学習アルゴリズムの幾何学的性質

大阪市立大学情報幾何研究集会

村田 昇

早稲田大学 理工学術院

2009 年 1 月 25 日

共同研究者:

- 竹之内 高志
  (奈良先端科学技術大学院大学)
- 金森 敬文
  (名古屋大学)
- 江口 真透
  (統計数理研究所)
- 藤本 悠
  (早稲田大学)

Part I

Boosting Algorithm

Outline

1. 集団学習と Boosting Algorithm
2. Bregman Divergence と判別モデル
3. Boosting Algorithm の幾何学的描像

Outline

1. **集団学習と Boosting Algorithm**
   - 集団学習の仕組み
   - Boosting の例 (AdaBoost)
   - Boosting Algorithm
   - Boosting の仕組み

集団学習の仕組み (その 1)

- 三人組で様々なジャンル (歴史，政治，芸能，スポーツなど 10 種類) の問題に答えるクイズ番組
  - そこそこできの良い三人組
  - それほどできの良くない三人組

集団学習の仕組み (その 1)

- 三人組で様々なジャンル (歴史，政治，芸能，スポーツなど 10 種類) の問題に答えるクイズ番組
  - そこそこできの良い三人組
    - 8 種類のジャンルに対応できる
    - 皆，芸能とスポーツが苦手
    - 間違いが比較的少ない
  - それほどできの良くない三人組
    - 6 種類のジャンルしか理解しない
    - 苦手なジャンルはバラバラ
    - 間違いが比較的多い

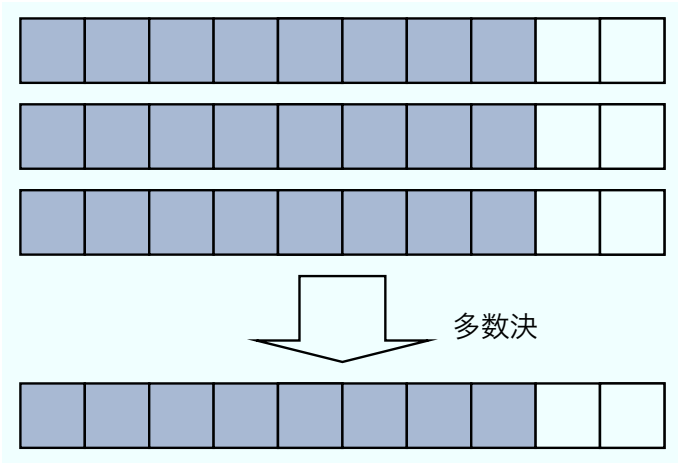
集団学習の仕組み (その 1)

- 三人組で様々なジャンル (歴史，政治，芸能，スポーツなど 10 種類) の問題に答えるクイズ番組
  - そこそこできの良い三人組
    - 8 種類のジャンルに対応できる
    - 皆，芸能とスポーツが苦手
    - 画一的な優等生
  - それほどできの良くない三人組
    - 6 種類のジャンルしか正解しない
    - 苦手なジャンルはバラバラ
    - 個性的な劣等生

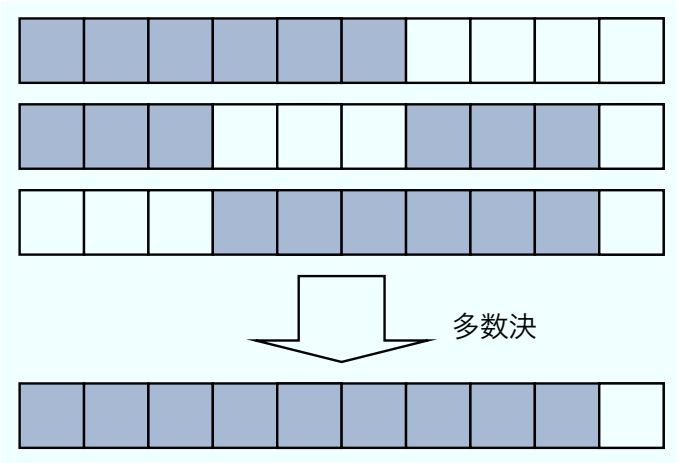
集団学習の仕組み (その 1)

- 三人組で様々なジャンル (歴史，政治，芸能，スポーツなど 10 種類) の問題に答えるクイズ番組
  - そこそこできの良い三人組
    - 8 種類のジャンルに対応できる
    - 皆，芸能とスポーツが苦手
    - 画一的な優等生
  - それほどできの良くない三人組
    - 6 種類のジャンルしか正解しない
    - 苦手なジャンルはバラバラ
    - 個性的な劣等生

集団学習の仕組み (その 2)



できの良い三人組

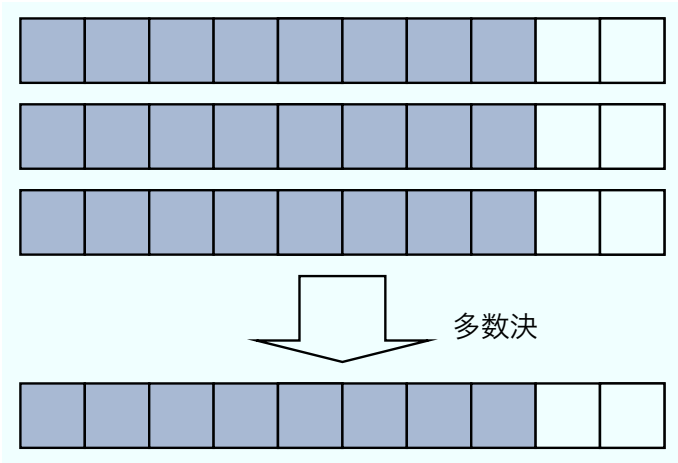


できの良くない三人組

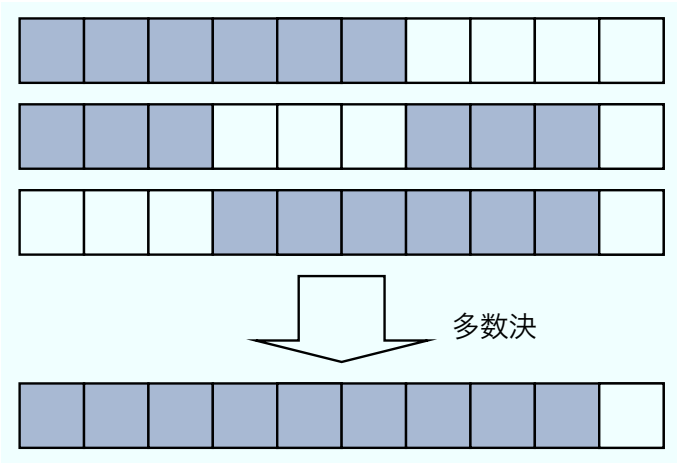
集団学習の要件

- できるだけ多様な人材を集める必要がある.
- 少なくとも半分くらいは正解してくれないと困る.

集団学習の仕組み (その 2)



できの良い三人組



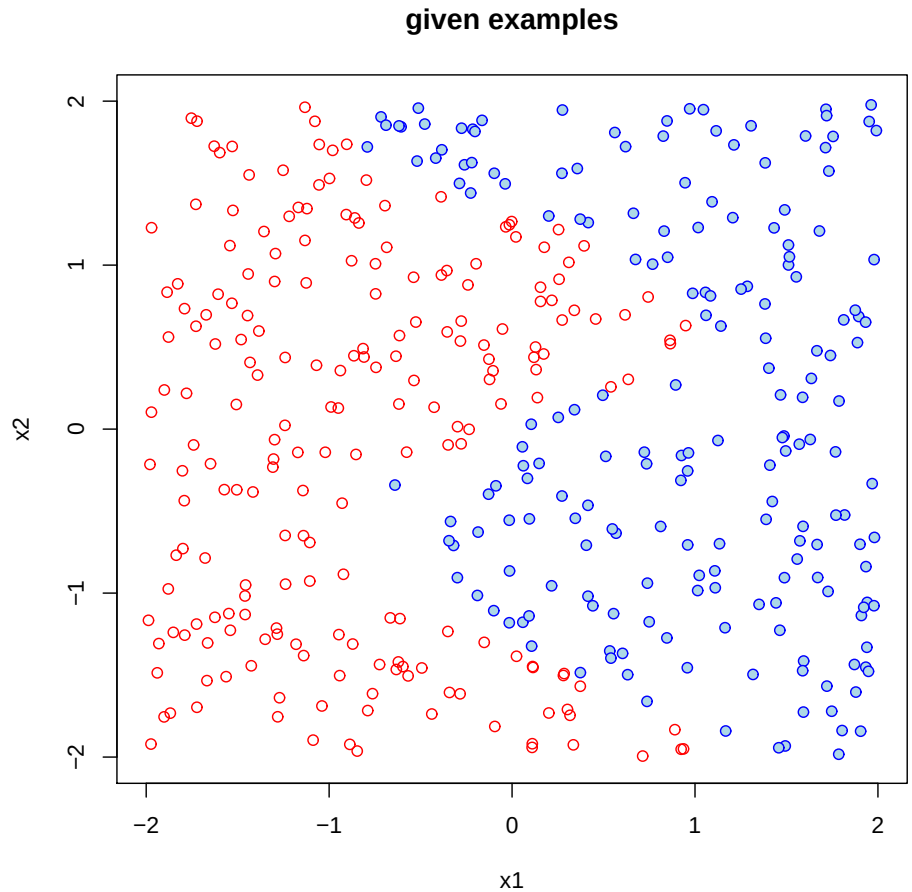
できの良くない三人組

集団学習の要件

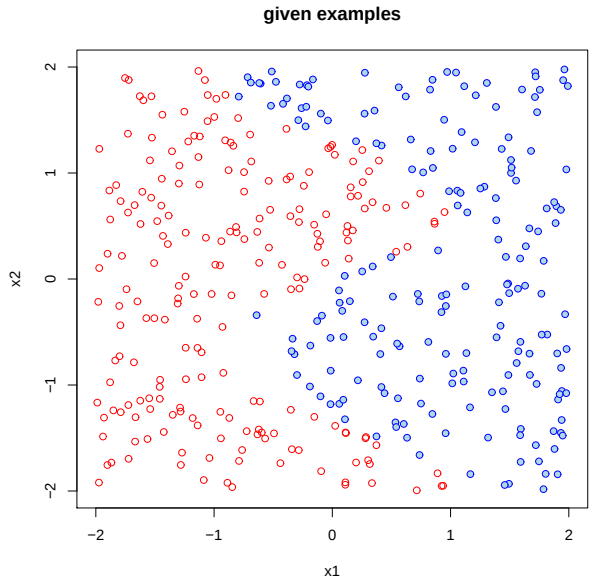
- できるだけ多様な人材を集める必要がある.
- 少なくとも半分くらいは正解してくれないと困る.

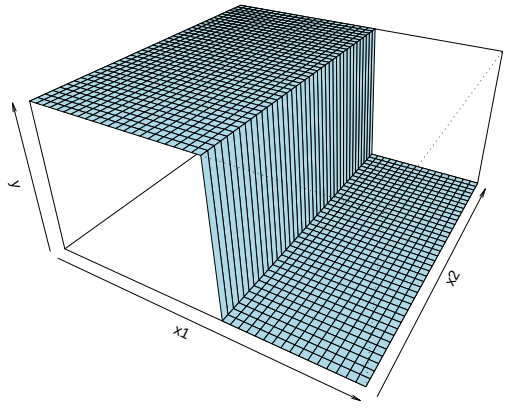
パタン認識の問題

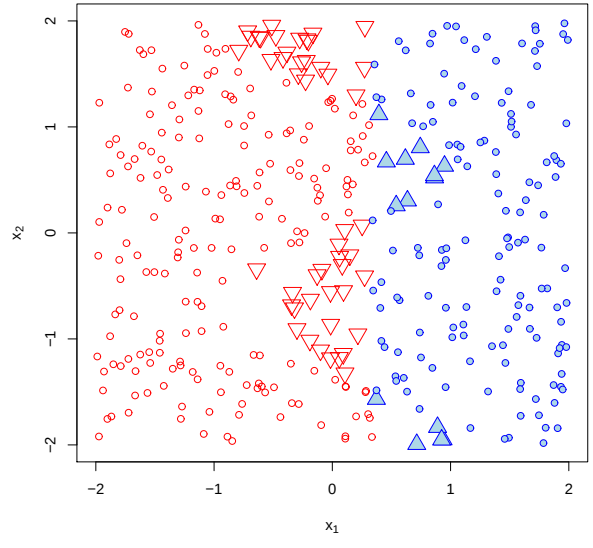
- 特徴量 $\boldsymbol{x} \in \mathcal{X}$ のラベル $y \in \mathcal{Y}$ を予測する.
- 有限個の例題から判別器 $h(x) = \hat{y}$ を構成する.



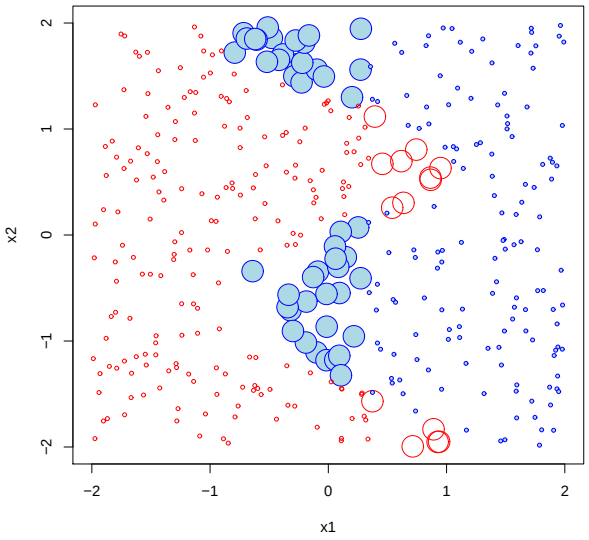
第 1 巡

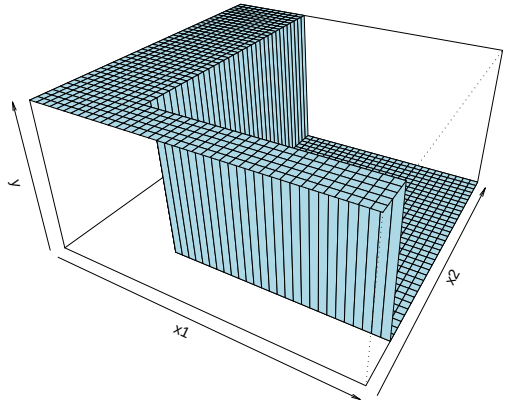


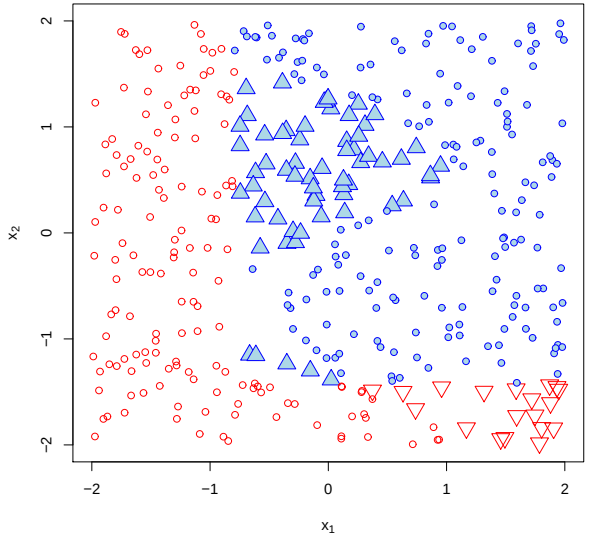




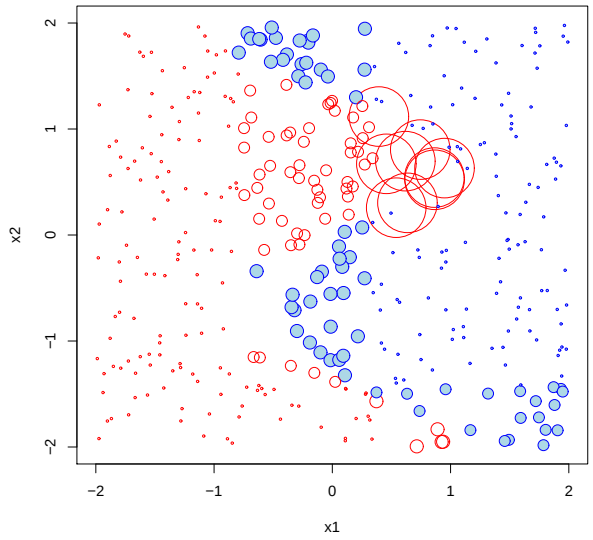
第 2 巡

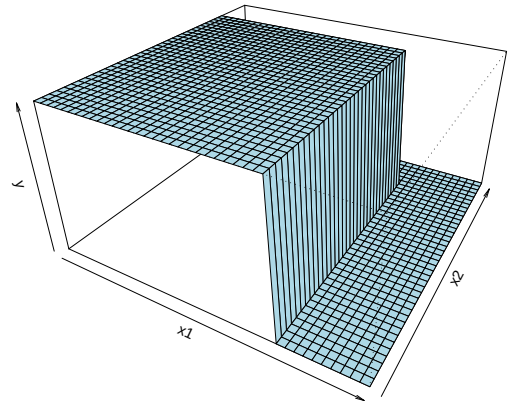


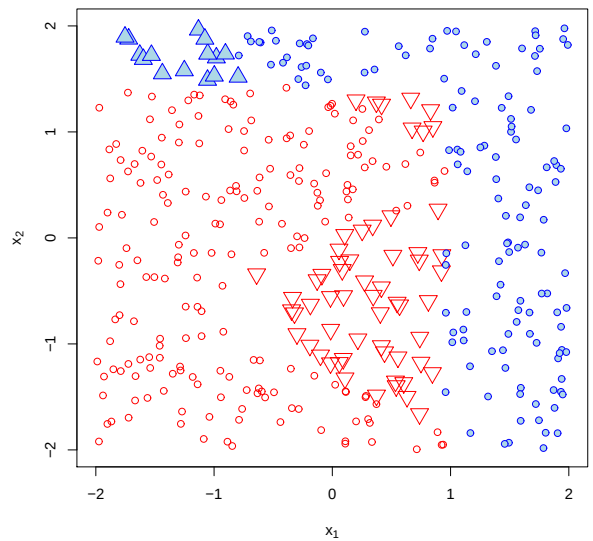




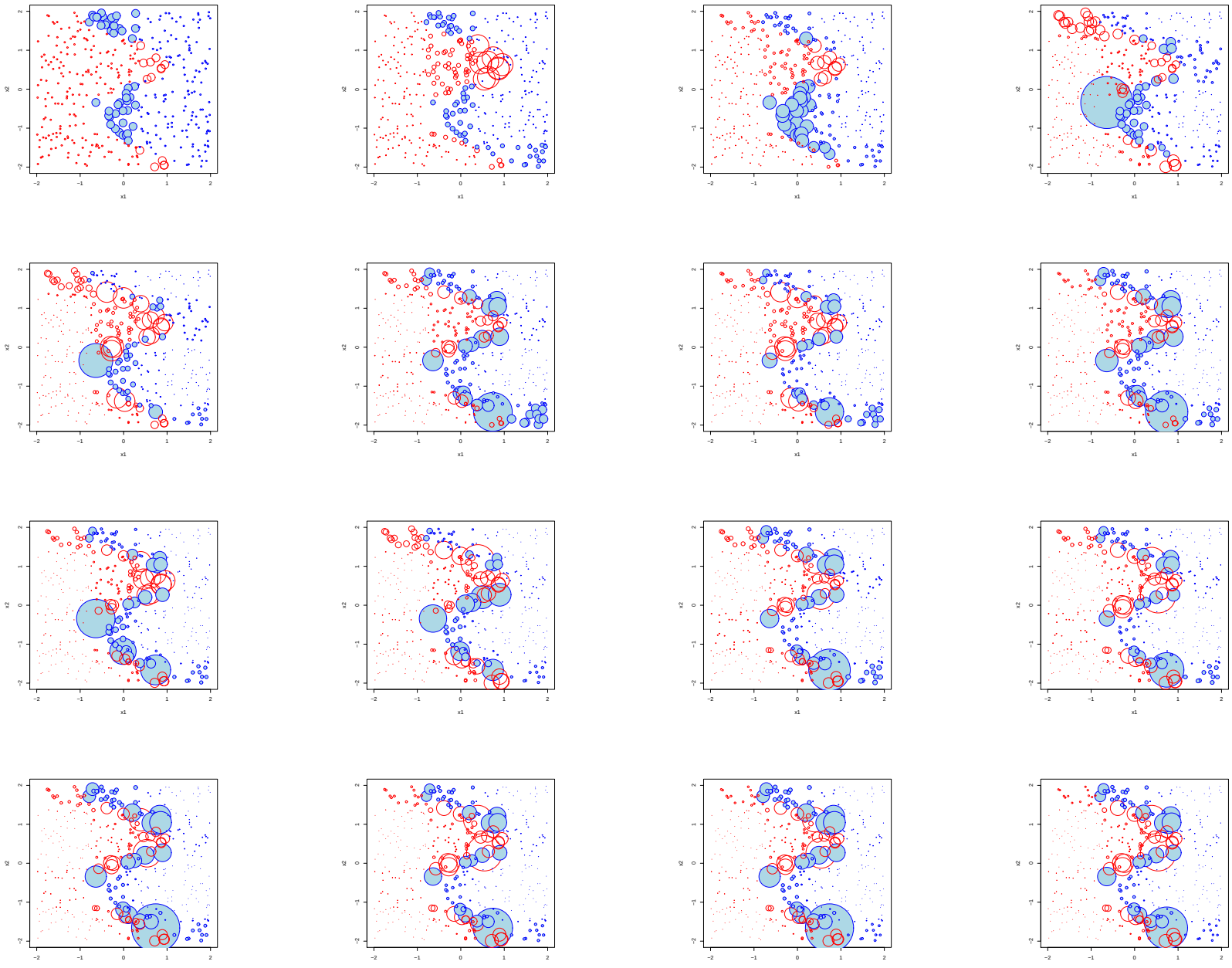
第 3 巡



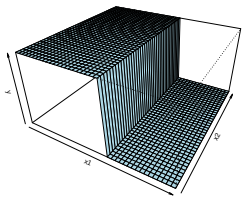
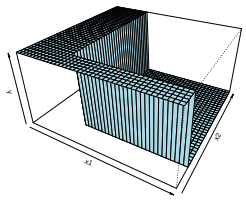
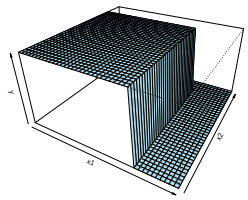
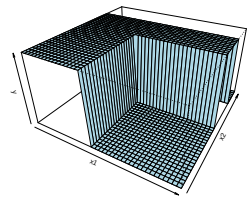
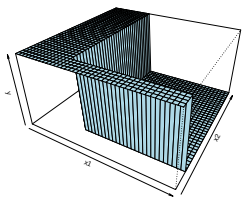
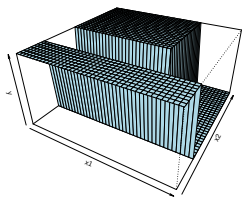
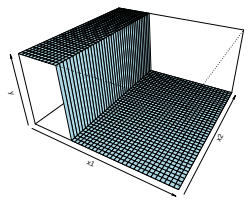
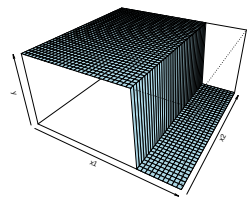
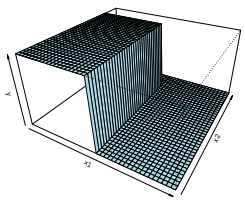
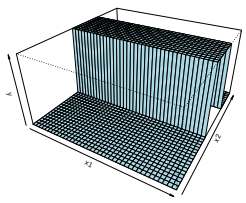
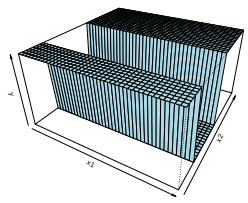
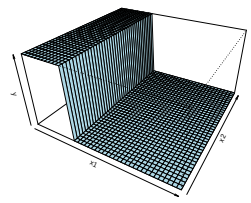
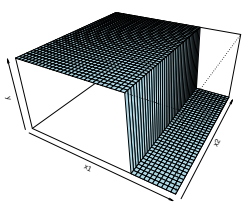
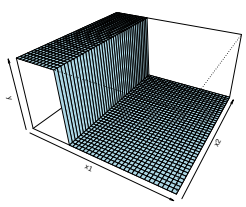
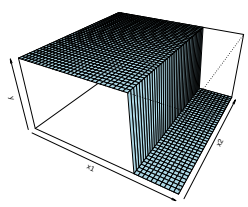
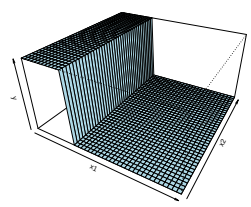




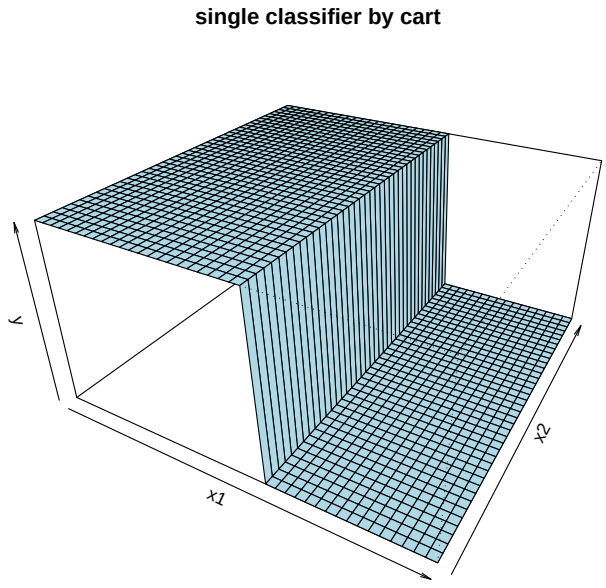
各巡の分布



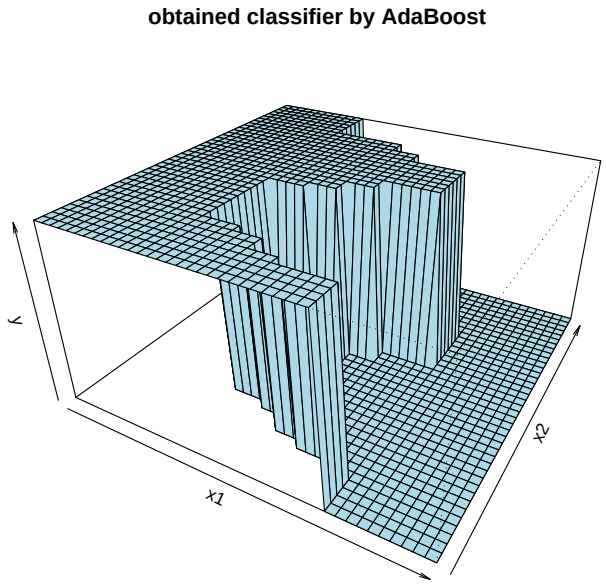
各巡の判別器

Boosting の効果 (判別器)

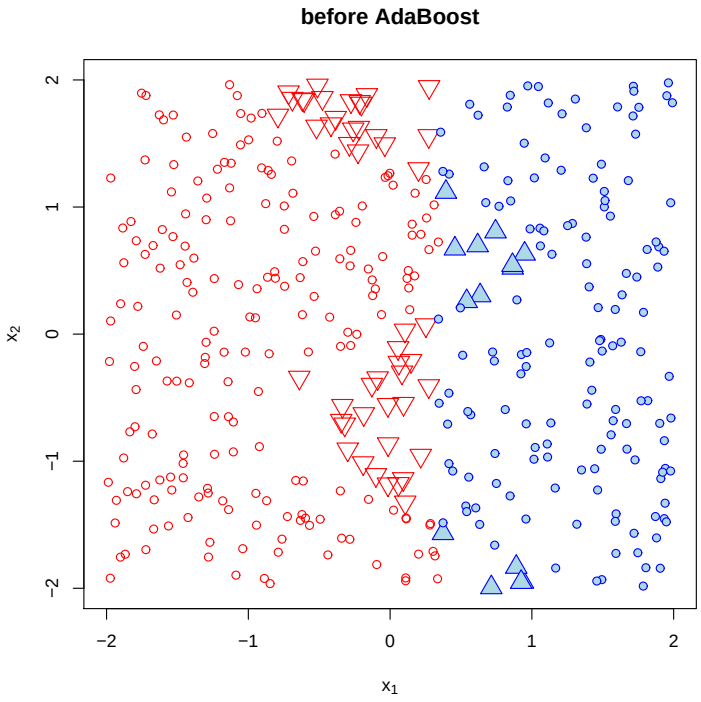


Boosting なし

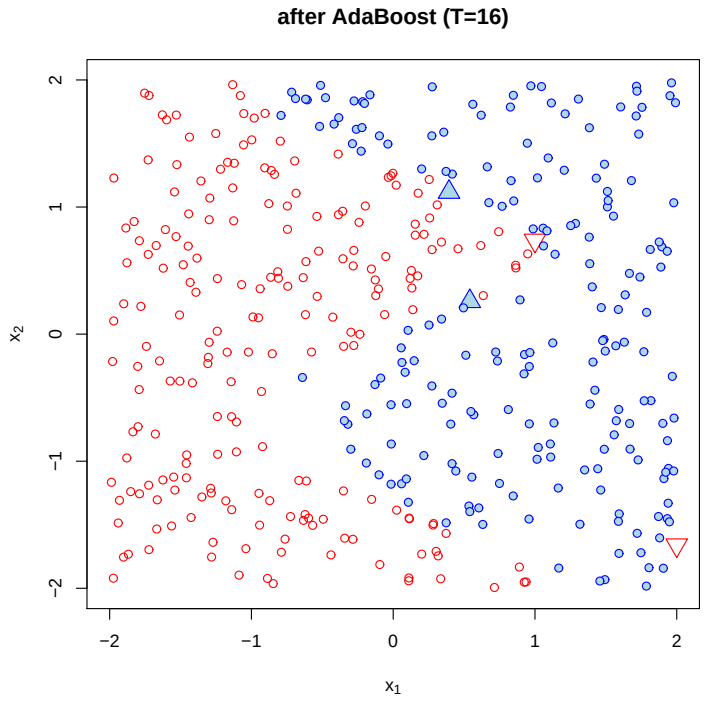


Boosting あり

Boosting の効果 (判別誤り)



Boosting なし



Boosting あり

記号の準備

- **問題**: 特徴量 $\boldsymbol{x} \in \mathcal{X}$ のラベル $y \in \mathcal{Y}$ を予測する
- **判別器**: 集合値関数 h

$$h : \boldsymbol{x} \in \mathcal{X} \mapsto \mathcal{C} \subset \mathcal{Y}$$

- **決定関数**: 判別器の別表現

$$f(\boldsymbol{x}, y) = \begin{cases} 1, & \text{if } y \in h(\boldsymbol{x}), \\ 0, & \text{otherwise}, \end{cases}$$

- **多数決判別器**: 複数の判別器の決定関数の線形結合

$$H(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y)$$

Boosting Algorithm (1)

- **入力:**
  n 個の例題 $\{(\boldsymbol{x}_i, y_i); \boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \ldots, n\}$,
  単調増加な凸関数 U.
- **初期化:**
  分布 $D_1(i, y) = 1/n(|\mathcal{Y}| - 1)$ $(i = 1, \ldots, n)$,
  結合決定関数 $F_0(\boldsymbol{x}, y) = 0$.
- **繰り返し:** 以下の step を繰り返す $(t = 1, \ldots, T)$.

Boosting Algorithm (2)

(繰り返しの中)

- **step 1:** 以下で定める分布 D_t のもとでの決定関数 f(判別器 h) の誤差 ϵ_t を (近似的に) 最小化する決定関数を 1 つ選ぶ.

$$\epsilon_t(f) = \sum_{i=1}^{n} \sum_{y \neq y_i} \frac{f(\boldsymbol{x}_i, y) - f(\boldsymbol{x}_i, y_i) + 1}{2} D_t(i, y)$$

$$f_t(\boldsymbol{x}, y) = \arg\min_{f \in \mathcal{F}} \epsilon_t(f).$$

Boosting Algorithm (3)

(繰り返しの中)

- **step 2:** 信頼度 α_t を計算する.

$$\alpha_t = \arg\min_{\alpha} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} U\Big(F_{t-1}(\boldsymbol{x}_i, y) + \alpha f_t(\boldsymbol{x}_i, y) - F_{t-1}(\boldsymbol{x}_i, y_i) - \alpha f_t(\boldsymbol{x}_i, y_i) \Big).$$

Boosting Algorithm (4)

(繰り返しの中)

- **step 3:** 結合決定関数 F_t と分布 D_t を更新する.

$$F_t(\boldsymbol{x}, y) = F_{t-1}(\boldsymbol{x}, y) + \alpha_t f_t(\boldsymbol{x}, y),$$

$$D_{t+1}(i, y) \propto U'\left(F_t(\boldsymbol{x}_i, y) - F_t(\boldsymbol{x}_i, y_i)\right),$$

$$\text{ただし} \sum_{i=1}^{n} \sum_{y \neq y_i} D_{t+1}(i, y) = 1.$$

Boosting Algorithm (5)

- **出力:**
  多数決判別器を構成する.

$$H(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} F_T(\boldsymbol{x}, y)$$

$$= \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y).$$

AdaBoost Algorithm

- $U(z) = \exp(z)$ の場合 (以下の step が簡略化される)
  - **step 2:**

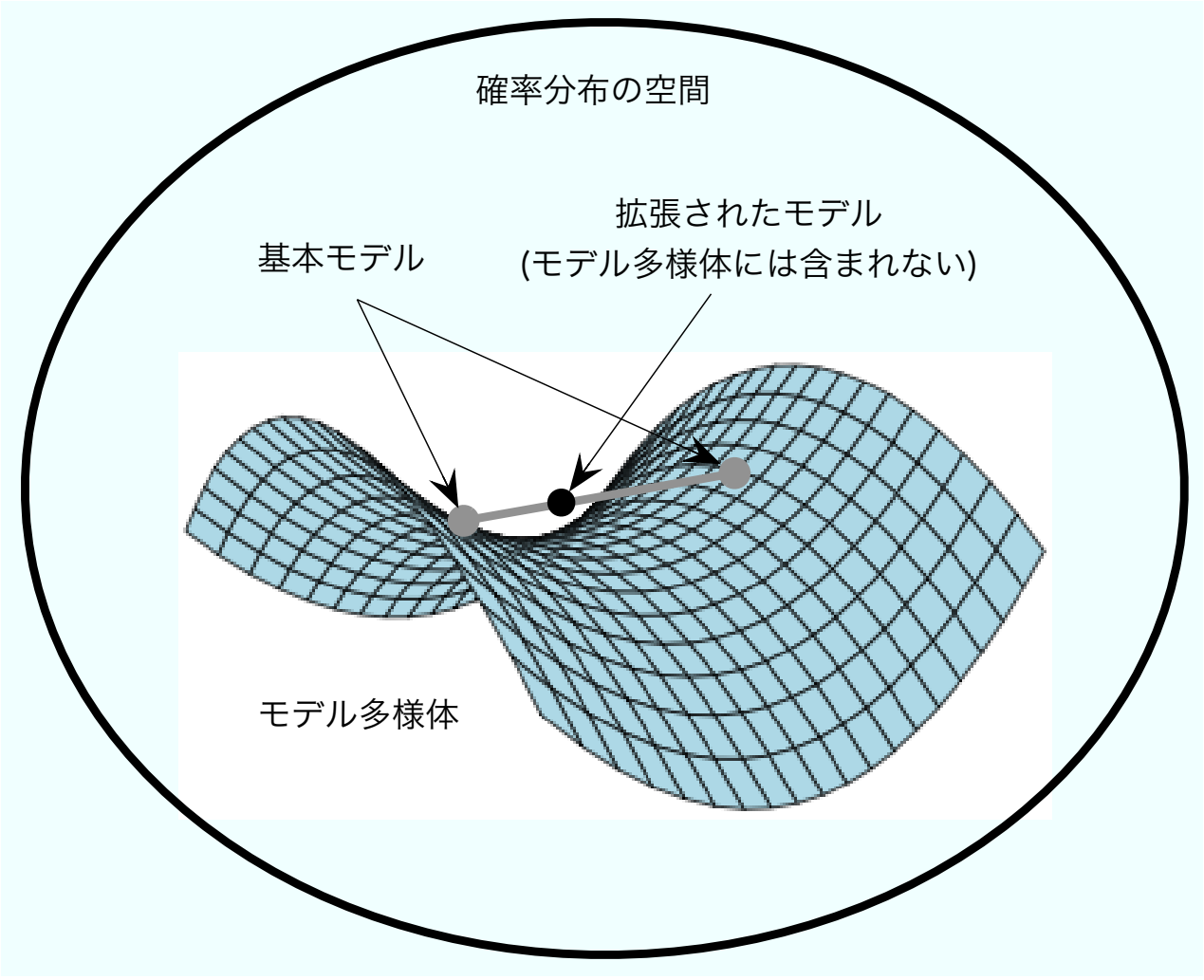
$$\alpha_t = \frac{1}{2} \log \frac{1 - \epsilon_t(f_t)}{\epsilon_t(f_t)},$$

  - **step 3:**

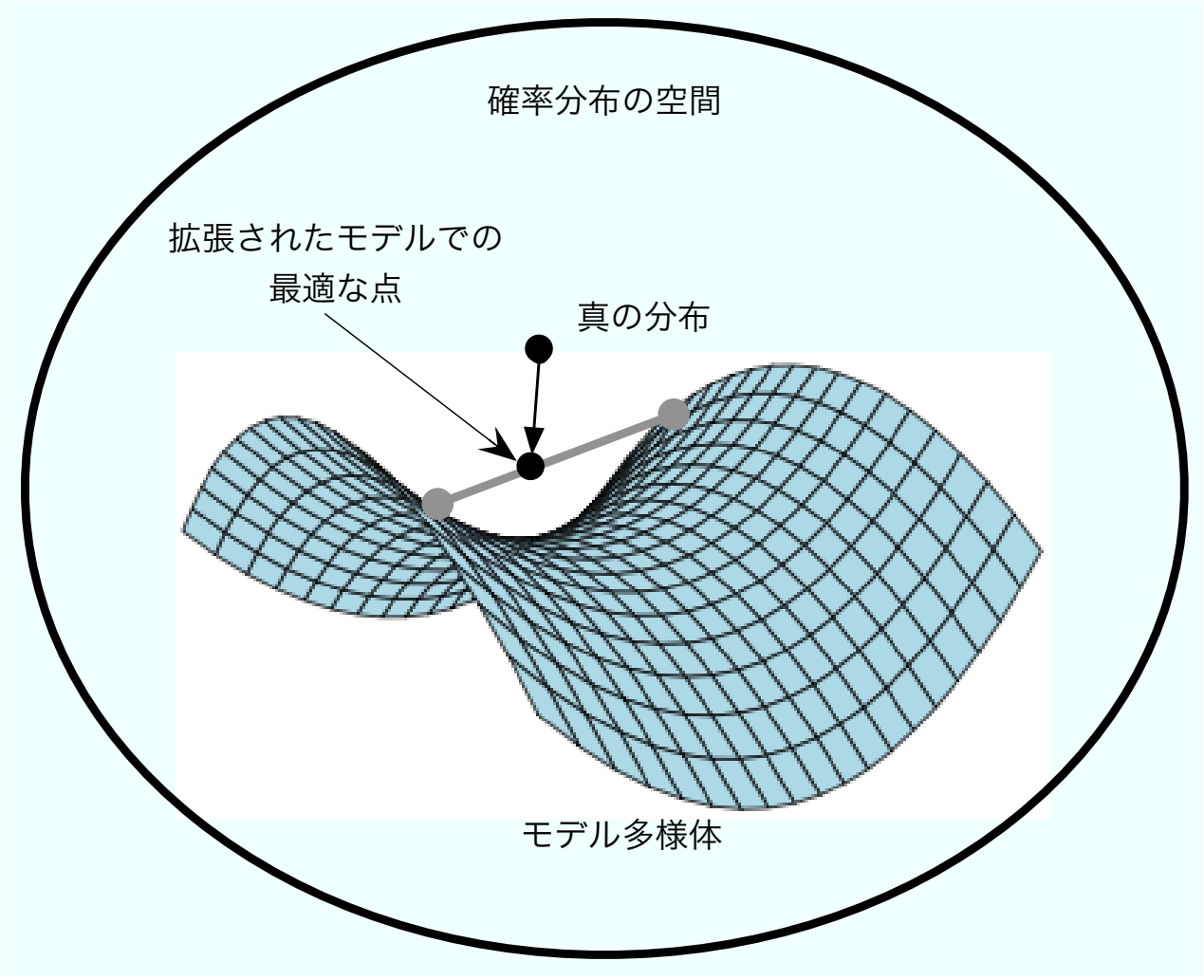
$$D_{t+1}(i, y) \propto \exp\{F_t(\boldsymbol{x}_i, y) - F_t(\boldsymbol{x}_i, y_i)\}$$

混合によるモデルの拡大

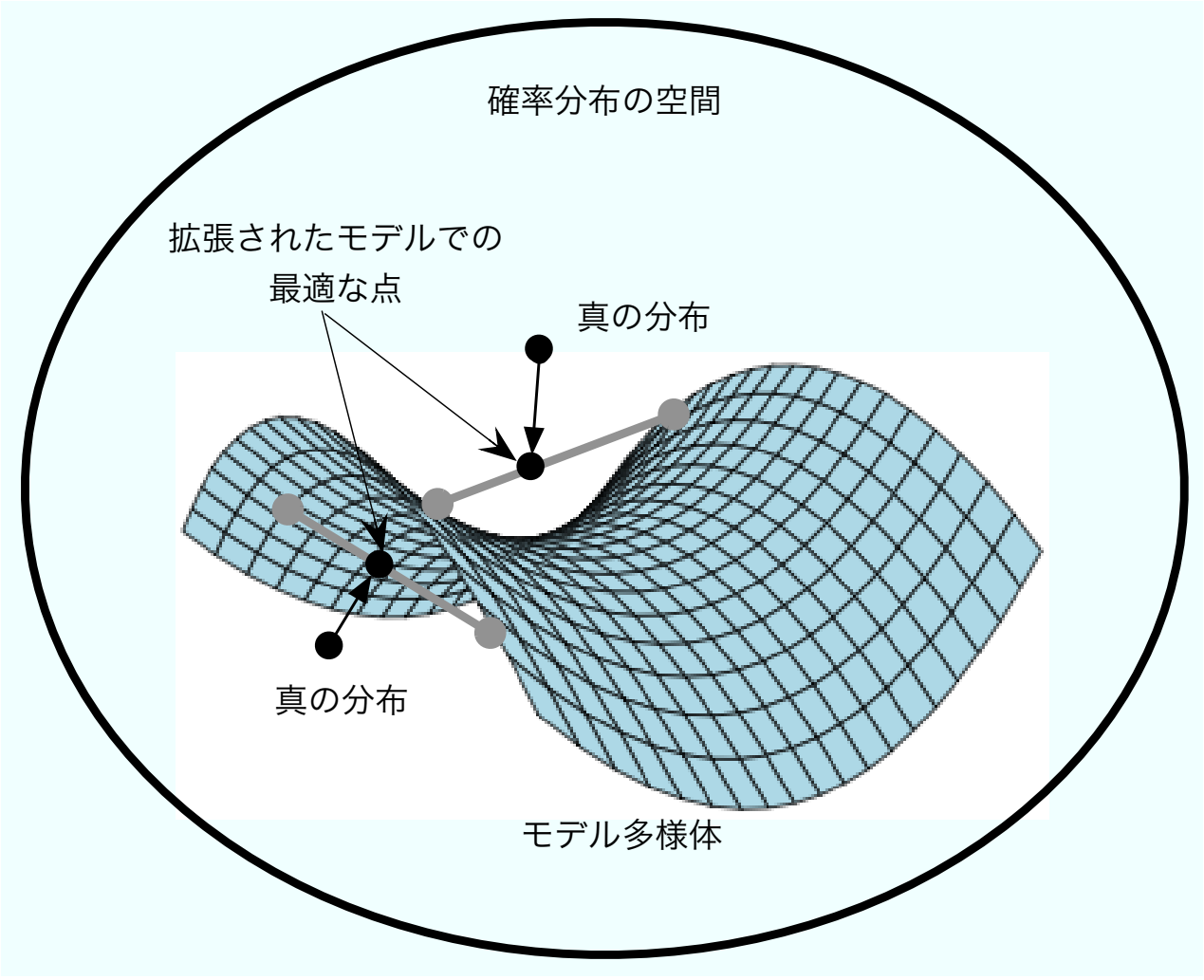
混合によるモデルの拡大



混合によるモデルの拡大



混合によるモデルの拡大



代理損失による解釈 (1)

- 判別関数:F

$$(x \text{ のラベルの予測値}) = \arg\max_{y \in \mathcal{Y}} F(\boldsymbol{x}, y)$$

- 判別誤り

$$I\left(\max_{y \in \mathcal{Y}} F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i)\right) = \begin{cases} 1, & \text{引数が正} \\ 0, & \text{それ以外} \end{cases}$$

- 誤り率

$$L_{0\text{-}1}(F) = \frac{1}{n} \sum_{i=1}^{n} I\left(\max_{y \in \mathcal{Y}} F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i)\right)$$

代理損失による解釈 (2)

- 誤り率の上界

$$L_{\text{0-1}}(F) \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} I\left(F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i)\right)$$

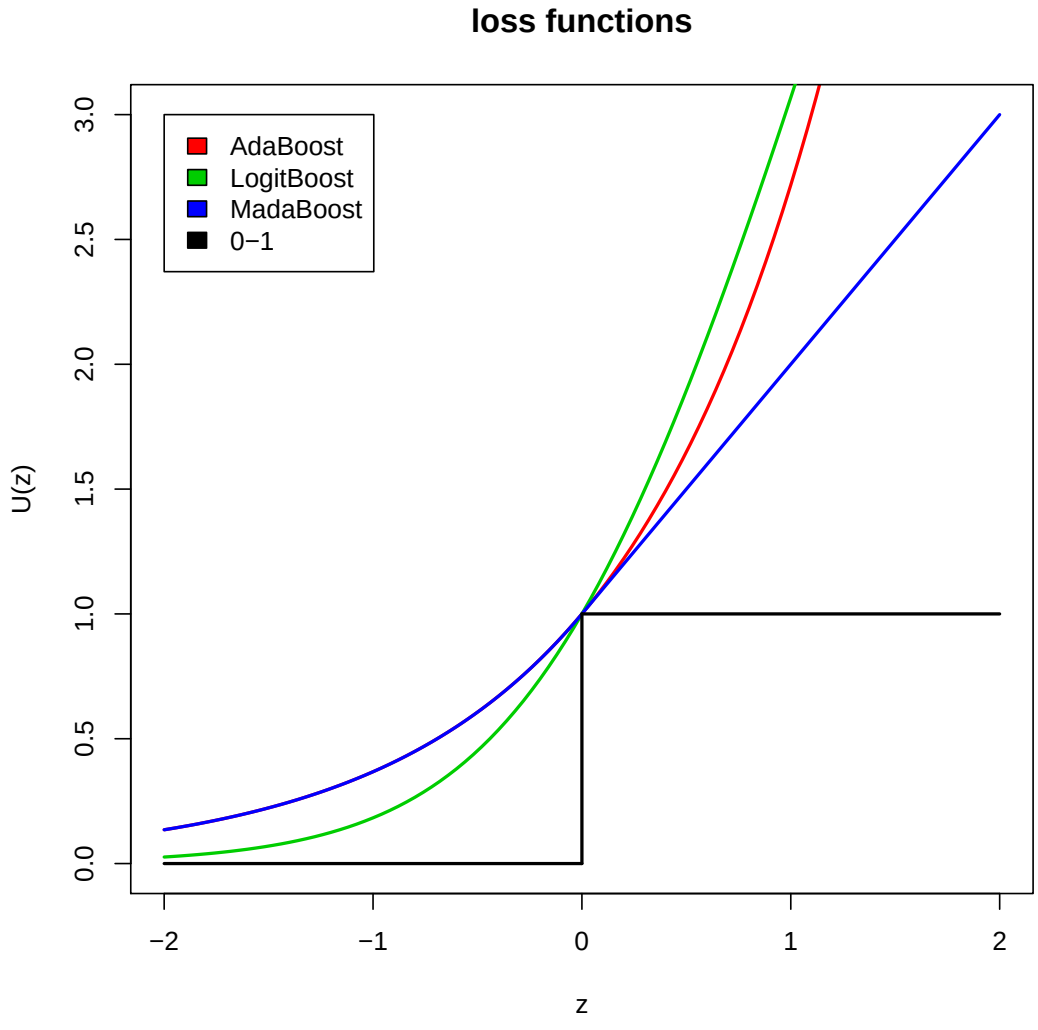
- 0-1 損失の代理損失 (surrogate loss)

$$L_U(F) = \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} U(F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i))$$

- 代理損失の最適化

$$L_{\text{0-1}}(F) \leq L_U(F) \rightarrow \text{最小化}$$

代理損失の例


loss functions
AdaBoost
LogitBoost
MadaBoost
0−1
U(z)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
−1
0
1
2
z

Outline

2. **Bregman Divergence と判別モデル**
   - 有限測度の空間
   - Bregman Divergence
   - ピタゴラスの定理
   - 平坦部分空間
   - 直交葉層化
   - Bregman Divergence に付随する判別モデル

有限測度の空間

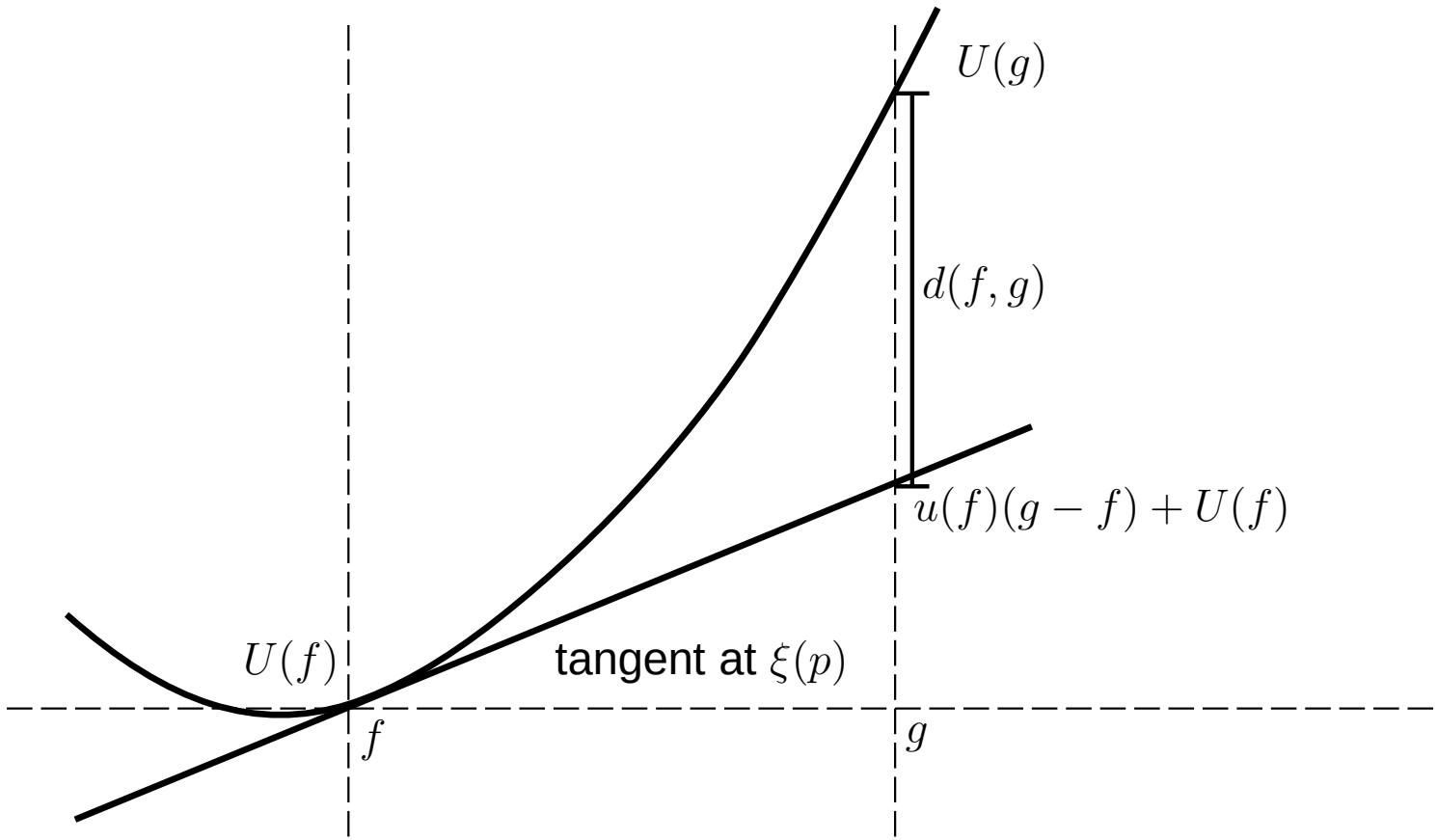
- 正値の条件付測度の空間

$$\mathcal{M} = \left\{ m(y|\boldsymbol{x}) \,\middle|\, \sum_{y \in \mathcal{Y}} m(y|\boldsymbol{x}) < \infty \text{ (a.e. } \boldsymbol{x}) \right\}$$

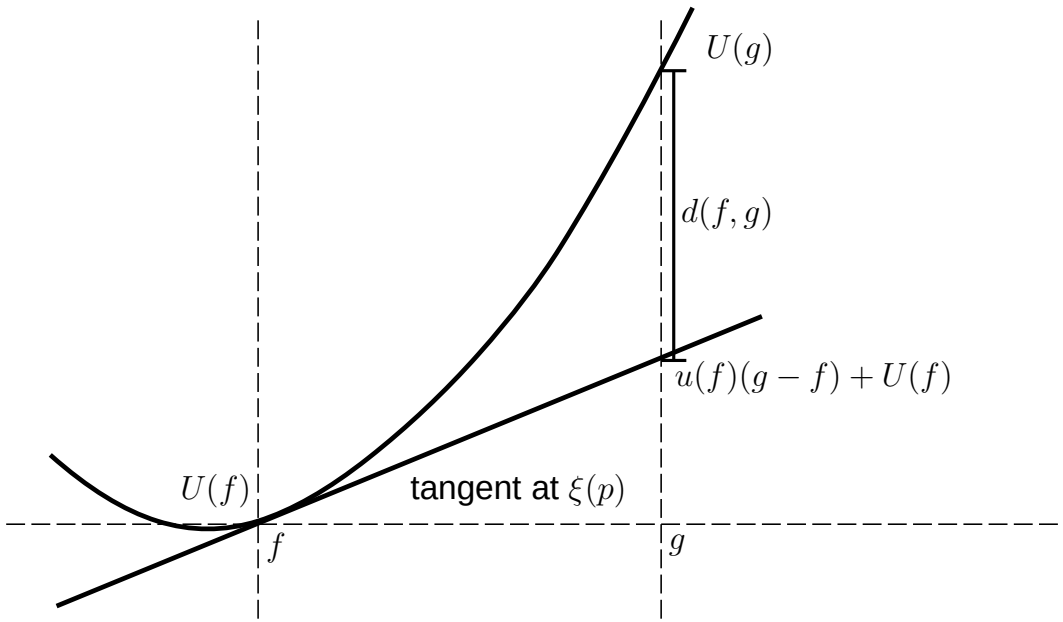
- 条件付確率の空間を含む

$$\mathcal{P} = \left\{ m(y|\boldsymbol{x}) \,\middle|\, \sum_{y \in \mathcal{Y}} m(y|\boldsymbol{x}) = 1 \text{ (a.e. } \boldsymbol{x}) \right\}$$

Bregman Divergence (1)



Bregman Divergence (2)

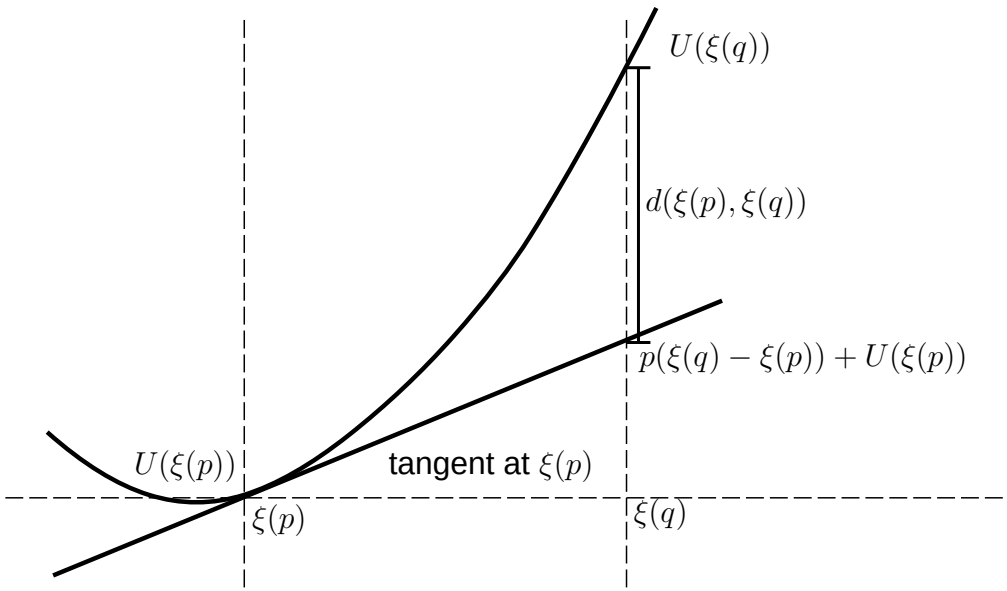


U: 凸関数 (下に)

$$d(f,g) = U(g) - U(f) - U'(f)(g-f) > 0$$

$$D(f,g) = \int d(f(x), g(x))dx \geq 0, \text{ (正値, 非対称)}$$

U-Divergence (1)



U: 凸関数 (下に)

$$u = U,$$
$$\xi = (U')^{-1}$$

$$\begin{aligned} d_U(p,q) &= U(\xi(q)) - U(\xi(p)) - p(\xi(q)-\xi(p)) \\ &= [U(\xi(q)) - p(\xi(q))] - [U(\xi(p)) - p(\xi(p))] \end{aligned}$$

$$D_U(p,q) = \int d_U(p(x),q(x))d\mu(x)$$

U-Divergence (2)

- U:R 上の凸関数 (導関数 $u = U'$, その逆関数 $\xi = u^{-1}$)
- $\mu(\boldsymbol{x})$ のもとでの $p, q \in \mathcal{M}$ の U-cross-entropy:

$$H_U(p, q; \mu) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} \{U(\xi(q(y|\boldsymbol{x}))) - p(y|\boldsymbol{x})\xi(q(y|\boldsymbol{x}))\}\, \mu(\boldsymbol{x}) d\boldsymbol{x}$$

- p から q への U-divergence (Bregman divergence の一種):

$$D_U(p, q; \mu) = H_U(p, q; \mu) - H_U(p, p; \mu)$$

ポテンシャル表現

- U の Legendre 変換

$$U^*(\zeta) = \sup_z \{\zeta z - U(z)\}$$

- 2 つの表現:
  - m-表現: p
  - U-表現: $\tilde{p} = \xi(p), \quad \xi = u^{-1} = U'^{-1}$
- ポテンシャル表現

$$d_U(p, q) = U^*(p) + U(\tilde{q}) - p\tilde{q}$$

$$D_U(p, q) = \int d_U(p(x), q(x)) d\mu(x)$$

計量と接続

- パラメトリックモデル

$$\left\{p(x;\theta),\ \theta \in \Theta \subset R^d\right\} \quad \partial_i = \frac{\partial}{\partial \theta^i}$$

- Riemann 計量

$$g^{(U)}(\partial_i, \partial_j)(\theta) = \int \partial_i p(x;\theta)\partial_j \xi(p(x;\theta))d\mu(x)$$

- 双対な接続

$$g^{(U)}(\nabla^{(U)}_{\partial_i}\partial_j, \partial_k)(\theta) = \int \partial_i\partial_j p(x;\theta)\partial_k \xi(p(x;\theta))d\mu(x)$$

$$g^{(U)}({}^*\nabla^{(U)}_{\partial_i}\partial_j, \partial_k)(\theta) = \int \partial_k p(x;\theta)\partial_i\partial_j \xi(p(x;\theta))d\mu(x)$$

特徴 (1)

- Bregman divergence の正値性より

$$H_U(p, q; \mu) \geq H_U(p, p; \mu)$$

- q に関する最小化の等価性

$$\arg\min_q D_U(p, q; \mu) = \arg\min_q H_U(p, q; \mu).$$

(KL-divergence の最小化と最尤推定の関係と同様)

特徴 (2)

- 経験分布が直接代入できる

$$\textbf{条件付:}\ \tilde{p}(y|\boldsymbol{x}) = \begin{cases} I(y = y_i), & \text{if } \boldsymbol{x} = \boldsymbol{x}_i, \\ \dfrac{1}{|\mathcal{Y}|}, & \text{otherwise,} \end{cases}$$

$$\textbf{周辺:}\ .\tilde{\mu}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\boldsymbol{x} - \boldsymbol{x}_i).$$

$$\begin{aligned} \tilde{q} &= \arg\min_{q\in\mathcal{Q}} H_U(\tilde{p}, q; \tilde{\mu}) \\ &= \arg\min_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{y\in\mathcal{Y}} U(\xi(q(y|\boldsymbol{x}_i))) - \xi(q(y_i|\boldsymbol{x}_i))\right] \end{aligned}$$

関数 U の例 (1)

- **exponential (Kullback-Leibler):**

$$U(z) = \exp(z),\ u(z) = \exp(z),\ \xi(z) = \log(z)$$

- **β-type:**

$$U(z) = \frac{(\beta z + 1)^{\frac{\beta+1}{\beta}}}{\beta + 1},\ u(z) = (\beta z + 1)^{\frac{1}{\beta}},\ \xi(z) = \frac{z^{\beta} - 1}{\beta}$$

- **η-type:**

$$U(z) = \exp(z) - \eta z,\ u(z) = \exp(z) - \eta,\ \xi(z) = \log(z + \eta)$$

関数 U の例 (2)

- **β-type ($\beta = 1$):**

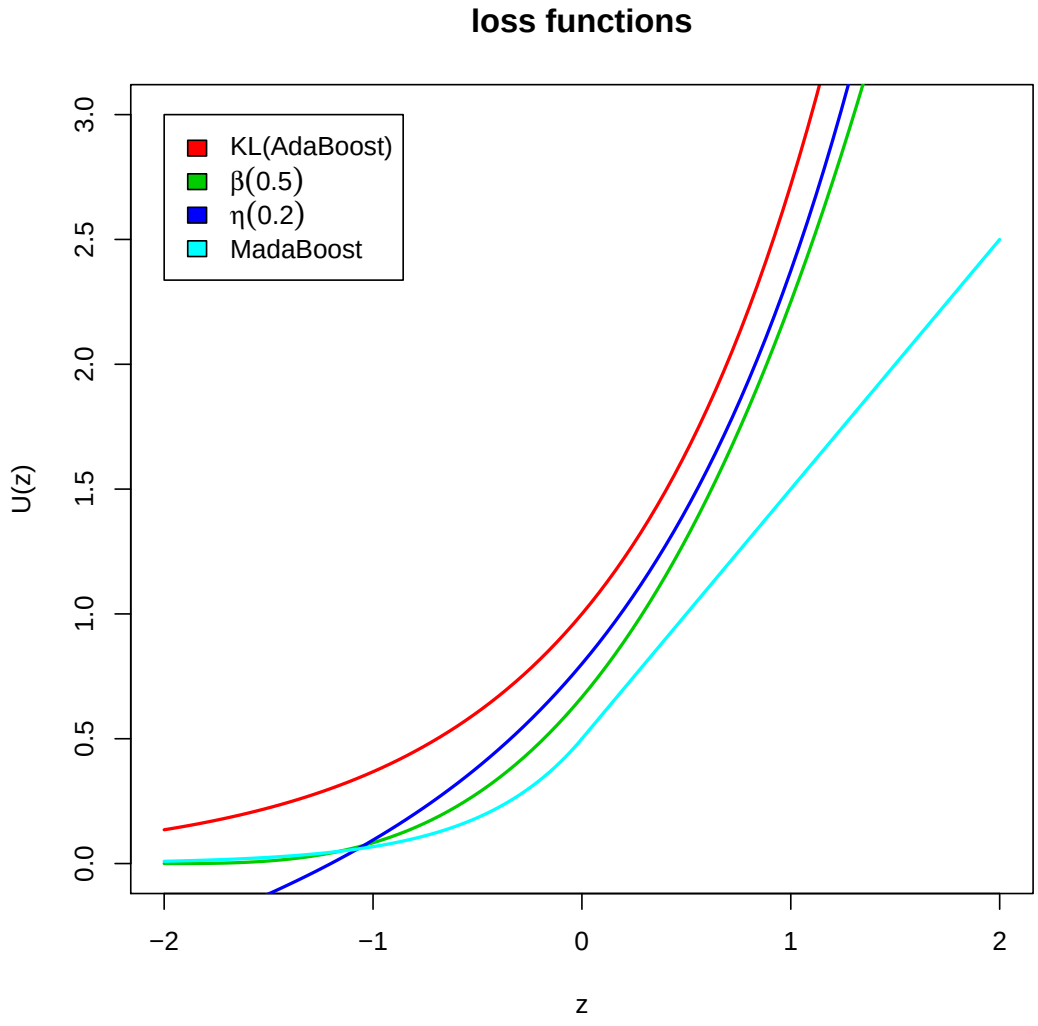
$$U(z) = \frac{1}{2}(z+1)^2,\ u(z) = z+1,\ \xi(z) = z-1$$

- **MadaBoost: (狭義の凸ではない)**

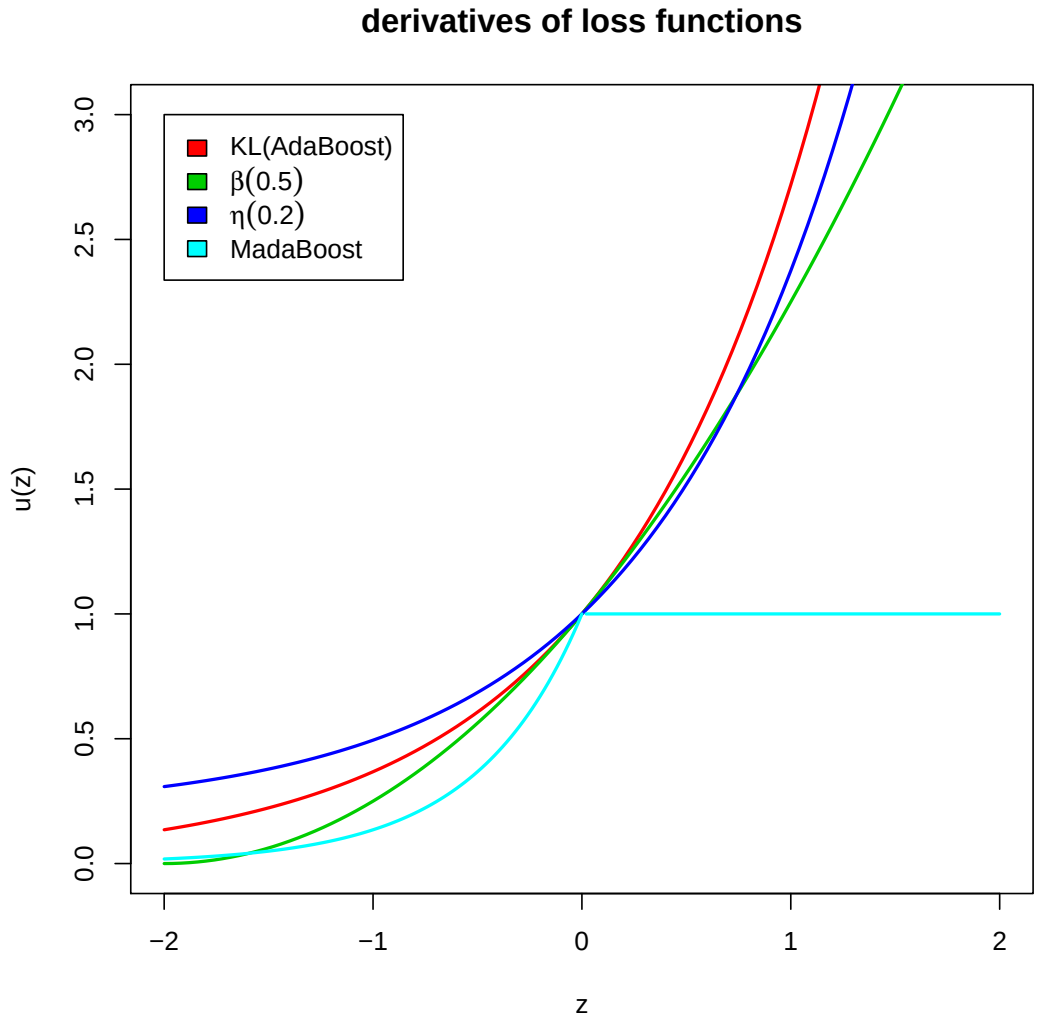
$$U(z) = \begin{cases} z + \frac{1}{2} & z \geq 0, \\ \frac{1}{2}\exp(2z) & z < 0, \end{cases}$$

$$u(z) = \begin{cases} 1 & z \geq 0, \\ \exp(2z) & z < 0 \end{cases} \quad \xi(z) = \frac{1}{2}\log(z) \quad (z \leq 1)$$

関数 U の比較


loss functions
KL(AdaBoost)
β(0.5)
η(0.2)
MadaBoost
U(z)
z
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
−1
0
1
2

関数 U の導関数


derivatives of loss functions
KL(AdaBoost)
β(0.5)
η(0.2)
MadaBoost
u(z)
z
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
−1
0
1
2

ピタゴラスの定理 (1)

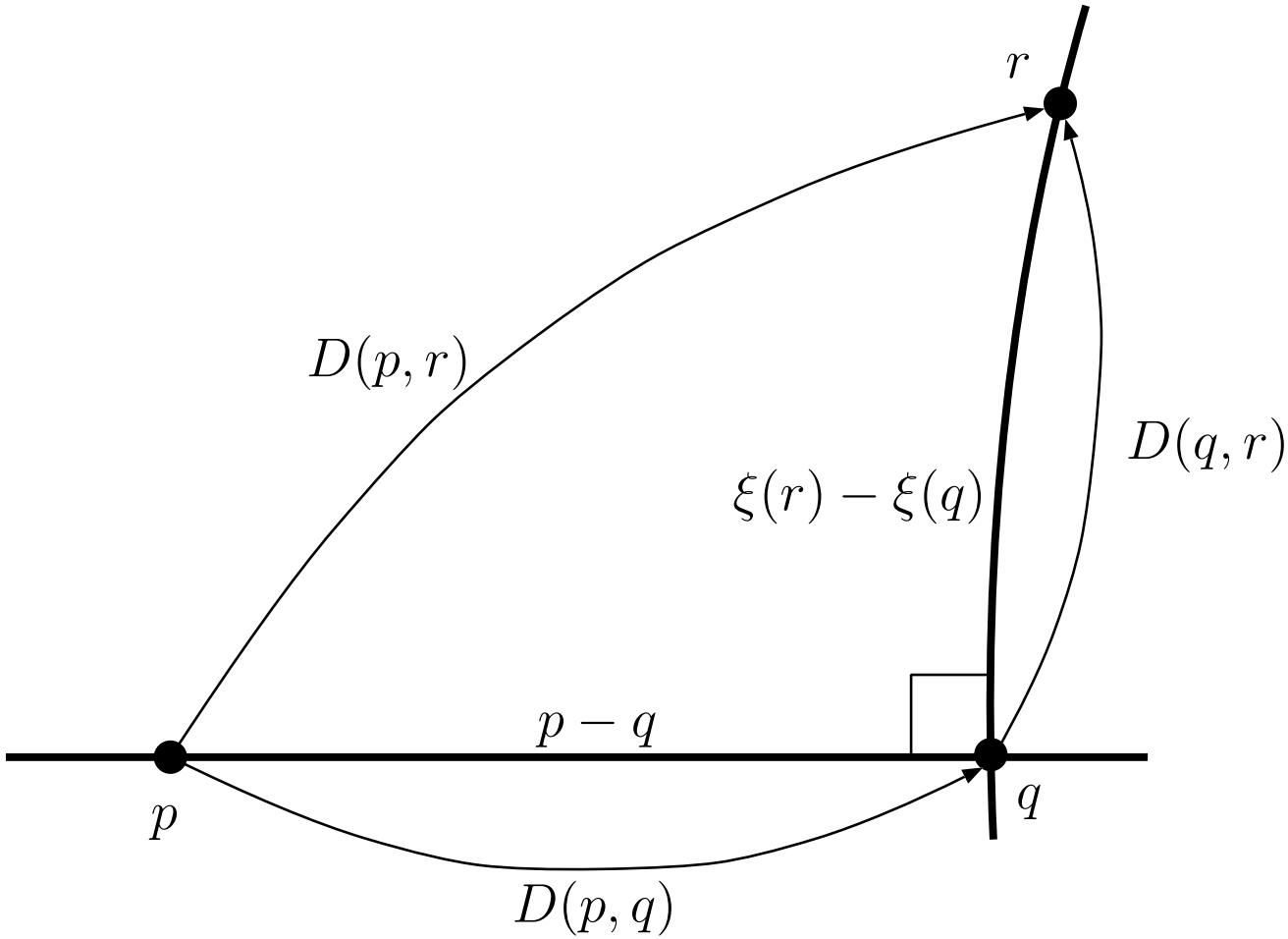
- 3 点 $p, q, r \in \mathcal{M}$ を考える
- $p - q$ と $\xi(r) - \xi(q)$ が μ のもとで直交する

$$\int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} (p(y|\boldsymbol{x}) - q(y|\boldsymbol{x})) \times (\xi(r(y|\boldsymbol{x})) - \xi(q(y|\boldsymbol{x}))) \mu(\boldsymbol{x}) d\boldsymbol{x} = \langle p - q, \xi(r) - \xi(q) \rangle_\mu = 0$$

ならば

$$D_U(p, r; \mu) = D_U(p, q; \mu) + D_U(q, r; \mu)$$

ピタゴラスの定理 (2)



平坦性

- m-測地線: m-表現による $p, q \in \mathcal{M}$ の内分点の集合

$$r(x;t) = (1-t) \cdot p(x) + t \cdot q(x), \quad 0 \leq t \leq 1,$$

- U-測地線: U-表現による $p, q \in \mathcal{M}$ の内分点の集合

$$\xi(r(x;t)) = (1-t) \cdot \xi(p(x)) + t \cdot \xi(q(x)), \quad 0 \leq t \leq 1,$$

- 2 つの自然な部分空間
  - m-**平坦部分空間:** m-測地線からなる集合.
  - U-**平坦部分空間:** U-測地線からなる集合.

U-平坦部分空間

- $q_0 \in \mathcal{M}$ と，Λ を有限の添字の集合とする決定関数の集合 $\mathcal{F} = \{f_\lambda(\boldsymbol{x}, y); \lambda \in \Lambda\}$ を用いて定義する

$$\mathcal{Q}_U(q_0, \mathcal{F}) = \Bigg\{ q \in \mathcal{M} \;\Bigg|\; q(y|\boldsymbol{x}) = u\Big(\xi(q_0(y|\boldsymbol{x})) + \sum_{\lambda \in \Lambda} \alpha_\lambda f_\lambda(\boldsymbol{x}, y)\Big),\ \alpha_\lambda \in R \Bigg\}$$

m-平坦部分空間

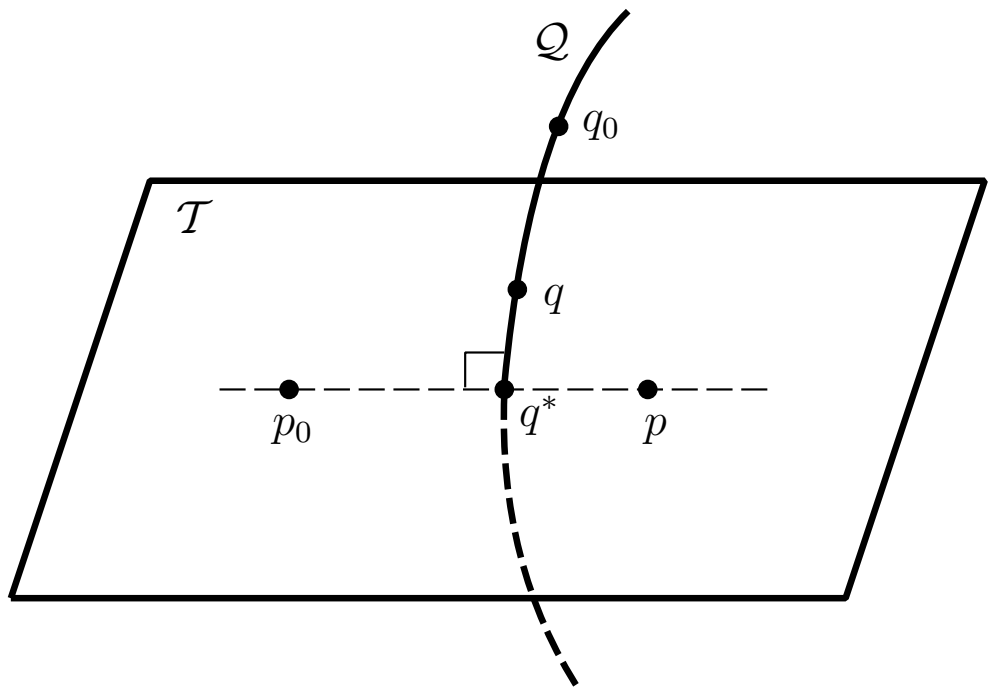
- 1 点 $p_0 \in \mathcal{M}$ を通り $\mathcal{Q}_U$ に垂直な $\mathcal{M}$ の部分空間

$$\mathcal{T}(p_0, \mu, \mathcal{F}) = \left\{ p \in \mathcal{M} \,\middle|\, \langle p - p_0, f_\lambda \rangle_\mu = 0,\ \forall \lambda \in \Lambda \right\}$$

平坦な部分空間の関係

- 交点の特徴付け

$$q^* = \underset{p \in \mathcal{T}}{\operatorname{argmin}}\, D_U(p, q_0) = \underset{q \in \mathcal{Q}_U}{\operatorname{argmin}}\, D_U(p_0, q),$$



双対な最適化問題

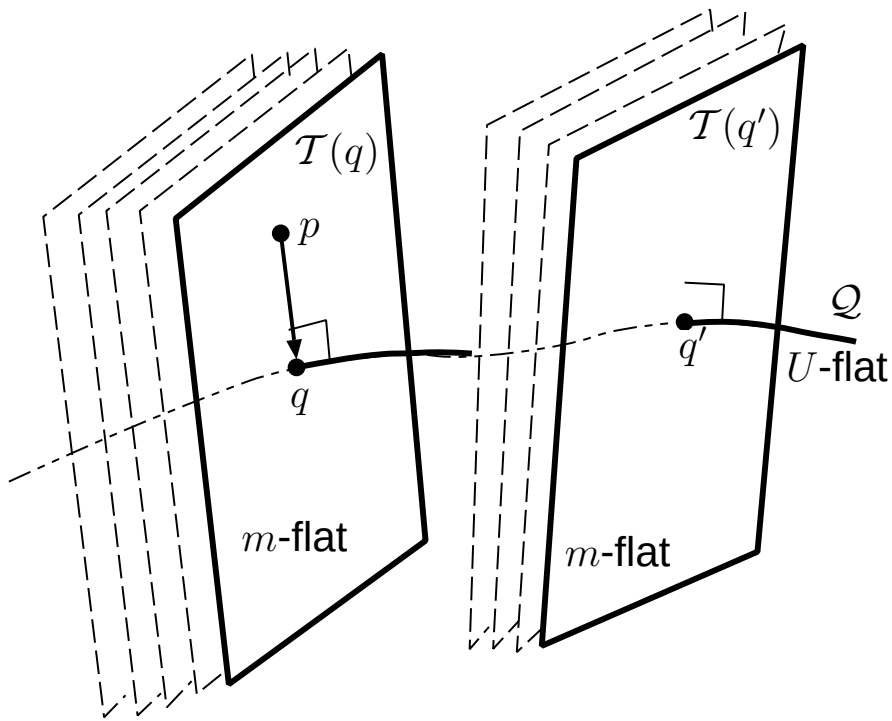
- minimize $D_U(p, q_0)$
  with respect to $p \in \mathcal{T}(p_0)$ for fixed q_0,
  (m-平坦部分空間内での最適化)
- minimize $D_U(p_0, q)$
  with respect to $q \in \mathcal{Q}_U(q_0)$ for fixed p_0,
  (U-平坦部分空間内での最適化)
- 二つの最適化問題の解

$$q^* = \operatorname*{argmin}_{p \in \mathcal{T}} D_U(p, q_0) = \operatorname*{argmin}_{q \in \mathcal{Q}_U} D_U(p_0, q),$$

直交葉層化

- $\mathcal{M}$ は $\mathcal{Q}$ と直交する $\mathcal{T}$ の集合に分割される。

$$\bigcup_{q \in Q} \mathcal{T}(q) = \mathcal{M} \text{ where } q \in \mathcal{T}(q)$$



判別モデルの不変性

- **シフト不変性:** $b(\boldsymbol{x})$ は任意の $\boldsymbol{x}$ の関数

$$\arg\max_{y\in\mathcal{Y}} q(y|\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} \left\{\xi\big(q(y|\boldsymbol{x})\big) - b(\boldsymbol{x})\right\}$$

- **スケール不変性:** $c(\boldsymbol{x})$ は任意の $\boldsymbol{x}$ の正値関数

$$\arg\max_{y\in\mathcal{Y}} q(y|\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} c(\boldsymbol{x})q(y|\boldsymbol{x})$$

- アルゴリズム構成に適した $b(\boldsymbol{x})$ や $c(\boldsymbol{x})$ を選んでよい

判別モデルの不変性

- **シフト不変性:** $b(\boldsymbol{x})$ は任意の $\boldsymbol{x}$ の関数

$$\arg\max_{y\in\mathcal{Y}} q(y|\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} \left\{\xi\big(q(y|\boldsymbol{x})\big) - b(\boldsymbol{x})\right\}$$

- **スケール不変性:** $c(\boldsymbol{x})$ は任意の $\boldsymbol{x}$ の正値関数

$$\arg\max_{y\in\mathcal{Y}} q(y|\boldsymbol{x}) = \arg\max_{y\in\mathcal{Y}} c(\boldsymbol{x})q(y|\boldsymbol{x})$$

- アルゴリズム構成に適した $b(\boldsymbol{x})$ や $c(\boldsymbol{x})$ を選んでよい

正規化 U モデル

- 確率分布となるように $b = \phi$ を選んだモデル

$$\mathcal{Q}_U^{\mathsf{norm}}(q_0, \mathcal{F}) = \Bigg\{ q \in \mathcal{P} \;\Bigg|\; \xi(q(y|\boldsymbol{x})) = \xi(q_0(y|\boldsymbol{x})) + \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y) - \phi(\boldsymbol{x}, \boldsymbol{\alpha}) \Bigg\}$$

- $F = \xi(q) + \phi$ の関数として書き直した $H_U(\tilde{p}, q; \tilde{\mu})$:

$$L_U^{\mathsf{norm}}(F) = \frac{1}{n} \sum_{i=1}^{n} \Big[\sum_{y \in \mathcal{Y}} U\big(F(\boldsymbol{x}_i, y) - \phi(\boldsymbol{x}_i, \boldsymbol{\alpha})\big) - \big\{F(\boldsymbol{x}_i, y_i) - \phi(\boldsymbol{x}_i, \boldsymbol{\alpha})\big\} \Big]$$

正規化 U モデル

- 確率分布となるように $b = \phi$ を選んだモデル

$$\mathcal{Q}_U^{\mathsf{norm}}(q_0, \mathcal{F}) = \Bigg\{ q \in \mathcal{P} \;\Bigg|\; \xi(q(y|\boldsymbol{x})) = \xi(q_0(y|\boldsymbol{x})) + \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y) - \phi(\boldsymbol{x}, \boldsymbol{\alpha}) \Bigg\}$$

- $F = \xi(q) + \phi$ の関数として書き直した $H_U(\tilde{p}, q; \tilde{\mu})$:

$$L_U^{\mathsf{norm}}(F) = \frac{1}{n} \sum_{i=1}^{n} \Bigg[\sum_{y \in \mathcal{Y}} U\big(F(\boldsymbol{x}_i, y) - \phi(\boldsymbol{x}_i, \boldsymbol{\alpha})\big) - \big\{F(\boldsymbol{x}_i, y_i) - \phi(\boldsymbol{x}_i, \boldsymbol{\alpha})\big\} \Bigg]$$

経験 U モデル

- cross-entropy の第 2 項が消えるように b を選んだモデル

$$\mathcal{Q}_U^{\mathsf{emp}}(q_0, \mathcal{F}) = \Bigg\{ q \in \mathcal{M} \;\Bigg|\; \xi(q(y|\boldsymbol{x})) = \xi(q_0(y|\boldsymbol{x})) + \sum_{t=1}^{T} \alpha_t \big(f_t(\boldsymbol{x}, y) - \tilde{f}_t(\boldsymbol{x}) \big) \Bigg\}.$$

($\tilde{f}$ は f の条件付経験分布による平均)

- $F = \xi(q)$ の関数として書き直した $H_U(\tilde{p}, q; \tilde{\mu})$:

$$L_U^{\mathsf{emp}}(F) = \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} U\big(F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i) \big)$$

経験 U モデル

- cross-entropy の第 2 項が消えるように b を選んだモデル

$$\mathcal{Q}_U^{\mathsf{emp}}(q_0, \mathcal{F}) = \Bigg\{ q \in \mathcal{M} \;\Bigg|\; \xi(q(y|\boldsymbol{x})) = \xi(q_0(y|\boldsymbol{x})) + \sum_{t=1}^{T} \alpha_t \big(f_t(\boldsymbol{x}, y) - \tilde{f}_t(\boldsymbol{x}) \big) \Bigg\}.$$

  ($\tilde{f}$ は f の条件付経験分布による平均)

- $F = \xi(q)$ の関数として書き直した $H_U(\tilde{p}, q; \tilde{\mu})$:

$$L_U^{\mathsf{emp}}(F) = \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} U\big(F(\boldsymbol{x}_i, y) - F(\boldsymbol{x}_i, y_i) \big)$$

Outline

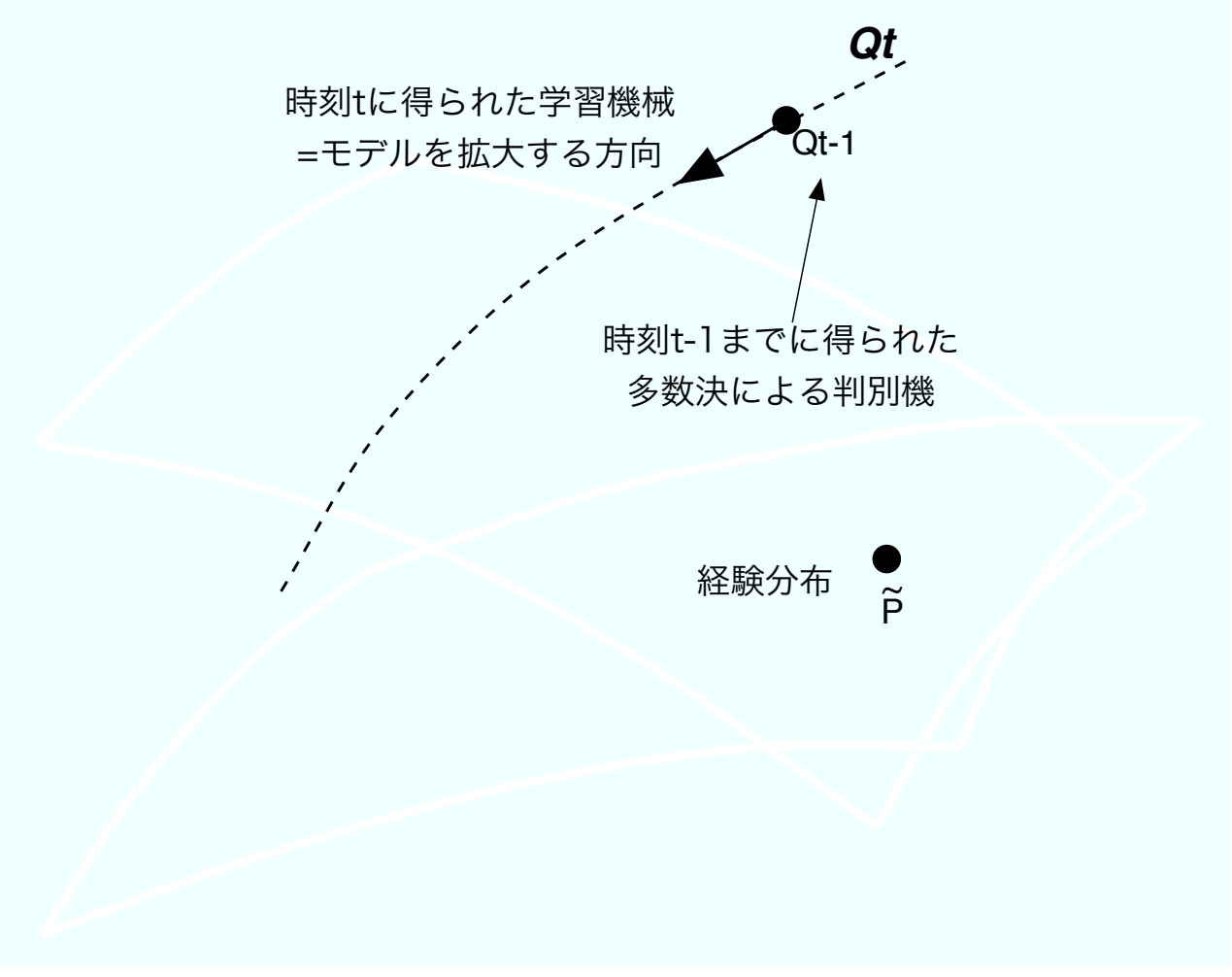
3 **Boosting Algorithm の幾何学的描像**
- Algorithm の幾何学的表現
- 統計的性質

U-Boost Algorithm (1)

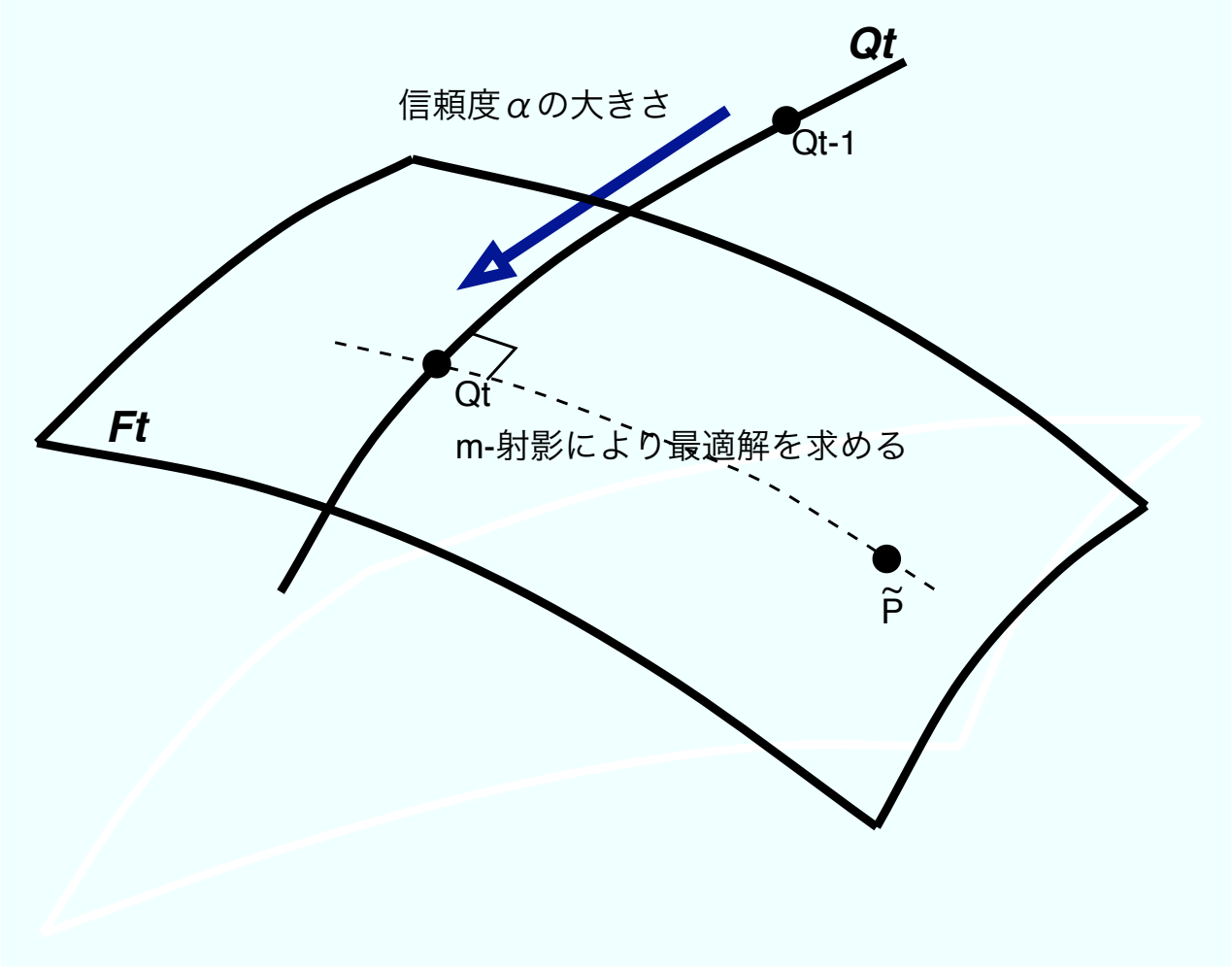
- **入力:**
  n 個の例題 $\{(\boldsymbol{x}_i, y_i); \boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \ldots, n\}$,
  単調増加な凸関数 U.
- **初期化:**
  $q_0(y|\boldsymbol{x})$ (通常，簡単のため $\xi(q_0) = 0$ とする).
- **繰り返し:** 以下の step を繰り返す $(t = 1, \ldots, T)$.

繰り返しの中の幾何学的解釈

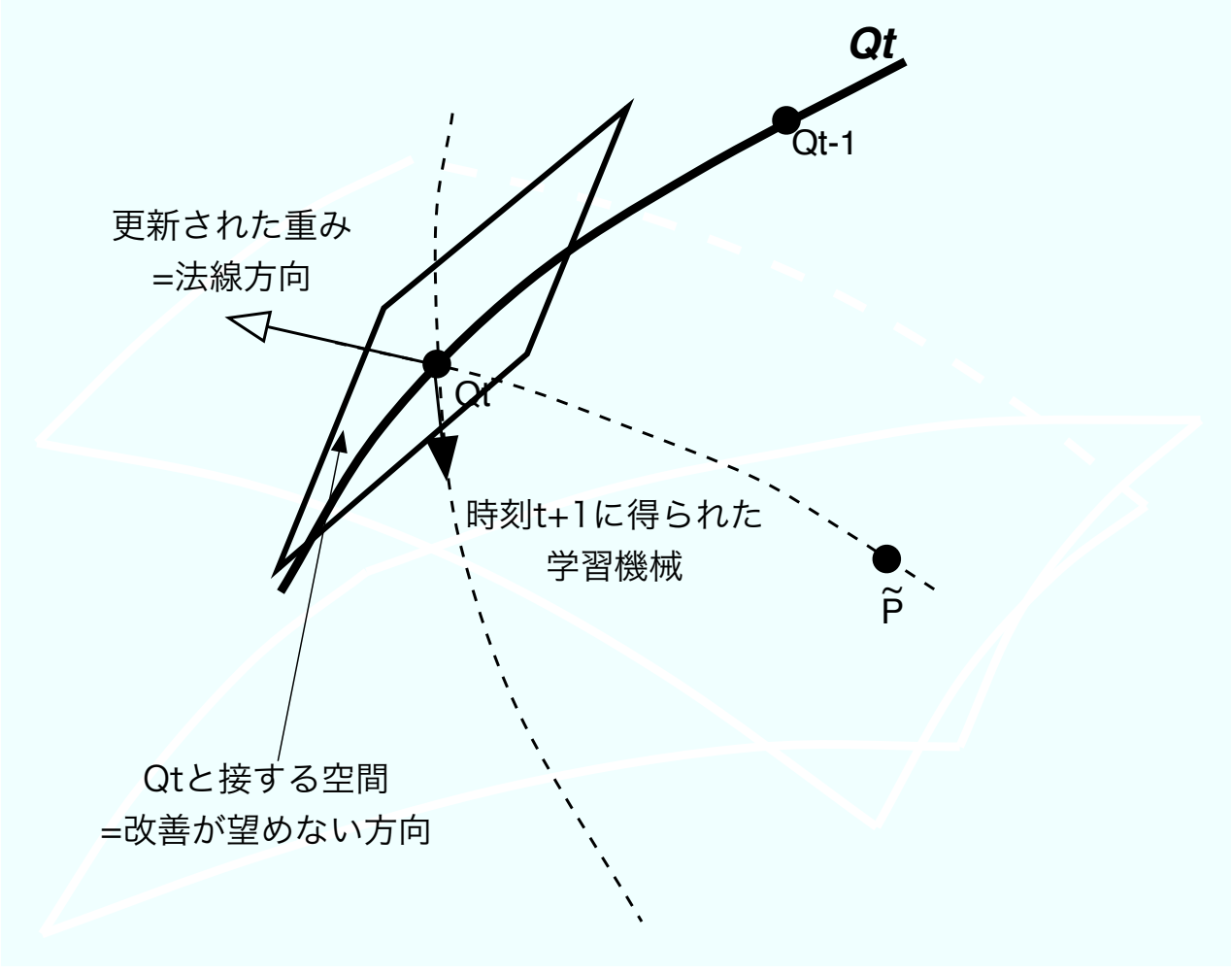
繰り返しの中の幾何学的解釈



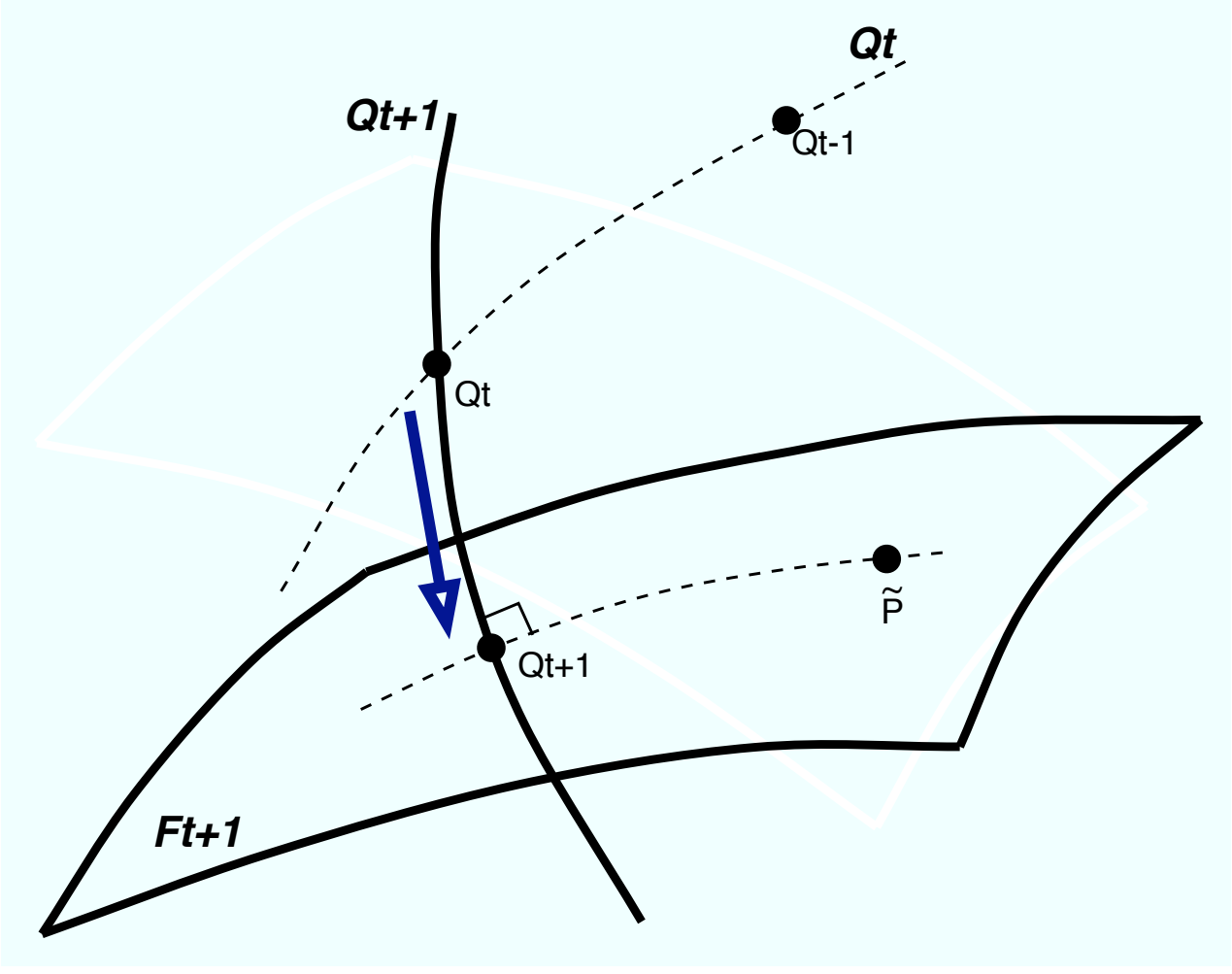
繰り返しの中の幾何学的解釈



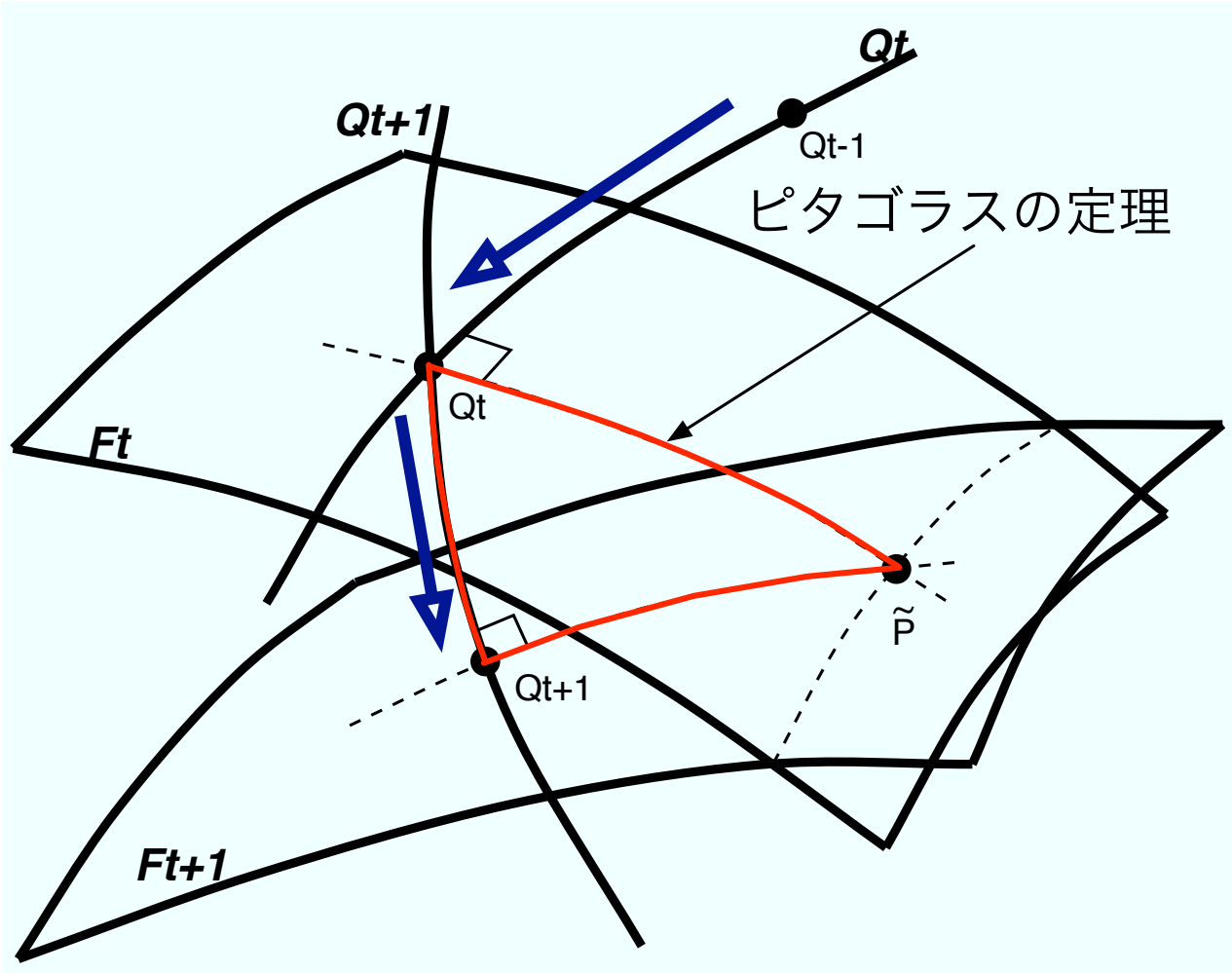
繰り返しの中の幾何学的解釈



繰り返しの中の幾何学的解釈



繰り返しの中の幾何学的解釈



U-Boost Algorithm (2)

(繰り返しの中)

- **step 1:** $f - b'$ ができるだけ $q_{t-1} - \tilde{p}$ と同じ方向を向くように決定関数 f_t(判別器 h_t) を選ぶ.

$$f_t(\boldsymbol{x}, y) = \arg\max_{f \in \mathcal{F}} \langle q_{t-1} - \tilde{p}, f - b' \rangle_{\tilde{\mu}}$$

$$= \arg\min_{f \in \mathcal{F}} \langle q_{t-1}, f - \bar{f} \rangle_{\tilde{\mu}}$$

$$D_t(i, y) \propto q_{t-1}(y|\boldsymbol{x}_i)$$

U-Boost Algorithm (3)

(繰り返しの中)

- **step 2:** 1 次元のモデル

$$\mathcal{Q}_t = \left\{ q \;\middle|\; \xi(q) = \xi(q_{t-1}) + \alpha f_t - b_t(\alpha),\ \alpha \in R \right\}$$

と直交葉層化 $\{\mathcal{T}(q); q \in \mathcal{Q}_t\}$ を構成し，経験分布 $\tilde{p}$ を含む葉と $\mathcal{Q}_t$ の交点から α_t を求める.

$$\alpha_t = \arg\min_{q \in \mathcal{Q}_t} \sum_{i=1}^{n} \left[\sum_{y \in \mathcal{Y}} U\left(\xi(q(y|\boldsymbol{x}_i))\right) - \xi(q(y_i|\boldsymbol{x}_i)) \right].$$

U-Boost Algorithm (4)

(繰り返しの中)

- **step 3:** q_t を更新する.

$$q_t(y|\boldsymbol{x}) = u\Big(\xi(q_{t-1}(y|\boldsymbol{x})) + \alpha_t f_t(\boldsymbol{x}, y) - b_t(\boldsymbol{x}, \alpha_t)\Big).$$

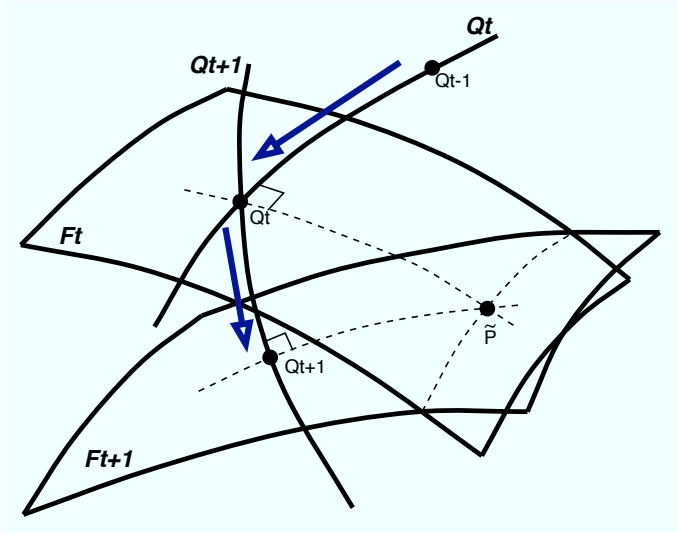
U-Boost Algorithm (5)

- **出力:**
  多数決判別器を構成する.

$$H(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} F_T(\boldsymbol{x}, y) = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y).$$

ブースティングの仕組み

- 大域的なモデルの拡大
  - 逐次的に例題の重みを変化させながら学習を行い，改善が望める方向にモデルを逐次的に拡大している．
  - 確率分布の外側に探索範囲を広げることによって，効率的なアルゴリズム (coordinate descent) を構成している．



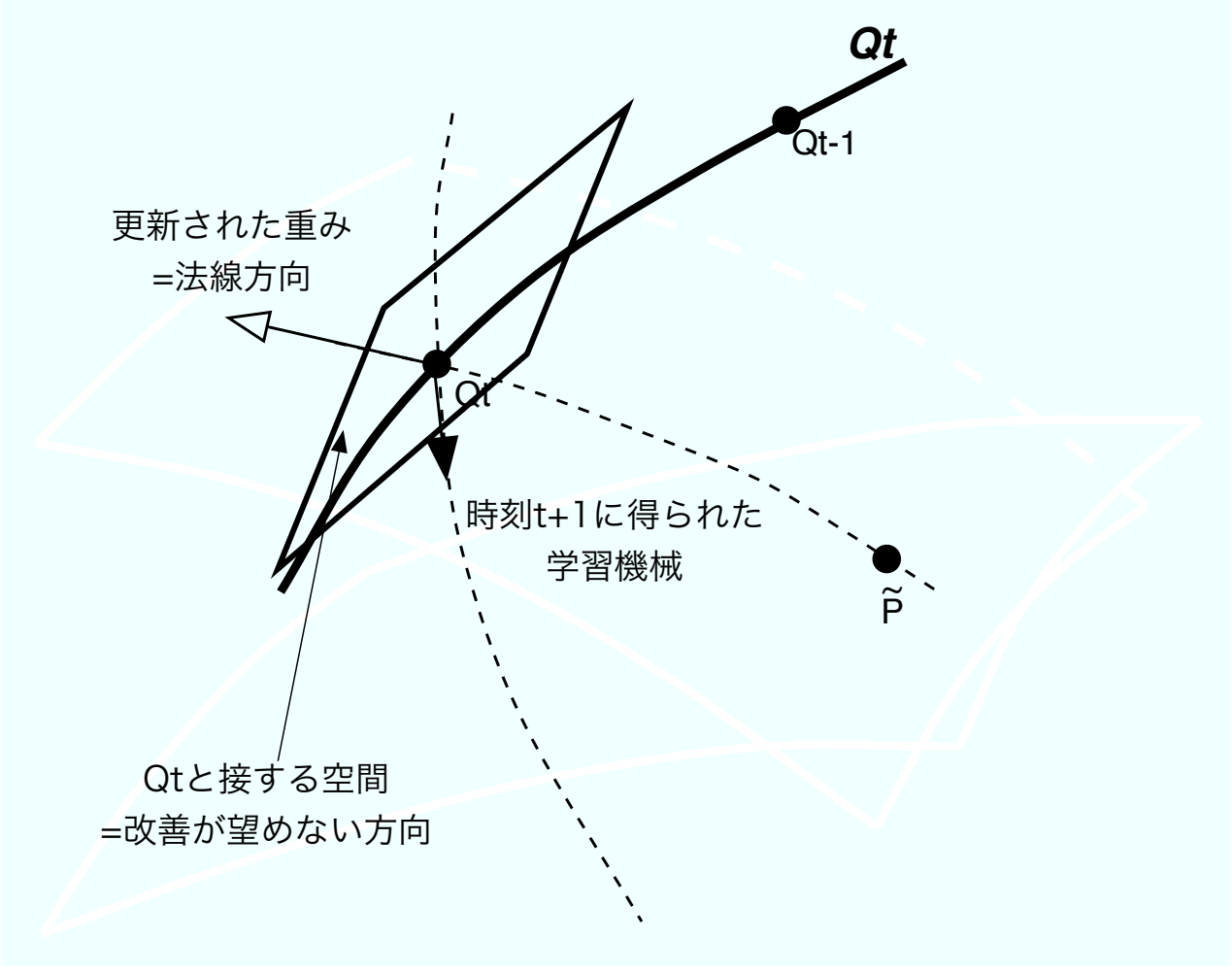
誤差の性質

- 選択された決定関数と次時刻の誤差との関係

$$\epsilon_{t+1}(f_t) = \frac{1}{2} \quad (\forall t = 1, 2, \ldots, T-1)$$

  - 時刻 t において選ばれた決定関数 f_t は，更新された分布 D_{t+1} のもとではランダムな判別と同等.
  - 1 つ前に選ばれた判別器が最も不得意となるように分布の更新が進む.

Random Guessの空間



Bayes 最適性 (2 値判別の場合)

- 真の分布のもとで最小値を与える F は Bayes 最適な判別器

$$\{\boldsymbol{x}|F(\boldsymbol{x})>0\}=\left\{\boldsymbol{x}\middle|\log\frac{p(+1|\boldsymbol{x})}{p(-1|\boldsymbol{x})}>0\right\}$$

- 以下の条件を満たす U では F は尤度比と一致 (Fisher consistent).

$$\log\frac{u(z)}{u(-z)}=2z\quad\Rightarrow\quad F(\boldsymbol{x})=\frac{1}{2}\log\frac{p(+1|\boldsymbol{x})}{p(-1|\boldsymbol{x})}$$

- U-Boost が一致性をもつのは
  - U-モデルが真の分布を含むとき
  - 判別器の線形結合が十分な表現力をもつとき

有効性

- 信頼度の漸近分散 $\hat{\boldsymbol{\alpha}}$:

$$\mathsf{Cov}(\hat{\boldsymbol{\alpha}}) = \frac{1}{n} H^{-1} G H^{-1} + o\left(\frac{1}{n}\right),$$

$$H = \int_{\mathcal{X}} \frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^{\tau}} r(\boldsymbol{x}, \boldsymbol{\alpha}^*) d\mu(\boldsymbol{x}),$$

$$G = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} p(y|\boldsymbol{x}) \Big(\frac{\partial}{\partial \boldsymbol{\alpha}} r(\boldsymbol{x}, \boldsymbol{\alpha}^*) - \boldsymbol{f}(\boldsymbol{x}, y) \Big)^2 d\mu(\boldsymbol{x})$$

$$r(\boldsymbol{x}, \boldsymbol{\alpha}) =$$

$$\sum_{y \in \mathcal{Y}} U\left(\xi(q_0) + \sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{x}, y) - b(\boldsymbol{x}, \boldsymbol{\alpha})\right) + b(\boldsymbol{x}, \boldsymbol{\alpha}).$$

- U モデルが一致性をもつ場合, LogitBoost が最小

頑健性

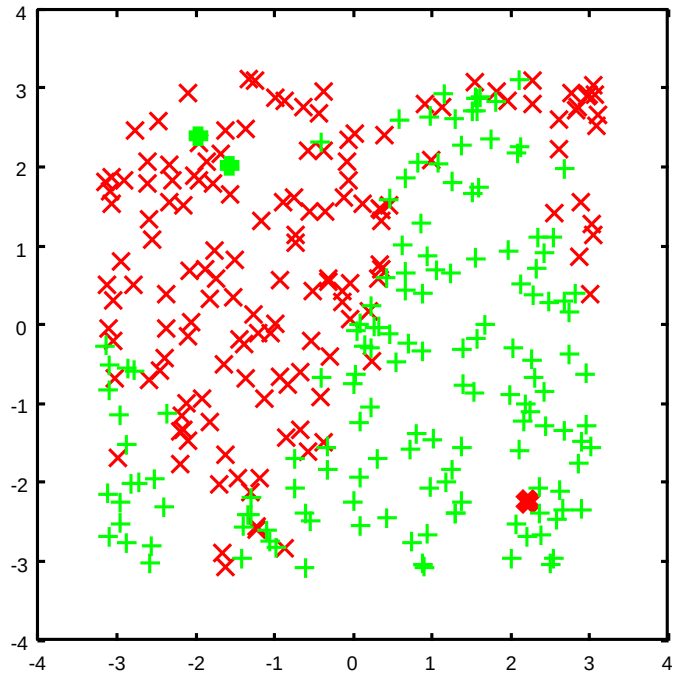
- 頑健性の指標: gross error sensitivity (Hampel et al., 1986)

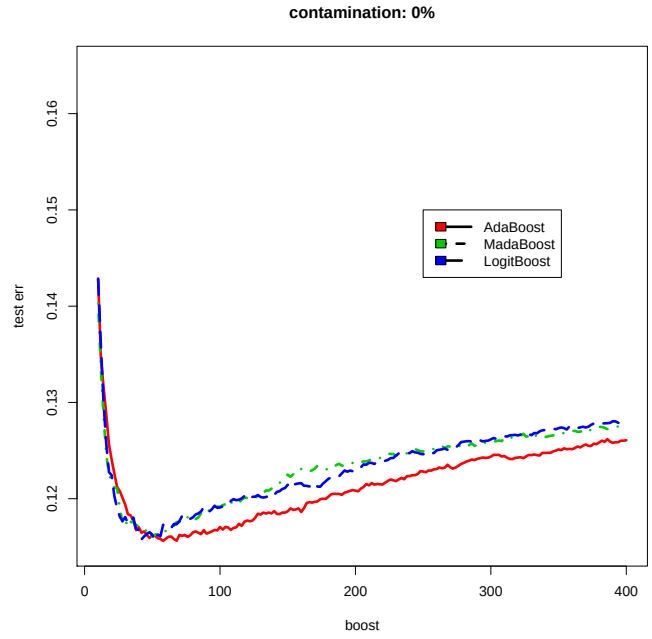
$$\gamma(U, p_0) = \sup_{(\tilde{\boldsymbol{x}}, \tilde{y})} \left\{ \lim_{\epsilon \to +0} \frac{1}{\epsilon} \Big[\alpha_U((1-\epsilon)\, p_0 + \epsilon\, \delta(\tilde{\boldsymbol{x}}, \tilde{y})) - \alpha_U(p_0) \Big] \right\}^2$$

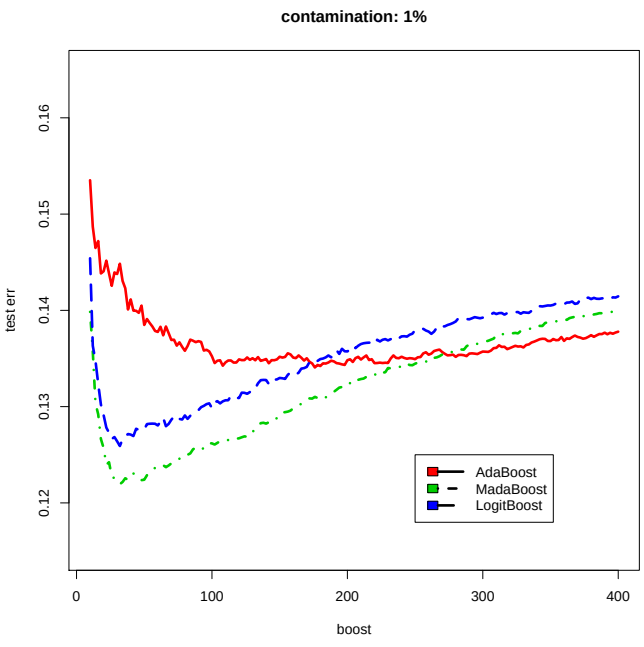
- Fisher 一致性をもつ関数 U の中で MadaBoost

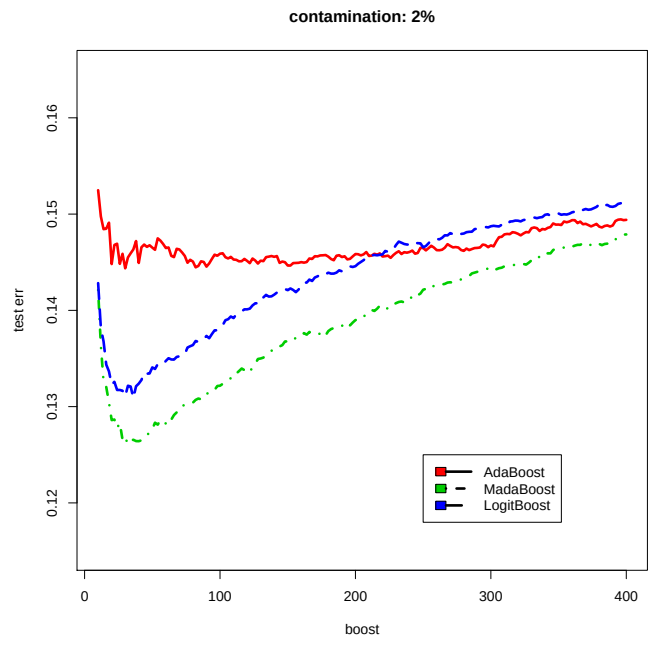
$$u(z) = \begin{cases} 1, & z \geq 0, \\ \exp(2z), & z < 0, \end{cases}$$

は最小の gross error sensitivity を与える.




contamination: 0%
test err
boost
AdaBoost
MadaBoost
LogitBoost


contamination: 1%
test err
boost
AdaBoost
MadaBoost
LogitBoost


contamination: 2%
test err
boost
AdaBoost
MadaBoost
LogitBoost

Part II

EM Algorithm

Outline

4. EM アルゴリズムの幾何学
5. EM アルゴリズムの拡張

Outline

4 **EM アルゴリズムの幾何学**
- EM アルゴリズム
- EM アルゴリズムの幾何学的描像
- em アルゴリズム

EM アルゴリズムの目的

- 確率変数の一部が観測できない状況
  - 観測できる確率変数 (顕在変数): X
  - 観測できない確率変数 (潜在変数，隠れ変数): Y
- 観測データ $\{x_1, x_2, \ldots, x_N\}$ から確率モデル $p(x, y; \theta)$ のパラメタ θ を決定
  - 欠損データのある問題
  - 単純なモデルを複合したモデルの推定
    (隠れマルコフモデル，混合モデルなど)
- 関連するアルゴリズム
  - Baum-Welch アルゴリズム
  - back-propagation アルゴリズムなどもその一種.

EM アルゴリズムの目的

- 確率変数の一部が観測できない状況
  - 観測できる確率変数 (顕在変数): X
  - 観測できない確率変数 (潜在変数，隠れ変数): Y
- 観測データ $\{x_1, x_2, \ldots, x_N\}$ から確率モデル $p(x, y; \theta)$ のパラメタ θ を決定
  - 欠損データのある問題
  - 単純なモデルを複合したモデルの推定
    (隠れマルコフモデル，混合モデルなど)
- 関連するアルゴリズム
  - Baum-Welch アルゴリズム
  - back-propagation アルゴリズムなどもその一種.

EM アルゴリズムの目的

- 確率変数の一部が観測できない状況
  - 観測できる確率変数 (顕在変数): X
  - 観測できない確率変数 (潜在変数，隠れ変数): Y
- 観測データ $\{x_1, x_2, \ldots, x_N\}$ から確率モデル $p(x, y; \theta)$ のパラメタ θ を決定
  - 欠損データのある問題
  - 単純なモデルを複合したモデルの推定
    (隠れマルコフモデル，混合モデルなど)
- 関連するアルゴリズム
  - Baum-Welch アルゴリズム
  - back-propagation アルゴリズムなどもその一種.

EM アルゴリズムの目的

- 確率変数の一部が観測できない状況
  - 観測できる確率変数 (顕在変数): X
  - 観測できない確率変数 (潜在変数，隠れ変数): Y
- 観測データ $\{x_1, x_2, \ldots, x_N\}$ から確率モデル $p(x, y; \theta)$ のパラメタ θ を決定
  - 欠損データのある問題
  - 単純なモデルを複合したモデルの推定
    (隠れマルコフモデル，混合モデルなど)
- 関連するアルゴリズム
  - Baum-Welch アルゴリズム
  - back-propagation アルゴリズムなどもその一種.

EM アルゴリズムの目的

- 確率変数の一部が観測できない状況
  - 観測できる確率変数 (顕在変数): X
  - 観測できない確率変数 (潜在変数，隠れ変数): Y
- 観測データ $\{x_1, x_2, \ldots, x_N\}$ から確率モデル $p(x, y; \theta)$ のパラメタ θ を決定
  - 欠損データのある問題
  - 単純なモデルを複合したモデルの推定
    (隠れマルコフモデル，混合モデルなど)
- 関連するアルゴリズム
  - Baum-Welch アルゴリズム
  - back-propagation アルゴリズムなどもその一種.

EM アルゴリズムの目的

- 確率変数の一部が観測できない状況
  - 観測できる確率変数 (顕在変数): X
  - 観測できない確率変数 (潜在変数，隠れ変数): Y
- 観測データ $\{x_1, x_2, \ldots, x_N\}$ から確率モデル $p(x, y; \theta)$ のパラメタ θ を決定
  - 欠損データのある問題
  - 単純なモデルを複合したモデルの推定
    (隠れマルコフモデル，混合モデルなど)
- 関連するアルゴリズム
  - Baum-Welch アルゴリズム
  - back-propagation アルゴリズムなどもその一種.

EM アルゴリズム

algorithm

- **入力:** 観測データ $x_1, \ldots, x_N$
- **反復:** 収束するまで以下の E-step と M-step を繰り返す.
  - **E-step:** $Q(\theta, \theta^t)$ を求める.

$$Q(\theta, \theta^t) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \int p(y|x_i; \theta^t) \log p(x_i, y; \theta) dy \right\}$$

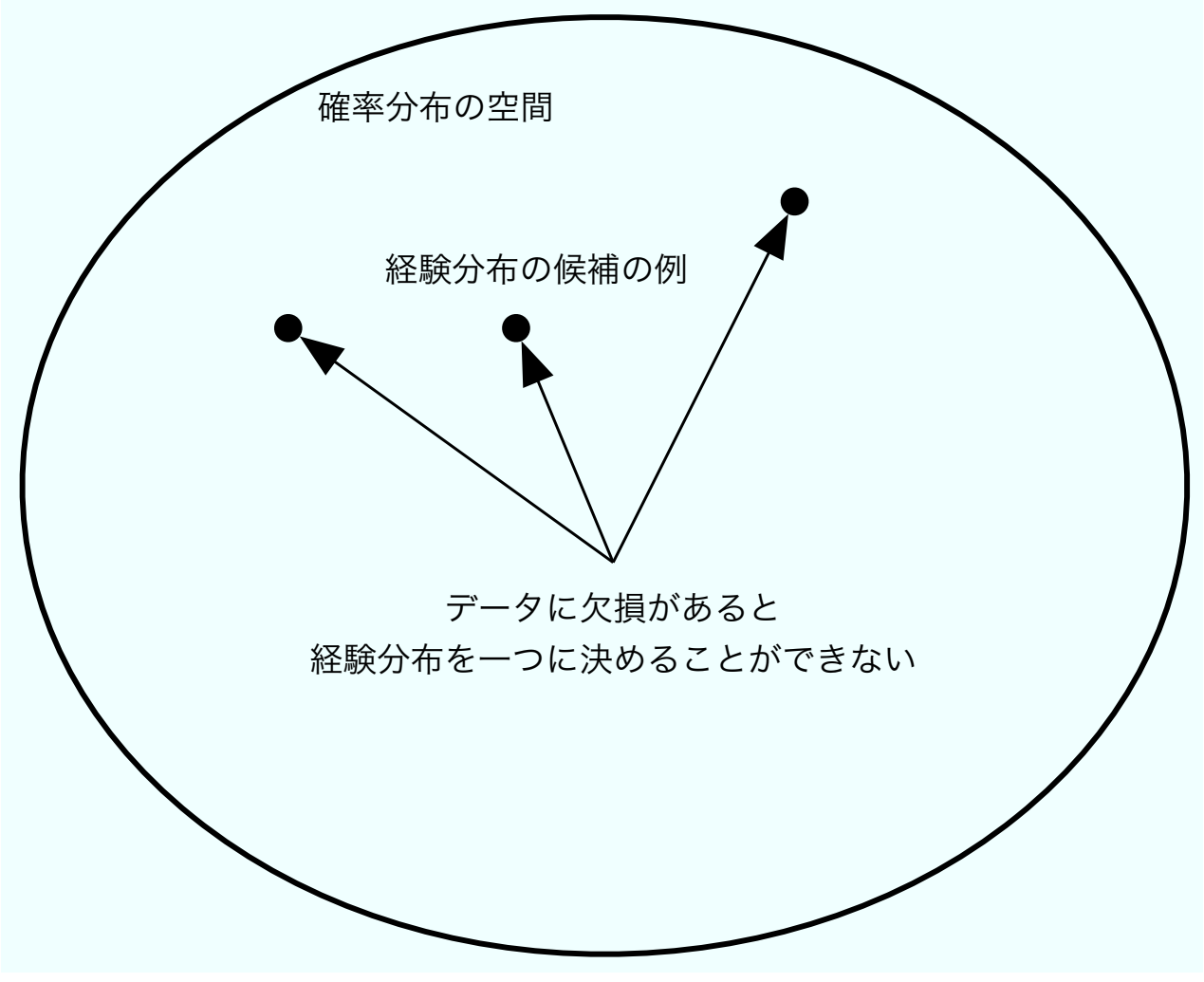
  - **M-step:** $Q(\theta, \theta^t)$ を最大にする θ^{t+1} を求める.

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^t)$$

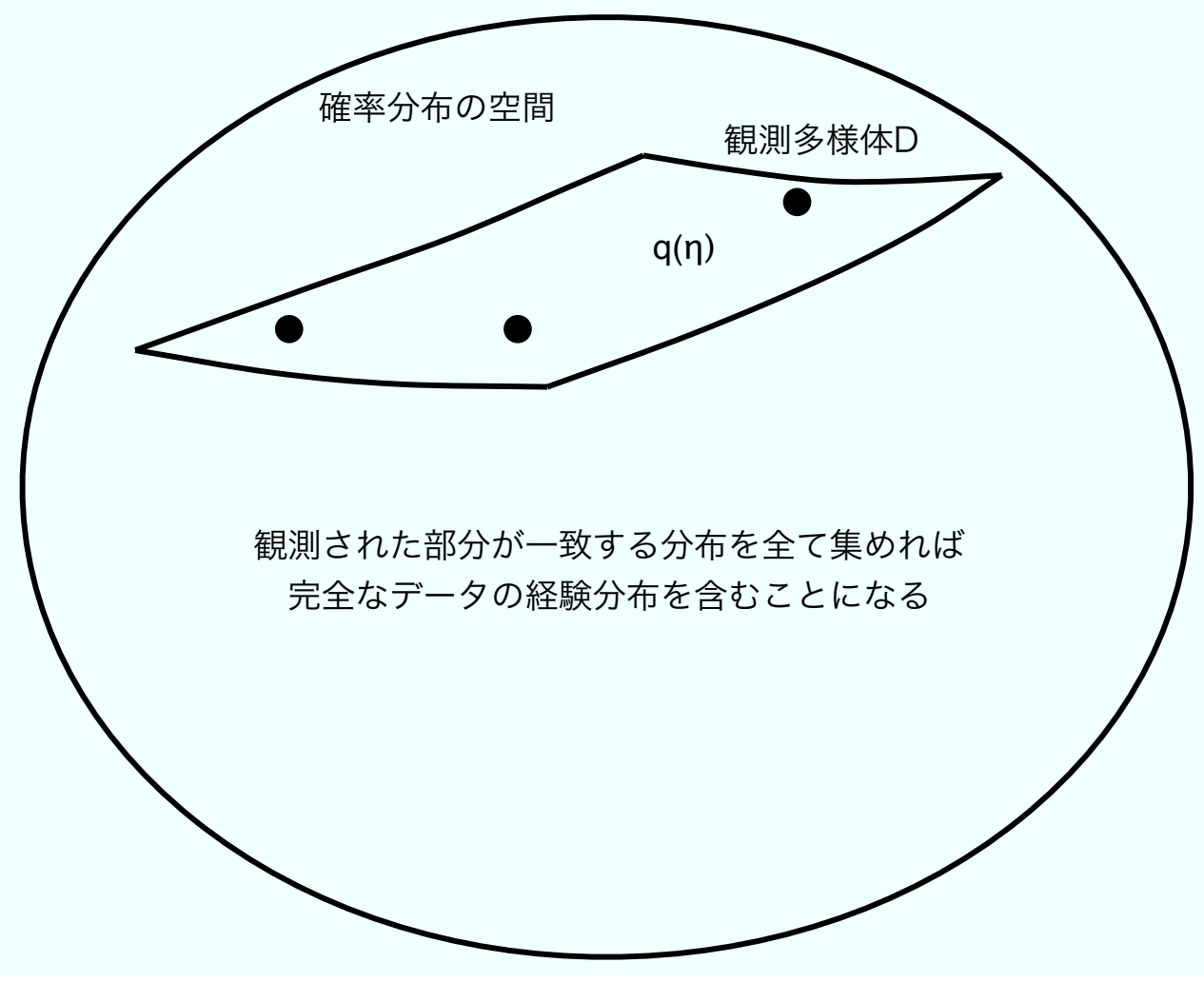
- **出力:** 収束したパラメタ θ.

EM アルゴリズムの幾何学的描像

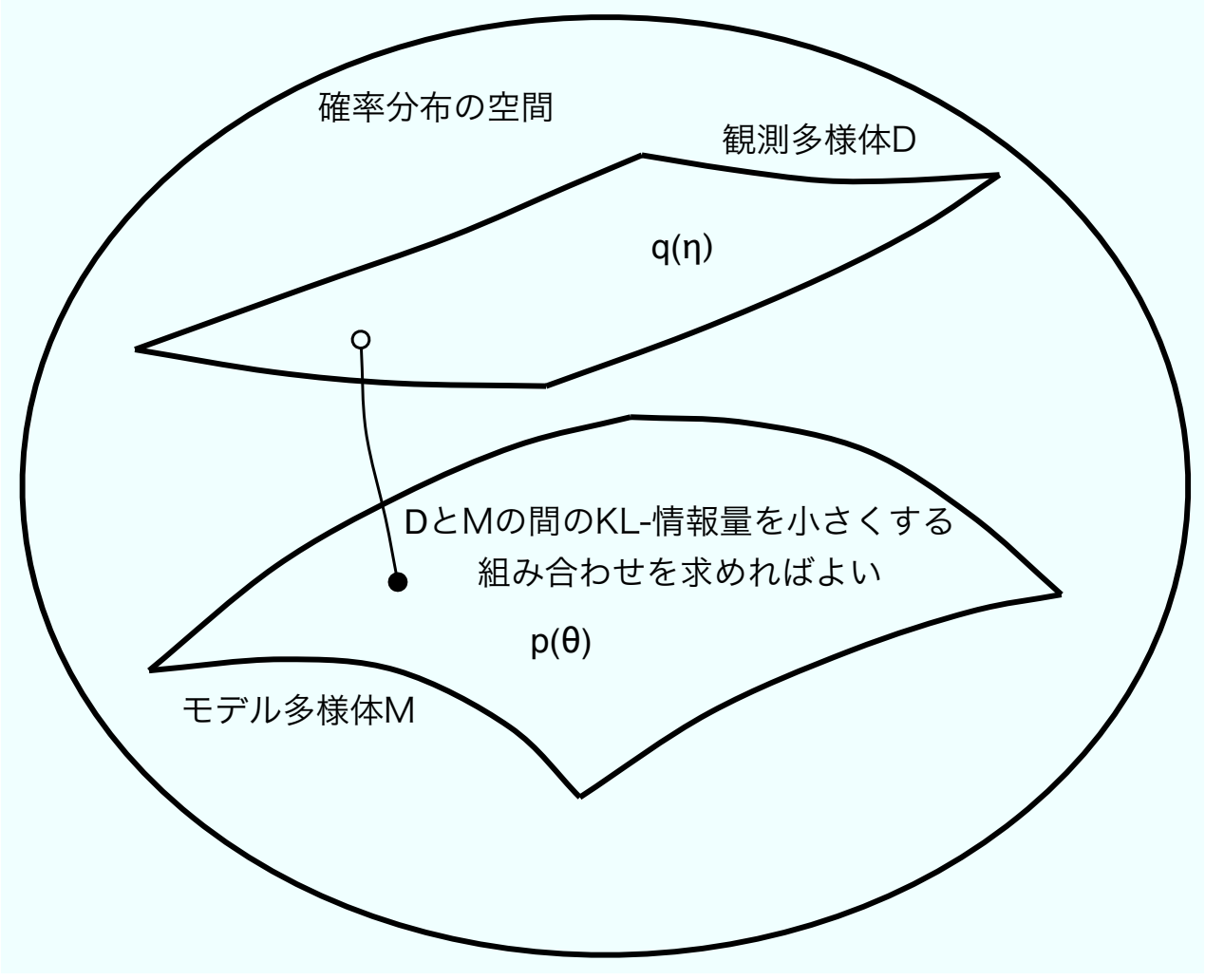
EM アルゴリズムの幾何学的描像



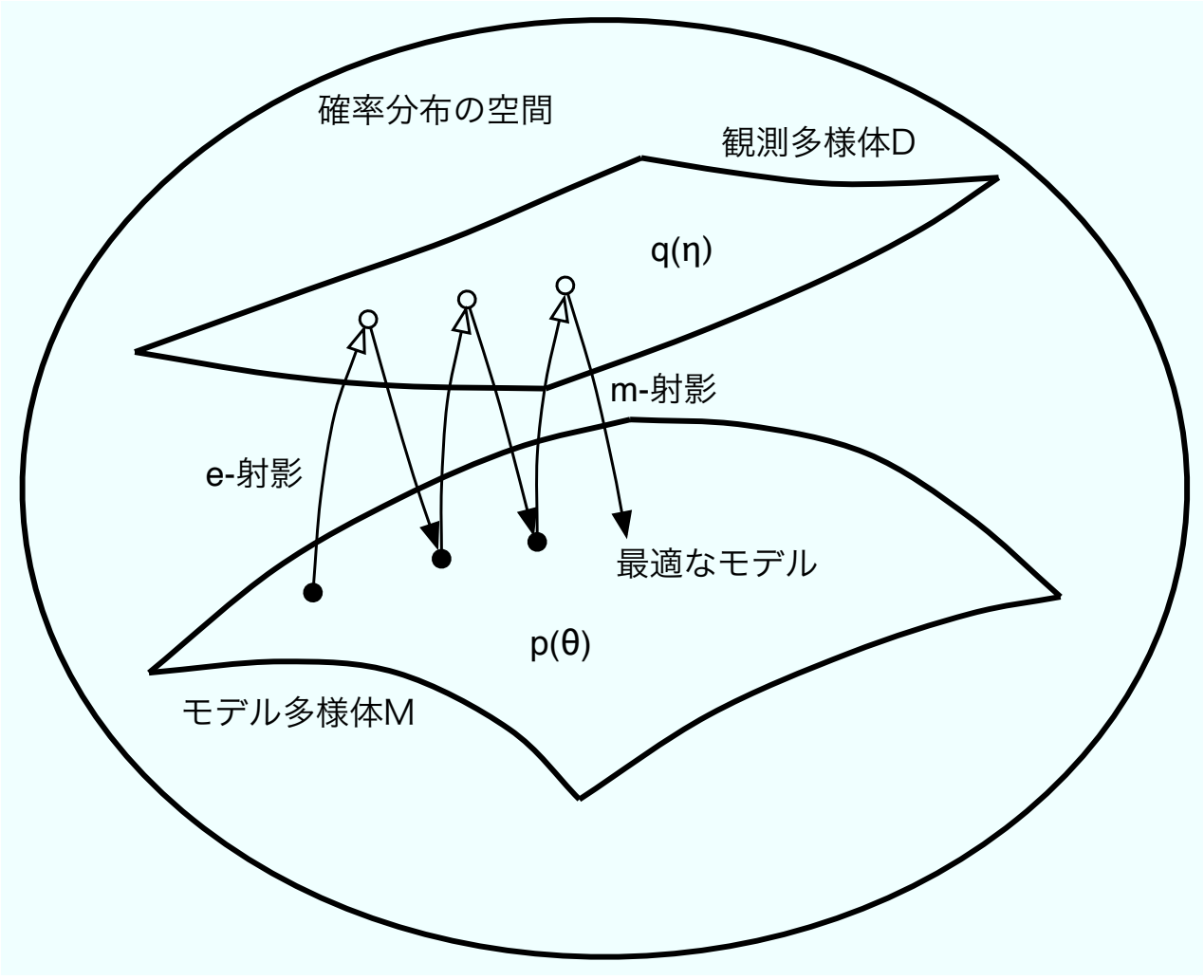
EM アルゴリズムの幾何学的描像



EM アルゴリズムの幾何学的描像



EM アルゴリズムの幾何学的描像



モデル多様体と観測多様体

- モデル多様体: 完全データ (X, Y) に対する確率モデル

$$\mathcal{M} = \{p(x, y; \theta), \theta \in \Theta\}$$

- 観測多様体

$$\mathcal{D} = \{\tilde{q}(x)q(y|x; \phi), \phi \in \Phi\}$$

$\tilde{q}(x)$: 観測データの経験分布

$q(y|x; \phi)$: ϕ で記述された条件付き分布

- 普通の EM では

$$q(y|x; \phi) = \frac{p(x, y; \phi)}{p(x; \phi)}.$$

モデル多様体と観測多様体

- モデル多様体: 完全データ (X, Y) に対する確率モデル

$$\mathcal{M} = \{p(x, y; \theta), \theta \in \Theta\}$$

- 観測多様体

$$\mathcal{D} = \{\tilde{q}(x)q(y|x; \phi), \phi \in \Phi\}$$

$\tilde{q}(x)$: 観測データの経験分布

$q(y|x; \phi)$: ϕ で記述された条件付き分布

- 普通の EM では

$$q(y|x; \phi) = \frac{p(x, y; \phi)}{p(x; \phi)}.$$

モデル多様体と観測多様体

- モデル多様体: 完全データ (X, Y) に対する確率モデル

$$\mathcal{M} = \{p(x, y; \theta), \theta \in \Theta\}$$

- 観測多様体

$$\mathcal{D} = \{\tilde{q}(x)q(y|x; \phi), \phi \in \Phi\}$$

$\tilde{q}(x)$: 観測データの経験分布

$q(y|x; \phi)$: ϕ で記述された条件付き分布

- 普通の EM では

$$q(y|x; \phi) = \frac{p(x, y; \phi)}{p(x; \phi)}.$$

モデル多様体と観測多様体

- モデル多様体: 完全データ (X, Y) に対する確率モデル

$$\mathcal{M} = \{p(x, y; \theta), \theta \in \Theta\}$$

- 観測多様体

$$\mathcal{D} = \{\tilde{q}(x)q(y|x; \phi), \phi \in \Phi\}$$

$\tilde{q}(x)$: 観測データの経験分布

$q(y|x; \phi)$: ϕ で記述された条件付き分布

- 普通の EM では

$$q(y|x; \phi) = \frac{p(x, y; \phi)}{p(x; \phi)}.$$

em アルゴリズム

algorithm

- **入力:** 観測データ $x_1, \ldots, x_N$
- **反復:** 収束するまで以下の e-step と m-step を繰り返す.
  - **e-step:**

$$\phi^{t+1} = \operatorname*{argmin}_{\phi} D(q(\phi), p(\theta^t)).$$

  - **m-step:**

$$\begin{aligned}\theta^{t+1} &= \operatorname*{argmin}_{\theta} D(q(\phi^{t+1}), p(\theta)) \\ &= \operatorname*{argmin}_{\theta} H(q(\phi^{t+1}), p(\theta)).\end{aligned}$$

- **出力:** 収束したパラメタ θ.

Outline

5 **EM アルゴリズムの拡張**
- UM アルゴリズム
- ロバスト化における応用例

UM アルゴリズム

algorithm

- **入力:** 観測データ $x_1, \ldots, x_N$
- **反復:** 収束するまで以下の U-step と M-step を繰り返す.
  - **U-step:**

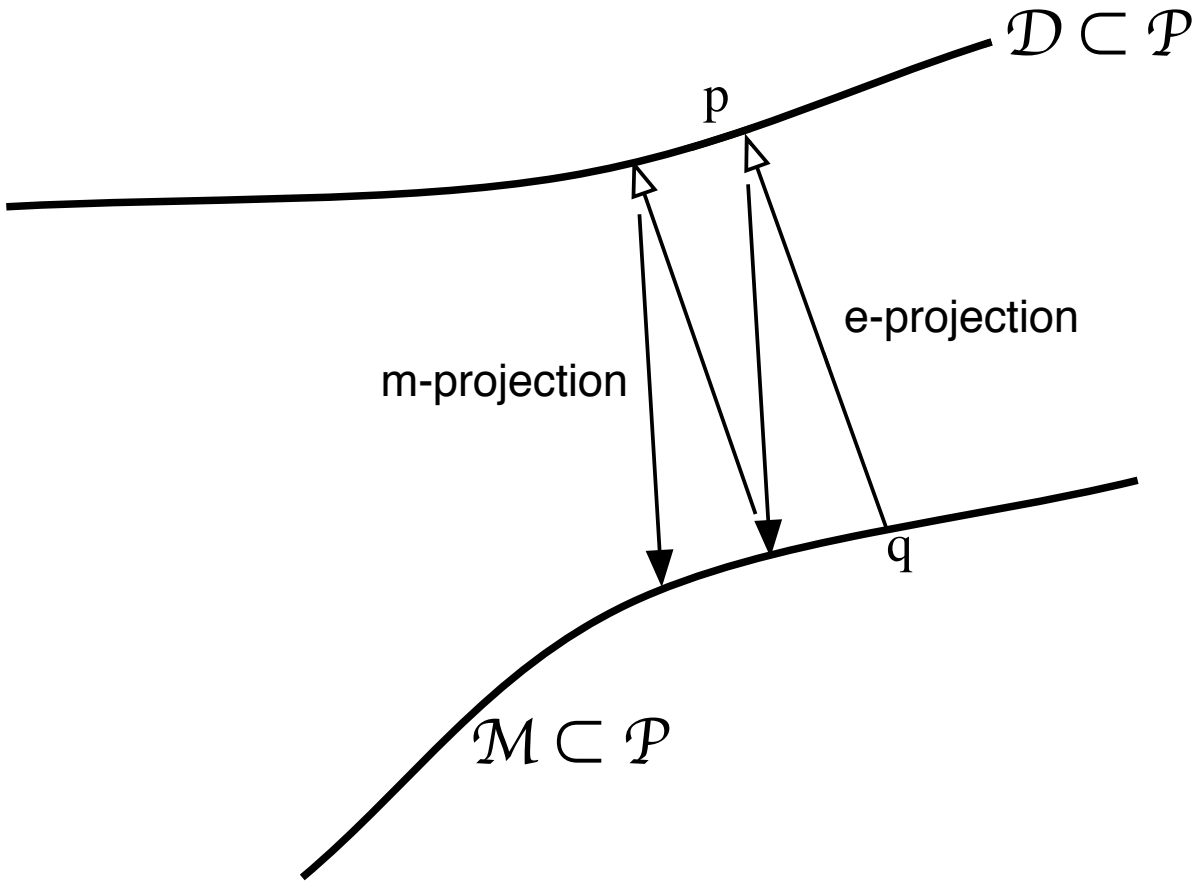
$$\phi^{t+1} = \operatorname*{argmin}_{\phi} D_U(p(\phi), q(\theta^t)).$$

  - **M-step:**

$$\begin{aligned}\theta^{t+1} &= \operatorname*{argmin}_{\theta} D_U(p(\phi^{t+1}), q(\theta)) \\ &= \operatorname*{argmin}_{\theta} H_U(p(\phi^{t+1}), q(\theta)).\end{aligned}$$

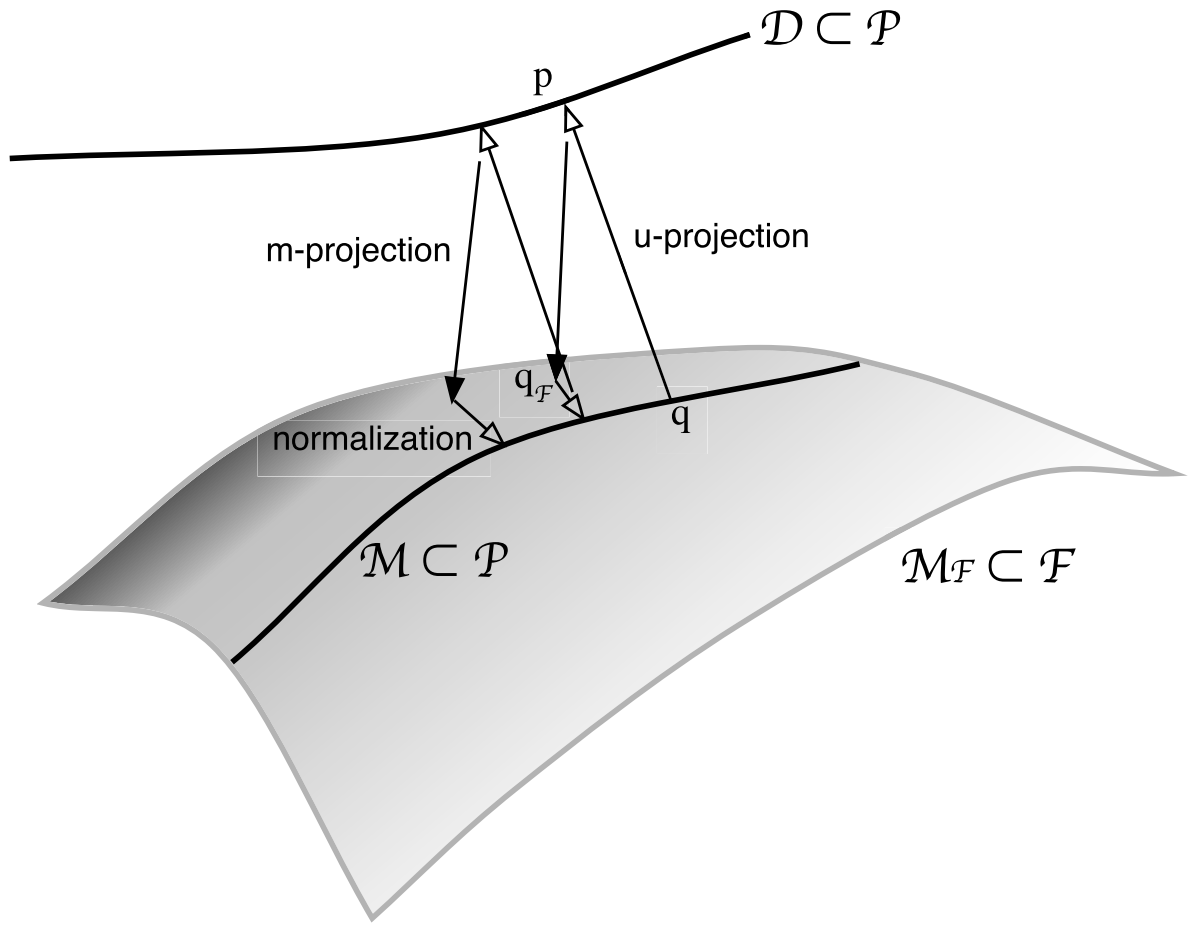
- **出力:** 収束したパラメタ θ.

UM アルゴリズムの幾何学的描像 (1)



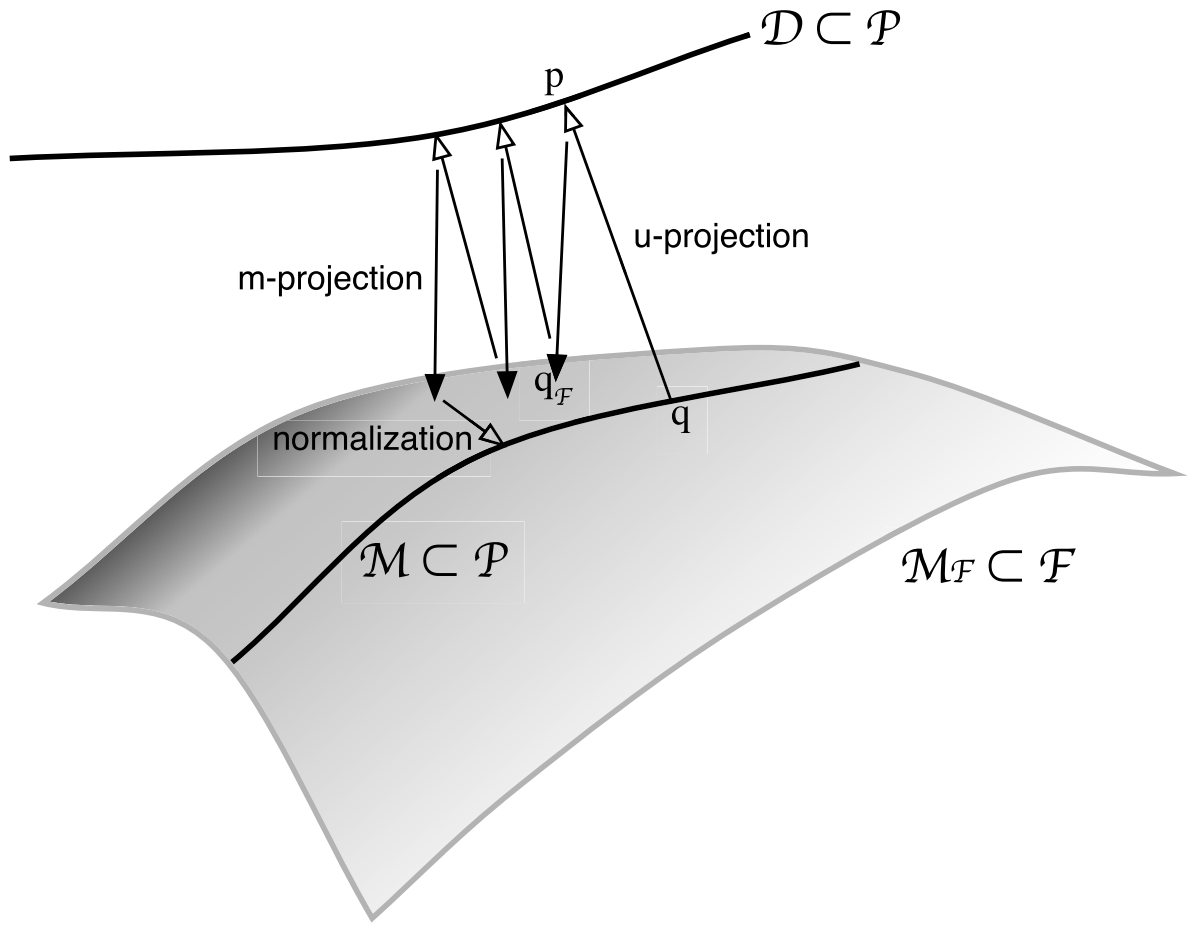
確率空間の中での最適化

UM アルゴリズムの幾何学的描像 (2)



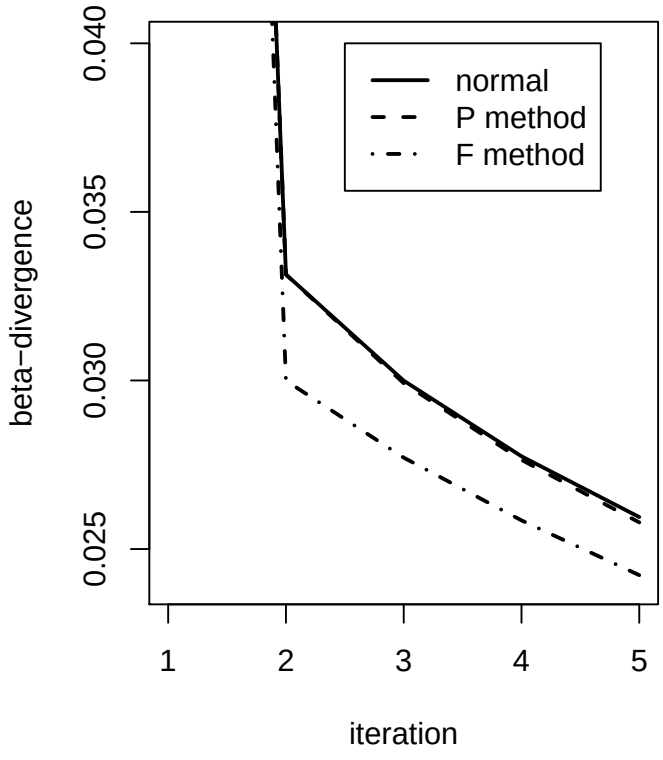
測度空間での最適化と逐次正規化 ($\mathcal{P}$ 法)

UM アルゴリズムの幾何学的描像 (3)

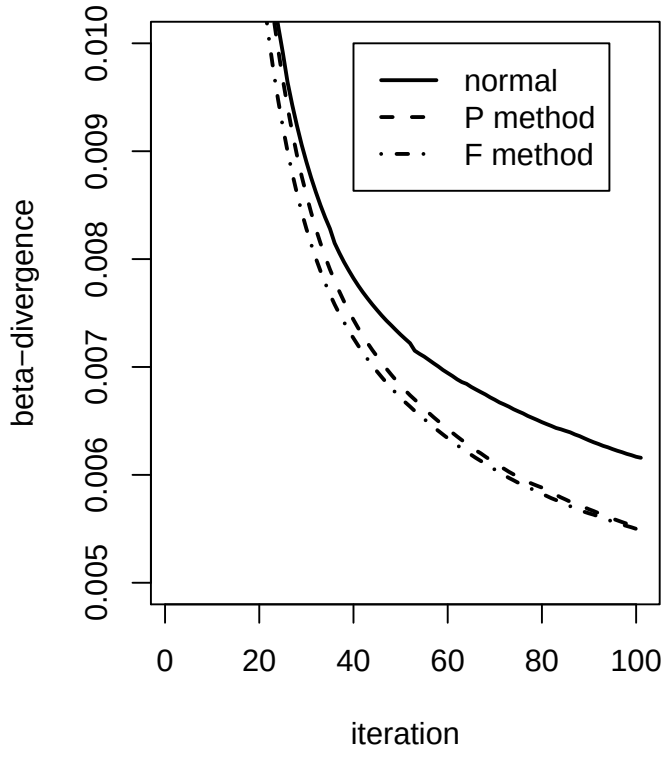


測度空間での最適化と正規化 ($\mathcal{F}$ 法)

UM アルゴリズムの収束特性



最初の 5 ステップの拡大図



100 ステップ

混合モデルの推定

- 混合モデル

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$$

- 隠れ変数 $y = (y_1, \ldots, y_K) \in \{0,1\}^K$: 指示変数

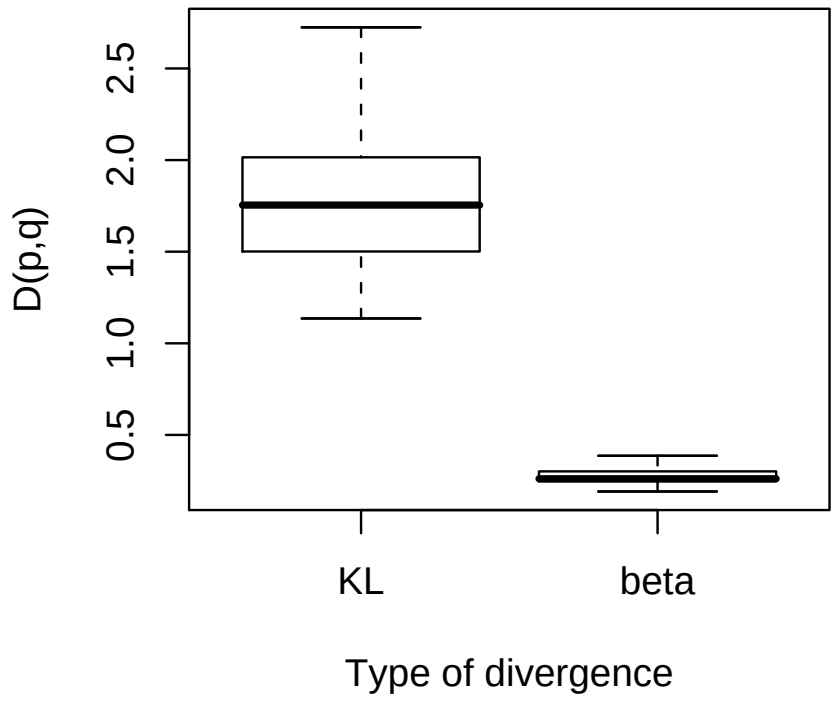
$$p(x,y) = \sum_{k=1}^{K} y_k \pi_k p_k(x) = \prod_{k=1}^{K} (\pi_k p_k(x))^{y_k} \quad (e\text{-flat})$$

M-step は重み付き最尤推定に帰着される.

- u-混合モデル

$$p(x,y) = u(\sum_{k=1}^{K} y_k f(x)) \quad (U\text{-flat})$$

β 情報量の頑健性



KL 情報量と β 情報量による推定結果

まとめ

- アルゴリズムと幾何学的描像
  - 情報の空間と Bregman divergence
  - ピタゴラスの定理と直交葉層化
  - 判別の不変性による測度空間での探索
  - 統計的推定問題としての統一的な視点
- アルゴリズムの理解と改善
  - 安定性解析
  - 頑健 (ロバスト) 化
  - 高速化・簡略化
  - モデルの拡大・縮小

まとめ

- アルゴリズムと幾何学的描像
  - 情報の空間と Bregman divergence
  - ピタゴラスの定理と直交葉層化
  - 判別の不変性による測度空間での探索
  - 統計的推定問題としての統一的な視点
- アルゴリズムの理解と改善
  - 安定性解析
  - 頑健 (ロバスト) 化
  - 高速化・簡略化
  - モデルの拡大・縮小

まとめ

- アルゴリズムと幾何学的描像
  - 情報の空間と Bregman divergence
  - ピタゴラスの定理と直交葉層化
  - 判別の不変性による測度空間での探索
  - 統計的推定問題としての統一的な視点
- アルゴリズムの理解と改善
  - 安定性解析
  - 頑健 (ロバスト) 化
  - 高速化・簡略化
  - モデルの拡大・縮小

まとめ

- アルゴリズムと幾何学的描像
  - 情報の空間と Bregman divergence
  - ピタゴラスの定理と直交葉層化
  - 判別の不変性による測度空間での探索
  - 統計的推定問題としての統一的な視点
- アルゴリズムの理解と改善
  - 安定性解析
  - 頑健 (ロバスト) 化
  - 高速化・簡略化
  - モデルの拡大・縮小

まとめ

- アルゴリズムと幾何学的描像
  - 情報の空間と Bregman divergence
  - ピタゴラスの定理と直交葉層化
  - 判別の不変性による測度空間での探索
  - 統計的推定問題としての統一的な視点
- アルゴリズムの理解と改善
  - 安定性解析
  - 頑健 (ロバスト) 化
  - 高速化・簡略化
  - モデルの拡大・縮小