統計的機械学習
演習問題解答集

詳細解説と採点基準

Yugo Nakayama

2026 年 2 月 10 日

目次

はじめに

本書は「統計的機械学習」教科書の演習問題に対する**詳細解答集**である。

特徴

- **完全解答**：すべての演習に対して採点可能なレベルの解答
- **詳細な証明**：定理の証明を丁寧に展開
- **数値例**：具体的な計算例で理解を深化
- **実装ノート**：PyTorch/Scikit-learn 実装への橋渡し
- **採点基準**：各解答の重要ポイントを明示

使い方

1. まず自力で演習問題に取り組む
2. 解答を確認し、自分の理解度をチェック
3. 証明の各ステップを追って理論的理解を深める
4. 実装ノートで実践的知識を獲得

構成

本書は教科書の章立てに対応：

- 第 1 章：教師あり学習の枠組み
- 第 2 章：線形分類器と最適化
- 第 3 章：カーネル法と RKHS
- 第 4 章：汎化理論
- 第 5 章：深層ニューラルネットワーク
- 第 6 章：確率的勾配法
- 第 7 章：スパース回帰
- 第 8 章：GLM 正則化
- 第 9 章：正規フローネットワーク

- 第 10 章：Transformer アーキテクチャ
- 付録 A：集中不等式

それでは、詳細な解答を見ていきましょう。

第 1 章

第 1 章：教師あり学習の枠組み

演習 1.1

(1) Huber 損失のロバスト性

Huber 損失は

$$\ell_\delta(t) = \frac{1}{2}t^2 \mathbf{1}_{|t|\le\delta} + \delta\left(|t| - \frac{\delta}{2}\right)\mathbf{1}_{|t|>\delta}$$

で定義される。ロバスト性は**勾配が大きな外れ値で飽和する**ことから示せる。

解答. まず各損失の勾配を計算する。

二乗損失の場合：

$$\ell(t) = \frac{1}{2}t^2 \Rightarrow \ell'(t) = t$$

よって $|t|$ が大きい外れ値ほど、勾配も線形に増加し、推定値が大きく引きずられる。

Huber 損失の場合：

$$\ell'_\delta(t) = \begin{cases} t & (|t| \le \delta) \\ \delta\,\mathrm{sign}(t) & (|t| > \delta) \end{cases}$$

したがって $|t| > \delta$ の外れ値に対しては、勾配の絶対値が常に δ に抑えられる。これにより：

1. 大きな外れ値の影響が**有界**（$|\ell'_\delta(t)| \le \delta$）
2. 推定値の変化も δ に比例する範囲に抑制
3. MSE と比較して外れ値の影響が二次的→一次的に軽減

これが**ロバスト性**の数学的根拠である。 □

注意 1.1 (実装上の注意)**.** Huber 損失は Scikit-learn では `HuberRegressor` として実装されており、δ は `epsilon` パラメータで制御する。一般に $\delta \approx 1.345\sigma$（$\sigma$ はノイズの標準偏差）が推奨される。

(2) VC 次元 3, n=100, δ =0.05 の汎化境界

解答. 定理 1.4 より、VC 次元 $V = 3$、サンプルサイズ $n = 100$、$\delta = 0.05$ とすると

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

係数 C, L を 1 と仮定して数値計算する：

Step 1: 対数値の計算

$$\log n = \log 100 \approx 4.605$$
$$\log(1/\delta) = \log 20 \approx 2.996$$

Step 2: 第 1 項

$$\sqrt{\frac{V \log n + \log(1/\delta)}{n}} = \sqrt{\frac{3 \times 4.605 + 2.996}{100}} = \sqrt{\frac{16.811}{100}} \approx 0.410$$

Step 3: 第 2 項

$$\sqrt{\frac{\log(1/\delta)}{n}} = \sqrt{\frac{2.996}{100}} \approx 0.173$$

Step 4: 合計

$$\text{誤差上乗せ} \approx 0.410 + 0.173 = 0.583$$

したがって、最良リスク $R^\star = 0.1$ とすると

$$R(\hat{f}_n) \lesssim 0.1 + 0.583 \approx \boxed{0.68}$$

□

注意 1.2 (境界の保守性)**.** この境界はかなり緩い（保守的）。実際の汎化誤差は通常これより遥かに小さい。これは VC 理論が**最悪ケース**を保証するためである。より精密な境界には Rademacher 複雑度や PAC-Bayes 境界を用いる。

(3) 過学習メカニズムと早期停止の正当化

解答. Bias-Variance 分解と集中不等式の観点から説明する。

(a) Bias-Variance トレードオフ

図 1.2 より、モデル複雑度が上がると

$$\text{Bias} \downarrow, \quad \text{Variance} \uparrow$$

訓練を進めると複雑度が増し、ある時点から汎化誤差が増加する。

(b) Hoeffding 不等式による解釈

固定の仮説 f に対し

$$P(|\hat{R}_n(f) - R(f)| \geq t) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

学習初期：

- 単純な仮説空間 $\mathcal{F}_{\text{simple}}$
- $\hat{R}_n(f) \approx R(f)$（経験誤差汎化誤差）

学習後期：

- 複雑な仮説空間 $\mathcal{F}_{\text{complex}}$
- 有効 VC 次元増加→一様収束の保証が弱まる
- $\hat{R}_n(f) \ll R(f)$（過学習）

(c) 早期停止の正当化
検証誤差最小時点で学習を止めることは：

1. **容量制御**：実効的な VC 次元を $V_{\text{eff}} < V_{\max}$ に制限
2. **正則化**：時間制約付き ERM とみなせる
3. **Hoeffding 保証**：制限された複雑度の範囲で一様収束が保証される

これは λ による明示的正則化

$$\min_{f \in \mathcal{F}} \hat{R}_n(f) + \lambda \|f\|^2$$

と本質的に同等である（implicit regularization）。 □

注意 1.3 (実装)**.** PyTorch では `EarlyStopping` コールバックで実装。検証誤差が `patience` エポック改善しなければ学習停止。これにより過学習を自動的に防止できる。

補遺：定理 1.4 の完全証明

本節では、教科書の定理 1.4（ERM の汎化保証）の**詳細証明**を展開する。

定理 1.4 (定理 1.4 再掲)**.** 仮説クラス $\mathcal{F}$ の VC 次元を $V < \infty$ とし、損失 ℓ が $[0,1]$ 上に抑えられた L-Lipschitz 連続関数とする。標本 $(X_i, Y_i)_{i=1}^n \overset{iid}{\sim} P$ に対する ERM

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f), \qquad \hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i)$$

を考える。このとき、ある定数 $C > 0$ が存在して、任意の $\delta \in (0,1)$ に対し確率少なくとも $1-\delta$ で

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

が成立する。ここで $R(f) = \mathbb{E}[\ell(f(X), Y)]$ は真のリスクである。

証明. 証明を 6 つのステップに分けて展開する。

Step 1: 分解 - ERM の定義からの基本不等式

任意の $f^\star \in \mathcal{F}$ に対し

$$R(\hat{f}_n) - R(f^\star) = \underbrace{R(\hat{f}_n) - \hat{R}_n(\hat{f}_n)}_{(A)} + \underbrace{\hat{R}_n(\hat{f}_n) - \hat{R}_n(f^\star)}_{\leq 0} + \underbrace{\hat{R}_n(f^\star) - R(f^\star)}_{(B)}$$

第 2 項は ERM の定義より $\hat{R}_n(\hat{f}_n) \leq \hat{R}_n(f^\star)$ なので非正。したがって

$$R(\hat{f}_n) - R(f^\star) \leq (A) + (B)$$

さらに

$$(A) = R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) \leq \sup_{f \in \mathcal{F}} (R(f) - \hat{R}_n(f))$$
$$(B) = \hat{R}_n(f^\star) - R(f^\star) \leq \sup_{f \in \mathcal{F}} (\hat{R}_n(f) - R(f))$$

よって

$$R(\hat{f}_n) - R(f^\star) \leq 2 \sup_{f \in \mathcal{F}} |R(f) - \hat{R}_n(f)|$$

Step 2: 固定関数に対する Hoeffding 不等式

単一の $f \in \mathcal{F}$ を固定する。$\ell(f(X_i), Y_i)$ は独立同分布で $[0, 1]$ に抑えられているので、Hoeffding 不等式より

$$\mathbb{P}\left(|\hat{R}_n(f) - R(f)| > t \right) \leq 2 \exp(-2nt^2), \qquad t > 0$$

Step 3: ε-net による有限化と union bound

VC 次元が有限 V であることから、Sauer-Shelah 補題より損失関数クラス $\mathcal{G} := \{\ell(f(\cdot), \cdot) : f \in \mathcal{F}\}$ の成長関数は

$$|\mathcal{G}|_{(X_1, \ldots, X_n)} \leq \left(\frac{en}{V} \right)^V$$

任意の $\varepsilon > 0$ について、標本上の ε-net の要素数 $|\mathcal{N}(\varepsilon)|$ が

$$|\mathcal{N}(\varepsilon)| \leq \left(\frac{An}{\varepsilon} \right)^V$$

となる（定数 $A > 0$）。

したがって、有限集合 $\mathcal{N}(\varepsilon)$ に対して union bound と Step 2 を適用すると

$$\mathbb{P}\left(\sup_{g \in \mathcal{N}(\varepsilon)} |\hat{R}_n(g) - R(g)| > t \right) \leq |\mathcal{N}(\varepsilon)| \cdot 2e^{-2nt^2}$$

Step 4: Net から元のクラスへの持ち上げ（Lipschitz 使用）

任意の $f \in \mathcal{F}$ に対し、$\mathcal{N}(\varepsilon)$ の中から標本上で近い g を 1 つ取る：

$$\max_{1\le i\le n} |\ell(f(X_i), Y_i) - \ell(g(X_i), Y_i)| \le L\varepsilon$$

すると

$$|\hat{R}_n(f) - \hat{R}_n(g)| \le L\varepsilon, \qquad |R(f) - R(g)| \le L\varepsilon$$

よって

$$|\hat{R}_n(f) - R(f)| \le |\hat{R}_n(g) - R(g)| + 2L\varepsilon$$

これを全ての $f \in \mathcal{F}$ に対してとると

$$\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| \le \sup_{g\in\mathcal{N}(\varepsilon)} |\hat{R}_n(g) - R(g)| + 2L\varepsilon$$

Step 5: 確率境界の具体化

Step 3, 4 を合わせ、$|\mathcal{N}(\varepsilon)|$ を代入すると

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)| > t + 2L\varepsilon\right) \le 2\left(\frac{An}{\varepsilon}\right)^V e^{-2nt^2}$$

ここで

- $\varepsilon = \sqrt{\frac{\log(1/\delta)}{n}}$
- $t = C_1\sqrt{\frac{V\log n + \log(1/\delta)}{n}}$

と置くと、右辺は $\le \delta$ となる。よって、適当な定数 C をとれば確率少なくとも $1-\delta$ で

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

Step 6: Step 1 との結合で定理を得る

任意の $f^\star \in \mathcal{F}$ に対し、Step 1 より

$$R(\hat{f}_n) \le R(f^\star) + 2\sup_{f\in\mathcal{F}} |\hat{R}_n(f) - R(f)|$$

右辺に Step 5 を代入し、$\inf_{f^\star\in\mathcal{F}}$ をとれば

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

が確率少なくとも $1-\delta$ で成立する。 □

注意 1.5 (証明の鍵となるアイデア)**.**

1. **ERM 分解**：訓練誤差と汎化誤差の差を経験過程の最大値で制御
2. **VC 次元**：無限クラスを有限 ε-net で近似
3. **Lipschitz 性**：近似誤差を $L\varepsilon$ で定量化
4. **Hoeffding**：各 net 要素の集中を保証
5. **Union bound**：有限要素への確率結合

第 2 章 演習解答

演習 2.1

(1) パーセプトロン更新が凸結合

定理 1.6 (パーセプトロンの凸結合表現)**.** 初期値 $w_0 = 0$ とし、パーセプトロン更新により得られる w_T は訓練サンプルの凸結合である。

証明. 誤分類されたサンプル (x_{i_t}, y_{i_t}) に対し

$$w_{t+1} = w_t + \eta y_{i_t} x_{i_t}$$

で更新する。$w_0 = 0$ から始めると

$$w_T = \sum_{t=0}^{T-1} \eta y_{i_t} x_{i_t}$$

各サンプル i が更新に使われた回数を n_i とすると

$$w_T = \sum_{i=1}^{n} (\eta n_i) y_i x_i = \sum_{i=1}^{n} \lambda_i y_i x_i$$

ここで $\lambda_i := \eta n_i \geq 0$。

正規化係数 $Z = \sum_j \lambda_j$ として $\tilde{\lambda}_i = \lambda_i / Z$ とおくと

$$w_T = Z \sum_i \tilde{\lambda}_i y_i x_i$$

係数 $\tilde{\lambda}_i$ は

$$\tilde{\lambda}_i \geq 0, \quad \sum_i \tilde{\lambda}_i = 1$$

を満たすので、w_T は $\{y_i x_i\}_{i=1}^n$ の**凸結合**である。 □

注意 1.7 (幾何的解釈)**.** 凸結合表現は、パーセプトロンが訓練データの「重み付き平均」として解を構成することを意味する。誤分類回数が多いサンプルほど重みが大きい。

(2) ロジスティック損失の上界

定理 1.8 (ロジスティック損失と Hinge 損失)**.** 二値ラベル $y \in \{\pm 1\}$、マージン $u = yf(x)$ に対し

$$\log(1 + e^{-u}) \le \max(0, -u) + \log 2$$

証明. 場合分けして証明する。

Case 1: $u \ge 0$ の場合

このとき $e^{-u} \le 1$ なので

$$\ell(u) = \log(1 + e^{-u}) \le \log(1 + 1) = \log 2$$

また $\max(0, -u) = 0$ なので

$$\ell(u) \le 0 + \log 2$$

Case 2: $u < 0$ の場合

$e^{-u} = e^{|u|}$ より

$$\begin{aligned} \ell(u) &= \log(1 + e^{|u|}) \\ &= \log\left(e^{|u|}(e^{-|u|} + 1)\right) \\ &= |u| + \log(1 + e^{-|u|}) \\ &\le |u| + \log 2 \end{aligned}$$

このとき $\max(0, -u) = -u = |u|$ なので

$$\ell(u) \le -u + \log 2$$

両ケースより

$$\log(1 + e^{-yf(x)}) \le \max(0, -yf(x)) + \log 2$$

□

注意 1.9 (解釈)**.** この不等式は、ロジスティック損失が Hinge 損失の**滑らかな上界**であることを示す。したがってロジスティック回帰で訓練すれば、SVM の Hinge 損失も間接的に最小化される。

(3) 加速勾配法が $O(1/\sqrt{K})$ 改善

定理 1.10 (Nesterov 加速の収束率)**.** 凸 L-滑らか関数に対し

- 通常の GD：$L(w_K) - L^\star = O(1/K)$
- Nesterov 加速：$L(x_K) - L^\star = O(1/K^2)$

解答. **(a) 通常の GD**

定理 2.6 より、学習率 $\eta = 1/L$ で

$$L(w_K) - L^\star \le \frac{L\|w_0 - w^\star\|^2}{2K}$$

(b) Nesterov 加速

補助変数 y_k を用いた更新

$$x_{k+1} = y_k - \eta \nabla L(y_k), \quad y_{k+1} = x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k)$$

により、定理 2.8 のアナログとして

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

(c) 誤差レートの比較

同じ目標誤差 ε を達成するために必要な反復回数：

- GD：$K = O(1/\varepsilon)$
- Nesterov：$K = O(1/\sqrt{\varepsilon})$

したがって、誤差レートの観点で $\sqrt{K}$ 倍の改善が得られる。 □

注意 1.11 (実装)**.** PyTorch では `torch.optim.SGD` で `momentum` と `nesterov=True` を設定することで実装可能。深層学習では通常 `momentum=0.9` を使用。

(4) MSE とクロスエントロピーの勾配が $y - \hat{y}$

定理 1.12 (出力層勾配の一致)**.** シグモイド出力 $\hat{y} = \sigma(z)$ に対し、MSE とクロスエントロピーの z に関する勾配は本質的に同じ形になる。

証明. (a) MSE

損失：$\ell_{\mathrm{MSE}} = \frac{1}{2}(y - \hat{y})^2$

勾配：

$$\begin{aligned}\frac{\partial \ell_{\mathrm{MSE}}}{\partial z} &= \frac{\partial \ell}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \\ &= (\hat{y} - y) \cdot \sigma'(z) \\ &= (\hat{y} - y) \cdot \hat{y}(1 - \hat{y})\end{aligned}$$

(b) クロスエントロピー

損失：$\ell_{\mathrm{CE}} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$

勾配：

$$\begin{aligned}\frac{\partial \ell_{\mathrm{CE}}}{\partial z} &= -y\frac{1}{\hat{y}}\sigma'(z) - (1-y)\frac{1}{1-\hat{y}}(-\sigma'(z)) \\ &= -y\frac{\hat{y}(1-\hat{y})}{\hat{y}} + (1-y)\frac{\hat{y}(1-\hat{y})}{1-\hat{y}} \\ &= -y(1-\hat{y}) + (1-y)\hat{y} \\ &= \hat{y} - y\end{aligned}$$

(c) 比較

出力層の誤差信号として：

- MSE：$(\hat{y}-y)\cdot\hat{y}(1-\hat{y})$（項 $\hat{y}(1-\hat{y})$ で減衰）
- CE：$\hat{y}-y$（直接的）

□

注意 1.13 (実装上の利点)**.** クロスエントロピーは勾配消失を起こしにくい。$\hat{y}\approx 0$ や $\hat{y}\approx 1$ でも MSE のように勾配が消えない。実装では数値安定性のため `BCEWithLogitsLoss`（シグモイド +CE を結合）を使う。

第 3 章 演習解答

演習 3.1

(1) 線形カーネルの正定値性

定理 1.14 (線形カーネルは正定値核)**.** $K(x, x') = x^\top x'$ は正定値核である。

証明. 任意の $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$、$\{c_i\}_{i=1}^n \subset \mathbb{R}$ に対し

$$\begin{aligned}\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) &= \sum_{i,j=1}^{n} c_i c_j x_i^\top x_j \\ &= \sum_{i=1}^{n} c_i x_i^\top \sum_{j=1}^{n} c_j x_j \\ &= \left(\sum_{i=1}^{n} c_i x_i\right)^\top \left(\sum_{j=1}^{n} c_j x_j\right) \\ &= \left\|\sum_{i=1}^{n} c_i x_i\right\|^2 \\ &\geq 0\end{aligned}$$

したがって正定値核の定義を満たす。 □

注意 1.15 (固有値分解)**.** カーネル行列 $K_{ij} = x_i^\top x_j$ は $K = XX^\top$（$X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$）と書けるので、固有値はすべて非負である。

(2) RBF 核の特徴写像と距離

定理 1.16 (RBF 核の距離保存性)**.** RBF 核 $K(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$ に対し、対応する特徴写像 $\phi(x)$ は

$$\|\phi(x) - \phi(x')\|^2 = 2 - 2K(x, x')$$

を満たす。

証明. RKHS の再現性より

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

また、RBF 核では $K(x, x) = 1$（自己内積が 1）なので

$$\begin{aligned}
\|\phi(x) - \phi(x')\|^2 &= \langle \phi(x) - \phi(x'), \phi(x) - \phi(x') \rangle \\
&= \langle \phi(x), \phi(x) \rangle + \langle \phi(x'), \phi(x') \rangle - 2\langle \phi(x), \phi(x') \rangle \\
&= K(x, x) + K(x', x') - 2K(x, x') \\
&= 1 + 1 - 2K(x, x') \\
&= 2 - 2K(x, x')
\end{aligned}$$

□

注意 1.17 (明示的特徴写像)**.** RBF 核の明示的な特徴写像は無限次元だが、Mercer 展開

$$K(x, x') = \sum_{k=1}^{\infty} \lambda_k \psi_k(x) \psi_k(x')$$

を用いて $\phi(x) = (\sqrt{\lambda_k}\psi_k(x))_{k \geq 1}$ と構成できる（理論的）。実装では内積 $K(x, x')$ のみ計算。

(3) 代表定理と n 次元性

定理 1.18 (代表定理の証明補完)**.** 最適解 $\hat{f}$ は $\mathrm{span}\{K(x_i, \cdot)\}_{i=1}^n$ に属する。

証明. 定理 3.6 の最適性条件より

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

右辺の第 1 項は明らかに

$$\sum_{i=1}^{n} \ell'(f(x_i), y_i) K(x_i, \cdot) \in \mathrm{span}\{K(x_i, \cdot)\}_{i=1}^n$$

両辺が等しいので

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

よって $f \in \mathrm{span}\{K(x_i, \cdot)\}_{i=1}^n$。

$\mathcal{H}_K$ はこの span の閉包として定義されるが、有限次元なので閉包は自身と一致。したがって最適解を

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

□

注意 1.19 (計算量削減)**.** 無限次元空間 $\mathcal{H}_K$ での最適化が、n 次元の係数ベクトル $\alpha \in \mathbb{R}^n$ の最適化に帰着される。これがカーネル法の強力な点。

(4) カーネル行列が特異なときの正則化

定理 1.20 (正則化による条件数改善)**.** カーネルリッジ回帰 $\alpha = (\frac{1}{n\lambda}K + I)^{-1}y$ において、正則化 $\lambda > 0$ は条件数を改善する。

証明. カーネル行列の固有値分解 $K = U\Lambda U^\top$ により

$$\frac{1}{n\lambda}K + I = U\left(\frac{1}{n\lambda}\Lambda + I\right)U^\top$$

固有値は

$$\mu_i = \frac{\lambda_i(K)}{n\lambda} + 1$$

Case 1: K **が特異**（$\lambda_{\min}(K) \approx 0$）
正則化なし：条件数 $\kappa(K) = \lambda_{\max}/\lambda_{\min} \to \infty$
正則化あり：

$$\kappa\left(\frac{1}{n\lambda}K + I\right) = \frac{\mu_{\max}}{\mu_{\min}} = \frac{\frac{\lambda_{\max}}{n\lambda} + 1}{\frac{\lambda_{\min}}{n\lambda} + 1} \approx \frac{\lambda_{\max}}{n\lambda} + 1$$

$\lambda_{\min} \approx 0$ でも分母が 1 に持ち上がるため、条件数が有界になる。
Case 2: 数値例
$\lambda_{\max} = 100$、$\lambda_{\min} = 10^{-6}$、$n = 100$、$\lambda = 0.01$ のとき

$$\kappa\left(\frac{1}{n\lambda}K + I\right) \approx \frac{100/(100 \times 0.01) + 1}{1} = 101$$

元の条件数 10^8 から 101 へ大幅改善。 □

注意 1.21 (実装)**.** Scikit-learn の `KernelRidge` では自動的に正則化が適用される。数値安定性のため `alpha`（λ に対応）は必ず正の値を使用。

(5) 多項式核と XOR

定理 1.22 (XOR の非線形分離可能性)**.** XOR 問題は 2 次多項式核 $K(x, x') = (x^\top x' + 1)^2$ により線形分離可能になる。

証明. (a) XOR 問題の定義

入力 $(x_1, x_2) \in \{0,1\}^2$、ラベル y：

$$\begin{aligned}(0,0) &\to y = 0\\ (1,0) &\to y = 1\\ (0,1) &\to y = 1\\ (1,1) &\to y = 0\end{aligned}$$

1 次元特徴では線形分離不可能。

(b) 2 次多項式特徴写像

$\phi : \mathbb{R}^2 \to \mathbb{R}^5$ を

$$\phi(x) = (1, x_1, x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

と定義すると

$$K(x, x') = \langle \phi(x), \phi(x') \rangle = (x^\top x' + 1)^2$$

(c) 線形分離可能性

特徴空間での座標：

$$\begin{aligned}\phi(0,0) &= (1,0,0,0,0,0)\\ \phi(1,0) &= (1,1,0,1,0,0)\\ \phi(0,1) &= (1,0,1,0,1,0)\\ \phi(1,1) &= (1,1,1,1,1,\sqrt{2})\end{aligned}$$

分離超平面 $w = (0,0,0,1,1,-\sqrt{2})$ とすると

$$\begin{aligned}w^\top\phi(0,0) &= 0 \quad (\text{class } 0)\\ w^\top\phi(1,0) &= 1 \quad (\text{class } 1)\\ w^\top\phi(0,1) &= 1 \quad (\text{class } 1)\\ w^\top\phi(1,1) &= 2 - \sqrt{2} \approx 0.59 \quad (\text{class } 0)\end{aligned}$$

閾値 0.5 で完全分離可能。 □

注意 1.23 (実装)**.** Scikit-learn では `SVC(kernel='poly', degree=2)` で実装。XOR のような非線形問題でも高精度で分類可能。

第 4 章 演習解答

演習 4.1

(1) Bernstein 不等式の証明

定理 1.24 (Bernstein 不等式)**.** 有界かつ分散制約付きの独立同分布確率変数 $X_1, \dots, X_n$ に対し、

$$\mathrm{Var}(X_i) \le \sigma^2, \quad |X_i - \mu| \le M, \quad \mu = \mathbb{E}[X_i]$$

とする。このとき任意の $t > 0$ について

$$\mathbb{P}\left(|\bar{X}_n - \mu| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2(\sigma^2 + Mt/3)}\right)$$

が成り立つ。

証明. 証明を 3 つのステップに分けて展開する。

Step 1: Chernoff 法による片側境界

上側確率について示せば、下側は X_i を $-X_i$ とすることで同様に得られる。$S_n = \sum_{i=1}^n (X_i - \mu)$ とおき、任意の $\lambda > 0$ に対し

$$\begin{aligned}\mathbb{P}(S_n \ge nt) &= \mathbb{P}(e^{\lambda S_n} \ge e^{\lambda nt}) \\ &\le e^{-\lambda nt}\,\mathbb{E}[e^{\lambda S_n}] \\ &= e^{-\lambda nt}\prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i-\mu)}]\end{aligned}$$

最後の等号で独立性を使用。したがって各 i に対して moment generating function (mgf) を評価すればよい。

Step 2: 有界変数の mgf 評価

$Y = X_i - \mu$ と置く。仮定より $|Y| \le M$、$\mathbb{E}[Y] = 0$、$\mathbb{E}[Y^2] \le \sigma^2$。

テイラー展開により

$$\mathbb{E}[e^{\lambda Y}] = 1 + \lambda\mathbb{E}[Y] + \frac{\lambda^2}{2}\mathbb{E}[Y^2] + \sum_{k\ge 3}\frac{\lambda^k}{k!}\mathbb{E}[Y^k]$$

$\mathbb{E}[Y] = 0$ より一次項は消える。高次項に対して $|Y^k| \le M^{k-2}Y^2$ なので

$$\left|\mathbb{E}[Y^k]\right| \le \lambda^{k-2}\mathbb{E}[Y^2] \le \lambda^{k-2}\sigma^2$$

したがって

$$\begin{aligned}\mathbb{E}[e^{\lambda Y}] &\le 1 + \frac{\lambda^2}{2}\sigma^2 + \sum_{k\ge 3}\frac{|\lambda|^k}{k!}M^{k-2}\sigma^2 \\ &= 1 + \frac{\lambda^2\sigma^2}{2} + \sigma^2\lambda^2\sum_{k\ge 3}\frac{(|\lambda|M)^{k-2}}{k!}\end{aligned}$$

$|\lambda|M < 1$ の範囲で指数関数級数の評価により

$$\sum_{k\ge 3}\frac{(|\lambda|M)^{k-2}}{k!} \le \frac{|\lambda|M}{6(1-|\lambda|M)}$$

よって

$$\mathbb{E}[e^{\lambda Y}] \le 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^3 M\sigma^2}{6(1-|\lambda|M)}$$

$\log(1+u) \le u$ を使って

$$\log \mathbb{E}[e^{\lambda Y}] \le \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^3 M\sigma^2}{6(1-|\lambda|M)}$$

$|\lambda|M \le 1/2$ に限定すれば $1-|\lambda|M \ge 1/2$、従って

$$\log \mathbb{E}[e^{\lambda Y}] \le \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^3 M\sigma^2}{3} \le \frac{\lambda^2\sigma^2}{2(1-\lambda M/3)}$$

Step 3: Chernoff bound への代入と最適化

Step 1, 2 より $\lambda M \le 1/2$ の範囲で

$$\mathbb{P}(S_n \ge nt) \le \exp\left(-\lambda nt + n\frac{\lambda^2\sigma^2}{2(1-\lambda M/3)}\right)$$

右辺を最小にする λ として

$$\lambda = \frac{t}{\sigma^2 + Mt/3}$$

を選ぶ。これを代入すると

$$-\lambda nt + n\frac{\lambda^2\sigma^2}{2(1-\lambda M/3)} = -\frac{nt^2}{2(\sigma^2 + Mt/3)}$$

従って

$$\mathbb{P}(S_n \ge nt) \le \exp\left(-\frac{nt^2}{2(\sigma^2 + Mt/3)}\right)$$

同様に $S_n \le -nt$ についても同じ bound が得られるから

$$\mathbb{P}(|S_n| \ge nt) \le 2\exp\left(-\frac{nt^2}{2(\sigma^2 + Mt/3)}\right)$$

$\bar{X}_n - \mu = S_n/n$ より

$$\mathbb{P}(|\bar{X}_n - \mu| \ge t) \le 2\exp\left(-\frac{nt^2}{2(\sigma^2 + Mt/3)}\right)$$

□

注意 1.25 (Hoeffding と Bernstein の比較)**.**

- **Hoeffding**(定理 1.3):$P(|\bar{X}_n - \mu| \geq t) \leq 2\exp\left(-\dfrac{2nt^2}{(b-a)^2}\right)$
- **Bernstein**:分散 σ^2 が小さい場合、より鋭い評価
- $t \ll \sigma^2/M$ のとき:Bernstein は $\exp(-nt^2/2\sigma^2)$ に近い
- $t \gg \sigma^2/M$ のとき:両者とも同程度

実用上、分散が既知で小さい場合は Bernstein が有用。

注意 1.26 (実装への応用)**.** 機械学習では勾配の分散 $\mathrm{Var}[\nabla L]$ を推定し、Bernstein 型信頼区間でステップサイズを適応的に調整するアルゴリズムが存在(例:AdaGrad, RMSProp)。

(2) VC=3, n=100 の汎化誤差上界と実用性

定理 1.27 (Rademacher 複雑度による汎化境界)**.** 定理 4.5 より、VC 次元 V を持つクラスに対し

$$\mathrm{Rad}_n(\mathcal{F}) \leq C\sqrt{\frac{V\log(n/V)}{n}}$$

が成り立ち、確率 $1-\delta$ で

$$R(\hat{f}_n) \leq R^\star + \mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

解答. $V=3$、$n=100$、$R^\star = 0.1$、$\delta = 0.05$ とする。

Step 1: Rademacher 項

$$\mathrm{Rad}_n(\mathcal{F}) \leq C\sqrt{\frac{3\log(100/3)}{100}} \approx C\sqrt{\frac{3\times 3.5}{100}} \approx 0.32C$$

Step 2: 確率項

$$t = \sqrt{\frac{\log(1/\delta)}{2n}} = \sqrt{\frac{\log 20}{200}} \approx \sqrt{\frac{3.0}{200}} \approx 0.12$$

Step 3: 合計定数 $C \approx 1$ として

$$R(\hat{f}_n) \leq 0.1 + 0.32 + 0.12 \approx 0.54$$

□

注意 1.28 (実用性の限界)**.** 95% 信頼で $R(\hat{f}_n) \leq 0.54$ が保証されるに過ぎない。現実には CV 等で 0.15-0.2 程度の性能が期待できる状況でも、理論界はかなり緩い。これは:

- VC 理論が**最悪ケース**を保証するため
- データ分布や学習アルゴリズムの特性を考慮していないため
- 実用的判断には直接使えず、「安全側の上限」を与えるのみ

したがって、モデル選択には交差検証が不可欠。

(3) 半空間の VC 次元：d=2 で 4 点シャタリング可能、5 点不可能

定理 1.29 (半空間の VC 次元)**.** d 次元半空間クラス $\mathcal{F} = \{x \mapsto \mathrm{sign}(w^\top x + b)\}$ の VC 次元は $d+1$。

解答：$d = 2$ の場合. **(a) 4 点をシャタリング可能**

一般位置にある 4 点（任意 3 点が非共線）を正方形の頂点

$$(0,0), \quad (1,0), \quad (0,1), \quad (1,1)$$

とする。

任意の $2^4 = 16$ 通りのラベル割り当てに対し、それを実現する直線が存在：

- **1 点のみ正例**：その点を通り他 3 点を分離する直線
- **対角ラベル**（例：$(0,0)$ と $(1,1)$ が正、他が負）：対角線で分離
- **隣接ラベル**：適切な傾きの直線で分離

幾何的に、一般位置の 4 点は任意の 2 点を区別する直線が存在するため、全ての 2^4 ラベルが半空間で表現可能。

(b) 5 点は不可能

平面の直線クラスの VC 次元は $d+1 = 3$ であり、任意の 5 点集合をシャタリングできない。

反例：凸五角形の頂点上の 5 点に対し、「中心の 1 点のみ負例」というラベルを考える。

- 負例点が凸包内部なら、どの直線も外側の正例 4 点すべてを同時に分離できない
- 負例点が凸包頂点なら、その点を分離する直線は他の正例点も切り離してしまう

したがって $\mathrm{VC}(\mathcal{F}) = 3$（一般に $d+1$）。 □

(4) Rademacher 複雑度の $\sqrt{V \log n / n}$ スケーリング

定理 1.30 (Dudley 積分による Rademacher 評価)**.** VC クラスに対し

$$\mathrm{Rad}_n(\mathcal{F}) \le \frac{C}{\sqrt{n}} \int_0^{\mathrm{diam}(\mathcal{F})} \sqrt{\log N(\varepsilon, \mathcal{F})}\, d\varepsilon$$

ここで $N(\varepsilon, \mathcal{F})$ は ε-covering 数。

解答スケッチ. **Step 1: Covering 数の VC 評価**

VC クラスでは半径 ε の covering 数が

$$N(\varepsilon) \lesssim \left(\frac{A}{\varepsilon}\right)^V$$

したがって

$$\log N(\varepsilon) \lesssim V \log(1/\varepsilon)$$

Step 2: 積分の評価

下限を t として

$$\int_t^1 \sqrt{V \log(1/\varepsilon)}\, d\varepsilon \sim \sqrt{V} \int_t^1 \sqrt{\log(1/\varepsilon)}\, d\varepsilon$$

Step 3: 経験分解能

「経験分解能」に対応するスケール $t \asymp 1/n$ にとると $\log(1/t) \asymp \log n$。

covering 数の主貢献は $\varepsilon \approx 1/n$ 近傍から来るので

$$\mathrm{Rad}_n(\mathcal{F}) = O\left(\sqrt{\frac{V \log(1/t)}{n}}\right) = O\left(\sqrt{\frac{V \log n}{n}}\right)$$

□

(5) Interpolation regime で VC 理論が破綻する理由

定理 1.31 (二重降下と VC 理論のギャップ)**.** Interpolation regime（$p \gg n$ で訓練誤差 0）において、古典 VC 理論は過学習を予測するが、実際には汎化性能が改善する（二重降下）。

解答. **(a) 古典 VC 理論の予測**

VC 次元 V が大きいほど汎化誤差上界

$$\sqrt{\frac{V \log n}{n}}$$

が増大し、**過学習一辺倒**の挙動を予測。

$V \gg n$ では理論上汎化不能と予測される。

(b) 現代深層学習の実際

過剰パラメータ化モデル（$V \gg n$）が補間領域で再び汎化性能を改善（二重降下現象）:

- 訓練誤差：0 を達成
- テスト誤差：過学習ピーク後に再び減少

(c) ギャップの原因

最小ノルム補間解

$$\hat{w} = \arg\min\{\|w\|_2 : \ Xw = y\}$$

が鍵となる：

1. **暗黙的正則化**：GD は全補間解の中で特にノルムが小さい解を選択

2. **特別なサブセット**：VC 理論は「クラス全体の最悪の仮説」を考えるが、実際のアルゴリズムは特別なサブセット（最小ノルム解）を選択
3. **一様境界の限界**：クラス全体に対する一様な VC 境界では、この選択のバイアスを捉えられない

したがって、アルゴリズム依存の解析（NTK 理論、暗黙的正則化理論）が必要となる。 □

注意 1.32 (現代的アプローチ)**.** 補間領域の汎化を説明する理論：

- **NTK 理論**：カーネル極限での最小ノルム解析
- **暗黙的正則化**：GD の選択バイアス
- **良性過学習**：高次元での幾何的構造

これらは VC 理論を補完・拡張する枠組み。

第 7 章 演習解答

演習 7.1

Lasso Oracle 不等式（定理 7.3）の完全証明

定理 1.33 (定理 7.3 再掲 : Lasso Oracle 不等式)**.** モデル

$$y = X\beta^\star + \varepsilon, \quad \varepsilon_i \overset{iid}{\sim} \text{sub-Gaussian}(\sigma^2)$$

設計行列の各列を $\|X_{\cdot j}\|_2 = \sqrt{n}$ に標準化、真の支持集合 $S = \mathrm{supp}(\beta^\star)$、疎性 $s = |S|$ とする。

Lasso 推定量

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

Restricted Eigenvalue (RE) 条件

$$\kappa(S, 3) := \inf_{\Delta \neq 0:\ \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1} \frac{\|X\Delta\|_2/\sqrt{n}}{\|\Delta_S\|_2} > 0$$

が成り立ち、正則化パラメータが

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

を満たすとき、高確率で

$$\frac{1}{n} \|X(\hat{\beta} - \beta^\star)\|_2^2 \leq \frac{9s\lambda^2}{\kappa(S,3)^2} \asymp \sigma^2 \frac{s \log p}{n}$$

が成り立つ。

証明. 証明を 4 つのステップに分けて展開する。$\kappa = \kappa(S, 3)$ と略記する。

Step 1: 基本不等式

Lasso の最適性より

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

$y = X\beta^\star + \varepsilon$ を代入し、$\Delta = \hat{\beta} - \beta^\star$ とおくと

$$\frac{1}{2n} \|X\Delta - \varepsilon\|_2^2 + \lambda \|\beta^\star + \Delta\|_1 \leq \frac{1}{2n} \|\varepsilon\|_2^2 + \lambda \|\beta^\star\|_1$$

したがって最適性条件は

$$\frac{1}{2n}\|\varepsilon - X\Delta\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda\|\beta^\star\|_1$$

左辺を展開：

$$\begin{aligned}&\frac{1}{2n}\left(\|\varepsilon\|_2^2 + \|X\Delta\|_2^2 - 2\varepsilon^\top X\Delta\right) + \lambda\|\hat{\beta}\|_1 \\ &\leq \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda\|\beta^\star\|_1\end{aligned}$$

$\frac{1}{2n}\|\varepsilon\|_2^2$ を両辺から消去して整理：

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

これが**基本不等式**である。

Step 2: ノイズ項の制御と円錐条件

ノイズ項を評価：

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

最後の不等式で仮定 $\lambda \geq 2\|(1/n)X^\top\varepsilon\|_\infty$ を使用。

ℓ_1 ノルムの三角不等式より

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

（$\beta^\star_{S^c} = 0$ を使用）

基本不等式に代入：

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

左辺は非負なので

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

これが**円錐条件**である。

この条件を基本不等式に戻すと

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

Step 3: RE 条件の適用と ℓ_2 ノルムへの変換

RE 条件の定義より、円錐条件を満たす任意の Δ に対し

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

また $\|\Delta_S\|_1 \leq \sqrt{s}\|\Delta_S\|_2$（Cauchy-Schwarz）なので

$$\begin{aligned}\frac{1}{2n}\|X\Delta\|_2^2 &\leq \frac{3\lambda}{2}\|\Delta_S\|_1 \\ &\leq \frac{3\lambda}{2}\sqrt{s}\,\|\Delta_S\|_2 \\ &\leq \frac{3\lambda}{2}\sqrt{s}\cdot\frac{\|X\Delta\|_2}{\kappa\sqrt{n}}\end{aligned}$$

両辺を $\|X\Delta\|_2/2n$ で割ると

$$\frac{1}{n}\|X\Delta\|_2 \leq \frac{3\lambda\sqrt{s}}{\kappa}$$

両辺に $\|X\Delta\|_2$ を掛けて

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

これが主張したい Oracle 不等式である。

Step 4: λ の選び方（確率評価）

$\lambda \geq 2\|(1/n)X^\top\varepsilon\|_\infty$ が高確率で成り立つ λ のオーダーを与える。

各成分 $\frac{1}{n}X_{\cdot j}^\top\varepsilon = \frac{1}{n}\sum_{i=1}^n X_{ij}\varepsilon_i$ は、列正規化と sub-Gaussian 性から

$$\mathrm{Var}\left[\frac{1}{n}X_{\cdot j}^\top\varepsilon\right] = \frac{\sigma^2}{n}\sum_{i=1}^n X_{ij}^2/n = \frac{\sigma^2}{n}$$

Hoeffding/Bernstein 型不等式と union bound により

$$\mathbb{P}\left(\left\|\frac{1}{n}X^\top\varepsilon\right\|_\infty > C\sigma\sqrt{\frac{\log p}{n}}\right) \leq p^{-c}$$

（定数 $C, c > 0$）

したがって

$$\lambda = K\sigma\sqrt{\frac{\log p}{n}}$$

で K を十分大きく取れば、$\lambda \geq 2\|(1/n)X^\top\varepsilon\|_\infty$ が確率 $1 - p^{-c}$ で成立。

これを Step 3 の結果に代入：

$$\frac{1}{n}\|X(\hat{\beta} - \beta^\star)\|_2^2 \leq \frac{9sK^2\sigma^2\log p}{\kappa^2 n} = O\left(\sigma^2\frac{s\log p}{n}\right)$$

□

注意 1.34 (Oracle 不等式の意味)**.** この不等式は、Lasso が真の支持集合 S を**知らなくても**、真のパラメータ $\beta^\star$ の予測誤差と同程度の性能を（対数的な劣化のみで）達成することを保証する。これが「Oracle」と呼ばれる理由。

注意 1.35 (RE 条件の役割)**.** RE 条件 $\kappa(S,3)>0$ は、ℓ_1 円錐上で設計行列 X が「十分にランク落ちしていない」ことを保証する。これは RIP（Restricted Isometry Property）の緩和版と考えられる。

注意 1.36 (実装)**.** Scikit-learn の `LassoCV` では、交差検証により最適な λ を自動選択。理論的には $\lambda \propto \sigma\sqrt{\log p/n}$ だが、実際は検証誤差で調整される。

第 8 章 演習解答

定理 8.3：Logistic Lasso の RSC 条件と Oracle 不等式（論文風完全証明）

定理 1.37 (定理 8.3 再掲)**.** $y_i \in \{0,1\}$、$\eta_i = X_i^\top \beta^\star$、$P(y_i = 1 \mid X_i) = \sigma(\eta_i)$、負対数尤度

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

とし、Logistic Lasso を

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

と定める。真のパラメータ $\beta^\star$ の支持を $S = \mathrm{supp}(\beta^\star)$、$s = |S|$ と書く。

行列

$$H(\beta) = \nabla^2 L_n(\beta) = \frac{1}{n} X^\top W(\beta) X, \quad W(\beta) = \mathrm{diag}(\sigma_i(1-\sigma_i))$$

に対し、ある $\kappa > 0$ が存在して

$$\Delta^\top H(\beta^\star)\Delta \geq \kappa \|\Delta_S\|_2^2 \quad \text{whenever } \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$$

（RSC 条件）が成り立つとする。このとき、$\lambda \geq 2\|\nabla L_n(\beta^\star)\|_\infty$ ならば

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

が高確率で成立する（$C > 0$ は絶対定数）。

証明. $\Delta = \hat{\beta} - \beta^\star$ と置く。証明を 4 つのステップに分ける。

Step 1: 基本不等式

最適性より

$$L_n(\hat{\beta}) + \lambda\|\hat{\beta}\|_1 \leq L_n(\beta^\star) + \lambda\|\beta^\star\|_1$$

Logistic 尤度に対する Bregman 発散を

$$D(\hat{\beta}\|\beta^\star) = L_n(\hat{\beta}) - L_n(\beta^\star) - \nabla L_n(\beta^\star)^\top \Delta$$

と定めると、最適性条件は

$$D(\hat{\beta}\|\beta^\star) + \nabla L_n(\beta^\star)^\top \Delta + \lambda(\|\beta^\star\|_1 - \|\hat{\beta}\|_1) \leq 0$$

Step 2: スコア項の上界と円錐条件

(i) スコア項：

$$-\nabla L_n(\beta^\star)^\top \Delta \le \|\nabla L_n(\beta^\star)\|_\infty \|\Delta\|_1 \le \frac{\lambda}{2}\|\Delta\|_1$$

最後の不等式で仮定 $\lambda \ge 2\|\nabla L_n(\beta^\star)\|_\infty$ を使用。

(ii) ℓ_1 項：三角不等式から

$$\begin{aligned}\|\beta^\star\|_1 - \|\hat\beta\|_1 &= \|\beta^\star_S\|_1 - \|\beta^\star_S + \Delta_S\|_1 - \|\Delta_{S^c}\|_1 \\ &\le \|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\end{aligned}$$

(i), (ii) を結合：

$$\begin{aligned}D(\hat\beta\|\beta^\star) &\le \frac{\lambda}{2}\|\Delta\|_1 + \lambda(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1) \\ &= \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1\end{aligned}$$

左辺は非負なので

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

（円錐条件）

これを基本不等式に戻すと

$$D(\hat\beta\|\beta^\star) \le \frac{3\lambda}{2}\|\Delta_S\|_1$$

Step 3: RSC 条件による下界

Logistic 損失の Hessian は

$$H(\beta^\star) = \nabla^2 L_n(\beta^\star) = \frac{1}{n} X^\top W^\star X, \quad W^\star = \mathrm{diag}(\sigma_i(1-\sigma_i))$$

テイラー展開により

$$D(\hat\beta\|\beta^\star) = \frac{1}{2}\Delta^\top H(\tilde\beta)\Delta$$

を満たす $\tilde\beta$ が線分 $[\beta^\star, \hat\beta]$ 上に存在。RSC 条件がこの近傍でも成り立つと仮定すると

$$D(\hat\beta\|\beta^\star) \ge \frac{\kappa}{2}\|\Delta_S\|_2^2$$

Cauchy-Schwarz より $\|\Delta_S\|_1 \le \sqrt{s}\,\|\Delta_S\|_2$ なので

$$D(\hat\beta\|\beta^\star) \ge \frac{\kappa}{2s}\|\Delta_S\|_1^2$$

Step 4: 上界と下界の結合

Step 2, 3 より

$$\frac{\kappa}{2s}\|\Delta_S\|_1^2 \le \frac{3\lambda}{2}\|\Delta_S\|_1$$

$\|\Delta_S\|_1 \neq 0$ として

$$\|\Delta_S\|_1 \leq \frac{3s\lambda}{\kappa}$$

これを Step 2 に代入：

$$D(\hat{\beta}\|\beta^\star) \leq \frac{3\lambda}{2} \cdot \frac{3s\lambda}{\kappa} = \frac{9s\lambda^2}{2\kappa}$$

また円錐条件より

$$\|\Delta\|_1 \leq 4\|\Delta_S\|_1 \leq \frac{12s\lambda}{\kappa}$$

従って

$$\lambda(\|\hat{\beta}\|_1 - \|\beta^\star\|_1) \leq \lambda\|\Delta\|_1 \leq \frac{12s\lambda^2}{\kappa}$$

定数をまとめれば

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

□

定理 8.6：GLM Lasso 統一 Oracle 不等式（論文風完全証明）

定理 1.38 (定理 8.6 再掲)**.** 一般化線形モデル

$$p(y \mid \eta) \propto \exp\left\{\frac{y\eta - b(\eta)}{a(\phi)V(y)}\right\}$$

負対数尤度

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{b(X_i^\top\beta) - y_i X_i^\top\beta\right\}$$

GLM Lasso $\hat{\beta} = \arg\min_\beta\{L_n(\beta) + \lambda\|\beta\|_1\}$ に対し、RSC 条件

$$D(\beta^\star + \Delta\|\beta^\star) \geq \frac{\kappa}{2}\|\Delta_S\|_2^2 \quad \text{if } \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$$

が成り立つとする。$\lambda \geq 2\|\nabla L_n(\beta^\star)\|_\infty$ ならば

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

が高確率で成立。

証明. 証明は定理 8.3 とほぼ同じ構造。GLM の負対数尤度に対する Bregman 発散

$$D(\hat{\beta}\|\beta^\star) = L_n(\hat{\beta}) - L_n(\beta^\star) - \nabla L_n(\beta^\star)^\top(\hat{\beta} - \beta^\star)$$

(1) 基本不等式と円錐条件

Lasso 最適性から

$$L_n(\hat{\beta}) + \lambda\|\hat{\beta}\|_1 \le L_n(\beta^\star) + \lambda\|\beta^\star\|_1$$

$\Delta = \hat{\beta} - \beta^\star$ として

$$D(\hat{\beta}\|\beta^\star) \le -\nabla L_n(\beta^\star)^\top \Delta + \lambda(\|\beta^\star\|_1 - \|\hat{\beta}\|_1)$$

定理 8.3 と同様に

$$D(\hat{\beta}\|\beta^\star) \le \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1$$

よって円錐条件 $\|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1$ が従う。

(2) RSC 条件と最終評価

RSC 条件により

$$D(\hat{\beta}\|\beta^\star) \ge \frac{\kappa}{2}\|\Delta_S\|_2^2 \ge \frac{\kappa}{2s}\|\Delta_S\|_1^2$$

上界と下界から

$$\|\Delta_S\|_1 \le \frac{3s\lambda}{\kappa}$$

従って

$$D(\hat{\beta}\|\beta^\star) \le C_1\frac{s\lambda^2}{\kappa}, \quad \lambda\|\hat{\beta} - \beta^\star\|_1 \le C_2\frac{s\lambda^2}{\kappa}$$

勾配の sub-Gaussian 性と union bound により

$$\|\nabla L_n(\beta^\star)\|_\infty = O_p\left(\sigma\sqrt{\frac{\log p}{n}}\right)$$

よって $\lambda \asymp \sigma\sqrt{\log p/n}$ を選べば

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

□

注意 1.39 (統一理論の意義)**.** この枠組みは線形回帰、ロジスティック回帰、Poisson 回帰等を統一的に扱う。鍵は Bregman 発散 $D(\cdot\|\cdot)$ と RSC 条件の組み合わせ。

定理 1.40 (定理 8.6 再掲：GLM Lasso Oracle 不等式)**.** 指数型分布族 GLM で、真のパラメータ $\beta^\star$ が s-スパース、RSC 条件が成り立ち、正則化パラメータ

$$\lambda \ge 2\|\nabla L_n(\beta^\star)\|_\infty$$

を満たすとき、高確率で

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

が成り立つ。ここで $D(\cdot\|\cdot)$ は損失に対応する Bregman 発散。

証明. 証明を 4 つのステップに分けて展開する。

Step 1: 基本不等式

GLM の負対数尤度 $L_n(\beta)$ に対し

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

$\Delta = \hat{\beta} - \beta^\star$ とおくと、最適性から

$$L_n(\hat{\beta}) + \lambda\|\hat{\beta}\|_1 \leq L_n(\beta^\star) + \lambda\|\beta^\star\|_1$$

Bregman 発散

$$D(\hat{\beta}\|\beta^\star) = L_n(\hat{\beta}) - L_n(\beta^\star) - \nabla L_n(\beta^\star)^\top \Delta$$

を用いると

$$D(\hat{\beta}\|\beta^\star) \leq -\nabla L_n(\beta^\star)^\top \Delta + \lambda(\|\beta^\star\|_1 - \|\hat{\beta}\|_1)$$

Step 2: ノイズ勾配と円錐条件

スコア $\nabla L_n(\beta^\star)$ は平均 0 の sub-Gaussian ベクトルなので、適切な λ のもとで

$$\|\nabla L_n(\beta^\star)\|_\infty \leq \frac{\lambda}{2}$$

が高確率で成り立つ。よって

$$-\nabla L_n(\beta^\star)^\top \Delta \leq \frac{\lambda}{2}\|\Delta\|_1$$

また $\|\beta^\star\|_1 - \|\hat{\beta}\|_1 \leq \|\Delta_S\|_1 - \|\Delta_{S^c}\|_1$（$\beta^\star_{S^c} = 0$ を使用）だから

$$\begin{aligned} D(\hat{\beta}\|\beta^\star) &\leq \frac{\lambda}{2}\|\Delta\|_1 + \lambda(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1) \\ &= \frac{\lambda}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) + \lambda(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1) \\ &= \frac{3\lambda}{2}\|\Delta_S\|_1 - \frac{\lambda}{2}\|\Delta_{S^c}\|_1 \end{aligned}$$

左辺が非負より

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

（円錐条件）

この条件を基本不等式に戻すと

$$D(\hat{\beta}\|\beta^\star) \leq \frac{3\lambda}{2}\|\Delta_S\|_1$$

Step 3: RSC 条件で Bregman 発散を二乗ノルムで下界

GLM の負対数尤度は局所的に二次式で近似でき、RSC 条件として

$$D(\beta^\star + \Delta\|\beta^\star) \geq \frac{\kappa}{2}\|\Delta_S\|_2^2 \quad \text{for } \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$$

が仮定されている。

$\|\Delta_S\|_1 \le \sqrt{s}\|\Delta_S\|_2$（Cauchy-Schwarz）なので

$$D(\hat{\beta}\|\beta^\star) \ge \frac{\kappa}{2}\|\Delta_S\|_2^2 \ge \frac{\kappa}{2s}\|\Delta_S\|_1^2$$

Step 4: 上界と下界の結合

Step 2, 3 より

$$\frac{\kappa}{2s}\|\Delta_S\|_1^2 \le D(\hat{\beta}\|\beta^\star) \le \frac{3\lambda}{2}\|\Delta_S\|_1$$

$\|\Delta_S\|_1 > 0$ として両辺を $\|\Delta_S\|_1$ で割ると

$$\|\Delta_S\|_1 \le \frac{3s\lambda}{\kappa}$$

これを下界に戻せば

$$D(\hat{\beta}\|\beta^\star) \le \frac{\kappa}{2s}\left(\frac{3s\lambda}{\kappa}\right)^2 = \frac{9s\lambda^2}{2\kappa}$$

また円錐条件と三角不等式から $\|\Delta\|_1 \le 4\|\Delta_S\|_1$ なので

$$\lambda\|\Delta\|_1 \le \frac{12s\lambda^2}{\kappa}$$

定数を整理：

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

最後に $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$ を代入：

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

□

注意 1.41 (線形回帰との統一)**.** この証明は線形回帰版（第 7 章）と本質的に同じ構造：

1. 最適性 → 基本不等式
2. ノイズ制御 → 円錐条件
3. RE/RSC 条件 → ℓ_2 変換
4. 確率評価 → 最終 Oracle 不等式

Bregman 発散 $D(\cdot\|\cdot)$ が二乗損失 $\|\cdot\|^2$ の一般化。

注意 1.42 (RSC 条件の役割)**.** RSC は「局所的な強凸性」を保証：

$$D(\beta + \Delta\|\beta) \ge \frac{\kappa}{2}\|\Delta_S\|_2^2$$

これにより、円錐上での Bregman 発散が ℓ_2 ノルムで下から評価可能。線形回帰の RE と同様の役割。

第 9 章・第 10 章 補遺：論文レベル補題集

第 9 章：RNN スペクトル解析の厳密補題

補題 9.1：ヤコビアンの積表示

補題 1.43 (勾配逆伝播の明示的表現)**.** Vanilla RNN $h_t = \sigma(Wh_{t-1} + Ux_t + b)$ に対し、任意の $s < t$ で

$$\frac{\partial h_t}{\partial h_s} = \prod_{\tau=s+1}^{t} W^\top \operatorname{diag}(\sigma'(z_\tau))$$

ここで $z_\tau = Wh_{\tau-1} + Ux_\tau + b$。

証明. 連鎖律により

$$\cdots \tfrac{\partial h_{s+1}}{\partial h_s}$$

各項は

$$\frac{\partial h_\tau}{\partial h_{\tau-1}} = \operatorname{diag}(\sigma'(z_\tau)) \cdot W = W^\top \operatorname{diag}(\sigma'(z_\tau))$$

（転置の順序に注意）

よって積表示が得られる。 □

補題 9.2：Gelfand の公式と長期挙動

補題 1.44 (スペクトル半径と漸近挙動)**.** $A_\tau = W^\top \operatorname{diag}(\sigma'(z_\tau))$ とおき、各 τ で $\|A_\tau - A\| \le \varepsilon$ を満たす行列 A が存在するとする。このとき

$$\limsup_{t\to\infty} \left\| \frac{\partial h_t}{\partial h_s} \right\|^{1/(t-s)} \le \rho(A) + o(1)$$

概要. Gelfand の公式 $\lim_{k\to\infty}\|A^k\|^{1/k} = \rho(A)$ と、摂動項 $\|A_\tau - A\|$ の sub-multiplicative な評価を組み合わせる。詳細は標準的なスペクトル理論の教科書を参照。 □

補題 9.3：Sigmoid 活性化によるスペクトル半径の縮小

補題 1.45 (Sigmoid 微分の上界)**.** Sigmoid 活性化 $\sigma(z) = 1/(1+e^{-z})$ に対し

$$\sigma'(z) = \sigma(z)(1-\sigma(z)) \in [0, 1/4], \quad \max_z \sigma'(z) = 1/4$$

証明. $f(z) = \sigma(z)(1-\sigma(z))$ の最大値を求める。微分すると

$$f'(z) = \sigma'(z)(1-2\sigma(z)) = 0 \iff \sigma(z) = 1/2 \iff z = 0$$

よって $\max f(z) = f(0) = 1/2 \cdot 1/2 = 1/4$。 □

系 1.46 (有効スペクトル半径)**.** 任意のベクトル v に対し

$$\|A_\tau v\| = \|W^\top \operatorname{diag}(\sigma'(z_\tau))v\| \le \|W\| \cdot \frac{1}{4}\|v\|$$

従って $\rho(A_\tau) \le \|W\|/4$。

たとえ $\|W\| \approx 1$（適切な初期化）でも $\rho(A_\tau) \le 1/4 < 1$ となり、長期依存を保持できない。

補題 9.4：LSTM の定数誤差伝播

補題 1.47 (LSTM 細胞状態の勾配)**.** LSTM の細胞状態更新 $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$ に対し

$$\frac{\partial c_t}{\partial c_s} = \prod_{\tau=s+1}^{t} \operatorname{diag}(f_\tau)$$

証明. 連鎖律より

$$\frac{\partial c_t}{\partial c_{t-1}} = \operatorname{diag}(f_t)$$

よって積表示が得られる。 □

注意 1.48 (長期依存の保持)**.** $f_\tau \approx 1$ に初期化（forget bias を大きく取る）すれば

$$\left\| \frac{\partial c_t}{\partial c_s} \right\| \approx 1$$

が長期的に維持され、勾配消失を回避できる。これが LSTM 成功の鍵。

第 10 章：Self-Attention の分散解析補題

補題 10.1：スケーリングによる分散安定化

補題 1.49 (縮尺付き内積の分散)**.** $q, k \in \mathbb{R}^d$ を独立 $\mathcal{N}(0, I_d)$ とする。スカラー

$$s = \frac{1}{\sqrt{d}} q^\top k$$

は平均 0、分散 1 の分布を持つ。

証明. 成分分解：

$$s = \frac{1}{\sqrt{d}} \sum_{m=1}^{d} q_m k_m$$

q_m, k_m 独立で $\mathbb{E}[q_m k_m] = 0$、$\mathbb{E}[(q_m k_m)^2] = \mathbb{E}[q_m^2]\mathbb{E}[k_m^2] = 1$。
よって

$$\mathrm{Var}(s) = \frac{1}{d} \sum_{m=1}^{d} \mathrm{Var}(q_m k_m) = \frac{1}{d} \cdot d = 1$$

□

補題 10.2：Softmax のリプシッツ定数

補題 1.50 (Softmax ヤコビアンの作用素ノルム)**.** Softmax $\alpha(z)_i = \exp(z_i)/\sum_j \exp(z_j)$ のヤコビアン $J_\alpha(z)$ に対し

$$\sup_z \|J_\alpha(z)\|_{op} \le \frac{1}{2}$$

スケッチ. ヤコビアンは

$$(J_\alpha)_{ij} = \begin{cases} \alpha_i(1-\alpha_i) & (i=j) \\ -\alpha_i \alpha_j & (i \neq j) \end{cases}$$

この行列の最大特異値は、α が一様分布に近いとき最大となり、その値は高々 $1/2$ 程度。
詳細な証明は行列解析の標準的手法による。 □

補題 10.3：RoPE による相対位置不変性

補題 1.51 (複素回転埋め込みの性質)**.** 複素回転埋め込み $q_m \mapsto q_m e^{im\theta}$、$k_n \mapsto k_n e^{in\theta}$ のもとで

$$\Re\left(q_m e^{im\theta} \cdot \overline{k_n e^{in\theta}}\right) = \Re(q_m \overline{k_n} e^{i(m-n)\theta})$$

となり、内積は位置差 $m-n$ のみに依存。

証明. 複素共役の性質より

$$\overline{k_n e^{in\theta}} = \overline{k_n} e^{-in\theta}$$

よって

$$q_m e^{im\theta} \cdot \overline{k_n e^{in\theta}} = q_m \overline{k_n} e^{i(m-n)\theta}$$

実部をとれば主張を得る。 □

注意 1.52 (実装への示唆)**.** RoPE は絶対位置情報を使わずに相対位置を符号化。これにより訓練時より長い系列への外挿が可能（length extrapolation）。GPT-NeoX, LLaMA 等で採用。

補題 9.A：Gelfand の公式と勾配消失/爆発

補題 1.53 (反復線形写像のノルム成長)**.** 任意の行列 $A \in \mathbb{R}^{d\times d}$ に対し

$$\lim_{t\to\infty} \|A^t\|^{1/t} = \rho(A)$$

が成り立つ。ここで $\rho(A)$ はスペクトル半径（最大固有値の絶対値）。

応用：*Vanilla RNN* の勾配. Vanilla RNN の隠れ状態遷移を線形化すると

$$h_t \approx A h_{t-1}, \quad A = W_{hh}\,\mathrm{diag}(\sigma'(z_t))$$

勾配の逆伝播は

$$\left\| \frac{\partial h_t}{\partial h_0} \right\| \approx \|A^t\| \sim \rho(A)^t$$

したがって：

- $\rho(A) < 1 \to$ 勾配は指数的に減衰（**勾配消失**）
- $\rho(A) > 1 \to$ 勾配は指数的に発散（**勾配爆発**）
- $\rho(A) \approx 1 \to$ 長期依存を保持（理想的だが不安定）

□

補題 9.B：Sigmoid 活性化によるスペクトル半径の縮小

補題 1.54 (Sigmoid の微分上界)**.** Sigmoid 活性化 $\sigma(z) = 1/(1+e^{-z})$ に対し

$$\sigma'(z) \in [0, 1/4], \quad \max_z \sigma'(z) = 1/4$$

応用：有効スペクトル半径. 任意のベクトル v に対し

$$\begin{aligned} \|Av\| &= \|W_{hh}\,\mathrm{diag}(\sigma'(z_t))v\| \\ &\le \|W_{hh}\| \cdot \max_i |\sigma'(z_{t,i})| \cdot \|v\| \\ &\le \|W_{hh}\| \cdot \frac{1}{4}\|v\| \end{aligned}$$

したがって $\rho(A) \leq \|W_{hh}\|/4$。
たとえ $\|W_{hh}\| \approx 1$（適切な初期化）でも

$$\rho(A) \leq 1/4 < 1$$

となり、長期依存を保持できない（勾配が指数的に減衰）。
これが Vanilla RNN の**根本的限界**であり、LSTM や GRU が必要な理由。 □

注意 1.55 (LSTM の解決策)**.** LSTM の細胞状態 c_t は

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$$

ゲート f_t（忘却ゲート）が 1 に近ければ

$$\frac{\partial c_t}{\partial c_0} \approx \prod_{s=1}^{t} f_s \approx 1$$

となり、スペクトル半径 ≈ 1 を保持可能。

第 10 章：Self-Attention の分散解析補題

補題 10.A：スケーリングによる分散安定化

補題 1.56 (縮尺付き内積の分散)**.** $q, k \in \mathbb{R}^d$ を独立 $\mathcal{N}(0, I_d)$ とする。スカラー

$$s = \frac{1}{\sqrt{d}} q^\top k$$

は平均 0、分散 1 の分布を持つ。

証明. 各成分に分解：

$$s = \frac{1}{\sqrt{d}} \sum_{m=1}^{d} q_m k_m$$

q_m, k_m 独立で $\mathbb{E}[q_m k_m] = 0$、$\mathbb{E}[(q_m k_m)^2] = \mathbb{E}[q_m^2]\mathbb{E}[k_m^2] = 1$。
よって

$$\begin{aligned} \mathrm{Var}(s) &= \frac{1}{d} \sum_{m=1}^{d} \mathrm{Var}(q_m k_m) \\ &= \frac{1}{d} \sum_{m=1}^{d} \mathbb{E}[(q_m k_m)^2] \\ &= \frac{d}{d} = 1 \end{aligned}$$

□

注意 1.57 (スケーリングの必要性)。スケーリング $1/\sqrt{d}$ がなければ

- $\mathrm{Var}(q^\top k) = d \to$ 分散が次元に比例して増大
- Softmax 入力が極端な値 → 勾配消失（ほぼ 1-hot）
- 訓練不安定

スケーリングにより**次元に依存しない安定な分布**が得られる。

補題 10.B：Softmax の勾配爆発回避

補題 1.58 (Softmax ヤコビアンの有界性)**.** Softmax $\alpha_i = \exp(z_i)/\sum_j \exp(z_j)$ のヤコビアン行列 J の最大特異値は

$$\sigma_{\max}(J) \le 1$$

かつ、入力分布が $\mathcal{N}(0,1)$ なら $\sigma_{\max}(J) \approx 1/4$ 程度。

スケッチ. Softmax のヤコビアンは

$$J_{ij} = \frac{\partial \alpha_i}{\partial z_j} = \begin{cases} \alpha_i(1-\alpha_i) & (i=j) \\ -\alpha_i\alpha_j & (i \ne j) \end{cases}$$

Case 1: 入力が $\mathcal{N}(0,1)$

ほとんどの $z_i \in [-3,3]$ に入り、指数関数の値が極端にならない。このとき各 α_i は比較的均等に分散し、ヤコビアンの特異値は 1 よりかなり小さい（$\approx 1/4$）。

Case 2: 分散が d_k に比例（スケーリングなし）

z_i の一部が極端に大きくなり、Softmax がほぼ 1-hot に：

$$\alpha_i \approx \begin{cases} 1 & (i = i^*) \\ 0 & (i \ne i^*) \end{cases}$$

このとき $J_{i^*i^*} \approx 0$、他も ≈ 0 となり**勾配消失**。 □

注意 1.59 (実装への示唆)**.** Scaled Dot-Product Attention では

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

$1/\sqrt{d_k}$ スケーリングにより：

1. Softmax 入力の分散を 1 に安定化
2. 勾配消失を回避
3. 深い Transformer でも訓練可能

これは Transformer 成功の**技術的キーポイント**の一つ。

第 2 章

ガウス過程演習解答

演習 Y.1 多変量ガウスの条件付き分布

問題.

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

とする。ブロック逆行列公式を用いて、条件付き分布

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

の平均と共分散

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

を導出せよ。

解答.

Step 1: 逆行列のブロック表示

逆行列のブロック表示を

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

とおく。

Step 2: 指数部の分解

密度の指数部を

$$Q(u, v) = \begin{pmatrix} u - m_u \\ v - m_v \end{pmatrix}^\top \Sigma^{-1} \begin{pmatrix} u - m_u \\ v - m_v \end{pmatrix}$$

と書き、展開すると

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

Step 3: 平方完成

v を固定して u に関する項だけを見ると、

となる。平方完成により

$$(u - m_{u|v})^\top A(u - m_{u|v}) + \text{定数 } (v),$$

となるような $m_{u|v}$ は

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

Step 4: ブロック逆行列公式の適用

ブロック逆行列公式から

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

が成り立つ。したがって

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

指数部が $(u - m_{u|v})^\top A(u - m_{u|v})$ の形なので、$u \mid v$ は平均 $m_{u|v}$、共分散 $\Sigma_{u|v}$ の正規分布となる。■

演習 Y.2 GP 回帰の事後平均・分散

問題. ガウス過程事前 $f \sim \mathcal{GP}(0, K)$、ノイズモデル $y_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ を仮定する。訓練点 $X = (x_1, \ldots, x_n)$、カーネル行列 K、新規点 $x_\star$ に対し、事後分布 $f(x_\star) \mid X, y$ の平均・分散が

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

であることを、演習 Y.1 の結果を用いて示せ。

解答.

Step 1: 事前分布の設定

潜在値ベクトル $f(X) = [f(x_1), \ldots, f(x_n)]^\top$ と $f_\star = f(x_\star)$ に対し、ガウス過程の定義より

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

Step 2: 観測モデル

観測 $y = f(X) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ なので

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

Step 3: 結合分布

$(y, f_\star)$ の結合分布は

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

Step 4: 条件付き分布の適用

演習 Y.1 の結果を $u = f_\star$, $v = y$ として適用する。$\Sigma_{vv} = K + \sigma^2 I$, $\Sigma_{uv} = k_\star^\top$, $\Sigma_{uu} = k_{\star\star}$ より

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

したがって主張が得られた。■

演習 Y.3 カーネルリッジ回帰の閉形式解

問題. RKHS $\mathcal{H}_K$ 上のカーネルリッジ回帰

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

を考える。代表定理により $\hat{f}(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$ と書けることを用いて、係数ベクトル α の閉形式を求めよ。

解答.

Step 1: ベクトル表示

代表定理より

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

であり、ベクトル表示で $\hat{f}(X) = K\alpha$。

Step 2: 目的関数

目的関数は

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

Step 3: 最適性条件

α で偏微分して 0 とおく：

$$\frac{\partial J}{\partial \alpha} = -\frac{2}{n} K^\top (y - K\alpha) + 2\lambda K\alpha = 0.$$

K 対称より $K^\top = K$ なので

$$K(y - K\alpha) = n\lambda K\alpha.$$

Step 4: 解の導出

K が半正定値かつ $K + n\lambda I$ が正則と仮定すると

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

よって

$\tilde{\lambda} = n\lambda$ と置き換えれば

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

これが閉形式解である。■

演習 Y.4 GP 回帰とカーネルリッジの同値性

問題. 演習 Y.2, Y.3 の結果を用いて、ガウス過程回帰の事後平均とカーネルリッジ回帰の予測値が、$\lambda = \sigma^2$ の対応のもとで一致することを示せ。

解答.

GP 回帰の事後平均（演習 Y.2 より）:

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

カーネルリッジの予測（演習 Y.3 より）:

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

同値性の証明

$\tilde{\lambda} = \sigma^2$ と選べば

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

したがって、**ガウス過程回帰の事後平均はカーネルリッジ回帰の解と一致する。**

解釈

この同値性は以下を意味する：

- 頻度論的な正則化付き ERM（カーネルリッジ）
- ベイズ的な事後推論（ガウス過程）

両者は点推定では同じ結果を与える。違いは、ガウス過程が事後分散による不確実性定量化を提供することである。■

演習 Y.5 1 次元トイデータでの数値確認

問題. 1 次元入力 $x_i \in [-1, 1]$ を等間隔に 10 点取り、真の関数を $f^\star(x) = \sin(2\pi x)$、ノイズ $\varepsilon_i \sim \mathcal{N}(0, 0.1^2)$ とする。RBF カーネル $K(x, x') = \exp(-(x - x')^2 / 2\ell^2)$ を用い、パラメータ $\ell = 0.3$, $\sigma^2 = 0.1^2$ とする。

1. Python 等で GP 回帰の事後平均 $m_\star(x)$ を計算し、$x \in [-1, 1]$ 上でプロットせよ。
2. 同じカーネル・$\lambda = \sigma^2$ を用いてカーネルリッジ回帰の予測 $\hat{f}(x)$ を計算し、GP の曲線と重ねて描け。
3. 数値的に両者が一致していることを確認せよ。

解答のポイント.

理論的な一致

演習 Y.4 より、両者は理論的に完全に同じ式で与えられる：

$$m_\star(x) = k(x)^\top (K + \sigma^2 I)^{-1} y, \quad \hat{f}(x) = k(x)^\top (K + \sigma^2 I)^{-1} y.$$

実装の手順

1. データ生成：

```
import numpy as np
X_train = np.linspace(-1, 1, 10)
y_train = np.sin(2*np.pi*X_train) + 0.1*np.random.randn(10)
X_test = np.linspace(-1, 1, 100)
```

2. カーネル行列の計算：

```
def rbf_kernel(X1, X2, ell=0.3):
    sqdist = np.sum(X1**2, 1).reshape(-1, 1) + \
             np.sum(X2**2, 1) - 2*np.dot(X1, X2.T)
    return np.exp(-0.5 * sqdist / ell**2)

K = rbf_kernel(X_train.reshape(-1,1), X_train.reshape(-1,1))
k_star = rbf_kernel(X_test.reshape(-1,1), X_train.reshape(-1,1))
```

3. 予測計算：

```
sigma2 = 0.01  # 0.1^2
alpha = np.linalg.solve(K + sigma2*np.eye(10), y_train)
y_pred = k_star.dot(alpha)
```

4. プロット：

```
import matplotlib.pyplot as plt
plt.plot(X_test, y_pred, 'b-', label='GP/KRR prediction')
plt.plot(X_train, y_train, 'ro', label='Training data')
plt.plot(X_test, np.sin(2*np.pi*X_test), 'g--',
         label='True function')
plt.legend()
```

数値確認

GP 回帰とカーネルリッジの予測を両方計算し、差のノルムを確認：

* 元々とも同じ式なので差は数値誤差のみ

```
diff = np.linalg.norm(y_pred_gp - y_pred_krr)
print(f"Difference: {diff}")  # ~1e-15 程度
```

数値誤差を無視すれば、完全に一致することが確認できる。■

補足：事後分散の可視化

ガウス過程の利点は事後分散による不確実性定量化である。追加演習として、事後分散

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

を計算し、$m_\star(x) \pm 2\sqrt{v_\star(x)}$ の信頼区間をプロットすることで、データから離れた領域での不確実性の増加を可視化できる。

```
# 事後分散の計算
k_starstar = rbf_kernel(X_test.reshape(-1,1),
                        X_test.reshape(-1,1)).diagonal()
v_star = k_starstar - np.sum(k_star *
         np.linalg.solve(K + sigma2*np.eye(10), k_star.T).T,
         axis=1)
std = np.sqrt(v_star)

# 信頼区間のプロット
plt.fill_between(X_test, y_pred - 2*std, y_pred + 2*std,
                 alpha=0.2, label='95% confidence')
```

これにより、カーネルリッジには無い「不確実性」の情報が得られることがわかる。