

はじめに

本書は、これから「機械学習」の仕組みを「最適化」の視点から学習しようとする方を対象にしています。

「機械学習」とは、コンピュータがデータをもとに学習し、経験を積むことで予測や判断ができるようになる技術のことで、私たちの生活や仕事をより便利にしてくれる人工知能（AI）の一分野として位置付けられています。「コンピュータがデータをもとに学習」することは、実は、ある関数を「最適化」することと同等です。

この関数が 1、2 変数からなる単純なものであれば代数計算により最適化が可能ですが、実用上使われる機械学習モデルでは、とてつもない数のデータと大規模で複雑な機械構造を利用するため、多変数からなる複雑な関数の最適化が必要となります。つまり、機械学習の理解には、多変数関数解析のような数学の知識が必要になります。

このような複雑な関数を最適化する手法、つまり、機械学習をするための手法（オプティマイザ）には、例えば、確率的勾配降下法があります。その他に、モーメンタム法、Adam（アダム）、といったオプティマイザが有名です。もしかしたら、良さそうなオプティマイザらしいので、なんとなく、これらを機械学習に用いた、という方もいらっしゃるのではないのでしょうか。

本書では、このようなオプティマイザがなぜ機械学習に適しているかを理論的観点から詳しく説明します。オプティマイザ構造の理論的理解により、なんとなくではなく、自信を持って対象の機械学習モデルに最も適したオプティマイザの選択と調整ができます。理解のためには、脳

内数学ではなく、泥臭い数学（手書き計算の習慣）がとても大切です。
本書の原稿に関して、学生目線から貴重な意見をくれた明治大学数理
最適化研究室の皆さんに感謝します。出版に関して、技術評論社書籍
編集部の皆様には終始お世話になりました。ここに深く御礼申し上げます。

最後に、機械学習と最適化の世界に挑もうとする若い人たちに、この
本を贈りたい。

2026年1月 生田にて
飯塚秀明

記号表..... 7

第1章 数学の復習

1-1 論理と集合..... 14
 1-2 ユークリッド空間..... 16
 1-3 線形代数..... 24
 1-4 微分積分..... 30
 1-5 確率・統計..... 38

第2章 機械学習モデルを訓練する

2-1 機械学習..... 48
 2-2 機械学習モデルを訓練するには?..... 51
 2-3 機械学習モデルの訓練は経験損失の最適化..... 59
 [章末問題]..... 66

第3章 機械学習モデルの最適化

3-1 経験損失の最小解..... 68
 3-2 経験損失は微分可能..... 74
 3-3 経験損失は局所的凸関数..... 83
 [章末問題]..... 90

第4章 識別の正解率を上げる

4-1 経験損失を下げる..... 92
 4-2 経験損失は平滑関数..... 100
 [章末問題]..... 104

第5章 ステップサイズを理解する

5-1 オプティマイザ(反復法)..... 108
 5-2 経験損失の平滑性を利用するステップサイズ..... 113
 5-3 経験損失を下げるステップサイズ..... 119
 [章末問題]..... 128



第6章 勾配降下法

6-1 オプティマイザの収束性	132
6-2 勾配降下法の収束性	135
6-3 非凸平滑経験損失の最適化	140
6-4 凸平滑経験損失の最適化	147
[章末問題]	157

第7章 訓練データの標本調査

7-1 勾配降下法の問題点	160
7-2 標本調査による推定量	163
7-3 ミニバッチ損失とミニバッチ勾配の不偏性	168
7-4 ミニバッチ勾配の分散	173
[章末問題]	182

第8章 確率的勾配降下法1

8-1 確率的勾配降下法の構成	184
8-2 確率的勾配降下法の探索方向	188
8-3 勾配降下法に対する更新の比較	197
[章末問題]	203

第9章 確率的勾配降下法2 定数バッチサイズ

9-1 定数バッチサイズによる収束性	206
9-2 非凸平滑経験損失の最適化	210
9-3 凸平滑経験損失の最適化	222
[章末問題]	229

第10章 確率的勾配降下法3 勾配降下法に近づける

10-1 確率的勾配降下法の利点	232
10-2 バッチサイズを大きくする	234
[章末問題]	241

第11章	確率的勾配降下法4 増加バッチサイズ	
	11-1 増加バッチサイズによる収束性	244
	11-2 非凸平滑経験損失の最適化 (定数・減少ステップサイズと増加バッチサイズ)	249
	11-3 非凸平滑経験損失の最適化 (増加ステップサイズと増加バッチサイズ)	258
	11-4 凸平滑経験損失の最適化	266
	[章末問題]	271

第12章	バッチサイズを理解する	
	12-1 確率的勾配降下法の計算量	274
	12-2 確率的勾配降下法の計算量とバッチサイズ	276
	12-3 確率的勾配計算量の最小化に基づいた 確率的勾配降下法	284
	[章末問題]	297

第13章	確率的勾配降下法を加速する	
	13-1 モーメンタム付き確率的勾配降下法	300
	13-2 適応手法	318
	13-3 特異値分解を利用するオプティマイザ	328
	[章末問題]	335

第14章	汎化性能を高める	
	14-1 経験損失に正則化項を加える	338
	14-2 平坦な最小解を見つける	345
	[章末問題]	361

	章末問題略解	362
	参考文献	377
	索引	379

記号表

論理と集合

$M \vee P$: 選言 (M または P)

$M \wedge P$: 連言 (M かつ P)

$\neg P$: 否定 (P ではない)

$M \Rightarrow P$: 推論 (M (が成り立つ) ならば P (が成り立つ))

$M \Leftrightarrow P$: 同値 ($(M \Rightarrow P) \wedge (P \Rightarrow M)$)

$M := P$: M を P として定義する ($M \stackrel{[定義]}{\Leftrightarrow} P$)

$\forall x(M(x) \Rightarrow P(x))$: すべての x に対して $M(x)$ ならば $P(x)$

$\exists x(M(x) \wedge P(x))$: $M(x)$ かつ $P(x)$ となる x が存在する

$x \in A$ ($A \ni x$) : x は集合 A の要素

$x \notin A$ ($A \not\ni x$) : x は集合 A の要素ではない

$A \subset B$ ($B \supset A$) : 集合 A は集合 B の部分集合

$A = B$: 集合 A は集合 B と同じ ($(A \subset B) \wedge (B \subset A)$)

$A = \{x: P(x)\}$: A は命題 $P(x)$ を満たす x の集合

$A \cup B$: 和集合 ($A \cup B := \{x: (x \in A) \vee (x \in B)\}$)

$A \cap B$: 積集合 ($A \cap B := \{x: (x \in A) \wedge (x \in B)\}$)

A^c : 補集合 ($A^c := \{x: \neg(x \in A)\} = \{x: x \notin A\}$)

$A \setminus B$: 差集合 ($A \setminus B := \{x: (x \in A) \wedge \neg(x \in B)\} = A \cap B^c$)

$\sup A$: 集合 A の上限 ($\sup A(x) := \sup\{A(x): x \in B\}$)

$\inf A$: 集合 A の下限 ($\inf A(x) := \inf\{A(x): x \in B\}$)

$\max A$: 集合 A の最大要素 ($\max A(x) := \max\{A(x): x \in B\}$)

$\min A$: 集合 A の最小要素 ($\min A(x) := \min\{A(x): x \in B\}$)

\mathbb{N} : 自然数全体の集合

$\bar{\mathbb{N}}$: 0 を含む自然数全体の集合 ($\bar{\mathbb{N}} := \{0\} \cup \mathbb{N}$)

\mathbb{R} : 実数全体の集合

\mathbb{R}_+ : 非負実数全体の集合 ($\mathbb{R}_+ := \{x \in \mathbb{R}: x \geq 0\}$)

$[T] : \{1, 2, \dots, T\}$ (ただし、 $T \in \mathbb{N}$)

$[S : T] : \{S, S + 1, \dots, T\}$ (ただし、 $S, T \in \mathbb{N} \wedge S \leq T$)

$\bigcup_{i=1}^{+\infty} A_i$: 集合 A_i の和集合 ($\bigcup_{i=1}^{+\infty} A_i := \{x : \exists i \in \mathbb{N} (x \in A_i)\}$)

$\bigcap_{i=1}^{+\infty} A_i$: 集合 A_i の積集合 ($\bigcap_{i=1}^{+\infty} A_i := \{x : \forall i \in \mathbb{N} (x \in A_i)\}$)

ユークリッド空間

\mathbb{R}^d : d 次元ユークリッド空間

\mathbf{x} : d 次元ユークリッド空間 \mathbb{R}^d のベクトル (点)

$\mathbf{0}$: d 次元ユークリッド空間 \mathbb{R}^d の零ベクトル

\mathbf{x}^\top : ベクトル \mathbf{x} の転置

$\langle \mathbf{x}, \mathbf{y} \rangle$: ベクトル \mathbf{x} と \mathbf{y} の内積 ($\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x}^\top \mathbf{y}$)

$\|\mathbf{x}\|$: ベクトル \mathbf{x} のノルム ($\|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$)

(\mathbf{x}_t) : d 次元ユークリッド空間 \mathbb{R}^d の点列 ($(\mathbf{x}_t) = (\mathbf{x}_1, \dots, \mathbf{x}_t, \dots)$)

$\lim_{t \rightarrow +\infty} \mathbf{x}_t = \mathbf{x}^*$: 点列 (\mathbf{x}_t) が \mathbf{x}^* に収束する ($\mathbf{x}_t \rightarrow \mathbf{x}^*$)

$B(\mathbf{0}; M)$: 中心 $\mathbf{0}$ と半径 M の球 ($B(\mathbf{0}; M) := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq M\}$)

(\mathbf{x}_{t_i}) : 点列 (\mathbf{x}_t) の部分列 ($(\mathbf{x}_{t_i}) = (\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_i}, \dots) \subset (\mathbf{x}_t)$)

$\lim_{i \rightarrow +\infty} \mathbf{x}_{t_i} = \mathbf{x}^*$: 部分列 (\mathbf{x}_{t_i}) が \mathbf{x}^* に収束する ($\mathbf{x}_{t_i} \rightarrow \mathbf{x}^*$)

$\lim_{s \downarrow 0} x(s)$: 正数 s が零に収束するときの $(x(s))$ の極限

$\overline{\lim}_{t \rightarrow +\infty} x_t$: 実数列 (x_t) の上極限

$\underline{\lim}_{t \rightarrow +\infty} x_t$: 実数列 (x_t) の下極限

$x_t = O(y_t)$: 大きい任意の t に対して $x_t \leq cy_t$ となる $c > 0$ が存在

($y_t \rightarrow 0$ のとき x_t は収束率 $O(y_t)$ をもつ)

$x_t = \mathcal{O}(y_t)$: $x_t \leq cy_t$ となる $t \in \mathbb{N}$ と $c > 0$ が存在

($O(y_t)$ が小さいとき x_t は収束率 $\mathcal{O}(y_t)$ をもつ)

$x = O(y)$: $x \leq cy$ となる $c > 0$ が存在

$x \approx y$: 実数 x と y は近似できる

$x \lesssim y$: 実数 x は y よりも近似的に小さい ($(x \leq y + z) \wedge (z \approx 0)$)

$x \ll y$: 実数 x は y よりも十分小さい

線形代数

$\mathbb{R}^{m \times n}$: $m \times n$ 行列 (m 行 n 列からなる行列) 全体の集合

$X = (x_{ij})$: (i, j) 成分が x_{ij} の行列

$\text{vec}(X)$: 行列 X の成分 x_{11}, \dots, x_{mn} を並べた mn 次ベクトル

X^{-1} : X の逆行列

I : 単位行列

O : 零行列

$\text{rank}(X)$: 行列 X の階数

$\text{Tr}(X)$: 正方行列の対角和

$X \bullet Y$: 行列 X と Y の内積 ($X \bullet Y := \text{Tr}(Y^T X)$)

$\|X\|_F$: 行列 X のフロベニウスノルム ($\|X\|_F := \sqrt{X \bullet X}$)

$\|X\|$: 行列 X の作用素ノルム ($\|X\| := \max_{\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{\|X\mathbf{x}\|}{\|\mathbf{x}\|}$)

\mathbb{O}^d : d 次直交行列全体の集合 ($\mathbb{O}^d := \{X \in \mathbb{R}^{d \times d} : X^T = X^{-1}\}$)

\mathbb{S}^d : d 次対称行列全体の集合 ($\mathbb{S}^d := \{X \in \mathbb{R}^{d \times d} : X^T = X\}$)

\mathbb{S}_+^d : d 次半正定値行列 (固有値が非負となる対称行列) 全体の集合

\mathbb{S}_{++}^d : d 次正定値行列 (固有値が正となる対称行列) 全体の集合

$\lambda_{\min}(X)$: 行列 X の最小固有値

$\lambda_{\max}(X)$: 行列 X の最大固有値

微分積分

$f: \mathbb{R}^d \rightarrow \mathbb{R}$: 定義域 \mathbb{R}^d から値域 \mathbb{R} への関数 f

$f \circ g: g: \mathbb{R}^d \rightarrow \mathbb{R}^d$ と $f: \mathbb{R}^d \rightarrow \mathbb{R}$ の合成関数 ($(f \circ g)(\mathbf{x}) := f(g(\mathbf{x}))$)

$\frac{\partial f(\mathbf{x})}{\partial x_i}$: x_i に関する $f(\mathbf{x})$ の偏微分係数

$f'(\mathbf{x}; \mathbf{d})$: 方向 \mathbf{d} に対する \mathbf{x} における f の方向微分係数

$\nabla f(\mathbf{x})$: \mathbf{x} における f の勾配

$\nabla f: \mathbb{R}^d \rightarrow \mathbb{R}^d$: f の (1 次) 導関数 (勾配)

$\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i}$: x_j に関する $\frac{\partial f(\mathbf{x})}{\partial x_i}$ の偏微分係数 ($f(\mathbf{x})$ の 2 次偏微分係数)

$\nabla^2 f(\mathbf{x})$: \mathbf{x} における f のヘッセ行列

確率・統計

$P(A)$: 事象 A の確率

A a.s. : 事象 A がほとんど至るところで (almost surely ; a.s.) 起きる

$DU(n)$: $1, \dots, n$ 上の離散一様分布

$DU_d(n)$: $1, \dots, n$ 上の d 次元離散一様分布

$\mathbb{E}_{\mathbf{X}}[\mathbf{Y}]$: \mathbf{X} に関する \mathbf{Y} の期待値

$\mathbb{E}_{\mathbf{X}}[\mathbf{Y}|\mathbf{z}], \mathbb{E}_{\mathbf{X}}[\mathbf{Y}|\mathbf{Z}]$: $\mathbf{Z} = \mathbf{z}$ を条件とする \mathbf{X} に関する \mathbf{Y} の期待値

$\mathbb{E}[\mathbf{Y}]$: \mathbf{Y} の全期待値

$\mathbb{V}_{\mathbf{X}}[\mathbf{Y}]$: \mathbf{X} に関する \mathbf{Y} の分散

$\mathbb{V}_{\mathbf{X}}[\mathbf{Y}|\mathbf{z}], \mathbb{V}_{\mathbf{X}}[\mathbf{Y}|\mathbf{Z}]$: $\mathbf{Z} = \mathbf{z}$ を条件とする \mathbf{X} に関する \mathbf{Y} の分散

$\lim_{t \rightarrow +\infty} \mathbf{X}_t = \mathbf{X}^*$ a.s. : 点列 (\mathbf{X}_t) が \mathbf{X}^* に概収束する ($\mathbf{X}_t \xrightarrow{\text{a.s.}} \mathbf{X}^*$)

機械学習モデル

n : 訓練データの総数

v : 検証データの総数

t : テストデータの総数

\mathbf{x}_i : i 番目の訓練データ

\mathbf{y}_i, y_i : i 番目の訓練データ \mathbf{x}_i の正解ラベル

(\mathbf{x}, \mathbf{y}) : 正解ラベル \mathbf{y} が付与されている訓練データ \mathbf{x}

θ : 機械学習モデル (ニューラルネットワーク) のパラメータ

d : 機械学習モデルのパラメータの次元 ($\theta \in \mathbb{R}^d$)

\hat{y}_i : 機械学習モデルによって出力された訓練データ \mathbf{x}_i に対する予測値

$f_i(\theta)$: 機械学習モデル θ における訓練データ $(\mathbf{x}_i, \mathbf{y}_i)$ に関する損失

$f(\theta)$: 機械学習モデル θ における全訓練データに関する経験損失

$f_T(\theta)$: 機械学習モデル θ における全テストデータに関する期待損失

$F_\lambda(\theta)$: 正則化付き経験損失 ($F_\lambda(\theta) := f(\theta) + \frac{\lambda}{2} \|\theta\|^2$)

$f_\rho(\theta)$: 経験損失 f の鋭さを考慮した修正経験損失 ($f_\rho(\theta) := \max_{\epsilon \in B(\mathbf{0}; \rho)} f(\theta + \epsilon)$)

$\tilde{f}_\rho(\theta)$: 鋭度正則化付き経験損失 ($\tilde{f}_\rho(\theta) := f(\theta) + \rho \|\nabla f(\theta)\|$)

機械学習モデルの最適化

θ^* , θ^* : 学習済みモデル (経験損失 f の最小解)

f^* , f^* : 学習済みモデルでの経験損失の値 (経験損失 f の最小値)

$\min_{\theta \in C} f(\theta)$: 集合 C 上での $f(\theta)$ の最小値

$\operatorname{argmin}_{\theta \in C} f(\theta)$: $f(\theta)$ を集合 C 上で最小にする点の集合

$\max_{\theta \in C} f(\theta)$: 集合 C 上での $f(\theta)$ の最大値

$\operatorname{argmax}_{\theta \in C} f(\theta)$: $f(\theta)$ を集合 C 上で最大にする点の集合

$N(\theta^*; \delta)$: θ^* の δ -近傍 ($N(\theta^*; \delta) := \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\| < \delta\}$)

L : 平滑関数 f (リプシッツ連続 ∇f) のリプシッツ定数

オブティマイザ

θ_t : 時刻 t における f の近似解

d_t : 時刻 t における探索方向 (f の降下方向)

η_t : 時刻 t でのステップサイズ (学習率)

η : 定数ステップサイズ

$\bar{\eta}$: ステップサイズの上限

$\underline{\eta}$: ステップサイズの下限

T : 総ステップ数

$\|\nabla f(\theta_t)\| \rightsquigarrow 0$: $\lim_{t \rightarrow +\infty} \|\nabla f(\theta_t)\| = 0$ ($\|\nabla f(\theta_t)\|$ の下極限が零)

$\mathcal{O}(x_t) \searrow \mathcal{O}(y_t)$: 収束率が $\mathcal{O}(x_t)$ から $\mathcal{O}(y_t)$ に変化する

ξ : 訓練データ $[n] = \{1, 2, \dots, n\}$ から無作為に一つ抽出された標本

$\xi \sim \text{DU}(n)$: 標本 ξ が離散一様分布 $\text{DU}(n)$ に従う

f_ξ : f の確率的損失 (標本 ξ に関する損失関数)

∇f_ξ : f の確率的勾配 (標本 ξ に関する損失関数の勾配)

$\xi_{t,i}$: 時刻 t に対する i 番目の標本

$f_{\xi_{t,i}}$: f の確率的損失 (標本 $\xi_{t,i}$ に関する損失関数)

$\nabla f_{\xi_{t,i}}$: f の確率的勾配 (標本 $\xi_{t,i}$ に関する損失関数の勾配)

b_t : 時刻 t に対する (ミニ) バッチサイズ (標本の大きさ)

b : 定数バッチサイズ

$\boldsymbol{\xi}_t$: 時刻 t に対する無作為標本 ($\boldsymbol{\xi}_t = (\xi_{t,1}, \dots, \xi_{t,b_t})^\top$)

B_t : $\boldsymbol{\xi}_t$ の成分からなる集合 ($B_t = \{\xi_{t,1}, \dots, \xi_{t,b_t}\}$)

f_{B_t} : 時刻 t におけるミニバッチ損失 ($f_{B_t} := \frac{1}{b_t} \sum_{i=1}^{b_t} f_{\xi_{t,i}}$)

∇f_{B_t} : 時刻 t におけるミニバッチ勾配 ($\nabla f_{B_t} = \frac{1}{b_t} \sum_{i=1}^{b_t} \nabla f_{\xi_{t,i}}$)

σ^2 : f の確率的勾配 ∇f_ξ の分散の上界

K : 1 エポック ($K = \lceil \frac{n}{b} \rceil = \min\{x \in \mathbb{N} : \frac{n}{b} \leq x\}$)

E : 総エポック数

\mathbf{x}^2 : \mathbf{x} の成分を2乗したベクトル ($\mathbf{x}^2 := \mathbf{x} \odot \mathbf{x} = (x_1^2, \dots, x_i^2, \dots, x_d^2)^\top$)

$\frac{1}{\sqrt{v}}$: $\frac{1}{\sqrt{v_1}}, \dots, \frac{1}{\sqrt{v_d}}$ を対角成分にもつ対角行列

$N(\eta_t, b_t)$: ステップサイズ η_t とバッチサイズ b_t に関するノイズ

B_T : バイアス項 ($f(\boldsymbol{\theta}_0) - f^*$ を含む項)

V_T, V : バリエーション項 (σ^2 を含む項)

$\|\nabla f(\boldsymbol{\theta}_t)\| \xrightarrow{\text{a.s.}} 0$: $\lim_{t \rightarrow +\infty} \|\nabla f(\boldsymbol{\theta}_t)\| = 0$ a.s.

$f \leftarrow g$: 関数 f に g を適用する

ε : 正精度

T_ε : 反復計算量

N_ε : 確率的勾配計算量

b_ε^* : クリティカルバッチサイズ

第 1 章

数学の復習

本書で必要な数学の内容を復習します。第 2 章と第 3 章で扱う機械学習モデルと第 4 章から第 6 章で扱う機械学習のための勾配降下法は、論理と集合（第 1-1 節）、ユークリッド空間（第 1-2 節）、線形代数（第 1-3 節）、微分積分（第 1-4 節）に基づいています。第 7 章以降で扱う機械学習のための確率的勾配降下法の理解のためには、勾配降下法と確率・統計（第 1-5 節）の知識が必要です。

本書で利用する重要な数学の性質をまとめています。これらの性質を利用することで、第 2 章以降を読み進めることができます。なお、本書で利用する記号については、前頁にまとめています。

1-1 論理と集合

論理

全称命題「すべての x に対して命題関数 $P(x)$ が成り立つ」ことを $\forall xP(x)$ と書きます。また、全称命題「すべての x に対して $M(x)$ が成り立つならば $P(x)$ が成り立つ」ことを $\forall x(M(x) \Rightarrow P(x))$ と書きます。

存在命題「 $P(x)$ が成り立つような x が存在する」ことを $\exists xP(x)$ と書きます。また、存在命題「 $M(x)$ と $P(x)$ が成り立つような x が存在する」ことを $\exists x(M(x) \wedge P(x))$ と書きます。

命題 $P(x)$ の否定命題を $\neg P(x)$ と書きます。否定命題については以下が成り立ちます。ただし、命題を結ぶ等号 $=$ は左辺の真理値と右辺の真理値が一致することを意味します。

[全称命題の否定命題と存在命題の否定命題]

1. $\neg(\forall x(M(x) \Rightarrow P(x))) = \exists x(M(x) \wedge \neg P(x))$
2. $\neg(\exists x(M(x) \wedge P(x))) = \forall x(M(x) \Rightarrow \neg P(x))$

本書で扱う証明について、以下で説明します。

- 背理法：否定命題 $\neg P$ が偽であることを示すことで元の命題 P が成り立つことを示す証明
- 反例：全称命題 $\forall xP(x)$ が成り立たないことを示すためにその否定 $\exists x(\neg P(x))$ が成り立つこと（つまり、 $P(x)$ が成り立たないような x の例を挙げること）を示す証明

集合

\mathbb{R} を実数全体の集合とし、 \mathbb{N} を自然数全体の集合とします。 \mathbb{R}_+ を非負実数全体の集合とし、 $\overline{\mathbb{N}}$ を 0 を含む自然数全体の集合 (つまり、 $\overline{\mathbb{N}} = \{0\} \cup \mathbb{N}$) とします。 $T \in \mathbb{N}$ に対して、 $[T] := \{1, 2, \dots, T\}$ とし、 $[0 : T] := \{0, 1, \dots, T\}$ とします (記号 “ $A := B$ ” は “ A を B として定義すること” を意味します)。

集合 A が命題 $P(x)$ を満たすような x の集合のとき、 $A = \{x : P(x)\}$ と書くことにします。ただし、集合を結ぶ等号 $=$ は左辺の集合と右辺の集合の要素が完全に一致することを意味します。 x が集合 A の要素 (または、元) であることを $x \in A$ 、または、 $A \ni x$ と書きます。 x と y が A の要素 (つまり、 $(x \in A) \wedge (y \in A)$) であるとき、 $x, y \in A$ と書きます。集合 A と B において、 A のすべての要素が B の要素となるとき、 A は B の部分集合といい、 $A \subset B$ 、または、 $B \supset A$ と書きます。 $A = B$ の必要十分条件は $(A \subset B) \wedge (B \subset A)$ です。

考察する対象全体の集合 U を全体集合、または、空間と呼び、その部分集合 A, B, \dots について考察します。 $(A \subset U) \wedge (B \subset U)$ を $A, B \subset U$ と書きます。要素を一つも含まない集合を空集合と呼び、 \emptyset と書きます。

四つの集合演算を以下でまとめます。

- 和 : $A \cup B := \{x : (x \in A) \vee (x \in B)\}$
- 積 : $A \cap B := \{x : (x \in A) \wedge (x \in B)\}$
- 補 : $A^c := \{x : \neg(x \in A)\} = \{x : x \notin A\}$
- 差 : $A \setminus B := \{x : (x \in A) \wedge \neg(x \in B)\} = A \cap B^c$

無限個の集合 A_i ($i \in \mathbb{N}$) の和と積は、それぞれ、 $\bigcup_{i=1}^{+\infty} A_i := \{x : \exists i \in \mathbb{N} (x \in A_i)\}$ と $\bigcap_{i=1}^{+\infty} A_i := \{x : \forall i \in \mathbb{N} (x \in A_i)\}$ のように定義されます。

1-2 ユークリッド空間

ベクトル空間

d 個の実数 x_1, x_2, \dots, x_d を成分にもつ \mathbf{x} の集合

$$\mathbb{R}^d := \left\{ \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_d \end{pmatrix} : \forall i \in [d] (x_i \in \mathbb{R}) \right\}$$

に以下で定義される和とスカラー倍の演算 (α は実数) を入れたとき、 \mathbb{R}^d をベクトル空間と呼びます。

$$[\text{和}] \mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_d \end{pmatrix} := \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_i + y_i \\ \vdots \\ x_d + y_d \end{pmatrix} \quad [\text{スカラー倍}] \alpha \mathbf{x} := \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_i \\ \vdots \\ \alpha x_d \end{pmatrix}$$

ベクトル空間 \mathbb{R}^d の要素 $\mathbf{x} (\in \mathbb{R}^d)$ をベクトルといいます。 $\mathbb{R} (= \mathbb{R}^1)$ の要素をスカラーと呼びます。ベクトル (複数の実数を縦にならべたもの) とスカラー (実数一つのもの) を区別するために、ベクトルを太字体 \mathbf{x} で、スカラーを細字体 x で表記します。ベクトル \mathbf{x} は転置記号 \top を用いて

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_d \end{pmatrix} = (x_1, \dots, x_i, \dots, x_d)^\top$$

と書くことができます。

● 内積とノルム

ベクトル空間 \mathbb{R}^d に

$$\begin{aligned}
 \text{[内積]} \langle \mathbf{x}, \mathbf{y} \rangle &:= \mathbf{x}^\top \mathbf{y} = (x_1, \dots, x_i, \dots, x_d) \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_d \end{pmatrix} & (1.1) \\
 &= x_1 y_1 + \dots + x_i y_i + \dots + x_d y_d = \sum_{i=1}^d x_i y_i
 \end{aligned}$$

を定義することができます。(1.1) で定義される内積は以下の四性質を満たすことが確認できます。ただし、 $\mathbf{0} := (0, 0, \dots, 0)^\top$ とします。

[内積の四性質]

1. $\forall \mathbf{x} (\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \wedge (\langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}))$
2. $\forall \mathbf{x} \forall \mathbf{y} \forall \mathbf{z} (\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle)$
3. $\forall \alpha \forall \mathbf{x} \forall \mathbf{y} (\langle \alpha \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle)$
4. $\forall \mathbf{x} \forall \mathbf{y} (\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle)$

内積の性質 2、3、および、4 は、(1.1) から、転置記号 \top を用いて、それぞれ、 $(\mathbf{x} + \mathbf{y})^\top \mathbf{z} = \mathbf{x}^\top \mathbf{z} + \mathbf{y}^\top \mathbf{z}$ 、 $(\alpha \mathbf{x})^\top \mathbf{y} = \alpha \mathbf{x}^\top \mathbf{y}$ 、 $\mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x}$ と書くことができます。さらに、内積の性質 3 と 4 から

$$\langle \mathbf{x}, \alpha \mathbf{y} \rangle \stackrel{[\text{性質 4}]}{=} \langle \alpha \mathbf{y}, \mathbf{x} \rangle \stackrel{[\text{性質 3}]}{=} \alpha \langle \mathbf{y}, \mathbf{x} \rangle \stackrel{[\text{性質 4}]}{=} \alpha \langle \mathbf{x}, \mathbf{y} \rangle \quad (1.2)$$

を満たすので、内積内部の右側にあるスカラー α を内積の外に出すことができます。(1.1) で定義される内積を有するベクトル空間 \mathbb{R}^d を **内積空間** と呼びます。

(1.1) で定義される内積は性質 1 ($\langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^\top \mathbf{x} \geq 0$) を有するので

$$[\text{ノルム}] \|\mathbf{x}\| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbf{x}^\top \mathbf{x}} \stackrel{(1.1)}{=} \sqrt{\sum_{i=1}^d x_i^2} \quad (1.3)$$

を定義することが可能です (ノルムはユークリッドノルム (距離) のことです)。 (1.3) で定義されるノルムは以下の三性質を満たすことが確認できます。

[ノルムの三性質]

1. $\forall \mathbf{x} (\|\mathbf{x}\| \geq 0 \wedge (\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}))$
2. $\forall \alpha \forall \mathbf{x} (\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|)$
3. $\forall \mathbf{x} \forall \mathbf{y} (\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|) \triangleleft$ [三角不等式]

ユークリッド空間

(1.3) で定義されるノルムを有する内積空間 \mathbb{R}^d を (d 次元) **ユークリッド空間** と呼びます。ユークリッド空間のベクトルを点ともいいます。以下で、本書で利用する等式や不等式をまとめます。

[ユークリッド空間で成り立つ等式や不等式]

$\alpha \in \mathbb{R}$ 、 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ とします。

1. $\|\mathbf{x} \pm \mathbf{y}\|^2 = \|\mathbf{x}\|^2 \pm 2\langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{y}\|^2$
2. $\|\alpha \mathbf{x} + (1-\alpha)\mathbf{y}\|^2 = \alpha\|\mathbf{x}\|^2 + (1-\alpha)\|\mathbf{y}\|^2 - \alpha(1-\alpha)\|\mathbf{x} - \mathbf{y}\|^2$
3. $|\|\mathbf{x}\| - \|\mathbf{y}\|| \leq \|\mathbf{x} - \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

点列の収束性

実数列 $(x_t) = (x_1, x_2, \dots, x_t, \dots)$ が実数 x^* に**収束**するとは、どのような (十分小さい) 正数 ε に対しても、ある自然数 t_0 が存在して自然数

t が t_0 よりも大きいならば x_t と x^* の距離が ε 未満になる、つまり

$$\forall \varepsilon > 0 \exists t_0 \in \mathbb{N} \forall t \in \mathbb{N} (t > t_0 \Rightarrow |x_t - x^*| < \varepsilon) \quad (1.4)$$

を満たすときをいい、 $\lim_{t \rightarrow +\infty} x_t = x^*$ または $x_t \rightarrow x^* (t \rightarrow +\infty)$ 、単に、 $x_t \rightarrow x^*$ と書きます。(1.4) は、実数列 (x_t) の添字 t が十分に大きいとき (ある自然数 t_0 よりも大きい自然数 t のとき)、 x_t は x^* に十分近い (x_t と x^* との距離 $|x_t - x^*|$ が十分小さい正数 ε よりも小さい) ことを表現しています。

正数 s を変数にもつ実数値関数を $x(s)$ とします。 s が零に収束するとき、 $x(s)$ がある実数 x_* に収束することを、(1.4) での議論と同様にして

$$\forall \varepsilon > 0 \exists s_0 > 0 \forall s > 0 (s < s_0 \Rightarrow |x(s) - x_*| < \varepsilon) \quad (1.5)$$

として定義し、これを $\lim_{s \downarrow 0} x(s) = x_*$ または $x(s) \rightarrow x_* (s \downarrow 0)$ と書きます。

ユークリッド空間 \mathbb{R}^d の点列 $(\mathbf{x}_t) = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t, \dots)$ が \mathbb{R}^d の点 \mathbf{x}^* に収束するとは

$$\lim_{t \rightarrow +\infty} \|\mathbf{x}_t - \mathbf{x}^*\| = 0 \quad (1.6)$$

つまり、 \mathbf{x}_t と \mathbf{x}^* の間のノルム $x_t := \|\mathbf{x}_t - \mathbf{x}^*\|$ で定義される実数列 (x_t) が零に収束することで、 $\lim_{t \rightarrow +\infty} \mathbf{x}_t = \mathbf{x}^*$ または $\mathbf{x}_t \rightarrow \mathbf{x}^* (t \rightarrow +\infty)$ 、単に、 $\mathbf{x}_t \rightarrow \mathbf{x}^*$ と書きます。 \mathbf{x}^* を (\mathbf{x}_t) の極限といいます。実数列の定義を利用することで

$$\mathbf{x}_t \rightarrow \mathbf{x}^* \Leftrightarrow x_t := \|\mathbf{x}_t - \mathbf{x}^*\| \rightarrow 0$$

$$\Leftrightarrow \forall \varepsilon > 0 \exists t_0 \in \mathbb{N} \forall t \in \mathbb{N} (t > t_0 \Rightarrow |x_t| := \|\mathbf{x}_t - \mathbf{x}^*\| < \varepsilon) \quad (1.4)$$

となります。

有界な点列

\mathbb{R}^d の点列 (\mathbf{x}_t) が有界であるとは、 \mathbf{x}_t と原点 $\mathbf{0}$ との間のノルム $\|\mathbf{x}_t\| = \|\mathbf{x}_t - \mathbf{0}\|$ がどのような添字 t であっても、ある値 M で抑えられるときをいいます。つまり

$$(\mathbf{x}_t) \subset \mathbb{R}^d \text{ が有界} \stackrel{\text{[定義]}}{\Leftrightarrow} \exists M \in \mathbb{R}_+ \forall t \in \mathbb{N} (\|\mathbf{x}_t\| \leq M)$$

となります。 (\mathbf{x}_t) が中心 $\mathbf{0}$ で半径 M の球 $B(\mathbf{0}; M) := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{0}\| \leq M\}$ に留まり続けるとき (つまり、 $(\mathbf{x}_t) \subset B(\mathbf{0}; M)$)、 \mathbf{x}_t のノルム $\|\mathbf{x}_t\|$ は発散しないことを示しています。つまり、有界な点列はノルムの意味で発散はしません。しかしながら、有界な点列が収束するとは限りません ([点列の有界性と収束性の関係 2] 参照)。

\mathbb{R}^d の点列 (\mathbf{x}_t) の部分的な点列

$$(\mathbf{x}_{t_i}) := (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \mathbf{x}_{t_3}, \dots) \subset (\mathbf{x}_t) = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots)$$

を (\mathbf{x}_t) の**部分列**といいます。ただし、 $t_1 < t_2 < t_3 < \dots$ とします。 (\mathbf{x}_t) の部分列 (\mathbf{x}_{t_i}) が点 \mathbf{x}^* に収束するとき、 $\lim_{i \rightarrow +\infty} \mathbf{x}_{t_i} = \mathbf{x}^*$ または $\mathbf{x}_{t_i} \rightarrow \mathbf{x}^* (i \rightarrow +\infty)$ 、単に、 $\mathbf{x}_{t_i} \rightarrow \mathbf{x}^*$ と書きます。

有界な点列とその収束性の関係を以下にまとめます。

[点列の有界性と収束性の関係]

(\mathbf{x}_t) はユークリッド空間 \mathbb{R}^d の点列とします。

1. (\mathbf{x}_t) が収束 $\Rightarrow (\mathbf{x}_t)$ は有界
2. (\mathbf{x}_t) が有界 \Rightarrow 収束する (\mathbf{x}_t) の部分列 (\mathbf{x}_{t_i}) が存在する
3. 有界な (\mathbf{x}_t) が収束 $\Rightarrow (\mathbf{x}_t)$ のすべての部分列が収束する
4. 有界な (\mathbf{x}_t) が収束 \Leftrightarrow 有界な (\mathbf{x}_t) の収束する部分列はすべて同じ点に収束する

上極限と下極限

有界な実数列 (x_t) は一般には収束しません。例えば

$$x_t := \begin{cases} 1 + \frac{1}{t} & (t \text{ が奇数}) \\ -\frac{1}{t} & (t \text{ が偶数}) \end{cases}$$

は、任意の $t \in \mathbb{N}$ に対して、 $-\frac{1}{2} \leq x_t \leq 2$ なので、有界な実数列です。しかしながら、 (x_t) は収束はしません。一方で、奇数番号の添字をもつ (x_t) の部分列 (x_1, x_3, x_5, \dots) は 1 に収束し、偶数番号の添字をもつ (x_t) の部分列 (x_2, x_4, x_6, \dots) は 0 に収束します。

一般に、有界な実数列 (x_t) の収束する部分列の収束先が最も大きいものを (x_t) の **上極限** といい、 $\overline{\lim}_{t \rightarrow +\infty} x_t$ と書きます。同様に、有界な実数列 (x_t) の収束する部分列の収束先が最も小さいものを (x_t) の **下極限** といい、 $\underline{\lim}_{t \rightarrow +\infty} x_t$ と書きます。上記の例では、 $\overline{\lim}_{t \rightarrow +\infty} x_t = 1$ 、 $\underline{\lim}_{t \rightarrow +\infty} x_t = 0$ となります。有界な実数列 (x_t) の上極限と下極限が一致するとき、それが (x_t) の極限となり、結果として、 (x_t) は収束します。このように、有界な実数列の極限の存在性は、一般には、保証されませんが、有界な実数列の上極限、および、下極限の存在性は保証されます。有界な実数列の収束性を証明したいときは、その上極限、および、下極限を利用することが定石です。

[上極限と下極限の性質]

(x_t) と (y_t) を有界な実数列とし、 x を実数とします。

1. $\forall t(x_t \leq y_t) \Rightarrow (\overline{\lim}_{t \rightarrow +\infty} x_t \leq \overline{\lim}_{t \rightarrow +\infty} y_t) \wedge (\underline{\lim}_{t \rightarrow +\infty} x_t \leq \underline{\lim}_{t \rightarrow +\infty} y_t)$
2. $\overline{\lim}_{t \rightarrow +\infty} (x_t + y_t) \leq \overline{\lim}_{t \rightarrow +\infty} x_t + \overline{\lim}_{t \rightarrow +\infty} y_t$
3. $\underline{\lim}_{t \rightarrow +\infty} (x_t + y_t) \geq \underline{\lim}_{t \rightarrow +\infty} x_t + \underline{\lim}_{t \rightarrow +\infty} y_t$
4. $x_t \rightarrow x \Rightarrow \begin{cases} \overline{\lim}_{t \rightarrow +\infty} (x_t + y_t) = x + \overline{\lim}_{t \rightarrow +\infty} y_t \\ \underline{\lim}_{t \rightarrow +\infty} (x_t + y_t) = x + \underline{\lim}_{t \rightarrow +\infty} y_t \end{cases}$

実数列の収束率

$(x_t), (y_t) \subset \mathbb{R}_+$ とします。ランダウの記号 O は以下のように定義されます。

$$x_t = O(y_t) \stackrel{[定義]}{\Leftrightarrow} \exists c > 0 \exists t_0 \in \mathbb{N} \forall t \in \mathbb{N} (t > t_0 \Rightarrow x_t \leq cy_t) \quad (1.7)$$

例えば、 $y_t = \frac{1}{t}$ とすると

$$\begin{aligned} x_t = O\left(\frac{1}{t}\right) &\Leftrightarrow \exists c > 0 \exists t_0 \in \mathbb{N} \forall t \in \mathbb{N} \left(t > t_0 \Rightarrow x_t \leq \frac{c}{t}\right) \\ &\Rightarrow 0 \leq x_t \leq \frac{c}{t} \rightarrow 0 \quad (t \rightarrow +\infty) \end{aligned}$$

を満たします。はさみうちの原理から、 $x_t = O(\frac{1}{t})$ で定義される正実数列 (x_t) は零に収束します。さらに、 (x_t) は $\frac{c}{t}$ が零に収束する速さと同程度であることも示しています。このことから、 $x_t = O(\frac{1}{t})$ で定義される正実数列 (x_t) は**収束率** $O(\frac{1}{t})$ を有すると呼ぶことにします。 $\frac{1}{t^2}$ は $\frac{1}{t}$ よりも速く零に収束します。そのことから、 $z_t = O(\frac{1}{t^2})$ で定義される正実数列 (z_t) は、 $x_t = O(\frac{1}{t})$ で定義される正実数列 (x_t) よりも速い収束率を有します。

ある特定の時刻 t に対して x_t が y_t の定数倍以下になるとき、ランダウの記号 O と区別して、 $x_t = \mathcal{O}(y_t)$ と書くことにします。正確な定義は以下のとおりです。

$$x_t = \mathcal{O}(y_t) \stackrel{[定義]}{\Leftrightarrow} \exists c > 0 \exists t \in \mathbb{N} (x_t \leq cy_t) \quad (1.8)$$

$x, y \in \mathbb{R}_+$ に対して、 x が y の定数倍以下になるとき、 $x = \mathcal{O}(y)$ と書くことにします。定義は以下のとおりです。

$$x = \mathcal{O}(y) \stackrel{[定義]}{\Leftrightarrow} \exists c > 0 (x \leq cy)$$

連続写像

写像 $f: \mathbb{R}^d \rightarrow \mathbb{R}^l$ が点 $\mathbf{x} \in \mathbb{R}^d$ で**連続**であるとは、 \mathbb{R}^d の点 \mathbf{y} ($\neq \mathbf{x}$) を連続的に \mathbf{x} に近づける (つまり、 $\mathbf{y} \rightarrow \mathbf{x}$) とき、 \mathbb{R}^l の点 $f(\mathbf{y})$ も連続的に $f(\mathbf{x})$ に近づく (つまり、 $f(\mathbf{y}) \rightarrow f(\mathbf{x})$) こととして定義されます。これを正確に記述すると

f が \mathbf{x} で連続 (1.9)

$$\Leftrightarrow \forall \varepsilon > 0 \exists \delta > 0 \forall \mathbf{y} \in \mathbb{R}^d (\|\mathbf{y} - \mathbf{x}\| < \delta \Rightarrow \|f(\mathbf{y}) - f(\mathbf{x})\| < \varepsilon)$$

[定義]

となります。さらに

$$f \text{ が } \mathbf{x} \text{ で連続} \Leftrightarrow [\mathbf{y}_t \rightarrow \mathbf{x} \Rightarrow f(\mathbf{y}_t) \rightarrow f(\mathbf{x})] \quad (1.10)$$

を示すことができます。つまり、点列 (\mathbf{y}_t) が $\mathbf{y}_1, \mathbf{y}_2, \dots$ のように離散的に \mathbf{x} に近づくとき、点列 $(f(\mathbf{y}_t))$ も離散的に $f(\mathbf{x})$ に近づくことと、 f の \mathbf{x} での連続性は同値です。

連続写像が最大値と最小値をもつための十分条件が連続関数の定義域がコンパクトになることです。

[最大値・最小値の定理]

ユークリッド空間 \mathbb{R}^d のコンパクト集合 C 上で定義される連続関数は C 上で最大値と最小値をもちます。

\mathbb{R}^d の部分集合 C が**コンパクト**であるとは

- C が有界、つまり、 $\exists M \in \mathbb{R} \forall \mathbf{x} \in C (\|\mathbf{x}\| \leq M)$ 、かつ、
- C が閉、つまり、 $(\mathbf{x}_t) \subset C (\mathbf{x}_t \rightarrow \mathbf{x}) \Rightarrow \mathbf{x} \in C$

のときをいいます。 C が有界のときは、 C に属するどのような点のノルム (距離) がある正数 M で抑えられることになるので、有界集合 C は無限に広がることのない有限な集合となります。 C が閉のときは、収束する C の点列の極限が必ず C に属することになるので、閉集合 C には境界が存在することになります。例えば、中心 $\mathbf{0}$ で半径 $r (> 0)$ の球 $B := \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x} - \mathbf{0}\| \leq r\} \stackrel{(1.3)}{=} \{\mathbf{x} \in \mathbb{R}^d: x_1^2 + \dots + x_d^2 \leq r^2\}$ は \mathbb{R}^d の有界閉集合です。

1-3 線形代数

行列全体からなるユークリッド空間

実数 x_{ij} を成分とする $m \times n$ 行列の集合は以下のように書くことができます。

$$\mathbb{R}^{m \times n} := \left\{ X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & \vdots & & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix} : \forall i \forall j (x_{ij} \in \mathbb{R}) \right\}$$

$m \times n$ 行列 X をその (i, j) 成分を用いて、 $X = (x_{ij})$ と書くことにします。 $\mathbb{R}^{m \times n}$ に以下で定義される和とスカラー倍の演算 (α は実数) を入れたとき、 $\mathbb{R}^{m \times n}$ はベクトル空間となります。

$$[\text{和}] X + Y = (x_{ij}) + (y_{ij}) := (x_{ij} + y_{ij}) \quad [\text{スカラー倍}] \alpha X := (\alpha x_{ij})$$

一方で、 $\mathbf{x}_j = (x_{1j}, \dots, x_{ij}, \dots, x_{mj})^\top$ を成分とするベクトル空間

$$\mathbb{R}^{mn} = \left\{ \text{vec}(X) := \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_j \\ \vdots \\ \mathbf{x}_n \end{pmatrix} : \forall j \left(\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{mj} \end{pmatrix} \in \mathbb{R}^m \right) \right\}$$

を定義することができます。 $X \in \mathbb{R}^{m \times n}$ を $\text{vec}(X) \in \mathbb{R}^{mn}$ へ、そして、 $\text{vec}(X) \in \mathbb{R}^{mn}$ を $X \in \mathbb{R}^{m \times n}$ に変換することができるので、ベクトル空間 $\mathbb{R}^{m \times n}$ はベクトル空間 \mathbb{R}^{mn} と同一視することができます。

ベクトル空間 \mathbb{R}^{mn} には、(1.1) (ただし、 $d = mn$) で定義される内積

$$\langle \text{vec}(X), \text{vec}(Y) \rangle := \text{vec}(X)^\top \text{vec}(Y) = \sum_{j=1}^n \langle \mathbf{x}_j, \mathbf{y}_j \rangle = \sum_{j=1}^n \sum_{i=1}^m x_{ij} y_{ij}$$

を定義することができます。 $\mathbb{R}^{m \times n}$ と $\mathbb{R}^{n \times m}$ が同一視できるので、上記の右辺を行列 $X = (x_{ij})$ と $Y = (y_{ij})$ で表現したものが $\mathbb{R}^{m \times n}$ の内積になります。行列 $Y = (y_{ij}) \in \mathbb{R}^{m \times n}$ の転置行列を転置記号 \top を用いて、 $Y^\top := (y_{ji}) \in \mathbb{R}^{n \times m}$ と定義します。このとき

$$\begin{aligned}
 Y^\top X &:= \begin{pmatrix} y_{11} & \cdots & y_{i1} & \cdots & y_{m1} \\ \vdots & & \vdots & & \vdots \\ y_{1j} & \cdots & y_{ij} & \cdots & y_{mj} \\ \vdots & & \vdots & & \vdots \\ y_{1n} & \cdots & y_{in} & \cdots & y_{mn} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & \vdots & & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix} \\
 &= \begin{pmatrix} \sum_{i=1}^m y_{i1}x_{i1} & \cdots & \sum_{i=1}^m y_{i1}x_{ij} & \cdots & \sum_{i=1}^m y_{i1}x_{in} \\ \vdots & \ddots & \vdots & & \vdots \\ \sum_{i=1}^m y_{ij}x_{i1} & \cdots & \sum_{i=1}^m y_{ij}x_{ij} & \cdots & \sum_{i=1}^m y_{ij}x_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ \sum_{i=1}^m y_{in}x_{i1} & \cdots & \sum_{i=1}^m y_{in}x_{ij} & \cdots & \sum_{i=1}^m y_{in}x_{in} \end{pmatrix}
 \end{aligned}$$

なので、 $n \times n$ 行列 $Y^\top X$ の対角和（対角成分の和）

$$\begin{aligned}
 \text{Tr}(Y^\top X) &:= \sum_{i=1}^m y_{i1}x_{i1} + \cdots + \sum_{i=1}^m y_{ij}x_{ij} + \cdots + \sum_{i=1}^m y_{in}x_{in} \\
 &= \sum_{j=1}^n \sum_{i=1}^m x_{ij}y_{ij}
 \end{aligned}$$

が、以下のような $\mathbb{R}^{m \times n}$ の内積 $X \bullet Y$ となります。

$$\begin{aligned}
 [\text{内積}] X \bullet Y &:= \text{Tr}(Y^\top X) = \text{Tr}(X^\top Y) = \sum_{j=1}^n \sum_{i=1}^m x_{ij}y_{ij} \\
 &= \langle \text{vec}(X), \text{vec}(Y) \rangle
 \end{aligned}$$

$\mathbb{R}^{m \times n}$ のノルムは、(1.3) のように内積の平方根で定義されるので

$$[\text{ノルム}] \|X\|_F := \sqrt{X \bullet X} = \sqrt{\text{Tr}(X^\top X)} = \sqrt{\sum_{j=1}^n \sum_{i=1}^m x_{ij}^2} \quad (1.11)$$

となります。(1.11)で定義されるノルムは**フロベニウスノルム**と呼ばれます。

以上のことから、フロベニウスノルム $\|\cdot\|_F$ を有する内積空間 $\mathbb{R}^{m \times n}$ はユークリッド空間となります。

● 作用素ノルム

フロベニウスノルム (1.11) の他にも行列のノルムを定義することができます。行列 $X \in \mathbb{R}^{m \times n}$ に対して、ユークリッド空間 \mathbb{R}^n からユークリッド空間 \mathbb{R}^m への写像 f を

$$\forall \mathbf{x} \in \mathbb{R}^n \quad (f(\mathbf{x}) := X\mathbf{x})$$

と定義します。このとき、 f は**線形作用素**、つまり、

$$\forall \mathbf{x} \in \mathbb{R}^n \quad \forall \mathbf{y} \in \mathbb{R}^n \quad \forall \alpha \in \mathbb{R} \quad \forall \beta \in \mathbb{R} \quad (f(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}))$$

を満たします。線形作用素 $f(\cdot) = X(\cdot) = X$ の**作用素ノルム**を

$$\|X\| := \sup_{\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}} \frac{\|X\mathbf{x}\|}{\|\mathbf{x}\|} \quad (1.12)$$

と定義します。(1.12)の右辺にある $\|X\mathbf{x}\|$ は \mathbb{R}^m のノルムであり、 $\|\mathbf{x}\|$ は \mathbb{R}^n のノルムです。このとき、作用素ノルム (1.12) がノルムの三性質を満たすことが確認できます。作用素ノルム (1.12) に対しては以下が成り立ちます。

$$\|X\| = \max_{\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}} \frac{\|X\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\| \leq 1} \|X\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|X\mathbf{x}\| \quad (1.13)$$

例えば、 $\|X\| = \max_{\|\mathbf{x}\|=1} \|X\mathbf{x}\|$ が成り立つので、 $\|X\mathbf{x}\|$ を $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$ 上で最大にする \mathbf{x}^* が存在し、 $\|X\| = \|X\mathbf{x}^*\|$ となることを意味します。(1.13)の成立には、連続関数 $\|f(\cdot)\| = \|X(\cdot)\|$ がコンパクト集合 $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1\}$ 上では最大値と最小値をもつことを保証する最大値・最小値の定理 (第 1-2 節) を利用します。

対称行列（半正定値行列と正定値行列）

d 次正方行列（行数と列数が共に d の $d \times d$ 行列） X がその転置 X^\top と一致するとき、 X を**対称行列**といいます。 d 次対称行列全体の集合を \mathbb{S}^d と書くことにします。例えば、2 回連続的の微分可能な関数のヘッセ行列は対称行列です（ヘッセ行列の重要な性質 2 参照）。 d 次正方行列 X に対して、 $X\mathbf{x} = \lambda\mathbf{x}$ を満たす $\mathbf{x} (\neq \mathbf{0})$ とスカラー λ が存在するとき、 \mathbf{x} と λ を、それぞれ、 X の**固有ベクトル**、 X の**固有値**と呼びます。 d 次正方行列 X の固有値を $\lambda_i(X)$ ($i \in [d]$) と書くことにします。また、 X の最大固有値と最小固有値を、それぞれ、 $\lambda_{\max}(X)$ 、 $\lambda_{\min}(X)$ と書くことにします。

以下で、機械学習や最適化の分野でよく現れる重要な対称行列とその性質をまとめます。

- $X \in \mathbb{S}^d$ (X が**対称行列**) $\Leftrightarrow \forall \mathbf{x} \forall \mathbf{y} (\langle X\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, X\mathbf{y} \rangle)$
 $\Rightarrow \forall i \in [d] (\lambda_i(X) \in \mathbb{R})$
- $X \in \mathbb{S}_+^d$ (X が**半正定値行列**) $\stackrel{[定義]}{\Leftrightarrow} \forall i \in [d] (\lambda_i(X) \geq 0)$
 $\Leftrightarrow \forall \mathbf{x} (\langle X\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{x}, X\mathbf{x} \rangle \geq 0)$
- $X \in \mathbb{S}_{++}^d$ (X が**正定値行列**) $\stackrel{[定義]}{\Leftrightarrow} \forall i \in [d] (\lambda_i(X) > 0)$
 $\Leftrightarrow \forall \mathbf{x} (\neq \mathbf{0}) (\langle X\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{x}, X\mathbf{x} \rangle > 0)$

このように、対称行列はその固有値の性質で特徴づけることができます。2 次形式と呼ばれる $\langle \mathbf{x}, X\mathbf{x} \rangle = \mathbf{x}^\top X\mathbf{x}$ は凸関数の議論で度々登場します（第 2 章；章末問題 2.1）。

(1.12) で定義される $X \in \mathbb{R}^{m \times n}$ の作用素ノルムは、固有値を用いて以下のように表現ができます。

1. $\|X\| = \sqrt{\lambda_{\max}(XX^\top)} = \sqrt{\lambda_{\max}(X^\top X)}$
2. $X \in \mathbb{S}_+^d \Rightarrow \|X\| = \sqrt{\lambda_{\max}(X^2)} = \lambda_{\max}(X)$

表現 1 は、 X の作用素ノルム $\|X\|$ が XX^\top の最大固有値の平方根（ X の最大特異値；(1.15) 参照）であることを意味します（ XX^\top の最大固有値は $X^\top X$ の最大固有値と一致します）。 XX^\top が m 次半正定値行列、そして、 $X^\top X$ が n 次半正定値行列になること（章末問題 3.4 参照）

から、それらの最大固有値 $\lambda_{\max}(XX^T)$ と $\lambda_{\max}(X^T X)$ は非負です。よって、それらの平方根は定義可能です。表現 2 は、 d 次半正定値行列 X の最大固有値がまさに X のノルムであることを意味します。機械学習や最適化の分野では、平滑関数（第 1-4 節や章末問題 4.3）の議論で作用素ノルムがよく登場します。

$X \in \mathbb{R}^{m \times n} \setminus \{O\}$ とします。行列 $XX^T \in \mathbb{S}_+^m$ の非負固有値 $\lambda_i(XX^T)$ ($i \in [m]$) を用いて $\sigma_i := \sqrt{\lambda_i(XX^T)}$ を定義します。 σ_i は以下のように並べることができます。

$$\exists r \geq 1 \quad (0 = \sigma_m = \cdots = \sigma_{r+1} < \sigma_r \leq \sigma_{r-1} \leq \cdots \leq \sigma_1) \quad (1.14)$$

を定義します。 r を X の階数 ($r = \text{rank}(X)$ と書きます) といい、特に、正の σ_i (XX^T の正固有値の平方根)、つまり、

$$\forall i \in [r] \quad \left(\sigma_i = \sqrt{\lambda_i(XX^T)} = \sqrt{\lambda_i(X^T X)} > 0 \right) \quad (1.15)$$

を X の特異値といいます。(1.14) で定義される σ_i を対角成分にもつ対角行列 Σ (対角成分 σ_i 以外がすべて零の行列) を用いて、 X は以下のように特異値分解が可能です。

$$\exists U \in \mathbb{O}^{m \times r} \quad \exists V \in \mathbb{O}^{n \times r} \quad \left(X = U \Sigma V^T \right) \quad (1.16)$$

ただし、 $\mathbb{O}^{m \times r}$ は $m \times r$ 次直交行列 ($U \in \mathbb{O}^{m \times r} \stackrel{[定義]}{\Leftrightarrow} U U^T = I$) 全体の集合を表します。

(1.11) で定義される $X \in \mathbb{R}^{m \times n}$ のフロベニウスノルムは、特異値を用いて以下のように表現ができます。

$$3. \quad \|X\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2} = \sqrt{\sum_{i=1}^r \lambda_i(XX^T)}$$

表現 3 は、 X のフロベニウスノルム $\|X\|_F$ が X の特異値の二乗和の平方根と一致することを意味します。この表現は、特異値分解 (1.16) と対角和 $\text{Tr}(\cdot)$ の性質を用いて示すことができます。

ここで、度々現れた $A := XX^T \in \mathbb{S}_+^m$ は、(1.16) を用いると

$$\begin{aligned} A := XX^T &\stackrel{(1.16)}{=} (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma V^T (V^T)^T \Sigma^T U^T \\ &= U\Sigma \underbrace{V^T V}_{=I} \Sigma U^T = U\Sigma^2 U^T \end{aligned} \quad (1.17)$$

のようになります。ただし、三つ目の等号は転置の性質 $((AB)^T = B^T A^T)$ から、そして、四つ目の等号は対角行列 Σ の対称性 $(\Sigma^T = \Sigma)$ と転置の対合律 $((A^T)^T = A)$ から成り立ちます。 $A = XX^T$ と同じくよく利用される $B := X^T X \in \mathbb{S}_+^n$ については、(1.17) を得るための同様の議論から、 $B = V\Sigma^2 V^T$ と書くことができます。さらに

$$A \stackrel{(1.17)}{=} U\Sigma^2 U^T \in \mathbb{S}_{++}^m \Rightarrow \begin{cases} \exists A^{-1} = U(\Sigma^2)^{-1} U^T =: U\Sigma^{-2} U^T \\ \exists A^{\frac{1}{2}} := U\Sigma^{\frac{1}{2}} U^T \quad \left(A = (A^{\frac{1}{2}})^2 \right) \end{cases}$$

の確認ができます。ただし、 $\Sigma^{\frac{1}{2}}$ は $\Sigma = (\Sigma^{\frac{1}{2}})^2$ を満たす行列です。以上の議論は、特異値分解を利用する最新のオプティマイザの解析でよく利用されます (第 13-2 節参照)。

以下に、行列の重要な性質をまとめます。

[行列の重要な性質]

点 \mathbf{x}, \mathbf{y} と行列 X に対して

1. $(X\mathbf{x})^T \mathbf{y} = \langle X\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, X^T \mathbf{y} \rangle = \mathbf{x}^T X^T \mathbf{y}$
2. $\|X\mathbf{x}\| \leq \|X\| \|\mathbf{x}\|$
3. $X \in \mathbb{S}^d \Rightarrow \lambda_{\min}(X) \|\mathbf{x}\|^2 \leq \langle \mathbf{x}, X\mathbf{x} \rangle \leq \lambda_{\max}(X) \|\mathbf{x}\|^2$
 $\Rightarrow \lambda_{\min}(X) = \min_{\mathbf{x} \neq \mathbf{0}} \frac{\langle \mathbf{x}, X\mathbf{x} \rangle}{\|\mathbf{x}\|^2}, \lambda_{\max}(X) = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\langle \mathbf{x}, X\mathbf{x} \rangle}{\|\mathbf{x}\|^2}$

1-4 微分積分

勾配 (1次導関数)

d 次元ユークリッド空間 \mathbb{R}^d 上で定義される実数値関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ が $\mathbf{x} \in \mathbb{R}^d$ で微分可能であるとは、ある (f と \mathbf{x} に依存する) 点 $\mathbf{g} \in \mathbb{R}^d$ が存在して

$$\lim_{\mathbf{h} \rightarrow 0} \frac{f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - \langle \mathbf{g}, \mathbf{h} \rangle}{\|\mathbf{h}\|} = 0 \quad (1.18)$$

が成り立つときをいいます。この点 \mathbf{g} を $\nabla f(\mathbf{x})$ と書くことにします。点 \mathbf{x} に対応する \mathbf{g} の関係は $\mathbb{R}^d \ni \mathbf{x} \mapsto \mathbf{g} = \nabla f(\mathbf{x}) \in \mathbb{R}^d$ で定義される写像として表現することができます。 $\nabla f(\mathbf{x})$ を f の \mathbf{x} における勾配と呼び、写像 $\nabla f: \mathbb{R}^d \rightarrow \mathbb{R}^d$ を f の (1次) 導関数、または、勾配と呼びます。 f の定義域の各点で微分可能なとき、 f は微分可能であるといえます。

ある方向ベクトル $\mathbf{d} (\in \mathbb{R}^d)$ に対して、 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ の $\mathbf{x} (\in \mathbb{R}^d)$ での方向微分係数を

$$f'(\mathbf{x}; \mathbf{d}) := \lim_{\eta \downarrow 0} \frac{f(\mathbf{x} + \eta \mathbf{d}) - f(\mathbf{x})}{\eta}$$

と定義します。一般の実数値関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ においては $f'(\mathbf{x}; \mathbf{d})$ の存在性 ($\frac{f(\mathbf{x} + \eta \mathbf{d}) - f(\mathbf{x})}{\eta}$ の極限の存在性) は保証されません。微分可能な関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ に対しては方向微分係数 $f'(\mathbf{x}; \mathbf{d})$ の存在が保証されます (勾配の重要な性質 2 参照)。方向ベクトル \mathbf{d} を第 i 成分が 1 でそれ以外が 0 の成分からなる単位ベクトル \mathbf{e}_i とすると

$$f'(\mathbf{x}; \mathbf{e}_i) = \lim_{\eta \downarrow 0} \frac{f(\mathbf{x} + \eta \mathbf{e}_i) - f(\mathbf{x})}{\eta} = \frac{\partial f(\mathbf{x})}{\partial x_i}$$

から、方向微分係数 $f'(\mathbf{x}; \mathbf{e}_i)$ は f の x_i に関する偏微分係数 $\frac{\partial f(\mathbf{x})}{\partial x_i}$ と一致します。つまり、微分可能な関数は偏微分可能です。

微分可能な関数の勾配の性質を以下でまとめます。

[勾配の重要な性質]

1. [勾配と偏微分係数の関係] $\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}$
2. [勾配と方向微分係数の関係] $\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle = f'(\mathbf{x}; \mathbf{d})$
3. [平均値の定理] $\forall \mathbf{x} \in \mathbb{R}^d \forall \mathbf{y} \in \mathbb{R}^d \exists \lambda \in (0, 1):$

$$f(\mathbf{y}) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle$$
4. [微分作用素 ∇ の線形性]

$$\forall \alpha_1 \in \mathbb{R} \forall \alpha_2 \in \mathbb{R} (\nabla(\alpha_1 f_1 + \alpha_2 f_2) = \alpha_1 \nabla f_1 + \alpha_2 \nabla f_2)$$

ただし、 f_1 と f_2 は微分可能とします。

勾配の重要な性質1は、 f の勾配 $\nabla f(\mathbf{x})$ は f の偏微分係数 $\frac{\partial f(\mathbf{x})}{\partial x_i}$ を縦に並べることで表現ができることを示しています。勾配の重要な性質2は、経験損失を下げるための方向の議論で重要な役割を演じます（第4-1節）。勾配の重要な性質3は、経験損失の最小解の最適性条件の議論で活躍します（第3-2節）。勾配の重要な性質4は、 n 個の微分可能な関数 f_i と実数 α_i ($i = 1, 2, \dots, n$)、および、 $\mathbf{x} \in \mathbb{R}^d$ に対して

$$\begin{aligned} \nabla \sum_{i=1}^n \alpha_i f_i(\mathbf{x}) &= \nabla (\alpha_1 f_1(\mathbf{x}) + \dots + \alpha_n f_n(\mathbf{x})) \\ &= \alpha_1 \nabla f_1(\mathbf{x}) + \dots + \alpha_n \nabla f_n(\mathbf{x}) = \sum_{i=1}^n \alpha_i \nabla f_i(\mathbf{x}) \end{aligned}$$

が成り立ちます。これは、微分作用素 ∇ と和演算 $\sum_{i=1}^n \alpha_i$ の入れ替えが可能であることを意味します。

微分可能な関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ の勾配 $\nabla f: \mathbb{R}^d \rightarrow \mathbb{R}^d$ が連続写像になるとき、 f は連続的微分可能、もしくは、 C^1 級であるといいます。

ヘッセ行列 (2次導関数)

d 次元ユークリッド空間 \mathbb{R}^d 上で定義される実数値関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ が $\mathbf{x} \in \mathbb{R}^d$ で 2 回微分可能であるとは、ある (f と \mathbf{x} に依存する) 行列 $H \in \mathbb{R}^{d \times d}$ が存在して

$$\lim_{\mathbf{h} \rightarrow \mathbf{0}} \frac{\nabla f(\mathbf{x} + \mathbf{h}) - \nabla f(\mathbf{x}) - H\mathbf{h}}{\|\mathbf{h}\|} = \mathbf{0} \quad (1.19)$$

が成り立つときをいいます。行列 H を $\nabla^2 f(\mathbf{x})$ と書くことにします。点 \mathbf{x} に対応する H の関係は $\mathbb{R}^d \ni \mathbf{x} \mapsto H = \nabla^2 f(\mathbf{x}) \in \mathbb{R}^{d \times d}$ で定義される写像として表現することができます。 $\nabla^2 f(\mathbf{x})$ を f の \mathbf{x} における 2 次微分係数と呼び、写像 $\nabla^2 f: \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ を f の 2 次導関数と呼びます。 f の定義域の各点で 2 回微分可能なとき、 f は 2 回微分可能であるといえます。

2 回微分可能な関数は 2 回偏微分可能、つまり、2 次偏微分係数

$$\frac{\partial}{\partial x_j} \left(\frac{\partial f(\mathbf{x})}{\partial x_i} \right) = \begin{cases} \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i} & (i \neq j) \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_i^2} & (i = j) \end{cases}$$

が存在します。このとき、 d^2 個の 2 次偏微分係数を成分にもつ $d \times d$ 行列

$$\begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d^2} \end{pmatrix} \quad (1.20)$$

を f の \mathbf{x} における **ヘッセ行列** と呼びます。2 回微分可能な関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ の 2 次導関数 $\nabla^2 f: \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ が連続写像になるとき、 f は **2 回連続的** 微分可能、もしくは、 **C^2 級** であるといえます。

2 回連続的微分可能な関数のヘッセ行列の性質を以下でまとめます。

[ヘッセ行列の重要な性質]

1. [ヘッセ行列と2次微分係数は一致]

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d^2} \end{pmatrix}$$

2. [ヘッセ行列は対称]
- $\nabla^2 f(\mathbf{x}) \in \mathbb{S}^d$
- 、つまり、
- $\nabla^2 f(\mathbf{x}) = \nabla^2 f(\mathbf{x})^\top$

3. [テイラーの定理]
- $\forall \mathbf{x} \in \mathbb{R}^d \forall \mathbf{y} \in \mathbb{R}^d \exists \lambda \in (0, 1)$
- :

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \langle \mathbf{y} - \mathbf{x}, \nabla^2 f(\mathbf{z}_\lambda)(\mathbf{y} - \mathbf{x}) \rangle$$

ただし、 $\mathbf{z}_\lambda := \mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})$

4. [勾配の展開式]

$$\nabla f(\mathbf{y}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) dt$$

5. [微分作用素
- ∇^2
- の線形性]

$$\forall \alpha_1 \in \mathbb{R} \forall \alpha_2 \in \mathbb{R} (\nabla^2(\alpha_1 f_1 + \alpha_2 f_2) = \alpha_1 \nabla^2 f_1 + \alpha_2 \nabla^2 f_2)$$

ただし、 f_1 と f_2 は2回微分可能とします。

ヘッセ行列の重要な性質1は、 f の2次微分係数 $\nabla^2 f(\mathbf{x})$ がヘッセ行列と一致することを意味しており、2次微分係数 $\nabla^2 f(\mathbf{x})$ は f の2回偏微分係数 $\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i}$ を規則正しく並べることで表現ができることを示しています。ヘッセ行列の重要な性質2により、ヘッセ行列の第 (i, j) 成分と第 (j, i) 成分が一致する、つまり、 $\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$ を満たすので、(1.20)で定義されるヘッセ行列(下式左辺)は

$$\begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d^2} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_d^2} \end{pmatrix}$$

のように、変数 x_i と x_j の入れ替えをしても差し支えがないこととなります。ヘッセ行列の重要な性質 3 は、経験損失の最小解の最適性条件の議論で活躍します (第 3-3 節)。ヘッセ行列の重要な性質 3 は関数 f の展開式であるのに対して、ヘッセ行列の重要な性質 4 は勾配 ∇f の展開式です。ヘッセ行列の重要な性質 4 にある記号 $\int \mathbf{x}(t)dt$ の定義は、 $\mathbf{x}(t) = (x_1(t), \dots, x_i(t), \dots, x_d(t))^T \in \mathbb{R}^d$ に対して

$$\int \mathbf{x}(t)dt := \left(\int x_1(t)dt, \dots, \int x_i(t)dt, \dots, \int x_d(t)dt \right)^T$$

です。ヘッセ行列の重要な性質 4 は、ヘッセ行列と勾配について考察するときによく利用されます (第 14-2 節)。ノルムによる積分評価

$$\left\| \int \mathbf{x}(t)dt \right\| \leq \int \|\mathbf{x}(t)\| dt \quad (1.21)$$

は最適化手法の解析で重宝されます。

● 平滑

C^1 級の関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ が**平滑**であるとは、 f の勾配 $\nabla f: \mathbb{R}^d \rightarrow \mathbb{R}^d$ が以下のような特別な連続性を満たすときをいいます。

$$\exists L > 0 \forall \mathbf{x} \in \mathbb{R}^d \forall \mathbf{y} \in \mathbb{R}^d (\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|) \quad (1.22)$$

(1.22) を満たす ∇f は**リプシッツ連続**であるといい、(1.22) にある正数 L を ∇f の**リプシッツ定数**といいます。リプシッツ定数を明確に示すとき、 f は L -平滑、もしくは、 ∇f は L -リプシッツ連続、のように書きます。以上のことから、 f が L -平滑であることと ∇f が L -リプシッツ連続であることは同値です。

平滑関数の性質を以下でまとめます。

[平滑関数の重要な性質]

1. [降下補題] $f: \mathbb{R}^d \rightarrow \mathbb{R}$ が L -平滑 \Rightarrow

$$\forall \mathbf{x} \forall \mathbf{y} \quad (f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2)$$

2. [平滑関数とヘッセ行列の関係] C^2 級関数 $f: \mathbb{R}^d \rightarrow \mathbb{R}$ が L -平滑 $\Leftrightarrow \forall \mathbf{x} \quad (\|\nabla^2 f(\mathbf{x})\| \leq L)$

平滑関数の重要な性質 2 から見てみます。 L -平滑性の定義 (1.22) から、固定した \mathbf{x} と異なる適当な \mathbf{y} に対して

$$\frac{\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|} \leq L$$

が成り立ちます。これは、 \mathbf{x} から \mathbf{y} への変化量に対する $\nabla f(\mathbf{x})$ から $\nabla f(\mathbf{y})$ への変化量の変化率 $\frac{\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|}$ が上からある正数 L で抑えられることを表しています。 $\mathbf{y} \rightarrow \mathbf{x}$ とすると変化率 $\frac{\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|}{\|\mathbf{x} - \mathbf{y}\|}$ が $\nabla^2 f(\mathbf{x})$ のノルムに近づき ((1.19) から直観的に理解ができるでしょう)、結果として、 $\nabla^2 f(\mathbf{x})$ のノルムが正数 L で抑えられます (正確な結果は平滑関数の重要な性質 2 を参照)。例えば、単純な 2 変数関数 $f(\mathbf{x}) = \frac{1}{2}(x_1^2 + x_2^2)$ は $\nabla f(\mathbf{x}) = (x_1, x_2)^\top$ 、 $\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ から、1-平滑関数です。 $f(\mathbf{x}) = \frac{1}{2}(x_1^2 + x_2^2)$ を図示してもわかるように、お椀型の滑らかな形状をもつことがわかります (図 4-1 参照)。

平滑関数の重要な性質 1 は、平滑関数が降下補題と呼ばれる不等式を満たすことを表しています。降下補題は関数値を下げるためのステップサイズの設定において重要な役割を演じます (第 5-2 節)。

凸関数

$f: \mathbb{R}^d \rightarrow \mathbb{R}$ が凸であるとは

$$\forall \lambda \in [0, 1] \forall \mathbf{x} \in \mathbb{R}^d \forall \mathbf{y} \in \mathbb{R}^d$$

$$(f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})) \quad (1.23)$$

が成り立つときをいいます。 f が狭義凸であるとは

$$\forall \lambda \in (0, 1) \forall \mathbf{x} \in \mathbb{R}^d \forall \mathbf{y} \in \mathbb{R}^d$$

$$(\mathbf{x} \neq \mathbf{y} \Rightarrow f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})) \quad (1.24)$$

が成り立つときをいいます。狭義凸関数は凸関数です（狭義凸関数のクラスは凸関数のクラスよりも強いです）。

微分可能性（勾配）と凸性の関係は以下のとおりです。

[微分可能凸関数の重要な性質]

1. [凸性の必要十分条件] $f: \mathbb{R}^d \rightarrow \mathbb{R}$ が凸
 $\Leftrightarrow \forall \mathbf{x} \forall \mathbf{y} (f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)$
 $\Leftrightarrow \forall \mathbf{x} \forall \mathbf{y} (\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0)$
2. [狭義凸性の必要十分条件] f が狭義凸
 $\Leftrightarrow \forall \mathbf{x} \forall \mathbf{y} (\mathbf{x} \neq \mathbf{y} \Rightarrow f(\mathbf{y}) > f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)$
 $\Leftrightarrow \forall \mathbf{x} \forall \mathbf{y} (\mathbf{x} \neq \mathbf{y} \Rightarrow \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle > 0)$

2回連続的微分可能性（ヘッセ行列）と凸性の関係は以下のとおりです。

[2回連続的微分可能凸関数の重要な性質]

1. [凸性の必要十分条件] $f: \mathbb{R}^d \rightarrow \mathbb{R}$ が凸 $\Leftrightarrow \forall \mathbf{x} (\nabla^2 f(\mathbf{x}) \in \mathbb{S}_+^d)$
2. [狭義凸性の十分条件] $\forall \mathbf{x} (\nabla^2 f(\mathbf{x}) \in \mathbb{S}_{++}^d) \Rightarrow f$ が狭義凸

狭義凸性の十分条件の逆命題 “ f が狭義凸 $\Rightarrow \forall \mathbf{x} (\nabla^2 f(\mathbf{x}) \in \mathbb{S}_{++}^d)$ ” は反例により成り立たないことが証明できます。 $f(x) = x^4$ は $f'(x) = 4x^3$

と $f''(x) = 12x^2$ を満たします。 $x_1 \neq x_2$ となる $x_1, x_2 \in \mathbb{R}$ に対して、 $(f'(x_2) - f'(x_1))(x_2 - x_1) = 4(x_2 - x_1)^2(x_2^2 + x_1x_2 + x_1^2) > 0$ により、 $f(x) = x^4$ は狭義凸となります（微分可能凸関数の重要な性質 2 参照）。一方で、 $f''(0) = 12 \cdot 0^2 = 0 \notin \mathbb{S}_{++}^1$ となるような $x = 0$ が存在します。よって、狭義凸性の十分条件の逆命題は成立しません。

2 回連続的微分可能凸関数の重要な性質は、2 回連続的微分可能な関数とその最小解の周りで凸関数になることを考察するうえで非常に役立つ性質です（第 3-3 節）。2 回連続的微分可能凸関数の重要な性質により、対象の関数が凸性を有するかどうかの判定はその関数のヘッセ行列の固有値の正負に関係していることがわかります。（半）正定値行列（固有値が非負（正）となる対称行列）の性質については、第 1-3 節を参照して下さい。

1-5 確率・統計

確率空間

対象のデータを要素としてもつ集合 Ω から無作為（ランダム）にデータを選ぶことを考えましょう。各データのことを**標本（サンプル）**といい、標本全体の集合 Ω を標本空間といいます。無作為にデータを選ぶという事柄を**事象**といい、事象の集まりを \mathcal{F} と書くことにします。事象（データが選ばれること）の起こりやすさを定量的に表すのが**確率** P です。一般に、標本空間、事象、確率の三組 (Ω, \mathcal{F}, P) が以下の三条件を満たすとき、**確率空間**といいます。

- (I) 標本空間 Ω は空でない集合です。
- (II) 事象の集合 $\mathcal{F} \subset 2^\Omega := \{A: A \subset \Omega\}$ は次の三性質を満たします。
 - (1) $\Omega \in \mathcal{F}$; (2) $\forall A \in \mathcal{F} (A^c \in \mathcal{F})$; (3) $\forall (A_i) \subset \mathcal{F} (\bigcup_{i=1}^{+\infty} A_i \in \mathcal{F})$
- (III) 確率 $P: \mathcal{F} \rightarrow [0, 1]$ は次の二性質を満たします。
 - (1) $\forall (A_i) \subset \mathcal{F} (\forall i \forall j (i \neq j \Rightarrow A_i \cap A_j = \emptyset) \Rightarrow P(\bigcup_{i=1}^{+\infty} A_i) = \sum_{i=1}^{+\infty} P(A_i))$; (2) $P(\Omega) = 1$

(III)(1) の仮定 ($i \neq j \Rightarrow A_i \cap A_j = \emptyset$) は事象 A_i と異なる事象 A_j は同時に起こらないこと、つまり、 A_i と A_j は排反であることを表します。例えば、 n 個のデータ集合 $\Omega := [n] = \{1, 2, \dots, n\}$ から無作為にデータの一つを選ぶことを考え、事象 A_i を $A_i := \{i\}$ 、つまり、「選ばれた一つのデータはデータ i である」と定義します（標本 i からなる事象 $\{i\}$ を根元事象といいます）。このとき、考えられる事象は A_1, \dots, A_n の n 個であり、また、 A_i と A_j は同時に起こることはないので、 A_i と A_j は排反 ($A_i \cap A_j = \emptyset$) です。定義 (III)(1) は、「排反事象 A_1, \dots, A_n のいずれかが起こる事象」 $\bigcup_{i=1}^n A_i$ の確率は「各事象 A_i が起こる確率の和」であることを意味します。いま、 $\Omega = \bigcup_{i=1}^n A_i$ なので、定義 (III)(2) か

ら、 A_1, \dots, A_n のいずれかが必ず起こります (確率 1)。ある事象 A が確率 1 で起きるとき、 A がほとんど至るところで (almost surely ; a.s.) 起きるといい、 A a.s. と書くことにします。

確率変数

標本空間を n 個の要素をもつ集合 $\Omega := \{\omega_1, \dots, \omega_i, \dots, \omega_n\}$ とし、確率空間を (Ω, \mathcal{F}, P) とします。関数 $X: \Omega \rightarrow \mathbb{R}$ が

$$\forall i \in [n] \exists x_i \in \mathbb{R} (X(\omega_i) = x_i)$$

を満たす、つまり、根元事象 $\{\omega_i\}$ を実数 $x_i = X(\omega_i)$ のように数値化できるとします。事象 $\{\omega \in \Omega: X(\omega) = x\}$ を、単に、 $X = x$ と書くことにします。このとき、 X が確率空間 (Ω, \mathcal{F}, P) の**離散型確率変数**であるとは

$$\forall i \in [n] (f(x_i) := P(X = x_i) \geq 0) \wedge \sum_{i=1}^n f(x_i) = 1 \quad (1.25)$$

を満たすときをいいます。これは、事象 $X = x_i$ (根元事象 $\{\omega_i\}$) が起こる確率 $P(X = x_i)$ は零以上で、排反事象 $X = x_1, \dots, X = x_n$ のいずれかが必ず起きること (確率 1) を表しており、確率の定義 (III) と合点がいきます。(1.25) で定義される関数 f を X の**離散型確率分布**といいます。

機械学習の分野では、以下で定義される**離散一様分布** (Discrete Uniform distribution) に基づいたデータの標本 (サンプル) がよく利用されます。

$$\forall i \in [n] \left(f(x_i) = P(X = x_i) = \frac{1}{n} \right) \quad (1.26)$$

離散一様分布に基づいたデータは特定のデータに偏ることなく等確率 $\frac{1}{n}$ で選ばれることを意味します。(1.26) を満たす離散型確率変数 X は離散一様分布 $DU(n)$ に従うといい、 $X \sim DU(n)$ と書きます。

(1.25) で定義される (1次元) 離散型確率変数を d 個利用することで

$$\mathbf{X} := (X_1, \dots, X_i, \dots, X_d)^\top$$

ただし、 X_i は離散型確率変数、つまり、 X_i の離散型確率分布は

$$\forall j \in [n] (f_i(x_j) := P(X_i = x_j) \geq 0) \wedge \sum_{j=1}^n f_i(x_j) = 1$$

を定義することができます。 \mathbf{X} が確率空間 (Ω, \mathcal{F}, P) の **d 次元離散型確率変数** であるとは、 $\mathbf{x} = (x_{j_1}, \dots, x_{j_i}, \dots, x_{j_d})^\top$ に対して

$$\begin{aligned} f(\mathbf{x}) &:= P(\mathbf{X} = \mathbf{x}) = P\left(\bigcap_{i=1}^d X_i = x_{j_i}\right) \geq 0 \\ \sum_{j_1=1}^n \cdots \sum_{j_d=1}^n P\left(\bigcap_{i=1}^d X_i = x_{j_i}\right) &= 1 \end{aligned} \tag{1.27}$$

を満たすときをいいます。ただし、 $\bigcap_{i=1}^d X_i = x_{j_i}$ は d 個の事象 $X_1 = x_{j_1}, \dots, X_i = x_{j_i}, \dots, X_d = x_{j_d}$ が同時に起こる事象を表します。(1.27) で定義される関数 f を \mathbf{X} の**離散型同時確率分布**とといいます。

(1.26) で定義される離散一様分布 $DU(n)$ に従う離散型確率変数 $X_i \sim DU(n)$ からなる **d 次元離散一様分布** $DU_d(n)$ を定義することができます。 $\mathbf{X} \sim DU_d(n)$ の離散型同時確率分布は以下のようになります。

$$f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{1}{n^d} \tag{1.28}$$

期待値・分散

確率空間 (Ω, \mathcal{F}, P) の d 次元離散型確率変数 $\mathbf{X} = (X_1, \dots, X_i, \dots, X_d)^\top$ の期待値 (平均) は

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] := \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_i] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^n x_j P(X_1 = x_j) \\ \vdots \\ \sum_{j=1}^n x_j P(X_i = x_j) \\ \vdots \\ \sum_{j=1}^n x_j P(X_d = x_j) \end{pmatrix} \quad (1.29)$$

として定義されます。 X_i の期待値 $\mathbb{E}[X_i] := \sum_{j=1}^n x_j P(X_i = x_j)$ は確率変数 X_i が取りうる値 x_1, x_2, \dots, x_n の確率重み付き平均として定義されます。複数の変数が混在するとき確率変数 X に関する期待値であることを明記する必要がある場合は、 $\mathbb{E}_X[\cdot]$ と書くことにします。 d 次元離散型確率変数 \mathbf{X}, \mathbf{Y} に対して

$$\mathbb{E}_{(\mathbf{X}, \mathbf{Y})}[\mathbf{X} + \mathbf{Y}] := (\mathbb{E}_{(X_1, Y_1)}[X_1 + Y_1], \dots, \mathbb{E}_{(X_d, Y_d)}[X_d + Y_d])^\top$$

のように定義されます。ただし、 $\mathbb{E}_{(X, Y)}[X + Y]$ は以下のとおりです。

$$\mathbb{E}_{(X, Y)}[X + Y] := \sum_{i=1}^n \sum_{j=1}^n (x_i + y_j) P(X = x_i \cap Y = y_j)$$

[期待値の重要な性質]

1. 離散型確率変数 X, Y と $c \in \mathbb{R}$ に対して

$$(1) \mathbb{E}_{(X,Y)}[X + Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y]$$

$$(2) \mathbb{E}_X[X + c] = \mathbb{E}_X[X] + c, \mathbb{E}_X[cX] = c\mathbb{E}_X[X]$$

2. d 次元離散型確率変数 \mathbf{X}, \mathbf{Y} と $\mathbf{c} \in \mathbb{R}^d$ に対して

$$(1) \mathbb{E}_{(\mathbf{X}, \mathbf{Y})}[\mathbf{X} + \mathbf{Y}] = \mathbb{E}_{\mathbf{X}}[\mathbf{X}] + \mathbb{E}_{\mathbf{Y}}[\mathbf{Y}]$$

$$(2) \mathbb{E}_{\mathbf{X}}[\mathbf{X} + \mathbf{c}] = \mathbb{E}_{\mathbf{X}}[\mathbf{X}] + \mathbf{c}, \mathbb{E}_{\mathbf{X}}[\langle \mathbf{c}, \mathbf{X} \rangle] = \langle \mathbf{c}, \mathbb{E}_{\mathbf{X}}[\mathbf{X}] \rangle$$

確率空間 (Ω, \mathcal{F}, P) の d 次元離散型確率変数を \mathbf{X} と \mathbf{Y} とします。事象 $\mathbf{Y} = \mathbf{y}$ が起きたときの \mathbf{X} の期待値を $\mathbf{Y} = \mathbf{y}$ を条件とする \mathbf{X} の条件付き期待値といい、 $\mathbb{E}_{\mathbf{X}}[\mathbf{X}|\mathbf{y}]$ と書きます。

一方で、 \mathbf{X} と \mathbf{Y} が独立であるとは

$$P(\mathbf{X} = \mathbf{x} \cap \mathbf{Y} = \mathbf{y}) = P(\mathbf{X} = \mathbf{x})P(\mathbf{Y} = \mathbf{y})$$

のように、 (\mathbf{X}, \mathbf{Y}) の同時確率分布 $P(\mathbf{X} = \mathbf{x} \cap \mathbf{Y} = \mathbf{y})$ が \mathbf{X} と \mathbf{Y} の確率分布の積になるときをいいます。 \mathbf{X} と \mathbf{Y} が独立のとき、 \mathbf{X} の期待値 $\mathbb{E}_{\mathbf{X}}[\mathbf{X}]$ は事象 $\mathbf{Y} = \mathbf{y}$ に関係なく定義できます。つまり

$$\mathbf{X} \text{ と } \mathbf{Y} \text{ が独立} \Rightarrow \mathbb{E}_{\mathbf{X}}[\mathbf{X}|\mathbf{y}] = \mathbb{E}_{\mathbf{X}}[\mathbf{X}] \quad (1.30)$$

が成り立ちます。

n 個の互いに独立な d 次元確率変数 $\mathbf{X}_1, \dots, \mathbf{X}_n$ からなる確率変数 $\mathbf{X} = \mathbf{X}(\mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_n)$ の期待値を定義しましょう。ただし、 $\mathbf{X}_i = \mathbf{X}_i(\mathbf{X}_1, \dots, \mathbf{X}_{i-1})$ は $[\mathbf{X}_{i-1}] := \{\mathbf{X}_1, \dots, \mathbf{X}_{i-1}\}$ を用いて定義できるものとします。つまり、 \mathbf{X}_i は $[\mathbf{X}_{i-1}]$ が得られたときに定義できます。このとき、条件 $[\mathbf{X}_{n-1}]$ のもとでの \mathbf{X}_n に関する \mathbf{X} の条件付き期待値

$$\mathbb{E}_{\mathbf{X}_n}[\mathbf{X} | [\mathbf{X}_{n-1}]] \stackrel{(1.30)}{=} \mathbb{E}_{\mathbf{X}_n}[\mathbf{X}]$$

は $[\mathbf{X}_{n-1}] = \{\mathbf{X}_1, \dots, \mathbf{X}_{n-1}\}$ の関数になることに注意します。 \mathbf{X}_{n-1}

に関する $\mathbb{E}_{\mathbf{X}_n}[\mathbf{X}]$ の期待値は

$$\mathbb{E}_{\mathbf{X}_{n-1}}[\mathbb{E}_{\mathbf{X}_n}[\mathbf{X}][\mathbf{X}_{n-2}]] \stackrel{(1.30)}{=} \mathbb{E}_{\mathbf{X}_{n-1}}[\mathbb{E}_{\mathbf{X}_n}[\mathbf{X}]] = \mathbb{E}_{\mathbf{X}_{n-1}}\mathbb{E}_{\mathbf{X}_n}[\mathbf{X}]$$

となります。これを繰り返す行うことで

$$\mathbb{E}[\mathbf{X}] := \mathbb{E}_{\mathbf{X}_1} \cdots \mathbb{E}_{\mathbf{X}_i} \cdots \mathbb{E}_{\mathbf{X}_n}[\mathbf{X}] \quad (1.31)$$

が定義できます。 $\mathbb{E} := \mathbb{E}_{\mathbf{X}_1} \cdots \mathbb{E}_{\mathbf{X}_i} \cdots \mathbb{E}_{\mathbf{X}_n}$ を確率変数 $\mathbf{X}_1, \dots, \mathbf{X}_n$ に関する **全期待値** といいます。

\mathbf{X} の**分散**とは、期待値（中心） $\boldsymbol{\mu} = \mathbb{E}_{\mathbf{X}}[\mathbf{X}]$ からの \mathbf{X} のばらつき度合いのことで、 \mathbf{X} と $\boldsymbol{\mu}$ の2乗ノルム $\|\mathbf{X} - \boldsymbol{\mu}\|^2$ の期待値（平均）

$$\mathbb{V}[\mathbf{X}] := \mathbb{E} \left[\|\mathbf{X} - \boldsymbol{\mu}\|^2 \right] = \mathbb{E} \left[\sum_{i=1}^d (X_i - \mu_i)^2 \right] \quad (1.32)$$

として定義されます。複数の変数が混在するとき確率変数 X に関する分散であることを明記する必要がある場合は、 $\mathbb{V}_X[\cdot]$ と書くことにします。 d 次元離散型確率変数 \mathbf{X}, \mathbf{Y} に対して

$$\mathbb{V}_{(\mathbf{X}, \mathbf{Y})}[\mathbf{X} + \mathbf{Y}] := \mathbb{E}_{(\mathbf{X}, \mathbf{Y})} \left[\left\| (\mathbf{X} + \mathbf{Y}) - \mathbb{E}_{(\mathbf{X}, \mathbf{Y})}[\mathbf{X} + \mathbf{Y}] \right\|^2 \right]$$

のように定義されます。

[分散の重要な性質]

1. 離散型確率変数 X, Y と、 $c \in \mathbb{R}$ に対して

$$(1) \quad \mathbb{V}_X[X + c] = \mathbb{V}_X[X], \quad \mathbb{V}_X[cX] = c^2\mathbb{V}_X[X]$$

$$(2) \quad X \text{ と } Y \text{ が独立} \Rightarrow \mathbb{V}_{(X, Y)}[X + Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y]$$

2. d 次元離散型確率変数 \mathbf{X}, \mathbf{Y} と $\mathbf{c} \in \mathbb{R}^d$ に対して

$$(1) \quad \mathbb{V}_{\mathbf{X}}[\mathbf{X}] = \mathbb{E}_{\mathbf{X}}[\|\mathbf{X}\|^2] - \|\mathbb{E}_{\mathbf{X}}[\mathbf{X}]\|^2,$$

$$\mathbb{V}_{\mathbf{X}}[\mathbf{X} + \mathbf{c}] = \mathbb{V}_{\mathbf{X}}[\mathbf{X}]$$

$$(2) \quad \mathbf{X} \text{ と } \mathbf{Y} \text{ が独立} \Rightarrow \mathbb{V}_{(\mathbf{X}, \mathbf{Y})}[\mathbf{X} + \mathbf{Y}] = \mathbb{V}_{\mathbf{X}}[\mathbf{X}] + \mathbb{V}_{\mathbf{Y}}[\mathbf{Y}]$$

確率空間 (Ω, \mathcal{F}, P) の d 次元離散型確率変数を \mathbf{X} と \mathbf{Y} とします。事象 $\mathbf{Y} = \mathbf{y}$ が起きたときの \mathbf{X} の分散を $\mathbf{Y} = \mathbf{y}$ を条件とする \mathbf{X} の**条件**

付き分散といい、 $\mathbb{V}_{\mathbf{X}}[\mathbf{X}|\mathbf{y}]$ と書きます。分散の定義と (1.30) から、以下が成り立ちます。

$$\mathbf{X} \text{ と } \mathbf{Y} \text{ が独立} \Rightarrow \mathbb{V}_{\mathbf{X}}[\mathbf{X}|\mathbf{y}] = \mathbb{V}_{\mathbf{X}}[\mathbf{X}] \quad (1.33)$$

🔵 標本調査と標本平均

本書では、標本空間 Ω の要素が多いことを対象とします (第 7 章参照)。例えば、飛行機、自動車などのカラー画像のデータセット CIFAR (Canadian Institute for Advanced Research) は 50,000 枚の訓練データと 10,000 枚のテストデータで構成されています (図 2-1 参照)。この例のような性質を知りたいデータ全体の集合を**母集団**といいます。母集団の要素をしらみつぶしに調べることができれば、母集団の性質を知ることができるでしょう。しかしながら、母集団の要素数がとても多いとき、これらをすべて調べることは易しくはありません。

標本 (母集団の一部) を取り出すことを**標本調査 (サンプリング)**とといいます。標本調査により得られる標本の性質を解明することで母集団の性質を知ろうと試みます。標本の値や性質は標本調査前には知ることができないので、標本は確率変数として扱います。標本 (確率変数) の期待値や分散を求めることで**母集団分布** (母集団の確率分布) の特性、特に、**母平均** (母集団の平均) と**母分散** (母集団の分散) を調査することができます。

母集団から標本調査をするとき、それが意図的では意味がないので無作為 (ランダム) に抽出される**無作為標本** (ランダムサンプル) を利用します。無作為標本は**独立同一分布** (independent and identically distributed ; i.i.d.)、つまり、独立で、かつ、母集団分布に従うことになります。

母集団の母平均と母分散を、それぞれ、 μ 、 σ^2 とします。母集団から抽出された大きさ b の無作為標本を $\mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_b$ とします。このとき、無作為標本 \mathbf{X}_i の平均 $\mathbb{E}_{\mathbf{X}_i}[\mathbf{X}_i]$ と分散 $\mathbb{V}_{\mathbf{X}_i}[\mathbf{X}_i]$ は、それぞれ、 μ 、 σ^2 となります。母集団の母平均 μ と母分散 σ^2 を求めるために

$$\bar{\mathbf{X}}_b := \frac{1}{b} \sum_{i=1}^b \mathbf{X}_i$$

で定義される**標本平均**の期待値と分散を調べます。

初めに、標本平均の期待値を調べてみましょう。[期待値の重要な性質 1(2), 2(1)] から

$$\mathbb{E}_{(\mathbf{X}_1, \dots, \mathbf{X}_b)} [\bar{\mathbf{X}}_b] = \mathbb{E}_{(\mathbf{X}_1, \dots, \mathbf{X}_b)} \left[\frac{1}{b} \sum_{i=1}^b \mathbf{X}_i \right] = \frac{1}{b} \sum_{i=1}^b \underbrace{\mathbb{E}_{\mathbf{X}_i} [\mathbf{X}_i]}_{\boldsymbol{\mu}} = \boldsymbol{\mu}$$

を満たすので、標本平均の期待値を調べることで母集団の平均を推定することができます。このとき、標本平均 $\bar{\mathbf{X}}_b$ は母集団の平均 $\boldsymbol{\mu}$ に対する**不偏推定量**であるといいます。不偏推定量とは標本の大きさ b の大小に関係なくその期待値が偏っていない（母集団の平均と一致する）推定量といえます。

次に、標本平均の分散を調べてみましょう。[分散の重要な性質 1(1), 2(2)] から

$$\mathbb{V}_{(\mathbf{X}_1, \dots, \mathbf{X}_b)} [\bar{\mathbf{X}}_b] = \mathbb{V}_{(\mathbf{X}_1, \dots, \mathbf{X}_b)} \left[\frac{1}{b} \sum_{i=1}^b \mathbf{X}_i \right] = \frac{1}{b^2} \sum_{i=1}^b \underbrace{\mathbb{V}_{\mathbf{X}_i} [\mathbf{X}_i]}_{\sigma^2} = \frac{\sigma^2}{b}$$

を満たすので、標本平均 $\bar{\mathbf{X}}_b$ は母集団の分散 σ^2 に対する不偏推定量とはなりません。しかしながら、この等式から、標本の大きさ b を大きくすることで標本平均 $\bar{\mathbf{X}}_b$ の散らばり度合い（分散）を抑えることができます（第7-4節参照）。

確率変数の収束性

確率空間 (Ω, \mathcal{F}, P) の d 次元確率変数の点列を (\mathbf{X}_t) とし、 $\mathbf{X}^* \in \mathbb{R}^d$ とします。点列 (\mathbf{X}_t) が \mathbf{X}^* に**概収束** (almost sure convergence) するとは、 (\mathbf{X}_t) が \mathbf{X}^* にノルムの意味で収束するような事象の確率が1のとき、つまり

$$P \left(\left\{ \omega \in \Omega : \lim_{t \rightarrow +\infty} \|\mathbf{X}_t(\omega) - \mathbf{X}^*\| = 0 \right\} \right) = 1$$

を満たすときをいい、 $\lim_{t \rightarrow +\infty} \mathbf{X}_t = \mathbf{X}^*$ a.s. または $\mathbf{X}_t \xrightarrow{\text{a.s.}} \mathbf{X}^*$ ($t \rightarrow +\infty$)、単に、 $\mathbf{X}_t \xrightarrow{\text{a.s.}} \mathbf{X}^*$ と書きます。概収束性に関連する以下の**優マルチンゲール収束定理** (The supermartingale convergence theorem) を紹介します。

[優マルチンゲール収束定理]

$(Y_t), (Z_t), (W_t)$ を非負確率変数の点列とし、 Y_t, Z_t, W_t は d 次元確率変数の集合 $[\mathbf{X}_t] := \{\mathbf{X}_1, \dots, \mathbf{X}_t\}$ に依存して定義されるものとします。さらに

$$\forall t \in \mathbb{N} \left(\mathbb{E}_{\mathbf{X}_{t+1}}[Y_{t+1} | [\mathbf{X}_t]] \leq Y_t - Z_t + W_t \right) \wedge \sum_{t=1}^{+\infty} W_t < +\infty \text{ a.s.}$$

が成り立つとします。このとき、 (Y_t) は概収束し、以下が成り立ちます。

$$\sum_{t=1}^{+\infty} Z_t < +\infty \text{ a.s.}$$

第2章

機械学習モデルを訓練する

機械学習モデルは、入力された訓練データを解析して、その予測値を出力します。この予測値と訓練データの正解値の誤差の大きさ（損失）をできるだけ小さくすることができる機械学習モデルは、訓練データを適切に識別できる理想の学習モデルです。

本章では、具体的な機械学習モデルを通して、全訓練データに関する損失の平均（経験損失）を陽に示します。さらに、最適化と呼ばれる最適化手法が経験損失を最小化することを、全訓練データを用いて機械学習モデルを訓練することとして定式化します。最後に、経験損失の形状や数学的性質についても考察します。

2-1 機械学習

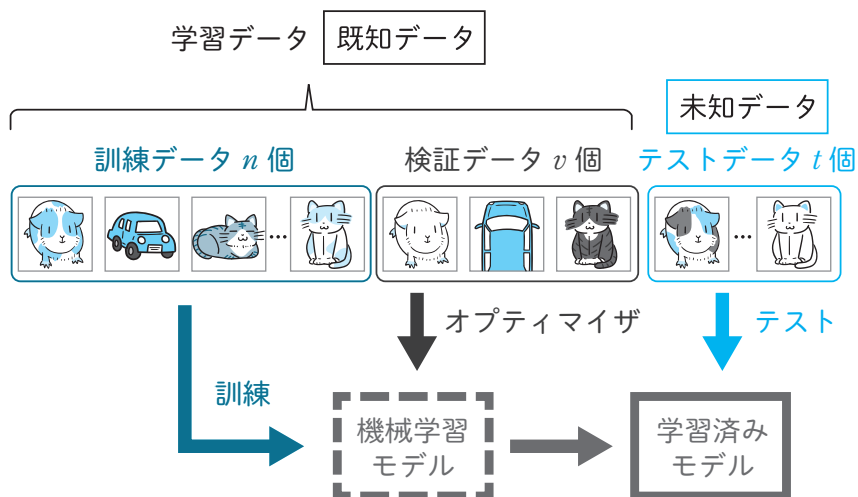
機械学習とは、人工知能（AI, Artificial Intelligence）開発のための手法の一つです。機械学習は、与えられたデータからルールやパターンを学習し、それらを用いて未知のデータに対する予測や判断を行うことができます。

① 学習データ（訓練データと検証データ）とテストデータ

機械学習をするために事前に与えられたデータを**学習データ**と呼ぶことにします。学習データを**訓練データ**と**検証データ**に分けます。

図 2-1 では、“モルモット”、“車”、“猫”の画像からなる学習データを例にしています。訓練データは、機械（機械学習モデル）を訓練するために使用するデータのことを指し、検証データは、機械学習モデルを訓練する最適化手法（オプティマイザ）の調整（ステップサイズなどのハ

図 2-1 学習データとテストデータを用いた機械学習



イパーパラメータチューニング) に使用します (詳細については、第 5-2 節を参照)。検証データによって調整されたオプティマイザが訓練データを用いて機械学習モデルを訓練します。図 2-1 を例にすると、オプティマイザが機械学習モデルを訓練するとは、“モルモット”の画像を高い確率で“モルモット”であると識別できるように、“車”の画像を高い確率で“車”であると識別できるように、“猫”の画像を高い確率で“猫”と識別できるように、繰り返し、繰り返し、学習することです。訓練後の機械学習モデルを**学習済みモデル**と呼ぶことにします。一度も使用していないデータに対してもこの学習済みモデルが適切に予測ができるかどうかを検査するために、**テストデータ**を利用します。テストデータは誰も知り得ない未知のデータであり、学習済みモデルの検査のときに初めて利用します。

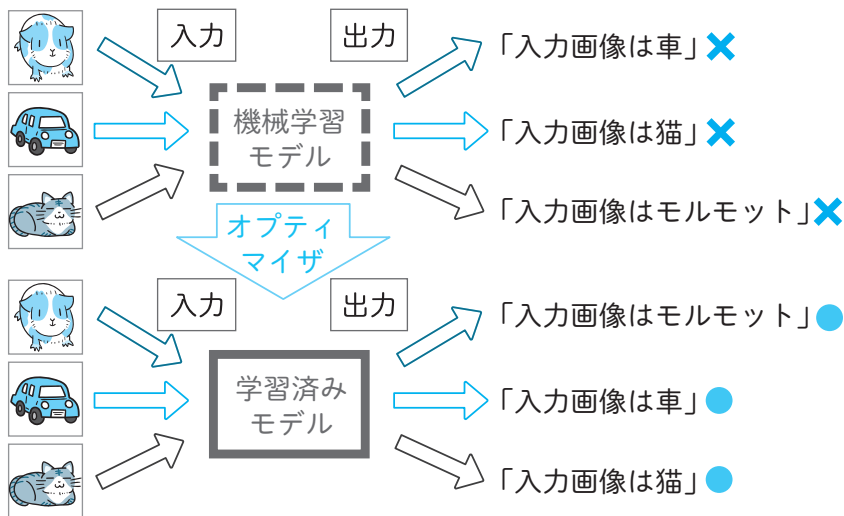
● 機械学習モデルとオプティマイザ

機械学習モデルとは、入力された訓練データを解析して結果を出力する仕組みのことです。例えば、図 2-2 のような訓練データを例にすると、訓練がされていない機械学習モデルは入力画像を適切に識別することができません。

しかしながら、**オプティマイザ**が訓練することで得られる学習済みモデルは入力画像を識別することができます。

それでは、**オプティマイザ**がどのようにして機械学習モデルを訓練するのかについて、次節で詳しく見ていきましょう。

図 2-2 機械学習モデルとオプティマイザ



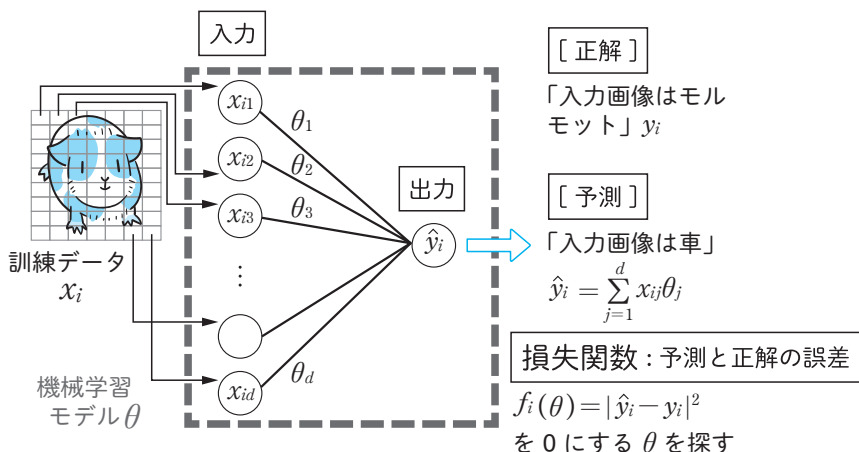
2-2 機械学習モデルを訓練するには？

ニューラルネットワーク（機械学習モデル）

ノード（図 2-3 の円形のこと）内に数値を代入し、エッジ（図 2-3 の円形と円形を繋ぐ線のこと）に重み（図 2-3 の θ_1 から θ_d のこと）が割り当てられているグラフをニューラルネットワークと呼びます。ニューラルネットワークは機械学習モデルの一つであり、ノードやエッジの数を多数にすることで大規模で複雑な機械学習モデルを生成することができます。

入力された訓練データを解析して結果を出力する機械学習モデルの構造を図 2-3 を例にして見ていきます。訓練データの画像“モルモット”をニューラルネットワークに入力するために、画像をピクセルで分割し、各ピクセルの色の度合いで数値化します。訓練データの総数を n とし、ピクセルの総数を d とすると、 i 番目 ($i = 1, 2, \dots, n$) の訓練データは入力ベクトル $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$ として表現ができます（例

図 2-3 機械学習モデル（ニューラルネットワーク）と損失関数



例えば、白黒画像を 28×28 個のピクセルで分割した場合は、白を 0、黒を 255 として表現することで、 $x_{ij} \in [0, 255]$ からなる $784 (= 28 \times 28)$ 次元のベクトル $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{i784})^\top$ となります。 i 番目の訓練データを \mathbf{x}_i と書くことを認めます。

訓練データ \mathbf{x}_i の第 j 成分 x_{ij} ($j = 1, 2, \dots, d$) はエッジの重み θ_j を掛けて足し込むことで \mathbf{x}_i に対する出力

$$\begin{aligned} \text{[予測値]} \hat{y}_i &= \hat{y}_i(\boldsymbol{\theta}) = x_{i1}\theta_1 + x_{i2}\theta_2 + \dots + x_{id}\theta_d \\ &= (x_{i1}, x_{i2}, \dots, x_{id})(\theta_1, \theta_2, \dots, \theta_d)^\top \quad (2.1) \\ &= \mathbf{x}_i^\top \boldsymbol{\theta} \end{aligned}$$

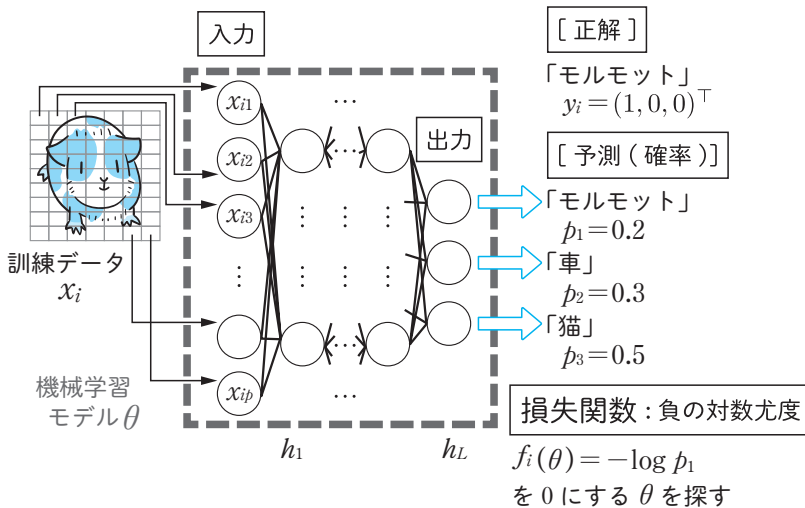
を得ます。ただし、(2.1) 内の三つ目の等号では内積の定義 (1.1) を利用します。このような操作で得られる \hat{y}_i を、訓練データ \mathbf{x}_i に対する機械学習モデル (ニューラルネットワーク) によって出力された**予測値**と呼ぶことにします。予測値は機械学習モデルの**パラメータ** $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)^\top$ の関数になります。つまり、機械学習モデルのパラメータ $\boldsymbol{\theta}$ を変えることで、予測値 $\hat{y}_i = \hat{y}_i(\boldsymbol{\theta})$ を修正することが可能です。機械学習モデルは端的にいえば、パラメータ $\boldsymbol{\theta}$ のことなので、混乱の生じない限り、機械学習モデル (ニューラルネットワーク) を $\boldsymbol{\theta}$ と書くことにします。一方で、訓練データには**正解ラベル** (正解値) が与えられているものとします。図 2-3 では、訓練データの画像が“車”や“猫”ではなく“モルモット” (正解) です。訓練データ \mathbf{x}_i の正解ラベルを \mathbf{y}_i と書くことにします (図 2-3 のように、正解ラベルが実数のときは y_i と書きます)。訓練データ \mathbf{x}_i に正解ラベル \mathbf{y}_i が付与されているとき、訓練データを $(\mathbf{x}_i, \mathbf{y}_i)$ と書くことにします。

図 2-4 のように、図 2-3 の単純なニューラルネットワークを複数繋げることで大規模で複雑な学習モデルを生成することができます。

図 2-3 の予測値 (2.1) $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta} = \boldsymbol{\theta}^\top \mathbf{x}_i$ (いい換えれば、ノードはエッジの重みベクトルと訓練データの内積を計算する) を利用して、図 2-4 の h_1 で得られる結果を考察します。訓練データ $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$ はエッジの重み w を掛けて足し込んだあと (いい換えれば、 $\mathbf{w}^\top \mathbf{x}_i$ を計算したあと) バイアス項 b を足す操作を各ノードで行うことで

$$\underbrace{\begin{pmatrix} u_{11} \\ \vdots \\ u_{1j} \\ \vdots \\ u_{1m} \end{pmatrix}}_{\mathbf{u}_1} \stackrel{(2.1)}{=} \begin{pmatrix} \mathbf{w}_1^\top \mathbf{x}_i \\ \vdots \\ \mathbf{w}_i^\top \mathbf{x}_i \\ \vdots \\ \mathbf{w}_m^\top \mathbf{x}_i \end{pmatrix} + \underbrace{\begin{pmatrix} b_{11} \\ \vdots \\ b_{1j} \\ \vdots \\ b_{1m} \end{pmatrix}}_{\mathbf{b}_1}$$

図 2-4 複雑な機械学習モデル (ニューラルネットワーク) と損失関数



$$= \underbrace{\begin{pmatrix} w_{11} & \cdots & w_{1j} & \cdots & w_{1p} \\ \vdots & & \vdots & & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{ip} \\ \vdots & & \vdots & & \vdots \\ w_{m1} & \cdots & w_{mj} & \cdots & w_{mp} \end{pmatrix}}_{W_1 = (\mathbf{w}_1^\top \cdots \mathbf{w}_i^\top \cdots \mathbf{w}_m^\top)^\top} \underbrace{\begin{pmatrix} x_{i1} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{pmatrix}}_{\mathbf{x}_i} + \mathbf{b}_1$$

を得ます。次に、 \mathbf{u}_i の各成分 u_{ij} をシグモイド関数 $\zeta(x) := (1 + e^{-x})^{-1}$ 、または、正規化線形関数 (rectified linear unit, ReLU) $x^+ := \max\{0, x\}$ の近似関数 $x^{++}(x) := \log(1 + e^x)$ のような非線形関数 s で変換します。以上のことから

$$\mathbf{h}_1 = s(\mathbf{u}_1) = s(W_1 \mathbf{x}_i + \mathbf{b}_1) = \underbrace{s \circ (W_1(\cdot) + \mathbf{b}_1)}_{T_1}(\mathbf{x}_i) \quad (2.2)$$

のように、訓練データ \mathbf{x}_i はエッジの重みを成分にもつ行列 W_1 とバイアス \mathbf{b} による変換を行った後に、非線形写像 s を作用させることで \mathbf{h}_1 を得ることになります。このような操作が \mathbf{h}_2 以降でも行われ、結果として、以下のような出力が得られます。

$$\mathbf{x}_i \mapsto \mathbf{h}_1 = T_1(\mathbf{x}_i) \mapsto \mathbf{h}_2 = T_2(\mathbf{h}_1) \mapsto \cdots \mapsto \mathbf{h}_L = T_L(\mathbf{h}_{L-1}) \quad (2.3)$$

ただし、 $T_i := s \circ (W_i(\cdot) + \mathbf{b}_i)$ ($i = 1, 2, \dots, L$) です。

$\mathbf{h}_L = (h_{L1}, \dots, h_{Lk})^\top$ の結果から、分類数 k (図 2-4 では“モルモット”、“車”、“猫”の 3 分類数) の分類クラス $(C_1, \dots, C_r, \dots, C_k)$ (図 2-4 では $(C_1, C_2, C_3) = (\text{モルモット}, \text{車}, \text{猫})$) に属する確率 p_r を出力として得ます。具体的には、以下で定義されるソフトマックス関数 σ を用いて計算します。

[予測値]

$$(p_1, \dots, p_r, \dots, p_k)^\top = (p_1(\boldsymbol{\theta}), \dots, p_r(\boldsymbol{\theta}), \dots, p_k(\boldsymbol{\theta}))^\top \quad (2.4)$$

$$= \sigma(\mathbf{h}_L) := \left(\frac{e^{h_{L1}}}{\sum_{i=1}^k e^{h_{Li}}}, \dots, \frac{e^{h_{Lr}}}{\sum_{i=1}^k e^{h_{Li}}}, \dots, \frac{e^{h_{Lk}}}{\sum_{i=1}^k e^{h_{Li}}} \right)^\top$$

図 2-4 のような機械学習モデル θ はエッジの重みからなる行列 W の成分とノード数分の成分をもつバイアス \mathbf{b} からなるベクトル

$$\theta = (\text{vec}(W_1), \dots, \text{vec}(W_L), \mathbf{b}_1, \dots, \mathbf{b}_L)^\top \in \mathbb{R}^d \quad (2.5)$$

となります。ただし、 $\text{vec}(W_1)$ は行列 W_1 のすべての成分を縦に並べたベクトルとします。(2.3) から \mathbf{h}_L は $W_1, \dots, W_L, \mathbf{b}_1, \dots, \mathbf{b}_L$ に依存したベクトルなので、予測値 p_r は θ の関数になります。パラメータ θ の次元 d はエッジ数とノード数が多ければ多いほどとてつもない大きさの自然数になることがわかります。例えば、対話型人工知能 ChatGPT (Chat Generative Pre-trained Transformer) のパラメータ総数 d は推定、一兆以上のようなのです。このように、予測や判断を適切に行うことができる学習済みモデルを生成するための機械学習モデルのパラメータ総数 d は大きいことがわかります。

● 損失関数 (予測と正解の誤差)

訓練データの正解ラベルと機械学習モデルによって出力される予測値との誤差の大きさを損失といいます。この損失は機械学習モデル θ の関数である予測値から定まるので、損失も機械学習モデル θ の関数となります。このことから、この損失を機械学習モデルにおける訓練データに関する**損失関数**と呼ぶことにします。

図 2-3 のようなニューラルネットワーク θ における訓練データ (\mathbf{x}_i, y_i) に関する損失関数 $f_i: \mathbb{R}^d \rightarrow \mathbb{R}_+$ は、例えば

$$f_i(\theta) = \begin{cases} |\hat{y}_i(\theta) - y_i| & (\text{微分不可能関数}) \\ |\hat{y}_i(\theta) - y_i|^2 & (\text{微分可能関数}) \end{cases} \quad (2.6)$$

と書くことができます。 $f_i(\theta) = |\hat{y}_i(\theta) - y_i|$ は $\hat{y}_i(\theta) = y_i$ となる θ で微分ができない関数 ($f(x) = |x|$ は原点 0 で微分不可能) ですが、 $f_i(\theta) = |\hat{y}_i(\theta) - y_i|^2$ は微分可能な関数です。導関数といった微分情報を利用するオプティマイザについて焦点を当てるので、図 2-3 では、**二乗誤差** $f_i(\theta) = |\hat{y}_i(\theta) - y_i|^2$ を損失関数として採用することにします。

次に、図 2-4 のようなニューラルネットワーク θ ((2.5) 参照) における訓練データ $(\mathbf{x}_i, \mathbf{y}_i)$ (図 2-4 では入力画像 \mathbf{x}_i が“モルモット”なので、分類クラス $(C_1, C_2, C_3) = (\text{モルモット}, \text{車}, \text{猫})$ に基づいて、正解ラベルは $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})^\top = (1, 0, 0)^\top$ となります) に関する損失関数を定義します。いま、訓練データ \mathbf{x}_i がクラス C_r (図 2-4 では、 $C_1 = \text{モルモット}$) に属するとします。このとき、(2.4) から、クラス C_r に属する確率 (尤度) $p_r(\theta)$ を得ることができます。この確率をできるだけ 1 に近づきたいので、対数尤度 $\log_e p_r(\theta)$ を予測値とすると、正解値は $\log_e y_{ir} = \log_e 1 = 0$ となります。以上のことから、損失関数 $f_i: \mathbb{R}^d \rightarrow \mathbb{R}_+$ は

$$\begin{aligned}
 f_i(\theta) &= -\underbrace{(\log_e p_r(\theta))}_{\text{予測値}} - \underbrace{(\log_e 1)}_{\text{正解値}} = -\log_e p_r(\theta) \\
 &= -\left(\underbrace{y_{i1}}_{=0} \log p_1(\theta) + \cdots + \underbrace{y_{ir}}_{=1} \log p_r(\theta) + \cdots + \underbrace{y_{ik}}_{=0} \log p_k(\theta)\right) \\
 &= -\sum_{j=1}^k y_{ij} \log p_j(\theta) \tag{2.7}
 \end{aligned}$$

と表現することができます。なお、損失関数値が非負になるように負の対数尤度に修正します。(2.7) で定義される $f_i(\theta)$ を**交差エントロピー誤差**と呼びます。図 2-4 では、 $f_i(\theta) = -\log p_1(\theta)$ となります。

経験損失

訓練データは一般的に多数からなるので、一つの訓練データに特化せずに、各訓練データに対して平等に配慮するような損失を考える必要があります。そこで、以下で定義される全損失関数の平均を考察します。

[経験損失：全損失関数の平均]

訓練データの総数を n とします。機械学習モデル θ における i 番目 ($i = 1, 2, \dots, n$) の訓練データ $(\mathbf{x}_i, \mathbf{y}_i)$ に関する損失関数を $f_i(\theta)$ とします ((2.6)、(2.7) 参照)。このとき、全損失関数の平均

$$f(\theta) := \frac{1}{n} \sum_{i=1}^n f_i(\theta) \quad (2.8)$$

を機械学習モデル θ における全訓練データに関する**経験損失**といいます。

図 2-3 での経験損失を表してみましょう。(2.6) から、訓練データ $(\mathbf{x}_i, \mathbf{y}_i)$ に関する損失関数は $f_i(\theta) = |\hat{y}_i(\theta) - y_i|^2$ なので、経験損失は

$$f(\theta) \stackrel{(2.8)}{=} \frac{1}{n} \sum_{i=1}^n |\hat{y}_i(\theta) - y_i|^2 \stackrel{(2.1)}{=} \frac{1}{n} \sum_{i=1}^n \left| \mathbf{x}_i^\top \theta - y_i \right|^2$$

となります。ここで、以下で定義される全訓練データ $\mathbf{x}_1, \dots, \mathbf{x}_n$ の情報を有する $n \times d$ 行列 X と正解値を成分にもつベクトル \mathbf{y}

$$X := \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1d} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{id} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_i^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}, \quad \mathbf{y} := \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

を用いて、損失関数を表現してみます。

$$X\boldsymbol{\theta} - \mathbf{y} = \begin{pmatrix} \sum_{j=1}^d x_{1j}\theta_j - y_1 \\ \vdots \\ \sum_{j=1}^d x_{ij}\theta_j - y_i \\ \vdots \\ \sum_{j=1}^d x_{nj}\theta_j - y_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \boldsymbol{\theta} - y_1 \\ \vdots \\ \mathbf{x}_i^\top \boldsymbol{\theta} - y_i \\ \vdots \\ \mathbf{x}_n^\top \boldsymbol{\theta} - y_n \end{pmatrix}$$

とノルムの定義 (1.3) から

$$\begin{aligned} \|X\boldsymbol{\theta} - \mathbf{y}\|^2 &= (\mathbf{x}_1^\top \boldsymbol{\theta} - y_1)^2 + \cdots + (\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2 + \cdots + (\mathbf{x}_n^\top \boldsymbol{\theta} - y_n)^2 \\ &= \sum_{i=1}^n (\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2 = n f(\boldsymbol{\theta}) \end{aligned}$$

つまり、[図 2-3](#) での経験損失である **平均二乗誤差** (Mean Squared Error ; MSE) は

$$f(\boldsymbol{\theta}) = \frac{1}{n} \|X\boldsymbol{\theta} - \mathbf{y}\|^2 \quad (2.9)$$

と表現ができます。[図 2-4](#) での経験損失である **(平均) 交差エントロピー誤差** (Cross-Entropy Loss ; CE Loss) は

$$f(\boldsymbol{\theta}) \stackrel{(2.8)}{=} \frac{1}{n} \sum_{i=1}^n f_i(\boldsymbol{\theta}) \stackrel{(2.7)}{=} -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k y_{ij} \log p_j(\boldsymbol{\theta}) \quad (2.10)$$

となります。ただし、 $p_j(\boldsymbol{\theta})$ は (2.4) で定義されます。

2-3 機械学習モデルの訓練は経験損失の最適化

オブティマイザがどのようにして機械学習モデルを訓練するのか、について明確にしましょう。図 2-3 での経験損失 (2.9)

$$f(\theta) = \frac{1}{n} \left\| \underbrace{X\theta}_{\text{予測}} - \underbrace{\mathbf{y}}_{\text{正解}} \right\|^2 (\geq 0)$$

を改めて見てみます。

- θ は機械学習モデルを表すパラメータ
- X は全訓練データの情報を有する行列
- $X\theta$ は n 個の訓練データに対する機械学習モデル θ による予測値
- \mathbf{y} は全訓練データの正解ラベル

なので、予測 $X\theta$ と正解 \mathbf{y} ができるだけ近い、つまり

$$\|X\theta^* - \mathbf{y}\|^2 = \min_{\theta \in \mathbb{R}^d} \|X\theta - \mathbf{y}\|^2$$

となるような θ^* が正に学習済みモデルです。これを経験損失 (2.9) を用いて表すと

$$\|X\theta^* - \mathbf{y}\|^2 = \min_{\theta \in \mathbb{R}^d} \|X\theta - \mathbf{y}\|^2 \Leftrightarrow f(\theta^*) = \min_{\theta \in \mathbb{R}^d} f(\theta)$$

となります¹。よって、以下のようにまとめることができます。

¹ ニューラルネットワーク θ を用いると入力 x の予測値は $x\theta$ です。例えば、1 番目の訓練データを $(x_1, y_1) = (1, 1)$ 、2 番目の訓練データを $(x_2, y_2) = (2, 3)$ とすると、 $f_1(\theta) = |\theta - 1|^2$ を最小にする点は $\theta_1^* = 1$ であり、 $f_2(\theta) = |2\theta - 3|^2$ を最小にする点は $\theta_2^* = \frac{3}{2}$ です。つまり、一般には、 i 番目の訓練データに関する損失関数 f_i を最小にする θ_i^* は j ($\neq i$) 番目の訓練データに関する損失関数 f_j を最小にできません。 $\|\theta(x_1, x_2)^\top - (y_1, y_2)^\top\|^2 = (\theta - 1)^2 + (2\theta - 3)^2$ から、経験損失 $f = \frac{1}{2}(f_1 + f_2)$ を最小にする点とその最小値は $\theta^* = \frac{7}{5}$ ($\neq \theta_1^* \neq \theta_2^*$) と $f(\theta^*) = \frac{1}{10} > 0 = f_1(\theta_1^*) = f_2(\theta_2^*)$ となります。

[機械学習モデルの訓練 ⇔ 経験損失の最適化：図 2-3]

オプティマイザが図 2-3 の機械学習モデルを訓練して学習済みモデル θ^* を生成する。

$$\Updownarrow \boxed{\|X\theta - \mathbf{y}\|^2 \rightarrow \text{最小化}}$$

オプティマイザが経験損失 (2.9) の最小解 θ^* を探す。

このことから、オプティマイザが経験損失を最適化することで、機械学習モデルを訓練することになります。オプティマイザとは経験損失の最適解 θ^* を探す最適化手法（最適化アルゴリズム）のことです。オプティマイザの具体的な構造や性質については、第 5-1 章で詳解します。

図 2-4 での経験損失 (2.10)

$$f(\theta) = -\frac{1}{n} \sum_{i=1}^n \underbrace{\sum_{j=1}^k y_{ij} \log p_j(\theta)}_{L_i(\theta): \text{対数尤度}} (\geq 0)$$

については

- θ は機械学習モデルを表すパラメータ
- $p_j(\theta)$ はクラス C_j に属する確率 ($j = 1, 2, \dots, k$)
- $L_i(\theta) := \sum_{j=1}^k y_{ij} \log p_j(\theta) = y_{ir} \log p_r(\theta)$ は機械学習モデル θ による訓練データ \mathbf{x}_i が正解クラス C_r に属する尤もらしさ
- n は訓練データの総数

から、全ての訓練データ $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_n, \mathbf{y}_n)$ の対数尤度 $L_1(\theta^*), L_2(\theta^*), \dots, L_n(\theta^*)$ をできるだけ最大にする θ^* 、つまり、訓練データが正解クラスに属する確率を最大にする（確率をできるだけ 1 にする）ような θ^* が学習済みモデルとなります。このとき

$$\sum_{i=1}^n L_i(\theta^*) = \max_{\theta \in \mathbb{R}^d} \sum_{i=1}^n L_i(\theta) \Leftrightarrow \frac{1}{n} \sum_{i=1}^n L_i(\theta^*) = \max_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n L_i(\theta)$$

となります。さらに

$$-\frac{1}{n} \sum_{i=1}^n L_i(\theta^*) = - \left\{ \max_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n L_i(\theta) \right\} = \min_{\theta \in \mathbb{R}^d} \left\{ -\frac{1}{n} \sum_{i=1}^n L_i(\theta) \right\}$$

のように、max および min に関するマイナス符号の入れ替えに注意す

ことで、学習済みモデル θ^* は (2.10) で定義される経験損失の最小解となります。よって、以下のようにまとめることができます。

[機械学習モデルの訓練 \Leftrightarrow 経験損失の最適化：図 2-4]

最適化が図 2-4 の機械学習モデルを訓練して学習済みモデル θ^* を生成する。

$$\Downarrow \left[\sum_{i=1}^n L_i(\theta) \rightarrow \text{最大化} \right]$$

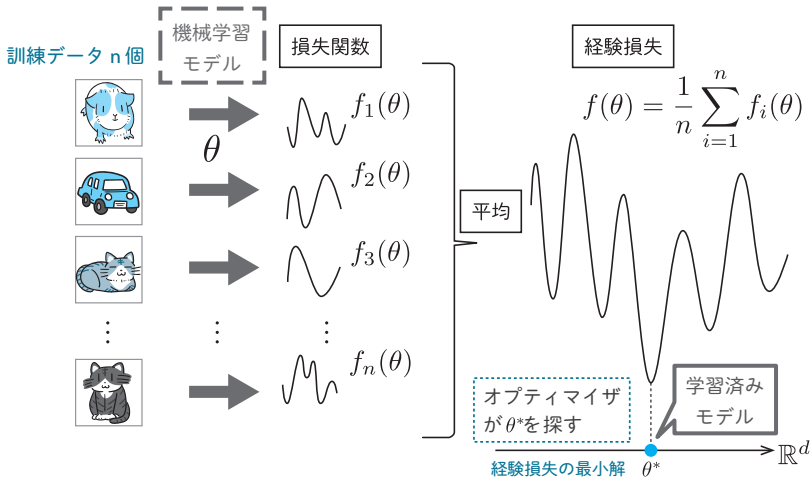
最適化が経験損失 (2.10) の最小解 θ^* を探す。

● 経験損失は一般には凸関数ではない

図 2-5 で、これまでの内容をまとめます。 n 個の訓練データと機械学習モデル θ が与えられたとき、機械学習モデル θ を変数にもつ損失関数 $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$ ($i = 1, 2, \dots, n$) が定義できます。

ここで、シグモイド関数 $\varsigma(x) := \frac{1}{1+e^{-x}}$ によって定義される非線形写像 s

図 2-5 学習済みモデル：経験損失の最小解



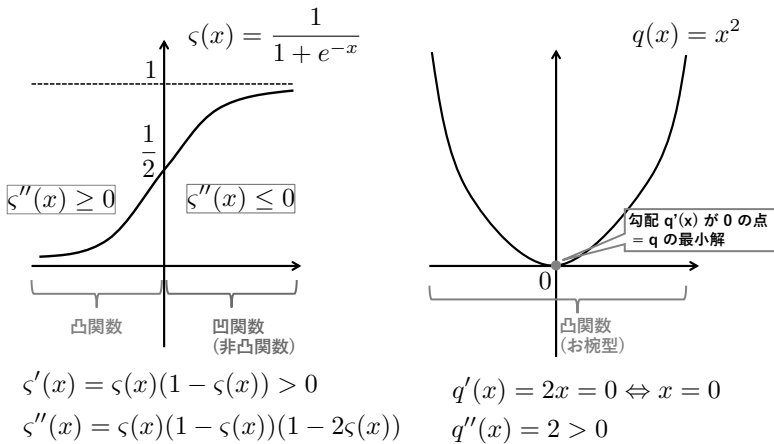
$$s(\boldsymbol{x}) := \begin{pmatrix} \varsigma(x_1) \\ \vdots \\ \varsigma(x_i) \\ \vdots \\ \varsigma(x_p) \end{pmatrix} = \begin{pmatrix} \frac{1}{1+e^{-x_1}} \\ \vdots \\ \frac{1}{1+e^{-x_i}} \\ \vdots \\ \frac{1}{1+e^{-x_p}} \end{pmatrix}$$

と(2.2)、(2.3)、(2.4)、および、(2.7)で定義される損失関数 f_i について考えます。まず初めに、 $\varsigma(x) = \frac{1}{1+e^{-x}}$ が凸関数になるかどうかを調べてみましょう (図 2-6)。

なぜ、対象の微分可能な関数が凸関数になるかどうかを調べるのかというと、お椀型の形状を有する凸関数の勾配が零になる点が正にその関数の最小解になるからです (図 2-6 の $q(x) = x^2$ の勾配 $q'(x) = 2x$ が零となる点が q の最小解 $x = 0$ になることがわかります)。凸関数は最小解を考察する上でとても扱いやすい関数ということになります。

さらに、2回微分可能な関数 $f: \mathbb{R} \rightarrow \mathbb{R}$ が凸であることの必要十分条

図 2-6 非凸関数 (左) と凸関数 (右)



件はその2次導関数 $f''(x)$ が非負になることです (\mathbb{R}^d で定義される凸関数の性質については、[2 回連続的微分凸関数の重要な性質 1] を参照)。
 $\varsigma(x) = \frac{1}{1+e^{-x}}$ の1次導関数 $\varsigma'(x)$ と2次導関数 $\varsigma''(x)$ は $\varsigma(x)$ を用いて

$$\begin{aligned} \varsigma'(x) &= \frac{e^{-x}}{(1+e^{-x})^2} = \varsigma(x)(1-\varsigma(x)) > 0 \\ \varsigma''(x) &= \varsigma'(x)(1-\varsigma(x)) - \varsigma(x)\varsigma'(x) \\ &= \underbrace{\varsigma(x)(1-\varsigma(x))}_{\varsigma'(x) > 0} (1-2\varsigma(x)) \begin{cases} \geq 0 & (x \leq 0; \varsigma(x) \leq \frac{1}{2}) \\ \leq 0 & (x \geq 0; \varsigma(x) \geq \frac{1}{2}) \end{cases} \end{aligned}$$

となるので、 $x \leq 0$ の範囲では、 $\varsigma(x)$ は凸関数ですが、 $x \geq 0$ の範囲では、凸関数ではありません (凹関数といいます)。そのため、シグモイド関数 $\varsigma(x)$ は定義域全体では凸関数ではありません。なお、 $\varsigma'(x) > 0$ なので、 $\varsigma(x)$ の勾配が零になることを利用して ς の最小解を得ることは不可能です (図 2-6 からわかるように $\varsigma(x)$ は $x \rightarrow -\infty$ とすることで、いくらでも 0 に近づけることはできますが、 $\varsigma(x)$ が丁度零になるような x は存在しません)。

非凸関数の一つでも含んでしまう関数は一般には非凸関数なので、 i 番目の訓練データに関する損失関数 f_i は、一般には、凸関数ではありません (図 2-5 の f_1, f_2, f_3 参照)。各訓練データに関する損失関数が凸関数ではないので、それらの平均として定義される経験損失 f も一般には凸関数ではありません (図 2-5 の f 参照)。山や谷が多く存在するような多峰性を有する経験損失を最小にする点が学習済みモデルを表現します (図 2-5 の θ^* は経験損失の最小解)。この最小解を見つけることがオプティマイザの役目です。

その一方で、機械学習モデルが単純であれば経験損失が凸関数になることも起こり得ます。例えば、図 2-3 の機械学習モデルにおける経験損失 $f(\theta) = \frac{1}{n} \|X\theta - \mathbf{y}\|^2$ ((2.9) 参照) は凸関数になります。ユークリッド空間 \mathbb{R}^d で定義される関数 f が凸関数であるとは、(1.23) から、任意の $\theta_1, \theta_2 \in \mathbb{R}^d$ と任意の $\lambda \in [0, 1]$ に対して

$$f(\lambda\theta_1 + (1-\lambda)\theta_2) \leq \lambda f(\theta_1) + (1-\lambda)f(\theta_2)$$

を満たすときをいいます。 $f(\boldsymbol{\theta}) = \frac{1}{n}\|X\boldsymbol{\theta} - \mathbf{y}\|^2$ が上記の不等式を満たすことを示してみましょう。スカラー α とベクトル \mathbf{x}, \mathbf{y} に対して、 $X(\mathbf{x} + \mathbf{y}) = X\mathbf{x} + X\mathbf{y}$ と $X(\alpha\mathbf{x}) = \alpha X\mathbf{x}$ (所謂、行列 X の線形性) が成り立つことから

$$\begin{aligned} f(\lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2) &= \frac{1}{n}\|X(\lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2) - \mathbf{y}\|^2 \\ &= \frac{1}{n}\|\lambda X\boldsymbol{\theta}_1 + (1-\lambda)X\boldsymbol{\theta}_2 - \mathbf{y}\|^2 \end{aligned}$$

となります。ここで、 $\mathbf{y} = \lambda\mathbf{y} + (1-\lambda)\mathbf{y}$ となることに注意をして

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

を得ます。上記の右辺はノルムの二乗展開 ([ユークリッド空間で成り立つ等式や不等式 2]) ができるので

$$\begin{aligned} &f(\lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2) \\ &\leq \frac{1}{n}\left\{ \lambda\|X\boldsymbol{\theta}_1 - \mathbf{y}\|^2 + (1-\lambda)\|X\boldsymbol{\theta}_2 - \mathbf{y}\|^2 - \underbrace{\lambda(1-\lambda)}_{\geq 0} \underbrace{\|X(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|^2}_{\geq 0} \right\} \\ &\leq \lambda\left(\frac{1}{n}\|X\boldsymbol{\theta}_1 - \mathbf{y}\|^2\right) + (1-\lambda)\left(\frac{1}{n}\|X\boldsymbol{\theta}_2 - \mathbf{y}\|^2\right) \\ &= \lambda f(\boldsymbol{\theta}_1) + (1-\lambda)f(\boldsymbol{\theta}_2) \end{aligned}$$

となり、確かに、 $f(\boldsymbol{\theta}) = \frac{1}{n}\|X\boldsymbol{\theta} - \mathbf{y}\|^2$ は凸関数となります。以上のような証明方法 (行列の線形性、ベクトルの凸結合性 ($\mathbf{y} = \lambda\mathbf{y} + (1-\lambda)\mathbf{y}$)、ノルムの二乗展開を用いる方法) は、最適化の議論でよく利用するので、ぜひ習得をしたい方法の一つです。[2 回連続的微分凸関数の重要な性質 1] を利用することで、 $f(\boldsymbol{\theta}) = \frac{1}{n}\|X\boldsymbol{\theta} - \mathbf{y}\|^2$ が凸関数になることもいえます。これについては、多次元ユークリッド空間で定義される関数の微分情報について詳解する次章で扱うことにします。

● 経験損失は部分的には凸関数

経験損失が一般には凸関数ではないので、凸関数の性質を理解しなくても良さそうに感じますが、 $q(x) = x^2$ のようなお椀型の形状 (図 2-6) を経験損失の谷の周りで有することから、経験損失は部分的 (局所的) にみれば凸関数です。そのため、経験損失を知る上で関数の凸性を理解することはとても大事なことだといえます。詳細については、次章で説明します。章末問題では、2 次関数の凸性について考察します。最適化でよく現れる議論なので、興味のある読者の方はぜひ立ち寄ってください。

まとめ

- 機械学習モデルを訓練するために必要なもの
訓練データ (正解ラベルが付与されている既知データ)
機械学習モデル (ノードとエッジからなるニューラルネットワーク)
オプティマイザ (検証データによって調整された最適化手法)
- 機械学習モデルは入力された訓練データの予測値を出力する
- 損失関数
機械学習モデルによる訓練データの予測値と訓練データの正解値との誤差の大きさのことで機械学習モデルを変数にもつ関数
- 経験損失
全訓練データに関する損失関数の平均のことで機械学習モデルを変数にもつ (非凸微分可能) 関数
- 機械学習モデルを訓練する \Leftrightarrow 経験損失を最適化する
オプティマイザが機械学習モデルを訓練することはオプティマイザが経験損失の最小解を探すことと同値

[問題]

2.1 $X = (x_{ij}) \in \mathbb{S}_+^d$ とする。このとき、 $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^\top \in \mathbb{R}^d$ に対して、関数 $q: \mathbb{R}^d \rightarrow \mathbb{R}$ を

$$q(\boldsymbol{\theta}) := \langle \boldsymbol{\theta}, X\boldsymbol{\theta} \rangle = \boldsymbol{\theta}^\top X \boldsymbol{\theta}$$

と定義する。

(1) $q(\boldsymbol{\theta})$ が以下のような θ_i の 2 次式で表現できることを示せ (この結果から、関数 q は 2 次関数と呼ばれる)。

$$q(\boldsymbol{\theta}) = \sum_{j=1}^d \sum_{i=1}^d x_{ij} \theta_i \theta_j$$

(2) $X \in \mathbb{S}_+^d$ の必要十分条件が $X = GG^\top$ となる $G \in \mathbb{R}^{d \times d}$ の存在条件であることを利用して

$$q(\boldsymbol{\theta}) = \left\| G^\top \boldsymbol{\theta} \right\|^2$$

となることを示せ。

(3) 2 次関数 q が凸関数になることを示せ。

2.2 $\mathbf{b} \in \mathbb{R}^d$ とし、 $c \in \mathbb{R}$ とする。 $l(\boldsymbol{\theta}) := \langle \mathbf{b}, \boldsymbol{\theta} \rangle$ と $c(\boldsymbol{\theta}) := c$ が凸関数になることを示せ。

2.3 有限個の凸関数の和が凸関数になることを示せ。このことを用いて、 $f(\boldsymbol{\theta}) := \frac{1}{2}q(\boldsymbol{\theta}) + l(\boldsymbol{\theta}) + c$ で定義される関数 f が凸関数になることを示せ。

略解は P.362