Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors

Geert-Jan Huizing, Laura Cantini, Gabriel Peyré

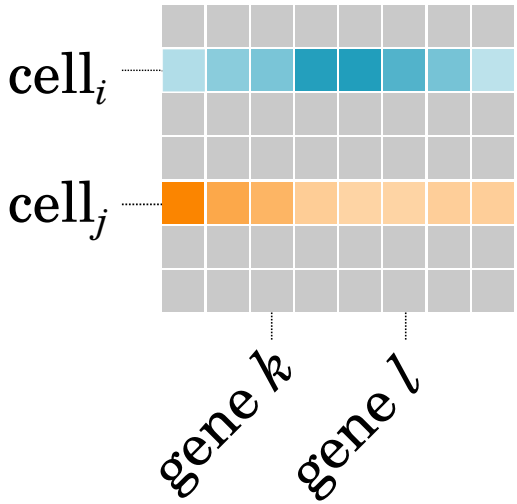
Comparing distributions

Defining **meaningful distances** between **distributions** is a fundamental topic in **machine learning**.

$a_i =$  , $a_j =$ 

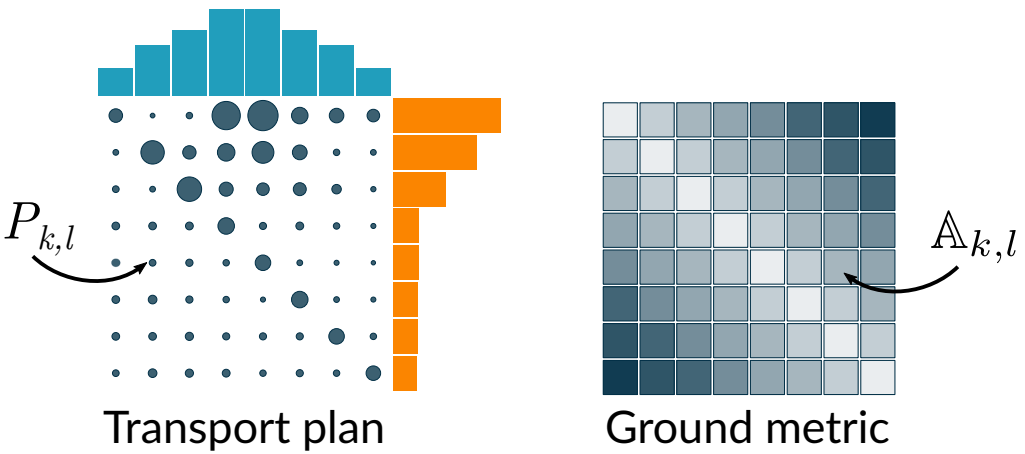
Example in **single-cell genomics**:

- **cells** as distributions over **genes**
- enables discovery of new cell types



Wasserstein Distance

Optimal Transport[1] computes the **transportation cost** from one *cell* to another. The **ground metric** $\mathbb{A}$ encodes a cost between *genes*.



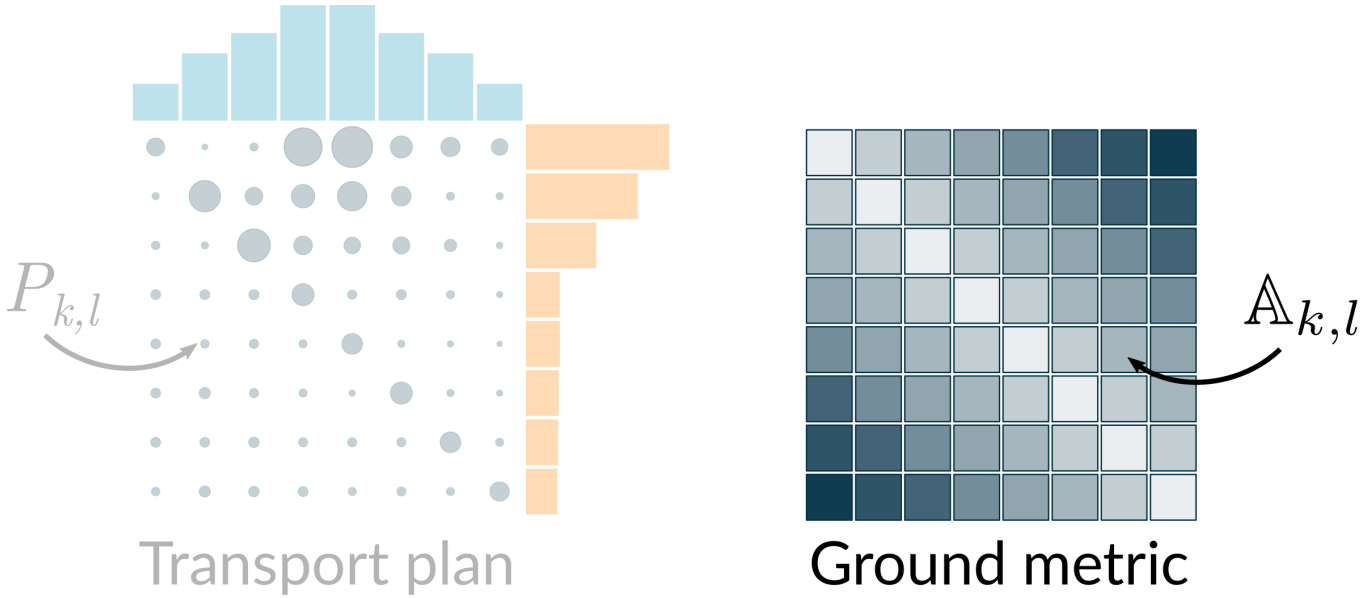
$$\mathrm{W}_{\mathbb{A}}(a_i, a_j) \overset{\text{def.}}{=} \min_{P \in \mathbb{R}_+^{n \times n}} \langle P, \mathbb{A} \rangle$$

$$\text{s.t. } P\mathbb{1}_n = a_i \text{ and } P^\top \mathbb{1}_n = a_j.$$

[1] Monge, 1781; Kantorovich, 1942

Ground Metric Learning

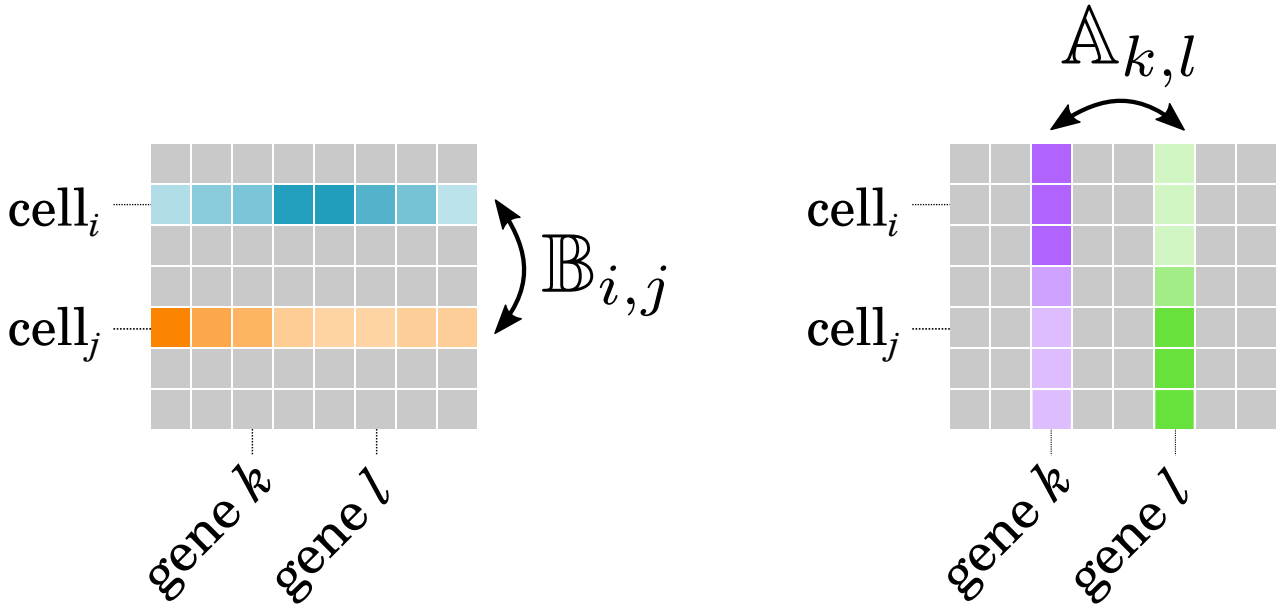
- No straightforward ground metric on genes.
- **Ground Metric Learning**[2] refers to learning the matrix $\mathbb{A}$.
- Unlabeled data motivates **unsupervised ground metric learning**.



[2] Cuturi & Avis, 2014; Wang & Guibas, 2012; Xu et al., 2021; Heitz et al., 2020

Distance on cells, distance on genes

- We call $\mathbb{B}$ the **Wasserstein distance matrix** between *cells* a_i, a_j.
- We require $\mathbb{A}$ to be a **distance matrix** between *genes* b_k, b_l.



Bootstrapping intuition

- Starting from some distance $\mathbb{A}$, we can define a **Wasserstein distance matrix** $\mathbb{B}$ between cells.

$$\mathbb{B}_{i,j} = \mathrm{W}_{\mathbb{A}}(a_i, a_j)$$

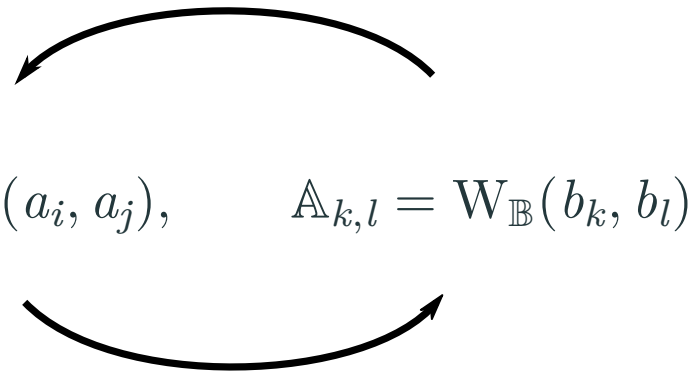
Bootstrapping intuition

- Starting from some distance $\mathbb{A}$, we can define a **Wasserstein distance matrix** $\mathbb{B}$ between cells.
- Then, we can use $\mathbb{B}$ to update $\mathbb{A}$.

$$\mathbb{B}_{i,j} = \mathrm{W}_{\mathbb{A}}(a_i, a_j), \qquad \mathbb{A}_{k,l} = \mathrm{W}_{\mathbb{B}}(b_k, b_l)$$

Bootstrapping intuition

- Starting from some distance $\mathbb{A}$, we can define a **Wasserstein distance matrix** $\mathbb{B}$ between cells.
- Then, we can use $\mathbb{B}$ to update $\mathbb{A}$. **And so on!**



$$\mathbb{B}_{i,j} = \mathrm{W}_{\mathbb{A}}(a_i, a_j), \qquad \mathbb{A}_{k,l} = \mathrm{W}_{\mathbb{B}}(b_k, b_l)$$

Wasserstein Singular Vectors

- More formally, we would like

$$\mu\mathbb{B} = \Phi_A(\mathbb{A}), \quad \lambda\mathbb{A} = \Phi_B(\mathbb{B}).$$

- Where Φ_A, Φ_B map ground costs to Wasserstein distance matrices

$$\Phi_A(\mathbb{A})_{i,j} = W_{\mathbb{A}}(a_i, a_j) + \tau R(a_i - a_j)$$

Theorem (Existence)
For $\tau > 0$, there exists a pair of Wasserstein Singular Vectors.

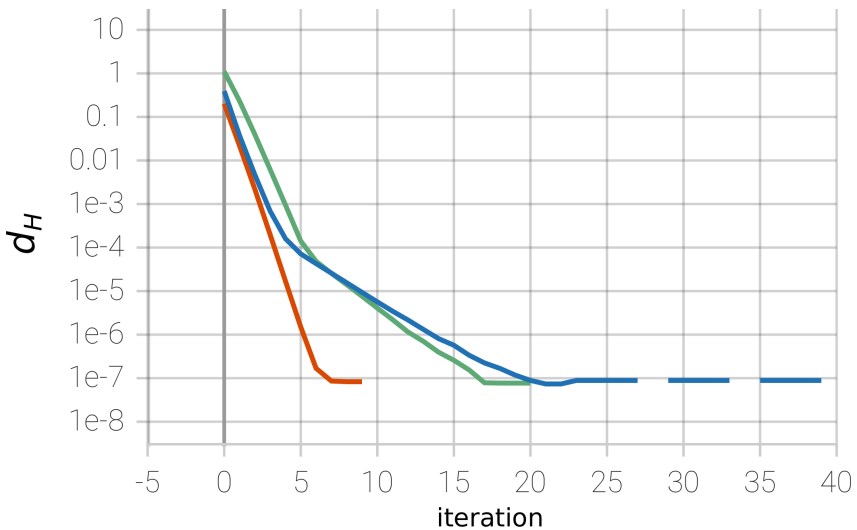
Power Iterations

This singular vectors problem can be solved using **power iterations**:

$$\mathbb{A}_{t+1} \overset{\text{def.}}{=} \frac{\Phi_B(\mathbb{B}_t)}{\|\Phi_B(\mathbb{B}_t)\|_\infty}, \qquad \mathbb{B}_{t+1} \overset{\text{def.}}{=} \frac{\Phi_A(\mathbb{A}_{t+1})}{\|\Phi_A(\mathbb{A}_{t+1})\|_\infty}.$$

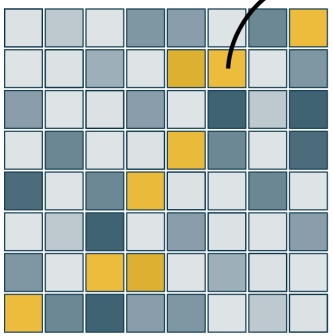
Theorem (Convergence)
For τ large enough, the power iterations ***converge linearly****.*

In practice, we observe convergence even for $\tau = 0$.



Stochastic Power Iterations

- **Issue:** computing n^2 distances per iteration is expensive.
- **Solution:** we show convergence for **stochastic power iterations**.

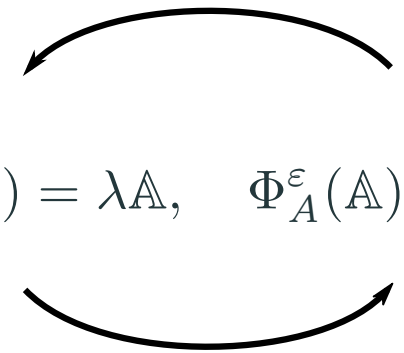


$\mathbb{B}_t$

$$(\tilde{\mathbb{B}}_t)_{i,j} \overset{\text{def.}}{=} \begin{cases} \Phi_A(\mathbb{A}_t)_{i,j}/\tilde{\mu}_t & \text{if } (i,j) \in \mathcal{I}, \\ (\mathbb{B}_t)_{i,j} & \text{otherwise.} \end{cases}$$

Sinkhorn Singular Vectors

- **Issue:** Optimal Transport is expensive in high dimension.
- **Solution:** we use **Sinkhorn Divergences**[3] instead.



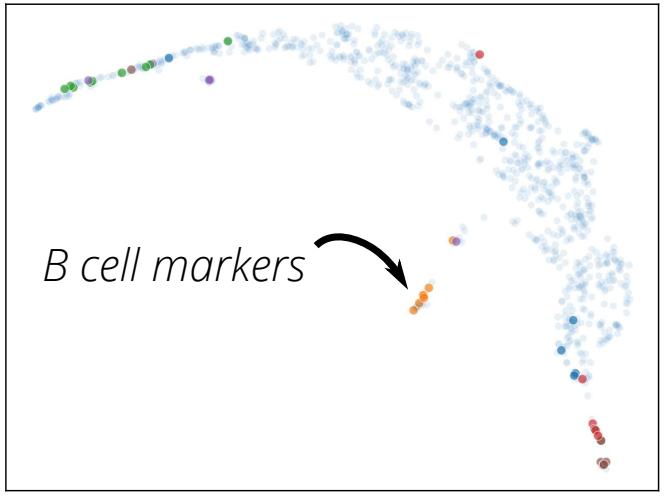
$$\Phi_B^\varepsilon(\mathbb{B}) = \lambda\mathbb{A}, \quad \Phi_A^\varepsilon(\mathbb{A}) = \mu\mathbb{B}$$

- Interpolation between **Wasserstein Singular Vectors** ($\varepsilon \to 0$) and **squared euclidean distances on PCA embeddings** ($\varepsilon \to \infty$).

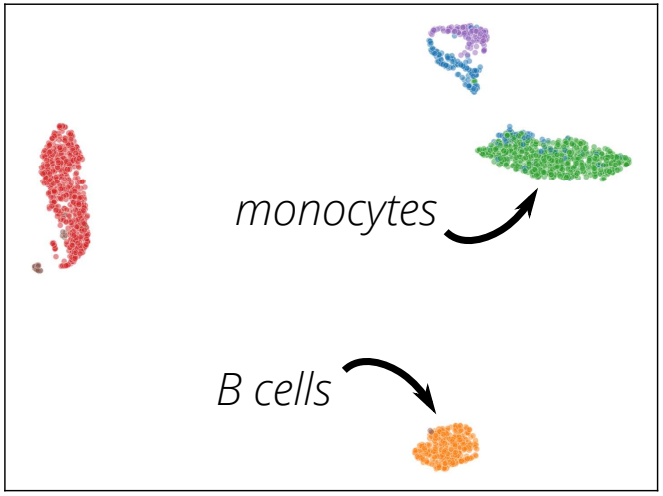
[3] Genevay et al., 2018

Sinkhorn Singular Vectors for single-cell genomics

- The method scales to $\sim$3,000 cells, and $\sim$1,000 genes.
- Biologically informative distances on cells *and* on genes.



UMAP projection of gene-gene distances $\mathbb{A}_{k,l}$



UMAP projection of cell-cell distances $\mathbb{B}_{i,j}$

Key takeaways

- **Wasserstein Singular Vectors** are a canonical pair of metrics on *samples* and on *features*
- **Stochastic** power iterations and **entropic OT** allow to scale the method to large datasets
- We demonstrated the method on **single-cell genomics**, but it can be applied to any positive dataset

Perspectives

- In-depth **theoretical analysis** when $\tau = 0$ and $\varepsilon > 0$
- Extension to **unbalanced optimal transport**
- Further results for **stochastic approximation**

Thanks!

 gjhuizing/wsingular

 pip install wsingular

 gjhuizing