Lecture 10, Monday, 4/18/2022

Today: More about Sinkhorn's algorithm.

- Finish the proof of convergence of Sinkhorn's alg.
- Convergence rate via Hilbert's metric.

Hilbert's projective metric. (Or: Cayley-Hilbert metric)

Denote $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x > 0\}$. Let $x, y \in \mathbb{R}^n_+$. Define

$$d_H(x,y) = \log \max_{i,j} \frac{x_i/y_i}{x_j/y_j} = \log \max_{i,j} \frac{y_i/x_i}{y_j/x_j}$$

$$= \log \max_{1 \le i,j \le n} \frac{x_i y_j}{x_j y_i} \left(= \max_{1 \le i,j \le n} \log \frac{x_i y_j}{x_j y_i} \right).$$

One can verify:

(1) $d_H(x,y) \ge 0$ $\forall x, y \in \mathbb{R}^n_+$. $d_H(x,y) = 0 \Leftrightarrow \frac{x}{\|x\|} = \frac{y}{\|y\|}$

($\Leftrightarrow \exists \alpha > 0$ such that $x = \alpha y$)

(2) $d_H(x,y) = d_H(y,x)$ $\forall x, y \in \mathbb{R}^n_+$

(3) $d_H(x,y) \le d_H(x,z) + d_H(z,y)$ $\forall x, y, z \in \mathbb{R}^n_+$.

Prove (3). $\max_{i,j} \frac{x_i y_j}{x_j y_i} = \frac{x_{i_0} y_{j_0}}{x_{j_0} y_{i_0}} = \frac{x_{i_0} z_{j_0}}{x_{j_0} z_{i_0}} \cdot \frac{y_{j_0} z_{i_0}}{y_{i_0} z_{j_0}}$

$\le \max_{i,j} \frac{x_i z_j}{x_j z_i} \cdot \max_{i,j} \frac{y_i z_j}{y_j z_i}$ (for some i_0, j_0). QED

Clearly, $\forall x, y \in \mathbb{R}^n_+$ $\forall \lambda, \mu > 0$,

$$d_H(x,y) = d_H\left(\frac{x}{\|x\|}, \frac{y}{\|y\|}\right) = d_H(\lambda x, \mu y).$$

$$d_H(x,y) = d_H\left(\frac{x}{y}, \mathbb{1}_n\right).$$

Therefore, d_H is a metric on $\mathbb{R}^n_+/\sim$ where $x \sim y \Leftrightarrow \frac{x}{\|x\|} = \frac{y}{\|y\|}$.

It is called Hilbert's projective metric on $\mathbb{R}^n_+$.

Proposition $(\mathbb{R}^n_+/\sim, d_H)$ is a complete metric space.

Proof Only the completeness. Assume $x^{(k)} \in \mathbb{R}^n_+$ $(k=1,2\cdots)$ and $d_H(x^{(k)}, x^{(l)}) \to 0$ as $k, l \to \infty$. Without loss of generality, we may assume $\|x^{(k)}\|=1$ (l^2-norm) for all k. For any $\varepsilon>0$, $d_H(x^{(k)}, x^{(l)}) \to 0$ means $\exists N$ s.t. $k, l \geq N$ $\Longrightarrow d_H(x^{(k)}, x^{(l)}) \leq \varepsilon$, i.e., $1 \leq \max_{i,j} \dfrac{x_i^{(k)} x_j^{(l)}}{x_j^{(k)} x_i^{(l)}} \leq 1+\varepsilon$. $(*)$

Since $\|x^{(k)}\|=1$, $\exists$ a subseq $\{x^{(k')}\}$ of $\{x^{(k)}\}$ s.t. $\|x^{(k')}-x\| \to 0$ for some $x \in \mathbb{R}^n$ with $\|x\|=1$. We show that $x>0$. From $(*)$, with k' replacing k, we get $x_i^{(k')} x_j^{(l)} \leq x_j^{(k')} x_i^{(l)}(1+\varepsilon)$ $\forall i,j$, $\forall k', l \geq N$. Setting $l=N$ and sending $k' \to \infty$, we get $x_i x_j^{(N)} \leq x_j x_i^{(N)}(1+\varepsilon)$ $\forall i,j$. If $x_{j_0}=0$ for some j_0, then all $x_i=0$. This is in contradiction with $\|x\|=1$. Hence $x>0$. From $(*)$, replacing k by k', and letting $k' \to \infty$, we get $1 \leq \max_{i,j} \dfrac{x_i x_j^{(l)}}{x_j x_i^{(l)}} \leq 1+\varepsilon$ $\forall l \geq N$. i.e., $d_H(x, x^{(l)}) \leq \log(1+\varepsilon) \leq \varepsilon$ if $l \geq N$. Hence, $d_H(x, x^{(l)}) \to 0$ as $l \to \infty$. QED.

Denote for $x \in \mathbb{R}^n$ $\|x\|_{var} = \max_i x_i - \min_i x_i$. Then

$$d_H(x,y) = \|\log x - \log y\|_{var},$$

where $\log x = (\log x_i)$. This can be shown by definition:

$$\begin{aligned} d_H(x,y) &= \max_{i,j} \log \frac{x_i y_j}{x_j y_i} \\ &= \max_{i,j} \left[(\log x_i - \log y_i) - (\log x_j - \log y_j)\right] \\ &= \max_i (\log x_i - \log y_i) - \min_j (\log x_j - \log y_j) \\ &= \|\log x - \log y\|_{var}. \end{aligned}$$

QED

Projective diameter and contraction ratio of a positive matrix

Now, denote $\mathbb{R}_+^{m\times n} = \{A = [A_{ij}] \in \mathbb{R}^{m\times n} : \text{all } A_{ij} > 0\}$.

Definition Let $A \in \mathbb{R}_+^{m\times n}$.

(1) Denote
$$\theta(A) = \sup\{d_H(Ax, Ay) : x, y \in \mathbb{R}_+^n\},$$
and call it *the projective diameter* of A.

(2) Denote
$$\kappa(A) = \sup\left\{\frac{d_H(Ax, Ay)}{d_H(x, y)} : x, y \in \mathbb{R}_+^n, \frac{x}{\|x\|} \neq \frac{y}{\|y\|}\right\},$$
call it the *contraction ratio* of A with respect to Hilbert's metric.

Proposition:
$$\theta(A) = \max_{i,j,k,l} \log \frac{a_{ik} a_{jl}}{a_{il} a_{jk}} = \log \max_{i,j,k,l} \frac{a_{ik} a_{jl}}{a_{il} a_{jk}}.$$

Proof Let $M = \max_{i,j,k,l} \frac{a_{ik} a_{jl}}{a_{il} a_{jk}}$. Then $a_{ik} a_{jl} \leq M a_{il} a_{jk}$ for all i, j, k, l. Now $\forall x, y \in \mathbb{R}_+^n$. We have for any i, j
$$\frac{(Ax)_i (Ay)_j}{(Ax)_j (Ay)_i} = \frac{(\sum_\alpha a_{i\alpha} x_\alpha)(\sum_\beta a_{j\beta} y_\beta)}{(\sum_\alpha a_{j\alpha} x_\alpha)(\sum_\beta a_{i\beta} y_\beta)} = \frac{\sum_{\alpha,\beta} a_{i\alpha} a_{j\beta} x_\alpha y_\beta}{\sum_{\alpha,\beta} a_{j\alpha} a_{i\beta} x_\alpha y_\beta}$$
$$\leq \frac{M \sum_{\alpha,\beta} a_{i\beta} a_{j\alpha} x_\alpha y_\beta}{\sum_{\alpha,\beta} a_{j\alpha} a_{i\beta} x_\alpha y_\beta} = M.$$

Hence, $d_H(Ax, Ay) = \log \max_{i,j} \frac{(Ax)_i (Ay)_j}{(Ax)_j (Ax)_i} \leq \log M = \theta(A)$.
$$\sup\{d_H(Ax, Ay) : x \in \mathbb{R}_+^n, y \in \mathbb{R}_+^n\} \leq \theta(A).$$

Conversely, assume $M = \frac{a_{ik} a_{jl}}{a_{il} a_{jk}}$ for some i, j, k, l.

Let $\varepsilon \in (0,1)$ and define $x, y \in \mathbb{R}^n_+$ by $x_k = 1$ and $x_j = \varepsilon$ if $j \neq k$, $y_\ell = 1$ and $y_j = \varepsilon$ if $j = \ell$. Then

$$\frac{(Ax)_i (Ay)_j}{(Ax)_j (Ay)_i} = \frac{(\sum_\alpha a_{i\alpha} x_\alpha)(\sum_\beta a_{j\beta} y_\beta)}{(\sum_\alpha a_{j\alpha} x_\alpha)(\sum_\beta a_{i\beta} y_\beta)} = \frac{(a_{ik} + O(\varepsilon))(a_{j\ell} + O(\varepsilon))}{(a_{jk} + O(\varepsilon))(a_{i\ell} + O(\varepsilon))} := M_\varepsilon$$

So, $e^{\Theta(A)} \geq e^{d_H(Ax, Ay)} \geq \frac{(Ax)_i (Ay)_j}{(Ax)_j (Ay)_i} = M_\varepsilon \to M$. <u>QED</u>

<u>Proposition</u> (1) $\Theta(A) = \Theta(A^T)$ and $\kappa(A) = \kappa(A^T)$.

(2) If $A \in \mathbb{R}^{m\times n}_+$ and $B \in \mathbb{R}^{n \times \ell}_+$, then $\kappa(AB) \leq \kappa(A)\kappa(B)$.

(3) If $A, B \in \mathbb{R}^{m\times n}_+$ are diagonally equivalence, defined by $A = \mathrm{diag}(u)\, B\, \mathrm{diag}(v)$ for some $u \in \mathbb{R}^m_+$ and $v \in \mathbb{R}^n_+$, then $\kappa(A) = \kappa(B)$.

<u>Proof</u> (1) The fact $\Theta(A) = \Theta(A^T)$ follows from the formula for $\Theta(A)$ (cf. Proposition above). This together with Birkhoff-Hopf Thm implies $\kappa(A^T) = \kappa(A)$.

(2) $\forall x, y \in \mathbb{R}^\ell_+$: $\dfrac{d_H(ABx, ABy)}{d_H(x,y)} = \dfrac{d_H(ABx, ABy)}{d_H(Bx, By)} \cdot \dfrac{d_H(Bx, By)}{d_H(x,y)}$

$\leq \kappa(A)\kappa(B)$. Hence, $\kappa(AB) \leq \kappa(A)\kappa(B)$.

(3) $\forall x, y \in \mathbb{R}^n_+$. Define $\hat{x} = v \odot x$ (i.e., $\hat{x}_i = v_i x_i$), $\hat{y} = v \odot y$.

Clearly, $\dfrac{\hat{x}_i \hat{y}_j}{\hat{x}_j \hat{y}_i} = \dfrac{x_i y_j}{x_j y_i}$. Hence $d_H(x,y) = d_H(\hat{x}, \hat{y})$.

Since $A_{ij} = u_i B_{ij} v_j$ $\forall i,j$, we have

$$\frac{(Ax)_i (Ay)_j}{(Ax)_j (Ay)_i} = \frac{(\sum_\alpha A_{i\alpha} x_\alpha)(\sum_\beta A_{j\beta} y_\beta)}{(\sum_\alpha A_{j\alpha} x_\alpha)(\sum_\beta A_{i\beta} y_\beta)}$$

$$= \frac{u_i (\sum_\alpha B_{i\alpha} v_\alpha x_\alpha)\, u_j (\sum_\beta B_{j\beta} v_\beta y_\beta)}{u_j (\sum_\alpha B_{j\alpha} v_\alpha x_\alpha)\, u_i (\sum_\beta B_{i\beta} v_\beta y_\beta)}$$

$$= \frac{(\sum_\alpha B_{i\alpha}\hat{x}_\alpha)(\sum_\beta B_{j\beta}\hat{y}_\beta)}{(\sum_\alpha B_{j\alpha}\hat{x}_\alpha)(\sum_\beta B_{i\beta}\hat{y}_\beta)} = \frac{(B\hat{x})_i (B\hat{y})_j}{(B\hat{x})_j (B\hat{y})_i}.$$

Hence, $d_H(Ax, Ay) = d_H(B\hat{x}, B\hat{y})$, and thus

$$\frac{d_H(Ax, Ay)}{d_H(x, y)} = \frac{d_H(B\hat{x}, B\hat{y})}{d_H(\hat{x}, \hat{y})}.$$

Since $(x, y) \to (\hat{x}, \hat{y})$ is a bijection, $k(A) = k(B)$. QED

Birkhoff (1957) - Hopf (1963) Thm. Let $A \in \mathbb{R}_+^{m\times n}$ and denote $\eta(A) = e^{\theta(A)} = \max_{i,j,k,l} \frac{a_{ik}\, a_{jl}}{a_{il}\, a_{jk}}$ (>1). Then

$$\kappa(A) = \frac{\sqrt{\eta(A)} - 1}{\sqrt{\eta(A)} + 1} \in (0, 1).$$

Given $a \in \mathcal{P}_m \cap \mathbb{R}_+^m$, $b \in \mathcal{P}_n \cap \mathbb{R}_+^n$, and $K \in \mathbb{R}_+^{m\times n}$.
Recall Sinkhorn's algorithm: Select $v^{(0)} \in \mathbb{R}_+^n$.

$$u^{(k)} = \frac{a}{K v^{(k-1)}} \quad \text{and} \quad v^{(k)} = \frac{b}{K^T u^{(k)}}, \quad k = 1, 2, \cdots$$

Set for $k = 1, 2, \cdots$:

$$A^{(k)} = \mathrm{diag}(u^{(k)})\, K\, \mathrm{diag}(v^{(k-1)}), \quad B^{(k)} = \mathrm{diag}(u^{(k)})\, K\, \mathrm{diag}(v^{(k)}).$$

The original Sinkhorn's construction (or process) uses the row sum and column sum to alternatively normalizing the matrix to have the row sum $= a$ and col. sum $= b$:

$A^{(k)} \to B^{(k)}$: calculate $\mu_j^{(k)} = \frac{1}{b_j}$ col-j sum of $A^{(k)}$, set $B_{ij}^{(k)} = A_{ij}^{(k)} / \mu_j^{(k)}$.

$B^{(k)} \to A^{(k+1)}$: calculate $\lambda_i^{(k)} = \frac{1}{a_i}$ row-i sum of $B^{(k)}$, set $A_{ij}^{(k+1)} = B_{ij}^{(k)} / \lambda_i^{(k)}$.

We have verified that (see Lecture 9)

$$\lambda^{(k)} = \frac{u^{(k)}}{u^{(k+1)}} \quad \text{and} \quad \mu^{(k)} = \frac{v^{(k-1)}}{v^{(k)}}, \quad k = 1, 2, \dots$$

Note that

row-sum of $A^{(k)} = a$, col. sum of $B^{(k)} = b$.

$\lambda^{(k)} \odot a =$ row sum of $B^{(k)}$, $\mu^{(k)} \odot b =$ col. sum of $A^{(k)}$.

By Sinkhorn's thm (or rather its proof),

$\lambda^{(k)} \to \mathbb{1}_m$ and $\mu^{(k)} \to \mathbb{1}_n$.

$u_i^{(k)} v_j^{(k)} \to u_i v_j \quad \forall i,j$ for some $u \in \mathbb{R}_+^m$, $v \in \mathbb{R}_+^n$.

$P := \operatorname{diag}(u)\, K \operatorname{diag}(v) \in \mathcal{A}(a,b)$

$A^{(k)} \to P$ and $B^{(k)} \to P$.

The following theorem provides the convergence rate of Sinkhorn's algorithm.

<u>Theorem</u> (Franklin-Lorenz 1989) We have

(1) $d_H(u^{(k)}, u) \le [\kappa(K)]^{2k-1} d_H(v^{(0)}, v)$,

$d_H(v^{(k)}, v) \le [\kappa(K)]^{2k} d_H(v^{(0)}, v)$.

(2) $d_H(\lambda^{(k)}, \mathbb{1}_m) \le [\kappa(K)]^{2k-2} d_H(\lambda^{(1)}, \mathbb{1}_m)$,

$d_H(\mu^{(k)}, \mathbb{1}_n) \le [\kappa(K)]^{2k} d_H(v^{(0)}, v)$.

(3) $$d_H(u^{(k)}, u) \le \frac{[\kappa(K)]^{2k-2}}{1 - [\kappa(K)]^2} d_H(\lambda^{(1)}, \mathbb{1}_m),$$

$$d_H(v^{(k)}, v) \le \frac{[\kappa(K)]^{2k-2}}{1 - [\kappa(K)]^2} d_H(\mu^{(1)}, \mathbb{1}_n).$$

(4) $\|\log B^{(k)} - \log P\|_\infty \le d_H(u^{(k)}, u) \le \frac{[\kappa(K)]^{2k-2}}{1-[\kappa(K)]^2} d_H(\lambda^{(1)}, \mathbb{1}_m)$.

$\|\log A^{(k+1)} - \log P\|_\infty \le d_H(v^{(k)}, v) \le \frac{[\kappa(K)]^{2k-2}}{1-[\kappa(K)]^2} d_H(\mu^{(1)}, \mathbb{1}_n)$.

Proof (1) We have for any $k \ge 1$,

$$d_H(u^{(k)}, u) = d_H\left(\frac{a}{K^T v^{(k-1)}}, \frac{a}{Kv}\right) = d_H(K^T v^{(k-1)}, K^T v)$$

$$\le \kappa(K^T) d_H(v^{(k-1)}, v) = \kappa(K) d_H\left(\frac{b}{Ku^{(k-1)}}, \frac{b}{Ku}\right)$$

$$= \kappa(K) d_H(Ku^{(k-1)}, Ku) \le [\kappa(K)]^2 d_H(u^{(k-1)}, u) \quad (*)$$

$$\le \cdots \le [\kappa(K)]^{2(k-1)} d_H(u^{(1)}, u) = [\kappa(K)]^{2k-2} d_H\left(\frac{a}{K^T v^{(0)}}, \frac{a}{K^T v}\right)$$

$$= [\kappa(K)]^{2k-2} d_H(K^T v^{(0)}, K^T v) \le [\kappa(K)]^{2k-2} \kappa(K^T) d_H(v^{(0)}, v)$$

$$= [\kappa(K)]^{2k-1} d_H(v^{(0)}, v).$$

Similarly,

$$d_H(v^{(k)}, v) = d_H\left(\frac{b}{Ku^{(k)}}, \frac{b}{Ku}\right) = d_H(Ku^{(k)}, Ku)$$

$$\le \kappa(K) d_H(u^{(k)}, u) \le [\kappa(K)]^{2k} d_H(v^{(0)}, v).$$

(2) $$d_H(\lambda^{(k)}, \mathbb{1}_m) = d_H\left(\frac{u^{(k)}}{u^{(k+1)}}, \mathbb{1}_m\right) = d_H(u^{(k)}, u^{(k+1)})$$

$$\le [\kappa(K)]^2 d_H(u^{(k-1)}, u^{(k)}) \le \cdots \le [\kappa(K)]^{2k-2} d_H(u^{(1)}, u^{(2)})$$

$$= [\kappa(K)]^{2k-2} d_H\left(\frac{u^{(1)}}{u^{(2)}}, \mathbb{1}_m\right) = [\kappa(K)]^{2k-2} d_H(\lambda^{(1)}, \mathbb{1}_m).$$

Similarly,

$$d_H(\mu^{(k)}, \mathbb{1}_n) \le [\kappa(K)]^{2k-2} d_H(\mu^{(1)}, \mathbb{1}_n).$$

(3) By the triangle inequality, we have

$$d_H(u^{(k)}, u) \le d_H(u^{(k)}, u^{(k+1)}) + d_H(u^{(k+1)}, u)$$

$$\overset{(*)}{=} d_H\left(\frac{u^{(k)}}{u^{(k+1)}}, \mathbb{1}_m\right) + [\kappa(K)]^2 d_H(u^{(k)}, u)$$

$$= d_H(\lambda^{(k)}, \mathbb{1}_m) + [\kappa(K)]^2 d_H(u^{(k)}, u).$$

Hence,

$$d_H(u^{(k)}, u) \le \frac{1}{1-[\kappa(K)]^2} d_H(\lambda^{(k)}, \mathbb{1}_m)$$

$$\le \frac{[\kappa(K)]^{2k-2}}{1-[\kappa(K)]^2} d_H(\lambda^{(1)}, \mathbb{1}_m).$$

Similarly,

$$d_H(v^{(k)}, v) \le \frac{[\kappa(K)]^{2k-2}}{1-[\kappa(K)]^2} d_H(\mu^{(1)}, \mathbb{1}_n).$$

(4) Now, fix $k \ge 1$. let $x_i = \frac{u_i^{(k)}}{u_i}$, $x(m) = \min_i x_i$, and $y_j = \frac{v_j^{(k)}}{v_j}$. Then, $\frac{B_{ij}^{(k)}}{P_{ij}} = \frac{u_i^{(k)} K_{ij} v_j^{(k)}}{u_i K_{ij} v_j} = x_i y_j$.

Let $\varepsilon = d_H(u^{(k)}, u)$. Then,

$$\varepsilon = d_H(u^{(k)}, u) = d_H\left(\frac{u^{(k)}}{u}, \mathbb{1}_m\right) = d_H(x, \mathbb{1}_m) = \log \max_{i,j} \frac{x_i}{x_j}.$$

Hence $x(m) \le x_i \le x_j e^{\varepsilon}$. Let $x_j = x(m)$. We get $x(m) \le x_i \le x(m) e^{\varepsilon}$. Hence $\frac{1}{x(m)} \ge \frac{1}{x_i} \ge \frac{1}{x(m)} e^{-\varepsilon}$. $(*)$

Since $P_{ij} y_j = x_i^{-1} B_{ij}^{(k)}$, col. sum of $P = b$, and col. sum of $B^{(k)} = b$, we have

$$y_j b_j = y_j \sum_i P_{ij} = \sum_i x_i^{-1} B_{ij}^{(k)},$$

$$\frac{e^{-\varepsilon} b_j}{x(m)} = \frac{1}{x(m)} e^{-\varepsilon} \sum_i B_{ij}^{(k)} \le \sum_i x_i^{-1} B_{ij}^{(k)} \le \frac{1}{x(m)} \sum_i B_{ij}^{(k)} = \frac{1}{x(m)} b_j.$$

Hence $\frac{e^{-\varepsilon}}{x(m)} b_j \le y_j b_j \le \frac{1}{x(m)} b_j$, so, $\frac{e^{-\varepsilon}}{x(m)} \le y_j \le \frac{1}{x(m)}$.
Thus, $e^{-\varepsilon} \le x_i y_j \le e^{\varepsilon}$. But $x_i y_j = B_{ij}^{(k)}/\beta_{ij}$. $\forall i,j$.
Hence $-\varepsilon \le \log B_{ij}^{(k)} - \log \beta_{ij} \le \varepsilon, \quad \forall i,j$.
Thus, $\| \log B^{(k)} - \log \beta \| \le d_H(u^{(k)}, u)$.
This and (3) imply the first inequality in (4). The second one is similar. QED