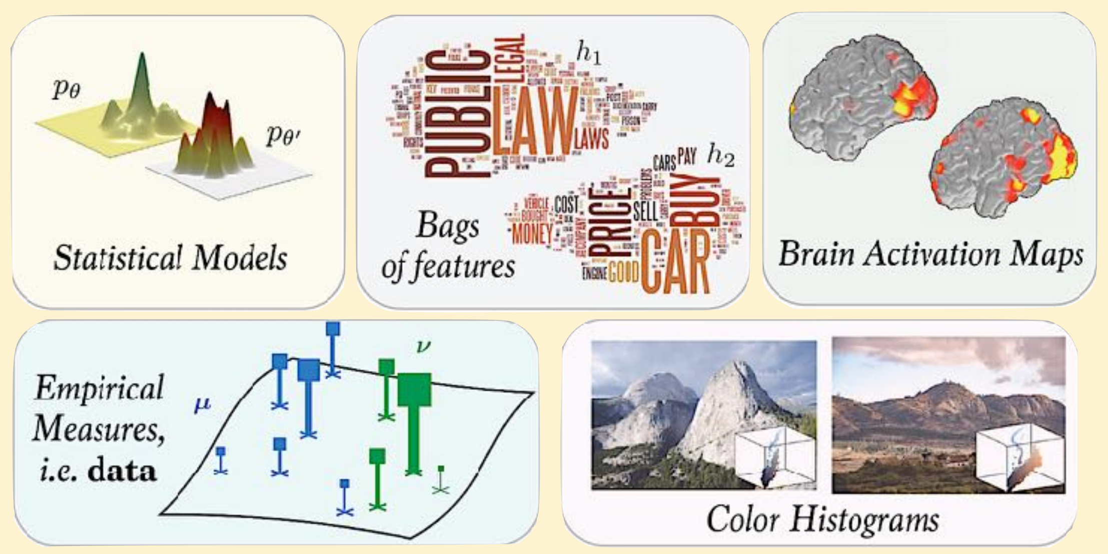
Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration

Neural Information Processing Systems Foundation

Jason Altschuler, Jonathan Weed, Philippe Rigollet

Massachusetts Institute of Technology

Optimal transport



Statistical primitive with many applications in machine learning and optimization

Given:

C: cost matrix $\in \mathbb{R}_+^{n\times n}$

r, c: probability distributions $\in \Delta_n$

OT distance: $\min \langle C, P\rangle$

$$\left.\begin{aligned} \text{s.t. } & P\in\mathbb{R}_+^{n\times n} \\ & P\mathbf{1} = r \\ & P^\top \mathbf{1} = c \end{aligned}\right\} \text{transport polytope } \mathcal{U}_{r,c}$$

Goal: find $\hat{P} \in \mathcal{U}_{r,c}$ satisfying

$$\langle C, \hat{P}\rangle \leq \min_{P\in\mathcal{U}_{r,c}} \langle C, P\rangle + \varepsilon$$

Images courtesy M. Cuturi

Algorithm

1. Approximately solve program with **entropic penalty**

$$\min_{P\in\mathcal{U}_{r,c}} \langle C, P\rangle - \eta^{-1} H(P) \qquad (*)$$

with $\eta \approx \varepsilon^{-1}\log n$

2. Round approximate solution to $\mathcal{U}_{r,c}$ (see paper!)

From entropy to scaling

Theorem [Cuturi 2013]**:** The penalized program (*) has a unique solution:

$$\operatorname*{argmin}_{P\in\mathcal{U}_{r,c}} \langle C, P\rangle - \eta^{-1}H(P) = \Pi_{\mathcal{S}}(A)$$

where:

$A = \exp(-\eta C)$ (entrywise)

$\Pi_{\mathcal{S}}(\cdot)$ is Sinkhorn (Bregman) projection onto $\mathcal{U}_{r,c}$

Penalized **OT** reduces to **matrix scaling**

Our contributions

Simple and **practical** algorithm to approximate OT distance between distributions on n points in $\tilde{O}(n^2\varepsilon^{-4})$ time.

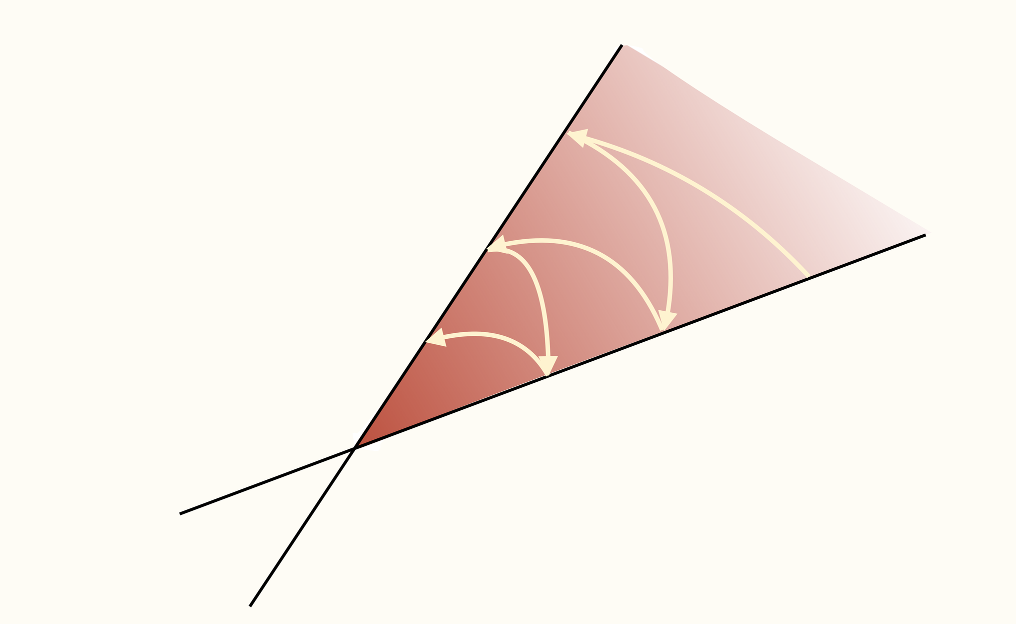
Based on **matrix scaling**, an approach pioneered by [Cuturi 2013]

We give:

New analysis of 50-year-old algorithm (**Sinkhorn scaling**)
New algorithm with much better performance in practice (**Greekhorn scaling**)

Same near-linear time convergence guarantee!

Sinkhorn scaling



Alternating Bregman projection onto polytopes

At each step, rescale *all* rows or *all* columns [$O(n^2)$ time]

Potential based analysis (based on dual program):

Nonnegative potential $f(A)$ decreases at each step by

$$\mathcal{K}(r \,\|\, A\mathbf{1}) + \mathcal{K}(c \,\|\, A^\top\mathbf{1})$$

While $\mathcal{K}(r \,\|\, A\mathbf{1}) + \mathcal{K}(c \,\|\, A^\top\mathbf{1}) \geq \varepsilon^2$,

$\tilde{O}(1)$ $\geq \varepsilon^2$ $\geq \varepsilon^2$ 0

(large improvement)

After $\tilde{O}(\varepsilon^{-2})$ iterations,

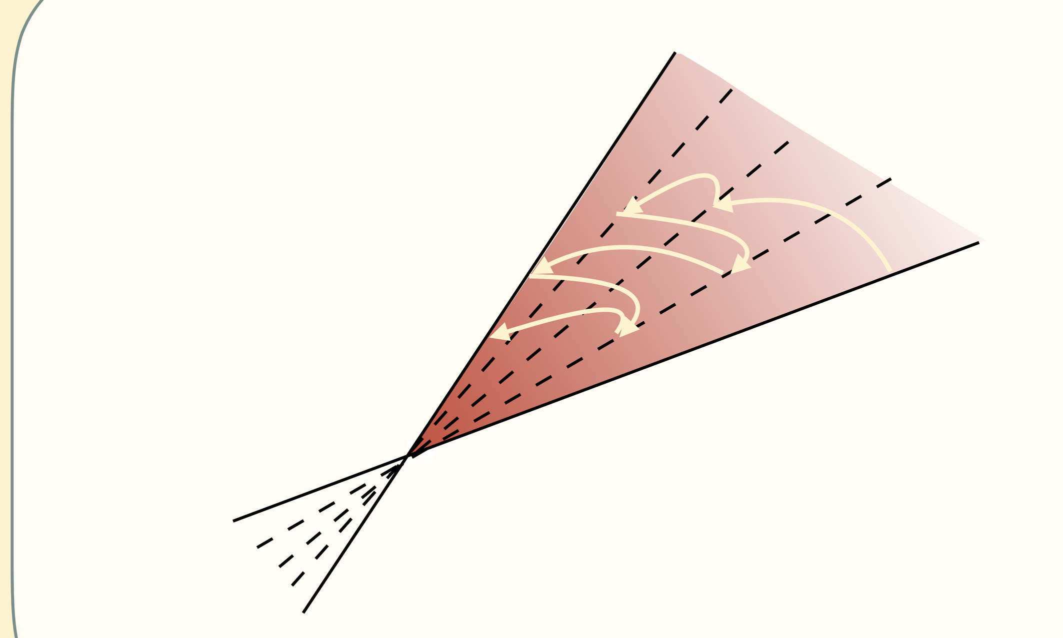
$$\|r - A\mathbf{1}\|_1 + \|c - A^\top\mathbf{1}\|_1 \lesssim \sqrt{\mathcal{K}(r \,\|\, A\mathbf{1}) + \mathcal{K}(c \,\|\, A^\top\mathbf{1})} \leq \varepsilon$$

(Pinsker)

Total runtime: $O(n^2)\cdot\tilde{O}(\varepsilon^{-2}) = \tilde{O}(n^2\varepsilon^{-2})$

cost per iteration number of iterations

Greenkhorn scaling



Greedy Bregman projection onto **simpler** polytopes

At each step, rescale *worst* row or column [$O(n)$ time]

Potential based analysis (based on dual program):

Nonnegative potential $f(A)$ decreases at each step by

$$\frac{1}{2n}[\rho(r \,\|\, A\mathbf{1}) + \rho(c \,\|\, A^\top\mathbf{1})]$$

appropriate generalization of $\mathcal{K}$

While $\rho(r \,\|\, A\mathbf{1}) + \rho(c \,\|\, A^\top\mathbf{1}) \geq \varepsilon^2$,

$\tilde{O}(1)$ $\geq \varepsilon^2/n$ $\geq \varepsilon^2/n$ 0

(large improvement)

After $\tilde{O}(n\varepsilon^{-2})$ iterations,

$$\|r - A\mathbf{1}\|_1 + \|c - A^\top\mathbf{1}\|_1 \lesssim \sqrt{\rho(r \,\|\, A\mathbf{1}) + \rho(c \,\|\, A^\top\mathbf{1})} \leq \varepsilon$$

(variant of Pinsker)

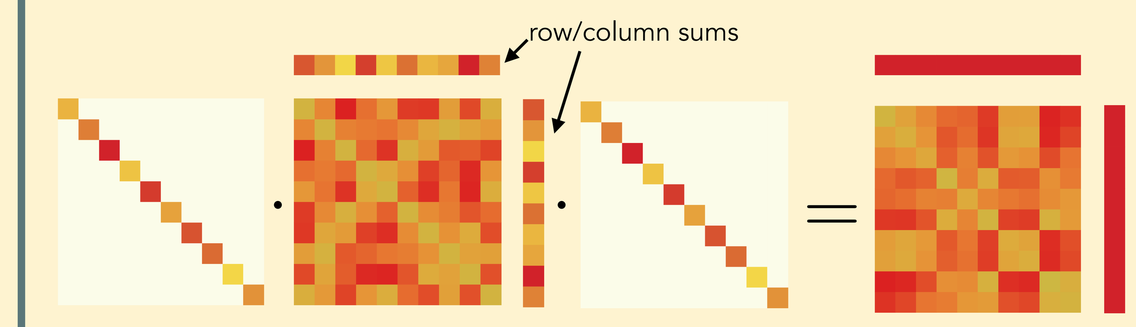
Total runtime: $O(n)\cdot\tilde{O}(n\varepsilon^{-2}) = \tilde{O}(n^2\varepsilon^{-2})$

cost per iteration number of iterations

References

J. Altschuler, J. Weed, P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30, 2017.*
M. Cuturi. Sinkhorn Distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, 2013.
R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 1967.

Matrix scaling



$$X \cdot A \cdot Y = \Pi_{\mathcal{S}}(A)$$

Matrix primitive with many applications in theoretical computer science and numerical linear algebra

Given:

A: matrix $\in \mathbb{R}_+^{n\times n}$

r, c: desired row/column sums $\in \mathbb{R}_+^n$

Goal: find positive diagonal matrices X, Y such that

$$XAY\mathbf{1} = r$$
$$(XAY)^\top\mathbf{1} = c$$

$\Pi_{\mathcal{S}}(A) := XAY$ is the **Sinkhorn projection**.

[Sinkhorn 1967] showed that it can be computed by alternating rescalings of rows and columns, but did not give effective bounds on rate of convergence

Empirical results