

2022 年度
応用統計学

大坂公立大学 大学院理学研究科 数学専攻

今野 良彦

2023 年 2 月 9 日

はじめに

2022 年 9 月 10 日更新

目次

はじめに	i
第 1 章 統計的学習の考え方	1
1 導入	1
2 教師あり学習と教師なし学習	2
3 教師あり学習における訓練誤差と汎化誤差	6
4 教師あり学習における兼ね合い (trade-off)	10
5 教師なし学習における兼ね合い (trade-off)	12
第 2 章 Monte Carlo 法	19
1 乱数の発生法	19
1.1 ベルヌーイ分布	19
1.2 2 項分布	19
1.3 標準正規分布	19
1.4 正規分布	21
1.5 確率の積分変換	21
1.6 受容・棄却法	21
2 Markov 連鎖 Monte Carlo 法 (MCMC 法)	23
2.1 基本的な考え方	24
2.2 推移核の Metropolis-Hastings による構成	26
3 ブートストラップ法	32
第 3 章 最尤推定値の計算と EM アルゴリズム	35
1 Newton-Raphson 法	35
2 Fisher のスコア法	38
3 EM アルゴリズム	40

第 4 章	線型回帰モデル	49
1	線型単回帰モデルと最小 2 乗推定量	49
2	線型重回帰モデル	53
3	変数選択の規準	57
3.1	自由度調整済み決定係数	57
3.2	Mallows C_p	58
3.3	AIC	59
3.4	交差検証法	62
3.5	BIC	64
4	補遺：Laplace 近似について	68
第 5 章	指数型分布族と一般化線型モデル	71
1	指数型分布族	71
2	指数型分布族の性質	73
3	一般化線型モデル (Generalized linear models(GLM))	76
4	GLM の最尤推定法	77
第 6 章	正則化とカーネル法	83
1	導入	83
2	正則化	85
3	カーネル関数	89
4	カーネル関数から内積空間の構成	91
5	例	96
5.1	連立方程式	96
5.2	補間スプライン	99
第 7 章	2 値判別分析と回帰	103
1	サポートベクターマシン (SMV)	103
1.1	SMV の考え方	103
第 A 章	基礎的な確率規則	107
1	基礎的な確率規則	107
2	確率変数	111

3	分位点関数	115
4	主な 1 次元分布	116
4.1	離散型確率変数	116
4.2	連続型確率変数	117
5	2 次元の分布	118
5.1	同時確率関数と密度関数	118
5.2	周辺分布	119
5.3	独立性な分布と条件付き分布	120
6	多次元分布と i.i.d. 標本	123
6.1	重要な多次元分布モデル	124
7	正規分布に関連した分布	129
7.1	ガンマ分布	129
7.2	χ^2 分布	131
7.3	F 分布	132
7.4	t 分布	134
第 B 章	期待値	137
1	期待値	137
2	確率ベクトルの期待値	140
3	分散と共分散	141
4	条件付き期待値	147
5	積率母関数	152
第 C 章	経験分布関数と統計的汎関数	155
1	経験分布関数	155
2	統計的汎関数	156
第 D 章	Hilbert 空間	159
1	Hilbert 空間の定義	159
第 E 章	線型代数	165
1	特異値分解	165
2	Moore-Penrose の一般化逆行列	173

第 F 章	凸最適化	177
1	凸関数とその性質	177
2	凸集合とその性質	186
3	共役凸関数と Young の不等式	196
4	非線形最適化と KKT 条件	206
5	Lagrange 乗数についてのさらなる議論	210
6	変分不等式	214
7	凸性と Lagrange 乗数	217
8	凸双対性	221
	8.1 双対問題	221
	8.2 Slater 条件 (再訪問)	223
9	凸双対性 (その 2)	227
	9.1 Fenchel 双対性	227
10	半正定値計画法	230
11	最適化アルゴリズム	234
	11.1 多変数関数の微分	235
12	勾配降下法	239
	12.1 降下方向	239
	12.2 最急降下方向	240
	12.3 ステップ幅の選択方法	240
	参考文献	244

第1章 統計的学習の考え方

1 導入

データサイエンスにおいて、データの視覚化や整理¹は重要であるが、主な課題はデータの数理的な解析である。データの不確実性(ランダム性)のモデル化や定量化を目的とするとき、これらのことを扱う分野を統計的推測/統計的学習といい、大規模データを用いた予測に重点をおく分野を機械学習またはデータマイニングという。何に重きをおくかの違いで、ふたつの文化²が存在する。

データのモデリングにはふたつの目的がある。

- (1) ある変量の将来の値を正確に予測すること。
- (2) データの中にある未知の面白いパターンを発見すること。

この目的を達成するために、数理科学の3つの重要な支柱からの知識を利用する。

- (1) 関数の近似: データに対して数理モデルを作ることは、ある変数が他の変数にどのように依存/関係しているかを理解することである。変数間の関連を表現する最も自然なやり方は、関数もしくは写像を用いることである。この関数ないしは写像が解析者³には完全には特定できないとき、データによる近似を試みる。
- (2) 最適化: 統計的モデルの族⁴が与えられたとき、この族の中で最もよいモデルをみつきたい。このために、効率的な探索と最適化の手続きが必要となる。最適化のステップは観測データへの関数のあてはめや校正と考えることができる。このステップでは、最適化アルゴリズムや効率的なプログラミ

¹いわゆる記述統計学の重要な内容。

²Breiman (2001): Statist Sci 16(3), pp. 199-231 を参照のこと。

³統計家、データサイエンティストなどデータを分析する人々の総称として解析者とよぶことにする。

⁴統計推測では、数理モデルの族を統計モデルという。「モデル」という言葉が何をさしているかを文脈から正確に理解することが大切である。

ングが必要となる。

- (3) 確率論と統計的推測: モデルをあてはめるデータはランダムな過程の実現値とみなす。ここでは、分布や確率が重要になる。将来の予測を行うときに内包する不確実性を定量化するために、確率論や統計的推測の知識が肝要である。

多次元のデータを扱う場合には、これらの知識に加えて、線型代数の知識が重要となる。

2 教師あり学習と教師なし学習

sec:a1-2

入力ベクトル x が与えられたとき、出力 y の予測をすることが機械学習の主な目的である。たとえば、 x をデジタル化された署名データとし、 y を 2 値の出力で、署名が「本人のもの」か「本人のものではない」のいずれかを表すとする。別の例では、入力 x は妊婦の体重と喫煙嗜好とし、出力 y を生まれてくる新生子の体重とする。入力 x による出力 y の予測値を関数 g を用いて、 $g(x)$ と表す。 g の定義域は、入力 x の取りうる値の集合を含むものであり、終域は y の取りうる値の集合である。関数 g のことを予測関数とよぶことにする。関数 g は、データがもつランダム性を除いた入力 x と出力 y の関連性についての情報を内包するものである。出力 y が実数値をとるとき、出力から入力を予測することを回帰問題といい、出力 y が 2 値もしくは有限集合に値をとるとき、判別問題⁵という。

入力 x による出力 y の予測値を $\hat{y} := g(x)$ と書いたとき、この予測値の精度を損失関数 $\text{Loss}(y, \hat{y})$ で測る。損失関数は非負値関数で、精度がわるいほどおおきな値をとるものである。2 値の判別問題では、損失関数を 0-1 損失とすることが基本的である。

$$\text{Loss}(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\} = \begin{cases} 1 & (y \neq \hat{y}) \\ 0 & (y = \hat{y}). \end{cases}$$

後で 2 値の判別問題でも別の損失関数を考えることになる。回帰問題では、2 乗

⁵分類問題ともいう。

損失

$$\text{Loss}(y, \hat{y}) = (y - \hat{y})^2$$

が最もよく使われる.

すべての入力と出力の組 (x, y) に対して正確な予測をすることができる予測関数を見つけることは一般に不可能である. これはデータのもつランダム性に由来する. さきほどの例において同じ値の入力に対して異なる出力をもつ組が存在することを考えればよい. すなわち各 (x, y) はある同時分布に従う確率ベクトル (X, Y) の実現値と考えることにする. このことで予測関数 g の精度を期待損失 (誤差/リスク) で定量化する.

$$\text{Err}(g) := \mathbb{E}[\text{Loss}(Y, g(X))].$$

ただし期待値 $\mathbb{E}[\cdot]$ は (X, Y) の同時分布に関するものである. $0-1$ 損失を用いた判別問題では, 誤差は誤差判別の確率となる.

$$\text{Err}(g) = \mathbb{P}[Y \neq g(X)].$$

この文脈では予測関数のことを判別器とよぶ.

(X, Y) の同時分布と損失関数を与えられると, 最良の判別器を求めることができる. $c \in \mathbb{N}$, $c \geq 2$ とし, Y は $\{0, 1, \dots, c-1\}$ に値を取るとする. このとき

$$g^* \in \underset{g}{\text{argmin}} \mathbb{E}[\text{Loss}(Y, g(X))]$$

を誤差を最小とする判別器とする. すなわち任意の判別器 g に対して,

$$\mathbb{E}[\text{Loss}(Y, g(X))] \geq \mathbb{E}[\text{Loss}(Y, g^*(X))]$$

が成立している. $X = x$ が与えられたときの Y の条件付き確率関数を $p(y|x)$ と書く. すなわち

$$p(y|x) = \Pr(Y = y | X = x).$$

このとき

$$g^*(x) = \underset{y \in \{0, 1, \dots, c-1\}}{\text{argmin}} p(y|x)$$

となる. 回帰問題で 2 乗損失を用いたとき最適の予測関数 g^* を回帰関数とよぶ.

定理 1.1. 回帰問題において 2 乗損失 $\text{Loss}(y, \hat{y}) = (y - \hat{y})^2$ を用いる. $\mathbb{E}[Y^2] < \infty$ を仮定する. このとき最適予測関数 g^* は次で与えられる.

$$g^*(\boldsymbol{x}) = \mathbb{E}[Y | X = \boldsymbol{x}].$$

すなわち $X = \boldsymbol{x}$ を与えたときの Y の条件付き期待値である.

Proof. $g^*(\boldsymbol{x}) = \mathbb{E}[Y | X = \boldsymbol{x}]$ とする. 任意の (2 乗可積分な) 関数 g を考える. すると

$$\begin{aligned} \mathbb{E}[\{Y - g(X)\}^2] &= \mathbb{E}[\{Y - g^*(X) + g^*(X) - g(X)\}^2] \\ &= \mathbb{E}[\{Y - g^*(X)\}^2] + \mathbb{E}[\{g^*(X) - g(X)\}^2] \\ &\quad + 2\mathbb{E}[\{Y - g^*(X)\}\{g^*(X) - g(X)\}] \\ &\geq \mathbb{E}[\{Y - g^*(X)\}^2] + 2\mathbb{E}[\{Y - g^*(X)\}\{g^*(X) - g(X)\}]. \end{aligned}$$

しかし

$$\begin{aligned} \mathbb{E}[\{Y - g^*(X)\}\{g^*(X) - g(X)\}] &= \mathbb{E}[\mathbb{E}[\{Y - g^*(X)\}\{g^*(X) - g(X)\} | X]] \\ &= \mathbb{E}[\{g^*(X) - g(X)\}\mathbb{E}[Y - g^*(X) | X]] \\ &= \mathbb{E}[\{g^*(X) - g(X)\} \underbrace{\{\mathbb{E}[Y | X] - g^*(X)\}}_{=0}] \\ &= 0. \end{aligned}$$

よって

$$\mathbb{E}[\{(Y - g(X))\}^2] \geq \mathbb{E}[\{Y - g^*(X)\}^2].$$

□

定理 ^{thm:a1-1}1.1 により $X = \boldsymbol{x}$ を与えたときの出力 Y を次のように表現できる.

$$Y = g^*(\boldsymbol{x}) + \epsilon(\boldsymbol{x}).$$

すると

$$\mathbb{E}[\epsilon(\boldsymbol{x}) | X = \boldsymbol{x}] = \mathbb{E}[Y - g^*(\boldsymbol{x}) | X = \boldsymbol{x}] = \mathbb{E}[Y | X = \boldsymbol{x}] - g^*(\boldsymbol{x}) = 0.$$

さらに

$$\mathbb{V}[\epsilon(X) | X = \mathbf{x}] =: \nu^2(\mathbf{x})$$

と書ける。 $\nu(\mathbf{x})$ は未知であり $\epsilon(\mathbf{x})$ の分布は平均が 0 であること以外は未知である。

一般に、最適予測関数 g^* は (X, Y) の同時分布に依存し、その分布は未知である。したがって g^* を実際の問題では用いることができない。 (X, Y) の同時分布 $p_{(X, Y)}(\mathbf{x}, y)$ に従う n 個の確率変数列を

$$\mathfrak{T}_n = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$$

と書き \mathfrak{T}_n を訓練集合という。この訓練集合の実現値を

$$\mathfrak{t}_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

と書く。

目標は訓練集合 \mathfrak{T}_n が与えられたとき、未知の最適予測関数 g^* を推定/学習することである。訓練集合 \mathfrak{T} に基づく g^* の近似/推定量を $g_{\mathfrak{T}}$ と書くこと⁶にする。ここで $g_{\mathfrak{T}}$ はランダムな関数であることに注意せよ。実現値 \mathfrak{t}_n により計算されたものが実現値になり、 $g_{\mathfrak{t}}$ と書く。訓練集合 \mathfrak{T}_n により未知の関数

$$g^* : \mathbf{x} \mapsto \hat{y}$$

を学習する学習器が $g_{\mathfrak{T}}$ である。教師が真の関係に基づき n 個の出力と入力対 \mathfrak{T}_n を見本として与え、その見本で学習器 $g_{\mathfrak{T}}$ を訓練する。新しい入力 X に対して教師によって与えられていない出力 Y を $g_{\mathfrak{T}}(X)$ で予測することになる。この設定を教師あり学習といい、入力のことを説明変数、出力のことを応答変数ということもある。

一方、教師なし学習では変数に説明変数と応答変数の区別はない。データの同時分布 $f(\mathbf{x}, y)$ を学習することになる。この変数を X と書くことにする。 X の分布 f の推定量 g に対して誤差は

$$\text{Err}(g) = \mathbb{E}[\text{Loss}(f(X), g(X))]$$

となる。教師なし学習の例としては、コンビニエンスストアの客の購買行動の解析を考える。100 個の商品があるとする。ある客が、 i ($i = 1, 2, \dots, 100$)

⁶本来ならば、 $g_{\mathfrak{T}_n}$ と書くべきだろうが、煩雑になるので、簡単に $g_{\mathfrak{T}}$ と書くことにした。

番目の商品を買うか買わないの結果を $0, 1$ に対応させる. 客の購買パターンは $x \in \{0, 1\}^{100}$ となる. 訓練集合の実現値 $t_{100} = \{x_1, x_2, \dots, x_{100}\}$ に基づいて, 客の購買パターンを見つけたい. 教師なし学習の手法としては, クラスタリング, 主成分分析, カーネル密度推定, 決定木等がある.

以下の節では, 教師あり学習についてみていくことにする.

3 教師あり学習における訓練誤差と汎化誤差

任意の予測関数 (以後は学習器とよぶことにする.) g が与えられたとき予測誤差を

$$\text{Err}(g) := \mathbb{E}[\text{Loss}(Y, g(X))]$$

で定義する. これを求めることは解析者は一般にはできない. なぜならば統計家にとって (X, Y) の同時分布は未知だからである. 学習器 g の予測誤差 $\text{Err}(g)$ を近似/推定するために訓練集合 \mathfrak{T}_n に基づく訓練誤差

$$\widehat{\text{Err}} := \frac{1}{n} \sum_{j=1}^n \text{Loss}(Y_j, g(X_j))$$

を用いる. 訓練集合 \mathfrak{T}_n は (X, Y) の同時分布に独立同一に従う確率変数列なので

$$\text{Err}(g) = \mathbb{E}[\widehat{\text{Err}}(g)]$$

が成立していることが簡単にわかる. すなわち $\widehat{\text{Err}}(g)$ は $\text{Err}(g)$ の不偏推定量である.

最適予測関数を g^* を近似/学習/推定するために, 学習器の候補の集まりである関数族を \mathcal{G} を選択する. すなわち, 訓練誤差を最小にする学習器を見つけることを目指す. 新しいデータ X の学習器の予測精度を測る尺度を導入する. 訓練集合の実現値 t_n を固定する. すなわち, $(X_j, Y_j) = (x_j, y_j)$ ($j = 1, 2, \dots, n$) とする.

訓練集合の実現値 t_n の基づく学習器 $g_t^{\mathcal{G}}$ を

$$g_t^{\mathcal{G}} \in \operatorname{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{j=1}^n \text{Loss}(y_j, g(x_j))$$

で定義する. (X, Y) を訓練集合 \mathcal{T}_n とは独立に同時分布 $p_{(X, Y)}(x, y)$ に従う確率ベクトルとしたとき学習器 g_t^G の汎化誤差 $\text{Err}(g_t^G)$ を

$$\text{Err}(g_t^G) := \mathbb{E}[\text{Loss}(Y, g_t^G(X))]$$

で定義する. ただし期待値 $\mathbb{E}[\cdot]$ は (X, Y) の同時分布 $p_{(X, Y)}(x, y)$ に関して取ったものである.

ランダムな訓練集合 \mathcal{T}_n に基づく学習器 $g_{\mathcal{X}}^G$ の汎化誤差 $\text{Err}(g_{\mathcal{X}}^G)$ はランダムになるので, \mathcal{T}_n の同時分布に関して期待値をとったものを期待汎化誤差という.

$$\mathbb{E}[\text{Err}(g_{\mathcal{X}}^G)] = \mathbb{E}[\text{Loss}(Y, g_{\mathcal{X}}^G(X))].$$

右辺の $\mathbb{E}[\cdot]$ は \mathcal{T}_n の同時分布に関する期待値であり, 左辺の $\mathbb{E}[\cdot]$ は (X, Y) と \mathcal{T}_n の同時分布に関する期待値である.

re:a1-2

注意 1.2. (多項式回帰) (U, Y) の同時分布を以下のように定める. $U \sim \text{Unif}(0, 1)$ とし $U = u$ ($0 < u < 1$) が与えられたときの Y の条件付き分布は $N(\theta, \sigma^2)$ とする. ただし,

$$\theta = 10 - 140u + 400u^2 - 250u^3, \quad \sigma^2 = 25$$

である. 2 乗損失を用いたとき最良予測関数は

$$g^*(u) = \mathbb{E}[Y|U = u] = 10 - 140u + 400u^2 - 250u^3$$

となる。これは以下の議論からわかる。

$$\begin{aligned}
 \text{Err}(g) &= \mathbb{E}[(Y - g(U))^2] = \mathbb{E}[\{Y - \mathbb{E}[Y|U] + \mathbb{E}[Y|U] + g(U)\}^2] \\
 &= \mathbb{E}[\{\mathbb{E}[Y|U] + g(U)\}^2] + \mathbb{E}[\{\mathbb{E}[Y|U] + g(U)\}^2] \\
 &\quad + 2\mathbb{E}[\{Y - \mathbb{E}[Y|U]\}\{\mathbb{E}[Y|U] + g(U)\}] \\
 &= \mathbb{E}[\{\mathbb{E}[Y|U] + g(U)\}^2] + \mathbb{E}[\{\mathbb{E}[Y|U] + g(U)\}^2] \\
 &\quad + 2\mathbb{E}\left[\mathbb{E}\left[\{Y - \mathbb{E}[Y|U]\}\{\mathbb{E}[Y|U] + g(U)\} \middle| U\right]\right] \\
 &= \mathbb{E}[\{\mathbb{E}[Y|U] + g(U)\}^2] + \mathbb{E}[\{\mathbb{E}[Y|U] + g(U)\}^2] \\
 &\quad + 2\mathbb{E}\left[\underbrace{\{\mathbb{E}[Y|U] + g(U)\} \mathbb{E}\left[Y - \mathbb{E}[Y|U] \middle| U\right]}_{=0}\right] \\
 &= \mathbb{E}[\{Y - \mathbb{E}[Y|U]\}^2] + \mathbb{E}[\{\mathbb{E}[Y|U] - g(U)\}^2] \\
 &\geq \mathbb{E}[\{\mathbb{E}[Y|U] - g(U)\}^2].
 \end{aligned}$$

$d \in \mathbb{N}$, $d \geq 2$ とする。 $d - 1$ 次元の u の多項式がなす関数族を考える。

$$\mathcal{G}_d := \{g(u) = \beta_0 + \beta_1 u + \beta_2 u^2 + \dots + \beta_{d-1} u^{d-1}; \beta_0, \beta_1, \dots, \beta_{d-1} \in \mathbb{R}\}.$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{d-1})^\top, \quad \boldsymbol{x} = (1, u, u^2, \dots, u^{d-1})^\top$$

とおくと

$$\mathcal{G}_d = \{g(u) = \boldsymbol{x}^\top \boldsymbol{\beta}; \boldsymbol{\beta} \in \mathbb{R}^d\}$$

と書ける。 $d \geq 4$ とする。

$$\boldsymbol{\beta}^* = (10, -140, 400, -250, 0, \dots, 0)^\top$$

としたとき $g^*(u) = \boldsymbol{x}^\top \boldsymbol{\beta}^*$ が最良予測関数となる。訓練データを

$$(u_1, y_1), (u_2, y_2), \dots, (u_n, y_n)$$

とし

$$\boldsymbol{x}_j = (1, u_j, u_j^2, \dots, u_j^{d-1})^\top$$

と書く。これらを行列とベクトルで表記する。

$$\underbrace{\mathbf{X}}_{n \times d \text{ 行列}} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^{d-1} \\ 1 & u_2 & u_2^2 & \cdots & u_2^{d-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_n & u_n^2 & \cdots & u_n^{d-1} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

とおく。すると学習器 $g(u) = \mathbf{x}^\top \boldsymbol{\beta}$ の訓練誤差は

$$\widehat{\text{Err}}(g) = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{2,n}^2$$

となる。ただし $\|\cdot\|_{2,n}$ は \mathbb{R}^n の Euclid ノルムである。すると

$$\widehat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathcal{G}_d}{\text{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{2,n}^2$$

とすれば $\widehat{\boldsymbol{\beta}}$ は最小 2 乗推定量となる。 \mathbf{X} の列で張られた線型部分空間を $\text{span}(\mathbf{X})$ と書く。 $\text{span}(\mathbf{X})$ への射影行列を \mathbf{P} とすれば $\widehat{\boldsymbol{\beta}}$ は

$$\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{y}$$

をみたすことがわかる。すると $d \times n$ の行列 \mathbf{X}^\dagger が存在⁷して

$$\mathbf{P} = \mathbf{X}\mathbf{X}^\dagger$$

と書けることが知られている。したがって

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{y}$$

と書けることがわかる。よって

$$g_t^{\mathcal{G}_d}(u) = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}, \quad \widehat{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{y}$$

となる。特に $\mathbf{X}^\top \mathbf{X}$ が正則のとき

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

となる。 □

⁷ \mathbf{X}^\dagger は \mathbf{X} の Moore-Penrose の一般逆行列と呼ばれるものである。補遺の [ch:e](#) 章を参照のこと。

4 教師あり学習における兼ね合い (trade-off)

教師ありの機械学習において汎化誤差

$$\text{Err}(g_t^{\mathcal{G}}) = \mathbb{E}[\text{Loss}(Y, g_t^{\mathcal{G}}(X))]$$

もしくは

$$\mathbb{E}[\text{Err}(g_t^{\mathcal{G}})] = \mathbb{E}[\mathbb{E}[\text{Loss}(Y, g_t^{\mathcal{G}}(X))]]$$

の最小化問題を考える。この問題を解くために関数族 \mathcal{G} を適切に選択することが肝要である。選択の最には以下の点を考慮する必要がある。

- 関数族の複雑さ (最適予測関数を近似するために十分な豊かさを持つことが望ましい。最適予測関数を含む族であればなおさらよい)。
- 最適問題

$$g_{\mathcal{G}}^* \in \operatorname{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{j=1}^n \text{Loss}(Y_j, g(X_j))$$

を解く学習器の訓練が容易である。

- 訓練誤差

$$\widehat{\text{Err}}(g) = \frac{1}{n} \sum_{j=1}^n \text{Loss}(Y_j, g(X_j))$$

が予測誤差

$$\text{Err}(g) = \mathbb{E}[\text{Loss}(Y, g(X))]$$

の推定量としてよい精度をもっているか?

- 特徴量のタイプは何か?

適切な関数族 \mathcal{G} の選択には衝突する因子間の兼ね合いが伴う。簡単な関数族からの学習器の訓練は容易であるが、最良予測関数 g^* をうまく近似/推定していないかもしれない。逆に g^* を含むかもしれない豊かな関数族 \mathcal{G} からの学習器の訓練には計算コストがかかる。関数族の複雑さ、計算コスト、推定精度の関係を理解するために汎化誤差を複数の因子に分解する。分解には二通りある。ひとつは近似と推定精度の兼ね合いを説明するもので、他方はバイアスと分散の兼ね合いを説明する分解である。

まず近似と推定誤差の兼ね合いを理解するための分解を考える.

$$\text{Err}(g_t^{\mathcal{G}}) = \text{Err}^* + \underbrace{\text{Err}(g^{\mathcal{G}}) - \text{Err}^*}_{\text{近似の精度}} + \underbrace{\text{Err}(g_t^{\mathcal{G}}) - \text{Err}(g^{\mathcal{G}})}_{\text{推定の精度}}. \quad (1.1)$$

eq:aa1-

ただし

$$\text{Err}^* = \text{Err}(g^*), \quad g^* \in \underset{g}{\text{argmin}} \text{Err}(g), \quad g^{\mathcal{G}} \in \underset{g \in \mathcal{G}}{\text{argmin}} \text{Err}(g).$$

(1.1) の右辺の第 2 項目を近似の精度といい, 誤差の下限と関数族 \mathcal{G} 中の最適な予測関数の誤差との差である. 関数族 \mathcal{G} の選択と \mathcal{G} 上の $\text{Err}(g)$ の最小化問題は関数解析と数値解析の問題である. 訓練データ t は関わらない. 近似精度をよくするためには, 関数族 \mathcal{G} を大きくとればよい. 第 3 項は統計的誤差になる. これは訓練データにかかわるものである. $g_t^{\mathcal{G}}$ がよりうまく $g^{\mathcal{G}}$ を推定するかどうかである. 標本数 n が無限大に近づくとこの精度は 0 に収束する. 近似と推定の精度をよくするためには互いに相反することが要求される. 推定精度をよくするためには \mathcal{G} を小さくすればよい. 一方, 近似精度をよくするためには関数族 \mathcal{G} を大きく取ればよい. したがって兼ね合いが肝要となる.

2 乗損失を考える. この場合には汎化誤差は

$$\text{Err}(g_t^{\mathcal{G}}) = \mathbb{E}[\{Y - g_t^{\mathcal{G}}(X)\}^2].$$

最良予測関数は

$$g^*(X) = \mathbb{E}[Y | X = \mathbf{x}]$$

であることを思い出そう. すると

- (1) 誤差の下限: $\text{Err}^* = \mathbb{E}[(Y - g^*(X))^2]$.
- (2) 第 2 因子の近似精度: $\text{Err}(g^{\mathcal{G}}) - \text{Err}(g^*) = \mathbb{E}[(g^{\mathcal{G}}(X) - g^*(X))^2]$.
- (3) 第 3 因子の推定精度: $\mathcal{G} = \{g(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}; \boldsymbol{\beta} \in \mathbb{R}^d\}$ としたとき,

$$\text{Err}(g_t^{\mathcal{G}}) - \text{Err}(g^{\mathcal{G}}) = \mathbb{E}[(g_t^{\mathcal{G}}(X) - g^{\mathcal{G}}(X))^2].$$

よって線型関数族 \mathcal{G} に対して

$$\begin{aligned} \text{Err}(g_t^{\mathcal{G}}) &= \mathbb{E}[(g_t^{\mathcal{G}}(X) - Y)^2] \\ &= \text{Err}^* + \mathbb{E}[(g^{\mathcal{G}}(X) - g^*(X))^2] + \mathbb{E}[(g_t^{\mathcal{G}}(X) - g^{\mathcal{G}}(X))^2] \end{aligned}$$

となる.

次に 2 乗損失の場合についてバイアスと分散の兼ね合いについて説明する.
まず

$$\begin{aligned}\text{Err}(g_t^G) &= \mathbb{E}[(g_t^G(X) - Y)^2] \\ &= \mathbb{E}[(g_t^G(X) - g^*(X) + g^*(X) - Y)^2] \\ &= \mathbb{E}[(Y - g^*(X))^2] + \mathbb{E}[(g_t^G(X) - g^*(X))^2] \\ &= \text{Err}^* + \mathbb{E}[D^2(X, t)].\end{aligned}$$

ただし

$$D(X, t) = g_t^G(X) - g^*(X)$$

の t に \mathfrak{T} を代入すると

$$\begin{aligned}\mathbb{E}[(g_{\mathfrak{T}}^G(x) - g^*(x))^2 | X = x] &= \mathbb{E}[D^2(x, \mathfrak{T}) | X = x] \\ &= \{\mathbb{E}[D(x, \mathfrak{T}) | X = x]\}^2 + \mathbb{V}[D(x, \mathfrak{T}) | X = x] \\ &\quad (\because \text{定理 } \text{B.28}^{\text{thm:2-29}}) \\ &= \{\mathbb{E}[g_{\mathfrak{T}}^G(x) - g^*(x) | X = x]\}^2 + \mathbb{V}[g_{\mathfrak{T}}^G | X = x].\end{aligned}$$

よって

$$\mathbb{E}[\text{Err}(g_{\mathfrak{T}}^G)] = \text{Err}^* + \underbrace{\{\mathbb{E}[g_{\mathfrak{T}}^G(X) | X] - g^*(X)\}^2}_{\text{バイアス}} + \underbrace{\mathbb{E}[\mathbb{V}[g_{\mathfrak{T}}^G(X) | X]]}_{\text{分散}}$$

と書ける.

5 教師なし学習における兼ね合い (trade-off)

応答変数 Y と説明変数ベクトル X の区別がある教師あり学習とは対比的に教師なし学習ではこのような変数の区別がない. 観測データ行列 $X^\top [x_1, x_2, \dots, x_n]$ から有用情報やパターンを抽出することが教師なし学習の主目的になる. データ x_1, x_2, \dots, x_n はある未知の確率分布からの n の独立同一標本の実現値と考え, データ行列からこの未知の確率分布を学習/推測することが教師なし学習は本質的に目標とする.

観測データはある未知の確率分布からの n 個のランダム標本の実現値であるときデータの経験分布は未知の確率分布に関する重要な情報を含んでいる。経験分布の概念とカーネル型密度推定については後ほど説明する。

教師なし学習における損失と誤差を導入する。訓練データを

$$\mathfrak{X}_n := \{X_1, X_2, \dots, X_n\}$$

とする。ここで X_j ($j = 1, 2, \dots, n$) は確率ベクトル $X \in \mathbb{R}^p$ の独立複製とする。また X の未知の同時 PDF を p と書くことにする。訓練データ \mathfrak{X}_n の実現値 $t_n = \{x_1, x_2, \dots, x_n\}$ と書くことにする。

教師なし学習では未知の同時 PDF f をうまく近似する PDF g をみつけることが目標である。すなわち $\text{Loss}(f(x), g(x))$ を損失関数としたときその汎化誤差

$$\text{Err}(g) := \mathbb{E}[\text{Loss}(f(X), g(X))]$$

を小さくする PDF g をみつけることである。ただし $\mathbb{E}[\cdot]$ は未知の真の確率分布 f に関する期待値である。損失関数を

$$\text{Loss}(f, g) = \log \frac{f}{g}$$

と取れば、汎化誤差は Kullback-Leibler 情報量 (K-L 情報量とかくことにする) と呼ばれるものになる。以下では K-L 情報量のもとで考える。K-L 情報量は

$$\text{Err}(g) = \mathbb{E}\left[\log \frac{f(X)}{g(X)}\right] = \mathbb{E}[\log f(X)] - \mathbb{E}[\log g(X)]$$

と書ける。上の式の最右辺の第 1 項目は g に関係ないので $\text{Err}(g)$ の最小化は

$$-\mathbb{E}[\log g(X)] = -\int f(x) \log g(x) dx$$

の最小化となる。記号を乱用して

$$\text{Err}(g) := -\mathbb{E}[\log g(X)] \tag{1.2}$$

eq:a1-1

と書くことにする。これをクロス・エントロピー誤差と呼ぶことにする。

次に探索の対象となる確率密度関数の集まりを \mathcal{G} となくことにする. この確率分布の集まりを統計的モデルといった. $f \in \mathcal{G}$ のときクロス・エントロピー誤差 $\text{Err}(g)$ を最小とするものは f となる. すなわち任意の $g \in \mathcal{G}$ に対して

$$\text{Err}(g) \geq \text{Err}(f)$$

が成立する. しかし (II.2) の最小化問題は一般に実行不可能である. なぜならばこの誤差は未知の分布 f に依存するからである. $\text{Err}(g)$ の代わりに訓練クロス・エントロピー誤差

$$\widehat{\text{Err}}_t := \frac{1}{n} \sum_{j=1}^n \log g(x_j)$$

の $g \in \mathcal{G}$ に関する最小化問題を考える.

ここで重要なステップは探索の対象となる統計的モデル \mathcal{G} の選択である. まず \mathcal{G} の元 g を添え字 θ で添え字付くことにする. 添え字 θ の全体の成す集合を Θ と書くことにする. すなわち集合 \mathcal{G} を集合 Θ により母数化する. 添え字の θ を母数といい, Θ を母数空間と呼ぶことにする. すると \mathcal{G} の選択は Θ の選択となる. 以下では $d \in \mathbb{N}$ とし, $\Theta \subset \mathbb{R}^d$ とする. すなわち

$$\mathcal{G} := \{g(\cdot | \theta); \theta \in \Theta \subset \mathbb{R}^d\}. \quad (\text{I.3})$$

この統計的モデルを母数モデルと統計学では伝統的に呼んでいる. 以下では (I.3) の中から f をよく近似する確率分布を学習することにする.

次にデータから近似モデルを探すときの情報な概念を導入する. データの実現値 x が与えられたときに関数

$$\theta \mapsto g(x | \theta) \quad (1.4)$$

を尤度関数という. この関数の $\theta = (\theta_1, \theta_2, \dots, \theta_d)^\top$ に関する勾配をスコア関数といい, $S(x | \theta)$ と書くことにする.

$$S(x | \theta) := \frac{1}{g(x | \theta)} \frac{\partial g(x | \theta)}{\partial \theta} := \frac{1}{g(x | \theta)} \left(\frac{\partial g(x | \theta)}{\partial \theta_1}, \frac{\partial g(x | \theta)}{\partial \theta_2}, \dots, \frac{\partial g(x | \theta)}{\partial \theta_d} \right)^\top.$$

$S(x | \theta)$ の x に X を代入しランダムにしたものを $S(X | \theta)$ と書くことにする.

すると

$$\begin{aligned}\mathbb{E}_{\theta}[S(X|\theta)] &= \int \frac{1}{g(x|\theta)} \left(\frac{\partial g(x|\theta)}{\partial \theta} \right) g(x|\theta) dx = \int \frac{\partial g(x|\theta)}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta} \int g(x|\theta) dx = \mathbf{0}\end{aligned}\quad (1.5) \quad \text{eq:a:1-5}$$

となる。ただし最後の等号は、微分記号と積分記号の交換が可能であることを仮定⁸している。すなわち $g(\cdot|\theta)$ の関する $S(X|\theta)$ の期待値は零ベクトルとなる。さらに $g(\cdot|\theta)$ の関する $S(X|\theta)$ の共分散行列を考える。これを Fisher 情報量行列といい、 $I(\theta)$ と記すことにする。

$$\begin{aligned}I(\theta) &:= \mathbb{E}_{\theta}[S(X|\theta)S^{\top}(X|\theta)] \\ &= \mathbb{E}_{\theta} \left[\begin{array}{cccc} \frac{\partial \log g(X|\theta)}{\partial \theta_1} \frac{\partial \log g(X|\theta)}{\partial \theta_1} & \frac{\partial \log g(X|\theta)}{\partial \theta_1} \frac{\partial \log g(X|\theta)}{\partial \theta_2} & \dots & \frac{\partial \log g(X|\theta)}{\partial \theta_1} \frac{\partial \log g(X|\theta)}{\partial \theta_d} \\ \frac{\partial \log g(X|\theta)}{\partial \theta_2} \frac{\partial \log g(X|\theta)}{\partial \theta_1} & \frac{\partial \log g(X|\theta)}{\partial \theta_2} \frac{\partial \log g(X|\theta)}{\partial \theta_2} & \dots & \frac{\partial \log g(X|\theta)}{\partial \theta_2} \frac{\partial \log g(X|\theta)}{\partial \theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \log g(X|\theta)}{\partial \theta_d} \frac{\partial \log g(X|\theta)}{\partial \theta_1} & \frac{\partial \log g(X|\theta)}{\partial \theta_d} \frac{\partial \log g(X|\theta)}{\partial \theta_2} & \dots & \frac{\partial \log g(X|\theta)}{\partial \theta_d} \frac{\partial \log g(X|\theta)}{\partial \theta_d} \end{array} \right]\end{aligned}$$

で定義する。さらに $-\log g(X|\theta)$ の Hesse 行列の真の分布 g に関する期待値を $J(\theta)$ を

$$\begin{aligned}J(\theta) &:= -\mathbb{E}_{\theta} \left[\frac{\partial S(X|\theta)}{\partial \theta} \right] \\ &:= -\mathbb{E}_{\theta} \left[\begin{array}{cccc} \frac{\partial^2 \log g(X|\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \log g(X|\theta)}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 \log g(X|\theta)}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial^2 \log g(X|\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log g(X|\theta)}{\partial \theta_2 \partial \theta_2} & \dots & \frac{\partial^2 \log g(X|\theta)}{\partial \theta_2 \partial \theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \log g(X|\theta)}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 \log g(X|\theta)}{\partial \theta_d \partial \theta_2} & \dots & \frac{\partial^2 \log g(X|\theta)}{\partial \theta_d \partial \theta_d} \end{array} \right]\end{aligned}$$

で定義する。ある $\theta^* \in \Theta$ が存在して $f = g(\cdot|\theta^*)$ となるとき

$$J(\theta^*) = I(\theta^*)$$

⁸無条件でこの交換が可能なのではない。通常は、交換が可能になるような正則条件を \mathcal{G} や f に課す。この議論では、厳密性に欠くが、正則条件を気にせずに議論をする。

となること⁹がわかる。

n が十分大きいとき行列 $I(\theta)$, $J(\theta)$ はクロス・エントロピー誤差の近似において重要な役割を担う。設定を記述するために, $g^G = g(\cdot | \theta^*)$ をクロス・エントロピー誤差

$$\text{Err}(\theta) := -\mathbb{E}[g(X|\theta)]$$

を最小にする点とする。すなわち任意の $\theta \in \Theta$ に対して

$$\text{Err}(\theta) \geq \text{Err}(\theta^*)$$

である。Err は θ の関数として「なめらか」な関数と仮定する。特に、これらの行列が θ^* の近傍で狭義凸で 2 回連続微分可能¹⁰なときこの仮定をみたしていることがわかる。 θ^* の定義より

$$\begin{aligned} 0 &= \left. \frac{\partial \text{Err}(\theta)}{\partial \theta} \right|_{\theta=\theta^*} = - \left. \frac{\partial \mathbb{E}[\log g(X|\theta)]}{\partial \theta} \right|_{\theta=\theta^*} = - \mathbb{E} \left[\left. \frac{\partial \log g(X|\theta)}{\partial \theta} \right| \right]_{\theta=\theta^*} \\ &= -\mathbb{E}[X(X|\theta^*)] \end{aligned}$$

となる。ただし微分記号と積分記号の交換が可能であることを仮定している。同じように $J(\theta)$ は Err の Hesse 行列であることがわかる。 $g(\cdot | \hat{\theta}_n)$ を訓練誤差

$$\text{Err}_{\hat{\tau}_n}(\theta) := \frac{1}{n} \sum_{j=1}^n \log g(X_j | \theta)$$

⁹ $d = 1$ として確認する。 $d \geq 2$ の場合も本質的には同じである。

$$\begin{aligned} J(\theta) &= -\mathbb{E} \left[\frac{\partial^2 \log g(X|\theta)}{\partial \theta^2} \right] = -\mathbb{E} \left[\frac{\partial}{\partial \theta} \left(\frac{1}{g(X|\theta)} \frac{\partial g(X|\theta)}{\partial \theta} \right) \right] \\ &= \mathbb{E} \left[\frac{1}{g^2(X|\theta)} \left(\frac{\partial g(X|\theta)}{\partial \theta} \right)^2 - \frac{1}{g(X|\theta)} \frac{\partial^2 g(X|\theta)}{\partial \theta^2} \right] \\ &= \mathbb{E} \left[\left(\frac{\partial \log g(X|\theta)}{\partial \theta} \right)^2 \right] - \mathbb{E} \left[\frac{1}{g(X|\theta)} \frac{\partial^2 g(X|\theta)}{\partial \theta^2} \right] \\ &= I(\theta) - \mathbb{E} \left[\frac{1}{g(X|\theta)} \frac{\partial^2 g(X|\theta)}{\partial \theta^2} \right]. \end{aligned}$$

しかし

$$\begin{aligned} \mathbb{E} \left[\frac{1}{g(X|\theta)} \frac{\partial^2 g(X|\theta)}{\partial \theta^2} \right] &= \int \frac{1}{g(x|\theta)} \frac{\partial^2 g(x|\theta)}{\partial \theta^2} g(x|\theta) dx \\ &= \int \frac{\partial^2 g(x|\theta)}{\partial \theta^2} dx = \frac{1}{\partial \theta^2} \int g(x|\theta) dx = 0. \end{aligned}$$

¹⁰たとえば g が正規分布のときにはこの条件をみたしている。

を最小にする点とする. ただし $\mathfrak{T}_n = \{X_1, X_2, \dots, X_n\}$ はランダム標本の集合である. Err^* をすべての確率分布に関するクロス・エントロピーの最小値とする. 明らかに $X \sim p$ のとき最小となる. すなわち $\text{Err}^* = -\mathbb{E}[\log f(X)]$ である. 教師あり学習の場合と同様に汎化誤差 $\text{Err}(\hat{\theta}_n)$ を分解する.

$$\begin{aligned} \text{Err}(\hat{\theta}_n) &= \text{Err}^* + \underbrace{\text{Err}(\theta^*) - \text{Err}^*}_{\text{近似精度}} + \underbrace{\text{Err}(\hat{\theta}_n) - \text{Err}(\theta^*)}_{\text{推定精度}} \\ &= \text{Err}(\theta^*) + \mathbb{E} \left[\frac{g(X|\theta^*)}{g(X|\hat{\theta}_n)} \right]. \end{aligned}$$

上の式の最左辺と最右辺の第 2 項目は \mathfrak{T}_n に依存しているのでランダムな量であることを注意せよ.

第2章 Monte Carlo 法

1 乱数の発生法

Monte Carlo 法は乱数生成法のことである。区間 $(0, 1)$ 上の一様乱数の列を発生させる。この列は決定的なアルゴリズムにより生成されるので真の意味ではランダム性を持たない。疑似乱数とよばれるものである。R では高速で安定した一様疑似乱数を発生するアルゴリズムである Mersenne twister 法を利用している。一様乱数を使って代表的な分布に従う疑似乱数を生成する。

1.1 ベルヌーイ分布

$0 < \theta < 1$ なる p に対して、 $\text{Ber}(\theta)$ に従う乱数を発生させる場合には

- 一様乱数 $U \sim \text{Unif}(0, 1)$ を発生させる。
- $0 < U \leq \theta \implies X = 1, \theta < U < 1 \implies X = 0$ とおく。

このとき $X \sim \text{Ber}(\theta)$ となる。

1.2 2 項分布

$n \in \mathbb{N}, 0 < \theta < 1$ とする。

- $X_1, X_2, \dots, X_n \sim \text{i.i.d. Ber}(\theta)$ を発生させる。
- $S = X_1 + X_2 + \dots + X_n$ とする。

このとき $S \sim \text{Bin}(n, \theta)$ となる。

1.3 標準正規分布

Box-Müller 法を用いる。

- $U_1, U_2 \sim \text{i.i.d. Unif}(0, 1)$ を発生させる.
- $R = \sqrt{-2 \log U_1}, \Theta = 2\pi U_2$ とする.
- $X = R \cos \Theta, Y = R \sin \Theta$ とする.

このとき $X, Y \sim \text{i.i.d. N}(0, 1)$ となる.

これは以下のことからわかる. まず R の PDF は

$$p_R(r) = \begin{cases} r \exp\left(-\frac{r^2}{2}\right) & (r > 0) \\ 0 & (r \leq 0) \end{cases}$$

となることに注意する. したがって (R, Θ) の同時 PDF は

$$p_{(R, \Theta)}(r, \theta) = \begin{cases} \frac{1}{2\pi} r \exp\left(-\frac{r^2}{2}\right) & (r > 0, 0 < \theta < 2\pi) \\ 0 & (\text{その他}) \end{cases} \quad (2.1) \quad \boxed{\text{eq:a2-1}}$$

である. ここで変換

$$\mathbf{h} = (h_1, h_2) : (r, \theta) \mapsto (r \cos \theta, r \sin \theta) = (x, y)$$

を考える.

$$r = \sqrt{x^2 + y^2}, \cos \theta = \frac{x}{\sqrt{x^2 + y^2}}, \sin \theta = \frac{y}{\sqrt{x^2 + y^2}}$$

である. 変換 \mathbf{h} の Jacobian は

$$J_{\mathbf{h}}(r, \theta) = \text{Det} \begin{pmatrix} \frac{\partial h_1}{\partial r} & \frac{\partial h_1}{\partial \theta} \\ \frac{\partial h_2}{\partial r} & \frac{\partial h_2}{\partial \theta} \end{pmatrix} = \text{Det} \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix} = r$$

となるので

$$J_{\mathbf{h}^{-1}}(x, y) = \frac{1}{r} = \frac{1}{\sqrt{x^2 + y^2}}.$$

(2.1) ^{eq:a2-1}に対して PDF の変換公式を用いれば

$$\begin{aligned} p_{(X, Y)}(x, y) &= p_{(R, \Theta)}\left(\sqrt{x^2 + y^2}, \arccos\left(\frac{x}{\sqrt{x^2 + y^2}}\right)\right) \frac{1}{\sqrt{x^2 + y^2}} \\ &= \frac{1}{2\pi} \sqrt{x^2 + y^2} \exp\left(-\frac{x^2 + y^2}{2}\right) \frac{1}{\sqrt{x^2 + y^2}} \\ &= \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) \end{aligned}$$

となることよりわかる.

1.4 正規分布

$-\infty < \mu < \infty, \infty > \sigma > 0$ とする. $N(\mu, \sigma^2)$ に従う疑似乱数を生成させるには

- $Z \sim N(0, 1)$ を発生させる.
- $X = \mu + \sigma Z$ とおく.

このとき $X \sim N(\mu, \sigma^2)$ となる.

1.5 確率の積分変換

連続型分布の CDF F に対してその逆関数 F^{-1} が陽にわかれば F に従う疑似乱数を発生させることができる.

- $U \sim \text{Unif}(0, 1)$ を発生させる.
- $X = F^{-1}(U)$ とおく.

このとき $X \sim F$ となる. なぜならば $\forall x \in \{x \in \mathbb{R}; F(x) > 0\}$ に対して

$$\Pr(X \leq x) = \Pr(F^{-1}(U) \leq x) = \Pr(U \leq F(x)) = F(x)$$

よりわかる.

1.6 受容・棄却法

疑似乱数を確率積分法等で発生させるのが難しい分布の PDF を p とし, 容易に乱数が生成できる分布の PDF を g とする. この g はつぎのように取ることができたとする. ある定数 $C \geq 1$ が存在して $\forall x$ に対して

$$p(x) \leq Cg(x)$$

をみtas. この g に従う疑似乱数を用いて p に従う疑似乱数を発生させることができる.

- $X \sim g$ を発生させる.

- $U \sim \text{Unif}(0, Cg(x))$ を発生させる.
- $f(X) \geq U \implies \tilde{X} = X, f(X) < U \implies X$ を捨てる.

このとき $\tilde{X} \sim p$ となる.

このことは以下からわかる. まず

$$X \sim g, \quad U|X=x \sim \text{Unif}(0, Cg(x))$$

に注意する. $-\infty < x < \infty, 0 < u < Cg(x)$ のとき (X, Y) の同時 PDF は

$$p_{(X,U)}(x, u) = p_{U|X}(u|x)g(x) = \frac{1}{Cg(x)}g(x) = \frac{1}{C}$$

となる. よって

$$p_{(X,U)}(x, u) = \begin{cases} \frac{1}{C} & (-\infty < x < \infty, 0 < u < Cg(x)) \\ 0 & (\text{その他}) \end{cases}$$

である. $t \in \mathbb{R}$ に対して,

$$\begin{aligned} \frac{d}{dt} \Pr(\tilde{X} \leq t) &= \frac{d}{dt} \Pr(X \leq t | f(X) \geq Y) = \frac{\frac{d}{dt} \Pr(X \leq t, p(X) \geq U)}{\mathbb{P}(f(X) \geq U)} \\ &= \frac{\frac{d}{dt} \int_{-\infty}^t \left\{ \int_0^{p(x)} \frac{1}{C} du \right\} dx}{\int_{-\infty}^{\infty} \left\{ \int_0^{p(x)} \frac{1}{C} du \right\} dx} = \frac{\frac{d}{dt} \int_{-\infty}^t \frac{p(x)}{C} dx}{\int_{-\infty}^{\infty} \frac{p(x)}{C} dx} \\ &= \frac{\frac{p(t)}{C}}{\frac{1}{C} \int_{-\infty}^{\infty} p(x) dx} = p(t). \end{aligned}$$

re:a2-1 注意 2.1. $C \geq 1$ としているが, $1/C$ の割合で X を採用しているため, C は 1 に近いほど乱数生成の計算効率はよいことがわかる. \square

ex:a2-2 例 2.2. $\alpha = 1.3, \lambda = 5.6$ としたときの $\text{Gam}(\alpha, \lambda)$ に従う乱数を発生させよう.

$$p(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} & (x \geq 0) \\ 0 & (x < 0). \end{cases}$$

ただし $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ である. g として $\text{Exp}(4)$ を用いる.

$$g(x) = \begin{cases} 4e^{-4x} & (x \geq 0) \\ 0 & (x < 0). \end{cases}$$

$C = 1.2$ とすれば $x \geq 0$ のとき

$$p(x) \leq Cg(x)$$

となることが計算機で確認することができる. $\text{Exp}(4)$ の CDF G は

$$G(x) = \begin{cases} 1 - e^{-4x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

となるので, $0 < u < 1$ に対して

$$G^{-1}(u) = -\frac{\log(1-u)}{4}$$

となる. さらに $1 - U \sim \text{Unif}(0, 1)$ であることに注意すれば

$$U \sim \text{Unif}(0, 1) \implies X = -\frac{\log U}{4} \sim \text{Exp}(4).$$

よって

- $V_1 \sim \text{Unif}(0, 1)$ を発生させる.
- $X = -\frac{\log V_1}{4}$ とおく.
- $V_2 \sim \text{Unif}(0, 1)$ を発生させ, $U = Cg(x)V_2$ とおく.
- $U \leq p(X) \implies \tilde{X} = X$ とする. $U > p(X)$ のときには X を捨てる.

この操作を繰り返せば, p に従う疑似乱数列を得る. □

2 Markov 連鎖 Monte Carlo 法 (MCMC 法)

μ を \mathbb{R}^n 上の確率測度とする. すなわち $\mu: \mathcal{B}(\mathbb{R}^n) \ni A \mapsto \mu(A) \in [0, 1]$ は以下の条件をみたす.

- $\mu(\mathbb{R}^n) = 1$.
- $A_1, A_2, \dots \in \mathcal{B}(\mathbb{R}^n)$, $A_i \cap A_j = \emptyset (i \neq j)$ に対して

$$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j).$$

ただし $\mathcal{B}(\mathbb{R}^n)$ は \mathbb{R}^n 上の Borel 集合族である.

$f: \mathbb{R}^n \rightarrow \mathbb{R}$ は μ 可積分関数としたとき積分

$$\int_{\mathbb{R}^n} f(\mathbf{x})\mu(d\mathbf{x})$$

を求めた.

求積法では台 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^N$ と重み $w_1, w_2, \dots, w_n (N \in \mathbb{N})$ をうまく選んで

$$\int_{\mathbb{R}^n} f(\mathbf{x})\mu(d\mathbf{x}) \approx \sum_{j=1}^N w_j f(\mathbf{x}_j)$$

と近似することを目指す.

Monte Carlo 積分では, 台 \mathbf{x}_j を確率測度 μ からランダムに生成させ, 重み w_j は分布 μ によって決める. X を μ に従う確率ベクトルとしたとき

$$\int_{\mathbb{R}^n} f(\mathbf{x})\mu(d\mathbf{x}) = \mathbb{E}[f(X)] \approx \frac{1}{N} \sum_{j=1}^n f(\mathbf{x}_j) \quad (2.2)$$

eq:mcmc

となることが期待される. ただし $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ は μ からのランダム標本の実現値である. MCMC 法では (2.2) ^{eq:mcmc-1} をみたすような標本を生成する体系的な方法である.

2.1 基本的な考え方

$\mathcal{B} := \mathcal{B}(\mathbb{R}^d)$ を \mathbb{R}^d 上の Borel 集合族とする. 写像

$$P: \mathbb{R}^d \times \mathcal{B} \rightarrow [0, 1]$$

は次の条件をみたす確率推移核 (probability transition kernel) とよぶ.

(1) 各 $B \in \mathcal{B}$ に対して写像

$$\mathbb{R}^d \ni \mathbf{x} \mapsto P(\mathbf{x}, B) \in [0, 1]$$

は可測関数.

(2) 各 $\mathbf{x} \in \mathbb{R}^d$ に対して写像

$$\mathcal{B} \ni B \mapsto P(\mathbf{x}, B) \in [0, 1]$$

は確率測度.

離散時間確率過程は確率変数 $X_j \in \mathbb{R}^d (j = 1, 2, \dots)$ の順序付けられた集合 $\{X_j\}_{j=1}^\infty$ である. 推移核 P をもつ時間等質的な Markov 連鎖とは確率過程 $\{X_j\}_{j=1}^\infty$ で次の性質を持つものである.

$$\begin{aligned} \mu_{X_{j+1}}(B_{j+1} | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j) &= \mu_{X_{j+1}}(B_{j+1} | \mathbf{x}_j) = P(\mathbf{x}_j, B_{j+1}) \\ & \quad (B_{j+1} \in \mathcal{B}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j \in \mathbb{R}^d). \end{aligned}$$

ただし X_j の確率測度を μ_{X_j} と書き, $X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2, \dots, X_j = \mathbf{x}_j$ を与えたときの X_{j+1} の条件付き確率測度を $\mu_{X_{j+1}}(\cdot | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j)$ と書いた. すなわち $X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2, \dots, X_j = \mathbf{x}_j$ を与えたときの X_{j+1} の条件付き分布は $X_j = \mathbf{x}_j$ にも依存する. そして X_{j+1} の分布 $\mu_{X_{j+1}}$ は時間に関して等質である. すると X_{j+1} の分布 $\mu_{X_{j+1}}$ は次のように表現される.

$$\mu_{X_{j+1}}(B_{j+1}) = \int_{\mathbb{R}^d} P(\mathbf{x}_j, B_{j+1}) \mu_{X_{j+1}}(d\mathbf{x}_j) =: \mu_{X_{j+1}} P(B_{j+1}).$$

さらに 3 つの概念を導入する. $\mu P(\cdot) = \int_{\mathbb{R}^n} P(\mathbf{x}, \cdot) d\mu(\mathbf{x})$ とおく.

- 測度 μ は推移核 $P(\mathbf{x}_j, B_{j+1})$ の不変測度であるとは

$$\mu P = \mu \iff \int_{\mathbb{R}^d} P(\mathbf{x}, B) d\mu(\mathbf{x}) = \mu(B) \quad (B \in \mathcal{B})$$

をみたすときをいう.

- 推移核 P は μ に関して既約であるとは各 $\mathbf{x} \in \mathbb{R}^d$ と $B \in \mathcal{B}$ で $\mu(B) > 0$ なるものに対して, ある正の整数 k が存在して

$$P^{(k)}(\mathbf{x}, B) > 0$$

をみたすときをいう。ただし

$$P^{(k)}(\mathbf{x}, B) = \int_{\mathbb{R}^d} P(\mathbf{x}, B) P^{(k-1)}(\mathbf{y}, d\mathbf{x})$$

$$P^{(1)}(\mathbf{x}, B) = P(\mathbf{x}, B) \quad (\mathbf{x} \in \mathbb{R}^d, B \in \mathcal{B})$$

である。

- P は既約な推移核とする。 P は周期的であるとはある整数 $m \geq 2$ が存在して互いに異なる空でない集合 E_j ($j = 1, 2, \dots, m$) と $\mathbf{x} \in E_j$ に対して

$$P(\mathbf{x}, E_{j+1(\text{mod } m)}) = 1$$

であるときをいう。 P が周期的でないとき 非周期的であるという。

pro: mh-3
命題 2.3. μ は \mathbb{R}^d 上の確率測度とし、 $\{X_j\}_{j=1}^\infty$ は推移核 P を持つ時間等質な Markov 連鎖とする。 さらに μ は推移核 P に関して不変測度とし P は既約で非周期的とする。 このとき $\mathbf{x} \in \mathbb{R}^d$ に対し

$$\lim_{N \rightarrow \infty} P^{(N)}(\mathbf{x}, B) = \mu(B) \quad (B \in \mathcal{B})$$

と μ 可積分な関数 f に対して

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N f(X_j) = \int_{\mathbb{R}^d} f(\mathbf{x}) d\mu(\mathbf{x})$$

がほとんど確実に成立する。

Proof. 信じることにする。 □

2.2 推移核の Metropolis-Hastings による構成

μ は \mathbb{R}^d 上の確率測度とし \mathbb{R}^d 上の Lebesgue 測度 $d\mathbf{x}$ に関して絶対連続と仮定する。 このことより \mathbb{R}^d 上の非負関数 π (PDF) が存在して

$$\frac{d\mu}{d\mathbf{x}} = \pi(\mathbf{x})$$

と書ける。

P は任意の推移核とする. 核 $x \in \mathbb{R}^d$ が与えられたとき推移核 P は点 x を異なる点 y に移動させるか, そのまま留める. このことから推移核 P は以下のように分解できると仮定する.

$$P(x, B) = \int_B K(x, \mathbf{y}) d\mathbf{y} + r(x)\mathbb{1}_B(x).$$

ただし

$$\mathbb{1}_B(x) = \begin{cases} 1 & (x \in B) \\ 0 & (\text{その他}) \end{cases}$$

である. 核 $K(x, \mathbf{y}) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$ に対して, x を \mathbf{y} を含む微小な領域 $d\mathbf{y}$ に移す確率は $K(x, \mathbf{y}) d\mathbf{y}$ とし, x に留めおく確率を $r(x)$ である. 条件 $P(x, \mathbb{R}^d) = 1$ から

$$r(x) = 1 - \int_{\mathbb{R}^d} K(x, \mathbf{y}) d\mathbf{y} \tag{2.3}$$

eq:mp-2

を得る.

次に μ が P に関して不変になるための条件を求める. まず $B \in \mathcal{B}$ に対して

$$\begin{aligned} \mu P(B) &= \int_{\mathbb{R}^d} P(x, B) \pi(x) dx \\ &= \int_{\mathbb{R}^d} \left\{ \int_B K(x, \mathbf{y}) d\mathbf{y} + r(x)\mathbb{1}_B(x) \right\} \pi(x) dx \\ &= \int_{\mathbb{R}^d} \left\{ \int_B K(x, \mathbf{y}) d\mathbf{y} \right\} \pi(x) dx + \int_{\mathbb{R}^d} r(x)\pi(x)\mathbb{1}_B(x) dx \\ &= \int_B \left\{ \int_{\mathbb{R}^d} K(x, \mathbf{y})\pi(x) dx \right\} d\mathbf{y} + \int_B r(\mathbf{y})\pi(\mathbf{y}) d\mathbf{y} \\ &\quad (\text{2 項目の積分の変数を } x \text{ から } \mathbf{y} \text{ に変更}) \\ &= \int_B \left\{ \int_{\mathbb{R}^d} K(x, \mathbf{y})\pi(x) dx + r(\mathbf{y})\pi(\mathbf{y}) \right\} d\mathbf{y} \end{aligned}$$

と書きかえられることに注意する. すると $\pi(x) dx$ が P に関して不変測度に

なるためには

$$\begin{aligned} \mu P = \mu &\iff \mu P(B) = \mu(B) \quad (\forall B \in \mathcal{B}) \\ &\iff \int_B \left\{ \int_{\mathbb{R}^d} K(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x}) d\mathbf{x} + r(\mathbf{y}) \pi(\mathbf{y}) \right\} d\mathbf{y} = \int_B \pi(\mathbf{y}) d\mathbf{y} \quad (\forall B \in \mathcal{B}) \\ &\iff \int_{\mathbb{R}^d} K(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x}) d\mathbf{x} + r(\mathbf{y}) \pi(\mathbf{y}) = \pi(\mathbf{y}) \quad (\forall \mathbf{y} \in \mathbb{R}^d) \\ &\iff \int_{\mathbb{R}^d} K(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x}) d\mathbf{x} = \pi(\mathbf{y}) (1 - r(\mathbf{y})) \quad (\forall \mathbf{y} \in \mathbb{R}^d) \end{aligned}$$

となる。^{eq:mp-2}(E.2) を考慮 (x を y に変更した式) に入れると

$$\begin{aligned} \int_{\mathbb{R}^d} K(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x}) d\mathbf{x} &= \pi(\mathbf{y}) \int_{\mathbb{R}^d} K(\mathbf{y}, \mathbf{x}) d\mathbf{x} \\ &\iff \int_{\mathbb{R}^d} K(\mathbf{x}, \mathbf{y}) \pi(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^d} K(\mathbf{y}, \mathbf{x}) \pi(\mathbf{y}) d\mathbf{x} \end{aligned} \quad (2.4)$$

を得る。この式を釣合方程式 (balance equation) とよぶ。 K が

$$\pi(\mathbf{x}) K(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}) K(\mathbf{y}, \mathbf{x}) \quad (\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d) \quad (2.5)$$

をみたすとき K は詳細的釣合方程式 (detailed balance equation) をみたすという。

Metropolis-Hastings アルゴリズムでは ^{eq:mp-4}(E.4) をみたす推移核 K を構成する。
 $q: \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$ は任意の関数で

$$\int q(\mathbf{x}, \mathbf{y}) d\mathbf{y} = 1$$

をみたす。この関数 q を用いて推移核 Q を

$$Q(\mathbf{x}, B) = \int_B q(\mathbf{x}, \mathbf{y}) d\mathbf{y} \quad (\forall B \in \mathcal{B})$$

で定める。 q が詳細的釣合方程式をみたしていれば

$$K(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}, \mathbf{y}), \quad r(\mathbf{x}) = 0$$

とする。みたしていなければ乗数倍で修正する。

$$K(\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{x}, \mathbf{y}) q(\mathbf{x}, \mathbf{y}). \quad (2.6)$$

ある $x, y \in \mathbb{R}^d$ に対して

$$\pi(y)q(y, x) < \pi(x)q(x, y)$$

ならば

$$\pi(y)\alpha(y, x)q(y, x) = \pi(x)\alpha(x, y)q(x, y)$$

となるように α を選ぶ. これは

$$\alpha(y, x) = 1, \quad \alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} < 1$$

とすればよいことがわかる. x と y を入れ替えることで

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\},$$

$$K(x, y) = \alpha(x, y)q(x, y)$$

とすると K は詳細的釣合方程式をみたすことがわかる.

候補分布 q が対称とする. すなわち

$$q(x, y) = q(y, x) \quad (\forall x, y \in \mathbb{R}^d)$$

とする. すると

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

となる. 対称な q の例としては

$$q(x, y) = g(x - y),$$

$$g: \mathbb{R}^d \rightarrow [0, \infty) \text{ で } g(x) = g(-x)$$

をみたすものがある.

以上の議論から繰り返し回数 K のアルゴリズムは以下ようになる.

Step 1: 初期値 $x_1 \in \mathbb{R}^d$ を選び, $k = 1$ とする.

Step 2: $y \sim q(x_k, y)$ で y を生成する.

Step 3: $\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x_k)}{\pi(x_k)q(x_k, y)}\right\}$ を計算する.

Step 4: $t \sim \text{Unif}(0, 1)$ で t を生成する.

Step 5 :

$$\alpha(\mathbf{x}_k, \mathbf{y}) \geq t \implies \mathbf{x}_{k+1} = \mathbf{y}.$$

$$\alpha(\mathbf{x}_k, \mathbf{y}) < t \implies \mathbf{x}_{k+1} = \mathbf{x}_k$$

とおく.

Step 6 : $k = K$ ならば \mathbf{x}_k を求めたい乱数として出力する. $k < K$ ならば $k \rightarrow k + 1$ として, Step 2 に戻る.

をみたまものがある.

:mcmc-4 注意 2.4. 上記アルゴリズムの Step 4 で得られた x_{k+1} の分布を確認する. 簡単のために $n = 1$ とする. $Y \sim q(y|x_k)$ かつ $U \sim \text{Unif}(0, 1)$ で Y と U は独立とする. このとき $t \in \mathbb{R}$ に対して

$$\begin{aligned} \frac{d}{dt} \mathbb{P}(x_{k+1} \leq t) &= \frac{d}{dt} \mathbb{P}(Y \leq t \mid \alpha(x_k, Y) \geq U) \\ &= \frac{\frac{d}{dt} \mathbb{P}(Y \leq t, \alpha(x_k, Y) \geq U)}{\mathbb{P}(\alpha(x_k, Y) \geq U)} \\ &= \frac{\frac{d}{dt} \int_{-\infty}^t \left\{ \int_0^{\alpha(x_k, y)} du \right\} q(x_k, y) dy}{\int_{-\infty}^{\infty} \left\{ \int_0^{\alpha(x_k, y)} du \right\} q(x_k, y) dy} \\ &= \frac{\frac{d}{dt} \int_{-\infty}^t \alpha(x_k, y) q(x_k, y) dy}{\int_{-\infty}^{\infty} \alpha(x_k, y) q(x_k, y) dy} \\ &= \frac{\alpha(x_k, t) q(x_k, t)}{\int_{-\infty}^{\infty} \alpha(x_k, y) q(x_k, y) dy} \end{aligned}$$

となり,

$$x_{k+1} \sim \frac{K(x_k, y)}{\int_{-\infty}^{\infty} k(x_k, y) dy}$$

となることがわかる. □

例 2.5. PDF

$$\pi(x) = Cx^{-4}(1 + |x|^3) \quad (x \in \mathbb{R})$$

から乱数を発生させることを考える. ただし

$$C^{-1} \int_{-\infty}^{\infty} x^{-4} (1 + |x|^3) dx$$

である. 候補 PDF として

$$Y | x_k \sim N(x_k, 1)$$

を採用する. すなわち

$$q(x, y) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(y-x)^2\right\}$$

である. ここで π は円周率である. すると

$$\alpha(x, y) = \min\left\{1, e^{-y^4+x^4} \frac{1+|y|^3}{1+|x|^3}\right\}$$

となる.

以上の議論から繰り返し回数 K のアルゴリズムは以下ようになる.

Step 1 : 初期値 $x_1 \in \mathbb{R}$ を選び, $k = 1$ とする.

Step 2 : $y \sim N(x_k, 1)$ で y を生成する.

Step 3 : $\alpha(x_k, y) = \min\left\{1, e^{-y^4+x_k^4} \frac{1+|y|^3}{1+|x_k|^3}\right\}$ を計算する.

Step 4 : $t \sim \text{Unif}(0, 1)$ で t を生成する.

Step 5 :

$$\alpha(x_k, y) \geq t \implies x_{k+1} = y.$$

$$\alpha(x_k, y) < t \implies x_{k+1} = x_k$$

とおく.

Step 6 : $k = K$ ならば x_k を求めたい乱数として出力する. $k < K$ ならば $k \rightarrow k + 1$ として, Step 2 に戻る.

□

3 ブートストラップ法

ブートストラップ法は推定量の標準誤差を推定する方法である。 F を未知の CDF とし、 $X_1, X_2, \dots, X_n \sim \text{i.i.d. } F$ とする。 未知の母数には依存しないある関数 g があって $S_n = g(X_1, X_2, \dots, X_n)$ とする。 すなわち、 S_n は統計量である。 このときに

$$T(F) = \mathbb{V}_F[S_n]$$

を推定したいとする。 一般にこの量は未知の分布 F に依存する。 たとえば $S_n = \bar{X}_n$ のとき、

$$\mathbb{V}_F[S_n] = \frac{\sigma^2}{n}$$

であり、未知の母数 σ に依存する。 ただし $\sigma^2 = \mathbb{E}[(X_1 - \mu)^2]$ 、 $\mu = \mathbb{E}[X_1]$ である。 $T(F)$ の推定量として、差し込み推定量を考える。

$$T(\hat{F}_n) = \mathbb{V}_{\hat{F}_n}[S_n].$$

ここで大数の法則を思い出す。 $B \in \mathbb{N}$ とし、 G を CDF とする。 $Y_1, Y_2, \dots, Y_B \sim \text{i.i.d. } G$ とする。 $B \rightarrow \infty$ のとき

$$\bar{Y}_B = \frac{1}{B} \sum_{j=1}^B Y_j \xrightarrow{P} \mathbb{E}[Y]$$

となる。 ただし $Y \sim G$ である。 さらに有限な期待値 $h(Y)$ をもつ関数 h に対して、 $B \rightarrow \infty$ のとき

$$\frac{1}{B} \sum_{j=1}^B h(Y_j) \xrightarrow{P} \mathbb{E}[h(Y)].$$

$\mathbb{V}_{\hat{F}_n}(S_n)$ をシミュレーションで近似することを考える。 $\mathbb{V}_{\hat{F}_n}(S_n)$ は「データが分布 \hat{F}_n に独立同一にしたがっているときの S_n の分散」と考えることができる。 $X_1^*, X_2^*, \dots, X_n^* \sim \text{i.i.d. } \hat{F}_n$ とし、 $S_n^* = g(X_1^*, X_2^*, \dots, X_n^*)$ を計算する。

- 現実の世界: $F \implies X_1, X_2, \dots, X_n \implies S_n = g(X_1, X_2, \dots, X_n)$.
- ブートストラップの世界: $\hat{F}_n \implies X_1^*, X_2^*, \dots, X_n^* \implies S_n^* = g(X_1^*, X_2^*, \dots, X_n^*)$.

では、 $X_1^*, X_2^*, \dots, X_n^*$ を \hat{F}_n からどのように発生させるのだろうか? \hat{F}_n は各 X_j ($j = 1, 2, \dots, n$) に確率 $1/n$ をのせたものであったことを思い出そう。

\hat{F}_n に従う確率変数 $X_j^* (j = 1, 2, \dots, n)$ はもとのデータ X_1, X_2, \dots, X_n から無作為抽出されたものと同じ分布になる

したがって $X_1^*, X_2^*, \dots, X_n^* \sim \text{i.i.d. } \hat{F}_n$ を発生させるためには, X_1, X_2, \dots, X_n からデータを復元抽出すればよい. 手続きをまとめると以下のようなになる.

- $X_1^*, X_2^*, \dots, X_n^* \sim \text{i.i.d. } \hat{F}_n$ を発生させる. X_1, X_2, \dots, X_n から復元抽出をするだけでよい.
- $S_n^* = g(X_1^*, X_2^*, \dots, X_n^*)$ を計算する.
- 上のふたつのステップを B 回繰り返して, $S_{n,1}^*, S_{n,2}^*, \dots, S_{n,B}^*$ を求める.

$$v_{\text{boot}} = \frac{1}{B} \sum_{j=1}^B \left(S_{n,j}^* - \frac{1}{B} \sum_{k=1}^B S_{n,k}^* \right)^2$$

を計算する.

このとき

$$V_F(S_n) \overset{\text{そんな色ない}}{\approx} V_{\hat{F}_n}(S_n) \overset{\text{小さい}}{\approx} v_{\text{boot}}.$$

$0 < \alpha < 1$ に対して, $z_{\alpha/2}$ を標準正規分布の両側 α 点とすれば, $T(F) = \mathbb{E}[S_n]$ とし

$$\hat{\text{se}}_{\text{boot}} := \sqrt{v_{\text{boot}}}$$

としたとき

$$S_n \pm z_{\alpha/2} \hat{\text{se}}_{\text{boot}}$$

は $T(F)$ の近似 $(1 - \alpha)\%$ 信頼区間となる. すなわち

$$\mathbb{P}(S_n - z_{\alpha/2} \hat{\text{se}}_{\text{boot}} \leq T(F) \leq S_n + z_{\alpha/2} \hat{\text{se}}_{\text{boot}}) \approx 1 - \alpha$$

となる.

第3章 最尤推定値の計算と EM アルゴリズム

ここでは最尤推定値を数値計算で求める方法を 3 つ紹介する.

1 Newton-Raphson 法

いま $g: \mathbb{R} \rightarrow \mathbb{R}$ を 2 回微分可能な関数とし, 方程式 $g(x) = 0$ をみたす解 $x = c$ をみつきたい. そのために c に近い x に対して, Taylor 展開をする.

$$0 = g(c) \approx g(x) + \dot{g}(x)(c - x)$$

ただし $\dot{g}(x) = dg/dx$ である. また「 \approx 」は「近い」と漠然と理解することにする.

$\dot{g}(x) \neq 0$ のとき $g(x) + \dot{g}(x)(x - c) \approx 0$ を c について解けば

$$c \approx x - \frac{g(x)}{\dot{g}(x)}$$

を得る. 初期値 x_0 を取り点列 $\{x_n\}_{n=1}^{\infty}$ を逐次的に

$$x_{n+1} = x_n - \frac{g(x_n)}{\dot{g}(x_n)}, \quad n = 0, 1, \dots \tag{3.1}$$

eq:newton1

で定義する. そして $|g(x_n)/\dot{g}(x_n)|$ が十分小さくなるまで操作を繰り返すとする.

区間 $I = [a, b]$ 上で $\ddot{g}(x) > 0$ かつ $g(a)g(b) < 0$ のとき, $g(x_0) > 0$ となる $x_0 \in (a, b)$ をひとつ見つければ, (3.1) で得られる点列 $\{x_n\}_{n=1}^{\infty}$ は I 上における零点 c に収束することが知られている¹. Newton-Raphson 法を用いて尤度方程式の解として定義される最尤推定値を求めることができる.

いま $\mathbf{X} = (X_1, X_2, \dots, X_n)$ は同時 PDF または PMF $p(\mathbf{x}|\theta)$ を持つとする. ただし $\theta \in \Theta \subset \mathbb{R}$ とする. $\mathbf{X} = \mathbf{x}$ が与えられたときの尤度関数を

¹杉浦「解析入門 I (東京大学出版会)」p.105 を参照.

$L_n(\theta | \mathbf{x}) = p(\mathbf{x} | \theta)$ と書くことにする。最尤推定値 $\hat{\theta}_n$ は尤度方程式

$$S_n(\theta | \mathbf{x}) = \frac{d}{d\theta} \log L_n(\theta | \mathbf{x}) = 0 \quad (3.2)$$

の解となる。さらに $m (m \in \mathbb{N})$ 回操作を行ったのちの θ の推定値を $\hat{\theta}_n^{(m)}(\mathbf{x})$ とする。^{eq:newton1}(3.1) から $H_n(\hat{\theta}_n^{(m)} | \mathbf{x}) \neq 0$ のとき

$$\hat{\theta}_n^{(m+1)}(\mathbf{x}) = \hat{\theta}_n^{(m)}(\mathbf{x}) + \frac{S_n(\hat{\theta}_n^{(m)}(\mathbf{x}) | \mathbf{x})}{H_n(\hat{\theta}_n^{(m)}(\mathbf{x}) | \mathbf{x})}$$

と書ける²。ただし

$$H_n(\theta | \mathbf{x}) = -\frac{d^2}{d\theta^2} \log L_n(\theta | \mathbf{x})$$

である。この操作を $\hat{\theta}_n^{(m+1)}(\mathbf{x})$ と $\hat{\theta}_n^{(m)}(\mathbf{x})$ との差が十分小さくなるまで繰り返す。

Newton-Raphson 法を用いるために初期推定値 $\hat{\theta}_n^{(0)}(\mathbf{x})$ が必要である。初期推定値に何をを用いるかによってアルゴリズムは収束したりしなかったりする。また尤度方程式 $S_n(\theta | \mathbf{x}) = 0$ が複数の解を持つ場合には、尤度方程式 ^{eq:equation}(3.2) の解は尤度関数の極小点、極大点、鞍馬点 (saddle point) に対応するので、 $\hat{\theta}_n^{(m)}(\mathbf{x})$ は最尤推定値とは異なる点に収束する可能性がある。収束先が最尤推定値と異なるかどうか不明な場合には、複数の初期値で試すとよい。また初期値として、別の推定値 (モーメント推定量などの別の推定量の実現値) を用いることもできる。

たとえば $\hat{\theta}_n^{(0)}(\mathbf{X})$ が θ の十分よい推定量ならば一段階推定量

$$\hat{\theta}_n^{(1)}(\mathbf{X}) = \hat{\theta}_n^{(0)}(\mathbf{X}) + \frac{S_n(\hat{\theta}_n^{(0)}(\mathbf{X}) | \mathbf{X})}{H_n(\hat{\theta}_n^{(0)}(\mathbf{X}) | \mathbf{X})}$$

は最尤推定量と同じ性質を漸近的には同じ性質をもつことが知られている。ただし、 $\hat{\theta}_n^{(0)}(\mathbf{X})$, $S(\hat{\theta}_n^{(0)} | \mathbf{X})$ と $H_n(\hat{\theta}_n^{(0)} | \mathbf{X})$ は $\hat{\theta}_n^{(0)}(\mathbf{x})$, $S_n(\hat{\theta}_n^{(0)} | \mathbf{x})$, $H_n(\hat{\theta}_n^{(0)} | \mathbf{x})$ の \mathbf{x} に \mathbf{X} を形式的に代入したものとす。正確に言えば $\sqrt{n}\{\hat{\theta}_n^{(0)}(\mathbf{X}) - \theta\}$ は正規分布に分布収束するならば $\sqrt{n}\{\hat{\theta}_n^{(1)}(\mathbf{X}) - \hat{\theta}_n^{(0)}(\mathbf{X})\} \xrightarrow{P} 0 (n \rightarrow \infty)$ となる。ただし $\hat{\theta}_n(\mathbf{X})$ は θ の最尤推定量である。

²(3.1) の右辺の第 2 項の「-」は H に入れていることに注意せよ。

例 3.1. この例では推定値のみを扱うので $\hat{\theta}_n(x)$ を $\hat{\theta}_n$ と書くことにする. $S(\theta|x)$, $H(\theta|x)$ も $S(\theta)$, $H(\theta)$ と書く. $\theta > 0$ とする. つぎの PDF を持つ Cauchy 分布からの大きさ n のランダム標本を X_1, X_2, \dots, X_n とする.

$$p(x|\theta) = \frac{1}{\pi\{1 + (x - \theta)^2\}} \quad (-\infty < x < \infty).$$

このとき実現値 x_1, x_2, \dots, x_n に対する対数尤度関数は

$$\log L_n(\theta) = - \sum_{i=1}^n \log\{1 + (x_i - \theta)^2\} - n \log \pi$$

となる. 最尤推定値 $\hat{\theta}_n$ は尤度方程式

$$S_n(\hat{\theta}_n) = \sum_{i=1}^n \frac{2(x_i - \hat{\theta}_n)}{1 + (x_i - \hat{\theta}_n)^2} = 0$$

の解である. $S_n(\theta)$ は θ の単調関数でないので, 与えられた (x_1, x_2, \dots, x_n) に対して尤度方程式は複数の解を持つ可能性がある. したがって適切な初期値 $\hat{\theta}^{(0)}$ を選ぶことが重要である. Cauchy 分布は $\mathbb{E}[X_1]$ が定義されないのので, 初期値として標本平均を用いるのは適当ではない. X_1 の分布は θ に関して対称であることに注目して, 標本中央値を初期値 $\hat{\theta}^{(0)}$ として用いる. これを用いて逐次的に $\hat{\theta}^{(m)}$ ($m = 1, 2, \dots$) を

$$\hat{\theta}_n^{(m)} = \hat{\theta}_n^{(m-1)} + \frac{S_n(\hat{\theta}_n^{(m-1)})}{H_n(\hat{\theta}_n^{(m-1)})}$$

で定める. ただし

$$H_n(\theta) = 2 \sum_{i=1}^n \frac{1 - (x_i - \theta)^2}{\{1 + (x_i - \theta)^2\}^2}$$

である. $\theta = 10$ の Cauchy 分布から標本の大きさ $n = 100$ のランダム標本に基づいて最尤推定値を求めた例が次である.

m	$\hat{\theta}_{100}^{(m)}$	$\log L_{100}(\hat{\theta}_{100}^{(m)}) + 100 \log(\pi)$
0	9.932387	11.95144
1	9.98055	11.9517
2	9.980323	11.9517
3	9.980323	11.9517

□

つぎに母数の次元が k の場合を考える. 母数 $\theta = (\theta_1, \theta_2, \dots, \theta_k)^\top$ が k 次元のとき, θ の最尤推定値 $\hat{\theta}_n(\boldsymbol{x})$ は尤度方程式 $S_n(\theta|\boldsymbol{x}) = \mathbf{0}$ の解として定義される. ただし

$$S_n(\theta|\boldsymbol{x}) = \left(\frac{\partial}{\partial\theta_1} \log L_n(\theta|\boldsymbol{x}), \frac{\partial}{\partial\theta_2} \log L_n(\theta|\boldsymbol{x}), \dots, \frac{\partial}{\partial\theta_k} \log L_n(\theta|\boldsymbol{x}) \right)^\top$$

である. さらに $H_n(\theta|\boldsymbol{x})$ を $p \times p$ の行列で, $i, j = 1, 2, \dots, k$ について $H_n(\theta|\boldsymbol{x})$ の (i, j) 成分は

$$H_{n,ij}(\theta|\boldsymbol{x}) = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log L_n(\theta)$$

で定義されるものとする. $H_n(\hat{\theta}_n^{(m-1)}|\boldsymbol{x})$ が正則のとき, m ($m \in \mathbb{N}$) 回目の逐次解 $\hat{\theta}_n^{(m)}(\boldsymbol{x})$ は

$$\hat{\theta}_n^{(m)}(\boldsymbol{x}) = \hat{\theta}_n^{(m-1)}(\boldsymbol{x}) + \{H_n(\hat{\theta}_n^{(m-1)}|\boldsymbol{x})\}^{-1} S_n(\hat{\theta}_n^{(m-1)}|\boldsymbol{x}) \quad (3.3)$$

eq:newt

で定義される.

2 Fisher のスコア法

Newton-Raphson アルゴリズムの簡単な修正として Fisher のスコアアルゴリズムがある. Fisher のスコアアルゴリズムは (3.3) 中の $H_n(\theta|\boldsymbol{x})$ の代わりに

$$H_n^*(\theta) = \mathbb{E}_\theta[H_n(\theta|\boldsymbol{X})] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta\partial\theta^\top} \log L_n(\theta|\boldsymbol{X}) \right]$$

である. ただし

$$\frac{\partial^2}{\partial\theta\partial\theta^\top} \log L_n(\theta|\boldsymbol{X})$$

の (i, j) 成分は

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log L_n(\theta|\boldsymbol{X}) \quad (i, j = 1, 2, \dots, k)$$

である.

したがって m ($m \in \mathbb{N}$) 回目の逐次解 $\hat{\theta}_n^{(m-1)}(x)$ は

$$\hat{\theta}_n^{(m)}(x) = \hat{\theta}_n^{(m-1)}(x) + [H_n^*(\hat{\theta}_n^{(m-1)})]^{-1} S_n(\hat{\theta}_n^{(m-1)} | x) \quad (3.4)$$

eq:fish

で定義される.

例 3.2. (例 [3.1](#) の続き) ex:cauchy-newton

$$H_n(\theta) = 2 \sum_{i=1}^n \frac{1 - (x_i - \theta)^2}{\{1 + (x_i - \theta)^2\}^2}$$

から

$$H_n^*(\theta) = \frac{2n}{\pi} \int_{-\infty}^{\infty} \frac{1 - (x - \theta)^2}{\{1 + (x - \theta)^2\}^3} dx = \frac{n}{2} \quad (3.5)$$

eq:keis

から Fisher のスコアアルゴリズムは

$$\hat{\theta}_n^{(m)} = \hat{\theta}_n^{(m-1)} + \frac{4}{n} \sum_{i=1}^n \frac{x_i - \hat{\theta}_n^{(m-1)}}{1 + (x_i - \hat{\theta}_n^{(m-1)})^2} \quad (m = 1, 2, \dots)$$

となる. 最後に [\(3.5\)](#) の計算をする. $z = x - \theta$ とおく. さらに $y = \tan \gamma$ とおけば

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{1 - y^2}{(1 + y^2)^3} dy &= \int_{-\infty}^{\infty} \frac{2}{(1 + y^2)^3} dy - \int_{-\infty}^{\infty} \frac{1}{(1 + y^2)^2} dy \\ &= 2 \int_{-\pi/2}^{\pi/2} \frac{1}{(1 + \tan^2 \gamma)^3} \frac{1}{\cos^2 \gamma} d\gamma - \int_{-\pi/2}^{\pi/2} \frac{1}{(1 + \tan^2 \gamma)^2} \frac{1}{\cos^2 \gamma} d\gamma \\ &= 2 \int_{-\pi/2}^{\pi/2} \cos^4 \gamma d\gamma - \int_{-\pi/2}^{\pi/2} \cos^2 \gamma d\gamma, \\ &= 2 \left[\frac{\cos 4\gamma + 1}{8} + \cos 2\gamma + \frac{1}{4} \right]_{-\pi/2}^{\pi/2} - \left[\frac{\cos 2\gamma + 1}{2} \right]_{-\pi/2}^{\pi/2} \\ &= \frac{\pi}{4} \end{aligned}$$

となることがわかる.

3 EM アルゴリズム

$(\mathcal{X}, \mathcal{A}, \mu)$ を測度空間とする。ここで μ は σ 有限な測度³とする。 \mathcal{X} 値確率変数 X は母数 $\theta (\theta \in \Theta)$ の確率測度 P_θ を持つ。さらに P_θ は μ に関する PDF $p(\cdot | \theta)$ をもつとする。

$$P_\theta(A) = \int_A p(\mathbf{x} | \theta) \mu(d\mathbf{x}) \quad (A \in \mathcal{A}).$$

いま $(\mathcal{X}, \mathcal{A})$ は隠れた空間で X のすべてを観測できないとする。実際にはある可測空間 $(\mathcal{Y}, \mathcal{B})$ と可測関数 $T : \mathcal{X} \mapsto \mathcal{Y}$ が存在して $Y = T(X)$ のみが観測できるとする。 T は $(\mathcal{Y}, \mathcal{B})$ 上の測度を誘導する。

$$Q_\theta(B) = P_\theta \circ T^{-1}(B) = P_\theta(T^{-1}(B)) \quad (B \in \mathcal{B})$$

とする。さらに ν を $(\mathcal{Y}, \mathcal{B})$ 上の σ 有限な測度としたとき、確率測度 Q_θ は測度 ν に関して PDF $q(\cdot | \theta)$ を持つとする。

$$Q_\theta(B) = \int_B q(\mathbf{y} | \theta) \nu(d\mathbf{y}) \quad (B \in \mathcal{B})$$

である。

EM アルゴリズムは観測 $Y = \mathbf{y}$ が与えられたとき、 θ の関数として $q(\mathbf{y} | \theta)$ を最大化することで θ の最尤推定値を求める方法である。

EM アルゴリズムは以下のように行う。 \mathbf{y} が与えられたとき θ の初期推定値の $\hat{\theta}^{(0)}(\mathbf{y})$ から始める。 $\hat{\theta}^{(0)}(\mathbf{y})$ より $P_{\hat{\theta}^{(0)}}$ と $Q_{\hat{\theta}^{(0)}} = P_{\hat{\theta}^{(0)}} \circ T^{-1}$ が初期の推定された確率測度となる。

E 段階 (E-Step): 各 $\theta \in \Theta$ に対し、条件付き期待値

$$\phi_1(\theta | \mathbf{y}) = \mathbb{E}_{P_{\hat{\theta}^{(0)}}} [\log p(\mathbf{X} | \theta) | T(\mathbf{X}) = \mathbf{y}] \quad (3.6)$$

を求める。ただし $\mathbb{E}_{P_{\hat{\theta}^{(0)}}}[\cdot]$ は $p(\mathbf{x} | \hat{\theta}^{(0)})$ に関して期待値を取ったものである。同様に $P_{\hat{\theta}^{(0)}}$ は $p(\mathbf{x} | \hat{\theta}^{(0)})$ のもとでの確率分布である。 $\phi_1(\theta)$ は \mathbf{y} の関数になることに注意する。 $P_{\hat{\theta}^{(0)}}(T(\mathbf{X}) = \mathbf{y}) > 0$ ⁴の場合には、 $T(\mathbf{X}) = \mathbf{y}$ が与えられた

³ある部分集合の列 $\{G_n\}_{n=1}^\infty$, $G_n \in \mathcal{A}$ で $\cup_{n=1}^\infty G_n = \mathcal{X}$ かつ各 n に対して $\mu(G_n) < \infty$ なるものが存在することである。

⁴ $P_{\hat{\theta}^{(0)}}(T(\mathbf{X}) = \mathbf{y}) := P_{\hat{\theta}^{(0)}}(\{\mathbf{x} \in \mathcal{X}; T(\mathbf{x}) = \mathbf{y}\})$ の意味である。

ときの X の条件付分布を求め, それに関して関数 $x \mapsto \log p(x|\theta)$ の期待値を求めればよい. $P_{\hat{\theta}^{(0)}}(T(\mathbf{X}) = \mathbf{y}) = 0$ のときは注意が必要であるが (3.6) の正当化は可能である. 以後は簡単のために $P_{\hat{\theta}^{(0)}}(T(\mathbf{X}) = \mathbf{y}) > 0$ の場合を考える.

M 段階 (M-Step): θ に関して

$$\phi_1(\theta|\mathbf{y})$$

を最大化する. 最大を与える点 (存在すれば) を $\hat{\theta}^{(1)}(\mathbf{y})$ とおく. つぎに $P_{\hat{\theta}^{(0)}}$ の代わりに $P_{\hat{\theta}^{(1)}}$ を用いる.

E 段階 (E-Step): 各 $\theta \in \Theta$ に対して

$$\phi_2(\theta|\mathbf{y}) = \mathbb{E}_{\hat{\theta}^{(1)}}[\log p(\mathbf{X}|\theta) | T(\mathbf{X}) = \mathbf{y}]$$

を求める.

M 段階 (M-Step): θ に関して

$$\phi_2(\theta|\mathbf{y})$$

を最大化する. 最大を与える点 (存在すれば) を $\hat{\theta}^{(2)}(\mathbf{y})$ とおく.

一般には m 段階 ($m = 1, 2, \dots$) は以下ようになる.

E 段階 (E-Step): 各 $\theta \in \Theta$ に対し, 条件付き期待値

$$\phi_m(\theta|\mathbf{y}) = \mathbb{E}_{\hat{\theta}^{(m-1)}}[\log p(\mathbf{X}|\theta) | T(\mathbf{X}) = \mathbf{y}] \quad (3.7)$$

を求める.

M 段階 (M-Step): θ に関して

$$\phi_m(\theta|\mathbf{y})$$

を最大化する. 最大を与える点 (存在すれば) を $\hat{\theta}^{(m)}(\mathbf{y})$ とおく.

この操作を $\hat{\theta}^{(m)}(\mathbf{y})$ が $\hat{\theta}^{(m-1)}(\mathbf{y})$ とほとんど変化がなくなるまで繰り返す.

例 3.3. $\theta > 0$ とする. X_1, X_2, \dots, X_n は独立同一に指数分布

$$p(x|\theta) = \frac{1}{\theta} e^{-x/\theta} \mathbb{1}_{(0, \infty)}(x), \quad \mathbb{1}_{(0, \infty)}(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x \leq 0) \end{cases}$$

に従う⁵とする。しかし各 X_i ($i = 1, 2, \dots, n$) は直接観測されず、各 X_i の整数部分のみが観測されるとする。すなわち $Y_i = \lfloor X_i \rfloor$ である。 Y_1, Y_2, \dots, Y_n の観測に基づいて θ の最尤推定値を求めよう。

この場合、 $\mathcal{X} = (0, \infty)^n$ 、 \mathcal{A} は $(0, \infty)^n$ 上の Borel 可測集合族⁶であり、 P_θ は \mathbb{R}^n 上の Lebesgue 測度⁷に関する PDF

$$p(\mathbf{x}|\theta) = p(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^n p(x_i|\theta) = \theta^{-n} \exp\left(-\sum_{i=1}^n x_i/\theta\right) \quad (3.8)$$

eq:ex3-

を持つ。また $\mathcal{Y} = \{0, 1, \dots\}^n$ で、 \mathcal{B} は \mathcal{Y} のすべての部分集合の集まりからなる σ 加法族である。さらに関数 $T: \mathcal{X} \mapsto \mathcal{Y}$ は

$$T(\mathbf{x}) = (\lfloor x_1 \rfloor, \lfloor x_2 \rfloor, \dots, \lfloor x_n \rfloor)$$

で定義される。

いま θ のもとではかった確率を P_θ と書けば

$$P_\theta(Y_i = y) = \int_y^{y+1} p(x|\theta) dx = e^{-y/\theta}(1 - e^{-1/\theta})$$

となる。このことより

$$q(\mathbf{y}|\theta) = q(y_1, y_2, \dots, y_n|\theta) = \prod_{i=1}^n e^{-y_i/\theta}(1 - e^{-1/\theta})$$

となることがわかる。直接 $q(\mathbf{y}|\theta)$ を θ に関して最大化して θ の最尤推定値を求めることはできるが、EM アルゴリズムを用いるとどうなるかを観てみよう。

まず $\lfloor X_i \rfloor = y$ が与えられたとき、 θ のもとでの X_i の条件付 PDF を求める。

$$k_\theta(x|y) = \frac{P_\theta(X_i = x, \lfloor X_i \rfloor = y)}{P_\theta(\lfloor X_i \rfloor = y)} = \frac{\theta^{-1} e^{-x/\theta} \mathbb{1}_{[y, y+1)}(x)}{e^{-y/\theta}(1 - e^{-1/\theta})}$$

⁵指数分布の PDF は

$$p(x|\lambda) = \lambda e^{-\lambda x} \mathbb{1}_{(0, \infty)}(x)$$

であるが、母数を $\lambda = 1/\theta$ としている。

⁶ $(0, \infty)^n$ の開集合族を含む最小の σ 加法族のこと。

⁷ dx と書くことにする。また \mathbb{R} 上の Lebesgue 測度は dx と書く。

これより

$$\begin{aligned}
\phi_1(\theta) &= \mathbb{E}_{\hat{\theta}^{(0)}} [\log p(\mathbf{X} | \theta) | T(\mathbf{X}) = \mathbf{y}] \\
&= \mathbb{E}_{\hat{\theta}^{(0)}} \left[-\frac{1}{\theta} \sum_{i=1}^n X_i - n \log \theta \mid T(\mathbf{X}) = \mathbf{y} \right] \quad (\text{eq:ex3-1を代入}) \\
&= -\frac{1}{\theta} \mathbb{E}_{\hat{\theta}^{(0)}} \left[\sum_{i=1}^n X_i \mid T(\mathbf{X}) = \mathbf{y} \right] - n \log \theta
\end{aligned}$$

と

$$\begin{aligned}
\mathbb{E}_{P_{\hat{\theta}^{(0)}}} [X_i | T(\mathbf{X}) = \mathbf{y}] &= \frac{1}{\hat{\theta}^{(0)} e^{-y/\hat{\theta}^{(0)}} (1 - e^{-1/\hat{\theta}^{(0)}})} \int_y^{y+1} x e^{-x/\hat{\theta}^{(0)}} dx \\
&= \frac{1}{e^{-y/\hat{\theta}^{(0)}} (1 - e^{-1/\hat{\theta}^{(0)}})} \left\{ [-x e^{-x/\hat{\theta}^{(0)}}]_y^{y+1} + \int_y^{y+1} e^{-x/\hat{\theta}^{(0)}} dx \right\} \\
&= \frac{1}{e^{-y/\hat{\theta}^{(0)}} (1 - e^{-1/\hat{\theta}^{(0)}})} \left\{ -(y+1) e^{-(y+1)/\hat{\theta}^{(0)}} + y e^{-y/\hat{\theta}^{(0)}} \right. \\
&\quad \left. - \hat{\theta}^{(0)} e^{-(y+1)/\hat{\theta}^{(0)}} + \hat{\theta}^{(0)} e^{-y/\hat{\theta}^{(0)}} \right\} \\
&= \frac{1}{e^{-y/\hat{\theta}^{(0)}} (1 - e^{-1/\hat{\theta}^{(0)}})} \left\{ -(y + \hat{\theta}^{(0)}) e^{-y/\hat{\theta}^{(0)}} (1 - e^{-1/\hat{\theta}^{(0)}}) - e^{-(y+1)/\hat{\theta}^{(0)}} \right\} \\
&= - \left(y - \frac{1}{e^{1/\hat{\theta}^{(0)}} - 1} + \hat{\theta}^{(0)} \right)
\end{aligned}$$

となる。よって E 段階は

$$\phi_1(\theta | \mathbf{y}) = n \left(-\log \theta + \frac{1}{\theta(e^{1/\hat{\theta}^{(0)}} - 1)} - \frac{\bar{y}_n + \hat{\theta}^{(0)}}{\theta} \right)$$

となる。ただし $\bar{y}_n = (1/n) \sum_{i=1}^n y_i$ である。次に M 段階は上の式を θ に関して最大化する：

$$\hat{\theta}^{(1)}(\mathbf{y}) = \arg \max_{\theta} \phi_1(\theta | \mathbf{y}) = \hat{\theta}^{(0)}(\mathbf{y}) + \bar{y}_n - \frac{1}{e^{1/\hat{\theta}^{(0)}(\mathbf{y})} - 1}$$

となる⁸。したがって EM アルゴリズムは

$$\hat{\theta}^{(m)}(\mathbf{y}) = \arg \max_{\theta} \phi_m(\theta | \mathbf{y}) = \hat{\theta}^{(m-1)}(\mathbf{y}) + \bar{y}_n - \frac{1}{e^{1/\hat{\theta}^{(m-1)}(\mathbf{y})} - 1}$$

⁸ $\arg \max_{\theta} \phi(\theta)$ は関数 $\theta \mapsto \phi_{\theta}$ を最大にする点の集合を表す。したがって厳密には $\hat{\theta}^{(1)} \in \arg \max_{\theta} \phi(\theta)$ と書くべきである。たとえば $\arg \max_{0 \leq \theta < 4\pi} \sin \theta = \{\pi/2, 5\pi/2\}$ となる。

で与えられる.

□

つぎに EM アルゴリズムがどうしてうまく働くかを観る. $Y = \mathbf{y}$ が与えられたとき $\hat{\theta}$ を θ の最尤推定値とし Θ の内部上で関数

$$\Theta \ni \theta \mapsto q(\mathbf{y}|\theta) \in \mathbb{R}$$

は微分可能とする. このとき $\hat{\theta}$ は関数 $\theta \mapsto q(\mathbf{y}|\theta)$ を最大にする点なので,

$$\left. \frac{\partial}{\partial \theta} q(\mathbf{y}|\theta) \right|_{\theta=\hat{\theta}} = 0$$

である. ただし x が変数の関数 $f(x)$ に x_0 を代入することを $f(x)|_{x=x_0}$ と書いている.

いま $\hat{\theta}^{(\infty)}(\mathbf{y})$ を Θ の内点とし, EM アルゴリズムの収束先とする. このとき, $\hat{\theta}^{(\infty)}(\mathbf{y})$ は関数

$$\theta \mapsto \mathbb{E}_{\hat{\theta}^{(\infty)}} [\log p(\mathbf{X}|\theta) | T(\mathbf{X}) = \mathbf{y}] \quad (3.9)$$

eq:2.11

を最大化する. (3.9) は Θ の内部で微分可能とし, 期待値と微分記号の交換が可能とすれば, $\theta = \hat{\theta}^{(\infty)}(\mathbf{y})$ において

$$\begin{aligned} \left. \frac{\partial}{\partial \theta} \mathbb{E}_{\hat{\theta}^{(\infty)}} [\log p(\mathbf{X}|\theta) | T(\mathbf{X}) = \mathbf{y}] \right|_{\theta=\hat{\theta}^{(\infty)}} &= \mathbb{E}_{\hat{\theta}^{(\infty)}} \left[\left. \frac{\partial}{\partial \theta} \log p(\mathbf{X}|\theta) \right| T(\mathbf{X}) = \mathbf{y} \right] \Big|_{\theta=\hat{\theta}^{(\infty)}} \\ &= 0 \end{aligned} \quad (3.10)$$

eq:ema1

となる. ここで

$$S(\theta|\mathbf{x}) = \frac{\partial}{\partial \theta} \log p(\mathbf{x}|\theta)$$

とおけば, 適当な仮定のもとで

$$\begin{aligned} \mathbb{E}_{\theta} [S(\theta|\mathbf{X}) | T(\mathbf{X}) = \mathbf{y}] &= \int_{T(\mathbf{x})=\mathbf{y}} S(\theta|\mathbf{x}) p(\mathbf{x}|\theta) \mu(d\mathbf{x}) \\ &= \int_{T(\mathbf{x})=\mathbf{y}} \left(\frac{\partial}{\partial \theta} p(\mathbf{x}|\theta) \right) \mu(d\mathbf{x}) \\ &= \frac{\partial}{\partial \theta} \int_{T(\mathbf{x})=\mathbf{y}} p(\mathbf{x}|\theta) \mu(d\mathbf{x}) \\ &= \frac{\partial}{\partial \theta} q(\mathbf{y}|\theta) \end{aligned}$$

となる。よって、 $q(\mathbf{y}|\theta) > 0$ となる \mathbf{y} に対して、^{eq:emalgo-3}(3.10) は

$$\frac{\partial}{\partial \theta} \log q(\mathbf{y}|\theta) \Big|_{\theta=\hat{\theta}^{(\infty)}} = 0 \quad (3.11)$$

を意味する。したがって、最尤推定値が一意に存在するならば、 $\hat{\theta} = \hat{\theta}^{(\infty)}$ となる。 $\theta = \hat{\theta}^{(\infty)}$ なる解をもつ方程式

$$\mathbb{E}_{\theta}\{S(\theta|\mathbf{X})|T(\mathbf{X}) = \mathbf{y}\} = 0$$

を自己一致方程式 (self-consistency equation) という。

以上の議論からつぎの場合には、EM アルゴリズムはうまく機能されるかは保証されていない。

1. 最大値を与える点が Θ の内点に含まれない。
2. 尤度関数とその最大を取る点で微分可能ではない。
3. スコア方程式 ^{eq:emalgo-4}(3.11) が複数の解を持ち、そのいくつかは尤度関数を最大にしない。

最後に EM アルゴリズムの各段階で、対数尤度関数 $\log q(\mathbf{y}|\theta)$ は非減少であることを示す：

$$\log q(\mathbf{y}|\hat{\theta}^{(m)}(\mathbf{y})) > \log q(\mathbf{y}|\hat{\theta}^{(m-1)}(\mathbf{y})), \quad m = 0, 1, \dots \quad (3.12)$$

$T(\mathbf{X}) = \mathbf{y}$ を与えたとき、 \mathbf{X} の条件付 PDF $k_{\theta}(\mathbf{x}|\mathbf{y})$ は次で与えられる：

$$k_{\theta}(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}|\theta)}{q(\mathbf{y}|\theta)} \mathbb{1}_{T^{-1}(\mathbf{y})}(\mathbf{x})$$

$T(\mathbf{x}) = \mathbf{y}$, $p(\mathbf{x}|\theta) > 0$, $q(\mathbf{y}|\theta) > 0$ の場合、

$$\log q(\mathbf{y}|\theta) = \log p(\mathbf{x}|\theta) - \log k_{\theta}(\mathbf{x}|\mathbf{y})$$

となる。

以下では最尤推定値の候補は $q(\mathbf{y}|\theta) > 0$ をみたしていなければいけないので、 $q(\mathbf{y}|\theta) > 0$ を仮定して議論を進める。これは、そうでなければ、 $\log q(\mathbf{y}|\theta) > 0$ を最大にしないことからわかる。

各 m に対し

$$\begin{aligned}
 \log q(\mathbf{y} | \theta) &= \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log q(\mathbf{y} | \theta) | T(\mathbf{X}) = \mathbf{y}] \\
 &= \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log p(\mathbf{X} | \theta) - \log k_{\theta}(\mathbf{X} | \mathbf{y}) | T(\mathbf{X}) = \mathbf{y}] \\
 &= \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log p(\mathbf{X} | \theta) | T(\mathbf{X}) = \mathbf{y}] - \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log k_{\theta}(\mathbf{X} | \mathbf{y}) | T(\mathbf{X}) = \mathbf{y}]
 \end{aligned} \tag{3.13}$$

となる。上式の最右辺の各項を別々に評価していく。

まず, $\hat{\theta}^{(m)}(\mathbf{y})$ の定義から

$$\mathbb{E}_{\hat{\theta}^{(m-1)}} [\log p(\mathbf{X} | \hat{\theta}^{(m)}) | T(\mathbf{X}) = \mathbf{y}] - \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log p(\mathbf{X} | \hat{\theta}^{(m-1)}) | T(\mathbf{X}) = \mathbf{y}] \geq 0 \tag{3.14}$$

がわかる。

次に, (3.13) の最右辺の 2 項目を評価する：

$$\begin{aligned}
 &\mathbb{E}_{\hat{\theta}^{(m-1)}} [\log k_{\hat{\theta}^{(m)}}(\mathbf{X} | \mathbf{y}) | T(\mathbf{X}) = \mathbf{y}] - \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log k_{\hat{\theta}^{(m-1)}}(\mathbf{X} | \mathbf{y}) | T(\mathbf{X}) = \mathbf{y}] \\
 &= \mathbb{E}_{\hat{\theta}^{(m-1)}} \left[\log \frac{k_{\hat{\theta}^{(m)}}(\mathbf{X} | \mathbf{y})}{k_{\hat{\theta}^{(m-1)}}(\mathbf{X} | \mathbf{y})} \middle| T(\mathbf{X}) = \mathbf{y} \right] \\
 &\leq \log \mathbb{E}_{\hat{\theta}^{(m-1)}} \left[\frac{k_{\hat{\theta}^{(m)}}(\mathbf{X} | \mathbf{y})}{k_{\hat{\theta}^{(m-1)}}(\mathbf{X} | \mathbf{y})} \middle| T(\mathbf{X}) = \mathbf{y} \right]
 \end{aligned} \tag{3.15}$$

となる。最後の不等号は Jensen の不等式よりわかる。いま,

$$g(\mathbf{y}) = \mathbb{E}_{\hat{\theta}^{(m-1)}} \left(\frac{k_{\hat{\theta}^{(m)}}(\mathbf{X} | \mathbf{y})}{k_{\hat{\theta}^{(m-1)}}(\mathbf{X} | \mathbf{y})} \middle| T(\mathbf{X}) = \mathbf{y} \right)$$

とおく。条件付き期待値の定義から任意の $B \in \mathcal{B}$ に対し

$$\int_B g(\mathbf{y}) dQ_{\hat{\theta}^{(m-1)}}(\mathbf{y}) = \int_{T^{-1}(B)} \frac{k_{\hat{\theta}^{(m)}}(\mathbf{X} | \mathbf{y})}{k_{\hat{\theta}^{(m-1)}}(\mathbf{X} | \mathbf{y})} dP_{\hat{\theta}^{(m-1)}}(\mathbf{x}) \tag{3.16}$$

となる。しかし

$$\frac{k_{\hat{\theta}^{(m)}}(\mathbf{x} | \mathbf{y})}{k_{\hat{\theta}^{(m-1)}}(\mathbf{x} | \mathbf{y})} = \frac{p(\mathbf{x} | \hat{\theta}^{(m)}(\mathbf{y}))}{q(T(\mathbf{x}) | \hat{\theta}^{(m)}(\mathbf{y}))} \cdot \frac{q(T(\mathbf{x}) | \hat{\theta}^{(m-1)}(\mathbf{y}))}{p(\mathbf{x} | \hat{\theta}^{(m-1)}(\mathbf{y}))}$$

に注意すれば ^{eq:emalgo-8}(3.16) の右辺は

$$\begin{aligned}
& \int_{T^{-1}(B)} \frac{k_{\hat{\theta}^{(m)}}(\mathbf{X}|\mathbf{y})}{k_{\hat{\theta}^{(m-1)}}(\mathbf{X}|\mathbf{y})} dP_{\hat{\theta}^{(m-1)}} \\
&= \int_{T^{-1}(B)} \frac{p(\mathbf{x}|\hat{\theta}^{(m)}(\mathbf{y}))}{q(T(\mathbf{x})|\hat{\theta}^{(m)}(\mathbf{y}))} q(T(\mathbf{x})|\hat{\theta}^{(m-1)}(\mathbf{y})) \mu(d\mathbf{x}) \\
&= \int_{T^{-1}(B)} \frac{q(T(\mathbf{x})|\hat{\theta}^{(m-1)}(\mathbf{y}))}{q(T(\mathbf{x})|\hat{\theta}^{(m)}(\mathbf{y}))} dP_{\hat{\theta}^{(m)}}(\mathbf{x}) \\
&= \int_B \frac{q(\mathbf{y}|\hat{\theta}^{(m-1)}(\mathbf{y}))}{q(\mathbf{y}|\hat{\theta}^{(m)}(\mathbf{y}))} dQ_{\hat{\theta}^{(m)}}(\mathbf{y}) \\
&= \int_B 1 \cdot dQ_{\hat{\theta}^{(m-1)}}(\mathbf{y})
\end{aligned}$$

となる。これより

$$g(\mathbf{y}) = \mathbb{E}_{\hat{\theta}^{(m-1)}} \left[\frac{k_{\hat{\theta}^{(m)}}(\mathbf{X}|\mathbf{y})}{k_{\hat{\theta}^{(m-1)}}(\mathbf{X}|\mathbf{y})} \middle| T(\mathbf{X}) = \mathbf{y} \right] = 1, \quad a.e. \quad Q_{\hat{\theta}^{(m-1)}}$$

となる。この式と ^{eq:emalgo-6a}(3.15) から

$$\mathbb{E}_{\hat{\theta}^{(m-1)}} [\log k_{\hat{\theta}^{(m)}}(\mathbf{X}|\mathbf{y}) | T(\mathbf{X}) = \mathbf{y}] - \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log k_{\hat{\theta}^{(m-1)}}(\mathbf{X}|\mathbf{y}) | T(\mathbf{X}) = \mathbf{y}] \leq 0 \quad (3.17)$$

がわかる。よって ^{eq:emalgo-8}(3.14) と ^{eq:emalgo-9}(3.17) から

$$\begin{aligned}
\log q(\mathbf{y}|\hat{\theta}^{(m)}) &= \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log p(\mathbf{X}|\hat{\theta}^{(m)}) | T(\mathbf{X}) = \mathbf{y}] \\
&\quad - \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log k_{\hat{\theta}^{(m)}}(\mathbf{X}|\mathbf{y}) | T(\mathbf{X}) = \mathbf{y}] \\
&= \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log p(\mathbf{X}|\hat{\theta}^{(m-1)}) | T(\mathbf{X}) = \mathbf{y}] \\
&\quad - \mathbb{E}_{\hat{\theta}^{(m-1)}} [\log k_{\hat{\theta}^{(m-1)}}(\mathbf{X}|\mathbf{y}) | T(\mathbf{X}) = \mathbf{y}] \\
&= \log q(\mathbf{y}|\hat{\theta}^{(m-1)})
\end{aligned}$$

となるので ^{eq:emalgo-5}(3.12) は示せた。

第4章 線型回帰モデル

1 線型単回帰モデルと最小 2 乗推定量

線型単回帰モデル (simple linear regression model) を考える. $n \in \mathbb{N}$ とし, n 個の観測の組を

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

とする. 各 y_j ($j = 1, 2, \dots, n$) は

$$y_j = \alpha^* + \beta^* x_j + \epsilon_j \quad (j = 1, 2, \dots, n)$$

なる線型構造を持って分布しているとする. ここで $\alpha^* \in \mathbb{R}$ を y 切片項, $\beta^* \in \mathbb{R}$ を回帰係数と呼ぶ. これらは未知の母数 (パラメータ) と仮定する. y_j を従属変数 (応答変数), x_j を独立変数 (説明変数) と呼ぶ. $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ は独立同一分布に従う確率変数列である.

この節において単回帰モデルに以下の仮定 (1) ~ (4) をおくことにする.

- (1) 説明変数 x_1, x_2, \dots, x_n は確率変数ではなく与えられた定数.
- (2) $\mathbb{E}[\epsilon_j] = 0$ ($j = 1, 2, \dots, n$).
- (3) $\mathbb{E}[\epsilon_j \epsilon_\ell] = 0$ ($j \neq \ell$).
- (4) $\mathbb{V}[\epsilon_j] = \sigma^2$. ただし分散 σ^2 ($\sigma > 0$) は未知とする.

未知の母数 α^*, β^* を推定するために最小 2 乗法を用いる. 推定量導出のために

$$h(\alpha, \beta) := \sum_{j=1}^n \{y_j - (\alpha + \beta x_j)\}^2$$

を考える. α, β に関して h の最小化問題を考える. 最小化問題の解 (存在すれば) を最小 2 乗推定量ということにする.

簡単のために,

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{j=1}^n y_j; & \bar{x} &= \frac{1}{n} \sum_{j=1}^n x_j; \\ Q_{xy} &= \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y}); \\ Q_{xx} &= \sum_{j=1}^n (x_j - \bar{x})^2; & Q_{yy} &= \sum_{j=1}^n (y_j - \bar{y})^2\end{aligned}$$

なる記号を導入する. 以下では $Q_{xx} \neq 0$ と仮定する.

関数 h を変形すれば

$$\begin{aligned}h(\alpha, \beta) &= Q_{xx} \left\{ \beta - \frac{Q_{xy}}{Q_{xx}} \right\}^2 + n \{ \bar{y} - \alpha - \beta \bar{x} \}^2 \\ &\quad + \frac{Q_{xx} Q_{yy} - Q_{xy}^2}{Q_{xx}}\end{aligned}\tag{4.1}$$

eq:2-1

と書ける.

$$\hat{\alpha} := \bar{y} - \hat{\beta} \bar{x}; \quad \hat{\beta} := \frac{Q_{xy}}{Q_{xx}}$$

とおく. 上の式より $h(\alpha, \beta)$ は $\alpha = \hat{\alpha}$, $\beta = \hat{\beta}$ のときに最小値をとる.

最小 2 乗推定量 (値) を用いて, 回帰直線 $y = \hat{\alpha} + \hat{\beta}x$ を引くことができる. 回帰直線上の点 $(x_j, \hat{\alpha} + \hat{\beta}x_j)$ ($j = 1, 2, \dots, n$) と観測 (x_j, y_j) との差

$$e_j := y_j - (\hat{\alpha} + \hat{\beta}x_j) \quad (j = 1, 2, \dots, n)$$

を残差といい

$$\text{RSS} := \sum_{j=1}^n e_j^2 = \sum_{j=1}^n \{ y_j - (\hat{\alpha} + \hat{\beta}x_j) \}^2$$

を残差平方和という. $n \geq 3$ のとき未知の分散 σ^2 を

$$\hat{\sigma}^2 = \frac{1}{n-2} \text{RSS}$$

で推定することができる.

- 命題 4.1. (1) $\mathbb{E}[\hat{\beta}] = \beta^*$; $\mathbb{V}[\hat{\beta}] = \frac{\sigma^2}{Q_{xx}}$.
 (2) $\mathbb{E}[\hat{\alpha}] = \alpha^*$; $\mathbb{V}[\hat{\alpha}] = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{Q_{xx}} \right\}$.
 (3) $\text{COV}[\hat{\alpha}, \hat{\beta}] = -\frac{\bar{x}\sigma^2}{Q_{xx}}$.
 (4) $\mathbb{E}[e_j] = 0$; $\sum_{j=1}^n \mathbb{V}[e_j] = (n-2)\sigma^2$ ($j = 1, 2, \dots, n$).

Proof. (1) の証明: 誤差項 ϵ_j ($j = 1, 2, \dots, n$) に対して

$$\bar{\epsilon} := \frac{1}{n} \sum_{j=1}^n \epsilon_j$$

とおく. すると

$$\begin{aligned} y_j - \bar{y} - \beta^*(x_j - \bar{x}) &= \epsilon_j - \bar{\epsilon}, \\ \sum_{j=1}^n (x_j - \bar{x})\bar{\epsilon} &= 0 \end{aligned}$$

となる. これより

$$\hat{\beta} - \beta^* = \frac{\sum_{j=1}^n (x_j - \bar{x})\epsilon_j}{Q_{xx}} \tag{4.2}$$

eq:2-2

と書ける. (4.2) より

$$\begin{aligned} \mathbb{E}[\hat{\beta} - \beta^*] &= 0, \\ \mathbb{V}[\hat{\beta}] &= \mathbb{E}[(\hat{\beta} - \beta^*)^2] = \frac{\sum_{j=1}^n (x_j - \bar{x})^2 \mathbb{E}[\epsilon_j^2]}{Q_{xx}^2} = \frac{\sigma^2}{Q_{xx}}. \end{aligned} \tag{4.3}$$

eq:2-3

(2) の証明: また

$$\hat{\alpha} - \alpha^* = \bar{\epsilon} - \bar{x}(\hat{\beta} - \beta^*) = \sum_{j=1}^n \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\} \epsilon_j$$

より

$$\begin{aligned} \mathbb{E}[\hat{\alpha} - \alpha^*] &= 0, \\ \mathbb{V}[\hat{\alpha}] &= \mathbb{E}[(\hat{\alpha} - \alpha^*)^2] = \sum_{j=1}^n \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\}^2 \mathbb{E}[\epsilon_j^2] = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{Q_{xx}} \right\}. \end{aligned}$$

(3) の証明: 同様に

$$\begin{aligned}\text{COV}[\hat{\alpha}, \hat{\beta}] &= \mathbb{E}[(\hat{\alpha} - \alpha^*)(\hat{\beta} - \beta^*)] \\ &= \frac{1}{Q_{xx}} \sum_{j=1}^n (x_j - \bar{x}) \left\{ \frac{1}{n} - \frac{(x_j - \bar{x})\bar{x}}{Q_{xx}} \right\} \mathbb{E}[\epsilon_j^2] = -\frac{\bar{x}\sigma^2}{Q_{xx}}.\end{aligned}$$

(4) の証明:

$$e_j = y_j - \bar{y} - \hat{\beta}(x_j - \bar{x}) = (\epsilon_j - \bar{\epsilon}) - (x_j - \bar{x})(\hat{\beta} - \beta^*)$$

と表さるので $\mathbb{E}[e_j] = 0$. また

$$\begin{aligned}\sum_{j=1}^n \mathbb{E}[e_j^2] &= \sum_{j=1}^n \mathbb{E}[(\epsilon_j - \bar{\epsilon})^2] - 2 \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[(\epsilon_j - \bar{\epsilon})(\hat{\beta} - \beta^*)] \\ &\quad + \sum_{j=1}^n (x_j - \bar{x})^2 \mathbb{E}[(\hat{\beta} - \beta^*)^2]\end{aligned}\tag{4.4}$$

eq:2-4

と書ける. (4.2) から

$$\begin{aligned}\sum_{j=1}^n \mathbb{E}[(\epsilon_j - \bar{\epsilon})^2] &= \sum_{j=1}^n \epsilon_j^2 - n\mathbb{E}[\bar{\epsilon}^2] = (n-1)\sigma^2, \\ \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[(\epsilon_j - \bar{\epsilon})(\hat{\beta} - \beta^*)] &= \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[\epsilon_j(\hat{\beta} - \beta^*)] \\ &\quad - \underbrace{\sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[\bar{\epsilon}(\hat{\beta} - \beta^*)]}_{=0} \\ &= \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}[\epsilon_j(\hat{\beta} - \beta^*)] \\ &= \sum_{j=1}^n (x_j - \bar{x}) \mathbb{E}\left[\frac{\epsilon_j \sum_{\ell=1}^n (x_\ell - \bar{x}) \epsilon_\ell}{Q_{xx}}\right] \\ &= \sigma^2\end{aligned}$$

となる. これらの 2 つの式と (4.3) を (4.4) に代入すれば

$$\sum_{j=1}^n \mathbb{E}[e_j^2] = (n-2)\sigma^2$$

を得る.

□

pro:2-2

命題 4.2. 正規性を仮定する. $n \geq 3$ のとき

- (1)
$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha^* \\ \beta^* \end{pmatrix}, \frac{\sigma^2}{Q_{xx}} \begin{pmatrix} \frac{Q_{xx}}{n} + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \right),$$
- (2)
$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\text{RSS}}{\sigma^2} \sim \chi_{n-2}^2,$$
- (3) $\hat{\sigma}^2$ と $(\hat{\alpha}, \hat{\beta})$ は独立.

Proof. (1), (2) は命題 ^{pro:2-1}4.1 と正規性の仮定より直ちにわかる. $\text{COV}[e_j, \hat{\alpha}] = \text{COV}[e_j, \hat{\beta}] = 0$ より $\hat{\sigma}^2$ と $(\hat{\alpha}, \hat{\beta})$ は独立 □

2 線型重回帰モデル

$d, n \in \mathbb{N}$ とし, n 個の観測の組を

$$(y_j, x_{j1}, x_{j2}, \dots, x_{jd}) \quad (j = 1, 2, \dots, n)$$

とする. ただし $x_{j\ell}$ ($j = 1, 2, \dots, n; \ell = 1, 2, \dots, d$) 変量は定数とする. さらに変数間には

$$y_j = \beta_0^* + \beta_1^* x_{j1} + \beta_2^* x_{j2} + \dots + \beta_d^* x_{jd} + \epsilon_j \quad (j = 1, 2, \dots, n) \quad (4.5)$$

eq:2-5

なるモデルを仮定する. ただし β_0^* は y 切片項, $\beta_1^*, \beta_2^*, \dots, \beta_d^*$ は重回帰係数という. これらは未知とする. また $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ は独立同一分布に従う確率変数列 (誤差項) であり, $\mathbb{V}[\epsilon_j] = \sigma^2$ ($j = 1, 2, \dots, n$) とする. ただし σ^2 ($\sigma > 0$) は未知とする. ^{eq:2-5}(4.5) を重回帰モデルという.

重回帰モデル ^{eq:2-5}(4.5) は

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{=\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}}_{=\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0^* \\ \beta_1^* \\ \vdots \\ \beta_d^* \end{pmatrix}}_{=\boldsymbol{\beta}^*} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{=\boldsymbol{\epsilon}}$$

と表される。これは

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \quad (4.6)$$

eq:2-6

と書ける。

回帰係数ベクトル $\boldsymbol{\beta}^*$ の最小 2 乗推定量は

$$h(\boldsymbol{\beta}) = h(\beta_0, \beta_1, \dots, \beta_d) = \sum_{j=1}^n \{y_j - (\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \dots + \beta_d x_{jd})\}^2$$

を最小化することにより得られる。関数 h は

$$h(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

と書ける。ただし行列 \mathbf{A} に対して \mathbf{A}^\top はその転置である。

以後、 $\mathbf{X}^\top \mathbf{X}$ は正則と仮定する。ここで

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

とおくと

$$h(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad (4.7)$$

eq:2-7

と書ける。このことより $h(\boldsymbol{\beta})$ は $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ で最小となる。よって $\hat{\boldsymbol{\beta}}$ は $\boldsymbol{\beta}^*$ の最小 2 乗推定量である。

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}$$

と書けるので

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta}^*, \\ \text{COV}[\hat{\boldsymbol{\beta}}] &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

さらに

$$\text{RSS} := (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^\top \{ \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \} \mathbf{y}$$

とする。トレース、期待値およびベキ等行列の性質より

$$\mathbb{E}[\text{RSS}] = \text{Tr} \left[\{ \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \} \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] \right] \quad (4.8)$$

$$= \sigma^2 \text{Tr} \left[\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right] \quad (4.9)$$

$$= (n - d - 1) \sigma^2. \quad (4.10)$$

eq:2-7

よって $n \geq d + 2$ のとき σ^2 の不偏推定量は

$$\hat{\sigma}^2 = \frac{1}{n - d - 1} \text{RSS}.$$

thm:2-1

定理 4.3. (Gauss-Markov の定理) 最小 2 乗推定量 $\hat{\beta}$ は最良線型不偏推定量 (best linear unbiased estimator, BLUE) である。

Proof. 任意の線型推定量は $(d + 1) \times n$ 行列 C を用いて Cy と表される。これが不偏推定量となるためには

$$\mathbb{E}[Cy] = \beta^* \iff CX = I_{d+1}$$

をみたさなければならない。さらに

$$\text{COV}[Cy] = \mathbb{E}[(Cy - \beta^*)(Cy - \beta^*)^\top] = C\mathbb{E}[uu^\top]C^\top = \sigma^2 CC^\top.$$

最小 2 乗推定量は $C^* := (X^\top X)^{-1} X^\top$ に対応しており、

$$\begin{aligned} CC^\top &= (C - C^* + C^*)(C - C^* + C^*)^\top \\ &= (C - C^*)(C - C^*)^\top + C^*(C^*)^\top \\ &\quad + (C - C^*)(C^*)^\top + C^*(C - C^*)^\top. \end{aligned}$$

しかし

$$(C - C^*)(C^*)^\top = \{C - (X^\top X)^{-1} X^\top\}X = CX - I_{d+1} = \mathbf{0}.$$

よって

$$\begin{aligned} \text{COV}[Cy] &= \sigma^2 CC^\top = \sigma^2 C^*(C^*)^\top + \sigma^2 (C - C^*)(C - C^*)^\top \\ &\succeq \sigma^2 C^*(C^*)^\top = \text{COV}[\hat{\beta}]. \end{aligned}$$

ただし正方行列 A, B に対して

$$A \succeq B \iff A - B \succeq \mathbf{0} \iff A - B \text{ は非負定値}$$

とした。したがって最小 2 乗推定量は共分散行列を最小 (上の \succeq の意味で最小) となるので線型不偏推定量の族の中で最良である。□

命題 4.4. $n \geq d + 2$ とする. $\epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ を仮定する. このとき以下のことが成り立つ.

(1) $\hat{\beta} \sim N_{d+1}(\beta^*, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$.

(2) $\frac{(n-d-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-d-1}^2$.

(3) $\hat{\beta}$ と $\hat{\sigma}^2$ は独立.

Proof. $z = (\mathbf{y} - \mathbf{X}\beta^*)/\sigma$ とし

$$U = (\mathbf{X}^\top \mathbf{X})^{-1/2} \mathbf{X}^\top z; \quad V = z^\top \{I_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\} z$$

とおく. ただし正則な行列 A に対して $A^{-1} = BB^\top$ をみたす行列 B を $A^{-1/2}$ ¹と記した. すると命題の主張は

(1a) $U \sim N_{d+1}(\mathbf{0}, I_{d+1})$,

(2a) $V \sim \chi_{n-d-1}^2$,

(3a) U と V は独立,

という形に簡略化される. \mathbf{X} は $n \times (d+1)$ 行列でランクが $d+1$ (フルランク)だから

$$\mathbf{X} = \mathbf{P}\mathbf{A}\mathbf{O}^\top$$

と分解できる. ただし \mathbf{P} は $n \times (d+1)$ 行列で $\mathbf{P}^\top \mathbf{P} = I_{d+1}$, \mathbf{A} は正の成分をもつ $d+1$ 次の対角行列で \mathbf{O} は $d+1$ 次の直交行列である. このとき

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \{\mathbf{O}\mathbf{A}\mathbf{P}^\top \mathbf{P}\mathbf{A}\mathbf{O}^\top\}^{-1} = \{\mathbf{O}\mathbf{A}^2 \mathbf{O}^\top\}^{-1} = \mathbf{O}\mathbf{A}^{-2} \mathbf{O}^\top$$

より

$$(\mathbf{X}^\top \mathbf{X})^{-1/2} = \mathbf{O}\mathbf{A}^{-1} \mathbf{O}^\top.$$

これらより

$$U = \mathbf{O}\mathbf{P}^\top z; \quad V = z^\top (\mathbf{I}_n - \mathbf{P}\mathbf{P}^\top) z$$

と書けることがわかる. ここで適当な $n \times (n-d-1)$ 行列 \mathbf{Q} をうまくとり $n \times n$ 行列 $\mathbf{H} := (\mathbf{P} : \mathbf{Q})$ が n 次直交行列になるようにする. すると $\mathbf{I}_n = \mathbf{H}\mathbf{H}^\top = \mathbf{P}\mathbf{P}^\top + \mathbf{Q}\mathbf{Q}^\top$ であることから $\mathbf{I}_n - \mathbf{P}\mathbf{P}^\top = \mathbf{Q}\mathbf{Q}^\top$. よって

$$V = z^\top \mathbf{Q}\mathbf{Q}^\top z = (\mathbf{Q}^\top z)^\top (\mathbf{Q}^\top z)$$

¹これは一意ではないことに注意せよ.

と表される.

$$H^\top z = \begin{pmatrix} P^\top z \\ Q^\top z \end{pmatrix} \sim N_n(\mathbf{0}, I_n)$$

であるから $U = P^\top z$ と $V = z^\top Q Q^\top z$ は独立で $U \sim N_{d+1}(\mathbf{0}, I_{d+1})$, $V \sim \chi_{n-d-1}^2$ となる. \square

3 変数選択の規準

説明変数 $x_{1j}, x_{2j}, \dots, x_{dj}$ の個数 d を増やすにつれて, 線型モデル

$$y_j = \beta_0^* + \beta_1^* x_{1j} + \beta_2^* x_{2j} + \dots + \beta_d^* x_{dj} + \epsilon_j \quad (j = 1, 2, \dots, n)$$

によるデータに対する説明力は増していき, 観測値に対するモデルの適合度は高くなる. しかし d を増やしていくにつれて未知母数である回帰係数の推定量の推定誤差は増していく. あとの章でわかるように

$$\mathbb{E}[\text{Tr}\{(X\hat{\beta} - X\beta^*)^\top (X\hat{\beta} - X\beta^*)\}] \leq C \frac{\sigma^2 d}{n}$$

と評価できる. ただし C は n, d に依らない定数である.

以下では $n \geq d+1$ とし, X はフルランクとする.

3.1 自由度調整済み決定係数

決定係数

$$R_d^2 = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}; \quad \hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \dots + \hat{\beta}_d x_{dj}$$

の値は説明変数の個数 d を増やせば 1 に近づいていく. そこで $\sum_{j=1}^n (y_j - \hat{y}_j)^2$ と $\sum_{j=1}^n (y_j - \bar{y})^2$ をそれらの自由度 $n-d-1, n-1$ で割ったもので置き換えたもの

$$R_d^{*2} = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2 / (n-d-1)}{\sum_{j=1}^n (y_j - \bar{y})^2 / (n-1)}$$

を自由度調整済み決定係数という. これを書き直せば

$$R_d^{*2} = 1 - \frac{n-1}{n-d-1} (1 - R_d^2)$$

となる。この形からわかるように d が大きくなると $1 - R_d^2$ は小さくなるものの分母の $n - d - 1$ が小さくなるので R_d^{2*} は必ずしも 1 には近づかない。 R_d^{2*} を最大にする説明変数の組 (d の選択) を選択すればよい。

3.2 Mallows C_p

モデルの誤差項 ϵ とは独立な確率ベクトル $\tilde{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ に基づく将来の観測を

$$\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}^* + \tilde{\epsilon}$$

とする。未来の観測 $\tilde{\mathbf{y}}$ を $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ で予測するとき、その平均 2 乗誤差は

$$\begin{aligned} \text{MSE}(\hat{\mathbf{y}}) &= \mathbb{E}[(\tilde{\mathbf{y}} - \hat{\mathbf{y}})^\top (\tilde{\mathbf{y}} - \hat{\mathbf{y}})] \\ &= \mathbb{E}[\{\tilde{\epsilon} - \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}^\top \{\tilde{\epsilon} - \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}] \\ &= n\sigma^2 + \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \mathbf{X}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)] \\ &= n\sigma^2 + \text{Tr}\{\mathbf{X}^\top \mathbf{X} \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top]\} \\ &= n\sigma^2 + (d+1)\sigma^2 = (n+d+1)\sigma^2 \end{aligned}$$

となる。一方、残差平方和 $\text{RSS}_d = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$ の期待値は

$$\mathbb{E}[\text{RSS}_d] = \mathbb{E}[\mathbf{y}^\top \{\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\} \mathbf{y}] = (n-d-1)\sigma^2$$

となるので、予測誤差は

$$\text{MSE}(\hat{\mathbf{y}}) = \mathbb{E}[\text{RSS}_d + 2(d+1)\sigma^2]$$

となる。よって σ^2 が既知の場合、 $\text{RSS}_d + 2(d+1)\sigma^2$ が予測誤差の不偏推定量になっていることが分かる。

候補となる最大のモデルの説明変数の組を

$$(x_{j1}, x_{j2}, \dots, x_{jd}, \dots, x_{jK}) \quad (1 \leq d \leq K)$$

とする。

$$\mathbf{X}_F = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2d} & \dots & x_{2K} \\ \vdots & & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} & \dots & x_{nK} \end{pmatrix}$$

としたとき分散 σ^2 の推定量を

$$\hat{\sigma}_F^2 = \mathbf{y}^\top \{ \mathbf{I}_n - \mathbf{X}_F (\mathbf{X}_F^\top \mathbf{X}_F)^{-1} \mathbf{X}_F^\top \} \mathbf{y}$$

とする. σ^2 の推定量 $\hat{\sigma}_F^2$ で割ったもの

$$C_p := \frac{\text{RSS}_d}{\hat{\sigma}_F^2} + 2(d+1)$$

を Mallows の C_p 規準といい, これを最小にする変数の組を選べばよい. これは

$$(\text{モデルの適合度}) + 2 \times (\text{モデルの母数の個数})$$

という形をしており, 第 2 項はモデルの複雑さに対する罰則項として機能する.

3.3 AIC

定義 4.5. (Kullback-Leibler 情報量) f, g を Lebesgue 測度 μ に関する確率密度関数とする. Kullback-Leibler 情報量を

$$\text{KL}(f, g) = \int \log \left(\frac{f}{g} \right) f d\mu$$

で定義する.

注意 4.6. $x \log x \geq x - 1$ ($x > 0$) に注意して

$$\begin{aligned} \text{KL}(f, g) &= \int \log \left(\frac{f}{g} \right) f d\mu = \int \frac{f}{g} \log \left(\frac{f}{g} \right) g d\mu \\ &\geq \int \left(\frac{f}{g} - 1 \right) g d\mu = 0. \end{aligned}$$

□

以下では $n \geq d + 4$ とする. $\mathbf{y} = \mathbf{X}\beta^* + \epsilon$ の確率密度関数を $p(\mathbf{y} | \mathbf{X}\beta^*, \sigma^2)$ とし, 将来の変数 $\tilde{\mathbf{y}} = \mathbf{X}\beta^* + \tilde{\epsilon}$ の確率密度関数を $p(\tilde{\mathbf{y}} | \mathbf{X}\beta^*, \sigma^2)$ と書くことにする. β^*, σ^2 の推定量 $\hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})$ を $p(\tilde{\mathbf{y}} | \mathbf{X}\beta^*, \sigma^2)$ に代入したもの

$$p(\tilde{\mathbf{y}} | \mathbf{X}\hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))$$

を推定されたモデルの分布とし、これで将来の分布 $p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2)$ を予測するとき、それらの分布間の Kullback-Leibler 情報量

$$\begin{aligned} & \text{KL}(p(\cdot | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2), p(\cdot | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))) \\ &= \int \cdots \int \log \left\{ \frac{p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2)}{p(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))} \right\} \cdot p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \end{aligned}$$

で測る。以後、ルベーク測度 $d\mu(\tilde{\mathbf{y}})$ を $d\tilde{\mathbf{y}}$ と書くことにする。これは \mathbf{y} に依存するランダムな量なので、 \mathbf{y} に関して期待値を取ったもの

$$\mathbb{E} \left[\text{KL}(p(\cdot | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2), p(\cdot | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))) \right]$$

を考える。するとこの関数が Mallows の C_p 規準で考えたところの平均 2 乗予測誤差に対応している。

$$\begin{aligned} & \mathbb{E} \left[\text{KL}(p(\cdot | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2), p(\cdot | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y}))) \right] \\ &= \mathbb{E} \left[\int \cdots \int \log \{ p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) \} p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \right] \\ &\quad - \mathbb{E} \left[\int \cdots \int \log \{ p(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) \} p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \right] \end{aligned}$$

と書き直せる。右辺の第 1 項目は推定されたモデルの分布に無関係なので、後者を 2 倍したものを

$$\text{AI}(\boldsymbol{\beta}^*, \sigma^2) := -2\mathbb{E} \left[\int \cdots \int \log \{ p(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) \} p(\tilde{\mathbf{y}} | \mathbf{X}\boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \right]$$

とおく。これを赤池情報量と呼ぶ。これの (漸近) 不偏推定量が AIC 規準となる。具体的に $\text{AI}(\boldsymbol{\beta}^*, \sigma^2)$ を計算してみると

$$-2 \log p(\tilde{\mathbf{y}} | \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) = n \log(2\pi\hat{\sigma}^2(\mathbf{y})) + \frac{(\tilde{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\tilde{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}))}{\hat{\sigma}^2(\mathbf{y})}$$

であることと命題 ^{pro:2-3} 4.4 に注意し $\tilde{\mathbf{y}}$ に関して積分²する。

$$\begin{aligned}
 & \int \cdots \int \left\{ -2 \log p(\tilde{\mathbf{y}} | \mathbf{X} \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) \right\} p(\tilde{\mathbf{y}} | \mathbf{X} \boldsymbol{\beta}^*, \sigma^2) d\tilde{\mathbf{y}} \\
 &= \int \cdots \int \left\{ n \log(2\pi \hat{\sigma}^2(\mathbf{y})) + \frac{\{\tilde{\boldsymbol{\epsilon}} + \mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}(\mathbf{y}))\}^\top \{\tilde{\boldsymbol{\epsilon}} + \mathbf{X}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}(\mathbf{y}))\}}{\hat{\sigma}^2(\mathbf{y})} \right\} \\
 & \quad \times p(\tilde{\boldsymbol{\epsilon}} | \mathbf{0}, \sigma^2) d\tilde{\boldsymbol{\epsilon}} \quad \left(\tilde{\mathbf{y}} = \mathbf{X} \boldsymbol{\beta}^* + \tilde{\boldsymbol{\epsilon}} \text{ と変換} \right) \\
 &= n \log(2\pi \hat{\sigma}^2(\mathbf{y})) + \frac{n\sigma^2 + (\hat{\boldsymbol{\beta}}(\mathbf{y}) - \boldsymbol{\beta}^*)^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}}(\mathbf{y}) - \boldsymbol{\beta}^*)}{\hat{\sigma}^2(\mathbf{y})}
 \end{aligned}$$

と書ける。命題 ^{pro:2-3} 4.4 から

$$\begin{aligned}
 \mathbb{E}[(\hat{\boldsymbol{\beta}}(\mathbf{y}) - \boldsymbol{\beta}^*)^\top \mathbf{X}^\top \mathbf{X} (\hat{\boldsymbol{\beta}}(\mathbf{y}) - \boldsymbol{\beta}^*)] &= (d+1)\sigma^2, \\
 \mathbb{E} \left[\frac{\sigma^2}{\hat{\sigma}^2(\mathbf{y})} \right] &= (n-d-1) \mathbb{E} \left[\frac{1}{\chi_{n-d-1}^2} \right] = \frac{n-d-1}{n-d-3}
 \end{aligned}$$

である。さらに、 $\hat{\boldsymbol{\beta}}(\mathbf{y})$ と $\hat{\sigma}^2(\mathbf{y})$ の独立性であることに注意すれば、

$$\text{AI}(\boldsymbol{\beta}^*, \sigma^2) = \mathbb{E}[n \log(2\pi \hat{\sigma}^2(\mathbf{y}))] + \frac{(n+d+1)(n-d-1)}{n-d-3} \quad (4.11)$$

eq:2-11

と書けることがわかる。

次に、対数尤度関数を計算する。

$$\begin{aligned}
 \hat{\sigma}^2(\mathbf{y}) &= \frac{1}{n-d-1} \mathbf{y}^\top \{ \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \} \mathbf{y} \\
 &= \frac{1}{n-d-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\mathbf{y}))
 \end{aligned}$$

に注意して、

$$\begin{aligned}
 & \mathbb{E} \left[-\log p(\mathbf{y} | \mathbf{X} \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) \right] \\
 &= \mathbb{E} \left[n \log(2\pi \hat{\sigma}^2(\mathbf{y})) + \frac{(\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\mathbf{y}))^\top (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}(\mathbf{y}))}{\hat{\sigma}^2(\mathbf{y})} \right] \\
 &= \mathbb{E}[n \log(2\pi \hat{\sigma}^2(\mathbf{y}))] + (n-d-1)
 \end{aligned}$$

²実際には、 $\tilde{\boldsymbol{\epsilon}} \sim N_n(\mathbf{0}, \sigma \mathbf{I}_n)$ に関して積分していることに注意せよ。

となる. これを (Eq:2-11) に代入すれば,

$$AI(\beta^*, \sigma^2) = \mathbb{E} \left[-2 \log p(\mathbf{y} | \mathbf{X} \hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) + 2(d+2) \frac{n-d-1}{n-d-3} \right]$$

と書けることがわかる. すなわち, 上の式の右辺の期待値記号の中身が $AI(\beta^*, \sigma^2)$ の不偏推定量になる.

$$\lim_{n \rightarrow \infty} \frac{n-d-1}{n-d-3} = 1$$

なので, 近似推定量として,

$$AIC := -2 \log p(\mathbf{y} | \mathbf{X} \hat{\beta}(\mathbf{y}), \hat{\sigma}^2(\mathbf{y})) + 2(d+2) \quad (4.12)$$

eq:2-12

が得られる. これを AIC 規準という. AIC を最小化する変数の組を選べばよい. Mallows の C_p 規準と同様, (Eq:2-12) の第 1 項目はデータの適合度, 第 2 項目はモデルの複雑さに対する罰則と解釈できる.

3.4 交差検証法

観測 $(x_1, y_1), \dots, (x_2, y_2), \dots, (x_n, y_n)$ から j 番目 ($j = 1, 2, \dots, n$) のデータ (x_j, y_j) を除いた残りのデータからの β^* の推定量を考える.

$$\begin{aligned} \hat{\beta}^{(j)} &= \{(\mathbf{X}^{(j)})^\top \mathbf{X}^{(j)}\}^{-1} (\mathbf{X}^{(j)})^\top \mathbf{y}^{(j)}, \\ \mathbf{y}^{(j)} &= (y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_n)^\top, \\ (\mathbf{X}^{(j)})^\top &= (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n). \end{aligned}$$

を考える. y_j の予測量 $\hat{y}_j = \mathbf{x}_j^\top \hat{\beta}^{(j)}$ を構成し, y_j に対する予測誤差

$$\{y_j - \mathbf{x}_j^\top \hat{\beta}^{(j)}\}^2$$

を計算する. この操作を繰り返し, 予測誤差

$$CV := \frac{1}{n} \sum_{j=1}^n \{y_j - \mathbf{x}_j^\top \hat{\beta}^{(j)}\}^2$$

を得る. これを交差検証法 (クロス・バリデーション) といい, CV を最小にする説明変数の組を選ぶ.

命題 4.7. CV は以下のように表現できる.

$$CV = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{y_j - \mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}^*}{1 - \mathbf{x}_j^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_j} \right\}^2.$$

Proof. 命題の主張を証明するために

$$\frac{y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}}{1 - \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_1} = y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}$$

を示せばよい.

$$\begin{aligned} \mathbf{X}^\top &= (\mathbf{x}_1; (\mathbf{X}^{(1)})^\top), \\ \mathbf{X}^\top \mathbf{X} &= \mathbf{x}_1 \mathbf{x}_1^\top + (\mathbf{X}^{(1)})^\top \mathbf{X}^{(1)} =: \mathbf{x}_1 \mathbf{x}_1^\top + A \end{aligned}$$

に注意する. 等式

$$(\mathbf{X}^\top \mathbf{X})^{-1} = A^{-1} - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \tag{4.13}$$

eq:2-13

より

$$\begin{aligned} 1 - \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_1 &= 1 - \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 + \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} \mathbf{x}_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\ &= \frac{1 - (\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2 + (\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\ &= \frac{1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \end{aligned} \tag{4.14}$$

eq:2-14

を得る. これらより

$$\begin{aligned}
\mathbf{x}_1^\top \widehat{\boldsymbol{\beta}} &= \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \mathbf{x}_1^\top \left\{ A^{-1} - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} (\mathbf{x}_1; (\mathbf{X}^{(1)})^\top)^\top \begin{pmatrix} y_1 \\ \mathbf{y}^{(1)} \end{pmatrix} \\
&= \mathbf{x}_1^\top \left\{ A^{-1} - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \{ \mathbf{x}_1 y_1 + (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \} \\
&= \mathbf{x}_1^\top \left\{ A^{-1} \mathbf{x}_1 y_1 - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} + A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \right. \\
&\quad \left. - \frac{A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \\
&= \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 - \frac{(\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2 y_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} + \mathbf{x}_1^\top A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)} \\
&\quad - \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top A^{-1} (\mathbf{X}^{(1)})^\top \mathbf{y}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\
&= \mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 - \frac{(\mathbf{x}_1^\top A^{-1} \mathbf{x}_1)^2 y_1}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)} - \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \\
&= \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \tag{4.15}
\end{aligned}$$

eq:2-15

を得る. (4.14) と (4.15) より

$$\begin{aligned}
\frac{y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}}{1 - \mathbf{x}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_1} &= (1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1) \left\{ y_1 - \frac{\mathbf{x}_1^\top A^{-1} \mathbf{x}_1 y_1 + \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \\
&= (1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1) \left\{ \frac{y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}}{1 + \mathbf{x}_1^\top A^{-1} \mathbf{x}_1} \right\} \\
&= y_1 - \mathbf{x}_1^\top \widehat{\boldsymbol{\beta}}^{(1)}
\end{aligned}$$

を得る. □

3.5 BIC

線型回帰モデル

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \text{COV}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n$$

において、 $(d+2)$ 個の未知母数 β^*, σ^2 に正則な事前分布 $\pi_{d+2}(\beta, \sigma^2)$ を仮定する。ここで、「正則」とは

$$\int \cdots \int \pi_{d+2}(\beta^*, \sigma^2) d\beta^* d\sigma^2 = 1$$

が成立していることである。すると \mathbf{y} の周辺分布

$$p_{\pi_{d+2}}(\mathbf{y}) = \int \cdots \int p(\mathbf{y} | \mathbf{X}\beta, \sigma^2) \pi_{d+2}(\beta, \sigma^2) d\beta d\sigma^2$$

で与えられる。これは σ^2 と $(d+1)$ 個の回帰係数 $\beta_0^*, \beta_1^*, \dots, \beta_d^*$ に事前分布を想定した Bayes 的周辺尤度である。これを最大にする説明変数の組 (Bayes 的周辺尤度 $-2 \log f_{\pi_{d+2}}(\mathbf{y})$ を最小にする説明変数の組) を選択する。基準となるモデルを定め、比較するモデルとの周辺尤度の比を Bayes 因子 (Bayes factor) と呼ぶ。

例えば、最も簡単なモデル

$$y_j = \beta_0^* + \epsilon_j \quad (j = 1, 2, \dots, n) \quad (4.16)$$

を考え、2 個の未知母数 β_0^*, σ^2 に事前分布 $\pi_2(\beta_0^*, \sigma^2)$ を想定した Bayes 的周辺尤度

$$p_{\pi_2}(\mathbf{y}) = \int \int p(\mathbf{y} | \beta_0, \sigma^2) \pi_2(\beta_0, \sigma^2) d\beta_0 d\sigma^2$$

を考える。ただし、 $f(\mathbf{y} | \beta_0^*, \sigma^2)$ はモデル (4.16) の確率密度関数である。これらの比

$$\frac{p_{\pi_{d+2}}(\mathbf{y})}{p_{\pi_2}(\mathbf{y})}$$

を考える。この値が $1/2$ を超えていれば、 d 個の説明変数は Bayes 的な意味を持つと解釈できる。これが Bayes 因子であり、この値を最大にする説明変数の組を選択すればよい。

Bayes 的周辺尤度や Bayes 因子の問題点は、これらが事前分布の取り方に依存していることである。ここでは、 $n \rightarrow \infty$ とすることによりその極限を求める。

そのために Laplace 近似を用いる。以下の議論は数学的な厳密性にかけるものであることに注意せよ。章末の補遺を参照のこと。一般に θ を k 次元の未知母数とし、 \mathbf{X} を n 次元確率変数とし、 θ を与えたときの \mathbf{X} の条件付き確率密度関数を

$$\mathbf{X} | \theta \sim p(\mathbf{x} | \theta)$$

eq:2-16

とし, θ の事前分布を

$$\theta \sim \pi(\theta)$$

とする. 対数尤度

$$\ell(\theta | \mathbf{x}) = \log p(\mathbf{x} | \theta)$$

を θ の最尤推定量³ $\hat{\theta}$ のまわりで Taylor 展開する.

$$\begin{aligned} \ell(\theta | \mathbf{x}) &\approx \ell(\hat{\theta} | \mathbf{x}) + \left(\frac{\partial}{\partial \theta} \ell(\theta | \mathbf{x}) \Big|_{\theta=\hat{\theta}} \right)^\top (\hat{\theta} - \theta) \\ &\quad + \frac{1}{2} (\hat{\theta} - \theta)^\top \left(\frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta | \mathbf{x}) \Big|_{\theta=\hat{\theta}} \right) (\hat{\theta} - \theta) \end{aligned}$$

と近似できる. ただし

$$\begin{aligned} \frac{\partial}{\partial \theta} \ell(\theta | \mathbf{x}) &= \left(\frac{\partial}{\partial \theta_1} \ell(\theta | \mathbf{x}), \dots, \frac{\partial}{\partial \theta_k} \ell(\theta | \mathbf{x}) \right)^\top \\ \frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta | \mathbf{x}) &= \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta | \mathbf{x}) \right)_{i,j=1,2,\dots,k} \\ \theta &= (\theta_1, \theta_2, \dots, \theta_k)^\top \end{aligned}$$

である. さらに

$$\hat{I}(\mathbf{x}) = -\frac{1}{n} \frac{\partial^2}{\partial \theta \partial \theta^\top} \ell(\theta | \mathbf{x}) \Big|_{\theta=\hat{\theta}}$$

とおいたとき $\hat{I}(X)$ はある正定値行列に確率収束すると仮定する. $\hat{\theta}$ は θ の最尤推定量だから

$$\frac{\partial}{\partial \theta} \ell(\theta | \mathbf{x}) \Big|_{\theta=\hat{\theta}} = \mathbf{0}$$

である. よって

$$\ell(\theta | \mathbf{x}) \approx \ell(\hat{\theta} | \mathbf{x}) - \frac{n}{2} (\hat{\theta} - \theta)^\top \hat{I}(\mathbf{x}) (\hat{\theta} - \theta)$$

と近似できる. $\pi(\theta)$ は θ に関して滑らかな関数で

$$\pi(\theta) \approx \pi(\hat{\theta})$$

³すなわち, 存在すれば,

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} \ell(\theta | \mathbf{x})$$

で定義する.

と近似できると仮定すると Bayes 的周辺分布は

$$\begin{aligned}
 p_{\pi}(\mathbf{x}) &= \int \cdots \int p(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
 &= \int \cdots \int \exp\{\ell(\boldsymbol{\theta}|\mathbf{x})\}\pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
 &\approx p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \frac{(2\pi)^{p/2}}{|n\hat{I}(\mathbf{x})|^{1/2}} \pi(\hat{\boldsymbol{\theta}}) \\
 &\quad \times \int \cdots \int \frac{|n\hat{I}(\mathbf{x})|^{1/2}}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\top} n\hat{I}(\mathbf{x})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \\
 &= p(\mathbf{x}|\hat{\boldsymbol{\theta}}) \frac{(2\pi)^{p/2}}{|n\hat{I}(\mathbf{x})|^{1/2}} \pi(\hat{\boldsymbol{\theta}})
 \end{aligned}$$

で近似できることが知られている。これを Laplace 近似という。ここで「 \approx 」の意味は以下である。 $a \in \mathbb{R}$ と $\{a_n\}_{n=1}^{\infty}$ に対して、

$$a \approx a_n \iff \lim_{n \rightarrow \infty} \frac{a}{a_n} = 1$$

である。したがって、

$$-2 \log p_{\pi}(\mathbf{x}) \approx -2 \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}) + p \log n + p \log \left(\frac{|\hat{I}(\mathbf{x})|^{1/(2p)}}{2\pi} \right) - 2 \log \pi(\hat{\boldsymbol{\theta}})$$

と書くことができる。 n が大きいとき、定式の右辺の 3, 4 項目

$$p \log \left(\frac{|\hat{I}(\mathbf{x})|^{1/(2p)}}{2\pi} \right) - 2 \log \pi(\hat{\boldsymbol{\theta}})$$

は $\log n$ に比較して無視できるので、

$$\text{BIC} := -2 \log p(\mathbf{x}|\hat{\boldsymbol{\theta}}) + p \log n$$

なる近似式を得る。これを Schwarz の Bayes 情報量規準という。

線型回帰モデルに適用してみると

$$\text{BIC} = -2 \log p(\mathbf{y}|\mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + (d+2) \log n$$

を得る。

2-2

注意 4.8. AIC と比較すると、モデルの複雑さへの罰則項が異なる。すなわち、AIC は $2(d+2)$ で、BIC は $(d+2)\log n$ である。これらの違いから以下の性質が知られている。BIC は真のモデルの選択に対する一貫性を持つ。一方、AIC はその性質を持たない。しかし、AIC は予測誤差を最小にするモデルを選択するが、BIC はそのような性質を持たない。□

4 補遺：Laplace 近似について

pro:2-5

命題 4.9. 連続関数 $h, g: \mathbb{R}^d \rightarrow \mathbb{R}$ と $\mathbf{x}_0 \in \mathbb{R}^d$ に以下の条件を仮定する。

(1) h は \mathbf{x}_0 において最小値を取り、

$$h(\mathbf{x}) > h(\mathbf{x}_0) \quad (\mathbf{x} \neq \mathbf{x}_0).$$

(2) h は \mathbf{x}_0 において、2 回連続微分可能で、

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} p(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} &= \mathbf{0}, \\ I(\mathbf{x}_0) &= \left(\frac{\partial^2}{\partial x_j \partial x_k} h(\mathbf{x}) \right)_{1 \leq j, k \leq d} \Big|_{\mathbf{x}=\mathbf{x}_0} \quad \text{は正定値,} \\ \mathbf{x} &= (x_1, x_2, \dots, x_d)^\top. \end{aligned}$$

(3) ある $\delta > 0$ と $\epsilon > 0$ が存在して、

$$\lim_{\lambda \rightarrow \infty} e^{\lambda h(\mathbf{x}_0) - \epsilon} g(\mathbf{x}_0) \int_{|\mathbf{x} - \mathbf{x}_0|_2 > \lambda} \exp\{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} = 0.$$

このとき、

$$\lim_{\lambda \rightarrow \infty} e^{\lambda h(\mathbf{x}_0)} \left(\frac{\lambda}{2\pi} \right)^{d/2} \int_{\mathbb{R}^d} e^{-\lambda h(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} = g(\mathbf{x}_0) |I(\mathbf{x}_0)|^{1/2}$$

が成立する。

Proof. まず、

$$h(\mathbf{x}) - h(\mathbf{x}_0) = \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) (1 + o(1)) \quad (\mathbf{x} \rightarrow \mathbf{x}_0).$$

これと g の連続性より, $\forall \epsilon > 0$ に対して, $\delta > 0$ がうまく取れて, $|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta$ ならば,

$$\begin{aligned} (1 - \epsilon) \left\{ \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) \right\} &\leq h(\mathbf{x}) - h(\mathbf{x}_0) \\ &\leq (1 + \epsilon) \left\{ \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) \right\}, \\ (1 - \epsilon)g(\mathbf{x}_0) &\leq g(\mathbf{x}) \leq (1 + \epsilon)g(\mathbf{x}_0) \end{aligned}$$

となる. これから

$$\begin{aligned} &\lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp \left\{ -\frac{\lambda}{2} (1 \pm \epsilon) (\mathbf{x} - \mathbf{x}_0)^\top I(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) \right\} (1 \pm \epsilon) g(\mathbf{x}_0) d\mathbf{x} \\ &= \int_{|\mathbf{y}|_2 \leq \sqrt{\lambda} \delta} \exp \left\{ -\frac{1}{2} (1 \pm \epsilon) \mathbf{y}^\top I(\mathbf{x}_0) \mathbf{y} \right\} (1 \pm \epsilon) g(\mathbf{x}_0) d\mathbf{y} \\ &\longrightarrow (1 \pm \epsilon)^{-d/2+1} (2\pi)^{d/2} \sqrt{|I(\mathbf{x}_0)|} g(\mathbf{x}_0) \quad (\lambda \rightarrow \infty). \end{aligned}$$

よって

$$\begin{aligned} &(1 - \epsilon)^{-d/2+1} g(\mathbf{x}_0) (2\pi)^{d/2} \sqrt{|I(\mathbf{x}_0)|} \\ &\leq \liminf_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp \{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} \\ &\leq \limsup_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{|\mathbf{x} - \mathbf{x}_0|_2 \leq \delta} \exp \{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} \\ &\leq (1 + \epsilon)^{-d/2+1} g(\mathbf{x}_0) (2\pi)^{d/2} \sqrt{|I(\mathbf{x}_0)|}. \end{aligned} \tag{4.17}$$

eq:2-17

次に, 容易にわかるように

$$\int_{|\mathbf{x} - \mathbf{x}_0|_2 > \delta} \exp \{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} = 0. \tag{4.18}$$

eq:2-18

よって, \mathbf{x}_0 の近くの評価 (4.17) と遠方での評価 (4.18) と $\epsilon > 0$ は任意に取ったことより

$$\lim_{\lambda \rightarrow \infty} \lambda^{d/2} \int_{\mathbb{R}^d} \exp \{-\lambda \{h(\mathbf{x}) - h(\mathbf{x}_0)\}\} g(\mathbf{x}) d\mathbf{x} = g(\mathbf{x}_0) (2\pi)^{d/2} |I(\mathbf{x}_0)|^{1/2}.$$

したがって,

$$\lim_{\lambda \rightarrow \infty} e^{\lambda h(\mathbf{x}_0)} \left(\frac{\lambda}{2\pi} \right)^{d/2} \int_{\mathbb{R}^d} \exp \{-\lambda h(\mathbf{x})\} g(\mathbf{x}) d\mathbf{x} = g(\mathbf{x}_0) |I(\mathbf{x}_0)|^{1/2}.$$

□

第5章 指数型分布族と一般化線型モデル

1 指数型分布族

$\Theta \subset \mathbb{R}$ を母数空間とし、母数 $\theta \in \Theta$ をもつ確率分布に従う確率変数 X を考える。

定義 5.1. 確率変数 X の PDF または PMF (これを p と記す) の対数をとったものが次の形で記述されるとき、その分布は指数型分布族に属すると言われる。

$$\log p(x|\theta) = a(x)b(\theta) + c(\theta) + d(x).$$

ただし $\log 0 = 0$ と約束し、 a, b, c, d は関数で

$$a: \mathcal{X} \rightarrow \mathbb{R}, \quad b: \Theta \rightarrow \mathbb{R}, \quad c: \Theta \rightarrow \mathbb{R}, \quad d: \mathcal{X} \rightarrow \mathbb{R}$$

とする。ただしこれらの関数形は既知とし、 $\mathcal{X} = \{x \in \mathbb{R}; p(x|\theta) > 0\}$ である。

$a(x) = x$ のとき、その分布は正準型 (canonical form) であると言われ、 $b(\theta)$ は自然母数 (natural parameter) と呼ばれる。

注意 5.2. もし関心のある未知母数 θ 以外に他の未知母数があるとき、それらを関数 b, c を構成する局外母数 (nuisance parameter) と呼ばれる。 □

例 5.3. (Bernoulli 分布) X は $\text{Ber}(\theta)$ ($0 < \theta < 1$) に従うとき、 X の PMF p は

$$p(x|\theta) = \theta^x(1-\theta)^{1-x} \quad (x = 0, 1)$$

と書ける。これは

$$\log p(x|\theta) = x \log \theta + (1-x) \log(1-\theta) = x \log \frac{\theta}{1-\theta} + \log(1-\theta)$$

Dobson (用して指
と一般化
ルを借用
岡留の『
の4章の
ころを

exp-1

exp-2

exp-3

と表現できるので

$$a(x) = x, \quad b(\theta) = \frac{\theta}{1-\theta}, \quad c(\theta) = \log(1-\theta), \quad d(\theta) = 0.$$

□

ex:exp-4

例 5.4. (ポアソン分布) X は $\text{Pois}(\theta)$ ($\theta > 0$) に従うとき, X の PMF p は

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \quad (x = 0, 1, \dots)$$

と書ける. ただし, $0! = 1$ と形式的に定める. これは

$$\log p(x|\theta) = x \log \theta - \theta - \log x!$$

と表現できるので

$$a(x) = x, \quad b(\theta) = \log \theta, \quad c(\theta) = -\theta, \quad d(x) = -\log x!.$$

□

ex:exp-5

例 5.5. (正規分布) X は $N(\theta, \sigma^2)$ ($\theta \in \mathbb{R}, 0 < \sigma < \infty$) に従うとき, X の PDF p は

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right] \quad (-\infty < x < \infty)$$

と書かれる. ここで, 関心のある母数は θ で, σ は局外母数で既知とする. このとき,

$$\log p(x|\theta) = -\frac{x^2}{2\sigma^2} + \frac{x\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2)$$

と表現できるので,

$$a(x) = x, \quad b(\theta) = \frac{\theta}{\sigma^2}, \quad c(\theta) = -\frac{\theta^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2), \quad d(x) = -\frac{x^2}{2\sigma^2}.$$

□

2 指数型分布族の性質

以下では連続型分布の場合で議論を進める。離散型分布の場合には積分記号を和の記号に書きかえればよい。さらに積分記号と微分記号の交換が可能であることが保証されていると仮定¹して議論を進める。

まず $p(x|\theta)$ は PDF であることから

$$\int_{\mathcal{X}} p(x|\theta) dx = 1. \quad (5.1)$$

ただし $\mathcal{X} = \{x \in \mathbb{R}; p(x|\theta) > 0\}$ である。 θ に関して (5.1) の両辺を微分すれば、

$$0 = \frac{d}{d\theta} 1 = \frac{d}{d\theta} \int_{\mathcal{X}} p(x|\theta) dx = \int_{\mathcal{X}} \frac{d}{d\theta} p(x|\theta) dx$$

から

$$\int_{\mathcal{X}} \frac{d}{d\theta} p(x|\theta) dx = 0 \quad (5.2)$$

を得る。同様に

$$\int_{\mathcal{X}} \frac{d^2 p(x|\theta)}{d\theta^2} dx = 0. \quad (5.3)$$

指数型分布族に属する分布の PDF を (5.1) に代入すると

$$\begin{aligned} \int_{\mathcal{X}} \frac{d}{d\theta} p(x|\theta) dx &= \int_{\mathcal{X}} \left(\frac{d}{d\theta} \log p(x|\theta) \right) p(x|\theta) dx \\ &= \int_{\mathcal{X}} \{a(x)\dot{b}(\theta) + \dot{c}(\theta)\} p(x|\theta) dx = 0. \end{aligned}$$

ただし

$$\dot{b}(\theta) = \frac{db(\theta)}{d\theta}, \quad \dot{c}(\theta) = \frac{dc(\theta)}{d\theta}.$$

期待値の定義から

$$\int_{\mathcal{X}} a(x)\dot{b}(\theta)p(x|\theta) dx = \dot{b}(\theta)\mathbb{E}[a(X)], \quad \int_{\mathcal{X}} \dot{c}(\theta)p(x|\theta) dx = \dot{c}(\theta)$$

¹実は、指数型分布族の場合には積分記号と微分記号の交換が保証されることを示すことができる。Lehmann and Romano (2022, pp.52-53) を参照のこと。

となるので,

$$\dot{b}(\theta)\mathbb{E}[a(X)] + \dot{c}(\theta) = 0. \quad (5.4)$$

よって $\dot{b}(\theta) \neq 0$ のとき

$$\mathbb{E}[a(X)] = -\frac{\dot{c}(\theta)}{\dot{b}(\theta)}. \quad (5.5)$$

同様に

$$\begin{aligned} \frac{d^2}{d\theta^2} p(x|\theta) &= \frac{d}{d\theta} \left[\left(\frac{d}{d\theta} \log p(x|\theta) \right) p(x|\theta) \right] \\ &= \left(\frac{d^2}{d\theta^2} \log p(x|\theta) \right) p(x|\theta) + \left(\frac{d}{d\theta} \log p(x|\theta) \right)^2 p(x|\theta) \\ &= \{a(x)\ddot{b}(\theta) + \ddot{c}(\theta)\} p(x|\theta) + \{a(x)\dot{b}(\theta) + \dot{c}(\theta)\}^2 p(x|\theta). \end{aligned} \quad (5.6)$$

ただし

$$\ddot{b}(\theta) = \frac{d^2 b(\theta)}{d\theta^2}, \quad \ddot{c}(\theta) = \frac{d^2 c(\theta)}{d\theta^2}.$$

(5.6) の最右辺の第 2 項に (5.5) を代入すると

$$\begin{aligned} \{a(x)\dot{b}(\theta) + \dot{c}(\theta)\} p(x|\theta) &= \{\dot{b}(\theta)\}^2 \left\{ a(x) + \frac{\dot{c}(\theta)}{\dot{b}(\theta)} \right\}^2 p(x|\theta) \\ &= \{\dot{b}(\theta)\}^2 \{a(x) - \mathbb{E}[a(X)]\}^2 p(x|\theta) \end{aligned}$$

と表され

$$\int_x \{a(x) - \mathbb{E}[a(X)]\}^2 p(x|\theta) dx = \mathbb{V}[a(X)]. \quad (5.7)$$

一方, (5.3) と (5.7) から (5.6) は以下のように書き直せる.

$$\begin{aligned} 0 &= \int_x \frac{d^2}{d\theta^2} p(x|\theta) dx = \int_x \{a(x)\ddot{b}(\theta) + \ddot{c}(\theta)\} p(x|\theta) dx + \{\dot{b}(\theta)\}^2 \mathbb{V}[a(X)] \\ &= \ddot{b}(\theta)\mathbb{E}[a(X)] + \ddot{c}(\theta) + \{\dot{b}(\theta)\}^2 \mathbb{V}[a(X)]. \end{aligned}$$

よって $\dot{b}(\theta) \neq 0$ のとき

$$\mathbb{V}[a(X)] = \frac{\ddot{b}(\theta)\dot{c}(\theta) - \ddot{c}(\theta)\dot{b}(\theta)}{\{\dot{b}(\theta)\}^3}. \quad (5.8)$$

次に $X = x$ が観測されたときの対数尤度関数

$$\ell(\theta|x) = \log p(x|\theta)$$

の導関数の x に確率変数 X を形式的に代入して確率変数にしたもの $\ell(\theta|X)$ (これも対数尤度関数と乱暴に呼ぶことにする) の期待値と分散を求める。

$$S(\theta|x) := \frac{\partial \ell(\theta|x)}{\partial \theta} = \frac{\partial}{\partial \theta} \{a(x)b(\theta) + c(\theta) + d(x)\} = a(x)\dot{b}(\theta) + \dot{c}(\theta).$$

うえの式の x に形式的に X を代入したものを $S(\theta|X)$ をスコア関数という。すると ^{eq:exp-5}(5.5) から

$$\begin{aligned} \mathbb{E}[S(\theta|X)] &= \mathbb{E}[a(X)\dot{b}(\theta) + \dot{c}(\theta)] = \dot{b}(\theta)\mathbb{E}[a(X)] + \dot{c}(\theta) \\ &= \dot{b} \left[-\frac{\dot{c}(\theta)}{\dot{b}(\theta)} \right] + \dot{c}(\theta) = 0. \end{aligned} \tag{5.9}$$

eq:exp-

$S(\theta|X)$ の分散を Fisher 情報量といい、 $I(\theta)$ と記すことにする。このとき

$$I(\theta) = \mathbb{V}[S(\theta|X)] = \mathbb{V}[a(X)\dot{b}(\theta) + \dot{c}(\theta)] = \{\dot{b}(\theta)\}^2 \mathbb{V}[a(X)].$$

^{eq:exp-7}(5.7) から

$$\mathbb{V}[S(\theta|X)] = \{\dot{b}(\theta)\}^2 \mathbb{V}[a(X)] = \frac{\ddot{b}(\theta)\dot{c}(\theta)}{\dot{b}(\theta)} - \ddot{c}(\theta).$$

さらに

$$\dot{S}(\theta|x) := \frac{dS(\theta|x)}{d\theta} = a(x)\ddot{b}(\theta) + \ddot{c}(\theta)$$

と ^{eq:exp-6}(5.6) から

$$\begin{aligned} \mathbb{E}[\dot{S}(\theta|X)] &= \ddot{b}(\theta)\mathbb{E}[a(X)] + \ddot{c}(\theta) = \ddot{b}(\theta) \left[-\frac{\dot{c}(\theta)}{\dot{b}(\theta)} \right] + \ddot{c}(\theta) \\ &= -\frac{\ddot{b}(\theta)\dot{c}(\theta)}{\dot{b}(\theta)} + \ddot{c}(\theta) = -I(\theta) = -\mathbb{V}[S(\theta|X)]. \end{aligned}$$

3 一般化線型モデル (Generalized linear models (GLM))

GLM は独立な確率変数列 Y_1, Y_2, \dots, Y_n に対する統計的モデルで以下で定める性質をみたく。

- 各 Y_j ($j = 1, 2, \dots, n$) の分布は同じ指数型分布族に属する。 Y_j の PDF または PMF $p_Y(y_j | \theta_j)$ は母数 $\theta_j \in \Theta \subset \mathbb{R}$ で添え字づけられ

$$p_Y(y_j | \theta_j) = \exp\{y_j b(\theta_j) + c(\theta_j) + d(y_j)\} \quad (5.10)$$

で与えられる。

- 母数 θ_j は直接興味ある対象ではなく、母数 $\beta_1, \beta_2, \dots, \beta_d$ ($d < n$) に興味がある。ここで

$$\mu_j = \mathbb{E}[Y_j]$$

と書いたとき

$$g(\mu_j) = \mathbf{x}_j^\top \boldsymbol{\beta}$$

と表現される。ただし $g: \mathbb{R} \rightarrow \mathbb{R}$ は狭義単調関数で連続微分可能である。したがって μ_j は θ_j の関数で (5.5) から

$$\mu_j = -\frac{\dot{c}(\theta_j)}{b(\theta_j)}$$

と書ける。さらに \mathbf{x}_j は非ランダムな説明変数 (既知のもの) で $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_d)^\top$ は未知の回帰係数ベクトルで

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix} \in \mathbb{R}^d, \quad \mathbf{x}_j = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jd} \end{bmatrix} \in \mathbb{R}^d,$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

である.

したがって (Y_1, Y_2, \dots, Y_n) の同時 PDF または PMF $p(y_1, y_2, \dots, y_n | \theta_1, \theta_2, \dots, \theta_n)$ は

$$\begin{aligned} p(y_1, y_2, \dots, y_n | \theta_1, \theta_2, \dots, \theta_n) &= \prod_{j=1}^n \exp\{y_j b(\theta_j) + c(\theta_j) + d(y_j)\} \\ &= \exp\left\{\sum_{j=1}^n y_j b(\theta_j) + \sum_{j=1}^n c(\theta_j) + \sum_{j=1}^n d(y_j)\right\} \end{aligned} \quad (5.11)$$

eq:glm-5

となる.

例 5.6. (ロジスティック回帰モデル)

書くこと.

□

ex:glm-1

4 GLM の最尤推定法

Y_1, Y_2, \dots, Y_n は (5.10) で与えられる GLM に従うとする. すると $Y_j = y_j$ ($j = 1, 2, \dots, n$) を観測したときの対数尤度関数 $\ell_j(\theta_j)$ は

$$\ell_j(\theta_j) = y_j b(\theta_j) + c(\theta_j) + d(y_j)$$

と書くことができる. さらに (5.5) と (5.8) から

$$\mathbb{E}[Y_j] = \mu_j = \frac{\dot{c}(\theta_j)}{\dot{b}(\theta_j)} \quad (5.12)$$

eq:glm-3

$$\mathbb{V}[Y_j] = \frac{\ddot{b}(\theta_j)\dot{c}(\theta_j) - \dot{c}(\theta_j)\ddot{b}(\theta_j)}{\{\dot{b}(\theta_j)\}^3} \quad (5.13)$$

eq:glm-4

$$g(\mu_j) = \mathbf{x}_j^\top \boldsymbol{\beta} =: \eta_j, \quad \mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jd})^\top \quad (5.14)$$

eq:glm-5

である. $(Y_1, Y_2, \dots, Y_n)^\top$ の対数尤度関数 ℓ は

$$\ell = \sum_{j=1}^n \ell_j(\theta_j) = \sum_{j=1}^n y_j b(\theta_j) + \sum_{j=1}^n c(\theta_j) + \sum_{j=1}^n d(y_j)$$

となる. β の最尤推定値を求めるために ℓ を β に関して偏微分をする. 連鎖偏微分律から

$$\frac{\partial \ell}{\partial \beta_k} = \sum_{j=1}^n \frac{\partial \ell_j}{\partial \beta_k} = \sum_{j=1}^n \frac{\partial \ell_j}{\partial \theta_j} \cdot \frac{\partial \theta_j}{\partial \mu_j} \cdot \frac{\partial \mu_j}{\partial \beta_k} \quad (k = 1, 2, \dots, d) \quad (5.15)$$

と書ける. 最右辺の各項を計算して行く. (5.12) に注意すると

$$\frac{\partial \ell_j}{\partial \theta_j} = y_j \dot{b}(\theta_j) + \dot{c}(\theta_j) = \dot{b}(\theta_j) \left\{ y_j + \frac{\dot{c}(\theta_j)}{\dot{b}(\theta_j)} \right\} = \dot{b}(\theta_j) \{y_j - \mu_j\} \quad (5.16)$$

を得る. 次に

$$\frac{\partial \theta_j}{\partial \mu_j} = 1 / \frac{\partial \mu_j}{\partial \theta_j}$$

であることと (5.12) に注意すると

$$\begin{aligned} \frac{\partial \mu_j}{\partial \theta_j} &= -\frac{\partial}{\partial \theta_j} \left(\frac{\dot{c}(\theta_j)}{\dot{b}(\theta_j)} \right) = -\frac{\ddot{c}(\theta_j)}{\dot{b}(\theta_j)} + \frac{\ddot{b}(\theta_j) \dot{c}(\theta_j)}{\{\dot{b}(\theta_j)\}^2} \\ &= \dot{b}(\theta_j) \left\{ \frac{\ddot{b}(\theta_j) \dot{c}(\theta_j) - \dot{b}(\theta_j) \ddot{c}(\theta_j)}{\{\dot{b}(\theta_j)\}^3} \right\} = \dot{b}(\theta_j) \mathbb{V}[Y_j] \end{aligned} \quad (5.17)$$

を得る. 最後の等号は (5.13) からわかる. 最後に

$$\frac{\partial \mu_j}{\partial \beta_k} = \frac{\partial \mu_j}{\partial \eta_j} \cdot \frac{\partial \eta_j}{\partial \beta_k} = \frac{\partial \mu_j}{\partial \eta_j} x_{jk} \quad (5.18)$$

を得る. (5.16)–(5.18) を (5.15) に代入することでスコア関数は

$$S_k := \sum_{j=1}^n \frac{(Y_j - \mu_j)}{\mathbb{V}[Y_j]} x_{jk} \left(\frac{\partial \mu_j}{\partial \eta_j} \right) \quad (k = 1, 2, \dots, d) \quad (5.19)$$

となることがわかる.

S_1, S_2, \dots, S_d の共分散行列を $J = (J_{kl})$ は

$$J_{kl} = \mathbb{E}[S_k S_\ell] \quad (k, \ell = 1, 2, \dots, d)$$

と書くことができる. すると (eq:g1m-10) から

$$\begin{aligned} J_{k\ell} &= \mathbb{E} \left[\sum_{i=1}^n \left\{ \frac{(Y_i - \mu_i)}{\sqrt{Y_i}} x_{ik} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right\} \sum_{j=1}^n \left\{ \frac{(Y_j - \mu_j)}{\sqrt{Y_j}} x_{j\ell} \left(\frac{\partial \mu_j}{\partial \eta_j} \right) \right\} \right] \\ &= \sum_{i=1}^n \frac{\mathbb{E}[(Y_i - \mu_i)^2]}{\{\mathbb{V}[Y_i]\}^2} x_{ik} x_{i\ell} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \end{aligned}$$

を得る. 最後から 2 番目の等号は $\mathbb{E}[(Y_i - \mu_i)(Y_j - \mu_j)] = 0 (i \neq j)$ からわかる. 結局

$$J_{k\ell} = \sum_{i=1}^n \frac{x_{ik} x_{i\ell}}{\mathbb{V}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \quad (k, \ell = 1, 2, \dots, d) \quad (5.20)$$

がわかる.

$m = 1, 2, \dots$ とする. $\hat{\beta}^{(0)}$ を β の初期推定値とし, 繰り返し計算で得られる m 回目の β の推定値を $\hat{\beta}^{(m)}$ と書くことにする. $J^{(m)} = (J_{k\ell}^{(m)})$ の $\beta = \hat{\beta}^{(m)}$ での値とし, $S^{(m)} = (S_1^{(m)}, S_2^{(m)}, \dots, S_d^{(m)})^\top$ を S_k の $\beta = \hat{\beta}^{(m)}$ での値とする. したがって (eq:fisher-1) から

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + [J^{(m-1)}]^{-1} S^{(m-1)} \quad (5.21)$$

を得る. 上式の両辺に $J^{(m-1)}$ を掛けると

$$J^{(m-1)} \hat{\beta}^{(m)} = J^{(m-1)} \hat{\beta}^{(m-1)} + S^{(m-1)} \quad (5.22)$$

を得る.

いま

$$W = (W_{ij})_{i,j=1,2,\dots,n}, \quad W_{ij} = \begin{cases} \frac{1}{\mathbb{V}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) & (i = j) \\ 0 & (i \neq j) \end{cases} \quad (5.23)$$

とおき, $\beta = \hat{\beta}^{(m)}$ での値を $W^{(m)}$ と書く. すると

$$J^{(m)} = X^\top W^{(m)} X$$

と書ける. (eq:g1m-13) の右辺のベクトルの第 j 成分 ($j = 1, 2, \dots, d$) は

$$\sum_{k=1}^d \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\mathbb{V}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \hat{\beta}_k^{(m-1)} + \sum_{i=1}^n \frac{(Y_i - \mu_i) x_{ij}}{\mathbb{V}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)$$

と書ける。ただし $\mathbb{V}[Y_i]$, μ_i , $\frac{\partial \mu_i}{\partial \eta_i}$ も $\beta = \hat{\beta}^{(m-1)}$ での値であることに注意せよ。さらに

$$\begin{aligned} \mathbf{z}^{(m)} &= (z_1^{(m)}, z_2^{(m)}, \dots, z_d^{(m)})^\top, \\ z_i^{(m)} &= \sum_{k=1}^d x_{ik} \hat{\beta}_k^{(m)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) + (Y_i - \mu_i) \end{aligned} \quad (5.24)$$

eq:g1m-

とおくと (5.22) の右辺は

$$\mathbf{X}^\top \mathbf{W}^{(m-1)} \hat{\mathbf{z}}^{(m-1)}$$

と書けるので

$$\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{X} \hat{\beta}^{(m)} = \mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}$$

を得る。なぜならば $\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}$ は第 j 成分を $\{\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}\}_j$ ($j = 1, 2, \dots, d$) と書くと

$$\begin{aligned} \{\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}\}_j &= \sum_{i=1}^n x_{ij} W_{ii}^{(m-1)} \hat{z}_i \\ &= \sum_{i=1}^n x_{ij} \frac{1}{\mathbb{V}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \left\{ \sum_{k=1}^d x_{ik} \hat{\beta}_k^{(m)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) + (Y_i - \mu_i) \right\} \\ &\quad \left(\because \text{eq:g1m-13} \text{ と } \text{eq:g1m-14} \text{ を代入} \right) \\ &= \sum_{k=1}^d \sum_{i=1}^n \frac{x_{ij} x_{ik}}{\mathbb{V}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 + \sum_{i=1}^n \frac{(Y_i - \mu_i) x_{ij}}{\mathbb{V}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \end{aligned}$$

からわかる。

例 5.7. (Poisson 回帰)

y_j	2	3	6	7	8	9	10	11	12
x_j	-1	-1	0	0	0	0	1	1	1

応答変数 Y_j は Poisson 分布に従う確率変数とする。すると

$$\mu_j = \mathbb{E}[Y_j] = \mathbb{V}[Y_j] \quad (j = 1, 2, \dots, n; n = 9) \quad (5.25)$$

eq:g1m-

となる. μ_j と x_j の間に線型関係を仮定する.

$$\mathbb{E}[Y_j] = \mu_j = \beta_1 + \beta_2 x_j = \mathbf{x}_j^\top \boldsymbol{\beta},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \mathbf{x}_j = \begin{bmatrix} 1 \\ x_j \end{bmatrix} \quad (j = 1, 2, \dots, n)$$

である. したがってリンク関数 g は恒等関数である.

$$g(\mu_j) = \mu_j = \mathbf{x}_j^\top \boldsymbol{\beta} =: \eta_j.$$

これより

$$\frac{\partial \mu_j}{\partial \eta_j} = 1$$

となる. [\(5.23\)](#) と [\(5.25\)](#) から

$$W_{jj} = \frac{1}{\mathbb{V}[Y_j]} = \frac{1}{\beta_1 + \beta_2 x_j} \quad (j = 1, 2, \dots, n)$$

となる. $\mu_j = \mathbf{x}_j^\top \hat{\boldsymbol{\beta}} = \hat{\beta}_1 + \hat{\beta}_2 x_j$ を [\(5.24\)](#) に代入すると

$$z_j = \hat{\beta}_1^{(m-1)} + \hat{\beta}_2^{(m-1)} x_j + y_j - \hat{\beta}_1^{(m-1)} - \hat{\beta}_2^{(m-1)} x_j = y_j$$

となる. また

$$\mathbf{J} = \mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{X} = \begin{bmatrix} \sum_{j=1}^n \frac{1}{\hat{\beta}_1^{(m-1)} + \hat{\beta}_2^{(m-1)} x_j} & \sum_{j=1}^n \frac{x_j}{\hat{\beta}_1^{(m-1)} + \hat{\beta}_2^{(m-1)} x_j} \\ \sum_{j=1}^n \frac{x_j}{\hat{\beta}_1^{(m-1)} + \hat{\beta}_2^{(m-1)} x_j} & \sum_{j=1}^n \frac{x_j^2}{\hat{\beta}_1^{(m-1)} + \hat{\beta}_2^{(m-1)} x_j} \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{z} = \begin{bmatrix} \sum_{j=1}^n \frac{y_j}{\hat{\beta}_1^{(m-1)} + \hat{\beta}_2^{(m-1)} x_j} \\ \sum_{j=1}^n \frac{x_j y_j}{\hat{\beta}_1^{(m-1)} + \hat{\beta}_2^{(m-1)} x_j} \end{bmatrix}$$

となる. よって

$$\mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{X} \hat{\boldsymbol{\beta}}^{(m)} = \mathbf{X}^\top \mathbf{W}^{(m-1)} \mathbf{y}$$

を得る.

$\hat{\beta}_1^{(0)} = 7, \hat{\beta}_2^{(0)} = 5$ とする.

$$\mathbf{y} = \mathbf{z} = \begin{bmatrix} 2 \\ 3 \\ \vdots \\ 12 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_9^\top \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

なので

$$\mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{X} = \begin{bmatrix} 1.821429 & -0.75 \\ -0.75 & 1.25 \end{bmatrix}, \quad \mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{y} = \begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix}$$

なので

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(1)} &= [\mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{W}^{(0)} \mathbf{y} \\ &= \begin{bmatrix} 0.729167 & 0.4375 \\ 0.4375 & 1.0625 \end{bmatrix} \begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix} = \begin{bmatrix} 7.4514 \\ 4.9375 \end{bmatrix} \end{aligned}$$

のように計算される。最尤推定値は

$$\hat{\beta}_1 = 7.45163, \quad \hat{\beta}_2 = 4.93530$$

となり、この値における J の逆行列の値は

$$\mathbf{J}^{-1} = \begin{bmatrix} 0.7817 & 0.4166 \\ 0.4166 & 1.1863 \end{bmatrix}$$

となる。繰り返し計算の結果は下記ようになった。

m	0	1	2	3
$\hat{\beta}_1^{(m)}$	7	7.45139	7.45163	7.45163
$\hat{\beta}_2^{(m)}$	5	4.93750	4.93531	4.93530

□

第6章 正則化とカーネル法

1 導入

訓練データの実現値を $t_n = \{(x_j, y_j); j = 1, 2, \dots, n\}$ とする. ただし $x_j \in \mathbb{R}^d, y_j \in \mathbb{R}$ である. 予測関数 g_t をみつけるために探索する関数族 $\mathcal{G} = \{g: \mathbb{R}^d \mapsto \mathbb{R}\}$ を定めて, $g \in \mathcal{G}$ に関して

$$\widehat{\text{Err}}_t(g) := \frac{1}{n} \sum_{j=1}^n (y_j - g(x_j))^2$$

を最小にする関数 $g_t \in \mathcal{G}$ をみつきたい. 関数族 \mathcal{G} をすべての \mathbb{R}^d 上の実数値関数全体がなす集合とすると

$$g(x_j) = y_j \quad (j = 1, 2, \dots, n)$$

をみたく任意の関数 g は $\widehat{\text{Err}}_t(g) = 0$ をみたく. しかしこの g の汎化誤差はよいことが期待できない. いわゆる過学習が起きていることになる.

予測の観点から問題を考える. (X, Y) は (x_j, y_j) を発生させて確率分布と同じ分布に従う確率ベクトルとする. 任意の関数 $g: \mathbb{R}^d \rightarrow \mathbb{R}$ に対して, 2乗誤差

$$\mathbb{E}[\{Y - g(X)\}^2]$$

を考える. このとき 2乗誤差を最小にする g は

$$g^*(x) = \mathbb{E}[Y | X = x]$$

で与えられた. 関数族 \mathcal{G} は解釈が容易であるような簡単なもので, かつ最適の g^* を含むような豊かなものが望ましい.

関数族 \mathcal{G} をよりよく理解するために \mathcal{G} を Hilbert 空間 (完備な内積空間) とみなす. たとえば \mathcal{G} を \mathbb{R}^d 上の線型関数すべてのなす集合とすれば, これは

Hilbert 空間とみなすことができる. $\mathcal{G} = \{g_\beta(x) = x^\top \beta; \beta \in \mathbb{R}^d\}$ の元 g_β と $\beta \in \mathbb{R}^d$ を同一視し, \mathcal{G} 上の内積を

$$\langle g_{\beta_1}, g_{\beta_2} \rangle = \beta_1^\top \beta_2,$$

で定義する. ただし $g_{\beta_1}(x) = x^\top \beta_1, g_{\beta_2}(x) = x^\top \beta_2, \beta_1, \beta_2 \in \mathbb{R}^d$ である. すなわち $\langle \cdot, \cdot \rangle$ は \mathbb{R}^d 上の Euclid 内積である. するとこの内積 $\langle \cdot, \cdot \rangle$ をもつ \mathcal{G} は Hilbert 空間になること¹がわかる.

ex:6-1

例 6.1. () 多項式回帰訓練データの実現値を

$$t_n = \{(u_1, y_1), (u_2, y_2), \dots, (u_n, y_n)\}$$

とする. ただし $u_j, y_j \in \mathbb{R} (j = 1, 2, \dots, n)$ である. $d \in \mathbb{N}, d \geq 2$ とし, 変換

$$\phi(u) := \begin{pmatrix} 1 \\ u \\ u^2 \\ \vdots \\ u^{d-1} \end{pmatrix}, \mathbb{R} \longrightarrow \mathbb{R}^d$$

を考える. すると $\beta \in \mathbb{R}^d$ としたとき, 関数

$$g_\beta : u \mapsto \phi(u)^\top \beta$$

全体のなす集合を \mathcal{G} としたとき, 対応

$$h_\beta \longleftrightarrow \beta$$

と \mathbb{R}^d 上の Euclid 内積により \mathcal{G} は Hilbert 空間となる. □

回帰問題において線型空間では単純すぎるが, 複雑すぎる関数の利用は避けたいところである. そのために, 問題そのものを「代入が内積で表現される空間 (再生核 Hilbert 空間)」に変換しその変換先での線型関数を考える方法がカーネル法である. カーネル法では, 半正定値行列を一般化したカーネル関数とよば

¹これは実数の完備性より直ちに確認できる.

れる 2 変数関数 k を考え、データ \mathbf{x}_j を $k\mathbf{x}_j = k(\cdot, \mathbf{x}_j)$ に変換することをカーネルトリックとよぶ。特に写像

$$\phi: \mathbf{x}_j \mapsto k\mathbf{x}_j$$

は特徴写像とよばれる。データを \mathbb{R}^d とは曲り具合の異なる高次元空間に埋め込むことに相当する。この高次元空間のことを特徴空間とよぶ。

2 正則化

sec:6-2

$d \in \mathbb{N}$, $d \geq 2$ とする。訓練データの実現値を $t_n = \{(\mathbf{x}_j, y_j); j = 1, 2, \dots, n\}$ とする。ただし $\mathbf{x}_j \in \mathbb{R}^d$, $y_j \in \mathbb{R}$ とする。 \mathbb{R}^d 上の実数値関数の族 \mathcal{G} に関してある損失関数 $\text{Loss}: \mathbb{R}^2 \rightarrow \mathbb{R}$ を用いた訓練誤差

$$\widehat{\text{Err}}_t(g) := \frac{1}{n} \sum_{j=1}^n \text{Loss}(g(\mathbf{x}_j), y_j) \quad (g \in \mathcal{G})$$

を最小にする $g_t \in \mathcal{G}$ をみつけたい。関数族 \mathcal{G} はある内積に関して Hilbert 空間になっているとする。 \mathcal{G} が豊かすぎると訓練誤差は零または零に近くなり、過学習を起こす。過学習をさけるためのひとつの方法はモデルの複雑さに罰則

$$J: \mathcal{G} \rightarrow [0, \infty) \cup \{+\infty\}$$

を導入し、罰則付きの訓練誤差の最小化問題を考えることである。すなわち

$$\min_{g \in \mathcal{G}} \{ \widehat{\text{Err}}_t(g) + J(g) : g \in \mathcal{G} \}$$

である。

ex:6-2

例 6.2. (リッジ回帰) リッジ回帰は 2 上ノルムの罰則項をもった線型回帰である。 $d, n \in \mathbb{N}$ とする。訓練データの実現値を $t_n = \{(\mathbf{x}_j, y_j); j = 1, 2, \dots, n\}$ とする。ただし $\mathbf{x}_j \in \mathbb{R}^d$, $y_j \in \mathbb{R}$ である。関数族

$$\mathcal{G} = \{g(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^d\}$$

を考え、 $\gamma > 0$ に対して

$$\min_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{j=1}^n (y_j - g(\mathbf{x}_j))^2 + \gamma \|g\|^2 \right\} \quad (6.1)$$

eq:6-1

を考える. ただし $\|\cdot\|$ は \mathcal{G} の内積から定義される \mathcal{G} 上のノルムである. $g \in \mathcal{G}$ に対して $\beta \in \mathbb{R}^d$ を対応させると

$$\|g\|^2 = \langle \beta, \beta \rangle = |\beta|_{2,d}^2$$

となる. ただし, $|\cdot|_{2,d}$ は Euclid \mathbb{R}^d のノルムである. したがって, 問題は

$$\min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_{2,n}^2 + \gamma |\beta|_{2,d}^2 \right\} = \min_{\beta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{j=1}^n (y_j - \mathbf{x}_j^\top \beta)^2 + \gamma |\beta|_{2,d}^2 \right\} \quad (6.2)$$

eq:6-2

と書きなおすことができる. ただし

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

とした. 最小化問題 (6.1) の解は

$$\mathbf{x} \mapsto \mathbf{x}^\top \beta$$

という形となり β^* は最小化問題 (6.1) の解である. $\gamma \rightarrow \infty$ のとき罰則項が支配的になり $g \equiv 0$ となる.

問題 (6.2) は凸問題 (対象の関数が凸関数) なので勾配を零とおく.

$$h(\beta) := \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_{2,n}^2 + \gamma |\beta|_{2,d}^2$$

とおいたとき

$$\frac{\partial h}{\partial \beta} = \frac{1}{n} \mathbf{X}^\top (\mathbf{X}\beta - \mathbf{y}) + \gamma \beta = 0 \quad (6.3)$$

eq:6-3

とおく. $\gamma = 0$ のとき上の方程式は正規方程式になる. $\mathbf{X}^\top \mathbf{X} + I_d$ が正則のとき

$$\hat{\beta}_\gamma := (\mathbf{X}^\top \mathbf{X} + \gamma I_d)^{-1} \mathbf{X}^\top \mathbf{y}$$

が方程式 (6.3) の解となる. $\hat{\beta}_\gamma$ のことをリッジ回帰推定量という. \square

eq:6-3

ある Hilbert 空間 \mathcal{G} に対して罰則項を用いるとき \mathcal{G} を 2 つの直交部分空間 \mathcal{H} と \mathcal{C} に分解するとよい. すなわち $\forall g \in \mathcal{G}$ に対して, ある $h \in \mathcal{H}$ と $c \in \mathcal{C}$ が一意的に存在して

$$g = h + c, \quad \langle h, c \rangle_{\mathcal{G}} = 0$$

となる. ただし $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ は Hilbert 空間 \mathcal{G} の内積である. Hilbert 空間 \mathcal{G} の任意の元がこのような分解をもつとき, \mathcal{G} は \mathcal{H} と \mathcal{C} の直和であるといい, $\mathcal{G} = \mathcal{H} \oplus \mathcal{C}$ と書く.

例 6.3. (例 ^{ex:6-2}6.2 の続き)

$$\tilde{\mathbf{x}}_j = (1, \mathbf{x}_j^\top)^\top \quad (j = 1, 2, \dots, n)$$

とする. 第 1 成分に対応する回帰係数(切片)には罰則を加えない. $\tilde{\mathbf{x}} = (1, \mathbf{x}^\top)^\top$ とおいたとき

$$\mathcal{G} := \{g(\tilde{\mathbf{x}}) = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}; \beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^d\}$$

とし

$$\widehat{\text{Err}}(g) = \frac{1}{n} \|\mathbf{y} - \tilde{\mathbf{X}} \tilde{\boldsymbol{\beta}}\|_{2,n}^2 + \gamma |\boldsymbol{\beta}|_{2,d}^2$$

の最小化問題を考える. ただし $\gamma > 0$ で

$$\tilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X}), \quad \tilde{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix}, \quad \mathbf{1} = \underbrace{(1, 1, \dots, 1)}_n^\top$$

である. すると $\gamma \rightarrow \infty$ のとき $g(\tilde{\mathbf{x}}) = \beta_0$ となる. $g \in \mathcal{G}$ は

$$\tilde{\mathbf{x}} \mapsto \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$$

である. さらに関数族 \mathcal{C} を

$$c: \tilde{\mathbf{x}} \mapsto \beta_0$$

とし, \mathcal{H} を

$$h: \tilde{\mathbf{x}} \mapsto \mathbf{x}^\top \boldsymbol{\beta}$$

とする. このとき \mathcal{G} の内積を $(1, \boldsymbol{\beta}^\top)^\top$ に対する \mathbb{R}^{d+1} 上の Euclid 内積とすれば

$$\mathcal{G} = \mathcal{C} \oplus \mathcal{H}$$

となる.

よって

$$\min_{g \in \mathcal{H} \oplus \mathcal{C}} \left\{ \frac{1}{n} \sum_{j=1}^n (y_j - g(\tilde{x}))^2 + \gamma \|g\|_{\mathcal{H}} \right\} \quad (6.4) \quad \text{eq:6-4}$$

となる. ただし $\|\cdot\|_{\mathcal{H}}$ は \mathcal{H} のノルムである. 問題 (6.4) は

$$\min_{\beta_0, \beta} \left\{ \frac{1}{n} \|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\beta\|_{2,n}^2 + \gamma \|\beta\|_{2,d}^2 \right\} =: \min_{\beta_0, \beta} h(\beta_0, \beta) \quad (6.5) \quad \text{eq:6-5}$$

と書きなおすことができる. $\gamma \rightarrow \infty$ のとき

$$g(\tilde{x}) \rightarrow \bar{y}_n = \frac{1}{n} \sum_{j=1}^n y_j$$

となる. 問題 (6.5) も凸なので

$$\frac{\partial h}{\partial \beta}(\beta_0, \beta) = \mathbf{X}^{\top}(\beta_0 \mathbf{1} + \mathbf{X}\beta - \mathbf{y}) + n\gamma\beta = \mathbf{0} \quad (6.6) \quad \text{eq:6-6}$$

$$\frac{\partial h}{\partial \beta_0}(\beta_0, \beta) = n\beta_0 - \mathbf{1}^{\top}(\mathbf{y} - \mathbf{X}\beta) = 0 \quad (6.7) \quad \text{eq:6-7}$$

を解けばよい. (6.7) の解を (6.6) に代入すれば β をみつけるためには方程式

$$(\mathbf{X}^{\top} \mathbf{X} - n^{-1} \mathbf{X}^{\top} \mathbf{1} \mathbf{1}^{\top} \mathbf{X} + n\gamma \mathbf{I}_d) \beta = (\mathbf{X}^{\top} - n^{-1} \mathbf{X}^{\top} \mathbf{1} \mathbf{1}^{\top}) \mathbf{y} \quad (6.8) \quad \text{eq:6-8a}$$

を解けばよい.

議論を簡単にするために $n \geq d$ かつ $\text{rank } \mathbf{X} = d$ ($\iff \mathbf{X}^{\top} \mathbf{X}$ は正則) を仮定する. すると

$$\begin{aligned} (6.6) \iff \beta_0 \mathbf{X}^{\top} \mathbf{1} + \mathbf{X}^{\top} \mathbf{X} \beta - \mathbf{X}^{\top} \mathbf{y} + n\gamma \beta &= \mathbf{0} \\ \iff (\mathbf{X}^{\top} \mathbf{X} + n\gamma \mathbf{I}_d) \beta &= \mathbf{X}^{\top} (\mathbf{y} - \beta_0 \mathbf{1}) \end{aligned}$$

なので β は x_1, x_2, \dots, x_n の一次結合で書ける. よって

$$\beta = \mathbf{X}^{\top} \alpha, \quad \alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^{\top} \in \mathbb{R}^n$$

とおく. これを (6.8) に代入すれば

$$(\mathbf{X} \mathbf{X}^{\top} - n^{-1} \mathbf{1} \mathbf{1}^{\top} \mathbf{X} \mathbf{X}^{\top} + n\gamma \mathbf{I}_n) \alpha = (\mathbf{I}_n - n^{-1} \mathbf{1} \mathbf{1}^{\top}) \mathbf{y}$$

を得る. $\mathbf{X}\mathbf{X}^\top - n^{-1}\mathbf{1}\mathbf{1}^\top + n\gamma\mathbf{I}_n$ は正則と仮定すると

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}\mathbf{X}^\top - n^{-1}\mathbf{1}\mathbf{1}^\top + n\gamma\mathbf{I}_n)^{-1}(\mathbf{I}_n - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{y} \quad (6.9)$$

eq:6-8

となる. ここで

$$\mathbf{X}\mathbf{X}^\top = \begin{pmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle_{2,d} & \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_{2,d} & \cdots & \langle \mathbf{x}_1, \mathbf{x}_n \rangle_{2,d} \\ \langle \mathbf{x}_2, \mathbf{x}_1 \rangle_{2,d} & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle_{2,d} & \cdots & \langle \mathbf{x}_2, \mathbf{x}_n \rangle_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{x}_n, \mathbf{x}_1 \rangle_{2,d} & \langle \mathbf{x}_n, \mathbf{x}_2 \rangle_{2,d} & \cdots & \langle \mathbf{x}_n, \mathbf{x}_n \rangle_{2,d} \end{pmatrix}$$

となることに注意すると $\hat{\boldsymbol{\alpha}}$ は $\{\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{2,d}; (i, j = 1, 2, \dots, n)\}$ のみを通して $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ に依存することがわかる. すなわち 各 $\mathbf{x}_j (j = 1, 2, \dots, n)$ の値がわからなくとも $\{\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{2,d}; (i, j = 1, 2, \dots, n)\}$ の値がわかれば $\hat{\boldsymbol{\alpha}}$ の値は計算できる. さらに (6.9) を (6.7) に代入すれば

$$\hat{\beta}_0 = \frac{1}{n}\mathbf{1}^\top(\mathbf{y} - \mathbf{X}\mathbf{X}^\top\hat{\boldsymbol{\alpha}})$$

となる. よって

$$g_i(\tilde{\mathbf{x}}) = \hat{\beta}_0 + \mathbf{x}^\top \mathbf{X}^\top \hat{\boldsymbol{\alpha}} = \hat{\beta}_0 + \sum_{j=1}^n \hat{\alpha}_j \langle \mathbf{x}, \mathbf{x}_j \rangle_{2,d}, \quad \hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_n)^\top$$

となる. $\hat{\beta}_0$ と $\hat{\alpha}_j (j = 1, 2, \dots, n)$ は内積 $\{\langle \mathbf{x}_i, \mathbf{x}_j \rangle_{2,d}; i, j = 1, 2, \dots, n\}$ の値がわかればよい. さらに $g_i(\tilde{\mathbf{x}})$ の値を求めるためには, $\{\langle \mathbf{x}_j, \mathbf{x}_j \rangle_{2,d}; j = 1, 2, \dots, n\}$ の値がわかればよい. 内積だけの計算になっていることが肝である.

3 カーネル関数

sec:6-3

df:6-4

定義 6.4. \mathcal{X} を空でない集合とし k を \mathcal{X} 上の 2 変数関数とする. k が次の 2 条件をみたすとき, k は \mathcal{X} 上のカーネル関数とよばれる.

(i) $\forall x, y \in \mathcal{X}$ に対し

$$k(x, y) = k(y, x). \quad (\text{対称性})$$

(ii) $\forall n \in \mathbb{N}, \{x_j\}_{j=1}^n \subset \mathcal{X}, \{c_j\}_{j=1}^n \subset \mathbb{R}$ に対し

$$\sum_{j, \ell=1}^n c_j c_\ell k(x_j, x_\ell) \geq 0. \quad (\text{半正定値性})$$

定義 $\overset{\text{df: 6-4}}{6.4(\text{ii})}$ は, $\forall n \in \mathbb{N}$ と $\{x_j\}_{j=1}^n \subset \mathcal{X}$ に対し

$$\begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{pmatrix} \text{は半正定値}$$

となる.

重要なカーネル関数の例をあげておく.

ex: 6-5

例 6.5. f を \mathcal{X} 上の関数とする. このとき

$$k(x, y) = f(x)f(y) \quad (x, y \in \mathcal{X})$$

はカーネル関数となる. 実際, $\forall n \in \mathbb{N}, \{x_j\}_{j=1}^n \subset \mathcal{X}, \{c_j\}_{j=1}^n \subset \mathbb{R}$ に対し

$$\sum_{j, \ell=1}^n c_j c_\ell k(x_j, x_\ell) = \sum_{j, \ell=1}^n c_j c_\ell f(x_j)f(x_\ell) = \left(\sum_{j=1}^n c_j f(x_j) \right)^2 \geq 0$$

より半正定値性はわかる. 対称性は定義より明らか. □

ex: 6-6

例 6.6. k_1, k_2 を \mathcal{X} 上のカーネル関数とする. このとき

$$(k_1 + k_2)(x, y) := k_1(x, y) + k_2(x, y) \quad (x, y \in \mathcal{X})$$

はカーネル関数. □

pro: 6-1

問 6.1. 例 $\overset{\text{ex: 6-6}}{6.6}$ の k の対称性と半正定値性を確認せよ.

ex: 6-7

例 6.7. $d \in \mathbb{N}, \mathcal{X}$ を空でない集合とし

$$\phi: \mathcal{X} \rightarrow \mathbb{R}^d$$

を \mathcal{X} から \mathbb{R}^d への写像とする. \mathbb{R}^d 上の任意の内積 $\langle \cdot, \cdot \rangle$ を考える. このとき

$$k(x, y) = \langle \phi(x), \phi(y) \rangle \quad (x, y \in \mathcal{X})$$

はカーネル関数である. 上式の右辺の ϕ のことを特徴写像, \mathbb{R}^d を特徴空間という.

pro:6-1

問 6.2. 内積の性質を用いて例 6.7 の k の対称性と半正定値性を確認せよ.

4 カーネル関数から内積空間の構成

sec:6-4

\mathcal{X} を空でない集合とし k を \mathcal{X} 上のカーネル関数とする. カーネル関数 k と $x \in \mathcal{X}$ に対し, \mathcal{X} 上の関数 k_x を

$$k_x : \mathcal{X} \longrightarrow \mathbb{R}, \quad y \mapsto k(x, y)$$

で定める. $\{k_x\}_{x \in \mathcal{X}}$ で張られる部分ベクトル空間 \mathcal{V} を

$$\mathcal{V} := \left\{ \sum_{j=1}^n c_j k_{x_j}; \text{ある } n \in \mathbb{N} \text{ があって, } \{x_j\}_{j=1}^n \subset \mathcal{X}, \{c_j\}_{j=1}^n \subset \mathbb{R} \right\}$$

で定める. $f = \sum_{j=1}^n a_j k_{x_j}$, $g = \sum_{j=1}^m b_j k_{y_j} \in \mathcal{V}$ ($a_j, b_j \in \mathbb{R}$, $x_j, y_j \in \mathcal{X}$, $m, n \in \mathbb{N}$) に対し, 写像 $\langle \cdot, \cdot \rangle_{\mathcal{V}} : \mathcal{V} \times \mathcal{V} \longrightarrow \mathbb{R}$ を

$$\langle f, g \rangle_{\mathcal{V}} := \left\langle \sum_{j=1}^n a_j k_{x_j}, \sum_{j=1}^m b_j k_{y_j} \right\rangle_{\mathcal{V}} = \sum_{j=1}^n \sum_{\ell=1}^m a_j b_{\ell} k(y_{\ell}, x_j)$$

と定める. するとカーネル関数 k の半正定値性より $\langle \cdot, \cdot \rangle_{\mathcal{V}} \geq 0$ となる. さらに $f, g \in \mathcal{V}$ の表し方に依らず $\langle f, g \rangle_{\mathcal{V}}$ は一意的に定まる. 実際, $x \in \mathcal{X}$ に対し

$$\sum_{j=1}^n a_j k_{x_j}(x) = \sum_{j=1}^{m'} a'_j k_{x'_j}(x), \quad \sum_{j=1}^n b_j k_{y_j}(x) = \sum_{j=1}^{m'} b'_j k_{y'_j}(x)$$

とする. このとき $k(x, y) = k_y(x)$ と $k(x, y) = k(y, x)$ より

$$\begin{aligned}
 \sum_{j, \ell} a_j b_\ell k(y_\ell, x_j) &= \sum_j a_j \sum_\ell b_\ell k(x_j, y_\ell) = \sum_j a_j \sum_\ell b_\ell k_{y_\ell}(x_j) \\
 &= \sum_j a_j \sum_\ell b'_\ell k_{y'_\ell}(x_j) = \sum_j a_j \sum_\ell b'_\ell k(x_j, y'_\ell) \\
 &= \sum_j a_j \sum_\ell b'_\ell k(y'_\ell, x_j) = \sum_\ell b'_\ell \sum_j a_j k(y'_\ell, x_j) \\
 &= \sum_\ell b'_\ell \sum_j a_j k_{x_j}(y'_\ell) = \sum_\ell b'_\ell \sum_j a_j k_{x'_j}(y'_\ell) \\
 &= \sum_{j, \ell} b'_\ell a'_j k(y'_\ell, x'_j)
 \end{aligned}$$

が成り立つ. よって $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ は f, g の表現の仕方によらず f, g で一意的に定まる.

次に, $f, g, h \in \mathcal{V}$ と $\forall \alpha \in \mathbb{R}$ に対し,

- (1) $\langle f, g + h \rangle_{\mathcal{V}} = \langle f, g \rangle_{\mathcal{V}} + \langle f, h \rangle_{\mathcal{V}}$,
- (2) $\langle f, g \rangle_{\mathcal{V}} = \langle g, f \rangle_{\mathcal{V}}$,
- (3) $\langle \alpha f, g \rangle_{\mathcal{V}} = \alpha \langle f, g \rangle_{\mathcal{V}}$

が成り立つ.

(2), (3) は明らか. (1) だけを示す.

(1) の証明: $f = \sum_{j=1}^n a_j k_{x_j}$, $g = \sum_{j=1}^m b_j k_{y_j}$, $\gamma k_z (a_j, b_j, z \in \mathbb{R}, x_j, y_j, z \in \mathcal{X})$ に対し, $b_{m+1} = \gamma$, $y_{m+1} = z$ とおく. すると

$$\begin{aligned}
 \langle f, g + h \rangle_{\mathcal{V}} &= \sum_{j=1}^n \sum_{\ell=1}^m a_j b_\ell k(y_\ell, x_j) + \sum_{j=1}^n a_j b_{m+1} k(y_{m+1}, x_j) \\
 &= \sum_{j=1}^n \sum_{\ell=1}^m a_j b_\ell k(y_\ell, x_j) + \sum_{j=1}^n a_j \gamma k(z, x_j) \\
 &= \langle f, g \rangle_{\mathcal{V}} + \langle f, \gamma k_z \rangle_{\mathcal{V}}
 \end{aligned}$$

が成り立つ. ここで示した等式を繰り返し用いれば, (1) が得られる.

(4) $\forall \mathcal{V}$ に対し

$$f(x) = \langle f, k_x \rangle_{\mathcal{V}} \quad (x \in \mathcal{V})$$

が成り立つ.

証明: $f = \sum_{j=1}^m a_j k_{x_j}$ とおく. このとき

$$\begin{aligned}\langle f, k_x \rangle_{\mathcal{V}} &= \left\langle \sum_{j=1}^m a_j k_{x_j}, k_x \right\rangle_{\mathcal{V}} = \sum_{j=1}^m a_j \langle k_{x_j}, k_x \rangle_{\mathcal{V}} = \sum_{j=1}^m a_j k(x, x_j) \\ &= \sum_{j=1}^m a_j k(x_j, x) = \sum_{j=1}^m a_j k_{x_j}(x) = f(x).\end{aligned}$$

(5) (Cauchy-Schwarz の不等式) $\forall f, g \in \mathcal{V}$ に対し

$$|\langle f, g \rangle_{\mathcal{V}}| \leq \langle f, f \rangle_{\mathcal{V}} \langle g, g \rangle_{\mathcal{V}}. \quad (6.10)$$

eq:6-9

等号成立は $f = cg$ ($c \in \mathbb{R}$) のとき.

証明: $\forall t \in \mathbb{R}$ に対し

$$\langle tf + g, tf + g \rangle_{\mathcal{V}} = t^2 \langle f, f \rangle_{\mathcal{V}} + 2t \langle f, g \rangle_{\mathcal{V}} + \langle g, g \rangle_{\mathcal{V}} \geq 0$$

である. $\langle f, f \rangle_{\mathcal{V}} = 0$ のとき,

$$2t \langle f, g \rangle_{\mathcal{V}} + \langle g, g \rangle_{\mathcal{V}} \geq 0$$

となる. t は任意であったので, $\langle f, g \rangle_{\mathcal{V}} = 0$. よって, (6.10) は成立.

$\langle f, f \rangle_{\mathcal{V}} \neq 0$ のとき,

$$\langle f, f \rangle_{\mathcal{V}} \left\{ t + \frac{\langle f, g \rangle_{\mathcal{V}}}{\langle f, f \rangle_{\mathcal{V}}} \right\}^2 + \frac{\langle f, f \rangle_{\mathcal{V}} \langle g, g \rangle_{\mathcal{V}} - \langle f, g \rangle_{\mathcal{V}}^2}{\langle f, f \rangle_{\mathcal{V}}} \geq 0$$

より

$$\frac{\langle f, f \rangle_{\mathcal{V}} \langle g, g \rangle_{\mathcal{V}} - \langle f, g \rangle_{\mathcal{V}}^2}{\langle f, f \rangle_{\mathcal{V}}} \geq 0 \iff |\langle f, g \rangle_{\mathcal{V}}|^2 \leq \langle f, f \rangle_{\mathcal{V}} \langle g, g \rangle_{\mathcal{V}}$$

がわかる.

以上の議論により, k によって構成された, \mathcal{V} は $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ を内積としてもつ内積空間となることがわかった. あとは, \mathcal{V} を完備化することになる. 完備化された空間を $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ と書くことにする.

定理 6.8. (Moore-Aronszajn の定理) \mathcal{X} 上のカーネル関数 k に対し, \mathcal{X} の再生核 Hilbert 空間 $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ で以下の 3 条件をみたすものが一意的に存在する.

- (1) $\forall x \in \mathcal{X}$ に対し, $k(\cdot, x) \in \mathcal{H}_k$.
- (2) $\text{span}\{k(\cdot, x) \in x \in \mathcal{X}\}$ は \mathcal{H}_k の稠密な部分集合. すなわち, $\forall \epsilon > 0$ と $\forall h \in \mathcal{H}$ に対し, ある $v \in \mathcal{V}$ が存在して,

$$\|f - v\|_{\mathcal{H}} < \epsilon$$

となる. ただし, $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.

- (3) k は \mathcal{H}_k の再生核. すなわち,

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \quad (\forall x \in \mathcal{X}, \forall f \in \mathcal{H}).$$

Proof. 福水 (2010, pp.18-19). □

命題 6.9. k を空ではない位相空間 \mathcal{X} 上の半正定値カーネル. \mathcal{H} を対応する再生核 Hilbert 空間とする. 関数 $x \mapsto k(x, x)$ は連続で, 任意の $y \in \mathcal{X}$ に対し, $x \mapsto \text{tok}(x, y)$ が $x = y$ において連続であれば, \mathcal{H} に属するすべての関数は \mathcal{X} 上で連続である. とくに $\mathcal{X} \times \mathcal{X}$ 上で連続な半正定値カーネルにより定まる再生核 Hilbert 空間は連続関数からなる.

Proof. 福水 (2010, p.19). □

定理 6.10. (リプレゼンタ定理) k を \mathcal{X} 上のカーネル関数とし, \mathcal{H} を k に対応する再生核 Hilbert 空間とする. $\mathfrak{X} = \{x_1, x_2, \dots, x_n\} \subset \mathcal{X}$ をデータとする. Loss を n 変数の関数で, pen は \mathbb{R} 上の非減少関数, $b \in \mathbb{R}$ とする. この設定で

$$\text{Loss}(f(x_1) + b, f(x_2) + b, \dots, f(x_n) + b) + \text{pen}(\|f\|_{\mathcal{H}}^2)$$

を \mathcal{H} の中で最小にするものは

$$f = \sum_{j=1}^n \xi_j k_{x_j} \quad (\xi_j \in \mathbb{R}, j = 1, 2, \dots, n)$$

と仮定してよい.

Proof. \mathbf{P} を $\text{span}\{k_{x_1}, k_{x_2}, \dots, k_{x_n}\} =: \mathcal{V}$ への直交射影とする。 \mathbf{P} は、
 $\langle \mathbf{P}f, g \rangle_{\mathcal{H}} = \langle f, \mathbf{P}g \rangle_{\mathcal{H}} (\forall f, g \in \mathcal{H})$ かつ $\mathbf{P}f = f (\forall f \in \mathcal{V})$ である。 $f \in \mathcal{H}$ に対し

$$f(x_j) = \langle f, k_{x_j} \rangle_{\mathcal{H}} = \langle f, \mathbf{P}k_{x_j} \rangle_{\mathcal{H}} = \langle \mathbf{P}f, k_{x_j} \rangle_{\mathcal{H}} = (\mathbf{P}f)(x_j)$$

となる。したがって、

$$\begin{aligned} & \text{Loss}(f(x_1) + b, f(x_2) + b, \dots, f(x_n) + b) + \text{pen}(\|f\|_{\mathcal{H}}^2) \\ &= \text{Loss}((\mathbf{P}f)(x_1) + b, (\mathbf{P}f)(x_2) + b, \dots, (\mathbf{P}f)(x_n) + b) + \text{pen}(\|f\|_{\mathcal{H}}^2) \end{aligned}$$

を得る。 $\mathcal{V}^\perp := \{f \in \mathcal{H}; \langle f, g \rangle = 0 (\forall g \in \mathcal{V})\}$ とする。このとき、 $\forall f \in \mathcal{H}$ に対し、

$$f = \mathbf{P}f + f^\perp \quad (f^\perp \in \mathcal{V}^\perp)$$

と一意的に分解できるので、

$$\|\mathbf{P}f\|_{\mathcal{H}}^2 \leq \|f\|_{\mathcal{H}}^2 \implies \text{pen}(\|\mathbf{P}f\|_{\mathcal{H}}^2) \leq \text{pen}(\|f\|_{\mathcal{H}}^2).$$

よって、

$$\begin{aligned} & \text{Loss}(f(x_1) + b, f(x_2) + b, \dots, f(x_n) + b) + \text{pen}(\|f\|_{\mathcal{H}}^2) + \text{pen}(\|f\|_{\mathcal{H}}^2) \\ & \geq \text{Loss}((\mathbf{P}f)(x_1) + b, (\mathbf{P}f)(x_2) + b, \dots, (\mathbf{P}f)(x_n) + b) \\ & \quad + \text{pen}(\|f\|_{\mathcal{H}}^2) + \text{pen}(\|\mathbf{P}f\|_{\mathcal{H}}^2) \end{aligned}$$

が成り立つので、

$$f = \mathbf{P}f = \sum_{j=1}^n \xi_j k_{x_j} \quad (\xi_j \in \mathbb{R}, j = 1, 2, \dots, n)$$

としてよい。 □

5 例

5.1 連立方程式

連立方程式

$$\begin{cases} x_1 + x_3 = 0 \\ x_2 = 1 \end{cases} \quad (6.11)$$

eq:6-10

を考える. 行列とベクトルで表現すると

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \iff \mathbf{R}^\top \mathbf{x} = \boldsymbol{\eta},$$

$$\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2), \mathbf{r}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \mathbf{r}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \boldsymbol{\eta} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$\mathbf{x} = \boldsymbol{\eta}$ と
いいたい.
である.
すると

$$\mathbf{x}_* = a \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} =: a \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} + \mathbf{x}_0 \quad (a \in \mathbb{R})$$

は方程式 ^{eq:6-10}6.11 の解となる.

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \in \text{span}(\mathbf{R}) = \{a\mathbf{r}_1 + b\mathbf{r}_2; a, b \in \mathbb{R}\},$$
$$\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \in \text{span}(\mathbf{R})^\perp := \{\mathbf{x} \in \mathbb{R}^3; \langle \mathbf{x}, \mathbf{y} \rangle = 0 (\forall \mathbf{y} \in \text{span}(\mathbf{R}))\}$$

であることに注意せよ. ベクトル $\mathbf{x}_0 = (0, 1, 0)^\top$ は

$$\|\mathbf{x}\|^2 := \langle \mathbf{x}, \mathbf{x} \rangle = x_1^2 + x_2^2 + x_3^2$$

を最小にする解である。すなわち、 $(0, 1, 0)^\top$ はノルムが最小となる解 (唯一存在) で、任意の解 x_* に対して

$$\mathbf{x}_0^\top \mathbf{x}_0 \leq \mathbf{x}_*^\top \mathbf{x}_*$$

をみたしている。

以上の考察より一般的に考える。 $d, n \in \mathbb{N}, n \geq d$ とし、 \mathbf{R} を $n \times d$ 実行列、 $\boldsymbol{\eta}$ を $d \times 1$ の縦ベクトルで、方程式

$$\mathbf{R}^\top \mathbf{x} = \boldsymbol{\eta} \tag{6.12}$$

eq:6-11

を解くことを考える。

$$\boldsymbol{\eta} \in \text{span}(\mathbf{R}^\top) \in \mathbb{R}^d$$

ならば、解 x_* は存在する。ただし、 $\text{span}(\mathbf{R}^\top)$ は \mathbf{R}^\top の列ベクトルで張られた \mathbb{R}^d の部分空間である。1 つの解 x_* が与えられたとき、他の解を x と書けば

$$\mathbf{R}^\top \mathbf{x}_* = \mathbf{R}^\top \mathbf{x} \iff \mathbf{R}^\top (\mathbf{x}_* - \mathbf{x}) = \mathbf{0} \iff \mathbf{x}_* - \mathbf{x} \in \text{span}(\mathbf{R})^\perp$$

となる。 $\mathbf{x}_1 := \mathbf{x}_* - \mathbf{x} \in \text{span}(\mathbf{R})^\perp$ とおけば、

$$\mathbb{R}^n = \text{span}(\mathbf{R}) \otimes \text{span}(\mathbf{R})^\perp$$

なので、解 x_* は

$$\mathbf{x}_* = \mathbf{x}_0 + \mathbf{x}_1 \quad (\mathbf{x}_0 \in \text{span}(\mathbf{R}), \mathbf{x}_1 \in \text{span}(\mathbf{R})^\perp)$$

と一意に書ける。さらに、

$$\mathbf{x}_0^\top \mathbf{x}_0 \leq \mathbf{x}_0^\top \mathbf{x}_0 + \mathbf{x}_1^\top \mathbf{x}_1$$

なので、 \mathbf{x}_0 は最ノルム小解である。よって、最小ノルム解は

$$\mathbf{x}_0 = \mathbf{R}\boldsymbol{\xi} \quad (\boldsymbol{\xi} \in \mathbb{R}^d)$$

と書ける。 $\mathbf{x}_0 = \mathbf{R}\boldsymbol{\xi}$ を方程式 eq:6-11(6.12) に代入すると

$$\mathbf{R}^\top \mathbf{R}\boldsymbol{\xi} = \boldsymbol{\eta}$$

となる. ξ とは異なるベクトル $\tilde{\xi} \in \mathbb{R}^d$ が

$$R^\top R\tilde{\xi} = \eta$$

をみたすとすると

$$R\xi = R\tilde{\xi}$$

となる. なぜならば, $R^\top x = \eta$ をみたし, $x \in \text{span}(R)$ となるものは唯一だからである.

例にもどれば

$$R^\top R = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

より

$$R^\top R\xi = \eta \iff \begin{cases} 2\xi_1 = 0 \\ \xi_2 = 1 \end{cases}$$

より, $(\xi_1, \xi_2) = (0, 1)$ となる. よって,

$$x_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

となる.

以上をまとめると,

$$r_k \in \mathbb{R}^d (k = 1, 2, \dots, d), \quad \eta = (\eta_1, \eta_2, \dots, \eta_n)^\top \in \mathbb{R}^n$$

とする. 方程式

$$\langle r_k, x \rangle = \eta_k \quad (k = 1, 2, \dots, n)$$

をみたす, $x \in \mathbb{R}^d$ でノルム

$$\|x\| = \sqrt{x^\top x}$$

を最小とするものを見つける問題の解は

$$x_s = \sum_{k=1}^n \xi_k r_k$$

で $(\xi_1, \xi_2, \dots, \xi_n)$ は

$$\sum_{k=1}^n \langle \mathbf{r}_i, \mathbf{r}_k \rangle \xi_k = \eta_i \quad (i = 1, 2, \dots, n)$$

をみたすものである。

この点は関数解析の観点からまとめなおすこと。参考文献は、堤正義『逆問題』(2012)の6章。小川英光『工学系の関数解析』(2010)の7章。Saitoh, S. and Sawano, Y (2016)の3章。

5.2 補間スプライン

点 (t_j, y_j) , $j = 0, \dots, n$, $y_0 = 0$, $0 = t_0 < t_1 < \dots < t_n = 1$ が与えられたとする。関数族

$$W_0^1 := \left\{ f \in L^2[0, 1]; f(0) = 0, \int_0^1 \{\dot{f}(t)\}^2 dt < \infty \right\}, \quad \dot{f}(t) = \frac{df}{dt}$$

の中で,

$$f(t_j) = y_j \quad (j = 1, 2, \dots, n)$$

をみたしたものの中で

$$\int_0^1 \{\dot{f}(t)\}^2 dt$$

を最小とするものを見つけたい。

W_0^1 の内積として, $f, g \in W_0^1$ に対し

$$\langle f, g \rangle := \int_0^1 \dot{f}(t)\dot{g}(t) dt, \quad \|f\| = \sqrt{\langle f, f \rangle}$$

とおく。 $t \in [0, 1]$ に対し

$$k_j(t) = \min\{t, t_j\} \quad (j = 1, 2, \dots, n)$$

とおく。このとき, $k_j(t) \in W_0^2$ であり,

$$\langle f, k_j \rangle = \int_0^1 \dot{f}(s)k_j(s) ds = \int_0^{t_j} \dot{f}(s) ds = f(t_j) - f(0) = f(t_j).$$

したがって、 $y_j = f(t_j)$ をみたま関数は

$$f(t_j) = \langle k_j, f \rangle = y_j \quad (j = 1, 2, \dots, n) \quad (6.13)$$

eq:6-12

条件 (6.13) をみたま最小ノルムの関数は、リプレゼンタ定理から

$$\hat{f}(s) = \sum_{j=1}^n \xi_j k_j(s), \quad s \in [0, 1], \xi_1, \xi_2, \dots, \xi_n \in \mathbb{R} \quad (6.14)$$

eq:6-13

の形でかける。(6.14) を (6.13) に代入すれば

$$\langle k_j, \sum_{\ell=1}^n \xi_\ell k_\ell \rangle = \sum_{\ell=1}^n \langle k_j, k_\ell \rangle \xi_\ell = y_j \quad (j = 1, 2, \dots, n)$$

となる。

$$\mathbf{K} := \begin{pmatrix} \langle k_1, k_1 \rangle & \langle k_1, k_2 \rangle & \cdots & \langle k_1, k_n \rangle \\ \langle k_2, k_1 \rangle & \langle k_2, k_2 \rangle & \cdots & \langle k_2, k_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle k_n, k_1 \rangle & \langle k_n, k_2 \rangle & \cdots & \langle k_n, k_n \rangle \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}$$

とおけば、

$$\mathbf{K}\boldsymbol{\xi} = \mathbf{y}$$

となる。したがって、 $0 < t_1 < t_2 < \cdots < t_n = 1$ なので、 \mathbf{K} は正定値になる。実際

$$\text{Det}(\mathbf{K}) = t_1 \prod_{j=2}^n (t_j - t_2)$$

と書けること²からわかる。さらに、 \mathbf{K}_i を 1 行目から j 行目に対する小行列とすると、すべての $i = 1, 2, \dots, n$ に対し $\text{Det} \mathbf{K}_i \gtrsim 0$ なので、 $\mathbf{K} \gtrsim 0$ がわかる。

たとえば、 $n = 5$ とし、

$$f(0) = 0, f(0.1) = 0.1, f(0.25) = 1, f(0.5) = 2, f(0.75) = 1.5, f(1) = 1.75$$

² $g(t_2, t_3, \dots, t_n) = \text{Det}(\mathbf{K})$ とおく。 $t_j = t_1$ ($j = 2, 3, \dots, n$) とおけば、1 行目と j 行目が等しくなるので、行列式の性質より $g(t_2, t_3, \dots, t_n) = 0$ 。よって、 g は $(t_j - t_1)$ ($j = 2, \dots, n$) という因子をもつことからわかる。

とすれば,

$$\begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.25 & 0.25 & 0.25 & 0.25 \\ 0.1 & 0.25 & 0.5 & 0.5 & 0.5 \\ 0.1 & 0.25 & 0.5 & 0.75 & 0.75 \\ 0.1 & 0.25 & 0.5 & 0.75 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ x_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 1 \\ 2 \\ 1.5 \\ 1.75 \end{pmatrix}$$

の解は

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{pmatrix} = \begin{pmatrix} -5 \\ -2 \\ 6 \\ -3 \\ 1 \end{pmatrix}$$

より

$$\hat{f}(t) = -5k_1(t) + 2k_2(t) + 6k_3(t) - 3k_4(t) + k_5(t) = \begin{cases} t & (0 \leq t \leq 0.1) \\ 6t - 0.5 & (0.1 \leq t \leq 0.25) \\ 4t & (0.25 \leq t \leq 0.5) \\ -2t + 3 & (0.5 \leq t \leq 0.75) \\ t + 0.75 & (0.75 \leq t \leq 1) \end{cases}$$

となる.

第7章 2 値判別分析と回帰

Christensen の資料を借用. 福水 (2010) も参照のこと.

1 サポートベクターマシン (SMV)

1.1 SMV の考え方

$j = 1, 2, \dots, n$ ($n \in \mathbb{N}$) とし $\mathbf{x}_j \in \mathbb{R}^d$, $y_j \in \{-1, 1\}$ を教師データとする. y_j の値によりデータ j を分類する. x_j は説明変数で, y_j は応答変数となる. ここで $y_j = 1$ と $y_j = -1$ に対応する説明変数の 2 つの集合

$$S_+ := \{j \in \{1, 2, \dots, n\}; y_j = 1\} \text{ と } S_- := \{j \in \{1, 2, \dots, n\}; y_j = -1\}$$

は \mathbb{R}^d の超平面で分離可能とする. すなわち, あるベクトル $\mathbf{w} \in \mathbb{R}^d$ と実数 $w_0 \in \mathbb{R}$ を上手く取り

$$y(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + w_0, \quad \mathbf{w} = (w_1, w_2, \dots, w_d)^\top \quad (7.1)$$

とおいたとき

$$y(\mathbf{x}_j) > 0 \implies j \in S_+, \quad y(\mathbf{x}_j) < 0 \implies j \in S_-$$

であるとする. したがって超平面 (7.1) は分類器となる. このような超平面は複数あるとき, 超平面 (7.1) に最も近いデータ \mathbf{x}_j との距離を最小にするような超平面を選択することを考える. 選ばれた超平面との距離が最小であるデータをサポートベクターという.

データ \mathbf{x}_j ($j = 1, 2, \dots, n$) と超平面 (7.1) との距離を求めよう. \mathbf{x}_j から (7.1) 下した垂線と超平面との交点を $\tilde{\mathbf{x}}$ とおく. 定義より $y(\tilde{\mathbf{x}}) = 0$ であることに注意せよ. 超平面 (7.1) と直交する単位ベクトルは

$$\frac{\mathbf{w}}{\|\mathbf{w}\|_2}$$

この節
志 (2010)
を参考に

eq:svm-

である. このことよりある実数 r_j があって

$$\mathbf{x}_j - \mathbf{x}_0 = r_j \frac{\mathbf{w}}{|\mathbf{w}|_2}$$

と書ける. すると

$$\begin{aligned} y(\mathbf{x}_j) &= \langle \mathbf{w}, \mathbf{x}_j \rangle + w_0 = \left\langle \mathbf{w}, \tilde{\mathbf{x}} + r_j \frac{\mathbf{w}}{|\mathbf{w}|_2} \right\rangle + w_0 = \langle \mathbf{w}, \tilde{\mathbf{x}} \rangle + w_0 + r_j |\mathbf{w}|_2 \\ &= y(\tilde{\mathbf{x}}) + r_j |\mathbf{w}|_2 = r_j |\mathbf{w}|_2 \end{aligned}$$

となることから

$$r_j = \frac{y(\mathbf{x}_j)}{|\mathbf{w}|_2}$$

がわかる. よって

$$|r_j| = \frac{y_j y(\mathbf{x}_j)}{|\mathbf{w}|_2}$$

を得る.

いま

$$\min_{j \in \{1, 2, \dots, n\}} |y(\mathbf{x}_j)| = \frac{1}{c}$$

とおく. データは分離可能なので c は有界である. ここで

$$z(\mathbf{x}) = \langle c\mathbf{w}, \mathbf{x} \rangle + cw_0$$

とおくと $y(\mathbf{x})$ と $z(\mathbf{x})$ は同じ直線を表し,

$$\begin{aligned} \min_{j \in \{1, 2, \dots, n\}} r_j &= \min_{j \in \{1, 2, \dots, n\}} \frac{y_j y(\mathbf{x}_j)}{|\mathbf{w}|_2} = \min_{j \in \{1, 2, \dots, n\}} \frac{y_j z(\mathbf{x}_j)}{|c\mathbf{w}|_2} \\ &= \frac{1}{c} \frac{1}{|\mathbf{w}|_2} \end{aligned}$$

となる. したがって $|r_j|$ の最小化問題は

$$\begin{aligned} &\text{minimize}_{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}} \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle \\ &\text{subject to } 1 - y_j (\langle \mathbf{w}, \mathbf{x}_j \rangle + w_0) \leq 0 \quad (j = 1, 2, \dots, n) \end{aligned} \tag{7.2}$$

eq:svm-

となる. 最適解を見つけるために Lagrange 関数と双対関数を

$$L(\mathbf{w}, w_0, \boldsymbol{\mu}) = \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle + \sum_{j=1}^n \mu_j \{1 - y_j (\langle \mathbf{w}, \mathbf{x}_j \rangle + w_0)\} \quad (7.3)$$

$$L^*(\boldsymbol{\mu}) = \max_{\mathbf{w} \in \mathbb{R}^d, w_0 \in \mathbb{R}} L(\mathbf{w}, w_0, \boldsymbol{\mu}) \quad (7.4)$$

とおく. ただし $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^\top$ である. KKT 条件 (7.2) から

$$\frac{\partial L(\mathbf{w}, w_0, \boldsymbol{\mu})}{\partial w_j} = 0 \quad (j = 0, 1, \dots, n)$$

から

$$\mathbf{w} = \sum_{j=1}^n \mu_j y_j \mathbf{x}_j, \quad \sum_{j=1}^n \mu_j y_j = 0 \quad (7.5)$$

を得る. 最適化問題 (7.2) の双対問題は

$$\text{maximize}_{\boldsymbol{\mu} \in \mathbb{R}^n} L^*(\boldsymbol{\mu}) \quad \text{subject to } \boldsymbol{\mu} \geq \mathbf{0} \quad (7.6)$$

となる. ここで (7.5) を (7.3) に代入すると

$$\begin{aligned} L^*(\boldsymbol{\mu}) &= \frac{1}{2} \left\langle \sum_{j=1}^n \mu_j y_j \mathbf{x}_j, \sum_{\ell=1}^n \mu_\ell y_\ell \mathbf{x}_\ell \right\rangle + \sum_{j=1}^n \mu_j \left(1 - y_j \left\langle \sum_{\ell=1}^n \mu_\ell y_\ell \mathbf{x}_\ell, \mathbf{x}_j \right\rangle - y_j w_0 \right) \\ &= \frac{1}{2} \sum_{j=1}^n \sum_{\ell=1}^n y_j \mu_j y_\ell \mu_\ell \langle \mathbf{x}_j, \mathbf{x}_\ell \rangle + \sum_{j=1}^n \mu_j - \sum_{j=1}^n \sum_{\ell=1}^n y_j \mu_j y_\ell \mu_\ell \langle \mathbf{x}_j, \mathbf{x}_\ell \rangle \\ &\quad - w_0 \sum_{j=1}^n y_j \mu_j \\ &= -\frac{1}{2} \sum_{j=1}^n \sum_{\ell=1}^n y_j \mu_j y_\ell \mu_\ell \langle \mathbf{x}_j, \mathbf{x}_\ell \rangle + \sum_{j=1}^n \mu_j \end{aligned}$$

となる. 最後の等号は (7.5) を用いた. したがって双対問題 (7.6) は 2 次最適問題となるので, 解くことができる. さらに定理 7.53 より強双対性が成立するので $\boldsymbol{\mu}^* = (\mu_1^*, \mu_2^*, \dots, \mu_n^*)^\top$ を双対問題 (7.6) の解とし, $\mathbf{w}^* = (w_1^*, w_2^*, \dots, w_d^*)^\top$, w_0^* を最適化問題 (7.2) の解としたとき

$$L^*(\boldsymbol{\mu}^*) = \frac{1}{2} \langle \mathbf{w}^*, \mathbf{w}^* \rangle$$

が成立している. 以上のことより

$$\mathbf{x}^* = \sum_{j=1}^n \mu_j^* y_j \mathbf{x}_j$$

を得る. さらに条件 leg: con-55 (E.57) から

$$\mu_j^* \neq 0 \iff y_j (\langle \mathbf{x}_j, \mathbf{w}^* \rangle + w_0^*) = 1$$

が成立している. ここで $J = \{j \in \{1, 2, \dots, n\}; \mu_j^* \neq 0\}$ とおく. 上の式の両辺に y_j を掛けると $j \in J$ に対して

$$\langle \mathbf{x}_j, \mathbf{w}^* \rangle + w_0^* = y_j$$

となるので,

$$w_0^* = \frac{1}{\#(J)} \sum_{j \in J} y_j \{ \langle \mathbf{w}^*, \mathbf{x}_j \rangle - y_j \}$$

を得る. ただし $\#(J)$ は集合 J の元の個数を表す.

第 A 章 基礎的な確率規則

この節では、統計的推測理論を理解する上で必要な確率論の基礎事項を簡潔に述べる。したがって、証明はしない場合がある。測度論の内容には立ち入らない。数学的な厳密性を犠牲にして、直観的な理解を目指す記述を心掛けた。

1 基礎的な確率規則

確率論はランダムな現象を扱う数学理論である。確率論で扱う行為を試行という。試行のありうる結果を集めた集合を標本空間といい、 Ω と記すことにする。 Ω の部分集合¹を事象という。事象には、標本空間 Ω と空事象 \emptyset (何も起こらないという事象) も含める。事象をすべて集めた集合族を \mathcal{F} と記す。 \mathcal{F} は以下で述べる σ 加法性 (完全加法性) をみたすことにする。

df:1-1

定義 A.1. Ω を空でない集合とし、 \mathcal{F} を Ω の部分集合族とする。 \mathcal{F} が次の 3 条件をみたすとき、それを σ 加法族と呼ぶ。

- (1) $\Omega \in \mathcal{F}$.
- (2) $A \in \mathcal{F} \implies A^c \in \mathcal{F}$.
- (3) $A_n \in \mathcal{F} (n = 1, 2, \dots) \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

ただし $A^c = \{\omega; \omega \notin A\}$ である。 Ω と \mathcal{F} の組 (Ω, \mathcal{F}) を可測空間と呼ぶ

lem:1-2

補題 A.2. (Ω, \mathcal{F}) を可測空間とする。このとき以下が成立する。

- (1) $\emptyset \in \mathcal{F}$.
- (2) $A_n \in \mathcal{F} (n = 1, 2, \dots) \implies \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$.

¹ Ω が可算集合ならば、事象はすべての部分集合としてよいが、 Ω が連続濃度のときには、すべての部分集合を対象にすることはしない。

Proof. (1) $\emptyset = \Omega^c \in \mathcal{F}$.

(2) $A_n^c \in \mathcal{F}$ と De Morgan の法則から

$$\bigcap_{n=1}^{\infty} A_n = \bigcap_{n=1}^{\infty} (A_n^c)^c = \left(\bigcup_{n=1}^{\infty} A_n^c \right)^c \in \mathcal{F}.$$

最後は定義 $\overset{\text{df:1-1}}{\text{A.1(3)}}$ を用いた。 □

re:1-3

注意 A.3. \mathcal{C} を Ω の部分集合族とする。 \mathcal{C} は σ 加法性をみたしてなくともよい。このとき

$$\sigma[\mathcal{C}] := \bigcap \{ \mathcal{A}; \mathcal{A} \supset \mathcal{C}, \mathcal{A} \text{ は } \sigma \text{ 加法族} \}$$

と定める。すると $\sigma[\mathcal{C}]$ は σ 加法族となる。さらに \mathcal{G} を \mathcal{A} を含む σ 加法族としたとき

$$\sigma[\mathcal{C}] \subseteq \mathcal{G}$$

となる。すなわち $\sigma[\mathcal{C}]$ は \mathcal{C} を含む最小 (包含関係の意味) の σ 加法族となる。 □

df:1-4

定義 A.4. $\Omega = \mathbb{R}$ とし

$$\mathcal{O} = \{ O \subset \mathbb{R}; O \text{ は } \mathbb{R} \text{ の開集合} \}$$

とする。 $\sigma[\mathcal{O}]$ を \mathbb{R} の Borel 集合族と呼び、 $\mathcal{B}(\mathbb{R})$ と記す。また

$$\mathcal{C} = \{ (-\infty, x) \subset \mathbb{R}; x \in \mathbb{R} \}$$

とする。このとき \mathcal{C} は \mathcal{O} の真部分集合であるが $\sigma[\mathcal{C}] = \sigma[\mathcal{O}]$ となる²。

注意 A.5. Ω が高々可算集合のときは、 $\mathcal{F} = 2^\Omega$ と取る。ただし 2^Ω は Ω のすべての部分集合からなる集合族で ^{べき} 冪集合という。 □

df:1-6

定義 A.6. (Ω, \mathcal{F}) を可測空間とする。 \mathcal{F} 上の関数

$$\text{Pr} : \mathcal{F} \ni A \mapsto \text{Pr}(A) \in [0, 1]$$

が次の 2 条件をみたすとき、 (Ω, \mathcal{F}) 上の確率測度³という。

² $\sigma[\mathcal{C}] \subset \sigma[\mathcal{O}]$ は明らかであるが、逆の包含関係も示すことができる。

³ 簡単に Ω 上の確率測度ともいう。

- (1) $\Pr(\Omega) = 1$.
 (2) 互いに排反⁴な事象列 $A_n \in \mathcal{F} (n = 1, 2, \dots)$ に対して

$$\Pr\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \Pr(A_n) := \lim_{N \rightarrow \infty} \sum_{n=1}^N \Pr(A_n).$$

これらの 3 つの組 $(\Omega, \mathcal{F}, \Pr)$ を確率空間という.

lem:1-7

補題 A.7. $(\Omega, \mathcal{F}, \Pr)$ を確率空間とする. このとき以下が成立する.

- (1) $\Pr(\emptyset) = 0$.
 (2) $A \in \mathcal{F}$ に対して, $\Pr(A^c) = 1 - \Pr(A)$.
 (3) $N \in \mathbb{N}$ とする. A_1, A_2, \dots, A_N が互いに排反ならば

$$\Pr\left(\bigcup_{n=1}^N A_n\right) = \sum_{n=1}^N \Pr(A_n).$$

- (4) $A, B \in \mathcal{F}, A \subset B \implies \Pr(B \setminus A) = \Pr(B) - \Pr(A)$.
 (5) (Boole の定理/ユニオン・バウンド) $A_n \in \mathcal{F} (n = 1, 2, \dots)$ に対して

$$\Pr\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \Pr(A_n).$$

- (6) $A_n \in \mathcal{F}$ が $A_n \subset A_{n+1} (n = 1, 2, \dots)$ をみたすならば

$$\lim_{n \rightarrow \infty} \Pr(A_n) = \Pr\left(\bigcup_{n=1}^{\infty} A_n\right).$$

- (7) $A_n \in \mathcal{F}$ が $A_n \supset A_{n+1} (n = 1, 2, \dots)$ をみたすならば

$$\lim_{n \rightarrow \infty} \Pr(A_n) = \Pr\left(\bigcap_{n=1}^{\infty} A_n\right).$$

Proof. 久保川 (2017, p.3 および p.7) を参照. □

注意 A.8. 補題 [A.7\(6\)\(7\)](#) ^{lem:1-7}において, $\{\Pr(A_n)\}_{n=1}^{\infty}$ は有界な非減少列もしくは非増加列なので, 左辺は必ず収束することに注意せよ.

⁴ $m \neq n$ ならば, $A_m \cap A_n = \emptyset$ が成立していること.

df:1-9

定義 A.9. $(\Omega, \mathcal{F}, \Pr)$ を確率空間とし, $A, B \in \mathcal{F}$ とする.

(1) (独立性):

$$A \text{ と } B \text{ は独立} \iff \Pr(A \cap B) = \Pr(A)\Pr(B).$$

(2) (条件付き確率) $\Pr(B) > 0$ のとき, B を与えたときの A の条件付き確率を

$$\Pr(A|B) := \frac{\Pr(A \cap B)}{\Pr(B)}$$

で定める.

re:1-10

注意 A.10. (1) $A, B \in \mathcal{F}$ とし, $\Pr(B) > 0$ とする. A と B が独立のとき

$$\Pr(A|B) = \Pr(A)$$

が成立する.

(2) $\Pr(B) > 0$ のとき

$$\Pr(A \cap B) = \Pr(B)\Pr(A|B)$$

となる. これを乗法の公式という.

(3) $\Pr(B) > 0$ のとき, 関数 $\Pr(\cdot|B) : \mathcal{F} \rightarrow [0, 1]$ は定義 [df:1-4](#) [A.4\(I\)](#) – (3) をみたす. すなわち (Ω, \mathcal{F}) 上の確率測度である. □

問 A.1. 注意 [re:1-10](#) [A.10\(I\)](#)(3) を証明せよ.

em:1-11

補題 A.11. (全確率の法則) $N \in \mathbb{N}$ とする. $A_1, A_2, \dots, A_N \in \mathcal{F}$ は Ω の分割とする. すなわち, すべての $m \neq n$ に対し, $A_m \cap A_n = \emptyset$ で, $\bigcup_{n=1}^N A_n = \Omega$ が成立している. このとき任意の $B \in \mathcal{F}$ に対して

$$\Pr(B) = \Pr\left(\bigcup_{n=1}^N (A_n \cap B)\right) = \sum_{n=1}^N \Pr(A_n \cap B).$$

Proof. $B = B \cap \Omega = B \cap \left(\bigcup_{n=1}^N A_n\right) = \bigcup_{n=1}^N (A_n \cap B)$ と $(A_1 \cap B), (A_2 \cap B), \dots, (A_N \cap B)$ は互いに排反であることに注意して, 定義 [df:1-6](#) [A.6\(2\)](#) を用いればよい. □

定理 A.12. (Bayes の定理) (1) $N \in \mathbb{N}$ とする. $A_1, A_2, \dots, A_N, B \in \mathcal{F}$ とし, A_1, A_2, \dots, A_N は Ω の分割とする. $j = 1, 2, \dots, N$ に対して, $\Pr(B) > 0, \Pr(A_j) > 0$ のとき

$$\Pr(A_j|B) = \frac{\Pr(A_j)\Pr(B|A_j)}{\sum_{n=1}^N \Pr(A_n)\Pr(B|A_n)}$$

が成立する.

(2) $A, B \in \mathcal{F}$ とし, $\Pr(A) > 0, \Pr(B) > 0$ とする. このとき

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$$

が成立する.

Proof. 条件付き確率の定義, 補題 [A.11](#), 定義 [A.4\(2\)](#) を用いればよい. □

例 A.13. 病気 D に対する検査の結果を $+$ と $-$ とし, 確率が以下であるとする.

$$\Pr(+|D) = 0.9, \quad \Pr(-|D^c) = 0.9, \quad \Pr(D) = 0.01.$$

Bayes の定理を用いて, 検査で $+$ に人が本当に D である確率を求めると

$$\begin{aligned} \Pr(D|+) &= \frac{\Pr(+ \cap D)}{\Pr(+)} = \frac{\Pr(D)\Pr(+|D)}{\Pr(D)\Pr(+|D) + \Pr(D^c)\Pr(+|D^c)} \\ &= \frac{0.01 \times 0.9}{0.01 \times 0.9 + (1 - 0.01) \times \{1 - 0.9\}} \approx 0.83. \end{aligned}$$

2 確率変数

前節では, 確率と事象を記述する数学的なモデルを導入した. しかし, 現実の現象を扱い統計学の対象は, 事象には直接結びつかないかもしれない数量的な情報である. 以下で定義する確率変数は, 事象と数量の間の橋渡しをする.

定義 A.14. 関数

$$X: \Omega \longrightarrow \mathbb{R} \cup \{\infty\}$$

が $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数であるとは, すべての $c \in \mathbb{R}$ に対して

$$X^{-1}((-\infty, c]) := \{\omega \in \Omega; X(\omega) \leq c\} \in \mathcal{F}$$

をみたすときをいう.

re:1-15 注意 A.15. 上の定義の条件は

$$X^{-1}(B) \in \mathcal{F} \quad (\forall B \in \mathcal{B}(\mathbb{R}))$$

と同値であることを示すことができる. □

df:1-16 定義 A.16. (1) 確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数 X に対して

$$F_X(x) := \Pr(X \leq x) := \Pr(\{\omega \in \Omega; X(\omega) \leq x\}) \quad (x \in \mathbb{R})$$

を X の累積分布関数 (cumulative distribution function(CDF)) という. また

$$P_X(B) := \Pr(X \in B) := \Pr(\{\omega \in \Omega; X(\omega) \in B\}) \quad (B \in \mathcal{B}(\mathbb{R}))$$

を X の分布という. $P_X((-\infty, x]) = F_X(x)$ である.

(2) X, Y を $(\Omega, \mathcal{F}, \Pr)$ の確率変数とし, それぞれの CDF を F_X, F_Y とする. このとき

$$F_X(x) = F_Y(x) \quad (\forall x \in \mathbb{R})$$

が成立するとき, X と Y の分布は同じであるという. これを $X \stackrel{d}{=} Y$ と書く.

(3) X が CDF F を持つとき, $X \sim F$ と書く.

re:1-17 注意 A.17. (1) P_X は可測空間 $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ 上の確率測度である. すなわち P_X は定義 [A.4](#) ^{df:1-4} をみたすことがわかる.

(2) CDF F が与えられると

$$F(x) = P((-\infty, x]) \quad (x \in \mathbb{R})$$

をみたす $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ 上の確率測度 P が一意的に定まることが知られている. このことにより X の CDF と分布を同一視する. さらに $X \sim F$ とし, CDF F から定まる確率測度を P としたとき, $X \sim P$ と書く.

(3) X と Y の分布が同じとき

$$P_X(B) = P_Y(B) \quad (\forall B \in \mathcal{B}(\mathbb{R}))$$

も成立する.

問 A.2. 注意 [A.17\(I\)](#) を示せ. 逆像の性質を利用すること.

hm: 1-18 定理 A.18. X を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数とし, F を X の CDF とする. このとき F は次をみたす.

(1) F は非減少関数: $x < y \implies F(x) \leq F(y)$.

(2) $\lim_{x \rightarrow \infty} F(x) = 1$: $\lim_{x \rightarrow -\infty} F(x) = 0$.

(3) F は右連続関数: $\lim_{y \rightarrow x+0} F(y) = F(x)$.

Proof. X の分布を P とする. (1) $(-\infty, x] \subset (-\infty, y]$ に注意して, [補題 A.7\(4\)](#) を P に適用すればよい.

(2) $A_n = (-\infty, n]$ として, P に [補題 A.7\(6\)\(7\)](#) を適用すればよい.

(3) $A_n = (-\infty, x + \frac{1}{n}]$ として, P に [補題 A.7\(7\)](#) を適用すればよい. \square

hm: 1-19 定理 A.19. X を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数とし, X の CDF を F とする. このとき次の (1) – (3) は同値である.

(1) F は \mathbb{R} 上の連続関数.

(2) $F(x) = F(x-) (\forall x \in \mathbb{R})$. ただし, $F(x-) := \sup\{F(y); y < x\}$ とした.

(3) $\Pr(X = x) = 0 (\forall x \in \mathbb{R})$.

Proof. $x_n = x - \frac{1}{n} (n = 1, 2, \dots)$ とすれば

$$\lim_{n \rightarrow \infty} F(x_n) = F(x-)$$

である. さらに

$$\{X \leq x_n\} \subset \{X \leq x_{n+1}\} (n = 1, 2, \dots), \quad \bigcup_{n=1}^{\infty} \{X \leq x_n\} = \{X < x\}$$

であるので, [補題 A.7\(6\)](#) より

$$F(x-) = \lim_{n \rightarrow \infty} F(x_n) = \lim_{n \rightarrow \infty} \Pr(X \leq x_n) = \Pr\left(\bigcup_{n=1}^{\infty} \{X \leq x_n\}\right) = \Pr(X < x)$$

を得る. さらに補題 ^{lem:1-7} A.7(4) より

$$\begin{aligned}\Pr(X = x) &= \Pr(\{X \leq x\} \setminus \{X < x\}) = \Pr(X \leq x) - \Pr(X < x) \\ &= F(x) - F(x-)\end{aligned}$$

であることからわかる. □

注意 A.20. F が \mathbb{R} 上で連続のとき, $\Pr(X = x) = 0 (\forall x \in \mathbb{R})$ なので, $a < b$ に対して

$$\Pr(a \leq X \leq b) = \Pr(a < X < b) = \Pr(a < X \leq b) = \Pr(a \leq X < b)$$

である. □

定義 A.21. X を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数とし, その CDF と分布をそれぞれ F と P と書く.

(1) X が高々可算個の集合 $\{x_1, x_2, \dots\}$ 上にしか値を取らないとき, X を離散型であるという. この場合には

$$p(x) := \Pr(X = x) = F(x) - F(x-) \quad (x \in \{x_1, x_2, \dots\})$$

で X の分布が特徴付けられる⁵. p を X の確率関数 (probability mass function(PMF)) と呼ぶ.

(2) $\Pr(X = x) = 0 (\forall x \in \mathbb{R})$ のとき, X を連続型であるという. さらにある非負値関数 f で

$$F(x) = \int_{-\infty}^x f(t) dt \quad (\forall x \in \mathbb{R})$$

をみたすものが存在するとき, f を X の確率密度関数 (probability density function(PDF)) という. 特に F がほとんどいたるところ⁶で微分可能ならばほとんどいたるところで

$$\dot{F}(x) = \frac{dF}{dx}(x) = f(x)$$

となる.

⁵ $p(x) = 0 (x \notin \{x_1, x_2, \dots\})$ となるので, p は \mathbb{R} 上の関数であり, $0 \leq p(x) \leq 1$ となる.

⁶この授業では, \mathbb{R} から可算個の点を除いた集合上で微分可能と理解して差し支えない.

例 A.22. (1) $\Omega = \{0, 1\}$, $\mathcal{F} = 2^\Omega$, $\Pr(\{0\}) = \Pr(\{1\}) = 1/2$ とし

$$X : \Omega \ni \omega \mapsto X(\omega) = \omega \in \mathbb{R}$$

と定義すれば

$$\Pr(X = 0) = \Pr(X = 1) = \frac{1}{2}$$

となる. X の CDF F と PMF p はそれぞれ

$$F(x) = \begin{cases} 0 & (x < 0) \\ \frac{1}{2} & (0 \leq x \leq 1) \\ 1 & (x > 1) \end{cases} \quad p(x) = \begin{cases} \frac{1}{2} & (x = 0, 1) \\ 0 & (\text{その他}) \end{cases}$$

となる.

(2) $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}(\mathbb{R}) \cap [0, 1]$ とし

$$P_X((a, b]) = b - a \quad (0 \leq a < b \leq 1)$$

とし

$$X : \Omega \ni \omega \mapsto X(\omega) = \omega \in \mathbb{R}$$

と定義すれば, X の CDF F と PDF f はそれぞれ

$$F(x) = \begin{cases} 0 & (x < 0) \\ x & (0 \leq x \leq 1) \\ 1 & (x > 1) \end{cases} \quad f(x) = \begin{cases} 1 & (0 \leq x \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

となる.

3 分位点関数

定義 A.23.

注意 A.24.

例 A.25.

加藤の講
用するこ

□

定理 A.26.

Proof.

□

系 A.27.

4 主な 1 次元分布

4.1 離散型確率変数

4.1.1 Bernoulli 分布

$0 \leq \theta \leq 1$ とする. $X \sim \text{Ber}(\theta)$ (X は母数 θ の Bernoulli 分布) に従うとは, X の PMF p が

$$p(x|\theta) = p(x) = \theta^x(1-\theta)^{1-x} \quad (x = 0, 1)$$

のときをいう.

4.1.2 2 項分布

$n \in \mathbb{N}, 0 \leq \theta \leq 1$ とする. $X \sim \text{Bino}(n, \theta)$ (X は母数 (n, θ) の 2 項分布) に従うとは, X の PMF p が

$$p(x|n, \theta) = p(x) = \binom{n}{x} \theta^x(1-\theta)^{n-x} \quad (x = 0, 1, \dots, n)$$

のときをいう. ただし

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}, \quad 0! = 1.$$

4.1.3 幾何分布

$0 < \theta < 1$ とする. $X \sim \text{Geom}(\theta)$ (X は母数 θ の幾何分布) に従うとは,

X の PMF p が

$$p(x|\theta) = p(x) = \theta(1-\theta)^{x-1} \quad (x = 1, 2, \dots)$$

のときをいう。

4.1.4 Poisson 分布

$\theta > 0$ とする. $X \sim \text{Pois}(\text{母数 } \theta \text{ の Poisson 分布})$ に従うとは, X の PMF p が

$$p(x|\theta) = p(x) = e^{-\theta} \frac{\theta^x}{x!} \quad (x = 0, 1, 2, \dots)$$

のときをいう。

4.2 連続型確率変数

4.2.1 正規分布

$\nu \in \mathbb{R}, 0 < \sigma < \infty$ とする. $X \sim \text{N}(\mu, \sigma^2)$ (平均 μ , 分散 σ^2 の正規分布) に従うとは, X の PDF f が

$$f(x|\mu, \sigma^2) = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

のときをいう. ただし $\exp(x) = e^x$ である. $\mu = 0, \sigma = 1$ のとき, 標準正規分布という。

4.2.2 ガンマ分布

$\alpha > 0, \beta > 0$ とする. $X \sim \text{Ga}(\alpha, \beta)$ (母数 (α, β) のガンマ分布) に従うとは, X の PDF f が

$$f(x|\alpha, \beta) = f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad (x > 0)$$

のときをいう. ただし

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

である. $\lambda > 0$ とする. $\text{Ga}(1, 1/\lambda)$ を母数 λ の指数分布といい, $\text{Exp}(\lambda)$ と書く. $p \in \mathbb{N}$ とする. $\text{Ga}(p/2, 2)$ を自由度 p の χ^2 分布といい, χ_p^2 と書く。

5 2次元の分布

5.1 同時確率関数と密度関数

確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の離散型確率変数 (X, Y) の対に対して、同時 PMF f を

	Y = 0	Y = 1	
X = 0	1/9	2/9	1/3
X = 1	2/9	4/9	2/3
	1/3	2/3	

で定める。たとえば

$$p(1, 1) = \Pr(X = 1, Y = 1) = \frac{4}{9}$$

である。

定義 A.28. 確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数 X, Y は連続型とする。 \mathbb{R}^2 上の非負値実数値関数 p が確率ベクトル (X, Y) の同時確率密度関数 (同時 PDF) であるとは、次の条件をみたすときをいう。

- (1) $p(x, y) \geq 0$ ($\forall (x, y) \in \mathbb{R}^2$).
- (2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) dx dy = 1$.
- (3) $\forall A \in \mathcal{B}(\mathbb{R}^2)$ に対して

$$\Pr((X, Y) \in A) = \int \int_A p(x, y) dx dy.$$

注意 A.29. (X, Y) の同時累積分布関数 (同時 CDF) F を

$$F(x, y) := \Pr(X \leq x, Y \leq y) := \Pr(\{\omega \in \Omega; X(\omega) \leq x, Y(\omega) \leq y\})$$

$$(\forall (x, y) \in \mathbb{R}^2)$$

で定義する。 (X, Y) の同時分布 P を

$$P(A) = \Pr((X, Y) \in A) \quad (\forall A \in \mathcal{B}(\mathbb{R}^2)) \tag{A.1}$$

で定義する。どの確率変数の CDF と分布であるかを明示したいときには、 $F_{(X, Y)}$, $P_{(X, Y)}$ と記す。細かなことであるが、(A.1) で P を定義すると $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ 上の確率測度を一意的に定めていることになる。 \square

例 A.30. 連続型確率変数のベクトル (X, Y) は同時 PDF

$$p(x, y) = \begin{cases} x + y & (0 \leq x \leq 1, 0 \leq y \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

を持つとする. このとき集合 $\{(x, y); p(x, y) > 0\}$ に注意すれば

$$\begin{aligned} \int_0^1 \int_0^1 p(x, y) dx dy &= \int_0^1 \left[\int_0^1 x dx \right] dy + \int_0^1 \left[\int_0^1 y dy \right] dx \\ &= \int_0^1 \frac{1}{2} dy + \int_0^1 \frac{1}{2} dx = 1. \end{aligned}$$

□

5.2 周辺分布

定義 A.31. (X, Y) を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上で定義された確率ベクトルとする. (1) (X, Y) が離散型で同時 PMF p を持つとする. X の周辺確率関数 (周辺 PMF) を

$$p_X(x) = \Pr(X = x) = \sum_{y \in S_Y} \Pr(X = y, Y = y) = \sum_{y \in S_Y} p(x, y) \quad (\forall x \in S_X)$$

で定義する. ただし

$$S_X := \{x \in \mathbb{R}; \text{ある } y \in \mathbb{R} \text{ に対して, } p(x, y) > 0\}$$

$$S_Y := \{y \in \mathbb{R}; \text{ある } x \in \mathbb{R} \text{ に対して, } p(x, y) > 0\}$$

である.

(2) (X, Y) は連続型とし, 同時 PDF p を持つとする. このとき X の周辺確率密度関数 (周辺 PDF) を

$$p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy \quad (\forall x \in \mathbb{R})$$

で定義し, Y の周辺確率密度関数 (周辺 PDF) を

$$p_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx \quad (\forall y \in \mathbb{R})$$

で定義する.

re:1-32

注意 A.32. 連続型確率ベクトル (X, Y) に対して

$$\begin{aligned}\Pr(X \leq x) &= \int \int_{(s,t) \in \mathbb{R}^2; s \leq x} p(s, t) ds dt = \int_{-\infty}^x \left\{ \int_{-\infty}^{\infty} p(s, t) ds \right\} dt \\ &= \int_{-\infty}^x p_X(s) ds\end{aligned}$$

となるので, X の周辺 PDF と X の PDF は同じである. □

re:1-33

注意 A.33. F を確率ベクトル (X, Y) の同時 CDF とし, F_X を X の周辺 CDF とする. このとき $x \in \mathbb{R}$ に対して

$$F_X(x) = \lim_{y \rightarrow \infty} F(x, y); F_Y(y) = \lim_{x \rightarrow \infty} F(x, y)$$

が成立する. □

lem:1-3

問 A.3. 注意 ^{re:1-33}A.33 の式を証明せよ.

5.3 独立な分布と条件付き分布

df:1-34

定義 A.34. 確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の 2 つの確率変数 X と Y は独立であるとは, $\forall A, B \in \mathcal{B}(\mathbb{R})$ に対して

$$\Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B)$$

が成り立つときをいう. 独立でないとき X と Y は従属であるという.

hm:1-35

定理 A.35. p を確率ベクトル (X, Y) の同時 PDF とし, p_X と p_Y を X と Y それぞれの周辺 PDF とする. このとき

$$X \text{ と } Y \text{ は独立} \iff p(x, y) = p_X(x)p_Y(y) \quad (\forall x, y \in \mathbb{R}).$$

Proof. 独立ならば, 同時 PDF が周辺 PDF の積で表現されることの証明は易しい. 逆はこの講義の範囲 (測度の拡張の議論が必要となる!) を超えるので, 信じることにする. □

re:1-36

注意 A.36. (1) 連続型の場合には、測度 0 の集合を除いて⁷PDF は定義されるので、定義 [A.34](#)^{df:1-34}の書き方はやや数学的な厳密性に欠ける。

(2) 離散型確率変数の場合には、PDF を PMF に置き換えればよい。 □

ex:1-37

例 A.37. 連続型確率変数 X と Y は独立で同じ PDF

$$p(x) = \begin{cases} 2x & (0 \leq x \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

を持つとする。このとき確率 $\Pr(X + Y \leq 1)$ を求めてみよう。独立性より (X, Y) の同時 PDF は

$$p_{(X, Y)}(x, y) = \begin{cases} 4xy & (0 \leq x \leq 1, 0 \leq y \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

となる。したがって

$$\begin{aligned} \Pr(X + Y \leq 1) &= \int \int_{x+y \leq 1} p_{(X, Y)}(x, y) \, dx dy = 4 \int_0^1 x \left\{ \int_0^{1-x} y \, dy \right\} dx \\ &= 4 \int_0^1 x \frac{(1-x)^2}{2} \, dx = \frac{1}{6}. \end{aligned}$$

□

hm:1-37

定理 A.38. 連続型確率ベクトル (X, Y) は同時 PDF $p(x, y)$ を持つとする。ただし $\{(x, y) \in \mathbb{R}^2; p(x, y) > 0\}$ は矩形⁸とする。このときある非負値関数 g と h が存在して

$$p(x, y) = h(x)g(y)$$

と書けるとき、 X と Y は独立である。

Proof. 積分をして確かめればよい。 □

⁷この用語は定義されていない。「有限個の点を除いて」と理解してもこの講義の内容の範囲では問題ない。

⁸有界でなくともよい。

ex: 1-39

例 A.39. 連続型確率ベクトル (X, Y) は同時 PDF

$$p(x, y) = \begin{cases} 2e^{-(x+y)} & (x \geq 0, y \geq 0) \\ 0 & (\text{その他}) \end{cases}$$

を持つとする. $\{(x, y) \in \mathbb{R}^2; f(x, y) > 0\} = [0, \infty)^2$ なので

$$g(x) = \begin{cases} 2e^{-x} & (x \geq 0) \\ 0 & (\text{その他}) \end{cases} \quad h(y) = \begin{cases} 2e^{-y} & (y \geq 0) \\ 0 & (\text{その他}) \end{cases}$$

と取れば

$$p(x, y) = g(x)h(y)$$

となるので, X と Y は独立. □

df: 1-39

定義 A.40. (1) 離散型確率ベクトル (X, Y) は同時 PMF $p(x, y)$ を持つとする. $p_Y(y) > 0$ なる y に対して, $Y = y$ を与えたときの X の条件付き確率関数 (条件付き PMF) を

$$p_{X|Y}(x|y) = \Pr(X = x | Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} = \frac{p(x, y)}{p_Y(y)} \quad (x \in \mathbb{R})$$

で定義する.

(2) 連続型確率ベクトル (X, Y) は同時 PDF $p(x, y)$ を持つとする. $p_Y(y) > 0$ なる y に対して, $Y = y$ を与えたときの X の条件付き確率密度関数 (条件付き PDF) を

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)} \quad (x \in \mathbb{R})$$

で定義する.

re: 1-41

注意 A.41. (X, Y) が連続型確率ベクトルのとき, $\forall A \in \mathcal{B}(\mathbb{R})$ に対して

$$\Pr(X \in A | Y = y) = \int_A p_{X|Y}(x|y) dx$$

となる. □

例 A.42. 連続型確率ベクトル (X, Y) は同時 PDF

$$p(x, y) = \begin{cases} x + y & (0 \leq x \leq 1, 0 \leq y \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

を持つとする. このとき $\Pr(X \leq 1/4 | Y = 1/3)$ を求めてみよう. まず $0 < y < 1$ に対して

$$p_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx = y + \int_0^1 y dx = y + \frac{1}{2}.$$

したがって

$$p_Y(y) = \begin{cases} y + \frac{1}{2} & (0 \leq y \leq 1), \\ 0 & (\text{その他}) \end{cases}$$

となる. よって $0 < y < 1$ と $0 \leq x \leq 1$ に対して

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)} = \frac{x + y}{y + \frac{1}{2}}$$

これより $0 < y < 1$ のとき,

$$p_{(X|Y)}(x|y) = \begin{cases} \frac{x + y}{y + \frac{1}{2}} & (0 \leq x \leq 1) \\ 0 & (\text{その他}) \end{cases}$$

となる. 注意 [re:1-41](#) [A.41](#) より

$$\Pr\left(X \leq \frac{1}{4} \mid Y = \frac{1}{3}\right) = \int_0^{1/4} p_{X|Y}\left(x \mid \frac{1}{3}\right) dx = \int_0^{1/4} \frac{x + \frac{1}{3}}{\frac{1}{3} + \frac{1}{2}} dx = \frac{11}{80}.$$

□

6 多次元分布と i.i.d. 標本

X_1, X_2, \dots, X_n を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数とし,

$$\mathbf{X} = (X_1, X_2, \dots, X_n)$$

と書く. \mathbf{X} を確率ベクトルという.

df:1-42a

定義 A.43. X_1, X_2, \dots, X_n が独立であるとは、すべての Borel 集合 $A_1, A_2, \dots, A_n \in \mathcal{B}(\mathbb{R})$ に対して

$$\Pr(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \prod_{j=1}^n \Pr(X_j \in A_j) \quad (\text{A.2}) \quad \text{eq:1-1}$$

が成立するときである。

X の同時 PDF を $p(x_1, x_2, \dots, x_n)$ と書き、各 X_j ($j = 1, 2, \dots, n$) の周辺 PDF を p_{X_j} と書くことにする。

re:1-43

注意 A.44. (A.2) を示すためには

$$p(x_1, x_2, \dots, x_n) = \prod_{j=1}^n p_{X_j}(x_j) \quad (\forall x_1, x_2, \dots, x_n \in \mathbb{R})$$

を示せばよい。 □

df:1-44

定義 A.45. X_1, X_2, \dots, X_n は独立で、各 X_j ($j = 1, 2, \dots, n$) は同じ CDF F を持つとき、 X_1, X_2, \dots, X_n は独立同一分布に従う (i.i.d. = identically and independently distributed) といい

$$X_1, X_2, \dots, X_n \sim \text{i.i.d. } F$$

と書く⁹。 X_1, X_2, \dots, X_n は累積分布関数 F から標本の大きさが n のランダム標本ともいう。

6.1 重要な多次元分布モデル

この本で取り上げる代表的な多次元分布モデルをあげておく。離散型分布は多項分布、連続型分布は多変量正規分布である。

⁹PDF p を使い

$$X_1, X_2, \dots, X_n \sim \text{i.i.d. } p$$

と書く。 \sim の右側には、確率測度/確率分布/PDF/PMF/ $N(0, 1)$ など分布を特定するものを書いてよいことにする。

6.1.1 多項分布

2 項分布を多変量にしたものが多項分布である. $d, n \in \mathbb{N}$ とし, $\mathbf{p} = (p_1, p_2, \dots, p_d)$, $p_j (0 \leq p_j \leq 1; j = 1, 2, \dots, d)$, $\sum_{j=1}^d p_j = 1$ とする. $\mathbf{X} = (X_1, X_2, \dots, X_d)$ は母数 (d, \mathbf{p}) の多項分布に従うとは, \mathbf{X} の PMF が

$$p(\mathbf{x}) = \binom{d}{x_1 x_2, \dots, x_d} p_1^{x_1} p_2^{x_2} \times \dots \times p_d^{x_d} \quad (x_j = 0, 1, \dots, n; j = 1, 2, \dots, d),$$

$$\binom{d}{x_1 x_2, \dots, x_d} = \frac{n!}{x_1! x_2! \dots x_d!}, \quad 0! = 1$$

で与えられたときをいう. これを $\mathbf{X} \sim \text{Multi}_d(n, \mathbf{p})$ と記す.

em: 1-46

補題 A.46. $\mathbf{X} \sim \text{Multi}_d(n, \mathbf{p})$ とする. ただし, $\mathbf{p} = (p_1, p_2, \dots, p_d)$ と $\mathbf{X} = (X_1, X_2, \dots, X_d)$ とする. このとき $X_j (j = 1, 2, \dots, d)$ の周辺分布は $\text{Bino}(n, p_j)$ である.

Proof. x_2 から x_d について同時 PMF の和をとればよい. □

6.1.2 多変量正規分布

$$\mathbf{Z} := \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_d \end{pmatrix}, \quad Z_1, Z_2, \dots, Z_d \sim \text{N}(0, 1)$$

とする¹⁰. \mathbf{Z} の同時 PDF は

$$p(\mathbf{z}) = p(z_1, z_2, \dots, z_d) = \prod_{j=1}^d \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} z_j^2\right\} = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2} \sum_{j=1}^d z_j^2\right\}$$

$$= \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2} \mathbf{z}^\top \mathbf{z}\right\} \quad (\mathbf{z} \in \mathbb{R}^d)$$

¹⁰この節ではベクトルは縦とする. 下の式では縦ベクトルと横ベクトルを混せて表現している. これは記号の乱用であるが, $f((z_1, z_2, \dots, z_d)^\top)$ などと書くのは煩わしい.

ただし $\mathbf{z} = (z_1, z_2, \dots, z_d)^\top$ である. これを $\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I}_d)$ と記す. ただし \mathbf{I}_d は d 次単位行列である. さらに定義より

$$\int \cdots \int p(z_1, z_2, \dots, z_d) dz_1 dz_2 \cdots dz_d = 1$$

となっていることもわかる.

次に

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix} \in \mathbb{R}^d, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{pmatrix}, \quad \begin{matrix} \sigma_{ij} = \sigma_{ji} \\ (i, j = 1, 2, \dots, d) \end{matrix}$$

とする. ただし $\boldsymbol{\Sigma}$ は正定値対称行列とする. このときある $d \times d$ の正則行列 \mathbf{A} が存在して

$$(1) \mathbf{A} \text{ は対称行列, } (2) \boldsymbol{\Sigma} = \mathbf{A}^2$$

と取れる. これを $\boldsymbol{\Sigma}$ の平方根といい, $\boldsymbol{\Sigma}^{1/2}$ と書くことにする. これを用いて

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \mathbf{Z}$$

と定めたとき, \mathbf{X} の分布を $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ と記す. このとき \mathbf{X} の同時 PDF は

$$p_{\mathbf{X}}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} \text{Det}(\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \quad (\mathbf{x} \in \mathbb{R}^d) \quad (\text{A.3})$$

eq:1-2

で与えられる.

re:1-46

注意 A.47. (A.3) の導出は以下のように行うことができる. 一般に $\mathbf{Z} = (Z_1, Z_2, \dots, Z_d)^\top$ を d 次元確率ベクトルとし, その同時 PDF を $o_{\mathbf{Z}}$ と書くことにする. さらに $S := \{\mathbf{z} \in \mathbb{R}^d; f_{\mathbf{Z}}(\mathbf{z}) > 0\}$, 関数

$$\mathbf{h}(\cdot) = (h_1(\cdot), h_2(\cdot), \dots, h_d(\cdot))^\top : S \longrightarrow \mathbf{h}(S)$$

は 1 対 1 とし

$$X_j = h_j(Z_1, Z_2, \dots, Z_d) \quad (j = 1, 2, \dots, d), \quad \mathbf{X} = (X_1, X_2, \dots, X_d)^\top$$

とおく. h は 1 対 1 なので, h の逆写像を

$$h^{-1} = (h_1^{-1}, h_2^{-1}, \dots, h_d^{-1})^\top : h(S) \longrightarrow S$$

が存在して, $Z = h^{-1}(X)$ となる. X の同時 PDF を求めるために, $h^{-1}(x)$ の Jacobian $J_{h^{-1}}(x)$ を

$$J_{h^{-1}}(x) = \text{Det} \begin{pmatrix} \frac{\partial h_1^{-1}(x)}{\partial x_1} & \dots & \frac{\partial h_1^{-1}(x)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_d^{-1}(x)}{\partial x_1} & \dots & \frac{\partial h_d^{-1}(x)}{\partial x_d} \end{pmatrix}$$

で定める. このとき

$$p_X(x) = |J_{h^{-1}}(x)| p_X(h^{-1}(x)) \quad (\text{A.4}) \quad \boxed{\text{eq:1-3}}$$

となる.

$$h(z) = \mu + \Sigma^{1/2}z \iff h^{-1}(x) = \Sigma^{-1/2}(x - \mu) \text{ より}$$

$$J_{h^{-1}}(x) = \text{Det}(\Sigma)^{-1/2} = \frac{1}{\sqrt{\text{Det}(\Sigma)}}$$

を (A.4) に代入すれば, (A.3) はわかる. 以上の議論から

$$\int \int \dots \int p_X(x) dx_1 dx_2 \dots dx_d = 1$$

となっていることもわかる. □

chm:1-47 定理 A.48. $\mathbf{X} = (X_1, X_2, \dots, X_d)^\top \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ とする. ただし $\boldsymbol{\mu} \in \mathbb{R}^d$ で, $\boldsymbol{\Sigma}$ は $d \times d$ の正定値行列である. このとき次が成立する.

- (1) $X_j (j = 1, 2, \dots, d) \sim N(\mu_j, \sigma_{jj})$. ただし μ_j は $\boldsymbol{\mu}$ の第 j 成分, σ_{jj} は $\boldsymbol{\Sigma}$ の第 j 対角成分である.
- (2) $X_\ell = x_\ell (\ell = 1, 2, \dots, d)$ を与えたときの $X_j (j \neq \ell)$ の条件付き分布は

$$X_j | X_\ell = x_\ell \sim N\left(\mu_j + \frac{\sigma_{\ell\ell}}{\sigma_{jj}}(x_\ell - \mu_\ell), \sigma_{jj} - \frac{\sigma_{j\ell}\sigma_{\ell j}}{\sigma_{\ell\ell}}\right).$$

(3) 定数ベクトル $\mathbf{c} \in \mathbb{R}^d$ ($\mathbf{c} \neq \mathbf{0}$) に対して

$$\mathbf{c}^\top \mathbf{X} \sim N(\mathbf{c}^\top \boldsymbol{\mu}, \mathbf{c}^\top \boldsymbol{\Sigma} \mathbf{c}).$$

(4) $V := (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi_d^2$.

Proof.

□

^[eq:1-3]
 (A.4) を用いた例題を述べておく

ex:1-49

例 A.49. 連続型確率ベクトル (X_1, X_2) は同時 PDF

$$p(x_1, x_2) = \begin{cases} 1 & (0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1), \\ 0 & (\text{その他}) \end{cases}$$

を持つとする.

$$\begin{aligned} y_1 &= h_1(x_1, x_2) = x_1 + x_2, y_2 = h_2(x_1, x_2) = x_2 \\ \iff x_1 &= h_1^{-1}(y_1, y_2), x_2 = h_2^{-1}(y_1, y_2) \end{aligned}$$

とし, $\mathbf{Y} = (h_1(X_1, X_2), h_2(X_1, X_2))^\top$, $\mathbf{y} = (y_1, y_2)^\top$ とする. このとき

$$J_{\mathbf{h}^{-1}}(\mathbf{y}) = \text{Det} \begin{pmatrix} \frac{\partial h_1^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial h_1^{-1}(\mathbf{y})}{\partial y_2} \\ \frac{\partial h_2^{-1}(\mathbf{y})}{\partial y_1} & \frac{\partial h_2^{-1}(\mathbf{y})}{\partial y_2} \end{pmatrix} = \text{Det} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = 1$$

より

$$\begin{aligned} p_{\mathbf{Y}}(\mathbf{y}) &= \begin{cases} 1 & (0 \leq y_1 - y_2 \leq 1, 0 \leq y_2 \leq 1), \\ 0 & (\text{その他}) \end{cases} \\ &= \begin{cases} 1 & (y_2 \leq y_1 \leq 1 + y_2, 0 \leq y_2 \leq 1), \\ 0 & (\text{その他}) \end{cases} \end{aligned}$$

を得る.

$$\int \int p_{\mathbf{Y}}(\mathbf{y}) dy_1 dy_2 = 1$$

となっていることに注意せよ.

□

ここで **Cockran の定理**

7 正規分布に関連した分布

この節では、正規分布に関連した分布であるガンマ分布、 χ^2 分布、 F 分布、 t 分布を定義し、これらの基本的な性質を述べる。これらの分布は、正規母集団から標本に基づく統計量の分布（いわゆる標本分布）の議論で重要な役割を担う。

まず、ガンマ関数を導入する。 $\alpha > 0$ に対してガンマ関数を

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

で定義する。 $\alpha > 1$ のとき、部分積分により

$$\begin{aligned}\Gamma(\alpha) &= \int_0^{\infty} x^{\alpha-1} e^{-x} dx = -x^{\alpha-1} e^{-x} \Big|_0^{\infty} + (\alpha-1) \int_0^{\infty} x^{\alpha-2} e^{-x} dx \\ &= \Gamma(\alpha-1).\end{aligned}$$

$\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$ なので、帰納法により、正の整数 n に対して

$$\Gamma(n) = (n-1)!.$$

7.1 ガンマ分布

df: 1-50 定義 A.50. (ガンマ分布) $\alpha, \beta > 0$ とする。連続型確率変数 X が母数 (α, β) のガンマ分布に従うとは、 X の PDF が

$$p(x|\alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & (x > 0) \\ 0 & (\text{その他}) \end{cases}$$

で与えられるときをいう。これを $X \sim \text{Ga}(\alpha, \beta)$ と記す。

pro: 1-4 問 A.4. (1) $X \sim \text{Ga}(\alpha, \beta)$ とする。 X の k 次の積率は以下で与えられることを示せ。

$$\mathbb{E}[X^k] = \frac{(a+k-1)(a+k-2)\cdots\alpha}{\beta^k}.$$

(2) 上の問いの結果から

$$\mathbb{E}[X] = \frac{\alpha}{\beta}, \quad \mathbb{E}[X^2] = \frac{(\alpha+1)\alpha}{\beta^2}, \quad \mathbb{V}[X] = \frac{\alpha}{\beta^2}$$

この問い
未問題

を示せ.

(3) X の積率母関数は以下で与えられることを示せ.

$$m_X(t) = \left(\frac{1}{1 - \beta t} \right)^\alpha \quad (t < 1/\beta).$$

em: 1-51 補題 A.51. $n \geq 2$ は自然数とする. X_1, X_2, \dots, X_n は独立な確率変数で, 各 $X_j \sim \text{Ga}(\alpha_j, \beta)$ ($j = 1, 2, \dots, n, \alpha_j > 0, \beta > 0$ としたとき

$$X_1 + X_2 + \dots + X_n \sim \text{Ga}(\alpha_1 + \alpha_2 + \dots + \alpha_n, \beta).$$

Proof. 証明は次の補題を用いてする. □

em: 1-52 補題 A.52. 連続型確率変数 X_1, X_2 は独立とし, それぞれの PDF を $f(x), g(x)$ とする. このとき $X_1 + X_2$ の PDF は

$$p_{X_1+X_2}(x) = \int_{-\infty}^{\infty} f(x-y)g(y) dy$$

で与えられる.

Proof. X_1, X_2 の同時 PDF は $f(x)g(y)$ になることに注意する. $t \in \mathbb{R}$ に対して

$$\begin{aligned} \Pr(X_1 + X_2 \leq t) &= \int \int_{x+y \leq t} f(x)g(y) dx dy \\ &= \int_{-\infty}^{\infty} g(y) \left\{ \int_{-\infty}^{t-y} f(x) dx \right\} y \\ &= \int_{-\infty}^{\infty} g(y) \left\{ \int_{-\infty}^t f(z-y) dz \right\} dy \quad (x = z - y \text{ と変換}) \\ &= \int_{-\infty}^t \left\{ \int_{-\infty}^{\infty} f(z-y)g(y) dy \right\} dz. \end{aligned}$$

よって微積分の基本定理より

$$p_{X_1+X_2}(t) = \frac{d}{dt} \Pr(X_1 + X_2 \leq t) = \int_{-\infty}^{\infty} f(t-y)g(y) dy$$

を得る. □

補題 $\frac{\text{lem:1-51}}{\text{A.51}}$ の証明. $X_1 \sim \text{Ga}(\alpha_1, \beta)$, $X_2 \sim \text{Ga}(\alpha_2, \beta)$ とし, X_1 と X_2 は独立とする. $\text{Ga}(\alpha, \beta)$ の PDF を $p(\cdot | \alpha, \beta)$ と書く. 補題 $\frac{\text{lem:1-52}}{\text{A.52}}$ より

$$\begin{aligned} p_{X_1+X_2}(x) &= \int_{-\infty}^{\infty} p(x-y | \alpha_1, \beta) p(y | \alpha_2, \beta) dy \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} e^{-\beta x} \int_0^x (x-y)^{\alpha_1-1} y^{\alpha_2-1} dy \\ &\quad (x-y > 0 \text{ かつ } y > 0 \text{ より } 0 < y < x \text{ となる.}) \\ &= \frac{\beta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1+\alpha_2-1} e^{-\beta x} \int_0^1 (1-z)^{\alpha_1-1} z^{\alpha_2-1} dz \quad (y = xz \text{ と変換}) \end{aligned}$$

を得る. ここで $p_{X_1+X_2}$ は PDF であることに注意すれば

$$\begin{aligned} 1 &= \int_0^{\infty} p_{X_1+X_2}(x) dx \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \left(\int_0^1 (1-z)^{\alpha_1-1} z^{\alpha_2-1} dz \right) \beta^{\alpha_1+\alpha_2} \int_0^{\infty} x^{\alpha_1+\alpha_2-1} e^{-\beta x} dx \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \left(\int_0^1 (1-z)^{\alpha_1-1} z^{\alpha_2-1} dz \right) \int_0^{\infty} w^{\alpha_1+\alpha_2-1} e^{-w} dx \\ &\quad (w = \beta z \text{ と変換}) \\ &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 (1-z)^{\alpha_1-1} z^{\alpha_2-1} dz \\ &\quad (\text{ガンマ関数の定義より}) \end{aligned}$$

より

$$\int_0^1 (1-z)^{\alpha_1-1} z^{\alpha_2-1} dz = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}$$

を得る. よって

$$X_1 + X_2 \sim \text{Ga}(\alpha_1 + \alpha_2, \beta)$$

がわかる. あとはこのことを繰り返せばよい.

7.2 χ^2 分布

定義 A.53. $Z_1, Z_2, \dots, Z_n \sim \text{i.i.d. } \text{N}(0, 1)$ とする.

$$S = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

の分布を自由度 n の χ^2 分布といい、 $S \sim \chi_n^2$ と記す。

$Z \sim N(0, 1)$ のとき $Y = Z^2$ の PDF は

$$p_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} y^{(1/2)-1} e^{-y/2} & (y > 0) \\ 0 & (\text{その他}) \end{cases}$$

となる。演習問題 [Ex:1-1](#) を参照。一方 $\text{Ga}(1/2, 1/2)$ の PDF は

$$p\left(y \mid \frac{1}{2}, \frac{1}{2}\right) = \begin{cases} \frac{1}{\Gamma(\frac{1}{2})} y^{(1/2)-1} e^{-y/2} & (y > 0) \\ 0 & (\text{その他}) \end{cases}$$

であった。これらはともに PDF なので、積分をすれば 1 となる。このことより

$$\sqrt{2}\Gamma\left(\frac{1}{2}\right) = \sqrt{2\pi} \iff \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

を得る。さらに補題 [A.51](#) より

$$\chi_n^2 = \text{Ga}\left(\frac{n}{2}, \frac{1}{2}\right)$$

がわかる。よって χ_n^2 の PDF は

$$p(x|n) = \begin{cases} \frac{\left(\frac{1}{2}\right)^n}{\Gamma\left(\frac{n}{2}\right)} x^{(n/2)-1} e^{-x/2} & (x > 0) \\ 0 & (\text{その他}) \end{cases}$$

である。問 [A.4](#) より $X \sim \chi_n^2$ のとき

$$\mathbb{E}[X] = n, \quad \mathbb{E}[X^2] = 2n$$

となる。

7.3 F 分布

定義 A.54. $k, m \in \mathbb{N}$ とする。 X と Y は独立で $X \sim \chi_k^2$ と $Y \sim \chi_m^2$ とする。このとき

$$F = \frac{X/k}{Y/m}$$

の分布を自由度 (k, m) の F 分布といい、 $F \sim F(k, m)$ と記す。

補題 A.55. $F \sim F(k, m)$ のとき, F の PDF は次で与えられる.

$$p(x|k, m) = \begin{cases} \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} k^{k/2} m^{m/2} x^{(k/2)-1} (m+kx)^{-(k+m)/2} & (x > 0) \\ 0 & (\text{その他}). \end{cases}$$

Proof. 証明は次の補題を用いて行う. □

補題 A.56. X, Y を正値の連続型確率変数とし, それぞれの PDF を $f(x)$ と $g(x)$ とする. このとき

$$Z = \frac{X}{Y}$$

の PDF は

$$h(x) = \begin{cases} \int_0^\infty f(xy)g(y)y dy & (x > 0) \\ 0 & (\text{その他}) \end{cases}$$

で与えられる.

Proof. $t > 0$ に対して

$$\begin{aligned} \Pr(Z \leq t) &= \int_0^\infty \left\{ \int_0^{ty} f(x)g(y) dx \right\} dy \\ &= \int_0^\infty \left\{ \int_0^t f(z)y g(y) dz \right\} dy \quad (x = zy \text{ と変換}) \\ &= \int_0^t \left\{ \int_0^\infty f(z)y g(y) dy \right\} dz. \end{aligned}$$

ここで微積分の基本定理を用いると

$$h(t) = \frac{d}{dt} \Pr(Z \leq t) = \int_0^\infty f(zy)g(y)y dy.$$

□

補題 Lem:1-55 A.55 の証明 補題 Lem:1-56 A.56 より, $w > 0$ のとき, $W = X/Y$ の PDF は

$$\begin{aligned}
 p_W(w) &= \int f(wy)g(y)y \, dy \\
 &= \frac{\left(\frac{1}{2}\right)^{k/2} \left(\frac{1}{2}\right)^{m/2}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} \int_0^\infty (wy)^{(k/2)-1} e^{-(wy)/2} y^{(m/2)-1} e^{-y} y \, dy \\
 &= \frac{\left(\frac{1}{2}\right)^{(k+m)/2}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} w^{(k/2)-1} \int_0^\infty y^{(k+m)/2-1} e^{-(w+1)y/2} \, dy \\
 &= \frac{\left(\frac{1}{2}\right)^{(k+m)/2}}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} w^{(k/2)-1} \int_0^\infty \left(\frac{2x}{w+1}\right)^{(k+m)/2-1} e^{-x} \frac{2}{z+1} \, dx \\
 &\quad (x = \frac{w+1}{2}y) \text{ と変換} \\
 &= \frac{1}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} w^{(k/2)-1} (1+w)^{-(k+m)/2} \int_0^\infty x^{(k+m)/2-1} e^{-x} \, dx \\
 &= \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} w^{(k/2)-1} (1+w)^{-(k+m)/2}.
 \end{aligned}$$

よって

$$p(x|k, m) = p_W\left(\frac{k}{m}z\right) \frac{k}{m}$$

よりわかる.

7.4 t 分布

df:1-57

定義 A.57. $Z_0, Z_1, Z_2, \dots, Z_n \sim \text{i.i.d. } N(0, 1)$ のとき

$$T = \frac{Z_0}{\sqrt{\frac{1}{n}(Z_1^2 + Z_2^2 + \dots + Z_n^2)}}$$

を自由度 n の t 分布といい, これを $T \sim t_n$ と記す.

Lem:1-58

補題 A.58. T の PDF は

$$p_T(x|n) = \begin{cases} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} & (x > 0) \\ 0 & (\text{その他}) \end{cases}$$

で与えられる

Proof. 定義より

$$T^2 \sim F(1, n)$$

であることをまず思い出す。 T の分布は対称なので

$$p_T(x|n) = p_T(-x|n) \quad (x \geq 0)$$

となる。 よって $x > 0$ に対して

$$2\Pr(0 \leq T \leq x) = \Pr(-x \leq T \leq x) = \Pr(T^2 \leq x^2).$$

よって $F(1, n)$ の PDF を $p_{1,n}$ と書けば

$$2 \int_0^x p_T(x|n) dx = \int_0^{x^2} p_{1,n}(x) dx.$$

$x = y^2$ と変換すると

$$\int_0^{x^2} p_{1,n}(x) dx = \int_0^t p_{1,n}(y^2) 2y dy.$$

したがって

$$p_T(x|n) = p_{1,n}(x^2)x = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

となる。 □

第 B 章 期待値

1 期待値

X を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の確率変数とする. X が離散型のとき, その PMF を p とし, X の取り得る値を x_1, x_2, \dots とする. 連続型のとき, その PDF を p とする.

df:2-1 定義 B.1. 関数 $g: \mathbb{R} \rightarrow \mathbb{R}$ とする. $g(X)$ の期待値 $\mathbb{E}[g(X)]$ を次のように定義する.

(1) $g \geq 0$ のとき

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{n=1}^{\infty} g(x_n)p(x_n) & (\text{離散型}) \\ \int_{-\infty}^{\infty} g(x)p(x) dx & (\text{連続型}) \end{cases}$$

と定義する. 右辺は ∞ を許せば, 必ず存在する.

(2) 一般の g に対して

$$g^+(x) = \max\{g(x), 0\}, \quad g^-(x) = \max\{-g(x), 0\}$$

と定義すれば, $g^+ \geq 0, g^- \geq 0$ となる. $\mathbb{E}[g^+(X)]$ または $\mathbb{E}[g^-(X)]$ のいずれかが有限ならば,

$$\mathbb{E}[g(X)] := \mathbb{E}[g^+(X)] - \mathbb{E}[g^-(X)]$$

と定義する. $\mathbb{E}[g^+(X)] = \mathbb{E}[g^-(X)] = \infty$ のときは, $g(X)$ の期待値は定義されない. $\mathbb{E}[g^+(X)] < \infty$ かつ $\mathbb{E}[g^-(X)] < \infty$ のとき, $\mathbb{E}[g(X)]$ は有限となる.

lem:2-2 補題 B.2. X を確率変数とする. 関数 $g, h: \mathbb{R} \rightarrow \mathbb{R}$ に対して, $\mathbb{E}[|g(X)|] < \infty, \mathbb{E}[|h(X)|] < \infty$ を仮定する.

(1) $a, b \in \mathbb{R}$ に対して

$$\mathbb{E}[ag(X) + bh(X)] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)].$$

(2) $g(x) \leq h(x) (\forall x \in \mathbb{R})$ ならば

$$\mathbb{E}[g(X)] \leq \mathbb{E}[h(X)].$$

Proof. 積分の性質に帰着される. □

lem:2-3

補題 B.3. $0 < q < r$ に対して

$$\mathbb{E}[|X|^r] < \infty \implies \mathbb{E}[|X|^q] < \infty.$$

Proof. 積分の性質に帰着される. □

df:2-4

定義 B.4. (1) $k = 1, 2, \dots$ に対して, $\mathbb{E}[|X|^k] < \infty$ のとき, $\mathbb{E}[X^k]$ を X の k 次モーメント (または積率) という.

(2) $\mathbb{E}[|X|] < \infty$ のとき $\mathbb{E}[X]$ を X の平均値¹という.

(3) $\mathbb{E}[X^2] < \infty$ のとき X の分散を

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

で定義する.

(4) $A \subset \mathbb{R}$ に対して $\mathbb{1}_A$ を A の指示関数という. $A \in \mathcal{B}(\mathbb{R})$ のとき

$$\mathbb{E}[\mathbb{1}_A(X)] = \Pr(X \in A)$$

となる.

re:2-2

注意 B.5. (1) X は確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上の離散型確率変数とする. p を X の PMF とし

$$S_X := \{x \in \mathbb{R}; p(x) > 0\}$$

とする. 関数 $g: \mathbb{R} \rightarrow \mathbb{R}$ に対して

$$Y = g(X)$$

¹簡単に「平均」ともいう.

とおく. $y \in \mathbb{R}$ に対して

$$A_y := \{x \in S_X; g(x) = y\}$$

とおき

$$S_Y := \{y \in \mathbb{R}; A_y \neq \emptyset\}$$

とする. $y \in S_Y$ に対して

$$q(y) := \Pr(Y = y) = \Pr(X \in A_y) = \sum_{x \in A_y} \Pr(X = x) = \sum_{x \in A_y} p(x)$$

と書ける. 正項級数は項の順番を並び替えてもその値は変わらないので

$$\sum_{y \in S_Y} |y|q(y) = \sum_{y \in S_Y} |y| \sum_{x \in S_y} p(x) = \sum_{y \in S_Y} \sum_{x \in A_y} |g(x)|p(x) = \sum_{x \in S_X} |g(x)|p(x)$$

となる. さらにいずれかの和が有限ならば

$$\sum_y yq(y) = \sum_y y \sum_{x \in S_y} p(x) = \sum_y \sum_{x \in S_y} g(x)p(x) = \sum_x g(x)p(x)$$

となる. よって

$$\mathbb{E}[Y] = \mathbb{E}[g(X)]$$

となる.

(2) X は連続型とする. p_X を X の PDF とし

$$S_X := \{x \in \mathbb{R}; p_X(x) > 0\}$$

とする. $g: S_X \rightarrow \mathbb{R}$ を狭義単調増加かつ C^1 級とする.

$$Y = g(X), \quad S_Y := \{y \in \mathbb{R}; y = g(x) (\exists x \in S_X)\}$$

とする. $g: S_X \rightarrow S_Y$ と制限²すれば g の逆関数 $g^{-1}: S_Y \rightarrow S_X$ が存在し

$$F_Y(y) := \Pr(Y \leq y) = \Pr(g^{-1}(Y) \leq g^{-1}(y)) = \Pr(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

²制限したものを同じ g を用いて表現するのは, 記号の乱用である. 記号が煩雑になるので, 記号を乱用した.

さらに $g(g^{-1}(y)) = y$ より

$$\frac{d}{dy}g^{-1}(y) = \frac{1}{\dot{g}(g^{-1}(y))}$$

となる。ただし $\dot{g}(y) = \frac{dg}{dy}(y)$ である。これらより, $y \in S_Y$ に対して

$$\begin{aligned} p_Y(y) &= \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = \dot{F}_X(g^{-1}(y)) \frac{d}{dy}g^{-1}(y) \\ &= p_X(g^{-1}(y)) \frac{1}{\dot{g}(g^{-1}(y))} \end{aligned}$$

となる。よって

$$p_Y(y) = p_X(g^{-1}(y)) \frac{1}{\dot{g}(g^{-1}(y))}$$

となる。 g が狭義単調減少の場合もふくめると

$$p_Y(y) = p_X(g^{-1}(y)) \frac{1}{|\dot{g}(g^{-1}(y))|}$$

となる。 □

2 確率ベクトルの期待値

X, Y を確率空間 $(\Omega, \mathcal{F}, \Pr)$ 上で定義された確率変数とする。 (X, Y) を確率ベクトルという。

(X, Y) が離散型るときその同時 PMF を $p(x, y)$ とし, 連続型るときその同時 PDF を $f(x, y)$ とする。

df: 2-6 定義 B.6. 関数 $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ に対して $g(X, Y)$ の期待値を次のように定義する。

(1) $g \geq 0$ のとき

$$\mathbb{E}[g(X, Y)] = \begin{cases} \sum_{x, y} g(x, y)p(x, y) & \text{(離散型)} \\ \int \int g(x, y)p(x, y) dx dy & \text{(連続型)}. \end{cases}$$

(2) 一般の g に対して

$$g^+(x, y) := \max\{g(x, y), 0\}, \quad g^-(x, y) := \max\{-g(x, y), 0\}$$

と定義すれば $g^+, g^- \geq 0$ となる. $\mathbb{E}[g^+(X, Y)]$ または $\mathbb{E}[g^-(X, Y)]$ のいずれかが有限ならば

$$\mathbb{E}[g(X, Y)] := \mathbb{E}[g^+(X, Y)] - \mathbb{E}[g^-(X, Y)]$$

で定義する. $\mathbb{E}[g^+(X, Y)] = \mathbb{E}[g^-(X, Y)] = \infty$ のときは, $g(X, Y)$ の期待値は定義されない. $\mathbb{E}[g^+(X, Y)] < \infty$ かつ $\mathbb{E}[g^-(X, Y)] < \infty$ のとき, $\mathbb{E}[g(X, Y)]$ は有限の値を取る.

re:2-7

注意 B.7. 3 つ以上の確率変数 X_1, X_2, \dots, X_n に対しても期待値を定義 ^{df:2-6} B.6 と同様に定義する. □

thm:2-8

定理 B.8. X_1, X_2, \dots, X_n を確率変数とし, 各 X_j ($j = 1, 2, \dots, n$) の期待値は有限とする. a_1, a_2, \dots, a_n を定数としたとき

$$\mathbb{E}\left[\sum_{j=1}^n a_j X_j\right] = \sum_{j=1}^n a_j \mathbb{E}[X_j].$$

Proof. 期待値の線型性 (補題 ^{lem:2-2} B.2(1)) よりわかる. □

ex:2-9

例 B.9. $0 < p < 1$ と $j = 1, 2, \dots, n$ に対して, $X_j \sim \text{Bino}(p)$ とする. このとき

$$\mathbb{E}[X_j] = \sum_{x=0}^1 x \Pr(X = x) = 0 \times (1 - p) + 1 \times p = p.$$

$S = \sum_{j=1}^n X_j$ としたとき

$$\mathbb{E}[S] = \sum_{j=1}^n \mathbb{E}[X_j] = np.$$

□

3 分散と共分散

sec:2-3

df:2-10

定義 B.10. X を確率変数とし, $\mathbb{E}[X^2] < \infty$ とする. X の分散を

$$\mathbb{V}[X] := \mathbb{E}[(X - \mu)^2]$$

で定義する. ただし $\mu = \mathbb{E}[X]$ と書いた. さらに $\sqrt{\mathbb{V}[X]}$ を X の標準偏差という.

re:2-11 注意 B.11. 分散 $\mathbb{V}[X]$ は X の分布の平均 μ まわりの散らばりを測る量である. 分散がおおきいほど分布は広がっていることになる. \square

thm:2-12 定理 B.12. 以下の確率変数は有限の 2 次の積率を持つとする. このとき, 次が成立する.

(1) $\mathbb{V}[X] = \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2$.

(2) 定数 $a, b \in \mathbb{R} (a \neq 0)$ に対して

$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X].$$

(3) X と Y は独立で, $\mathbb{E}[|XY|] < \infty$ とする. このとき

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

(4) X_1, X_2, \dots, X_n は独立とし, $\mathbb{E}[X_j^2] < \infty (j = 1, 2, \dots, n)$ とする. a_1, a_2, \dots, a_n は定数としたとき,

$$\mathbb{V}[a_1X_1 + a_2X_2 + \dots + a_nX_n] = a_1^2\mathbb{V}[X_1] + a_2^2\mathbb{V}[X_2] + \dots + a_n^2\mathbb{V}[X_n].$$

Proof. (1), (2) は分散を期待値の記号を用いて表現し, 期待値の線型性を用いて計算すればよい. (3) については, 連続型の場合を示す. (X, Y) の同時 PDF は, X と Y の周辺 PDF p_X と p_Y を用いて $p_X(x)p_Y(y)$ という形でかけるので,

$$\mathbb{E}[XY] = \int \int xyp_X(x)p_Y(y) dx dy = \int xp_X(x) dy \int yp_Y(y) dy = \mathbb{E}[X]\mathbb{E}[Y].$$

(4) は, 分散を期待値の記号を用いて表現し, (3) に注意して, 期待値の線型性を用いて計算すればよい. \square

例 B.13. (例 ^{ex:2-9}B.9 の続き) 例 ^{ex:2-9}B.9 の設定に加えて X_1, X_2, \dots, X_n は独立とする. 定理 ^{thm:2-12}B.12(1) に注意すれば

$$\begin{aligned} \mathbb{V}[X_j] &= \mathbb{E}[X_j^2] - \{\mathbb{E}[X_j]\}^2 = \sum_{x=0}^2 x^2 \Pr(X = x) - p^2 \\ &= 0 \times (1 - p) + 1 \times p - p^2 = p(1 - p). \end{aligned}$$

これと定理 $\frac{\text{thm:2-12}}{\text{B.12(4)}}$ から

$$\mathbb{V}[S] = \sum_{j=1}^n \mathbb{V}[X_j] = np(1-p).$$

□

hm:2-14

定理 B.14. $n \geq 2$ とし, X_1, X_2, \dots, X_n は i.i.d. 確率変数列とし

$$\mathbb{E}[X_1] = \mu, \quad \mathbb{V}[X_1] = \sigma^2, \quad 0 < \sigma < \infty$$

とする. X_1, X_2, \dots, X_n に基づく標本平均を

$$\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$$

で定義し, 標本 (不偏) 分散を

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2$$

で定義する. このとき

$$(1) \mathbb{E}[\bar{X}_n] = \mu, \quad (2) \mathbb{V}[\bar{X}_n] = \frac{\sigma^2}{n}, \quad (3) \mathbb{E}[S_n^2] = \sigma^2.$$

Proof. (1), (2) は定理 $\frac{\text{thm:2-8}}{\text{B.8}}, \frac{\text{thm:2-12}}{\text{B.12(4)}}$ よりわかる. (3) を証明するために

$$\sum_{j=1}^n (X_j - \bar{X}_n)^2 = \sum_{j=1}^n (X_j - \mu)^2 - n(\bar{X}_n - \mu)^2$$

に注意する. 定理 $\frac{\text{thm:2-12}}{\text{B.12(1)}}$ より

$$\mathbb{E} \left[\sum_{j=1}^n (X_j - \mu)^2 \right] = n\sigma^2.$$

(1) と (2) より

$$\mathbb{E}[(\bar{X}_n - \mu)^2] = \mathbb{V}[\bar{X}_n] = \frac{\sigma^2}{n}.$$

したがって

$$\begin{aligned}\mathbb{E}[S_n^2] &= \frac{1}{n-1} \left\{ \sum_{j=1}^n \mathbb{E}[(X_j - \mu)^2] - n\mathbb{E}[(\bar{X}_n - \mu)^2] \right\} \\ &= \frac{1}{n-1} \{n\sigma^2 - \sigma^2\} = \sigma^2.\end{aligned}$$

□

df: 2-15

定義 B.15. X と Y は確率変数とし

$$\mathbb{E}[X] = \mu_X, \quad \mathbb{V}[X] = \sigma_X^2, \quad \mathbb{E}[Y] = \mu_Y, \quad \mathbb{V}[Y] = \sigma_Y^2$$

とする. ただし $\mu_X, \mu_Y \in \mathbb{R}$, $0 < \sigma_X, \sigma_Y < \infty$ とする. このとき X と Y の共分散を

$$\text{COV}[X, Y] = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

で定義し, X と Y の (Pearson) の相関係数を

$$\rho := \rho[X, Y] := \frac{\text{COV}[X, Y]}{\sigma_X \sigma_Y}$$

で定義する.

thm: 2-16

定理 B.16. X, Y, Z は 2 次の積率が有限な確率変数とする.

(1) 共分散は

$$\text{COV}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

(2) $\text{COV}[X, Y] = \text{COV}[Y, X]$.

(3) 定数 $a, b \neq 0$ に対して

$$\text{COV}[aX + bY, Z] = a\text{COV}[X, Z] + b\text{COV}[Y, Z].$$

(4) 相関係数は

$$-1 \leq \rho[X, Y] \leq 1$$

をみたす.

(5) ある定数 a, b が存在して $Y = aX + b$ となったとき

$$a > 0 \implies \rho[X, Y] = 1,$$

$$a < 0 \implies \rho[X, Y] = -1.$$

(6) X と Y が独立のとき

$$\text{COV}[X, Y] = 0.$$

Proof. (1) は期待値の線型性からわかる. (2)(3) は共分散の定義と期待値の線型性よりわかる. (4) は $\mathbb{V}[X] = \mathbb{V}[Y] = 0$ のときは, 明らかなので, $\mathbb{V}[X] \neq 0$ として, 証明する. 補題 [B.2\(2\)](#) より

$$\mathbb{V}[X]t^2 - 2\text{COV}[X, Y]t + \mathbb{V}[Y] = \mathbb{E}[\{t(X - \mathbb{E}[X]) + Y - \mathbb{E}[Y]\}^2] \geq 0$$

となる. したがって, 判別式をとれば

$$\{\text{COV}[X, Y]\}^2 - \mathbb{V}[X]\mathbb{V}[Y] \leq 0$$

がわかる. (5) は (2) からわかる. (6) は定理 [B.12\(2\)](#) からわかる. □

注意 B.17. (6) の逆は一般に正しくない. 反例は下の問いをみよ. □

問 B.1. Z を $[0, 2\pi]$ 上の一様分布とし

$$X = \cos Z, \quad Y = \sin Z$$

とおく.

- (1) $\mathbb{E}[X] = 0, \mathbb{E}[Y] = 0, \mathbb{E}[XY] = 0$ を確かめよ.
- (2) $\Pr(X \leq 1/2, Y \leq 1/2) = \Pr((5/4)\pi \leq Z \leq (6/4)\pi)$ を確かめよ.
- (3) $\Pr(X \leq 1/2) = \Pr((1/4)\pi \leq Z \leq (7/4)\pi)$ を確かめよ³.

定理 B.18. (1) X, Y は有限な 2 次の積率を持つ確率変数とする. このとき $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\text{COV}[X, Y]$.

³これらより, $\Pr(X \leq 1/2, Y \leq 1/2) \neq \Pr(X \leq 1/2)\Pr(Y \leq 1/2)$ となり, $\text{COV}[X, Y] = 0$ だが, X と Y は従属であることがわかる.

(2) $d \geq 2$ とする. X_1, X_2, \dots, X_d は有限な 2 次の積率を持つ確率変数とする. このとき

$$\mathbb{V}\left[\sum_{j=1}^d a_j X_j\right] = \sum_{j=1}^d a_j^2 \mathbb{V}[X_j] + 2 \sum_{j=1}^d \sum_{\ell=j+1}^d a_j a_\ell \text{COV}[X_j, X_\ell].$$

Proof. 共分散を期待値で表現し, 展開して期待値の線型性を用いればよい. \square

問 B.2. 定理 [B.18\(1\)](#) の証明を具体的に書け.

X_1, X_2, \dots, X_d を有限な 2 次の積率を持つ確率変数とし

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}$$

と書く. このとき確率ベクトル \mathbf{X} の期待値を

$$\mathbb{E}[\mathbf{X}] := \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

で定義する. \mathbf{X} の共分散を

$$\mathbb{V}[\mathbf{X}] := \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top]$$

で定義する⁴. ただし $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_d)^\top := \mathbb{E}[\mathbf{X}]$ である. これは

$$\mathbb{V}[\mathbf{X}] = \begin{pmatrix} \mathbb{V}[X_1] & \text{COV}[X_1, X_2] & \cdots & \text{COV}[X_1, X_d] \\ \text{COV}[X_2, X_1] & \mathbb{V}[X_2] & \cdots & \text{COV}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \text{COV}[X_d, X_1] & \text{COV}[X_d, X_2] & \cdots & \mathbb{V}[X_d] \end{pmatrix}$$

である.

⁴ 確率ベクトルと同様に確率変数を成分とする行列を確率行列という. 確率行列の期待値はそれぞれの成分の期待値を取ったものを配置した行列と定義している.

em:2-19

補題 B.19. $X = (X_1, X_2, \dots, X_d)^\top$ は確率ベクトルで各成分は有限な 2 次の積率を持つとし

$$\mathbb{E}[X] = \mu, \quad \mathbb{V}[X] = \Sigma$$

とする。ただし $\mu \in \mathbb{R}^d$, Σ は $d \times d$ の半正定値行列⁵である。

(1) 任意の定数ベクトル $a \in \mathbb{R}^d$ に対して

$$\mathbb{E}[a^\top X] = a^\top \mu, \quad \mathbb{V}[a^\top X] = a^\top \Sigma a.$$

(2) 任意の定数の d 次正方行列 A に対して

$$\mathbb{E}[AX] = A\mathbb{E}[X], \quad \mathbb{V}[AX] = A\mathbb{V}[X]A^\top.$$

Proof. 期待値の線型性を用いて計算すればよい。 □

問 B.3. $d = 2$ として、補題 ^{em:2-19}B.19 を確認せよ。

co:2-20

系 B.20. 確率ベクトル X は、任意の定数ベクトル $a (\neq 0)$ に対して、

$$\Pr(a^\top X = 0) = 0$$

をみたす。このとき、 Σ は正定値である。

re:2-21

注意 B.21. X の共分散行列 Σ は、補題 ^{em:2-19}B.19(2) から半正定値であることがわかる。系 ^{co:2-20}B.20 の仮定は、確率ベクトル X が d 次元空間より次元の低い空間に集中しないことを意味している。したがって、確率ベクトル X の値を取る空間が退化していなければ、その共分散行列は正定値となることがわかる。 □

4 条件付き期待値

df:2-22

定義 B.22. (1) X と Y を確率変数とし、条件付き PDF(条件付き PMF) を $p_{X|Y}(p_{X|Y})$ とする。 $Y = y$ を与えたときの X の条件付き期待値を

$$\mathbb{E}[X|Y = y] = \begin{cases} \sum_x x p_{X|Y}(x|y) & (\text{離散型}) \\ \int x p_{X|Y}(x|y) dx & (\text{連続型}) \end{cases}$$

⁵ $d \times d$ の対称行列 A が半正定値であるとは、 $\forall a \in \mathbb{R}^d$ に対して $a^\top A a \geq 0$ が成立するときをいう。また $d \times d$ の対称行列 A が正定値であるとは、 $\forall a \in \mathbb{R}^d (a \neq 0)$ に対して $a^\top A a > 0$ が成立するときをいう。

で定義する. ただし考えている $Y = y$ で条件付き PDF または PMF は定義され $\mathbb{E}[|X|] < \infty$ とする.

(2) (Borel 可測) 関数 $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ に対して, $g(X, Y)$ の条件付き期待値を

$$\mathbb{E}[g(X, Y)|Y = y] = \begin{cases} \sum_x g(x, y)p_{X|Y}(x|y) & (\text{離散型}) \\ \int g(x, y)p_{X|Y}(x|y) dx & (\text{連続型}) \end{cases}$$

で定義する. ただし考えている $Y = y$ で条件付き PDF または PMF は定義され $\mathbb{E}[|g(X, Y)|] < \infty$ とする.

ex:2-24

注意 B.23. $\mathbb{E}[X]$ は定数であるが, $\mathbb{E}[X|Y = y]$ は一般に y の関数である.

$$h(y) := \mathbb{E}[X|Y = y]$$

とおいたときに $h(y)$ に Y を代入したものは確率変数になる. これを

$$\mathbb{E}[X|Y] := h(Y)$$

と定義する. したがって $\omega \in \Omega$ に対して, $y = Y(\omega)$ と書けば

$$\mathbb{E}[X|Y]: \Omega \ni \omega \mapsto \mathbb{E}[X|Y(\omega)] = \mathbb{E}[X|Y = y] \in \mathbb{R}$$

は可測となる. □

ex:2-25

例 B.24. 連続型確率変数 Y は PDF

$$p(y) = \begin{cases} 1 & (0 < y < 1), \\ 0 & (\text{その他}) \end{cases}$$

を持つとする. $Y = y$ ($0 < y < 1$) を観測したとき

$$X|Y = y \sim \text{Unif}(y, 1)$$

とする. すなわち $0 < y < 1$ のとき

$$p_{X|Y}(x|y) = \begin{cases} \frac{1}{1-y} & (y < x < 1), \\ 0 & (\text{その他}) \end{cases}$$

である. よって

$$\mathbb{E}[X|Y=y] = \int_y^1 yp_{X|Y}(x|y) dx = \frac{1}{1-y} \int_y^1 dx = \frac{1+y}{2}.$$

これより

$$\mathbb{E}[X|Y] = \frac{1+Y}{2}$$

となる. □

hm: 2-26

定理 B.25. (1) 有限な期待値を持つ確率変数 X, Y と確率変数 Z に対して

$$\mathbb{E}[X+Y|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z].$$

(2) 有限な期待値を持つ確率変数 X, Y に対して

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y], \quad \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

(3) 一般の (Borel 可測) 関数 $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ を考える. $\mathbb{E}[|g(X, Y)|] < \infty$ のとき

$$\mathbb{E}[\mathbb{E}[g(X, Y)|Y]] = \mathbb{E}[g(X, Y)].$$

(4) $\mathbb{E}[XY|Y] = Y\mathbb{E}[X|Y]$.

Proof. (1) は期待値の線型性よりわかる. 連続型の場合について (2) の第 1 番目の等式を示す. 他の場合もほとんど同じように証明できる. f を (X, Y) の同時 PDF とする. このとき

$$p(x, y) = p_Y(y)p_{X|Y}(x|y)$$

となる. ただし p_X は X の周辺 PDF で $p_{Y|X}$ は $X = x$ を与えたときの Y の

条件付き PDF である. すると

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y]] &= \int \mathbb{E}[X|Y=y]p_Y(y) dy \\ &= \int \left\{ \int yp_{X|Y}(x|y) dx \right\} p_Y(y) dy \\ &= \int \int yp_{X|Y}(x|y)p_Y(y) dx dy \\ &= \int \int x \left\{ \int p(x, y) dy \right\} dx \\ &= \int xp_X(x) dx \\ &= \mathbb{E}[X].\end{aligned}$$

積分の順序交換は $g(X, Y)$ が有限の期待値を持つことから保証される. □

例 B.26. (例 B.24^{ex:2-25}の続き)

$$\mathbb{E}[X] = \mathbb{E}\left[\frac{1+Y}{2}\right] = \frac{3}{4}$$

となる. 一方 $0 < x < y < 1$ に対して

$$p(x, y) = p_{X|Y}(x|y)p_Y(y) = \frac{1}{1-y}.$$

よって

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 \left\{ \int_y^1 x \frac{1}{1-y} dx \right\} dy \\ &= \int \frac{1}{1-y} \left[\frac{x^2}{2} \right]_y^1 dy \\ &= \int_0^1 \frac{1}{1-y} \frac{1-y^2}{2} dy = \left[\frac{(1+y)^2}{4} \right]_0^1 = \frac{3}{4}.\end{aligned}$$

□

定義 B.27. X は有限の 2 次の積率を持つとする. $Y = y$ を与えたときの X の条件付き PDF $f_{X|Y}$ (PMF $p_{X|Y}$) が定義できる y を考える. このとき, $Y = y$ を与えたときの条件付き分散を

$$\mathbb{V}[X|Y = y] = \begin{cases} \sum_x \{x - \mu(y)\}^2 p_{X|Y}(x|y) & (\text{離散型}) \\ \int \{x - \mu(y)\}^2 p_{X|Y}(x|y) & (\text{連続型}) \end{cases}$$

で定義する. ただし $\mu(y) = \mathbb{E}[X|Y = y]$ である. これは

$$\mathbb{V}[X|Y] = \mathbb{E}[X^2|Y] - \{\mathbb{E}[X|Y]\}^2$$

とも書ける.

定理 B.28. X, Y を確率変数とし $\mathbb{E}[X^2] < \infty$ とする. このとき

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X|Y]] + \mathbb{V}[\mathbb{E}[X|Y]]$$

が成立する. ただし $h(y) = \mathbb{V}[X|Y = y]$ としたとき $\mathbb{V}[X|Y] := h(Y)$ と定義した.

Proof.

$$\begin{aligned} \mathbb{V}[X] &= \mathbb{E}[\{X - \mathbb{E}[X]\}^2] \\ &= \mathbb{E}[\{X - \mathbb{E}[X|Y] + \mathbb{E}[X|Y] - \mathbb{E}[X]\}^2] \\ &= \mathbb{E}[\{X - \mathbb{E}[X|Y]\}^2] + \mathbb{E}[\{\mathbb{E}[X|Y] - \mathbb{E}[X]\}^2] \\ &\quad + 2\mathbb{E}[\{X - \mathbb{E}[X|Y]\}\{\mathbb{E}[X|Y] - \mathbb{E}[X]\}]. \end{aligned}$$

しかし

$$\begin{aligned}\mathbb{E}[\{X - \mathbb{E}[X|Y]\}^2] &= \mathbb{E}[X^2 - 2X\mathbb{E}[X|Y] + \{\mathbb{E}[X|Y]\}^2] \\ &= \mathbb{E}\left[\mathbb{E}[X^2 - 2X\mathbb{E}[X|Y] + \{\mathbb{E}[X|Y]\}^2 \mid Y]\right] \\ &= \mathbb{E}\left[\mathbb{E}[X^2|Y] - 2\mathbb{E}[X|Y]\mathbb{E}[X|Y] + \{\mathbb{E}[X|Y]\}^2\right] \\ &= \mathbb{E}\left[\mathbb{E}[X^2|Y] - \{\mathbb{E}[X|Y]\}^2\right] \\ &= \mathbb{E}\left[\mathbb{V}[X|Y]\right], \\ \mathbb{E}\left[\{\mathbb{E}[X|Y] - \mathbb{E}[X]\}^2\right] &= \mathbb{E}\left[\{\mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Y]]\}^2\right] = \mathbb{V}\left[\mathbb{E}[X|Y]\right], \\ \mathbb{E}[\{X - \mathbb{E}[X|Y]\}\{\mathbb{E}[X|Y] - \mathbb{E}[X]\}] &= \mathbb{E}\left[\mathbb{E}[\{X - \mathbb{E}[X|Y]\}\{\mathbb{E}[X|Y] - \mathbb{E}[X]\} \mid Y]\right] \\ &= \mathbb{E}\left[\{\mathbb{E}[X|Y] - \mathbb{E}[X]\} \underbrace{\mathbb{E}[X - \mathbb{E}[X|Y] \mid Y]}_{=0}\right] \\ &= 0.\end{aligned}$$

最後から 2 番目の等号は定理 [B.25\(3\)](#) を用いた。 □

5 積率母関数

df: 2-30

定義 B.29. X を確率変数とし, ある $t_0 > 0$ が存在して, $\mathbb{E}[e^{tX}] < \infty$ ($\forall |t| < t_0$) とする. このとき, X の積率母関数 (Moment Generating Function (MGF)) を

$$m_X(t) := \mathbb{E}[e^{tX}] \quad (-t_0 < t < t_0)$$

と定義する.

re: 2-29

能がここ
メントを

注意 B.30. $m_X(t)$ が存在するとき, 期待値と微分の記号の入れ替えが保証され,

$$\begin{aligned} \dot{m}_X(0) &= \dot{m}_X(t)|_{t=0} = \left[\frac{d}{dt} m_X(t) \right] \Big|_{t=0} = \left[\frac{d}{dt} \mathbb{E}[e^{tX}] \right] \Big|_{t=0} = \mathbb{E} \left[\frac{d}{dt} e^{tX} \right] \Big|_{t=0} \\ &= \mathbb{E}[X e^{tX}] \Big|_{t=0} = \mathbb{E}[X] \end{aligned}$$

となる. この議論を繰り返せば $k = 2, 3, \dots$ に対して

$$m_X^{(k)}(0) = \mathbb{E}[X^k].$$

□

ex:2-32

例 B.31. $X \sim \text{Exp}(1)$ とする. $t < 1$ に対して

$$m_X(t) = \mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} e^{-x} dx = \int_0^\infty e^{-(1-t)x} dx = \frac{1}{1-t}.$$

$t \geq 1$ のときは, e^{tX} の期待値は発散する. したがって

$$m_X(t) = \frac{1}{1-t} \quad (t < 1)$$

となる. いま

$$\dot{m}_X(0) = 1, \quad \ddot{m}_X(0) = 2$$

なので

$$\mathbb{E}[X] = 1, \quad \mathbb{E}[X^2] = 2, \quad \mathbb{V}[X] = \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2 = 1.$$

□

em:2-33

補題 B.32. (1) $a, b \in \mathbb{R}$ ($a \neq 0$) とする. $Y = aX + b$ としたとき

$$m_Y(t) = e^{tb} m_X(at).$$

(2) X_1, X_2, \dots, X_d は独立とし $Y = \sum_{j=1}^d Y_j$ とする. このとき

$$m_Y(t) = \prod_{j=1}^d m_{X_j}(t).$$

Proof. (1) は指数関数と期待値の性質よりわかる. (2) は指数関数の性質と定理 B.12(3) よりわかる. \square

ex:2-34

定理 B.33. X と Y を確率変数とする. ある数 $t_0 > 0$ が存在して

$$m_X(t) = m_Y(t) \quad (|t| < t_0)$$

ならば

$$X \stackrel{d}{=} Y.$$

Proof. これは信じることにする. \square

ex:2-33

例 B.34. $n_1, n_2 \in \mathbb{N}$, $0 < p < 1$ とし $X_1 \sim \text{Bino}(n_1, p)$, $X_2 \sim \text{Bino}(n_2, p)$ は独立とする. $Y = X_1 + X_2$ としたとき

$$m_Y(t) = m_{X_1}(t)m_{X_2}(t) = (pe^t + q)^{n_1}(pe^t + q)^{n_2} = (pe^t + q)^{n_1+n_2}.$$

ただし $q = 1 - p$. よって $Y \sim \text{Bino}(n_1 + n_2, p)$. \square

ex:2-34

例 B.35. $\lambda_1, \lambda_2 > 0$ とし, $X_1 \sim \text{Pois}(\lambda_1)$, $X_2 \sim \text{Pois}(\lambda_2)$ は独立とする. $Y = X_1 + X_2$ としたとき

$$m_Y(t) = m_{X_1}(t)m_{X_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

より $Y \sim \text{Pois}(\lambda_1 + \lambda_2)$ となる. \square

第 C 章 経験分布関数と統計的汎関数

経験分布関数の定義とその性質を述べる。証明はしない。さらに、統計的汎関数を導入する。

1 経験分布関数

\mathbb{R} 上の実数値関数 F を CDF とし、 $X_1, X_2, \dots, X_n \sim F$ とする。

df:c-1 定義 C.1. 経験分布関数 \hat{F}_n を各データ X_j ($j = 1, 2, \dots, n$) 上に確率 $1/n$ をのせた分布の CDF とする。すなわち、

$$\hat{F}_n(x) = \frac{\sum_{j=1}^n \mathbb{1}_{[X_j, \infty)}(x)}{n}.$$

ただし、

$$\mathbb{1}_{[a, \infty)}(x) = \begin{cases} 1 & (x \geq a) \\ 0 & (x < a). \end{cases}$$

thm:C-2 定理 C.2. 固定した x において、

$$\mathbb{E}[\hat{F}_n(x)] = F(x),$$

$$\mathbb{V}[\hat{F}_n(x)] = \frac{F(x)(1 - F(x))}{n},$$

$$\hat{F}_n(x) \xrightarrow{P} F(x).$$

すなわち、 $\forall \epsilon > 0$ に対して、 $\lim_{n \rightarrow \infty} \mathbb{P}(|\hat{F}_n(x) - F(x)| > \epsilon) = 0$ となる。

Proof. 最初のふたつの主張の証明はやさしい。最後の主張の証明は Chebyshev の不等式を用いればよい。□

定理 C.3. (Glivenko-Cantelli の定理) $X_1, X_2, \dots, X_n \sim F$ とする. このとき,

$$\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| \xrightarrow{P} 0.$$

Proof. 難しいので, 省略. □

定理 C.4. (Dvoretzky-Kiefer-Wolfowitz の定理) $X_1, X_2, \dots, X_n \sim F$ とする. このとき, $\forall \epsilon > 0$ に対して,

$$\mathbb{P}(\sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F(x)| > \epsilon) \leq 2 \exp^{-2n\epsilon^2}.$$

Proof. 難しいので, 省略. □

2 統計的汎関数

F を CDF とする. F の汎関数 $T(F)$ を統計的汎関数という. F が導関数 f をもつとき, 平均や分散

$$\mu = \int x f(x) dx, \quad \sigma^2 = \int (x - \mu)^2 f(x) dx$$

は統計的汎関数である. また, F が $\{x \in \mathbb{R}; F(x) > 0\}$ 上で狭義単調増加のとき, 中央値

$$m = F^{-1}(1/2)$$

も統計的汎関数である.

定義 C.5. $\theta := T(F)$ の差し込み推定量 $\widehat{\theta}_n$ を

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

で定義する.

df:C-6 定義 C.6. F を分布関数とし, $X \sim F$ とする. ある可積分な関数 $r(x)$ に対して,

$$T(F) = \mathbb{E}[r(X)]$$

とかけるとき, T を線型汎関数という.

re:C-7 注意 C.7. T が線型汎関数ならば, F, G を CDF, $a, b \in \mathbb{R}$ に対して,

$$T(aG + bF) = aT(G) + bT(F)$$

となることに注意せよ. ただし, 線型汎関数を定める関数 r は F, G の分布に関して可積分としている. \square

thm:C-8 定理 C.8. F を分布関数とし, $X \sim F$ とする. 線型汎関数 $T(F) = \mathbb{E}[r(X)]$ に対する差し込み推定量は

$$T(\widehat{F}_n) = \frac{1}{n} \sum_{j=1}^n r(X_j)$$

となる.

Proof. Lebesgue-Stieltes 積分の知識があると明らかである. 難しいので, 省略. \square

re:C-9 注意 C.9. $r(x) = x$ とすれば,

$$T(F) = \mu$$

となる. 定理 ^{thm:C-8}C.8 より, $\mu = T(F)$ の差し込み推定量は

$$T(\widehat{F}_n) = \frac{1}{n} \sum_{j=1}^n X_j$$

となる. $r(x) = (x - \mathbb{E}[X])^2$ とすれば, $\sigma^2 = T(F)$ の差し込み推定量は

$$\begin{aligned} T(\widehat{F}_n) &= \frac{1}{n} \sum_{j=1}^n \left\{ X_j - \frac{1}{n} \sum_{k=1}^n X_k \right\}^2 \\ &= \frac{1}{n} \sum_{j=1}^n X_j^2 - \left(\frac{1}{n} \sum_{j=1}^n X_j \right)^2 \\ &= \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^2 \end{aligned}$$

となる. ただし, $\bar{X}_n = (1/n) \sum_{j=1}^n X_j$ である. □

第 D 章 Hilbert 空間

1 Hilbert 空間の定義

df:d-1 定義 D.1. \mathcal{H} をベクトル空間とする. 関数 $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$ が任意の $x, y, z \in \mathcal{H}$ と $a \in \mathbb{R}$ に対し次の条件 (1) ~ (3) をみたすとき, $\langle \cdot, \cdot \rangle$ は内積であるという.

- (1) $\langle x, x \rangle \geq 0$. 更に, $\langle x, x \rangle = 0 \iff x = \mathbf{0}$. ただし, $\mathbf{0}$ は \mathcal{H} の零ベクトル.
- (2) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.
- (3) $\langle x, y \rangle = \langle y, x \rangle$.

ベクトル空間 \mathcal{H} と内積 $\langle \cdot, \cdot \rangle$ の組 $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ を内積空間という.

定義 D.2. $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ を内積空間とする. 任意の $x \in \mathcal{H}$ に対し

$$\|x\| = \sqrt{\langle x, x \rangle}$$

とし, $\|\cdot\|$ を x のノルムという.

thm:d-2 定理 D.3. $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ を内積空間とする. $x, y \in \mathcal{H}$ と $a \in \mathbb{R}$ に対し, 次が成立する.

- (4) $\|x\| \geq 0$.
- (5) $\|x\| = 0 \iff x = \mathbf{0}$.
- (6) $\|ax\| = |a| \|x\|$.
- (7) (三角不等式): $\|x + y\| \leq \|x\| + \|y\|$.
- (8) (Cauchy-Schwarz の不等式): $|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$.
- (9) (中線定理): $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$.

df:d-4 定義 D.4. 内積空間 $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ がノルム $\|\cdot\|$ に関して完備であるとき, \mathcal{H} を

Hilbert 空間という。すなわち,

$$\|x_m - x_n\| \rightarrow 0 \quad (m, n \rightarrow \infty)$$

をみたす \mathcal{H} 内の列 $\{x_n\}_{n=1}^{\infty}$ を Cauchy 列といい, \mathcal{H} の任意の Cauchy 列に対し

$$\|x_n - x\| \rightarrow 0 \quad (n \rightarrow \infty)$$

をみたす $x \in \mathcal{H}$ が存在する.

定義 D.5. ベクトル空間の部分集合 C が $\forall x, y \in C$ と $0 < \forall t < 1$ に対し

$$(1-t)x + ty \in C$$

をみたすとき, C は凸集合であるという.

定理 D.6. $C \neq \emptyset$ は Hilbert 空間 $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ の閉凸集合¹で, $x \in \mathcal{H}$ であるとき, $x_0 \in C$ で

$$\|x - x_0\| = \inf_{y \in C} \|x - y\|$$

をみたすものが唯一存在する. x_0 を x からの C への射影という. C への射影を P_C と書くことにする. すなわち,

$$P_C : \mathcal{H} \rightarrow C, \quad x \mapsto x_0$$

である. 射影の定義から $\forall c \in C$ に対し $P_C(c) = c$ となっていることに注意せよ.

Proof. $a = \inf_{y \in C} \|x - y\|$ とおき, C の点列 $\{y_n\}_{n=1}^{\infty}$ で $\{\|x - y_n\|\}_{n=1}^{\infty}$ が a に収束するものをとる. このとき中線定理より

$$\|x - y_m + x - y_n\|^2 + \|(x - y_m) - (x - y_n)\|^2 = 2\|x - y_m\|^2 + 2\|x - y_n\|^2 \quad (\text{D.1})$$

¹ C は定義 D.5 の条件をみたし, さらに, C は閉集合であるので, $x_n \in C (n \in \mathbb{N})$ に対し, $\|x_n - x\| \rightarrow 0 (n \rightarrow \infty)$ ならば $x \in C$ が成り立つ.

なので,

$$\begin{aligned}\|\mathbf{y}_m - \mathbf{y}_n\|^2 &= 2\|\mathbf{x} - \mathbf{y}_m\|^2 + 2\|\mathbf{x} - \mathbf{y}_n\|^2 - 4\left\|\mathbf{x} - \frac{1}{2}(\mathbf{y}_m + \mathbf{y}_n)\right\|^2 \\ &\leq 2\|\mathbf{x} - \mathbf{y}_m\|^2 + 2\|\mathbf{x} - \mathbf{y}_n\|^2 - 4a^2 \\ &\rightarrow 0 \quad (n, m \rightarrow \infty)\end{aligned}$$

となり², $\{\mathbf{y}_n\}_{n=1}^{\infty}$ は Cauchy 列である. \mathcal{H} は完備で C は閉集合なので, $\{\mathbf{y}_n\}_{n=1}^{\infty}$ は C 内で収束し, その極限を x_0 とおくと $\|\mathbf{x} - x_0\| = a$ が成り立つ. $x'_0 \in C$ も同じ条件をみたすものとする, 再び中線定理より

$$\|x_0 - x'_0\|^2 = 2\|\mathbf{x} - x_0\|^2 + 2\|\mathbf{x} - x'_0\|^2 - 4\left\|\mathbf{x} - \frac{1}{2}(x_0 + x'_0)\right\|^2$$

となり, $x_0 = x'_0$ となる. □

以下では, \mathcal{H} は, その内積が $\langle \cdot, \cdot \rangle$ の Hilbert 空間とする.

df:d-7 定義 D.7. $x, \mathbf{y} \in \mathcal{H}$ が $\langle x, \mathbf{y} \rangle = 0$ をみたすとき, x と \mathbf{y} は直交するといい, $x \perp \mathbf{y}$ と書くことにする. $S \subset \mathcal{H}$ に対し

$$S^\perp := \{x \in \mathcal{H} : \langle x, \mathbf{y} \rangle = 0 (\mathbf{y} \in S)\}$$

とおき, S の直交補空間という.

定義の直接的な結果を列挙しておく.

- $x \perp \mathbf{y}$ のとき,

$$\|x + \mathbf{y}\|^2 = \|x\|^2 + \|\mathbf{y}\|^2. \quad (\text{Pythagoras の定理})$$

- S, S_1, S_2 を \mathcal{H} の部分集合とする. このとき, $(S^\perp)^\perp \supset S$. $S_1 \subset S_2 \implies S_2^\perp \subset S_1^\perp$.

² C は凸なので, $(1/2)(\mathbf{y}_m + \mathbf{y}_n) \in C$ であるので,

$$4\left\|\mathbf{x} - \frac{1}{2}(\mathbf{y}_m + \mathbf{y}_n)\right\|^2 \geq 4a^2$$

となることに注意せよ.

- 内積は連続であり,

$$S^\perp = \bigcap_{\mathbf{y} \in S} \{\mathbf{x} \in \mathcal{H}; \langle \mathbf{x}, \mathbf{y} \rangle = 0\}$$

なので, S^\perp は \mathcal{H} の閉集合.

pro:d-8

命題 D.8. \mathcal{K} を Hilbert 空間 \mathcal{H} の閉部分空間³であるとき,

- (1) $\mathcal{H} = \mathcal{K} \oplus \mathcal{K}^\perp$.
- (2) $(\mathcal{K}^\perp)^\perp = \mathcal{K}$.

Proof. 泉 (2021, p.19) を参照のこと. □

ここで直交射影について書く.

Hilbert 空間 \mathcal{H} 上で定義された写像 $\varphi: \mathcal{H} \rightarrow \mathbb{R}$ が

$$\varphi(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha \varphi(\mathbf{x}) + \beta \varphi(\mathbf{y}) \quad (\mathbf{x}, \mathbf{y} \in \mathcal{H}, \alpha, \beta \in \mathbb{R})$$

をみたすとき, φ は \mathcal{H} 上の線型汎関数という. \mathcal{H} 上の線型汎関数全体の集合を \mathcal{H}^* と書くことにする. $\varphi \in \mathcal{H}^*$ のノルム⁴を

$$\|\varphi\| = \sup_{\mathbf{x} \in \mathcal{H}; \|\mathbf{x}\| \leq 1} |\varphi(\mathbf{x})|$$

と定める. この量が有限のとき, φ は有界という.

pro:d-9

命題 D.9. $\varphi \in \mathcal{H}^*$ とする. このとき, 次の 2 条件は同値.

- (1) φ は有界.
- (2) φ は連続. すなわち, $\mathbf{x}, \mathbf{x}_n \in \mathcal{H} (n \in \mathbb{N})$ に対し, $\mathbf{x}_n \rightarrow \mathbf{x} (n \rightarrow \infty)$ ⁵ならば, $\varphi(\mathbf{x}_n) \rightarrow \varphi(\mathbf{x}) (n \rightarrow \infty)$ が成り立つ.

³ $\forall \mathbf{x}, \mathbf{y} \in \mathcal{K}$ と $\forall \alpha, \beta \in \mathbb{R}$ に対し $\alpha \mathbf{x} + \beta \mathbf{y} \in \mathcal{K}$.

⁴ \mathcal{H} と \mathcal{H}^* 上のノルム (みかけの定義では異なるもの) を同じ記号を用いて表記する理由はあとでわかる.

⁵ $\|\mathbf{x}_n - \mathbf{x}\| \rightarrow 0 (n \rightarrow \infty)$ が成立していること.

Proof. 瀬戸 (2021, p.56) を参照. □

$y \in \mathcal{H}$ に対し $\varphi_y(x) = \langle x, y \rangle$ と定めると, φ_y は \mathcal{H} の線型汎関数であり, Cauchy-Schwarz の不等式より, φ_y は有界であるので, $\varphi_y \in \mathcal{H}^*$ となる. Hilbert 空間では, その逆も成り立ち, \mathcal{H} と \mathcal{H}^* の元の間に対応関係がつく.

hm:d-10 定理 D.10. (Riesz の表現定理) \mathcal{H} を Hilbert 空間とすると, $\forall \varphi \in \mathcal{H}^*$ に対して, $x_\varphi \in \mathcal{H}$ で, $\forall x \in \mathcal{H}$ に対し

$$\varphi(x) = \langle x, x_\varphi \rangle$$

をみたすものが唯一存在する. さらに

$$\|\varphi\| = \|x_\varphi\|$$

が成立する.

Proof. 泉 (2021, p.20) または瀬戸 (2021, p.57) を参照. □

df:d-11 定義 D.11. $\mathcal{X} (\neq \emptyset)$ を集合とし, 以下の 2 条件をみたす Hilbert 空間 \mathcal{H} を考える.

(1) \mathcal{H} は \mathcal{X} 上の関数からなる Hilbert 空間である.

(2) $\forall x \in \mathcal{X}$ と $f, f_n \in \mathcal{H} (n \in \mathbb{N})$ に対し

$$\|f_n - f\| \longrightarrow 0 (n \rightarrow \infty) \implies f_n(x) \longrightarrow f(x) (n \rightarrow \infty)$$

である.

このような Hilbert 空間を再生核 Hilbert 空間という.

ここで, $x \in \mathcal{X}$ に対し

$$\varphi_x(f) = f(x) \quad (f \in \mathcal{H})$$

とさだめると, φ_x は線型汎関数になる. このことに注意すれば, Riesz の表現定理より, (2) は次の (2)' と同値になる.

(2)' $\forall x \in \mathcal{X}$ と $f \in \mathcal{H}$ に対し

$$f(x) = \langle f, k_x \rangle$$

となる $k_x \in \mathcal{H}$ が唯一存在する. すなわち, 代入が内積で表現される. この k_x を再生核という. さらに, $\mathcal{X} \times \mathcal{X}$ 上の関数 k を

$$k_x(y) = k(x, y) \quad (x, y \in \mathcal{X})$$

と書いたとき, k を \mathcal{H} の核関数といい, この再生核 Hilbert 空間を \mathcal{H}_k と書くことにする.

(2) と (2)' は同値であることは以下のように考えればよい. (2) は汎関数 φ_x が連続であることを言っている. したがって, 汎関数 φ_x について

$$\varphi_x \text{ は連続} \iff \varphi_x \text{ は有界}$$

である. 斉藤 (2002) の定理 2.2.1(page 30) により

$$\varphi_x \text{ は有界} \iff \mathcal{X} \text{ 上の関数からなる Hilbert 空間 } \mathcal{K} \text{ において} \\ \text{核関数 } k(x, y) \text{ (} x, y \in \mathcal{X} \text{) が存在}$$

がわかる.

第 E 章 線型代数

ch:e

1 特異値分解

e-e1

定義 E.1. $n, p \geq \mathbb{N}$, $n \geq p$ とし, $V \subset \mathbb{R}^n$ をベクトル空間とする. $v_1, v_2, \dots, v_p \in V$ とする.

(1) v_1, v_2, \dots, v_p はベクトル空間 V の基底であるとは, 任意の $v \in V$ に対して, 唯一の $a_1, a_2, \dots, a_p \in \mathbb{R}$ があって,

$$v = a_1 v_1 + a_2 v_2 + \dots + a_p v_p$$

とかけるときをいう. この p をベクトル空間 V の次元といい, $\dim(V)$ と書く.

(2) v_1, v_2, \dots, v_p は線型独立 (1 次独立) であるとは,

$$a_1 v_1 + a_2 v_2 + \dots + a_p v_p = \mathbf{0} \quad (a_1, a_2, \dots, a_p \in \mathbb{R}) \implies a_1 = a_2 = \dots = a_p = 0$$

が成り立つときをいう. ただし, $\mathbf{0} = (0, 0, \dots, 0)^\top \in \mathbb{R}^n$ である.

df:e-0

定義 E.2. (1) $n, p \in \mathbb{N}$, $n \geq p$ とする. ベクトル $x_1, x_2, \dots, x_p \in \mathbb{R}^n$ で張られた \mathbb{R}^n の部分空間 $\text{span}\{x_1, x_2, \dots, x_p\}$ を

$$\text{span}\{x_1, x_2, \dots, x_p\} = \{a_1 x_1 + a_2 x_2 + \dots + a_p x_p; a_1, a_2, \dots, a_p \in \mathbb{R}\}$$

と定義する.

(2) $m, n \in \mathbb{N}$ とし, $A \in \text{Mat}(m, n; \mathbb{R})$ とする. A と A^\top の像と核をそれぞれ

$$R(A) = \{Ax; x \in \mathbb{R}^n\} \subset \mathbb{R}^m,$$

$$\ker(A) = \{x; Ax = \mathbf{0} \in \mathbb{R}^m\} \subset \mathbb{R}^n,$$

$$R(A^\top) = \{A^\top y; y \in \mathbb{R}^m\} \subset \mathbb{R}^n,$$

$$\ker(A^\top) = \{y; A^\top y = \mathbf{0} \in \mathbb{R}^m\} \subset \mathbb{R}^m$$

で定める.

df:e-e2

定義 E.3. $n \in \mathbb{N}$ とする. 行列式は関数 $\det : \text{Mat}(n; \mathbb{R}) \rightarrow \mathbb{R}$ で以下のように定義する.

(1) 1×1 の行列 (a) の行列式を

$$\det(a) = a.$$

(2) $n \in \mathbb{N}$, $n \geq 2$ とき, 行列

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \in \text{Mat}(n; \mathbb{R})$$

の行列式を以下のように定める. $i, j = 1, 2, \dots, n$ に対し \mathbf{A}_{ij} を \mathbf{A} から i 行と j 列をとった $(n-1) \times (n-1)$ 行列とする. $n \times n$ 行列 \mathbf{A} の行列式を

$$\det(\mathbf{A}) = \sum_{k=1}^n (-1)^{k+1} a_{1k} \det(\mathbf{A}_{1k})$$

で定める.

pro:e-e3

命題 E.4. $n \times n$ の行列 $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \in \text{Mat}(n; \mathbb{R})$ に対して, 行列式は以下の性質をみたく.

(1) $k = 1, 2, \dots, n$ と $d \in \mathbb{R}$ に対し,

$$\det(\mathbf{a}_1, \dots, \lambda \mathbf{a}_k, \dots, \mathbf{a}_n) = \lambda \det(\mathbf{a}_1, \dots, \mathbf{a}_k, \dots, \mathbf{a}_n).$$

(2) $j \neq k$ に対し,

$$\det(\mathbf{a}_1, \dots, \mathbf{a}_k + \lambda \mathbf{a}_j, \dots, \mathbf{a}_n) = \det(\mathbf{a}_1, \dots, \mathbf{a}_k, \dots, \mathbf{a}_n).$$

(3) $\det(\mathbf{I}_n) = 1$.

(4) $\mathbf{A}, \mathbf{B} \in \text{Mat}(n; \mathbb{R})$ に対し

$$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B}).$$

df:e-e4

定義 E.5. $A \in \text{Mat}(n; \mathbb{R})$ に対し, ある $B \in \text{Mat}(n; \mathbb{R})$ が存在し,

$$AB = BA = I_n$$

が成り立つとき, A は可逆であるといい, B を A の逆行列といい, これを A^{-1} と記す.

eq:e-e6

命題 E.6. $A \in \text{Mat}(n; \mathbb{R})$ とする. このとき

$$A \text{ は可逆} \iff \det(A) \neq 0.$$

thm:e-1

定理 E.7. (特異値分解) $m, n \in \mathbb{N}$ とする. 任意の $A \in \text{Mat}(m, n; \mathbb{R})$ は次の分解をもつ.

$$A = UDV^{\top}.$$

ただし $U \in \text{Mat}(m; \mathbb{R})$, $V \in \text{Mat}(n; \mathbb{R})$ は直交行列で, $D \in \text{Mat}(m, n; \mathbb{R})$ は対角行列で, その対角成分は $d_1 \geq d_2 \geq \cdots \geq d_{\min(m, n)} \geq 0$ である. d_j ($j = 1, 2, \dots, \min(m, n)$) を行列 A の特異値といい, $\{(d_j, \mathbf{u}_j, \mathbf{v}_j); j = 1, 2, \dots, \min(m, n)\}$ を特異値系¹という.

re:e-2

注意 E.8. $n \geq m$ のとき

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_m & 0 & \cdots & 0 \end{pmatrix}$$

¹ $j = 1, 2, \dots, \min(m, n)$ に対して

$$A\mathbf{v}_j = d_j\mathbf{u}_j$$

となっている.

である. $m \geq n$ のとき,

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \\ 0 & 0 & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

である. □

thm:e-1
定理 E.7 の証明 $\mathbf{y} \in \mathbb{R}^m$ に対し,

$$|\mathbf{y}|_2 = \sqrt{\mathbf{y}^\top \mathbf{y}}$$

とする. ただし, 次元の異なる Euclid 空間のノルムも同じ記号を流用する. 行列 A のノルム

$$\|A\| := \sup_{\mathbf{x} \in \mathbb{R}^n} \frac{|A\mathbf{x}|_2}{|\mathbf{x}|_2}$$

で定義すると

$$d_1 := \|A\| := \sup_{\mathbf{x} \in \mathbb{R}^n, |\mathbf{x}|_2=1} |A\mathbf{x}|_2$$

となる. $d_1 \neq 0$ と仮定する. そうでなければ, 定理の証明は自明となる. あるベクトル $\mathbf{x} \in \mathbb{R}^n$ ($|\mathbf{x}|_2 = 1$) が存在して

$$|A\mathbf{x}|_2 = d_1$$

をみたすとする.

$$\mathbf{y} = \frac{1}{d_1} A\mathbf{x} \in \mathbb{R}^m$$

とおけば

$$|\mathbf{y}|_2^2 = \mathbf{y}^\top \mathbf{y} = \frac{1}{d_1^2} \mathbf{x}^\top A^\top A \mathbf{x} = \frac{|A\mathbf{x}|_2^2}{d_1^2} = 1$$

より $\|y\|_2 = 1$ となる. $v_2, v_3, \dots, v_n \in \mathbb{R}^n$ と $u_2, u_3, \dots, u_m \in \mathbb{R}^m$ をうまく選んで, $\{x, v_2, \dots, v_n\}$ と $\{y, u_2, \dots, u_m\}$ はそれぞれ \mathbb{R}^n と \mathbb{R}^m の正規直交基底となるようにする.

$$U_1 = (y, u_2, \dots, u_m) \in \text{Mat}(m; \mathbb{R}), \quad V_1 = (x, v_2, \dots, v_n) \in \text{Mat}(n; \mathbb{R}),$$

$$A_1 = U_1^\top A V_1$$

とおけば,

$$\begin{aligned} A_1 = U_1^\top A V_1 &= \begin{bmatrix} y^\top \\ u_2^\top \\ \vdots \\ u_m^\top \end{bmatrix} A [x, v_2, \dots, v_n] = \begin{bmatrix} y^\top \\ u_2^\top \\ \vdots \\ u_m^\top \end{bmatrix} [Ax, Av_2, \dots, Av_n] \\ &= \begin{bmatrix} y^\top \\ u_2^\top \\ \vdots \\ u_m^\top \end{bmatrix} [d_1 y, Av_2, \dots, Av_n] = \begin{bmatrix} d_1 y^\top y & y^\top Av_2 & \dots & y^\top Av_n \\ 0 & u_2^\top Av_2 & \dots & u_2^\top Av_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & u_m^\top Av_2 & \dots & u_m^\top Av_n \end{bmatrix} \\ &=: \begin{bmatrix} d_1 & w^\top \\ 0 & B \end{bmatrix} \end{aligned}$$

と書き直せる. ただし

$$w = \begin{bmatrix} y^\top Av_2 \\ y^\top Av_3 \\ \vdots \\ y^\top Av_n \end{bmatrix} \in \mathbb{R}^{n-1},$$

$$B = \begin{bmatrix} u_2^\top Av_2 & u_2^\top Av_3 & \dots & u_2^\top Av_n \\ u_3^\top Av_2 & u_3^\top Av_3 & \dots & u_3^\top Av_n \\ \vdots & \vdots & \ddots & \vdots \\ u_m^\top Av_2 & u_m^\top Av_3 & \dots & u_m^\top Av_n \end{bmatrix} \in \text{Mat}(m-1, n-1; \mathbb{R})$$

である. 上記の形に注意して計算すると

$$A_1 \begin{bmatrix} d_1 \\ w \end{bmatrix} = \begin{bmatrix} d_1 & w^\top \\ 0 & B \end{bmatrix} \begin{bmatrix} d_1 \\ w \end{bmatrix} = \begin{bmatrix} d_1^2 + w^\top w \\ Bw \end{bmatrix}$$

となる。したがって

$$\left\| \mathbf{A}_1 \begin{bmatrix} d_1 \\ \mathbf{w} \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} d_1^2 + \mathbf{w}^\top \mathbf{w} \\ \mathbf{B}\mathbf{w} \end{bmatrix} \right\|_2^2 = \{d_1^2 + \mathbf{w}^\top \mathbf{w}\}^2 + \mathbf{w}^\top \mathbf{B}^\top \mathbf{B} \mathbf{w} \geq \{d_1^2 + |\mathbf{w}|_2^2\}^2$$

から

$$\left\| \mathbf{A}_1 \begin{bmatrix} d_1 \\ \mathbf{w} \end{bmatrix} \right\|_2 \geq d_1^2 + |\mathbf{w}|_2^2, \quad \left\| \frac{1}{\sqrt{d_1^2 + |\mathbf{w}|_2^2}} \begin{bmatrix} d_1 \\ \mathbf{w} \end{bmatrix} \right\|_2 = 1$$

がわかる。よって

$$\|\mathbf{A}_1\| = \sup_{\mathbf{x} \in \mathbb{R}^n, |\mathbf{x}|_2=1} |\mathbf{A}_1 \mathbf{x}|_2 \geq \frac{1}{\sqrt{d_1^2 + |\mathbf{w}|_2^2}} \left\| \mathbf{A}_1 \begin{bmatrix} d_1 \\ \mathbf{w} \end{bmatrix} \right\|_2 \geq \sqrt{d_1^2 + |\mathbf{w}|_2^2}. \quad (\text{E.1})$$

eq: sig-

直交変換に関して $\|\cdot\|$ は不変²なので、

$$d_1 = \|\mathbf{A}\| = \|\mathbf{A}_1\|.$$

である。これと (E.1) と合わせると

$$d_1 = \|\mathbf{A}\| = \|\mathbf{A}_1\| \geq \sqrt{d_1^2 + |\mathbf{w}|_2^2} \geq d_1$$

となるので、

$$|\mathbf{w}|_2 = 0 \iff \mathbf{w} = \mathbf{0}.$$

よって

$$\mathbf{U}_1^\top \mathbf{A} \mathbf{V}_1 = \mathbf{A}_1 = \begin{bmatrix} d_1 & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{B} \end{bmatrix}$$

と書けることがわかった。

² $\forall \mathbf{x} \in \mathbb{R}^n$ ($|\mathbf{x}|_2 = 1$) に対して $|\mathbf{V}_1 \mathbf{x}|_2 = 1$ に注意すれば、

$$\begin{aligned} \|\mathbf{A}\| &= \|\mathbf{U}_1 \mathbf{A}_1 \mathbf{V}_1\| = \sup_{\mathbf{x} \in \mathbb{R}^n; |\mathbf{x}|_2=1} |\mathbf{U}_1 \mathbf{A}_1 \mathbf{V}_1 \mathbf{x}|_2 = \sup_{\mathbf{x} \in \mathbb{R}^n; |\mathbf{x}|_2=1} |\mathbf{U}_1 \mathbf{A}_1 \mathbf{x}|_2 \\ &= \sup_{\mathbf{x} \in \mathbb{R}^n; |\mathbf{x}|_2=1} \sqrt{\mathbf{x}^\top \mathbf{A}_1^\top \mathbf{U}_1^\top \mathbf{U}_1 \mathbf{A}_1 \mathbf{x}} = \sup_{\mathbf{x} \in \mathbb{R}^n; |\mathbf{x}|_2=1} \sqrt{\mathbf{x}^\top \mathbf{A}_1^\top \mathbf{A}_1 \mathbf{x}} \\ &= \sup_{\mathbf{x} \in \mathbb{R}^n; |\mathbf{x}|_2=1} |\mathbf{A}_1 \mathbf{x}|_2 = \|\mathbf{A}_1\| \end{aligned}$$

からわかる。

次に, $d_2 = \|B\|$ とおけば,

$$\begin{aligned} d_2 &= \sup_{\mathbf{x} \in \mathbb{R}^{n-1}; |\mathbf{x}|_2=1} |\mathbf{B}\mathbf{x}|_2 = \sup_{\mathbf{x} \in \mathbb{R}^{n-1}; |\mathbf{x}|_2=1} \left\| \begin{bmatrix} d_1 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{B} \end{bmatrix} \begin{bmatrix} 0 \\ \mathbf{x} \end{bmatrix} \right\|_2 \\ &\leq \sup_{\mathbf{z} \in \mathbb{R}^n; |\mathbf{z}|_2=1} |\mathbf{A}_1 \mathbf{z}|_2 = \|\mathbf{A}_1\| = \|\mathbf{A}\| = d_1. \end{aligned}$$

$d_2 = 0$ のとき, $B = \mathbf{0}$ より証明はおわり. $d_2 > 0$ と仮定して議論を進める. 前と同じようにすれば,

$$d_2 = \|B\|$$

とおき,

$$|\mathbf{B}\mathbf{x}|_2 = 1, \mathbf{x} \in \mathbb{R}^{n-1} \text{ s.t. } |\mathbf{x}|_2 = 1, \mathbf{y} = \frac{1}{d_2} \mathbf{B}\mathbf{x} \in \mathbb{R}^{m-1}$$

と取り, さきほどの議論を繰り返すと $\tilde{U}_2 \in \text{Mat}(m-1; \mathbb{R})$, $\tilde{V}_2 \in \text{Mat}(n-1; \mathbb{R})$ で

$$\tilde{U}_2^\top \mathbf{B} \tilde{V}_2 = \begin{bmatrix} d_2 & \mathbf{0}^\top \\ \mathbf{0} & \mathbf{C} \end{bmatrix}, \mathbf{C} \in \text{Mat}(m-2, n-2; \mathbb{R})$$

とできて,

$$U_2 = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{U}_2 \end{bmatrix}, V_2 = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{V}_2 \end{bmatrix}$$

とおけば, U_2, V_2 は直交行列で

$$U_2^\top U_1^\top \mathbf{A} V_1 V_2 = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & \mathbf{C} \end{bmatrix}$$

となる. さらに $U_1 U_2 \in \text{Mat}(m; \mathbb{R})$ と $V_1 V_2 \in \text{Mat}(n; \mathbb{R})$ も直交行列であることに注意する.

以上の操作を繰り返せば, 定理は証明される. □

thm:e-2 定理 E.9. $A = UDV^\top$ と特異値分解されたとする. ただし, $p \leq \min(m, n)$ に対し

$$d_1 \geq d_2 \geq \cdots \geq d_p > d_{p+1} = d_{p+2} = \cdots = d_{\min(m, n)} = 0$$

とする.

$$U = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m), \quad V = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$$

としたとき,

$$\begin{aligned} \ker(\mathbf{A}) &= \text{span}\{\mathbf{v}_{p+1}, \mathbf{v}_{p+2}, \dots, \mathbf{v}_n\} = \mathbf{R}(\mathbf{A})^\perp, \\ \ker(\mathbf{A}^\top) &= \text{span}\{\mathbf{u}_{p+1}, \mathbf{u}_{p+2}, \dots, \mathbf{u}_m\} = \mathbf{R}(\mathbf{A}^\top)^\perp. \end{aligned}$$

Proof. U の直交性より

$$\mathbf{A}\mathbf{x} = \mathbf{0} \quad (\mathbf{x} \in \mathbb{R}^n) \iff \mathbf{D}\mathbf{V}^\top \mathbf{x} = \begin{bmatrix} d_1 \mathbf{v}_1^\top \mathbf{x} \\ d_2 \mathbf{v}_2^\top \mathbf{x} \\ \vdots \\ d_p \mathbf{v}_p^\top \mathbf{x} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}.$$

よって $\ker(\mathbf{A})$ の任意の元 \mathbf{x} は $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p$ と直交するので,

$$\ker(\mathbf{A}) = \text{span}\{\mathbf{v}_{p+1}, \mathbf{v}_{p+2}, \dots, \mathbf{v}_n\}.$$

さらに $\mathbf{A}^\top = \mathbf{V}\mathbf{D}^\top\mathbf{U}^\top$ より

$$\ker(\mathbf{A}^\top) = \text{span}\{\mathbf{u}_{p+1}, \mathbf{u}_{p+2}, \dots, \mathbf{u}_m\} = \mathbf{R}(\mathbf{A})^\perp.$$

一方, $\forall \tilde{\mathbf{x}} \in \mathbf{R}(\mathbf{A}^\top)$ とすれば, ある $\mathbf{y} \in \mathbb{R}^m$ があって, $\tilde{\mathbf{x}} = \mathbf{A}^\top \mathbf{y}$ とかける. このことより,

$$\mathbf{x} \in \ker(\mathbf{A}) \iff \mathbf{x}^\top \tilde{\mathbf{x}} = (\mathbf{A}\mathbf{x})^\top \mathbf{y} = 0$$

より, \mathbf{x} は $\mathbf{R}(\mathbf{A}^\top)$ と直交する. 同様に, $\forall \tilde{\mathbf{y}} \in \mathbf{R}(\mathbf{A})$ をとる. ある $\mathbf{x} \in \mathbb{R}^n$ があって, $\tilde{\mathbf{y}} = \mathbf{A}\mathbf{x}$ とかける. このことより

$$\mathbf{y} \in \ker(\mathbf{A}^\top) \iff \mathbf{y}^\top \tilde{\mathbf{y}} = (\mathbf{A}^\top \mathbf{y})^\top \mathbf{x} = 0$$

より, \mathbf{y} は $\mathbf{R}(\mathbf{A})$ と直交することもわかる. □

re:e-4 注意 E.10.

$$U_1 = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p], U_2 = [\mathbf{u}_{p+1}, \mathbf{u}_{p+2}, \dots, \mathbf{u}_m], \\ V_1 = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_p], V_2 = [\mathbf{v}_{p+1}, \mathbf{v}_{p+2}, \dots, \mathbf{v}_n]$$

とおく. すなわち

$$U = [U_1, U_2], \quad V = [V_1, V_2]$$

である. このとき, \mathbb{R}^n から $\ker(A)$ への直交射影を P , \mathbb{R}^m から $\ker(A^\top)$ への直交射影を Q とおくと

$$P = V_2 V_2^\top, \quad Q = U_2 U_2^\top$$

となる. □

2 Moore-Penrose の一般化逆行列

$A \in \text{Mat}(m, n; \mathbb{R})$ で $A \neq \mathbf{0}$ とする. 定理 ^{thm:e-1}E.7 から $1 \leq p \leq \min(m, n)$ があって

$$A = \tilde{U} \tilde{D} \tilde{V}^\top,$$

$\tilde{U} \in \text{Mat}(m; \mathbb{R})$, $\tilde{V} \in \text{Mat}(n; \mathbb{R})$ は直交行列,

$\tilde{D} \in \text{Mat}(m, n; \mathbb{R})$ は対角行列 (非対角成分は 0) で

その正の対角成分は $d_1 \geq d_2 \geq \dots \geq d_p > 0$

と書ける. ここで \tilde{U} , \tilde{D} , \tilde{V} を分解し

$$\tilde{U} = [U, U^\perp], \quad U \in \text{Mat}(m, p; \mathbb{R}),$$

$$\tilde{V} = [V, V^\perp], \quad V \in \text{Mat}(n, p; \mathbb{R}),$$

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{bmatrix} \in \text{Mat}(p; \mathbb{R})$$

とする. すると

$$\begin{aligned}\tilde{U}^\top \tilde{U} &= \begin{bmatrix} U^\top U & U^\top U^\perp \\ (U^\perp)^\top U & (U^\perp)^\top U^\perp \end{bmatrix} = I_m, \\ \tilde{V}^\top \tilde{V} &= \begin{bmatrix} V^\top V & V^\top V^\perp \\ (V^\perp)^\top V & (V^\perp)^\top V^\perp \end{bmatrix} = I_n\end{aligned}$$

から

$$U^\top U = I_p, U^\top U^\perp = \mathbf{0}, V^\top V = I_p, V^\top V^\perp = \mathbf{0}$$

となる. よって

$$A = \tilde{U} \tilde{D} \tilde{V}^\top = \begin{bmatrix} U & U^\perp \end{bmatrix} \begin{bmatrix} D & \mathbf{0}_{p \times (n-p)} \\ \mathbf{0}_{(m-p) \times n} \end{bmatrix} \begin{bmatrix} V^\top \\ (V^\perp)^\top \end{bmatrix} = U D V^\top \quad (\text{E.2})$$

eq:mp-2

となる. ただし $U \in \text{Mat}(m, p; \mathbb{R})$, $V \in \text{Mat}(n, p; \mathbb{R})$, $D \in \text{Mat}(p; \mathbb{R})$ で

$$U^\top U = V^\top V = I_p, D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{bmatrix}$$

となる.

(E.2) を踏まえて, $A^\dagger \in \text{Mat}(n, m; \mathbb{R})$ を

$$A^\dagger := V D^{-1} U^\top, D^{-1} = \begin{bmatrix} d_1^{-1} & 0 & \cdots & 0 \\ 0 & d_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p^{-1} \end{bmatrix} \quad (\text{E.3})$$

eq:mp-3

と定める. A^\dagger を A の Moore-Penrose の一般逆行列と呼ぶことにする.

すると A^\dagger は下記の関係式をみたす $n \times m$ の行列となる.

$$A A^\dagger A = A, \quad (\text{E.4})$$

eq:mp-4

$$A^\dagger A A^\dagger = A^\dagger, \quad (\text{E.5})$$

eq:mp-5

$$A A^\dagger = (A A^\dagger)^\top, \quad (\text{E.6})$$

eq:mp-6

$$A^\dagger A = (A^\dagger A)^\top \quad (\text{E.7})$$

eq:mp-7

が成立する. 上記 ^{eq:mp-4}(E.4)–^{eq:mp-7}(E.7) は ^{eq:mp-2}(E.2) と ^{eq:mp-3}(E.3) よりわかる. 実際

$$\begin{aligned} AA^\dagger A &= UDV^\top VD^{-1}U^\top UDV^\top = UDV^\top = A, \\ A^\dagger AA^\dagger &= VD^{-1}U^\top UDV^\top VD^{-1}U^\top = VD^{-1}U^\top = A^\dagger, \\ AA^\dagger &= UDV^\dagger VD^{-1}U^\dagger = UU^\top = (UU^\top)^\top = (AA^\dagger)^\top, \\ A^\dagger A &= VD^{-1}U^\top UDV^\top = VV^\top = (VV^\top)^\top = (A^\dagger A)^\top \end{aligned}$$

からわかる. さらに

$$\begin{aligned} P &= AA^\dagger : \mathbb{R}^m \longrightarrow \text{ran}(A \subset \mathbb{R}^m), \\ Q &= A^\dagger A : \mathbb{R}^n \longrightarrow \text{ran}(A^\top) \subset \mathbb{R}^n \end{aligned}$$

は射影行列になっている. すなわち $P \in \text{Mat}(m; \mathbb{R})$ と $Q \in \text{Mat}(n; \mathbb{R})$

$$P^2 = P, P^\top = P, Q^2 = Q, Q^\top = Q$$

をみたしている.

最後に A^\dagger の一意性を示す. B^\dagger も A の Moore-Penrose の一般逆行列とする. すると A^\dagger を B と替えた

$$ABA = A, \tag{E.8}$$

$$BAB = B, \tag{E.9}$$

$$AB = (AB)^\top, \tag{E.10}$$

$$BA = (BA)^\top \tag{E.11}$$

eq:mp-8

eq:mp-9

eq:mp-10

eq:mp-11

も成立する. すると

$$\begin{aligned}
 A^\dagger &= A^\dagger A A^\dagger && (\because (E.4)) \\
 &= A^\dagger (A B A) A^\dagger && (\because (E.8)) \\
 &= A^\dagger (A B)^\top (A A^\dagger)^\top && (\because (E.6) \text{ と } (E.10)) \\
 &= A^\dagger B^\top (A A^\dagger A)^\top \\
 &= A^\dagger B^\top A^\top && (\because (E.4)) \\
 &= A^\dagger (A B)^\top \\
 &= A^\dagger A B && (\because (E.10)) \\
 &= A^\dagger A B A B && (\because (E.8)) \\
 &= (A^\dagger A)^\top (B A)^\top B && (\because (E.7) \text{ と } (E.11)) \\
 &= (A A^\dagger A)^\top B^\top B \\
 &= A^\top B^\top B && (\because (E.4)) \\
 &= (B A)^\top B \\
 &= B A B && (\because (E.11)) \\
 &= B && (\because (E.8))
 \end{aligned}$$

からわかる. よって一意性を証明できた.

注意 E.11. 多くの本では, $A \in \text{Mat}(m, n; \mathbb{R})$ に対して $(E.4) = (E.7)$ をみだす $n \times m$ の行列 A^\dagger を A の Moore-Penrose の一般逆行列と定義している. \square

第 F 章 凸最適化

ch:f

1 凸関数とその性質

lf:con-1

定義 F.1. $a, b \in \mathbb{R}$, $a < b$ とする. 関数 $f: [a, b] \rightarrow \mathbb{R}$ が $x, y \in [a, b]$, $0 \leq \lambda \leq 1$ に対して

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

をみたすとき凸関数であるという. とくに $x \neq y$, $0 < \lambda < 1$ のとき

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y)$$

が成立する場合は狭義凸関数であるという.

le:con-2

注意 F.2. $-f$ が凸関数 (狭義凸関数) のとき, f は凹関数 (狭義凹関数) であるという. □

lm:con-3

定理 F.3. 区間 $[a, b]$ 上の凸関数は开区間 (a, b) 上の各点で連続である.

Proof. $\forall x \in [a, b]$ は

$$x = \lambda a + (1 - \lambda)b \quad (0 \leq \lambda \leq 1)$$

と表せるから凸関数の定義より

$$f(x) \leq \lambda f(a) + (1 - \lambda)f(b) \leq \max\{f(a), f(b)\} =: K.$$

したがって f は上に有界である. $\forall x_0 \in (a, b)$ とし, $f(x)$ は $x = x_0$ で連続であることを示すために関数 g を

$$g(t) := f(x_0 + t) - f(x_0) \quad (a - x_0 < t < b - x_0)$$

と定める. このとき $g(t)$ は $(a - x_0, b - x_0) \ni \{0\}$ 上で凸関数であり, $g(0) = 0$ である. いま $\forall \epsilon > 0$ に対して正数 ϵ' を

$$0 < \epsilon' < \min \left\{ \frac{\epsilon}{K + |f(x_0)|}, 1 \right\} \quad (\text{F.1})$$

をみたすように選ぶ. さらに正数 ρ と δ を

$$0 < \rho < \min\{x_0 - a, b - x_0\}, \quad 0 < \delta < \rho\epsilon'$$

をみたすように定める. $0 < \epsilon' \leq 1$ だから

$$|t| < \delta \implies |t| < \rho.$$

したがって $|t| < \delta$ に対して, $g(t)$ は定義されている. しかも

$$|t| < \delta \implies \left| \frac{t}{\epsilon'} \right| < \rho \implies x_0 + \frac{t}{\epsilon'} \in (a, b)$$

となり

$$g\left(\frac{t}{\epsilon'}\right) = f\left(x_0 + \frac{t}{\epsilon'}\right) - f(x_0) \leq \left| f\left(x_0 + \frac{t}{\epsilon'}\right) \right| + |f(x_0)| \leq K + |f(x_0)| =: K_1$$

を得る. ところで

$$t = (1 - \epsilon') \times 0 + \epsilon' \left(\frac{t}{\epsilon'} \right)$$

と表されることと (F.1) から

$$g(t) \leq (1 - \epsilon')g(0) + \epsilon'g\left(\frac{t}{\epsilon'}\right) \leq \epsilon'g\left(\frac{t}{\epsilon'}\right) \leq \epsilon'K_1 < \epsilon. \quad (\text{F.2})$$

他方

$$0 = \frac{\epsilon'}{1 + \epsilon'} \left(-\frac{t}{\epsilon'} \right) + \frac{1}{1 + \epsilon'} \cdot t$$

と表されるので

$$\begin{aligned} 0 \leq g(0) &\leq \frac{\epsilon'}{1 + \epsilon'} g\left(-\frac{t}{\epsilon'}\right) + \frac{1}{1 + \epsilon'} g(t) \leq \frac{\epsilon'K_1}{1 + \epsilon'} + \frac{1}{1 + \epsilon'} g(t) \\ &< \frac{\epsilon}{1 + \epsilon'} + \frac{1}{1 + \epsilon'} g(t) \end{aligned}$$

となり

$$-\epsilon < g(t) \tag{F.3}$$

を得る. よって (F.2) と (F.3) より

$$|t| < \delta \implies |g(t)| < \epsilon$$

である. $t = x - x_0$ と $g(t)$ の定義より

$$|x - x_0| < \delta \implies |f(x) - f(x_0)| < \epsilon$$

となり, $f(x)$ は点 x_0 で連続となる. □

am:con-4

定理 F.4. $\varphi: [a, b] \rightarrow \mathbb{R}$ が単調増加関数ならば

$$f(x) := C + \int_a^x \varphi(t) dt \quad (a \leq x \leq b)$$

は凸関数である. ただし C は定数.

Proof. $x, y \in [a, b]$, $x < y$, $0 \leq \lambda \leq 1$ のとき

$$x \leq \lambda x + (1 - \lambda)y \leq y.$$

だから φ の単調性に注意すれば

$$\begin{aligned} \int_{\lambda x + (1 - \lambda)y}^y \varphi(t) dt &\geq \{y - \lambda x - (1 - \lambda)y\} \varphi(\lambda x + (1 - \lambda)y), \\ \int_x^{\lambda x + (1 - \lambda)y} \varphi(t) dt &\leq \{\lambda x + (1 - \lambda)y - x\} \varphi(\lambda x + (1 - \lambda)y) \end{aligned}$$

であるから

$$\begin{aligned} &\lambda f(x) + (1 - \lambda)f(y) - f(\lambda x + (1 - \lambda)y) \\ &= (1 - \lambda)\{f(y) - f(\lambda x + (1 - \lambda)y)\} - \lambda\{f(\lambda x + (1 - \lambda)y) - f(x)\} \\ &= (1 - \lambda) \int_{\lambda x + (1 - \lambda)y}^y \varphi(t) dt - \lambda \int_x^{\lambda x + (1 - \lambda)y} \varphi(t) dt \\ &\geq (1 - \lambda)\{y - \lambda x - (1 - \lambda)y\} \varphi(\lambda x + (1 - \lambda)y) \\ &\quad - \lambda\{\lambda x + (1 - \lambda)y - x\} \varphi(\lambda x + (1 - \lambda)y) \\ &= 0 \end{aligned}$$

となる。つまり

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(\lambda x + (1 - \lambda)y).$$

□

定理 F.5. f が区間 $[a, b]$ 上の凸関数のとき, 开区間 (a, b) の各点 x で右微分係数 $\dot{f}_+(x)$ が存在し, 関数

$$\dot{f}_+ : (a, b) \ni x \mapsto \dot{f}_+(x) \in \mathbb{R}$$

は単調増加である.

同様に左微分係数 $\dot{f}_-(x)$ が存在し, 関数

$$\dot{f}_- : (a, b) \ni x \mapsto \dot{f}_-(x) \in \mathbb{R}$$

も単調増加である. さらに各点 $x \in (a, b)$ において

$$\dot{f}_-(x) \leq \dot{f}_+(x)$$

が成立する.

Proof. $\forall x \in (a, b)$ に対して実数 η と ξ を

$$0 < \eta < \min\{x - a, b - x\}, \quad 0 < \xi < \eta$$

をみたすように選ぶ. このとき

$$0 < \frac{\xi}{\eta} < 1, \quad x + \xi = \frac{\xi}{\eta}(x + \eta) + \left(1 - \frac{\xi}{\eta}\right)x$$

と書ける. f は凸関数であることから

$$f(x + \xi) \leq \frac{\xi}{\eta}f(x + \eta) + \left(1 - \frac{\xi}{\eta}\right)f(x) = \frac{\xi}{\eta}\{f(x + \eta) - f(x)\} + f(x)$$

となる. つまり $0 < \xi < \eta$ のとき

$$\frac{f(x + \xi) - f(x)}{\xi} \leq \frac{f(x + \eta) - f(x)}{\eta}$$

となり, $\xi \searrow 0$ とすれば

$$f_+(x) \leq \frac{f(x+\eta) - f(x)}{\eta}. \quad (\text{F.4})$$

同様に

$$x - \xi = \frac{\xi}{\eta}(x - \eta) + \left(1 - \frac{\xi}{\eta}\right)x$$

と表せるから

$$f(x - \xi) \leq \frac{\xi}{\eta}f(x - \eta) + \left(1 - \frac{\xi}{\eta}\right)f(x) = \frac{\xi}{\eta}\{f(x - \eta) - f(x)\} + f(x).$$

したがって $0 < \xi < \eta$ のとき

$$\frac{f(x) - f(x - \eta)}{\eta} \leq \frac{f(x) - f(x - \xi)}{\xi}$$

となり, $\xi \searrow 0$ とすれば

$$f_-(x) \geq \frac{f(x) - f(x - \eta)}{\eta}. \quad (\text{F.5})$$

さらに

$$f(x) = f\left(\frac{(x + \xi) + (x - \xi)}{2}\right) \leq \frac{f(x + \xi) + f(x - \xi)}{2}$$

となり

$$f(x) - f(x - \xi) \leq f(x + \xi) - f(x).$$

だから

$$\frac{f(x) - f(x - \xi)}{\xi} \leq \frac{f(x + \xi) - f(x)}{\xi}$$

が得られ, $\xi \searrow 0$ とすると

$$f_-(x) \leq f_+(x). \quad (\text{F.6})$$

~~(F.4) = (F.6)~~ を考慮すれば

$$\frac{f(x) - f(x - \eta)}{\eta} \leq f_-(x) \leq f_+(x) \leq \frac{f(x + \eta) - f(x)}{\eta}. \quad (\text{F.7})$$

よって $\dot{f}_-(x)$ と $\dot{f}_+(x)$ は有界.

次に, 単調性を示す. そのために $a < x_1 < x_2 < b$ とする. (F.7) の

$$f_-(x) \leq f_+(x) \leq \frac{f(x+\eta) - f(x)}{\eta}$$

において $x = x_1, x + \eta = x_2$ とおくと

$$f_-(x_1) \leq f_+(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

を得る. 同様に (F.7) の

$$\frac{f(x) - f(x-\eta)}{\eta} \leq f_-(x) \leq f_+(x)$$

において $x = x_2, x - \eta = x_1$ とおくと

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq f_-(x_2) \leq f_+(x_2)$$

を得る. これらの式を合わせると

$$\dot{f}_-(x_1) \leq \dot{f}_+(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \dot{f}_-(x_2) \leq \dot{f}_+(x_2) \quad (\text{F.8})$$

となる. よって

$$\dot{f}_-(x_1) \leq \dot{f}_-(x_2), \quad \dot{f}_+(x_1) \leq \dot{f}_+(x_2)$$

が得られる. □

定理 F.6. 区間 (a, b) 上で定義された凸関数 f は, 単調増加で右連続関数 g と点 $c \in (a, b)$ によって

$$f(x) = f(c) + \int_c^x g(t) dt \quad (a < x < b)$$

と表せる.

Proof. $a < c < x < b$ とし, 閉区間 $[c, x]$ の分割を

$$c = x_0 < x_1 < \cdots < x_n = x$$

とする. このとき (F.8) より, $k = 1, 2, \dots, n$ に対し

$$\dot{f}_-(x_{k-1}) \leq \dot{f}_+(x_{k-1}) \leq \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \leq \dot{f}_-(x_k) \leq \dot{f}_+(x_k).$$

これらに $x_k - x_{k-1}$ を掛ければ

$$\begin{aligned} \dot{f}_-(x_{k-1})(x_k - x_{k-1}) &\leq \dot{f}_+(x_{k-1})(x_k - x_{k-1}) \leq f(x_k) - f(x_{k-1}) \\ &\leq \dot{f}_-(x_k)(x_k - x_{k-1}) \leq \dot{f}_+(x_k)(x_k - x_{k-1}) \end{aligned}$$

が成立し, 1 から n まで k について和をとると

$$\begin{aligned} \sum_{k=1}^n \dot{f}_-(x_{k-1})(x_k - x_{k-1}) &\leq \sum_{k=1}^n \dot{f}_+(x_{k-1})(x_k - x_{k-1}) \leq f(x) - f(c) \\ &\leq \sum_{k=1}^n \dot{f}_-(x_k)(x_k - x_{k-1}) \leq \sum_{k=1}^n \dot{f}_+(x_k)(x_k - x_{k-1}) \end{aligned}$$

が成立する. \dot{f}_- と \dot{f}_+ は単調増加だから Riemann 積分可能. よって $\max_{1 \leq k \leq n} \{x_k - x_{k-1}\} \rightarrow 0 (n \rightarrow \infty)$ とすれば

$$\begin{aligned} \sum_{k=1}^n \dot{f}_-(x_{k-1})(x_k - x_{k-1}) &\longrightarrow \int_c^x \dot{f}_-(t) dt \\ \sum_{k=1}^n \dot{f}_+(x_{k-1})(x_k - x_{k-1}) &\longrightarrow \int_c^x \dot{f}_+(t) dt \\ \sum_{k=1}^n \dot{f}_-(x_k)(x_k - x_{k-1}) &\longrightarrow \int_c^x \dot{f}_-(t) dt \\ \sum_{k=1}^n \dot{f}_+(x_k)(x_k - x_{k-1}) &\longrightarrow \int_c^x \dot{f}_+(t) dt \end{aligned}$$

となり

$$f(x) - f(c) = \int_c^x \dot{f}_-(t) dt = \int_c^x \dot{f}_+(t) dt.$$

他方 $a < x < b$ のときも閉区間 $[x, c]$ について同様の論理を展開して

$$f(c) - f(x) = \int_x^c \dot{f}_-(t) dt = \int_x^c \dot{f}_+(t) dt$$

を示すことができる。したがって

$$g(x) = \dot{f}_+(x+0)$$

とおけば, g は右連続で単調増加で

$$f(x) = f(c) + \int_c^x g(t) dt$$

となる。 □

定理 F.7. f が区間 (a, b) 上の凸関数であるための必要十分条件は, $\forall x_0 \in (a, b)$ について関数

$$(a, b) \setminus \{x_0\} \ni x \mapsto \frac{f(x) - f(x_0)}{x - x_0} \in \mathbb{R}$$

が単調増加関数となることである。

Proof. 必要性: f が开区間 (a, b) 上で凸関数ならば, 定理 [F.6](#) ^{thm: con-6} により, 右連続な単調増加関数 g が存在して

$$f(x) - f(x_0) = \int_{x_0}^x g(t) dt$$

と表せる。よって $x \neq x_0$ のとき

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{1}{x - x_0} \int_{x_0}^x g(t) dt$$

は $(a, b) \setminus \{x_0\}$ 上で単調増加である。

十分性: $\forall x, y \in (a, b), x < y, 0 < \lambda < 1$ に対し

$$x_0 = \lambda x + (1 - \lambda)y$$

とおく。仮定より $x < x_0 < y$ のとき

$$\frac{f(x_0) - f(x)}{x_0 - x} \leq \frac{f(y) - f(x_0)}{y - x_0}.$$

この式に $x_0 = \lambda x + (1 - \lambda)y$ を代入すると

$$\frac{f(\lambda x + (1 - \lambda)y) - f(x)}{(1 - \lambda)(y - x)} \leq \frac{f(y) - f(\lambda x + (1 - \lambda)y)}{\lambda(y - x)}.$$

両辺に $\lambda(1-\lambda)(y-x)$ を掛けて

$$\lambda\{f(\lambda x + (1-\lambda)y) - f(x)\} \leq (1-\lambda)\{f(y) - f(\lambda x + (1-\lambda)y)\}.$$

これを整理すれば

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y).$$

$\lambda = 0, 1$ のときは自明だから、 f は (a, b) 上で凸関数である。 □

lf:con-8 定義 F.8. f を (a, b) 上の凸関数とする. 点 $x_0 \in (a, b)$ において適当に実数 m を定めると $\forall x \in (a, b)$ に対して

$$S(x) := m(x - x_0) + f(x_0) \leq f(x)$$

が成立するとき、直線 $y = S(x)$ を点 x_0 において f を下から支える直線という。

m:con-9 定理 F.9. f が区間 (a, b) 上の凸関数であるための必要十分条件は、 (a, b) の各点 x_0 において f を下から支える直線 $y = S(x)$ が存在することである。

Proof. 必要性: f は开区間 (a, b) 上の凸関数で $x_0 \in (a, b)$ なので、定理 thm:con-6
F.6 より、ある右連続な単調増加関数 g が存在して

$$f(x) = f(x_0) + \int_{x_0}^x g(t) dt$$

と書ける. このとき

$$f(x) \geq f(x_0) + g(x_0) \int_{x_0}^x dt = f(x_0) + g(x_0)(x - x_0).$$

したがって

$$S(x) := f(x_0) + g(x_0)(x - x_0)$$

は点 x_0 で $f(x)$ を下から支える直線である.

十分性: $x, y \in (a, b)$, $x < y$, $0 \leq \lambda \leq 1$ とし

$$x_0 = \lambda x + (1-\lambda)y$$

とおく. 仮定より x_0 において f を下から支える直線 $S(x)$ が存在する. すると $S(x) \leq f(x)$ と S の線型性から

$$f(\lambda x + (1 - \lambda)y) = f(x_0) = S(x_0) = \lambda S(x) + (1 - \lambda)S(y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

となる. よって f は (a, b) 上の凸関数である. \square

2 凸集合とその性質

定義 F.10. C を \mathbb{R}^n の空でない部分集合とする. 部分集合 C は凸であるとは次の条件をみたすときをいう.

$$\forall x, y \in C \text{ と } 0 \leq \lambda \leq 1 \implies \lambda x + (1 - \lambda)y \in C.$$

定義 F.11. $p \in \mathbb{N}$ とし, $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p\} \subset \mathbb{R}^n$ とする.

(1) $\lambda_k \geq 0 (k = 1, 2, \dots, p)$ かつ $\sum_{k=1}^p \lambda_k = 1$ のとき

$$\sum_{k=1}^p \lambda_k \mathbf{a}_k$$

を $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p$ の凸結合という.

(2) $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p$ によって生成された凸多面体を

$$\langle \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p \rangle = \left\{ \sum_{k=1}^p \lambda_k \mathbf{a}_k; \lambda_k \geq 0, \sum_{k=1}^p \lambda_k = 1 \right\} \subset \mathbb{R}^n.$$

定理 F.12. (Caratheodory の定理) $p \in \mathbb{N}$ とし, $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p\} \subset \mathbb{R}^n$ とする. $\mathbf{y} \in \langle \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p \rangle$ は以下のように表せる.

$$\mathbf{y} = \sum_{k=1}^{n+1} \lambda_k \mathbf{a}_{j_k},$$

$$1 \leq j_1 < j_2 < \dots < j_{n+1} \leq p, \quad \lambda_k \geq 0, \quad \sum_{k=1}^{n+1} \lambda_k = 1.$$

取れない

Proof. 後で書くこと.

□

1: con-13

補題 F.13. C を空でない \mathbb{R}^n 凸閉部分集合とし, $\mathbf{0} \notin C$ とする.

(1) $\mathbf{x}_0 \in C$ があって

$$|\mathbf{x}_0|_2 = \min\{|\mathbf{x}|_2; \mathbf{x} \in C\} > 0.$$

(2) さらに

$$0 \leq \langle \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0 \rangle \quad (\forall \mathbf{x} \in C). \tag{F.9}$$

eq: con-

Proof.

$$\delta = \inf\{|\mathbf{x}|_2; \mathbf{x} \in C\} > 0$$

とおく. 点列 $\{\mathbf{x}_k\}_{k=1}^\infty \subset C$ を選んで

$$\delta = \lim_{k \rightarrow \infty} |\mathbf{x}_k|_2$$

となるようにする. $\{\mathbf{x}_k\}_{k=1}^\infty$ は有界なので Bolzano-Weirstrass の定理より収束する部分列 $\{\mathbf{x}_{k(j)}\}_{j=1}^\infty$ が存在するので

$$\lim_{j \rightarrow \infty} \mathbf{x}_{k(j)} =: \mathbf{x}_0$$

とおく. 結果

$$|\mathbf{x}_0|_2 = \lim_{j \rightarrow \infty} |\mathbf{x}_{k(j)}|_2 = \delta.$$

C は閉集合なので $\mathbf{x}_0 \in C$. $\mathbf{0} \notin C$ なので $\delta > 0$.

(2) $\mathbf{x}_0 \in C$ は $|\mathbf{x}_0|_2 = \delta$ をみたす唯一の点であることを示すために $\mathbf{x}_1 \in C$ で $|\mathbf{x}_1|_2 = \delta$ とする. C の凸性から

$$\frac{1}{2}(\mathbf{x}_0 + \mathbf{x}_1) \in C.$$

したがって

$$\left| \frac{1}{2}(\mathbf{x}_0 + \mathbf{x}_1) \right|_2 \geq \delta.$$

しかし

$$4\delta^2 \leq |\mathbf{x}_0 + \mathbf{x}_1|_2^2 \leq |\mathbf{x}_1 - \mathbf{x}_0|_2^2 + |\mathbf{x}_1 + \mathbf{x}_0|_2^2 \leq 2\{|\mathbf{x}_1|_2^2 + |\mathbf{x}_0|_2^2\} = 4\delta^2$$

より

$$|\mathbf{x}_1 - \mathbf{x}_0|_2^2 = 0 \implies \mathbf{x}_1 = \mathbf{x}_0.$$

leg: con-8a
(F.9) を示すために $\forall \mathbf{x} \in C$ を取る. このとき $(1 - \lambda)\mathbf{x}_0 + \lambda\mathbf{x} \in C$ ($0 \leq \lambda \leq 1$).
したがって

$$|\mathbf{x}_0|_2^2 \leq |(1 - \lambda)\mathbf{x}_0 + \lambda\mathbf{x}|_2^2 = |\mathbf{x}_0|_2^2 + \lambda^2|\mathbf{x} - \mathbf{x}_0|_2^2 + 2\lambda\langle \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0 \rangle.$$

結果

$$0 \leq \lambda|\mathbf{x} - \mathbf{x}_0|_2^2 + 2\langle \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0 \rangle.$$

$\lambda \rightarrow 0$ とすれば leg: con-8a
(F.9) がわかる. □

con-14 定義 F.14. S_1 と S_2 を \mathbb{R}^n の空でない部分集合とする. $\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}$ とする.

(1) 超平面 $\langle \mathbf{a}, \mathbf{x} \rangle + b$ は S_1 と S_2 を分離するとは次の条件をみたすときをいう.

$$\forall \mathbf{x} \in S_1 \implies \langle \mathbf{a}, \mathbf{x} \rangle + b \geq 0,$$

$$\forall \mathbf{x} \in S_2 \implies \langle \mathbf{a}, \mathbf{x} \rangle + b \leq 0.$$

(2) 超平面 $\langle \mathbf{a}, \mathbf{x} \rangle + b$ は S_1 と S_2 を厳密に分離するとは次の条件をみたすときをいう.

$$\forall \mathbf{x} \in S_1 \implies \langle \mathbf{a}, \mathbf{x} \rangle + b > 0,$$

$$\forall \mathbf{x} \in S_2 \implies \langle \mathbf{a}, \mathbf{x} \rangle + b < 0.$$

con-15 定理 F.15. $C \subset \mathbb{R}^n$ を空でない凸閉部分集合とし, $\mathbf{y} \notin C$ とする. このとき C と \mathbf{y} を厳密に分離する超平面 $\langle \mathbf{a}, \mathbf{x} \rangle + b$ ($\mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}$) が存在する.

Proof. 一般性を失うことなく $\mathbf{y} = \mathbf{0}$ としてよい. 左の仮定から $\mathbf{0} \notin C$ に注意して補題 lem: con-13 F.13 を用いると $\mathbf{x}_0 \in C$ があって

$$0 < \delta = |\mathbf{x}_0|_2 = \min\{|\mathbf{x}|_2; \mathbf{x} \in C\}.$$

ここで $z = (1/2)x_0$, $a = x_0/\delta$ とおく. 超平面 $\langle a, x - z \rangle = 0$ を考える. すなわち

$$\langle a, x \rangle + b = 0, \quad b = -\langle a, z \rangle.$$

$y = 0$ に対して

$$\langle a, y \rangle + b = b = -\langle a, z \rangle = -\left\langle \frac{x_0}{\delta}, \frac{x_0}{2} \right\rangle = -\frac{|x_0|^2}{2\delta} = -\frac{\delta}{2} < 0. \quad (\text{F.10})$$

$x \in C$ に対して (F.9) から

$$0 \leq \langle x_0, x - x_0 \rangle.$$

よって

$$0 \leq \langle \delta a, x - 2z \rangle.$$

上の式の両辺を $\delta > 0$ で割れば

$$\begin{aligned} 0 &\leq \langle a, x - z \rangle - \langle a, z \rangle = \langle a, x - z \rangle - \left\langle \frac{x_0}{\delta}, \frac{x_0}{2} \right\rangle \\ &= \langle a, x - z \rangle - \frac{|x_0|^2}{2\delta} \\ &= \langle a, x - z \rangle - \frac{\delta}{2}. \end{aligned}$$

よって

$$\langle a, x - z \rangle = \langle a, x \rangle - \langle a, z \rangle = \langle a, x \rangle + b \geq \frac{\delta}{2} > 0. \quad (\text{F.11})$$

(F.10) と (F.11) より $\langle a, x \rangle + b$ は $y = 0$ と C を厳密に分離することが示された. \square

定義 F.16. $\{a_1, a_2, \dots, a_p\} \subset \mathbb{R}^n$ とする. 集合

$$C = \left\{ \sum_{j=1}^p \lambda_j a_j; \lambda_j \geq 0 (j = 1, 2, \dots, p) \right\}$$

を $\{a_1, a_2, \dots, a_p\}$ によって生成された有限錐という.

con-17

定義 F.17. $\{a_1, a_2, \dots, a_p\}$ が線型独立のとき, このベクトルで生成された有限錐を基本錐という.

con-18

補題 F.18. $\{a_1, a_2, \dots, a_p\} \subset \mathbb{R}^n$ は有限錐 C を生成しているとする. C_1, C_2, \dots, C_q を $\{a_1, a_2, \dots, a_p\}$ の部分集合で生成されるすべての基本錐とする. このとき

$$C = \bigcup_{j=1}^q C_j$$

となる.

Proof. $C_j \subset C$ ($j = 1, 2, \dots, q$) なので

$$\bigcup_{j=1}^q C_j \subset C.$$

いま $b \in C = \{A\beta; \beta = (\beta_1, \beta_2, \dots, \beta_p)^\top, \beta_j \geq 0 (j = 1, 2, \dots, p)\}$ を選ぶ. ただし $A = (a_1, a_2, \dots, a_p)$ は $n \times p$ 行列である. このとき

$$A\beta = b, \quad \beta \geq 0$$

なので, 基本解 $\beta^* = (\beta_1^*, \beta_2^*, \dots, \beta_p^*)^\top \in \mathbb{R}^p$ が存在して

$$A\beta^* = b.$$

すなわち $\exists m \in \mathbb{N}, (1 \leq m \leq n)$ があって, $1 \leq j_1 < \dots < j_m \leq p$ は

$$\beta_{j_\ell} > 0 (j = 1, 2, \dots, m)$$

で

$$\{a_{j_1}, a_{j_2}, \dots, a_{j_m}\}$$

は線型独立である. よって b は $\{a_{j_1}, a_{j_2}, \dots, a_{j_m}\}$ で生成された基本錐に含まれるので

$$b \in \bigcup_{j=1}^q C_j.$$

よって

$$C \subset \bigcup_{j=1}^q C_j.$$

□

在について
と書くこ

定理 F.19. $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p\} \subset \mathbb{R}^n$ とし, C はこのベクトルで生成された有限錐とする. このとき C は凸かつ閉部分集合である.

Proof. 凸性: $\mathbf{c}_1, \mathbf{c}_2 \in C, 0 \leq \lambda \leq 1$ とする. このとき $\beta_1, \beta_2 \in \mathbb{R}^p$ が存在して

$$\mathbf{c}_1 = A\beta_1, \quad \mathbf{c}_2 = A\beta_2.$$

ただし $A = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p), \beta_1 \geq \mathbf{0}, \beta_2 \geq \mathbf{0}$. したがって

$$(1 - \lambda)\mathbf{c}_1 + \lambda\mathbf{c}_2 = A((1 - \lambda)\beta_1 + \lambda\beta_2).$$

さらに $(1 - \lambda)\beta_1 + \lambda\beta_2 \geq \mathbf{0}$. よって

$$(1 - \lambda)\mathbf{c}_1 + \lambda\mathbf{c}_2 \in C$$

より C は凸集合.

C は閉集合: $C_j (j = 1, 2, \dots, q)$ を独立なベクトル

$$\{\mathbf{a}_{j_1}, \mathbf{a}_{j_2}, \dots, \mathbf{a}_{j_\ell}\} \subset \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p\}$$

によって生成される基本錐とする. $\{\mathbf{c}_k\}_{k=1}^\infty$ は C_j の収束する点列とし, $\mathbf{c}_0 \in \mathbb{R}^n$ に収束するとする.

$$\lim_{k \rightarrow \infty} \mathbf{c}_k = \mathbf{c}_0.$$

$\mathbf{c}_0 \in C_j$ を示そう. $\{\mathbf{a}_{j_1}, \mathbf{a}_{j_2}, \dots, \mathbf{a}_{j_\ell}\}$ は線型独立なので, 任意の $\mathbf{u} = (u_1, u_2, \dots, u_\ell)^\top \in \mathbb{R}^\ell$ に対し

$$\sum_{k=1}^{\ell} u_k \mathbf{a}_{j_k} = \mathbf{0} \implies \mathbf{u} = \mathbf{0}.$$

よって上の主張の対偶を取ると $\mathbf{u} \in \mathbb{R}^\ell$ で $|\mathbf{u}|_2 = 1$ なるものに対して

$$\sum_{k=1}^{\ell} u_k \mathbf{a}_{j_k} \neq \mathbf{0}.$$

よって C_j は凸集合なので, 補題 [F.18](#) lem: con-18 より

$$\min \left\{ \left| \sum_{k=1}^{\ell} u_k \mathbf{a}_{j_k} \right|_2 ; |\mathbf{u}|_2 = 1 \right\} =: \epsilon > 0.$$

したがって ϵ の定義に注意すると $\mathbf{v} = (v_1, v_2, \dots, v_\ell)^\top \in \mathbb{R}^\ell$ に対し

$$\left| \sum_{k=1}^{\ell} v_k \mathbf{a}_{jk} \right|_2 = |\mathbf{v}|_2 \left| \sum_{j=1}^{\ell} \frac{v_k}{|\mathbf{v}|_2} \mathbf{a}_{jk} \right|_2 \geq \epsilon |\mathbf{v}|_2 \quad (\text{F.12})$$

eq: con-

となる. $\mathbf{c}_k \in C_j$ ($k = 1, 2, \dots$) に対して

$$\mathbf{c}_k = \mathbf{A}\boldsymbol{\beta}_k$$

とおけば

$$\{j_1, j_2, \dots, j_\ell\} = \{j \in \{1, 2, \dots, p\}; \beta_j > 0, \boldsymbol{\beta}_k = (\beta_1, \beta_2, \dots, \beta_p)^\top\}.$$

このとき

$$\mathbf{c}_k = \sum_{j=1}^{\ell} \beta_{j_k} \mathbf{a}_{jk}.$$

(eq: con-8
(F.12) より)

$$\frac{1}{\epsilon} |\mathbf{c}_k|_2 \geq |\boldsymbol{\beta}_k|_2 \quad (k = 1, 2, \dots).$$

よって $\{\mathbf{c}_k\}_{k=1}^{\infty}$ は収束するので, $\{\boldsymbol{\beta}_k\}_{k=1}^{\infty}$ は有界. よって Bolzano-Weierstrass の定理より収束する部分列 $\{\boldsymbol{\beta}_{k_j}\}_{j=1}^{\infty}$ が取れ

$$\lim_{j \rightarrow \infty} \boldsymbol{\beta}_{k_j} =: \boldsymbol{\beta}_0$$

とする. このとき $\beta_0 \geq 0$ で

$$\mathbf{A}\boldsymbol{\beta}_0 = \lim_{j \rightarrow \infty} \mathbf{A}\boldsymbol{\beta}_{k_j} = \lim_{j \rightarrow \infty} \mathbf{c}_{k_j} = \mathbf{c}_0.$$

β_0 の j_k 成分 ($k = 1, 2, \dots, \ell$) 以外は零なので, $\mathbf{c}_0 \in C_j$. よって C_j は閉集合. だから

$$C = \bigcup_{j=1}^q C_j$$

も閉集合. □

1: con-20

定理 F.20. (Farkas の交代定理) \mathbf{A} を $m \times n$ 行列とし, $\mathbf{b} \in \mathbb{R}^m$ とする. このとき次のいずれかが成立し, 同時には成立しない.

- (1) $Ax = b, x \geq 0$ は解 $x^* \in \mathbb{R}^n$ を持つ。
 (2) $A^\top y \geq 0, \langle y, b \rangle < 0$ は解 $y^* \in \mathbb{R}^n$ を持つ。

ただし $w = (w_1, w_2, \dots, w_n)^\top, z = (z_1, z_2, \dots, z_n)^\top \in \mathbb{R}^n$ のとき

$$w \geq z \iff w_j \geq z_j \quad (j = 1, 2, \dots, n).$$

Proof. $x^* \in \mathbb{R}^m$ は (1) の解, $y^* \in \mathbb{R}^n$ は (2) の解とする。このとき

$$0 \leq \langle x^*, A^\top y^* \rangle = \langle Ax^*, y^* \rangle = \langle b, y^* \rangle < 0.$$

よって矛盾。(1) と (2) は同時に起こらないことがわかる。

(1) が起こらないとき, (2) が起こることを示す。仮定より

$$b \notin C := \{Ax; x \geq 0\}.$$

C は閉凸集合なので超平面が存在して C と b を厳密に分離する。すなわち $a \in \mathbb{R}^m$ と $c \in \mathbb{R}$ があって

$$z \in C \implies \langle a, z \rangle + c > 0 \tag{F.13}$$

かつ

$$\langle a, b \rangle + c < 0. \tag{F.14}$$

$x \geq 0, \lambda > 0$ とし

$$z := \lambda Ax = A(\lambda x) \in C.$$

eq: con-9
(F.13) より

$$\langle a, \lambda Ax \rangle + c > 0.$$

$\lambda > 0$ で割って $\lambda \rightarrow \infty$ とすれば

$$0 \leq \langle a, Ax \rangle = \langle A^\top a, x \rangle \quad (\forall x \geq 0).$$

したがって

$$A^\top a \geq 0.$$

いま $y = a$ とおけば, $A^\top y \geq 0$. eq: con-9
(F.13) において $z = 0 \in C$ とおけば

$$c > 0.$$

eq: con-10
(F.14) より

$$\langle \mathbf{b}, \mathbf{y} \rangle = \langle \mathbf{b}, \mathbf{a} \rangle < -c < 0.$$

よって $\langle \mathbf{b}, \mathbf{y} \rangle < 0$. よって, (1) が起こらなければ (2) が起こる. 対偶をとれば (2) が起こらなければ (1) が起こる. \square

記法: \mathbf{A} を $m \times n$ 行列とする.

$$\ker(\mathbf{A}) := \{\mathbf{x}; \mathbf{A}\mathbf{x} = \mathbf{0}_m\} \subset \mathbb{R}^n,$$

$$\text{Ran}(\mathbf{A}) := \{\mathbf{A}\mathbf{x}; \mathbf{x} \in \mathbb{R}^m\} \subset \mathbb{R}^m.$$

$S \subset \mathbb{R}^m$ としたとき

$$S^\perp := \{\mathbf{x} \in \mathbb{R}^m; \langle \mathbf{x}, \mathbf{s} \rangle = 0 (\forall \mathbf{s} \in S)\}.$$

定理 F.21. \mathbf{A} を $m \times n$ 行列, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ とする. このとき次は同値である.

(1) $\mathbf{A}\mathbf{x} = \mathbf{b}$ は解 \mathbf{x}^* を持つ.

(2) $\mathbf{b} \in \ker(\mathbf{A}^\top)^\perp$.

Proof. $\mathbf{A}\mathbf{x} = \mathbf{b}$ が解を持つための必要十分条件は

$$\tilde{\mathbf{A}} = [\mathbf{A}, -\mathbf{A}] \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \mathbf{b}, \quad \mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}$$

は解を持つ. これは

$$\tilde{\mathbf{A}} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \mathbf{b}, \quad \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \geq \mathbf{0}$$

とした Farkas(1) である. よって

$$\text{「}\mathbf{A}\mathbf{x} = \mathbf{b} \text{ は解 } \mathbf{x}^* \text{ を持つ」} \iff \text{「Farkas (1) が成立」} \quad (\text{F.15})$$

となる.

一方 Farkas(2) が成立するとき

$$\tilde{\mathbf{A}}^\top \mathbf{y} \geq \mathbf{0}, \quad \langle \mathbf{y}, \mathbf{b} \rangle < 0$$

が解を持つ. これは

$$\begin{bmatrix} A^\top \\ -A^\top \end{bmatrix} \mathbf{y} \geq \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

よって $A^\top \mathbf{y} \geq \mathbf{0}$ かつ $-A^\top \mathbf{y} \geq \mathbf{0}$ より $A^\top \mathbf{y} = \mathbf{0}$ なので

$$A^\top \mathbf{y} = \mathbf{0}, \quad \langle \mathbf{y}, \mathbf{b} \rangle < 0$$

は解をもつ.

以上の議論から

1. 「 $Ax = b$ は解 x^* を持つ」が成立する.
2. $\tilde{A}^\top \mathbf{y} \geq \mathbf{0}, \quad \langle \mathbf{y}, \mathbf{b} \rangle < 0$ は解をもつ.

のいずれかが成立 (同時には起こらない) することがわかる.

さらに

$$\text{Farkas (2) が起こらない} \iff \mathbf{b} \in \ker(A^\top)^\perp$$

である. なせならば $\mathbf{b} \in \ker(A^\top)^\perp$ ならば

$$A^\top \mathbf{y} = \mathbf{0} \quad \text{かつ} \quad \langle \mathbf{y}, \mathbf{b} \rangle = 0$$

なので (2) は成立しない. 逆に (2) が成立しないならば Farkas(1) が成立する.

(eq: con-15a
(F.15) より)

$$\text{「Farkas (1) が起こる」} \iff \text{「} Ax = b \text{ は解 } x^* \text{ を持つ」}$$

なので $\mathbf{b} = Ax^*$ と書ける. よって $\forall \mathbf{y} \in \ker(A^\top)$ に対して

$$\langle \mathbf{y}, \mathbf{b} \rangle = \langle \mathbf{y}, Ax^* \rangle = \langle A^\top \mathbf{y}, x^* \rangle = 0$$

から $\mathbf{b} \in \ker(A^\top)^\perp$ がわかる.

以上の議論から

$$\text{「} Ax = b \text{ は解 } x^* \text{ を持つ」} \iff \text{「Farkas(1) が成立」}$$

$$\iff \text{「Farkas(2) が成立しない」}$$

$$\iff \mathbf{b} \in \ker(A^\top)^\perp$$

より定理は証明された. □

3 共役凸関数と Young の不等式

con-22

定義 F.22. 関数 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ が凸であるとは, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ と $0 \leq \lambda \leq 1$ に対して

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

が成り立つときをいう.

con-23

注意 F.23. 関数 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ が凸であるための必要十分条件はそのエピグラフ

$$E = \left\{ \begin{bmatrix} \mathbf{x} \\ y \end{bmatrix} : y \geq f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n \right\} \subset \mathbb{R}^{n+1}$$

が凸集合となることである.

$m \in \mathbb{N}$ とする. f が凸のとき, 帰納法により $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m \in \mathbb{R}^n$ と $\lambda_1, \lambda_2, \dots, \lambda_m \geq 0$ で $\sum_{k=1}^m \lambda_k = 1$ に対して

$$f\left(\sum_{k=1}^m \lambda_k \mathbf{x}_k\right) \leq \sum_{k=1}^m \lambda_k f(\mathbf{x}_k)$$

が成り立つ. □

con-24

定理 F.24. (1) $f : \mathbb{R}^n \rightarrow \mathbb{R}$ は連続微分可能とする. f が凸関数となるための必要十分条件は $\forall \mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^n$ に対して

$$f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle \leq f(\mathbf{x})$$

が成り立つことである. ただし

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right)^\top,$$

$$\mathbf{x} = (x_1, x_2, \dots, x_n)^\top$$

である.

(2) $f : \mathbb{R}^n \rightarrow \mathbb{R}$ は 2 回連続微分可能とする. f が凸であるための必要十分条件は $\forall \mathbf{x} \in \mathbb{R}^n$ に対して

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$$

が成り立つことである。 $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ は $\forall \mathbf{y} = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^n$ に対して

$$\mathbf{y}^\top \nabla^2 f(\mathbf{x}) \mathbf{y} = \sum_{i,j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) y_i y_j \geq 0$$

が成り立つことである。

Proof. まず $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は凸関数であるための必要十分条件は $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ に対して

$$\phi(t) := f(\mathbf{x} + t\mathbf{y}) \quad (t \in \mathbb{R}) \tag{F.16}$$

eq: con-

が凸関数であることである。

まず f は凸関数であるとする。 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n, 0 \leq \lambda \leq 1, s, t \in \mathbb{R}$ を取り、

$$\mathbf{w} := \mathbf{x} + s\mathbf{y}, \quad \mathbf{z} := \mathbf{x} + t\mathbf{y}$$

とおく。このとき

$$f(\lambda\mathbf{w} + (1-\lambda)\mathbf{z}) \leq \lambda f(\mathbf{w}) + (1-\lambda)f(\mathbf{z}) = \lambda\phi(s) + (1-\lambda)\phi(t). \tag{F.17}$$

eq: con-

また

$$f(\lambda\mathbf{w} + (1-\lambda)\mathbf{z}) = f(\mathbf{x} + \{\lambda t + (1-\lambda)s\}\mathbf{y}) = \phi(\lambda t + (1-\lambda)s) \tag{F.18}$$

eq: con-

^{eq: con-15} (F.17) と ^{eq: con-16} (F.18) を考慮すると ϕ は凸関数であることがわかる。

逆を示す。 $0 \leq \lambda \leq 1, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ とする。 ^{eq: con-14} (F.16) において、 \mathbf{x}, \mathbf{y} は任意だったので、 \mathbf{y} と $\mathbf{x} - \mathbf{y}$ とする。すると

$$\phi(t) = f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$$

とおく。 ϕ は凸関数だから

$$\begin{aligned} f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) = \phi(\lambda) \\ &= \phi((1-\lambda) \times 0 + \lambda \times 1) \\ &\leq (1-\lambda)\phi(0) + \lambda\phi(1) \\ &= (1-\lambda)f(\mathbf{y}) + \lambda f(\mathbf{x}). \end{aligned}$$

よって $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ に対して関数

$$\mathbb{R} \ni t \mapsto \phi(t) = f(\mathbf{x} + t\mathbf{y})$$

が凸関数ならば f も凸関数であることがわかる。

2. 凸関数 $\phi(t) = f(\mathbf{x}_0 + t\mathbf{y})$ に対して

$$\phi(t) + \dot{\phi}(t)(s - t) \leq \phi(s) \quad (\forall s, t \in \mathbb{R})$$

である。ただし $\dot{\phi}(t) = \frac{d\phi}{dt} = \langle \nabla f(\mathbf{x}_0 + t\mathbf{y}), \mathbf{y} \rangle$ である。ここで $s = 1, t = 0$ とおくと

$$\phi(0) + \dot{\phi}(0) \leq \phi(1) \iff f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{y} \rangle \leq f(\mathbf{x}_0 + \mathbf{y}).$$

となる。上式に $\mathbf{y} = \mathbf{x} - \mathbf{x}_0$ を代入すると

$$f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle \leq f(\mathbf{x})$$

を得る。

3. 凸関数 $\phi(t) = f(\mathbf{x}_0 + t\mathbf{y})$ は

$$\ddot{\phi}(t) \geq 0$$

をみたす。このとき

$$\dot{\phi}(t) = \langle \nabla f(\mathbf{x}_0 + t\mathbf{y}), \mathbf{y} \rangle, \quad \ddot{\phi}(t) = \mathbf{y}^\top \nabla^2 f(\mathbf{x}_0 + t\mathbf{y}) \mathbf{y}.$$

ただし $\nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{i, j=1, 2, \dots, n}$ である。 $t = 0$ とおくと

$$\mathbf{y}^\top \nabla^2 f(\mathbf{x}) \mathbf{y} \geq 0 \quad (\forall \mathbf{y} \in \mathbb{R}^n).$$

□

: con-25

定義 F.25. $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は凸関数とする。各 $\mathbf{x} \in \mathbb{R}^n$ に対して

$$\partial f(\mathbf{x}) = \{ \mathbf{r} \in \mathbb{R}^n; f(\mathbf{x}) + \langle \mathbf{r}, \mathbf{y} - \mathbf{x} \rangle \leq f(\mathbf{y}) \ (\forall \mathbf{y} \in \mathbb{R}^n) \}$$

とおく。この集合を点 \mathbf{x} における f の劣微分という。

注意 F.26. 凸関数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は点 \boldsymbol{x} で連続微分可能なとき

$$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}.$$

一般に, $\partial f(\boldsymbol{x})$ は 1 つの元よりも多い元をもつことがある. □

例 F.27. (1) $n = 1$, $f(x) = |x|$ とする. このとき

$$\partial f(x) = \begin{cases} \{1\} & (x > 0) \\ [-1, 1] & (x = 0) \\ \{-1\} & (x < 0). \end{cases}$$

(2) $n > 1$, $f(\boldsymbol{x}) = |\boldsymbol{x}|_2$ とする. このとき

$$\partial f(\boldsymbol{x}) = \begin{cases} \left\{ \frac{\boldsymbol{x}}{|\boldsymbol{x}|_2} \right\} & (\boldsymbol{x} \neq \mathbf{0}) \\ B(1) & (\boldsymbol{x} = \mathbf{0}). \end{cases}$$

ただし $B(1) := \{\boldsymbol{x} \in \mathbb{R}^n; |\boldsymbol{x}|_2 = 1\}$ である.

実際, $\forall \boldsymbol{r} \in \partial f(\mathbf{0})$ は $\forall \boldsymbol{x} \in \mathbb{R}^n$ に対して

$$|\mathbf{0}|_2 + \langle \boldsymbol{r}, \boldsymbol{x} - \mathbf{0} \rangle \leq |\boldsymbol{x}|_2$$

を意味する. $|\langle \boldsymbol{r}, \boldsymbol{x} \rangle| \leq |\boldsymbol{r}|_2 \times |\boldsymbol{x}|_2$ から

$$\langle \boldsymbol{r}, \boldsymbol{x} \rangle \leq |\boldsymbol{x}|_2 \iff |\boldsymbol{r}|_2 \leq 1.$$

□

定理 F.28. 関数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は凸とする. 各 $\boldsymbol{x} \in \mathbb{R}^n$ に対して

$\partial f(\boldsymbol{x})$ は凸閉集合である.

Proof. 明らかに $\partial f(\boldsymbol{x})$ は凸集合. これが閉集合であることを示すために

$$\{\boldsymbol{r}_k\}_{k=1}^{\infty} \subset \partial f(\boldsymbol{x}) \quad \text{かつ} \quad \boldsymbol{r}_0 = \lim_{k \rightarrow \infty} \boldsymbol{r}_k$$

を取る. $\boldsymbol{r}_k \in \partial f(\boldsymbol{x})$ ($k = 1, 2, \dots$) だから $\forall \boldsymbol{y} \in \mathbb{R}^n$ に対して

$$f(\boldsymbol{x}) + \langle \boldsymbol{r}_k, \boldsymbol{y} - \boldsymbol{x} \rangle \leq f(\boldsymbol{x}).$$

$k \rightarrow \infty$ とすれば, $\forall \mathbf{y} \in \mathbb{R}^n$ に対して

$$f(\mathbf{x}) + \langle \mathbf{r}_0, \mathbf{y} - \mathbf{x} \rangle \leq f(\mathbf{y})$$

となるから $\mathbf{r}_0 \in \partial f(\mathbf{x})$ となる. よって $\partial f(\mathbf{x})$ は閉集合. □

con-29 定理 F.29. 関数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は凸とする. このとき, 各 $\mathbf{x} \in \mathbb{R}^n$ に対して $\partial f(\mathbf{x}) \neq \emptyset$.

Proof. 凸関数 f のエピグラフ¹ E は凸閉集合なので正の整数 k に対して

$$\mathbf{e}_k := \begin{bmatrix} \mathbf{x} \\ f(\mathbf{x}) - \frac{1}{k} \end{bmatrix} \notin E$$

である. 分離定理 (定理 thm:con-15) よりある超平面 $\langle \mathbf{a}_k, \mathbf{x} \rangle + d_k$ ($\mathbf{a}_k \in \mathbb{R}^{n+1}$, $d_k \in \mathbb{R}$) は \mathbf{e}_k と E を厳密に分離する.

$$\mathbf{a}_k = \begin{bmatrix} \mathbf{b}_k \\ c_k \end{bmatrix}, \quad \mathbf{b}_k \in \mathbb{R}^n, \quad c_k \in \mathbb{R}$$

と書くと

$$\langle \mathbf{a}_k, \mathbf{e}_k \rangle + d_k < 0, \tag{F.19}$$

$$\langle \mathbf{a}_k, \mathbf{z} \rangle + d_k > 0 \quad (\forall \mathbf{z} \in E) \tag{F.20}$$

となる. eq:con-17 (F.19) は

$$\langle \mathbf{b}_k, \mathbf{x} \rangle + c_k \left(f(\mathbf{x}) - \frac{1}{k} \right) + d_k < 0 \tag{F.21}$$

$\forall \mathbf{y} \in \mathbb{R}^n$ に対して

$$\mathbf{z} = \begin{bmatrix} \mathbf{y} \\ f(\mathbf{y}) \end{bmatrix} \in E$$

なので

$$\langle \mathbf{b}_k, \mathbf{y} \rangle + c_k f(\mathbf{y}) + d_k > 0. \tag{F.22}$$

¹epigraph の epi は「上の」という意味の接頭語.

^{eq:con-19}(F.21) を -1 倍して ^{eq:con-20}(F.22) に加わえると $\forall \mathbf{y} \in \mathbb{R}^n$ に対して

$$\langle \mathbf{b}_k, \mathbf{y} - \mathbf{x} \rangle + c_k \left(f(\mathbf{y}) - f(\mathbf{x}) + \frac{1}{k} \right) > 0. \quad (\text{F.23})$$

^{eq:con-21}(F.23) に $\mathbf{y} = \mathbf{x}$ を代入すると

$$\frac{c_k}{k} > 0 \implies c_k > 0.$$

したがって ^{eq:con-21}(F.23) は $\forall \mathbf{y} \in \mathbb{R}^n$ に対して

$$f(\mathbf{y}) + \frac{1}{k} \geq - \left\langle \frac{\mathbf{b}_k}{c_k}, \mathbf{y} - \mathbf{x} \right\rangle + f(\mathbf{x}) \quad (\text{F.24})$$

を意味する。ここで

$$\mathbf{r}_k = -\frac{1}{c_k} \mathbf{b}_k$$

とおく。

次に $\{\mathbf{r}_k\}_{k=1}^\infty \subset \mathbb{R}^n$ は有界であることを示す。 $|\mathbf{r}_k|_2 \neq 0$ ならば ^{eq:con-22}(F.24) において

$$\mathbf{y} = \mathbf{x} + \frac{\mathbf{r}_k}{|\mathbf{r}_k|_2}$$

とおく²と

$$f\left(\mathbf{x} + \frac{\mathbf{r}_k}{|\mathbf{r}_k|_2}\right) + \frac{1}{k} \geq |\mathbf{r}_k|_2 + f(\mathbf{x}).$$

よって $k = 1, 2, \dots$ に対して

$$|\mathbf{r}_k|_2 \leq 1 + \max_{|\mathbf{u}|_2=1} |f(\mathbf{x} + \mathbf{u})| + f(\mathbf{x}) =: M$$

となる。よって $\{\mathbf{r}_k\}_{k=1}^\infty$ は有界であることがわかる。

Bolzano-Weierstrass の定理より収束する部分列 $\{\mathbf{r}_{k(j)}\}_{j=1}^\infty$ が取れる。

$$\mathbf{r}_0 := \lim_{j \rightarrow \infty} \mathbf{r}_{k(j)}$$

²定義から $|\mathbf{b}_k|^2 = c_k^2 |\mathbf{r}_k|^2$ である。

$$\left\langle \frac{\mathbf{b}_k}{c_k}, \frac{\mathbf{r}_k}{|\mathbf{r}_k|_2} \right\rangle = \left\langle \frac{\mathbf{b}_k}{c_k}, -\frac{\mathbf{b}_k}{c_k |\mathbf{r}_k|_2} \right\rangle = -\frac{|\mathbf{b}_k|_2^2}{c_k^2 |\mathbf{r}_k|_2} = -|\mathbf{r}_k|_2$$

になることに注意する。

とおく. ^{eq: con-22}(F.24)において $k = k_j \rightarrow \infty (j \rightarrow \infty)$ とすれば

$$f(\mathbf{y}) \geq \langle \mathbf{r}_0, \mathbf{y} - \mathbf{x} \rangle + f(\mathbf{x}) \quad (\forall \mathbf{y} \in \mathbb{R}^n)$$

を得る. よって $\mathbf{r}_0 \in \partial f(\mathbf{x})$ がわかる. したがって $\partial f(\mathbf{x}) \neq \emptyset$ が示せた. \square

以降, この節では関数 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は凸で

$$\lim_{\|\mathbf{x}\|_2 \rightarrow \infty} \frac{f(\mathbf{x})}{\|\mathbf{x}\|_2} = +\infty \quad (\text{F.25})$$

eq: con-

をみたすとする. これを超線型成長条件という.

con-30

定義 F.30. $\mathbf{y} \in \mathbb{R}^n$ に対し

$$f^*(\mathbf{y}) = \max_{\mathbf{x} \in \mathbb{R}^n} \{ \langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x}) \}$$

とする. f^* を f の凸共役関数という.

con-31

注意 F.31. $n = 1, f(x) = x^2/2 (x \in \mathbb{R})$ とする. このとき $y \in \mathbb{R}$ に対し

$$f^*(y) = \max_x \left(xy - \frac{x^2}{2} \right) = \frac{y^2}{2}.$$

(2) $n = 1, 1 < p < \infty, f(x) = |x|^p/p$ とする. このとき $y \in \mathbb{R}$ に対し

$$f^*(y) = \max_x \left(xy - \frac{|x|^p}{p} \right) = \frac{|y|^q}{q}, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

$\therefore y$ を固定し

$$g(x) = xy - \frac{|x|^p}{p}$$

とする. $\dot{g}(x) = y - |x|^{p-1} \text{sgn}(x)$. ただし $\text{sgn}(x) = \begin{cases} 1 & (x > 0) \\ -1 & (x < 0) \\ 0 & (x = 0) \end{cases}$ とし

た. したがって

$$0 = \dot{g}(x) \iff y = |x|^{p-1} \text{sgn}(x) \implies x = |y|^{1/(p-1)} \text{sgn}(y).$$

結果

$$\begin{aligned} f^*(y) &= (|y|^{1/(p-1)} \operatorname{sgn}(y))y - \frac{|y|^{1/(p-1)} \operatorname{sgn}(y)|y|^p}{p} = |y|^{\frac{1}{1-p}+1} - \frac{|y|^{\frac{p}{p-1}}}{p} \\ &= \left(1 - \frac{1}{p}\right) |y|^{\frac{p}{p-1}} = \frac{|y|^q}{q}. \end{aligned}$$

□

con-33

補題 F.32. $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ に対し

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq f(\mathbf{x}) + f^*(\mathbf{y}).$$

これを Fenchel-Young 双対という.

Proof.

□

con-33

定理 F.33. (1) $f^* : \mathbb{R}^n \rightarrow \mathbb{R}$ は凸関数.

$$(2) \lim_{\|\mathbf{y}\|_2 \rightarrow \infty} \frac{|f^*(\mathbf{y})|}{\|\mathbf{y}\|_2} = +\infty.$$

$$(3) f^{**} = f.$$

Proof. $\mathbf{y}, \mathbf{w} \in \mathbb{R}^n, 0 \leq \lambda \leq 1$ に対し

$$\begin{aligned} f^*(\lambda \mathbf{y} + (1-\lambda)\mathbf{w}) &= \max_{\mathbf{x} \in \mathbb{R}^n} \left(\langle \lambda \mathbf{y} + (1-\lambda)\mathbf{w}, \mathbf{x} \rangle - f(\mathbf{x}) \right) \\ &= \max_{\mathbf{x} \in \mathbb{R}^n} \left(\lambda \{ \langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}) \} + (1-\lambda) \{ \langle \mathbf{w}, \mathbf{x} \rangle - f(\mathbf{x}) \} \right) \\ &\leq \lambda \max_{\mathbf{x} \in \mathbb{R}^n} \left(\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}) \right) + (1-\lambda) \max_{\mathbf{x} \in \mathbb{R}^n} \left(\langle \mathbf{w}, \mathbf{x} \rangle - f(\mathbf{x}) \right) \\ &= \lambda f^*(\mathbf{y}) + (1-\lambda) f^*(\mathbf{w}). \end{aligned}$$

したがって f^* は凸関数.

2. $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ に対して

$$f(\mathbf{x}) + f^*(\mathbf{y}^*) \geq \langle \mathbf{x}, \mathbf{y} \rangle$$

であった. $\mathbf{y} \neq \mathbf{0}$, $\lambda > 0$ に対し

$$\mathbf{x} = \frac{\lambda \mathbf{y}}{\|\mathbf{y}\|_2}$$

とおく. このとき

$$f^*(\mathbf{y}) \geq \left\langle \frac{\lambda \mathbf{y}}{\|\mathbf{y}\|_2}, \mathbf{y} \right\rangle - f\left(\frac{\lambda \mathbf{y}}{\|\mathbf{y}\|_2}\right) \geq \lambda \|\mathbf{y}\|_2 - \max_{\|\mathbf{u}\|_2=\lambda} f(\mathbf{u}).$$

したがって

$$\frac{f^*(\mathbf{y})}{\|\mathbf{y}\|_2} \geq \lambda - \frac{1}{\|\mathbf{y}\|_2} \max_{\|\mathbf{u}\|_2=\lambda} f(\mathbf{u}).$$

ここで

$$\liminf_{\|\mathbf{y}\|_2 \rightarrow \infty} \frac{f^*(\mathbf{y})}{\|\mathbf{y}\|_2} \geq \lambda > 0$$

を得る. $\lambda \rightarrow \infty$ とすれば, (2) がわかる.

(3). $f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle$ から

$$f(\mathbf{x}) \geq \max_{\mathbf{y} \in \mathbb{R}^n} \left(\langle \mathbf{x}, \mathbf{y} \rangle - f^*(\mathbf{y}) \right) = f^{**}(\mathbf{x}). \quad (\text{F.26})$$

逆に定理 [F.29](#) から $\partial f(\mathbf{x}) \neq \emptyset$ だったので $\mathbf{r} \in \partial f(\mathbf{x})$ を取る. すると $\mathbf{z} \in \mathbb{R}^n$ に対して

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \langle \mathbf{r}, \mathbf{z} - \mathbf{x} \rangle.$$

上の式を書きかえると

$$\langle \mathbf{r}, \mathbf{x} \rangle - f(\mathbf{x}) \geq \langle \mathbf{r}, \mathbf{z} \rangle - f(\mathbf{z}).$$

を得る. よって

$$\langle \mathbf{r}, \mathbf{x} \rangle - f(\mathbf{x}) = \max_{\mathbf{z} \in \mathbb{R}^n} \left(\langle \mathbf{r}, \mathbf{z} \rangle - f(\mathbf{z}) \right) = f^*(\mathbf{r}). \quad (\text{F.27})$$

[eq:con-24](#)
(F.27) から

$$f^{**}(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^n} \left(\langle \mathbf{x}, \mathbf{y} \rangle - f^*(\mathbf{y}) \right) \geq \langle \mathbf{x}, \mathbf{r} \rangle - f^*(\mathbf{r}) = f(\mathbf{x}). \quad (\text{F.28})$$

[eq:con-25](#) [eq:con-26](#)
(F.26) と (F.28) から $f^{**} = f$ がわかる. □

定理 F.34. $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ に対して次の条件は同値である.

$$(1) \langle \mathbf{x}, \mathbf{y} \rangle = f(\mathbf{x}) + f^*(\mathbf{y}).$$

$$(2) \mathbf{y} \in \partial f(\mathbf{x}).$$

$$(3) \mathbf{x} \in \partial f^*(\mathbf{y}).$$

Proof. (1) \implies (2) の証明. $\mathbf{z} \in \mathbb{R}^n$ に対し

$$\langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x}) = f^*(\mathbf{y}) \geq \langle \mathbf{y}, \mathbf{z} \rangle - f(\mathbf{z})$$

となる. 上の式の最左辺と最右辺から $\forall \mathbf{z} \in \mathbb{R}^n$ に対し

$$f(\mathbf{z}) \geq \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle + f(\mathbf{x})$$

を得る. よって

$$\mathbf{y} \in \partial f(\mathbf{x})$$

が示せた.

(2) \implies (1) の証明. $\mathbf{y} \in \partial f(\mathbf{x})$ のとき $\forall \mathbf{z} \in \mathbb{R}^n$ に対し

$$f(\mathbf{z}) \geq \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle + f(\mathbf{x}) \implies \langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x}) \geq \langle \mathbf{y}, \mathbf{z} \rangle - f(\mathbf{z})$$

となる. よって

$$\langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x}) \geq \max_{\mathbf{z} \in \mathbb{R}^n} \left(\langle \mathbf{y}, \mathbf{z} \rangle - f(\mathbf{z}) \right) = f^*(\mathbf{y})$$

から

$$\langle \mathbf{x}, \mathbf{y} \rangle \geq f(\mathbf{x}) + f^*(\mathbf{y})$$

を得る. Fenchel-Young の不等式から

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq f(\mathbf{x}) + f^*(\mathbf{y})$$

なので

$$\langle \mathbf{x}, \mathbf{y} \rangle = f(\mathbf{x}) + f^*(\mathbf{y})$$

が示せた.

(1) \iff (3) も同様に示すことができる.

□

4 非線形最適化と KKT 条件

$n, p \in \mathbb{N}$ とし, $f, h_1, h_2, \dots, h_p: \mathbb{R}^n \rightarrow \mathbb{R}$ は連続微分可能とする. 次の記号を導入する.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_p \end{bmatrix},$$

$$\nabla \mathbf{h} = \begin{bmatrix} (\nabla h_1)^\top \\ (\nabla h_2)^\top \\ \vdots \\ (\nabla h_p)^\top \end{bmatrix} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \cdots & \frac{\partial h_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_p}{\partial x_1} & \frac{\partial h_p}{\partial x_2} & \cdots & \frac{\partial h_p}{\partial x_n} \end{bmatrix}.$$

この節では, 制約付き最適化問題

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } \mathbf{h} \leq \mathbf{0} \quad (\text{F.29})$$

の解 \mathbf{x}^* をみつきたい. ただし $\mathbf{h}(\mathbf{x}) \leq \mathbf{0}$ は $\forall j \in \{1, 2, \dots, p\}$ に対して $h_j(\mathbf{x}) \leq 0$ が成立していることである.

$\mathbf{h}(\mathbf{x}) \leq \mathbf{0}$ をみたす点 \mathbf{x}^* を実行可能解という. 実行可能解 \mathbf{x}^* に対して $h_j(\mathbf{x}^*) = 0$ となる制約式を有効制約式という. 解 \mathbf{x}^* に対する有効制約式の添え字集合を J と書く. すなわち

$$J = \{j \in \{1, 2, \dots, p\}; h_j(\mathbf{x}^*) = 0\}.$$

記号 関数 $r: \mathbb{R} \rightarrow \mathbb{R}^n$ に対して

$$r = o(t) \iff \lim_{t \rightarrow 0^+} \frac{|r(t)|_2}{t} = 0.$$

定義 F.35. \mathbf{x}^* における制約 qualification 条件 (CQ 条件) が成立するとは

$$\langle \mathbf{y}, \nabla h_j(\mathbf{x}) \rangle \leq 0 \quad (j \in J) \quad (\text{F.30})$$

をみたす $y \in \mathbb{R}^n$ にたいしてある正数 t_0 と連続曲線 $\{x(t), 0 \leq t < t_0\}$ があって

$$h(x) \leq 0 \quad (0 \leq t < t_0) \quad (\text{F.31})$$

かつ

$$x(t) = x^* + ty + o(t) \quad (t \rightarrow 0+) \quad (\text{F.32})$$

が成立するときをいう。

注意 F.36. 条件 (F.31) は $x(t)$ ($0 \leq t < t_0$) は実行可能解であることを保証している. (F.32) は $x(t)$ の $t = 0$ における右微分係数ベクトル $\dot{x}(0)$ が存在し

$$\dot{x}(0) = y$$

であることを保証している。□

定理 F.37. (Karush-Kuhn-Tucker(KKT) 条件) x^* は問題 (F.29) の解とし, x^* において CQ 条件は成立しているとする. このとき実数 $\lambda_1^*, \lambda_2^*, \dots, \lambda_p^*$ が存在して

$$\nabla f(x^*) + \sum_{j=1}^p \lambda_j^* \nabla h_j(x^*) = 0 \quad (\text{F.33})$$

が成立する. さらに $\lambda^* := (\lambda_1^*, \lambda_2^*, \dots, \lambda_p^*)^\top$ は

$$\lambda^* \geq 0, \quad \langle \lambda^*, h(x^*) \rangle = 0 \quad (\text{F.34})$$

をみたしている;

注意 F.38. 条件 (F.33) と (F.34) を合わせて Karush-Kuhn-Tucker 条件 (KKT 条件) という. (F.34) は

$$\nabla f(x^*) + \nabla h(x^*)^\top \lambda^* = 0$$

とも表すことができる. λ^* は Lagrange 乗数という. $\langle \lambda^*, h(x^*) \rangle = 0$ を相補性条件という. 定理の証明からわかることであるが $\lambda_j^* = 0$ か $h_j(x^*) = 0$ のいずれかが成立しているのこのように呼んでいる。□

Proof. 定理 ^{thm:con-37}F.37 の証明. ベクトル $\mathbf{y} \in \mathbb{R}^n$ は ^{eq:con-30}(F.32) をみたしているとする. CQ 条件で保証されている \mathbf{y} に対応する曲線を $\{\mathbf{x}(t); 0 \leq t < t_0\}$ とする. いま

$$\phi(t) := f(\mathbf{x}(t))$$

とおく. このとき \mathbf{x}^* は問題 ^{eq:con-27}(F.29) の解なので

$$\phi(0) = f(\mathbf{x}^*) \leq f(\mathbf{x}(t)) = \phi(t) \quad (0 \leq t < t_0)$$

が成り立つ. したがって ϕ は $t=0$ で最小値を取るので

$$\dot{\phi}(0) \geq 0$$

となる.

$$\dot{\phi}(0) = \langle \nabla f(\mathbf{x}^*), \dot{\mathbf{x}}(0) \rangle = \langle \nabla f(\mathbf{x}^*), \mathbf{y} \rangle$$

から ^{eq:con-28}(F.30) をみたく $\mathbf{y} \in \mathbb{R}^n$ に対し

$$\langle \nabla f(\mathbf{x}^*), \mathbf{y} \rangle \geq 0$$

となる.

$$\langle \mathbf{y}, \nabla h_j(\mathbf{x}^*) \rangle \leq 0 \quad (j \in J) \implies \langle \mathbf{y}, \nabla f(\mathbf{x}^*) \rangle \geq 0 \quad (\text{F.35})$$

eq:con-

を示せた. Farkas の交代定理を思い出す. 以下の (1) か (2) のいずれかは成立する.

- (1) $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$ は解 \mathbf{x} を持つ.
- (2) $A^\top \mathbf{y} \geq \mathbf{0}$, $\langle \mathbf{y}, \mathbf{b} \rangle < 0$ は解を持つ.

さらに (1) と (2) は同時に成立しない. $J = \{j_1, j_2, \dots, j_k\}$ と書き

$$\begin{aligned} A &= -[\nabla h_{j_1}(\mathbf{x}^*), \nabla h_{j_2}(\mathbf{x}^*), \dots, \nabla h_{j_k}(\mathbf{x}^*)], \\ \mathbf{b} &= \nabla f(\mathbf{x}^*) \end{aligned}$$

に対して Farkas の交代定理を適用する. ^{eq:con-33}(F.35) は

$$A^\top \mathbf{y} \geq \mathbf{0} \implies \langle \mathbf{y}, \mathbf{b} \rangle \geq 0$$

を主張しているので, Farkas 交代定理の (2) は成立しない. 結果 Farkas 交代定理の (1) が成立する. つまり $\sigma_j \geq 0 (j \in J)$ が存在して

$$-\sum_{j \in J} \sigma_j \nabla h_j(\mathbf{x}^*) = \nabla f(\mathbf{x}^*)$$

が成り立つ. そこで $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*)^\top$ を

$$\lambda_j^* = \begin{cases} \sigma_j & (j \in J) \\ 0 & (j \notin J) \end{cases}$$

と定めると, $j \in J$ のとき $h_j(\mathbf{x}^*) = 0$, $j \notin J$ のとき, $\lambda_j^* = 0$ なので

$$\langle \lambda^*, \mathbf{h}(\mathbf{x}^*) \rangle = 0$$

となる. さらに Farkas の交代定理の (1) から

$$\lambda^* \geq \mathbf{0},$$

$$\nabla f(\mathbf{x}^*) + \sum_{j=1}^p \lambda_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}$$

もわかる. □

re: con-

注意 F.39. CQ 条件をみたす制約の例を上げておく

- (1) $\{h_j\}_{j=1}^p$ は \mathbf{x} の線型関数.
- (2) $\{\nabla h_j(\mathbf{x}^*)\}_{j \in J}^p$ は \mathbb{R}^n の線型独立なベクトル.

これらが CQ 条件をみたすことの証明は Evans の pp.115-117 を参照のこと.

□

5 Lagrange 乗数についてのさらなる議論

$f, g_1, g_2, \dots, g_m, h_1, h_2, \dots, h_p : \mathbb{R}^n \rightarrow \mathbb{R}$ とし

$$\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_p \end{bmatrix}$$

とおく. この節では, 下記の最小化問題を考える.

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } \mathbf{g}(\mathbf{x}) = \mathbf{0}, \mathbf{h}(\mathbf{x}) \leq \mathbf{0}. \quad (\text{F.36})$$

定理 F.40. $\mathbf{x}^* \in \mathbb{R}^n$ は問題 (F.36) の解とする. このとき実数の組

$$\gamma^*, \boldsymbol{\mu}^* = (\mu_1^*, \mu_2^*, \dots, \mu_m^*)^\top, \boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_p^*)^\top$$

が存在し, これはすべては 0 に同時になることはなく, さらに

$$\gamma^* \nabla f(\mathbf{x}^*) + \sum_{k=1}^m \mu_k^* \nabla g_k(\mathbf{x}^*) + \sum_{j=1}^p \lambda_j^* \nabla h_j(\mathbf{x}^*) \quad (\text{F.37})$$

と

$$\gamma^* \geq 0, \quad \boldsymbol{\lambda} \geq \mathbf{0}, \quad \langle \boldsymbol{\lambda}, \mathbf{h}(\mathbf{x}^*) \rangle = 0 \quad (\text{F.38})$$

をみたす.

注意 F.41. (1) (F.37) は

$$\gamma^* \nabla f(\mathbf{x}^*) + \nabla \mathbf{g}(\mathbf{x}^*)^\top \boldsymbol{\mu}^* + \nabla \mathbf{h}(\mathbf{x}^*) \boldsymbol{\lambda}^* = \mathbf{0}$$

と書き直すことができる. ただし

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_p \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix},$$

$$\nabla \mathbf{h} = \begin{bmatrix} (\nabla h_1)^\top \\ (\nabla h_2)^\top \\ \vdots \\ (\nabla h_p)^\top \end{bmatrix} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \cdots & \frac{\partial h_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_p}{\partial x_1} & \frac{\partial h_p}{\partial x_2} & \cdots & \frac{\partial h_p}{\partial x_n} \end{bmatrix},$$

$$\nabla \mathbf{g} = \begin{bmatrix} (\nabla g_1)^\top \\ (\nabla g_2)^\top \\ \vdots \\ (\nabla g_m)^\top \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{bmatrix}.$$

(2) $\gamma^* = 1$ のとき [\(F.37\)](#) と [\(F.38\)](#) は

$$\nabla f(\mathbf{x}^*) + \nabla \mathbf{g}(\mathbf{x}^*)^\top \boldsymbol{\mu}^* + \nabla \mathbf{h}(\mathbf{x}^*) \boldsymbol{\lambda}^* = \mathbf{0} \quad (\text{F.39})$$

と

$$\boldsymbol{\lambda} \geq \mathbf{0}, \quad \langle \boldsymbol{\lambda}, \mathbf{h}(\mathbf{x}^*) \rangle = 0 \quad (\text{F.40})$$

となり KKT 条件となる.

$\gamma^* \neq 0$ のときは γ^* または $-\gamma^*$ で割ることで [\(F.37\)](#) と [\(F.38\)](#) を KKT 条件に直すことができる.

$\gamma^* = 0$ のとき, 異常な乗数とよぶ. □

Proof. (定理 [F.40](#) の証明). $k = 1, 2, \dots$ に対して

$$F_k(\mathbf{x}) := f(\mathbf{x}) + \frac{k}{2} \left(\|\mathbf{g}(\mathbf{x})\|_2^2 + \|\mathbf{h}^+(\mathbf{x})\|_2^2 \right)^2 + \frac{1}{2} \|\mathbf{x} - \mathbf{x}^*\|_2$$

とおく. ただし $\ell = 1, 2, \dots, p$ に対し

$$h_\ell^+(\mathbf{x}) = \begin{cases} h_\ell(\mathbf{x}) & (h_\ell(\mathbf{x}) \geq 0 \text{ のとき}) \\ 0 & (h_\ell(\mathbf{x}) = 0 \text{ のとき}) \end{cases}$$

$B = \{\mathbf{x} \in \mathbb{R}^n; |\mathbf{x} - \mathbf{x}^*|_2 \leq 1\}$ とする. すると B はコンパクトで F_k は連続関数なので, 点 $\mathbf{x}_k \in B$ があって

$$F_k(\mathbf{x}_k) = \max_{\mathbf{x} \in B} F_k(\mathbf{x})$$

が成り立つ. したがって

$$F_k(\mathbf{x}_k) \leq F_k(\mathbf{x}^*)$$

である. さらに \mathbf{x}^* は問題 (F.36) ^{eq:con-34} の解なので, $g(\mathbf{x}^*) = \mathbf{0}$, $h(\mathbf{x}^*) \leq \mathbf{0}$ となる. よって h_ℓ^+ の定義から $h^+(\mathbf{x}^*) = \mathbf{0}$ となる. したがって

$$f(\mathbf{x}_k) + \frac{k}{2} \left(|g(\mathbf{x}_k)|_2^2 + |h^+(\mathbf{x}_k)|_2^2 \right) + \frac{1}{2} |\mathbf{x}_k - \mathbf{x}^*|_2^2 \leq f(\mathbf{x}^*) \quad (\text{F.41})$$

が成り立つので $\{\frac{k}{2}|g(\mathbf{x}_k)|_2^2\}_{k=1}^\infty$ と $\{\frac{k}{2}|h^+(\mathbf{x}_k)|_2^2\}_{k=1}^\infty$ は有界なので

$$\lim_{k \rightarrow \infty} g(\mathbf{x}_k) = \mathbf{0}, \quad \lim_{k \rightarrow \infty} h^+(\mathbf{x}_k) = \mathbf{0} \quad (\text{F.42})$$

となる. さらに Bolzano-Weirstrass の定理から $\{\mathbf{x}_k\}_{k=1}^\infty \subset B$ から収束する部分列 $\{\mathbf{x}_{k(j)}\}_{j=1}^\infty$ を選ぶ.

$$\lim_{j \rightarrow \infty} \mathbf{x}_{k(j)} = \bar{\mathbf{x}} \in B$$

と書く. (F.41) ^{eq:con-39} は

$$f(\mathbf{x}_k) + \frac{1}{2} |\mathbf{x}_k - \mathbf{x}^*|_2^2 \leq f(\mathbf{x}^*)$$

となる. したがって

$$f(\bar{\mathbf{x}}) + \frac{1}{2} |\bar{\mathbf{x}} - \mathbf{x}^*|_2^2 \leq f(\mathbf{x}^*) \quad (\text{F.43})$$

となる. しかし (F.42) ^{eq:con-40} から

$$h(\bar{\mathbf{x}}) \leq \mathbf{0}, \quad g(\bar{\mathbf{x}}) = \mathbf{0}$$

となるので, \bar{x} も実行可能となる. したがって

$$f(\mathbf{x}^*) \leq f(\bar{x}) \quad (\text{F.44})$$

を得る. (F.43) と (F.44) を合わせると $\mathbf{x}^* = \bar{x}$ となる. 収束する部分列 $\{\mathbf{x}_{k(j)}\}_{j=1}^{\infty}$ はすべて同じ収束先をもつので

$$\lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}^* \quad (\text{F.45})$$

がわかる.

$\{\mathbf{x}_k\}_{k=1}^{\infty}$ は \mathbf{x}^* に収束するので, k を十分大きくとれば \mathbf{x}_k は B の境界から離れている. このことより関数

$$B \ni \mathbf{x} \mapsto F_k(\mathbf{x}) \in \mathbb{R}$$

は制約なしの局所解 \mathbf{x}_k を持つ. よって

$$\begin{aligned} \mathbf{0} &= \nabla F_k(\mathbf{x}_k) \\ &= \nabla f(\mathbf{x}_k) + k \left((\nabla \mathbf{g}(\mathbf{x}_k))^{\top} \mathbf{g}(\mathbf{x}_k) + (\nabla \mathbf{h}^+(\mathbf{x}_k))^{\top} \mathbf{h}^+(\mathbf{x}_k) \right) + \mathbf{x}^* - \mathbf{x}_k \\ &= \nabla f(\mathbf{x}_k) + k \left((\nabla \mathbf{g}(\mathbf{x}_k))^{\top} \mathbf{g}(\mathbf{x}_k) + (\nabla \mathbf{h}(\mathbf{x}_k))^{\top} \mathbf{h}(\mathbf{x}_k) \right) + \mathbf{x}^* - \mathbf{x}_k \end{aligned} \quad (\text{F.46})$$

となる. 最後の等号は $h_{\ell}^+ \neq 0$ ($\ell = 1, 2, \dots, p$) のとき $\nabla h_{\ell}^+ = \nabla h_{\ell}$ からわかる. つぎに定数

$$\gamma_k := (1 + k^2 |\mathbf{g}(\mathbf{x}_k)|_2^2 + k^2 |\mathbf{h}(\mathbf{x}_k)|_2^2)^{-1/2} > 0$$

を (F.46) の最左辺に掛けると

$$\mathbf{0} = \gamma_k \nabla f(\mathbf{x}_k) + (\nabla \mathbf{g}(\mathbf{x}_k))^{\top} \boldsymbol{\mu}_k + (\nabla \mathbf{h}(\mathbf{x}_k))^{\top} \boldsymbol{\lambda}_k + \gamma_k (\mathbf{x}_k - \mathbf{x}^*) \quad (\text{F.47})$$

を得る. ただし

$$\boldsymbol{\mu}_k := \gamma_k k \mathbf{g}(\mathbf{x}_k), \quad \boldsymbol{\lambda}_k := \gamma_k k \mathbf{h}^+(\mathbf{x}_k) \geq \mathbf{0}$$

である. 記号の定義から

$$\gamma_k^2 + |\boldsymbol{\mu}_k|_2^2 + |\boldsymbol{\lambda}_k|_2^2 = 1$$

なので $\{(\gamma_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)\}_{k=1}^\infty$ は有界である。したがって収束する部分列 $\{(\gamma_{k(j)}, \boldsymbol{\mu}_{k(j)}, \boldsymbol{\lambda}_{k(j)})\}_{j=1}^\infty$ を取れる。すなわち $j \rightarrow \infty$ のとき

$$\gamma_{k(j)} \rightarrow \gamma_0 \geq \in \mathbb{R}, \quad \boldsymbol{\mu}_{k(j)} \rightarrow \boldsymbol{\mu}_0 \in \mathbb{R}^m, \quad \boldsymbol{\lambda}_{k(j)} \rightarrow \boldsymbol{\lambda}_0 \in \mathbb{R}^p.$$

このとき

$$\gamma_0^2 + |\boldsymbol{\mu}_0|_2^2 + |\boldsymbol{\lambda}_0|_2^2 = 1$$

なので

$$(\gamma_0, \boldsymbol{\mu}_0, \boldsymbol{\lambda}_0) \neq (0, \mathbf{0}, \mathbf{0})$$

である。[\(F.47\)](#) で $k = k(j) \rightarrow \infty (j \rightarrow \infty)$ とすると [\(F.45\)](#) から [\(F.37\)](#) を得る。

次に $\mathbf{h}^*(\mathbf{x}^*) \geq \mathbf{0}$ と $\boldsymbol{\lambda}^* \geq \mathbf{0}$ に注意する。ある $\ell \in \{1, 2, \dots, p\}$ に対して

$$h_\ell(\mathbf{x}^*) < 0$$

ならば十分大きな k に対して

$$h_\ell(\mathbf{x}_k) < 0$$

となる。よって $h_\ell^+(\mathbf{x}_k) = 0$ なので λ_k の第 ℓ 成分は 0 となるので $\lambda_\ell^* = 0$ となる。よって

$$\langle \boldsymbol{\lambda}^*, \mathbf{h}(\mathbf{x}^*) \rangle = 0$$

が成り立つ。よって [\(F.38\)](#) が示された。□

単調増加関数の不連続点は高々可算個であることの証明をのせる。

6 変分不等式

$C \subset \mathbb{R}^n$ とする。基本的な問題

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } \mathbf{x} \in C \tag{F.48}$$

の解 \mathbf{x}^* をみつけることを考える。

eq: con-

: con-42

定理 F.42. (i) $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は連続微分可能で, x^* は問題 (F.48) の解のとき

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad (x \in C) \quad (\text{F.49})$$

eq: con-

が成立する.

(ii) f は凸関数とする. x^* が (F.49) をみたすとき, x^* は問題 (F.48) の解である.

: con-43

注意 F.43. (F.49) を変分不等式という. これは制約問題 (F.48) の 1 次変分の形になっている. \square

定理 F.42 の証明: (i) $x \in C$ とする. このとき $0 \leq t \leq 1$ に対して

$$x^* + t(x - x^*) = tx + (1-t)x^* \in C$$

である. x^* は問題 (F.48) の解なので

$$\phi(t) := f(x_0 + t(x - x_0)) \quad (0 \leq t \leq 1)$$

は $t=0$ で最小となる. したがって

$$\dot{\phi}(0) = \left. \frac{d\phi}{dt} \right|_{t=0} \geq 0$$

である. ただし $\dot{\phi}(0)$ は $t=0$ における ϕ の右微分係数である. しかし

$$\dot{\phi}(t) = \langle \nabla f(x^* + t(x - x^*)), x - x^* \rangle$$

なので

$$\dot{\phi}(0) = \langle \nabla f(x^*), x - x^* \rangle \geq 0$$

となる.

(ii) f は点 x^* で連続微分可能なので

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle$$

である. (F.49) より

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0$$

なので

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) \quad (\mathbf{x} \in C)$$

となる. よって \mathbf{x}^* は問題 ^{eq:con-46}(F.48) の解である.

eq:con-44

例 F.44. $f: \mathbb{R}^n \rightarrow \mathbb{R}$ は連続微分可能とし, $\mathbf{A} \in \text{Mat}(m, n; \mathbb{R})$, $\mathbf{b} \in \mathbb{R}^m$ とする. 問題

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{b} \quad (\text{F.50})$$

eq:con-

を考える.

\mathbf{x}^* は問題 ^{eq:con-48}(F.50) の解のとき ^{eq:con-47}(F.49) は

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \quad (\forall \mathbf{x} \in \mathbb{R}^n; \mathbf{A}\mathbf{x} = \mathbf{b})$$

である. $\mathbf{x} \in \mathbb{R}^n$ は

$$\mathbf{x} = \mathbf{x}^* + \mathbf{w} \quad (\mathbf{x} \in \ker(\mathbf{A}))$$

の形でかけるものとする. 結果

$$\langle \nabla f(\mathbf{x}^*), \mathbf{w} \rangle \geq 0 \quad (\forall \mathbf{w} \in \ker(\mathbf{A}))$$

となる. \mathbf{w} を $-\mathbf{w}$ に置き換えると

$$\langle \nabla f(\mathbf{x}^*), \mathbf{w} \rangle \leq 0 \quad (\forall \mathbf{w} \in \ker(\mathbf{A}))$$

となるので

$$\langle \nabla f(\mathbf{x}^*), \mathbf{w} \rangle = 0 \quad (\forall \mathbf{w} \in \ker(\mathbf{A}))$$

がわかる. したがって

$$\nabla f(\mathbf{x}^*) \in (\ker(\mathbf{A}))^\perp = \text{R}(\mathbf{A}^\top).$$

したがってある $\lambda^* \in \mathbb{R}^m$ が存在して

$$\nabla f(\mathbf{x}^*) + \mathbf{A}^\top \lambda^* = \mathbf{0}$$

となる. このとき $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_p^*)^\top$ は制約 $\mathbf{A}\mathbf{x} = \mathbf{b}$ の Lagrange 乗数となる. □

7 凸性と Lagrange 乗数

問題

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } \mathbf{h} \leq \mathbf{0} \quad (\text{F.51})$$

の解 \mathbf{x}^* をみつけるときの Lagrange 乗数の存在を定理 [F.37](#) は保証している。これは関数の凸性の仮定を用いていない。しかし多くの場合 qualification 条件を確認することが容易でない。この節では強い仮定

$$f, h_1, h_2, \dots, h_p : \mathbb{R}^n \longrightarrow \mathbb{R} \text{ は凸関数} \quad (\text{F.52})$$

を課す。

まず凸関数にたいして、KKT 条件は最適解の十分条件になることを示す。

定理 F.45. $f, h_1, h_2, \dots, h_p : \mathbb{R}^n \longrightarrow \mathbb{R}$ は凸関数とする。 $\mathbf{x}^* \in \mathbb{R}^n$ は

$$\mathbf{h}(\mathbf{x}^*) \leq \mathbf{0} \quad (\text{F.53})$$

をみたす。さらにある $\boldsymbol{\mu}^* = (\mu_1^*, \mu_2^*, \dots, \mu_p^*)^\top \in \mathbb{R}^p$ が存在して KKT 条件

$$\nabla f(\mathbf{x}^*) + \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0} \quad (\text{F.54})$$

$$\boldsymbol{\mu}^* \geq \mathbf{0}, \quad \langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}^*) \rangle \geq 0 \quad (\text{F.55})$$

が成立するとする。このとき \mathbf{x}^* は問題 [\(F.51\)](#) の解である。

Proof. $C := \{\mathbf{x} \in \mathbb{R}^n; \mathbf{h}(\mathbf{x}) \leq \mathbf{0}\}$ は実行可能解の集合とする。各 h_j ($j = 1, 2, \dots, p$) は凸関数なので

$$C_j := \{\mathbf{x} \in \mathbb{R}^n; h_j(\mathbf{x}) \leq 0\}$$

は凸集合なので

$$C = \bigcap_{j=1}^p C_j$$

も凸集合となる。^{eq:con-51}(F.54)は

$$\begin{aligned} 0 &= \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle + \sum_{j=1}^p \mu_j^* \langle \nabla h_j(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \\ &= \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle + \left\langle \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \right\rangle \quad (\forall \mathbf{x} \in \mathbb{R}^n) \end{aligned} \quad \text{(F.56)}$$

となる。 $\mathbf{x} \in C$ のとき h_j ($j = 1, 2, \dots, p$) の凸性と^{eq:con-51a}(F.53)より

$$h_j(\mathbf{x}^*) + \langle \nabla h_j(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \leq h_j(\mathbf{x}) \leq 0.$$

上式の両辺を μ_j^* (≥ 0) 倍して和をとれば

$$\left\langle \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \right\rangle \geq 0 \quad (\forall \mathbf{x} \in C) \quad \text{(F.57)}$$

となる。上式と f は凸関数であることを考慮すると

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq f(\mathbf{x}^*) \quad (\forall \mathbf{x} \in C)$$

を得る。したがって \mathbf{x}^* は問題^{eq:con-49}(F.51)の解であることが示せた。□

定義 F.46. 問題^{eq:con-49}(F.51)を考える。

$$\text{ある } \bar{\mathbf{x}} \in \mathbb{R}^n \text{ が存在して } h(\bar{\mathbf{x}}) < 0 \quad \text{(F.58)}$$

が成り立つとき, 問題^{eq:con-49}(F.51)に対する Slater 条件が成立するという。

定理 F.47. \mathbf{x}^* は問題^{eq:con-49}(F.51)の解とし

$$f, h_1, h_2, \dots, h_p : \mathbb{R}^n \longrightarrow \mathbb{R}$$

は凸関数とする。

(i) このとき $\mu^* \in \mathbb{R}^p$ が存在して KKT 条件^{eq:con-51}(F.54)と^{eq:con-52}(F.55)が成り立つ。

(ii) さらに

$$\mathbf{x} \mapsto f(\mathbf{x}) + \langle \mu^*, \mathbf{h}(\mathbf{x}) \rangle$$

は点 \mathbf{x}^* で最小となる。

Proof. 定理 [F.37](#) の [John](#) の formulation から, ある 非負実数の組 $\gamma^*, \mu_1^*, \mu_2^*, \dots, \mu_p^*$ が存在 (すべてが同時に零になることはない.) して

$$\gamma^* \nabla f(\mathbf{x}^*) + \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0} \quad (\text{F.59})$$

$$\gamma^* \geq 0, \quad \mu^* \geq \mathbf{0}, \quad \langle \mu^*, \mathbf{h}(\mathbf{x}^*) \rangle \geq 0 \quad (\text{F.60})$$

が成立する.

まず

$$\gamma^* \neq 0 \quad (\text{F.61})$$

を示す. そのために $\gamma^* = 0$ を仮定して矛盾を導く. すべては同時に零にならないので $\mu^* \neq \mathbf{0}$ で

$$\sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*) = \mathbf{0}. \quad (\text{F.62})$$

したがって

$$\left\langle \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \right\rangle = 0 \quad (\text{F.63})$$

かつ $h(\bar{\mathbf{x}}) < 0$ である. しかし h_j ($j = 1, 2, \dots, p$) は凸関数なので

$$h_j(\mathbf{x}^*) + \langle \nabla h_j(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \rangle \leq h_j(\bar{\mathbf{x}})$$

が成り立つ. したがって

$$\langle \mu^*, \mathbf{h}(\mathbf{x}^*) \rangle + \left\langle \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \right\rangle = \langle \mu^*, \mathbf{h}(\mathbf{x}^*) \rangle$$

と [\(F.63\)](#) と Slater 条件 (i), $\mu^* \geq \mathbf{0}$, $\mu^* \neq \mathbf{0}$ より

$$\langle \mu^*, \mathbf{h}(\mathbf{x}^*) \rangle \leq \langle \mu^*, \mathbf{h}(\bar{\mathbf{x}}) \rangle < 0$$

となる. これは [\(F.62\)](#) と矛盾する.

よって $\gamma^* > 0$ が示せた. (F.59) を $\gamma^* > 0$ で割り

$$\nabla f(\mathbf{x}^*) + \sum_{j=1}^p \frac{\mu_j^*}{\gamma^*} \nabla h_j(\mathbf{x}^*) = \mathbf{0}$$

を考える. $\tilde{\mu}_j^* = \mu_j^*/\gamma^*$ とおけば

$$\begin{aligned} f(\mathbf{x}^*) + \sum_{j=1}^p \tilde{\mu}_j^* \nabla h_j(\mathbf{x}^*) &= \mathbf{0}, \\ \tilde{\boldsymbol{\mu}} &= (\tilde{\mu}_1^*, \tilde{\mu}_2^*, \dots, \tilde{\mu}_p^*) \geq \mathbf{0}, \\ \langle \tilde{\boldsymbol{\mu}}, \mathbf{h}(\mathbf{x}^*) \rangle &= 0 \end{aligned}$$

を得る. 再度 h_j ($j = 1, 2, \dots, p$) の凸性より

$$h_j(\mathbf{x}^*) - \langle \nabla h_j(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \rangle \leq h_j(\bar{\mathbf{x}})$$

を得る. さらに f の凸性から

$$f(\bar{\mathbf{x}}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \rangle$$

なので

$$\begin{aligned} f(\bar{\mathbf{x}}) + \langle \tilde{\boldsymbol{\mu}}^*, \mathbf{h}(\bar{\mathbf{x}}) \rangle &\geq f(\bar{\mathbf{x}}) + \langle \nabla f(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \rangle + \langle \tilde{\boldsymbol{\mu}}^*, \mathbf{h}(\mathbf{x}^*) \rangle \\ &\quad + \left\langle \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \right\rangle \\ &\geq f(\bar{\mathbf{x}}) + \langle \tilde{\boldsymbol{\mu}}^*, \mathbf{h}(\mathbf{x}^*) \rangle + \left\langle \nabla f(\mathbf{x}^*) + \sum_{j=1}^p \mu_j^* \nabla h_j(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \right\rangle \\ &= f(\mathbf{x}^*) + \langle \tilde{\boldsymbol{\mu}}^*, \mathbf{h}(\mathbf{x}^*) \rangle \end{aligned}$$

を得る. よって (ii) は示された. □

8 凸双対性

8.1 双対問題

与えられた関数 $f, g_1, g_2, \dots, g_m, h_1, h_2, \dots, h_p: \mathbb{R}^n \rightarrow \mathbb{R}$ に対して

$$\mathbf{h} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_m \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_p \end{bmatrix}$$

と書くことにする.

問題

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } g(\mathbf{x}) = \mathbf{0}, h(\mathbf{x}) \leq \mathbf{0} \quad (\text{F.64})$$

eq:con-

の解を \mathbf{x}^* をみつきたいとする.

定義 F.48. 問題 (F.64) ^{eq:con-64} に対応する Lagrange 問題 $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ ($\mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\mu} \in \mathbb{R}^p$) を

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \langle \boldsymbol{\lambda}, g(\mathbf{x}) \rangle + \langle \boldsymbol{\mu}, h(\mathbf{x}) \rangle$$

で定める. $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m)^\top, \boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)^\top$ と書いたとき

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^m \lambda_j g_j(\mathbf{x}) + \sum_{j=1}^p \mu_j h_j(\mathbf{x})$$

とも書ける.

例 F.49. f は凸関数とし, 問題

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0} \quad (\text{F.65})$$

eq:con-

を考える。ただし $A \in \text{Mat}(q, n; \mathbb{R})$, $b \in \mathbb{R}^m$ とする。すると

$$\begin{aligned} L^*(\lambda, \mu) &= \min_{x \in \mathbb{R}^n} \left\{ f(x + \langle \lambda, g(x) \rangle) + \langle \mu, h(x) \rangle \right\} \\ &= \min_{x \in \mathbb{R}^n} \left\{ f(x) + \langle \lambda, Ax - b \rangle - \langle \mu, x \rangle \right\} \\ &= -\langle \lambda, b \rangle + \min_{x \in \mathbb{R}^n} \left\{ f(x) + \langle A^\top \lambda - \mu, x \rangle \right\} \\ &= -\langle \lambda, b \rangle - \max_{x \in \mathbb{R}^n} \left\{ \langle \mu - A^\top \lambda, x \rangle - f(x) \right\} \end{aligned}$$

となる。したがって

$$L^*(\lambda, \mu) = -\langle \lambda, b \rangle - f^*(\mu - A^\top \lambda)$$

となる。ただし

$$f^*(x) = \max_{y \in \mathbb{R}^n} \left\{ \langle x, y \rangle - f(y) \right\}$$

は凸関数 f の双対関数である。 □

定義 F.50. 問題 $(F.64)$ ^{eq: con-64} に対する双対問題は

$$\text{minimize } L^*(\lambda, \mu) \quad \text{subject to } \mu \geq 0 \tag{F.66}$$

の解 $\lambda^* \in \mathbb{R}^m$, $\mu^* \in \mathbb{R}^p$ をみつけることである。

例 F.51. $f(x) = \langle x, a \rangle$ とする。ただし $a \in \mathbb{R}^n$ は定数ベクトルである。線型計画法

$$\text{minimize } \langle x, a \rangle \quad \text{subject to } Ax = b, x \geq 0$$

を考える。ただし $A \in \text{Mat}(m, n; \mathbb{R})$, $b \in \mathbb{R}^m$ である。いま

$$f^*(y) = \max_{x \in \mathbb{R}^n} \left\{ \langle y - a, x \rangle \right\} = \begin{cases} 0 & (y = a) \\ \infty & (\text{その他}) \end{cases}$$

と定める。したがって

$$L^*(\lambda, \mu) = -\langle \lambda, b \rangle - f^*(\mu - A^\top \lambda) = \begin{cases} -\langle \lambda, b \rangle & (\mu - A^\top \lambda = a) \\ 0 & (\text{その他}) \end{cases}$$

となる。このとき双対問題は

$$\text{minimize } -\langle \boldsymbol{\lambda}, \mathbf{b} \rangle \quad \text{subject to } \boldsymbol{\mu} - \mathbf{A}^\top \boldsymbol{\lambda} = \mathbf{a} = \mathbf{0}, \boldsymbol{\mu} \geq \mathbf{0}$$

となる。 □

con-52

定理 F.52. (弱双対不等式)

$$\sup_{\boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\lambda} \in \mathbb{R}^m} L^*(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \inf_{\mathbf{g}(\mathbf{x})=\mathbf{0}, \mathbf{h} \leq \mathbf{0}} f(\mathbf{x}). \quad (\text{F.67})$$

eq: con-

この不等式が厳密に成立するとき、双対ギャップがあるという。

Proof. $\boldsymbol{\mu} \geq \mathbf{0}, \mathbf{g} = \mathbf{0}, \mathbf{h} \leq \mathbf{0}$ とする。このとき

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \langle \boldsymbol{\mu}, \mathbf{h}(\mathbf{x}) \rangle + \underbrace{\langle \boldsymbol{\mu}, \mathbf{h}(\mathbf{x}) \rangle}_{\leq 0} + \underbrace{\langle \boldsymbol{\lambda}, \mathbf{g}(\mathbf{x}) \rangle}_{=0} \leq f(\mathbf{x})$$

となる。

$$L^*(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\mathbf{x} \in \mathbb{R}^n} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

より

$$L^*(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq f(\mathbf{x}).$$

よって (F.67) が示された。 □

8.2 Slaker 条件 (再訪問)

$\mathbf{A} \in \text{Mat}(m, n; \mathbb{R}), m \leq n, \mathbf{b} \in \mathbb{R}^m$ とし、

$$\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$$

とおく。さらに

$$\text{rank}(\mathbf{A}) = m$$

とする。

主問題

$$\text{minimize } f(\mathbf{x}) \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{h}(\mathbf{x}) \leq \mathbf{0} \quad (\text{F.68})$$

eq: con-

の解 $x^* \in \mathbb{R}^n$ をみつけた. 対応する双対問題

$$\text{maximize } L^*(\lambda, \mu) \quad \text{subject to } \mu \geq 0 \quad (\text{F.69})$$

の解 $\mu^* \in \mathbb{R}^m, \lambda \in \mathbb{R}^p$ をみつけたい. ただし

$$L^*(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu),$$

$$L(x, \lambda, \mu) = f(x) + \langle \lambda, Ax + b \rangle + \langle \mu, h(x) \rangle$$

である.

定理 F.53. $h_1, h_2, \dots, h_p: \mathbb{R}^n \rightarrow \mathbb{R}$ は凸関数とする. 修正 Slater 条件が成立するとする.

$$\text{ある } \bar{x} \in \mathbb{R}^n \text{ があって } h(\bar{x}) \leq 0, A\bar{x} = b \text{ をみたす.} \quad (\text{F.70})$$

このとき以下が成り立つ.

(i) x^* は主問題 (F.68) の解のとき

$$\text{組 } (\lambda^*, \mu^*) \text{ は双対問題 (F.69) の解である.} \quad (\text{F.71})$$

(ii) さらに

$$L^*(\lambda^*, \mu^*) = f(x^*) \quad (\text{F.72})$$

が成立する. したがって強双対性

$$\max_{\mu \in \mathbb{R}^p, \lambda \in \mathbb{R}^m, \mu \geq 0} L^*(\lambda^*, \mu^*) = \min_{g(x)=0, h(x) \geq 0} f(x)$$

が成立する.

Proof. 定理 F.40 から $(\gamma^*, \lambda^*, \mu^*) \neq (0, 0_p, 0_m)$ があって F. John の条件をみたす.

$$\gamma^* \nabla f(x^*) + A^\top \lambda^* + \nabla h(x^*)^\top \mu^* = 0_n,$$

$$\gamma^* \geq 0, \mu \geq 0_m, \lambda \geq 0_p, \langle \mu^*, h(x^*) \rangle = 0.$$

まず背理法を用いて

$$\gamma^* \neq 0 \tag{F.73}$$

を示す. そのために $\gamma = 0$ と仮定すると

$$\mathbf{A}^\top \boldsymbol{\lambda}^* + \nabla \mathbf{h}(\mathbf{x}^*)^\top \boldsymbol{\mu}^* = \mathbf{0}_m. \tag{F.74}$$

^{eq:con-70}(F.70) をみたく $\bar{\mathbf{x}}$ を用いて $\mathbf{x}^* - \bar{\mathbf{x}}$ と上の式の内積をとると

$$\begin{aligned} 0 &= \langle \mathbf{A}^\top \boldsymbol{\lambda}^*, \bar{\mathbf{x}} - \mathbf{x}^* \rangle + \langle \nabla \mathbf{h}(\mathbf{x}^*)^\top \boldsymbol{\mu}^*, \bar{\mathbf{x}} - \mathbf{x}^* \rangle \\ &= \langle \boldsymbol{\lambda}^*, \mathbf{A}\bar{\mathbf{x}} - \mathbf{A}\mathbf{x}^* \rangle + \langle \boldsymbol{\mu}^*, \nabla \mathbf{h}(\mathbf{x}^*)(\bar{\mathbf{x}} - \mathbf{x}^*) \rangle \end{aligned}$$

となる. しかし

$$\mathbf{A}\bar{\mathbf{x}} - \mathbf{A}\mathbf{x}^* = \mathbf{b} - \mathbf{b} = \mathbf{0}_m$$

なので

$$\langle \boldsymbol{\mu}^*, \nabla \mathbf{h}(\mathbf{x}^*)(\bar{\mathbf{x}} - \mathbf{x}^*) \rangle = 0 \tag{F.75}$$

を得る. 背理法の仮定から $\gamma \neq 0$ なので KKT 条件は

$$\nabla \mathbf{h}(\mathbf{x}^*) + \mathbf{A}^\top \boldsymbol{\lambda}^* + \nabla \mathbf{h}(\mathbf{x}^*)^\top \boldsymbol{\mu}^* = \mathbf{0}_n \tag{F.76}$$

となる. この等式が ^{eq:con-73}(F.73) を導くことを示そう. $f, h_j (j = 1, 2, \dots, p)$ は凸関数であることと ^{eq:con-75}(F.75) を用いると

$$\begin{aligned} f(\mathbf{x}^*) &= f(\mathbf{x}^*) + \underbrace{\langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x}^* - \mathbf{b} \rangle}_{=0} + \underbrace{\langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}^*) \rangle}_{=0} \\ &\leq \underbrace{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}^*), \mathbf{x}^* - \mathbf{x} \rangle}_{f \text{ の凸性から}} + \langle \boldsymbol{\lambda}^*, (\mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x}) + (\mathbf{A}\mathbf{x} - \mathbf{b}) \rangle \\ &\quad + \underbrace{\langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}) + \nabla \mathbf{h}(\mathbf{x}^*)(\mathbf{x}^* - \mathbf{x}) \rangle}_{h_j \text{ の凸性と } \boldsymbol{\mu}^* \geq \mathbf{0}_m} \\ &= f(\mathbf{x}) + \langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \\ &\quad + \langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}) \rangle + \underbrace{\langle \nabla f(\mathbf{x}^*) + \mathbf{A}^\top \boldsymbol{\lambda}^* + \nabla \mathbf{h}(\mathbf{x}^*)^\top \boldsymbol{\mu}^*, \mathbf{x}^* - \bar{\mathbf{x}} \rangle}_{=0_n \text{ : (F.76) } \supseteq \text{eq:con-76}} \\ &= f(\mathbf{x}) + \langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}) \rangle \end{aligned}$$

となる. さらに h_j ($j = 1, 2, \dots, p$) の凸性から

$$h_j(\mathbf{x}^*) + \langle \nabla h_j(\mathbf{x}^*), \bar{\mathbf{x}} - \mathbf{x}^* \rangle \leq h_j(\bar{\mathbf{x}})$$

なので, (F.75) と合わせると

$$\langle \boldsymbol{\mu}^*, \mathbf{h}^*(\mathbf{x}^*) \rangle \leq \langle \boldsymbol{\mu}^*, \mathbf{h}(\bar{\mathbf{x}}) \rangle \quad (\text{F.77})$$

となる. $\boldsymbol{\mu}^* \neq \mathbf{0}_m$ のとき, 修正 Slater 条件より

$$\langle \boldsymbol{\mu}^*, \mathbf{h}(\bar{\mathbf{x}}) \rangle < 0 \quad (\text{F.78})$$

となる. (F.77) と (F.78) から

$$\langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}^*) \rangle < 0$$

となり, $\langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}^*) \rangle = 0$ と矛盾する. よって $\gamma^* \neq 0$ ならば $\boldsymbol{\mu} \neq \mathbf{0}_m$ とはならない.

次に $\boldsymbol{\mu} = \mathbf{0}_m$ の場合を考える. すると (F.76) から

$$A\boldsymbol{\lambda}^* = \mathbf{0}_m$$

となる. A は非退化なので $\boldsymbol{\lambda}^* = \mathbf{0}_p$ となり, $(\boldsymbol{\lambda}^*, \boldsymbol{\lambda}, \boldsymbol{\mu}) = (0, \mathbf{0}_m, \mathbf{0}_p)$ となるので矛盾する. よって $\gamma^* \neq 0$ となる. したがって

$$f(\mathbf{x}^*) \leq \inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) + \langle \boldsymbol{\lambda}^*, A\mathbf{x} - \mathbf{b} \rangle + \langle \boldsymbol{\mu}^*, \mathbf{h}(\mathbf{x}) \rangle \right\} = L^*(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$$

となる. 逆向きの不等式は定理 (F.52) よりわかる. □

9 凸双対性 (その2)

9.1 Fenchel 双対性

2 つの凸かつ下半連続関数³

$$f: \mathbb{R}^n \longrightarrow (-\infty, \infty) \cup \{+\infty\}, \quad g: \mathbb{R}^m \longrightarrow (-\infty, \infty) \cup \{+\infty\}$$

と $m \times n$ 行列 A が与えられたとする.

con-54

定義 F.54. f と g の領域をそれぞれ

$$\text{dom } f = \{x \in \mathbb{R}^n; f(x) < \infty\}$$

$$\text{dom } g = \{y \in \mathbb{R}^m; g(y) < \infty\}$$

で定義する.

con-55

仮定 F.55. f と g は $-\infty$ の値は取らず

$$\text{dom } f \neq \emptyset, \quad \text{dom } g \neq \emptyset$$

とする. このことが成り立つとき, これらの関数は proper であるという.

f, g は凸関数なので, Young の不等式 (補題 ^{lem:con-33}F.32 を利用すると, $x \in \mathbb{R}^n$ と $y \in \mathbb{R}^m$ に対して

$$f(x) + f^*(A^\top y) \geq \langle x, A^\top y \rangle,$$

$$g(Ax) + g^*(-y) \geq -\langle Ax, y \rangle$$

が成立するとする.

これらの不等式の辺々を加えると

$$f(x) + g(Ax) \geq -f(A^\top y) - g^*(-y) \quad (x \in \mathbb{R}^n, y \in \mathbb{R}^m) \quad (\text{F.79})$$

eq:con-

³ f は点 x で ^{かはん}下半連続であるとは, 任意の $\{x_k\}_{k=1}^\infty \subset \mathbb{R}^n$ で $\lim_{k \rightarrow \infty} x_k = x$ なるものに対して

$$\liminf_{k \rightarrow \infty} f(x_k) = f(x)$$

をみたすときをいう.

となる. このことを踏まえて以下の主問題と双対問題を導入することにする.

$$\text{minimize } f(\mathbf{x}) + g(\mathbf{Ax}) \quad (\mathbf{x} \in \mathbb{R}^n) \quad (\text{F.80})$$

と双対問題

$$\text{maximize } -f^*(\mathbf{A}^\top \mathbf{y}) - g^*(-\mathbf{y}) \quad (\mathbf{y} \in \mathbb{R}^m). \quad (\text{F.81})$$

このとき (F.79) は弱双対性を保証する.

以上の議論から以下の定理を得る.

定理 F.56.

$$\sup_{\mathbf{y} \in \mathbb{R}^m} \left\{ -f^*(\mathbf{A}^\top \mathbf{y}) - g^*(-\mathbf{y}) \right\} \leq \inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) + g(\mathbf{Ax}) \right\}$$

が成立する.

定義 F.57. 値関数 $v: \mathbb{R}^m \rightarrow \mathbb{R} \cup \{+\infty\}$ を

$$v(\mathbf{a}) := \inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) + g(\mathbf{Ax} + \mathbf{a}) \right\} \quad (\mathbf{a} \in \mathbb{R}^m)$$

で定める.

補題 F.58. 値関数 $v: \mathbb{R}^m \rightarrow (-\infty, \infty) \cup \{+\infty\}$ は凸かつ下半連続である.

Proof. $\mathbf{a}, \tilde{\mathbf{a}} \in \mathbb{R}^m, 0 \leq \lambda \leq 1$ に対して

$$g(\mathbf{Ax} + (\lambda \mathbf{a} + (1 - \lambda)\tilde{\mathbf{a}})) \leq \lambda g(\mathbf{Ax} + \mathbf{a}) + (1 - \lambda)g(\mathbf{Ax} + \tilde{\mathbf{a}})$$

から

$$\begin{aligned} v(\lambda \mathbf{a} + (1 - \lambda)\tilde{\mathbf{a}}) &\leq \inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ \mathbf{Ax} + \lambda g(\mathbf{Ax} + \mathbf{a}) + (1 - \lambda)g(\mathbf{Ax} + \tilde{\mathbf{a}}) \right\} \\ &= \inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ \lambda(\mathbf{Ax} + g(\mathbf{Ax} + \mathbf{a})) + (1 - \lambda)(\mathbf{Ax} + g(\mathbf{Ax} + \tilde{\mathbf{a}})) \right\} \\ &\leq \lambda \inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ \mathbf{Ax} + g(\mathbf{Ax} + \mathbf{a}) \right\} \\ &\quad + (1 - \lambda) \inf_{\mathbf{x} \in \mathbb{R}^n} \left\{ \mathbf{Ax} + g(\mathbf{Ax} + \tilde{\mathbf{a}}) \right\} \\ &= \lambda v(\mathbf{a}) + (1 - \lambda)v(\tilde{\mathbf{a}}) \end{aligned}$$

から v は凸関数であることがわかる.

$a \in \mathbb{R}^m$ とし, $\{a_k\}_{k=1}^{\infty} \subset \mathbb{R}^m$ は $\lim_{k \rightarrow \infty} a_k = a$ となる任意の点列とする.

$$\begin{aligned} \liminf_{k \rightarrow \infty} v(a_k) &= \liminf_{k \rightarrow \infty} \left\{ f(x + g(Ax + a_k)) \right\} \\ &= f(x + \liminf_{k \rightarrow \infty} g(Ax + a_k)) \\ &= f(x) + g(Ax + a) \quad (\because g \text{ は下半連続}) \end{aligned}$$

を得る. よって v は下半連続である. □

定義 F.59. 関数 f, g と行列 A の組が双対条件をみたすとは

$$\text{凸関数 } v \text{ は原点の近くで } +\infty \text{ の値を取らない} \quad (\text{F.82})$$

をみたすときをいう.

注意 F.60. 条件 (F.82) は次のように書きかえることができる. ある小さな正数 $\delta > 0$ があって $a \in \mathbb{R}^m$ が $|a|_2 \leq \delta$ をみたすときある $y \in \text{dom } g, x \in \text{dom } f$ があって

$$a = y - Ax \quad (\text{F.83})$$

とかけることである. したがって

$$v(a) \leq f(x) + g(Ax + a) = f(x) + g(y) < \infty$$

となる. □

定理 F.61. $x^* \in \text{mathbbR}^n$ が主問題 (F.80) の解とし, f, g, A は双対条件 (F.82) をみたすとする. このとき以下が成立する.

(i) $y^* \in \mathbb{R}^m$ が存在して

$$f(x^*) + g(A^\top x^*) = -f^*(A^\top y^*) - g^*(-y^*)$$

をみたす.

(ii) 強双対性が成立する.

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) + g(Ax) \right\} = \max_{y \in \mathbb{R}^m} \left\{ -f^*(A^\top y^*) - g^*(-y^*) \right\} \quad (\text{F.84})$$

Proof. 注意 [re:con-60](#) より, 双対条件 ([F.82](#)) は原点の近くで有限の値を取ることがわかる. よって v は凸関数であることを合わせて考えると $\partial v(\mathbf{0}) \neq \emptyset$ となることがわかる. そこで

$$-\mathbf{y}^* \in \partial v(\mathbf{0})$$

を取る. このとき $\mathbf{x} \in \mathbb{R}^n$ と $\mathbf{a} \in \mathbb{R}^m$ に対して

$$\begin{aligned} f(\mathbf{x}^*) + g(\mathbf{Ax}) &= v(\mathbf{0}) \leq v(\mathbf{a}) + \langle \mathbf{a}, \mathbf{y}^* \rangle \quad (\because -\mathbf{y}^* \in \partial v(\mathbf{0})) \\ &\leq f(\mathbf{x}) + g(\mathbf{Ax} + \mathbf{a}) + \langle \mathbf{a}, \mathbf{y}^* \rangle \\ &= -\left(\langle \mathbf{A}^\top \mathbf{y}^*, \mathbf{x} \rangle - f(\mathbf{x})\right) - \left(\langle \mathbf{y}^*, \mathbf{Ax} + \mathbf{a} \rangle - g(\mathbf{Ax} + \mathbf{a})\right) \end{aligned}$$

となる. \mathbf{a} に関して infimum を取ると

$$f(\mathbf{x}^*) + g(\mathbf{Ax}) \leq -\left(\langle \mathbf{A}^\top \mathbf{y}^*, \mathbf{x} \rangle - f(\mathbf{x})\right) - g^*(-\mathbf{y}^*)$$

となる. さらに \mathbf{x} に関して infimum を取ると

$$f(\mathbf{x}^*) + g(\mathbf{Ax}^*) \leq -f(\mathbf{A}^\top \mathbf{y}^*) - g^*(-\mathbf{y}^*)$$

を得る.

逆向きの不等式は定理 [thm:con-56](#) からわかる. □

10 半正定値計画法

記号 (i) $\text{Sym}(n; \mathbb{R})$ を $n \times n$ 対称行列全体の成す線型部分空間とし

$$\text{Sym}^+(n; \mathbb{R}) = \{\mathbf{A} \in \text{Sym}(n; \mathbb{R}); \mathbf{A} \succeq \mathbf{0}\}$$

とする. ただし

$$\mathbf{A} \succeq \mathbf{0} \iff \mathbf{x}^\top \mathbf{Ax} \geq 0 \quad (\forall \mathbf{x} \in \mathbb{R}^n)$$

である.

(ii) $A, B \in \text{Sym}(n; \mathbb{R})$ の内積を

$$\langle A, B \rangle = \sum_{i,j=1}^n a_{ij}b_{ij} \quad (A = (a_{ij}), B = (b_{ij}))$$

で定める.

与えられた $A_k, C \in \text{Sym}(n; \mathbb{R})$ ($k = 1, 2, \dots, m$) と $b_k \in \mathbb{R}^m$ に対して半正定値計画問題

$$\begin{aligned} & \text{minimize } \langle C, X \rangle \\ & \text{subject to } \langle A_1, X \rangle = b_1, \langle A_2, X \rangle = b_2, \dots, \langle A_m, X \rangle = b_m, X \succeq 0 \end{aligned} \quad (\text{F.85})$$

を考える.

双対性. まず写像 $A : \text{Sym}(n; \mathbb{R}) \rightarrow \mathbb{R}^m$ を

$$AX = [\langle A_1, X \rangle, \langle A_2, X \rangle, \dots, \langle A_m, X \rangle]^\top$$

と定める. 主問題 (F.85) を

$$\text{minimize } f(X) + g(AX) \quad (\text{F.86})$$

と書きかえる. ただし

$$f(X) = \begin{cases} \langle C, X \rangle & (X \succeq 0) \\ +\infty & (\text{その他}) \end{cases}, \quad g(y) = \begin{cases} 0 & (y = b) \\ +\infty & (\text{その他}) \end{cases} \quad (\text{F.87})$$

である. このとき $z \in \mathbb{R}^m$ に対し

$$g^*(z) = \sup_{y \in \mathbb{R}^m} \{ \langle z, y \rangle - g(y) \} = \langle z, b \rangle \quad (\because y \neq b \text{ のとき } g(y) = +\infty) \quad (\text{F.88})$$

となる. さらに $W \in \text{Sym}(n; \mathbb{R})$ に対して

$$f^*(W) = \sup_{X \in \text{Sym}(n; \mathbb{R})} \{ \langle W, X \rangle - f(X) \} \quad (\text{F.89})$$

$$= \sup_{X \succeq 0} \{ \langle W - C, X \rangle \} \quad (\because X \succeq 0 \text{ でなければ } f(X) = +\infty) \quad (\text{F.90})$$

$$= \begin{cases} 0 & (W \preceq 0) \\ +\infty & (\text{その他}) \end{cases} \quad (\text{F.91})$$

となる。ただし

$$W \preceq C \iff C - W \succeq 0$$

である。さらに $\forall X \in \text{Sym}(n; \mathbb{R})$ と $\mathbf{y} = (y_1, y_2, \dots, y_m)^\top \in \mathbb{R}^m$ に対し

$$\begin{aligned} \langle \mathbf{A}^\top \mathbf{y}, \mathbf{X} \rangle &= \langle \mathbf{y}, \mathbf{A}\mathbf{X} \rangle = \left\langle \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \begin{bmatrix} \langle \mathbf{A}_1, \mathbf{X} \rangle \\ \langle \mathbf{A}_2, \mathbf{X} \rangle \\ \vdots \\ \langle \mathbf{A}_m, \mathbf{X} \rangle \end{bmatrix} \right\rangle = \sum_{k=1}^m y_k \langle \mathbf{A}_k, \mathbf{X} \rangle \\ &= \left\langle \sum_{k=1}^m y_k \mathbf{A}_k, \mathbf{X} \right\rangle \end{aligned}$$

から

$$\mathbf{A}^\top \mathbf{y} = \sum_{k=1}^m y_k \mathbf{A}_k$$

がわかる。

Fenchel 双対問題.

$$f^*(\mathbf{A}^\top \mathbf{y}) - g^*(-\mathbf{y})$$

を最大化することは半正定値双対問題

$$\text{minimize } \langle \mathbf{y}, \mathbf{b} \rangle \quad \text{subject to } \sum_{k=1}^m y_k \mathbf{A}_k \preceq C (\iff f^*(\mathbf{A}^\top \mathbf{y}) = 0) \quad (\text{F.92})$$

となる。これは線型計画法の双対問題のアプローチとなる。ただし制約は対称行列に対する不等式制約である。さらに主問題 (F.85) は対称行列 \mathbf{X} が未知であり、双対問題 (F.86) はベクトル $\mathbf{y} \in \mathbb{R}^m$ が未知である。

定理 F.62. 主問題 (F.85) の実行可能解 $\bar{\mathbf{X}}$ で

$$\bar{\mathbf{X}} \succeq 0 \quad (\text{F.93})$$

をみたすものが存在したとする。さらに写像

$$\mathbf{A} : \text{Sym}(n; \mathbb{R}) \longrightarrow \mathbb{R}^m$$

は全射とする. このとき主問題 ^{leq:con-85}(F.85) の解を X^* としたとき双対問題 ^{leq:con-86}(F.86) の実行解 y^* が存在し

$$\langle C, X^* \rangle = \langle b, y^* \rangle$$

をみたす.

Proof. ^{leq:con-87}(F.92) を考慮すると, 十分小さな正数 $\delta > 0$ が存在して

$$\langle Y, Y \rangle \leq \lambda^2 \implies X = \bar{X} + Y \in \text{Sym}^+(n; \mathbb{R})$$

とできる. 写像

$$A : \text{Sym}(n; \mathbb{R}) \longrightarrow \mathbb{R}^m$$

は全射なので, ある正数 $\eta > 0$ が存在して

$$A(\mathbb{B}_{\text{Sym}}(\mathbf{0}, 1)) \supset \mathbb{B}_{\mathbb{R}^m}(\mathbf{0}, \lambda) = \{y \in \mathbb{R}^m; |y|_2 \leq \lambda\}$$

とできる. ただし $\mathbb{B}_{\text{Sym}}(\mathbf{0}, 1) = \{A \in \text{Sym}(n; \mathbb{R}); \langle A, A \rangle \leq 1\}$ である. よって各 $a \in \mathbb{B}_{\mathbb{R}^m}(\mathbf{0}, \lambda\eta)$ に対して $X = \bar{X} + Y \in \text{Sym}^+(n; \mathbb{R})$ かつ $Y \in \mathbb{B}_{\text{Sym}}(\mathbf{0}, \lambda)$ が存在して

$$AX = A\bar{X} + AY = b - a$$

となる.

関数 f, g の定義 ^{leq:con-86a}(F.87) から

$$\text{dom } f = \text{Sym}^+(n; \mathbb{R}), \quad \text{dom } g = \{b\}$$

である. したがって $\lambda = \delta\eta$ かつ $|a|_2 \leq \delta$ のとき ^{leq:con-83}(F.83) から

$$a = b - AX$$

と書ける. 結局双対条件 ^{leq:con-82}(F.82) をみたす. すると値関数

$$v(a) = \inf_{X \in \text{Sym}(n; \mathbb{R})} \{f(X) + g(AX + a)\}$$

は原点近くで有限の値を取るので, $\partial v(\mathbf{0}) \neq \emptyset$ となる. そこで

$$-y^* \in \partial v(\mathbf{0})$$

を取る. このとき $\forall X \in \text{Sym}(n; \mathbb{R})$ と $\mathbf{a} \in \mathbb{R}^m$ に対して

$$\begin{aligned} f(\mathbf{X}^*) + g(\mathbf{A}\mathbf{X}) &= v(\mathbf{0}) \leq v(\mathbf{a}) + \langle \mathbf{a}, \mathbf{y}^* \rangle \\ &\leq f(\mathbf{X}) + g(\mathbf{A}\mathbf{X} + \mathbf{a}) + \langle \mathbf{a}, \mathbf{y}^* \rangle \\ &= -\left(\langle \mathbf{A}^\top \mathbf{y}, \mathbf{X} \rangle - f(\mathbf{X})\right) - \left(-\langle \mathbf{y}, \mathbf{A}\mathbf{X} + \mathbf{a} \rangle - g(\mathbf{A}\mathbf{X} + \mathbf{a})\right) \end{aligned}$$

を得る. 上の式の最右辺を \mathbf{a} に関して infimum を取ると

$$f(\mathbf{X}^*) + g(\mathbf{A}\mathbf{X}^*) \leq -\left(\langle \mathbf{A}^\top \mathbf{y}, \mathbf{X} \rangle - f(\mathbf{X})\right) - g^*(-\mathbf{y}^*)$$

を得る. さらに上の式の右辺を \mathbf{X} に関して infimum を取ると

$$f(\mathbf{X}^*) + g(\mathbf{A}\mathbf{X}^*) \leq -f^*(\mathbf{A}^\top \mathbf{y}^*) - g^*(-\mathbf{y}^*)$$

を得る. 逆向きの不等式は常に成立するので,

$$f(\mathbf{X}^*) + g(\mathbf{A}\mathbf{X}^*) = -f^*(\mathbf{A}^\top \mathbf{y}^*) - g^*(-\mathbf{y}^*) \quad (\text{F.94})$$

を得る. $\frac{\text{eq:con-86}}{(\text{F.87})} - \frac{\text{eq:con-86c}}{(\text{F.91})}$ から

$$\begin{aligned} f(\mathbf{X}^*) &= \langle \mathbf{C}, \mathbf{X}^* \rangle, \quad g(\mathbf{A}\mathbf{X}^*) = \mathbf{0} (\iff \mathbf{A}\mathbf{X}^* = \mathbf{b}), \\ f^*(\mathbf{A}^\top \mathbf{y}^*) &= \mathbf{0} (\iff \mathbf{A}^\top \mathbf{y} \preceq \mathbf{C}), \quad g^*(-\mathbf{y}^*) = -\langle \mathbf{y}^*, \mathbf{b} \rangle \end{aligned}$$

となるので, これらを $\frac{\text{eq:con-90}}{(\text{F.94})}$ に代入すると

$$\langle \mathbf{C}, \mathbf{X}^* \rangle = \langle \mathbf{y}^*, \mathbf{b} \rangle$$

がわかる. □

11 最適化アルゴリズム

この節では, 3 つの基本的なアルゴリズムである最急降下法 (steepest-descent method), Newton 法, 共役勾配法 (conjugate-gradient method) について説明

をする。最初の 2 つのアルゴリズムに基づいて多くの最適化がなされている。最急降下法は収束のスピードは遅く実用的ではないが、簡単で信頼性が高い。Newton 法は非線型方程式の解を見つける一般的な方法である。しかし解の近傍に初期値がないと停留点に収束しない場合がある。共役勾配法は凸 2 次計画法における最適化に用いると効果を発揮する。

11.1 多変数関数の微分

定義 F.63. $U \subset \mathbb{R}^d$ は空でない開部分集合とし, $f: U \rightarrow \mathbb{R}$ は関数とする。 $\mathbf{x} = (x_1, x_2, \dots, x_d)^\top \in U$ に対し, 極限

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) := \lim_{t \rightarrow 0} \frac{f(x_1, \dots, x_{i-1}, x_i + t, x_{i+1}, \dots, x_d) - f(\mathbf{x})}{t} \quad (i = 1, 2, \dots, d)$$

が存在するとき, $\frac{\partial f}{\partial x_i}(\mathbf{x})$ 点 \mathbf{x} における x_i に関する f の偏微分係数という。すべての変数に関して偏微分係数が存在するとき, ベクトル

$$\nabla f(\mathbf{x}) := \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_d}(\mathbf{x}) \right)^\top$$

は f の勾配 (gradient) という。

$\mathbf{b} = (b_1, b_2, \dots, b_d)^\top \in \mathbb{R}^d$ とし

$$\mathbf{e}_i := \underbrace{(0, 0, \dots, 0, 1, 0, \dots, 0)}_{i-1}^\top \quad (i = 1, 2, \dots, d)$$

とする。すると

$$\mathbf{b} = b_1 \mathbf{e}_1 + b_2 \mathbf{e}_2 + \dots + b_d \mathbf{e}_d$$

となる。すると

定義 F.64. 方向 $\mathbf{d} \in \mathbb{R}^d$ に沿った点 $\mathbf{x} \in U$ における f の方向微分を

$$f'(\mathbf{x}; \mathbf{d}) = \lim_{t \searrow 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}$$

で定める。ただし $t > 0$ が 0 に近づくとき上の式の右辺の極限が存在するときに方向微分を定める。

明らかに $\alpha \geq 0$ に対して

$$\dot{f}(x; \alpha d) = \alpha \dot{f}(x; d)$$

となる。したがって $\dot{f}(x; -d) = -\dot{f}(x; d)$ であるので

$$\dot{f}(x; d) = \lim_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t}$$

となる。実際

$$\dot{f}(x; d) = -\dot{f}(x; -d) = -\lim_{t \searrow 0} \frac{f(x - td) - f(x)}{t} = \lim_{s \nearrow 0} \frac{f(x + sd) - f(x)}{s}$$

よりわかる。

lf:alg-3

定義 F.65. 関数 $f : U \rightarrow \mathbb{R}$ は点 $x \in U$ で Gâteaux 微分可能であるとは、任意の方向 $d \in \mathbb{R}^d$ に対して、方向微分 $\dot{f}(x; d)$ は存在し、それが d の線型関数であるときをいう。 f が U の任意の点 x で Gâteaux 微分可能のとき、 f は U 上で Gâteaux 微分可能という。

f が点 x において Gâteaux 微分可能のとき、 $d = d_1 e_1 + d_2 e_2 + \cdots + d_d e_d$ であり、 $\dot{f}(x; d)$ は d の線型関数なので

$$\begin{aligned} \dot{f}(x; d) &= \dot{f}(x; d_1 e_1 + d_2 e_2 + \cdots + d_d e_d) \\ &= d_1 \dot{f}(x; e_1) + d_2 \dot{f}(x; e_2) + \cdots + d_d \dot{f}(x; e_d) \\ &= \langle d, \nabla f(x) \rangle \end{aligned}$$

となる。

lf:alg-4

定義 F.66. 関数 $f : U \rightarrow \mathbb{R}$ は点 x で Féchet 微分可能であるとは、ある線型関数 $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$, $\ell(x) = \langle \ell, x \rangle$, $\ell \in \mathbb{R}^d$ が存在して

$$\lim_{|h|_2 \rightarrow 0} \frac{f(x + h) - f(x) - \langle \ell, h \rangle}{|h|_2} = 0 \quad (\text{F.95})$$

eq:alg-

が成り立つときをいう。

ベクトル $o(h) \in \mathbb{R}^d$ は

$$\lim_{|h|_2 \rightarrow 0} \frac{o(h)}{|h^*|_2} = 0$$

をみたまものとする。この記号を用いると f が点 x で F chet 微分可能であることは

$$f(x + h) = f(x) + \langle \ell, h \rangle + o(h) \tag{F.96}$$

eq:alg-

と書けることがわかる。

$h = te_i (i = 1, 2, \dots, d), t \neq 0$ と取ると (F.95) は、 $\ell = (\ell_1, \ell_2, \dots, \ell_d)^\top$ としたとき

$$\lim_{t \rightarrow 0} \frac{f(x + te_i) - f(x) - t\ell_i}{t} = 0$$

と書きなおすことができる。このとき (F.96) は

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(h)$$

となる。

定理 F.67. $U \subset \mathbb{R}^d$ は空でない開部分集合とし、 $f : U \rightarrow \mathbb{R}$ は点 $x \in U$ で F chet 微分可能とする。このとき f は点 x で G teaux 微分可能である。

定義 F.68. $x, y \in \mathbb{R}^d$ は点とする。 $[x, y]$ は x と y の間の線分とする。すなわち

$$[x, y] = \{x + t(y - x); 0 \leq t \leq 1\}$$

である。同様に

$$(x, y) = \{x + t(y - x); 0 < t < 1\}$$

とする。 $x = y$ のとき

$$[x, y] = (x, y) = \{x\}$$

と定める。

補題 F.69. $U \subset \mathbb{R}^d$ は空でない開部分集合とし、 $f : U \rightarrow \mathbb{R}$ は U 上の G teaux 微分可能とする。 $x, y \in U$ は異なる点で $[x, y] \subset U$ とする。このとき点 $z \in (x, y)$ が存在して

$$f(y) = f(x) + \langle \nabla f(z), y - x \rangle$$

とすることができる。

Proof. $t \in \mathbb{R}$ に対し

$$h(t) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$$

とおく. f は Gâteaux 微分可能なので, $h(t)$ は微分可能で

$$\dot{h}(t) = \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle$$

となる. 1 変数の中間値の定理から $0 < \bar{t} < 1$ が存在して

$$h(1) - h(0) = \dot{h}(\bar{t})$$

となる. $z = \mathbf{x} + \bar{t}(\mathbf{y} - \mathbf{x})$ とおく. $h(0) = f(\mathbf{x})$, $h(1) = f(\mathbf{y})$ なので

$$f(\mathbf{x}) - f(\mathbf{y}) = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

となる. □

m:alg-8

定理 F.70. $U \subset \mathbb{R}^d$ は空でない開部分集合とし, $f: U \rightarrow \mathbb{R}$ は関数とする. f は点 $\mathbf{x}_0 \in U$ で Gâteaux 微分可能で, $\partial f / \partial x_i$ ($i = 1, 2, \dots, d$) は点 \mathbf{x}_0 で連続微分可能ならば f は点 \mathbf{x}_0 で Féchet 微分可能である.

Proof. 補題 ^{Lem:alg-7}F.69 より任意の $\mathbf{h} \in \mathbb{R}^d$ に対して $\bar{\mathbf{x}} \in (\mathbf{x}, \mathbf{x} + \mathbf{h})$ が存在し

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle = \langle \nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}_0), \mathbf{h} \rangle$$

と書ける. ここで ∇f は \mathbf{x}_0 で連続であるので

$$\nabla f(\bar{\mathbf{x}}) - \nabla f(\mathbf{x}_0) \rightarrow 0 \quad (\mathbf{h} \rightarrow \mathbf{0}).$$

したがって

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{h} \rangle = o(\mathbf{h})$$

となり, f は点 \mathbf{x}_0 で Féchet 微分可能となる. □

12 勾配降下法

12.1 降下方向

alg-9 定義 F.71. $x \in M$ は $\nabla f(x) \neq 0$ とする. したがって点 x は f の停留点 (critical point) ではない. f の降下方向 (descent direction) は $d \in \mathbb{R}^d (\neq 0)$ で, ある $\bar{t} > 0$ が存在して

$$f(x + td) < f(x) \quad (0 < \forall t < \bar{t})$$

をみたすときをいう. したがって十分小さなステップ幅 $t > 0$ に対して半直線 $\mathbb{R}_{++} := \{x + td; t > 0\}$ 上で f は厳密に減少している.

alg-10 補題 F.72. $x \in U$ は f の停留点ではないとし, $d \in \mathbb{R}^d$ は $d \neq 0$ とする. $\langle \nabla f(x), d \rangle < 0$ のとき d は点 x における f の降下方向となる.

逆に d が点 x における降下方向としたとき $\langle \nabla f(x), d \rangle \leq 0$ となる.

Proof. f は Gâteaux 微分可能なので

$$f(x + td) = f(x) + t\langle \nabla f(x), d \rangle + o(t) \quad (\text{F.97})$$

となる. d は

$$\langle \nabla f(x), d \rangle < 0$$

をみたすので十分小さな $t > 0$ に対して

$$f(x + td) < f(x)$$

となる.

逆に小さな $t > 0$ に対して

$$f(x + td) < f(x)$$

のとき, $t \downarrow 0$ とすると

$$\langle \nabla f(x), d \rangle \leq 0$$

となる. □

12.2 最急降下方向

(F.97) から十分小さな $t > 0$ に対して

$$f(\mathbf{x} + t\mathbf{d}) \approx f(\mathbf{x}) + t\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle$$

と書ける. $t > 0$ を固定する. $\mathbf{d} \in \mathbb{R}^d$ で $|\mathbf{d}|_2 = 1$ に対して $\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle$ を最小にするものが f を最小にする方向である.

$\nabla f(\mathbf{x})$ と \mathbf{d} の角度を θ とする. このとき

$$\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle = |\nabla f(\mathbf{x})|_2 \cdot |\mathbf{d}|_2 \cos \theta = |\nabla f(\mathbf{x})|_2 \cos \theta$$

となるので $\cos \theta = -1$ と取ると

$$\mathbf{d} := -\frac{\nabla f(\mathbf{x})}{|\nabla f(\mathbf{x})|_2}$$

となる. このことより方向 $\mathbf{d} = -\nabla f(\mathbf{x})$ を点 \mathbf{x} における f の最急降下方向 (steepest descent direction) とよぶ.

12.3 ステップ幅の選択方法

$k = 1, 2, \dots$ とする. \mathbf{x}_k における降下方向 \mathbf{d}_k が選択されると, ステップ幅 t_k を適切に選択して, 次の点 \mathbf{x}_{k+1} を

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t\mathbf{d}_k \quad (t_k > 0)$$

と定める. この方法を直接降下法と呼ぶ. t_k を適切に探索しないと \mathbf{x}_k が停留点に収束する保証が得られないことがある.

代表的なステップ幅の選択方法である Armijo 法, Goldstein 法, Wolfe 法を説明する.

Armijo 法: $s > 0, 0 < \beta < 1, 0 < \sigma < 1$ を固定する. $i = 0, 1, 2, \dots$ に対して

$$f(\mathbf{x}_k) - f(\mathbf{x}_k + \beta^i s\mathbf{d}_k) \geq -\sigma(\beta^i s)\langle \nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle \quad (\text{F.98})$$

をテストする.

$i = 0$ とすると

$$f(\mathbf{x}_k) - f(\mathbf{x}_k + s\mathbf{d}_k) \geq -\sigma\langle \nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle \quad (\text{F.99})$$

をテストすることになる. 不等式 (F.99)^{eq:alg-5} が成り立てば

$$\mathbf{x}_{k+1} = \mathbf{x}_k + s\mathbf{d}_k$$

とする. (F.99)^{eq:alg-5} が成立しなければ, ステップ幅が大きすぎるので, $i = 1$ の (F.98)^{eq:alg-4} をテストする.

$$f(\mathbf{x}_k) - f(\mathbf{x}_k + \beta s\mathbf{d}_k) \geq -\beta s \langle \nabla f(\mathbf{x}_k), \beta s\mathbf{d}_k \rangle$$

である. i をおおくしていくとある i で (F.98)^{eq:alg-4} が必ず成立する. このことは以下の議論からわかる. i が十分おおいとき, $t := \beta^i s > 0$ は十分小さい. したがって (F.98)^{eq:alg-4} において

$$f(\mathbf{x}_k + t\mathbf{d}_k) < f(\mathbf{x}_k) + \sigma t \langle \nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle$$

であったとする.

$$f(\mathbf{x}_k + \mathbf{d}_k) = f(\mathbf{x}_k) + t \langle \nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle + o(t)$$

と比較すると

$$(\sigma - 1) \langle \nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle + \frac{o(t)}{t} < 0$$

となる. $t \rightarrow 0$ とすると

$$(\sigma - 1) \langle \nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle \leq 0$$

となり, $\sigma - 1 < 0$ なので

$$\langle \nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle < 0$$

と矛盾することから (F.98)^{eq:alg-4} をみたま i が存在することがわかる.

$k = 1, 2, \dots, \{\mathbf{x}_k\}_{k=1}^{\infty}$ は最急降下法により求められる点列とし, \mathbf{d}_k を \mathbf{x}_k における f の降下方向とする. $-\nabla f(\mathbf{x}_k)$ と \mathbf{d}_k の角度を θ_k と書く. するとある $\epsilon > 0$ が存在して

$$\cos \theta_k := \frac{\langle -\nabla f(\mathbf{x}_k), \mathbf{d}_k \rangle}{|\nabla f(\mathbf{x}_k)|_2} \in (\epsilon, 1] \quad |\mathbf{d}_k|_2 = 1 \quad (\text{F.100})$$

となる.

eq:alg-

定理 F.73. $\{x_k\}_{k=1}^{\infty}$ を降下法によって生成された点列で

$$x_{k+1} = x_k + \alpha_k d_k$$

とする. ただし d_k と (F.98) をみだし, α_k はパラメータ s, β, σ に対する Armijo 条件をみたしているとする. $\lim_{k \rightarrow \infty} x_k = x^*$ としたとき x^* は f の停留点になる. すなわち

$$\nabla f(x^*) = 0$$

をみたす.

Proof. 必要であれば $\{d_k\}_{k=1}^{\infty}$ の部分列を取ることで $\{d_{k_i}\}_{i=1}^{\infty}$ はある点 d^* に収束する. $|d^*|_2 = 1$ である. i 段階において Armijo 条件が成立していたとする.

$$\begin{aligned} f(x_{k_i}) - f(x_{k_i} + \alpha_{k_i} d_{k_i}) &\geq -\sigma \alpha_{k_i} \langle \nabla f(x_{k_i}), d_{k_i} \rangle \\ &= \sigma \alpha_{k_i} |\nabla f(x_{k_i})|_2 \cos \theta_{k_i} \\ &\geq \epsilon \sigma \alpha_{k_i} |\nabla f(x_{k_i})|_2 \end{aligned} \quad (\text{F.101})$$

となる. α_{k_i} で Armijo 条件が成立しているので

$$f(x_{k_i}) - f\left(x_{k_i} + \frac{\alpha_{k_i}}{\beta} d_{k_i}\right) < -\sigma \frac{\alpha_{k_i}}{\beta} \langle \nabla f(x_{k_i}), d_{k_i} \rangle \quad (\text{F.102})$$

となる. $\nabla f(x^*) \neq 0$ と仮定する. $x_{k_i} \rightarrow x^* (i \rightarrow \infty)$ で $\{f(x_{k_i})\}_{i=1}^{\infty}$ は減少列なので

$$f(x_{k_i}) \searrow f(x^*)$$

となり, (F.101) の左辺は $i \rightarrow \infty$ のとき 0 に収束する. なぜならば $f(x_{k_i}) - f(x_{k_{i+1}}) \rightarrow 0 (i \rightarrow \infty)$ からわかる. さらに $\alpha_{k_i} = |x_{k_i} - x_{k_{i+1}}|_2 \rightarrow 0 (i \rightarrow \infty)$ なので, (F.101) の他の項も $i \rightarrow \infty$ のときに 0 に収束する. 仮定 (F.100) の仮定より

$$\langle \nabla f(x^*), d^* \rangle < 0, \quad \alpha_{k_i} = |x_{k_i} - x_{k_{i+1}}|_2 \rightarrow 0 (i \rightarrow \infty) \quad (\text{F.103})$$

□

となる. i が十分大きいとき x_{k_i} においてバックトラックが必要となることがわかる. (F.102) から矛盾を導く. 中間値の定理から $z_{k_i} \in (x_{k_i}, x_{k_i} + \frac{\alpha_{k_i}}{\beta} d_{k_i})$

が存在して

$$f(\mathbf{x}_{k_i}) - f\left(\mathbf{x}_{k_i} + \frac{\alpha_{k_i}}{\beta} \mathbf{d}_{k_i}\right) = -\frac{\alpha_{k_i}}{\beta} \langle \nabla f(\mathbf{z}_{k_i}), \mathbf{d}_{k_i} \rangle$$

をみたく。 (F.102) ^{leq:alg-8} を考慮すると

$$-\frac{\alpha_{k_i}}{\beta} \langle \nabla f(\mathbf{z}_{k_i}), \mathbf{d}_{k_i} \rangle < -\sigma \frac{\alpha_{k_i}}{\beta} \langle \nabla f(\mathbf{x}_{k_i}), \mathbf{d}_{k_i} \rangle$$

を得る。 $i \rightarrow \infty$ のとき $\lim \mathbf{z}_{k_i} = \mathbf{x}^*$ となるので、上の不等式より

$$\langle \nabla f(\mathbf{x}^*), \mathbf{d}^* \rangle \geq 0$$

を得る。これは (F.103) ^{leq:alg-9} と矛盾。 よって $\nabla f(\mathbf{x}^*) = \mathbf{0}$ が示せた。

□

関連図書

- I [1] 泉正己 (2021), 『数理科学のための関数解析』, サイエンス社.
- OK [2] 岡留剛 (2022), 『機械学習 1』, 共立出版.
- KA [3] 加藤賢吾 (2015), 『数理統計 講義ノート』, Available at
https://drive.google.com/file/d/0B7C_CufYq6j6ekFINVhTS09lZzQ/view?resourcekey=0-zaYo9XudjhK_K6kMfpq-XQ (accessed 2022/09/22).
- KU [4] 久保川達也 (2017), 『現代数理統計学の基礎』, 共立出版.
- SIH [5] 瀬戸道生・伊吹竜也・畑中建志 (2021), 『機械学習のための関数解析入門』, 内田老鶴圃.
- T1996 [6] 高橋陽一郎 (1996), 『実関数と Fourier 解析 1』, 岩波書店.
- TSU [7] 堤正義 (2012), 『逆問題－理論および数理科学への応用－』, 朝倉書店.
- HARA [8] 原啓介 (2017), 『測度・確率・ルベーグ積分』, 講談社.
- FU [9] 藤岡敦 (2021), 『入門情報幾何』, 共立出版.
- Fu2010 [10] 福水健次 (2010), 『カーネル法入門』, 朝倉書店.
- YO [11] 吉田伸生 (2021), 『ルベーグ積分入門』, 日本評論社.
- B [12] Bisgard, James (2021), Analysis and Linear Algebra: The Singular Value Decomposition and Applications, AMS.
- C2019 [13] Christensen, Ronald (2019), Advanced Linear Modelling, 2nd ed., Springer.

- DO** [14] Dobson, Annette J. (2008), 『一般化線形モデル入門』原著 第2版, 田中豊・森川敏彦・山中竹春・富田誠訳, 共立出版.
- EVANS** [15] Evans, Wrence Craig (2021), *Mathematical Methods for Optimization: Finite Dimensional Optimization*, Available at <https://math.berkeley.edu/~evans/math%20170%20notes.pdf> (accessed 2022/09/22).
- GA** [16] Garrity, Thomas, A. (2002), *All the Mathematics You Missed*, Cambridge University Press.
- ISL** [17] James, D., Witten D., Hastie, T., and Tibshirani, R. (2021), *An Introduction to Statistical Learning*, 2nd ed., Springer. Available at <https://link.springer.com/book/10.1007/978-1-0716-1418-1> (大学内).
- LE** [18] Lederer, Johannes (2021), *Fundamentals of high-dimensional Statistics with Exercises and R Labs*, Springer text in Statistics. Available at <https://link.springer.com/book/10.1007/978-3-030-73792-4>
- KS** [19] Kaipio, Jari and Somersalo, Erkki (2005), *Statistical and Computational Inverse Problems*, Springer.
- KBTV** [20] Kroses, Dirk P., Botev, Zdravko I., Taimre, Thomas, and Vaisman, Radislav (2020), *Data Science and Machine Learning*, Chapman & Hall/CRC.
- RI** [21] Rigollet, Philippe and Hütter, Jan-Christian (2019), *High dimensional Statistics*, Lecture note. Available at <https://math.mit.edu/~rigollet/PDFs/RigNotes17.pdf>
- SS** [22] Saitoh, Saburo and Sawano, Yoshihiro (2016), *Theory of Reproducing Kernels and Applications*, Springer.
- W** [23] Wasserman, Larry (2004), *All of Statistics*, Springer.