機械学習と凸共役の交わり

2023年1月6日（金）

最適化手法とアルゴリズム研究部会（SOMA 第7回）

包 含（京都大学）

自己紹介

- 名前: 包 含（🇯🇵 つつみ ふくむ, 🇨🇳 BAO Han）
- 経歴
  - 2017年4月 - 2019年3月: 東京大学大学院 情報理工学系研究科 修士課程
  - 2019年4月 - 2022年3月: 東京大学大学院 情報理工学系研究科 博士課程
  - 2022年4月 - 現在　　: 京都大学 白眉センター 特定助教
- 研究の興味
  - 機械学習におけるロバスト性、転移学習、および**損失関数**の設計
  - （関連して**凸解析**や情報幾何）
- 好きなこと
  - 旅行（本を読みながら電車旅など）
  - アルコール（クラフトビールとウイスキー）

自己紹介

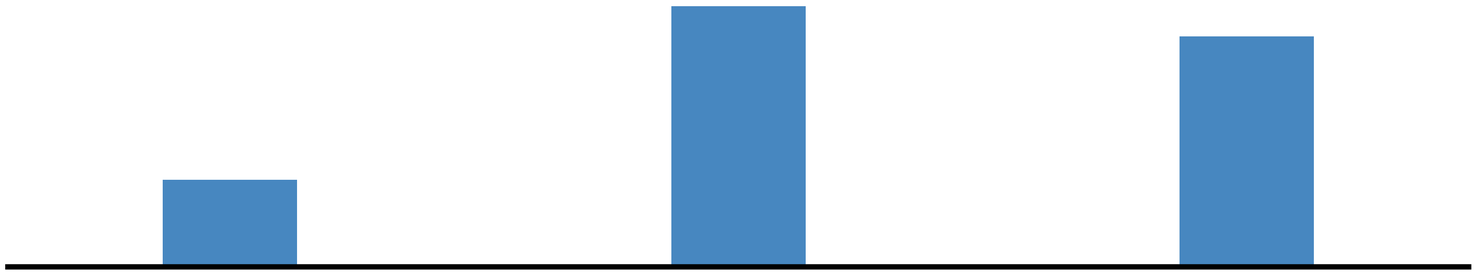
- 名前: 包 含（🇯🇵 つつみ ふくむ, 🇨🇳 BAO Han）
- 経歴
  - 2017年4月 - 2019年3月: 東京大学大学院 情報理工学系研究科 修士課程
  - 2019年4月 - 2022年3月: 東京大学大学院 情報理工学系研究科 博士課程
  - 2022年4月 - 現在　　: 京都大学 白眉センター 特定助教
- 研究の興味
  - 機械学習におけるロバスト性、転移学習、および**損失関数**の設計
  - （関連して**凸解析**や情報幾何）
- 好きなこと
  - 旅行（本を読みながら電車旅など）
  - アルコール（クラフトビールとウイスキー）

Bandit Brewery (@Tronto) のスタウト

意思決定の枠組み

🤔

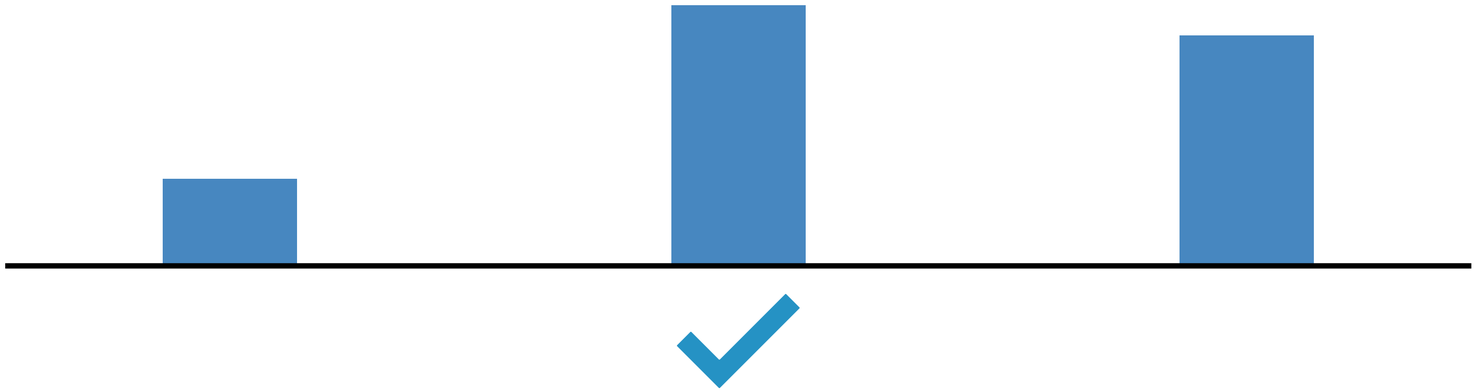
意思決定の枠組み



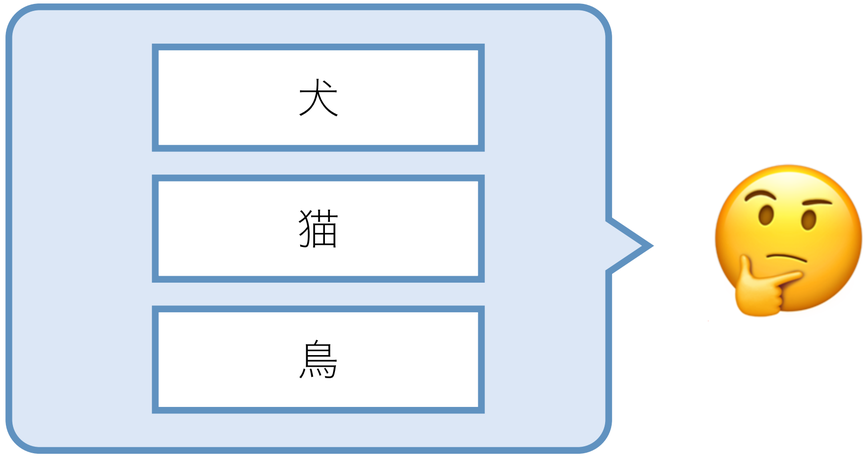
🤔

意思決定の枠組み

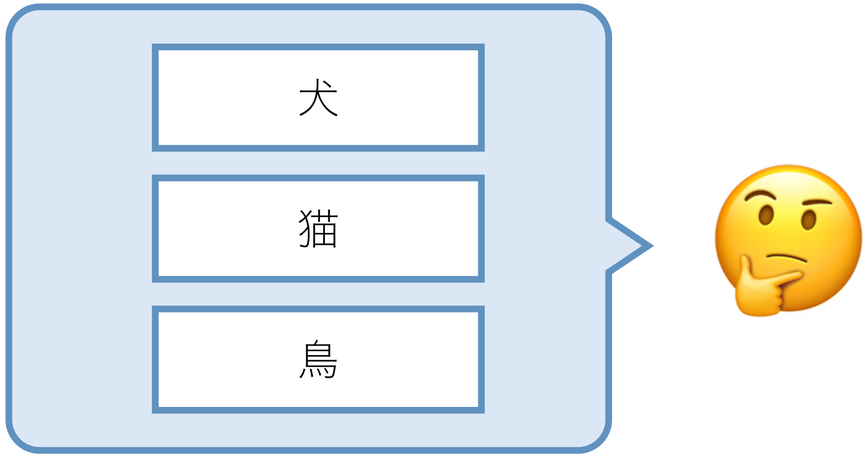
🤔



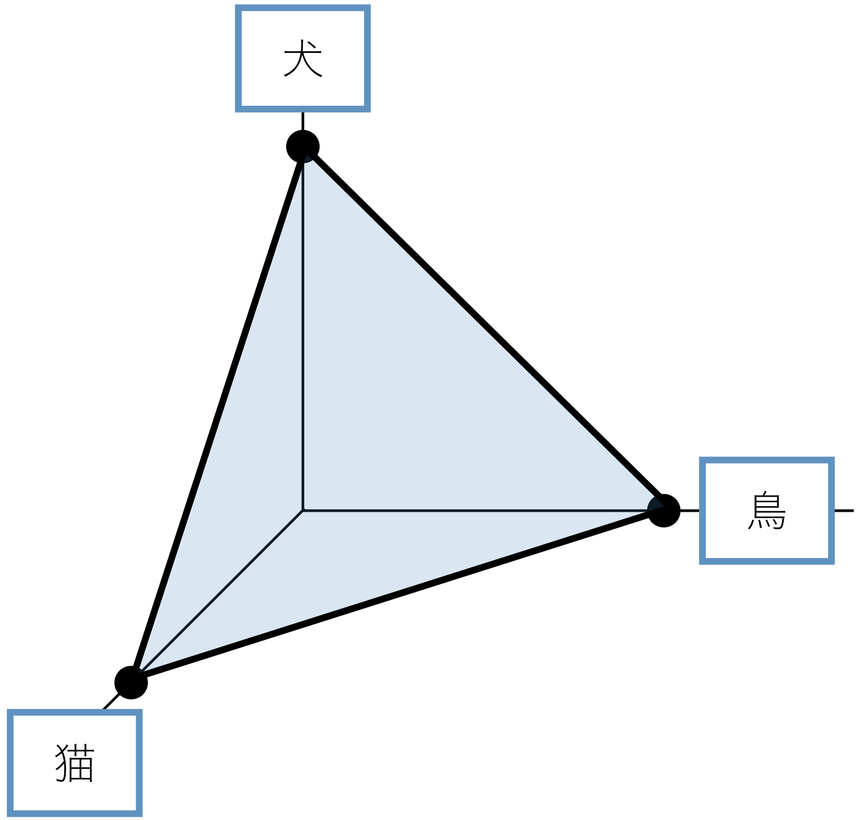
意思決定の枠組み（例: 分類問題）



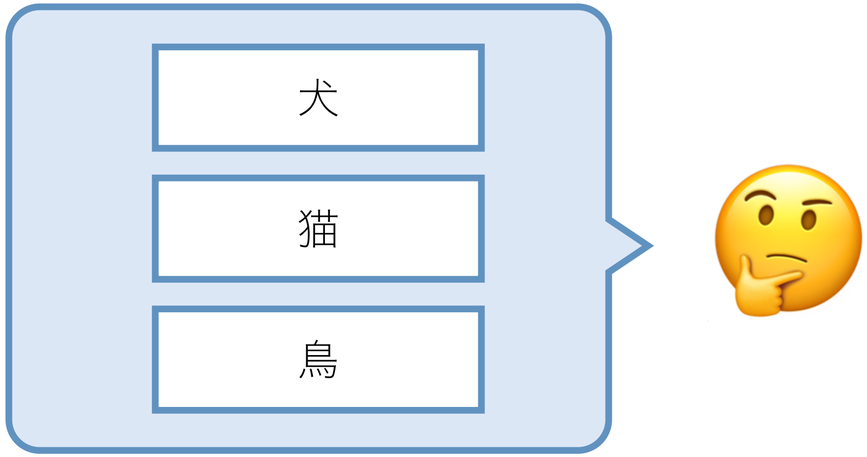
意思決定の枠組み（例: 分類問題）



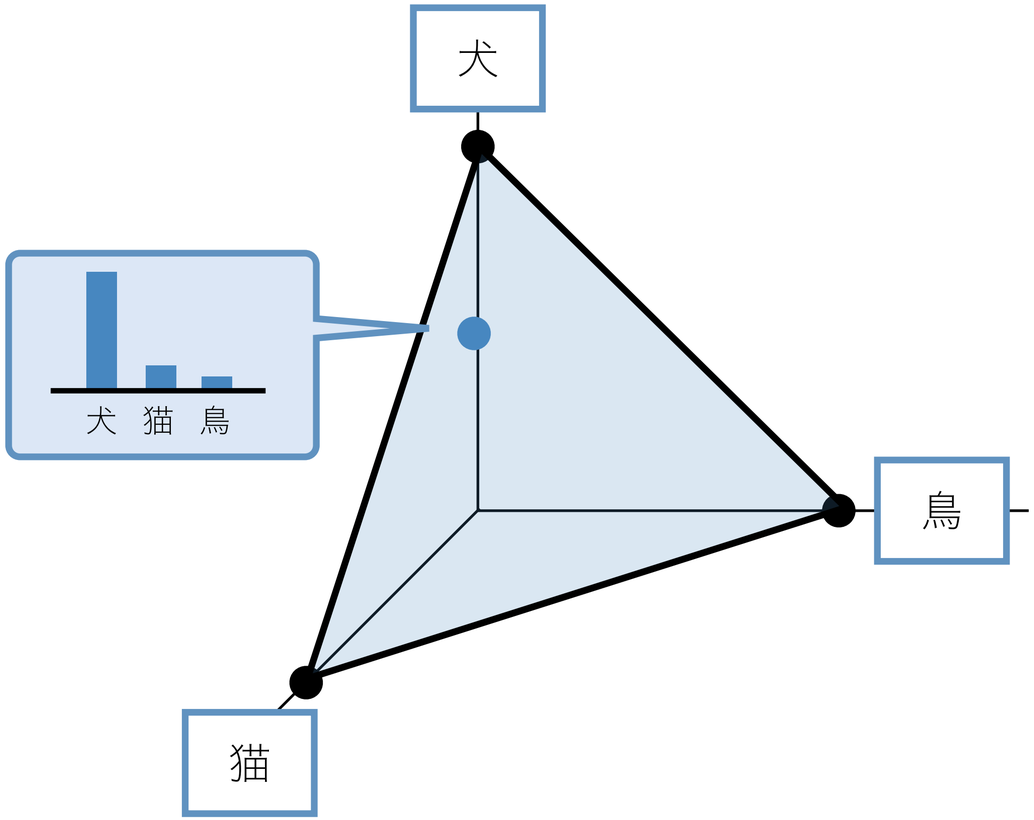
🤔



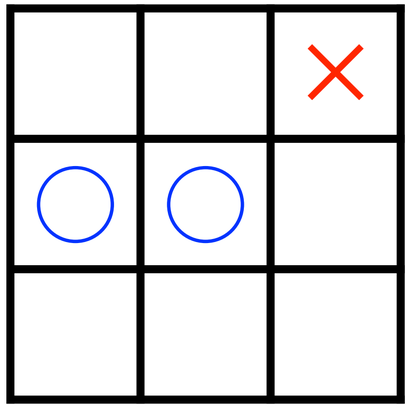
意思決定の枠組み（例: 分類問題）

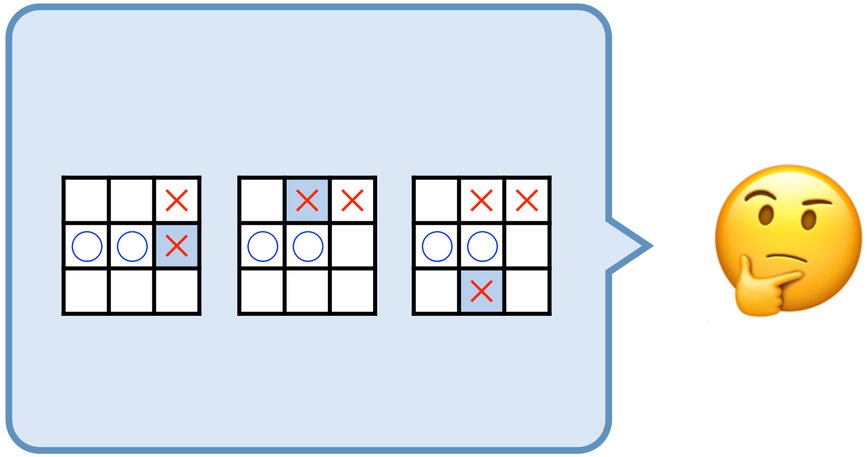


🤔

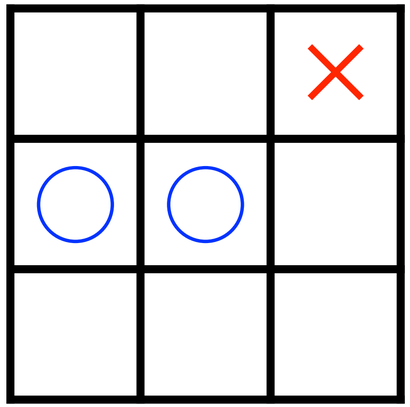


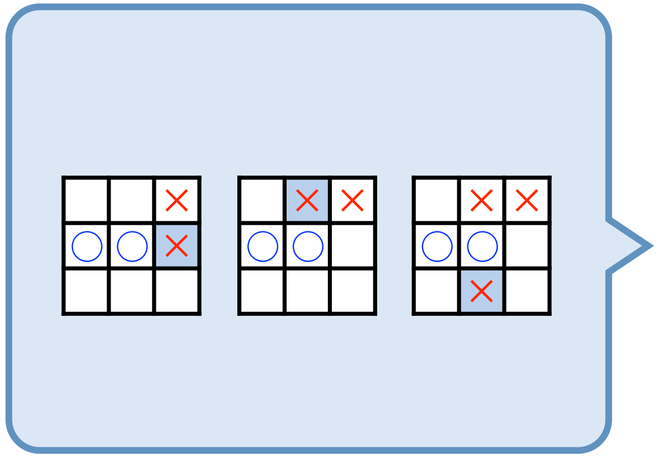
意思決定の枠組み（例: 強化学習）

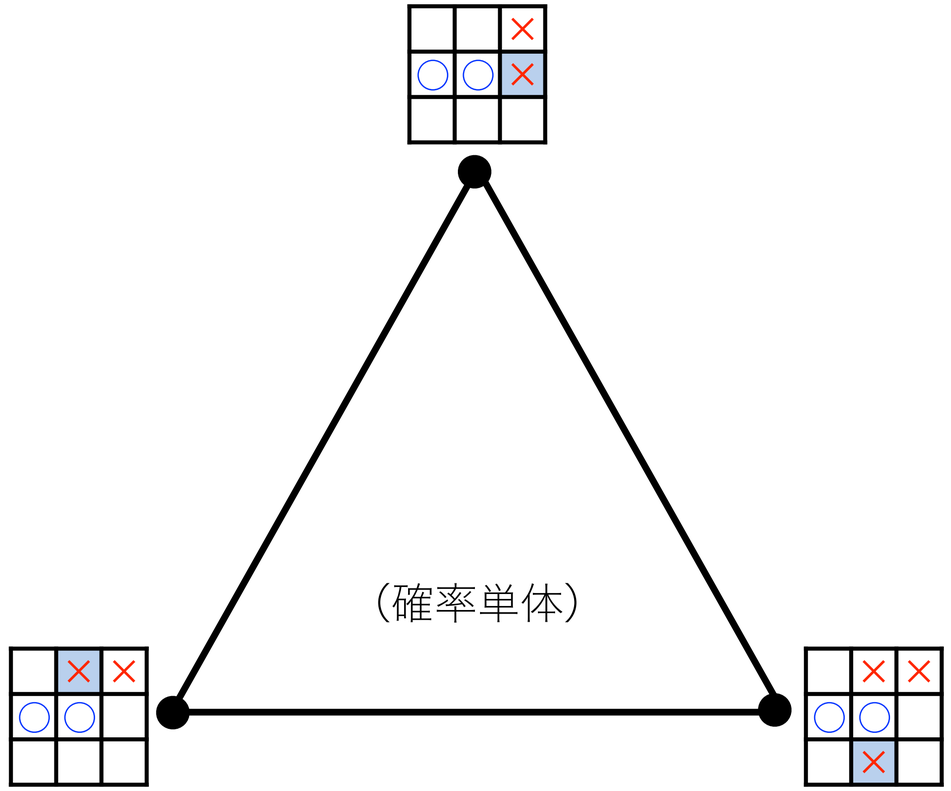




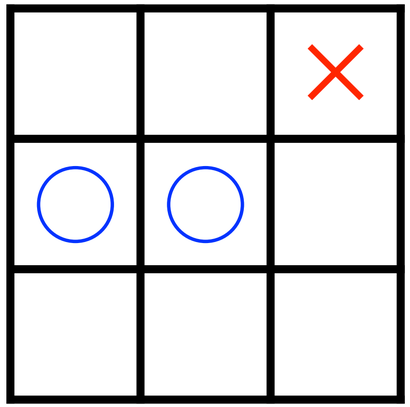
意思決定の枠組み（例: 強化学習）

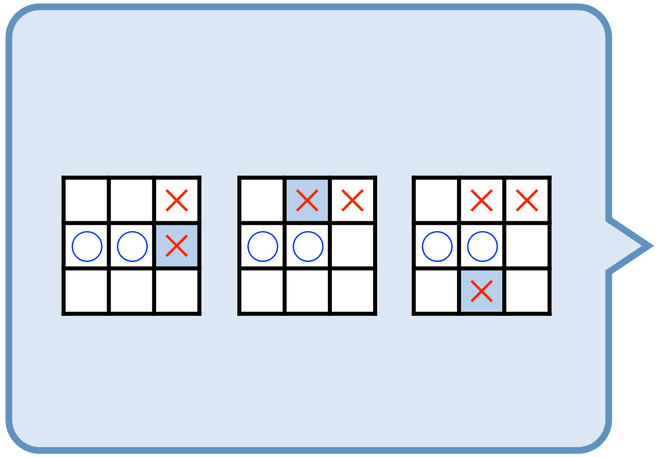


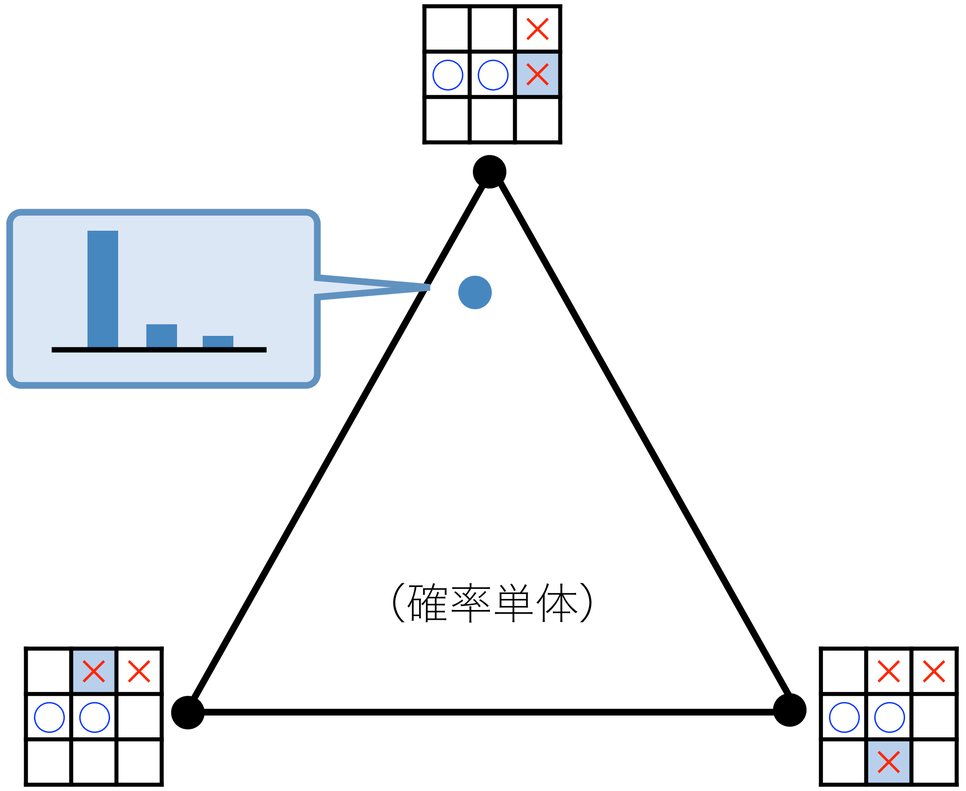




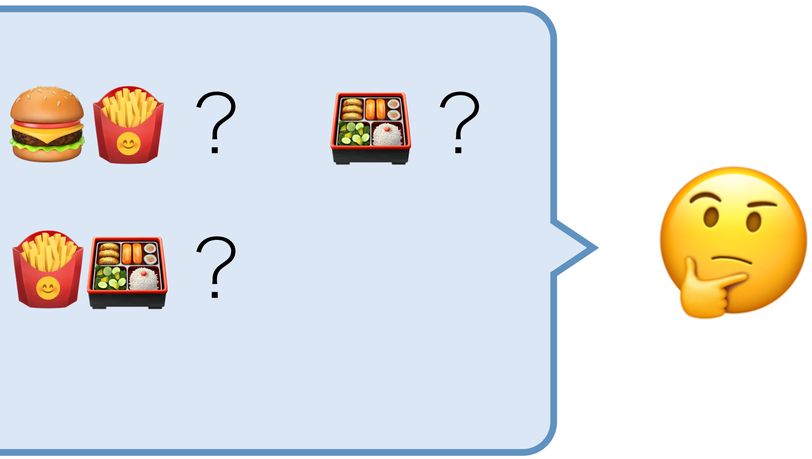
意思決定の枠組み（例: 強化学習）





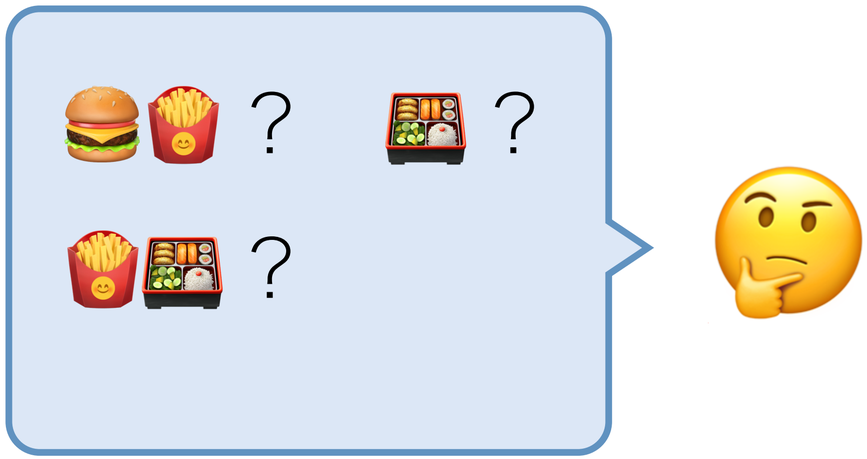


意思決定の枠組み（例: マルチラベル予測）

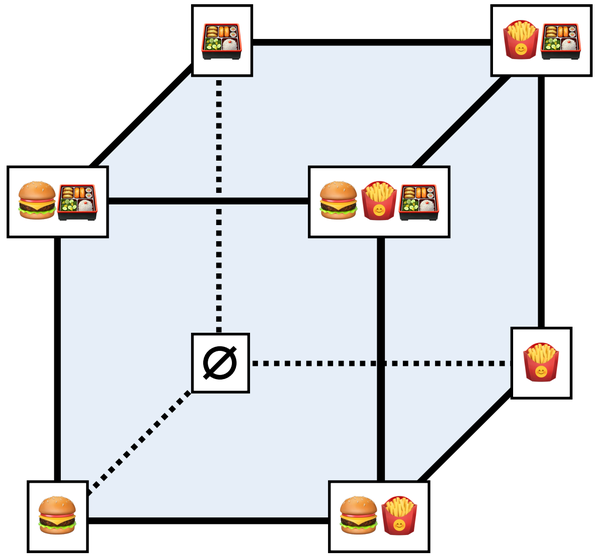


意思決定の枠組み（例: マルチラベル予測）

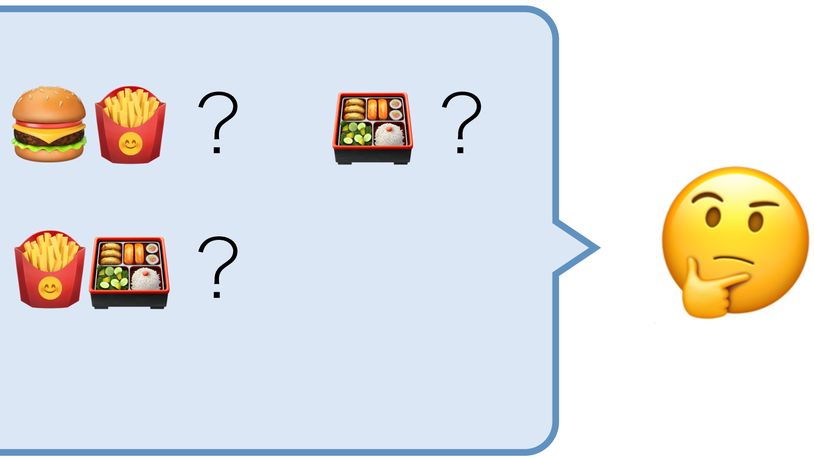




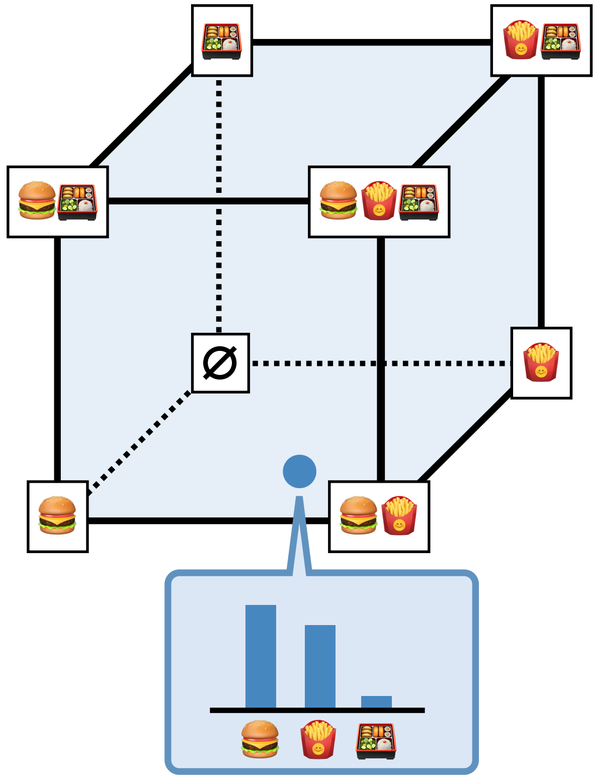
（単位立方体）



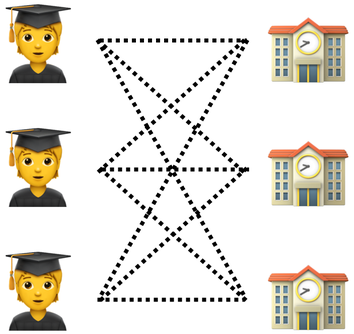
意思決定の枠組み（例: マルチラベル予測）

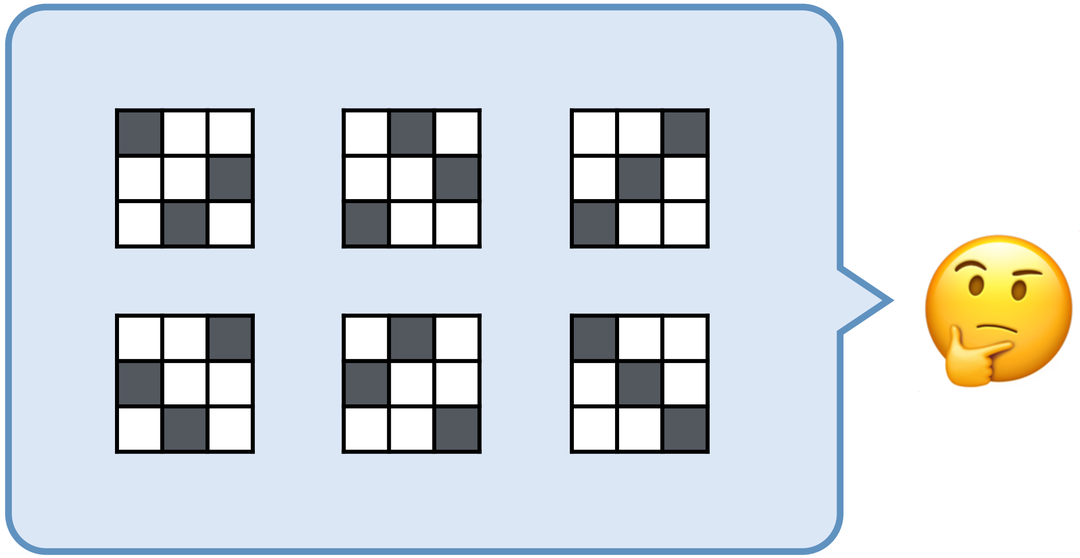


（単位立方体）

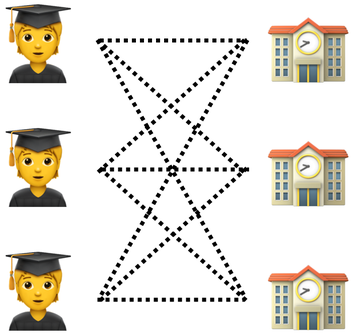


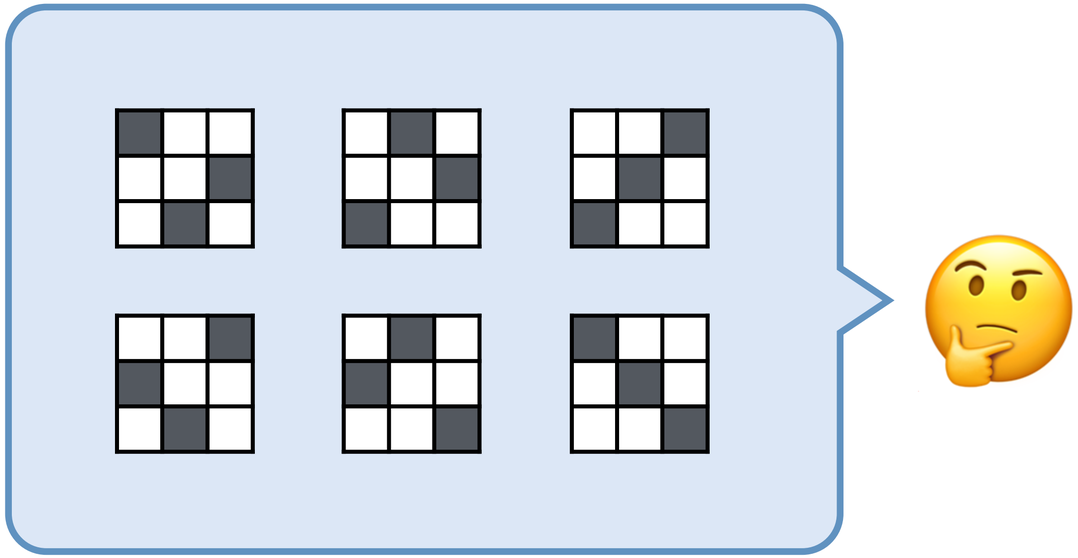
意思決定の枠組み（例: マッチング・最適輸送）

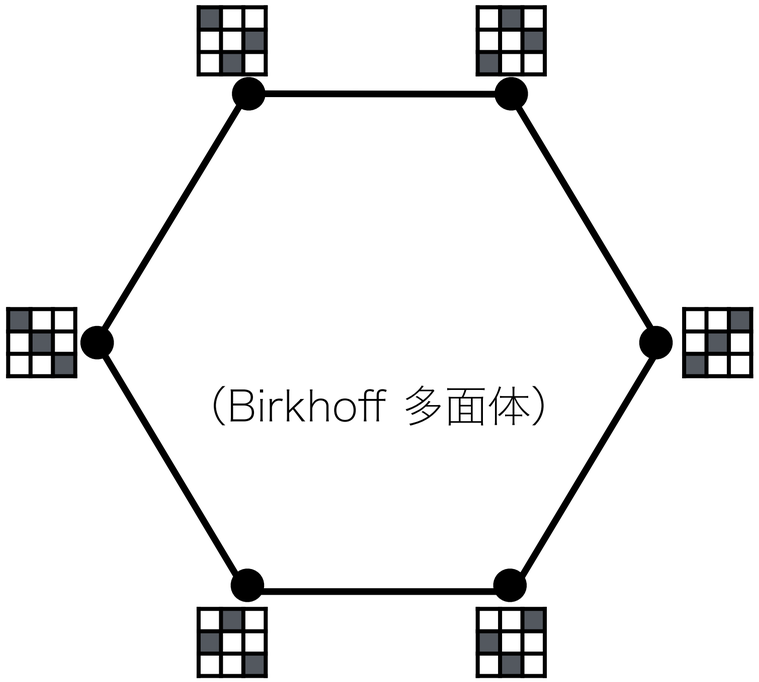




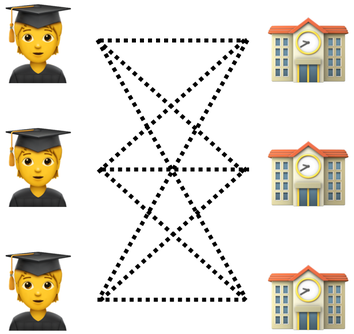
意思決定の枠組み（例: マッチング・最適輸送）

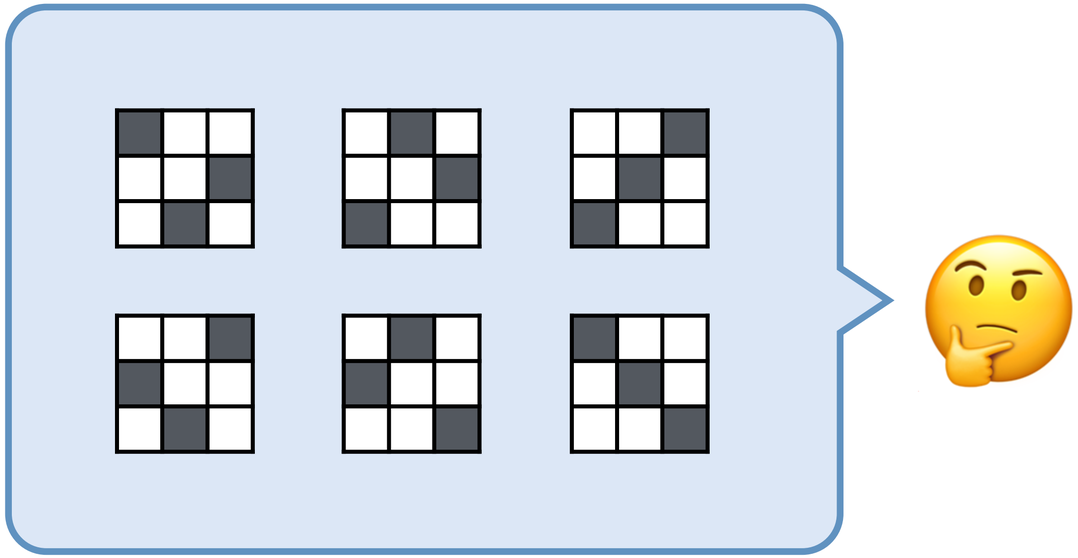


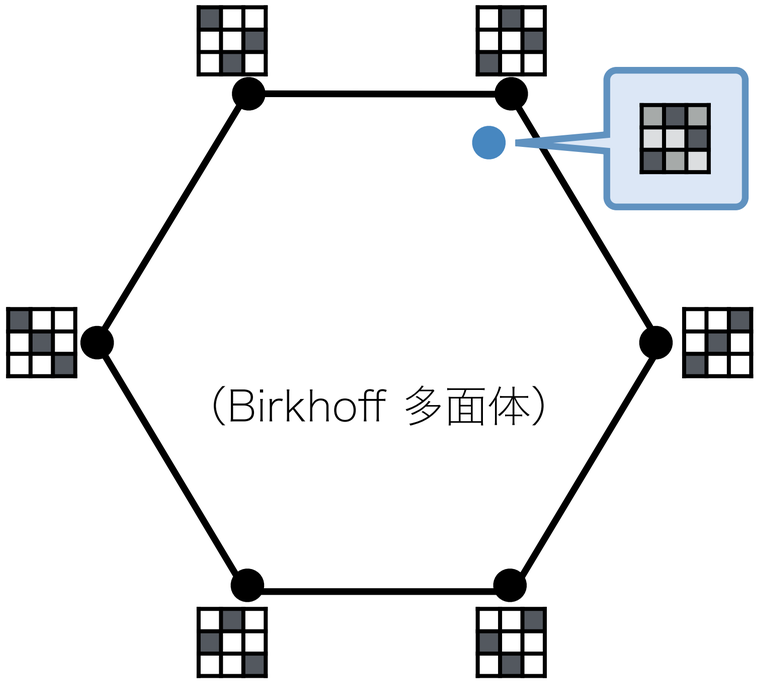




意思決定の枠組み（例: マッチング・最適輸送）

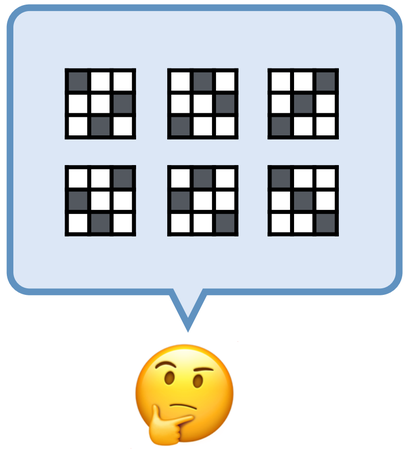




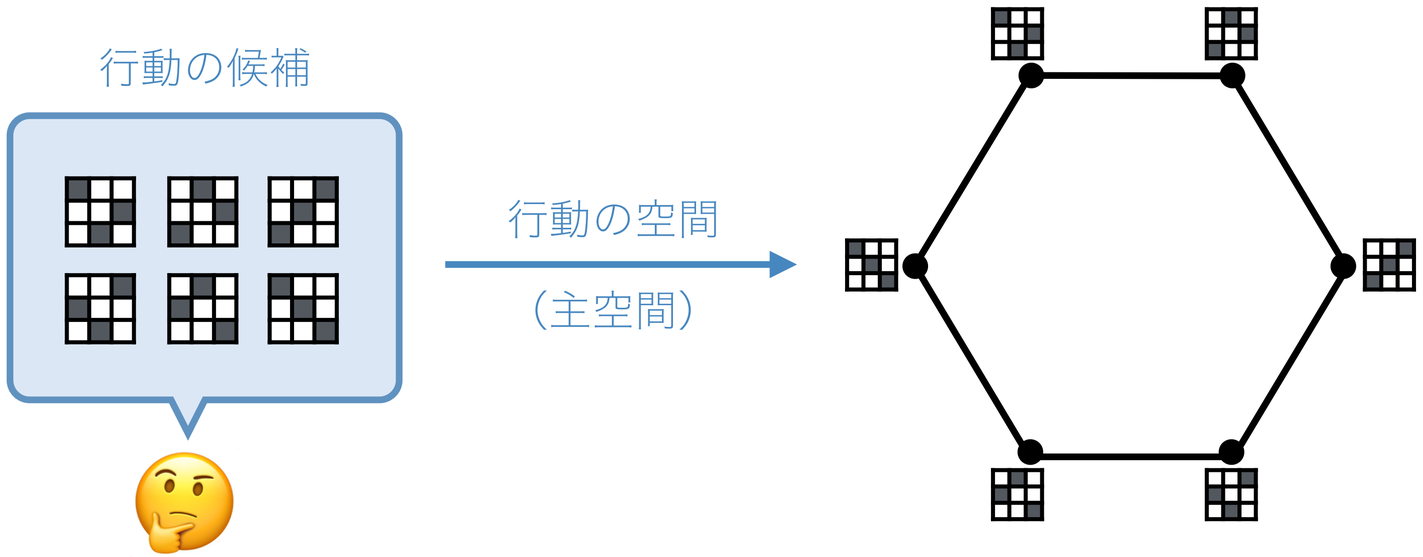


意思決定の枠組み: 機械学習の視点から

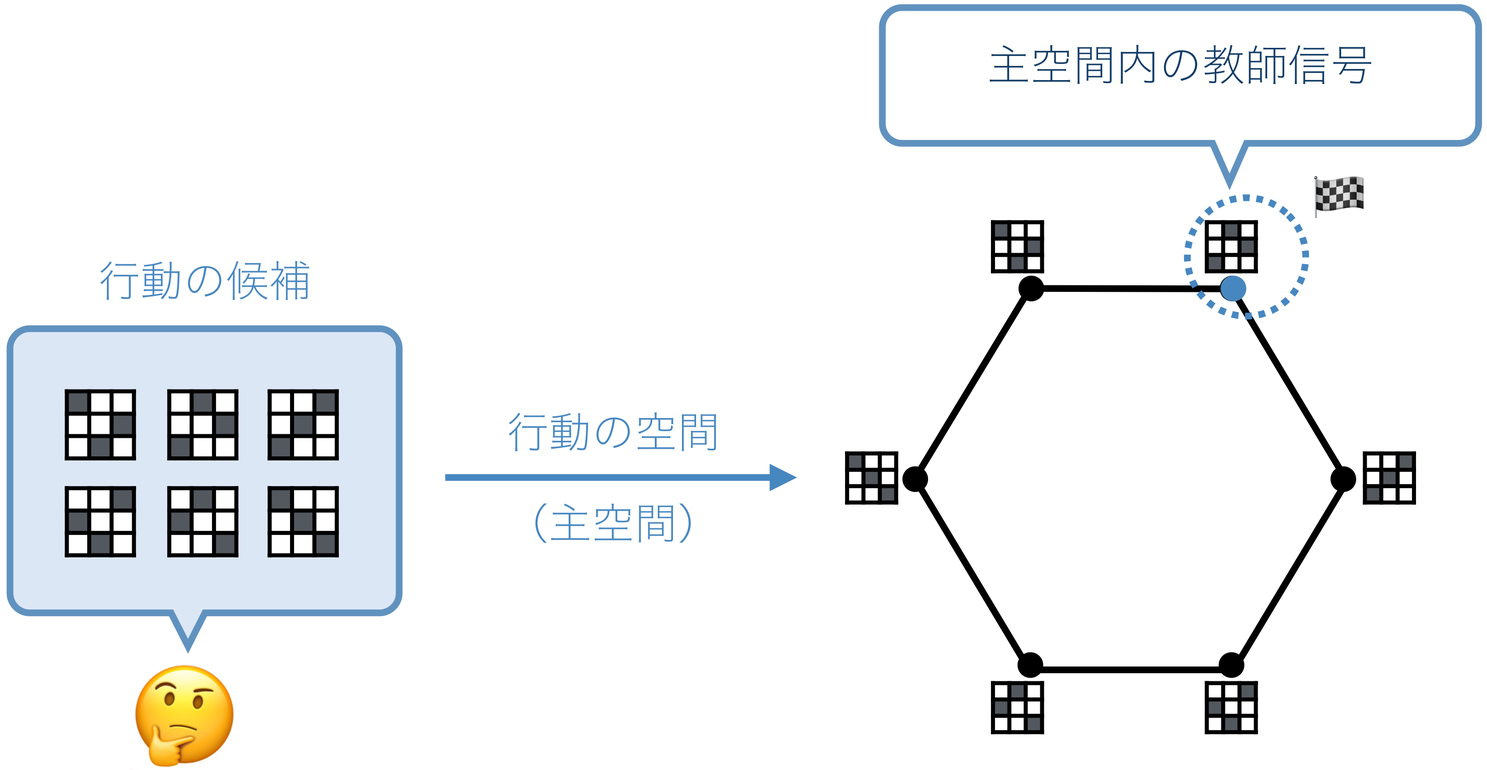
行動の候補



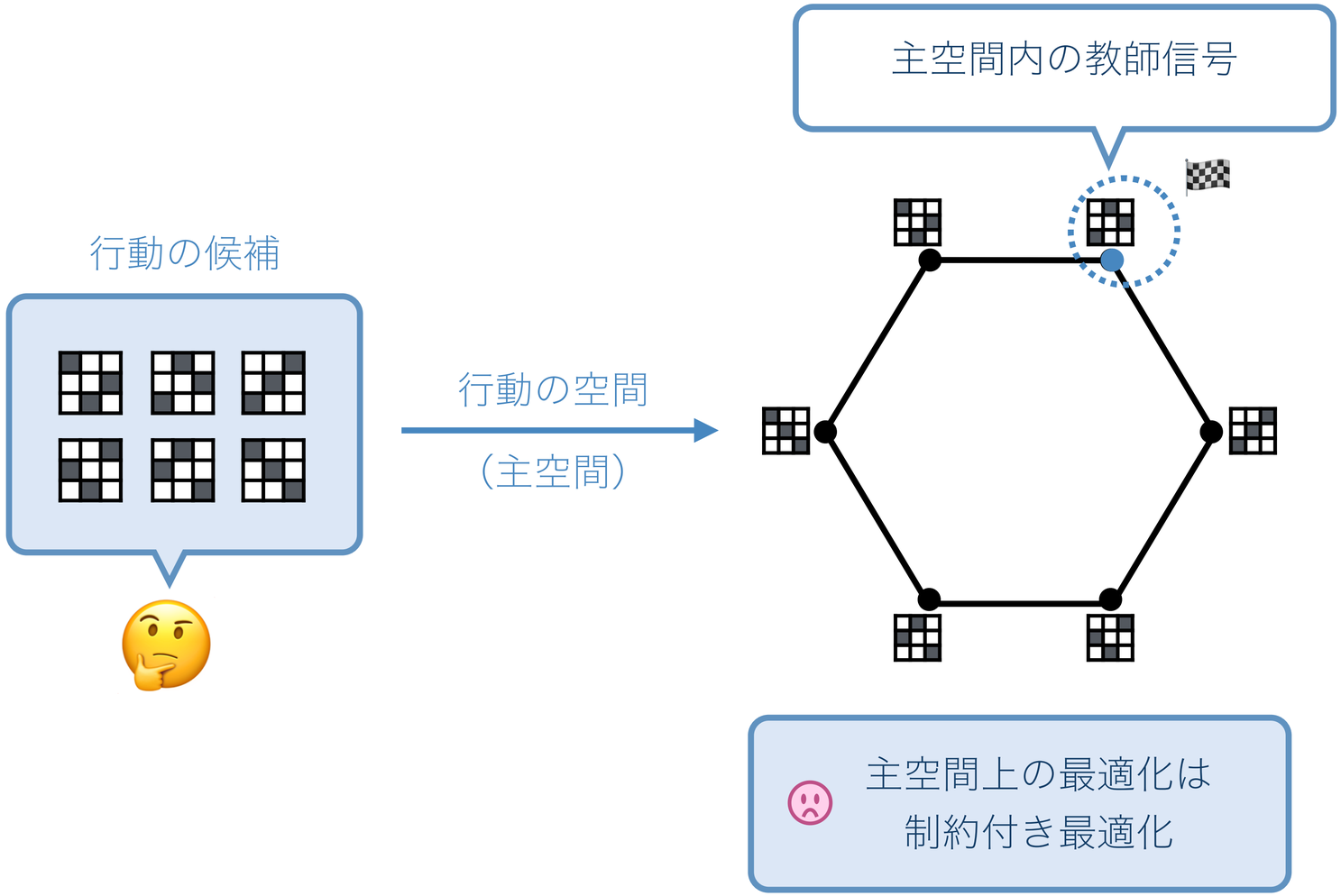
意思決定の枠組み: 機械学習の視点から



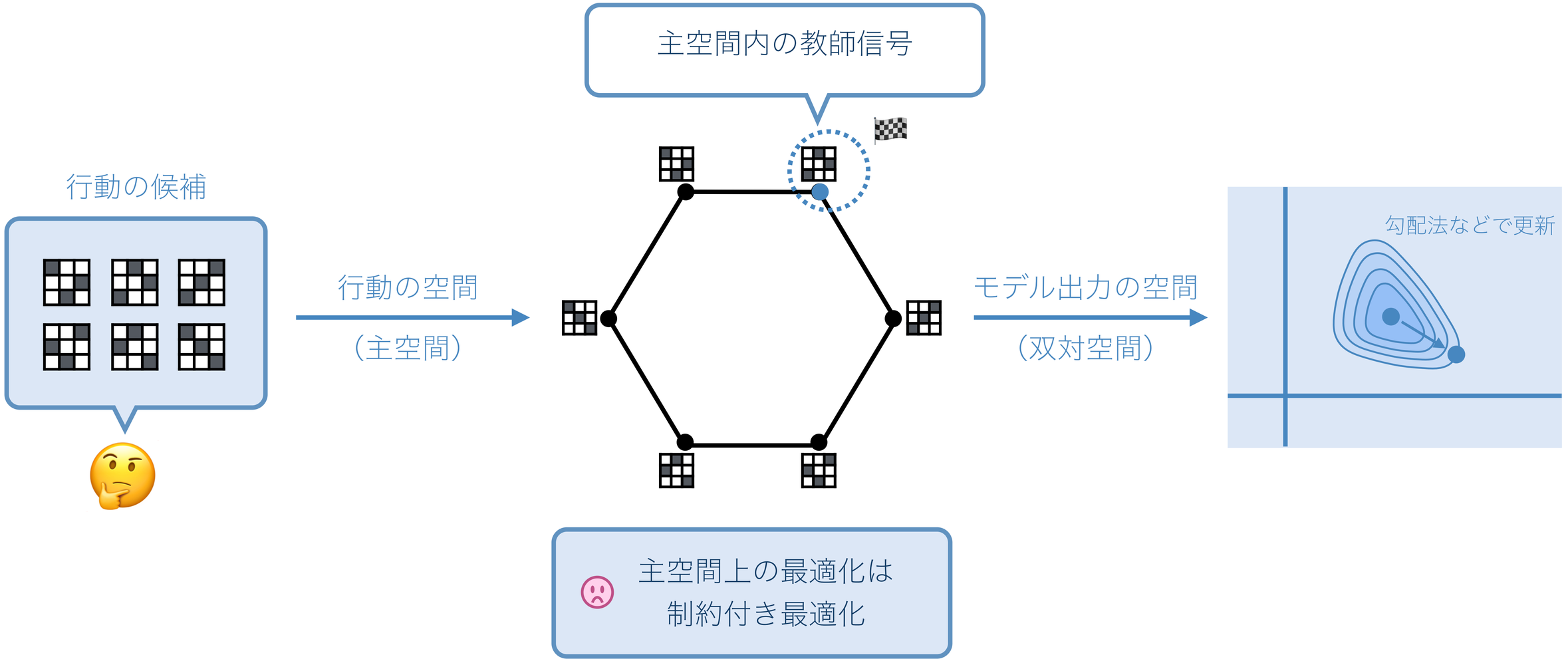
意思決定の枠組み: 機械学習の視点から



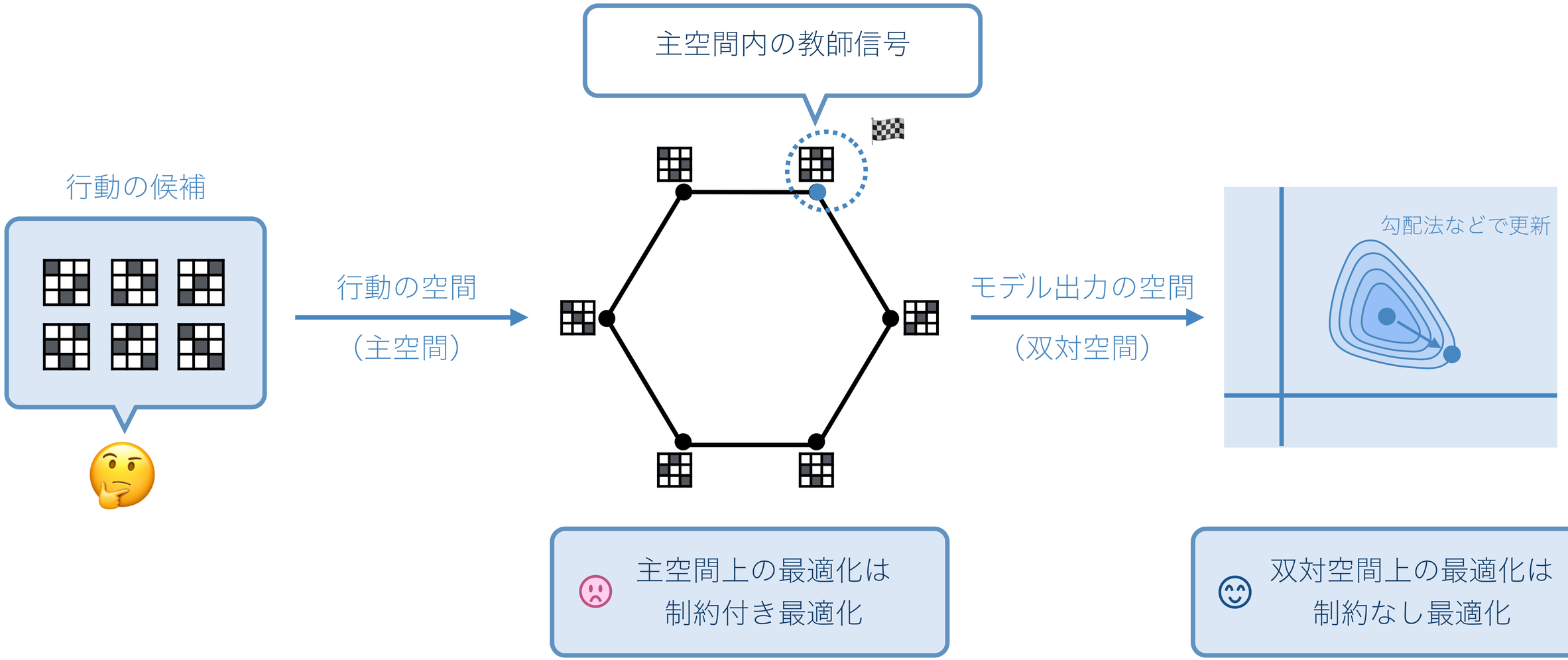
意思決定の枠組み: 機械学習の視点から



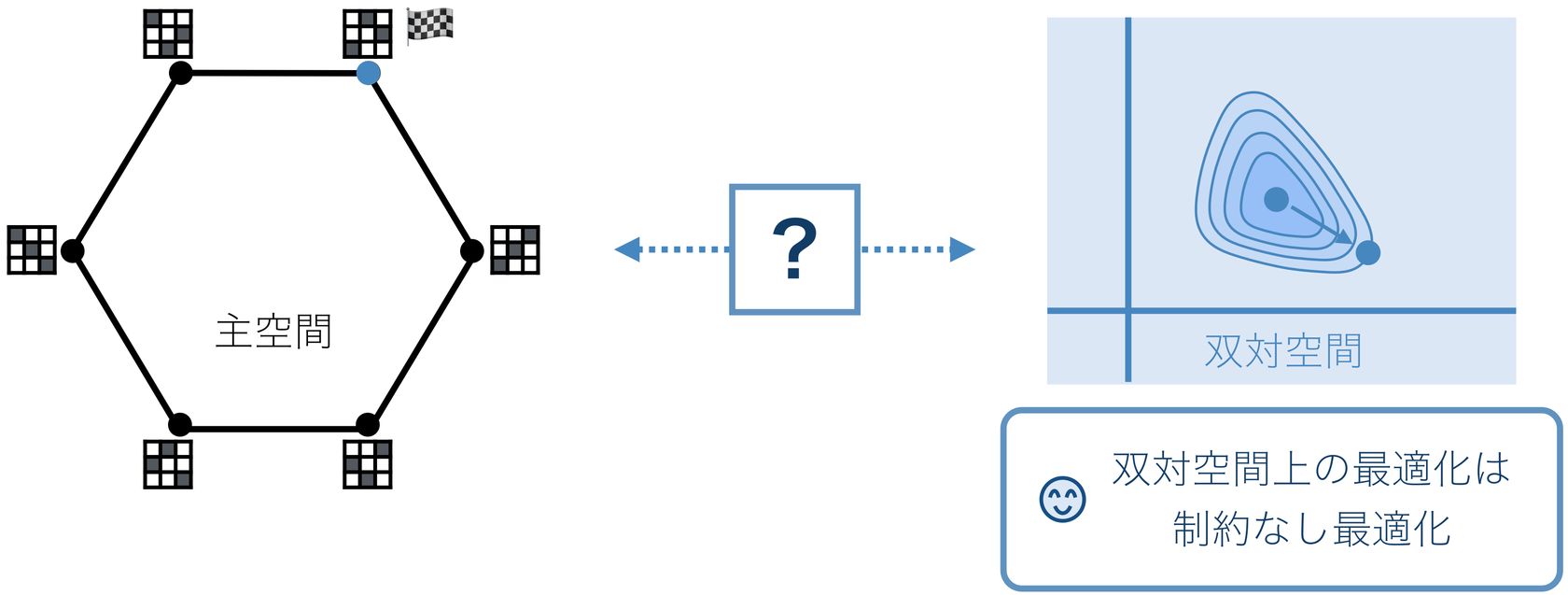
意思決定の枠組み: 機械学習の視点から



意思決定の枠組み: 機械学習の視点から



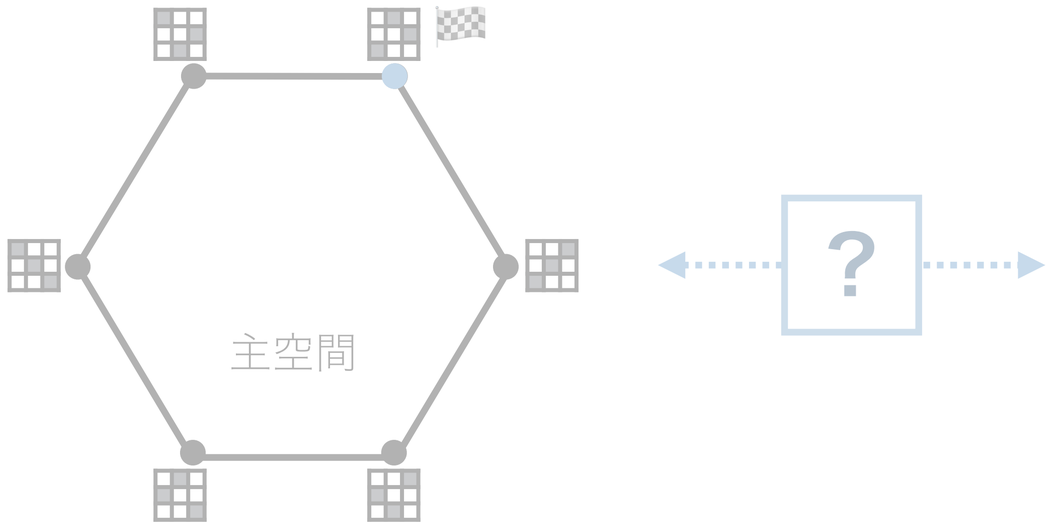
意思決定の枠組み: 機械学習の視点から



今日の主要なテーマ

- **Q1 (主双対の関係)** 主空間と双対空間はどのように行き来する？
- **Q2 (主双対の距離)** 双対変数と主変数の近さはどう測る？
  - ❖ 双対空間にはどのような目的関数を入れればよい？

補遺: 最適化に対するアプローチ



双対空間上の最適化は
制約なし最適化

- 機械学習コミュニティでの最適化の捉え方とは？ **（以下の感覚は大いに個人差があるので注意！）**
  - 離散最適化 → 連続緩和したい
  - 制約付き最適化 → 未定乗数法で制約なし最適化にしたい
  - 制約なし最適化 → 勾配法 + α で最適化すればヨシ
  - 非凸最適化 → 停留点が求まればとりあえず OK

本当は厳密な近似保証・収束判定が必要
だけど解の厳密な最適性以外にも
気になる性質が多い…
⇒ 最適化屋さんの参入は大歓迎です！

機械学習と凸共役の交わり（目次）

前半

- **二値分類問題: 主空間の観点から**
- 二値分類問題: 双対空間の観点から
- 応用: 非対称リンク関数を用いた二値応答回帰

  Bao & Sugiyama (2021) “Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation”

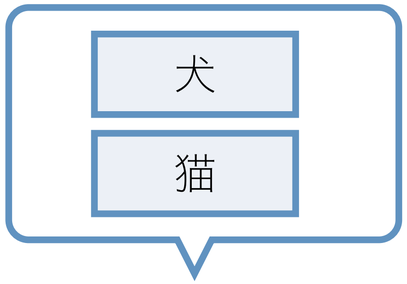
後半

- 最適輸送問題: 双対空間の観点から
- 応用: q-指数分布を用いたスパース最適輸送

  Bao & Sakaue (2022) “Sparse Regularized Optimal Transport with Deformed q-Entropy”
- その他の問題

最初の例: ロジスティック回帰

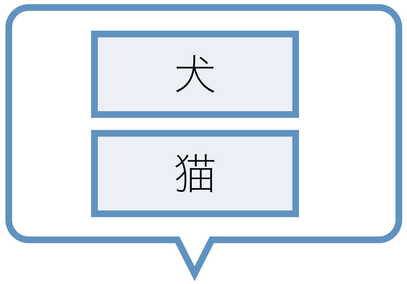
問題: 二値ラベルの決定



重要事項: このパートでは入力特徴量 $\mathbf{x}$ は省略し、$\mathbf{x}$ 一点固定のもとで議論を進めていることが多い

最初の例: ロジスティック回帰

問題: 二値ラベルの決定



主空間: 1次元確率単体

$$\triangle^1 = [0,1]$$

重要事項: このパートでは入力特徴量 $\mathbf{x}$ は省略し、$\mathbf{x}$ 一点固定のもとで議論を進めていることが多い

最初の例: ロジスティック回帰

問題: 二値ラベルの決定

犬

猫

🤔

主空間: 1次元確率単体

$$\triangle^1 = [0,1]$$

犬

1

0

猫

双対空間: 1次元 Euclid 空間

😊 制約なし最適化

$f(\mathbf{x})$

$\mathbb{R}$

(ロジット)

重要事項: このパートでは入力特徴量 $\mathbf{x}$ は省略し、$\mathbf{x}$ 一点固定のもとで議論を進めていることが多い

最初の例: ロジスティック回帰

問題: 二値ラベルの決定

犬

猫

🤔

主空間: 1次元確率単体

$\triangle^1 = [0,1]$

犬

1

0

猫

双対空間: 1次元 Euclid 空間

😊 制約なし最適化

$f(\mathbf{x})$

(ロジット)

$\mathbb{R}$

Q1. 主双対の関係

重要事項: このパートでは入力特徴量 $\mathbf{x}$ は省略し、$\mathbf{x}$ 一点固定のもとで議論を進めていることが多い

最初の例: ロジスティック回帰

問題: 二値ラベルの決定

犬

猫

🤔

主空間: 1次元確率単体

$\triangle^1 = [0,1]$

犬

1

0

猫

双対空間: 1次元 Euclid 空間

😊 制約なし最適化

Q2. 目的関数

$f(\mathbf{x})$

(ロジット)

$\mathbb{R}$

Q1. 主双対の関係

重要事項: このパートでは入力特徴量 $\mathbf{x}$ は省略し、$\mathbf{x}$ 一点固定のもとで議論を進めていることが多い

最初の例: ロジスティック回帰

Q1. 主双対の関係

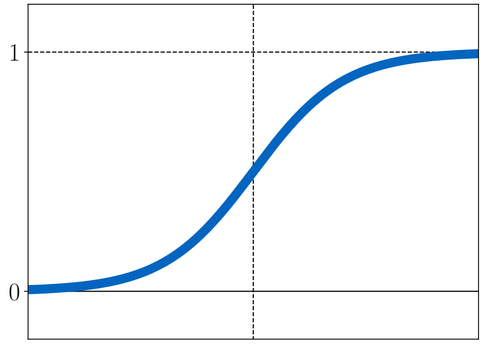
A1. 逆ロジットリンク関数

$$\pi = \psi^{-1}(\theta) := \frac{1}{1+\exp(-\theta)}$$

(シグモイド関数)

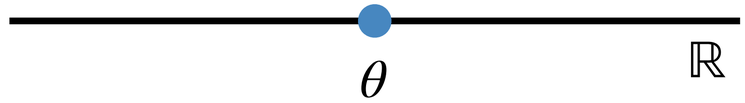
主空間: 1次元確率単体

$$\triangle^1 = [0,1]$$



逆ロジットリンク

双対空間: 1次元 Euclid 空間



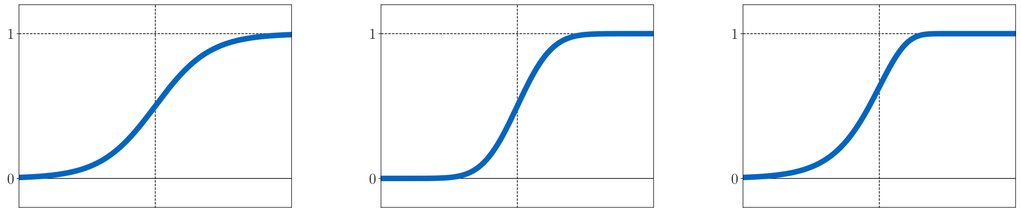
補足: 一般化線形モデルとの関係

- 一般化線形モデル = 指数型分布族 + 線形予測子 + リンク関数

ベルヌーイ分布で二値応答の確率をモデリング

$$Y|\mathbf{x} \sim \text{Bernoulli}(\pi)$$

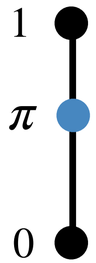
$\in$ 指数型分布族



逆ロジットリンク　逆プロビットリンク　逆 cloglog リンク

平均パラメータの予測

平均パラメータ



逆リンク関数による非線形変換

$$\pi = \psi^{-1}(\theta) := \frac{1}{1+\exp(-\theta)}$$

その他のリンク関数の例

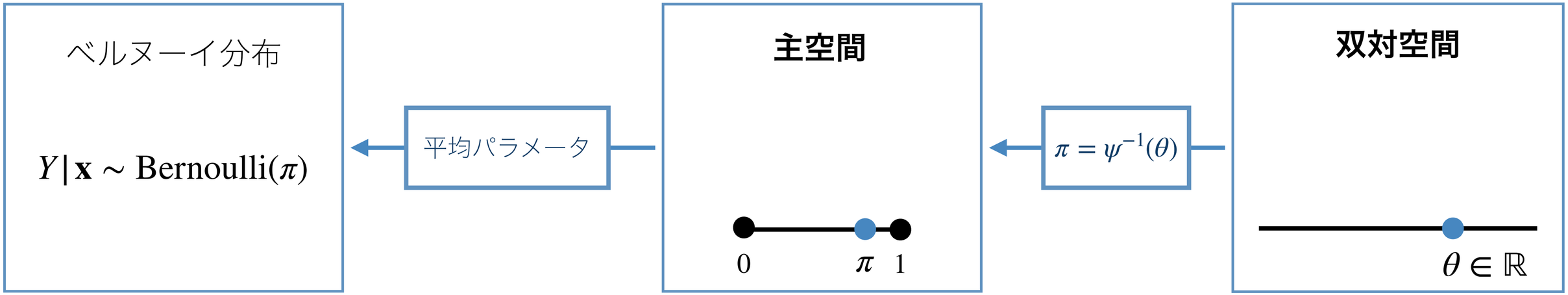
- プロビットリンク関数: ガウス分布の CDF の逆関数
- cloglog リンク関数: $\psi(\pi) = \ln(-\ln(1-\pi))$

ロジット

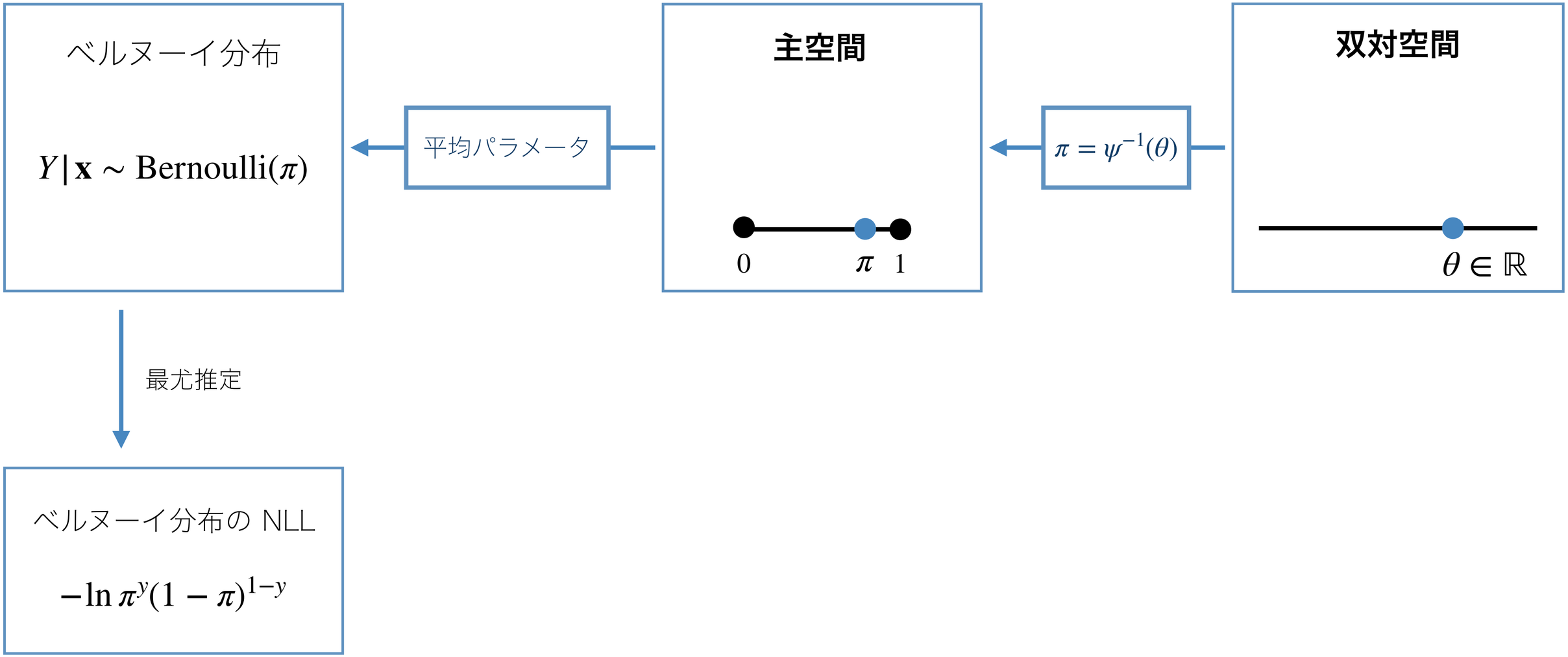
（線形予測子）

$$\theta = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}$$

最初の例: ロジスティック回帰



最初の例: ロジスティック回帰



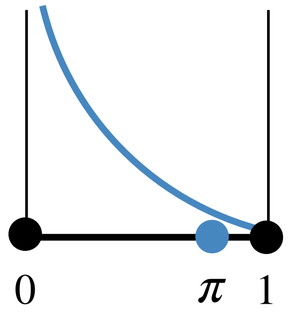
最初の例: ロジスティック回帰

ベルヌーイ分布

$$Y \mid \mathbf{x} \sim \text{Bernoulli}(\pi)$$

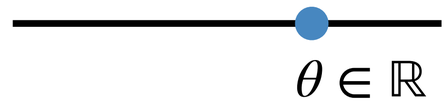
平均パラメータ

主空間



$\pi = \psi^{-1}(\theta)$

双対空間



最尤推定

ベルヌーイ分布の NLL

$$-\ln \pi^{y}(1-\pi)^{1-y}$$

A2. 対数損失

$$\begin{aligned} -\ln \pi \quad & (y=1) \\ -\ln(1-\pi) \quad & (y=0) \end{aligned}$$

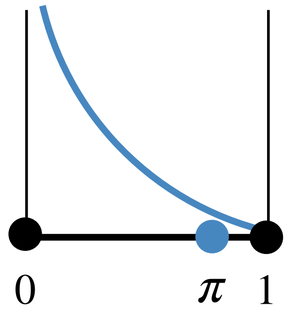
最初の例: ロジスティック回帰

ベルヌーイ分布

$$Y \mid \mathbf{x} \sim \text{Bernoulli}(\pi)$$

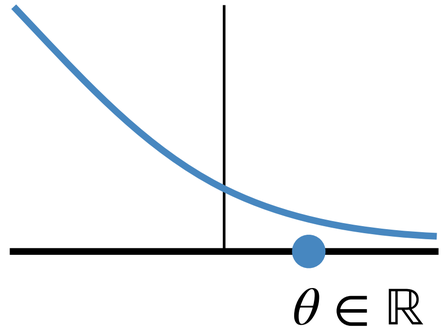
平均パラメータ

主空間



$\pi = \psi^{-1}(\theta)$

双対空間



最尤推定

ベルヌーイ分布の NLL

$$-\ln \pi^{y}(1-\pi)^{1-y}$$

Q2. 損失関数

A2. 対数損失

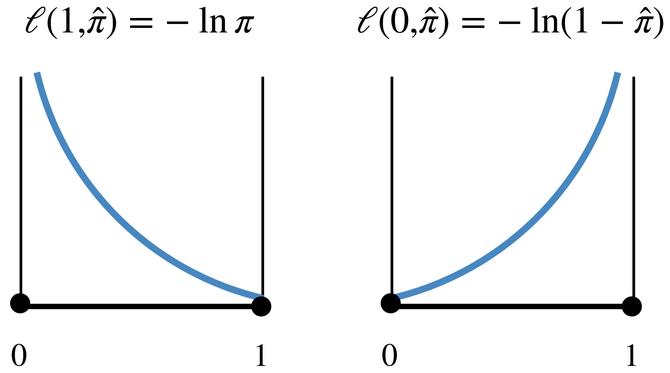
$$\begin{aligned} -\ln \pi \quad & (y=1) \\ -\ln(1-\pi) \quad & (y=0) \end{aligned}$$

A2. ロジスティック損失

$$\begin{aligned} \ln(1+\exp(-\theta)) \\ \ln(1+\exp(\theta)) \end{aligned}$$

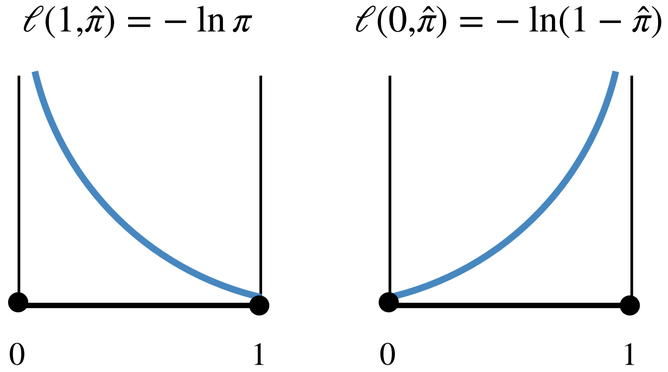
二値応答の損失関数

- 確率の推定 $\hat{\pi} \in [0,1]$ (主変数) が二値応答 $y \in \{0,1\}$ にどれくらい近いか？
  - 二値の損失関数 $\ell(y, \pi)$
  - 例: 対数損失 $\ell(1,\hat{\pi}) = -\ln \pi$, $\ell(0,\hat{\pi}) = -\ln(1-\hat{\pi})$
    - 逆ロジットリンク関数 $\hat{\pi} = (1+\exp(-\theta))^{-1}$ を代入すればロジスティック損失

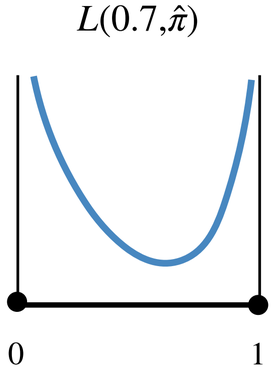


二値応答の損失関数

- 確率の推定 $\hat{\pi} \in [0,1]$ (主変数) が二値応答 $y \in \{0,1\}$ にどれくらい近いか？
  - 二値の損失関数 $\ell(y, \pi)$
  - 例: 対数損失 $\ell(1,\hat{\pi}) = -\ln \pi$, $\ell(0,\hat{\pi}) = -\ln(1-\hat{\pi})$
    - 逆ロジットリンク関数 $\hat{\pi} = (1+\exp(-\theta))^{-1}$ を代入すればロジスティック損失



- 確率の推定 $\hat{\pi}$ (主変数) が真のクラス確率 π (主変数) にどれくらい近いか？
  - (クラス) 条件付きリスク $L(\pi, \hat{\pi}) := \mathbb{E}_{Y \sim \mathrm{Bernoulli}(\pi)}[\ell(Y, \hat{\pi})] = \pi \ell(1,\hat{\pi}) + (1-\pi)\ell(0,\hat{\pi})$
  - 達成可能な下界: ベイズ条件付きリスク $\underline{L}(\pi) := \inf_{\hat{\pi} \in [0,1]} L(\pi, \hat{\pi})$
    - 対数損失の場合は Shannon エントロピーに一致



「良い」損失関数とは？

条件付きリスク $L(\pi, \hat{\pi}) := \mathbb{E}_{Y \sim \mathrm{Bernoulli}(\pi)}[\ell(\hat{\pi}, Y)] = \pi\ell(1,\hat{\pi}) + (1-\pi)\ell(0,\hat{\pi})$

ベイズリスク $\underline{L}(\pi) := \inf_{\hat{\pi} \in [0,1]} L(\pi, \hat{\pi})$

「良い」損失関数とは？

条件付きリスク $L(\pi, \hat{\pi}) := \mathbb{E}_{Y\sim\mathrm{Bernoulli}(\pi)}[\ell(\hat{\pi}, Y)] = \pi\ell(1,\hat{\pi}) + (1-\pi)\ell(0,\hat{\pi})$　　ベイズリスク $\underline{L}(\pi) := \inf_{\hat{\pi}\in[0,1]} L(\pi, \hat{\pi})$

定義 (Proper loss). 二値損失 $\ell(\pi, \hat{\pi})$ が (strictly) proper $\iff L(\pi,\pi) = \underline{L}(\pi) \quad \forall\pi \in [0,1]$

「良い」損失関数とは？

条件付きリスク $L(\pi, \hat{\pi}) := \mathbb{E}_{Y\sim\mathrm{Bernoulli}(\pi)}[\ell(\hat{\pi}, Y)] = \pi\ell(1,\hat{\pi}) + (1-\pi)\ell(0,\hat{\pi})$　　ベイズリスク $\underline{L}(\pi) := \inf_{\hat{\pi}\in[0,1]} L(\pi, \hat{\pi})$

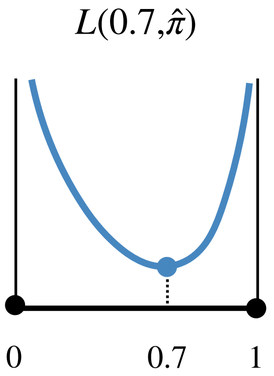
定義 (Proper loss). 二値損失 $\ell(\pi, \hat{\pi})$ が (strictly) proper $\iff L(\pi, \pi) = \underline{L}(\pi) \quad \forall \pi \in [0,1]$

- Proper loss の直感: 推定 $\hat{\pi}$ が真の π と一致するときのみ損失値が下界を達成
  - 必要最小限な性質
- 性質: Proper loss は **weight function** と一対一対応

$$\ell \text{ is proper} \iff \frac{-\ell'(1,\hat{\pi})}{1-\hat{\pi}} = \frac{\ell'(0,\hat{\pi})}{\hat{\pi}} := w(\hat{\pi})$$

Weight function

  - 意味づけは後ほど



その予測、どれくらい最適？

条件付きリスク $L(\pi, \hat{\pi}) := \mathbb{E}_{Y\sim\mathrm{Bernoulli}(\pi)}[\ell(\hat{\pi}, Y)] = \pi\ell(1,\hat{\pi}) + (1-\pi)\ell(0,\hat{\pi})$　　ベイズリスク $\underline{L}(\pi) := \inf_{\hat{\pi}\in[0,1]} L(\pi, \hat{\pi})$

- 条件付きリスク vs. ベイズリスク
  - 条件付きリスク = 推定 $\hat{\pi}$ に対する罰則　　(“一般化” クロスエントロピー)
  - ベイズリスク = 真値が π の際の達成可能な下界　　(“一般化” エントロピー)
- **リグレット** = 条件付きリスク − ベイズ条件付きリスク

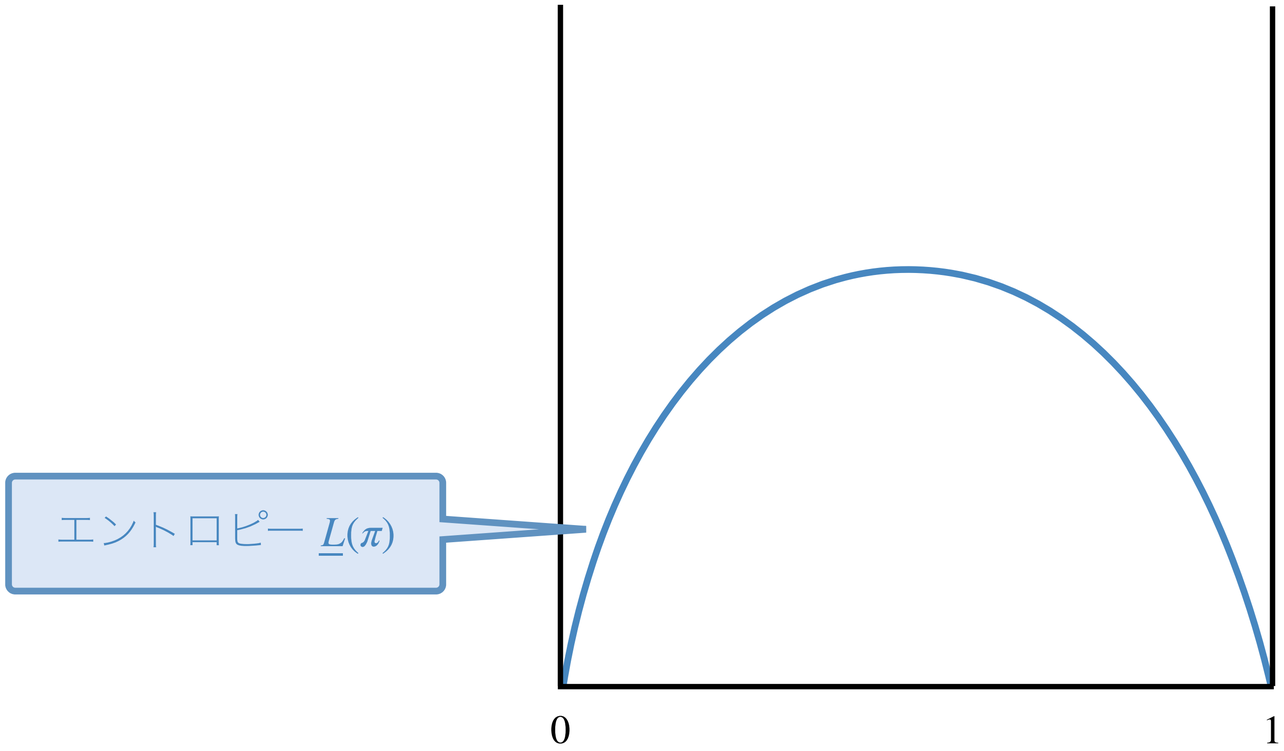
$$R(\pi, \hat{\pi}) = L(\pi, \hat{\pi}) - \underline{L}(\pi)$$

クロスエントロピー　　エントロピー

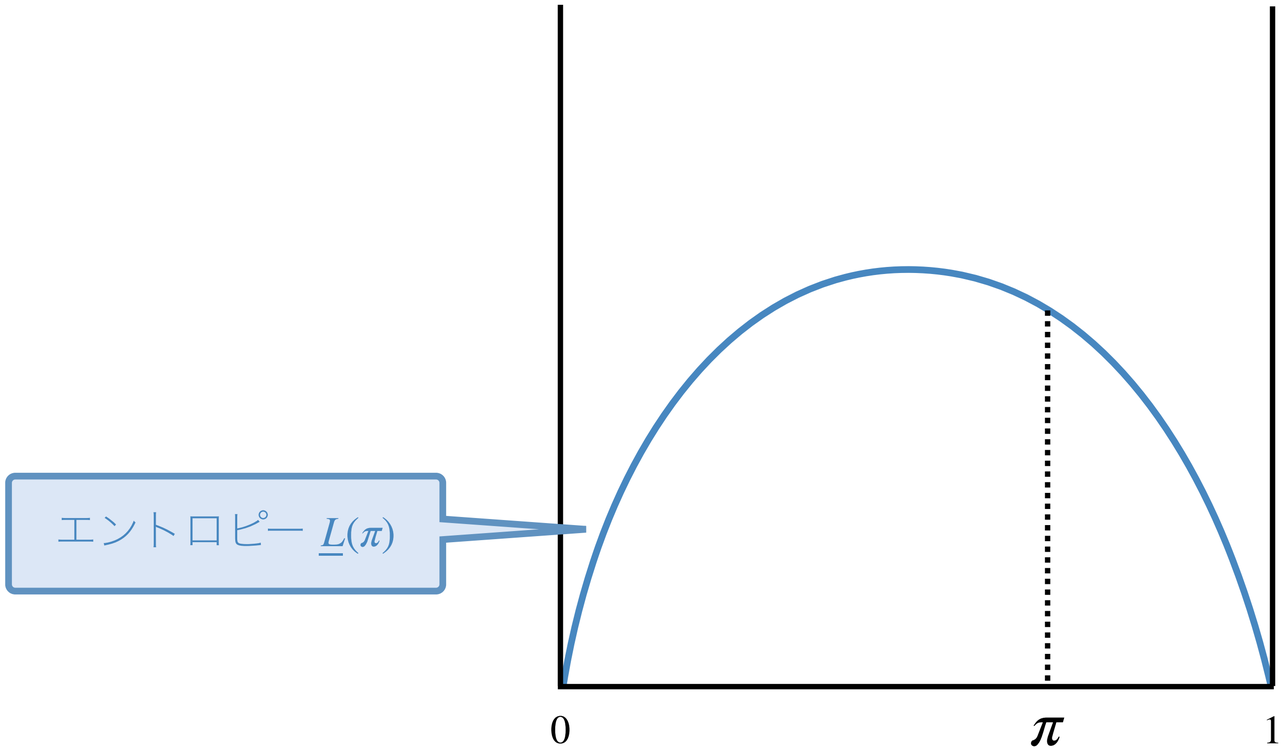
  - (各点 $\mathbf{x}$ における) 推定 $\hat{\pi}$ の最適性
- 対数損失の場合は「リグレット = KL ダイバージェンス」

$$R(\pi, \hat{\pi}) = \pi \ln\left(\frac{\pi}{\hat{\pi}}\right) + (1-\pi)\ln\left(\frac{1-\pi}{1-\hat{\pi}}\right)$$

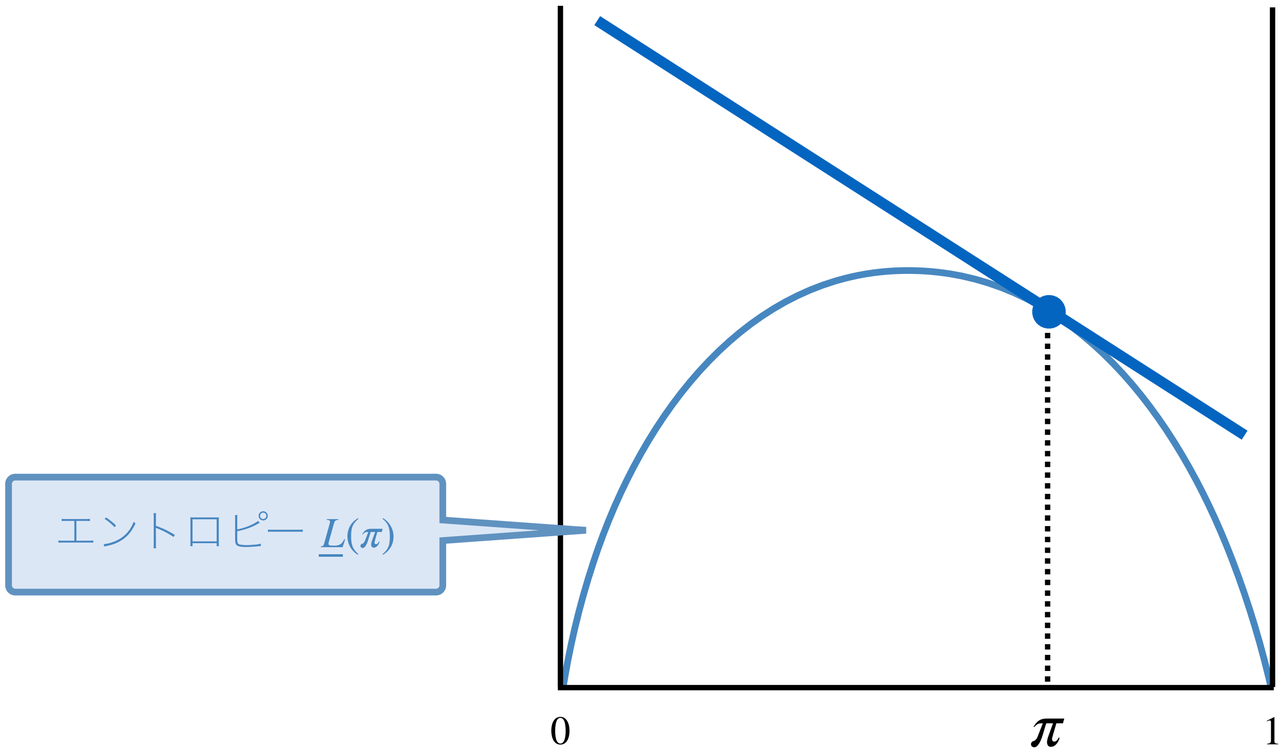
Bregman ダイバージェンスの観点から



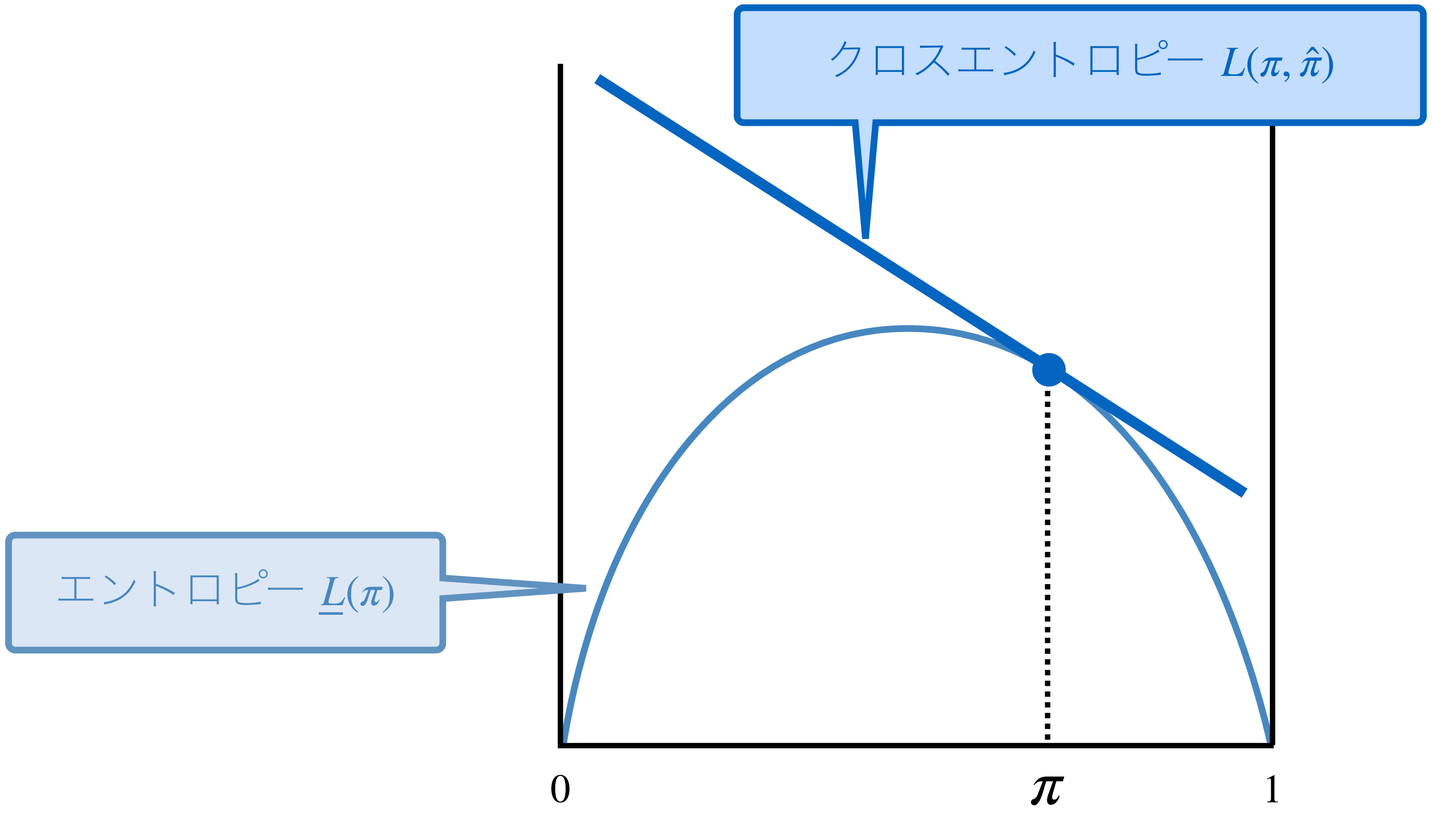
Bregman ダイバージェンスの観点から



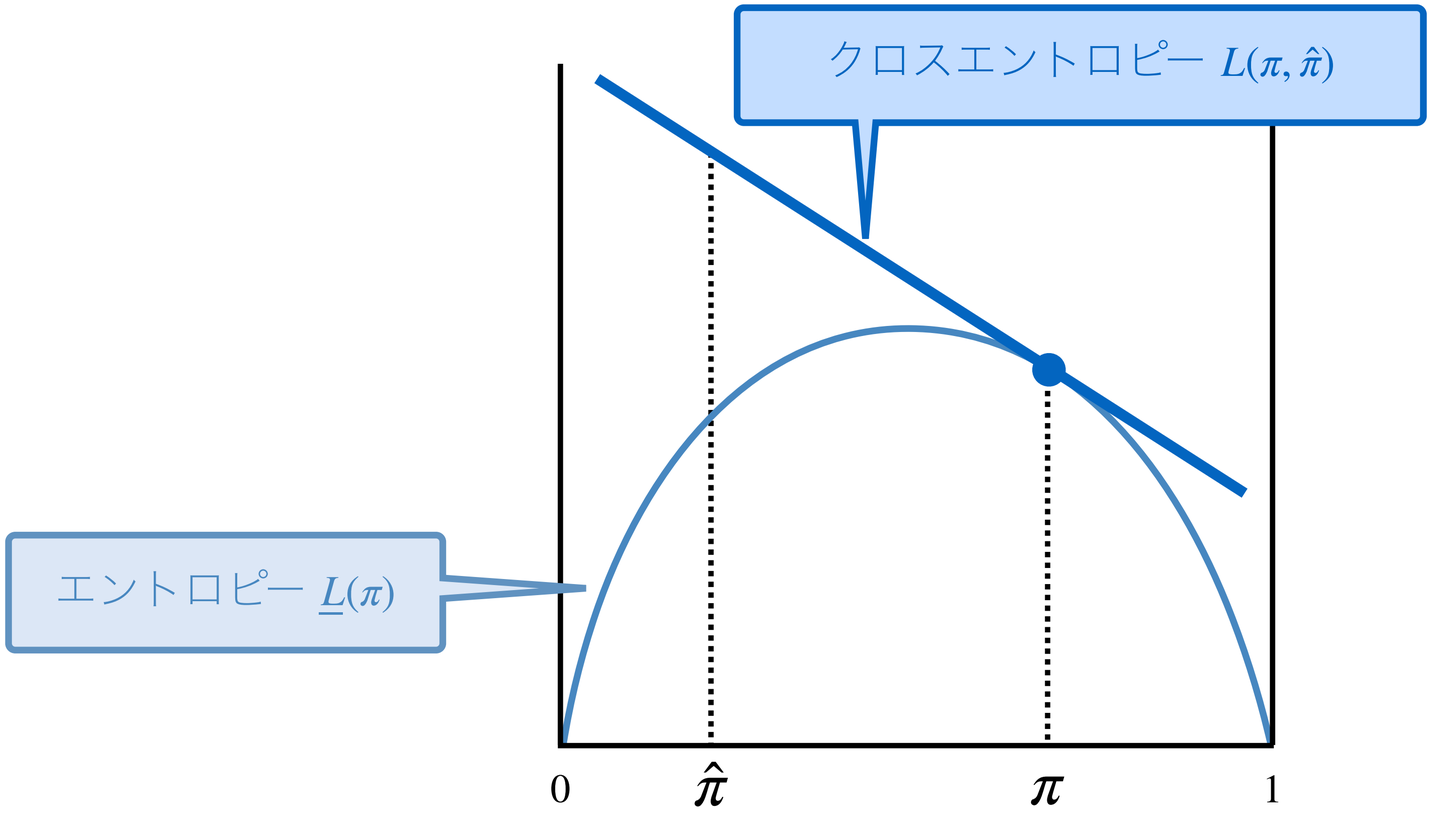
Bregman ダイバージェンスの観点から



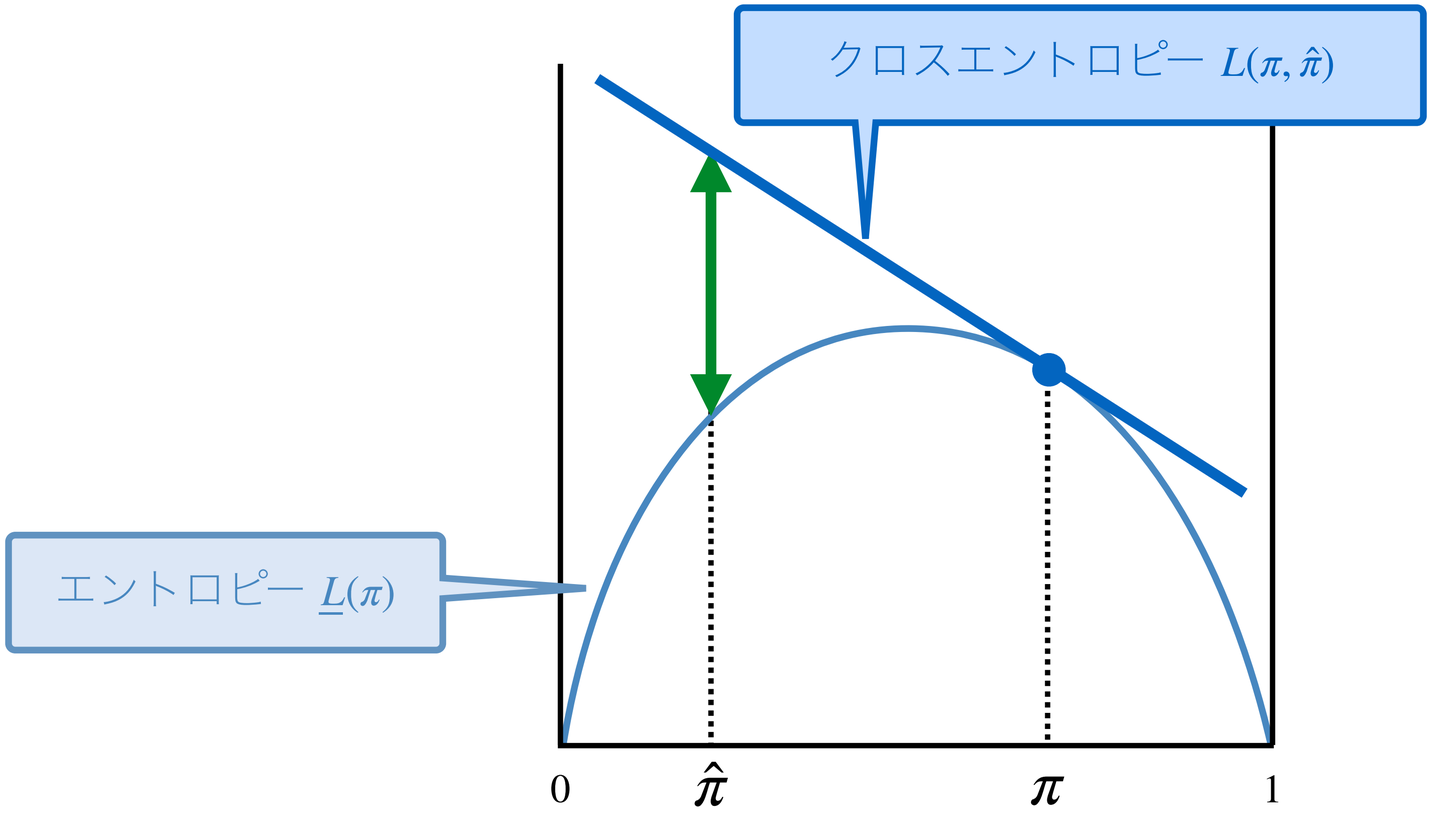
Bregman ダイバージェンスの観点から



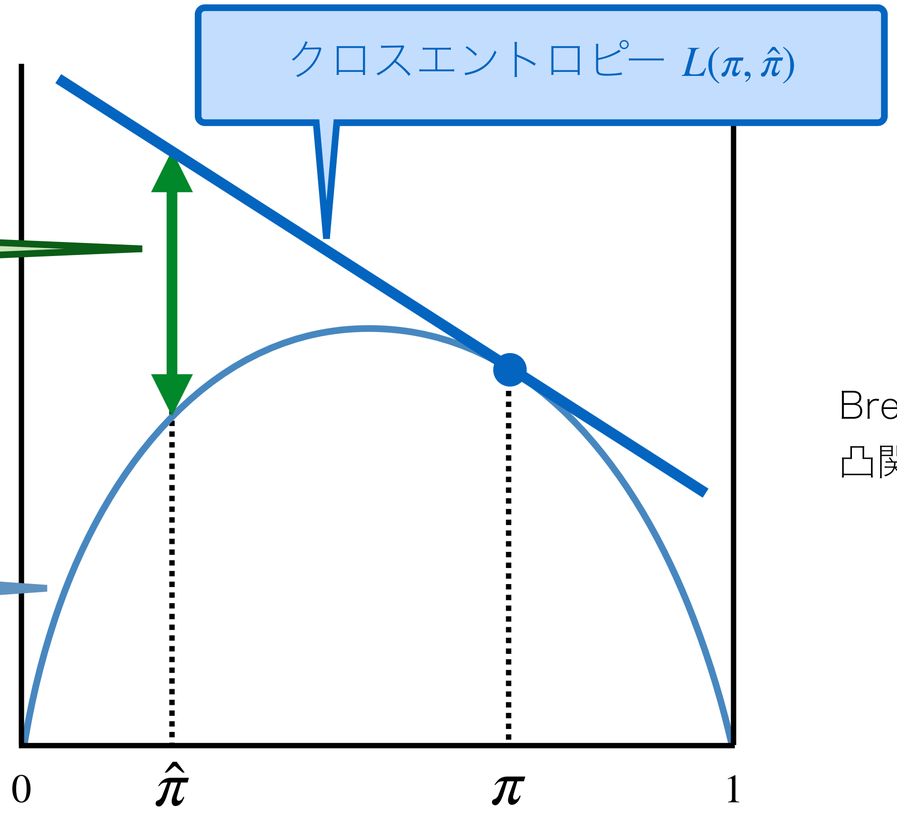
Bregman ダイバージェンスの観点から



Bregman ダイバージェンスの観点から

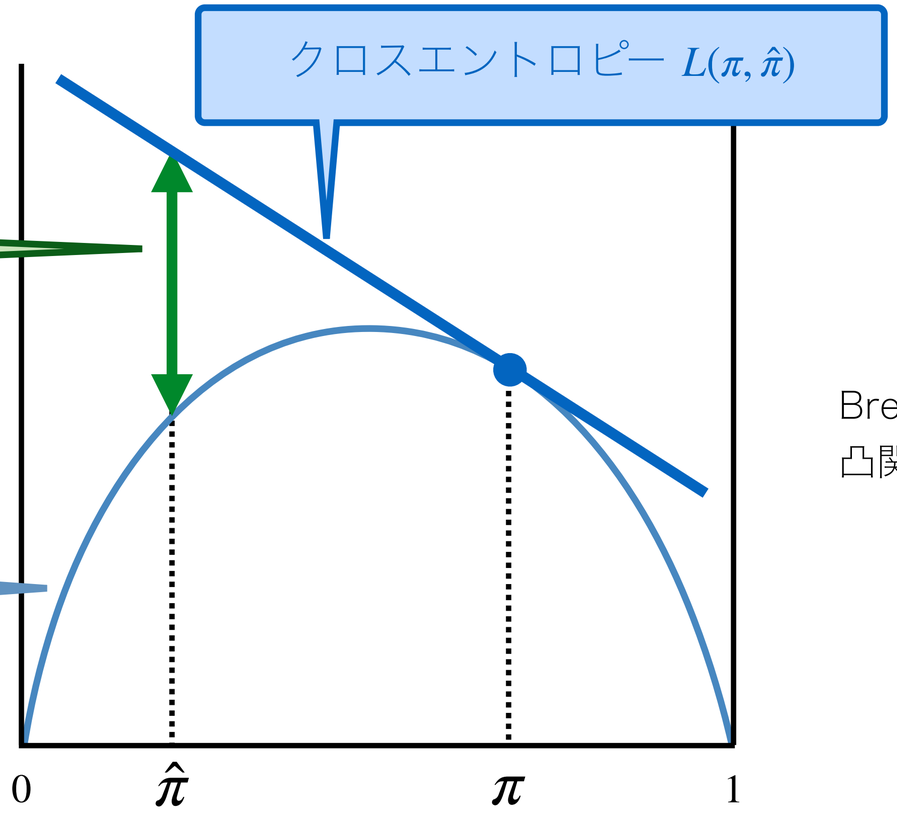


Bregman ダイバージェンスの観点から



Bregman ダイバージェンスは
凸関数の「1次近似誤差」として解釈可能

Bregman ダイバージェンスの観点から



Bregman ダイバージェンスは
凸関数の「1次近似誤差」として解釈可能

リグレット = (負の) ベイズ条件付きリスクを generator とする Bregman ダイバージェンス

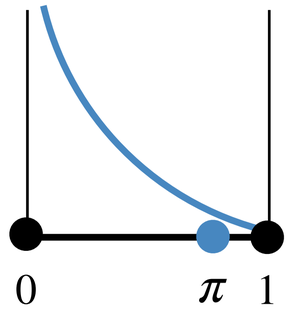
まとめ: Proper loss の枠組み

ベルヌーイ分布で
二値応答をモデリング

$$Y \mid \mathbf{x} \sim \mathrm{Bernoulli}(\pi)$$

平均パラメータの
モデリング

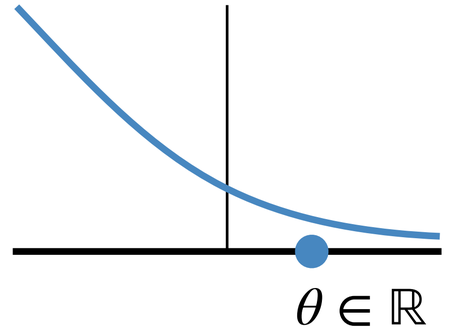
主空間 (平均の推定)



逆リンク関数

$$\pi = \psi^{-1}(\theta)$$

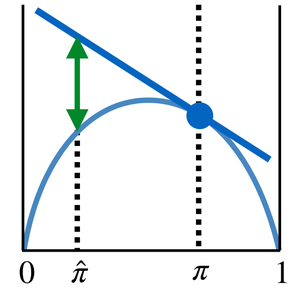
双対空間 (最適化 in ロジット)



損失: 主空間内での条件付きリスク

$$L(\pi, \hat{\pi}) := \mathbb{E}_{Y \sim \pi}[\ell(\hat{\pi}, Y)]$$

Bregman ダイバージェンス
として解釈可能



Q. 損失関数 & リンク関数の組合せはどのように選ぶ？

機械学習と凸共役の交わり（目次）

前半

- 二値分類問題: 主空間の観点から
- **二値分類問題: 双対空間の観点から**
- 応用: 非対称リンク関数を用いた二値応答回帰

  Bao & Sugiyama (2021) “Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation”

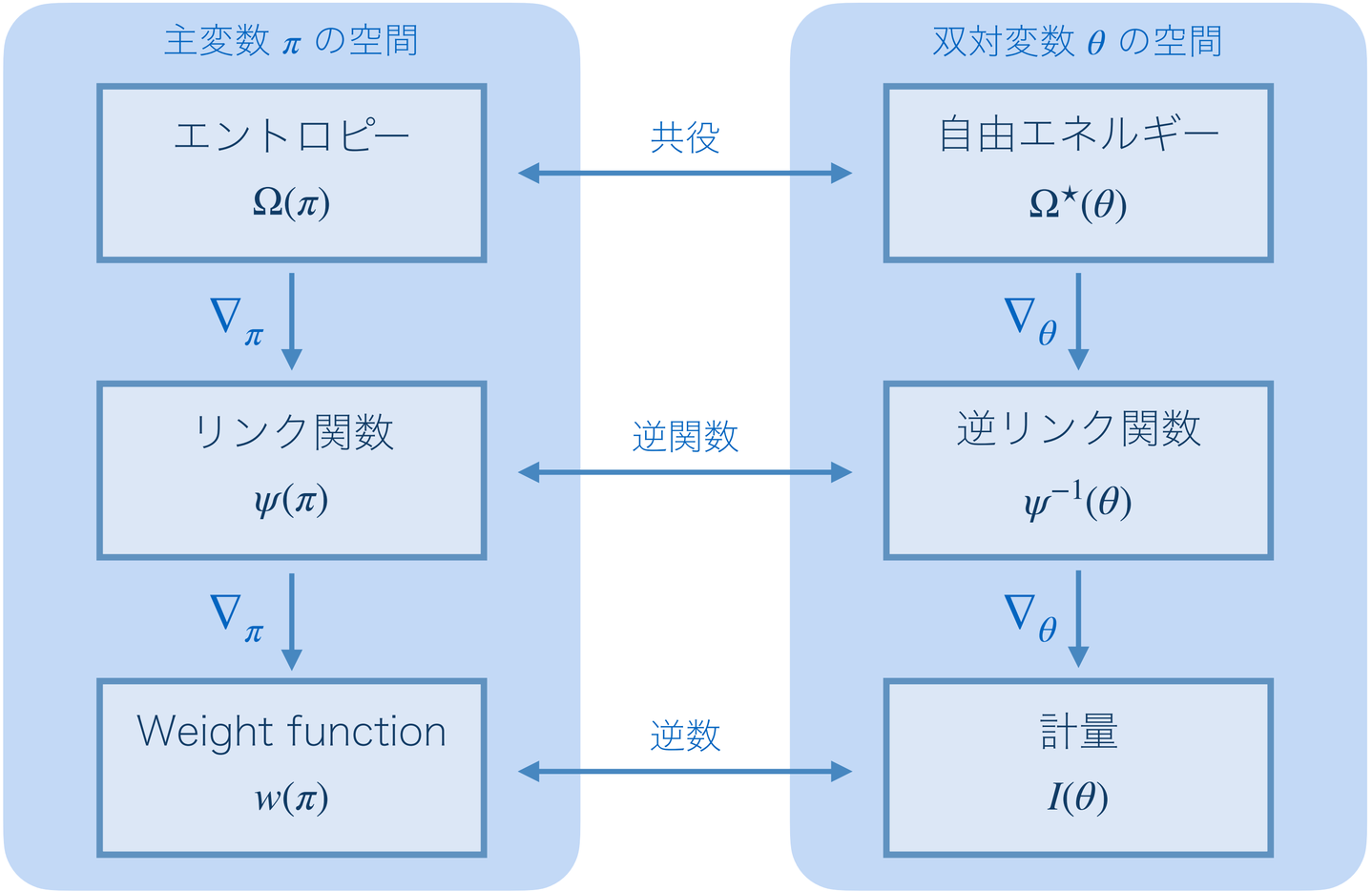
後半

- 最適輸送問題: 双対空間の観点から
- 応用: q-指数分布を用いたスパース最適輸送

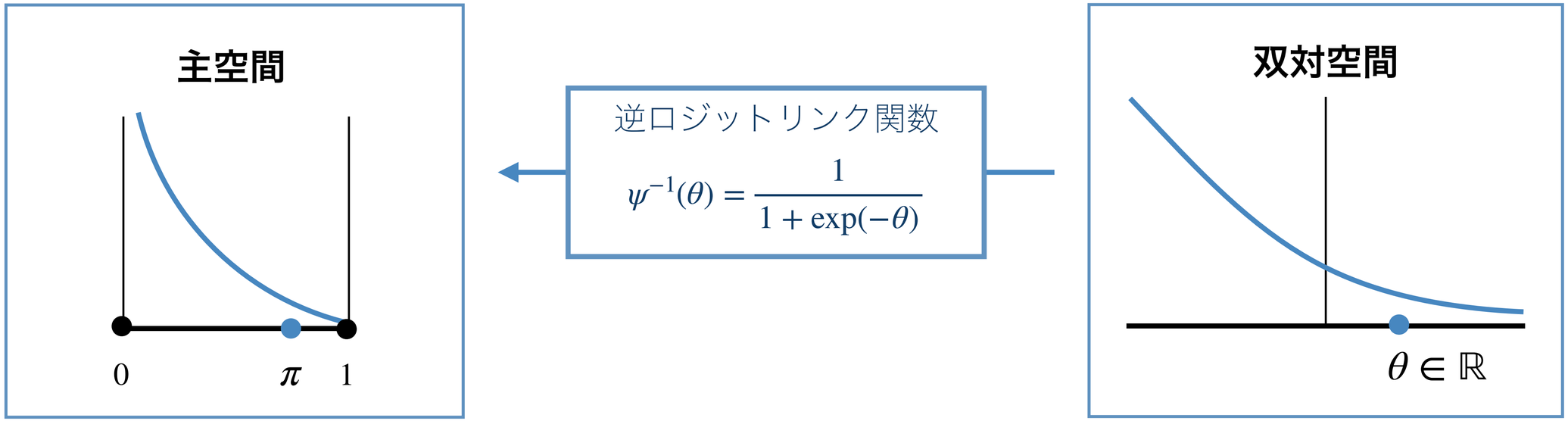
  Bao & Sakaue (2022) “Sparse Regularized Optimal Transport with Deformed q-Entropy”
- その他の問題

このパートの目標

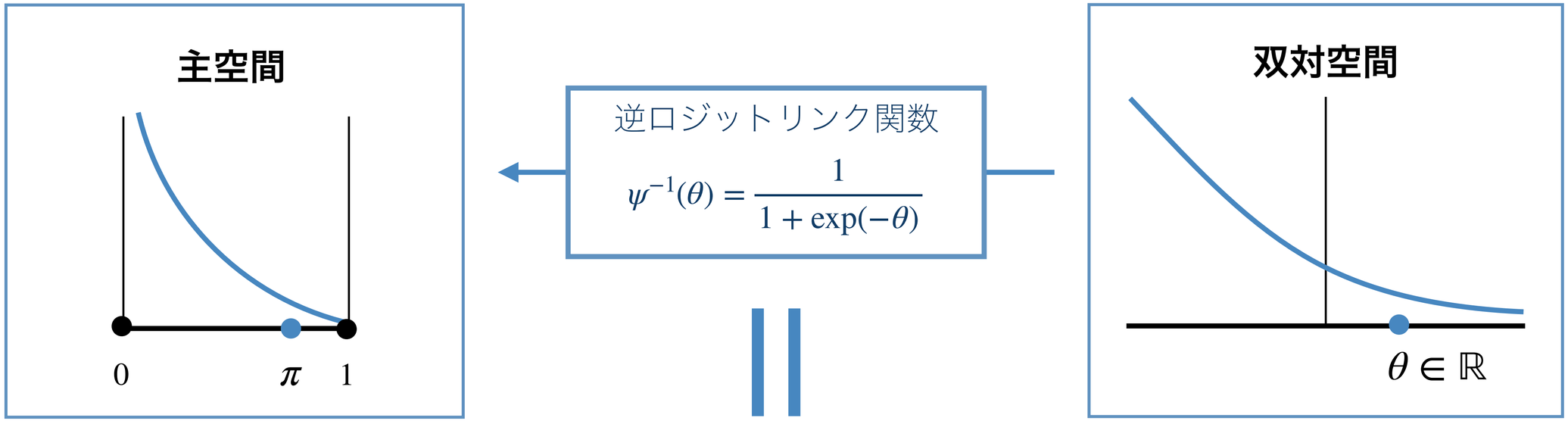
- 以下の関係の理解



再訪: ロジスティック回帰



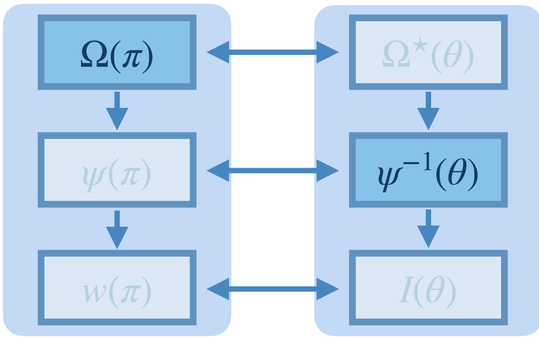
再訪: ロジスティック回帰



逆リンク関数: ある Ω に対する **最大エントロピーモデル** として解釈可能

$$\psi^{-1}(\theta) = \arg\max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

ただし $\Omega(\pi) = \pi \ln \pi + (1-\pi)\ln(1-\pi)$ は負の Shannon エントロピー

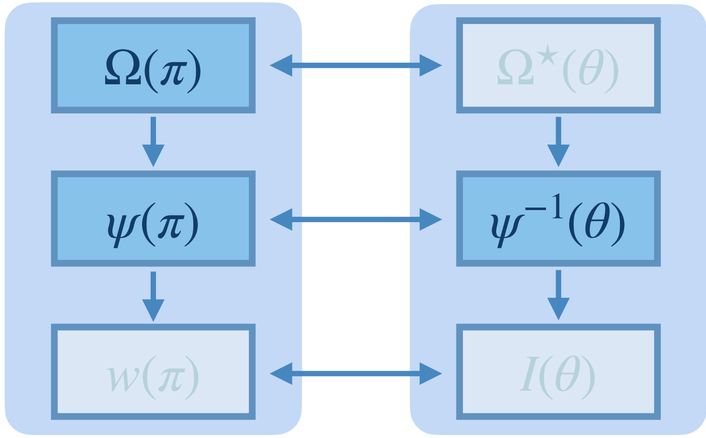


凸共役としての最大エントロピーモデル

最大エントロピーモデル

$$\psi^{-1}(\theta) = \underset{\pi \in [0,1]}{\arg\max}\, \theta\pi - \Omega(\pi)$$

1次の最適性条件

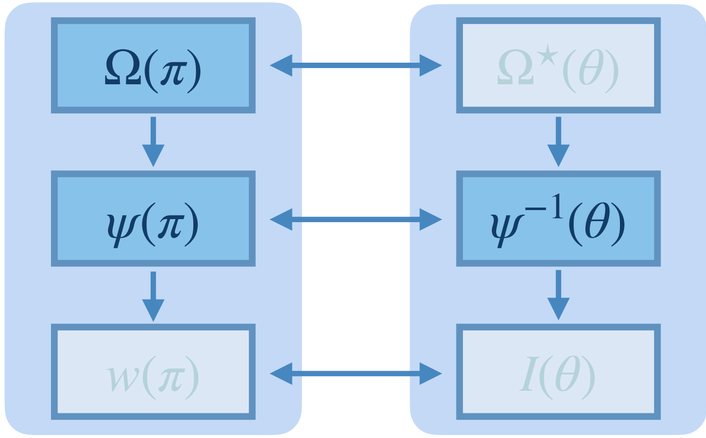


凸共役としての最大エントロピーモデル

最大エントロピーモデル

$$\psi^{-1}(\theta) = \underset{\pi \in [0,1]}{\arg\max}\, \theta\pi - \Omega(\pi)$$

1次の最適性条件 → $\theta = \nabla\Omega(\pi)$



凸共役としての最大エントロピーモデル

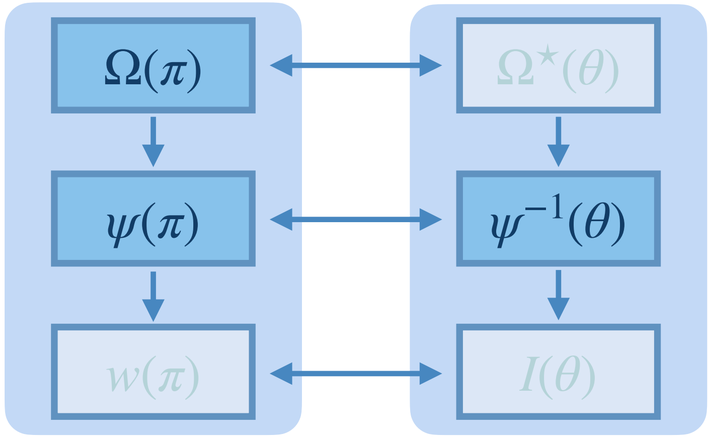
最大エントロピーモデル

$$\psi^{-1}(\theta) = \underset{\pi \in [0,1]}{\arg\max}\, \theta\pi - \Omega(\pi)$$

1次の最適性条件

$$\theta = \nabla\Omega(\pi)$$

$$\pi = [\nabla\Omega]^{-1}(\theta)$$



凸共役としての最大エントロピーモデル

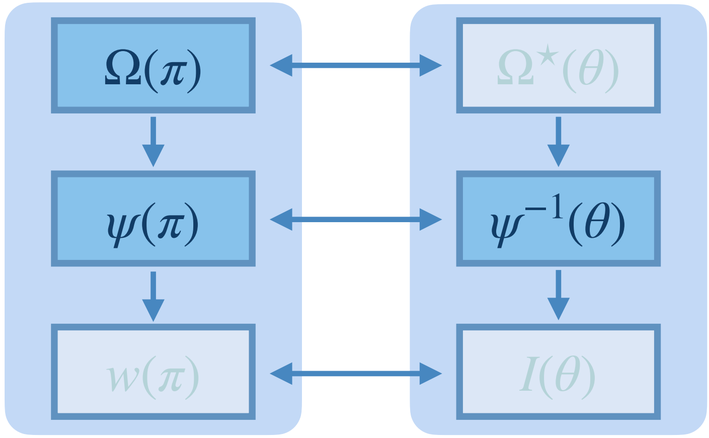
最大エントロピーモデル

$$\psi^{-1}(\theta) = \underset{\pi \in [0,1]}{\arg\max}\ \theta\pi - \Omega(\pi)$$

1次の最適性条件

$$\theta = \nabla\Omega(\pi)$$

$$\pi = [\nabla\Omega]^{-1}(\theta) = \psi^{-1}$$



凸共役としての最大エントロピーモデル

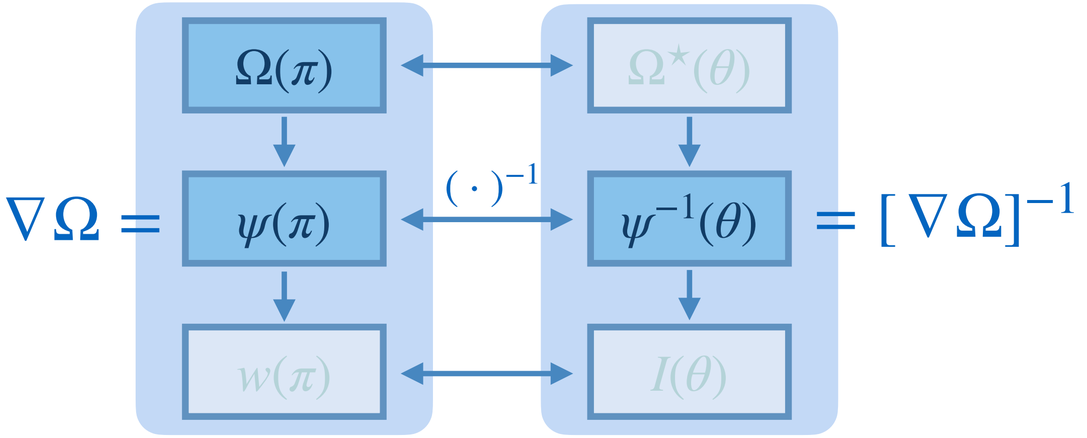
最大エントロピーモデル

$$\psi^{-1}(\theta) = \underset{\pi \in [0,1]}{\arg\max}\ \theta\pi - \Omega(\pi)$$

1次の最適性条件

$$\theta = \nabla\Omega(\pi)$$

$$\pi = [\nabla\Omega]^{-1}(\theta) = \psi^{-1}$$



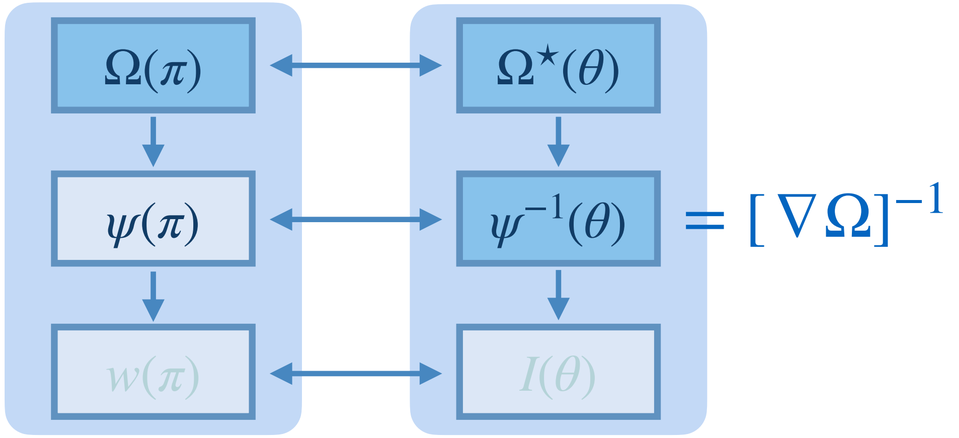
凸共役としての最大エントロピーモデル

最大エントロピーモデル

$$\psi^{-1}(\theta) = \arg\max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

$$\theta = \nabla\Omega(\pi)$$

$\theta = \nabla\Omega(\pi)$ を最大エントロピーモデルに代入



凸共役としての最大エントロピーモデル

最大エントロピーモデル

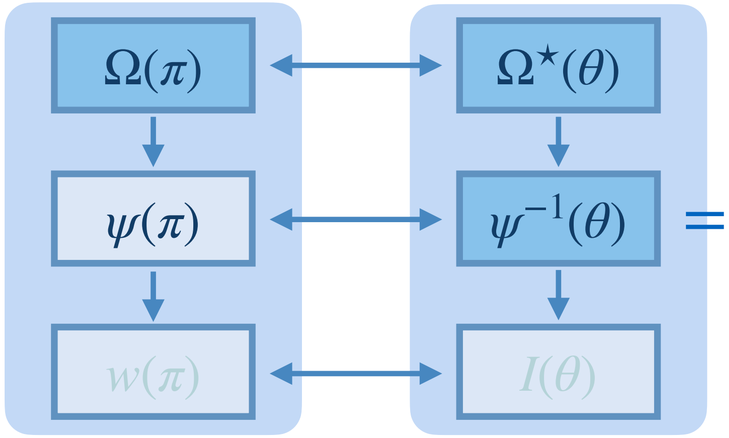
$$\psi^{-1}(\theta) = \arg\max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

$$\theta = \nabla\Omega(\pi)$$

$\theta = \nabla\Omega(\pi)$ を最大エントロピーモデルに代入

凸共役関数

$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$



凸共役としての最大エントロピーモデル

最大エントロピーモデル

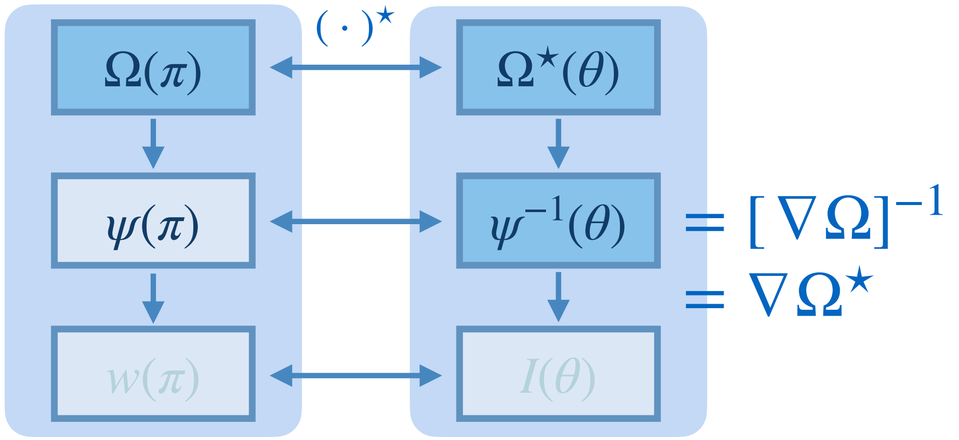
$$\psi^{-1}(\theta) = \arg\max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

$$\theta = \nabla\Omega(\pi)$$

$\theta = \nabla\Omega(\pi)$ を最大エントロピーモデルに代入

凸共役関数

$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$



凸共役としての最大エントロピーモデル

最大エントロピーモデル

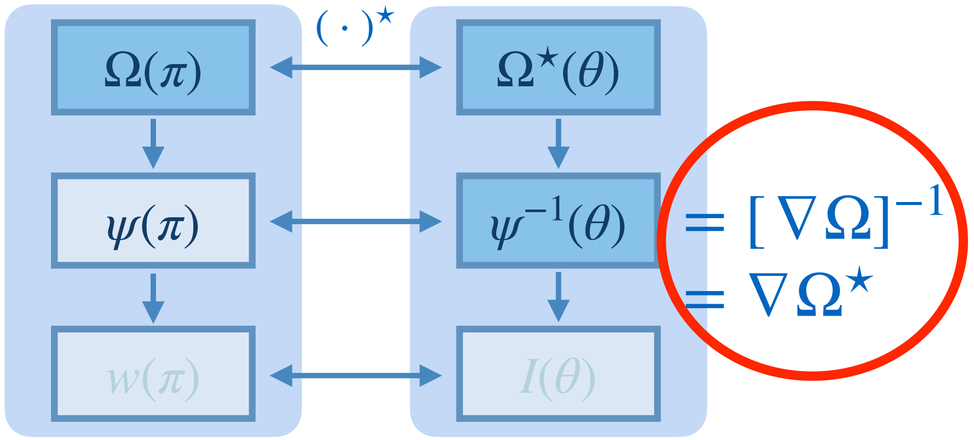
$$\psi^{-1}(\theta) = \arg\max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

$$\theta = \nabla\Omega(\pi)$$

$\theta = \nabla\Omega(\pi)$ を最大エントロピーモデルに代入

凸共役関数

$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$



関係式 $\nabla\Omega^\star = [\nabla\Omega]^{-1}$ の証明 (Danskin の定理)

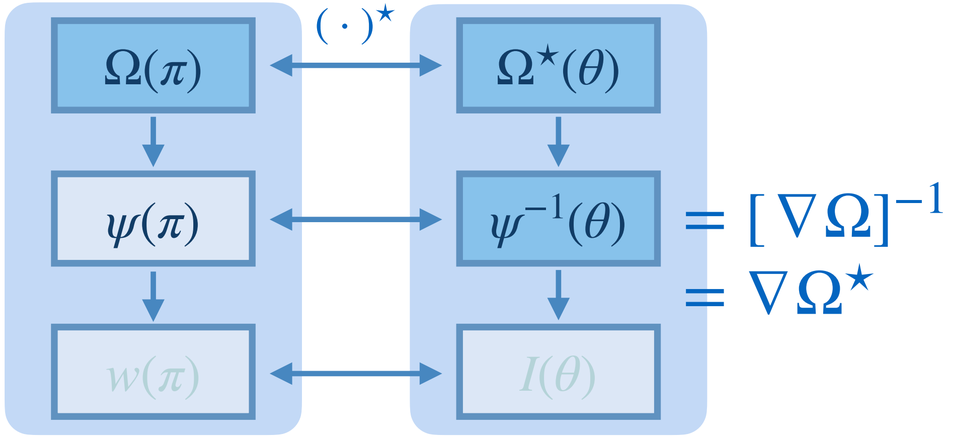
最大エントロピーモデル

$$\psi^{-1}(\theta) = \arg\max_{\pi\in[0,1]} \theta\pi - \Omega(\pi)$$

凸共役関数

$$\Omega^\star(\theta) = \max_{\pi\in[0,1]} \theta\pi - \Omega(\pi)$$

- **Fenchel-Young の不等式**: $\Omega(\pi) + \Omega^\star(\theta) \geq \theta\pi$ (定義より従う)



関係式 $\nabla\Omega^\star = [\nabla\Omega]^{-1}$ の証明 (Danskin の定理)

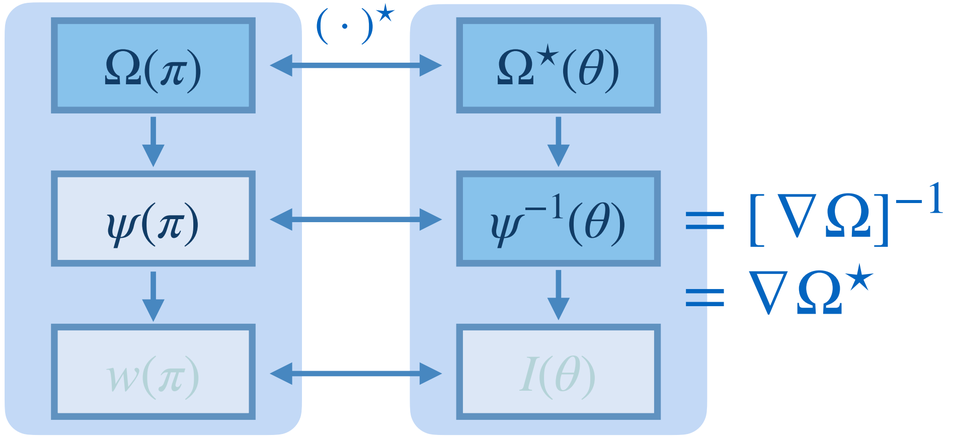
最大エントロピーモデル

$$\psi^{-1}(\theta) = \arg\max_{\pi\in[0,1]} \theta\pi - \Omega(\pi)$$

凸共役関数

$$\Omega^\star(\theta) = \max_{\pi\in[0,1]} \theta\pi - \Omega(\pi)$$

- **Fenchel-Young の不等式**: $\Omega(\pi) + \Omega^\star(\theta) \geq \theta\pi$ (定義より従う)
  - 等号成立条件: $\Omega^\star(\theta) = \theta\pi^* - \Omega(\pi^*)$ ただし $\pi^* = \psi^{-1}(\theta)$



関係式 $\nabla\Omega^\star = [\nabla\Omega]^{-1}$ の証明 (Danskin の定理)

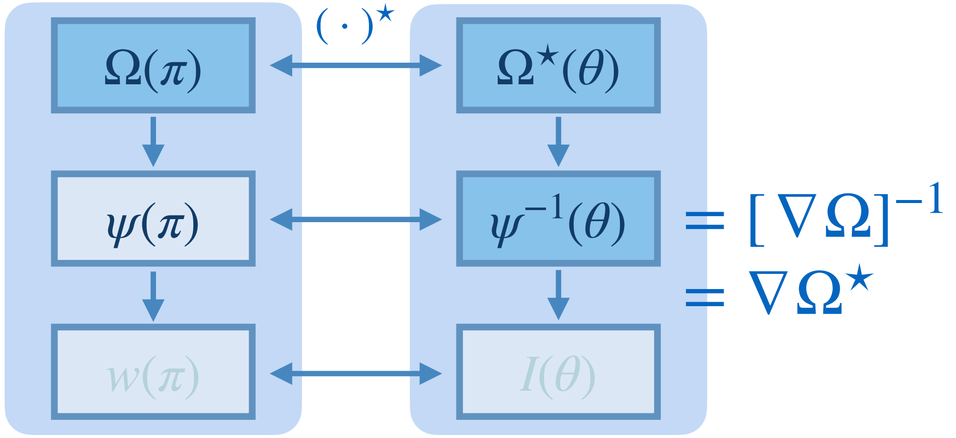
最大エントロピーモデル

$$\psi^{-1}(\theta) = \arg\max_{\pi\in[0,1]} \theta\pi - \Omega(\pi)$$

凸共役関数

$$\Omega^\star(\theta) = \max_{\pi\in[0,1]} \theta\pi - \Omega(\pi)$$

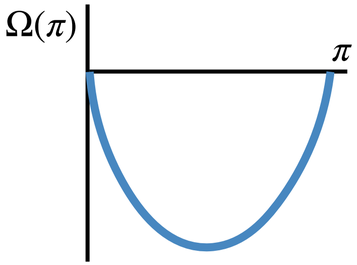
- **Fenchel-Young の不等式**: $\Omega(\pi) + \Omega^\star(\theta) \geq \theta\pi$ (定義より従う)
  - 等号成立条件: $\Omega^\star(\theta) = \theta\pi^* - \Omega(\pi^*)$ ただし $\pi^* = \psi^{-1}(\theta)$
  - 両辺を (θ について) 微分: $\nabla\Omega^\star(\theta) = \pi^* = [\nabla\Omega]^{-1}(\theta)$

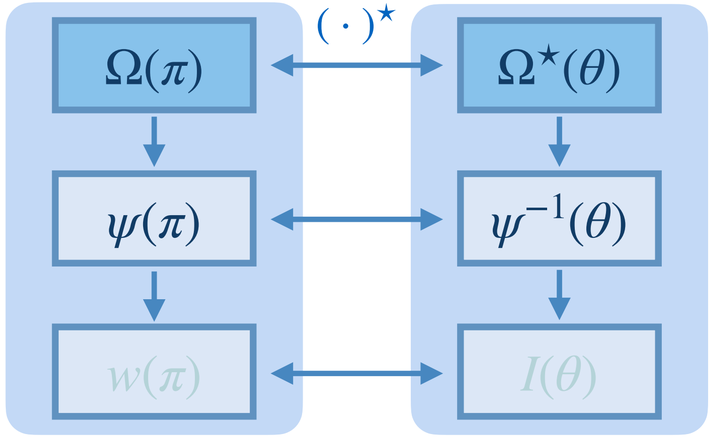


凸共役の意味づけ

- 再訪: ロジスティック回帰の場合

$$\Omega(\pi) = \pi \ln \pi + (1-\pi)\ln(1-\pi)$$

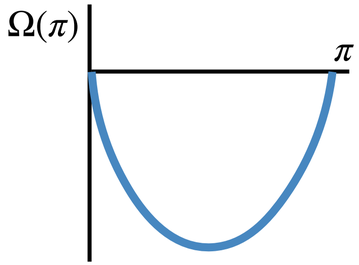




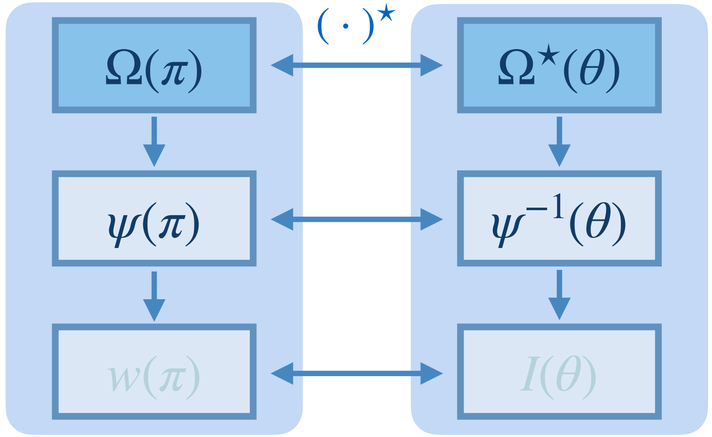
凸共役の意味づけ

- 再訪: ロジスティック回帰の場合

$$\Omega(\pi) = \pi \ln \pi + (1-\pi)\ln(1-\pi)$$



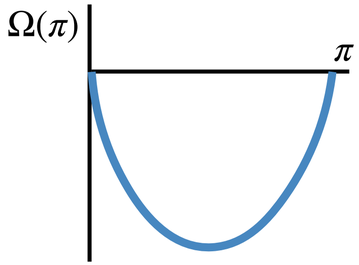
$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$



凸共役の意味づけ

- 再訪: ロジスティック回帰の場合

$$\Omega(\pi) = \pi \ln \pi + (1-\pi)\ln(1-\pi)$$

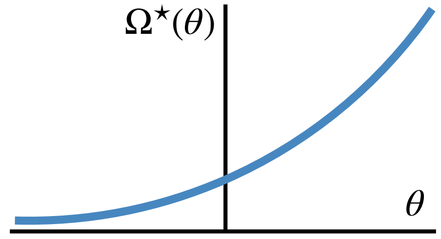


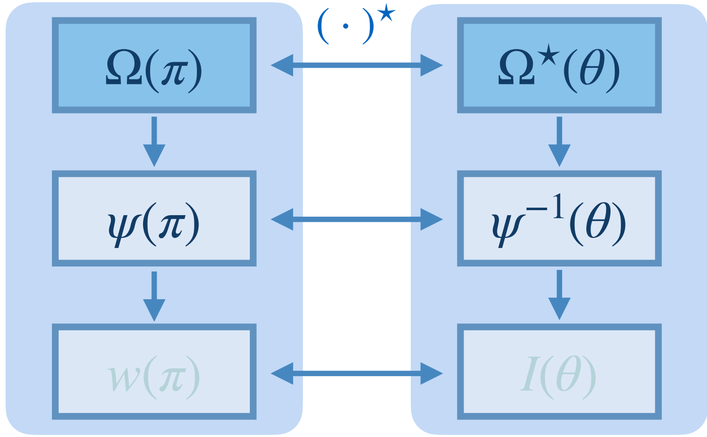
$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

Log-sum-exp 関数

(対数分配関数、自由エネルギー、etc.)

$$\Omega^\star(\theta) = \ln(1+\exp(\theta))$$

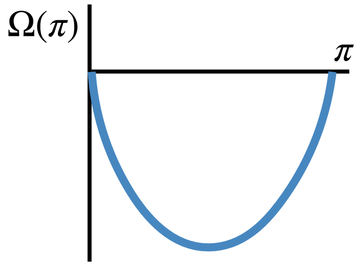




凸共役の意味づけ

- 再訪: ロジスティック回帰の場合

$$\Omega(\pi) = \pi \ln \pi + (1-\pi)\ln(1-\pi)$$

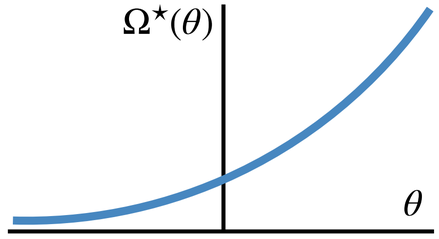


$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

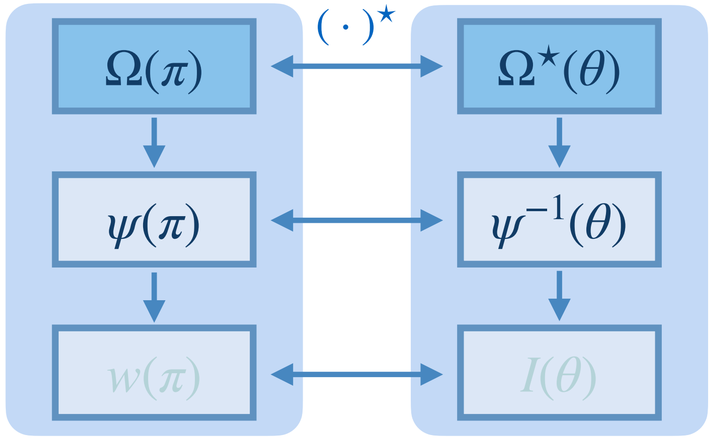
Log-sum-exp 関数

(対数分配関数、自由エネルギー、etc.)

$$\Omega^\star(\theta) = \ln(1+\exp(\theta))$$



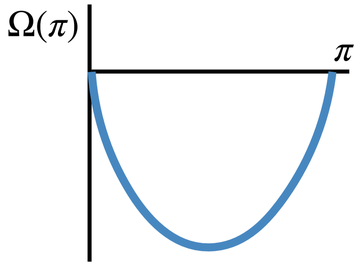
- Log-sum-exp 関数から損失関数が自然に誘導



凸共役の意味づけ

- 再訪: ロジスティック回帰の場合

$$\Omega(\pi) = \pi \ln \pi + (1-\pi)\ln(1-\pi)$$

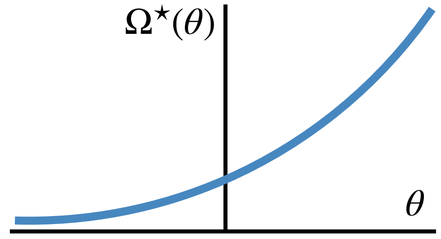


$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

Log-sum-exp 関数

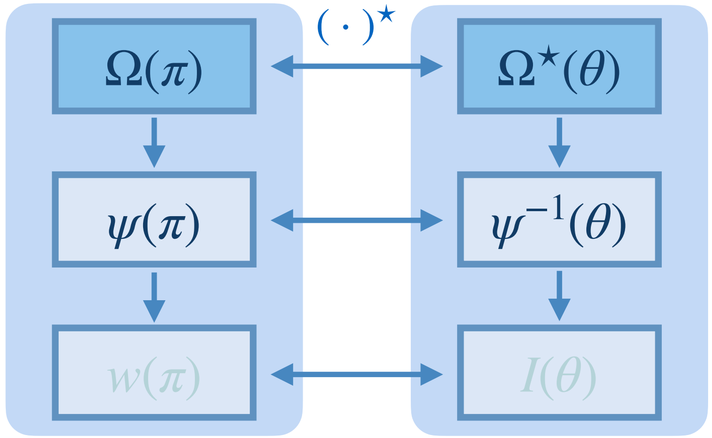
(対数分配関数、自由エネルギー、etc.)

$$\Omega^\star(\theta) = \ln(1+\exp(\theta))$$



- Log-sum-exp 関数から損失関数が自然に誘導

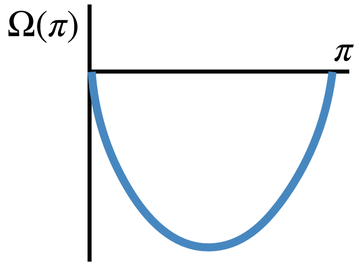
$$\Omega^\star(\theta) - \theta = \ln(1+\exp(-\theta))$$



凸共役の意味づけ

- 再訪: ロジスティック回帰の場合

$$\Omega(\pi) = \pi \ln \pi + (1-\pi)\ln(1-\pi)$$

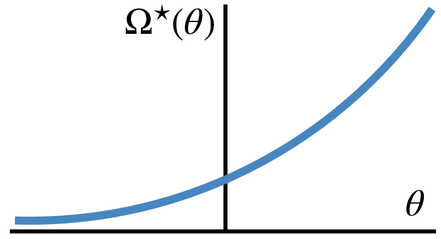


$$\Omega^\star(\theta) = \max_{\pi \in [0,1]} \theta\pi - \Omega(\pi)$$

Log-sum-exp 関数

(対数分配関数、自由エネルギー、etc.)

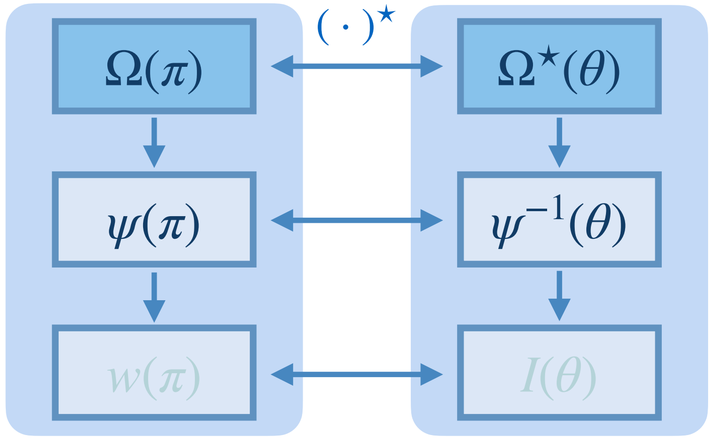
$$\Omega^\star(\theta) = \ln(1+\exp(\theta))$$



- Log-sum-exp 関数から損失関数が自然に誘導

$$\Omega^\star(\theta) - \theta = \ln(1+\exp(-\theta))$$

ロジスティック損失

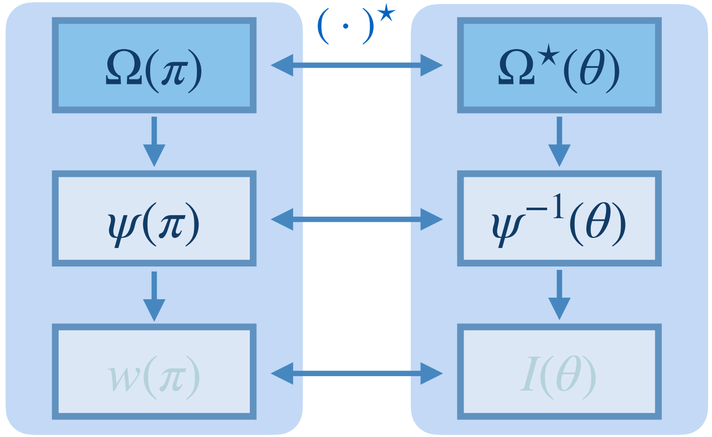


凸共役から Bregman ダイバージェンスが誘導

- 凸関数 Ω が誘導する Bregman ダイバージェンス

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi \quad \text{(主双対型ダイバージェンス)}$$

  - 例: Total variation distance, L2 distance, Mahalanobis distance



凸共役から Bregman ダイバージェンスが誘導

- 凸関数 Ω が誘導する Bregman ダイバージェンス

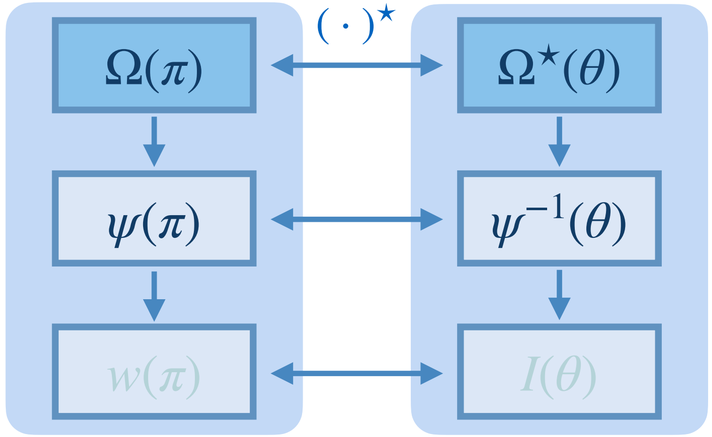
$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi \quad \text{(主双対型ダイバージェンス)}$$

  - 例: Total variation distance, L2 distance, Mahalanobis distance

- 主空間のダイバージェンス (proper loss の観点): 逆リンク関数を代入

$$D_\Omega(\pi;\theta) = \pi \ln\left(\frac{\pi}{\pi^*}\right) + (1-\pi)\ln\left(\frac{1-\pi}{1-\pi^*}\right) \quad \text{ただし } \pi^* = \psi^{-1}(\theta)$$

主変数型ダイバージェンス



凸共役から Bregman ダイバージェンスが誘導

- 凸関数 Ω が誘導する Bregman ダイバージェンス

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi \quad \text{(主双対型ダイバージェンス)}$$

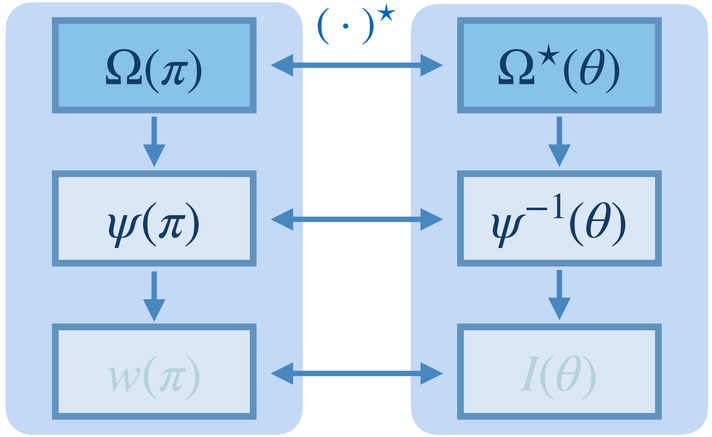
  - 例: Total variation distance, L2 distance, Mahalanobis distance

- 主空間のダイバージェンス (proper loss の観点): 逆リンク関数を代入

$$D_\Omega(\pi;\theta) = \pi \ln\left(\frac{\pi}{\pi^*}\right) + (1-\pi)\ln\left(\frac{1-\pi}{1-\pi^*}\right) \quad \text{ただし } \pi^* = \psi^{-1}(\theta)$$

主変数型ダイバージェンス

- 主双対ダイバージェンスから損失関数が誘導



凸共役から Bregman ダイバージェンスが誘導

- 凸関数 Ω が誘導する Bregman ダイバージェンス

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi \quad \text{(主双対型ダイバージェンス)}$$

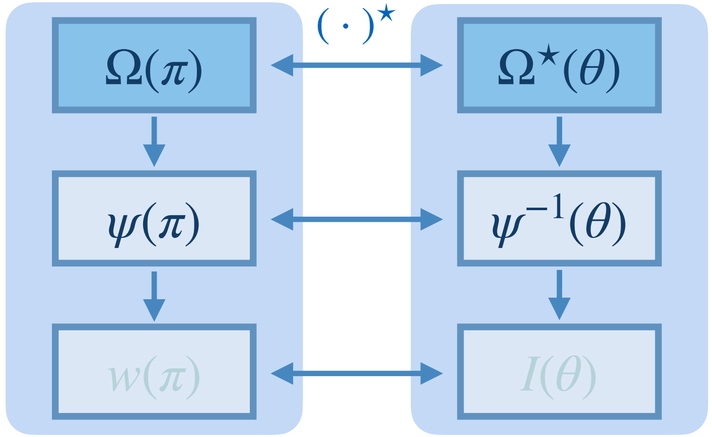
  - 例: Total variation distance, L2 distance, Mahalanobis distance

- 主空間のダイバージェンス (proper loss の観点): 逆リンク関数を代入

$$\underbrace{D_\Omega(\pi;\theta) = \pi \ln\left(\frac{\pi}{\pi^*}\right) + (1-\pi)\ln\left(\frac{1-\pi}{1-\pi^*}\right)}_{\text{主変数型ダイバージェンス}} \quad \text{ただし } \pi^* = \psi^{-1}(\theta)$$

- 主双対ダイバージェンスから損失関数が誘導

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi$$



凸共役から Bregman ダイバージェンスが誘導

- 凸関数 Ω が誘導する Bregman ダイバージェンス

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi \quad \text{(主双対型ダイバージェンス)}$$

  - 例: Total variation distance, L2 distance, Mahalanobis distance

- 主空間のダイバージェンス (proper loss の観点): 逆リンク関数を代入

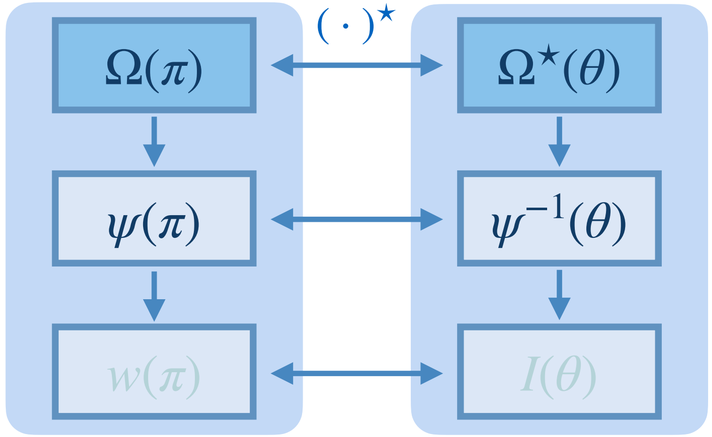
$$D_\Omega(\pi;\theta) = \pi \ln\left(\frac{\pi}{\pi^*}\right) + (1-\pi)\ln\left(\frac{1-\pi}{1-\pi^*}\right) \quad \text{ただし } \pi^* = \psi^{-1}(\theta)$$

主変数型ダイバージェンス

- 主双対ダイバージェンスから損失関数が誘導

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi$$

↓ $\pi = 1$ (one-hot ラベル) を代入



凸共役から Bregman ダイバージェンスが誘導

- 凸関数 Ω が誘導する Bregman ダイバージェンス

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi \quad \text{(主双対型ダイバージェンス)}$$

  - 例: Total variation distance, L2 distance, Mahalanobis distance

- 主空間のダイバージェンス (proper loss の観点): 逆リンク関数を代入

$$D_\Omega(\pi;\theta) = \pi \ln\left(\frac{\pi}{\pi^*}\right) + (1-\pi)\ln\left(\frac{1-\pi}{1-\pi^*}\right) \quad \text{ただし } \pi^* = \psi^{-1}(\theta)$$

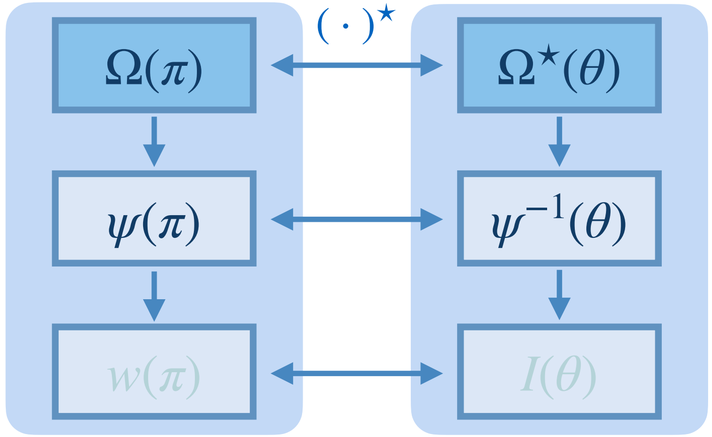
主変数型ダイバージェンス

- 主双対ダイバージェンスから損失関数が誘導

$$D_\Omega(\pi;\theta) = \Omega(\pi) + \Omega^\star(\theta) - \theta\pi$$

$\pi = 1$ (one-hot ラベル) を代入

$$= \Omega^\star(\theta) - \theta$$

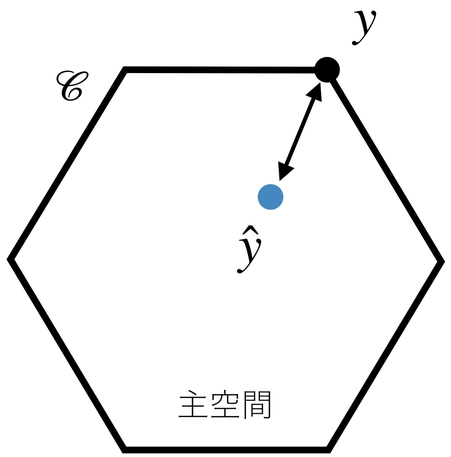


主双対で定義される損失関数

定義 (Fenchel-Young loss). 凸関数 $\Omega : \mathscr{C} \to \mathbb{R}$ に対して次のように Fenchel–Young 損失を定義:

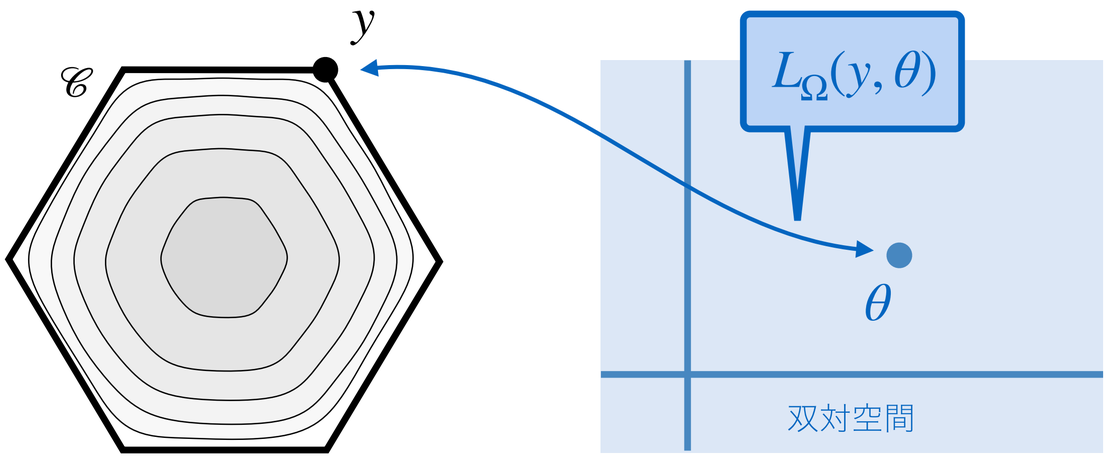
$$L_\Omega(y, \theta) := \Omega(y) + \Omega^\star(\theta) - \langle \theta, y \rangle$$

(一般の台集合 $\mathscr{C}$ に対して定義)



主空間ダイバージェンス (proper loss) は
主空間内の 2点を比較

⇒ **リンク関数の選択に恣意性あり** ☹



(主空間に Ω で計量を導入)

主双対ダイバージェンス (FY-loss) は
主変数と双対変数を直接比較

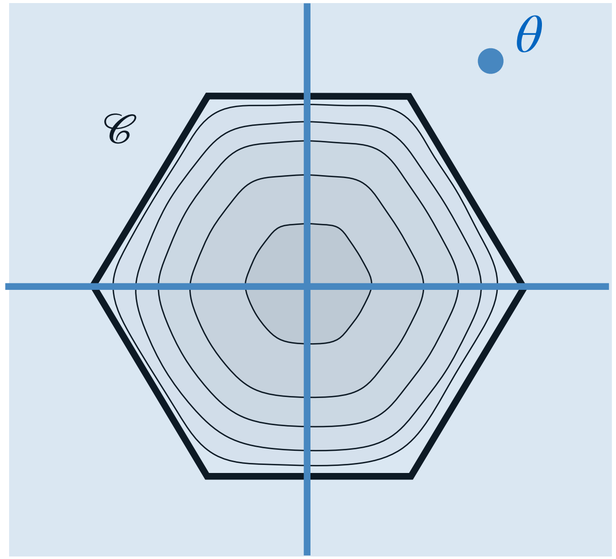
⇒ **自然なリンク関数が自動的に選ばれる** ☺

主双対で定義される損失関数

定義 (Fenchel-Young loss). 凸関数 $\Omega : \mathscr{C} \to \mathbb{R}$ に対して次のように Fenchel–Young 損失を定義:

$$L_\Omega(y, \theta) := \Omega(y) + \Omega^\star(\theta) - \langle \theta, y \rangle$$

主双対の関係の直感



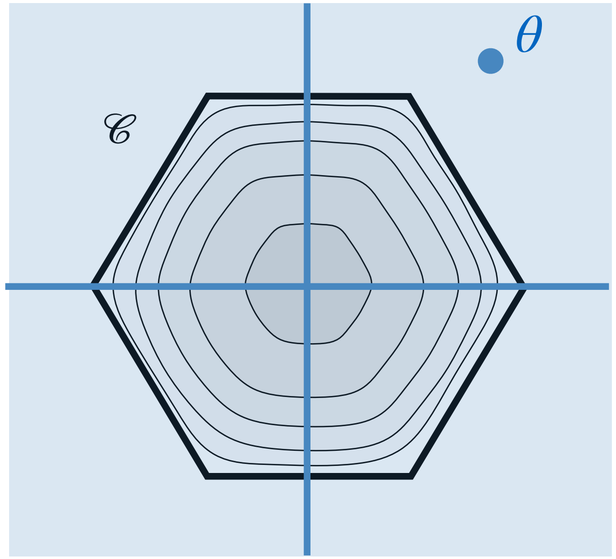
① 主空間・双対空間を重ね合わせる

主双対で定義される損失関数

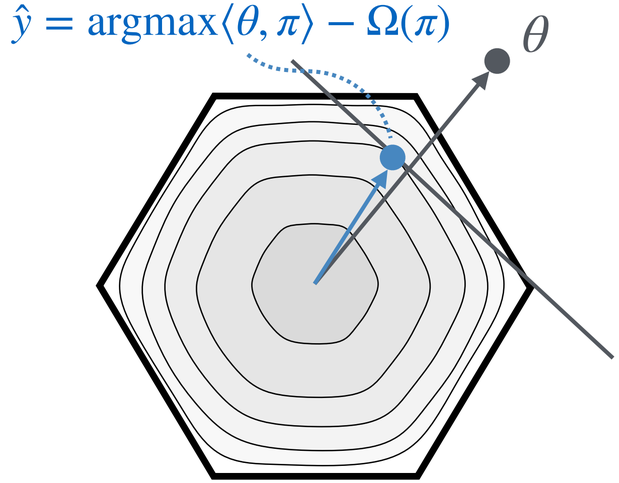
定義 (Fenchel-Young loss). 凸関数 $\Omega : \mathscr{C} \to \mathbb{R}$ に対して次のように Fenchel–Young 損失を定義:

$$L_\Omega(y, \theta) := \Omega(y) + \Omega^\star(\theta) - \langle \theta, y \rangle$$

主双対の関係の直感



① 主空間・双対空間を重ね合わせる



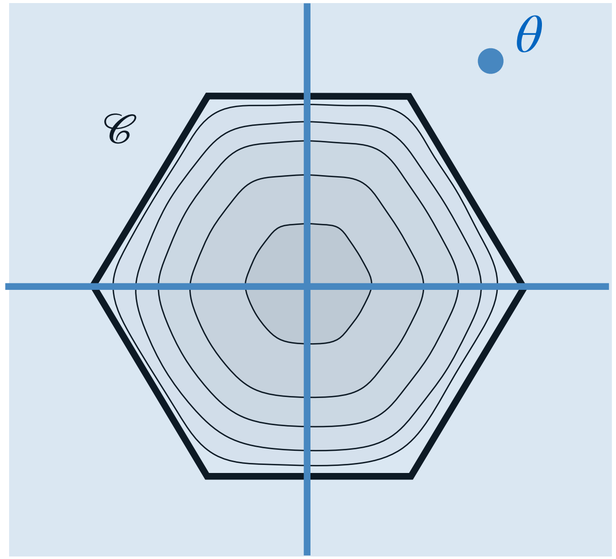
② 双対変数と類似度の高い $\hat{y}$ を出力

主双対で定義される損失関数

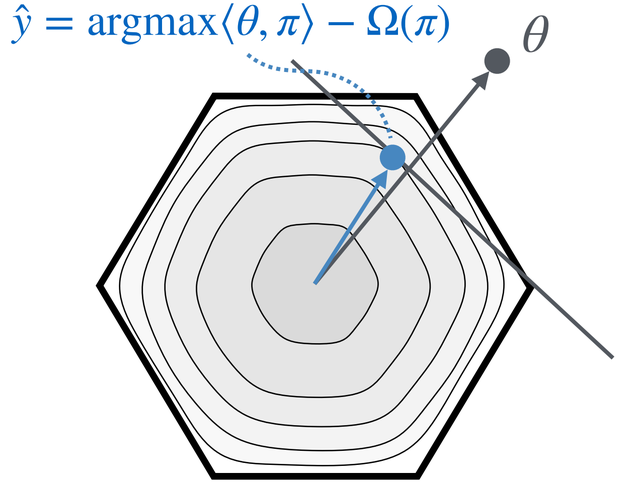
定義 (Fenchel-Young loss). 凸関数 $\Omega : \mathscr{C} \to \mathbb{R}$ に対して次のように Fenchel–Young 損失を定義:

$$L_\Omega(y, \theta) := \Omega(y) + \Omega^\star(\theta) - \langle \theta, y \rangle$$

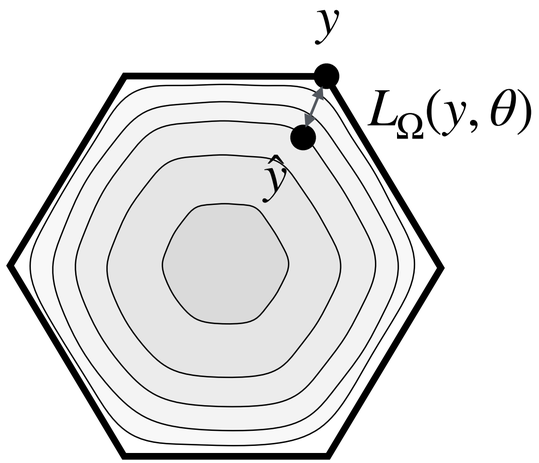
主双対の関係の直感



① 主空間・双対空間を重ね合わせる



② 双対変数と類似度の高い $\hat{y}$ を出力



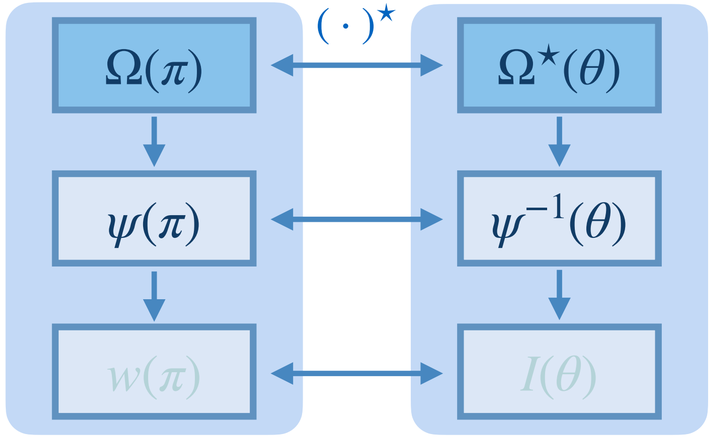
③ 主空間で距離を測る

主双対で定義される損失関数

定義 (Fenchel-Young loss). 凸関数 $\Omega : \mathscr{C} \to \mathbb{R}$ に対して次のように Fenchel–Young 損失を定義:

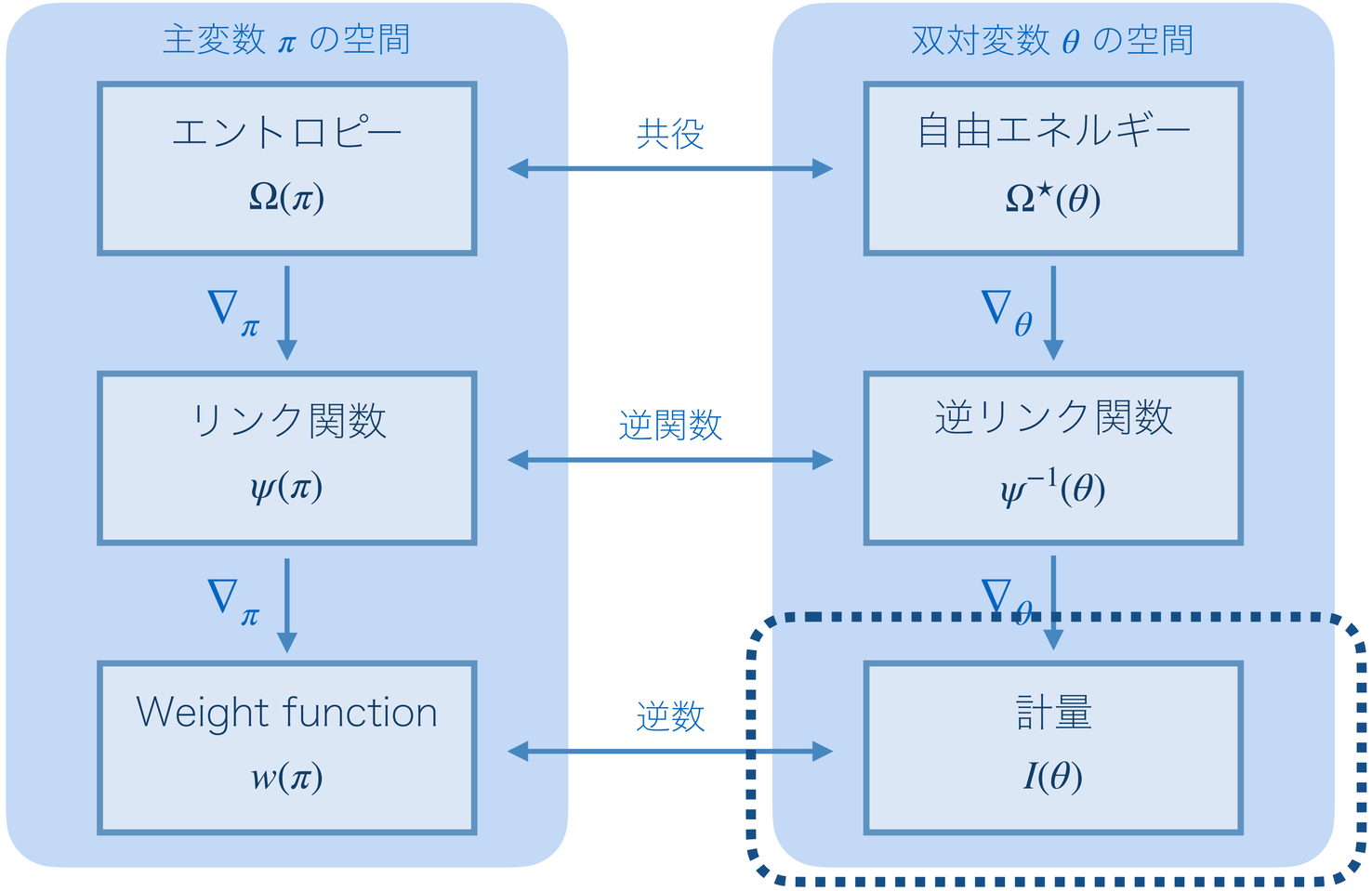
$$L_\Omega(y, \theta) := \Omega(y) + \Omega^\star(\theta) - \langle \theta, y \rangle$$

- 重要な性質
  - 双対変数 θ に関する**凸性**
  - **勾配** = 残差 $\psi^{-1}(\theta) - y \in \partial L_\Omega(y, \theta)$
  - 予測が正しいときのみ損失は**最適**: $L_\Omega(y, \theta) = 0 \iff y = \psi^{-1}(\theta)$
- Proper loss に対するメリット
  - リンク関数が凸関数から自然に定まる
  - 双対空間内での制約なし & 凸最適化
  - 台集合 $\mathscr{C}$ が 1次元確率単体よりはるかに複雑でも自然に損失を設計可能



このパートの目標

- 以下の関係の理解



二次最適化による学習

- ロジスティック回帰: 最尤推定 = Fenchel–Young 損失の最小化

$$\min_{\theta} \; \Omega^\star(\theta) - \theta$$

(各点の) ロジスティック損失 $L_\Omega(y, \theta)$

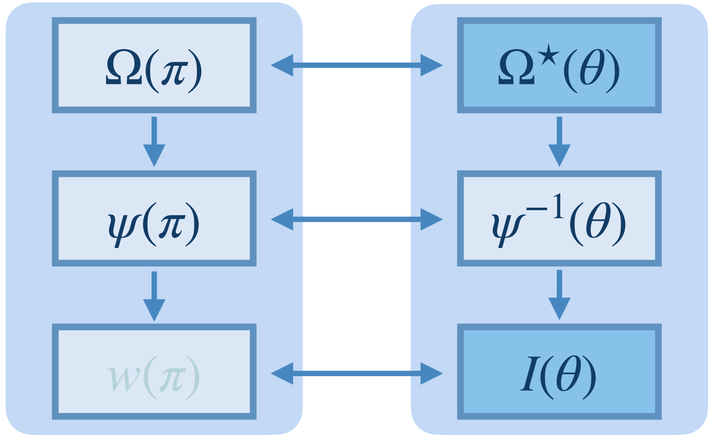
- ニュートン法による最適化 (Fisher スコア法)

$$\theta' \leftarrow \theta - \left[\nabla L^2(\theta)\right]^{-1} \nabla L(\theta)$$

ロジスティック損失の 2階微分

= Fisher 情報量 $= \nabla^2 \Omega^\star(\theta)$

- 二次最適化のまとめ
  - ニュートン法:　目的関数の 2階微分で更新方向を修正
  - 鏡像勾配法:　凸共役の 2階微分で更新方向を修正



鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

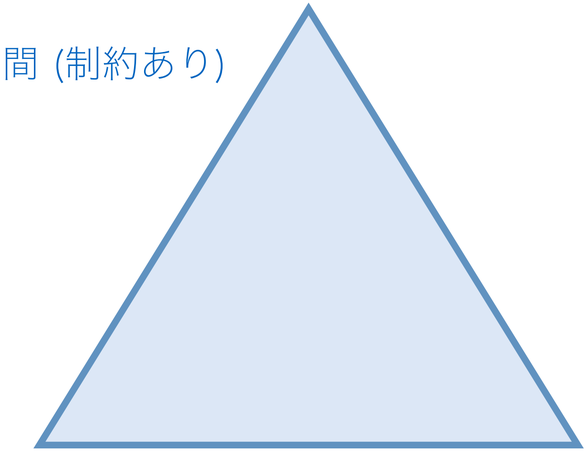
$\pi \in \triangle^K$

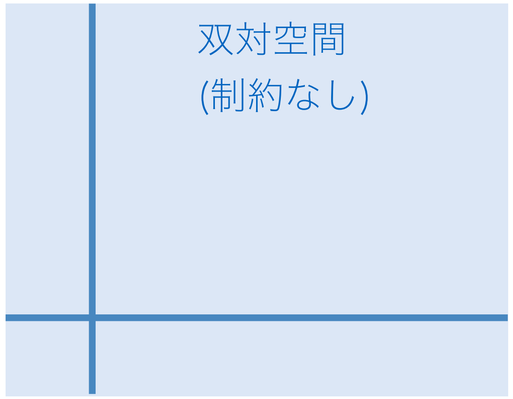
主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

双対変数

$\theta \in \mathbb{R}^K$

主空間 (制約あり)





鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

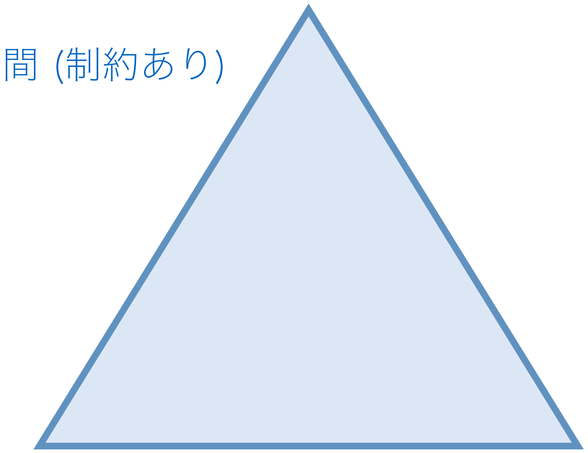
$$\pi \in \triangle^K$$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

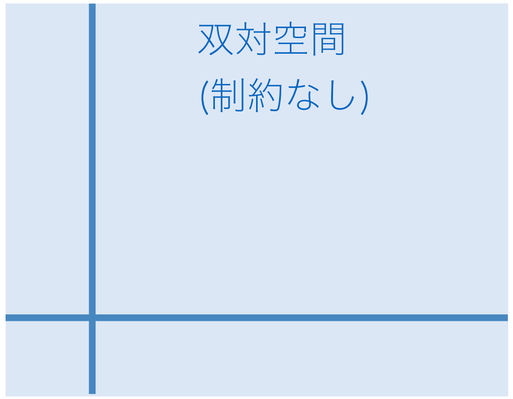
双対変数

$$\theta \in \mathbb{R}^K$$

主空間 (制約あり)



双対空間
(制約なし)



近接更新: $\underset{\pi \in \triangle^K}{\arg\min} \left\{ -\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t) \right\}$

鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

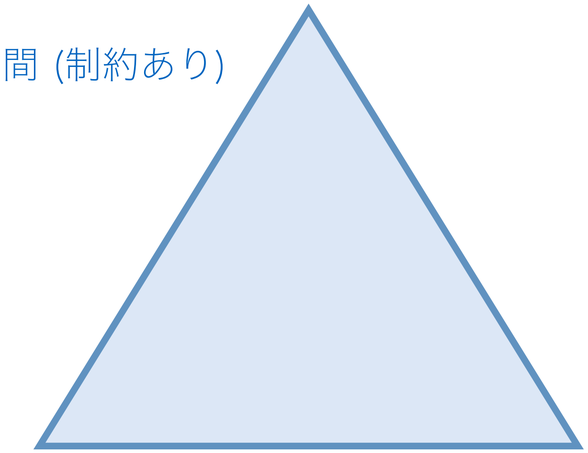
$\pi \in \triangle^K$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

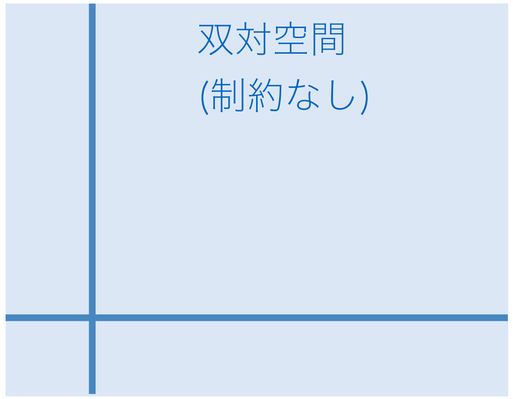
双対変数

$\theta \in \mathbb{R}^K$

主空間 (制約あり)



双対空間
(制約なし)



近接更新: $\arg\min_{\pi \in \triangle^K} \left\{ -\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t) \right\}$

更新が遠くなりすぎないように

鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

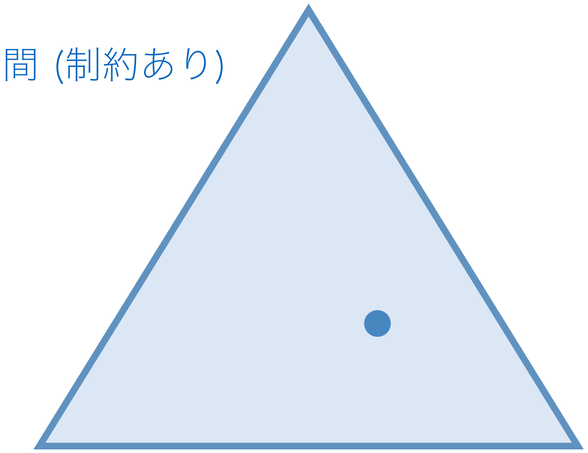
$$\pi \in \triangle^K$$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

双対変数

$$\theta \in \mathbb{R}^K$$

主空間 (制約あり)



双対空間
(制約なし)

近接更新: $\arg\min_{\pi \in \triangle^K} \left\{ -\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t) \right\}$

更新が遠くなりすぎないように

鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

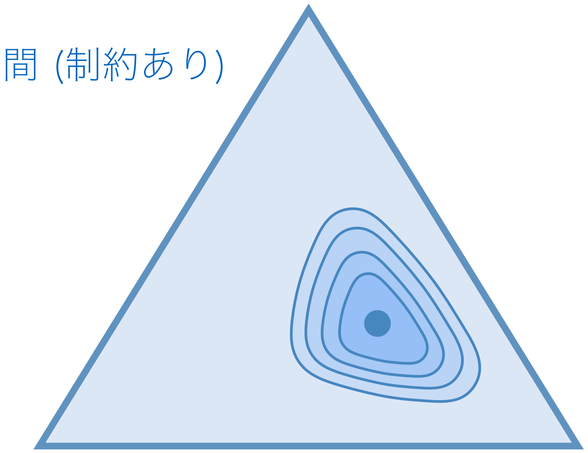
$$\pi \in \triangle^K$$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

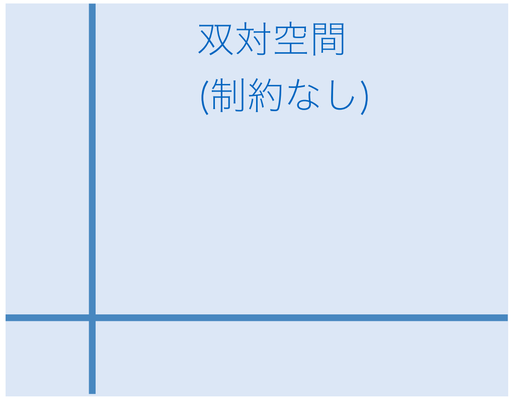
双対変数

$$\theta \in \mathbb{R}^K$$

主空間 (制約あり)



双対空間
(制約なし)



近接更新: $\arg\min_{\pi \in \triangle^K} \left\{ -\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t) \right\}$

更新が遠くなりすぎないように

鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

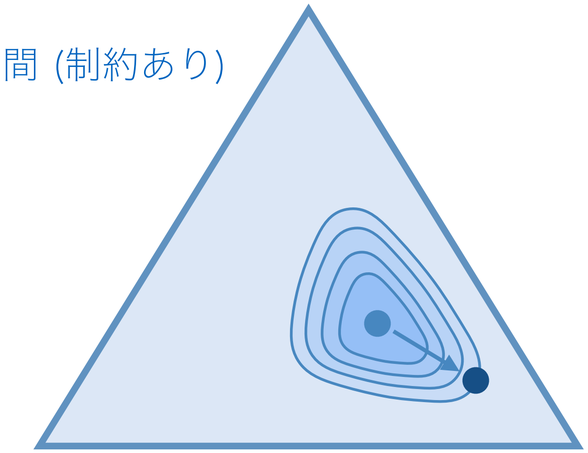
$$\pi \in \triangle^K$$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

双対変数

$$\theta \in \mathbb{R}^K$$

主空間 (制約あり)



双対空間
(制約なし)

近接更新: $\underset{\pi \in \triangle^K}{\arg\min} \left\{ -\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t) \right\}$

更新が遠くなりすぎないように

鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

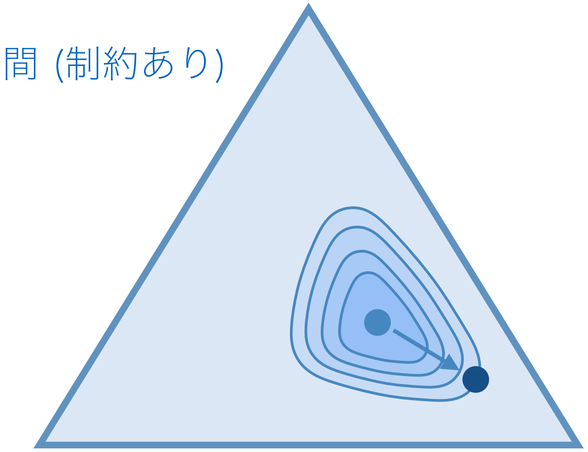
$$\pi \in \triangle^K$$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

双対変数

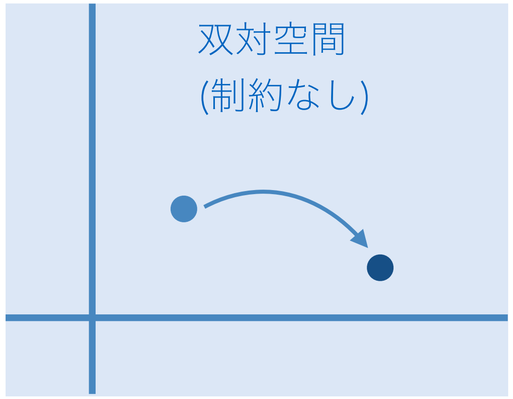
$$\theta \in \mathbb{R}^K$$

主空間 (制約あり)



近接更新: $\underset{\pi \in \triangle^K}{\arg\min} \left\{ -\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t) \right\}$

更新が遠くなりすぎないように



2次の更新: $-[\nabla L^2(\theta)]^{-1} \nabla L(\theta)$

鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

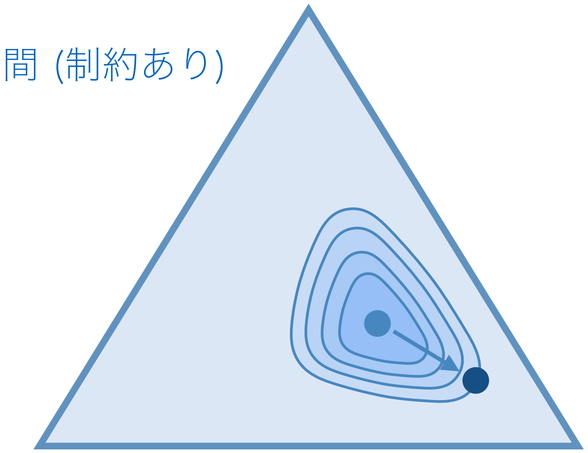
$\pi \in \triangle^K$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

双対変数

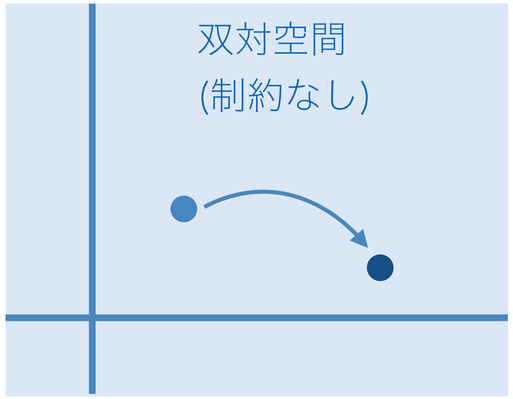
$\theta \in \mathbb{R}^K$

主空間 (制約あり)



近接更新: $\arg\min_{\pi \in \triangle^K} \{-\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t)\}$

更新が遠くなりすぎないように



2次の更新: $-[\nabla L^2(\theta)]^{-1} \nabla L(\theta)$

- 双対最適化は制約なしで容易 (である場合が多い)
- 鏡像勾配法の観点: 主空間での近接更新 = 双対空間での 2次最適化 (証明は割愛)

鏡像勾配法 (Mirror descent) （詳細は省略）

主変数

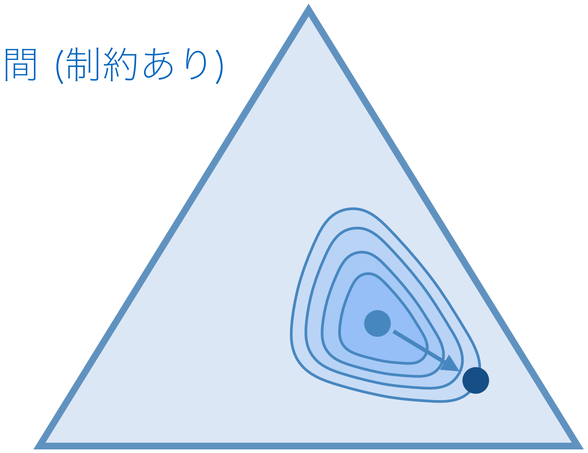
$\pi \in \triangle^K$

主双対変換 $\sigma(\theta_k) = \dfrac{\exp(\theta_k)}{\sum_{j=1}^{K} \exp(\theta_j)}$

双対変数

$\theta \in \mathbb{R}^K$

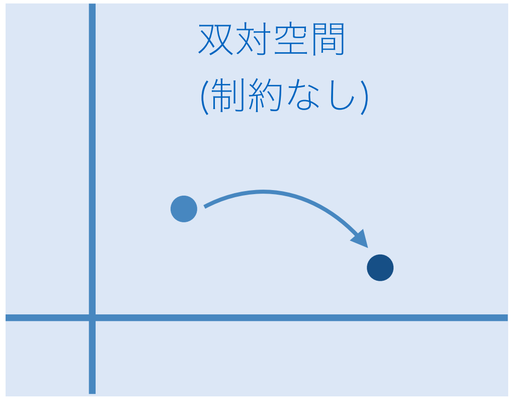
主空間 (制約あり)



★Point

MD は主空間で D_Ω を最小化

双対空間
(制約なし)



近接更新: $\arg\min_{\pi \in \triangle^K} \left\{ -\langle \theta, \pi \rangle + D_\Omega(\pi \| \pi_t) \right\}$

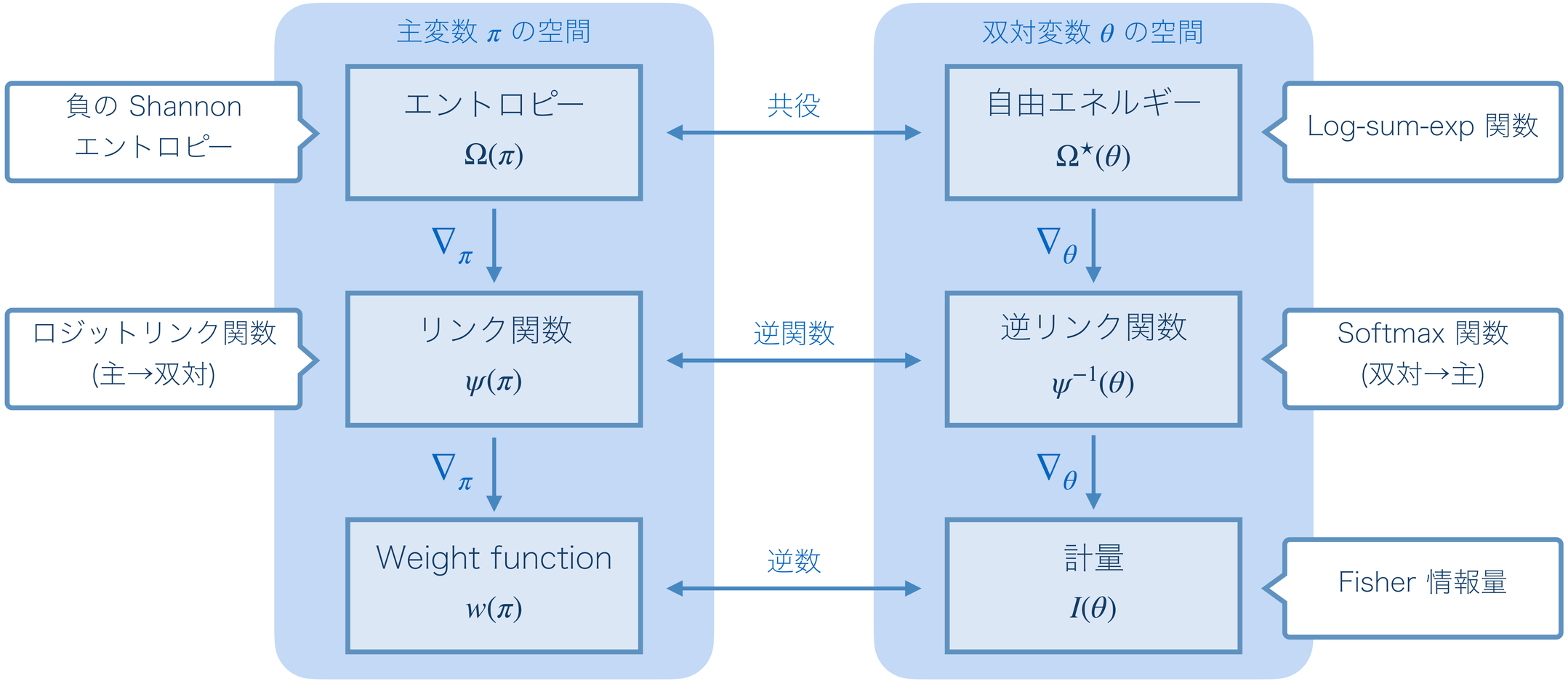
更新が遠くなりすぎないように

2次の更新: $-[\nabla L^2(\theta)]^{-1} \nabla L(\theta)$

- 双対最適化は制約なしで容易 (である場合が多い)
- 鏡像勾配法の観点: 主空間での近接更新 = 双対空間での 2次最適化 (証明は割愛)

このパートの目標

- 以下の関係の理解



機械学習と凸共役の交わり（目次）

前半

- 二値分類問題: 主空間の観点から
- 二値分類問題: 双対空間の観点から
- **応用: 非対称リンク関数を用いた二値応答回帰**

  Bao & Sugiyama (2021) “Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation”

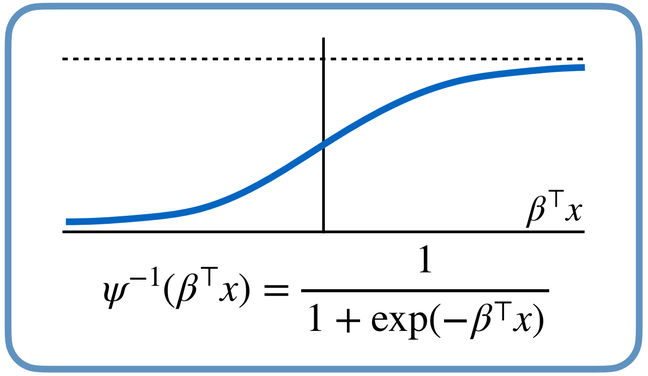
後半

- 最適輸送問題: 双対空間の観点から
- 応用: q-指数分布を用いたスパース最適輸送

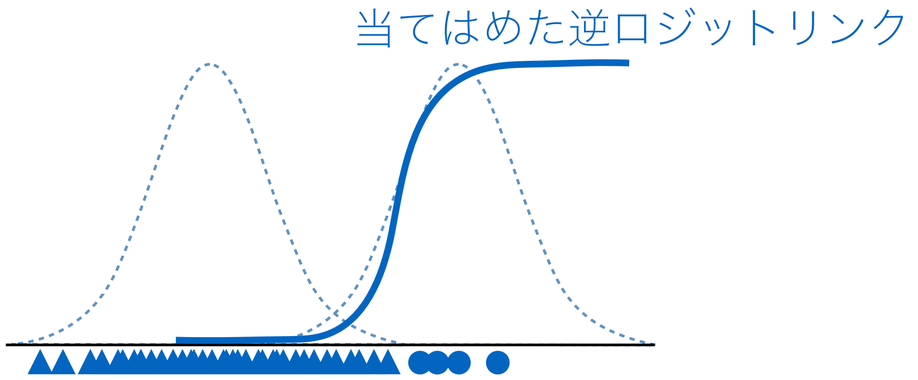
  Bao & Sakaue (2022) “Sparse Regularized Optimal Transport with Deformed q-Entropy”
- その他の問題

ロジスティック回帰の弱点 ①：クラス不均衡

- 逆リンク関数によって $\mathbb{P}(Y=1\,|\,x)$ をモデリング
  - ロジットリンク関数は対称



- 希少クラスに対して学習がうまくいかないことがある
  - 頻出クラスに重きをおくため、希少クラスが無視される傾向に



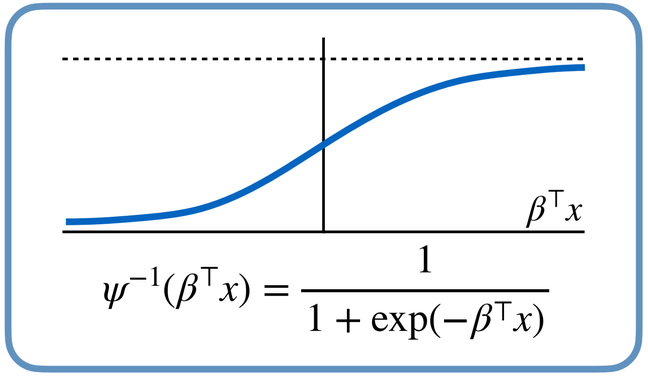
- (真の) クラス事前確率 $\mathbb{P}(Y=1)$ が 0.5 から遠くなるほどロジスティック回帰のバイアスは増大

[King & Zeng 2001]

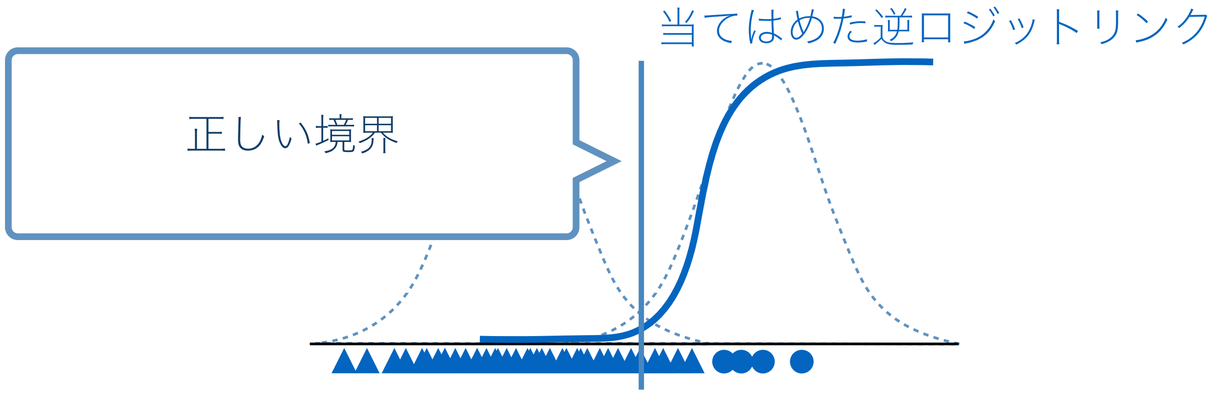
King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137-163.

ロジスティック回帰の弱点 ①：クラス不均衡

- 逆リンク関数によって $\mathbb{P}(Y=1\,|\,x)$ をモデリング
  - ロジットリンク関数は対称



- 希少クラスに対して学習がうまくいかないことがある
  - 頻出クラスに重きをおくため、希少クラスが無視される傾向に



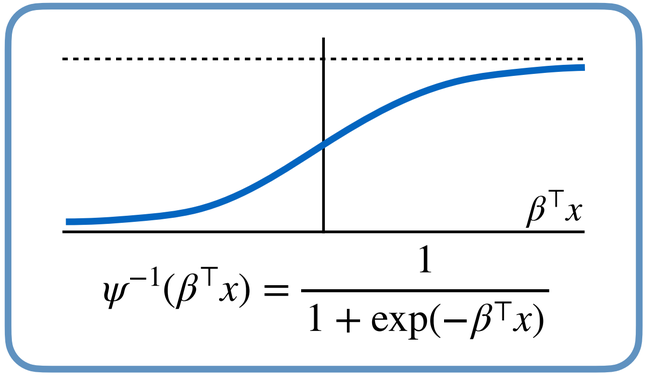
- (真の) クラス事前確率 $\mathbb{P}(Y=1)$ が 0.5 から遠くなるほどロジスティック回帰のバイアスは増大

[King & Zeng 2001]

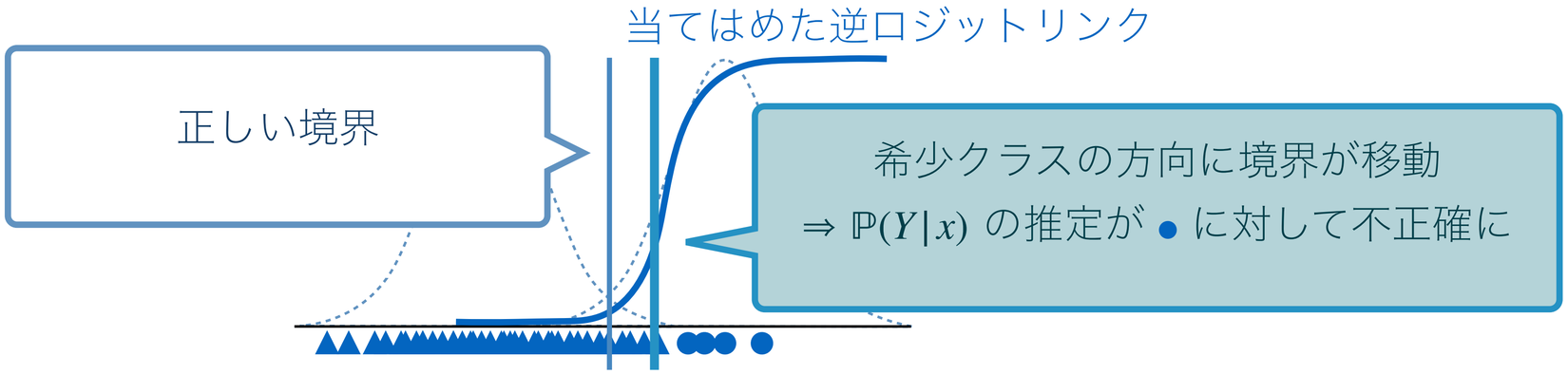
King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137-163.

ロジスティック回帰の弱点 ①：クラス不均衡

- 逆リンク関数によって $\mathbb{P}(Y=1|x)$ をモデリング
  - ロジットリンク関数は対称



- 希少クラスに対して学習がうまくいかないことがある
  - 頻出クラスに重きをおくため、希少クラスが無視される傾向に



- (真の) クラス事前確率 $\mathbb{P}(Y=1)$ が 0.5 から遠くなるほどロジスティック回帰のバイアスは増大

[King & Zeng 2001]

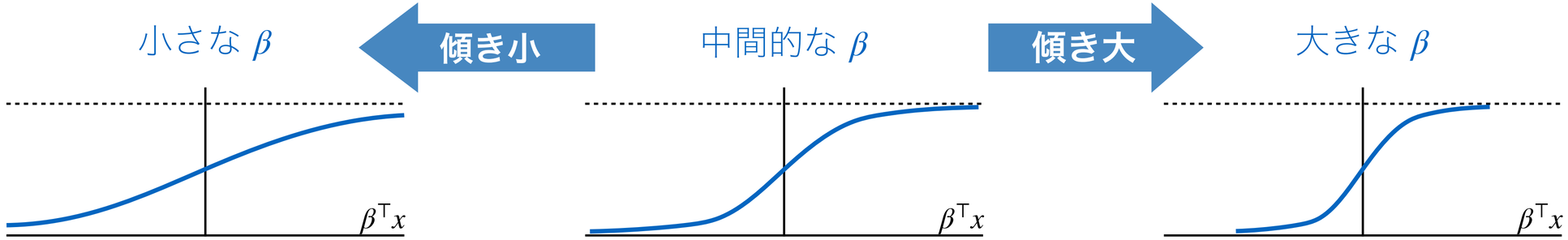
King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137-163.

ロジスティック回帰の弱点 ②：モデル誤設定

- ロジスティック回帰ではベルヌーイモデルを仮定

$$Y|x \sim \text{Bernoulli}(\mu) \quad \text{where} \quad \mu = \psi^{-1}(\beta_*^\top x) := \frac{1}{1+\exp(-\beta_*^\top x)}$$

- リンク関数自体が誤設定されている可能性あり

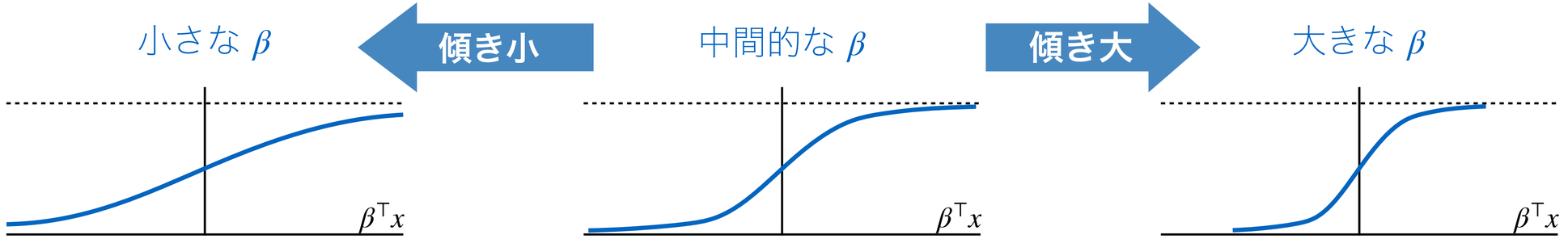


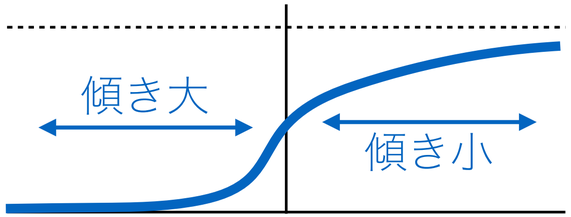
ロジスティック回帰の弱点 ②：モデル誤設定

- ロジスティック回帰ではベルヌーイモデルを仮定

$$Y|x \sim \text{Bernoulli}(\mu) \quad \text{where} \quad \mu = \psi^{-1}(\beta_*^\top x) := \frac{1}{1+\exp(-\beta_*^\top x)}$$

- リンク関数自体が誤設定されている可能性あり



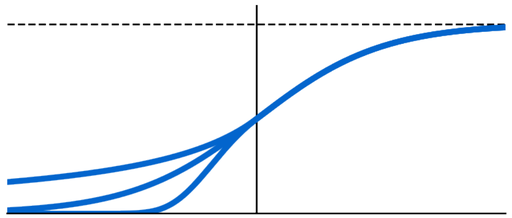


真のリンク関数が歪んでいる場合は
β の傾きを変えるだけでは不十分

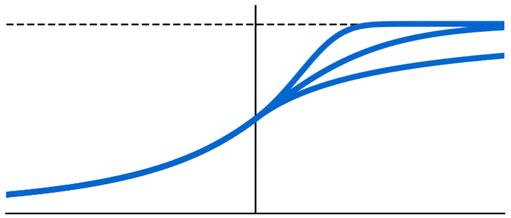
解決策: より柔軟なリンク関数

アイデア: 柔軟な歪度を取りうるリンク関数を設計（→ モデル選択を行う）

- 一般化ロジットリンク [Stukel JASA1988]
  - 2つのハイパーパラメータがそれぞれ正・負の歪度を調整

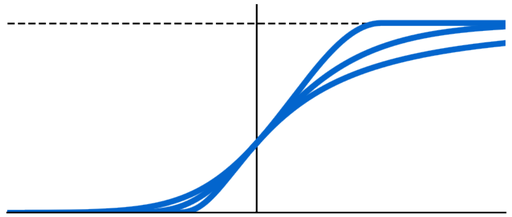


負の領域の歪度を調整



正の領域の歪度を調整

- GEV 分布 (一般化極値分布) の累積分布関数 [Wang & Dey AoAS2010]



$$F_\xi(x) = \exp\left((1+\xi x)_+^{-1/\xi}\right)$$

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association*, *83*(402), 426-431.
Wang et al. (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics.*

最尤推定 + 非対称リンク関数 → 非凸最適化 ☹

- ロジスティック回帰 = 最尤推定 + ロジットリンク関数

$$\log\left(1+\exp(-\bar{y}_i\beta^\top x_i)\right) = -\log\ \psi_{\log}^{-1}\left(\beta^\top x_i\right)^{y_i}\ \left(1-\psi_{\log}^{-1}\left(\beta^\top x_i\right)\right)^{1-y_i}$$

ロジスティック損失

- 一般のリンク関数も最尤推定 (対数損失) と組合せ可能

$$-\log F_\xi\left(\beta^\top x_i\right)^{y_i}\left(1-F_\xi\left(\beta^\top x_i\right)\right)^{1-y_i}$$

- ただし $\beta^\top x$ に関して凸とは限らない

最尤推定 + 非対称リンク関数 → 非凸最適化 ☹

- ロジスティック回帰 = 最尤推定 + ロジットリンク関数

$$\log\left(1+\exp(-\bar{y}_i\beta^\top x_i)\right) = -\log\ \psi_{\log}^{-1}\left(\beta^\top x_i\right)^{y_i}\ \left(1-\psi_{\log}^{-1}\left(\beta^\top x_i\right)\right)^{1-y_i}$$

ロジスティック損失　　対数損失

- 一般のリンク関数も最尤推定 (対数損失) と組合せ可能

$$-\log F_\xi\left(\beta^\top x_i\right)^{y_i}\left(1-F_\xi\left(\beta^\top x_i\right)\right)^{1-y_i}$$

- ただし $\beta^\top x$ に関して凸とは限らない

最尤推定 + 非対称リンク関数 → 非凸最適化 ☹

- ロジスティック回帰 = 最尤推定 + ロジットリンク関数

$$\underbrace{\log\left(1+\exp(-\bar{y}_i\beta^\top x_i)\right)}_{\text{ロジスティック損失}} = \underbrace{-\log}_{\text{対数損失}} \underbrace{\psi_{\log}^{-1}\left(\beta^\top x_i\right)^{y_i}}_{\text{逆リンク関数}} \left(1-\psi_{\log}^{-1}\left(\beta^\top x_i\right)\right)^{1-y_i}$$

- 一般のリンク関数も最尤推定 (対数損失) と組合せ可能

$$-\log F_\xi\left(\beta^\top x_i\right)^{y_i}\left(1-F_\xi\left(\beta^\top x_i\right)\right)^{1-y_i}$$

- ただし $\beta^\top x$ に関して凸とは限らない

最尤推定 + 非対称リンク関数 → 非凸最適化 ☹

- ロジスティック回帰 = 最尤推定 + ロジットリンク関数

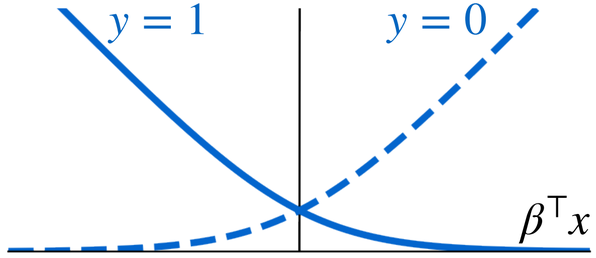
$$\log\left(1+\exp(-\bar{y}_i\beta^\top x_i)\right) = -\log \psi_{\log}^{-1}\left(\beta^\top x_i\right)^{y_i}\left(1-\psi_{\log}^{-1}\left(\beta^\top x_i\right)\right)^{1-y_i}$$

ロジスティック損失　対数損失　逆リンク関数

- 一般のリンク関数も最尤推定 (対数損失) と組合せ可能

$$-\log F_\xi\left(\beta^\top x_i\right)^{y_i}\left(1-F_\xi\left(\beta^\top x_i\right)\right)^{1-y_i}$$

- ただし $\beta^\top x$ に関して凸とは限らない



対数損失 + ロジットリンク

最尤推定 + 非対称リンク関数 → 非凸最適化 ☹

- ロジスティック回帰 = 最尤推定 + ロジットリンク関数

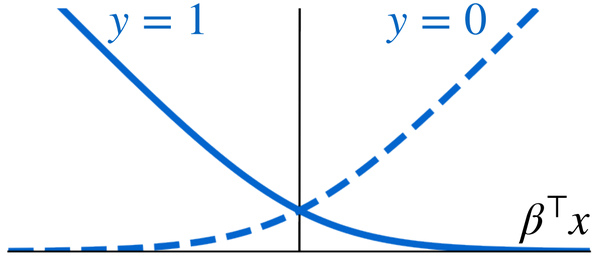
$$\log\left(1+\exp(-\bar{y}_i\beta^\top x_i)\right) = -\log \psi_{\log}^{-1}\left(\beta^\top x_i\right)^{y_i}\left(1-\psi_{\log}^{-1}\left(\beta^\top x_i\right)\right)^{1-y_i}$$

ロジスティック損失　対数損失　逆リンク関数

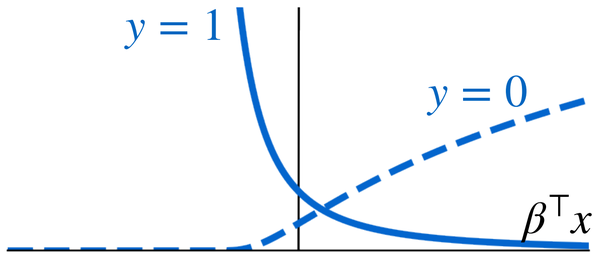
- 一般のリンク関数も最尤推定 (対数損失) と組合せ可能

$$-\log F_\xi\left(\beta^\top x_i\right)^{y_i}\left(1-F_\xi\left(\beta^\top x_i\right)\right)^{1-y_i}$$

- ただし $\beta^\top x$ に関して凸とは限らない

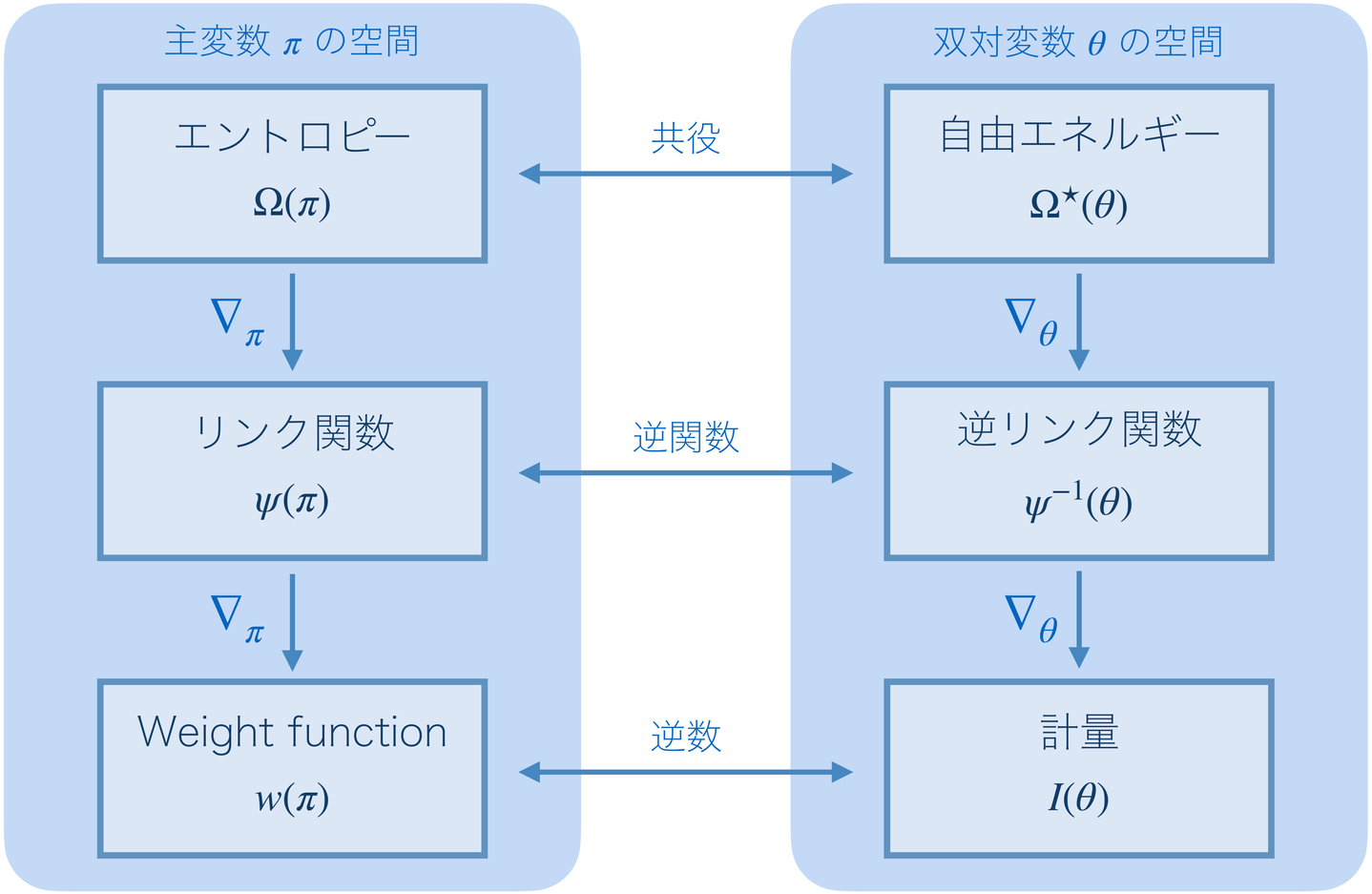


対数損失 + ロジットリンク

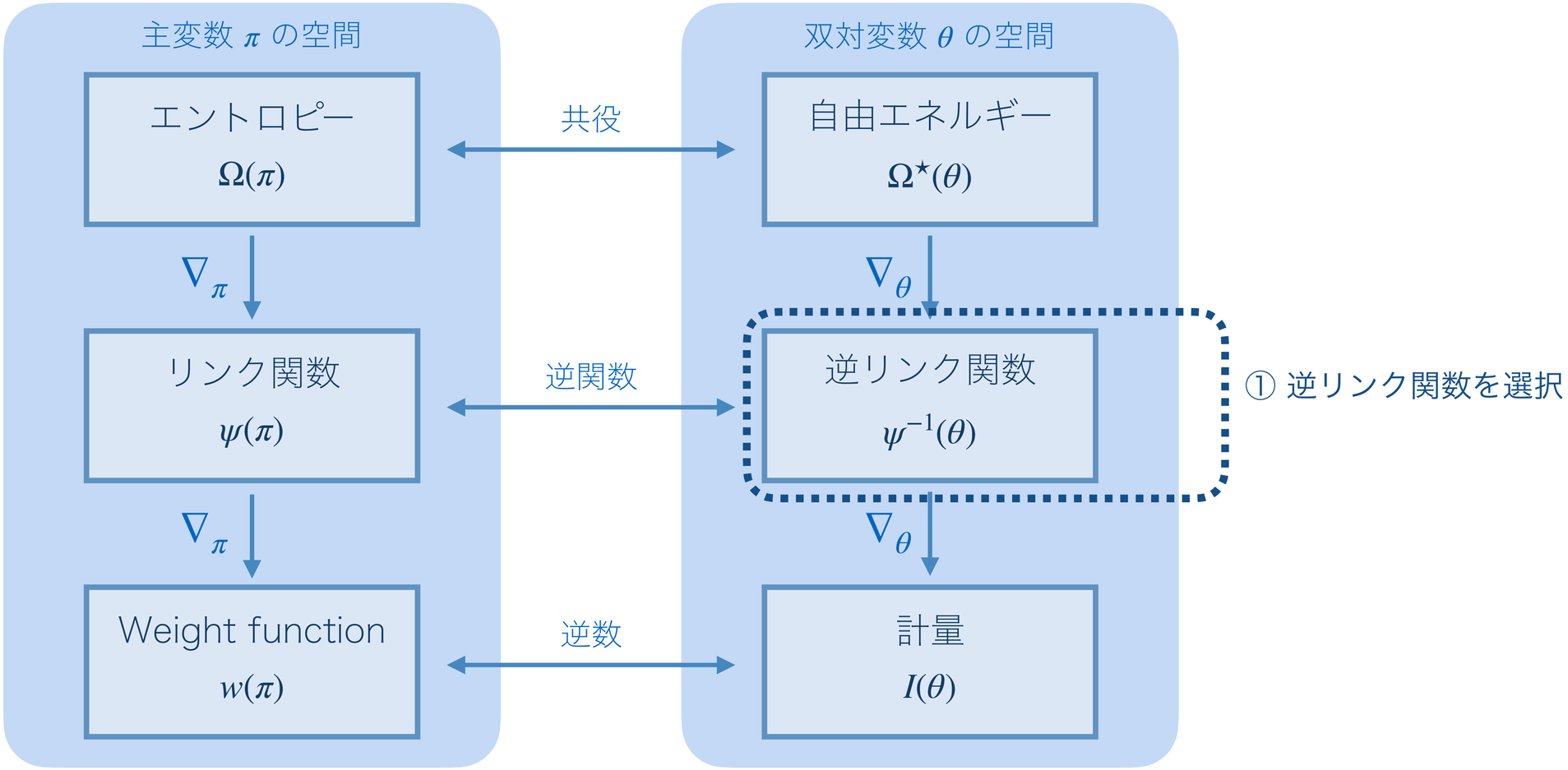


対数損失 + GEV リンク ($\xi = 0.5$)

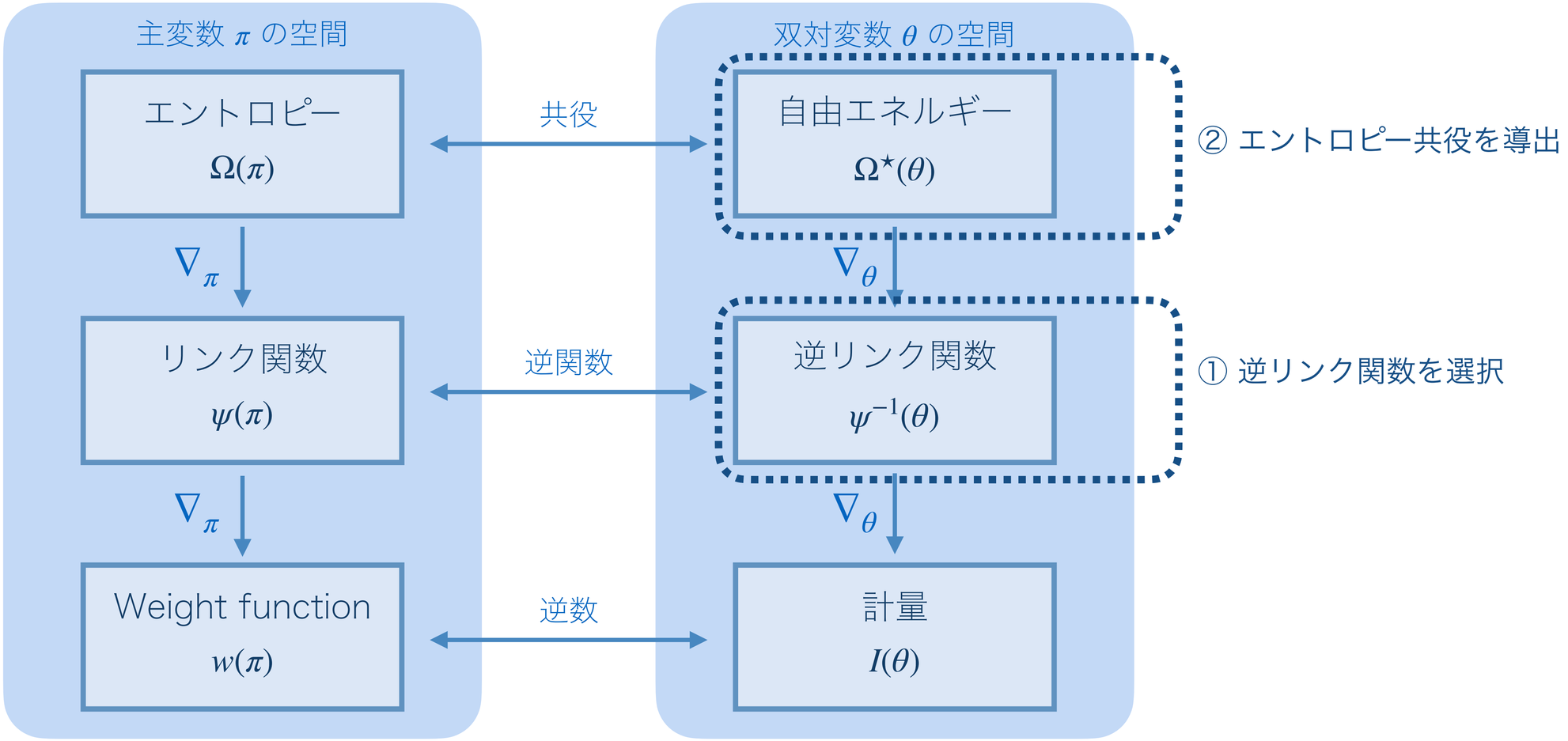
アイデア: 非対称リンクから FY 損失を構成



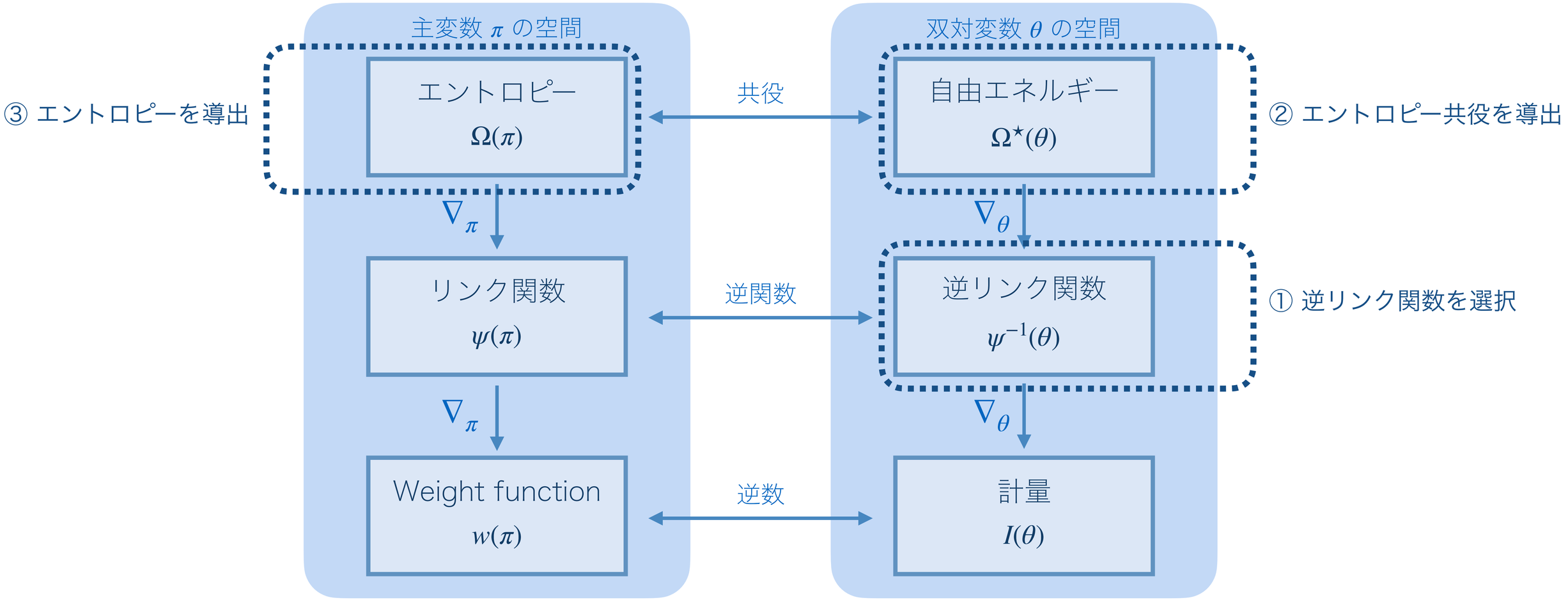
アイデア: 非対称リンクから FY 損失を構成



アイデア: 非対称リンクから FY 損失を構成

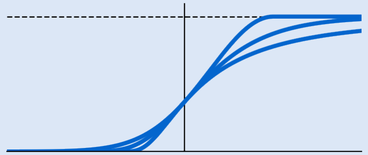


アイデア: 非対称リンクから FY 損失を構成



GEV リンク関数を用いた Fenchel–Young 損失

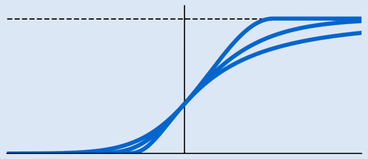
逆 GEV リンク関数



$$F_\xi(x) = \exp\left((1+\xi x)_+^{-1/\xi}\right)$$

GEV リンク関数を用いた Fenchel–Young 損失

逆 GEV リンク関数



$$F_\xi(x) = \exp\left((1+\xi x)_+^{-1/\xi}\right)$$

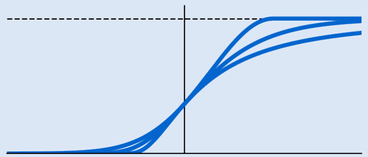
積分

エントロピー共役

$$\Omega^\star(\theta) = \int_{\inf \mathrm{supp}(F_\xi)}^{\theta} F_\xi(s)\mathrm{d}s$$

GEV リンク関数を用いた Fenchel–Young 損失

逆 GEV リンク関数



$$F_\xi(x) = \exp\left((1+\xi x)_+^{-1/\xi}\right)$$

逆関数
+ 積分

エントロピー

$$\Omega(p) = \int_0^p F_\xi^{-1}(q)\mathrm{d}q$$

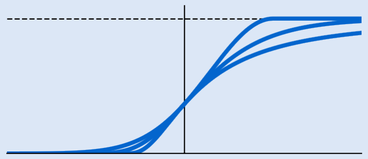
積分

エントロピー共役

$$\Omega^\star(\theta) = \int_{\inf \mathrm{supp}(F_\xi)}^{\theta} F_\xi(s)\mathrm{d}s$$

GEV リンク関数を用いた Fenchel–Young 損失

逆 GEV リンク関数



$$F_\xi(x) = \exp\left((1+\xi x)_+^{-1/\xi}\right)$$

逆関数
+ 積分

エントロピー

$$\Omega(p) = \int_0^p F_\xi^{-1}(q)\mathrm{d}q$$

積分

エントロピー共役

$$\Omega^\star(\theta) = \int_{\inf \mathrm{supp}(F_\xi)}^{\theta} F_\xi(s)\mathrm{d}s$$

Fenchel–Young 損失

$$\Omega^\star(\theta) + \Omega(y) - \langle \theta, y \rangle$$

二値分類問題の先へ

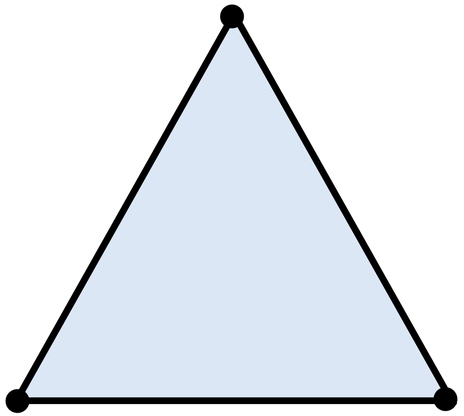
- Fenchel–Young 損失の利点: 予測空間さえ定まれば自然にダイバージェンスを定義可能

$$\arg\max_{\boldsymbol{\pi}\in\mathscr{C}} \langle \boldsymbol{\theta}, \boldsymbol{\pi} \rangle - \Omega(\boldsymbol{\pi})$$

台集合 $\mathscr{C}$ への射影オラクルがあれば OK

- 多クラス分類

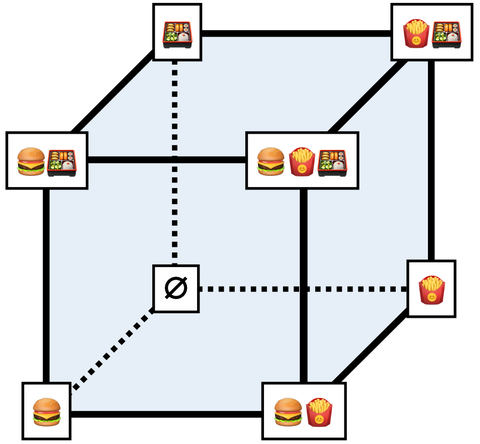
$\mathscr{C} =$



確率単体

- マルチラベル予測

$\mathscr{C} =$



単位立方体

Fenchel–Young 損失の利点と未解決の課題

- 😊 予測空間 (主空間) さえ定まれば自然にダイバージェンスを定義可能
- 😊 双対空間内では制約なし最適化
- ☹ 予測空間が複雑になればなるほどどのような凸関数を用いればよいか自明でない
- ☹ 凸関数ごとの差分がよくわからない

機械学習と凸共役の交わり（目次）

前半

- 二値分類問題: 主空間の観点から
- 二値分類問題: 双対空間の観点から
- 応用: 非対称リンク関数を用いた二値応答回帰

  Bao & Sugiyama (2021) “Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation”

Done 🙌

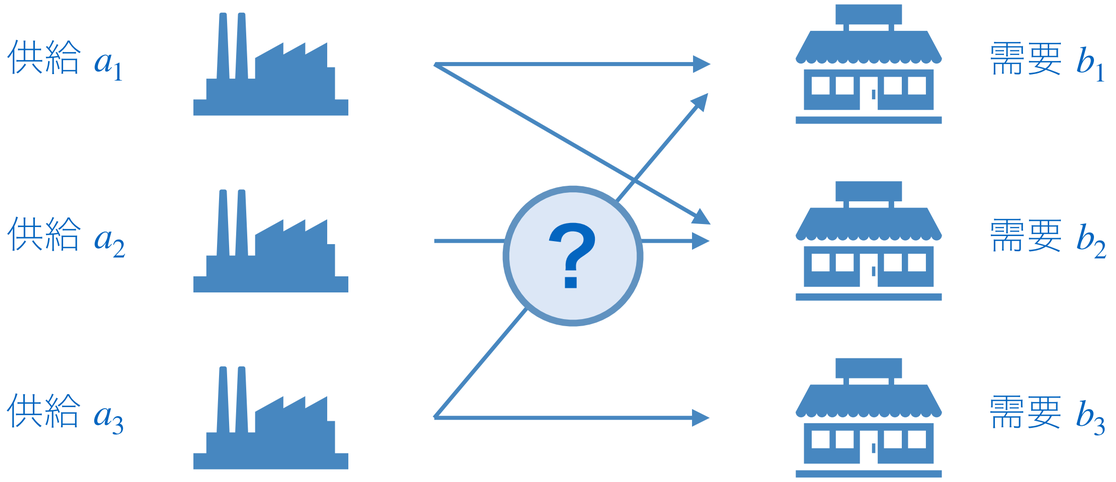
後半

- ➡ **最適輸送問題: 双対空間の観点から**
- 応用: q-指数分布を用いたスパース最適輸送

  Bao & Sakaue (2022) “Sparse Regularized Optimal Transport with Deformed q-Entropy”
- その他の問題

導入: 最適輸送問題

- 目標: 輸送コストが定まっているときに、需給を満たしつつ総輸送コストが最適な輸送計画を求める



- 線形計画による定式化 (Kantorovich の凸緩和)

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle \quad \text{where} \quad U(\mu,\nu) = \left\{ \mathbf{\Pi} \in \mathbb{R}^{n \times n} \;\middle|\; \mathbf{\Pi 1} = \mathbf{a},\ \mathbf{\Pi}^{\mathsf{T}} \mathbf{1} = \mathbf{b} \right\}$$

$\mathbf{D}_{ij}$: (i, j) 間のコスト

工場の供給量をすべて満たす

店舗の需要量をすべて満たす

輸送多面体

導入: 最適輸送問題

- 線形計画問題として解くと計算量が $O(n^3 \ln n)$

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle \quad \text{where} \quad U(\mu, \nu) = \left\{ \mathbf{\Pi} \in \mathbb{R}^{n \times n} \;\middle|\; \mathbf{\Pi 1} = \mathbf{a},\ \mathbf{\Pi}^\top \mathbf{1} = \mathbf{b} \right\}$$

  - cf. ネットワーク単体法、内点法
  - 機械学習の応用では高速化できると望ましい (パイプライン内で最適輸送を何度も解きたいため)

導入: 最適輸送問題

- 線形計画問題として解くと計算量が $O(n^3 \ln n)$

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle \quad \text{where} \quad U(\mu, \nu) = \left\{ \mathbf{\Pi} \in \mathbb{R}^{n \times n} \;\middle|\; \mathbf{\Pi} \mathbf{1} = \mathbf{a},\ \mathbf{\Pi}^\top \mathbf{1} = \mathbf{b} \right\}$$

  - cf. ネットワーク単体法、内点法
  - 機械学習の応用では高速化できると望ましい (パイプライン内で最適輸送を何度も解きたいため)

- 解決策 (Sinkhorn アルゴリズム): **凸正則化** + 座標降下法によって計算量 $O(n^2)$ に

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \lambda \sum_{i,j} H\left(\mathbf{\Pi}_{ij}\right) \quad \text{where} \quad H(\pi) = \pi \ln \pi - \pi \ \text{(負の Shannon エントロピー)}$$

  - $O(n^2)$ よりも高速なアルゴリズムも存在するが、Sinkhorn アルゴリズムは GPU による高速な行列演算で実現可能なのが利点

一般の正則化付き最適輸送の定式化

- 主問題 ($\Omega : \mathbb{R} \to \mathbb{R}$ は凸正則化)

$$\inf_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega\big(\mathbf{\Pi}_{ij}\big)$$

- 双対問題

$$\sup_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{R}^n} -\langle \mathbf{a}, \boldsymbol{\alpha} \rangle - \langle \mathbf{b}, \boldsymbol{\beta} \rangle - \sum_{i,j} \Omega^\star\big(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j\big)$$

  - ラグランジアン + 強双対性から従う ($\boldsymbol{\alpha}, \boldsymbol{\beta}$: 未定定数)
  - 注意: 双対問題は制約なし最適化

- 主双対変数の関係 (逆リンク関数)

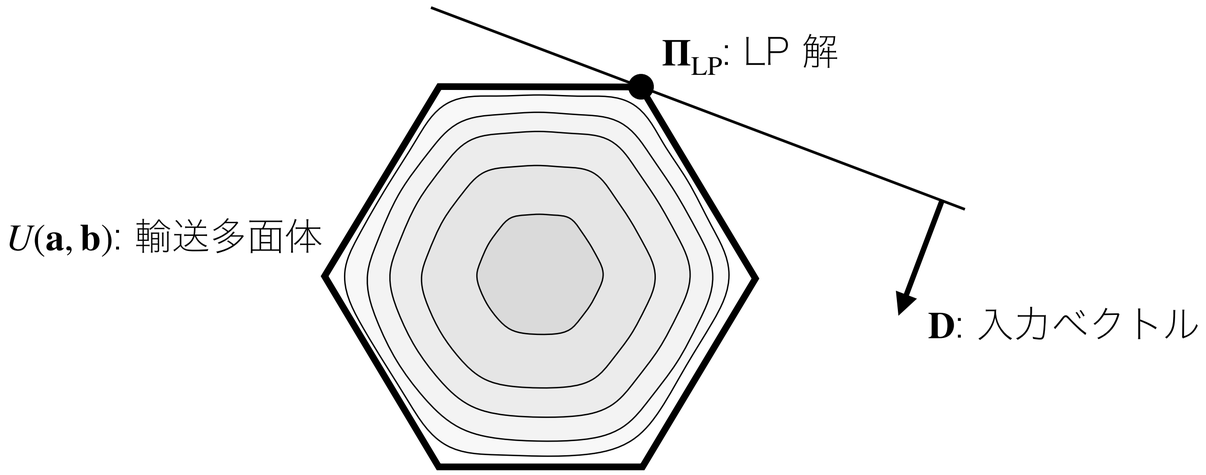
$$\mathbf{\Pi}_{ij} = \nabla\Omega^\star\big(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j\big) \quad \text{for all i, j}$$

  - KKT 条件から従う

Bregman 射影としての最適輸送

定理 [Dessein *et al*. 2018]**.** 正則化付き最適輸送の解 $\mathbf{\Pi}^* = \arg\min_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$ は制約なしの解 $\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi} \in \mathbb{R}^{n \times n}} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$ の Bregman 射影:

$$\mathbf{\Pi}^* = \arg\min_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} D_\Omega(\mathbf{\Pi}, \tilde{\mathbf{\Pi}})$$



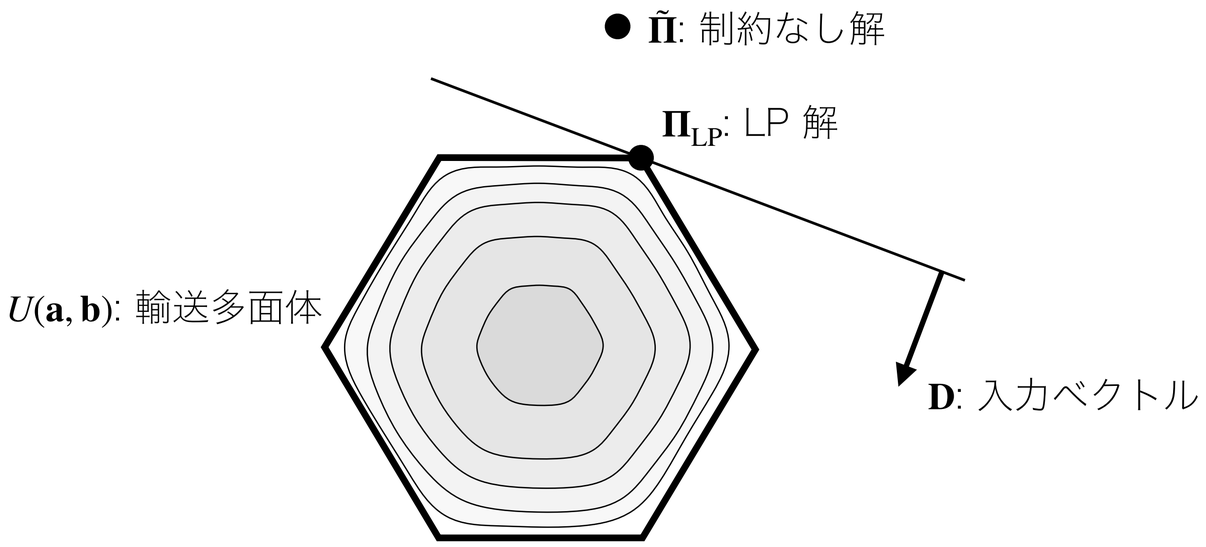
ここでは Bregman ダイバージェンスは正測度空間 $\mathbb{R}^{n \times n}_{\geq 0}$ 上で定義

Dessein, Arnaud, Nicolas Papadakis, and Jean-Luc Rouas. "Regularized optimal transport and the rot mover's distance." *The Journal of Machine Learning Research* 19.1 (2018): 590-642.

Bregman 射影としての最適輸送

定理 [Dessein *et al*. 2018]**.** 正則化付き最適輸送の解 $\mathbf{\Pi}^* = \arg\min_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega\left(\mathbf{\Pi}_{ij}\right)$ は制約なしの解 $\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi} \in \mathbb{R}^{n \times n}} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$ の Bregman 射影:

$$\mathbf{\Pi}^* = \underset{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})}{\arg\min} D_\Omega(\mathbf{\Pi}, \tilde{\mathbf{\Pi}})$$



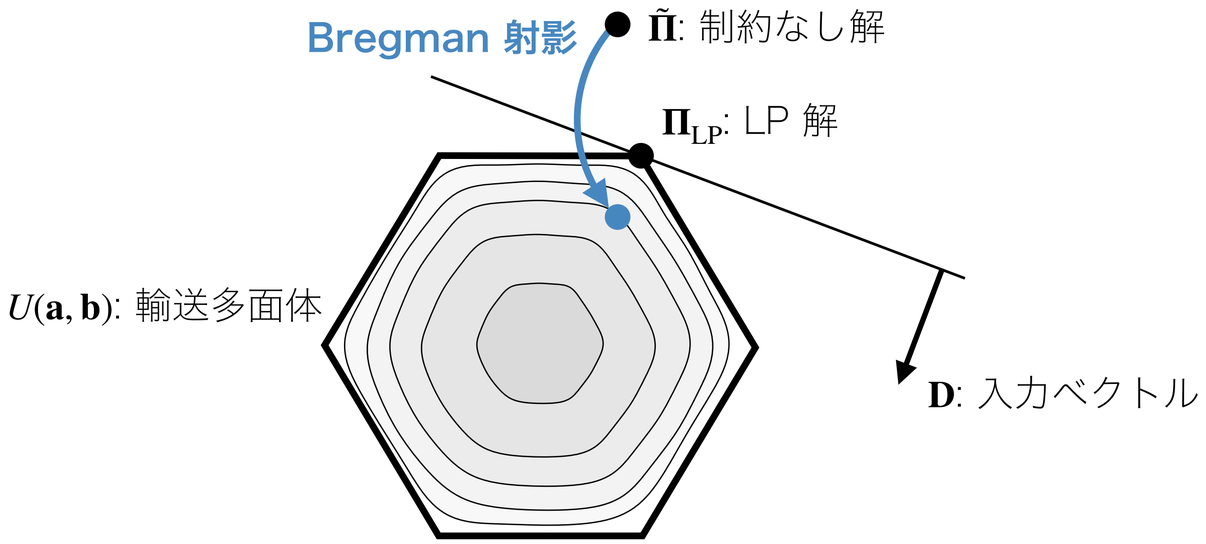
ここでは Bregman ダイバージェンスは正測度空間 $\mathbb{R}^{n \times n}_{\geq 0}$ 上で定義

Dessein, Arnaud, Nicolas Papadakis, and Jean-Luc Rouas. "Regularized optimal transport and the rot mover's distance." *The Journal of Machine Learning Research* 19.1 (2018): 590-642.

Bregman 射影としての最適輸送

定理 [Dessein *et al*. 2018]**.** 正則化付き最適輸送の解 $\mathbf{\Pi}^* = \arg\min_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$ は制約なしの解 $\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi} \in \mathbb{R}^{n \times n}} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$ の Bregman 射影:

$$\mathbf{\Pi}^* = \underset{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})}{\arg\min} D_\Omega(\mathbf{\Pi}, \tilde{\mathbf{\Pi}})$$



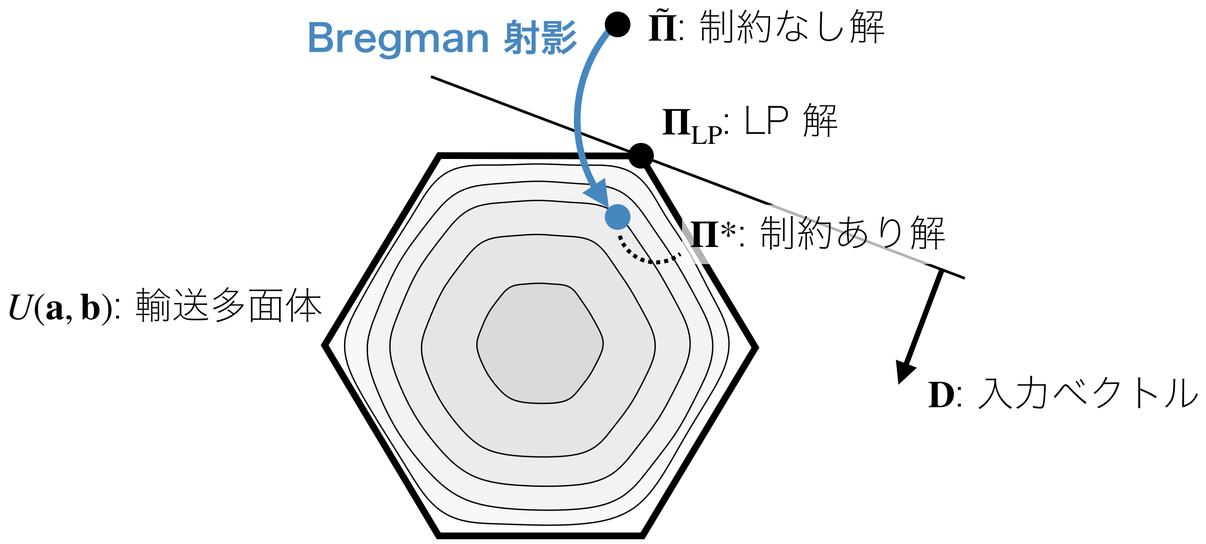
ここでは Bregman ダイバージェンスは正測度空間 $\mathbb{R}^{n \times n}_{\geq 0}$ 上で定義

Dessein, Arnaud, Nicolas Papadakis, and Jean-Luc Rouas. "Regularized optimal transport and the rot mover's distance." *The Journal of Machine Learning Research* 19.1 (2018): 590-642.

Bregman 射影としての最適輸送

定理 [Dessein *et al*. 2018]**.** 正則化付き最適輸送の解 $\mathbf{\Pi}^* = \arg\min_{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$ は制約なしの解 $\tilde{\mathbf{\Pi}} = \arg\min_{\mathbf{\Pi} \in \mathbb{R}^{n \times n}} \langle \mathbf{D}, \mathbf{\Pi} \rangle + \sum_{i,j} \Omega(\mathbf{\Pi}_{ij})$ の Bregman 射影:

$$\mathbf{\Pi}^* = \underset{\mathbf{\Pi} \in U(\mathbf{a},\mathbf{b})}{\arg\min} D_\Omega(\mathbf{\Pi}, \tilde{\mathbf{\Pi}})$$



ここでは Bregman ダイバージェンスは正測度空間 $\mathbb{R}^{n \times n}_{\geq 0}$ 上で定義

Dessein, Arnaud, Nicolas Papadakis, and Jean-Luc Rouas. "Regularized optimal transport and the rot mover's distance." *The Journal of Machine Learning Research* 19.1 (2018): 590-642.

Sinkhorn アルゴリズムの導出

双対問題 $\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n} -\langle \mathbf{a},\boldsymbol{\alpha}\rangle - \langle \mathbf{b},\boldsymbol{\beta}\rangle - \sum_{i,j}\Omega^\star\left(-\mathbf{D}_{ij}-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j\right)$

逆リンク関数 $\boldsymbol{\Pi}_{ij} = \nabla\Omega^\star\left(-\mathbf{D}_{ij}-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j\right)$

- Shannon エントロピー $\Omega(\pi) = \lambda(\pi\ln\pi - \pi)$ を正則化として用いた場合

$$\boldsymbol{\Pi}_{ij} = \exp\left(\frac{-\mathbf{D}_j - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j}{\lambda}\right) = \exp\left(-\frac{\boldsymbol{\alpha}_i}{\lambda}\right)\exp\left(-\frac{\mathbf{D}_{ij}}{\lambda}\right)\exp\left(-\frac{\boldsymbol{\beta}_j}{\lambda}\right)$$

分解可能

- 行列形に書き直す: $\boldsymbol{\Pi} = \mathbf{ADB}$ $(\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha}),\ \mathbf{B} = \mathrm{diag}(\boldsymbol{\beta}))$

- 行列スケーリングに帰着 (Sinkhorn アルゴリズム)
  - 行和制約に関する更新 $\boldsymbol{\beta}^{(k+1)} = \mathbf{a} \oslash (\boldsymbol{\Pi}\boldsymbol{\alpha}^{(k)})$
  - 列和制約に関する更新 $\boldsymbol{\alpha}^{(k+1)} = \mathbf{b} \oslash (\boldsymbol{\Pi}^\top\boldsymbol{\beta}^{(k+1)})$

] それぞれ計算量 $O(n^2)$ の行列-ベクトル積

一般の正則化付き最適輸送

双対問題 $\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n} -\langle \mathbf{a},\boldsymbol{\alpha}\rangle - \langle \mathbf{b},\boldsymbol{\beta}\rangle - \sum_{i,j} \Omega^\star\left(-\mathbf{D}_{ij}-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j\right)$

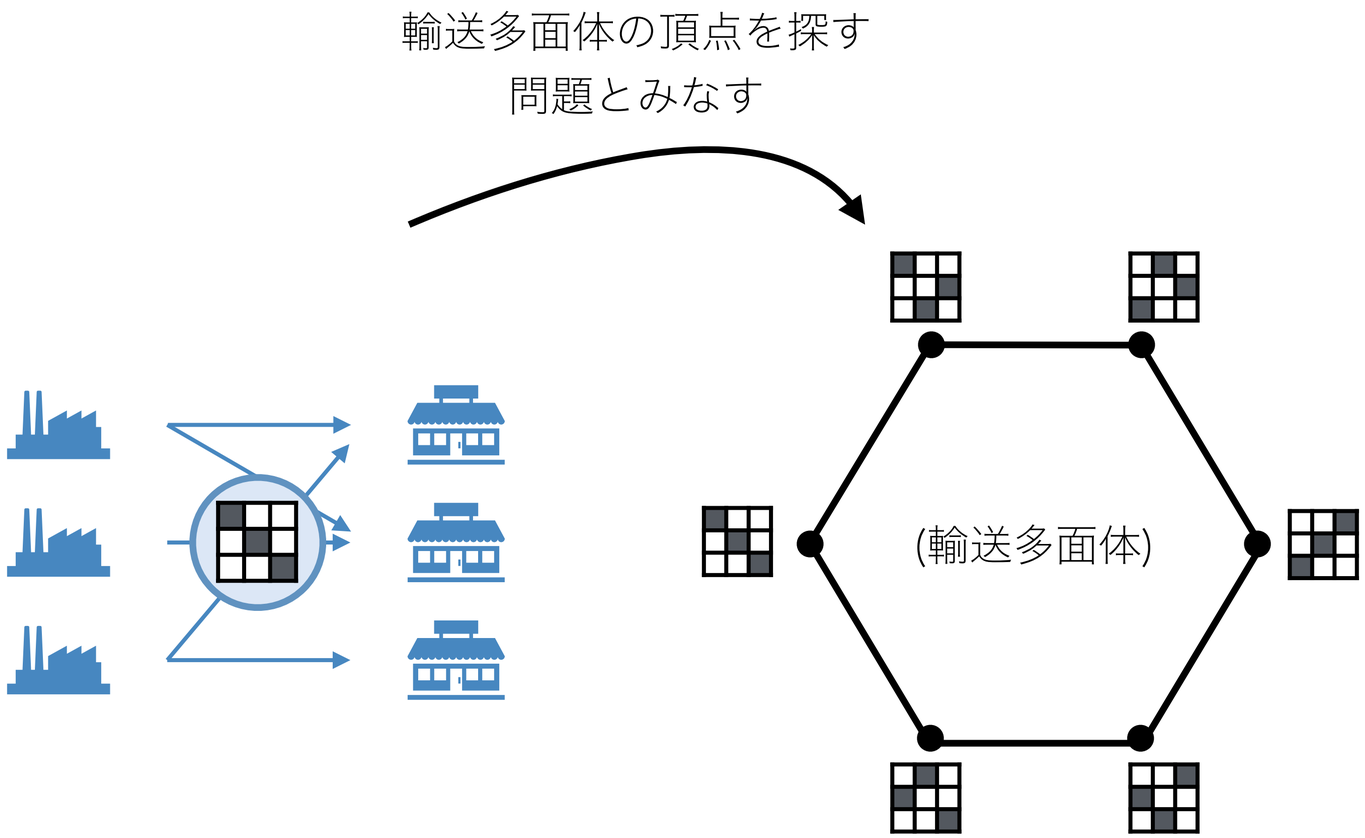
逆リンク関数 $\boldsymbol{\Pi}_{ij} = \nabla\Omega^\star\left(-\mathbf{D}_{ij}-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j\right)$

- 一般の逆リンク関数の場合は最適解が分解可能とは限らない
  - cf. Sinkhorn の場合 (分解可能)

$$\boldsymbol{\Pi}_{ij} = \exp\left(\frac{-\mathbf{D}_j-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j}{\lambda}\right) = \exp\left(-\frac{\boldsymbol{\alpha}_i}{\lambda}\right)\exp\left(-\frac{\mathbf{D}_{ij}}{\lambda}\right)\exp\left(-\frac{\boldsymbol{\beta}_j}{\lambda}\right)$$

- 解法: 双対問題に勾配法を適用
  - 例えば BFGS を適用すると「更新方向の計算」「Hessian 推定の更新」それぞれ rank-1 更新で可能 ($O(n^2)$ の計算量)

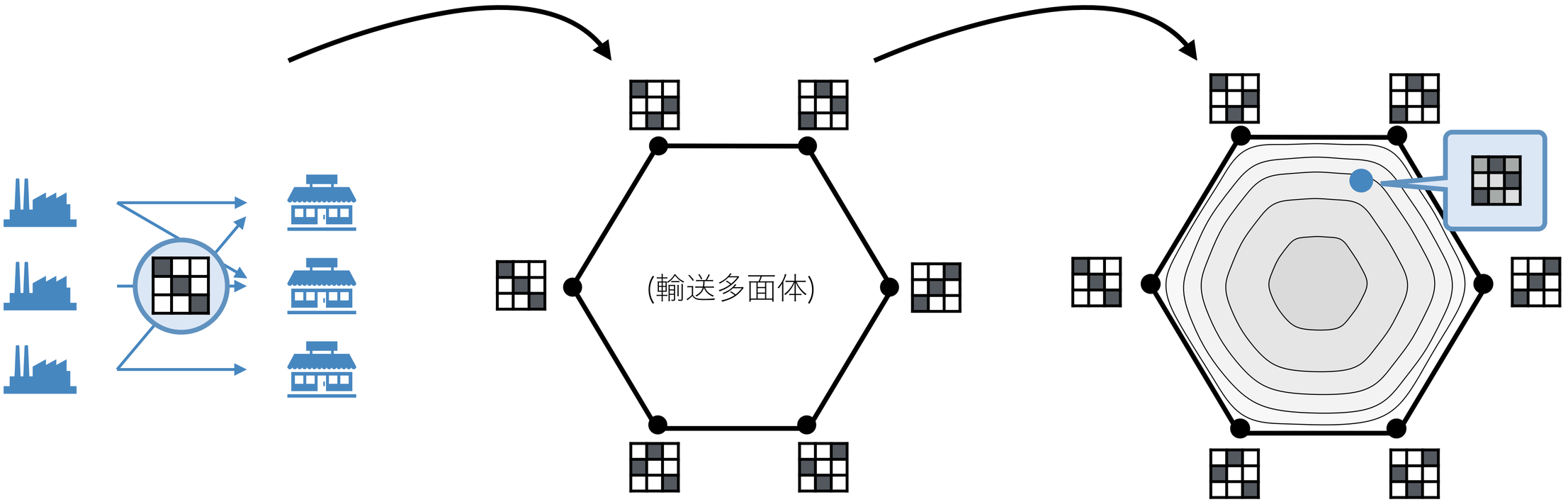
正則化付き最適輸送の別の見方



正則化付き最適輸送の別の見方

輸送多面体の頂点を探す
問題とみなす

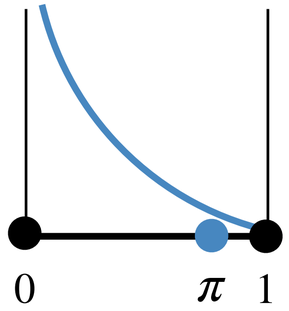
最適な輸送行列のかわりに
尤もらしい確率分布を探索



正則化付き最適輸送の別の見方

分類

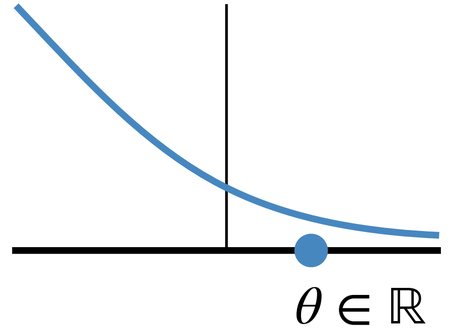
主空間 (平均の推定)



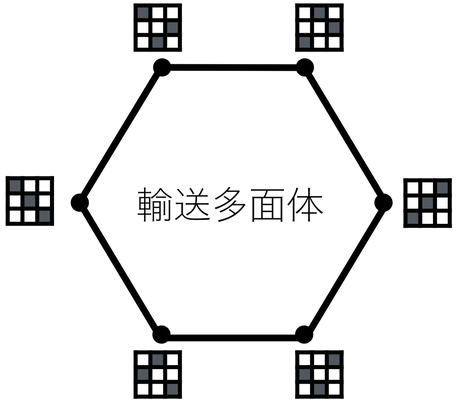
逆リンク関数

$\pi = \psi^{-1}(\theta)$

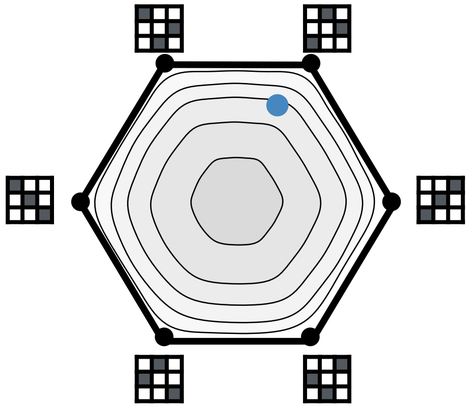
双対空間 (最適化 in ロジット)



最適輸送



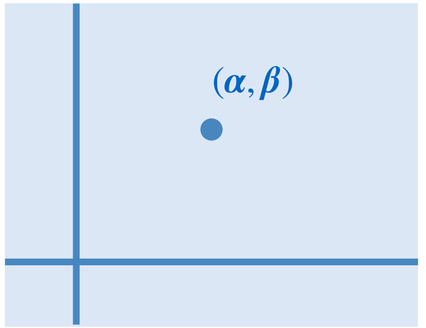
主空間 (輸送行列の探索)



逆リンク関数

$\mathbf{\Pi}_{ij} = \nabla\Omega^{\star}(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j)$

双対空間 (未定定数の決定)



😊 双対空間上の最適化は制約なし最適化

機械学習と凸共役の交わり（目次）

前半

- 二値分類問題: 主空間の観点から
- 二値分類問題: 双対空間の観点から
- 応用: 非対称リンク関数を用いた二値応答回帰

  Bao & Sugiyama (2021) “Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation”

後半

- 最適輸送問題: 双対空間の観点から
- **応用: q-指数分布を用いたスパース最適輸送**

  Bao & Sakaue (2022) “Sparse Regularized Optimal Transport with Deformed q-Entropy”
- その他の問題

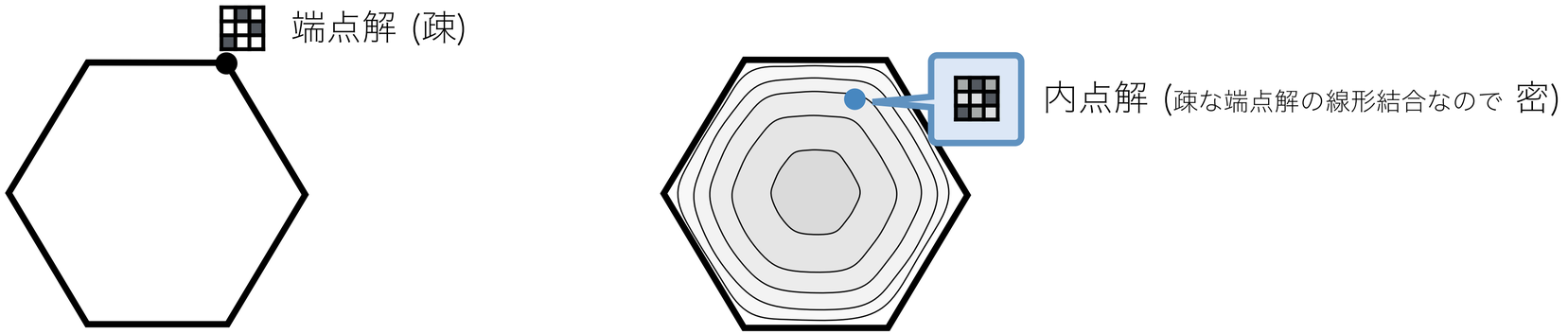
正則化はどのように選ぶ？

- LP 解と Sinkhorn 解の比較

解釈性の高さ & LP 解との定性的類似性ゆえに
スパース解が好まれることが多い

各々 30 × 30 の輸送行列の解を図示

- 輸送多面体上での理解



Sinkhorn 解が密であることの別の理解

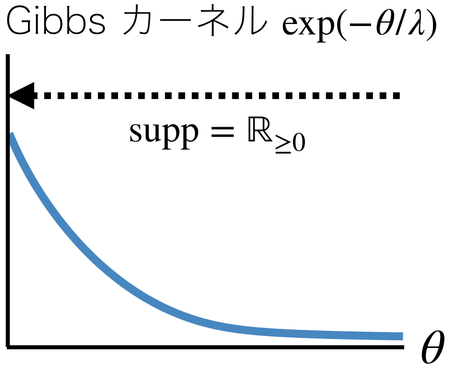
双対問題 $\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n} -\langle \mathbf{a},\boldsymbol{\alpha}\rangle - \langle \mathbf{b},\boldsymbol{\beta}\rangle - \sum_{i,j} \Omega^\star\big(-\mathbf{D}_{ij}-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j\big)$

逆リンク関数 $\boldsymbol{\Pi}_{ij} = \nabla\Omega^\star\big(-\mathbf{D}_{ij}-\boldsymbol{\alpha}_i-\boldsymbol{\beta}_j\big)$

- Shannon エントロピー $\Omega(\pi) = \lambda(\pi\ln\pi - \pi)$ を正則化として用いた場合

$$\boldsymbol{\Pi}_{ij} = \exp\left(\frac{-\mathbf{D}_j - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j}{\lambda}\right)$$

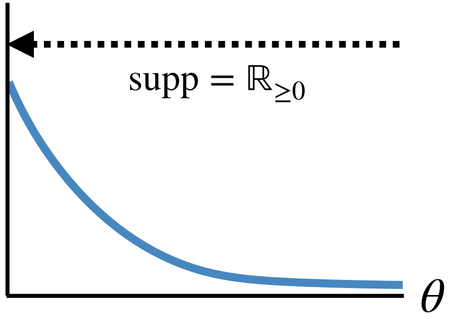
- 逆リンク関数の台が非有界であるため解は必ず密行列



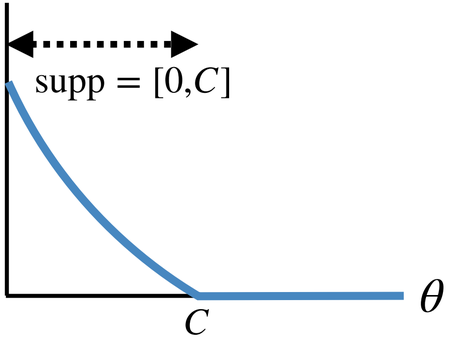
アイデア: 台が有界なリンク関数を利用

- 逆リンク関数の有界性はスパース解の必要条件 (十分とは限らないが)

Gibbs カーネル $\exp(-\theta/\lambda)$



有界な逆リンク関数 $\nabla\Omega^\star(-\theta)$



逆リンク関数 $\boldsymbol{\Pi}_{ij} = \nabla\Omega^\star\big(-(\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j)\big)$

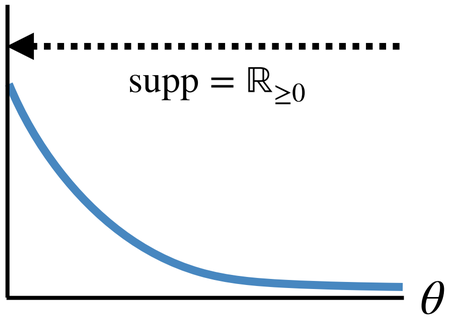
コスト $\mathbf{D}_{ij}$ が十分大きかったら

対応する辺 (i, j) は輸送を考えない

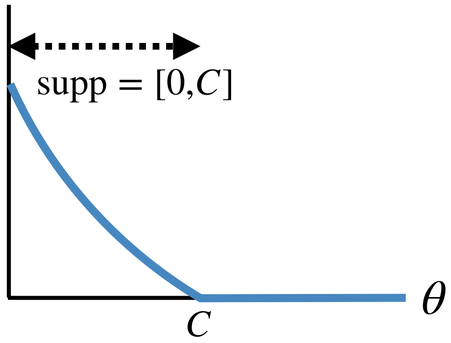
アイデア: 台が有界なリンク関数を利用

- 逆リンク関数の有界性はスパース解の必要条件 (十分とは限らないが)

Gibbs カーネル $\exp(-\theta/\lambda)$



有界な逆リンク関数 $\nabla\Omega^\star(-\theta)$



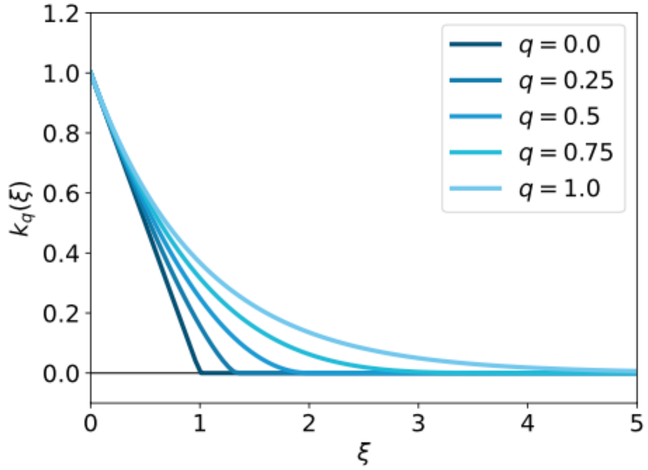
逆リンク関数 $\boldsymbol{\Pi}_{ij} = \nabla\Omega^\star\big(-(\mathbf{D}_{ij} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j)\big)$

コスト $\mathbf{D}_{ij}$ が十分大きかったら

対応する辺 (i, j) は輸送を考えない

- ここでは q-指数分布を用いる

q-Gibbs カーネル $\exp_q\left(-\frac{\theta}{\lambda}\right) = \left[1 + (1-q)\left(-\frac{\theta}{\lambda}\right)\right]_+^{1/(1-q)}$



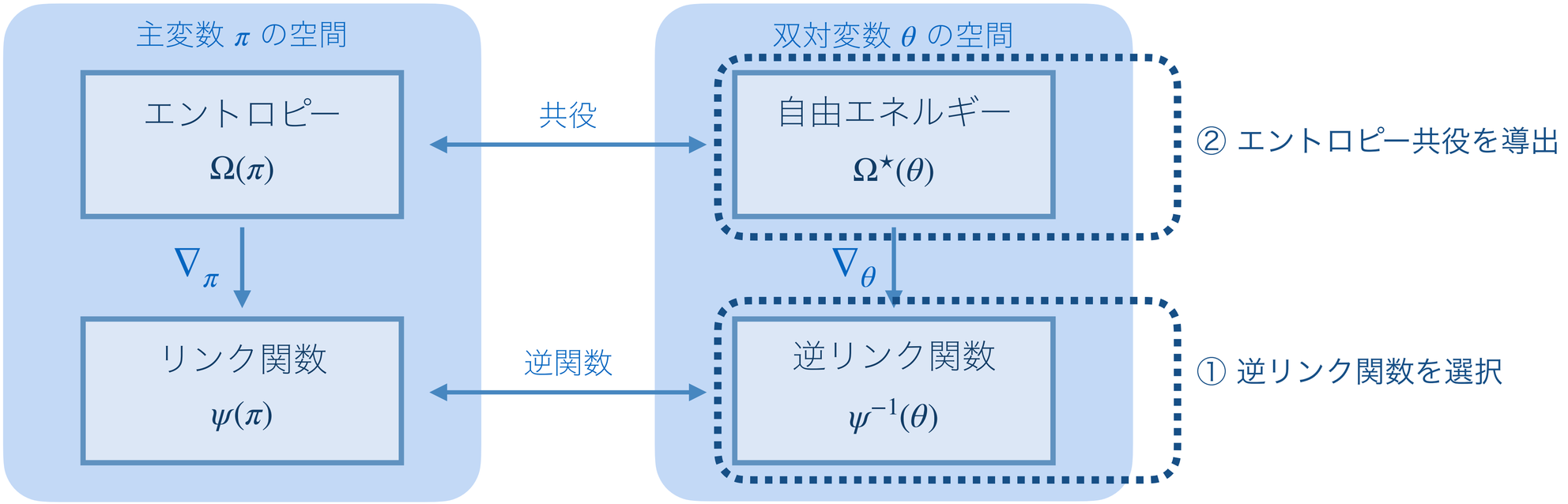
補足: q-指数関数と q-類似

- 線形常微分方程式の非線形拡張
  - $\frac{\mathrm{d}y}{\mathrm{d}x} = y$ の解は $y = \exp(x)$
  - $\frac{\mathrm{d}y}{\mathrm{d}x} = y^q$ の解は $y = \exp_q(x)$

q-指数関数 $\exp_q(x) := [1 + (1-q)x]_+^{1/(1-q)}$

- 指数関数に成立する演算の拡張
  - 加法性 $\exp(x+y) = \exp(x) \cdot \exp(y)$ → 疑似加法性 $\exp_q(x+y) = \exp_q(x) \otimes_q \exp_q(y)$
  - 逆関数 $\ln_q(x) = \frac{x^{1-q} - 1}{1-q}$
- Tsallis 統計力学での活用
  - Gibbs 分布 $P(z) \propto \exp(-\beta z)$ → q-Gibbs 分布 $P(z) \propto \exp_q(-\beta z)$
  - Shannon エントロピー $H(p) = -\langle p, \ln p \rangle$ → Tsallis エントロピー $H_q(p) = -\langle p, \ln_q p \rangle_q$

提案法: q-指数型のリンク関数に基づく最適輸送



- q-指数分布を逆リンク関数としてエントロピー共役 (双対問題の正則化項) を導出

$$\Omega(\pi) = \frac{\lambda}{2-q}(\pi \log_q(\pi) - \pi)$$

  - λ: 正則化の強さ, q: 逆リンク関数の台の広さを調整

双対問題の収束解析: q は収束にどう影響？

- 双対問題 $\sup_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n} -\langle \mathbf{a},\boldsymbol{\alpha}\rangle - \langle \mathbf{b},\boldsymbol{\beta}\rangle - \sum_{i,j} \Omega^\star\big(-\mathbf{D}_{ij} - \boldsymbol{\alpha}_i - \boldsymbol{\beta}_j\big)$ を BFGS で解くことを想定

定理. 適当な条件の下で、BFGS の K 回目の更新で得られる勾配の 2-ノルムは以下で抑えられる:

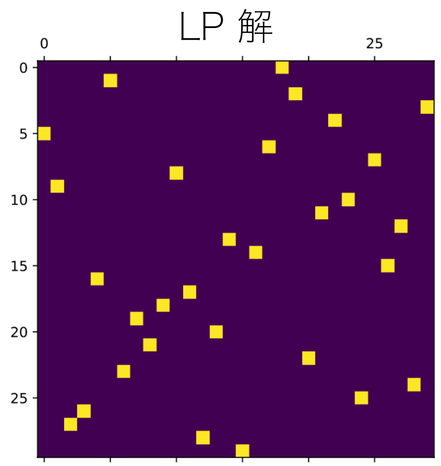
$$\sqrt{\frac{Cn\tau^q}{\lambda}} r^K$$

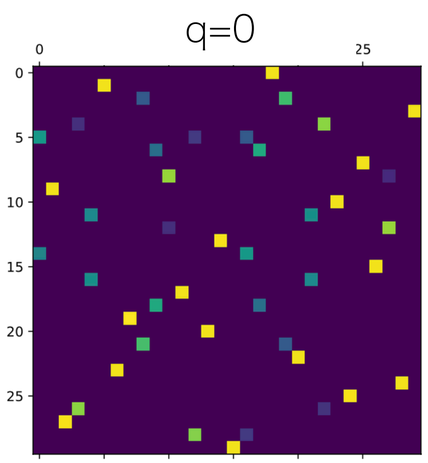
ただし、収束率は $0 < r < 1$ を満たす定数であり、τ は $(0,1)$ に含まれる適当な定数である。

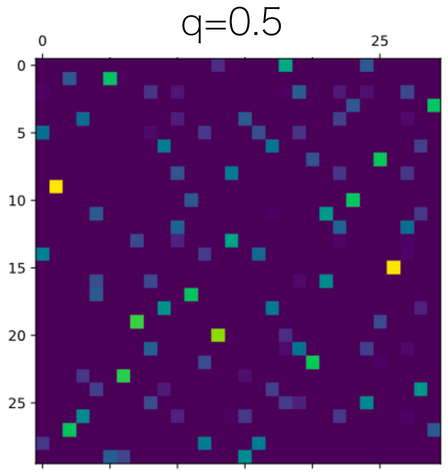
- ポイント: q が小さいほど収束が速い
- 行列スケーリングは利用できない
  - 疑似加法性 $\exp_q(x+y) = \exp_q(x) \otimes_q \exp_q(y)$ は分配則を満たさないため

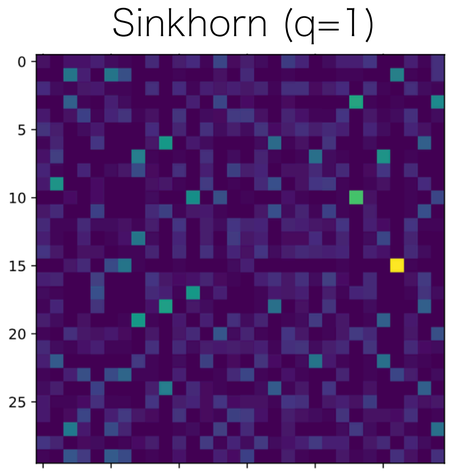
実験: 輸送行列解のスパース性

- q → 0 に近づけると解は LP 解に近づく







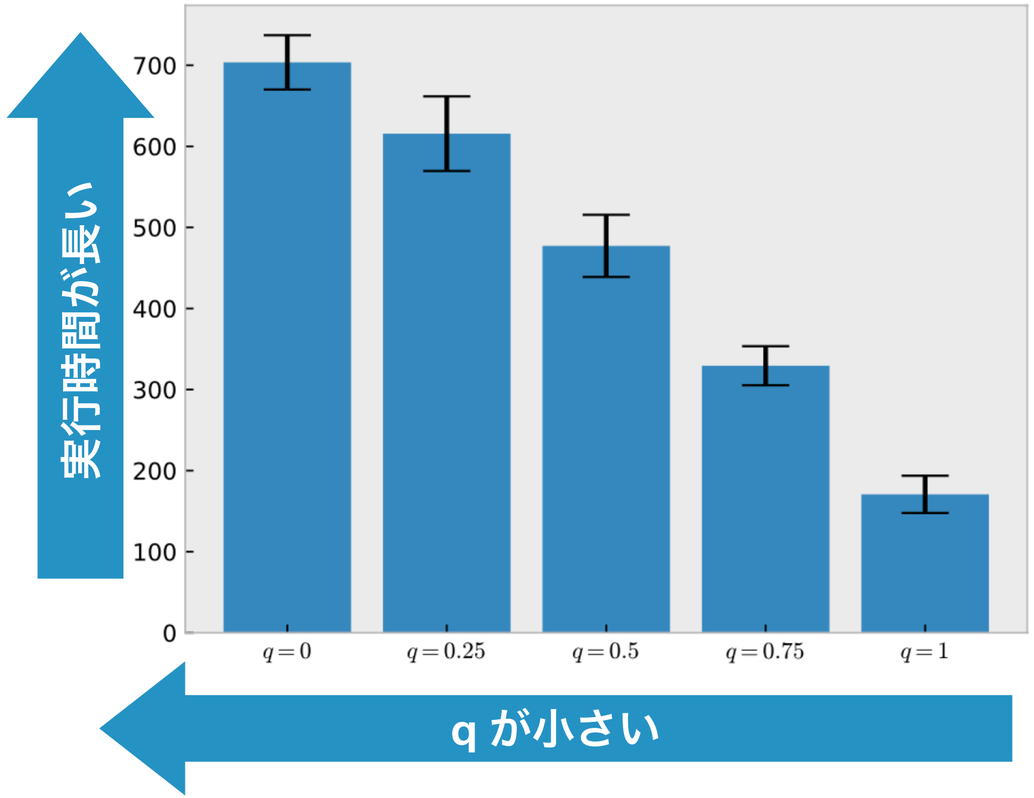


- 定量的な結果
  - q = 1: 完全に密行列 (非ゼロ割合 = 0)
  - q < 1: q = 1 の場合に比べてかなりスパース (非ゼロ割合 > 0.6 くらいであることが多い)

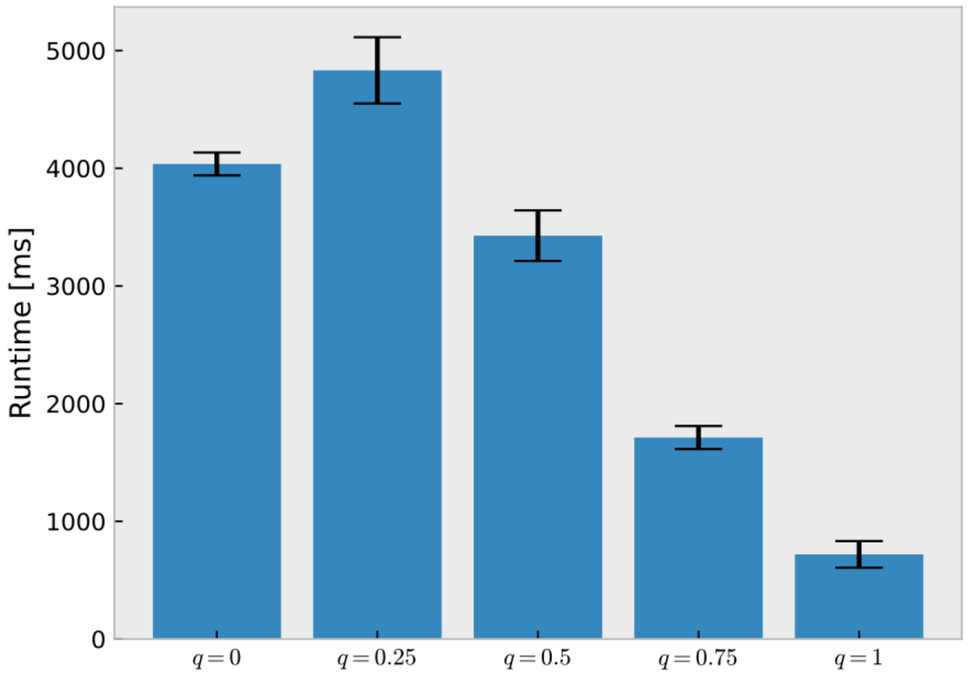
実験: 実行時間の比較

- q が小さいほど実行時間が長くなる傾向あり
  - スパース性と実行時間のトレードオフ

データセットの大きさ = 100



データセットの大きさ = 300



未解決の課題

- 正則化の導入による近似誤差
  - エントロピーの場合は調べられている [Weed COLT2018]
  - ただし機械学習コミュニティではエントロピー正則化付きの最適輸送自体を距離とみなす動きも出つつある
- 「最適」な正則化の設計
  - 計算量を最適にするにはどのような凸正則化を導入すればよい？

Weed, Jonathan. "An explicit analysis of the entropic penalty in linear programming." *Conference on Learning Theory*. PMLR, 2018.

機械学習と凸共役の交わり（目次）

前半

- 二値分類問題: 主空間の観点から
- 二値分類問題: 双対空間の観点から
- 応用: 非対称リンク関数を用いた二値応答回帰

  Bao & Sugiyama (2021) “Fenchel-Young Losses with Skewed Entropies for Class-posterior Probability Estimation”

後半

- 最適輸送問題: 双対空間の観点から
- 応用: q-指数分布を用いたスパース最適輸送

  Bao & Sakaue (2022) “Sparse Regularized Optimal Transport with Deformed q-Entropy”

Done 🙌

➜ **その他の問題**

凸共役による双対構造は様々な問題に見られる

- **分類問題（今日の前半）**
- **最適輸送（今日の後半）**
- ブースティング
- 強化学習・Bellman 方程式
- バンディット問題
- 「微分可能」線形計画問題
- 陰的正則化

凸共役による双対構造は様々な問題に見られる

- **分類問題（今日の前半）**
- **最適輸送（今日の後半）**
- ブースティング
- 強化学習・Bellman 方程式
- バンディット問題
- 「微分可能」線形計画問題
- 陰的正則化

分類問題

- 確率単体上の正しい点を推定するために Bregman 射影が有用
- 非対称リンク関数からでも自然な損失を導出可能

最適輸送

- 輸送多面体上の最適化効率の向上のために凸緩和が有用
- 解のスパース性をリンク関数から自然にデザイン可能

凸共役による双対構造は様々な問題に見られる

- **分類問題（今日の前半）**
- **最適輸送（今日の後半）**
- ブースティング
- 強化学習・Bellman 方程式
- バンディット問題
- 「微分可能」線形計画問題
- 陰的正則化

強化学習

- 行動空間上の**探索を促進**するためにエントロピー正則化が有効

バンディット問題

- 敵対者に対して**敏感になりすぎない**よう内点解に留まるのが有用

「微分可能」線形計画

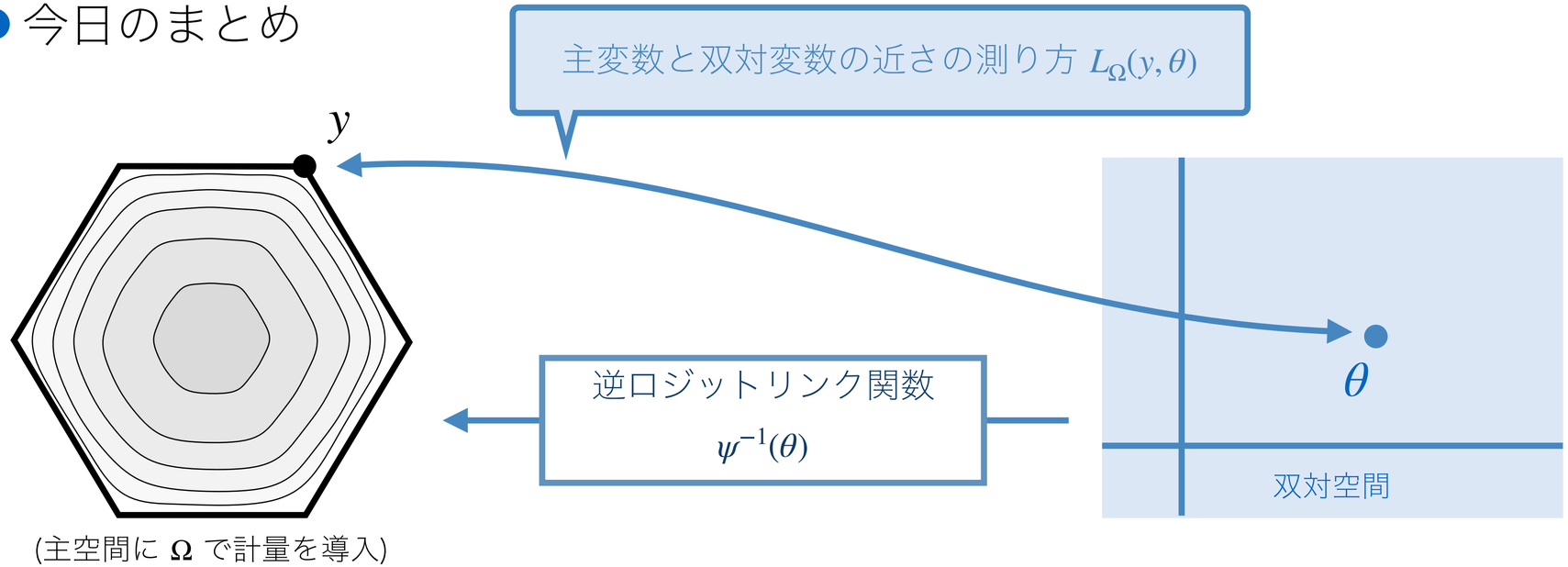
- 線形計画の凸緩和によって解が入力に対して**敏感になりすぎない**

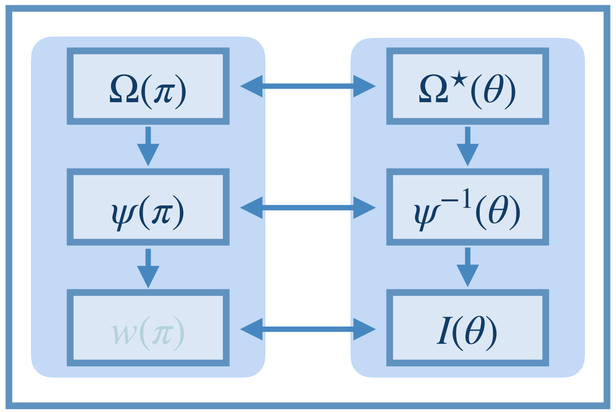
陰的正則化

- 深層学習の補間レジームでは、モデルアーキテクチャから定まる凸関数にもとづく正則化が陰的にかかっていると**解釈可能**

機械学習と凸共役の交わり

- 今日のまとめ





- 双対構造の活用方法
  - ❖ スパース性、入力敏感性の低減、探索の促進、etc.
- 機械学習の様々な問題の背後に潜む双対構造に関する研究はまだまだ途上
  - ❖ わかっていないことがたくさんある！