2024 理論談話会 #11　6月6日

Sinkhorn Distances:
Lightspeed Computation of Optimal Transport

Cuturi, M. (2013). Advances in neural information processing systems, 26.

B4 薬師神晴悟 / Seigo Yakushijin

記法

- 同じサイズの2つの行列の内積 $\langle \cdot, \cdot \rangle : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \to \mathbb{R}$

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{A}_{ij} \boldsymbol{B}_{ij}$$

1. Abstract

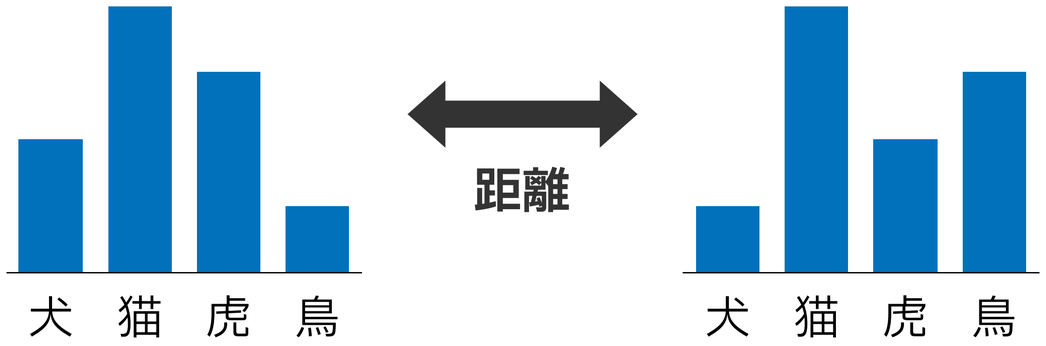
Abstract

- 確率分布やヒストグラムを比較する上での基本的な距離である最適輸送距離は優れた性質を持っていて定式化も直感的だが、測度の台のサイズやヒストグラムの次元が数百を超えると計算コストが急激に増大する。
- 従来の最適輸送問題にエントロピー正則化項を導入し、最適輸送距離がSinkhornアルゴリズムを用いて従来の手法よりもずっと早く計算できることを示す。
- 正則化された輸送距離を用いることで、MNISTの分類能が向上することも示す。

2. Introduction

確率分布間の距離

- 機械学習分野では、確率分布の比較を行う（＝確率分布間の**距離**を考える）場面が多くある。
- 適切な比較方法（**距離**）を選ぶことで、効果的な評価や訓練、検定を行うことができる。
  - よく使われるものの1つとして**KLダイバージェンス**（Kullback-Leibler divergence）がある。



離散分布に対するKLダイバージェンス

$$\mathrm{KL}(\boldsymbol{a}||\boldsymbol{b}) = \sum_{i=1}^{n} \boldsymbol{a}_i \log \frac{\boldsymbol{a}_i}{\boldsymbol{b}_i} - \boldsymbol{a}_i + \boldsymbol{b}_i$$

- しかしKLダイバージェンスは不都合な特徴がある。そこでこれに代わるものとして**最適輸送**を考える。

確率分布間の距離

- 他にも、Hellinger距離、Total Variation Distance、χ^2などがある。

離散分布に対するKLダイバージェンス

$$\mathrm{KL}(\boldsymbol{a}||\boldsymbol{b}) = \sum_{i=1}^{n} \boldsymbol{a}_i \log \frac{\boldsymbol{a}_i}{\boldsymbol{b}_i} - \boldsymbol{a}_i + \boldsymbol{b}_i$$

Total Variation Distance

$$\delta(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

離散分布に対するHellinger距離

$$H(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} \left(\sqrt{\boldsymbol{a}_i} - \sqrt{\boldsymbol{b}_i}\right)^2}$$

χ^2距離

$$\chi^2 = \frac{1}{2} \sum_{i=1}^{n} \frac{(\boldsymbol{a}_i - \boldsymbol{b}_i)^2}{\boldsymbol{a}_i + \boldsymbol{b}_i}$$

- 確率空間が距離空間のときは**最適輸送距離（Earth Mover’s Distance）**が強力である。

最適輸送の直感的なイメージ

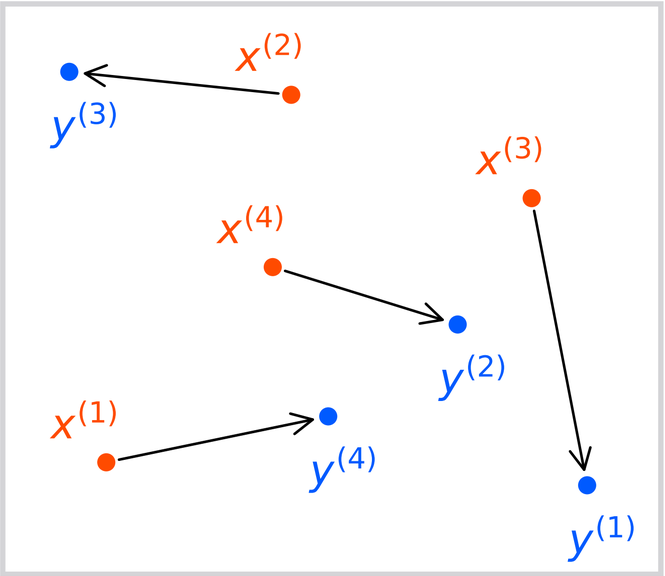
- 直感的には「**一方の分布の質量をもう一方の分布に合わせるのに必要な輸送コスト**」で定義される。

土砂を運ぶイメージから、

Earth Mover's Distance（EMD）

とも言う。

- n個の点群を比較する場合は、各点に質量$1/n$の砂山があると考え、その輸送距離を考える。
  - **最もコストが小さくなるような**輸送方法を選ぶ。



最適輸送距離の利点

- 最適輸送距離は以下の3点でKLダイバージェンスよりも優れている。

    1. **距離構造を捉えられる**
    2. **距離の公理を満たす**
    3. **サポートが一致していなくても定義できる**

1. 距離構造を捉えられる｜最適輸送距離の利点

- KLダイバージェンスでは、ヒストグラムのラベル間の距離を考慮した分布の差異を捉えられない。
  - 例：KLダイバージェンスだと、猫→犬と猫→虎が同じ「距離」になってしまうが、直感的でない。
- 最適輸送距離では、直感に合う自然な分布どうしの距離を考えられる。

2. 距離の公理を満たす｜最適輸送距離の利点

- KLダイバージェンスは対称ではなく、距離の公理も満たさない。

3. サポートが一致していなくても定義できる｜最適輸送距離の利点

サポート

確率分布が定義されている全体空間$\mathcal{X}$の点のうち、近傍の確率が正となる点からなる集合。

定義より、サポートが重ならない確率分布どうしのKLダイバージェンスは無限大である。これでは距離の情報として役に立たない。

最適輸送距離ならサポートが重ならなくても「近くにある点は似ている」とみなすことができる。

最適輸送距離の欠点

最適輸送距離は今述べたように強力であるものの、**計算コストが嵩む**という欠点がある。

現在提案されているどのようなアルゴリズムを用いたとしても、次元dのヒストグラム、もしくは距離空間上のd個の点群同士を比較するとき、最適輸送距離の計算には$O(d^3 \log d)$の計算量が必要。

埋め込み

距離空間の間の写像 $\phi : X \to Y$ が（distortion $C > 0$ の）**埋め込み**とは

$$Ld_X(x, y) \leq d_Y(\phi(x), \phi(y)) \leq CLd_X(x, y)$$

がある定数 $L > 0$ に対して成り立つことをいう。

最適輸送距離の欠点

Metric probability spaceが$\mathbb{R}^n$に埋め込めて、nが比較的小さいとき、最適輸送距離の計算コストは比較的小さく済む。

$n = 1$のとき	計算量は$O(d \log d)$で済む。
$n \geq 2$のとき	• 測度の埋め込みにより線形時間で近似的に求めることができる。 • Network Simplex法を修正することにより2乗の時間で求めることができる。
$n > 4$のとき	nが大きくなるに伴って埋め込みのdistortion（歪み）が大きくなり、計算コストも**指数関数的に増大する**。

最適輸送距離の欠点

- 任意の距離空間において数百の点／ビンでサポートされている測度間の距離を計算するのに、単一のCPUで数秒以上かかる。
- この欠点により、高次元のヒストグラムや高次元空間における測度が広く用いられる機械学習分野に、最適輸送距離を適用するのが困難になっている。

提案する手法

最適輸送問題にエントロピー項を導入して正則化する。

- 確率空間の性質（低次元のユークリッド空間への埋め込み可能性等）に依らず適用することができる。
- 正則化は最適輸送問題の幾何的性質を考えると直感的で、かつ交通パターンの推定のために交通理論でも長く使われてきた。
- 最適輸送問題を狭義凸計画問題に書き換えることにより、Sinkhorn アルゴリズムを含むmatrix scaling algorithm で解ける。
- 並列化によりGPGPUアーキテクチャを用いることができる。
- MNIST分類実験においても正則化により性能が良くなり、高速化が実現される。

2. Reminders on Optimal Transport

最適輸送問題の定式化

（点群の場合）

点xから点yに一単位の質量を輸送するのにかかるコスト

$$\underset{\boldsymbol{P}\in\mathbb{R}^{d\times d}}{\text{minimize}} \quad \sum_{i=1}^{d}\sum_{j=1}^{d} C(x_i, y_j)\boldsymbol{P}_{ij}$$

$$\text{subject to} \quad \boldsymbol{P}_{ij} \geq 0$$

輸送量は非負

$$\sum_{j=1}^{d} \boldsymbol{P}_{ij} = \boldsymbol{r}$$

$$\sum_{i=1}^{d} \boldsymbol{P}_{ij} = \boldsymbol{c}$$

質量保存制約

制約条件を満たす行列全体の集合

輸送多面体

$$U(r,c) := \{P \in \mathbb{R}_{+}^{d\times d} \mid P\mathbf{1}_d = r, P^T\mathbf{1}_d = c\}$$

書き換えると…

$$d_M(r,c) = \underset{P\in U(r,c)}{\text{minimize}} \langle P, M\rangle$$

輸送多面体の解釈

確率$\boldsymbol{P}_{ij}$で(i,j)をとる分布を考えると、制約条件より$\boldsymbol{r},\boldsymbol{c}$は$\boldsymbol{P}$の周辺分布となる。

すなわち、$\boldsymbol{P}_{ij}$ は $\boldsymbol{r},\boldsymbol{c}$ の同時確率。

$\boldsymbol{M}_{ij} := C(x_i, y_j)$ の期待値は$\langle \boldsymbol{P}, \boldsymbol{M}\rangle$である。

最適輸送問題の定式化

$$d_M(r,c) = \underset{P \in U(r,c)}{\text{minimize}} \langle P, M \rangle$$

Mが距離空間であるとき、$d_M(r,c)$は r, c 間の距離である。

Mが距離空間である

Mが距離の公理を満たすような行列の錐$\mathcal{M}$：

$$\mathcal{M} = \{M \in \mathbb{R}_+^{d \times d} : \forall i,j \leq d, m_{ij} = 0 \Leftrightarrow i = j,\ \forall i,j,k \leq d, m_{ij} \leq m_{ik} + m_{kj}\}$$

に属する。

一般に行列Mについて、現在提案されているアルゴリズムだと、計算量は$O(d^3 \log d)$である。

（＝**Super Cubic**）

3. Sinkhorn Distances:

Optimal Transport with Entropic Constraints

エントロピーの導入

$P, Q \in U(r,c), r \in \sum_d$ について、

$$h(r) = -\sum_{i=1}^{d} r_i \log r_i, \quad h(P) = -\sum_{i,j=1}^{d} p_{ij} \log p_{ij}, \quad \mathbf{KL}(P||Q) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

と定める。エントロピーに関する不等式

$$\forall r, c \in \sum_d, \forall P \in U(r,c), h(P) \leq h(r) + h(c)$$

は、r, c が独立であるので厳密に成立。

$$\begin{aligned}
h(P) &= -\sum_{i,j=1}^{d} p_{ij} \log p_{ij} \\
&\leq -\sum_{i,j=1}^{d} p_{ij} \log(r_i c_j) \\
&= -\sum_{i,j=1}^{d} p_{ij} \log r_i - \sum_{i,j=1}^{d} p_{ij} \log c_i \\
&= -\sum_{i=1}^{d} r_i \log r_i - \sum_{i=1}^{d} c_i \log c_i \\
&= h(r) + h(c)
\end{aligned}$$

エントロピーの凹性から、次のような凸集合が定義できる。

$$\begin{aligned}
U_\alpha(r,c) &:= \{P \in U(r,c) \mid \mathbf{KL}(P||rc^T) \leq \alpha\} \\
&= \{P \in U(r,c) \mid h(P) \geq h(r) + h(c) - \alpha\} \subset U(r,c)
\end{aligned}$$

Sinkhorn Distanceの定義

$$U_\alpha(r,c) := \{P \in U(r,c) \mid \mathbf{KL}(P||rc^T) \leq \alpha\} = \{P \in U(r,c) \mid h(P) \geq h(r) + h(c) - \alpha\} \subset U(r,c)$$

- 2つの確率変数 (X, Y) が同時確率Pに従うときの相互情報量 $I(X||Y)$ と解釈できる。
- したがって、rc^Tとの間のKLダイバージェンスが一定の値（α）以下であるような行列Pの集合$U_\alpha(r,c)$は、$h(r)$と$h(c)$に関して十分なエントロピー、つまり十分小さい相互情報を持つ$U(r,c)$の同時確率Pの集合と解釈できる。

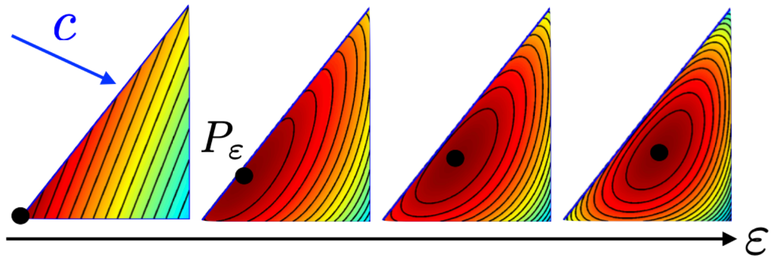
これを用いて、**Sinkhorn Distance** を次のように定める。

Sinkhorn Distance

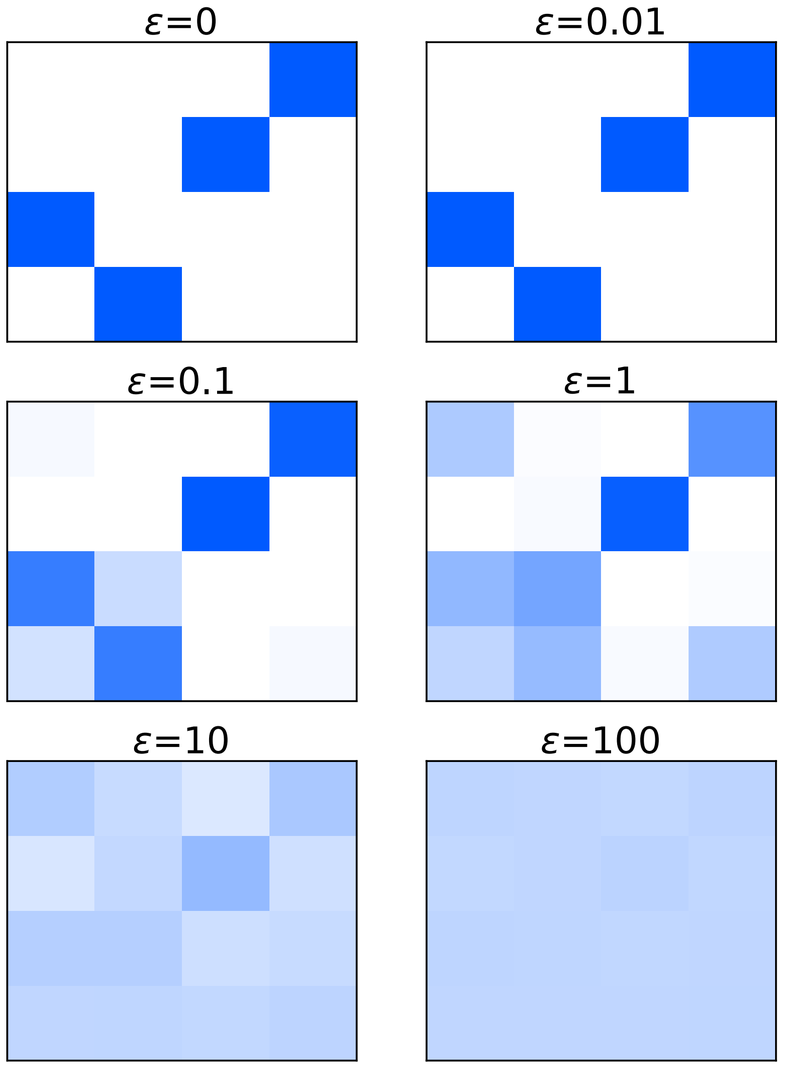
$$d_{M,\alpha}(r,c) := \min_{P \in U_\alpha(r,c)} \langle P, M \rangle$$

なぜエントロピー制約を考慮するのか？

- 計算上の利点がある（**後述します**）。
- 通常の最適輸送では最適解は多面体の頂点に存在するがこのとき最適解は疎になる。一方で、エントロピー正則化を導入すると最適解が多面体の中央に寄っていき、最適解は密になる。
  - 最適解が多面体の中央に寄ると、線形計画問題の解が一意に定まる（頂点だと面全体で最適を取ることがある）。



  - 最適解が密になると、最適解から遠く離れた場所でも解の更新に有用な勾配情報を得られる。



Sinkhorn Distanceの性質

1. αが十分に大きいとき、Sinkhorn Distance $d_{M,\alpha}$は最適輸送距離d_Mに一致する。

   エントロピー正則化されていない従来のもの

2. $\alpha = 0$のとき、Mがユークリッド距離行列（Euclidean distance matrix）であれば、$d_{M,0} = r^T Mc$は**閉形式**で**負定値カーネル**となる。また、$e^{-td_{M,0}}$はすべての$t > 0$について**正定値**となる。

αが十分小さい場合、$h(r) > 0$ となるような r について $d_{M,\alpha}(r, r) > 0$ となる。→ **距離の公理✕**

↓ $d_{M,\alpha}$ に $\mathbf{1}_{r \neq c}$ をかける

定理1

任意の $\alpha \geq 0, M \in \mathcal{M}$ について、$d_{M,\alpha}$ は対称かつ三角不等式を満たす。

また、関数 $(r, c) \mapsto \mathbf{1}_{r \neq c} d_{M,\alpha}(r, c)$ は距離の公理を3つすべて満たす。

定理1の証明

補題1　**（エントロピー制約付きの貼り合わせ補題）**

$\alpha \geq 0$ と $x, y, z \in \sum_{d}$ について、$P \in U_\alpha(x, y), Q \in U_\alpha(y, z)$ とし、S を $s_{ik} := \sum_{j} \frac{p_{ij} q_{jk}}{y_j}$ を満たすような $d \times d$ 行列とするとき、$S \in U_\alpha(x, z)$ が成り立つ。

これを用いることにより、最適輸送距離が実際に（距離の公理を満たす）**距離**であることを示すことができる（**次頁参照**）。

定理1の証明

定理1

任意の $\alpha \geq 0, M \in \mathcal{M}$ について、$d_{M,\alpha}$ は対称かつ三角不等式を満たす。

また、関数 $(r,c) \mapsto \mathbf{1}_{r \neq c} d_{M,\alpha}(r,c)$ は距離の公理を3つすべて満たす。

$d_{M,\alpha}$の対称性はMの対称性から直ちに言える。

x, y, z を$\sum_d$の3つの要素とし、$P \in U_\alpha(x,y), Q \in U_\alpha(y,z)$ は $d_{M,\alpha}(x,y), d_{M,\alpha}(y,z)$ それぞれの最適輸送問題の解であるとする。前頁で示した補題1で与えたSを用いて、次のように三角不等式を示すことができる：

$$
\begin{aligned}
d_{M,\alpha}(x,z) &= \min_{P \in U_\alpha(x,z)} \langle P, M \rangle \leq \langle S, M \rangle = \sum_{ik} m_{ik} \sum_j \frac{p_{ij} q_{jk}}{y_j} \leq \sum_{ijk} (m_{ij} + m_{jk}) \frac{p_{ij} q_{jk}}{y_j} \\
&= \sum_{ijk} m_{ij} \frac{p_{ij} q_{jk}}{y_j} + m_{jk} \frac{p_{ij} q_{jk}}{y_j} = \sum_{ij} m_{ij} p_{ij} \sum_k \frac{q_{jk}}{y_j} + \sum_{jk} m_{jk} q_{jk} \sum_i \frac{p_{ij}}{y_j} \\
&= \sum_{ij} m_{ij} p_{ij} + \sum_{jk} m_{jk} q_{jk} = d_{M,\alpha}(x,y) + d_{M,\alpha}(y,z)
\end{aligned}
$$

P, Q は同時確率行列であるためこの部分は 1

4. Computing Regularized Transport with Sinkhorn’s Algorithm

Sinkhorn Distanceのエントロピー制約のLagrange乗数

Sinkhorn Distanceのエントロピー制約に関するLagrange乗数を考える。

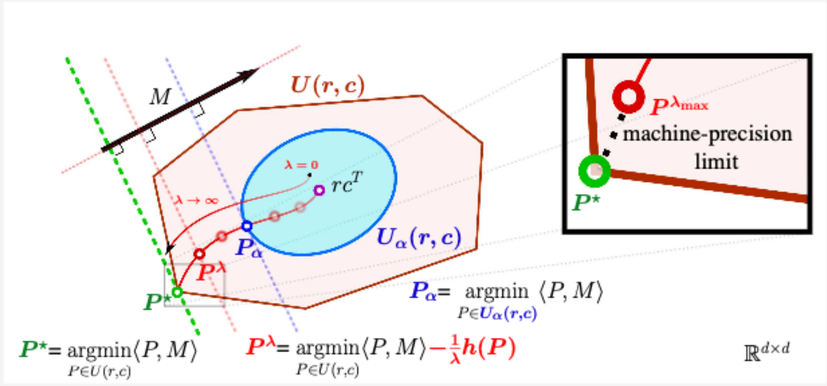
$$\text{For } \lambda > 0, d_M^\lambda(r,c) := \langle P^\lambda, M \rangle, \text{where } P^\lambda = \underset{P \in U(r,c)}{\operatorname{argmin}} \langle P, M \rangle - \frac{1}{\lambda} h(P)$$

双対定理より、それぞれの α に $d_{M,\alpha}(r,c) = d_M^\lambda(r,c)$ となるような $\lambda \in [0, \infty]$ が対応する。

主問題と双対問題のいずれか一方が最適解を持つなら、もう一方も最適解を持ち、主問題の最小値と双対問題の最大値は一致する。

d_M^λ を**dual-Sinkhorn divergence**と呼ぶことにする。

これは、もともとの最適輸送問題 d_M に比べてずっと計算が楽であるという嬉しい性質がある。



d_M^λをmatrix-scalingアルゴリズムで計算する

補題2

$\lambda > 0$ において解P^λは一意に定まり、$P^\lambda = \mathrm{diag}(u)K\mathrm{diag}(v)$ の形で表すことができる。ただし、u, vは非負の$\mathbb{R}^d$ベクトルで、$K := e^{-\lambda M}$ は$-\lambda M$の項別の指数である。

- P^λの存在とその一意性は$U(r, c)$の有界性とマイナス（-）エントロピーの狭義凸性から従う。
- 前頁の式のラグランジアン$\mathcal{L}(P, \alpha, \beta)$を、$P\mathbf{1}_d = r, P^T\mathbf{1}_d = c$という$U(r, c)$の2つの制約条件に対する双対変数$\alpha, \beta \in \mathbb{R}^d$を用いて書く。

$$\mathcal{L}(P, \alpha, \beta) = \sum_{ij} \frac{1}{\lambda} p_{ij} \log p_{ij} + p_{ij} m_{ij} + \alpha^T (P\mathbf{1}_d - r) + \beta^T (P^T\mathbf{1}_d - c)$$

$\partial\mathcal{L}/\partial p_{ij} = 0$ より、$p_{ij} = e^{-1/2-\lambda\alpha_i} e^{-\lambda m_{ij}} e^{-1/2-\lambda\beta_j}$ となる。K は真に正の成分からなる行列なので、Sinkhornの定理によれば$U(r, c)$に属する $P^\lambda = \mathrm{diag}(u)K\mathrm{diag}(v)$ が $u, v \geq \mathbf{0}_d$において一意に存在する。

d_M^λをmatrix-scalingアルゴリズムで計算する

- P^λに**Sinkhorn's fixed point iteration** $(u, v) \leftarrow (r./Kv, c./K'u)$が適用できる。

Sinkhorn's fixed point iteration

$P^\lambda = \mathrm{diag}(u)K\mathrm{diag}(v)$の行和、列和がそれぞれ$r, c$となるように逐次更新していく。

```
1 Input M,λ,r,C := [c1,··· ,cN].
2 I = (r > 0);r = r(I);M = M(I,:);K = exp(-λM)
3 u = ones(length(r), N )/length(r);
4 K˜ = bsxfun(@rdivide, K, r) % equivalent to K˜ = diag(1./r)K while u changes or any other relevant stopping criterion do
5 u = 1./(K˜(C./(K´u))) end while
6 v = C./(K´u)
7 d = sum(u. * ((K. * M )v)
```

上記の疑似コードは$\boldsymbol{d} = \left[d_M^\lambda(r, c_1), \ldots, d_M^\lambda(r, c_N)\right]$を求めるためのもの。

並列化、収束判定基準

- 前頁の疑似コードで示したように、Sinkhornアルゴリズムはベクトル化してN個のヒストグラム$c_1, \ldots, c_N$に適用することができる。$N > 1$ のとき計算は**並列化**することができる。
  - これによりGPGPUアーキテクチャを活用することができる。
- Franklin and Lorenz は scaling factor u, v の収束は線形であり、その上界は

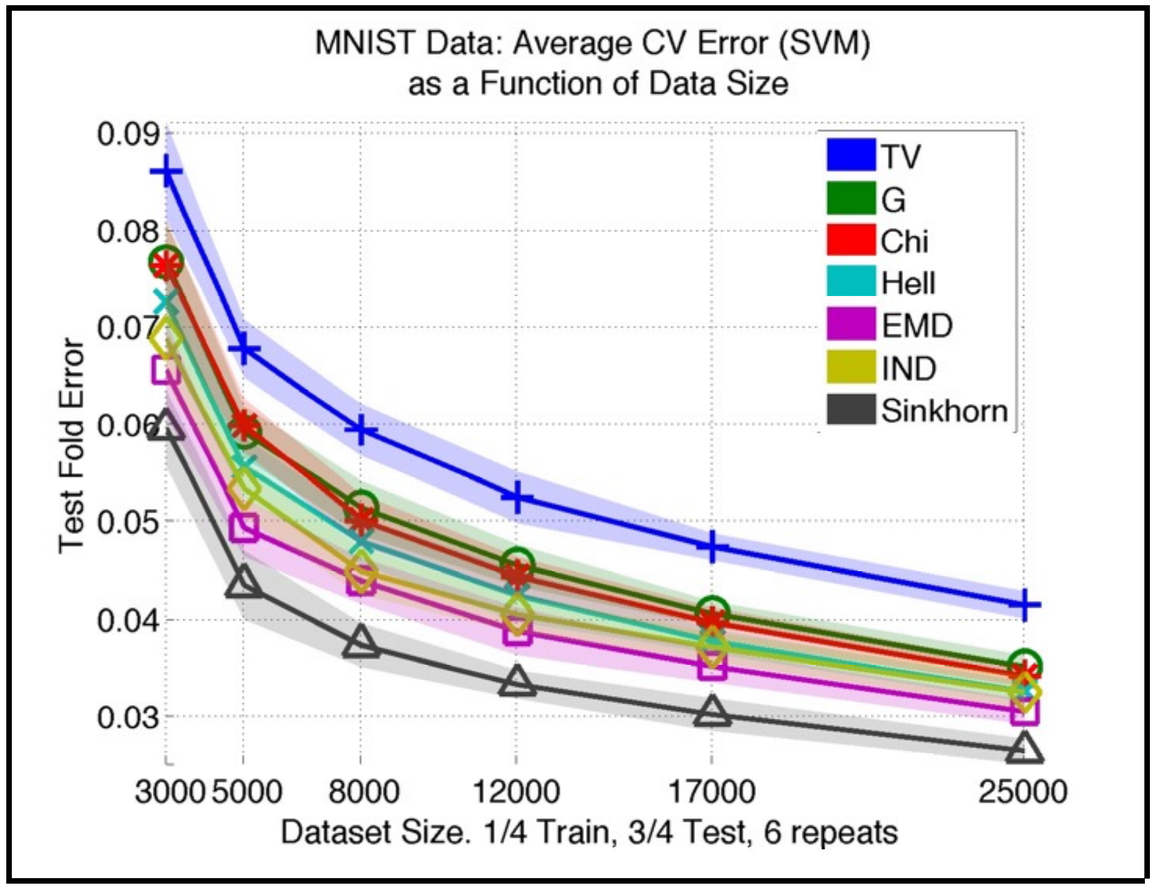
$$\kappa(K) = \frac{\sqrt{\theta(K)} - 1}{\sqrt{\theta(K)} + 1} < 1, \text{and } \theta(K) = \max_{i,j,l,m} \frac{K_{il}K_{jm}}{K_{jl}K_{im}}$$

としたとき、$\kappa(K)^2$であると示した。

> $\kappa(K)$は λ が大きくなるにつれて 1 に近づき、P^{λ}がP^{*}に近づくにつれて収束は遅くなる。
>
> そこで、**ここでは異なる収束判定基準を検討する。**

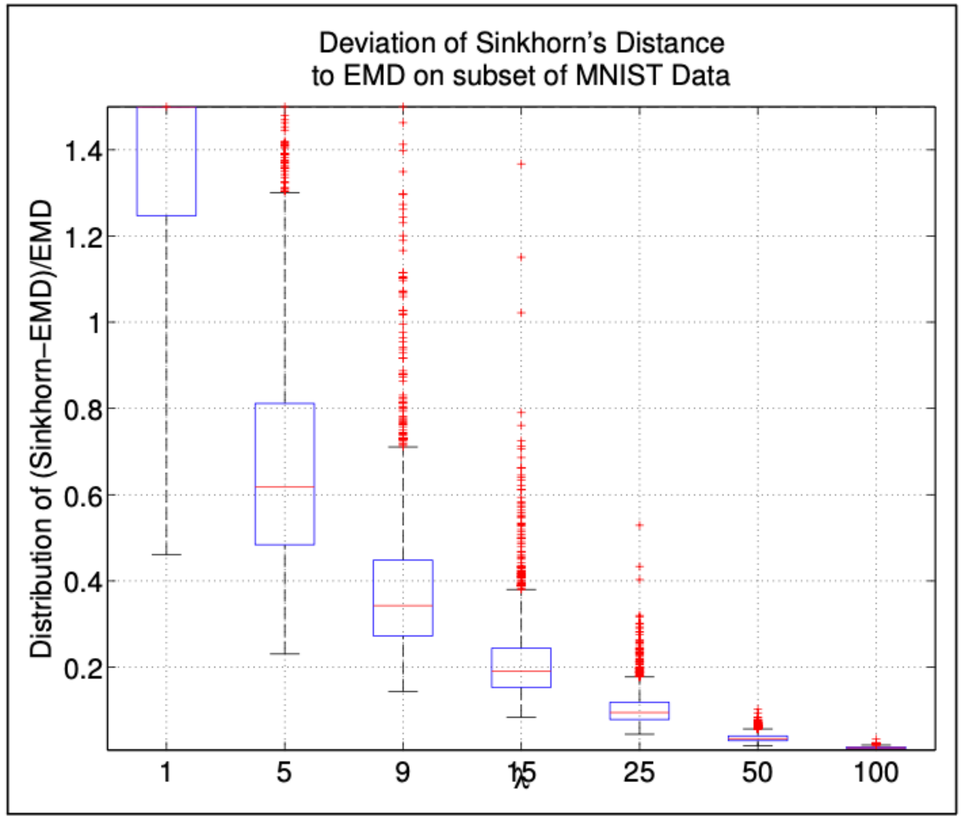
5. Experimental Results

MNIST Digitsの識別実験



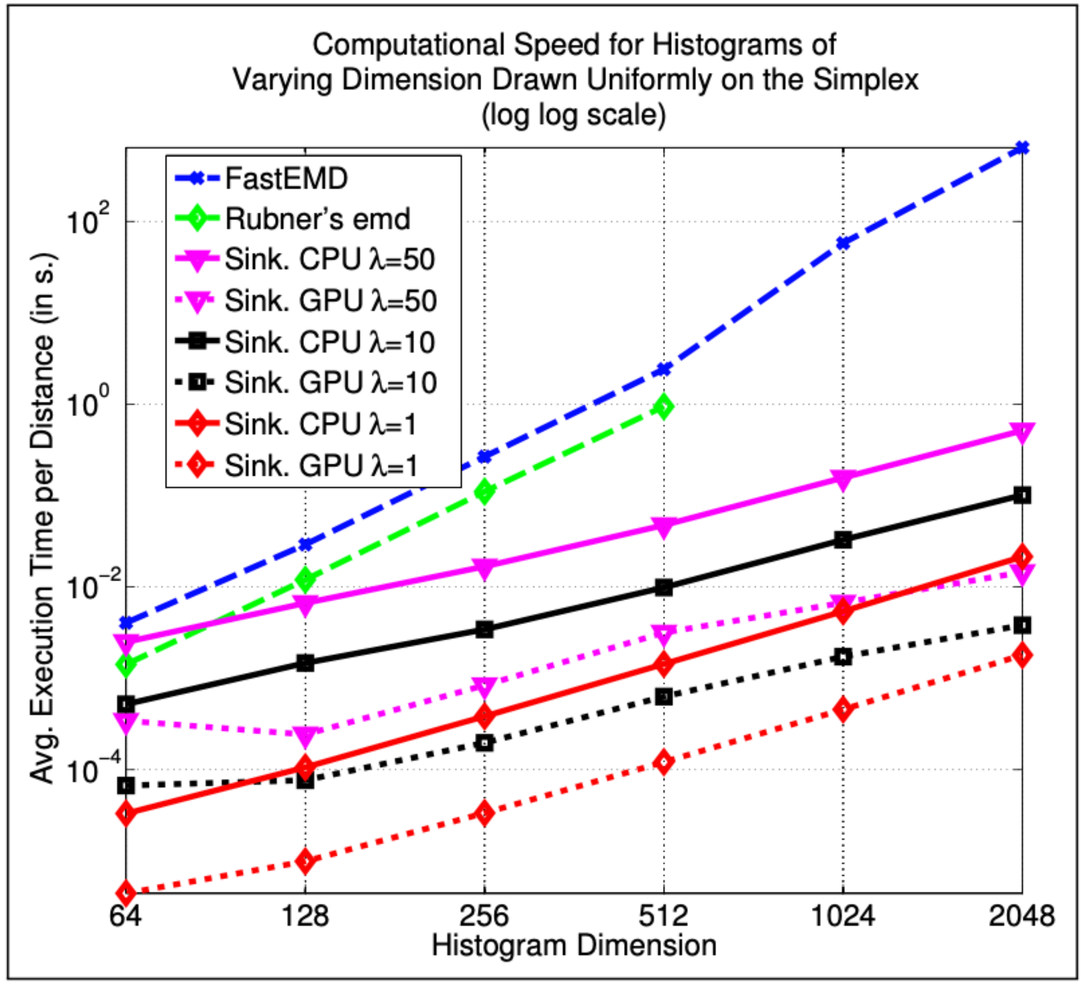
- Dual-Sinkhorn Divergence は他の距離（従来の最適輸送距離：EMD 含む）よりも**交差検証の結果が優れている**（エラーが少ない）。

Dual-Sinkhorn DivergenceはEMDと合致するのか？



- λが大きくなるにつれて2つの差は小さくなるという仮説。
- λが50を超えるとd_M^λはEMDを高い精度で推定することができる。
- この実験を含むこれ以降の実験では、$\boldsymbol{d}$のN個のどの要素の変化量の絶対値も $1/10000$ 未満になったとき、反復を停止することとする（**収束判定**）。

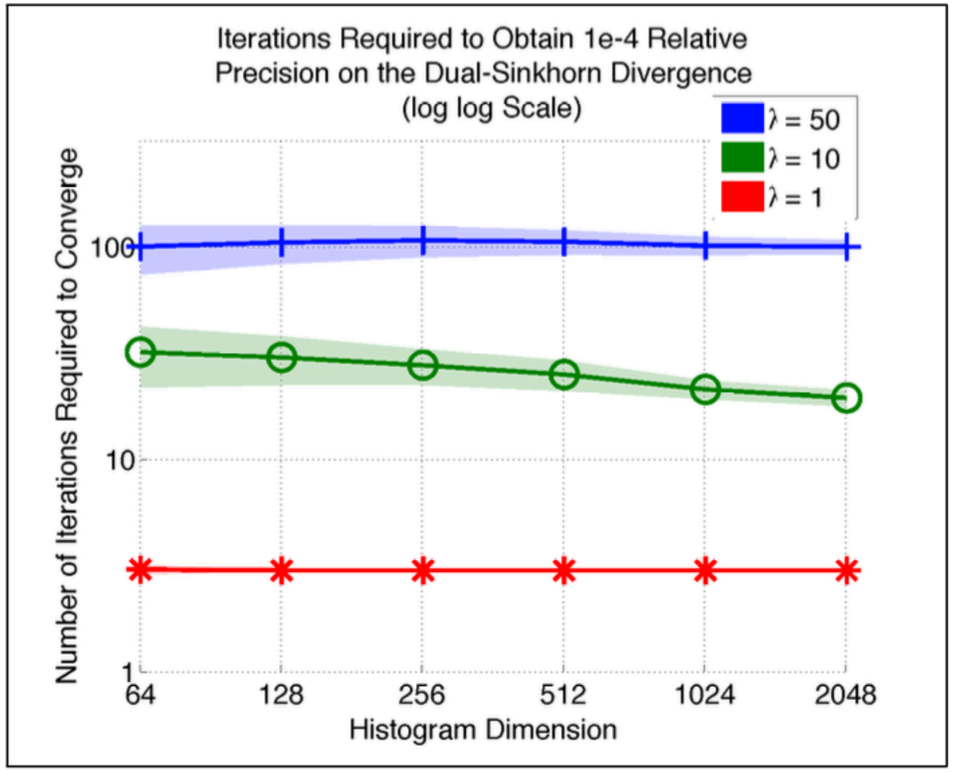
実際EMDよりも速いのか？



（λが小さいほど正則化項が支配的になる）

- Dual-Sinkhorn DivergenceはEMDよりも速く、その差は次元が大きくなるにつれて拡大する。
- GPUを使うと更なる高速化が実現できる。

収束に必要な反復回数



- 収束に必要な反復回数は$e^{-\lambda M}$が対角優位になるにつれて増大するが、次元には依存しないように見える。
- このことから、誤差を追跡する（$\boldsymbol{d}$の変化を逐次求める）ようなことは必要ではなく、反復回数をあらかじめ定めておくので十分である。
  また、このようなコストのかかる操作を省略することで高速化も期待できる。

6. Conclusion

まとめ

- 最適輸送問題にエントロピー正則化項を導入することにより、最適輸送を計算する上での新たな数理的な手法が拓けた。
- 本手法は確率空間の性質に依らず適用することができ、EMD と同様に機能するばかりか、応用上は EMD よりも強力であると言える。
- λ は計算と性能の両方のバランスを考えて設定されるべきであるが、そのトレードオフは検討されていない。
  - 適度に小さなλは大きなλよりも性能が良い。

所感

- 測度論や確率論など数学的に初見の概念が多く、数式を追うのに大変苦労した。
- 不幸中の幸いか、本論文に関する書籍が複数出版されていて、視覚的にわかりやすい図も多用されていたため、なんとなくのイメージを持つことは比較的容易だった。
- 夏研究では車と公共交通機関の間の相転移を扱おうとしているが、車から公共交通機関への手段選択の確率分布の遷移に関して最適輸送が適用できるのではないかと感じた。
    - 次元が大きくなる場合は今回の Sinkhorn Distance が有効であろう。
- Feeling regretful that I wasn’t able to have enough time to put in English annotations.

References

- 佐藤竜馬．最適輸送の理論とアルゴリズム．講談社，2023，308p，機械学習プロフェッショナルシリーズ．
- 佐藤竜馬．“最適輸送入門”．2022-03，https://www.slideshare.net/joisino/ss-251328369．
- 佐藤竜馬．“最適輸送の解き方”．2021-06，https://www.slideshare.net/joisino/ss-249394573．
- Gabriel Peyré and Marco Cuturi, Computational Optimal Transport, ArXiv:1803.00567, 2018.