

統計的学習理論チュートリアル: 基礎から応用まで

† 鈴木 大慈

† 東京大学
情報理工学研究科
数理情報学専攻

IBIS 2012@筑波大学東京キャンパス文京校舎
2012 年 11 月 7 日

構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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

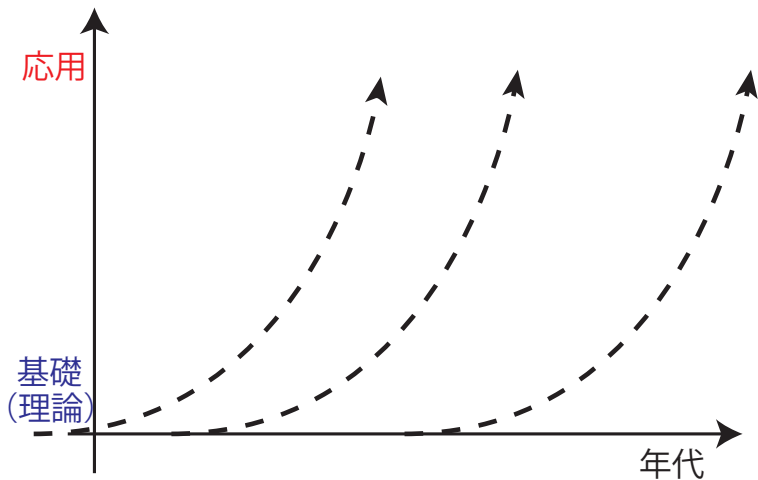
構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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

そもそも理論は必要？

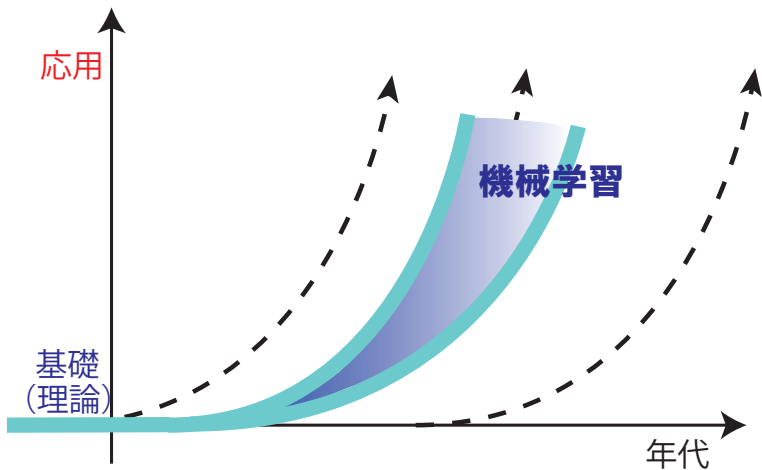
- VC 次元とか再生核の理論とか知らなくても SVM は動かせる.
- 測度論とか知らなくてもノンパラベイズの実装はできる.
- そもそも理論家は 役に立たない 難しい話をこねくり回しているだけでは？

基礎研究と応用の関係



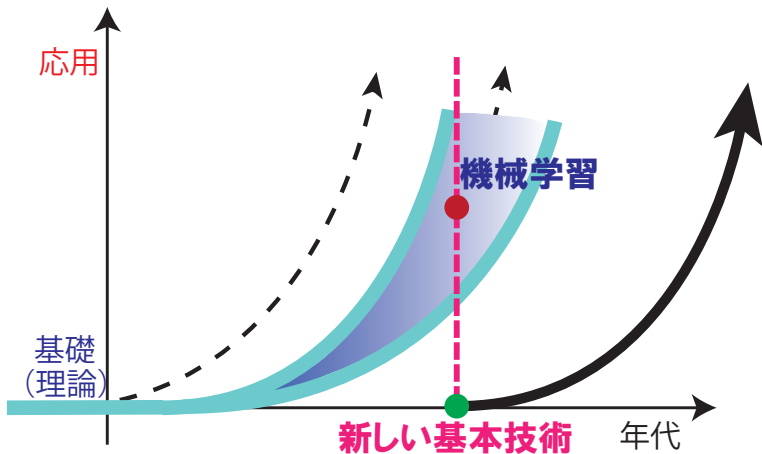
常に数多の基礎研究が応用化されている

基礎研究と応用の関係



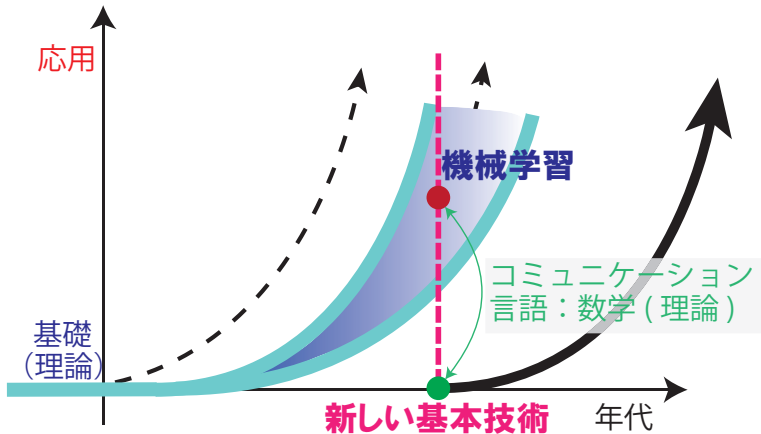
機械学習業界の変遷

基礎研究と応用の関係



未来の応用につながる基礎研究

基礎研究と応用の関係



時に異なるレベル間のコミュニケーションが新しい発見を導く

歴史にみる成功例

- SVM (Vapnik, 1998, Cortes and Vapnik, 1995): VC 次元
- AdaBoost (Freund and Schapire, 1995): 弱学習機による学習可能性
- Dirichlet process (Ferguson, 1973): 確率論, 測度論

歴史にみる成功例

- SVM (Vapnik, 1998, Cortes and Vapnik, 1995): VC 次元
- AdaBoost (Freund and Schapire, 1995): 弱学習機による学習可能性
- Dirichlet process (Ferguson, 1973): 確率論, 測度論
- Lasso (Tibshirani, 1996)
- AIC (Akaike, 1974)
- 圧縮センシング (Candès, Tao and Donoho, 2004)
- などなど

大事なものは本質の理解
→新しい手法
(そのための本セッション！)

より直接的な効能

学習理論を知ることのより直接的な有用性

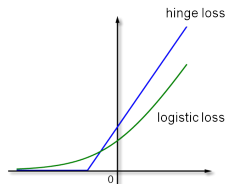
- ① **手法の意味:** そもそも何をやっている手法なのか→正しい使い方
- ② **手法の正当性:** ちゃんとした解が得られるか（「本当に収束するのか」）
- ③ **手法の最適性:** ある尺度に関して最適な手法か→安心して使える

①手法の意味: 例. ロス関数の選択

二値判別:

ヒンジロスとロジスティックロス, どちらを使うべき?

$$\min_f \frac{1}{n} \sum_{i=1}^n \phi(-y_i f(x_i)) \quad (y_i \in \{\pm 1\})$$



① 両者とも判別誤差を最小化

ϕ が凸の時

「 ϕ は判別一致性をもつ $\Leftrightarrow \phi$ が原点で微分可能かつ $\phi'(0) > 0$ 」

(Bartlett et al., 2006)

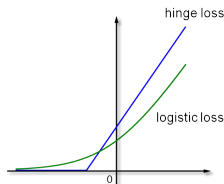
判別一致性: 期待リスク最小化関数 ($\arg \min_f E[\phi(-Yf(X))]$) が Bayes 最適.

①手法の意味: 例. ロス関数の選択

二値判別:

ヒンジロスとロジスティックロス, どちらを使うべき?

$$\min_f \frac{1}{n} \sum_{i=1}^n \phi(-y_i f(x_i)) \quad (y_i \in \{\pm 1\})$$



① 両者とも判別誤差を最小化

ϕ が凸の時

「 ϕ は判別一致性をもつ $\Leftrightarrow \phi$ が原点で微分可能かつ $\phi'(0) > 0$ 」

(Bartlett et al., 2006)

判別一致性: 期待リスク最小化関数 ($\arg \min_f E[\phi(-Yf(X))]$) が Bayes 最適.

② サポートベクターの数 vs 条件付き確率 $p(Y|X)$ の推定能力

- ヒンジ: スパースな解 (条件付き確率は全く推定できない)
- ロジスティック: 条件付き確率が求まる (一方, 全サンプルがサポートベクター)

「サポートベクターの数と条件付き確率の推定能力との間にはトレードオフがある. 両者を完全に両立させることはできない。」

(Bartlett and Tewari, 2007)

②手法の正当性・③手法の最適性

2 手法の正当性

例: 一致性

$$\hat{f} \text{ (推定量)} \xrightarrow{P} f^* \text{ (真の関数)}$$

3 手法の最適性

収束するとしてその速さは最適？

- 許容性
- minimax 性

今日はここらへんを中心に話します.

統計的學習理論

構成

- 1 はじめに: 理論の役割
- 2 統計的学習理論と経験過程
- 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

統計的学習理論の立ち位置

厳密評価

パラメトリック

正則

Fisher情報量が正則, 微分可能
漸近理論
漸近正規性→検定論・情報量規準
普遍性
サンプル数 $\gg$ 次元

非正則

Fisher情報量が退化, 微分不可能
•ニューラルネット
•隠れ状態モデル
漸近正規性×
錐の幾何学, 代数幾何...
サンプル数 $\gg$ 次元

ノンパラ

カーネル密度推定, Nadaraya-Watson推定量, スプライン回帰...
漸近正規性・漸近展開

厳密評価の難しい状況:

微分不可能(ヒンジロス, 0-1ロス), 高次元(スパース)

分布の仮定は最小限

リスクの上界および裾確率の評価

→不等式評価が主

パラメトリック・ノンパラメトリック問わない(一般性)

統計的学習理論の立ち位置

厳密評価

パラメトリック

伝統的数理統計

Fisher情報量が正則, 微分可能
漸近理論
漸近正規性→検定論・情報量規準
普遍性
サンプル数 $\gg$ 次元

非正則

Fisher情報量が退化, 微分不可能
•ニューラルネット
•隠れ状態モデル
漸近正規性×
錐の幾何学, 代数幾何...
サンプル数 $\gg$ 次元

ノンパラ

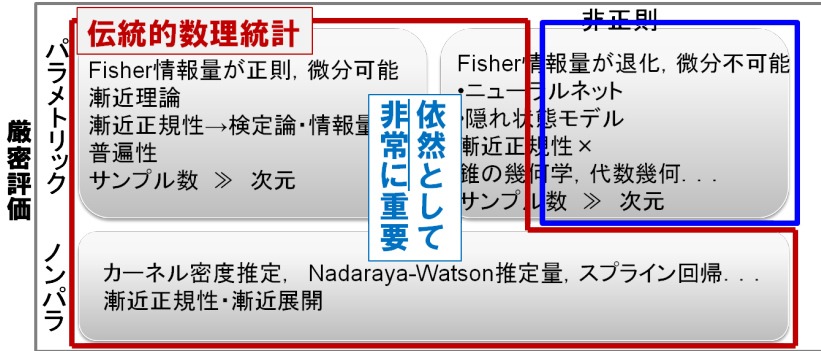
カーネル密度推定, Nadaraya-Watson推定量, スプライン回帰...
漸近正規性・漸近展開

学習理論

厳密評価の難しい状況:
微分不可能(ヒンジロス, 0-1ロス), 高次元(スパース)
分布の仮定は最小限
リスクの上界および裾確率の評価
→不等式評価が主
パラメトリック・ノンパラメトリック問わない(一般性)

※実際のところ, 境界は非常に曖昧.

統計的学習理論の立ち位置



学習理論

厳密評価の難しい状況:
微分不可能 (ヒンジロス, 0-1ロス), 高次元 (スパース)
分布の仮定は最小限
リスクの上界および裾確率の評価
→不等式評価が主
パラメトリック・ノンパラメトリック問わない (一般性)

今日の内容

※実際のところ, 境界は非常に曖昧.

(今回お話しする) 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

構成

- 1 はじめに: 理論の役割
- 2 統計的学習理論と経験過程
- 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: $D_n = \{(x_1, y_1), \dots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ 入力と出力の i.i.d. 系列

ロス関数: $\ell(\cdot, \cdot) : \mathcal{Y} \times \mathbb{R} \rightarrow \mathbb{R}_+$ 間違いへのペナルティ

仮説集合 (モデル): $\mathcal{F}$ $\mathcal{X} \rightarrow \mathbb{R}$ 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):

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

- 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):

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

正則化付き経験誤差最小化 (RERM):

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

- 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): ☆

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

正則化付き経験誤差最小化 (RERM):

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

- 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

ほとんどのバウンズの導出は次の式から始まる:

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

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

ほとんどのバウンズの導出は次の式から始まる:

$$\begin{aligned} \hat{L}(\hat{f}) &\leq \hat{L}(f^*) \quad (\because \text{経験誤差最小化}) \\ \Rightarrow \underbrace{L(\hat{f}) - L(f^*)}_{\text{汎化誤差}} &\leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{?} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n}) \text{ (後述)}} \end{aligned}$$

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

ほとんどのバウンズの導出は次の式から始まる:

$$\begin{aligned} \hat{L}(\hat{f}) &\leq \hat{L}(f^*) && (\because \text{経験誤差最小化}) \\ \Rightarrow \underbrace{L(\hat{f}) - L(f^*)}_{\text{汎化誤差}} &\leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{?} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n}) \text{ (後述)}} \end{aligned}$$

安易な解析

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

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

ほとんどのバウンズの導出は次の式から始まる:

$$\begin{aligned} \hat{L}(\hat{f}) &\leq \hat{L}(f^*) \quad (\because \text{経験誤差最小化}) \\ \Rightarrow \underbrace{L(\hat{f}) - L(f^*)}_{\text{汎化誤差}} &\leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{?} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n}) \text{ (後述)}} \end{aligned}$$

安易な解析

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

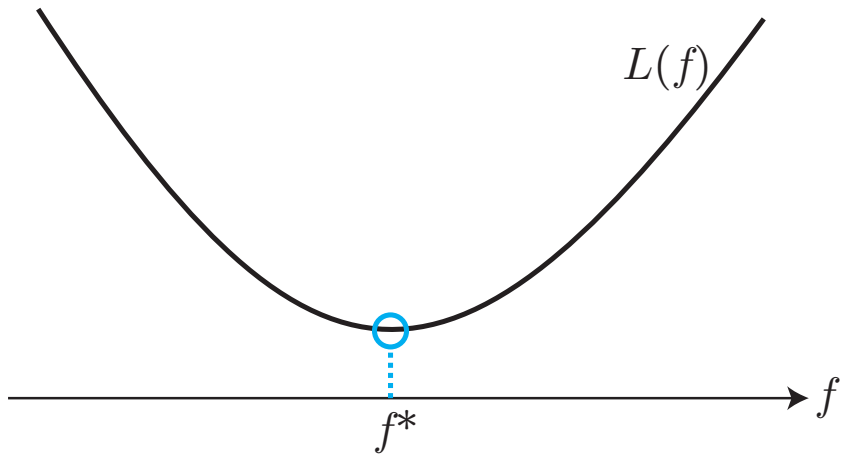
楽勝 !!!

ダメです

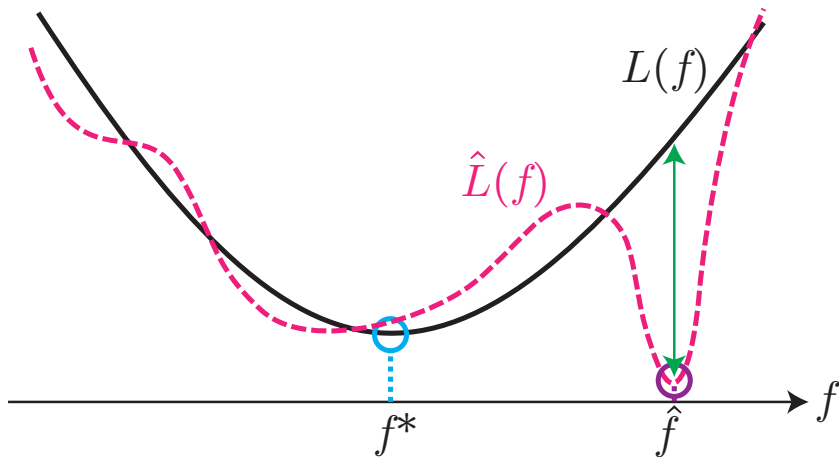
$\hat{f}$ と教師データは独立ではない

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？

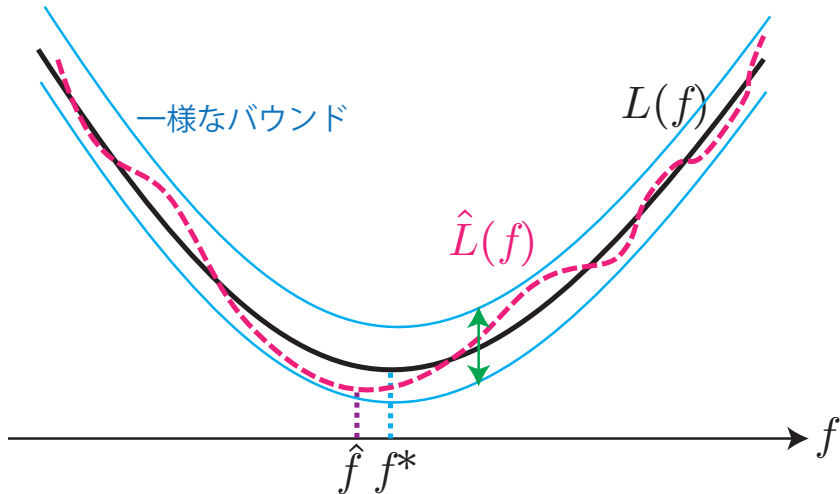


なにが問題か？



“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

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

一様に リスクを抑えることが重要

構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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 ($i = 1, \dots, n$): 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 ($i = 1, \dots, n$): 独立で (同一とは限らない) 期待値 0 の確率変数 s.t.
 $E[Z_i^2] = \sigma_i^2$, $|Z_i| \leq M$

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

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

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

-
- Bernstein の不等式

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

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

(ヒルベルト空間版もある)

有限集合の一樣バウンド 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)

$\mathcal{F} = \{f_m \ (m = 1, \dots, M)\}$ 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)

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

一樣バウンド

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

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

有限集合の一樣バウンド 2: Bernstein の不等式版

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

Bernstein の不等式

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

一樣バウンド

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

※ 一樣バウンドはせいぜい $\sqrt{\log(M)}$ オーダで増える.

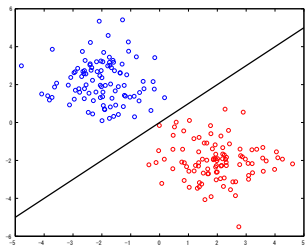
構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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

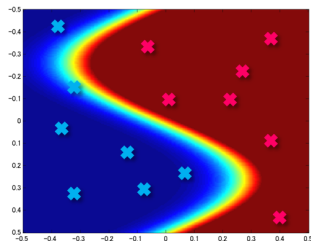
有限から無限へ

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？

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

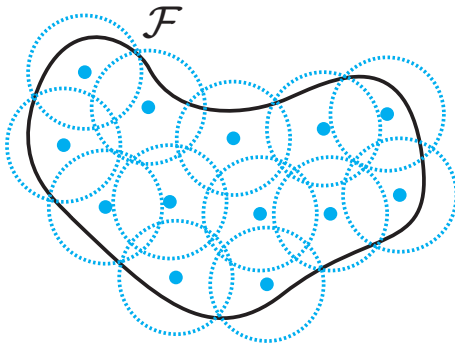


$$\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$$



基本的なアイデア

有限個の元で代表させる



Rademacher 複雑さ

$\epsilon_1, \epsilon_2, \dots, \epsilon_n$: 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:

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

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

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

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.,
 $|\psi(f) - \psi(f')| \leq B|f - f'|,$

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

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

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

Rademacher 複雑さの各種性質

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

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

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

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

特に最初の性質が有り難い.

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

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

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.

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

カバリングナンバー

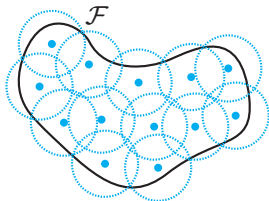
Rademacher complexity を抑える方法.

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,

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

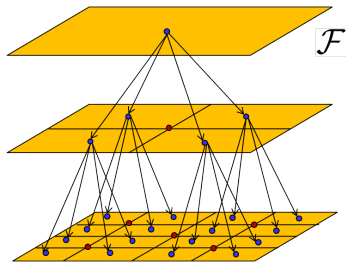
Dudley 積分のイメージ

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

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た、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

$$\begin{aligned} \hat{L}(\hat{f}) &\leq \hat{L}(f^*) \quad (\because \text{経験誤差最小化}) \\ \Rightarrow L(\hat{f}) - L(f^*) &\leq \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{\text{これを抑えたい}} + \underbrace{\hat{L}(f^*) - L(f^*)}_{O_p(1/\sqrt{n}) \text{ (Hoeffding)}} \end{aligned}$$

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,

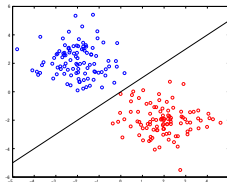
$$\begin{aligned} L(\hat{f}) - \hat{L}(\hat{f}) &\leq \sup_{f \in \mathcal{F}} (L(f) - \hat{L}(f)) \\ &\leq R(\ell(\mathcal{F})) + \sqrt{\frac{t}{n}} \quad (\text{with prob. } 1 - e^{-t}) \\ &\leq R(\mathcal{F}) + \sqrt{\frac{t}{n}} \quad (\text{contraction ineq., Lipschitz 連続}) \\ &\leq \frac{1}{\sqrt{n}}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

※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

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

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



すると, 0-1 ロス ℓ に対し

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

例: VC 次元

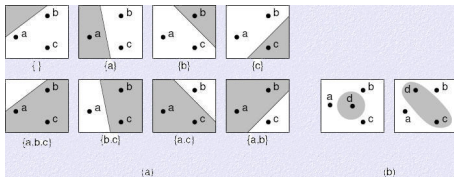
$\mathcal{F}$ は指示関数の集合: $\mathcal{F} = \{\mathbf{1}_C \mid C \in \mathcal{C}\}$.

$\mathcal{C}$ 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 $X_n = \{x_1, \dots, x_n\}$ を細分する
 $\Leftrightarrow$ 任意のラベル $Y_n = \{y_1, \dots, y_n\}$ ($y_i \in \{\pm 1\}$) 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.

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

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



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.

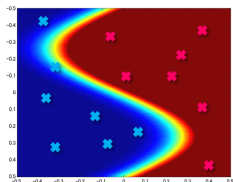
例: カーネル法

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

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.

$$\begin{aligned} \sum_{i=1}^n \epsilon_i f(x_i) &= \langle \sum_{i=1}^n \epsilon_i k(x_i, \cdot), f \rangle_{\mathcal{H}} \leq \left\| \sum_{i=1}^n \epsilon_i k(x_i, \cdot) \right\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &\leq \left\| \sum_{i=1}^n \epsilon_i k(x_i, \cdot) \right\|_{\mathcal{H}} \text{ を使う.} \end{aligned}$$

$$\begin{aligned} R(\mathcal{F}) &= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{|\sum_{i=1}^n \epsilon_i f(x_i)|}{n} \right] \leq \mathbb{E} \left[\frac{\left\| \sum_{i=1}^n \epsilon_i k(x_i, \cdot) \right\|_{\mathcal{H}}}{n} \right] \\ &= \mathbb{E} \left[\frac{\sqrt{\sum_{i,j=1}^n \epsilon_i \epsilon_j k(x_i, x_j)}}{n} \right] \leq \frac{\sqrt{\mathbb{E} \left[\sum_{i,j=1}^n \epsilon_i \epsilon_j k(x_i, x_j) \right]}}{n} \quad (\text{Jensen}) \\ &= \frac{\sqrt{\sum_{i=1}^n k(x_i, x_i)}}{n} \leq \frac{1}{\sqrt{n}} \end{aligned}$$

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 $|a_{ij}| \leq 1$ なる確率変数.

A の作用素ノルム $\|A\| := \max_{\substack{\|z\| \leq 1 \\ z \in \mathbb{R}^q}} \|Az\| = \max_{\substack{\|w\| \leq 1, \|z\| \leq 1 \\ w \in \mathbb{R}^p, z \in \mathbb{R}^q}} w^\top Az$.

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

$n = pq$ 個のサンプルがあるとみなす.

$$\|f_{w,z} - f_{w',z'}\|_n^2 = \frac{1}{pq} \sum_{i,j=1}^{p,q} |a_{ij}(w_i z_j - w'_i z'_j)|^2 \leq \frac{2}{pq} (\|w - w'\|^2 + \|z - z'\|^2)$$

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

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

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).

$$\text{モデル: } Y = X\beta^* + \xi.$$

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

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 $E[e^{\tau \xi_i}] \leq e^{\sigma^2 \tau^2 / 2}$ ($\forall \tau > 0$) 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, 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？

有限個の一樣バウンドからやってきた.

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

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

$\log(p)$ はどこからやってきたか？

有限個の一致バウンドからやってきた.

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

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

Hoeffding の不等式由来の一致バウンドにより, 確率 $1 - \delta$ で

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

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 $\{x_1, \dots, x_n\} \in \mathbb{R}^d$ 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 isometry (Baraniuk et al., 2008, Candès, 2008)

- Gaussian concentration inequality, concentration inequality on product space (Ledoux, 2001)

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

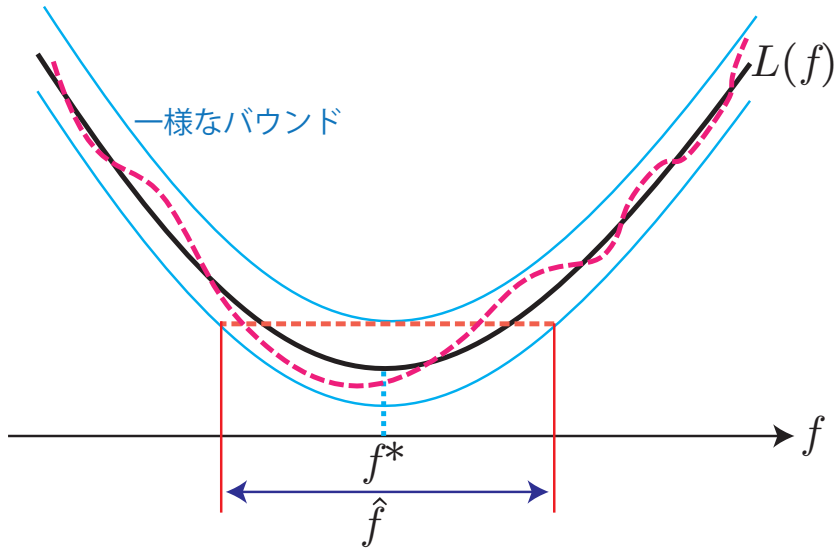
- Majorizing measure: ガウシアンプロセスにまつわる上界, 下界 (Talagrand, 2000).

構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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

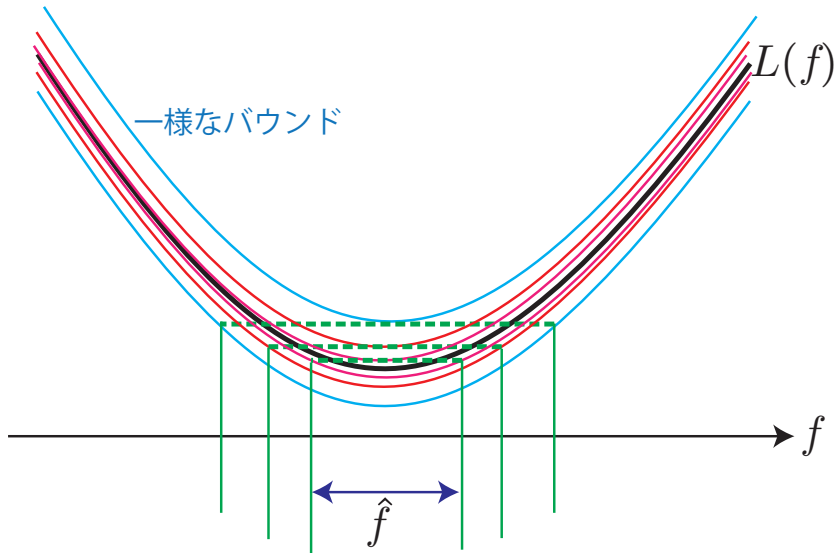
$O_p(1/\sqrt{n})$ オーダより速いレートは示せる？

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

次の条件を仮定してみる.

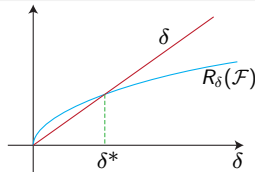
- $\mathcal{F}$ は 1 で上から抑えられている: $\|f\|_\infty \leq 1$ ($\forall f \in \mathcal{F}$).
- ℓ は Lipschitz 連続かつ 強凸:
 $\mathbf{E}[\ell(Y, f(X))] - \mathbf{E}[\ell(Y, f^*(X))] \geq B\mathbf{E}[(f - f^*)^2]$ ($\forall f \in \mathcal{F}$).

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

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

構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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.

構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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: $\{P_\theta | \theta \in \Theta\}$

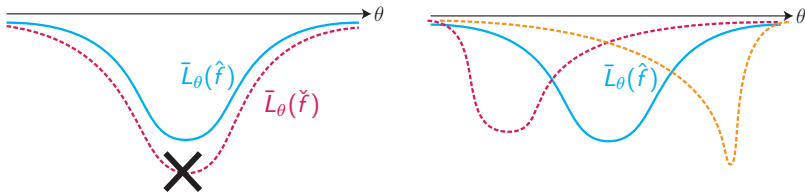
P_θ における推定量 $\check{f}$ のリスクの期待値:

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

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 $D_n = \{(x_1, \dots, x_n)\} \sim P_\theta^n$ 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 $L(\theta_0, \hat{P})$

$$\hat{P} = \arg \min_{\hat{P}: \text{推定量}} \int E_{D_n \sim P_{\theta_0}} [L(\theta_0, \hat{P})] \pi(\theta_0) d\theta_0.$$

- 二乗リスク $L(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$: $\hat{\theta} = \int \theta \pi(\theta | D_n) d\theta$ (事後平均)
- KL-リスク $L(\theta, \hat{P}) = \text{KL}(P_\theta \| \hat{P})$: $\hat{P} = \int P(\cdot | \theta) \pi(\theta | D_n) d\theta$ (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.

構成

- 1 はじめに: 理論の役割
- 2 統計的学習理論と経験過程
- 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)

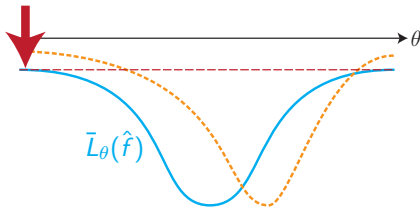
$\hat{f}$ が minimax 最適

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

学習理論では定数倍を許すことが多い: $\exists C$ で

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

そういう意味で「minimax レート」を達成する」と言ったりする.

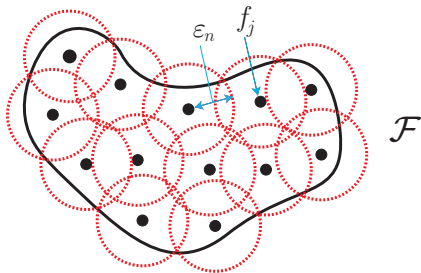


minimax レートを求める方法

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)

$$\{f_1, \dots, f_{M_n}\} \subseteq \mathcal{F}$$



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.

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. (2012), Suzuki and Sugiyama (2012).

構成

- ① はじめに: 理論の役割
- ② 統計的学習理論と経験過程
- 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

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

(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)

まとめ

一様バウンドが重要

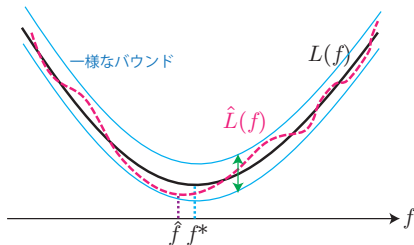
$$\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) - \mathbb{E}[\ell(Y, f(X))] \right\}$$

- 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



- H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- 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.
- P. L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, 8:775–790, 2007.
- 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.
- 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.
- C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.
- 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.
- T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT '95*, pages 23–37, 1995.
- 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.
- 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 Annual Conference on Computational Learning Theory*, pages 230–234, 1998.
- D. McAllester. PAC-Bayesian model averaging. In *the Annual 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, 2012.

- 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.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.
- 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.
- V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5):2109–2144, 2009.