数理統計学 (Mathematical Statistics)

Yugo Nakayama

2026 年 2 月 9 日

目次

第 I 部

数理統計学 (Mathematical Statistics)

第 1 章

確率空間と確率変数

統計学とは、観測されたデータ（確率変数の実現値）から、その背後にある確率分布の性質を推論する学問である。その議論を数学的に厳密に行うためには、まず「確率とは何か」「確率変数とは何か」を集合論と測度論の言葉で定義する必要がある。

本章では、現代確率論の基礎となる**確率空間（Probability Space）**と、その上に定義される**確率変数（Random Variable）**の構造を明らかにする。

1.1　確率空間の公理的定義

我々が「確率」と呼ぶものは、標本空間（起こりうる事象の全体）の部分集合に対して、0 から 1 の値を割り当てる「関数」である。しかし、すべての部分集合に確率を定義しようとすると数学的な矛盾（バナッハ＝タルスキーのパラドックス等）が生じるため、確率を定義できる「計測可能な集合」を限定する必要がある。これが σ-加法族である。

定義 1.1 (σ-加法族 / σ-algebra)**.** 集合 Ω の部分集合族 $\mathcal{F} \subset 2^{\Omega}$ が以下の条件を満たすとき、$\mathcal{F}$ を Ω 上の **σ-加法族** と呼ぶ。

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$ （補集合について閉じている）
3. $A_1, A_2, \cdots \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ （可算和について閉じている）

$\mathcal{F}$ の要素を**事象 (Event)** と呼ぶ。「可算個の操作で閉じている」という性質は、極限操作（$n \to \infty$）を扱う統計学において不可欠である。

定義 1.2 (確率測度 / Probability Measure)**.** $(\Omega, \mathcal{F})$ を可測空間とする。集合関数 $P : \mathcal{F} \to \mathbb{R}$ が以下の公理（コルモゴロフの公理）を満たすとき、P を**確率測度**と呼ぶ。

1. **非負性**: 任意の $A \in \mathcal{F}$ に対し、$P(A) \geq 0$
2. **正規性**: $P(\Omega) = 1$
3. **完全加法性 (Countable Additivity)**: 互いに素な（$A_i \cap A_j = \emptyset, i \neq j$）事象列 $A_1, A_2, \cdots \in \mathcal{F}$ に対し、
$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

このとき、組 $(\Omega, \mathcal{F}, P)$ を**確率空間 (Probability Space)** と呼ぶ。

1.2 確率変数の可測性

統計学において主役となるのは、標本空間 Ω そのものではなく、そこから得られる数値データである。これを確率変数と呼ぶが、厳密には「単なる変数」ではなく「関数」である。

定義 1.3 (確率変数 / Random Variable)**.** 確率空間 $(\Omega, \mathcal{F}, P)$ 上の関数 $X : \Omega \to \mathbb{R}$ が、任意のボレル集合 $B \in \mathcal{B}(\mathbb{R})$ に対して

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}$$

を満たすとき、X を**確率変数**と呼ぶ（厳密には $\mathcal{F}$-可測関数）。

【解説：なぜ「可測性」が必要か】「確率変数 X の値が区間 $[a, b]$ に入る確率」を計算したいとする。それは $P(a \le X \le b)$ と書かれるが、これは定義に戻ると

$$P(\{\omega \in \Omega \mid a \le X(\omega) \le b\})$$

を計算することになる。この集合 $\{\omega \mid \dots\}$ が事象族 $\mathcal{F}$ に含まれていなければ、確率 P は定義されない。つまり、**「確率が計算できる」ことを保証する条件が可測性**である。

定理 1.4 (確率変数の生成する σ-加法族)**.** 確率変数 X に対し、

$$\sigma(X) = \{X^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\}$$

は Ω 上の σ-加法族となる。これを X が生成する σ-加法族と呼ぶ。$\sigma(X)$ は、「X の値を観測することで区別できる事象の全体」という情報的な意味を持つ。

1.3 分布関数と確率密度

確率変数の挙動は、元の確率空間 $(\Omega, \mathcal{F}, P)$ に立ち戻らなくとも、実数上の関数である分布関数によって完全に記述できる。

定義 1.5 (累積分布関数 / CDF)**.** 確率変数 X に対し、関数 $F_X : \mathbb{R} \to [0, 1]$ を

$$F_X(x) = P(X \le x) = P(\{\omega \in \Omega \mid X(\omega) \le x\})$$

と定義する。これを**累積分布関数**と呼ぶ。

定理 1.6 (分布関数の性質)**.** 任意の分布関数 $F(x)$ は以下の性質を持つ。

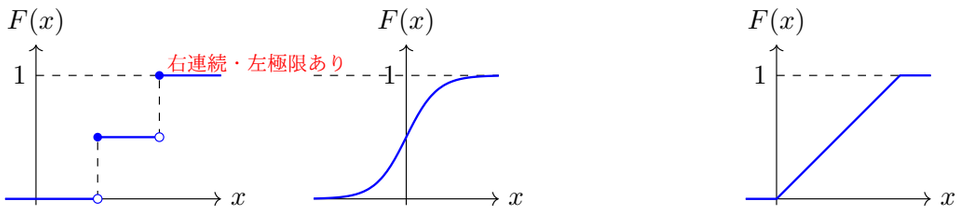
1. **単調非減少**: $x < y \implies F(x) \le F(y)$
2. **右連続性**: $\lim_{h \downarrow 0} F(x + h) = F(x)$
3. **境界条件**: $\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1$

定義 1.7 (確率密度関数 / PDF)**.** もし分布関数 $F_X(x)$ が絶対連続であるならば、ある非負可測関数 $f_X(x)$ が存在して

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

と書ける（ラドン＝ニコディムの定理）。この $f_X(x)$ を**確率密度関数**と呼ぶ。

【実社会・高次元統計への接続】通常の統計学では $f_X(x)$ の存在を前提とすることが多いが、高次元データや特異モデルにおいては、分布が絶対連続でない（密度関数を持たない）場合がある。例えば、**高次元空間に**



(a) 離散型（階段関数） (b) 連続型（滑らか） (c) 一様型（線形）

図 1.1　累積分布関数 (CDF) の典型例：定義より単調非減少・右連続・境界値 $0, 1$ を満たす。

おける低次元部分多様体上に分布するデータ（マニフォールド学習の想定）では、全空間に対する通常の密度関数は定義できない（ルベーグ測度 0 のため）。したがって、CDF $F_X(x)$ や確率測度 P そのものを扱う姿勢は、高次元データ解析においてより本質的となる。

第 1 章のまとめ

- 確率は、 σ-加法族上の測度として定義される。
- 確率変数は、可測空間から実数への可測関数である。
- 可測性は、「確率が計算可能であること」を保証する数理的要件である。

第 2 章

分布と期待値

本章では「確率変数の分布（測度）」と「期待値（積分）」を同一の枠組みで定義し、以後の大数・中心極限定理・推定理論の計算基盤を作ります。キーワードは、押し出し測度（law）とルベーグ積分としての期待値です。

2.1　分布（law）と押し出し測度

定義 2.1 (分布・押し出し測度)**.** 確率空間 $(\Omega, \mathcal{F}, P)$ 上の確率変数 $X : \Omega \to \mathbb{R}$ に対し、$\mathbb{R}$ 上の確率測度 P_X を

$$P_X(B) := P(X \in B) = P(X^{-1}(B)), \qquad B \in \mathcal{B}(\mathbb{R})$$

で定める。これを X の**分布**（law）または P の X による**押し出し測度**という。

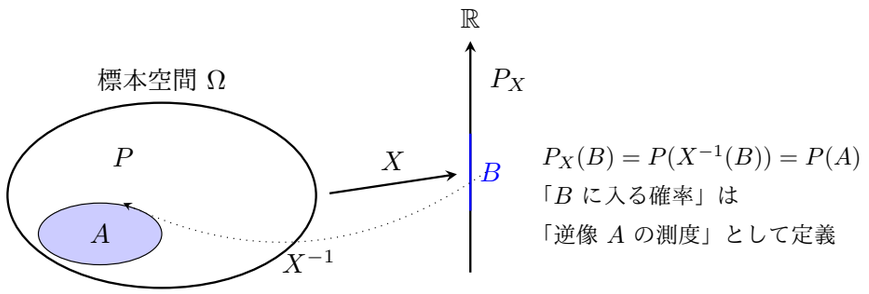


図 2.1　確率変数の分布（Law）と押し出し測度：元の空間 Ω の測度 P を X で $\mathbb{R}$ 上に移す。

定理 2.2 (分布は確率測度)**.** 上で定めた P_X は $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ 上の確率測度である。

証明. (1) 非負性：任意の B に対し $P_X(B) = P(X^{-1}(B)) \geq 0$。(2) 正規性：$P_X(\mathbb{R}) = P(X \in \mathbb{R}) = P(\Omega) = 1$。(3) 可算加法性：互いに素な $B_1, B_2, \ldots$ に対し $X^{-1}(B_n)$ も互いに素で、

$$P_X\Big(\bigcup_{n=1}^{\infty} B_n\Big) = P\Big(X^{-1}\big(\bigcup_{n=1}^{\infty} B_n\big)\Big) = P\Big(\bigcup_{n=1}^{\infty} X^{-1}(B_n)\Big) = \sum_{n=1}^{\infty} P(X^{-1}(B_n)) = \sum_{n=1}^{\infty} P_X(B_n).$$

よって確率測度である。

【補足（「モデル」としての分布）】統計学で「未知パラメータ θ を推定する」とは、多くの場合「観測 X の分布 P_θ（あるいは密度 p_θ）を仮定し、その θ をデータから同定する」ことに他ならない。

2.2 絶対連続性とラドン＝ニコディム微分

測度論において、二つの測度の関係性を記述する重要な概念に「絶対連続性」と「ラドン＝ニコディム微分」があります。これは確率密度関数の存在保証や、尤度比の定義に直結します。

直感的なイメージ：雪の深さとおにぎり

ある土地 Ω（標本空間）に対して、二つの「測り方」があるとします。

- 測度 μ：土地の「面積」を測るものさし。
- 測度 ν：その土地に積もった「雪の総量」を測るものさし。

このとき、各点 $x \in \Omega$ における「雪の深さ（密度）」を表す関数 $f(x)$ が存在すれば、

$$\nu(A) = \int_A f(x)\,\mu(dx)$$

のように、「雪の総量 ν」は「面積 $\mu\times$ 雪の深さ f」の積分で書けるはずです。この密度関数 f こそが、μ に対する ν の**ラドン＝ニコディム微分**であり、記号

$$f = \frac{d\nu}{d\mu}$$

で表されます。「測度同士の比率（微分）」という直感的な記法です。

ここで重要な前提条件があります。「面積がゼロの場所には、雪も積もらない」はずです。すなわち、

$$\mu(A) = 0 \implies \nu(A) = 0$$

という条件が必要です。これを「ν は μ に対して**絶対連続**である」といい、記号 $\nu \ll \mu$ で表します。

確率論での具体例

確率論や統計学の様々な場面で、この「測度の比」が登場します。

表 2.1　ラドン＝ニコディム微分の具体例

分野	記号 $\frac{d\nu}{d\mu}$	意味
確率密度関数	$f = \frac{dP}{d\lambda}$	ルベーグ測度（長さ）λ に対する確率 P の密度
ベイズ統計	$\frac{dP(\text{post})}{dP(\text{prior})}$	事前分布から事後分布への「重みの更新比率」
数理ファイナンス	$Z = \frac{dQ}{dP}$	実確率 P からリスク中立確率 Q への変換係数（状態価格密度）
尤度比	$\frac{dP_{\theta_1}}{dP_{\theta_0}}$	パラメータ θ_0 に対する θ_1 の「もっともらしさ」の比

計算のルール

微分の記法 $\frac{d\nu}{d\mu}$ を使うと、通常の分数のように直感的な計算が可能です。

- **連鎖律（チェインルール）**：$\nu \ll \mu \ll \lambda$ のとき

$$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu} \cdot \frac{d\mu}{d\lambda} \quad (\lambda\text{-a.e.})$$

- **逆数の関係**: μ と ν が互いに絶対連続（$\mu \sim \nu$）のとき

$$\frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu}\right)^{-1} \quad (\nu\text{-a.e.})$$

このように、測度の変換を「関数の積」に落とし込んで計算できるのが、ラドン＝ニコディム微分の強力な点です。

2.3 期待値：ルベーグ積分としての定義

定義 2.3 (非負可測関数の積分＝期待値の原型)**.** 非負可測関数 $Y : \Omega \to [0, \infty]$ に対し、単関数近似を用いて

$$\int_\Omega Y \, dP := \sup \left\{ \int_\Omega s \, dP \;\middle|\; 0 \le s \le Y,\ s \text{ は単関数} \right\}$$

で定義する。確率論ではこれを $\mathbb{E}[Y]$ と書く。

定義 2.4 (可積分性と一般の期待値)**.** 一般の可測関数 Y に対し、正負部分 $Y^+ = \max(Y, 0)$, $Y^- = \max(-Y, 0)$ を用いる。$\mathbb{E}[Y^+] < \infty$ かつ $\mathbb{E}[Y^-] < \infty$ のとき Y は可積分（$Y \in L^1(P)$）で、期待値を

$$\mathbb{E}[Y] := \mathbb{E}[Y^+] - \mathbb{E}[Y^-]$$

で定める。

定理 2.5 (線形性)**.** $Y, Z \in L^1(P)$、$a, b \in \mathbb{R}$ に対し $aY + bZ \in L^1(P)$ かつ

$$\mathbb{E}[aY + bZ] = a\mathbb{E}[Y] + b\mathbb{E}[Z]$$

が成立する。

証明. (非負の場合) まず $Y, Z \ge 0$ とし、単関数近似 $s_n \uparrow Y$, $t_n \uparrow Z$ をとる。単関数の段階では $\int (as_n + bt_n)\, dP = a\int s_n \, dP + b\int t_n \, dP$ が定義から従う。単調収束定理により

$$\int (aY + bZ)\, dP = \lim_{n\to\infty} \int (as_n + bt_n)\, dP = a \lim_{n\to\infty} \int s_n \, dP + b \lim_{n\to\infty} \int t_n \, dP = a\int Y \, dP + b\int Z \, dP.$$

(一般の場合) $Y = Y^+ - Y^-$, $Z = Z^+ - Z^-$ と分解し、各項に非負の場合を適用して整理すればよい。

定理 2.6 (LOTUS：分布による期待値計算)**.** 可測関数 $g : \mathbb{R} \to \mathbb{R}$ について $g(X) \in L^1(P)$ なら

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, P_X(dx)$$

が成り立つ（Law of the Unconscious Statistician）。

証明. まず $g \ge 0$ の単関数 $g = \sum_{k=1}^m c_k \mathbf{1}_{B_k}$（$c_k \ge 0$）に対し、

$$\mathbb{E}[g(X)] = \sum_{k=1}^m c_k\, \mathbb{E}[\mathbf{1}_{\{X \in B_k\}}] = \sum_{k=1}^m c_k\, P(X \in B_k) = \sum_{k=1}^m c_k\, P_X(B_k) = \int_{\mathbb{R}} g(x)\, P_X(dx).$$

次に一般の $g \ge 0$ は単関数列 $g_n \uparrow g$ で近似し、単調収束定理により等式を極限に渡して得る。最後に一般の g は $g = g^+ - g^-$ に分けて同様に示す。

【実社会イメージ（測度としての分布）】保険数理や品質管理では「平均損失」や「平均不良率」は $\mathbb{E}[g(X)]$ の形で現れるが、現場では X の“値の式”よりも「分布 P_X がどう歪むか（裾が重いか等）」のほうが本質になることが多い。この見方が、外れ値耐性（ロバスト性）やリスク管理（VaR/CVaR）に直結する。

2.4 分散・共分散と基本不等式

定義 2.7 (分散・共分散)**.** $X \in L^2(P)$（すなわち $\mathbb{E}[X^2] < \infty$）のとき、

$$\mathrm{Var}(X) := \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

と定める。さらに $X, Y \in L^2(P)$ に対して

$$\mathrm{Cov}(X, Y) := \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$$

と定める。

定理 2.8 (Cauchy - Schwarz と分散の非負性)**.** $X, Y \in L^2(P)$ なら

$$|\mathbb{E}[XY]| \le \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$$

が成り立ち、特に $\mathrm{Var}(X) \ge 0$ である。

証明. 任意の $t \in \mathbb{R}$ に対し $\mathbb{E}[(X + tY)^2] \ge 0$ より、

$$0 \le \mathbb{E}[X^2] + 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2]$$

は t の二次式であり判別式が非正：

$$(2\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \le 0.$$

よって $|\mathbb{E}[XY]| \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$。分散については $X - \mathbb{E}[X] \in L^2$ なので $\mathbb{E}[(X - \mathbb{E}[X])^2] \ge 0$。

定理 2.9 (Markov と Chebyshev)**.** (1) $Y \ge 0$, $\mathbb{E}[Y] < \infty$ なら任意の $a > 0$ で

$$P(Y \ge a) \le \frac{\mathbb{E}[Y]}{a}.$$

(2) $X \in L^2(P)$ なら任意の $t > 0$ で

$$P(|X - \mathbb{E}[X]| \ge t) \le \frac{\mathrm{Var}(X)}{t^2}.$$

証明. (1) $\mathbf{1}_{\{Y \ge a\}} \le Y/a$（$Y \ge 0$ より）なので期待値をとって

$$P(Y \ge a) = \mathbb{E}[\mathbf{1}_{\{Y \ge a\}}] \le \mathbb{E}[Y]/a.$$

(2) (1) を $Y = (X - \mathbb{E}[X])^2$, $a = t^2$ に適用する。

【実社会イメージ（「平均＋分散」だけで外れ確率を抑える）】 Chebyshev は分布の形を仮定せず、「平均と分散だけ」で大外れの確率に上界を与える。例えば製造ラインで、分布形状が時間変動しても、分散推定さえ安定なら最低限の安全側評価ができる。

2.5 モーメント母関数・特性関数

定義 2.10 (mgf と cf)**.** 確率変数 X に対し、モーメント母関数（mgf）を

$$M_X(t) := \mathbb{E}[e^{tX}]$$

（定義できる t の範囲で）とする。特性関数（cf）を

$$\varphi_X(t) := \mathbb{E}[e^{itX}], \qquad t \in \mathbb{R}$$

とする（ここで $i^2 = -1$）。

定理 2.11 (独立和の積：特性関数)**.** X, Y が独立なら

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

が成り立つ。

証明. 独立性より $\sigma(X)$ と $\sigma(Y)$ は独立であり、可測関数の積の期待値が分解する：

$$\varphi_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}e^{itY}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}] = \varphi_X(t)\varphi_Y(t).$$

【実社会イメージ（「畳み込み」を計算しないための道具）】ノイズが独立に足されるシステム（計測誤差、通信雑音、工程誤差）の合成分布は畳み込みになるが、特性関数に移すと積になる。この「空間を変えて計算を単純化する」発想は、フーリエ解析・信号処理・ランダム行列など多方面に連結する。

2.6 例題（手を動かす）

例題 2.12 (Bernoulli の期待値と分散)**.** $X \sim \mathrm{Bernoulli}(p)$（$P(X=1)=p, P(X=0)=1-p$）のとき、$\mathbb{E}[X]$ と $\mathrm{Var}(X)$ を求めよ。

解答. $\mathbb{E}[X] = 1\cdot p + 0\cdot(1-p) = p$。また $X^2 = X$ より $\mathbb{E}[X^2] = p$、従って $\mathrm{Var}(X) = p - p^2 = p(1-p)$。

例題 2.13 (Markov から Chebyshev を導け)**.** 定理 2.5 の (1) から (2) を導け。

解答. 上の証明の通り、$Y = (X - \mathbb{E}X)^2$ と $a = t^2$ を代入する。

例題 2.14 (LOTUS：密度がある場合の形)**.** X が密度 f を持つとき、$\mathbb{E}[g(X)]$ を f を用いて書け。

解答. $P_X(dx) = f(x)\,dx$ と見なせるので $\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x)\,dx$。

第 3 章

確率変数の収束

統計的推論の正当性は、サンプルサイズ n を無限大にしたときの極限挙動（漸近理論）によって保証される。しかし、確率変数 X_n は「関数」であるため、その収束には複数のモード（定義）が存在する。

本章では、**概収束・確率収束・法則収束**の厳密な定義とその階層構造を明らかにし、大数の法則（LLN）と中心極限定理（CLT）を証明する。

3.1　収束のモード（定義）

確率空間 $(\Omega, \mathcal{F}, P)$ 上の確率変数列 $X_1, X_2, \ldots$ と確率変数 X を考える。

定義 3.1 (概収束 / Almost Sure Convergence)**.** X_n が X に**概収束する**（$X_n \xrightarrow{a.s.} X$）とは、

$$P\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$$

が成立することをいう。これは「例外集合の測度が 0」という最強の収束である。

定義 3.2 (確率収束 / Convergence in Probability)**.** X_n が X に**確率収束する**（$X_n \xrightarrow{p} X$）とは、任意の $\epsilon > 0$ に対して

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

が成立することをいう。統計的推定量の**一致性（Consistency）**はこの概念である。

定義 3.3 (法則収束 / Convergence in Distribution)**.** X_n, X の分布関数を F_n, F とする。X_n が X に**法則収束する**（$X_n \xrightarrow{d} X$ または $X_n \xrightarrow{\mathcal{L}} X$）とは、$F$ のすべての連続点 x において

$$\lim_{n\to\infty} F_n(x) = F(x)$$

が成立することをいう。※確率変数の値そのものが近づくのではなく、「分布の形」が近づく最も弱い収束である。

3.2　収束の関係とボレル・カンテリの補題

定理 3.4 (収束の階層構造)**.** 以下の包含関係が成立する。

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

※逆は一般に成り立たない。ただし、X が定数の場合は $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$ が成立する。

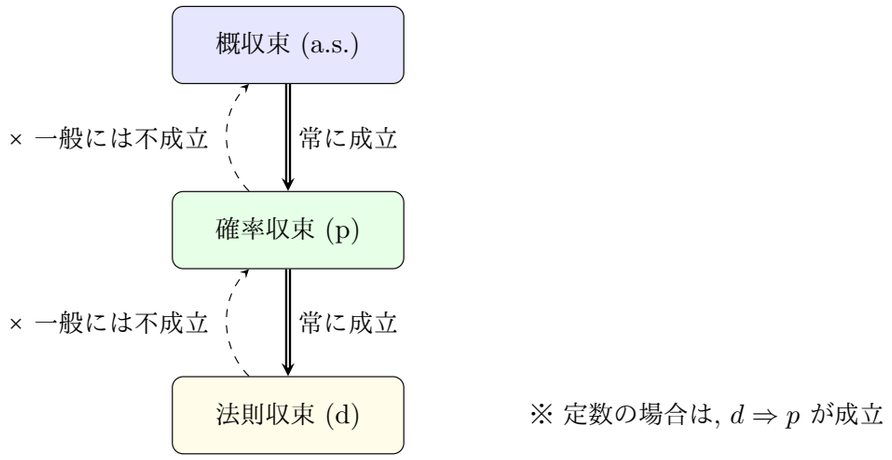


図 3.1　収束の階層構造：強さの順序

定理 3.4 の証明（Implications）

- **(1) 概収束 $\implies$ 確率収束** $(X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X)$

  証明. 概収束の定義より、イベント $A = \{\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$ の確率は $P(A) = 1$ である。これは、「任意の $\epsilon > 0$ に対して、ある N が存在し、すべての $n \geq N$ で $|X_n - X| < \epsilon$ となる」ような ω の集合が確率 1 であることを意味する。集合列を用いて書くと、

  $$P\left(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{|X_n - X| < \epsilon\}\right) = 1$$

  これは、補集合（収束しない集合）の確率が 0 であることと同値である：

  $$P\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{|X_n - X| \geq \epsilon\}\right) = 0$$

  確率測度の連続性（単調減少列の極限）より、

  $$\lim_{m\to\infty} P\left(\bigcup_{n=m}^{\infty} \{|X_n - X| \geq \epsilon\}\right) = 0$$

  ここで、単純に $n = m$ の項だけを見れば $\{|X_m - X| \geq \epsilon\} \subset \bigcup_{n=m}^{\infty} \{|X_n - X| \geq \epsilon\}$ であるから、

  $$P(|X_m - X| \geq \epsilon) \leq P\left(\bigcup_{n=m}^{\infty} \{|X_n - X| \geq \epsilon\}\right) \to 0 \quad (m \to \infty)$$

  よって $X_n \xrightarrow{p} X$ が成り立つ。

- **(2) 確率収束 $\implies$ 法則収束** $(X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X)$

  証明. F を X の分布関数とし、x を F の連続点とする。任意の $\epsilon > 0$ に対し、以下の包含関係が成り立つ：

  $$\{X_n \leq x\} \subseteq \{X \leq x + \epsilon\} \cup \{|X_n - X| > \epsilon\}$$
  $$\{X \leq x - \epsilon\} \subseteq \{X_n \leq x\} \cup \{|X_n - X| > \epsilon\}$$

  これらより確率の不等式を作ると、

  $$P(X_n \leq x) \leq F(x + \epsilon) + P(|X_n - X| > \epsilon)$$

$$F(x-\epsilon) - P(|X_n - X| > \epsilon) \leq P(X_n \leq x)$$

$n \to \infty$ とすると、$P(|X_n - X| > \epsilon) \to 0$ なので、

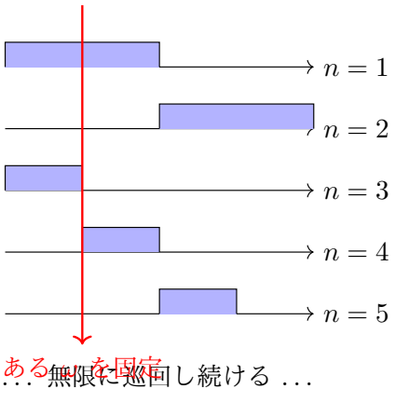
$$F(x-\epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F(x+\epsilon)$$

$\epsilon \downarrow 0$ とすると、x は F の連続点なので $F(x-\epsilon) \to F(x)$ かつ $F(x+\epsilon) \to F(x)$。したがって $\lim_{n\to\infty} F_{X_n}(x) = F(x)$ となり、法則収束する。

逆が成り立たない反例（Counterexamples）

■(1) 確率収束するが、概収束しない例 $(X_n \xrightarrow{p} 0 \not\Longrightarrow X_n \xrightarrow{a.s.} 0)$ 「**タイプライター数列（Typewriter Sequence）**」と呼ばれる有名な例です。区間 $[0,1]$ 上の一様分布を考えます。確率変数 X_n を、区間内を巡回する「幅が縮小する矩形波」として定義します。

- **定義:**
  - $X_1 = \mathbf{1}_{[0,1/2]}$
  - $X_2 = \mathbf{1}_{[1/2,1]}$
  - $X_3 = \mathbf{1}_{[0,1/4]}, X_4 = \mathbf{1}_{[1/4,2/4]}, \ldots$
  - 一般に、区間 $[0,1]$ を 2^k 等分した区間を順に動いていく指示関数とする。



どんな ω も無限回 “当たり” (1) になる, $\Rightarrow$ 0 に収束しない

図 3.2　タイプライター数列：確率は減るが、当たりが逃げ回る

- **性質（図 3.2 参照）:**
  - **確率収束する** $(X_n \xrightarrow{p} 0)$**:** 各 n で「当たり」の区間幅（確率）は $1/2, 1/4, \ldots$ と 0 に収束するため。
  - **概収束しない** $(X_n \not\xrightarrow{a.s.} 0)$**:** どのような ω を固定しても、区間が無限に巡回してくるため、無限回 $X_n(\omega) = 1$ となり 0 に収束しない。

■(2) 法則収束するが、確率収束しない例 $(X_n \xrightarrow{d} X \not\Longrightarrow X_n \xrightarrow{p} X)$ 法則収束は分布（値の散らばり具合）しか見ておらず、確率変数の値そのものの近さは問わないため、容易に反例が作れます。

- **例 1（符号反転）:** $X \sim N(0,1)$（標準正規分布）とする。$X_n = -X$ と置く。正規分布の対称性より $-X$ も $N(0,1)$ に従うため、分布関数は完全に一致し $X_n \xrightarrow{d} X$ である。しかし、$|X_n - X| = |-X - X| = |2X|$ であり、$X \neq 0$ a.s. なので

$$P(|X_n - X| > \epsilon) = P(|2X| > \epsilon) > 0$$

となり、確率収束しない。

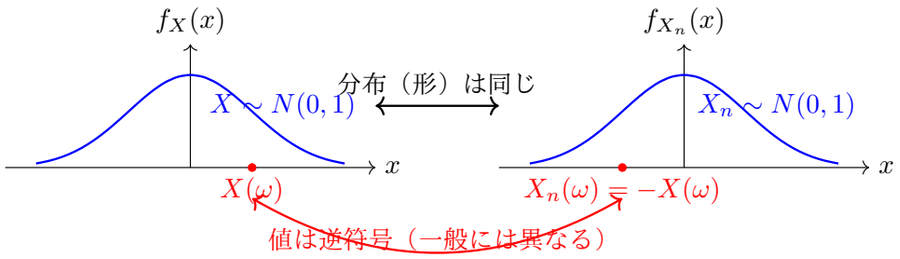


図 3.3　分布収束するが確率収束しない例 ($X_n = -X$)

- **例 2（独立なコピー）：** $X, X_1, X_2, \ldots$ をすべて独立で、$P(X=0) = 1/2, P(X=1) = 1/2$ であるベルヌーイ分布に従うとする。分布は同じなので当然 $X_n \xrightarrow{d} X$。しかし、差の絶対値を考えると

$$P(|X_n - X| = 1) = P(X_n = 1, X = 0) + P(X_n = 0, X = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

  となり、0 に収束しないため $X_n \xrightarrow{p} X$ ではない。

■【直感的なイメージ】クラスのテストと生徒

- **法則収束（分布収束）：** あるクラス（例えば 1 年 A 組）のテストの点数分布を毎年見ている状況。毎年平均点は 60 点でヒストグラムの形も同じであれば、統計データとしては「収束」している（分布は変わらない）。
- **確率収束せず：** しかし、去年の出席番号 1 番の生徒（80 点）と今年の出席番号 1 番の生徒（30 点）は別人である。分布（クラス全体）は似ていても、個々の生徒（ω）レベルで見れば値は一致していない。

■【補足】定数の場合　画像にある通り、収束先が定数 c （つまり $X = c$ a.s.）の場合に限り、法則収束から確率収束が言えます。$X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$。これは、定数への収束の場合、分布が一点に集中していくことと、値がその点に集まることが同値になるためです。

概収束の判定には、事象列の上極限（lim sup）を用いた以下の補題が強力な武器となる。

定理 3.5 (ボレル・カンテリの第一補題)**.** 事象列 $\{A_n\}$ に対し、$\sum_{n=1}^{\infty} P(A_n) < \infty$ ならば、

$$P\left(\limsup_{n\to\infty} A_n\right) = P(A_n \text{ i.o.}) = 0$$

である。ここで A_n i.o. (infinitely often) は「A_n が無限回起こる」事象を表す。

証明. $\limsup A_n = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n$ である。任意の k に対し、確率の劣加法性より

$$P\left(\bigcup_{n=k}^{\infty} A_n\right) \le \sum_{n=k}^{\infty} P(A_n)$$

仮定より級数は収束するので、右辺は $k \to \infty$ で 0 に収束する。測度の連続性より

$$P\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n\right) = \lim_{k\to\infty} P\left(\bigcup_{n=k}^{\infty} A_n\right) = 0$$

【実社会・高次元統計への接続】高次元統計学（例えば Lasso の変数選択一致性）では、「誤った変数を選択する確率」が $n \to \infty$ で 0 になることを示すが、単に $P(\text{error}) \to 0$ だけでなく、その**減衰スピード**が重要になる。減衰が $1/n^2$ や e^{-n} のように速ければ、級数が収束し、ボレル・カンテリの補題により「有限回のミスの後に、永続的に正解し続ける（概収束）」ことが保証される。

3.3 大数の法則 (LLN)

標本平均 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ が真の平均 μ に収束することを保証する定理。

定理 3.6 (大数の弱法則 / WLLN)**.** $X_1, X_2, \ldots$ が独立同分布（i.i.d.）で、平均 μ、分散 $\sigma^2 < \infty$ を持つとする。このとき、

$$\bar{X}_n \xrightarrow{p} \mu$$

が成立する。

証明 (チェビシェフの不等式による)**.** 期待値の線形性より $\mathbb{E}[\bar{X}_n] = \mu$。独立性より $\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$。任意の $\epsilon > 0$ に対し、チェビシェフの不等式（定理 2.5）を適用すると

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

$n \to \infty$ で右辺は 0 に収束する。

※より強い条件なしで $\bar{X}_n \xrightarrow{a.s.} \mu$ を示すのが**大数の強法則（SLLN）**である。

3.4 中心極限定理 (CLT)

「誤差の積み重ねがなぜ正規分布になるのか」を説明する、統計学で最も美しい定理。

3.5 スラツキーの定理 (Slutsky's Theorem) と応用

法則収束と確率収束を組み合わせた演算に関する重要な定理である。漸近理論において、推定量の複雑な変換後の分布を導出するために必須となる。

定理 3.7 (スラツキーの定理)**.** $X_n \xrightarrow{d} X$ かつ $Y_n \xrightarrow{p} c$ （c は定数）ならば、以下が成立する。

1. **和:** $X_n + Y_n \xrightarrow{d} X + c$
2. **積:** $X_n Y_n \xrightarrow{d} cX$
3. **商:** $Y_n \xrightarrow{p} c \neq 0$ ならば、$X_n / Y_n \xrightarrow{d} X/c$

スラツキーの定理の演算例 10 選

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(0,1)$ とし、$\bar{X}_n$ を標本平均、S_n^2 を標本分散とする。以下の事実を既知とする：

$$\sqrt{n}\bar{X}_n \xrightarrow{d} N(0,1), \quad \bar{X}_n \xrightarrow{p} 0, \quad S_n^2 \xrightarrow{p} 1$$

1. **基本形 1（和）:** $\bar{X}_n + S_n^2 \xrightarrow{d} 0 + 1 = 1$
2. **基本形 2（積）:** $\bar{X}_n S_n^2 \xrightarrow{d} 0 \cdot 1 = 0$
3. **基本形 3（商）:** $\bar{X}_n / S_n^2 \xrightarrow{d} 0/1 = 0$
4. **t 統計量（最も重要）:**

$$t_n = \frac{\sqrt{n}\bar{X}_n}{S_n} = \frac{\sqrt{n}\bar{X}_n}{\sqrt{S_n^2}} \xrightarrow{d} \frac{N(0,1)}{\sqrt{1}} = N(0,1)$$

これは t 分布がサンプルサイズ大で標準正規分布に近似することを示す。

5. **標本分散の変換:** $n(\bar{X}_n^2 + S_n^2 - 1) \xrightarrow{d} 0$

6. **分散安定化変換:** $\sqrt{n}\sin(\bar{X}_n) \xrightarrow{d} \sqrt{n}\sin(0) = 0$ （デルタ法の一種）
7. **比率推定量:** $\frac{\bar{X}_n+1}{S_n^2+2} \xrightarrow{d} \frac{0+1}{1+2} = \frac{1}{3}$
8. **漸近正規性:** $\sqrt{n}(\bar{X}_n^2) = (\sqrt{n}\bar{X}_n)\cdot\bar{X}_n \xrightarrow{d} Z\cdot 0 = 0$
9. **尤度比統計量:** $-2\log\Lambda_n \approx n\bar{X}_n^2/S_n^2 \xrightarrow{d} \chi_1^2/1 = \chi_1^2$
10. **多変量版:** $(\bar{X}_n, S_n^2)^\top \xrightarrow{d} (0,1)^\top$

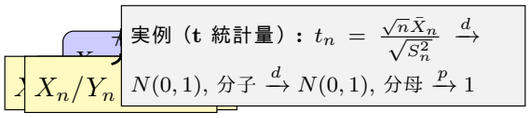


図 3.4　スラツキーの定理のイメージ：一方が定数に確率収束すれば、和・積・商の法則収束が保たれる。

3.6　中心極限定理 (CLT) Continued

定理 3.8 (リンドバーグ・レヴィの中心極限定理)**.** $X_1, X_2, \ldots$ が i.i.d. で、平均 μ、分散 $\sigma^2 < \infty$ を持つとする。$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ と置くと、

$$Z_n \xrightarrow{d} N(0,1)$$

が成立する。

証明 (特性関数によるアプローチ)**.** 一般性を失わず $\mu = 0, \sigma = 1$ とする（標準化後の変数 $Y_i = (X_i - \mu)/\sigma$ を考えればよい）。目標は $Z_n = \frac{1}{\sqrt{n}}\sum X_i$ の特性関数 $\varphi_{Z_n}(t)$ が、標準正規分布の特性関数 $e^{-t^2/2}$ に各点収束することを示すことである（レヴィの連続性定理）。

X_i の特性関数を $\varphi(t)$ とする。$\mathbb{E}[X] = 0, \mathbb{E}[X^2] = 1$ なので、Taylor 展開により

$$\varphi(t) = \mathbb{E}[e^{itX}] = 1 + it\mathbb{E}[X] + \frac{(it)^2}{2!}\mathbb{E}[X^2] + o(t^2) = 1 - \frac{t^2}{2} + o(t^2)$$

独立性より、和の特性関数は積になるから

$$\varphi_{Z_n}(t) = \mathbb{E}\left[\exp\left(it\frac{1}{\sqrt{n}}\sum X_i\right)\right] = \left(\varphi\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Taylor 展開を代入すると

$$\varphi_{Z_n}(t) = \left(1 - \frac{(t/\sqrt{n})^2}{2} + o\left(\frac{t^2}{n}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n$$

有名な極限公式 $\lim_{n\to\infty}(1 + \frac{c}{n})^n = e^c$ において $c = -t^2/2$ とみなせば、

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = e^{-t^2/2}$$

これは標準正規分布 $N(0,1)$ の特性関数である。レヴィの連続性定理により、$Z_n \xrightarrow{d} N(0,1)$ が示された。

3.7　デルタ法 (Delta Method)

中心極限定理は $\bar{X}_n$ の漸近分布を与えるが、実務では $\bar{X}_n^2$ や $\log\bar{X}_n$ など、推定量の関数の分布を知りたいことが多い。これを可能にするのがデルタ法である。

定理 3.9 (デルタ法)**.** $\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ とし、$g(u)$ を θ の近傍で微分可能な関数とする。このとき、

$$\sqrt{n}(g(U_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2)$$

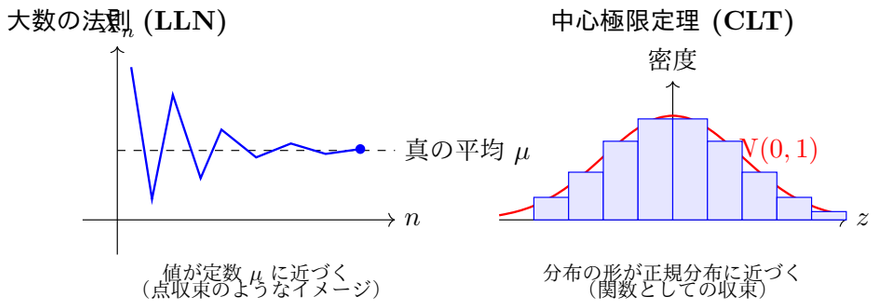


図 3.5　LLN と CLT のイメージの違い：LLN は「値の収束」、CLT は「分布形の収束」を主張する。

が成立する。

直感的には、テイラー展開 $g(U_n) \approx g(\theta) + g'(\theta)(U_n - \theta)$ により、誤差が $g'(\theta)$ 倍に拡大縮小されるため、分散は $[g'(\theta)]^2$ 倍になる。

デルタ法の演算例 10 選

$X_1, \ldots, X_n \overset{i.i.d.}{\sim} (\mu, \sigma^2)$ とし、$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ を出発点とする。

表 3.1　実務で頻出するデルタ法の適用例

推定対象	変換関数 $g(x)$	漸近分散 (Asymptotic Variance)
1. 分散推定 (基礎)	x^2	$4\mu^2\sigma^2$
2. 標準偏差	$\sqrt{x}$	$\sigma^2/(4\mu)$
3. 対数変換 (相対誤差)	$\log x$	σ^2/μ^2
4. 逆数 (レート)	$1/x$	σ^2/μ^4
5. オッズ比	$x/(1+x)$	$\sigma^2/(1+\mu)^4$
6. 対数オッズ	$\log(x/(1-x))$	$\sigma^2/[\mu(1-\mu)]^2$
7. 指数変換 (ハザード)	e^x	$e^{2\mu}\sigma^2$
8. Fisher の Z 変換	$\frac{1}{2}\log\frac{1+x}{1-x}$	$1/(1-\rho^2)^2$ (相関係数の場合)
9. 平方根変換 (Poisson)	$\sqrt{x}$	$1/4$ (分散安定化)
10. 尤度比統計量	$-2\log x$	$4\sigma^2$

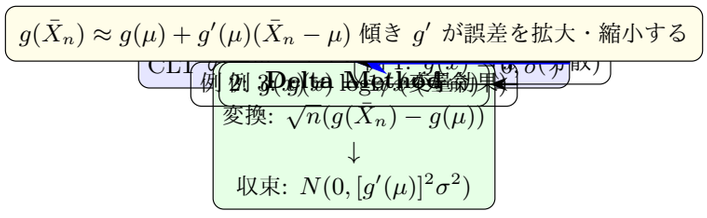


図 3.6　デルタ法のイメージ：一次近似により正規性が保存され、分散が勾配の二乗倍になる。

3.8　高次元統計の新漸近枠組み：HDLSS (High Dimension, Low Sample Size)

古典的な漸近理論（$n \to \infty, d$ 固定）が崩壊する現代の高次元データ解析において、新しい極限理論が必要とされている。それが **HDLSS (High Dimension, Low Sample Size)** 枠組みである。

3.8.1 HDLSS 設定の定義と問題点

定義 3.10 (HDLSS 漸近枠組み)**.** 次元 d とサンプルサイズ n が共に無限大に発散し、その比がある正の定数に収束する状況を考える。

$$n \to \infty, \quad \frac{d}{n} \to \gamma \in (0, \infty)$$

これは、ゲノムデータ（$d \approx 10^4, n \approx 10^2$）や画像解析などの状況を数理的にモデル化したものである。

この枠組みでは、古典的な統計学の常識（一致性や漸近正規性）が成立しない「次元の呪い」が顕著に現れる。

現象 1: 標本共分散行列の崩壊 (Marchenko-Pastur 則)

標本共分散行列 $\hat{\Sigma}$ の固有値分布は、真の固有値分布に収束せず、広がりを持った **Marchenko-Pastur (MP) 分布** に従うことが知られている。例えば、真の共分散が単位行列 I であっても、標本固有値は区間 $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$ に広く分布してしまうため、通常の主成分分析 (PCA) は機能しない。

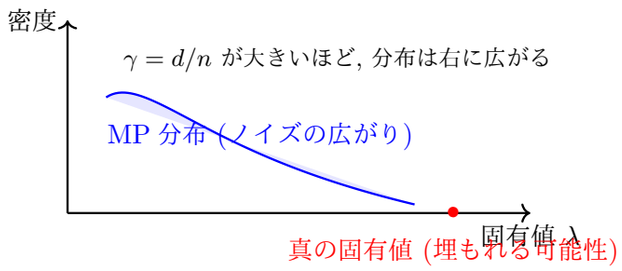


図 3.7 Marchenko-Pastur 分布のイメージ：ノイズだけで固有値が広がってしまう現象

現象 2: 最尤推定量の不一致性 (Neyman-Scott 問題)

パラメータ数 d が n と同オーダーで増えるため、最尤推定量 (MLE) は一致性を持たないことがある。

$$X_i \sim N(\theta_i, 1), \quad i = 1, \ldots, d \quad \text{(各次元独立)}$$

このとき、$\|\hat{\theta}_{MLE} - \theta\|^2/d \xrightarrow{p} 1 \neq 0$ となり、誤差が消えない。これに対処するために、Lasso などのスパース推定（9 章参照）や、Cross-Data-Matrix 法などの新手法が必要となる。

3.8.2 HDLSS 向けの新手法・理論

- **Noise-Reduction (NR) Methodology:** 標本共分散行列に含まれるバイアス（ノイズ成分）を幾何学的な性質を用いて除去し、真の固有構造を復元する手法（青嶋・矢田らによる研究）。これにより、条件 $\frac{d}{n} \to \infty$ の超高次元下でも PCA やクラスタリングが可能になる。
- **Cross-Data-Matrix Methodology:** データを分割して独立な共分散行列の積 $S_{12} = \frac{1}{n}\sum X_i^{(1)}(X_i^{(2)})^\top$ を考えることで、対角成分以外のノイズを相殺させる手法。
- **スパース推定 (Lasso / RE 条件):** 高次元空間の中でも、真に重要な変数は少数（スパース）であると仮定し、正則化項を加えることで推定を安定化させる。これは第 9 章、第 10 章で詳述する。

図 3.8　古典漸近論 vs HDLSS 漸近論：次元の呪いとその克服

第 4 章

統計的モデルと十分性

統計学では、観測データは「ある未知パラメータ θ を持つ確率分布からの標本」として生成されると仮定し、その θ を推論する。このとき、推論に本質的な情報を**落とさずに**データを要約する概念が「十分統計量」であり、指数型分布族と並んで数理統計学の中核をなす。

4.1　統計的モデルと尤度

定義 4.1 (統計的モデル)**.** 可測空間 $(\mathcal{X}, \mathcal{A})$ 上の確率測度族

$$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$$

を**統計的モデル**という。ここで Θ はパラメータ空間であり、観測 X はある θ のもとで $X \sim P_\theta$ に従うとする。

定義 4.2 (支配測度と密度)**.** ある σ-有限測度 μ が存在して、すべての θ について $P_\theta \ll \mu$（絶対連続）なら、モデルは μ により**支配される**という。このときラドン＝ニコディム微分

$$p_\theta(x) = \frac{dP_\theta}{d\mu}(x)$$

を（μ に関する）**密度**と呼ぶ。

定義 4.3 (尤度)**.** 観測値 $x \in \mathcal{X}$ が得られたとき、

$$L(\theta; x) := p_\theta(x)$$

を θ の関数として見たものを**尤度関数**という（定数倍の不定性は本質ではない）。

実社会のイメージ：逆問題としての推定センサー観測（例：温度、材料特性、信号波形）は「真の状態 θ があって、雑音を通して X が出る」と見なせる。尤度 $L(\theta; x)$ は「その θ ならこのデータがどれくらい自然か」を定量化するスコアである。

4.2　十分統計量：情報を落とさない要約

直感的に、十分統計量とは「データ X を圧縮した値 $T(X)$ だけ残しても、θ の推論に必要な情報が失われない」ような要約である。これを図形的に表現すると以下のようになる。

69: 70: **たとえ話（クイズ大会）**: 71: 参加者 100 人のクイズ正解率 θ を知りたいとき、「誰がいつ正解したか」という詳細な全履歴データ X は不要である。「合計何人が正解したか」という数字 $T(X)$ さえあれば、θ の推論には十分である。個々の正解のタイミングや順序は θ にとって「ノイズ」に過ぎない。72: 73:

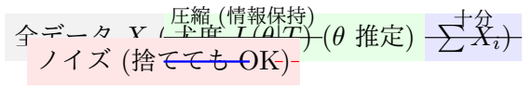


図 4.1　十分統計量のイメージ：情報は $T(X)$ に凝縮され、ノイズは捨てられる。

定義 4.4 (十分性：条件付き分布による定義)**.** 統計量 $T : \mathcal{X} \to \mathcal{T}$ に対し、$\sigma(T)$ を T が生成する σ-加法族とする。T が θ に関して**十分**であるとは、（正則条件付き分布が存在するとして）条件付き分布

$$\mathcal{L}(X \mid T)$$

が θ に依存しないこと、すなわち任意の可測集合 $A \in \mathcal{A}$ について

$$P_\theta(X \in A \mid T) = g(A, T) \quad (\theta \text{ に依らない})$$

が成り立つことをいう。

コメントこの定義は最も概念的だが、実際の判定には次節の「因子分解定理」が便利である。

4.3　因子分解定理（Fisher - Neyman）

以下では支配測度 μ があり密度 p_θ が存在する（支配された）モデルを仮定する。

定理 4.5 (因子分解定理)**.** 統計量 $T(X)$ が θ に関して十分であることと、ある非負可測関数 $g_\theta(\cdot), h(\cdot)$ が存在して

$$p_\theta(x) = g_\theta(T(x))\, h(x) \quad (\mu\text{-a.e.})$$

と因子分解できることは同値である。

証明. **(⇒)** T が十分であると仮定する。支配されたモデルでは、条件付き密度（あるいは条件付き測度）を用いて

$$p_\theta(x) = p_\theta(x \mid T(x))\; p_\theta^{T}(T(x))$$

という分解を期待したくなる（ここで p_θ^T は T の分布の密度）。十分性により $p_\theta(x \mid T)$ の θ 依存性が消えるので、ある $h(x)$ が存在して

$$p_\theta(x \mid T) = h(x) \quad (\theta \text{ に依らない})$$

と書け、残りは T のみに依存する項 $g_\theta(T(x)) := p_\theta^T(T(x))$ に吸収できる。よって

$$p_\theta(x) = g_\theta(T(x))h(x)$$

が得られる。

(⇐) 因子分解

$$p_\theta(x) = g_\theta(T(x))h(x)$$

を仮定する。任意の可測集合 $A \in \mathcal{A}$ と任意の（十分に良い）集合 $B \subset \mathcal{T}$ に対し

$$P_\theta(X \in A,\, T \in B) = \int_{A \cap T^{-1}(B)} p_\theta(x)\, \mu(dx) = \int_{A \cap T^{-1}(B)} g_\theta(T(x))h(x)\, \mu(dx).$$

ここで $A \cap T^{-1}(B)$ 上では $T(x) \in B$ なので、$g_\theta(T(x))$ は T の値だけに依存する「外側の重み」として働く。この構造により、条件付き確率 $P_\theta(X \in A \mid T)$ は $h(x)$ のみにより決まり、θ に依存しない（形式的には、$\sigma(T)$ 上の正則条件付き分布を構成して確認できる）。よって T は十分である。

実社会のイメージ：無損失圧縮十分統計量 $T(X)$ は、推定・検定に関して「情報を落とさないデータ圧縮」になっている。分散計算や最適化を高速化したい現場では、まず X を $T(X)$ に落としてから推論を回すのが定石になる。

4.4 指数型分布族と十分統計量

定義 4.6 (指数型分布族)**.** $\mathbb{R}^k$ 値統計量 $T(x)$ と関数 $h(x)$、自然パラメータ $\eta(\theta) \in \mathbb{R}^k$、正規化関数 $A(\theta)$ が存在して

$$p_\theta(x) = h(x) \exp\Big(\eta(\theta)^\top T(x) - A(\theta)\Big)$$

と書けるとき、$\{P_\theta\}$ を（k 次元の）**指数型分布族**という。

定理 4.7 (指数型分布族では T が十分)**.** 指数型分布族では、上の表示に現れる $T(X)$ は θ に関して十分である。

証明. 密度を

$$p_\theta(x) = \underbrace{\exp\big(\eta(\theta)^\top T(x) - A(\theta)\big)}_{g_\theta(T(x))} \, \underbrace{h(x)}_{h(x)}$$

と因子分解できるので、因子分解定理より T は十分である。

4.5 例題（典型例を手で確認）

例題 4.8 (Bernoulli / Binomial)**.** $X_1, \dots, X_n$ を i.i.d. Bernoulli(p) とし、観測 $x_1, \dots, x_n \in \{0, 1\}$ を得た。密度（確率質量関数）は

$$p_p(x_1, \dots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

ここで

$$T(x) = \sum_{i=1}^{n} x_i$$

とおけば

$$p_p(x) = g_p(T(x))\, h(x), \quad g_p(t) = p^t(1-p)^{n-t},\ h(x) \equiv 1.$$

よって因子分解定理より $T = \sum X_i$ は十分統計量である。

現場の意味：成功回数だけ分かれば、並び順などの詳細は p の推論には不要である（無損失要約）。

例題 4.9 (正規分布：平均未知・分散既知)**.** $X_1, \dots, X_n$ を i.i.d. $N(\mu, \sigma^2)$（σ^2 既知）とする。同時密度は

$$p_\mu(x) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right).$$

指数部を展開すると

$$\sum (x_i - \mu)^2 = \sum x_i^2 - 2\mu \sum x_i + n\mu^2,$$

よって

$$p_\mu(x) = \underbrace{\exp\left(\frac{\mu}{\sigma^2} \sum x_i - \frac{n\mu^2}{2\sigma^2}\right)}_{g_\mu(T(x))} \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum x_i^2\right)}_{h(x)}.$$

従って $T(x) = \sum_{i=1}^n x_i$（同値に $\bar{X}$）は μ に関して十分である。

例題 4.10 (正規分布：平均・分散とも未知)**.** $X_1, \dots, X_n$ を i.i.d. $N(\mu, \sigma^2)$ をとする。同時密度は

$$p_{\mu,\sigma^2}(x) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right).$$

上と同様に展開して

$$p_{\mu,\sigma^2}(x) = \underbrace{(\sigma^2)^{-n/2} \exp\left(\frac{\mu}{\sigma^2} \sum x_i - \frac{n\mu^2}{2\sigma^2} \right)}_{g_{\mu,\sigma^2}(T_1(x))} \underbrace{\exp\left(-\frac{1}{2\sigma^2} \sum x_i^2 \right)}_{\text{まだ } \sigma^2 \text{ が残る}}$$

となるが、$\sum x_i^2$ も統計量として入れれば

$$T(x) = \left(\sum_{i=1}^n x_i,\ \sum_{i=1}^n x_i^2 \right)$$

で因子分解が成立する。よってこの T は (μ, σ^2) に関して十分である。（実務上は同値な形として $(\bar{X},\ \sum (X_i - \bar{X})^2)$ を使う。分散推定・信頼区間の計算がこの 2 量だけで完結する。）

例題 4.11 (一様分布 $U(0,\theta)$)**.** $X_1, \dots, X_n \sim U(0,\theta)$ とする。密度関数は

$$p_\theta(x) = \frac{1}{\theta^n} \cdot \mathbf{1}_{\{0 \le \min x_i, \max x_i \le \theta\}}$$

これを因子分解すると

$$p_\theta(x) = \underbrace{\frac{1}{\theta^n} \mathbf{1}_{\{\max x_i \le \theta\}}}_{g_\theta(T(x))} \cdot \underbrace{\mathbf{1}_{\{0 \le \min x_i\}}}_{h(x)}$$

よって $T(X) = \max X_i$ は θ に関して十分統計量である。

4.6 十分統計量のまとめ

主な分布の十分統計量を以下の表にまとめる。これらはすべて「全データの情報を無損失圧縮」したものである。

表 4.1 代表的な分布の十分統計量

分布	十分統計量 $T(X)$	意味
Bernoulli(p)	$\sum X_i$	成功の総回数
Poisson(λ)	$\sum X_i$	事象の総発生数
Normal$(\mu, \sigma^2$ 既知$)$	$\bar{X}$ (または $\sum X_i$)	標本平均
Normal$(\mu, \sigma^2$ 未知$)$	$(\bar{X}, \sum (X_i - \bar{X})^2)$	平均と分散
Uniform$(0,\theta)$	$\max X_i$	観測された最大値

演習 4.1.

1. $X_1, \dots, X_n$ i.i.d. Poisson(λ) とする。$\sum X_i$ が λ に関して十分であることを因子分解定理で示せ。
2. 支配測度 μ を変えて密度表示を変えても、「十分性」という性質が不変であることを説明せよ（ヒント：尤度比を考える）。
3. 正規モデル $N(\mu, \sigma^2)$ において、$(\sum X_i, \sum X_i^2)$ と $(\bar{X}, \sum (X_i - \bar{X})^2)$ が互いに一対一変換であることを示せ。

第 5 章

推定理論：点推定の最適性

第 4 章で、我々はデータを情報の損失なく圧縮する「十分統計量」を手に入れた。本章では、その圧縮された情報を用いて、未知パラメータ θ の真の値を一点で言い当てる**点推定（Point Estimation）**の問題を扱う。

本章の構成と理論の流れを以下の図に示す。

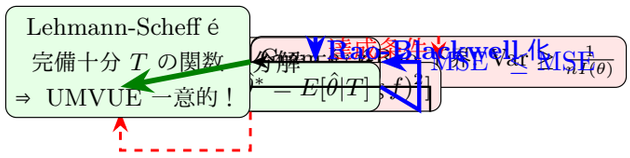


図 5.1　点推定の理論フロー：評価基準から最適化への道筋

推定量の良し悪しを測る基準（MSE、不偏性）を定義し、理論的に到達可能な精度の限界（クラメール・ラオの下界）を導出する。そして、その限界を達成する「最強の推定量（UMVUE）」を構成する方法論（ラオ・ブラックウェル化）を確立する。

5.1　推定量の評価基準

定義 5.1 (推定量と推定誤差)**.** 統計的モデル $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ において、データ X からパラメータ θ（またはその関数 $\tau(\theta)$）の値を推測する可測関数 $\hat{\theta}(X)$ を**推定量（Estimator）**と呼ぶ。実現値 $\hat{\theta}(x)$ を**推定値（Estimate）**と呼ぶ。

推定量 $\hat{\theta}$ の性能は、真のパラメータ θ との距離（損失）の期待値である**リスク関数**で評価される。最も代表的なのが平均二乗誤差である。

定義 5.2 (平均二乗誤差 / MSE)**.**

$$\mathrm{MSE}(\theta, \hat{\theta}) := \mathbb{E}_\theta\left[(\hat{\theta}(X) - \theta)^2\right]$$

定理 5.3 (バイアス・バリアンス分解)**.** MSE は「偏り（Bias）」の 2 乗と「分散（Variance）」の和に分解できる。

$$\mathrm{MSE}(\theta, \hat{\theta}) = \underbrace{(\mathbb{E}_\theta[\hat{\theta}] - \theta)^2}_{\mathrm{Bias}^2} + \underbrace{\mathrm{Var}_\theta(\hat{\theta})}_{\mathrm{Variance}}$$

証明. $\mu = \mathbb{E}_\theta[\hat{\theta}]$ と置く。

$$\begin{aligned}
\mathbb{E}[(\hat{\theta} - \theta)^2] &= \mathbb{E}[(\hat{\theta} - \mu + \mu - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \mu)^2] + 2(\mu - \theta)\underbrace{\mathbb{E}[\hat{\theta} - \mu]}_{0} + (\mu - \theta)^2 \\
&= \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias})^2
\end{aligned}$$

定義 5.4 (不偏推定量 / Unbiased Estimator)**.** すべての $\theta \in \Theta$ に対してバイアスが 0、すなわち

$$\mathbb{E}_\theta[\hat{\theta}(X)] = \theta$$

が成立するとき、$\hat{\theta}$ を**不偏推定量**と呼ぶ。不偏推定量のクラスの中で、分散を一様に最小にするものを **一様最小分散不偏推定量 (UMVUE)** と呼ぶ。

【実社会・高次元統計への接続】古典統計学では「不偏性」は美徳とされるが、高次元統計学（特に機械学習）では、あえてバイアスを受け入れて分散を劇的に減らす「縮小推定（Shrinkage Estimation）」が主流となる（例：Lasso, Ridge）。MSE の分解式は、この **Bias-Variance Trade-off** の数理的根拠である。

5.2 フィッシャー情報量とクラメール・ラオの下界

不偏推定量ならば分散はどこまでも小さくできるのか？ その「理論限界」を与えるのがクラメール・ラオの不等式である。これを導くために、情報の「量」を定義する。

定義 5.5 (スコア関数とフィッシャー情報量)**.** モデルの密度 $p_\theta(x)$ が θ に関して微分可能であるとする。対数尤度の勾配

$$V(X, \theta) := \frac{\partial}{\partial \theta} \log p_\theta(X) = \frac{\frac{\partial}{\partial \theta} p_\theta(X)}{p_\theta(X)}$$

を**スコア関数**と呼ぶ。スコア関数の分散

$$I(\theta) := \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log p_\theta(X) \right)^2 \right]$$

を**フィッシャー情報量（Fisher Information）**と呼ぶ。

補題 5.6 (スコアの性質)**.** 正則条件（積分と微分の交換が可能）の下で、

1. $\mathbb{E}_\theta[V(X, \theta)] = 0$
2. $I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log p_\theta(X) \right]$ （ヘッセ行列の期待値の符号反転）

定理 5.7 (クラメール・ラオの不等式 / Cramé r-Rao Lower Bound)**.** θ が 1 次元で、正則条件を満たすとする。$\hat{\theta}(X)$ を任意の不偏推定量とすると、以下の不等式が成立する。

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$$

ここで $I_n(\theta)$ はサンプルサイズ n の全データが持つフィッシャー情報量である（i.i.d. なら $nI_1(\theta)$）。

証明. 不偏性 $\mathbb{E}[\hat{\theta}] = \theta$ の両辺を θ で微分する（微分の交換を仮定）。

$$\frac{\partial}{\partial \theta} \int \hat{\theta}(x) p_\theta(x) d\mu(x) = 1$$

左辺に積の微分を実行：

$$\int \hat{\theta}(x) \frac{\partial p_\theta}{\partial \theta} d\mu = \int \hat{\theta}(x) \left(\frac{\partial \log p_\theta}{\partial \theta} \right) p_\theta(x) d\mu = \mathbb{E}[\hat{\theta} V] = 1$$

ここで、$\mathbb{E}[V] = 0$ より $\mathbb{E}[\hat{\theta} V] = \mathrm{Cov}(\hat{\theta}, V)$ である。Cauchy-Schwarz の不等式（定理 2.4）より

$$|\mathrm{Cov}(\hat{\theta}, V)|^2 \leq \mathrm{Var}(\hat{\theta}) \mathrm{Var}(V)$$

$\mathrm{Cov}(\hat{\theta}, V) = 1$、$\mathrm{Var}(V) = I_n(\theta)$ を代入すると

$$1^2 \leq \mathrm{Var}(\hat{\theta}) I_n(\theta) \implies \mathrm{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$$

【幾何学的解釈】情報幾何学において、フィッシャー情報量 $I(\theta)$ はモデル多様体の計量（リーマン計量）である。クラメール・ラオの不等式は、「推定の誤差（分散）は、統計的モデルの曲率（情報の密度）の逆数で制限される」という幾何学的な制約を示している。

5.3 ラオ・ブラックウェルの定理と UMVUE

クラメール・ラオの下界は強力だが、常に達成可能とは限らない。達成できない場合でも「ベストな推定量」を作る構成法が存在する。それが十分統計量への射影である。

定理 5.8 (ラオ・ブラックウェルの定理)**.** $\hat{\theta}$ をある不偏推定量とし、T を θ に対する十分統計量とする。このとき、条件付き期待値

$$\tilde{\theta}(T) := \mathbb{E}[\hat{\theta} \mid T]$$

は、以下の性質を持つ。

1. **不偏性**: $\mathbb{E}[\tilde{\theta}] = \theta$
2. **分散縮小**: $\mathrm{Var}(\tilde{\theta}) \leq \mathrm{Var}(\hat{\theta})$

等号成立は $\hat{\theta}$ が T の関数であるときに限る。

証明. (1) 重期待値の法則より $\mathbb{E}[\mathbb{E}[\hat{\theta}|T]] = \mathbb{E}[\hat{\theta}] = \theta$。(2) 分散の分解公式 $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathrm{Var}(Y|X)]$ を用いると、

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\tilde{\theta}) + \underbrace{\mathbb{E}[\mathrm{Var}(\hat{\theta} \mid T)]}_{\geq 0} \geq \mathrm{Var}(\tilde{\theta})$$

この定理は、「十分統計量 T を使っていない推定量には、改善の余地（ノイズ）が含まれている」ことを意味する。推定量は T の関数であるべきだ。

定理 5.9 (レーマン・シェフェの定理)**.** T が**完全十分統計量 (Complete Sufficient Statistic)** であり、$\tilde{\theta}(T)$ が不偏推定量であるならば、$\tilde{\theta}(T)$ は一様最小分散不偏推定量 (UMVUE) である。

完備十分統計量とは：ノイズゼロの要約

十分統計量 T が「完備（Complete）」であるとは、

$$\forall \theta, \quad \mathbb{E}_\theta[g(T)] = 0 \implies g(T) = 0 \quad (P_\theta\text{-a.s.})$$

が成り立つことを指す。これは「T の関数で、平均が 0 になるような余計な揺らぎ（ノイズ）が存在しない」という究極の情報を意味する。指数型分布族の自然十分統計量は、多くの場合この性質を満たす。

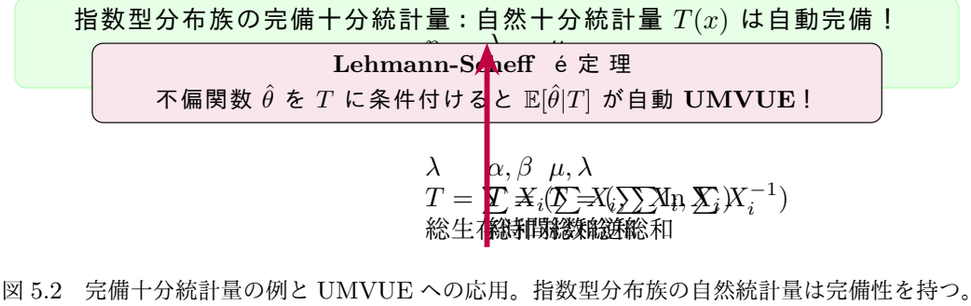


図 5.2 完備十分統計量の例と UMVUE への応用。指数型分布族の自然統計量は完備性を持つ。

5.4 例題（理論の実践）

例題 5.10 (正規分布の平均の UMVUE)**.** $X_1, \dots, X_n \sim N(\mu, \sigma^2)$（$\sigma^2$ 既知）とする。

1. μ のフィッシャー情報量を求めよ。
2. 標本平均 $\bar{X}$ が UMVUE であることを示せ。

解答. 1. 対数尤度は $l(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i - \mu)^2$。微分してスコア関数は $V = \frac{1}{\sigma^2}\sum(x_i - \mu) = \frac{n}{\sigma^2}(\bar{x} - \mu)$。2 階微分は $-\frac{n}{\sigma^2}$。よって $I_n(\mu) = -E[-\frac{n}{\sigma^2}] = \frac{n}{\sigma^2}$。

2. クラメール・ラオの下界は $\frac{1}{I_n(\mu)} = \frac{\sigma^2}{n}$。一方、$\bar{X}$ の分散は $\mathrm{Var}(\frac{1}{n}\sum X_i) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}$。下界を達成しているので、$\bar{X}$ は有効推定量（Efficient Estimator）であり、当然 UMVUE である。

例題 5.11 (一様分布と正則条件の破れ)**.** $X_1, \dots, X_n \sim U(0, \theta)$ とする。最大値 $X_{(n)} = \max X_i$ を考える。このモデルでは、定義域 $(0, \theta)$ がパラメータに依存するため、微分と積分の交換ができず、クラメール・ラオの下界は成立しない。実際、不偏化した推定量 $\hat{\theta} = \frac{n+1}{n}X_{(n)}$ の分散は $O(1/n^2)$ のオーダーで減少する（通常の正則モデルは $O(1/n)$）。これは「超有効（Super-efficiency）」と呼ばれる現象の一種であり、サポート境界に情報が集中していることを示唆する。

第 6 章

検定理論：仮説の最適選択

推定理論では未知パラメータの値を点や区間で求めたが、科学や工学の現場では「値そのもの」よりも、ある仮説（新薬に効果があるか、システムは正常か）が正しいか否かの**二値決定**を迫られることが多い。

本章では、統計的仮説検定を「決定理論（Decision Theory）」の一部として厳密に定式化する。直感的な判断ではなく、誤り確率（Type I/II Error）を制御しつつ、検出力（Power）を最大化する**最強検定（Most Powerful Test）**の存在とその構成法（ネイマン・ピアソンの補題）を導出する。

6.1　統計的仮説検定の定式化

定義 6.1 (統計的仮説)**.** パラメータ空間 Θ の互いに素な分割 $\Theta = \Theta_0 \cup \Theta_1$ に対し、

- **帰無仮説 (Null Hypothesis)** $H_0 : \theta \in \Theta_0$
- **対立仮説 (Alternative Hypothesis)** $H_1 : \theta \in \Theta_1$

を設定する。Θ_0 が 1 点のみからなる場合（例：$\theta = \theta_0$）を**単純仮説**、広がりを持つ場合（例：$\theta > \theta_0$）を**複合仮説**と呼ぶ。

定義 6.2 (検定関数と棄却域)**.** 観測データ x に基づき、H_0 を棄却する確率を与える関数 $\phi : \mathcal{X} \to [0, 1]$ を**検定関数 (Test Function)** と呼ぶ。

$$\phi(x) = \begin{cases} 1 & (\text{reject } H_0) \\ 0 & (\text{accept } H_0) \end{cases}$$

非決定論的な検定（$0 < \phi(x) < 1$）は**無作為化検定**と呼ばれるが、理論的整備のために重要である。$R = \{x \in \mathcal{X} \mid \phi(x) = 1\}$ を**棄却域 (Rejection Region)** と呼ぶ。

定義 6.3 (第一種・第二種の過誤と検出力)**.** 検定の良さは 2 種類の誤りで評価される。

1. **第一種の過誤 (Type I Error)**: H_0 が正しいのに棄却してしまう誤り（「冤罪」）。確率：$\alpha(\theta) = E_\theta[\phi(X)] \quad (\theta \in \Theta_0)$
2. **第二種の過誤 (Type II Error)**: H_1 が正しいのに受容してしまう誤り（「見逃し」）。確率：$\beta(\theta) = 1 - E_\theta[\phi(X)] \quad (\theta \in \Theta_1)$

検出力関数 (Power Function) を $\pi(\theta) = E_\theta[\phi(X)]$ と定義する。我々の目標は、**サイズ（有意水準）** α を一定以下に抑えつつ、検出力を最大化することである。

$$\sup_{\theta \in \Theta_0} E_\theta[\phi(X)] \leq \alpha \quad \text{かつ} \quad E_{\theta_1}[\phi(X)] \to \max$$

6.2 ネイマン・ピアソンの補題

単純仮説対単純仮説（$H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$）において、尤度比に基づく検定が「最強」であることを示す、検定理論の基本定理。

定理 6.4 (ネイマン・ピアソンの補題 / Neyman-Pearson Lemma)**.** P_0, P_1 をそれぞれの仮説の下での確率分布とし、密度 $p_0(x), p_1(x)$ を持つとする。有意水準 α に対し、以下の形の検定関数 ϕ^* を考える。

$$\phi^*(x) = \begin{cases} 1 & \text{if } p_1(x) > k p_0(x) \\ \gamma & \text{if } p_1(x) = k p_0(x) \\ 0 & \text{if } p_1(x) < k p_0(x) \end{cases}$$

ここで定数 $k \geq 0$ と $\gamma \in [0, 1]$ は、$E_{P_0}[\phi^*(X)] = \alpha$ となるように選ばれる。このとき、ϕ^* はレベル α の検定の中で**最強力検定 (Most Powerful Test, MP test)** である。すなわち、任意のレベル α の検定 ϕ に対し、$E_{P_1}[\phi^*(X)] \geq E_{P_1}[\phi(X)]$ が成立する。

証明. ϕ^* と、任意のレベル α 検定 ϕ（すなわち $E_{P_0}[\phi] \leq \alpha$）を考える。最大化したい量は検出力 $E_{P_1}[\phi]$ である。ここで、以下の積分を考える。

$$\Delta = \int (\phi^*(x) - \phi(x))(p_1(x) - k p_0(x)) d\mu(x)$$

被積分関数を x の領域ごとに評価する。

1. $p_1(x) > k p_0(x)$ **の領域**: $\phi^*(x) = 1$ なので、$\phi^*(x) - \phi(x) = 1 - \phi(x) \geq 0$。また $p_1(x) - k p_0(x) > 0$ なので、積は非負。
2. $p_1(x) < k p_0(x)$ **の領域**: $\phi^*(x) = 0$ なので、$\phi^*(x) - \phi(x) = -\phi(x) \leq 0$。また $p_1(x) - k p_0(x) < 0$ なので、積は非負（負×負）。
3. $p_1(x) = k p_0(x)$ **の領域**: 第 2 項が 0 なので、積は 0。

以上より、全領域で被積分関数は非負であり、積分値 $\Delta \geq 0$ である。これを展開すると

$$\int \phi^* p_1 - \int \phi p_1 - k \left(\int \phi^* p_0 - \int \phi p_0 \right) \geq 0$$

$$E_{P_1}[\phi^*] - E_{P_1}[\phi] \geq k(E_{P_0}[\phi^*] - E_{P_0}[\phi])$$

定義より $E_{P_0}[\phi^*] = \alpha$ かつ $E_{P_0}[\phi] \leq \alpha$ なので、右辺は $k(\alpha - (\leq \alpha)) \geq 0$。したがって $E_{P_1}[\phi^*] \geq E_{P_1}[\phi]$。

【実社会・高次元統計への接続】この定理は「尤度比こそが情報を最も効率的に使うスコアである」ことを示している。異常検知（Anomaly Detection）において、正常データの分布 p_0 と異常データの分布 p_1 が既知ならば、判定閾値は尤度比 $p_1(x)/p_0(x)$ で決めるのが数学的に最適である。機械学習の二値分類（ロジスティック回帰など）も、本質的にはこの尤度比の境界を学習しているに過ぎない。

6.3 一様最強力検定 (UMP) と単調尤度比

対立仮説が複合仮説 $H_1 : \theta > \theta_0$ の場合、すべての $\theta \in \Theta_1$ に対して最強となる検定を**一様最強力検定 (Uniformly Most Powerful Test, UMP test)** と呼ぶ。UMP 検定は常に存在するとは限らないが、「単調尤度比」という良い性質を持つモデルでは存在する。

定義 6.5 (単調尤度比 / MLR)**.** 1 次元パラメータ θ を持つモデル $\{p_\theta\}$ が、ある統計量 $T(x)$ に関して**単調尤度比 (Monotone Likelihood Ratio)** を持つとは、任意の $\theta_1 < \theta_2$ に対して、比

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$$

が $T(x)$ の単調非減少関数となることをいう。

定理 6.6 (カーリン・ルービンの定理 / Karlin-Rubin Theorem)**.** モデルが $T(x)$ に関して MLR を持つとする。仮説 $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$ に対する検定

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > c \\ \gamma & \text{if } T(x) = c \\ 0 & \text{if } T(x) < c \end{cases}$$

（ここで c, γ は $E_{\theta_0}[\phi] = \alpha$ で定まる）は、サイズ α の UMP 検定である。

解説: 指数型分布族（正規分布の平均、ベルヌーイ分布など）は自然パラメータに関して MLR を持つため、単純な「$T(x) > c$」という棄却域が常に最強となる。

6.4 尤度比検定と漸近理論（Wilks の定理）

複雑なモデル（多次元パラメータなど）では UMP 検定が存在しないことが多い。その場合、汎用的な手法として**尤度比検定 (Likelihood Ratio Test, LRT)** が用いられる。

定義 6.7 (尤度比統計量)**.**

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}$$

検定統計量として、通常 $-2 \log \lambda(x)$ を用いる。

定理 6.8 (ウィルクスの定理 / Wilks' Theorem)**.** 正則条件の下で、$n \to \infty$ のとき、帰無仮説 H_0 が正しければ

$$-2 \log \lambda(X) \xrightarrow{d} \chi^2_k$$

が成立する。ここで自由度 $k = \dim(\Theta) - \dim(\Theta_0)$ である（制約の数）。

証明の方針対数尤度を最尤推定量 $\hat{\theta}$ の周りで Taylor 展開し、スコア関数の漸近正規性とフィッシャー情報量の関係を用いる（第 7 章の漸近理論で詳細を扱う）。

Wilks の定理の直感的理解：「制約のペナルティ」

Wilks の定理は、「パラメータに制約を課すことで失われる尤度の量」が、漸近的にカイ二乗分布に従うことを示している。

実務的価値:

- **汎用性:** どんな複雑なモデルでも、MLE さえ計算できれば検定可能。
- **計算容易:** $-2 \log \lambda$ を計算し、χ^2 表と比較するだけ。
- **ネスト検定:** モデル選択（変数選択、次数選択）に直接応用可能。
- **多次元対応:** Wald 検定や Score 検定と並ぶ「三大漸近検定」の一つ。
- **情報幾何的解釈:** 統計多様体上の「測地線距離」に対応。

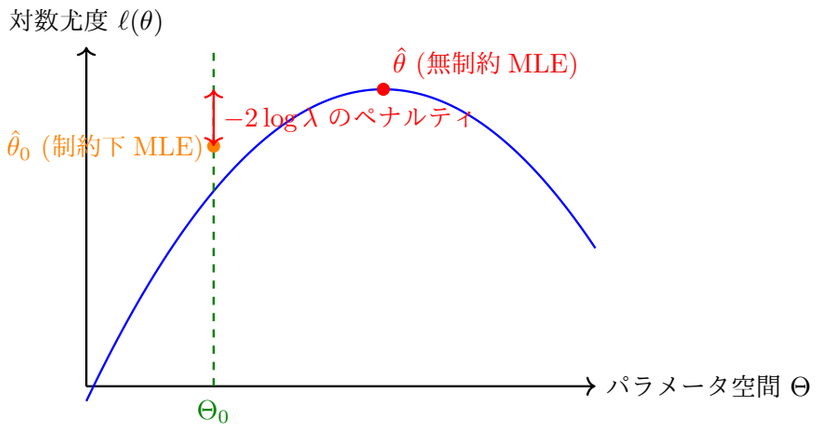


図 6.1　Wilks の定理の幾何学的イメージ：制約により失われる尤度が χ^2 分布に従う。

三大漸近検定の比較：Wald, Score, Wilks

大標本理論において、同じ帰無仮説を検定する 3 つの方法が存在し、いずれも漸近的に χ^2 分布に従う。

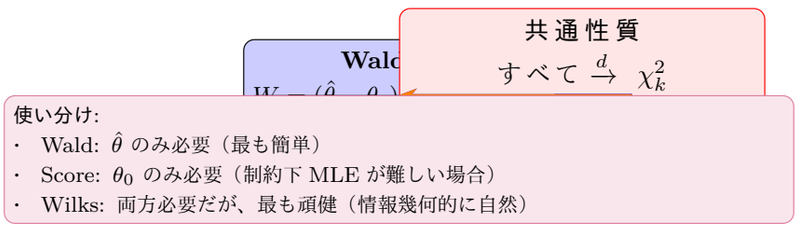


図 6.2　三大漸近検定の比較：異なる量を測るが、漸近的に同じ χ^2 分布に収束する。

6.5　P 値と多重比較の問題

P 値の正しい解釈

P 値（p-value）は「帰無仮説が正しいと仮定したとき、観測されたデータと同等かそれ以上に極端な結果が得られる確率」である。

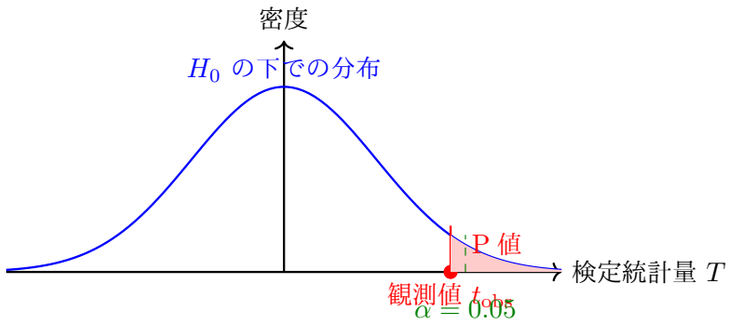


図 6.3　P 値の幾何学的意味：観測値より極端な領域の確率。

重要な注意:

- P 値 < 0.05 は「H_0 が間違っている確率が 95%」ではない！

- P 値は「データの極端さ」の指標であり、効果の大きさではない。
- 小さい P 値 $=$ 統計的有意 $\neq$ 実務的重要性

多重比較の罠と Bonferroni 補正

複数の検定を同時に行うと、偶然による「偽陽性」が急増する。

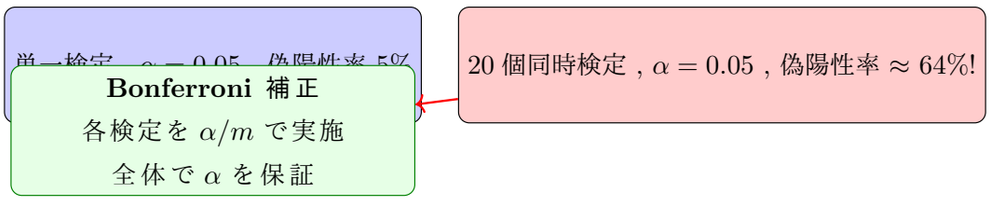


図 6.4　多重比較の問題：検定数が増えると偽陽性が急増する。Bonferroni 補正で制御可能。

6.6　演習問題

演習 6.1.

1. **正規分布の分散検定**: $X_1, \ldots, X_n \sim N(0, \sigma^2)$ とする。$H_0 : \sigma^2 = 1$ vs $H_1 : \sigma^2 = 2$ の最強力検定の棄却域が $\sum X_i^2 > c$ の形になることをネイマン・ピアソンの補題を用いて示せ。
2. **検出力の計算**: 上記の問題で $n = 10, \alpha = 0.05$ のとき、棄却限界値 c を χ^2 分布表から求め、対立仮説 $\sigma^2 = 2$ での検出力を計算式で表せ。
3. **トレードオフ**: サンプルサイズ n を固定したまま α を小さくすると、β（第二種の過誤）が増大することを図形的に説明せよ。

第 7 章

漸近理論：無限の彼方での真実

第 3 章で導入した収束概念（LLN, CLT）を、第 5・6 章で構築した推定・検定の文脈に適用する。有限のサンプルサイズ n での厳密な分布導出は、正規分布などごく一部のモデルでしか不可能である。しかし、$n \to \infty$ としたとき、推定量は驚くべき普遍的な挙動を示す。

本章では、**最尤推定量の漸近正規性**と、関数の変換を扱う**デルタ法**、そして第 6 章で保留した**ウィルクスの定理の証明**を与える。これらは現代統計解析の「近似計算」の理論的支柱である。

7.1　最尤推定量の漸近的性質

最尤推定量（MLE）$\hat{\theta}_n$ は、モデルが正則であれば、漸近的に「不偏」かつ「最小分散（有効）」となる。

前提条件（正則性条件）

パラメータ θ は開集合 $\Theta \subset \mathbb{R}$ 内にあり、密度 $p(x|\theta)$ は θ に関して 3 回微分可能、かつ積分と微分の交換が可能であるとする（Cramé r の正則条件）。

定理 7.1 (MLE の一致性 / Consistency)**.** フィッシャー情報量 $I(\theta)$ が正定値であるとき、尤度方程式の解 $\hat{\theta}_n$ は真値 θ_0 に確率収束する。

$$\hat{\theta}_n \xrightarrow{p} \theta_0$$

証明の指針（Wald のアプローチ）：対数尤度関数 $l_n(\theta) = \frac{1}{n}\sum \log p(X_i|\theta)$ は、大数の法則により期待値関数 $\mathbb{E}_{\theta_0}[\log p(X|\theta)]$ に各点収束する。この期待値関数はカルバック・ライブラー情報量により $\theta = \theta_0$ で一意に最大となる。一様収束性を仮定すれば、最大値を与える点も真値に収束する。

定理 7.2 (MLE の漸近正規性 / Asymptotic Normality)**.** $\hat{\theta}_n$ を一致性を持つ最尤推定量とする。

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$$

すなわち、MLE は漸近的に**クラメール・ラオの下界を分散として達成する（漸近有効性）**。

証明. 対数尤度の導関数（スコア関数）を $S_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log p(X_i|\theta)$ とする。最尤推定量は $S_n(\hat{\theta}_n) = 0$ を満たす。$S_n(\theta)$ を真値 θ_0 の周りで一次の Taylor 展開を行う（平均値の定理）。

$$0 = S_n(\hat{\theta}_n) = S_n(\theta_0) + S_n'(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

ここで $\tilde{\theta}_n$ は $\hat{\theta}_n$ と θ_0 の間の点である。式を変形して $\sqrt{n}$ を掛ける：

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\frac{1}{\sqrt{n}} S_n(\theta_0)}{\frac{1}{n} S_n'(\tilde{\theta}_n)}$$

分子の挙動: $S_n(\theta_0) = \sum U_i$（スコアの和）であり、$\mathbb{E}[U_i] = 0, \mathrm{Var}(U_i) = I(\theta_0)$。中心極限定理（CLT）より、

$$\frac{1}{\sqrt{n}} S_n(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$$

分母の挙動: $S_n'(\theta) = \sum \frac{\partial^2}{\partial \theta^2} \log p(X_i|\theta)$ である。大数の法則（LLN）より、$\frac{1}{n} S_n'(\theta_0) \xrightarrow{p} \mathbb{E}[p''/p] = -I(\theta_0)$。$\hat{\theta}_n \xrightarrow{p} \theta_0$ より $\tilde{\theta}_n \xrightarrow{p} \theta_0$ なので、連続写像定理により

$$\frac{1}{n} S_n'(\tilde{\theta}_n) \xrightarrow{p} -I(\theta_0)$$

全体の挙動: スラツキーの定理（Slutsky's Theorem：$X_n \to^d X, Y_n \to^p c \implies X_n/Y_n \to^d X/c$）を適用する。

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \frac{N(0, I(\theta_0))}{-I(\theta_0)} = N\left(0, \frac{I(\theta_0)}{I(\theta_0)^2}\right) = N\left(0, I(\theta_0)^{-1}\right)$$

7.2 デルタ法：関数の漸近分布

我々はしばしば θ そのものではなく、その関数 $g(\theta)$（例：オッズ比 $p/(1-p)$、変動係数 σ/μ）の推定に興味がある。

定理 7.3 (デルタ法 / Delta Method)**.** 数列 $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ と、微分可能な関数 $g : \mathbb{R} \to \mathbb{R}$（ただし $g'(\theta) \neq 0$）に対し、以下が成立する。

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2 \sigma^2)$$

証明. $g(T_n)$ を θ の周りで Taylor 展開する。

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + o_p(|T_n - \theta|)$$

$$\sqrt{n}(g(T_n) - g(\theta)) = g'(\theta) \underbrace{\sqrt{n}(T_n - \theta)}_{\xrightarrow{d} N(0,\sigma^2)} + \underbrace{\sqrt{n} o_p(|T_n - \theta|)}_{\xrightarrow{p} 0}$$

線形変換の性質より、極限分布は $g'(\theta) N(0, \sigma^2) = N(0, [g'(\theta)]^2 \sigma^2)$ となる。

例題 7.4. $X_i \sim \mathrm{Bernoulli}(p)$。オッズ $\psi = p/(1-p)$ の推定量を $\hat{\psi} = \hat{p}/(1-\hat{p})$ とする。その漸近分布を求めよ。

解答. $\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p(1-p))$。$g(p) = p/(1-p)$ とすると、$g'(p) = 1/(1-p)^2$。デルタ法より、分散は $[g'(p)]^2 \cdot p(1-p) = \frac{p}{(1-p)^3}$。よって $\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N\left(0, \frac{p}{(1-p)^3}\right)$。

7.3 ウィルクスの定理の証明

第 6 章の尤度比検定統計量 $-2 \log \lambda(X)$ がカイ二乗分布に従うことの証明。これにより、複雑なモデル比較（AIC/BIC の基礎）が可能になる。

定理 7.5 (ウィルクスの定理の再掲と証明)**.** 帰無仮説 $H_0 : \theta = \theta_0$（単純仮説）の下で、

$$-2 \log \lambda(X) = 2(l_n(\hat{\theta}) - l_n(\theta_0)) \xrightarrow{d} \chi_1^2$$

（多変量の場合は $\chi^2_{\dim \Theta}$）

証明. 対数尤度 $l_n(\theta_0)$ を MLE $\hat{\theta}$ の周りで 2 次まで Taylor 展開する。ここで $l_n'(\hat{\theta}) = 0$（極値条件）であることに注意する。

$$l_n(\theta_0) \approx l_n(\hat{\theta}) + (\theta_0 - \hat{\theta}) \underbrace{l_n'(\hat{\theta})}_{0} + \frac{1}{2}(\theta_0 - \hat{\theta})^2 l_n''(\hat{\theta})$$

移項して 2 倍すると

$$2(l_n(\hat{\theta}) - l_n(\theta_0)) \approx -(\hat{\theta} - \theta_0)^2 l_n''(\hat{\theta})$$

右辺を変形する：

$$= (\sqrt{n}(\hat{\theta} - \theta_0))^2 \cdot \left(-\frac{1}{n} l_n''(\hat{\theta}) \right)$$

1. 第 1 項は $(\sqrt{n}(\hat{\theta} - \theta_0))^2 \xrightarrow{d} [N(0, I^{-1})]^2 = I^{-1}\chi_1^2$ （※厳密には Z^2/I の形）。
2. 第 2 項は観測フィッシャー情報量であり、$\xrightarrow{p} I(\theta_0)$。

これらを合わせると、$I^{-1}\chi_1^2 \cdot I = \chi_1^2$。よって尤度比統計量は漸近的に自由度 1 のカイ二乗分布に従う。

7.4 高次元統計における「漸近理論の崩壊」

本章の理論はすべて $n \to \infty$ でパラメータ数 d が固定されていることを前提としている。しかし、現代のデータ解析（ゲノム、画像、LLM）では、データ数 n と共に次元 d も増大する（HDLSS: High Dimension Low Sample Size）。

HDLSS 設定（$d/n \to \gamma > 0$）でのパラドックス:

- サンプル共分散行列の固有値分布は、真の固有値に収束せず、**Marchenko-Pastur 則**に従って広がってしまう。
- MLE は一致性を持たず、バイアスが消えない（例：Neyman-Scott 問題）。
- 本章の χ^2 近似も成立しなくなる。

専門とする**スパース推定（Lasso）やランダム行列理論**は、この「古典的漸近理論が壊れた世界」で、いかにして新たな「集中現象（Concentration of Measure）」を見つけ出すかという試みである。次章以降では、その入り口となる数理を紹介する。

第 8 章

高次元確率論の基礎：集中現象とランダム行列への入口

高次元では「平均からのズレ」が指数関数的に抑えられる集中現象が支配的になり、統計理論の主役が極限定理から不等式へ移ります。本章では、Chernoff（指数型）評価 →Hoeffding/Bernstein→ε-net による行列ノルム評価、という**研究で使う最短経路**を数式で整備します。

8.1 Chernoff 法とサブガウシアン

定理 8.1 (Chernoff bound)**.** 確率変数 Z と任意の $t > 0$ に対し

$$P(Z \geq x) = P(e^{tZ} \geq e^{tx}) \leq e^{-tx}\,\mathbb{E}[e^{tZ}]$$

が成り立つ。

証明. $e^{tZ} \geq 0$ に Markov の不等式を適用する：

$$P(Z \geq x) = P(e^{tZ} \geq e^{tx}) \leq \frac{\mathbb{E}[e^{tZ}]}{e^{tx}}.$$

この 1 行が、以後の集中不等式の共通テンプレートになる。

定義 8.2 (サブガウシアン：mgf 型)**.** 平均 0 の確率変数 X が **sub-Gaussian** であるとは、ある $\sigma > 0$ が存在して任意の $t \in \mathbb{R}$ で

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right)$$

が成り立つことをいう。

このとき Chernoff を最適化すると、ガウス型の尾確率が得られる：

$$P(X \geq x) \leq \exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad x > 0.$$

例題 8.3 (標準正規)**.** $G \sim N(0,1)$ は $\mathbb{E}[e^{tG}] = e^{t^2/2}$ より sub-Gaussian（$\sigma = 1$）である。

8.2 Hoeffding 不等式（有界独立和）

補題 8.4 (Hoeffding の補題)**.** X が $a \leq X \leq b$ を満たし、$\mathbb{E}[X] = 0$ とする。すると任意の $t \in \mathbb{R}$ で

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

証明. 指数関数 e^{tx} は凸なので、任意の $x \in [a,b]$ に対し、端点を結ぶ線形補間で上から抑えられる：

$$e^{tx} \le \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}.$$

両辺の期待値をとると

$$\mathbb{E}[e^{tX}] \le \frac{b-\mathbb{E}[X]}{b-a}e^{ta} + \frac{\mathbb{E}[X]-a}{b-a}e^{tb} = \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb},$$

ここで $\mathbb{E}[X]=0$ を用いた。右辺を $a=-c, b=c$ となるよう平行移動しても上界は悪化しない（より厳密には、固定幅 $b-a$ のもとで右辺最大は中心化されたときに達成される）。よって $a=-c, b=c$ として示せば十分で、このとき

$$\mathbb{E}[e^{tX}] \le \frac{1}{2}(e^{-tc}+e^{tc}) = \cosh(tc) \le \exp\left(\frac{t^2c^2}{2}\right)$$

（$\cosh u \le e^{u^2/2}$ は級数比較で示せる）。$c=(b-a)/2$ を代入して結論：

$$\mathbb{E}[e^{tX}] \le \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

定理 8.5 (Hoeffding 不等式)**.** 独立な確率変数 $X_1,\dots,X_n$ が $a_i \le X_i \le b_i$ を満たし、$\mathbb{E}[X_i]=0$ とする。このとき任意の $x>0$ で

$$P\left(\sum_{i=1}^n X_i \ge x\right) \le \exp\left(-\frac{2x^2}{\sum_{i=1}^n (b_i-a_i)^2}\right).$$

証明 (Chernoff ＋補題)**.** $S=\sum X_i$ とおく。任意の $t>0$ で Chernoff より

$$P(S \ge x) \le e^{-tx}\mathbb{E}[e^{tS}] = e^{-tx}\prod_{i=1}^n \mathbb{E}[e^{tX_i}] \le e^{-tx}\prod_{i=1}^n \exp\left(\frac{t^2(b_i-a_i)^2}{8}\right) = \exp\left(-tx + \frac{t^2}{8}\sum (b_i-a_i)^2\right).$$

右辺を t で最小化する。二次式の最小は

$$t^\star = \frac{4x}{\sum (b_i-a_i)^2}.$$

これを代入して

$$P(S \ge x) \le \exp\left(-\frac{2x^2}{\sum (b_i-a_i)^2}\right).$$

Hoeffding Bound の完全証明と視覚的理解

Hoeffding の不等式は、有界独立変数の和が平均から大きく外れる確率を、ガウス型の tail で抑える強力な道具である。

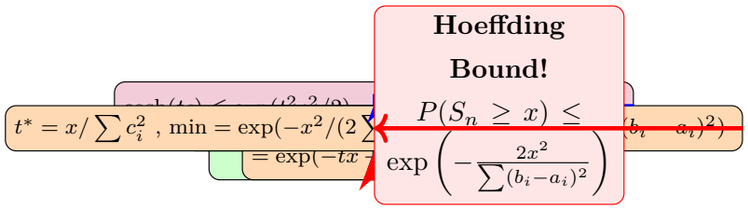


図 8.1　Hoeffding Bound の証明フロー：cosh 上界が心臓部、t 最適化で最強 bound を達成。

Hoeffding が「なぜ最強か？」の直感:

- **有界変数に特化**: Bernoulli, $[-1,1]$ 変数など、箱の幅 $(b-a)$ で統一制御。箱が狭いほど tail が速く減る。

- **次章 Bernstein への橋渡し**: Hoeffding は全有界変数に効く（緩い）、Bernstein は moment 条件で強化（きつい）。
- **Lasso 理論の土台**: 9 章で X の列が有界 $\Rightarrow$ Hoeffding で $\|X^T\varepsilon\|$ を制御 $\Rightarrow$ Lasso の Oracle 不等式の証明に必須！

実例（$n = 1000$, 変数 $\in [-1, 1]$）：

$$\text{Hoeffding: } P(\bar{X}_n \geq 0.6) \leq \exp(-200) \approx 10^{-87}, \quad \text{Markov: } P \geq 0.1 \text{ （桁違い！）}$$

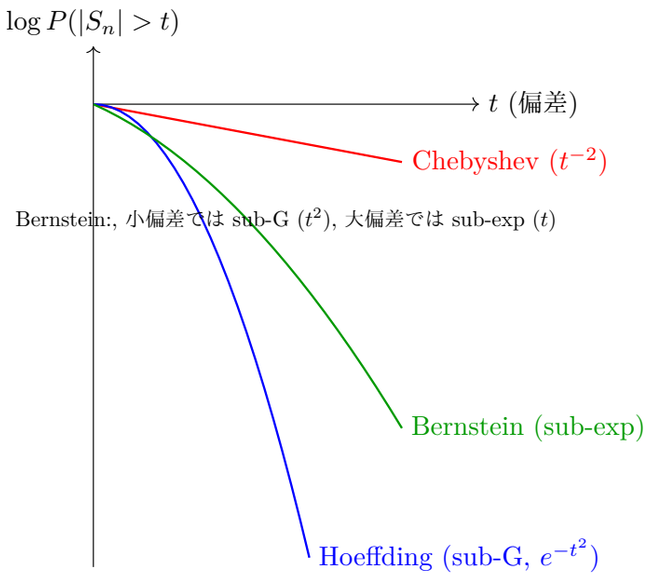


図 8.2　集中不等式の比較：Chebyshev（多項式減衰）に対し、指数型不等式（Hoeffding, Bernstein）は劇的に速く減衰する。Bernstein は 小偏差で正規分布的、大偏差で指数分布的な挙動をつなぐ。

実社会イメージ（多重比較と union bound）高次元では「d 個の誤差の最大値」を抑える必要があり、各座標に Hoeffding をかけて和で抑える（union bound）という戦略が頻出する。たとえば「全特徴量で同時に誤差 $\leq \epsilon$」を保証したいなら、片側確率を α/d に落として閾値を上げる設計になる。

8.3　Bernstein 不等式（分散も使う：サブ指数尾）

有界性は強すぎることが多い。二乗可積分で、尾が指数関数的に減るクラス（sub-exponential）では Bernstein 型が自然に出る。

定義 8.6 (sub-exponential：mgf 局所条件)**.** 平均 0 の X が **sub-exponential** であるとは、ある (ν, b) が存在して $|t| < 1/b$ の範囲で

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{\nu^2 t^2}{2}\right)$$

が成り立つことをいう。

定理 8.7 (Bernstein 不等式：代表形)**.** 独立な平均 0 の $X_1, \ldots, X_n$ が sub-exponential (ν_i, b_i) であるとする。$S = \sum X_i$、$\nu^2 = \sum \nu_i^2$、$b = \max b_i$ とおくと任意の $x > 0$ で

$$P(S \geq x) \leq \exp\left(-\frac{1}{2}\min\left(\frac{x^2}{\nu^2}, \frac{x}{b}\right)\right).$$

証明 (2 領域最適化)**.** $|t| < 1/b$ で

$$\mathbb{E}[e^{tS}] = \prod \mathbb{E}[e^{tX_i}] \le \prod \exp\left(\frac{\nu_i^2 t^2}{2}\right) = \exp\left(\frac{\nu^2 t^2}{2}\right).$$

Chernoff より

$$P(S \ge x) \le \exp\left(-tx + \frac{\nu^2 t^2}{2}\right), \qquad 0 < t < 1/b.$$

無制約最小は $t = x/\nu^2$。(i) もし $x/\nu^2 \le 1/b$ ならそれを採用して $\exp(-x^2/(2\nu^2))$。(ii) もし $x/\nu^2 > 1/b$ なら制約端 $t = 1/b$ を入れて

$$\exp\left(-\frac{x}{b} + \frac{\nu^2}{2b^2}\right) \le \exp\left(-\frac{x}{2b}\right)$$

(ここで $x > \nu^2/b$ を用いた)。両者をまとめて stated bound が得られる。

Matrix Bernstein Bound：Lasso 理論への架け橋

ランダム行列の和のスペクトルノルムを制御する Matrix Bernstein 不等式は、Lasso の Restricted Eigenvalue (RE) 条件の証明に不可欠である。

定理 8.8 (Matrix Bernstein Bound)**.** 独立な対称ランダム行列 $X_1, \dots, X_n \in \mathbb{R}^{d\times d}$ が $\mathbb{E}[X_i] = 0$ かつ $\|X_i\|_{\mathrm{op}} \le L$ (a.s.) を満たすとする。分散パラメータを

$$\sigma^2 = \left\| \sum_{i=1}^n \mathbb{E}[X_i^2] \right\|_{\mathrm{op}}$$

と定義すると、任意の $t > 0$ に対し

$$P\left(\left\| \sum_{i=1}^n X_i \right\|_{\mathrm{op}} \ge t \right) \le 2d \exp\left(-\frac{t^2/2}{\sigma^2 + Lt/3} \right).$$

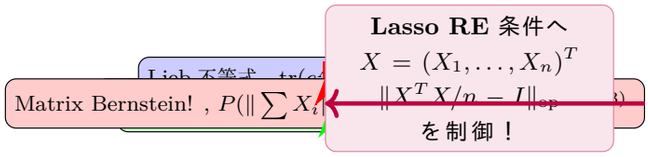


図 8.3　Matrix Bernstein の証明フロー：Lieb 不等式が心臓部、Union bound で次元依存 $2d$ が出現。Lasso RE 条件の証明に直結。

Matrix Bernstein の威力:

- **スカラー Bernstein の行列版**: 分散 σ^2 と有界性 L の両方を使う最適 bound。
- **次元依存**: $2d$ の係数が出現するが、$t \gg \sqrt{d}$ で exponential tail が効く。
- **Lasso 理論の土台**: 9 章で $X^T X/n$ のスペクトルノルムを制御 $\Rightarrow$ RE 条件成立 $\Rightarrow$ Oracle 不等式の証明完成！
- **ランダム行列理論への橋渡し**: Marchenko-Pastur 分布（3 章 HDLSS）の厳密版。

実例（Lasso 設計行列 $X \in \mathbb{R}^{n\times d}$, $d/n \to \gamma$）:

$$\text{Matrix Bernstein: } P\left(\left\| \frac{X^T X}{n} - I \right\|_{\mathrm{op}} \ge \epsilon \right) \le 2d \exp(-cn\epsilon^2)$$

これにより、$n \gg d \log d$ で $X^T X/n \approx I$ が高確率で成立 $\Rightarrow$ Lasso の RE 条件が満たされる！

8.4 ε-net とランダム行列（スペクトルノルム評価）

高次元統計の行列評価は、スカラー不等式＋離散化（net）で「ほぼ」片が付くことが多い。

補題 8.9 (ε-net の存在とサイズ)**.** 単位球面 $S^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ には、任意の $\varepsilon \in (0,1)$ に対し有限集合 $\mathcal{N}_\varepsilon \subset S^{d-1}$ が存在して

$$\forall u \in S^{d-1}\ \exists v \in \mathcal{N}_\varepsilon :\ \|u - v\|_2 \le \varepsilon, \qquad |\mathcal{N}_\varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^d.$$

証明 (体積比較)**.** 半径 $\varepsilon/2$ の開球を、互いに交わらないように球面上に最大個配置する（極大集合をとる）。その中心集合を $\mathcal{N}_\varepsilon$ とする。極大性より、半径 ε の球で球面全体を被覆する（被覆性）。一方、互いに交わらない半径 $\varepsilon/2$ 球は、半径 $1 + \varepsilon/2$ の球に含まれるので

$$|\mathcal{N}_\varepsilon| \cdot \mathrm{vol}(B(\varepsilon/2)) \le \mathrm{vol}(B(1 + \varepsilon/2)).$$

体積は半径の d 乗に比例するから

$$|\mathcal{N}_\varepsilon| \le \left(\frac{1 + \varepsilon/2}{\varepsilon/2}\right)^d = \left(1 + \frac{2}{\varepsilon}\right)^d.$$

補題 8.10 (対称行列のノルム：net 近似)**.** 対称行列 $A \in \mathbb{R}^{d \times d}$ と $\varepsilon \in (0, 1/2)$ に対し

$$\|A\|_{\mathrm{op}} = \sup_{\|u\|_2 = 1} |u^\top A u| \le \frac{1}{1 - 2\varepsilon} \sup_{v \in \mathcal{N}_\varepsilon} |v^\top A v|.$$

証明. 任意の $\|u\| = 1$ に対し、net より $\|u - v\| \le \varepsilon$ となる $v \in \mathcal{N}_\varepsilon$ をとる。$u = v + (u - v)$ を用いて

$$u^\top A u - v^\top A v = (u - v)^\top A u + v^\top A (u - v).$$

よって Cauchy - Schwarz と $\|u\| = \|v\| = 1$ から

$$|u^\top A u - v^\top A v| \le \|u - v\|\, \|A\|_{\mathrm{op}} \|u\| + \|v\|\, \|A\|_{\mathrm{op}} \|u - v\| \le 2\varepsilon \|A\|_{\mathrm{op}}.$$

したがって

$$|u^\top A u| \le |v^\top A v| + 2\varepsilon \|A\|_{\mathrm{op}}.$$

両辺で u 上の上限をとると

$$\|A\|_{\mathrm{op}} \le \sup_{v \in \mathcal{N}_\varepsilon} |v^\top A v| + 2\varepsilon \|A\|_{\mathrm{op}}.$$

整理して結論。

定理 8.11 (標本共分散のスペクトルノルム：ガウス例)**.** $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(0, I_d)$ とし

$$S = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$$

を標本共分散とする。すると任意の $\delta \in (0,1)$ に対し、ある普遍定数 $C > 0$ が存在して

$$\|S - I_d\|_{\mathrm{op}} \le C \left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n} \right)$$

が確率少なくとも $1 - \delta$ で成り立つ。

証明 (スカラー集中＋ net ＋ union bound)**.** 対称行列 $A = S - I_d$ とおく。補題 8.6 より、$\varepsilon = 1/4$ として

$$\|A\|_{\mathrm{op}} \le 2 \sup_{v \in \mathcal{N}} |v^\top A v| \quad (\mathcal{N} = \mathcal{N}_{1/4}).$$

各固定 $v \in S^{d-1}$ について、$v^\top X_i \sim N(0,1)$ なので $(v^\top X_i)^2$ は χ^2_1。よって

$$v^\top S v = \frac{1}{n} \sum_{i=1}^{n} (v^\top X_i)^2$$

は χ^2_1 の平均であり、$v^\top A v = v^\top S v - 1$ は平均 0 の sub-exponential（実際、$\chi^2_1 - 1$ は sub-exponential）として Bernstein 型の評価ができる。従ってある定数 $c > 0$ が存在して

$$P\big(|v^\top A v| \ge t\big) \le 2 \exp\big(-cn \min(t^2, t)\big).$$

次に union bound：

$$P\left(\sup_{v \in \mathcal{N}} |v^\top A v| \ge t \right) \le |\mathcal{N}| \cdot 2 \exp\big(-cn \min(t^2, t)\big).$$

補題 8.5 より $|\mathcal{N}| \le 9^d$。右辺を $\le \delta$ にするよう t を選ぶと

$$cn \min(t^2, t) \gtrsim d + \log(1/\delta).$$

したがって

$$t \lesssim \sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n}.$$

最後に $\|A\|_{\mathrm{op}} \le 2 \sup_{v \in \mathcal{N}} |v^\top A v|$ を戻して結論。

実社会イメージ（なぜ共分散推定が難しいか）この形は「誤差がおおむね $\sqrt{d/n}$」で効いてくることを意味し、d が n と同程度だと共分散推定が本質的に不安定になる。ゆえに高次元では、スパース性や低ランク性などの構造仮定（正則化）が必要になる。

8.5 演習問題

1. （Hoeffding）$X_i \in [-1, 1]$ 独立・平均 0 のとき、$\bar{X}_n$ の両側尾確率 $P(|\bar{X}_n| \ge \epsilon)$ を導け。
2. （Bernstein）$\chi^2_1 - 1$ が sub-exponential であることを、mgf を直接計算して示せ（$|t| < 1/2$ の範囲で評価せよ）。
3. （net）補題 8.6 の係数 $\frac{1}{1-2\varepsilon}$ が自然に出る理由を、証明中の不等式を丁寧に追って説明せよ。
4. （標本共分散）定理 8.7 の証明で「$(v^\top X_i)^2 - 1$ に Bernstein が使える」部分を、mgf の上界から完全に埋めよ。

第 9 章

スパース推定と Oracle 不等式（Lasso）

高次元（$p \gg n$）では最小二乗推定が不適切（不定・過学習）になりやすいので、「ℓ_1 正則化」により疎な解を選ぶ。ここでは線形回帰を主題に、Lasso の誤差評価（Oracle 不等式）を数学的に導く。

9.1　設定と Lasso の定義

設定（高次元線形回帰）

観測 (y, X) は

$$y = X\beta^\star + \varepsilon$$

で生成されるとする。ここで $y \in \mathbb{R}^n$、設計行列 $X \in \mathbb{R}^{n\times p}$、真の係数 $\beta^\star \in \mathbb{R}^p$、ノイズ $\varepsilon \in \mathbb{R}^n$ は平均 0（例えば独立 sub-Gaussian）とする。

真の疎性を

$$S := \mathrm{supp}(\beta^\star), \qquad s := |S|$$

で表す。

定義 9.1 (Lasso)**.** $\lambda > 0$ に対し、Lasso 推定量 $\hat\beta$ を

$$\hat\beta \in \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}$$

で定める。

Lasso の解析の要点は「凸最適化の最適性条件」から、推定誤差 $\Delta := \hat\beta - \beta^\star$ がある円錐（cone）に入ることを示し、そこに設計行列の下界条件（RE 条件など）を組み合わせて誤差上界を得る点にある。

9.2　最適性条件（KKT）とソフト閾値化

定理 9.2 (KKT 条件)**.** 目的関数

$$f(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

の最小解 $\hat\beta$ は、ある劣勾配ベクトル $\hat z \in \partial\|\hat\beta\|_1$ を用いて

$$-\frac{1}{n}X^\top(y - X\hat\beta) + \lambda\hat z = 0$$

を満たす。すなわち

$$\frac{1}{n}X^\top(y - X\hat\beta) = \lambda\hat z, \quad \hat z_j = \begin{cases} \mathrm{sign}(\hat\beta_j) & (\hat\beta_j \neq 0) \\ u,\ |u| \le 1 & (\hat\beta_j = 0) \end{cases}$$

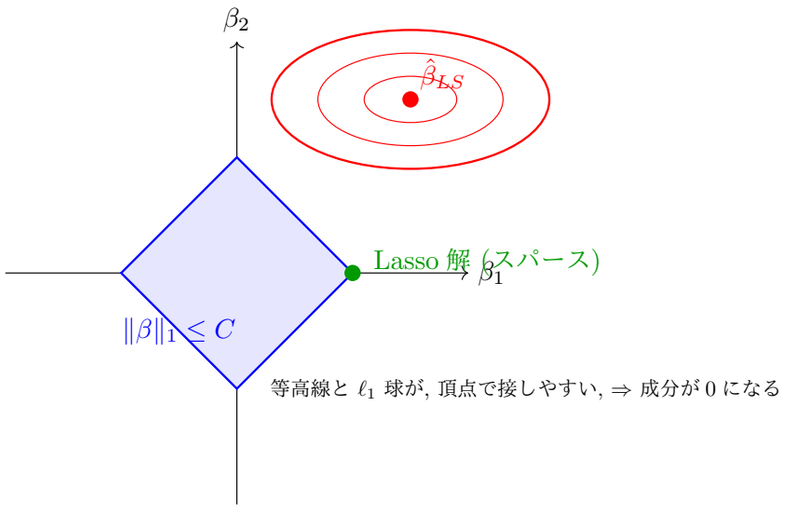


図 9.1　Lasso の幾何学的解釈：ℓ_1 制約領域（正方形）の「角」が等高線と接するため、スパースな解（いくつかの成分が厳密に 0）が得られやすい。

証明. 滑らかな項 $g(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2$ は $\nabla g(\beta) = -(1/n)X^\top(y - X\beta)$ を持つ。凸関数 $h(\beta) = \lambda\|\beta\|_1$ の劣勾配集合は $\partial h(\beta) = \lambda\partial\|\beta\|_1$。凸最適化の一階最適性条件 $0 \in \nabla g(\hat{\beta}) + \partial h(\hat{\beta})$ より結論を得る。

例題 9.3 (直交設計：ソフト閾値化)**.** もし $X^\top X = nI_p$（列が直交で正規化）なら、Lasso は成分ごとに解ける。目的関数を展開すると

$$\frac{1}{2n}\|y - X\beta\|^2 = \frac{1}{2n}\|y\|^2 - \frac{1}{n}\beta^\top X^\top y + \frac{1}{2}\|\beta\|^2$$

（定数項は無視できる）なので

$$\hat{\beta} = \arg\min_\beta \left\{ \frac{1}{2}\|\beta\|_2^2 - \beta^\top \underbrace{\left(\frac{1}{n}X^\top y\right)}_{=:u} + \lambda\|\beta\|_1 \right\} = \arg\min_\beta \sum_{j=1}^p \left\{ \frac{1}{2}(\beta_j - u_j)^2 + \lambda|\beta_j| \right\}.$$

よって各成分は

$$\hat{\beta}_j = \mathcal{S}_\lambda(u_j), \quad \mathcal{S}_\lambda(t) := \mathrm{sign}(t)\,(|t| - \lambda)_+$$

（ソフト閾値化）となる。

9.3　Oracle 不等式（基本不等式 → 円錐条件 →RE 条件）

ここが理論の核である。以下、$\Delta = \hat{\beta} - \beta^\star$、また集合の制限を Δ_S, Δ_{S^c} と書く。

補題 9.4 (基本不等式)**.** 任意の β に対し、特に $\beta = \beta^\star$ を代入して

$$\frac{1}{2n}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \le \frac{1}{2n}\|y - X\beta^\star\|_2^2 + \lambda\|\beta^\star\|_1$$

が成り立つ。これを整理すると

$$\frac{1}{2n}\|X\Delta\|_2^2 \le \frac{1}{n}\varepsilon^\top X\Delta + \lambda(\|\beta^\star\|_1 - \|\beta^\star + \Delta\|_1).$$

（ここで $y = X\beta^\star + \varepsilon$ を用いた。）

証明. $\hat{\beta}$ は最小解なので $f(\hat{\beta}) \le f(\beta^\star)$。そこから $y = X\beta^\star + \varepsilon$ を代入し、二乗ノルムの差を展開すればよい。

補題 9.5 (円錐条件：ℓ_1 の分解可能性)**.** もし

$$\lambda \geq 2 \left\| \frac{1}{n} X^\top \varepsilon \right\|_\infty$$

ならば、Δ は

$$\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$$

を満たす（cone condition）。

証明. 基本不等式において内積項を H ö lder で抑える：

$$\frac{1}{n}\varepsilon^\top X\Delta = \left\langle \frac{1}{n}X^\top\varepsilon,\ \Delta \right\rangle \leq \left\| \frac{1}{n}X^\top\varepsilon \right\|_\infty \|\Delta\|_1 \leq \frac{\lambda}{2}\|\Delta\|_1.$$

次に ℓ_1 ノルムの差を支持集合で分解する。$S = \mathrm{supp}(\beta^\star)$ なので $(\beta^\star)_{S^c} = 0$。三角不等式より

$$\|\beta^\star\|_1 - \|\beta^\star + \Delta\|_1 = \|\beta^\star_S\|_1 - \|\beta^\star_S + \Delta_S\|_1 - \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1 - \|\Delta_{S^c}\|_1.$$

これらを基本不等式に代入して

$$\frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\lambda}{2}\|\Delta\|_1 + \lambda(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1).$$

右辺を Δ_S, Δ_{S^c} に分けると

$$\frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\lambda}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) + \lambda\|\Delta_S\|_1 - \lambda\|\Delta_{S^c}\|_1 = \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1.$$

左辺は非負なので

$$0 \leq \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1 \quad \Rightarrow \quad \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1.$$

定義 9.6 (Restricted Eigenvalue: RE 定数)**.** 設計行列 X について、与えられた S と定数 $c > 0$ に対し

$$\kappa(S, c) := \inf_{\Delta \neq 0:\ \|\Delta_{S^c}\|_1 \leq c\|\Delta_S\|_1} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_S\|_2}$$

を RE 定数と呼ぶ（$\kappa(S,c) > 0$ が欲しい）。

直感的には「ℓ_1 円錐上で X がつぶれない」ことを意味し、$p > n$ でも疎な方向に限れば同定できる、という条件である。

定理 9.7 (Oracle 不等式：予測誤差と ℓ_1 誤差)**.** $\lambda \geq 2\|(1/n)X^\top\varepsilon\|_\infty$ を仮定し、さらに $\kappa = \kappa(S, 3) > 0$ とする。すると Lasso 誤差 $\Delta = \hat\beta - \beta^\star$ は

$$\frac{1}{n}\|X\Delta\|_2^2 \leq \frac{9\lambda^2 s}{\kappa^2}, \qquad \|\Delta\|_1 \leq \frac{12\lambda s}{\kappa^2}$$

を満たす。

証明. 補題 9.2 の基本不等式から出発する。先ほどと同様に内積項を抑え、さらに ℓ_1 差を分解すると

$$\frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\lambda}{2}\|\Delta\|_1 + \lambda(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1) = \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1 \leq \frac{3\lambda}{2}\|\Delta_S\|_1.$$

よって

$$\frac{1}{n}\|X\Delta\|_2^2 \leq 3\lambda\|\Delta_S\|_1. \tag{9.1}$$

一方、補題 9.3 より Δ は円錐 $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$ に属するので、RE 定義から

$$\|X\Delta\|_2 \geq \kappa\sqrt{n}\|\Delta_S\|_2. \tag{9.2}$$

さらに $\|\Delta_S\|_1 \le \sqrt{s}\|\Delta_S\|_2$ より、(9.2) は

$$\|\Delta_S\|_1 \le \sqrt{s}\|\Delta_S\|_2 \le \frac{\sqrt{s}}{\kappa\sqrt{n}}\|X\Delta\|_2. \tag{9.3}$$

(9.1) と (9.3) を組み合わせる：

$$\frac{1}{n}\|X\Delta\|_2^2 \le 3\lambda \cdot \frac{\sqrt{s}}{\kappa\sqrt{n}}\|X\Delta\|_2.$$

両辺が非負なので $\|X\Delta\|_2$ を約して

$$\frac{1}{\sqrt{n}}\|X\Delta\|_2 \le \frac{3\lambda\sqrt{s}}{\kappa} \quad \Rightarrow \quad \frac{1}{n}\|X\Delta\|_2^2 \le \frac{9\lambda^2 s}{\kappa^2}.$$

次に ℓ_1 誤差は円錐条件より $\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \le 4\|\Delta_S\|_1$。よって (9.3) と上の予測誤差境界から

$$\|\Delta\|_1 \le 4\|\Delta_S\|_1 \le \frac{4\sqrt{s}}{\kappa\sqrt{n}}\|X\Delta\|_2 \le \frac{4\sqrt{s}}{\kappa} \cdot \frac{3\lambda\sqrt{s}}{\kappa} = \frac{12\lambda s}{\kappa^2}.$$

Restricted Eigenvalue (RE) 条件の幾何学的理解

RE 条件は「スパース方向で設計行列 X が縮退しない」ことを保証する。

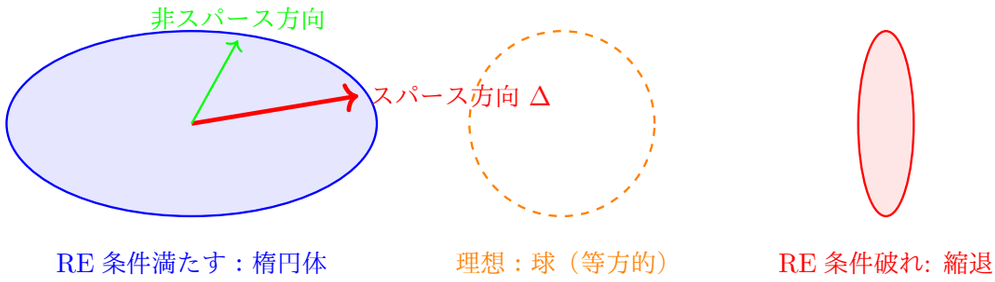


図 9.2　RE 条件の幾何学：スパース方向で X が縮退しないことを保証。X^TX/n の固有値がスパース部分空間で下から抑えられる。

RE 条件の直感:

- **スパース方向の安定性**: $\|X\Delta\|_2 \ge \kappa\sqrt{n}\|\Delta_S\|_2$ は「スパース成分が小さくても予測誤差は大きい」ことを意味し、Lasso が真の支持集合を回復できる根拠。
- **Matrix Bernstein との接続**: 8 章の Matrix Bernstein で $\|X^TX/n - I\|_{\mathrm{op}} \le \epsilon$ が高確率で成立 $\Rightarrow$ RE 条件が満たされる！
- **サンプル数条件**: $n \gg s\log p$ で RE 条件が高確率で成立（ランダム設計）。

Irrepresentable Condition：変数選択の一致性

Lasso が真の支持集合を正確に選択する（sign consistency）ためには、より強い条件が必要。

定義 9.8 (Irrepresentable Condition)**.** 設計行列 X の Gram 行列を $\Sigma = X^TX/n$ とし、$S = \mathrm{supp}(\beta^\star)$ とする。Irrepresentable 条件とは

$$\|\Sigma_{S^c,S}\Sigma_{S,S}^{-1}\mathrm{sign}(\beta_S^\star)\|_\infty < 1$$

が成立することをいう。

Irrepresentable の直感:

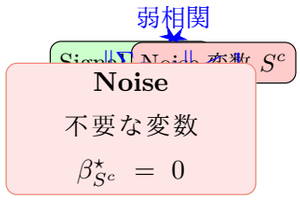


図 9.3 Irrepresentable Condition：Signal 変数と Noise 変数の相関が弱いことを要求。これにより Lasso が真の支持集合を正確に選択できる。

- **変数間の独立性**: 真に重要な変数と不要な変数が強く相関していると、Lasso は誤った変数を選んでしまう。
- **RE vs Irrepresentable**: RE 条件は予測精度、Irrepresentable 条件は変数選択の正確性を保証。
- **実務的課題**: 高相関変数が多い場合、Irrepresentable 条件は破れやすい ⇒ Elastic Net などの改良手法が必要。

Oracle 性能達成への道筋

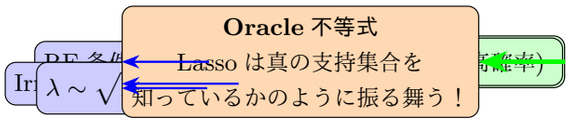


図 9.4 Lasso Oracle 性能への道筋：RE 条件で予測精度、Irrepresentable 条件で変数選択、適切な λ 選択で Oracle 性能を達成。

9.4 スパース推定手法の比較：Lasso, SCAD, MCP, Elastic Net

Lasso は強力だが、バイアスや変数選択の限界がある。様々な改良手法が提案されている。

表 9.1 スパース推定手法の比較

手法	ペナルティ	バイアス	変数選択	計算
Lasso	$\lambda\|\beta\|$	大	可	凸（高速）
SCAD	非凸	小	可	非凸（遅）
MCP	非凸	小	可	非凸（遅）
Adaptive Lasso	$\lambda_j\|\beta_j\|$	中	可	凸
Elastic Net	$\lambda_1\|\beta\| + \lambda_2\beta^2$	中	可（群）	凸
Group Lasso	$\lambda\|\beta_g\|_2$	中	群選択	凸

実務的選択ガイド:

- **高速・大規模データ**: Lasso（凸最適化、並列化容易）
- **高相関変数**: Elastic Net（相関変数を群として選択）
- **理論的厳密性**: SCAD/MCP（Oracle 性能、unbiased selection）
- **群構造**: Group Lasso（遺伝子パスウェイ、カテゴリ変数）
- **適応的重み**: Adaptive Lasso（二段階推定、Oracle 性能）

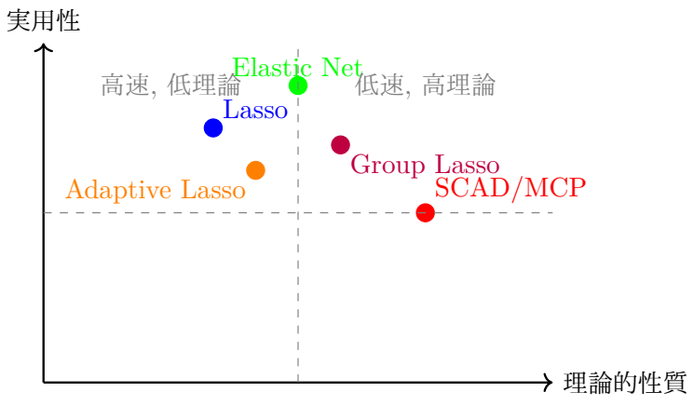


図 9.5　スパース推定手法の「理論 vs 実用」トレードオフ：Lasso は計算効率が高いが理論的にはバイアスあり。SCAD/MCP は理論的に優れるが計算コスト高。Elastic Net はバランス型。

注意 9.9 (λ の選び方)**.** 典型的に ε_i が sub-Gaussian、かつ列正規化 $\|X_{\cdot j}\|_2^2/n = 1$ を仮定すると、

$$\left\| \frac{1}{n} X^\top \varepsilon \right\|_\infty \lesssim \sigma \sqrt{\frac{\log p}{n}}$$

が高確率で成り立つので、$\lambda \asymp \sigma\sqrt{(\log p)/n}$ が自然となる。これを定理 9.4 に代入すると

$$\frac{1}{n}\|X(\hat{\beta} - \beta^\star)\|_2^2 \lesssim \frac{\sigma^2 s \log p}{n}$$

という、現代統計で最も基本的な高次元レートが得られる。

9.5　変数選択一致性と演習

定義 9.10 (符号一致・支持回復)**.**

$$\operatorname{sign}(\hat{\beta}) = \operatorname{sign}(\beta^\star)$$

が高確率で成立することを符号一致（support recovery）という。これは予測誤差の小ささより強い性質であり、設計行列に追加条件（例：irrepresentable condition）を要することが多い。

実社会イメージ（「当てる」 vs 「選ぶ」） 予測（$\|X(\hat{\beta} - \beta^\star)\|$）が良いことと、真に効いている特徴量を正しく同定することは別問題である。材料探索や医療のように「解釈」が重要な場合は支持回復条件が本質になる一方、推薦・予測では Oracle 不等式のような予測誤差評価が中心になる。

演習 9.1.

1. **基本不等式の練習**: 補題 9.2 の展開を、$\|a-b\|^2 = \|a\|^2 + \|b\|^2 - 2a^\top b$ から丁寧に導け。
2. **直交設計**: 例 9.1 の導出を、KKT 条件から直接示せ。
3. **円錐条件の一般化**: 補題 9.3 の定数「3」が、$\lambda \geq (1+\eta)\|(1/n)X^\top\varepsilon\|_\infty$ のときどう変わるか計算せよ。
4. **RE 条件の意味**: $\kappa(S, 3) = 0$ となる具体例（X の列に完全な線形従属がある状況）を作れ。
5. **レート**: $\lambda \asymp \sigma\sqrt{(\log p)/n}$ を仮定したとき、定理 9.4 の境界から ℓ_2 誤差 $\|\Delta\|_2$ の上界を導け（ヒント：$\|\Delta_S\|_2$ と $\|X\Delta\|_2$ を結び、$\|\Delta_{S^c}\|_2 \leq \|\Delta_{S^c}\|_1/\sqrt{s}$ を使う）。

第 10 章

一般化線形モデルと ℓ_1 正則化（Logistic Lasso）

2 値データ $Y \in \{0,1\}$ に対する代表的モデルがロジスティック回帰であり、高次元 $p \gg n$ では ℓ_1 正則化によりスパース性を導入する。

10.1　ロジスティック回帰：モデルと尤度

定義 10.1 (ロジスティックモデル)**.** 観測 $(x_i, y_i) \in \mathbb{R}^p \times \{0,1\}$（$i = 1, \dots, n$）に対し、条件付き分布を

$$P(Y_i = 1 \mid x_i) = \sigma(x_i^\top \beta), \qquad \sigma(t) := \frac{1}{1+e^{-t}}$$

とする。ここで $\beta \in \mathbb{R}^p$ は未知パラメータである。

命題 10.2 (対数尤度と負の対数尤度)**.** ロジスティック回帰の（平均）対数尤度は

$$\ell_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Big(y_i x_i^\top \beta - \log(1 + e^{x_i^\top \beta}) \Big),$$

従って負の対数尤度（損失）は

$$L_n(\beta) := -\ell_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \Big(\log(1 + e^{x_i^\top \beta}) - y_i x_i^\top \beta \Big).$$

定理 10.3 (勾配・ヘッセ行列、凸性)**.** $p_i(\beta) := \sigma(x_i^\top \beta)$ とおくと

$$\nabla L_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} (p_i(\beta) - y_i) x_i = \frac{1}{n} X^\top (p(\beta) - y),$$

$$\nabla^2 L_n(\beta) = \frac{1}{n} X^\top W(\beta) X, \qquad W(\beta) = \operatorname{diag}\big(p_i(\beta)(1 - p_i(\beta))\big).$$

特に $\nabla^2 L_n(\beta) \succeq 0$ より L_n は凸である。

証明. 各 i について $\frac{d}{dt} \log(1 + e^t) = \sigma(t)$ と $\sigma'(t) = \sigma(t)(1 - \sigma(t))$ を用い、合成関数の微分（チェインルール）で成分ごとに計算する。得られた $\nabla^2 L_n(\beta)$ は $W(\beta) \succeq 0$ なので半正定値。

10.2　Logistic Lasso の定義

定義 10.4 (Logistic Lasso)**.** $\lambda > 0$ に対し、Logistic Lasso 推定量 $\hat{\beta}$ を

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \Big\{ L_n(\beta) + \lambda \|\beta\|_1 \Big\}$$

で定める。

注意 10.5 (線形回帰 Lasso との違い)**.** 線形回帰では二乗誤差が「強凸」になりやすいが、ロジスティック損失はデータ配置により曲率が弱くなり得る（分離可能性・確率 $p_i(\beta)$ の飽和など）。そのため誤差解析では「制限付き強凸性（RSC）」のような条件が要になる。

10.3 KKT 条件と基本不等式（円錐条件まで）

以後、真のパラメータを $\beta^\star$、誤差を $\Delta := \hat\beta - \beta^\star$、支持集合を $S = \mathrm{supp}(\beta^\star)$、$s = |S|$ とする。

定理 10.6 (KKT 条件)**.** ある $\hat z \in \partial\|\hat\beta\|_1$ が存在して

$$\nabla L_n(\hat\beta) + \lambda \hat z = 0, \qquad \hat z_j = \begin{cases} \mathrm{sign}(\hat\beta_j) & (\hat\beta_j \neq 0), \\ u,\ |u| \le 1 & (\hat\beta_j = 0) \end{cases}$$

が成り立つ。

証明. L_n は凸かつ微分可能、$\|\cdot\|_1$ は凸。凸最適化の一階最適性条件 $0 \in \nabla L_n(\hat\beta) + \lambda \partial\|\hat\beta\|_1$ より従う。

補題 10.7 (基本不等式)**.** $\hat\beta$ の最適性より

$$L_n(\beta^\star + \Delta) - L_n(\beta^\star) + \lambda(\|\beta^\star + \Delta\|_1 - \|\beta^\star\|_1) \le 0. \tag{10.1}$$

証明. 目的関数 $L_n(\beta) + \lambda\|\beta\|_1$ で $\hat\beta$ が最小なので、比較点 $\beta^\star$ を代入するだけ。

補題 10.8 (スコアの ℓ_∞ 制御 $\to$ 円錐条件)**.** もし

$$\lambda \ge 2\|\nabla L_n(\beta^\star)\|_\infty \tag{10.2}$$

ならば

$$\|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1 \tag{10.3}$$

が成り立つ。

証明. 凸性より

$$L_n(\beta^\star + \Delta) - L_n(\beta^\star) \ge \langle \nabla L_n(\beta^\star), \Delta \rangle.$$

これを (10.1) に代入して

$$\langle \nabla L_n(\beta^\star), \Delta \rangle + \lambda(\|\beta^\star + \Delta\|_1 - \|\beta^\star\|_1) \le 0.$$

H ö lder より $\langle \nabla L_n(\beta^\star), \Delta\rangle \ge -\|\nabla L_n(\beta^\star)\|_\infty \|\Delta\|_1 \ge -(\lambda/2)\|\Delta\|_1$。また $(\beta^\star)_{S^c} = 0$ を用い、三角不等式で

$$\|\beta^\star + \Delta\|_1 - \|\beta^\star\|_1 = \|\beta^\star_S + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 - \|\beta^\star_S\|_1 \ge -\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1.$$

ゆえに

$$-\frac{\lambda}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) + \lambda(-\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) \le 0,$$

すなわち $(\lambda/2)\|\Delta_{S^c}\|_1 - (3\lambda/2)\|\Delta_S\|_1 \le 0$。整理して (10.3)。

10.4 RSC（制限付き強凸性）と Oracle 不等式

定義 10.9 (RSC：ここでは簡約形)**.** ある $\kappa > 0$ が存在して、円錐

$$\mathcal{C} := \{\Delta \neq 0 :\ \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1\}$$

上で

$$\Delta^\top \nabla^2 L_n(\tilde\beta)\Delta = \frac{1}{n}\|W(\tilde\beta)^{1/2}X\Delta\|_2^2 \ge \kappa\|\Delta\|_2^2 \tag{10.4}$$

が成り立つと仮定する（$\tilde\beta$ は $\beta^\star$ と $\hat\beta$ の間の点）。

直感的には「推定に関係する疎な方向では、ロジスティック損失が十分に曲がっている」ことを意味する。

定理 10.10 (Logistic Lasso の基本 Oracle 型上界)**.** (10.2) と RSC (10.4) を仮定する。このとき

$$\|\Delta\|_2 \le \frac{C_2\,\lambda\sqrt{s}}{\kappa}, \qquad \|\Delta\|_1 \le \frac{C_1\,\lambda s}{\kappa} \tag{10.5}$$

となる（普遍定数 $C_1, C_2 > 0$）。

証明. (1) (10.1) に対し、$\beta^\star$ 周りの 2 次展開（平均値の定理）で

$$L_n(\beta^\star+\Delta) - L_n(\beta^\star) = \langle \nabla L_n(\beta^\star), \Delta\rangle + \frac{1}{2}\Delta^\top\nabla^2 L_n(\tilde\beta)\Delta$$

となる $\tilde\beta$ が存在する。(2) (10.2) で $\langle \nabla L_n(\beta^\star),\Delta\rangle \ge -(\lambda/2)\|\Delta\|_1$。(3) ℓ_1 差は前節と同様に

$$\|\beta^\star+\Delta\|_1 - \|\beta^\star\|_1 \ge -\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1.$$

これらを (10.1) に代入し、$\|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1$ を使うと

$$\frac{1}{2}\Delta^\top\nabla^2 L_n(\tilde\beta)\Delta \le \frac{3\lambda}{2}\|\Delta_S\|_1. \tag{10.6}$$

(4) RSC より左辺 $\ge (\kappa/2)\|\Delta\|_2^2$。(5) $\|\Delta_S\|_1 \le \sqrt{s}\|\Delta_S\|_2 \le \sqrt{s}\|\Delta\|_2$ を (10.6) に適用して

$$\frac{\kappa}{2}\|\Delta\|_2^2 \le \frac{3\lambda}{2}\sqrt{s}\|\Delta\|_2.$$

$\|\Delta\|_2 \ge 0$ なので約して $\|\Delta\|_2 \le (3\lambda\sqrt{s})/\kappa$。よって $C_2 = 3$ が取れる。(6) $\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \le 4\|\Delta_S\|_1 \le 4\sqrt{s}\|\Delta\|_2$ より $\|\Delta\|_1 \le 12(\lambda s)/\kappa$。よって $C_1 = 12$。

注意 10.11 (λ のスケール)**.** $\nabla L_n(\beta^\star) = (1/n)X^\top(p(\beta^\star)-y)$ は「ノイズのような和」になっているので、sub-Gaussian 条件と列正規化の下で $\|\nabla L_n(\beta^\star)\|_\infty$ は概ね $\sqrt{(\log p)/n}$ スケールで抑えられ、$\lambda \asymp \sqrt{(\log p)/n}$ が自然になる。

10.5 例題・演習

例題 10.12 (勾配・ヘッセ行列の手計算)**.**

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^n \Big(\log(1+e^{x_i^\top\beta}) - y_i x_i^\top\beta\Big)$$

から $\nabla L_n(\beta)$, $\nabla^2 L_n(\beta)$ を導け。

解答. 定理 10.1 の通り。要点は $\frac{d}{dt}\log(1+e^t) = \sigma(t)$ と $\sigma'(t) = \sigma(t)(1-\sigma(t))$。

例題 10.13 (分離可能性と MLE の発散)**.** もしデータが完全分離（ある β で $x_i^\top\beta > 0$ が $y_i = 1$ の全点、$x_i^\top\beta < 0$ が $y_i = 0$ の全点で成立）なら、（ℓ_1 なしの）MLE が有限に存在しないことがある。一方、ℓ_1 正則化は $\|\beta\|_1$ を罰するため、解の存在（少なくとも目的関数の下界）を与える方向に働く。

演習 10.1. 1. **凸性**: 任意の $a \in \mathbb{R}$ で $\log(1+e^a)$ が凸であることを、2 階微分 $\sigma(a)(1-\sigma(a)) \ge 0$ から示せ。

2. **KKT**: $\hat{\beta}$ が解であることと、定理 10.2 の KKT 条件が同値であることを（凸解析の一般論を使わず）示せ。
3. **円錐条件の一般化**: 補題 10.4 を、$\lambda \geq (1+\eta)\|\nabla L_n(\beta^\star)\|_\infty$ の形に一般化し、係数「3」がどう変わるか計算せよ。
4. **RSC**: $\nabla^2 L_n(\beta) = (1/n) X^\top W(\beta) X$ で $0 < w_i(\beta) \leq 1/4$ を用い、$W(\beta) \succeq w_{\min} I$ が成り立つ領域では「加重設計行列」の RE 条件に帰着することを示せ。
5. **推定したい量の変換**: $\theta = g(\beta)$（例：ある座標の odds ratio）を推定したいとき、$\hat{\theta} = g(\hat{\beta})$ の誤差評価がデルタ法（第 7 章）とどう接続するか説明せよ。

付録 A

集中不等式の道具箱

（Orlicz ノルム・同値な定義・matrix Bernstein）

A.1　Orlicz ノルムと確率変数クラス

定義 A.1 (ψ_2 Orlicz ノルム：sub-Gaussian)**.** 確率変数 X に対し

$$\|X\|_{\psi_2} := \inf\left\{K>0:\ \mathbb{E}\exp\!\left(X^2/K^2\right)\le 2\right\}$$

と定める。$\|X\|_{\psi_2}<\infty$ のとき X を sub-Gaussian と呼ぶ。

定義 A.2 (ψ_1 Orlicz ノルム：sub-exponential)**.**

$$\|X\|_{\psi_1} := \inf\left\{K>0:\ \mathbb{E}\exp\!\left(|X|/K\right)\le 2\right\}$$

と定める。$\|X\|_{\psi_1}<\infty$ のとき X を sub-exponential と呼ぶ。

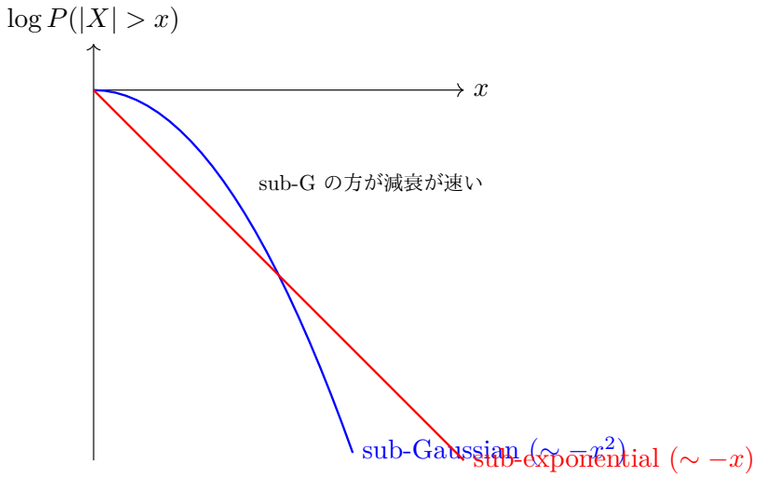


図 A.1　Orlicz ノルムと尾確率の減衰：ψ_2 (sub-Gaussian) は正規分布のように二乗で減衰し、ψ_1 (sub-exponential) は指数分布のように線形で減衰する。Matrix Bernstein ではこの二つの混合形が現れる。

補題 A.3 (基本性質：スケーリング・三角不等式型)**.** 任意の $a\in\mathbb{R}$ について $\|aX\|_{\psi_\alpha}=|a|\|X\|_{\psi_\alpha}$ $(\alpha=1,2)$。またある普遍定数 $C>0$ が存在して

$$\|X+Y\|_{\psi_2}\le C(\|X\|_{\psi_2}+\|Y\|_{\psi_2}),\qquad \|X+Y\|_{\psi_1}\le C(\|X\|_{\psi_1}+\|Y\|_{\psi_1})$$

が成り立つ（厳密な「ノルム」ではなく準ノルムであることに注意）。

証明. スケーリングは定義に $X \mapsto aX$ を代入して直ちに従う。加法については $\exp((x+y)^2) \le \exp(2x^2)\exp(2y^2)$ と Hölder を組み合わせ、定義の「≤ 2」を満たす係数を調整すればよい（ψ_1 も同様に $\exp(|x+y|) \le \exp(|x|)\exp(|y|)$ を使う）。

A.2 sub-Gaussian の同値性（mgf・尾・モーメント）

定理 A.4 (sub-Gaussian の同値な特徴付け)**.** 平均 0 の確率変数 X について、以下は互いに同値であり、定数は普遍定数倍で行き来できる。

1. （Orlicz）$\|X\|_{\psi_2} \le K$.
2. （尾確率）ある $c>0$ が存在して任意の $t \ge 0$ で
$$\mathbb{P}(|X| \ge t) \le 2\exp\!\left(-c\,\frac{t^2}{K^2}\right).$$
3. （モーメント増大）ある $C>0$ が存在して任意の $q \ge 1$ で
$$(\mathbb{E}|X|^q)^{1/q} \le CK\sqrt{q}.$$
4. （mgf）ある $C>0$ が存在して任意の $s \in \mathbb{R}$ で
$$\mathbb{E}\exp(sX) \le \exp(CK^2 s^2).$$

証明. (1)⇒(2)：任意の $t \ge 0$ に対し Markov より
$$\mathbb{P}(|X| \ge t) = \mathbb{P}\left(\exp(X^2/K^2) \ge e^{t^2/K^2}\right) \le e^{-t^2/K^2}\mathbb{E}e^{X^2/K^2} \le 2e^{-t^2/K^2}.$$
よって (2) が $c=1$ で成立。

(2)⇒(3)：分布関数の積分表示（tail integration）を使う：
$$\mathbb{E}|X|^q = \int_0^\infty q t^{q-1}\mathbb{P}(|X| \ge t)\,dt \le 2q\int_0^\infty t^{q-1}\exp\!\left(-c\frac{t^2}{K^2}\right)dt.$$
変数変換 $u = ct^2/K^2$ により右辺は $\lesssim K^q q(q/2-1)!$ 型になり、Stirling を使って $(\mathbb{E}|X|^q)^{1/q} \lesssim K\sqrt{q}$ を得る。

(3)⇒(4)：級数で mgf を評価する：
$$\mathbb{E}e^{sX} = 1 + \sum_{m\ge 2}\frac{s^m \mathbb{E}X^m}{m!} \le 1 + \sum_{m\ge 2}\frac{|s|^m\mathbb{E}|X|^m}{m!}.$$
(3) より $\mathbb{E}|X|^m \le (CK\sqrt{m})^m$。これを代入すると
$$\frac{|s|^m\mathbb{E}|X|^m}{m!} \le \frac{(|s|CK)^m m^{m/2}}{m!}$$
であり、Stirling $m! \asymp (m/e)^m\sqrt{m}$ から級数は $\exp(C'K^2s^2)$ で抑えられる。

(4)⇒(1)：(4) から X のガウス型尾確率（Chernoff）を得て (2) を経由してもよいし、直接 $\mathbb{E}e^{X^2/K^2}$ をモーメントで展開して収束半径を示してもよい。いずれも定数調整で $\|X\|_{\psi_2} \lesssim K$ が従う。

A.3 sub-exponential と Bernstein 型評価

定理 A.5 (sub-exponential の同値性：ψ_1)**.** 確率変数 X について $\|X\|_{\psi_1} \le K$ であることは、ある $c>0$ が存在して
$$\mathbb{P}(|X| \ge t) \le 2\exp\!\left(-c\,\frac{t}{K}\right), \qquad t \ge 0$$

が成り立つことと同値である（定数は普遍定数倍で行き来）。

証明. ψ_2 と同様に Markov で (Orlicz)⇒(尾) を示し、逆は tail integration により $\mathbb{E}\exp(|X|/CK) \le 2$ を構成すればよい。

定理 A.6 (独立 sub-exponential 和の Bernstein)**.** 独立で平均 0 の $X_1, \dots, X_n$ が $\|X_i\|_{\psi_1} \le K$ を満たすとき、ある普遍定数 $c, C > 0$ が存在して任意の $t \ge 0$ で

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i \right| \ge t \right) \le 2 \exp\left[-cn \min\left(\frac{t^2}{K^2}, \frac{t}{K} \right) \right]$$

が成り立つ。

証明. mgf の局所上界（$|s| \le c/K$ で $\mathbb{E}e^{sX_i} \le \exp(Cs^2K^2)$）を sub-exponential の同値性から導く。独立性で積に分解し、Chernoff により $\exp(-st + Cs^2K^2)$ の最小化を「二次領域（t 小）/一次領域（t 大）」に分けて行うと結論が出る。

A.4　最大値評価と union bound の定石

補題 A.7 (sub-Gaussian 最大値の上界)**.** $X_1, \dots, X_p$ が（独立でなくてもよいが）同一の尾上界 $\mathbb{P}(|X_j| \ge t) \le 2\exp(-t^2/(2K^2))$ を満たすとき

$$\mathbb{P}\left(\max_{1 \le j \le p} |X_j| \ge K\sqrt{2\log(2p/\delta)} \right) \le \delta.$$

証明. union bound より

$$\mathbb{P}\left(\max_{j} |X_j| \ge t \right) \le \sum_{j=1}^{p} \mathbb{P}(|X_j| \ge t) \le 2p \exp(-t^2/(2K^2)).$$

右辺を $\le \delta$ とおくように t を解けばよい。

（Lasso の $\|(1/n)X^\top \varepsilon\|_\infty$ 抑え込みが、まさにこの型で出てきます。）

A.5　Matrix Bernstein（自己共役行列の集中）

ここからは「非可換性（行列の積が可換でない）」が障害になるが、行列版のラプラス変換法で同型の結果が得られる。

補題 A.8 (行列版 Markov：ラプラス変換法)**.** 自己共役行列 Y と $t > 0$ に対し

$$\mathbb{P}\big(\lambda_{\max}(Y) \ge u\big) = \mathbb{P}\big(e^{t\lambda_{\max}(Y)} \ge e^{tu}\big) \le e^{-tu} \operatorname{tr} \mathbb{E}[e^{tY}]$$

が成り立つ。

証明. $\lambda_{\max}(Y) \ge u \Rightarrow \operatorname{tr}(e^{tY}) \ge e^{tu}$（最大固有値の寄与で trace が下から抑えられる）を用い、Markov を $\operatorname{tr}(e^{tY})$ に適用する。

定理 A.9 (Matrix Bernstein：簡潔版)**.** 独立な自己共役ランダム行列 $X_1, \dots, X_n \in \mathbb{R}^{d \times d}$ が

$$\mathbb{E}[X_k] = 0, \qquad \|X_k\|_{\mathrm{op}} \le R \text{ a.s.}$$

を満たすとする。分散パラメータを

$$\sigma^2 := \left\| \sum_{k=1}^{n} \mathbb{E}[X_k^2] \right\|_{\mathrm{op}}$$

とおくと、任意の $u \geq 0$ で

$$\mathbb{P}\left(\left\| \sum_{k=1}^{n} X_k \right\|_{\mathrm{op}} \geq u \right) \leq 2d \cdot \exp\left(-\frac{u^2}{2\sigma^2 + \frac{2Ru}{3}} \right). \tag{A.1}$$

証明 (主要ステップを明示)**.** (1) まず片側を抑える。$Y = \sum X_k$ とし、補題 A.6 より

$$\mathbb{P}(\lambda_{\max}(Y) \geq u) \leq e^{-tu} \operatorname{tr} \mathbb{E} e^{tY}. \tag{A.2}$$

(2) 次に mgf を逐次条件付き期待値で処理する。独立性により

$$\mathbb{E} e^{t \sum_{k=1}^{n} X_k} = \mathbb{E}\Big[\mathbb{E}\big[e^{t(\sum_{k=1}^{n-1} X_k)} e^{tX_n} \,\big|\, X_1, \dots, X_{n-1} \big] \Big] = \mathbb{E}\Big[e^{t(\sum_{k=1}^{n-1} X_k)} \, \mathbb{E} e^{tX_n} \Big].$$

非可換性のため単純に積の形にはならないが、「trace と指数関数」に関する標準不等式（Lieb 型の凸性）により、最終的に

$$\operatorname{tr} \mathbb{E} e^{t \sum_{k=1}^{n} X_k} \leq \operatorname{tr} \exp\left(\sum_{k=1}^{n} \log \mathbb{E} e^{tX_k} \right) \tag{A.3}$$

が得られる（ここが行列版の核心補題）。

(3) 各 X_k の mgf をスカラーの Bernstein と同じく二次で抑える。条件 $\|X_k\|_{\mathrm{op}} \leq R$ と $\mathbb{E} X_k = 0$ から、$|t| < 3/R$ の範囲で

$$\mathbb{E} e^{tX_k} \preceq \exp\left(\frac{t^2}{2(1 - \frac{Rt}{3})} \, \mathbb{E}[X_k^2] \right) \tag{A.4}$$

が示せる（テイラー展開と $X_k^m \preceq R^{m-2} X_k^2$ 型の順序評価で作る）。

(4) (A.3)(A.4) をまとめると

$$\operatorname{tr} \mathbb{E} e^{tY} \leq \operatorname{tr} \exp\left(\frac{t^2}{2(1 - \frac{Rt}{3})} \sum_{k=1}^{n} \mathbb{E}[X_k^2] \right) \leq d \cdot \exp\left(\frac{t^2}{2(1 - \frac{Rt}{3})} \sigma^2 \right). \tag{A.5}$$

最後の不等式は $\operatorname{tr}(e^A) \leq d\, e^{\lambda_{\max}(A)}$ を用いた。

(5) (A.2)(A.5) より

$$\mathbb{P}(\lambda_{\max}(Y) \geq u) \leq d \cdot \exp\left(-tu + \frac{t^2}{2(1 - \frac{Rt}{3})} \sigma^2 \right), \qquad 0 < t < 3/R.$$

右辺を t で最小化すると、(A.1) の片側版が出る。もう一方は $\lambda_{\min}(Y) \leq -u$ に同様に適用し、和をとって係数 $2d$ を得る。

A.6 例題

例題 A.10 (ガウスベクトルの線形汎関数)**.** $g \sim N(0, I_d)$, 固定 $u \in \mathbb{R}^d$ に対し $X = u^\top g$ は $N(0, \|u\|_2^2)$。したがって $\|X\|_{\psi_2} \asymp \|u\|_2$ を確認せよ（定理 A.2 のいずれかの特徴付けを使う）。

例題 A.11 (中心化 χ^2 は sub-exponential)**.** $Z \sim N(0,1)$ とし $X = Z^2 - 1$。mgf $\mathbb{E} e^{t(Z^2-1)}$ を $|t| < 1/2$ で評価し、$\|X\|_{\psi_1} < \infty$ を示せ。

A.7 演習問題

演習 A.1. 1. 定理 A.2 の (2)⇒(3) を、変数変換まで含めて定数を追って示せ。

2. sub-Gaussian の同値性から「X sub-Gaussian $\Rightarrow X^2 - \mathbb{E}X^2$ sub-exponential」を証明せよ。
3. 補題 A.5 を用いて、独立 sub-Gaussian $X_1, \dots, X_p$ に対し $\mathbb{E}\max_j |X_j| \lesssim K\sqrt{\log p}$ を示せ（積分表示 $\mathbb{E}U = \int_0^\infty \mathbb{P}(U \ge t)\,dt$ を使う）。
4. 定理 A.7 のステップ (3)（行列 mgf の二次上界）を、テイラー展開と $\|X_k\|_{\mathrm{op}} \le R$ から自力で埋めよ（ヒント：偶数次と奇数次を分け、作用素順序で抑える）。
5. **応用**: 第 8 章の標本共分散の評価を、(A.1) を使って「ε-net を使わない」方針で組み直せ（どこで d が現れるかを比較せよ）。