日本応用数理学会論文誌
Vol. 23, No. 3, 2013, pp. 537~561

統計的学習理論概説

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

概要. 本稿では統計的学習理論における基本的な道具立てを概説する．統計的学習理論は，学習手法の意味や正当性およびその最適性といった問題を議論する．特に，教師データを増やしていった時いかに汎化誤差が収束してゆくかという問題は主要な興味の対象である．汎化誤差の振る舞いを解析するにあたり，経験過程の理論が重要な役割を果たす．また，minimax リスクの下限を導出する情報理論的技法も紹介する．

Overview of Statistical Learning Theory

Taiji Suzuki*
*University of Tokyo

Abstract. In this article, an overview of basic tools in statistical learning theory is given. The aim of learning theory is to clarify the meaning of a learning method, justify the method and show an optimality of that. In particular, analyzing the behavior of generalization error is one of the most important issues. To analyze it, the empirical process theory plays a vital role. Technical tools such as Rademacher complexity, covering number and Dudley's integral are useful in the analysis. Finally, minimax optimality is discussed. A theoretic technique to give a lower bound of the minimax risk is presented.

1. 統計的学習理論とは

統計的学習理論とは，機械学習における諸手法の性質を数学的に厳密に解明し，そこで得られた知見を新しい手法の構築へ役立たせるための理論体系である．学習理論においては統計学や確率論が主要な道具になる．これはなぜかというと，そもそも「機械学習」とはデータから何らかの知識を引き出すための手法であり，それそのものが統計的推論と非常に近しい概念であるからである．機械学習では「学習」という言葉が頻繁に用いられるが，それは多くの場合統計学における「推定」という言葉に置き換えても差支えない．これら二つの概念は，裏に流れる哲学や動機・応用先を異にするために全く同じものであると言いきることは難しいが，技術的なレベルで見る分には共通する部分が多い．また，今日における広い意味でのデータサイエンスの興隆により，両者の境界はますます薄くなってきている．このことより，本稿では学習を推定とほぼ同じ意味で用いる．

理論面について述べると，統計的学習理論と数理統計学の理論はやや色合いを異にする．数理統計学と言うと概して Fisher 流の統計学を指すことが多いが，本稿で紹介する学習理論は Kolmogorov らの確率論の流れを強く受けており，その理論展開には大きな違いが見てとれる．いわゆる統計的学習理論 ("Statistical Learning Theory") と言ったと

き，Vapnik による VC 次元の概念の導入から始まった一連の理論体系を指し，その理論の発端からして経験過程の理論と密接にかかわっている．経験過程の理論は Fisher 流の統計学にも推定量の漸近正規性や一致性といった性質の理論的正当化において不可欠である．しかし，実際の理論展開においてはその用いられ方に差が見られる．例えば，学習理論は有限サンプルでのリスクの評価を主に行い，数理統計学では漸近理論が主役である．また，学習理論ではなるべく分布に仮定を入れない理論展開が多いが，数理統計学では分布の性質を詳細に調べ，それを利用した理論展開が多い．これらの違いは経験過程の理論を前面に押し出した理論展開をするかどうかの違いをもたらす．学習理論ではやや「厄介な」問題を扱うことが多く，そのため経験過程の理論による帰結をそのまま用いることが多く，一方で数理統計学では経験過程の理論による結果を前提として，より詳細な議論に踏み込むことが多い．

上では学習理論と数理統計学の違いを明快にするために，その相違点に焦点を当てたが，それらの境界はやはり曖昧であり，どこまでが学習理論でどこまでが数理統計学かという線引きは難しく，また無理に線引きすること自体，意味をなさない．実際，機械学習手法と統計的推定手法の境界が曖昧になるにつれ，それらにまつわる理論もまた中間領域に位置するものが増えてきている．

では，学習理論は具体的になにを明らかにするのであろうか．学習理論における問題意識は大別して次の三つの要素に分けられる: (1) 手法の意味，(2) 手法の正当性，(3) 手法の最適性．まず (1) 手法の意味であるが，ある手法があったとして，その手法が「実は何をしているのか」ということは自明でない場合が多い．それを明確にするのが (1) である．例えば，二値判別においては凸ロス関数の選択によって，得られる推定量の意味が変わってくる．ロス関数の選択がいかに推定量を特徴付けるかといった問題は，[3, 4] といった文献で扱われている．次に (2) 手法の正当性であるが，十分サンプルを多くとってくれば，これだけの精度を保証できるというような正当性を与えることである．例えば PAC 学習 (Probably Approximately Correct) 可能性はこの問題を扱っている．(3) 手法の最適性とは，手法に正当性があるとしてそれがある規準のもと最適であるどうかを議論することである．これらの研究を通して手法を理解し，新しい手法の提案に還元するのが学習理論の目指すところである．本稿では主に (2) と (3) についてその基本的な事項を述べる．

2. 学習理論と経験過程の理論

前節で述べたように，学習理論と経験過程の理論には密接な関係がある．その関係を経験過程の理論の歴史とともに眺めてみよう．まず，経験過程の理論は 1933 年に Glivenko と Catelli によって証明された一様大数の法則 (Glivenko–Catelli の定理) から始まる [23]．これは経験分布関数が真の分布関数に収束すると主張するものである．ここで，「一様」という単語が用いられていることに注意されたい．分布関数が滑らか

であれば，ある一点での経験分布関数の値は，大数の法則により真の分布関数の値に確率収束する．この収束が全ての点で一様に確率収束することから「一様」という形容が付けられている．次いで，Kolmogorov がその収束レートと漸近分布を導出した [28]. Kolmogorov–Smirnov 検定はこの結果から構成される．また，この論文は一様中心極限定理の最も単純な形を与えている．その後，1952 年に Donsker が一様中心極限定理の一般化を試みている (Donsker の定理) [15] (ただし，後に証明に可測性に関する不備があることが指摘された)．さらに 1967 年には Dudley が Dudley 積分の概念を導入し，それにより一様バウンドの上界およびガウシアンプロセスの正則性の十分条件を与えている [16].

1968 年に学習理論において重要な役割を果たす VC 次元の概念が Vapnik と Chervonenkis により提案された [60]．VC 次元は一様 Glivenko–Catelli の必要十分条件を与え (6 節を参照)，分布によらない学習可能性に関する特徴付けを与えている．VC 次元は学習理論において今でも頻繁に用いられている重要な概念である．

経験過程の理論に関する詳細は，[17] や [59] が参考になるであろう．本稿で扱ういくつかの内容は，これらの本にその詳細が記述されている．

3. 問題設定: 経験誤差最小化

ここから問題設定を述べる．本稿では基本的に経験リスク最小化を考える．そのため，問題設定として教師有り学習を考えるのが分かりやすいであろう．今，教師データ $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ が与えられているとする．ここで，(x_i, y_i) は入力と応答の組であり，確率分布 P から i.i.d. で生成されているとする．これに対し，仮説集合 $\mathcal{F}$ なる $\mathcal{X}$ から $\mathbb{R}$ への関数の集合を用意し，そこからデータをよく説明する関数 $f \in \mathcal{F}$ を学習する．そのため，データへの当てはまりを測るロス関数 $\ell(\cdot, \cdot) : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ を用意する．例えば，二値判別ではヒンジロス $\ell(y, f) = \max\{1 - yf, 0\}$ やロジスティックロス $\ell(y, f) = \log(1 + \exp(-yf))$ が，回帰問題では二乗ロス $\ell(y, f) = (y - f)^2$ がよく用いられる．

ここで，教師データ D_n から何らかの方法で $\hat{f} \in \mathcal{F}$ を学習したとする．学習の目的は次で定義される「汎化誤差」をなるべく小さくすることである:

$$\mathrm{E}_{(X,Y)}[\ell(Y, \hat{f}(X))] - \inf_{f:\text{可測関数}} \mathrm{E}_{(X,Y)}[\ell(Y, f(X))].$$

ここで，$\mathrm{E}_{(X,Y)}$ は教師データとは独立同一なテストデータでの期待値である．統計的学習理論は次のような点を明らかにすることを目的にする．

- 汎化誤差は収束するか？
- その収束レートは？
- $\hat{f}$ に最適性はあるか？

本稿では，上のような疑問に応えるための数学的ツールを概説する．

3.1 Bias-Variance の分解

汎化誤差は未知の分布 P による期待値を含んでいるため，実際に知ることはできない．そこで，学習に際しては次で定義される経験リスクを用いる:

$$\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

一方で，期待リスク（真のリスク）を次のように定義する:

$$L(f) = \mathrm{E}_{(X,Y)}[\ell(Y, f(X))].$$

この表記を用いると，汎化誤差は次のように分解される:

$$L(\hat{f}) - \inf_{f:\text{可測関数}} L(f) = \left(L(\hat{f}) - \inf_{f\in\mathcal{F}} L(f)\right) + \left(\inf_{f\in\mathcal{F}} L(f) - \inf_{f:\text{可測関数}} L(f)\right).$$

ここで，右辺最初の二項 $L(\hat{f}) - \inf_{f\in\mathcal{F}} L(f)$ を推定誤差と呼び，残りの二項 $\inf_{f\in\mathcal{F}} L(f) - \inf_{f:\text{可測関数}} L(f)$ をモデル誤差と呼ぶ．推定誤差とモデル誤差にはトレードオフがある．すなわち，仮説集合 $\mathcal{F}$ が広ければ推定誤差が大きくなる一方でモデル誤差は小さくなり，仮説集合が小さければ推定誤差が小さくなるがモデル誤差が大きくなる．このトレードオフは非常に重要であり，ちょうどよいバランスを取るための手法として交差確認法や情報量規準，モデル平均といった方法がある．また，カーネル法におけるモデル誤差の取り扱いについては interpolation space の理論を使った理論 [5, 18, 43] がある．

しかし，本稿では簡単のためモデル誤差は十分小さく無視できるものとする．また，$f^* \in \mathcal{F}$ が存在して $\inf_{f\in\mathcal{F}} L(f) = L(f^*)$ とする．

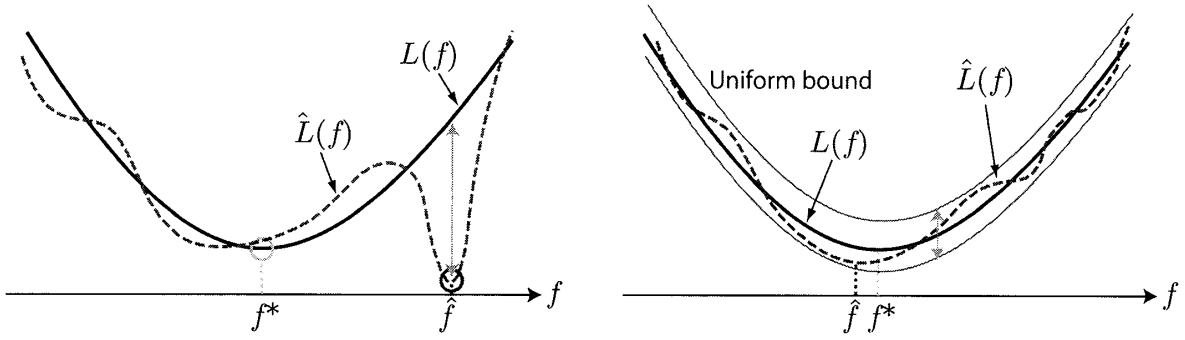
3.2 経験誤差最小化

$\hat{f}$ の学習方法として最も代表的な方法は，経験誤差最小化 (Empirical Risk Minimization) である:

$$\hat{f} = \operatorname*{arg\,min}_{f\in\mathcal{F}} \quad \hat{L}(f).$$

最尤推定は経験誤差最小化の一つの例である．ただし，単純に経験誤差を最小化すると過学習 (Fig. 1(a)) を起こしてしまうことがあるため，次のように正則化項 $\psi(f)$ を加えた正則化付き経験誤差最小化 (Regularized Empirical Risk Minimization) も考えられる:

$$\hat{f} = \operatorname*{arg\,min}_{f\in\mathcal{F}} \quad \hat{L}(f) + \psi(f).$$



(a) Over-fitting (b) Uniform bound of risk difference

Fig. 1. Relation between empirical risk and expected risk

SVM やカーネルロジスティック回帰などはこの定式化に含まれる．正則化付き経験誤差最小化に関する学習理論については，例えば [39,42] が参考になるだろう．本稿では基礎的なツールを紹介するという目的から，経験誤差最小化に限って話を進める．

これから経験誤差最小化の汎化誤差を導出しよう．まず，ほとんどのバウンドの導出の出発点である次の式から始めよう:

$$\begin{aligned} & \hat{L}(\hat{f}) \leq \hat{L}(f^*) \\ \Rightarrow \quad & L(\hat{f}) - L(f^*) \leq \left(L(\hat{f}) - \hat{L}(\hat{f})\right) + \left(\hat{L}(f^*) - L(f^*)\right). \end{aligned}$$

ここで，第一式は $\hat{f}$ が経験誤差を最小化している事実による．さて，第二式の左辺は汎化誤差であることに注意すると，汎化誤差を抑えるためには右辺を評価すれば良いことになる．右辺第三・四項 $(\hat{L}(f^*) - L(f^*))$ は後述の Hoeffding の不等式などから適当な条件のもと $\hat{L}(f^*) - L(f^*) = O_p(1/\sqrt{n})$ であることが示せる．問題は，右辺の第一・二項 $(L(\hat{f}) - \hat{L}(\hat{f}))$ である．一見して，これにも大数の法則などを当てはめれば良いように思える．しかし，教師データと $\hat{f}$ は独立ではないため，安直にそのような解析は適用できない．たまたまサンプルとの相性がよく，データに良く当てはまってしまった $\hat{f}$ が $L(\hat{f})$ も小さくするとは限らないからである (過学習，Fig. 1(a))．よって，そのような過学習が起きていないということを保証する必要がある．そこで，次のように経験リスクと期待リスクの差をその上限で抑えることを考える:

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

右辺 $\sup_{f \in \mathcal{F}}\{L(f) - \hat{L}(f)\}$ を抑えることによって，過学習が起きていないことを保証するのである[*1]．これの上界を一様バウンドと呼ぶ (Fig. 1(b))．一様バウンドを求めるために経験過程の理論が必要になる．

[*1] これ以降 $\sup_{f \in \mathcal{F}} f(X)$ なる量が頻繁に現れるが，一般にはこれは可測であるとは限らない．しかし，本稿ではこれが可測であるような $\mathcal{F}$ のみを扱う．

4. 有限集合における一様バウンド

ここでは $\mathcal{F}$ の要素数が有限個の場合の一様バウンドの導出を行う.

4.1 基本的な不等式

まずは, 一様バウンドを導出する前に, 測度集中に関する基本的な不等式を紹介する.

定理 4.1 (Hoeffding の不等式) Z_i $(i = 1, \ldots, n)$ を独立で (同一とは限らない) 期待値 0 かつ $[-m_i, m_i]$ に値を取る実確率変数とする. すると $\forall t > 0$ で次の不等式が成り立つ:

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

証明は, 例えば [59] を参照されたい. Hoeffding の不等式における条件 $|Z_i| \leq m_i$ は sub-Gaussian 性 $\mathrm{E}[e^{\tau Z_i}] \leq e^{m_i^2\tau^2/2}$ $(\forall \tau > 0)$ に一般化することができる. Z_i の値域だけでなく分散も分かっている場合, 次の Bernstein の不等式は有用である.

定理 4.2 (Bernstein の不等式) Z_i $(i = 1, \ldots, n)$ を独立で (同一とは限らない) 期待値 0 かつ $\mathrm{E}[Z_i^2] = \sigma_i^2$, $|Z_i| \leq M$ なる確率変数とする. すると $\forall t > 0$ で次の不等式が成り立つ:

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

$m_i = M$ $(\forall i)$ のとき, Bernstein の不等式は Hoeffding の不等式よりも $\{t \mid 0 < t < \frac{\sqrt{n}}{M}(M^2 - \frac{1}{n}\sum_{i=1}^n \sigma_i^2)\}$ なる領域でタイトである. なお, $\frac{1}{\sqrt{n}}Mt$ の項は $\frac{1}{3\sqrt{n}}Mt$ へ改善することができる. また, $|Z_i| \leq M$ なる条件はモーメント条件 $\mathrm{E}[|Z_i|^k] \leq \frac{k!}{2}\sigma^2 M^{k-2}$ に拡張できる. 詳しくは [59] や [42] を参照されたい.

これらの不等式より, 適当なモーメント条件のもと $\frac{\sum_{i=1}^n f(X_i)}{n} - \mathrm{E}[f(X)]$ は $O_p(1/\sqrt{n})$ で減少することに注意されたい.

4.2 一様バウンド

本節では上記の Hoeffding の不等式と Bernstein の不等式から導かれる一様バウンドを紹介する. ここではロス関数を明記しないが, これから紹介する一様バウンドに $g_m = \ell(y, f_m(x)) - \mathrm{E}[\ell(Y, f_m(X))]$ $(f_m \in \mathcal{F})$ を代入すればリスクに関する一様バウンドが導かれる.

4.2.1 Hoeffding の不等式版一様バウンド

$\mathcal{G} = \{g_m : \mathcal{X} \to \mathbb{R}\ (m = 1, \ldots, M)\}$ を有限個の関数集合とし，どれも期待値 0 であるとする $(\mathrm{E}[g_m(X)] = 0\ (\forall m))$．また，$\{X_1, \ldots, X_n\}$ を $\mathcal{X}$ に値を取る i.i.d. 確率変数列であるとする．すると，Hoeffding の不等式に $Z_i = g_m(X_i)$ を代入することにより，

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

を得る．これより直ちに次の一様バウンドを得る．

定理 4.3 $X_i \in \mathcal{X}\ (i = 1, \ldots, n)$ を i.i.d. 確率変数としたとき，全ての $0 < \delta \leq 1$ で次の不等式が成り立つ:

$$P\left(\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n g_m(X_i)|}{\sqrt{n}} > \max_m \|g_m\|_\infty \sqrt{2\log(2M/\delta)}\right) \leq \delta.$$

これより期待値に関しては，ある普遍定数 C が存在して次の不等式を得る:

$$\mathrm{E}\left[\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n g_m(X_i)|}{\sqrt{n}}\right] \leq C \max_m \|g_m\|_\infty \sqrt{\log(1+M)}.$$

証明

$$P\left(\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n g_m(X_i)|}{\sqrt{n}} > t\right) = P\left(\bigcup_{1\leq m\leq M} \frac{|\sum_{i=1}^n g_m(X_i)|}{\sqrt{n}} > t\right)$$

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

$$\leq 2M\exp\left(-\frac{t^2}{2\max_m \|g_m\|_\infty^2}\right).$$

ここで $t = \max_m \|g_m\|_\infty \sqrt{2\log(2M/\delta)}$ を代入することにより題意を得る． □

4.2.2 Bernstein の不等式版一様バウンド

Bernstein の不等式からも一様バウンドが得られる．

定理 4.4 $X_i \in \mathcal{X}\ (i = 1, \ldots, n)$ を i.i.d. 確率変数としたとき，普遍定数 C が存在して次の不等式が成り立つ:

$$P\left(\max_{1\leq m\leq M} \frac{|\sum_{i=1}^n g_m(X_i)|}{\sqrt{n}}\right.$$

$$> C\left\{\frac{\max_m \|g_m\|_\infty}{\sqrt{n}}\log\left(\frac{1+M}{\delta}\right)+\max_m\|g_m\|_{L_2(P)}\sqrt{\log\left(\frac{1+M}{\delta}\right)}\right\}\right) \le \delta,$$

$$\mathrm{E}\left[\max_{1\le m\le M}\frac{|\sum_{i=1}^n g_m(X_i)|}{\sqrt{n}}\right]$$

$$\le C\left\{\frac{1}{\sqrt{n}}\max_m\|g_m\|_\infty\log(1+M)+\max_m\|g_m\|_{L_2(P)}\sqrt{\log(1+M)}\right\}.$$

期待値に関する証明は [55] の Lemma 19.33 を参照されたい．裾確率は，定理 4.3 の証明のように直接求めても良いし，後で述べる Talagrand の不等式と期待値評価から求めることもできる．

定理 4.3 と定理 4.4 において重要なことは，$\mathcal{G}$ の要素数 M を大きくしていった場合でも，$\|g_m\|_\infty$ や $\|g_m\|_{L_2(P)}$ が一様に定数で抑えられるなら，一様バウンドがせいぜい $\log(M)/\sqrt{n}+\sqrt{\log(M)}$ のオーダでしか増えないことである．このように，M は一様バウンドにゆるやかにしか影響しないことより，多くの学習手法は実際に過学習せずに機能しているわけである．

5. Rademacher 複雑さと Dudley 積分

前節では有限の仮説集合を考えたが，仮説集合の要素が無限個あったらどうなるであろうか．また，可算集合でもなく連続濃度をもっていた場合どうすればよいであろうか．このような状況はありふれたものである．例えば線形判別や連続パラメータのパラメトリックモデルなどは連続濃度を持った仮説集合である．

このような状況を扱うために，仮説集合の「複雑さ」を定義する．複雑さの指標として Rademacher 複雑さとメトリックエントロピーの概念は非常に重要である．

5.1 Rademacher 複雑さ

まず，複雑さの指標として基本的な Rademacher 複雑さを導入しよう．Rademacher 変数 ϵ を $\{\pm 1\}$ に値を取る確率変数とし，$P(\epsilon=1)=P(\epsilon=-1)=\frac{1}{2}$ なるものとする．n 個の Rademacher 変数の i.i.d. 列 $\epsilon_1,\epsilon_2,\ldots,\epsilon_n$ を考える．これを用いて，Rademacher 複雑さは次のように定義される．

定義 5.1 (Rademacher 複雑さ) 関数集合 $\mathcal{F}$ の Rademacher 複雑さ $R(\mathcal{F})$ は次のように定義される:

$$R(\mathcal{F}) := \mathrm{E}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(X_i)\right|\right].$$

ここで期待値は Rademacher 変数列 $\{\epsilon_i\}_{i=1}^n$ と入力確率変数列 $\{X_i\}_{i=1}^n$ に関してとる．

Rademacher 複雑さの直観的な意味は，完全にランダムな系列 $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ と $\mathcal{F}$ との相関である．ランダムな系列と高い相関をもつのならそれだけ複雑な振る舞いを $\mathcal{F}$ が再現することができることを意味する．

Rademacher 複雑さを用いると一様バウンドが次のように得られる．これを対称化と呼ぶ．$\mathcal{F}_c := \{f - \mathrm{E}[f] \mid f \in \mathcal{F}\}$ とすると，最大値の期待値に関しては，

$$\frac{1}{2}R(\mathcal{F}_c) \leq \mathrm{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}(f(x_i) - \mathrm{E}[f])\right|\right] \leq 2R(\mathcal{F})$$

が成り立ち．さらに，もし $\|f\|_\infty \leq 1$ $(\forall f \in \mathcal{F})$ なら次の裾確率評価が成り立つ:

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

よって，Rademacher 複雑さを抑えられれば一様バウンドが得られることになる．これらの証明は，期待値に関しては [59] に詳細が記されており，裾確率に関しては期待値評価と McDiamid の不等式から導かれる [8].

命題 5.2 (Rademacher 複雑さの性質) Rademacher 複雑さには次のような性質がある．

- Contraction inequality: $\psi_i : \mathbb{R} \to \mathbb{R}$ $(i = 1, \ldots, n)$ が $\psi_i(0) = 0$ かつ B-Lipschitz 連続 $(\exists B$ で $|\psi_i(f) - \psi_i(f')| \leq B|f - f'|$, $\forall f, f' \in \mathbb{R})$ なら，

  $$\mathrm{E}\left[\frac{1}{n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\psi_i(f(X_i))\right|\right] \leq 2BR(\mathcal{F}).$$

  特に全ての ψ_i が等しい時 $(\psi = \psi_i)$,

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

  が成り立つ．
- 凸包の Rademacher 複雑さ: $\mathrm{conv}(\mathcal{F})$ を $\mathcal{F}$ の元の凸結合全体からなる集合とする $(\mathrm{conv}(\mathcal{F}) = \{\sum_{i=1}^{I}\lambda_i f_i \mid I \geq 1, \lambda_i \geq 0, \sum_{i=1}^{I}\lambda_i = 1, f_i \in \mathcal{F}\})$．するとその Rademacher 複雑さは次の性質を満たす:

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

特に contraction inequality は有用である．というのも，ロス関数がリプシッツ連続 $|\ell(y, f) - \ell(y, f')| \leq |f - f'|$ かつ $\ell(y, 0) = 0$ $(\forall y \in \mathcal{Y})$[*2]なら，

$$\mathrm{E}\left[\sup_{f\in\mathcal{F}}|\hat{L}(f) - L(f)|\right] \leq 2R(\ell(\mathcal{F})) \leq 4R(\mathcal{F}). \tag{5.1}$$

[*2] $\ell(y, 0) = 0$ $(\forall y)$ を満たさない場合は $\ell(y, f) \leftarrow \ell(y, f) - \ell(y, 0)$ と置きなおせば良い．

ただし，$\ell(\mathcal{F}) = \{\ell(\cdot, f(\cdot)) \mid f \in \mathcal{F}\}$．よって $\mathcal{F}$ の Rademacher complexity を抑えれば十分であることが分かる．

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．

例 1 再生核ヒルベルト空間の単位球を考え，その Rademacher 複雑さを求めてみよう．今，カーネル関数 k に対して，それに付随した再生核ヒルベルト空間 $\mathcal{H}_k$ を考える．その単位球を $\mathcal{B}(\mathcal{H}_k)$ とおくと，その Rademacher 複雑さは次のように抑えられる:

$$
\begin{aligned}
R(\mathcal{B}(\mathcal{H}_k)) =& \mathrm{E}\left[\sup_{f\in\mathcal{B}(\mathcal{H}_k)} \frac{|\sum_{i=1}^n \epsilon_i f(x_i)|}{n}\right] = \mathrm{E}\left[\sup_{f\in\mathcal{B}(\mathcal{H}_k)} \frac{|\sum_{i=1}^n \epsilon_i \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}|}{n}\right] \\
=& \mathrm{E}\left[\sup_{f\in\mathcal{B}(\mathcal{H}_k)} \left|\left\langle f, \frac{\sum_{i=1}^n \epsilon_i k(\cdot, x_i)}{n}\right\rangle_{\mathcal{H}_k}\right|\right] \\
\leq& \mathrm{E}\left[\sup_{f\in\mathcal{B}(\mathcal{H}_k)} \|f\|_{\mathcal{H}_k} \left\|\frac{\sum_{i=1}^n \epsilon_i k(\cdot, x_i)}{n}\right\|_{\mathcal{H}_k}\right] \\
=& \mathrm{E}\left[\left\|\frac{\sum_{i=1}^n \epsilon_i k(\cdot, x_i)}{n}\right\|_{\mathcal{H}_k}\right] = \frac{1}{n}\mathrm{E}\left[\sqrt{\textstyle\sum_{i,j=1}^{n,n} \epsilon_i\epsilon_j k(x_i, x_j)}\right] \\
\leq& \frac{1}{n}\sqrt{\mathrm{E}\left[\textstyle\sum_{i,j=1}^{n,n} \epsilon_i\epsilon_j k(x_i, x_j)\right]} \quad (\because \text{Jensen の不等式}) \\
=& \frac{1}{n}\sqrt{\sum_{i=1}^n \mathrm{E}_X[k(X, X)]} = \frac{1}{\sqrt{n}}\sqrt{\mathrm{E}_X[k(X, X)]}.
\end{aligned}
$$

よって，カーネル関数が $k(x, x) \leq 1$ $(\forall x)$ の時 (例えばガウシアンカーネルはこれを満たす)，$R(\mathcal{B}(\mathcal{H}_k)) \leq \frac{1}{\sqrt{n}}$ を得る．

5.2 被覆数

仮説集合の複雑さとして Rademacher 複雑さに加え，被覆数と呼ばれる概念がよく用いられる．これは，Rademacher 複雑さを上から抑える際に有用である．被覆数を用いた一様バウンドの導出の基本的な発想は，有限集合でない $\mathcal{F}$ を有限個の元で近似し，有限の場合で用いた論理を適用して一様バウンドを導出することである．

被覆数は，経験過程の理論においては Glivenko–Cantelli クラス (一様大数の法則が成り立つクラス) や Donsker クラス (一様中心極限定理が成り立つクラス) の特徴付けに極めて重要な役割を果たす．この点に関しては [59] や [33] に詳細が記されている．

定義 5.3 ϵ-被覆数 $(\mathcal{F}, d)$ を実関数 $f : \mathcal{X} \to \mathbb{R}$ のノルム d が備わったあるノルム空間の部分集合であるとする．このとき，$\mathcal{F}$ の ϵ-被覆数 $N(\mathcal{F}, \epsilon, d)$ は，ノルム d で定まる半径 ϵ

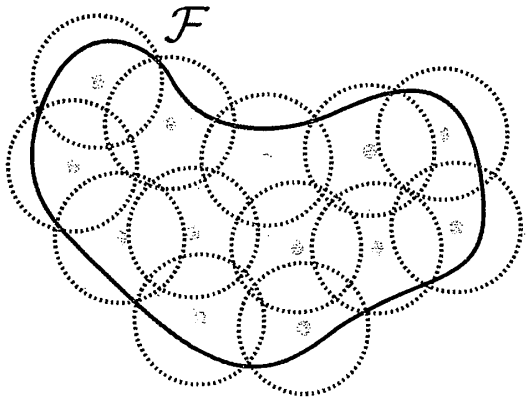


Fig. 2. ϵ-covering number

のボールで $\mathcal{F}$ を覆うために必要な最小のボールの数と定義される．ただし，各ボールの中心は $\mathcal{F}$ に含まれている必要はない．

被覆数は有限個の元で $\mathcal{F}$ を近似するのに最低限必要な個数といえる．これを用いて，有限集合で用いた論理を一般の集合に拡張するわけである．また，被覆数は後述の局所 Rademacher 複雑さの評価に効いてくる．

$\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f(x_i)^2$ とする．すると次に示すように Dudley 積分を用いて Rademacher 複雑さの上界が与えられる．

定理 5.4 (Dudley 積分) $\mathcal{F}$ に 0 関数 $(f(x) = 0\ \forall x)$ が含まれているする．すると，ある普遍定数 C が存在して次の不等式が成り立つ:

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

証明は [59] の Corollary 2.2.8 を参照されたい．Dudley 積分のイメージは有限個の元で $\mathcal{F}$ を近似する際，その解像度をだんだん荒くしてゆき，似ている元をまとめ上げてゆくというものである．積分の中身には $\sqrt{\log(N(\mathcal{F}, \epsilon, \|\cdot\|_n))}$ が現れており，これは有限集合の一様バウンドで現れた $\sqrt{\log(M)}$ と対応するものである．また，解像度 ϵ を変化させて誤差を積み上げてゆく操作が積分を取ることに対応する．この操作をチェイニングと呼ぶ．なお，ジェネリックチェイニング (generic chaining) という技法を用いれば Dudley 積分よりもタイトなバウンドを導くことができる [50].

Dudley 積分においては被覆数を用いたが，一方で次で定義されるエントロピー数 [61] を用いた方が便利な場合もある (例 3 を見よ).

定義 5.5 (エントロピー数) $(\mathcal{F}, d)$ を実関数 $f : \mathcal{X} \to \mathbb{R}$ のノルム d が備わったあるノルム空間の部分集合であるとする．このとき，$\mathcal{F}$ のエントロピー数 $e_i(\mathcal{F}, d)$ は，$e_i(\mathcal{F}, d) := \inf\{\epsilon > 0 \mid N(\mathcal{F}, \epsilon, d) \leq 2^{i-1}\}$ として定義される．

すると，ある普遍定数 C が存在して次の不等式が成り立つ:

$$\int_0^\infty \sqrt{\log(N(\mathcal{F},\epsilon,d))}\mathrm{d}\epsilon \le C\sum_{j=0}^{\infty} 2^{j/2}e_{2^j}(\mathcal{F},d).$$

これより，定理 5.4 から次の不等式を得る:

$$R(\mathcal{F}) \le \frac{C}{\sqrt{n}}\mathrm{E}_{D_n}\left[C\sum_{j=0}^{\infty} 2^{j/2}e_{2^j}(\mathcal{F},\|\cdot\|_n)\right]. \tag{5.2}$$

例 2 Dudley 積分を用いて実際に一様バウンドを出してみよう．ℓ が 1-Lipschitz ($|\ell(y,f)-\ell(y,f')| \le |f-f'|$) かつ $\ell(y,0)=0$ $(\forall y)$ と仮定する．また，$\|f\|_\infty \le 1$ $(\forall f\in\mathcal{F})$ で 0 関数が $\mathcal{F}$ に含まれているとき，

$$\begin{aligned} L(\hat{f})-\hat{L}(\hat{f}) &\le \sup_{f\in\mathcal{F}}(L(f)-\hat{L}(f)) \\ &\le 2R(\ell(\mathcal{F}))+\sqrt{\frac{t}{n}} \quad (\text{with probability } 1-e^{-t}) \\ &\le 4R(\mathcal{F})+\sqrt{\frac{t}{n}} \quad (\text{contraction inequality (5.1)}) \\ &\le \frac{C}{\sqrt{n}}\mathrm{E}_{D_n}\left[\int_0^\infty \sqrt{\log N(\mathcal{F},\epsilon,\|\cdot\|_n)}\mathrm{d}\epsilon\right]+\sqrt{\frac{t}{n}} \quad (\text{Dudley 積分}) \end{aligned} \tag{5.3}$$

なる評価を得る．

例 3 (カーネル法) 例 1 では，再生核ヒルベルト空間の単位球の Rademacher 複雑さを直接求めたが，ここでは被覆数やエントロピー数の性質について述べる．今 d 次元ユークリッド空間上の半径 B の閉超球 $\{x\in\mathbb{R}^d \mid \|x\|\le B\}$ 上で定義されたカーネル関数 k とそれに付随した再生核ヒルベルト空間 $\mathcal{H}_k$ を考える．今，カーネル関数 k が m 回連続微分可能な場合，B,m,d に依存した定数 $c_{B,m,d}$ が存在して，

$$\log(N(\mathcal{B}(\mathcal{H}_k),\epsilon,\|\cdot\|_\infty)) \le c_{B,m,d}\epsilon^{-2d/m}$$

が成り立つ [42, Theorem 6.26]．ここで，$N(\mathcal{B}(\mathcal{H}_k),\epsilon,\|\cdot\|_n) \le N(\mathcal{B}(\mathcal{H}_k),\epsilon,\|\cdot\|_\infty)$ であることに注意されたい．

また，もし $0<p<1$，$a\ge 1$ なる定数が存在して，$L_2(P)$ ノルムに関するエントロピー数が

$$e_i(\mathcal{B}(\mathcal{H}_k),\|\cdot\|_{L_2(P)}) \le ai^{-\frac{1}{2p}} \quad (\forall i\ge 1) \tag{5.4}$$

を満たすなら，$\|\cdot\|_n$ ノルムに関するエントロピー数の期待値は次のように抑えられる [42, Corollary 7.31]:

$$\mathrm{E}_{D_n}[e_i(\mathcal{B}(\mathcal{H}_k),\|\cdot\|_n)] \le c_p a(\min\{i,n\})^{\frac{1}{2p}}i^{-\frac{1}{p}} \quad (\forall i\ge 1).$$

ただし $c_p > 0$ は p に依存した定数である．この関係式は式 (5.2) の右辺を評価するのに有用である．式 (5.4) の十分条件としてカーネルの固有値を用いた特徴付けがある．今，カーネル関数 k に付随した積分作用素 $T_k : L_2(P) \to L_2(P)$ を

$$T_k f(x) = \int_{\mathcal{X}} k(x, x') f(x') \mathrm{d}P(x')$$

と定める．T_k の i 番目に大きな固有値を $\mu_i(T_k)$ とおくと，全ての $q > 0$ に対して定数 c'_q が存在し，全ての $m \geq 1$ において次の関係が成り立つ [43, Theorem 15]:

$$\sup_{i \leq m} i^{1/q} e_i(\mathcal{B}(\mathcal{H}_k), \|\cdot\|_{L_2(P)}) \leq c'_q \sup_{i \leq m} i^{1/q} \mu_i^{\frac{1}{2}}(T_k),$$
$$\mu_i^{\frac{1}{2}}(T_k) \leq 2 e_i(\mathcal{B}(\mathcal{H}_k), \|\cdot\|_{L_2(P)}).$$

これより $\mu_i(T_k) \leq a i^{-\frac{1}{p}}$ $(p \in (0,1))$ なら，

$$e_i(\mathcal{B}(\mathcal{H}_k), \|\cdot\|_{L_2(P)}) \leq c'_{2p} a i^{-\frac{1}{2p}}$$

が成り立つ．

6. VC 次元

本節では学習理論において重要な意味を持つ VC 次元について述べる．VC 次元の概念は，被覆数の評価や Rademacher 複雑さを抑えらるのに使えるのに加え，後述の一様 Glivenko-Cantelli の必要十分条件も与える．VC 次元は確率分布に依らない量である．よって，VC 次元で特徴付けられる性質は全て特定の分布に依存しない性質であることに注意されたい．

$\mathcal{F}$ を指示関数の集合とする: ある集合族 $\mathcal{C}$ に対して，$\mathcal{F} = \{\mathbf{1}_C \mid C \in \mathcal{C}\}$ (例: 半空間の集合)．ここで集合 C の指示関数 $\mathbf{1}_C$ は $\mathbf{1}_C(x) = 1$ $(x \in C)$, 0 $(x \notin C)$ なる関数である．

定義 6.1 (VC 次元) $\mathcal{F}$ がある与えられた有限集合 $X_n = \{x_1, \ldots, x_n\}$ を細分するとは，任意のラベル $Y_n = \{y_1, \ldots, y_n\}$ $(y_i \in \{\pm 1\})$ に対してある $f \in \mathcal{F}$ が存在し，$f(x_i) = 1$ $(\forall i \text{ s.t. } y_i = 1)$ かつ $f(x_i) = 0$ $(\forall i \text{ s.t. } y_i = -1)$ を満たすことと定義する．$\mathcal{F}$ の VC 次元 $V_{\mathcal{F}}$ とは，$\mathcal{F}$ が細分できる集合 X_n が存在しえない n の最小値とする．

VC 次元有限な仮説集合 $\mathcal{F}$ の被覆数は次のように抑えられる [59, Theorem 2.6.4]: ある普遍定数 K が存在して，

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

この関係式と定理 5.4 より,

$$R(\mathcal{F}) = O(\sqrt{V_{\mathcal{F}}/n}) \tag{6.1}$$

となることが分かる.

続いて, 一様 Glivenko-Cantelli クラスの特徴付けについて述べる. $\mathcal{F}$ が一様 Glivenko–Cantelli クラスであるとは

$$\lim_{n\to\infty} \sup_P \mathrm{E}_{D_n\sim P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathrm{E}_P[f]\right|\right] = 0$$

を満たすことである. ここで $\sup_P$ は全ての確率測度の上でとる.

定理 6.2 $\mathcal{F}$ が一様 Glivenko–Cantelli クラスであるための必要十分条件は $\mathcal{F}$ の VC 次元が有限であることである.

十分条件は式 (6.1) から明らか. 必要条件は [17, Theorem 6.4.5] を参照されたい.

例 4 (線形判別機) $\mathcal{F}$ を線形判別機の集合とする: $\mathcal{F} = \{\mathbf{1}_C \mid C = \{x \in \mathbb{R}^d \mid x^\top\beta + c \geq 0\},\ \beta \in \mathbb{R}^d,\ c \in \mathbb{R}\}$. この時, $\mathcal{F}$ の VC 次元は $d+2$ である [59, Exercise 2.6.14].

7. 局所 Rademacher 複雑さ

これまでの章で導出された収束レートは基本的に $O_p(1/\sqrt{n})$ より遅いものであったが, ロス関数の強凸性を仮定するとより速いレートが達成できることが示される. 例えば, 正則モデルでの最尤推定量はリスクの期待値が $O(d/n)$ (d はパラメータの次元) となることはよく知られた事実である. この章ではこれを一般化した議論を進める. そこで重要になってくるのが局所 Rademacher 複雑さという概念である.

定義 7.1 (局所 Rademacher 複雑さ) ある $\delta > 0$ に対して, 局所 Rademacher 複雑さを期待リスク最小化元 f^* を用いて次のように定義する:

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

$R_\delta(\mathcal{F}) := R(\{f - g \mid \mathrm{E}[(f-g)^2] \leq \delta, f, g \in \mathcal{F}\})$ のように定義する場合もある. ここで, 次の条件を仮定してみよう.

- $\mathcal{F}$ の元の絶対値ノルムは 1 で上から抑えられている: $\|f\|_\infty \leq 1$ $(\forall f \in \mathcal{F})$.
- ℓ は Lipschitz 連続かつ強凸である: $|\ell(y,f) - \ell(y,f')| \leq |f - f'|$ $(\forall f, f' \in \mathbb{R}, \forall y \in \mathcal{Y})$ かつ $\mathrm{E}[\ell(Y, f(X))] - \mathrm{E}[\ell(Y, f^*(X))] \geq B\mathrm{E}[(f-f^*)^2]$ $(\forall f \in \mathcal{F})$.

すると経験誤差最小化元 $\hat{f}$ の汎化誤差は次のように抑えられる.

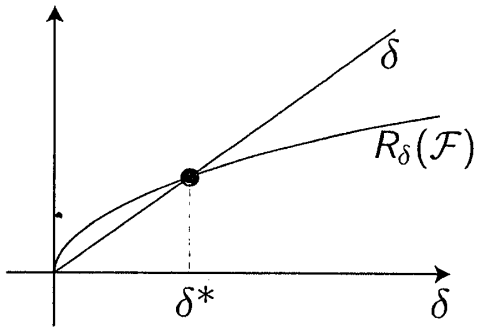


Fig. 3. δ^* as a fixed point

定理 7.2 (Fast learning rate [2, Corollary 5.3])

$$\delta^* := \inf\{\delta > 0 \mid \delta \geq R_\delta(\mathcal{F})\}$$

と定義すると，B に依存した定数 C が存在して確率 $1 - e^{-t}$ で

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

が成り立つ.

これを Fast learning rate と言う．なお，$\delta^* \leq R(\mathcal{F})$ は常に成り立つ．なぜなら $R(\mathcal{F}) \geq R_{R(\mathcal{F})}(\mathcal{F})$ が常に成り立つからである．δ^* は Fig. 3 のように不動点となっていることに注意されたい.

例 5 $\sup_Q \log N(\mathcal{F}, \epsilon, \|\cdot\|_{L_2(Q)}) \leq C\epsilon^{-2p}$ $(0 < p < 1)$ なる仮説集合 $\mathcal{F}$ を考える．ここで $\sup_Q$ は全ての有限離散確率測度に関して取る．このような仮説集合の例として例 3 を参照されたい．この時,

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

が成り立つ [38, Lemma 2.5]．すると，δ^* の定義から確率 $1 - e^{-t}$ で次が成り立つ:

$$L(\hat{f}) - L(f^*) \leq C\left(n^{-\frac{1}{1+p}} + \frac{t}{n}\right). \tag{7.1}$$

これは $O_p(1/\sqrt{n})$ よりタイトである.

例 6 有限次元パラメトリックモデルを考えてみよう．簡単のため線形回帰を考える．$e_1, \ldots, e_d$ を $L_2(P(X))$ における d 次元正規直交系とし，$\Theta \in \mathbb{R}^d$ をコンパクトな d 次元パラメータ空間とする．これに付随する線形関数のモデルを $\mathcal{F} = \{f_\theta(x) = \sum_{k=1}^d e_k\theta_k \mid$

$\theta \in \Theta\}$ とおく．すると，$\mathrm{E}[(f_\theta - f_{\theta'})^2] = \|\theta - \theta'\|^2$ となる．これより，期待リスク最小化元 f^* に対応するパラメータを θ^* とすると，Cauchy-Schwarz の不等式を用いて，

$$\begin{aligned} R_\delta(\mathcal{F}) &= \mathrm{E}\left[\sup_{\theta\in\Theta:\|\theta-\theta^*\|^2\le\delta}\left|\frac{\sum_{i=1}^n \epsilon_i(f_\theta(x_i) - f_{\theta^*}(x_i))}{n}\right|\right] \\ &= \mathrm{E}\left[\sup_{\theta\in\Theta:\|\theta-\theta^*\|^2\le\delta}\left|\frac{\sum_{k=1}^d(\theta_k - \theta_k^*)\sum_{i=1}^n \epsilon_i e_k(x_i)}{n}\right|\right] \\ &\le \sqrt{\delta}\left\{\sum_{k=1}^d \mathrm{E}\left[\left(\frac{\sum_{i=1}^n \epsilon_i e_k(x_i)}{n}\right)^2\right]\right\}^{1/2} = \sqrt{\frac{\delta d}{n}} \end{aligned}$$

を得る．これより，$\delta^* \le d/n$ となり，良く知られた有限次元パラメトリックモデルにおけるリスクの収束レートを得る．

その他の具体例は [29] が参考になる．またカーネル法における fast learning rate は [42] に詳細がまとめられている．

局所 Rademacher 複雑さに関する参考文献を挙げておこう．まず，局所 Rademacher 複雑さの一般論は [2,29] で展開されている．また，判別問題において Tsybakov の低雑音条件と呼ばれる条件を課すと $O_p(1/\sqrt{n})$ より速いレートが示せることが [52] によって証明されている．この結果は，単純な条件を課すだけで判別問題における minimax レートである $O_p(1/\sqrt{n})$ を改善できることを指摘した点で非常に重要である．また，これに対し後に局所 Rademacher 複雑さを用いた導出が与えられた [29]．[3] では，Tsybakov の低雑音条件のもとでの凸ロス関数最小化を議論している．また，局所 Rademacher 複雑さと本質的に同値な概念は peeling device として以前から知られていた．Peeling device という命名は van de Geer によるもので，[54] にその全容がまとめられている．

8. その他のトピック

ここでは，上で扱いきれなかったトピックをいくつか紹介する．それらの中には，現在進行中であり今後の発展が期待されるものも含まれる．

8.1 高次元推定問題における理論

高次元推定問題は現在非常に広く研究されている分野である．ここでは，その中でも最も基本的な Lasso [51] の収束レートに関する結果を紹介しよう．デザイン行列を $X = (X_{ij}) \in \mathbb{R}^{n\times p}$ とおく．ここで，p は次元であり，サンプル数 n よりも大きいことを許す．$n \gg p$ の状況は古典的な線形回帰の漸近論の範疇だが，$p \gg n$ となるとそのような手法は適用できず工夫が必要になる．

今，真の重みベクトルを $\beta^* \in \mathbb{R}^p$ とおき，β^* の非ゼロ要素の個数はたかだか d 個であ

るようなスパースな状況を考える．β^* を用いて次のような生成モデルを考える:

$$Y = X\beta^* + \xi.$$

ここで，$\xi = [\xi_1, \ldots, \xi_n]^\top$ は雑音であり，観測量は (X, Y) である．観測量 (X, Y) から β^* を推定する．Lasso は次の最適化問題で定義される推定量である:

$$\hat{\beta} \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_1. \tag{8.1}$$

デザイン行列 X に次で定義される Restricted eigenvalue condition を仮定する [6]:

$$\kappa(s, c_0) := \min_{\substack{J_0 \subseteq \{1,2,\ldots,p\} \\ |J_0| \le s}} \min_{\substack{\delta \neq 0, \\ \|\delta_{J_0^c}\|_1 \le c_0 \|\delta_{J_0}\|_1}} \frac{\|X\delta\|}{\sqrt{n}\|\delta_{J_0}\|} > 0. \tag{8.2}$$

ただし，インデックス集合 J に対して $\delta_J = (\delta_i)_{i \in J}$ なるベクトルとする．すると次の収束レートを得る．

定理 8.1 (Lasso の収束レート [6, 63]) デザイン行列が Restricted eigenvalue condition (8.2) を $\kappa(2d, 3) > 0$ で満たし，かつ $\max_{i,j} |X_{ij}| \le 1$ であり，ノイズが $\mathrm{E}[e^{\tau \xi_i}] \le e^{\sigma^2 \tau^2 / 2}$ $(\forall \tau > 0)$ を満たすとする．すると，$\forall \delta \in (0, 1)$ に対し $\lambda_n = 4\sigma\sqrt{2\log(2p/\delta)/n}$ とすると，ある普遍定数 C が存在して，確率 $1 - \delta$ で

$$\|\hat{\beta} - \beta^*\|_2^2 \le \frac{C\sigma^2}{\kappa(2d, 3)} \frac{d \log(p/\delta)}{n}$$

が成り立つ．

これより，次元が高くてもたかだか $\log(p)$ オーダしか影響せず，実質的な次元 d の方が支配的であることがわかる．

p の影響が $\log(p)$ で済んでいるのはなぜだろうか．これは有限個の一様バウンドから導くことができる．まず $\hat{\beta}$ が式 (8.1) なる最適化問題の最適解であることより，

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

が成り立つ．これを式変形し，$\xi^\top X(\hat{\beta} - \beta^*) \le \|X^\top \xi\|_\infty \|\hat{\beta} - \beta^*\|_1$ を使うと，

$$\frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda_n \|\hat{\beta}\|_1 \le \frac{2}{n}\|X^\top \xi\|_\infty \|\hat{\beta} - \beta^*\|_1 + \lambda_n \|\beta^*\|_1$$

が得られる．ここで，$\frac{1}{n}\|X^\top \xi\|_\infty = \max_{1 \le j \le p} \left|\frac{1}{n}\sum_{i=1}^n X_{ij}\xi_i\right|$ に注意すると，Hoeffding の不等式由来の一様バウンド (定理 4.3 の拡張，定理 4.1 直後の記述を参照) により，確率 $1 - \delta$ で

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

が成り立つ．ここに現れる $\log(p)$ が定理 8.1 の収束レートに現れているわけである．

近年，restricted eigenvalue condition を仮定しなくても，Bayes 的推定量を用いれば関数値 $X\hat{\beta}$ に関して上で述べたような収束レートを達成することがわかってきた [1,14,40,41]．また，ノンパラメトリックスパース加法モデルにおいても Bayes 的推定量が有用であることが [13,24,46] で示されている．このように Bayes 的手法の高次元推定問題への応用は精力的に研究されているが，変数選択の一致性といった事後分布のより詳細な性質に関してはまだまだ研究の余地がある．また，他の様々なスパース学習においても Bayes 的手法を考案し，その統計的性質を調べるという研究は有益であると考えられる．

ノンパラメトリック加法モデルの推定について，もう少し詳しく述べておこう．ノンパラメトリック加法モデルの推定方法として Multiple Kernel Learning (MKL) と呼ばれる手法がある [32]．MKL の収束レートは盛んに研究されており，例えば [11,26,27] では ℓ_p ノルムを正則化項に用いた MKL の収束レートを出している．これらは局所 Rademacher 複雑さを用いておらず，fast learning rate は出していない．一方で，[30,31,37] においては局所 Rademacher 複雑さを用いて fast learning rate を導出している．さらに，これらの研究を拡張して [47,48] では，真の関数のノルムや滑らかさに応じたより速い収束レートを導いている．また，その導出されたレートが minimax 最適であることも示されている．これらの研究は真がスパースな場合を扱っていたが，スパースでない場合において，全ての単調増加な混合ノルム型正則化に適用できる fast learning rate が [44,45] で導出されている．そこでは，導出されたバウンドを用いて最適な正則化項の選択が議論されている．これは，[25] で導出された ℓ_p ノルム正則化における fast learning rate を包含している．

8.2 Talagrand の concentration inequality

Talagrand の concentration inerquality は非常に汎用性の高い不等式である．実際，局所 Rademacher 複雑さを用いた fast learning rate の導出に用いられている．

定理 8.2 (Talagrand の concentration inequality [7,34,49]) $\sup_{f\in\mathcal{F}}\|f\|_\infty \leq 1$ を仮定する．$\sigma := \sup_{f\in\mathcal{F}} \mathrm{E}[f(X)^2]$, $P_n f := \frac{1}{n}\sum_{i=1}^n f(x_i)$, $Pf := \mathrm{E}[f(X)]$ とする．すると次の不等式が成り立つ:

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

9. Minimax 最適性

これまでは，汎化誤差の上界を評価する方法を紹介してきたが，本節では手法の最適性を調べる方法を紹介しよう．最適性には様々な規準があるが，統計学では主に「許容性」

と「minimax 最適性」を扱うことが多い．学習理論においては minimax 最適性を議論することが多いので，本稿ではそれについて述べる．

簡単のために回帰問題に限定して話を進める．仮説集合 $\mathcal{F}$ は実数値関数 $f : \mathcal{X} \to \mathbb{R}$ の集合とする．ある真の関数 $f^* \in \mathcal{F}$ が存在して，

$$y_i = f^*(x_i) + \xi_i \quad (i = 1, \ldots, n)$$

なるモデルを仮定する．ここで x_i $(i = 1, \ldots, n)$ はある確率分布 P から i.i.d. で生成されているとし，ξ_i $(i = 1, \ldots, n)$ は x_i とは独立な i.i.d. ガウス雑音 (平均 0，標準偏差 σ) であるとする．この時，ある推定量 $\hat{f}$ が minimax 最適であるとは

$$\sup_{f^*} \mathrm{E}_{D_n}[\|\hat{f} - f^*\|^2_{L_2(P)}] = \inf_{\tilde{f}:\text{推定量}} \sup_{f^*} \mathrm{E}_{D_n}[\|\tilde{f} - f^*\|^2_{L_2(P)}] \tag{9.1}$$

を満たすことと定義する．なお，$\inf_{\tilde{f}:\text{推定量}}$ は教師データ D_n から構成される全ての推定量の中で取る．右辺の値を minimax リスクと呼ぶ．

ここでの目的は minimax リスクを評価することである．下限を導出する技法は [53] に詳しい．また [62] においては情報理論における Fano の不等式を用いた明快な理論が展開されている．ここでは [62] による手法を回帰問題を用いて紹介する．Fano の不等式と minimax 最適性の関係を簡単に述べておこう．まず，minimax リスクの下限を導出するために仮説集合から何らかの方法で有限個の代表点を取り出して，その中から真の関数を推定するという，より簡単な問題を考える．問題が簡単な分，minimax リスクもより小さくなることに注意されたい．この問題はデータから真の情報源を有限個の候補から探す情報源復号化の問題と捉えることができる．よって，その誤り確率を評価するには，復号誤り確率の下限を与える Fano の不等式 [12, 19] を適用すれば良い．なお，ちょうど良い有限個の代表点を取り出すところで，被覆数の評価が重要になってくる．

ある $0 < \alpha < 1$ に対して，

$$\liminf_{\epsilon \to 0} \log(N(\mathcal{F}, \alpha\epsilon, \|\cdot\|_{L_2(P)}))/\log(N(\mathcal{F}, \epsilon, \|\cdot\|_{L_2(P)})) > 1 \tag{9.2}$$

が成り立っていると仮定する．すると次の定理が成り立つ．

定理 9.1 仮説集合 $\mathcal{F}$ が式 (9.2) と $\|f\|_\infty \leq 1$ $(\forall f \in \mathcal{F})$ を満たしているとする．ϵ_n を

$$\log(N(\mathcal{F}, \epsilon_n, \|\cdot\|_{L_2(P)})) = n\epsilon_n^2$$

なる実数とすると，

$$\inf_{\tilde{f}:\text{推定量}} \sup_{f^*} \mathrm{E}_{D_n}[\|\tilde{f} - f^*\|^2_{L_2(P)}] \asymp \epsilon_n^2$$

が成り立つ．

よって例3で見たように，$0 < p < 1$ なる定数が存在して，

$$\log(N(\mathcal{F}, \epsilon_n, \|\cdot\|_{L_2(P)})) \asymp \epsilon^{-2p}$$

が成り立つと仮定すると，そのminimaxリスクは

$$\inf_{\tilde{f}:\text{推定量}} \sup_{f^*} \mathrm{E}_{D_n}[\|\tilde{f} - f^*\|^2_{L_2(P)}] \asymp n^{-\frac{1}{1+p}}$$

となる．これより，局所Rademacher複雑さから導かれた汎化誤差の上界（式(7.1)）はタイトであることがわかる．例えば二乗ロス $\ell(y, f) = (y - f)^2$ を用いれば，経験リスク最小化は式(7.1)から（定数倍を除いて）minimax最適であることがわかる．

一般的なminimaxリスクの評価方法に関しては元論文[62]を参照されたい．

10. まとめ

本稿では，学習理論における基本的な技法を紹介した．汎化誤差の評価にはRademacher複雑さが有用であり，さらにそれを抑えるための道具として被覆数とDudley積分が有用であった．仮説集合の複雑さとしてVC次元は分布に依らない指標を与え，VC次元が有限であればどのような分布においても汎化誤差が収束することを見た．また，ロス関数が強凸な場合，局所Rademacher複雑さの概念を用いることで $O(1/\sqrt{n})$ より速い収束レート（fast learning rate）を示せることを紹介した．最後にminimaxリスクの下限の導出方法を簡単な回帰問題において紹介した．

本稿では詳しくは述べられなかったが，ノンパラメトリックBayesにおけるリスクの評価も盛んに研究されている[9, 10, 20–22, 35, 36, 56–58]．経験誤差最小化だけでなく，Bayes推定の学習理論も今後さらなる発展を見せるだろう．今後，ビッグデータ時代の到来とともに，高次元大標本における学習手法および統計的手法の性質の解明がますます重要になってくると考えられる．そこでは，本稿で紹介したような技法は強力なツールとなりえるだろう．一方で，数理統計学の理論はその長い歴史の中で多くの明快にして強力な結果を残し，大きな成功を収めてきた．今後も学習理論の発展には数理統計学から学ぶことは多いであろう．より強力な帰結を得るために学習理論にはますますの発展が期待される．

謝辞 鈴木大慈はJSPS科研費25730013（若手B）の助成を受けています．

参考文献

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

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

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

[4] P. L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, 8:775–790, 2007.

[5] C. Bennett and R. Sharpley. *Interpolation of Operators.* Academic Press, Boston, 1988.

[6] 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.

[7] 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.

[8] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 169–207, Berlin/Heidelberg, 2004. Springer.

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

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

[11] C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., N. Y., 1991.

[13] A. Dalalyan, Y. Ingster, and A. B. Tsybakov. Statistical inference in compound functional models, 2012. arXiv:1208.6402.

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

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

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

[17] R. M. Dudley. *Uniform Central Limit Theorems.* Cambridge University Press, 1999.

[18] M. Eberts and I. Steinwart. Optimal learning rates for least squares SVMs using Gaussian kernels. In *Advances in Neural Information Processing Systems 25*, 2012.

[19] R. M. Fano. *Transmission of Information: A Statistical Theory of Communication.* MIT Press, 1961.

[20] 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.

[21] 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.

[22] J. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics.* Springer, 2003.

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

[24] B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models, 2012. arXiv:1208.1211.

[25] M. Kloft and G. Blanchard. On the convergence rate of ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 13:2465–2501, 2012.

[26] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, 2011.

[27] M. Kloft, U. Rückert, and P. L. Bartlett. A unifying view of multiple kernel learning. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2010.

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

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

[30] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of the Annual Conference on Learning Theory*, pages 229–238, 2008.

[31] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *The Annals*

of Statistics, 38(6):3660–3695, 2010.

[32] G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[33] M. Ledoux and M. Talagrand. *Probability in Banach Spaces. Isoperimetry and Processes.* Springer, New York, 1991. MR1102015.

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

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

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

[37] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

[38] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48:1977–1991, 2002.

[39] 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.

[40] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[41] P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

[42] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer, 2008.

[43] 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.

[44] T. Suzuki. Fast learning rate of non-sparse multiple kernel learning and optimal regularization strategies, 2011. arXiv:1111.3781.

[45] T. Suzuki. Unifying framework for fast learning rate of non-sparse multiple kernel learning. In *Advances in Neural Information Processing Systems 24*, pages 1575–1583, 2011. NIPS2011.

[46] 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).

[47] T. Suzuki. Fast learning rate of non-sparse multiple kernel learning and optimal regularization strategies. *The Annals of Statistics*, 2013. to appear.

[48] 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).

[49] M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

[50] M. Talagrand. *The generic chaining.* Springer, 2000.

[51] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

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

[53] A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, 2008.

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

[55] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 1998.

[56] 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, 2008.

[57] 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, 2008. IMS Collections.

[58] 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.

[59] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York, 1996.

[60] 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.

[61] R. C. Williamson, A. J. Smola, and B. Shöelkof. Entropy numbers, operators and support vector kernels. In *Advances in Kernel Methods : Support Vector Learning*, pages 127–144. MIT Press, 1998.

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

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

鈴木 大慈 (正会員) 〒113-8656 東京都文京区本郷 7-3-1

2009 年東京大学大学院情報理工学系研究科において博士号（情報理工学）を取得. 2009 年より東京大学大学院情報理工学系研究科数理情報学専攻助教. 現在に至る.

(2013年3月3日受付)

(2013年5月30日最終稿受付)