\boldsymbol{P}) \propto -\boldsymbol{C_1}\boldsymbol{P}\boldsymbol{C_2} \qquad \boldsymbol{C_1} = \boldsymbol{A}\boldsymbol{A}^T \qquad \boldsymbol{C_2} = \boldsymbol{B}\boldsymbol{B}^T$$

$$\boldsymbol{P}_{t+1} \leftarrow \underset{\substack{\boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}\\ \mathrm{rank}(\boldsymbol{P})=r}}{\mathrm{argmin}} \langle\nabla\mathcal{E}(\boldsymbol{P}_t), \boldsymbol{P}\rangle$$

$$Q_{t+1}, \boldsymbol{g}_{t+1}, R_{t+1} \leftarrow \arg\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g}),\\ \boldsymbol{g}\in\Sigma_r}} -\langle \overbrace{\boldsymbol{C_1}\, Q_t D(\boldsymbol{g}_t)^{-1}\, R_t^T \boldsymbol{C_2}}^{\boldsymbol{P}_t}, \overbrace{Q D(1/\boldsymbol{g})\, R^T}^{\boldsymbol{P}}\rangle.$$

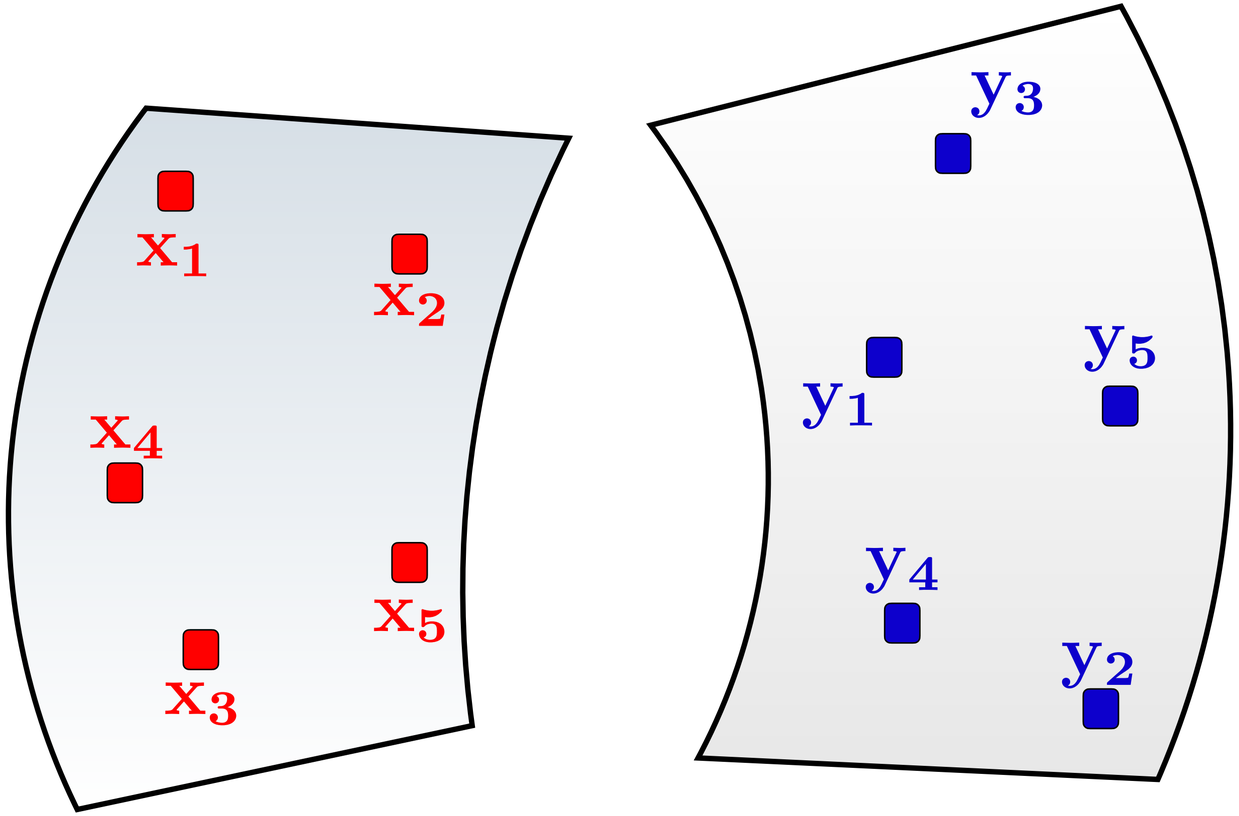
$$Q_{t+1}, \boldsymbol{g}_{t+1}, R_{t+1} \leftarrow \arg\min_{\substack{Q\in U(\boldsymbol{a},\boldsymbol{g}),\\ R\in U(\boldsymbol{b},\boldsymbol{g}),\\ \boldsymbol{g}\in\Sigma_r}} -\langle \underbrace{\boldsymbol{A}\boldsymbol{A}^T\, Q_t D(\boldsymbol{g}_t^{-1/2})}_{\boldsymbol{A}_t} \underbrace{D(\boldsymbol{g}_t^{-1/2})\, R_t^T \boldsymbol{B}^T \boldsymbol{B}}_{\boldsymbol{B}_t}, Q D(1/\boldsymbol{g})\, R^T\rangle.$$

LRGW

Fused Gromov-Wasserstein

$$\mathcal{E}(P) := \sum_{i,i',j,j'} P_{ij} P_{i'j'} \left(c_1(\mathbf{x}_i, \mathbf{x}_{i'}) - c_2(\mathbf{y}_j, \mathbf{y}_{j'})\right)^2$$

[Vayer+'18]

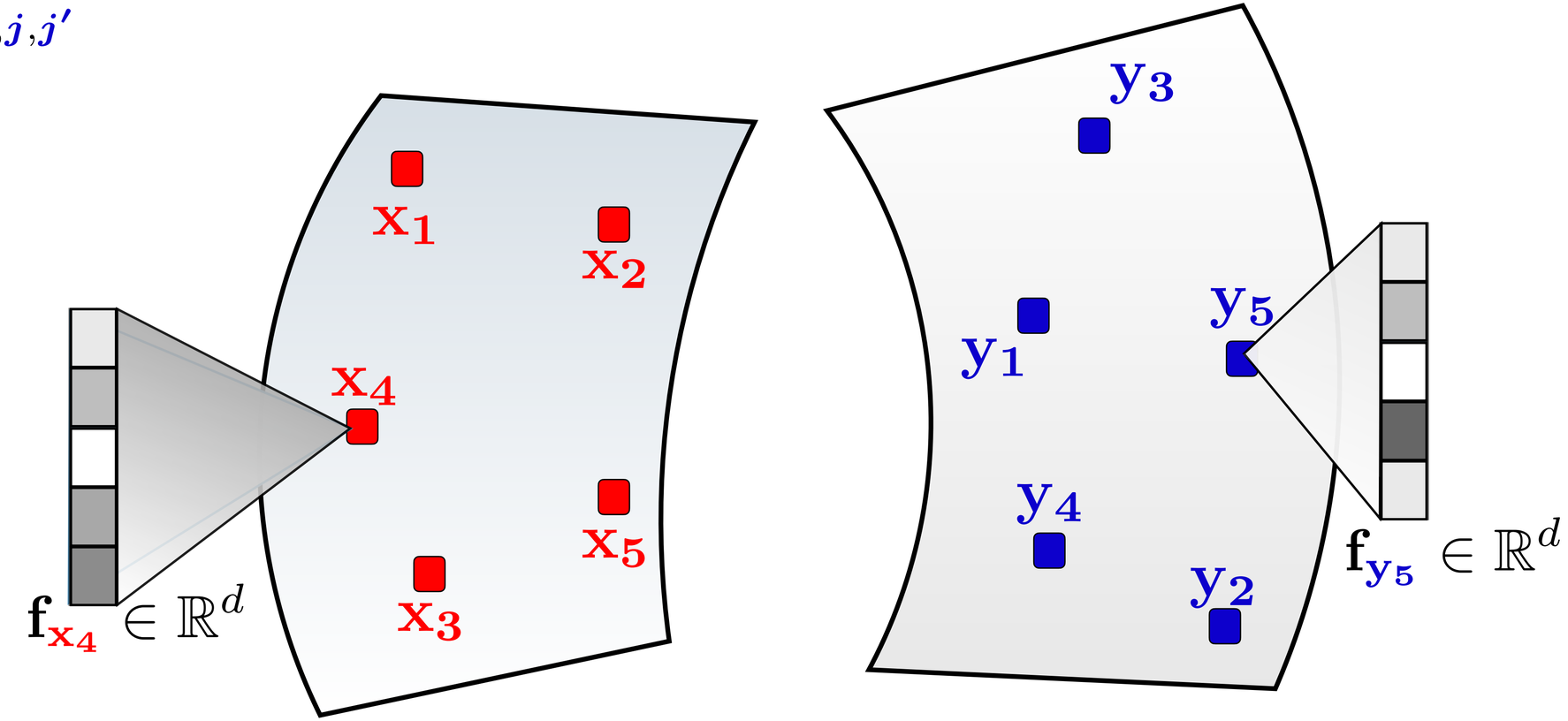


$$P_{t+1} \leftarrow \underset{P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}{\text{argmin}} \ \langle -C_1 P_t C_2 + \alpha\, C_3, P \rangle$$

Fused Gromov-Wasserstein

$$\mathcal{E}(\boldsymbol{P}) := \sum_{i,i',j,j'} \boldsymbol{P}_{ij}\boldsymbol{P}_{i'j'}\left(\boldsymbol{c}_1(\mathbf{x}_i,\mathbf{x}_{i'}) - \boldsymbol{c}_2(\mathbf{y}_j,\mathbf{y}_{j'})\right)^2$$

[Vayer+'18]

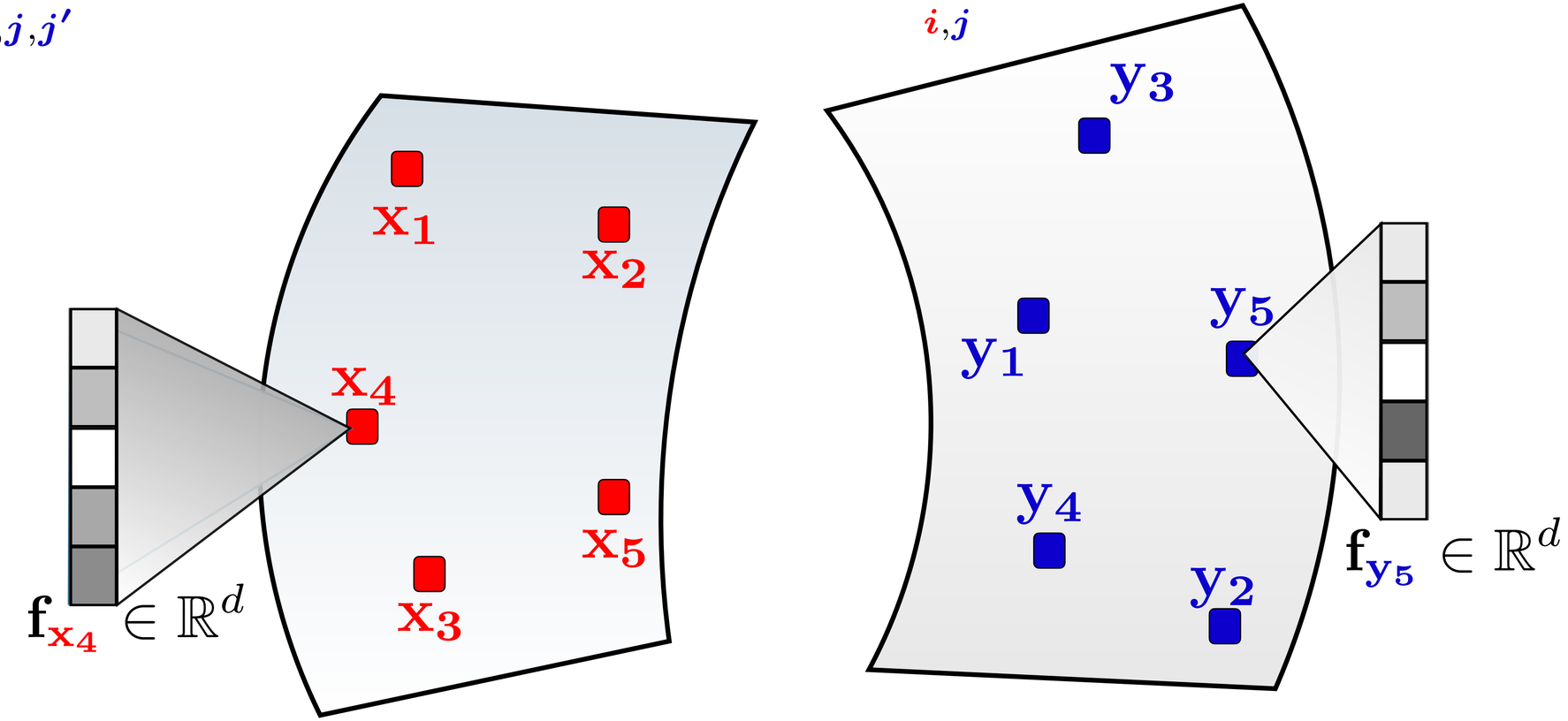


$$\boldsymbol{P}_{t+1} \leftarrow \underset{\boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}{\operatorname{argmin}} \langle -\boldsymbol{C}_1\boldsymbol{P}_t\boldsymbol{C}_2 + \alpha\, \boldsymbol{C}_3, \boldsymbol{P}\rangle$$

Fused Gromov-Wasserstein

$$\mathcal{E}(\boldsymbol{P}) := \sum_{i,i',j,j'} \boldsymbol{P}_{ij}\boldsymbol{P}_{i'j'}\left(\boldsymbol{c}_1(\mathbf{x}_i,\mathbf{x}_{i'}) - \boldsymbol{c}_2(\mathbf{y}_j,\mathbf{y}_{j'})\right)^2 + \alpha \sum_{i,j} \boldsymbol{P}_{ij}\, \boldsymbol{c}_3(\mathbf{f}_{\mathbf{x}_i},\mathbf{f}_{\mathbf{y}_j})$$

[Vayer+'18]

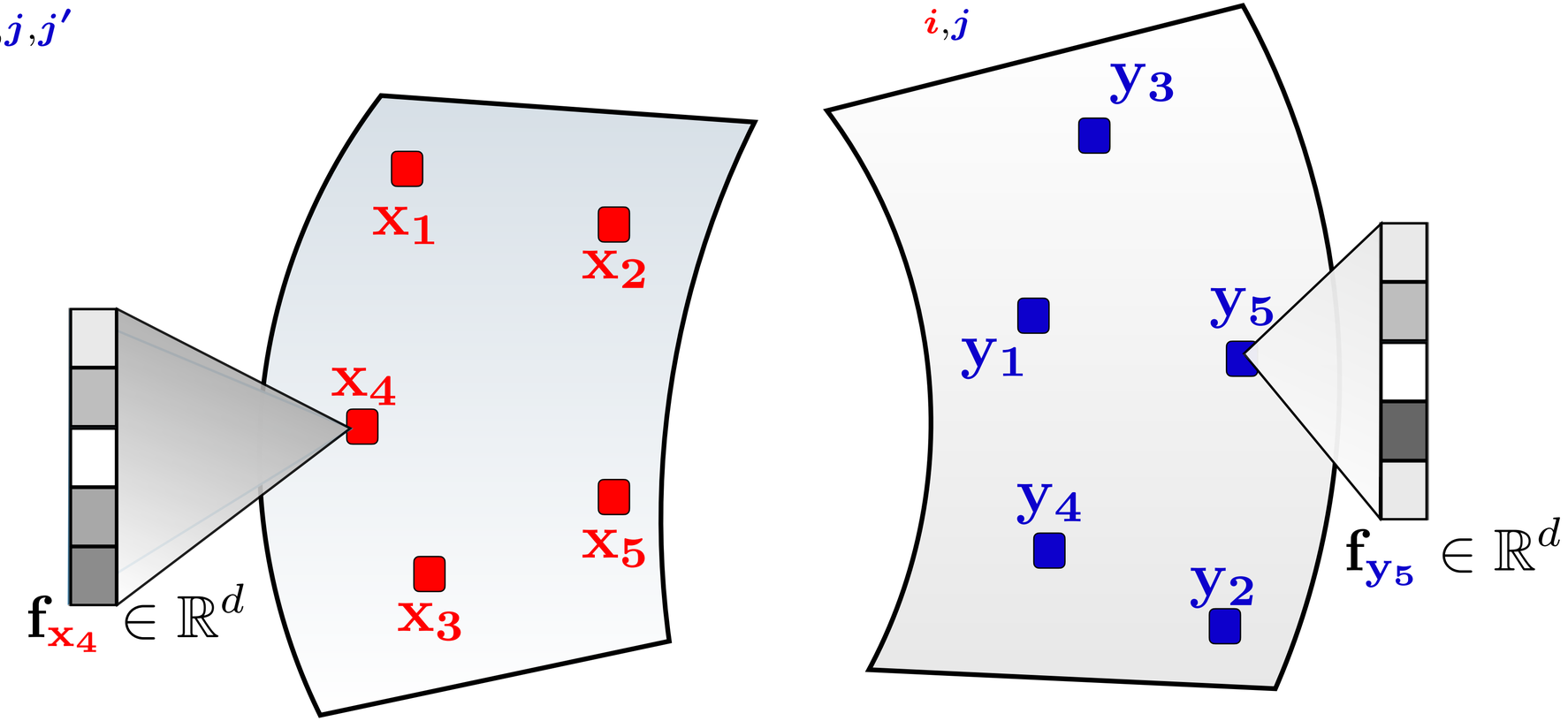


$$\boldsymbol{P}_{t+1} \leftarrow \underset{\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}}{\operatorname{argmin}} \ \langle -\boldsymbol{C}_1\boldsymbol{P}_t\boldsymbol{C}_2 + \alpha\, \boldsymbol{C}_3, \boldsymbol{P}\rangle$$

Fused Gromov-Wasserstein

$$\mathcal{E}(\boldsymbol{P}) := \sum_{i,i',j,j'} \boldsymbol{P}_{ij}\boldsymbol{P}_{i'j'}\left(\boldsymbol{c_1}(\mathbf{x}_i,\mathbf{x}_{i'}) - \boldsymbol{c_2}(\mathbf{y}_j,\mathbf{y}_{j'})\right)^2 + \alpha \sum_{i,j} \boldsymbol{P}_{ij}\, \boldsymbol{c_3}(\mathbf{f}_{\mathbf{x}_i},\mathbf{f}_{\mathbf{y}_j})$$

[Vayer+'18]



$$\boldsymbol{P}_{t+1} \leftarrow \underset{\boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}{\operatorname{argmin}} \ \langle -\boldsymbol{C_1}\boldsymbol{P}_t\boldsymbol{C_2} + \alpha\ \boldsymbol{C}_3, \boldsymbol{P}\rangle$$

... also compatible with LR formulation

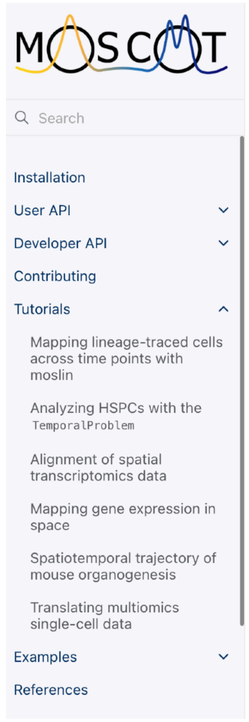
MOSCOT FGW

Rough Picture of What is Doable and What is Not

Problem \ Solver	Linear Program	Entropic Regularization $-\varepsilon E(\boldsymbol{P})$	**Low-Rank** constraint $\mathrm{rank}(\boldsymbol{P}) = r$
Linear OT Problem	N < 10k, unstable, no GPU	**quadratic (N~100k)** **Convex (Sinkhorn), GPU** [Wilson'62, Cuturi'13]	**linear (N ~ 500k)** **not cvx, GPU friendly** [Scetbon+'21]
Fused-GW Problem	Small instances, e.g. graphs	**quadratic / cubic (N~20k)** **not cvx, GPU.** [Peyré+'16]	**linear (N ~ 500k)** **not cvx, GPU** [Scetbon+'22]
Unbalanced Linear OT	Partial OT formulations	**quadratic (N~100k)** **cvx (Sinkhorn), GPU** [Frogner+'15, Chizat+16,...]	**linear (N ~ 500k)** **not cvx, GPU** **This paper, UL(F)(G)W**
Unbalanced Fused-GW	NA	**quadratic (N < 20k)** **not cvx, GPU** [Séjourné+'22]	

Application: **Reconstruct Cell Differentiation Processes**

Part 1



Moscot - Multiomics Single-cell Optimal Transport

`moscot` is a framework for optimal transport applications in single cell genomics.

If you find a model useful for your research, please consider citing the `moscot` manuscript (in preparation) as well as the publication introducing the model, which can be found in the corresponding documentation.

Installation
Learn how to install `moscot`.

User API
Find a detailed documentation of `moscot`.

Contributing
Add a functionality or report a bug.

Examples
Find brief and concise examples of `moscot`.

Tutorials
Check out how to use `moscot` for data analysis.

Manuscript
Please have a look at our manuscript [Klein *et al.*, 2023] to learn more.

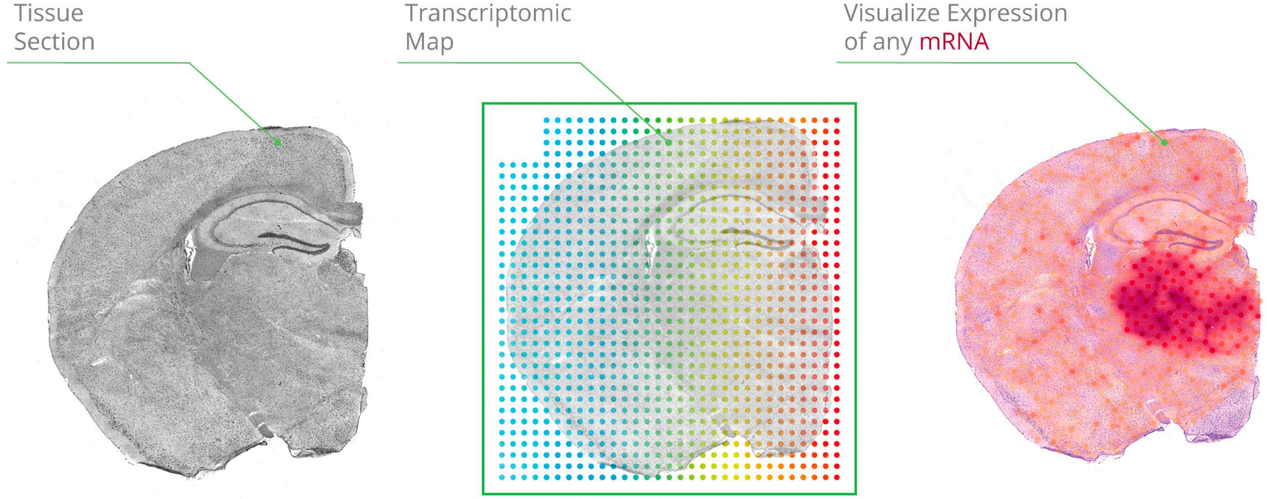
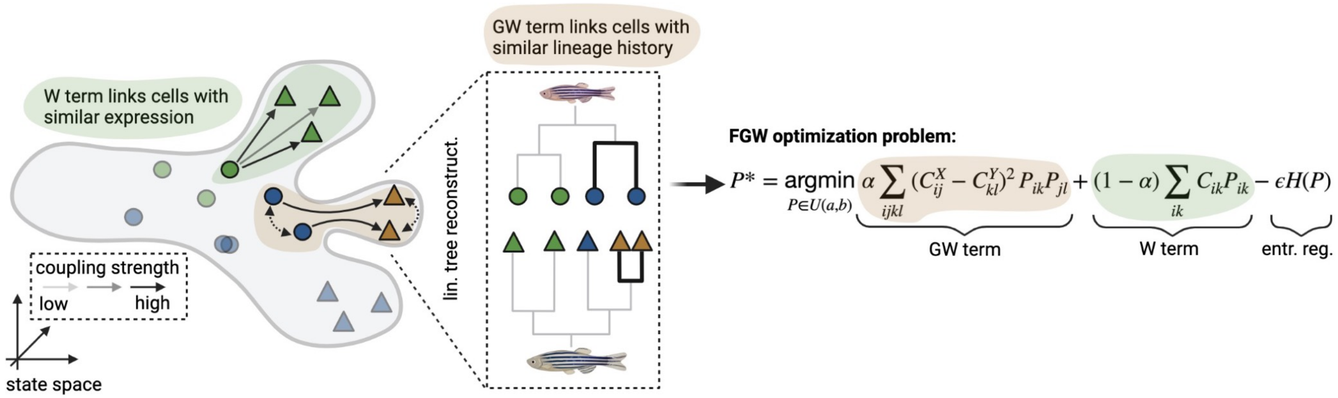


Illustration: https://spatialtranscriptomics.com/technology/

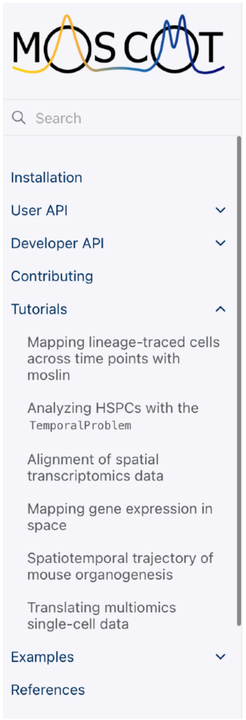


FGW optimization problem:

$$P^* = \underset{P \in U(a,b)}{\operatorname{argmin}} \underbrace{\alpha \sum_{ijkl} (C^X_{ij} - C^Y_{kl})^2 P_{ik} P_{jl}}_{\text{GW term}} + \underbrace{(1-\alpha) \sum_{ik} C_{ik} P_{ik}}_{\text{W term}} - \underbrace{\epsilon H(P)}_{\text{entr. reg.}}$$

Application: **Reconstruct Cell Differentiation Processes**

Part 1



Moscot - Multiomics Single-cell Optimal Transport

`moscot` is a framework for optimal transport applications in single cell genomics.

MOSCOT

If you find a model useful for your research, please consider citing the `moscot` manuscript (in preparation) as well as the publication introducing the model, which can be found in the corresponding documentation.

Installation
Learn how to install `moscot`.

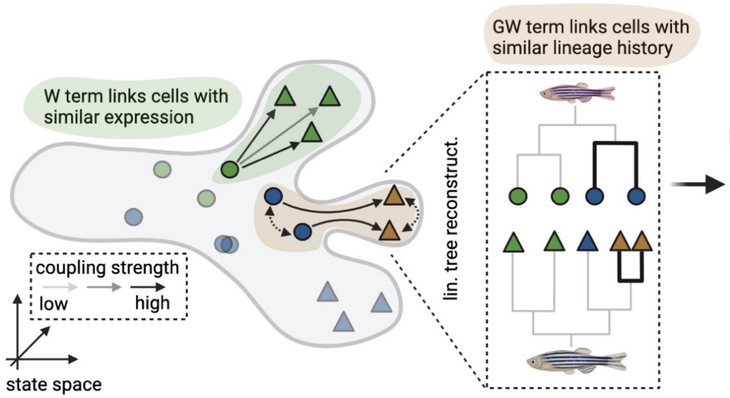
User API
Find a detailed documentation of `moscot`.

Contributing
Add a functionality or report a bug.

Examples
Find brief and concise examples of `moscot`.

Tutorials
Check out how to use `moscot` for data analysis.

Manuscript
Please have a look at our manuscript [Klein *et al.*, 2023] to learn more.



FGW optimization problem:

$$P^* = \underset{P \in U(a,b)}{\operatorname{argmin}} \underbrace{\alpha \sum_{ijkl} (C^X_{ij} - C^Y_{kl})^2 P_{ik} P_{jl}}_{\text{GW term}} + \underbrace{(1-\alpha) \sum_{ik} C_{ik} P_{ik}}_{\text{W term}} - \underbrace{\epsilon H(P)}_{\text{entr. reg.}}$$

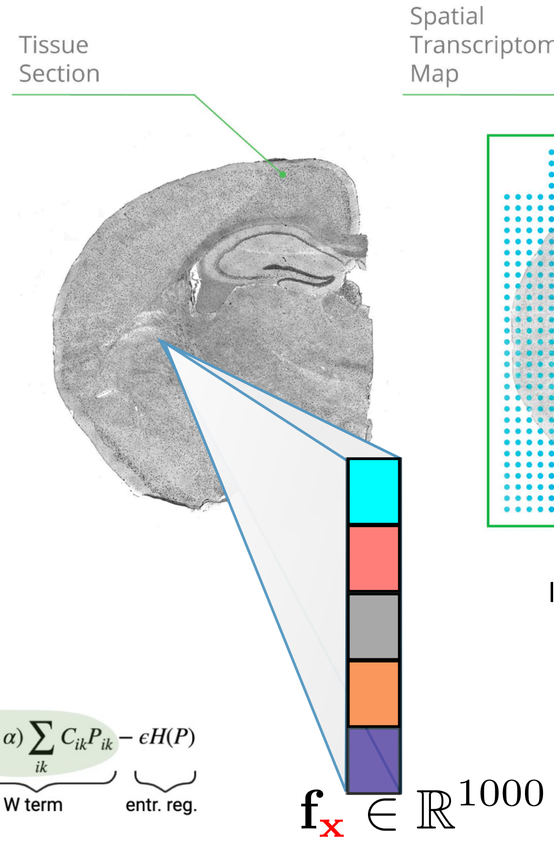


Illustration: https://spatialtranscriptomics.com/technology/

Perspective

These solvers are used as subroutines, tricks to speed/stabilize them matter

- (Nonlinear) acceleration of fixed point iterations, and their differentiation — Accel
- Faster kernel multiplications (with caveat that they must remain >0) — Grid
- Amortized and/or clever initializations, notably for LR approaches — MetaOT

Theory and statistics help setting right levels of regularization

- Stats: Sinkhorn vs. LP plugin [Altschuler+17, Genevay+18, Mena+19, Stromme+22]
- Debiasing approaches $\hat{W}_{\mathrm{reg}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = W_{\mathrm{reg}}(\boldsymbol{\mu}, \boldsymbol{\nu}) - \frac{1}{2}(W_{\mathrm{reg}}(\boldsymbol{\mu}, \boldsymbol{\mu}) + W_{\mathrm{reg}}(\boldsymbol{\nu}, \boldsymbol{\nu}))$
  [Genevay+17, Feydy+18, Chizat+22] — SinkDiv

More structured losses, beyond linear/quadratic [Alvarez-Melis+19, Redko+20]

Choosing cost-functions adaptively [Cuturi+11, Paty+19, Lin+20, Chuang+23]

Outline of the Tutorial

Prelude — Warm-Up: Starting with Optimal Matchings

Part 1 — Kantorovich Formulation of OT, Discrete Computations

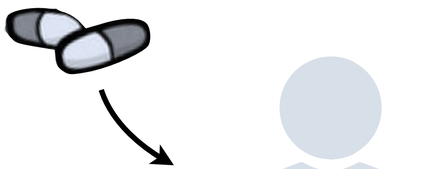
Part 2 — **Monge Formulation, Duality, Gangbo-McCann-Brenier Theorems**

- Monge Formulation and Kantorovich duality
- c-concavity, Gangbo-McCann and Brenier theorems
- Entropic map estimators, learning neural Monge maps

Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile

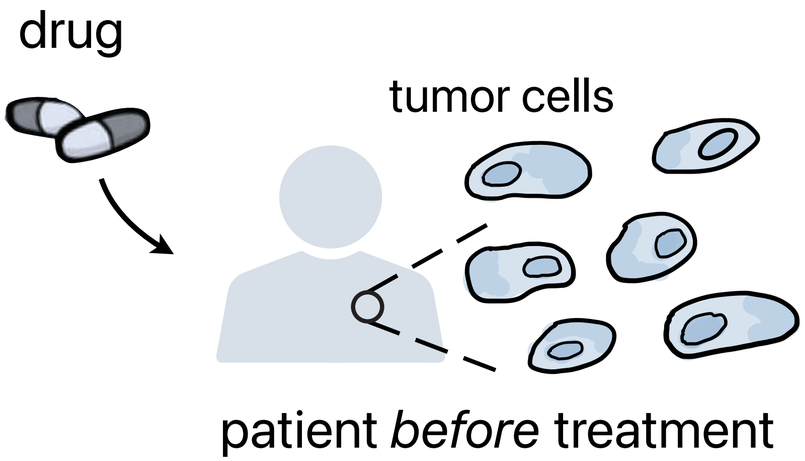
drug



patient *before* treatment

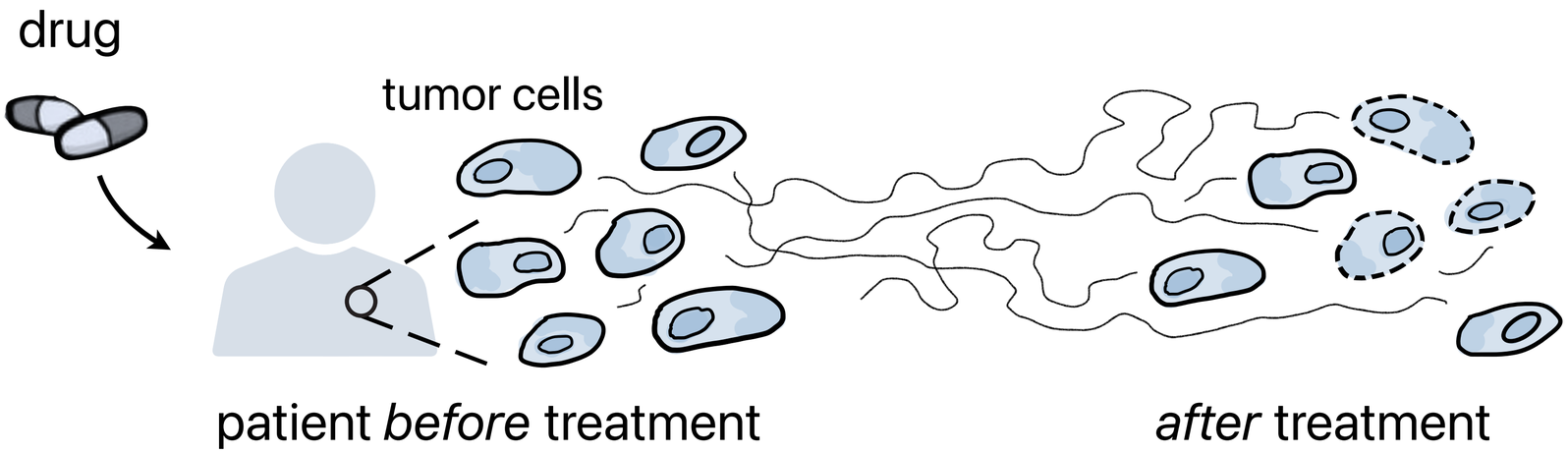
Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile



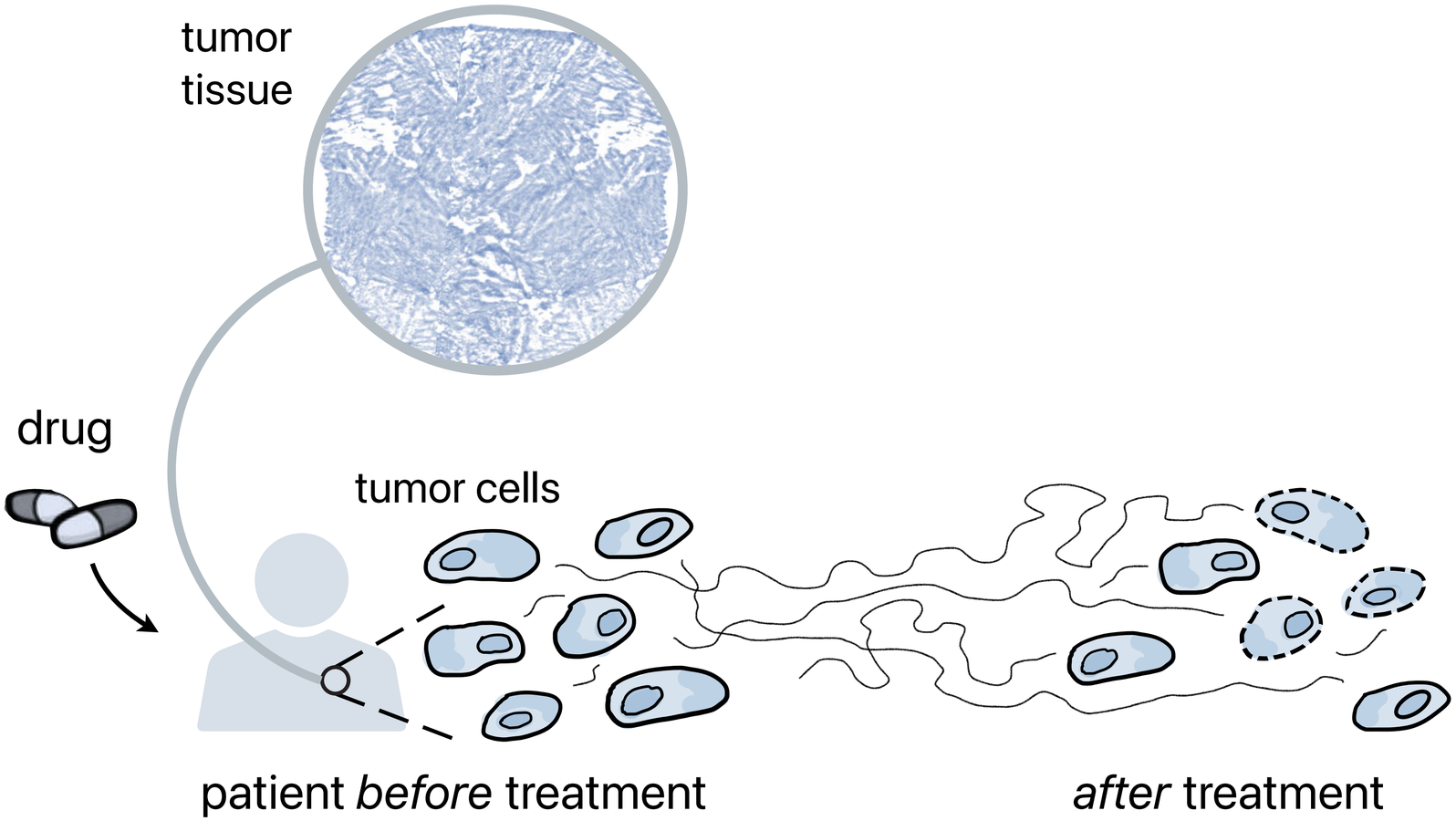
Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile



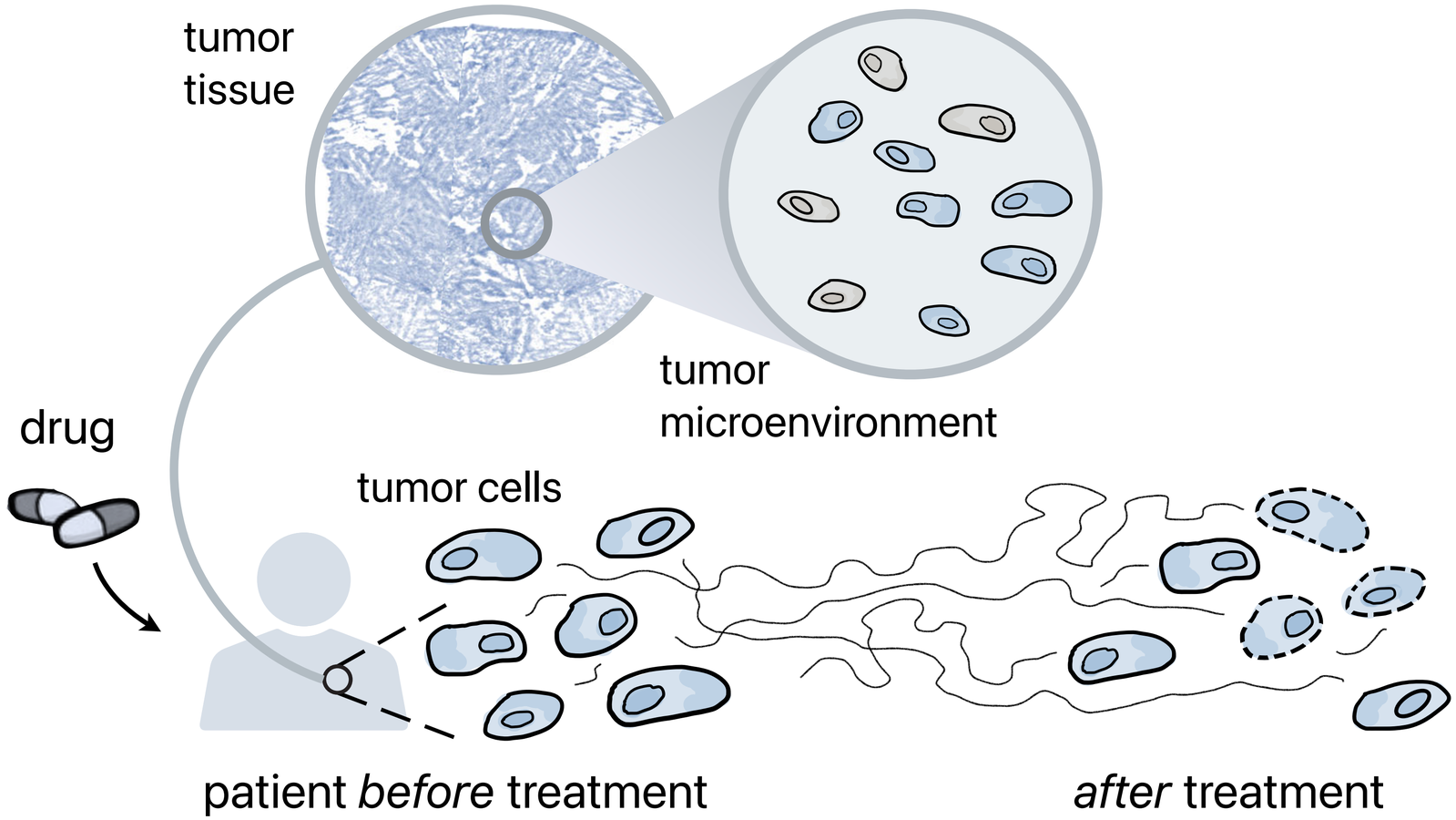
Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile



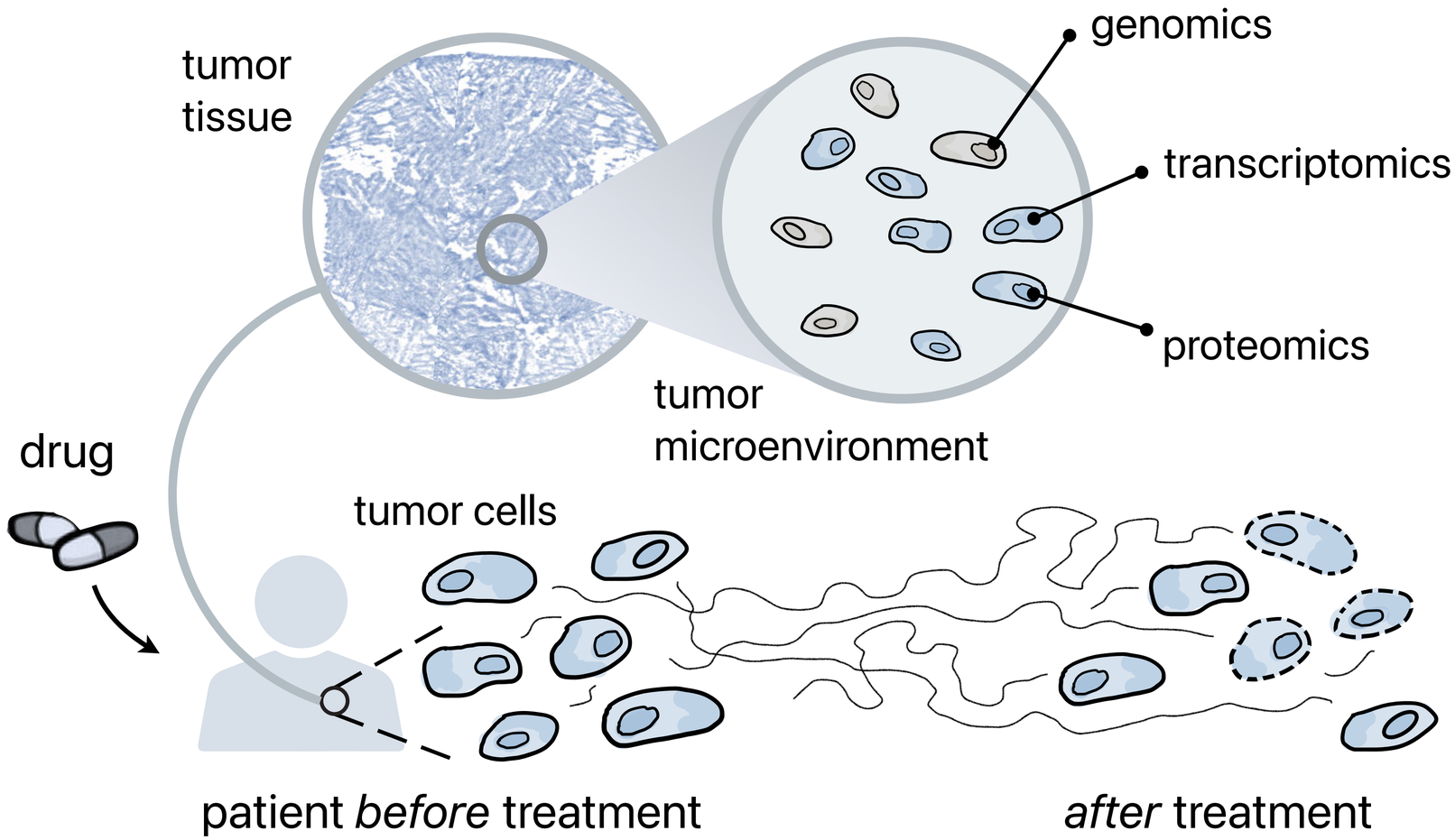
Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile



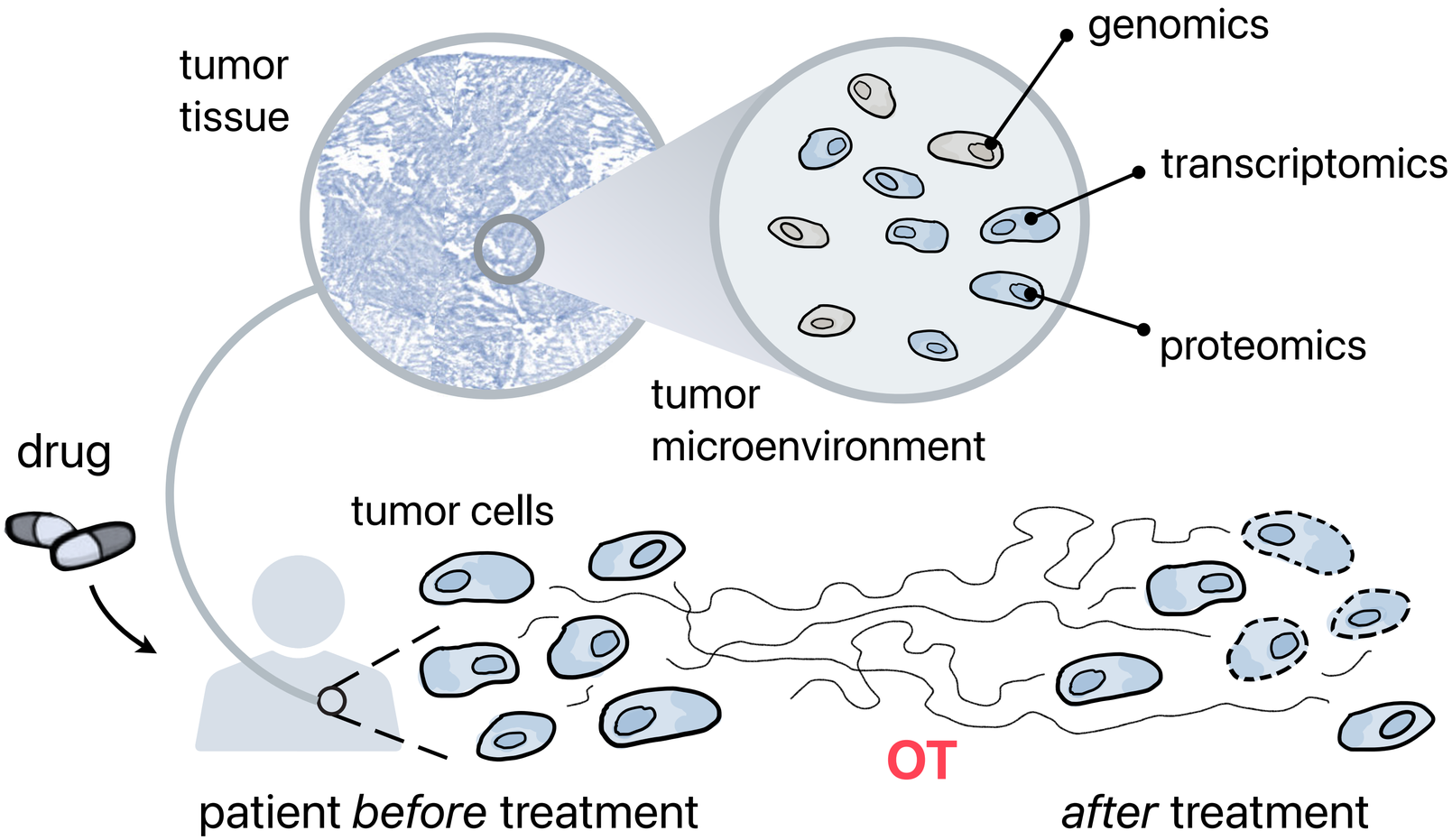
Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile



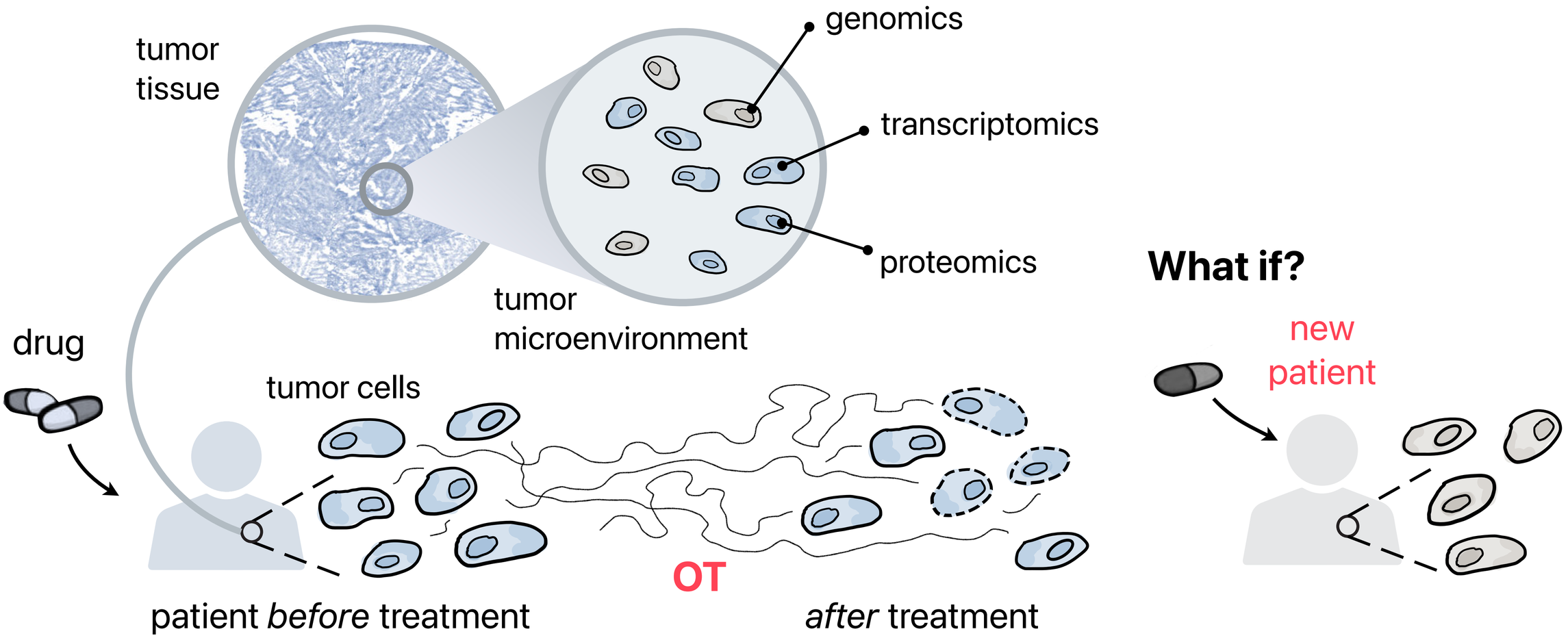
Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile



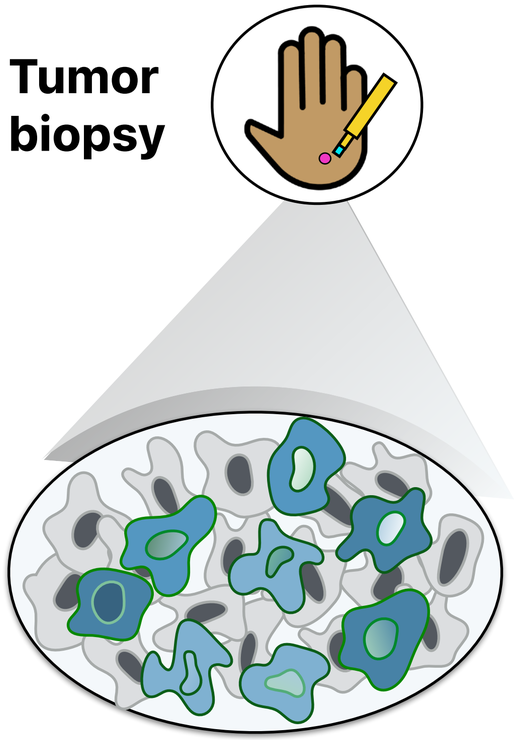
Motivating Example: **Personalized Medecine**

Goal: give patients **targeted therapies** tailored to their molecular profile

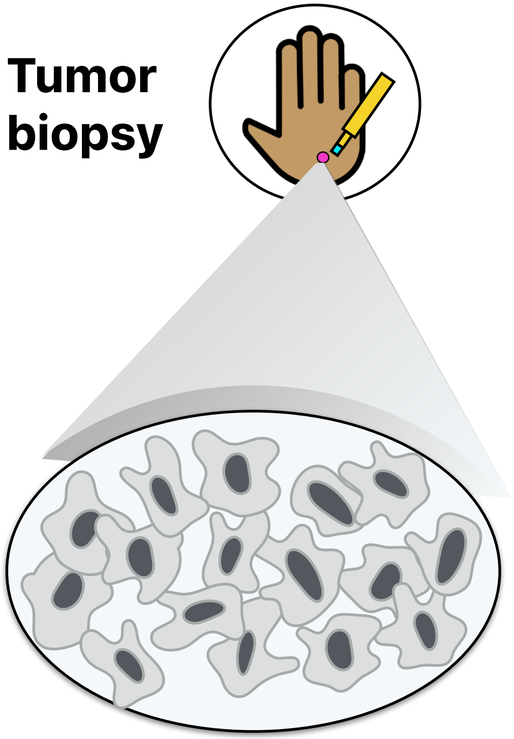


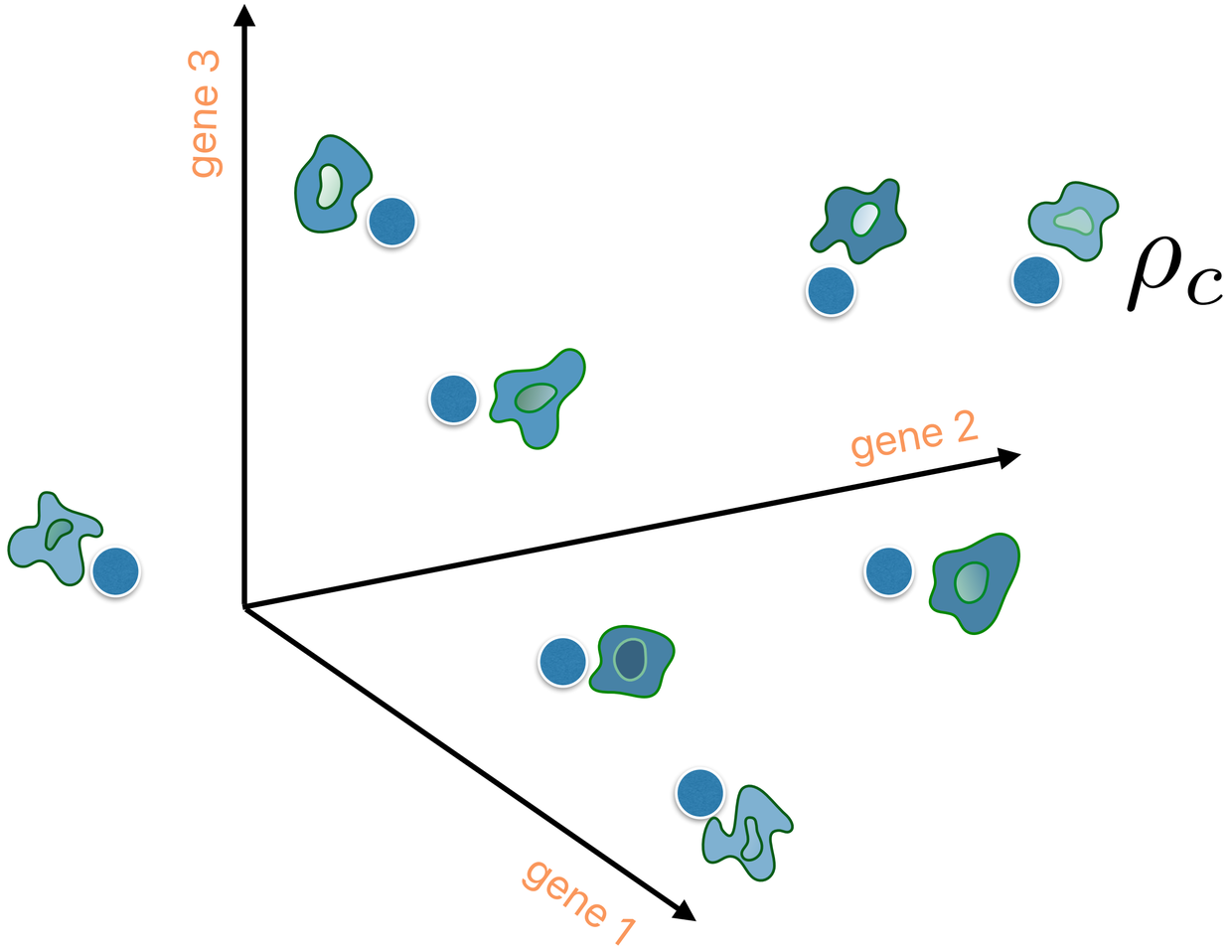
Motivating Example: Single-Cell Omics

Tumor biopsy



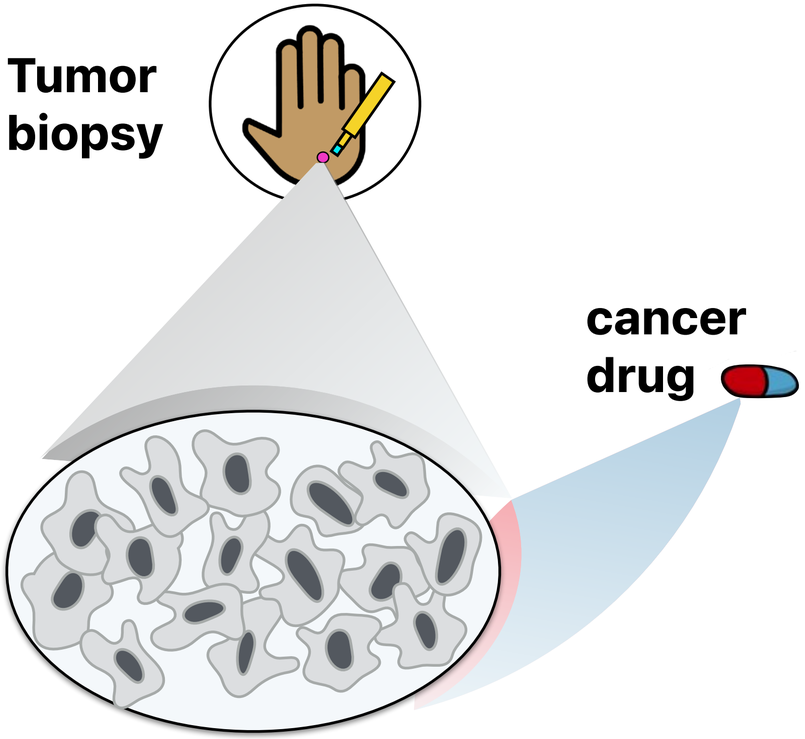
Motivating Example: Single-Cell Omics

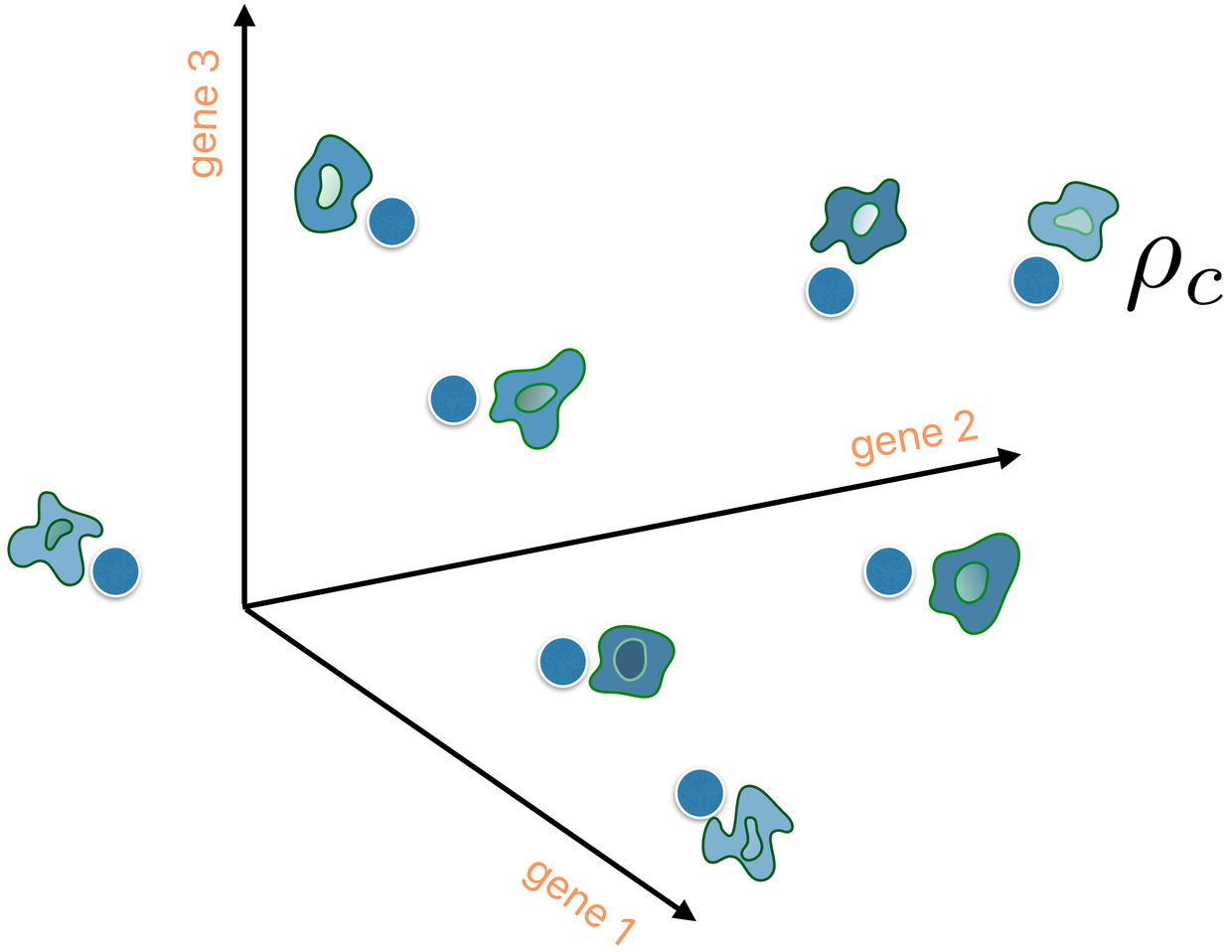




$d = 35{,}000$

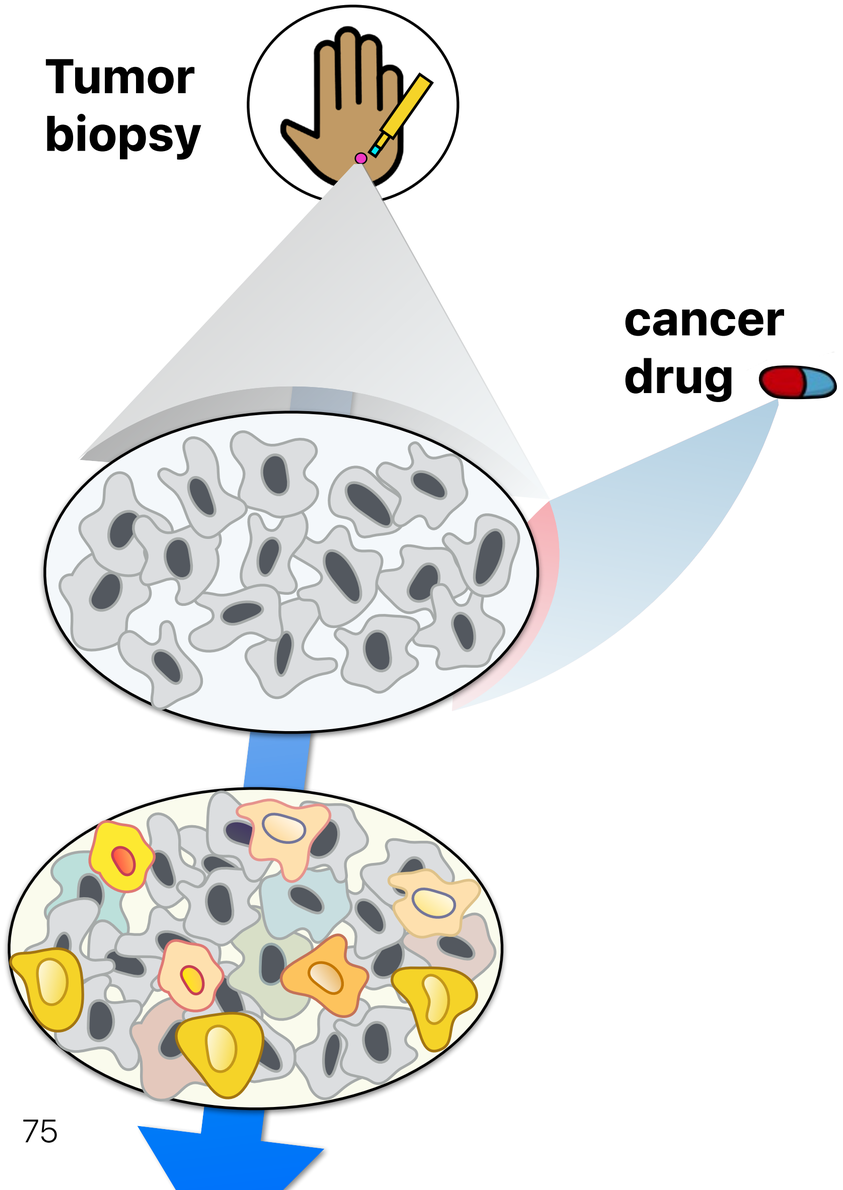
Motivating Example: Single-Cell Omics

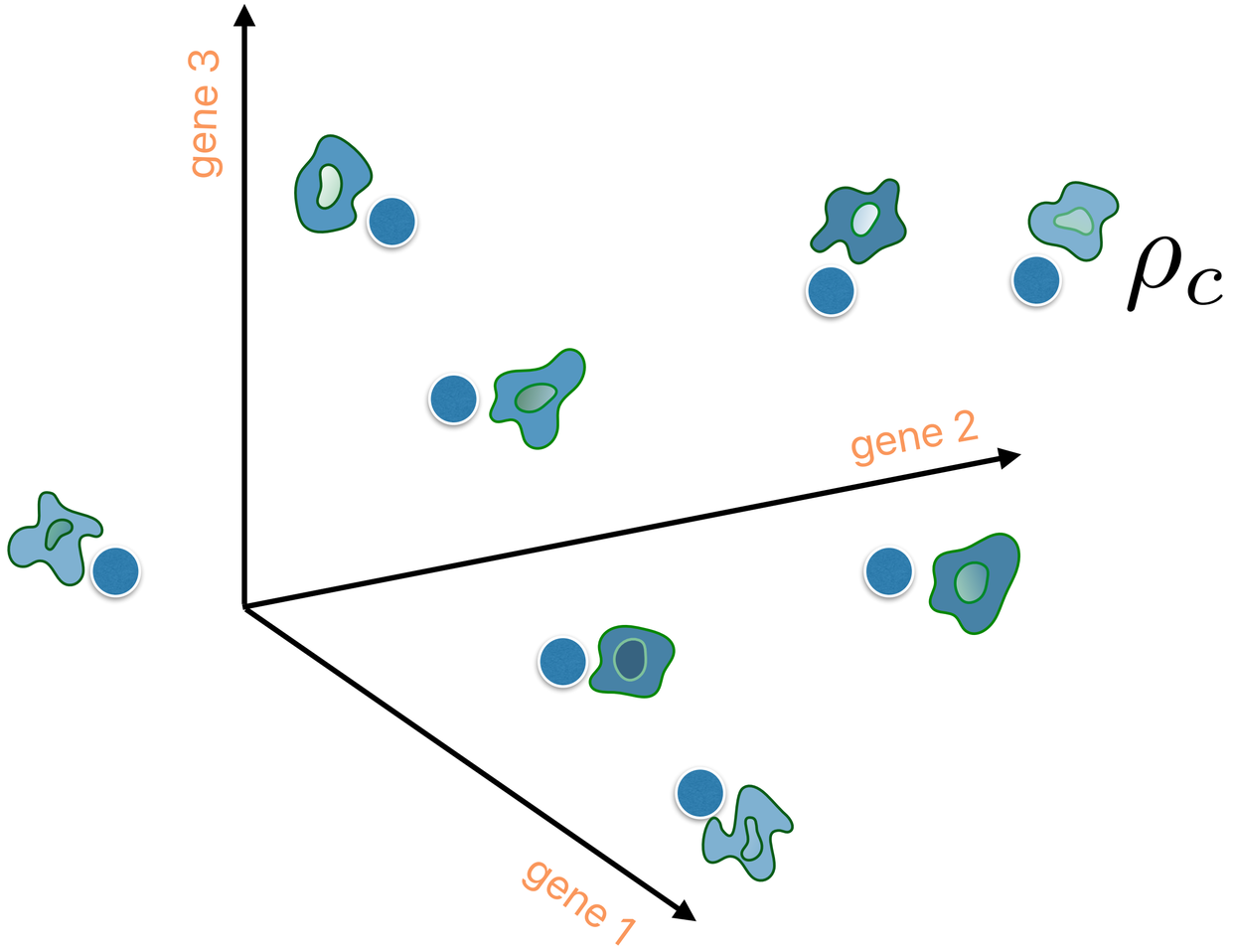




$d = 35{,}000$

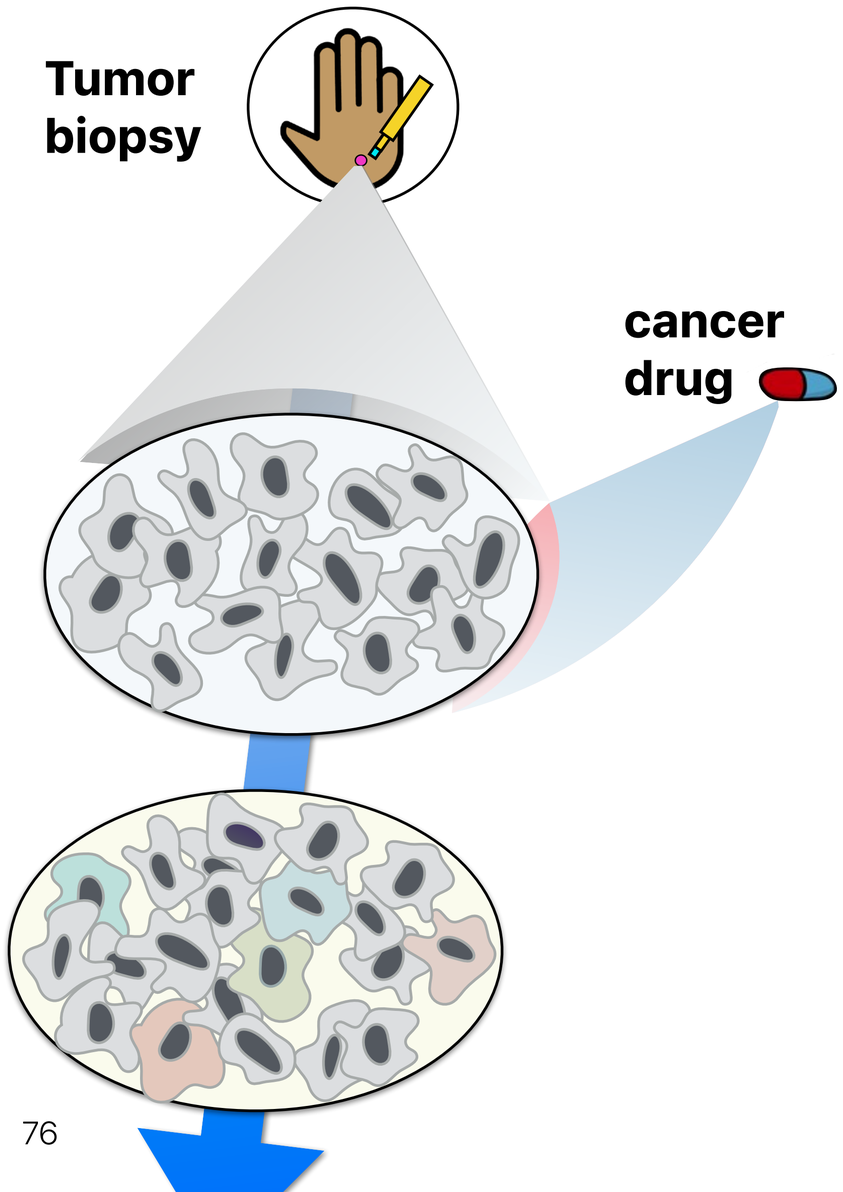
Motivating Example: Single-Cell Omics



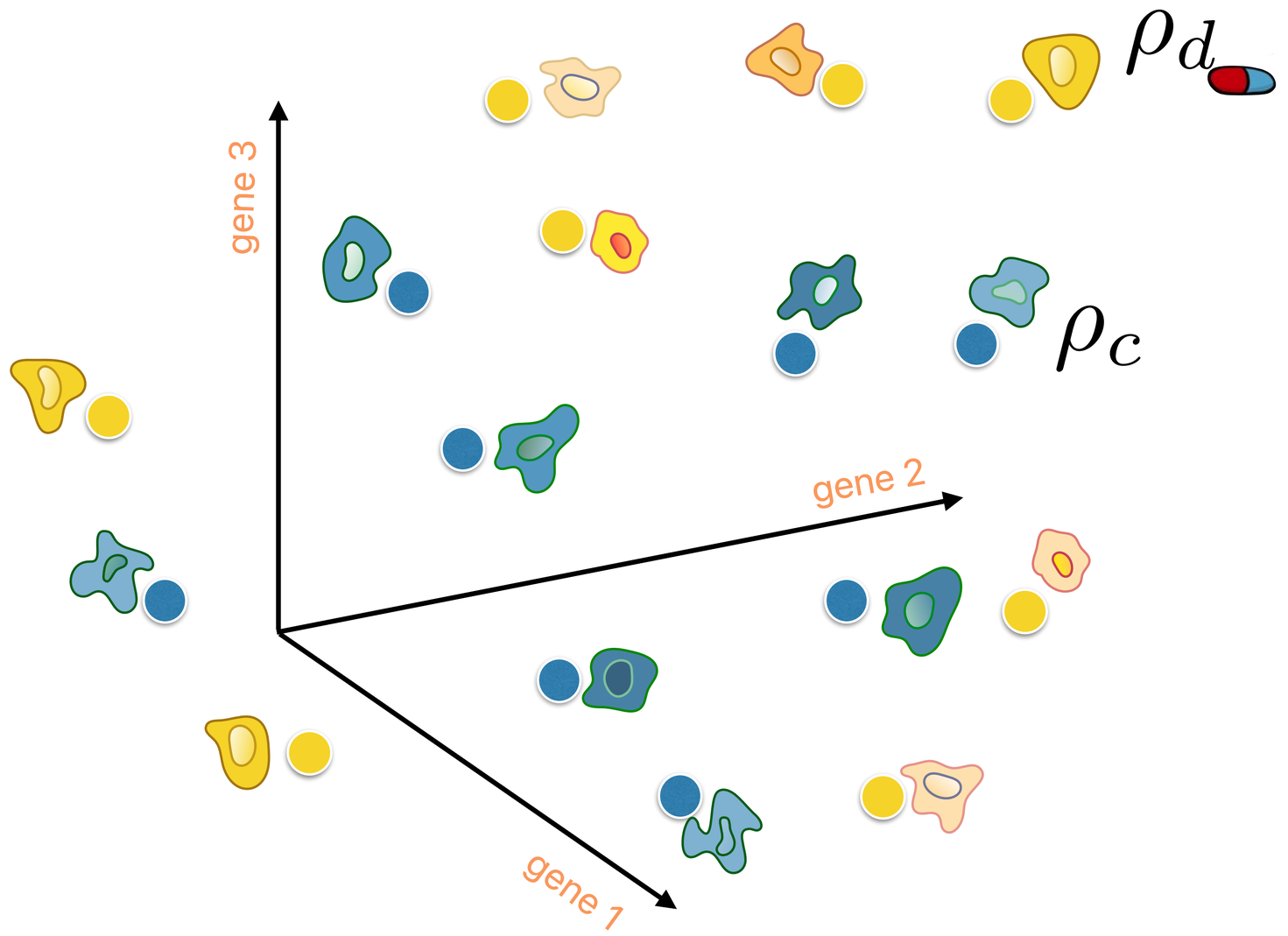


$d = 35{,}000$

Motivating Example: Single-Cell Omics

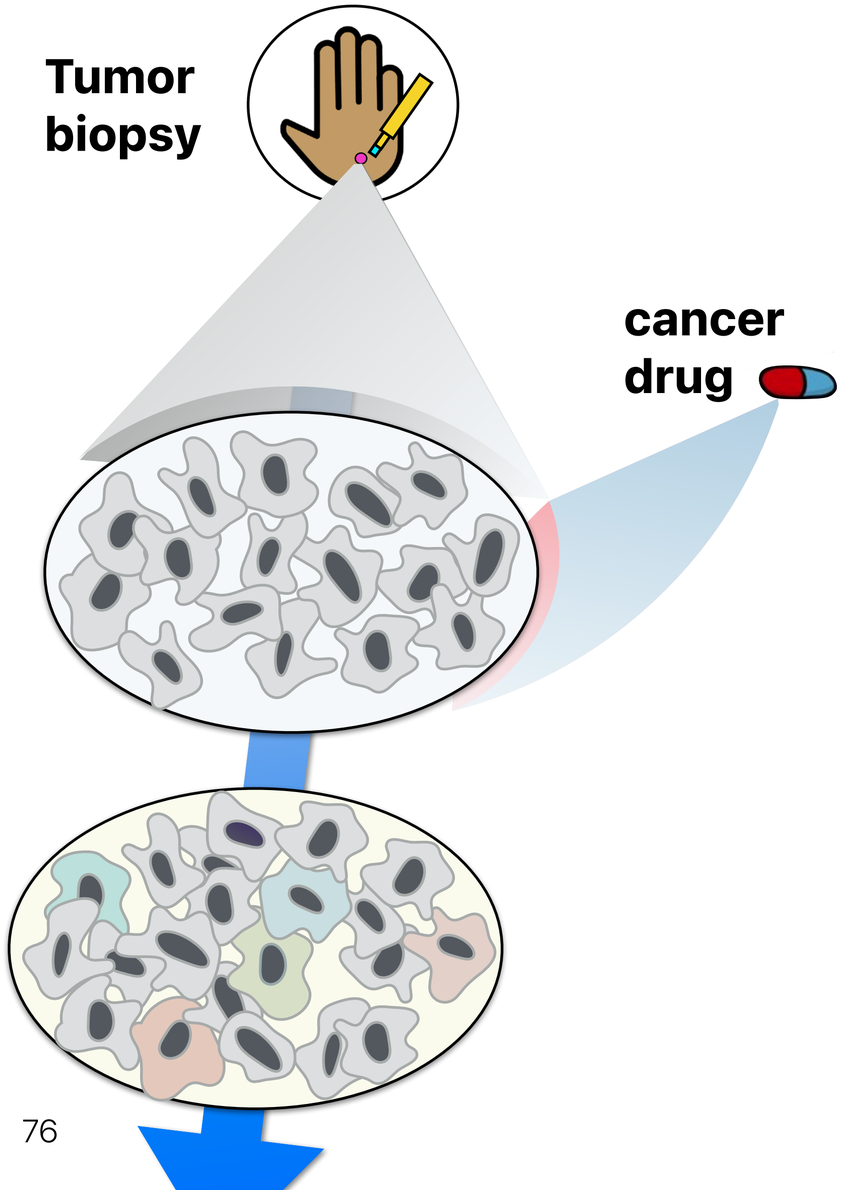


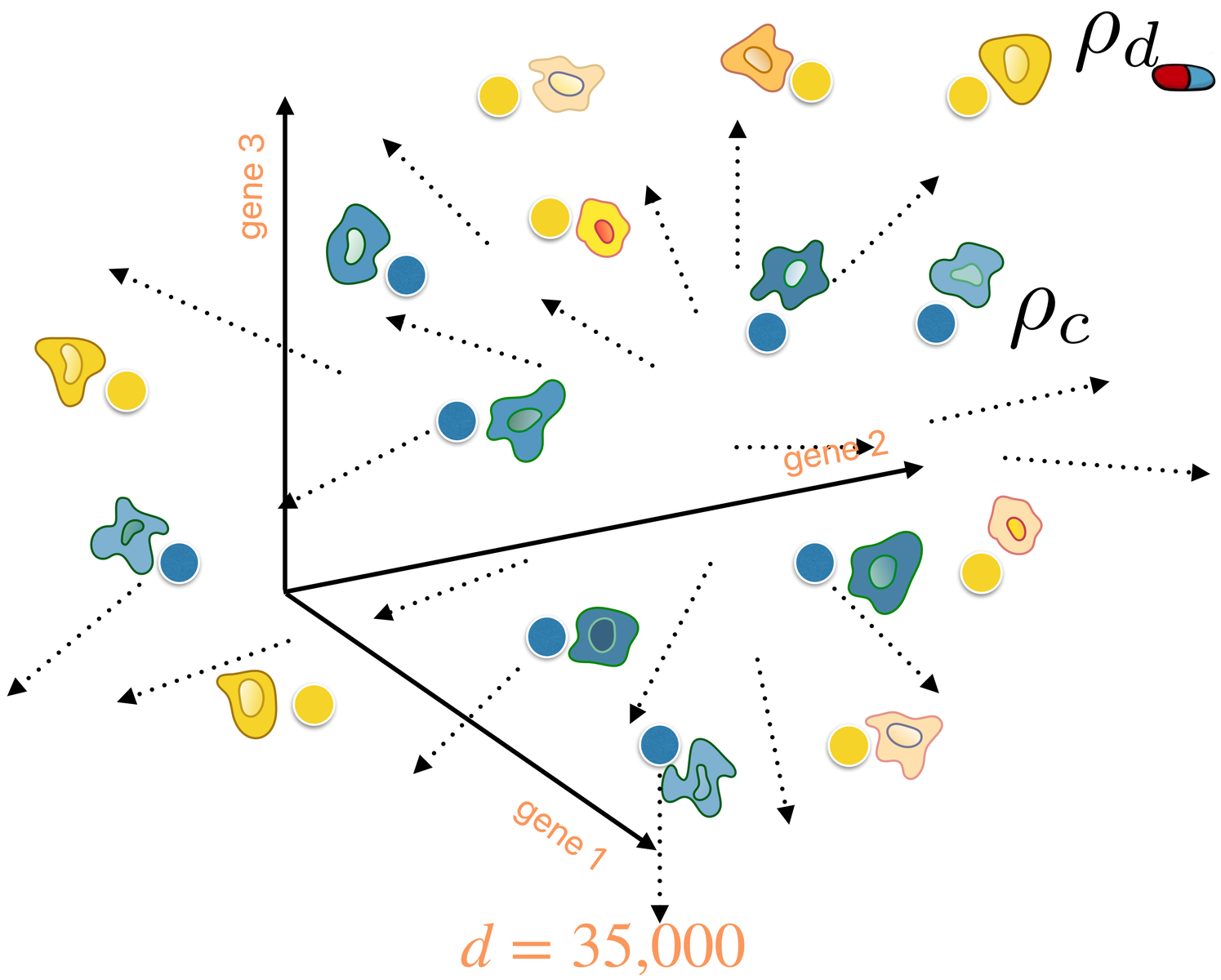
cancer
drug



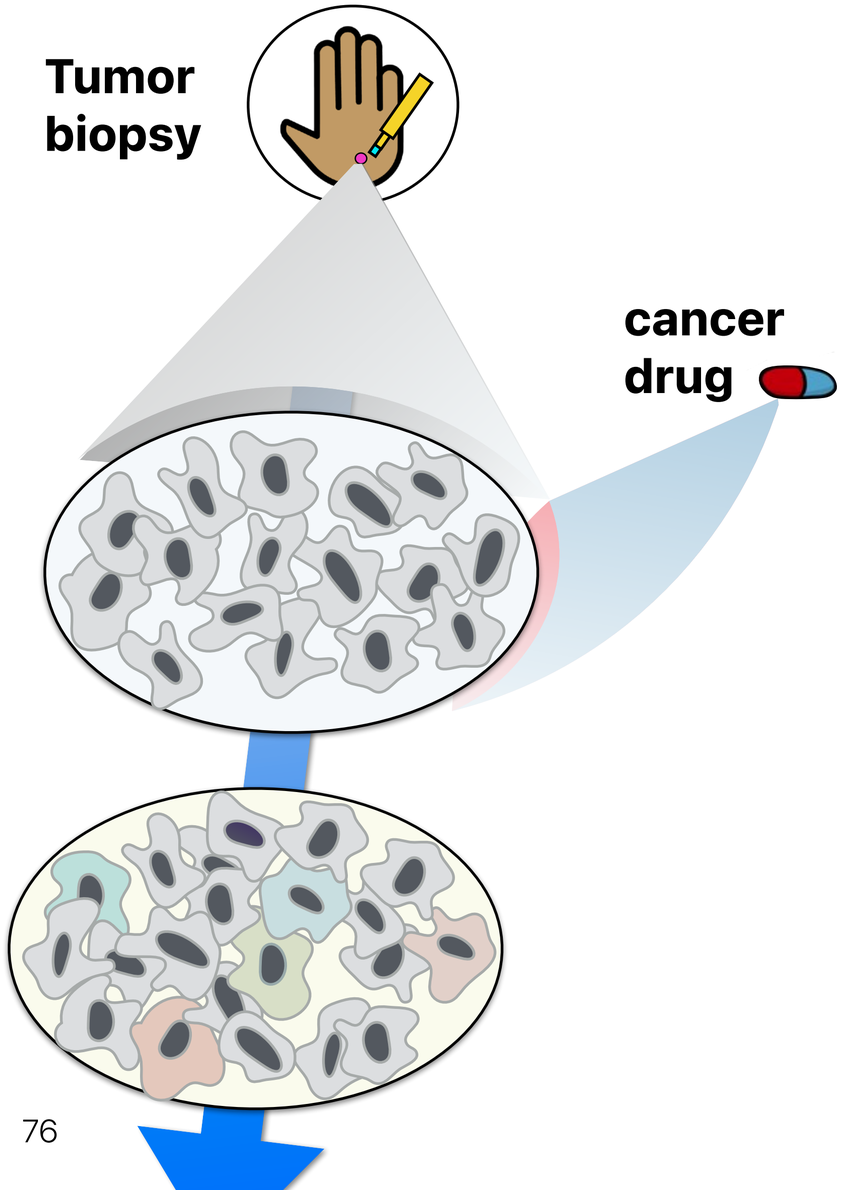
$d = 35{,}000$

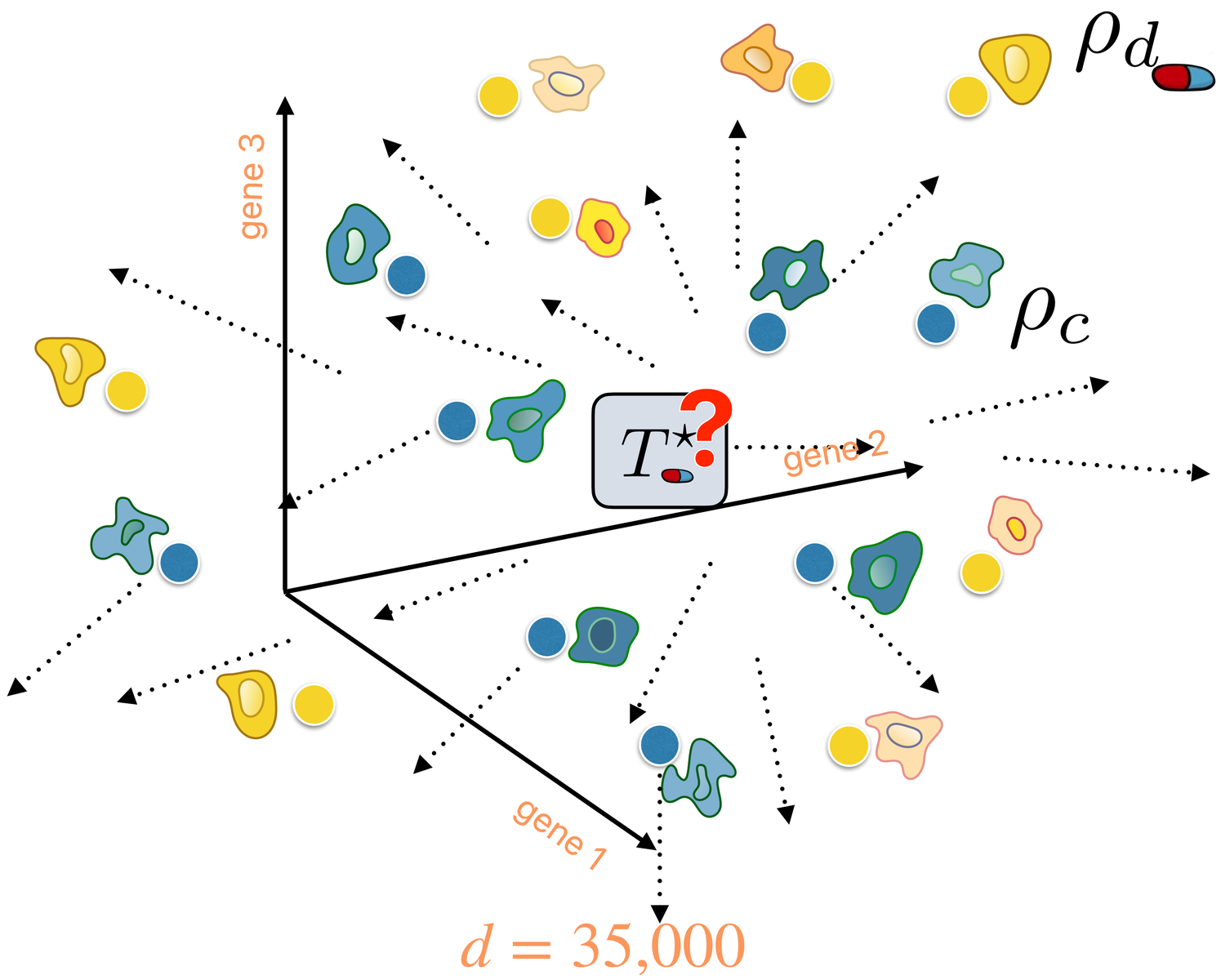
Motivating Example: Single-Cell Omics



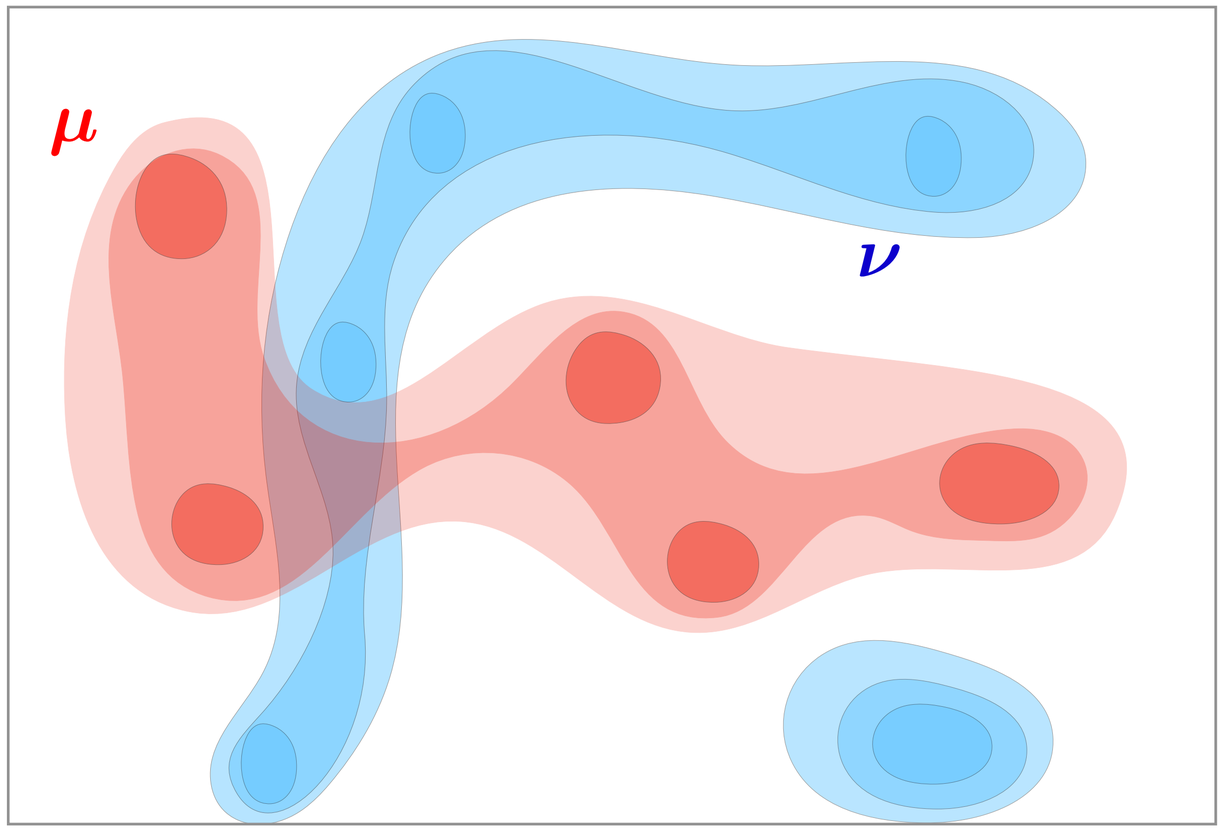


Motivating Example: Single-Cell Omics





Monge Maps




Monge

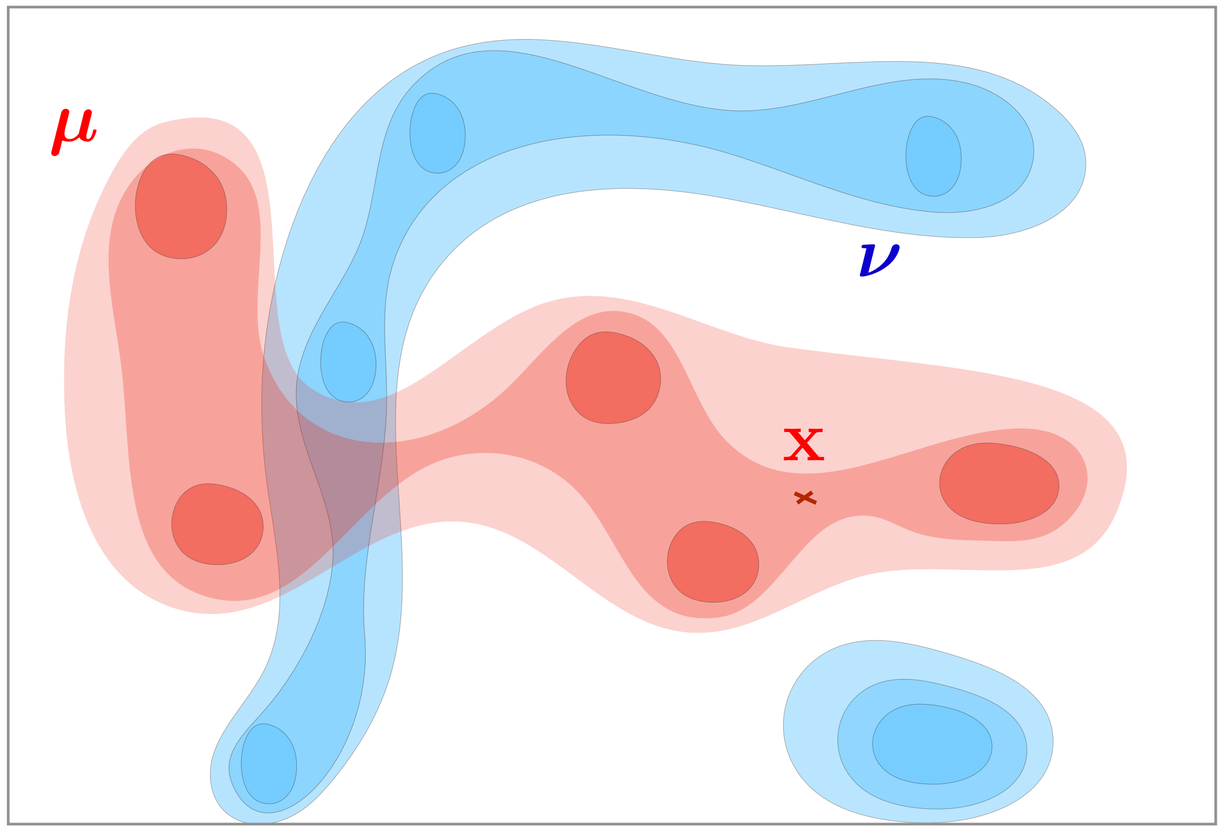

Brenier


Gangbo


McCann


Caffarelli

Monge Maps




Monge

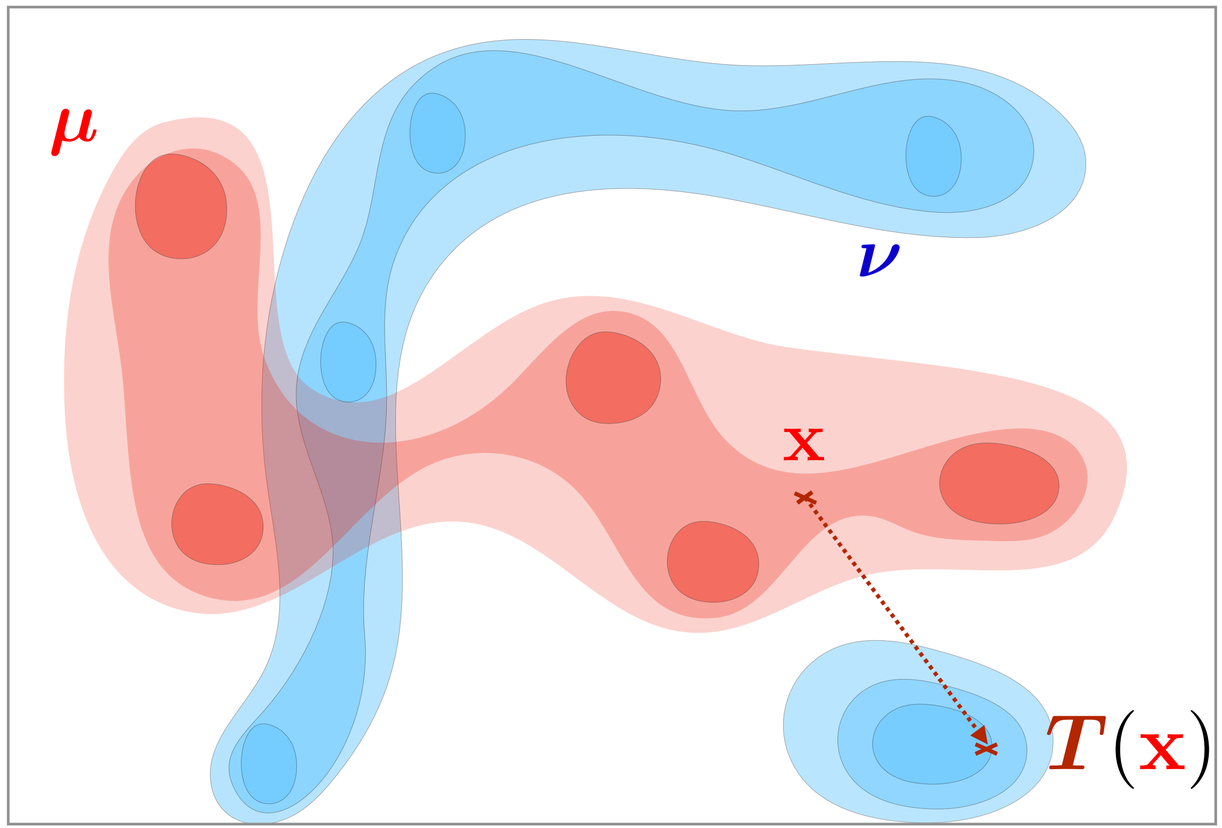

Brenier


Gangbo


McCann


Caffarelli

Monge Maps




Monge

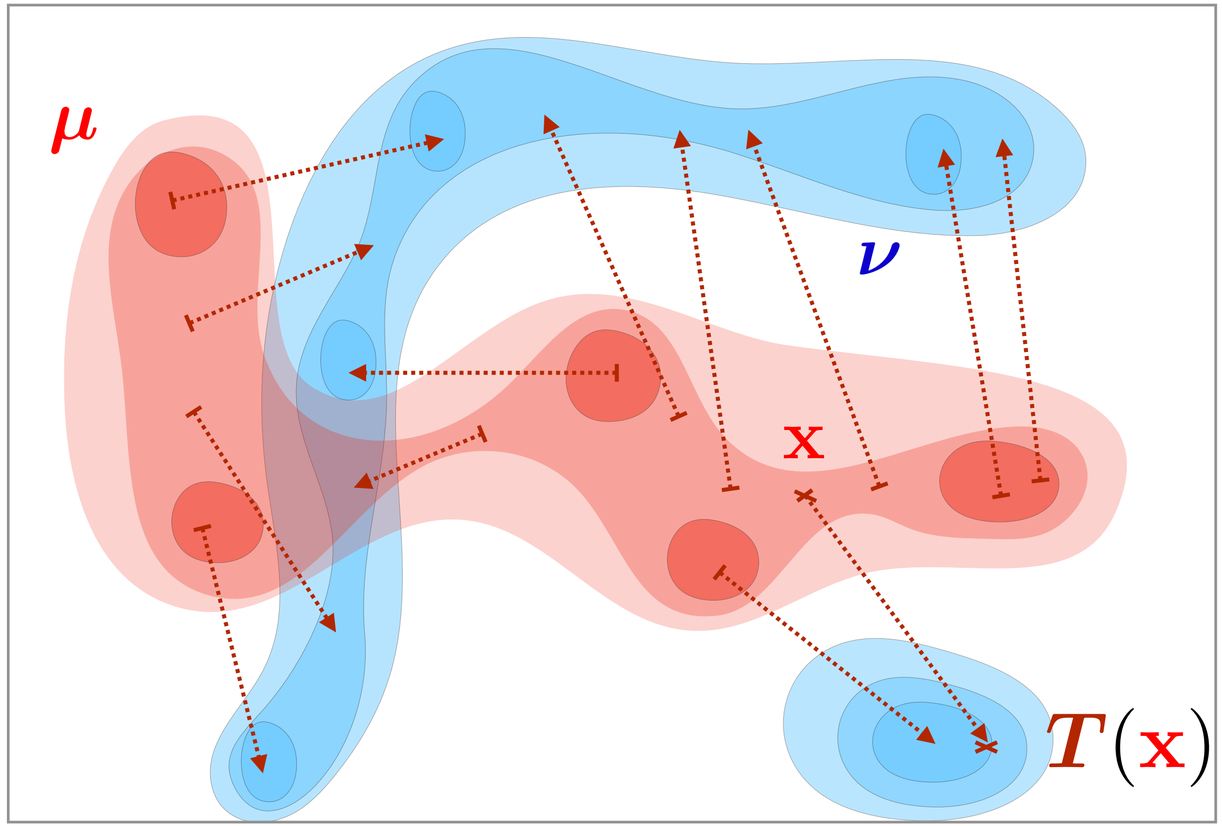

Brenier


Gangbo


McCann


Caffarelli

Monge Maps




Monge

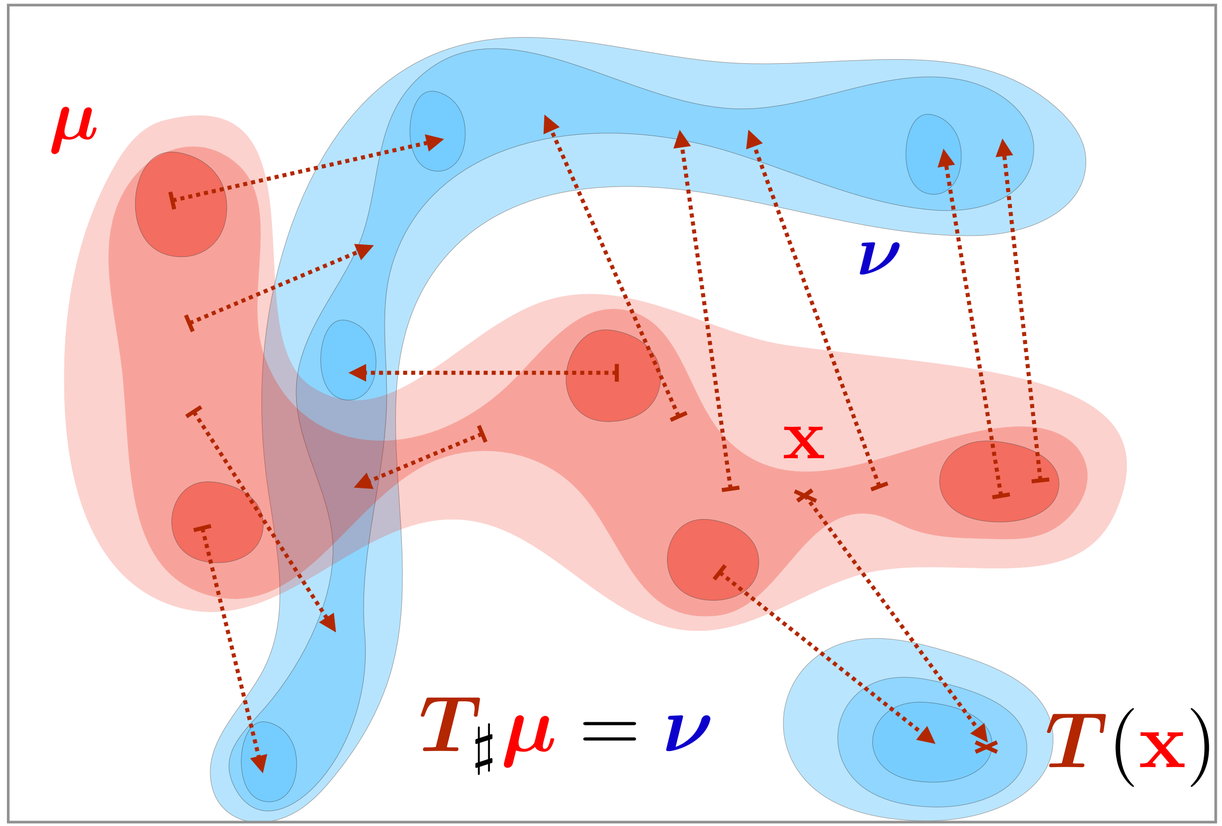

Brenier


Gangbo


McCann


Caffarelli

Monge Maps




Monge

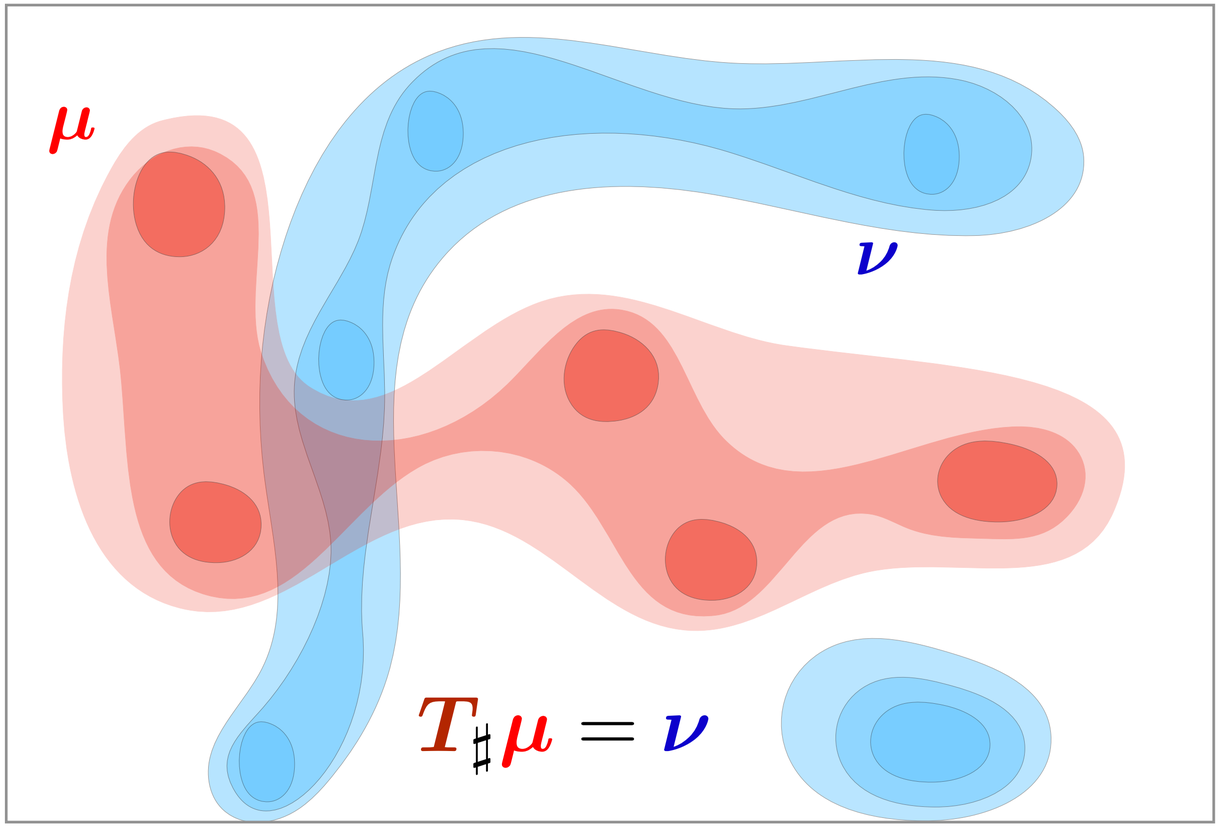

Brenier


Gangbo


McCann


Caffarelli

Monge Maps




Monge


Brenier


Gangbo


McCann

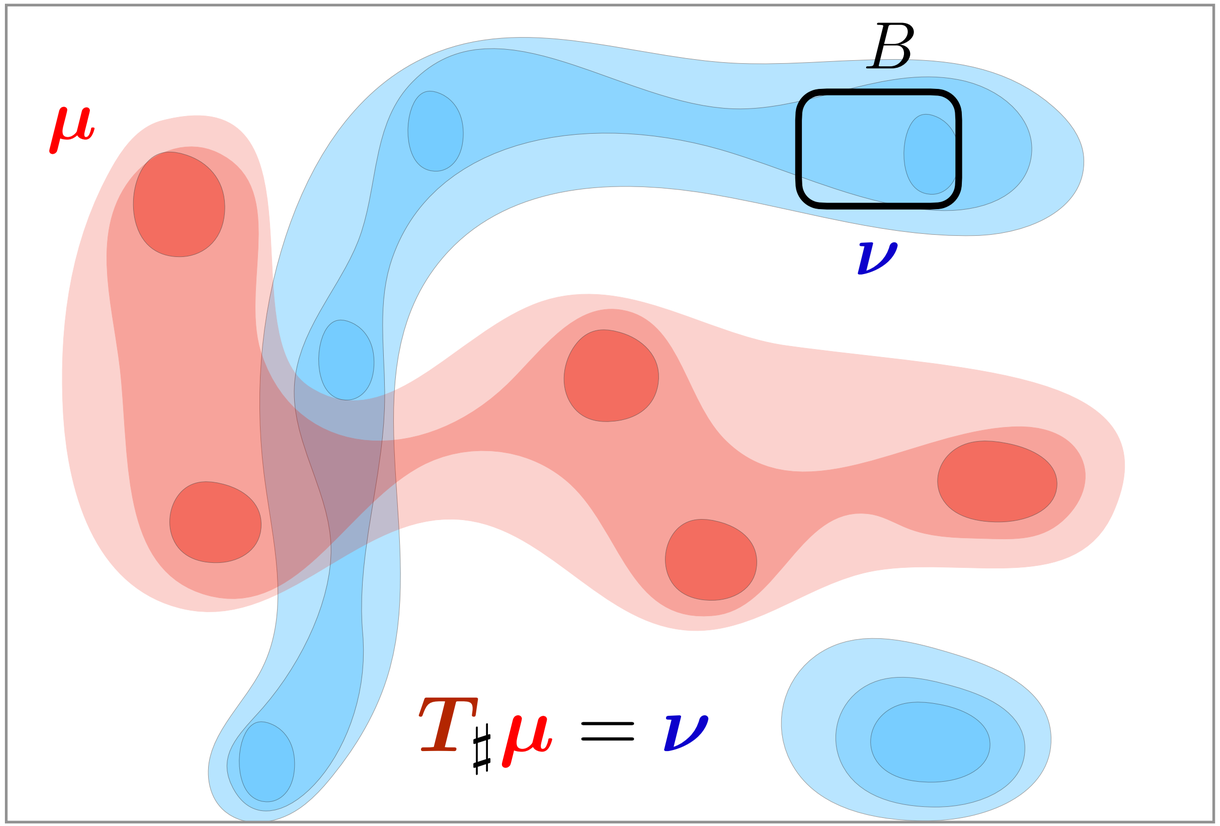

Caffarelli

$$T_\sharp \mu = \nu$$

$$\Updownarrow$$

$$\forall B \subset \mathbb{R}^d, \nu(B) = \mu(T^{-1}(B))$$

Monge Maps




Monge


Brenier


Gangbo


McCann

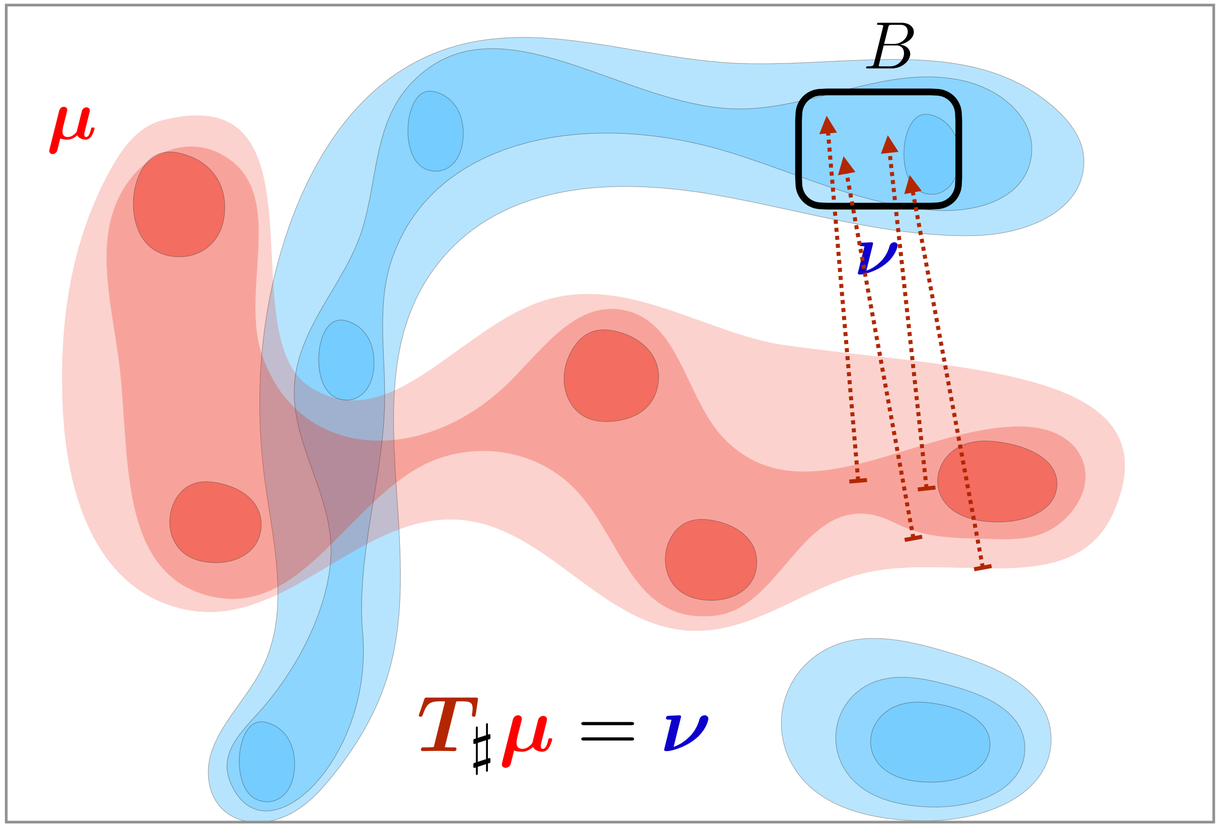

Caffarelli

$$T_\sharp \mu = \nu$$

$$\Updownarrow$$

$$\forall B \subset \mathbb{R}^d, \nu(B) = \mu(T^{-1}(B))$$

Monge Maps




Monge


Brenier


Gangbo


McCann

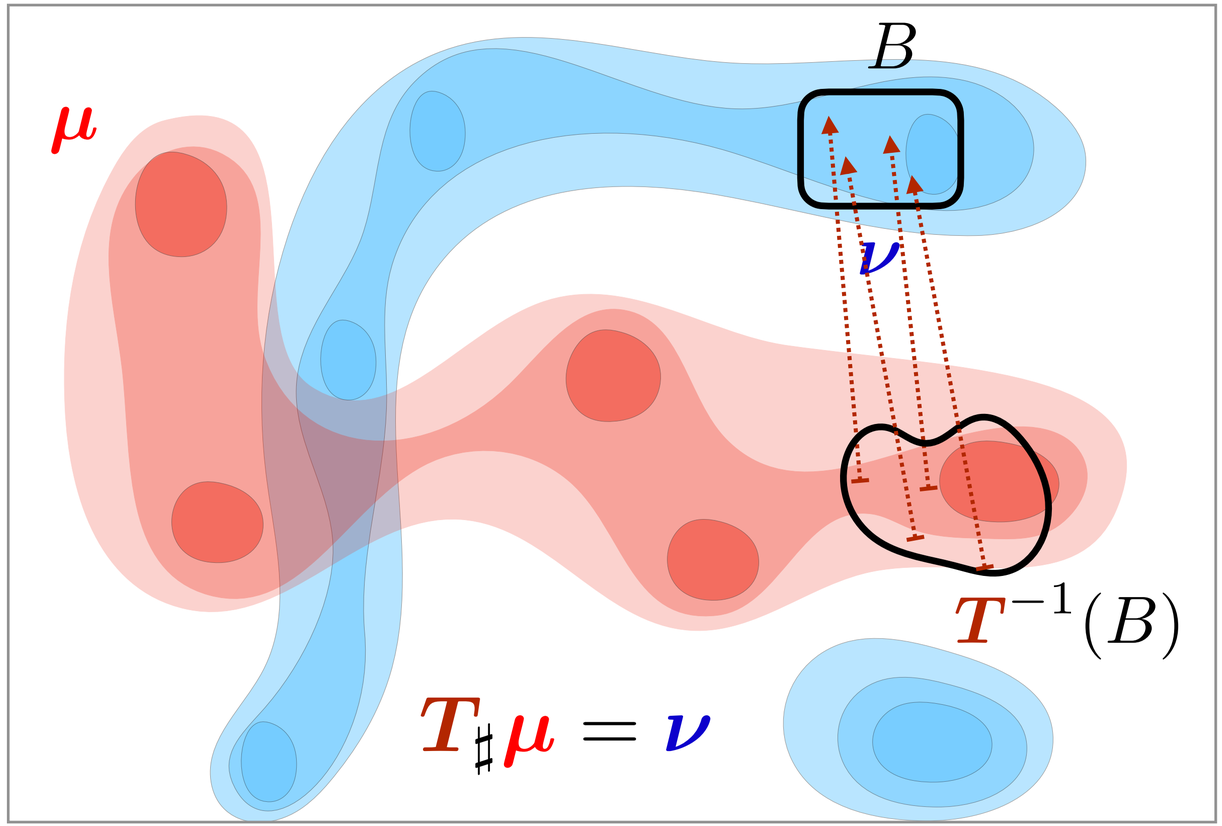

Caffarelli

$$T_\sharp \mu = \nu$$

$$\Updownarrow$$

$$\forall B \subset \mathbb{R}^d, \nu(B) = \mu(T^{-1}(B))$$

Monge Maps




Monge


Brenier


Gangbo


McCann

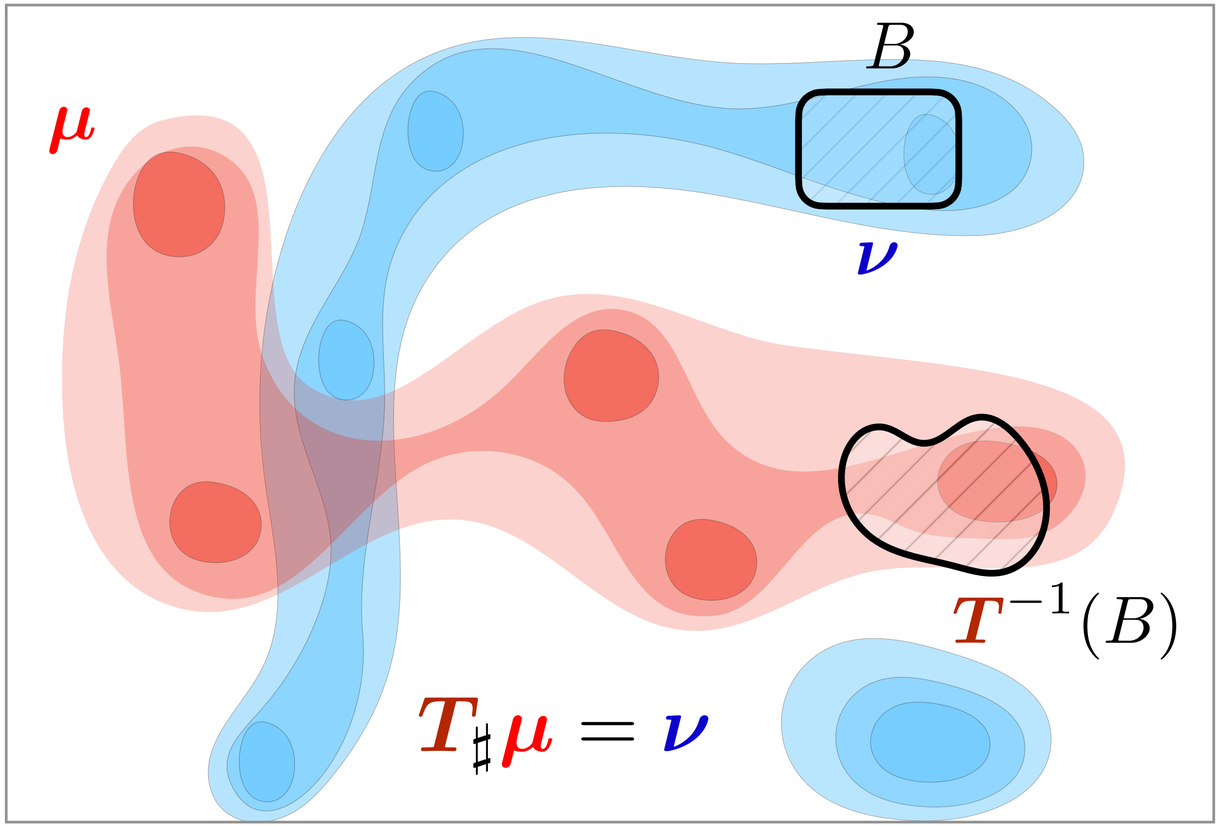

Caffarelli

$$T_\sharp \mu = \nu$$

$$\Updownarrow$$

$$\forall B \subset \mathbb{R}^d, \nu(B) = \mu(T^{-1}(B))$$

Monge Maps




Monge


Brenier


Gangbo


McCann

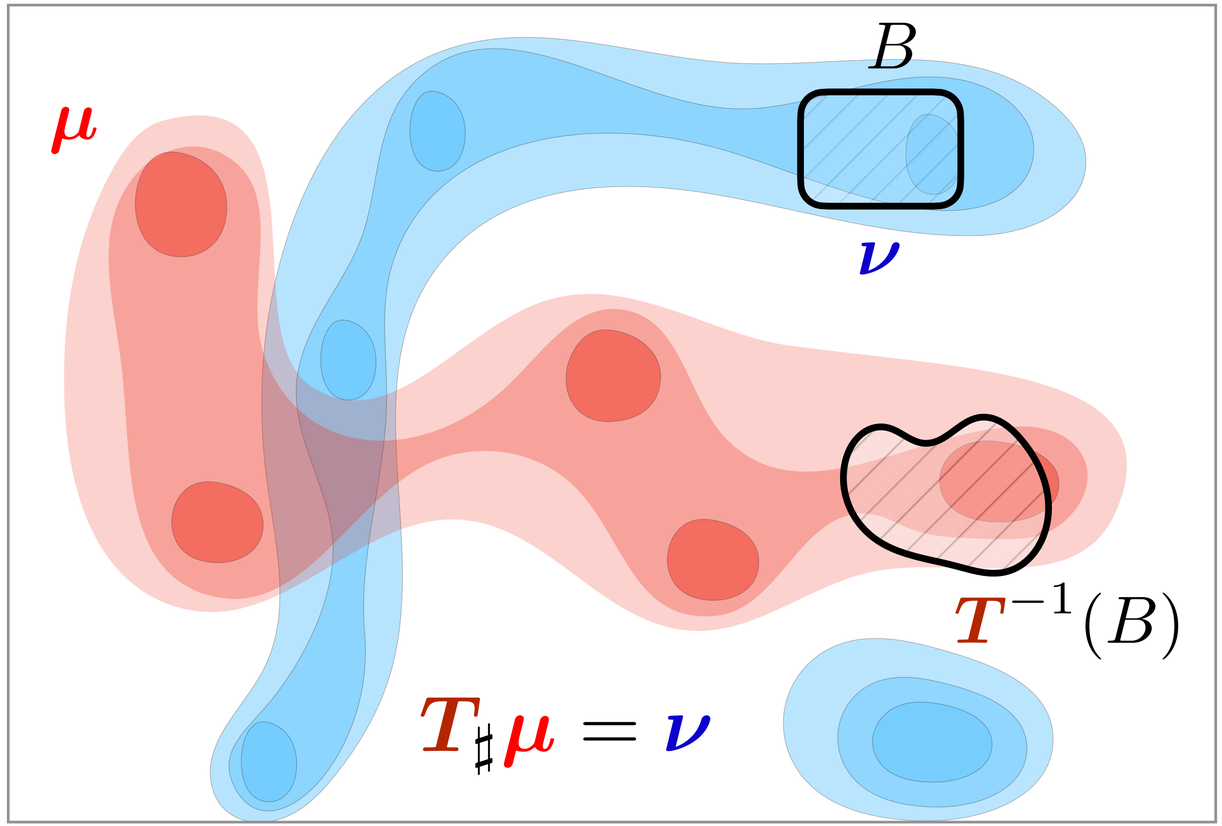

Caffarelli

$$T_\sharp \mu = \nu$$

$$\Updownarrow$$

$$\forall B \subset \mathbb{R}^d, \nu(B) = \mu(T^{-1}(B))$$

Monge Maps




Monge


Brenier


Gangbo


McCann


Caffarelli

$$T_\sharp \mu = \nu$$

$$\Updownarrow$$

$$\forall B \subset \mathbb{R}^d, \nu(B) = \mu(T^{-1}(B))$$

$$T^\star = \underset{\substack{T:\mathbb{R}^p \to \mathbb{R}^p \\ T_\sharp \mu = \nu}}{\operatorname{argmin}} \int c(\mathbf{x}, T(\mathbf{x})) \mathrm{d}\mu(x)$$

Monge and Kantorovich Problems

Monge

$$T^\star = \underset{\substack{T:\mathbb{R}^p \to \mathbb{R}^p \\ T_\sharp \mu = \nu}}{\operatorname{argmin}} \int c(\mathbf{x}, T(\mathbf{x}))\mathrm{d}\mu(x)$$

Kantorovich

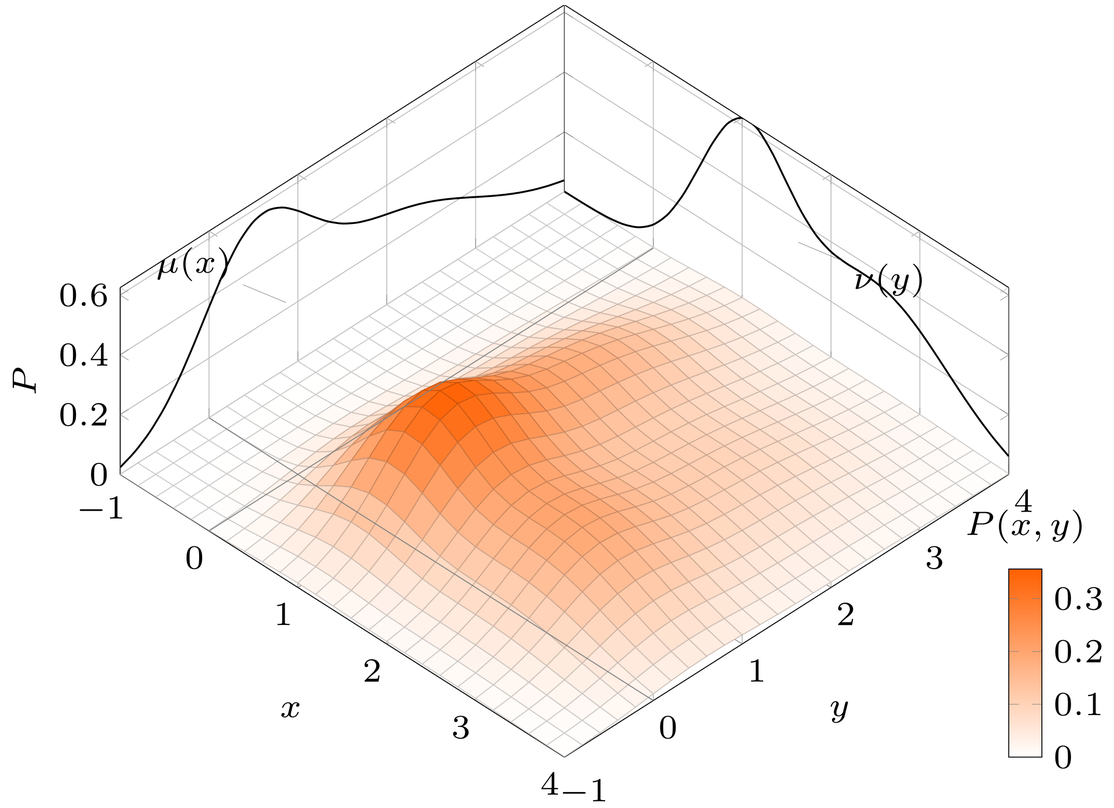
$$\min_{\substack{P \in \mathbb{R}_+^{n \times m} \\ P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}} \langle P, M_{XY} \rangle$$

Kantorovich-Dual

$$\max_{\substack{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m \\ \alpha_i + \beta_j \leq c(x_i, y_j)}} \alpha^T a + \beta^T b$$

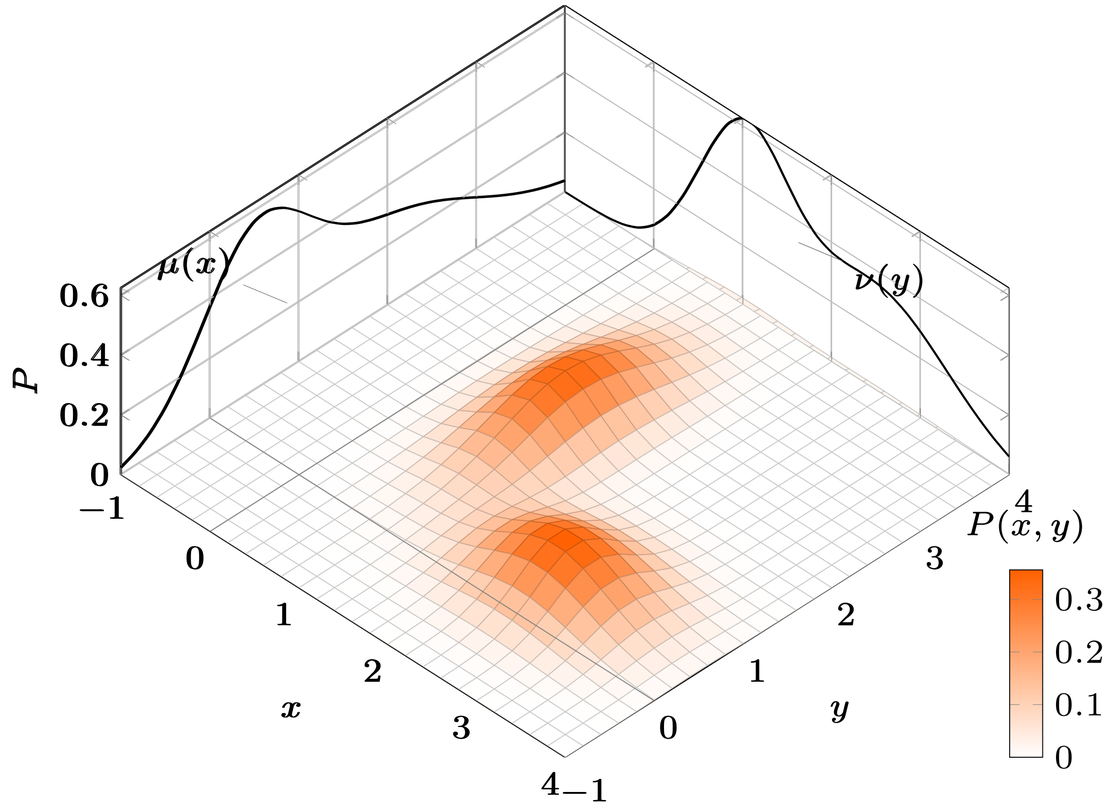
Couplings in Continuous Settings

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) | \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega, \\ \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$



Couplings in Continuous Settings

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) | \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega, \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$



Kantorovich Problem

Def. Given $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$; a cost function $\boldsymbol{c}$ on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{c}(x, y) \boldsymbol{P}(dx, dy).$$

PRIMAL

Kantorovich Problem

Def. Given $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$; a cost function $\boldsymbol{c}$ on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{c}(x, y) \boldsymbol{P}(dx, dy).$$

PRIMAL

$$\sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{c}(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Kantorovich Problem

Def. Given $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$; a cost function $\boldsymbol{c}$ on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{c}(x, y) \boldsymbol{P}(dx, dy).$$

PRIMAL

For two real-valued functions $\boldsymbol{\varphi}, \boldsymbol{\psi}$ on Ω,

$$(\boldsymbol{\varphi} \oplus \boldsymbol{\psi})(x, y) \stackrel{\text{def}}{=} \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y)$$

Kantorovich Problem

Def. Given $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$; a cost function $\boldsymbol{c}$ on $\Omega \times \Omega$, the Kantorovich problem is

$$\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{c}(x, y) \boldsymbol{P}(dx, dy).$$

PRIMAL

$$\sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi} \oplus \boldsymbol{\psi} \leq \boldsymbol{c}}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Deriving Kantorovich Duality

$$\text{Let } \boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}, \text{ and } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

$$\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} = ?$$

Deriving Kantorovich Duality

Let $\boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}$, and $\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$.

$$\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} = ?$$

$$\iint \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{P}(\boldsymbol{x}, y) + \iint \boldsymbol{\psi}(\boldsymbol{y}) d\boldsymbol{P}(x, \boldsymbol{y})$$

Deriving Kantorovich Duality

$$\text{Let } \boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}, \text{ and } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

$$\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} = ?$$

$$\iint \boldsymbol{\varphi}(\boldsymbol{x}) d\boldsymbol{P}(\boldsymbol{x}, y) + \iint \boldsymbol{\psi}(\boldsymbol{y}) d\boldsymbol{P}(x, \boldsymbol{y})$$

Deriving Kantorovich Duality

$$\text{Let } \boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}, \text{ and } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

$$\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} = \mathbf{0}$$

Deriving Kantorovich Duality

$$\text{Let } \boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}, \text{ and } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

$$\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} = \mathbf{0}$$

$$\text{Let } \boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}, \text{ and } \boldsymbol{P} \in \mathcal{P}_+(\Omega^2).$$

Deriving Kantorovich Duality

$$\text{Let } \boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}, \text{ and } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

$$\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} = \mathbf{0}$$

$$\text{Let } \boldsymbol{\varphi}, \boldsymbol{\psi} : \Omega \to \mathbb{R}, \text{ and } \boldsymbol{P} \in \mathcal{P}_+(\Omega^2).$$

$$\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} =$$

$$\int \boldsymbol{\varphi} d \underbrace{(\boldsymbol{\mu} - \boldsymbol{P}_{\boldsymbol{X}})}_{\neq 0} + \int \boldsymbol{\psi} d \underbrace{(\boldsymbol{\nu} - \boldsymbol{P}_{\boldsymbol{Y}})}_{\neq 0}$$

$$\text{and / or}$$

Deriving Kantorovich Duality

$$\iota_\Pi(\boldsymbol{P}) = \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \left[\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} \right]$$

$$= \begin{cases} 0 & \text{if } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}), \\ +\infty & \text{otherwise.} \end{cases}$$

Deriving Kantorovich Duality

$$\iota_\Pi(\boldsymbol{P}) = \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \left[\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} \right]$$

$$= \begin{cases} 0 & \text{if } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}), \\ +\infty & \text{otherwise.} \end{cases}$$

$$\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \iint \boldsymbol{c}\, d\boldsymbol{P}$$

Deriving Kantorovich Duality

$$\iota_\Pi(\boldsymbol{P}) = \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \left[\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} \right]$$

$$= \begin{cases} 0 & \text{if } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}), \\ +\infty & \text{otherwise.} \end{cases}$$

$$\cancel{\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})}} \iint \boldsymbol{c}\, d\boldsymbol{P}$$

$$\inf_{\boldsymbol{P} \in \mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \boxed{\iota_\Pi(\boldsymbol{P})}$$

Deriving Kantorovich Duality

$$\iota_\Pi(\boldsymbol{P}) = \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \left[\int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P} \right]$$

$$= \begin{cases} 0 & \text{if } \boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu}), \\ +\infty & \text{otherwise.} \end{cases}$$

$$\inf_{\boldsymbol{P} \in \mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_\Pi(\boldsymbol{P})$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P}$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \iint \boldsymbol{c}\, d\boldsymbol{P} + \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu} - \iint \boldsymbol{\varphi} \oplus \boldsymbol{\psi} d\boldsymbol{P}$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \iint \boldsymbol{c}\, d\boldsymbol{P} - \iint \boldsymbol{\varphi}\oplus\boldsymbol{\psi} d\boldsymbol{P} + \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \iint (\boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi}) d\boldsymbol{P} \qquad + \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint (\boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi}) d\boldsymbol{P} \qquad + \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint (\boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi}) d\boldsymbol{P} \qquad + \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint (\boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi}) d\boldsymbol{P} \qquad + \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}$$

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega)} \iint (\boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi}) d\boldsymbol{P} = \begin{cases} 0 & \text{if } \boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi} \geq 0. \\ -\infty & \text{otherwise} \end{cases}$$

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\sup_{\boldsymbol{\varphi},\boldsymbol{\psi}} \inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint (\boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi}) d\boldsymbol{P} \qquad + \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}$$

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega)} \iint (\boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi}) d\boldsymbol{P} = \begin{cases} 0 & \text{if } \boldsymbol{c} - \boldsymbol{\varphi} \oplus \boldsymbol{\psi} \geq 0. \\ -\infty & \text{otherwise} \end{cases}$$

$$\sup_{\boldsymbol{\varphi}\oplus\boldsymbol{\psi}\leq\boldsymbol{c}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Deriving Kantorovich Duality

$$\inf_{\boldsymbol{P}\in\mathcal{P}_+(\Omega^2)} \iint \boldsymbol{c}\, d\boldsymbol{P} + \iota_{\Pi}(\boldsymbol{P})$$

$$\sup_{\boldsymbol{\varphi}\oplus\boldsymbol{\psi}\leq\boldsymbol{c}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Prop: Primal-dual relationship:

$$\boldsymbol{P}^\star(\boldsymbol{x_0}, \boldsymbol{y_0}) > 0 \Leftrightarrow \boldsymbol{\varphi}^\star(\boldsymbol{x_0}) + \boldsymbol{\psi}^\star(\boldsymbol{y_0}) = \boldsymbol{c}(\boldsymbol{x_0}, \boldsymbol{y_0}).$$

Wasserstein Distances

Let $p \geq 1$. Let $\boldsymbol{c}(x, y) := \boldsymbol{D}^p(x, y)$, a metric.

Def. The p-**Wasserstein distance** between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$ is

$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \left(\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy) \right)^{1/p}.$$

Wasserstein Distances

Let $p \geq 1$. Let $\boldsymbol{c}(x, y) := \boldsymbol{D}^p(x, y)$, a metric.

Def. The p-**Wasserstein distance** between $\boldsymbol{\mu}, \boldsymbol{\nu}$ in $\mathcal{P}(\Omega)$ is

$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \left(\inf_{\boldsymbol{P} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \iint \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})^p \boldsymbol{P}(dx, dy) \right)^{1/p}.$$

Kantorovich Duality

$$W_c(\mu, \nu) = \sup_{\varphi \oplus \psi \leq c} \int \varphi d\mu + \int \psi d\nu.$$

DUAL

- Dual is computationally easier: easier to study 2 functions than a coupling.
- c-transforms: useful machinery to consider only **two** instead of **one** dual potential, For instance, when p = 1

$$W_1(\mu, \nu) = \sup_{\varphi \text{ 1-Lipschitz}} \int \varphi(d\mu - d\nu).$$

W1

c transforms

$$W_c(\mu, \nu) = \sup_{\varphi \oplus \psi \leq c} \int \varphi d\mu + \int \psi d\nu.$$

DUAL

Suppose we fix function φ, can we propose a “good” function for ψ?

c transforms

$$W_c(\mu, \nu) = \sup_{\varphi \oplus \psi \leq c} \int \varphi d\mu + \int \psi d\nu.$$

DUAL

Suppose we fix function φ, can we propose a “good” function for ψ?

$$\varphi(x) + \psi(y) \leq c(x, y)$$

c transforms

$$W_c(\mu,\nu) = \sup_{\varphi\oplus\psi\leq c} \int \varphi d\mu + \int \psi d\nu.$$

DUAL

Suppose we fix function φ, can we propose a “good” function for ψ?

$$\varphi(x) + \psi(y) \leq c(x,y)$$

$$\psi(y) \leq c(x,y) - \varphi(x)$$

c transforms

$$W_c(\mu, \nu) = \sup_{\varphi \oplus \psi \leq c} \int \varphi d\mu + \int \psi d\nu.$$

DUAL

Suppose we fix function φ, can we propose a "good" function for ψ?

$$\varphi(x) + \psi(y) \leq c(x, y)$$

$$\psi(y) \leq c(x, y) - \varphi(x)$$

$$\psi(y) \leq \inf_x c(x, y) - \varphi(x)$$

c transforms

$$W_{\boldsymbol{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi} \oplus \boldsymbol{\psi} \leq \boldsymbol{c}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

For given $\boldsymbol{\varphi}$, cannot get a better $\boldsymbol{\psi}$ than

$$\overline{\boldsymbol{\varphi}}(\boldsymbol{y}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{x}} \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x}).$$

c transforms

$$W_{\boldsymbol{c}}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi} \oplus \boldsymbol{\psi} \leq \boldsymbol{c}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

For given $\boldsymbol{\varphi}$, cannot get a better $\boldsymbol{\psi}$ than

$$\overline{\boldsymbol{\varphi}}(\boldsymbol{y}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{x}} \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x}).$$

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \overline{\boldsymbol{\varphi}} d\boldsymbol{\nu}.$$

SEMI-DUAL

c transforms

$$\overline{\varphi}(\boldsymbol{y}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{x}} \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x}).$$

$$\overline{\boldsymbol{\psi}}(\boldsymbol{x}) = \inf_{\boldsymbol{y}} \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\psi}(\boldsymbol{y}).$$

c transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x c(x, y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y c(x, y) - \psi(y).$$

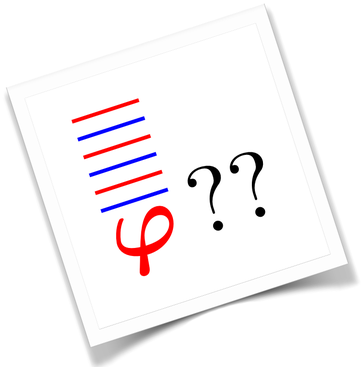
$$W_p^p(\mu, \nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

c transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y c(x,y) - \psi(y).$$

$$W_p^p(\mu,\nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$



c transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y c(x,y) - \psi(y).$$

$$W_p^p(\mu,\nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

For all φ, we have $\overline{\overline{\overline{\varphi}}} = \overline{\varphi}$

c transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y c(x,y) - \psi(y).$$

$$W_p^p(\mu,\nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

For all φ, we have $\overline{\overline{\overline{\varphi}}} = \overline{\varphi}$

φ is c-concave if $\exists \phi : \varphi = \overline{\phi}$

φ is c-concave $\Leftrightarrow \overline{\overline{\varphi}} = \varphi$

c transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x c(x, y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y c(x, y) - \psi(y).$$

$$W_p^p(\mu, \nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

$$W_p^p(\mu, \nu) = \sup_{\varphi \text{ is } c\text{-concave}} \int \varphi d\mu + \int \overline{\varphi} d\nu.$$

c transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x, y) - \varphi(x)$ is 1-Lipschitz.
$\overline{\varphi}_x(y) - \overline{\varphi}_x(y') = D(x, y) - D(x, y') \leq D(y, y')$

c transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x, y) - \varphi(x)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

c transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \overset{\text{def}}{=} D(x, y) - \varphi(x)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \leq D(x, y)$

c transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \overset{\text{def}}{=} D(x, y) - \varphi(x)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \leq D(x, y)$
$\Rightarrow -\overline{\varphi}(x) \leq D(x, y) - \overline{\varphi}(y)$

c transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \le D(x,y)$

$\Rightarrow -\overline{\varphi}(x) \le D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y)$

c transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \overset{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \le D(x,y)$

$\Rightarrow -\overline{\varphi}(x) \le D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y) \le -\overline{\varphi}(x)$

c transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \overset{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \le D(x,y)$

$\Rightarrow -\overline{\varphi}(x) \le D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y) \le -\overline{\varphi}(x)$

$\Rightarrow -\overline{\varphi}(x) \le \overline{\overline{\varphi}}(x) \le -\overline{\varphi}(x)$ and $\overline{\varphi}(x) = -\varphi(x)$

C transforms, W_1

$$W_1(\mu,\nu) = \sup_{\varphi \text{ is } D\text{-concave}} \int \varphi d\mu + \int \overline{\varphi} d\nu.$$

SEMI-DUAL

Prop. If $c = D$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

$$W_1(\mu,\nu) = \sup_{\varphi \text{ 1-Lipschitz}} \int \varphi(d\mu - d\nu).$$

W1

Roadmap to Solve Monge Problem

Monge

$$T^\star = \underset{\substack{T:\mathbb{R}^p \to \mathbb{R}^p \\ T_\sharp \mu = \nu}}{\operatorname{argmin}} \int c(\mathbf{x}, T(\mathbf{x}))\mathrm{d}\mu(x)$$

Kantorovich

$$P^\star = \arg \inf_{P \in \Pi(\mu,\nu)} \iint c(x,y) P(\mathrm{d}x, \mathrm{d}y).$$

Kantorovich-Dual

$$f^\star, g^\star = \arg \sup_{\substack{f,g:\mathbb{R}^d \to \mathbb{R} \\ f(x)+g(y) \le c(x,y)}} \int_{\mathbb{R}^d} f \,\mathrm{d}\mu + \int_{\mathbb{R}^d} g \,\mathrm{d}\nu \,.$$

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

$$P^\star(x_0, y_0) > 0 \Leftrightarrow f^\star(x_0) + g^\star(y_0) = c(x_0, y_0).$$

Primal-dual relationship (KKT)

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

$$P^\star(x_0,y_0) > 0 \Leftrightarrow f^\star(x_0) + g^\star(y_0) = c(x_0,y_0).$$
$$\Leftrightarrow f^\star(x_0) + \overline{f^\star}(y_0) = c(x_0,y_0).$$

Primal-dual relationship (KKT)

c-transforms

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

$$P^\star(x_0,y_0) > 0 \Leftrightarrow f^\star(x_0) + g^\star(y_0) = c(x_0,y_0).$$
$$\Leftrightarrow f^\star(x_0) + \overline{f^\star}(y_0) = c(x_0,y_0).$$
$$\Leftrightarrow \overline{f^\star}(y_0) = c(x_0,y_0) - f^\star(x_0).$$

Primal-dual relationship (KKT)

c-transforms

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

$$P^\star(x_0,y_0) > 0 \Leftrightarrow f^\star(x_0) + g^\star(y_0) = c(x_0,y_0).$$ Primal-dual relationship (KKT)

$$\Leftrightarrow f^\star(x_0) + \overline{f^\star}(y_0) = c(x_0,y_0).$$ **c**-transforms

$$\Leftrightarrow \overline{f^\star}(y_0) = c(x_0,y_0) - f^\star(x_0).$$

$$\text{yet, } \overline{f^\star}(y_0) = \inf_z c(z,y_0) - f^\star(z).$$ definition

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

$$P^\star(x_0,y_0) > 0 \Leftrightarrow f^\star(x_0) + g^\star(y_0) = c(x_0,y_0).$$ Primal-dual relationship (KKT)

$$\Leftrightarrow f^\star(x_0) + \overline{f^\star}(y_0) = c(x_0,y_0).$$ **c**-transforms

$$\Leftrightarrow \overline{f^\star}(y_0) = c(x_0,y_0) - f^\star(x_0).$$

$$\text{yet, } \overline{f^\star}(y_0) = \inf_z c(z,y_0) - f^\star(z).$$ definition

$$\Rightarrow x_0 \text{ is minimizer, } \nabla_1 c(x_0,y_0) = \nabla f^\star(x_0)$$

Gangbo-McCann Theorem

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

$P^\star(x_0, y_0) > 0 \Leftrightarrow f^\star(x_0) + g^\star(y_0) = c(x_0, y_0).$ Primal-dual relationship (KKT)

$\Leftrightarrow f^\star(x_0) + \overline{f^\star}(y_0) = c(x_0, y_0).$ **c**-transforms

$\Leftrightarrow \overline{f^\star}(y_0) = c(x_0, y_0) - f^\star(x_0).$

yet, $\overline{f^\star}(y_0) = \inf_z c(z, y_0) - f^\star(z).$ definition

$\Rightarrow x_0$ is minimizer, $\nabla_1 c(x_0, y_0) = \nabla f^\star(x_0)$

therefore $y_0 = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$

Translation Invariant Costs Generated by a Convex Functional

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

Translation Invariant Costs Generated by a Convex Functional

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$c(x,y) = h(x-y), h : \mathbb{R}^d \to \mathbb{R}, \text{convex}$$

Translation Invariant Costs Generated by a Convex Functional

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$c(x,y) = h(x-y), h : \mathbb{R}^d \to \mathbb{R}, \text{convex}$$
$$\nabla_1 c(x,y) = \nabla h(x-y)$$

Translation Invariant Costs Generated by a Convex Functional

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$c(x,y) = h(x-y), h : \mathbb{R}^d \to \mathbb{R}, \text{convex}$$
$$\nabla_1 c(x,y) = \nabla h(x-y)$$
$$\nabla_1 c(x,\cdot) = \nabla h(x-\cdot)$$

Translation Invariant Costs Generated by a Convex Functional

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$c(x,y) = h(x-y), h : \mathbb{R}^d \to \mathbb{R}, \text{convex}$$
$$\nabla_1 c(x,y) = \nabla h(x-y)$$
$$\nabla_1 c(x,\cdot) = \nabla h(x-\cdot)$$
$$\nabla_1 c(x,\cdot)^{-1} = x - (\nabla h)^{-1}(\cdot)$$

Translation Invariant Costs Generated by a Convex Functional

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

$$c(x,y) = h(x-y), h : \mathbb{R}^d \to \mathbb{R}, \text{convex}$$
$$\nabla_1 c(x,y) = \nabla h(x-y)$$
$$\nabla_1 c(x,\cdot) = \nabla h(x-\cdot)$$
$$\nabla_1 c(x,\cdot)^{-1} = x - (\nabla h)^{-1}(\cdot)$$
$$\nabla_1 c(x,\cdot)^{-1} = x - \nabla h^*(\cdot)$$

$$h^*(z) = \sup_x \langle x, z\rangle - h(x)$$

Gangbo-McCann and Brenier Theorems

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

General cost function c

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

Translation invariant cost h

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x)$$

Squared-Euclidean Cost

$$T^\star(x) = x - \nabla f^\star(x) = \nabla(\tfrac{1}{2}\|\cdot\|_2^2 - f^\star)(x)$$

$h = \frac{1}{2}\|\cdot\|_2^2 \to \nabla h = \nabla h^* = \mathrm{Id}$

Gangbo-McCann and Brenier Theorems

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

General cost function c

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

Translation invariant cost h

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x)$$

Squared-Euclidean Cost

$$h = \tfrac{1}{2}\|\cdot\|_2^2 \to \nabla h = \nabla h^* = \mathrm{Id}$$

$$T^\star(x) = x - \nabla f^\star(x) = \nabla(\underbrace{\tfrac{1}{2}\|\cdot\|_2^2 - f^\star}_{\text{can prove this is convex}})(x)$$

Gangbo-McCann and Brenier Theorems

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

General cost function c

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

Translation invariant cost h

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x)$$

Squared-Euclidean Cost

$$T^\star(x) = \nabla \varphi^\star(x)$$

$h = \frac{1}{2}\|\cdot\|_2^2 \to \nabla h = \nabla h^* = \mathrm{Id}$

Gangbo-McCann and Brenier Theorems

c-transform:

$$\overline{f}(y) \stackrel{\text{def}}{=} \inf_x c(x,y) - f(x).$$
$$\overline{g}(x) \stackrel{\text{def}}{=} \inf_y c(x,y) - g(y).$$

f is a **c**-concave function if exists g such that $f = \overline{g}$.

$f^\star$ is **necessarily** **c**-concave

General cost function c

$$T^\star(x) = \nabla_1 c(x,\cdot)^{-1} \circ \nabla f^\star(x)$$

Translation invariant cost h

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x)$$

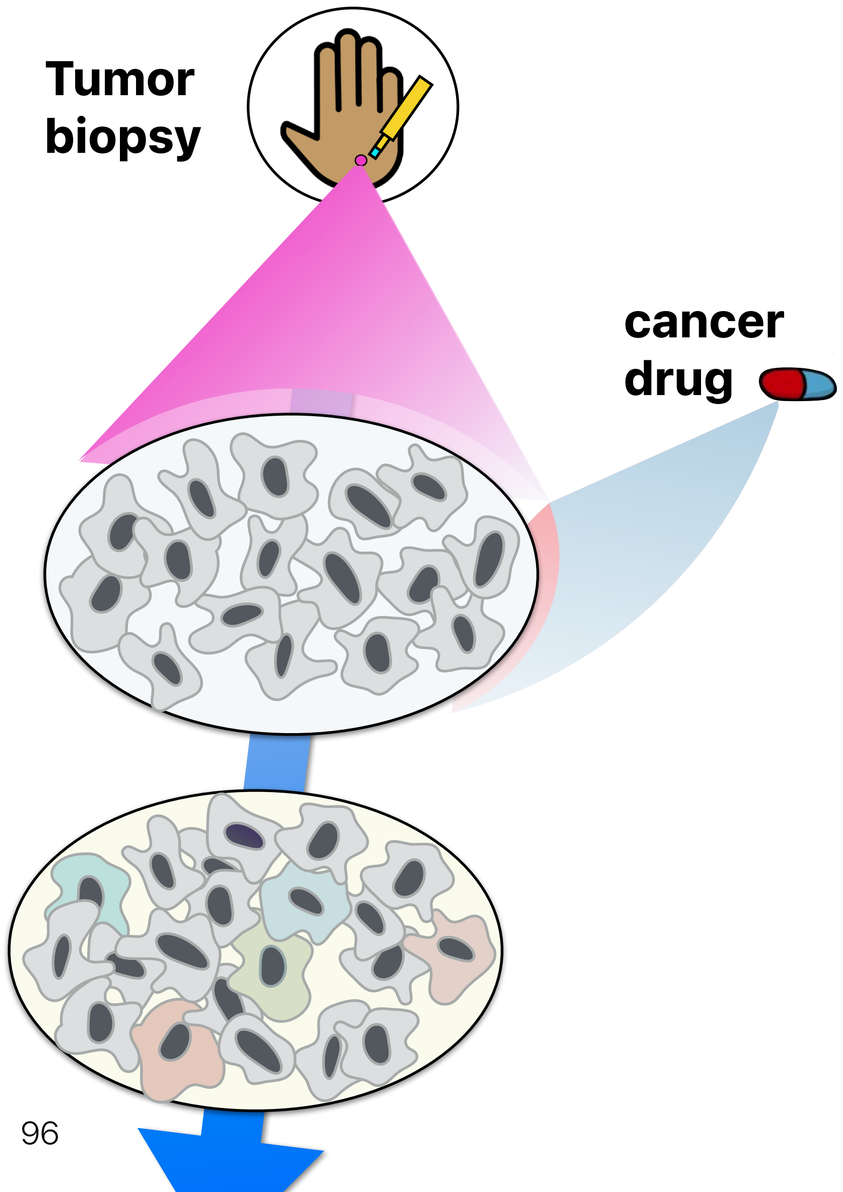
Squared-Euclidean Cost

$$T^\star(x) = \nabla \varphi^\star(x)$$

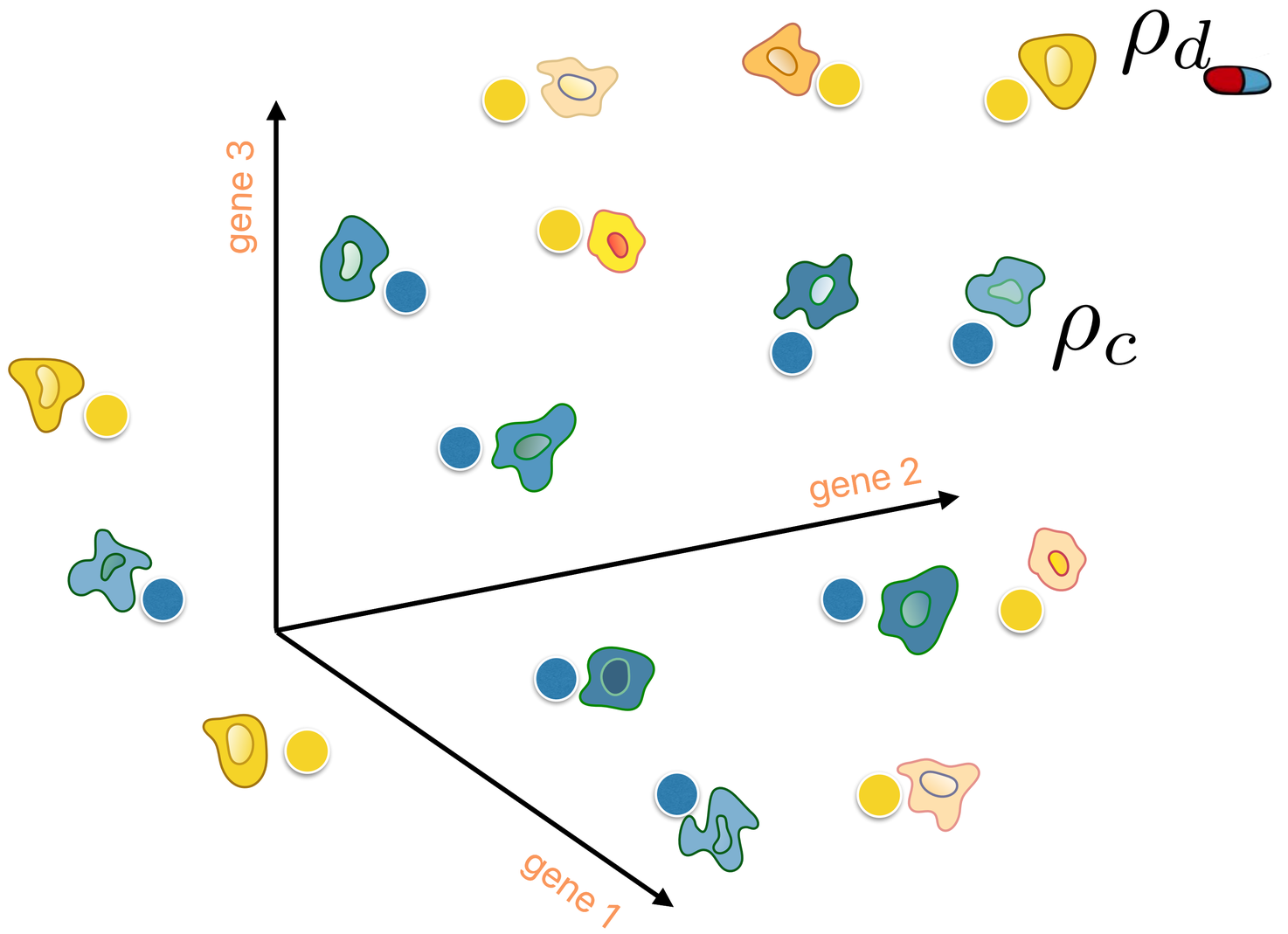
$h = \frac{1}{2}\|\cdot\|_2^2 \to \nabla h = \nabla h^* = \mathrm{Id}$

$$\varphi^\star = \operatorname*{argmin}_{\varphi\,\mathrm{cvx}} \int \varphi \mathrm{d}\mu + \int \varphi^* \mathrm{d}\nu$$

Motivating Example: Personalised Medicine

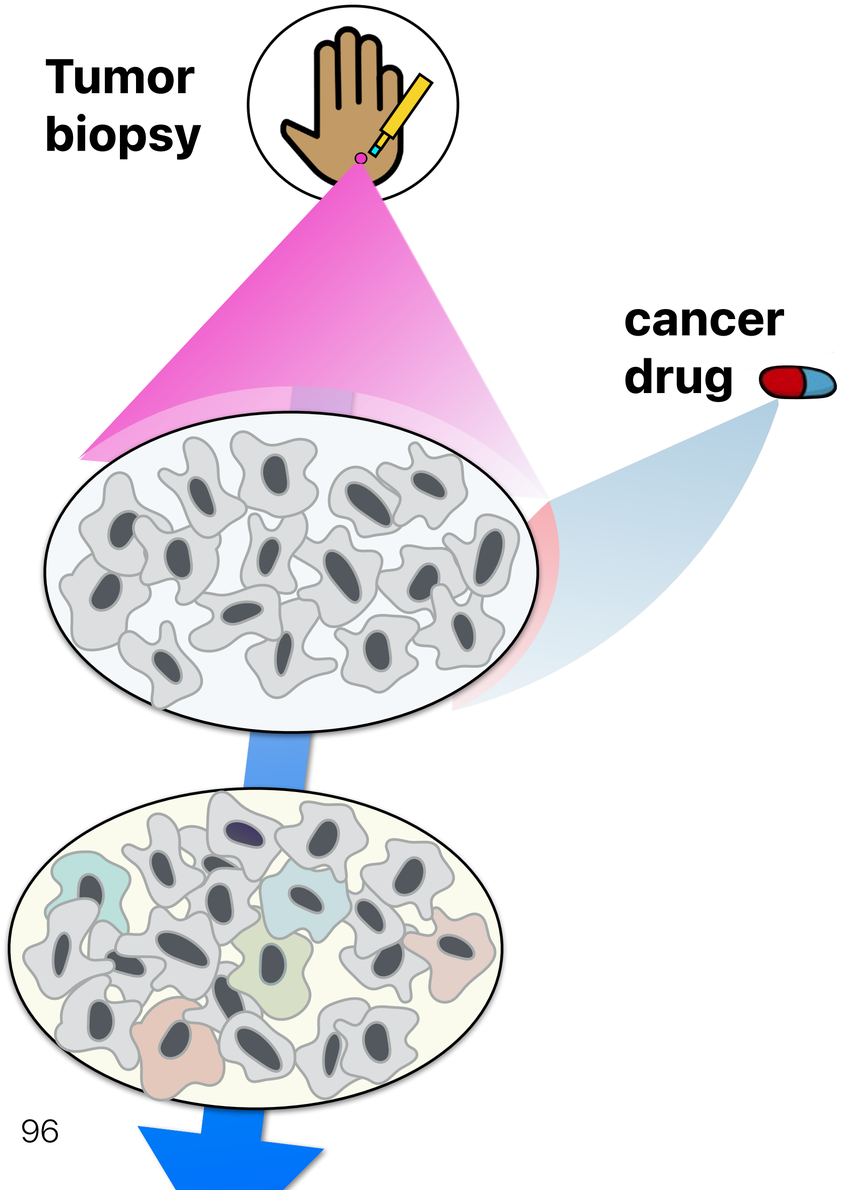


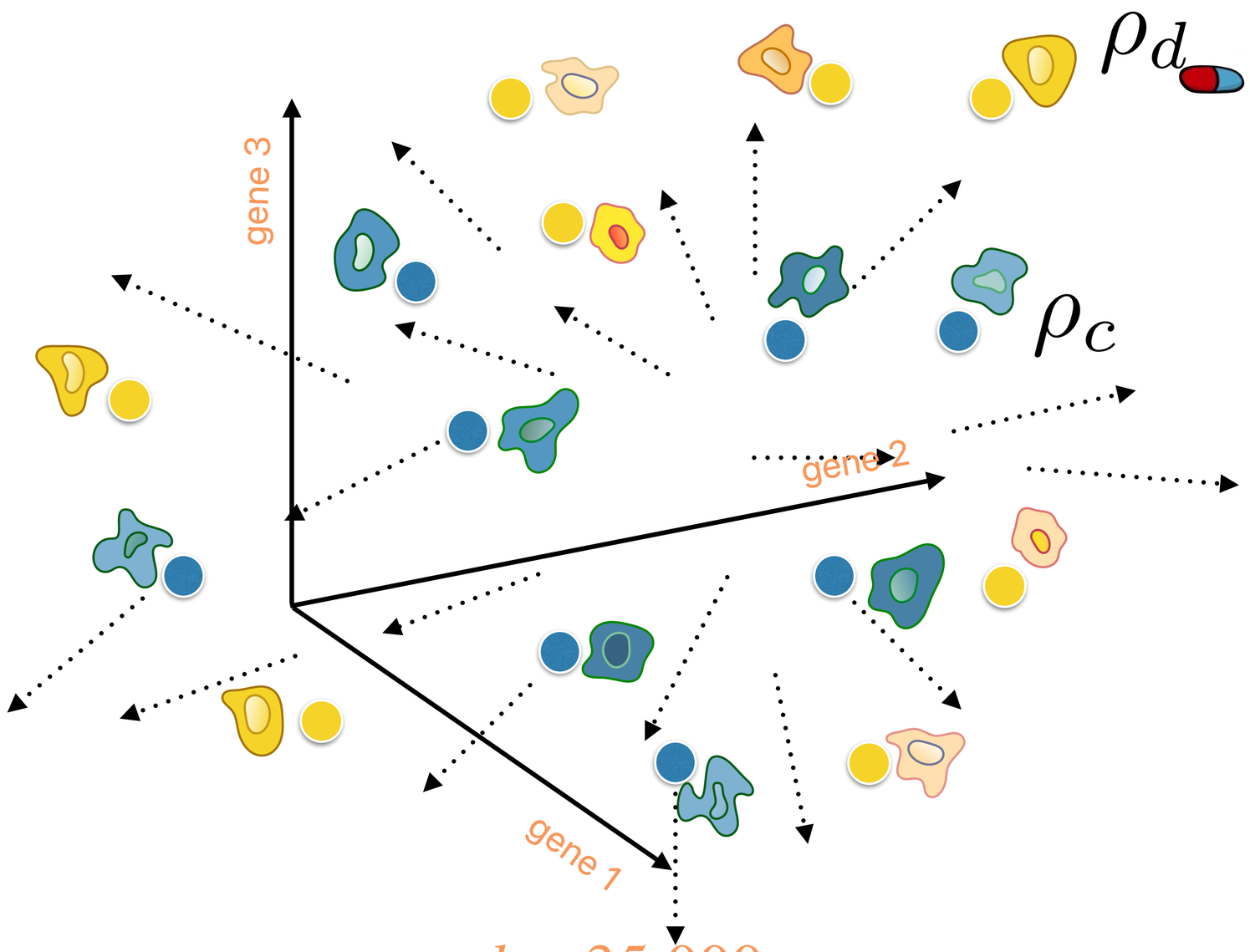
cancer
drug



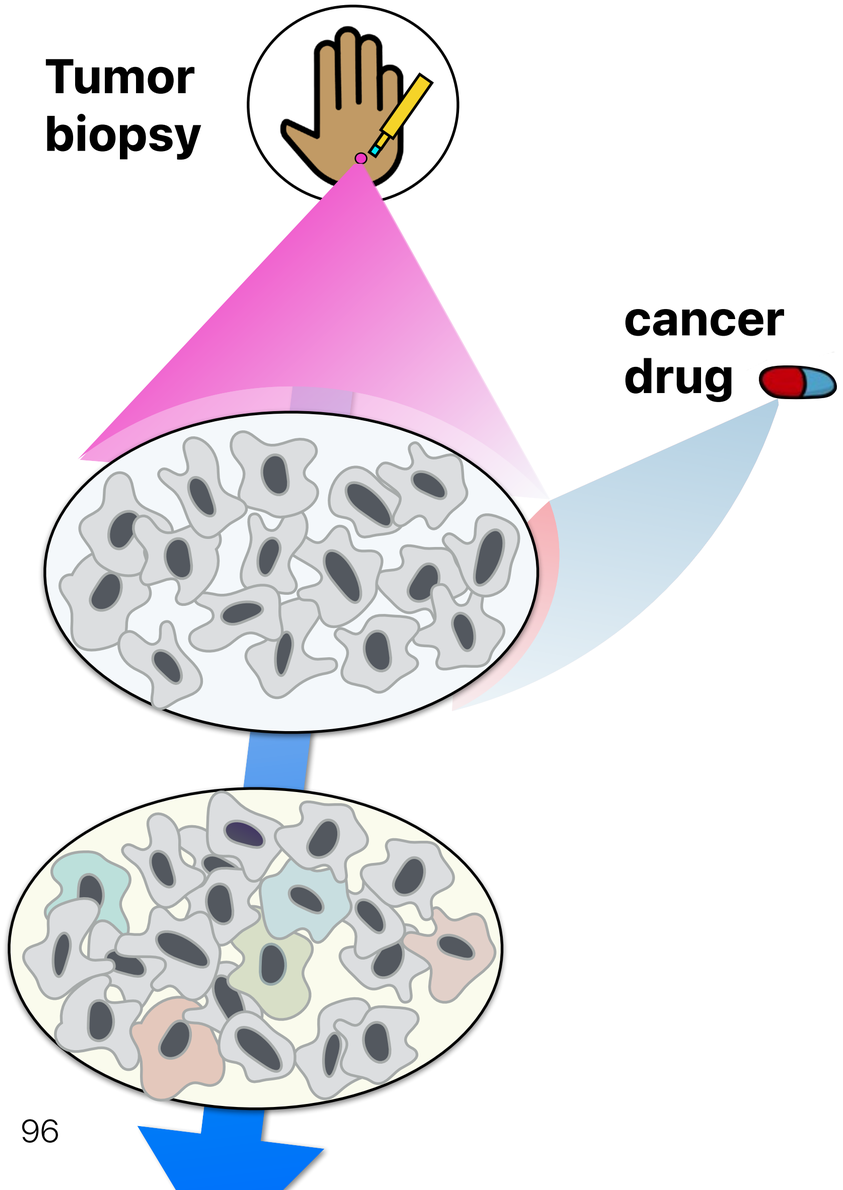
$d = 35{,}000$

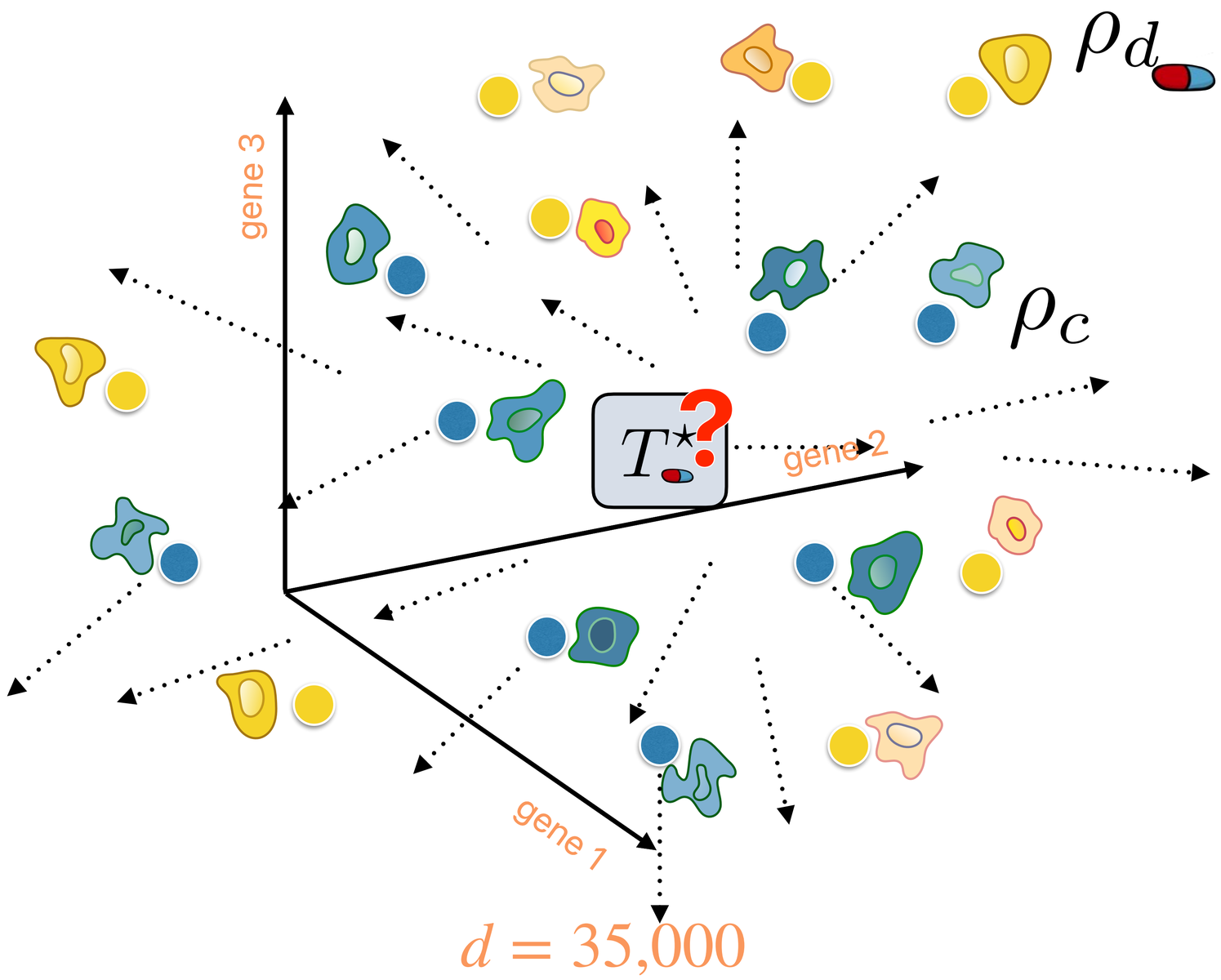
Motivating Example: Personalised Medicine



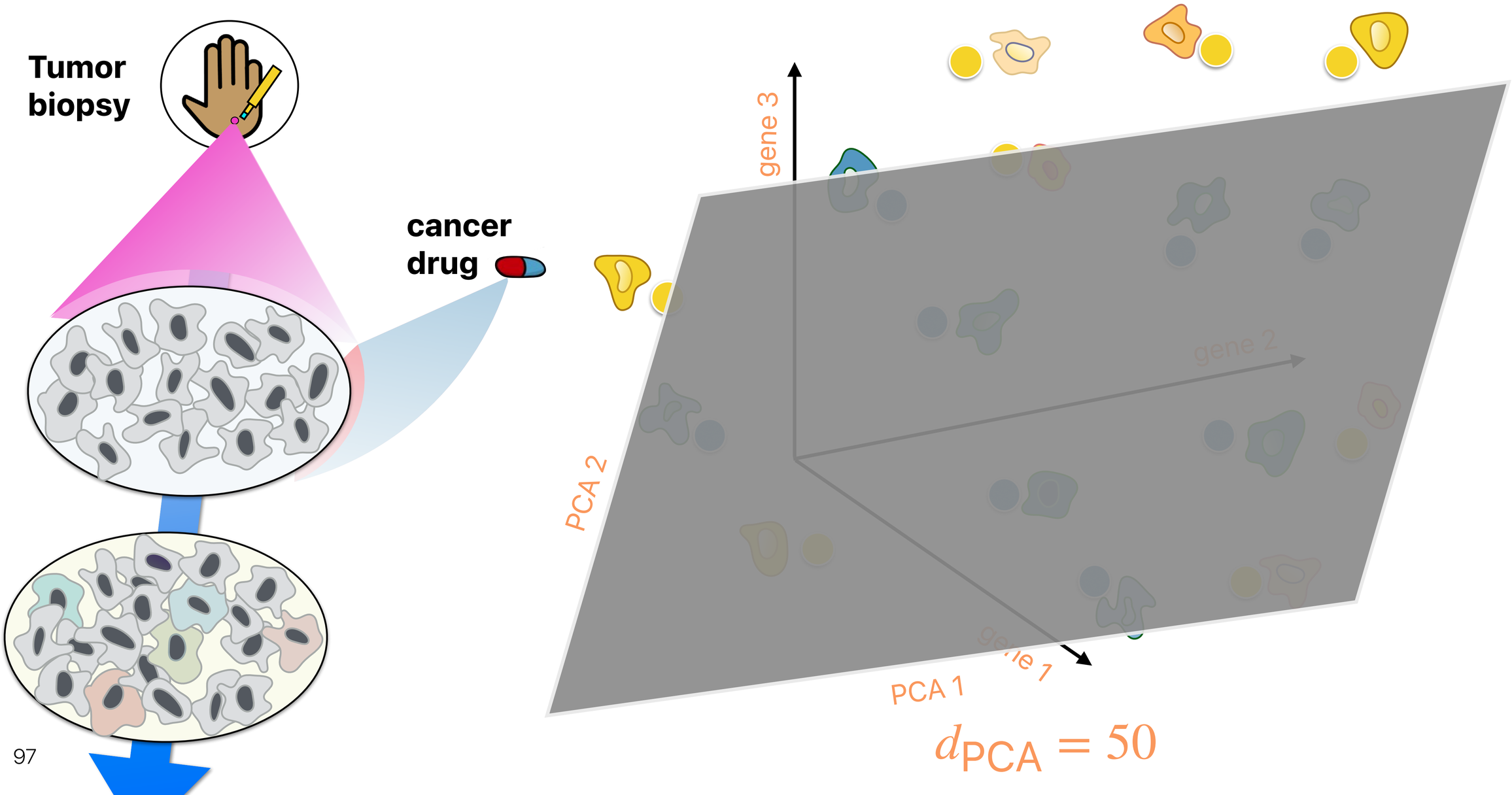


Motivating Example: Personalised Medicine

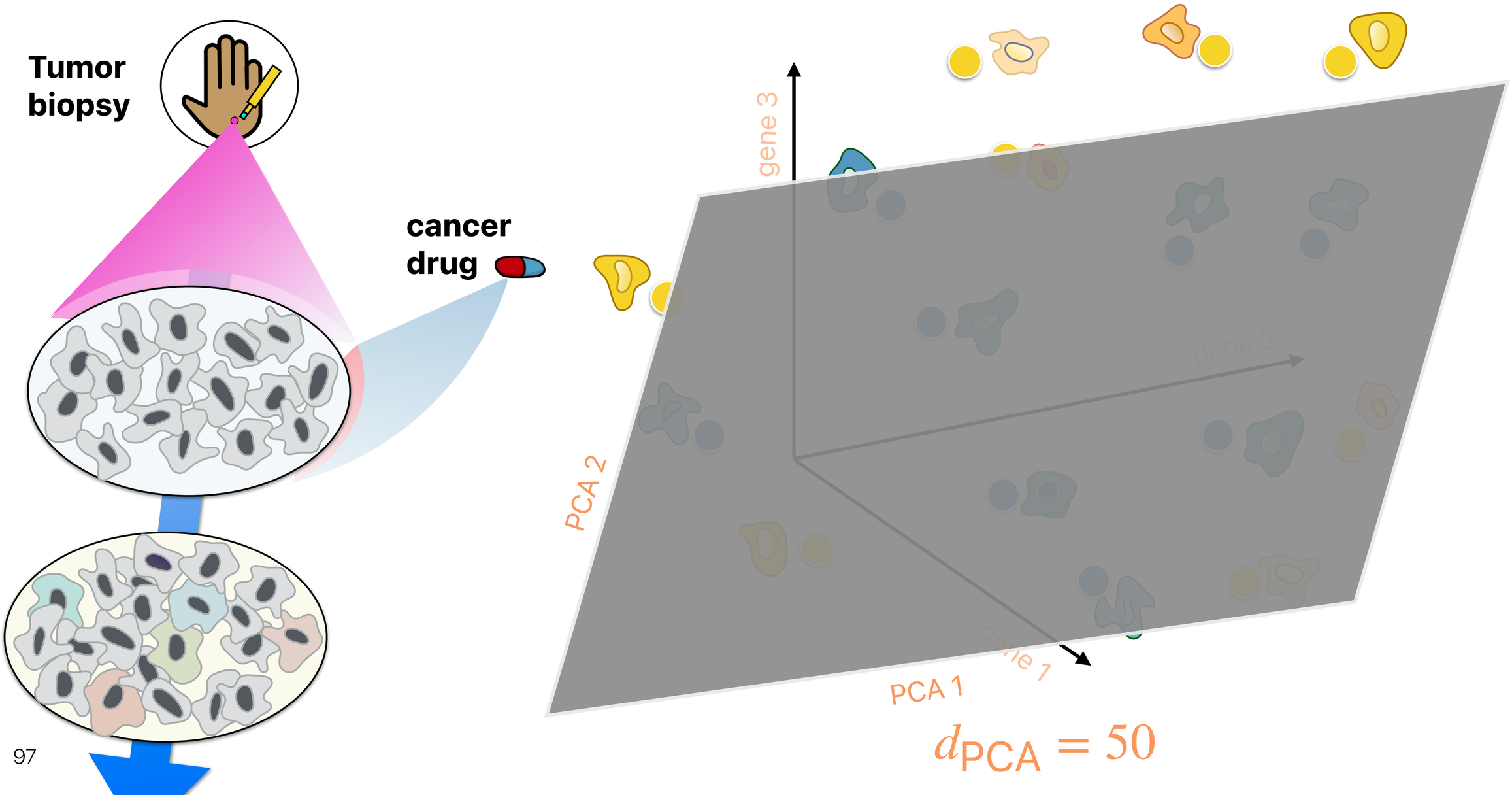




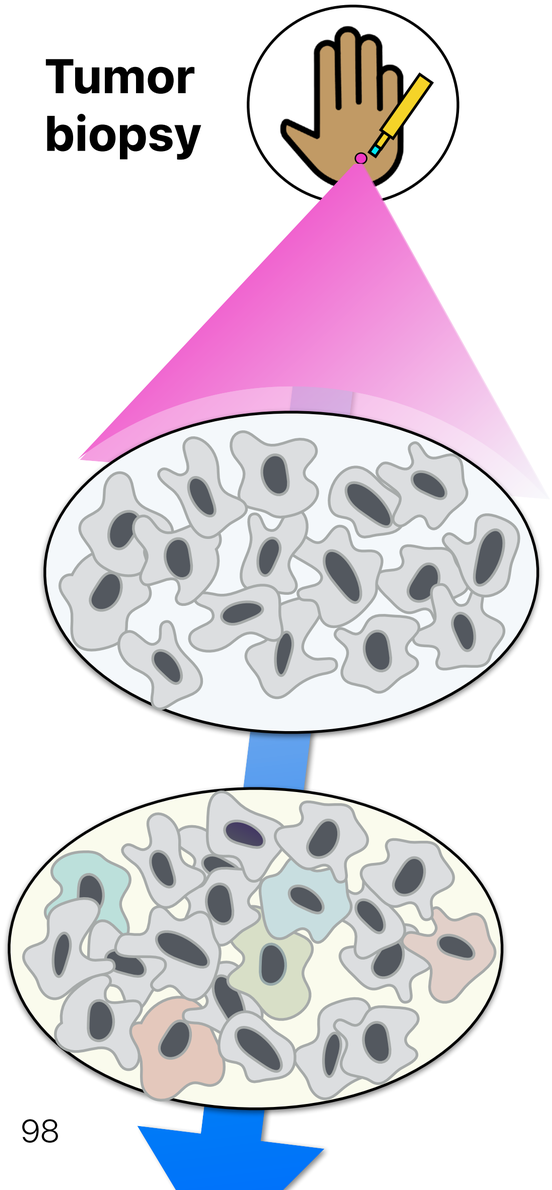
Motivating Example: Personalised Medicine

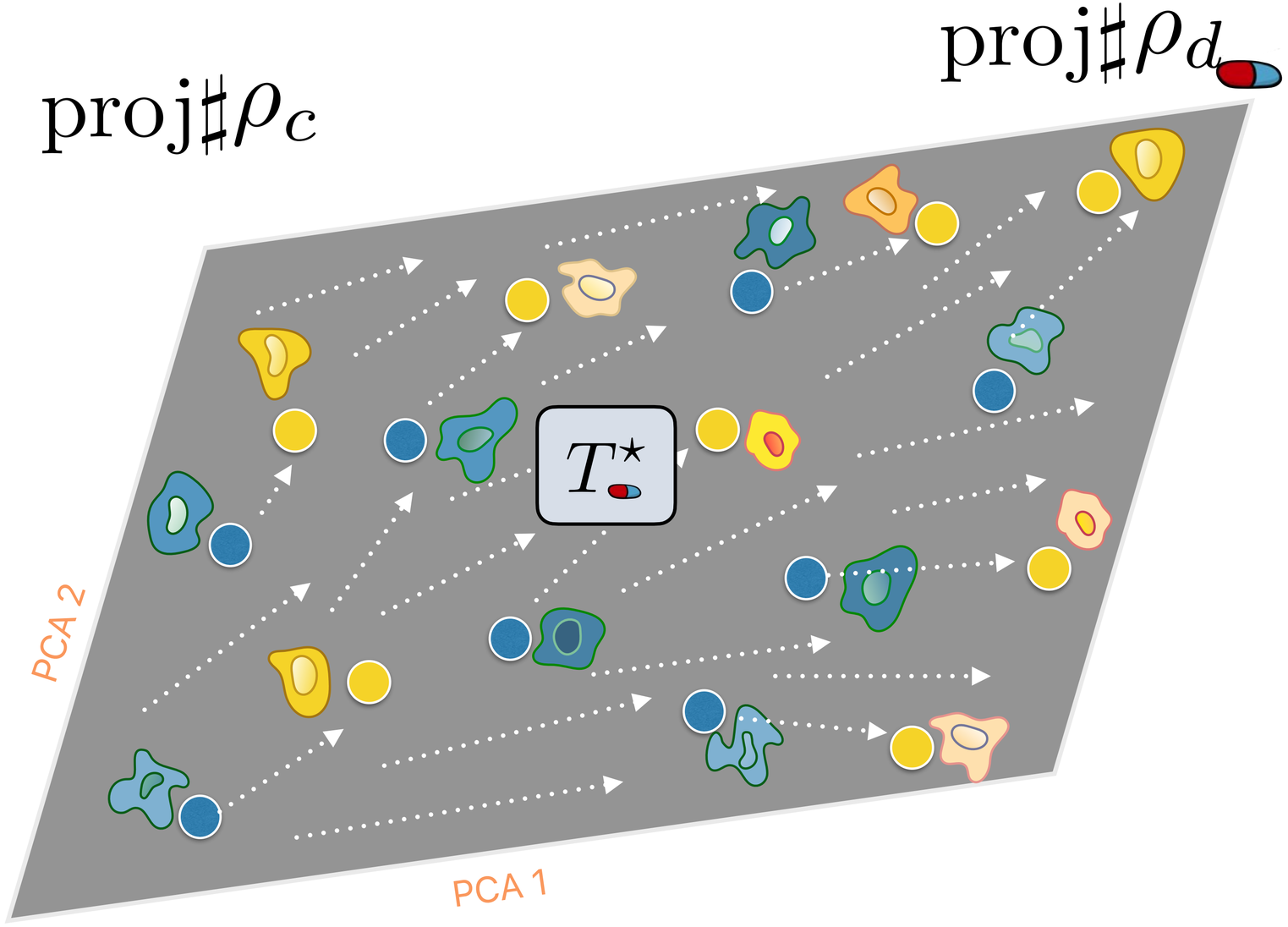


Motivating Example: Personalised Medicine



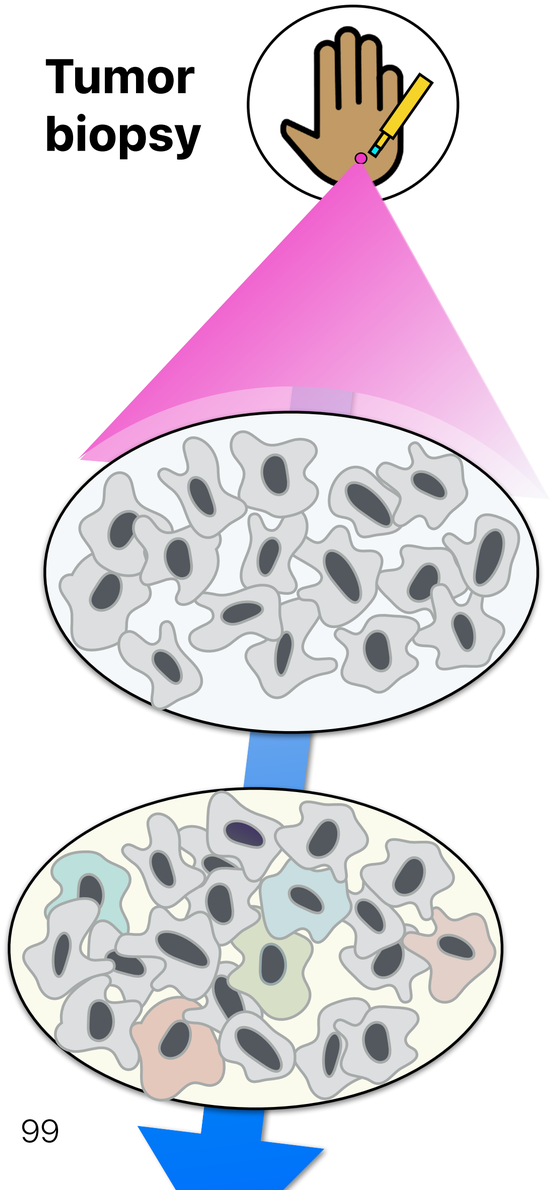
Motivating Example: Personalised Medicine





$d_{\text{PCA}} = 50$

Motivating Example: Personalised Medicine



$\text{proj}\sharp\rho_c$

$\text{proj}\sharp\rho_d$

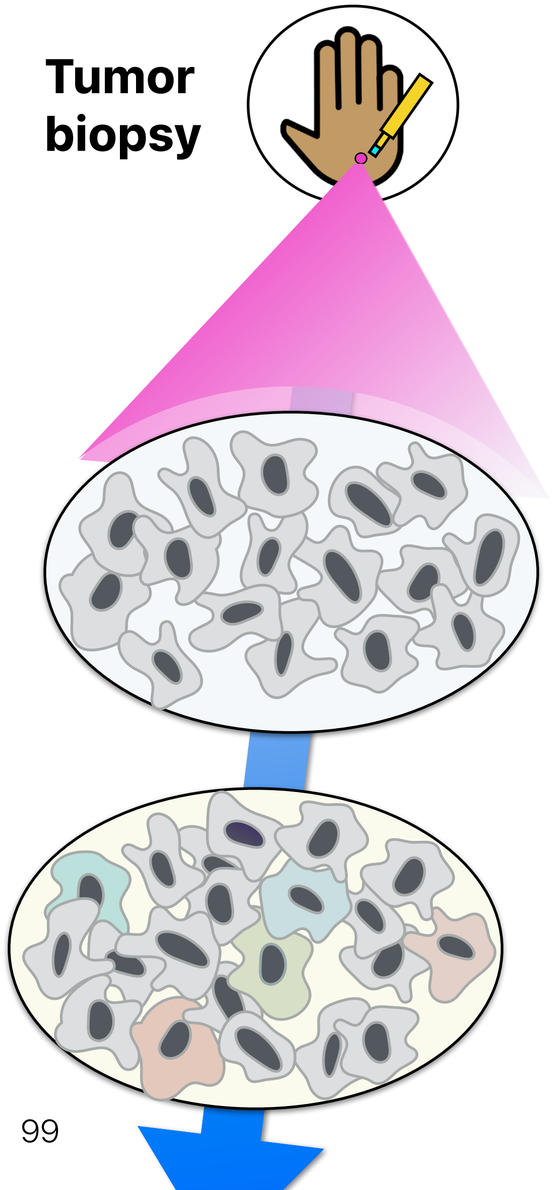
$$T^\star \leftarrow \inf_{\substack{T:\mathbb{R}^d \to \mathbb{R}^d \\ T\sharp\rho_c \approx \rho_d}} \int_{\mathbb{R}^d} \|x - T(x)\|_2^2 d\rho_c(x)$$

PCA 2

PCA 1

$d_{\text{PCA}} = 50$

Motivating Example: Personalised Medicine



$\text{proj}\sharp\rho_c$

$\text{proj}\sharp\rho_d$

$$T^\star \leftarrow \inf_{\substack{T:\mathbb{R}^d \to \mathbb{R}^d \\ T\sharp\rho_c \approx \rho_d}} \int_{\mathbb{R}^d} \|x - T(x)\|_2^2 d\rho_c(x)$$

PCA 2

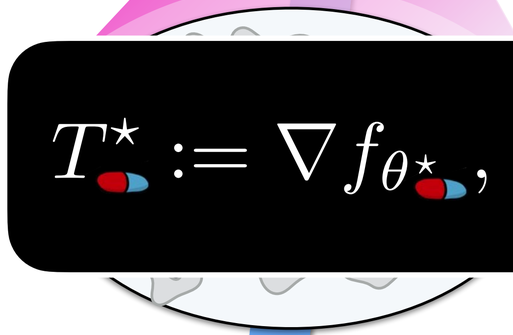
PCA 1

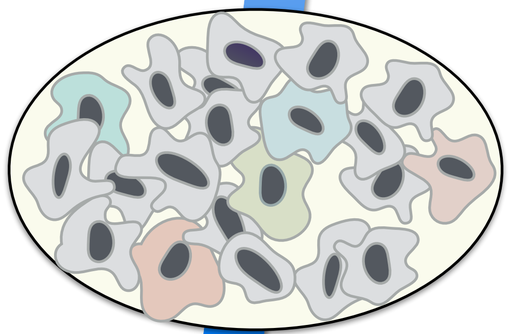
$d_{\text{PCA}} = 50$

Motivating Example: Personalised Medicine

Tumor biopsy

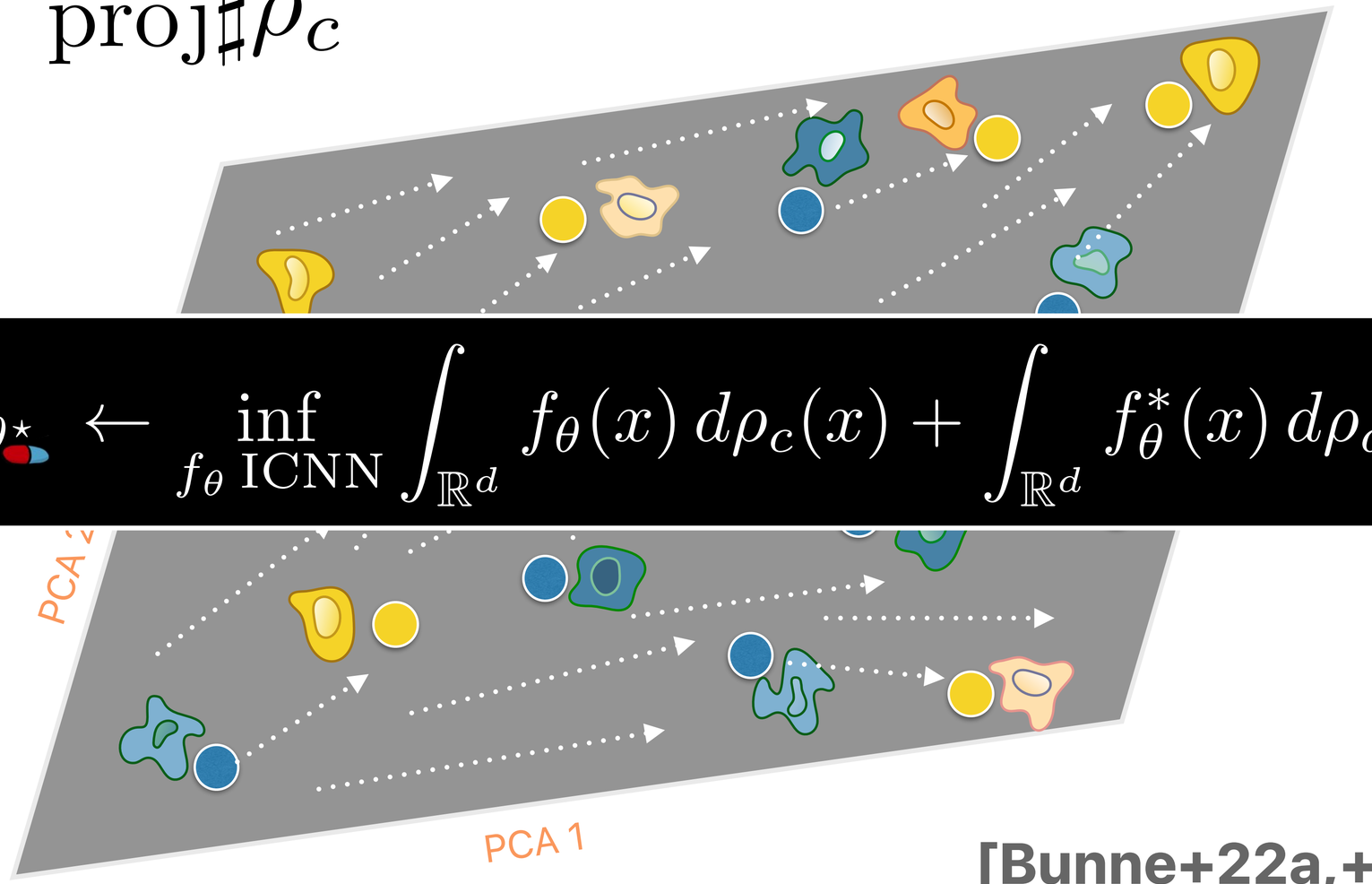






$\text{proj}\sharp\rho_c$

$\text{proj}\sharp\rho_d$



$$T^\star := \nabla f_{\theta^\star}, \text{ where } f_{\theta^\star} \leftarrow \inf_{f_\theta \text{ ICNN}} \int_{\mathbb{R}^d} f_\theta(x)\, d\rho_c(x) + \int_{\mathbb{R}^d} f_\theta^*(x)\, d\rho_d(x)$$

$d_{\text{PCA}} = 50$

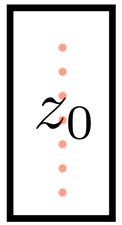
[Bunne+22a,+22b]

Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$

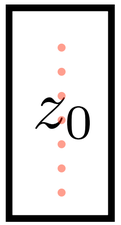
Parameterize Convex Functions: ICNN

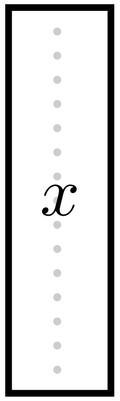
$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



Parameterize Convex Functions: ICNN

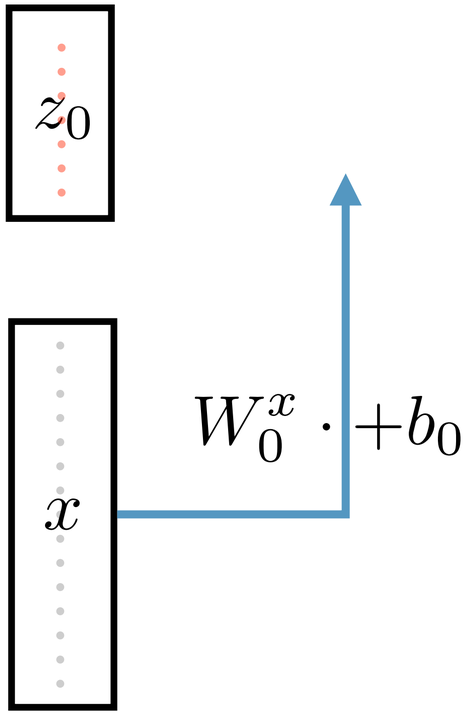
$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$





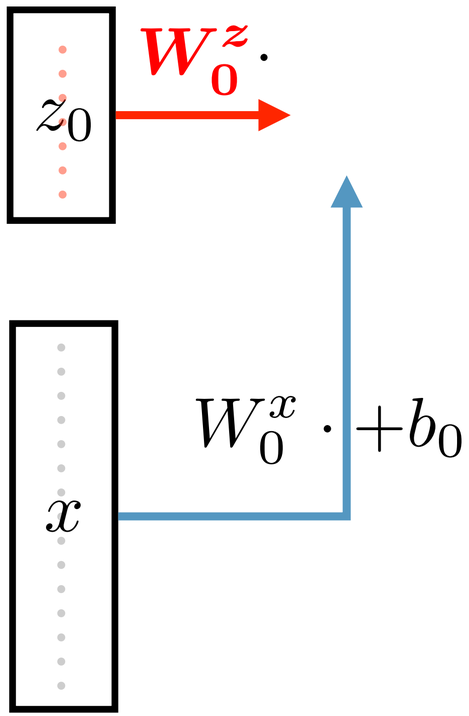
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



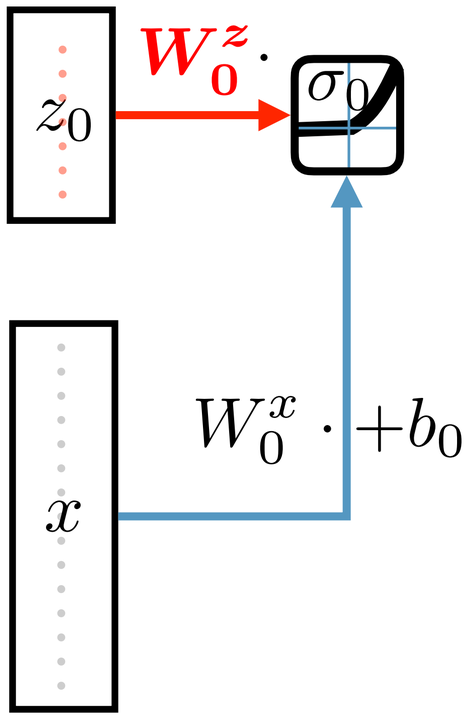
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



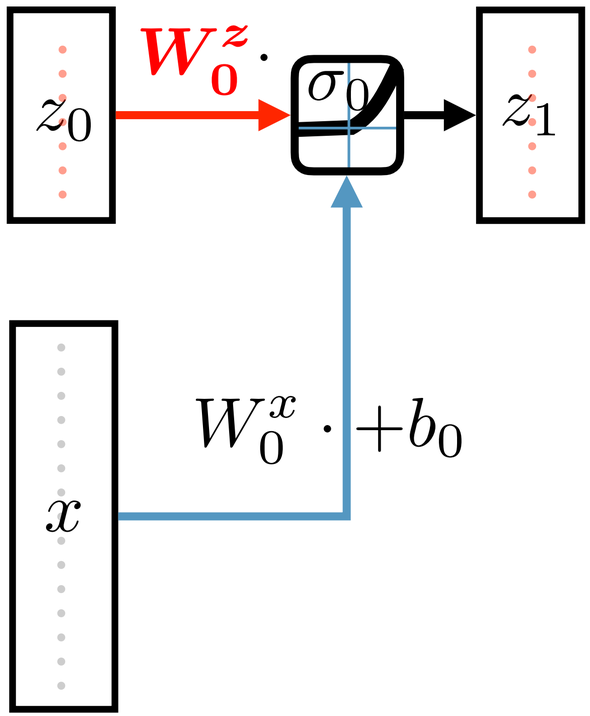
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



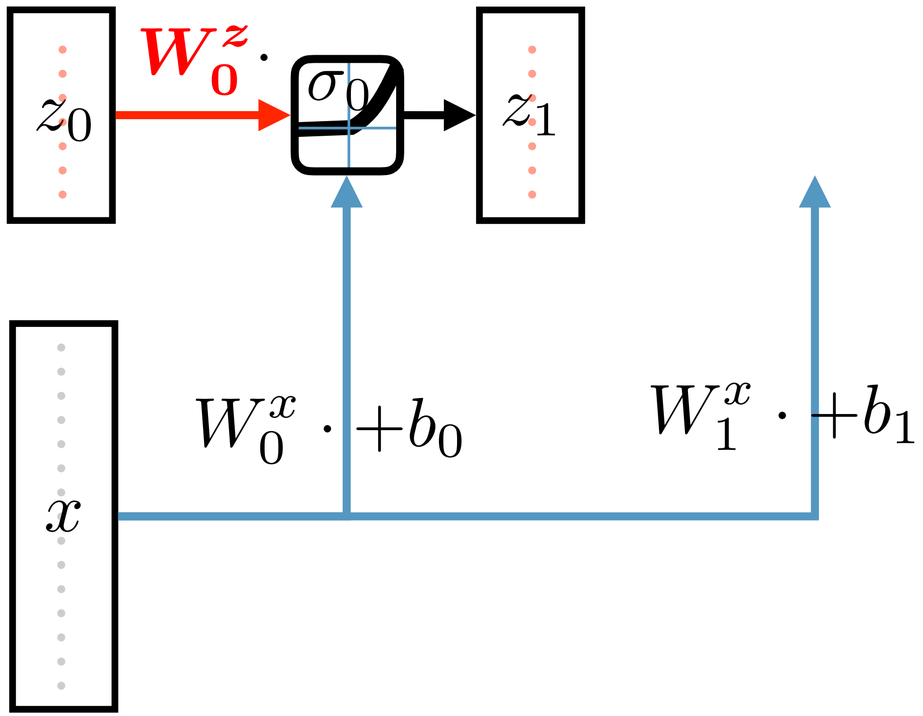
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



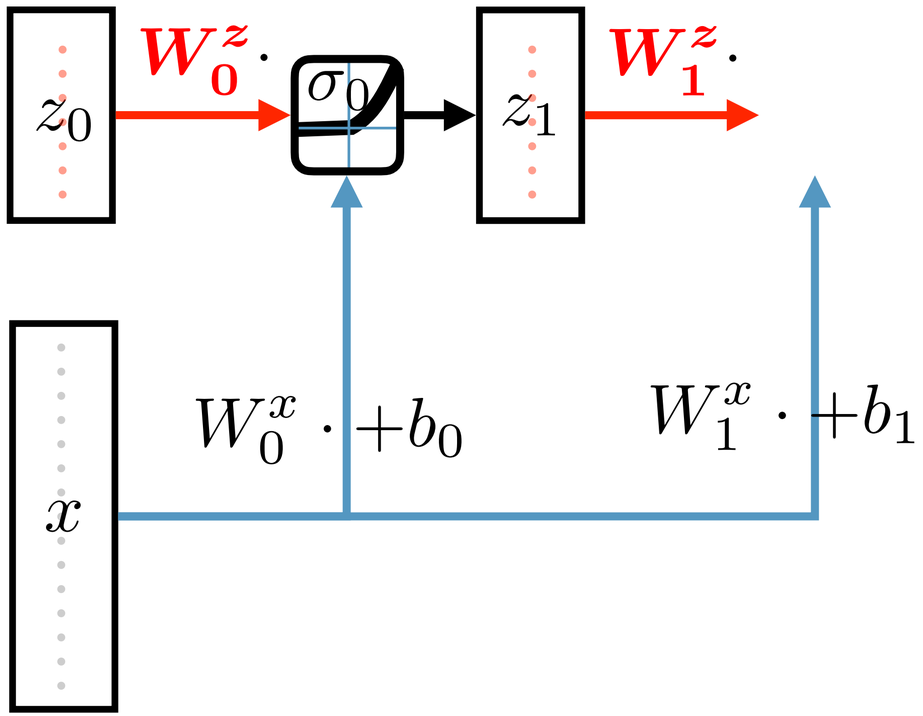
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



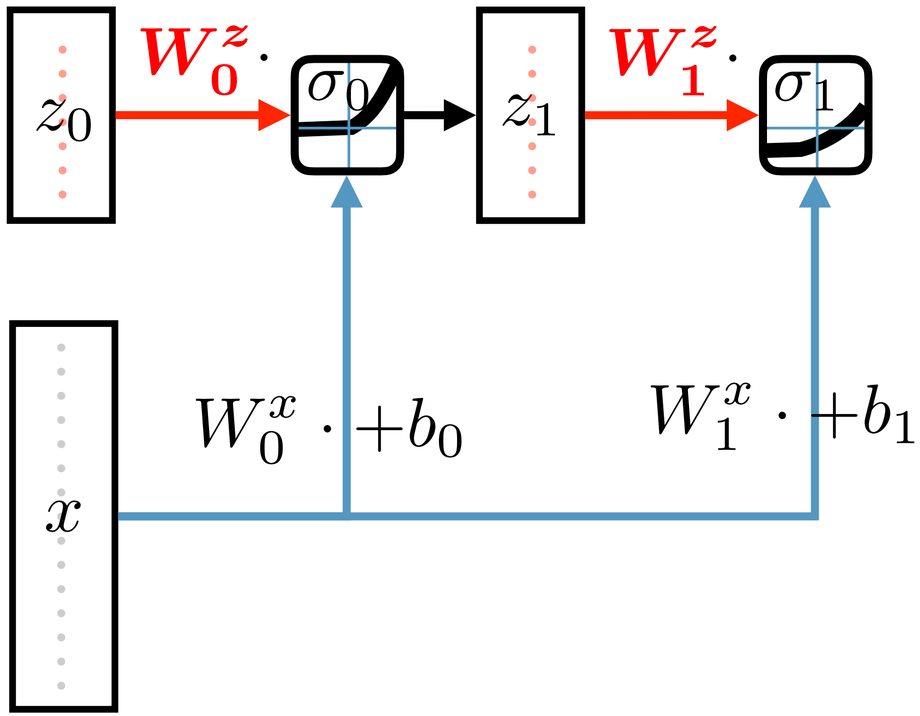
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



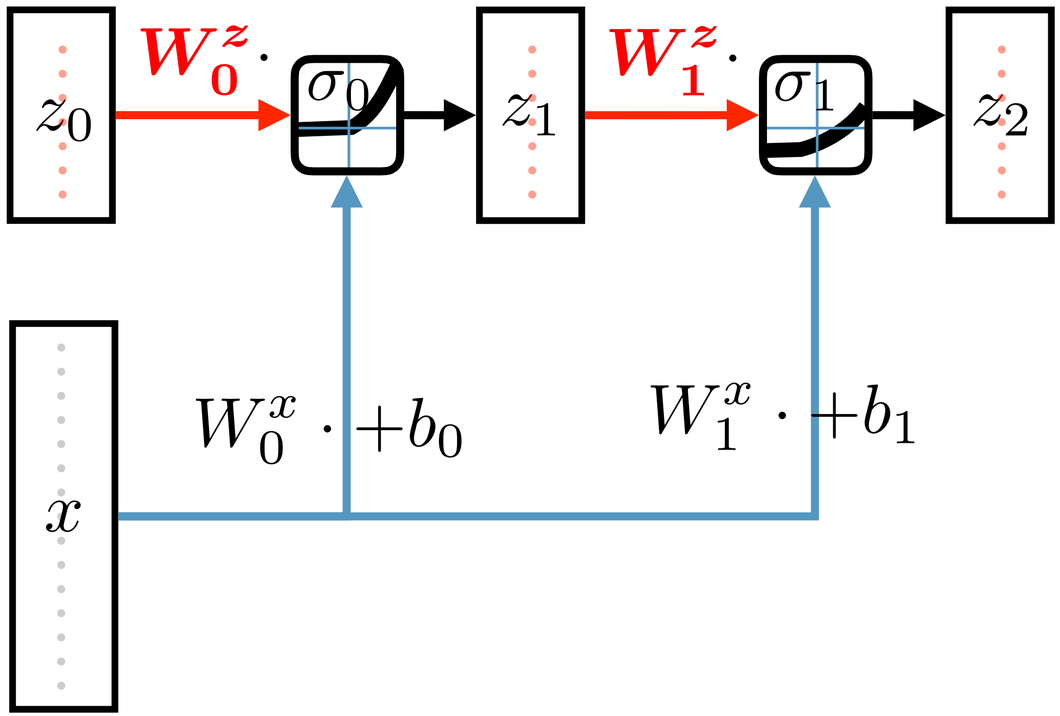
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



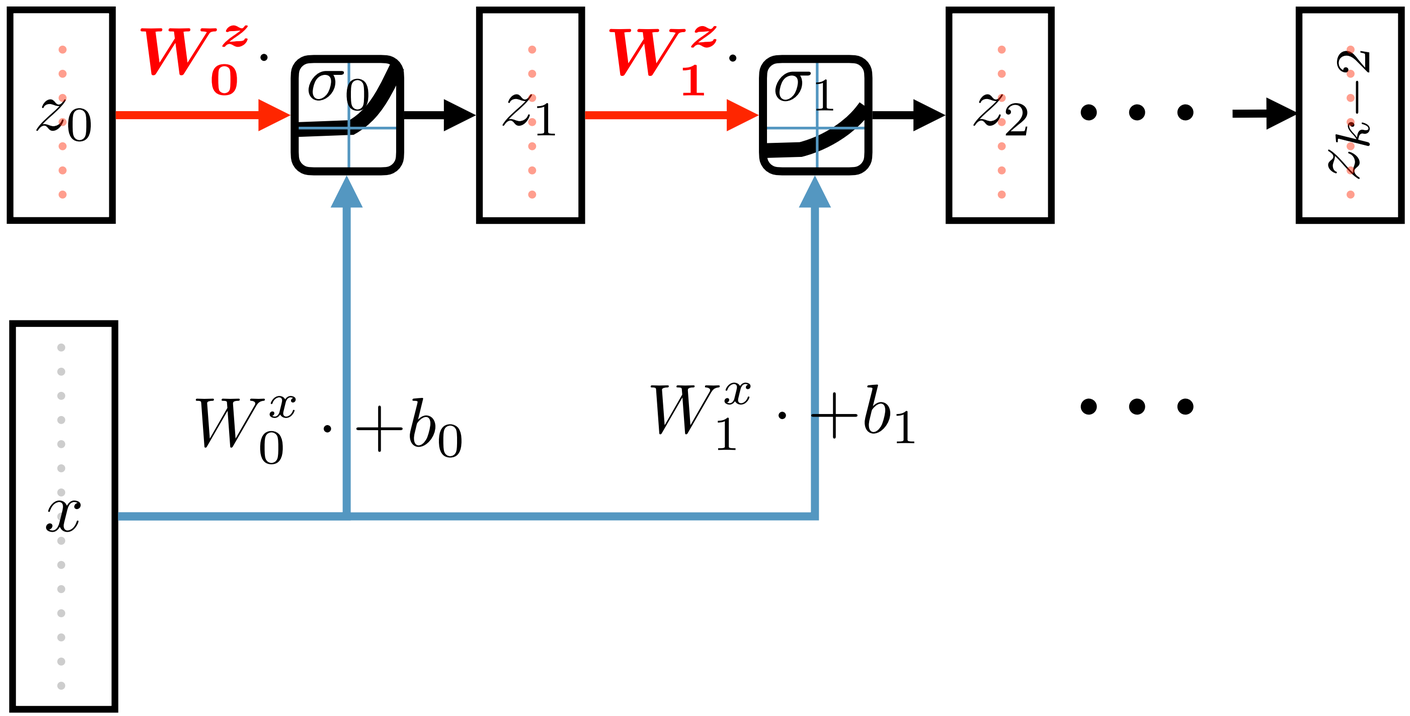
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



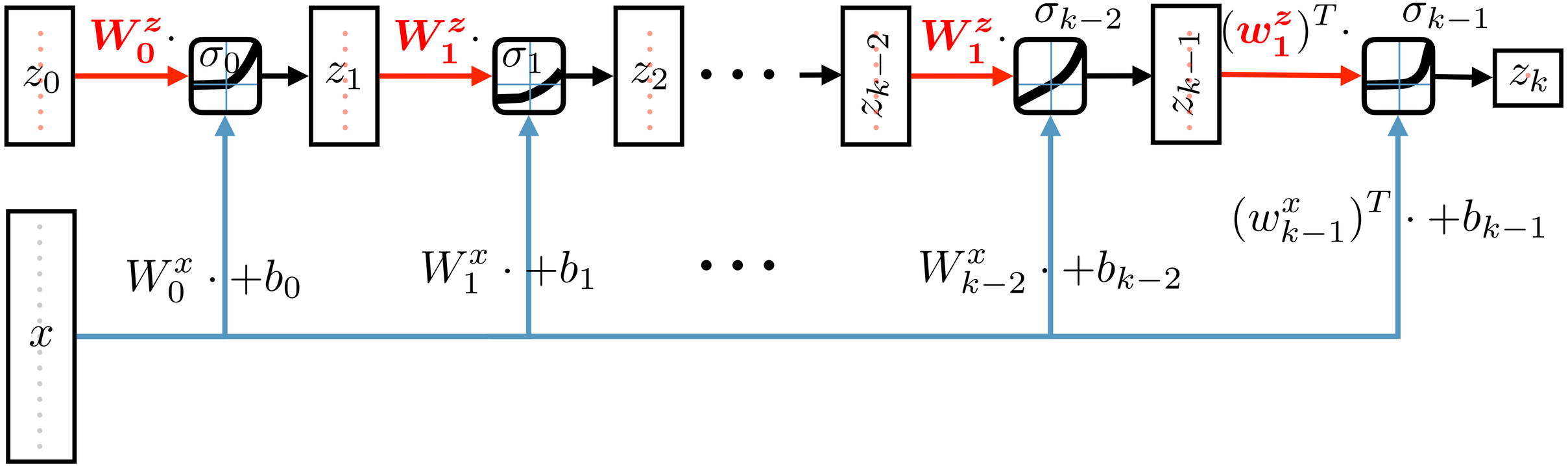
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



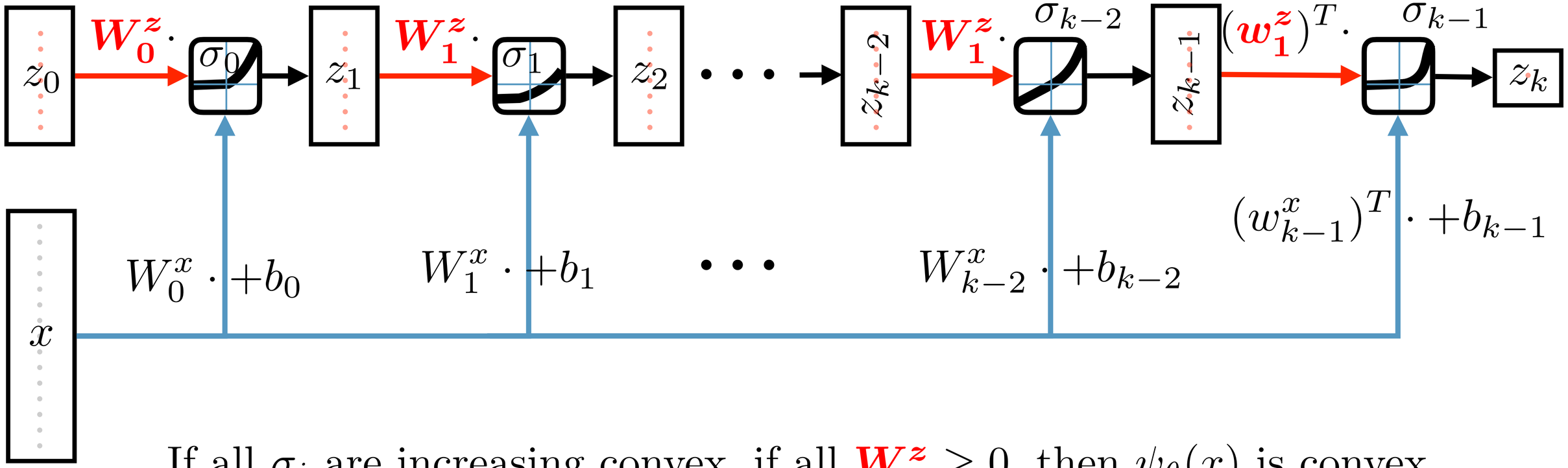
Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



Parameterize Convex Functions: ICNN

$$z_0 = \mathbf{0} \qquad z_{i+1} = \sigma_i\left(W_i^x x + \boldsymbol{W_i^z} z_i + b_i\right) \qquad \psi_\theta(x) = z_k$$



If all σ_i are increasing convex, if all $\boldsymbol{W_i^z} \geq 0$, then $\psi_\theta(x)$ is convex.

[Amos+17]

In Practice, using Surrogate Convex Conjugate

$$T^{\star} := \nabla f_{\theta^{\star}}, \text{ where } f_{\theta^{\star}} \leftarrow \inf_{f_\theta \text{ ICNN}} \int_{\mathbb{R}^d} f_\theta(x)\, d\rho_c(x) + \int_{\mathbb{R}^d} f_\theta^*(x)\, d\rho_d(x)$$

$$f_{\theta^\star} \leftarrow \arg_{f_\theta} \sup_{f_\theta \text{ ICNN}} \inf_{g_\omega \text{ ICNN}} - \int_{\mathbb{R}^d} f_\theta(x) d\rho_c(x) - \int_{\mathbb{R}^d} \langle y, \nabla g_\omega(y) \rangle - f_\theta(\nabla g_\omega(y))\, d\rho_d(y)$$

In Practice, using Surrogate Convex Conjugate

$$T^\star := \nabla f_{\theta^\star}, \text{ where } f_{\theta^\star} \leftarrow \inf_{f_\theta \text{ ICNN}} \int_{\mathbb{R}^d} f_\theta(x)\, d\rho_c(x) + \int_{\mathbb{R}^d} f_\theta^*(x)\, d\rho_d(x)$$

$$f_{\theta^\star} \leftarrow \arg_{f_\theta} \sup_{f_\theta \text{ ICNN}} \inf_{g_\omega \text{ ICNN}} - \int_{\mathbb{R}^d} f_\theta(x) d\rho_c(x) - \int_{\mathbb{R}^d} \langle y, \nabla g_\omega(y) \rangle - f_\theta(\nabla g_\omega(y))\, d\rho_d(y)$$

Training this in practice is challenging:

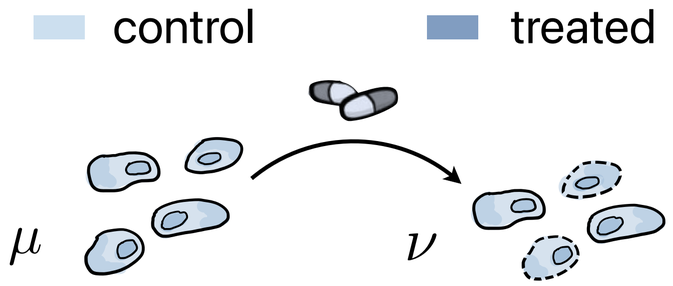
- **Constraints** in weight matrices
- Challenge of **handling conjugate** (usually drop convexity, see also **[Amos'22]** for amortized approaches)
- Challenge of **initialisation**. (See **[Bunne+22b]**)

Neural Solver

Application: Predicting Treatment Responses to Drugs

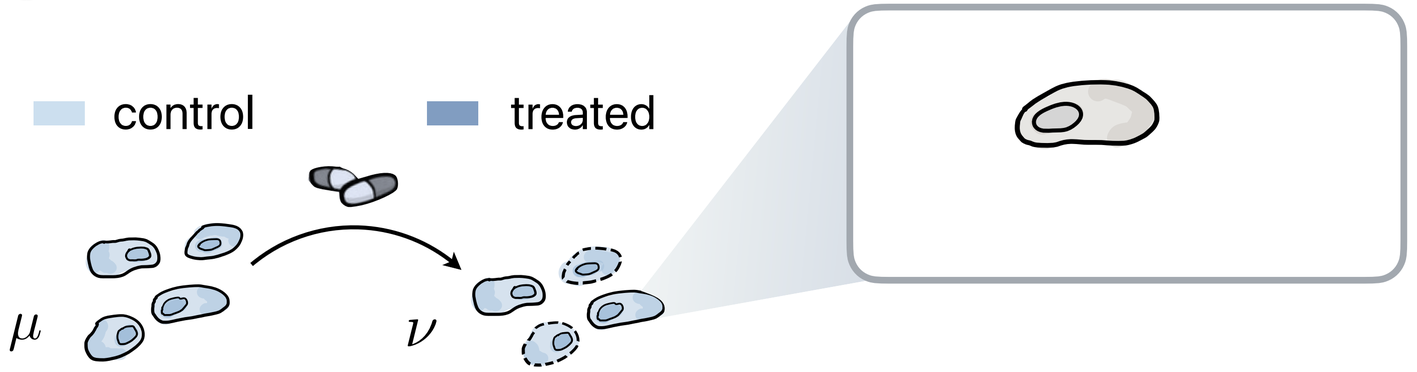
Part 2

Example of a **drug** on a **mixture of two cell lines**



Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**

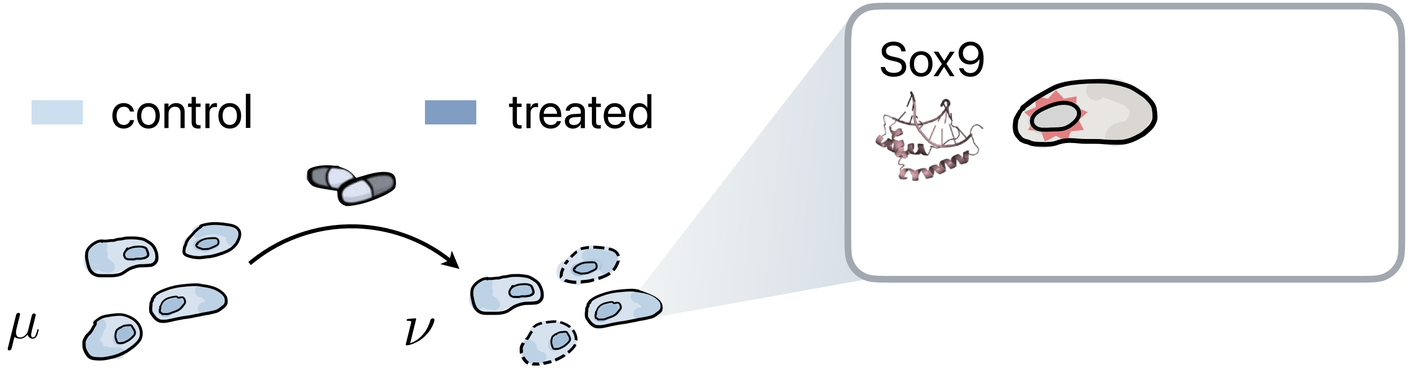


4i microscopy

[Gut+18]

Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**

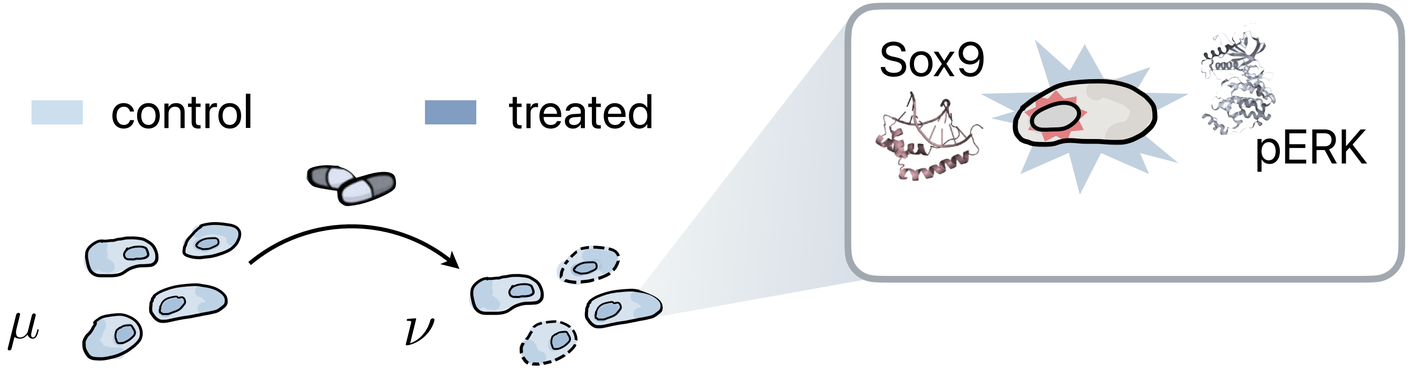


4i microscopy

[Gut+18]

Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**

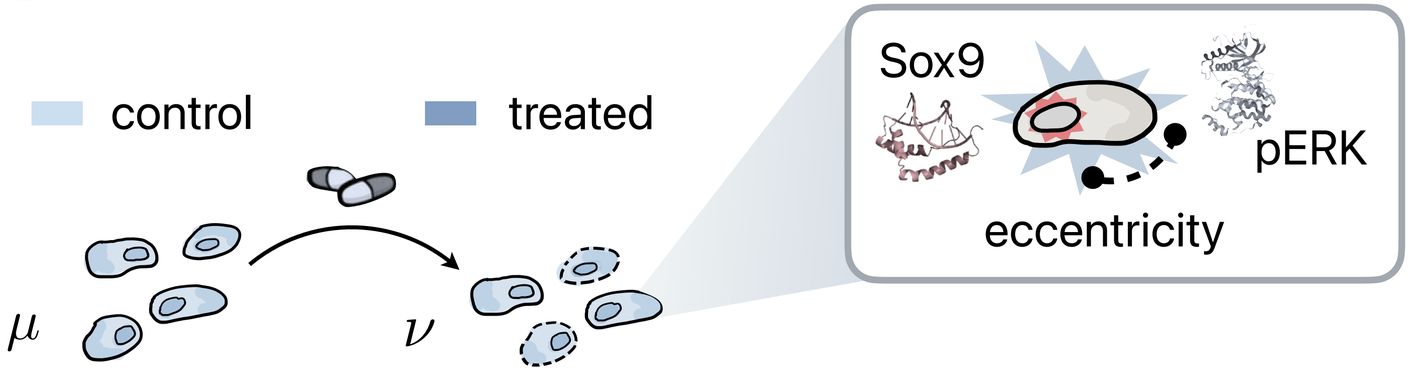


4i microscopy

[Gut+18]

Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**

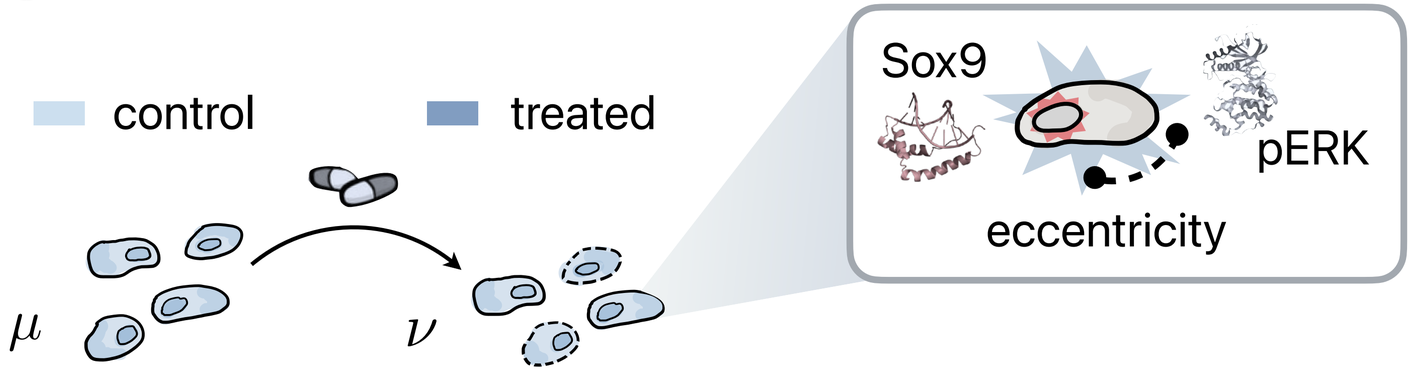


4i microscopy

[Gut+18]

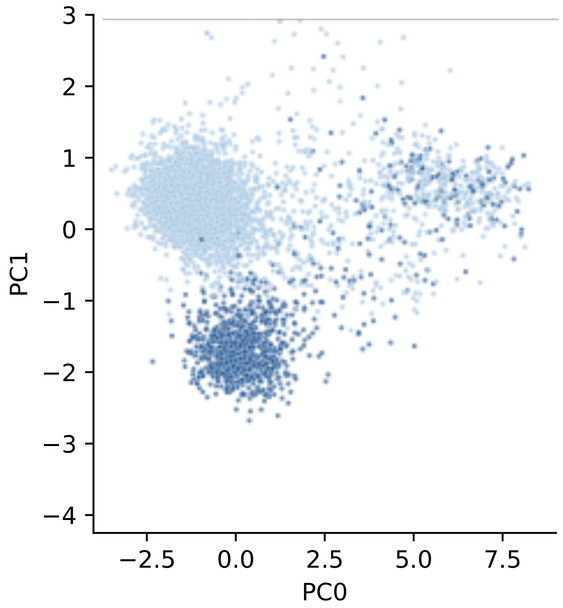
Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**



4i microscopy

[Gut+18]

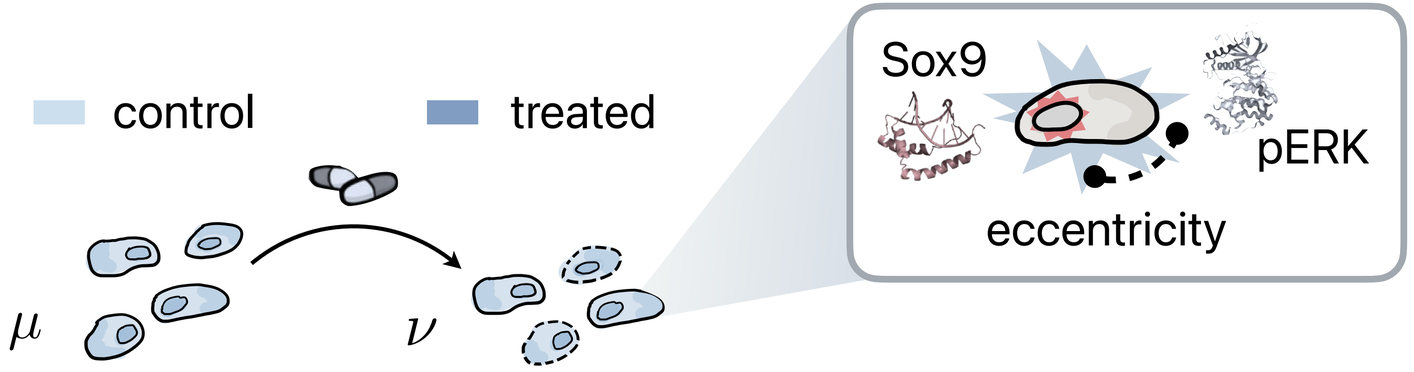


control

treated

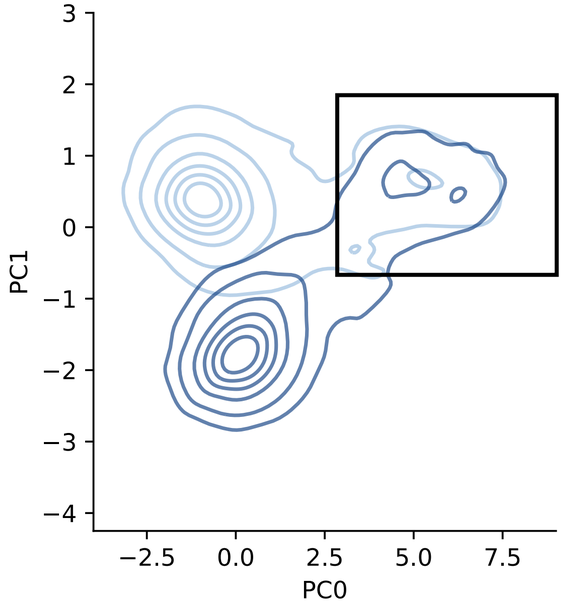
Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**



4i microscopy

[Gut+18]

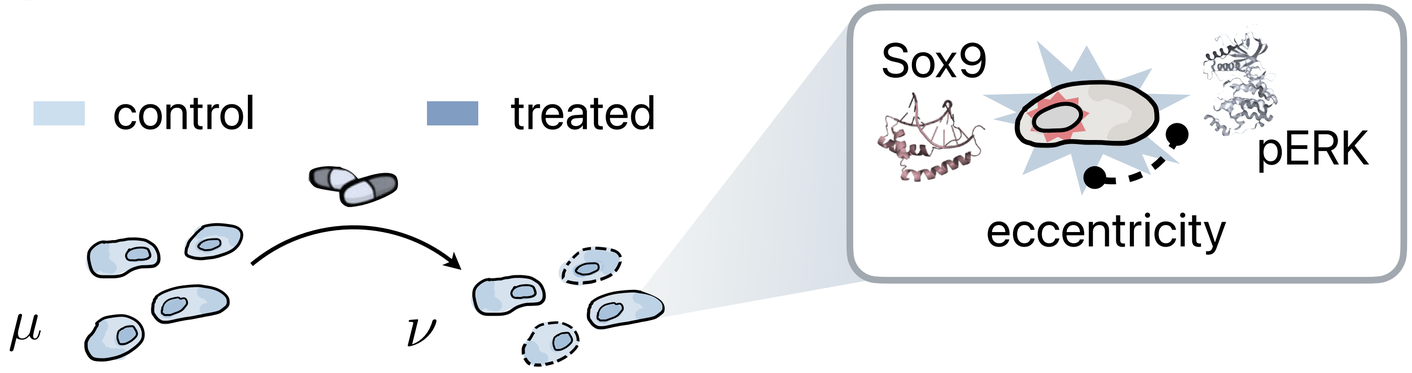


control

treated

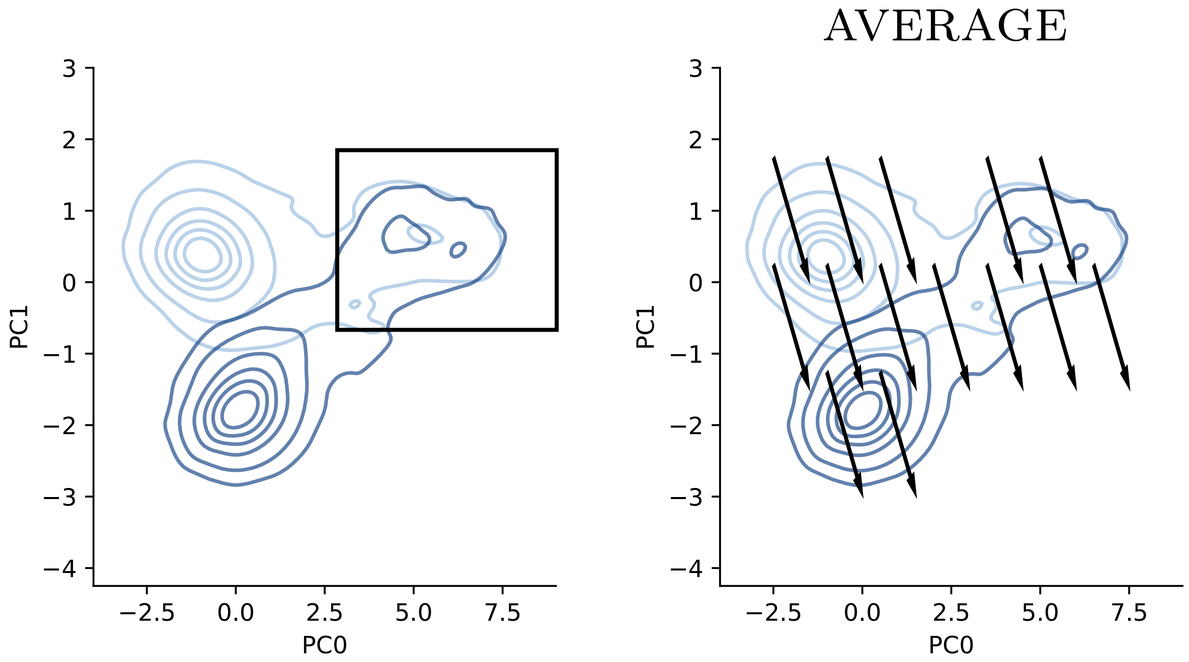
Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**



4i microscopy

[Gut+18]

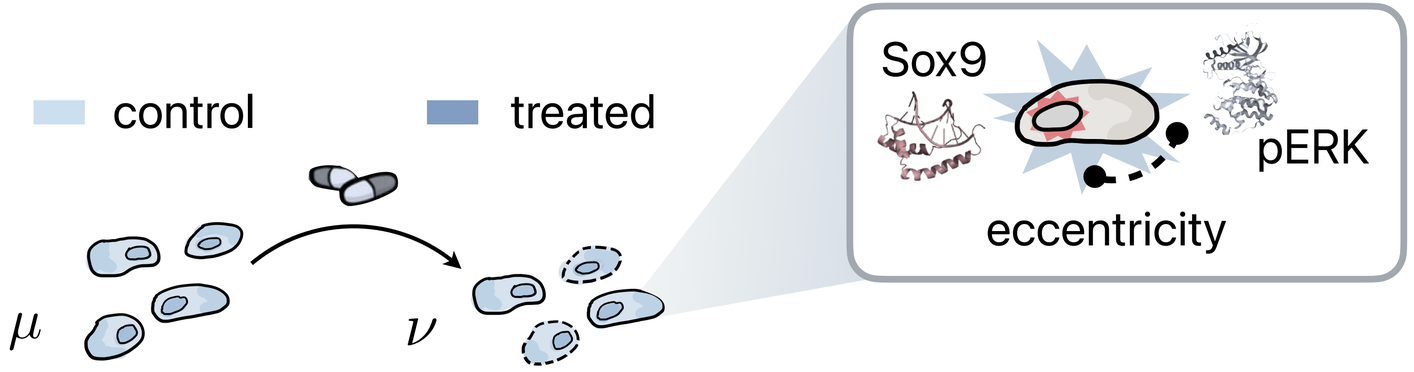


control

treated

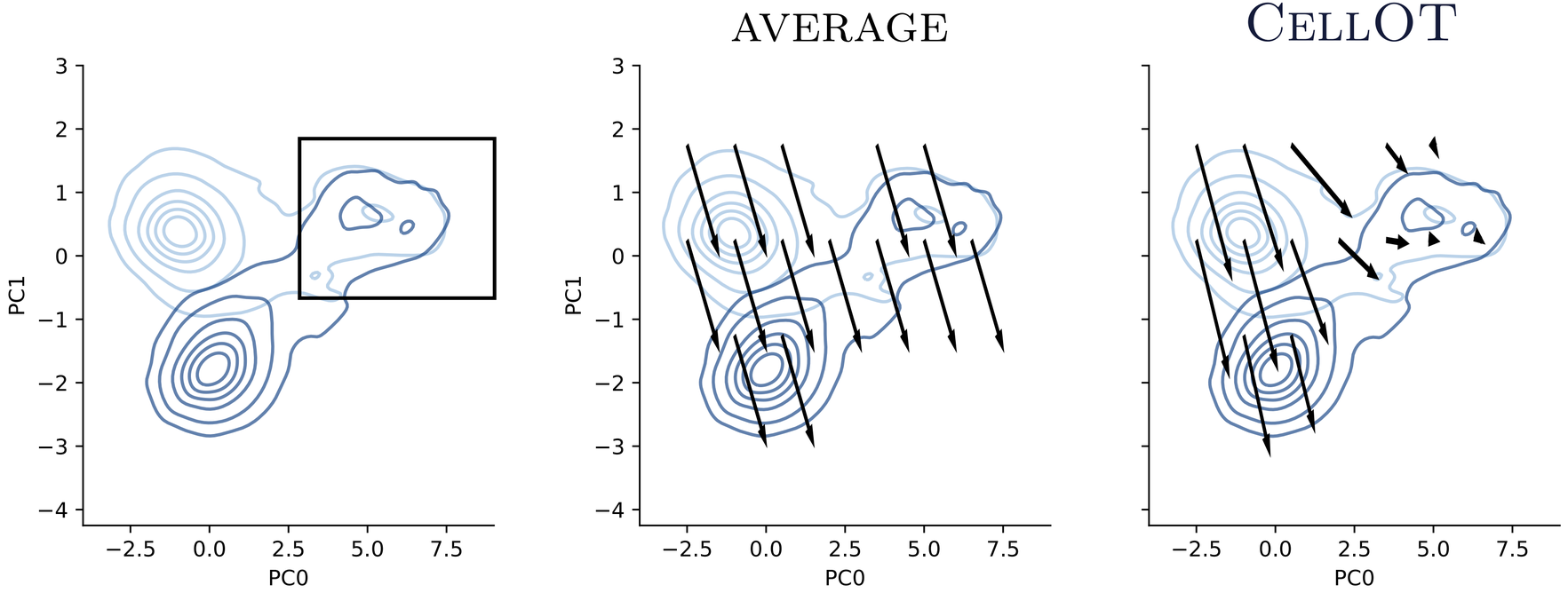
Application: Predicting Treatment Responses to Drugs

Example of a **drug** on a **mixture of two cell lines**



4i microscopy

[Gut+18]



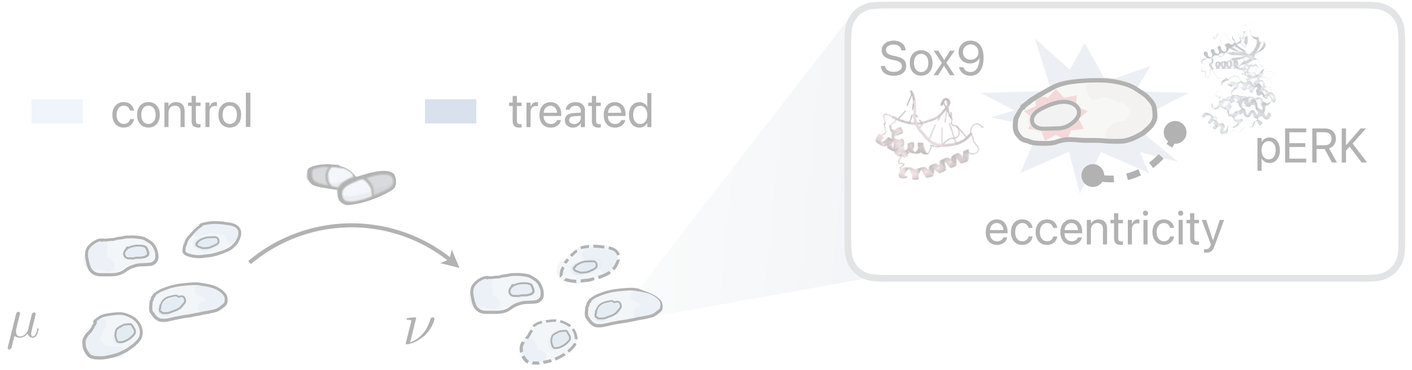
[Bunne+23]

control

treated

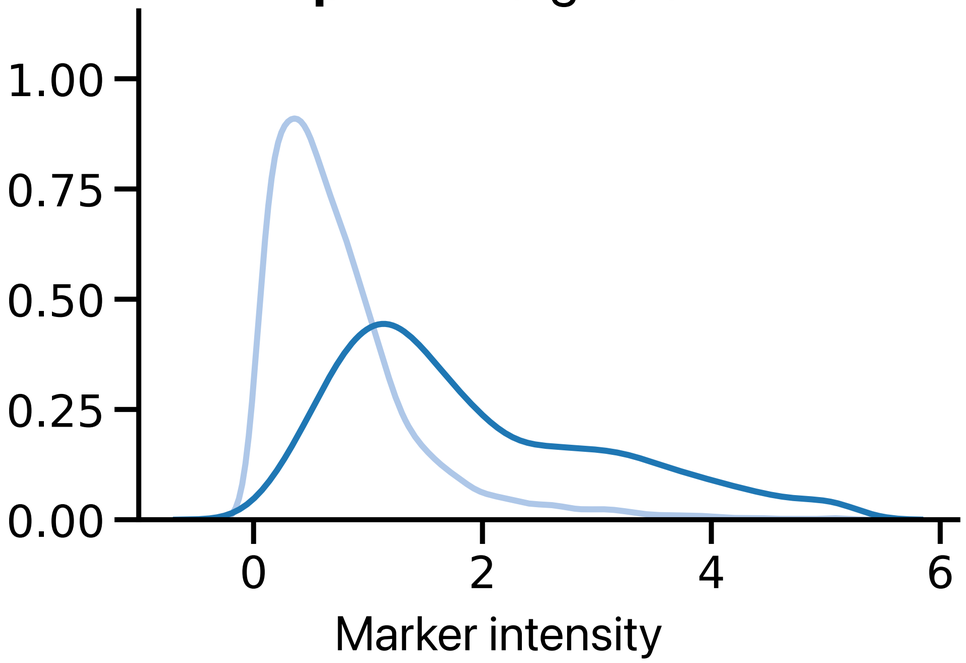
Application: Predicting Treatment Responses to Drugs

… do we capture **changes in molecular markers**?

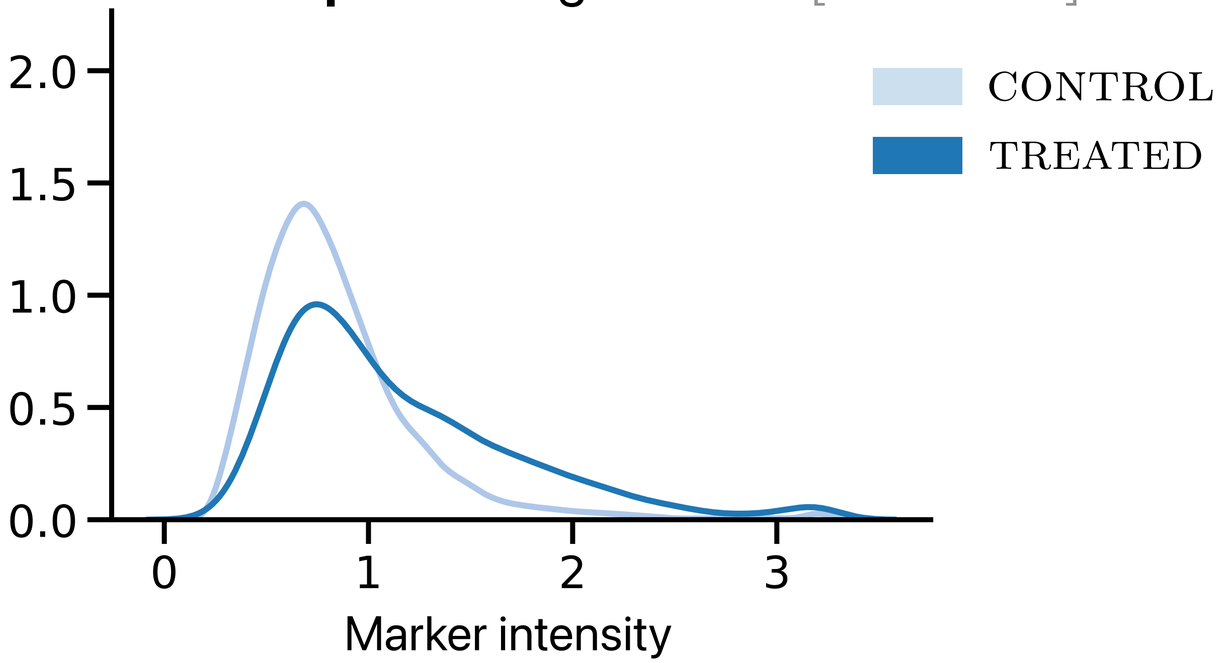


4i microscopy
[Gut+18]

pAKT Marginals



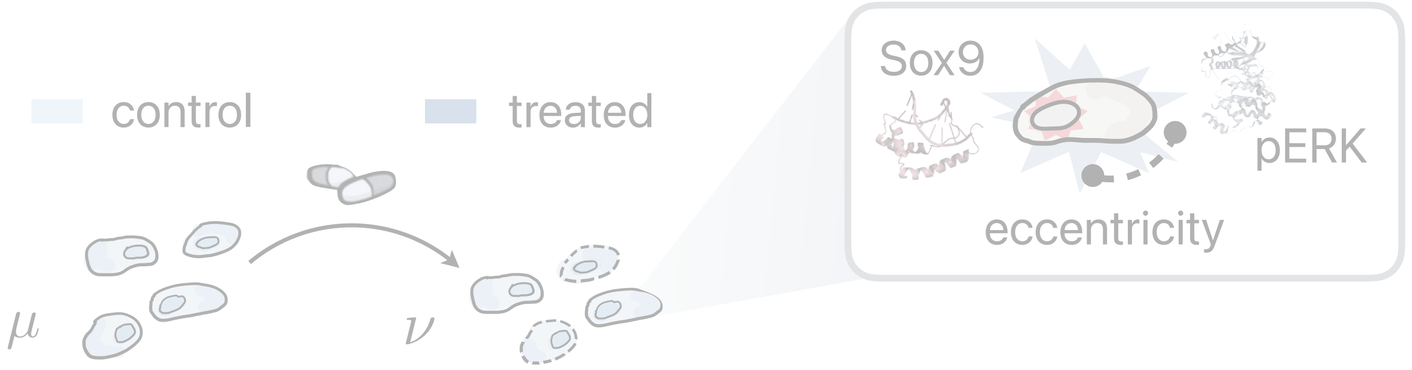
pERK Marginals [Bunne+23]



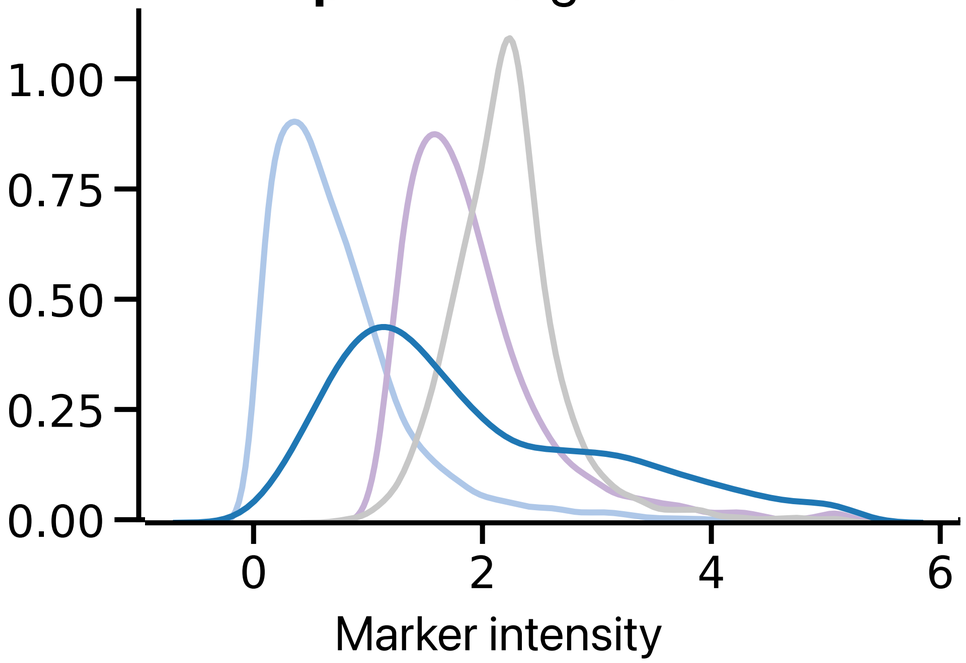
Application: Predicting Treatment Responses to Drugs

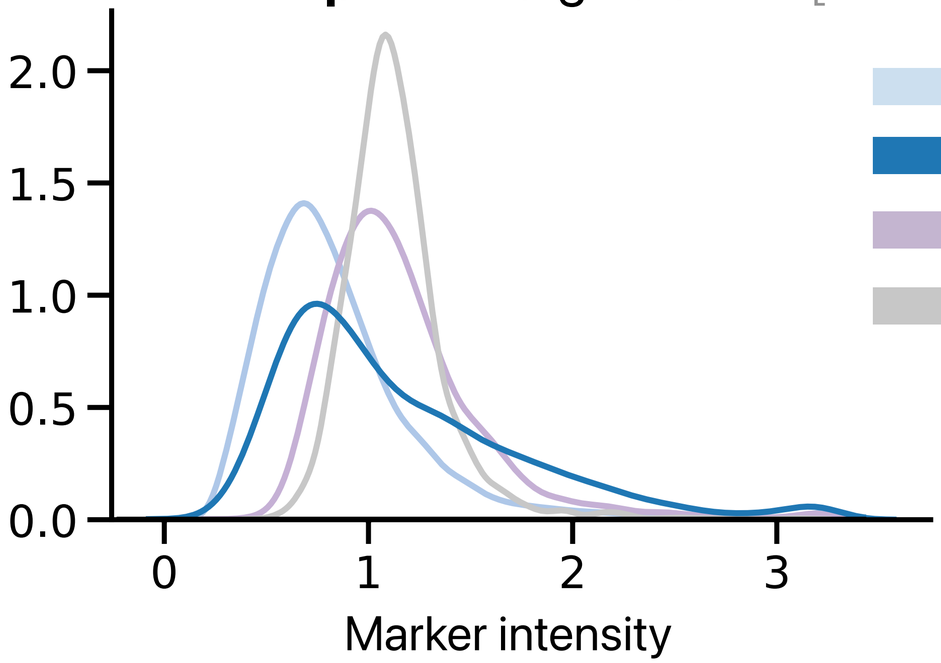
Part 2

... do we capture **changes in molecular markers**?



4i microscopy
[Gut+18]



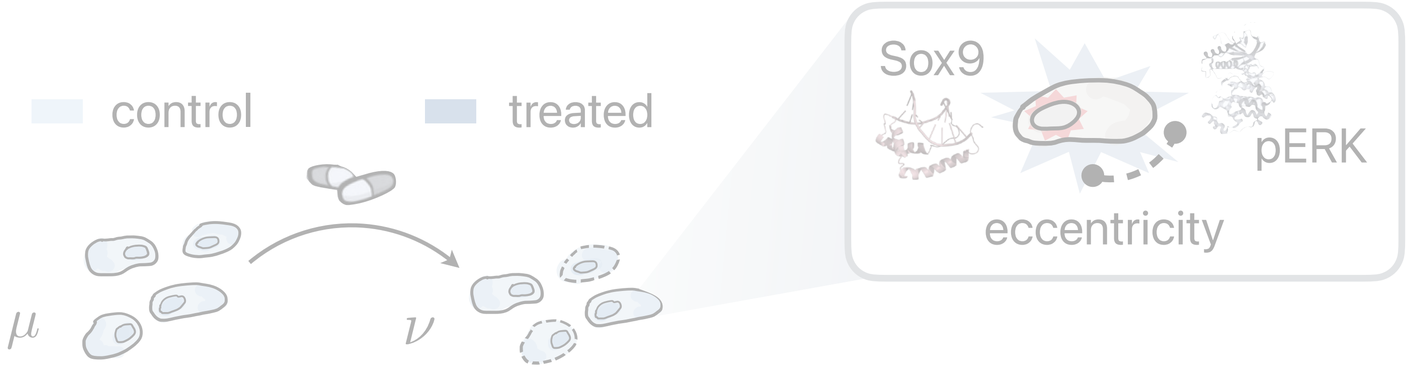


[Bunne+23]

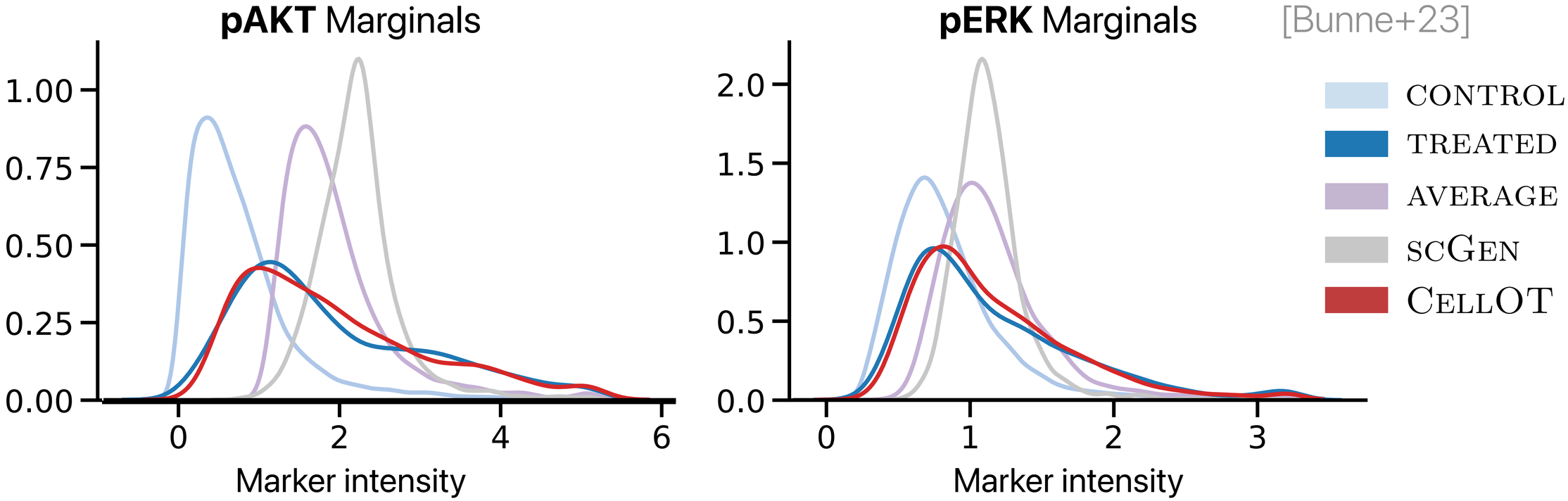
CONTROL
TREATED
AVERAGE
SCGEN

Application: Predicting Treatment Responses to Drugs

… do we capture **changes in molecular markers**?

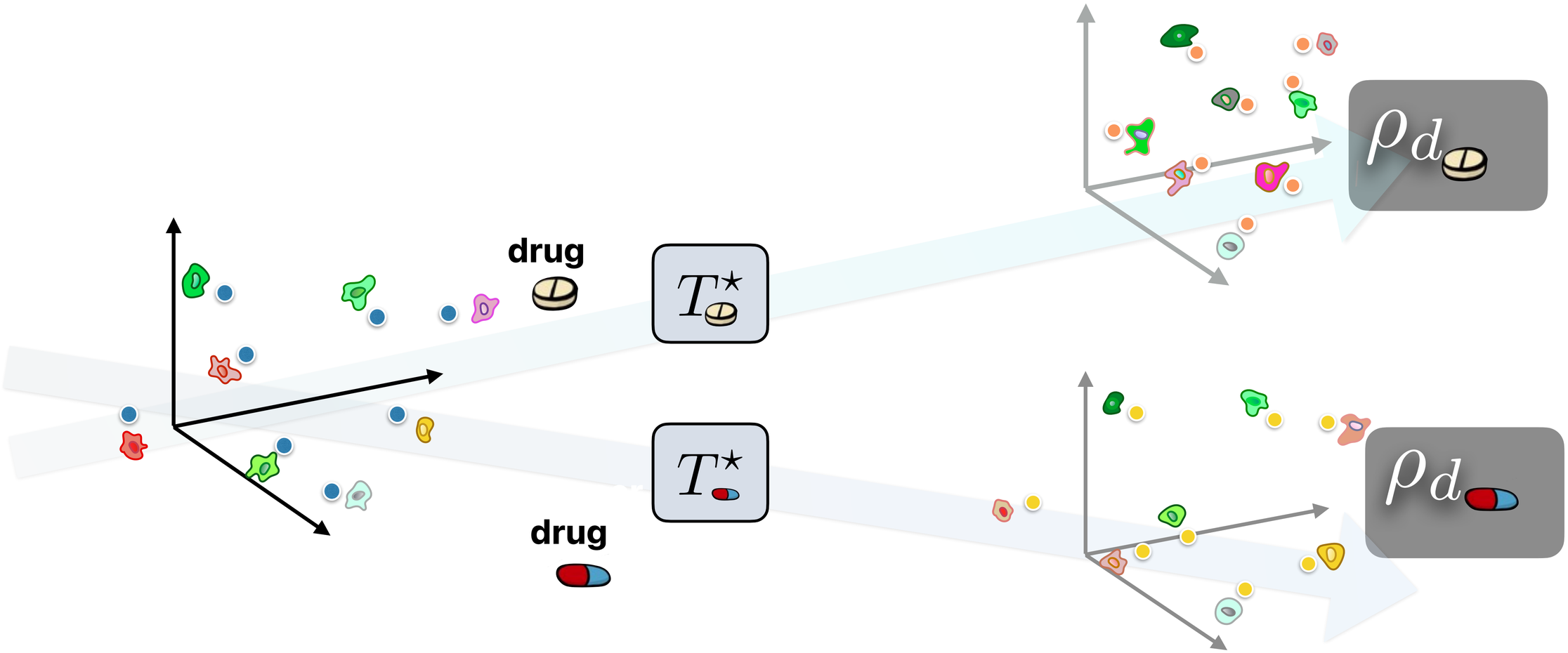


4i microscopy
[Gut+18]

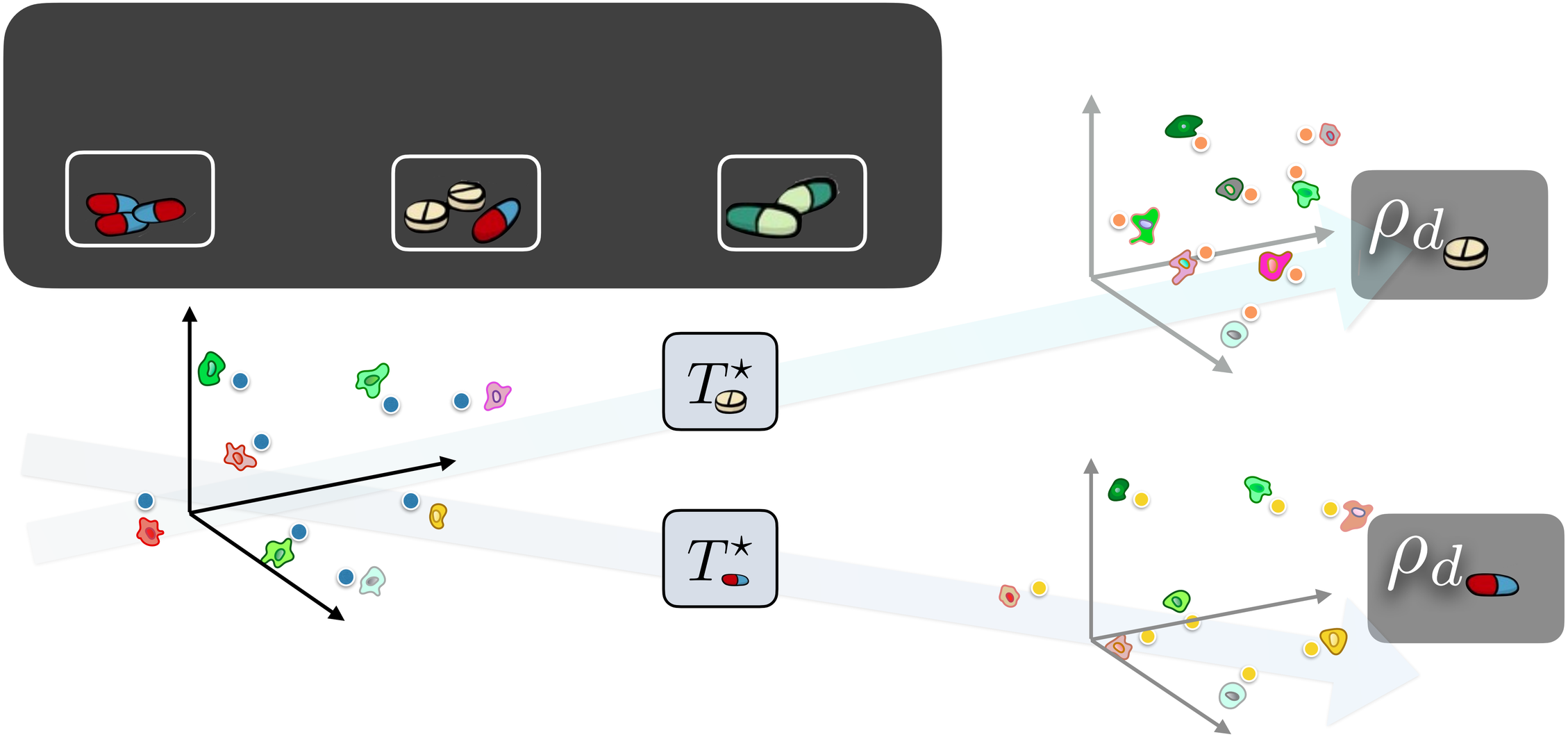


[Bunne+23]

Problem: Learning a map per single drug does not scale

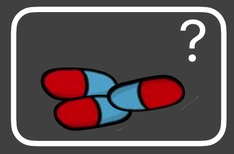


Extrapolating to Unknown Drugs (or Combinations)



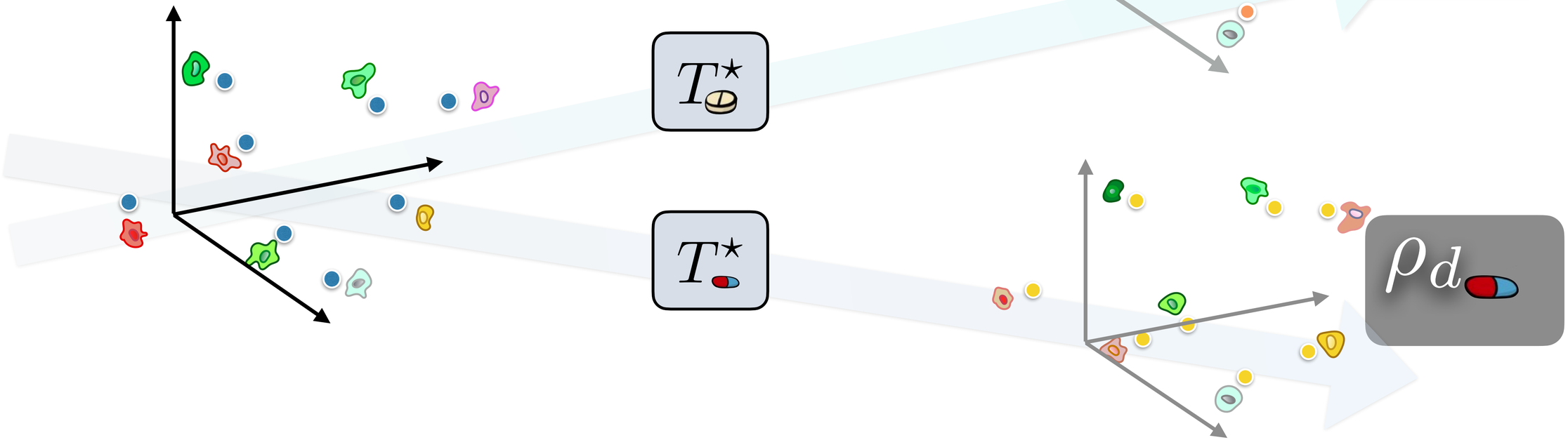
Extrapolating to Unknown Drugs (or Combinations)

Can we use the *same* datasets, to extrapolate the effects of unseen combinations?



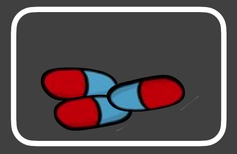


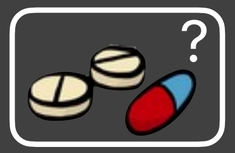




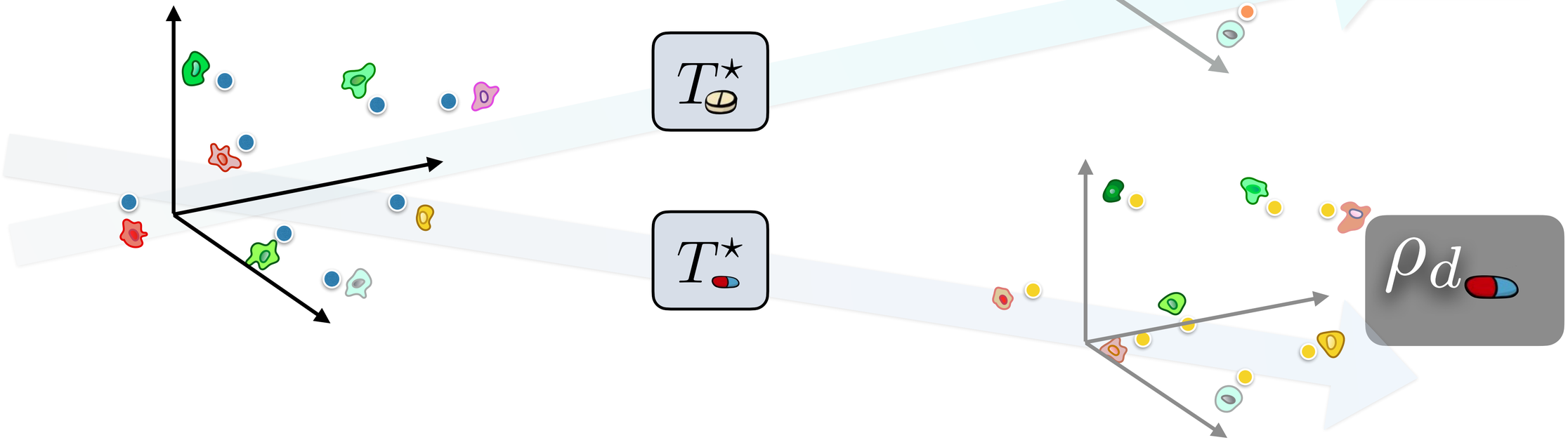
Extrapolating to Unknown Drugs (or Combinations)

Can we use the *same* datasets, to extrapolate the effects of unseen combinations?



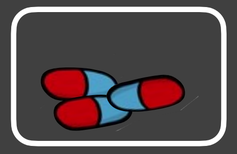






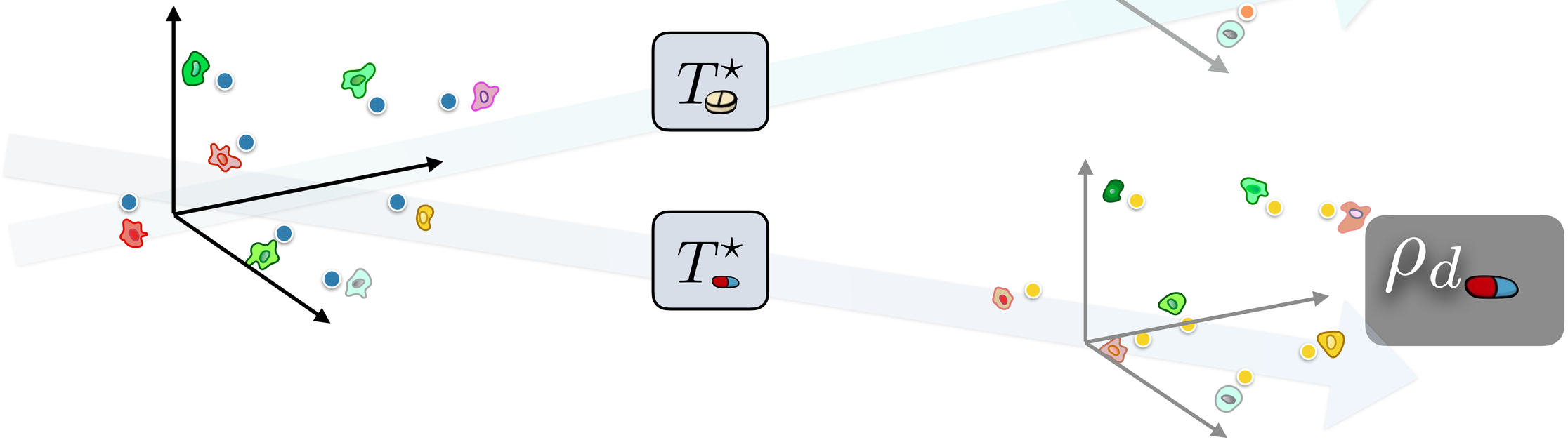
Extrapolating to Unknown Drugs (or Combinations)

Can we use the *same* datasets, to extrapolate the effects of unseen combinations?



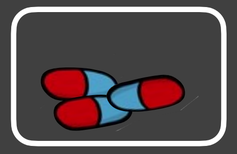






Extrapolating to Unknown Drugs (or Combinations)

Can we use the *same* datasets, to extrapolate the effects of unseen combinations?





Parameterization of conditional OT? PICNN

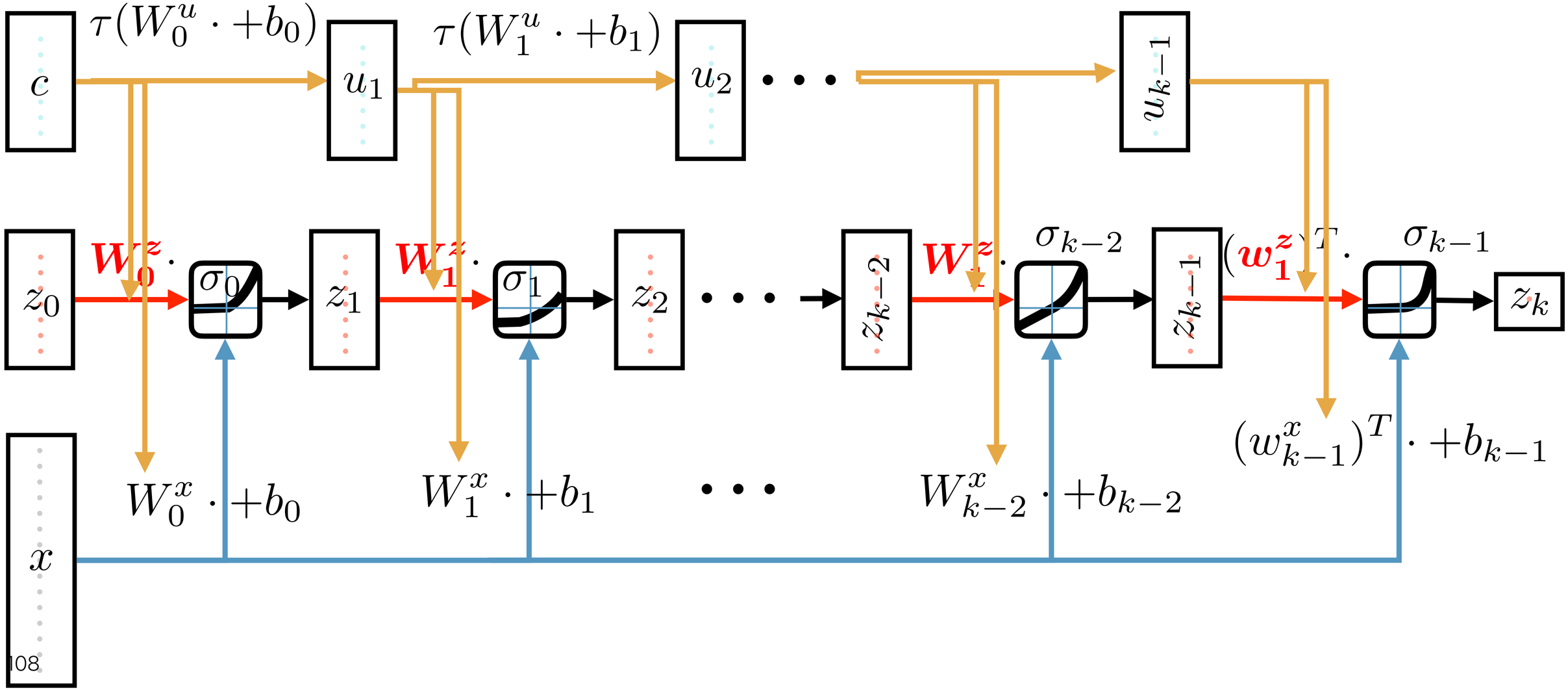
$$f_\theta(x) \longrightarrow f_\theta(x, c)$$ convex w.r.t x for all c

$$
\begin{aligned}
u_{k+1} &= \tau_k \left(V_k u_k + v_k\right) \\
h_{k+1} &= a_k \left(W_k^h \left(h_k \circ \left[W_k^{hu} u_k + b_k^h\right]_+ \right) + W_k^x \left(x \circ \left[W_k^{xu} u_k + b_k^x\right] \right) + W_k^u u_k + b_k^u \right) \\
f_\theta(x, c) &= h_k.
\end{aligned}
$$

[Amos+17]

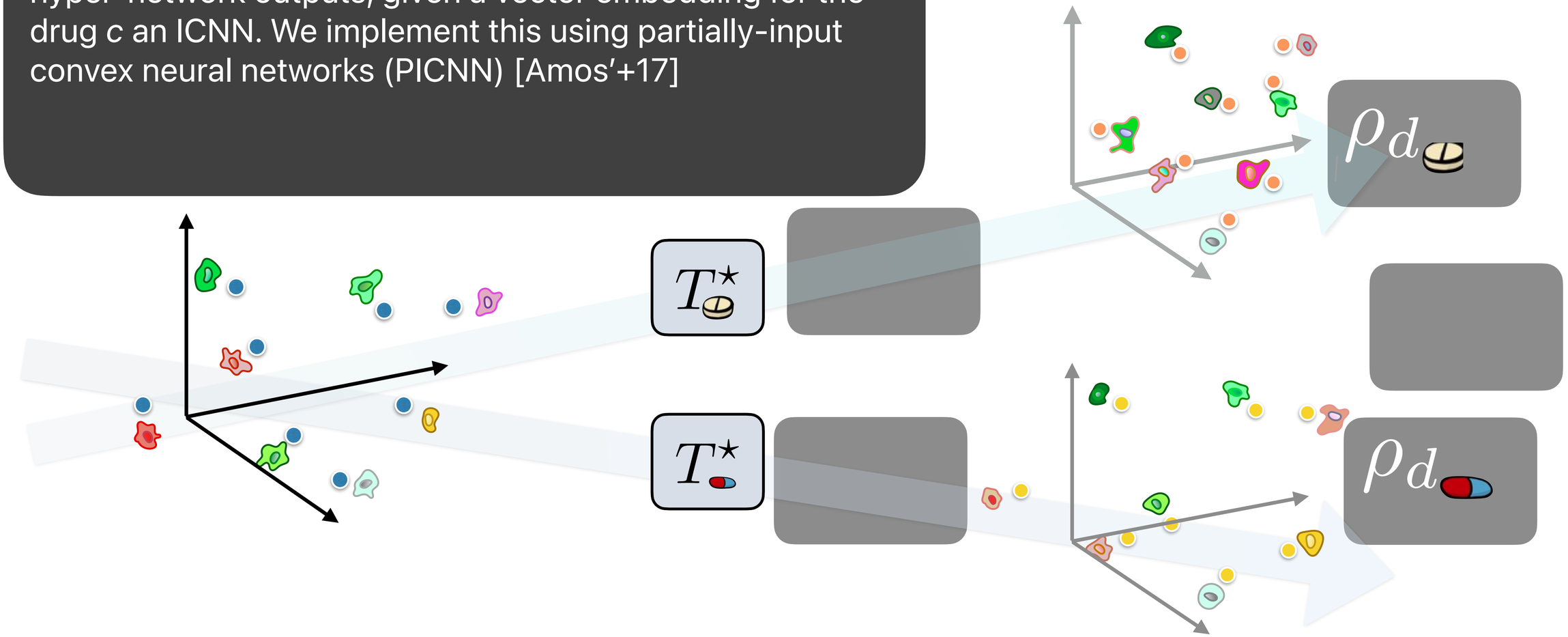
This can be seen as a *"hypernetwork"* producing an ICNN depending on context *c*.

From ICNN to PICNN



Implementing Conditional Monge Maps as ∇_1PICNN(x,c)

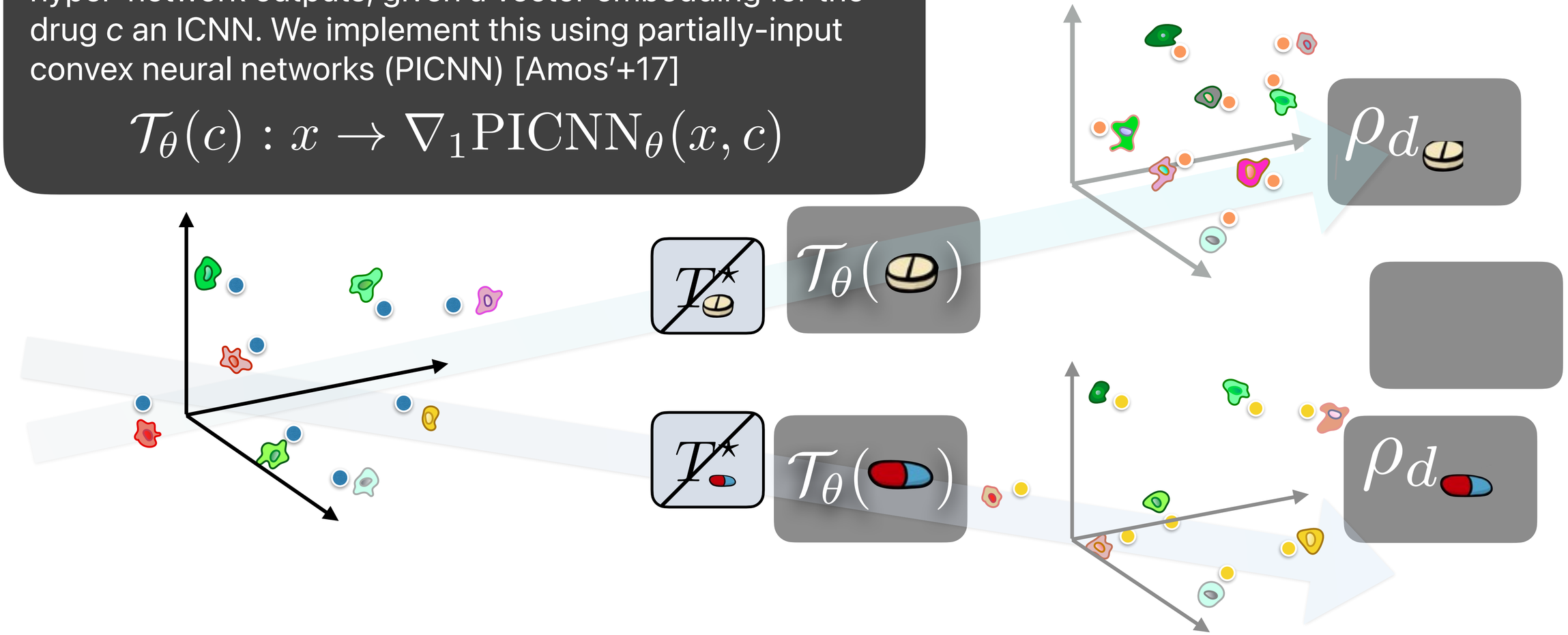
Our solution: joint learning of a single hyper-network. That hyper-network outputs, given a vector embedding for the drug c an ICNN. We implement this using partially-input convex neural networks (PICNN) [Amos'+17]



Implementing Conditional Monge Maps as ∇_1PICNN(*x*,*c*)

Our solution: joint learning of a single hyper-network. That hyper-network outputs, given a vector embedding for the drug *c* an ICNN. We implement this using partially-input convex neural networks (PICNN) [Amos'+17]

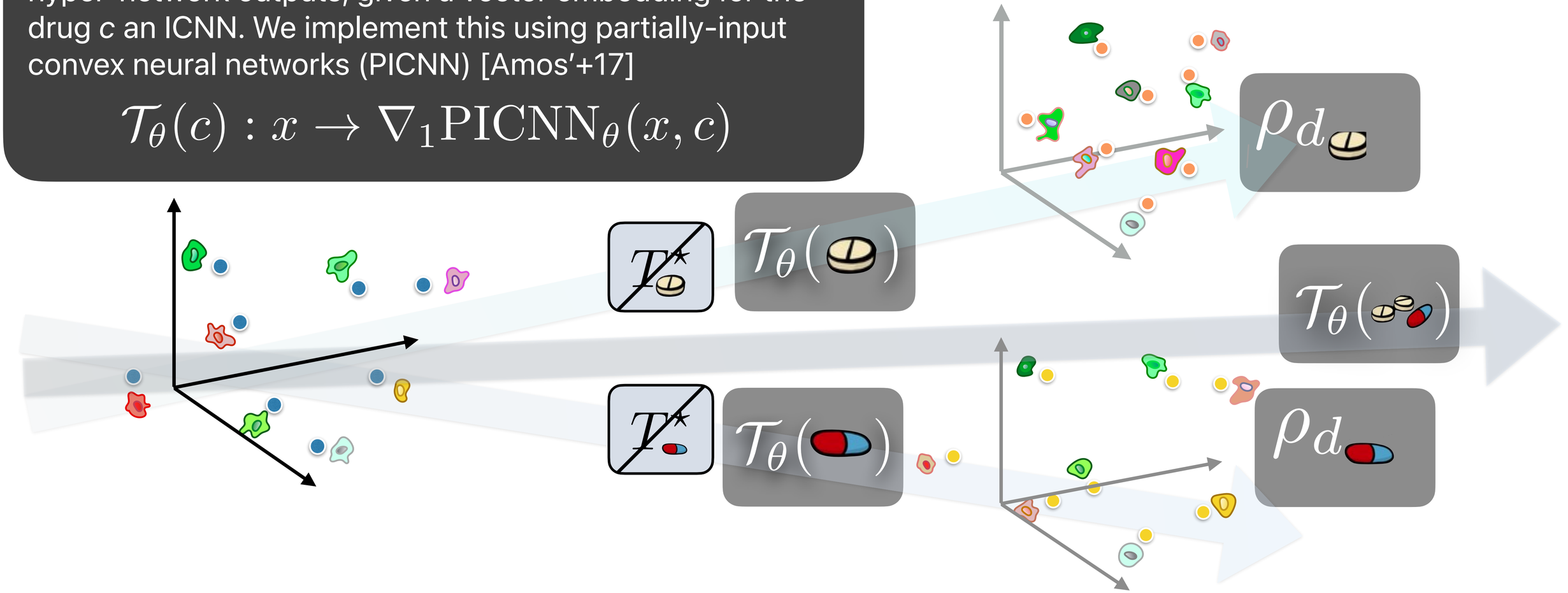
$$\mathcal{T}_\theta(c) : x \to \nabla_1 \mathrm{PICNN}_\theta(x, c)$$



Implementing Conditional Monge Maps as ∇_1PICNN(*x*,*c*)

Our solution: joint learning of a single hyper-network. That hyper-network outputs, given a vector embedding for the drug *c* an ICNN. We implement this using partially-input convex neural networks (PICNN) [Amos'+17]

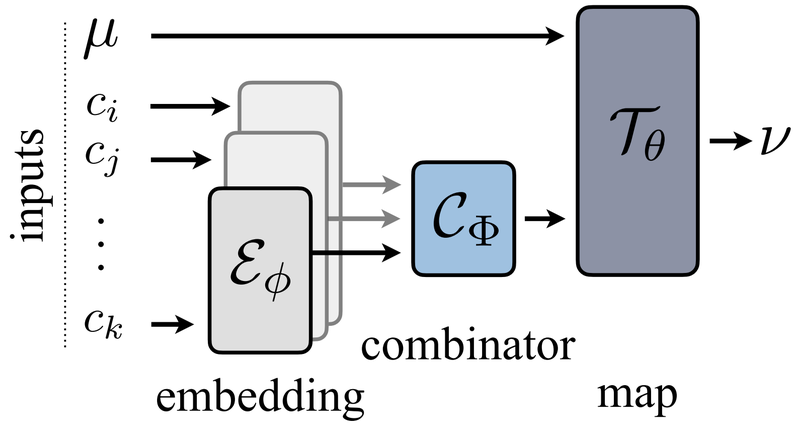
$$\mathcal{T}_\theta(c) : x \to \nabla_1 \mathrm{PICNN}_\theta(x, c)$$



Entire Conditional OT (CondOT) includes embeddings for *c*

Model Architecture

Model Components

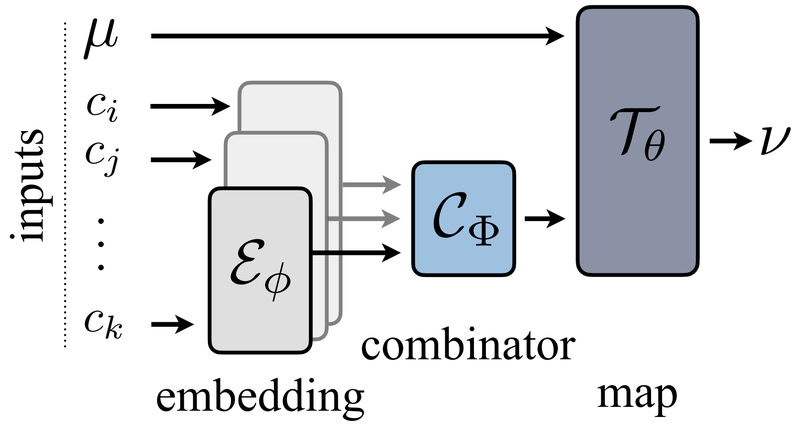


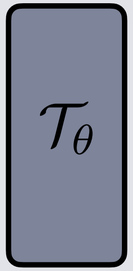
$$\mathcal{T}_\theta = \nabla_1 \mathbf{PICNN}_\theta$$

Entire Conditional OT (CondOT) includes embeddings for *c*

Model Architecture

Model Components





.. parameterized via
partial input convex neural network
(PICNN)

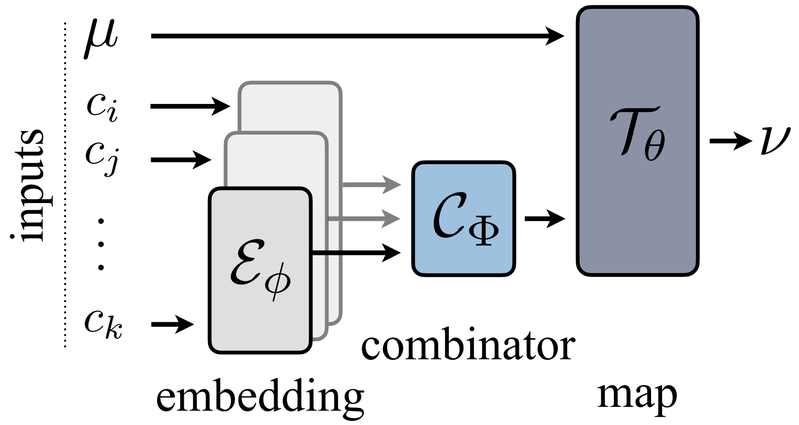
$$\mathcal{T}_\theta = \nabla_1 \mathbf{PICNN}_\theta$$

.. observation: **partial gradient.**

.. can be learned using [Makkuva+20]

Entire Conditional OT (CondOT) includes embeddings for *c*

Model Architecture



Model Components

$\mathcal{E}_\phi$

.. taking a context variable c,
embed $\mathcal{E}_\phi(c)$

.. e.g., one-hot, mode-of-action embedding

$\mathcal{T}_\theta$

.. parameterized via
partial input convex neural network
(PICNN)

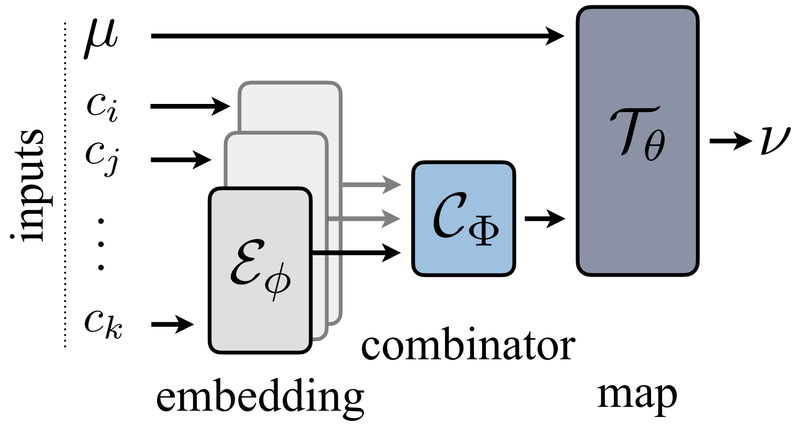
$$\mathcal{T}_\theta = \nabla_1 \mathbf{PICNN}_\theta$$

.. observation: **partial gradient.**

.. can be learned using [Makkuva+20]

Entire Conditional OT (CondOT) includes embeddings for *c*

Model Architecture



Model Components

$\mathcal{E}_\phi$

.. taking a context variable c, embed $\mathcal{E}_\phi(c)$

.. e.g., one-hot, mode-of-action embedding

$\mathcal{C}_\Phi$

.. take arbitrary number of embedded context variables, get combination via $\hat{c}_i = \mathcal{C}_\Phi\left(\mathcal{E}_\phi\left(c^1\right), \mathcal{E}_\phi\left(c^2\right), \ldots, \mathcal{E}_\phi\left(c^k\right)\right)$

.. e.g., multi-hot, deep set combinator

$\mathcal{T}_\theta$

.. parameterized via partial input convex neural network (PICNN)

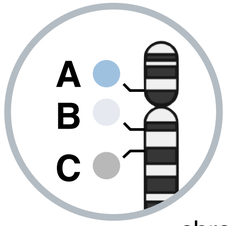
$$\mathcal{T}_\theta = \nabla_1 \mathbf{PICNN}_\theta$$

.. observation: **partial gradient.**

.. can be learned using [Makkuva+20]

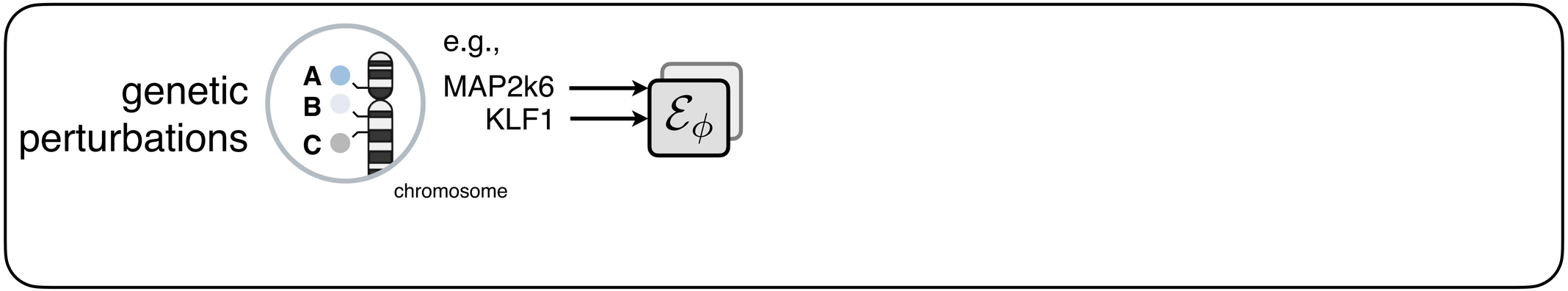
Results on Predicting Combinations of Perturbations

genetic perturbations

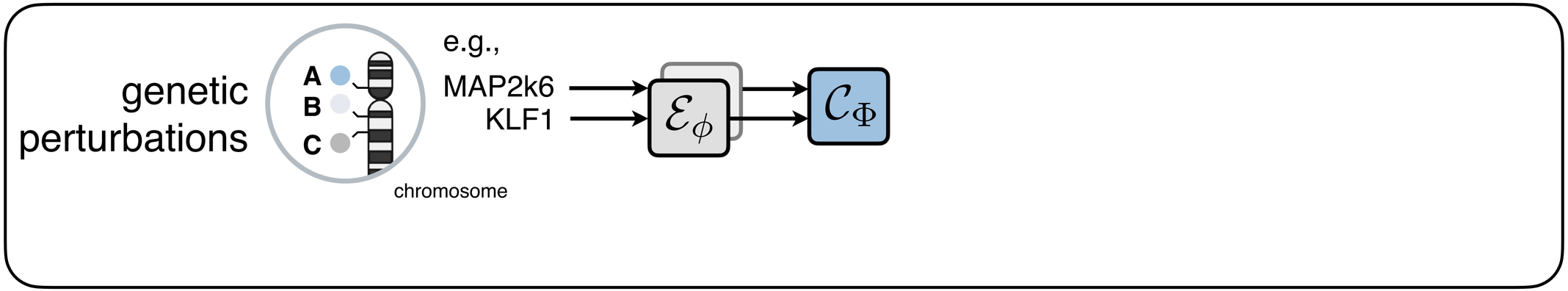


chromosome

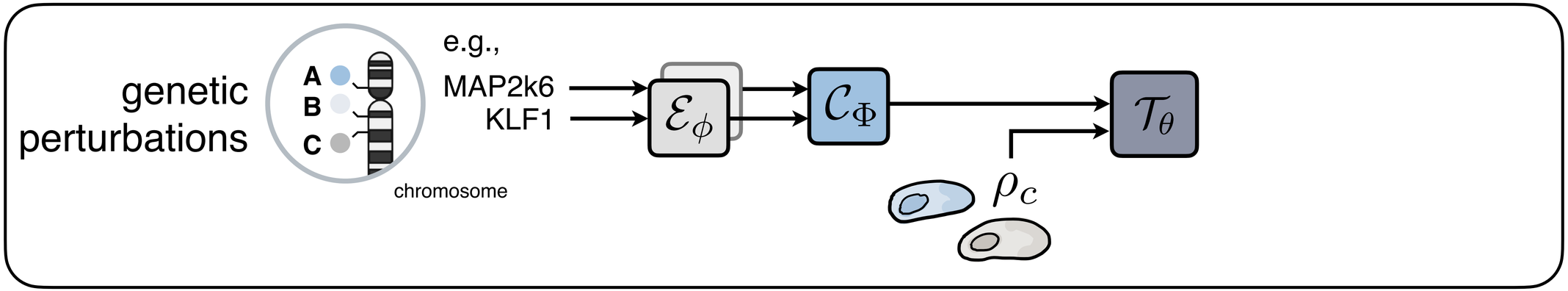
Results on Predicting Combinations of Perturbations



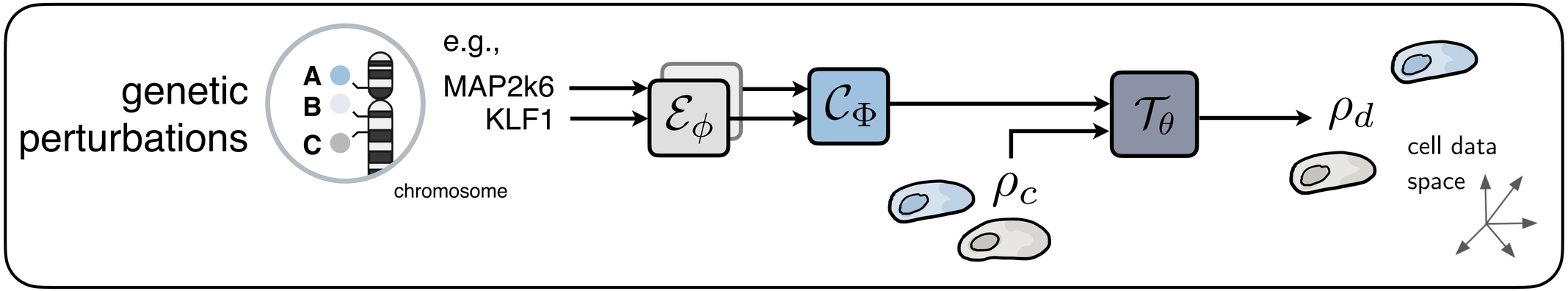
Results on Predicting Combinations of Perturbations



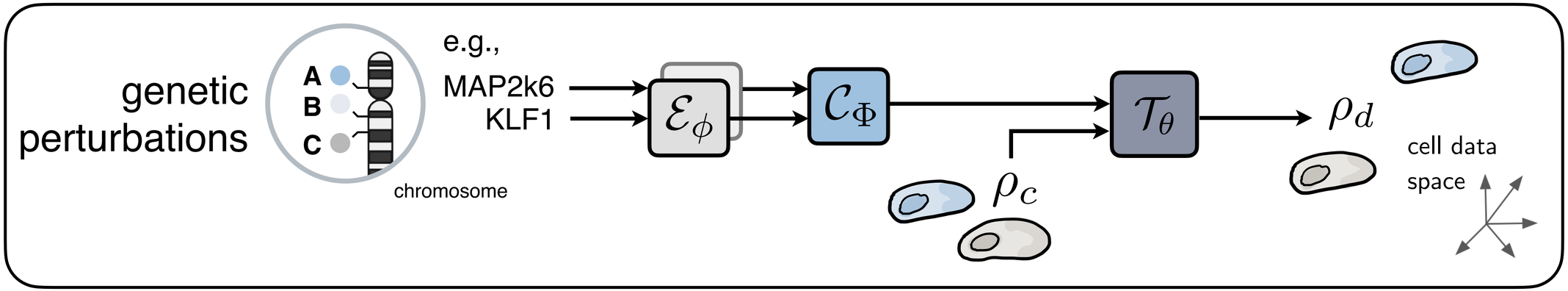
Results on Predicting Combinations of Perturbations

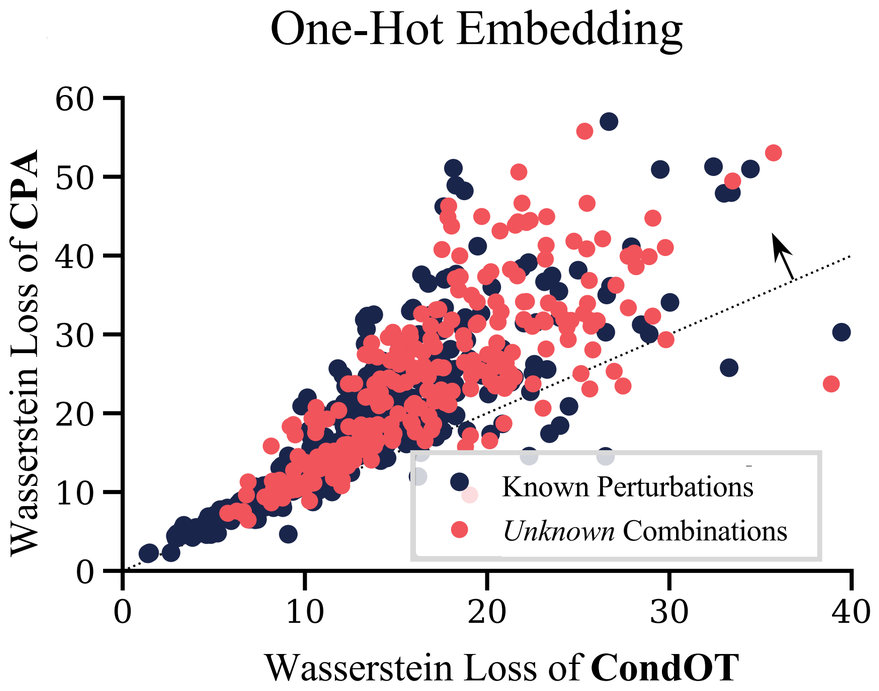


Results on Predicting Combinations of Perturbations

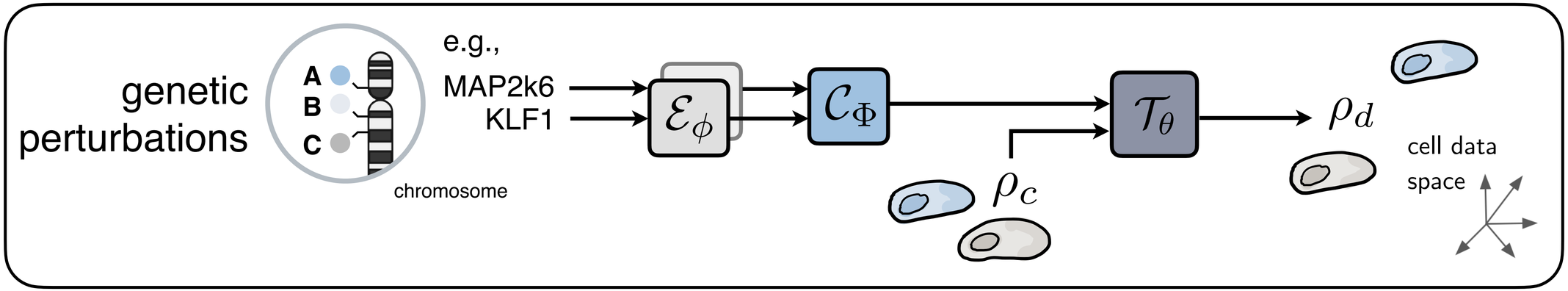


Results on Predicting Combinations of Perturbations

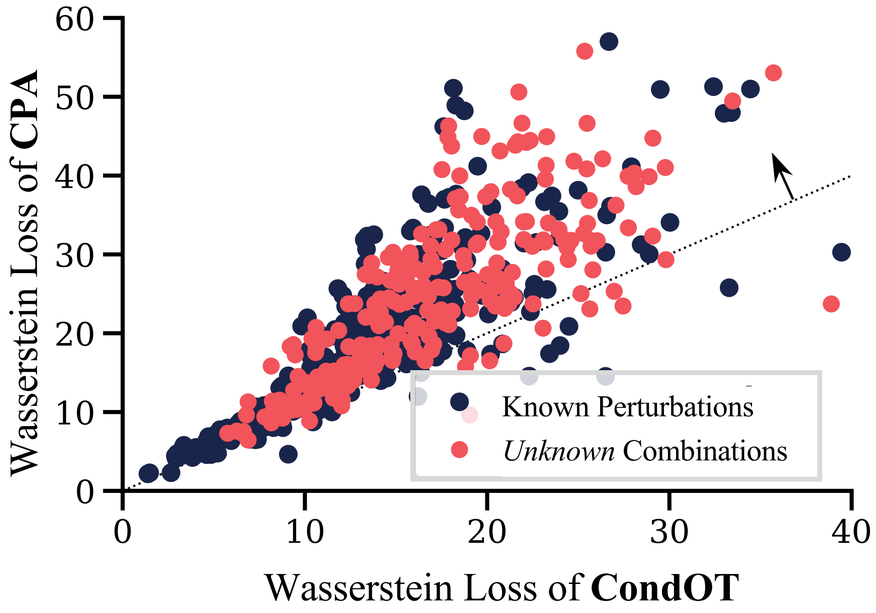




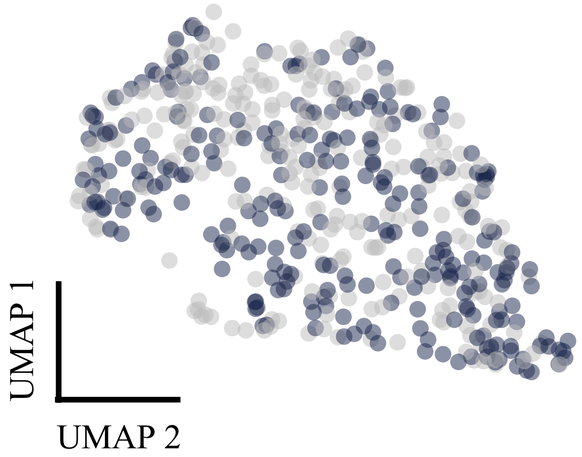
Results on Predicting Combinations of Perturbations



One-Hot Embedding

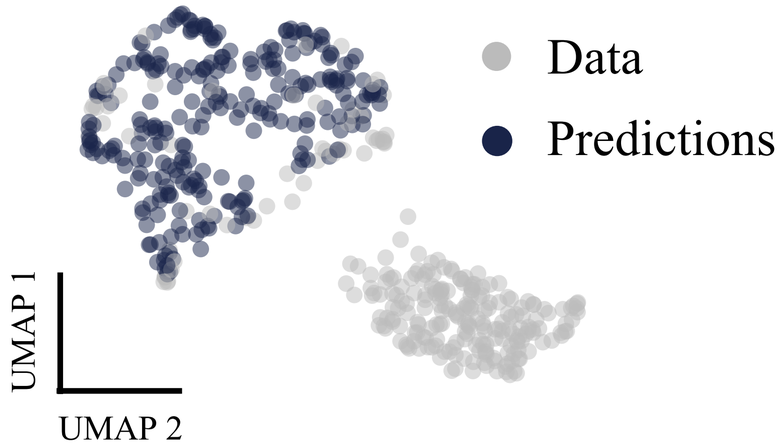


CondOT

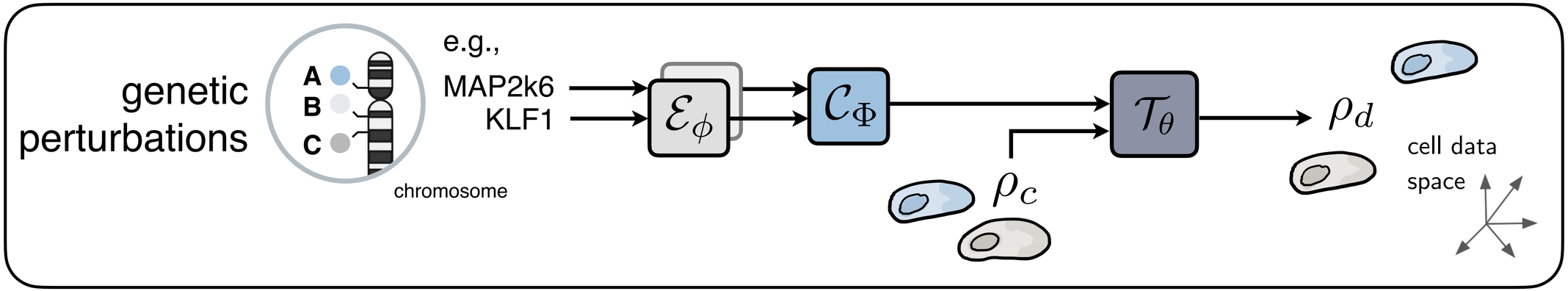


CPA

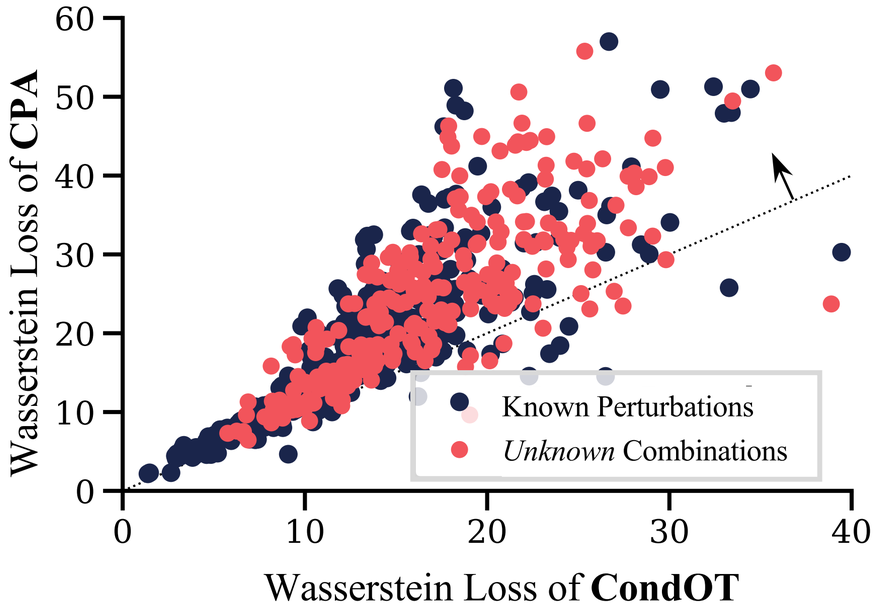
(Lotfollahi et al., 2021)



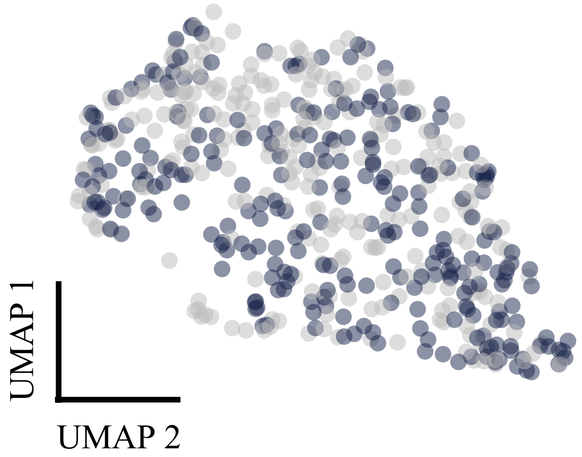
Results on Predicting Combinations of Perturbations



One-Hot Embedding

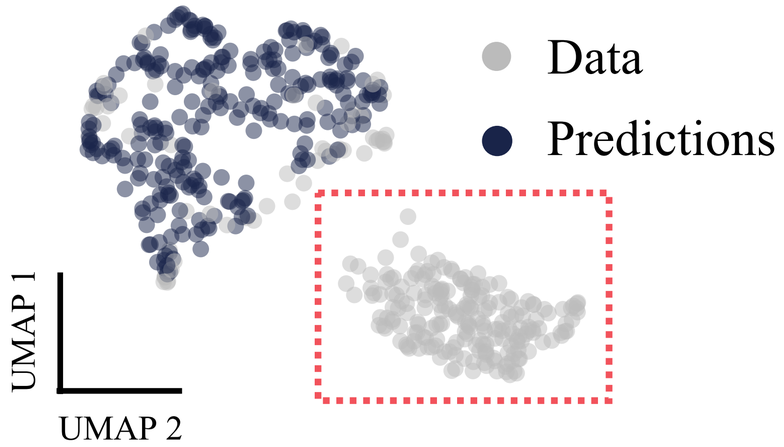


CondOT



CPA

(Lotfollahi et al., 2021)



The Neural Approach to OT: Use Bernier Theorem

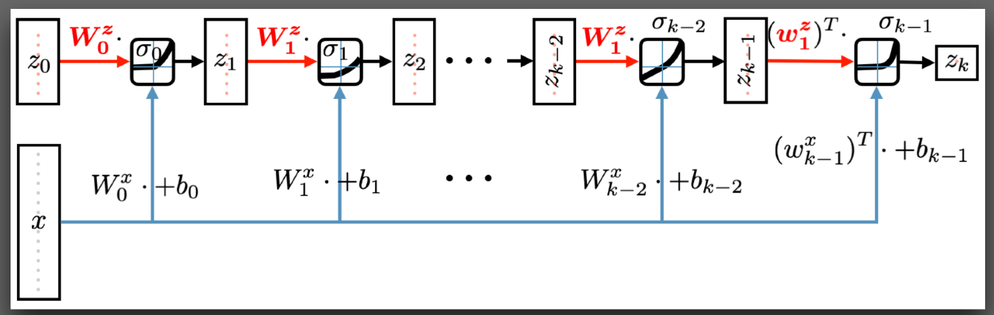
$$T^\star(\mathbf{x}) = \mathbf{x} - \nabla h^* \circ \nabla f^\star(\mathbf{x}).$$

The Neural Approach to OT: Use Bernier Theorem

$$T^\star(\mathbf{x}) = \mathbf{x} - \nabla h^* \circ \nabla f^\star(\mathbf{x}).$$

$$h(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|_2^2; \nabla h = \nabla h^* = \mathrm{Id}$$

Thanks to the Bernier theorem, one can use an OT point of view by parameterizing *T* as the *gradient* of an input convex NN.

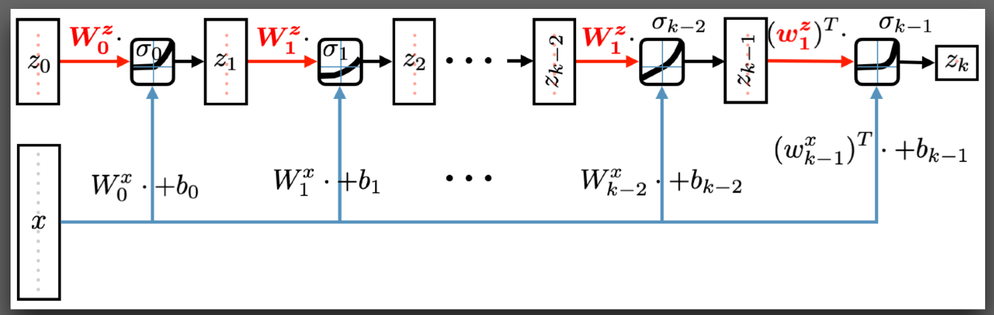


The Monge Gap: A regularizer to compute “OT like” maps.

$$T^\star(\mathbf{x}) = \mathbf{x} - \nabla h^* \circ \nabla f^\star(\mathbf{x}).$$

$$h(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|_2^2; \nabla h = \nabla h^* = \mathrm{Id}$$

Thanks to the Bernier theorem, one can use an OT point of view by parameterizing *T* as the *gradient* of an input convex NN.

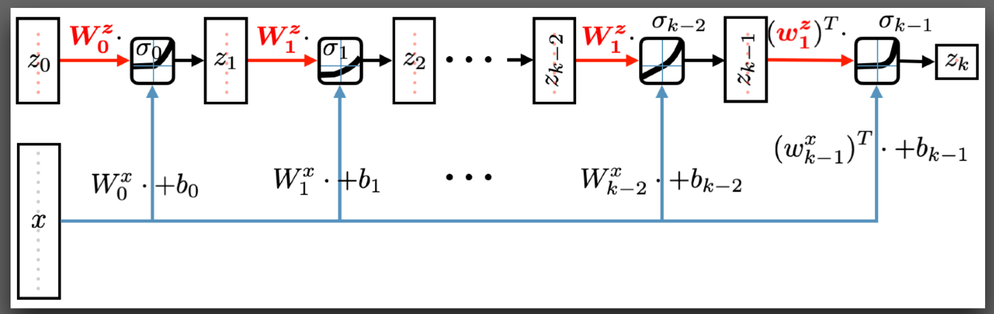


The Monge Gap: A regularizer to compute "OT like" maps.

$$T^\star(\mathbf{x}) = \mathbf{x} - \nabla h^* \circ \nabla f^\star(\mathbf{x}).$$

$$h(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|_2^2; \nabla h = \nabla h^* = \mathrm{Id}$$

Thanks to the Bernier theorem, one can use an OT point of view by parameterizing *T* as the *gradient* of an input convex NN.

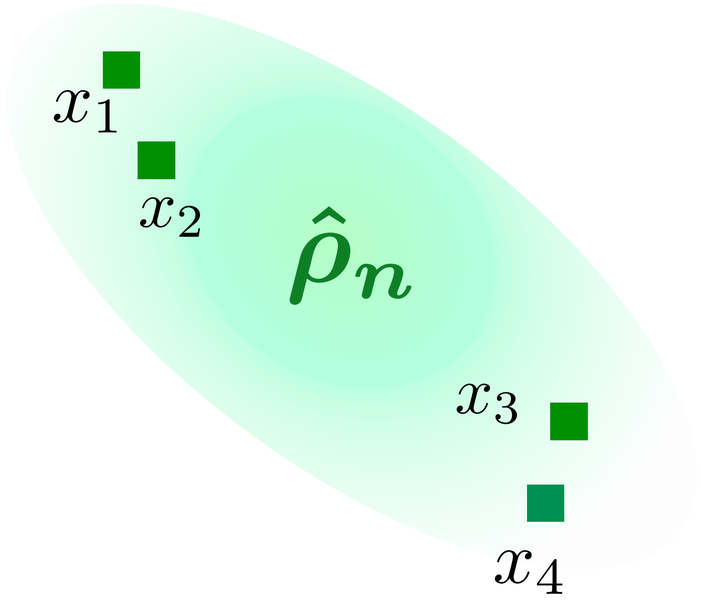


Parameterize *T* as as *any* neural network architecture, for *any cost*.

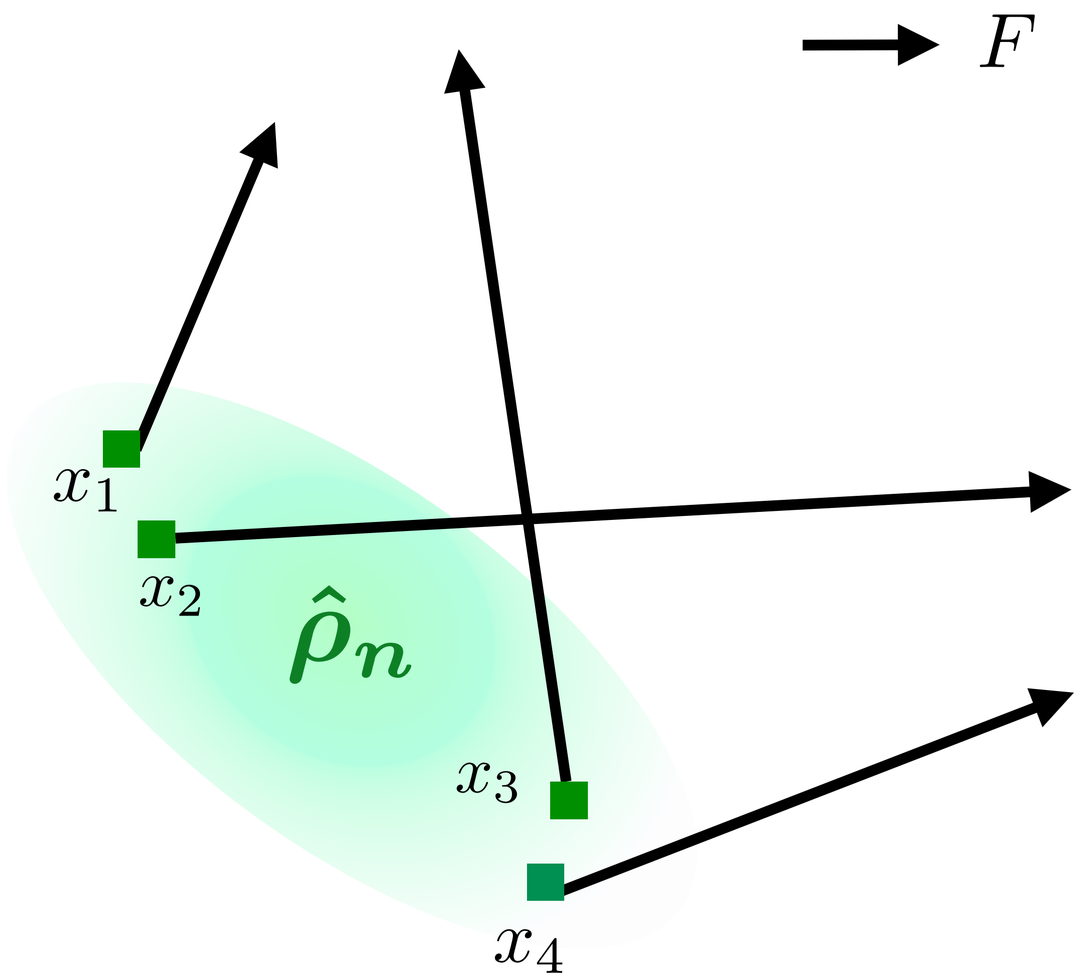
Add this simple regularization on top the training loss to promote a NN map *F* that *"looks"* like an OT map. Valid for *any* cost

$$\mathcal{M}_\rho^c(F) := \int_\Omega c(\mathbf{x}, F(\mathbf{x}))\mathrm{d}\rho(\mathbf{x}) - W_c(\rho, F\sharp\rho)\,.$$

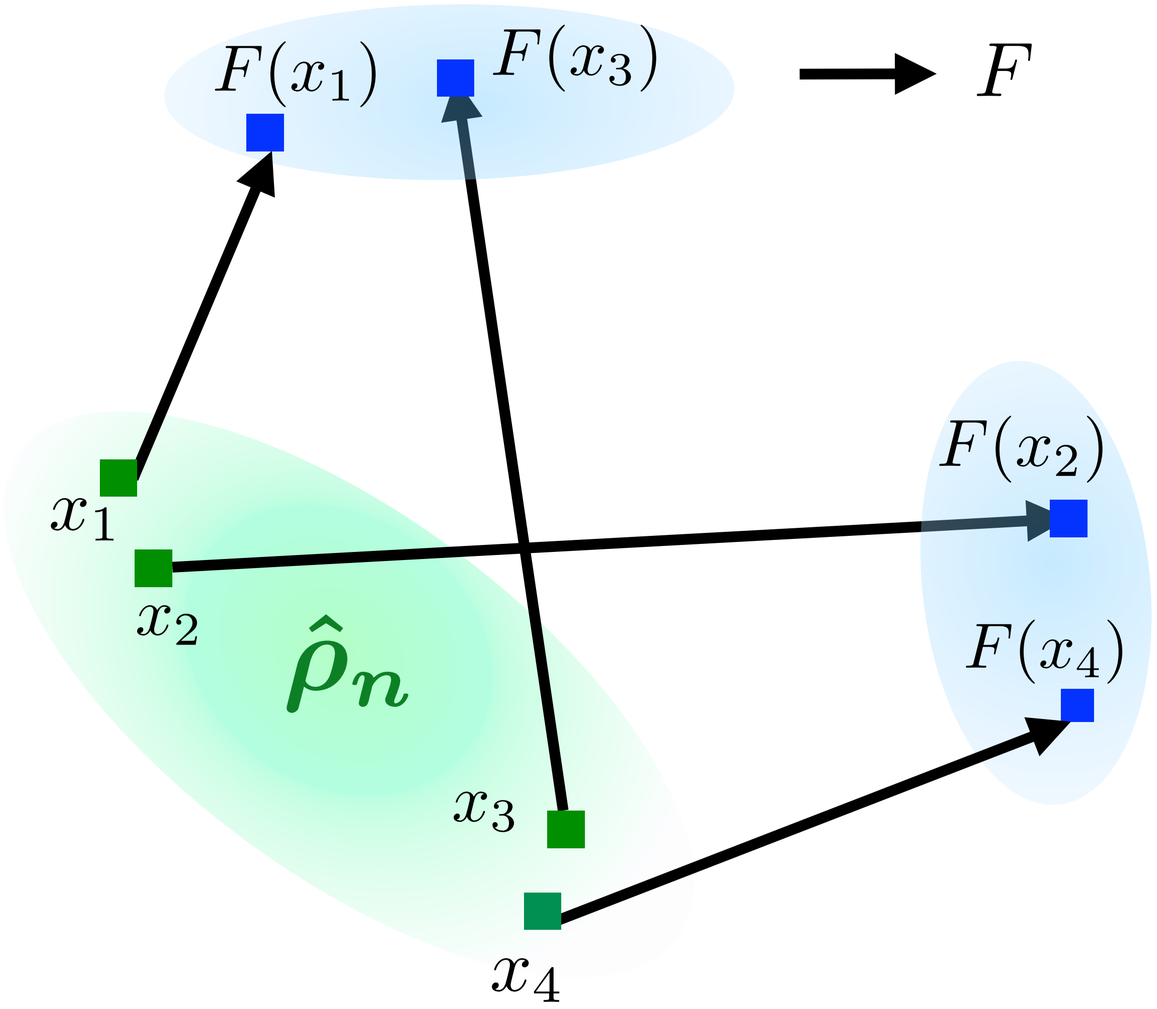
The Monge Gap: A regularizer to compute “OT like” maps.



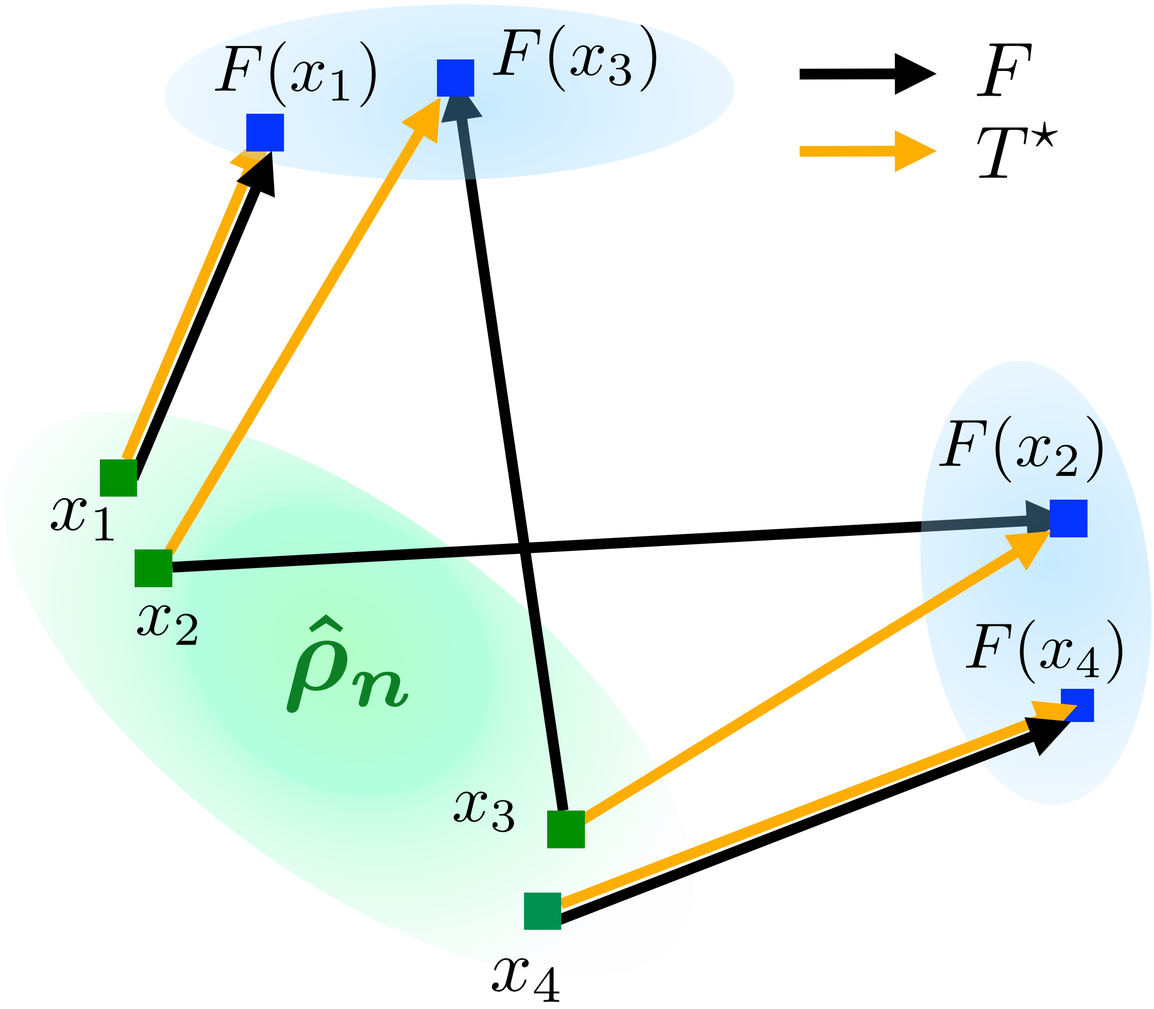
The Monge Gap: A regularizer to compute “OT like” maps.



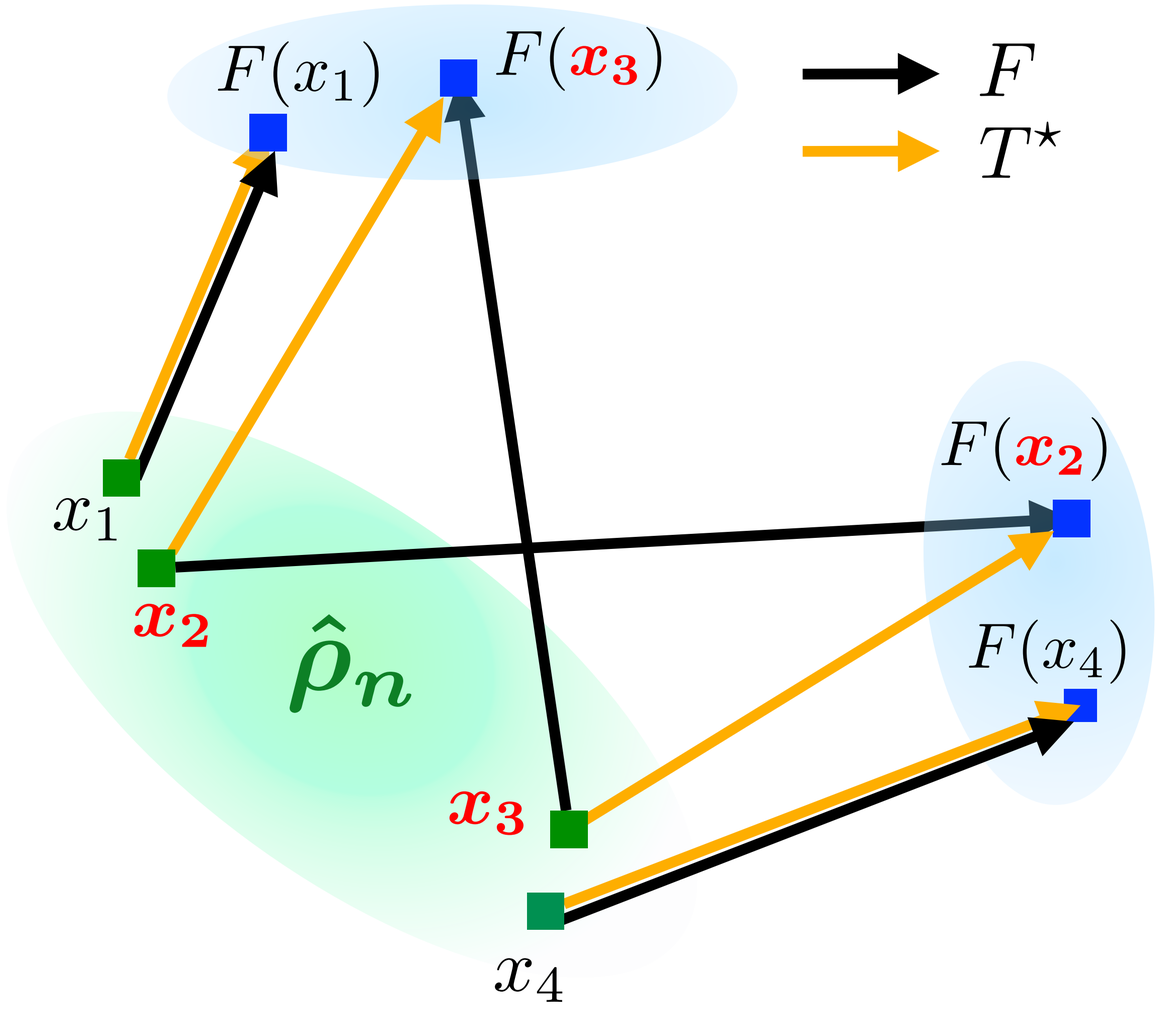
The Monge Gap: A regularizer to compute “OT like” maps.



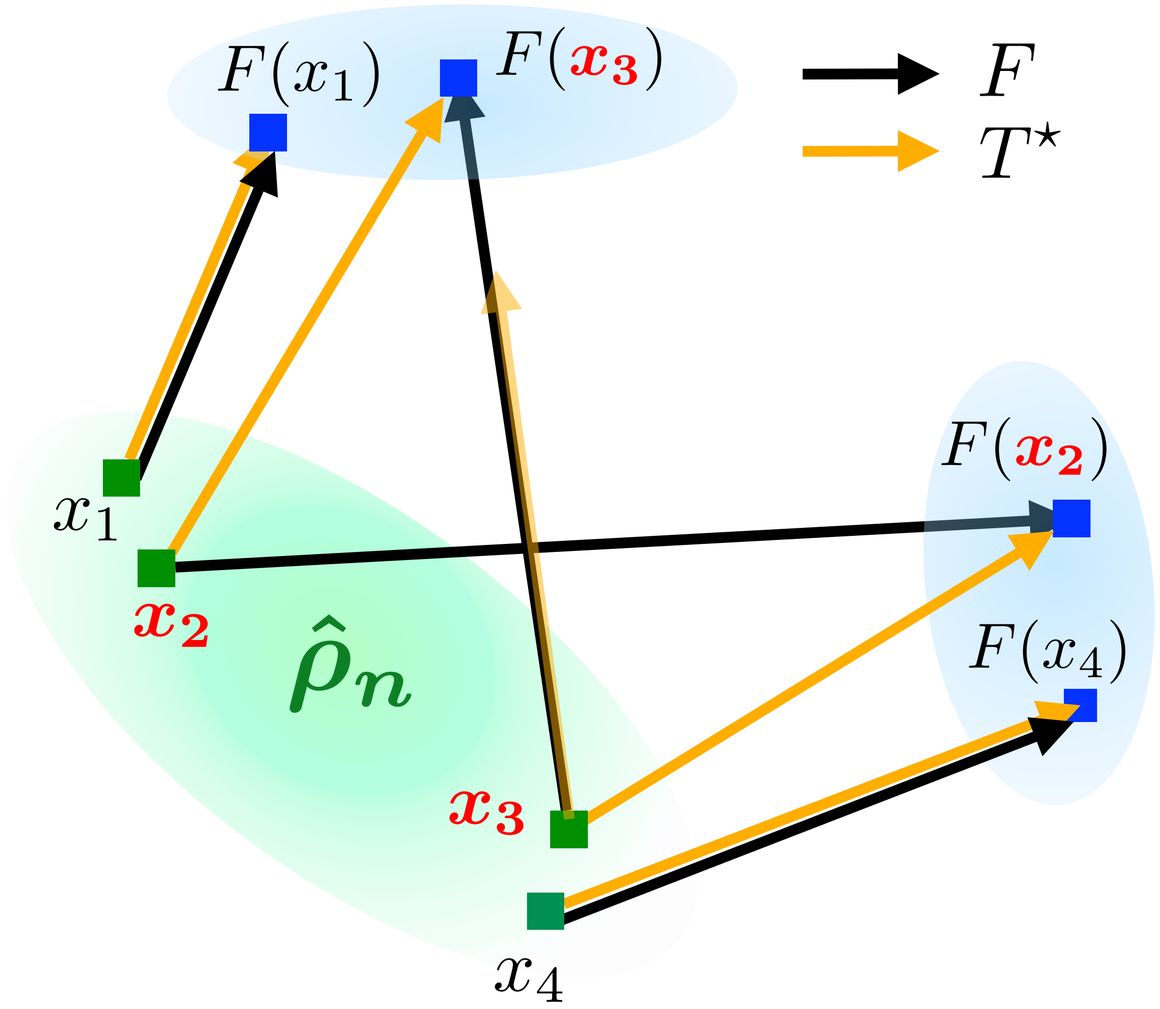
The Monge Gap: A regularizer to compute “OT like” maps.



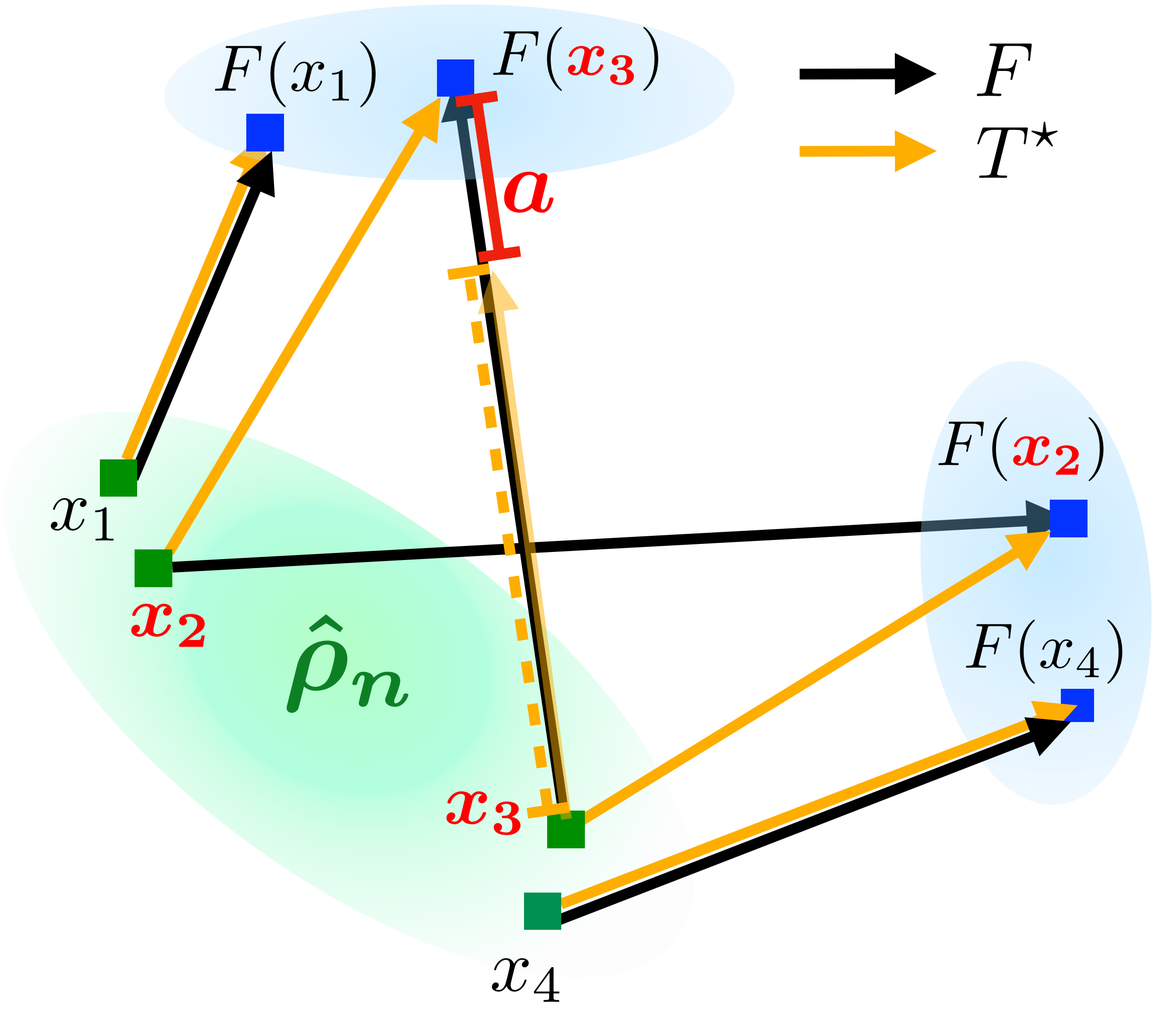
The Monge Gap: A regularizer to compute “OT like” maps.



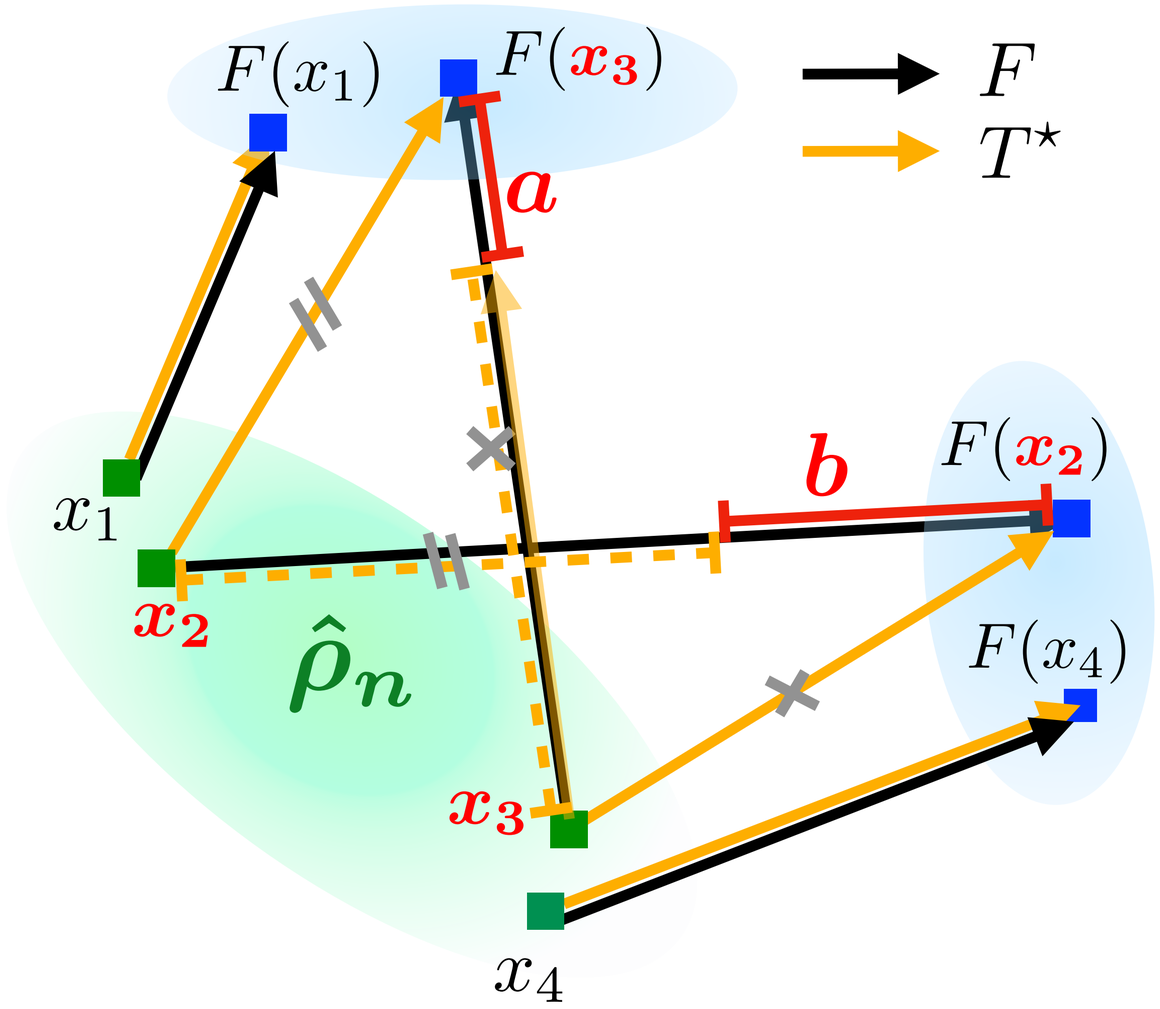
The Monge Gap: A regularizer to compute “OT like” maps.



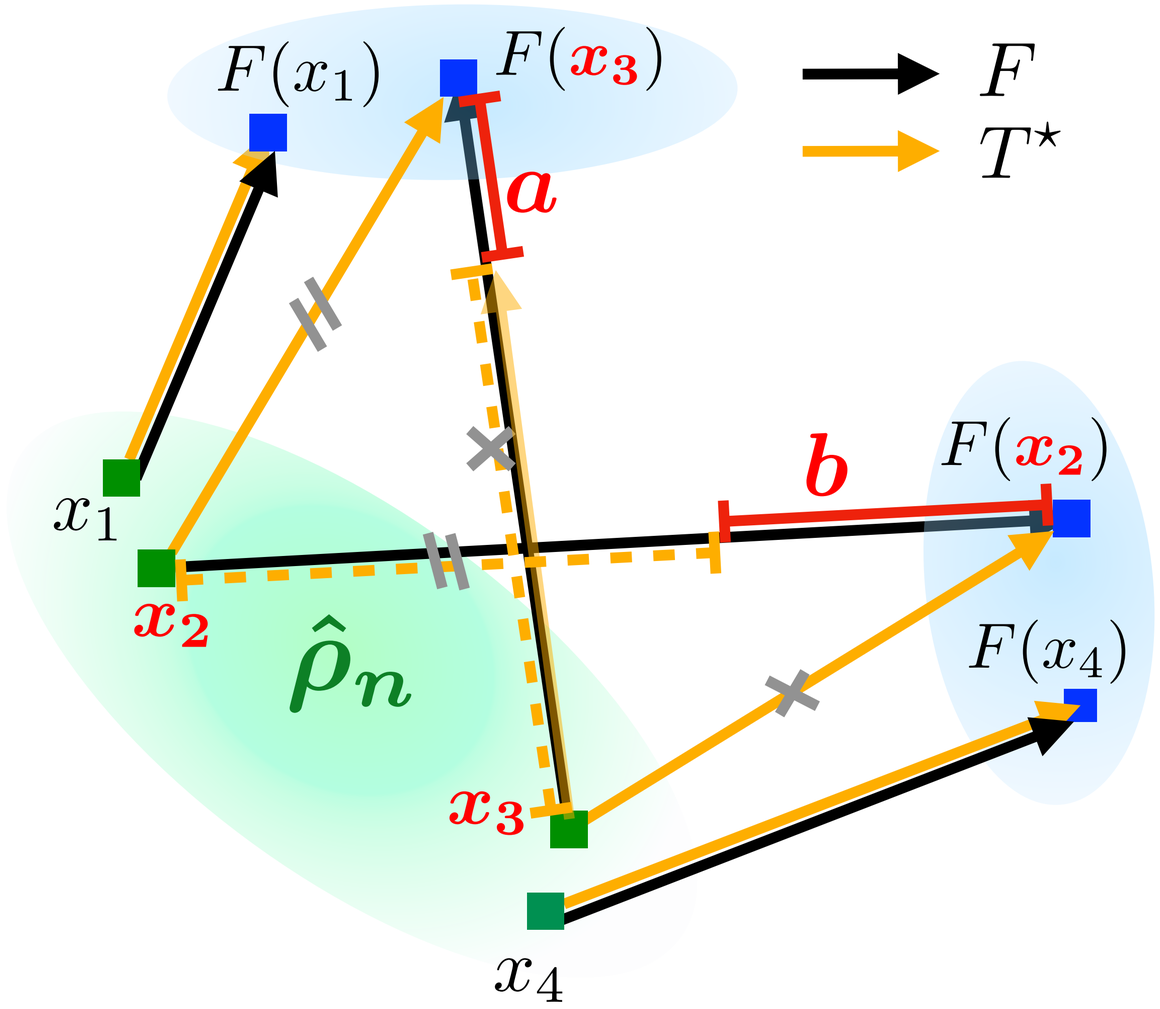
The Monge Gap: A regularizer to compute “OT like” maps.



The Monge Gap: A regularizer to compute “OT like” maps.



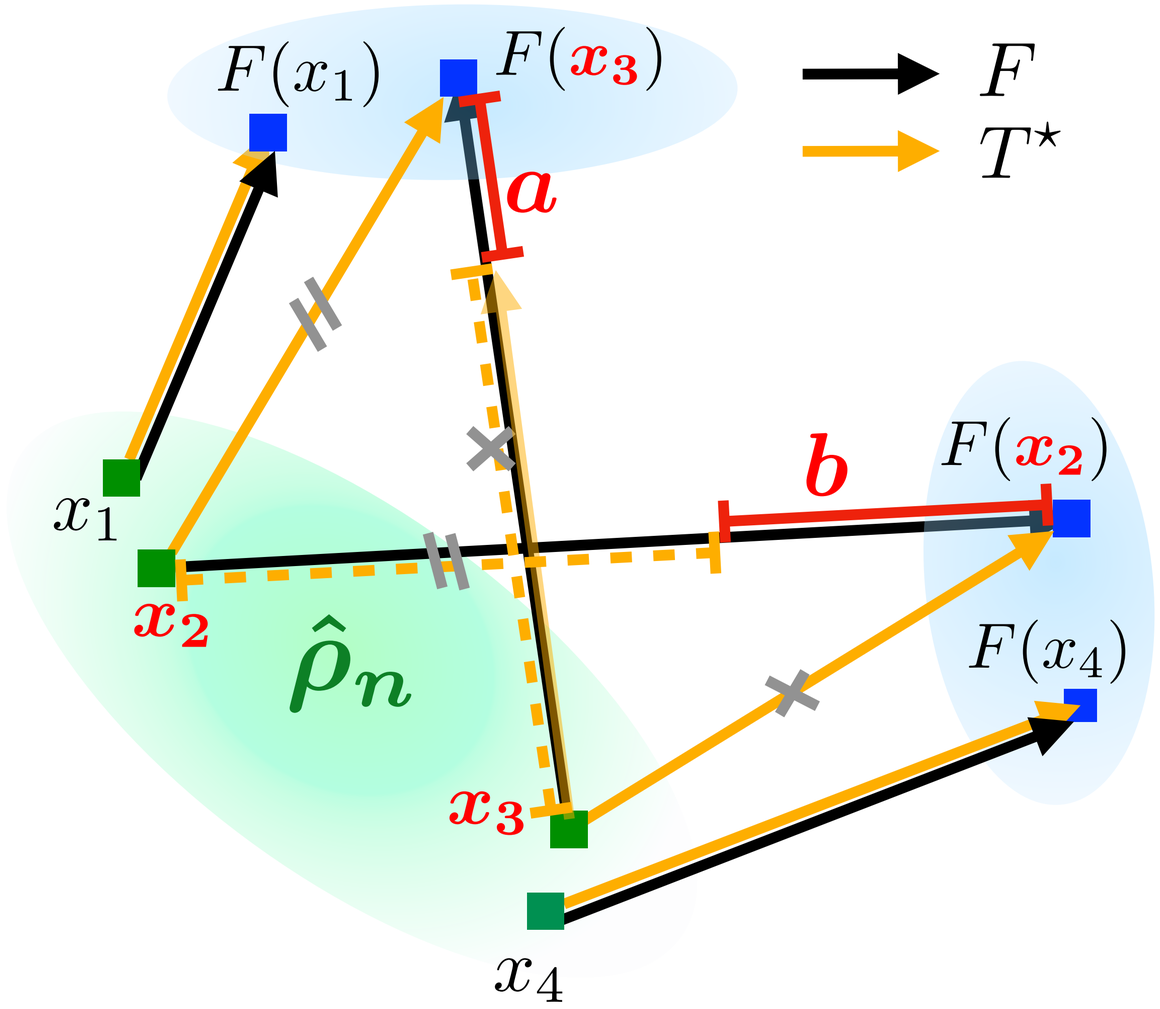
The Monge Gap: A regularizer to compute “OT like” maps.



Conceptually:

$$\mathcal{M}^1_{\hat{\rho}_n}(F) = \frac{1}{4}(a + b)$$

The Monge Gap: A regularizer to compute "OT like" maps.



Conceptually:

$$\mathcal{M}^1_{\hat{\rho}_n}(F) = \frac{1}{4}(a + b)$$

In practice, this formula is equivalent but can be computed/differentiated efficiently:

$$\mathcal{M}^c_\rho(F) := \int_\Omega c(\mathbf{x}, F(\mathbf{x}))\mathrm{d}\rho(\mathbf{x}) - W_c(\rho, F\sharp\rho)\,.$$

Properties

- Monge gap is ≥0, and 0 iff the map is optimal.
- Monge gap approximated using Sinkhorn algorithm remains ≥0.
- For squared-Euclidean costs:
  - Monge Gap is convex
  - Approximation using entropic regularization is too.
- Easy to compute/differentiate with OTT-JAX (using Danskin)

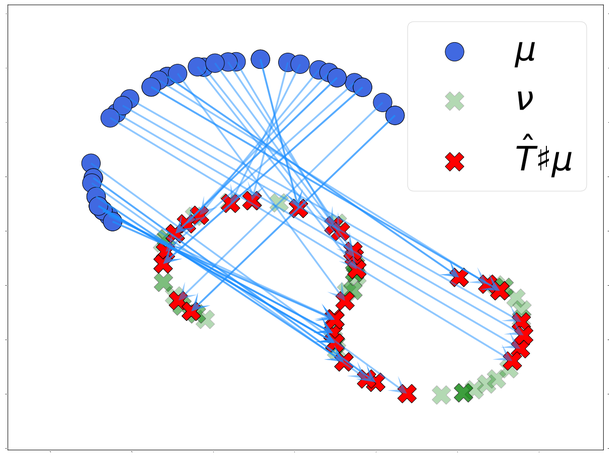


Monge Gap

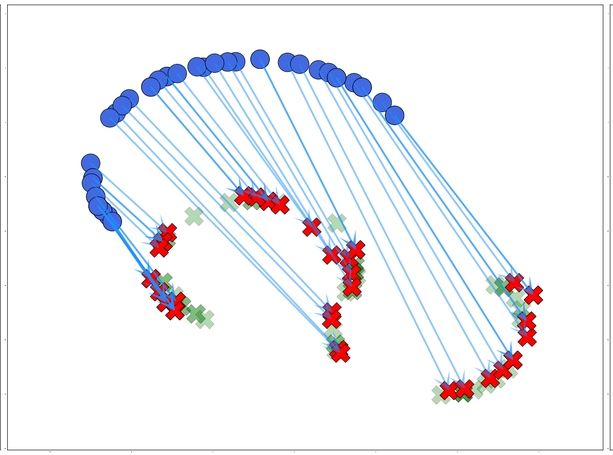
The Monge Gap in Practice

$$\mathcal{M}_\rho^c(T) := \int_\Omega c(\mathbf{x}, T(\mathbf{x}))\mathrm{d}\rho(\mathbf{x}) - W_c(\rho, T\sharp\rho)\,.$$

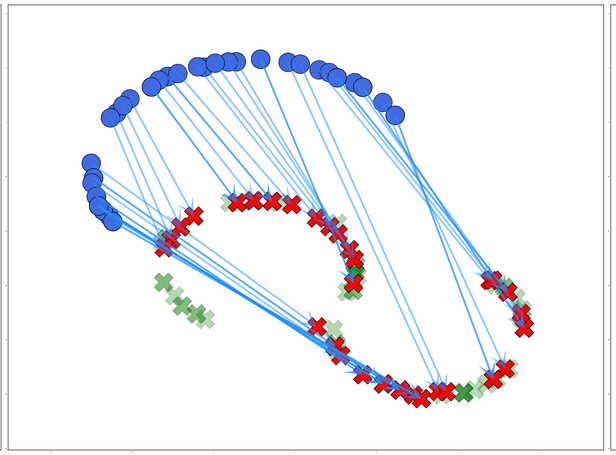
No Monge Gap



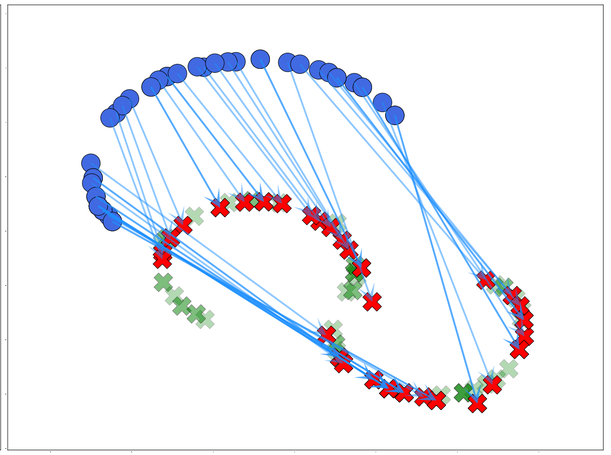
Monge Gap: $c(x, y) = |x - y|_2$



Monge Gap: $c(x, y) = \frac{1}{1.5}|x - y|_{1.5}^{1.5}$

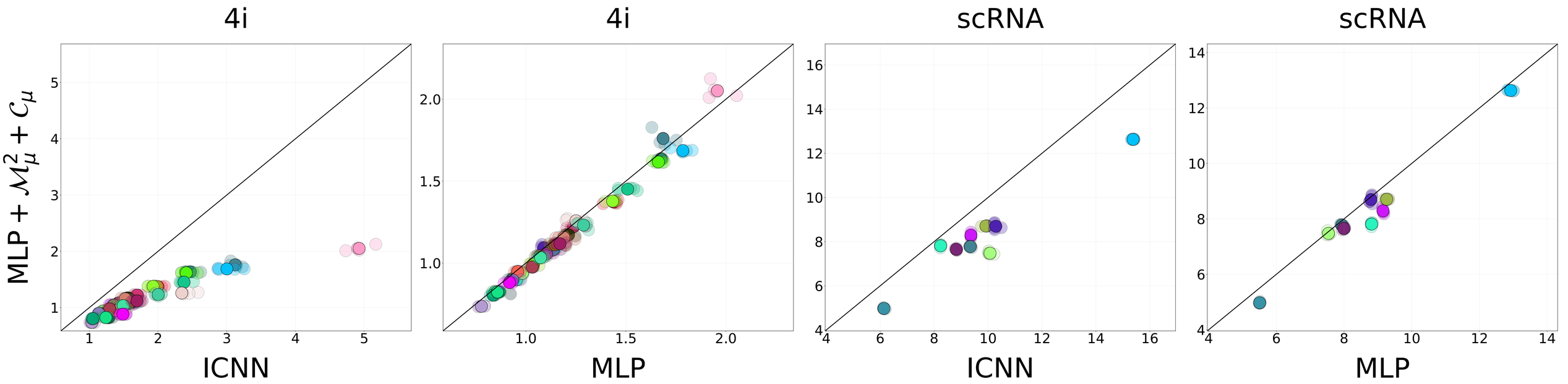


Monge Gap: $c(x, y) = \frac{1}{2}|x - y|_2^2$



The Monge Gap: A regularizer to compute OT like maps.

$$\mathcal{M}_\rho^c(T) := \int_\Omega c(\mathbf{x}, T(\mathbf{x}))\mathrm{d}\rho(\mathbf{x}) - W_c(\rho, T\sharp\rho)\,.$$



Simpler Attempt using Sinkhorn

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

- Use two samples of points from each measure + regularization parameter,

$$\mathbf{x_1}, \ldots, \mathbf{x}_{\boldsymbol{n}}, \mathbf{y_1}, \ldots, \mathbf{y}_{\boldsymbol{m}}, \gamma$$

- Run Sinkhorn algorithm to get $\boldsymbol{f}_{\boldsymbol{i}}^\star, \boldsymbol{g}_{\boldsymbol{j}}^\star$ **point-wise** estimates for dual values (these correspond to the $\boldsymbol{\alpha}_{\boldsymbol{i}}^\star, \boldsymbol{\beta}_{\boldsymbol{j}}^\star$ values in log-space returned by Sinkhorn).

- Extend these values ***out-of-sample***, with a "soft" h-transform, to form the estimator:

$$\hat{\boldsymbol{f}}_{\boldsymbol{n}}(\mathbf{x}) = -\gamma \log \sum_j e^{\frac{-\boldsymbol{h}(\mathbf{x}-\mathbf{y}_j)+\boldsymbol{g}_j^\star}{\gamma}}$$

[Pooladian+21]

Monge / Bregman Entropic Maps

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

$$\nabla \hat{\boldsymbol{f}}_{\boldsymbol{n}}(\mathbf{x}) = \sum_j p_j(\mathbf{x}) \nabla \boldsymbol{h}(\mathbf{x} - \mathbf{y}_{\boldsymbol{j}}) \qquad p_j(\mathbf{x}) := \frac{e^{(-\boldsymbol{h}(\mathbf{x}-\mathbf{y}_{\boldsymbol{j}})+\boldsymbol{g}_{\boldsymbol{j}}^\star)/\gamma}}{\sum_k e^{(-\boldsymbol{h}(\mathbf{x}-\mathbf{y}_{\boldsymbol{k}})+\boldsymbol{g}_{\boldsymbol{k}}^\star)/\gamma}}$$

Monge / Bregman Entropic Maps

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

$$\nabla \hat{\boldsymbol{f}}_{\boldsymbol{n}}(\mathbf{x}) = \sum_j p_j(\mathbf{x}) \nabla \boldsymbol{h}(\mathbf{x} - \mathbf{y}_{\boldsymbol{j}}) \qquad p_j(\mathbf{x}) := \frac{e^{(-\boldsymbol{h}(\mathbf{x}-\mathbf{y}_{\boldsymbol{j}}) + \boldsymbol{g}_{\boldsymbol{j}}^\star)/\gamma}}{\sum_k e^{(-\boldsymbol{h}(\mathbf{x}-\mathbf{y}_{\boldsymbol{k}}) + \boldsymbol{g}_{\boldsymbol{k}}^\star)/\gamma}}$$

$$D_{\boldsymbol{h}}(\mathbf{z}|\mathbf{w}) = \boldsymbol{h}(\mathbf{z}) - \boldsymbol{h}(\mathbf{w}) - \langle \nabla \boldsymbol{h}(\mathbf{w}), \mathbf{z} - \mathbf{w} \rangle$$ Bregman divergence

$$\nabla \boldsymbol{h}^* \left(\sum_i \lambda_i \nabla \boldsymbol{h}(\mathbf{z}_i) \right) = \operatorname*{argmin}_{\mathbf{z}} \sum_i \lambda_i D_{\boldsymbol{h}}(\mathbf{z}|\mathbf{z_i})$$ Bregman barycenter

Monge / Bregman Entropic Maps

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

$$\nabla \hat{\boldsymbol{f}}_{\boldsymbol{n}}(\mathbf{x}) = \sum_j p_j(\mathbf{x}) \nabla \boldsymbol{h}(\mathbf{x} - \mathbf{y}_{\boldsymbol{j}}) \qquad p_j(\mathbf{x}) := \frac{e^{(-\boldsymbol{h}(\mathbf{x}-\mathbf{y}_{\boldsymbol{j}})+\boldsymbol{g}_{\boldsymbol{j}}^\star)/\gamma}}{\sum_k e^{(-\boldsymbol{h}(\mathbf{x}-\mathbf{y}_{\boldsymbol{k}})+\boldsymbol{g}_{\boldsymbol{k}}^\star)/\gamma}}$$

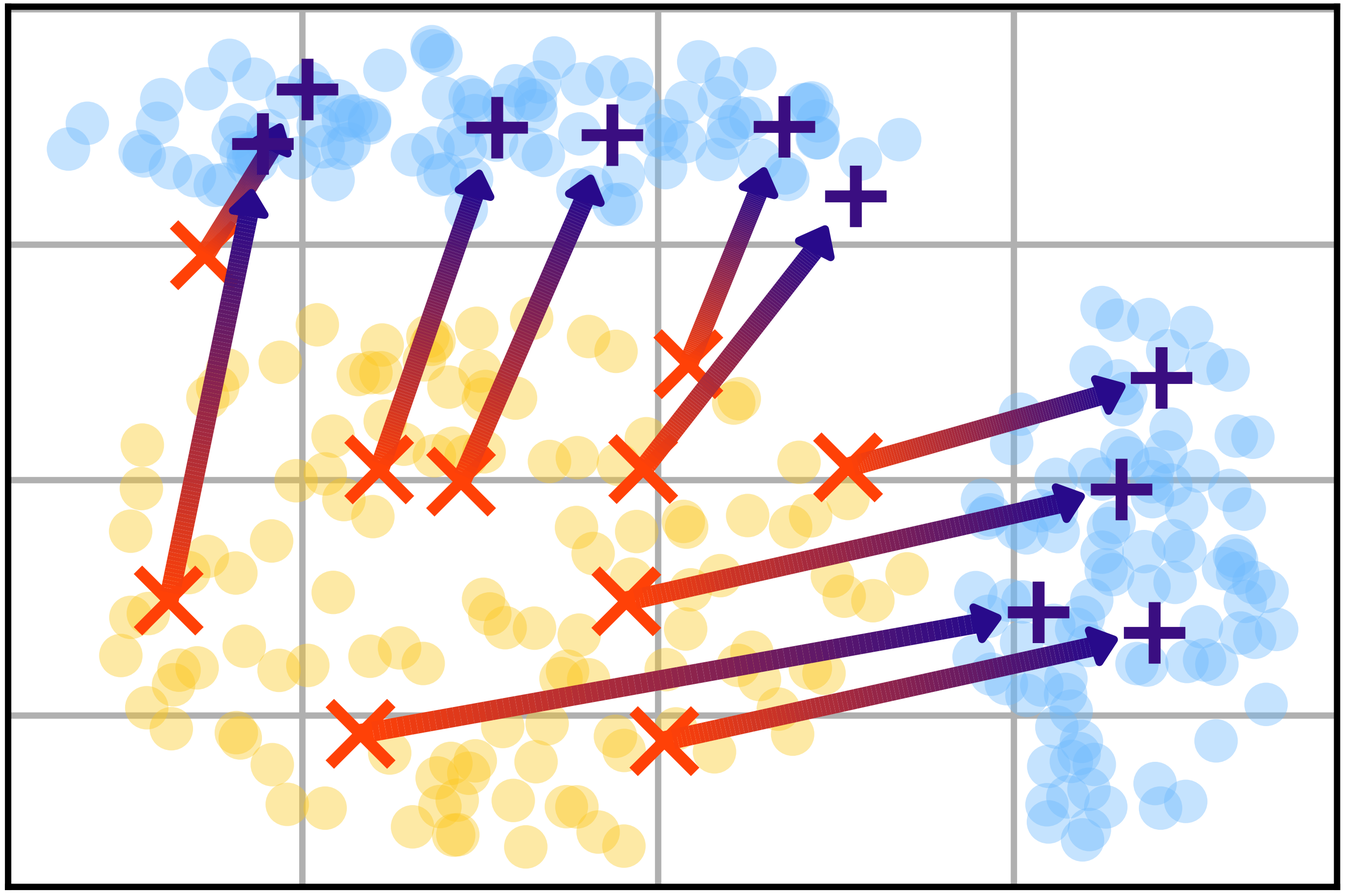
$$D_{\boldsymbol{h}}(\mathbf{z}|\mathbf{w}) = \boldsymbol{h}(\mathbf{z}) - \boldsymbol{h}(\mathbf{w}) - \langle \nabla \boldsymbol{h}(\mathbf{w}), \mathbf{z} - \mathbf{w} \rangle$$ Bregman divergence

$$\nabla \boldsymbol{h}^* \left(\sum_i \lambda_i \nabla \boldsymbol{h}(\mathbf{z}_i) \right) = \operatorname*{argmin}_{\mathbf{z}} \sum_i \lambda_i D_{\boldsymbol{h}}(\mathbf{z}|\mathbf{z}_\mathbf{i})$$ Bregman barycenter

Displacement $\nabla \boldsymbol{h}^* \circ \nabla \hat{\boldsymbol{f}}_n(\mathbf{x})$ is a Bregman barycenter

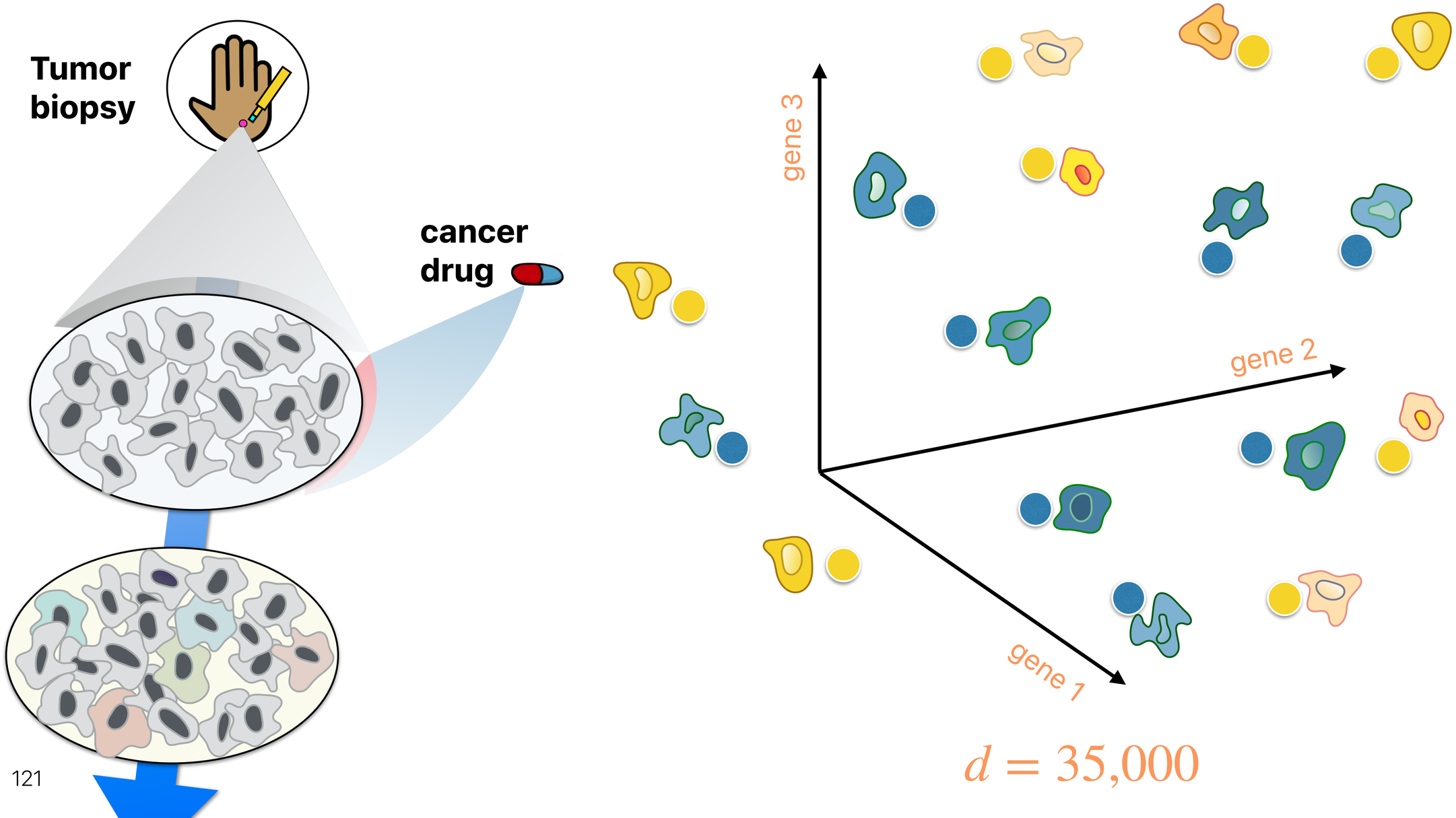
Monge / Bregman Entropic Maps

$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2$$

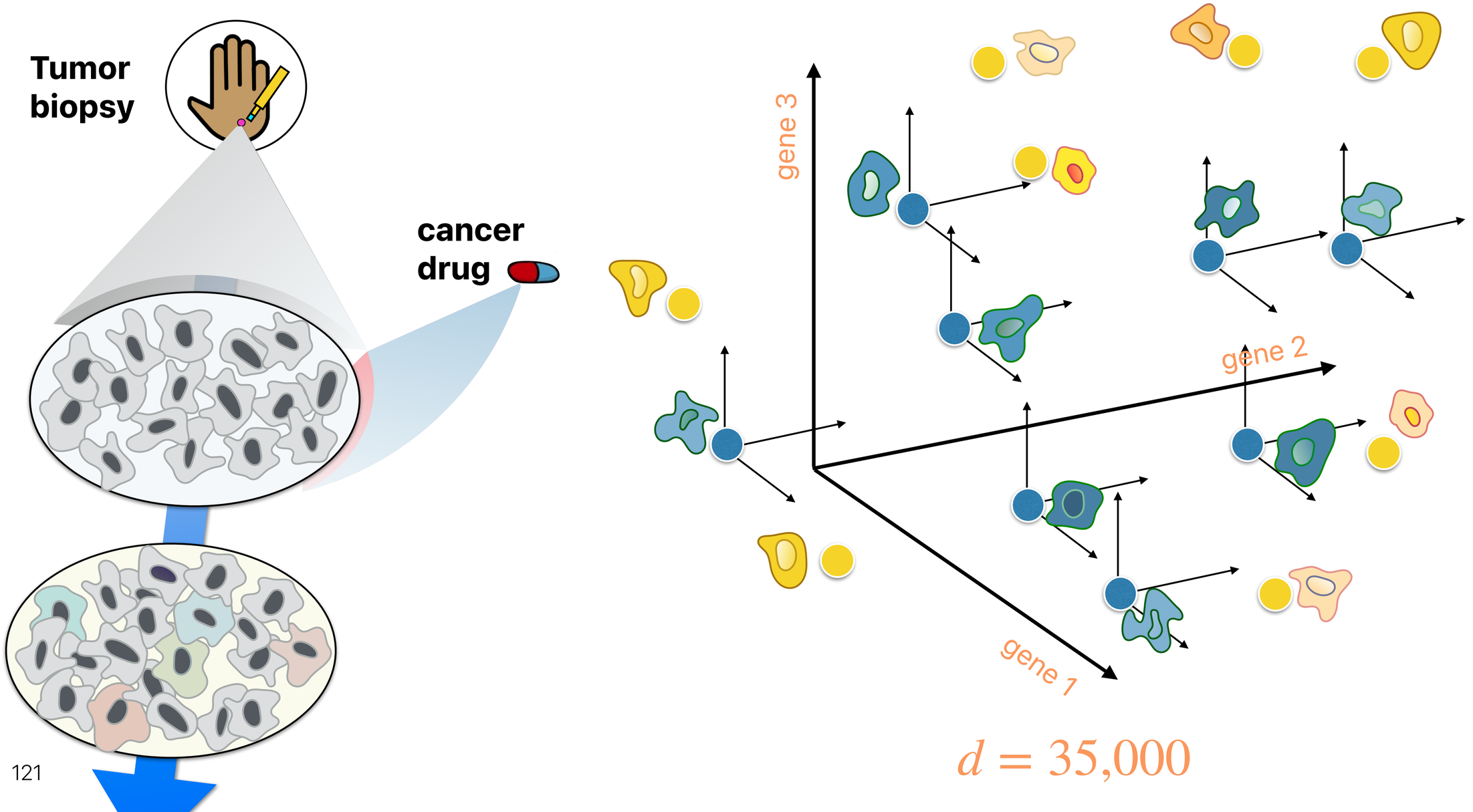


Monge / Bregman Entropic Maps

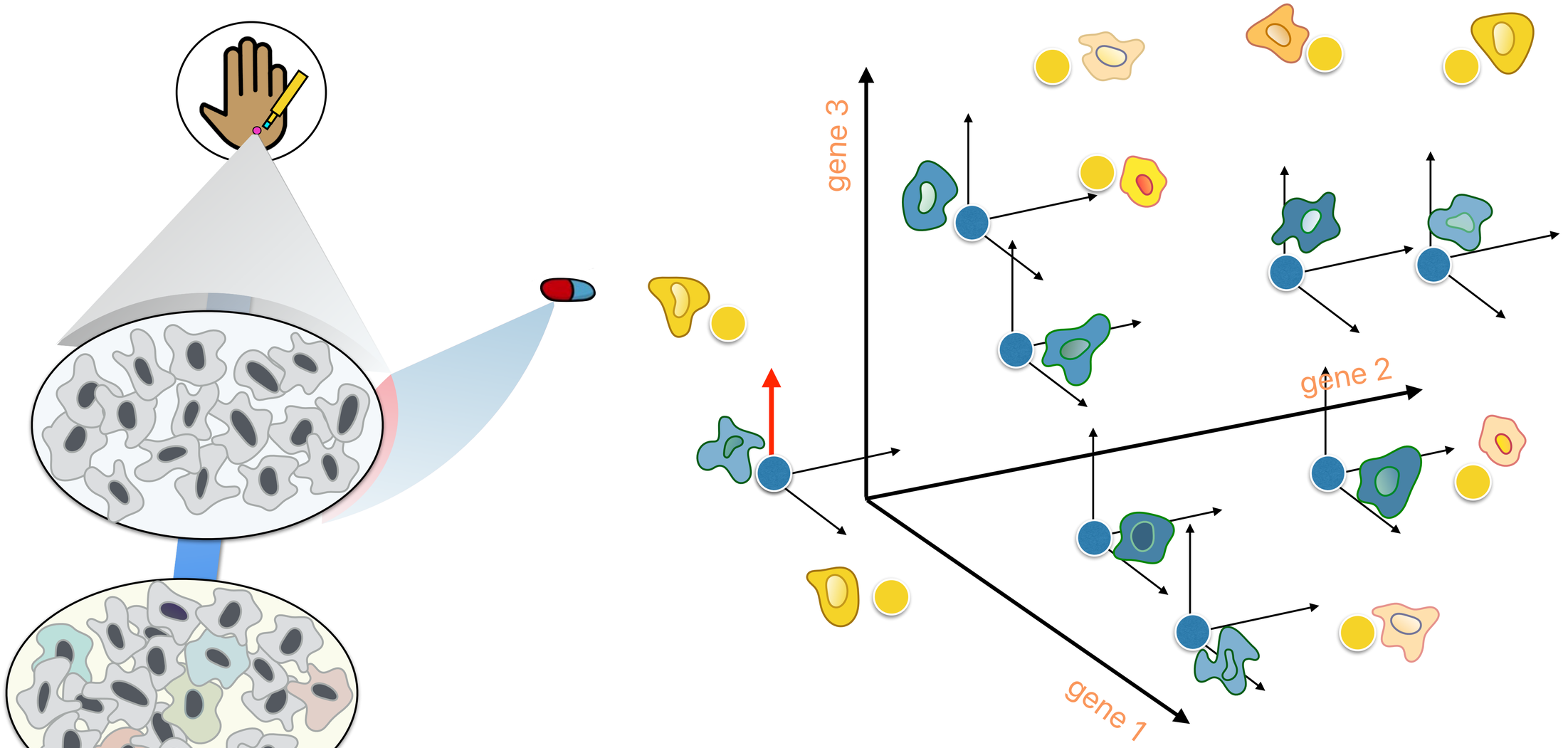
Recent Contribution: Interpretable Sparse Actions



Recent Contribution: Interpretable Sparse Actions

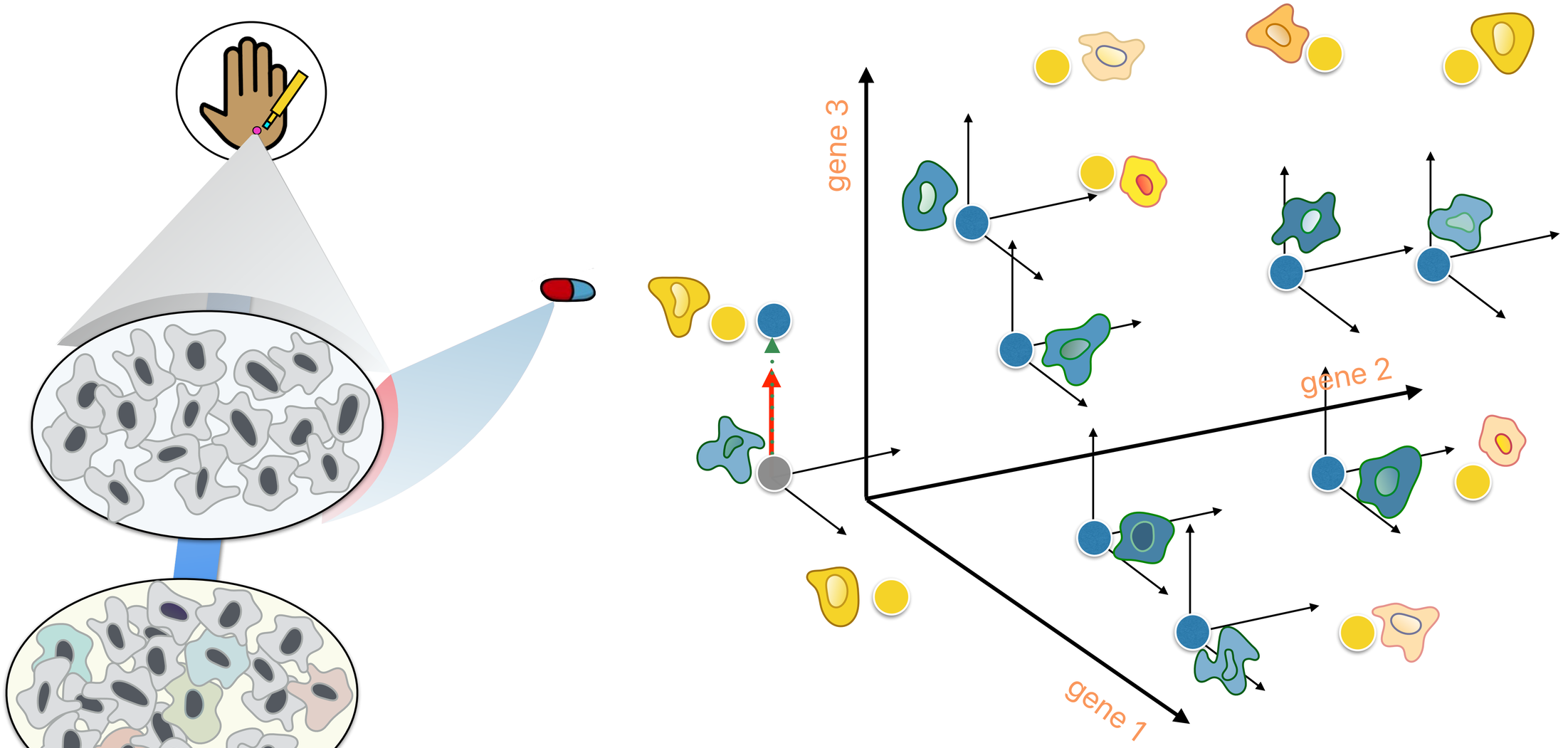


Interpretable Sparse Actions



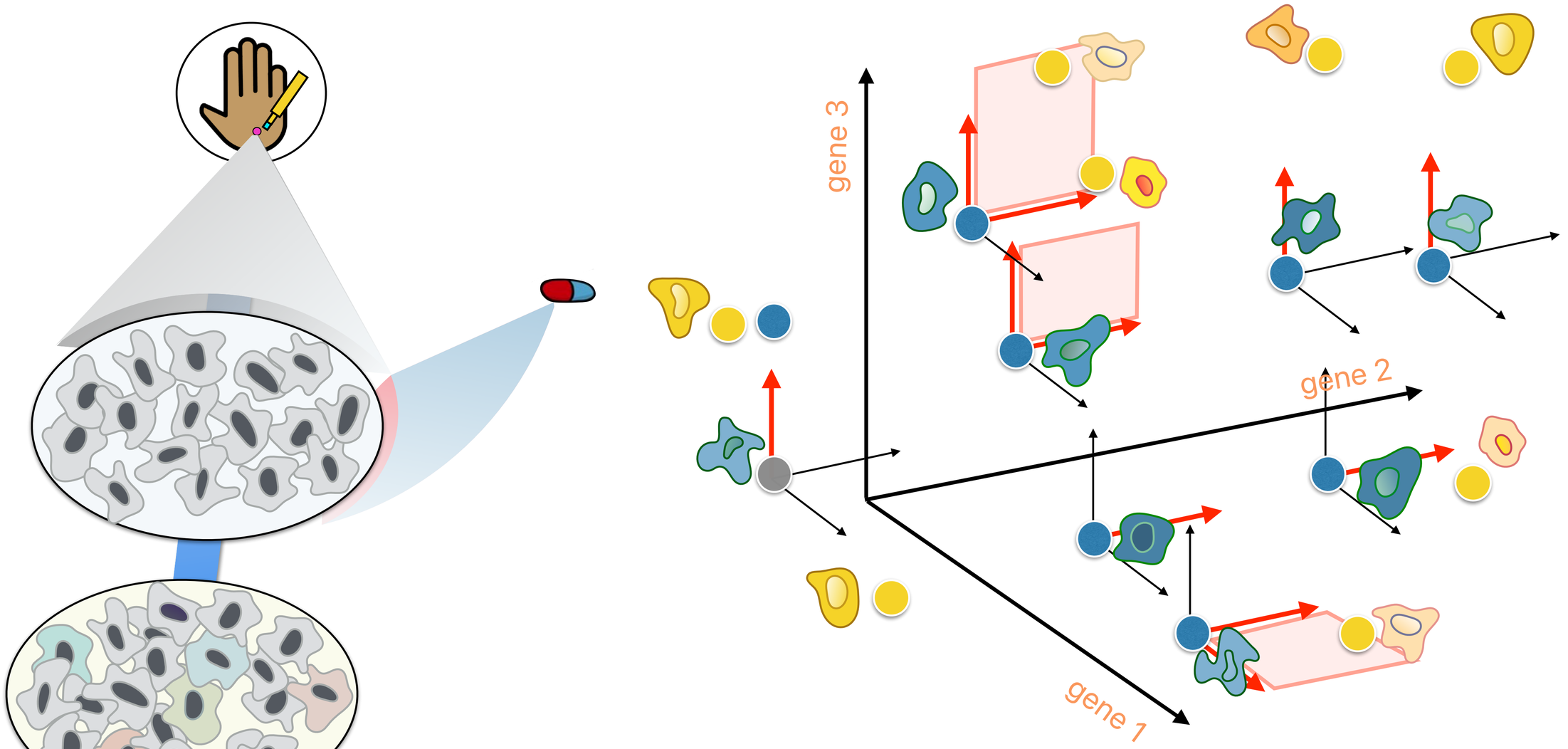
$d = 35{,}000$

Interpretable Sparse Actions



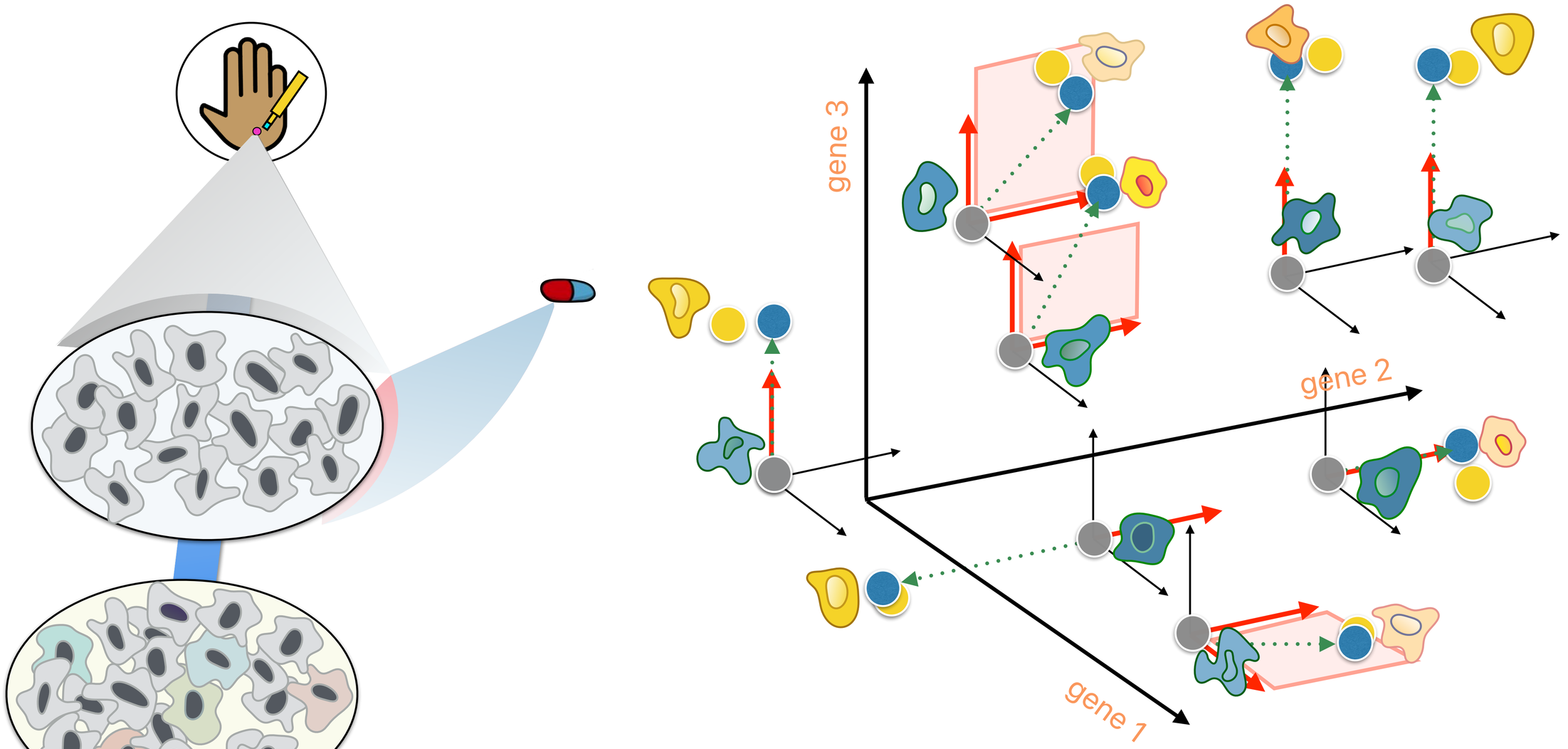
$d = 35{,}000$

Interpretable Sparse Actions



$d = 35{,}000$

Interpretable Sparse Actions



$d = 35{,}000$

Monge / Bregman / Occam : Costs *h* of *Elastic-Type*

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

If $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 + \tau(\mathbf{x})$ then $\nabla h^* = \mathrm{prox}_\tau$

where $\mathrm{prox}_\tau(\mathbf{z}) := \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{z} - \mathbf{w}\|^2 + \tau(\mathbf{w})$

Monge / Bregman / Occam : Costs *h* of *Elastic-Type*

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

$$\text{If } h(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|_2^2 + \tau(\mathbf{x}) \text{ then } \nabla h^* = \text{prox}_\tau$$

$$\text{where } \text{prox}_\tau(\mathbf{z}) := \arg\min_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{z} - \mathbf{w}\|^2 + \tau(\mathbf{w})$$

$$h^*(\mathbf{w}) = \sup_{\mathbf{z}} \mathbf{w}^T\mathbf{z} - \tfrac{1}{2}\|\mathbf{z}\|^2 - \tau(\mathbf{z})$$

Monge / Bregman / Occam : Costs *h* of *Elastic-Type*

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

$$\text{If } h(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{x}\|_2^2 + \tau(\mathbf{x}) \text{ then } \nabla h^* = \text{prox}_\tau$$

$$\text{where } \text{prox}_\tau(\mathbf{z}) := \arg\min_{\mathbf{w}} \tfrac{1}{2}\|\mathbf{z} - \mathbf{w}\|^2 + \tau(\mathbf{w})$$

$$h^*(\mathbf{w}) = \sup_{\mathbf{z}} \mathbf{w}^T\mathbf{z} - \tfrac{1}{2}\|\mathbf{z}\|^2 - \tau(\mathbf{z}) \;\; = -\inf_{\mathbf{z}} -\mathbf{w}^T\mathbf{z} + \tfrac{1}{2}\|\mathbf{z}\|^2 + \tau(\mathbf{z})$$

Monge / Bregman / Occam : Costs *h* of *Elastic-Type*

$$T^\star(\boldsymbol{x}) = \boldsymbol{x} - \nabla \boldsymbol{h}^* \circ \nabla \boldsymbol{f}^\star(\boldsymbol{x})$$

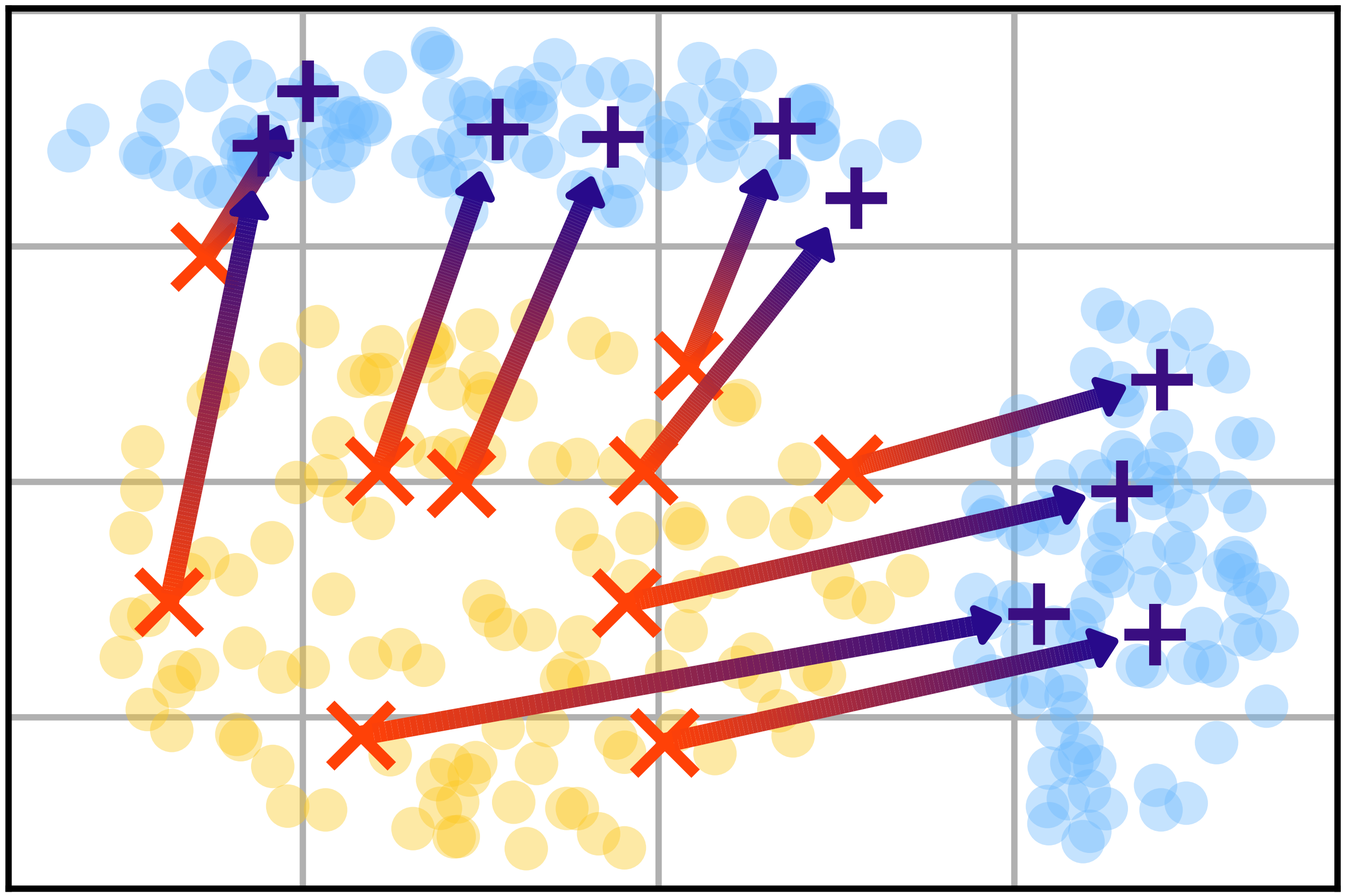
If $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 + \tau(\mathbf{x})$ then $\nabla h^* = \mathrm{prox}_\tau$

where $\mathrm{prox}_\tau(\mathbf{z}) := \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{z} - \mathbf{w}\|^2 + \tau(\mathbf{w})$

$$h^*(\mathbf{w}) = \sup_{\mathbf{z}} \mathbf{w}^T\mathbf{z} - \tfrac{1}{2}\|\mathbf{z}\|^2 - \tau(\mathbf{z}) \;= -\inf_{\mathbf{z}} -\mathbf{w}^T\mathbf{z} + \tfrac{1}{2}\|\mathbf{z}\|^2 + \tau(\mathbf{z})$$
$$= \tfrac{1}{2}\|\mathbf{w}\|^2 - \inf_{\mathbf{z}} \tfrac{1}{2}\|\mathbf{z} - \mathbf{w}\|^2 + \tau(\mathbf{z}).$$

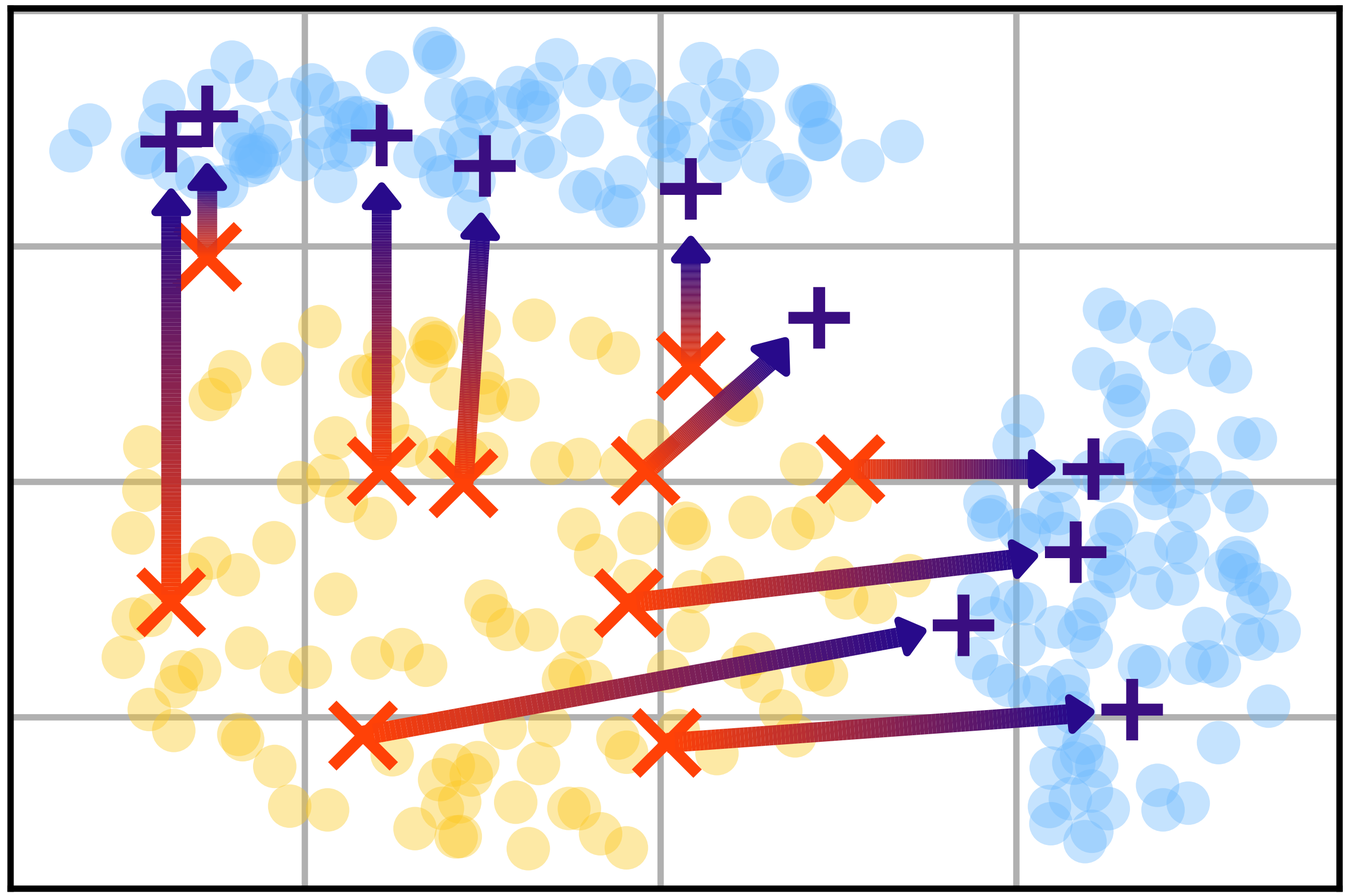
Monge / Bregman / Occam : In Practice

$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2$$



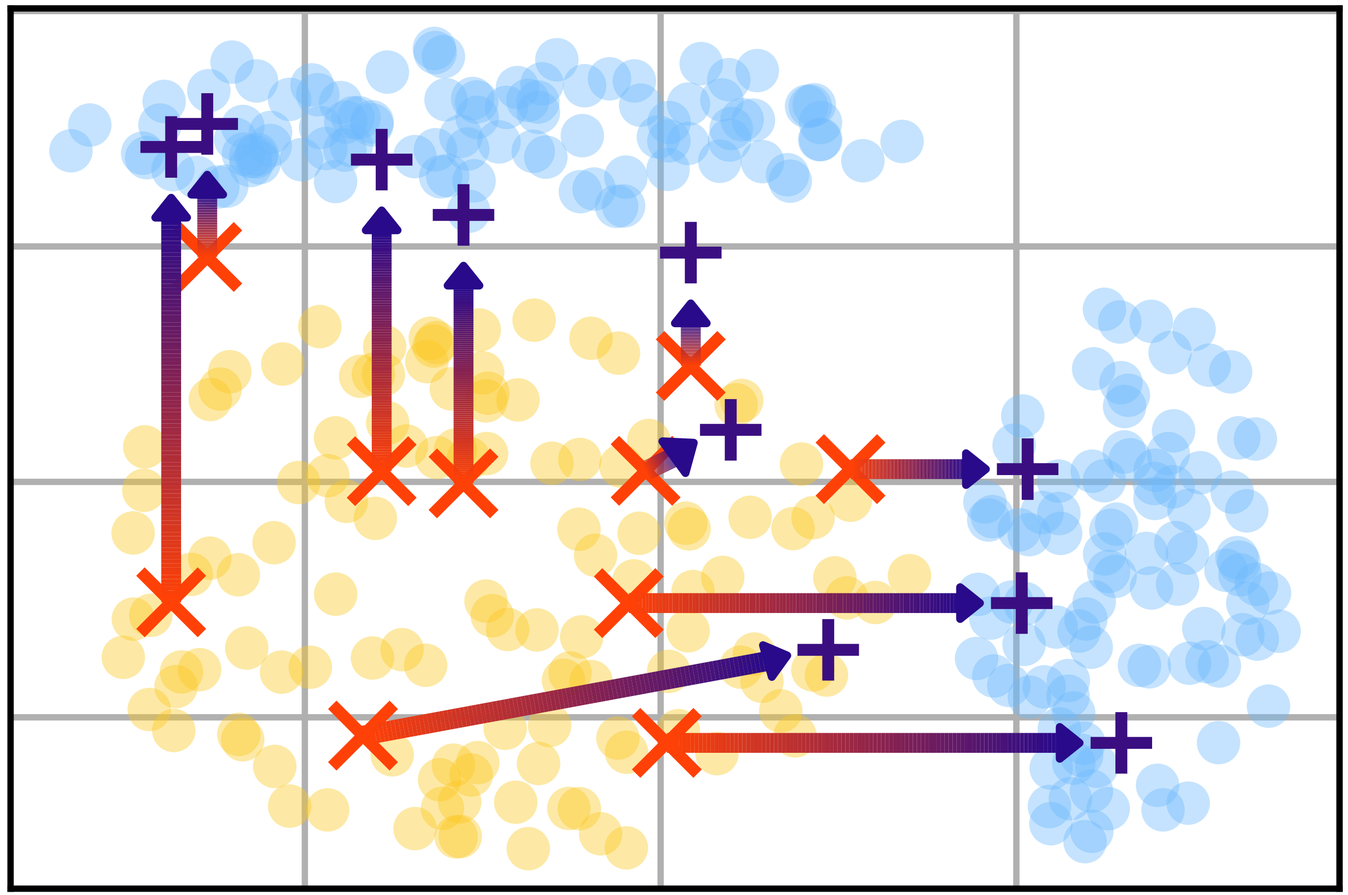
Monge / Bregman / Occam : In Practice

$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2 + 1\|\mathbf{z}\|_1$$



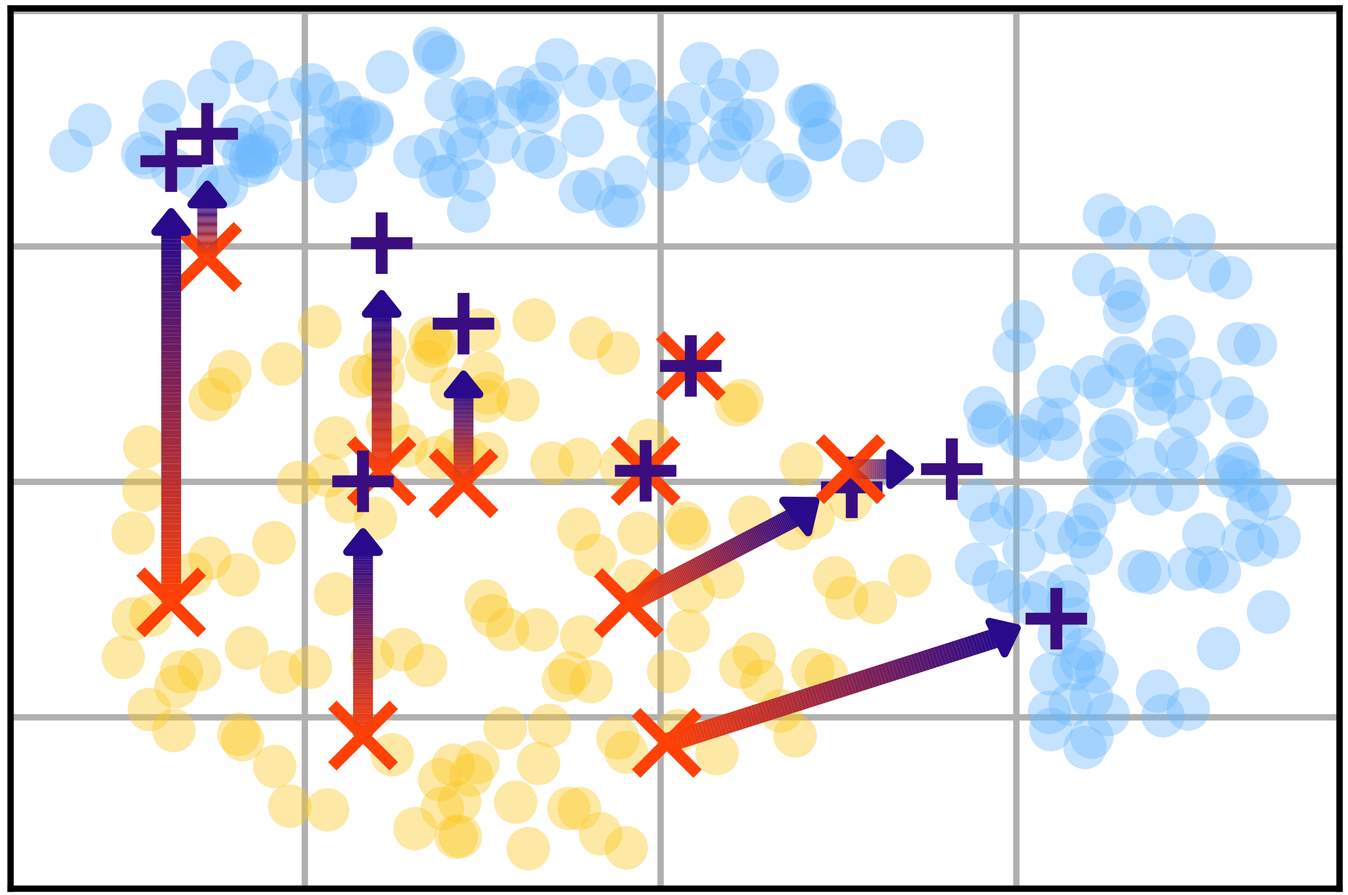
Monge / Bregman / Occam : In Practice

$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2 + 2\|\mathbf{z}\|_1$$



Monge / Bregman / Occam : In Practice

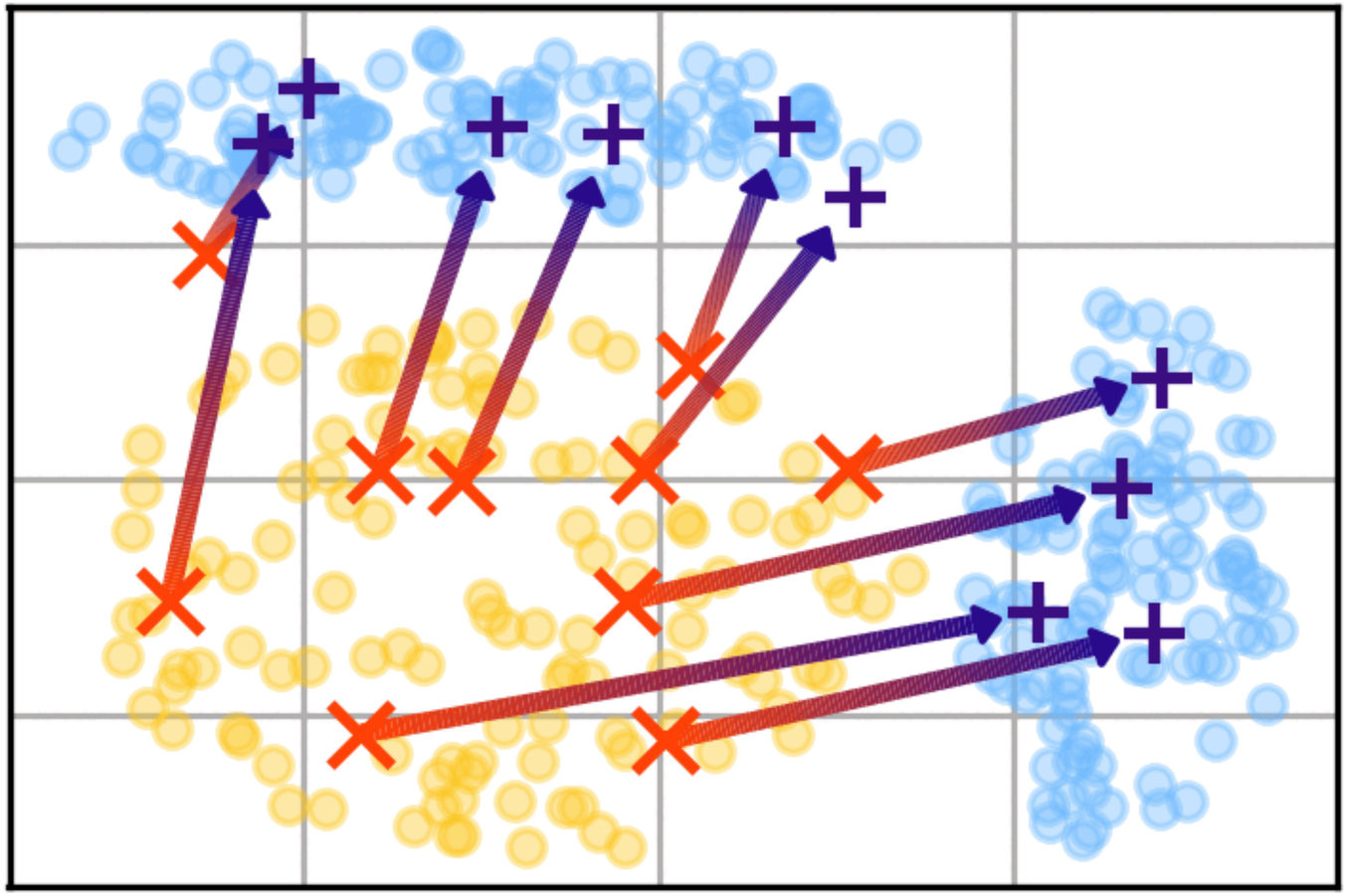
$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2 + 5\|\mathbf{z}\|_1$$



Monge / Bregman / Occam : In Practice

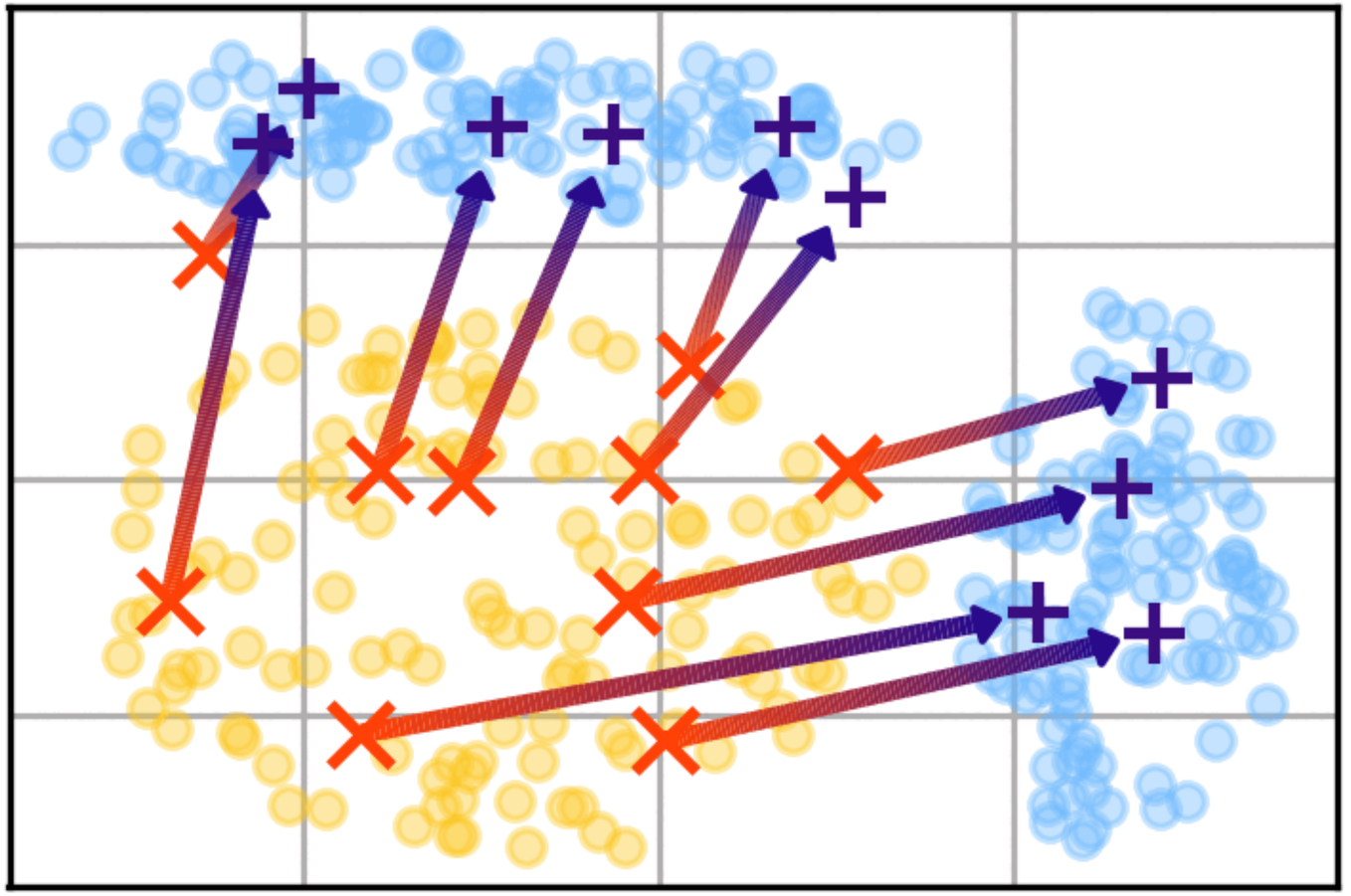
Monge / Bregman / Occam : In Practice

$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2 + 0\|\mathbf{z}\|_{\mathrm{ov}k}^2$$



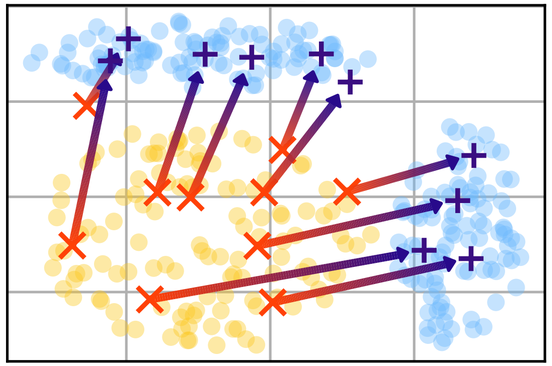
Monge / Bregman / Occam : In Practice

$$h(\mathbf{z}) = \tfrac{1}{2}\|\mathbf{z}\|_2^2 + 0\|\mathbf{z}\|_{\mathrm{ov}k}^2$$

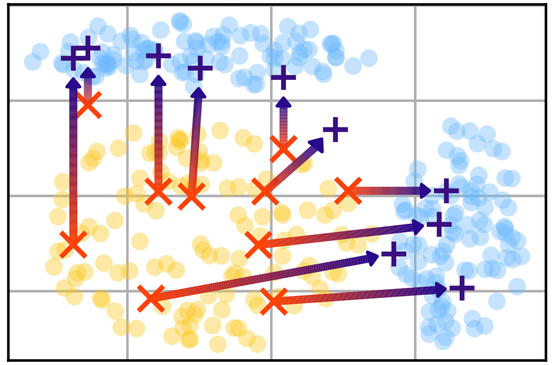


Monge / Bregman / Occam : In Practice

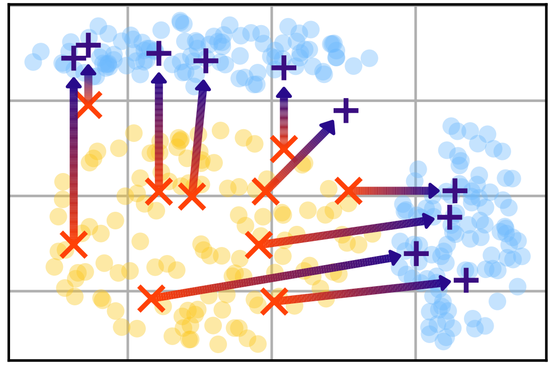
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2$



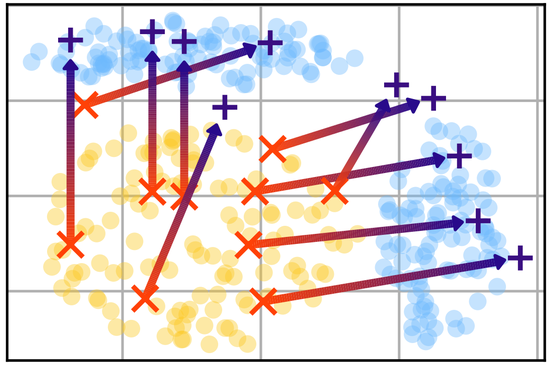
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 1\|\mathbf{z}\|_1$



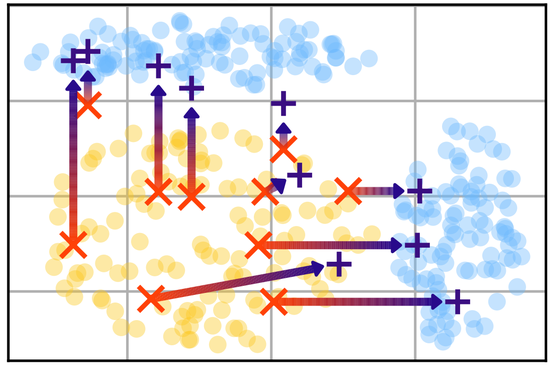
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 0.5\tau_{\text{stvs}}(\mathbf{z})$



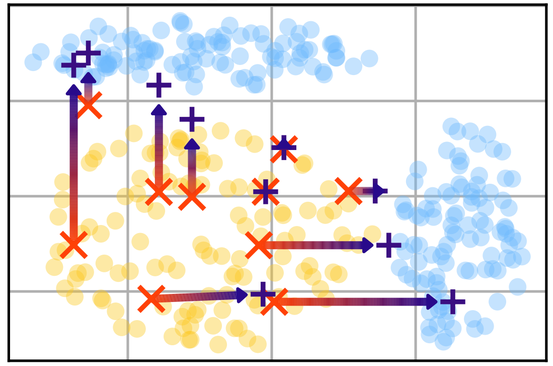
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 10\|\mathbf{z}\|_{\text{ov}k}^2$



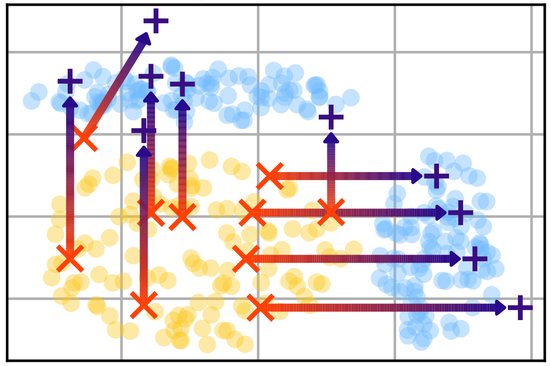
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 2\|\mathbf{z}\|_1$



$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 2\tau_{\text{stvs}}(\mathbf{z})$



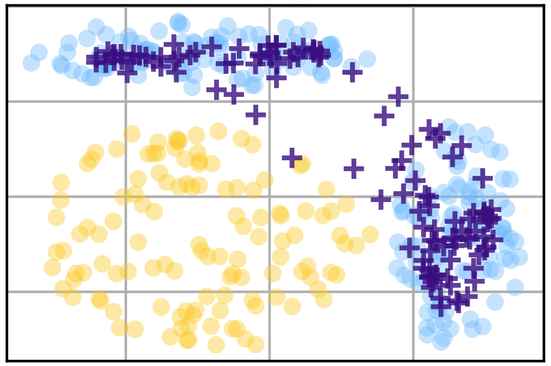
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 100\|\mathbf{z}\|_{\text{ov}k}^2$



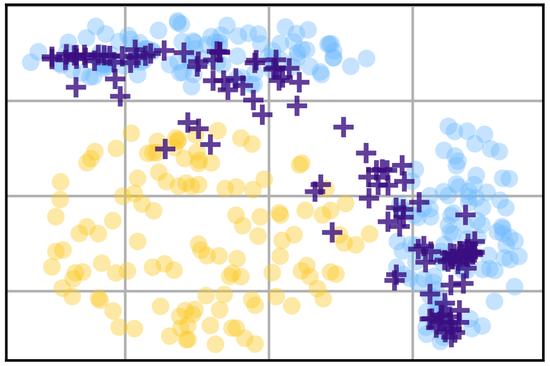
Base measure $\mathbf{x}^i$ Target measure $\mathbf{y}^j$ Points $\mathbf{x}$ Entropic map $T_{h,\varepsilon}(\mathbf{x})$

Monge / Bregman / Occam : In Practice

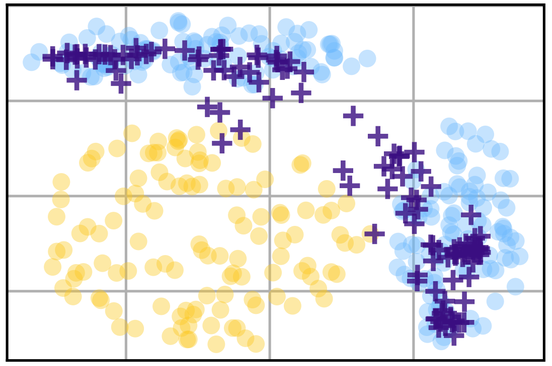
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2$



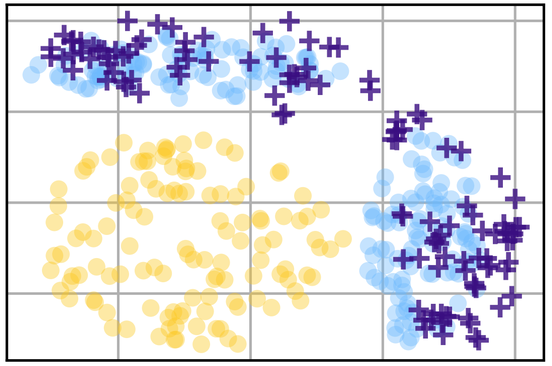
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 1\|\mathbf{z}\|_1$



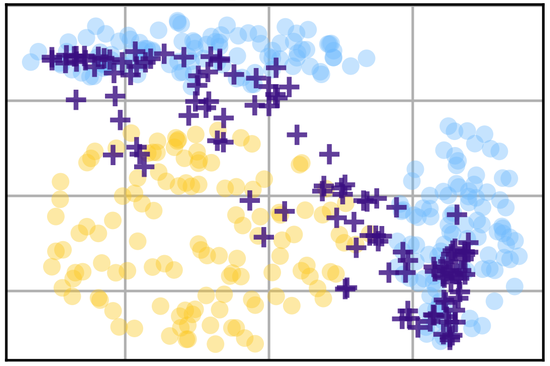
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 0.5\tau_{\text{stvs}}(\mathbf{z})$



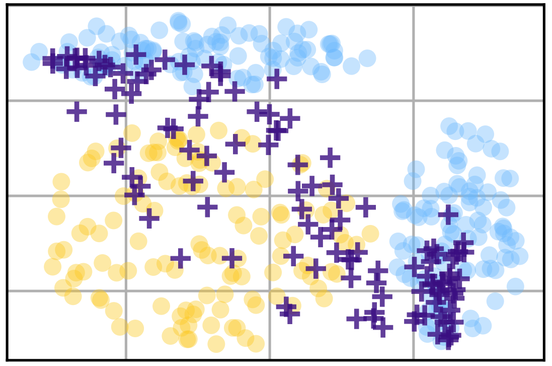
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 10\|\mathbf{z}\|_{\text{ov}k}^2$



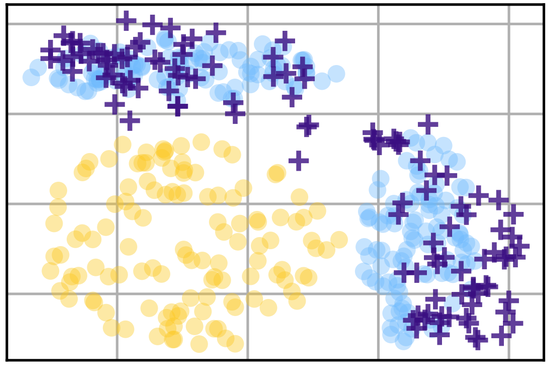
$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 2\|\mathbf{z}\|_1$



$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 2\tau_{\text{stvs}}(\mathbf{z})$



$h(\mathbf{z}) = \frac{1}{2}\|\mathbf{z}\|_2^2 + 100\|\mathbf{z}\|_{\text{ov}k}^2$



Base measure $\mathbf{x}^i$ Target measure $\mathbf{y}^j$ Entropic map distribution $(T_{h,\varepsilon}(\mathbf{x}^i))$

Coding Perspective

```
# jit first a Sinkhorn solver.
solver = jax.jit(sinkhorn.Sinkhorn())

def entropic_map(x, y, cost_fn: costs.TICost) -> jnp.ndarray:
    geom = pointcloud.PointCloud(x, y, cost_fn=cost_fn)
    output = solver(linear_problem.LinearProblem(geom))
    dual_potentials = output.to_dual_potentials()
    return dual_potentials.transport
```

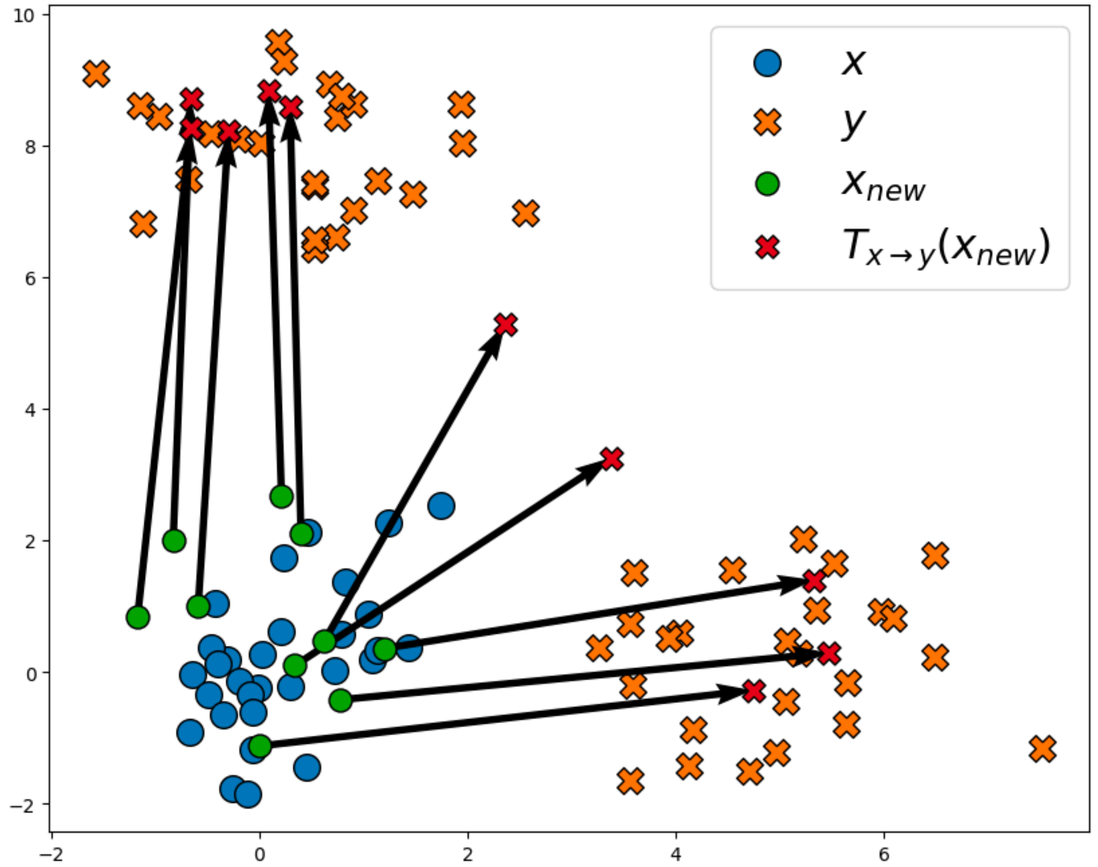
Coding Perspective

```
# jit first a Sinkhorn solver.
solver = jax.jit(sinkhorn.Sinkhorn())

def entropic_map(x, y, cost_fn: costs.TICost) -> jnp.ndarray:
    geom = pointcloud.PointCloud(x, y, cost_f
    output = solver(linear_problem.LinearProb
    dual_potentials = output.to_dual_potentia
    return dual_potentials.transport
```

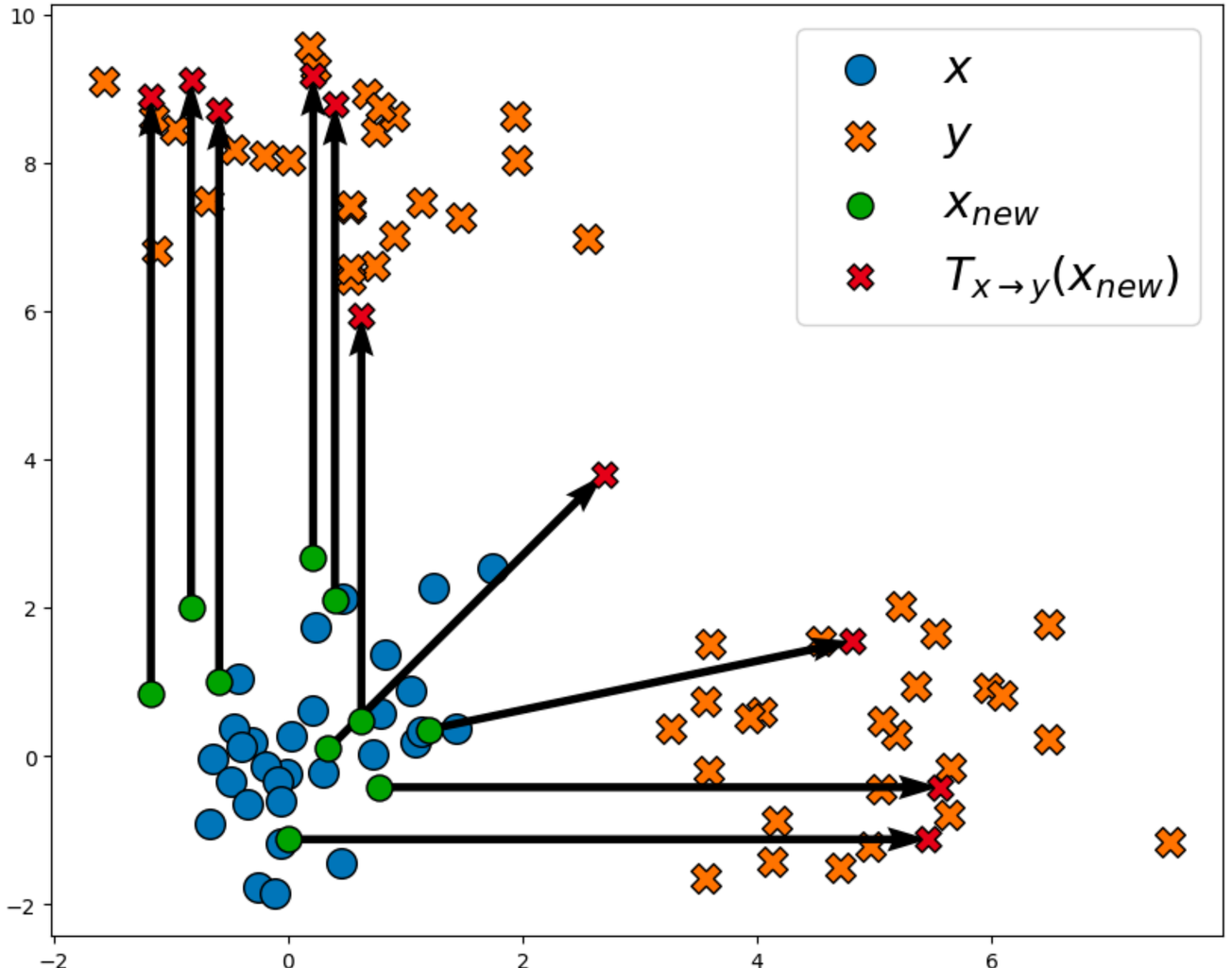
```
map = entropic_map(x,y,costs.SqEuclidean())

plot_map(x, y, x_new, map(x_new))
```

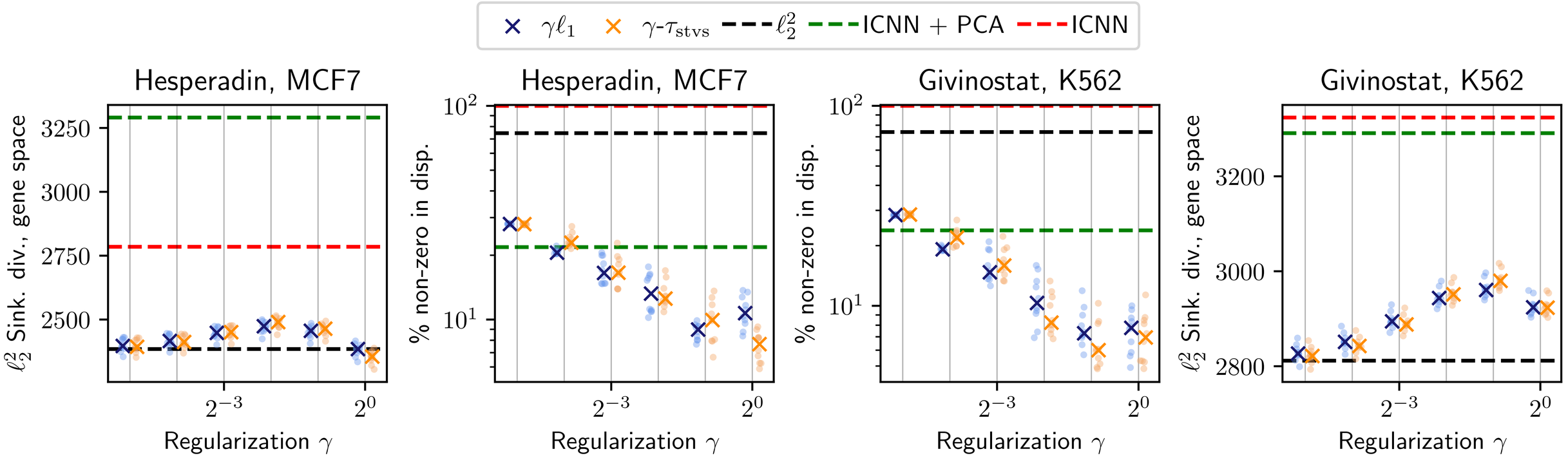


Coding Perspective

```
map = entropic_map(x, y, costs.ElasticSqKOverlap(k=1, scaling_reg=1.0))
```

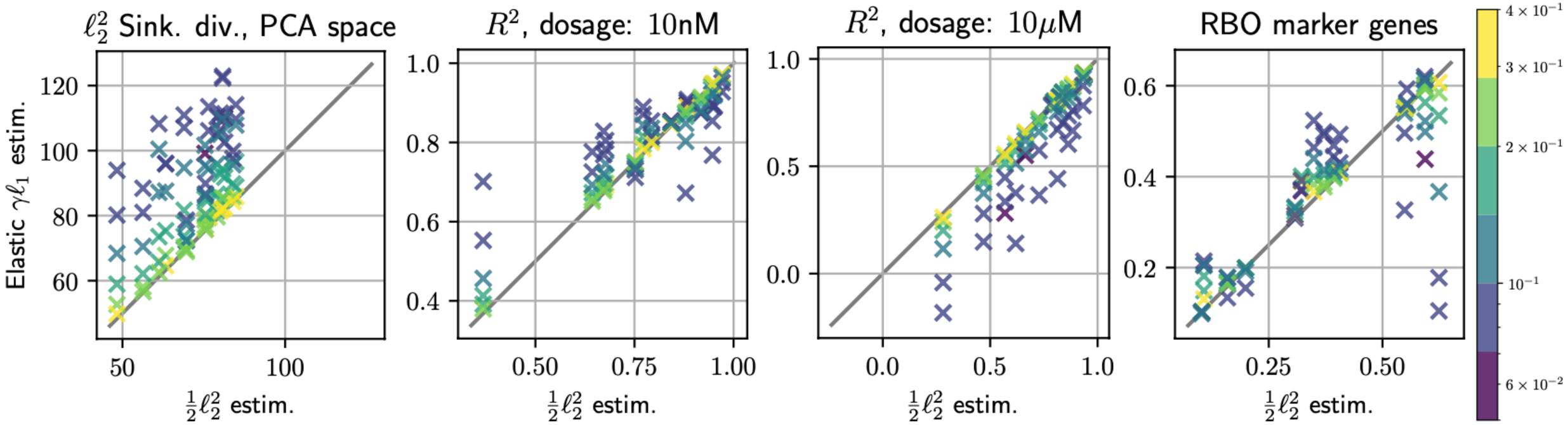


Monge / Bregman / Occam : High-Dim OT with sparsity



	Cont.	Dac.	Giv.	Bel.	Hes.	Quis.
A549	3274	558	703	669	436	475
K562	3346	388	589	656	624	339
MCF7	6346	1562	1805	1684	882	1520

Monge / Bregman / Occam : High-Dim OT with sparsity



Perspective



Benchmark [Korotin+21]

Extensions

- to structured and general cost functionals [Klein+23, Korotin+22, Fan+21]
- to unbalanced problems [Lübeck+22, Eyring+22]
- to incomparable spaces [Nekrashevich+23, Bunne+19, Dumont+22]
- to conditional settings [Bunne+22]
- to barycenter estimation [Korotin+22, Fan+21]
- to Riemannian manifolds [Rezende+21, Scarvelis+22]

Statistics of entropic maps [Rigollet+22, Pooladian+21]

Applications

- e.g., image super-resolution [Gazdieva+22]

Resources

Books

G. Peyré and M. Cuturi. **Computational Optimal Transport: With Applications to Data Science.** Foundations and Trends® in Machine Learning 11.5-6 (2019)

C. Villani. **Topics in Optimal Transportation.** GSM Vol. 58, AMS, 2009.

F. Santambrogio. **Optimal Transport for Applied Mathematicians.** Birkhäuser, 2015.

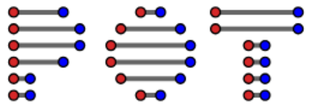
Review Articles

Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. **Optimal Transport in Systems and Control.** *Annual Review of Control, Robotics, and Autonomous Systems* Vol. 4 (2021).

Python Libraries



https://ott-jax.readthedocs.io/



https://pythonot.github.io/



https://www.kernel-operations.io/geomloss/