Computational Optimal Transport for Machine and Deep Learning

Introductive course

Mathurin Massias, Titouan Vayer, Quentin Bertrand.

ENS DE LYON

November 20, 2024

Table of contents

About this course

Generalities

- About us: three researchers in machine learning/computer science.
- Course about the **computational aspects of optimal transport** and **its applications**.
- Three practical labs (Python).
- All details of the course here `https://mathurinm.github.io/otml/`.

Evaluation

- 50 % homeworks (6 homeworks: 4 small/ 2 longer).
- 50 % one project: paper presentation and extension of a selected research article and the associated code applied on real data.
- Bonus points: scribing (one per session, max 2 per person).

Acknowledgments

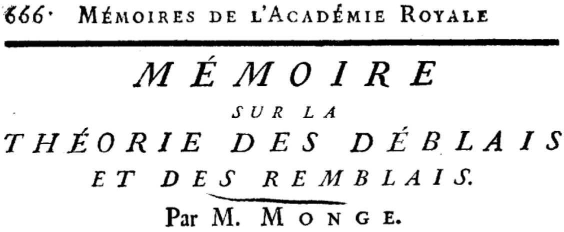
Some slides adapted from those of Rémi Flamary & Nicolas Courty.

Table of contents

A brief history

The natural geometry of probability measures

The origins of optimal transport

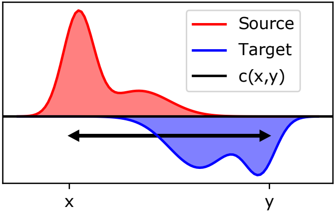

666· MÉMOIRES DE L'ACADÉMIE ROYALE

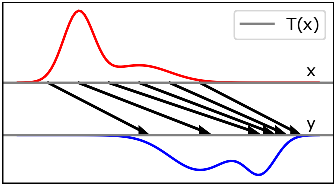
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. MONGE.

Problem Monge 1781

- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort ?

The origins of optimal transport

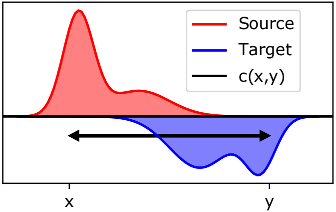


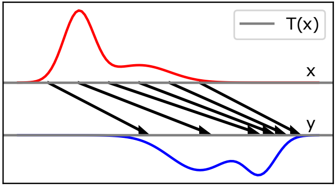


Problem Monge 1781

- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort ?

The origins of optimal transport

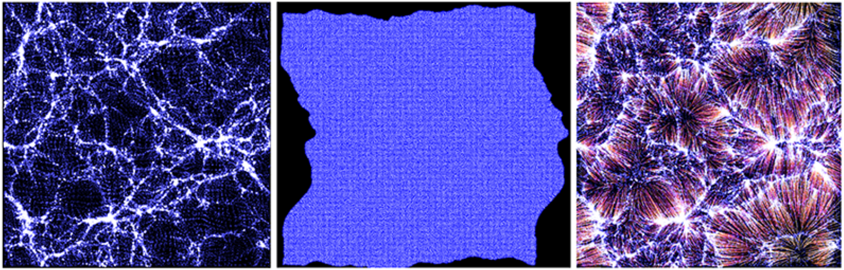




Problem Monge 1781

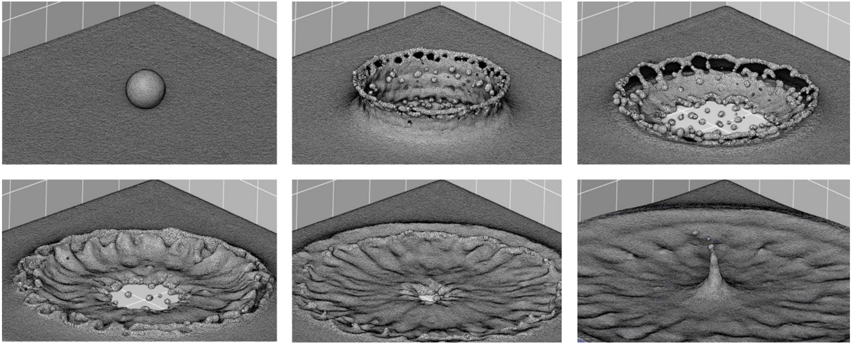
- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort ?
- Condorcet about Monge 1781: "Ainsi, l'on voit dans les Sciences, tantôt des théories brillantes, mais longtemps inutiles, devenir tout à coup le fondement des applications les plus importantes, et tantôt des applications très simples en apparence, faire naître l'idée de théories abstraites dont on n'avait pas encore le besoin, diriger vers les théories des travaux des Géomètres, et leur ouvrir une carrière nouvelle."

Some applications



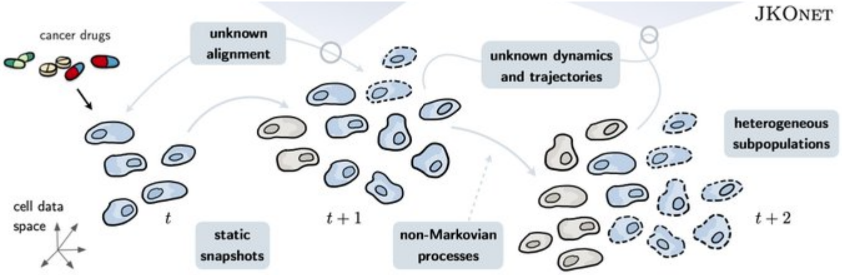
- Reconstruction of the early universe Levy, Mohayaee, and von-Hausegger 2021

Some applications



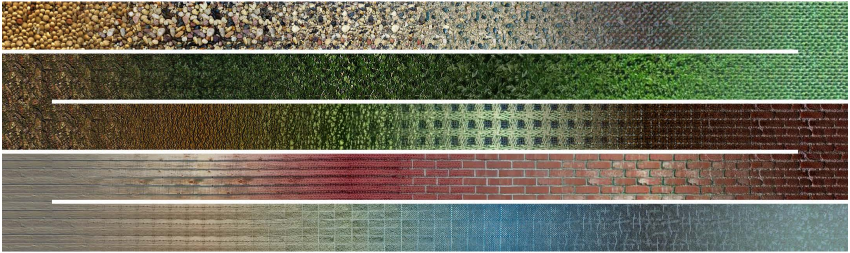
- Reconstruction of the early universe Levy, Mohayaee, and von-Hausegger 2021
- Fluid dynamics Lévy 2022

Some applications



- Reconstruction of the early universe Levy, Mohayaee, and von-Hausegger 2021
- Fluid dynamics Lévy 2022
- Cells analysis Bunne et al. 2024

Some applications



- Reconstruction of the early universe Levy, Mohayaee, and von-Hausegger 2021
- Fluid dynamics Lévy 2022
- Cells analysis Bunne et al. 2024
- Computer graphics, computer vision Bonneel and Digne 2023

Some applications

- Reconstruction of the early universe Levy, Mohayaee, and von-Hausegger 2021
- Fluid dynamics Lévy 2022
- Cells analysis Bunne et al. 2024
- Computer graphics, computer vision Bonneel and Digne 2023
- And machine learning !

Table of contents

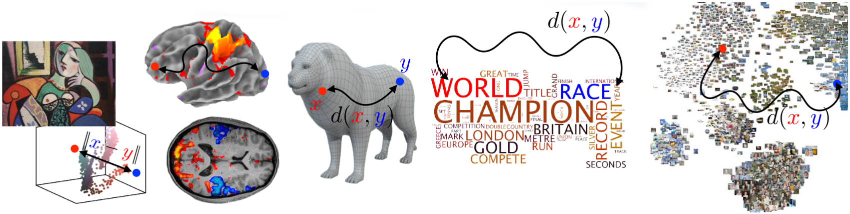
Distributions are everywhere

Distributions are everywhere in machine learning

- Images, vision, graphics, Time series, text, genes, proteins.
- Many datum and datasets can be seen as distributions.
- Important questions:
  - How to compare distributions?
  - How to use the geometry of distributions?
- Optimal transport provides many tools that can answer those questions.

Illustration from the slides of Gabriel Peyré.

Distributions are everywhere



Distributions are everywhere in machine learning

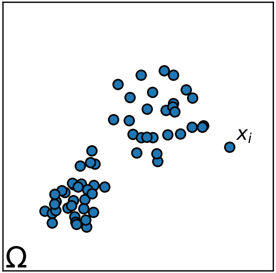
- Images, vision, graphics, Time series, text, genes, proteins.
- Many datum and datasets can be seen as distributions.
- Important questions:
  - How to compare distributions?
  - How to use the geometry of distributions?
- Optimal transport provides many tools that can answer those questions.

Illustration from the slides of Gabriel Peyré.

Discrete distributions: Empirical vs Histogram

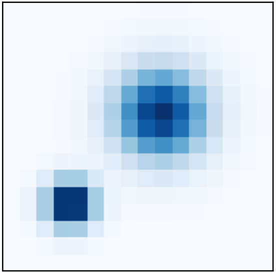
Discrete measure: $$\alpha = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}, \quad \mathbf{x}_i \in \Omega, \quad \sum_{i=1}^{n} a_i = 1$$

Lagrangian (point clouds)



- Constant weight: $a_i = \frac{1}{n}$
- Quotient space: Ω^n, Σ_n

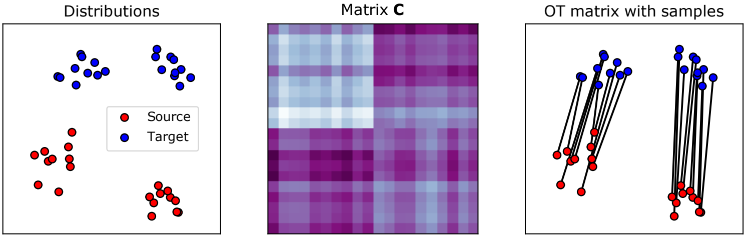
Eulerian (histograms)



- Fixed positions $\mathbf{x}_i$ e.g. grid
- Convex polytope Σ_n (simplex): $\{(a_i)_i \geq 0; \sum_i a_i = 1\}$

Table of contents

Optimal transport between discrete distributions



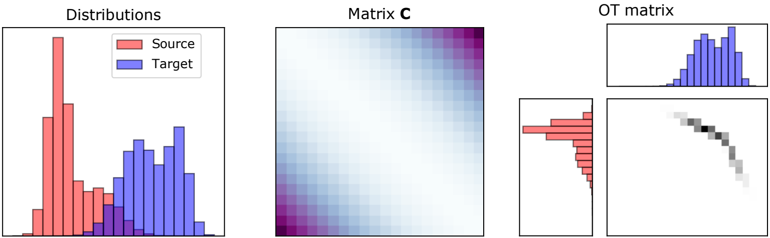
A matching problem

When $\alpha = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i}$ and $\beta = \frac{1}{n}\sum_{j=1}^{n}\delta_{\mathbf{y}_j}$

$$\min_{\sigma \in \mathrm{Perm}(n)} \sum_{i=1}^{n} C_{i,\sigma(i)}$$

where $\mathbf{C}$ is a cost matrix with $C_{i,j} = c(\mathbf{x}_i, \mathbf{y}_j)$.

Optimal transport between discrete distributions



Kantorovitch formulation : OT Linear Program

When $\alpha = \sum_{i=1}^n a_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{j=1}^m b_j \delta_{\mathbf{y}_j}$

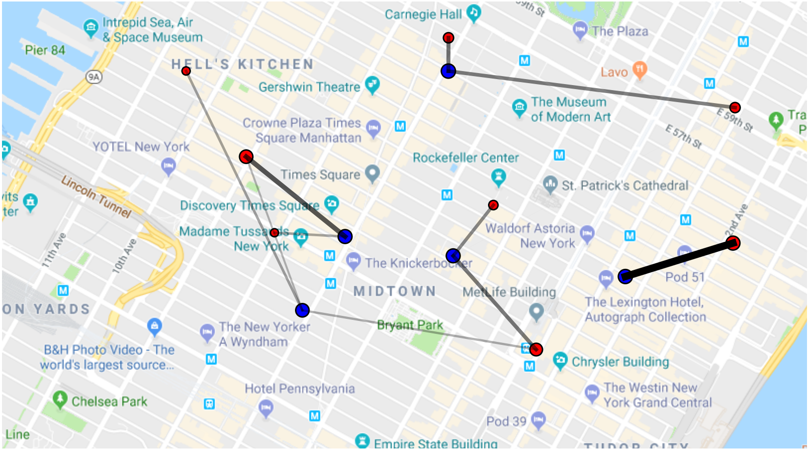
$$\min_{\boldsymbol{P} \in U(\mathbf{a},\mathbf{b})} \left\{ \langle \boldsymbol{P}, \mathbf{C} \rangle_F = \sum_{i,j} P_{i,j} C_{i,j} \right\}$$

where $\mathbf{C}$ is a cost matrix with $C_{i,j} = c(\mathbf{x}_i, \mathbf{y}_j)$ and

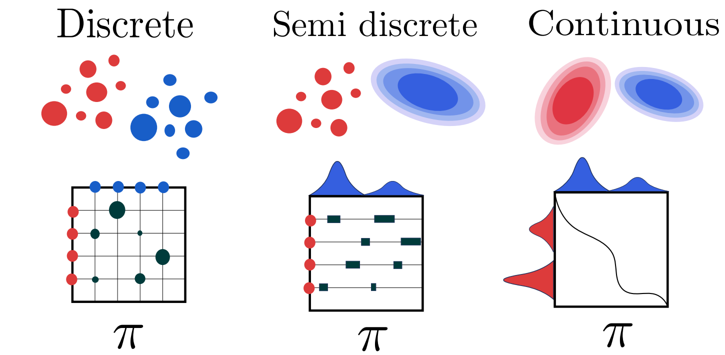
$$U(\mathbf{a}, \mathbf{b}) = \left\{ \boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \, \boldsymbol{P}\mathbf{1}_m = \mathbf{a}, \boldsymbol{P}^T \mathbf{1}_n = \mathbf{b} \right\}$$

- $(n = m)$ Solving OT with network simplex is $O(n^3 \log(n))$.

Boulangeries & Cafés



Wasserstein distance



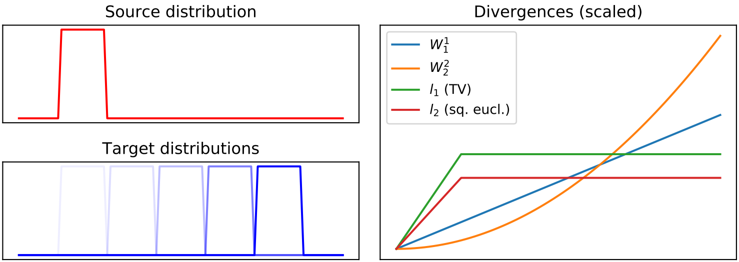
Wasserstein distance

Distance between two **arbitrary** prob. distributions $\alpha \in \mathcal{P}(\Omega)$ and $\beta \in \mathcal{P}(\Omega)$

$$W_p(\alpha, \beta) = \left(\min_{\pi \in U(\alpha,\beta)} \int_{\Omega \times \Omega} \|\mathbf{x} - \mathbf{y}\|^p d\pi(\mathbf{x}, \mathbf{y}) \right)^{1/p} = \left(\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\pi}[\|\mathbf{x} - \mathbf{y}\|^p] \right)^{1/p}$$

- $(\mathcal{P}(\Omega), W_p)$ is a metric space.
- Works for continuous and discrete distributions (histograms, empirical).

Wasserstein distance



Wasserstein distance

Distance between two **arbitrary** prob. distributions $\alpha \in \mathcal{P}(\Omega)$ and $\beta \in \mathcal{P}(\Omega)$

$$W_p(\alpha, \beta) = \left(\min_{\pi \in U(\alpha, \beta)} \int_{\Omega \times \Omega} \|\mathbf{x} - \mathbf{y}\|^p d\pi(\mathbf{x}, \mathbf{y}) \right)^{1/p} = \left(\mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \pi} [\|\mathbf{x} - \mathbf{y}\|^p] \right)^{1/p}$$

- $(\mathcal{P}(\Omega), W_p)$ is a metric space.
- Works for continuous and discrete distributions (histograms, empirical).

Table of contents

Some properties of optimal couplings

The Monge-Mather shortening principle

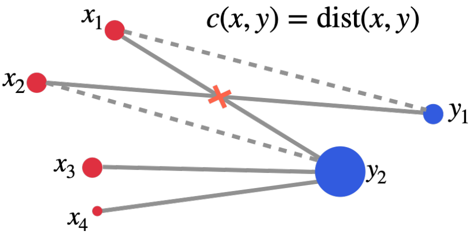
Let

$$\mathrm{supp}(\mathbf{P}) = \{(i,j) \in [n] \times [m] : P_{ij} > 0\}\,. \tag{1}$$

If $\mathbf{P}$ is an optimal coupling and $c(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$, then for any $(i_1,j_1),(i_2,j_2) \in \mathrm{supp}(\mathbf{P})^2$,

$$[\mathbf{x}_{i_1}, \mathbf{y}_{j_1}] \text{ and } [\mathbf{x}_{i_2}, \mathbf{y}_{j_2}] \text{ do not cross, except maybe at their endpoints}\,.$$

- Monge 1781 "Lorsque le transport du deblai se fait de manière que la somme des produits des molécules par l'espace parcouru est un minimum, les routes de deux points quelconques A & B, ne doivent plus se couper entre leurs extrémités, car la somme Ab + Ba des routes qui se coupent est toujours plus grande que la somme Aa + Bb de celles qui ne se coupent pas."



Some properties of optimal couplings

The Monge-Mather shortening principle

Let

$$\mathrm{supp}(\mathbf{P}) = \{(i,j) \in [n] \times [m] : P_{ij} > 0\}\,. \tag{1}$$

If $\mathbf{P}$ is an optimal coupling (and whatever $\mathbf{C}$)

$$\forall (i_1, j_1), (i_2, j_2) \in \mathrm{supp}(\mathbf{P})^2,\, C_{i_1,j_1} + C_{i_2,j_2} \leq C_{i_1,j_2} + C_{i_2,j_1}\,. \tag{2}$$

Some properties of optimal couplings

The Monge-Mather shortening principle

Let

$$\mathrm{supp}(\mathbf{P}) = \{(i,j) \in [n] \times [m] : P_{ij} > 0\}\,. \tag{1}$$

If $\mathbf{P}$ is an optimal coupling (and whatever $\mathbf{C}$)

$$\forall (i_1, j_1), (i_2, j_2) \in \mathrm{supp}(\mathbf{P})^2, C_{i_1,j_1} + C_{i_2,j_2} \leq C_{i_1,j_2} + C_{i_2,j_1}\,. \tag{2}$$

The main theorem of OT: cyclical monotonicty

A coupling $\mathbf{P} \in U(\mathbf{a}, \mathbf{b})$ is optimal **if and only if** for any $N \in \mathbb{N}^*, (i_1, j_1), \cdots, (i_N, j_N) \in \mathrm{supp}(\mathbf{P})^N$ and permutation $\sigma \in \mathrm{Perm}(N)$,

$$\sum_{k=1}^{N} C_{i_k,j_k} \leq \sum_{k=1}^{N} C_{i_k,j_{\sigma(k)}}\,. \tag{3}$$

Dual OT problem

The OT problem

$$\min_{\boldsymbol{P} \in U(\mathbf{a},\mathbf{b})} \langle \boldsymbol{P}, \mathbf{C} \rangle , \tag{Primal}$$

admits the dual formulation

$$\max_{\substack{\mathbf{f} \in \mathbb{R}^n, \mathbf{g} \in \mathbb{R}^m \\ \forall (i,j) \in [n] \times [m], f_i + g_j \leq C_{i,j}}} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle . \tag{Dual}$$

- If $\boldsymbol{P}^\star$ is a solution of (Primal) and $(\mathbf{f}^\star, \mathbf{g}^\star)$ is a solution of (Dual) then $\langle \boldsymbol{P}^\star, \mathbf{C} \rangle = \langle \mathbf{f}^\star, \mathbf{a} \rangle + \langle \mathbf{g}^\star, \mathbf{b} \rangle$
- Also for any $(i,j) \in \mathrm{supp}(\boldsymbol{P}^\star)$, $f_i^\star + g_j^\star = C_{i,j}$.

Table of contents

Maximal coupling and total variation

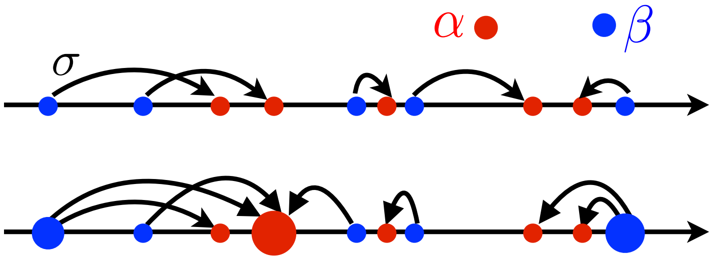
A simple special case

- When $n = m$ and $\alpha = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$ and $\beta = \sum_{j=1}^{n} b_j \delta_{\mathbf{y}_j}$.
- Cost $C_{i,j} = 1 - \delta_i(j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{otherwise} \end{cases}$.
- One optimal coupling is the "maximal coupling"

$$P_{ii} = \min(a_i, b_i) \text{ and } i \neq j, P_{ij} = \frac{(a_i - \min(a_i, b_i))(b_j - \min(a_j, b_j))}{1 - \sum_k \min(a_k, b_k)} \tag{4}$$

- Smallest OT cost is the total variation $\min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle = \frac{1}{2} \|\mathbf{a} - \mathbf{b}\|_1$.

Special case: 1D distribution



A important special case

When $x_i, y_j \in \mathbb{R}$ and $c(x, y) = h(x - y)$ where h is convex.

- Example $h(x - y) = |x - y|^2$.
- If $x_1 \leq x_2$ and $y_1 \leq y_2$, we can check that

$$c(x_1, y_1) + c(x_2, y_2) \leq c(x_1, y_2) + c(x_2, y_1) \tag{5}$$

- **Optimal plan respects the ordering of the elements**.
- Wery simple algorithm to compute the transport in $O(\max\{n, m\} \log(\max\{n, m\}))$, by sorting both x_i and y_j.

Special case: 1D distribution

The north-west corner rule

Initialize $\bar{\mathbf{a}} = \mathbf{a}, \bar{\mathbf{b}} = \mathbf{b}$, and $(i, j) = (1, 1)$.

While $i \leq n, j \leq m$ do:

- Send as much mass possible from i to j: $P_{ij} = \min\{\bar{a}_i, \bar{b}_j\}$.
- Adjust marginals $\bar{a}_i \leftarrow \bar{a}_i - P_{ij}$, $\bar{b}_j \leftarrow \bar{b}_j - P_{ij}$.
- If $\bar{a}_i = 0$ (marginal is saturated) then $i \leftarrow i + 1$.
- Si $\bar{b}_j = 0$ (marginal is saturated) then $j \leftarrow j + 1$.

Return $\mathbf{P}$.

This algorithm runs in $O(n + m)$ operations.

Ex 1

$$\begin{array}{cc} \begin{bmatrix} \; & \; & \; \\ \; & \; & \; \\ \; & \; & \; \end{bmatrix} & \overset{\mathbf{a}}{\begin{pmatrix} 0.3 \\ 0.4 \\ 0.3 \end{pmatrix}} \\ \mathbf{b}\ (0.5 \quad 0.3 \quad 0.2) & \end{array}$$

iter 1

$$\begin{array}{ccc} \begin{bmatrix} 0.3 & 0 & 0 \\ & & \\ & & \end{bmatrix} & \begin{pmatrix} 0.3 \\ 0.4 \\ 0.3 \end{pmatrix} & \overset{\bar{\mathbf{a}}}{\begin{pmatrix} 0 \\ 0.4 \\ 0.3 \end{pmatrix}} \\ (0.5 \quad 0.3 \quad 0.2) & & \\ \bar{\mathbf{b}}\ (0.2 \quad 0.3 \quad 0.2) & & \end{array}$$

iter 2 (i_ℓ, j_ℓ)

$$\begin{array}{ccc} \begin{bmatrix} 0.3 & 0 & 0 \\ 0.2 & & \\ & & \end{bmatrix} & \begin{pmatrix} 0.3 \\ 0.4 \\ 0.3 \end{pmatrix} & \begin{pmatrix} 0 \\ 0.2 \\ 0.3 \end{pmatrix} \\ (0.5 \quad 0.3 \quad 0.2) & & \\ (0 \quad 0.3 \quad 0.2) & & \end{array}$$

Special case: 1D distribution

The north-west corner rule

Initialize $\bar{\mathbf{a}} = \mathbf{a}, \bar{\mathbf{b}} = \mathbf{b}$, and $(i, j) = (1, 1)$.

While $i \leq n, j \leq m$ do:

- Send as much mass possible from i to j: $P_{ij} = \min\{\bar{a}_i, \bar{b}_j\}$.
- Adjust marginals $\bar{a}_i \leftarrow \bar{a}_i - P_{ij}$, $\bar{b}_j \leftarrow \bar{b}_j - P_{ij}$.
- If $\bar{a}_i = 0$ (marginal is saturated) then $i \leftarrow i + 1$.
- Si $\bar{b}_j = 0$ (marginal is saturated) then $j \leftarrow j + 1$.

Return $\mathbf{P}$.

This algorithm runs in $O(n + m)$ operations.

Ex 2

$$\begin{array}{c} \boxed{\begin{matrix} & & \\ & & \\ & & \end{matrix}} \\ (0.5 \quad 0.3 \quad 0.2) \end{array} \begin{pmatrix} 0.5 \\ 0.4 \\ 0.1 \end{pmatrix} \qquad \begin{array}{c} \boxed{\begin{matrix} 0.5 & 0 & 0 \\ 0 & & \\ 0 & & \end{matrix}} \\ (0.5 \quad 0.3 \quad 0.2) \\ (0 \quad 0.3 \quad 0.2) \end{array} \begin{pmatrix} 0.5 \\ 0.4 \\ 0.1 \end{pmatrix} \begin{pmatrix} 0 \\ 0.4 \\ 0.3 \end{pmatrix} \qquad \begin{array}{c} \boxed{\begin{matrix} 0.5 & 0 & 0 \\ 0 & 0.3 & \\ 0 & & \end{matrix}} \\ (0.5 \quad 0.3 \quad 0.2) \\ (0 \quad 0 \quad 0.2) \end{array} \begin{pmatrix} 0.5 \\ 0.4 \\ 0.1 \end{pmatrix} \begin{pmatrix} 0 \\ 0.1 \\ 0.3 \end{pmatrix}$$

Special case: 1D distribution

The north-west corner rule

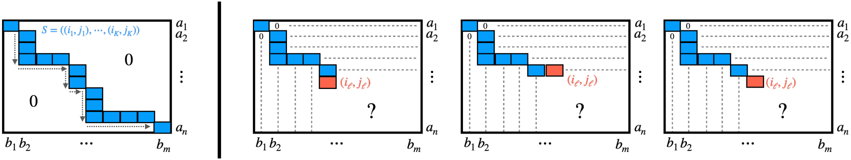
Initialize $\bar{\mathbf{a}} = \mathbf{a}, \bar{\mathbf{b}} = \mathbf{b}$, and $(i, j) = (1, 1)$.

While $i \leq n, j \leq m$ do:

- Send as much mass possible from i to j: $P_{ij} = \min\{\bar{a}_i, \bar{b}_j\}$.
- Adjust marginals $\bar{a}_i \leftarrow \bar{a}_i - P_{ij}$, $\bar{b}_j \leftarrow \bar{b}_j - P_{ij}$.
- If $\bar{a}_i = 0$ (marginal is saturated) then $i \leftarrow i + 1$.
- Si $\bar{b}_j = 0$ (marginal is saturated) then $j \leftarrow j + 1$.

Return $\mathbf{P}$.

This algorithm runs in $O(n + m)$ operations.



Special case: 1D distribution

Monge matrices

A matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$ is a Monge matrix if

$$\forall (i,j) \in [n] \times [m], C_{i,j} + C_{i+1,j+1} \leq C_{i+1,j} + C_{i,j+1} \tag{6}$$

- When $x_1 \leq \cdots \leq x_n, y_1 \leq \cdots \leq y_m$ then $\mathbf{C} = \left(|x_i - y_j|^2\right)_{i,j}$ is a Monge matrix.
- More generally, $\mathbf{C} = (h(x_i - y_j))_{i,j}$ with h convex.
- It is equivalent to

$$\forall 1 \leq i < r \leq n, 1 \leq j < s \leq m,\ \ C_{i,j} + C_{r,s} \leq C_{i,s} + C_{r,j} \tag{7}$$

Main result

If $\mathbf{C}$ is a Monge matrix the north-west corner rule produces an optimal coupling.

- Corollary: in 1D you can solve OT in $O(\max\{n,m\} \log(\max\{n,m\}))$.

References I

Bonneel, Nicolas and Julie Digne (2023). “A survey of optimal transport for computer graphics and computer vision”. In: *Computer Graphics Forum*. Vol. 42. 2. Wiley Online Library, pp. 439–460.

Bunne, Charlotte et al. (2024). “Optimal transport for single-cell and spatial omics”. In: *Nature Reviews Methods Primers* 4.1, p. 58.

Levy, Bruno, Roya Mohayaee, and Sebastian von-Hausegger (June 2021). “A fast semidiscrete optimal transport algorithm for a unique reconstruction of the early Universe”. In: *Monthly Notices of the Royal Astronomical Society* 506.1, pp. 1165–1185.

Lévy, Bruno (2022). “Partial optimal transport for a constant-volume Lagrangian mesh with free boundaries”. In: *Journal of Computational Physics* 451, p. 110838.

Monge, Gaspard (1781). *Mémoire sur la théorie des déblais et des remblais*. De l’Imprimerie Royale.