ReLU DNNの二値化によるモデル圧縮

Model Compression of ReLU DNN by binarization

長瀬 准平 (NAGASE Jumpei)[1], 石渡 哲哉 (ISHIWATA Tetsuya)[2]
[1] 電気通信大学 (University of electro-communications)
[2] 芝浦工業大学 (Shibaura Institute of Technology)
e-mail : jnagase@uec.ac.jp

1 深層ニューラルネットワークモデルについて

深層学習は近年注目されている機械学習や人工知能技術の手法の一種であり，**深層ニューラルネットワークモデル**（以下，**DNN**）と呼ばれる大量の学習パラメータからなる大規模モデルを同定（学習）するという特徴をもつ．DNN は，多次元の入力ベクトル x に対し，**層**と呼ばれる線型変換（厳密にはアフィン変換）と非線型変換を繰り返すことにより構成される．線型変換を行う層は係数行列を学習パラメータにもち，期待される構造に応じて呼び名と制約が異なる．本発表では特に制約のない**全結合層**を対象とする．全結合層はその名の通り，入力と出力が全て結合するように線型変換を行う層であり，特に制約のない任意の行列をパラメータとする．非線型変換を行う層は，一般に**活性化層**と呼ばれることが多く，モデル自体の出力の構造に基づく制約や重み付けのための非線型変換などの特殊なものを除き，基本的に要素ごとの非線型変換を行うことが一般的である．近年では様々な非線型変換が活性化関数として提案されているが，**ReLU 関数**（Rectified Linear Unit function）と呼ばれる区分線型関数；$\mathrm{ReLU(x)} := \max(x, 0)$ をベースとして，その派生の関数が広く用いられている．全結合層と活性化層を繰り返すことで構成され，内部に循環をもたない順伝播型の DNN は**多層パーセプトロン**（Multi-Layer Perceptron; MLP）と呼ばれることが多い．本発表では活性化関数に ReLU 関数を仮定した多層パーセプトロン；ReLU MLP を対象とした結果を主に紹介する．

2 ReLU MLP の表現能力に関するこれまでの成果

DNN はパラメータを学習させることで様々な関数を近似可能である．実際，十分に多くのパラメータをもつ DNN は任意の関数を任意の精度で近似できることがよく知られており，その近似の収束の性質によって各 DNN の特徴づけができる．一方で，有限の学習パラメータを用いて実装されている現実のモデル同士の対応や特徴づけを明らかにする研究は多くない．そこで，発表者らはモデルの表現能力を次のように定義し，その包含関係によってモデルの比較を行った．

定義 1 (表現能力)**.** ある学習モデル M が学習パラメータ $\theta \in \Theta$ によって関数 f を定めるものとする．学習パラメータが $\tilde{\Theta} \subset \Theta$ の範囲内を動くとしたときに学習モデル M が取りうる関数の集合；

$$\mathcal{R}(M, \tilde{\Theta}) := \left\{ M(\theta) | \theta \in \tilde{\Theta} \right\}$$

を学習モデル M の $\tilde{\Theta}$ における表現能力と呼ぶ．また，ある関数 g について $g \in \mathcal{R}(M, \tilde{\Theta})$ が成り立つとき，g は M によって $\tilde{\Theta}$ において設計可能であるという．また，$\tilde{\Theta} = \Theta$ であるとき，$\mathcal{R}(M, \Theta)$ を単に学習モデル M の表現能力といい，g は M によって設計可能であるという．

ReLU MLP の表現能力について，発表者らの次の結果が知られている．

定理 2 (JJIAM 2022)**.** ReLU 関数とアフィン関数の連結，加算，合成によって設計される任意の関

数は *ReLU MLP* によって設計可能である．

定理 3 (情報処理学会論文誌 2023)**.** $\mathbb{R}$ 上で定義される連続な区分線型関数のうち，異なる 2 つの傾きをもつ区間がある関数の集合を G とする．ある連続区分線型関数 $g \in G$ を活性化関数とした多層パーセプトロンは *ReLU MLP* で設計可能である．また，逆も成り立つ．

以上の結果から，ReLU MLP および区分線型 MLP は，区分線型関数とアフィン関数の連結と合成と加算によって構成される任意の関数および DNN を設計可能であり，その意味で，ReLU MLP を改めて ReLU DNN と呼ぶこととする．

3 ReLU DNN の二値化

前節で紹介した結果はパラメータに依らず，広く一般的な DNN の表現能力に関する対応関係を示すものである．そこで，アフィン関数のパラメータに制約を入れた DNN の表現能力について議論し，発表者らは次の結果を得た．

定理 4. $\mathrm{ReLU}_{[N]}$ は N 次元に ReLU 関数を適用する関数，$\mathbb{B} = \{-1, 0, 1\}$ とする．任意のパラメータ $A_0 \in \mathbb{R}^{m\times n}$, $A_1 \in \mathbb{R}^{l\times m}$, $A_2 \in \mathbb{R}^{k\times l}$ と任意の入力 $\boldsymbol{x}$ について，

$$A_2\mathrm{ReLU}_{[l]}\left(A_1\mathrm{ReLU}_{[m]}\left(A_0\boldsymbol{x}\right)\right) = B_2\mathrm{ReLU}_{[kl]}\left(B_1\mathrm{ReLU}_{[klm]}\left(A^*\boldsymbol{x}\right)\right)$$

をみたす $A^* \in \mathbb{R}^{klm\times n}$, $B_1 \in \mathbb{B}^{kl\times klm}$, $B_2 \in \mathbb{B}^{k\times kl}$ が存在する．

すなわち，十分なパラメータ数を持つ二値パラメータの ReLU 多層パーセプトロンは入力に適切な線型変換を施すことで実数パラメータの ReLU MLP を設計可能であり，逆に，任意の ReLU MLP は二値のパラメータをもつ多層パーセプトロンに帰着できる．さらに，この証明は構成的であることから，パラメータを具体的に書き下すことができ，一見すると大規模なパラメータ数に見える B_1, B_2 についても，A_1, A_2 と同数のパラメータ以外はすべて零であることがわかる．したがって，入力層のみ kl 倍のパラメータ数となるものの，それ以外のパラメータは $lm + kl$ 個のパラメータを二値に圧縮できる．また，この結果は任意の層数の場合に一般化でき，各層の次元数を d_i，層数を L としたとき，入力層を $\prod_{i=2}^{L} d_i$ 倍することで，$\sum_{i=2}^{L} d_i d_{i-1}$ 個のパラメータが二値に圧縮できることがわかる．一般の計算効率を考えるとこの変形は推奨されるものではないが，入力層の変換と実際の推論を別の計算機で実施することなどにより，圧縮の有用性が考えられる．ここで，活性化関数である ReLU 関数の出力が非負であることに着目すると，非負の値に -1 と $+1$ を掛けて足し合わせた結果をまた ReLU 関数に入力するという操作が繰り返されることになる．もし，-1 と $+1$ を掛けて足し合わせた結果が負となる場合，その結果は ReLU 関数を通る際に 0 となるため，実質的にそれ以降のその変数を用いたパラメータは不要となり，圧縮可能であることがわかる．このような ReLU 関数と二値化を組み合わせた圧縮可能性の議論として，本発表では次の二つの性質を紹介する．

- 非負の入力に -1, $+1$ を掛けて足し合わせた結果がどのような条件で正となるのか．

  → 例えば，-1, $+1$ の順序と個数および，入力の順序について．

- 入力に順序付き制約のある MLP と等価（互いに設計可能）な MLP が存在するのか．

また，この性質によって，実質的に圧縮されるパラメータの数についても議論する．

C^* 環を用いたカーネル法の拡張による関数データの解析

Kernel methods with C^*-algebra for analyzing functional data

橋本 悠香 (Yuka Hashimoto)[1]
[1]NTT 株式会社 (NTT, Inc.)
e-mail : yuka.hashimoto@ntt.com

1 概要

カーネル法は，データの非線形性を効果的に扱うことができるという性質から，さかんに研究が行われてきた．本講演では，C^* 環を用いてカーネル法を拡張し，複雑なデータから特徴抽出を行う枠組みについて述べる．特に関数データに対して本枠組みを適用し，その局所的な特徴と大域的な特徴を同時に抽出できることを示す．

2 はじめに

C^* 環は複素数の空間を一般化した概念であり，関数，行列，作用素等を統一的に扱うことができる．一方，カーネル法は機械学習における最も基本的なツールの一つであり [1]，データの非線形性を効果的に扱うことができるという性質から，さかんに研究が行われてきた．カーネル法においては，カーネルと呼ばれる関数を決定し，それを用いて再生核 Hilbert 空間と呼ばれる関数空間を構成する．C^* 環を用いてカーネル法を拡張することで，関数，行列，作用素等を用いて表されるデータの解析を効率的に行うことが可能となる [2]．本稿では，その中でも特に関数データに焦点を当てる．関数データを出力とするモデルをカーネル法で構築する場合，関数値のカーネルが必要となるが，どのような関数値カーネルを選択すべきかは，非自明である．

本研究では，spectral truncation という概念に基づき，関数を入出力とするモデルを構築するためのカーネルを提案する．提案するカーネルは，既存の典型的な 2 種類のカーネルのギャップを埋めるものであり，関数の局所的な特徴と大域的な特徴の抽出を，truncation パラメータと呼ばれる値により制御することが可能である．

3 再生核 Hilbert C^*-module

再生核 Hilbert C^*-module (RKHM) は再生核 Hilbert 空間の C^* 環を用いた一般化である．RKHM 上でデータ解析を行うことで，関数，行列，作用素等を用いて表されるデータを効率的に扱うことができる [2]．$\mathcal{A}$ を C^* 環，$\mathcal{X}$ を空でない集合とする．

定義 1（$\mathcal{A}$ 値正定値カーネル） $\mathcal{A}$ に値を取る写像 $k\colon \mathcal{X}\times\mathcal{X}\to\mathcal{A}$ が以下の 2 条件を満たすとき，$\mathcal{A}$ 値正定値カーネルと呼ばれる．

1) $x, y\in\mathcal{X}$ に対して，$k(x,y)=k(y,x)^*$,
2) $n\in\mathbb{N}$, $c_i\in\mathcal{A}$, $x_i\in\mathcal{X}$ に対して，$\sum_{i,j=1}^n c_i^* k(x_i,x_j)c_j\geq_{\mathcal{A}} 0$.

写像 $\phi:\mathcal{X}\to\mathcal{A}^{\mathcal{X}}$ を，$x\in\mathcal{X}$ に対して $\phi(x)=k(\cdot,x)$ で定義する．これを用いて $\mathcal{A}$ に値を持つ関数により構成される加群 $\mathcal{M}_{k,0}=\left\{\sum_{i=1}^n \phi(x_i)c_i \middle|\ n\in\mathbb{N},\ c_i\in\mathcal{A},\ x_i\in\mathcal{X}\right\}$ を考える．さらに，$\mathcal{A}$

値写像 $\langle\cdot,\cdot\rangle_{\mathcal{M}_k} : \mathcal{M}_{k,0} \times \mathcal{M}_{k,0} \to \mathcal{A}$ を以下のように定義する．

$$\left\langle \sum_{i=1}^{n} \phi(x_i)c_i, \sum_{j=1}^{l} \phi(y_j)d_j \right\rangle_{\mathcal{M}_k} = \sum_{i=1}^{n}\sum_{j=1}^{l} c_i^* k(x_i, y_j)d_j.$$

定義 1 より，$\langle\cdot,\cdot\rangle_{\mathcal{M}_k}$ は，内積を拡張した概念である $\mathcal{A}$ 値内積となる．空間 $\mathcal{M}_{k,0}$ の完備化 $\mathcal{M}_k$ は k に関する RKHM と呼ばれる．

4　関数値カーネルの構成

$\mathbb{T}$ を 1 次元トーラスとし，$\mathcal{A} = C(\mathbb{T})$ とおく．$j \in \mathbb{Z}$，$z \in \mathbb{T}$ に対して，$e_j(z) = \mathrm{e}^{\mathrm{i}jz}$ とおく．ただし，i は虚数単位である．関数 $x \in C(\mathbb{T})$ に対して，$R_n(x)_{j,l} = \int_{\mathbb{T}} x(t)\mathrm{e}^{-\mathrm{i}(j-l)t}\mathrm{d}t$ と定義する．また，行列 $A \in \mathbb{C}^{n\times n}$ に対して，$S_n(A) \in C(\mathbb{T})$ を，$S_n(A)(z) = 1/n \sum_{j,l=0}^{n-1} A_{j,l}\mathrm{e}^{\mathrm{i}(j-l)z}$ と定義する．ただし，$A_{j,l}$ は A の (j,l) 成分である．S_n は，行列 $R_n(x)$ を，元の関数 x に戻す役割を果たす．R_n，S_n を用いた関数の変換を，spectral truncation と呼ぶ．

上で定義した R_n，S_n を用いて，以下のようにカーネルを構成する．

定義 2　$q \in \mathbb{N}$，$\alpha_i \geq 0$，$\tilde{k}_{i,j} : \mathbb{C}\times\mathbb{C} \to \mathbb{C}$ を，複素数値正定値カーネルとする．また，$\tilde{k}_{i,j}(x,y)$ により，写像 $z \mapsto \tilde{k}_{i,j}(x(z), y(z))$ を表す．$x = [x_1, \ldots, x_d], y = [y_1, \ldots, y_d] \in \mathcal{A}^d$ と $z \in \mathbb{T}$ に対して，以下のように定義する．

$$k_n^{\mathrm{poly},q}(x,y) = S_n\Big(\sum_{i=1}^{d} \alpha_i (R_n(x_i)^*)^q R_n(y_i)^q\Big),$$

$$k_n^{\mathrm{prod},q}(x,y) = S_n\Big(\prod_{j=1}^{q} R_n(\tilde{k}_{1,j}(x,y))^* \prod_{j=1}^{q} R_n(\tilde{k}_{2,j}(x,y))\Big).$$

例えば，$k_n^{\mathrm{poly},q}(x,y)$ の場合，関数同士の積 $(x_i^*)^q y_i^q$ は可換であり，各点ごとの積となるが，R_n を用いて Toeplitz 行列 $R_n(x_i)$ と $R_n(y_i)$ を構成することにより，$(R_n(x_i)^*)^q R_n(y_i)^q$ は非可換な積となる．これにより，構成されるカーネルの $z \in \mathbb{T}$ における値 $k_n^{\mathrm{poly},q}(x,y)(z)$ は，$x_i(z)$ や $y_i(z)$ だけでなく，$z \neq w$ となる $x_i(w)$ や $y_i(w)$ にも依存する．パラメータ n を truncation パラメータと呼び，$n = 1$ の場合は separable カーネル，$n = \infty$ の場合は commutative カーネルと呼ばれる既存のカーネルとなる．separable カーネルは入力関数 x, y の大局的な依存関係を抽出する一方，commutative カーネルは局所的な依存関係を抽出する．よって，$1 < n < \infty$ の場合は大局的な依存関係と局所的な依存関係を同時に抽出し，n の値が小さいほど大局的な依存関係を重視する．この意味で n は，大局的な依存関係と局所的な依存関係をコントロールするパラメータとなっている．

謝辞　本研究は，Ayoub Hafid 氏，池田正弘氏，Hachem Kadri 氏との共同研究に基づく．

参考文献

[1] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2001.

[2] Y. Hashimoto et al. Reproducing kernel Hilbert C^*-module and kernel mean embeddings. *JMLR*, 22(267):1–56, 2021.

φ^4 スカラー場マシンによる多峰的モーメントマッチング問題の可解性について

On The Solvability of Some Multimodal Moment Matching Problems via the φ^4 Scalar Field Machine

加治佐 貴大 (KAJISA Takahiro)[1]
[1] 電気通信大学大学院情報理工学研究科 (The University of Electro–Communications, Graduate School of Informatics and Engineering)
e-mail : t.kajisa[at]uec.ac.jp

1 概要

ニューラルネットワークの一種であるボルツマンマシンは，そのユニット数や相互作用のパラメータを調整することによって目的分布を近似する生成的機械学習モデルである．しかし学習の更新に必要なダイバージェンスの計算量は NP 困難であることが知られており [1]，実用の際にはこの点が大きな障害となる．本研究では，ボルツマンマシンの「軟らかい」拡張である φ^4 スカラー場模型 [2] を用いることによって，ユニット数を増やすことなくボルツマンマシンの表現力の向上を図る．また実際に，これがある種の多峰的モーメントマッチング問題の解を与えることを示す．

2 φ^4–スカラー場マシン

スカラー場の量子論のトイモデルとしてよく使われる φ^4 模型を用いて，各ユニットが二峰的な分布の $\mathbb{R}$ 値確率変数をを持つようなボルツマンマシンを導入する．

定義 1. $T_i \geq 0,\ v_i > 0,\ m_i \geq 0,\ c_i \geq 0, c_N = 0\ i = 1, \ldots, N$ とする．このとき，φ_1^4–スカラー場マシン $\gamma_{T,v,m,c}$ は，ユニットごとのアプリオリな分布を

$$d\mu_i = \begin{cases} \frac{1}{\sqrt{\pi T_i}} \exp\left[-\frac{1}{T_i}(x_i^2 - v_i^2)^2\right] dx_i, & (T_i > 0), \\ \frac{1}{2}\left[\delta(x_i - v_i) + \delta(x_i + v_i)\right] dx_i, & (T_i = 0), \end{cases}$$

としたとき，

$$d\gamma_{T,v,m,c} = \frac{1}{Z_{v,m,c}} \exp\left[+\sum_{i=1,\ldots,N-1} c_i x_i x_{i+1} + \sum_{i=1,\ldots,N} m_i x_i\right] \prod_{i=1,\ldots,N} d\mu_i,$$

によって与えられる．ただし $Z_{v,m,c}$ は規格化定数である．また超関数の意味で，次が成り立つ．すなわち μ_i は $T_i = 0$ で連続である（図 1 参照）．

$$\lim_{T_i \to 0} \frac{1}{\sqrt{\pi T_i}} \exp\left[-\frac{1}{T_i}(x_i^2 - v_i^2)^2\right] dx_i = \frac{1}{2}\left[\delta(x_i - v_i) + \delta(x_i + v_i)\right] dx_i.$$

3 二峰的モーメントマッチング問題

次に本研究で取り扱うモーメントマッチング問題を導入する．

定義 2. μ を確率ベクトル $X = (X_1, \ldots, X_N) \in \mathbb{R}^N$ の N 次元分布とし，各 1 次元の周辺分布が二峰的であるとする．さらに μ は次のモーメントの制約を持つ：

$$\mathbb{E}_\mu[X_i^2] = V_i > 0, \qquad i = 1, \ldots, N, \tag{1}$$
$$\mathbb{E}_\mu[X_i] = M_i \geq 0, \qquad i = 1, \ldots, N, \tag{2}$$
$$\mathbb{E}_\mu[X_i X_{i+1}] = C_i \geq 0, \qquad i = 1, \ldots, N-1. \tag{3}$$

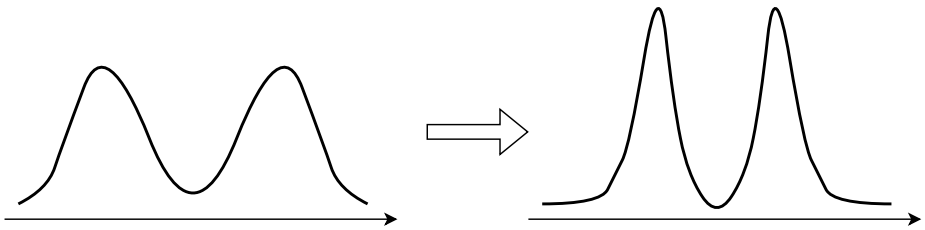


図 1. $d\mu_i$ の変化の様子．T_i を 0 へ近づける，すなわち低温へ近づけると，Ising 模型の分布が表現される．この意味で φ^4–スカラー場は Ising 模型のソフトな拡張となっている．

このとき，上のすべての条件を満たすような分布 μ を構成することを，*BFN-MMP (Bimodal Ferromagnetic Nearest–Neighbor Moment Matching Problem)* と呼ぶ．

次が本研究の主定理である：

定理 3. 任意の *BFN–MMP* に対して，φ^4–スカラー場マシンによる解が存在する．

証明には多次元の中間値の定理と構成的場の量子論における不等式を用いる．

注意 4. この理論は多峰的な場合にも適用できる．例えば *[3, Theorem 4]* によれば，

$$P(s) := qs^2(s-1)^2(s+1)^2 + 2\varepsilon(s^2 - s^4/2),$$

は超関数の意味で

$$\sqrt{\frac{q}{\pi}}e^{-P(s)} \to \delta(s) + \frac{\delta(s-1)+\delta(s+1)}{2e^{\varepsilon}}, \quad (q \to \infty),$$

に収束する．これを用いて φ^4–スカラー場マシンの場合と同様の議論を行うことができる．

4　結びに代えて

本研究ではここまでボルツマンマシンの拡張とモーメントマッチング問題の可解性について理論的な解析を行ってきた．本結果は木グラフや，分布の並進不変性を仮定した場合には多次元整数格子にも拡張することができる．一方，その実用性については（著者の力量不足によって）未知である．主な課題として，(i) 従来型ボルツマンマシンとのベンチマーク比較を行うこと，(ii) 多次元中間値定理の求根アルゴリズムを開発すること，および (iii) 面白い応用先を見つけることが挙げられる．

謝辞　本研究は、科学技術振興機構 戦略的創造研究推進事業 ERATO「竹内超量子もつれ」（研究代表者：竹内繁樹）（JPMJER2402）の一環として実施されました。

参考文献

[1] A. Bhattacharyya, S. Gayen, et al.“Computational explorations of total variation distance.” 13th ICLR, 2025.

[2] R. Fernández, J. Fröhlich, A. D. Sokal. *Random walks, critical phenomena, and triviality in quantum field theory.* Springer Science & Business Media, 2013.

[3] B. Simon, R.B. Griffiths. “The φ^4_2 Field Theory as a Classical Ising Model.” *Commun. Math. Phys.* **33**, 145–164 (1973).

最適輸送問題が成すストリング・ダイアグラム[*1]

String Diagrams of Optimal Transports

磯部 伸 (Noboru Isobe)
理化学研究所 革新知能統合研究センター (RIKEN AIP)
e-mail : noboru.isobe@riken.jp

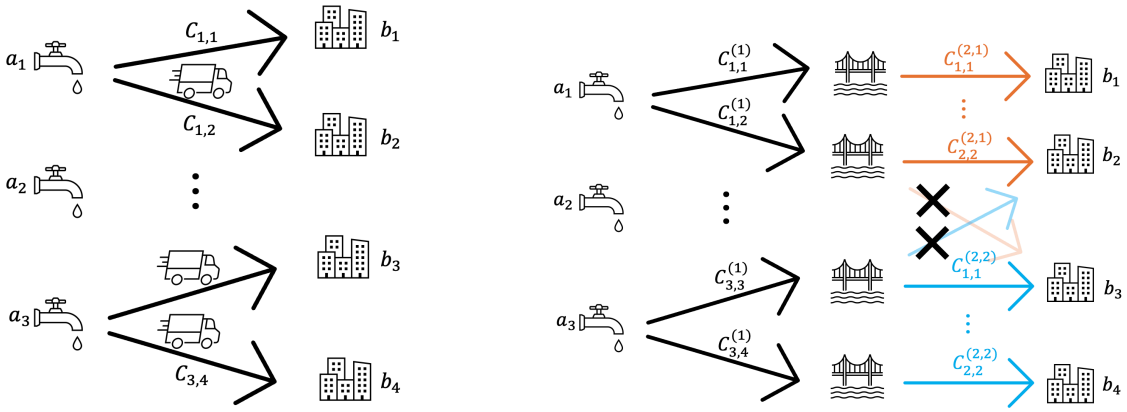
1　概要

最適輸送問題は，確率分布を「移動」させるのに係る最小の費用を計算する枠組みであり，オペレーションズリサーチや機械学習など，数理科学において幅広く応用されている最適化問題である．数学的には，二つの離散型確率分布 $\mathbf{a} = (a_i) \in \mathbb{R}^m_{\geq 0}$, $\mathbf{b} = (b_j) \in \mathbb{R}^n_{\geq 0}$ と，その間に定まるコスト関数（行列）$\mathbf{C} \in \mathbb{R}^{m \times n}$ に対して，（離散）最適輸送問題 $\mathrm{OT}(\mathbf{C}, \mathbf{a}, \mathbf{b})$ は

$$\min_{\mathbf{P} \in \Pi(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle, \Pi(\mathbf{a}, \mathbf{b}) \coloneqq \left\{ \mathbf{P} = (P_{ij}) \in \mathbb{R}^{m \times n}_{\geq 0} \;\middle|\; \forall (i,j) \in [m] \times [n], \sum_{j=1}^{n} P_{ij} = a_i, \sum_{i=1}^{m} P_{ij} = b_j \right\}$$

で定義される．ここで $\langle \cdot, \cdot \rangle$ は行列の Frobenius 内積であり，整数 n に対し $[n] \coloneqq \{ j \in \mathbb{Z} \mid 1 \leq j \leq n \}$ とした．この問題を実際に解く際には，輸送元 $\mathbf{a}$ の台の各点 $i \in [m]$ から，輸送先 $\mathbf{b}$ の台の各点 $j \in [n]$ までの輸送費用 C_{ij} が，全てのペア (i, j) について既知である必要がある（図 1a）．したがって，この最適輸送問題は，図 1b のように，始点 i から終点 j までに中間地点があったり（逐次的構造），輸送する方法が複数ある（並列的構造）ような場合には，素直には適用することが難しい．

本講演では，**ストリングダイアグラム**と呼ばれる代数的対象を用いた新たな最適輸送問題を提案する．ストリングダイアグラムは，逐次・並列的構造によって構築される階層構造を表現する方法の一つである．このモデルは，階層構造を持つような輸送経路を考慮した最適輸送問題を定式化することを可能にする．さらに我々は，ストリングダイアグラムの構造が誘導する，コスト行列たちがなす代数を利用することによって，効率的なアルゴリズムが導出できることを発見した．



(a) 通常の最適輸送問題が表す輸送　　(b) 提案する問題が表し得る輸送

図 1: 通常の最適輸送問題と提案する問題が表現し得る輸送の違いの例．

2　問題設定–合成的最適輸送問題–

考える問題を単純化して述べる：より一般的な定義は [1] をみよ．コスト行列のサイズ $m^{\mathcal{B}}, n^{\mathcal{B}} \in \mathbb{N}$ とコスト行列 $\mathbf{C}^{\mathcal{B}} \in \mathbb{R}^{m^{\mathcal{B}} \times n^{\mathcal{B}}}$ の組を $\mathcal{B}$ と表し，open OT (oOT) と呼ぶ．oOT の族 $\mathbb{D} \coloneqq ((\mathcal{B}_{kl})_{l=1}^{L_k})_{k=1}^{H}$ $(H, L_1, \dots, L_H \in \mathbb{N})$ が，条件 $\sum_{l \in [L_k]} n^{\mathcal{B}_{kl}} = \sum_{l' \in [L_{k+1}]} m^{\mathcal{B}_{(k+1)l'}}$ $(k \in [H-1])$ を満たすとき，$\mathbb{D}$ をストリングダイアグラムと呼ぶ．ストリングダイアグラム $\mathbb{D}$ と $m \coloneqq \sum_{l=1}^{L_1} m^{\mathcal{B}_{1l}}$ 次元確率分布 $\mathbf{a}$, $n \coloneqq \sum_{l=1}^{L_H} n^{\mathcal{B}_{Hl}}$ 次元確率分布 $\mathbf{b}$ に対し，**合成的最適輸送問題** $\mathrm{OT}(\mathbb{D}, \mathbf{a}, \mathbf{b})$ を

$$\min_{(\mathbf{P}^{\mathcal{B}_{kl}})_{kl} \in \Pi(\mathbb{D}, \mathbf{a}, \mathbf{b})} \sum_{k,l} \left\langle \mathbf{C}^{\mathcal{B}_{kl}}, \mathbf{P}^{\mathcal{B}_{kl}} \right\rangle$$

と定める．ここで $\Pi(\mathbb{D}, \mathbf{a}, \mathbf{b}) \subset \prod_{k,l} \mathbb{R}_{\geq 0}^{m^{\mathcal{B}_{kl}} \times n^{\mathcal{B}_{kl}}}$ は，次の三条件 $\sum_j P^{\mathcal{B}_1}_{ij} = a_i$, $\sum_i P^{\mathcal{B}_H}_{ij} = b_j$, $\sum_{i'} P^{\mathcal{B}_k}_{i'\bullet} = \sum_{j'} P^{\mathcal{B}_{k+1}}_{\bullet j'}$ $(k \in [H-1])$ を表す集合であり，$\mathbf{P}^{\mathcal{B}_k} \coloneqq \mathrm{diag}\left(\mathbf{P}^{\mathcal{B}_{k1}}, \dots, \mathbf{P}^{\mathcal{B}_{kL_k}}\right)$ である．直感的には，添字 k が逐次合成，l が並列合成を表している．

3　主定理，アルゴリズム

合成的最適輸送問題は，最適化問題としては変数が多く，愚直な線形計画法のソルバーでは大規模な確率分布を扱うことは現実的ではない．次の定理は，合成的最適輸送問題を（通常の）最適化問題に帰着させ解くことができることを可能にする．

定理 1 ([1, Corollary 4.6])**.** 合成的最適輸送問題 $\mathrm{OT}(\mathbb{D}, \mathbf{a}, \mathbf{b})$ に実行可能解が存在するとき，

$$\mathrm{OT}(\mathbb{D}, \mathbf{a}, \mathbf{b}) = \mathrm{OT}(\mathbf{C}^{\mathbb{D}}, \mathbf{a}, \mathbf{b}), \quad \mathbf{C}^{\mathbb{D}} \coloneqq (\mathbf{C}^{\mathcal{B}_{1,1}} \otimes \cdots \otimes \mathbf{C}^{\mathcal{B}_{1,L_1}}) \fatsemi \cdots \fatsemi (\mathbf{C}^{\mathcal{B}_{H,1}} \otimes \cdots \otimes \mathbf{C}^{\mathcal{B}_{H,L_H}})$$

が成り立つ．ここで，二つの演算 $\otimes, \fatsemi$ は，oOT $\mathcal{A} = (m^{\mathcal{A}}, n^{\mathcal{A}}, \mathbf{C}^{\mathcal{A}})$ と $\mathcal{B} = (m^{\mathcal{B}}, n^{\mathcal{B}}, \mathbf{C}^{\mathcal{B}})$ に対して，それぞれ

$$\mathbf{C}^{\mathcal{A}} \otimes \mathbf{C}^{\mathcal{B}} \coloneqq \begin{pmatrix} \mathbf{C}^{\mathcal{A}} & \infty \\ \infty & \mathbf{C}^{\mathcal{B}} \end{pmatrix} \in \overline{\mathbb{R}}^{(m^{\mathcal{A}}+m^{\mathcal{B}}) \times (n^{\mathcal{A}}+n^{\mathcal{B}})}, \qquad (\mathbf{C}^{\mathcal{A}} \fatsemi \mathbf{C}^{\mathcal{B}})_{ij} \coloneqq \min_k \{C^{\mathcal{A}}_{ik} + C^{\mathcal{B}}_{kj}\}$$

で定義される．ただし，$\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{+\infty\}$ であり，演算 $\fatsemi$ は，$n^{\mathcal{A}} = m^{\mathcal{B}}$ を満たす $\mathcal{A}, \mathcal{B}$ に対して定義される．

演算 $\fatsemi$ は，min-トロピカル半環 $(\overline{\mathbb{R}}, +_t, \cdot_t)$ 上の行列積とみなせる．ここで，加法 $r_1 + r_2$ は $r_1 +_t r_2 \coloneqq \min\{r_1, r_2\}$ として定義され，乗法 $r_1 \cdot_t r_2$ は $r_1 \cdot_t r_2 \coloneqq r_1 + r_2$ として定義する．したがって，$\mathbf{C}^{\mathcal{A}} \fatsemi \mathbf{C}^{\mathcal{B}}$ の合成に係る計算量は $\mathrm{O}(m^{\mathcal{A}} n^{\mathcal{A}} n^{\mathcal{B}})$ である．当日の講演で時間があれば，この計算量を削減する近似解法として，合成的最適輸送問題のエントロピー正則化に基づく Sinkhorn アルゴリズム [2] についても紹介する．

[*1] 本講演の内容は渡邊知樹（国立情報学研究所）との共同研究の内容 [1, 2] に基づく．

参考文献

[1] K. Watanabe and N. Isobe. *String Diagram of Optimal Transports*. 2025. arXiv: 2408.08550v2 [cs.AI].

[2] K. Watanabe and N. Isobe. *Sinkhorn Algorithm for Sequentially Composed Optimal Transports*. 2025. arXiv: 2412.03120v4 [cs.DS].

