カーネル法と深層学習の数理

鈴木大慈

東京大学大学院情報理工学系研究科数理情報学専攻
理研 AIP

2020 年 8 月 28 日/29 日
@広島市立大学集中講義

Outline

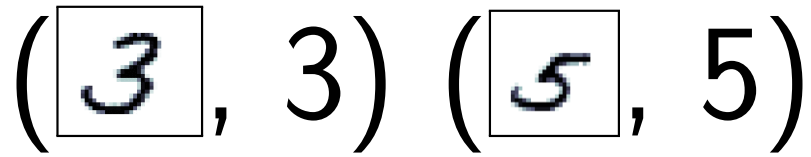
1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

機械学習の問題設定

- **教師あり学習**
  データが入力とそれに対するラベルの組で与えられる.
  新しい入力が来た時に対応するラベルを予測する問題.
  問題の例：回帰，判別

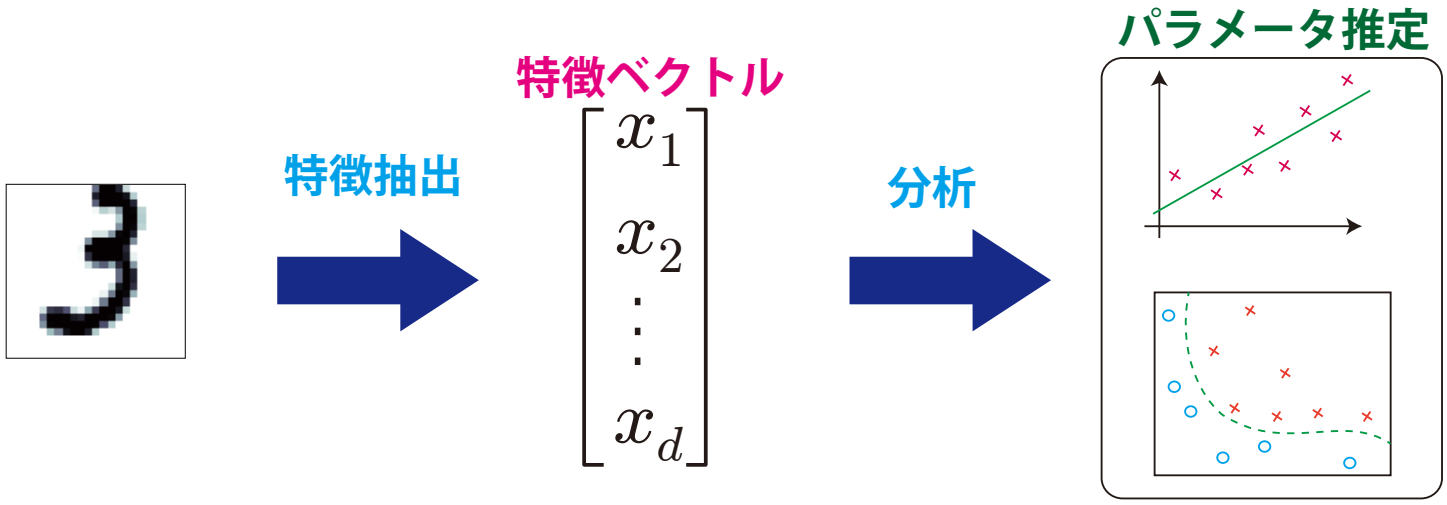


- **教師なし学習**
  データにラベルが付いていない.
  問題の例：クラスタリング，音源分離，異常検知

- **半教師あり学習**
  一部のデータにラベルが付いている.
- **強化学習**
  試行錯誤しながら自分でデータを集める.

機械学習の流れ

- **特徴抽出:** 画像などの対象を何らかの方法でベクトルに変換．(分野ごとのノウハウ)
- 一度特徴ベクトルに変換してしまえばあとは統計の問題．



予測モデルの構築 (θ: モデルのパラメータ)

（教師有り学習） $y = f(x; \theta)$

※深層学習は特徴抽出の部分をネットワーク構造を工夫することで学習に組み込んでいる．

損失関数を用いた定式化

教師あり/なし学習，いずれも損失関数の最小化として定式化できる．

- データの構造を表すパラメータ $\theta \in \Theta$ (Θ は仮説集合 (モデル))
  ← 「学習」$\approx \theta$ の推定
- 損失関数: パラメータ θ がデータ $z = (x, y)$ をどれだけよく説明しているか;

$$\ell(z, \theta) \quad (= \ell(y, f(x; \theta))).$$

汎化誤差 (期待誤差)：損失の期待値 → 汎化誤差最小化 ≈「学習」

$$\min_{\theta \in \Theta} \mathbb{E}_Z[\ell(Z, \theta)].$$

訓練誤差（経験誤差）：観測されたデータで代用,

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, \theta).$$

※ 訓練誤差と汎化誤差に差があることが機械学習における最適化の特徴.

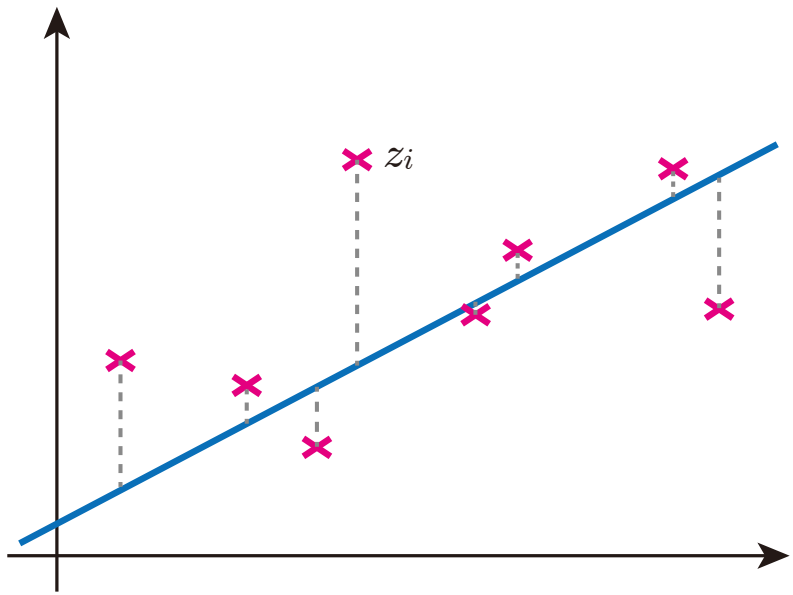
モデルの例 (教師あり)

- 回帰

$$z = (x, y) \in \mathbb{R}^{d+1}$$

$$\ell(z, \theta) = (y - \theta^\top x)^2 \qquad \text{(二乗誤差)}$$

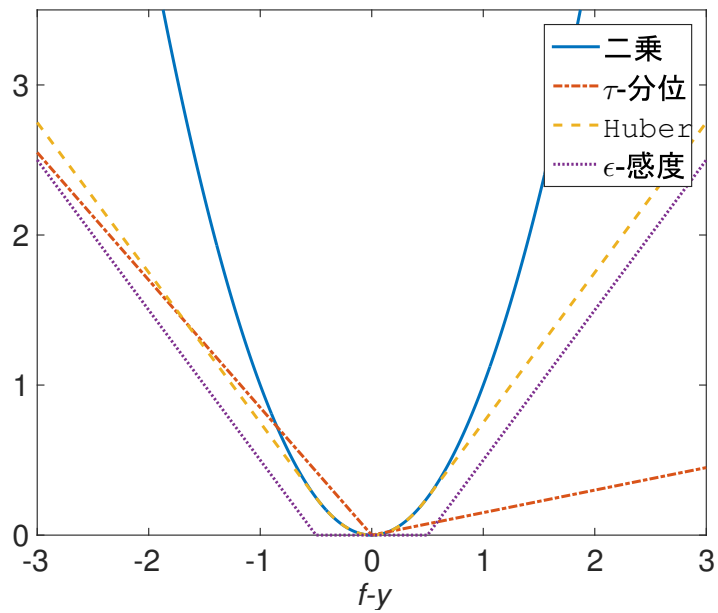
$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, \theta) = \min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2 \quad \text{(最小二乗法)}$$



教師あり学習の損失関数（回帰）

のデータ $z = (x, y)$ における $f = x^\top \theta$ の損失.

- 二乗損失: $\ell(y, f) = \frac{1}{2}(y - f)^2$.
- τ-分位点損失: $\ell(y, f) = (1 - \tau)\max\{f - y, 0\} + \tau \max\{y - f, 0\}$.
  ただし，$\tau \in (0, 1)$. 分位点回帰に用いられる.
- ϵ-感度損失: $\ell(y, f) = \max\{|y - f| - \epsilon, 0\}$,
  ただし，$\epsilon > 0$. サポートベクトル回帰に用いられる.

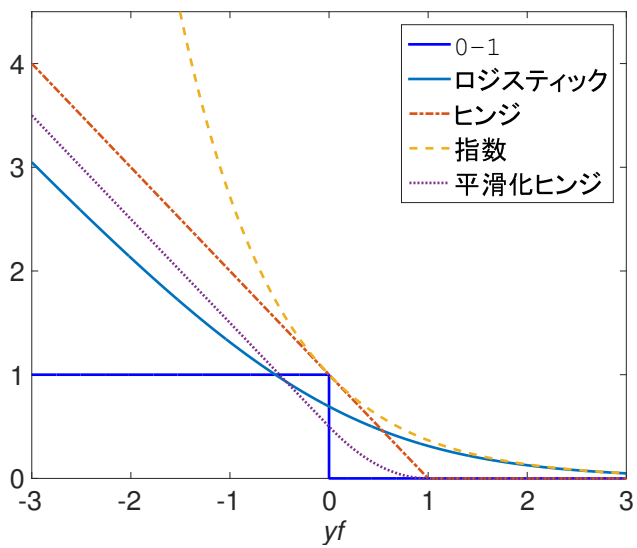


教師あり学習の損失関数（判別）

$y \in \{\pm 1\}$

- ロジスティック損失: $\ell(y, f) = \log((1 + \exp(-yf))/2)$.
- ヒンジ損失: $\ell(y, f) = \max\{1 - yf, 0\}$.
- 指数損失: $\ell(y, f) = \exp(-yf)$.
- 平滑化ヒンジ損失:

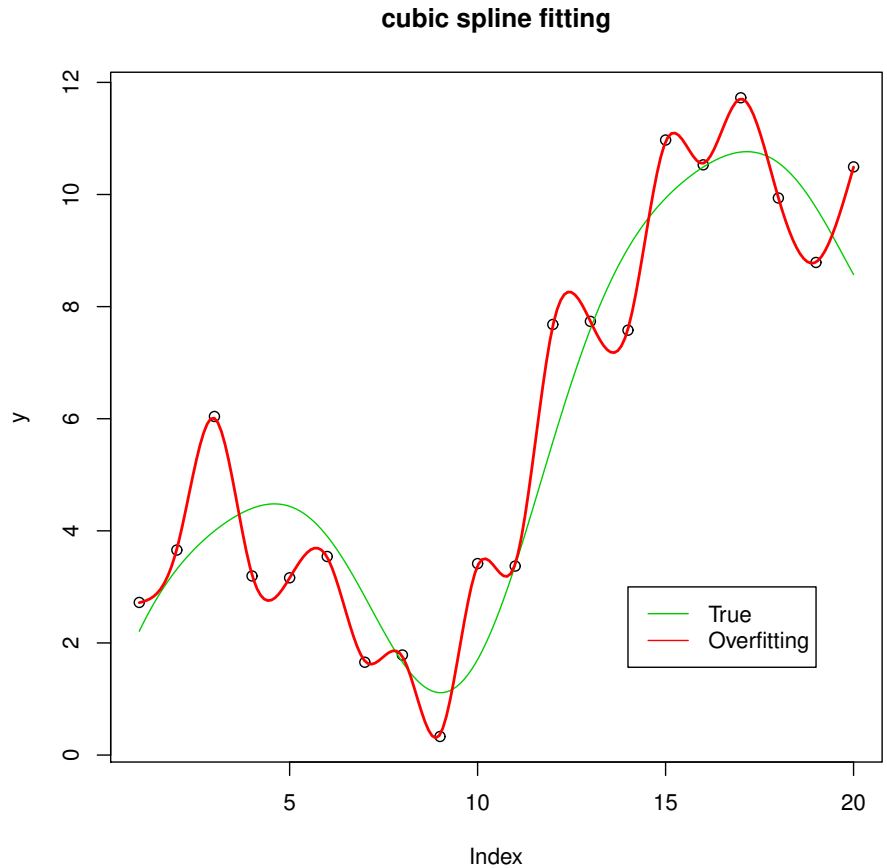
$$\ell(y, f) = \begin{cases} 0, & (yf \geq 1), \\ \frac{1}{2} - yf, & (yf < 0), \\ \frac{1}{2}(1 - yf)^2, & (\text{otherwise}). \end{cases}$$



過学習

経験誤差最小化と汎化誤差最小化には大きなギャップがある.

単なる経験誤差最小化は「過学習」を引き起こす.



正則化法

普通のロス関数 (負の対数尤度) 最小化:

$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i).$$

正則化付き損失関数最小化:

$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \underbrace{\psi(\beta)}_{\text{正則化項}}.$$

正則化項の例:

- リッジ正則化 (ℓ_2-正則化): $\psi(\beta) = \lambda \|\beta\|_2^2$
- ℓ_1-正則化: $\psi(\beta) = \lambda \|\beta\|_1$
- トレースノルム正則化: $\psi(W) = \mathrm{Tr}[(W^\top W)^{1/2}]$ $(W \in \mathbb{R}^{N \times M}$: 行列)

正則化法

普通のロス関数 (負の対数尤度) 最小化:

$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i).$$

正則化付き損失関数最小化:

$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \underbrace{\psi(\beta)}_{\text{正則化項}}.$$

正則化項の例:

- リッジ正則化 (ℓ_2-正則化): $\psi(\beta) = \lambda \|\beta\|_2^2$
- ℓ_1-正則化: $\psi(\beta) = \lambda \|\beta\|_1$
- トレースノルム正則化: $\psi(W) = \mathrm{Tr}[(W^\top W)^{1/2}]$ ($W \in \mathbb{R}^{N \times M}$: 行列)

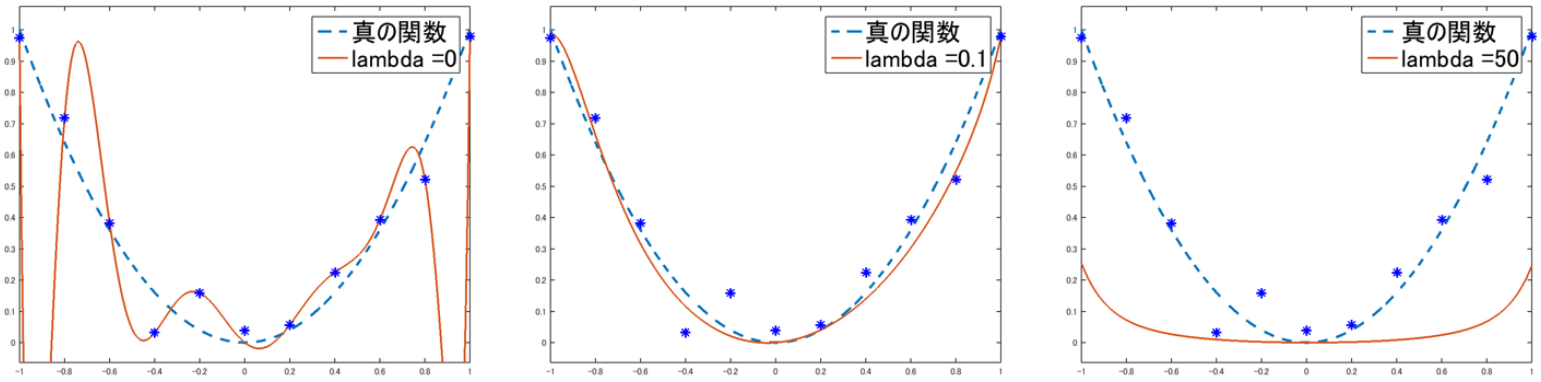
- 正則化項により分散が抑えられ，過学習が防がれる.
- その分，バイアスが乗る.

→ 適切な正則化の強さ (λ) を選ぶ必要がある.

正則化の例： リッジ正則化と過学習

多項式回帰（15 次多項式）

$$\min_{\theta \in \mathbb{R}^{15}} \frac{1}{n} \sum_{i=1}^{n} \{y_i - (\beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{15} x_i^{15})\}^2 + \lambda \|\beta\|_2^2$$

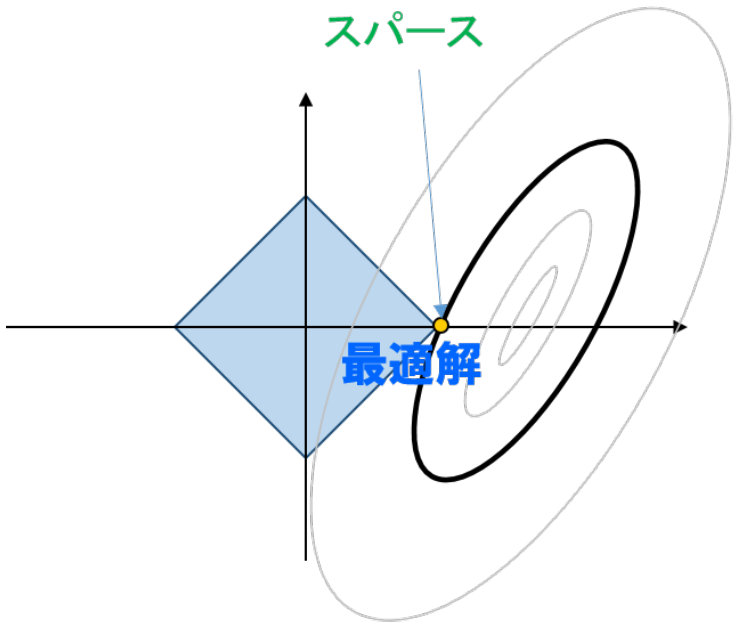


正則化の例：ℓ_1-正則化（スパース推定）

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

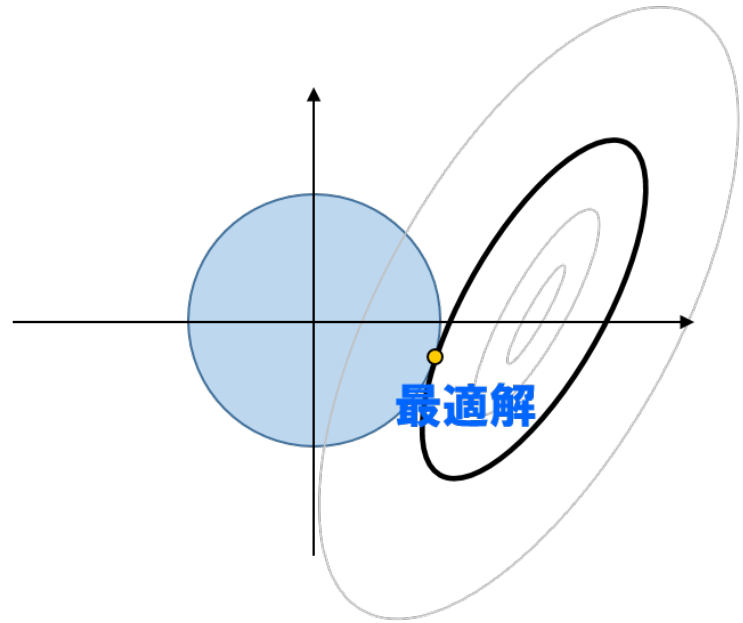
L1正則化

$$\|\beta\|_1 = |\beta_1| + \cdots + |\beta_p|$$



L2正則化（リッジ正則化）

$$\|\beta\|_2^2 = \beta_1^2 + \cdots + \beta_p^2$$



座表軸の上に乗りやすい

R. Tsibshirani (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267–288.

スパース性の恩恵

$y_i = x_i^\top \beta^* + \epsilon_i$ $(i = 1, \ldots, n)$. β^*：真のベクトル.

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

$x_i \in \mathbb{R}^p$ (p 次元), $d = \|\beta^*\|_0$ (真の非 0 要素の数) とする.

Theorem (Lasso の収束レート)

ある条件のもと，ある定数 C が存在して

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C \frac{d \log(p)}{n}.$$

※全体の次元 p はたかだか $O(\log(p))$ でしか影響しない！
実質的次元 d が支配的.

$$\text{(Lasso)} \quad \frac{d \log(p)}{n} \ll \frac{p}{n} \quad \text{(最小二乗法)}$$

制限固有値条件 (Restricted eigenvalue condition)

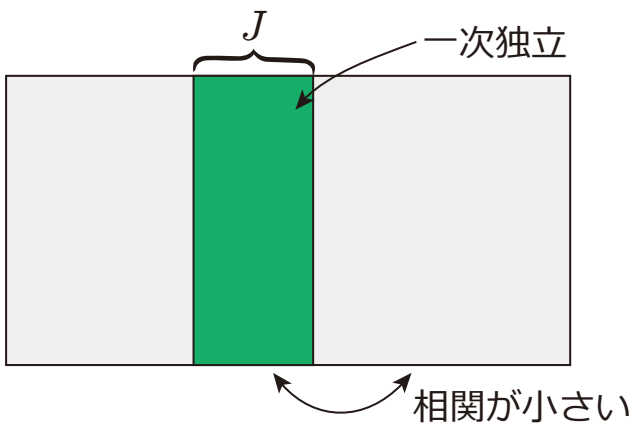
$A = \frac{1}{n} X^\top X$ とする.

Definition (制限固有値条件 (RE(k', C)))

$$\phi_{\mathrm{RE}}(k', C) = \phi_{\mathrm{RE}}(k', C, A) := \inf_{\substack{J \subseteq \{1,\dots,n\}, v \in \mathbb{R}^p : \\ |J| \leq k', C\|v_J\|_1 \geq \|v_{J^c}\|_1}} \frac{v^\top A v}{\|v_J\|_2^2}$$

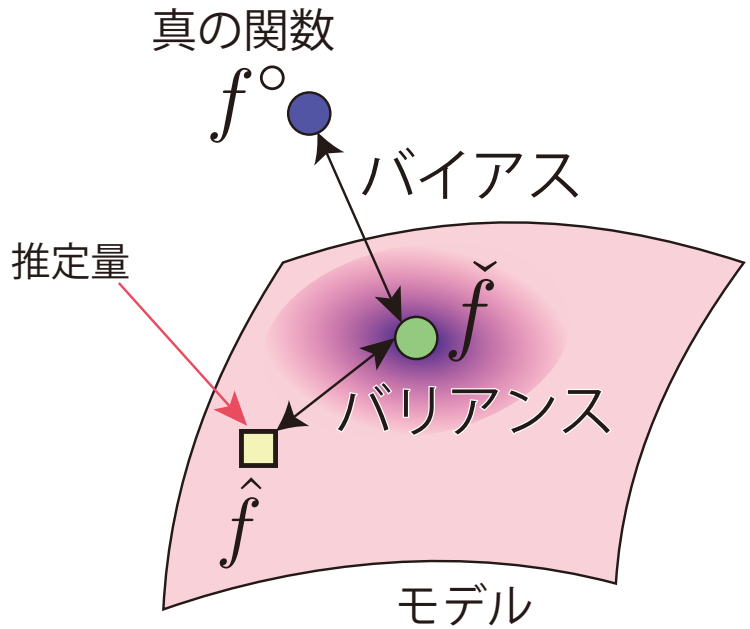
に対し，$\phi_{\mathrm{RE}} > 0$ が成り立つ.

- ほぼスパースなベクトルに制限して定義した最小固有値.
- $k' = 2d$ で成り立っていればよい.
- ランダムな X に対して高確率で成り立つことが示せる: Johnson Lindenstrauss の補題 (Johnson et al., 1986, Dasgupta and Gupta, 1999, Rudelson and Zhou, 2013).

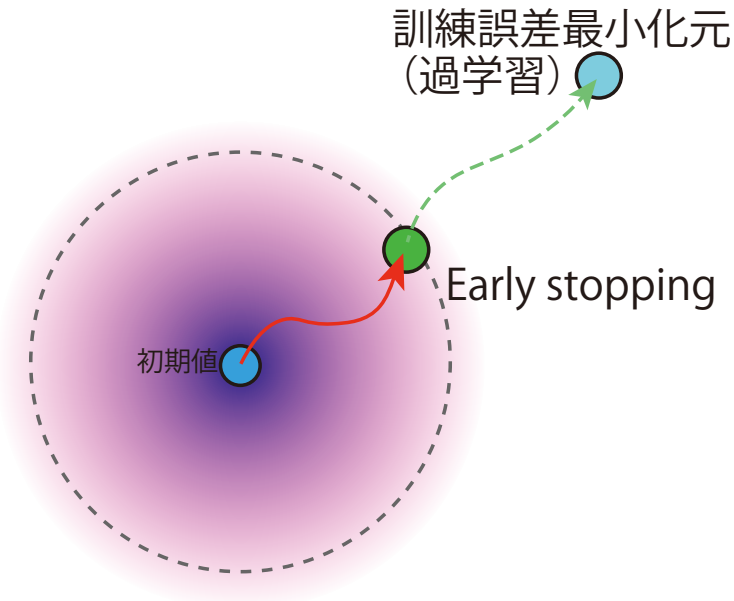


正則化と最適化

モデルの制限による正則化



Early stopping による正則化

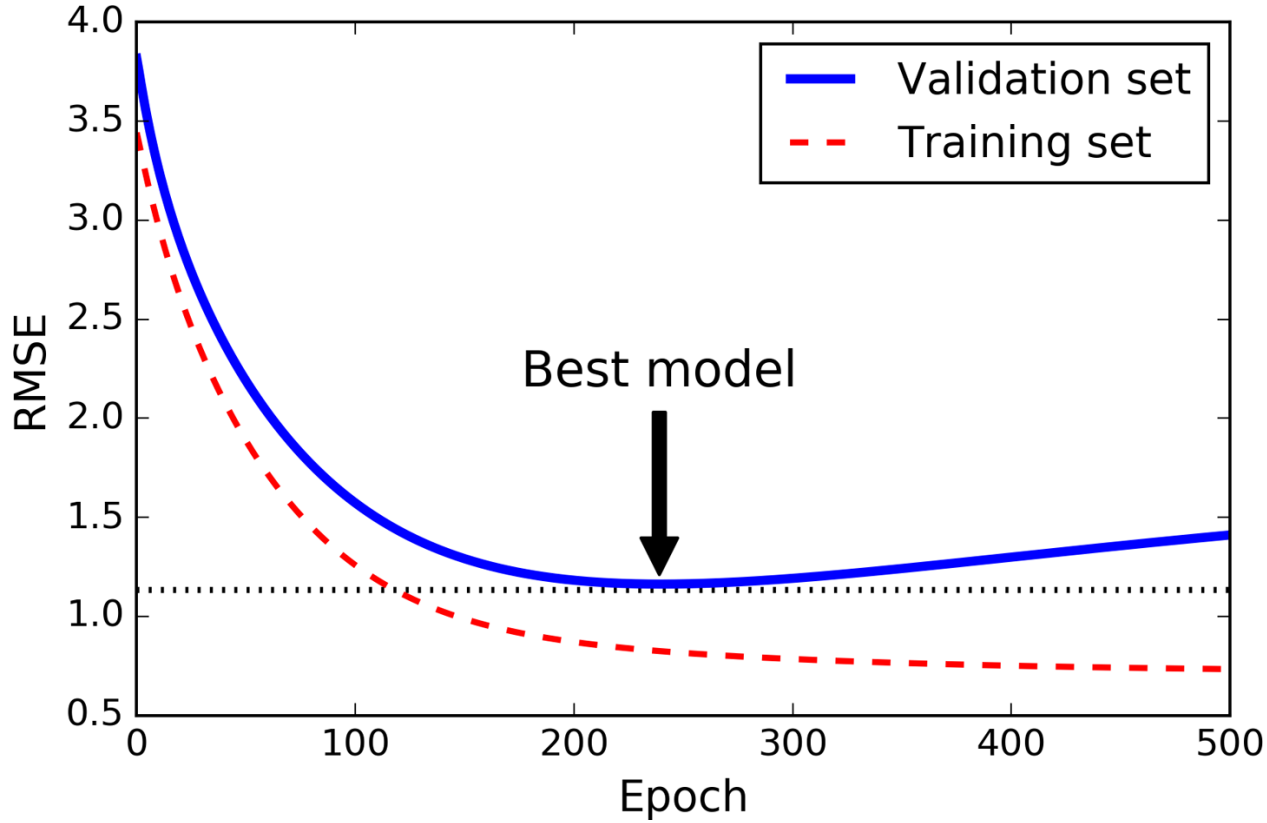


バイアス-バリアンス分解

$$\underbrace{\|f^{\circ} - \hat{f}\|_{L_2(P_X)}}_{\text{Estimation error}} \leq \underbrace{\|f^{\circ} - \check{f}\|_{L_2(P_X)}}_{\substack{\text{Approximation error} \\ \text{(bias)}}} + \underbrace{\|\check{f} - \hat{f}\|_{L_2(P_X)}}_{\substack{\text{Sample deviation} \\ \text{(variance)}}}$$

訓練誤差最小化元に達する前に止める (early stopping) ことで正則化が働く.
→ 深層学習，Boosting の常套手段.

Early stopping による過学習の回避



Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurlien Gron.
Chapter 4. Training Models.
https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ch04.html

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

（今からお話しする）学習理論 $\approx$ 経験過程の理論

$$\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}[f] \right\}$$

の評価が重要.

歴史: 経験過程の理論

1933	**Glivenko, Cantelli**	**Glivenko-Catelli** の定理 (一様大数の法則)
1933	**Kolmogorov**	**Kolmogorov-Smirnov** 検定 (収束レート，漸近分布)
1952	**Donsker**	**Donsker** の定理 (一様中心極限定理)
1967	**Dudley**	**Dudley** 積分
1968	**Vapnik, Chervonenkis**	**VC** 次元 (一様収束の必要十分条件)
1996a	**Talagrand**	**Talagrand** の不等式

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

問題設定

教師有り学習

教師データ: $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$　入力と出力の i.i.d. 系列
ロス関数: $\ell(\cdot, \cdot) : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$　間違いへのペナルティ
仮説集合 (モデル): $\mathcal{F}$　$\mathcal{X} \to \mathbb{R}$ なる関数の集合

$\hat{f}$: 推定量. サンプル $(x_i, y_i)_{i=1}^n$ から構成される $\mathcal{F}$ の元.

抑えたい量 (汎化誤差):

$$\underbrace{\mathbb{E}_{(X,Y)}}_{\text{テストデータ}} [\ell(Y, \hat{f}(X))] - \inf_{f:\text{可測関数}} \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$$

- 汎化誤差は収束する？
- その速さは?

Bias-Variance の分解

経験リスク: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$,
期待リスク: $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$

$$
\begin{aligned}
\text{汎化誤差} &= L(\hat{f}) - \inf_{f:\text{可測関数}} L(f) \\
&= \underbrace{L(\hat{f}) - \inf_{f\in\mathcal{F}} L(f)}_{\text{推定誤差}} + \underbrace{\inf_{f\in\mathcal{F}} L(f) - \inf_{f:\text{可測関数}} L(f)}_{\text{モデル誤差}}
\end{aligned}
$$

簡単のため $f^* \in \mathcal{F}$ が存在して $\inf_{f\in\mathcal{F}} L(f) = L(f^*)$ とする.

Bias-Variance の分解

経験リスク: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$,
期待リスク: $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$

$$
\begin{aligned}
\text{汎化誤差} &= L(\hat{f}) - \inf_{f:\text{可測関数}} L(f) \\
&= \underbrace{L(\hat{f}) - \inf_{f\in\mathcal{F}} L(f)}_{\text{推定誤差}} + \underbrace{\inf_{f\in\mathcal{F}} L(f) - \inf_{f:\text{可測関数}} L(f)}_{\text{モデル誤差}}
\end{aligned}
$$

簡単のため $f^* \in \mathcal{F}$ が存在して $\inf_{f\in\mathcal{F}} L(f) = L(f^*)$ とする.

※モデル誤差については今回は触れない.
しかし, モデリングの問題は非常に重要.

- Sieve 法, Cross validation, 情報量規準, モデル平均, ...
- カーネル法におけるモデル誤差の取り扱い: interpolation space の理論 (Steinwart et al., 2009, Eberts and Steinwart, 2012, Bennett and Sharpley, 1988).

以降, モデル誤差は十分小さいとする.

経験誤差最小化

経験誤差最小化 **(ERM):**

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \quad \hat{L}(f)$$

正則化付き経験誤差最小化 **(RERM):**

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \quad \hat{L}(f) + \underbrace{\psi(f)}_{\text{正則化項}}$$

- RERM に関する研究も非常に沢山ある (Steinwart and Christmann, 2008, Mukherjee et al., 2002).
- ERM の延長線上.

経験誤差最小化

経験誤差最小化 **(ERM):** ☆

$$\hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \quad \hat{L}(f)$$

正則化付き経験誤差最小化 **(RERM):**

$$\hat{f} = \underset{f \in \mathcal{F}}{\arg\min} \quad \hat{L}(f) + \underbrace{\psi(f)}_{\text{正則化項}}$$

- RERM に関する研究も非常に沢山ある (Steinwart and Christmann, 2008, Mukherjee et al., 2002).
- ERM の延長線上.

出発点

ほとんどのバウンドの導出は次の式から始まる:

$$\hat{L}(\hat{f}) \leq \hat{L}(f^*) \qquad (\because \text{経験誤差最小化})$$
$$\Rightarrow \quad L(\hat{f}) - L(f^*) \leq L(\hat{f}) - \hat{L}(\hat{f}) + \hat{L}(f^*) - L(f^*)$$

Reminder: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$, $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$

出発点

ほとんどのバウンドの導出は次の式から始まる:

$$\hat{L}(\hat{f}) \leq \hat{L}(f^*) \qquad (\because \text{経験誤差最小化})$$

$$\Rightarrow \quad \underbrace{L(\hat{f}) - L(f^*)}_{\text{汎化誤差}} \leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{?} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n})\ \text{(後述)}}$$

Reminder: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$, $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$

出発点

ほとんどのバウンドの導出は次の式から始まる:

$$\hat{L}(\hat{f}) \leq \hat{L}(f^*) \qquad (\because \text{経験誤差最小化})$$

$$\Rightarrow \quad \underbrace{L(\hat{f}) - L(f^*)}_{\text{汎化誤差}} \leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{?} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n})\ \text{(後述)}}$$

安易な解析

$$L(\hat{f}) - \hat{L}(\hat{f}) \begin{cases} \to 0 & (\because \text{大数の法則!!}) \\ = O_p(1/\sqrt{n}) & (\because \text{中心極限定理!!}) \end{cases}$$

楽勝 !!!

Reminder: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$, $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$

出発点

ほとんどのバウンドの導出は次の式から始まる:

$$\hat{L}(\hat{f}) \leq \hat{L}(f^*) \qquad (\because \text{経験誤差最小化})$$

$$\Rightarrow \quad \underbrace{L(\hat{f}) - L(f^*)}_{\text{汎化誤差}} \leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{?} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n})\ \text{(後述)}}$$

安易な解析

$$L(\hat{f}) - \hat{L}(\hat{f}) \begin{cases} \to 0 & (\because \text{大数の法則!!}) \\ = O_p(1/\sqrt{n}) & (\because \text{中心極限定理!!}) \end{cases}$$

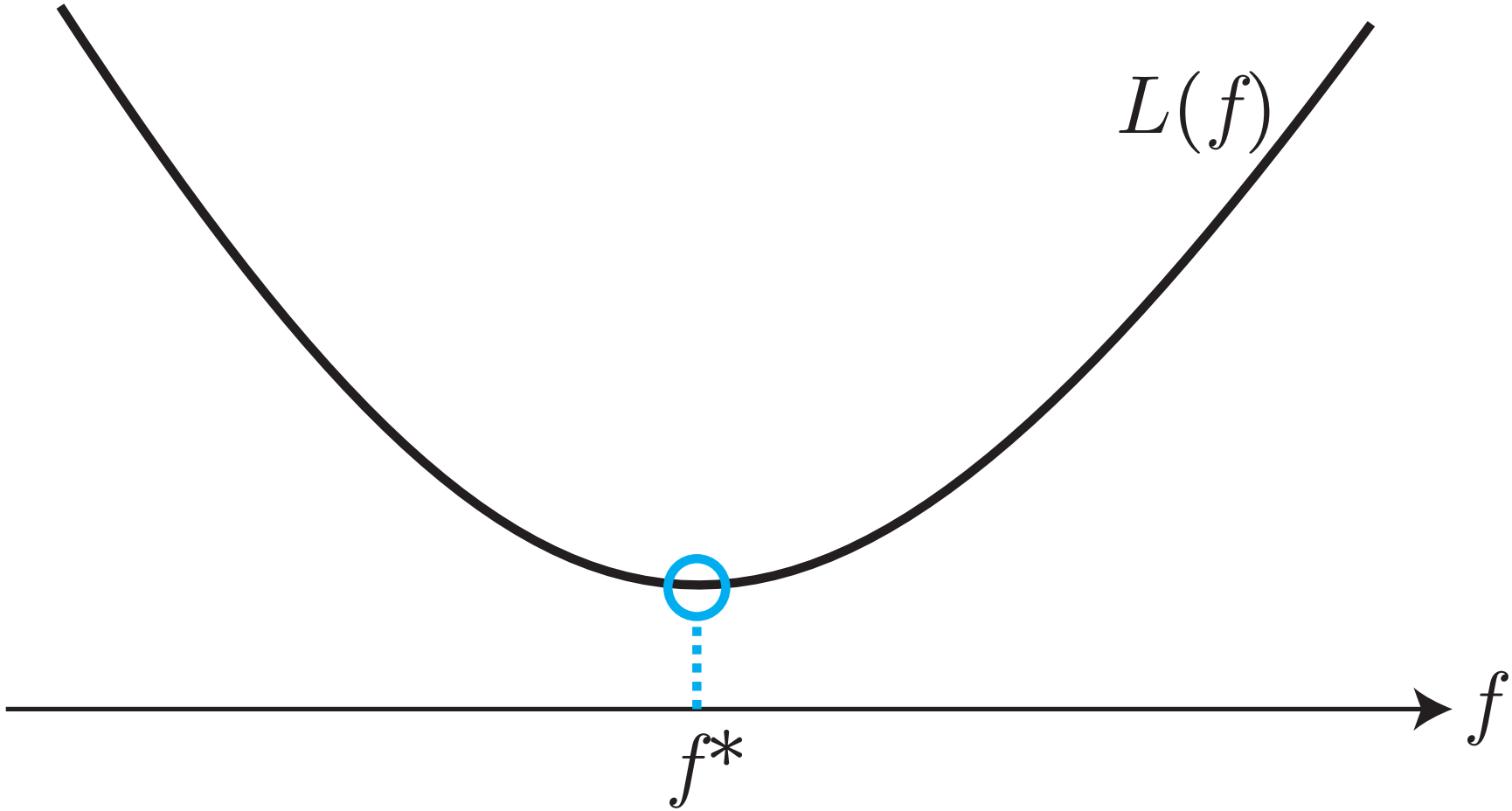
楽勝 !!!

ダメです

$\hat{f}$ と教師データは独立ではない

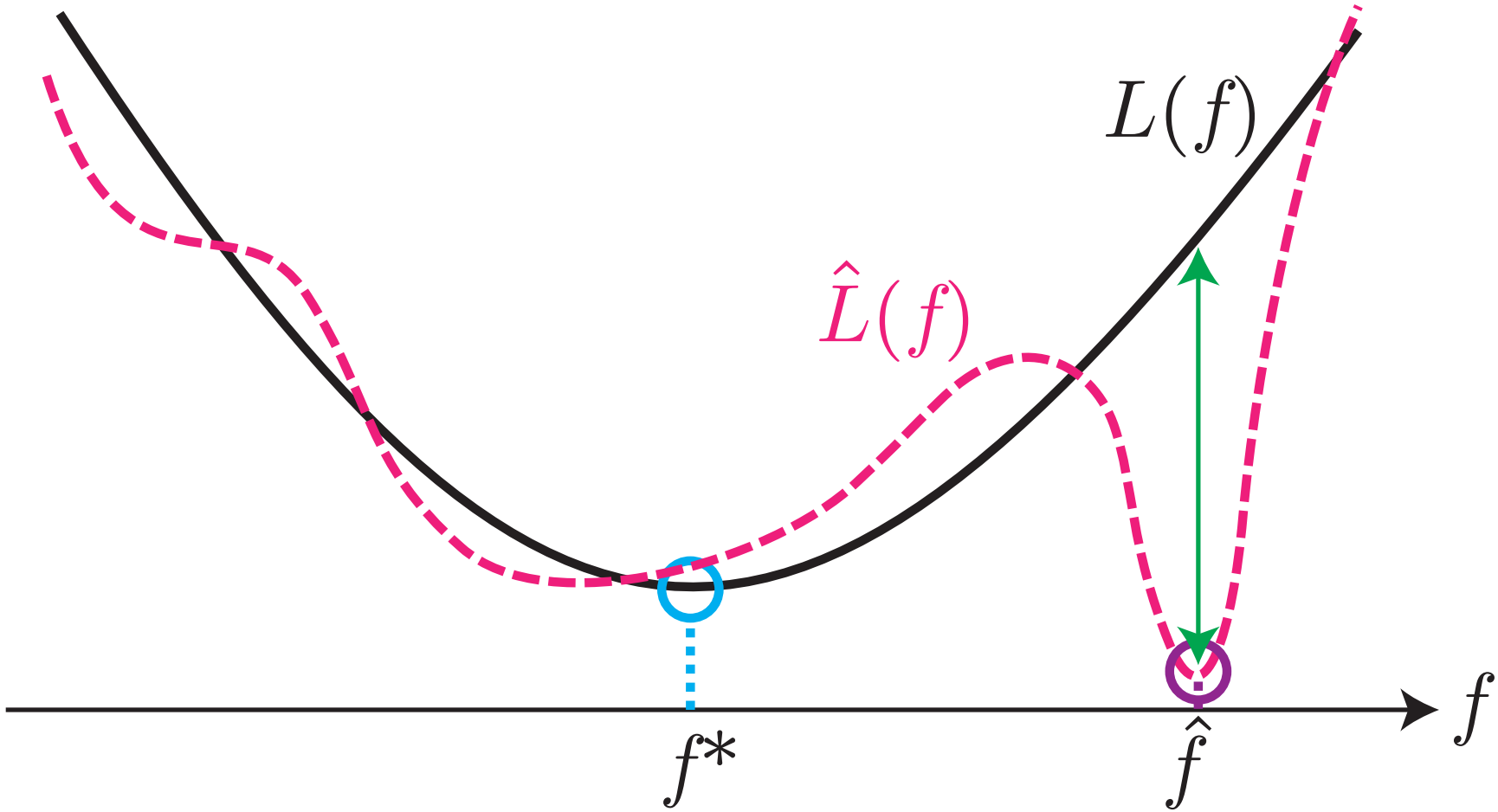
Reminder: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$, $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$

なにが問題か？



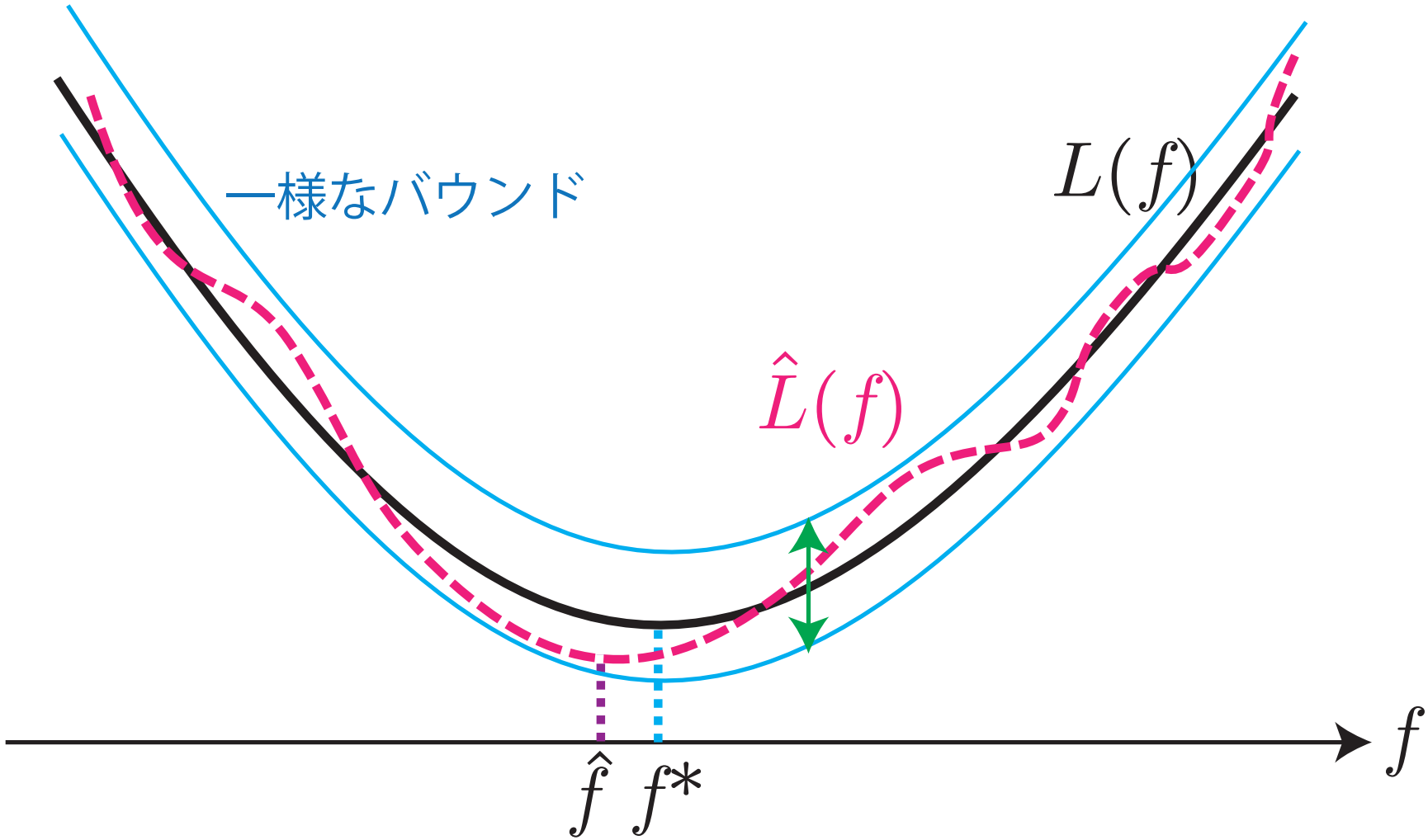
$$L(f) = \mathbb{E}[\ell(Y, f(X))], \quad \widehat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$$

なにが問題か？



“たまたま” うまくいくやつがいる (過学習) かもしれない.
実際，$\mathcal{F}$ が複雑な場合収束しない例が

なにが問題か？



一様なバウンドによって「たまたまうまくいく」が (ほとんど) ないことを保証
それは自明ではない (経験過程の理論)

一様バウンド

$$L(\hat{f}) - \hat{L}(\hat{f}) \leq \sup_{f \in \mathcal{F}} \left\{ L(f) - \hat{L}(f) \right\} \leq (?)$$

一様にリスクを抑えることが重要

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

まずは有限から

$|\mathcal{F}| < \infty$

有用な不等式

- Hoeffding の不等式
  Z_i $(i = 1, \ldots, n)$: 独立で (同一とは限らない) 期待値 0 の確率変数 s.t. $|Z_i| \leq m_i$

$$P\left(\frac{|\sum_{i=1}^n Z_i|}{\sqrt{n}} > t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n m_i^2/n}\right)$$

- Bernstein の不等式
  Z_i $(i = 1, \ldots, n)$: 独立で (同一とは限らない) 期待値 0 の確率変数 s.t. $\mathbb{E}[Z_i^2] = \sigma_i^2$, $|Z_i| \leq M$

$$P\left(\frac{|\sum_{i=1}^n Z_i|}{\sqrt{n}} > t\right) \leq 2\exp\left(-\frac{t^2}{2(\frac{1}{n}\sum_{i=1}^n \sigma_i^2 + \frac{1}{\sqrt{n}}Mt)}\right)$$

分散の情報を利用

有用な不等式: 拡張版

- Hoeffding の不等式 (sub-Gaussian tail)
  Z_i $(i = 1, \ldots, n)$: 独立で (同一とは限らない) 期待値 0 の確率変数 s.t.
  $\mathbb{E}[e^{\tau Z_i}] \leq e^{\sigma_i^2 \tau^2 / 2}$ $(\forall \tau > 0)$

$$P\left(\frac{|\sum_{i=1}^n Z_i|}{\sqrt{n}} > t \right) \leq 2 \exp\left(- \frac{t^2}{2 \sum_{i=1}^n \sigma_i^2 / n} \right)$$

- Bernstein の不等式
  Z_i $(i = 1, \ldots, n)$: 独立で (同一とは限らない) 期待値 0 の確率変数 s.t.
  $\mathbb{E}[Z_i^2] = \sigma_i^2$, $\mathbb{E}|Z_i|^k \leq \frac{k!}{2} \sigma^2 M^{k-2}$ $(\forall k \geq 2)$

$$P\left(\frac{|\sum_{i=1}^n Z_i|}{\sqrt{n}} > t \right) \leq 2 \exp\left(- \frac{t^2}{2(\frac{1}{n} \sum_{i=1}^n \sigma_i^2 + \frac{1}{\sqrt{n}} M t)} \right)$$

(ヒルベルト空間版もある)

有限集合の一様バウンド 1: Hoeffding の不等式版

これだけでも知っていると有用. ($f \leftarrow \ell(y, g(x)) - \mathbb{E}\ell(Y, g(X))$ として考える)

$\mathcal{F} = \{f_m \ (m = 1, \ldots, M)\}$ 有限個の関数集合: どれも期待値 0 ($\mathbb{E}[f_m(X)] = 0$).

Hoeffding の不等式 ($Z_i = f_m(X_i)$ を代入)

$$P\left(\frac{|\sum_{i=1}^n f_m(X_i)|}{\sqrt{n}} > t\right) \leq 2\exp\left(-\frac{t^2}{2\|f_m\|_\infty^2}\right)$$

一様バウンド

- $P\left(\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n f_m(X_i)|}{\sqrt{n}} > \max_m \|f_m\|_\infty \underline{\sqrt{2\log(2M/\delta)}}\right) \leq \delta$
- $\mathbb{E}\left[\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n f_m(X_i)|}{\sqrt{n}}\right] \leq C \max_m \|f_m\|_\infty \underline{\sqrt{\log(1+M)}}$

(導出) $P\left(\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n f_m(X_i)|}{\sqrt{n}} > t\right) = P\left(\bigcup_{1\leq m\leq M} \left\{\frac{|\sum_{i=1}^n f_m(X_i)|}{\sqrt{n}} > t\right\}\right) \leq 2\sum_{m=1}^M \exp\left(-\frac{t^2}{2\|f_m\|_\infty^2}\right)$

有限集合の一様バウンド 2: Bernstein の不等式版

$\mathcal{F} = \{f_m \ (m = 1, \ldots, M)\}$ 有限個の関数集合: どれも期待値 0 ($\mathbb{E}[f_m(X)] = 0$).

Bernstein の不等式

$$P\left(\frac{|\sum_{i=1}^n f_m(X_i)|}{\sqrt{n}} > t\right) \leq 2\exp\left(-\frac{t^2}{2(\|f_m\|_{L_2}^2 + \frac{1}{\sqrt{n}}\|f_m\|_\infty t)}\right)$$

一様バウンド

$$\begin{aligned}
&\mathbb{E}\left[\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n f_m(X_i)|}{\sqrt{n}}\right] \\
&\lesssim \frac{1}{\sqrt{n}} \max_m \|f_m\|_\infty \log(1+M) + \max_m \|f_m\|_{L_2} \sqrt{\log(1+M)}
\end{aligned}$$

※ 一様バウンドはせいぜい $\sqrt{\log(M)}$ オーダで増える.

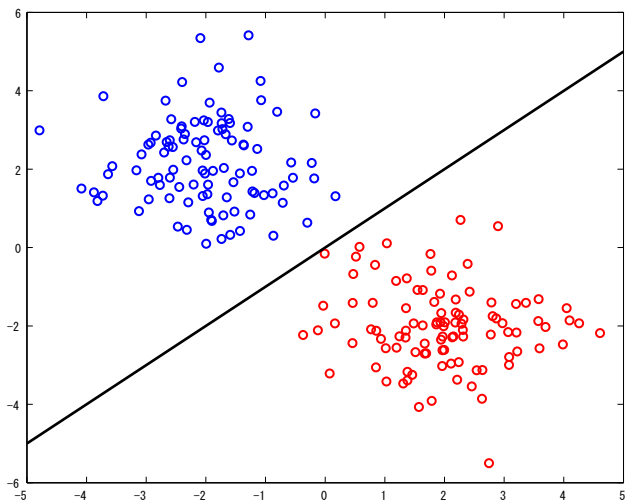
Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

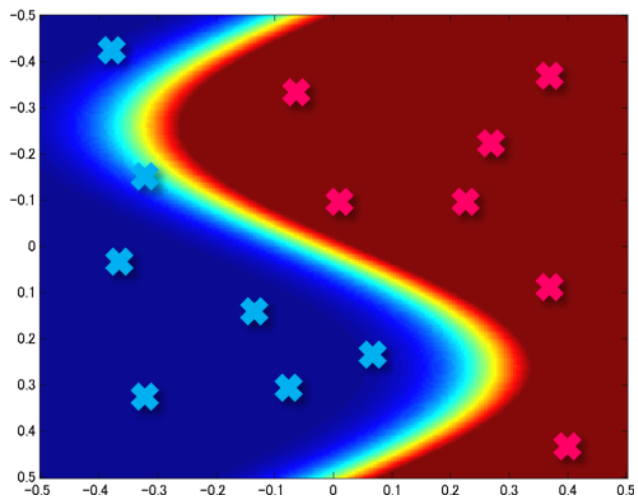
有限から無限へ

仮説集合の要素が無限個あったら？
連続濃度をもっていたら？

$\mathcal{F} = \{x^\top \beta \mid \beta \in \mathbb{R}^d, \|\beta\| \leq 1\}$

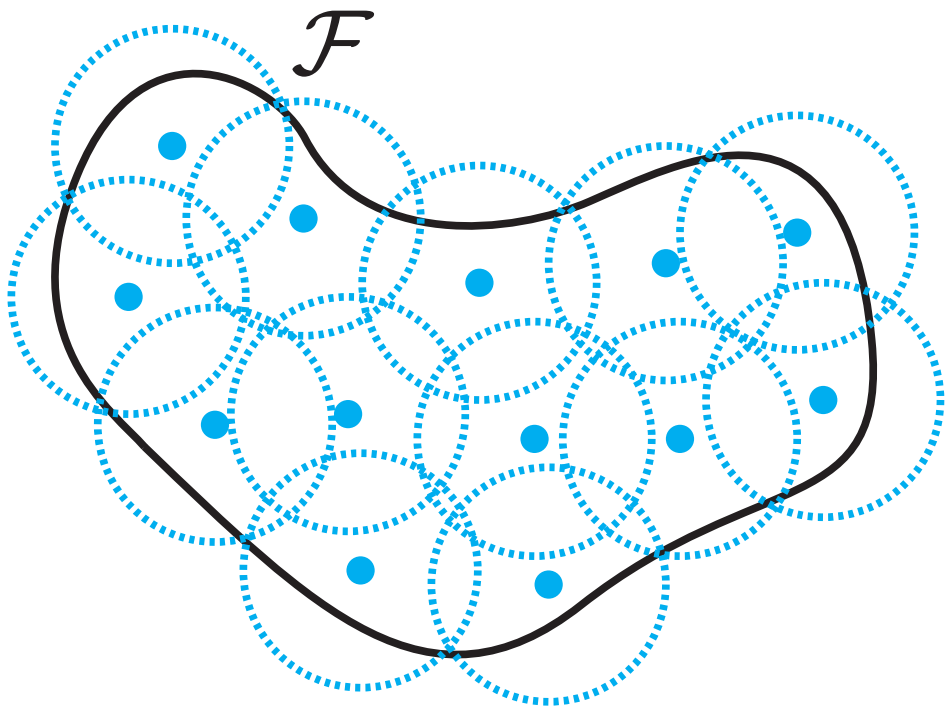


$\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$



基本的なアイディア

有限個の元で代表させる



Rademacher 複雑さ

$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$: Rademacher 変数, i.e., $P(\epsilon_i = 1) = P(\epsilon_i = -1) = \frac{1}{2}$.

Rademacher 複雑さ

$$R(\mathcal{F}) := \mathbb{E}_{\{\epsilon_i\},\{x_i\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right]$$

対称化:

$$\text{(期待値)} \qquad \mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} (f(x_i) - \mathbb{E}[f]) \right| \right] \leq 2R(\mathcal{F}).$$

もし $\|f\|_\infty \leq 1$ $(\forall f \in \mathcal{F})$ なら

$$\text{(裾確率)} \qquad P \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} (f(x_i) - \mathbb{E}[f]) \right| \geq 2R(\mathcal{F}) + \sqrt{\frac{t}{2n}} \right) \leq e^{-t}.$$

Rademacher 複雑さを抑えれば一様バウンドが得られる！

Rademacher 複雑さの各種性質

- Contraction inequality: もし ψ が Lipschitz 連続なら, i.e., $|\psi(f) - \psi(f')| \leq B|f - f'|$,

$$R(\{\psi(f) \mid f \in \mathcal{F}\}) \leq 2BR(\mathcal{F}).$$

- 凸包: $\mathrm{conv}(\mathcal{F})$ を $\mathcal{F}$ の元の凸結合全体からなる集合とする.

$$R(\mathrm{conv}(\mathcal{F})) = R(\mathcal{F})$$

Rademacher 複雑さの各種性質

- Contraction inequality: もし ψ が Lipschitz 連続なら, i.e., $|\psi(f) - \psi(f')| \leq B|f - f'|$,

$$R(\{\psi(f) \mid f \in \mathcal{F}\}) \leq 2BR(\mathcal{F}).$$

- 凸包: $\mathrm{conv}(\mathcal{F})$ を $\mathcal{F}$ の元の凸結合全体からなる集合とする.

$$R(\mathrm{conv}(\mathcal{F})) = R(\mathcal{F})$$

特に最初の性質が有り難い.

$|\ell(y, f) - \ell(y, f')| \leq |f - f'|$ なら,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |\hat{L}(f) - L(f)|\right] \leq 2R(\ell(\mathcal{F})) \leq 4R(\mathcal{F}),$$

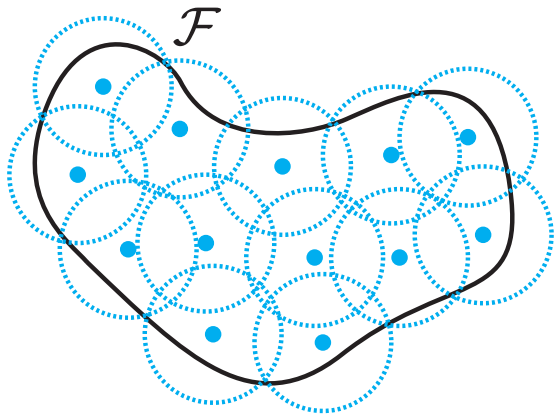
ただし, $\ell(\mathcal{F}) = \{\ell(\cdot, f(\cdot)) \mid f \in \mathcal{F}\}$.

よって $\mathcal{F}$ の Rademacher complexity を抑えれば十分！
Lipschitz 連続性はヒンジロス, ロジスティックロスなどで成り立つ. さらに y と $\mathcal{F}$ が有界なら二乗ロスなどでも成り立つ.

Reminder: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))$, $L(f) = \mathbb{E}_{(X,Y)}[\ell(Y, f(X))]$

カバリングナンバー

Rademacher complexity を抑える方法.
カバリングナンバー: 仮説集合 $\mathcal{F}$ の複雑さ・容量.



ϵ-カバリングナンバー

$$N(\mathcal{F}, \epsilon, d)$$

ノルム d で定まる半径 ϵ のボールで $\mathcal{F}$ を覆うために必要な最小のボールの数.

有限個の元で $\mathcal{F}$ を近似するのに最低限必要な個数.

Theorem (Dudley 積分)

$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f(x_i)^2$ とすると,

$$R(\mathcal{F}) \leq C \inf_{\alpha>0}\left\{\alpha + \frac{1}{\sqrt{n}}\mathbb{E}_{D_n}\left[\int_\alpha^\infty \sqrt{\log(N(\mathcal{F}, \epsilon, \|\cdot\|_n))}\mathrm{d}\epsilon\right]\right\}.$$

(Boucheron et al. (2013) の Lemma 11.4 や Wainwright (2019) の Theorem 5.22 を参照)

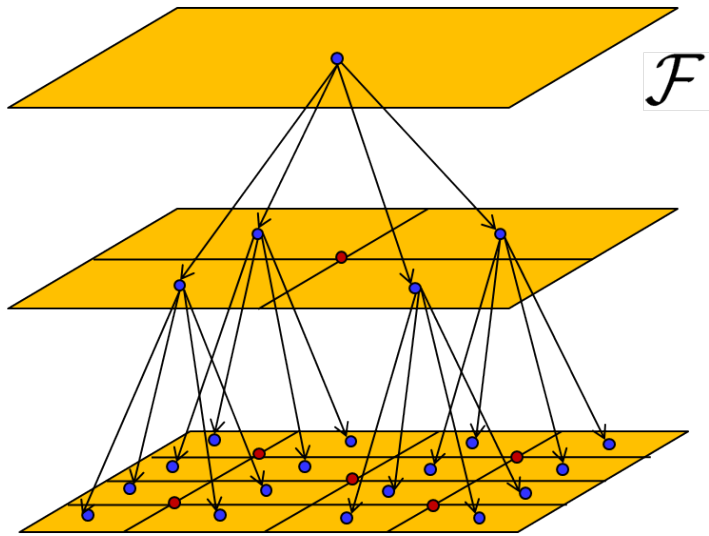
Dudley 積分のイメージ

$$R(\mathcal{F}) \leq \frac{C}{\sqrt{n}} \mathbb{E}_{D_n} \left[\int_0^\infty \sqrt{\log(N(\mathcal{F}, \epsilon, \|\cdot\|_n))} \mathrm{d}\epsilon \right].$$

有限個の元で $\mathcal{F}$ を近似する.

その解像度を細かくしていって，似ている元をまとめ上げてゆくイメージ.

チェイニングという.



これまでのまとめ

$$\hat{L}(\hat{f}) \leq \hat{L}(f^*) \qquad (\because \text{経験誤差最小化})$$

$$\Rightarrow \quad L(\hat{f}) - L(f^*) \leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{\text{これを抑えたい}} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n})\ \text{(Hoeffding)}}$$

ℓ が 1-Lipschitz ($|\ell(y,f) - \ell(y,f')| \leq |f - f'|$) かつ $\|f\|_\infty \leq 1$ ($\forall f \in \mathcal{F}$) のとき,

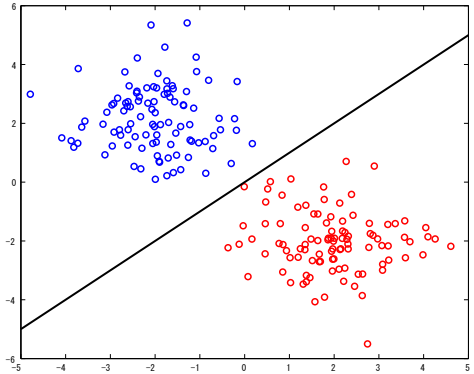
$$\begin{aligned}
L(\hat{f}) - \hat{L}(\hat{f}) &\leq \sup_{f \in \mathcal{F}} (L(f) - \hat{L}(f)) \\
&\leq 2R(\ell(\mathcal{F})) + \sqrt{\frac{t}{n}} \quad (\text{with prob. } 1 - e^{-t}) \\
&\leq 4R(\mathcal{F}) + \sqrt{\frac{t}{n}} \quad (\text{contraction ineq., Lipschitz 連続}) \\
&\lesssim \frac{1}{\sqrt{n}} \mathbb{E}_{D_n} \left[\int_0^\infty \sqrt{\log N(\mathcal{F}, \epsilon, \|\cdot\|_n)} \mathrm{d}\epsilon \right] + \sqrt{\frac{t}{n}} \quad (\text{Dudley 積分}).
\end{aligned}$$

※カバリングナンバーが小さいほどリスクは小さい→ Occam's Razor

例：線形判別関数

$$\mathcal{F} = \{f(x) = \mathrm{sign}(x^\top \beta + c) \mid \beta \in \mathbb{R}^d, c \in \mathbb{R}\}$$

$$N(\mathcal{F}, \epsilon, \|\cdot\|_n) \leq C(d+2)\left(\frac{c}{\epsilon}\right)^{2(d+1)}$$



すると，0-1 ロス ℓ に対し

$$\begin{aligned} L(\hat{f}) - \hat{L}(\hat{f}) &\leq O_p\left(\frac{1}{\sqrt{n}}\mathbb{E}_{D_n}\left[\int_0^1 \sqrt{\log N(\mathcal{F}, \epsilon, \|\cdot\|_n)}\mathrm{d}\epsilon\right]\right) \\ &\leq O_p\left(\frac{1}{\sqrt{n}}\int_0^1 C\sqrt{d\log(1/\epsilon) + \log(d)}\mathrm{d}\epsilon\right) \\ &\leq O_p\left(\sqrt{\frac{d}{n}}\right). \end{aligned}$$

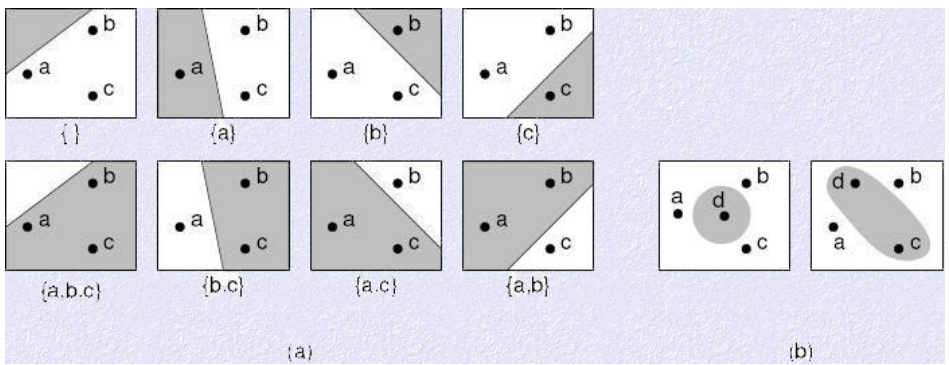
例: VC 次元

$\mathcal{F}$ は指示関数の集合: $\mathcal{F} = \{\mathbf{1}_C \mid C \in \mathcal{C}\}$.
$\mathcal{C}$ はある集合族 (例: 半空間の集合)

- 細分: $\mathcal{F}$ がある与えられた有限集合 $X_n = \{x_1, \ldots, x_n\}$ を細分する
  $\Leftrightarrow$ 任意のラベル $Y_n = \{y_1, \ldots, y_n\}$ $(y_i \in \{\pm 1\})$ に対して X_n を $\mathcal{F}$ が正しく判別できる.
- VC 次元 $V_{\mathcal{F}}$: $\mathcal{F}$ が細分できる集合が存在しない n の最小値.

$$N(\mathcal{F}, \epsilon, \|\cdot\|_n) \leq K V_{\mathcal{F}} (4e)^{V_{\mathcal{F}}} \left(\frac{1}{\epsilon}\right)^{2(V_{\mathcal{F}}-1)}$$

$$\Rightarrow \text{汎化誤差} = O_p(\sqrt{V_{\mathcal{F}}/n})$$

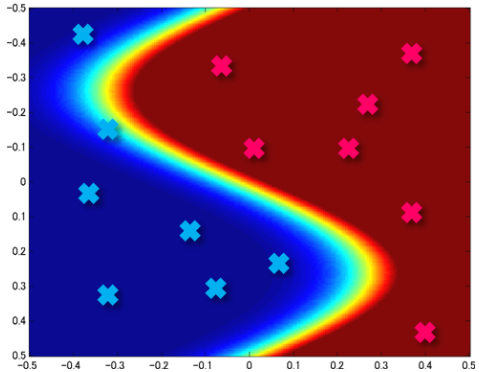


http://www.tcs.fudan.edu.cn/rudolf/Courses/Algorithms/Alg_ss_07w/Webprojects/Qinbo_diameter/e_net.htm から拝借

VC 次元が有限であることが ERM の期待リスクが収束することの必要十分条件.

例: カーネル法

$\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$



カーネル関数 k
再生核ヒルベルト空間 $\mathcal{H}$
$k(x,x) \leq 1$ $(\forall x \in \mathcal{X})$ を仮定, e.g., ガウスカーネル.

直接 Rademacher 複雑さを評価してみる.
$\sum_{i=1}^n \epsilon_i f(x_i) = \langle \sum_{i=1}^n \epsilon_i k(x_i,\cdot), f\rangle_{\mathcal{H}} \leq \|\sum_{i=1}^n \epsilon_i k(x_i,\cdot)\|_{\mathcal{H}} \|f\|_{\mathcal{H}}$
$\leq \|\sum_{i=1}^n \epsilon_i k(x_i,\cdot)\|_{\mathcal{H}}$ を使う.

$$R(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{|\sum_{i=1}^n \epsilon_i f(x_i)|}{n}\right] \leq \mathbb{E}\left[\frac{\|\sum_{i=1}^n \epsilon_i k(x_i,\cdot)\|_{\mathcal{H}}}{n}\right]$$

$$= \mathbb{E}\left[\frac{\sqrt{\sum_{i,j=1}^n \epsilon_i\epsilon_j k(x_i,x_j)}}{n}\right] \leq \frac{\sqrt{\mathbb{E}\left[\sum_{i,j=1}^n \epsilon_i\epsilon_j k(x_i,x_j)\right]}}{n} \quad \text{(Jensen)}$$

$$= \frac{\sqrt{\sum_{i=1}^n k(x_i,x_i)}}{n} \leq \frac{1}{\sqrt{n}}$$

例: ランダム行列の作用素ノルム

$A = (a_{ij})$: $p \times q$ 行列で各 a_{ij} は独立な期待値 0 かつ $|a_{ij}| \le 1$ なる確率変数.
A の作用素ノルム $\|A\| := \max_{\substack{\|z\| \le 1 \\ z \in \mathbb{R}^q}} \|Az\| = \max_{\substack{\|w\| \le 1, \|z\| \le 1 \\ w \in \mathbb{R}^p, z \in \mathbb{R}^q}} w^\top A z.$

$$\mathcal{F} = \{ f_{w,z}(a_{ij}, (i,j)) = a_{ij} w_i z_j \mid w \in \mathbb{R}^p, z \in \mathbb{R}^q \} \ \Rightarrow \ \|A\| = \sup_{f \in \mathcal{F}} \sum_{i,j} f(a_{ij}, (i,j))$$

$n = pq$ 個のサンプルがあるとみなす.
$\|f_{w,z} - f_{w',z'}\|_n^2 = \frac{1}{pq} \sum_{i,j=1}^{p,q} |a_{ij}(w_i z_j - w'_i z'_j)|^2 \le \frac{2}{pq}(\|w - w'\|^2 + \|z - z'\|^2)$

$$\therefore N(\mathcal{F}, \epsilon, \|\cdot\|_n) \begin{cases} \le C(\sqrt{pq}\epsilon)^{-(p+q)}, & (\epsilon \le 2/\sqrt{pq}), \\ = 1, & (\text{otherwise}). \end{cases}$$

$$\mathbb{E}\left[\frac{1}{pq} \sup_{w,z} w^\top A z \right] \le \frac{C}{\sqrt{pq}} \int_0^{\frac{1}{\sqrt{pq}}} \sqrt{(p+q)\log(C/\sqrt{pq}\epsilon)}\, \mathrm{d}\epsilon \le \frac{\sqrt{p+q}}{pq}$$

よって，A の作用素ノルムは $O_p(\sqrt{p+q})$.
→ 低ランク行列推定, Robust PCA, ...
詳しくは Tao (2012), Davidson and Szarek (2001) を参照.

例: Lasso の収束レート

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n \times p}$. p (次元) $\gg n$ (サンプル数).
真のベクトル $\beta^* \in \mathbb{R}^p$: 非ゼロ要素の個数がたかだか d 個 (スパース).

$$\text{モデル}: \quad Y = X\beta^* + \boldsymbol{\xi}.$$

$$\hat{\beta} \leftarrow \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_1.$$

Theorem (Lasso の収束レート (Bickel et al., 2009, Zhang, 2009))

デザイン行列が *Restricted eigenvalue condition (Bickel et al., 2009)* かつ $\max_{i,j} |X_{ij}| \leq 1$ を満たし，ノイズが $\mathbb{E}[e^{\tau \xi_i}] \leq e^{\sigma^2 \tau^2 / 2}$ $(\forall \tau > 0)$ を満たすなら，確率 $1 - \delta$ で

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C \frac{d \log(p/\delta)}{n}.$$

※次元が高くても，たかだか $\log(p)$ でしか効いてこない．実質的な次元 d が支配的.

$\log(p)$ はどこからやってきたか？

有限個の一様バウンドからやってきた.

$$\frac{1}{n}\|X\hat{\beta} - Y\|_2^2 + \lambda_n\|\hat{\beta}\|_1 \leq \frac{1}{n}\|X\beta^* - Y\|_2^2 + \lambda_n\|\beta^*\|_1$$

$$\Rightarrow \quad \frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda_n\|\hat{\beta}\|_1 \leq \frac{2}{n}\underbrace{\|X^\top \boldsymbol{\xi}\|_\infty}_{\text{これ}}\|\hat{\beta} - \beta^*\|_1 + \lambda_n\|\beta^*\|_1$$

$$\frac{1}{n}\|X^\top \boldsymbol{\xi}\|_\infty = \max_{1\leq j\leq p}|\frac{1}{n}\sum_{i=1}^{n} X_{ij}\xi_i|$$

$\log(p)$ はどこからやってきたか？

有限個の一様バウンドからやってきた.

$$\frac{1}{n}\|X\hat{\beta} - Y\|_2^2 + \lambda_n\|\hat{\beta}\|_1 \leq \frac{1}{n}\|X\beta^* - Y\|_2^2 + \lambda_n\|\beta^*\|_1$$

$$\Rightarrow \quad \frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda_n\|\hat{\beta}\|_1 \leq \frac{2}{n}\underbrace{\|X^\top \boldsymbol{\xi}\|_\infty}_{\text{これ}}\|\hat{\beta} - \beta^*\|_1 + \lambda_n\|\beta^*\|_1$$

$$\frac{1}{n}\|X^\top \boldsymbol{\xi}\|_\infty = \max_{1\leq j\leq p}|\frac{1}{n}\sum_{i=1}^n X_{ij}\xi_i|$$

Hoeffding の不等式由来の一様バウンドにより, 確率 $1-\delta$ で

$$\max_{1\leq j\leq p}|\frac{1}{n}\sum_{i=1}^n X_{ij}\xi_i| \leq \sigma\sqrt{\frac{2\log(2p/\delta)}{n}}.$$

Talagrand の concentration inequality

汎用性の高い不等式.

Theorem (Talagrand (1996b), Massart (2000), Bousquet (2002))

$\sigma := \sup_{f \in \mathcal{F}} \mathbb{E}[f(X)^2]$, $P_n f := \frac{1}{n} \sum_{i=1}^n f(x_i)$, $Pf := \mathbb{E}[f(X)]$ とする.

$$P\left[\sup_{f \in \mathcal{F}}(P_n f - Pf) \geq C\left(\mathbb{E}\left[\sup_{f \in \mathcal{F}}(P_n f - Pf)\right] + \sqrt{\frac{t}{n}}\sigma + \frac{t}{n}\right)\right] \leq e^{-t}$$

Fast learning rate を示すのに有用.

その他のトピック

- Johnson-Lindenstrauss の補題 (Johnson and Lindenstrauss, 1984, Dasgupta and Gupta, 1999)
  n 個の点 $\{x_1, \ldots, x_n\} \in \mathbb{R}^d$ を k 次元空間へ射影する. $k \geq c_\delta \log(n)$ なら, k 次元へのランダムプロジェクション $A \in \mathbb{R}^{k \times d}$ (ランダム行列) は

  $$(1-\delta)\|x_i - x_j\| \leq \|Ax_i - Ax_j\| \leq (1+\delta)\|x_i - x_j\|$$

  を高い確率で満たす.
  → restricted isometory (Baraniuk et al., 2008, Candès, 2008)
- Gaussian concentration inequality, concentration inequality on product space (Ledoux, 2001)

  $$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(x_i) \quad (x_i : \text{ガウス分布など})$$

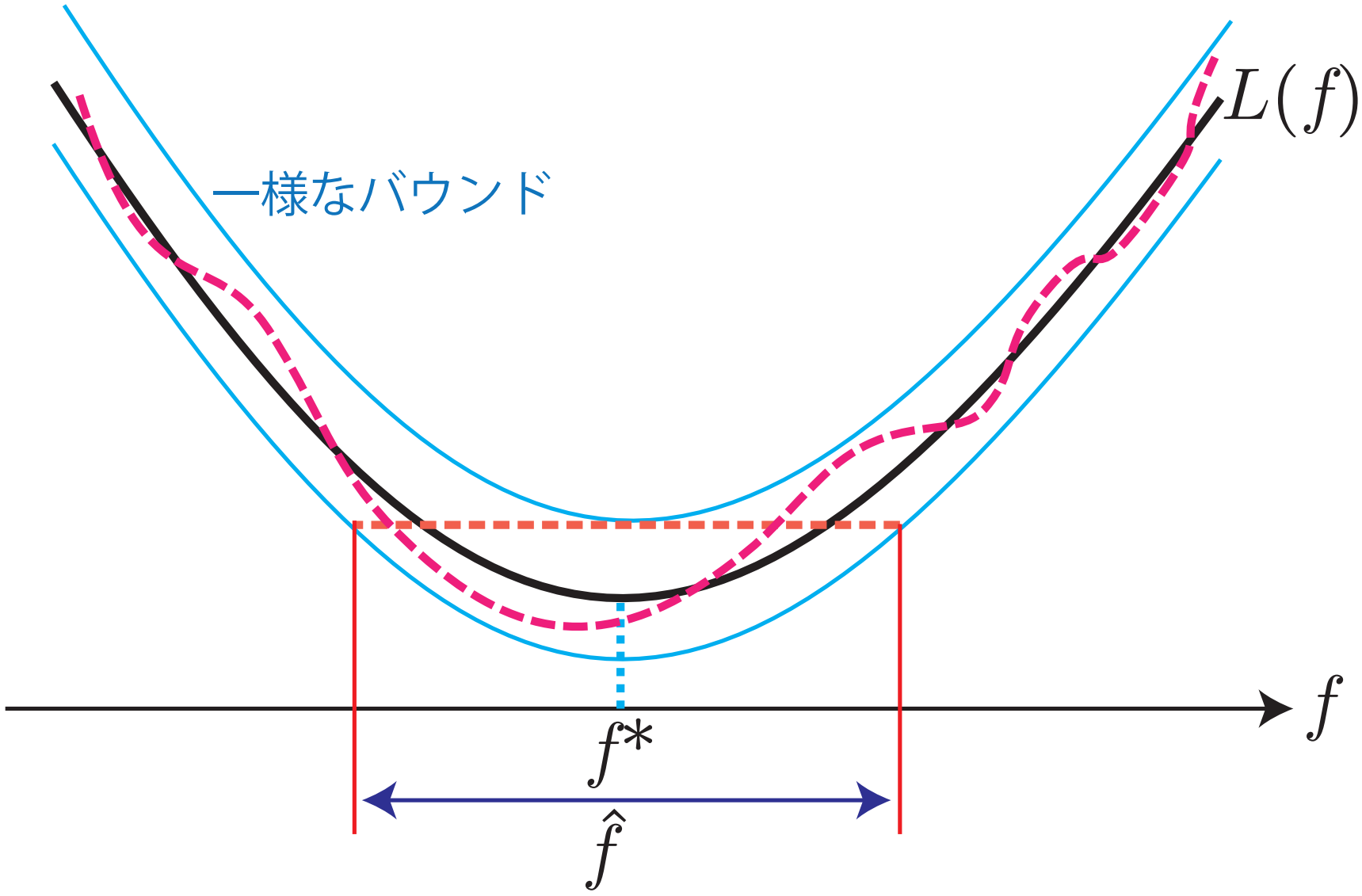
- Majorizing measure: ガウシアンプロセスにまつわる上界, 下界 (Talagrand, 2000).

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

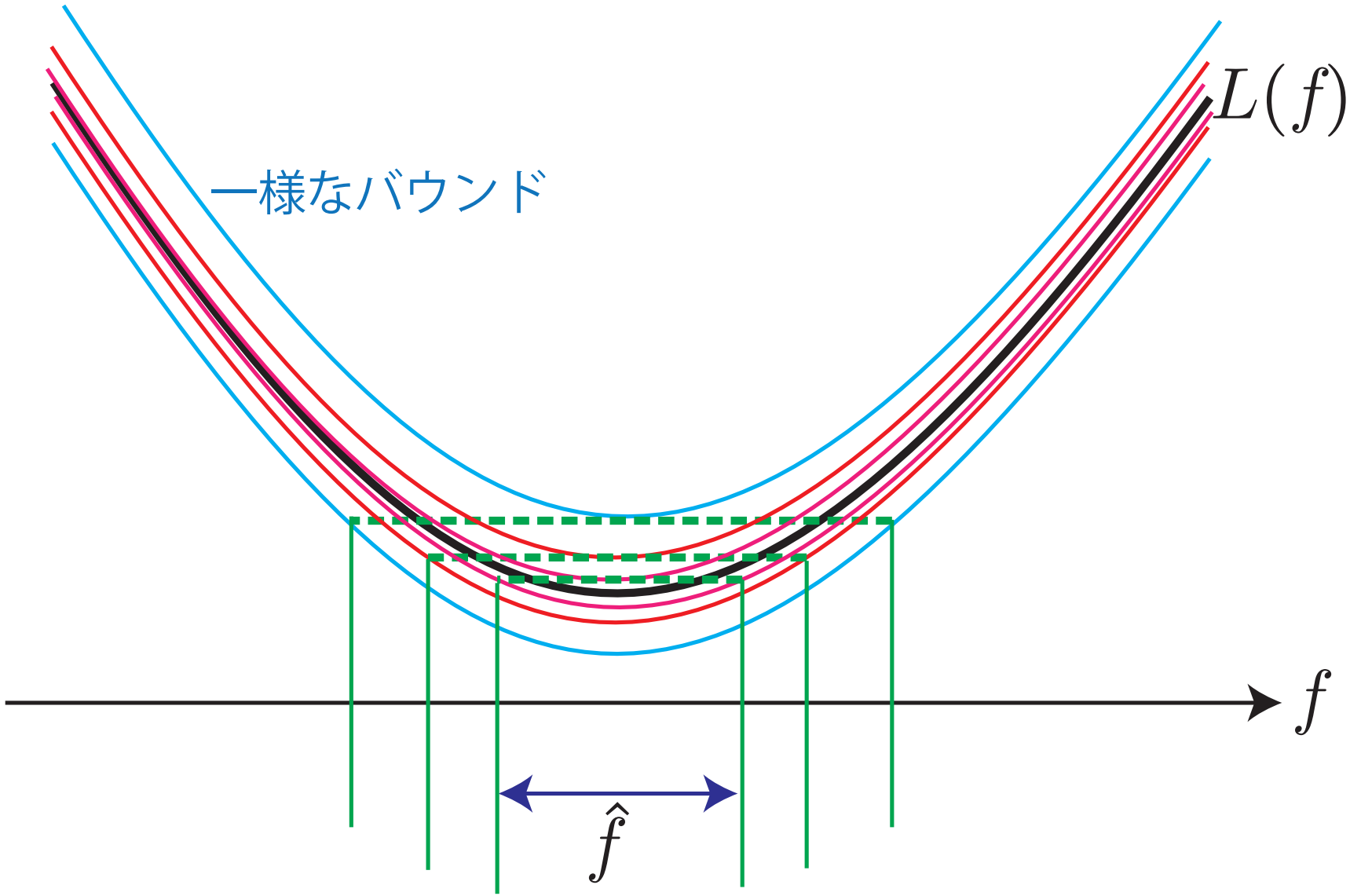
$O_p(1/\sqrt{n})$ オーダより速いレートは示せる？

ロス関数の強凸性を積極的に利用



ロスの強凸性を使うと $\hat{f}$ の存在範囲が制限される→よりきついバウンド

ロス関数の強凸性を積極的に利用



同じ論理を何度も適用させることによって $\hat{f}$ のリスクが小さいことを示す.
$\hat{f}$ が f^* に近いことを利用→ “局所” Rademacher 複雑さ

局所 Rademacher 複雑さ

局所 Rademacher 複雑さ: $R_\delta(\mathcal{F}) := R(\{f \in \mathcal{F} \mid \mathbb{E}[(f - f^*)^2] \leq \delta\})$.

次の条件を仮定してみる.

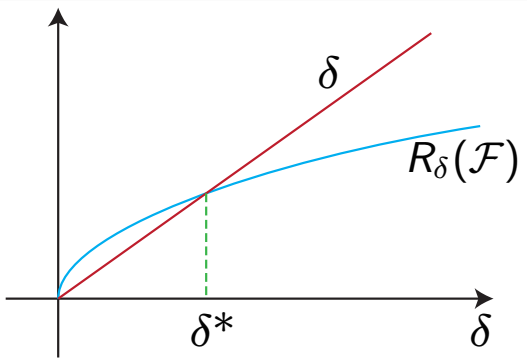
- $\mathcal{F}$ は 1 で上から抑えられている: $\|f\|_\infty \leq 1$ $(\forall f \in \mathcal{F})$.
- ℓ は Lipschitz 連続かつ強凸:
  $\mathbb{E}[\ell(Y, f(X))] - \mathbb{E}[\ell(Y, f^*(X))] \geq B\mathbb{E}[(f - f^*)^2]$ $(\forall f \in \mathcal{F})$.

Theorem (Fast learning rate (Bartlett et al., 2005))

$\delta^* = \inf\{\delta \mid \delta \geq R_\delta(\mathcal{F})\}$ とすると，確率 $1 - e^{-t}$ で

$$L(\hat{f}) - L(f^*) \leq C\left(\delta^* + \frac{t}{n}\right).$$

$\delta^* \leq R(\mathcal{F})$ は常に成り立つ (右図参照).
これを Fast learning rate と言う.



Fast learning rate の例

$\boxed{\log N(\mathcal{F}, \epsilon, \|\cdot\|_n) \leq C\epsilon^{-2\rho}}$ のとき，

$$R_\delta(\mathcal{F}) \leq C\left(\frac{\delta^{\frac{1-\rho}{2}}}{\sqrt{n}} \vee n^{-\frac{1}{1+\rho}}\right),$$

が示され，δ^* の定義から確率 $1 - e^{-t}$ で次が成り立つ:

$$L(\hat{f}) - L(f^*) \leq C\left(n^{-\frac{1}{1+\rho}} + \frac{t}{n}\right).$$

※ $1/\sqrt{n}$ よりタイト！

参考文献

- 局所 Rademacher 複雑さの一般論: Bartlett et al. (2005), Koltchinskii (2006)
- 判別問題, Tsybakov の条件: Tsybakov (2004), Bartlett et al. (2006)
- カーネル法における fast learning rate: Steinwart and Christmann (2008)
- Peeling device: van de Geer (2000)

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

最適性

ある学習方法が「最適」とは？

どの学習方法もデータの分布に応じて得意不得意がある.
「この場合はうまくいくがこの場合はうまくいかない」

主な最適性の規準

- 許容性
  常に性能を改善させる方法が他にない.
- **minimax** 最適性
  一番不得意な場面でのリスクが最小.

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

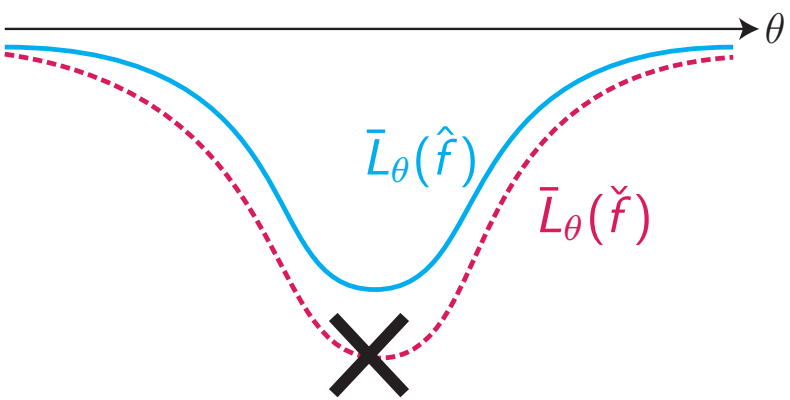
許容性

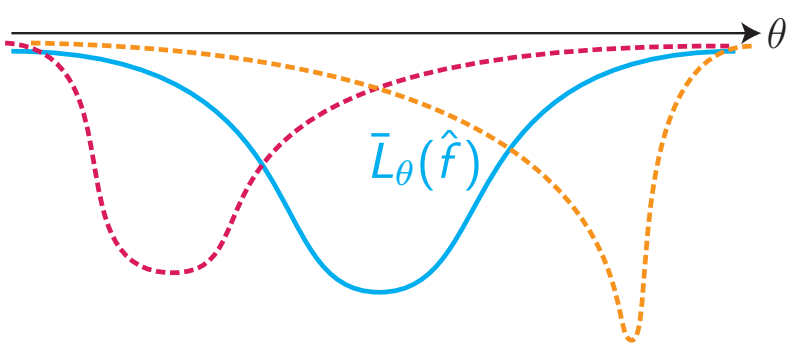
分布のモデル: $\{P_\theta | \theta \in \Theta\}$
P_θ における推定量 $\check{f}$ のリスクの期待値:

$$\bar{L}_\theta(\check{f}) := \mathbb{E}_{D_n \sim P_\theta}[\mathbb{E}_{(X,Y)\sim P_\theta}[\ell(Y, \check{f}(X))]]$$

Definition (許容性)

$\hat{f}$ が許容的 (admissible)
$\Leftrightarrow$ $\bar{L}_\theta(\check{f}) \leq \bar{L}_\theta(\hat{f})$ $(\forall \theta \in \Theta)$ かつ, ある $\theta' \in \Theta$ で $\bar{L}_{\theta'}(\check{f}) < \bar{L}_{\theta'}(\hat{f})$ なる推定量 $\check{f}$ が存在しない.





例

簡単のためサンプル $D_n = \{(x_1, \ldots, x_n)\} \sim P_\theta^n$ から P_θ $(\theta \in \Theta)$ を推定する問題を考える．

- 一点賭け: ある θ_0 を常に用いる．その θ_0 に対する当てはまりは最良だが他の θ には悪い．
- ベイズ推定量: 事前分布 $\pi(\theta)$, リスク $L(\theta_0, \hat{P})$

$$\hat{P} = \underset{\hat{P}:\text{推定量}}{\arg\min} \int \mathbb{E}_{D_n \sim P_{\theta_0}}[L(\theta_0, \hat{P})]\pi(\theta_0)\mathrm{d}\theta_0.$$

  - 二乗リスク $L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$: $\hat{\theta} = \int \theta\pi(\theta|D_n)\mathrm{d}\theta$ (事後平均)
  - KL-リスク $L(\theta, \hat{P}) = \mathrm{KL}(P_\theta||\hat{P})$: $\hat{P} = \int P(\cdot|\theta)\pi(\theta|D_n)\mathrm{d}\theta$ (ベイズ予測分布)

  ベイズ推定量の定義より，常にリスク $L(\theta, \hat{P})$ を改善する推定量は存在しない．

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

minimax 最適性

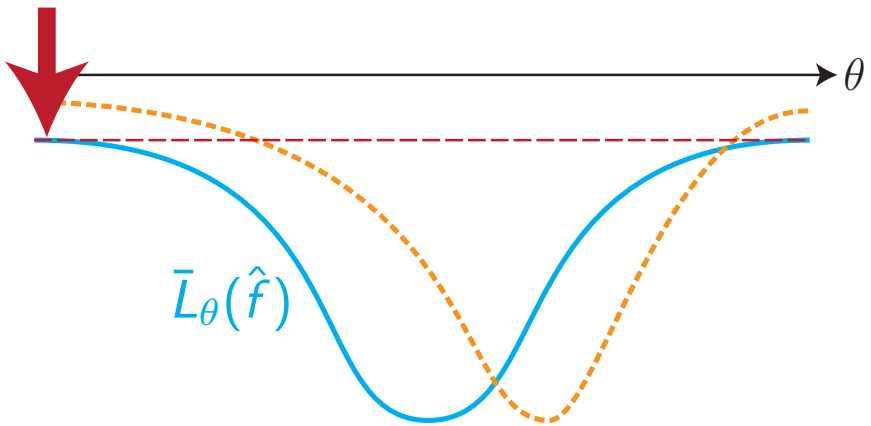
Definition (minimax 最適性)

$$\hat{f} \text{ が minimax 最適}$$
$$\Leftrightarrow \max_{\theta\in\Theta} \bar{L}_\theta(\hat{f}) = \min_{\check{f}:\text{推定量}} \max_{\theta\in\Theta} \bar{L}_\theta(\check{f}).$$

学習理論では定数倍を許すことが多い: $\exists C$ で

$$\max_{\theta\in\Theta} \bar{L}_\theta(\hat{f}) \le C \min_{\check{f}:\text{推定量}} \max_{\theta\in\Theta} \bar{L}_\theta(\check{f}) \quad (\forall n).$$

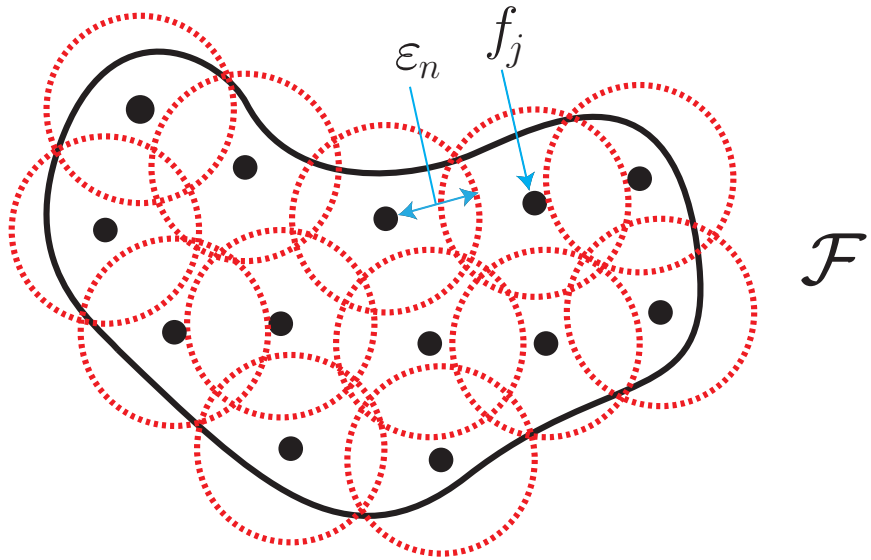
そういう意味で「minimax レートを達成する」と言ったりする.



minimax レートを求める方法

Introduction to nonparametric estimation (Tsybakov, 2008) に詳しい記述.

$\mathcal{F}$ を有限個の元で代表させ，そのうち一つ最良なものを選ぶ問題を考える.
(もとの問題より簡単→リスクの下限を与える)
$\{f_1, \ldots, f_{M_n}\} \subseteq \mathcal{F}$



個数 M_n と誤差 ε_n のトレードオフ: M_n が小さい方が最適な元を選ぶのが簡単になるが誤差 ε_n が大きくなる.

cf. Fano の不等式, Assouad の補題.

ノイズがガウスのノンパラメトリック回帰 minimax レート

Yang and Barron (1999) の情報理論的方法
(Raskutti et al. (2012b) でより整理された形を利用).

- $\varepsilon_n > 0$: 任意
- $\log(N) = \log(N(\mathcal{F}, \varepsilon_n, \|\cdot\|_{L_2(P_X)}))$

に対して,

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{D_n}[\|\hat{f} - f^*\|^2_{L_2(P_X)}] \geq \frac{\varepsilon_n^2}{2}\left(1 - \frac{\log(N) + \frac{n}{2\sigma^2}\varepsilon_n^2 + \log(2)}{8\log(N)}\right).$$

$\Rightarrow$ $\epsilon_n^2 \simeq \frac{\log(N)}{n}$ となるように ϵ_n を選べば,

$$\text{minimax レート} \gtrsim \epsilon_n^2.$$

スパース推定の minimax レート

Theorem (Raskutti and Wainwright (2011))

ある条件のもと，確率 1/2 以上で，

$$\min_{\hat{\beta}:\text{推定量}} \max_{\beta^*:d\text{-スパース}} \|\hat{\beta} - \beta^*\|^2 \geq C \frac{d \log(p/d)}{n}.$$

Lasso は minimax レートを達成する ($\frac{d \log(d)}{n}$ の項を除いて).

この結果を Multiple Kernel Learning に拡張した結果: Raskutti et al. (2012a), Suzuki and Sugiyama (2012).

Outline

1. 機械学習概要
2. 学習理論概要
3. 一様バウンド
   - 基本的な不等式
   - Rademacher 複雑さと Dudley 積分
   - 局所 Rademacher 複雑さ
4. 最適性
   - 許容性
   - minimax 最適性
5. ベイズの学習理論

ベイズの学習理論

ノンパラベイズの統計的性質

- 教科書: Ghosh and Ramamoorthi (2003), *Bayesian Nonparametrics*. Springer, 2003.
- 収束レート
  - 一般論: Ghosal et al. (2000)
  - Dirichlet mixture: Ghosal and van der Vaart (2007)
  - Gaussian process: van der Vaart and van Zanten (2008a,b, 2011).

ベイズの学習理論

ノンパラベイズの統計的性質

- 教科書: Ghosh and Ramamoorthi (2003), *Bayesian Nonparametrics*. Springer, 2003.
- 収束レート
  - 一般論: Ghosal et al. (2000)
  - Dirichlet mixture: Ghosal and van der Vaart (2007)
  - Gaussian process: van der Vaart and van Zanten (2008a,b, 2011).

PAC-Bayes

$$L(\hat{f}_\pi) \leq \inf_\rho \left\{ \int L(f)\rho(\mathrm{d}f) + 2\left[\frac{\lambda C^2}{n} + \frac{\mathrm{KL}(\rho||\pi) + \log\frac{2}{\epsilon}}{\lambda} \right] \right\}$$

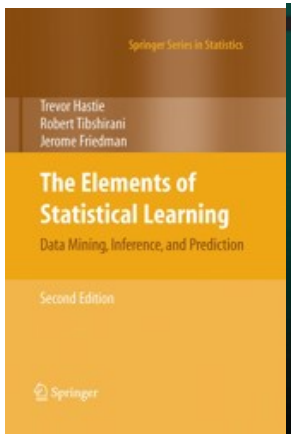
(Catoni, 2007)

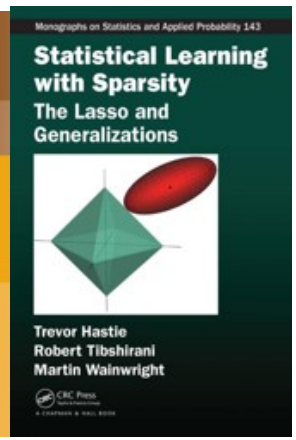
- 元論文: McAllester (1998, 1999)
- オラクル不等式: Catoni (2004, 2007)
- スパース推定への応用: Dalalyan and Tsybakov (2008), Alquier and Lounici (2011), Suzuki (2012)

スパース推定文献情報

- R の glmnet, LiblineaR およぼ，python の scikit-learn.
- SPAMS: 最適化ソルバー, C++で実装． matlab, R, python インターフェイスが付いている．
  `http://spams-devel.gforge.inria.fr/`

- 冨岡亮太著『スパース性に基づく機械学習』，講談社.
- 鈴木大慈著『確率的最適化』，講談社.
- Hastie, Tibshirani, Friedman: The Elements of Statistical Learning.
- Hastie, Tibshirani, Wainwright: Statistical Learning with Sparsity: The Lasso and Generalizations.





まとめ

一様バウンドが重要

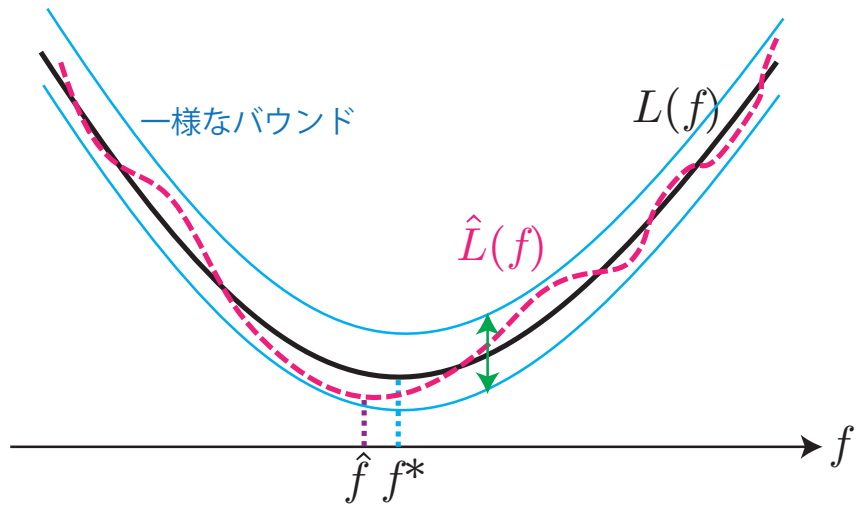
$$\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) - \mathbb{E}[\ell(Y, f(X))] \right\}$$

- Rademacher 複雑さ
- カバリングナンバー

仮説集合が単純であればあるほど，速い収束.

最適性規準

- 許容性
- minimax 最適性



P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. Electronic Journal of Statistics, 5:127–145, 2011.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. Constructive Approximation, 28(3):253–263, 2008.

P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. The Annals of Statistics, 33:1487–1537, 2005.

P. Bartlett, M. Jordan, and D. McAuliffe. Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101:138–156, 2006.

C. Bennett and R. Sharpley. Interpolation of Operators. Academic Press, Boston, 1988.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 37(4):1705–1732, 2009.

S. Boucheron, G. Lugosi, and P. Massart. Concentration Inequalities: A Nonasymptotic Theory of Independence. OUP Oxford, 2013. ISBN 9780199535255. URL https://books.google.co.jp/books?id=koNqWRluhP0C.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. C. R. Acad. Sci. Paris Ser. I Math., 334:495–500, 2002.

E. Candès. The restricted isometry property and its implications for compressed sensing. Compte Rendus de l'Academie des Sciences, Paris, Serie I, 346: 589–592, 2008.

F. P. Cantelli. Sulla determinazione empirica della leggi di probabilità. G. Inst. Ital. Attuari, 4:221–424, 1933.

O. Catoni. Statistical Learning Theory and Stochastic Optimization. Lecture Notes in Mathematics. Springer, 2004. Saint-Flour Summer School on Probability Theory 2001.

O. Catoni. PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning). Lecture Notes in Mathematics. IMS, 2007.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. Machine Learning, 72:39–61, 2008.

S. Dasgupta and A. Gupta. An elementary proof of the johnson-lindenstrauss lemma. Technical Report 99–006, U.C. Berkeley, 1999.

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces, volume 1, chapter 8, pages 317–366. North Holland, 2001.

M. Donsker. Justification and extension of doob's heuristic approach to the kolmogorov-smirnov theorems. Annals of Mathematical Statistics, 23:277–281, 1952.

R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. J. Functional Analysis, 1:290–330, 1967.

M. Eberts and I. Steinwart. Optimal learning rates for least squares svms using gaussian kernels. In Advances in Neural Information Processing Systems 25, 2012.

S. Ghosal and A. W. van der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. The Annals of Statistics, 35(2):697–723, 2007.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. The Annals of Statistics, 28(2):500–531, 2000.

J. Ghosh and R. Ramamoorthi. Bayesian Nonparametrics. Springer, 2003.

V. I. Glivenko. Sulla determinazione empirica di probabilità. G. Inst. Ital. Attuari, 4:92–99, 1933.

W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In Conference in Modern Analysis and Probability, volume 26, pages 186–206, 1984.

W. B. Johnson, J. Lindenstrauss, and G. Schechtman. Extensions of lipschitz maps into banach spaces. Israel Journal of Mathematics, 54(2):129–138, 1986.

A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. G. Inst. Ital. Attuari, 4:83–91, 1933.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. The Annals of Statistics, 34:2593–2656, 2006.

M. Ledoux. The concentration of measure phenomenon. American Mathematical Society, 2001.

P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. The Annals of Probability, 28(2):863–884, 2000.

D. McAllester. Some PAC-Bayesian theorems. In the Anual Conference on Computational Learning Theory, pages 230–234, 1998.

D. McAllester. PAC-Bayesian model averaging. In the Anual Conference on Computational Learning Theory, pages 164–170, 1999.

S. Mukherjee, R. Rifkin, and T. Poggio. Regression and classification with regularization. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, editors, Lecture Notes in Statistics: Nonlinear Estimation and Classification, pages 107–124. Springer-Verlag, New York, 2002.

G. Raskutti and M. J. Wainwright. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Transactions on Information Theory, 57(10):6976–6994, 2011.

G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Journal of Machine Learning Research, 13:389–427, 2012a.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. The Journal of Machine Learning Research, 13(1):389–427, 2012b.

M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. IEEE Transactions of Information Theory, 39, 2013.

I. Steinwart and A. Christmann. Support Vector Machines. Springer, 2008.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In Proceedings of the Annual Conference on Learning Theory, pages 79–93, 2009.

T. Suzuki. Pac-bayesian bound for gaussian process regression and multiple kernel additive model. In JMLR Workshop and Conference Proceedings, volume 23, pages 8.1–8.20, 2012. Conference on Learning Theory (COLT2012).

T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. In JMLR Workshop and Conference Proceedings 22, pages 1152–1183, 2012. Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS2012).

M. Talagrand. New concentration inequalities in product spaces. Invent. Math., 126:505–563, 1996a.

M. Talagrand. New concentration inequalities in product spaces. Inventiones Mathematicae, 126:505–563, 1996b.

M. Talagrand. The generic chaining. Springer, 2000.

T. Tao. Topics in random matrix theory. American Mathematical Society, 2012.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. Annals of Statistics, 35:135–166, 2004.

A. B. Tsybakov. Introduction to nonparametric estimation. Springer Series in Statistics. Springer, 2008.

S. van de Geer. Empirical Processes in M-Estimation. Cambridge University Press, 2000.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. The Annals of Statistics, 36(3): 1435–1463, 2008a.

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, 3:200–222, 2008b. IMS Collections.

A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric gaussian process methods. Journal of Machine Learning Research, 12: 2095–2119, 2011.

V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Soviet Math. Dokl., 9:915–918, 1968.

M. Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. The Annals of Statistics, 27(5):1564–1599, 1999.

T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. The Annals of Statistics, 37(5):2109–2144, 2009.

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. Electronic Journal of Statistics, 5:127–145, 2011.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. Constructive Approximation, 28(3):253–263, 2008.

P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. The Annals of Statistics, 33:1487–1537, 2005.

P. Bartlett, M. Jordan, and D. McAuliffe. Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101:138–156, 2006.

C. Bennett and R. Sharpley. Interpolation of Operators. Academic Press, Boston, 1988.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 37(4):1705–1732, 2009.

S. Boucheron, G. Lugosi, and P. Massart. Concentration Inequalities: A Nonasymptotic Theory of Independence. OUP Oxford, 2013. ISBN 9780199535255. URL https://books.google.co.jp/books?id=koNqWRluhP0C.

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical process. C. R. Acad. Sci. Paris Ser. I Math., 334:495–500, 2002.

E. Candès. The restricted isometry property and its implications for compressed sensing. Compte Rendus de l'Academie des Sciences, Paris, Serie I, 346: 589–592, 2008.

F. P. Cantelli. Sulla determinazione empirica della leggi di probabilità. G. Inst. Ital. Attuari, 4:221–424, 1933.

O. Catoni. Statistical Learning Theory and Stochastic Optimization. Lecture Notes in Mathematics. Springer, 2004. Saint-Flour Summer School on Probability Theory 2001.

O. Catoni. PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning). Lecture Notes in Mathematics. IMS, 2007.

A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. Machine Learning, 72:39–61, 2008.

S. Dasgupta and A. Gupta. An elementary proof of the johnson-lindenstrauss lemma. Technical Report 99–006, U.C. Berkeley, 1999.

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces, volume 1, chapter 8, pages 317–366. North Holland, 2001.

M. Donsker. Justification and extension of doob's heuristic approach to the kolmogorov-smirnov theorems. Annals of Mathematical Statistics, 23:277–281, 1952.

R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. J. Functional Analysis, 1:290–330, 1967.

M. Eberts and I. Steinwart. Optimal learning rates for least squares svms using gaussian kernels. In Advances in Neural Information Processing Systems 25, 2012.

S. Ghosal and A. W. van der Vaart. Posterior convergence rates of dirichlet mixtures at smooth densities. The Annals of Statistics, 35(2):697–723, 2007.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. The Annals of Statistics, 28(2):500–531, 2000.

J. Ghosh and R. Ramamoorthi. Bayesian Nonparametrics. Springer, 2003.

V. I. Glivenko. Sulla determinazione empirica di probabilità. G. Inst. Ital. Attuari, 4:92–99, 1933.

W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In Conference in Modern Analysis and Probability, volume 26, pages 186–206, 1984.

W. B. Johnson, J. Lindenstrauss, and G. Schechtman. Extensions of lipschitz maps into banach spaces. Israel Journal of Mathematics, 54(2):129–138, 1986.

A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. G. Inst. Ital. Attuari, 4:83–91, 1933.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. The Annals of Statistics, 34:2593–2656, 2006.

M. Ledoux. The concentration of measure phenomenon. American Mathematical Society, 2001.

P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. The Annals of Probability, 28(2):863–884, 2000.

D. McAllester. Some PAC-Bayesian theorems. In the Anual Conference on Computational Learning Theory, pages 230–234, 1998.

D. McAllester. PAC-Bayesian model averaging. In the Anual Conference on Computational Learning Theory, pages 164–170, 1999.

S. Mukherjee, R. Rifkin, and T. Poggio. Regression and classification with regularization. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, editors, Lecture Notes in Statistics: Nonlinear Estimation and Classification, pages 107–124. Springer-Verlag, New York, 2002.

G. Raskutti and M. J. Wainwright. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE Transactions on Information Theory, 57(10):6976–6994, 2011.

G. Raskutti, M. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Journal of Machine Learning Research, 13:389–427, 2012a.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. The Journal of Machine Learning Research, 13(1):389–427, 2012b.

M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. IEEE Transactions of Information Theory, 39, 2013.

I. Steinwart and A. Christmann. Support Vector Machines. Springer, 2008.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In Proceedings of the Annual Conference on Learning Theory, pages 79–93, 2009.

T. Suzuki. Pac-bayesian bound for gaussian process regression and multiple kernel additive model. In JMLR Workshop and Conference Proceedings, volume 23, pages 8.1–8.20, 2012. Conference on Learning Theory (COLT2012).

T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. In JMLR Workshop and Conference Proceedings 22, pages 1152–1183, 2012. Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS2012).

M. Talagrand. New concentration inequalities in product spaces. Invent. Math., 126:505–563, 1996a.

M. Talagrand. New concentration inequalities in product spaces. Inventiones Mathematicae, 126:505–563, 1996b.

M. Talagrand. The generic chaining. Springer, 2000.

T. Tao. Topics in random matrix theory. American Mathematical Society, 2012.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. Annals of Statistics, 35:135–166, 2004.

A. B. Tsybakov. Introduction to nonparametric estimation. Springer Series in Statistics. Springer, 2008.

S. van de Geer. Empirical Processes in M-Estimation. Cambridge University Press, 2000.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. The Annals of Statistics, 36(3): 1435–1463, 2008a.

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, 3:200–222, 2008b. IMS Collections.

A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric gaussian process methods. Journal of Machine Learning Research, 12: 2095–2119, 2011.

V. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Soviet Math. Dokl., 9:915–918, 1968.

M. Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. The Annals of Statistics, 27(5):1564–1599, 1999.

T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. The Annals of Statistics, 37(5):2109–2144, 2009.