行列スケーリングと非線形最適化

相馬　輔

行列スケーリングは，工学，経済学，統計学といったさまざまな分野に共通に現れる問題であり，Sinkhorn アルゴリズムと呼ばれる単純な反復解法で解くことができる．本稿では，非線形最適化の観点から，Sinkhorn の定理，Sinkhorn–Knopp の定理など，行列スケーリングの基礎を解説する．また，作用素スケーリングなどの近年の拡張についても簡単に紹介する．

キーワード：行列スケーリング，交互最適化，二部グラフ，完全マッチング

1. 行列スケーリングとは

A を各成分が非負な $m \times n$ 行列とする．また，$r \in \mathbf{R}^m$ と $c \in \mathbf{R}^n$ を $\sum_i r_i = \sum_j c_j$ を満たす非負ベクトルとする．**行列スケーリング (matrix scaling)** とは，A の各行・各列に正のスカラーを掛けて，行和を r，列和を c にせよという問題である．すなわち，正の対角成分をもつ対角行列 L, R で，

$$(LAR)\mathbf{1} = r, \quad (LAR)^\top \mathbf{1} = c$$

を満たすものを求めよという問題である．ここで $\mathbf{1}$ はすべての成分が 1 であるベクトルを表す．行列スケーリングの解が存在するとき，A は**スケーリング可能 (scalable)** であるという．

行列スケーリングの歴史は古く，工学，経済学，統計学といったさまざまな分野で現れ，独立に再発見が繰り返されてきた．歴史については Idel のサーベイ論文 [1] に詳しい．ここでは，行列スケーリングが現れる例を二つ紹介しよう．

例 1：エントロピー正則化つき輸送問題　m 地点の工場から，n 地点の店舗へ商品を輸送することを考えよう．工場 i では r_i 単位の商品を生産し，店舗 j では c_j 単位の商品を消費するとする．また，工場 i から店舗 j へ 1 単位の商品を輸送するのにかかるコストを $w_{ij} \in (0, +\infty]$ とする．ここで，$w_{ij} = +\infty$ は工場 i から店舗 j への輸送が不可能であることを表す．工場 i から店舗 j への輸送量を p_{ij} とすると，総輸送コストを最小にする問題は

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j} w_{ij} p_{ij} \\ \text{subject to} \quad & \sum_j p_{ij} = r_i \quad (i = 1, \dots, m) \\ & \sum_i p_{ij} = c_j \quad (j = 1, \dots, n) \\ & p_{ij} \geq 0 \quad (i = 1, \dots, m, j = 1, \dots, n) \end{aligned}$$

と書ける．これは **Hitchcock 型輸送問題**と呼ばれる典型的な線形計画問題で，効率よく解くことができるが，ここではさらに目的関数にエントロピー正則化項を付加した問題を考える．

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j} (w_{ij} p_{ij} + H(p_{ij})) \\ \text{subject to} \quad & \sum_j p_{ij} = r_i \quad (i = 1, \dots, m) \\ & \sum_i p_{ij} = c_j \quad (j = 1, \dots, n) \end{aligned}$$

ここで，$H : \mathbf{R} \to \mathbf{R} \cup \{+\infty\}$ は

$$H(p_{ij}) = \begin{cases} p_{ij} \log p_{ij} - p_{ij} & (p_{ij} > 0) \\ 0 & (p_{ij} = 0) \\ +\infty & (p_{ij} < 0) \end{cases}$$

なる凸関数である．この問題は非線形凸最適化問題になる．双対変数を x_i $(i = 1, \dots, m)$, y_j $(j = 1, \dots, n)$ とおくと，ラグランジアンは

$$\begin{aligned} L(P, x, y) = & \sum_{i,j} (w_{ij} p_{ij} + H(p_{ij})) \\ & - \sum_i x_i (r_i - \sum_j p_{ij}) - \sum_j y_j \sum_i (p_{ij} - c_j) \end{aligned}$$

であるから，1 次の最適性条件は

$$w_{ij} + \log p_{ij} - x_i - y_j = 0$$

そうま　たすく
統計数理研究所
〒 190–8562 東京都立川市緑町 10–3
soma@ism.ac.jp

となる．これより，(i,j) 成分が

$$p_{ij} = e^{-w_{ij}+x_i+y_j}$$

と書ける P で実行可能なもの，すなわち行和が r で列和が c になるものが求められれば，それは最適解である．非負行列 A を $a_{ij} = e^{-w_{ij}}$ で定め，A の i 行を e^{x_i} 倍し，j 列を e^{y_j} 倍した行列を考えると，(i,j) 成分は $e^{-w_{ij}+x_i+y_j}$ になるので，結局 A, r, c に関する行列スケーリングが解ければ，エントロピー正則化つき輸送問題が解けることがわかる．エントロピー正則化つき輸送問題は，Wilson [2] により 1969 年に導入されたが，2010 年代から機械学習の分野で Wasserstein 距離を近似的に高速に計算する手法として再注目されている [3].

例 2：産業連関表の推定　産業に関わる重要な統計として**産業連関表**というものがある．これはある地域（たとえば都道府県）のある期間における，各産業の財・サービスの生産額と，各産業間の移動量などをまとめたものである．ここでは，簡単のため，図 1 のように，行に m 個の供給側（売り手）の産業，列に n 個の需要側（買い手）の産業があり，(i,j) 成分は産業 i から産業 j へ財・サービスが移動した量を表す $m \times n$ の表を考える．

このような表は定期的な統計調査により作成されるわけだが，mn 通りの産業間の移動量を調査する必要があり，非常に手間がかかる．他方，各産業の総生産量（行和）と総消費量（列和）は $m+n$ 個しか要素がないため，比較的容易に求められる．そこで，前年度の表 A と，今年度の行和 r，列和 c を用いて，今年度の表 P を推定する問題を考える．さまざまな考え方がありうるが，ここでは **Kullback–Leibler ダイバージェンス**（KL ダイバージェンス）を最小にする，次のような問題を考える．

$$\begin{aligned} \text{minimize} \quad & D(P:A) \\ \text{subject to} \quad & P\mathbf{1} = r \\ & P^\top \mathbf{1} = c \end{aligned} \tag{1}$$

ここで，

$$D(P:A) = \sum_{i,j} (p_{ij} \log \frac{p_{ij}}{a_{ij}} - p_{ij} + a_{ij})$$

は KL ダイバージェンスである．ただし，$p_{ij} < 0$ であるか，$a_{ij} = 0$ かつ $p_{ij} > 0$ となる (i,j) が存在するときは，$D(P:A) = +\infty$ と定める．KL ダイバージェンスは P に関して凸関数であるから，(1) は凸最適化問題である．先程と同様に，双対変数を x_i $(i = 1, \ldots, m)$，y_j $(j = 1, \ldots, n)$ とおき，1 次の最適性条件から最適な p を求めて代入すると，ラグランジュ双対問題は

供給側＼需要側	産業 1	産業 2	…	産業 n	行和
産業 1	a_{11}	a_{12}	…	a_{1n}	r_1
産業 2	a_{21}	a_{22}	…	a_{2n}	r_2
⋮	⋮	⋮	⋱	⋮	⋮
産業 m	a_{m1}	a_{m2}	…	a_{mn}	r_m
列和	c_1	c_2	…	c_n	

図 1　産業連関表の模式図

$$f(x,y) = \sum_{i,j} a_{ij} e^{x_i+y_j} - \sum_i r_i x_i - \sum_j c_j y_j \tag{2}$$

を最小化する無制約最適化問題 $\inf_{x \in \mathbf{R}^n, y \in \mathbf{R}^m} f(x,y)$ となる[1]．また，f の勾配は

$$\frac{\partial f}{\partial x_i}(x,y) = \sum_j a_{ij} e^{x_i+y_j} - r_i \quad (i = 1, \ldots, m)$$
$$\frac{\partial f}{\partial y_j}(x,y) = \sum_i a_{ij} e^{x_i+y_j} - c_j \quad (j = 1, \ldots, n)$$

で与えられる．ここで，$x \in \mathbf{R}^m, y \in \mathbf{R}^n$ に対して，行と列をスケールした行列 $\tilde{A}$ を $\tilde{a}_{ij} = a_{ij} e^{x_i+y_j}$ で定める．すると，勾配ベクトルは

$$\nabla_x f(x,y) = \tilde{A}\mathbf{1} - r$$
$$\nabla_y f(x,y) = \tilde{A}^\top \mathbf{1} - c$$

となり，(x,y) が f の停留点であることと，$\tilde{A}$ が行和 r，列和 c をもつことは等価である．したがって，f の停留点を求めれば，行列スケーリングの解が求められる．f は凸関数であるから，f の停留点は，f を最小化すれば求めることができる．双対問題の最適解 x^*, y^* が求められれば，主問題の最適解は $p_{ij} = a_{ij} e^{x_i^*+y_j^*}$ で与えられる．これは，上で定めた行列 $\tilde{A}$ にほかならない．

2. Sinkhorn の定理

行和と列和がともに 1 である非負行列を**二重確率行列 (doubly stochastic matrix)** という．Sinkhorn [4] は**正行列**の二重確率行列へのスケーリング可能性に関する次の定理を証明した．

[1] 正確には，主問題が最小化問題であるから，$-f$ を最大化する最適化問題が双対問題になるのだが，後の都合で双対問題も最小化問題の形にしておく．

$$A^{(0)} = \begin{array}{c} \\ 1.2000 \\ 0.9089 \\ 0.7675 \\ 1.1235 \end{array} \begin{array}{c} \begin{array}{cccc} 1.0000 & 1.0000 & 1.0000 & 1.0000 \end{array} \\ \begin{bmatrix} 0.3062 & 0.4189 & 0.0214 & 0.4535 \\ 0.1533 & 0.1564 & 0.4889 & 0.1104 \\ 0.3142 & 0.0410 & 0.2224 & 0.1899 \\ 0.2263 & 0.3838 & 0.2672 & 0.2462 \end{bmatrix} \end{array} \quad A^{(1)} = \begin{array}{c} \\ 1.0000 \\ 1.0000 \\ 1.0000 \\ 1.0000 \end{array} \begin{array}{c} \begin{array}{cccc} 1.0346 & 0.9161 & 1.0834 & 0.9659 \end{array} \\ \begin{bmatrix} 0.2552 & 0.3490 & 0.0179 & 0.3779 \\ 0.1687 & 0.1720 & 0.5379 & 0.1214 \\ 0.4094 & 0.0534 & 0.2898 & 0.2475 \\ 0.2014 & 0.3416 & 0.2379 & 0.2191 \end{bmatrix} \end{array}$$

$$A^{(2)} = \begin{array}{c} \\ 1.0354 \\ 0.9730 \\ 0.9776 \\ 1.0140 \end{array} \begin{array}{c} \begin{array}{cccc} 1.0000 & 1.0000 & 1.0000 & 1.0000 \end{array} \\ \begin{bmatrix} 0.2466 & 0.3810 & 0.0165 & 0.3912 \\ 0.1630 & 0.1878 & 0.4965 & 0.1257 \\ 0.3957 & 0.0583 & 0.2675 & 0.2562 \\ 0.1947 & 0.3729 & 0.2195 & 0.2269 \end{bmatrix} \end{array} \quad A^{(3)} = \begin{array}{c} \\ 1.0000 \\ 1.0000 \\ 1.0000 \\ 1.0000 \end{array} \begin{array}{c} \begin{array}{cccc} 1.0025 & 0.9884 & 1.0163 & 0.9928 \end{array} \\ \begin{bmatrix} 0.2382 & 0.3680 & 0.0159 & 0.3779 \\ 0.1675 & 0.1930 & 0.5103 & 0.1292 \\ 0.4047 & 0.0596 & 0.2736 & 0.2620 \\ 0.1920 & 0.3678 & 0.2165 & 0.2237 \end{bmatrix} \end{array}$$

$$A^{(25)} = \begin{array}{c} \\ 1.0000 \\ 1.0000 \\ 1.0000 \\ 1.0000 \end{array} \begin{array}{c} \begin{array}{cccc} 1.0000 & 1.0000 & 1.0000 & 1.0000 \end{array} \\ \begin{bmatrix} 0.2358 & 0.3703 & 0.0155 & 0.3784 \\ 0.1682 & 0.1970 & 0.5036 & 0.1312 \\ 0.4050 & 0.0607 & 0.2691 & 0.2652 \\ 0.1910 & 0.3720 & 0.2118 & 0.2252 \end{bmatrix} \end{array}$$

図 2　$r = c = \mathbf{1}$ に対する Sinkhorn アルゴリズムの実行例
25 反復の後に二重確率行列に収束し，停止した．

定理 1 (Sinkhorn [4])**.**　A を $n \times n$ 正行列（すなわち，すべての i, j に対し $a_{ij} > 0$）とする．このとき，ある正対角行列 L, R が存在して，LAR を二重確率行列にできる．

この定理を非線形最適化の道具を使って証明してみよう．Slater 条件と強双対定理を思い出すと，主問題が（相対的）内点実行可能解をもてば，ラグランジュ双対問題に最適解が存在し，その最適値は主問題の最適値に一致するのであった．

ここでは，主問題として $m = n$, $r = c = \mathbf{1}$ としたときの KL ダイバージェンス最小化問題 (1) を考える．$p_{ij} = 1/n$ とすれば，これは明らかに内点実行可能解であるから，主問題は Slater 条件を満たす．よって，双対変数 x_i^*, y_j^* が存在し，$p_{ij} = a_{ij} e^{x_i^* + y_j^*}$ は主問題の最適解になる．L, R をそれぞれ $e^{x_i^*}, e^{y_j^*}$ を並べた対角行列とすると，LAR は二重確率行列になる．

3. Sinkhorn アルゴリズム

行列スケーリングに対する最も単純なアルゴリズムとして，**Sinkhorn アルゴリズム**がある．これは，以下のような反復解法である．まず，A の各列を正規化し，A の列和は c であるようにしておく．$A^{(0)} = A$ とおく．各 $t = 0, 1, 2, \ldots$ に対して，t が偶数の場合は $A^{(t)}$ を行正規化して行和を r にし，t が奇数の場合は列正規化して列和を c にする．すなわち，各 i, j に対して，t が偶数のときは

$$a_{ij}^{(t+1)} = \frac{a_{ij}^{(t)} r_i}{(A^{(t)}\mathbf{1})_i}$$

t が奇数のときは

$$a_{ij}^{(t+1)} = \frac{a_{ij}^{(t)} c_j}{(A^{(t)\top}\mathbf{1})_j}$$

という更新式で行列 $A^{(t)}$ を定める．図 2 に実行例を示した．

このアルゴリズムは Kruithof [5] が 1937 年に輸送問題の文脈で提案したものが最初とされるが，その後もさまざまな分野で独立に再発見されたため，**RAS 法**や **iterative proportional fitting procedure** と呼ばれることもある．

交互最適化としての解釈　Sinkhorn アルゴリズムは，双対問題 (2) に対する**交互最適化法**として解釈できる．$(x^{(0)}, y^{(0)}) = (\mathbf{0}, \mathbf{0})$ からはじめて，t が偶数のときは y を固定したまま x について最小化し，t が奇数のときは x を固定したまま y について最小化する．式で書くと，$t = 0, 1, 2, \ldots$ に対して，t が偶数のときは

$$x^{(t+1)} = \operatorname*{argmin}_{x \in \mathbf{R}^m} f(x, y^{(t)})$$
$$y^{(t+1)} = y^{(t)}$$

とし，t が奇数のときは

$$x^{(t+1)} = x^{(t)}$$
$$y^{(t+1)} = \operatorname*{argmin}_{y \in \mathbf{R}^n} f(x^{(t)}, y)$$

と $(x^{(t)}, y^{(t)})$ を更新する．各ステップでの最適化問題は，勾配の式から，結局 $A^{(t)}$ の行和・列和を正規化する問題と等価である．

交互射影としての解釈　さらに強い結果として，Csiszár [6] は，Sinkhorn アルゴリズムは KL ダイバージェンスに関する**交互射影**であることを示した．すなわち，t が偶数のときは

$$A^{(t+1)} = \operatorname*{argmin}_{A \geq O : A\mathbf{1} = r} D(A : A^{(t)})$$

t が奇数のときは

$$A^{(t+1)} = \operatorname*{argmin}_{A \geq O : A^\top \mathbf{1} = c} D(A : A^{(t)})$$

が成り立つ．この性質を利用して，Sinkhorn アルゴリズムの収束解析ができる．A^* を KL ダイバージェンス最小化問題 (1) の最適解とする．一般に，次の**一般化ピタゴラスの定理**が成り立つ．

$$D(A^{(t)} : A^*) = D(A^{(t)} : A^{(t+1)}) + D(A^{(t+1)} : A^*)$$

これより，

$$\sum_{t=0}^{T-1} D(A^{(t)} : A^{(t+1)}) = D(A : A^*) - D(A^{(T)} : A^*)$$

また，**Pinsker の不等式** [7] などから，以下の不等式が示せる．詳細は割愛する．

$$D(A^{(t)} : A^{(t-1)}) \geq \frac{1}{2} \max\{\|A^{(t)}\mathbf{1} - r\|_1^2, \|A^{(t)\top}\mathbf{1} - c\|_1^2\}$$

任意の $\varepsilon > 0$ を取る．いま，$t = 1, \ldots, T$ に対して，

$$\max\{\|A^{(t)}\mathbf{1} - r\|_1, \|A^{(t)\top}\mathbf{1} - c\|_1\} \geq \varepsilon$$

と仮定すると，上式を組み合わせて

$$\frac{T\varepsilon^2}{2} \leq D(A : A^*)$$

となる．したがって，$T > 2D(A : A^*)/\varepsilon^2$ であれば，上の式は成り立たないので，背理法より，ある $t \in \{1, \ldots, T\}$ で

$$\max\{\|A^{(t)}\mathbf{1} - r\|_1, \|A^{(t)\top}\mathbf{1} - c\|_1\} < \varepsilon$$

が成り立つ．すなわち，$A^{(t)}$ は誤差 ε 以内で行和 r，列和 c をもつ．

4. Sinkhorn–Knopp の定理

Sinkhorn の定理は $n \times n$ 正行列 A と $r = c = \mathbf{1}$ に対して，行列スケーリングの解が必ず存在することを示した．ここでは，Sinkhorn の定理を非負行列へ拡張した **Sinkhorn–Knopp の定理**を紹介しよう．

非負行列においては，行列スケーリングの解が必ずしも存在するとは限らない．たとえば，

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \tag{3}$$

に対しては行列スケーリングの解は存在しない．実際，A を二重確率行列にしようとすると，(1,2) 成分と (2,1) 成分の 1 を残したまま，(1,1) 成分を 0 にしなければならない．しかし，行や列をスケーリングして非零成分を 0 にすることはできないので，これは不可能である．一方，任意の $\varepsilon > 0$ に対して，

$$\begin{bmatrix} \varepsilon & \\ & 1 \end{bmatrix} A \begin{bmatrix} 1 & \\ & 1/\varepsilon \end{bmatrix} = \begin{bmatrix} \varepsilon & 1 \\ 1 & 0 \end{bmatrix}$$

となるから，$\varepsilon \to 0$ とすればいくらでも二重確率行列に近づけられる．ここで，スケーリング行列には $1/\varepsilon$ が含まれるため，$\varepsilon = 0$ とはできないことに注意する．非線形最適化の観点からは，A が非負行列の場合，KL ダイバージェンス最小化問題 (1) の Slater 条件が成り立たないことがあることに対応する．この場合，たとえ双対問題の目的関数値が有界であっても，最小値を達成する解が存在するとは限らない．

非負行列 A が**近似スケーリング可能 (approximately scalable)** であるとは，任意の $\varepsilon > 0$ に対して，ある正対角行列 L, R が存在して，

$$\|(LAR)\mathbf{1} - \mathbf{1}\| < \varepsilon, \quad \|(LAR)^\top \mathbf{1} - \mathbf{1}\| < \varepsilon$$

とできることをいう．すなわち，A をスケーリングによりいくらでも二重確率行列に近づけられることをいう．式 (3) の行列 A はスケーリング不可能だが近似スケーリング可能な例である．

次に，非負行列 A の**台グラフ (support graph)** とは，頂点集合 $V^+ = V^- = \{1, \ldots, n\}$ と，枝集合 $E = \{ij : a_{ij} \neq 0\}$ をもつ二部グラフ $G = (V^+, V^-; E)$ である．また，二部グラフにおける**完全マッチング (perfect matching)** とは，枝部分集合 $M \subseteq E$ で，どの頂点にも 1 本だけ M の枝が接続しているものである．図 3 に例がある．

Sinkhorn–Knopp の定理は，双対問題 (2) の有界性と，近似スケーリング可能性と，台グラフにおける完全マッチングの存在とを特徴づける定理である．

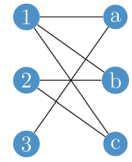
定理 2 (Sinkhorn–Knopp [8])**.** A を $n \times n$ 非負行列とする．次の条件は等価．

1. $\inf_{x,y \in \mathbf{R}^n} f(x, y) > -\infty$
2. A は近似スケーリング可能である
3. A の台グラフに完全マッチングが存在する

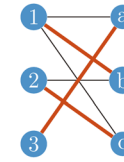
この定理を非線形最適化を用いて証明してみよう．まず，$1 \implies 2$ であるが，これは Sinkhorn アルゴリ

$$A = \begin{bmatrix} .3 & .2 & .5 \\ 0 & .8 & .5 \\ .7 & 0 & 0 \end{bmatrix}$$

(a) 非負行列



(b) 台グラフ



(c) 完全マッチング

図 3　非負行列，台グラフ，完全マッチングの例

ズムの解析を行うことで示される．いま，$x, y \in \mathbf{R}^n$ があるとき，y を固定して x について最小化することを考える．最小化した解を x^+ とおく．このとき，関数値の改善量に関して

$$f(x,y) - f(x^+,y) \geq D(A^{(t)}\mathbf{1} : \mathbf{1})$$

が成り立つ．再び Pinsker の不等式より，

$$D(A^{(t)}\mathbf{1} : \mathbf{1}) \geq \frac{1}{2}\|A^{(t)}\mathbf{1} - \mathbf{1}\|_1^2$$

である[2]．よって，任意の $\varepsilon > 0$ に対し，$\|A^{(t)}\mathbf{1} - \mathbf{1}\|_1 \geq \varepsilon$ である限り，関数値は $O(\varepsilon^2)$ だけ毎回改善する．ところが，仮定より目的関数値は有界なので，有限の t で $\|A^{(t)}\mathbf{1} - \mathbf{1}\|_1 < \varepsilon$ となる．列和に対しても同様．よって，A は近似スケーリング可能である．

次に 2 $\Longrightarrow$ 3 を示す．二部グラフの完全マッチングの存在条件である **Hall の条件**が成り立つことを示せばよい．Hall の条件とは，任意の部分集合 $X \subseteq V^+$ に対して，$|X| \leq |\Gamma(X)|$ が成り立つことである．ここで，$\Gamma(X) \subseteq V^-$ は X に接続する頂点の集合である．いま，仮定より，任意の $\varepsilon > 0$ に対して，正対角行列 L, R が存在して，$\tilde{A} = LAR$ は行和，列和ともに $1 \pm \varepsilon$ 以内にできる．正対角行列をかけても非零成分は変化しないので，$\tilde{A}$ と A の台グラフは等しいことに注意する．すると，任意の $X \subseteq V^+$ に対して，

$$\begin{aligned} \frac{1}{1-\varepsilon}|X| &\leq \sum_{i \in X} \sum_{j \in \Gamma(X)} \tilde{a}_{ij} \\ &= \sum_{j \in \Gamma(X)} \sum_{i \in X} \tilde{a}_{ij} \\ &\leq \sum_{j \in \Gamma(X)} \sum_{i \in V^-} \tilde{a}_{ij} \\ &\leq (1+\varepsilon)|\Gamma(X)| \end{aligned}$$

である．$\varepsilon \to +0$ とすれば，Hall の条件が得られる．

最後に 3 $\Longrightarrow$ 1 を示す．仮定より，A の台グラフに完全マッチング M が存在する．そこで，

$$p_{ij} = \begin{cases} 1 & ij \in M \\ 0 & ij \notin M \end{cases}$$

とすれば，P は主問題の実行可能解で，その目的関数値は有限である．したがって，弱双対性から双対問題の目的関数は有界である．

以上で，Sinkhorn–Knopp の定理が示された．なお，凸解析における**後退関数 (recession function)** を用いた直接的な証明もある．文献 [9] を参照．

5. 発展的な話題

最後に，発展的話題についていくつか紹介する．

5.1 Hilbert 距離を用いた Sinkhorn アルゴリズムの収束解析

Sinkhorn アルゴリズムは正行列に対して **1 次収束**することが知られている．解析においては，**Hilbert 距離**と呼ばれる特殊な距離が用いられる．正ベクトル $u, v \in \mathbf{R}^n_{++}$ の Hilbert 距離[3]は

$$d_H(u,v) = \log \max_{i,j} \frac{u_i v_j}{v_i u_j}$$

で与えられる．また，成分ごとの対数を取って

$$d_H(u,v) = \|\log u - \log v\|_{\mathrm{var}}$$

とも書ける．ここで，

$$\|f\|_{\mathrm{var}} = (\max_i f_i) - (\min_i f_i)$$

である．Franklin and Lorenz [10] は，Sinkhorn アルゴリズムの各反復が Hilbert 距離に関する縮小写像であることを示し，その帰結として 1 次収束性を示した．具体的には，双対問題 (2) の最適解を (x^*, y^*) とするとき，ある定数 $c \in (0,1)$ に対して，

[2] ここで，m 次元ベクトル同士の KL ダイバージェンスは，ベクトルを $m \times 1$ 行列とみなして定める．

[3] 正確には，正象限の錐 $\mathbf{R}^n_{++}$ に正のスカラー倍による同値関係 $\sim$ を入れ，商集合を取った集合 $\mathbf{R}^n_{++}/\sim$ 上の距離となる．

$$\|x_t - x^*\|_{\mathrm{var}} \le c^t \|x_0 - x^*\|_{\mathrm{var}}$$
$$\|y_t - y^*\|_{\mathrm{var}} \le c^t \|y_0 - y^*\|_{\mathrm{var}}$$

となる．

5.2 作用素スケーリング

行列スケーリングの拡張として，Gurvits [11] は2004年に**作用素スケーリング (operator scaling)**と呼ばれる問題を導入した．$A_1, \ldots, A_k$ を $m \times n$ 行列とする．このとき，$m \times m$ 正則行列 L と $n \times n$ 正則行列 R で，

$$\sum_{i=1}^{k} (LA_iR)(LA_iR)^\top = \frac{1}{m} I_m$$
$$\sum_{i=1}^{k} (LA_iR)^\top (LA_iR) = \frac{1}{n} I_n$$

を満たすものを求めよという問題である．

Gurvits [11] は，行列スケーリングに対するSinkhornアルゴリズムの拡張である**作用素 Sinkhorn アルゴリズム**を提案し，Sinkhorn–Knoppの定理の作用素スケーリング版も示している．その後，より詳細な収束解析がGarg et al. [12] によってなされたが，その中では**非可換ランク (noncommutative rank)**などの高度な代数的道具立てが用いられている．作用素スケーリングは，Edmonds問題やBrascamp–Lieb不等式，TylerのM推定量などに応用が見つかりつつある．また，Franks [13] は単位行列ではなく，一般の正定値行列をマージナルにもつ作用素スケーリングを提案し，作用素Sinkhornアルゴリズムのさらなる拡張を提案している．Matsuda and Soma [14] では，行列スケーリングに対するCsiszarの結果を拡張し，作用素Sinkhornアルゴリズムに対する量子情報幾何的特徴づけを与えている．

5.3 テンソルスケーリング，非可換最適化理論

Bürgisser et al. [15] は，作用素スケーリングをさらに一般化した問題として**テンソルスケーリング (tensor scaling)**を提案した．この問題に対しても，Sinkhorn型のアルゴリズムが提案され，収束解析が行われている．さらに，これらのスケーリング系の問題群に対する統一的な枠組みとして**非可換最適化理論**がBürgisser et al. [16] により提案されている．

本稿で見てきたとおり，行列スケーリングにおいては通常のユークリッド空間上の凸最適化が有効であったが，これらの一般化されたスケーリング問題においては，**非正曲率空間上の凸最適化**が本質的な役割を果たすことがHirai [9] により指摘されており，非線形最適化における新たなフロンティアを切り拓きつつある．これらの一般化された空間上の最適化問題に対して，効率的なアルゴリズムが存在するのか否かが，最先端の研究トピックとなっている．

謝辞　本稿執筆の機会をいただきました奥野貴之先生に感謝いたします．

参考文献

[1] M. Idel, "A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps," *arXiv*, arxiv:1609.06349, 2016.
[2] A. G. Wilson, "The use of entropy maximising models, in the theory of trip distribution, mode split and route split," *Journal of Transport Economics and Policy*, **3**, pp. 108–126, 1969.
[3] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends® in Machine Learning*, **11**, pp. 355–607, 2019.
[4] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *The Annals of Mathematical Statistics*, **35**, pp. 876–879, 1964.
[5] J. Kruithof, "Telefoonverkeersrekening," *De Ingenieur*, **52**, pp.15–25, 1937.
[6] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, **3**, pp. 146–158, 1975.
[7] R. W. Yeung, *Information Theory and Network Coding*, Springer Science & Business Media, 2008.
[8] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, **21**, pp. 343–348, 1967.
[9] H. Hirai, "Convex analysis on Hadamard spaces and scaling problems," *Foundations of Computational Mathematics*, https://doi.org/10.1007/s10208-023-09628-5, 2023
[10] J. Franklin and J. Lorenz, "On the scaling of multidimensional matrices," *Linear Algebra and its Applications*, **114-115**, pp. 717–735, 1989.
[11] L. Gurvits, "Classical complexity and quantum entanglement," *Journal of Computer and System Sciences*, **69**, pp. 448–484, 2004.
[12] A. Garg, L. Gurvits, R. Oliveira and A. Wigderson, "Operator scaling: Theory and applications," *Foundations of Computational Mathematics*, **20**, pp. 223–290, 2020.
[13] C. Franks, "Operator scaling with specified marginals," In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 190–203, 2018.
[14] T. Matsuda and T. Soma, "Information geometry of operator scaling," *Linear Algebra and its Applications*, **649**, pp. 240–267, 2022.
[15] P. Bürgisser, A. Garg, R. Oliveira, M. Walter and A. Wigderson, "Alternating minimization, scaling algorithms, and the null-cone problem from invariant theory," In *Proceedings of 9th Innovations in Theoretical Computer Science Conference (ITCS)*, **94**, pp. 24：1–24：20, 2018.

[16] P. Bürgisser, C. Franks, A. Garg, R. Oliveira, M. Walter and A. Wigderson, "Towards a theory of non-commutative optimization: Geodesic 1st and 2nd order methods for moment maps and polytopes," In *Proceedings of the 60th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 845–861, 2019.