深層神経回路における勾配降下の複雑度展開と勾配の退化性

椙山女学園大学 現代マネジメント学部
石井雅治
Masaharu ISHII
School of Modern Management, Sugiyama Jogakuen University

1. はじめに

深層神経回路は，現在の様々の応用において従来手法より高い性能に到達している．この応用の多くでは教師あり学習が用いられ，学習は主に勾配降下と呼ばれる反復解法（とその派生手法）が利用される．この勾配降下では，次の奇妙とも思われる現象が観察されている．

I. 学習は, 目標関数の複雑度の低い成分から順に進む（例えば Zhi-Qin John Xu et.al 2019）.

II. “二重降下”1 : パラメータ次元が教師テータのサイズを大きく超える場合，反復回数を増やしていくと，最初は汎化誤差が減少するが或る回数から増大し再び減少する（Preetum Nakkiran et.al 2019a）.

III. “二重降下”2 : 同じ教師データに対し，パラメータ次元を増やしていくと，最初は汎化誤差が減少するが，パラメータ次元が教師テータのサイズに近づくにつれて増大し，その付近を超えると再び減少する（Preetum Nakkiran et.al 2019b）.

これらの現象は理論的に興味深いだけでなく，モデルの汎化能力を規定するので実用上も大変重要である．上記 I.の現象に関し，従来の理論的研究において，深層神経回路に関しそのバイアスの存在は或る程度示されているが（Nasim Rahaman et.al 2019），その力学は知られていない．上記 II.，III.の現象は特に奇妙であり，中でも III.は，統計理論の AIC や WAIC により, パラメータ次元が増大し続けると汎化誤差が増大すると考えられることに反する．現在のところ，実用的な条件において，二重降下を十分説明する理論はないようである（線型または 2 層神経回路に対し，n,m を無限にした極限においては，“漸近リスク解析”による説明がある（Ryumei Nakada1 and Masaaki Imaizumi 2021）.

我々は, モデル関数の勾配の空間を張る, 空間周波数で定義された“複雑度(関数)基底”と, 勾配降下の非線型特異値分解を導入し，大きな特異値と低い複雑度がその順に概対応することを証明した．この結果，特異値に関わる力学をモデルの複雑度に関わる力学とみることが可能になった．この枠組みを利用して深層神経回路の勾配降下においてI.の現象が生じることを証明できる．ここでは二重降下に関し，この枠組みから，パラメータ次元よりも，勾配が関数空中でどの程度退化しているかまたそれが反復回数の増大につれてどのように解消されるかがその力学にとって本質的であることを指摘する．

本稿の構成は次の通り．第 2 章で記号法を定め，第 3 章で上述の現象を解説する．第 4 章から第 7 章までで, 我々の枠組みを簡潔に紹介する．第 8 章でこの枠組みを用いて, I. - III.の現象に力学的説明を与える．特に二重降下については, 勾配の退化性についての考察を基に，証明としては十分でないものの，前章までの結果と整合的な仮説を立てる．なお本稿では，発表時のミスを修正し用語法を整理した．

2. 記号法

基本的な記号法を定める．定義域をトーラス T^M（$\cong[-\pi,\pi)^M$ の張り合わせ）とし，その上で目標関数 $g:T^M\to\mathbb{R}$，モデル関数 $f(w):T^M\to\mathbb{R}$（$w=(w_1,\ldots,w_n)\ (\in\mathbb{R}^n)$ はパラメータ)をとる．

$f(w), g$ は次を内積とする Hilbert 空間 L に属する．a,b が複素数値をとる場合も考える．

$$<a,b>=\int_{T^M} a(x)b(x)\rho(x)\,d^Mx \qquad \left(\int_{T^M}\rho(x)\,d^Mx=(2\pi)^M,\ \rho(x)\geq 0\right).$$

ここで $d^Mx=dx_1dx_2\cdots dx_M$，$\rho(x)$ は重み（δ 関数の場合も考える）である．$<\cdot,\cdot>$ に関する L の退化空間を次とする．

$$Kr(\rho)=\{a\in L\,|\,\forall b\in L, <a,b>=0\}$$

また，通常の内積を

$$<a,b>_*=\int_{T^M} a(x)b(x)\,d^Mx\ ,$$

とし，m 個のサンプリング点 $x_1,\ldots,x_m$（これと g を合わせ“教師データ”という）上での内積を次とする．

$$< a,b >_{S(m)} = \frac{(2\pi)^M}{m} \sum_{i=1}^{m} a(x_i)\, b(x_i)\,.$$

例えば，$\rho(x) = \rho_{S(m)}(x) = \dfrac{(2\pi)^M}{m} \sum_{i=1}^{m} \delta(x_i)$ とおいた場合，

$$< a,b >_{S(m)} = < a, b\rho_{S(m)} >_* = < a,b > \ .$$

$\hat{a}(k)\ (k \in \mathbb{Z}^M)$ で $a(x)\ (x \in \mathbb{R}^M)$ の Fourier 係数を表す．また Fourier 係数の内積と畳込みを次で表す．

$$\ll \hat{a}, \hat{b} \gg = \sum\nolimits_{k \in \mathbb{Z}^M} \hat{a}(k)\hat{b}^*(k), \qquad (\hat{a} * \hat{b})(k) = \sum\nolimits_{\ell \in \mathbb{Z}^M} \hat{a}(\ell - k)\hat{b}(k)\,.$$

残差を $\Delta(w;x) = f(w;x) - g(x)$ とし，誤差を次で定める．

$$E(w) = \frac{1}{2} < \Delta(w), \Delta(w) > \qquad \text{(汎化(/テスト)誤差)},$$

$$E_{S(m)}(w) = \frac{1}{2} < \Delta(w), \Delta(w) >_{S(m)} \quad \text{(訓練誤差)}.$$

モデル関数の勾配 $\partial f(w)/\partial w_i\ (i = 1,\ldots,n) \in L$ の 1 つの成分のみは定数関数とし，$f(w)$ の接空間[1] $T_w f$ と，内積の退化を考慮した実効的な接空間 $\tilde{T}_w f$ とを次で定める．ここで $P_{Kr(\rho)}$ は $Kr(\rho)$ への（通常の）垂直射影．

$$T_w f = \mathbb{R}\partial f/\partial w_1 + \cdots + \mathbb{R}\partial f/\partial w_n\ (\subset L), \qquad \tilde{T}_w f = (1 - P_{Kr(\rho)})T_w f\ (\subset L)\,.$$

目的． w を調節して g を $f(w)$ で近似する（回帰問題）．このため $E(w)$ を最小化したいが，m は有限なので，$E_{S(m)}(w)$ を小さくすることによってこれを近似的に実行する．

注 1： 実際には $E(w)$ は計算できないので（ρ が未知で計算量が大），訓練とは別のサンプルリング点を用いて $E_{S(m)}(w)$ との差を推定（“交差検証”）する．

注 2： 従来 $n \ll m$ が前提であったが，近年は，極端に大きい n の系も珍しくない[2].

注 3： $E_{S(m)}(w)$ の最小化（“学習”）は，通常は非線型問題になるので，下に定義する勾配降下またはその派生手法を用いて解く．

次の力学系の数値的反復解法により $w(t)$ を求め，$E_{S(m)}(w)$ の極小点を探査する方法を，**“勾配降下”**という．

---

[1] 微分作用素ではない．

[2] $n \gg m$ の超過少決定系が実際に利用されていて，2021 年では n が数千億オーダーの系もある．

$$\frac{dw}{dt} = -\frac{\partial E_{S(m)}(w)}{\partial w} = -<\Delta(w), \frac{\partial f(w)}{\partial w}>_{S(m)} \quad \left(\frac{\partial f(w)}{\partial w} = \left(\frac{\partial f(w)}{\partial w_1}, \cdots, \frac{\partial f(w)}{\partial w_n}\right)\right).$$

初期値を w_0（通常，適当な乱数で生成），数値的な収束先を w_* で表す．

3. 観察された学習特性

深層神経回路に関する勾配降下では，反復回数（epochs）の増大に伴って $E_{S(m)}(w)$ はほぼ単調に降下して（ノイズ有り）極小値 $E_{S(m)}(w_*)$ に収束する．また $E_{S(m)}(w_*)$ は，n の増大につれてほぼ単調に減少し，n が m 程度以上の場合ではほとんど 0 となる．

他方，$E(w)$ は，パラメータ次元 n の違いによって，多くの例で学習特性に違いが観察される．反復回数の増大に応じて次の特性を示す．

i) 小さい n の場合：$E(w)$ はほぼ単調に降下して一定値に漸近する（図 1.左で最も上のグラフ．グラフは Preetum Nakkiran et. al 2019a より）．

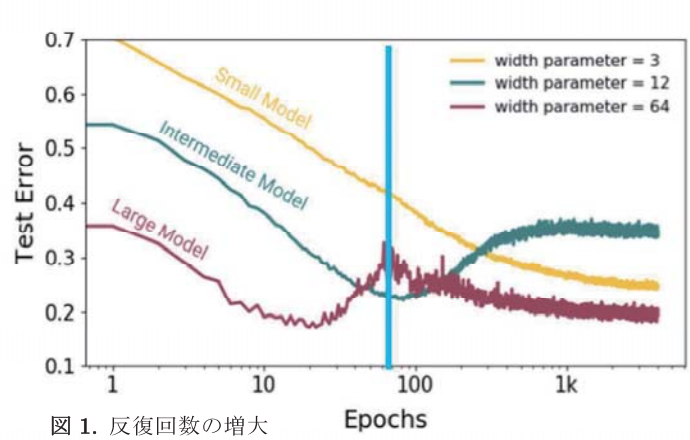

図 1. 反復回数の増大

ii) 中程度の n ($< m$) の場合：$E(w)$ は一旦降下した後に上昇する．この現象は“過学習”と呼ばれる．（図 1. 左で中のグラフ).

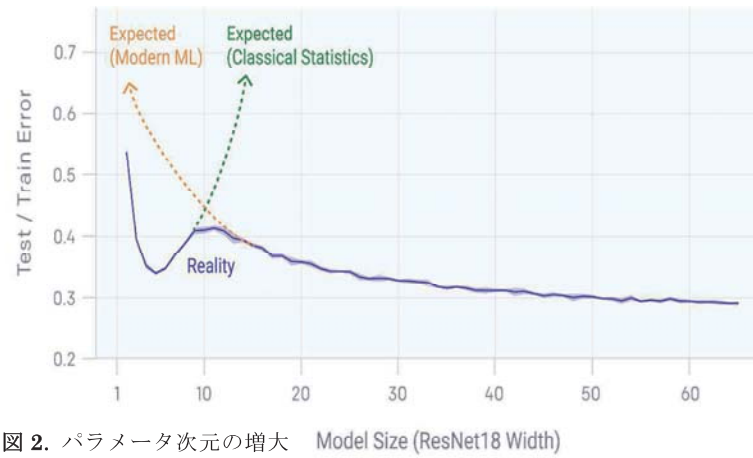

図 2. パラメータ次元の増大

iii) $n \gg m$ の場合：$E(w)$ は一旦降

下した後に上昇し再び降下する（図 1.左で最も下のグラフ）．この現象は（深層）“**二重降下**”と呼ばれる．

また，パラメータ次元の増大に応じて，$E(w_*)/E_{S(m)}(w_*)$ は，図 2.の左で最も下のグラフの特性を示す（グラフは Preetum Nakkiran et.al 2019b より）．特に，$n \gg m$ の場合，統計理論の AIC や WAIC からの類推で，過学習が発生しこの比が増大すると考えられる（図 2.右上がりの破線のグラフ）のだが，実際には降下する．この現象も二重降下と呼ばれる．

いずれの二重降下においても再降下が始まるのは，$n \approx m$ 程度のときで，この“≈”は 10 倍から数 10 倍程度とされる[3]．

4. 設計者の要求と内積の重み

従来の，統計理論に基づく枠組みの妥当性を検討し，暗黙の前提等の幾つかを明確化して要請として置き，後に利用する．

通常，重み ρ は確率密度関数として解釈され，x_i は分布 ρ によって独立に生成され，$E_{S(m)}(w) \to E(w)\ (m \to \infty)$（大数の法則）が成り立つものと仮定される．しかし実用上，この仮定が成り立っていないことも多く，どの程度成り立っているかを検証することも困難である．また，これに，設計者のデータへ対する暗黙の仮定，都合や要求が混合されて，結局，様々な出自の，異質な要素が ρ に皺寄せされているのである．

まず，次の画像識別の例で暗黙の前提をみてみよう．

Ex.1. 画像識別

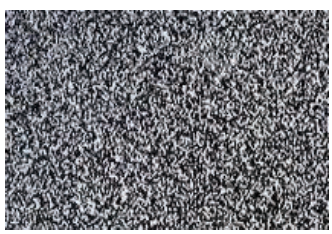

図 3. 画像の 99.99･･･%.

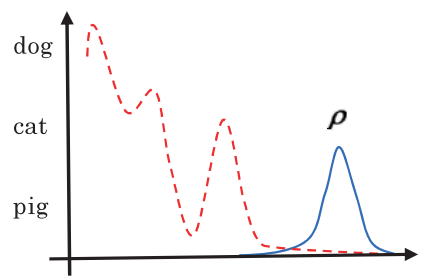

図 4. 破線がデータの分布，実線が ρ .

[3] 同程度とされる幅がかなり広い理由は，n が本来の規定要因でなく，$\dim(\tilde{T}_w f)$ が規定要因であることによる．このことについては，第 8 章で議論する．

理論上，画像の 99.99･･･%は図 3.のようなノイズとなるが，この上で ρ が大きな値をとっていては効率的な学習ができない．逆に，図 4.のように，対象となる画像が存在しない領域で ρ が大きな値をとっても意味がない．また，ρ がこの上で激しく振動することは（仮に実際のデータがそうであったとしても）技術的に望ましくない場合が多い．また，サンプリング点の集合は，画像全体を大まかには表して欲しい．

以上を次で表し，要請として置く．

i) $\rho(x)$ は，対象の範囲外では 0 に近く，対象の範囲上では速く変化しない．

ii) $\Delta\rho = \rho_{S(m)} - \rho$ とおいたとき，低周波域では $|\rho| \gg |\Delta\rho|$ が成り立つ．

上の $\rho_{S(m)}$ は周波数領域ではしばしば連続関数になり，簡単な例では $\widehat{\Delta\rho}$ を具体的に計算できる．この例を次に示す．

Ex.2. ρ と $\rho_{S(m)}$ を次として，その Fourier 係数を示す．

$$M = 1,$$

$$\rho(x) = \begin{cases} 1/2 & (|x| < 1) \\ 0 & (Otherwise) \end{cases},$$

$$\rho_{S(m)} = \frac{2\pi}{m} \sum_{i=1}^{m} \delta(x - 2(i-1)/(m-1) + 1).$$

$$\hat{\rho}(k) = \frac{1}{\sqrt{2\pi}} \frac{\sin k}{k}, \qquad \hat{\rho}_{S(m)}(k) = \frac{\sqrt{2\pi}}{m} \sin\frac{k\,m}{m-1} \Big/ \sin\frac{k}{m-1}.$$

ρ， $\rho_{S(m)}$ と $\widehat{\Delta\rho}$ を図示する．

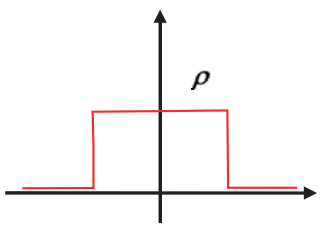

図 5. ρ.

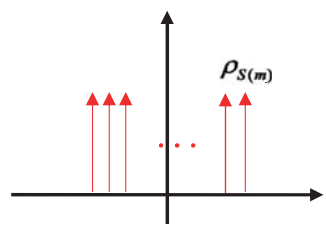

図 6. $\rho_{S(m)}$.

この例では，残差 Δ の高周波成分が十分速く減衰している場合，m の増大に伴って，$|\Delta\rho|$

がほぼ 0 になる低周波の範囲が広がれば，$E_{S(m)}(w) \to E(w)\ (m \to \infty)$ が成り立つ[4]ことがわかる．これを一般化して次の要請を置く．

iii) $m \to \infty$ のとき，$|\Delta\rho| \to 0$ となる低周波域が広がる．

iv) $g(x)$ の高周波成分は十分早く減衰する．

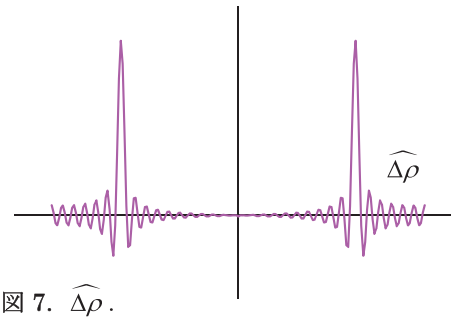

図 7. $\widehat{\Delta\rho}$.

5. 複雑度基底と複雑度展開

定義 1. $k=(k_1,\ldots,k_M)^T\ (\in \mathbb{Z}^M)$，$\varphi_k(x)=e^{-\sqrt{-1}\,k^T x}/\sqrt{2\pi}^M$．$\tilde{\varphi}_{\tilde{k}}\ (\tilde{k} \in \mathbb{Z}^M)$ を，

$$\tilde{\varphi}_{\tilde{k}} = \sum_{|k| \le |\tilde{k}|} \alpha_k \varphi_k \ (\tilde{k} \in \mathbb{Z}^M, \alpha_k \in \mathbb{C}, \alpha_k^* = \alpha_{-k})$$

を満たし，$<\varphi_k - \tilde{\varphi}_k, \varphi_k^* - \tilde{\varphi}_k^*>_*$ を最小化する $<\cdot,\cdot>$ の直交系で，0 または単位ベクトルとする．ここで，a^* は a の複素共役．さらに，$|k| = \sqrt{k_1^2 + \cdots + k_n^2}$ とおいて，双線型形式 $MF: L^2 \to \mathbb{C}$ を

$$MF(h,h') = \sum_{k \in \mathbb{Z}^M} |k| < h', \tilde{\varphi}_k^* >< h, \tilde{\varphi}_k^* >^*,$$

で定め，$h(x)\ (\in L)$ の“**複雑度**”を次で定義する．

$$cmp(h) = MF(h,h)/< h, h >.$$

注 1： $cmp(h)$ は，ほぼ h の平均周波数．

注 2： φ_k は一般に $<\cdot,\cdot>$ の正規直交系にならないので，これに近い直交系 $\tilde{\varphi}_k$ をとって，この類似を構成した．

仮定 2. 十分大きい $K\ (>0)$ が存在して，$k'\ (\in \mathbb{Z}^M)$ が $|k'| < K$ を満たせば，$\tilde{\varphi}_{k'}(x) \neq 0$ であり，$|\alpha_{k'}|$ が $|\alpha_k|\ (k \neq k')$ の和に比べ十分大きい．（k' がドミナントの周波数になっている）．

定義 3. $N = dim(\tilde{T}_w f)$ とし，$<\cdot,\cdot>$ についての $\tilde{T}_w f$ の正規直交基底 $\tilde{e}_i(x)\ (\in L, i = 1,\ldots,N)$ で次を満たすものをとる．

---

[4] サンプリング点を上手くとる必要がある．

$$MF(\tilde{e}_i, \tilde{e}_j) = \lambda_i \delta_{ij} \quad (i, j = 1, \ldots, N,\ 0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N).$$

$\partial f / \partial w_i$ は 1 つだけ定数で $MF(\cdot, \cdot)$ は半正定値２次形式なので，このような $\tilde{e}_i$ がとれる．$N < n$ である場合，$\tilde{T}_w f$ の $<\cdot, \cdot>$ についての直交補空間から互いに直交する $n-N$ 個の単位ベクトル $\tilde{e}_{N+1}, \cdots, \tilde{e}_n$ をとる．以下，$\tilde{e}_1, \cdots, \tilde{e}_n$ を（$\tilde{T}_w f$ の）**“複雑度(関数)基底”**という．

定義 4. $n \times n$ 実行列 A を $A_{ji}(w) = <\tilde{e}_j, \partial f(w) / \partial w_i>$，$\tilde{\mathbf{e}} = (\tilde{e}_1, \ldots, \tilde{e}_n)^T$ とおき，

$$\frac{\partial f(w)}{\partial w} = \tilde{\mathbf{e}}^T A$$

の右辺を（勾配の）“**複雑度展開**”といい，$<\tilde{e}_j, \partial f(w) / \partial w>$ を（複雑度 $cmp(\tilde{e}_j)$ の）“**複雑度成分**”という．

定義 5. $f(w; x)$，$f'(w'; x)$ に対し，それぞれのパラメータを $w = (w_1, \ldots, w_n)$，$w' = (w'_1, \ldots, w'_{n'})$ $(n < n')$ とする．適当な $w^*_{n+1}, \ldots, w^*_{n'}$ とれば次が成り立つとき，$f'(w'; x)$ は $f(w; x)$ を **“真に包含”**するという．

$$f(w; x) = f'(w_1, \ldots, w_n, w^*_{n+1}, \ldots, w^*_{n'}; x),$$
$$\tilde{T}_w f \subset \tilde{T}_{w'} f',\quad \dim(\tilde{T}_w f) < \dim(\tilde{T}_{w'} f').$$

定理 6. $f'(w'; x)$ は $f(w; x)$ を真に包含するとする．$N = \dim(\tilde{T}_w f)$，$N' = \dim(\tilde{T}_{w'} f'_w)$ とおき，それぞれの複雑度基底を $\tilde{e}_i$，$\tilde{e}'_j$ で表す．このとき，$cmp(\tilde{e}_2) \geq cmp(\tilde{e}_2{}')$ であり，等号成立は $\tilde{e}_2 = \tilde{e}'_2$ を満たす $\tilde{e}'_2$ がとれるとき．また，$cmp(\tilde{e}_N) \leq cmp(\tilde{e}'_{N'})$ であり，等号成立は $\tilde{e}_{N-1} = \tilde{e}'_{N'-1}$ を満たす $\tilde{e}'_{N'-1}$ がとれるとき．

注： $cmp(\tilde{e}_1) = cmp(\tilde{e}'_1) = 0$ なので，元のモデルを真に包含するモデルでは，通常，0 を除いた複雑度の下限は下がり，上限は上がる．

6. 勾配降下の特異値分解とその複雑度対応

A の特異値分解の 1 つを

$$A = U \Sigma V$$

とする．ここで，U, V は $n \times n$ 実直交[5]行列，$\Sigma = diag(\sigma_1, \ldots, \sigma_n)$，$\sigma_i$ は特異値と呼ばれる

[5] この“直交”は通常の意味．

実数で $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$．ただし，$N = \dim(\tilde{T}_w f) < n$ である場合，$\sigma_{N+1} = \cdots = \sigma_n = 0$．

定義 7. $\mathbf{e}^T = \tilde{\mathbf{e}}^T U$ とおくと，$e_1, \ldots, e_n$ $(e_i = (\mathbf{e})_i)$ も $<\cdot,\cdot>$ について $\tilde{T}_w f$ の正規直交基底になる．e_i を**“特異値（関数）基底”**という．

$\partial f(w)/\partial w = \mathbf{e}^T \Sigma V$ より，勾配降下は次のように特異値分解される[6]．ここで，$<\Delta, \mathbf{e}^T>_{S(m)}$ $= (<\Delta, e_1>_{S(m)}, \ldots, <\Delta, e_n>_{S(m)})$，$<\Delta, \mathbf{e}>_{S(m)} = (<\Delta, \mathbf{e}^T>_{S(m)})^T$ とした．

$$\frac{dw}{dt} = -V^T \Sigma <\Delta, \mathbf{e}>_{S(m)}.$$

よって残差について，次の特異値分解された微分方程式系が成り立つ．

$$\frac{d\Delta(w)}{dt} = -\mathbf{e}^T(w)\,\Sigma^2(w) <\Delta(w), \mathbf{e}(w)>_{S(m)} \quad \left(= -\sum_{i=1}^{n} e_i \sigma_i^2 <\Delta, e_i>_{S(m)}\right).$$

注 1： $\mathbf{e}, \Sigma$ は w に依存するので，系は通常は非線型である．ただし，これらが大きく変化する点は，$f(w)$ だけで決まり，目標関数や固定点[7]とは独立である．

注 2： NTK（神経接核）理論（Jaehoon Lee et.al 2019）では，$m \ll n$（超過少決定）のとき，適当な w_0 について次を証明している．

i) $\mathbf{e}(w), \Sigma(w)$ はほぼ定数（従ってほぼ線型系になる）．

ii) w_* は w_0 に極めて近い．

上の残差の方程式から次が主張できる．

命題 8. w の挙動の詳細に依らず，特異値の大きい順に $\Delta(w)$ の $e_i(w)$ 成分が，不可逆に減少していく．

命題 9. 極小点の近くでは，$\dim(\tilde{T}_w f) > m$ である場合，第 m 番までの特異値の大きい成分が関数近似に関与し，それ以後の小さい成分はほぼ関与しない．$\dim(\tilde{T}_w f) \leq m$ である場合，全ての成分が関数近似に関与する．

大変興味深いことに，深層神経回路の勾配降下では，適当な仮定の下で大きい特異値と

---

[6] 以下では，明記せず右辺にノイズによる小さい等方的な摂動が加わった系を取り扱う．よって w は安定性の小さい $E_{S(m)}(w)$ の極小点からは脱出し，安定性の高い極小点に収束する．

[7] 固定点は $<\Delta(w), \mathbf{e}(w)>_{S(m)} = 0$ の解であるから，サンプリング点にも依存する．

低い複雑度が概ねその順に対応し, 特異値は複雑度成分の大きさであることが証明される. よって, 上述の特異値で分解された系の挙動は, 複雑度で分解された挙動とみることが可能である.

仮定 10. 任意の i, j に対し, $\sum_{1\le k\le n} A_{ik}\, A_{jk} \approx 0$ が成り立つ. すなわち, $\partial f(w)/\partial w$ の異なる複雑度成分同士は, ほぼ相関を持たない.

定義 11. $f(w)$, $\partial f(w)/\partial w$ の高周波成分が十分速く減衰するとき, $f(w)$ を**"高周波減衰モデル"**という.

注： ReLU を活性化関数に持つ神経回路（深層なものを含む）は高周波減衰モデルであることが示されている（Nasim Rahaman et.al 2019).

定理 12. $f(w)$ が高周波減衰モデルであり, 仮定 10 が満たされているとき, 次が成り立つ. ここで $i \le \dim(\tilde{T}_w f)$ である.

i) 低い複雑度では, $\sigma_i \approx |< \tilde{e}_i, \partial f(w)/\partial w >|$, $e_i \approx \tilde{e}_i$ が成り立つ.

ii) 複雑度が高くなるにつれて, i)の対応は徐々に失われる.

以下では, w_0 は, モデル関数の複雑度が十分低くなるように選ぶとする. 以上の結果より, 勾配降下の基本的な挙動として次が主張できる.

定理 13. 勾配降下は, 概ね, 目標関数の低複雑度成分から高複雑度成分の順に学習する.

7. 臨界複雑度

第 4 章の要請 i) - iv)が成り立っていれば,

$$< \Delta, e_i > = < \Delta, e_i >_{S(m)} - \ll \hat{\Delta}, \hat{e}_i * \widehat{\Delta\rho} \gg$$

の関係より, 勾配降下における汎化誤差の挙動に関して次の命題が成り立つ.

定義 14. i を増やしていったとき, $|\hat{e}_i * \hat{\rho}| + \mu \approx |\hat{e}_i * \widehat{\Delta\rho}|$ となり始める $cmp(e_i)$ を**"臨界複雑度"**といい, C_C で表す. ここで $\mu(>0)$ は許容ノイズレベル.

注 1： Ex.2.では, $C_C \approx$ Nyquist 周波数.

注 2： この定義は暫定的に置いた大雑把な目安であって, より精密にする必要がある.

命題 15. 訓練誤差と汎化誤差の差は，主にC_C以上の複雑度の成分により生じ，勾配降下の反復がこの成分の残差$\Delta(w)$を減らしている期間に増大し，それ以外の期間ではほぼ一定である．また，C_C未満の複雑度では，$<\Delta, e_i>$と$<\Delta, e_i>_{S(m)}$の差は小さい．

8. 学習特性の力学

第 4 章から前章までの枠組みを使うことによって，これまで十分には理解されてこなかった深層神経回路の幾つかの特性を捉えることができる．ここでは，学習の順序や 2 種類の二重降下を含む以下の 8.1. - 8.3.の学習特性について，力学的説明を与える．なお，二重降下については，証明として十分ではないものの，前章までの結果と整合的な仮説を立てる．

この枠組みより，学習の挙動において系の挙動を基本的に規定するのは，パラメータ次元nよりも，実効的な接空間の次元$\dim(\tilde{T}_w f)$であり，臨界複雑度の上下にどのように複雑度が出現するかであることを指摘できる．例えば，二重降下における再降下の開始には，nとmの大小関係ではなく，$\dim(\tilde{T}_w f)$とmの大小関係が意味を持つと考えられる[8]．

$\tilde{T}_w f$に関しては，しばしば，w_0において大半の成分が退化しているが，反復回数の増大につれて退化が徐々に解消され，$\dim(\tilde{T}_w f)$が増大する傾向がみられる．ただし，w_*においても退化は著しい（Levent Sagun et.al 2016）．このような動的な退化解消の一般的理論としては，Masaharu Ishii & Y. hirata 2019 がある．動的な退化解消による複雑度の数と範囲の増大が，反復回数を増やした場合の二重降下の原因であると予想される．

以上の考慮に基づいて，それぞれの学習特性に説明を与える．

8.1. 目標関数の複雑度の低い成分から順に学習が進む特性.

この力学は，定理 12 の特異値 - 複雑度対応によって説明される（定理 13 を参照）.

8.2. 反復回数の増大で観測される学習特性（図 1.を参照）

i) 小さいnの場合：勾配にC_C以上の複雑度の成分が存在せず，ほぼ単調に汎化誤差が訓練誤差と共に減少する（定理 6 と命題 15 より）.

ii) 中程度の$n\,(<m)$の場合：nが或る程度大きいので，勾配にC_C以上の複雑度の成分が現れる（定理 6 より）．まず，初期の反復では低複雑度成分が学習され，汎化誤差は訓練

[8] まだ証明はできていない.

誤差と共に減少する（命題 8 と定理 12 より）．反復回数が或る一定値を超えた辺りから，C_C を超える高複雑度の成分が学習され，訓練誤差は減少するが，汎化誤差は増大する（命題 8，定理 12 と命題 15 より）．ただし，C_C を超える複雑度の成分が現れないときは，汎化誤差は増大しない（命題 9 と定理 12 より）．

次は仮説であり，今後の検証が必要である．

iii) $n \gg m$ の場合：前半では ii)の場合と同じ挙動を示す．更に反復回数を増やすと $\dim(\tilde{T}_w f)$ が増え，或る回数から $\dim(\tilde{T}_w f) > m$ となって，C_C より低い複雑度の成分が増える．この結果，主に関数近似を行う成分に C_C より低い複雑度のものが増えていき，汎化誤差が減少する（命題 9，定理 12 と命題 15 より．ただし，$\dim(\tilde{T}_w f)$ の増大と，C_C より低い複雑度の成分がどのように増えるかを示していないので，不十分）．

8.3. パラメータ次元の増大で観測される学習特性（図 2.を参照）

i) 最初の降下区間：n が小さいので，n の増大につれて $\dim(\tilde{T}_w f)$ が増え，勾配の持つ複雑度成分の数が増えるが，その最大の複雑度がまだ C_C より小さいので，訓練誤差と共に汎化誤差も減少する（定理 6 と命題 15 より）．

ii) 増大区間： i)の区間を超えて n が増大すると，C_C を越える複雑度の成分が増え出す（定理 6 より）．この区間では，概ね $\dim(\tilde{T}_w f) \leq m$ であるから，C_C より高い複雑度の成分も訓練誤差の低減に寄与し，その結果，汎化誤差が増大する（命題 15 より）．ただし，C_C を超える複雑度の成分が現れないときは，汎化誤差は増大しない（命題 9 と定理 12 より）．

次は仮説であり，今後の検証が必要である．

iii) 再降下区間：ii)の区間を超えて n が増大すると，C_C を越える複雑度の成分だけでなく，より低い複雑度の成分も増える（定理 6 より）．この区間では，$\dim(\tilde{T}_w f) > m$ となって，主に関数近似を行う成分に C_C より低い複雑度のものが増えていき，汎化誤差が減少する．（命題 9，定理 12 と命題 15 より．ただし，C_C より低い複雑度の成分がどのように増えるかを示していないので，不十分）．

なお，早期終了[9]を行った場合，n の増大につれて，C_C を越える複雑度の成分が現れても，これが汎化誤差を増やし始めると，学習は終了する．また，この成分は，C_C より低い複雑度の成分が十分あるときには，学習に寄与しない．従って，汎化誤差は（二重降下なしに）ほぼ単調に減少することがいえる（定理 6，命題 9，定理 12 と命題 15 より）．

謝辞

本研究の一部は，椙山女学園から令和 3 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深く感謝を申し上げる．

参照

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein and Jeffrey Pennington, 2019, “Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent”, arXiv:1902.06720v4.

Levent Sagun, L´eon Bottou and Yann LeCun, 2016, “Singularity of the hessian in deep learning”, arXiv:1611.07476v1.

Masaharu Ishii and Yoshihiro Hirata, 2019, “A proof of existence of attractors for gradient systems with singular manifolds, and its application to neural networks”, RIMS Research Conference 2019. (Oral presentation).

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio and Aaron Courville, 2019, “On the Spectral Bias of Neural Networks”, arXiv:1806.08734v3.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak and Ilya Sutskever, 2019a, “Deep Double Descent: Where Bigger Models and More Data Hurt”, arXiv:1912.02292v1.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak and Ilya Sutskever, 2019b, “Deep Double Descent”, https://openai.com/blog/deep-double-descent/.

[9] 汎化誤差が増大し始めたら反復を終了し，学習結果として，終了前に汎化誤差を最小にした w をとる処理．

Ryumei Nakada1 and Masaaki Imaizumi, 2021, “Asymptotic risk of overparameterized likelihood models: Double descent theory for deep neural networks”, arXiv:2103.00500v1.

Zhi-Qin John Xu, Yaoyu Zhanget, Tao Luo, Yanyang Xiao and Zheng Ma, 2019, “Training behavior of deep neural network in frequency domain”, arXiv:1807.01251v6.