混沌系工学特論 配布資料 #7

担当：井上 純一 (情報科学研究科棟 8-13)
URL：http://chaosweb.complex.eng.hokudai.ac.jp/~j_inoue/

平成 17 年 1 月 17 日

目 次

3 例題からの学習と汎化 **101**
3.1 学習機械パーセプトロン 102
3.1.1 拡張 Hebb 則とその収束性 102
3.1.2 パーセプトロンの収束定理 102
3.2 教師機械の導入 104
3.3 学習の 2 つの指標 — 訓練誤差と汎化誤差 — 105
3.4 統計力学による性能評価 107
3.4.1 解空間とギブス学習 107
3.4.2 学習曲線の評価 108
3.4.3 例題数無限大での漸近評価 111

3 例題からの学習と汎化

ここからは神経回路網の**例題からの学習 (learning from examples)** を統計力学の方法を用いて調べて行くことにする. この「学習」を既に述べたように「外界からの信号 (環境) に応じて自らの構造を変化させていくこと」と定義するのであれば, 神経回路網の場合には素子間の結合を変えること (調節すること) がここでの学習に相当する. そのときの学習の仕方として大きく次の 2 種類 :

- **教師あり学習 (supervised learning)**
- **教師なし学習 (un-supervised learning)**

が考えられる. 前者は回路網に与える入力信号と, それに対する望ましい出力が陽に与えられる場合の学習であり, この入出力関係のことを**例題 (examples, あるいは training sets)** と呼ぶ. 一方, そのような例題が陽に与えられない学習を考えることもできる. 例えば, [**連想記憶の数理**] のときに学んだシナプス結合の学習では, 記憶させたいパターン $\boldsymbol{\xi}^\mu$ を各結合 w_{ij} に分散的に蓄えさせる段階で

$$w_{ij}^{\mu+1} \quad = \quad w_{ij}^{\mu} + \frac{1}{N}\,\xi_i^\mu \xi_j^\mu \tag{1}$$

のような学習則を採用した. この各々のパターン $\boldsymbol{\xi}^\mu$ に対する望ましい出力 (回路網の状態) : $\boldsymbol{S}$ はそのパターン $\boldsymbol{\xi}^\mu$ そのものであるが, そのこと自体を学習則 : (1) 式は用いているわけではない (実際の回路網の出力値に応じて結合を変えているわけではなく, パターンの個数だけ (1) 式で $\xi_i^\mu \xi_j^\mu$ を足しあげているだけ

である). このような学習を**教師なし学習**と呼んでいる. ここでは主に前者の教師あり学習を考えることにする.

また, 与えられる例題の用い方として

- **バッチ学習 (Batch learning)**
- **オンライン学習 (On-line learninig)**

の 2 つの様式が考えられる. 前者は学習機械がいくつかの例題をまとめて受け取り, その例題全てに正解を与えるように自分の結合を変えていくような学習様式であり, 一方, 後者は学習機械に一つずつ例題が与えられ, その例題に対する結合変化を終えた段階でその例題は破棄され, 新たな例題が一つ与えられる, というように逐次的に学習が進んで行く. ここでははじめにバッチ学習について学び, その後にオンライン学習について見ていくことにしょう.

計算時間 (処理時間) の観点から言って, ロボット制御などに神経回路網の学習を適用する際には, いくつかの例題をかき集めてその学習機械に与え, その全てに正解するように結合を調節することは, リアルタイム処理が重要となるこの種の課題には向いておらず, オンライン型の学習を用いた方が好ましい. しかし, 一般的にオンライン学習は後に見る「汎化誤差」という指標で評価される精度がバッチ学習に比べて落ちる場合が多く, 学習則の工夫等が必要となってくる.

3.1 学習機械パーセプトロン

まずは代表的な学習機械である**パーセプトロン**について復習し, その学習を簡単に述べよう. パーセプトロン自体は [**連想記憶の数理**] で既に学んだように, 出力 y が入力ベクトル $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)$ に対して

$$y \quad = \quad \mathrm{sgn}(\boldsymbol{w} \cdot \boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N} w_i x_i\right) \tag{2}$$

で与えられるものである. このパーセプトロンに対して, 入出力の組：$(\boldsymbol{x}, y)$ を与え, それを実現させるようにある規則によって結合を変化させることをここでの学習と呼ぶとすれば, どのような規則が存在し, その規則で状態更新した結合が有限の状態更新で収束するのか否かを確認しておくことは意味のあることであろう. そこで, ここでは**拡張 Hebb 則**という学習則を導入し, それの収束性を議論してみることにしょう.

3.1.1 拡張 Hebb 則とその収束性

入出力 $(\boldsymbol{x}, y)$ に対して, パーセプトロンの結合ベクトル $\boldsymbol{w}$ は次の更新式に従うものとする.

$$\boldsymbol{w}(n+1) \quad = \quad \boldsymbol{w}(n) + y\,\boldsymbol{x} \tag{3}$$

つまり, その時点での入力 $\boldsymbol{x}$ に対応する出力の符号に応じて 1 ステップ手前の結合に入力ベクトルを足して (引いて) 行くわけである.

このとき, 更新式 (3) が収束するのか, 言い方を換えれば, 線形分離可能な望ましい入出力の組：$(\boldsymbol{x}, y)$ の全てに対して正しい結果を出すような結合を得るためにはどのくらいのステップ数 t_{need} が必要になるのかを調べることが重要になる.

3.1.2 パーセプトロンの収束定理

この問題を明らかにするため, まずは入力ベクトル $\boldsymbol{x}$ が $|\boldsymbol{x}| = \sqrt{x_1^2 + \cdots + x_N^2} = 1$ のように規格化されているものとし, 入出力の組：$(\boldsymbol{x}, y)$ の全てに対して正しい結果を出すような結合を $\boldsymbol{w}^*$ と名づけることに

しょう (図 1 参照). 簡単のため, この結合も $|\boldsymbol{w}^*| = \sqrt{(w_1^*)^2 + (w_N^*)^2} = 1$ のように規格化されているものとする. このとき, δ という量を次で定義しておく.

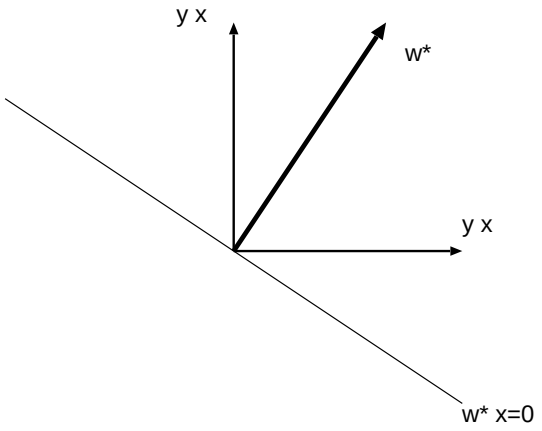

図 1: 結合 $\boldsymbol{w}^*$ は全ての入出力に対して, $y\,(\boldsymbol{w}^* \cdot \boldsymbol{x}) > 0$ となり, 正解を与える結合である.

$$\delta \quad = \quad \min_{(\boldsymbol{x},y)\in F}\{y \cdot (\boldsymbol{w}^* \cdot \boldsymbol{x})\} > 0 \tag{4}$$

幾何学的に見ると, この δ という量は任意の入力ベクトル $\boldsymbol{x}$ と「正解」「不正解」の分離面：$\boldsymbol{w}^* \cdot \boldsymbol{x} = 0$ 間の距離のうち, 最も短いものとなっている. また, F は線形分離可能な入出力の集合であり, $\boldsymbol{w}^*$ は全ての入出力に対して正しい答えを出す「正解」であるから, 全ての $(\boldsymbol{x}, y) \in F$ に対して $y(\boldsymbol{w}^* \cdot \boldsymbol{x}) > 0$ であることに注意しておく (図 1 参照).

$\boldsymbol{w}(n)$ を n ステップ目の修正で得られる結合ベクトルとするのであれば

$$\begin{aligned} \boldsymbol{w}^* \cdot \boldsymbol{w}(i+1) \quad &= \quad \boldsymbol{w}^* \cdot (\boldsymbol{w}(i) + y\,\boldsymbol{x}) = \boldsymbol{w}^* \cdot \boldsymbol{w}(i) + y\,(\boldsymbol{w}^* \cdot \boldsymbol{x}) \\ &\geq \quad \boldsymbol{w}^* \cdot \boldsymbol{w}(i) + \delta \end{aligned} \tag{5}$$

が成立する. 従って, 上記の操作を $i = 0$ から $n-1$ まで繰り返し行えば

$$\boldsymbol{w}^* \cdot \boldsymbol{w}(n) \quad \geq \quad n\delta + \epsilon_0 \tag{6}$$

が満たされる. ここで, $\epsilon_0 = \boldsymbol{w}^* \cdot \boldsymbol{w}(0)$ である. これは学習スタート時における正解と学習機械の結合の内積である. 以下では $\epsilon_0 = 0$ として議論を進める. これは $\boldsymbol{w}(0) = 0$ であると考えても良いが, $\boldsymbol{w}(0)$ が有限の場合でも, $\epsilon_0 = 0$ は正解から最も遠い初期条件から学習をスタートさせることを意味するので, この最も難しい場合に関して収束までに必要なステップ数を評価すると考えれば我々の目的に即している. しかし, 以下では簡単のため $\boldsymbol{w}(0) = 0$ として話を進めよう.

このとき $|\boldsymbol{w}(i)|^2$ は $y\,(\boldsymbol{w}(i) \cdot \boldsymbol{x})$ は常に正であることを考えれば学習の状態更新式 (3) に対して

$$\begin{aligned} |\boldsymbol{w}(i+1)|^2 \quad &= \quad |\boldsymbol{w}(i) + y\,\boldsymbol{x}|^2 = |\boldsymbol{w}(i)|^2 + 2y\,(\boldsymbol{w}(i) \cdot \boldsymbol{x}) + y^2|\boldsymbol{x}|^2 \\ &< \quad |\boldsymbol{w}(i)|^2 + 1 \end{aligned} \tag{7}$$

となり, 従って, これを繰り返し用いることにより

$$|\boldsymbol{w}_n| \quad < \quad \sqrt{n} \tag{8}$$

が成り立つ. ところで, $|\boldsymbol{w}^*| = 1$ であるから, 任意のベクトル $\boldsymbol{w}$ に対し,

$$G(\boldsymbol{w}) \quad = \quad \frac{\boldsymbol{w}^* \cdot \boldsymbol{w}}{|\boldsymbol{w}|} \leq 1 \tag{9}$$

が成り立つことに注意すれば, (6)(8) 式から

$$G(\boldsymbol{w}(n)) \quad = \quad \frac{\boldsymbol{w}^* \cdot \boldsymbol{w}(n)}{|\boldsymbol{w}(n)|} \geq \sqrt{n}\delta \tag{10}$$

となる. 従って, $\sqrt{n}\delta > 1$, すなわち $n > \delta^{-2}$ であれば $G(\boldsymbol{w}^n) > 1$ となり G は常に 1 以下であることに矛盾してしまうので, $n \leq t_{\rm need} = \delta^{-2}$ である. 従って, パーセプトロンについての学習則が収束するためには, どんなに多くとも $t_{\rm need} = \delta^{-2}$ ステップあれば良く, アルゴリズムはこのステップ内で正解を得て終了する.

ところで, 上記のように全ての例題に対して正解を与える結合の更新式に関する反復数がわかったとは言うものの, 得られた結合が一体何を意味するのかということが気になる. そもそも, 全ての可能なパターン 2^N 個の中でどのように線形分離なパターンを選んだらよいのであろうか？

3.2　教師機械の導入

ここでは簡単のため, 入力層と出力層のみからなる**単純パーセプトロン (simple perceptron)** を考える. また, 学習の様式として教師機械[1]　及び生徒機械を用意し, 生徒機械が教師機械から学ぶ, 教師あり学習を考えることにする. このように, 教師機械であるパーセプトロンからの入出力 $(\sigma_T, \boldsymbol{\xi})$ を例題とするのであれば, 生徒機械が受取る例題は全て線形分離可能なものであることは明らかである. なお, 全ての可能な入出力の数 2^N 個の中のいくつがパーセプトロンの実現できる入出力なのかが気になるところであるが, グラフ理論, 組み合わせ理論, 統計力学的手法など様々なアプローチにより, およそ $2N$ 個であることがわかっている[2] .

さて, 入力数が N の単純パーセプトロンはその結合ベクトルのみで特徴付けられるのであるから, 教師機械の結合ベクトルを $\boldsymbol{T} = (T_1, T_2, \cdots, T_N)$ とし, 生徒機械の結合ベクトルを $\boldsymbol{J} = (J_1, J_2, \cdots, J_N)$ と置こう. このとき教師あり学習の枠組みでは, 生徒機械は教師の提示する P 個の入出力の組 : $\{\boldsymbol{\xi}^\mu, \sigma_T^\mu\}, \quad \mu = 1, \cdots, P$ から自分の結合ベクトルを変更して行く (学習して行く). ここで $\boldsymbol{\xi}^\mu$ 及び σ_T^μ は $\mathrm{sgn}(\cdots)$ を符号関数として

$$\sigma_T^\mu \quad = \quad \mathrm{sgn}\left(\boldsymbol{T} \cdot \boldsymbol{\xi}^\mu\right) = \left(\sum_{i=1}^{N} T_i \xi_i^\mu\right) \tag{11}$$

を満たさなければならない.

このとき生徒機械の行う学習は, 例題 $\{\boldsymbol{\xi}^\mu, \sigma_T^\mu\}$ から結合 $\boldsymbol{J}$ を決めることであり, P 個の例題全てに対して正しい答えを出す結合 $\boldsymbol{J}$ を学習により獲得できたとすれば

$$\sigma^\mu \quad = \quad \mathrm{sgn}\left(\boldsymbol{J} \cdot \boldsymbol{\xi}^\mu\right) = \left(\sum_{i=1}^{N} J_i \xi_i^\mu\right) \tag{12}$$

に対して

$$\sigma^\mu \quad = \quad \sigma_T^\mu \quad \forall \ (\mu \in 1, \cdots, P) \tag{13}$$

が成立する (図 2 参照).

[1] 「機械」という言葉を頻繁に用いるが, この講義では全て単純パーセプトロンを指していると思ってよい.
[2] このことから, 教師機械の出力からではなく, ランダムに線形分離可能なパターンを選び出すことがいかに難しいかがわかるであろう. そのようなパターンが選び出される確率は $2N/2^N$ であり, 入力次元が大きくなればこの確率は限りなくゼロである.

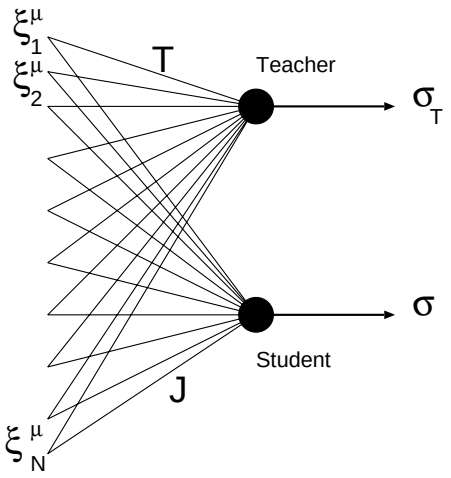

図 2: ここで扱う「教師あり学習」. 教師, 生徒共に単純パーセプトロンであり, 入力 $\boldsymbol{\xi}^\mu$ に対し, $\sigma_T = \sigma$ となるように生徒機械は結合 $\boldsymbol{J}$ を調節して行く.

3.3　学習の 2 つの指標 — 訓練誤差と汎化誤差 —

ここまでで簡単にではあるが, 教師あり学習の問題設定を行った. 次にすべきなのは, 学習がどの程度巧く行われたかに関する指標を決めることである. この指標の一つは**訓練誤差 (training error)** と呼ばれるものであり

$$\epsilon_t(\boldsymbol{J};\{\boldsymbol{\xi}^\mu\},\boldsymbol{T}) = \frac{1}{P}\sum_{\mu=1}^{P} d(\boldsymbol{J};\{\boldsymbol{\xi}^\mu\},\boldsymbol{T}) \tag{14}$$

$$d(\boldsymbol{J};\{\boldsymbol{S}\},\boldsymbol{T}) = \Theta(-\sigma_T\sigma) \tag{15}$$

で与えられる. ここで $\Theta(\cdots)$ は階段関数である. これは見ての通り, 与えられた例題に対する誤り率である.

しかし, この訓練誤差のみからは生徒機械がどの程度巧く教師機械を「模倣できた」かは解らない. 特に例題数が少ない場合には, **その少数の例題に対してのみ正しい入出力を返す機械が出来上がっている可能性がある**からである (図 3 参照.). そこで有限個の例題に対して結合 $\boldsymbol{J}$ を獲得した回路網に対し, その**汎化誤差 (generalization error)** と呼ばれる量 :

$$\epsilon(\boldsymbol{J};\boldsymbol{T}) = \sum_{\{\boldsymbol{S}\}} P_S(\boldsymbol{S}) d(\boldsymbol{J};\{\boldsymbol{S}\},\boldsymbol{T}) \tag{16}$$

を導入する. ここで, 入力ベクトルに関する分布 $P_S(\boldsymbol{S})$ で d の平均をとっているのは, **提示された例題を含む全ての入出力に対する誤差**を学習の指標とするためである[3] .

訓練誤差 ϵ_t と汎化誤差 ϵ の定義より, 訓練誤差は提示された有限個の例題 $\boldsymbol{\xi}^\mu$ から求められた汎化誤差の「推定値」とみなすことができる. また, この両者の差は例題数の増加とともに (確率的に) 小さくなることが予想される. つまり, ある正の数 δ に対して, 次のように定義される確率 $\mathcal{P}$:

$$\mathcal{P} = \mathrm{Prob}(|\epsilon_t - \epsilon| > \delta) \tag{17}$$

は例題数 P の増加とともに減少して行くことになる.

[3] 真の意味で平均的なパフォーマンスを求めるためには結合 $\boldsymbol{J}$ 及び $\boldsymbol{T}$ に関する分布 $P_J(\boldsymbol{J}), P_T(\boldsymbol{T})$ に関する平均操作も行うべきであることを注意しておく. 例えばこれらの結合が一様分布から生成されたものであれば $P_S(\boldsymbol{J}) = \delta(\boldsymbol{J}^2 - N), P_T(\boldsymbol{T}) = \delta(\boldsymbol{T}^2 - N)$ と置くことになる.

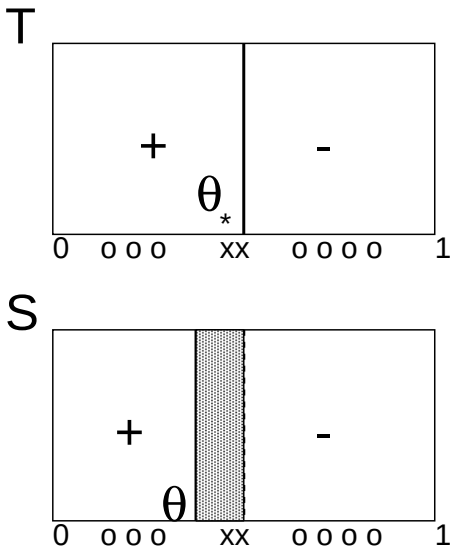

図 3: 簡単な 1 次元閾値入出力機の学習. 教師機械はパラメータ θ_* を有し, 入力 x が $0 \leq x < \theta_*$ の場合には $+1$ を, $\theta_* \leq x \leq 1$ の場合には -1 を返すとする. このとき, この入出力関係から生徒機械は可変なパラメータ θ を調節してゆくのだが, これがこの機械の「学習」である. 図のような θ を生徒機械が学習により獲得したとすると, 丸印を付けた例題に対しては確かに教師機械と同じ出力を返してうまくいっているように見えるが, しかし, バツ印をつけた問題に対しては間違った答えを返す (網模様部分に落ちた入力に対しては全て間違う). 有限個の例題の学習に成功しているからといって, それが必ずしも信頼のおける機械であるとは言えないことの一例である.

問 10 :
図 3 の教師機械, 生徒機械の t 番目の入力ベクトル $x(t)$ に対する出力 $\sigma_T(t), \sigma_S(t)$ は式で書けば

$$\sigma_T(t) = \mathrm{sgn}[\theta_* - x(t)] \tag{18}$$
$$\sigma_S(t) = \mathrm{sgn}[\theta(t) - x(t)] \tag{19}$$

となる. そこで, 生徒機械の持つパラメータ $\theta(t)$ の学習則を

$$\theta(t+1) = \theta(t) - \lambda\, \sigma_T(t) \Theta[-\sigma_T(t)) \sigma_S(t)] \tag{20}$$

で与えることにしょう. ここに $\Theta(\cdots)$ は階段関数である. これは生徒機械/教師機械の出力値が異なる場合のみ θ の状態更新が行われ, 更新される場合には教師機械の出力とは逆符号で λ だけ θ が調整されることを意味する.
そこで, 各ステップ t で入力 $x(t)$ を $[0,1]$ の一様乱数として選び, 教師/生徒両機械に与えることにより, 更新式 (20) を計算機上でシミュレートし, $\theta(t)$ がどのような確率過程に従うのかを $\theta(t)$ をプロットすることにより確かめよ. また, 学習の各ステップでエラー : $E(t) = (\theta_* - \theta(t))^2$ を測定し, この振る舞いも同様にプロットせよ. ここで, 学習係数 λ は小さな値で一定値に選んでも, λ 自体にもステップ依存性を持たせ, $\lambda = \lambda_0\, \mathrm{e}^{-at}$ のような「減衰項」として扱っても良い. この選び方は各自の考察に任せるが, 少なくとも

(a) $\lambda =$ **比較的小さな一定値**
(b) $\lambda = \lambda_0\, \mathrm{e}^{-at}$

の 2 通りは必ず確かめてみること.
なお, 正解 θ_*, 及び, 初期値 $\theta(0)$ は $[0,1]$ の間で各自が適当に設定せよ.

3.4 統計力学による性能評価

以上で教師あり学習の枠組みで学習と汎化を定義した．ある学習則の下に出来上がった学習機械が「どの程度信頼のおけるものであるかどうか」の評価指標は汎化誤差により与えられる．従って汎化誤差の評価, つまり, 例題数の増加と共に汎化誤差がどのように振舞うか — **学習曲線 (learning curve)** — を求めることは, 学習理論におけるメインテーマの一つである[4]．そこで, ここではこの学習曲線を統計力学的手法に基づいて評価する手続きを紹介することにしょう．ただし, 以下の議論では入力次元 (結合ベクトルの次元) N と例題数 P が両者とも無限大に取れる場合を扱い, 両者の比を α とすると $P, N \to \infty$ で

$$\alpha = \frac{P}{N} \simeq \mathcal{O}(1) \tag{21}$$

であることを仮定して議論を進める．

3.4.1 解空間とギブス学習

簡単のため生徒機械, 教師機械の結合ベクトルは次の規格化条件：

$$\boldsymbol{J}^2 = \sum_{i=1}^{N} J_i^2 = N \tag{22}$$

$$\boldsymbol{T}^2 = \sum_{i=1}^{N} T_i^2 = N \tag{23}$$

を満たしているものとする．これは言い方を変えれば, $\boldsymbol{J}, \boldsymbol{T}$ がデルタ関数 $\delta(\cdots)$ を用いて分布：

$$P_J(\boldsymbol{J}) = \delta\left(\sum_{i=1}^{N} J_i^2 - N\right) \tag{24}$$

$$P_T(\boldsymbol{T}) = \delta\left(\sum_{i=1}^{N} T_i^2 - N\right) \tag{25}$$

から生成されたということである．この場合, N 次元球面上に $\boldsymbol{J}, \boldsymbol{T}$ があるということだから, 汎化誤差は球の中心を通り $\boldsymbol{J}$ に垂直な面と, 球の中心を通り $\boldsymbol{T}$ に垂直な面に囲まれる部分の全球面に対する比で与えられる[5]．従って, $\boldsymbol{J}$ と $\boldsymbol{T}$ のなす角度を θ とすれば

$$\epsilon = \frac{\theta}{\pi} \tag{26}$$

となる．これはまた, 結合ベクトル $\boldsymbol{J}, \boldsymbol{T}$ の内積を入力数 N で割ったものを R とすると

$$\begin{aligned} R &= \frac{1}{N}(\boldsymbol{J}\cdot\boldsymbol{T}) = \frac{1}{N}\sum_{i=1}^{N} J_i T_i \\ &= \frac{1}{N}|\boldsymbol{J}||\boldsymbol{T}| = \frac{1}{N}\sqrt{N}\sqrt{N}\cos\theta = \cos\theta \end{aligned} \tag{27}$$

と書けるから, 汎化誤差はこの R を用いて

$$\epsilon = \frac{1}{\pi}\mathrm{arc}\cos R \tag{28}$$

と書き直すこともできる．

[4] 「学習曲線」とはもともと心理学から来た用語らしい．
[5] 分かりづらい場合には 2 次元の場合を図示してみること．図 4 のようになります．

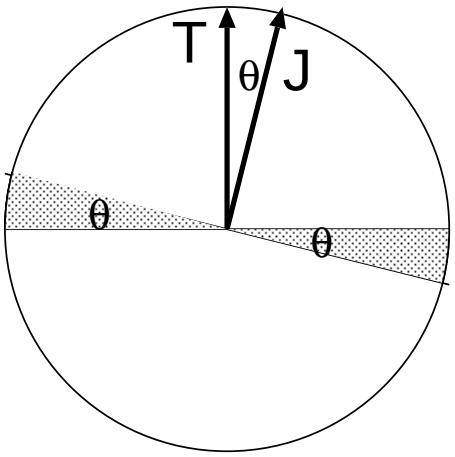

図 4: 解空間. 斜線部に落ちた例題に関しては誤った答えを出す.

ところで, ある有限個の例題の答えを正確に出力することのできる $\boldsymbol{J}$ は (22) 式を満たす空間の中のある有限部分空間 (**version space**, あるいは**解空間**と呼ばれる) を作るが, この中で, どの $\boldsymbol{J}$ を選ぶかによって異なる汎化能力を持つ機械が出来上がる. この解空間において一つの $\boldsymbol{J}$ をランダムに選ぶ戦略 (学習則) を**ギブス学習 (Gibbs learning)** と呼ぶ. 次の節ではギブス学習の学習曲線を具体的に求めてみることにしょう.

3.4.2 学習曲線の評価

前節の最後で, ある例題に対して正解を与える $\boldsymbol{J}$ は (22) 式を満たす全空間の中で, 「解空間」と呼ばれる部分空間を形作ることを見た. 当然のことながら, この解空間の「体積」は例題数の増加とともに減少して行く (図 5 参照.) [6] . そこで, $\Omega_0(\epsilon)$ を $\boldsymbol{J}, \boldsymbol{T}$ の内積が $R = \cos(\pi\epsilon)$ のときの $\boldsymbol{J}$ 空間の体積とすれば, 生徒機械と教師機械が同じ出力をする確率は ϵ の定義から $1-\epsilon$ であるわけだから, P 個の例題が独立に与えられたとすれば, その時点での解空間の体積 $\Omega_P(\epsilon)$ は

$$\Omega_P(\epsilon) \quad = \quad \Omega_{P-1}(\epsilon)(1-\epsilon) = \Omega_{P-2}(\epsilon)(1-\epsilon)^2 = \cdots = \Omega_0(\epsilon)(1-\epsilon)^P \tag{29}$$

で与えられる. 実際, $0 < \epsilon < 1$ であるから $1-\epsilon < 1$ であり, $\Omega_P(\epsilon)$ は例題数の増加とともに減少して行くことに注意しょう.

さて, $\Omega_P(\epsilon)$ を $N, P \to \infty$ の極限で評価することになる. まず, 定義より $\Omega_0(\epsilon)$ は

$$\Omega_0(\epsilon) \quad = \quad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} d\boldsymbol{J}\delta(\boldsymbol{J}^2 - N)\delta\left(\frac{1}{N}(\boldsymbol{J}\cdot\boldsymbol{T}) - \cos(\pi\epsilon)\right) \tag{30}$$

と書ける. ただし, $d\boldsymbol{J} \equiv \prod_j dJ_j$ と定義した. この式の中のデルタ関数をフーリエ変換で書き直すと

$$\delta(\boldsymbol{J}^2 - N) \quad = \quad \int_{-\infty}^{\infty} \frac{d\hat{J}}{2\pi} \mathrm{e}^{i\hat{J}(\sum_j J_j^2 - N)} \tag{31}$$

$$\delta\left(\frac{1}{N}(\boldsymbol{J}\cdot\boldsymbol{T}) - \cos(\pi\epsilon)\right) \quad = \quad \int_{-\infty}^{\infty} \frac{d\hat{T}}{2\pi/N} \mathrm{e}^{i\hat{T}(\sum_j J_j T_j - N\cos\pi\epsilon)} \tag{32}$$

[6] 極端な場合として, 無限個の例題に対し正解を与えることのできる $\boldsymbol{J}$ は解空間の中で $\boldsymbol{J} = \boldsymbol{T}$ となる 1 点にまで収縮していることを確認してみるとよい.

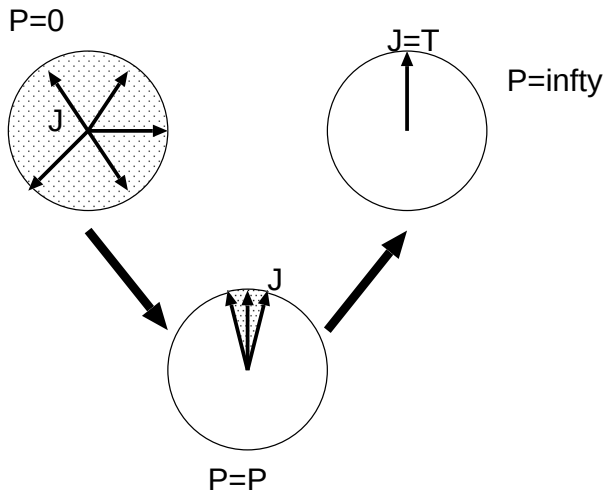

図 5: 解空間は例題数の増加とともに収縮して行く.

となるから, (30) 式は

$$\begin{aligned}
\Omega_0(\epsilon) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-i\infty}^{i\infty}\int_{-i\infty}^{i\infty}\frac{d\boldsymbol{J}d\hat{J}d\hat{T}}{(2\pi)^2/N}\exp\left(-\hat{J}\sum_j(J_j)^2-\hat{T}\sum_j J_jT_j+N\hat{J}+N\hat{T}\cos\pi\epsilon\right)\\
&= \int_{-i\infty}^{i\infty}\int_{-i\infty}^{i\infty}\frac{d\hat{J}d\hat{T}}{(2\pi i)^2/N}\,\mathrm{e}^{N\{\hat{J}+\hat{T}\cos(\pi\epsilon)\}}\prod_{j=1}^{N}\int_{-\infty}^{\infty}dJ_j\exp\left[-\hat{J}(J_j)^2+\hat{T}T_jJ_j\right]\\
&= \int_{-i\infty}^{i\infty}\int_{-i\infty}^{i\infty}\frac{d\hat{J}d\hat{T}}{(2\pi i)^2/N}\,\mathrm{e}^{N\{\hat{J}+\hat{T}\cos(\pi\epsilon)\}}\exp\left[\frac{\hat{T}^2}{4\hat{J}^2}\sum_{j=1}^{N}(T_j)^2\right]\prod_{j=1}^{N}\sqrt{\frac{\pi}{\hat{J}}} \qquad (33)
\end{aligned}$$

と書き換えることができる. ただし, $\hat{J}\to i\hat{J},\hat{T}\to i\hat{T}$ なる変換を施し, この中の J_j に関するガウス積分を各 j 毎に独立に実行した. すると, 最終的に $\Omega_0(\epsilon)$ は

$$\Omega_0(\epsilon) = \int_{-i\infty}^{i\infty}\int_{-i\infty}^{i\infty}\frac{d\hat{J}d\hat{T}}{(2\pi)^2}\exp\left(N\Phi(\hat{J},\hat{T})\right) \qquad (34)$$

$$\Phi(\hat{J},\hat{T}) = \frac{1}{2}\log(\pi/\hat{J})+\frac{\hat{T}^2}{4\hat{J}}+\hat{J}+\hat{T}\cos\pi\epsilon \qquad (35)$$

のように書ける. 残る $\hat{J},\hat{T}$ に関する積分はそのままでは実行できないが, P,N が無限大の極限において, これらの積分は既に説明してある鞍点法で評価できる. つまり

$$\frac{\partial\Phi}{\partial\hat{J}}=0,\qquad \frac{\partial\Phi}{\partial\hat{T}}=0 \qquad (36)$$

を満たす $\hat{J},\hat{T}$ を $\hat{J}_*,\hat{T}_*$ とすれば積分はこの点 $(\hat{J}_*,\hat{T}_*)$ での被積分関数の値

$$\Omega_0(\epsilon) = \exp\left(N\Phi(\hat{J}_*,\hat{T}_*)\right) \qquad (37)$$

で評価できる. 実際に (36) 式を用いて鞍点を求めてみると

$$\hat{J}_* = \frac{1}{2}\frac{1}{1-\cos^2\pi\epsilon} \qquad (38)$$

$$\hat{T}_* = -\frac{\cos\pi\epsilon}{1-\cos^2\pi\epsilon} \qquad (39)$$

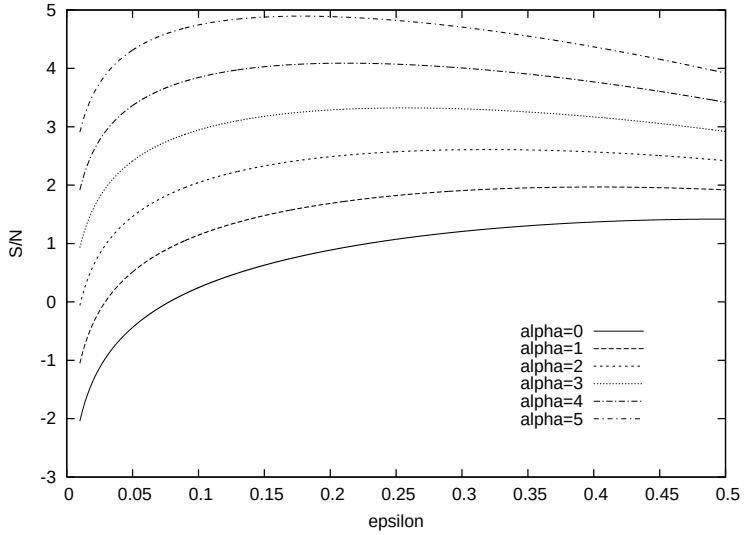

図 6: 上から $\alpha = 0, 1, 2, 3, 4, 5$ に対する単位入力当たりのエントロピー S/N の汎化誤差 ϵ 依存性. α の増加と共に, S/N の極値を与える ϵ は 0.5 から単調に減少していることに注意しょう. また, エントロピーが「状態数の対数」であるとするならば, 状態数の最小値は 1 であるから, エントロピーの最小値は $\log 1 = 0$ のはずである. しかし, このグラフをみると, エントロピーが負にもなり得ている. この理由は状態空間 (解空間) の体積の対数をエントロピーとみなしたため, $|\Omega_P(\epsilon)| < 1$, つまり, $S < 0$ となり得ることがあるからである.

となるから，これを (35) に代入して

$$\Omega_0(\epsilon) \quad = \quad \exp\left(\frac{N}{2}[1 + \log 2\pi + \log \sin^2 \pi\epsilon]\right) \tag{40}$$

が得られる. 従って解空間の体積 $\Omega_P(\epsilon)$ は (29) 式から

$$\begin{aligned}
\Omega_P(\epsilon) &= \exp\left(\frac{N}{2}[1 + \log 2\pi + \log \sin^2 \pi\epsilon]\right) \mathrm{e}^{N\alpha \log(1-\epsilon)} \\
&= \exp\left(N\left[\frac{1}{2}(1 + \log 2\pi) + \frac{1}{2}\log \sin^2(\pi\epsilon) + \alpha \log(1-\epsilon)\right]\right) \\
&\equiv \exp\left(S(\epsilon, \alpha)\right) \qquad (41) \\
S(\epsilon, \alpha) &\equiv \frac{1}{2}(1 + \log 2\pi) + \frac{1}{2}\log \sin^2(\pi\epsilon) + \alpha \log(1-\epsilon) \qquad (42)
\end{aligned}$$

と決定される ($P = N\alpha$ に注意).

$\Omega_P(\epsilon)$ は汎化誤差 ϵ を与える解空間の体積 — 大雑把には $\boldsymbol{J}$ の取りうる場合の数[7] — であるから, これの対数をとることにより得られる $S(\epsilon, \alpha)$ は**解空間のエントロピー**に相当する量である. 図 6 にいくつかの α の値に対して, 単位入力数あたりのエントロピー, つまり S/N をプロットしたものを載せる. 各 α の値に対してエントロピーはある値 ϵ で極大値を持つが, 極値を与える ϵ_* とそれ以外の ϵ のエントロピーの差は $\mathcal{O}(N)$ であることから, 前に述べた「アンサンブル」の考え方に従えば, $N \to \infty$ の極限では圧倒的割合で汎化誤差 ϵ_* を持つ学習機械が現ることになる. それ以外の汎化誤差を持つ学習機械も出現することはするが, それは極めて希なことであるというわけである. よって, 各 α に対して S/N の極値を与える ϵ をプロットした $\epsilon(\alpha)$ がギブス学習の学習曲線となる. 式で書けば

$$\epsilon(\alpha) \quad = \quad \arg\max_{\epsilon}\left[\frac{1}{2}\log \sin^2(\pi\epsilon) + \alpha \log(1-\epsilon)\right] \tag{43}$$

[7] $\boldsymbol{J}$ は連続値をとる変数だから, 「個数」とするのは大雑把な見方である.

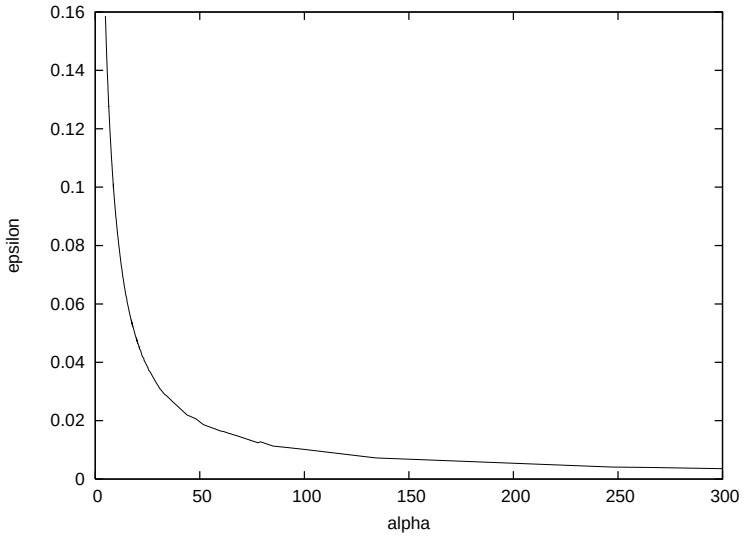

図 7: エントロピー最大化条件 (45) を数値的に解くことにより求めた学習曲線. 例題数 α の増加とともに汎化誤差は単調に減少する.

である.

ところで図 6 より, $\alpha = 0$, つまり, 例題を与えていない状況では $\epsilon = 1/2$ となり, 所謂 **random guess** と同じ性能の学習機械が得られることになり, 直観と合っている. また, 例題数 α の増加とともに, S/N の極値を与える ϵ は単調に減少することも分かり, これも直観と合っている. これらを統合すると, 例題数の増加とともに汎化誤差は 0.5 から単調に減少するというのがギブス学習の学習曲線であることが分かる.

3.4.3 例題数無限大での漸近評価

(43) 式を具体的な極値条件として書き下せば

$$\frac{\pi \cos(\pi\epsilon)}{\sin(\pi\epsilon)}(1-\epsilon) = \alpha, \tag{44}$$

つまり,

$$1-\epsilon \quad = \quad \frac{\alpha}{\pi}\tan(\pi\epsilon) \tag{45}$$

が得られる. 図 7 に (45) 式を数値的に解き, ϵ を α の関数としてプロットしたものを載せよう. この図より, 例題数 α の増加とともに ϵ は単調にゼロへと近づくことがわかる. ここで取り上げた例では原理的に教師機械の与える例題から $\epsilon = 0$ となる規則, つまり, $\boldsymbol{T}$ を見つけることができるという意味において「学習可能」な課題である. もちろん, 現実的には教師信号にノイズが加わった場合や, 生徒機械と学習機械の構造上のミスマッチがある場合の方が圧倒的に多いであろう. この場合, $\Omega_P(\epsilon)$ 自体が単調には減少しなくなり, ここに述べた解析が困難になるので, 適時工夫が必要となる.

さて, $\alpha \to \infty$ では汎化誤差は非常に小さくなっているはずだから，無限個の例題を与えたとき, $\epsilon \ll 1$ として (45) 式左辺を展開し, 主要項のみ残す. すると

$$1-\epsilon \quad = \quad \frac{\alpha}{\pi}\tan(\pi\epsilon) \simeq \frac{\alpha}{\pi}\pi\epsilon = \alpha\epsilon \tag{46}$$

つまり

$$\epsilon \quad = \quad \frac{1}{1+\alpha} \simeq \frac{1}{\alpha} \tag{47}$$

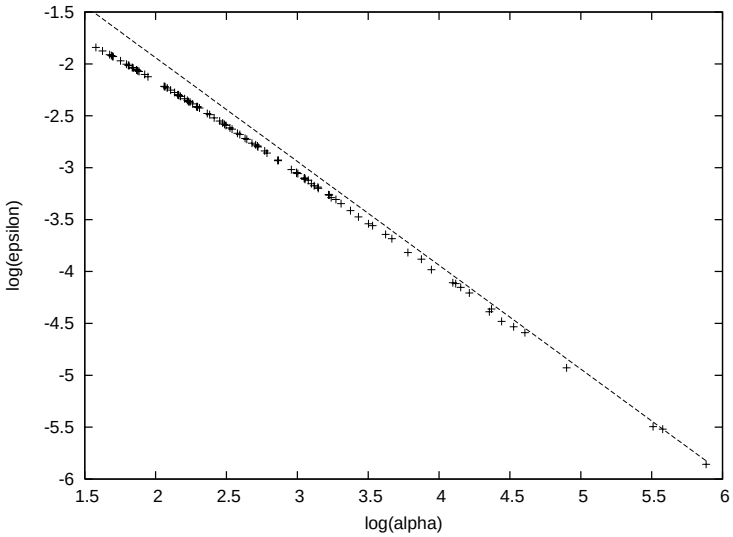

図 8: 学習曲線の α が十分大きい場合のスケーリング則. α が十分に大きければ α^{-1} 則に従う.

がギブス学習の学習曲線の例題数無限大での漸近形であり, 例題数の逆べきでゼロに漸近する. 図 8 に図 7 の数値計算で得た結果を対数プロットしたものと, 傾き -1 の直線を重ねて描く. この図より, α が大きければ (47) 式の漸近形に従うことが実際に確認できる.

$\alpha = \mathcal{O}(1)$ の領域での学習曲線は個々の学習機械の構造や用いる訓練データの統計的性質に依存するが, α が十分大きな漸近領域での振る舞いはこれらの詳細には依らない, 「ユニバーサル」なものであることが多い. 従って, 学習曲線の漸近形の評価及び, それに基づく学習曲線の分類は学習理論において非常に重要な課題であると考えられており, 様々異なるアプローチから精力的な研究が現在まで進められてきている.

問 11：
この講義でみたように, $\Omega_P(\epsilon)$ が例題数の増加とともに単調に減少せず, 例題を 1 つ学習するごとに Ω が確率 λ で小さくならない (単調非増加) 場合を考えよう. このとき以下の問いに答えよ.

(1) $\Omega_{\mu+1}(\epsilon)$ と $\Omega_\mu(\epsilon)$ の間に成り立つ関係式を求め, $\Omega_P(\epsilon)$ を $\Omega_0(\epsilon)$ と ϵ, λ を用いて表せ.
(2) 講義ノートにならって, 解空間のエントロピー $S(\epsilon, \alpha)$ を計算し, ϵ を α の関数としてプロットせよ. また, α が十分大きな場合の学習曲線の漸近形を求めよ.