線型不良設定問題における早期停止付き最急降下法

石井雅治 *Masaharu ISHII*

Abstract

Consider relaxing an ill-posed system of linear equations to a minimization problem and finding its approximate minimum solution. It has been empirically known that using the steepest descent method with early stopping often gives a better solution than using the regularization method.

To study this empirical rule systematically, we have expanded the scope of analysis not only to the mathematical structure of the solution method but also to the observation situation, and introduced an "orthogonal subspaces with evaluation values" to formulate the goodness of the solution required by the observer. Here, the orthogonal subspaces with evaluation values are defined by subspaces with the observer's evaluation values which compose a direct sum decomposition of the observed space.

As this result, it was proved that it is necessary and sufficient to obtain a good solution satisfying the requirements of the observer for that the following conditions hold in the observation situation. This condition is that (i) a direct sum of some orthogonal subspaces with evaluation values is the same a certain subspace spanned by some left-singular vectors of the data generation process, (ii) the order of the evaluation values and the order of the singular values match, and (iii) the observed data mainly distribute in this subspace.

キーワード：□最急降下法　□早期停止　□正則化法
□特異値分解　□ロバスト性　□ニューラルネット
□神経回路網　□汎化能力　□信号復元

1 はじめに

次の線型連立方程式を考える．

$$y = Ax. \tag{1.1}$$

ここで，x, yはそれぞれm，n次元の実ベクトル，Aは$m \times n$の実行列であって，$rank\ A$が，m，nのいずれよりも小さい（縮退している）とする．信号復元などの分野で，ノイズを

含んだyを観測してxを推定しようとするとき，このような問題がしばしば現れる．しかしこの問題では，ノイズを無視したとしても，$rank\ A$が小さいので一般には解が存在しないし，仮に存在したとても一意に定まらない．このような問題は“不良設定問題”とよばれる．

そこで(1.1)の問題をやや緩和して，次の$E(x)$を最小化する問題を解くことを考える．

(1.2)　$$E(x) = ||y - Ax||^2.$$

ここで$||\bullet||$はベクトルの長さを表す．この最小化問題はxについて 2 次の問題なので，解が存在することは明らかであるが，この最小解は一意性を持たず，$m - rank\ A$次元の$\mathbb{R}^m$のアフィン部分空間を成す．

工学的問題では多くの場合，最終的に解を 1 つに絞り込む必要があるので，$E(x)$に“正則化項”とよばれる項を付加した次を最小化する問題を解く手法が広く用いられている．

(1.3)　$$E(x) + \varepsilon||x||^2$$

ここでεは十分小さい正の実数である．この問題では，通常の線型連立方程式の解法によって唯一つの解が得られる．

他方，$E(x)$の最小化問題の解法に，反復解法の 1 つである最急降下法が用いられる場合がある．渡辺は，“最急降下法を適用し，適当な回数で停止すると比較的良好な結果が得られるという経験則は･･･知られていたそうで･･･（(1.3)の）最小化よりもよい結果が得られることがある”（渡辺 2001, p. 41. 省略と括弧部は引用者），“直感的な説明は「A^TAの固有値の大きさと雑音レベルの大小」によって行われることが多いが，十分な説明は未だ確立されていない”（渡辺 2001, p. 41），と述べている．

最急降下法は，基本的に（人工）神経回路網等の非線型最小化の問題に適用される手法であり，この問題においても早期停止によって汎化性能の高い解が得られることがしばしば観察され議論されている．このため，深層神経回路網の教師有り学習において，最急降下法（正確にはその派生方法）と早期停止との組み合わせは，実用上不可欠な，極めてスタンダードな手法となっている．

線型の場合に，3 層神経回路網について，或る仮定の下でこの解法が確率的に高い汎化性能を持つことを証明した研究はあるが（福水 1998），この解法が(1.3)に比べよい解を与え得ることの十分な説明はないようである．

本研究は，$E(x)$の最小化問題に関し，早期停止付き最急降下法が“よい解”を導くという経験則を解明しようとするものである．まず，従来から知られている結果の整理として，Aの特異値分解を利用し，早期停止付き最急降下法は，特異値の大きさとAの誤差のレベルの大小に応じて，ロバストな解を与えることをみた．また，(1.3)の解法は，εが小さいとき不安定になる．或る程度大きいときも，誤差に埋もれた特異値に対応するyの射影成分が，他の成分と同程度の影響を解に及ぼし，早期停止付き最急降下法に比べロバ

スト性が低い場合があることをみた．

早期停止付き最急降下法が導く解のロバスト性は，解法の数学的構造によって保証される．しかしながら，このロバスト性は，通常$E(x)$を大きくするから，無条件で解の“良さ”を保証するものではない．更に，ロバスト性とはやや異なる汎化性能等の“良さ”が要求される場合もある．従って，解の良さは，単に解法の数学的構造に由来するものでない．この良さに先立ち，この良さを与えるような，解法には直接現れない何か潜在的な条件[1]に支えられているはずである．その解明は，理論的に興味深いだけでなく，技術上の大きな貢献となるだろう．しかし従来の研究では，断片的な言及を除き，この潜在的条件を体系的に明らかにしようとするものはなかった．

我々は，解析の対象を解法の数学的構造から，観測状況[2]にまで拡げることによって，この潜在的条件を体系的に解析した．ここで，観測状況とは，観測者が，誤差を含むAを通して，観測値yを観測し，これを生成する内部値xを推定するという状況を定式化したものである．すなわち，

a．観測者の要求，

b．（内部値から観測値を生成する）データ生成過程の特性A，

c．観測データの特性，

の組を“観測状況”とし，特に“評価値付き直交部分空間”を導入して観測者の要求する解の良さを定式化し，解析を行った．ここで，評価値付き直交部分空間とは，観測値の空間を直和分解する，観測者の評価値が付与された直交部分空間の組である．ただし，観測者の要求に関しては，観測者が，その存在そのものや，要求する評価値付き直交部分空間を明確に意識していないことも多いと考えられる．このことが，この条件を潜在化したままにさせてきた大きな要因であったのだろう．

以上によって我々は，観測者の要求を満たす良い解が得られることと，観測状況について次の条件が成立することとが必要十分であることを証明した．この条件とは，(i)評価値付き部分空間の一部の直和と，Aの左特異値ベクトルの一部が張る部分空間が一致し，(ii)評価値と特異値の順序が一致し，更に，(iii)観測データは主にこの部分空間に分布するというものである．

この条件は，解法の数学的構造の中ではなく，より広い，観測状況における整合性の中で始めて見出されることに注意しなければならない．

以下，第 2 章で，既に知られていると思われる，早期停止付き最急降下法や関連する解法の特性，これら相互の関係や誤差に対するロバスト性を整理する．第 3 章で，観測状況を定義し，評価値付き直交部分空間を導入して，この解法が良い解を導くための必要十分条件を定理として示す．第 4 章で，この定理の意義を明確化するため，この解法

---

[1] その要素の 1 つは，しばしば“バイアス”とよばれる．

[2] 観測状況以外の状況であっても，各要素を数学的に等価なものに置換可能であれば，以下の議論は成り立つ．必ずしも，観測系に限られたそれ固有の議論を行うものではない．

が，ロバストではあっても良い解を導かないような観測状況を示す．第 5 章で，結論をまとめ議論を行う．

2 従来の結果の整理

最急降下法と早期停止の関係について，従来から知られている結果がまとまって記述されたものがないようなので，この解法と関連する解法の特性，これらの解相互の関係や誤差に対するロバスト性を整理しておく．以下では，議論の単純化のため$m \leq n$とする．

最急降下法の基になる勾配系の微分方程式は次である．

$$\frac{dx}{dt} = -\frac{\partial E(x)}{\partial x} = -2A^T Ax + 2A^T y. \tag{2.1}$$

以下で，最急降下法の各ステップにおける値は，この微分方程式の解を十分よく近似しているものとする．

2.1. 各解法による解とその関係

Aの特異値分解を用いると，(2.1)の解，(1.2)の最小解，(1.3)の最小解等が明示されこれら相互の関係が分かる．特異値分解の 1 つを$\mathrm{A} = \mathrm{V}\Sigma U$とする．ここで，$V$は$n \times n$の実直交行列，$U$は$m \times m$の実直交行列，$\Sigma$は$n \times m$の次の実行列であって，対角成分のみが非負値$\sigma_i (i = 1, \ldots, m)$をとり，その他の成分は 0 であるものである．

$$\Sigma = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

各σ_iは“特異値”とよばれ，$\sigma_1 \geq \cdots \geq \sigma_m \geq 0$を満たす．$r = rank A$とおくと，$r < m$のとき，$\sigma_i > 0\ (1 \leq i \leq r)$であり他の$\sigma_i$は 0 である．以下では，対角成分以外 0 である$n \times m$行列を$diag_{n\times m}(\ )$と表記し，括弧内の列で対角成分を表す．例えば，$\Sigma = diag_{n\times m}(\sigma_1, \ldots, \sigma_m)$である．

ここで，$\tilde{x} = Ux$，$\tilde{y} = V^T y$とおくと，特異値分解の定義より(1.2)に関し次が成り立つ．

$$y - Ax = V(\tilde{y} - \Sigma\tilde{x}).$$

この関係を用いて，各種の解法による解とこれら相互の関係を示し，命題としてまとめる．

命題 2.1：次が成り立つ．

(A) $E(x) = ||\tilde{y} - \Sigma\tilde{x}||^2$であるから，この最小解$\tilde{x}_{\min}$は，$m - r$次元アフィン空間

$$\widetilde{X}_{\min} = \{(\tilde{y}_1\sigma_1^{-1}, \dots, \tilde{y}_r\sigma_r^{-1}, \tilde{x}_{r+1}, \dots, \tilde{x}_m)^T \mid \tilde{x}_i \in \mathbb{R}\ (i = r+1, \dots, m)\ \}$$

の任意の点$\tilde{x}_{\min}$. 最小値は$\tilde{y}_{r+1}^2 + \cdots + \tilde{y}_n^2$. よって(1.2)の最小解は$U^T\tilde{x}_{\min}$. 以下では, この解空間の点で, 解自身の大きさが最小であるものを$\tilde{x}^*_{\min}$ $(\tilde{x}^*_{\min\, i} = 0, (i = r+1, \dots, m)$で表す.

(B) (1.3)を最小化する問題は, 非退化な線型連立方程式

$$(diag_{n\times m}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0)) + \varepsilon 1_n)\ \tilde{x} - \Sigma^T\tilde{y} = 0$$

を解く問題に帰着し, この解$\tilde{x}_\varepsilon$は, 任意のε (> 0)に対し次で与えられる.

$$\tilde{x}_{\varepsilon i} = \frac{\sigma_i\tilde{y}_i}{\sigma_i^2 + \varepsilon} \quad (i = 1, \dots, m).$$

よって(1.3)の解は$U^T\tilde{x}_\varepsilon$であり, $\lim_{\varepsilon \to 0} \tilde{x}_\varepsilon = \tilde{x}^*_{\min}$が成り立つ.

(C) (2.1)は,

$$\frac{d\tilde{x}}{dt} = -2\left(diag_{n\times m}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0)\ \tilde{x} - \Sigma^T\tilde{y}\right)$$

と等価であるから, この解$\tilde{x}(t)$は次で与えられる.

$$\tilde{x}_i(t) = \begin{cases} \tilde{x}_i(0)e^{-2\sigma_i^2 t} + \tilde{y}_i\sigma_i^{-1}\left(1 - e^{-2\sigma_i^2 t}\right) & (i = 1, \dots, r). \\ \tilde{x}_i(0) \quad (i = r+1, \dots, m) \end{cases} \tag{2.2}$$

よって(2.1)の解は$x(t) = U^T\tilde{x}(t)$であり,

$$E(x(t)) = \sum_{i=1}^{r}(\sigma_i\tilde{x}_i(0) - \tilde{y}_i)^2 e^{-4\sigma_i^2 t} + \sum_{j=r+1}^{n}\tilde{y}_j^2$$

が成り立つ. $\tilde{x}(0) = 0$とおいた解を$\tilde{x}^*(t)$で表すと, $\lim_{t \to \infty} \tilde{x}^*(t) = \tilde{x}^*_{\min}$が成り立つ.

(D) 或る整数k $(1 \le k < r)$が存在して, $\sigma_k > \sigma_{k+1}$が成り立つ場合, $\tilde{x}_{\min}$の第$k+1$成分以降を0に置き換えた,

$$\tilde{x}^{(k)}_{\min} = \left(\tilde{y}_1\sigma_1^{-1}, \cdots, \tilde{y}_k\sigma_k^{-1}, 0, \cdots, 0\right)^T\ (\in \mathbb{R}^m)$$

について, $E(U^T\tilde{x}^{(k)}_{\min}) = \tilde{y}_{k+1}^2 + \cdots + \tilde{y}_n^2$である. よって, $|\tilde{y}_i\sigma_i^{-1}|$ $(i = k+1, \dots, r)$が十分小さいとき, $\tilde{x}^{(k)}_{\min} \approx \tilde{x}^*_{\min}$が成り立つ. また, $\sigma_k \gg \sigma_{k+1}$であるとき, $\tilde{x}^{(k)}_{\min} \approx \tilde{x}^*(t)$が成り立つような$t = T$ $(\ll \infty)$が存在する.

証明: (D)の最後の主張のみを証明する. (他はほぼ自明) εを, $\left(\sigma_k/\sigma_{k+1}\right)^2 = -log\,\varepsilon\,/\varepsilon$を満たすようにとる. $\sigma_k \gg \sigma_{k+1}$なので, この$\varepsilon$は十分小さい. このとき,

$$1 - e^{-2\sigma_{k+1}^2 t} = \varepsilon$$

を成り立たせる$t = T$は, $T \approx \varepsilon/(2\sigma_{k+1}^2)$であるから, $-2\sigma_k^2 T \approx \log\varepsilon$となり, 次が成り立つ.

$$1 - e^{-2\sigma_k^2 T} \approx 1 - \varepsilon.$$

また，$\tilde{x}_{\min}^{(k+1)} \big/ \tilde{x}_{\min}^{(k)} = 0$であるところ，

$$\frac{\tilde{x}_{k+1}^{*}}{\tilde{x}_{k}^{*}} = \frac{\sigma_k \tilde{y}_{k+1}}{\sigma_{k+1} \tilde{y}_k} \cdot \frac{\varepsilon}{1-\varepsilon} = \frac{\tilde{y}_{k+1}}{\tilde{y}_k} \cdot \sqrt{\frac{-\varepsilon \, log \, \varepsilon}{(1-\varepsilon)^2}}$$

が成り立つ．σ_k/σ_{k+1}を大きくしたとき，εは単調に 0 に近づくので，上の 3 つの式と(2.2)より，$\tilde{x}^*(T)$は$\tilde{x}_{\min}^{(k)}$に単調に近づく．■

注 1：(D)で $\tilde{x}_{\min}^{(k)}$を求めるには，Aの特異値分解を明示的に計算しなければならないので，計算量が多い．

注 2：(D)の最後の主張に関連するが，実用上は，σ_k/σ_{k+1}があまり大きな値でない場合であっても，$\tilde{x}(T)$は$\tilde{x}_{min}^{(k)}$の良い近似値を与える．例えば，適当なTに対し$1 - e^{-2\sigma_{k+1}^2 T} = 0.01$，$\sigma_k/\sigma_{k+1} = 22$のとき，$1 - e^{-2\sigma_k^2 T} \approx 0.99$となり，$\tilde{x}_{\min}^{(k)}$と$\tilde{x}(T)$の各成分の誤差は$\tilde{x}_{\min}$を基準にして，1 %以下である．また，また，$1 - e^{-2\sigma_{k+1}^2 T} = 0.05$，$\sigma_k/\sigma_{k+1} = 8$のとき，$1 - e^{-2\sigma_k^2 T} \approx 0.96$となり，誤差は 5 %以下である．

注 3：(D)の最後の主張より，早期停止付き最急降下法は，特異値分解による解$\tilde{x}_{\min}^{(k)}$を近似的に求めるアルゴリズムであるといえる．$\tilde{x}(T)$の計算には，Aの明示的な特異値分解を使わないので，$\tilde{x}_{\min}^{(k)}$に比べ極めて少ない計算量しか必要ない．

2.2. 各解法の解への誤差の影響

Aが誤差$\mu\delta A$ ($||\delta A|| = 1$)を持つ場合，よく知られているように，$A + \mu\delta A$の特異値は，重複がなければ，誤差の大きさ$|\mu|$の最大値μ_{max}が十分小さいとき，ほぼ

(2.3)　　$\sigma_i + \mu\sigma_i'$

で表される．ここでσ_i'はδAによって決まる実数値である．以下では誤差が(2.3)で表せると仮定して考察を進める．この仮定を明確化しておこう．

仮定2.1：Aの誤差$\mu\delta A$ ($||\delta A|| = 1$)に関し，(i) 誤差の大きさの最大値$\mu_{\max}$は(2.3)が近似的に成り立つ範囲の値であり，(ii) Aの特異値に重複はない，とする．

注：計算の簡単化のため，Aの特異値は，0を含め，誤差によってそれぞれ$\mu_{\max}$以下の微少に異なる特異値であったと前提する．

以下では仮定 2.1 をおく．このとき命題 2.1 より，(A)から(D)までそれぞれの解法の解への誤差の影響を明確化できる．これを命題としてまとめる．

命題2.2：或る 1 以上の整数kが存在して，添え字i ($1 \le i \le k$)について，任意の$\mu\delta A$に対し$|\mu\sigma_i'| \ll \sigma_i$であり，これ以外の添え字（存在すれば）について，この不等式が成り立たないとする．（このとき，$\sigma_k \gg \mu_{\max}$と$\mu_{\max} \approx \sigma_{K+1}$または$\mu_{\max} > \sigma_{K+1}$が成り立っていることに注意．）

このとき，命題 2.1 における(A)から(D)までの解と，誤差$\mu\delta A$の関係について，それぞれ次の(A')から(D')までが成り立つ．また，(C')または(D')が成り立てば，上の条件

を満たすkが存在する．

（A’）$k < m$の場合，$\tilde{x}_{\min j}$ $(k+1 \le j \le m)$の因数σ_j^{-1}が発散を含む影響を受け得るので，$\tilde{x}_{\min j}$は大きく変動する可能性を持つ．

（B’）$k < m$の場合，添え字j $(k+1 \le j \le m)$について$\varepsilon < |\sigma_j + \mu_{\max}|^2$が成り立つとき，$\tilde{x}_{\varepsilon j}$の因数$\sigma_j/(\sigma_j^2+\varepsilon)$が発散を含む影響を受け得るので，$\tilde{x}_{\varepsilon j}$は大きく変動する可能性を持つ．$\varepsilon \gg |\sigma_{k+1} + \mu_{\max}|^2$であるとき，$\tilde{x}_\varepsilon$はロバスト．

（C’）適当なT (>0)をとれば，$\tilde{x}(T)$はロバスト．（$\sigma_k \gg \sigma_{k+1}$と（D）の最後の主張より，）

（D’）$\tilde{x}_{\min}^{(k)}$はロバスト．（$\sigma_k \gg \sigma_{k+1}$より）

注１：kの存在についての前提条件の否定は，どの特異値も，適当な$\mu\delta A$をとると，$|\mu\sigma_i'| \ll \sigma_i$が成り立たない，という条件になる．これは，どの特異値も誤差と同程度のオーダーであるということを意味し，Aは実質的に無意味であり，当然，ロバストな解は存在しない．

注２：$k < r$の場合，$\tilde{y}_{k+1}^2 + \cdots + \tilde{y}_r^2$ が大きいと，$E(x)$の近似的最小解であって，上記の誤差に対しロバストであり同時に高精度であるものは存在しない．精度を高くするために$E(x)$を小さくしようとすると，$E(x) = ||\tilde{y} - \Sigma\tilde{x}||^2$より，$\tilde{y}_j$ $(k+1 \le j \le r)$成分を消去する必要があるところ，これには$\tilde{x}_j$の計算に発散を含む影響を受け得る因数σ_j^{-1}が必要であり，不安定性を生じるからである．

2.3. 解の良さに関する従来の“直感的な説明”の検討

A^TAの固有値をλ_iとすると，$A^TA = U^T\Sigma^T\Sigma U$より，$\sigma_i^2 = \lambda_i$である．よって，命題 2.2 の（C’）と注 2 から，固有値の（平方根）の大きさσ_iと雑音レベル$\mu\sigma_i'$の大小および$\tilde{y}_{k+1}^2 + \cdots + \tilde{y}_m^2$の大小によって，$U^T\tilde{x}^*(T)$は，良い解であり得る．すなわちこの従来の説明が，$A$の誤差に対するロバスト性に関しては妥当であることがわかる．

次に，このロバストな$\tilde{x}^*(T)$と$\tilde{x}_\varepsilon$をとり，両者のロバスト性を比較する．命題 2.1 の注 2 より，$\sigma_k/\sigma_{k+1} = 22$とした場合，

$$\tilde{x}_k^*(T) \approx 0.99\tilde{x}_{\min k},\ \tilde{x}_{k+1}^*(T) \approx 0.01\tilde{x}_{\min k+1}$$

であるところ，$\tilde{x}_{\varepsilon i} = (1 + (\varepsilon/\sigma_i^2))^{-1}\tilde{x}_{\min i}$であるから，$\varepsilon = \sigma_k\sigma_{k+1}$とおくと[3]，

$$\tilde{x}_{\varepsilon k} \approx 0.96\tilde{x}_{\min k},\quad \tilde{x}_{\varepsilon k+1} \approx 0.04\tilde{x}_{\min k+1}$$

を得る．$\tilde{x}_\varepsilon$に比べ$\tilde{x}^*(T)$では，誤差による変動の少ない$\tilde{x}_{\min k}$の係数が 1 に近く，変動が多い$\tilde{x}_{\min\ k+1}$の係数は 0 に近いので，よりロバスト性が高い．$\tilde{x}_\varepsilon$もより低いがロバスト性

3 このとき，$\sigma_k/\sigma_{k+1} \gg 1$であれば，$\sigma_k^2 \gg \varepsilon \gg \sigma_{k+1}^2$，$\tilde{x}_{\varepsilon k} \approx (1 - \sigma_{k+1}/\sigma_k)\tilde{x}_{\min k}$，$\tilde{x}_{\varepsilon\ k+1} \approx (\sigma_{k+1}/\sigma_k)\tilde{x}_{\min k+1}$,が成り立つ．ここで$\sigma_k^2 \gg \varepsilon$が必要なのは，これがないと，$E(U^T\tilde{x}_\varepsilon)$が大きくなって，$\tilde{x}_\varepsilon$が近似的最小値でなくなり得るからである．

を持っている．

しかし両者で，k成分と$k+1$成分との比を比較すると，大きな違いが現れる．

$$\frac{||\tilde{x}_{\varepsilon\ k+1}||}{||\tilde{x}_{\varepsilon\ k}||} \approx 1 \cdot \frac{||\tilde{y}_{k+1}||}{||\tilde{y}_k||}, \quad \frac{||\tilde{x}^*_{k+1}(T)||}{||\tilde{x}^*_k(T)||} \approx 0.22 \cdot \frac{||\tilde{y}_{k+1}||}{||\tilde{y}_k||}.$$

$\tilde{y}_{k+1}$は$\tilde{x}$の推定に意味を持ち得ない値であるが，$\tilde{x}_\varepsilon$はこの影響を排除せず，$\tilde{x}^*(T)$は低減するのである．同様な条件下でσ_k/σ_{k+1}を大きくしたとき，$\tilde{x}_\varepsilon$では$||\tilde{y}_{k+1}||/||\tilde{y}_k||$の係数は不変だが，$\tilde{x}^*(T)$では命題 2.1 の証明より 0 に近づく．

このように，早期停止付き最急降下法の方が正則化法よりも良い解を与える場合がある，という従来の感触は，ロバスト性に関しては妥当であるといえる．

3 早期停止付き最急降下法が良い解を導く条件

ここでは，観測者が，誤差を含むAを通して，yを観測し，これを生成する内部値xを推定するという典型的な状況の下に，(1.2)の最小化問題があるものとする．その上で，解析の対象を解法の数学的構造から，この観測の状況にまで拡げることによって，早期停止付き最急降下法が良い解を導く条件を体系的に明らかにする．すなわち，

a. 観測者の要求，

b.（内部値から観測値を生成する）データ生成過程の特性A，

c. 観測データの特性，

の組[4]を“観測状況”とし，特に“評価値付き直交部分空間”（直後に定義する）を導入して観測者の要求する解の良さを定式化し，解析を行う．

以下では，観測過程の誤差について仮定 2.1 をおく[5]．

a. 観測者の要求：観測者は，最急降下法の解$\tilde{x}^*(t)$（簡単化のためこの解に限る）に，通常，次の要求をする．

(i) 適当な$t = T$をとったとき，$E(x^*(T))$は小さい．

(ii) 上のTに対し，$x^*(T)$はAの誤差に対しロバストである．

加えて，観測者は，次のように観測値の空間を分解し，評価についての要求を行う場合がある．まず，観測値の空間を，1 次元の評価値付き直交部分空間の直和

$$D_1 \oplus \cdots \oplus D_n \ \ (D_i \cap D_j = \{0\},\ D_i \perp D_j\ (i \neq j))$$

---

[4] 定式の明確化のため，yの誤差は無視する．なお，命題 2.1 より，yの誤差は推定値xの分子にしか寄与しないので，通常はAの誤差に比べ影響は少ない．

[5] 誤差についてこの程度の前提が成り立たないとすると，データ生成過程を行列Aでモデル化することの意義があまりないことになる．従って，この前提をおくことは，自然であるとまではいえないにしても，解析を目的とした人為的なものではないとはいえるだろう．

に分解する．ここで，評価値付き直交部分空間とは，D_iに評価値$w(D_i)$[6] (≥ 0)を付随させたものであり，$w(D_i) \geq w(D_j)$ $(i < j)$とする．この上で，観測者は，次を要求する．
(iii) $P(D_i)$をD_iへの直交射影としたとき，

$$(3.1) \qquad ||P(D_i)(y - Ax^*(t))||^2$$

は，iが小さい（評価値が高い）ほど，tの増大によって速く収束するか最初から十分に小さく，しかもこの速さの大きさはiの増大によって速く減少する．

注１：上の (i)，(ii) が成り立つだけで，$x^*(T)$は (1.1) の近似解として，数学的には“良い”．(iii) まで成り立つと，観測データの内で高い評価の部分が$x(T)$によって生成できることになるから，実用的にも価値が高くなる．

注２：(iii) の，観測者の（時に潜在的な）要求の本体を定式化する条件こそが，従来の研究では十分明確にされてこなかったものである．

b. データ生成過程の特性：Aは，特異値分解は$V\Sigma U$を持ち，誤差$\mu\delta A$ $(||\delta A|| = 1)$を含む．

c. 観測データの特性：観測値yは，適当な線型部分空間Sをとれば，近似的にSの元とみなせる．すなわち，$y = y_S + y_\perp (y_S \in S,\ y_\perp \in S^\perp)$の分解について，$||y_S|| \gg ||y_\perp||$が成り立つ．

この 3 者が整合して良い解が得られるための条件を定理として示す．

定理3.1：観測者の要求 (i) と (ii) が成り立つことは，次の (I) と (II) が成り立つことと必要十分である．更に，観測者の要求 (i) から (iii) までが成り立つことは，次の (I) から (IV) までが成り立つことと必要十分である．
(I) 或る 1 以上の整数kが存在して，添え字i $(1 \leq i \leq k)$について，任意の$\mu\delta A$に対し$|\mu\sigma_i'| \ll \sigma_i$が成り立ち，これ以外の添え字（存在すれば）について，この不等式が成り立たない．
(II) $S \subset \mathrm{Im}\, A^{(k)}$．ここで$A^{(k)}$は，$k = m$の場合$A$であり，$k < m$の場合$A$の特異値$\sigma_{k+1}, \ldots, \sigma_m$を 0 とした次の行列である．

$$V\ diag_{n\times m}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)\ U.$$

(III) $v_i = (V_{1\,i}, \ldots, V_{n\,i})^T \in \mathbb{R}^n$とおいたとき，$D_i = \mathbb{R}v_i$ $(i = 1, \ldots, k)$．ただし，同じ値のσ_iが存在する場合は，その中でv_iの添え字を適当に入れ替えてこれが成り立てばよい．
(IV) σ_i^2は，iの増大によって（$i \leq k$の間）速く減少する．

証明：まず，(i), (ii) $\leftrightarrow$ (I), (II) を示す．命題 2.2 の (C') についての主張と注 2 の内容に注意すればいえる．

次に，(iii), (I) $\rightarrow$ (III), (IV) を示す．(I) より，収束の速い 1 次元部分空間の個数はkで

[6] 値そのものは不要で，部分空間を順序付けるため，その順序のみを使う．

あり，加えて(iii)における(3.1)の収束の速さの大きさの順序は，(2.2)よりσ_iの値の順序と一致しなければならない．従って，これらの順序に対応する部分空間について(III)が成り立つ．この一致と(iii)における(3.1)の収束の速さのi依存性から，(IV)が成り立つ．

最後に，(I), (II), (III), (IV)→(iii)を示す．(III), (IV)より，添え字i $(1 \le i \le k)$については，iが小さくなるほど(3.1)が速く収束しその速さの大きさも速く増大するので，(iii)が成り立つ．$k < n$である場合に，残りの添え字$i(k+1 \le i \le n)$についても，(iii)が成り立つことを示す．(III)より

$$D_1 \oplus \cdots \oplus D_k = \mathbb{R}v_1 \oplus \cdots \oplus \mathbb{R}v_k = \mathrm{Im}\, A^{(k)}$$

であるから，直交補空間をとれば次が成り立つ．

$$D_{k+1} \oplus \cdots \oplus D_n = \mathbb{R}v_{k+1} \oplus \cdots \oplus \mathbb{R}v_n = (Im\, A^{(k)})^{\perp}.$$

これと(I)と(2.2)より，添え字i $(k+1 \le i \le n)$について，(3.1)は速く収束しないか定数．他方，(II)より$S^{\perp} \supset (\mathrm{Im}\, A^{(k)})^{\perp}$であるから，この添え字について(3.1)は最初から小さい．すなわち，(iii)が成り立つ．■

注１：実用上，(I)から(IV)までが厳密に成立していなくても，或る程度成立していれば，早期停止付き最急降下法はそれなりに良い解を与える．

注２：(I)は解がロバスト性を持つための，データ生成過程の特性に関する条件であり，Aがモデルとして無意味でなければ成り立っている．詳しくは命題 2.2 の注 1 を参照．

4 早期停止付き最急降下法が良い解を導かない例

ここでは，観測状況における整合性がどのようなものであるかを明確化するため，早期停止付き最急降下法が，ロバストであるにも関わらず良い解を与えない観測状況を示す．特に，観測者の要求とデータ生成過程の特性あるいは観測者の要求と観測データの特性といった，２者だけの整合では良い解は得られず，３者が整合することが必要不可欠であることをみる．

まず，観測状況の中でデータ生成過程の特性を具体的に定義し，その特異値分解を与える．これを用いて，幾つかの観測状況について解の良さを議論する．

データ生成過程の特性：観測されたデータを（近似的に）生成する或る線型モデルについて，このパラメータを推定する最小化問題を与え，これに関する特異値分解を求める．

まず，Aを与える．或る直交多項式$P_i(X)$ $(i = 1, \ldots, m,\ P_i: \mathbb{R} \to \mathbb{R})$に対し，或る$X_j$ $(\in \mathbb{R},\ j = 1, \ldots, n,\ X_j < X_k\ (j<k))$ について，

$$p_i = (P_i(X_1), \ldots, P_i(X_n))^T \quad (\in \mathbb{R}^n,\ i = 1, \ldots, m)$$

とおいたとき，$p_1, \ldots, p_m$が次の直交性を満たしている．

$$(4.1)\qquad p_i^T p_j = \begin{cases} 0 & (i \neq j) \\ \alpha_i^2 & (i = j,\ \alpha_i \geq 0) \end{cases}.$$

また, 次が成り立っていて, α_iは, 0 でなければ添え字の増大によって急激に変化する.

$$(4.2)\qquad \alpha_1 \leq \cdots \leq \alpha_m.$$

このとき, モデル$x_1 P_1(X) + \cdots + x_m P_m(X)$によって観測データが生成される. 内部値は$x = (x_1, \ldots, x_m)$であり, 点$X_j$において観測された値を観測値$y$ $(\in \mathbb{R}^n)$のj成分とする. このとき, $A = (p_1, \ldots, p_m)$とおく.

次に, Aの特異値分解を与える. α_iが大きい順に$p_1, \ldots, p_m$の順序を置換したものを$\tilde{p}_1, \ldots, \tilde{p}_m$とすると, この置換は$m \times m$の直交行列で表され, これを$U$とおく. 同様に$\alpha_1, \ldots, \alpha_m$の順序を入れ替え$\tilde{\alpha}_1, \ldots, \tilde{\alpha}_m$とする. また, r $(3 \leq r < m)$は, $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_m$の内で 0 でないものの個数であり, $\sigma_i = \tilde{\alpha}_i$ $(i = 1, \ldots, r)$, $v_i = \tilde{\alpha}_i^{-1} \tilde{p}_i$ $(i = 1, \ldots, r)$ とおく. $v_{r+1}, \ldots, v_n$ $(\in \mathbb{R}^n)$を, $v_1, \ldots, v_n$が正規直交系を成すようにとる. 更に, $n \times m$行列Σを$diag_{n \times m}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$で, $n \times n$行列Vを$V_{ij} = v_{ji}$で定める.

このとき$A = V \Sigma U$が成り立っている. いま, Uは定義より, Vは(4.1)より直交行列であるから, $V \Sigma U$はAの特異値分解の 1 つである.

この特異値分解においては, (4.2)より, 特異値の順序とこれに対応するp_iの順序が逆であるから, Uは順序を逆にする置換である. すなわち, $P_1(X), \ldots, P_m(X)$に対応する特異値はそれぞれ$\sigma_m, \ldots, \sigma_1$である. また, 或る整数$k$ $(1 < k < r)$が存在して, $\sigma_{k+1}, \ldots, \sigma_r$は相対的に 0 に近く, 定理 3.1 の (I) が成り立っている.

観測状況 I :

a. 観測者の要求 : (i) 適当な$t = T$をとったとき, $E(x^*(T))$は十分小さく, (ii) $x^*(T)$はAの小さい誤差$\mu \delta A$に対しロバストである. (iii) $P_1(X), \ldots, P_m(X)$の順に観測値の重要な成分を表現している。そこで, 評価値付き直交部分空間を$D_i = \mathbb{R} p_i'$ $(i = 1, \ldots, n)$で定める. ここで, $p_i' = p_{m-r+i}$ $(i = 1, \ldots, r)$であり, $p_{r+1}', \ldots, p_n' (\in \mathbb{R}^n)$を, $p_1', \ldots, p_n'$が正規直交系を成すようにとる. また, $w(D_i) \geq w(D_j)$ $(i < j)$とする。次式は, iが小さいほど, tの増大によって速く収束するか最初から十分に小さく, しかもこの速さの大きさはiの増大によって速く減少する.

$$||P(D_i)(\ y - Ax^*(t)\ )||^2.$$

b. データ生成過程の特性 : 上記の通り.

c. 観測データの特性 : $y_1, \ldots, y_n$は区間 $[0, 1]$ 上の一様乱数として生成される.

このとき, a., b., c.の 3 者は整合しない, すなわち定理 3.1 の (II), (III), (IV) のいずれも成り立たない. このため, $x^*(T)$はロバストであるにも関わらず, $E(x^*(T))$は大きい上に, 観測者にとって重要度の高い$P_1(X), \ldots, P_{m-k}(X)$に対応した情報成分が欠落している.

観測状況 II : 観測状況 I において, **b.**の一部を次に置き換え **b'.**とする. このとき, **a.**と

b'.は整合し，定理 3.1 の(III)と(IV)が成り立つが，**b.**と **c.**が整合しないので定理 3.1 の(II)が成り立たない．このため$E(x^*(T))$は通常小さくならない．

・(4.2)　$\alpha_1 \geq \cdots \geq \alpha_m$，　$p_i' = p_i$ $(i = 1, \ldots, r)$.（このとき，σ_iの順序とこれに対応するp_iの順序が一致するので，Uは単位行列.）

観測状況III：観測状況 I において，**c.**を次の **c'.**に置き換える．このとき，**a.**と **c'.**は整合し，定理 3.1 の(II)が成り立つが，**a.**と **b.**が整合しないので定理 3.1 の(III)と(IV)が成り立たない．このため観測者にとって重要度の高い情報成分が欠落している．
c'. $||P(D_i)y||^2$は，iの増大によって急激に減少する．

観測状況IV：観測状況 I において，**b.**と **c.**をそれぞれ上記の **b'.**と **c'.**に置き換える．このとき，**a.**, **b'.**, **c'.**の3者が整合し，定理3.1の(II), (III), (IV)の全てが成り立つ．よって，$E(x^*(T))$は小さく，$x^*(T)$はロバストで，観測者にとって重要度の高い$P_1(X), \ldots, P_k(X)$に対応した情報成分が取り出されるので，$x^*(T)$は良い解である．

5 結論と議論

(1.1)の$y = Ax$においてAが縮退しているか縮退に近い状態にあるとき，xを近似的に計算するため，(1.2)の$E(x)$を最小化するxを求める問題を考える．本研究は，この最小化問題に関し，早期停止付き最急降下法が“よい解”を導くという経験則を解明しようとするものである．

5.1. 従来の結果の整理

最急速降下法と早期停止の関係について，従来から知られている結果をAの特異値分解を利用し整理し，この解法や正則化項付きの(1.3)の解法等の特性，これらの解相互の関係や誤差に対するロバスト性をみた．

この結果，早期停止付き最急降下法は，Aの誤差があまり大きくなければ，従来の説明のように，特異値の大きさとAの誤差のレベルの大小に応じて，ロバストな解を導くことを確認した（命題 2.2 の(C')）．ただしこのロバスト性は，通常$E(x)$を大きくするから，無条件では解の“良さ”を保証するわけではない．

他方，“重み減衰”あるいは“L^2正則化”ともよばれる(1.3)の解法を厳密に適用すると，Aの誤差に対しこの解は，正則化パラメータεが小さいとき不安定になるが，十分大きなεをとったときロバストになることをみた（命題 2.2 の(B')）．

この正則化法と早期停止付き最急降下法とのロバスト性を比較し，Aの誤差に対し後者の方がロバスト性が高いことを示した．また，前者では，誤差に埋もれた特異値に対応するyの射影成分が，他の成分と同程度の影響を解に及ぼすが，後者では，この影響を低減できる場合があることをみた．従来からいわれてきたように，早期停止付き最急降

下法の方が良い解を与える場合があるのである．

5.2. 正則化法との比較について従来の議論が有する問題

ディープネットを含む神経回路網の教師有り学習において，従来の議論では，極小解の近傍で(1.2)が近似として成立することを用い，小さい特異値がεに比べ十分小さい場合，上記の正則化法が早期停止付き最急降下法とほぼ同じ解を導くとした（例えば，Goodfellow, Bengio & Courville (2016), §7.8 を参照）．しかしこの議論には本質的な矛盾がある．目標となる極小解の座標値が特異値によらない定数であると暗黙の内に仮定している点である．この座標値は，実際には特異値の逆数に比例する（命題 2.1 の(A)を参照）．小さい特異値が存在すると，座標値に絶対値の大きな成分が表れ，この結果，両解法による近似解に大きな差異が生じてしまうのである（第 2.3 節を参照）．

一方，近年の神経回路網の研究では，A^TAに対応する極小解におけるヘシアンが，0 または 0 に近い固有値（この平方根が特異値に対応）を多く持つとする結果がある（Sagun, Bottou & LeCun (2016), Garipov et al. (2018)）．これが神経回路網の常態であっても，従来の議論とは異なるメカニズムによって，実用上は 2 つの解法でほぼ同じ結果が得られる．神経回路網において(1.3)に対応する解は，最急降下法かその派生解法によって近似的に計算されるので，(1.2)に対応する最急降下法との違いは，特異値のわずかな違いとみなすことができる（命題 2.1 の(B)と(C)を参照）．この結果，適当なステップ数で計算を停止すれば，極小解でヘシアンが退化していても，(1.2)の最急降下法とあまり変らない解が得られるのである．

5.3. 良い解が得られるための条件

これまで，早期停止付き最急降下法が，ロバストである以上に良い解を導く理由は十分解明されていなかった．その中で，我々は解析の対象を解法の数学的構造から，観測状況にまで拡げることによって，この解法が良い解を導く条件を体系的に解析した．ここで，観測状況とは，観測者が，誤差を含むAを通して，観測値yを観測し，これを生成する内部値xを推定するという状況を定式化したものである．すなわち，

a. 観測者の要求，
b.（内部値から観測値を生成する）データ生成過程の特性A，
c. 観測データの特性，

の組を“観測状況”とし，特に“評価値付き直交部分空間”を導入して観測者の要求する解の良さを定式化し，解析を行った．評価値付き直交部分空間とは，観測値の空間を直和分解する，観測者の評価値が付与された直交部分空間の組である．

この結果，①Aの誤差があまり大きくなく，②誤差より十分大きな特異値が存在するという前提の下で，観測者が要求するような“良さ”を持つ解が得られることと，観測状況について次の条件が成立することとが必要十分であることが明らかになった（定理 3.1）．この条件とは，(i)評価値付き部分空間の一部の直和と，Aの左特異値ベクトルの一

部が張る部分空間が一致し，(ii)評価値と特異値の順序が一致し，更に，(iii)観測データは主にこの部分空間に分布するというものである．

この条件は，解法の数学的構造の中ではなく，より広い，観測状況における整合性の中で始めて見出されることに注意しなければならない．

次に，この整合性がどのようなものであるかを明確化するため，早期停止付き最急降下法が，ロバストであるにも関わらず良い解を与えない観測状況を示した．特に，**a.**と**b.**あるいは**a.**と**c.**といった，２者だけの整合では良い解は得られず，3 者が整合することが必要不可欠であることをみた．

5.4. 残された課題と今後の展望

我々の結果の応用として，例えば，早期停止付き最急降下法が“よい解”を導くか否かの判定がある．まず，**a.**と**b.**がどのようなものであり整合しているかを明確化し，更に．**c.**がこれらと整合しているか否かを実験的または理論的に調べればよい．工学的には，**a.**と**c.**のみが整合している場合，データ生成過程を取り替えるという判断にも利用できる．信号復元の分野等への寄与が期待される．

また，非線型の最小化問題への拡張は，残された不可避の課題である．非線型問題では，Aは計算ステップ毎に変化するとみなせるから，計算中にこの整合性が維持されるような観測状況は有るのか，有る場合にそれがどのようなものなのかが基本的な問いとなる．この問いに関する研究は，非線型の最小化法における解の良さの解明を大きく進める可能性を持ったものであり，理論上も実用上も大きなしかも基礎としての意義を有するだろう．特に，本研究の枠組みは，観測値の空間が無限次元ヒルベルト空間であってもほぼ適用できるように構成されており，このような今後の研究に寄与することが期待できる．

謝辞

本研究の一部は，椙山女学園から令和 2 年度学園研究費助成金(B)を受けてなされたものである．助成いただいた椙山女学園に，感謝を申し上げる．

参考文献

福水健次（1998）．多層ニューラルネットワークのバッチ学習における過学習の存在，1998年情報論的学習理論ワークショップ．

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., & Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of DNNs. arXiv preprint arXiv: 1802.10026v4.

Goodfellow, I., Bengio, & Y., Courville, A. (2016). Deep Learning, The MIT Press. (岩澤有祐他訳 (2018), 深層学習, ドワンゴ)

Sagun, L., Bottou, L., & LeCun, Y. (2016). Eigenvalues of the hessian in deep learning:

singularity and beyond. arXiv preprint arXiv:1611.07476v2.
渡辺澄夫（2001）．第 2 章 学習と統計的推測，データ学習アルゴリズム，共立出版.

【著者略歴】

石井　雅治（いしい　まさはる）

1963年　福岡県生まれ

所　属・現　職　椙山女学園大学現代マネジメント学部現代マネジメント学科・教授

最終学歴・学位　名古屋大学大学院理学研究科博士課程後期課程物理学専攻修了・博士（理学）

所　属　学　会　日本応用数理学会，日本物理学会，情報処理学会，電子情報通信学会，日本神経回路学会

主　要　業　績　「ニューロ多様体の正則領域における勾配流の構造」，『日本応用数理学会論文誌』19(2)（2009），日本応用数理学会，pp. 143–157.
「Chebyshev 多項式を拡張した高次元可換多項式写像」，『日本応用数理学会論文誌』25(2)（2015），日本応用数理学会，pp. 59–90.
「人工知能とビッグデータ」，『社会とマネジメント』15（2018），椙山女学園大学現代マネジメント学部，pp. 1–12.