修士論文

無限次元異方的平滑度を持つ関数に対する畳み込みネットワークによる関数近似・推定誤差の解析

48196208　奥本 翔

指導教員　鈴木 大慈 准教授

2021 年 1 月

東京大学大学院情報理工学系研究科数理情報学専攻

概要

近年，深層学習が様々なタスクにおいて高いパフォーマンスを示すことが実験で明らかになり，その理論的な性質解明にも注目が集まっている．特に，拡張畳み込みニューラルネットワークは，次元の大きな入力を伴う音声認識や画像処理，自然言語処理といったタスクで活躍している．現在，深層学習の関数推定能力を調べた研究は多くなされているが，入力データの次元 d がサンプルサイズ n に対して，$d \ll n$ なる状況を仮定しているものが多い．しかし，上述の拡張畳み込みニューラルネットワークが扱うような応用上よく用いられるデータでは，$d \gg n$，あるいは無限次元のデータが入力として与えられる．そこで，本研究では拡張畳み込みニューラルネットワークに対して無限次元のデータが入力された際の近似・推定誤差の解析を行う．この解析により，既存のヘルダー空間やベゾフ空間を用いた解析では，近似・推定誤差は次元の呪いの影響を受けるが，異方的平滑度を持つ関数族を考えることにより，近似・推定精度には次元が現れず関数の滑らかさにのみ依存することを示す．さらに，滑らかさにスパース性がある場合に拡張畳み込みが重要な役割を果たすことを示す．

目次

第 1 章

イントロダクション

近年，深層学習を用いたモデルが，画像認識，音声認識，自然言語処理等のタスクで高いパフォーマンスを発揮している．特に，畳み込みニューラルネットワーク (CNN) や，拡張畳み込みニューラルネットワークは，高次元データを伴うタスクで活躍している [19, 7, 16, 22]．しかし，その理論的な性質については明らかになっていない部分も多いため，関連する理論研究が注目を集めている．その中に，深層学習の関数近似・推定能力を解析する研究がある．例えば，適当な幅の 2 層全結合ニューラルネットワークモデルを用いれば，コンパクトサポートを持つ任意の連続関数を任意の精度で近似できることはよく知られている [1, 9]．また，ある関数クラスを近似するために必要な表現能力の解析もされている．例えば [21] においては，ヘルダー滑らかさ β を持つ関数を近似するニューラルネットワークが構成されている．また，同様の関数クラスについて，[10] では，滑らかさ β と次元 d により特徴づけられるレートを達成する推定誤差が導出されている．ほかにも，[18] では滑らかさ s を持つベゾフ空間について，同様に滑らかさと次元により特徴づけられる近似・推定レートが導出されている．これらの研究で導出されたレートでは，次元 d が n に対して非常に小さい ($d \ll n$) 場合を仮定して近似・推定レートが導出されているが，実際のタスクにおいては $d \gg n$ や $d = \infty$ であるケースが多々存在する．例えば，画像認識や自然言語処理等では，次元 d がサンプル数 n と比較して非常に大きいデータが与えられるケースが多い．また，関数データ解析や音声認識のタスクでは，入力される信号データが無限次元であるとみなすことができる．そこで，関数クラスや入力データにどのような条件があれば，次元の呪いを避けられるのかという観点での研究がなされている．例えば，[10] では，与えられたデータの次元が d であるが，それがより小さな次元 d' の多様体に埋め込まれる場合を仮定している．この場合，次元 d' に依存したレートが導出され，d のレートに対する影響を小さくできることが示されている．[11] では，与えられたデータの確率分布のサポートが，ミンコフスキー次元 d' の空間の部分集合である場合，d' と関数の滑らかさによってレートが決定されることが示されている．これら 2 つの研究では，データが実質的に低次元の情報量のみを有していると仮定した解析を行っているが，[17, 18] では，データを低次元に埋め込むことができなくても，混合平滑，あるいは異方平滑といった，軸ごとにスパースであったり，異なったスムーズネスを持つ場合には次元 d の影響が緩和されることが示されている．しかし，これらの研究においても，次元 d に依存する影

響を完全に解消したわけではなく，$d \gg n$ や $d = \infty$ である場合に，どのような条件を付ければ次元の呪いを回避できるのかということは解明されておらず，重要なトピックとなる．一方で，深層学習以外のモデルを用いた場合において，無限次元データが与えられた場合の近似・推定誤差に関する研究は既にいくつか存在している．例えば，[2] では，$[0,1]^\infty$ 上にサポートをもつ関数の，周波数成分の落ち方と近似精度との間の関係性について議論している．[23] では，無限次元異方ソボレフ空間上の信号に対する推定精度が，軸ごとの滑らかさの逆数和に依存することを示している．また，[13, 14] では，入力が関数空間や分布空間である場合の推定レートと推定手法を提案している．[6] では，関数が入力として与えられた場合の k-近傍法や Nadaraya-Watson 推定量の推定誤差解析がなされている．本研究では，無限次元のデータが与えられた状況を想定して，次の結果をす：

1. 無限次元列 $X = (x_1, \dots, x_i, \dots)$ が入力されたと仮定する．この時，ある特定の条件を満たす混合もしくは異方平滑な関数に対して，適当な全結合ニューラルネットワークが存在して，滑らかさとサンプル数のみに依存した近似・推定誤差を達成できる．
2. ある条件の下で，CNN は X 内のインデックス選択が可能であり，スムーズネスとサンプル数に依存した近似・推定誤差と達成できる．また，関数の滑らかさにスパース性があり，広い範囲のインデックスを選択することが必要ま場合でも，拡張畳み込み CNN を用いることで，それが達成可能になる．

これらの結果から，データの次元 d がデータ数 n に比べて非常に大きい場合 $(n \ll d)$ や，入力データが無限次元であるような場合でも，ある特定の滑らかさを持つ関数クラスを考えれば，その滑らかさのみに依存した推定誤差のオーダーが導出可能であることが示される．つまり，ある滑らかさの下では，次元の大きさは本質的ではなく，関数の滑らかさこそが本質的であるということができる．

第 2 章

既存研究と本研究の比較

本章では，既存研究の紹介と，本研究結果との比較をする．

2.1　関数の滑らかさと次元の呪い

関数の滑らかさ，入力次元と最適推定精度との間に深い関係性があることは既存研究で明らかにされている．一般的に滑らかな関数ほど推定しやすく，高次元入力を受け取る関数ほど推定しにくい．例えば，d 次元の入力を受け取る滑らかさ β のヘルダー空間は次のように定義される：

定義 1. ヘルダー空間 $d \in \mathbb{N},\ \beta \in \mathbb{R},\ \beta > 0$ を用いて，

$$\mathcal{C}^{\beta}(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} : \sum_{\alpha \in \mathbb{N}^d : |\alpha| < \beta} \|\partial^{\alpha} f\|_{\infty} + \sum_{\alpha \in \mathbb{N}^d : |\alpha| = \lceil \beta \rceil} \sup_{x \neq y} \frac{|\partial^{\alpha} f(x) - \partial^{\alpha} f(y)|}{|x - y|_{\infty}^{\beta - \lceil \beta \rceil}} < \infty \right\}.$$

ただし，$|a| = \sum_{i=1}^{d} a_i$，$\lceil \beta \rceil$ は β より小さい整数の中で最大のものとした．

ヘルダー空間の定義では，微分値の有界性やリプシッツ性により滑らかさを定義している．また，このほかに，主要な関数空間としてソボレフ空間，ベゾフ空間と呼ばれる空間では，それぞれ異なる滑らかさの定義を導入している．ソボレフ空間では，微分値の積分値，ベゾフ空間ではスプライン基底の係数の減衰レートにより，それぞれの滑らかさが定義される．既存研究では，関数空間の滑らかさと次元に依存した深層学習の学習レートがいくつか導出されている (表 2.1 参照)．ヘルダー空間やベゾフ空間に対するレートは，次元 d に強く依存することが分かる．これら関数空間の推定を考える場合，膨大な数のサンプル数 n が必要になることが示されている．こういった現象は，一般的に「次元の呪い」と呼ばれる．

Hölder space ($\mathcal{C}^{\beta}$) [8]	Besov space (B^{s}_{pq}) [17]
$n^{-\frac{2\beta}{2\beta+d}}$	$n^{-\frac{2s}{2s+d}}$

表 2.1. 滑らかさと次元に依存したレート

2.2　多様体仮説と異方，混合滑らかさ

前節では，関数の滑らかさと次元がレートに与える影響について解説した．前節の結果が示唆する通り，一般的に，次元の大きなデータから学習することは非常に難しい．しかし，現実の問題では，高次元データの学習がうまくいく例が多々存在する．そのため，どのような仮定の下で次元の呪いが回避できるか調べることが理論的に重要となる．その仮定の一つに，「多様体仮説」が存在する [5]．高次元データがその情報を失うことなく低次元に埋め込むことが可能であるとき，多様体仮説が成り立つと言う．多様体仮説の下で導出された学習レートを 2 つ紹介する．

Minkowski Dimension d' $(\mathcal{C}^\beta)$ [11]	Lower manifold d' [10]
$n^{-\frac{2\beta}{2\beta+d'}}$	$n^{-\frac{2\beta}{2\beta+d'}}$

表 2.2. 多様体仮説下での推定誤差

[11] の設定では，入力されるデータの次元が d が大きい場合でも，ミンコフスキー次元と呼ばれる量 d' が d より小さな値であれば，レートの次元 d への依存性がなくなり，代わりに d' に依存することが示されている．また，[10] では，データが d' 次元の多様体に埋め込み可能である場合に [11] と同様のことが示されている．これらの結果より，多様体仮説の下では，高次元データが与えられた場合でも，より小さい次元のみに依存した学習が可能であることが分かる．一方，多様体仮説が成立しない場合でも，次元の呪いを回避できるケースが存在することも示されている．[17, 18] では，混合・異方ベゾフ空間に対する深層学習の推定誤差の解析を行っている．その中で，関数空間に混合，異方滑らかさがある場合には，その滑らかさのみに依存した学習レートが導出され，次元の呪いが回避できることが示されている．

mixed Besov space (MB^s_{pq}) [17]	anisotropic Besov space (B^α_{pq}) [18]
$n^{-\frac{2s}{2s+1}}$	$n^{-\frac{2\tilde{\alpha}}{2\tilde{\alpha}+1}}$

表 2.3. 異方・混合滑らかさの下での推定誤差

2.3　本研究結果との比較

前節で紹介した既存研究では，$d \ll n$ の仮定の下で推定誤差の上界が導出されている．しかし，応用上の問題においては，音声データや分布回帰問題のように関数データを入力とする場合 $(d = \infty)$ や，画像データや自然言語データのように超高次元 $(d \gg n)$ であるケースが多々見られる．そこで，次元 $d \ll n$ の仮定を外した場合に，どのような関数クラスであれば次元の呪いを回避できるのか調べることが重要となる．本研究では，関数空間の周波数成分の減衰に着目し滑らかさを定義することにより，次元 d が無限次元の場合であっても，適切な滑らかさを持つ関数クラスに対しては，その滑らかさのみに依存した推定誤差が達成可能であること

を示す．本研究では，多様体仮説を仮定していないため，元の次元が超高次元であっても適当な関数の滑らかさの下では，速い学習レートが達成可能であることが示されている．

混合滑らかさ $(d \ll n)$	著者	異方滑らかさ $(d \ll n)$	著者
$n^{-\frac{2s}{2s+1}}$	Suzuki (2019) [17]	$n^{-\frac{2\frac{1}{\alpha}}{2\frac{1}{\alpha}+1}}$	Suzuki and Nitanda (2019) [18]
混合滑らかさ $(d = \infty)$	著者	異方滑らかさ $(d = \infty)$	著者
$n^{-\frac{2(a_1-v)}{2(a_1-v)+1}}$	本研究	$n^{-\frac{2(\frac{1}{\alpha}-v)}{2(\frac{1}{\alpha}-v)+1}}$	本研究

表 2.4. 本研究との比較

第 3 章

問題設定

これ以降，$\mathbb{R}_{>0} := \{s \in \mathbb{R} : s > 0\}$，$\mathbb{R}^\infty := \{(s_1, \dots s_i, \dots) : s_i \in \mathbb{R}\}$ とする．また，$s \in \mathbb{R}^\infty$ に対して，$\mathrm{supp}(s) = \{i \in \mathbb{N} : s_i \neq 0\}$，$\mathbb{N}_0^\infty := \{l \in (\mathbb{N} \cup \{0\})^\infty : \mathrm{supp}(l) < \infty\}$ と定義する．また，$\mathbb{Z}_0^\infty$，$\mathbb{R}_0^\infty$，$\mathbb{R}_{>0}^\infty$ についても同様に定義する．さらに，$s \in \mathbb{R}_0^\infty$ に対して，$2^s := 2^{\sum_{i=1}^\infty s_i}$ とする．また，$L \in \mathbb{N}$ に対して，$[L] = \{i : i = 1, \dots, L\}$ とし，$a \in \mathbb{R}$ に対して，$\lceil a \rceil$ は，a より小さい整数の中で最大のものとする．

3.1　無限次元入力と回帰問題

本節では，これ以降の議論で用いられる関数ドメインの定義と，回帰問題に関する解説を与える．$[0,1]$ 上の一様確率測度を λ とし，λ^∞ を λ の無限直積測度とする．ここで，P_X を $[0,1]^\infty$ 上の，λ^∞ について絶対連続な確率測度で，$\frac{\mathrm{d}P_X}{\mathrm{d}\lambda} < \infty$ を満たすとする．次に，回帰問題について説明する．X を分布 P_X に従う確率変数とする．確率変数 Y について，ある関数 $f : \mathbb{R}^\infty \to \mathbb{R}$ を用いて，

$$Y = f(X) + \xi \ \left(\xi \sim \mathrm{N}(0, \sigma^2),\ \sigma > 0\right)$$

が成り立つと仮定し，(X, Y) の従う分布を P とする．得られた，n 個の観測値 $\{(X_i, Y_i)\}_{i=1}^n$ から関数 f を推定する問題を回帰問題と呼ぶ．また，$\mathrm{E}_P[(f(X) - Y)^2]$ を汎化誤差，$\frac{1}{n}\sum_{i=1}^n (f(X_i) - Y_i)^2$ を経験誤差と呼ぶ．あるモデルの集合 $\mathcal{F}$ (線形回帰モデル，CNN etc.) の中で，経験誤差を最小にするようなモデルを選択する学習方法を経験誤差最小化法 (Empirical Risk Minimization, ERM) と呼び，$\hat{f}$ で表す．これ以降，ERM によって推定されたモデルの収束オーダーについて議論する．

3.2　無限次元空間上の関数クラスと滑らかさ

本節では，本研究で議論の対象となる関数クラスについて説明する．$l \in \mathbb{Z}_0^\infty$ を用いて $x_i \in \mathbb{R}$ に対して，

$$\psi_{l_i}(x_i) = \begin{cases} \sqrt{2}\cos(2\pi|l_i|x_i) & (l_i < 0) \\ \sqrt{2}\sin(2\pi|l_i|x_i) & (l_i > 0) \\ \phi_0(x_i) = 1 & (l_i = 0) \end{cases}$$

として，$\psi_l(X) := \prod_{i=1}^\infty \psi_{l_i}(x_i)$ と定義する．また，$f, g : \mathbb{R}^\infty \to \mathbb{R}$ を用いて，

$$\langle f, g \rangle := \int_{[0,1]^\infty} fg \mathrm{d}\lambda^\infty$$

とする定義する．ここで，

$$L^2([0,1]^\infty) := \left\{ f : \mathbb{R}^\infty \to \mathbb{R} : \int_{[0,1]^\infty} f^2 \mathrm{d}\lambda^\infty < \infty \right\}$$

とすれば，$f \in L^2([0,1]^\infty)$ は $l \in \mathbb{Z}_0^\infty$ を用いて，$f(X) = \sum_{l \in \mathbb{Z}_0^\infty} \langle f, \psi_l \rangle \psi_l(X)$ で展開できる．l は各成分の周波数を表している．$s \in \mathbb{N}_0^\infty$ を用いて $\delta_s(f, X) = \sum_{\lceil 2^{s_i - 1} \rceil \leq |l_i| < 2^{s_i}} \langle f, \psi_l \rangle \psi_l(X)$ として，次の関数空間を定義する：

定義 2 (滑らかさ γ を持つ関数空間)**.** $\gamma(s) : \mathbb{N}_0^\infty \to \mathbb{R}_{>0}$ は，s の各要素 s_i について狭義単調増加であると仮定する．この時，$p \geq 1,\ \theta \geq 1$ を用いて，

$$\mathcal{F}_{p,\theta}^\gamma([0,1]^\infty) := \left\{ f = \sum_{l \in \mathbb{N}_0^\infty} \langle f, \phi_l \rangle \phi_l : \left(\sum_{s \in \mathbb{N}_0^\infty} 2^{\theta\gamma(s)} \|\delta_s(f)\|_p^\theta \right)^{1/\theta} < \infty \right\}$$

と定義し，$\|f\|_{\mathcal{F}_{p,\theta}^\gamma} := \left(\sum_{s \in \mathbb{N}_0^\infty} 2^{\theta\gamma(s)} \|\delta_s(f)\|_p^\theta \right)^{1/\theta}$ とする．この関数空間を，γ-平滑周波空間と呼ぶ．ただし，$\|f\|_p := \left(\int_{[0,1]^\infty} f^p \mathrm{d}\lambda^\infty \right)$ とした．また，これ以降，$U(F_{p,\theta}^\gamma) := \left\{ f \in \mathcal{F}_{p,\theta}^\gamma : \|f\|_{\mathcal{F}_{p,\theta}^\gamma} \leq 1 \right\}$ で定義し，$\mathcal{F}_{p,\theta}^\gamma$ と略記した場合には，$\mathcal{F}_{p,\theta}^\gamma([0,1]^\infty)$ を表すものとする．

$\delta_s(f)$ は，特定の領域に入る周波数成分を表している．この定義では，周波数成分の減衰度合いが滑らかさ $\gamma(s)$ により決まっている．本定義と類似する滑らかさの定義としては，ウェーブレット基底を用いたベゾフ空間の定義等に見ることができる [4]．また，有限次元の場合について，類似の定義による関数クラスの研究がいくつかなされている [20]．ここで，これ以降重要になる 2 つの滑らかさを導入する:

定義 3 (混合滑らかさと異方滑らかさ)**.** ある単調増加数列 $a = (a_1, \dots a_i, \dots) \in \mathbb{R}_{>0}^\infty$ が与えられたとき，

$$\gamma(s) = \langle a, s \rangle$$

の場合を混合周波空間，

$$\gamma(s) = \max_i \{a_i s_i\}_{i=1}^{\infty}$$

の場合を異方周波空間と呼ぶ．ただし，$\langle a, s\rangle := \sum_{i=1}^{\infty} a_i s_i$ である．

この時，$a = (a_1, \ldots, a_i, \ldots)$ のそれぞれの成分 a_i は，データ X 中の変数 x_i 方向の滑らかさを表している．例えば，a が単調増加数列であれば，インデックスの大きな成分ほど滑らかさが大きいことを意味している．入力空間が有限次元の場合においては，これらの滑らかさをもつベゾフ関数に対する，ニューラルネットワークを用いた推定誤差が議論されている [18, 17].

3.3　既存の関数空間との関係

ここでは，有限次元混合ベゾフ空間・無限次元異方ソボレフ空間と，$\mathcal{F}_{p,\theta}^{\gamma}$ の関係について議論する．関数の平滑係数は次のように定義される．

定義 4 (平滑係数)**.** 関数 $f : [0,1]^d \to \mathbb{R}$ の r 階平滑係数は，

$$\Delta_h^r(f)(x) := \begin{cases} \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} f(x + jh) & (x \in [0,1]^d,\ x + jh \in [0,1]^d) \\ 0 & (\text{otherwise}) \end{cases}$$

を用いて，

$$w_{r,p}(f, t) = \sup_{\|h\|_2 \le t} \|\Delta_h^r(f)\|_p$$

で定義される．

平滑係数は，関数の滑らかさを評価する指標として，様々な関数近似理論の研究において用いられている．これを用いてベゾフ空間の定義を与える．

定義 5 (ベゾフ空間)**.** $0 < p, q \le \infty$，$s > 0$, $r = \lceil s \rceil + 1$ を用いて，ベゾフノルムを，

$$|f|_{B_{p,q}^s} := \begin{cases} \left(\int_{[0,1]^d} (t^{-s} w_{r,p}(f,t))^q \frac{dt}{t} \right)^{1/q} & (q < \infty) \\ \sup_t t^{-s} w_{r,p}(f,t) & (q = \infty) \end{cases}$$

とする．これを用いて，ベゾフ空間 $B_{p,q}^s$ は，

$$B_{p,q}^s := \left\{ f : [0,1]^d \to \mathbb{R} : \|f\|_p + |f|_{B_{p,q}^s} < \infty \right\}$$

で定義される．

ベゾフ空間に属する関数の近似に関する研究は，近似理論の分野において頻繁になされている．また，ベゾフ空間に属する関数を対象とした回帰問題の統計的解析は，[17] においてなされている．次に，混合・異方ベゾフ空間について説明する．混合平滑係数は次のように定義される．

定義 6 (混合平滑係数)**.** $e \subset \{1,\dots,d\}$，$h \in \mathbb{R}^d_{>0}, r \in \mathbb{N}^d$ を用いて，r 階混合平滑係数は，$x \in [0,1]^d$ に対して，

$$\Delta_{h_i}^{r_i,i}(f)(x) = \Delta_{h_i}^{r_i}\left(f(x_1,\dots,\cdot,\dots,x_d)(x_i)\right),\ \Delta_h^{r,e}(f) := \left(\prod_{i\in e}\Delta_{h_i}^{r_i,i}\right)(f)$$

として，$t>0$ に対して，

$$w_{r,p}^e(f,t) := \sup_{|h_i|\le t_i, i\in e} \|\Delta_h^{r,e}(f)\|_p$$

で定義される．

また，$s \in \mathbb{R}^d_{>0}$，$r_i = [a_i]+1$ を用いて，

$$|f|_{MB^{a,e}_{p,q}} := \begin{cases} \left\{\int_{x\in[0,1]^d}\left[\left(\prod_{i\in e} t_i^{-r_i}\right) w_{r,p}^e(f,t)\right]^q \frac{dt}{\prod_{i\in e} t_i}\right\}^{1/q} & (1\le q<\infty) \\ \sup_{t\in[0,1]^d}\left(\prod_{i\in e} t_i^{-r_i}\right) w_{r,p}^e(f,t) & (q=\infty) \end{cases}$$

と定義する．これを用いて，f の混合ノルムは，

$$|f|_{MB^s_{p,q}} := \|f\|_p + \sum_{e\subset\{1,\dots,d\}} |f|_{MB^{a,e}_{p,q}}$$

で定義される．また，混合ベゾフ空間は，

$$MB^a_{p,q} = \left\{f : [0,1]^d \to \mathbb{R} : |f|_{MB^a_{p,q}} < \infty\right\}$$

で定義される．また，有限次元の混合ベゾフノルムに関して，$s \in \mathbb{N}_0^\infty$ を用いて，

$$\delta_s(f) = \sum_{s\in\mathbb{N}_0^\infty : \lceil 2^{s_i-1}\rceil \le \sum_{i=1}^d a_i s_i < 2^{s_i}} \langle f, \phi_l\rangle \phi_l$$

とすれば，

$$|f|_{MB^a_{p,q}} \sim \left(\sum_{s\in\mathbb{N}_0^d} (2^{\langle a,s\rangle}\|\delta_s(f)\|_p)^\theta\right)^{1/\theta}$$

が成り立つことが知られている [3]．したがって，混合周波空間は，有限次元混合ベゾフ空間の無限次元への拡張とみなすことができる．また，無限次元の異方ソボレフ空間は，単調増加な正の無限列 $a \in \mathbb{R}^\infty$ を用いて，

$$\mathcal{S}_{a,\infty} := \left\{f \in L^2([0,1]^\infty) : \sum_{i=1}^\infty \left\|\frac{\partial^{a_i} f}{\partial x_i^{a_i}}\right\|_2^2 < \infty\right\}$$

で定義される．また，[23] によれば，$f \in \mathcal{S}_{a,\infty}$ と，

$$\sum_{i=1}^\infty c_l^2 \langle f,\phi_l\rangle^2 < \infty$$

が成立することが同値である．ただし，任意の $l \in \mathbb{N}_0^\infty$ について，

$$c_l^2 = \sum_{i=1}^{\infty} (2\pi l_i)^{2a_i}$$

と定義した．ここで，

$$\|\delta_s\|^2 = \sum_{l \in \mathbb{N}_0^\infty : 2^{s_i - 1} \leq |l_i| < 2^{s_i}} \langle f, \phi_l \rangle^2$$

であり，

$$\begin{aligned} &\sum_{l \in \mathbb{N}_0^\infty : 2^{s_i - 1} \leq |l_i| < 2^{s_i}} \left(\min_{l \in \mathbb{N}_0^\infty : 2^{s_i - 1} \leq |l_i| < 2^{s_i}} c_l \right) \langle f, \phi_l \rangle^2 \\ \leq &\sum_{l \in \mathbb{N}_0^\infty : 2^{s_i - 1} \leq |l_i| < 2^{s_i}} c_l^2 \langle f, \phi_l \rangle^2 \end{aligned}$$

である．$s \in \mathbb{N}_0^\infty$ に対して，

$$\begin{aligned} &\min_{l \in \mathbb{N}_0^\infty : 2^{s_i - 1} \leq |l_i| < 2^{s_i}} c_l \\ &= \max_{i \in \mathbb{N}} \left\{ (2\pi)^{2a_i} 2^{2a_i (s_i - 1)} \right\}_i \\ &= \max_{i \in \mathbb{N}} \left\{ (\pi)^{2a_i} 2^{2a_i s_i} \right\}_i \\ &> 2^{2 \max\{a_i s_i\}_{i=1}^\infty} \end{aligned}$$

なので，

$$\sum_{s \in \mathbb{N}_0^\infty} 2^{2 \max\{a_i s_i\}_{i=1}^\infty} \|\delta_s\|_2^2 < \infty$$

である．よって，$f \in \mathcal{S}_{a,\infty}$ ならば，f は $p = 2, \theta = 2$ の異方周波空間に属することが分かる．このように，本研究で定義した，混合・異方周波空間は既存の混合・異方性の性質を持つ関数クラスを含むことが分かる．

3.4　拡張畳み込みニューラルネットワーク (拡張 CNN) の定義

この節では，本研究で用いるモデルを定義する．$A_i \in \mathbb{R}^{d_{i+1} \times d_i}$，$b_i \in \mathbb{R}^{d_{i+1}}$，$\eta(x) = \max\{x, 0\}$ を用いて，

$$(A_L \eta(\cdot) + b_L) \circ \cdots \circ (A_i \eta(\cdot) + b_i) \circ \cdots \circ (A_1 x + b_1)$$

で表されるモデルを，全結合 ReLU ニューラルネットワーク (ReLU FNN) と呼ぶ．ただし，η はベクトルの要素ごとに適用されるものとする．ここで，活性化関数 $\eta(\cdot)$ のことを ReLU と呼ぶ．また，ある定数 $L,\ W,\ S \in \mathbb{N},\ B > 0$ を用いて，集合 $\Phi(L, W, B, S)$ を，

$$\begin{aligned} \Phi(L, W, B, S) := \{ &f(x) = A_L \eta(\cdot) + b_L \circ \cdots \circ A_i \eta(\cdot) + b_i \circ \cdots \circ A_1 x + b_1 : \\ &\max_{i=1,\ldots,L} \|A_i\|_\infty \vee \|b_i\|_\infty \leq B,\ \sum_{i=1}^{L} \|A_i\|_0 + \|b_i\|_0 \leq S,\ \max_{i=1,\ldots L} d_i \leq W \Bigg\} \end{aligned}$$

で定義する．ただし，$\|\cdot\|_\infty$ はベクトルや行列の要素の中での最大値，$\|\cdot\|_0$ は要素中の非ゼロパラメータの数する． 次に，拡張 CNN の定義をする．$C \in \mathbb{N}$, $\mathbb{R}^{C\times\infty} := \left\{(x_1, \dots x_i, \dots) : x_i \in \mathbb{R}^C\right\}$ とする．無限次元ベクトル列 $X \in \mathbb{R}^{C\times\infty}$ が与えられたとき，$T \in \mathbb{N}$, $w \in \mathbb{R}^{C\times T}$, $h \in \mathbb{N}$ を用いて，$w \star_h X : \mathbb{R}^{C\times\infty} \to \mathbb{R}^\infty$ を，$(w \star_h X)_k = \sum_{i=1}^{C}\sum_{j=1}^{T} w_{i,j}x_{i,h(j-1)+k}$ と定義し，$w \star_h X$ を間隔 h，幅 T の拡張畳み込みと呼ぶ．$h = 1$ の場合は通常の畳み込みという．また，フィルター $F \in \mathbb{R}^{C'\times C\times T}$ が与えらえたとき，$\mathrm{Conv}_{h,F} : \mathbb{R}^{C\times\infty} \to \mathbb{R}^{C'\times\infty}$ を，

$$\mathrm{Conv}_{h,W}(X) = \begin{pmatrix} F_{1,:,:} \star_h X \\ \vdots \\ F_{C',:,:} \star_h X \end{pmatrix}$$

と定義する．ここで，C はチャネル数と呼ばれる量である．さらに，拡張 CNN は次のように定義される：

定義 7 (拡張 CNN)**.** L', $W' \in \mathbb{N}$, $l \in [L']$, $C_l \in \mathbb{N}$, $F_l \in \mathbb{R}^{C_{l+1},\times C_l\times W'}$，FNN をある全結合ニューラルネットワークとして，

$$f(X) := \left(\mathrm{FNN} \circ \mathrm{Conv}_{W'L'-1,F_{L'}} \circ \cdots \circ \mathrm{Conv}_{W'^l,F_l} \circ \cdots \circ \mathrm{Conv}_{1,F_1} \circ X\right)_1$$

の形で表されるモデルを拡張 CNN と呼ぶ．ただし，FNN は無限列の各要素に適用されるものとする．

$L', L, W', C, S, W \in \mathbb{N}$, $B', B > 0$ を用いて，拡張 CNN の集合，$\mathcal{P}(L', B', W', C, L, W, B, S)$ を，

$$\begin{aligned}&\mathcal{P}(L', B', W', C, L, W, B, S)\\ &= \left\{\left(\mathrm{FNN} \circ \mathrm{Conv}_{W'L'-1,F_{L'}} \circ \cdots \circ \mathrm{Conv}_{W'^l,F_l} \circ \cdots \circ \mathrm{Conv}_{1,F_1} \circ X\right)_1 :\right.\\ &\left. F_l \in \mathbb{R}^{C\times C\times H},\ \|F\|_\infty \le B',\ \mathrm{FNN} \in \Phi(L, W, B, S)\right\}\end{aligned}$$

で定義する．簡単のために，$\mathcal{P}$ と略記する場合もある．また，$T = 1$ の場合，通常の CNN と一致する．本研究では，すべての証明においてチャネル数が全層を通して一定の値の拡張 CNN を考えれば十分である．そのため，$\mathcal{P}$ はチャネル数が全層通して一定の拡張 CNN の集合とした．また，これ以降，推定精度の議論をする際に，$\|\cdot\|_\infty$ ノルムに関して有界なモデルを考える必要性が出てくる．そのため，ある $B_f > 0$ に関して clipping された拡張 CNN を，$L', L, W', C, S, W \in \mathbb{N}$, $B', B > 0$ を用いて，

$$\begin{aligned}&\bar{\mathcal{P}}(B_f, L', B', W', C, L, W, B, S)\\ &:= \left\{\bar{f} : X \mapsto \max\{-B_f, \min\{B_f, f(X)\}\} : f \in \mathcal{P}(L', B', W', C, L, W, B, S)\right\}\end{aligned}$$

で定義する．

第 4 章

ニューラルネットワークによる近似誤差の解析

4.1　全結合ニューラルネットワークによる近似誤差解析

本節では，全結合ニューラルネットワークを用いた場合の近似誤差解析を行う．これ以降，$T > 0,\ \gamma : \mathbb{N}_0^\infty \to \mathbb{R}_{>0}$ に対して，

$$I(T,\gamma) = \{i \in \mathbb{N} : \exists s \in \mathbb{N}_0^\infty,\ s_i \neq 0,\ \gamma(s) \leq T\}$$

とする．また，近似誤差の解析を行うにあたって重要となるのが，次の量である．

定義 8. 周波数方向複雑さと軸方向複雑さ

$$d_{\max}(T,\gamma) := |I(T,\gamma)|$$

を軸方向複雑さ，

$$f_{\max}(T,\gamma) := \max_{s \in \mathbb{N}_0^\infty : \gamma(s) \leq T} \max_{i \in \mathbb{N}} s_i$$

を周波数方向複雑さと呼ぶ．

これら 2 つの概念の意味を説明する．軸方向複雑さは，ある特定の近似誤差を達成するためには，与えられた無限次元列データ $X = (x_1, \ldots, x_i, \ldots)$ から何個の成分を，周波数方向複雑さは，どの周波数までを抜き出せばよいかをそれぞれ定義している．これらの量は近似誤差解析において重要である．次に，FNN による近似定理について説明する．これ以降，

$$v := \left(\frac{1}{p} - \frac{1}{2}\right)_+,\ \alpha(\gamma) := \sup_{s \in \mathbb{N}_0^\infty} \frac{\sum_{i=1}^\infty s_i}{\gamma(s)},\ S(T,\gamma) := \sum_{s \in \mathbb{N}_0^\infty : \gamma(s) < T} 2^s,$$

とする．ただし，$(x)_+ := \max\{x, 0\}$ と定義する．

定理 9 (FNN による γ-平滑周波空間の近似誤差)**.** 関数 $\gamma,\ \gamma' : \mathbb{N}_0^\infty \to \mathbb{R}_{>0}$ が，

$$\gamma'(s) < \gamma(s),\ v\alpha(\gamma) < 1,\ v\alpha(\gamma') < 1,$$

を満たすと仮定する．このとき，ある定数 $B_f > 0$ を用いて，$\|f\|_{L^\infty} \leq B_f$，$f \in \mathcal{F}^{\gamma}_{p,\theta}$ $(p \geq 1,\ \theta \geq 1)$ であれば，

$$d_{\max} = \begin{cases} d_{\max}(\gamma) & (1 \leq \theta \leq 2) \\ d_{\max}(\gamma') & (2 < \theta) \end{cases},\ f_{\max} = \begin{cases} f_{\max}(\gamma) & (1 \leq \theta \leq 2) \\ f_{\max}(\gamma') & (2 < \theta) \end{cases},$$
$$S = \begin{cases} S(T,\gamma) & (1 \leq \theta \leq 2) \\ S(T,\gamma') & (2 < \theta) \end{cases},$$

と，ある正の定数 $K,\ K'$ を用いて，

$$\begin{aligned} L &= 2K \max\left\{d_{\max}^2, T^2, (\log S)^2, \log f_{\max}\right\}, \\ W &= 21 d_{\max} S, \\ S &= 1764 K d_{\max}^2 \max\left\{d_{\max}^2, T^2, (\log S)^2, \log f_{\max}\right\} S, \\ B &= (\sqrt{2})^{d_{\max}} K', \end{aligned}$$

とすれば，$(x_i : i \in I(T,\gamma))$ を入力として受け取る $d_{\max}$ 次元の入力をもつニューラルネットワーク

$$\hat{R}_T \in \Phi(L, W, S, B)$$

が存在して，$f' : [0,1]^\infty \to \mathbb{R}$ を，

$$f'(X) := \hat{R}_T\left((x_i : i \in I(T,\gamma))\right)$$

とすれば，

$$\|f - f'\|_2 \lesssim \begin{cases} 2^{-(1-v\alpha(\gamma))T} \|f\|_{\mathcal{F}^{\gamma}_{p,\theta}} & (1 \leq \theta \leq 2) \\ 2^{-(1-v\alpha(\gamma'))T} \left(\sum_{T \leq \gamma'(s)} 2^{\frac{2\theta}{\theta-2}(\gamma'(s)-\gamma(s))}\right)^{1/2-1/\theta} \|f\|_{\mathcal{F}^{\gamma}_{p,\theta}} & (2 < \theta) \end{cases}$$

が成立する．

この定理では，入力データ X から必要となる $d_{\max}$ 個のデータを抜き出すことができた場合には，上記の近似誤差を達成することを主張している．本定理により，関数の周波数成分の減衰度合いがニューラルネットワークによる近似精度に対して与える影響が明らかになっている．また，$S(T,\gamma)$，$d_{\max}(\gamma)$，$f_{\max}(\gamma)$ により近似精度が決定されるため，例え大きな次元が入力されたとしても，その次元に依存せず，この 3 つの量が本質的に重要であることが分かる．つまり，滑らかさを定義する関数 $\gamma(s) : \mathbb{N}_0^\infty \to \mathbb{R}_{>0}$ によって，近似に必要なネットワークの複雑さが決定されるということである．具体的には，近似に必要な基底の数，$\gamma(s) < T$ を満たすような $s \in \mathbb{N}_0^\infty$ の個数，近似に必要な最大周波数により，近似の難しさが決定されるということである．本定理は，$f \in \mathcal{F}^{\gamma}_{p,\theta}$ を近似するようなニューラルネットワークを構成することにより証明されている．具体的には，三角多項式をある精度で近似するニューラルネットワークを構成することが可能であること，$f \in \mathcal{F}^{\gamma}_{p,\theta}$ が三角多項式で分解されていることを用いた．多変数の掛け算がニューラルネットワークにより任意の精度で近似できることは，[21] で証明されている．また，三角多項式があるニューラルネットワークにより所望の精度で近似可能なことは，[15] により示されている．本定理は，これらの事実を用いて証明されている．

4.2　有限次元入力を受け取る場合

本節では，$\mathcal{F}^{\gamma}_{p,\theta}([0,1]^{\infty})$ が有限次元入力を受け取る関数を表現可能であることを説明する．

$$J_d := \{s \in \mathbb{N}_0^{\infty} : s_i = 0 \ (i = d+1, \dots)\}$$

とする．ここで，$\gamma(s) : \mathbb{N}^{\infty} \to \mathbb{R}_{>0}$ を

$$\begin{cases} \gamma(s) < \infty & (s \in J_d) \\ \gamma(s) = \infty & (s \notin J_d) \end{cases}$$

とする．$d \in \mathbb{N}$ を用いて，

$$\mathcal{F}^{\gamma}_{p,\theta,d}([0,1]^{\infty}) := \left\{ f = \sum_{l \in \mathbb{N}_0^{\infty}} \langle f, \phi_l \rangle \phi_l : \left(\sum_{s \in J_d} 2^{\theta\gamma(s)} \|\delta_s(f)\|_p^{\theta} \right)^{1/\theta} < \infty, \ \forall s \notin J_d, \ \delta_s(f) = 0 \right\}$$

と定義する．任意の $f \in \mathcal{F}^{\gamma}_{p,\theta,d}([0,1]^{\infty})$ に関して，$\forall s \notin J_d, \ \delta_s(f) = 0$ が成り立つことから，

$$f(x_1, \dots, x_d, x_{d+1}, \dots, x_i, \dots) = f(x_1, \dots, x_d, x'_{d+1}, \dots, x'_i \dots) \ (x_i, x'_i \in [0,1], \ i = d+1, \dots)$$

が成立する．したがって，$f \in \mathcal{F}^{\gamma}_{p,\theta,d}([0,1]^{\infty})$ ならば，関数 f の値は，$(x_1, \dots x_d)$ の値のみに依存して決定されていると言える．ここで，

$$f_d(x_1, \dots x_d) := f(x_1, \dots, x_d, 0, \dots)$$

で定義すれば，$f, f' \in \mathcal{F}^{\gamma}_{p,\theta,d}([0,1]^{\infty})$ に対して，

$$\|f - f'\|_2 = \|f_d - f'_d\|_2$$

となる．ここで，$\mathcal{F}^{\gamma}_{p,\theta,d}([0,1]^{\infty})$ に関して定理 9 の証明と全く同じ議論をすれば，定理 9 が $\mathcal{F}^{\gamma}_{p,\theta,d}([0,1]^{\infty})$ に対しても適用可能であることが分かる．したがって，$[0,1]^d$ 上の γ—平滑関数に関しても，定理 9 が成立することが分かる．これにより，$d \gg n$ を満たすような状況であっても，平滑度 $\gamma(s)$ のみに依存して近似の難しさが決定されることが分かる．また，$\forall s \notin J_d$ に対して，$\gamma(s) = \infty$ という状況は，f は $d+1$ 番目以降のインデックスに一切依存しない，つまり d 次元の入力により決定される関数であることを示している．

4.3　定理 9 の証明

Proof. $R_T(f)$ を次のように定める：

$$R_T(f) := \begin{cases} \sum_{s \in \mathbb{N}_0^{\infty} : \gamma(s) < T} \delta_s(f) & (1 \le \theta \le 2) \\ \sum_{s \in \mathbb{N}_0^{\infty} : \gamma'(s) < T} \delta_s(f) & (2 < \theta). \end{cases}$$

この時，δ_s は，$c_k > 0$ と虚数 i を用いて，

$$\delta_s(f)(x) = \sum_{\lceil 2^{s_i - 1} \rceil \le |k_i| < 2^{s_i}} c_k \exp(\mathrm{i}\langle k, x \rangle)$$

とあらわされるので，$1 \leq p \leq 2$ の時，[12] の定理 1 より，$v = (\frac{1}{p} - \frac{1}{2})_+$ を用いて，

$$\|\delta_s(f)\|_2 \leq 2^{vs}\|\delta_s(f)\|_p \tag{4.1}$$

が成立する．ただし，$2^{vs} = 2^{v\sum_{i=1}^{\infty} s_i}$ としている．さらに，$2 < p$ の場合，コーシー・シュワルツの不等式より

$$\begin{aligned}
&\|\delta_s(f)\|_2^2 \\
&\qquad = \int_{[0,1]^\infty} \delta_s(f)^2 \mathrm{d}\lambda^\infty \\
&\qquad \leq \left(\int_{[0,1]^\infty} \delta_s(f)^p \mathrm{d}\lambda^\infty \right)^{2/p} \left(\int_{[0,1]^\infty} \mathrm{d}\lambda^\infty \right)^{1-2/p} \\
&\qquad = \|\delta_s(f)\|_p^2
\end{aligned}$$

である．したがって，(4.1) が $1 \leq p < \infty$ で成立する．次に，f を R_T で近似することを考える．$1 \leq \theta \leq 2$ と $2 < \theta$ の場合に分けて議論をする．

1. $1 \leq \theta \leq 2$:
   δ_s の直交性を用いれば，

$$\begin{aligned}
\|f - R_T(f)\|_2^\theta = (\|f - R_T(f)\|_2^2)^{\theta/2} & \\
&\leq \left(\sum_{T \leq \gamma(s)} \|\delta_s(f)\|_2^2 \right)^{\theta/2} && (4.2) \\
&\leq \sum_{T \leq \gamma(s)} (2^{vs}\|\delta_s(f)\|_p)^\theta \\
&= \sum_{T \leq \gamma(s)} (2^{\gamma(s)} 2^{vs-\gamma(s)}\|\delta_s(f)\|_p)^\theta && (4.3)
\end{aligned}$$

   が成立する．ここで，(4.2) の変形では (4.1) を用いた．次に，$v\alpha < 1$ の仮定を用いると，$s \in \mathbb{N}_0^\infty$ ならば，$T \leq \gamma(s)$ において，

$$\begin{aligned}
2^{vs-\gamma(s)} &\leq 2^{(v\alpha-1)\gamma(s)} \\
&\leq 2^{(v\alpha-1)T}
\end{aligned}$$

   と評価できる．よって，(4.3) にこれを適用すると，

$$\begin{aligned}
\sum_{T \leq \gamma(s)} (2^{\gamma(s)} 2^{vs-\gamma(s)}\|\delta_s(f)\|_p)^\theta &\leq 2^{-\theta(1-v\alpha)T} \sum_{T \leq \gamma(s)} \left(2^{\gamma(s)}\|\delta_s(f)\|_p \right)^\theta \\
&\leq 2^{-\theta(1-v\alpha)T}\|f\|_{\mathcal{F}^\gamma_{p,\theta}}^\theta
\end{aligned}$$

   が得られる．

2. $2 < \theta$:

$\delta_s(f)$ の直交性を用いれば，

$$\|f - R_T(f)\|_2^2 = \sum_{T \le \gamma'(s)} \|\delta_s(f)\|_2^2$$

である．また，

$$\begin{aligned}\sum_{T \le \gamma'(s)} \|\delta_s(f)\|_2^2 &\le \sum_{T \le \gamma'(s)} \left(2^{vs}\|\delta_s(f)\|_p\right)^2 \\ &= \sum_{T \le \gamma'(s)} \left(2^{vs-\gamma'(s)}2^{\gamma'(s)-\gamma(s)}2^{\gamma(s)}\|\delta_s(f)\|_p\right)^2\end{aligned}$$

が得られる．ここで，$1 \le \theta \le 2$ の証明の場合と同様に $v\alpha' < 1$ の仮定を用いると，

$$\begin{aligned}2^{vs-\gamma'(s)} &\le 2^{(v\alpha'-1)\gamma'(s)} \\ &\le 2^{(v\alpha'-1)T}\end{aligned}$$

と評価できて，これを適用すると，

$$\begin{aligned}&\sum_{T \le \gamma'(s)} \left(2^{vs-\gamma'(s)}2^{\gamma'(s)-\gamma(s)}2^{\gamma(s)}\|\delta_s(f)\|_p\right)^2 \\ &\le 2^{-2(1-v\alpha')T} \sum_{T \le \gamma'(s)} \left(2^{\gamma'(s)-\gamma(s)}2^{\gamma(s)}\|\delta_s(f)\|_p\right)^2\end{aligned}$$

となる．ただし，$\alpha' = \alpha(\gamma')$ である．ここで，コーシーシュワルツの不等式を用いれば，

$$\begin{aligned}&2^{-2(1-v\alpha')T} \sum_{T \le \gamma'(s)} \left(2^{\gamma'(s)-\gamma(s)}2^{\gamma(s)}\|\delta_s(f)\|_p\right)^2 \\ &\le 2^{-2(1-v\alpha')T} \left[\sum_{T \le \gamma'(s)} \left(2^{\gamma(s)}\|\delta_s(f)\|_p\right)^\theta\right]^{2/\theta} \times \left[\sum_{T \le \gamma'(s)} \left(2^{\gamma'(s)-\gamma(s)}\right)^{2/(1-2/\theta)}\right]^{(1-2/\theta)} \\ &\lesssim 2^{-2(1-v\alpha')T} \left[\sum_{T \le \gamma'(s)} 2^{\frac{2\theta}{\theta-2}(\gamma'(s)-\gamma(s))}\right]^{1-2/\theta} \|f\|^2_{\mathcal{F}^\gamma_{p,\theta}}\end{aligned}$$

が得られる．

ここで，$1 \le \theta \le 2$ の場合に，R_T をニューラルネットワークで近似することを考える．[15] の定理 4.1 より，C_1, C_2 をある 0 より大きい定数，

$$L_{\hat{\psi}} = C_1 \left[\left(\log \frac{1}{\epsilon}\right)^2 + \log\left(f_{\max}\right)\right]$$

として，あるニューラルネットワーク $\hat{\psi}_{l_i} \in \Phi(L_{\hat{\psi}},\ 21,\ C_2,\ 21^2 L_{\hat{\psi}})$ が存在して，

$$\|\psi_{l_i} - \hat{\psi}_{l_i}\|_{L^\infty([0,1])} \leq \epsilon$$

が成り立つ．さらに，$\psi_{l_i} \in [-\sqrt{2}, \sqrt{2}]$ であることを考慮すれば，

$$\|\max\{-\sqrt{2}, \min\{\sqrt{2}, \hat{\psi}_{l_i}\}\} - \psi_{l_i}\|_{L^\infty([0,1])} \leq \epsilon$$

である．また，$x \in \mathbb{R}$ に対して，

$$\max\{-\sqrt{2}, \min\{\sqrt{2}, x\}\} = \Big[\eta(x) - \eta(x - \sqrt{2})\Big] + \Big[-\eta(-x) + \eta(-x + \sqrt{2})\Big]$$

であるので，$\max\{-\sqrt{2}\min\{\sqrt{2}, x\}\} \in \Phi(2, 4, \sqrt{2}, 16)$ となる．したがって，

$$\max\{-\sqrt{2}, \min\{\sqrt{2}, \hat{\psi}_{l_i}\}\} \in \Phi(L_{\hat{\psi}} + 2,\ 21,\ \max\{C_2, \sqrt{2}\},\ 21^2(L_{\hat{\psi}} + 2))$$

である．これにより，$[\sqrt{2}, \sqrt{2}]$ に値をとるニューラルネットワークが構成できる．これ以降，

$$\hat{\psi}_{l_i} = \max\{-\sqrt{2}, \min\{\sqrt{2}, \hat{\psi}_{l_i}\}\},\ L_{\hat{\psi}} = C_1 \left[\left(\log \frac{1}{\epsilon}\right)^2 + \log\left(f_{\max}\right)\right] + 2$$

として話を進める．また，[21] の結果により，$x_i \in [0, 1]$ $(i = 1 \ldots d_{\max})$ の下で，

$$L_\times = \left\lceil \log\left(\frac{3^{d_{\max}}}{\epsilon} + 5\right)\right\rceil \lceil \log d_{\max} \rceil,\ W_\times = 6 d_{\max},\ S_\times = L_\times W_\times^2$$

と，ある定数 $B_\times > 0$ を用いると，あるニューラルネットワーク $\phi_\times \in \Phi(L_\times, W_\times, B_\times, S_\times)$ が存在して，

$$\left\| \phi_\times - \prod_{i=1}^{d_{\max}} x_i \right\|_{L^\infty([-1,1]^{d_{\max}})} \leq \epsilon$$

が成立する．$\hat{\psi}_{l_i} \leq \sqrt{2}$ であるから，

$$\left\| \phi_\times \left(\frac{\hat{\psi}_{l_1}}{\sqrt{2}}, \ldots, \frac{\hat{\psi}_{l_{d_{\max}}}}{\sqrt{2}}\right) - \prod_{i=1}^{d_{\max}} \frac{\hat{\psi}_{l_i}}{\sqrt{2}} \right\|_{L^\infty([0,1]^\infty)} \leq \epsilon$$

$$\Rightarrow \left\| (\sqrt{2})^{d_{\max}} \phi_\times \left(\frac{\hat{\psi}_{l_1}}{\sqrt{2}}, \ldots, \frac{\hat{\psi}_{l_{d_{\max}}}}{\sqrt{2}}\right) - \prod_{i=1}^{d_{\max}} \hat{\psi}_{l_i} \right\|_{L^\infty([0,1]^\infty)} \leq \sqrt{2}^{d_{\max}} \epsilon$$

である．また，

$$\begin{aligned}
&\left\| \prod_{i=1}^{d_{\max}} \hat{\psi}_{l_i} - \prod_{i=1}^{d_{\max}} \psi_{l_i} \right\|_{L^\infty([0,1]^\infty)} \\
&= \left\| \sum_{j=0}^{d_{\max}-1} \prod_{i=1}^{j} \hat{\psi}_{l_i} \prod_{i=j+1}^{d_{\max}} \psi_{l_i} - \prod_{i=1}^{j+1} \hat{\psi}_{l_i} \prod_{i=j+2}^{d_{\max}} \psi_{l_i} \right\|_{L^\infty([0,1]^\infty)} \\
&= \left\| \sum_{j=0}^{d_{\max}-1} \prod_{i=1}^{j} \hat{\psi}_{l_i} \prod_{i=j+2}^{d_{\max}} \psi_{l_i} \left(\hat{\psi}_{l_{j+1}} - \psi_{l_{j+1}} \right) \right\|_{L^\infty([0,1]^\infty)} \\
&\leq \sqrt{2}^{d_{\max}} \sum_{j=0}^{d_{\max}-1} \left\| \hat{\psi}_{l_{j+1}} - \psi_{l_{j+1}} \right\|_{L^\infty([0,1]^\infty)} \\
&\leq \sqrt{2}^{d_{\max}} (d_{\max}+1)\epsilon
\end{aligned}$$

である．したがって，三角不等式より，

$$\begin{aligned}
&\left\| (\sqrt{2})^{d_{\max}} \phi_\times \left(\frac{\hat{\psi}_{l_1}}{\sqrt{2}}, \ldots, \frac{\hat{\psi}_{l_{d_{\max}}}}{\sqrt{2}} \right) - \prod_{i=1}^{d_{\max}} \psi_{l_i} \right\|_{L^\infty([0,1]^\infty)} \\
&\leq \left\| (\sqrt{2})^{d_{\max}} \phi_\times \left(\frac{\hat{\psi}_{l_1}}{\sqrt{2}}, \ldots, \frac{\hat{\psi}_{l_{d_{\max}}}}{\sqrt{2}} \right) - \prod_{i=1}^{d_{\max}} \hat{\psi}_{l_i} \right\|_{L^\infty([0,1]^\infty)} + \left\| \prod_{i=1}^{d_{\max}} \hat{\psi}_{l_i} - \prod_{i=1}^{d_{\max}} \psi_{l_i} \right\|_{L^\infty([0,1]^\infty)} \\
&\leq (\sqrt{2})^{d_{\max}} (d_{\max}+1)\epsilon
\end{aligned}$$

が成立する．ここで，

$$\hat{R}_T(f) := \sum_{\gamma(s)<T} \sum_{l \in J(s)} (\sqrt{2})^{d_{\max}} \langle f, \psi_l \rangle \phi_\times \left(\frac{\hat{\psi}_{l_1}}{\sqrt{2}}, \ldots, \frac{\hat{\psi}_{l_{d_{\max}}}}{\sqrt{2}} \right)$$

を用いると，

$$\begin{aligned}
&\|\hat{R}_T - R_T\|_{L^\infty([-1,1]^d_{\max})} \\
&\leq \left\| \sum_{s \in \mathbb{N}_0^\infty : \gamma(s)<T} \sum_{l \in J(s)} (\sqrt{2})^{d_{\max}} \langle f, \phi_l \rangle \left(\phi_\times (\frac{\hat{\psi}_{l_1}}{\sqrt{2}}, \ldots, \frac{\hat{\psi}_{l_{d_{\max}}}}{\sqrt{2}}) - \frac{\phi_l}{\sqrt{2}^{d_{\max}}} \right) \right\|_{L^\infty([-1,1]^{d_{\max}})} \\
&\leq (\sqrt{2})^{d_{\max}} B_f (d_{\max}+1) S(\gamma, T)\epsilon
\end{aligned}$$

となる．ただし，最後の変形において $\langle f, \psi_l \rangle \leq \|f\|_2 \leq B_f$ を用いた．したがって，

$$\epsilon = \frac{2^{-T}}{(\sqrt{2})^{d_{\max}} B_f (d_{\max}+1) S(\gamma, T)} \tag{4.4}$$

とすれば，

$$\begin{aligned}
&\|f - \hat{R}_T\|_2 \\
&\leq \|f - R_T\|_2 + \|\hat{R}_T - R_T\|_{L^\infty([-1,1]^{d_{\max}})} \\
&\lesssim 2^{-T} \|f\|_{\mathcal{F}^\gamma_{p,\theta}}
\end{aligned}$$

が成り立つ．また，$\hat{R}_T$ はニューラルネットワーク $\phi_\times\left(\frac{\hat{\psi}_{l_1}}{\sqrt{2}},\ldots,\frac{\hat{\psi}_{l_{d_{\max}}}}{\sqrt{2}}\right)$ の線形結合である．ここで，

$$\begin{aligned}
L &= L_{\hat{\phi}} + L_\times + 1,\\
W &= 21 d_{\max} S(\gamma, T),\\
S &= (21^2 d_{\max} L_{\hat{\phi}} + L_\times W_\times^2 + 1) S(\gamma, T),\\
B &= \max\{(\sqrt{2})^{d_{\max}} B_f, B_\times, C_2\},
\end{aligned}$$

とすれば，$\hat{R}_T \in \Phi(L, W, S, B)$ となる．よって，式 (4.4) を，$L_\times$，$L_{\hat{\phi}}$ に代入すれば，

$$\begin{aligned}
L_{\hat{\psi}} &= C_1 \left\lceil \left(\log \frac{1}{\epsilon}\right)^2 + \log(f_{\max}) + 2 \right\rceil \\
&= C_1 \left\lceil \left(T \log 2 + d_{\max} \log \sqrt{2} + \log S(\gamma, T) + \log B_f + \log d_{\max} + 1\right)^2 + \log f_{\max} \right\rceil \\
&\leq \left\lceil C_1 \left(6 \max\{\log B_f, \log 2\}\right)^2 \right\rceil \left\lceil \left(\max\{d_{\max}^2, T^2, (\log S(\gamma, T))^2, \log f_{\max}, 2\}\right) \right\rceil
\end{aligned}$$

であり，

$$\begin{aligned}
L_\times &= \left\lceil \log\left(\frac{3^{d_{\max}}}{\epsilon} + 5\right) \right\rceil \lceil \log d_{\max} \rceil \\
&\leq \left\lceil \max\left\{ \log\left(\frac{3^{d_{\max}}}{\epsilon}\right), \log 5 \right\} + \log 2 \right\rceil \lceil \log d_{\max} \rceil \\
&\leq \lceil 6 \max\{\log B_f, \log 5\} \rceil \lceil \max\{d_{\max}, T, \log S(\gamma, T)\} \rceil \lceil \log d_{\max} \rceil
\end{aligned}$$

ここで，$K = 2 \max\left\{ \lceil 6 \max\{\log B_f, \log 5\} \rceil, \left\lceil C_1 \left(6 \max\{\log B_f, \log 2\}\right)^2 \right\rceil \right\}$ とすれば，

$$L \leq 2K \max\left\{d_{\max}^2, T^2, (\log S(\gamma, T))^2, \log f_{\max}\right\}$$

また，S について，

$$\begin{aligned}
S &= (21^2 d_{\max} L_{\hat{\phi}} + L_\times W_\times^2 + 1) S(\gamma, T) \\
&\leq 2K(21^2 d_{\max} + 36 d_{\max}^2) \max\left\{d_{\max}^2, T^2, (\log S(\gamma, T))^2, \log f_{\max}\right\} S(\gamma, T) \\
&\leq 4 \times 21^2 K d_{\max}^2 \max\left\{d_{\max}^2, T^2, (\log S(\gamma, T))^2, \log f_{\max}\right\} S(\gamma, T)
\end{aligned}$$

また，$2 < \theta$ の場合も同様の手順で証明できる．したがって，補題が示された． □

第 5 章

畳み込みニューラルネットワークによる推定誤差の解析

5.1　CNN・拡張 CNN による近似・推定誤差解析

本章では，CNN・拡張 CNN による近似・推定誤差について議論をする．前章の定理 9 では，$\mathcal{F}^{\gamma}_{p,\theta}$ 中の関数を，全結合ニューラルネットワークにより近似することを考えた．それによれば，ある近似誤差を達成するために必要なインデックスが $I(T,\gamma)$ により定まり，X 中のそのインデックスに対応するデータのみを入力として受け取ればよい．しかし，実際にデータ X がデータとして得られた場合，どのインデックスが入力として必要かは与えられておらず，データから学習する必要がある．本章において，ある条件の下で，CNN・拡張 CNN 型のアーキテクチャを用いれば，必要なインデックスをデータから学習することが可能であることを示す．また，これらのアーキテクチャを用いれば，混合・異方周波空間に属する関数に対して，それら関数クラスの滑らかさに依存した近似・推定誤差が達成可能であることを示す．

5.2　多項式オーダーで滑らかさが上昇する場合

ここでは，滑らかさを定める単調増加かつ正の数列 $a = (a_1, \dots, a_i, \dots)$ について，ある $0 < q < \infty$ が存在して $a_i = \Omega(i^q)$ となるような場合を考える．

定理 10 (多項式オーダーの下での近似定理)**.** ある $0 < q < \infty$ が存在して，$a_i = \Omega(i^q)$ であれば，

1. $\gamma(s) = \langle a, s \rangle$ の場合:

$v/a_1 < 1$ であれば，ある定数 $B' > 0$ を用いて，

$$L' = 1,\ W' \sim T^{\frac{1}{q}},\ C \sim T^{\frac{1}{q}},\ L \sim \max\left\{T^{\frac{2}{q}}, T^2\right\},\ W \sim \left(\prod_{i=2}^{\infty} \frac{1}{1-2^{\frac{-(a_i-a_1)}{a_1}}},\ \right) T^{\frac{1}{q}} 2^{\frac{T}{a_1}},$$

$$S \sim \left(\prod_{i=2}^{\infty} \frac{1}{1-2^{\frac{-(a_i-a_1)}{a_1}}}\right) T^{\frac{2}{q}} \max\left\{T^{\frac{2}{q}}, T^2\right\} 2^{\frac{T}{a_1}},\ B \sim (\sqrt{2})^{T^{\frac{1}{q}}},$$

とすれば，$f \in U(\mathcal{F}^{\gamma}_{p,\theta})$ $(p \geq 1,\ 1 \leq \theta)$ と，ある $B_f > 0$ が存在して，$\|f\|_{\infty} \leq B_f$，が成り立つような任意の関数 f に対して，$f' \in \mathcal{P}(L', B', W', C, L, W, B, S)$ が存在して，

$$\|f' - f\|_2 \lesssim 2^{-\left(1-\frac{v}{a_1}\right)T}$$

が成立する．

2. $\gamma(s) = \max_i\{a_i s_i\}$ の場合:
   $\tilde{a} = \sum_{i=1}^{\infty} \frac{1}{a_i}$ とする．$v\tilde{a} < 1$ の下で，ある定数 $B' > 0$ を用いて，

$$L' = 1,\ W' \sim T^{\frac{1}{q}},\ C \sim T^{\frac{1}{q}},\ L \sim \max\left\{T^{\frac{2}{q}}, T^2\right\},\ W \sim T^{\frac{1}{q}} 2^{\tilde{a}T},$$
$$S \sim T^{\frac{2}{q}} \max\left\{T^{\frac{2}{q}}, T^2\right\} 2^{\tilde{a}T},\ B \sim (\sqrt{2})^{T^{\frac{1}{q}}},$$

   とすれば，$f \in U(\mathcal{F}^{\gamma}_{p,\theta})$ $(p \geq 1,\ 1 \leq \theta \leq 2)$ と，ある $B_f > 0$ が存在して，$\|f\|_{\infty} \leq B_f$，が成り立つような任意の関数 f に対して，$f' \in \mathcal{P}(L', B', W', C, L, W, B, S)$ が存在して，

$$\|f' - f\|_2 \lesssim 2^{-(1-v\tilde{a})T}$$

   が成立する．

この定理より，層数，幅，パラメータ数，パラメータの大きさはともに T とスムーズネスに依存する量によって決定されることが見て取れる．また，前章の定理 9 では，近似に必要なインデックス集合 $I(T, \gamma)$ が与えられたと仮定して，近似誤差を導出していた．本定理では，そのような仮定はしておらず，代わりに $a_i = \Omega(i^q)$ という条件を課している．この条件の下で，CNN は必要なインデックス $I(T, \gamma)$ を抽出する役割を負っている．次に，これらモデルの回帰問題における推定誤差について考察する．回帰問題では，観測値 $(X_i, y_i)_{i=1}^n$ が，$\sigma > 0,\ f^{\circ} : \mathbb{R}^{\infty} \to \mathbb{R}$，を用いて定義されるモデル，

$$y_i = f^{\circ}(X_i) + \xi_i\ (X_i \sim P_X,\ \xi_i \sim \mathrm{N}(0, \sigma^2))$$

により生成される．この時，モデル $\mathcal{P}$ の中での経験誤差最小化元

$$\hat{f} = \mathrm{argmin}_{\hat{f} \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - y_i)^2$$

により，関数 f を推定することを考える．汎化誤差の観測値に関する期待値について，

$$\mathrm{E}_{(X_i,y_i)_{i=1}^n\sim P^n}[\mathrm{E}_{P_X}[(\hat{f}(X)-Y)^2]]=\mathrm{E}_{(X_i,y_i)_{i=1}^n\sim P^n}[\mathrm{E}_{P_X}[(\hat{f}(X)-f^\circ(X))^2]]+\sigma^2$$

である．したがって，$\mathrm{E}_{(X_i,y_i)_{i=1}^n\sim P^n}[\mathrm{E}_{P_X}[(\hat{f}(X)-f^\circ(X))^2]]$ の上界を得れば，汎化誤差の期待値を上から抑えることができる．これ以降，$\mathrm{E}_{P_X}[(\hat{f}(X)-f^\circ(X))^2]:=\|\hat{f}-f^\circ\|_{P_X}^2$ とする．この時，次の定理が成立する．

定理 11 (多項式オーダーの下での推定誤差)**.** ある $0<q<\infty$ が存在して，$a_i=\Omega(i^q)$ であれば，

1. $\gamma(s)=\langle a,s\rangle$ の場合:
   $v/a_1<1$ の下で，$f^\circ\in U(\mathcal{F}_{p,\theta}^{\gamma})$ $(p\geq 1,\ 1\leq\theta)$ と，ある $B_f>0$ が存在して $\|f^\circ\|_\infty\leq B_f$，が成り立つような任意の関数 f° に対して，ある定数 $B'>0$ と，

   $$L'=1,\ W'\sim(\log n)^{\frac{1}{q}},\ C\sim(\log n)^{\frac{1}{q}},\ L\sim\max\{(\log n)^{\frac{2}{q}},(\log n)^2\},\ W\sim n^{\frac{1}{2(a_1-v)+1}},$$

   $$S\sim(\log n)^{\frac{2}{q}}\max\{(\log n)^{\frac{2}{q}},(\log n)^2\}n^{\frac{1}{2(a_1-v)+1}},\ B\sim(\sqrt{2})^{(\log n)^{\frac{1}{q}}},$$

   を用いて定義される $\bar{\mathcal{P}}(B_f,L',B',W',C,L,W,S,B)$ の中での経験誤差最小化元 $\hat{f}$ について，

   $$\mathrm{E}_{P^n}[\|\hat{f}-f^\circ\|_{P_X}^2]\lesssim\left(\prod_{i=2}^{\infty}\frac{1}{1-2^{\frac{-(a_i-a_1)}{a_1}}}\right)n^{-\frac{2(a_1-v)}{2(a_1-v)+1}}(\log n)^{\frac{2}{q}+2}\max\{(\log n)^{\frac{4}{q}},(\log n)^4\}$$

   が成立する．

2. $\gamma(s)=\max_i\{a_is_i\}$ の場合:
   $v\tilde{a}<1$ であれば，$f^\circ\in U(\mathcal{F}_{p,\theta}^{\gamma})$ $(p\geq 1,\ 1\leq\theta\leq 2)$ と，ある $B_f>0$ が存在して $\|f^\circ\|_\infty\leq B_f$，が成り立つような任意の関数 f° に対して，ある正の定数 $B'>0$ と

   $$L'=1,\ W'\sim(\log n)^{\frac{1}{q}},\ C\sim(\log n)^{\frac{1}{q}},\ L\sim\max\{(\log n)^{\frac{2}{q}},(\log n)^2\},\ W\sim n^{\frac{1}{2(\frac{1}{\tilde{a}}-v)+1}},$$

   $$S\sim(\log n)^{\frac{2}{q}}\max\{(\log n)^{\frac{2}{q}},(\log n)^2\}n^{\frac{1}{2(\frac{1}{\tilde{a}}-v)+1}},\ B\sim(\sqrt{2})^{(\log n)^{\frac{1}{q}}},$$

   を用いて定義される $\bar{\mathcal{P}}(B_f,L',B',W',C,L,W,S,B)$ の中での経験誤差最小化元 $\hat{f}$ について，

   $$\mathrm{E}_{P^n}[\|\hat{f}-f^\circ\|_{P_X}^2]\lesssim n^{-\frac{2(\frac{1}{\tilde{a}}-v)}{2(\frac{1}{\tilde{a}}-v)+1}}(\log n)^{\frac{2}{q}+2}\max\{(\log n)^{\frac{4}{q}},(\log n)^4\}$$

   が成立する．

これにより，入力データの次元が無限次元であっても，特定の滑らかさを持つ関数の場合には本質的な次元がその滑らかさにより決定されることが分かる．また，滑らかさに依存した量

が収束レートのオーダーに大きな影響を持つことが分かる．この解析により，データ数に比べてはるかに大きなデータが入力されるタスクにおいても，関数の滑らかさがある条件を満たせば次元への依存を完全に回避することが分かる．これによって，インデックスが大きくなればなるほどスムーズネスが大きくなるような状況では，通常の CNN で次元非依存の多項式レートでの収束が達成可能であることが示されている．また，本定理では，$I(\gamma, T)$ が与えられていない下で，推定誤差を導出している．本証明では，CNN は必要なインデックス $I(\gamma, T)$ を観測値から学習する役目を負っている．また，[17] や [18] において導出された結果では，推定・近似誤差の解析において B-spline 基底を用いている．この時，条件として，$d \ll n$ が仮定されている．一方，本研究では三角関数を用いた関数空間の特徴づけにより，$d \gg n$ や $d = \infty$ なる状況であっても，混合・異方性を持つような関数空間の CNN による近似・推定誤差を滑らかさによって特徴づけることができて，次元の影響が全くないことを示している．

5.3 滑らかさにスパース性がある場合

次に，関数の滑らかさを表す数列 a において，スパース性がある場合を考える．前章で考えたようにインデックスが大きい成分のデータが重要ではないという仮定の下では，通常の CNN により滑らかさに依存する多項式レートが達成可能であった．ここでは，列 a にスパース性がある場合であっても，拡張 CNN を用いれば同様のレートが達成可能であることを示す．まずは，滑らかさのスパース性を定義する．

定義 12 (スパースな滑らかさ)**.** $a = (a_1, \ldots, a_i, \ldots,)$ が与えられたとき，a の項を昇順に並び変えた列 $0 < a_{i_1} < a_{i_2} < \cdots$ を考えたとき，ある $0 < q < \infty$ を用いて定義される

$$\|a\|_{wl^q} := \sup_j j^q a_{i_j}^{-1}$$

が有界であるとき，a はスパース性を持つという．ここで，$\|a\|_{wl^q}$ が小さいということは，直感的には，多くの a_i が非常に大きく，近似の際に重要となるインデックスが少ないことを示している．

$\|\cdot\|_{wl^q}$ は，a の要素を昇順に並び変えたときの増大度合いを表している．このオーダーは，a_i のオーダーによらない．

定理 13 (滑らかさにスパース性がある場合の近似定理)**.** ある $0 < q < \infty$ に対して，$\|a\|_{wl^q} < 1$，$a_i = \Omega(\log i)$ の仮定の下で，

1. $\gamma(s) = \langle a, s \rangle$ の場合:

$$L' \sim T,\ W' = 3,\ C \sim T^{\frac{1}{q}},\ L \sim \max\left\{T^{\frac{2}{q}}, T^2\right\},\ W \sim \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{\frac{-(a_i - a_{i_1})}{a_{i_1}}}},\right) T^{\frac{1}{q}} 2^{\frac{T}{a_1}},$$

$$S \sim \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{\frac{-(a_i - a_{i_1})}{a_{i_1}}}}\right) T^{\frac{2}{q}} \max\left\{T^{\frac{2}{q}}, T^2\right\} 2^{\frac{T}{a_1}},\ B \sim (\sqrt{2})^{T^{\frac{1}{q}}},$$

とすれば，$f \in U(\mathcal{F}^{\gamma}_{p,\theta})$ $(p \geq 1,\ 1 \leq \theta)$ と，ある $B_f > 0$ が存在して $\|f\|_\infty \leq B_f$，が成り立つような任意の関数 f に対して，$f' \in \mathcal{P}(L', B', W', C, L, W, B, S)$ が存在して，

$$\|f' - f\|_2 \lesssim 2^{-(1 - \frac{1}{a_{i_1}})T}$$

が成立する．

2. $\gamma(s) = \max_i\{a_i s_i\}_i$ の場合:

$$L' \sim T,\ W' = 3,\ C \sim T^{\frac{1}{q}},\ L \sim \max\left\{T^{\frac{2}{q}}, T^2\right\},\ W \sim T^{\frac{1}{q}} 2^{\tilde{a}T},$$
$$S \sim T^{\frac{2}{q}} \max\left\{T^{\frac{2}{q}}, T^2\right\} 2^{\tilde{a}T},\ B \sim (\sqrt{2})^{T^{\frac{1}{q}}},$$

とすれば，とすれば，$q > 1$ の時，$f \in U(\mathcal{F}^{\gamma}_{p,\theta})$ $(p \geq 1,\ 1 \leq \theta \leq 2)$ と，ある $B_f > 0$ が存在して $\|f\|_\infty \leq B_f$，が成り立つような任意の関数 f に対して，$f' \in \mathcal{P}(L', B', W', C, L, W, B, S)$ が存在して，

$$\|f' - f\|_2 \lesssim 2^{-(1 - v\tilde{a})T}$$

が成立する．

定理 14 (滑らかさにスパース性がある場合の推定誤差)**.** ある $0 < q < \infty$ について，$a_i = \Omega(i^q)$ であれば，

1. $\gamma(s) = \langle a, s \rangle$ の場合:
$v/a_{i_1} < 1$ の下で，$f^\circ \in U(\mathcal{F}^{\gamma}_{p,\theta})$ $(p \geq 1,\ 1 \leq \theta)$ と，ある $B_f > 0$ が存在して $\|f^\circ\|_\infty \leq B_f$，が成り立つような任意の関数 f° に対して，ある定数 $B' > 0$ と，

$$L' \sim T,\ W' = 3,\ C \sim (\log n)^{\frac{1}{q}},\ L \sim \max\{(\log n)^{\frac{2}{q}}, (\log n)^2\},\ W \sim n^{\frac{1}{2(a_{i_1} - v) + 1}},$$
$$S \sim (\log n)^{\frac{2}{q}} \max\{(\log n)^{\frac{2}{q}}, (\log n)^2\} n^{\frac{1}{2(a_{i_1} - v) + 1}},\ B \sim (\sqrt{2})^{(\log n)^{\frac{1}{q}}},$$

を用いて定義される $\bar{\mathcal{P}}(B_f, L', B', W', C, L, W, S, B)$ の中での経験誤差最小化元 $\hat{f}$ について，

$$\mathrm{E}_{P^n}[\|\hat{f} - f^\circ\|^2_{P_X}] \lesssim \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{\frac{-(a_i - a_{i_1})}{a_{i_1}}}}\right) n^{-\frac{2(a_{i_1} - v)}{2(a_{i_1} - v) + 1}} (\log n)^{\frac{2}{q} + 2} \max\{(\log n)^{\frac{4}{q}}, (\log n)^4\}$$

が成立する．

2. $\gamma(s) = \max_i\{a_i s_i\}$ の場合:
$v\tilde{a} < 1$ であれば，$f^\circ \in U(\mathcal{F}^{\gamma}_{p,\theta})$ $(p \geq 1,\ 1 \leq \theta \leq 2)$ と，ある $B_f > 0$ が存在して $\|f^\circ\|_\infty \leq B_f$，が成り立つような任意の関数 f° に対して，ある正の定数 $B' > 0$ と

$$L' \sim T,\ W' = 3,\ C \sim (\log n)^{\frac{1}{q}},\ L \sim \max\{(\log n)^{\frac{2}{q}}, (\log n)^2\},\ W \sim n^{\frac{1}{2(\frac{1}{a}-v)+1}},$$
$$S \sim (\log n)^{\frac{2}{q}} \max\{(\log n)^{\frac{2}{q}}, (\log n)^2\} n^{\frac{1}{2(\frac{1}{a}-v)+1}},\ B \sim (\sqrt{2})^{(\log n)^{\frac{1}{q}}},$$

を用いて定義される $\bar{\mathcal{P}}(B_f, L', B', W', C, L, W, S, B)$ の中での経験誤差最小化元 $\hat{f}$ について，

$$\mathrm{E}_{P^n}[\|\hat{f} - f^\circ\|^2_{P_X}] \lesssim n^{-\frac{2(\frac{1}{a}-v)}{2(\frac{1}{a}-v)+1}} (\log n)^{\frac{2}{q}+2} \max\{(\log n)^{\frac{4}{q}}, (\log n)^4\}$$

が成立する．

これらの定理より，軸方向に多項式オーダーで上昇する滑らかさの場合には，通常の CNN で基底選択が行えている一方，関数の滑らかさにスパース性がある場合には，拡張 CNN が基底選択において重要な役割をしていることが分かる．この定理により，長期依存性のあるデータを抽出する際に，拡張 CNN を用いれば次元への依存を回避した学習レートが達成可能であることが示されている．

5.4 定理の証明

5.4.1 定理 10 の証明

a を各要素が正の狭義単調増加列とすれば，次の補題が成立する．

補題 15. $a' = (a'_i)_{i=1}^{\infty}$ について，$1 = a_1 = a'_1$，$a'_1 < a'_2 < a'_3 < \ldots$ であり，ある $\beta > 0$ について，

$$\prod_{i=2}^{\infty} \frac{1}{1 - 2^{-\beta(a_i - a'_i)}} < \infty$$

であると仮定すると，

$$\sum_{s \in \mathbb{N}_0^\infty : \langle a', s\rangle \geq T} 2^{-\beta\langle a, s\rangle} \leq 2(1 - 2^{-\beta})^{-1} \left(\prod_{i=2}^{\infty} \frac{1}{1 - 2^{-\beta(a_i - a'_i)}} \right) 2^{-\beta T}$$

が成立する．また，

$$\sum_{s \in \mathbb{N}_0^\infty : \langle a, s\rangle \leq T} 2^{s} \leq 16 \left(\prod_{i=2}^{\infty} \frac{1}{1 - 2^{-(a_i - 1)}} \right) 2^{T}$$

が成立する．

Proof.

$$\begin{aligned}
&\sum_{s\in\mathbb{N}_0^\infty:\langle a',s\rangle\geq T} 2^{-\beta\langle a,s\rangle} \\
&= \left(\sum_{s_1=0}^{\infty} 2^{-\beta s_1}\right) \times \left(\sum_{(s_i)_{i=2}^\infty\in\mathbb{N}_0^\infty:\sum_{i=2}^\infty a_i' s_i\geq T} 2^{-\beta\sum_{i=2}^\infty a_i s_i}\right) \\
&+ \sum_{(s_i)_{i=2}^\infty\in\mathbb{N}_0^\infty:\sum_{i=2}^\infty a_i' s_i< T} 2^{-\beta\sum_{i=2}^\infty a_i s_i} \left(\sum_{s_1\in\mathbb{N}\cup\{0\}:s_1\geq T-\sum_{i=2}^\infty a_i' s_i} 2^{-\beta s_1}\right)
\end{aligned}$$

である．$T\leq\sum_{i=2}^\infty a_i' s_i$ において，

$$\sum_{s_1\in\mathbb{N}\cup\{0\}:s_1\geq T-\sum_{i=2}^\infty a_i' s_i} 2^{-\beta s_1} = \frac{2^{-\beta T}2^{\beta\sum_{i=2}^\infty a_i' s_i}}{1-2^{-\beta}}$$

$$\sum_{s_1=0}^{\infty} 2^{-\beta s_1} = (1-2^{-\beta})^{-1}$$

であり，また

$$\begin{aligned}
&\sum_{(s_i)_{i=2}^\infty\in\mathbb{N}_0^\infty:\sum_{i=2}^\infty a_i' s_i\geq T} 2^{-\beta\sum_{i=2}^\infty a_i s_i} \\
&\leq \sum_{(s_i)_{i=2}^\infty\in\mathbb{N}_0^\infty:\sum_{i=2}^\infty a_i' s_i\geq T} 2^{-\beta\sum_{i=2}^\infty (a_i-a_i') s_i} 2^{-\beta\sum_{i=2}^\infty a_i' s_i} \\
&\leq 2^{-\beta T} \sum_{(s_i)_{i=2}^\infty\in\mathbb{N}_0^\infty:\sum_{i=2}^\infty a_i' s_i\geq T} 2^{-\beta\sum_{i=2}^\infty (a_i-a_i') s_i}
\end{aligned}$$

が成立する．また，

$$\begin{aligned}
&\sum_{s\in\mathbb{N}_0^\infty} 2^{-\beta\sum_{i=2}^\infty (a_i-a_i') s_i} \\
&= \prod_{i=2}^{\infty}\left(\sum_{s_i=0}^{\infty} 2^{-\beta(a_i-a_i')s_i}\right)
\end{aligned}$$

が成立する．これらを用いれば，

$$\begin{aligned}
\sum_{s\in\mathbb{N}_0^\infty:\langle a',s\rangle\geq T} 2^{-\beta\langle a,s\rangle} &\leq 2(1-2^{-\beta})^{-1}2^{-\beta T}\prod_{i=2}^{\infty}\frac{1}{1-2^{-\beta(a_i-a_i')}} \\
&\leq 2(1-2^{-\beta})^{-1}\left(\prod_{i=2}^{\infty}\frac{1}{1-2^{-\beta(a_i-a_i')}}\right)2^{-\beta T}
\end{aligned}$$

で評価できる．また，

$$I_T := \sum_{s\in\mathbb{N}_0^\infty:T-1<\langle a,s\rangle\leq T} 2^{\sum_{i=1}^\infty s_i}$$

について，

$$\begin{aligned} I_T &\leq 2^{2T} \sum_{s\in\mathbb{N}_0^\infty : T-1\langle a,s\rangle \leq T} 2^{\sum_{i=1}^\infty s_i - 2\langle a,s\rangle} \\ &\leq 2^{2T} \sum_{s\in\mathbb{N}_0^\infty : T-1\langle a,s\rangle \leq T} 2^{-\sum_{i=1}^\infty (2a_i-1)s_i} \end{aligned}$$

が成立する．ここで，$a_1 = 2a_1 - 1 = 1,\ a_i < 2a_i - 1\ (i > 1)$ であるので，

$$\begin{aligned} \sum_{s\in\mathbb{N}_0^\infty : T-1\langle a,s\rangle \leq T} 2^{-\sum_{i=1}^\infty (2a_i-1)s_i} &\leq \sum_{s\in\mathbb{N}_0^\infty : T-1\langle a,s\rangle \leq T} 2^{-\langle a,s\rangle} \\ &\leq 8\left(\prod_{i=2}^\infty \frac{1}{1-2^{-(a_i-1)}}\right) 2^{-T} \end{aligned}$$

である．よって，

$$2^{2T} \sum_{s\in\mathbb{N}_0^\infty : T-1\langle a,s\rangle \leq T} 2^{-\sum_{i=1}^\infty (2a_i-1)s_i} \leq 8\left(\prod_{i=2}^\infty \frac{1}{1-2^{-(a_i-1)}}\right) 2^{T}$$

が成立する．したがって，

$$I_T \leq 8\left(\prod_{i=2}^\infty \frac{1}{1-2^{-(a_i-1)}}\right) 2^{T}$$

なので，

$$\begin{aligned} \sum_{s\in\mathbb{N}_0^\infty : \langle a,s\rangle \leq T} 2^{s} &\leq 8\left(\prod_{i=2}^\infty \frac{1}{1-2^{-(a_i-1)}}\right) \sum_{t=0}^{T} 2^{t} \\ &\leq 16\left(\prod_{i=2}^\infty \frac{1}{1-2^{-(a_i-1)}}\right) 2^{T} \end{aligned}$$

が成り立つ． □

定理 *10* の証明*.*

まずは，$\gamma(s) = \langle a, s\rangle$ の場合を示す．$1 \leq \theta \leq 2$ の場合，補題 15 を用いれば，

$$S(\langle a,s\rangle, T) \leq 16\left(\prod_{i=2}^\infty \frac{1}{1-2^{\frac{-(a_i-a_1)}{a_1}}}\right) 2^{\frac{T}{a_1}}$$

が導かれる．また，

$$\alpha = \sup_{s\in\mathbb{N}_0^\infty} \frac{\sum_{i=1}^\infty s_i}{\langle a,s\rangle} = \frac{1}{a_1}$$

が成立する．$a_i = \Omega(i^q)$ より，$d_{\max} \sim T^{1/q},\ f_{\max} \sim T$ である．ここで，フィルター

$$w_i = \begin{cases} (w_{ij} : w_{ii} = 1,\ \forall i \neq j,\ w_{ij} = 0)_{j=1}^\infty\ (i \leq d_{\max}) \\ (0, \dots)\ (i > d_{\max}) \end{cases}$$

を用いれば，

$$(\mathrm{Conv}_{1,w}(X))_1 = \begin{pmatrix} x_1 \\ \vdots \\ x_{d_{\max}} \end{pmatrix}$$

とできる．また，定理 9 より，ある定数 $K, K' > 0$ が存在して，

$$\begin{aligned} L &= 2K \max\left\{T^{\frac{2}{q}}, T^2\right\}, \\ W &= 21 \left(\prod_{i=2}^{\infty} \frac{1}{1 - 2^{\frac{-(a_i - a_1)}{a_1}}}\right) T^{\frac{1}{q}} 2^{\frac{T}{a_1}}, \\ S &= 1764K \left(\prod_{i=2}^{\infty} \frac{1}{1 - 2^{\frac{-(a_i - a_1)}{a_1}}}\right) T^{\frac{2}{q}} \max\max\left\{T^{\frac{2}{q}}, T^2\right\} 2^{\frac{T}{a_1}}, \\ B &= (\sqrt{2})^{d_{\max}} K', \end{aligned}$$

とすれば，任意の $f \in U(\mathcal{F}_{p,\theta}^{\langle a,s\rangle})$ に対して，あるニューラルネットワーク $\hat{R}_T \in \Phi(L, W, S, B)$ が存在して，

$$f'(X) := \hat{R}_T(x_1, \ldots, x_{d_{\max}})$$

とすれば，

$$\|f' - f\|_2 \leq 2^{-(1 - \frac{v}{a_1})T}$$

が成り立つ．ここで，a は狭義単調増加列であるから，$d_{\max}$ 番目以降の要素が非 0 であれば，$\langle a, s\rangle \geq T$ となる事実を用いた．よって，$f'(X) = \left(\hat{R}_T \circ \mathrm{Conv}_{1,w}(X)\right)_1$ となるので，

$$\left\| \left(\hat{R}_T \circ \mathrm{Conv}_{1,w}(X)\right)_1 - f \right\|_2 \leq 2^{-(1 - \frac{v}{a_1})T}$$

となり，題意が示される．

また，$2 < \theta$ の場合，$a'_1 = \frac{a_1}{2}$ とし，ここで $\beta = \inf_{i \in \mathbb{N}}(a_{i+1} - a_i)$ として，$2 < k < 2 + \frac{2\beta}{a_1}$ を満たすようなある k を用いて，$a'_i = \frac{a_i}{k}$ と定義する．この時，$a'_1 < a'_2 < \ldots$ を満たす．また，仮定より任意の $c > 0$ に対して，

$$\begin{aligned} &\prod_{i=2}^{\infty} \frac{1}{1 - 2^{-c\left(\frac{a_i}{a_1} - \frac{2a'_i}{a_1}\right)}} \\ &= \prod_{i=2}^{\infty} \frac{1}{1 - 2^{-c\left(\frac{1}{a_1} - \frac{2}{ka_1}\right)a_i}} \\ &< \infty \end{aligned}$$

であるので，

$$\begin{aligned} &\sum_{s \in \mathbb{N}_0^\infty : \langle a', s\rangle > T} 2^{\frac{2\theta}{\theta - 2}\langle a' - a, s\rangle} \\ &\leq \sum_{s \in \mathbb{N}_0^\infty : \langle 2a'/a_1, s\rangle > 2T/a_1} 2^{-\frac{a_1\theta}{\theta - 2}\langle a/a_1, s\rangle} \end{aligned}$$

この式の右辺に対して，補題 15 が適用可能であり，

$$\sum_{s\in\mathbb{N}_0^\infty:\langle 2a'/a_1,s\rangle>2T/a_1} 2^{-\frac{a_1\theta}{\theta-2}\langle a/a_1,s\rangle} \lesssim 2^{-\frac{2\theta}{\theta-2}T}$$

と評価できる．ここで，$\gamma'(s)=\langle a',s\rangle$ として，定理 9 の結果を用いて，$1\le\theta\le 2$ の場合と同様の議論により，CNN の集合 $\mathcal{P}$ を定義すれば，$\forall f\in U(\mathcal{F}_{p,\theta}^{\gamma})$ に対して，ある CNN，$f'\in\mathcal{P}$ が存在して，

$$\|f'-f\|_2 \lesssim 2^{-2(1-\frac{v}{a_1})T}$$

が成立する．さらに，$\alpha'=\frac{2}{a_1}$ であり，$S(\langle a',s\rangle,T)\sim 2^{\frac{2T}{a_1}}$ なので，これ以降は，$1\le\theta\le 2$ の場合と同様の手順で，結果を得る．

次に，$\gamma(s)=\max_i\{a_is_i\}_i$，$1\le\theta\le 2$ の場合を考える．まず，

$$\alpha=\max_{s\in\mathbb{N}_0^\infty}\frac{\sum_{i=1}^\infty s_i}{\max_i\{a_is_i\}} \tag{5.1}$$

について考える．

$$\begin{aligned}\max_{s\in\mathbb{N}_0^\infty}\frac{\sum_{i=1}^\infty s_i}{\max_i\{a_is_i\}} &\le \max_{s\in\mathbb{R}_{>0}^\infty}\frac{\sum_{i=1}^\infty s_i}{\max_i\{a_is_i\}}\\ &= \max_{T>0}\max_{\{s\in\mathbb{R}_{>0}:\max_i\{a_is_i\}=T\}}\frac{\sum_{i=1}^\infty s_i}{T}\end{aligned}$$

である．ここで，$\max_i\{a_is_i\}\le T$ は任意の i について，$s_i\le\frac{T}{a_i}$ が成り立つことと同値であるから，(5.1) は

$$\alpha\le\sum_{i=1}^\infty\frac{1}{a_i}$$

と評価できる．よって，$1\le\theta\le 2$ の場合，定理 9 より $\alpha=\sum_{i=1}^\infty\frac{1}{a_i}$ を用いて，

$$\|R_T(f)-f\|_2\le 2^{-(1-\delta\alpha)T}\|f\|_{\mathcal{F}_{p,\theta}^{\max_i\{a_is_i\}}}$$

が成立する．さらに，

$$\begin{aligned}\sum_{s\in\mathbb{N}_0^\infty:\gamma(s)\le T}2^s &\le \prod_{i=1}^\infty\left(\sum_{s_i=0}^{\lceil\frac{T}{a_i}\rceil}2^{s_i}\right)\\ &\le 2^{\sum_{i=1}^\infty\lceil\frac{T}{a_i}\rceil}\end{aligned}$$

であるから，$S(\gamma,T)\sim 2^{\sum_{i=1}^\infty\lceil\frac{T}{a_i}\rceil}$．また多項式オーダーでスムーズネスが落ちることを考慮すると $\gamma(s)=\langle a,s\rangle$ の場合と同様の手順を踏むことにより結果を得る．これにより，定理 10 が証明された． □

5.4.2　定理 11 の証明

証明において，次の概念を用いる：

定義 16 (カバリングナンバー)**.** あるノルム空間 $\mathcal{F}$ のノルム $\|\cdot\|$ に関する ϵ—カバリングナンバーは，

$$\mathcal{N}(\mathcal{F}, \delta, \|\cdot\|) := \inf\{n \in \mathbb{N} : \exists (f_1, \dots f_n) \in \mathcal{F}^n,\ \forall f \in \mathcal{F},\ \exists i \in [n],\ \|f_i - f\| \leq \epsilon\}$$

で定義される．

カバリングナンバーは，多くの既存研究で関数クラスの複雑さを図る量として用いられている．また，拡張 CNN のカバリングナンバーについて，次の補題が成立する：

補題 17. 拡張 CNN，$\mathcal{P}(L', B', W', C, L, W, B, S)$ について，

$$\log \mathcal{N}(\mathcal{P}, \delta, \|\cdot\|_\infty) \lesssim (S + W'C)(L + L') \log\left(\frac{LL'B'BCW'W}{\delta}\right)$$

が成立する．

Proof. $w \in \mathbb{R}^{C \times W'}$ を用いた任意の拡張幅 $h \in \mathbb{N}$ の畳み込みについて，

$$\begin{aligned}
&\|w \star_h X - w' \star_h X\|_\infty \\
&\leq \|(w - w') \star_h X\|_\infty \\
&\leq W'C\|w - w'\|_\infty \|X\|_\infty, \\
&\|w \star_h X - w \star_h X'\|_\infty \\
&\leq W'C\|w\|_\infty \|X - X'\|_\infty,
\end{aligned}$$

が成立する．したがって，$F,\ F' \in \mathbb{R}^{C \times C \times W'}$ に関して，

$$\begin{aligned}
&\|\mathrm{Conv}_{h,F} \circ X - \mathrm{Conv}_{h,F'} \circ X\|_\infty \\
&\leq \max_{i=1}^{C} \|(F_{i,:,:} \star_h X) - (F'_{i,:,:} \star_h X)\|_\infty \\
&\leq W'C \left(\max_{i=1}^{C} \|F_{i,:,:} - F'_{i,:,:}\|_\infty\right) \|X\|_\infty \\
&\leq W'C\|F - F'\|_\infty \|X\|_\infty, \qquad (5.2) \\
&\|\mathrm{Conv}_{h,F} \circ X - \mathrm{Conv}_{h,F} \circ X'\|_\infty \\
&\leq \max_{i=1}^{C} \|(F_{i,:,:} \star_h X) - (F_{i,:,:} \star_h X')\| \\
&\leq W'C \left(\max_{i=1}^{C} \|F_{i,:,:}\|_\infty\right) \|X - X'\|_\infty \\
&\leq W'C\|F\|_\infty \|X - X'\|_\infty, \qquad (5.3)
\end{aligned}$$

が成立する．ここで，

$$\begin{aligned}
f(X) &= \mathrm{Conv}_{W'^{L'}, F_{L'}} \circ \cdots \circ \mathrm{Conv}_{W'^{l}, F_l} \circ \cdots \circ \mathrm{Conv}_{1, F_1} \circ X, \\
g(X) &= \mathrm{Conv}_{W'^{L'}, F'_{L'}} \circ \cdots \circ \mathrm{Conv}_{W'^{l}, F'_l} \circ \cdots \circ \mathrm{Conv}_{1, F'_1} \circ X,
\end{aligned}$$

と置き，

$$\mathcal{A}_l(f) = \mathrm{Conv}_{W'^{l-1},F_{l-1}} \circ \cdots \circ \mathrm{Conv}_{1,F_1} \circ X,$$
$$\mathcal{B}_l(g) = \mathrm{Conv}_{W'^{L'},F'_{L'}} \circ \cdots \circ \mathrm{Conv}_{W'^{l+1},F'_{l+1}},$$

で定義する．ただし，$\mathcal{A}_1(f) = \mathcal{B}_{L'}(f) = X$ とする．これらを用いると，

$$\begin{aligned}&|f(X) - g(X)| \\ &\leq \sum_{l=1}^{L'} \|\mathcal{B}_l(g) \circ \mathrm{Conv}_{W'^{l},F_l} \circ \mathcal{A}_l(f) - \mathcal{B}_l(g) \circ \mathrm{Conv}_{W'^{l},F'_l} \circ \mathcal{A}_l(f)\|_\infty\end{aligned}$$

である．(5.3), (5.2) を用いれば，

$$\begin{aligned}|f(X) - g(X)| &\leq L'(W'C)^{L'}\|X\|_\infty \left(\prod_{i=1}^{L'} \|F_i\|_\infty\right) \max_{l=1,\ldots L'} \|F_l - F'_l\|_\infty \\ &\leq L'(W'CB')^{L'} \max_{l=1,\ldots L'} \|F_l - F'_l\|_\infty\end{aligned}$$

が成立する．ここで，$\mathrm{FNN} \in \Phi(L, W, B, S)$ について，

$$|\mathrm{FNN}(x) - \mathrm{FNN}(x')| \leq (BW)^L \|x - x'\|_\infty$$

が成り立つので，

$$\|\mathrm{FNN} \circ f(X) - \mathrm{FNN} \circ g(X)\|_\infty \leq L'(BW)^L (TCB')^{L'} \max_{l=1,\ldots L'} \|F_l - F'_l\|_\infty$$

である．一方で，$\mathrm{FNN}, \mathrm{FNN}' \in \Phi(L, W, B, S)$ にを，

$$\begin{aligned}\mathrm{FNN}(x) &= (A_L\eta(\cdot) + b_L) \circ \cdots \circ (A_l\eta(\cdot) + b_l) \circ \cdots \circ \ldots (A_1 x + b_1), \\ \mathrm{FNN}'(x) &= (A'_L\eta(\cdot) + b_L) \circ \cdots \circ (A'_l\eta(\cdot) + b'_l) \circ \cdots \circ \ldots (A'_1 x + b'_1),\end{aligned}$$

とおく．$\|g(X)\|_\infty \leq (W'C)^{L'}$ なので，[17] の補題 3 と同様の議論により，

$$\max_{l=1,\ldots,L'} \max\{\|A_l - A'_l\|_\infty, \|b_l - b'_l\|_\infty\} \leq \delta \tag{5.4}$$

の下で，

$$\|\mathrm{FNN} \circ g(X) - \mathrm{FNN}' \circ g(X)\| \leq \delta L(W'C)^{L'}\{(B+1)(W+1)\}^L$$

が成立する．また，三角不等式を用いれば，

$$\|\mathrm{FNN}\circ f(X) - \mathrm{FNN}'\circ g(X)\|_\infty \leq \|\mathrm{FNN}\circ f(X) - \mathrm{FNN}\circ g(X)\|_\infty + \|\mathrm{FNN}\circ g(X) - \mathrm{FNN}'\circ g(X)\|_\infty$$

であるから，式 (5.4) と $\max_{l=1,\ldots L'} \|F_l - F'_l\|_\infty \leq \delta$ の下で，

$$\begin{aligned}&\|\mathrm{FNN} \circ f(X) - \mathrm{FNN}' \circ g(X)\|_\infty \\ &\leq \delta L'(BW)^L (W'CB')^{L'} + \delta L(W'C)^{L'}\{(B+1)(W+1)\}^L\end{aligned} \tag{5.5}$$

が成立する．ここで，[17] の補題 3 と同様の議論により，$\Phi(L, W, B, S)$ に属する関数の非ゼロパラメータのパターン数は，$(W+1)^{LS}$ で抑えられる．したがって，(5.5) により，

$$\log \mathcal{N}(\mathcal{P}, \delta, \|\cdot\|_\infty) \lesssim (S + W'C)(L + L') \log\left(\frac{LL'B'BCW'W}{\delta}\right)$$

が成立する．　□

[8]，[17] において，次の結果が証明されている：

補題 18. $F > 0$ が存在して $\|f^\circ\|_\infty < F$ かつ，すべての ERM 推定量 $\hat{f} \in \mathcal{F}$ において，$\|\hat{f}\|_\infty < F$ が成り立つとき，すべての $\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty) \geq 3$ が成り立つような $0 < \delta < 1$ に対して，

$$\mathrm{E}_{P^n}[\|f - f^\circ\|_{P_X}^2] \lesssim \inf_{f \in \mathcal{F}} \|f - f^\circ\|_{P_X}^2 + \frac{\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty)}{n} + \delta F^2$$

が成り立つ．

また，$\frac{\mathrm{d}P_X}{\mathrm{d}\lambda^\infty} < \infty$ の条件の下で，右辺第一項目の $\|\cdot\|_{P_X}$ は，$\|\cdot\|_2$ で置き換えることができる．これらの補題を用いて，定理 11 を証明する．

定理 *11* の証明*.*

定理 10 より，混合滑らかさを持つ場合，ある定数 $B' > 0$ が存在して，

$$\begin{aligned}
L' &= 1, \\
W' &\sim T^{\frac{1}{q}}, \\
C &\sim T^{\frac{1}{q}}, \\
L &\sim \max\left\{T^{\frac{2}{q}} T^2\right\}, \\
W &\sim \left(\prod_{i=2}^{\infty} \frac{1}{1 - 2^{\frac{-(a_i - a_1)}{a_1}}}\right) T^{\frac{1}{q}} 2^{\frac{T}{a_1}}, \\
S &\sim \left(\prod_{i=2}^{\infty} \frac{1}{1 - 2^{\frac{-(a_i - a_1)}{a_1}}}\right) T^{\frac{2}{q}} \max\left\{T^{\frac{2}{q}}, T^2\right\} 2^{\frac{T}{a_1}}, \\
B &\sim (\sqrt{2})^{T^{\frac{1}{q}}},
\end{aligned}$$

とすれば，$\forall f^\circ \in U(\mathcal{F}_{p,\theta}^\gamma)$ に対して，ある $f \in \mathcal{P}(L', B', W', C, L, W, B, S)$ を用いれば，

$$\|f - f^\circ\|_2^2 \leq 2^{-2(1 - v/a_1)T}$$

である．また，補題 17 より，

$$\log\left(\mathcal{N}(\mathcal{P}, \delta, \|\cdot\|_\infty)\right) \lesssim \left(\prod_{i=2}^{\infty} \frac{1}{1 - 2^{\frac{-(a_i - a_1)}{a_1}}}\right) 2^{\frac{T}{a_1}} T^{\frac{2}{q}+1} \max\left\{T^{\frac{4}{q}}, T^4\right\} \log\left(\frac{T}{\delta}\right)$$

である．また，仮定により，$\|f^\circ\|_\infty \leq B_f$ であるから，ここで，$\hat{f}_c \in \bar{\mathcal{P}}(B_f, L', B', W', C, L, W, B, S)$ を $\bar{\mathcal{P}}$ の中の ERM 推定量として，補題 18 を用いれば，

$$\mathrm{E}_{P^n}[\|\hat{f}_c - f^\circ\|_{P_X}^2] \lesssim 2^{-2(1-\frac{v}{a_1})T} + \frac{\left(\prod_{i=2}^{\infty} \frac{1}{1-2^{\frac{-(a_i-a_1)}{a_1}}}\right) 2^{\frac{T}{a_1}} T^{\frac{2}{q}+1} \max\left\{T^{\frac{4}{q}}, T^4\right\} \log(\frac{T}{\delta})}{n} + \delta B_f^2$$

が成立する．ここで，$N = 2^{\frac{T}{a_1}}$ と置けば，

$$\begin{aligned}&\mathrm{E}_{P^n}[\|\hat{f}_c - f^\circ\|_{P_X}^2] \\ &\lesssim N^{-2(a_1-v)} + \frac{\left(\prod_{i=2}^{\infty} \frac{1}{1-2^{\frac{-(a_i-1)}{a_1}}}\right) N(\log N)^{\frac{2}{q}+1} \max\left\{(\log N)^{\frac{4}{q}}, (\log N)^4\right\} \log(\frac{\log N}{\delta})}{n} + \delta B_f^2\end{aligned}$$

となる．ここで，$N = n^{\frac{1}{2(a_1-v)+1}}, \delta = \frac{1}{n}$ とすれば，

$$\mathrm{E}_{P^n}[\|\hat{f}_c - f^\circ\|_{P_X}^2] \lesssim \left(\prod_{i=2}^{\infty} \frac{1}{1-2^{\frac{-(a_i-a_1)}{a_1}}}\right) n^{-\frac{2(a_1-v)}{2(a_1-v)+1}} (\log n)^{\frac{2}{q}+2} \max\{(\log n)^{\frac{4}{q}}, (\log n)^4\}$$

である．よって，題意が示された．
次に，$\gamma = \max_i\{a_i s_i\}_i$ の場合について，ある定数 $B' > 0$ と

$$\begin{aligned}L' &= 1, \\ W' &\sim T^{\frac{1}{q}}, \\ L &\sim \max\left\{T^{\frac{2}{q}}, T^2\right\}, \\ W &\sim T^{\frac{1}{q}} 2^{\tilde{a}T}, \\ S &\sim T^{\frac{2}{q}} \max\left\{T^{\frac{2}{q}}, T^2\right\} 2^{\tilde{a}T}, \\ B &\sim (\sqrt{2})^{T^{\frac{1}{q}}},\end{aligned}$$

とする．$1 \leq \theta \leq 2$, $f^\circ \in \mathcal{F}_{p,\theta}^{\gamma}$ に対して，ある $f \in \mathcal{P}(L', B', W', C, L, W, B, S)$ を用いれば，

$$\|f - f^\circ\|_2^2 \lesssim 2^{-2(1-\tilde{a}v)T}$$

が成り立つ．したがって，混合滑らかさの場合と同様の手順を踏めば，

$$\mathrm{E}_{P^n}[\|\hat{f}_c - f^\circ\|_{P_X}^2] \lesssim^{-\frac{2(\frac{1}{\tilde{a}}-v)}{2(\frac{1}{\tilde{a}}-v)+1}} (\log n)^{\frac{2}{q}+2} \max\{(\log n)^{\frac{4}{q}}, (\log n)^4\}$$

が成立する．したがって題意が示される． □

5.4.3 定理 14 の証明

まずは，次の補題を証明する：

補題 19. 無限次元入力 $X = (x_1, \dots x_i \dots) \in \mathbb{R}^\infty$ が入力されたとき，$i_j < T^{L'}$, $i_1 < \dots i_j < \cdots < i_N$ を満たすようなインデックスをもつ任意の部分列 $(x_{i_1}, \dots x_{i_N})$ について，

$$(x_{i_1}, \dots x_{i_N}) = \left(\mathrm{Conv}_{T^{L'-1}, W_{L'}} \circ \cdots \circ \mathrm{Conv}_{T^{l-1}, W_l} \circ \cdots \circ \mathrm{Conv}_{1, W_1} \circ X\right)_1$$

となるような L' 層，幅 T，各層のチャネル数が N，$\|W_l\|_\infty \leq 1$ $(l = 1, \dots, L')$ の拡張畳み込みが存在する．

Proof.

$$\mathcal{A}_l(X) = \mathrm{Conv}_{T^{l-1}, W_l} \circ \cdots \circ \mathrm{Conv}_{1, W_1} \circ X$$

とする．ここで，

$$\mathcal{A'}_{l,i}(X) = \left(\mathcal{A}_l(X)_{iT^l+1}, \dots \mathcal{A}_l(X)_{iT^l+jT^{l-1}+1}, \dots, \mathcal{A}_l(X)_{iT^l+(T-1)T^{l-1}+1}\right)$$
$$(j = 0, \dots, T-1, i \in \{0\} \cup \mathbb{N})$$

を用いると，

$$\begin{aligned}
&\mathcal{A}_{l+1}(X)_{iT^l+1} \\
&= \mathrm{Conv}_{T^l, W_{l+1}} \circ \mathcal{A}_l(X) \\
&= ((W_{l+1})_{1,:,:} \star_1 \mathcal{A'}_{l,i}(X), \dots, (W_{l+1})_{i,:,:} \star_1 \mathcal{A'}_{l,i}(X), \dots, (W_{l+1})_{N,:,:} \star_1 \mathcal{A'}_{l,i}(X))
\end{aligned}$$

である．ここで，$\mathcal{A'}_{l,i}(X)$ が，$iT^l + 1 \leq i_j \leq (i+1)T^l$ を満たすような x_{i_j} をすべて含んでいると仮定する．また，仮定よりそのような要素の個数は N 以下であるから，重み W_{l+1} を適切に設定すれば，$i = 1, \dots, T^{L'-l} - 1$ において，$\mathcal{A}_{l+1}(X)_{iT^l+1}$ が $iT^l + 1 \leq i_j \leq (i+1)T^l$ を満たすような要素をすべて含むようにできる．このとき，$\mathcal{A'}_{l+1,i}(X)$ は，$iT^{l+1} \leq i_j \leq (i+1)T^{l+1}$ を満たす要素をすべて含む．これにより，$\mathcal{A}_{L'}(X)_1$ は，$1 \leq i_j \leq T^{L'}$ を満たすような x_{i_j} をすべて含むようにできることが分かる．　□

定理 *13,14* の証明.

インデックス $(i_1, \dots, i_j, \dots)$ は，無限列 a において，$a_{i_1} < a_{i_2} < \cdots$ を満たすとする．まずは，$\gamma(s) = \langle a, s \rangle$ の場合について考察する．$\|a\|_{wl^q} \leq 1$ であることから，$a_{i_j} \geq j^q$ が成り立つので，$T > 0$ について，

$$\begin{aligned}
I(\gamma, T) &= \{i : \exists s \in \mathbb{N}_0^\infty,\ s_i \neq 0,\ \langle a, s \rangle \leq T\} \\
&= \left\{i_j : 1 \leq j \leq T^{\frac{1}{q}}\right\}
\end{aligned}$$

が成立する．したがって，$d_{\max} = T^{\frac{1}{q}}$, $f_{\max} = T^{\frac{1}{q}}$ が成立する．また，$d_{\max}$ 次元の入力を受け取るニューラルネットワークを $\hat{R}_T$ とする．$a_i = \Omega(\log i)$ であることから，ある定数 $Q > 0$ が存在し，

$$a_i \geq Q \log i$$

なので，$a_i \leq T$ ならば，$i \leq \exp \frac{T}{Q}$ である．補題 19 により，$I(\gamma, T)$ に含まれるインデックスの要素のみを抽出する畳み込みを，

$$\mathrm{Conv}_{3^{\lceil \frac{T}{Q} \rceil}, W_{\lceil \frac{T}{Q} \rceil}} \circ \cdots \circ \mathrm{Conv}_{1, W_1} \circ X$$

とすれば，

$$\left(\hat{R}_T \circ \mathrm{Conv}_{3^{\lceil \frac{T}{Q} \rceil}, W_{\lceil \frac{T}{Q} \rceil}} \circ \cdots \circ \mathrm{Conv}_{1,W_1} \circ X\right)_1$$

は，$I(\gamma, T)$ に含まれるインデックスのみを取り出し，ニューラルネットワークに入力するような拡張 CNN となる．仮定より，

$$d_{\max} \leq T^{\frac{1}{q}},\ f_{\max} \leq T^{\frac{1}{q}}$$

が成立する．ここでは，$\gamma(s) = \langle a, s \rangle$ の場合について考える．補題 15 により，

$$\begin{aligned}
&\sum_{s \in \mathbb{N}_0^\infty : \langle a, s \rangle \leq T} 2^s \\
&= \sum_{s \in \mathbb{N}_0^\infty : \sum_{j=1}^\infty a_{i_j} s_{i_j} \leq T} 2^s \\
&\lesssim \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{-\frac{(a_i - a_{i_1})}{a_{i_1}}}} \right) 2^{\frac{1}{a_{i_1}}}
\end{aligned}$$

である．したがって，定理 9 により，

$$\begin{aligned}
&L \sim \max\{T^{\frac{2}{q}}, T^2\},\ W \sim 21 \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{-\frac{(a_i - a_{i_1})}{a_{i_1}}}} \right) T^{\frac{1}{q}} 2^{\frac{T}{a_{i_1}}}, \\
&S \sim \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{-\frac{(a_i - a_{i_1})}{a_{i_1}}}} \right) \max T^{\frac{2}{q}} \{T^{\frac{2}{q}}, T^2\} 2^{\frac{T}{a_{i_1}}},\ B \sim \sqrt{2}^{T^{\frac{1}{q}}},
\end{aligned}$$

とすれば，任意の $f \in U(\mathcal{F}_{p,\theta}^{\gamma})$ に対して，$\hat{R}_T \in \Phi(L, W, B, S)$ が存在して，$f'(X) = \hat{R}_T((x_i : i \in I(\gamma, T))$ とすれば，

$$\|f' - f\|_2^2 \leq 2^{-(1 - \frac{v}{a_{i_1}})T}$$

である．ここで，最初に述べた事実によりある正の定数 $B' > 0$ と

$$\begin{aligned}
&L' \sim T,\ W' = 3,\ C \sim T^{\frac{1}{q}}, \\
&L \sim \max\{T^{\frac{2}{q}}, T^2\},\ W \sim 21 \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{-\frac{(a_i - a_{i_1})}{a_{i_1}}}} \right) T^{\frac{1}{q}} 2^{\frac{T}{a_{i_1}}}, \\
&S \sim \left(\prod_{i \neq i_1} \frac{1}{1 - 2^{-\frac{(a_i - a_{i_1})}{a_{i_1}}}} \right) \max T^{\frac{2}{q}} \{T^{\frac{2}{q}}, T^2\} 2^{\frac{T}{a_{i_1}}},\ B \sim \sqrt{2}^{T^{\frac{1}{q}}},
\end{aligned}$$

を用いて，$\mathcal{P}(L', B', W', C, L, W, B, S)$ とすれば，ある拡張 CNN，$f' \in \mathcal{P}$ により，

$$\|f' - f\|_2^2 \leq 2^{-(1 - \frac{v}{a_{i_1}})T}$$

が成立する．推定誤差の議論は，定理 11 の場合と同様である．また，$\gamma(s) = \max\{a_i s_i\}_i$ の場合も同様の議論により，題意が示される．

□

第 6 章

数値実験による検証

6.1　次元非依存性の検証

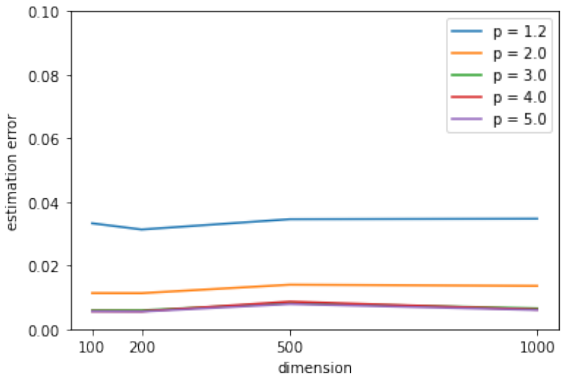

図 6.1. 多項式オーダーの場合

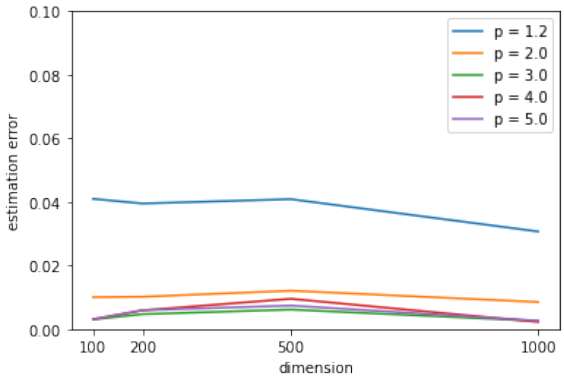

図 6.2. スパースな場合

本節では，異方周波空間に属する関数を用いて，CNN の次元非依存性を数値実験により検証する．次のような設定で実験を行った：

1. 平滑度が多項式オーダーで上昇するケース
   実験で用いた関数：
   $$f^\circ(x) = \sum_{k=1}^{d} \frac{\sqrt{2}^k}{\sum_{i=1}^{k} 2^{ip}} \prod_{i=1}^{k} \cos(2\pi x_i)$$
   データの生成モデル：$y_i = f^\circ(x_i)\ (i = 1, \ldots, n)$
   パラメータの設定：$n = 128,\ p = 1.2, 2.0, 3.0, 4.0, 5.0,\ d = 100, 200, 500, 1000$
   用いたモデル：カーネル幅 10, チャンネル数 10 の CNN
2. 平滑度にスパース性があるケース
   実験で用いた関数：
   $$f^\circ(x) = \sum_{k=1}^{[d/10]} \frac{\sqrt{2}^k}{\sum_{i=1}^{k} 2^{ip}} \prod_{i=1}^{k} \cos(2\pi x_{10i})$$

データの生成モデル：$y_i = f^\circ(x_i)$ $(i = 1, \ldots, n)$
パラメータの設定：$n = 128,\ p = 1.2, 2.0, 3.0, 4.0, 5.0,\ d = 100, 200, 500, 1000$
用いたモデル：層数 3, カーネル幅 3，拡張幅 3，チャンネル数 10 の拡張 CNN

それぞれの実験において，得られたデータ $(x_i, y_i)_{i=1}^n$ を用いて，CNN および拡張 CNN をにより f° 推定した．その結果が，図 6.1 および図 6.2 である．これらの結果より，p が一定の下では，データの次元が増えることによる推定誤差への悪影響がほとんど見られないことが分かる．これらの結果は，本研究で導出したように，理論的な上界が次元非依存であり，滑らかさに依存することと一致するが分かる．また，スパース性がある場合でも拡張 CNN を用いることにより，次元非依存に推定ができていることが分かる．また，滑らかさとデータ数により推定誤差が決定されているため，100 次元のデータを受け取る場合であっても，1000 次元の場合より推定が難しくなることがあることが見て取れる．

第 7 章

結論

本研究では，超高次元データ $(n \ll d)$ や無限次元データ $(d = \infty)$ が学習データとして得られた場合に，次元の呪いを回避した学習が可能な状況の一つとして，関数空間の平滑さに関する条件を与えた．本研究により，次元がサンプル数 n に比べて大きなデータであっても，関数の平滑さが学習レートの決定において本質的な役割を果たしていることを示した．今後の方針としては，例えば滑らかさがデータに依存する場合についての考察をすることや，$\mathcal{F}^{\gamma}_{p,\theta}$ が他の関数クラスとどのような関係があるのか，これまでに提案されている概念との関係性について考察すること等が考えられる．

謝辞

参考文献

[1] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

[2] D. Dung and M. Griebel. Hyperbolic cross approximation in infinite dimensions. *Journal of Complexity*, 33:55–88, 2016.

[3] D. Dung, V. Temlyakov, and T. Ullrich. *Hyperbolic Cross Approximation*. Springer International Publishing, 2018.

[4] G. Evarist and N. Richard. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

[5] C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[6] F. Ferraty, I. Van Keilegom, and P. Vieu. Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109:10–28, 2012.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] S. Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

[9] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[10] C. Minshuo, J. Haoming, L. Wenjing, and Z. Tuo. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In *Advances in Neural Information Processing Systems*, volume 32, pages 8174–8184, 2019.

[11] R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

[12] R. Nessel and G. Wilmes. Nikolskii-type inequalities for trigonometric polynomials and entire functions of exponential type. *Journal of the Australian Mathematical Society*, 25(1):7–18, 1978.

[13] J. Oliva, B. Poczos, and J. Schneider. Distribution to distribution regression. In S. Dasgupta and D. McAllester, editors, *Proceedings of the International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1049–1057, 2013.

[14] J. Oliva, W. Neiswanger, B. Póczos, E. Xing, H. Trac, S. Ho, and J. Schneider. Fast function to function regression. *ArXiv*, abs/1410.7414, 2015.

[15] D. Perekrestenko, P. Grohs, D. Elbrächter, and H. Bölcskei. The universal approximation power of finite-width deep ReLU networks. *CoRR*, abs/1806.01528, 2018.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[17] T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

[18] Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. *arXiv preprint arXiv:1910.12799*, 2019.

[19] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

[20] S. Yanchenko. Best approximation of the functions from anisotropic nikolskii besov classes. *Ukrainian Mathematical Journal*, 70(4):661–670, 2018.

[21] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

[22] K. Yoon. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014.

[23] I. Yuri and S. Natalia. Estimation and detection of functions from anisotropic sobolev classes. *Electronic Journal of Statistics*, 5:484–506, 2011.