

目次

- Wasserstein距離から導かれる確率分布空間の幾何構造
- Wasserstein統計多様体と引き戻し計量
- グラフ上のWasserstein統計多様体
- まとめと今後の課題

確率分布空間の幾何構造

- 確率分布間の近さを測ることはデータサイエンスで重要な役割を持つ
- 多くの問題が次のような形式を取る

$$\min_{\rho \in \mathcal{P}_\theta} d(\rho, \rho_{\text{target}})$$

\mathcal{P}_θ : パラメトリックな確率分布の集合

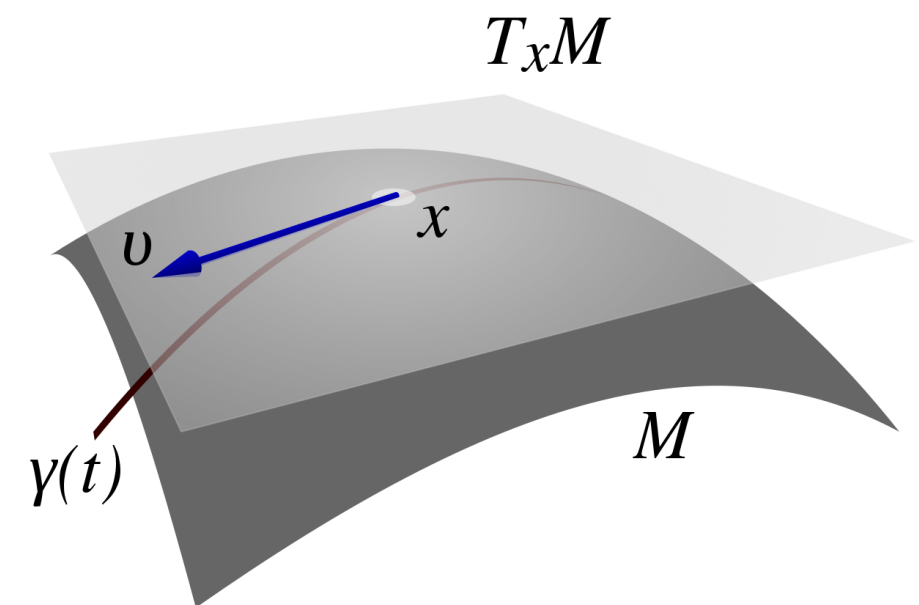
- d としてよく用いられるのがKLダイバージェンスである
- d から確率分布空間の幾何構造を導入したい！

内積と計量

\mathbb{R}^n の内積 $\langle \cdot, \cdot \rangle$ を考える

- 標準内積 $\langle x, y \rangle = x^T I y$
- 正定値行列 G に対して $\langle x, y \rangle = x^T G y$ も内積
- G を一定ではなく各点で依存させる (テンソル場)
- 計量: 多様体 M の各点 x での接空間に正定値行列 $G(x)$ が定義されるイメージ

多様体: 各点の近傍では
ユークリッド空間とみなせるもの



接空間のイメージ

内積と計量

- $G = \Sigma^{-1}$ とすればマハラノビス距離になる
- パラメータ空間 Θ において $G(\theta) =$ フィッシャー情報行列
とすれば統計多様体と呼ばれるものになる
- 自然勾配法:最急降下法

$$x^{t+1} = x^t - \gamma \nabla f(x^t)$$

ではなく空間の計量も加味して最も目的関数が下がる方向に降下する、特にパラメータの学習においてパラメータ空間のフィッシャー計量を用いたものを自然勾配法と呼ぶ

$$\theta^{t+1} = \theta^t - \gamma G^{-1}(\theta^t) \nabla f(\theta^t)$$

KLダイバージェンスからフィッシャー計量へ

ρ_θ と $\rho_{\theta+h}$ の間のKLダイバージェンスは以下のようにかける

$$\text{KL}(\rho_\theta | \rho_{\theta+h}) = \frac{1}{2} h^T I(\theta) h + o(\|h\|^2)$$

$I(\theta)$: フィッシャー情報行列

→ KLダイバージェンスからはフィッシャー計量が導かれる

→ 統計学や機械学習のアルゴリズムの解釈

→ 自然勾配法の開発

• 他の d ではどうなる？

Wasserstein距離

Def.1 Wasserstein距離

$\forall \mu, \nu \in \mathcal{P}_p(X)$ (p次モーメントが存在する X 上の確率測度全体)
に対して

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} d(x, y)^p d\pi(x, y) \right)^{1/p}$$

を μ, ν の p -Wasserstein距離と定義する

Wasserstein距離は最も効率の良い確率分布の輸送コストを表している

主に $p = 1, 2$ を扱う

カップリング

Def.2 カップリング

(X, Σ) : 確率空間, $\mathcal{P}(X)$: X 上の確率分布全体の集合

$\mu, \nu \in \mathcal{P}(X)$ に対して $\pi \in \mathcal{P}(X \times X)$ が μ, ν のカップリング
であるとは

$$\forall A \in \Sigma, \pi([A, X]) = \mu(A), \pi([X, A]) = \nu(A)$$

となることでありカップリング全体の集合を $\Pi(\mu, \nu)$ と表す

Wasserstein距離の特徴

- 元の空間の構造が反映されている

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} \underline{d(x, y)^p} d\pi(x, y) \right)^{1/p}$$

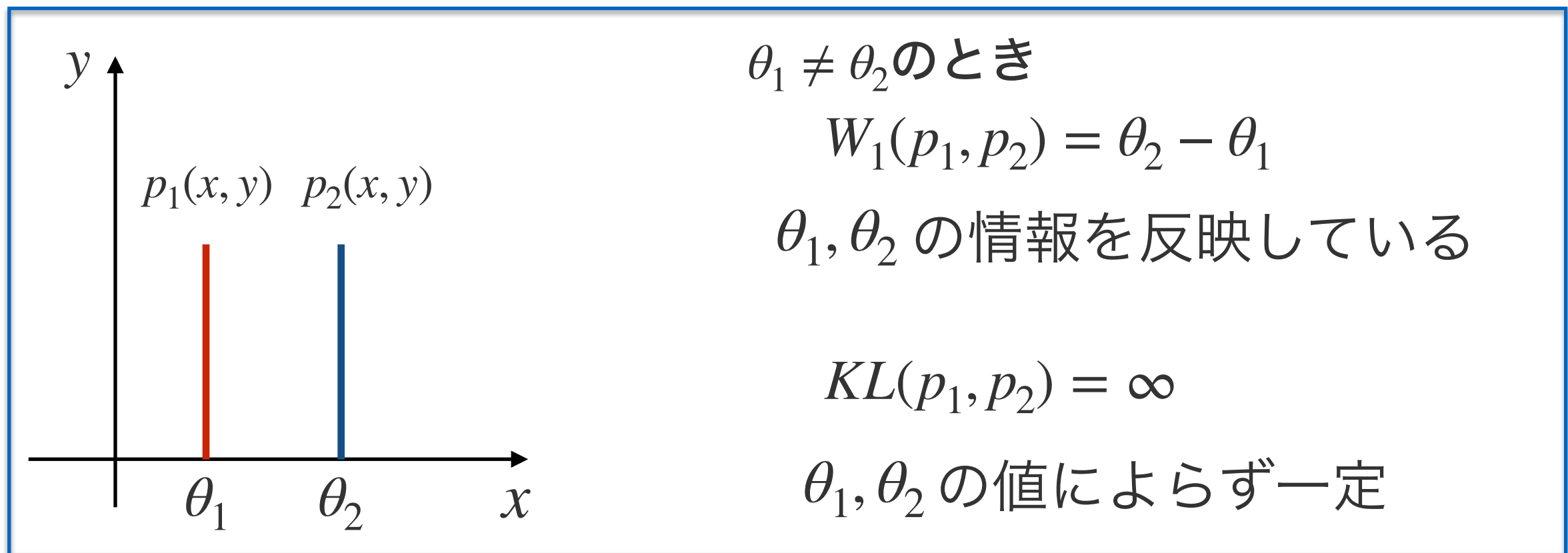
X の距離が定義に含まれている $\rightarrow X$ の空間構造が重要な場合有用

$$\text{KL}(\mu | \nu) = \int_X p(x) \log \frac{p(x)}{q(x)} dx$$

X の距離に関わらずその上の確率密度の値だけで定まっている

Wasserstein距離の特徴

- Wasserstein距離は**台を共有しない確率分布間の距離を計ることができる**(元の空間の距離を用いているから)
- 点群や経験分布同士の比較や高次元空間上の低次元多様体に集中する分布（多様体仮説）の比較に有用
- KLダイバージェンスなどでは不可能



$\mathcal{P}(X)$ の幾何構造

- $\mathcal{P}(X)$ に何らかの幾何構造を導入したい
 - Wasserstein距離→Wasserstein計量
 - KLダイバージェンス→フィッシャー計量
- Brenier-Benamouの公式から導入する

Thm.1 Benamou-Brenierの公式[Villani,2009]

$$W_2(\rho_0, \rho_1)^2 = \inf_{\Phi_t} \int_0^1 \int_X \|\nabla \Phi(t, x)\|^2 \rho(t, x) dx dt$$

inf は連続方程式を満たす (ρ, Φ) の組でとる

$$\frac{\partial \rho(t, x)}{\partial t} + \nabla \cdot (\rho(t, x) \nabla \Phi(t, x)) = 0, \quad \rho(0, x) = \rho_0(x), \quad \rho(1, x) = \rho_1(x)$$

$\mathcal{P}(X)$ の接空間

$$\mathcal{P}_+(X) = \left\{ \rho \in C^\infty : \rho(x) > 0, \int_X \rho(x) dx = 1 \right\} \cap \mathcal{P}_2(X)$$

$\mathcal{P}_2(X)$: X 上の2次モーメントを持つ確率分布全体

を考える

$\mathcal{P}_+(X)$ の各点での接空間を次のように定義する

$$T_\rho \mathcal{P}_+(X) = \left\{ \sigma \in C^\infty(X); \int_X \sigma(x) dx = 0 \right\}$$

楕円型作用素

$$\Phi \in C^\infty(X), \quad V_\rho(\Phi)(x) := -\nabla \cdot (\rho(x) \nabla \Phi(x))$$

と定義する。これは楕円型作用素の一種である

$$\mathcal{F}(X) = \left\{ \Phi \in C^\infty(X); \int_X V_\rho(\Phi)(x) dx = 0 \right\}$$

とおく。これは**zero flux condition**とも呼ばれる

Zero flux conditionより

$$\Phi \in \mathcal{F}(X) \Rightarrow V_\rho(\Phi) \in T_\rho \mathcal{P}_+(X)$$

Wasserstein計量テンソル

Thm.2[Li,2018]

$\mathcal{F}(X)/\mathbb{R} \rightarrow T_\rho \mathcal{P}_+(X), \Phi \mapsto V_\rho(\Phi)$ は同相写像である

Def.3 Wasserstein計量テンソル[Li,2018]

$g_\rho : T_\rho \mathcal{P}_+(X) \times T_\rho \mathcal{P}_+(X) \rightarrow \mathbb{R}$ を以下で定義する

$$g_\rho(\sigma_1, \sigma_2) := \int_X \nabla \Phi_1(x) \cdot \nabla \Phi_2(x) \rho(x) dx = \int_X \sigma_1 V_\rho^{-1}(\sigma_2) dx$$

ここで Φ_1, Φ_2 は $\sigma_1 = V(\Phi_1), \sigma_2 = V(\Phi_2)$ を満たすもの

$(\mathcal{P}_+(X), g)$ を density manifold と呼ぶ

density manifoldの勾配作用素

$\mathcal{P}_+(X)$ 上の汎関数 F が次のように定義されているとする

$$F(\rho) = \int_X U(\rho(x)) dx$$

この時Density manifold上の勾配 $\text{grad}_W F$ を定義する

勾配作用素の定義より $\text{grad}_W F$ は次の等式を満たす

$$g_\rho(\text{grad}_W F, \sigma) = \int_X U'(\rho(x)) \sigma(x) dx$$

上の等式より次式が成り立つことが知られている

$$\text{grad}_W F = V_\rho(U'(\rho))$$

Wasserstein勾配流

Def.4 Wasserstein勾配流[Villani,2009]

微分方程式

$$\frac{d\rho_t}{dt} = \text{grad}_W F(\rho_t)$$

の解をWasserstein勾配流と呼ぶ

- Wasserstein勾配流の例
 - F :エントロピー \Leftrightarrow 熱方程式
 - F :Tsallisエントロピー \Leftrightarrow 多孔媒質方程式
 - F :KLダイバージェンス \Leftrightarrow Fokker Plank方程式
- 応用では F としては何らかの損失関数や目的関数を想定している

目次

- Wasserstein距離から導かれる確率分布空間の幾何構造
- Wasserstein統計多様体と引き戻し計量
- グラフ上のWasserstein統計多様体
- まとめと今後の課題

Wasserstein統計多様体

- Wasserstein計量をパラメトリックモデルのパラメータ空間の計量に反映させたい！

X : サンプル空間, Θ : パラメータ空間(有限次元リーマン多様体)

$\rho(\cdot, \theta)$: θ により定まる確率分布, $\langle \cdot, \cdot \rangle_\theta$: Θ での内積

に対してWasserstein計量をパラメータ空間に引き戻せばよい
すなわち引き戻し計量を考える

$$g_\theta(\xi, \eta) := g_{\rho(\cdot, \theta)}(d_\theta \rho(\cdot, \xi), d_\theta \rho(\cdot, \eta))$$

ただし $d_\theta \rho(\cdot, \xi) = \langle \nabla_\theta \rho(\cdot, \theta), \xi \rangle$, $d_\theta \rho(\cdot, \eta) = \langle \nabla_\theta \rho(\cdot, \eta), \eta \rangle$

パラメータ空間のWasserstein計量テンソル

Def.5 パラメータ空間のWasserstein計量テンソル[Li,2018]

いくつかの仮定の下

$$\xi, \eta \in T_{\theta}\Theta, \quad g_{\theta}(\xi, \eta) := \int_X \nabla \Phi_{\xi}(x) \cdot \nabla \Phi_{\eta}(x) \rho(x) dx$$

Φ_{ξ}, Φ_{η} は $\langle \nabla_{\theta} \rho(\cdot, \theta), \xi \rangle_{\theta} = V_{\rho}(\Phi_{\xi}), \langle \nabla_{\theta} \rho(\cdot, \theta), \eta \rangle_{\theta} = V_{\rho}(\Phi_{\eta})$ を満たす

Thm.3[Li,2018]

$$A_{ij}(\theta) = \int_X \partial_{\theta_i} \rho(x, \theta) V^{-1}(\partial_{\theta_j} \rho(x, \theta)) dx \text{ とする}$$

Θ の計量テンソルを G_{θ} とし $G_W = G_{\theta}^T A(\theta) G_{\theta}$ とすると

$$g_{\theta}(\xi, \eta) = \xi^T G_W \eta.$$

Wasserstein計量テンソルとフィッシャー計量

Thm.4 $X = \mathbb{R}$ のときのWasserstein計量テンソル[Li,2018]

Θ をユークリッド空間とする : i.e. $G_\theta = I$

$X = \mathbb{R}$ のとき

$$G_W(\theta) = \int_{\mathbb{R}} \frac{1}{\rho(x, \theta)} \nabla_{\theta} F(x, \theta)^T \nabla_{\theta} F(x, \theta) dx .$$

ここで $F(x, \theta)$ は $\rho(x, \theta)$ の確率分布関数

参考: フィッシャー計量

$$G_{\text{Fisher}}(\theta) = \int_{\mathbb{R}} \frac{1}{\rho(x, \theta)} \nabla_{\theta} \rho(x, \theta)^T \nabla_{\theta} \rho(x, \theta) dx .$$

フィッシャー計量は確率密度関数によって定まるのに対して
Wasserstein計量テンソルは確率分布関数によって定まる！

証明

$X = \mathbb{R}$ のとき $g_{\theta}(\xi, \eta) = \int_X \Phi'_{\xi}(x) \cdot \Phi'_{\eta}(x) \rho(x) dx$ となり

Φ_{ξ} は $\langle \nabla \rho(\cdot, \theta), \xi \rangle_{\theta} = V_{\rho}(\Phi_{\xi}) = \left(\rho \Phi'_{\xi} \right)'$ を満たすから両辺積分し

$$\int_{-\infty}^y \langle \nabla_{\theta} \rho(x, \theta), \xi \rangle dx = \rho(x, \theta) \Phi'_{\xi}(y)$$

積分と微分の順序交換をして

$$\Phi'_{\xi}(x) = \frac{1}{\rho(x, \theta)} \langle \nabla_{\theta} F(x, \theta), \xi \rangle .$$

Φ'_{η} も同様に求まる. $g_{\theta}(\xi, \eta)$ に代入して定理を得る

パラメータ空間での勾配

- (Θ, g_θ) を Wasserstein 統計多様体と呼ぶ
- パラメータの関数 $R(\theta)$ の Wasserstein 統計多様体上での勾配 $\text{grad}_{g_\theta} R(\theta)$ は勾配作用素の定義より

$$\forall \xi \in T_\theta \Theta, \text{grad}_{g_\theta} R(\theta) G_W(\theta) \xi = \nabla_\theta R(\theta) \cdot \xi$$

が成り立ち

$$\text{grad}_{g_\theta} R(\theta) = G_W(\theta)^{-1} \nabla_\theta R(\theta)$$

となる

Wasserstein自然勾配法

Def.6 パラメータ空間でのWasserstein勾配流[Li,2018]

微分方程式

$$\frac{d\theta}{dt} = -G_W(\theta)^{-1} \nabla_{\theta} R(\theta)$$

の解をパラメータ空間でのWasserstein勾配流と呼ぶ

パラメータを逐次更新して勾配流を求めることができる

$$\theta^{n+1} = \theta^n - \gamma G_W(\theta^n) \nabla_{\theta} R(\theta^n), \quad \tau : \text{学習率}$$

上式で G_W ではなく G_{Fisher} を用いたものは自然勾配法と呼ばれる
それと対比してこれを**Wasserstein自然勾配法**と呼ぶ

実験1:混合正規分布

$X = \mathbb{R}$ においてパラメータ θ の混合正規分布のモデルを考える

$$\rho(x, \theta) = \frac{a}{\sigma_1 \sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu_1)^2}{2\sigma_1^2} \right\} + \frac{1-a}{\sigma_2 \sqrt{2\pi}} \exp \left\{ -\frac{(x - \mu_2)^2}{2\sigma_2^2} \right\}$$

$$\theta = (a, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

サンプルの経験分布を $\frac{1}{N} \sum_{i=1}^N \delta_{x_i}(\cdot)$ とおき次の最適化問題を考える

$$\min_{\theta} d \left(\rho(\cdot, \theta), \frac{1}{N} \sum_{i=1}^N \delta_{x_i}(\cdot) \right)$$

dとしてKLダイバージェンスを用いれば最尤推定法となる

しかし色々な問題がある

実験1:混合正規分布[Li,2018]

- d としてWasserstein距離を用いWasserstein自然勾配法で最適な θ を求める

$$\frac{1}{2}W_2^2\left(\rho(\cdot, \theta), \frac{1}{N}\sum_{i=1}^N \delta_{x_i}(\cdot)\right)$$

Thm.5

カントロヴィッチ双対性より次が成り立つ

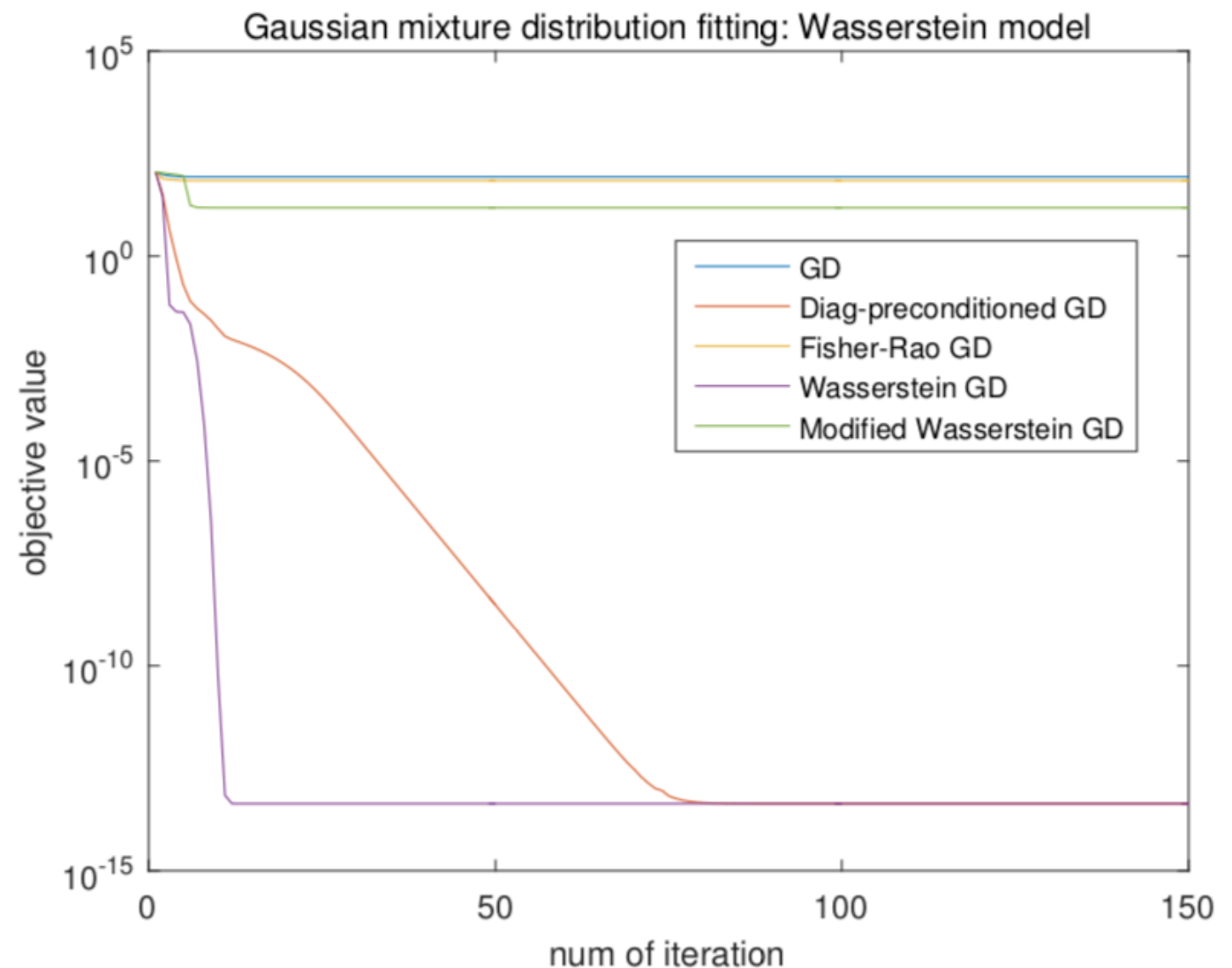
$$\frac{1}{2}\nabla_{\theta}W_2^2 = \int_{\mathbb{R}} \phi(x) \nabla_{\theta}\rho(x, \theta)dx$$

$$\text{ここで } \phi(x) = \int_{-\infty}^x (y - F_{\text{em}}^{-1}(F(x))) \nabla_{\theta}\rho(x, \theta)dx$$

F_{em} : 経験分布の確率分布関数, $F : \rho(x, \theta)$ の確率分布関数

実験1:混合正規分布[Li,2018]

- $N=300$ での実験結果
- Wasserstein自然勾配法が収束までのイテレーションも d の値も最も小さくなっていることがわかる



目次

- Wasserstein距離から導かれる確率分布空間の幾何構造
- Wasserstein統計多様体と引き戻し計量
- グラフ上のWasserstein統計多様体
- まとめと今後の課題

グラフへの拡張

- これまでは X はユークリッド空間とリーマン多様体を考えた
 - X を重み付きグラフに変え、楕円型作用素 V_ρ をグラフラプラシアンに置き換えることでほぼ同様の議論が重み付きグラフとその上のパラメトリックな確率分布に対してできる
- グラフ構造が扱えるようになる&有限次元なので計算がしやすい

重み付きグラフラプラシアン

Def.7 重み付きグラフラプラシアン

$G = (I, E, w), I = \{1, 2, \dots, n\} : n$ 頂点の重み付きグラフ

$$L(a) = D^\top \Lambda(a) D, \quad a = (a_i)_{i=1}^n \in \mathbb{R}^n$$

を重み付きグラフラプラシアンと呼ぶ

$$D_{(i,j) \in E, k \in V} = \begin{cases} \sqrt{\omega_{ij}} & \text{if } i = k, i > j, \\ -\sqrt{\omega_{ij}} & \text{if } j = k, i > j, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Lambda(a)_{(i,j) \in E, (k,l) \in E} = \begin{cases} \frac{a_i + a_j}{2} & \text{if } (i, j) = (k, l) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

グラフ上のWasserstein計量テンソル

連続のときと同様に

$$\mathcal{P}_+(G) = \left\{ \rho \in \mathbb{R}^n : \rho_i > 0, \sum_{i=1}^n \rho_i = 1 \right\}$$

$$T_\rho \mathcal{P}_+(G) = \left\{ \sigma \in \mathbb{R}^n; \sum_{i=1}^n \sigma_i = 0 \right\}$$

と定義する

Def.8 グラフ上のWasserstein計量テンソル[Li,2018]

$$\forall \sigma, \tilde{\sigma} \in T_\rho \mathcal{P}_+(G), \quad g_\rho^W(\sigma, \tilde{\sigma}) := \sigma^T L(p)^\dagger \tilde{\sigma}$$

ここで $L(p)^\dagger$ は $L(p)$ の擬逆行列

グラフ上のWasserstein統計多様体

G上のパラメトリックモデル $p(\theta) \in \mathbb{R}^n, \theta \in \Theta \subset \mathbb{R}^d$ を考える
 $d < n$ とする, 連続のときと同様に引き戻し計量を考える

Def.9 グラフ上のWasserstein計量テンソル[Li,2018]

$$\forall a, b \in T_\rho \Theta, g_\theta^W(a, b) := a^T J_\theta p(\theta)^T L(p(\theta))^\dagger J_\theta p(\theta) b.$$

$$J_\theta p(\theta)_{ij} = \frac{\partial p_i(\theta)}{\partial \theta_j}, \text{ ヤコビ行列}$$

$$G_W(\theta) := J_\theta p(\theta)^T L(p(\theta))^\dagger J_\theta p(\theta) \text{ と略記する}$$

グラフ上のWasserstein勾配流

Def.10 パラメータ空間でのWasserstein勾配流[Li,2018]

パラメータの関数 $R(\theta)$ に対して微分方程式

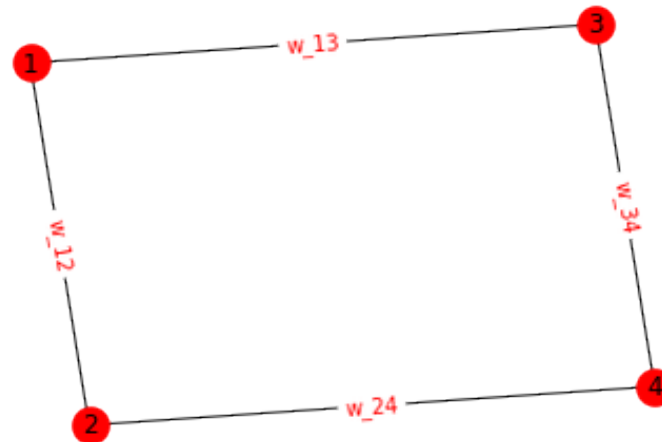
$$\frac{d\theta}{dt} = -G_W(\theta)^{-1} \nabla_{\theta} R(\theta)$$

の解をパラメータ空間でのWasserstein勾配流と呼ぶ

Wasserstein自然勾配法も同様に定義できる

実験2:独立モデル[Li,2018]

次のような4頂点の重み付きグラフ



と二次元のパラメータ $a = (a_1, a_2)$ で定まる確率分布

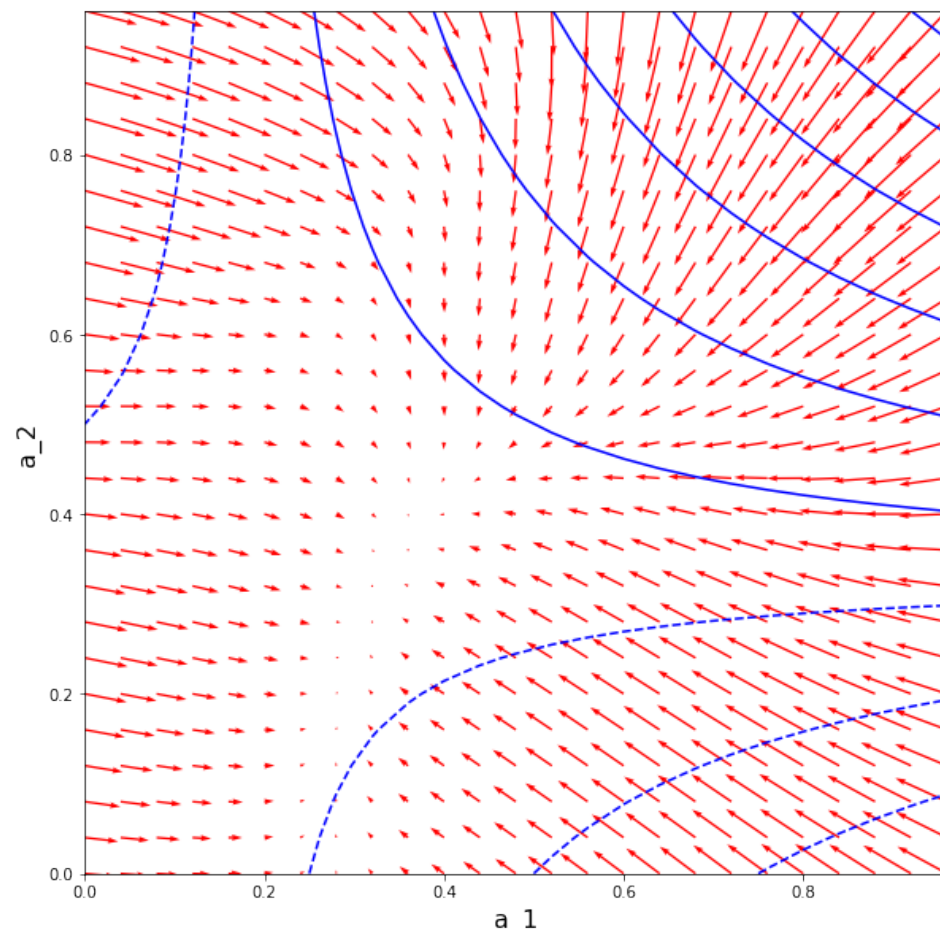
$$p(a)(i) = \begin{cases} (1 - a_1)(1 - a_2) & i = 1 \\ (1 - a_1)a_2 & i = 2 \\ a_1(1 - a_2) & i = 3 \\ a_1a_2 & i = 4 \end{cases}, \quad 0 \leq a_1, a_2 \leq 1$$

を考える

実験2:独立モデル[Li,2018]

次のような関数を考えパラメータ空間での
Wasserstein勾配を計算すると以下のようなになる

$$F(p(a)) = \sum_{i \in I} f(i)p(a)(i) = -4a_1 - 2a_2 + 12a_1a_2 \quad f(i) = \begin{cases} 0 & i = 1 \\ -2 & i = 2 \\ -4 & i = 3 \\ 6 & i = 4 \end{cases}$$



青い線は関数の等高線
赤い矢印はWasserstein計量テンソル
での勾配ベクトルを表している

実験3:KLダイバージェンス勾配流[Li,2018]

- 汎関数としてKLダイバージェンスを考えれば最尤推定になる

$$\operatorname{argmin} \operatorname{KL}(q \mid p(\theta)) = \operatorname{argmin} \sum_{x \in I} q_x \log \frac{q_x}{p_x(\theta)}$$

- 対応する勾配流は次のように計算される

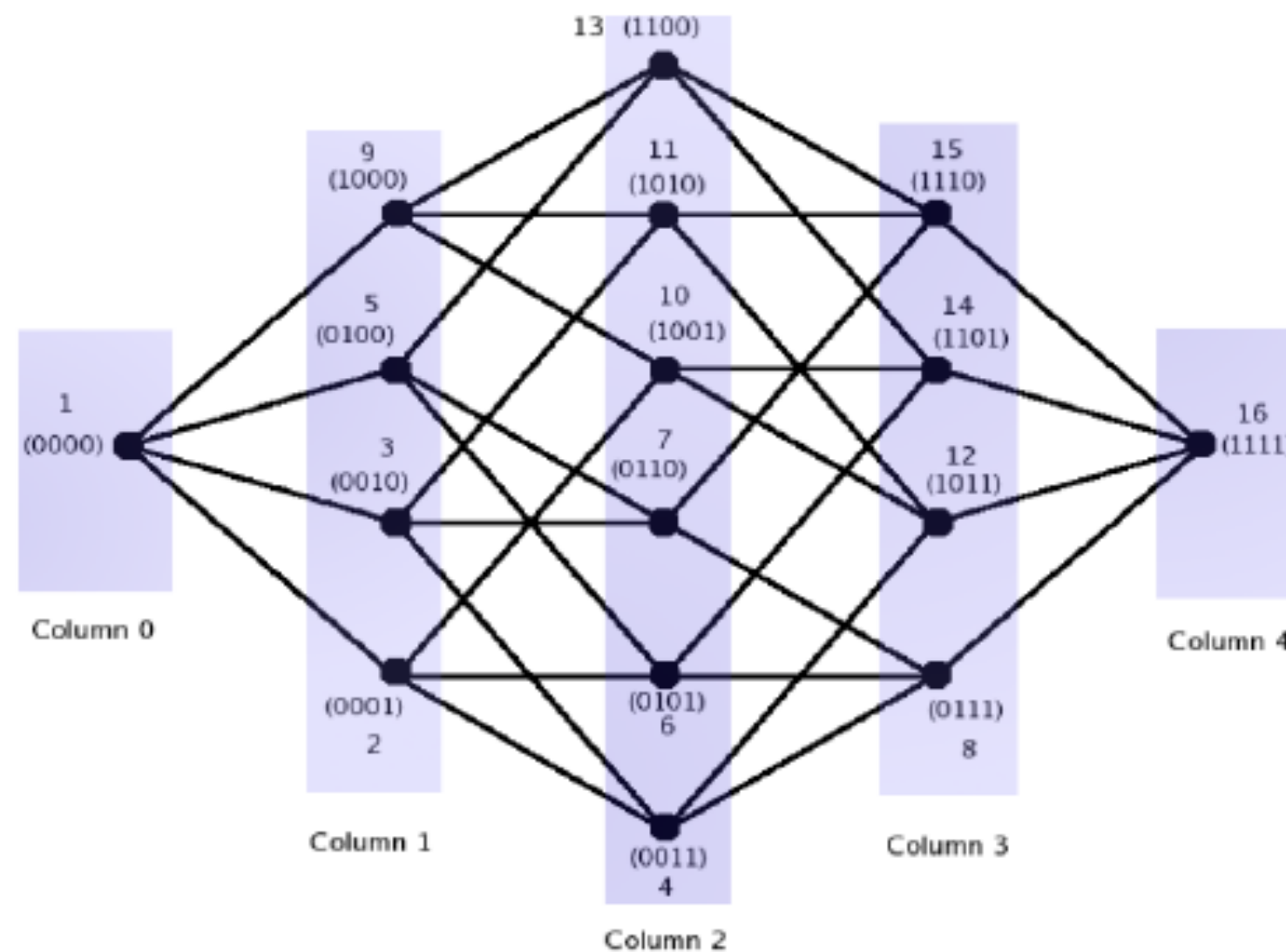
$$\frac{d\theta}{dt} = (J_{\theta} p(\theta)^T L(p(\theta))^{\dagger} J_{\theta} p(\theta)^T \left(\frac{q}{p(\theta)} \right)$$

q : ターゲットの分布、適当に定める

- Binary cube graph**を仮定してn個の0-1変数の相互作用モデルで最尤推定を行う

binary cube graph

- 頂点はn桁のビットで表される数全て
- ハミング距離が1の頂点のみを辺で結ぶ(重みは均一)



0-1変数の相互作用モデル

次のような指数型分布族を考える

$$S \subset \mathfrak{P}(I) \quad p_x(\theta) = \frac{1}{Z(\theta)} \exp \left(\sum_{\lambda \in S} \theta_\lambda \phi_\lambda(x) \right), \quad x \in \{0,1\}^n$$

λ : 相互作用するグループ

ϕ_λ の例として主に次の二つがある

$$\sigma_\lambda(x) = \prod_{i \in \lambda} (-1)^{x_i}, \quad x \in \{0,1\}^n$$

$$\pi_\lambda(x) = \prod_{i \in \lambda} x_i, \quad x \in \{0,1\}^n \quad \text{論理積}$$

$S = \{\lambda \subset (1,2,\dots,n) \mid |\lambda| \leq k\}$ のとき k-interaction model という
 k が大きければ $p_x(\theta)$ の表現力は大きくなる

KLダイバージェンスの最小化を勾配降下法で計算する

$$\theta_{t+1} = \theta_t - \gamma_t G(\theta)^{-1} \nabla \text{KL}(q | p(\theta_t))$$

計量Gは**Euclid, Fisher, Wasserstein**で比較する

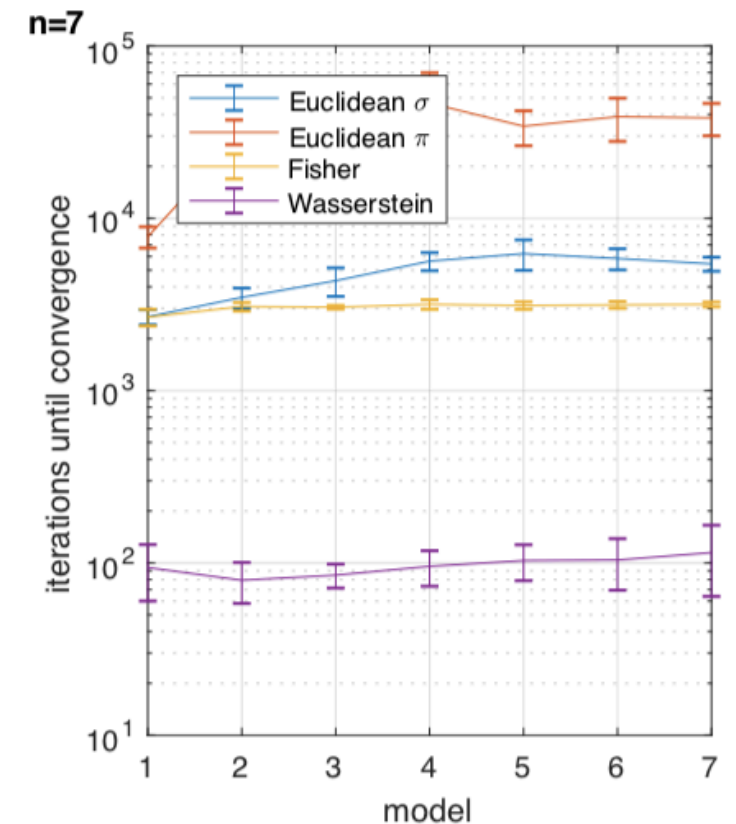
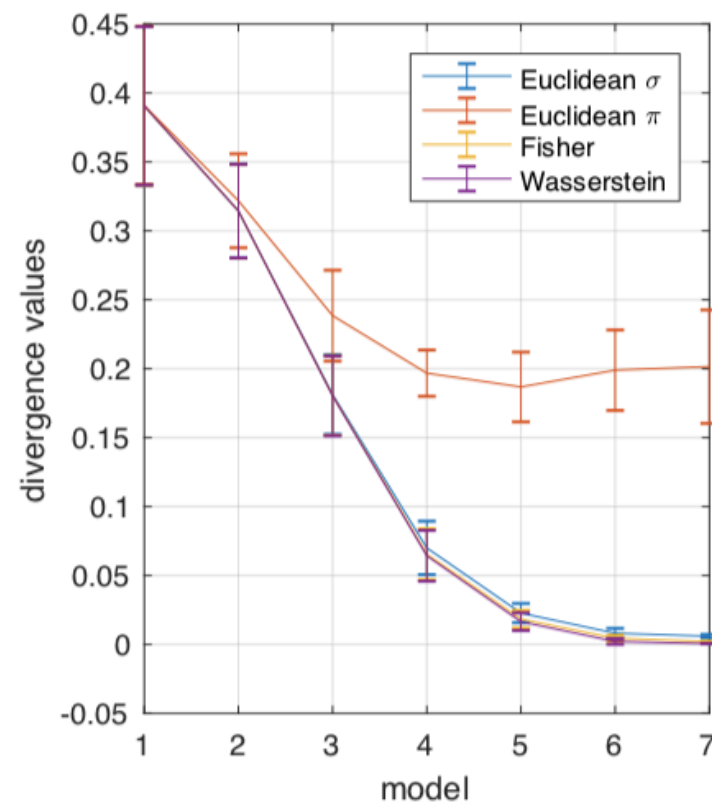
ここで γ_t は学習率で学習途中でKLダイバージェンスが減少しなければ0.75倍する(iterative method)

目的分布qはディリクレ分布から適当に作成する

k=1から7までのモデルでの学習の様子を比較

数値例3:KLダイバージェンス勾配流[Li,2018]

- k が大きいほどモデルの表現力が上がるためKLダイバージェンスは減少するが通常の勾配法では学習に失敗している
- Wasserstein自然勾配法は学習に成功しかつ収束までの反復回数も最も小さい
- Binary cube graphの構造が学習に寄与しているから？



横軸: k の値

目次

- Wasserstein距離から導かれる確率分布空間の幾何構造
- Wasserstein統計多様体と引き戻し計量
- グラフ上のWasserstein統計多様体
- まとめと今後の課題

今後の課題

- 確率分布族の間のWasserstein測地線が閉じているかどうか
 - 閉じていればパラメータ空間での測地線の長さ とWasserstein距離が一致する
- 正規分布間のWasserstein測地線の各点は常に正規分布である
 - (数値実験を見る限り) 混合正規分布はそうではない
 - (数値実験を見る限り) ガンマ分布はそうかもしれない？
- 指数型分布族はどうか？
- パラメトリックモデルが特異モデルの場合Wasserstein計量テンソルは擬リーマン計量となるがその時の振る舞いはどうなるか？

今後の課題

- グラフ上でのWasserstein計量テンソルとフィッシャー計量の関係
 - Wasserstein計量テンソルがフィッシャー計量と一致するようなグラフ構造が存在するか？
- グラフの重みの決め方
 - はじめに固定するか、学習の途中でグラフの重みを変えるか
- カントロヴィッチ双対性と情報幾何学における規範ダイバージェンスの関係 **[Wong,2109]**
- 実データを用いた応用：画像処理、敵対的生成モデル、マルチラベリング、ドメイン適合

参考文献

- Y.Chen and W.Li.Natural gradient in Wasserstein statistical manifold.arXiv1805.08380,2018.
- W.Li and G.Montufar. Natural gradient via optimal transport. arXiv: 1803.07033,2018.
- C. Villani. Optimal Transport: Old and New. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.
- T.L.Wong and J.Yang.Optimal transport and information geometry.arXiv:1906.00030,2019.