Greedy stochastic algorithms for entropy-regularized optimal transport problems

Brahim Khalil Abid and Robert M. Gower
brahim-khalil.abid@polytechnique.edu robert.gower@telecom-paristech.fr

1. Introduction

Probability distributions are the backbone of machine learning and statistics. Optimal Transport (OT) provides a meaningful notion of distance between probability distributions and histograms. Here we develop a family of fast and practical stochastic algorithms for solving the optimal transport problem with an entropic penalization.

2. The discrete OT problem

The regularised discrete OT problem can be seen as an optimal resource allocation given by:

$$P^*_\lambda = \arg\min_{P\in\mathbb{R}^{n\times n}_+} \langle P, C\rangle - \frac{1}{\lambda}E(P), \quad \text{subject to} \quad P\mathbf{1} = r,\ P^\top\mathbf{1} = c, \quad (1)$$

where the entropy is $E(P) = \sum_{i,j=1}^n -P_{ij}\log(P_{ij})$, $r, c \in \Delta_n \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \,|\, \sum_{i=1}^n x_i = 1\}$ are respectively the initial and target distributions, $C \in \mathbb{R}^{n\times n}_+$ the transport cost matrix and $\mathbf{1} \in \mathbb{R}^n$ is a vector of all ones.

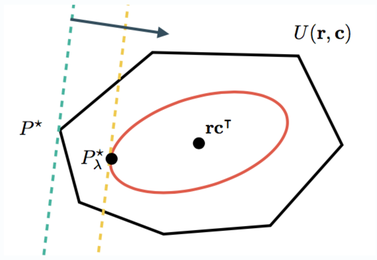

Figure 1: Regularized transport polytope (thanks to Michiel Stock)

3. Equivalence to matrix scaling

Let $A \stackrel{\text{def}}{=} e^{-\lambda C}$. The dual formulation of (1) is

$$(x^*, y^*) = \arg\max_{x,y} \sum_{i,j=1}^n A_{ij}e^{x_i+y_j} - \langle r, x\rangle - \langle c, y\rangle,$$

where $P^*_\lambda = D(e^{x^*})AD(e^{y^*})$. Let $u = e^x$ and $v = e^y$. Writing out the first order optimality conditions of the dual gives the *matrix scaling* problem:

$$D(u)AD(v)\mathbf{1} = r \quad \text{and} \quad D(v)A^\top D(u)\mathbf{1} = c$$

We design new stochastic methods for finding the (u, v) solution to this matrix scaling problem.

4. Transport Polytope

Let $r(P) = P\mathbf{1}$ and $c(P) = P^\top\mathbf{1}$ denote the *row sum* and *column sum* vectors of P. Let $U_{r,c}$ be the transport polytope defined by

$$U_{r,c} \stackrel{\text{def}}{=} \{P \in \mathbb{R}^{n\times n}_+ \,|\, r(P) = r,\ c(P) = c\}.$$

To measure the distance from the transport polytope we use

$$dist(A, U_{r,c}) \stackrel{\text{def}}{=} \|r(A) - r\|_1 + \|c(A) - c\|_1 \quad (2)$$

5. Sinkhorn Algorithm

The Sinkhorn algorithms efficiently solves the matrix scaling problem using only matrix vector products (Cuturi 2013)

Input: $A = e^{-\lambda C} \in \mathbb{R}^{n\times n}_+$, $r, c \in \mathbb{R}^n_+$, $\epsilon > 0$
Initialization: $u,v = \mathbf{1}$
while $\text{dist}(D(u)AD(v),U_{r,c}) \geq \epsilon$ **do**
  $u = r./(Av)$
  $v = c./(A^\top u)$
Output: $u, v \in R^n_+$.

7. Greedy Stochastic Sinkhorn

We propose the **Greedy Stochastic Sinkhorn (GSS)** algorithm based on selecting rows / columns according to an increasing probability function.

Input: $A = e^{-\lambda C} \in \mathbb{R}^{n\times n}_+$, $r, c \in \mathbb{R}^n_+$, $\epsilon > 0$
Initialization: $u,v = \mathbf{1}$
while $\text{dist}(D(u)AD(v),U_{r,c}) \geq \epsilon$ **do**
  Let $p = \Psi(\rho(D(u)AD(v))) \in \Delta_{2n}$
  Sample $i \sim p_i$ where $i \in \{1, 2, \ldots, 2n\}$
  if $i \leq n$ (row update) **then**
    $u_i = r_i./(Av)_i$
  else if $i > n$ (column update) **then**
    $v_{i-n} = c_{i-n}./(A^\top u)_{i-n}$
Output: $u, v \in \mathbb{R}^n_+$

8. Convergence theorem

Theorem 2 *Let $l = \min_{i,j}|A_{ij}|$, $s = \|A\|_1$. and $A^k \stackrel{def}{=} D(u^k)AD(v^k)$ be the iterates produced by the Greedy Stochastic Sinkhorn Algorithm. For a given $\epsilon > 0$ and every increasing probability function Ψ, we have that there exists $k \in \mathbb{N}$ such that*

$$k \leq \frac{28n}{\epsilon^2}\log\left(\frac{s}{l}\right) \quad \Rightarrow \quad \mathbf{E}\left[dist(A^k, U_{r,c})\right] \leq \epsilon.$$

6. Sampling rows/columns

To measure violations of each row/column of a matrix with respect to the transport polytope we use

$$\rho(a,b) = b - a + a\log(\frac{a}{b}), \quad \text{for } a, b \in \mathbb{R}_+$$
$$\rho(P) = \big(\rho(r_1, r_1(P)), \ldots, \rho(r_n, r_n(P)), \rho(c_1, c_1(P)), \ldots, \rho(c_n, c_n(P))\big) \in \mathbb{R}^{2n}$$

Using $\rho(P)$ we now define probability distributions that prioritize the most violated rows or columns.

Definition 1 *Let $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a positive and increasing function. We say that Ψ where*

$$\Psi(h) = \left(\frac{g(h_k)}{\sum_{i=1}^{2n} g(h_i)}\right)_{k=1..2n} \in \Delta_{2n}, \quad \forall h \in \mathbb{R}^{2n}_+$$

is an increasing probability function.

Several examples of an increasing probability function are given as follows

$$\Psi(h) = \left(\frac{h_i^\alpha}{\sum_{j=1..2n} h_j^\alpha}\right)_{i=1,\ldots,2n}, \quad (3)$$

$$\Psi(h) = \left(\frac{e^{(h_i/T)}}{\sum_{j=1..2n} e^{(h_j/T)}}\right)_{i=1,\ldots,2n}, \quad (4)$$

where $T, \alpha \geq 0$ are parameters. If $A^k = D(u^k)AD(v^k)$ is our current best guess for solving the matrix scaling problem, then $\Psi(\rho(A^k)) = p \in \Delta_{2n}$. When $\rho(A^k)_i$ is large, the probability p_i of selecting the corresponding column or row of A^k will be large.

6. Greenhorn Algorithm

The Greenhorn (*Greedy Sinkhorn*) algorithm proposed by Altschuler, Weed, and Rigollet 2017 is a limiting case of the GSS algorithm when $\alpha \to \infty$ or $T = 0$ is used together with Ψ defined by in (3) or in (4), respectively. On the other extreme $\alpha = 0$ in (3) or $T \to \infty$ in (4) gives uniform distribution.

9. Numerics

We compare Sinkhorn, Greenkhorn and several other variates of Greedy Stochastic Sinkhorn.

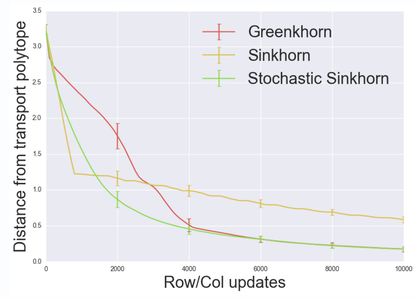

Figure 2: The GSS performs best in regimes of low penalization ($\lambda = 10$) on MNIST dataset. For the x-axis, one should read "number of row and column updates" in the sense that one iteration on the x-axis represents one update of a row or a column.

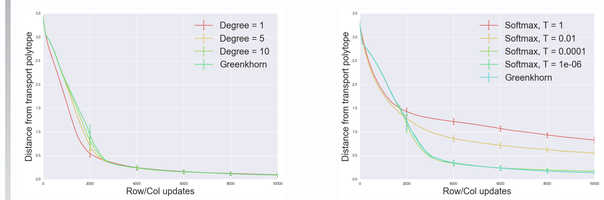

Figure 3: Greedy Stochastic Sinkhorn with different probability functions, and Greenkhorn as limiting case. Left: polynomial probabilities (3), Right: softmax probabilities (4).

References

Altschuler, Jason, Jonathan Weed, and Philippe Rigollet (2017). "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *CoRR* abs/1705.09634.

Cuturi, Marco (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems 26*, pp. 2292–2300.

Acknowledgements

Both authors are indebted to Marco Cuturi for teaching both of them about OT and many inspiring discussions.