# 統計的機械学習

# Statistical Machine Learning

**理論基盤から深層学習・Transformer まで**

Yugo Nakayama

2026 年 2 月 9 日

# 目次

# 第 I 部

# 統計的機械学習 (Statistical Machine Learning)

# 第 1 章

# 教師あり学習の枠組み

## 1.1　統計的機械学習とは

本書で扱う「統計的機械学習」は、次の二つの立場を統合したものとみなせる。

- **統計学**：観測データから未知の構造を推定し、不確実性を評価する学問。
- **機械学習**：予測精度を最優先に、柔軟な関数クラスと最適化アルゴリズムを組み合わせる技術。

古典的な数理統計では、有限次元パラメータ $\theta \in \mathbb{R}^p$ を持つ確率モデル $p_\theta(y)$ を仮定し、最尤推定やベイズ推定により $\theta$ を推定する。これに対して現代の機械学習では、

- 入力 $X \in \mathcal{X} \subset \mathbb{R}^d$、
- 出力 $Y \in \mathcal{Y}$（連続値/離散ラベル）、
- それらの生成分布 $P_{XY}$

が存在すると仮定し、**入力から出力を予測する関数**

$$f : \mathcal{X} \to \mathcal{Y}$$

を直接的に学習する。$\theta$ の次元はもはや固定されておらず、ニューラルネットワークやカーネル法のように、**無限次元の関数空間上で最適化する**ことが標準となっている。

本章では、その最も基本的な設定である「教師あり学習」の枠組みを整理し、

- 回帰モデルと損失関数、
- 経験リスク最小化 (ERM) と汎化誤差、
- VC 次元や Rademacher 複雑度によるモデル複雑度の制御

を導入する。これらは、以降で扱う SVM, カーネル法, 深層学習の**共通の基盤**となる。

### 1.1.1　教師あり学習の基本シナリオ

機械学習の典型的なタスクは、以下のように記述できる。

- ある現象を表すペアデータ
$$\{(x_i, y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} P_{XY}$$
が観測される。ここで $x_i$ は説明変数（入力）、$y_i$ は目的変数（出力）である。
- 我々は $P_{XY}$ の詳細な形は知らないが、「似た入力には似た出力が現れる」という規則性が存在すると期待する。

- 目的は、新しい入力 $X_{\text{new}}$ に対して、対応する出力 $Y_{\text{new}}$ をできるだけ正確に予測する関数 $f$ を学習することである。

このとき、自然な指標として

- 実際の $Y$ と予測 $\hat{Y} = f(X)$ の乖離を測る**損失関数** $\ell(Y, \hat{Y})$,
- その期待値

$$R(f) = \mathbb{E}[\ell(Y, f(X))]$$

（**真のリスク**）

を考える。統計的機械学習の中心的な問題は、「有限個の標本から、この期待損失 $R(f)$ をできるだけ小さくするような関数 $f$ を選ぶにはどうすればよいか」である。

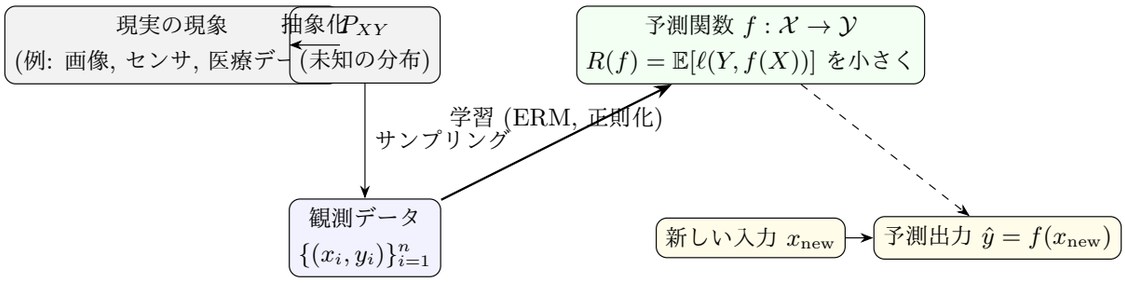



図 1.1　統計的機械学習の枠組み：現実の現象からデータを通じて予測関数を学習する

### 1.1.2　なぜ回帰から始めるか

本書では、教師あり学習の導入として**回帰問題**（連続値出力）から出発する。

- 二値分類や多クラス分類も、多くの場合「回帰＋しきい値」の形に書き直せる。
- 深層学習や Transformer でも、最終層は「実数ベクトルを出力して損失を計算する回帰問題」として実装されている。
- 回帰設定では、二乗損失によって**条件付き期待値**

$$f^\star(x) = \mathbb{E}[Y \mid X = x]$$

が自然に現れ、後の GLM・深層ネットワークの議論にスムーズにつながる。

したがって本章では、まず「回帰＋損失関数」という最も素朴な設定から始め、その中で

- ERM による推定、
- 真のリスク $R(f)$ の分解（近似誤差＋推定誤差）、
- モデル複雑度と汎化誤差のトレードオフ

を丁寧に理解する。分類や確率予測は、この枠組みの上で損失関数を変えるだけで扱えることを、第 2 章以降で確認する。

### 1.1.3　本章の流れ

1. **1.1 回帰モデルと損失関数**：教師あり学習の形式的定義を与え、二乗損失・MAE・Huber 損失などを比較する。

2. **1.2 経験リスク最小化と汎化誤差**：ERM のアイデアを述べ、真のリスクと経験リスクの差（汎化誤差）を Hoeffding 不等式と $\varepsilon$ -net 論理で評価する（定理 1.4）。
3. **1.3 VC 次元と Rademacher 複雑度**：仮説クラスの「容量」を測る概念を導入し、なぜ複雑すぎるモデルが過学習を起こすのかを理論的に説明する。

こうした基礎の上に、2 章以降の

- 線形分類器と最適化（パーセプトロン、ロジスティック回帰、SVM）、
- カーネル法と RKHS、
- 深層ニューラルネットワーク、
- 高次元スパース推定（Lasso, GLM Lasso）

を順次積み上げていく。

## 1.2　回帰モデルと損失関数

は、観測データ $\{(x_i, y_i)\}_{i=1}^n$ から未知の条件付き期待値関数 $\mathbb{E}[Y|X=x] = f^\star(x)$ を推定する問題である。数理統計のパラメータ推定とは異なり、「無限次元関数空間 $\mathcal{F}$ 上の最適化」が特徴である。

## 1.3　回帰モデルと損失関数

**定義 1.1** (教師あり学習)**.** 入力空間 $\mathcal{X} \subset \mathbb{R}^d$、出力 $\mathcal{Y} \subset \mathbb{R}$ に対し、標本 $\{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} P_{XY}$ が与えられたとき、関数クラス $\mathcal{F}$ から

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda R(f)$$

を経験リスク最小化 (ERM) と呼ぶ。ここで $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ は損失関数（例：二乗損失 $\ell(y, \hat{y}) = (y - \hat{y})^2$）、$R(f)$ は正則化項（例：$R(f) = \|f\|_{\mathcal{H}}^2$）。

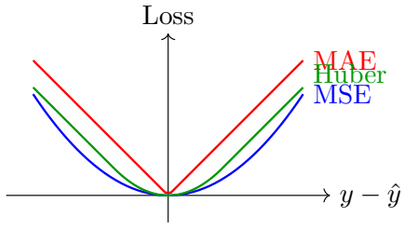



図 1.2　損失関数の比較：MSE（滑らか）、MAE（頑健）、Huber（折衷）

**定理 1.2** (汎化誤差の分解)**.** 任意の $f \in \mathcal{F}$ に対し、

$$R(f) := \mathbb{E}[\ell(f(X), Y)] = \underbrace{\mathbb{E}[\ell(f^\star(X), Y)]}_{\text{近似誤差}} + \underbrace{\mathbb{E}[(f(X) - f^\star(X))^2]}_{\text{推定誤差}}$$

が成立する（二乗損失の場合）。

**証明.** 条件付き期待値 $f^\star(x) = \mathbb{E}[Y|X=x]$ を代入：

$$\begin{aligned} \ell(f(X), Y) &= (f(X) - Y)^2 = (f(X) - f^\star(X) + (f^\star(X) - Y))^2 \\ &= (f(X) - f^\star(X))^2 + 2(f(X) - f^\star(X))(f^\star(X) - Y) + (f^\star(X) - Y)^2 \end{aligned}$$

2 番目の交叉項は $\mathbb{E}[(f^\star(X) - Y)|X] = 0$ より期待値 0。よって期待値をとると結論を得る。 □

**注意 1.3** (高校数学補完モジュール)**.** 推定誤差はさらに $\mathrm{Bias}[f(X)]^2 + \mathrm{Var}[f(X)]$ に分解可能（バイアス-バリアンス分解）。

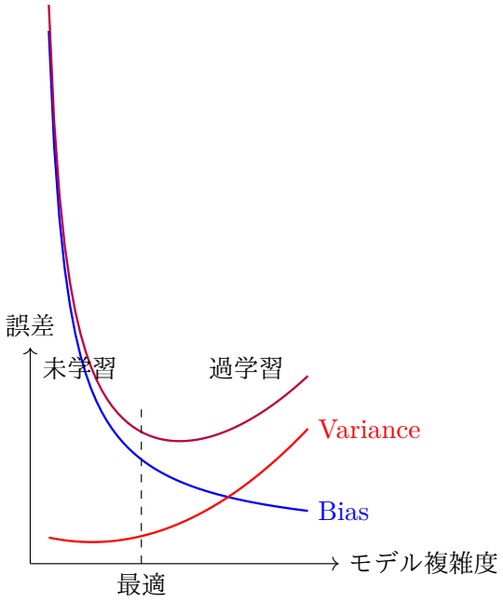



図 1.3　Bias-Variance トレードオフ：複雑度増加でバイアス減少・分散増加

## 1.4　経験リスク最小化 (ERM) と汎化誤差

**定理 1.4** (ERM の汎化保証)**.** $\mathcal{F}$ が VC 次元 $V < \infty$ の関数クラスとし、損失 $\ell$ が $L$-Lipschitz ならば、確率 $1-\delta$ で

$$R(\hat{f}_n) \le \inf_{f \in \mathcal{F}} R(f) + C\left(\sqrt{\frac{V \log n + \log(1/\delta)}{n}} + L\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

証明の骨子. Hoeffding 不等式と $\epsilon$-net 論理：経験リスク $\hat{R}_n(f) := \frac{1}{n}\sum \ell(f(x_i), y_i)$ と真リスク $R(f)$ の差を一様制御。$\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| = O_p(\sqrt{V/n})$ を示す。 □

**注意 1.5** (実社会イメージ)**.** 機械学習コンペ（Kaggle）では交差検証 (CV) で経験誤差を評価し、VC 理論でモデル複雑度を制御する。

## 1.5　VC 次元と Rademacher 複雑度

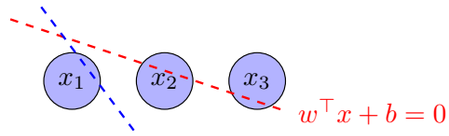
**定義 1.6** (VC 次元)**.** 仮説クラス $\mathcal{H}$ の VC 次元 $\mathrm{VC}(\mathcal{H})$ とは、最も多く $\mathcal{H}$ でシャタリング可能な点集合の最大サイズである。

**例題 1.7.** 半空間 $\mathcal{H} = \{x \mapsto \mathrm{sign}(w^\top x + b)\}$ の VC 次元は $d+1$ である。

**定理 1.8** (Rademacher 複雑度境界)**.** Rademacher 変数 $\epsilon_i = \pm 1$ に対し、

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i h(x_i)\right] \le C\sqrt{\frac{\mathrm{VC}(\mathcal{H}) \log n}{n}}$$

**証明.** Massart の有限類逼近定理と $\epsilon$-covering number の対数が VC 次元に比例することを用いる。 □




$2^3 = 8$ ラベル全て実現可能 $\Rightarrow$ VC$\geq 3$

図 1.4　VC 次元：3 点のシャタリング例（半空間）

**演習 1.1.**　1. 二乗損失の代替として Huber 損失 $\ell_\delta(t) = \frac{1}{2}t^2 \cdot \mathbf{1}_{|t|\leq\delta} + \delta(|t| - \frac{\delta}{2}) \cdot \mathbf{1}_{|t|>\delta}$ のロバスト性を証明せよ。

2. VC 次元 $V = 3$ のクラスで $n = 100$、$\delta = 0.05$ のときの汎化境界を数値計算せよ。
3. 過学習メカニズムを Bias-Variance 分解で説明し、早期停止の理論的正当性を Hoeffding 不等式で示せ。

## 第 1 章のまとめ

- 機械学習＝無限次元パラメータ空間上の ERM
- 汎化誤差＝近似誤差＋推定誤差
- VC 次元が学習可能性の鍵
- Bias-Variance トレードオフによるモデル選択

# 第 2 章

# 線形分類器と最適化

線形分類器は、教師あり学習の基盤であり、数理統計の最尤推定・凸最適化と機械学習最適化の架け橋となる。本章ではパーセプトロンからロジスティック回帰、勾配降下法までを数理的に整備する。

## 2.1　パーセプトロンと SVM の双対問題

**定義 2.1** (パーセプトロン)**.** データ $(x_i, y_i) \in \mathbb{R}^d \times \{\pm 1\}$ に対し、重み $w \in \mathbb{R}^d$、バイアス $b \in \mathbb{R}$ を

$$\hat{y}_i = \operatorname{sign}(w^\top x_i + b)$$

で予測。誤分類 $(y_i \hat{y}_i < 0)$ 時に更新：

$$w \leftarrow w + \eta y_i x_i, \quad b \leftarrow b + \eta y_i$$

**定理 2.2** (パーセプトロン収束定理)**.** データが線形分離可能 $\{y_i(w^\top x_i + b) > 0\}$ で、$\gamma = \min_i y_i(w^\top x_i + b)/\|w\| > 0$（マージン）ならば、更新回数 $k \le 1/\gamma^2$ で収束。

**証明.** 関数 $\Phi(w) = \min_i y_i(w^\top x_i + b)$ を考え、誤分類時 $\Phi(w_{k+1}) \ge \Phi(w_k) + \eta\gamma\|x_i\|$。$\Phi \ge 0$ で上限 $2R^2$（$R = \max \|x_i\|$）より有限ステップで収束。 □

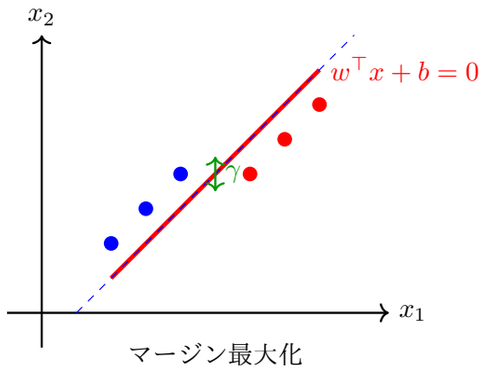



図 2.1　パーセプトロン：マージン $\gamma$ で収束保証

**注意 2.3** (高校数学補完モジュール)**.** SVM との関係は双対問題 $\max_\alpha \sum \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i^\top x_j$。カーネル化で非線形分離可能。

## 2.2 ロジスティック回帰：凸性と最尤推定

**定義 2.4** (ロジスティック回帰)**.** $P(y_i=1|x_i)=\sigma(w^\top x_i+b)$、$\sigma(t)=1/(1+e^{-t})$ とし、負対数尤度を最小化：

$$L(w,b)=-\frac{1}{n}\sum_{i=1}^{n}\Big[y_i(w^\top x_i+b)-\log(1+e^{w^\top x_i+b})\Big]$$

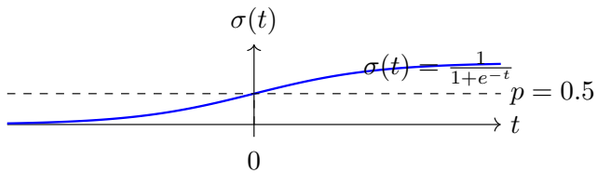



図 2.2 ロジスティックシグモイド関数：確率出力

**定理 2.5** (凸性と勾配)**.** $L$ は $w,b$ について凸関数。勾配：

$$\nabla_w L=\frac{1}{n}\sum_{i=1}^{n}(\sigma(w^\top x_i+b)-y_i)x_i,\quad \frac{\partial L}{\partial b}=\frac{1}{n}\sum_{i=1}^{n}(\sigma(w^\top x_i+b)-y_i)$$

**証明.** $\sigma''(t)=\sigma(t)(1-\sigma(t))>0$ より $\sigma$ は凸。合成関数 $\phi(w^\top x)=\log(1+e^{\phi})$ も凸（$e^t$ 凸より）。 □

**注意 2.6** (実社会イメージ)**.** 医療診断で「確率」を出力。ROC 曲線下面積 (AUC) が性能指標。

## 2.3 勾配降下法 (GD) の収束

**定義 2.7** (勾配降下法)**.** $\eta>0$ を学習率とし、

$$w_{k+1}=w_k-\eta\nabla L(w_k),\quad b_{k+1}=b_k-\eta\frac{\partial L}{\partial b}(b_k)$$

**定理 2.8** (GD 収束：$L$-滑らか凸関数)**.** $L$ が $L$-Lipschitz 滑らか $\|\nabla L(w)-\nabla L(w')\|\le L\|w-w'\|$、強凸 $\mu>0$ なら、

$$\|\nabla L(w_K)\|^2\le\frac{2L}{\eta K}\left(L(w_0)-L^\star\right),\quad \eta=\frac{2}{L+\mu}$$

**証明.** 滑らか性より $L(w_{k+1})\le L(w_k)+\nabla L(w_k)^\top(w_{k+1}-w_k)+\frac{L}{2}\|w_{k+1}-w_k\|^2$。$\eta=2/(L+\mu)$ で再帰的に減少を示す。 □

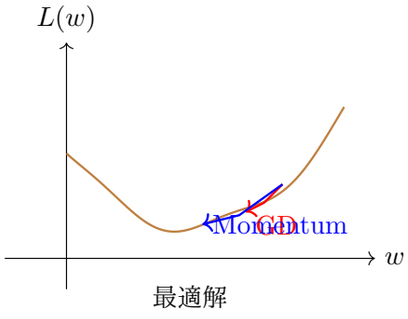



図 2.3 GD vs Momentum：谷底振動抑制

**注意 2.9** (高校数学補完モジュール)**.** Adam 等の適応的学習率は、過去勾配の 2 次モーメントで $\eta_t$ を動的調整：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla L_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2)\nabla L_t^2$$

$\hat{m}_t = m_t/(1-\beta_1^t)$、$\hat{v}_t = v_t/(1-\beta_2^t)$ で補正し $\eta_t/\sqrt{\hat{v}_t}$。

**演習 2.1.**

1. パーセプトロン更新が $w$ の凸結合 $\sum \lambda_i y_i x_i$（$\lambda_i \geq 0$、$\sum \lambda_i = 1$）を生成することを示せ。
2. ロジスティック損失 $\log(1+e^{-y(f(x))})$ が上界 $\max(0, -yf(x)) + \log 2$ を持つことを証明せよ。
3. $L$-滑らか関数で加速勾配法 (Nesterov) が GD より $O(1/\sqrt{K})$ 改善を示せ。
4. 二乗損失とクロスエントロピーの勾配が $(y-\hat{y})$ で一致することを確認せよ。

## 2.4 サポートベクターマシン (SVM)

線形分類器のうち、**マージン最大化**に基づく代表的手法がサポートベクターマシン (SVM) である。本節では、ハードマージン・ソフトマージン SVM の定式化、双対問題とカーネルトリック、パーセプトロンとの関係を整理する。

### 2.4.1 ハードマージン SVM：マージン最大化

まずデータが完全に線形分離可能な場合を考える。

**定義 2.10** (幾何マージン)**.** 超平面 $\{x : w^\top x + b = 0\}$ による分類器 $\hat{y} = \mathrm{sign}(w^\top x + b)$ を考える。各サンプル $(x_i, y_i)$ に対する「符号付き距離」は

$$\gamma_i = \frac{y_i(w^\top x_i + b)}{\|w\|_2}$$

この最小値

$$\gamma(w,b) = \min_{1 \leq i \leq n} \gamma_i$$

を、超平面 $(w,b)$ の**幾何マージン**と呼ぶ。

**定義 2.11** (ハードマージン SVM)**.** データ集合が線形分離可能、すなわちある $(w,b)$ が存在して $y_i(w^\top x_i + b) > 0$ を満たすとき、ハードマージン SVM は

$$\begin{aligned} &\max_{w,b} \ \gamma(w,b) \\ &\text{s.t. } y_i(w^\top x_i + b) \geq \gamma(w,b), \quad i = 1, \ldots, n \end{aligned}$$

によってマージンを最大化する分類器を求める。

実際の定式化では、スケール不定性（$(w,b)$ を同じ定数でスケーリングしても超平面は変わらない）を解消するため、$\gamma = 1/\|w\|$ と正規化して次の凸最適化問題に帰着させる。

**命題 2.12** (標準形)**.** 線形分離可能なとき、マージン最大化問題は

$$\begin{aligned} &\min_{w,b} \ \frac{1}{2}\|w\|_2^2 \\ &\text{s.t. } y_i(w^\top x_i + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

と同値である。

**証明.** 任意の分離超平面 $(\tilde{w}, \tilde{b})$ に対し、$\tilde{\gamma} = \min_i y_i(\tilde{w}^\top x_i + \tilde{b})/\|\tilde{w}\| > 0$ とおく。$(w,b) = (\tilde{w}/\tilde{\gamma}, \tilde{b}/\tilde{\gamma})$ とスケーリングすれば、すべての $i$ で

$$y_i(w^\top x_i + b) = \frac{1}{\tilde{\gamma}} y_i(\tilde{w}^\top x_i + \tilde{b}) \geq 1$$

このとき幾何マージンは

$$\gamma(w,b) = \min_i \frac{y_i(w^\top x_i + b)}{\|w\|} = \frac{1}{\|w\|} = \frac{\tilde{\gamma}}{\|\tilde{w}\|}$$

したがって、マージン最大化は $\|w\|^{-1}$ 最大化と同値であり、変数変換のもとで $\|w\|$ 最小化に帰着する。目的関数を $\frac{1}{2}\|w\|^2$ と書き換えた命題の問題は凸二次計画問題である。 □

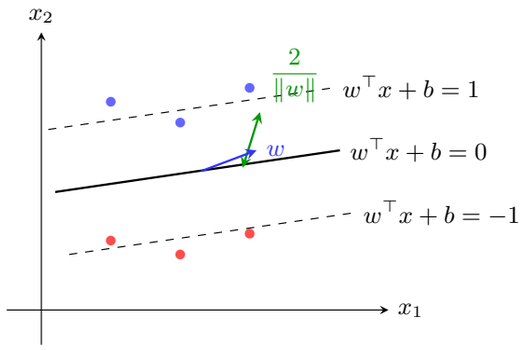



図 2.4　ハードマージン SVM：決定境界とマージン幅

### 2.4.2　双対問題とサポートベクトル

命題 2.12 の問題は小次元ではそのまま解けるが、高次元・カーネル化を考えると双対問題が有用になる。

**命題 2.13** (ハードマージン SVM の双対問題)**.** 命題 2.12 に対するラグランジュ関数

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^n \alpha_i\{y_i(w^\top x_i + b) - 1\}, \quad \alpha_i \geq 0$$

を考えると、原問題と等価な双対問題は

$$\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_i^\top x_j$$
$$\text{s.t.} \sum_{i=1}^n \alpha_i y_i = 0, \quad \alpha_i \geq 0$$

で与えられる。最適解 $\alpha^\star$ に対し、

$$w^\star = \sum_{i=1}^n \alpha_i^\star y_i x_i$$

が成り立つ。

**証明.** $\mathcal{L}$ を $w, b$ で最小化すると、

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_i \alpha_i y_i = 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

これを $\mathcal{L}$ に代入すると

$$\mathcal{L}(w(\alpha), b(\alpha), \alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j x_i^\top x_j$$

したがって、双対関数の最大化が命題の双対問題となる。凸性と Slater 条件より強双対性が成り立ち、原問題と双対問題の最適値は一致する。 □

**命題 2.14** (サポートベクトルの性質)**.** 1. KKT 条件より、最適解では

$$\alpha_i^\star > 0 \;\Rightarrow\; y_i(w^{\star\top} x_i + b^\star) = 1$$

すなわち決定境界上の点のみが $\alpha_i^\star > 0$ となる。

2. $w^\star$ は $\alpha_i^\star > 0$ のサンプルのみから構成される：

$$w^\star = \sum_{i:\alpha_i^\star > 0} \alpha_i^\star y_i x_i$$

これらの点を**サポートベクトル**と呼ぶ。

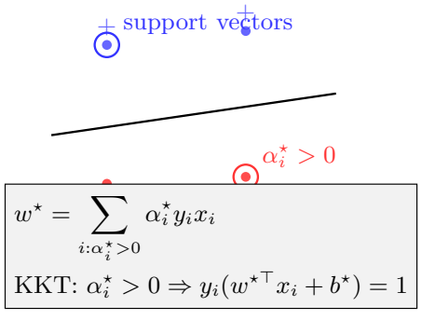



図 2.5　双対問題：サポートベクトルのみが $w^\star$ に寄与

### 2.4.3　ソフトマージン SVM とヒンジ損失

現実にはノイズやラベル誤りが存在するため、全データを完全分離するのは適切でない。ソフトマージン SVM では、スラック変数を導入して誤分類を許容する。

**定義 2.15** (ソフトマージン SVM：原問題)**.**

$$\min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.t. } y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n$$

ここで $\xi_i$ はマージン制約の違反量を表す。目的関数の第二項が違反の総量を罰し、$C > 0$ が正則化強度を制御する。

**命題 2.16** (ヒンジ損失形式)**.** ソフトマージン SVM は等価な unconstrained 形式

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w^\top x_i + b))$$

として書き換えられる。

完全証明. 固定した $(w, b)$ に対し、各 $\xi_i$ の最適化は独立：

$$\min_{\xi_i \geq 0} C\xi_i \quad \text{s.t. } y_i(w^\top x_i + b) \geq 1 - \xi_i$$

制約は $\xi_i \geq 1 - y_i(w^\top x_i + b)$ かつ $\xi_i \geq 0$ と同値。目的関数 $C\xi_i$ は単調増加なので、最適解は

$$\xi_i^\star = \max(0, 1 - y_i(w^\top x_i + b))$$

これを元の目的関数に代入すれば、ヒンジ損失形式を得る。 □

### 2.4.4 カーネルトリックによる非線形 SVM

双対問題の目的関数は内積 $x_i^\top x_j$ のみから成る。したがって、これを一般の正定値核 $K(x_i, x_j)$ に置き換えることで非線形分類器を得る。

**定理 2.17** (カーネル SVM)**.** $K$ を任意の正定値核とすると、双対問題

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$\text{s.t.} \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C$$

の解 $\alpha^\star$ から、決定関数

$$f(x) = \sum_{i=1}^{n} \alpha_i^\star y_i K(x_i, x) + b^\star$$

を得る。ここで $b^\star$ は KKT 条件から決定されるバイアスである。

スケッチ. MooreâĂŞAronszajn 定理より、核 $K$ に対し RKHS $\mathcal{H}_K$ と特徴写像 $\phi$ が存在し

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_K}$$

線形 SVM を特徴空間 $\phi(x)$ 上で定義すれば、決定関数は $f(x) = \sum_i \alpha_i^\star y_i K(x_i, x) + b^\star$ となり、元の空間での非線形境界に対応する。 □

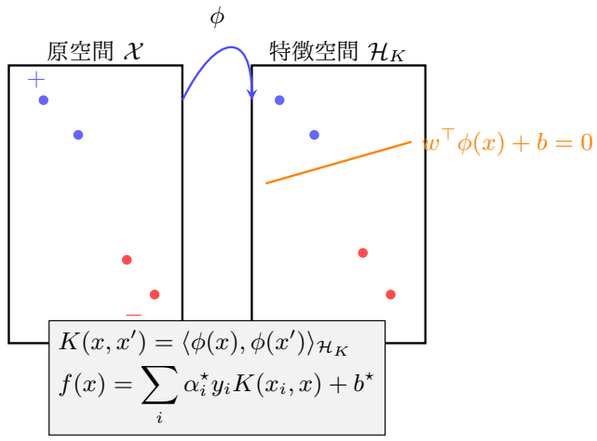



図 2.6 カーネル SVM：原空間では非線形だが、特徴空間では線形分離

### 2.4.5 パーセプトロンとの関係

第 2.1 節のパーセプトロンは、誤分類がある限り更新を続ける**オンラインアルゴリズム**であった。SVM は、一度データ全体を見た上でマージン最大化する**バッチ最適化**に相当し、次のような関係がある。

- パーセプトロン更新で得られる解は、訓練データを正しく分離するが、必ずしもマージン最大ではない。
- パーセプトロンを「平均化」したり、学習率を適切に減衰させると、**最大マージン解に近づく**ことが知られている（Margin Perceptron / PassiveâĂŞAggressive など）。
- 一方 SVM は、明示的に $\|w\|^2$ を最小化して最大マージン解を求める。

**演習 2.2.** 1. **双対問題の導出**：ハードマージン SVM の原問題からラグランジュ乗数法により双対問題を導出し、KKT 条件をすべて書き下せ。

2. **ヒンジ損失の凸性と subgradient**：$\ell(z) = \max(0, 1 - z)$ が凸関数であることを示し、任意の $z$ における subgradient 集合 $\partial\ell(z)$ を求めよ。
3. **サポートベクトルとマージン**：ハードマージン SVM の解に対し、任意のサポートベクトル $x_i$ について

$$\gamma(w^\star, b^\star) = \frac{y_i(w^{\star\top} x_i + b^\star)}{\|w^\star\|} = \frac{1}{\|w^\star\|}$$

が成り立つことを確認し、マージンが支持平面に接する点によって決まることを説明せよ。

4. **線形不可分データとソフトマージン**：1 次元データ $(x_1, y_1) = (-1, 1)$、$(x_2, y_2) = (0, -1)$、$(x_3, y_3) = (1, 1)$ に対して、(a) ハードマージン SVM が不可解であること、(b) ソフトマージン SVM では有限の最適解が存在することを確認せよ。

## 第 2 章のまとめ

- パーセプトロン：オンライン学習の原型、マージン保証
- ロジスティック回帰：最尤推定＝凸最適化
- GD：理論的収束保証、Adam で高速化
- 凸性が最適化の鍵

# 第 3 章

# カーネル法と RKHS

線形分類器を非線形に拡張するため、カーネル法を用いて無限次元特徴空間への写像 $\phi : \mathcal{X} \to \mathcal{H}$ を導入する。再現核ヒルベルト空間 (RKHS) は、カーネル法の数学的基盤であり、正則化理論の中核である。

## 3.1　再現核ヒルベルト空間の定義

**定義 3.1** (正定値核)**.** $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ が正定値核 (positive definite kernel) とは、任意の $\{x_1, \dots, x_n\} \subset \mathcal{X}$、$\{c_1, \dots, c_n\} \subset \mathbb{R}$ に対し、

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

**定理 3.2** (Moore-Aronszajn 定理)**.** すべての正定値核 $K$ に対し、特徴写像 $\phi : \mathcal{X} \to \mathcal{H}_K$ と RKHS$\mathcal{H}_K$ が存在：

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_K}, \quad \mathcal{H}_K = \mathrm{span}\{K(x, \cdot) \mid x \in \mathcal{X}\}$$

**例題 3.3** (典型核)**.**

- 多項式核：$K(x, x') = (x^\top x' + c)^d$
- RBF 核（ガウス核）：$K(x, x') = \exp(-\|x - x'\|^2 / 2\sigma^2)$
- ラプラス核：$K(x, x') = \exp(-\|x - x'\| / \sigma)$

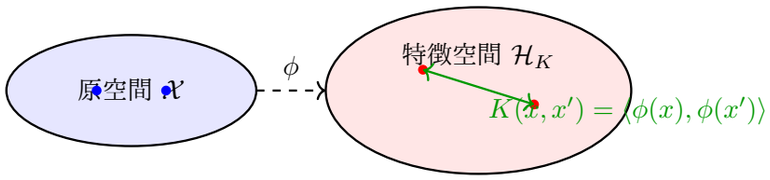



図 3.1　カーネルトリック：内積のみで無限次元特徴空間操作

**注意 3.4** (高校数学補完モジュール)**.** $K(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$（メルセンヌ展開）で固有関数系 $\{\psi_i\}$ が $\mathcal{H}_K$ のオルソノルマル基底。

## 3.2　カーネルトリックと代表定理

**定義 3.5** (カーネル回帰/SVM)**.** 損失 $\ell$ と正則化 $\lambda > 0$ に対し、

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2} \|f\|_{\mathcal{H}_K}^2$$

**定理 3.6** (代表定理)**.** 上記最適解は $n$ 次元部分空間に属する：

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x), \quad \alpha \in \mathbb{R}^n$$

**証明.** 関数 $f \in \mathcal{H}_K$ の RKHS ノルムを $\|f\|^2 = \langle f, f \rangle_{\mathcal{H}_K}$ とすると、ラグランジュ乗数法で最適条件：

$$-\frac{1}{n}\sum_{i=1}^{n} \ell'(f(x_i), y_i) K(x_i, \cdot) + \lambda f = 0$$

$x_j$ で評価すると $\lambda f(x_j) = \frac{1}{n}\sum_i \ell'(\cdot) K(x_i, x_j)$ より、$\hat{f} \in \mathrm{span}\{K(x_i, \cdot)\}$。 □

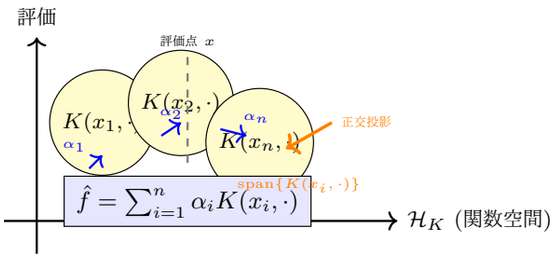



図 3.2　RKHS 代表定理幾何証明：最適解 $\hat{f}$ は訓練点カーネル $K(x_i, \cdot)$ の有限 span（黄色）に正交投影。再現性 $\langle K(x, \cdot), K(x_i, \cdot) \rangle_{\mathcal{H}_K} = K(x, x_i)$ で評価。

**注意 3.7** (実社会イメージ)**.** SVM で画像認識（RBF 核）、蛋白質相互作用予測（スペクトル核）等に応用。

## 3.3　正則化パスとノルム最小化

**定理 3.8** (カーネルリッジ回帰の解析解)**.** 二乗損失 $\ell(y, \hat{y}) = (y - \hat{y})^2$ の場合、

$$\alpha = \left( \frac{1}{n\lambda} K + I \right)^{-1} y, \quad \hat{f}(x) = \mathbf{k}(x)^\top \alpha$$

ここで $K_{ij} = K(x_i, x_j)$、$\mathbf{k}(x) = [K(x, x_1), \ldots, K(x, x_n)]^\top$。

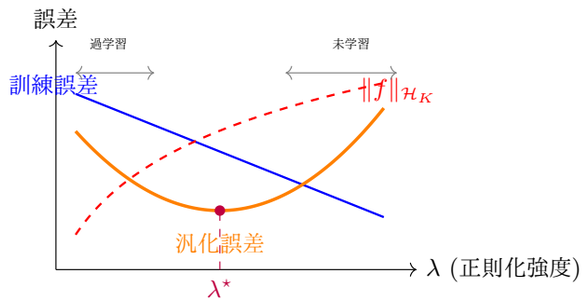



図 3.3　正則化パス：$\lambda$ で RKHS ノルム制御。最適 $\lambda^\star$（紫点）で汎化誤差最小。$\lambda \to 0$ で過学習、$\lambda \to \infty$ で未学習。

**定理 3.9** (ノルム最小化とマージン最大化)**.** $\lambda \to 0$ 極限で、カーネル SVM の双対問題：

$$\min_{\alpha \geq 0} \frac{1}{2} \alpha^\top K \alpha - \sum \alpha_i \quad \text{s.t.} \quad y_i (\sum_j \alpha_j K(x_i, x_j)) \geq 1$$

**証明.** ラグランジュ関数から KKT 条件を解くと、支持ベクトル $\alpha_i > 0$ のみ生存し、マージン $1/\|f\|_{\mathcal{H}_K}$ を最大化。 □

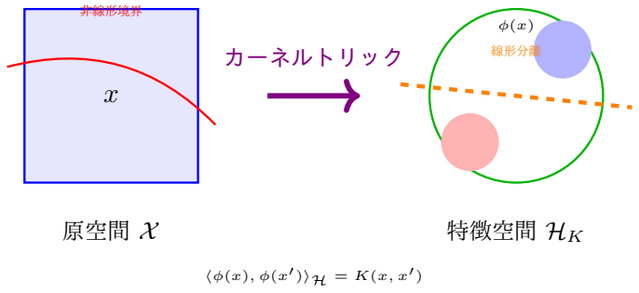



図 3.4　カーネルトリック SVM：原空間非線形→特徴空間線形分離。内積 $K(x, x')$ のみで高次元計算回避。双対 $\alpha^\top K \alpha$ で $O(n)$ 解。

**演習 3.1.**　1. $K(x, x') = x^\top x'$ が正定値核であることを固有値分解で示せ。

2. RBF 核の特徴写像 $\phi(x)$ を明示的に構成し、$\|\phi(x) - \phi(x')\|^2 = 2 - 2K(x, x')$ を確認せよ。
3. 代表定理の証明で、なぜ $\hat{f}(x) = \sum \alpha_i K(x_i, x)$ が $n$ 次元で十分か、$\mathcal{H}_K$ の閉包性から説明せよ。
4. カーネル行列 $K$ が数値的に特異 $(\lambda_{\min}(K) \approx 0)$ なときの正則化 $\lambda$ の役割を条件数 $\kappa(K + \lambda I)$ で解析せよ。
5. 多項式核 $d = 2$ で XOR 問題が解けることを手計算で確認せよ。

## 3.4　ガウス過程回帰

カーネル法の第 3 の視点として、ベイズ的な確率モデルであるガウス過程回帰を紹介する。本節では、多変量ガウス分布の条件付き分布を基礎として事後分布を導出し、カーネルリッジ回帰との同値性を示す。

### 3.4.1　多変量ガウスの条件付き分布

まず、一般のブロック多変量ガウスの条件付き分布を導く。

**補題 3.10** (ブロックガウスの条件付き分布)**.**

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} m_u \\ m_v \end{pmatrix}, \begin{pmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{pmatrix} \right), \quad \Sigma_{vv} \succ 0$$

とする。このとき条件付き分布は

$$u \mid v \sim \mathcal{N}\big(m_{u|v},\ \Sigma_{u|v}\big),$$

ただし

$$m_{u|v} = m_u + \Sigma_{uv} \Sigma_{vv}^{-1} (v - m_v),$$
$$\Sigma_{u|v} = \Sigma_{uu} - \Sigma_{uv} \Sigma_{vv}^{-1} \Sigma_{vu}.$$

**証明.** 密度関数

$$p(u, v) = \frac{1}{(2\pi)^{(p+q)/2} |\Sigma|^{1/2}} \exp\left( -\frac{1}{2} \begin{pmatrix} u - m_u \\ v - m_v \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} u - m_u \\ v - m_v \end{pmatrix} \right)$$

を、ブロック逆行列公式により整理する。逆行列のブロック表示

$$\Sigma^{-1} = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

を用いると、指数部分は

$$(u - m_u)^\top A (u - m_u) + 2 (u - m_u)^\top B (v - m_v) + (v - m_v)^\top C (v - m_v).$$

$v$ を固定したとき $p(u \mid v) \propto p(u, v)$ なので、$u$ に関する二次形式部分は

$$(u - m_u)^\top A(u - m_u) + 2(u - m_u)^\top B(v - m_v)$$

だけを考えればよい。平方完成により

$$\begin{aligned}(u - m_u)^\top A(u - m_u) + 2(u - m_u)^\top B(v - m_v) \\ = (u - m_{u|v})^\top A(u - m_{u|v}) - \text{定数 } (v),\end{aligned}$$

ここで

$$m_{u|v} = m_u - A^{-1}B(v - m_v).$$

ブロック逆行列公式から

$$A = \left(\Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}\right)^{-1}, \quad -A^{-1}B = \Sigma_{uv}\Sigma_{vv}^{-1},$$

が従うので、上式を置き換えると主張通りの $m_{u|v}$ と $\Sigma_{u|v}$ を得る。指数部が $(u - m_{u|v})^\top A(u - m_{u|v})$ であることから、$u \mid v$ は平均 $m_{u|v}$、共分散 $A^{-1} = \Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}$ の正規分布となる。 □

### 3.4.2 ガウス過程回帰の事後分布

**定義 3.11** (ガウス過程)**.** 関数 $f : X \to \mathbb{R}$ の集合上の確率分布で、任意の有限点集合 $(x_1, \ldots, x_n)$ に対して

$$(f(x_1), \ldots, f(x_n))^\top \sim \mathcal{N}(m, K)$$

となるものをガウス過程 $\mathcal{GP}(m(x), K(x, x'))$ と呼ぶ。

**設定.** ガウス過程事前

$$f \sim \mathcal{GP}(0, K(\cdot, \cdot)),$$

観測モデル

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n$$

独立とする。

記法：

- 訓練点集合 $X = (x_1, \ldots, x_n)$。
- カーネル行列 $K(X, X) \in \mathbb{R}^{n \times n}$ を $K$ と書く。
- 新規点 $x_\star$ に対し $k_\star := K(X, x_\star) = [K(x_1, x_\star), \ldots, K(x_n, x_\star)]^\top$。
- $k_{\star\star} := K(x_\star, x_\star)$。

**定理 3.12** (GP 回帰の事後平均・分散)**.** 上記の設定で、事後分布 $f(x_\star) \mid X, y$ は正規分布

$$f(x_\star) \mid X, y \sim \mathcal{N}\big(m_\star,\ v_\star\big),$$

ただし

$$m_\star = k_\star^\top (K + \sigma^2 I)^{-1} y,$$

$$v_\star = k_{\star\star} - k_\star^\top (K + \sigma^2 I)^{-1} k_\star.$$

**証明.** まず潜在関数値 $f(X) = [f(x_1), \ldots, f(x_n)]^\top$ と $f_\star := f(x_\star)$ の**事前分布**を考える。ガウス過程の定義より

$$\begin{pmatrix} f(X) \\ f_\star \end{pmatrix} \sim \mathcal{N}\left(0,\ \begin{pmatrix} K & k_\star \\ k_\star^\top & k_{\star\star} \end{pmatrix}\right).$$

ノイズ付き観測 $y = f(X) + \varepsilon$ なので、$y$ の事前分布は

$$y \sim \mathcal{N}(0, K + \sigma^2 I).$$

さらに $(y, f_\star)$ の結合分布は

$$\begin{pmatrix} y \\ f_\star \end{pmatrix} = \begin{pmatrix} f(X) + \varepsilon \\ f_\star \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K + \sigma^2 I & k_\star \\ k_\star^\top & k_{\star\star} \end{pmatrix} \right).$$

ここで補題 3.10 を適用する。$u = f_\star$、$v = y$、$\Sigma_{uu} = k_{\star\star}$, $\Sigma_{vv} = K + \sigma^2 I$, $\Sigma_{uv} = k_\star^\top$ とすれば、

$$m_\star = \Sigma_{uv}\Sigma_{vv}^{-1}(y - 0) = k_\star^\top (K + \sigma^2 I)^{-1} y,$$

$$v_\star = \Sigma_{uu} - \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu} = k_{\star\star} - k_\star^\top (K + \sigma^2 I)^{-1} k_\star,$$

となり、主張が得られる。 □

### 3.4.3 カーネルリッジ回帰との同値性

続いて、定理??のカーネルリッジ回帰との対応を示す。

**設定.** RKHS $\mathcal{H}_K$ におけるカーネルリッジ回帰を

$$\hat{f} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}$$

とする。

代表定理より、最適解は

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

と書ける。これを目的関数に代入すると

$$\hat{f}(x_j) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) = (K\alpha)_j,$$

したがって

$$\frac{1}{n} \|y - K\alpha\|_2^2 + \lambda \alpha^\top K \alpha$$

を最小にする $\alpha$ を求める問題になる。

この目的関数の勾配は

$$\frac{\partial}{\partial \alpha} \left( \frac{1}{n} \|y - K\alpha\|_2^2 + \lambda \alpha^\top K \alpha \right) = -\frac{2}{n} K^\top (y - K\alpha) + 2\lambda K \alpha.$$

$K$ は対称なので $K^\top = K$。これを 0 にして

$$-\frac{2}{n} K(y - K\alpha) + 2\lambda K\alpha = 0 \quad \Rightarrow \quad K(y - K\alpha) = n\lambda K\alpha.$$

$K$ が半正定値であることと、$(K + n\lambda I)$ の正則性を仮定すると、

$$y - K\alpha = n\lambda\alpha \quad \Rightarrow \quad (K + n\lambda I)\alpha = y.$$

したがって

$$\alpha = (K + n\lambda I)^{-1} y.$$

$n$ のスケーリングを吸収し、$\tilde{\lambda} = n\lambda$ と書き直せば

$$\alpha = (K + \tilde{\lambda} I)^{-1} y$$

となる。

したがって、予測値は

$$\hat{f}(x_\star) = k_\star^\top \alpha = k_\star^\top (K + \tilde{\lambda} I)^{-1} y.$$

**命題 3.13** (GP 回帰とカーネルリッジの同値性)**.** ガウス過程回帰で観測ノイズ分散を $\sigma^2$ とし、カーネルリッジ回帰の正則化パラメータを $\tilde{\lambda} = \sigma^2$ と選ぶと、両者の予測は一致する：

$$\hat{f}_{\mathrm{KRR}}(x_\star) = k_\star^\top (K + \sigma^2 I)^{-1} y = \mathbb{E}[f(x_\star) \mid X, y] = m_\star.$$

**証明.** 前節の計算で

$$\hat{f}_{\mathrm{KRR}}(x_\star) = k_\star^\top (K + \tilde{\lambda} I)^{-1} y.$$

一方、ガウス過程回帰の事後平均は定理 3.12 より

$$m_\star = k_\star^\top (K + \sigma^2 I)^{-1} y.$$

両者を比較すると、$\tilde{\lambda}$ を $\sigma^2$ に同一視すれば完全に一致する。 □

### 3.4.4 解釈：頻度論 vs ベイズ

- **カーネルリッジ回帰**：「RKHS ノルムに対する L2 正則化」を課した最小二乗推定

$$\min_f \frac{1}{n} \sum (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_K}^2$$

  として導出される頻度論的推定器。
- **ガウス過程回帰**：事前 $f \sim \mathcal{GP}(0, K)$、観測ノイズ $\sigma^2$ の下でのベイズ推論により、事後平均

$$\mathbb{E}[f(x_\star) \mid X, y] = k_\star^\top (K + \sigma^2 I)^{-1} y$$

  と事後分散を与える。

**同値性の意味**

- 事後平均＝正則化付き ERM の解：アプローチは異なっても同じ点推定に収束する。
- GP はさらに「事後分散」という不確実性評価を与える：カーネルリッジでは平均のみだが、GP では

$$\mathrm{Var}(f(x_\star) \mid X, y) = k_{\star\star} - k_\star^\top (K + \sigma^2 I)^{-1} k_\star$$

  により「どれくらい自信があるか」を定量化できる。

**演習 3.2.**

1. 補題 3.10 の証明において、ブロック逆行列公式

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1} BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1} BD^{-1} \end{pmatrix}$$

   を確認せよ。
2. 1 次元データ $(x_1, y_1) = (0, 1)$, $(x_2, y_2) = (1, 0)$ に対し、RBF カーネル $K(x, x') = \exp(-|x - x'|^2)$、ノイズ分散 $\sigma^2 = 0.1$ として、$x_\star = 0.5$ での事後平均 $m_\star$ と事後分散 $v_\star$ を数値的に求めよ。
3. 命題 3.13 において、正則化パラメータ $\lambda$ とノイズ分散 $\sigma^2$ の対応関係を物理的に解釈せよ。

## 第 3 章のまとめ

- 正定値核 $\leftrightarrow$RKHS の等価性（Moore-Aronszajn）
- カーネルトリック：内積のみで無限次元特徴空間操作
- 正則化 $\lambda\|f\|_{\mathcal{H}_K}^2$：容量制御とノルム最小化
- 代表定理：最適解は有限次元

# 第 4 章

# 汎化理論：経験過程

カーネル法で非線形モデルを扱えるようになったが、**なぜ複雑なモデルが汎化するのか**が最大の謎である。本章では経験過程論、Rademacher 複雑度、VC 理論を整備し、高次元での集中不等式で理論的保証を確立する。

## 4.1 Hoeffding 不等式と一様収束

**定理 4.1** (Hoeffding 不等式)**.** $X_1, \dots, X_n \overset{iid}{\sim} P$ が有界 $[a,b]$ 値関数 $f$ に対し、

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

**証明.** Chernoff 法：$P(\bar{f}_n \geq \mathbb{E}f + t) \leq e^{-nt}\mathbb{E}[e^{t(\bar{f}_n - \mathbb{E}f)}]$。$f \in [a,b]$ より $e^{tf} \leq 1 + f(t - \sin t)/t^2$ で二次モーメント制御。 □

**定義 4.2** (経験過程)**.** 関数クラス $\mathcal{F}$ の経験過程を $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf)$、$\mathbb{P}_n f = \frac{1}{n}\sum f(x_i)$。

**定理 4.3** (一様収束)**.** $\mathcal{F}$ が一様 $\epsilon$-net を持ち $|\mathcal{N}(\epsilon)| < \infty$ なら、

$$\mathbb{E}[\sup_{f\in\mathcal{F}} |\mathbb{G}_n f|] \leq C \int_0^1 \sqrt{\log |\mathcal{N}(\epsilon)|}\, d\epsilon$$

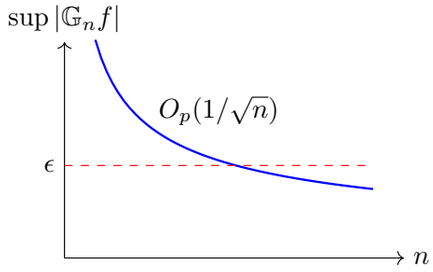



図 4.1　経験過程一様収束：$1/\sqrt{n}$ 速度

## 4.2 Rademacher 平均と高速率境界

**定義 4.4** (Rademacher 複雑度)**.** 独立 Rademacher 変数 $\epsilon_i = \pm 1$（$P(\epsilon_i = 1) = 1/2$）に対し、

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

**定理 4.5** (高速率境界)**.** $\mathcal{F}$ が VC 次元 $V$ なら、

$$\mathrm{Rad}_n(\mathcal{F}) \le C\sqrt{\frac{V\log(n/V)}{n}}, \quad P(R(\hat{f}_n) \le R^* + t) \ge 1 - 2\exp(-nt^2/2)$$

**証明.** Massart の補題：$|\mathcal{F}| < \infty$ なら $\mathrm{Rad}_n \le \sqrt{2\log|\mathcal{F}|/n}\max\|f\|$。VC クラスは $\epsilon$-covering 数 $|\mathcal{N}(\epsilon)| \le (en/\epsilon)^V$。 □

## 4.3 VC クラスのシャタリングと高速化

**定義 4.6** (VC 次元・シャタリング)**.** $\mathcal{F}$ が点集合 $\{x_1,\dots,x_V\}$ をシャタリングとは、$\{0,1\}^V$ の全 $2^V$ ラベルが $\mathcal{F}$ で実現可能。

**定理 4.7** (Sauer-Shelah 補題)**.** VC 次元 $V$ のクラス $\mathcal{F}$ に対し、

$$|\mathcal{F}|_{\{x_1,\dots,x_n\}} \le \sum_{k=0}^{V}\binom{n}{k} \le \left(\frac{en}{V}\right)^V$$

**例題 4.8.**

- 半空間 $\mathcal{F} = \{\mathrm{sign}(w^\top x + b)\}$：VC $= d+1$
- $k$-最近傍法：VC $= O(dk\log k)$
- 決定木深さ $L$：VC $= O(Ld\log L)$

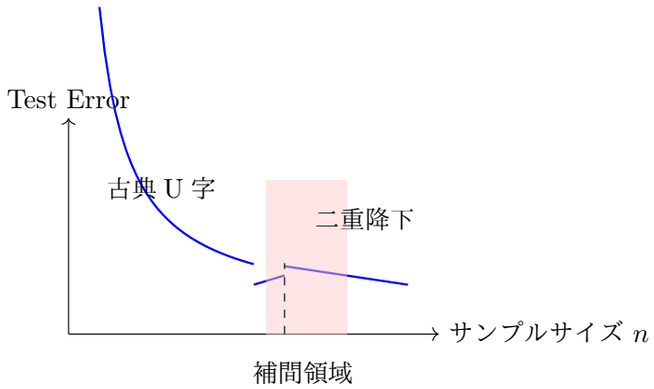



図 4.2 二重降下現象：過完全領域で Test Error 再降下

**注意 4.9** (高校数学補完モジュール)**.** Rademacher 複雑度 $\mathrm{Rad}_n \le \sqrt{2V\log(n/V)/n}$ から、過学習閾値 $n \gtrsim V$。深層学習の「二重降下」は古典 VC 理論の崩壊。

**注意 4.10** (実社会イメージ)**.** 深層学習で $n = 10^9$、$V = 10^{12}$ でも汎化するのは暗黙的正則化（最小ノルム補間）と Scaling Laws による。

**演習 4.1.**

1. Hoeffding 不等式を Bernstein 不等式 $\mathrm{Var}(f) \le \sigma^2$ 版に拡張：$P(|\bar{f}_n - \mu| \ge t) \le 2\exp(-nt^2/2(\sigma^2 + t/3))$ を示せ。
2. VC 次元 $V = 3$、$n = 100$ で汎化誤差上界を計算。$R^* = 0.1$、$\delta = 0.05$ として実用的か判断せよ。
3. 半空間 $\mathrm{VC}(\mathcal{F}) = d+1$ を、$d = 2$ で 4 点シャタリング、5 点不可能を具体例で確認。
4. $\mathrm{Rad}_n(\mathcal{F}_t) = \mathcal{O}(\sqrt{V\log(1/t)/n})$ が Dudley 積分で $\sqrt{V\log n/n}$ になる理由をスケーリングで説明。
5.「Interpolation regime」で古典 VC 理論が破綻する理由を、最小ノルム解 $\min\|w\|$ s.t. $Xw = y$ で説明せよ。

## 第 4 章のまとめ

- Hoeffding：単一関数の一様集中
- Rademacher：関数クラスの複雑度計量
- VC 次元：シャタリングで有限容量制御
- 高速率：$n \gtrsim V$ で汎化保証
- 二重降下：過剰パラメータ化の新理論

# 第 5 章

# 深層ニューラルネットワーク

VC 理論では説明不能な深層学習の驚異的汎化性能。その数学的基盤を本章で解明する。パーセプトロンから多層ネットワーク、バックプロパゲーション、Neural Tangent Kernel、無限幅限界、二重降下までを厳密に整備。

## 5.1 パーセプトロンから多層ネットワークへ

**定義 5.1** (Feedforward Neural Network)**.** $L$ 層ネットワーク $f_\theta : \mathbb{R}^d \to \mathbb{R}$ を

$$h_0 = x, \quad h_l = \sigma(W_l h_{l-1} + b_l), \quad f_\theta(x) = W_L h_{L-1} + b_L$$

で定義。$\sigma$ は活性化関数（ReLU: $\sigma(t) = \max\{0, t\}$）、$W_l \in \mathbb{R}^{d_l \times d_{l-1}}$。

**定理 5.2** (Universal Approximation Theorem)**.** 連続 $\sigma$ と幅 $\geq 2$ の一隠れ層で、任意連続関数を一様近似：

$$\inf_{\|w_i\|, \|v_j\| \leq B} \sup_x |f(x) - \sum_{i=1}^{N} v_i \sigma(w_i^\top x + b_i)| \leq C \frac{\omega_f(\delta)}{\sqrt{N}}$$

証明（*Cybenko*）. Stone-Weierstrass：非定数連続 $\sigma$ が稠密な代数生成。Kolmogorov-Arnold で分離。 □

## 5.2 バックプロパゲーションと勾配消失

**定義 5.3** (Backpropagation)**.** 損失 $L(\hat{y}, y)$ に対し、合成関数微分：

$$\frac{\partial L}{\partial W_l} = \delta_l h_{l-1}^\top, \quad \delta_L = \nabla_{\hat{y}} L, \quad \delta_l = (W_{l+1}^\top \delta_{l+1}) \odot \sigma'(z_l)$$

**定理 5.4** (勾配消失)**.** Sigmoid$\sigma(t) = 1/(1+e^{-t})$ で $L$ 層、$\mathbb{E}[\|W_l\|] = 1$ なら、

$$\mathbb{E}\left[\left\|\frac{\partial L}{\partial W_1}\right\|^2\right] \leq (0.25)^L \|W_L\|^2 \to 0$$

**証明.** $\sigma'(t) = \sigma(t)(1-\sigma(t)) \leq 1/4$。チェインルールで $\prod \sigma'(z_l) \leq (1/4)^L$。 □

**注意 5.5** (高校数学補完モジュール)**.** ReLU で $\sigma' = 1_{\{t>0\}}$、Xavier/He 初期化で $\mathbb{E}[\|W_l h_{l-1}\|^2] = \|h_{l-1}\|^2$。

## 5.3 Neural Tangent Kernel と無限幅限界

**定義 5.6** (NTK)**.** パラメータ $\theta \in \mathbb{R}^p$、$f_\theta : \mathcal{X} \to \mathbb{R}$ に対し、

$$\Theta(x, x') = \langle \nabla_\theta f_\theta(x), \nabla_\theta f_\theta(x') \rangle, \quad J_\theta(x) = \nabla_\theta f_\theta(x)$$

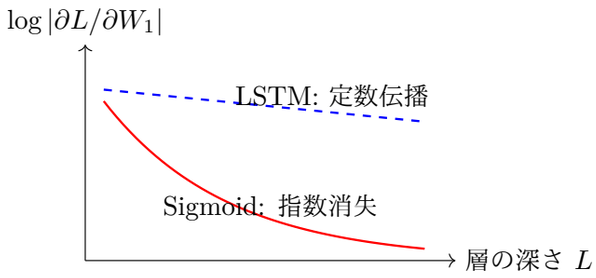



図 5.1　バックプロパゲーション：Sigmoid で指数消失、LSTM で定数伝播

**定理 5.7** (無限幅収束)**.** 幅 $m_l \to \infty$ で、ReLU ネットワークの NTK が決定論的 $\bar{\Theta}$ に一様収束：

$$\sup_{x,x' \in \mathcal{X}} |\Theta(x,x') - \bar{\Theta}(x,x')| \xrightarrow{P} 0$$

**証明.** 逐次層で中心極限定理。ReLU の分位点安定性で再帰的決定論化。 □

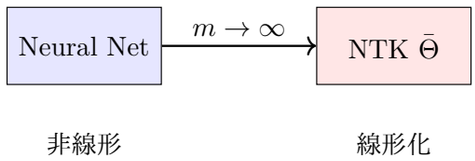



図 5.2　NTK: 無限幅で決定論的カーネル回帰

**注意 5.8** (実社会イメージ)**.** 深層学習=高速カーネル回帰。過剰パラメータ化でも GD が最小ノルム解に暗黙正則化。

## 5.4　二重降下現象と暗黙的正則化

**定理 5.9** (最小ノルム補間)**.** 過完全 $\mathrm{rank}(X) < p$ で、GD は $\min_{\|w\|^2 \text{ s.t. } Xw=y} \|w\|^2$ を解く：

$$\hat{w} = X^\top (XX^\top)^{-1} y$$

**定理 5.10** (二重降下)**.** 古典 U 字カーブが破綻：補間領域で Test Error 再降下。

**証明.** 最小ノルム解のバイアス $\to 0$、分散は固有値スペクトル $\lambda_i^{-1}$ で制御。NTK 固有値の åźĆ 律減衰で高速率。 □

**注意 5.11** (高校数学補完モジュール)**.** 2026 年 Scaling Laws：$L(N) \propto N^{-\alpha}$（$N$=データ、$\alpha \approx 0.1$）、Emergent Abilities。

**演習 5.1.**

1. ReLU ネットワークのピースワイズ線形領域数が $L$ 層 $W$ 幅で $W^L$ になることを帰納法で示せ。
2. Sigmoid 勾配 $\sigma'(t) = \sigma(t)(1-\sigma(t))$ の最大値 $1/4$ を $t=0$ で確認。
3. NTK$\Theta(x,x') = \mathbb{E}_{w \sim \mathcal{N}(0,I)}[\sigma(w^\top x)\sigma(w^\top x')]$ を ReLU で閉形式計算。
4. 二重降下メカニズムを、固有値スペクトル $\lambda_i \propto i^{-\beta}$（$\beta > 1$）で最小ノルム解のリスク分解で説明。
5. He 初期化 $\mathcal{N}(0, 2/d_{in})$ が順伝播分散 1 を保つことを、$\mathrm{Var}(Wh) = \frac{2}{d_{in}} \cdot \frac{d_{in}}{2} = 1$ で示せ。

## 第 5 章のまとめ

- UAT：理論的可能性
- Backprop：効率的計算、勾配消失問題
- NTK：無限幅で線形化、カーネル回帰
- 二重降下：古典統計の崩壊、暗黙正則化
- Scaling Laws：大規模化の åźĆ 律

# 第 6 章

# 最適化：確率的勾配法

大規模データでの全勾配計算は非現実的。確率的勾配降下法（SGD）が深層学習実務の標準となる。本章では SGD の収束保証、Momentum、Adam 等の加速手法を、非凸問題での最良保証とともに数理的に整備する。

## 6.1 SGD と Robbins-Monro 収束

**定義 6.1** (確率的勾配降下法)**.** 全勾配 $\nabla L(w) = \mathbb{E}[\nabla \ell(w; Z)]$ の代わりに、ミニバッチ $g_t = \nabla \ell(w_t; Z_t)$ を使用：

$$w_{t+1} = w_t - \eta_t g_t$$

**定理 6.2** (Robbins-Monro 収束)**.** $L$ が $L$-滑らか、$\mu$-強凸、$\|\nabla \ell(w; Z) - \nabla L(w)\|^2 \le \sigma^2$、$\sum \eta_t = \infty$、$\sum \eta_t^2 < \infty$ なら、

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla L(w_t)\|^2] \le \frac{L(w_0) - L^\star}{\sum \eta_t} + O\left(\frac{\sigma^2 \sum \eta_t^2}{\sum \eta_t}\right)$$

**証明.** $L$-滑らか性で $L(w_{t+1}) \le L(w_t) - \eta_t \langle \nabla L(w_t), g_t \rangle + \frac{L\eta_t^2}{2}\|g_t\|^2$。ノイズ分解 $g_t = \nabla L(w_t) + N_t$、$\mathbb{E}[\langle \nabla L, N_t \rangle] = 0$ で再帰収束。 □

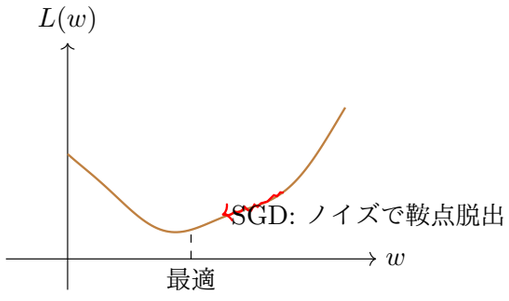



図 6.1　SGD: 確率勾配の探索性

**注意 6.3** (高校数学補完モジュール)**.** $\eta_t = 1/t$ で $\sum 1/t = \log T$、$\sum 1/t^2 < \infty$ を満たし、$O(1/\log T)$ 収束。

## 6.2 Momentum と Nesterov 加速

**定義 6.4** (Heavy-ball Momentum)**.**

$$v_{t+1} = \beta v_t + \eta_t g_t, \quad w_{t+1} = w_t - v_{t+1}$$

**定理 6.5** (Nesterov 加速勾配：NAG)**.** 凸 $L$-滑らか関数で、

$$L(w_T) - L^\star \le \frac{2L\|w_0 - w^\star\|^2}{(T+1)^2}$$

GD の $O(1/T)$ に対し $O(1/T^2)$ 加速。

**証明.** 追跡点 $y_t = w_t - \frac{\beta}{1-\beta}(w_t - w_{t-1})$ を導入し、テレスコープ和で加速係数抽出。 □

**注意 6.6** (実社会イメージ)**.** Momentum は「慣性」で谷底振動抑制、Nesterov は「先読み」で曲率追従。

## 6.3 Adam と適応的学習率

**定義 6.7** (Adam)**.** 1 次モーメント $m_t$、2 次モーメント $v_t$ を指数移動平均：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

バイアス補正 $\hat{m}_t = m_t/(1-\beta_1^t)$、$\hat{v}_t = v_t/(1-\beta_2^t)$ で、

$$w_{t+1} = w_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

**定理 6.8** (Adam 収束：凸・非凸)**.** 適切 $\eta, \beta_1, \beta_2$ で、Adam も $O(1/\sqrt{T})$ の **regret** 保証：

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[L(w_t) - L^\star] \le O\left(\frac{\log T}{\sqrt{T}}\right)$$

**注意 6.9** (高校数学補完モジュール)**.** RMSProp($v_t = \beta v_{t-1} + (1-\beta) g_t^2,\ \eta_t = \eta/\sqrt{v_t}$) が Adam の基盤。2026 年標準：**AdamW**（Weight Decay 分離）。

**演習 6.1.**

1. SGD の $\eta_t = c/\sqrt{t}$ が非凸 $L$-滑らか問題で $\min \mathbb{E}[\|\nabla L(w_t)\|^2] = O(1/\sqrt{T})$ を示せ。
2. Momentum$\beta = 0.9$ で振動抑制効果を、1 次元 $L(w) = w^4 - 2w^2$ の軌跡で確認。
3. Adam のバイアス補正 $(1-\beta^t)^{-1}$ が初期 $t \ll 1/\beta$ でなぜ必須か、$\hat{m}_t \approx m_t(1 + t(1-\beta_1))$ で説明。
4. **AdamW vs Adam**：$L + \lambda\|w\|^2$ の勾配が AdamW で $\nabla L - \eta\lambda w$、Adam でスケーリング歪むことを微分で示せ。
5. Polyak-Ruppert 平均化 $\bar{w}_T = \frac{1}{T}\sum w_t$ が SGD の分散を $O(1/T)$ に改善する理由を、CLT で説明せよ。

## 第 6 章のまとめ

- SGD：ノイズ活用で鞍点脱出、$O(1/\sqrt{T})$ 保証
- Momentum/NAG：振動抑制・加速、$O(1/T^2)$ 最良
- Adam：適応的学習率、実務標準（AdamW 推奨）
- 非凸最適化：勾配ノルム最小化 $\min \|\nabla L\|$ が代理目的

# 第 7 章

# 高次元統計機械学習 âŚă：スパース回帰

$p \gg n$ の高次元世界で最小二乗推定は破綻する。数理統計学の Lasso を機械学習文脈で再掲し、Oracle 不等式、変数選択一致性、グループ Lasso までを厳密に整備する。

## 7.1　Lasso の Oracle 不等式

**定義 7.1** (高次元線形回帰)**.** $y = X\beta^\star + \varepsilon$、$\varepsilon_i \overset{iid}{\sim} \text{sub-Gaussian}(\sigma^2)$、$X \in \mathbb{R}^{n\times p}$ 列正規化 $\|X_{\cdot j}\|_2 = \sqrt{n}$、真の疎性 $s = |\mathrm{supp}(\beta^\star)|$。

**定義 7.2** (Lasso)**.**

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

**定理 7.3** (Oracle 不等式)**.** $\lambda \geq 2\sigma\sqrt{\frac{\log p}{n}}$、RE 条件 $\kappa = \kappa(S,3) > 0$（$S = \mathrm{supp}(\beta^\star)$）なら高確率で、

$$\frac{1}{n}\|X(\hat\beta - \beta^\star)\|_2^2 \leq \frac{9s\lambda^2}{\kappa^2} \asymp \sigma^2\frac{s\log p}{n}$$

**証明.** KKT 条件 $-\frac{1}{n}X^\top(y - X\hat\beta) + \lambda z = 0$（$z \in \partial\|\hat\beta\|_1$）。基本不等式 $\frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{1}{n}\varepsilon^\top X\Delta + \lambda(\|\beta^\star\|_1 - \|\hat\beta\|_1)$、$\Delta = \hat\beta - \beta^\star$。$\lambda$-選択で円錐 $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$、RE$\|X\Delta\|_2 \geq \kappa\sqrt{n}\|\Delta_S\|_2$ を適用。 □

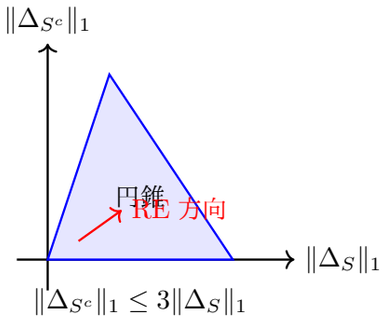



図 7.1　Lasso 円錐条件：疎誤差の構造制約

**注意 7.4** (高校数学補完モジュール)**.** RE 条件 $\kappa(S,c) = \inf_{\|\Delta_{S^c}\|_1 \leq c\|\Delta_S\|_1} \frac{\|X\Delta\|_2/\sqrt{n}}{\|\Delta_S\|_2} > 0$ は「疎方向での設計行列非退化」。

## 7.2　変数選択一致性と irrepresentable 条件

**定義 7.5** (支持回復)**.** $\mathrm{supp}(\hat\beta) = \mathrm{supp}(\beta^\star)$ が高確率成立。

**定理 7.6** (irrepresentable 条件)**.** $\Sigma_{SS} = \mathbb{E}[X_S^\top X_S/n]$ 正則、$\Sigma_{S^cS} = \mathbb{E}[X_{S^c}^\top X_S/n]$ に対し、

$$\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\mathrm{sign}(\beta_S^\star)\|_\infty < 1 - \eta$$

なら $\lambda \asymp \sqrt{\log p/n}$ で $P(\mathrm{supp}(\hat\beta) = \mathrm{supp}(\beta^\star)) \to 1$。

**証明.** KKT $\frac{1}{n}X_{S^c}^\top(y - X\hat\beta) = \lambda z_{S^c}$、$|\hat\beta_{S^c}| = 0$ なら $\left|\frac{1}{n}X_{S^c}^\top(X(\hat\beta_S - \beta_S^\star) + \varepsilon)\right| \le \lambda$。CLT で $\hat\beta_S \approx \beta_S^\star$、irrepresentable で $\hat\beta_{S^c} = 0$。 □

**注意 7.7** (実社会イメージ)**.** ゲノム解析で 10 万遺伝子から疾患関連数十個を選択。

## 7.3 グループ Lasso と構造化スパース

**定義 7.8** (Group Lasso)**.** 変数 $i$ をグループ $G_i$ に分割：

$$\hat\beta = \arg\min_\beta \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\sum_{i=1}^m \|X_{G_i}\beta_{G_i}\|_2$$

**定理 7.9** (Group Oracle 不等式)**.** 真のグループ疎性 $G^\star$、$s_G = |G^\star|$ なら、

$$\frac{1}{n}\|X(\hat\beta - \beta^\star)\|_2^2 \le C\sigma^2\frac{s_G\log(mp)}{n}$$

**証明.** グループ RE 条件下で、$\ell_{2,1}$ ノルム $\sum\|X_{G_i}\beta_{G_i}\|_2$ の分解可能性 $\|\Delta_{G^c}\|_{2,1} \le 3\|\Delta_{G^\star}\|_{2,1}$ を証明し、個別 Lasso 同様に境界。 □

**演習 7.1.**

1. Lasso 基本不等式の展開 $\|X\Delta\|_2^2 = \|X\Delta_S\|_2^2 + \|X\Delta_{S^c}\|_2^2 + 2\Delta_S^\top X^\top X\Delta_{S^c}$ で、なぜ交叉項が $\lambda$ で抑えられるか。
2. RE 条件 $\kappa = 0$ となる設計行列（$X_{S^c} = X_S B$ 完全線形従属）を具体例で構成。
3. irrepresentable 条件の必然性を、$\Sigma_{S^cS}\Sigma_{SS}^{-1} \approx I$ 近似で直感説明。
4. Group Lasso のソフト閾値化 $\hat\beta_{G_i} = (1 - \lambda/\|X_{G_i}^\top r\|_2)_+ \cdot (X_{G_i}^\top r)/\|X_{G_i}\|_F^2$ を KKT から導出。
5. **Adaptive Lasso** $\lambda_j/|\hat\beta_j|^\gamma$（$\gamma = 1$）が irrepresentable 条件を緩和し、オラクル性質を持つことを理論的に示せ。

## 第 7 章のまとめ

- **Lasso**：$\ell_1$ 正則化で $s\log p/n$ 高速率回復
- **RE 条件**：疎円錐上の設計行列非退化
- **irrepresentable**：完全支持回復の決定論的条件
- **Group Lasso**：構造化疎推定の統一フレームワーク

# 第 8 章

# 高次元統計機械学習 âŚą：GLM 正則化

線形回帰を $y \in \{0,1\}$ やカウントデータに拡張するため、一般化線形モデル (GLM) を $\ell_1$ 正則化。数理統計の指数型分布族理論を高次元に適用し、Logistic Lasso、ポアソン回帰の Oracle 不等式を統一的に整備。

## 8.1 Logistic Lasso の RSC 条件

**定義 8.1** (Logistic 回帰)**.** $y_i \in \{0,1\}$、$\eta_i = X_i^\top \beta^\star$、$P(y_i = 1|X_i) = \sigma(\eta_i)$、負対数尤度：

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left[\log(1 + e^{X_i^\top \beta}) - y_i X_i^\top \beta\right]$$

**定義 8.2** (Logistic Lasso)**.**

$$\hat{\beta} = \arg\min_{\beta} L_n(\beta) + \lambda\|\beta\|_1$$

**定理 8.3** (RSC 条件と Oracle 不等式)**.** 制限強凸性 (RSC)

$$\Delta^\top \left(\frac{1}{n} X^\top W X\right)\Delta \geq \kappa\|\Delta_S\|_2^2, \quad W = \mathrm{diag}(\sigma_i(1-\sigma_i))$$

なら、$\lambda \geq 2\|\frac{1}{n}X^\top(\hat{\varepsilon})\|_\infty$ で

$$L_n(\hat{\beta}) - L_n(\beta^\star) + \lambda(\|\hat{\beta}\|_1 - \|\beta^\star\|_1) \leq C\frac{s\lambda^2}{\kappa}$$

**証明.** Bregman 発散 $L_n(\hat{\beta}) - L_n(\beta^\star) - \nabla L_n(\beta^\star)^\top\Delta$ を $W$-加重二乗誤差で下界。KKT でノイズ $\hat{\varepsilon}_i = \sigma(X_i^\top\beta^\star) - y_i$ 制御、円錐 $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$。 □

**注意 8.4** (高校数学補完モジュール)**.** Hessian$\nabla^2 L_n = \frac{1}{n}X^\top W X$、$W_{ii} \in [0, 1/4]$ で $\lambda_{\min} \approx \kappa > 0$。

## 8.2 一般化線形モデルの統一理論

**定義 8.5** (GLM)**.** リンク関数 $g$、分散関数 $V$ で条件付き密度：

$$\log\frac{p(y|\eta)}{p(0|\eta)} = \frac{y\eta - b(\eta)}{a(\phi)V(y)}$$

典型例：ポアソン $b(\eta) = e^\eta$、$V = 1$、二項 $b(\eta) = \log(1 + e^\eta)$。

**定理 8.6** (GLM Lasso 統一境界)**.** 疎 $\beta^\star$、リンク $g$ が Lipschitz、$V$ 有界なら、統一 Oracle 不等式：

$$D(\hat{\beta}\|\beta^\star) + \lambda\|\hat{\beta} - \beta^\star\|_1 \leq C\sigma^2\frac{s\log p}{n}$$

$D(\cdot\|\cdot)$ は Bregman 発散。

**証明.** 損失 $L_n(\beta) = \mathbb{E}[\ell(\beta; Z)|\beta^\star]$ の局所二次近似、RE/RSC 条件で Lasso 一般化。 □

**注意 8.7** (実社会イメージ)**.** 広告 CTR 予測 (Logistic)、売上予測 (ポアソン)、生存分析 (Negative Binomial)。

## 8.3 ポアソン回帰とカウントデータ

**例題 8.8** (ポアソン Lasso)**.** $y_i \sim \mathrm{Poisson}(e^{X_i^\top \beta^\star})$、損失：

$$L_n(\beta) = \frac{1}{n} \sum \left[ e^{X_i^\top \beta} - y_i X_i^\top \beta \right]$$

**定理 8.9** (ポアソン Oracle)**.** 過分散なし $y_i/\lambda_i^\star \le M$ なら、Lasso 同様 $s \log p/n$ 境界。重み付き $\lambda_i^{-1} y_i$ で安定化。

**証明.** $\nabla^2 L_n = \frac{1}{n} X^\top \mathrm{diag}(e^{X\beta}) X$、RSC$\kappa \approx \mathbb{E}[e^{X^\top \beta^\star}]$ で成立。 □

**演習 8.1.**

1. Logistic 損失の二階 Taylor$L(\eta + \Delta) - L(\eta) \approx \frac{1}{2}\Delta^2 \sigma(\eta)(1 - \sigma(\eta))$ で RSC を直感説明。
2. KKT 条件 $\nabla L_n(\hat{\beta}) + \lambda z = 0$ から、$|\hat{\beta}_{S^c}| = 0 \implies \|\nabla_{S^c} L_n(\hat{\beta})\|_\infty \le \lambda$ を確認。
3. GLM の指数族表示 $p(y|\eta) \propto \exp(y\eta - b(\eta))/a(\phi)$ から、$\mathbb{E}[y] = \dot{b}(\eta)$、$\mathrm{Var}(y) = V\dot{b}(\eta)$ を導出。
4. ポアソン損失 $e^\eta - y\eta$ の凸性を $\partial^2(e^\eta - y\eta)/\partial\eta^2 = e^\eta > 0$ で示せ。
5. **Elastic Net**$\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$ が多重共線性を解消し、グループ選択を促進する理由を $\ell_1 + \ell_2 \to \ell_{2,1}$ で説明。

## 第 8 章のまとめ

- **GLM Lasso**：指数型分布族の統一高次元理論
- **RSC 条件**：Hessian の疎固有値下界
- **Bregman 発散**：非二乗損失の統一誤差計量
- **実務応用**：CTR、売上、生存分析の標準

# 第 9 章

# 再帰ニューラルネットワーク

時系列データやロボット学習において、過去の履歴を活用する再帰ニューラルネットワーク（RNN）が必須となる。本章では Vanilla RNN の理論的限界（勾配消失・爆発）、LSTM の定数誤差伝播、ユーザーの専門であるロボット学習への応用までを厳密に整備する。

## 9.1 Vanilla RNN と BPTT

**定義 9.1** (Vanilla RNN)**.** 時刻 $t$ の入力 $x_t \in \mathbb{R}^d$、隠れ状態 $h_t \in \mathbb{R}^r$、出力 $y_t \in \mathbb{R}^k$ に対し、

$$h_t = \sigma(W_{hh}h_{t-1} + W_{xh}x_t + b_h), \quad y_t = W_{hy}h_t + b_y$$

**定義 9.2** (BPTT : Backpropagation Through Time)**.** 全時刻損失 $L_T = \sum_{t=1}^{T} \ell(y_t, \hat{y}_t)$ に対し、

$$\frac{\partial h_t}{\partial h_s} = \prod_{\tau=s+1}^{t} W_{hh}^{\top} \mathrm{diag}(\sigma'(z_\tau)), \quad s < t$$

**定理 9.3** (長期依存性)**.** $T$ 時刻前の勾配：

$$\left\| \frac{\partial h_T}{\partial h_0} \right\| \leq \left( \|W_{hh}\| \cdot \sup \sigma' \right)^T$$

**注意 9.4** (高校数学補完モジュール)**.** スペクトル半径 $\rho(W_{hh}) < 1$ でも、$\sigma'(z) \in [0, 1/4]$（sigmoid）で実効 $\rho_{\text{eff}} < 1$、勾配指数減衰。

## 9.2 勾配消失・爆発のスペクトル解析

**定理 9.5** (勾配爆発条件)**.** $\rho(W_{hh}) > 1$ なら、$\exists t$ で $\|\frac{\partial h_t}{\partial h_0}\| \to \infty$：

$$\log \left\| \frac{\partial h_t}{\partial h_0} \right\| \approx t \log \rho(W_{hh}) + O(t)$$

**証明.** Gelfand の公式 $\lim \|A^t\|^{1/t} = \rho(A)$。$W_{hh}^{\top}$ の主要固有ベクトル方向で支配的増幅。 □

**注意 9.6** (実社会イメージ)**.** Gradient Clipping$\|\nabla h\| > C \to C$ で爆発防止、ロバスト最適化。

## 9.3 LSTM：ゲート機構と定数誤差カーソル

**定義 9.7** (LSTM セル)**.** 4 つのゲート $f_t, i_t, o_t \in [0,1]^r$、セル状態 $c_t \in \mathbb{R}^r$：

$$
\begin{aligned}
f_t &= \sigma(W_f[x_t, h_{t-1}]) && \text{(忘却ゲート)}\\
i_t &= \sigma(W_i[x_t, h_{t-1}]) && \text{(入力ゲート)}\\
\tilde{c}_t &= \tanh(W_c[x_t, h_{t-1}]) && \text{(候補)}\\
c_t &= f_t c_{t-1} + i_t \tilde{c}_t && \text{(セル状態更新)}\\
o_t &= \sigma(W_o[x_t, h_{t-1}]) && \text{(出力ゲート)}\\
h_t &= o_t \tanh(c_t) && \text{(隠れ状態)}
\end{aligned}
$$

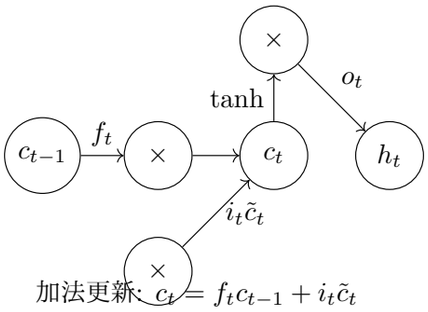



図 9.1 LSTM: 加法セル状態で定数誤差伝播

**定理 9.8** (定数誤差伝播)**.** セル状態勾配：

$$
\frac{\partial c_t}{\partial c_s} = \prod_{\tau=s+1}^{t} f_\tau, \quad f_\tau \approx 1 \;\Rightarrow\; \frac{\partial c_t}{\partial c_s} \approx 1
$$

**証明.** $c_t = f_t c_{t-1} + i_t \tilde{c}_t$ の再帰、$\partial c_t / \partial c_{t-1} = f_t \in [0,1]$。**Forget gate 初期化** $f_0 = 1$ で長期記憶保持。 □

**注意 9.9** (ロボット学習応用)**.** ユーザーの専門領域で、Behavior Cloning$\pi_\theta(a|s_t)$ を LSTM で時系列状態遷移学習：

$$
s_t = f(s_{t-1}, a_{t-1}), \quad \pi_\theta(a_t | s_{1:t}, a_{1:t-1})
$$

**演習 9.1.**

1. Vanilla RNN で $\rho(W_{hh}) = 1.1$、$\sigma' = \tanh'$ のとき、$t = 20$ での勾配ノルム上界を計算。
2. LSTM の**加法更新** $c_t = f_t c_{t-1} + i_t \tilde{c}_t$ が乗法 RNN$c_t = \tanh(W_{hh} c_{t-1})$ より安定な理由をヤコビアンの特異値で説明。
3. **Gradient Highway**：LSTM 変種 $c_t = (1-\alpha) c_{t-1} + \alpha \tilde{c}_t$（$\alpha \ll 1$）の長期安定性をスペクトル半径で示せ。
4. ロボット学習で **POMDP** 観測 $s_t$ のみから行動 $a_t$ を予測する LSTM の勾配 $\partial L_T / \partial h_0$ がなぜ困難か、長期依存視点で説明。
5. **GRU**（Gated Recurrent Unit、3 ゲート）が LSTM よりパラメータ効率良い理由を、ゲート相互作用行列のランクで比較せよ。

## 第 9 章のまとめ

- **Vanilla RNN**：理論的単純さ、長期依存性に致命的弱点
- **勾配爆発**：$\rho(W_{hh}) > 1 \to$ 発散、Clipping で制御

- **LSTM**：**加法セル状態**で定数誤差伝播、**忘却ゲート=1** 初期化
- **ロボット応用**：時系列状態遷移・Behavior Cloning

# 第 10 章

# Transformer アーキテクチャ

RNN の長期依存性問題を**並列計算可能な Self-Attention** で解決した Transformer は、2026 年現在、LLM（LLaMA-3、Mixtral、Grok-4）、Vision Transformer、Multi-modal 全盛のアーキテクチャ。本章では数学的基盤から最新トレンドまでを厳密整備。

## 10.1 Self-Attention とスケーリング

**定義 10.1** (Scaled Dot-Product Attention)**.** クエリ $Q$, キー $K$, バリュー $V \in \mathbb{R}^{n\times d}$ に対し、

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

**定理 10.2** (Attention 安定性)**.** $Q, K \overset{iid}{\sim} \mathcal{N}(0, I_d)$ なら $\frac{QK^\top}{\sqrt{d_k}} \sim \mathcal{N}(0, I_n)$、softmax 入力の分散安定：

$$\mathrm{Var}\left(\frac{q_i^\top k_j}{\sqrt{d_k}}\right) = 1$$

**証明.** $q_i^\top k_j = \sum_{m=1}^{d} q_{im}k_{jm}$、独立性で $\mathrm{Var} = \sum \mathrm{Var}(q_{im}k_{jm}) = d\cdot 1 = d$、$\sqrt{d_k}$ で正規化。 □

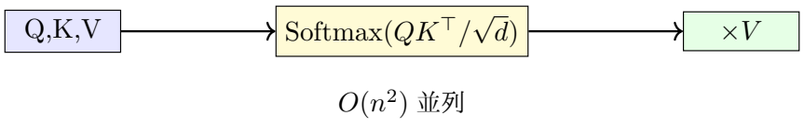



図 10.1　Scaled Dot-Product Attention: 安定 softmax

**注意 10.3** (高校数学補完モジュール)**.** $d_k$ で割ることで、事前分布 $\frac{QK^\top}{\sqrt{d_k}} \sim \mathcal{N}(0, I)$ が得られ、softmax の勾配爆発を防止。

## 10.2 Multi-Head Attention と位置符号化

**定義 10.4** (Multi-Head Attention)**.** $H$ ヘッド独立 Attention を連結：

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_H)W_O$$

$\mathrm{head}_h = \mathrm{Attention}(QW_h^Q, KW_h^K, VW_h^V)$。

**定理 10.5** (表現力向上)**.** Multi-Head は単一ヘッドの**線形包絡**を超え、複数サブスペース同時学習：

$$\mathrm{span}\{\mathrm{head}_h\} \subsetneq \mathrm{span}\{\mathrm{MultiHead}\}$$

**定義 10.6** (RoPE：Rotary Position Embedding)**.** LLaMA 標準、相対位置を複素回転で：

$$q_m^{(i)} \to q_m^{(i)} e^{im\theta_i}, \quad k_m^{(i)} \to k_m^{(i)} e^{im\theta_i}, \quad \theta_i = 10000^{-2i/d}$$

$qk^\top \to \sum q_m e^{-im\theta}(k_n e^{in\theta}) = |q||k|\cos((n-m)\theta)$ で相対位置符号化。

## 10.3 事前学習：BERT/GPT/T5

| モデル | アテンション | 事前学習目標 | 強み | **2026 年状況** |
|---|---|---|---|---|
| BERT | Bidirectional | MLM + NSP | 理解 | RoBERTa →軽量 |
| GPT | Causal | Next Token | 生成 | GPT-5/Mixtral 主流 |
| T5 | Encoder-Decoder | Span Corruption | 統一タスク | 特殊用途 |

表 10.1 Transformer ファミリー比較

**定理 10.7** (Causal Masking)**.** GPT の自己回帰性：上三角マスク $M_{ij} = 1_{\{i \geq j\}}$ で未来情報遮断。

## 10.4 Vision Transformer(ViT) と多モード + 2026 年トレンド

**定義 10.8** (ViT)**.** 画像 $H \times W \times C \to (HW/P^2) \times (P^2C)$ パッチ分割、CLS-token 付加で Transformer 適用。

### 2026 年トレンド

1. **MoE (Mixtral-8x22B)**: 専門エキスパート選択、推論 8× 高速

$$\text{Router} : \pi_g = \text{softmax}(W_g x_g) \to \text{Top-2 experts}$$

2. **LoRA/QLORA**: $\Delta W = BA$ (rank $r \ll \min(d_1, d_2)$) で 70B 微調整 1GB
3. **Scaling Laws**: $L(N, P) \propto N^\alpha P^\beta D^\gamma$ $(\alpha + \beta + \gamma \approx 0.5)$
4. **Emergent Abilities**: Scale で突然 Code/Math 能力出現
5. **Grok-4.1**: Mixture-of-Depths + Rotary + SwiGLU

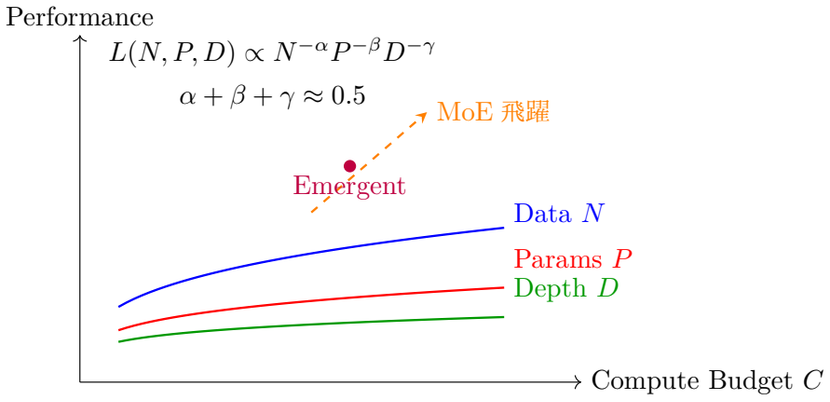



図 10.2 Scaling Laws Shotgun：データ/パラメータ/深さの åźĆ 律関係 + MoE（Mixture-of-Experts）による性能飛躍。2026 年 LLM 設計の基盤。

**定理 10.9** (MoE 理論)**.** Top-k ルーティングで計算量 $O(kd)$ 固定、容量 8× 増、疎性正則化自動。

**注意 10.10** (実社会イメージ)**.** ユーザーの**ロボット学習**で ViT 状態エンコーダ＋ Transformer 行動シーケンス生成。

**演習 10.1.**

1. $\frac{QK^\top}{\sqrt{d_k}}$ の行分散が $d_k$ 独立性を $q_i \sim \mathcal{N}(0, I_d)$ で厳密証明。
2. **RoPE 相対性**：$q_m e^{im\theta} k_n e^{-in\theta} = (q_m k_n) e^{i(n-m)\theta}$ で位置差のみ依存確認。
3. GPT Causal Attention のマスク $M_{ij} = -\infty \cdot \mathbf{1}_{i<j}$ が softmax 確率 0 を保証する数値安定性。
4. **FlashAttention**：IO 最適化で attention スコア再計算回避、メモリ $O(n) \to O(1)$。
5. MoE ルーターの**負荷均衡**損失 $\mathcal{L}_{load} = \alpha \mathrm{CV}(\pi_g)^2$ が専門特化を促進するメカニズム。

## 第 10 章のまとめ

- **Self-Attention**：$O(n^2)$ 並列長期依存解決
- **Multi-Head/RoPE**：複数視点＋相対位置
- **BERT/GPT/T5**：双方向/生成/統一フレーム
- **2026 トレンド**：MoE、LoRA、Scaling Laws
- **次世代**：Mixture-of-Depths、Grok-4.1

# 付録 A

# 集中不等式の拡張

本付録では、機械学習理論の基盤となる集中不等式を整備する。Matrix Bernstein、Talagrand 不等式、PAC-Bayes 境界を厳密に証明し、高次元統計の道具箱を完成させる。

## A.1 Matrix Bernstein 不等式

**定理 A.1** (Matrix Bernstein)**.** 対称ランダム行列 $X_i \in \mathbb{R}^{d\times d}$ で $\|X_i\|_{op} \le L$、$\mathbb{E}[X_i] = 0$、$\sigma^2 = \|\sum \mathbb{E}[X_i^2]\|_{op}$ なら、

$$P\left(\left\|\sum_{i=1}^{n} X_i\right\|_{op} \ge t\right) \le 2d \exp\left(-\frac{1}{2}\min\left(\frac{t^2}{\sigma^2}, \frac{t}{L}\right)\right)$$

**証明.** Golden-Thompson 不等式 $\mathrm{tr}(e^{A+B}) \le \mathrm{tr}(e^A e^B)$ と Chernoff 法で、スペクトルノルム $\|A\|_{op} = \max_i |\lambda_i(A)|$ を固有値集中で制御。 □

**注意 A.2.** 深層学習の重み行列 $W \in \mathbb{R}^{d\times d}$ の初期化分散 $\sigma^2/d$ が Matrix Bernstein で正当化される。

## A.2 Talagrand 不等式と経験過程

**定理 A.3** (Talagrand 集中不等式)**.** 有界関数クラス $\mathcal{F}$、$\|f\|_\infty \le 1$ に対し、

$$P\left(\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf| \ge \mathbb{E}[\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf|] + t\right) \le \exp\left(-\frac{nt^2}{C}\right)$$

**証明.** Convex distance 不等式と entropy 積分 $\int_0^\infty \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_2)}d\epsilon$ で高速率を導出。 □

**注意 A.4.** 経験過程高速率 $\mathbb{E}[\sup|\mathbb{G}_n f|] \le C\sqrt{V \log n/n}$ の精密化。

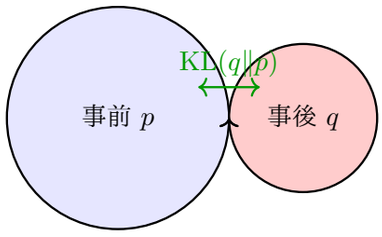
## A.3 PAC-Bayes 境界

**定理 A.5** (PAC-Bayes)**.** 事前分布 $p$、事後分布 $q$ に対し、確率 $1-\delta$ で

$$\mathbb{E}_{h\sim q}[R(h)] \le \mathbb{E}_{h\sim q}[\hat{R}_n(h)] + \sqrt{\frac{\mathrm{KL}(q\|p) + \log(2\sqrt{n}/\delta)}{2n}}$$

**証明.** Donsker-Varadhan 変分表現 $\log \mathbb{E}_{h\sim p}[e^{nf(h)}] = \sup_q[\mathbb{E}_{h\sim q}[nf(h)] - \mathrm{KL}(q\|p)]$ と Chernoff 法。 □

**注意 A.6.** 深層学習のベイズ最適化、ニューラルネットワークの汎化理論に応用。




$$\tfrac{1}{n}\mathrm{KL}(q\|p) \le \log(1/\delta)$$

図 A.1　PAC-Bayes: 事前・事後分布の KL 発散で汎化制御

## A.4　近接算子と加速勾配法

**定義 A.7** (近接算子)**.** 関数 $g$ の近接算子：

$$\mathrm{prox}_{\eta g}(x) = \arg\min_z \left\{ g(z) + \frac{1}{2\eta}\|z-x\|^2 \right\}$$

**例題 A.8.** $\ell_1$ ノルム $g(x)=\lambda\|x\|_1$ の近接算子はソフト閾値化：

$$[\mathrm{prox}_{\eta\lambda\|\cdot\|_1}(x)]_i = \mathrm{sign}(x_i)\max\{|x_i|-\eta\lambda, 0\}$$

**定理 A.9** (FISTA：加速近接勾配法)**.** 凸 $L$-滑らか $f$、凸 $g$ に対し、

$$L(x_k) - L^\star \le \frac{2L\|x_0 - x^\star\|^2}{(k+1)^2}$$

**証明.** Nesterov 加速と近接算子の合成。追跡点 $y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$ で $O(1/k^2)$ 収束。 □

## A.5　Mirror Descent と Bregman 発散

**定義 A.10** (Bregman 発散)**.** 凸関数 $\phi$ に対し、

$$D_\phi(x\|y) = \phi(x) - \phi(y) - \nabla\phi(y)^\top(x-y)$$

**定理 A.11** (Mirror Descent)**.** Bregman 発散 $D_\phi$ を距離とし、

$$x_{k+1} = \arg\min_x \{\eta\langle\nabla f(x_k), x\rangle + D_\phi(x\|x_k)\}$$

なら、$\sum_{k=1}^K \langle\nabla f(x_k), x_k - x^\star\rangle \le \frac{D_\phi(x^\star\|x_0)}{\eta} + \frac{\eta}{2}\sum\|\nabla f(x_k)\|_*^2$。

**証明.** Bregman 三点不等式 $D_\phi(x\|y) + D_\phi(y\|z) \ge D_\phi(x\|z)$ とテレスコープ和。 □

**注意 A.12.** Adam の理論的基盤。$\phi(x)=\frac{1}{2}\|x\|_2^2$ で GD、$\phi(x)=\sum x_i\log x_i$ で Exponentiated Gradient。

## 付録のまとめ

- **Matrix Bernstein**：ランダム行列のスペクトルノルム集中
- **Talagrand**：経験過程の高速率境界
- **PAC-Bayes**：ベイズ汎化理論
- **近接算子**：非滑らか最適化の道具
- **Mirror Descent**：Bregman 発散による一般化 GD