学習機械の予測誤差

池田和司

(1)

学習機械の予測誤差

池田 和司

目次

第 1 章

序論

人間に限らず多くの生物は，環境から情報を得て自らを変化させ，より高い能力を身につけることができる．これを学習と呼ぶ．この能力を機械にも持たせることができれば，機械は環境に適応して内部のパラメータを自動的に調整するので，人間が環境を事細かに調べる必要がなくなる．また，環境が変化した場合でも，柔軟に対応することができる．機械に学習を行わせる時に問題となるのは，学習した機械がどれ程の能力をもっているか，ということである．本論は，機械をダイナミクスのない入出力装置に限定し，機械が学習する時の能力の向上のスピードについて論ずるものである．ここでいう入出力装置とは，入力が与えられると出力を出す機械のことである．入出力装置は確定的機械と確率的機械に大別される．確定的機械は与えられた入力と内部のパラメータによって出力が一意に決定されるもので，確率的機械は与えられた入力と内部のパラメータによって出力の確率分布が決定されるものである．確定的機械は確率的機械の特別なものと見ることができる．

今，機械に実現させたい入出力関係が指定されているとしよう．パラメータをうまく選ぶとこの入出力関係を実現することができる時，この入出力関係は学習可能であるといい，できないときは学習不可能であるという．本論では入出力関係は学習可能であるとし，この入出力関係を実現するパラメータを真のパラメータと呼ぶ．真のパラメータをもつ機械を真の機械と呼ぶ．また，真の機械によって作られた入力と出力の組を例題と呼ぶ．今，機械にいくつかの例題とテスト用の新規の入力を一つ与える．機械は例題の入出力を見て，新規の入力に対する真の機械の出力を予測する．この機械を学習機械と呼ぶことにする．学習機械が例題を有効に利用すれば，学習機械の予測は与えられる例題の数が多いほど真の機械の出力と一致する確率が高くなる．このように予測を改善することを学習と呼び，具体的な予測方法を学習アルゴリズムと呼ぶ．多くの学習アルゴリズムでは，例題から直接テスト入力に対する出力を予測するのではなく，例題からパラメータを推定し，そのパラメー

タをもつ機械の出力を予測とする．この場合には学習アルゴリズムは，パラメータの逐次推定法とみなすことができる．

学習に関する研究は，心理学の分野で始まった．そして動物の脳に関する解剖学的，生理学的知見から，生体の神経系のモデルとして神経回路網が考案された．神経回路網は，比較的単純な入出力関係を持つ多数の神経素子が相互に結合して信号をやりとりするネットワークであり，生体の神経系ではシナプスの結合荷重の変化で記憶や学習が起こると考えられていた．その学習アルゴリズムを初めて提唱したのは Hebb [10] である．Hebb の提唱した学習アルゴリズムは，出力がアクティブである二つのニューロン間の結合は強められる，という局所的で単純なものであったが，現在でも，いわゆる教師無し学習のアルゴリズムの基本となっている．

1961 年に Rosenblatt [16] は，例題からの学習アルゴリズムとして単純パーセプトロン学習を提案した．単純パーセプトロンは神経素子の簡単なモデルであり，本質的には空間を超平面で 2 分するだけの簡単な機械であるが，Block [6] がその収束性を示したことにより理論的な裏付けを得た．しかし単純パーセプトロンの能力は線形分離に限られており，また複雑な回路を実際の工学系として実現するだけの技術もなかったため，神経回路網の研究はその後活気を失った．現在の神経回路網の研究の隆盛のきっかけとなったのは，Rumelhart *et al.* [17] による一般の多層ネットワークの学習法の提案であった．この学習法はエネルギー関数の最急降下法の一種であり，Back-Propagation 学習と呼ばれている．Back-Propagation 学習は以前の研究 [1] の枠組みに含まれており，さらに学習は極小点で止まるためエネルギー関数は必ずしも最小化されないという欠点も指摘されていた．にもかかわらず，計算機の能力の向上により画像処理，音声認識，システム制御等様々な実例に応用することができ，さらにそれらにおいて有効性をみせたことが多くの研究者の興味をひく原因となった．

その他にも，エネルギー関数の極小点に捕らわれないように温度パラメータを用いる焼きなまし法 [12] など，学習方法を改善する研究が行われたが，それに並行して学習による汎化能力の評価の研究も進められた．汎化能力とは，例題として与えられなかった入力も含め，新規の入力に対してどれだけ正しい出力を出せるかを評価したものである．学習方法の研究が，与えられた例題を機械にいかに学習させるかを問題とするのに対し，汎化能力の研究は，与えられる例題の個数と機械の能力の関係を問題にしている．一般に学習機械は与えられる例題の数が多いほど，汎化能力は増大する．逆に例題数が少ない場合には，学習機械の汎化能力には限界がある．学習機械を実用に耐えうるものにするためには，どれだけの例題を与えればどれだけの汎化能力が期待できるのかを明らかにしておくことは重要である．

Baum and Haussler [5] は例題数と汎化能力の関係について，Valiant [19] が提案したPAC 学習と呼ばれる枠組みに機械の複雑さを表す VC 次元 [20] を導入し，学習に必要な例題数のバウンドを求めた．しかしこの枠組みは最悪評価を行うため，評価値が実際値よりもかなり大きくなり，実用的でないという批判もある．

最悪評価でなく平均的な学習の振舞いを示す基準として，平均予測損失の評価がある．この枠組では，学習機械に例題とテスト用の入力を与え，テスト入力に対する真の出力を予測させ，予測の良さを予め定めた損失関数の値によって評価する．テスト入力をランダムに選んだ時の損失関数の期待値を予測損失と呼ぶ．予測損失は例題に依存するので，例題も確率的に選ばれるとしてその平均を評価したものが，平均予測損失である．この時，平均予測損失は例題数の関数となり，これは学習曲線と呼ばれる．平均予測損失が例題数に反比例することは，2 分割機械の特殊な場合について Cover [7] が，一般の場合について Murata *et al.* [14] が示した．

以下では，単純パーセプトロンを一般化したものである，2 分割機械について考察する．2 分割機械は，入力とパラメータによって $+1$ か -1 のいずれかを出力する機械である．2 分割機械について最も自然に導かれる損失関数の一つとして，誤り確率がある．誤り確率は，テスト入力に対する出力の予測が誤りである確率である．誤り確率のテスト入力に関する平均を予測誤差と呼び，予測誤差の例題に関する平均を平均予測誤差という．

平均予測誤差の評価の研究にはいくつかのアプローチがあるが，いずれも真の値を求めるには至っていない．一つは統計力学の手法を用いるものである [9,13,15,18]．この手法では統計力学とニューラルネットワークのアナロジーから統計力学の一手法であるレプリカ法を用いて，入力の次元 m，例題数 t が $m \to \infty, \alpha = t/m \to \infty$ という極限において平均予測誤差が評価された．しかしこのアプローチは，レプリカ法が数学的に裏付けのある計算手法ではないことと，入力の次元が小さい場合に適応できないという欠陥がある．

また，Amari *et al.* [2–4] は Bayes 統計に基づき，予測誤差の代わりに予測誤差の上のバウンドである予測エントロピを評価した．このアプローチにより平均予測誤差の上のバウンドは求められたが，平均予測誤差が明示的に計算される場合について比較すると，平均予測エントロピは平均予測誤差とは一致しないことがわかっている．

本論文は，上記のいずれとも異なるアプローチで，平均予測誤差の評価を行うものである．

第 2 章では，一般の学習機械について成り立つ事柄を述べる．まず，パラメータの事後分布を用いて，Gibbs アルゴリズム，Bayes アルゴリズム，重心アルゴリズ

ム，最尤推定アルゴリズムの各アルゴリズムを定義する．そして，Gibbs アルゴリズム及び Bayes アルゴリズムの予測誤差が，事後分布に関連して定義される全確率を用いて表されることを示す．

第 3 章では，確定的機械の学習について言及する．まず，パラメータが冗長でない確定的機械は漸近的に単純パーセプトロンとみなすことができることを示す．次に，例題の入力及びテスト入力が超球面上の一様分布に従って選ばれる場合について，単純パーセプトロンの許容領域の性質を統計幾何学的に評価する．確定的機械は，入力とパラメータによって出力が一意に定まるので，与えられた例題の入力に対して例題の出力と同じ出力を出すことができるパラメータの集合が存在する．この集合を許容領域と呼ぶ．許容領域は例題が与えられると一意に定まる．パラメータの事後分布の確率密度は，許容領域に含まれるパラメータは事前分布に比例し，許容領域に含まれないパラメータは 0 であるので，許容領域と学習アルゴリズムは密接な関係がある．この関係を利用して，許容領域の統計幾何学的性質から単純パーセプトロンの平均予測誤差のバウンドを求める．このバウンドは，従来よりも良いバウンドとなっている．また計算機実験により，学習に Gibbs アルゴリズムと重心アルゴリズムを用いた時の平均予測誤差を求める．さらに，パーセプトロン学習をした場合の平均予測誤差を求め，パーセプトロン学習が Gibbs アルゴリズムの近似となることを実験的に示す．

第 4 章では，確率的機械の平均予測誤差を導出する．まず，損失関数を予測誤差に限定せず，一般の場合について平均予測損失の漸近的特性を導く．ここで学習とは，損失関数の最小化によってパラメータを推定することとしておく．ここで，予測の良さを評価するのに用いる損失関数と学習のための損失関数は同一である必要はない．それぞれを，予測損失関数，学習損失関数と呼ぶ．次に，第 2 章で定義した各アルゴリズムは，学習損失関数と予測損失関数をうまく選ぶことにより，上で定義した学習，すなわち損失関数の最小化によるパラメータ推定の枠組みに帰着されることを示し，上で求めた学習特性を利用して，確率的 2 分割機械の予測誤差を求める．さらに，確率的 2 分割機械の例として，確率的パーセプトロンの予測誤差を具体的に求める．確率的パーセプトロンは，入力とパラメータの内積をパラメータとする確率分布に従って出力が決まる 2 分割機械であり，単純パーセプトロンは確率的パーセプトロンの極限であるとみなすことができる．

第 2 章

学習アルゴリズムと全確率

2.1 2 分割学習機械

2 分割機械は，入力 $\boldsymbol{x}$ に対して $+1$ か -1 のいずれかを出力する装置である．出力 y は，入力 $\boldsymbol{x}$ とパラメータ $\boldsymbol{w}$ によって定まる確率分布に従って選ばれる．入力が $\boldsymbol{x}$，パラメータが $\boldsymbol{w}$ の時に y が出力される確率密度を，$p(y|\boldsymbol{x},\boldsymbol{w})$ と表す．

以下では確率を表すのに確率密度 $p(\,\cdot\,)$ を用い，$p(\,\cdot\,|\,\cdot\,)$ で条件つき確率密度を表す．ただし特定の値を表す場合には，例えば $p(y=+1|\boldsymbol{x},\boldsymbol{w})$ のように書く．

今，学習機械に t 個の例題が与えられるとする．ここで例題とは，学習機械が学ぶべき入出力関係によって生成された入力と出力の組である．学ぶべき入出力関係は確率的であるとし，その確率密度はあるパラメータ $\boldsymbol{w}^*$ を用いて $p(y|\boldsymbol{x},\boldsymbol{w}^*)$ と表されるとする．$p(y|\boldsymbol{x},\boldsymbol{w}^*)$ という入出力関係を持った機械を真の機械と呼び，真の機械が持つパラメータ $\boldsymbol{w}^*$ を，真のパラメータと呼ぶ．以下では i 番目の例題を $\overset{i}{\xi} = (\overset{i}{\boldsymbol{x}},\overset{i}{y})$ で表し，t 番目までの例題をまとめて $\xi^{(t)}$ で表す．すなわち $\xi^{(t)} = \{(\overset{i}{\boldsymbol{x}},\overset{i}{y}), i = 1,\ldots,t\}$ である．また，$\boldsymbol{x}^{(t)} = \{\overset{i}{\boldsymbol{x}}, i = 1,\ldots,t\}$ で t 番目までの例題の入力を表し，$y^{(t)} = \{\overset{i}{y}, i = 1,\ldots,t\}$ で t 番目までの例題の出力を表す．例題の入力 $\overset{i}{\boldsymbol{x}}$ は，同一の確率分布 $p(\boldsymbol{x})$ に従って独立に選ばれるとする．

本論では Bayes の立場をとり，真のパラメータ $\boldsymbol{w}^*$ に事前確率分布 $p(\boldsymbol{w})$ を仮定する．例題の入力 $\boldsymbol{x}^{(t)}$ が固定されている時，機械のパラメータが $\boldsymbol{w}$ で，かつ出力 $y^{(t)}$ が生成される確率密度 $p\left(y^{(t)},\boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right)$ は

$$p\left(y^{(t)},\boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right) = p(\boldsymbol{w})\prod_i p\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\boldsymbol{w}\right) \tag{2.1}$$

である．ここで

$$Z_t = Z(\xi^{(t)}) = \int p\left(y^{(t)},\boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right)\,\mathrm{d}\boldsymbol{w} \tag{2.2}$$

とすれば，Bayes の定理により $\boldsymbol{w}$ の事後分布の確率密度 $p(\boldsymbol{w}|\xi^{(t)})$ は

$$p(\boldsymbol{w}|\xi^{(t)}) = \frac{p\left(y^{(t)}, \boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right)}{Z_t} \tag{2.3}$$

で与えられる．Z_t は，例題の入力が $\boldsymbol{x}^{(t)}$ である時に例題の出力が $y^{(t)}$ である確率密度 $p\left(y^{(t)}\middle|\boldsymbol{x}^{(t)}\right)$ を表している．Z_t を全確率と呼ぶことにする．

t 個の例題 $\boldsymbol{x}^{(t)}$ が与えられた学習機械に，例題の入力と同一の分布 $p(\boldsymbol{x})$ に従って選ばれた新規の入力 ${}^{t+1}\boldsymbol{x}$ を与える．この新規の入力 ${}^{t+1}\boldsymbol{x}$ をテスト入力と呼ぶ．学習機械は t 個の例題を参考にして，テスト入力 ${}^{t+1}\boldsymbol{x}$ に対して真の機械が出す出力 ${}^{t+1}y$ を予測する．この時，学習機械の予測した出力が誤りである確率を誤り確率と呼ぶ．誤り確率は，例題 $\xi^{(t)}$ とテスト入力 ${}^{t+1}\boldsymbol{x}$，真の機械の出力 ${}^{t+1}y$ の関数となる．誤り確率の，テスト入力 ${}^{t+1}\boldsymbol{x}$ と真の機械の出力 ${}^{t+1}y$ に関する平均を予測誤差と呼ぶ．今後しばしば，テスト入力 ${}^{t+1}\boldsymbol{x}$ と真の機械の出力 ${}^{t+1}y$ を $t+1$ 番目の例題と同一視する．例えば，例題 $\xi^{(t)}$ とテスト入力 ${}^{t+1}\boldsymbol{x}$，真の機械の出力 ${}^{t+1}y$ をまとめて $\xi^{(t+1)}$ と書く．以下では，$\langle\cdot\rangle_X$ で確率変数 X に関する平均を表すとする．$\langle\cdot\rangle_{\boldsymbol{x}^{(t)}}$ は $\boldsymbol{x}^{(t)} = \{{}^{i}\boldsymbol{x}, i = 1,\ldots,t\}$ に関する平均であり，$\langle\cdot\rangle_{\xi^{(t+1)}}$ は $\xi^{(t+1)} = \{({}^{i}\boldsymbol{x}, {}^{i}y), i = 1,\ldots,t+1\}$ に関する平均である．

2.2 学習アルゴリズムの定義

与えられた例題から予測を決める方法を学習アルゴリズムと呼ぶ．ここでは四つの学習アルゴリズムを定義する．これらは2種類に大別される．一方は例題 $\xi^{(t)}$ から真の機械の持つパラメータを推定し，推定パラメータ $\boldsymbol{w}_t$ を持つ機械がテスト入力 ${}^{t+1}\boldsymbol{x}$ に対して出す出力を予測とするアルゴリズムである．ここでは，Gibbs アルゴリズム，重心アルゴリズム，最尤推定アルゴリズムがこの種類に属する．もう一方は，例題 $\xi^{(t)}$ とテスト入力 ${}^{t+1}\boldsymbol{x}$ から確定的に予測を作るアルゴリズムである．ここでは，Bayes アルゴリズムがこの種類に属する．

定義 2.1 (Gibbs アルゴリズム) 事後確率分布 $p(\boldsymbol{w}|\xi^{(t)})$ に従って選んだパラメータを推定パラメータ $\boldsymbol{w}_t$ とするアルゴリズムを Gibbs アルゴリズムと呼ぶ．

Gibbs アルゴリズムの予測が y である確率密度は

$$p\left(y\middle|{}^{t+1}\boldsymbol{x}, \boldsymbol{w}_t\right), \quad \text{ただし } \boldsymbol{w}_t \sim p(\boldsymbol{w}|\xi^{(t)}), \tag{2.4}$$

である．

定義 2.2 (Bayes アルゴリズム) テスト入力 ${}^{t+1}\boldsymbol{x}$ に対する真の機械の出力 ${}^{t+1}y$ について，例題が $\xi^{(t)}$ である時に ${}^{t+1}y = +1$ である事後確率と ${}^{t+1}y = -1$ である事後確率を比較し，確率の高い方の出力を予測とするアルゴリズムを，Bayes アルゴリズムと呼ぶ．

Bayes アルゴリズムの予測は

$$y = +1 \quad \text{if} \int p\left(y = +1 \middle| {}^{t+1}\boldsymbol{x}, \boldsymbol{w}\right) p(\boldsymbol{w}|\xi^{(t)})\, \mathrm{d}\boldsymbol{w} > \frac{1}{2}, \tag{2.5}$$

$$y = -1 \quad \text{otherwise}, \tag{2.6}$$

である．

定義 2.3 (重心アルゴリズム) パラメータの事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ に関する重心を推定パラメータ $\boldsymbol{w}_t$ とするアルゴリズムを，重心アルゴリズムを呼ぶ．

重心アルゴリズムの予測が y である確率は

$$p\left(y \middle| {}^{t+1}\boldsymbol{x}, \boldsymbol{w}_t\right), \quad \text{ただし } \boldsymbol{w}_t = \int \boldsymbol{w}\; p(\boldsymbol{w}|\xi^{(t)})\, \mathrm{d}\boldsymbol{w}, \tag{2.7}$$

である．しかし一般には，パラメータの重心がパラメータ空間内にあるとは限らず，そのような場合には重心アルゴリズムは意味をなさない．

定義 2.4 (最尤推定アルゴリズム) 事後確率密度 $p(\boldsymbol{w}|\xi^{(t)})$ を最大にするパラメータを推定パラメータとするアルゴリズムを最尤推定アルゴリズムと呼ぶ．

事後確率 $p(\boldsymbol{w}|\xi^{(t)})$ を最大にするパラメータを最尤推定パラメータと呼ぶ．最尤推定アルゴリズムの予測が y である確率は

$$p\left(y \middle| {}^{t+1}\boldsymbol{x}, \boldsymbol{w}_t\right), \quad \text{ただし } \boldsymbol{w}_t = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\xi^{(t)}), \tag{2.8}$$

である．しかし一般には，$p(y|\boldsymbol{x}, \boldsymbol{w})$ を最大にするパラメータが一意に定まるとは限らず，そのような場合には最尤推定アルゴリズムは意味をなさない．

2.3 全確率と予測誤差

Gibbs アルゴリズムの予測誤差と Bayes アルゴリズムの予測誤差は，全確率 Z_t を用いて表すことができる．

定理 2.1 Gibbs アルゴリズムの平均予測誤差は全確率 Z_t を用いて，

$$\langle \mathrm{GA} \rangle_{\xi^{(t+1)}} = 1 - \left\langle \frac{Z_{t+1}}{Z_t} \right\rangle_{\xi^{(t+1)}} \tag{2.9}$$

と表される．

証明 Gibbs アルゴリズムの誤り確率を GA とすると

$$\begin{aligned}
&\langle \mathrm{GA} \rangle_{\xi^{(t+1)}} \\
&= \left\langle \int \left(1 - p\left(\overset{t+1}{y} \middle| \overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}\right)\right) p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} \right\rangle_{\xi^{(t+1)}} && (2.10)\\
&= 1 - \left\langle \frac{1}{Z_t} \int p\left(\overset{t+1}{y} \middle| \overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}\right) \prod_{i=1}^{t} p\left(\overset{i}{y} \middle| \overset{i}{\boldsymbol{x}}, \boldsymbol{w}\right) p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} \right\rangle_{\xi^{(t+1)}} \\
&= 1 - \left\langle \frac{Z_{t+1}}{Z_t} \right\rangle_{\xi^{(t+1)}} && (2.11)
\end{aligned}$$

が成り立つ． □

$\left\langle \frac{Z_{t+1}}{Z_t} \right\rangle_{\xi^{(t+1)}}$ は確率変数の商の期待値となっており，このことが予測誤差の計算が困難である原因となる．

定理 2.2 Bayes アルゴリズムの平均予測誤差は全確率 Z_t を用いて，

$$\langle \mathrm{BA} \rangle_{\xi^{(t+1)}} = \left\langle \Theta\left[\frac{Z_{t+1}}{Z_t} < \frac{1}{2}\right] \right\rangle_{\xi^{(t+1)}} \tag{2.12}$$

と表される．ここで $\Theta[x]$ は式 x が真の時に 1 であり，偽の時に 0 である関数である．

証明 Bayes アルゴリズムの誤り確率を BA とすると

$$\begin{aligned}
&\langle \mathrm{BA} \rangle_{\xi^{(t+1)}} \\
&= \left\langle \Theta\left[\int p\left(\overset{t+1}{y} \middle| \overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}\right) p(\boldsymbol{w}|\xi^{(t)})\, \mathrm{d}\boldsymbol{w} < \frac{1}{2}\right] \right\rangle_{\xi^{(t+1)}} && (2.13)\\
&= \left\langle \Theta\left[\frac{1}{Z_t} \int p\left(\overset{t+1}{y} \middle| \overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}\right) \prod_{i=1}^{t} p\left(\overset{i}{y} \middle| \overset{i}{\boldsymbol{x}}, \boldsymbol{w}\right) p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w} < \frac{1}{2}\right] \right\rangle_{\xi^{(t+1)}} \\
&= \left\langle \Theta\left[\frac{Z_{t+1}}{Z_t} < \frac{1}{2}\right] \right\rangle_{\xi^{(t+1)}} && (2.14)
\end{aligned}$$

が成り立つ． □

第 3 章

確定的機械の予測誤差

確定的機械とは入力とパラメータが与えられれば出力が一意に決まる入出力装置である．確定的2分割機械では，例題が一つ与えられた時，パラメータ空間は2分割される．一方は例題と同じ出力を与える機械のパラメータの集合であり，もう一方は例題と異なる出力を与える機械のパラメータの集合である．与えられる例題の個数が多くなるにつれ，すべての例題に合致する機械のパラメータの集合はだんだん小さくなる．パラメータが離散値をとる場合には，例題数が十分大きければこの集合に含まれるパラメータは高い確率で真のパラメータだけになる．しかしパラメータが連続値をとる場合には，この集合に含まれるパラメータは減少してゆくが，決して一つにはならない．このような場合の平均予測誤差は，最も簡単な機械である単純パーセプトロンについても，非常に特殊な場合を除いて求められていない．本章では，まず始めに冗長なパラメータを持たない確定的機械が漸近的に単純パーセプトロンとみなすことができることを示し，その後に，統計幾何学的手法により単純パーセプトロンの予測誤差のバウンドを求める．

3.1 問題設定

まず，確定的2分割機械を定義する．確定的2分割機械は，入力が $\boldsymbol{x}$，パラメータが $\boldsymbol{w}$ の時に y が出力される確率が1か0であるような，2分割機械である．よって次のように定義できる:

定義 3.1 (確定的 2 分割機械) 次のような入出力関係を持つ2分割機械を，確定的2分割機械と呼ぶ:

$$y = \text{sign } f(\boldsymbol{x}, \boldsymbol{w}). \tag{3.1}$$

ただし sign は引数の符号に応じて ± 1 のいずれかを出力する関数である．引数が0の時は -1 を出力するものとしておく．

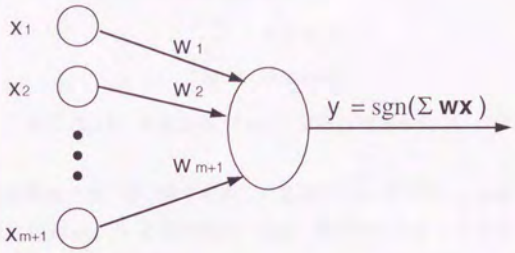


図 3.1: 単純パーセプトロン

もっとも簡単な確定的2分割機械は単純パーセプトロンであり，$f(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w} \cdot \boldsymbol{x}$ である (図3.1).

3.2 許容領域と学習アルゴリズム

確定的2分割機械では，一つの入力 $\boldsymbol{x}$ に応じてパラメータの空間を二つの集合にわけることができる．一方は $f(\boldsymbol{x}, \boldsymbol{w})$ が正であるパラメータの集合であり，この集合に含まれる任意のパラメータを持つ機械は入力 $\boldsymbol{x}$ に対して $+1$ を出力する．もう一方は $f(\boldsymbol{x}, \boldsymbol{w})$ が負であるパラメータの集合であり，この集合に含まれる任意のパラメータを持つ機械は入力 $\boldsymbol{x}$ に対して -1 を出力する．よって一つの例題 $(\overset{i}{\boldsymbol{x}}, \overset{i}{y})$ が与えられると，例題の入力 $\overset{i}{\boldsymbol{x}}$ はパラメータ空間を2分割し，例題の出力 $\overset{i}{y}$ はその一方の集合を示すと考えることができる．真のパラメータはこの集合に含まれており，またこの集合に含まれる任意のパラメータをもつ機械は，入力 $\overset{i}{\boldsymbol{x}}$ に対して $\overset{i}{y}$ を出力する．

例題の数が増えていくと，与えられた例題すべてに対して真の機械と同じ出力を出す機械のパラメータの集合は狭められてゆく．これを許容領域と呼び，次のように定義する:

定義 3.2 (許容領域) 例題 $\xi^{(t)} = \{(\overset{1}{\boldsymbol{x}}, \overset{1}{y}), \ldots, (\overset{t}{\boldsymbol{x}}, \overset{t}{y})\}$ に関して真のパラメータ $\boldsymbol{w}^*$ と同じ出力を出すパラメータの集合を許容領域と呼び A_t で表す．すなわち

$$A_t = \{\boldsymbol{w} | \overset{i}{y} f(\overset{i}{\boldsymbol{x}}, \boldsymbol{w}) > 0, i = 1, \ldots, t\} \tag{3.2}$$

である．

この時，真のパラメータの事後分布の確率密度は

$$\frac{p(\boldsymbol{w})}{Z_t} \quad \text{if } \boldsymbol{w} \in A_t \tag{3.3}$$

$$0 \quad \text{otherwise} \tag{3.4}$$

となるので，全確率 Z_t は，事前分布 $p(\boldsymbol{w})$ で重み付けされた A_t の体積になっている．

確定的2分割機械では，第2章で定義した各学習アルゴリズムによる予測は，パラメータの事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ と符号関数 sign を用いて表すことができる．すなわち，Gibbs アルゴリズムの予測は

$$\text{sign } f(\overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}_t), \quad \text{ただし } \boldsymbol{w}_t \sim p(\boldsymbol{w}|\xi^{(t)}), \tag{3.5}$$

Bayes アルゴリズムの予測は

$$\text{sign } \left(\int \text{sign } f(\overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}) p(\boldsymbol{w}|\xi^{(t)}) \, \mathrm{d}\boldsymbol{w} \right), \tag{3.6}$$

重心アルゴリズムの予測は

$$\text{sign } f(\overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}_t), \quad \text{ただし } \boldsymbol{w}_t = \int \boldsymbol{w} p(\boldsymbol{w}|\xi^{(t)}) \, \mathrm{d}\boldsymbol{w}, \tag{3.7}$$

最尤推定アルゴリズムの予測は

$$\text{sign } f(\overset{t+1}{\boldsymbol{x}}, \boldsymbol{w}_t), \quad \text{ただし } \boldsymbol{w}_t = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|\xi^{(t)}), \tag{3.8}$$

である．Gibbs アルゴリズムによる予測は確率的であるが，他のアルゴリズムによる予測は確定的に定まる．

3.3 確定的機械のパーセプトロンによる近似

冗長なパラメータを持たない確定的2分割機械は，漸近的に単純パーセプトロンと同じ振舞いをすることを示す．まず，単純パーセプトロンを定義する (図3.1)．

定義 3.3 (単純パーセプトロン) 次のような入出力関係を持つ確定的2分割機械を，単純パーセプトロンと呼ぶ:

$$y = \text{sign } (\boldsymbol{w} \cdot \boldsymbol{x}) \tag{3.9}$$

確定的2分割機械 $y = \text{sign } f(\boldsymbol{x}, \boldsymbol{w})$ について，$f(\boldsymbol{x}, \boldsymbol{w})$ はなめらかな関数であり $\boldsymbol{w}$ の各成分で微分可能とする．また，パラメータは冗長でないことを仮定する．冗長でないとは，$\boldsymbol{w}^*$ の十分小さな近傍において，任意のパラメータ $\boldsymbol{w}$ について

$$c_1 \leq \frac{\int \Theta\left[f(\boldsymbol{x}, \boldsymbol{w}^*) \cdot f(\boldsymbol{x}, \boldsymbol{w}) \leq 0\right] p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}}{|\boldsymbol{w} - \boldsymbol{w}^*|} \leq c_2 \tag{3.10}$$

をみたす正数 c_1, c_2 が存在することである．この仮定により，次の補題が成り立つ:

補題 3.1 パラメータの事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ に関する $|\boldsymbol{w}-\boldsymbol{w}^*|^k$ の平均は $\mathrm{O}\left(\frac{1}{t^k}\right)$ である．

証明　$|\boldsymbol{w}-\boldsymbol{w}^*|=\Delta$ とすると，

$$\left\langle \int |\boldsymbol{w}-\boldsymbol{w}^*|^k p(\boldsymbol{w}|\xi^{(t)})\,\mathrm{d}\boldsymbol{w} \right\rangle_{\xi^{(t)}} \tag{3.11}$$

$$= \left\langle \frac{\int \Delta^k \prod \mathrm{Prob}\left[y^i f(\boldsymbol{w},\boldsymbol{x}^i)>0\right] p(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}}{\int \prod \mathrm{Prob}\left[y^i f(\boldsymbol{w},\boldsymbol{x}^i)>0\right] p(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}} \right\rangle_{\xi^{(t)}} \tag{3.12}$$

$$\leq \frac{\int \Delta^k (1-c_1\Delta)^t\,\mathrm{d}\boldsymbol{w}}{\int (1-c_2\Delta)^t\,\mathrm{d}\boldsymbol{w}} \tag{3.13}$$

$$= \frac{\mathrm{O}\left(\frac{1}{t^{m+k}}\right)}{\mathrm{O}\left(\frac{1}{t^m}\right)} \tag{3.14}$$

$$= \mathrm{O}\left(\frac{1}{t^k}\right) \tag{3.15}$$

が導かれる．ただし m はパラメータ $\boldsymbol{w}$ の次元である．　□

補題3.1により t が十分大きい時には，推定パラメータ $\boldsymbol{w}_t$ は $\boldsymbol{w}^*$ の近傍だけ考えればよい．そこで，次の定理が成り立つ．

定理 3.1 確定的学習機械はパラメータが冗長でない時，漸近的に単純パーセプトロンとみなすことができる．

証明　補題3.1により $\boldsymbol{w}-\boldsymbol{w}^*$ は0に収束する．そこで $f(\boldsymbol{w},\boldsymbol{x})$ を $\boldsymbol{w}^*$ のまわりで展開すると

$$f(\boldsymbol{w},\boldsymbol{x}) = f(\boldsymbol{w}^*,\boldsymbol{x}) + \partial_{\boldsymbol{w}} f(\boldsymbol{w}^*,\boldsymbol{x})\cdot(\boldsymbol{w}-\boldsymbol{w}^*) + \mathrm{o}(|\boldsymbol{w}-\boldsymbol{w}^*|) \tag{3.16}$$

であり，高次の項を無視すると複号同順で

$$f(\boldsymbol{w}^*,\boldsymbol{x}) + \partial_{\boldsymbol{w}} f(\boldsymbol{w}^*,\boldsymbol{x})\cdot(\boldsymbol{w}-\boldsymbol{w}^*) \gtrless 0 \tag{3.17}$$

$$\Leftrightarrow\quad W(\boldsymbol{w})\cdot X(\boldsymbol{x}) \gtrless 0 \tag{3.18}$$

である．ここで W, X は

$$W(\boldsymbol{w}) = \begin{pmatrix} \boldsymbol{w}-\boldsymbol{w}^* \\ 1 \end{pmatrix}, \qquad X(\boldsymbol{x}) = \begin{pmatrix} \partial_{\boldsymbol{w}} f(\boldsymbol{w}^*,\boldsymbol{x}) \\ f(\boldsymbol{w}^*,\boldsymbol{x}) \end{pmatrix} \tag{3.19}$$

という，$\boldsymbol{w}$ よりも1だけ次元の高いベクトルである．よって確定的2分割機械は，単純パーセプトロンとみなすことができる．

□

3.4 許容領域の統計的性質

以下では学習機械は，単純パーセプトロン $y = \mathrm{sign}\,(\boldsymbol{w}\cdot\boldsymbol{x})$ であるとする．入力 $\boldsymbol{x}$ は $m+1$ 次元ユークリッド空間の点とし，正規分布 $N(0, E_{m+1})$ に従って独立に選ばれる．ここで E_{m+1} は $m+1$ 次単位行列である．また，パラメータ $\boldsymbol{w}$ は m 次元超球面上の点とし，パラメータの事前分布の確率密度 $p(\boldsymbol{w})$ はなめらかであるとする．入力の分布が，縮退していない正規分布 $N(0,\Sigma)$ ならば，$\Sigma = TT^{\mathrm{t}}$ なる行列 T を用いて

$$\boldsymbol{w}' = T^{\mathrm{t}}\boldsymbol{w} \tag{3.20}$$

$$\boldsymbol{x}' = T^{-1}\boldsymbol{x} \tag{3.21}$$

と座標変換することにより，入力 x が一様分布である場合と同じ議論ができる．入力の分布が正規分布でない場合については未解決である．

単純パーセプトロンでは，$(+\overset{i}{\boldsymbol{x}}, -1)$ という例題と $(-\overset{i}{\boldsymbol{x}}, +1)$ という例題は，同じパラメータの集合を与えるという意味で等価である．よって出力が -1 であるような例題については，この例題と等価で出力が $+1$ であるような例題が与えられたとみなすことにより，例題の出力 $\overset{i}{y}$ は常に $+1$ であるとしても一般性を失わない．また，入力 $\boldsymbol{x}$ を正数倍しても出力は変わらないので，入力ベクトルの大きさについて $|\boldsymbol{x}| = 1$ としてよい．以上のことから，入力の空間は m 次元半超球面 $\boldsymbol{S}_+^m$ とできる．この時入力と真のパラメータとの内積は常に正となり，例題の出力及びテスト入力に関する真の出力は常に $+1$ である．よって例題の入力を例題と呼ぶこととし，$\boldsymbol{S}_+^m$ を例題空間と呼ぶことにする．入力は $N(0, E_{m+1})$ に従うとしたので，$\boldsymbol{S}_+^m$ 上では一様に分布する．以上をまとめて仮定としておく．

仮定 3.1 入力 $\boldsymbol{x}$ は例題空間 $\boldsymbol{S}_+^m$ から一様分布に従って独立に選ばれ，真のパラメータ $\boldsymbol{w}^*$ を持つ機械による出力は常に $+1$ である．

以上の仮定により i 番目の例題は $\overset{i}{\boldsymbol{x}}$ と略記でき，t 番目までの例題はまとめて $\boldsymbol{x}^{(t)}$ で表される．また，許容領域 A_t は

$$A_t = \{\boldsymbol{w} | \boldsymbol{w}\cdot\overset{i}{\boldsymbol{x}} > 0, i = 1, \ldots, t\} \tag{3.22}$$

と表される (図3.2)．以下では，許容領域 A_t の統計的性質について論じる．

まず，パラメータ空間に $(r,\boldsymbol{\omega})$ 座標を導入する．これは 真のパラメータ $\boldsymbol{w}^*$ を原点とする S^m 上の極座標であり，r が動径，$\boldsymbol{\omega}$ が方向を表す．$(r,\boldsymbol{\omega})$ 座標を用いると許容領域 A_t は，$\boldsymbol{\omega}$ 方向の原点から境界までの距離を表す関数 $l(\boldsymbol{\omega})$ を用いて

$$A_t = \{(r,\boldsymbol{\omega}) | 0 \le r \le l(\boldsymbol{\omega})\} \tag{3.23}$$

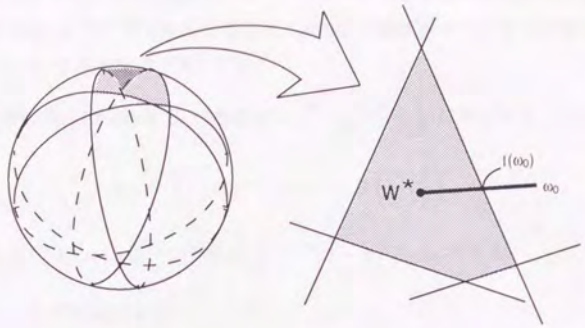


図 3.2: 許容領域 A_t と $(r,\boldsymbol{\omega})$ 座標

と表すことができる (図 3.2). ここで距離とは，S^m 上の測地距離のことである．$(r,\boldsymbol{\omega})$ 座標において方向を表す成分 $\boldsymbol{\omega}$ を $\boldsymbol{\omega}_0$ に固定したとき，次の定理が成り立つ:

定理 3.2 原点から許容領域の境界までの $\boldsymbol{\omega}_0$ 方向の測地距離 $l(\boldsymbol{\omega}_0)$ は，漸近的に指数分布 $\mathrm{Ex}(t/\pi)$ に従う．

証明 方向を表す $\boldsymbol{\omega}$ を定方向 $\boldsymbol{\omega}_0$ に固定する．図3.2 の $\boldsymbol{w}^*$ と $(l(\boldsymbol{\omega}_0),\boldsymbol{\omega}_0)$ を結ぶ測地線分に交わらないような例題が一つ選ばれる確率は，この測地線分の長さが l の時に $\left(1-\dfrac{l}{\pi}\right)$ である．よって測地線分の長さ l が l_0 より大きくなる確率は

$$\mathrm{Prob}\left[l_0 < l\right] = 1 - \left(1-\frac{l_0}{\pi}\right)^t \tag{3.24}$$

であり，測地線分の長さが l になる確率密度は t が大きい時，

$$f(l) = \frac{t}{\pi}\left(1-\frac{l}{\pi}\right)^{t-1} \approx \frac{t}{\pi}e^{-\frac{t}{\pi}l} \tag{3.25}$$

となる．これは l が漸近的に指数分布 $\mathrm{Ex}(t/\pi)$ に従うことを示している．□

上の定理から直ちに次の系が得られる．

系 3.3 原点から許容領域の境界までの S^m 上の距離を $l(\boldsymbol{\omega}_0)$ とすると，$l(\boldsymbol{\omega}_0)$ の k 次モーメント $\left\langle l(\boldsymbol{\omega}_0)^k \right\rangle_{\boldsymbol{x}^{(t)}}$ は t が大きくなる時，$\dfrac{k!\pi^k}{t^k}$ に漸近する．

単純パーセプトロンでは全確率 Z_t は，パラメータの事前確率 $p(\boldsymbol{w})$ による重み付きの，許容領域 A_t の体積を意味している．$p(\boldsymbol{w})$ が超球面 $\boldsymbol{S}^m$ 上で一様分布の時，全確率 Z_t について次の定理が成り立つ:

定理 3.4 許容領域の全確率 Z_t の期待値は $\dfrac{(m-1)!\pi^m}{2I_m}\dfrac{1}{t^m}$ に漸近する．ただし

$$I_m = \int_0^{\pi/2} \sin^{m-1}\phi \, \mathrm{d}\phi = \mathrm{B}\left(\frac{m}{2}, \frac{1}{2}\right) \tag{3.26}$$

である．ここで $\mathrm{B}(s,t)$ はベータ関数 $\displaystyle\int_0^1 x^{s-1}(1-x)^{t-1}\,\mathrm{d}x$ である．

証明 Z_t の定義により

$$\langle Z_t \rangle_{\boldsymbol{x}^{(t)}} = \left\langle \int_{\boldsymbol{S}^m} \prod_i \Theta\left[\boldsymbol{w}\cdot\overset{i}{\boldsymbol{x}} > 0\right] \mathrm{d}\boldsymbol{w} \right\rangle_{\boldsymbol{x}^{(t)}} \tag{3.27}$$

である．$\boldsymbol{w}$ を $(r,\boldsymbol{\omega})$ 座標で表記し，式(3.27)の体積を計算する積分と例題に関する平均操作の順序を入れ換えると，t が大きい時には

$$\begin{aligned}
\langle Z_t \rangle_{\boldsymbol{x}^{(t)}} &= \left\langle \int_{\boldsymbol{S}^m} \prod_i \Theta\left[\boldsymbol{w}\cdot\overset{i}{\boldsymbol{x}} > 0\right] \mathrm{d}\boldsymbol{w} \right\rangle_{\boldsymbol{x}^{(t)}} && (3.28)\\
&= \frac{\displaystyle\int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega} \int_0^{\pi} \left(1-\frac{r}{\pi}\right)^t \sin^{m-1} r \, \mathrm{d}r}{\displaystyle\int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega} \int_0^{\pi} \sin^{m-1} r \, \mathrm{d}r} && (3.29)\\
&= \frac{1}{2I_m}\int_0^{\pi} \left(1-\frac{r}{\pi}\right)^t \sin^{m-1} r \, \mathrm{d}r && (3.30)\\
&= \frac{(m-1)!\pi^m}{2I_m}\frac{1}{t^m} + \mathrm{o}\left(\frac{1}{t^m}\right) && (3.31)
\end{aligned}$$

である． □

許容領域の統計的性質を用いると，Gibbs アルゴリズムの予測誤差のバウンドが得られる．

定理 3.5 Gibbs アルゴリズムの予測誤差の期待値 $\langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}}$ について，漸近的に

$$\frac{1}{m+1}\frac{m}{t} \leq \langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}} \leq \frac{m}{t} \tag{3.32}$$

が成り立つ．

証明 まず $\dfrac{1}{m+1}\dfrac{m}{t} \leq \langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}}$ を示す．パラメータ空間 $\boldsymbol{S}^m$ に $(r,\boldsymbol{\omega})$ 座標を導入する．推定パラメータ $\boldsymbol{w}_t$ が $\boldsymbol{w}_t = (r,\boldsymbol{\omega})$ である時，$\boldsymbol{w}_t$ による予測が誤りとなるようなテスト入力 $\overset{t+1}{\boldsymbol{x}}$ が与えられる確率は r/π であ

る．Gibbs アルゴリズムでは，推定パラメータ $\boldsymbol{w}_t$ は事後分布 $p(\boldsymbol{w}|\boldsymbol{x}^{(t)})$ に従って選ばれるので，予測誤差 GA の ${}^{t+1}\boldsymbol{x}$ に関する平均は t が大きい時

$$\langle \mathrm{GA} \rangle_{\boldsymbol{x}}^{t+1} = 1 - \frac{\int_{A_t} \left(1 - \frac{r}{\pi}\right) p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w}}{\int_{A_t} p(\boldsymbol{w})\, \mathrm{d}\boldsymbol{w}} \tag{3.33}$$

$$= \frac{\int_{\boldsymbol{S}^{m-1}} \int_0^{l(\boldsymbol{\omega})} \frac{r}{\pi} p(r,\boldsymbol{\omega}) \sin^{m-1} r\, \mathrm{d}r\, \mathrm{d}\boldsymbol{\omega}}{\int_{\boldsymbol{S}^{m-1}} \int_0^{l(\boldsymbol{\omega})} p(r,\boldsymbol{\omega}) \sin^{m-1} r\, \mathrm{d}r\, \mathrm{d}\boldsymbol{\omega}} \tag{3.34}$$

$$\approx \frac{\int_{\boldsymbol{S}^{m-1}} \int_0^{l(\boldsymbol{\omega})} \frac{r}{\pi} p(0,\boldsymbol{\omega}) r^{m-1}\, \mathrm{d}r\, \mathrm{d}\boldsymbol{\omega}}{\int_{\boldsymbol{S}^{m-1}} \int_0^{l(\boldsymbol{\omega})} p(0,\boldsymbol{\omega}) r^{m-1}\, \mathrm{d}r\, \mathrm{d}\boldsymbol{\omega}} \tag{3.35}$$

$$= \frac{m}{(m+1)\pi} \frac{\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})^{m+1}\, \mathrm{d}\boldsymbol{\omega}}{\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})^{m}\, \mathrm{d}\boldsymbol{\omega}} \tag{3.36}$$

である．ここで

$$\bar{l} = \frac{\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}}{\int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega}} \tag{3.37}$$

とすると

$$\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega}_1)^{m+1}\, \mathrm{d}\boldsymbol{\omega}_1 \int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega}_2 - \int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega}_1)^{m}\, \mathrm{d}\boldsymbol{\omega}_1 \int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega}_2)\, \mathrm{d}\boldsymbol{\omega}_2 \tag{3.38}$$

$$= \int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega}_2 \int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega}_1)^{m} \left(l(\boldsymbol{\omega}_1) - \bar{l}\right)\, \mathrm{d}\boldsymbol{\omega}_1 \tag{3.39}$$

$$\geq \int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega}_2 \int_{\boldsymbol{S}^{m-1}} \bar{l}^{m} \left(l(\boldsymbol{\omega}_1) - \bar{l}\right)\, \mathrm{d}\boldsymbol{\omega}_1 \tag{3.40}$$

$$= 0 \tag{3.41}$$

であるので

$$\frac{\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})^{m+1}\, \mathrm{d}\boldsymbol{\omega}}{\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})^{m}\, \mathrm{d}\boldsymbol{\omega}} \geq \frac{\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}}{\int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega}} \tag{3.42}$$

が成り立ち，また

$$\left\langle \frac{\int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega}}{\int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega}} \right\rangle_{\boldsymbol{x}^{(t)}} = \frac{\left\langle \int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})\, \mathrm{d}\boldsymbol{\omega} \right\rangle_{\boldsymbol{x}^{(t)}}}{\left\langle \int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega} \right\rangle_{\boldsymbol{x}^{(t)}}}, \tag{3.43}$$

$$\left\langle \int_{\boldsymbol{S}^{m-1}} l(\boldsymbol{\omega})^k \, \mathrm{d}\boldsymbol{\omega} \right\rangle_{\boldsymbol{x}^{(t)}} = \int_{\boldsymbol{S}^{m-1}} \left\langle l(\boldsymbol{\omega})^k \right\rangle_{\boldsymbol{x}^{(t)}} \mathrm{d}\boldsymbol{\omega} \tag{3.44}$$

$$= \frac{k!\pi^k}{t^k} \int_{\boldsymbol{S}^{m-1}} \mathrm{d}\boldsymbol{\omega} \tag{3.45}$$

であるので，

$$\frac{1}{m+1}\frac{m}{t} \leq \langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}} \tag{3.46}$$

が得られる．ここで式(3.45) において系3.3を用いた．

次に $\langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}} \leq \dfrac{m}{t}$ を示す．$-\log \dfrac{Z_{t+1}}{Z_t}$ を予測エントロピと定義すると，log の凸性により

$$1 - \frac{Z_{t+1}}{Z_t} \leq -\log \frac{Z_{t+1}}{Z_t} \tag{3.47}$$

であるから，予測エントロピは予測誤差の上のバウンドになっている．以下では，予測エントロピ $-\log \dfrac{Z_{t+1}}{Z_t}$ の例題に関する平均が $\dfrac{m}{t}$ に等しいことを，Amari [2] の手法によって示す．

まず，許容領域の体積 Z_t の k 次のモーメント ${Z_t}^k$ について考察する．${Z_t}^k$ は k 個のパラメータ $\overset{1}{\boldsymbol{w}}, \overset{2}{\boldsymbol{w}}, \ldots, \overset{k}{\boldsymbol{w}}$ を用いて

$${Z_t}^k = \int_{A_t} p(\overset{1}{\boldsymbol{w}}) \cdots p(\overset{k}{\boldsymbol{w}}) \, \mathrm{d}\overset{1}{\boldsymbol{w}} \cdots \mathrm{d}\overset{k}{\boldsymbol{w}} \tag{3.48}$$

と表され，その例題に関する平均 $\left\langle {Z_t}^k \right\rangle_{\boldsymbol{x}^{(t)}}$ はある定数 α_k を用いて

$$\left\langle {Z_t}^k \right\rangle_{\boldsymbol{x}^{(t)}} = \frac{\alpha_k}{t^{km}} \tag{3.49}$$

と表される．よって $Y_t = t^m Z_t$ という確率変数を考えると，各モーメントが定数に収束するから Y_t はある確率変数 Y に分布収束する．ここで

$$\beta(t) = \langle \log Y_t - \log Y \rangle_{\boldsymbol{x}^{(t)}} \tag{3.50}$$

とすると $\beta(t)$ は $t \to \infty$ の時，0 に収束する．この収束が単調であると仮定すれば，

$$\beta(t+1) - \beta(t) = \mathrm{O}\left(\frac{\beta(t)}{t}\right) \tag{3.51}$$

であるので

$$\left\langle -\log \frac{Z_{t+1}}{Z_t} \right\rangle_{\boldsymbol{x}^{(t+1)}} = \langle \log Z_t - \log Z_{t+1} \rangle_{\boldsymbol{x}^{(t+1)}} \tag{3.52}$$

$$= m\log(t+1) - m\log t + \mathrm{O}\left(\frac{\beta(t)}{t}\right) \tag{3.53}$$

$$= \frac{m}{t} + \mathrm{o}(\frac{1}{t}) \tag{3.54}$$

が得られる． □

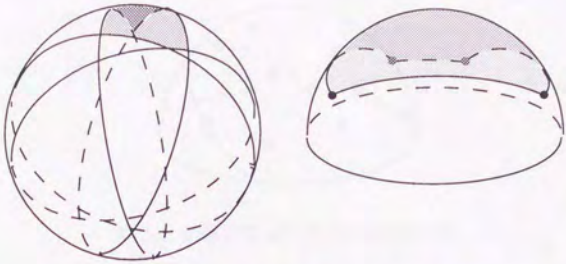


図 3.3: パラメータ空間 $\boldsymbol{S}^m$ 内の許容領域と例題空間 $\boldsymbol{S}^m_+$ 内の凸包

一般に $x \approx 1$ の時には $1 - x \approx -\log x$ であるが，ここでは

$$\left\langle 1 - \frac{Z_{t+1}}{Z_t} \right\rangle_{\boldsymbol{x}^{(t+1)}} \approx \left\langle -\log \frac{Z_{t+1}}{Z_t} \right\rangle_{\boldsymbol{x}^{(t+1)}} \tag{3.55}$$

ではないことに注意する．これは $\left\langle \dfrac{Z_{t+1}}{Z_t} \right\rangle_{\boldsymbol{x}^{(t+1)}} \approx 1$ は成り立つが，$\dfrac{Z_{t+1}}{Z_t} \approx 1$ は成り立たないことに起因する．$\dfrac{Z_{t+1}}{Z_t}$ は高い確率で 1 であるが，1 でない時には 1 に近いとは限らず，これらが平均された結果として $\left\langle \dfrac{Z_{t+1}}{Z_t} \right\rangle_{\boldsymbol{x}^{(t+1)}} \approx 1$ となっているのである．$\left\langle \dfrac{Z_{t+1}}{Z_t} \right\rangle_{\boldsymbol{x}^{(t+1)}}$ は確率変数の商の期待値となっており，このことが予測誤差の計算が困難である原因となっている．

3.5 許容領域の幾何学的性質

許容領域 A_t は例題 $\boldsymbol{x}^{(t)}$ が超球面上に作る t 個の超平面のうちのいくつかによって囲まれた，$\boldsymbol{S}^m$ 上の超多面体である．ここでは許容領域 A_t の，幾何学的性質について考察する．許容領域 A_t の頂点数について次の定理が成り立つ:

定理 3.6 ランダムに選ばれた t 個の例題が作る許容領域 A_t の頂点数 V_t の例題に関する期待値 $\langle V_t \rangle_{\boldsymbol{x}^{(t)}}$ は $\dfrac{\pi^m}{mI_m}$ に漸近する．

証明 パラメータ空間と例題空間は互いに双対であるので，パラメータ空間に例題がつくる許容領域 A_t は例題空間に例題がつくる凸包と互いに双対になっている (図 3.3)．よって許容領域の頂点数 V_t は凸包の境界面数と等しく，また許容領域の境界面数 F_t は凸包の頂点数に等しい．従っ

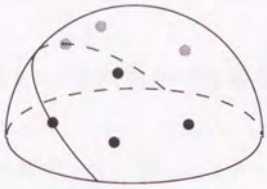


図 3.4: 例題空間 S^m_+ 内の凸包の境界面

て，許容領域の頂点数の例題に関する期待値 $\langle V_t \rangle_{\boldsymbol{x}^{(t)}}$ を求めるには，凸包の境界面数の例題に関する期待値を求めればよい．

閉領域内のランダム点群の作る凸包の頂点数については，閉領域が2次元平面上の円及び正多角形の場合について Efron [8] が求めた．ここではその手法を m 次元超球面上に応用する．まず，m 個の例題 $\overset{1}{\boldsymbol{x}}, \ldots, \overset{m}{\boldsymbol{x}}$ が例題空間 S^m_+ 上に作る超平面 $\boldsymbol{w}_m$ が凸包の境界面になる確率を考える．超平面 $\boldsymbol{w}_m$ は例題空間を2分しているので，残りの $t-m$ 個の例題がすべて $\boldsymbol{w}_m$ の片側から選ばれることが $\boldsymbol{w}_m$ が凸包の境界面になることと同じである (図 3.4)．$\boldsymbol{w}_m$ がパラメータ空間の $(r, \boldsymbol{\omega})$ 座標で $\boldsymbol{w}_m = (r, \boldsymbol{\omega})$ と表される時，$\boldsymbol{w}_m$ が凸包の境界面になる確率は

$$\left(1 - \frac{r}{\pi}\right)^{t-m} + \left(\frac{r}{\pi}\right)^{t-m} \tag{3.56}$$

である．

m 個の例題 $\overset{1}{\boldsymbol{x}}, \ldots, \overset{m}{\boldsymbol{x}}$ は例題空間 S^m_+ 上で一様に分布するので $\boldsymbol{w}_m$ はパラメータ空間 S^m 上で一様に分布する．よって $\boldsymbol{w}_m$ が境界面になる確率の期待値は t が大きい時

$$\int_{S^m} \left\{ \left(1 - \frac{r}{\pi}\right)^{t-m} + \left(\frac{r}{\pi}\right)^{t-m} \right\} \mathrm{d}\boldsymbol{w} \tag{3.57}$$

$$= \frac{(m-1)!\pi^m}{I_m t(t-1)\cdots(t-m+1)} + \mathrm{o}(\frac{1}{t^m}) \tag{3.58}$$

となる．t 個の点群が作る $m-1$ 次元超平面は ${}_t\mathrm{C}_m$ 個あるから，凸包の境界面数の例題に関する期待値は $\boldsymbol{w}_m$ が境界面になる確率の ${}_t\mathrm{C}_m$ 倍となり，$\dfrac{\pi^m}{mI_m}$ に漸近する．従って許容領域の頂点数 V_t の期待値 $\langle V_t \rangle_{\boldsymbol{x}^{(t)}}$ も $\dfrac{\pi^m}{mI_m}$ に漸近する． □

$m \leq 3$ の場合は，許容領域の頂点数の期待値を用いて許容領域の境界面数の期待値を求めることができる．

系 3.7 ランダムに選ばれた t 個の例題が作る許容領域 A_t の境界面数 F_t の例題に関する期待値 $\langle F_t \rangle_{\boldsymbol{x}^{(t)}}$ は

$$m = 1\text{の時}, \qquad \langle F_t \rangle_{\boldsymbol{x}^{(t)}} = 2, \tag{3.59}$$

$$m = 2\text{の時}, \qquad \langle F_t \rangle_{\boldsymbol{x}^{(t)}} = \frac{\pi^2}{2}, \tag{3.60}$$

$$m = 3\text{の時}, \qquad \langle F_t \rangle_{\boldsymbol{x}^{(t)}} = \frac{2\pi^2}{3} + 2 \tag{3.61}$$

である．また，一般の m について

$$m + 1 \leq F_t \leq V_t \tag{3.62}$$

が確率 1 で成り立つ．

証明　$m = 1$ の時，許容領域は線分であるので境界面数は常に 2 である．

$m = 2$ の時，許容領域は多角形であるので境界面数は頂点数に等しい．

$m = 3$ の時，許容領域は多面体であるので Euler-Poincaré の公式が成り立ち

$$\text{頂点数} - \text{辺数} + \text{面数} = 2$$

である．また，すべての頂点は確率 1 で 3 平面の交点であるので，各頂点はすべて 3 辺の端点となっている．よって辺数は頂点数の $\frac{3}{2}$ 倍であり，

$$V_t - \frac{3}{2}V_t + F_t = 2 \tag{3.63}$$

から

$$F_t = \frac{1}{2}V_t + 2 \tag{3.64}$$

が導かれる．これと定理3.6の結果から，境界面数の期待値が得られる．

次に式(3.62) を示す．m 次元多面体の最小頂点数は $m + 1$ であるから，$m + 1 \leq F_t$ が成り立つ．また，各境界面上にある頂点数の，すべての境界面についての和を σ とすると，一つの頂点は確率 1 で m 個の境界面によって作られることから，

$$\sigma = mV_t \tag{3.65}$$

である．また，境界面は $m-1$ 次元多面体であるから，一つの境界面上には m 個以上の頂点がある．境界面は F_t 個あるから

$$\sigma \geq mF_t \tag{3.66}$$

である．式(3.65)と式(3.66)から，$F_t \geq V_t$ が導かれる． □

系3.7の結果は，例題数 t がいくら増加しても，機械にとって覚えなければならない例題の数の期待値はたかだか定数であることを示している．すなわち，例題がどんなに多く与えられても，そのうちいくつかを選び出してそれだけ学習すれば，残りの例題はすべて間違いなく判別できるということである．このように，許容領域の境界面となっている例題を，有効例題と呼ぶ．有効例題の期待値が定数であることを利用すると，ランダムに選ばれた入力の取捨選択について議論することができる(付録参照)．

定理3.6及び系3.7の結果を確認し，また分布がどのようになっているのかを調べるため，計算機実験を行った．

実験 3.1 パラメータ次元 $m = 2, 3, 4, 5, 6$ について，100, 300, 1000, 3000, 10000, 30000, 100000, 300000 個の例題を与えた時の許容領域の頂点数と境界面数を調べた．実験は各 m について 200 回ずつ行った．

結果 図3.5は，例題数 t と許容領域の境界面数 F_t の関係を次元 m が $m = 3, m = 6$ の場合について示したものである．エラーバーは200回の実験の標準偏差である．この図から，例題数が 100 以上においては例題数に関わらず境界面数は定数となっていることがわかる．

図3.6はパラメータの次元 m と許容領域の境界面数 F_t の関係を示したものである．ここで実線は理論値であり，実験の結果は $m \leq 3$ において理論値と一致している．

境界面数 F_t は，どのような分布になっているのであろうか．図3.7, 3.8 は $m = 3$, 図3.9, 3.10 は $m = 6$ において例題数が 100 の時の，それぞれ境界面数の度数分布，累積度数を示したものである．これらの図から，パラメータの次元がいずれであっても，境界面数はほぼ対称に分布していることがわかる．

図3.11は，例題数 t と許容領域の頂点数 V_t の関係を次元 m が $m = 3, m = 6$ の場合について示したものである．エラーバーは200回の実験の標準偏差である．この図から，例題数が 100 以上においては例題数に関わらず頂点数は定数となっていることがわかる．

図3.12はパラメータの次元 m と許容領域の頂点数 V_t の関係を示したものである．ここで実線は理論値であり，各 m について実験の結果は理論値と一致している．

図3.13 は $m = 3$, 図3.14 は $m = 6$ の，例題数が 100 の時の頂点数の累積度数を示したものである．いずれの次元においても，似た分布となっていることがわかる．

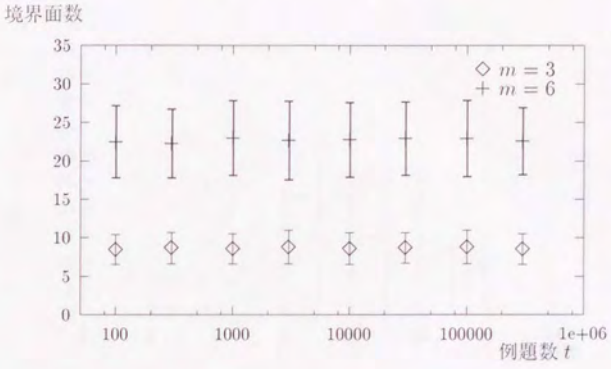


図 3.5: 与えられた例題数と境界面数 ($m = 3, m = 6$)

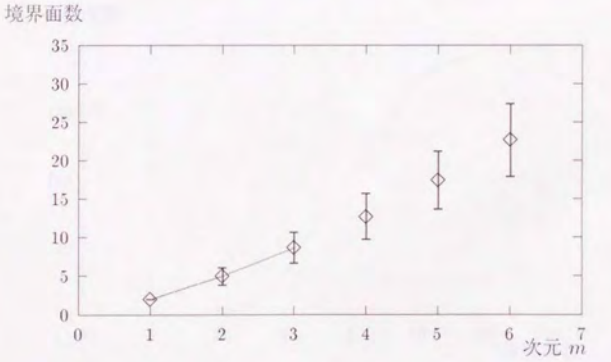


図 3.6: 次元 m と境界面数

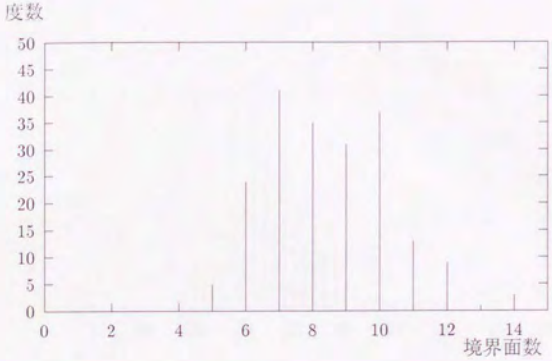


図 3.7: $m = 3$ における境界面数の度数分布

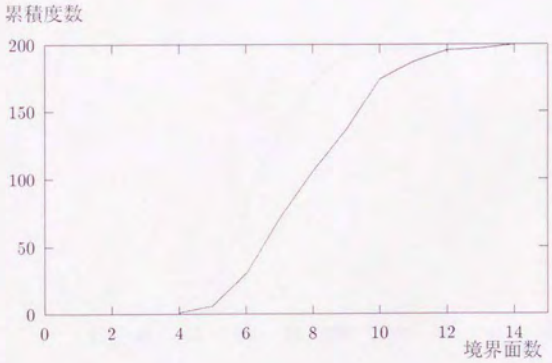


図 3.8: $m = 3$ における境界面数の累積度数

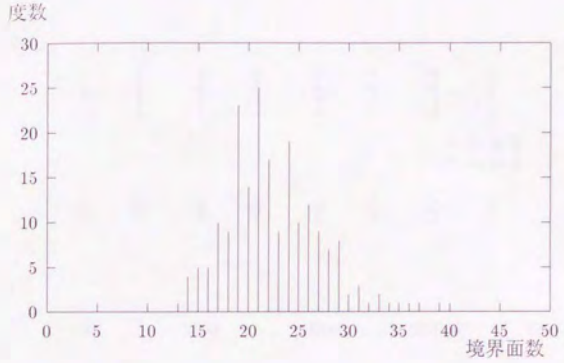


図 3.9: $m = 6$ における境界面数の度数分布

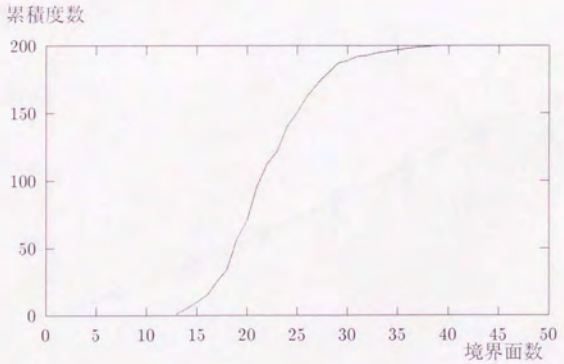


図 3.10: $m = 6$ における境界面数の累積度数

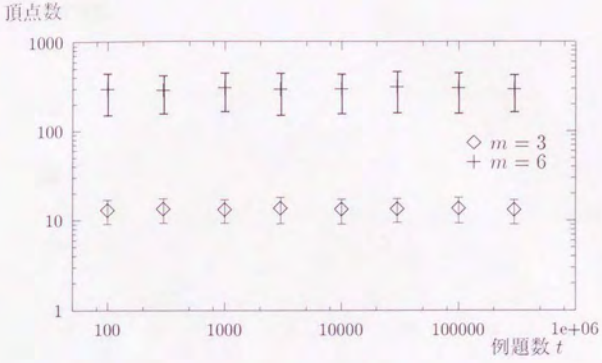


図 3.11: 与えられた例題数と頂点数 ($m = 3, m = 6$)

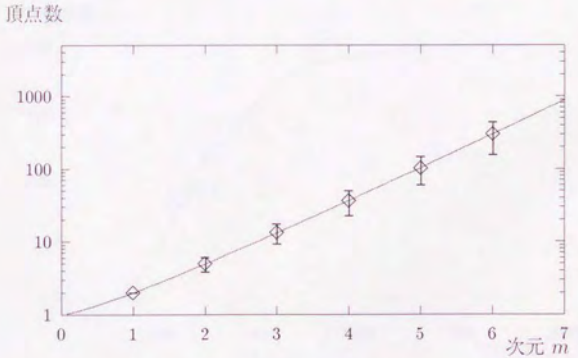


図 3.12: 次元 m と頂点数

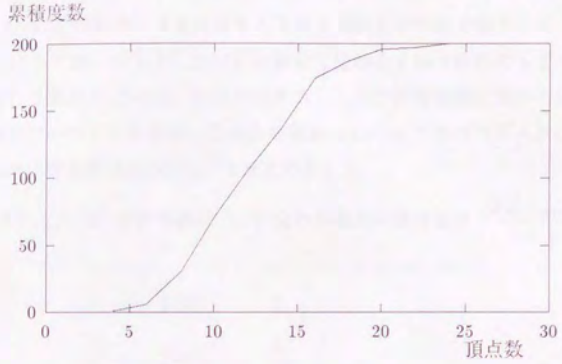


図 3.13: $m = 3$ における頂点数の累積度数

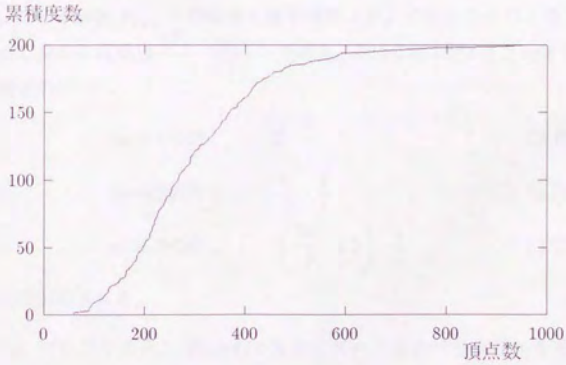


図 3.14: $m = 6$ における頂点数の累積度数

3.6 条件付予測誤差を用いた予測誤差の評価

平均予測誤差は，テスト入力 ${}^{t+1}\boldsymbol{x}$ が許容領域 A_t に交わる確率と，テスト入力 ${}^{t+1}\boldsymbol{x}$ が許容領域 A_t に交わるという条件のもとでの予測誤差の平均の積として，表すことができる．以下では，テスト入力が許容領域に交わるという条件のもとでの予測誤差を，条件付予測誤差と呼ぶ．本節ではテスト入力が許容領域に交わる確率及び条件付予測誤差のバウンドを求め，これらの積から Gibbs アルゴリズム及び Bayes アルゴリズムの平均予測誤差のバウンドを求める [11].

補題 3.2 テスト入力 ${}^{t+1}\boldsymbol{x}$ が許容領域 A_t に交わる確率の期待値は $\dfrac{\langle F_{t+1}\rangle_{\boldsymbol{x}^{(t+1)}}}{t+1}$ であり，

$$m=1\text{の時} \qquad 2\cdot\frac{1}{t}, \tag{3.67}$$

$$m=2\text{の時} \qquad \frac{\pi^2}{2}\cdot\frac{1}{t}, \tag{3.68}$$

$$m=3\text{の時} \qquad \left(\frac{2\pi^2}{3}+2\right)\cdot\frac{1}{t}, \tag{3.69}$$

に漸近的する．

証明 例題とテスト入力は同じ分布に従って独立に選ばれるので，テスト入力 ${}^{t+1}\boldsymbol{x}$ が許容領域 A_t に交わる確率は，${}^{t+1}\boldsymbol{x}$ が $t+1$ 個の入力 $\boldsymbol{x}^{(t+1)}$ が作る許容領域 A_{t+1} の境界面である確率と等しく，その期待値は許容領域 A_{t+1} の境界面数 F_{t+1} の期待値を超平面数 $t+1$ で割ったものと等しい．よって交わる確率は $\dfrac{\langle F_{t+1}\rangle_{\boldsymbol{x}^{(t+1)}}}{t+1}$ であり，系3.7および $t+1\approx t$ により，漸近的に

$$m=1\text{の時}, \qquad 2\cdot\frac{1}{t} \tag{3.70}$$

$$m=2\text{の時}, \qquad \frac{\pi^2}{2}\cdot\frac{1}{t} \tag{3.71}$$

$$m=3\text{の時}, \qquad \left(\frac{2\pi^2}{3}+2\right)\cdot\frac{1}{t} \tag{3.72}$$

であることが示される． □

補題 3.3 Gibbs アルゴリズム，Bayes アルゴリズムの条件付予測誤差をそれぞれ GA′, BA′ とすると，次の式が成り立つ:

$$\frac{3m+1}{2(m+1)(2m+1)} \le \langle \mathrm{GA}'\rangle_{\boldsymbol{x}^{t+1}} \le \frac{1}{3}, \tag{3.73}$$

$$\frac{1}{2(m+1)} \le \langle \mathrm{BA}'\rangle_{\boldsymbol{x}^{t+1}} \le \frac{1}{4}. \tag{3.74}$$

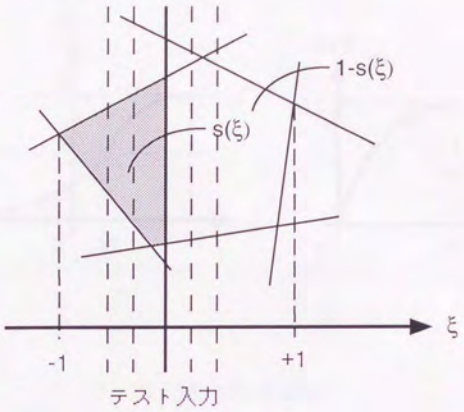


図 3.15: 許容領域とテスト入力

証明　テスト入力が許容領域 A_t を，体積比が $s:1-s$ になるように分けているとする (図 3.15)．この時，Gibbs アルゴリズムの条件付予測誤差 GA$'$ は $2s(1-s)$ で与えられる．なぜなら，予測が誤りであるのは，真のパラメータと推定パラメータが異なる領域にある時だからである．また，Bayes アルゴリズムの条件付予測誤差 BA$'$ は $\min(s,1-s)$ で与えられる．なぜなら，予測が誤りであるのは，真のパラメータが半分より小さい領域にある時だからである．今，許容領域 A_t は小さいとすれば，テスト入力 ${}^{t+1}\boldsymbol{x}$ は定方向に平行に一様に分布するとしてよい．この時 図 3.15 のように ξ 座標をとれば，s は ξ の関数 $s(\xi)$ となり，$s(-1)=0, s(+1)=1$ を満たす．また，$s'(\xi)$ は連続で正値をとり，上に凸となる．

$s(\xi)=\dfrac{1}{2}$ なる ξ を ξ_0 とすると，$s(\xi)$ は $g_i(0)=0, g_i(1)=\dfrac{1}{2}, i=1,2$, なる関数 $g_1(\eta), g_2(\eta)$ を用いて

$$s(\xi)=\begin{cases}\dfrac{1}{2}-g_1\left(\dfrac{\xi_0-\xi}{1+\xi_0}\right) & \text{if } \xi\le\xi_0\\ \dfrac{1}{2}+g_2\left(\dfrac{\xi-\xi_0}{1-\xi_0}\right) & \text{otherwise}\end{cases} \tag{3.75}$$

と書ける (図 3.16)．この時，条件付予測誤差 GA$'$, BA$'$ のテスト入力 ${}^{t+1}\boldsymbol{x}$

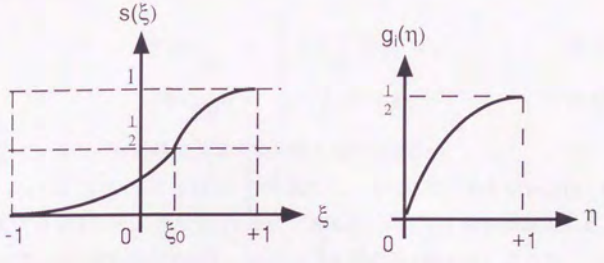


図 3.16: $s(\xi)$ と $g_i(\eta)$

に関する平均はそれぞれ

$$\begin{aligned}
\langle \mathrm{GA}' \rangle_{\boldsymbol{x}}^{t+1} &= \frac{1}{2}\int_{-1}^{1} 2s(\xi)(1-s(\xi))\ \mathrm{d}\xi && (3.76)\\
&= \frac{1}{2} - (1+\xi_0)\int_0^1 g_1(\eta)^2\ \mathrm{d}\eta - (1-\xi_0)\int_0^1 g_2(\eta)^2\ \mathrm{d}\eta\\
&= \frac{1}{2} - \int_0^1 \left\{ g_1(\eta)^2 + g_2(\eta)^2 \right\}\ \mathrm{d}\eta\\
&\quad -\xi_0 \int_0^1 \left\{ g_1(\eta)^2 - g_2(\eta)^2 \right\}\ \mathrm{d}\eta, && (3.77)\\
\langle \mathrm{BA}' \rangle_{\boldsymbol{x}}^{t+1} &= \frac{1}{2}\int_{-1}^{1} \min\left(s(\xi), 1-s(\xi)\right)\ \mathrm{d}\xi && (3.78)\\
&= \frac{1}{2} - \frac{1+\xi_0}{2}\int_0^1 g_1(\eta)\ \mathrm{d}\eta - \frac{1-\xi_0}{2}\int_0^1 g_2(\eta)\ \mathrm{d}\eta\\
&= \frac{1}{2} - \frac{1}{2}\int_0^1 \left\{ g_1(\eta) + g_2(\eta) \right\}\ \mathrm{d}\eta\\
&\quad -\frac{\xi_0}{2}\int_0^1 \left\{ g_1(\eta) - g_2(\eta) \right\}\ \mathrm{d}\eta && (3.79)
\end{aligned}$$

となる．式 (3.77) 及び式 (3.79) から，$\langle \mathrm{GA}' \rangle_{\boldsymbol{x}}^{t+1}$ 及び $\langle \mathrm{BA}' \rangle_{\boldsymbol{x}}^{t+1}$ は複号同順で，

$$\int g_1(\eta)^2\ \mathrm{d}\eta \gtrless \int g_2(\eta)^2\ \mathrm{d}\eta \qquad (3.80)$$

の時に $\xi_0 = \pm 1$ とすると最小となり，$\xi_0 = \mp 1$ とすると最大となる．この時，$g_1 = g_2$ としても同じ値となり，さらに $\xi_0 = 0$ としても同じ値となる．すなわち，$g_1 = g_2, \xi_0 = 0$ という条件のもとでの最大値及び最小値は，この条件がない時の最大値，最小値と一致する．よって $\langle \mathrm{GA}' \rangle_{\boldsymbol{x}}^{t+1}$

及び $\langle \mathrm{GA}' \rangle_{\boldsymbol{x}}^{t+1}$ の最大値及び最小値を求めるには

$$\langle \mathrm{GA}' \rangle_{\boldsymbol{x}}^{t+1} = \frac{1}{2} - 2\int_0^1 g(\eta)^2 \,\mathrm{d}\eta, \tag{3.81}$$

$$\langle \mathrm{BA}' \rangle_{\boldsymbol{x}}^{t+1} = \frac{1}{2} - \int_0^1 g(\eta) \,\mathrm{d}\eta \tag{3.82}$$

として，それぞれの最大値及び最小値を求めればよい．

$g(\eta)$ は $g(0) = 0, g(1) = \dfrac{1}{2}$ を満たし，また体積であるから $g'(\eta)$ は連続で正値をとり，また $s'(\xi)$ が凸であるから，$g'(\eta)$ は単調減少する．よって $g(\eta)$ は上に凸であり，$g(\eta) \geq \dfrac{1}{2}\eta$ である (図3.16)．よって

$$\langle \mathrm{GA}' \rangle_{\boldsymbol{x}}^{t+1} = \frac{1}{2} - 2\int_0^1 g(\eta)^2 \,\mathrm{d}\eta \tag{3.83}$$

$$\leq \frac{1}{2} - 2\int_0^1 \left(\frac{1}{2}\eta\right)^2 \mathrm{d}\eta = \frac{1}{3}, \tag{3.84}$$

$$\langle \mathrm{BA}' \rangle_{\boldsymbol{x}}^{t+1} = \frac{1}{2} - \int_0^1 g(\eta) \,\mathrm{d}\eta \tag{3.85}$$

$$\leq \frac{1}{2} - \int_0^1 \left(\frac{1}{2}\eta\right) \mathrm{d}\eta = \frac{1}{4} \tag{3.86}$$

が成り立つ．

また A_t が十分小さい時には m 次元 Euclid 空間内の図形とみなすことができる．許容領域 A_t は凸図形なので，相似な図形を考えることにより，$-1 \leq \xi_1 < \xi_2 \leq 0$ の時に

$$s(\xi_1) \geq \frac{(1+\xi_1)^m}{(1+\xi_2)^m} s(\xi_2) \tag{3.87}$$

が成り立つ．よって

$$\frac{s(\xi_1)}{(1+\xi_1)^m} \geq \frac{s(\xi_2)}{(1+\xi_2)^m} \geq s(0) = \frac{1}{2} \tag{3.88}$$

であるので

$$s(\xi) \geq \frac{1}{2}(1+\xi)^m \tag{3.89}$$

であり，これからただちに

$$g(\eta) = \frac{1}{2} - s(-\eta) \leq \frac{1}{2} - \frac{1}{2}(1-\eta)^m \tag{3.90}$$

が導かれる．これを用いると

$$\langle \mathrm{GA}' \rangle_{\boldsymbol{x}}^{t+1} = \frac{1}{2} - 2\int_0^1 g(\eta)^2 \,\mathrm{d}\eta \tag{3.91}$$

$$\geq \frac{1}{2} - 2\int_0^1 \left(\frac{1}{2}\eta\right)^2 \mathrm{d}\eta \tag{3.92}$$

$$= \frac{3m+1}{2(m+1)(2m+1)}, \tag{3.93}$$

$$\langle \mathrm{BA}' \rangle_{\boldsymbol{x}}^{t+1} = \frac{1}{2} - \int_0^1 g(\eta)\, \mathrm{d}\eta \tag{3.94}$$

$$\geq \frac{1}{2} - \int_0^1 \left(\frac{1}{2}\eta\right) \mathrm{d}\eta \tag{3.95}$$

$$= \frac{1}{2(m+1)} \tag{3.96}$$

である．□

テスト入力が許容領域に交わる確率と条件付予測誤差の積が予測誤差なので，補題3.2，補題3.3から，次の定理が導かれる：

定理 3.8 Gibbs アルゴリズムの予測誤差 GA について漸近的に

$$m = 1\text{の時}, \qquad \langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}} = \frac{2}{3} \cdot \frac{1}{t} \tag{3.97}$$

$$m = 2\text{の時}, \qquad 0.5757 \cdot \frac{2}{t} \leq \langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}} \leq 0.8225 \cdot \frac{2}{t} \tag{3.98}$$

$$m = 3\text{の時}, \qquad 0.5107 \cdot \frac{3}{t} \leq \langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}} \leq 0.9533 \cdot \frac{3}{t} \tag{3.99}$$

が成り立ち，また Bayes アルゴリズムの予測誤差 BA について漸近的に

$$m = 1\text{の時}, \qquad \langle \mathrm{BA} \rangle_{\boldsymbol{x}^{(t+1)}} = \frac{1}{2} \cdot \frac{1}{t} \tag{3.100}$$

$$m = 2\text{の時}, \qquad 0.4112 \cdot \frac{2}{t} \leq \langle \mathrm{BA} \rangle_{\boldsymbol{x}^{(t+1)}} \leq 0.6169 \cdot \frac{2}{t} \tag{3.101}$$

$$m = 3\text{の時}, \qquad 0.3575 \cdot \frac{3}{t} \leq \langle \mathrm{BA} \rangle_{\boldsymbol{x}^{(t+1)}} \leq 0.7150 \cdot \frac{3}{t} \tag{3.102}$$

が成り立つ．

この結果は，$m = 1$ については従来の計算方法による結果と一致している．

3.7 計算機実験による予測誤差の評価

確定的機械については $m = 1$ の場合を除いて予測誤差は求められていない．ここでは計算機実験により平均予測誤差を求めた．また，パーセプトロン学習を行った場合の平均予測誤差を実験的に求めた．

実験 3.2 パラメータ次元 $m = 1, 2, 3, 4, 5$ について，100, 200, 300, 500, 1000, 2000, 3000, 5000, 10000 個の例題を与えた時の Gibbs アルゴリズム及び重心アルゴリズムの予測誤差を求めた．実験は各 m について 30 回ずつ行った．

この実験では，予測誤差を評価する際，実際にテスト入力をランダムに選ぶのではなく，推定パラメータがもつ誤り確率を利用した．すなわち，推定パラメータが $(r,\boldsymbol{\omega})$ の時には予測が誤りであるようなテスト入力が選ばれる確率は $\frac{r}{\pi}$ であるので，推定パラメータに対して $\frac{r}{\pi}$ を評価して，予測誤差の平均を求めた．この計算方法はBayes アルゴリズムに関しては評価できないという欠点があるが，実際にテスト入力を多数選んで誤り確率を実測するよりも高い精度で予測誤差が評価できる．

Gibbs アルゴリズムでは推定パラメータを許容領域内からランダムに選ぶので，実験では許容領域内からランダムに選んだ 500 個のパラメータの誤り確率を平均したものを予測誤差とした．重心アルゴリズムでは許容領域の重心を推定パラメータとするので，許容領域内からランダムに選んだ 500 個のパラメータの重心を求め，そのパラメータの誤り確率を予測誤差とした．

結果　図3.17は，例題数と Gibbs アルゴリズムの予測誤差の関係を次元 m が $m=1, m=3, m=5$ の場合について示したものである．エラーバーは30回の実験の標準偏差である．また，図3.18はパラメータの次元と Gibbs アルゴリズムの予測誤差の関係を示したものである．この二つのグラフから，Gibbs アルゴリズムの予測誤差の期待値 $\langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}}$ は

$$\langle \mathrm{GA} \rangle_{\boldsymbol{x}^{(t+1)}} = 0.66\frac{m}{t} \tag{3.103}$$

であるという結果が得られた．

図3.19は，例題数と重心アルゴリズムの予測誤差の関係を次元 m が $m=1, m=3, m=5$ の場合について示したものである．エラーバーは30回の実験の標準偏差である．また，図3.20はパラメータの次元と重心アルゴリズムの予測誤差の関係を示したものである．この二つのグラフから，重心アルゴリズムの予測誤差の期待値 $\langle \mathrm{GCA} \rangle_{\boldsymbol{x}^{(t+1)}}$ は

$$\langle \mathrm{GCA} \rangle_{\boldsymbol{x}^{(t+1)}} = 0.50\frac{m}{t} \tag{3.104}$$

であるという結果が得られた．

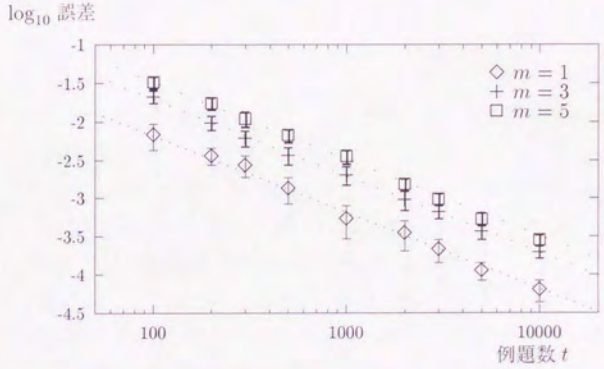


図 3.17: 与えられた例題数と予測誤差 (Gibbs アルゴリズム) ($m = 1, m = 3, m = 5$)

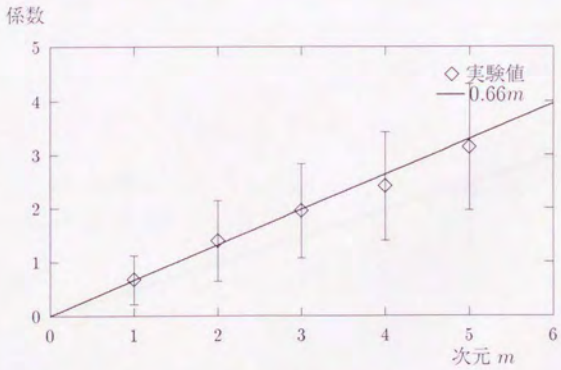


図 3.18: 次元 m と予測誤差の係数 (Gibbs アルゴリズム)

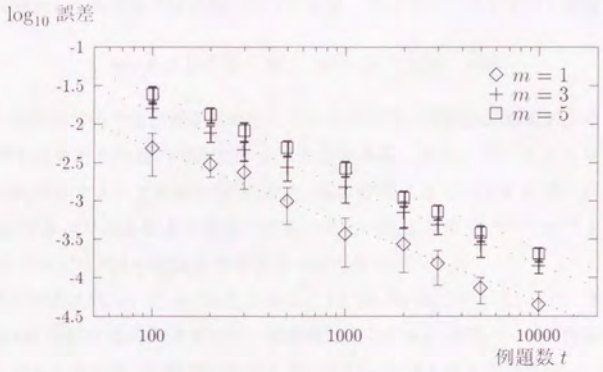


図 3.19: 与えられた例題数と予測誤差 (重心アルゴリズム) ($m = 1, m = 3, m = 5$)

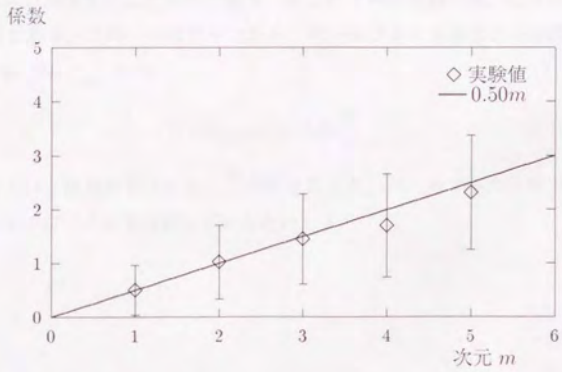


図 3.20: 次元 m と予測誤差の係数 (重心アルゴリズム)

実験 3.3 実験3.2では有効例題をすべて記憶し，さらに新しい例題が許容領域に交わる時にはその交点をすべて求める必要があるため，計算量が次元数に対して指数関数的に増加して高次元では計算できなくなる．そこでパーセプトロン学習

$$\boldsymbol{w}\cdot\overset{i}{\boldsymbol{x}}<0\text{の時}\quad \boldsymbol{w}\leftarrow\boldsymbol{w}-(1+\epsilon)(\boldsymbol{w}\cdot\overset{i}{\boldsymbol{x}})\overset{i}{\boldsymbol{x}} \tag{3.105}$$

を行い，高次元での予測誤差を求めた．パーセプトロン学習は計算量が少ないので，高次元でも実用的な時間内で計算することができる．また，パーセプトロン学習が止まった時のパラメータが許容領域内で一様に分布すると仮定すれば，パラメータは Gibbs アルゴリズムによる推定パラメータと一致するので，パーセプトロン学習は Gibbs アルゴリズムの近似とみなすことができる．

パラメータの次元 m を $m=1,2,3,5,7,10,20,30,50,70$ とし，100, 300, 1000, 3000, 10000, 30000 個の例題を与え，初期値はランダムに選び，一つの例題の組についてランダムに選んだ 20 個の初期値を用いてパーセプトロン学習を行い，学習が停止した時のパラメータがもつ誤り確率を 20 個のパラメータついて平均したものを予測誤差とした．実験では $\epsilon=0.5$ とし，各 m について例題を 50 組とった．

結果　図 3.21は，例題数と予測誤差の関係を次元 m が $m=1, m=3, m=10, m=30$ の場合について示したものである．エラーバーは 50 回の実験の標準偏差である．また，図 3.22はパラメータの次元と予測誤差の係数の関係を示したものであり，図 3.22を両対数軸で表したものが図 3.23である．この三つのグラフから，パーセプトロン学習の予測誤差の期待値 $\langle\mathrm{Perc}\rangle_{\boldsymbol{x}^{(t+1)}}$ は

$$\langle\mathrm{Perc}\rangle_{\boldsymbol{x}^{(t+1)}}=0.69\frac{m}{t} \tag{3.106}$$

であるという結果が得られた．この結果により，パーセプトロン学習は Gibbs アルゴリズムを近似しているといえる．

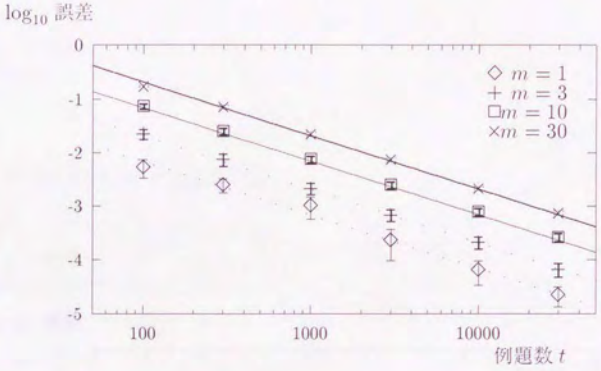


図 3.21: 与えられた例題数と予測誤差 ($m = 1, m = 3, m = 10, m = 30$)

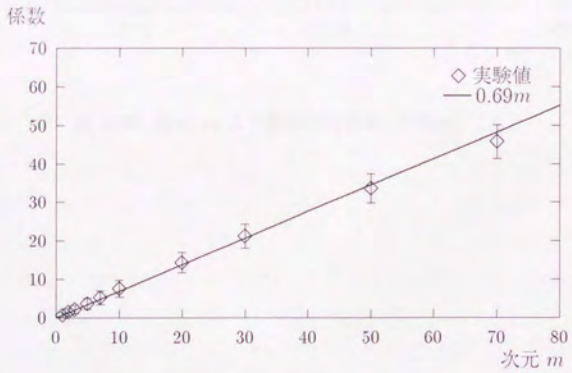


図 3.22: 次元 m と予測誤差の係数

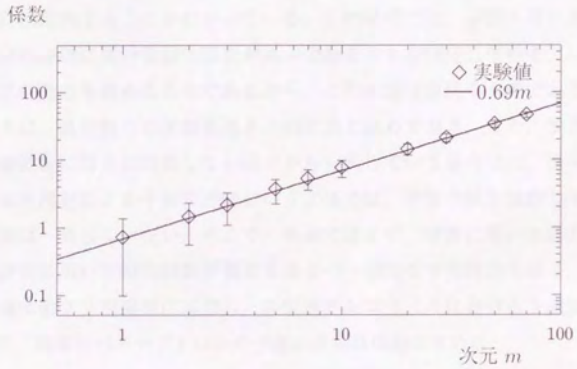


図 3.23: 次元 m と予測誤差の係数 (対数軸)

第 4 章

確率的機械の予測誤差

確率的機械の学習特性は Murata *et al.* [14] によって解析されており，平均予測損失が例題数に反比例することがわかっている．この研究では，学習に用いる損失関数と汎化能力の評価に用いる損失関数が同一であることを仮定している．一般に学習の目的は汎化能力を高めることであるから，この仮定は自然なものである．しかし本論のように，汎化能力の評価基準を予測誤差と決めておき，また，学習アルゴリズムを予測誤差には直接関係しない観点から定義している場合には，損失関数による学習と損失関数による予測の評価という立場では，学習の損失関数と予測の評価の損失関数は一致していない．そこで，本章ではまず，学習に用いる損失関数と汎化能力の評価に用いる損失関数が異なる場合の一般的な学習特性を導く．次に，その結果を確率的 2 分割機械に適用し，各学習アルゴリズムにおける予測誤差を求める．最後に，確率的パーセプトロンの予測誤差を具体的に求める．

4.1 問題設定

機械学習においては一般に，学習アルゴリズムが一つに定まるとは限らない．また，汎化能力の評価基準も一つではない．パーセプトロン学習においても，学習アルゴリズムには Gibbs アルゴリズムや Bayes アルゴリズムなどがあり，また汎化能力の評価基準も予測誤差や予測エントロピなどがある．すなわち，学習アルゴリズムは必ずしも汎化能力を最大にするものが用いられるわけではない．

本章では学習を，損失関数の小さくするパラメータを推定することと考える．推定パラメータを用いて予測を行う学習アルゴリズムがこの枠組みに含まれることは，後に示す．学習に用いる損失関数を学習損失関数と呼び，$d^{\mathrm{T}}(\boldsymbol{x}, y, \boldsymbol{w})$ と表す．機械の学習は，学習損失関数を用いて次のように定義される．

定義 4.1 学習とは，例題 $\xi^{(t)}$ によって作られる経験分布 $p_t(\boldsymbol{x}, y)$ に関する学習損失関数 $d^{\mathrm{T}}(\boldsymbol{x}, y, \boldsymbol{w})$ の期待値を小さくするパラメータを選ぶことである．

経験分布 $p_t(\boldsymbol{x}, y)$ は与えられた例題 $\xi^{(t)}$ が作る分布で

$$p_t(\boldsymbol{x}, y) = \frac{1}{t} \sum_i \delta(\boldsymbol{x} - \overset{i}{\boldsymbol{x}}, y - \overset{i}{y}) \tag{4.1}$$

である．ここで $\delta(\cdot, \cdot)$ は Dirac のデルタ関数である．上の定義から，経験分布に関する学習損失関数の期待値 d_t^{T} は，例題 $\xi^{(t)}$ とパラメータ $\boldsymbol{w}$ の関数として

$$\begin{aligned} d_t^{\mathrm{T}}(\xi^{(t)}, \boldsymbol{w}) &= \int d^{\mathrm{T}}(\boldsymbol{x}, y, \boldsymbol{w}) \frac{1}{t} \sum_i \delta(\boldsymbol{x} - \overset{i}{\boldsymbol{x}}, y - \overset{i}{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}y & (4.2) \\ &= \frac{1}{t} \sum_i d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}) & (4.3) \end{aligned}$$

と表される．$d_t^{\mathrm{T}}(\xi^{(t)}, \boldsymbol{w})$ を学習損失と呼ぶ．経験分布 $p_t(\boldsymbol{x}, y)$ は例題数 t が大きくなるにつれて真のパラメータ $\boldsymbol{w}^*$ による入出力の分布

$$p(\boldsymbol{x}, y \,|\, \boldsymbol{w}) = p(y \,|\, \boldsymbol{x}, \boldsymbol{w}^*) \, p(\boldsymbol{x})$$

に近づくので，真の分布に関する学習損失関数の期待値

$$d^{\mathrm{T}}(\boldsymbol{w}) = \int d^{\mathrm{T}}(\boldsymbol{x}, y, \boldsymbol{w}) p(y \,|\, \boldsymbol{x}, \boldsymbol{w}^*) \, p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}y \tag{4.4}$$

が，真のパラメータ $\boldsymbol{w}^*$ で最小値をとることを仮定する．

学習について，次のような仮定をする

仮定 4.1 推定パラメータ $\boldsymbol{w}_t$ は t が大きくなるにつれて，平均が $\boldsymbol{w}_t^*$，分散が $V^{\kappa\lambda}$ であるような正規分布に漸近する．ここで $\boldsymbol{w}_t^*$ は $d_t^{\mathrm{T}}(\xi^{(t)}, \boldsymbol{w})$ を最小にするパラメータである．

学習損失 $d_t^{\mathrm{T}}(\xi^{(t)}, \boldsymbol{w})$ を最小にするパラメータ $\boldsymbol{w}_t^*$ を，経験分布に対して d^{T}- 最適なパラメータと呼ぶ．また，分散 $V^{\kappa\lambda}$ を学習のばらつきと呼ぶ．

汎化能力を評価する基準は，学習に用いる損失関数と一致するとは限らない．汎化能力の評価に用いる損失関数を，予測損失関数と呼び，$d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w})$ と表す．推定パラメータ $\boldsymbol{w}_t$ をもつ機械の汎化能力は，真のパラメータ $\boldsymbol{w}^*$ による入出力の分布 $p(\boldsymbol{x}, y \,|\, \boldsymbol{w}^*)$ に関する予測損失関数 $d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w}_t)$ の期待値

$$d^{\mathrm{P}}(\boldsymbol{w}_t) = \int d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w}_t) p(y \,|\, \boldsymbol{x}, \boldsymbol{w}^*) \, p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}y \tag{4.5}$$

によって評価される．$d^{\mathrm{P}}(\boldsymbol{w}_t)$ を予測損失と呼び，予測損失の例題に関する平均を平均予測損失と呼ぶ．以下では，仮定4.1に従って推定パラメータ $\boldsymbol{w}_t$ が選ばれる時の平均予測損失を評価し，例題数 t と平均予測損失の関係を明らかにする．

以下では簡単のため，真のパラメータを持つ機械による入出力の分布に関する平均を $\langle f(\boldsymbol{x},y)\rangle_\xi$ と表す．すなわち

$$\langle f(\boldsymbol{x},y)\rangle_\xi = \int f(\boldsymbol{x},y)p(y|\boldsymbol{x},\boldsymbol{w}^*)\,p(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}y \tag{4.6}$$

である．この標記を用いると予測損失 $d^{\mathrm{P}}(\boldsymbol{w}_t)$ は

$$d^{\mathrm{P}}(\boldsymbol{w}_t) = \left\langle d^{\mathrm{P}}(\boldsymbol{x},y,\boldsymbol{w}_t)\right\rangle_\xi \tag{4.7}$$

と表される．

4.2 最適パラメータの分布と平均予測損失

本節では推定パラメータ $\boldsymbol{w}_t$ の分布を求め，それを利用して平均予測損失を求める．∂_κ をパラメータの第 κ 成分による微分とする，すなわち

$$\partial_\kappa d^{\mathrm{T}}(\boldsymbol{x},y,\boldsymbol{w}) = \frac{\partial}{\partial \boldsymbol{w}^\kappa} d^{\mathrm{T}}(\boldsymbol{x},y,\boldsymbol{w}) \tag{4.8}$$

とし，$G_{\kappa\lambda}(\boldsymbol{w}), Q_{\kappa\lambda}(\boldsymbol{w}), S_{\kappa\lambda\mu}(\boldsymbol{w})$ を

$$G_{\kappa\lambda}(\boldsymbol{w}) = \left\langle \partial_\kappa d^{\mathrm{T}}(\boldsymbol{x},y,\boldsymbol{w})\partial_\lambda d^{\mathrm{T}}(\boldsymbol{x},y,\boldsymbol{w})\right\rangle_\xi, \tag{4.9}$$

$$Q_{\kappa\lambda}(\boldsymbol{w}) = \left\langle \partial_\kappa\partial_\lambda d^{\mathrm{T}}(\boldsymbol{x},y,\boldsymbol{w})\right\rangle_\xi, \tag{4.10}$$

$$S_{\kappa\lambda\mu}(\boldsymbol{w}) = \left\langle \partial_\kappa\partial_\lambda\partial_\mu d^{\mathrm{T}}(\boldsymbol{x},y,\boldsymbol{w})\right\rangle_\xi, \tag{4.11}$$

と定義する．$G_{\kappa\lambda} = G_{\kappa\lambda}(\boldsymbol{w}^*), Q_{\kappa\lambda} = Q_{\kappa\lambda}(\boldsymbol{w}^*), S_{\kappa\lambda\mu} = S_{\kappa\lambda\mu}(\boldsymbol{w}^*)$ とすれば，推定パラメータの分布について，次の定理が成り立つ:

定理 4.1 推定パラメータ $\boldsymbol{w}_t$ は漸近的に，正規分布

$$N\left(\boldsymbol{w}^{*\kappa} - \frac{1}{2t}Q^{\kappa\kappa'}S_{\kappa'\lambda'\mu'}Q^{\lambda\lambda'}Q^{\mu\mu'}G_{\lambda\mu}, V^{\kappa\lambda} + \frac{1}{t}Q^{\kappa\kappa'}Q^{\lambda\lambda'}G_{\kappa'\lambda'}\right) \tag{4.12}$$

に従う．ただし添字は Einstein の記法であり，同じものについて和をとるものとする．

証明 まず，例題が真の機械による分布 $p(\boldsymbol{x},y|\boldsymbol{w}^*)$ で独立に選ばれる時，経験分布に対して d^{T}- 最適なパラメータ $\boldsymbol{w}_t^*$ がどのように分するのかを調べる．以下では θ^λ で $(\boldsymbol{w}_t^* - \boldsymbol{w}^*)$ の 第 λ 成分を表す．

$\boldsymbol{w}_t^*$ は 学習損失 $d_t^{\mathrm{T}}(\boldsymbol{w})$ を最小とするからパラメータに関する微係数は 0 である．すなわち

$$\begin{aligned}\partial_\kappa d_t^{\mathrm{T}}(\xi^{(t)}, \boldsymbol{w}_t^*) &= \int \partial_\kappa d^{\mathrm{T}}(\boldsymbol{x}, y, \boldsymbol{w}_t^*) \frac{1}{t} \sum_i \delta(\boldsymbol{x} - \overset{i}{\boldsymbol{x}}, y - \overset{i}{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}y \\ &= \frac{1}{t} \sum_i \partial_\kappa d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}_t^*) \qquad (4.13) \\ &= 0\end{aligned}$$

が成り立つ．これを真のパラメータ $\boldsymbol{w}^*$ のまわりで展開すると

$$\begin{aligned}0 &= \frac{1}{t} \sum_i \partial_\kappa d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}^*) + \frac{1}{t} \sum_i \partial_\kappa \partial_\lambda d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}^*) \theta^\lambda \\ &\quad + \frac{1}{2t} \sum_i \partial_\kappa \partial_\lambda \partial_\mu d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}^*) \theta^\lambda \theta^\mu + \mathrm{o}(|\theta|^2) \qquad (4.14)\end{aligned}$$

が得られる．ここで例題 $\xi^{(t)}$ に関する平均をとると，t が大きい時は大数の法則により

$$\left\langle \frac{1}{t} \sum_i \partial_\kappa \partial_\lambda d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}^*) \right\rangle_{\xi^{(t)}} = Q_{\kappa\lambda} + \mathrm{o}(1) \qquad (4.15)$$

$$\left\langle \frac{1}{t} \sum_i \partial_\kappa \partial_\lambda \partial_\mu d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}^*) \right\rangle_{\xi^{(t)}} = S_{\kappa\lambda\mu} + \mathrm{o}(1) \qquad (4.16)$$

である．また，中心極限定理により

$$\frac{1}{\sqrt{t}} \sum_i \partial_\kappa d^{\mathrm{T}}(\overset{i}{\boldsymbol{x}}, \overset{i}{y}, \boldsymbol{w}^*) \qquad (4.17)$$

の分布は平均 0, 分散 $G_{\lambda\mu}$ の正規分布に漸近する．よって高次の項を無視すると，例題に対して d^{T}- 最適なパラメータ $\boldsymbol{w}_t^*$ の分布は，正規分布

$$N\left(\boldsymbol{w}^{*\kappa} - \frac{1}{2t} Q^{\kappa\kappa'} S_{\kappa'\lambda'\mu'} Q^{\lambda\lambda'} Q^{\mu\mu'} G_{\lambda\mu}, \frac{1}{t} Q^{\kappa\kappa'} Q^{\lambda\lambda'} G_{\kappa'\lambda'}\right) \qquad (4.18)$$

に漸近する．

推定パラメータ $\boldsymbol{w}_t$ は仮定 4.1 において，平均 $\boldsymbol{w}_t^*$, 分散 $V^{\kappa\lambda}$ の正規分布に従うとしたので，推定パラメータの分布は漸近的に

$$N\left(\boldsymbol{w}^{*\kappa} - \frac{1}{2t} Q^{\kappa\kappa'} S_{\kappa'\lambda'\mu'} Q^{\lambda\lambda'} Q^{\mu\mu'} G_{\lambda\mu}, V^{\kappa\lambda} + \frac{1}{t} Q^{\kappa\kappa'} Q^{\lambda\lambda'} G_{\kappa'\lambda'}\right) \qquad (4.19)$$

となる． □

推定パラメータの分布が求められたので，平均予測損失，すなわち予測損失の例題に関する平均を求めることができる．

定理 4.2 予測損失関数が $d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w})$ の時，平均予測損失 $\left\langle d^{\mathrm{P}}(\boldsymbol{w}_t)\right\rangle_{\xi^{(t)}}$ は

$$d^{\mathrm{P}}(\boldsymbol{w}^*) + \frac{1}{2}Q'_{\lambda\mu}V^{\lambda\mu} + \frac{1}{2t}\left(Q'_{\lambda\mu} - A_\kappa Q^{\kappa\kappa'}S_{\kappa'\lambda\mu}\right)Q^{\lambda\lambda'}Q^{\mu\mu'}G_{\lambda'\mu'} \tag{4.20}$$

に漸近する．ただし

$$d^{\mathrm{P}}(\boldsymbol{w}^*) = \left\langle d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w}^*)\right\rangle_{\xi}, \tag{4.21}$$

$$A_\kappa = \left\langle \partial_\kappa d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w}^*)\right\rangle_{\xi}, \tag{4.22}$$

$$Q'_{\kappa\lambda} = \left\langle \partial_\kappa\partial_\lambda d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w}^*)\right\rangle_{\xi}, \tag{4.23}$$

とする．

証明 平均予測損失 $\left\langle d^{\mathrm{P}}(\boldsymbol{w}_t)\right\rangle_{\xi^{(t)}}$ は

$$\left\langle d^{\mathrm{P}}(\boldsymbol{w}_t)\right\rangle_{\xi^{(t)}} = \int d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w}_t)p(y|\boldsymbol{x}, \boldsymbol{w}^*)\,p(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}\,\mathrm{d}y \tag{4.24}$$

であるので，予測損失関数 $d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w}_t)$ を真のパラメータ $\boldsymbol{w}^*$ の周りで展開すると，

$$\begin{aligned}\left\langle d^{\mathrm{P}}(\boldsymbol{w}_t)\right\rangle_{\xi^{(t)}} &= d^{\mathrm{P}}(\boldsymbol{w}^*) + A_\kappa\langle(\boldsymbol{w}_t - \boldsymbol{w}^*)^\kappa\rangle_{\xi^{(t)}} \\ &\quad + \frac{1}{2}Q'_{\kappa\lambda}\left\langle(\boldsymbol{w}_t - \boldsymbol{w}^*)^\kappa(\boldsymbol{w}_t - \boldsymbol{w}^*)^\lambda\right\rangle_{\xi^{(t)}} \qquad (4.25)\\ &= d^{\mathrm{P}}(\boldsymbol{w}^*) + \frac{1}{2}Q'_{\lambda\mu}V^{\lambda\mu} \\ &\quad + \frac{1}{2t}\left(Q'_{\lambda\mu} - A_\kappa Q^{\kappa\kappa'}S_{\kappa'\lambda\mu}\right)Q^{\lambda\lambda'}Q^{\mu\mu'}G_{\lambda'\mu'} \qquad (4.26)\end{aligned}$$

であり，定理が示される．ただし，定理4.1で得られた $\boldsymbol{w}_t$ の分布を用いている． □

真のパラメータ $\boldsymbol{w}^*$ は，真の分布に対して d^{T}-最適なパラメータ $\boldsymbol{w}^*$ であることを仮定していた．この $\boldsymbol{w}^*$ が，真の分布に対して d^{P}-最適なパラメータにもなっている時は，パラメータに関する微分係数の期待値が $\mathrm{o}(1)$ になるので，$A_\kappa = 0$ としてよい．この時，定理 4.2 の平均予測損失は $S_{\kappa\lambda\mu}$ の項が消え，$G_{\kappa\lambda}, Q_{\kappa\lambda}, Q'_{\kappa\lambda}$ だけで表される．これを系としてまとめておく．

系 4.3 真の分布に対して d^{T}-最適なパラメータ $\boldsymbol{w}^*$ が，真の分布に対して d^{P}-最適なパラメータにもなっている時，平均予測損失は

$$\left\langle d^{\mathrm{P}}(\boldsymbol{w}_t)\right\rangle_{\xi^{(t)}} = d^{\mathrm{P}}(\boldsymbol{w}^*) + \frac{1}{2}Q'_{\lambda\mu}V^{\lambda\mu} + \frac{1}{2t}Q'_{\lambda\mu}Q^{\lambda\lambda'}Q^{\mu\mu'}G_{\lambda'\mu'} \tag{4.27}$$

に漸近する．

学習損失関数 $d^{\mathrm{T}}(\boldsymbol{x}, y, \boldsymbol{w})$ と予測損失関数 $d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w})$ が同一の場合，すなわち系4.3において $Q_{\kappa\lambda} = Q'_{\kappa\lambda}$ の場合には Murata *et al.* [14] の結果と一致し，

$$\left\langle d^{\mathrm{P}}(\boldsymbol{w}_t) \right\rangle_{\xi^{(t)}} = d^{\mathrm{P}}(\boldsymbol{w}^*) + \frac{1}{2} Q'_{\lambda\mu} V^{\lambda\mu} + \frac{1}{2t} Q^{\lambda\mu} G_{\lambda\mu} \tag{4.28}$$

となる．

4.3 学習アルゴリズムと損失関数

本節では，第2.2節で定義した各アルゴリズムによる学習が仮定4.1の枠組にあてはまることを示し，系4.3を用いて確率的2分割機械の平均予測誤差を求める．まず，二つの損失関数を定義する．

定義 4.2 二つの損失関数 $d_1(\boldsymbol{x}, y, \boldsymbol{w}), d_2(\boldsymbol{x}, y, \boldsymbol{w})$ を

$$d_1(\boldsymbol{x}, y, \boldsymbol{w}) = -\log p(y|\boldsymbol{x}, \boldsymbol{w}), \tag{4.29}$$

$$d_2(\boldsymbol{x}, y, \boldsymbol{w}) = 1 - p(y|\boldsymbol{x}, \boldsymbol{w}). \tag{4.30}$$

と定義し，それぞれ対数尤度，誤り確率と呼ぶ．

各損失関数から導かれる G, Q をそれぞれ G, Q, G', Q' と表すことにする．この時，$G = Q$ が成り立つ．

$d_1(\boldsymbol{x}, y, \boldsymbol{w})$ は尤度の対数であるので，経験分布に関して d_1-最適なパラメータは最尤推定パラメータとなる．また $d_2(\boldsymbol{x}, y, \boldsymbol{w})$ を予測損失関数として用いると，予測損失 $DDP(\boldsymbol{w}_t)$ は

$$d^{\mathrm{P}}(\boldsymbol{w}_t) = \langle (1 - p(y|\boldsymbol{x}, \boldsymbol{w}_t)) \rangle_{\xi} \tag{4.31}$$

であり，これは $\boldsymbol{w}_t$ を推定パラメータとして予測をした時に予測が誤りである確率の期待値すなわち予測誤差となっている．よって以下では，予測損失関数 $d^{\mathrm{P}}(\boldsymbol{x}, y, \boldsymbol{w})$ として $d_2(\boldsymbol{x}, y, \boldsymbol{w})$ を用いる．

各アルゴリズムはパラメータの事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ を用いて定義されているので，まずはじめに，パラメータの事後分布を求める．

定理 4.4 (パラメータの事後分布) パラメータの事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ は最尤推定パラメータ $\hat{\boldsymbol{w}}$ を中心とした正規分布 $N\left(\hat{\boldsymbol{w}}^{\kappa}, \frac{1}{t} Q(\hat{\boldsymbol{w}})^{\kappa\lambda}\right)$ に漸近する．

証明 $\xi^{(t)}$ が固定されているとする．機械のパラメータが $\boldsymbol{w}$ で，かつ出力 $y^{(t)}$ が生成される確率密度 $p\left(y^{(t)}, \boldsymbol{w} \middle| \boldsymbol{x}^{(t)}\right)$ は

$$p\left(y^{(t)}, \boldsymbol{w} \middle| \boldsymbol{x}^{(t)}\right) = p(\boldsymbol{w}) \prod_i p\left(\begin{matrix} i \\ y \end{matrix} \middle| \begin{matrix} i \\ \boldsymbol{x} \end{matrix}, \boldsymbol{w}\right) \tag{4.32}$$

であるので，$\boldsymbol{w}$ の事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ は Bayes の定理により

$$p(\boldsymbol{w}|\xi^{(t)}) = \frac{p(\boldsymbol{w})\prod_i p\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\boldsymbol{w}\right)}{Z_t}$$

である．ここで Z_t は

$$Z_t = \int p(\boldsymbol{w})\prod_i p\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\boldsymbol{w}\right)\,\mathrm{d}\boldsymbol{w}$$

である．

$\log p(\,\cdot\,)$ を $l(\,\cdot\,)$ で略記すると，$\log p\left(y^{(t)},\boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right)$ を例題 $\xi^{(t)}$ に関する最尤推定パラメータ $\hat{\boldsymbol{w}}$ のまわりでの展開は

$$\begin{aligned}
l\left(y^{(t)},\boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right) &= \log p\left(y^{(t)},\boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right) && (4.33)\\
&= \log p(\boldsymbol{w}) + \sum_i l\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\boldsymbol{w}\right) && (4.34)\\
&= \log p(\boldsymbol{w}) + \sum_i l\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\hat{\boldsymbol{w}}\right)\\
&\quad + \sum_i \partial_\kappa l\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\hat{\boldsymbol{w}}\right)(\boldsymbol{w}-\hat{\boldsymbol{w}})^\kappa\\
&\quad + \frac{1}{2}\sum_i \partial_\kappa\partial_\lambda l\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\hat{\boldsymbol{w}}\right)(\boldsymbol{w}-\hat{\boldsymbol{w}})^\kappa(\boldsymbol{w}-\hat{\boldsymbol{w}})^\lambda\\
&\quad + \mathrm{o}(|\boldsymbol{w}-\hat{\boldsymbol{w}}|^2) && (4.35)
\end{aligned}$$

となる．$\hat{\boldsymbol{w}}$ は最尤推定パラメータなので $\sum_i \partial_\kappa l\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\hat{\boldsymbol{w}}\right) = 0$ であり，また大数の法則により

$$\frac{1}{t}\sum_i \partial_\kappa\partial_\lambda l\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\hat{\boldsymbol{w}}\right) \approx -Q(\hat{\boldsymbol{w}}) \qquad (4.36)$$

である．よって

$$\begin{aligned}
p\left(y^{(t)},\boldsymbol{w}\middle|\boldsymbol{x}^{(t)}\right) &= p(\boldsymbol{w})\prod_i p\left(\overset{i}{y}\middle|\overset{i}{\boldsymbol{x}},\hat{\boldsymbol{w}}\right)\\
&\quad \times \exp\left[-\frac{t}{2}Q(\hat{\boldsymbol{w}})(\boldsymbol{w}-\hat{\boldsymbol{w}})^2\right] && (4.37)
\end{aligned}$$

であり，これを $\boldsymbol{w}$ に関して積分して Z_t を求めると，事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ は漸近的に

$$p(\boldsymbol{w}|\xi^{(t)}) = t^{\frac{m}{2}}|Q|^{\frac{1}{2}}\exp\left[-\frac{t}{2}Q(\hat{\boldsymbol{w}})(\boldsymbol{w}-\hat{\boldsymbol{w}})^2\right] \qquad (4.38)$$

である． □

以下では Gibbs アルゴリズム，最尤推定アルゴリズム，重心アルゴリズムについて，学習が各アルゴリズムと一致するように学習損失関数 $d^{\Gamma}(\boldsymbol{x}, y, \boldsymbol{w})$ と学習のばらつき $V^{\kappa\lambda}$ を定める．

Gibbs アルゴリズムでは，事後分布 $p(\boldsymbol{w}|\xi^{(t)})$ に従って選んだパラメータを推定パラメータ $\boldsymbol{w}_t$ として予測をする．パラメータの事後分布の平均は定理 4.4により $\hat{\boldsymbol{w}}$ なので学習損失関数を対数尤度 $d_1(\boldsymbol{x}, y, \boldsymbol{w})$ とし，パラメータの事後分布の分散は定理4.4により $\frac{1}{t}Q(\hat{\boldsymbol{w}})^{\kappa\lambda}$ なので学習のばらつき $V^{\kappa\lambda}$ を $V^{\kappa\lambda} = \frac{1}{t}Q(\hat{\boldsymbol{w}})^{\kappa\lambda}$ とすれば，学習損失関数を用いた学習による推定パラメータの分布は Gibbs アルゴリズムによる推定パラメータの分布と一致し，

$$N\left(\hat{\boldsymbol{w}}^{\kappa}, \frac{1}{t}\left(Q(\hat{\boldsymbol{w}})^{\kappa\lambda} + Q(\boldsymbol{w}^*)^{\kappa\lambda}\right)\right)$$

となる．よって系4.3に代入し，さらに $Q(\hat{\boldsymbol{w}})^{\kappa\lambda}$ を $\boldsymbol{w}^*$ の回りで展開して $\xi^{(t)}$ に関する平均をとると，平均予測誤差 $\langle d_2(\boldsymbol{w}_t)\rangle_{\xi^{(t)}}$ は

$$\langle d_2(\boldsymbol{w}_t)\rangle_{\xi^{(t)}} = d_2(\boldsymbol{w}^*) + \frac{1}{t}Q'_{\kappa\lambda}Q^{\kappa\lambda} \tag{4.39}$$

となる．

最尤推定アルゴリズムでは，最尤推定したパラメータを推定パラメータ $\boldsymbol{w}_t$ として予測をする．最尤推定パラメータ $\hat{\boldsymbol{w}}$ をそのまま推定パラメータ $\boldsymbol{w}_t$ とするので，学習損失関数を対数尤度 d_1 とし，学習のばらつき $V^{\kappa\lambda}$ を $V^{\kappa\lambda} = 0$ とすれば，学習損失関数を用いた学習の推定パラメータの分布は最尤推定アルゴリズムの推定パラメータの分布と一致し，

$$N\left(\hat{\boldsymbol{w}}^{\kappa}, \frac{1}{t}Q(\boldsymbol{w}^*)^{\kappa\lambda}\right)$$

となる．よって平均予測誤差 $\langle d_2(\boldsymbol{w}_t)\rangle_{\xi^{(t)}}$ は

$$\langle d_2(\boldsymbol{w}_t)\rangle_{\xi^{(t)}} = d_2(\boldsymbol{w}^*) + \frac{1}{2t}Q'_{\kappa\lambda}Q^{\kappa\lambda} \tag{4.40}$$

となる．

重心アルゴリズムでは，事後分布 $p(\boldsymbol{w}, \xi^{(t)})$ で重み付けした時のパラメータの重心を推定パラメータ $\boldsymbol{w}_t$ とするアルゴリズムである．定理 4.4により，パラメータの事後分布は正規分布 $N\left(\hat{\boldsymbol{w}}^{\kappa}, \frac{1}{t}Q(\hat{\boldsymbol{w}})^{\kappa\lambda}\right)$ に漸近するので，パラメータの重心は最尤推定パラメータ $\hat{\boldsymbol{w}}$ と一致する．よって重心アルゴリズムの予測は最尤推定アルゴリズムの予測と一致する．

4.4 確率的パーセプトロンの予測誤差

本章の議論では確率密度関数 $p(y|\boldsymbol{x},\boldsymbol{w})$ がパラメータ $\boldsymbol{w}$ で微分できることが必要であるから，そのまま確定的機械に当てはめることはできない．そこで，温度パラメータ β を持つ，次のような確率的2分割機械(確率的パーセプトロン) について，具体的に予測誤差を求める．確率的パーセプトロンは $\beta \to \infty$ の極限で，確定的機械である単純パーセプトロンと一致する．

定義 4.3 (確率的パーセプトロン) 次のような入出力関係を持つ確率的2分割機械を確率的パーセプトロンと呼ぶ：

$$p(y=+1|\boldsymbol{x},\boldsymbol{w}) = \frac{1}{2}(1+k(\beta\boldsymbol{x}\cdot\boldsymbol{w})), \tag{4.41}$$

$$p(y=-1|\boldsymbol{x},\boldsymbol{w}) = \frac{1}{2}(1-k(\beta\boldsymbol{x}\cdot\boldsymbol{w})). \tag{4.42}$$

ここで β は温度の逆数を表すパラメータであり，k は $k:(-\infty,\infty)\to(-1,1)$ の単調増加する2階微分可能な奇関数とする．

まず $d_2(\boldsymbol{w}^*)$ を求める．パラメータ空間 $\boldsymbol{S}^m$ に $\boldsymbol{w}^*$ を原点とする球面上の極座標 $(r,\boldsymbol{\omega})$ を導入し，積分変数の変換を行うと

$$d_2(\boldsymbol{w}^*) = \frac{1}{I_m}\int_0^{\frac{\pi}{2}}\frac{1}{2}\left(1-k(\beta\cos r)^2\right)\sin^{m-1}r\,\mathrm{d}r \tag{4.43}$$

$$= \frac{1}{2I_m}\int_0^1\left(1-k(\beta f)^2\right)\left(1-f^2\right)^{\frac{m-2}{2}}\,\mathrm{d}f \tag{4.44}$$

が得られる．

次に，$G_{\kappa\lambda}, G'_{\kappa\lambda}, Q'_{\kappa\lambda}$ を求める．まず，入力 $\boldsymbol{x}$ 及びパラメータ $\boldsymbol{w}$ を $\boldsymbol{R}^{m+1}$ のベクトルとみなし，それぞれ

$$\boldsymbol{x} = \begin{pmatrix} X \\ \sqrt{1-|X|^2} \end{pmatrix} \qquad \boldsymbol{w} = \begin{pmatrix} W \\ \sqrt{1-|W|^2} \end{pmatrix}$$

という座標を入れる．ただし

$$\boldsymbol{w}^* = \begin{pmatrix} \boldsymbol{0} \\ 1 \end{pmatrix}$$

とする．この時 X は m 次元超球 $\boldsymbol{D}^m$ の元である．

$$\boldsymbol{w}\cdot\boldsymbol{x} = W\cdot X + \sqrt{1-|W|^2}\sqrt{1-|X|^2} \tag{4.45}$$

であるので，$G_{\kappa\lambda}(\boldsymbol{w})$ は $\boldsymbol{w}=\boldsymbol{w}^*$ では

$$G_{\kappa\lambda}(\boldsymbol{w}^*) = \left\langle \langle \partial_\kappa l \partial_\lambda l \rangle_y \big|_{\boldsymbol{w}=\boldsymbol{0}} \right\rangle_{\boldsymbol{x}} \tag{4.46}$$

$$= \int \frac{\beta^2 k'(\beta\sqrt{1-|X|^2})^2}{1-k(\beta\sqrt{1-|X|^2})^2} X_\kappa X_\lambda p(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \tag{4.47}$$

$$= g\ \delta_{\kappa\lambda} \tag{4.48}$$

と書ける．ここで $\delta_{\kappa\lambda}$ は $\kappa=\lambda$ の時に 1，そうでない時に 0 となる関数である．X に極座標 $(\sin r, \boldsymbol{\omega})$ を導入して考えると，次の補題が成り立つ:

補題 4.1 半径 $\sin r$ の $\boldsymbol{S}^{m-1}$ 上での X_κ^2 の平均は

$$\left\langle X_\kappa^2 \right\rangle = \frac{\sin^2 r}{m} \tag{4.49}$$

である．

証明

$$\left\langle X_\kappa^2 \right\rangle = \frac{\int_0^{\frac{\pi}{2}} \cos^2 r \sin^{m-2} r\ \mathrm{d}r}{\int_0^{\frac{\pi}{2}} \sin^{m-2} r\ \mathrm{d}r} \tag{4.50}$$

$$= \frac{I_{m-1}-I_{m+1}}{I_{m-1}} \tag{4.51}$$

であり，さらに

$$I_{m-1}-I_{m+1} = \int_0^{\frac{\pi}{2}} \cos^2 r \sin^{m-2} \tag{4.52}$$

$$= \frac{1}{m-1} I_{m+1} \tag{4.53}$$

により，

$$\left\langle X_\kappa^2 \right\rangle = \frac{\sin^2 r}{m} \tag{4.54}$$

である． □

この補題を利用すると，g が計算でき，

$$g = \frac{1}{2I_m} \int_0^\pi \frac{\beta^2 k'(\beta\cos r)^2}{1-k(\beta\cos r)^2} \left\langle X_\kappa^2 \right\rangle \sin^{m-1} r\ \mathrm{d}r \tag{4.55}$$

$$= \frac{1}{2mI_m} \int_0^\pi \frac{\beta^2 k'(\beta\cos r)^2}{1-k(\beta\cos r)^2} \sin^{m+1} r\ \mathrm{d}r \tag{4.56}$$

$$= \frac{\beta^2}{2mI_m} \int_{-1}^1 \frac{k'(\beta f)^2}{1-k(\beta f)^2} \left(1-f^2\right)^{\frac{m}{2}}\ \mathrm{d}f \tag{4.57}$$

が得られる．

同様にして $Q'_{\kappa\lambda}(\boldsymbol{w}^*)$ は

$$Q'_{\kappa\lambda}(\boldsymbol{w}^*) = \int -\frac{\beta^2}{2} k(\beta\sqrt{1-|X|^2}) k''(\beta\sqrt{1-|X|^2}) X_\kappa X_\lambda p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad (4.58)$$

$$= q' \, \delta_{\kappa\lambda} \quad (4.59)$$

と書け，g と同様の計算により

$$q' = -\frac{\beta^2}{4mI_m} \int_{-1}^{1} k(\beta f) k''(\beta f) \left(1-f^2\right)^{\frac{m}{2}} \, \mathrm{d}f \quad (4.60)$$

が得られる．

以上をまとめると，次の定理が得られる．

定理 4.5 Gibbs アルゴリズム，最尤推定アルゴリズムにおける確率的パーセプトロンの予測誤差について，次の式が漸近的に成り立つ:

$$d_2(\boldsymbol{w}^*) = \frac{1}{2I_m\beta} \int_0^1 \left(1-k(\beta f)^2\right) \left(1-f^2\right)^{\frac{m-2}{2}} \, \mathrm{d}f, \quad (4.61)$$

$$\langle \mathrm{GA} \rangle = d_2(\boldsymbol{w}^*) + \frac{m}{t} \cdot \frac{-\frac{1}{2}\int_0^1 k(\beta f) k''(\beta f) \left(1-f^2\right)^{\frac{m}{2}} \, \mathrm{d}f}{\int_0^1 \frac{f k'(\beta f)^2}{1-k(\beta f)^2} \left(1-f^2\right)^{\frac{m}{2}} \, \mathrm{d}f}, \quad (4.62)$$

$$\langle \mathrm{MLEA} \rangle = d_2(\boldsymbol{w}^*) + \frac{m}{2t} \cdot \frac{-\frac{1}{2}\int_0^1 k(\beta f) k''(\beta f) \left(1-f^2\right)^{\frac{m}{2}} \, \mathrm{d}f}{\int_0^1 \frac{f k'(\beta f)^2}{1-k(\beta f)^2} \left(1-f^2\right)^{\frac{m}{2}} \, \mathrm{d}f}. \quad (4.63)$$

例えば $k(x)$ を Sigmond 関数 $k(x) = \tanh x$ とすると，数値積分により，各値は表4.1のようになる．表4.1から，次元 m が変わっても $\frac{m}{t}$ の係数はほとんど変わらないことがわかる．確率的パーセプトロンは温度パラメータ $\beta \to \infty$ の極限で単純パーセプトロンに収束する．しかし β が大きくなると統計的漸近理論が破綻するため，確率的パーセプトロンの予測誤差が $\beta \to \infty$ の極限で単純パーセプトロンの予測誤差に収束するとはいえない．今回求めた確率的パーセプトロンの予測誤差と単純パーセプトロンの予測誤差との関係は，未だ明らかにはなっていない．

	GA	MLEA
m=2	$0.2500\frac{1}{\beta}+0.3333\frac{m}{t}$	$0.2500\frac{1}{\beta}+0.1667\frac{m}{t}$
3	$0.3183\frac{1}{\beta}+0.3333\frac{m}{t}$	$0.3183\frac{1}{\beta}+0.1667\frac{m}{t}$
5	$0.4244\frac{1}{\beta}+0.3333\frac{m}{t}$	$0.4244\frac{1}{\beta}+0.1667\frac{m}{t}$
10	$0.6152\frac{1}{\beta}+0.3333\frac{m}{t}$	$0.6152\frac{1}{\beta}+0.1667\frac{m}{t}$
50	$1.4034\frac{1}{\beta}+0.3333\frac{m}{t}$	$1.4034\frac{1}{\beta}+0.1667\frac{m}{t}$
200	$2.8172\frac{1}{\beta}+0.3333\frac{m}{t}$	$2.8172\frac{1}{\beta}+0.1667\frac{m}{t}$
500	$4.4571\frac{1}{\beta}+0.3333\frac{m}{t}$	$4.4571\frac{1}{\beta}+0.1666\frac{m}{t}$

表 4.1: 確率的パーセプトロンの予測誤差 ($\beta = 1000$)

第 5 章

まとめ

本論では，パラメータが冗長でない確定的機械は単純パーセプトロンで近似できることを示し，単純パーセプトロンについて許容領域の性質を統計幾何学的に解析して許容領域の頂点数及び境界面数の期待値は有界であることを示した．また，その結果と幾何学的考察から，予測誤差の新しいバウンドを求めた．今回求めた $m \leq 3$ については従来より良いバウンドを与えたが，この手法は，必ずしも良いバウンドを与えるとは限らない．また，初等幾何的な性質を利用するため，高次元への応用が難しいという欠点もある．しかし，予測誤差以外の問題への応用の余地はあると考えている．

また，統計幾何学的手法では求められなかった，パラメータが高次元の場合の予測誤差を，計算機実験によって求めた．さらに，パーセプトロン学習が Gibbs アルゴリズムのよい近似となっていることを実験的に示した．

確率的機械については，既存の理論を拡張し，2 分割機械に限らず一般の機械について，学習損失関数と予測損失関数が異なる場合についての損失の期待値を評価した．この枠組みは非常に広いものであり，他の問題へも応用が可能である．しかし，確率的パーセプトロンの例にみられるように，実際の計算は容易とは限らないという問題がある．

確率的パーセプトロンは，確定的出力をする単純パーセプトロンの近似とみなすことができる．しかし確率的パーセプトロンでは漸近理論を利用しているため，単に温度の逆数パラメータ β を無限大にして予測誤差の極限を求めても，単純パーセプトロンの予測誤差に収束するとは限らない．この点については慎重な議論が必要であるが，確率的パーセプトロンの予測誤差がなんらかのバウンドを与える可能性は大きい．確率的機械の方が確定的機械よりも扱い易いことを考慮すると，今後検討していくべき課題である．

謝辞

大学学部生の時から大学院の修士課程，博士課程と長きにわたり大変お世話になった吉澤修治教授に感謝します．吉澤先生は，研究の方向性を決めかねていた私にさりげなくプレッシャーをかけ，研究を進める原動力となってくださいました．学習の統計的理論という研究分野を紹介して下さり，また多忙にも関わらず快く議論につきあってくださった甘利俊一教授に感謝します．また，本論文の審査にあたり，ご助言を下さった伏見正則教授，中野馨助教授，合原一幸助教授に感謝します．

本研究を進めるにあたり，多くの有益な助言，提案をくださった村田昇助手，そして川鍋元明君に感謝します．

5年間の研究室での生活を楽しく思い出深いものとしてくれた，藤原彰夫助手，院生諸氏，秘書の方々に感謝します．

最後に，9年にわたる大学，大学院での学生生活に，惜しみない援助を与えて下さった両親に感謝します．

参考文献

[1] Amari, S.: Theory of Adaptive Pattern Classifiers, *IEEE Trans. EC*, Vol. 16 (1967), 299–307.

[2] Amari, S.: A Universal Theorem on Learning Curves, *Neural Networks*, Vol. 6 (1993), 161–166.

[3] Amari, S., Fujita, N. and Shinomoto, S.: Four Types of Learning Curves, *Neural Computation*, Vol. 4 (1992), 604–617.

[4] Amari, S. and Murata, N.: Statistical Theory of Learning Curves under Entropic Loss Criterion, *Neural Computation*, Vol. 5 (1993), 140–153.

[5] Baum, E. B. and Haussler, D.: What Size Net Gives Valid Generalization?, *Neural Computation*, Vol. 1 (1989), 151–160.

[6] Block, H. D.: The Perceptron — A Model for Brain Functioning I, *Review of Modern Physics*, Vol. 34 (1962), 123–135.

[7] Cover, T. M.: Geometrical and Statistical Properties of Linear Threshold Devices, SEL 64-052, Stanford Univ., 1964.

[8] Efron, B.: The Convex Hull of a Random Set of Points, *Biometrika*, Vol. 52 (1965), 331–343.

[9] Haussler, D., Kearns, M. and Schapire, R.: Bounds on the Sample Complexity of Bayesian Learning, UCSC-CRL 91-44, UCSC, 1992.

[10] Hebb, D. O.: *The Organization of Behavior*, Wiley, New York, 1949.

[11] Ikeda, K., Amari, S. and Yoshizawa, S.: Prediction Error and Consistent Parameter Area in Neural Learning, *Proc. Int'l Joint Conf. Neural Networks*, (1993), 1633–1636.

[12] Kirkpatrick, S.: Optimization by Simulated Annealing: Quantitative Studies, *J. Statist. Physics*, Vol. 34 (1984), 975–986.

[13] Levin, E., Tishby, N. and Solla, S. A.: A Statistical Approach to Learning and Generalization in Layered Neural Networks, *Proc. IEEE*, Vol. 78 (1990), 1568–1574.

[14] Murata, N., Yoshizawa, S. and Amari, S.: Learning Curves, Model Selection and Complexity of Neural Networks, in *Advances in Neural Information Processing Systems 5*, Morgan Kaufmann, 1993.

[15] Opper, M. and Haussler, D.: Calculation of the Learning Curve of Bayes Optimal Classification on Algorithm for Learning a Perceptron with Noise, *Proc. Ann. Workshop Comp. Learning Theory*, Vol. 4 (1991), 75–87.

[16] Rosenblatt, F.: *Principle of Neurodynamics*, Spartan, 1961.

[17] Rumelhart, D., McClelland, J. L. and the PDP Research Group, : *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, 1986.

[18] Seung, H. S., Sompolinsky, H. and Tishby, N.: Statistical Mechanics of Learning from Examples, *Physical Review A*, Vol. 45 (1992), 6056–6091.

[19] Valiant, L. G.: A Theory of the Learnable, *Comm. ACM*, Vol. 27 (1984), 1134–1142.

[20] Vapnik, V. N. and Chervonenkis, A. Y.: On the Uniform Convergence of Relative Frequencies of Events to their Probabilities, *Theory of Probability and its Applications*, Vol. 16 (1971), 264–280.

付録

例題のフィルタリング

学習機械に例題が与えられる場合，例題の与え方は大別して3通りある．一つめはこれまでのように，独立にランダムに選ばれた入力とそれに対する真の出力が例題として機械に与えられる場合である．二つめは，機械は与えられたランダムな入力列を順番に見て，真の出力を知りたいものを選び出す．この時，選ばれた入力とその入力に対する真の出力が例題として与えられ，残りの例題は捨てられる場合である．この入力の取捨選択をフィルタリングと呼ぶ．最後に，入力列そのものを機械が決め，その入力に対する真の出力が例題として与えられる場合である．これを探査と呼ぶ．ここでは，フィルタリングにおいて機械が入力列から選ぶ入力の数について議論する．

どのような入力は，真の出力を得る必要があるだろうか．確率的機械では出力が確率的に決まるため，入力が決まってもその入力に対する出力は一意には決まらない．よってすべての入出力の組がパラメータ推定に寄与することができる．しかし確定的機械では出力は入力とパラメータにより一意に定まるため，パラメータ推定に寄与しない入力が存在する．具体的にいえば，許容領域に交わらない入力に対しては，許容領域内のどのパラメータをもつ機械も同じ出力を出し，この出力は真の出力と一致している．すなわち，この入力に対する真の出力を学習機械は既に知っているので，真の機械から与えられる必要がない．逆に許容領域に交わる例題については，二分割されたパラメータの領域のどちら側に真のパラメータがあるのかがわからない．よって真の出力を例題として与えてもらう必要がある．

以上のことから，確定的機械のフィルタリングは，新規の入力が許容領域に交わる確率に深く関係していることがわかる．ここでは第3章で求めた新規の入力が許容領域に交わる確率を利用して，フィルタリングにおいて機械が入力列から選ぶ入力の数の期待値を求める．この問題は，先生の機械に正解を教えてもらう場合には

コストがかかり，そうでない場合にはコストがかからないとした時の，コストの期待値を求める問題である．

入出力機械は第3章と同じ単純パーセプトロンとする．第3章では，許容領域を成す例題を有効例題と定義した．与えられた t 個の例題について，有効例題になったことのある例題の数を累積有効例題数と呼ぶ．フィルタリングにおける入力数の問題は，累積有効例題数の期待値を求めることと同じである．

入力が k 個与えられた時の累積有効例題数の期待値を c_k とし，p_k を $k+1$ 番目の入力が k 個の例題の作る許容領域に交わる確率の期待値とすると

$$c_{t+1} = c_t + p_t \tag{.1}$$

即ち

$$c_t = \sum_{k=1}^{t-1} p_k \tag{.2}$$

が成り立つ．k 個の例題の作る凸包の頂点数の期待値を V_k とすると $p_k = V_k/k$ であるから，漸近的に

$$c_t = \sum_{k=1}^{t-1} \frac{V_k}{k} = V \log t + \mathrm{o}(\log t) \tag{.3}$$

が成り立つ．ここで $V = \lim_{k-\infty} V_k$ とした．

まとめると，長さ t の独立でランダムな入力列がある時，有効な (= 出力を教えてもらいたい) 入力は平均して $c_t = V \log t$ 個である．このことは，許容領域は最終的には $\mathrm{O}(1)$ 個の有効例題で決まるが，どの $\mathrm{O}(1)$ 個かを選ぶには $\mathrm{O}(\log t)$ 個について出力を調べなければならないことを意味している．

実験 .1 以上の結果を確認するため，計算機実験を行った．パラメータ次元 $m = 1,2,3,4,5$ について，100, 200, 300, 500, 1000, 2000, 3000, 5000, 10000 個の例題を与えた時の有効例題数を調べた．実験は各 m について 30 回ずつ行った．

結果 図.1, 図.2は，例題数と有効例題数の関係を次元 m がそれぞれ $m = 1, m = 3, m = 5$ 及び $m = 2, m = 4$ の場合について示したものである．エラーバーは30回の実験の標準偏差を表し，直線は $V \log t$ を適当に平行移動したものである．これらの図から，実験値は理論値とよく一致していることがわかる．

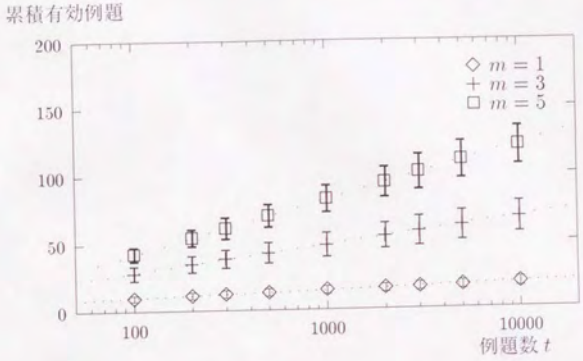


図 .1: 与えられた例題数と累積有効例題数 ($m = 1, m = 3, m = 5$)

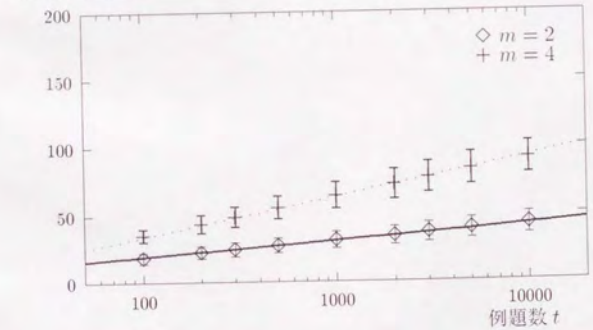
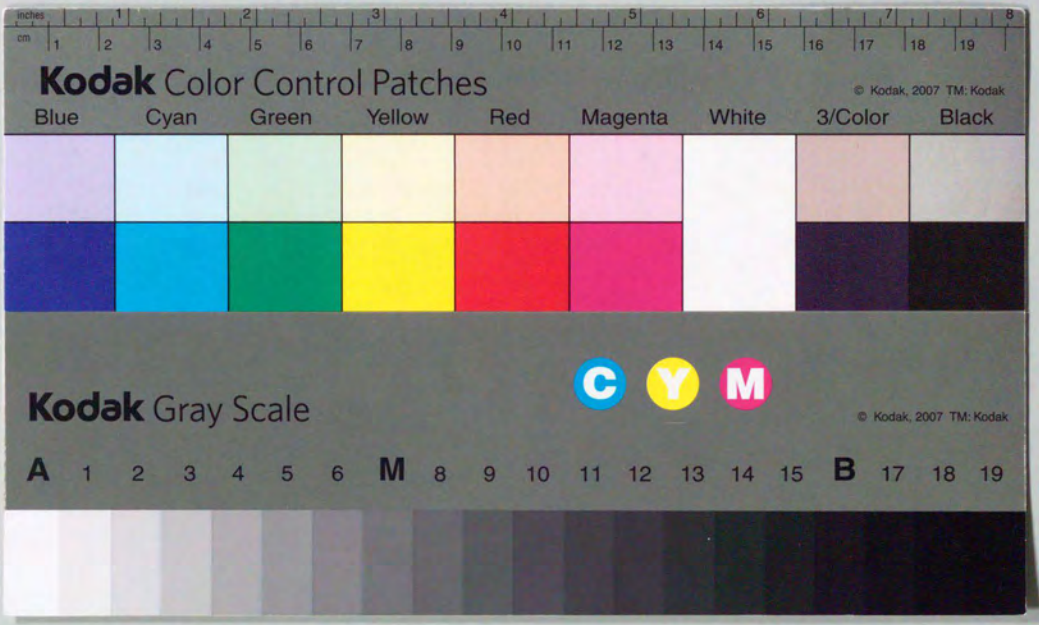


図 .2: 与えられた例題数と累積有効例題数 ($m = 2, m = 4$)


inches
cm
Kodak Color Control Patches
© Kodak, 2007 TM: Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
C
Y
M
Kodak Gray Scale
© Kodak, 2007 TM: Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19