Optimal Transport

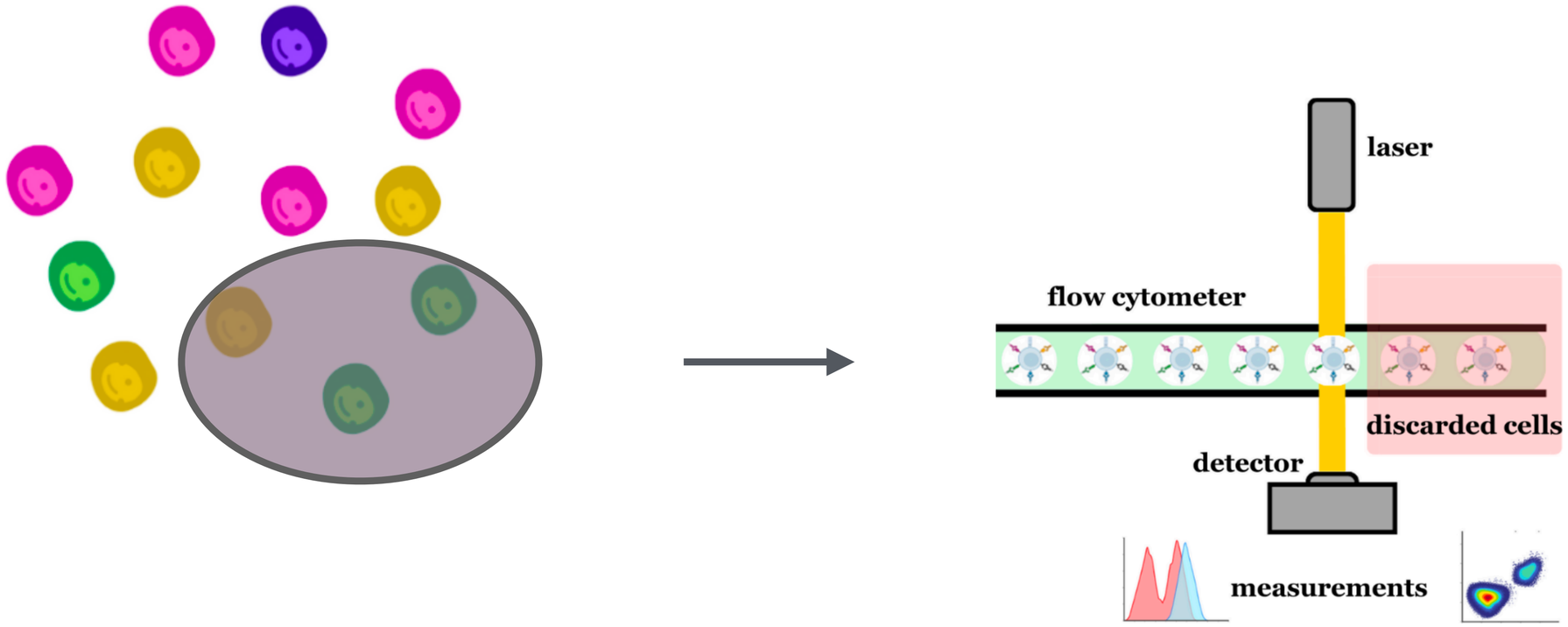
Darin Momayezi

Table of Contents

Introduction

Motivation

Cannot measure the same cell twice : need to work on the level of populations

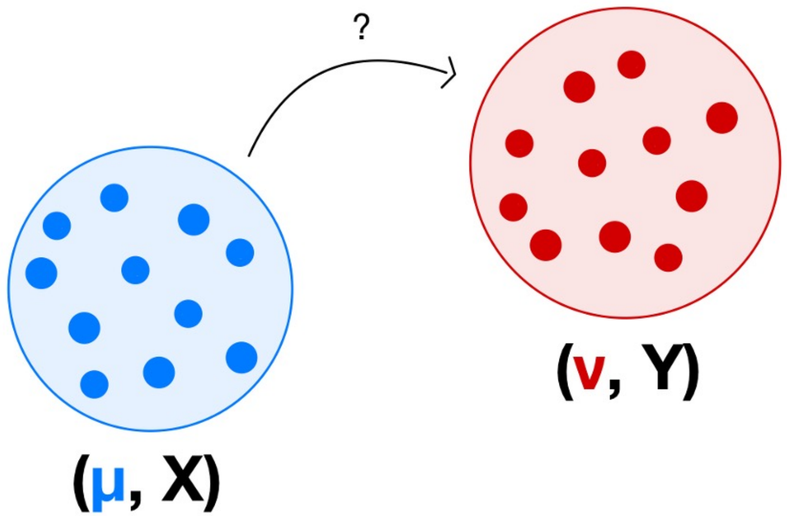


Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Introduction

Problem Statement

- We have data …
- But how are points in X and Y connected ?
- I.e., how do we transform points m(x) in X to points n(y) in Y ?



μ, ν : measures

$X, Y \in \mathbb{R}^d$: spaces

$m(x), n(y)$: distributions

Distances

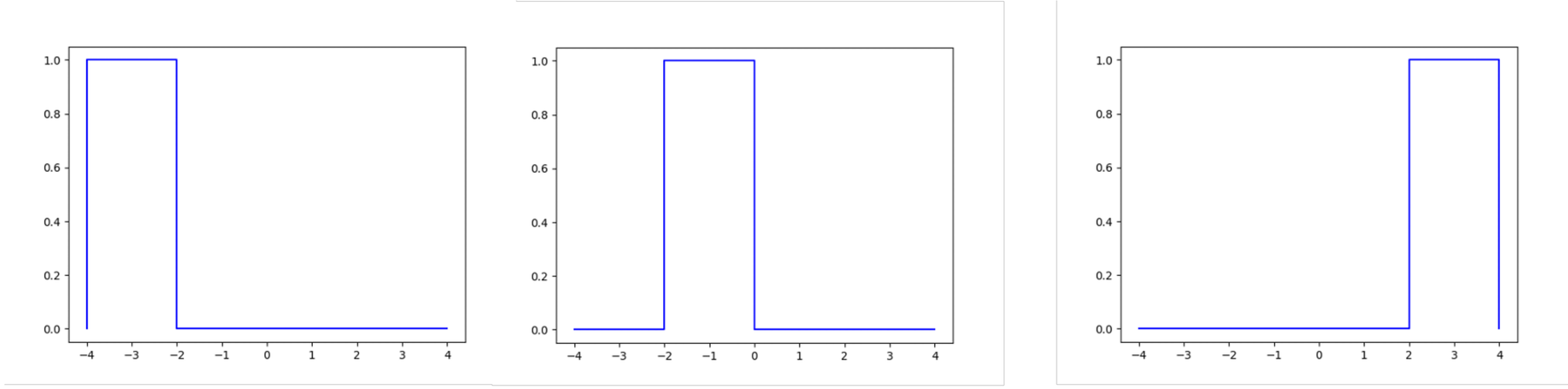
Why do we need an abstract notion of distance ?

- Essentially, quantify how much effort is needed to move points from m(x) to n(y). We could ask about …
- Some form of least action principle : minimal effort path
- Similarity of distributions : “distance” between distributions
- Clearly, we need a notion of distance between distributions. We have options …

Distances

Comparing options

- Consider the following histograms
- Total variation : $TV(P,Q) = \sup_{A \subseteq X} |P(A) - Q(A)| = \frac{1}{2}\sum_i |p_i - q_i|$ says distance same for all histograms
- Wasserstein Distance : $W_1(\mu,\nu) = \int_{\mathbb{R}} |F_\mu(x) - F_\nu(x)|\,dx$ says 1 and 2 closer

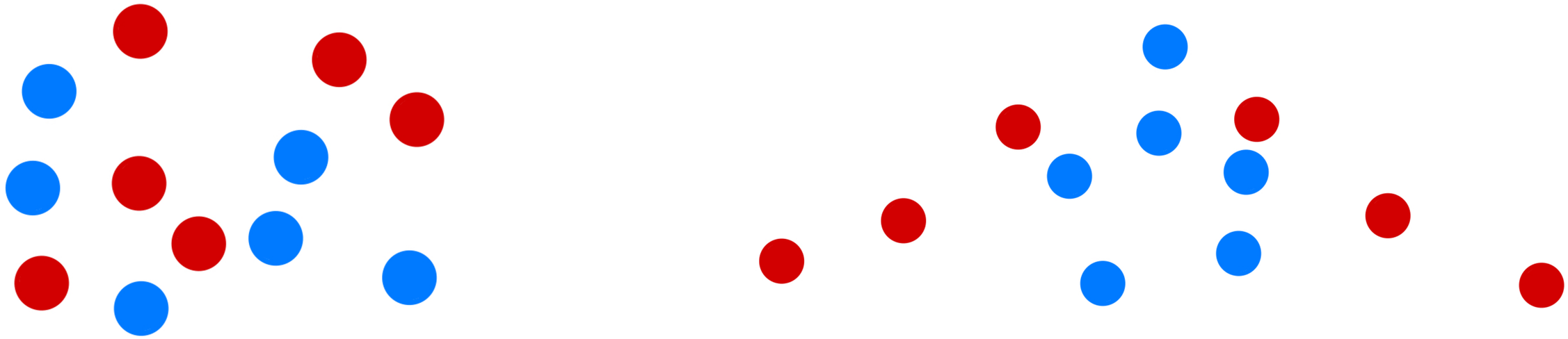


1 2 3

Distances

Comparing options

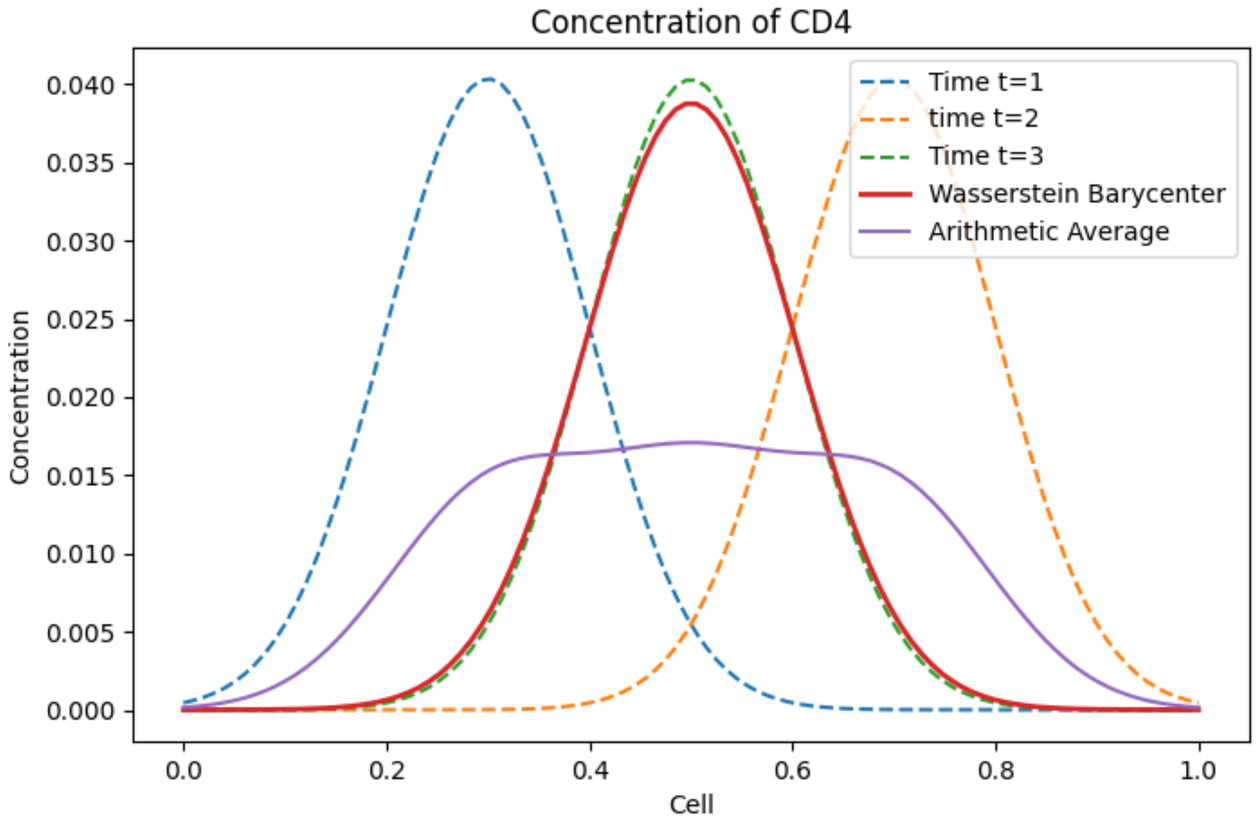
- When the Wasserstein distance is small that means the original distribution is “perturbed” by a small amount
- When the Wasserstein distance is large the original distribution is “perturbed” by a large amount



Distances

Wasserstein distance preserves structure

- The Wasserstein distance is a much better summary than a simple average.



Distances

Wasserstein distance is a metric

- $W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} d(x,y)^p d\pi(x,y) \right)^{1/p}$: we will be interested in the squared Euclidean distance.
- The Wasserstein distance is a metric :
  - Non-negative : $W_p(\mu,\nu) > 0$ if $\mu \neq \nu$
  - Symmetric : $W_p(\mu,\nu) = W_p(\nu,\mu)$
  - Obeys the triangle inequality : $W_p(\mu,\nu) \leq W_p(\mu,\gamma) + W_p(\gamma,\nu)$
- The Wasserstein distance gives a meaningful distance between measures : reliable way to interpolate between measures, leads to dynamic formulation over space of measures (given a metric structure with the Wasserstein distance) that minimize a total length

Table of Contents

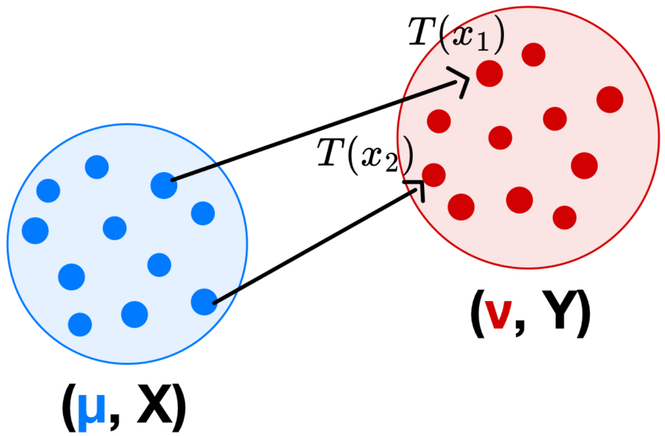
Monge problem

Definition

Gaspard Monge

(1746-1818)

- Find a mapping between points in X to points in Y
- To do this optimally find a map $T : X \rightarrow Y$ that minimizes
- $$\inf_T \left\{ \int_X c(x, T(x))\, d\mu(x)\ ;\ T_{\#}\mu = \nu \right\}$$
- Discrete measures can be defined as $\mu = \sum_{i=1}^{M} m_i \delta_{x_i}$ and $\nu = \sum_{i=1}^{N} n_i \delta_{y_i}$



μ, ν : measures

$c(x, y)$: cost function

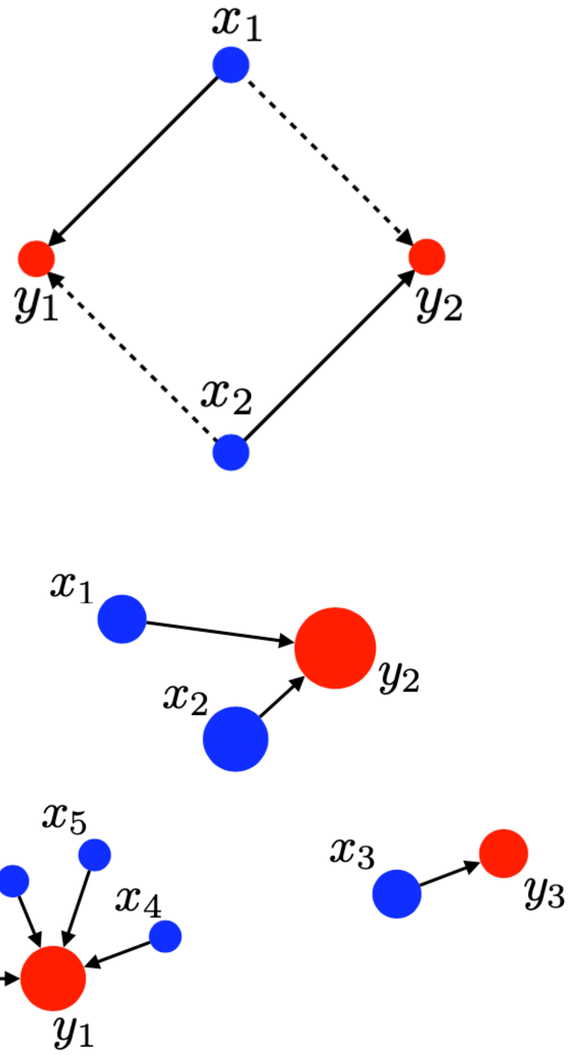
$T_{\#}[\ \cdot\ ]$: push forward

Monge, Gaspard, *Sur la théorie des déblais et des remblais*

Monge Problem

Drawbacks

- Deterministic/surjective mapping - can't assume each measurement has the same amount of cells
- Minimization over maps (may not be unique) - expensive
- It is essentially combinatorial (maps can be thought of as permutation matrices) and non-convex
  - $[x_1, x_2, \ldots, x_n]$
  - $[y_1, y_2, \ldots, y_n]$
- This is a hard optimization problem, but the constraints can be relaxed

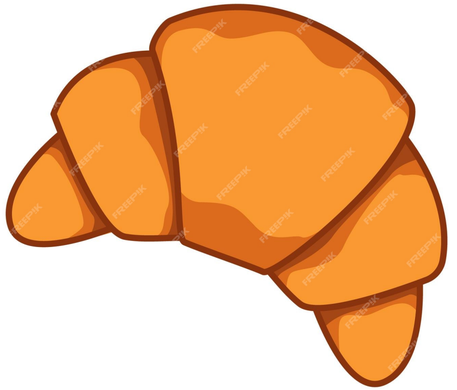


Cuturi et al., Computational Optimal Transport

Monge Problem

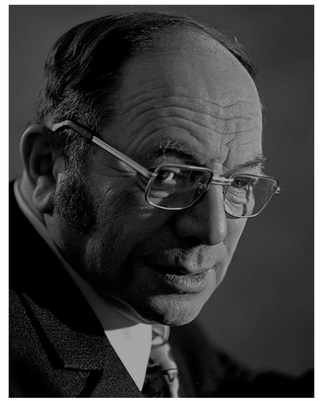
Why is convexity important ?

- Convex set : a set of points that contains every line segment between two points in the set
- Consequence : a local minimum is the global minimum



Kantorovich relaxation

From deterministic to probabilistic



Leonid Kantorovich
(1912-1986)

- Replace maps with coupling matrix $\mathbf{P}$ such that
- $\mathbf{U}(\mu,\nu) = \left\{\mathbf{P} \in \mathbb{R}_+^{n\times m} \ : \ \mathbf{P}\mathbb{1}_m = \mu \ \text{ and } \ \mathbf{P}^T\mathbb{1}_n = \nu\right\}$
- $L_{\mathbf{C}}(\mu,\nu) := \min_{\mathbf{P}\in\mathbf{U}(\mu,\nu)} \langle \mathbf{C}, \mathbf{P}\rangle := \sum_{i,j} \mathbf{C}_{i,j}\mathbf{P}_{i,j}$
- In the continuous setting find a joint distribution π such that
- $\mathbf{U}(\mu,\nu) = \left\{\pi \in \mathscr{M}_+^1(X\times Y) \ : \ P_{X\#}\pi = \mu \ \text{ and } \ P_{Y\#}\pi = \nu\right\}$
- $L_{\mathbf{C}}(\mu,\nu) := \min_{\pi\in\mathbf{U}(\mu,\nu)} \int_{X\times Y} c(x,y)\ d\pi(x,y)$ loss function

When the cost is the p-th power of a metric $c(x,y) = d(x,y)^p$ the minimum transport cost defines the p-Wasserstein distance. We're interested in obtaining π .

$$W_p(\mu,\nu) = \left(\min_{\pi\in\mathbf{U}(\mu,\nu)} \int_{X\times Y} d(x,y)^p \ d\pi(x,y)\right)^{1/p}$$

μ, ν : measures
$\mathbf{C}$, $c(x,y)$: cost
$\mathbf{U}$: set of coupling matrices
$L_{\mathbf{C}}$: value of optimization
$P_{X\#}$: push forward of projection

Kantorovich relaxation

Advantages

- Linear program over space of joint distributions $\pi(x, y)$ - convex (can use convex optimization techniques, e.g., dual formulation)
- The structure of $\pi(x, y)$ and the marginal constraints allow for mass splitting
  - $\int_Y d\pi(x, y) = d\mu(x)$ and $\int_X d\pi(x, y) = d\nu(y)$
- Natural connection to the Wasserstein distance - connection to convex analysis, differential geometry, PDEs, etc.

Entropic regularization

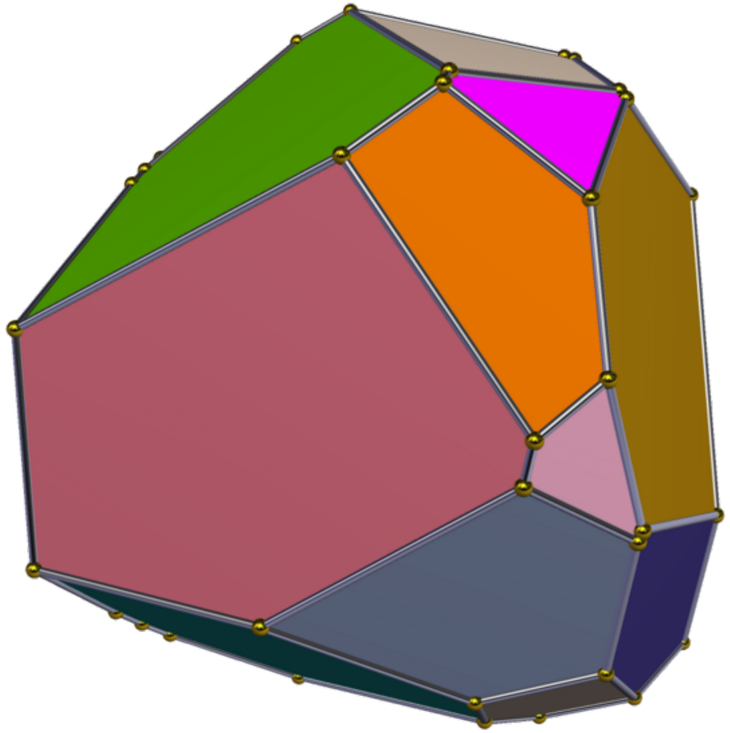
Geometry of optimal transport

- While the Kantorovich formulation allows for a more probabilistic model this is not guaranteed.
- The Kantorovich problem will find optimal, but extreme Monge-like solutions. Regularization makes solutions easier to find and more regular.
- Key idea : Use the Kullback-Leibler (KL) divergence between the optimal transport plan and the maximum entropy solution as a regularizing function. Ensure that the joint distribution is more probabilistic by adding entropy.

Entropic regularization

Geometry of optimal transport

- $\mathbf{U}(\mu, \nu)$ is $m \times n$ space of transport plans π that obeys the marginal constraints
- The space is carved up by $m + n$ constraints that define a convex polytope.
- In high dimensions, $m \times n \gg m + n - 1$ and there are many possible solutions. The optimal solutions live at the vertices of the polytope.



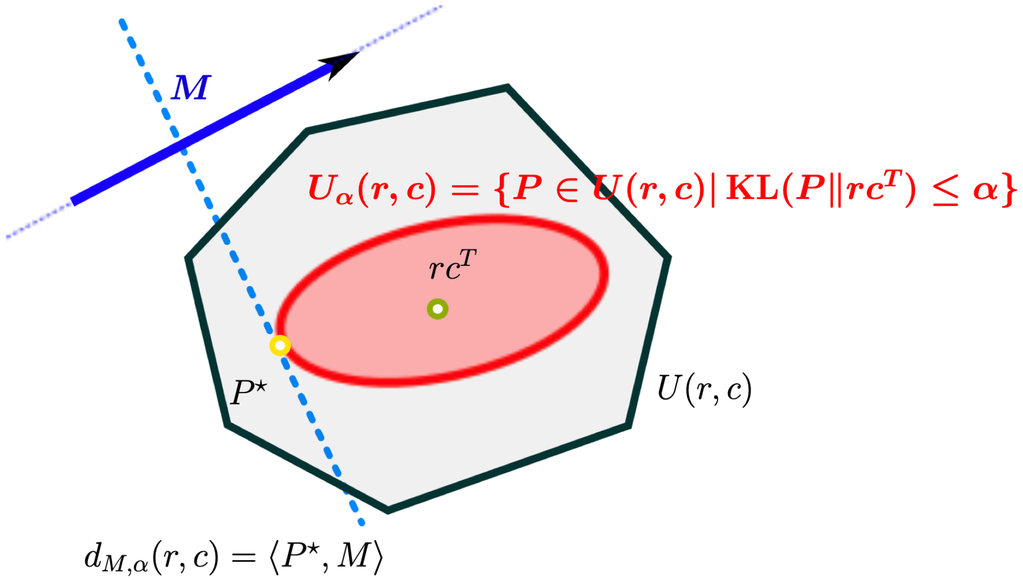
Entropic regularization

Geometry of optimal transport

- The highest entropy/most regular distribution at the center of the polytope is called the independent coupling/transport plan $\pi^I = \mu \otimes \nu = \sum_{i,j} \mu_i \nu_j \delta_{x_i} \otimes \delta_{y_j}$

- Entropic regularization moves the optimal transport plan from the boundary of the polytope to the center

- Add regularization term : $\epsilon KL(\pi \,|\, \mu \otimes \nu)$



Cuturi, Marco, *Sinkhorn distances : Lightspeed computation of optimal transportation distances*

Dual formulation

Making the Kantorovich problem easier to solve

- Replace Kantorovich's convex problem with a dual concave problem
- $W(\mu,\nu) = \sup_{(f,g)\in\Phi} \int f d\mu + \int g d\nu$, $\Phi := \{(f,g) \in C(X) \times C(Y) \ : \ f(x) + g(x) \leq c(x,y)\}$
- How to choose these potentials ?
- Definition (c-transform) : $\forall y \in Y,\ f^c(y) := \inf_{x\in X} c(x,y) - f(x)$
  - $f(x) + g(y) \leq c(x,y)$
  - $g(y) \leq c(x,y) - f(x)$
  - $g^*(y) = \inf_{x\in X} c(x,y) - f(x) = f^c(x)$

$$W(\mu,\nu) = \sup_{(f,g)\in\Phi} \int f d\mu + \int f^c d\nu$$

μ, ν : measures

$c(x,y)$: cost

f, g : dual potentials

Sinkhorn Iteration

Making the Kantorovich problem easier to solve

- Sinkhorn algorithm says : $\mathbf{P}_{i,j} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j$ subject to marginal constraints $\mathbf{P}\mathbb{1}_m = \text{diag}(\mathbf{u})\mathbf{K}\text{diag}(\mathbf{v})\mathbb{1}_m = \mathbf{m}$ and $\mathbf{P}^T\mathbb{1}_n = \text{diag}(\mathbf{v})\mathbf{K}^T\text{diag}(\mathbf{u})\mathbb{1}_n = \mathbf{n}$
- Using the dual potentials $\mathbf{f}, \mathbf{g}$ the Lagrangian of the optimization problem $\left(\min_{\mathbf{P} \in \mathbf{U}(\mu,\nu)} \langle \mathbf{P}, \mathbf{C} \rangle - \epsilon \mathbf{H}(\mathbf{P}) \right)$ is $\mathscr{L}(\mathbf{P}, \mathbf{f}, \mathbf{g}) = \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}) - \langle \mathbf{f}, \mathbf{P}\mathbb{1}_m - \mathbf{m} \rangle - \langle \mathbf{g}, \mathbf{P}^T\mathbb{1}_n - \mathbf{n} \rangle$
- Optimal solution obtained via
  - $\dfrac{\partial \mathscr{L}(\mathbf{P}, \mathbf{f}, \mathbf{g})}{\partial \mathbf{P}_{i,j}} = \mathbf{C}_{i,j} + \varepsilon \log \mathbf{P}_{i,j} - \mathbf{f}_i - \mathbf{g}_j = 0$ and solving for $\mathbf{P}_{i,j}$
  - $\mathbf{P}_{i,j} = e^{\mathbf{f}_i/\varepsilon} e^{-\mathbf{C}_{i,j}/\varepsilon} e^{\mathbf{g}_j/\varepsilon} = \mathbf{u}_i \mathbf{K}_{i,j} \mathbf{v}_j$ and marginal constraints require $\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{m}$ and $\mathbf{v} \odot (\mathbf{K}^T\mathbf{u}) = \mathbf{n}$
  - Update step : $\mathbf{u}^{(l+1)} := \dfrac{\mathbf{m}}{\mathbf{K}\mathbf{v}^{(l)}}$ and $\mathbf{v}^{(l+1)} := \dfrac{\mathbf{n}}{\mathbf{K}^T\mathbf{u}^{(l+1)}}$

$\mathbf{f}, \mathbf{g}$: dual potential functions

$\mathbf{P}$: coupling

$\mathbf{H}(\cdot)$: entropy

Theory of Optimal Transport

Summary

- In summary, the goal is to find a transport plan $\boldsymbol{\pi}$ that minimizes a cost function including (i) the transport cost, (ii) the entropy of the transport plan and (iii) the marginal constraints.

$$\mathscr{L}[\pi;\ \varepsilon, \lambda_1, \lambda_2] := \int_{X\times Y} dxdy\, c(x,y)\pi(x,y) + \varepsilon \int_{X\times Y} dxdy\, \pi(x,y)\ \log \pi(x,y) + \lambda_1 \mathrm{KL}\left(\int dy \pi(x,y) \,|\, m(x)\right) + \lambda_2 \mathrm{KL}\left(\int dx \pi(x,y) \,|\, n(x)\right)$$

$$\pi^*(x,y) := \min_{\mathbf{U}(m,n)} \mathscr{L}[\pi; \varepsilon, \lambda_1, \lambda_2]$$

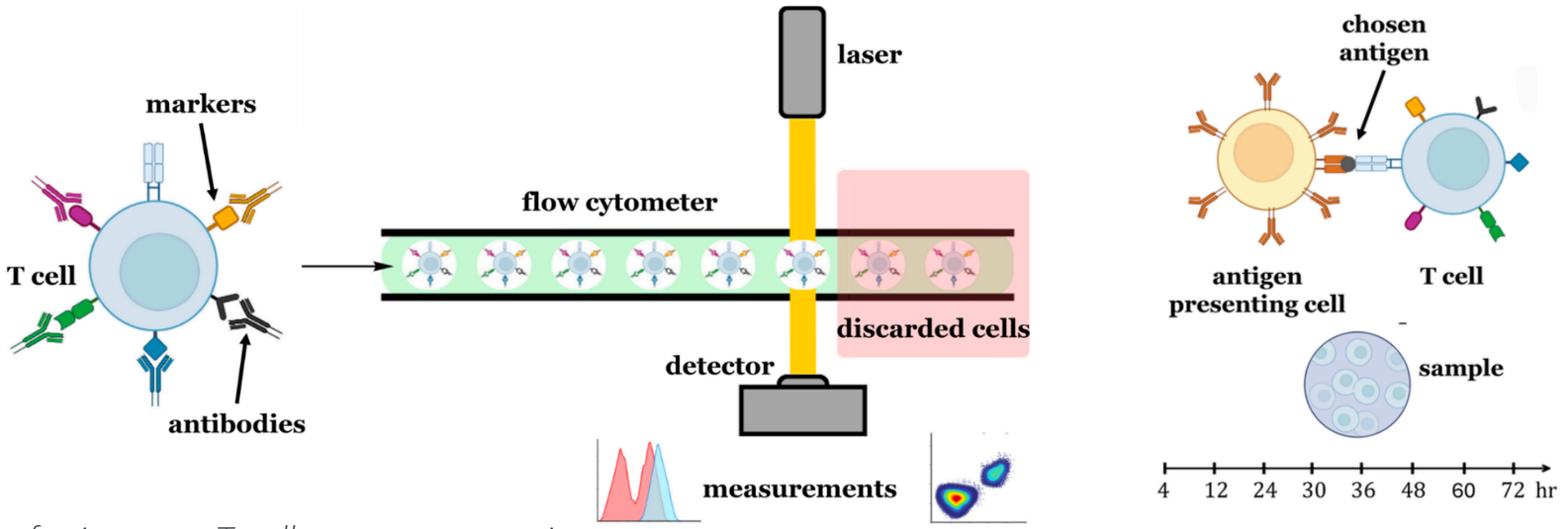
- λ_1 and λ_2 need to be chosen : limited knowledge about marginals

Table of Contents

Inferring T-Cell trajectories

How do we measure them ?

- T-cells perform different functions based on the markers they present in response to their environment, namely antigens
- Remove a representative sample from the population at regular intervals and measure surface markers using flow cytometry
- We measure 23 surface markers (and two shape markers and a proliferation score) at each time point

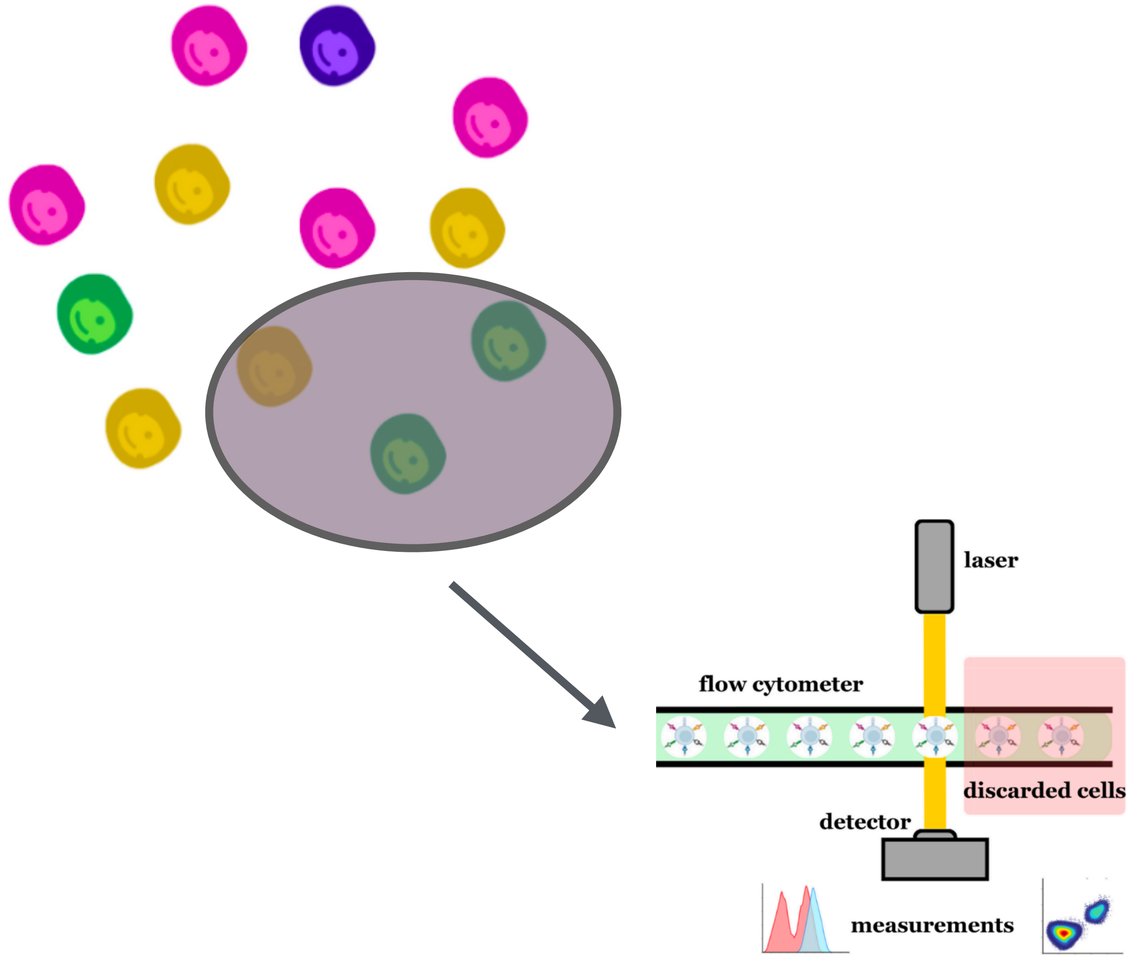


Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Inferring T-Cell trajectories

Central problem

- We cannot measure the same cell twice ; cells are destroyed during measurement
- We only have the concentrations of surface markers of a subset of T-cells at each time
- We have developed the OT tools to infer trajectories between cells in sampled data



Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Inferring T-Cell trajectories

What is the data ?

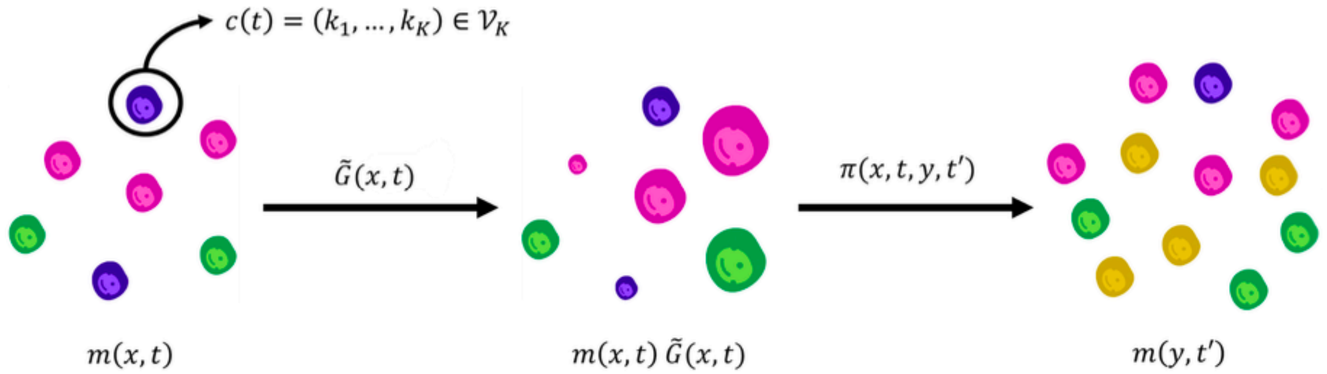
				Marker	CD8a	CD2
Peptide	**Concentration**	**Replicate**	**Time**	**Event**		
Q4	**10pM**	**1**	**4.0**	**1**	259	340
			...	...	...	...
		2	**72.0**	**10912**	490	273

- 24 markers tracked and each cell identified by marker concentration $c := [k_1, \ldots, k_K] \in \mathscr{V}_K$
- For a given sample : mass density $m(x,t) = \sum_{i=1}^{N(t)} \delta(x - c_i(t))$ and PDF $p(x,t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} \delta(c - c_i(t))$ satisfying the marginal constraints $\int_{\mathscr{V}_K} dx\, m(x,t) = N(t)$ and $\int_{\mathscr{V}_K} dy\, m(y,t') = N(t')$
- This is an unbalanced problem ($N(t) \neq N(t')$)

Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Inferring T-Cell trajectories

An unbalanced problem



- Data is noisy sample of population : sample fluctuations
- Cells proliferate : birth process model $G(x,t) = \exp[\beta(x,t)(t'-t)]$, where $\beta(x,t)$ is birth rate and $m(x,t') \approx m(x,t)G(x,t)$
- Define a modified growth parameter $\tilde{G}(x,t) = \dfrac{N(t')}{N(t)} \dfrac{G(x,t)}{\langle G(x,t) \rangle_{p(x,t)}} \longrightarrow \tilde{m}(x,t) = m(x,t)G(x,t)$
- $p_i^{(0)}(t) = \dfrac{1}{N(t)} \dfrac{G_i^{(0)}(t)}{\langle \mathbf{G}^{(0)}(t) \rangle}$ and $q_j(t') = \dfrac{1}{N(t')}$

Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Inferring T-Cell trajectories

What is the algorithm ?

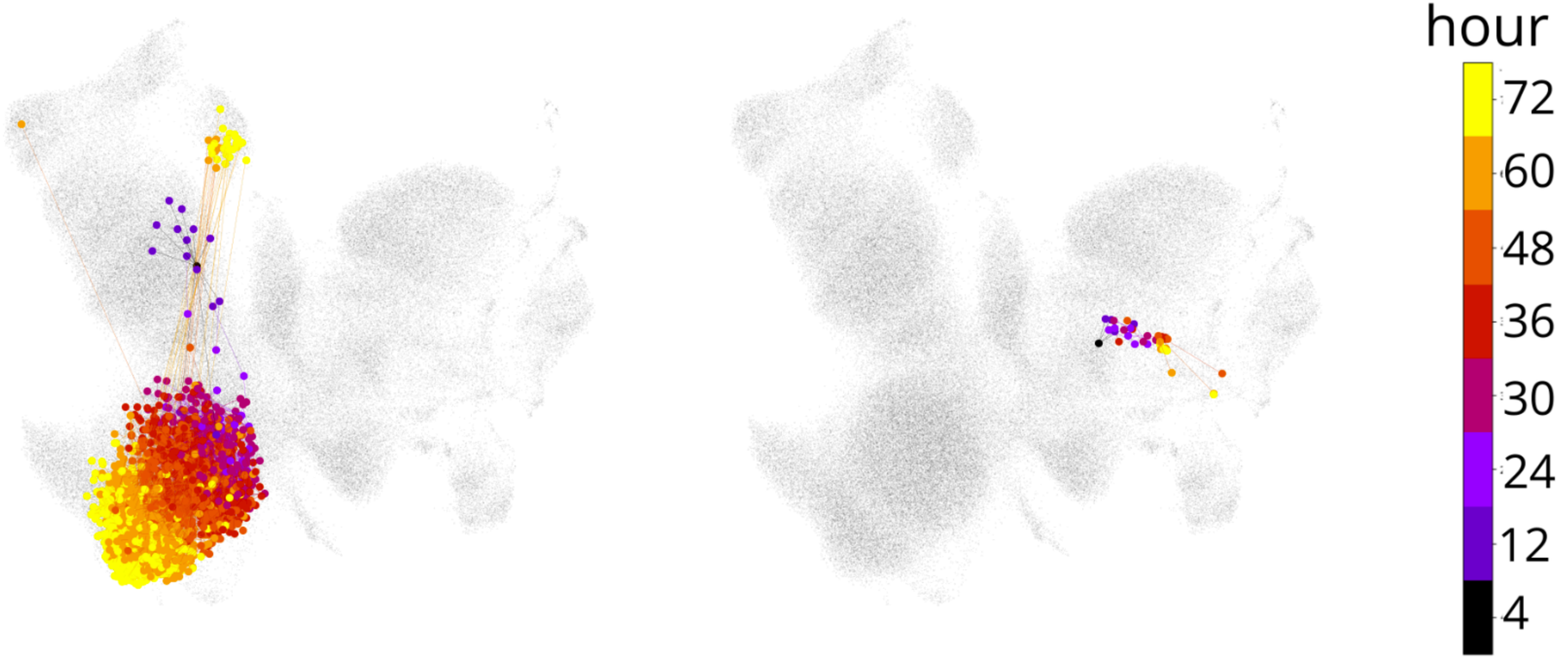
- Convert unbalanced problem to balanced problem : estimate growth parameter for each cell
- Handle numerical overflows in the Sinkhorn-Knopp algorithm : increase entropic regularization (trade off with accuracy) or perform calculations in the log domain
- Analyze the resulting transport maps (and cell differentiation through information quantities, e.g., mutual information, path entropies, etc.)

Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Inferring T-Cell trajectories

Results and conclusions : T-cell progenies

An ancestor of a cell c_j is a cell c_i^* that satisfies $c_i^* := \arg\max_N \pi_{i,j}$ at time t



Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Inferring T-Cell trajectories

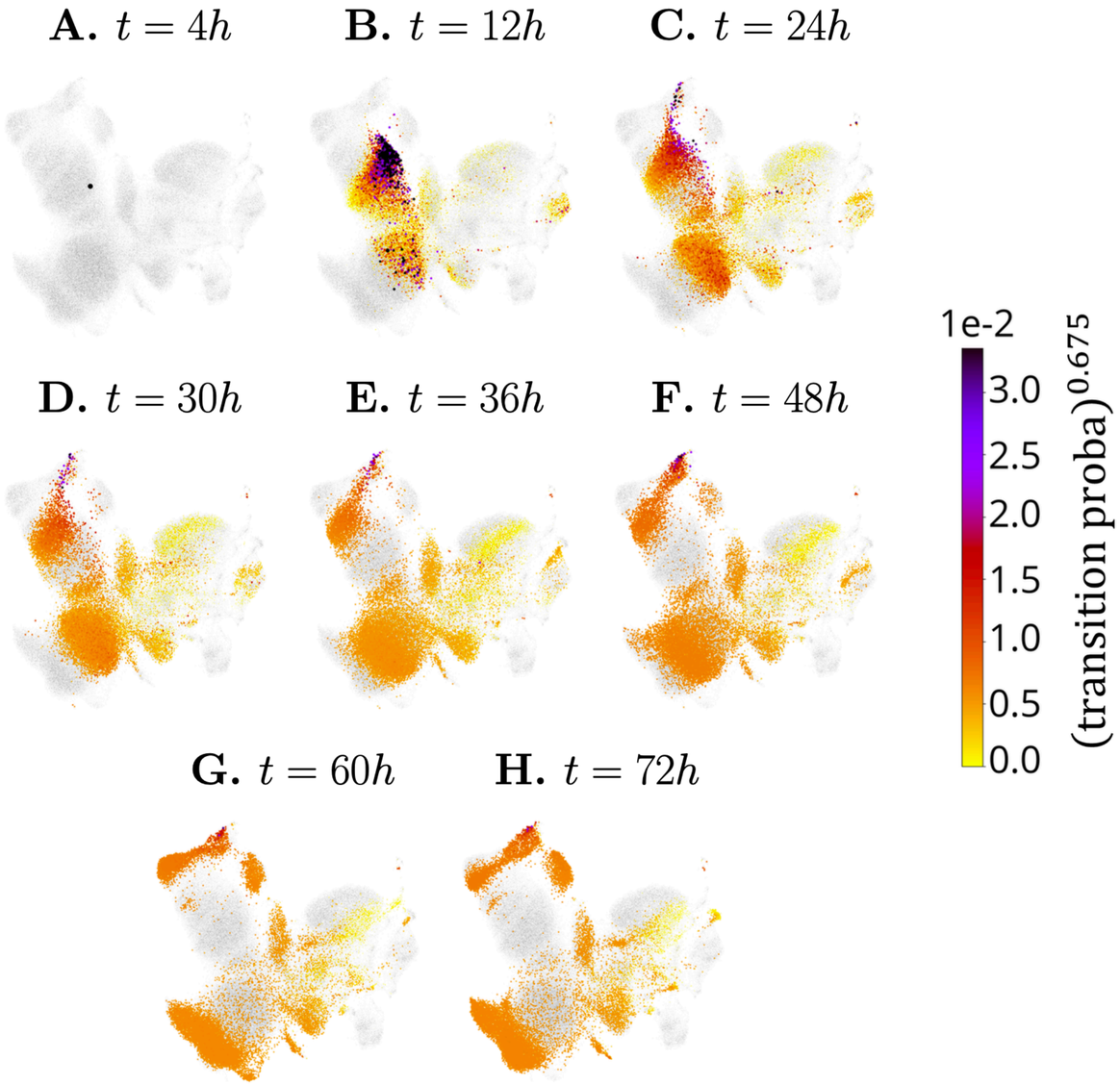
Single cell evolution

$$\mathbf{T}(t_j \mid t_i) := \pi(t_i, t_j) \circ /\tilde{\mathbf{p}}(t_i) := \frac{\pi_{i,j}(t_i, t_j)}{p_i(t_i)}$$

$$\mathbf{T}(t_k \mid t_0) = \mathbf{T}(t_1 \mid t_0) \ldots \mathbf{T}(t_k \mid t_{k-1})$$

$\mathbf{T}(t_j \mid t_i)$ represents a conditional probability

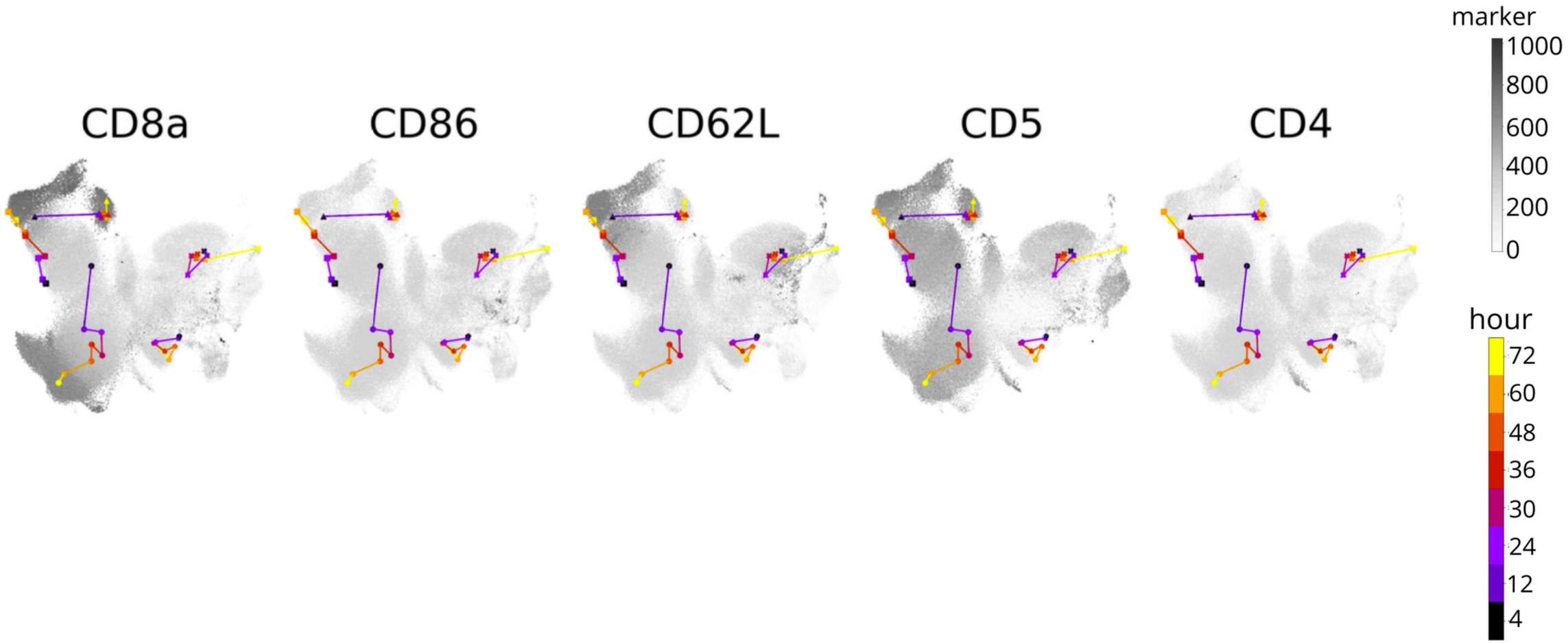
$\pi(t_i, t_j)$ represents a joint probability



Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Inferring T-Cell trajectories

Evolution of marker concentration



Hoyos, Pablo, *Optimal transport for immune T cell response to antigens*

Table of Contents

Dynamic Optimal Transport

Motivation

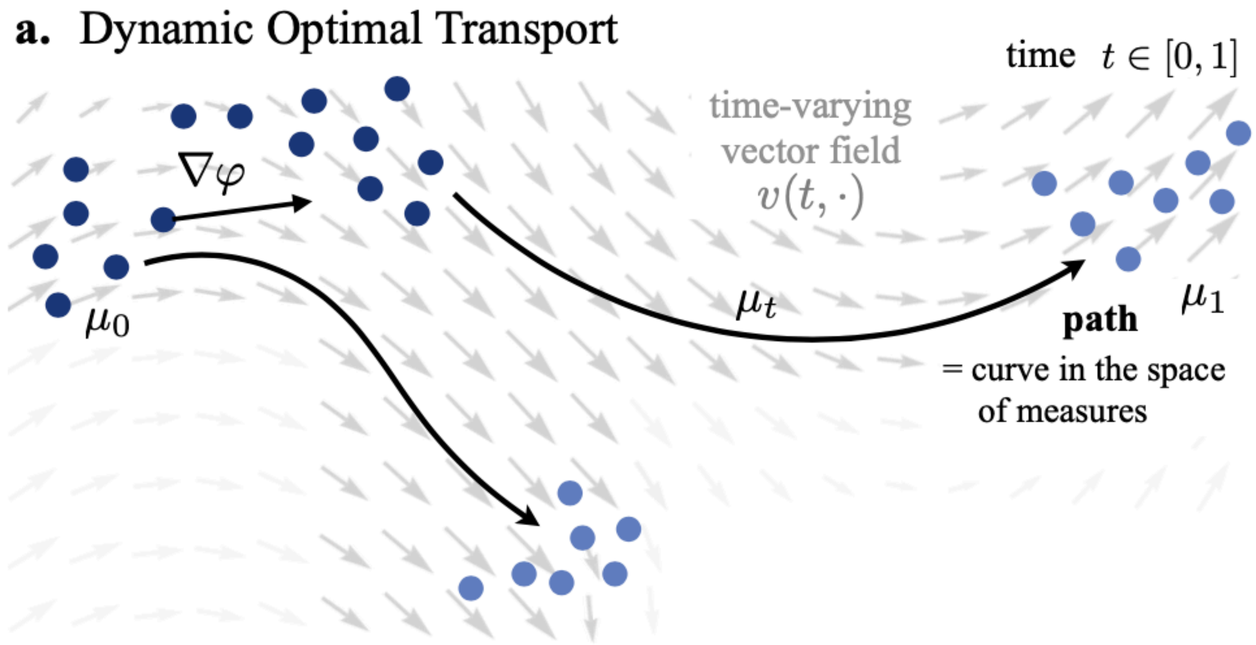
- The system is continuous in time and requires a dynamic description
- Our data subsamples a much larger state space $c := [k_1, \ldots, k_K] \in \mathscr{V}_K$
- The space is continuous and requires a continuous description
- Smooth and continuous inferred trajectories

Dynamic Optimal Transport

Continuous deformation with a flow field

- Recall in the K $y)\,dx\,dy$
- Brenier theore $= \nabla\varphi^*$
- By using the c $p_\nu(\nabla\varphi(x))$
- First contact v eying

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mathbf{v}\mu_t)$$



- Wasserstein distance becomes $W_2^2(\mu_0,\mu_1) = \min_{(\mu_t,v_t)} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \, d\mu_t(x)\, dt$

φ : smooth, continuous potential

T : Monge map

μ_t : measure

$\mathbf{v}$: vector field

Bunne, Charlotte, *Neural Optimal Transport for Dynamical Systems*

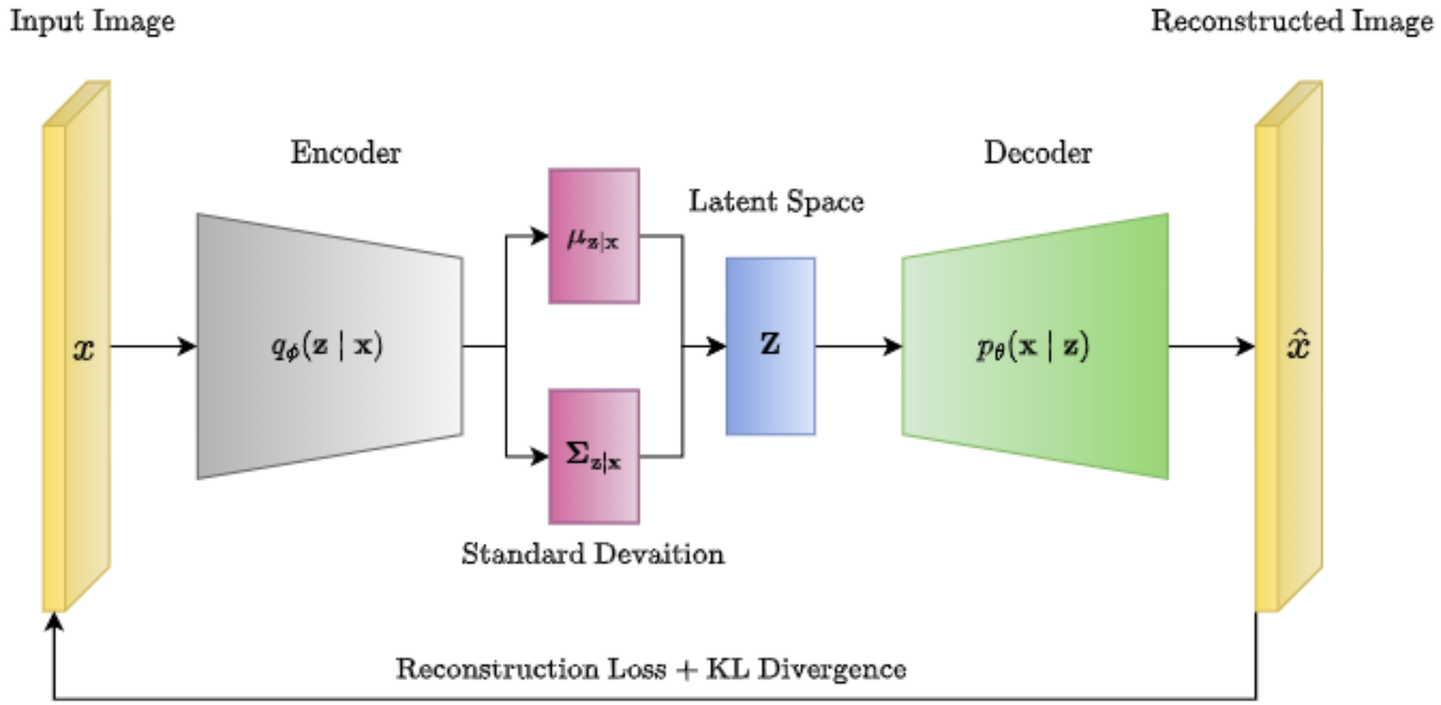
Dynamic Optimal Transport

Continuous description with a variational autoencoder (VAE)

- Our data is discrete but the space is continuous
- The cells are labeled by their surface markers $c := [k_1, \ldots, k_K] \in \mathscr{V}_K$; think of these as states in a large state space
- There is numerical bias due to sampling
- The data is big and there are memory issues. It's possible to represent the data with a lower dimensional probability distribution using a VAE
- Learning an effective latent representation means learning a greater state space

Dynamic Optimal Transport

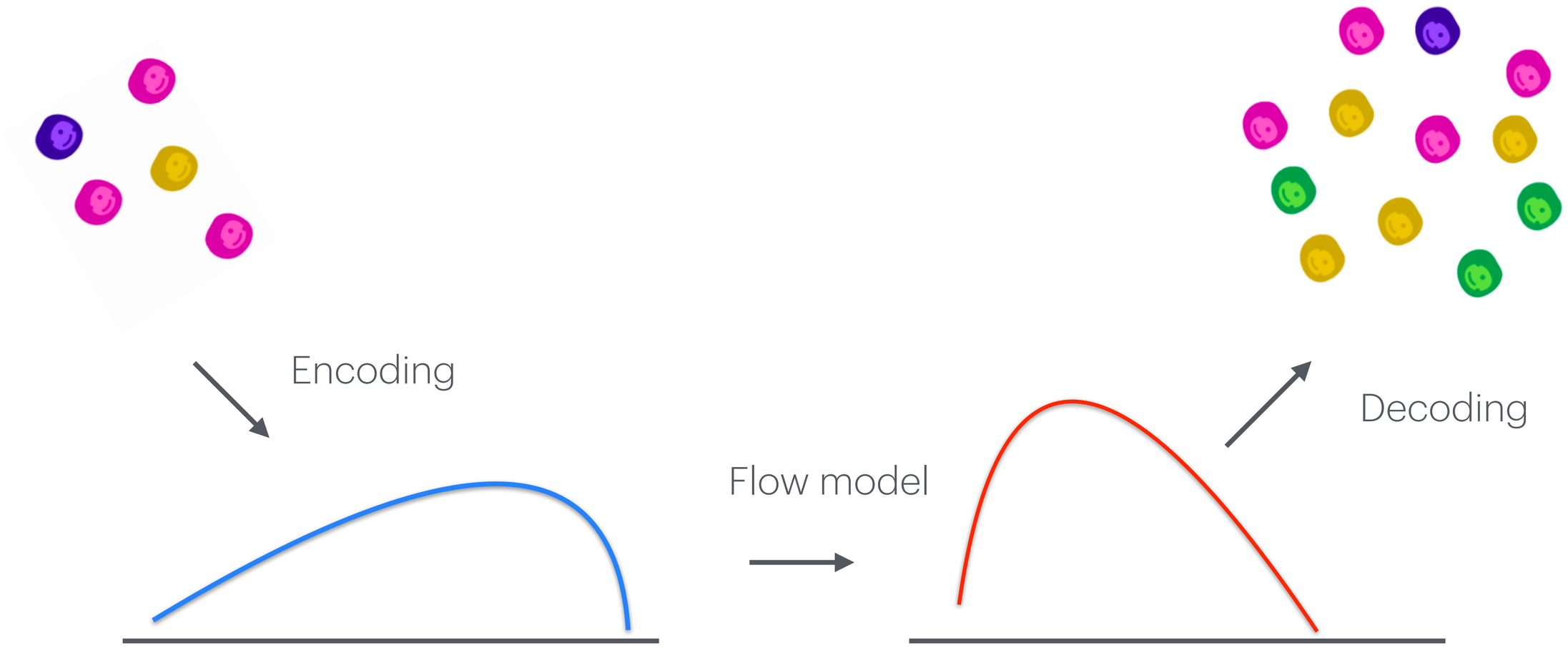
Review of VAEs



https://sushant-kumar.com/blog/vae-variational-autoencoder

Dynamic Optimal Transport

OT in the latent space



Dynamic Optimal Transport

OT in the latent space

- The prior measure is encoded in the latent space as a distribution
- Evolve the latent representation with a learnable kernel $\mathbf{k}_\theta$
- Define the flow field as $v(z,t) = \int \mathbf{k}_\theta(z,z')\,(z-z')\,d\rho_t(z')$
- The loss in the style of Benamou and Brenier is

$$\mathscr{L}_{\text{kinetic}} = \int_0^1 \int_{\mathbb{R}} d\rho_t(z)\,dt \frac{1}{2} \left| \int d\rho_t(z')\,\mathbf{k}_\theta(z,z')\,(z-z') \right|^2$$

- $\mathscr{L}_{\text{recon}} = \mathbb{E}_{z_1 \sim \rho_1'}[\,|\,\text{dec}(z_1) - z_{\text{target}}\,|^2\,] + \beta D_{\text{KL}}(\rho_1(z)\,\|\,p(z)) + \mathscr{L}_{\text{kinetic}}$

$\mathbf{k}_\theta$: learnable kernel
$\rho_t(z)$: latent distribution
$p(z)$: Gaussian distribution
$\text{dec}(\cdot)$: decoder
$D_{\text{KL}}(\cdot\|\cdot)$: KL divergence

Dynamic Optimal Transport

Variational autoencoder :

- Encode input distribution in latent space as probability distribution
- Evolve latent space distribution via learned flow field
- Sample transformed latent space distribution to reconstruct target distribution (no longer any problem with balanced or unbalanced)
- Wasserstein distance replaced with Benamou-Brenier formulation
- KL divergence between evolved T-cell population and real T-cell population