学習の統計的漸近理論

村田昇

①

学習の統計的漸近理論

村田　昇

目次

第 1 章

序論

生体の神経系が行なう記憶や学習は，シナプス (synapse) の結合荷重が適応的に変化することによって起こると古くから考えられていた．この考えを明確な形で仮説として初めて提唱したのは心理学者 Hebb (1949) [18] であった．この仮説を構成的な研究として最初に具体的な形にしたのは心理学者 Rosenblatt である．彼は 1961 年にパーセプトロン (Perceptron) の学習アルゴリズム [47] を提案し，学習機械の研究に大きな影響を与えた．与えられた例題から信号空間の二分割を学習するパーセプトロンは，決定問題の基本的モデルであるが，Block (1962) [14] によって収束定理が証明されることにより，理論的背景が与えられた．パーセプトロン流の決定識別型の学習機械についてはいろいろな議論 [15, 40] があるが，Minsky and Papert (1971) [32] はその図形処理能力を計算幾何学に基づき議論している．一方，例題を与えながらパラメタを更新するタイプの学習を，パターン認識における識別関数と結び付けて確率論の観点から統一的に扱う理論もある．Amari (1967) [2] はこの立場にたち，学習速度と精度の間の関係などを論じている．また最近では Heskes and Kappen (1991) [19] がマルコフ過程を用いて学習のダイナミクスを解析している．学習機械の理論的研究は 1960 年代の初期の研究以降はしばらく沈滞していたが，1986 年に Rumelhart, Hinton and Williams [48, 49] によって発表された Connectionist Model によって再び活性化された．特に彼らの提案した Multi Layer Network の Back-Propagation 学習法は神経回路網の潜在能力の高さを一般に知らしめた．この回路網と学習法はすでに初期の研究により指摘されてはいたが，回路構造が比較的一様であるために計算機で実現が容易であり，なおかつ最近の計算機能力の向上により非常に大規模の回路網が構成できるようになったため，初めて実用に耐えるようになったといえる．彼らは，画像処理，音声認識，データ

圧縮，ロボットをはじめとする非線形システムの同定や制御などさまざまな実例を示すことにより，パーセプトロンでは限られていた応用範囲を広げ数々の分野への応用を促した．また中間層に形成される特徴抽出機構あるいは内部表現と，実際の生物が持っている特徴抽出機構との対応づけなど生物が行なっている情報処理の裏に潜むある種の構造の解明といった魅力的な問題をも提示した．彼らの研究に触発されて現在では様々な機械と学習法が提案され，精力的に研究されている．

実際の問題に対して適切な回路網を設計，あるいは選択しようとした場合，次の 2 つが重要な問題として浮かび上がってくる．ひとつは機械の大きさ，すなわち自由に更新できるパラメタの数をどのくらいにすれば良いのかという問題であり，もうひとつはどのくらいの例題を集めて学習を行なえば良いかという問題である．この 2 つの問題に関して従来は具体的な解決策は与えられておらず，設計者の主観，経験に頼ったり，試行錯誤を繰り返したりしていたが，ようやく最近になっていくつかの研究が報告されるようになった．

機械の大きさの問題は，たとえば Multi Layer Network であれば中間層の素子数を決めることに対応する．素子数を増やして回路を大きくすれば，回路網の表現力は豊かになり，与えられた例題に対しては正解を出しやすくなるが，学習に要する時間は長くなり，また例題以外の新しい入力に対しては非常に大きな間違いを犯してしまう可能性が高くなる．逆に素子数が不十分であれば例題すら満足に表現できないことになる．このように機械の大きさは機械の汎化能力，学習時間と密接に関わってくるため，適切な機械の大きさを何らかの方法で求めることは，学習機械の能力を十分に発揮させるためにも重要な問題である．この問題に関しては，多くの研究者により回路網の素子数を可変にするアルゴリズムが提案されている．しかしこうしたアプローチは一般に問題に依存する場合が多い．より一般的な枠組での評価としては統計の分野でモデル選択の規準として用いられる AIC や MDL 規準量を学習機械の選択に応用しようという試みが，栗田 (1990) [30]，和田および川人 (1991) [61]，Forgel (1991) [17] などにより提案されている．しかしこうしたアプローチも学習と規準量の本質的な関係を議論していない場合が多く不十分な点が多い．

例題数の問題に関しては，まず Baum and Haussler (1989) [13] が Valiant [60] に始まる PAC 学習の枠組に VC 次元を導入し，機械の複雑さを定量化することによって，学習に必要な例題数の条件を求めている．しかしながら彼らの与えた値は最悪評価であるために，一般にその評価値はかなり大きくなり，実用向きでないという批判も多

い．また，Levin, Tishby and Solla (1990) [31] は多数の均一な素子が相互に作用しながら動作する神経回路網の特徴に着目して，推定されるパラメータの分布に Gibbs 分布を仮定し，統計力学を用いた解析を行なっている．彼らは t 個の例題で学習した機械が新規の入力に対して犯す誤りを generalization error と定義し，ベイズの立場からその平均的な振舞いを論じている．また Rissanen (1986) [43] により提案された predictive minimum description length method との関係にも言及している．ただし実際問題として generalization error は計算が困難である場合が多いため，annealed approximation という近似計算に頼ることになるため，実用的な指針としてはまだ不十分である．このような統計力学的なアプローチは Seung, Sompolinsky and Tishby [54, 55, 56] などによっても進められている．

本研究は，機械の良さを計る評価関数として損失関数を導入して学習則を定式化することにより，機械の大きさと例題数の問題を統一的に考えるものである．機械の大きさ，構造および例題数が損失関数に与える影響を明らかにすることによって機械の設計・選択等に対してひとつの規準を与えることを目的とする．

本論文の構成は次のとおりである．

第 2 章では本研究で対象とするシステムと，それを近似するためのモデルの概念を定義する．またシステムとモデルの間の差を計る擬距離関数として損失関数および平均損失関数を定義し，これに基づいて例題からの学習という概念を明確にする．

第 3 章では損失関数を用いた学習の中で，提示される例題の順序に依存する逐次型学習を扱う．逐次型学習には多くの方法が考えられるが，ここでは損失関数を用いた最も基本的な方法として確率的降下法を用いる．まずシステムが時不変な場合について学習法の基本特性として速度と精度を調べ，その収束のための条件を明らかにする．学習の特性としては，例題による学習を終えた機械が，例題以外の新しい入力に対してどの程度の損失を持つかを考える．この損失を予測損失と言い，さらにこれを例題数の関数としてみたとき学習曲線と言う．本論文では確率論の手法を用いてこの学習曲線の漸近特性を明らかにする．またシステムが時変な場合については，学習法の追従特性を調べることができるが，これを用いて学習曲線の動特性についても考える．

第 4 章では損失関数を用いた学習の中で，提示される例題の順序によらない非逐次型学習をとりあげ，その学習特性を解析する．学習の特性としては，逐次型学習と同様に予測損失に基づく学習曲線を考える．また例題により学習した機械が，例題そのもの

に対しては持つ損失として学習損失を定義し，これを予測損失と同時に考え，2つの損失の漸近特性を明らかにする．またこの結果に基づいて最尤推定におけるモデル選択規準である AIC 規準量を拡張し，平均損失関数を用いた学習でのモデル選択規準を提案する．

第 5 章では与えられた例題だけでは最適なモデルのパラメタを一意に決めることができないような，特殊な場合の学習を考える．この種の問題の学習曲線の特性を求めることは一般に未解決ではあるが，問題を入力信号空間の二分割問題に限定すると，自然に導かれる損失に関しては緩い上限を求めることができる．これを用いると，信号空間の二分割問題における学習曲線の漸近特性に関して最悪評価を行なうことができる．

第 2 章

学習系と擬距離関数

2.1 システムとモデル

入力 x を受け，出力 y を生成する特定の計算機構を対象とし，あるモデルの中からその機構を正確に，あるいは近似的に表現するものを選択することを考える．まずシステムの概念を述べておく．入出力関係を実現する計算機構はその中に決まった規則を有し，与えられた入力からその規則にしたがって出力を算出する．また，入力は計算機構を取り囲む環境によって生成されるとする．以下では環境と計算機構をまとめてシステムと考え，次のように定義する．

定義 2.1 入力を $x \in \boldsymbol{R}^n$ ，出力を $y \in \boldsymbol{R}^{n'}$ で表すものとする．このとき次の 2 つの確率密度関数の組をシステム S

$$S = (\, p(x),\, p(y|x)\,) \tag{2.1}$$

と定義する．

確率密度 $p(x)$ は入力を支配する環境の性質を表し，条件つき確率密度 $p(y|x)$ は入出力関係を決定するシステムの本質的な部分を表す．すなわち入力 x は確率 $p(x)$ に従い発生され，出力 y は条件つき確率 $p(y|x)$ に従い生成される (図 2.1)．以下では確率密度 $p(x)$ を環境，条件つき確率密度 $p(y|x)$ を入出力関係と呼ぶこともある．また混乱のない限り x と y の同時確率

$$p(x, y) = p(y|x)\, p(x) \tag{2.2}$$

をシステム，あるいはシステムの確率と呼ぶこともある．

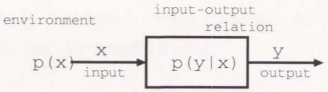


図 2.1: システムの概念.

システムは確率密度により表現されるが，デルタ関数 (delta function) を導入することにより入出力関係が特定の関数によって表現されるシステムも同様に扱うことができる．すなわち，入出力関係が

$$y = f(x) \tag{2.3}$$

で与えられるシステム S は

$$S = (\, p(x),\, \delta(y - f(x)) \,) \tag{2.4}$$

のように表すことができる．

一方，モデルとしては次の 2 つを考えることにする．

定義 2.2 入力 x に対して出力 y が一意に決定される機械を確定的機械 (deterministic machine) という．パラメタ $\theta \in \boldsymbol{R}^m$ を持つ確定的機械を関数

$$y = f(x;\theta) \tag{2.5}$$

で表現する．このときパラメタ θ を持つ関数の族を確定的な m 次元モデル

$$M_d = \{\, f(x;\theta) \,|\, \theta \in \boldsymbol{R}^m \,\} \tag{2.6}$$

と定義する．

定義 2.3 入力 x に対して出力 y が確率的に生成される機械を確率的機械 (stochastic machine) という．パラメタ $\theta \in \boldsymbol{R}^m$ を持つ確率的機械を条件つき確率密度

$$p(y|x;\theta) \tag{2.7}$$

で表現する．このときパラメタ θ を持つ条件つき確率の族を確率的な m 次元モデル

$$M_s = \{\, p(y|x;\theta) \,|\, \theta \in \boldsymbol{R}^m \,\} \tag{2.8}$$

と定義する．

確率的なモデルは入力 x に対して出力する y は確率変数となる．一方確定的なモデルは入力 x が与えられると特定の y，いわば出力の代表点となる値を計算し出力する．機械が確定的な場合でもデルタ関数を用いれば，

$$p(y|x;\theta) = \delta(y - f(x;\theta))$$

のように確率密度として表現できる．あるいは θ に依存しない加法的雑音 $n(x)$ を用い，$f(x;\theta)$ によって決定される確定値を確率変数 $f(x;\theta)+n(x)$ として，確率的な機械として扱うこともできる．このようにして確定的な機械を便宜的に確率的に動作するものとして扱えば，確定的なモデルも確率的なモデルもともにパラメタ θ で特徴づけられる．また，次節で定義する損失関数は機械が確定的であるか確率的であるかということに関わらず定義される．以下では必要のない限り 2 つのモデルを区別せずにモデル M としてあつかう．

定義 2.4 確定的なモデル M_d または確率的なモデル M_s のパラメタ θ の族を m 次元モデル

$$M = \{\, \theta \,|\, \theta \in \boldsymbol{R}^m \,\} \tag{2.9}$$

と定義する．

なお以下ではパラメタ θ の各要素は

$$\theta = (\theta^1, \theta^2, \cdots, \theta^m) \tag{2.10}$$

のように表す．パラメタ θ の各要素に関する偏微分は

$$\nabla = (\partial_1, \partial_2, \cdots, \partial_m) = (\frac{\partial}{\partial\theta^1}, \frac{\partial}{\partial\theta^2}, \cdots, \frac{\partial}{\partial\theta^m}) \tag{2.11}$$

のように略号で表す．またパラメタの要素および偏微分の添字に関しては予告なしに Einstein の記法を用いる．

次にモデルの具体的な例をあげておく．

例 1 (Multi Layer Network) 次の規則にしたがって入力 $x \in \boldsymbol{R}^{n_0}$ から出力 $y \in \boldsymbol{R}^{n_l}$ を計算する機械を考える．

$$\begin{aligned} z_i^k &= \tanh\left(\sum_{j=1}^{n_{k-1}} W_{ij}^k z_j^{k-1} + h_i^k\right), \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, l \\ x &= (z_1^0, \cdots, z_{n_0}^0) \\ y &= (z_1^l, \cdots, z_{n_l}^l) \end{aligned} \tag{2.12}$$

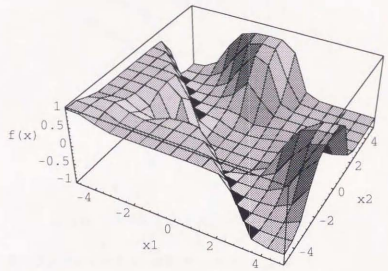


図 2.2: Multi Layer Network の例.

ただし,

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.13}$$

である. これは feed-forward 型の Multi Layer Network と呼ばれる, 最も頻繁に利用されるタイプの神経回路網である. 各 $z_i^k, k = 1, \ldots, l-1$ は中間素子 (hidden unit) と呼ばれることもある. また同一の k に属する素子をまとめて中間層 (hidden layer) と呼ぶこともある.

この神経回路網の入出力関係は明示的に書き下すこともできるが, 以下では

$$\theta = (W_{ij}^k, h_i^k), \quad i = 1, \ldots, n_k, j = 1, \ldots, n_{k-1}, k = 1, \ldots, l \tag{2.14}$$

とまとめて簡単に

$$y = f(x; \theta) \tag{2.15}$$

と書くこともある.

具体的な例として 2 入力 1 出力型の関数を図 2.2 に示す. 中間層は 1 層, 中間素子

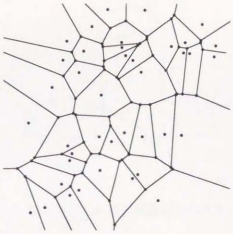


図 2.3: 競合系による 2 次元 Voronoi 分割の例.

は 4 素子，各パラメタは $l=2$, $n_0=2$, $n_1=4$, $n_2=1$ で

$$(W_{ij}^1) = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 2 & -1 \\ 1 & -2 \end{pmatrix} \quad (h_i^1) = \begin{pmatrix} -2 \\ 4 \\ 3 \\ -3 \end{pmatrix}$$

$$(W_{ij}^2) = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \quad (h_i^2) = \begin{pmatrix} 1 \end{pmatrix} \tag{2.16}$$

となっている.

例 2 (Kohonen Map) 入力を $x \in \boldsymbol{R}^n$, 出力を $y \in \boldsymbol{R}^{n'}$ とする. $v_i \in \boldsymbol{R}^{n'}$, $w_i \in \boldsymbol{R}^n$ なる k 個のベクトルの組

$$(v_i, w_i), \quad i = 1, \ldots, k \tag{2.17}$$

を考える. 機械は次の規則にしたがって出力を計算する.

$$y = v_j, \quad j = \arg\min(\|w_i - x\|) \tag{2.18}$$

これは競合系あるいは Kohonen Map と呼ばれる神経回路網による関数近似法である. この機械は入力信号空間を Voronoi 領域に分割し (図 2.3), 各領域ごとに出力値を決める (図 2.4) 一種の階段関数である.

以下では

$$\theta = (v_1, \cdots, v_k, w_1, \cdots, w_k) \tag{2.19}$$

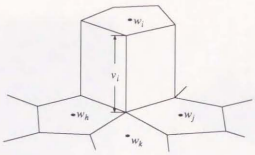


図 2.4: 競合系による階段関数近似の概念図.

として，機械の入出力関係を簡単に

$$y = g(x;\theta) \tag{2.20}$$

と書くこともある.

例 3 (Deterministic Dichotomy of R^n) 入力が n 次元，出力が 1 次元の Multi Layer Network $f(x;\theta)$ を考える．次式にしたがって入力信号空間 $\boldsymbol{R}^n$ を二分割する.

$$D_+ = \{x | f(x,\theta) \geq 0\}, \qquad D_- = \{x | f(x,\theta) < 0\}$$
$$D_+ \cap D_- = \emptyset, \qquad D_+ \cup D_- = \boldsymbol{R}^n \tag{2.21}$$

この分割にしたがって入力信号 x を二値 $\{-1,1\}$ に変換する機械 $h(x;\theta)$ を考える．機械の出力を y としたとき

$$y = h(x;\theta) = \begin{cases} 1, & x \in D_+ \\ -1, & x \in D_- \end{cases} \tag{2.22}$$

と記述できる．単純パーセプトロンでは信号空間の超平面による分割しか許さなかった．そのため，二分割される信号が空間内で線形分離可能でなければいけないという強い制約が課せられていた．Multi Layer Network を用いるとその制約はやや緩和され，空間内を曲面で分割することになる．図 2.5 は式 (2.16) の Multi Layer Network $f(x;\theta)$ を用いた 2 次元空間の二分割である．右図の出力に応じて入力空間を二分割している．分割は左図の濃淡によって示している.

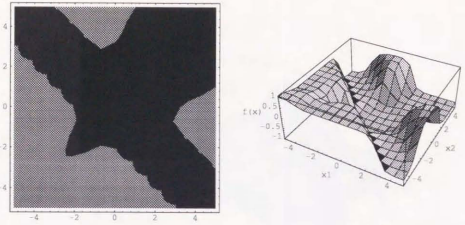


図 2.5: 2 次元空間の確定的な分割の例.

例 4 (Stochastic Dichotomy of R^n) 入力が n 次元，出力が 1 次元の Multi Layer Network $f(x;\theta)$ を考える．次式に示す確率にしたがって $\{-1,1\}$ を出力する機械 $p_f(y|x;\theta)$ を考える．

$$\begin{aligned} p_f(1|x;\theta) &= k(f(x;\theta)) \\ p_f(-1|x;\theta) &= 1-k(f(x;\theta)) \end{aligned} \tag{2.23}$$

ただし，関数 k は β を正の定数 (温度パラメタ) とし，

$$k(x)=\frac{1}{1+\exp(-\beta x)} \tag{2.24}$$

で定められるとする．この機械は入力信号空間を確率的に分割することになる．例 3 では確定的に空間を分割していたが，この機械は境界付近では ± 1 の出る確率がほぼ 5 分 5 分となり，境界領域がぼやけた分割を行なうことになる．なお，$\beta \to \infty$ においてその動作は例 3 の確定的な機械に近付いていく．

図 2.6, 2.7 は式 (2.16) の Multi Layer Network $f(x;\theta)$ を用いた 2 次元空間の確率的な分割である．図 2.6 は $\beta=0.5$，図 2.7 は $\beta=100$ とし，$y=1$ の確率密度を図示したものである．$\beta=100$ ではほとんど確定的に動作していることがわかる．

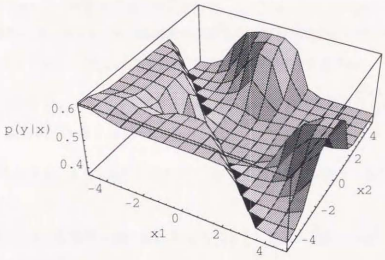


図 2.6: 2 次元空間の確率的な分割の例 ($\beta = 0.5$).

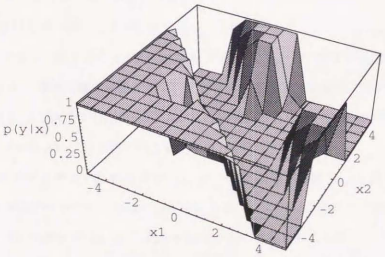


図 2.7: 2 次元空間の確率的な分割の例 ($\beta = 100$).

2.2 損失関数と平均損失関数

モデルの中からシステムを近似するために適当な機械を選択する際，その良否を判断する基準をあらかじめ決めておく必要がある．ここでは，システム $p(x,y)$ からモデルの 1 点 θ への隔たりを計る一種の擬距離関数として平均損失関数を導入する．

まず，システムの一つの入出力の組に対して決まる損失を定義する．

定義 2.5 入出力 (x,y) に対してモデル M 上の 1 点 θ が負う損失を $d(x,y;\theta)$ で表し，これを損失関数 (loss function) という．

損失関数には次のような条件を課しておく．これは以降の解析を行なうために必要な条件となる．

対象とするシステムの確率分布の族を $\mathcal{S}$ としたとき，任意の分布 $p(x,y) \in \mathcal{S}$ に対しての次の値が確定する．

$$(LF1) \qquad \int |d(x,y;\theta)|^2 p(x,y)dxdy < \infty$$

$$(LF2) \qquad \int |\partial_i d(x,y;\theta)\partial_j d(x,y;\theta)|p(x,y)dxdy < \infty$$

$$(LF3) \qquad \int |\partial_{i_1}\cdots\partial_{i_k} d(x,y;\theta)|p(x,y)dxdy < \infty, \quad k = 1,\ldots,5$$

注．条件 $(LF3)$ は必ずしも $k = 5$ までは必要としない．後述する意味での学習を定義するためには $k = 1$ までで十分であるが，第 4 章や第 3 章において学習の諸特性を解析するためには $k = 5$ まで必要となる．

損失関数はモデルの 1 点 θ が入出力 (x,y) を表現するのにどのくらい損失を持つかを表すものである．確定的なモデルでは機械 θ が入力 x から予測する出力 y' とシステムの出力 y との間の一種の誤差であり，モデルの予測に対する罰金にあたる．モデルの予測出力 y' とシステムの出力 y が近ければ罰金は小さく，差が開けば罰金を大きくするようにすればよい．具体的には 2 乗誤差などを考えればよい．

一方確率的なモデルでは機械 θ が (x,y) の発生は確率密度 $p(y|x,\theta)$ であると予測したその予測の確からしさをはかる規準になる．システムが実際に入出力 (x,y) を行なったとき，モデルが (x,y) の発生確率は小さく滅多に起こらないことだと予測していたら罰金を大きく，発生確率は大きく良く起こることと予測していたら罰金を小さくすればよい．例えば符合を考慮した対数尤度などが考えられる．いずれにしても損失関数は特定の入出力に関してモデルが行う予測に対する損失を表している．具体的な $d(x,y;\theta)$ の形状はシステム S の構造やモデル M の設計・評価方法に依存する．

次にシステム $p(x,y)$ とモデルの 1 点 θ との隔たりを評価する関数を定義する.

定義 2.6 平均損失関数 (averaged loss function) を

$$D(p,\theta)=\int d(x,y;\theta)\,p(y|x)\,p(x)dxdy \tag{2.25}$$

で定義する.

注. ここで p はシステムの分布 (入出力の同時分布) $p(x,y)$ を表している.

平均損失関数は損失関数をシステム $p(x,y)$ によって平均化したものである. 損失関数がシステムのただ一つの入出力の組に対して定義されていたのに対し, 平均損失関数はシステムのあらゆる入出力の組に対する重みつき積分で定義される. 損失関数はある入出力 (x,y) に対してモデルが行う予測に対する損失を表している. 出力 y は確率密度 $p(y|x)$ にしたがって生成されるので, この入力 x に対してモデル θ は平均的に

$$d_y(x;\theta)=\int d(x,y;\theta)p(y|x)dy \tag{2.26}$$

だけの損失をもつ. 同様に入力 x は確率密度 $p(x)$ にしたがって生成されるので, システムをモデル θ で近似したときには, 平均的に

$$D(p,\theta)=\int d_y(x;\theta)p(x)dx \tag{2.27}$$

だけの損失を持つわけである.

平均損失関数に基づいた評価規準に基づき, 最適な機械は次のように定義される.

定義 2.7 平均損失関数を最小化するパラメタ θ_{opt}

$$D(p,\theta_{opt})=\min_{\theta}\{D(p,\theta)\} \tag{2.28}$$

をシステム S の最適パラメタと呼ぶ.

確定的なモデルは, デルタ関数を用いるか, あるいは加法的雑音を加えるかすることによって確率的なモデルと見做せることはすでに述べた. モデルを確率密度関数の形に表したとき, モデルの忠実度 (fidelity) という概念を自然に導入できる.

定義 2.8 システムの入出力関係を $p(y|x)$, モデルを $\{p(y|x;\theta)\mid\theta\in\boldsymbol{R}^m\}$ と表したとする.

$$p(y|x)\in\{p(y|x;\theta)\} \tag{2.29}$$

が成り立つとき, モデルは忠実であるという.

これはシステム S の入出力関係をモデル M が完全に含んでいて，モデルが真に正しい入出力関係を表現できることを意味する．逆にモデルがシステムの入出力関係を含まず，近似的にしかシステムを表現できないとき，このモデルを非忠実なモデルという．また忠実なモデル M においては損失関数の条件を次のように書き直しておく．

$$(LF1') \qquad \int |d(x,y;\theta)|^2 p(y|x,\theta')p(x)dxdy < \infty$$

$$(LF2') \qquad \int |\partial_i d(x,y;\theta)\partial_j d(x,y;\theta)|p(y|x,\theta')p(x)dxdy < \infty$$

$$(LF3') \qquad \int |\partial_{i_1}\cdots\partial_{i_k} d(x,y;\theta)|p(y|x,\theta')p(x)dxdy < \infty, \quad k=1,\ldots,5$$

ただし $\theta,\theta' \in M$ であり，$p(x)$ は対象とするシステムの任意の環境である．

モデル M が忠実な場合，$\{\theta \in \boldsymbol{R}^m\}$ の中にシステムの入出力関係を記述するパラメタが存在する．このときには次のように一致性 (consistency) の概念を導入できる．

定義 2.9 モデル M を忠実とし，システムの入出力関係を記述するパラメタを $\theta_{sys} \in \boldsymbol{R}^m$ とする．任意の環境 $p(x)$ のもとでモデルの最適パラメタ θ_{opt} とシステムを記述するパラメタ θ_{sys} が一致するとき，損失関数は一致性を持つという．

一般に，モデルが忠実ではあるが一致性を持たない損失関数，またはモデルが非忠実な場合には最適パラメタはシステムの環境の変化に応じて変化してしまう．とくにモデルが非忠実な場合には，同じモデルを用いても損失関数のとり方によって最適パラメタは異なる．

以下に前節で述べた例の損失関数をあげておく．

例 1 (Multi Layer Network) 式 (2.12) で定義した Multi Layer Network を考える．このとき 2 乗誤差によって損失関数を定義する．

$$d(x,y;\theta) = \frac{1}{2}\|y - f(x;\theta)\|^2 \tag{2.30}$$

システムの入出力関係が確定的で

$$S = (p(x), \delta(y - f(x;\theta_{opt}))) \tag{2.31}$$

と書けるときには，明らかにモデルは忠実で，しかも上の損失関数は一致性を持つ．また，期待値が 0 で共分散が有界な加法的雑音 $\eta(x) \in \boldsymbol{R}^{n_l}$ があって，システムの入出力関係が

$$y = f(x;\theta_{opt}) + \eta(x) \tag{2.32}$$

と書ける．すなわち $\eta(x)$ の分布を $p(\eta|x)$ としたときシステムが，

$$S = (p(x), p(y - f(x;\theta_{opt})|x)) \tag{2.33}$$

で表されるならば，上の損失関数は忠実で一致性を持つ．

例 2 (Kohonen Map) 式 (2.18) で定義した Kohonen Map $g(x;\theta)$ の損失関数としては p 乗誤差を考えることにする．

$$d(x,y;\theta) = \frac{1}{p}\|y - g(x;\theta)\|^p \tag{2.34}$$

平均損失の最小化は，$p = 2$ なら通常の最小 2 乗規準に一致し，p が十分大きければ，近似的に損失の最大値を最小にすることになるのでミニマックス規準に近付いていく．ただしこの損失関数では 1 階微分が不連続となるため，損失関数の条件 $(LF3)$ が完全には満たされない．

例 3 (Deterministic Dichotomy of R^n) 信号空間を二分割する機械 $h(x;\theta)$ の損失関数を Multi Layer Network $f(x;\theta)$ を用いて次式で定義する．

$$d(x,y;\theta) = \begin{cases} \frac{1}{p}|f(x;\theta)|^p, & yf(x;\theta) < 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.35}$$

例題を生成するシステムがこの機械に含まれるとき，すなわちこのモデルが忠実ならば，この損失関数は明らかに一致性を持つ．しかし，有限個の例題に対してこの損失を 0 にするようなパラメタ θ は無数に存在する場合がある．この問題についての考察は第5章で行なう．

例 4 (Stochastic Dichotomy of R^n) 信号空間を分割する機械 $p_f(y|x,\theta)$ の損失関数を次式で定義する．

$$\begin{aligned} d(x,y;\theta) &= -\log p_f(y|x,\theta) \\ &= \delta_1(y)\log(1+\exp(-\beta f(x;\theta))) \\ &\quad +\delta_{-1}(y)\log(1+\exp(\beta f(x;\theta))) \end{aligned} \tag{2.36}$$

ただし

$$\delta_i(y) = \begin{cases} 1, & y = i \\ 0, & \text{otherwise} \end{cases} \tag{2.37}$$

である．この損失関数はシステムの入出力関係 $p(y|x)$ から見て Kullback-Leibler の divergence を最小にするようにパラメタを決定する規準になっている．Divergence の性質から，モデルが忠実ならば，この損失関数は一致性を持つ．

2.3 例題からの学習

あるシステムへの入出力が観測できる状況下で，システムの内部構造や性質が完全にわかっているならば，その知識を基にモデルを構成することは容易である．しかし通常システムに関して得られる知識はほんのわずかで，モデルを構成するために実際に利用できるものはシステムを観測して得られる入出力の組だけである場合が多い．例題の定義を明確にしておく．

定義 2.10 システムの入出力の観測値の組 $\xi = (x, y)$ を例題という．例題はシステムの確率 $p(x, y)$ に従い生成される．独立な t 個の例題を

$$\xi^t = \{(x_i, y_i); i = 1, \ldots, t\} \tag{2.38}$$

によって表す．

本研究で扱う学習を次のように定義する．

定義 2.11 例題の組 ξ^t から最適パラメタ θ_{opt} を探索することを学習と定義する．

例題のみを用いたモデルの決定を学習と定義するのである．もちろん有限個のデータしか使えなければ，平均損失関数を最小化する最適パラメタを完全に求めることはできない．我々は何らかの方法でその近似値を求めるに過ぎない．

学習は，例題 ξ の提示される順序によって推定されるパラメタが異なるか否かによって 2 つの場合に分けられる．パラメタの順序に依存する学習を逐次型，依存しない学習を非逐次型と呼ぶ．非逐次型学習の場合，観測の間例題は完全に同一の分布にしたがっていると考え，与えられた例題 ξ^t から経験分布

$$p_t(x, y) = \frac{1}{t} \sum_{i=1}^{t} \delta(x - x_i, y - y_i) \tag{2.39}$$

を構成し，これを疑似的にシステムと見做し，最適パラメタを推定することになる．具体的な手続きは第 4 章で述べるが，学習の間与えられた例題を全て記憶しておかなけれ

ばならない．また，例題を観測する間にシステムが変動していくような場合，有効なパラメタ推定が難しくなる．一方，逐次型学習は例題の系列に応じてできるだけ良いパラメタ推定を行なうように設計できるので，システムが変動する場合に有効である．また観測と学習を同時に行なえば，観測した例題の全てを記憶しておく必要がない学習系を作ることもできる．ただし，システムが変動しない場合には，例題を全て記憶しておく非逐次型学習の方がより高い精度の学習が行なえる．このような事情に関しては第 4章で議論する．

一般に広く使われている神経回路網の学習方式は本質的に非逐次型学習である場合が多い．はじめに例題を多数集めておいてパラメタを推定するか，もしくはその中から一つずつ例題を提示して，疑似的に逐次型学習を行ないパラメタを推定する．また多くの場合 off line で非逐次型学習を行い，実際に使うときには学習を行わない場合が多い．神経回路網の学習能力と 2 つの型の学習の特性を考えると，非逐次型学習で神経回路網を初期化しておき，逐次型学習を行いながら on line で利用するのがもっとも望ましいであろう．

2.4 学習曲線と学習関数

心理学，特に数理心理学と呼ばれるの分野では，学習の進行状況を知るために学習曲線というものを描く [22] ．学習曲線 (learning curve) とは，横軸に試行回数とか練習時間など学習に要した時間をとり，縦軸には誤答数とか，一定量の作業を行なうのに必要な所要時間などの学習の成績を評価する学習の測度 (measure of learning) をとって描いた曲線のことである．学習曲線を見ることにより，学習の進行状況，行動改善の程度，課題の難易度，学習者の能力などを直観的に知ることができる．また人間などの学習の進行過程を表した学習曲線にはある規則性が見られるため，これを数式化しようという試みが古くからなされてきた．学習曲線は，試行回数あるいは練習時間を t とし，学習成績を l とすれば，

$$l = e(t) \tag{2.40}$$

と表すことができるが，このような関数 e を学習関数 (learning function) とよんでいる．

本研究では心理学で利用されるこの学習曲線の概念を学習系の評価に持ち込み，これによってモデルの能力を評価する．我々は平均損失関数を定義し，それを最小化する

ことによって学習を定義したのであるから，縦軸にあたる学習の測度，すなわち成績の評価には平均損失関数を用いるのが自然であろう．横軸には例題数をとることにする．非逐次型学習における学習曲線については第 4章で，逐次型学習における学習曲線については第 3章で議論する．また第 5 章では特殊な場合として，有限個の例題では最適パラメタが一意に決められないような損失関数を用いた学習における学習曲線について議論する．

第 3 章

逐次型学習の特性

3.1 逐次型学習と確率的降下法

提示される例題の順序によって，推定されるパラメタが異なる学習方法を逐次型学習と定義した．時刻 t において t 番目の例題 ξ_t が与えられるような状況を考えたとき，逐次型学習としては現時刻から k 個前までの例題 $\xi_{t-k+1}, \ldots, \xi_t$ を用いて平均損失を最小化するようにパラメタを更新していく方法が考えられる．ここでは，その最も単純な場合として $k = 1$，すなわちその時刻において与えられた例題のみを用いてパラメタの更新を行なうことを考え，具体的な学習方法として以下に述べるような確率的降下法を導入する．

以下では損失関数 $d(x, y; \theta)$ がパラメタ θ に関して微分可能であることを仮定し，Amari [2] によって提案された確率的降下法の定義を述べる．この方法は後述する例によっても示すが，Multi Layer Netwrok に用いられる Back-Propagation [48, 49] や，Kohonen Map に用いられる Learning Vector Quantization [29] などの一般化になってる．以降でおこなう推定されるパラメタの期待値，分散等の学習の特性解析は Amari [2] の解析法に基づいて行なう．なお，分散に関する解析は Amari [2] の解析を精密化している．

モデルのパラメタ θ を，与えられた例題 (x, y) に基づいて逐次修正していく学習系を考える．例題 (x, y) はシステムの入出力の同時分布 $p(x, y)$ に従って独立に生成されるものとする．修正前のパラメタを θ，修正後のパラメタを θ'，修正項を $\delta\theta$ とする．

$$\theta' = \theta + \delta\theta(x, y, \theta) \tag{3.1}$$

修正項 $\delta\theta$ は提示された例題 (x, y) と修正前のパラメタ θ に依存する．修正項 $\delta\theta$ は $D(p, \theta)$ を減少させるように選びたいが，平均損失関数 $D(p, \theta)$ が未知であるため $D(p, \theta)$ を減少させる方向を完全に知ることはできない．そこで与えられる例題の分布 $p(x, y)$ に基づいた期待値の意味で $D(p, \theta)$ を減少させるように修正項 $\delta\theta$ を選ぶことを考える．

修正項 $\delta\theta$ の期待値は

$$E_\xi(\delta\theta) = \int \delta\theta(x, y, \theta)\, p(x, y) dx dy \tag{3.2}$$

と表すことができる．記号 E_ξ は $\xi = (x, y)$ の分布 $p(x, y)$ に関して期待値を計算することを表す．同様に V_ξ により分布 $p(x, y)$ に関して分散を計算することを表す．1 回の修正による平均損失関数 $D(p, \theta)$ の変化は

$$\begin{aligned} \delta D(p, \theta) &= D(p, \theta') - D(p, \theta) \\ &= \nabla D(p, \theta)^{\mathrm{T}} \delta\theta + O(\|\delta\theta\|^2) \end{aligned} \tag{3.3}$$

と書ける．$\delta\theta$ が微小量であるとし，2 次以上の微小量を無視すれば変化分 $\delta D(p, \theta)$ の期待値は

$$E_\xi(\delta D(p, \theta)) = \nabla D(p, \theta)^{\mathrm{T}} E_\xi(\delta\theta) \tag{3.4}$$

となる．パラメタの更新により平均損失関数を最小化するためにはこの値を負にしなければならないが，そのためには $\nabla D(p, \theta)$ と $E_\xi(\delta\theta)$ とが鈍角をなさなければならない．このためには適当な正の数 ε と適当な正定値行列 $C = (c^{ij})$ を用いて

$$E_\xi(\delta\theta) = -\varepsilon C \nabla D(p, \theta) \tag{3.5}$$

ならば十分である．

$$\begin{aligned} \nabla D(p, \theta) &= \nabla \int d(x, y; \theta) p(x, y) dx dy \\ &= E_\xi(\nabla d(x, y; \theta)) \end{aligned} \tag{3.6}$$

であるから，式 (3.5) を満たすためには，

$$\delta\theta(x, y, \theta) = -\varepsilon C \nabla d(x, y; \theta) \tag{3.7}$$

とすればよい．以上より確率的降下法を次のように定義する [2]．

定義 3.1 提示された t 個の例題 ξ^t から推定されたパラメタを θ_t で表す．新たに例題 (x_{t+1}, y_{t+1}) が与えられたとき，パラメタ θ_t の修正を次式によって定める．

$$\theta_{t+1} = \theta_t - \varepsilon_t C_t \nabla d(x_{t+1}, y_{t+1}; \theta_t) \tag{3.8}$$

ただし，ε_t は正の実数で学習係数という．また C_t は適当な正定値行列である．

一般に ε, C は時刻 t に依存するが，以下の議論ではそれぞれ定数，定数行列として扱う．すなわち

$$\varepsilon_t = \varepsilon, \quad C_t = C, \quad t = 1, 2, \ldots \tag{3.9}$$

とする．このとき $t \to \infty$ においてパラメタ θ_t は一つの値に収束しない．収束を厳密に保証するためには，時刻 t における学習係数 ε_t を条件

$$\begin{cases} \lim\limits_{t\to\infty} \varepsilon_t = 0 \\ \sum\limits_{t=0}^{\infty} \varepsilon_t = \infty \end{cases} \tag{3.10}$$

を満たすように選べば十分であることが，確率的近似法によって示される [46]．直観的に説明すると，十分な時間がたったとき学習係数が 0 に近付くので修正項は限りなく小さくなっていくが，修正の和は発散するため，パラメタ空間のあらゆる場所に到達することができるのである．ただし，学習係数をこのように選ぶと収束が非常に遅くなる．また，あとで議論するようにシステムが時変であるとき，これを追従することが困難となる．このため，学習係数に対する上のような制約条件は実用上あまり意味がない．

注. $d(x, y, \theta)$ は θ に関して少なくとも 1 階微分可能であれば確率的降下法は適用できる．ただし以降の学習特性の解析においては高階微分が存在することを仮定している．

以下に前章に例として載せたモデルの修正式を示す．

例 1 (Multi Layer Network) 損失関数の偏微分は $\tanh x$ の微分が $1 - \tanh^2 x$ であることを利用すると簡単に計算できる．まず，e_i^k を次のように定める．

$$e_i^l = -(y_i - z_i^l) \tag{3.11}$$

$$e_i^k = \sum_{p=1}^{n_{k+1}} e_p^{K+1} \{1 - (z_p^{k+1})^2\} W_{pi}^{k+1}, \quad k = 1, \ldots, l-1 \tag{3.12}$$

これを用いると偏微分は

$$\frac{\partial}{\partial W_{ij}^k} d(x,y;\theta) = \epsilon_i^k \{1-(z_i^k)^2\} z_j^{k-1} \tag{3.13}$$

$$\frac{\partial}{\partial h_i^k} d(x,y;\theta) = \epsilon_i^k \{1-(z_i^k)^2\} \tag{3.14}$$

で与えられる.

このは Back-Propagation と呼ばれる計算方法である [48, 49]. 各パラメタの更新量は Feed-forward 型の Multi Layer Network を逆向きに辿ることによって計算できるため, 非常に効率のよい並列演算を構成することができる.

損失関数 $d(x,y;\theta)$ の 2 階以上の微分も同様にして計算できるが, これらは 1 階微分ほど単純な形で表現することはできない.

例 2 (Kohonen Map) Kohonen Map を $g(x;\theta)$ とし, 損失関数として 2 乗誤差を考える. このとき, 損失関数の偏微分は次のように計算される.

入力 x が $\{w_i\}$ の形成する Voronoi 領域の内部にあり, $j = \arg\min\{\|w_i - x\|\}$ のとき,

$$\frac{\partial}{\partial v_i} d(x,y;\theta) = \begin{cases} (v_j - x), & i = j \\ 0, & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial w_i} d(x,y;\theta) = 0 \tag{3.15}$$

入力 x が $\{w_i\}$ の形成する Voronoi 領域の境界上にあり, $j,k = \arg\min\{\|w_i - x\|\}$ のとき,

$$\frac{\partial}{\partial v_i} d(x,y;\theta) = 0$$

$$\frac{\partial}{\partial w_i} d(x,y;\theta) = \begin{cases} \dfrac{\|v_j - x\| - \|v_k - x\|}{\|w_j - w_k\|}(x - w_j), & i = j \\ \dfrac{\|v_k - x\| - \|v_j - x\|}{\|w_k - w_j\|}(x - w_k), & i = k \\ 0, & \text{otherwise} \end{cases} \tag{3.16}$$

境界領域での偏微分に関する議論は甘利 (1968) [8] などにある.

これは Learning Vector Quantization 2 と呼ばれる学習法と同値である [29]. ただし通常の場合境界部分に入力信号がある確率は 0 なので実際には, 入力信号が境界を中心とする微小な帯状領域に入ったときに学習を行なうことになる. このため学習効率は一般によくない. これに関する考察は阪口・村田 (1990) [51] などにある.

例 3 (Deterministic Dichotomy of R^n) 機械 $h(x;\theta)$ の損失関数の偏微分はパラメタ θ の第 i 成分を θ^i としたとき次式のようになる．

$$\frac{\partial}{\partial\theta^i}d(x,y;\theta)=\begin{cases}-y|f(x,\theta)|^{p-1}\dfrac{\partial}{\partial\theta^i}f(x,\theta), & yf(x,\theta)<0\\ 0, & \text{otherwise}\end{cases} \tag{3.17}$$

$p=2$ ならば BP と同様に，Network を逆向きに辿るようにして偏微分が計算できる．なお，この損失関数には一般に p 階まで微分が存在し，(p-1) 階までは連続となる．

例 4 (Stochastic Dichotomy of R^n) 機械 $p_f(y|x,\theta)$ の損失関数の偏微分はパラメタ θ の第 i 成分を θ^i としたとき次式のようになる．

$$\begin{aligned}&\frac{\partial}{\partial\theta^i}d(x,y;\theta)\\ &=\left\{\frac{-\beta\delta_1(y)}{1+\exp(\beta f(x,\theta))}+\frac{\beta\delta_{-1}(y)}{1+\exp(-\beta f(x,\theta))}\right\}\frac{\partial}{\partial\theta^i}f(x,\theta)\end{aligned} \tag{3.18}$$

なお，$\beta\to\infty$ において機械の動作は例 3 の確定的な機械に近付いていくが，偏微分は 0 となってしまうので，この損失関数ではパラメタの推定ができなくなってしまう．

3.2 確率的降下法の基本的特性

この節では確率的降下法の基本的な性質を調べるために，ある初期パラメタから出発した学習系が時刻 t においてもつパラメタの期待値と分散を求める．この二つの量はそれぞれ学習の速度と精度を表す一つの指標となる．パラメタの期待値がどのように最適パラメタに近付いていくかを見ることによって学習の進行状況を調べ，パラメタの分散の変化を見ることによって学習されたパラメタが最適パラメタのまわりにどのように広がっているかを調べることになる．前節で述べたように ε は微小な値に，C は適当な正定値行列に固定し，そのときの学習特性を解析する．

パラメタの期待値や分散を求める前に，まず微分可能な θ の関数 $f(\theta)$ に関して一般に成り立つ補題を証明しておく．これは関数 $f(\theta)$ が学習の進行にともなってどのように変化していくかを表している．時刻 t に与えられる例題を $\xi_t=(x_t,y_t)$ と書く．

時刻 t に得られるパラメタ θ_t はパラメタの初期値 θ_0 と提示される例題の列

$$\begin{aligned}\xi^t &= \{\xi_1, \xi_2, \cdots, \xi_t\} \\ &= \{(x_1, y_1), (x_2, y_2), \cdots, (x_t, y_t)\} \end{aligned} \tag{3.19}$$

の関数

$$\theta_t = \theta_t(\theta_0, \xi_1, \cdots, \xi_t) \tag{3.20}$$

である．例題の組 ξ^t は各要素がシステムの同時分布 $p(x, y)$ に従って確率的に選ばれるので，初期値 θ_0 を固定したとき，θ_t は ξ^t に依存する確率変数となる．以下では θ_t の確率密度を $q_t(\theta_t|\theta_0)$ とし，この分布に関する平均を計算する操作を $E_{\theta_t|\theta_0}$ で表す．このとき $f(\theta_t)$ も ξ^t に依存する確率変数となる．時刻 t における関数 $f(\theta_t)$ の期待値を $\hat{f}_t$ で表すことにする．

$$\begin{aligned}\hat{f}_t &= \hat{f}_t(\theta_0) \\ &= E_{\theta_t|\theta_0}(f(\theta_t)) = \int f(\theta) q_t(\theta|\theta_0) d\theta \end{aligned} \tag{3.21}$$

また次の二つの量を

$$\begin{aligned} g_{ij}(\theta) &= E_\xi(\partial_i d(x, y; \theta) \partial_j d(x, y; \theta)) \\ q_{ij}(\theta) &= E_\xi(\partial_i \partial_j d(x, y; \theta)) \end{aligned} \tag{3.22}$$

のように定義し，これから二つの行列

$$\begin{aligned} G(\theta) &= E_\xi\left(\nabla d(x, y; \theta) \nabla d(x, y; \theta)^{\mathrm{T}}\right) \\ Q(\theta) &= E_\xi(\nabla\nabla d(x, y; \theta)) \,(= \nabla\nabla D(p, \theta)) \end{aligned} \tag{3.23}$$

を定義する．ただし，$\nabla\nabla$ は $\partial_i \partial_j$ を i 行 j 列成分とする行列型の演算子である．また以下では tr は行列のトレース (trace) を計算する演算子とする．以上の記号のもとで次の補題が成り立つ [2]．

補題 3.1 $\hat{f}_t$ は漸化式

$$\begin{aligned}\hat{f}_{t+1} &= \hat{f}_t - \varepsilon E_{\theta_t|\theta_0}\left((\nabla f(\theta_t))^{\mathrm{T}} C \nabla D(p, \theta_t)\right) \\ &\quad + \frac{\varepsilon^2}{2} \operatorname{tr} E_{\theta_t|\theta_0}\left(CG(\theta_t) C^{\mathrm{T}} \nabla\nabla f(\theta_t)\right) + O(\varepsilon^3) \end{aligned} \tag{3.24}$$

を満たす．

証明　時刻 t におけるパラメタを θ，時刻 $t+1$ におけるパラメタを θ'，時刻 $t+1$ に提示される例題を $\xi=(x,y)$ で表すことにする．例題 ξ の確率密度 $p(\xi)$ はシステムの入出力の同時分布 $p(x,y)$ に等しい．このとき

$$\theta'(\xi,\theta)=\theta+\delta\theta(\xi,\theta) \tag{3.25}$$

と表せる．まず θ を固定したときの θ' の確率密度，すなわち条件つき確率密度 $q(\theta'|\theta)$ を求める．パラメタが $\boldsymbol{R}^m$ の矩形領域

$$[\theta'^1,\theta'^1+d\theta'^1]\times\cdots\times[\theta'^m,\theta'^m+d\theta'^m] \tag{3.26}$$

にある確率は $q(\theta'|\theta)d\theta'$ であるが，θ' と ξ が積分変換の関係式

$$d\theta'=\frac{\partial\theta'(\xi,\theta)}{\partial\xi}d\xi \tag{3.27}$$

を満たすときには

$$q(\theta'|\theta)d\theta'=p(\xi(\theta,\theta'))d\xi(\theta,\theta') \tag{3.28}$$

と書ける．ここでは ξ の積分領域が θ と θ' によって決まるため，明示的に ξ を θ と θ' の関数として書いている．式 (3.28) の右辺は例題 $\xi(\theta,\theta')=(x,y)$ が領域

$$\begin{aligned}&[x^1,x^1+dx^1]\times\cdots\times[x^n,x^n+dx^n]\\&\times[y^1,y^1+dy^1]\times\cdots\times[y^{n'},y^{n'}+dy^{n'}]\end{aligned} \tag{3.29}$$

に入る確率を表している．パラメタ θ は分布 $q_t(\theta|\theta_0)$ に従う確率変数であるから，式 (3.28) の両辺を $q_t(\theta|\theta_0)$ に関して平均することにより時刻 $t+1$ におけるパラメタの確率密度が得られる．すなわち

$$\begin{aligned}q_{t+1}(\theta'|\theta_0)d\theta' &= \left(\int_{\theta\in\boldsymbol{R}^m}q(\theta'|\theta)q_t(\theta|\theta_0)d\theta\right)d\theta'\\&= \int_{\theta\in\boldsymbol{R}^m}p(\xi(\theta,\theta'))d\xi(\theta,\theta')q_t(\theta|\theta_0)d\theta\end{aligned} \tag{3.30}$$

である．$\xi(\theta,\theta')$ は θ を固定しても θ' が $\boldsymbol{R}^m$ の全域を動くとき $\boldsymbol{R}^{n+n'}$ の全域を動くから，$\boldsymbol{R}^m$ 全域の積分は $\boldsymbol{R}^{n+n'}$ 全域の積分に置き換えられ，

$$\begin{aligned}\hat{f}_{t+1} &= \int_{\theta'\in\boldsymbol{R}^m}f(\theta')q_{t+1}(\theta'|\theta_0)d\theta'\\&= \int_{\xi\in\boldsymbol{R}^{n+n'}}\int_{\theta\in\boldsymbol{R}^m}f(\theta'(\xi,\theta))p(\xi(\theta,\theta'))d\xi(\theta,\theta')q_t(\theta|\theta_0)d\theta\\&= \int\left(\int f(\theta'(\xi,\theta))p(\xi)d\xi\right)q_t(\theta|\theta_0)d\theta\\&= E_{\theta|\theta_0}\left(E_\xi\left(f(\theta'(x,y,\theta))\right)\right)\end{aligned} \tag{3.31}$$

が成り立つ．θ' が $\theta+\delta\theta$ であることを用い，さらに $\delta\theta$ が $O(\varepsilon)$ であることに注意すると，

$$\begin{aligned}
E_\xi\left(f(\theta')\right) &= E_\xi\left(f(\theta+\delta\theta)\right)\\
&= E_\xi\left(f(\theta)-\partial_i f(\theta)\delta\theta^i+\frac{1}{2}\partial_i\partial_j f(\theta)\delta\theta^i\delta\theta^j+O(\varepsilon^3)\right)\\
&= f(\theta)-\varepsilon\partial_i f(\theta)c^{ij}E_\xi\left(\partial_j d(x,y;\theta)\right)\\
&\quad+\frac{\varepsilon^2}{2}\partial_i\partial_j f(\theta)c^{ik}c^{jl}E_\xi\left(\partial_k d(x,y;\theta)\partial_l d(x,y;\theta)\right)+O(\varepsilon^3)\\
&= f(\theta)-\varepsilon\partial_i f(\theta)c^{ij}\partial_j D(p,\theta)+\frac{\varepsilon^2}{2}\partial_i\partial_j f(\theta)c^{ik}c^{jl}g_{kl}(\theta)\\
&\quad+O(\varepsilon^3)\\
&= f(\theta)-\varepsilon\nabla f(\theta)^{\mathrm{T}}C\nabla D(p,\theta)+\frac{\varepsilon^2}{2}\operatorname{tr}CG(\theta)C^{\mathrm{T}}\nabla\nabla f(\theta)\\
&\quad+O(\varepsilon^3)
\end{aligned}\tag{3.32}$$

が示される．この両辺を $q_t(\theta|\theta_0)$ で平均すれば補題が得られる． ∎

上の補題を用いると時刻 t におけるパラメタ θ の期待値と分散を求めることができる．定理の説明の前にいくつかの記号を定義する．時刻 t におけるパラメタ θ_t の期待値と共分散行列をそれぞれ $\hat{\theta}_t$, $\hat{V}_t$ で表す．

$$\hat{\theta}_t = E_{\theta_t|\theta_0}\left(\theta_t\right)\tag{3.33}$$

$$\hat{V}_t = E_{\theta_t|\theta_0}\left((\theta_t-\hat{\theta}_t)^2\right)\tag{3.34}$$

また，システム $p(x,y)$ に対する最適パラメタ θ_{opt} における行列 $G(\theta_{opt})$, $Q(\theta_{opt})$ を G, Q と書くことにする．

$$G = (g_{ij}) = (E_\xi\left(\partial_i d(x,y;\theta_{opt})\partial_j d(x,y;\theta_{opt})\right))\tag{3.35}$$

$$Q = (q_{ij}) = (E_\xi\left(\partial_i\partial_j d(x,y;\theta_{opt})\right))\tag{3.36}$$

このとき行列 G は定義より非負定値行列となる．また行列 Q は平均損失関数 $D(p,\theta)$ がパラメタ θ_{opt} で最小化されることから非不定値行列となるが，ここでは正定値行列となることを仮定しておく．さらに行列に作用する 2 つの線形演算子を次のように定義する．

定義 3.2 M, A を $m\times m$ の正方行列とする．線形演算子 Ξ_A, Φ_A を次のように定義する．

$$\Xi_A M = AM+(AM)^{\mathrm{T}}\tag{3.37}$$

$$\Phi_A M = AMA^{\mathrm{T}} \tag{3.38}$$

$A \in \boldsymbol{R}^{m\times m}$ を正則，a を任意の実数とするとき，この二つの演算子は次のような性質を持つ．

$$\Xi_A \Phi_A = \Phi_A \Xi_A \tag{3.39}$$

$$\Phi_{A^{-1}} \Xi_A = \Xi_A \Phi_{A^{-1}} \tag{3.40}$$

$$\Xi_{aA} = a\Xi_A \tag{3.41}$$

$$\Phi_{aA} = a^2 \Phi_A \tag{3.42}$$

$$(\Phi_A)^n = \Phi_{A^n} \tag{3.43}$$

特に二つの演算子が対称行列 $M \in \boldsymbol{R}^{m\times m}$ に作用する場合には

$$\Xi_{E-A} M = (2E - \Xi_A) M \tag{3.44}$$

$$\Phi_{E-A} M = (E - \Xi_A + \Phi_A) M \tag{3.45}$$

なる関係が成り立つ．またそれぞれの演算子の固有値に関しては以下の補題が成り立つ [25] .

補題 3.2 行列 $A \in \boldsymbol{R}^{m\times m}$ の固有値を $\lambda_i, i = 1, \ldots, m$ とする．このとき Ξ_A の固有値は

$$\lambda_i + \lambda_j \quad (i, j = 1, \ldots, m) \tag{3.46}$$

で表される．

証明　まず，行列の直積と列展開を定義する．$B, C \in \boldsymbol{R}^{m\times m}$ とし，行列 A の i 行 j 列成分を (a_{ij}) で表す．このとき，行列 A, B の直積 $A \otimes B$ および行列 A の列展開 $\mathrm{cs}\, A$ を次のように定義する．

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mm}B \end{pmatrix} \in \boldsymbol{R}^{mm\times mm} \tag{3.47}$$

$$\mathrm{cs}\, A = (a_{11}, a_{21}, \cdots, a_{m1}, a_{12}, a_{22}, \cdots, a_{mm})^{\mathrm{T}} \in \boldsymbol{R}^{mm} \tag{3.48}$$

次に ABC の列展開を考える．C の第 i 列を c_i とすると，

$$\mathrm{cs}\,(ABC) = (ABc_1, ABc_2, \cdots, ABc_m)^{\mathrm{T}} \tag{3.49}$$

である．ABc_i にだけ注目し，A の第 j 行を a_j，B の第 k 列を b_k として，

$$\begin{aligned} ABc_i &= \begin{pmatrix} a_1b_1 & \cdots & a_1b_m \\ \vdots & \ddots & \vdots \\ a_mb_1 & \cdots & a_mb_m \end{pmatrix} c_i = \begin{pmatrix} a_1c_{1i} & \cdots & a_1c_{mi} \\ \vdots & \ddots & \vdots \\ a_mc_{1i} & \cdots & a_mc_{mi} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \\ &= (c_i{}^{\mathrm{T}} \otimes A)\,\mathrm{cs}\,B \end{aligned} \tag{3.50}$$

ゆえに

$$\mathrm{cs}\,(ABC) = (C^{\mathrm{T}} \otimes A)\,\mathrm{cs}\,B \tag{3.51}$$

が成り立つ．

演算子 Ξ_A は $m \times m$ 行列から $m \times m$ 対称行列への線形演算子であるから，固有値を求めるためには，対称行列を $M \in \boldsymbol{R}^{m\times m}$ として，行列方程式

$$AM + (AM)^{\mathrm{T}} = \lambda M \tag{3.52}$$

を考えれば良い．

$$AME + EMA^{\mathrm{T}} = \lambda M \tag{3.53}$$

として列展開を考えると，式 (3.52) は

$$(E \otimes A + A \otimes E)\,\mathrm{cs}\,M = \lambda\,\mathrm{cs}\,M \tag{3.54}$$

と等価であることがわかる．行列 $(E\otimes A+A\otimes E) \in \boldsymbol{R}^{mm\times mm}$ の固有値を考えると，

$$\lambda_i + \lambda_j \quad (i, j = 1, \ldots, m) \tag{3.55}$$

であるから，これが演算子 Ξ_A の固有値となることがわかる．∎

補題 3.3 行列 $A \in \boldsymbol{R}^{m\times m}$ の固有値を $\lambda_i, i = 1, \ldots, m$ とする．このとき Φ_A の固有値は

$$\lambda_i\lambda_j \quad (i, j = 1, \ldots, m) \tag{3.56}$$

で表される．

証明　前補題の証明と同様に m 次正方行列を $M \in \boldsymbol{R}^{m\times m}$ として，行列方程式

$$AMA^{\mathrm{T}} = \lambda M \tag{3.57}$$

を考える．両辺を列展開すると，等価な式

$$(A \otimes A)\,\mathrm{cs}\,M = \lambda\,\mathrm{cs}\,M \tag{3.58}$$

が得られる．ここで行列 $(A \otimes A) \in \boldsymbol{R}^{mm \times mm}$ の固有値を考えると，

$$\lambda_i \lambda_j \quad (i, j = 1, \ldots, m) \tag{3.59}$$

が Φ_A の固有値となることがわかる．　∎

注．　補題から明らかなように Ξ_A は対称行列に作用し，なおかつ A の任意の 2 つの固有値の和が 0 とならないときには逆演算子が存在する．一般の正方行列は対称行列と交代行列（歪対称行列）の和に分けられるが，交代行列部分に関しては逆演算子は不定となる．また Φ_A は A が正則ならば逆演算子が存在する．

以上の記号と補題 3.1 を用いると θ_{opt} の近傍における θ_t の振舞いを論じることができる．はじめにパラメタ θ_t の期待値に関する定理を述べる [2]．

定理 3.1 時刻 t におけるパラメタ θ_t の期待値 $\hat{\theta}_t$ は

$$\hat{\theta}_t = \theta_{opt} + (E - \varepsilon CQ)^t(\theta_0 - \theta_{opt}) \tag{3.60}$$

となる．

証明　まず $D(p, \theta)$ が θ_{opt} において極値をとることに注意すると $\nabla D(p, \theta_{opt}) = 0$，また $\nabla\nabla D(p, \theta_{opt}) = Q$ であるから，

$$D(p, \theta) = D(p, \theta_{opt}) + \frac{1}{2}(\theta - \theta_{opt})^{\mathrm{T}} Q(\theta - \theta_{opt}) + O(\|\theta - \theta_{opt}\|^3) \tag{3.61}$$

と展開できる．θ が θ_{opt} の近傍で $\|\theta - \theta_{opt}\|$ が十分小さいとし，高次の微小量 $O(\|\theta - \theta_{opt}\|^3)$ を省略すると，$D(p, \theta)$ の gradient は

$$\nabla D(p, \theta) = Q(\theta - \theta_{opt}) \tag{3.62}$$

と書ける．また補題 3.1 において

$$f(\theta) = \theta \tag{3.63}$$

とすると，時刻 t における f の期待値は

$$\hat{f}_t = \hat{\theta}_t \tag{3.64}$$

となり，関数 $f(\theta)$ の微分は

$$\nabla f(\theta) = E \tag{3.65}$$

$$\nabla\nabla f(\theta) = 0 \tag{3.66}$$

となる．ε に関する高次の項を省略すると $\hat{\theta}_t$ の漸化式は

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \varepsilon CQ(\hat{\theta}_t - \theta_{opt}) \tag{3.67}$$

とかける．θ_0 を初期値としてこの漸化式を解くことにより定理が証明される．∎

同様にしてパラメタ θ_t の分散，に関しては次の定理が成り立つ．

定理 3.2 時刻 t におけるパラメタ θ_t の分散 $\hat{V}_t$ は

$$\begin{aligned}\hat{V}_t = {} & \{E-(E-\varepsilon\Xi_{CQ})^t\}\{\varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}}\} \\ & -\{(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t-(E-\varepsilon\Xi_{CQ})^t\}(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}}\end{aligned} \tag{3.68}$$

となる．

証明　まず補題 3.1 において

$$f(\theta) = \theta\theta^{\mathrm{T}} \tag{3.69}$$

とすると，時刻 t における f の期待値は

$$\hat{f}_t = E_{\theta_t|\theta_0}\left(\theta_t\theta_t{}^{\mathrm{T}}\right) \tag{3.70}$$

となる．ここで $f(\theta)$ の i 行 j 列成分を

$$f^{ij}(\theta) = \theta^i\theta^j \tag{3.71}$$

で表すことにすれば，

$$\begin{aligned}\partial_k f^{ij}(\theta) &= \partial_k(\theta^i\theta^j) \\ &= \delta^i_k\theta^j + \delta^j_k\theta^i\end{aligned} \tag{3.72}$$

$$\begin{aligned}\partial_k\partial_l f^{ij}(\theta) &= \partial_k\partial_l(\theta^i\theta^j) \\ &= \delta^i_l\delta^j_k + \delta^i_k\delta^j_l\end{aligned} \tag{3.73}$$

となる．ただし δ_j^i は Kronecker のデルタである．前定理の証明と同様に高次の微小量 $O(\|\theta-\theta_{opt}\|^3)$ を省略して

$$\nabla D(p,\theta) = Q(\theta-\theta_{opt}) \tag{3.74}$$

とし，$CQ(\theta-\theta_{opt})$ の第 i 成分を $(CQ(\theta-\theta_{opt}))^i$ と書くことにすれば，

$$\begin{aligned}
\nabla f(\theta)^{\mathrm{T}} C \nabla D(p,\theta) &= \left(\partial_k f^{ij}(\theta)(CQ(\theta-\theta_{opt}))^k\right) \\
&= \left((\delta_k^i\theta^j + \delta_k^j\theta^i)(CQ(\theta-\theta_{opt}))^k\right) \\
&= \left((CQ(\theta-\theta_{opt}))^i\theta^j + (CQ(\theta-\theta_{opt}))^j\theta^i\right) \\
&= CQ(\theta-\theta_{opt})\theta^{\mathrm{T}} + (CQ(\theta-\theta_{opt})\theta^{\mathrm{T}})^{\mathrm{T}} \\
&= \Xi_{CQ}(\theta-\theta_{opt})\theta^{\mathrm{T}}
\end{aligned} \tag{3.75}$$

となる．また $CG(\theta)C^{\mathrm{T}}$ の i 行 j 列成分を $(CG(\theta)C^{\mathrm{T}})^{ij}$ で表すことにすれば，

$$\begin{aligned}
CG(\theta)C^{\mathrm{T}}\nabla\nabla f(\theta) &= \left((CG(\theta)C^{\mathrm{T}})^{kl}\partial_k\partial_l f^{ij}\right) \\
&= \left((CG(\theta)C^{\mathrm{T}})^{kl}(\delta_l^i\delta_k^j + \delta_k^i\delta_l^j)\right) \\
&= \left((CG(\theta)C^{\mathrm{T}})^{ij} + (CG(\theta)C^{\mathrm{T}})^{ji}\right) \\
&= 2CG(\theta)C^{\mathrm{T}}
\end{aligned} \tag{3.76}$$

となる．分布 $q_t(\theta_t|\theta_0$ に関して平均を計算するとき，$G(\theta)$ を θ_{opt} のまわりで展開して高次の項を省略し，さらに上の二つ式を用いると，補題 3.1 より

$$\begin{aligned}
E_{\theta_{t+1}|\theta_0}\left(\theta_t\theta_t{}^{\mathrm{T}}\right) &= E_{\theta_t|\theta_0}\left(\theta_t\theta_t{}^{\mathrm{T}}\right) - \varepsilon\Xi_{CQ}E_{\theta_t|\theta_0}\left((\theta_t-\theta_{opt})\theta_t{}^{\mathrm{T}}\right) + \varepsilon^2 CGC^{\mathrm{T}} \\
&= (E-\varepsilon\Xi_{CQ})E_{\theta_t|\theta_0}\left(\theta_t\theta_t{}^{\mathrm{T}}\right) + \varepsilon\Xi_{CQ}\theta_{opt}\hat{\theta}_t{}^{\mathrm{T}} + \varepsilon^2 CGC^{\mathrm{T}}
\end{aligned} \tag{3.77}$$

が得られる．一方，定理 3.1 の証明より

$$\begin{aligned}
\hat{\theta}_{t+1}\hat{\theta}_{t+1}{}^{\mathrm{T}} &= \{(E-\varepsilon CQ)\hat{\theta}_t + \varepsilon CQ\theta_{opt}\}\{(E-\varepsilon CQ)\hat{\theta}_t + \varepsilon CQ\theta_{opt}\}^{\mathrm{T}} \\
&= (E-\varepsilon\Xi_{CQ})\hat{\theta}_t\hat{\theta}_t{}^{\mathrm{T}} + \varepsilon\Xi_{CQ}\theta_{opt}\hat{\theta}_t{}^{\mathrm{T}} + \varepsilon^2\Phi_{CQ}(\hat{\theta}_t-\theta_{opt})(\hat{\theta}_t-\theta_{opt})^{\mathrm{T}}
\end{aligned} \tag{3.78}$$

が計算される．よって式 (3.77), (3.78) から $\hat{V}_t$ に関する漸化式

$$\hat{V}_{t+1}$$

$$
\begin{aligned}
&= E_{\theta_{t+1}|\theta_0}\left(\theta_t\theta_t{}^{\mathrm{T}}\right) - \hat{\theta}_{t+1}\hat{\theta}_{t+1}{}^{\mathrm{T}} \\
&= (E-\varepsilon\Xi_{CQ})(E_{\theta_t|\theta_0}\left(\theta_t\theta_t{}^{\mathrm{T}}\right) - \hat{\theta}_t\hat{\theta}_t{}^{\mathrm{T}}) \\
&\quad +\varepsilon^2 CGC^{\mathrm{T}} - \varepsilon^2\Phi_{CQ}(\hat{\theta}_t-\theta_{opt})(\hat{\theta}_t-\theta_{opt})^{\mathrm{T}} \\
&= (E-\varepsilon\Xi_{CQ})\hat{V}_t + \varepsilon^2 CGC^{\mathrm{T}} - \varepsilon^2\Phi_{CQ}(\hat{\theta}_t-\theta_{opt})(\hat{\theta}_t-\theta_{opt})^{\mathrm{T}} \qquad (3.79)
\end{aligned}
$$

が得られる．ここで定理 3.1 と演算子 Ξ, Φ の性質を用いると，

$$
\begin{aligned}
&\Phi_{CQ}(\hat{\theta}_t-\theta_{opt})(\hat{\theta}_t-\theta_{opt})^{\mathrm{T}} \\
&= \Phi_{CQ}\Phi_{(E-\varepsilon CQ)^t}(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}} \\
&= \Phi_{CQ}(\Phi_{E-\varepsilon CQ})^t(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}} \\
&= \Phi_{CQ}(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}} \\
&= (E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t\Phi_{CQ}(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}} \qquad (3.80)
\end{aligned}
$$

であるから，

$$
v_t = \hat{V}_t - \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} \qquad (3.81)
$$

$$
u_t = \varepsilon^2(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t\Phi_{CQ}(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}} \qquad (3.82)
$$

とおくと，

$$
v_{t+1} = (E-\varepsilon\Xi_{CQ})v_t - u_t \qquad (3.83)
$$

$$
u_{t+1} = (E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})u_t \qquad (3.84)
$$

つまり

$$
\begin{pmatrix} v_{t+1} \\ u_{t+1} \end{pmatrix} = \begin{pmatrix} E-\varepsilon\Xi_{CQ} & -E \\ 0 & (E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ}) \end{pmatrix}\begin{pmatrix} v_t \\ u_t \end{pmatrix} \qquad (3.85)
$$

なる 2 項間漸化式が得られる．演算子の可換性に注意してこの漸化式を解くと

$$
\begin{pmatrix} v_t \\ u_t \end{pmatrix} = R^t\begin{pmatrix} v_0 \\ u_0 \end{pmatrix} \qquad (3.86)
$$

となる．ただし，R^t の i 行 j 列成分 r_{ij}^n は

$$
\begin{aligned}
r_{11}^t &= (E-\varepsilon\Xi_{CQ})^t \\
r_{12}^t &= -(\varepsilon^2\Phi_{CQ})^{-1}\{(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t-(E-\varepsilon\Xi_{CQ})^t\} \\
r_{21}^t &= 0 \\
r_{22}^t &= (E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t \qquad (3.87)
\end{aligned}
$$

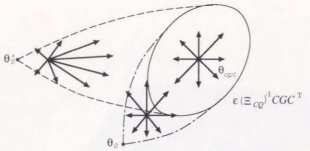


図 3.1: 逐次型学習の進行にともなう分散の変化.

で与えられ，初期値 v_0, u_0 は

$$v_0 = -\varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} \tag{3.88}$$

$$u_0 = \varepsilon^2\Phi_{CQ}(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}} \tag{3.89}$$

である．よって

$$v_t = \hat{V}_t-\varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} \tag{3.90}$$

$$= -(E-\varepsilon\Xi_{CQ})^t\varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} - \{(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t-(E-\varepsilon\Xi_{CQ})^t\}(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}} \tag{3.91}$$

となり，これを整理することによって定理が証明される．　∎

定理によって分散の広がり方は初期値を最適パラメタから遠い点にとった場合ほど遅くなることがわかる．最適パラメタから離れているときには，どのような例題が与えられてもほとんどの場合パラメタは最適パラメタの方向に移動していくのに対し，最適パラメタに近いときにはいろいろな例題に対して平均をとるとパラメタはほとんど動かない，つまり最適パラメタを中心に広がっていく方向にパラメタが移動するという学習の性質を表している (図 3.1).

例 5 (計算機シミュレーション 1.)　入力 $x=(x_1,x_2)$ が 2 次元，出力 y が 1 次元で

$$p(x) : [-1,1]\times[-1,1]\text{上で一様分布} \tag{3.92}$$

$$p(y|x) : y=\tanh(x_1+x_2)+\eta,\quad \eta\sim N(0,0.5^2) \tag{3.93}$$

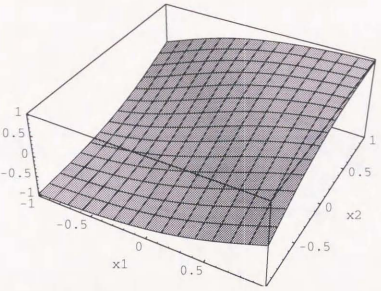


図 3.2: 最適パラメタにおける機械の入出力関係.

であるようなシステムを，確定的モデル

$$M_d = \{\tanh(\theta^1 x_1 + \theta^2 x_2) \mid \theta = (\theta^1, \theta^2) \in \boldsymbol{R}^2\} \tag{3.94}$$

で近似する問題を考える．損失関数として

$$d(x, y; \theta) = \frac{1}{2}(y - \tanh(\theta^1 x_1 + \theta^2 x_2))^2 \tag{3.95}$$

を用いて学習を行なうと，これはある入力 x に対するシステムの期待値を出力する機械が得られる．ここでは逐次型学習によるパラメタの変化を追った計算機シミュレーションの結果を示す．なお，学習に用いたパラメタは

$$C = E, \quad \varepsilon = 0.01$$

で，初期パラメタ，および最適パラメタは

$$\theta_0 = (0.58, 1.0), \quad \theta_{opt} = (1.0, 1.0)$$

である．

図 3.2 に最適パラメタでの機械の入出力関係を示す．

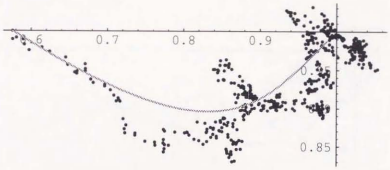


図 3.3: 逐次型学習により 1 台の機械のパラメタが変化する様子.

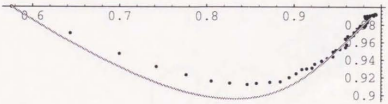


図 3.4: 逐次型学習により 100 台の機械のパラメタの平均値が変化する様子.

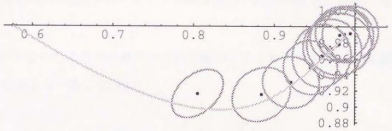


図 3.5: 逐次型学習による機械のパラメタ変化の平均値と分散.

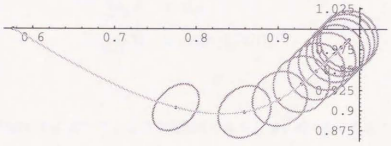


図 3.6: 逐次型学習による機械のパラメタの理論的期待値と分散.

図 3.3 は 1 台の機械のパラメタ変化を例題 10 個毎に表示したものである．実線は理論的に導かれたパラメタの期待値の軌跡を表している．

図 3.4 は 100 台の機械のパラメタの平均値を例題 100 個毎に表示したものである．実線は理論値である．

図 3.5 は 100 台の機械のパラメタの平均値と分散を例題 500 個毎に表示したものである．

図 3.6 は理論的に導かれるパラメタの期待値と分散を例題 500 個毎に表示したものである．

例題数が少ないところでは高次の項の影響のためずれが大きいが，例題数が十分多いところではかなり一致している．

3.3　確率的降下法の収束

確率的降下法による学習は少なくとも $\varepsilon \to 0$ としない限り推定されるパラメタが一点に収束することはない．

しかし，前節の 2 つの定理から直ちに導かれる次の系から，推定されるパラメタの分布についてはその平均と分散が収束することがわかる．

系 3.3 ε が十分小さければ，確率的降下法により推定されるパラメタの期待値 $\hat{\theta}_t$ と分散 $\hat{V}_t$ はそれぞれ

$$\lim_{t\to\infty} \hat{\theta}_t = \theta_{opt} \tag{3.96}$$

$$\lim_{t\to\infty} \hat{V}_t = \varepsilon(\Xi_{CQ})^{-1} CGC^{\mathrm{T}} \tag{3.97}$$

に近づく．

証明　まず行列 $A \in \boldsymbol{R}^{m\times m}$，および線形作用素 Ξ_A, Φ_A のノルムを次のように定義する．

$$\|A\| = \sup_{|x|\le 1} |Ax| \tag{3.98}$$

$$\|\Xi_A\| = \sup_{\|X\|\le 1} \|\Xi_A X\| \tag{3.99}$$

$$\|\Phi_A\| = \sup_{\|X\|\le 1} \|\Phi_A X\| \tag{3.100}$$

ただし $x \in \boldsymbol{R}^m$, $X \in \boldsymbol{R}^{m\times m}$ である．

定義より行列 C, Q の固有値はすべて正となる．したがって行列 CQ の固有値も正となるが，これを小さい順に並べて $\lambda_i,\ i = 1, \ldots, m$ とする．補題 3.2, 3.3 に注意すると

$$\|E - \varepsilon CQ\| = |1 - \varepsilon\lambda_1| \tag{3.101}$$

$$\|E - \varepsilon\Xi_{CQ}\| = |1 - 2\varepsilon\lambda_1| \tag{3.102}$$

$$\|E - \varepsilon\Xi_{CQ} + \varepsilon^2\Phi_{CQ}\| = \|\Phi_{E-\varepsilon CQ}\| = |1 - \varepsilon\lambda_1|^2 \tag{3.103}$$

となり，ε が微小ならいずれのノルムも 1 より小さくなる．よって

$$\lim_{t\to\infty}(E - \varepsilon CQ)^t = 0 \tag{3.104}$$

$$\lim_{t\to\infty}(E - \varepsilon\Xi_{CQ})^t = 0 \tag{3.105}$$

$$\lim_{t\to\infty}(E - \varepsilon\Xi_{CQ} + \varepsilon^2\Phi_{CQ})^t = 0 \tag{3.106}$$

が成り立つ．以上より系が証明される．　∎

この系は，適当な初期値から十分な時間学習を行なうと，パラメタは初期値によらず最適パラメタを中心に $O(\varepsilon)$ の分散を持ってばらつくことを意味している．θ_t は例題を与えられる限り修正され続けるので，最適パラメタのまわりでゆらいでおり，特定の確率変数に収束することはない．この系によれば，パラメタ θ_t はある弱定常系列に収束していくことがわかる．次の定理は特性関数を用いてさらに強い意味でパラメタ θ_t の法則収束を保証するものである．

定理 3.4 $t \to \infty, \varepsilon \to 0$ においてパラメタ θ_t の分布は正規分布

$$N(\theta_{opt}, \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}}) \tag{3.107}$$

に近づく．

証明　以下では簡単のためパラメタ θ は最適パラメタ θ_{opt} を原点とするような座標系で考えることする．時刻 t におけるパラメタ θ_t の r 次モーメントの $i_1, i_2, \ldots, i_r$ 成分を $(\theta^{i_1}\theta^{i_2}\cdots\theta^{i_r})_t$ と書くことにすると，補題 3.1 より

$$\begin{aligned}(\theta^{i_1}\cdots\theta^{i_r})_{t+1} \;=\; & (\theta^{i_1}\cdots\theta^{i_r})_t - \varepsilon\sum_{s=1}^{r} c^{i_s k}q_{kj}(\theta^{i_1}\cdots\theta^{j_s-1}\theta^{j}\theta^{i_s+1}\cdots\theta^{i_r})_t \\ & + \frac{\varepsilon^2}{2}\sum_{s,s'=1}^{r} c^{i_s j}c^{i_{s'}k}g_{jk}(\theta^{i_1}\cdots\theta^{j_s-1}\theta^{i_s+1}\cdots\theta^{i_{s'}-1}\theta^{i_{s'}+1}\cdots\theta^{i_r})_t \\ & + O(\varepsilon^3)\end{aligned} \tag{3.108}$$

となり，r 次モーメントは $(r-2)$ 次モーメントを用いた漸化式で表すことができる．ただし $\partial_j D(p,\theta)$, $g_{jk}(\theta)$ は原点のまわりで展開し，高次の項を省略して $q_{jk}\theta^k$, g_{jk} とした．$\sum_{s=1}^{r} c^{i_s k} q_{kj}$ を演算子としてみたとき，その固有値は行列 CQ の固有値 $\lambda_i, i = 1,\ldots,m$ を用いて

$$\lambda_{i_1} + \lambda_{i_2} + \cdots + \lambda_{i_r}, \quad (i_1, i_2, \cdots, i_r = 1, \ldots, m) \tag{3.109}$$

となる．これは $\{(\theta^{i_1}\cdots\theta^{i_{s-1}}\theta^j\theta^{i_{s+1}}\cdots\theta^{i_r})_t; j = 1,\ldots,m\}$ に対する固有値を全ての $s = 1,\ldots,r$ について考えることにより示される．よって ε を十分小さく選ぶことにより $E - \varepsilon\sum_{s=1}^{r} c^{i_s k} q_{kj}$ の固有値は 1 より小さくなる．$r-2$ 次のモーメントについて有界性と収束を仮定することにより，帰納的に各モーメントの収束を示すことができる．

さて $z = (z_i) \in \boldsymbol{R}^m$ として

$$f(\theta) = e^{iz_i\theta^i} \tag{3.110}$$

とおく．時刻 t における関数 f の期待値は分布 $q_t(\theta_t|\theta_0)$ の特性関数となるが，これを $\varphi_t(z)$ で表す．

$$E_{\theta_t|\theta_0}\left(f(\theta_t)\right) = \varphi_t(z) \tag{3.111}$$

またパラメタ θ_t の各モーメントが $t \to \infty$ において収束することから，$\varphi_t(z)$ の極限が存在するのでこれを $\varphi(z)$ で表す．

$$\lim_{t\to\infty} E_{\theta_t|\theta_0}\left(f(\theta_t)\right) = \varphi(z) \tag{3.112}$$

補題 3.1 より

$$\begin{aligned}
\varphi_{t+1}(z) &= \varphi_t(z) - \varepsilon E_{\theta_t|\theta_0}\left(\partial_j f(\theta_t) c^{jk} q_{kl}\theta_t{}^l\right) \\
&\quad + \frac{\varepsilon^2}{2} E_{\theta_t|\theta_0}\left(\partial_j\partial_k f(\theta_t) c^{jr}c^{ks} g_{rs}\right) + O(\varepsilon^3) \\
&= \varphi_t(z) - \varepsilon E_{\theta_t|\theta_0}\left(iz_j e^{iz_i\theta_t{}^i} c^{jk} q_{kl}\theta_t{}^l\right) \\
&\quad - \frac{\varepsilon^2}{2} E_{\theta_t|\theta_0}\left(z_j z_k e^{iz_i\theta_t{}^i} c^{jr}c^{ks} g_{rs}\right) + O(\varepsilon^3) \\
&= \varphi_t(z) - \varepsilon z_j c^{jk} q_{kl} E_{\theta_t|\theta_0}\left(\frac{\partial}{\partial z_l} e^{iz_i\theta_t{}^i}\right) \\
&\quad - \frac{\varepsilon^2}{2} z_j z_k c^{jr}c^{ks} g_{rs} E_{\theta_t|\theta_0}\left(e^{iz_i\theta_t{}^i}\right) + O(\varepsilon^3) \\
&= \varphi_t(z) - \varepsilon z_j c^{jk} q_{kl} \frac{\partial}{\partial z_l}\varphi_t(z) \\
&\quad - \frac{\varepsilon^2}{2} z_j z_k c^{jr}c^{ks} g_{rs}\varphi_t(z) + O(\varepsilon^3)
\end{aligned} \tag{3.113}$$

である．$t \to \infty$ の極限をとると

$$z_j c^{jk} q_{kl} \frac{\partial}{\partial z_l} \varphi(z) = -\frac{\varepsilon}{2} z_j z_k c^{jr} c^{ks} g_{rs} \varphi(z) + O(\varepsilon^2) \tag{3.114}$$

となる．

$$\varphi(z) = e^{h_0(z) + \varepsilon h_1(z) + O(\varepsilon^2)} \tag{3.115}$$

として $h_0(z)$, $h_1(z)$ の満たすべき条件式を求めると

$$z_j c^{jk} q_{kl} \frac{\partial}{\partial z_l} h_0(z) = 0 \tag{3.116}$$

$$z_j c^{jk} q_{kl} \frac{\partial}{\partial z_l} h_1(z) = -\frac{\varepsilon}{2} z_j z_k c^{jr} c^{ks} g_{rs} \tag{3.117}$$

となる．これが任意の z で成り立つためには $h_0(z)$ は明らかに 0．また $h_1(z)$ は z に関して 2 次式であることがわかるから

$$h_1(z) = -\frac{1}{2} z_j z_k v^{jk} \tag{3.118}$$

とおくと，

$$z_j z_k c^{jr} q_{rs} v^{sk} = \frac{\varepsilon}{2} z_j z_k c^{jr} c^{ks} g_{rs} \tag{3.119}$$

が成り立たなければならない．右辺が対称であることに注意すると

$$c^{jr} q_{rs} v^{sk} + c^{kr} q_{rs} v^{sj} = \varepsilon c^{jr} c^{ks} g_{rs} \tag{3.120}$$

すなわち $V = (v^{jk})$ としたとき

$$CQV + (CQV)^{\mathrm{T}} = \varepsilon C G C^{\mathrm{T}} \tag{3.121}$$

が成り立てばよい．これを解くと

$$V = \varepsilon (\Xi_{CQ})^{-1} C G C^{\mathrm{T}} \tag{3.122}$$

となるから，分布 $q_t(\theta_t|\theta_0)$ の特性関数は分散が V で与えられる正規分布の特性関数に漸近することがわかる．以上の議論は最適パラメタを原点としていたことを考慮すると，分布 $q_t(\theta_t|\theta_0)$ は正規分布

$$N(\theta_{opt}, \varepsilon (\Xi_{CQ})^{-1} C G C^{\mathrm{T}}) \tag{3.123}$$

に近づく．∎

次に述べる定理は上の定理とは別の意味で確率的降下法の収束を保証している*．

*甘利 (1968) [8] には分散を用いないで学習の定義から直接これを示す厳密な証明がある．

定理 3.5 t が十分大きいときには，任意の正の数 δ に対して学習係数 ε を十分小さく選ぶことにより，パラメタ θ_t が最適パラメタ θ_{opt} から δ 以上離れている確率をいくらでも小さくできる．すなわち

$$\lim_{\varepsilon \to 0} \lim_{t \to \infty} \mathrm{Prob}\{\|\theta_t - \theta_{opt}\| \geq \delta\} = 0 \tag{3.124}$$

証明　Čebyšev の不等式より，$a > 0$ に対し，

$$\mathrm{Prob}\{\|\theta_t - \theta_{opt}\| > a(\,\mathrm{tr}\,\hat{V}_t)^{\frac{1}{2}}\} \leq \frac{1}{a^2} \tag{3.125}$$

が成り立つ．前系より ε が十分小さいとき，十分大きな T に対して正の定数 c が存在して，$t > T$ ならば

$$\mathrm{tr}\,\hat{V}_t < \varepsilon\,\mathrm{tr}\,(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} + c\varepsilon \tag{3.126}$$

とすることができる．任意の δ に対して

$$a = \frac{\delta}{\sqrt{\varepsilon\,\mathrm{tr}\,(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} + c\varepsilon}} \tag{3.127}$$

と選べば，$\delta > a(\,\mathrm{tr}\,\hat{V}_t)^{\frac{1}{2}}$ であるから，

$$\mathrm{Prob}\{\|\theta_t - \theta_{opt}\| \geq \delta\} < \mathrm{Prob}\{\|\theta_t - \theta_{opt}\| > a(\,\mathrm{tr}\,\hat{V}_t)^{\frac{1}{2}}\} \leq \frac{1}{a^2} \tag{3.128}$$

が成り立つ．よって

$$\lim_{t \to \infty} \mathrm{Prob}\{\|\theta_t - \theta_{opt}\| \geq \delta\} \leq \frac{\varepsilon\,\mathrm{tr}\,(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} + c\varepsilon}{\delta^2} \tag{3.129}$$

ここで $\varepsilon \to 0$ とすれば定理が得られる．　∎

3.4　学習時間と精度

前節で求めた学習の速度と精度に関する定理から，実用上必要とされる学習時間を概算することができる．系 3.3 より，学習の結果最終的に残るパラメタの分散は $O(\varepsilon)$ であるから，パラメタの期待値と最適パラメタの偏差がこれより高いオーダーになればよい．十分大きな t に対して ε が微小ならば，行列 CQ の最小固有値 λ_1 を用いて

$$\|(E - \varepsilon CQ)^t\| = O(\exp(-\varepsilon\lambda_1 t)) \tag{3.130}$$

$$\|(E - \varepsilon\Xi_{CQ})^t\| = O(\exp(-2\varepsilon\lambda_1 t)) \tag{3.131}$$

$$\|(E - \varepsilon\Xi_{CQ} + \varepsilon^2\Phi_{CQ})^t\| = O(\exp(-2\varepsilon\lambda_1 t)) \tag{3.132}$$

と評価することができる．このことから，

$$t = O\left(\sim \frac{1}{\lambda_1}\left(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}\right)\right) \tag{3.133}$$

とすれば，定理 3.1 を用いてパラメタの期待値は

$$\hat{\theta}_t = \theta_{opt} + O(\varepsilon) \tag{3.134}$$

となり，定理 3.2 を用いてパラメタの分散は

$$\hat{V}_t = \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} + O(\varepsilon^2) \tag{3.135}$$

と計算される．このとき最適パラメタからの偏差は

$$\|\hat{\theta}_t - \theta_{opt}\|^2 = O(\varepsilon^2) \tag{3.136}$$

となり，パラメタの分散より小さくすることができる．すなわちパラメタの期待値の最適パラメタからのずれを学習に付随する揺らぎの中に閉じ込めることができる．

以上の計算は平均評価に基づく概算であるが，実用上は式 (3.133) の右辺のオーダー内の数倍程度の学習時間をとれば十分であろう．特に神経回路網のなどのパラメタの次数が高いモデルにおいては，高次元分布の特性から推定されたパラメタはほとんど共分散行列で表される超楕円球面の近くに局在するので，このような評価はかなり有効であると思われる．また，最終的に得られる機械の精度 $O(\varepsilon)$ と必要とされる学習時間 $O(\log\varepsilon/\varepsilon)$ の間の関係は，モデルの設計における一つの指標となると考えられる．

3.5 学習曲線とその性質

学習曲線を定義するために，まず予測損失を次のように定義する．

定義 3.3 パラメタ θ_t を持つ機械が，新しい例題 (x, y) に対して持つ平均損失を予測損失 (predictive loss) $L_P(\theta_t)$ という．

$$\begin{aligned} L_P(\theta_t) &= \int d(x, y; \theta_t)p(y|x)p(x)dxdy \\ &= D(p, \theta_t) \end{aligned} \tag{3.137}$$

注．上で定義した予測損失と同時に，学習に用いられた特定の例題に対して持つ損失を考えることができるが，これは逐次型学習の場合にはあまり意味がない．なぜなら，

パラメタの推定値 θ_t は与えられる例題の順序によるので，例題の組 $\{x_i, y_i\}$ と θ_t の間には非常に複雑な関係があり，例題 (x_1, y_1) が θ_t に及ぼした影響と例題 (x_t, y_t) が θ_t に及ぼした影響は同一視できないからである．このため全ての例題に対する損失を同一の重み付けで加算平均をとって評価することは妥当ではない．また適当な重み付けを考えて平均を考えることもできるが，これは計算を複雑にするだけであるし，あまり意味のないことであろう．

次に予測誤差を用いて学習曲線を定義する．

定義 3.4 予測損失を ξ^t およびパラメタの初期値の分布 $q_0(\theta_0)$ に関して平均したものを平均予測損失という．平均予測損失を t の関数としてみたとき，これを学習曲線と呼ぶ．

確率的降下法の基本的な特性から学習曲線を具体的に求めることができる．

定理 3.6 予測損失の期待値は漸近的に

$$E_{\xi^t,\theta_0}\left(L_P(\theta_t)\right) = D(p,\theta_{opt}) + \frac{1}{2}\varepsilon\ \mathrm{tr}\,QV + \frac{1}{2}\,\mathrm{tr}\,(E-\varepsilon\Xi_{CQ})^t(V_0Q-\varepsilon QV) \quad (3.138)$$

で与えられる．ただし，

$$V = (\Xi_{CQ})^{-1}CGC^{\mathrm{T}} \quad (3.139)$$

$$V_0 = \int(\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}}d\theta_0 \quad (3.140)$$

である．

証明　ξ^t の分布およびパラメタの初期値の分布 $q_0(\theta_0)$ に関する平均操作は独立に行なえることを注意しておく．まず，$\theta_t-\theta_{opt}$ の 1 次，および 2 次モーメントを調べる．定理 3.1，3.2 より

$$\begin{aligned}
&E_{\xi^t,\theta_0}\left(\theta_t-\theta_{opt}\right)\\
&\quad = (E-\varepsilon CQ)^t(\theta_0-\theta_{opt}) \quad (3.141)\\
&E_{\xi^t,\theta_0}\left((\theta_t-\theta_{opt})(\theta_t-\theta_{opt})^{\mathrm{T}}\right)\\
&\quad = E_{\theta_0}\left(E_{\xi^t}\left((\theta_t-\hat{\theta}_t)(\theta_t-\hat{\theta}_t)^{\mathrm{T}}\right)\right) + E_{\theta+0}\left((\hat{\theta}_t-\theta_{opt})(\hat{\theta}_t-\theta_{opt})^{\mathrm{T}}\right)\\
&\quad = \{E-(E-\varepsilon\Xi_{CQ})^t\}\varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}}
\end{aligned}$$

$$
\begin{aligned}
&-\{(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t-(E-\varepsilon\Xi_{CQ})^t\}E_{\theta_0}\left((\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}}\right)\\
&+(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t E_{\theta_0}\left((\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}}\right)\\
=\ & \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}}\\
&+(E-\varepsilon\Xi_{CQ})^t\left\{E_{\theta_0}\left((\theta_0-\theta_{opt})(\theta_0-\theta_{opt})^{\mathrm{T}}\right)-\varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}}\right\}\\
=\ & \varepsilon V+(E-\varepsilon\Xi_{CQ})^t(V_0-\varepsilon V) \qquad (3.142)
\end{aligned}
$$

と求められる．これを用いて予測損失を最適パラメタ θ_{opt} のまわりで展開すると

$$
\begin{aligned}
&E_{\xi^t,\theta_0}\left(L_P(\theta_t)\right)\\
&=\ E_{\xi^t,\theta_0}\left(D(p,\theta_{opt}+(\theta_t-\theta_{opt}))\right)\\
&=\ D(p,\theta_{opt})+\partial_i D(p,\theta_{opt})E_{\xi^t,\theta_0}\left((\theta_t-\theta_{opt})^i\right)\\
&\quad+\frac{1}{2}\partial_i\partial_j D(p,\theta_{opt})E_{\xi^t,\theta_0}\left((\theta_t-\theta_{opt})^i(\theta_t-\theta_{opt})^j\right)\\
&=\ D(p,\theta_{opt})+\frac{1}{2}q_{ij}(\varepsilon V+(E-\varepsilon\Xi_{CQ})^t(V_0-\varepsilon V))^{ij}\\
&=\ D(p,\theta_{opt})+\frac{1}{2}\,\mathrm{tr}\,(\varepsilon V+(E-\varepsilon\Xi_{CQ})^t(V_0-\varepsilon V))Q \qquad (3.143)
\end{aligned}
$$

となり，定理が得られる．　∎

定理の第 3 項に現れる $V_0Q-\varepsilon QV$ は ε が初期値の偏差に比べて十分小さければ通常は正定値行列となるので，学習曲線は $\|E-\varepsilon\Xi_{CQ}\|^t=O(\exp(-2\varepsilon\lambda_1 t))$ 程度の速さで減衰していくことがわかる．

例 5 (計算機シミュレーション 1.(つづき))　図 3.7 は 1 台の機械の予測損失の変化を示したものである．図 3.8 は 100 台の機械の予測損失を平均して求めた学習曲線である．実線は理論的に導かれた学習曲線を表している．

3.6　確率的降下法の動特性

ここまではシステムの確率構造が時不変だとして議論を進めてきたが，現実にはシステムは時間とともに変動し，学習によってそれを追従する場合が多い．以下では学習系のステップ応答 (step response) と周波数応答 (frequency response) を調べ，その動特性を考察する．

まずはじめにステップ応答について考える．パラメタの期待値 $\bar{\theta}_t$ のステップ応答を求めるには，最適パラメタに定常偏差を加えて，期待値に関する漸化式を解き直せば

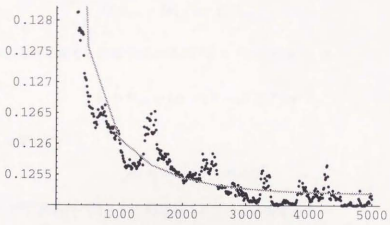


図 3.7: 1 台の機械の予測損失の変化.

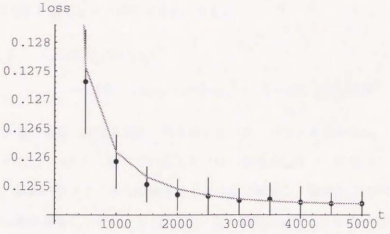


図 3.8: 100 台の機械の予測損失の平均値の変化.

よい．具体的には時刻 0 におけるパラメタの初期状態の期待値を θ_{opt}，最適パラメタを $\theta_{opt}+\delta\theta$ として漸化式 (3.67) を解く．ただしここでは簡単のため $\delta\theta$ は十分小さいとして

$$G(\theta_{opt}+\delta\theta) \simeq G(\theta_{opt}) = G \tag{3.144}$$

$$Q(\theta_{opt}+\delta\theta) \simeq Q(\theta_{opt}) = Q \tag{3.145}$$

を仮定する．これを解くと時刻 t におけるパラメタの期待値 $\hat{\theta}_t$ は

$$\hat{\theta}_t = \theta_{opt} + (E-(E-\varepsilon CQ)^t)\delta\theta \tag{3.146}$$

となる．したがって，

$$S_t = E-(E-\varepsilon CQ)^t \tag{3.147}$$

すれば，S_t が学習系のパラメタの期待値のステップ応答を示す．

同様に定常偏差を与えられたときパラメタの分散がどう変化するか調べる．時刻 0 におけるパラメタの初期状態を最適パラメタ θ_{opt} に対する定常状態，すなわち期待値 θ_{opt}，分散 $\varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}}$ とし，最適パラメタが $\theta_{opt}+\delta\theta$ に変移したとして式 (3.79) を解けばよい．これは結局，

$$v_0 = 0, \quad u_0 = -\varepsilon^2\Phi_{CQ}\delta\theta\delta\theta^{\mathrm{T}} \tag{3.148}$$

として式 (3.86) を解くことと同等である．ゆえに

$$\begin{aligned}\hat{V}_t &= \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} \\ &\quad -\left\{(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t-(E-\varepsilon\Xi_{CQ})^t\right\}\delta\theta\delta\theta^{\mathrm{T}}\end{aligned} \tag{3.149}$$

となる．第 2 項が変動に対する過渡応答を表している．パラメタの分散は，パラメタが最適パラメタ $\theta_{opt}+\delta\theta$ から離れているときには一旦減少しながら最適パラメタに近付いていくが，最適パラメタにある程度近付くと今度は増加し，最終的には学習系が安定するもとの分散に戻ることがわかる (図 3.9)．この理由は定理 3.2 における分散の変化の仕方と同様に説明される．

次に周波数応答を求める．最適パラメタが周期 $2\pi/\omega_i$ で変動しているとする．すなわち，時刻 t における最適パラメタを $\theta_{opt}+\delta\theta_t$ としたとき，

$$\theta_{opt}+\delta\theta_t = \theta_{opt}+\eta_i\sin\omega_i t \tag{3.150}$$

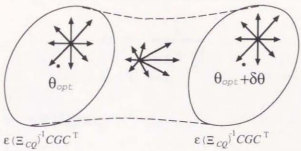


図 3.9: 最適パラメタが変動した場合の逐次型学習系の分散の変化.

とする．ここに，η_i は行列 CQ の固有値 λ_i に対する固有ベクトルとする．

$$CQ\eta_i = \lambda_i \eta_i \tag{3.151}$$

また $\delta\theta_t$ は十分小さいとして

$$G(\theta_{opt} + \delta\theta_t) \simeq G(\theta_{opt}) = G \tag{3.152}$$

$$Q(\theta_{opt} + \delta\theta_t) \simeq Q(\theta_{opt}) = Q \tag{3.153}$$

とする．パラメタの期待値 $\hat{\theta}_i$ の漸化式は，式 (3.67) から

$$\hat{\theta}_{t+1} = (E - \varepsilon CQ)\hat{\theta}_t + \varepsilon CQ\theta_{opt} + \varepsilon\lambda_i\eta_i \sin\omega_i t \tag{3.154}$$

となる．定常振動解を

$$\hat{\theta}_t = \theta_{opt} + a_i\eta_i \sin(\omega_i t + \phi_i) \tag{3.155}$$

とおいて，振幅の拡大率 a_i と位相差 ϕ_i を求める方程式をたてると，

$$\varepsilon\lambda_i \cos\phi_i = a_i\{\cos\omega_i - (1 - \varepsilon\lambda_i)\} \tag{3.156}$$

$$\varepsilon\lambda_i \sin\phi_i = -a_i \sin\omega_i \tag{3.157}$$

が得られる．これを解くと

$$a_i = \frac{\varepsilon\lambda_i}{\sqrt{(\cos\omega_i - (1 - \varepsilon\lambda_i))^2 + (\sin\omega_i)^2}} \tag{3.158}$$

$$\tan\phi_i = -\frac{\sin\omega_i}{\cos\omega_i - (1 - \varepsilon\lambda_i)} \tag{3.159}$$

となる．ω_i が微小，すなわち変動が非常に緩やかなときには，

$$\alpha_i = \frac{\omega_i}{\varepsilon\lambda_i} \tag{3.160}$$

おき，α_i が十分小さいと仮定してその高次の項を省略すると，

$$a_i = \frac{1}{\sqrt{1+\alpha_i^2}} \tag{3.161}$$

$$\phi_i = -\alpha_i \tag{3.162}$$

と近似できる．よって定常振動解は

$$\hat{\theta}_t = \theta_{opt} + \frac{1}{\sqrt{1+\alpha_i^2}}\eta_i \sin(\omega_i t - \alpha_i) \tag{3.163}$$

である．すなわち，最適パラメタが周期 $2\pi/\omega_i$ で CQ の固有値 λ_i の固有ベクトルの方向に変動するときには，振幅が $1/\sqrt{1+\alpha_i^2}$ 倍になり，位相が α_i だけ遅れて追従することになる．$\delta\theta_t$ の変動の方向が一般の場合には行列 CQ の固有ベクトルを重ね合わせればよい．いずれにせよ α_i が小さい，すなわち

$$\omega_i \ll \varepsilon\lambda_i, \quad (i = 1, \ldots, m) \tag{3.164}$$

ならば，学習系は変動を十分よく追従することがわかる．

同様に分散の周波数応答を調べる．式 (3.150) の変動を加えたとき，α_i が十分小さければパラメタの期待値と最適パラメタの差は

$$\begin{aligned}\hat{\theta}_t - (\theta_{opt} + \delta\theta_t) &= \eta_i \left(\frac{1}{\sqrt{1+\alpha^2}} \sin(n\omega_i - \alpha_i) - \sin(n\omega_i) \right) \\ &= \eta_i \left(\alpha_i \cos(n\omega_i) + O(\alpha_i^2) \right)\end{aligned} \tag{3.165}$$

となるから，パラメタの共分散の漸化式は，式 (3.79) を用いて

$$\hat{V}_{t+1} = (E - \varepsilon\Xi_{CQ})\hat{V}_t + \varepsilon^2 CGC^{\mathrm{T}} - \varepsilon^2\alpha_i^2 \cos^2(\omega_i t)\Phi_{CQ}\eta_i\eta_i^{\mathrm{T}} \tag{3.166}$$

となる．$\hat{V}_t$ の定常振動解を

$$\hat{V}_t = \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} + \left\{ b_i \cos(2n\omega_i + \phi_i') + c_i \right\} \eta_i\eta_i^{\mathrm{T}} \tag{3.167}$$

とおき，CQ の固有ベクトルと CQ から作られる演算子の関係式

$$\Xi_{CQ}\eta_i\eta_i^{\mathrm{T}} = 2\lambda_i\eta_i\eta_i^{\mathrm{T}} \tag{3.168}$$

$$\Phi_{CQ}\eta_i\eta_i^{\mathrm{T}} = \lambda_i^2\eta_i\eta_i^{\mathrm{T}} \tag{3.169}$$

を用いて，振幅の拡大率 b_i と直流バイアス成分 c_i と位相差 ϕ_i' を求める方程式をたてると，

$$\frac{\varepsilon^2\lambda_i^2\alpha_i^2}{2}\cos\phi_i' = b_i(1-2\varepsilon\lambda_i-\cos 2\omega_i) \tag{3.170}$$

$$\frac{\varepsilon^2\lambda_i^2\alpha_i^2}{2}\sin\phi_i' = b_i\sin 2\omega_i \tag{3.171}$$

$$\frac{\varepsilon^2\lambda_i^2\alpha_i^2}{2} = -2\varepsilon\lambda_i c_i \tag{3.172}$$

となる．これを解いて

$$b_i = \frac{\varepsilon^2\lambda_i^2\alpha_i^2}{2\sqrt{(1-2\varepsilon\lambda_i-\cos 2\omega_i)^2+(\sin 2\omega_i)^2}} \tag{3.173}$$

$$c_i = -\frac{1}{4}\varepsilon\lambda_i\alpha_i^2 \tag{3.174}$$

$$\tan\phi_i' = \frac{\sin 2\omega_i}{1-2\varepsilon\lambda_i-\cos 2\omega_i} \tag{3.175}$$

であるが，期待値の場合と同様に式 (3.160) を用いて高次の項を省略すると

$$b_i = \frac{\varepsilon\lambda_i\alpha_i^2}{4\sqrt{1+\alpha_i^2}} \tag{3.176}$$

$$c_i = -\frac{\varepsilon\lambda_i\alpha_i^2}{4} \tag{3.177}$$

$$\phi_i' = \alpha_i \tag{3.178}$$

と近似できる．よって定常振動解は

$$\hat{V}_t = \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}} - \frac{\varepsilon\lambda_i\alpha_i^2}{4}\left(1-\frac{1}{\sqrt{1+\alpha_i^2}}\cos(2\omega_i t+\alpha_i)\right)\eta_i\eta_i^{\mathrm{T}} \tag{3.179}$$

となる．第 2 項は常に負であるから分散は最適パラメタが動かない場合よりわずかに小さくなることがわかる．

3.7 学習曲線の動特性

前節で行なった動特性の解析結果を用いると学習曲線についても，そのステップ応答と周波数応答が求められる．

まずステップ応答について考える．逐次型学習の結果，時刻 0 まで定常状態にあった学習系で時刻 1 においてシステムが p から $p+\delta p$ に変化し，最適ベクトルが θ から $\theta+\delta\theta$ に変わったとする．このときの学習曲線の変動を求める．各時刻に例題がひ

とつずつ提示されるとすれば，

$$
\begin{aligned}
&E_{\xi^t,\theta_0}\left((\theta_t-(\theta+\delta\theta))(\theta_t-(\theta+\delta\theta))^{\mathrm{T}}\right)\\
&= E_{\xi^t,\theta_0}\left((\theta_t-\hat{\theta}_t)(\theta_t-\hat{\theta}_t)^{\mathrm{T}}\right)+E_{\xi^t,\theta_0}\left((\hat{\theta}_t-(\theta+\delta\theta))(\hat{\theta}_t-(\theta+\delta\theta))^{\mathrm{T}}\right)\\
&= \hat{V}_t+(E-\varepsilon\Xi_{CQ}+\varepsilon^2\Phi_{CQ})^t\delta\theta\delta\theta^{\mathrm{T}}\\
&= \varepsilon V+(E-\varepsilon\Xi_{CQ})^{n-1}\delta\theta\delta\theta^{\mathrm{T}} \qquad (3.180)
\end{aligned}
$$

であるから，$L_T(\theta)$ を $\theta+\delta\theta$ のまわりで展開して

$$
\begin{aligned}
&E_{\xi^t,\theta_0}(L_T(\theta))\\
&= D(p+\delta p,\theta+\delta\theta)+\frac{1}{2}\operatorname{tr}QE_{\xi^t,\theta_0}\left((\theta_t-(\theta+\delta\theta))(\theta_t-(\theta+\delta\theta))^{\mathrm{T}}\right))\\
&= D(p+\delta p,\theta+\delta\theta)+\frac{1}{2}\varepsilon\operatorname{tr}QV\\
&\quad+\frac{1}{2}\operatorname{tr}(E-\varepsilon\Xi_{CQ})^tQ\delta\theta\delta\theta^{\mathrm{T}} \qquad (3.181)
\end{aligned}
$$

となる．第 3 項が過渡応答の部分である．

次に周波数応答について考える．時刻 t におけるシステムが $p+\delta p_t$ で表され，最適ベクトルが

$$\theta_{opt}+\delta\theta_t=\theta_{opt}+\eta_i\sin\omega_it \qquad (3.182)$$

となったとする．ただし，η_i は行列 CQ の固有値 λ_i に対する固有ベクトルで，

$$CQ\eta_i=\lambda_i\eta_i \qquad (3.183)$$

を満たしているのは前節と同様である．このときの学習曲線の定常振動状態を求める．確率的降下法の周波数特性の関係から，

$$
\begin{aligned}
&E_{\theta_t}\left((\theta_t-(\theta+\delta\theta_t))(\theta_t-(\theta+\delta\theta_t))^{\mathrm{T}}\right)\\
&= E_{\theta_t}\left((\theta_t-\hat{\theta}_t)(\theta_t-\hat{\theta}_t)^{\mathrm{T}}\right)+E_{\theta_t}\left((\hat{\theta}_t-(\theta+\delta\theta_t))(\hat{\theta}_t-(\theta+\delta\theta_t))^{\mathrm{T}}\right)\\
&= \hat{V}_t+(\alpha_i\cos\omega_it)^2\eta_i\eta_i^{\mathrm{T}}\\
&= \varepsilon(\Xi_{CQ})^{-1}CGC^{\mathrm{T}}-\frac{\varepsilon\lambda_i\alpha_i^2}{4}(1-\frac{1}{\sqrt{1+\alpha_i^2}}\cos(2\omega_it+\alpha_i))\eta_i\eta_i^{\mathrm{T}}\\
&\quad+\frac{\alpha_i^2}{2}(\cos2\omega_it+1)\eta_i\eta_i^{\mathrm{T}}\\
&\simeq \varepsilon V+\left\{(\frac{1}{2}-\frac{\varepsilon\lambda_i}{4})+(\frac{1}{2}+\frac{\varepsilon\lambda_i}{4\sqrt{1+\alpha_i^2}})\cos2\omega_it\right\}\alpha_i^2\eta_i\eta_i^{\mathrm{T}} \qquad (3.184)
\end{aligned}
$$

であるから，$L_T(\theta_t)$ を $\theta+\delta\theta_t$ のまわりで展開して

$$
\begin{aligned}
& E_{\theta_t}\left(L_T(\theta_t)\right) \\
&= D(p+\delta p_t, \theta+\delta\theta_t) + \frac{1}{2}\,\mathrm{tr}\,Q E_{\theta_t}\left((\theta_t-(\theta+\delta\theta_t))(\theta_t-(\theta+\delta\theta_t))^{\mathrm{T}}\right) \\
&= D(p+\delta p, \theta+\delta\theta_t) + \frac{1}{2}\varepsilon\,\mathrm{tr}\,QV \\
&\quad + \frac{1}{2}\left\{\left(\frac{1}{2}-\frac{\varepsilon\lambda_i}{4}\right) + \left(\frac{1}{2}+\frac{\varepsilon\lambda_i}{4\sqrt{1+\alpha_i^2}}\right)\cos 2\omega_i t\right\}\alpha_i^2\,\mathrm{tr}\,Q\eta_i\eta_i^{\mathrm{T}} \qquad (3.185)
\end{aligned}
$$

となる．第 3 項が最適パラメタの変動による学習曲線の変動を表しており，そのオーダーは

$$
O\left(\frac{\omega_i^2}{(\varepsilon\lambda_i)^2}\right) \qquad (3.186)
$$

である．このため ε に比べ $\omega_i^2/(\varepsilon\lambda_i)^2$ が十分に小さい場合には機械は損失関数で評価した性能を損なわずに，いいかえると安定した性能を維持したままシステムの変動を追従できることになる．また第 3 項は $\cos 2\omega_i t$ が -1 をとる近くでは負の値をとるため，通常残る誤差 $1/2\varepsilon\,\mathrm{tr}\,QV$ より誤差が減る場合があり得るが，これは α_i が十分小さいときにはほとんど無視できる．

第 4 章

非逐次型学習の特性

4.1 非逐次型の学習

t 個の例題を用いた学習の中で，提示される例題の順序に依存しない方法を非逐次型の学習と定義した．ここでは損失関数を用いる立場から，具体的な手続きとして非逐次型学習を定義する．

まず，与えられた t 個の例題

$$\xi^t = \{(x_1, y_1), (x_2, y_2), \cdots, (x_t, y_t)\} \tag{4.1}$$

から経験分布 (empirical distribution)

$$p^t(x, y) = \frac{1}{t} \sum_{i=1}^{t} \delta(x - x_i, y - y_i) \tag{4.2}$$

を構成する．このとき分布 $p^t(x, y)$ は正確には例題 ξ^t の関数 $p^t(x, y; \xi^t)$ となるが，混乱のない限り ξ^t を省略して表記する．経験分布に対する最適パラメタと非逐次型学習を次のように定義する．

定義 4.1 経験分布に対する平均損失関数

$$\begin{aligned} D(p^t, \theta) &= \int d(x, y; \theta) p^t(x, y) dx dy \\ &= \frac{1}{t} \sum_{i=1}^{t} d(x_i, y_i; \theta) \end{aligned} \tag{4.3}$$

を最小にする最適パラメタ θ^t_{opt}

$$D(p^t, \theta^t_{opt}) = \min_{\theta} D(p^t, \theta) \tag{4.4}$$

を経験分布 p^t に対する最適パラメタという．

このときパラメタ θ^t_{opt} を探索する過程を非逐次型学習と定義し，推定されたパラメタを θ^t と書く．

パラメタ θ^t_{opt} も ξ^t の関数 $\theta^t_{opt}(\xi^t)$ となるが，この場合も混乱のない限り $p^t(x,y)$ と同様に ξ^t を省略して表記する．

非逐次型学習は例題から作られる経験分布 $p^t(x,y)$ によって真のシステムの分布 $p(x,y)$ を近似し，パラメタの推定を行うことと定義することもできる．よく知られているように，t が十分大きければ $p^t(x,y)$ は $p(x,y)$ の性質をよく表すことができる．もう少し正確に言えば，f を x, y の関数とするとき，分布 $p(x,y)$ に関する $f(x,y)$ の分散が有限であれば，t を大きくしたとき $p^t(x,y)$ に関する平均を $p(x,y)$ に関する平均にいくらでも近付けることができる．

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t} f(x_i,y_i)=\int f(x,y)p(x,y)dxdy \tag{4.5}$$

このため t が十分大きければ，経験分布に対する最適パラメタ θ^t_{opt} は真の最適パラメタ θ_{opt} の十分良い推定値となる．

注． 実際に最適パラメタを探索するには，式 (4.3) を θ に関して偏微分し，各微係数を 0 とおいた m 元連立方程式を解くことになる．

$$\partial_j D(p^t,\theta^t_{opt})=\frac{1}{t}\sum_{i=1}^{t}\partial_j d(x_i,y_i;\theta^t_{opt})=0,\quad j=1,\ldots,m \tag{4.6}$$

ただし，式 (4.4) と (4.6) は同値ではない．なぜなら最小値を与えない θ でも極小値 (local minima) となるなら式 (4.6) を満たすからである．以下では最適パラメタ θ_{opt} の近傍において諸々の関数の振舞いを考えるが，これらの議論のほとんどが，そのまま極小値の近傍でもなりたつことを注意しておく．

また式 (4.6) の方程式は一般には非線形連立方程式となるため，特殊な場合を除いて解析的に解くことは難しい．このため反復法 (iteration) を用いた近似計算によって θ^t_{opt} を求めることになる．反復法は大きく分けると次の 2 つの方法が考えられる．ひとつは，最急降下法 (gradient discent method) やニュートン法 (Newton's method) などの数値演算手法を用いて式 (4.6) を直接解く方法である．これを一括型と呼ぶことにする．もうひとつは，経験分布 $p^t(x,y)$ を疑似的なシステムと見做し，これからランダムに例題を生成する，すなわち ξ^t からランダムに例題を選んで与えることにより確率的降下法，あるいはその発展型を用いる方法である．これを確率型と呼ぶことにする．

一般に統計学で行われるモデル推定などでは，前者の数値演算手法が用いられる．一方，神経回路網の学習では後者の手法，あるいはそれを高速化したり，極小値に捕まらないような工夫をしたものを用いることが多い．これは非常に多数の素子を用いる神経回路網において通常の数値演算手法を用いると膨大な記憶領域を必要とするのに対し，Back-propagation や Learning Vector Quantization はパラメタの更新が局所演算によって行なえるため，計算に必要とされる記憶領域が少なくてすむからである．二つの場合のどちらを選ぶかは，設計したモデルの性質，利用できる計算機の能力等によって決められるべきものである．

非逐次型学習により推定されたパラメタは，探索に数値演算手法を用いた場合には初期値の選び方，繰り返し演算の打ち切り回数，丸め誤差等の計算誤差からある範囲に散らばる．また確率的降下法を用いた場合には，前章で議論したように十分な学習時間を経た後にはほぼ正規分布になっている．いづれにせよ推定パラメタ θ^t は学習の特性により確率変数となる．例題の組 ξ^t を固定したときの θ^t の分布を $q^t(\theta^t|\xi^t)$ と書くことにし，この分布に関して平均と分散をとる操作を $E_{\theta^t|\xi^t}$, $V_{\theta^t|\xi^t}$ で表す．以下では分布 $q^t(\theta^t|\xi^t)$ に関するパラメタ θ^t の高次モーメントは有界であることを仮定し，特に平均と分散については次のように書けるとする．

$$E_{\theta^t|\xi^t}\left(\theta^t\right) = \theta^t_{opt} \tag{4.7}$$

$$V_{\theta^t|\xi^t}\left(\theta^t\right) = \varepsilon V(\theta^t_{opt}) \tag{4.8}$$

ここに ε は学習の精度を表すパラメタである．特に確率的降下法を用いて十分学習を行なった場合には，

$$\varepsilon V(\theta^t_{opt}) = \varepsilon(\Xi_{CQ^t(\theta^t_{opt})})^{-1} C G^t(\theta^t_{opt}) C^{\mathrm{T}} \tag{4.9}$$

$$G^t(\theta^t_{opt}) = (g^t_{ij}(\theta^t_{opt})) = \frac{1}{t}\sum_{k=1}^{t} \partial_i d(x_k, y_k; \theta^t_{opt}) \partial_j d(x_k, y_k; \theta^t_{opt}) \tag{4.10}$$

$$Q^t(\theta^t_{opt}) = (q^t_{ij}(\theta^t_{opt})) = \frac{1}{t}\sum_{k=1}^{t} \partial_i \partial_j d(x_k, y_k; \theta^t_{opt}) \tag{4.11}$$

となり，推定パラメタの分散は θ^t_{opt} の関数として損失関数を用いて明示的に書ける．

4.2 予測損失と学習損失

学習によって得たパラメタの推定値 θ^t に対して次の 2 つの損失を考える．

定義 4.2 パラメタ θ^t を持つ機械が，新しい例題 (x, y) に対して持つ平均損失を予測損失 (predictive loss) $L_P(\theta^t)$ という．

$$\begin{aligned} L_P(\theta^t) &= \int d(x, y; \theta^t) p(y|x) p(x) dx dy \\ &= D(p, \theta^t) \end{aligned} \tag{4.12}$$

定義 4.3 パラメタ θ^t をもつ機械が，学習に用いた例題 ξ^t に対して持つ平均損失を学習損失 (training loss) $L_T(\theta^t)$ という．

$$\begin{aligned} L_T(\theta^t) &= \frac{1}{t} \sum_{i=1}^{t} d(x_i, y_i; \theta^t) \\ &= D(p^t, \theta^t) \end{aligned} \tag{4.13}$$

予測損失は，学習を終えた機械が現実の環境の中におかれたときどれだけの能力を発揮するかの指標となる．つまり，学習時に与えられた例題以外の入力が与えられたとき，機械が予測する出力がどれだけの損失を持っているかの平均評価になっている．一方学習損失は，学習を終えた機械が学習時と同じ環境，つまり学習に用いた例題しか与えられないような状況に置かれたとき機械の予測する出力がどれだけの損失を持つかを表している．予測損失を最も小さくするパラメタはシステムの分布 $p(x, y)$ に対する最適パラメタ θ_{opt} であり，そのときの予測損失はシステムに対する平均損失関数の最小値に一致する．また学習損失を最も小さくするパラメタは経験分布 $p^t(x, y)$ に対する最適パラメタ θ^t_{opt} であり，このときの学習損失は一般にシステムに対する平均損失関数の最小値より小さくなる．これは学習損失に関しては

$$L_T(\theta^t_{opt}) \leq L_T(\theta_{opt}) \tag{4.14}$$

が成り立っているが，上式の右辺は中心極限定理によりシステムの平均損失関数の最小値 $D(p, \theta_{opt})$ に近付くからである．

次に学習曲線を定義する．

定義 4.4 予測損失を θ^t と ξ^t の分布に関して平均したものを平均予測損失という．平均予測損失を t の関数としてみたとき，これを学習曲線と呼ぶ．

学習曲線は例題数を増やしていくと学習された機械の損失が平均的にどのくらい減少していくかを表したものである．もう少し直観的に言うと，たくさんの機械があって，それぞれに別々の t 個の例題を与え逐次型学習を行なうと，これらの機械は平均

損失の意味で最も良い機械のまわりにばらつく．このばらつきを予測損失で測ったものが，平均予測損失になる．t を増やせば増やすほど機械は最も良い機械のまわりに集まってくるが，この集まり具合を平均予測損失によって見てやろうと言うのが，学習曲線の考え方である．

4.3 学習曲線の特性

学習曲線の具体的な形を求める前に，経験分布 $p^t(x,y)$ に対する最適パラメタ θ^t_{opt} の性質に関して成り立つ補題を証明しておく．

補題 4.1 θ^t_{opt} の分布は漸近的に正規分布

$$\theta^t_{opt} \sim N(\theta_{opt}, \frac{1}{t}Q^{-1}GQ^{-1}) \tag{4.15}$$

に一致する．

証明 θ^t_{opt} が $p^t(x,y)$ の平均損失関数を最小にすることより，

$$\begin{aligned} \nabla D(p^t, \theta^t_{opt}) &= E_{p^t}\left(\nabla d(x,y;\theta^t_{opt})\right) \\ &= \frac{1}{t}\sum_{i=1}^{t}\nabla d(x_i,y_i;\theta^t_{opt}) \\ &= 0 \end{aligned} \tag{4.16}$$

が成り立つ．t が十分大きければ $|\xi^t - \theta_{opt}|$ は微小なので上の式を θ_{opt} のまわりで展開して高次の項を省略すると，

$$\frac{1}{t}\sum_{i=1}^{t}\nabla d(x_i,y_i;\theta_{opt}) + \frac{1}{t}\sum_{i=1}^{t}\nabla\nabla d(x_i,y_i;\theta_{opt})(\theta^t_{opt}-\theta_{opt}) = 0 \tag{4.17}$$

が得られる．十分大きな t に対しては大数の法則により

$$\frac{1}{t}\sum_{i=1}^{t}\nabla\nabla d(x_i,y_i;\theta_{opt}) \simeq Q \tag{4.18}$$

としてよい．また

$$E_p\left(\frac{1}{\sqrt{t}}\nabla d(x_i,y_i;\theta_{opt})\right) = \frac{1}{\sqrt{t}}\nabla D(p,\theta_{opt}) = 0 \tag{4.19}$$

$$V_p\left(\frac{1}{\sqrt{t}}\nabla d(x_i,y_i;\theta_{opt})\right) = \frac{1}{t}G \tag{4.20}$$

であるから，中心極限定理により

$$\sum_{i=1}^{t} \frac{1}{\sqrt{t}} \nabla d(x_i, y_i; \theta_{opt}) \tag{4.21}$$

は平均 0 共分散行列 G の正規分布に漸近する．すなわち

$$\sum_{i=1}^{t} \frac{1}{\sqrt{t}} \nabla d(x_i, y_i; \theta_{opt}) \sim N(0, G) \tag{4.22}$$

よって 式 (4.17) を整理して，

$$-\left(\frac{1}{t}\sum_{i=1}^{t} \nabla\nabla d(x_i, y_i; \theta_{opt})\right) \sqrt{t}(\theta^t_{opt} - \theta_{opt}) = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} \nabla d(x_i, y_i; \theta_{opt}) \tag{4.23}$$

から

$$-Q\sqrt{t}(\theta^t_{opt} - \theta_{opt}) \sim N(0, G) \tag{4.24}$$

がいえる．ゆえに

$$(\theta^t_{opt} - \theta_{opt}) \sim N(0, \frac{1}{t} Q^{-1} G Q^{-1}) \tag{4.25}$$

が漸近的に成り立つ． ∎

推定されたパラメタ θ^t は ξ^t の関数であるから，分布 $q^t(\theta^t|\xi^t)$ に関してとった平均は，ξ^t の関数として確率変数になっていることに注意しておく．またシステムの最適パラメタのまわりでの推定値の分散 $\varepsilon V(\theta^t_{opt})$ の最適パラメタ θ_{opt} での値を εV で表す．

$$V = V(\theta_{opt}) \tag{4.26}$$

以上の準備のもとで予測損失に関して次の定理が証明される．

定理 4.1 予測損失の期待値は漸近的に

$$E_{\xi^t, \theta^t}\left(L_P(\theta^t)\right) = D(p, \theta_{opt}) + \frac{1}{2t} \operatorname{tr} G Q^{-1} + \frac{\varepsilon}{2} \operatorname{tr} QV \tag{4.27}$$

で与えられる．

証明 定理の式の証明の前に $\theta^t - \theta_{opt}$ の分布を考える．平均は明らかに 0 であり，分散は

$$V_{\xi^t, \theta^t}\left((\theta^t - \theta_{opt})\right)$$

$$
\begin{aligned}
&= E_{\xi^t,\theta^t}\left((\theta^t-\theta^t_{opt}+\theta^t_{opt}-\theta_{opt})(\theta^t-\theta^t_{opt}+\theta^t_{opt}-\theta_{opt})^{\mathrm{T}}\right)\\
&= E_{\xi^t}\left(E_{\theta^t|\xi^t}\left((\theta^t-\theta^t_{opt})(\theta^t-\theta^t_{opt})^{\mathrm{T}}\right)\right)\\
&\quad +2E_{\xi^t}\left(E_{\theta^t|\xi^t}\left((\theta^t-\theta^t_{opt})\right)(\theta^t_{opt}-\theta_{opt})^{\mathrm{T}}\right)\\
&\quad +E_{\xi^t}\left((\theta^t_{opt}-\theta_{opt})(\theta^t_{opt}-\theta_{opt})^{\mathrm{T}}\right)\\
&= \varepsilon V(\theta^t_{opt})+\frac{1}{t}Q^{-1}GQ^{-1}\\
&= \varepsilon V+\frac{1}{t}Q^{-1}GQ^{-1}+O(\frac{\varepsilon}{t})
\end{aligned}
\tag{4.28}
$$

となる．

行列 Q, Q^{-1}, G, V の第 i 行第 j 列成分をそれぞれ q_{ij}, q^{ij}, g_{ij}, v^{ij} で表す．ただしこれらの行列はいずれも対称であることに注意する．

$$
\theta=\theta^t-\theta_{opt} \tag{4.29}
$$

として予測損失を最適パラメタ θ_{opt} のまわりで展開し，これに関する高次の項を省略すると

$$
\begin{aligned}
&E_{\xi^t,\theta^t}\left(L_P(\theta^t)\right)\\
&= E_{\xi^t,\theta^t}\left(E_\xi\left(d(x,y;\theta^t)\right)\right)\\
&= E_{\xi^t,\theta^t}\left(E_\xi\left(d(x,y;\theta_{opt}+\theta)\right)\right)\\
&= E_{\xi^t,\theta^t}\left(E_\xi\left(d(x,y;\theta_{opt})\right)\right)+E_{\xi^t,\theta^t}\left(E_\xi\left(\partial_i d(x,y;\theta_{opt})\theta^i\right)\right)\\
&\quad +\frac{1}{2}E_{\xi^t,\theta^t}\left(E_\xi\left(\partial_i\partial_j d(x,y;\theta_{opt})\theta^i\theta^j\right)\right)\\
&= E_\xi\left(d(x,y;\theta_{opt})\right)+E_\xi\left(\partial_i d(x,y;\theta_{opt})\right)E_{\xi^t,\theta^t}\left(\theta^i\right)\\
&\quad +\frac{1}{2}E_\xi\left(\partial_i\partial_j d(x,y;\theta_{opt})\right)E_{\xi^t,\theta^t}\left(\theta^i\theta^j\right)\\
&= D(p,\theta_{opt})+\frac{1}{2}q_{ij}(\varepsilon v^{ij}+\frac{1}{t}q^{ik}g_{kl}q^{lj})\\
&= D(p,\theta_{opt})+\frac{1}{2}(q_{ij}\varepsilon v^{ij}+\frac{1}{t}q^{ij}g_{ij})\\
&= D(p,\theta_{opt})+\frac{1}{2}\,\mathrm{tr}\,(\varepsilon QV+\frac{1}{t}GQ^{-1})
\end{aligned}
\tag{4.30}
$$

となり，定理が証明される．∎

確率的降下法のように学習により得られるパラメタの分布が正規分布と見做せる場合には，予測損失の分散を評価することができる．

定理 4.2 予測損失の分散は漸近的に

$$V_{\xi^t,\theta^t}\left(L_P(\theta^t)\right) = \frac{1}{2t^2}\,\mathrm{tr}\,GQ^{-1}GQ^{-1} + \frac{\varepsilon^2}{2}\,\mathrm{tr}\,QVQV + \frac{\varepsilon}{t}\,\mathrm{tr}\,GV \tag{4.31}$$

に従う．

証明　まず 2 つの量

$$t_{ijk} = E_\xi\left(\partial_i\partial_j\partial_k d(x,y;\theta_{opt})\right) \tag{4.32}$$

$$s_{ijkl} = E_\xi\left(\partial_i\partial_j\partial_k\partial_l d(x,y;\theta_{opt})\right) \tag{4.33}$$

を定義して，予測損失を 4 次の項まで展開する．

$$\begin{aligned}
&L_P(\theta^t)\\
&= E_\xi\left(d(x,y;\theta_{opt}+\theta)\right)\\
&= E_\xi\left(d(x,y;\theta_{opt})\right) + E_\xi\left(\partial_i d(x,y;\theta_{opt})\right)\theta^i\\
&\quad +\frac{1}{2}E_\xi\left(\partial_i\partial_j d(x,y;\theta_{opt})\right)\theta^i\theta^j + \frac{1}{6}E_\xi\left(\partial_i\partial_j\partial_k d(x,y;\theta_{opt})\right)\theta^i\theta^j\theta^k\\
&\quad +\frac{1}{24}E_\xi\left(\partial_i\partial_j\partial_k\partial_l d(x,y;\theta_{opt})\right)\theta^i\theta^j\theta^k\theta^l + O(\|\theta\|^5)\\
&= D(p,\theta_{opt}) + \frac{1}{2}q_{ij}\theta^i\theta^j + \frac{1}{6}t_{ijk}\theta^i\theta^j\theta^k + \frac{1}{24}s_{ijkl}\theta^i\theta^j\theta^k\theta^l + O(\|\theta\|^5)
\end{aligned} \tag{4.34}$$

さてこれを用いて分散を 4 次の項まで求めると

$$\begin{aligned}
&V_{\xi^t,\theta^t}\left(L_P(\theta^t)\right)\\
&= E_{\xi^t,\theta^t}\left(L_P(\theta^t)L_P(\theta^t)\right) - E_{\xi^t,\theta^t}\left(L_P(\theta^t)\right)^2\\
&= \Big\{D(p,\theta_{opt})^2 + D(p,\theta_{opt})q_{ij}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\right) + \frac{1}{3}D(p,\theta_{opt})t_{ijk}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\theta^k\right)\\
&\quad +\frac{1}{4}q_{ij}q_{kl}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\theta^k\theta^l\right) + \frac{1}{12}D(p,\theta_{opt})s_{ijkl}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\theta^k\theta^l\right) + O(\|\theta\|^5)\Big\}\\
&\quad -\Big\{D(p,\theta_{opt})^2 + D(p,\theta_{opt})q_{ij}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\right) + \frac{1}{3}D(p,\theta_{opt})t_{ijk}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\theta^k\right)\\
&\quad +\frac{1}{4}q_{ij}q_{kl}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\right)E_{\xi^t,\theta^t}\left(\theta^k\theta^l\right) + \frac{1}{12}D(p,\theta_{opt})s_{ijkl}E_{\xi^t,\theta^t}\left(\theta^i\theta^j\theta^k\theta^l\right)\\
&\quad +O(\|\theta\|^5)\Big\}\\
&= \frac{1}{4}q_{ij}q_{kl}\left\{E_{\xi^t,\theta^t}\left(\theta^i\theta^j\theta^k\theta^l\right) - E_{\xi^t,\theta^t}\left(\theta^i\theta^j\right)E_{\xi^t,\theta^t}\left(\theta^k\theta^l\right)\right\} + O(\|\theta\|^5)
\end{aligned} \tag{4.35}$$

となる．θ は正規分布で，その共分散行列の第 i 行第 j 列成分は高次の項を省略すると

$$v_*^{ij} = \varepsilon v^{ij} + \frac{1}{t}q^{ik}q^{jl}g_{kl} \tag{4.36}$$

と書ける．下にあげる補題 4.2 から

$$
\begin{aligned}
& V_{\xi^t,\theta^t}\left(L_P(\theta^t)\right) \\
&= \frac{1}{4} q_{ij} q_{kl} (v_*^{ik} v_*^{jl} + v_*^{il} v_*^{jk}) \\
&= \frac{1}{2} q_{ij} q_{kl} v_*^{ik} v_*^{jl} \\
&= \frac{\varepsilon^2}{2} q_{ij} q_{kl} v^{ik} v^{jl} + \frac{1}{2t^2} g_{ij} g_{kl} q^{ik} q^{jl} + \frac{\varepsilon}{t} g_{ij} v^{ij} \\
&= \frac{\varepsilon^2}{2} \operatorname{tr} QVQV + \frac{1}{2t^2} \operatorname{tr} GQ^{-1}GQ^{-1} + \frac{\varepsilon}{t} \operatorname{tr} GV
\end{aligned}
\tag{4.37}
$$

となり，定理が証明される．∎

補題 4.2 m 次元の確率変数ベクトル $\theta = (\theta^1, \cdots, \theta^m)$ が，平均 0，共分散行列 $V = (v^{ij})$ の正規分布 $N(0, V)$ に従うとき，その 4 次のモーメント $E\left(\theta^i \theta^j \theta^k \theta^l\right)$ は

$$
E\left(\theta^i \theta^j \theta^k \theta^l\right) = v^{ij} v^{kl} + v^{ik} v^{jl} + v^{il} v^{jk} \tag{4.38}
$$

で与えられる．

証明　共分散行列の逆行列を $V^{-1} = (v_{ij})$ と書くことにする．θ の積率母関数 $\phi(z)$ を求める．ただし $z = (z_1, \cdots, z_m)$ である．確率密度は

$$
p(\theta) = \frac{1}{\sqrt{2\pi}^m |V|^{\frac{1}{2}}} e^{-\frac{1}{2} v_{ij} \theta^i \theta^j} \tag{4.39}
$$

と書け，

$$
\begin{aligned}
\phi(z) &= E\left(e^{z_i \theta^i}\right) \\
&= \frac{1}{\sqrt{2\pi}^m |V|^{\frac{1}{2}}} \int e^{z_i \theta^i} e^{-\frac{1}{2} v_{ij} \theta^i \theta^j} d\theta \\
&= \frac{1}{\sqrt{2\pi}^m |V|^{\frac{1}{2}}} \int e^{-\frac{1}{2} v_{ij} (\theta^i - v^{ik} z_k)(\theta^j - v^{jl} z_l) + \frac{1}{2} v^{ij} z_i z_j} d\theta \\
&= e^{\frac{1}{2} v^{ij} z_i z_j}
\end{aligned}
\tag{4.40}
$$

となる．よって

$$
\begin{aligned}
& E\left(\theta^i \theta^j \theta^k \theta^l\right) \\
&= \left[\frac{\partial^4}{\partial z_i \partial z_j \partial z_k \partial z_l} e^{\frac{1}{2} v^{ab} z_a z_b} \right]_{z_1, \cdots, z_m = 0} \\
&= [\{ v^{ij} v^{kl} + v^{ik} v^{jl} + v^{il} v^{jd}
\end{aligned}
$$

$$
\begin{aligned}
&+v^{ij}v^{k\epsilon}z_\epsilon v^{lf}z_f + v^{ic}z_\epsilon v^{jd}z_d v^{kl} + v^{ik}v^{jd}z_d v^{lf}z_f \\
&+v^{ic}z_\epsilon v^{k\epsilon}z_\epsilon v^{jl} + v^{il}v^{jd}z_d v^{k\epsilon}z_\epsilon + v^{ic}z_\epsilon v^{lf}z_f v^{jk} \\
&+v^{ic}z_\epsilon v^{jd}z_d v^{k\epsilon}z_\epsilon v^{lf}z_f\}e^{\frac{1}{2}v^{ab}z_a z_b}]_{z_1,\cdots,z_m=0} \\
=\ & v^{ij}v^{kl} + v^{ik}v^{jl} + v^{il}v^{jd} \qquad (4.41)
\end{aligned}
$$

■

以上により，逐次型学習における学習曲線の漸近特性がわかった．分散に関する定理では学習の揺らぎが正規分布と見做せることを条件としたが，例えば有界な台を持つ一様分布のように，積率母関数が計算できるような分布であれば同様の手法で計算できる．ただし 4 次のモーメントを用いるため，分散特性の表現は複雑になる．数値計算により経験分布に対する最適パラメタを求める場合などは，最適値付近に計算誤差の程度に一様分布するとして，その影響を調べればよい．

次に前節に置いて予測損失と比較するために定義した学習損失の期待値の特性に関する定理を述べる．定理の前に次の補題を証明しておく．

補題 4.3 2つの例題の集合 ξ^t と $\xi^{t+1} = \{\xi^t, (x_{t+1}, y_{t+1})\}$ が与えられたとき，それぞれの経験分布に対する最適パラメタ θ^t_{opt}, θ^{t+1}_{opt} の間には

$$
\theta^{t+1}_{opt} - \theta^t_{opt} = -\frac{1}{t}Q(\theta^t_{opt})^{-1}\nabla d(x_{t+1}, y_{t+1}; \theta^t_{opt}) + O(\frac{1}{t^2}) \qquad (4.42)
$$

なる関係が成り立つ．

証明　それぞれの最適パラメタの性質から

$$
\sum_{i=1}^{t} \nabla d(x_i, y_i; \theta^t_{opt}) = 0 \qquad (4.43)
$$

$$
\sum_{i=1}^{t+1} \nabla d(x_i, y_i; \theta^{t+1}_{opt}) = 0 \qquad (4.44)
$$

が成り立つ．ここで

$$
\triangle\theta = \theta^{t+1}_{opt} - \theta^t_{opt} \qquad (4.45)
$$

とおいて，高次の項を省略すると

$$
\begin{aligned}
&\sum_{i=1}^{t+1} \nabla d(x_i, y_i; \theta^t_{opt} + \triangle\theta) \\
&= \sum_{i=1}^{t} \nabla d(x_i, y_i; \theta^t_{opt}) + \nabla d(x_{t+1}, y_{t+1}; \theta^t_{opt}) + \sum_{i=1}^{t+1} \nabla\nabla d(x_i, y_i; \theta^t_{opt})\triangle\theta \\
&= 0 \qquad (4.46)
\end{aligned}
$$

とでき，第 2 式の第 1 項が 0 となることに注意すれば，

$$\nabla d(x_{t+1}, y_{t+1}; \theta^t_{opt}) = -\sum_{i=1}^{t+1} \nabla\nabla d(x_i, y_i; \theta^t_{opt}) \triangle\theta \tag{4.47}$$

が成り立つ．t が十分大きいときには大数の法則により，

$$\sum_{i=1}^{t+1} \nabla\nabla d(x_i, y_i; \theta^t_{opt}) \simeq (t+1) Q(\theta^t_{opt}) \tag{4.48}$$

としてよいから，

$$\triangle\theta = -\frac{1}{t+1} Q(\theta^t_{opt})^{-1} \nabla d(x_{t+1}, y_{t+1}, \theta^t_{opt}) \tag{4.49}$$

が成り立つ． ∎

定理 4.3 学習損失の期待値の漸近特性は

$$E_{\xi^t, \theta^t}\left(L_T(\theta^t)\right) = D(p, \theta_{opt}) - \frac{1}{2t} \operatorname{tr} G Q^{-1} + \frac{1}{2} \varepsilon \operatorname{tr} QV \tag{4.50}$$

で与えられる．

証明 非逐次型学習では，推定されるパラメタが例題の与えられる順序に依らないから，

$$\begin{aligned} E_{\xi^t, \theta^t}\left(L_T(\theta^t)\right) &= \frac{1}{t}\sum_{i=1}^{t} E_{\xi^t, \theta^t}\left(d(x_i, y_i; \theta^t)\right) \\ &= E_{\xi^t, \theta^t}\left(d(x_t, y_t; \theta^t)\right) \end{aligned} \tag{4.51}$$

としてよい．

$$E_{\theta^t|\xi^t}\left(\|\theta^t - \theta^t_{opt}\|^3\right) = O(\varepsilon^{\frac{3}{2}}) \tag{4.52}$$

であることを用いて学習のゆらぎによる学習損失の変動を ε の 1 次の項まで求める．

$$\begin{aligned} & E_{\xi^t, \theta^t}\left(L_T(\theta^t)\right) \\ &= E_{\xi^t, \theta^t}\left(d(x_t, y_t; \theta^t_{opt} + (\theta^t - \theta^t_{opt}))\right) \\ &= E_{\xi^t}\left(d(x_t, y_t; \theta^t_{opt})\right) + E_{\xi^t}\left(\partial_i d(x_t, y_t; \theta^t_{opt})\ E_{\theta^t|\xi^t}\left((\theta^t - \theta^t_{opt})^i\right)\right) \\ &\quad + \frac{1}{2} E_{\xi^t}\left(\partial_i\partial_j d(x_t, y_t; \theta^t_{opt})\ E_{\theta^t|\xi^t}\left((\theta^t - \theta^t_{opt})^i (\theta^t - \theta^t_{opt})^j\right)\right) \\ &\quad + O(\varepsilon^{\frac{3}{2}}) \\ &= E_{\xi^t}\left(d(x_t, y_t; \theta^t_{opt})\right) + \frac{1}{2}\varepsilon E_{\xi^t}\left(\partial_i\partial_j d(x_t, y_t; \theta^t_{opt}) v^{ij}(\theta^t_{opt})\right) + O(\varepsilon^{\frac{3}{2}}) \end{aligned} \tag{4.53}$$

さて，x, y, θ の関数 $f(x,y,\theta)$ で，θ に関して 4 階微分可能なものについては次の関係式が成り立つ．まず，補題 4.3 を用いて θ^t_{opt} を θ^{t-1}_{opt} のまわりで展開する．

$$E_{\xi^{t-1}}\left(\|\theta^t_{opt}-\theta^{t-1}_{opt}\|^2\right)=O(\frac{1}{t^2}) \tag{4.54}$$

であるので，

$$\begin{aligned}
&E_{\xi^t}\left(f(x_t,y_t,\theta^t_{opt})\right)\\
&= E_{\xi^t}\left(f(x_t,y_t,\theta^{t-1}_{opt}+(\theta^t_{opt}-\theta^{t-1}_{opt}))\right)\\
&= E_{\xi^{t-1},\xi_t}\left(f(x_t,y_t,\theta^{t-1}_{opt})\right)+E_{\xi^{t-1},\xi_t}\left(\partial_i f(x_t,y_t,\theta^{t-1}_{opt})(\theta^t_{opt}-\theta^{t-1}_{opt})^i\right)\\
&\quad +O(\frac{1}{t^2})\\
&= E_{\xi^{t-1},\xi}\left(f(x,y,\theta^{t-1}_{opt})\right)\\
&\quad -\frac{1}{t}E_{\xi^{t-1},\xi}\left(q^{ij}(\theta^{t-1}_{opt})\partial_i f(x,y,\theta^{t-1}_{opt})\partial_j d(x,y,\theta^{t-1}_{opt})\right)+O(\frac{1}{t^2})
\end{aligned} \tag{4.55}$$

となる．さらに $\theta^{t-1}_{opt}-\theta_{opt}$ が正規分布に従い，

$$E_{\xi^{t-1}}\left(\theta^{t-1}_{opt}-\theta_{opt}\right)=0 \tag{4.56}$$

$$E_{\xi^{t-1}}\left((\theta^{t-1}_{opt}-\theta_{opt})^i(\theta^{t-1}_{opt}-\theta_{opt})^j\right)=\frac{1}{t}q^{ik}g_{kl}q^{lj} \tag{4.57}$$

$$E_{\xi^{t-1}}\left((\theta^{t-1}_{opt}-\theta_{opt})^i(\theta^{t-1}_{opt}-\theta_{opt})^k(\theta^{t-1}_{opt}-\theta_{opt})^k\right)=0 \tag{4.58}$$

$$E_{\xi^{t-1}}\left(\|\theta^{t-1}_{opt}-\theta_{opt}\|^4\right)=O(\frac{1}{t^2}) \tag{4.59}$$

ことを用いると，θ_{opt} のまわりで展開できて，

$$\begin{aligned}
&E_{\xi^{t-1},\xi}\left(f(x,y,\theta_{opt}+(\theta^{t-1}_{opt}-\theta_{opt}))\right)\\
&-\frac{1}{t}E_{\xi^{t-1},\xi}\Big(q^{ij}(\theta_{opt}+(\theta^{t-1}_{opt}-\theta_{opt}))\\
&\qquad \partial_i f(x,y,\theta_{opt}+(\theta^{t-1}_{opt}-\theta_{opt}))\partial_j d(x,y,\theta_{opt}+(\theta^{t-1}_{opt}-\theta_{opt}))\Big)\\
&+O(\frac{1}{t^2})\\
=\ & E_\xi\left(f(x,y,\theta_{opt})\right)\\
&+E_\xi\left(\partial_i f(x,y,\theta_{opt})\right)\ E_{\xi^{t-1}}\left((\theta^{t-1}_{opt}-\theta_{opt})^i\right)\\
&+\frac{1}{2}E_\xi\left(\partial_i\partial_j f(x,y,\theta_{opt})\right)\ E_{\xi^{t-1}}\left((\theta^{t-1}_{opt}-\theta_{opt})^i(\theta^{t-1}_{opt}-\theta_{opt})^j\right)\\
&-\frac{1}{t}E_\xi\left(q^{ij}(\theta_{opt})\partial_i f(x,y,\theta_{opt})\partial_j d(x,y,\theta_{opt})\right)\\
&\quad -\frac{1}{t}E_\xi\left(\partial_k q^{ij}(\theta_{opt})\partial_i f(x,y,\theta_{opt})\partial_j d(x,y,\theta_{opt})\right)\ E_{\xi^{t-1}}\left((\theta^{t-1}_{opt}-\theta_{opt})^k\right)
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{t}E_\xi\left(q^{ij}(\theta_{opt})\partial_i\partial_k f(x,y,\theta_{opt})\partial_j d(x,y,\theta_{opt})\right)\ E_{\xi^{t-1}}\left((\theta_{opt}^{t-1}-\theta_{opt})^k\right)\\
&-\frac{1}{t}E_\xi\left(q^{ij}(\theta_{opt})\partial_i f(x,y,\theta_{opt})\partial_j\partial_k d(x,y,\theta_{opt})\right)\ E_{\xi^{t-1}}\left((\theta_{opt}^{t-1}-\theta_{opt})^k\right)\\
&+O(\frac{1}{t^2})\\
=\ & E_\xi\left(f(x,y,\theta_{opt})\right)+\frac{1}{2t}E_\xi\left(\partial_i\partial_j f(x,y,\theta_{opt})\right)q^{ik}g_{kl}q^{lj}\\
&-\frac{1}{t}q^{ij}E_\xi\left(\partial_i f(x,y,\theta_{opt})\partial_j d(x,y,\theta_{opt})\right)+O(\frac{1}{t^2})
\end{aligned}
\tag{4.60}
$$

となる．上式において $f(x,y,\theta)$ を $d(x,y;\theta)$，$\partial_i\partial_j d(x,y;\theta)$ として式 (4.53) を整理すると，

$$
\begin{aligned}
&E_{\xi^t,g^t}\left(L_T(\theta^t)\right)\\
=\ & E_\xi\left(d(x,y;\theta_{opt})\right)+\frac{1}{2t}E_\xi\left(\partial_i\partial_j d(x,y,\theta_{opt})\right)q^{ik}g_{kl}q^{lj}\\
&-\frac{1}{t}q^{ij}E_\xi\left(\partial_i d(x,y,\theta_{opt})\partial_j d(x,y,\theta_{opt})\right)+\frac{1}{2}\varepsilon E_\xi\left(\partial_i\partial_j d(x,y;\theta_{opt})\right)v^{ij}\\
&+O(\frac{1}{t^2})+O(\varepsilon^{\frac{3}{2}})+O(\frac{\varepsilon}{t})\\
=\ & D(p,\theta_{opt})+\frac{1}{2t}q^{ij}g_{ij}-\frac{1}{t}q^{ij}g_{ij}+\frac{1}{2}\varepsilon q_{ij}v^{ij}+O(\frac{1}{t^2})+O(\varepsilon^{\frac{3}{2}})+O(\frac{\varepsilon}{t})\\
=\ & D(p,\theta_{opt})-\frac{1}{2t}\operatorname{tr}GQ^{-1}+\frac{1}{2}\varepsilon\operatorname{tr}QV+O(\frac{1}{t^2})+O(\varepsilon^{\frac{3}{2}})+O(\frac{\varepsilon}{t})
\end{aligned}
\tag{4.61}
$$

となり，定理が証明される． ∎

注．パラメタ θ^t の分布 $q^t(\theta^t|\xi^t)$ に関する 3 次のモーメントが 0 なら定理の証明において $O(\varepsilon^{\frac{3}{2}})$ は $O(\varepsilon^2)$ に置き換えられる．

学習損失の分散については次の定理が成り立つ．

定理 4.4 学習損失の分散は漸近的に

$$
V_{\xi^t,g^t}\left(L_T(\theta^t)\right)=\frac{1}{t}V_\xi\left(d(x,y;\theta_{opt})\right) \tag{4.62}
$$

に従う．

証明 期待値の証明と同じように，非逐次型学習では推定されるパラメタが例題の与えられる順序に依らないため，

$$
\begin{aligned}
&V_{\xi^t,g^t}\left(L_T(\theta^t)\right)\\
=\ & E_{\xi^t,g^t}\left(\left\{\frac{1}{t}\sum_{i=1}^{t}d(x_i,y_i;\theta^t)\right\}^2\right)-E_{\xi^t,g^t}\left(\frac{1}{t}\sum_{i=1}^{t}d(x_i,y_i;\theta^t)\right)^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{t^2} E_{\xi^t,\theta^t}\left(\left\{\sum_{i=1}^{t} d(x_i,y_i;\theta^t)^2 + \sum_{\substack{i,j=1\\ i\neq j}}^{t} d(x_i,y_i;\theta^t)d(x_j,y_j;\theta^t)\right\}^2\right)\\
&\quad -\frac{1}{t^2} E_{\xi^t,\theta^t}\left(\sum_{i=1}^{t} d(x_i,y_i;\theta^t)\right)^2\\
&= \frac{1}{t} E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)^2\right) + \frac{t-1}{t} E_{\xi^t,\theta^t}\left(d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right)\\
&\quad -E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)\right)^2\\
&= \frac{1}{t} E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)^2 - d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right)\\
&\quad +\left\{E_{\xi^t,\theta^t}\left(d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right) - E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)\right)^2\right\} \quad (4.63)
\end{aligned}
$$

としてよい．

上式第 1 の項は $1/t$ の 1 次の項となっているので，

$$
E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)^2 - d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right)
$$

では ε, $1/t$ に関して $O(1)$ になる項に注意すればよい．前定理の証明と同様に，まず θ^t_{opt} のまわりで各項を展開すると，

$$
\begin{aligned}
&E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)^2 - d(x_{t-1},y_{t-1};\theta^t)d(x_i,y_i;\theta^t)\right)\\
&= E_{\xi^t}\left(d(x_t,y_t;\theta^t_{opt})^2 - d(x_{t-1},y_{t-1};\theta^t_{opt})d(x_t,y_t;\theta^t_{opt})\right)\\
&\quad +2E_{\xi^t}\left(d(x_t,y_t;\theta^t_{opt})\partial_i d(x_t,y_t;\theta^t_{opt})\right)\ E_{\theta^t|\xi^t}\left((\theta^t-\theta^t_{opt})^i\right)\\
&\quad -2E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta^t_{opt})\partial_i d(x_t,y_t;\theta^t_{opt})\right)\ E_{\theta^t|\xi^t}\left((\theta^t-\theta^t_{opt})^i\right)\\
&\quad +O(\varepsilon)\\
&= E_{\xi^t}\left(d(x_t,y_t;\theta^t_{opt})^2 - d(x_{t-1},y_{t-1};\theta^t_{opt})d(x_t,y_t;\theta^t_{opt})\right) + O(\varepsilon) \qquad (4.64)
\end{aligned}
$$

となる．ただし，ここでも θ^t_{opt} が ξ_k の順序によらないことを用いている．次に θ^{t-1}_{opt} のまわりで損失関数を展開すると

$$
\begin{aligned}
&E_{\xi^t}\left(d(x_t,y_t;\theta^t_{opt})^2 - d(x_{t-1},y_{t-1};\theta^t_{opt})d(x_t,y_t;\theta^t_{opt})\right)\\
&= E_{\xi^t}\left(d(x_t,y_t;\theta^{t-1}_{opt})^2 - d(x_{t-1},y_{t-1};\theta^{t-1}_{opt})d(x_t,y_t;\theta^{t-1}_{opt})\right) + O(\frac{1}{t}) \quad (4.65)
\end{aligned}
$$

となり，さらに θ^{t-2}_{opt} のまわりで展開すると

$$
\begin{aligned}
&E_{\xi^t}\left(d(x_t,y_t;\theta^{t-1}_{opt})^2 - d(x_{t-1},y_{t-1};\theta^{t-1}_{opt})d(x_t,y_t;\theta^{t-1}_{opt})\right)\\
&= E_{\xi^{t-2},\xi_{t-1},\xi_t}\left(d(x_t,y_t;\theta^{t-2}_{opt})^2 - d(x_{t-1},y_{t-1};\theta^{t-2}_{opt})d(x_t,y_t;\theta^{t-2}_{opt})\right) + O(\frac{1}{t})\\
&= E_{\xi^{t-2}}\left(V_\xi\left(d(x,y;\theta^{t-2}_{opt})\right)\right) + O(\frac{1}{t}) \qquad (4.66)
\end{aligned}
$$

となる．最後に θ_{opt} のまわりで展開して

$$\begin{aligned} & E_{\xi^{t-2}}\left(V_\xi\left(d(x,y;\theta_{opt}^{t-2})\right)\right) \\ &= V_\xi\left(d(x,y;\theta_{opt})\right) + \partial_i V_\xi\left(d(x,y;\theta_{opt})\right)\ E_{\xi^{t-2}}\left((\theta_{opt}^{t-2}-\theta_{opt})^i\right) \\ &\quad +O(\frac{1}{t}) \\ &= V_\xi\left(d(x,y;\theta_{opt})\right) + O(\frac{1}{t}) \end{aligned} \tag{4.67}$$

を得る．よって第 1 項は

$$\begin{aligned} & \frac{1}{t}E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)^2 - d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right) \\ &= \frac{1}{t}V_\xi\left(d(x,y;\theta_{opt})\right) + O(\frac{1}{t^2}) + O(\frac{\varepsilon}{t}) \end{aligned} \tag{4.68}$$

となる．

同様にして第 2 項も求めることができる．ε および $1/t$ に関して 1 次以下の項に注意して

$$E_{\xi^t,\theta^t}\left(d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right) \tag{4.69}$$

を展開していく．まず θ_{opt}^t のまわりで展開する．

$$\begin{aligned} & E_{\xi^t,\theta^t}\left(d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right) \\ &= E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^t)d(x_t,y_t;\theta_{opt}^t)\right) \\ &\quad +2E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^t)\partial_i d(x_t,y_t;\theta_{opt}^t)\ E_{\theta^t|\xi^t}\left((\theta^t-\theta_{opt}^t)^i\right)\right) \\ &\quad +E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^t)\partial_i\partial_j d(x_t,y_t;\theta_{opt}^t)\ E_{\theta^t|\xi^t}\left((\theta^t-\theta_{opt}^t)^i(\theta^t-\theta_{opt}^t)^j\right)\right) \\ &\quad +E_{\xi^t}\left(\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^t)\partial_j d(x_t,y_t;\theta_{opt}^t)\ E_{\theta^t|\xi^t}\left((\theta^t-\theta_{opt}^t)^i(\theta^t-\theta_{opt}^t)^j\right)\right) \\ &\quad +O(\varepsilon^{\frac{3}{2}}) \\ &= E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^t)d(x_t,y_t;\theta_{opt}^t)\right) \\ &\quad +\varepsilon E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^t)\partial_i\partial_j d(x_t,y_t;\theta_{opt}^t)v^{ij}(\theta_{opt}^t)\right) \\ &\quad +\varepsilon E_{\xi^t}\left(\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^t)\partial_j d(x_t,y_t;\theta_{opt}^t)v^{ij}(\theta_{opt}^t)\right) \\ &\quad +O(\varepsilon^{\frac{3}{2}}) \end{aligned} \tag{4.70}$$

つぎに各項を θ_{opt}^{t-1}, θ_{opt}^{t-2}, θ_{opt} のまわりで順次展開する．

$$E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^t)d(x_t,y_t;\theta_{opt}^t)\right)$$

$$
\begin{aligned}
&= E_{\xi^{t-1}\xi_t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^{t-1})d(x_t,y_t;\theta_{opt}^{t-1})\right)\\
&\quad -\frac{1}{t}E_{\xi^{t-1}\xi_t}\Big(q^{ij}(\theta_{opt}^{t-1})\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^{t-1})d(x_t,y_t;\theta_{opt}^{t-1})\partial_j d(x_t,y_t;\theta_{opt}^{t-1})\\
&\qquad -q^{ij}(\theta_{opt}^{t-1})d(x_{t-1},y_{t-1};\theta_{opt}^{t-1})\partial_i d(x_t,y_t;\theta_{opt}^{t-1})\partial_j d(x_t,y_t;\theta_{opt}^{t-1})\Big)\\
&\quad +O(\frac{1}{t^2})\\
&= E_{\xi^{t-2}\xi_{t-1}\xi_t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})d(x_t,y_t;\theta_{opt}^{t-2})\right)\\
&\quad -\frac{1}{t}E_{\xi^{t-2}\xi_{t-1}\xi_t}\Big(q^{ij}(\theta_{opt}^{t-2})\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})d(x_t,y_t;\theta_{opt}^{t-2})\partial_j d(x_t,y_t;\theta_{opt}^{t-2})\\
&\qquad -q^{ij}(\theta_{opt}^{t-2})d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})\partial_i d(x_t,y_t;\theta_{opt}^{t-2})\partial_j d(x_t,y_t;\theta_{opt}^{t-2})\\
&\qquad -q^{ij}(\theta_{opt}^{t-2})\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})d(x_t,y_t;\theta_{opt}^{t-2})\partial_j d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})\\
&\qquad -q^{ij}(\theta_{opt}^{t-2})d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})\partial_i d(x_t,y_t;\theta_{opt}^{t-2})\partial_j d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})\Big)\\
&\quad +O(\frac{1}{t^2})\\
&= E_{\xi^{t-2}}\left(E_\xi\left(d(x,y;\theta_{opt}^{t-2})\right)^2\right)\\
&\quad -\frac{2}{t}E_{\xi^{t-2}}\left(q^{ij}(\theta_{opt}^{t-2})g_{ij}(\theta_{opt}^{t-2})E_\xi\left(d(x,y;\theta_{opt}^{t-2})\right)\right)\\
&\quad +O(\frac{1}{t^2})\\
&= E_\xi\left(d(x,y;\theta_{opt})\right)^2\\
&\quad +2E_\xi\left(d(x,y;\theta_{opt})\right)E_\xi\left(\partial_i d(x,y;\theta_{opt})\right)E_{\xi^{t-2}}\left((\theta_{opt}^{t-2}-\theta_{opt})^i\right)\\
&\quad +E_\xi\left(d(x,y;\theta_{opt})\right)E_\xi\left(\partial_i\partial_j d(x,y;\theta_{opt})\right)E_{\xi^{t-2}}\left((\theta_{opt}^{t-2}-\theta_{opt})^i(\theta_{opt}^{t-2}-\theta_{opt})^j\right)\\
&\quad +E_\xi\left(\partial_i d(x,y;\theta_{opt})\right)E_\xi\left(\partial_j d(x,y;\theta_{opt})\right)E_{\xi^{t-2}}\left((\theta_{opt}^{t-2}-\theta_{opt})^i(\theta_{opt}^{t-2}-\theta_{opt})^j\right)\\
&\quad -\frac{2}{t}q^{ij}(\theta_{opt})g_{ij}(\theta_{opt})E_\xi\left(d(x,y;\theta_{opt})\right)\\
&\quad +O(\frac{1}{t^2})\\
&= E_\xi\left(d(x,y;\theta_{opt})\right)^2\\
&\quad +\frac{1}{t}E_\xi\left(d(x,y;\theta_{opt})\right)q_{ij}q^{ik}q^{lj}g_{kl}\\
&\quad -\frac{2}{t}q^{ij}(\theta_{opt})g_{ij}(\theta_{opt})E_\xi\left(d(x,y;\theta_{opt})\right)\\
&\quad +O(\frac{1}{t^2})\\
&= D(p,\theta_{opt})^2-\frac{1}{t}D(p,\theta_{opt})q^{ij}(\theta_{opt})g_{ij}(\theta_{opt})+O(\frac{1}{t^2}) \qquad (4.71)
\end{aligned}
$$

$$
\begin{aligned}
&\varepsilon E_{\xi^t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^t)\partial_i\partial_j d(x_t,y_t;\theta_{opt}^t)v^{ij}(\theta_{opt}^t)\right)\\
&\quad = \varepsilon E_{\xi^{t-1}\xi_t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^{t-1})\partial_i\partial_j d(x_t,y_t;\theta_{opt}^{t-1}v^{ij}(\theta_{opt}^{t-1})\right)+O(\frac{\varepsilon}{t})
\end{aligned}
$$

$$
\begin{aligned}
&= \varepsilon E_{\xi^{t-2},\xi_{t-1},\xi_t}\left(d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})\partial_i\partial_j d(x_t,y_t;\theta_{opt}^{t-2})v^{ij}(\theta_{opt}^{t-2})\right)+O(\frac{\varepsilon}{t})\\
&= \varepsilon E_{\xi_{t-1},\xi_t}\left(d(x_{t-1},y_{t-1};\theta_{opt})\partial_i\partial_j d(x_t,y_t;\theta_{opt})v^{ij}(\theta_{opt})\right)+O(\frac{\varepsilon}{t})\\
&= \varepsilon E_{\xi}\left(d(x,y;\theta_{opt})\right)E_{\xi}\left(\partial_i\partial_j d(x,y;\theta_{opt})\right)v^{ij}(\theta_{opt})+O(\frac{\varepsilon}{t})\\
&= \varepsilon D(p,\theta_{opt})q_{ij}v^{ij}+O(\frac{\varepsilon}{t}) \qquad (4.72)
\end{aligned}
$$

$$
\begin{aligned}
&\varepsilon E_{\xi^t}\left(\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^{t})\partial_j d(x_t,y_t;\theta_{opt}^{t})v^{ij}(\theta_{opt}^{t})\right)\\
&= \varepsilon E_{\xi^{t-1},\xi_t}\left(\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^{t-1})\partial_j d(x_t,y_t;\theta_{opt}^{t-1})v^{ij}(\theta_{opt}^{t-1})\right)+O(\frac{\varepsilon}{t})\\
&= \varepsilon E_{\xi^{t-2},\xi_{t-1},\xi_t}\left(\partial_i d(x_{t-1},y_{t-1};\theta_{opt}^{t-2})\partial_j d(x_t,y_t;\theta_{opt}^{t-2})v^{ij}(\theta_{opt}^{t-2})\right)+O(\frac{\varepsilon}{t})\\
&= \varepsilon E_{\xi_{t-1},\xi_t}\left(\partial_i d(x_{t-1},y_{t-1};\theta_{opt})\partial_j d(x_t,y_t;\theta_{opt})v^{ij}(\theta_{opt})\right)+O(\frac{\varepsilon}{t})\\
&= \varepsilon E_{\xi}\left(\partial_i d(x,y;\theta_{opt})\right)^2 v^{ij}(\theta_{opt})+O(\frac{\varepsilon}{t})\\
&= 0+O(\frac{\varepsilon}{t}) \qquad (4.73)
\end{aligned}
$$

一方，前定理より

$$
\begin{aligned}
&E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)\right)^2\\
&= D(p,\theta_{opt})^2-\frac{1}{t}D(p,\theta_{opt})g_{ij}q^{ij}+\varepsilon D(p,\theta_{opt})q_{ij}v^{ij}\\
&\quad +O(\frac{1}{t^2})+O(\frac{\varepsilon}{t})+O(\varepsilon^{\frac{3}{2}}) \qquad (4.74)
\end{aligned}
$$

これらをまとめると第 2 項は

$$
\begin{aligned}
E_{\xi^t,\theta^t}\left(d(x_{t-1},y_{t-1};\theta^t)d(x_t,y_t;\theta^t)\right) &- E_{\xi^t,\theta^t}\left(d(x_t,y_t;\theta^t)\right)^2\\
&= O(\frac{1}{t^2})+O(\frac{\varepsilon}{t})+O(\varepsilon^{\frac{3}{2}}) \qquad (4.75)
\end{aligned}
$$

となる．

以上の結果をまとめて

$$
\begin{aligned}
&V_{\xi^t,\theta^t}\left(L_T(\theta^t)\right)\\
&= \frac{1}{t}V_{\xi}\left(d(x,y;\theta_{opt})\right)+O(\frac{1}{t^2})+O(\frac{\varepsilon}{t})+O(\varepsilon^{\frac{3}{2}}) \qquad (4.76)
\end{aligned}
$$

が得られ，定理が証明される．　∎

注．学習損失の期待値に関する証明と同様にパラメタ θ^t の分布 $q^t(\theta^t|\xi^t)$ に関する 3 次のモーメントが 0 なら $O(\varepsilon^{\frac{3}{2}})$ は $O(\varepsilon^2)$ に置き換えられる．

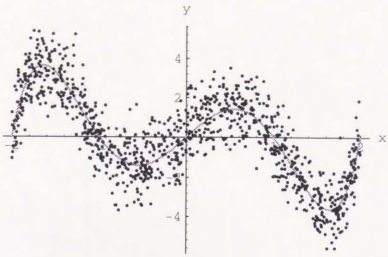


図 4.1: システムから観測した例題.

V は Q, G と C （定数行列） で記述されるため，予測損失，学習損失の特性を決定するのは Q, G という 2 つの量と C, ε という 2 つの学習パラメタであることがわかる．またもし ε を十分に小さく選んでやることができるのなら，Q, G の 2 つの量のみで予測損失，学習損失の特性は完全に記述される．

例 6 (計算機シミュレーション 2.)

入力 x が 1 次元，出力 y が 1 次元で

$$p(x) \quad : \quad [-2, 2] \text{上で一様分布} \tag{4.77}$$

$$p(y|x) \quad : \quad y = (x-2)(x-1)x(x+1)(x+2) + \eta, \quad \eta \sim N(0, 1.0^2) \tag{4.78}$$

であるようなシステムを，確定的モデル

$$M_d = \{\theta^1 x^5 + \theta^2 x^4 + \theta^3 x^3 + \theta^4 x^2 + \theta^5 x + \theta^6; | \, \theta = (\theta^i) \in R^6\} \tag{4.79}$$

で近似する問題を考える．損失関数として

$$d(x, y; \theta) = \frac{1}{2}(y - \theta^1 x^5 + \theta^2 x^4 + \theta^3 x^3 + \theta^4 x^2 + \theta^5 x + \theta^6)^2 \tag{4.80}$$

を用いる．t 個の例題を与えた場合，この問題では連立方程式を解くことによって経験分布に対する最適パラメタを容易に求めることができる．このため学習による推定パラ

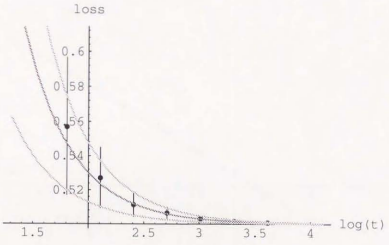


図 4.2: 予測損失の平均値と分散.

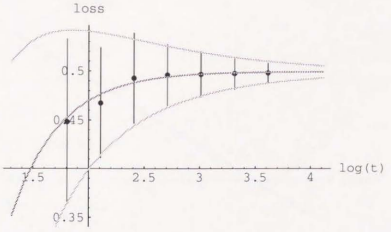


図 4.3: 学習損失の平均値と分散.

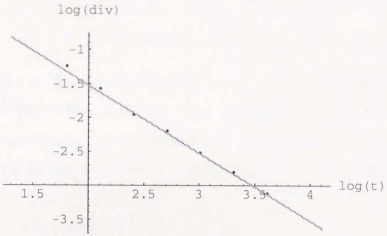


図 4.4: 予測損失の平均値の偏差.

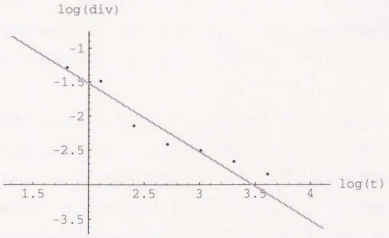


図 4.5: 学習損失の平均値の偏差.

メタの分散は計算誤差に由来するものだけとなり，ほとんど 0 である．以下では

$$\varepsilon = 0 \tag{4.81}$$

として理論値を計算している．

図 4.1 にシステムから観測したの 1000 個の例題を示す．実線は関数

$$y = (x-2)(x-1)x(x+1)(x+2) \tag{4.82}$$

を表す．図 4.2，図 4.3 は異なる例題を与え学習させた 128 台の機械の予測損失と学習損失の平均値と分散をそれぞれ表している．例題数は 64, 128, 256, 512, 1024, 2048, 4096 の 6 通りである．また実線は理論期待値とその分散を表している．横軸には例題数の対数を縦軸には損失をとっている．例題数の大きいところでは実験値と理論値が非常に良く一致していることがわかる．

図 4.4，図 4.5 は偏差をそれぞれ

$$\text{div} = E_{\xi^t}\left(L_P(\theta^t)\right) - \lim_{t\to\infty} E_{\xi^t}\left(L_P(\theta^t)\right) \tag{4.83}$$

$$\text{div} = \lim_{t\to\infty} E_{\xi^t}\left(L_T(\theta^t)\right) - E_{\xi^t}\left(L_T(\theta^t)\right) \tag{4.84}$$

で定義し，予測損失，学習損失それぞれについて表示したものである．横軸には例題数の対数を，縦軸には偏差の対数をとっている．また実線は理論値を表している．

予測損失については理論と非常に良く一致している．学習損失は理論からも予測される通り分散が非常に大きいため，機械ごとのばらつきが大きく実験値は理論値のまわりにゆらいでいるが，傾向としては非常に良くあっている．

例 7 (計算機シミュレーション 3.) ここでは計算機シュミレーション 1. のシステムとモデルおよび損失関数を用いて，非逐次型学習を行なった場合の計算機シミュレーションの結果を示す．学習は t 個の例題の中から一様ランダムに例題を提示する確率的降下法による確率型とした．学習の条件は計算機シュミレーション 1. と同様である．ただしパラメタの初期値には真の機械のパラメタを与え，学習回数は 6000 回とした．

図 4.6，図 4.7 は異なる例題を与え学習させた 100 台の機械の予測損失と学習損失の平均値と分散をそれぞれ表している．例題数は 200, 400, 800, 1600 の 4 通りである．また実線は理論期待値とその分散を表している．横軸には例題数の対数を縦軸には損失をとっている．予測損失は実験値と理論値が良く一致しているが，学習損失は分散

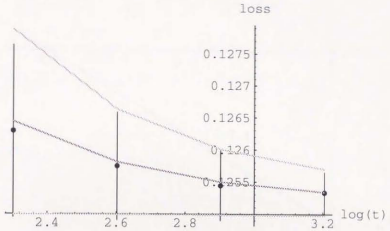


図 4.6: 予測損失の平均値と分散.

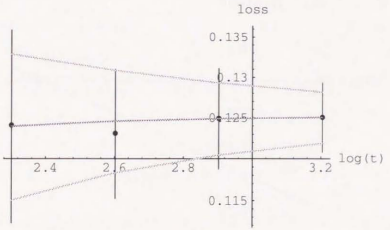


図 4.7: 学習損失の平均値と分散.

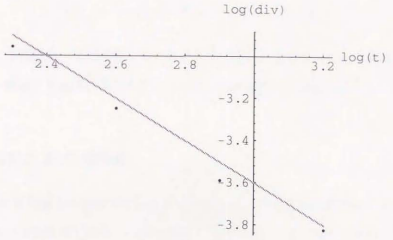


図 4.8: 予測損失の平均値の偏差.

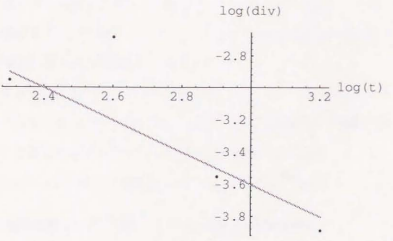


図 4.9: 学習損失の平均値の偏差.

が非常に大きく，平均をとる機械の数が少ないためため実験値と理論値はあまり一致していない．

図 4.8，図 4.9 は計算機シミュレーション 2. と同様に，偏差をそれぞれ

$$\text{div} = E_{\xi^t}\left(L_P(\theta^t)\right) - \lim_{t\to\infty} E_{\xi^t}\left(L_P(\theta^t)\right) \tag{4.85}$$

$$\text{div} = \lim_{t\to\infty} E_{\xi^t}\left(L_T(\theta^t)\right) - E_{\xi^t}\left(L_T(\theta^t)\right) \tag{4.86}$$

で定義し，予測損失，学習損失それぞれについて表示したものである．横軸には例題数の対数を，縦軸には偏差の対数をとっている．また実線は理論値を表している．

4.4 学習系の AIC 規準量

有限個の観測値から統計モデルを推定する場合，モデルの次数を高くするにしたがい，モデルの自由度が上がり，一般に表現力は高くなる．このため，推定に用いた観測値を表現する際の誤差は減少するが，与えられた観測値以外に対しては誤差が減少するとは限らず，むしろ増大する場合がある．これは，真の分布を観測による経験分布に置き換えたために推定パラメタにずれが生じ，モデルの次数が高いほどこのずれに対する感度が鋭敏になるための悪影響である．Akaike [1] は最尤推定を行なう際に生じる推定パラメタのずれを平均的に評価し，モデルの次数=パラメタを増やしたことによる影響が対数尤度に与える損失を導き，最適なパラメタ数の規準を求めた．これが情報規準量 AIC と呼ばれるものである．ここではこの考えを平均損失関数に拡張し，損失関数を用いた学習系における情報規準量を導出する．

経験分布 $p^t(x,y)$ に対してできる限り最適となるように求めたパラメタ θ^t を実際のシステムの分布 $p(x,y)$ の中で使うときに起きる損失 $D(p,\theta^t)$ を経験分布に対する損失 $D(p^t,\theta^t)$ で推定するのがそもそもの狙いである (図 4.10)．

前述したいくつかの定理と補題を用いると次の定理が証明できる．

定理 4.5 t 個の例題 ξ^t から学習した θ^t の平均損失の期待値は

$$E_{\xi^t,\theta^t}\left(D(p,\theta^t)\right) = E_{\xi^t,\theta^t}\left(D(p^t,\theta^t)\right) + \frac{1}{t}\,\mathrm{tr}\,GQ^{-1} \tag{4.87}$$

で与えられる．

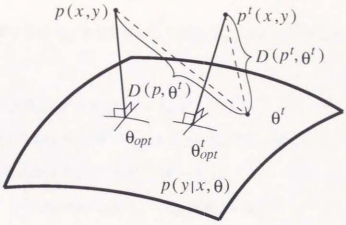


図 4.10: システムの分布と経験分布およびパラメタの幾何学的関係

証明　まず，平均損失関数 $D(p,\theta^t)$ を最適パラメタ θ_{opt} のまわりで展開する．

$$\begin{aligned} & D(p,\theta^t) \\ &= D(p,\theta_{opt}) + \partial_i D(p,\theta_{opt})(\theta^t-\theta_{opt})^i \\ &\quad +\frac{1}{2}\partial_i\partial_j D(p,\theta_{opt})(\theta^t-\theta_{opt})^i(\theta^t-\theta_{opt})^j + O(\|\theta^t-\theta_{opt}\|^3) \end{aligned} \quad (4.88)$$

以下では式 (4.88) の展開の各項を詳しく見ていく．まず，第 1 項は

$$\begin{aligned} & D(p,\theta_{opt}) \\ &= D(p,\theta_{opt}) - D(p^t,\theta_{opt}) + D(p^t,\theta_{opt}) \\ &\quad -D(p^t,\theta^t_{opt}) + D(p^t,\theta^t_{opt}) - D(p^t,\theta^t) + D(p^t,\theta^t) \\ &= D(p^t,\theta^t) + \{D(p,\theta_{opt}) - D(p^t,\theta_{opt})\} \\ &\quad +\{D(p^t,\theta^t_{opt}+(\theta_{opt}-\theta^t_{opt})) - D(p^t,\theta^t_{opt})\} \\ &\quad -\{D(p^t,\theta^t_{opt}-(\theta^t_{opt}-\theta^t)) - D(p^t,\theta^t_{opt})\} \\ &= D(p^t,\theta^t) + \{D(p,\theta_{opt}) - D(p^t,\theta_{opt})\} \\ &\quad +\frac{1}{2}\partial_i\partial_j D(p^t,\theta^t_{opt})(\theta_{opt}-\theta^t_{opt})^i(\theta_{opt}-\theta^t_{opt})^j \\ &\quad -\frac{1}{2}\partial_i\partial_j D(p^t,\theta^t_{opt})(\theta^t_{opt}-\theta^t)^i(\theta^t_{opt}-\theta^t)^j \\ &\quad +O(\|\theta_{opt}-\theta^t_{opt}\|^3) + O(\|\theta^t_{opt}-\theta^t\|^3) \\ &= D(p^t,\theta^t) + \{D(p,\theta_{opt}) - D(p^t,\theta_{opt})\} \\ &\quad +\frac{1}{2}\partial_i\partial_j D(p^t,\theta_{opt})\{(\theta_{opt}-\theta^t_{opt})^i(\theta_{opt}-\theta^t_{opt})^j - (\theta^t_{opt}-\theta^t)^i(\theta^t_{opt}-\theta^t)^j\} \end{aligned}$$

$$+O((\|\theta_{opt}-\theta^t_{opt}\|+\|\theta^t_{opt}-\theta^t\|)^3) \tag{4.89}$$

となる．次に第 2 項は θ_{opt} の最適性から $\nabla D(p,\theta_{opt})=0$ より 0 となる．最後に第 3 項は

$$\begin{aligned}
&\frac{1}{2}\partial_i\partial_j D(p,\theta_{opt})(\theta^t-\theta_{opt})^i(\theta^t-\theta_{opt})^j\\
&= \frac{1}{2}\partial_i\partial_j D(p,\theta_{opt})(\theta^t-\theta^t_{opt}+\theta^t_{opt}-\theta_{opt})^i(\theta^t-\theta^t_{opt}+\theta^t_{opt}-\theta_{opt})^j\\
&= \frac{1}{2}\partial_i\partial_j D(p,\theta_{opt})(\theta^t-\theta^t_{opt})^i(\theta^t-\theta^t_{opt})^j\\
&\quad +\frac{1}{2}\partial_i\partial_j D(p,\theta_{opt})(\theta^t_{opt}-\theta_{opt})^i(\theta^t_{opt}-\theta_{opt})^j\\
&\quad +\partial_i\partial_j D(p,\theta_{opt})(\theta^t-\theta^t_{opt})^i(\theta^t_{opt}-\theta_{opt})^j\\
&= \frac{1}{2}\,\mathrm{tr}\left[Q\{(\theta^t-\theta^t_{opt})(\theta^t-\theta^t_{opt})^{\mathrm{T}}+(\theta^t_{opt}-\theta_{opt})(\theta^t_{opt}-\theta_{opt})^{\mathrm{T}}\right.\\
&\quad \left.+2(\theta^t_{opt}-\theta_{opt})(\theta^t-\theta^t_{opt})^{\mathrm{T}}\}\right]
\end{aligned} \tag{4.90}$$

となる．大数の法則により，

$$\nabla\nabla D(p^t,\theta_{opt})\simeq Q \tag{4.91}$$

となることと，

$$M(\theta^t_{opt})=D(p,\theta_{opt})-D(p^t,\theta_{opt}) \tag{4.92}$$

と置いたとき

$$E_{\xi^t}\left(M(\theta^t_{opt})\right)=0 \tag{4.93}$$

となることを用いると，式 (4.88) は

$$\begin{aligned}
&D(p,\theta^t)\\
&= D(p^t,\theta^t)\\
&\quad +\mathrm{tr}\left[Q\{(\theta_{opt}-\theta^t_{opt})(\theta_{opt}-\theta^t_{opt})^{\mathrm{T}}+(\theta^t_{opt}-\theta_{opt})(\theta^t-\theta^t_{opt})^{\mathrm{T}}\}\right]\\
&\quad +O((\|\theta_{opt}-\theta^t\|+\|\theta^t_{opt}-\theta^t\|)^3)
\end{aligned} \tag{4.94}$$

となるが，θ^t と ξ^t に関して平均をとると

$$E_{\xi^t}\left((\theta^t_{opt}-\theta_{opt})E_{\theta^t|\xi^t}\left((\theta^t-\theta^t_{opt})^{\mathrm{T}}\right)\right)=0 \tag{4.95}$$

であるから，高次の項を無視して

$$E\left(D(p,\theta^t)\right) = E\left(D(p^t,\theta^t)\right) + \frac{1}{t}\,\mathrm{tr}\,QQ^{-1}GQ^{-1} \tag{4.96}$$

となり，定理の証明を終る．∎

この結果は予測損失と学習損失の関係からも説明することができる．前節で求めた予測損失と学習損失の期待値の関係から，t 個の例題から求められたパラメタの学習損失を用いて予測損失を推定することができる．

$$E\left(L_P(\theta^t)\right) = E\left(L_T(\theta^t)\right) + \frac{1}{t}\,\mathrm{tr}\,GQ^{-1} \tag{4.97}$$

となり，これは上の定理の結果と一致する．すなわち，AIC 規準は予測損失を最小にする規準であることがわかる．

上の定理は，t 個の例題からなるさまざまな例題の組 ξ^t_λ, $\lambda \in \Lambda$ があって，それぞれの組に対してパラメタ θ^t_λ が得られたときの平均評価である．しかし実際の応用の場面では，例題の組はひとつだけしか与えられず，モデルをいくつか構成して最も良さそうなモデルを選ぶような場合が多い．具体的な定理の適用法を次の系によって示す．まず部分モデルを定義する．

定義 4.5 *2*つのモデル M_1, M_2 があって，包含関係

$$M_1 \subset M_2 \tag{4.98}$$

を満たすとき，M_1 を M_2 の部分モデルと呼ぶ．

上の定理から包含関係が成り立つモデルとその部分モデルの間の損失に関して次の系が導ける (図 4.11)．

系 4.6 モデル $M_1 = \{\theta_1\}$ がモデル $M_2 = \{\theta_2\}$ の部分モデルとする．同一の t 個の例題 ξ^t を用いた逐次型学習により，それぞれパラメタ θ^t_i, $i = 1,2$ を得たとする．このとき，次式で与えられる量

$$D(p^t,\theta^t_i) + \frac{1}{t}\,\mathrm{tr}\,G^t_i Q^{t-1}_i \quad i = 1,2 \tag{4.99}$$

を最小にするモデルが平均損失の意味で最適なモデルとなる．ただし，

$$G^t_i = \frac{1}{t}\sum_{j=1}^{t} \nabla d_i(x_j,y_j,\theta^t_i)\nabla d_i(x_j,y_j,\theta^t_i)^{\mathrm{T}}$$

$$Q^t_i = \frac{1}{t}\sum_{j=1}^{t} \nabla\nabla d(x_j,y_j,\theta^t_i)$$

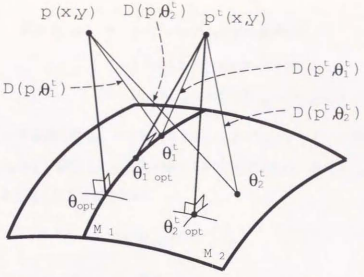


図 4.11: 包含関係にある 2 つのモデルと平均損失関数の関係

であり，$d_i(x,y;\theta_i)$ はそれぞれモデル M_i に対する損失関数で，θ_1 を M_2 の座標で見たとき $\theta_2(\theta_1)$ と表すとすれば

$$d_1(x,y;\theta_1) = d_2(x,y;\theta_2(\theta_1)) \tag{4.100}$$

を満たすものとする．

この系は上の定理において平均評価をとる前の量に対して，行列 G, Q を与えられた例題 ξ^t と学習の結果得たパラメタ θ_i^t で推定した行列 G_i^t, Q_i^t に置き換えることによって得られる．

注. 定理 4.5 の証明において，式 (4.92) の項は ξ^t の分布に関して平均したとき 0 となり定理には現れなかったが，平均を操作を行なう前には残る項である．システムの分布 $p(x,y)$ は未知であるので実際には求められない項であるが，部分モデルとの比較を行なう場合には次のような理由から評価する必要がなくなる．

まずモデル M_1 の最適パラメタ $\theta_{1\ opt}$ とモデル M_2 の最適パラメタ $\theta_{2\ opt}$ が図 4.11 のように一致しているときには明らかである．次に二つのパラメタが一致していない場合を考える．経験分布に対するモデル M_2 の最適パラメタ $\theta_{2\ opt}^t$ は M_2 の最適パラメタ $\theta_{2\ opt}$ を中心とした半径 $O(1/\sqrt{t})$ 程度のところに分布している．M_1 の最適パラメタ $\theta_{1\ opt}$ が $\theta_{2\ opt}$ から $O(1/\sqrt{t})$ より離れていればモデル M_2 を採択すべきであろう．

このとき M_1、M_2 の座標を区別せずに書くと

$$\begin{aligned} D(p,\theta_{1\ opt}) &= \int d_1(x,y,\theta_{1\ opt})p(x,y)dxdy \\ &= \int d_2(x,y,\theta_{1\ opt})p(x,y)dxdy \\ &= D(p,\theta_{2\ opt}) + O(\|\theta_{1\ opt} - \theta_{2\ opt}\|^2) \end{aligned} \tag{4.101}$$

となり，M_2 の平均損失が M_1 のそれより $O(1/t)$ より低いオーダーで増加する．よって系の規準量は M_2 を採択することを示唆するので問題はない．逆に $\theta_{1\ opt}$ が $\theta_{2\ opt}$ から $O(1/\sqrt{t})$ より近くにある場合には，

$$|M(\theta_{opt})| = O\left(\frac{1}{\sqrt{t}}\right) \tag{4.102}$$

が中心極限定理からいえることに注意すると $|M(\theta_{1\ opt}) - M(\theta_{2\ opt})|$ は $O(1/t)$ より高次のオーダーとなり，規準量には影響しないことがわかる．

ただし $\theta_{1\ opt}$ と $\theta_{2\ opt}$ の差がちょうど $O(1/\sqrt{t})$ となるときには，モデルの能力は与えられた t 個の例題で比較した場合にはほぼ等しくなり，この規準量では決められなくなる．

この系によって包含関係にあるモデル間の能力の差を具体的に数量化することができる．例えば Multi Layer Network における中間層の素子の数を増減したいとき，系 4.6 に与えた量を計算して素子数を変化させればよい．

例 8 (計算機シミュレーション 4.) 入力 $x = (x_1, x_2)$ が 2 次元，出力 y が 1 次元で

$$p(x) \ : \ [-5,5]\times[-5,5] \text{上で一様分布} \tag{4.103}$$

$$p(y|x) \ : \ y = f(x;\theta) + \eta, \quad \eta \sim N(0, 1.0^2) \tag{4.104}$$

であるシステムを考える．ただし $f(x;\theta)$ は例 1 の式 (2.16) の Multi Layer Network である．確定的モデルを

$$\begin{aligned} M_d &= \{f_n(x;\theta)\} \\ f_n(x;\theta) &= \tanh\left(\sum_{j=1}^{n} W_j^2 \tanh\left(\sum_{i=1}^{2} W_{ji}^1 x_i + h_j^1\right) + h^2\right), \end{aligned} \tag{4.105}$$

とし，損失関数

$$d_n(x,y;\theta) = \frac{1}{2}(y - f_n(x;\theta))^2 \tag{4.106}$$

の意味で最適な近似を求める問題を考える．

中間素子の個数 n を 2, 3, 4, 5, 6 と変え，1000 個の固定した例題の中から一様ランダムに例題を 50000 回提示する確率的降下法により学習を行なった．学習に用いたパラメタは

$$C = E, \quad \varepsilon = 0.01$$

で，初期パラメタは，W_{ji}^1, h_j^1 は $[-2, 2]$ から，W_j^2, h^2 は $[-1, 1]$ から一様ランダムに選んだ．

図 4.12 は学習終了後の各機械の出力を表示したものである．(a) は真の機械，(b), (c), (d), (e), (f) はそれぞれ中間素子の個数が 2, 3, 4, 5, 6 の場合である．

ここで用いた損失関数では，加法的雑音が入力よらず一定の場合に限り，その分散 σ^2 を用いて

$$G_n = \sigma^2 Q_n \tag{4.107}$$

となることが容易に計算される．ただし，添字 n は中間素子の数を表すとする．このとき系 4.6 に与えた規準量

$$D(p^t, \theta_n^t) + \frac{1}{t} \operatorname{tr} G_n^t Q_n^{t\,-1} \tag{4.108}$$

は次のようにして G_n^t, Q_n^t を直接求めることなく計算できる．

$$m = 4n + 1 \tag{4.109}$$

$$\begin{aligned} & D(p^t, \theta_n^t) + \frac{1}{t} \operatorname{tr} G_n^t Q_n^{t\,-1} \\ &= D(p^t, \theta_n^t) + \frac{m}{t}\sigma^2 \\ &= D(p^t, \theta_n^t) + \frac{2m}{t} D(p^t, \theta_{n\ opt}^t) \\ &= D(p^t, \theta_n^t) + \frac{2m}{t} D(p^t, \theta_{n\ opt}^t) \\ &= D(p^t, \theta_n^t) + \frac{2m}{t} D(p^t, \theta_n^t) + O\left(\frac{1}{t^2}\right) \\ &= D(p^t, \theta_n^t)\left(1 + \frac{2m}{t}\right) + O\left(\frac{1}{t^2}\right) \end{aligned} \tag{4.110}$$

図 4.13 は規準量と中間層の素子数の関係を示したものである．横軸には中間素子の個数を，縦軸には規準量をとっている．この結果中間素子数が 4 個のモデルが選択されるが，これは真の機械 (システム) の構造と一致している．

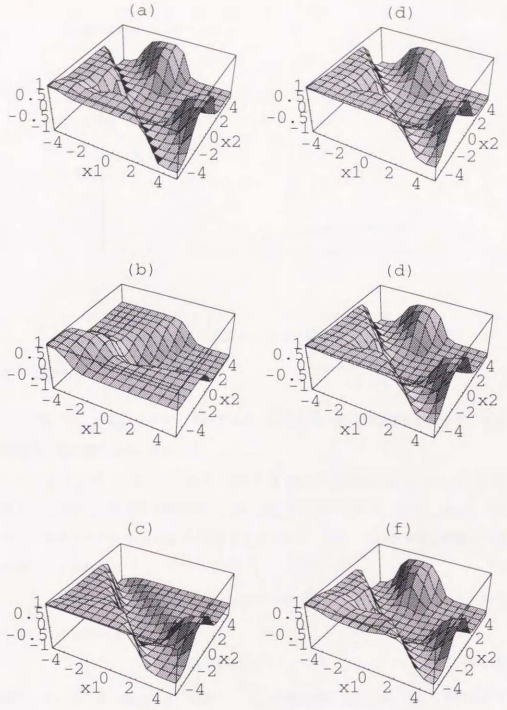


図 4.12: 学習後の入出力関係. (a) 最適な機械, (b) 中間素子数 $n = 2$ の機械,(c) $n = 3$,(d) $n = 4$,(e) $n = 5$,(f) $n = 6$.

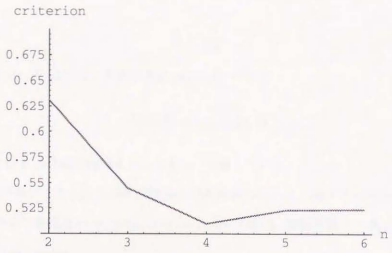


図 4.13: 中間素子の個数と規準量.

4.5 部分モデルにおける学習

この節では，忠実なモデルと一致性を持つ損失関数の場合に議論を限定して，その幾何学的な構造を考えることにする．

システムがモデルと同じパラメタで記述できる場合を考える．すなわち，モデルをパラメタ θ で表したとき，システムの入出力関係は θ をパラメタとして $p(y|x,\theta)$ で表され，システムの分布は $p(x,y,\theta)$ で表されるとする．また，このとき損失関数は一致性を持つ，すなわち

$$D(p(\theta),\theta) = \min_{\theta'} \{D(p(\theta),\theta')\} \tag{4.111}$$

あるいはこれを微分して

$$E_\theta\left(\partial_i d(x,y;\theta)\right) = 0 \tag{4.112}$$

を満たしていることを仮定する．ただし，E_θ は確率密度 $p(x,y,\theta)$ による平均をあらわす．

次に 4 つの量を定義する．

$$g_{ij} = E_\theta\left(\partial_i d(x,y;\theta)\partial_j d(x,y;\theta)\right) \tag{4.113}$$

$$q_{ij} = E_\theta\left(\partial_i \partial_j d(x,y;\theta)\right) \tag{4.114}$$

$$T_{ijk} = E_\theta\left(\partial_i d(x,y;\theta)\partial_j d(x,y;\theta)\partial_k l(y|x,\theta)\right) \tag{4.115}$$

$$H_{ijk} = E_\theta\left(-\partial_i\partial_j d(x,y;\theta)\partial_k l(y|x,\theta)\right) \tag{4.116}$$

ただし

$$\partial_i = \frac{\partial}{\partial\theta^i}$$

であり，$l(y|x,\theta)$ は条件つき確率密度 $p(y|x,\theta)$ の対数

$$l(y|x,\theta) = \log p(y|x,\theta)$$

である．なお θ の各成分は添え字 $i,j,k,\cdots$ で表している．

これらの量がどのような座標変換をうけるか調べてみる．以下では $\theta = (\theta^i)$, $i = 1,\ldots,m$ が別の座標系 $w = (w^\alpha)$, $\alpha = 1,\ldots,m$ で書け，座標変換 $\theta = \theta(w)$ が滑らかで，Jacobian matrix

$$B^i_\alpha(\theta) = \frac{\partial\theta^i}{\partial w^\alpha}, \quad i,\alpha = 1,\ldots,m \tag{4.117}$$

が full rank であるとする．また

$$\partial_\alpha = \frac{\partial}{\partial w^\alpha}$$

とし，w の各成分は添え字 $\alpha,\beta,\gamma,\cdots$ で表すことにする．

g_{ij}, q_{ij} は前述の定義ではシステム $p(x,y,\theta)$ に対して定められる行列 G, Q の第 i 行第 j 列成分にあたり，いずれも対称テンソルとなる．すなわち，g_{ij}, q_{ij} は以下のような座標変換を受ける．

$$\begin{aligned} g_{\alpha\beta} &= E_\theta\left(\partial_\alpha d(x,y;\theta(w))\partial_\beta d(x,y;\theta(w))\right) \\ &= E_\theta\left(B^i_\alpha\partial_i d(x,y;\theta)B^j_\beta\partial_j d(x,y;\theta)\right) \\ &= B^i_\alpha B^j_\beta E_\theta\left(\partial_i d(x,y;\theta)\partial_j d(x,y;\theta)\right) \\ &= B^i_\alpha B^j_\beta g_{ij} \end{aligned} \tag{4.118}$$

$$\begin{aligned} q_{\alpha\beta} &= E_\theta\left(\partial_\alpha\partial_\beta d(x,y;\theta(w))\right) \\ &= E_\theta\left(\partial_\alpha B^i_\beta\partial_i d(x,y;\theta) + B^i_\alpha B^j_\beta\partial_i\partial_j d(x,y;\theta)\right) \\ &= \partial_\alpha B^i_\beta E_\theta\left(\partial_i d(x,y;\theta)\right) + B^i_\alpha B^j_\beta E_\theta\left(\partial_i\partial_j d(x,y;\theta)\right) \\ &= B^i_\alpha B^j_\beta q_{ij} \end{aligned} \tag{4.119}$$

(4.120)

対称性は明らかである．

T_{ijk} は i,j に関して対称なテンソルとなるが，H_{ijk} は i,j に関して対称ではあるがテンソルとはならない．すなわち，T_{ijk}, H_{ijk} は座標変換に対して以下の変換を受ける．

$$\begin{aligned}
T_{\alpha\beta\gamma} &= E_\theta\left(\partial_\alpha d(x,y;\theta(w))\partial_\beta d(x,y;\theta(w))\partial_\gamma l(y|x,\theta(w))\right)\\
&= E_\theta\left(B^i_\alpha\partial_i d(x,y;\theta)B^j_\beta\partial_j d(x,y;\theta)B^k_\gamma\partial_k l(y|x,\theta)\right)\\
&= B^i_\alpha B^j_\beta B^k_\gamma E_\theta\left(\partial_i d(x,y;\theta)\partial_j d(x,y;\theta)\partial_k l(y|x,\theta)\right)\\
&= B^i_\alpha B^j_\beta B^k_\gamma T_{ijk} \qquad (4.121)\\
H_{\alpha\beta\gamma} &= E_\theta\left(-\partial_\alpha\partial_\beta d(x,y;\theta(w))\partial_\gamma l(y|x,\theta(w))\right)\\
&= E_\theta\left(-\{\partial_\alpha B^i_\beta\partial_i d(x,y;\theta)+B^i_\alpha B^j_\beta\partial_i\partial_j d(x,y;\theta)\}B^k_\gamma\partial_k l(y|x,\theta)\right)\\
&= -\partial_\alpha B^i_\beta B^j_\gamma E_\theta\left(\partial_i d(x,y;\theta)\partial_j l(y|x,\theta)\right)\\
&\quad + B^i_\alpha B^j_\beta B^k_\gamma E_\theta\left(-\partial_i\partial_j d(x,y;\theta)\partial_k l(y|x,\theta)\right)\\
&= \partial_\alpha B^i_\beta B^j_\gamma q_{ij} + B^i_\alpha B^j_\beta B^k_\gamma H_{ijk} \qquad (4.122)
\end{aligned}$$

ただし，最後の式変形は式 (4.112) を偏微分して得られる

$$E_\theta\left(\partial_i\partial_j d(x,y;\theta)\right)+E_\theta\left(\partial_i d(x,y;\theta)\partial_j l(y|x,\theta)\right)=0 \qquad (4.123)$$

による．対称性は明らかである．また，H_{ijk} の座標変換は q_{ij} を計量と見たときの接続のうける座標変換と同型である．

さて θ で表されるモデルの中に $\theta(u)$ なる部分モデルを考えたとする．u の成分は添字 $a,b,c,\cdots$ で表示する．部分モデルの座標系 $u=(u^a)$ に添字 $\kappa,\lambda,\cdots$ で成分が表示される適当な座標系 $v=(v^\kappa)$ を付加してやることにより，$w=(u,v)$ を θ の座標系とできる．このとき w の成分は添字 $p,q,r,\cdots$ で表示する．また，u 上に点 θ をとり，別に u 上にない点 θ' をとる．2 点は

$$\theta'=\theta+\delta\theta \qquad (4.124)$$

の関係を満たしているとし，$\delta\theta$ を θ の座標で $(\delta\theta^i)$ と表し，w の座標で (δw^p) と表しておく．次の 2 つの量を考える．

$$\begin{aligned}
g'_{ab} &= E_{\theta'}\left(\partial_a d(x,y,\theta(u))\partial_b d(x,y,\theta(u))\right) \qquad (4.125)\\
q'_{ab} &= E_{\theta'}\left(\partial_a\partial_b d(x,y,\theta(u))\right) \qquad (4.126)
\end{aligned}$$

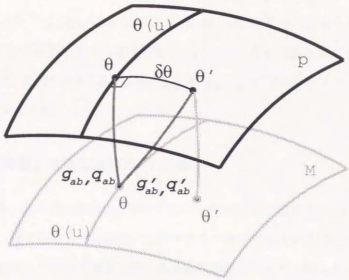


図 4.14: g, q と g', q' の関係.

∂_a は u^a に関する偏微分である. このとき

$$g'_{ab} = g_{ab} + T_{abp}\delta w^p + O(\|\delta\theta\|^2) \tag{4.127}$$

$$q'_{ab} = q_{ab} - H_{abp}\delta w^p + O(\|\delta\theta\|^2) \tag{4.128}$$

と展開できるので,

$$g'_{ab}q'^{ab} = g_{ab}q^{ab} + (q^{ab}T_{abp} + g_{ab}q^{ac}q^{bd}H_{cdp})\delta w^p + O(\|\delta\theta\|^2) \tag{4.129}$$

なる関係が成り立つ.

真のシステムが θ' で, 最適パラメタが θ であるとしてみる. このときモデル u は非忠実なモデルとなる. g'_{ab}, q'_{ab} はモデルの学習能力, あるいはモデル次数選択の規準量を計算するために必要とされる量であった. 右辺第 1 項はモデルが忠実であった場合に得られる $\mathrm{tr}GQ^{-1}$ に対応する. これは明らかに座標変換にはよらないスカラ量となる. 第 2 項はシステムが $\delta\theta$ ずれた分に対する補正項にあたる. 特に左の項は座標変換にはよらない項であり, 右の項は座標変換に影響される項であることに注意しておく. このためモデルが非忠実なとき, 座標のとり方によっても学習能力に差が生じる場合があり得ることがわかる.

図 4.14 は g_{ab}, q_{ab} と g'_{ab}, q'_{ab} の関係を模式的に書いたものである. システムの族を p で, モデルを M で表しており, 平均損失関数によってシステム θ はモデル θ に対応

づけられる．システム θ' に対して部分モデル $\{\theta(u)\}$ の中で最適なものが θ であった場合を考えている．g'_{ab}, q'_{ab} は p 上の点 θ' と M 上の点 θ から決められる量である．g_{ab}, q_{ab} はモデルが忠実であった場合の量で，p 上の点 θ と M 上の点 θ から決められる量である．式 (4.127), (4.128) は g_{ab}, q_{ab} と g'_{ab}, q'_{ab} の差をシステム θ, θ' の差で表したものになっている．

4.6 逐次型学習と非逐次型学習との比較

逐次型学習と確率型の非逐次型学習において確率的降下法を用いた同一の学習法を行なった場合，逐次型学習では観測したデータを一度学習に用いただけで捨ててしまうため，例題を記憶しておいて何度も用いる非逐次型学習に比べて明らかに損をする．これは学習の精度を制限した場合，すなわち同程度の正確さでパラメタの推定を行なおうとしたときには，学習に必要な例題の数の差となって現れる．非逐次型学習である精度を達成するのに必要な例題数を t_{ns}，逐次型学習に必要な例題数を t_s とする．2 種類の学習の学習曲線を比べることによって，非逐次型学習と逐次型学習で起こる損失を同程度にするためには

$$\mathrm{tr}\,(E - \varepsilon \Xi_{CQ})^{t_s} \tilde{V} Q \sim \frac{1}{t_{ns}} \mathrm{tr}\, G Q^{-1} \tag{4.130}$$

つまり

$$\exp\{-2\varepsilon \lambda_1 t_s\} \sim \frac{1}{t_{ns}} \tag{4.131}$$

であることがわかる．このため

$$\log t_{ns} \sim \varepsilon \lambda_1 t_s \tag{4.132}$$

の関係を満たさなければならないことがわかる．実際問題として必要な例題数は，学習の結果最終的に残る損失の誤差 $\frac{1}{2}\varepsilon\,\mathrm{tr}\,QV$ より例題数の不足のため現れる誤差が小さくなるように選べばよいから，必要な例題数の条件は，

$$t_{ns} \gg \frac{1}{\varepsilon} \tag{4.133}$$

$$t_s \gg \frac{1}{\lambda_1} \frac{1}{\varepsilon} \log \frac{1}{\varepsilon} \tag{4.134}$$

となる．

逐次型学習における学習の速度 (最終的に得られる平均損失からの偏差) は $O(e^{-\varepsilon t})$ であり，非逐次型学習における学習の速度は $O(1/t)$ であるから，たとえ ε がどんな

に小さくてもこの項を考える限りは逐次型学習の方が早く収束することになる．ところが確率的降下法の場合どんなに例題数を多くとっても最後に残ってしまう学習の‘ゆらぎ’があるため，ある程度大きな数以上の例題を用いても学習成果は上がらなくなる．t を徐々に増やしていったとき逐次型学習を行なった機械の平均損失は非逐次型学習を行なった機械の平均損失に近付いていくが，非逐次型学習が追いつく前に逐次型学習の平均損失は学習の‘ゆらぎ’の中に入ってしまい，学習が終了することになる．

第 5 章

例題からパラメタを一意に決定できない学習の特性

5.1 問題の記述

第 4 章での非逐次型学習の特性解析は，与えられた例題に対して最適パラメタが一意に決定する場合を扱ってきた．この場合には経験分布に対する平均損失関数 $D(p^t, \theta)$ を最適パラメタ θ^t_{opt} のまわりで展開して 2 次関数で近似することによってさまざまな特性を議論することができた．しかし平均損失関数 $D(p^t, \theta)$ において，最小値を達成するパラメタが無数に存在し，それらがパラメタ空間で測度 0 でない連結した領域を形成する場合にはこうした解析は行えない．このときの平均損失関数 $D(p^t, \theta)$ は底が平らなすり鉢状になると考えればよい (図 5.1)．そうした状況で第 4 章のような解析を行なうと，例題に対する平均損失関数 $D(p^t, \theta)$ の 2 階微分がパラメタ θ のある領域内で 0 になり，Q の逆行列が存在しなくなるため，議論の途中で破綻が生じることになる．

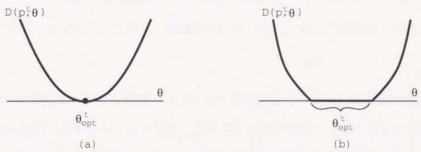


図 5.1: 経験分布に対する平均損失関数の形状の概念図．(a): 最適パラメタ θ^t_{opt} が一意に決まる場合．(b): 最適パラメタ θ^t_{opt} が一意に決まらない場合．

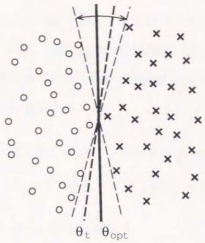


図 5.2: 二分割問題における例題を用いたパラメタ選択の冗長性.

具体的な例として次のような単純パーセプトロンが考えられる.

例 9 (Perceptron) 入力を $x \in \boldsymbol{R}^m$, 出力を $y = \{1, -1\}$ とし, 機械は次式にしたがい出力を計算するものとする.

$$y = \mathrm{sign}(\theta \cdot x) \tag{5.1}$$

ただし θ は機械を特徴づけるパラメタで $\theta \in \boldsymbol{R}^m$ であり, $\theta \cdot x$ は θ と x の内積を表す. この機械は $\boldsymbol{R}^m$ の中の信号を θ を法線ベクトルとする超平面の上下に分割する. さて, 例題はパラメタ θ_{opt} をもつ機械から t 個発生されたとし, この t 個の例題に矛盾しないように機械のパラメタを選ぶことにする. 通常この条件を満たすパラメタは無数に存在するので, この中から適当に一つ取り出すことにし, これを θ_t とする (図 5.2). こうして得られた機械 θ_t は, 例題の中にあった入力に対しては必ず正解の出力を出すことができる. しかし, 例題の中にない入力に対しては正解を出すとは限らない.

例題が無数にあるのなら最適パラメタは一意に決定されるが, 有限個の例題に対してはそれら全てに正解を与えるようなパラメタが, 真の最適パラメタのまわりにいくらでも存在し得る場合があるため, 経験分布に対する平均損失関数を最適パラメタのまわりで展開するという手法は使えなくなる.

単純パーセプトロンに限らず, こうした問題は様々な状況で起こり得る. これを一般的な問題として取り扱うのは非常に困難なので, ここでは入力信号空間を二分割する

単純な機械に限って議論し，機械の犯す誤りが例題の数の増加とともにどのように減少していくかを調べる．これは上で述べた例 9 の単純パーセプトロンや例 3 の確定的な分割を行なう機械を含んでいる．具体的な問題の記述は次のようになる．

パラメタ $\theta \in \boldsymbol{R}^m$ を持つ $\boldsymbol{R}^n$ 内の集合の定義関数 $\pi(x,\theta)$ を考える．この定義関数により $\boldsymbol{R}^n$ は 2 つの集合に分割される．

$$D_+ = \{x|\pi(x,\theta)=1, x\in\boldsymbol{R}^n\}, \quad D_- = \{x|\pi(x,\theta)=0, x\in\boldsymbol{R}^n\}$$
$$D_+\cap D_- = \emptyset, \quad D_+\cup D_- = \boldsymbol{R}^n \tag{5.2}$$

これを用いて条件つき確率 $p(y|x,\theta)$ を定義する．

$$p(y|x,\theta) = \begin{cases} \pi(x,\theta), & y=1 \\ 1-\pi(x,\theta), & y=-1 \end{cases} \qquad x\in\boldsymbol{R}^n \tag{5.3}$$

あとの記述をわかりやすくするために，便宜上条件つき確率の形で表現するが，実際は $p(y|x,\theta)$ は 0,1 の 2 値しかとらない．いいかえると $p(1|x,\theta)$, $p(-1|x,\theta)$ は x の関数としてみた場合，それぞれ D_+, D_- の定義関数になっている．

さて，条件つき確率の族 $\{p(y|x,\theta)\}$ の中にシステムの入出力を実現する真の機械が存在するとする．これを $p(y|x,\theta_{opt})$ とする．入力 x の分布を $p(x)$ とすればシステムは

$$S = (p(x), p(y|x,\theta_{opt})) \tag{5.4}$$

で表され，モデルとして

$$M = \{p(y|x,\theta)\} \tag{5.5}$$

を用いることにすれば，これは忠実なモデルとなる．例題は $p(y|x,\theta_{opt})$ から独立に t 個発生されるとし，いままでと同様に

$$\xi^t = \{(x_1,y_1),\cdots,(x_t,y_t)\} \tag{5.6}$$

で表す．また Bayes の立場をとり，パラメタ θ の事前分布 $q(\theta)$ を仮定する．以上の準備のもとで，次にあげる確率を考える．

定義 5.1

$Q(\theta,\xi^t)$

$=$ Prob{機械 θ が選択され，かつ例題 ξ^t が生成される}

$$= q(\theta)\prod_{i=1}^{t} p(y_i|x_i,\theta)p(x_i), \tag{5.7}$$

$p(\xi^t)$

$$= \text{Prob}\{\text{例題 } \xi^t \text{ が生成される}\}$$

$$= \int Q(\theta,\xi^t)d\theta$$

$$= Z_t(\xi^t)\prod_{i=1}^{t} p(x_i), \tag{5.8}$$

$Z_t(\xi^t)$

$$= \text{Prob}\{\text{ランダムに選ばれた機械が 例題 } \xi^t \text{ に対して正しい答を出す}\}$$

$$= \int q(\theta)\prod_{i=1}^{t} p(y_i|x_i,\theta)d\theta \tag{5.9}$$

上で定義された確率を用いてこの問題での学習則，つまり機械の選び方を次のように定める．

定義 5.2 機械は例題 ξ^t に矛盾しないものの中から，事前分布にしたがって選ばれるものとする．すなわち，例題の組 ξ^t が与えられたとき，パラメタ θ の選ばれる確率は

$$Q(\theta|\xi^t) = \frac{Q(\theta,\xi^t)}{p(\xi^t)} \tag{5.10}$$

で表される．

例題に矛盾しない θ の領域に，事前分布による重みをつけたもので機械を選ぶのであるから，事前分布が一様ならば確率密度関数 $Q(\theta|\xi^t)$ は例題に矛盾しない θ の定義関数の定数倍になる．

例題の組 ξ^t を用い，上の規則にしたがって選ばれる機械の集合に対しては，新たな例題 (x_{t+1},y_{t+1}) に正解する機械の確率として予測分布 $p(y_{t+1}|x_{t+1};\xi^t)$ を定めることができる．

定義 5.3

$$p(y_{t+1}|x_{t+1};\xi^t)$$

$$= \text{Prob}\{\xi^t \text{ に矛盾しない機械の中で } (x_{t+1},y_{t+1}) \text{ に矛盾しない}\} \tag{5.11}$$

予測分布 $p(y_{t+1}|x_{t+1};\xi^t)$ は次のように確率 Z_t で表される．

補題 5.1 予測分布は

$$p(y_{t+1}|x_{t+1};\xi^t) = \frac{Z_{t+1}(\xi^{t+1})}{Z_t(\xi^t)} \tag{5.12}$$

で与えられる．但し，

$$\xi^{t+1} = \{\xi^t, (x_{t+1}, y_{t+1})\}$$

証明

$$\begin{aligned} p(y_{t+1}|x_{t+1};\xi^t) &= \int p(y_{t+1}|x_{t+1},\theta)Q(\theta|\xi^t)d\theta \\ &= \frac{\int Q(\theta,\xi_t)p(y_{t+1}|x_{t+1},\theta)d\theta}{Z_t(\xi^t)\prod_{i=1}^t p(x_i)} \\ &= \frac{\int q(\theta)\prod_{i=1}^{t+1} p(y_i|x_i,\theta)d\theta}{Z_t(\xi^t)} \\ &= \frac{Z_{t+1}(\xi^{t+1})}{Z_t(\xi^t)} \end{aligned} \tag{5.13}$$

より明らか． ∎

注． (x,y) を固定し θ の関数としてみたとき方程式 $p(y|x,\theta) = 1$ が完全に解けているのなら，定義 5.2 の式 (5.10) にしたがって θ を求めることはできる．しかし，式 (5.10) の計算が明示的に行なえない場合には，定義 5.2 に定めた規則で実際に θ を求めることは非常に難しい．一般にはパラメタ空間を事前分布にしたがって無作為探索するしか方法はない．しかし，例 3 のように平均損失関数 $D(p^t,\theta)$ が広がりを持ったパラメタの領域で一様に極小となっても，損失関数 $d(x,y;\theta)$ の 1 階微分が存在すれば逐次型学習によって ξ^t に矛盾しない θ を求めることは可能である．実際パーセプトロンの学習アルゴリズムはこの方法に基づいており，有限回の操作でその収束が保証されている．ただし，この場合には ξ^t に矛盾しない θ の領域の境界近傍で学習は終了する．

5.2 学習曲線と予測エントロピー

出力 y が ± 1 の二値なので予測損失として予測誤差が自然に定義できる．

定義 5.4 予測誤差を次式で定義する．

$$\begin{aligned} e(t) &= 1 - E_{\xi_{t+1}}\left(p(y_{t+1}|x_{t+1};\xi^t)\right) \\ &= 1 - \frac{E_{\xi_{t+1}}\left(Z_{t+1}(\xi^{t+1})\right)}{Z_t(\xi^t)} \end{aligned} \tag{5.14}$$

定義 5.2 にしたがって選ばれた機械が，新たな例題 (x_{t+1}, y_{t+1}) に対して矛盾してしまう確率は

$$1 - p(y_{t+1}|x_{t+1}; \xi^t) \tag{5.15}$$

である．これを新たに出現する例題で平均した期待値が予測誤差になっている．しかし，上に定義した予測誤差は計算が困難であり，漸近的な振舞いさえ調べるのが難しいので，別の規準にしたがってもうひとつの予測損失を定義する．

定義 5.5 予測分布の対数を対数予測誤差と定義する．

$$\begin{aligned} e^*(t) &= -\log p(y_{t+1}|x_{t+1}, \xi^t) \\ &= \log Z_t(\xi^t) - E_{\xi_{t+1}}\left(\log Z_{t+1}(\xi^{t+1})\right) \end{aligned} \tag{5.16}$$

対数予測誤差は機械が多数あった場合，新たな例題に対する機械の正答率の平均を情報量に換算したものであり，この意味での損失となっていることを注意しておく．したがって，一つ一つの機械が持つ情報量の平均とは意味が異なる．特にこの問題設定では一つの機械が一つの例題に対して持つ情報量はうまく定義できない．またこのとき

$$1 - x \leq -\log x$$

であるから，2 つの予測損失の間には

$$e(t) \leq e^*(t) \tag{5.17}$$

なる関係が成り立つ．すなわち対数予測誤差は予測誤差の緩い上限を与えていることがわかる．

上で定義した 2 つの予測損失は例題 ξ^t の関数として確率変数になっている．そこで例題 ξ^t の分布に関して平均をとった量で学習の特性を記述することにする．

定義 5.6 予測誤差の期待値により平均予測誤差を定義し，これを学習曲線という．

$$\begin{aligned} E_{\xi^t}(e(t)) &= 1 - E_{\xi^t}\left(\frac{E_{\xi_{t+1}}(Z_{t+1}(\xi^{t+1}))}{Z_t(\xi^t)}\right) \\ &= 1 - E_{\xi^{t+1}}\left(\frac{Z_{t+1}(\xi^{t+1})}{Z_t(\xi^t)}\right) \end{aligned} \tag{5.18}$$

定義 5.7 対数予測誤差の期待値により予測エントロピーを定義する．

$$\begin{aligned} E_{\xi^t}(e^*(t)) &= E_{\xi^t}\left(\log Z_t(\xi^t) - E_{\xi_{t+1}}\left(\log Z_{t+1}(\xi^{t+1})\right)\right) \\ &= E_{\xi^t}\left(\log Z_t(\xi^t)\right) - E_{\xi^{t+1}}\left(\log Z_{t+1}(\xi^{t+1})\right) \end{aligned} \tag{5.19}$$

このとき，2 つの予測損失の間に成り立つ関係から，

$$E_{\xi^t}(e(t)) \leq E_{\xi^t}(e^*(t)) \tag{5.20}$$

なる関係が成り立ち，予測エントロピーは平均予測誤差，すなわち学習曲線の上限を与えることがわかる．

5.3 予測エントロピーの性質

この問題設定で学習曲線の漸近特性を求めることは未解決である．以下では学習曲線の上限となる予測エントロピーの性質について調べる．

予測エントロピーの漸近特性については次の定理が成り立つ．

定理 5.1 予測エントロピーの漸近特性は

$$E_{\xi^t}(e^*(t)) = \frac{m}{t} \tag{5.21}$$

で与えられる．

この定理の証明には次の補題を用いる．

補題 5.2

$$Y_t = t^m Z_t, \quad t = 1, 2, \ldots \tag{5.22}$$

とおくと Y_t はある確率変数 Y に法則収束する．

証明 パラメタ θ を持つ機械が真の機械 θ_{opt} と同じ出力を与える確率を $s(\theta)$ とすると，これは

$$s(\theta) = \int p(y|x,\theta)p(y|x,\theta_{opt})p(x)dxdy \tag{5.23}$$

で与えられる．各例題 (x_i, y_i) は独立であるからこれを用いると

$$\begin{aligned} E_{\xi^t}\left(Z_t(\xi^t)\right) &= \int q(\theta)\prod_{i=1}^{t} p(y_i|x_i,\theta)d\theta \prod_{i=1}^{t} p(y_i|x_i,\theta_{opt})dx_i dy_i \\ &= \int s(\theta)^t q(\theta)d\theta \end{aligned} \tag{5.24}$$

と書き直すことができる．同様に k 個の機械がそれぞれパラメタ $\theta_1, \cdots, \theta_k$ を持っているとし，これらが同時に真の機械 θ_{opt} と同じ出力を与える確率を $s(\theta_1, \cdots, \theta_k)$ とすると，

$$s(\theta_1, \cdots, \theta_k) = \int \prod_{i=1}^{k} p(y|x,\theta_i)p(y|x,\theta_{opt})p(x)dxdy \tag{5.25}$$

となり，

$$E_{\xi^t}\left((Z_t(\xi^t))^k\right) = \int s(\theta_1,\cdots,\theta_k)^t \prod_{i=1}^{k} q(\theta_i)d\theta_i \tag{5.26}$$

と書ける．このとき明らかに

$$s(\theta_1,\cdots,\theta_k) \leq s(\theta_{opt},\cdots,\theta_{opt}) = 1 \tag{5.27}$$

が成り立っている．等号成立は $\theta_1 = \cdots = \theta_k = \theta_{opt}$ のときである．ここで

$$\theta_i = \theta_{opt} + r_i e_i, \quad i = 1,\ldots,k$$

とおく．r_i は正の実数であり，e_i は $\boldsymbol{R}^m$ 内の単位ベクトルとする．r_i が小さいとき $s(\theta_1,\cdots,\theta_k)$ は次のように展開できる．

$$s(\theta_1,\cdots,\theta_k) = 1 - \sum_{i=1}^{k} a_i(e_1,\cdots,e_k)r_i \tag{5.28}$$

θ_i に関する積分を極座標 (r_i,Ω_i) に書き換え，t が十分大きいとして鞍点法 (saddle point approximation) を用いると，$1/t$ に関して最も低次の項は

$$\begin{aligned} E_{\xi^t}\left((Z_t(\xi^t))^k\right) &= \int (1 - \sum_{i=1}^{k} a_i(e_1,\cdots,e_k)r_i)^t \prod_{j=1}^{k} q(\theta_j)d\theta_j \\ &= c_k' \int \exp\{-t\sum_{i=1}^{k} a_i(\Omega_1,\cdots,\Omega_k)r_i\} \prod_{i=1}^{k} r_i^{m-1} dr_i d\Omega_i \\ &= c_k' \int \prod_{i=1}^{k} \frac{\Gamma(m)d\Omega_i}{t^m a_i(\Omega_1,\cdots,\Omega_k)^m} \\ &= \frac{c_k}{t^{mk}} \end{aligned} \tag{5.29}$$

となる．これは

$$Y_t = t^m Z_t, \quad t = 1,2,\ldots \tag{5.30}$$

とおいたとき Y_t の k 次のモーメントが c_k に収束することを表している．つまり Y_t の確率分布を $\Phi_t(Y_t)$ とすると

$$\lim_{t\to\infty} \Phi_t(Y_t) = \Phi(Y) \tag{5.31}$$

なる確率分布が存在することを表している． ∎

これを用いて定理の証明に移る．

証明 補題 5.2 より

$$E_{\xi^t}\left(\log Z_t(\xi^t)\right) \sim E_{\xi^t}\left(\log(t^{-m}Y)\right) = E_{\xi^t}\left(\log Y\right) - m\log t \tag{5.32}$$

よって,

$$\begin{aligned} E_{\xi^t}(e^*(t)) &= E_{\xi^t}\left(\log Z_t(\xi^t)\right) - E_{\xi^{t+1}}\left(\log Z_{t+1}(\xi^{t+1})\right) \\ &\sim m\log t - m\log(t+1) \\ &= \frac{m}{t} + O(\frac{1}{t^2}) \end{aligned} \tag{5.33}$$

以上より定理が証明される. ∎

確率変数 Y_t は一般に概収束しないので, この補題を用いて平均予測誤差を求めることはできない. 確率変数 Y_t が概収束するならば, 次のような手続きで平均予測誤差の漸近特性を知ることができる. 確率変数 Y_t の概収束先を Y とする. t が十分大きいとすれば, $Y_t \sim Y$ だから $Z_t \sim t^{-m}Y$ と表すことができる. よって

$$\begin{aligned} E_{\xi^t}(e(t)) &\sim 1 - E_{\xi^t}\left(\frac{(t+1)^{-m}Y}{t^{-m}Y}\right) \\ &= 1 - \left(1+\frac{1}{t}\right)^{-m} \\ &= \frac{m}{t} + O(\frac{1}{t^2}) \end{aligned} \tag{5.34}$$

である. これは漸近特性として予測エントロピーが平均予測誤差の厳密な上限になることを示している. ところが, これに矛盾する例を次のように簡単に作ることができる.

例 10 (Dichotomy of S^1) 以下では入力 $x \in R$ を mod2 で考える. 次のような計算をする機械を考える.

$$f(x,\theta) = \begin{cases} 1, & x \in (\theta, \theta+1) \\ -1, & \text{otherwise} \end{cases} \qquad \theta \in R \tag{5.35}$$

明らかに f は台の大きさが 1 となる定義関数を用いて表すことができる. さて, 真の機械は $\theta_{opt} = 0$ であるとする. すなわち

$$y(x) = f(x,0) = \begin{cases} 1, & 0 < x < 1 \\ -1, & 1 \le x \le 2 \end{cases} \tag{5.36}$$

となる. 入力 x の出現確率 $p(x)$ は $0 < x < 1$ 上で一様とし, θ の事前分布 $q(\theta)$ は $0 < \theta \le 2$ で一様とする. 真の機械 $y(x)$ より独立に生成された t 個の例題は全て $y = 1$ となるが, 入力の組を

$$X^t = \{x_1, \cdots, x_t\} \tag{5.37}$$

とおく．また

$$\min_{X^t} x_i = \underline{x}, \quad \max_{X^t} x_i = \overline{x} \tag{5.38}$$

とすると，

$$\begin{aligned} S_t &= \{\theta|\xi^t \text{ に正解を与える}\} \\ &= \{\theta|X^t \text{ に対して } 1 \text{ を出力する}\} \\ &= \{\theta|\overline{x} - 1 < \theta < \underline{x}\} \end{aligned} \tag{5.39}$$

となり，Z_t は

$$Z_t = \frac{1}{2}|S_t| = \frac{1}{2}((1-\overline{x}) + \underline{x}) \tag{5.40}$$

と表される．以下では

$$1 - \overline{x} = \alpha, \quad \underline{x} = \beta \tag{5.41}$$

とする．β の分布関数を $F(\beta)$ とすると，x が $0 < x < 1$ で一様なことから

$$1 - F(\beta) = \text{Prob}\{\min_{X^t} x_i > \beta\} = (1-\beta)^t \tag{5.42}$$

となり，確率密度は漸近的に

$$p(\beta) = t(1-\beta)^{t-1} \simeq te^{-t\beta} \tag{5.43}$$

となる．α の分布も同様に

$$p(\alpha) = t(1-\alpha)^{t-1} \simeq te^{-t\alpha} \tag{5.44}$$

である．つぎに Z_t の分布を求める．まず $Z^* = \alpha + \beta$ とすると

$$\begin{aligned} p(Z^*) &= \int_0^{Z^*} te^{-t\alpha} \cdot te^{-t(Z^*-\alpha)} d\alpha \\ &= t^2 e^{-tZ^*} \int_0^{Z^*} d\alpha \\ &= Z^* t^2 e^{-tZ^*} \end{aligned} \tag{5.45}$$

ここで $Z^* = 2Z_t$ であるから

$$p(Z_t) = 4Z_t t^2 e^{-tZ_t} \tag{5.46}$$

となり，

$$\begin{aligned} E_{\xi^t}(Z_t) &= \int 4x^2t^2e^{-2zt}dz \\ &= \frac{1}{2t}\int u^2e^{-u}du \\ &= \frac{1}{t} \end{aligned} \tag{5.47}$$

$$\begin{aligned} E_{\xi^t}\left((Z_t - E_{\xi^t}(Z_t))^2\right) &= \int 4x^3t^2e^{-2zt}dz - \frac{1}{t^2} \\ &= \frac{1}{2t}\int u^2e^{-u}du - \frac{1}{t^2} \\ &= \frac{1}{2t^2} \end{aligned} \tag{5.48}$$

が漸近的に成り立つ．また

$$Z_{t+1} = \begin{cases} Z_t, & \underline{x} < x_{t+1} < \overline{x} \\ \frac{1}{2}(1-\overline{x}+x_{t+1}), & x_{t+1} < \underline{x} \\ \frac{1}{2}(1-x_{t+1}+\underline{x}), & x_{t+1} > \overline{x} \end{cases} \tag{5.49}$$

であるので，

$$\begin{aligned} & E_{\xi_{t+1}}(Z_{t+1}) \\ &= (\overline{x}-\underline{x})Z_t + \frac{1}{2}\underline{x}(1-\overline{x}+\frac{1}{2}\underline{x}) + \frac{1}{2}(1-\overline{x})(\frac{1}{2}(1-\overline{x})+\underline{x}) \\ &= \frac{1}{2}(\alpha+\beta) - \frac{1}{4}(\alpha^2+\beta^2) \end{aligned} \tag{5.50}$$

したがって，

$$\frac{E_{\xi_{t+1}}[(_{t+1})}{Z_t} = 1 - \frac{1}{2}\frac{\alpha^2+\beta^2}{\alpha+\beta} \tag{5.51}$$

となる．$t\alpha = u$, $t\beta = v$ と置換して

$$E_{\xi^t}\left(\frac{E_{\xi_{t+1}}(Z_{t+1})}{Z_t}\right) = 1 - \frac{1}{2t}\int_0^\infty\int_0^\infty \frac{u^2+v^2}{u+v}e^{-(u+v)}dudv \tag{5.52}$$

さらに $x = u+v$, $y = u-v$ と置換して

$$\begin{aligned} \int_0^\infty\int_0^\infty \frac{u^2+v^2}{u+v}e^{-(u+v)}dudv &= \frac{1}{2}\int_0^\infty dx\int_{-x}^{x} dy\frac{x^2+y^2}{2x}e^{-x} \\ &= \frac{3}{2}\int_0^\infty x^2e^{-x}dx \\ &= \frac{4}{3} \end{aligned} \tag{5.53}$$

が得られる．よって

$$E_{\xi^t}(e(t)) = 1 - E_{\xi^t}\left(\frac{E_{\xi_{t+1}}(Z_{t+1})}{Z_t}\right) = \frac{2}{3t} < \frac{1}{t} \tag{5.54}$$

となり，1 次元の場合の予測エントロピーより小さくなることがわかる．

なお，$p(x)$ を $0 < x \leq 2$ で一様としたとき，θ を $-1 < \theta < 1$ で考えると，S_t は次の区間になる．

$$S_t : -\alpha < \theta < \beta \tag{5.55}$$

$$-\alpha = \max_{\xi^t} \left\{ x_i^{(1)} - 1, x_i^{(-1)} - 2 \right\}$$

$$\beta = \max_{\xi^t} \left\{ x_i^{(1)}, x_i^{(-1)} - 1 \right\}$$

ただし，$x_i^{(1)}$ は $y_i = 1$ となる $x_i \in \xi^t$ であり，$x_i^{(-1)}$ は $y_i = -1$ となる $x_i \in \xi^t$ である．α, β の分布関数 F および確率密度 p を求めると

$$1 - F(\alpha) = \mathrm{Prob}\left\{ \max_{\xi^t} \left\{ (x_i^{(1)} - 1, x_i^{(-1)} - 2 \right\} < -\alpha \right\}$$

$$= \left\{ \frac{1}{2}(1-\alpha) + \frac{1}{2}(1-\alpha) \right\}^t$$

$$= (1-\alpha)^t \tag{5.56}$$

$$p(\alpha) = t(1-\alpha)^{t-1} \simeq te^{-t\alpha} \tag{5.57}$$

$$1 - F(\beta) = \mathrm{Prob}\left\{ \max_{\xi^t} \left\{ x_i^{(1)}, x_i^{(-1)} - 1 \right\} > \beta \right\}$$

$$= \left\{ \frac{1}{2}(1-\beta) + \frac{1}{2}(1-\beta) \right\}^t$$

$$= (1-\beta)^t \tag{5.58}$$

$$p(\beta) = t(1-\beta)^{t-1} \simeq te^{-t\beta} \tag{5.59}$$

となり，上と同様に議論できる．

以上より，予測エントロピーが平均予測誤差（学習曲線）の厳密な上限にはなっていないことがわかる．学習曲線の漸近特性を求めるためには，例題の組 ξ^t に矛盾しないパラメタの領域の面積*$Z_t(\xi^t)$ が，新たに提示された例題により減少していく過程を厳密に評価していかなくてはならない(図 5.3)．これは m 次元空間を超平面により分割していく問題と見做すことができ，幾何学的な見方も含めた洞察が必要とされるであろう．

*事前分布で重みづけが行なわれているので正しくは面積ではない．

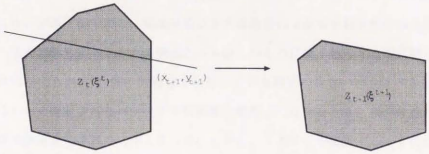


図 5.3: 新たに与えられた例題により $Z_t(\xi^t)$ が減少していく様子.

第 6 章

結論

本研究では一つ一つの入出力の組に対して定義される損失関数と，それをシステムの入出力の同時分布で平均した平均損失関数を考えることにより，学習機械で問題となる例題数と機械の大きさの関係を統一的に見渡してきた．ここで扱った学習は例題の順序に依存しない非逐次型学習と，確率近似法から導出される最も単純な形の逐次型学習であった．逐次型学習が与えられる例題のうち新しいものに強く依存するのに対し，非逐次型学習は与えられた t 個の例題全てに対して最も損失の少ないパラメタを得ようとする．このため推定されるパラメタを例題数の関数として見たとき，真の最適パラメタへの収束は非逐次型学習の方がよかった．しかし，システムが変動するときには，例題を一つの増やしたときの計算量が圧倒的に少なくなる逐次型学習の方が有利である．この有利さは，パラメタの更新が機械を構成する素子毎に局所的な計算で行なえる Multi Layer Network や Kohonen map では顕著に表れる．こうした観点から見ると一般に非逐次型学習は off-line の学習，逐次型学習は on-line の学習に向いているといえるだろう．学習の性格がこのように著しく違うことは，本論文で調べた学習曲線の漸近特性によく現れている．しかしながらこの漸近特性を決める量が，

$$G = (g_{ij}) = E_p[\partial_i d(x, y; \theta_{opt}) \partial_j d(x, y; \theta_{opt})] \tag{6.1}$$

$$Q = (q_{ij}) = E_p[\partial_i \partial_j d(x, y; \theta_{opt})] \tag{6.2}$$

という 2 つの行列から計算される量によって特徴づけられることは注目に値する．いいかえると，この 2 つの行列が学習機械そのものの主たる特性を表していることがわかる．

本研究で扱ったように損失関数が定義されるタイプの学習の問題は，統計学の推定

関数の問題，特にポテンシャルを持つ推定関数の問題と深くつながってる．最尤推定などの一般の推定法が行なえない場合に利用される推定関数は，幾何学的にも興味深い特徴を持つ．行列 G, Q はモデルの空間で計量としての性質を持っている．通常の推定論で扱われる最尤推定は，損失関数として $-\log p$ を使うことに相当し，モデルが忠実な場合には G, Q は一致して見かけ上計量は一つしか現れず，様々な議論は一つの計量に対して行なわれることが多い．しかし一般の損失関数では G, Q は一致しないで残るので，2 つの計量を持つ空間の幾何学として興味深い問題であると思われる．

なお，本研究の延長としては次のような問題が残されている．

ひとつは高次の学習系の問題である．ここでいう高次の学習とは，逐次型学習における ε や C などの学習則を司るものそのものの学習や，機械の素子数や結合の局所化を自動的に行なうような機械の構造の学習を指す．学習則の学習に関しては Amari (1968) [8] にすでに試みがあるものの，その解析等は不十分である．過去に与えられた例題の影響が未来まで複雑な形で残るような確率過程の問題としての考察が必要である．構造の学習は，素子数や結合の変更によって行列 G, Q の成分ではなく，大きさそのものが変化してしまうという意味で別の難しさがある．多くの研究者は損失関数そのものを工夫することにより構造の変化を学習に回帰させようとしているが，未だ著しい結果は得られていないように思う．これらの問題に関してはより深い洞察に基づいた工夫が必要なように思われる．

もうひとつはサンプリング (sampling) の問題である．これは，ある規準に沿って学習系に必要な例題を自ら選んで観測し，それに基づいてパラメタを更新していくような学習を考えようというものである．本研究で扱ってきたのは環境に依存するタイプの学習であった．環境が生成する入力の確率 $p(x)$ で重み付けされた意味で最適なパラメタを定義していた．このため一致性のない損失関数を用いたり，忠実でないモデルを用いると環境によって最適パラメタが変化してしまう．必然的に環境からよく受ける入力 x に対して損失が少なくなるようにパラメタが選ばれることになるため，選ばれるパラメタは環境から決まってしまうことになる．しかし現実の生物の学習を考えると，この戦略が必ずしも正しいとはいえなくなる．つまり現実問題では判断結果が致命的なものとなるような問題は稀にしか起こらないが，それに対する損失ができるだけ小さくなるようでなければ生き残ることはできないからである．本研究の枠組でも，例えば損失関数として p-norm が用いられているタイプのものであれば，p を大きくとることによって近似的に ∞-norm，すなわち max norm を用い，最悪事態の損失をできるかぎり小

さくするような学習法といったものを導出することはできる．しかしこの場合も本質的に環境の重み付けによっていることに変わりはなく，本当に危険性を持つ例題の出現確率が低ければ，こうした損失関数も実質上あまり意味を持たなくなる．むしろ致命的な影響を与えそうな例題をすすんで観測するような学習法を考えるべきであろう．このサンプリングの問題は，例題が損失に関して持つ情報量をできるだけ稼ぐように，自ら例題を選択するという意味で，受動的な学習から能動的な学習への移行という見方もできる．例題の持つ情報量という問題については最近 Rissanen and Yu [44] によって興味深い試みが報告されている．彼らは，確率的な空間の二分割問題において境界のみを求めるためには，例題をどのようなヒストグラムに分ければよいかという問題を MDL 規準量から解き，1 次元問題の場合，集められる例題数が n のとき，$n^{1/3}$ 個のヒストグラムに分けると最も効率よく境界を求めることができるという結果を与えている．これは見方をかえれば，境界を見つけ出すためには，境界付近の $O(n^{2/3})$ 個の例題が必要とされることを意味している．つまり，このような問題設定の場合，能動的に例題を集めることができるのなら，$O(n^{2/3})$ 個の例題で受動的な場合と同等の推定が行なえる，あるいは同じ例題数なら受動的な場合の精度の 3/2 乗分よい精度で推定が行なえる可能性を示唆している．こうした議論を一般化することができれば，能動的に例題を観測する学習系を構成することができるであろう．

謝辞

修士課程，博士課程の五年間に渡り研究をはじめ様々な面で丁寧に御指導戴いた吉澤修治助教授に感謝します．本研究を進めるにあたり適切な助言と示唆を与え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，貴重な御助言を戴いた鈴木良次教授，広津千尋教授，中野馨助教授に感謝します．

自分の考えをまとめていく上で非常に有益な話を聞かせていただいたり，貴重な時間を割いて議論につきあってくださった銅谷賢治助手，伊藤尚史助手，阪口豊助手に感謝します．

常に暖かく見守り励ましてくれた瀬部昇氏，津村幸治氏，新妻素直氏をはじめとする全ての友人たちに感謝します．

研究内容をはじめとしていろいろな面で相談にのってくださった本研究室の大濱靖匡氏，池田和司君，小林慶一郎君，藤田直毅君，村松正和君，川鍋元明君，山下敬君，齋藤渉君，藤原彰夫君，堀玄君，水田成仁君，秘書の濱川ゆかりさん，杉本敦子さん，さらに鳥居真君をはじめとする当学科の大学院生諸君，学生諸君に感謝します．

最後になりましたが，大学院の五年を含め二十年を越える長い学生生活の間，この我儘な子の言葉に真剣に耳を傾け，理解を示し，全ての面で惜しむことなく援助し，励まし続けて下さった両親とそして智美に感謝の意を表します．

参考文献

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans.*, Vol. AC–19, No. 6, pp. 716–723, 1974.

[2] S. Amari. Theory of adaptive pattern classifiers. *IEEE Trans.*, Vol. EC–16, No. 3, pp. 299–307, 1967.

[3] S. Amari. *Differential-Geometrical Methods in Statistics*, Vol. 28 of *Lecture Notes in Statistics*. Springer-Verlag, 1985.

[4] S. Amari. Mathematical foundations of neurocommmputing. *Proc. IEEE*, Vol. 78, No. 9, pp. 1443–1463, 1990.

[5] S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. METR 91-04, University of Tokyo, 1991.

[6] S. Amari and M. Kumon. Estimation in the presence of infintely many nuisance parameters – geometry of estimating functions. *Ann. Statist.*, Vol. 16, No. 3, pp. 1044–1068, 1988.

[7] S. Amari and N. Murata. Statistical theory of learning curves under entropic loss criterion. METR 91-12, University of Tokyo, 1991.

[8] 甘利俊一. 情報理論 II – 情報の幾何学的理論 –, 情報科学講座, 第 A.2.5 巻. 共立出版, 1968.

[9] 甘利俊一. 神経回路網の数理. 産業図書, 1978.

[10] 甘利俊一, 村田昇. 予測誤差と学習曲線. 日本応用数理学会平成 3 年度年会研究発表予稿集, pp. 65–66, October 1991.

[11] 甘利俊一, 篠本滋, 藤田直毅, 村田昇. 一般化誤差と学習曲線. 日本応用数理学会平成 3 年度年会研究発表予稿集, p. 64, October 1991.

[12] 甘利俊一, 篠本滋, 村田昇, 藤田直毅. 一般化誤差, 予測エントロピー, 学習曲線. 神経回路学会第 2 回全国大会講演論文集, p. 26, December 1991.

[13] E. B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, Vol. 1, pp. 151–160, 1989.

[14] H. D. Block. The Perceptron, a model for brain functioning I. *Rev. of Modern Physics*, Vol. 34, No. 1, pp. 123–135, 1962.

[15] H. D. Block, B. W. Knight, and F. Rosenblatt. Analysis of four-layer series coupled Perceptron II. *Rev. of Modern Physics*, Vol. 34, No. 1, pp. 135–142, 1962.

[16] 鍋谷賢治, 吉澤修治. 神経回路における運動パターンの記憶. 電子情報通信学会技術研究報告, Vol. MBE 87, pp. 293–300, 1988.

[17] D. B. Forgel. An information criterion for optimal neural network selection. *IEEE Trans.*, Vol. NN-2, No. 5, pp. 490–497, 1991.

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

[19] T. M. Heskes and B. Kappen. Learning processes in neural networks. Submitted to Physical Review A, 1991.

[20] G. E. Hinton. 認知・学習のコネクショニストモデル. 科学, Vol. 57,, pp. 228–237, 1987.

[21] G. E. Hinton and T. J. Sejnowski. Learning and relearning in Boltzmann machines. In D. Rumelhart, J. L. McClelland, and the PDP Reserch Group, editors, *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, Vol. 1: Foundations, chapter 1, pp. 282–317. The MIT Press, 1986.

[22] 池田央, 小野茂ほか. 数理心理学. 広中平祐ほか (編) , 現代数理科学事典, 第 VI 章, pp. 353–382. 大阪書籍, 1991.

[23] 伊藤清. 確率論, 現代数学, 第 14 巻. 岩波書店, 1953.

[24] 伊藤清. 確率論. 岩波基礎数学選書. 岩波書店, 1991.

[25] 児玉慎三, 須田信英. システム制御のためのマトリクス理論. 計測自動制御学会, 1978.

[26] T. Kohonen. Correlation matrix memories. *IEEC Trans.*, Vol. C-12, No. 4, pp. 353–359, 1972.

[27] T. Kohonen. *Associative Memory*. Springer Verlag, 1977.

[28] T. Kohonen. Self-organized formation of topographically correct feature maps. *Biol. Cybern.*, Vol. 43, pp. 56–69, 1982.

[29] T. Kohonen. *Self-organization and associative memory*, Vol. 8 of *information sciences*. Springer, 1985.

[30] 栗田多喜夫. 情報量基準による 3 層ニューラルネットの隠れ層ユニット数の決定法. 電子情報通信学会論文誌, Vol. J73-D-II, pp. 1872–1878, 1990.

[31] E. Levin, N. Tishby, and S. A. Solla. A statistical approach to learning and generalization in layered neural networks. *Proc. of IEEE*, Vol. 78, No. 10, pp. 1568–1574, 1990.

[32] M. Minsky and S. Papert. *Perceptrons – An Introduction to Computational Geometry (Expanded Edition)*. MIT Press, 1988.

[33] 森田啓義. 算術符号から MDL 規準へ. 数理科学, No. 290, pp. 25–31, 1987.

[34] N. Murata, K. Doya, and S. Yoshizawa. Learning spatiotemporal patterns in a neural network with lateral inhibitory connections. In M. Caudill, editor, *Proc. of the International Joint Conference on Neural Networks*, Vol. 1, pp. 177–180. Lawrence Erlbaum Associates, Publishers, January 1990.

[35] N. Murata, S. Yoshizawa, and S. Amari. A criterion for determining the number of parameters for an artificial neural network model. unpublished, 1991.

[36] N. Murata, S. Yoshizawa, and S. Amari. A criterion for determining the number of parameters in an artificial neural network model. In T. Kohonen, et al., editors, *Artificial Neural Networks*, pp. 9–14. Elsevier Science Publishers, July 1991.

[37] 村田昇, 銅谷賢治, 吉澤修治. 非線形関数を記憶する神経回路網. 第 27 回計測自動制御学会講演会予稿集, pp. 361–362, 1988.

[38] 村田昇, 銅谷賢治, 吉澤修治. 時間パターンを学習する神経回路網モデル. 電子情報通信学会技術研究報告, Vol. NC89-11, pp. 1–8, 1989.

[39] 中村嘉男, 酒田英夫編. 脳の科学, I, II. 朝倉書店, 1983.

[40] N. J. Nilsson. *Learning Machines*. McGraw-Hill, 1965.

[41] 野田淳彦, 南雲仁一. システムの学習的同定法. 計測と制御, Vol. 7, No. 9, pp. 1–9, 1968.

[42] M. Opper and D. Haussler. Calculation of the learning curve of Bayes optimal classification algorithm for learning a Perceptron with noise. to appear, 1991.

[43] J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, Vol. 14, pp. 1080–1100, 1986.

[44] J. Rissanen and Bin Yu. MDL learning, 1991.

[45] H. Ritter and K. Schultem. Convergence properties of Kohonen's topology conserving maps: Fluctuations, stability, and dimension selection. *Biol. Cybern.*, Vol. 60, pp. 59–71, 1988.

[46] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, Vol. 22, pp. 400–4007, 1951.

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

[48] D. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. Rumelhart, J. L. McClelland, and the PDP Reserch Group, editors, *Parallel Distributed Processing : Explorations in*

the Microstructure of Cognition, Vol. 1: Foundations, chapter 8, pp. 318–362. The MIT Press, 1986.

[49] D. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, Vol. 323, pp. 533–536, 1986.

[50] Y. Sakaguchi. Topographic organization of nerve field with teacher signal. *Neural Networks*, Vol. 3, pp. 411–421, 1990.

[51] 阪口豊, 村田昇. 出力信号の構造を反映した競合学習型神経回路モデル – 競合的な中間層をもつ多層神経回路による非線型関数の実現 –. 電子情報通信学会技術研究報告, Vol. NC89-54, pp. 33–38, 1990.

[52] J. S. Albus, 亀井宏行 訳)(小杉幸夫. ロボティクス – ニューロンから知能ロボットへ. 啓学出版, 1984.

[53] 佐々木重夫. 微分幾何学. 岩波基礎数学選書. 岩波書店, 1991.

[54] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples I. general formulation and annealed approximation. Submitted to Physical Review A, 1991.

[55] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples II. quenched theory and unrealizable rules. Submitted to Physical Review A, 1991.

[56] H. Sompolinsky, S. Seung, and N. Tishby. Learning from examples in large neural networks. to be published, 1991.

[57] 竹内啓. 数理統計学 – データ解析の方法 –. 東洋経済, 1963.

[58] 竹内啓. 情報統計量の分布とモデルの適切さの規準. 数理科学, No. 153, pp. 12–18, 1976.

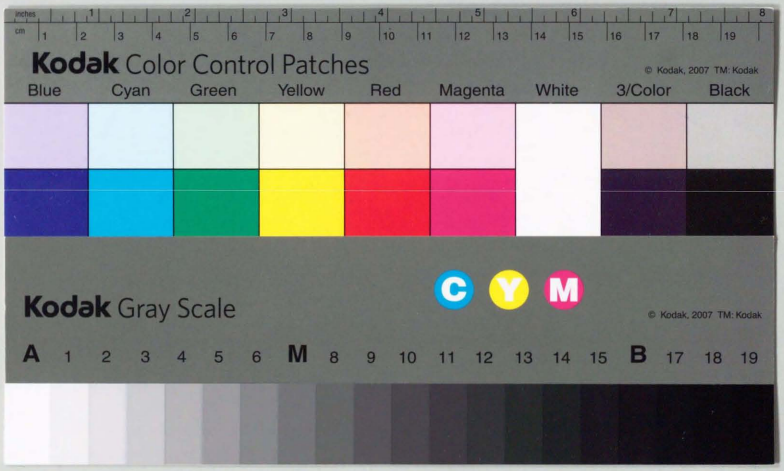
[59] 田中宏一良, 中野馨. 適応スプランフィルタによるマニピュレータの制御. 電子情報通信学会技術研究報告, Vol. NC89-47, pp. 29–34, 1989.

[60] L. G. Valiant. A theory of the learnable. *Comm. ACM*, Vol. 27, No. 11, pp. 1134–1142, 1984.

[61] 和田安弘, 川人光男. 新しい情報量規準と Cross Validation による汎化能力の推定. 電子通信学会論文誌, Vol. J74-D-II, No. 7, pp. 955–965, 1991.

[62] D. H. Wolpert. A mathematical theory of generalization: Part II. *Complex Systems*, Vol. 4, pp. 201–249, 1990.

[63] K. Yamanishi. A loss bound model for on-line stochastic prediction algorithms. Submitted to Information and Computation, 1991.


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