Sinkhorn Algorithm as a Special Case of Stochastic Mirror Descent

Konstantin Mishchenko

KAUST

Problem 1: Matrix Scaling

Given a matrix $X^0 \in \mathbb{R}^{n\times n}_{++}$, find vectors $u, v \in \mathbb{R}^n_+$ such that

$$W \overset{\text{def}}{=} \text{diag}(u)X^0\text{diag}(v)$$

is doubly stochastic, i.e. $W1 = 1$ and $W^\top 1 = 1$.

Motivation

- Matrix preconditioning for improved linear algebra operations such as solving $X^0 w = b$.
- Ranking web page significance: take network connectivity matrix and find the stationary distribution of its doublly-stochastic form
- Estimation of transition probabilities in Markov chains; traffic and transportation planning; network optimization (see [2] for more details).

Sinkhorn Algorithm

Algorithm 1: Sinkhorn Algorithm.

Input : X^0;
for $k = 1, \dots$ **do**
  $X^{k+1}_{i:} = X^k_{i:}/\|X^k_{i:}\|_1$ for all i;
  $X^{k+2}_{:j} = X^{k+1}_{:j}/\|X^{k+1}_{:j}\|_1$ for all j;
end

Note that

$$\log X^{k+1} = \log X^k + \text{diag}(u_1^k, \dots, u_n^k)11^\top,$$
$$\log X^{k+2} = \log X^{k+1} + 11^\top \text{diag}(v_1^{k+1}, \dots, v_n^{k+1}).$$

This is very helpful for showing the equivalence.
Can be trivially generalized to finding W such that $W1 = p$, $W^\top 1 = q$ for any $p, q \in \mathbb{R}^d_+$.

[1] Marco Cuturi.
Sinkhorn distances: Lightspeed computation of optimal transport.
In *Advances in neural information processing systems*, 2013.

[2] Bahman Kalantari, Isabella Lari, Federica Ricca, and Bruno Simeone.
On the complexity of general matrix scaling and entropy minimization via the ras algorithm.
Mathematical Programming, 2008.

[3] Richard Sinkhorn.
Diagonal equivalence to matrices with prescribed row and column sums.
The American Mathematical Monthly, 1967.

Problem 2: Entropy Regularization

Introduce entropy penalty:

$$\min_{X\in\mathbb{R}^{n\times n}} \sum_{i,j=1}^n (C_{ij}X_{ij} + \gamma X_{ij}\log X_{ij})$$
$$\text{s.t.} \quad X1 = p, \; X^\top 1 = q.$$

Linear Programming

Given a matrix $C \in \mathbb{R}^{n\times n}_{++}$ and vectors $p, q \in \mathbb{R}^n_+$ such that $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$ solve

$$\min_{X\in\mathbb{R}^{n\times n}} \sum_{i,j=1}^n C_{ij}X_{ij}$$
$$\text{s.t.} \quad X1 = p, \; X^\top 1 = q, \; X \geqslant 0.$$

Motivation: discrete optimal transport.

Problem 3: Nonsmooth Minimization

Given matrices $A_1, \dots, A_m \in \mathbb{R}^{n_i\times d}_{++}$ and vectors $b_1, \dots, b_m \in \mathbb{R}^{n_i}_{++}$ solve

$$\min_{x\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m \mathcal{KL}(A_i x \| b_i).$$

Problem Properties

For $f_i(x) \overset{\text{def}}{=} \mathcal{KL}(A_i x\|b_i)$ the gradients are given by

$$\nabla f_i(x) = A_i^\top \log\frac{A_i x}{b_i},$$

where log and division are taken coordinate-wise. Note f_i is **nonsmooth**, but is **relatively** smooth w.r.t. ω, i.e. $D_{f_i}(x,y) \leqslant L_i D_\omega(x,y)$ with $L_i = \max_{1\leqslant j\leqslant n_i}\sum_{p=1}^d (A_i)_{jp}$.

Main result

Stochastic Mirror Descent $\overset{\text{new}}{=}$ Sinkhorn algorithm $\overset{\text{known}}{=}$ Method of Stochastic Bregman projections .

Bregman Projections

$$\sum_{i,j=1}^n (C_{ij}X_{ij} + \gamma X_{ij}\log X_{ij}) = \mathcal{KL}(X\|X^0) + \text{const},$$

where $X^0 \overset{\text{def}}{=} \exp(-C/\gamma)$.
Let $\omega\colon \mathbb{R}^d \to \mathbb{R}$ be a strictly convex function. The associated **Bregman divergence** is

$$D_\omega(x,y) \overset{\text{def}}{=} \omega(x) - \omega(y) - \langle\nabla\omega(y), x-y\rangle.$$

If $\omega(x) = \sum_{i=1}^d x_i(\log x_i - 1)$, then $D_\omega(x,y)$ is the Kullback-Leibler divergence,

$$\mathcal{KL}(x\|y) \overset{\text{def}}{=} \sum_{i=1}^n (x_i \log\frac{x_i}{y_i} - x_i + y_i).$$

Thus, we are interested in projecting onto the intersection of some sets $C_1, \dots, C_m$

$$\min_{x\in\bigcap_{i=1}^m C_i} D_\omega(x, x^0).$$

Algorithm 2: Stochastic projections.

Input : x^0;
for $k = 1, \dots$ **do**
  Sample $i \in \{1, \dots, m\}$;
  $x^{k+1} = \text{argmin}_{x\in C_i} D_\omega(x, x^k)$;
end

Algorithm 3: Stochastic Mirror Descent.

Input : x^0, $\{\gamma_k\}_k$;
for $k = 1, \dots$ **do**
  Sample $i \in \{1, \dots, m\}$;
  $\nabla\omega(x^{k+1}) = \nabla\omega(x^k) - \gamma_k \nabla f_i(x^k)$;
end

Intuition

Since $\omega(x) = \sum_{i=1}^d x_i(\log x_i - 1)$, $\nabla\omega(x) = \log(x)$. Then, **the iterates live in a certain range space**,

$$\log(x^{k+1}) \in \log(x^0) + \text{Range}(A^\top),$$

where $A = (A_1^\top, \dots, A_m^\top)^\top$.
To show equivalence with Problems 1-2, we set $x = \text{vec}(X)$, $d = n^2$, $A_1, A_2 \in \{0,1\}^d$, $A_1 x = X1$, $A_2 x = X^\top 1$, $b_1 = p$, $b_2 = q$.
New insight: there is no guarantee for convergence of stochastic mirror descent on that problem, because $\mathcal{KL}(\cdot\|b)$ is a nonsmooth function. Moreover, Problem 3 is not constrained, so there is **no strong convexity**. This is **a gap in theory of stochastic mirror descent**.