深層学習・人工知能の数理

2024/03/22

名古屋大学

今泉允聡

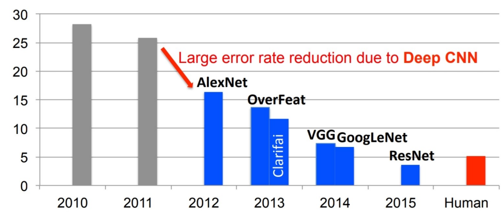
(東京大学/理化学研究所)

導入

深層学習・人工知能の発展

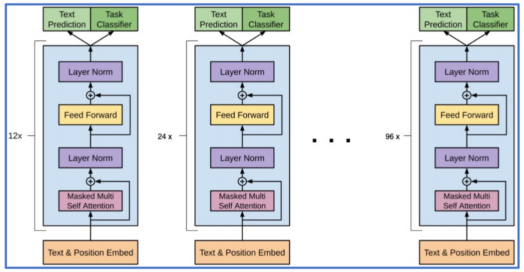
基礎研究 ~2000
技術的課題

深層学習実用化 2012



10Mパラメータ〜
画像分類など

Transformer登場 2017



200Mパラメータ〜
自然言語処理など

ChatGPT公開 2022

100Bパラメータ〜
汎用目的

大規模モデルによる現代的データ科学の発展
⇒ 原理の解明はまだ発展途上

内部の解釈や効率的運用に向けて

深層学習とは何か

多層ニューラルネットワーク(NN)によるデータ解析

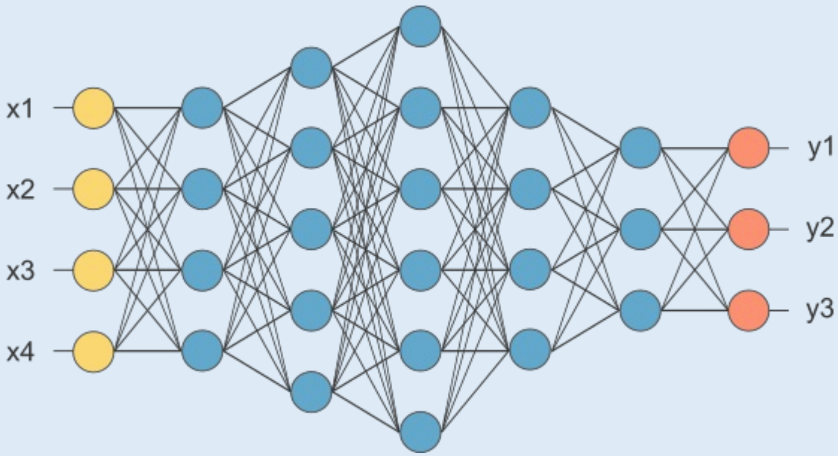
- NN：入力された数値ベクトルを変換する関数モデル

入力（例：画像）

変換

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

ベクトル $\boldsymbol{x}$



多層ニューラルネットワーク

出力（例：情報）

これは
茶色い猫です

変換

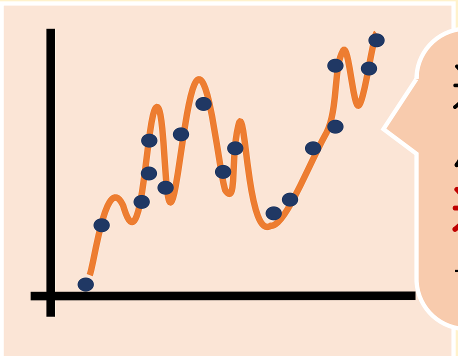
$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

ベクトル $\boldsymbol{y}$

深層学習の原理には謎が多い

明らかになっていない謎の例

<u>従来のデータ解析理論</u>



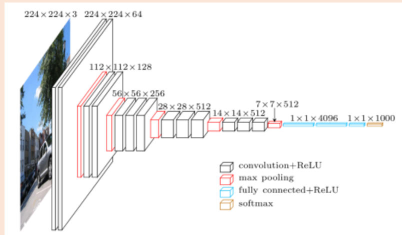
過剰な
パラメタは
過学習する
→<u>性能悪化</u>

誤差は $\frac{p(\text{パラメタ数})}{n(\text{データ数})}$ に比例

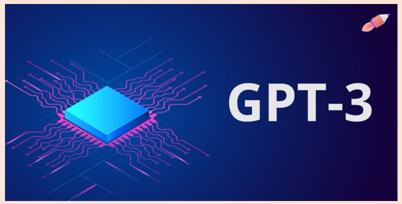
深層学習登場前の**常識**

矛盾

<u>巨大深層学習の成功</u>



<u>VGG19 Net</u>
1億パラメタ



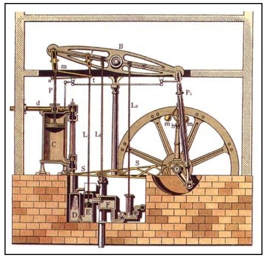
<u>GPT-3</u>
千億パラメタ

パラメタを増やすほど
予測精度が向上

→ 深層学習を理解するためには**理論研究**が必要

“発見”を理論で記述すること

- 歴史的には共通の現象



蒸気機関の発明
(1769年)

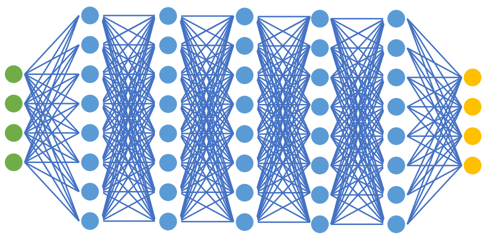
➡

熱力学の
成立

飛行機の発明
(1903年)

➡

航空力学の
成立



深層学習の発明
(2012年)

➡

？

問：深層学習は記述・理解できる理論は構築できるか？

深層学習入門

深層学習の基本構造は関数

- **入力**に対して、適切な**出力**を出すシステム

深層学習システムの中身

多層ニューラルネットワーク

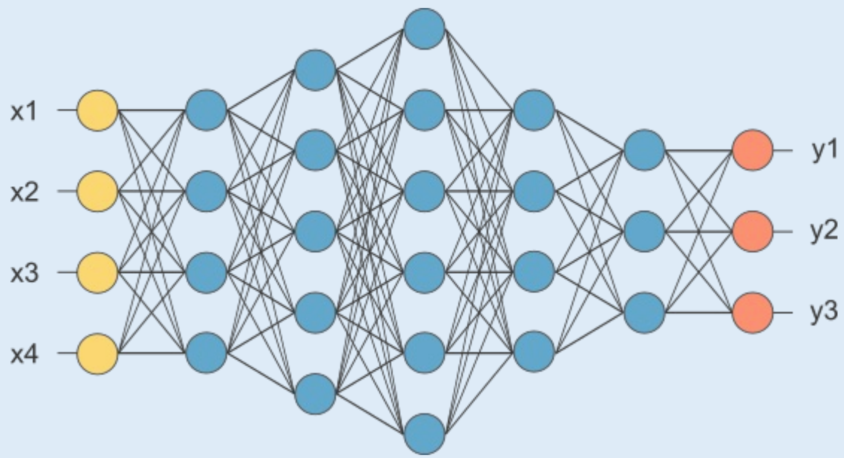
- 入力ベクトルを変換する関数のモデル

入力（例：画像）

変換

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

ベクトル x



多層ニューラルネットワーク

出力（例：情報）

これは
茶色い猫です

変換

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

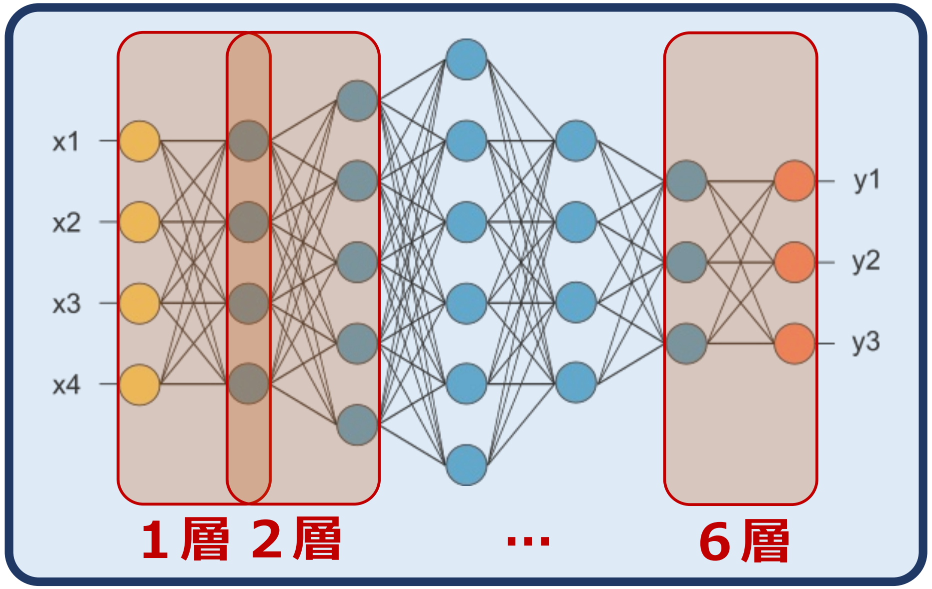
ベクトル y

深層学習システムの中身

ベクトルの変換を**層の数だけ**繰り返す

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

ベクトル x



$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

ベクトル y

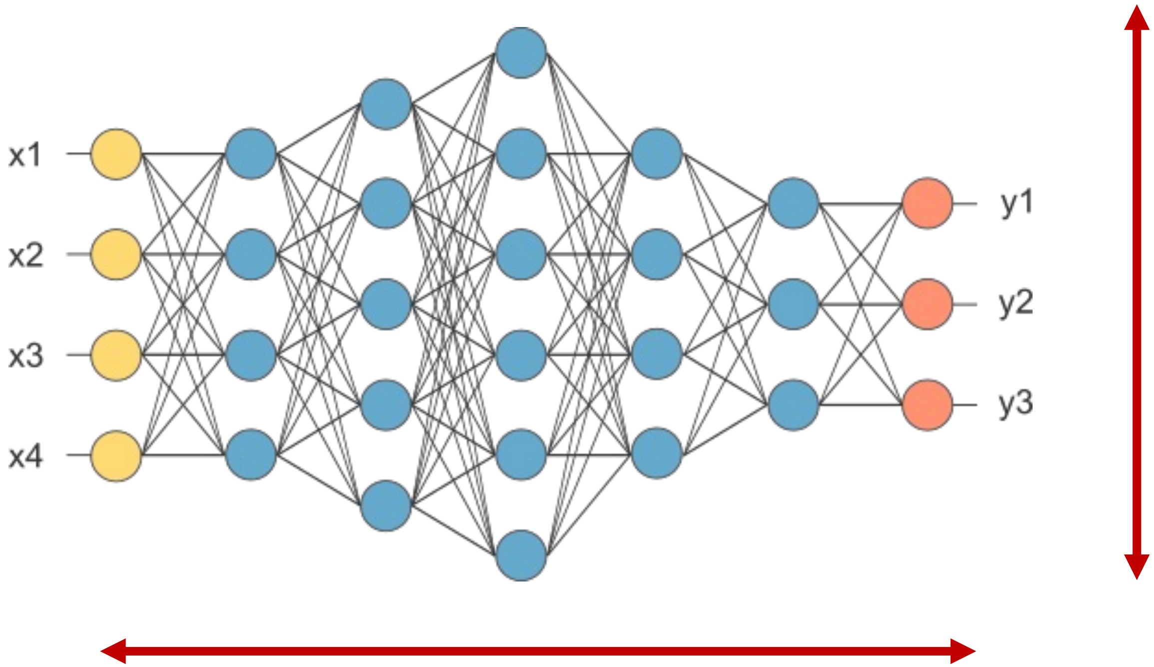
ベクトルの変換

1層目	$z_1 = \sigma(A_1 x + b_1)$
2層目	$z_2 = \sigma(A_2 z_1 + b_2)$
⋮	⋮
6層目	$y = A_6 z_5 + b_6$

A: パラメタ (行列)
b: パラメタ (ベクトル)
σ: 非線型変換

深さと幅

- ネットワークの大きさを特徴付ける量



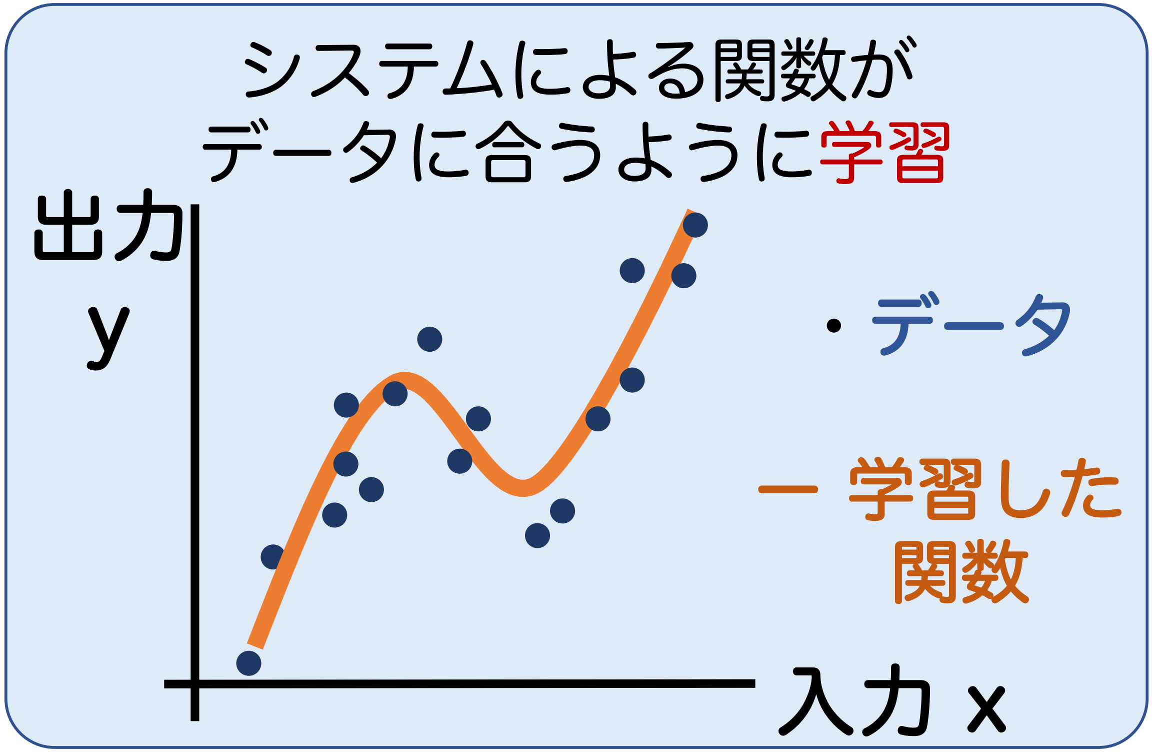
幅（一層あたりノード数の最大数）

$$\max_{\ell} d_{\ell}$$

深さL（層の数）

膨大なパラメタはデータから学習

パラメタ：システムが機能するために必要
- データの構造を再現できるように学習



損失最小化

θ：パラメタ

$\ell(y, y')$：損失関数

$(y_i, x_i)_{i=1}^{n}$：データ

$$\min_\theta \sum_i \ell\big(y_i, f_\theta(x_i)\big)$$

損失
＝システムによる関数と
データのズレ

定式化

- 訓練データからパラメータを学習

訓練データ（n個） $D_n = \{(x_i, y_i)\}_{i=1}^{n}$

- 経験誤差を最小化し、汎化誤差を評価

経験誤差(訓練誤差)

$$R_n(\theta) = n^{-1}\Sigma_{i=1}^{n}\ell\left(y_i, f_\theta(x_i)\right)$$

ℓ: 損失関数

$R_n(\theta)$を小さくするように$\hat{\theta}$を学習

汎化誤差(≈テスト誤差)

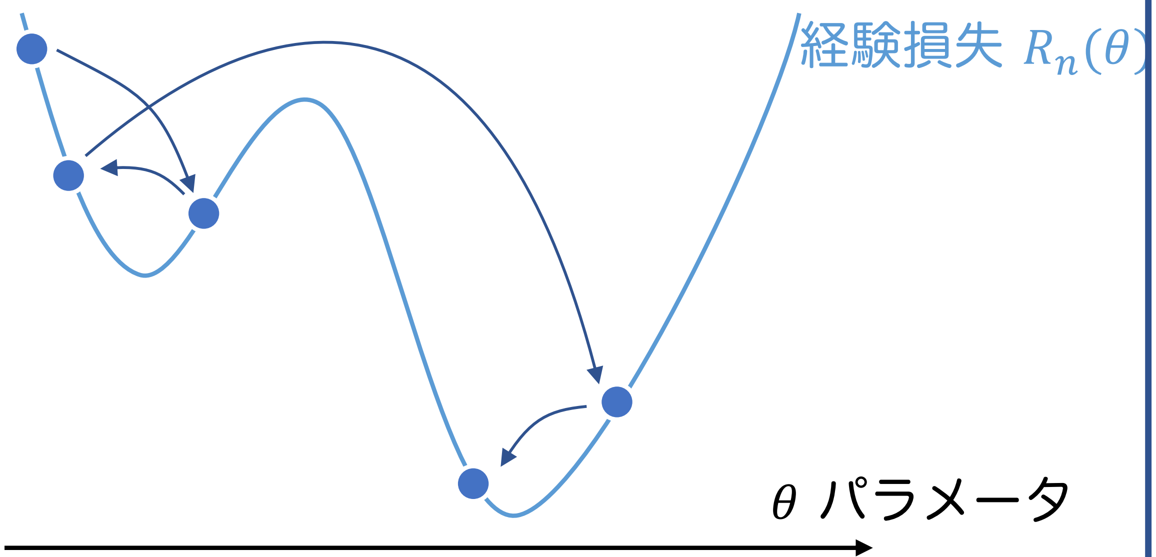
$$R(\hat{\theta}) = E[\ell\left(y, f_{\hat{\theta}}(x)\right)]$$

期待値で性能評価
（新しいデータ上の平均値）

パラメータの学習アルゴリズム

確率的勾配降下法(SGD)

- 勾配＋摂動で損失最小解を探索

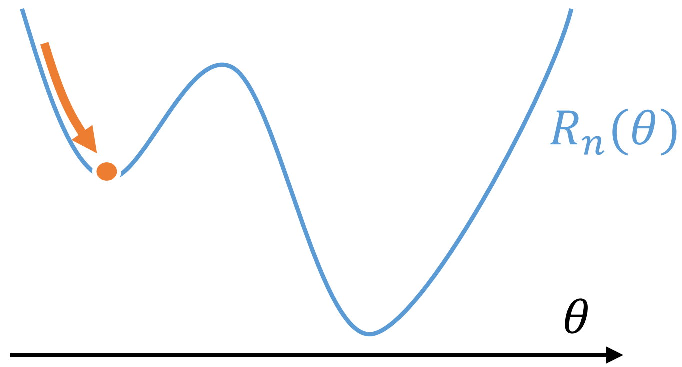


パラメータの反復更新 ($\eta_t > 0$: 学習率)

$\hat{R}\ (\theta)$: 摂動つき経験損失

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \hat{R}\ (\theta)$$

(非確率的)
勾配降下法



摂動がない場合は
浅い谷(局所解)にとどまる

深層学習の理論入門

深層学習の理論の概要

深層学習は多くの要素の組み合わせ

- 以下の3要素へ分解して個別解析

$$R(\hat{\theta}) = \inf_{\theta} R_n(\theta) + R(\hat{\theta}) - R_n(\hat{\theta}) + R_n(\hat{\theta}) - \inf_{\theta} R_n(\theta)$$

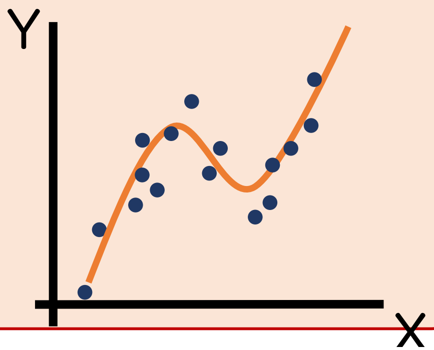
汎化誤差
(予測誤差)

近似誤差
NNの表現力

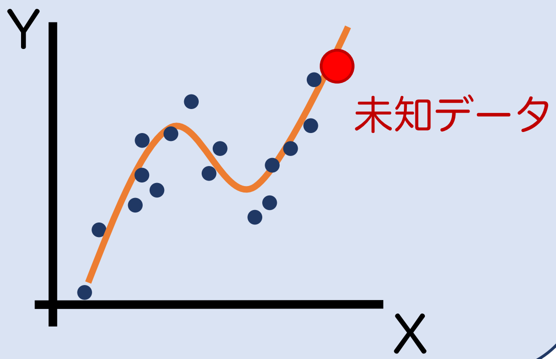
複雑性誤差
予測の安定性

最適化誤差
学習がうまくいくか

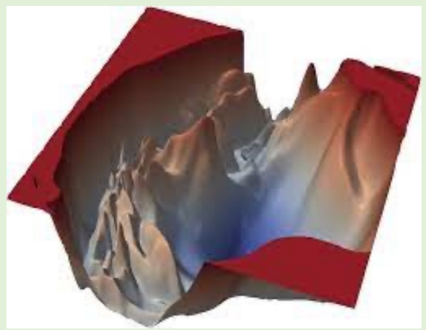
近似誤差
DNNは複雑なデータに
適合できる？



複雑性誤差
なぜ未知データを予測？



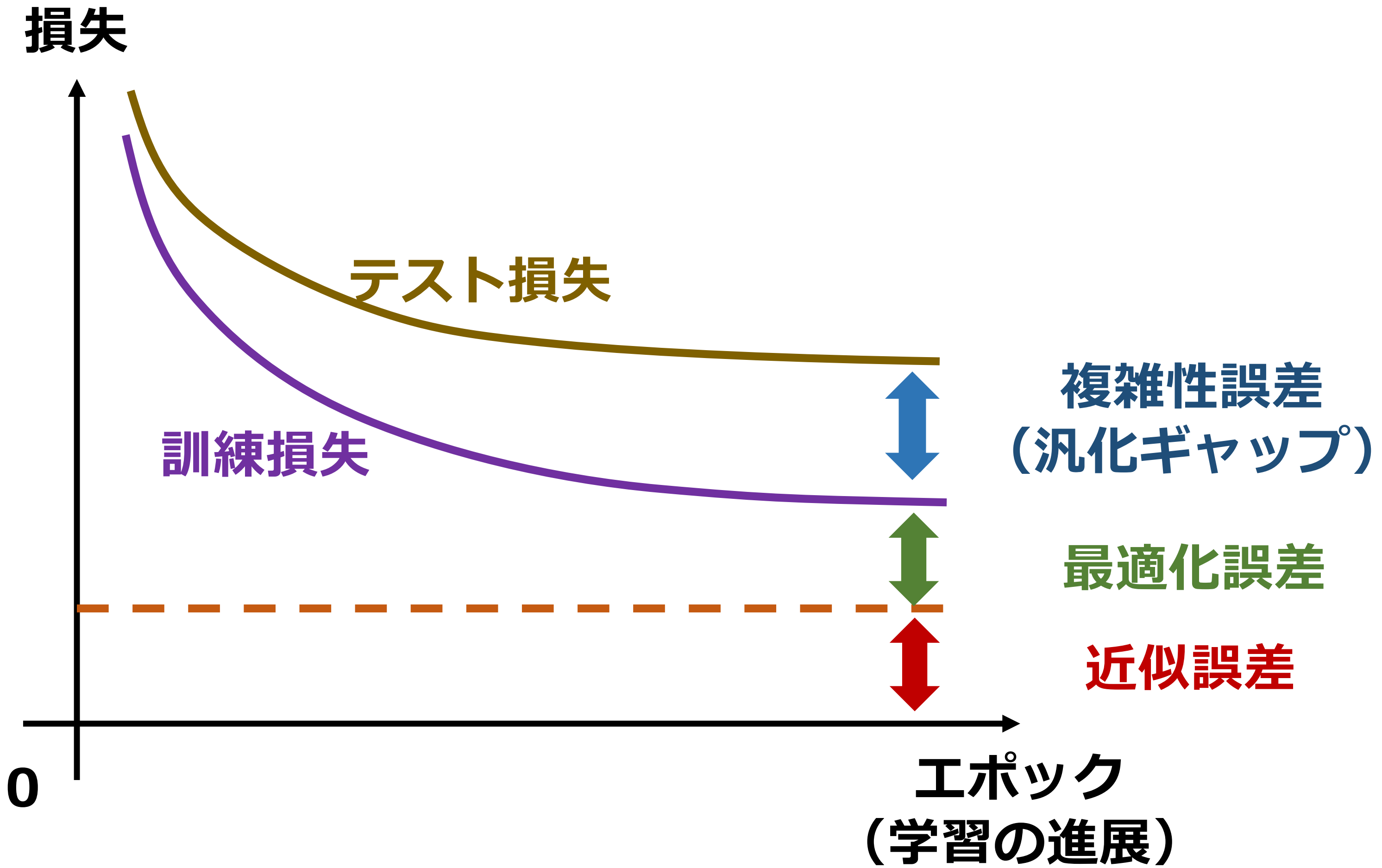
最適化誤差
なぜ学習が成功？



深層モデルの非凸損失

汎化誤差の分解

- 実際の学習との対応

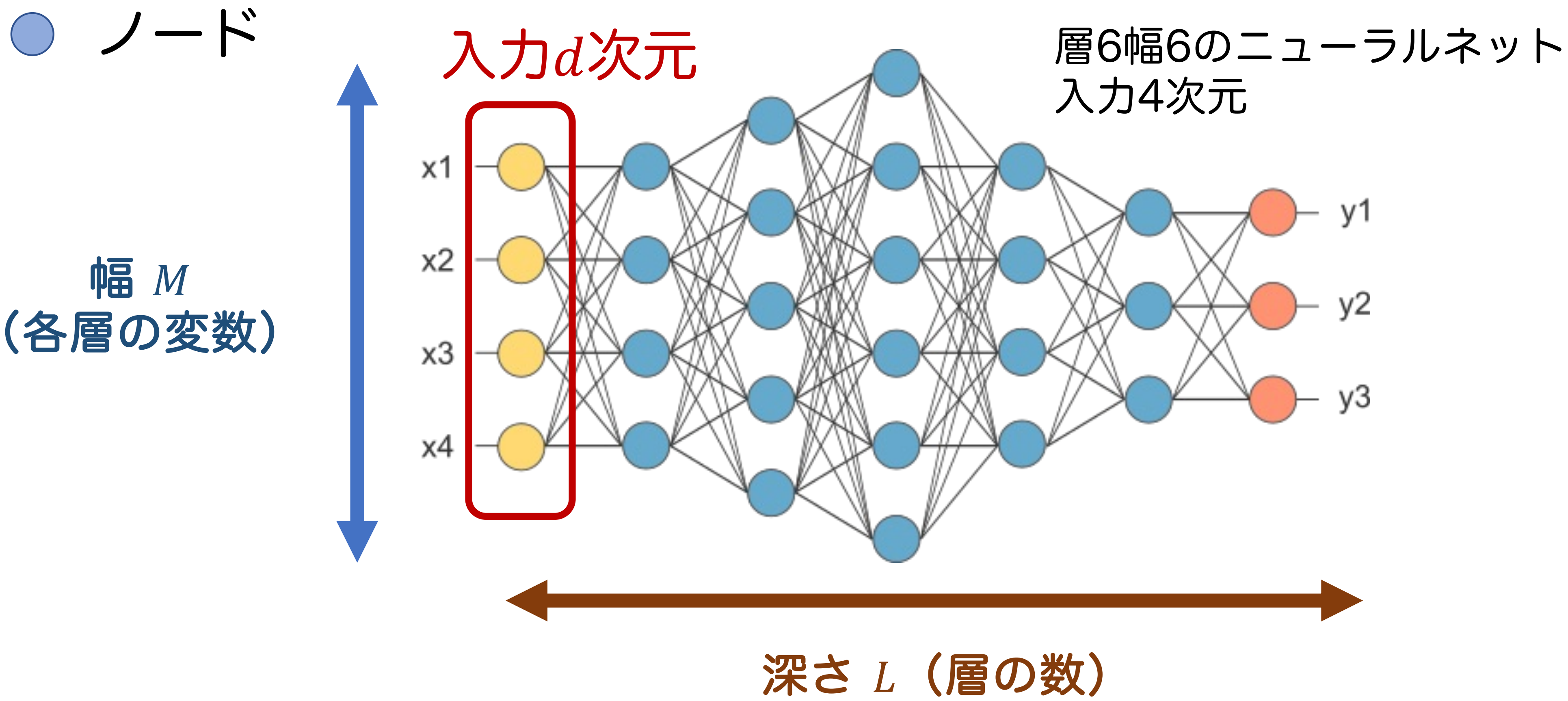


トピック一覧

	近似誤差	複雑性誤差	最適化誤差
従来からの学習理論	普遍近似定理 NNはどんな関数も近似可	汎化ギャップ解析 ギャップ≦パラメータ数	凸・非凸最適化 大域解・停留点への収束
深層に特化した理論	近似レート解析 深層構造でレート改善	暗黙的正則化 アルゴリズムの影響？ 平坦最小解 PAC-Bayesで解析	鞍点・損失曲面解析 SGDの鞍点脱出 定常分布解析 初期値近傍の解析
過剰パラメータ解析		良性過適合 スペクトル解析 二重降下 誤差の再減少	学習レジーム 初期値近傍の解析 特徴学習 ゼロ損失は大域解

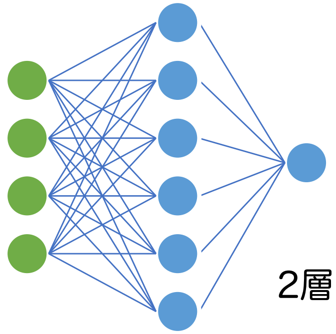
多層ニューラルネットワーク

- ネットワークの性質を深さと幅で測る



この講演では、単純化のために出力は1次元に限定

普遍近似定理



2層NN（隠れ1層）

主張：

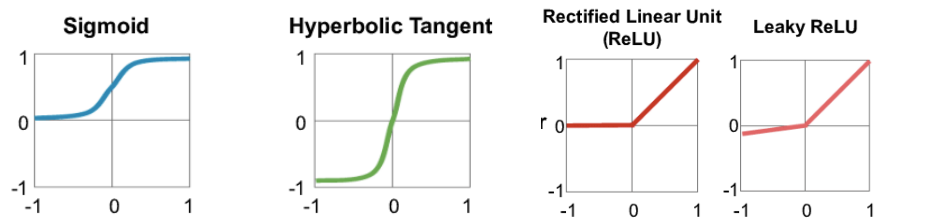
- ニューラルネット(NN)はどんな連続関数でも近似可

普遍近似定理 (Leshno (1993)など)
2層NNの活性化関数 $\sigma: \mathbb{R} \to \mathbb{R}$ が多項式でないとする。
この時、任意の連続関数 $f: [0,1]^d \to \mathbb{R}$ と 任意の誤差 $\varepsilon > 0$に対して、以下の不等式を満たすNNの関数g_θが存在する：

$$\sup_{x \in [0,1]^d} |f(x) - g_\theta(x)| \leq \varepsilon.$$

- 大抵の活性化関数は
  多項式($\sigma(x) = x^a$)ではない

多項式→有限次数の導関数が0になる
非多項式→導関数が0にならない

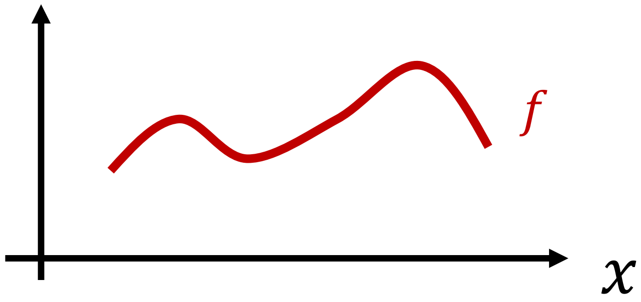


$\sigma(x) = \dfrac{1}{1+\exp(-x)}$ $\sigma(x) = \dfrac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$ $\sigma(x) = \max\{x, 0\}$ $\sigma(x) = \max\{0.1x, x\}$

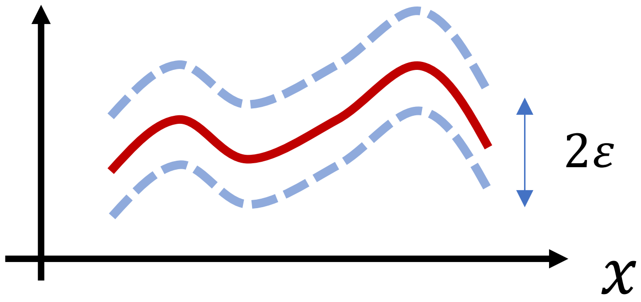
Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*

どういうこと？

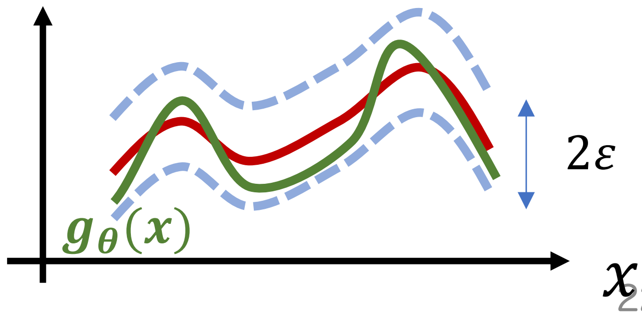
- ある連続な関数 f がある
  - ニューラルネットで
    これを作りたい



- 許容する誤差$\varepsilon > 0$を決める
  - 正ならどんなに小さくても良い



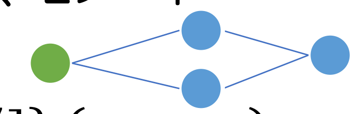
- ニューラルネットは
  この誤差内の関数を必ず作れる
  - 幅がいくらでも増えて良い場合



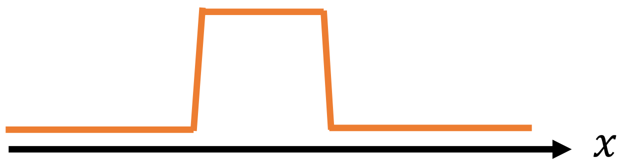
シンプルな証明（$d = 1$次元の場合）

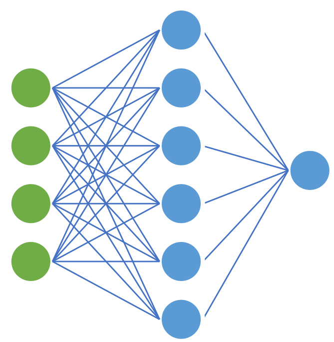
シグモイド活性化を考える: $\sigma(x) = \dfrac{1}{1+\exp(-x)}$, $d = 1$

2層、2ノード



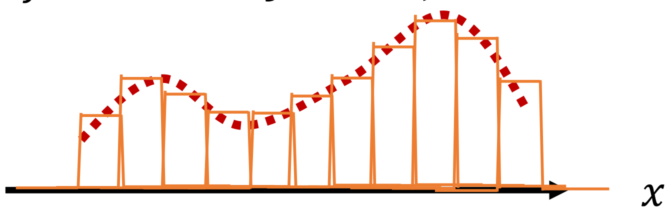
- <u>ステップ 1</u>: 長方形関数を作る
  - $g(x) \coloneqq \sigma\big(a(x+b)\big) - \sigma\big(a(x+b')\big) \approx 1\{x \in [-b, -b']\}\ (a \to \infty)$
    - $1\{\mathrm{A}\}$: 指示関数 (命題 A が真の時に1、それ以外の時に0)





2層、2Jノード

- <u>ステップ 2</u>: 区分定数関数を作る
  - c_j, w_j, e_j: 適切に選ばれた係数
  - 階段状関数（区分定数関数）で連続関数fを近似
  - $\sum_{j=1}^{J} e_j g\big(c_j x - w_j\big)$ はfに収束 ($J \to \infty$).



普遍近似定理のバリエーション

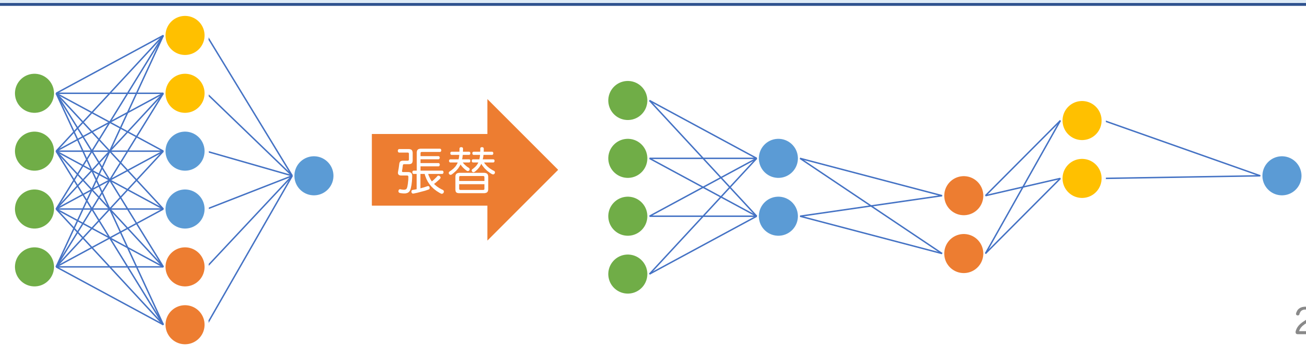
解析対象の変遷

- 1990年代→層が少なく幅が広いニューラルネット
- 2019年代→層が多い幅が狭いニューラルネット

普遍近似定理 (Park (2021)など)

活性化関数 $\sigma: \mathbb{R} \to \mathbb{R}$ が多項式でないとする。
この時、任意の連続関数 $f: [0,1]^{d_x} \to [0,1]^{d_y}$ について、幅が $\max\{d_x + 2, d_y + 2\}$の多層ニューラルネットf_θは普遍近似性 ($\sup_x |f(x) - f_\theta(x)| \le \varepsilon$)を満たす。

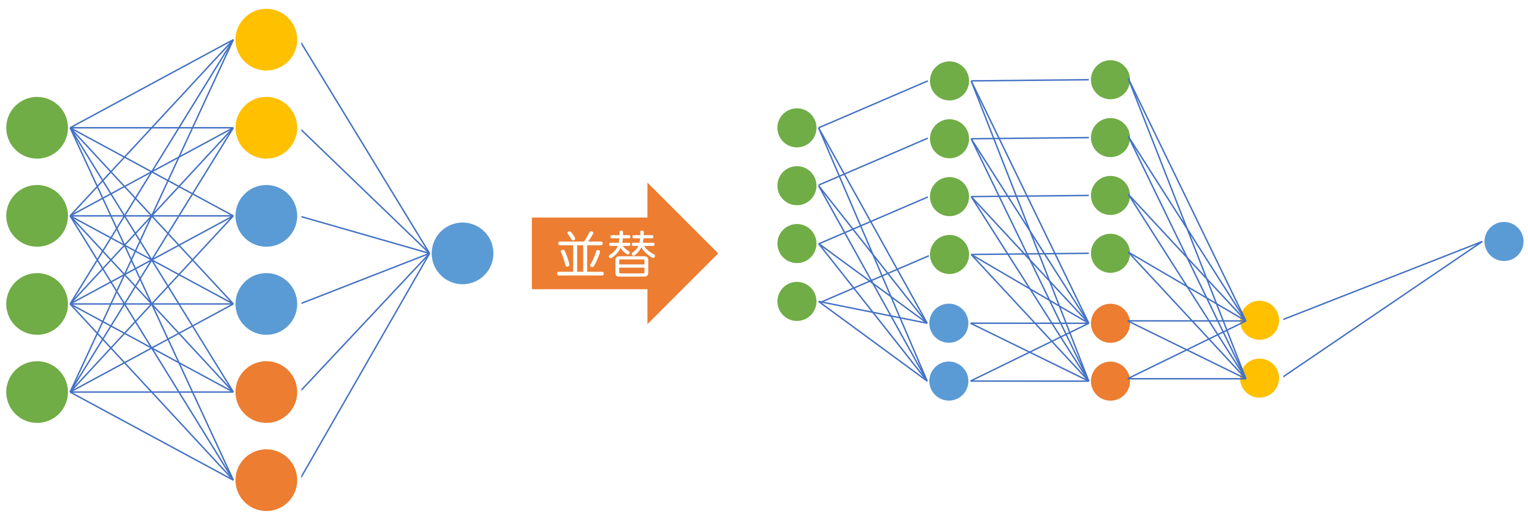
- 証明の直感



Park, S., Yun, C., Lee, J., & Shin, J. (2921). Minimum width for universal approximation. *ICLR*.

証明の概要

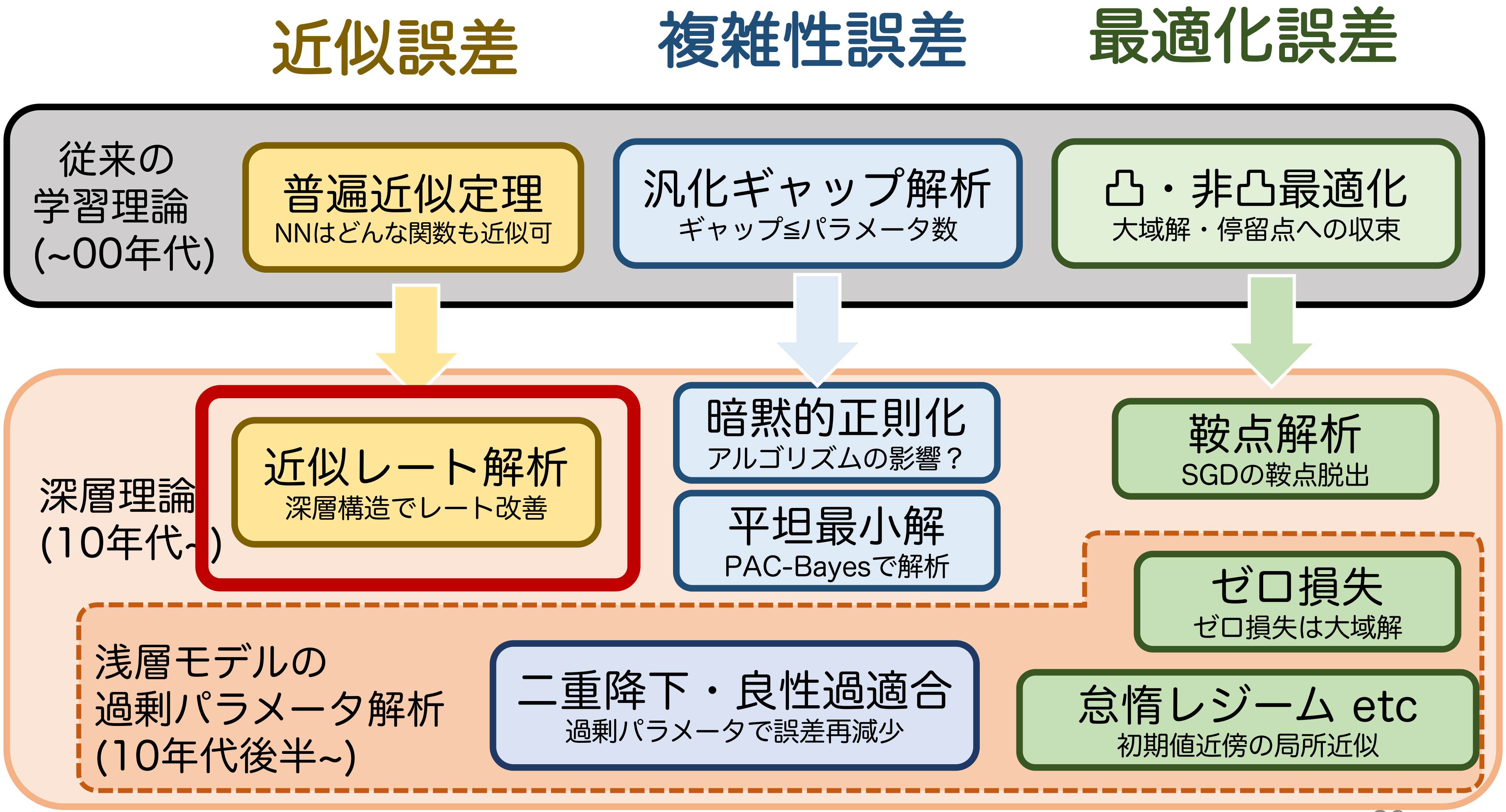
- 方針：
  - 最初に、関数を近似する2層ネットワークを構成
  - 各ノードを並び替えて、1層あたり少数のノード計算だけを行うようにする



元になる、層の少ない
ニューラルネットワーク

並び替えで作った、層の多い
ニューラルネットワーク

深層学習理論マップ



近似誤差
複雑性誤差
最適化誤差
従来の
学習理論
(~00年代)
普遍近似定理
NNはどんな関数も近似可
汎化ギャップ解析
ギャップ≦パラメータ数
凸・非凸最適化
大域解・停留点への収束
深層理論
(10年代~)
近似レート解析
深層構造でレート改善
暗黙的正則化
アルゴリズムの影響？
平坦最小解
PAC-Bayesで解析
鞍点解析
SGDの鞍点脱出
ゼロ損失
ゼロ損失は大域解
浅層モデルの
過剰パラメータ解析
(10年代後半~)
二重降下・良性過適合
過剰パラメータで誤差再減少
怠惰レジーム etc
初期値近傍の局所近似

より詳細に近似を調べるには

<u>復習：ランダウのO記号</u>

$x_N = O(y_N), (N \to \infty)$ とは

$\exists N', C > 0 \text{ s.t. } N > N' \Rightarrow |x_N| \le C|y_N|$

<u>近似誤差の減衰レート</u>

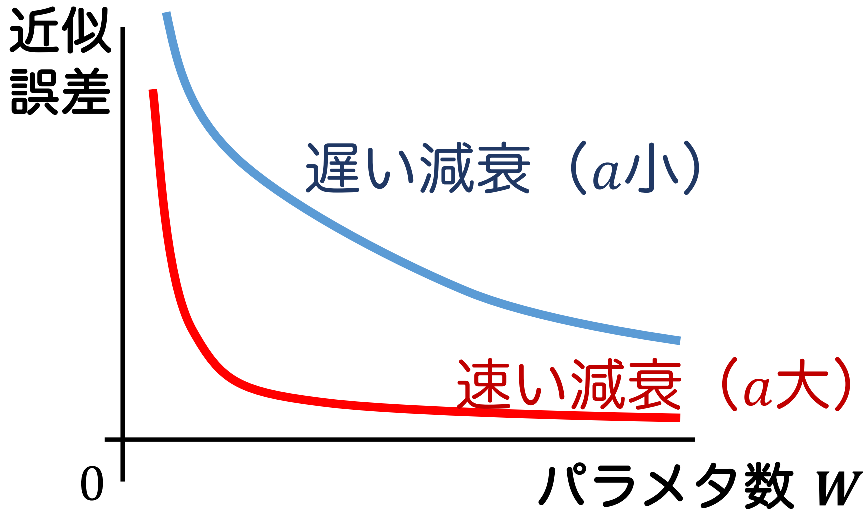
- パラメタ(エッジ)が増えるときの誤差減少スピード

近似誤差の減衰レート a

f^*: 近似対象の関数

f_θ: W個のパラメタを持つDNN

$$\underbrace{\inf_\theta \|f^* - f_\theta\|}_{\text{近似誤差}} = \underbrace{O(W^{-a})}_{\text{パラメタ増で減少}}$$



- レートを出すには、f^*が滑らかである必要
  ➡ f^*が微分可能である状況を調べる

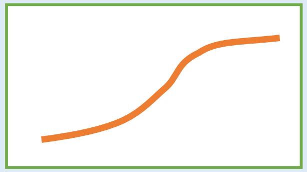
滑らかな関数に対する近似レート

f^*：近似対象（入力d次元、β回微分可能）

f_θ：DNN（L層, パラメタW個、活性化関数σ）

σがsigmoid等の場合 (Mhaskar (1996)など)

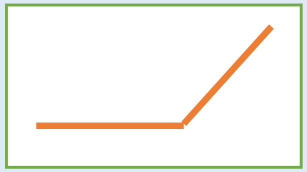
DNNは $L=2$ のもとで以下を達成：



$$\inf_\theta \|f^* - f_\theta\| = O(W^{-\beta/d})$$

活性化関数がReLUの場合 (Yarotsky (2017)など)

L層のDNNは以下を達成：



$$\inf_\theta \|f^* - f_\theta\| = O(W^{-\beta/d} + 2^{-L})$$

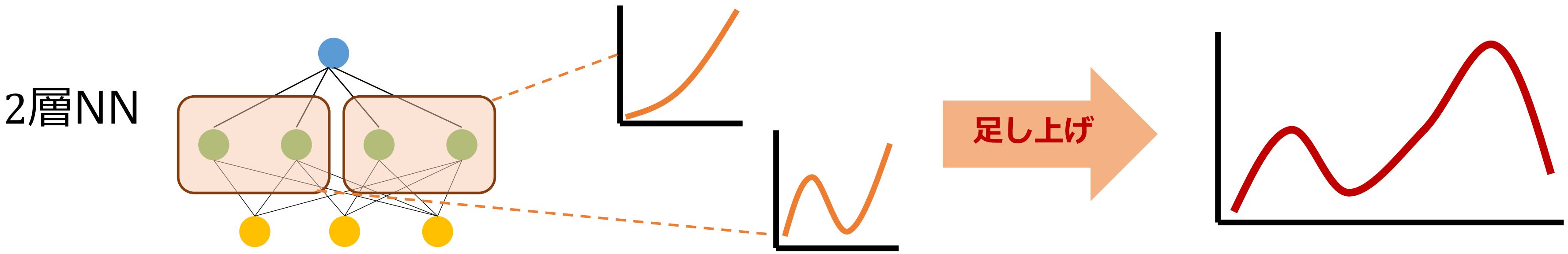
ReLUの尖りから来る影響

誤差レートβ/dは、 f^*の滑らかさで増加、入力次元で減少。

Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*.
Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*.

活性化関数σによる違い

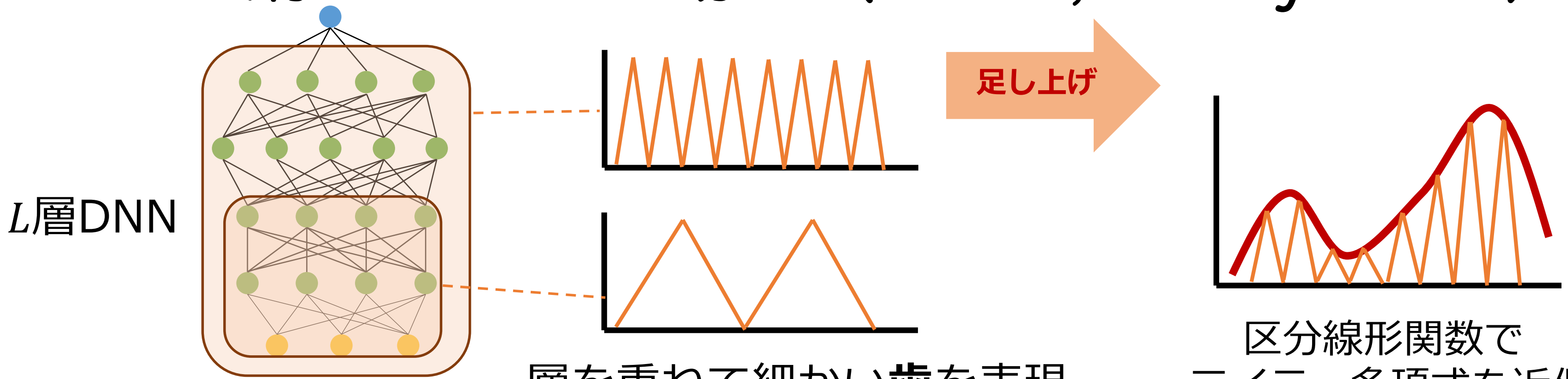
- σが滑らかな場合（sigmoid, softplus）



ノード毎に多項式近似

テイラー多項式を表現

- σが滑らかでない場合（ReLU, LeakyReLU）



層を重ねて細かい**歯**を表現

区分線形関数で
テイラー多項式を近似

深さと幅を用いた近似誤差

- パラメータ数Wではなく層数・幅数による解析

活性化関数がReLUの場合 (Lu et al. (2021)など)
L層かつ幅NのDNNは、 β回微分可能な関数f^*に対して

$$\inf_{\theta} \|f^* - f_\theta\| = O(N^{-2\beta/d} L^{-2\beta/d})$$

- パラメータ数Wと層数・幅数の関係
  - $W \leq LN(N+1) = O(LN^2)$
- 層数Lが定数だと思うと、上のレートは$O(W^{-\beta/d})$
  - 既存理論と整合的
- 層数Lが増えると、 Wの増加を抑制できる

Lu, J., Shen, Z., Yang, H., & Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, *53*(5), 5465-5506.

良いレートなの？

すごいぞ！
やはりDNN
は
最適なんだ！

- 誤差レートβ/dは理論的に最適

近似誤差の最適性 (DeVore+ (1989)など)
近似誤差レート$O(W^{-\beta/d})$は理論上の最適値。

- しかし、他手法も同じように最適

他のも最適だから
結局同じ？

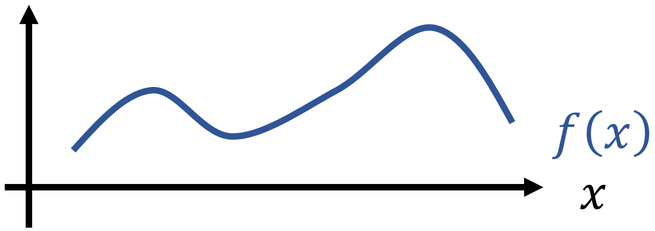
他手法の近似レート (Newman+ (1964)など)
フーリエ基底、多項式基底などによる近似は
レート$O(W^{-\beta/d})$を達成する。

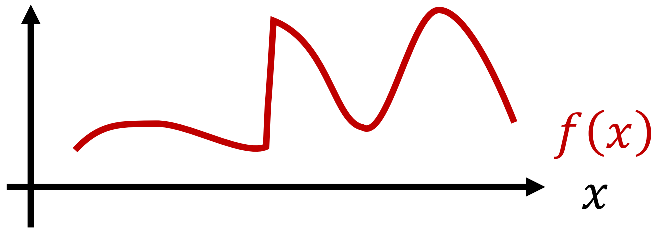
f^*が滑らかなら、DNNと他手法の理論的性能は同等。

DeVore, R. A., Howard, R., & Micchelli, C. (1989). Optimal nonlinear approximation. *Manuscripta mathematica*,

層の多さが必要になる状況

- 局所構造を持つ関数fの近似

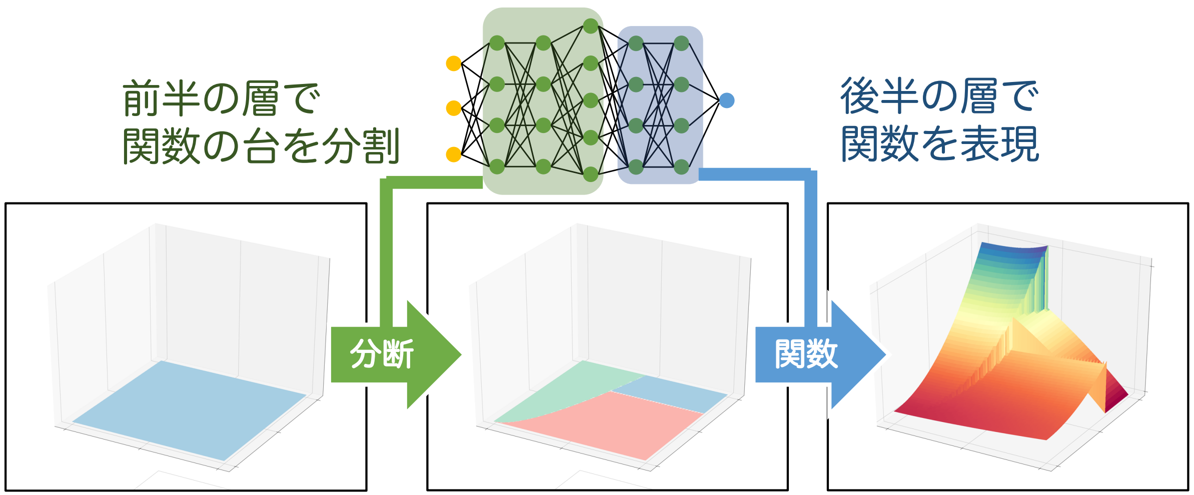




局所構造なし(滑らかさが一定)
→非深層手法でOK

局所構造あり(左右で滑らかさが異なる)
→DNNが有利

  - 深層構造が局所性の表現に貢献



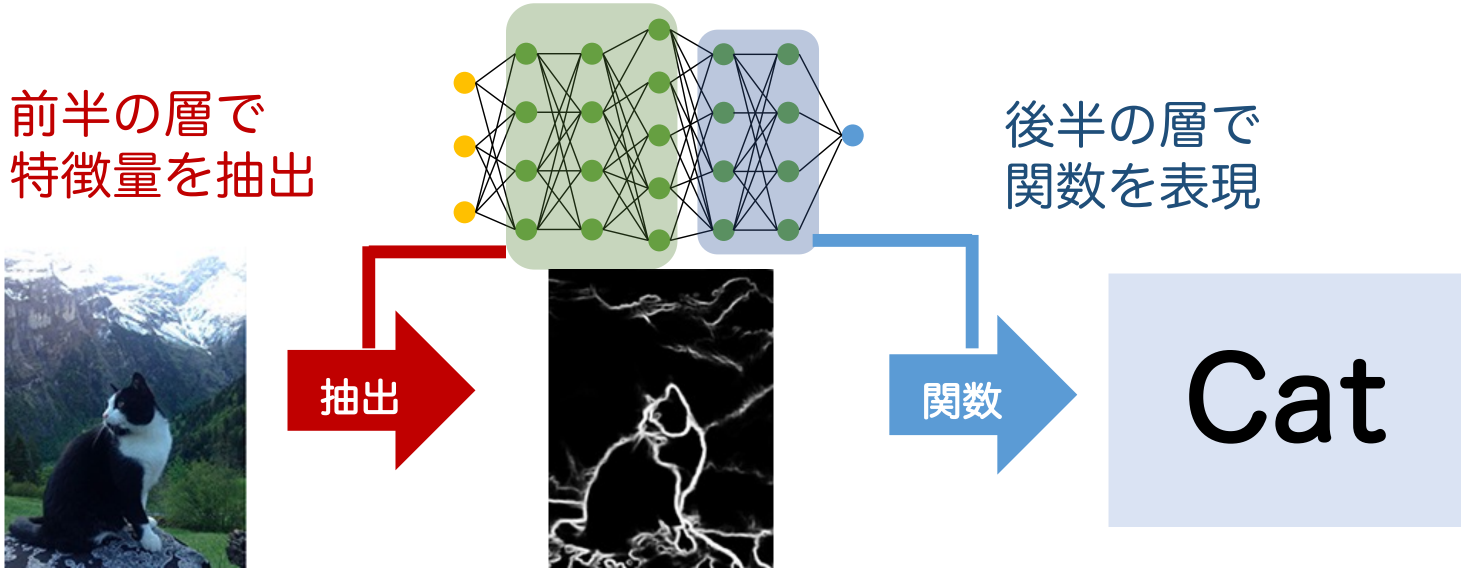
DNNが速いレートを達成

DNNの近似レート：
$O(\max\{W^{-\beta/d}, W^{-\alpha/2(d-1)}\})$
他手法の近似レート：
$O(\max\{W^{-\beta/d}, W^{-\alpha/4(d-1)}\})$
α: 境界線の滑らかさ

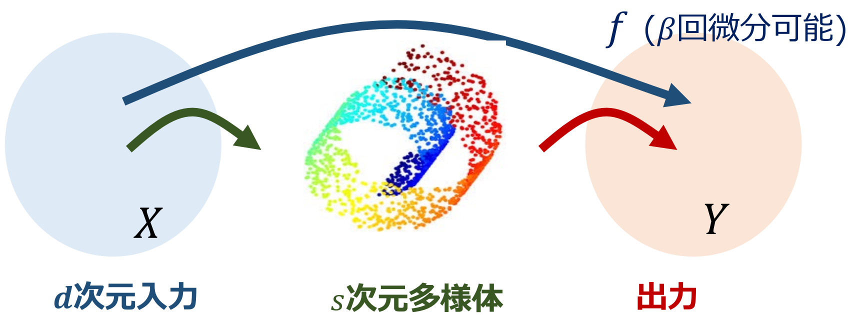
Imaizumi, M., & Fukumizu, K. (2022). Advantage of Deep Neural Networks for Estimating Functions with Singularity on Hypersurfaces. *JMLR*.

層の多さが必要になる状況

- 特徴量抽出が有効な関数fの近似



- 特徴量の次元によってレート改善



多様体次元sのレート

DNNのレート：
$\tilde{O}(W^{-\beta/s})$
一般的なレート：
$O(W^{-\beta/d})$

Nakada, R., & Imaizumi, M. (2020). Adaptive Approximation and Generalization of Deep Neural Network with Intrinsic Dimensionality. *JMLR*.

ここまでのまとめ

- 近似誤差には多くの研究
  - 連続関数ならなんでも近似可能
  - なめらかな関数なら最適な誤差レート

- 21世紀の近似研究の特徴
  - 層の多さを扱う（より深く、より幅が少なく）
  - 滑らかでない活性化関数（ReLU）がより調査

- 深層学習の理論研究の中では、
近似誤差はわかっていることが多い。

複雑性誤差

汎化誤差の分解

深層学習は多くの要素の組み合わせ

- 以下の3要素へ分解して個別解析

$$R(\hat{\theta}) = \inf_{\theta} R_n(\theta) + R(\hat{\theta}) - R_n(\hat{\theta}) + R_n(\hat{\theta}) - \inf_{\theta} R_n(\theta)$$

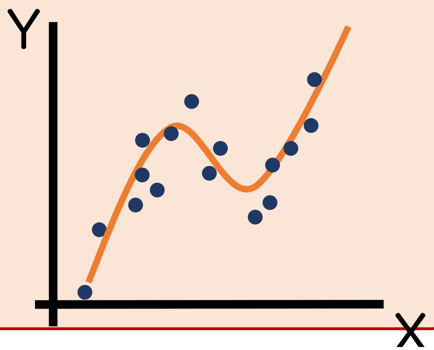
汎化誤差 (予測誤差)

近似誤差 NNの表現力

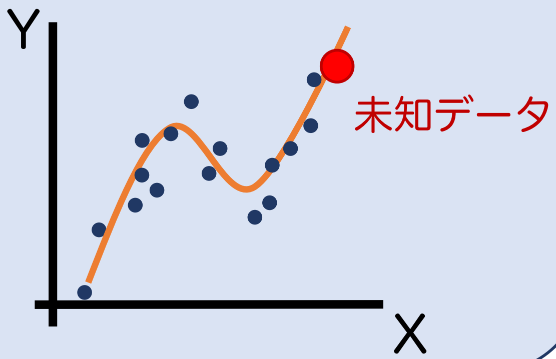
複雑性誤差 予測の安定性

最適化誤差 学習がうまくいくか

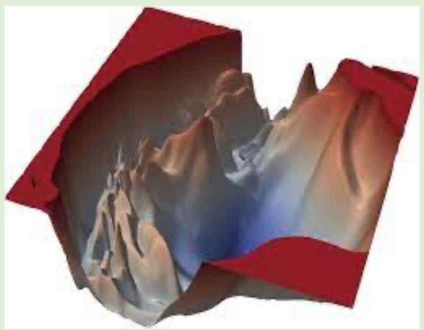
近似誤差
多層は複雑なデータに適合できる？



複雑性誤差
なぜ未知データを予測？

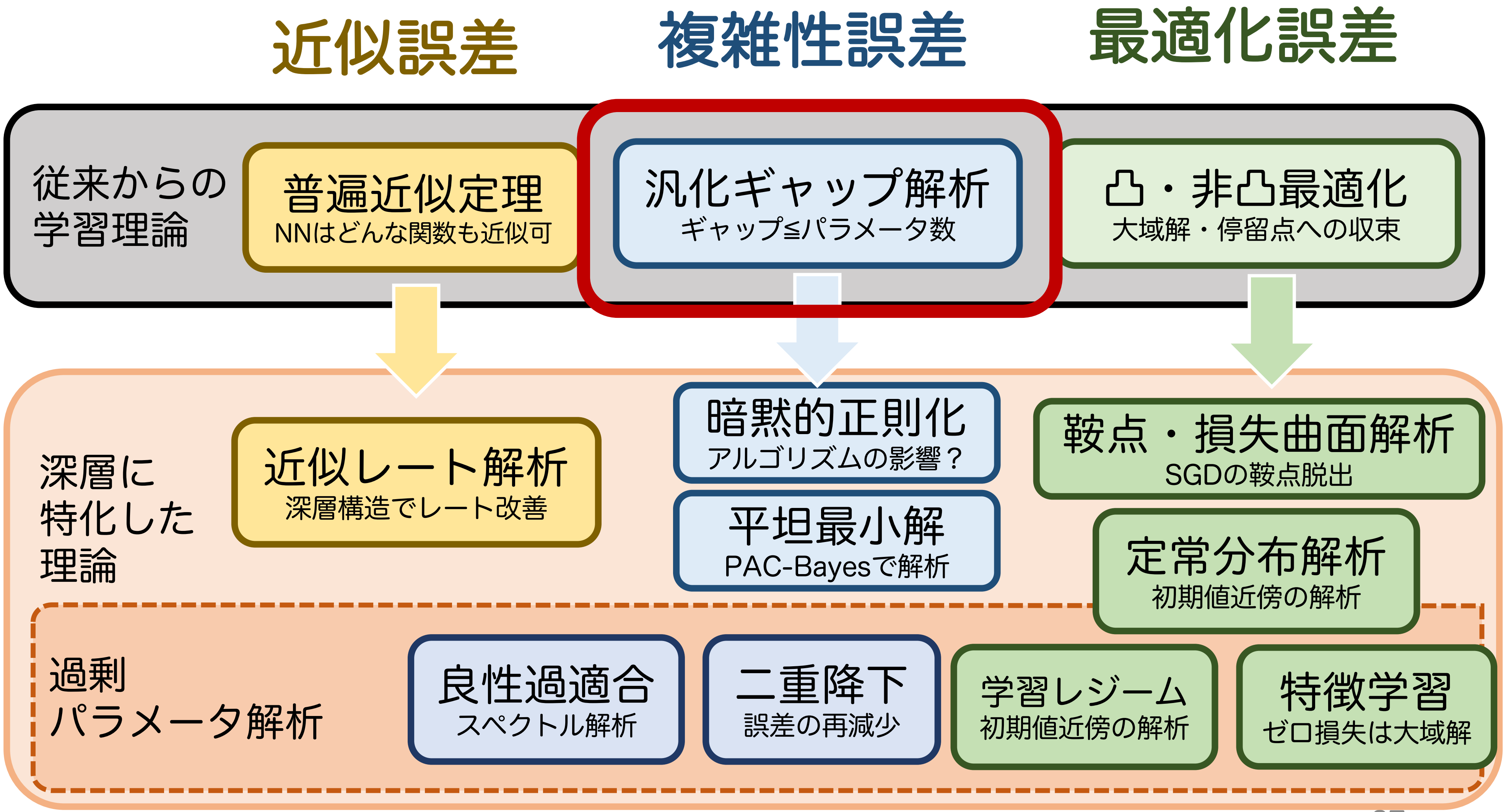


最適化誤差
なぜ学習が成功？



深層モデルの非凸損失

今日話すトピック


近似誤差
複雑性誤差
最適化誤差
従来からの学習理論
普遍近似定理
NNはどんな関数も近似可
汎化ギャップ解析
ギャップ≦パラメータ数
凸・非凸最適化
大域解・停留点への収束
深層に特化した理論
近似レート解析
深層構造でレート改善
暗黙的正則化
アルゴリズムの影響？
平坦最小解
PAC-Bayesで解析
鞍点・損失曲面解析
SGDの鞍点脱出
定常分布解析
初期値近傍の解析
過剰パラメータ解析
良性過適合
スペクトル解析
二重降下
誤差の再減少
学習レジーム
初期値近傍の解析
特徴学習
ゼロ損失は大域解

古典的な理論

汎化ギャップの評価

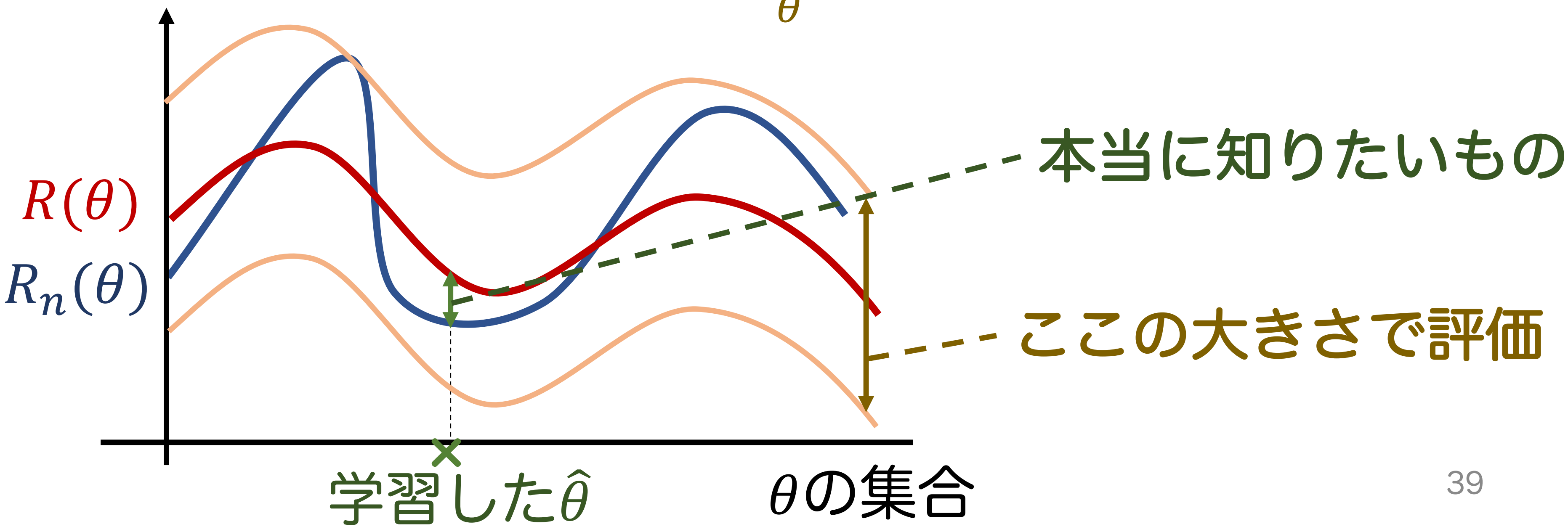
そもそも複雑性誤差とは？

- 汎化誤差(期待値)と訓練誤差(経験平均)との差

$$\left|R(\hat{\theta}) - R_n(\hat{\theta})\right|$$

<u>既存理論</u>：**可能な全てのθ**上での$|R(\theta) - R_n(\theta)|$

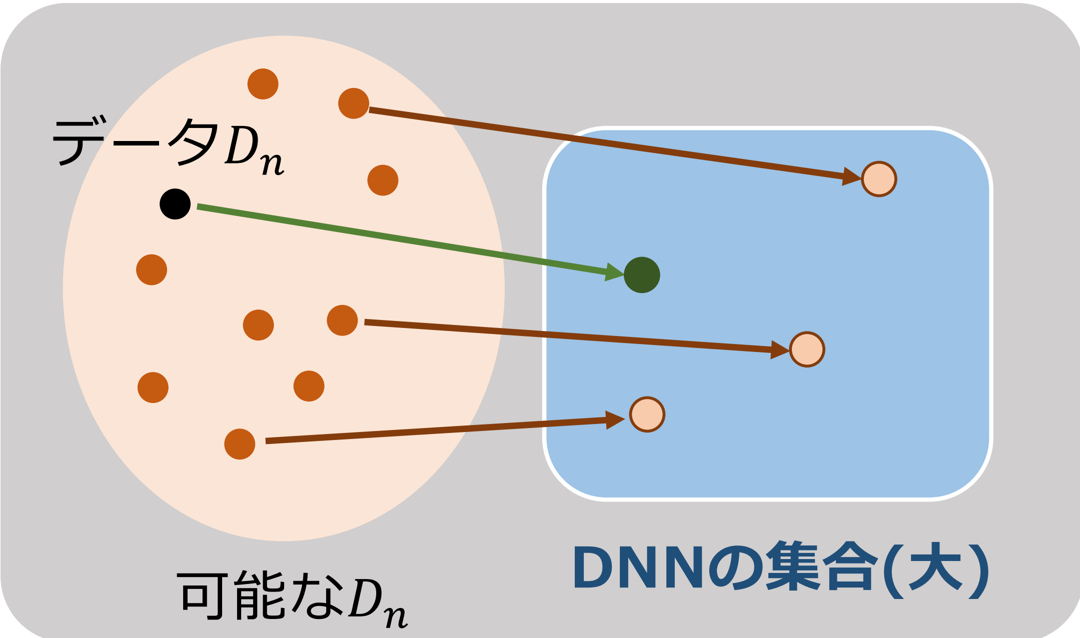
$$\left|R(\hat{\theta}) - R_n(\hat{\theta})\right| \leq \sup_{\theta}|R(\theta) - R_n(\theta)|$$



既存の理論の考え方

モデルの大きさが重要

- 複雑性誤差＝DNN関数f_θの集合　　の大きさ



可能なD_nから定まる$f_{\hat{\theta}}$を
すべて考慮

可能な$f_{\hat{\theta}}$の候補集合が
大きいほど
複雑性誤差が増加

複雑性評価の数学的方法

レートは改善可だが
<u>大きさ</u>への依存は不可

評価の理論 (e.g. Anthony & Bartlett (1999))

n: データ数
W: パラメタ数

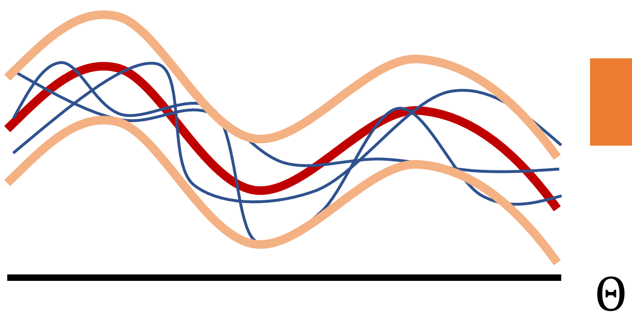
$$\sup_{\theta}|R(\theta) - R_n(\theta)| = O\left(\frac{1}{\sqrt{n}} \int_0^{\infty} \sqrt{\log N_\delta}\, d\delta\right)$$

可能なf_Θの集合の大きさ

<u>**導出の流れ**</u>

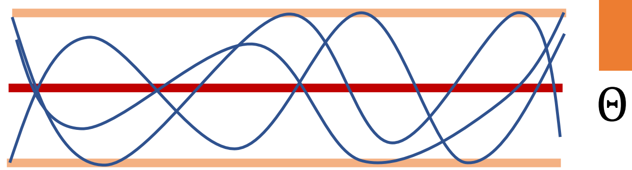
一様誤差

$$\sup_{\theta}|R(\theta) - R_n(\theta)|$$



Rademacher複雑性

$$n^{-1/2}\mathrm{E}\left[\sup_{\theta} \Sigma_{i=1}^{n} \sigma_i \ell\big(y_i, f_\theta(x_i)\big)\right]$$

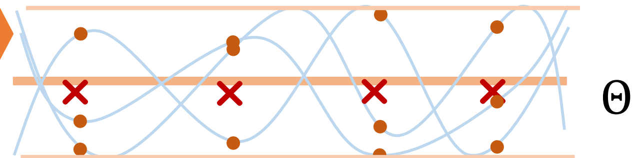


σ_i: Rademacher変数

Dudley積分

$$n^{-1/2}\int_0^{\infty} \sqrt{\log N_\delta}\, d\delta$$

→集合の大きさを評価



N_δ: $\{f_\theta\}$の最小δ被覆数
×: 離散点（被覆球の中心）

複雑性はパラメタ数が主

DNNの複雑性評価 (e.g. Anthony & Bartlett (1999))

$$\left|R(\hat{\theta}) - R_n(\hat{\theta})\right| = O\left(\frac{\sqrt{LW \log W}}{\sqrt{n}}\right)$$

➡ パラメタ数 W が主な要素

- この理論は深層学習の実性能を説明できない

大量のパラメタは
複雑性誤差を上げる

矛盾

高精度DNNは
膨大なパラメタ数
Alex Net →6千万
VGG Net →1億

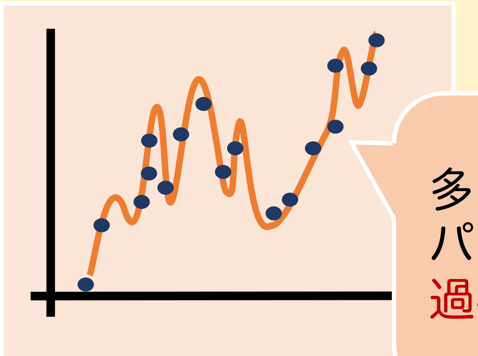
統計・学習理論の(大)原則

Anthony, T. F. M., & Bartlett, P. L. Neural Network Learning Theoretical Foundations.

深層学習をめぐる謎

既存理論と深層学習の実際の矛盾

既存理論

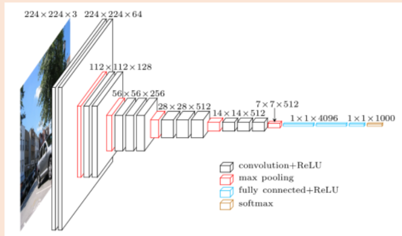


多すぎる
パラメータは
過学習！

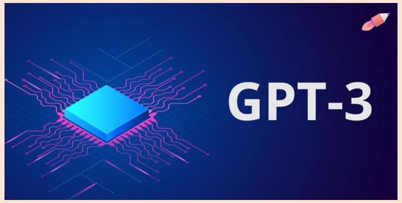
$$誤差 \propto \sqrt{\frac{パラメータ数}{データ数}}$$

矛盾

深層学習の成功



VGG19 Net
100 million~
parameters



GPT-3
100 billion~
parameters

パラメータが多いほど
精度が上がる

→ 理論の再考: 理論的な理解の再構築の必要性

過剰パラメータの理論

今日話すトピック

	近似誤差	複雑性誤差	最適化誤差
従来からの学習理論	普遍近似定理 NNはどんな関数も近似可	汎化ギャップ解析 ギャップ≦パラメータ数	凸・非凸最適化 大域解・停留点への収束
深層に特化した理論	近似レート解析 深層構造でレート改善	暗黙的正則化 アルゴリズムの影響？ 平坦最小解 PAC-Bayesで解析	鞍点・損失曲面解析 SGDの鞍点脱出 定常分布解析 初期値近傍の解析
過剰パラメータ解析	良性過適合 スペクトル解析	二重降下 誤差の再減少	学習レジーム 初期値近傍の解析 特徴学習 ゼロ損失は大域解

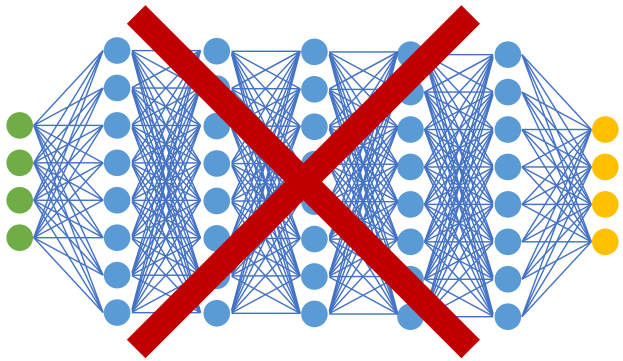
過剰パラメータ理論への関心

良：関心の高まり

- $p \gg n$　（パラメタ数≫データ数）な状況への注目

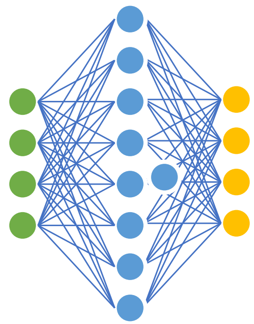
悪：理論解析の限界

- 深層モデルは数学的に解析できない
  →1層 or 2層モデルが解析の対象



$$f(x) = A_L\sigma(A_{L-1}\sigma(A_{L-2}\cdots A_2\sigma(A_1 x))$$

多層構造で線形性がなくなる

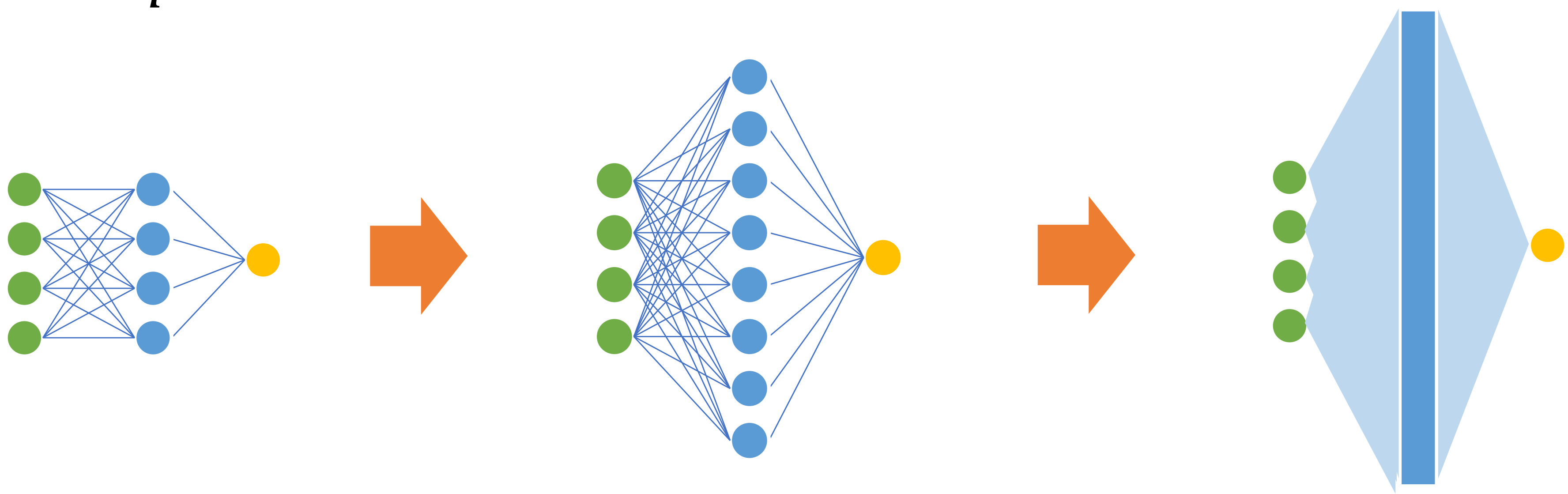


$$f(x) = A_2\sigma(A_1 x)$$

特徴量 $\sigma(A_1 x)$ の線形関数でかける

大規模モデルの極限

- 無限個のパラメータとデータを考える
  - p: パラメータ数、 n: データ数



p: 小, n: 小

p: 大, n: 大

p: 無限, n: 無限
p/n が一定値

有限のpやnに依存しない**極限**でモデルの性質を解析

主たる関心：線形モデル

線形回帰

- 訓練データ $D_n = \{(x_i, y_i)\}_{i=1}^n$, X_iはp次元ベクトル
- 線形回帰モデル

$$y_i = {\beta^*}^\top x_i + \varepsilon_i, \quad \beta^* \text{は} p \text{次元パラメータ}$$

2層ニューラルネット

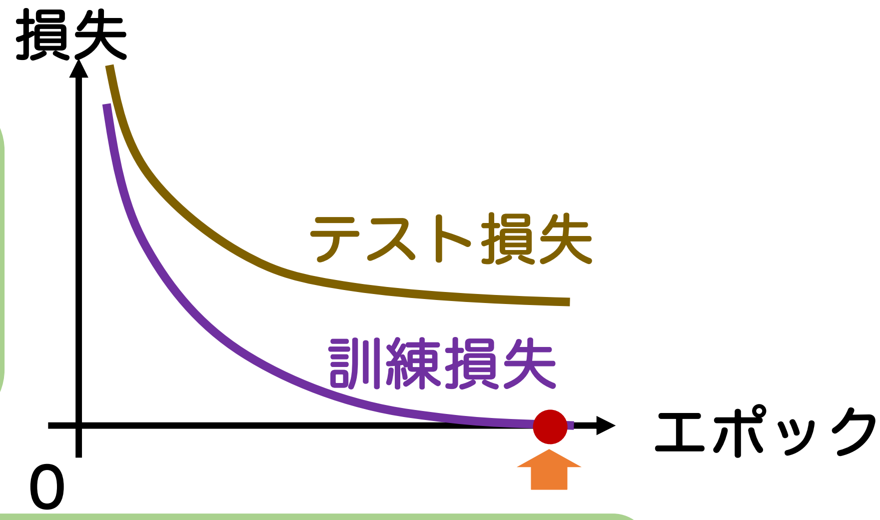
- 訓練データ $D_n = \{(x_i, y_i)\}_{i=1}^n$, X_iはd次元ベクトル
- $A \in \mathbb{R}^{p \times d}$: 1層目の重み行列
- σ: 活性化関数

$$y_i = {\beta^*}^\top \sigma(A x_i) + \varepsilon_i, \quad \beta^* \text{は} p \text{次元パラメータ}$$

- Aをランダムに生成すると、線形回帰と似た性質を持つ

二重降下

過剰パラメータ



補間量 (interpolator)
訓練損失をゼロにする学習器
（訓練データに完全フィットする）

過剰パラメタ化 (over-parameterization)
学習器のパラメタ数を過剰に増やすこと

近年の発見 (線形モデル)

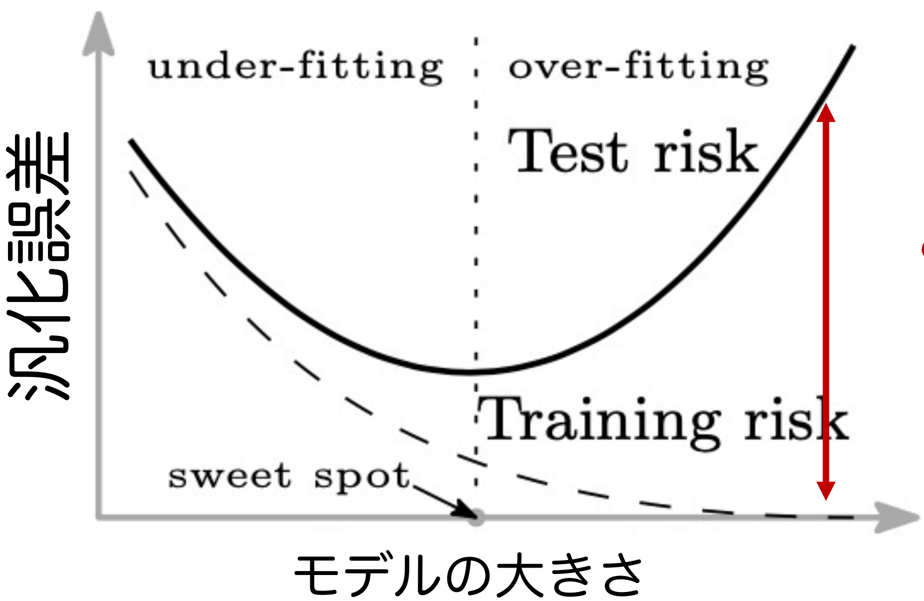
- 補間量の複雑性誤差は、過剰パラメタ化のもとで減少

二重降下という考え

二重降下

- モデルを過剰に大きくすると、
  バリアンス（複雑性誤差）が逆に減少すること

既存理論の考え



このギャップが複雑性誤差
（テスト誤差-訓練誤差）

図はBelkin+ 2019より

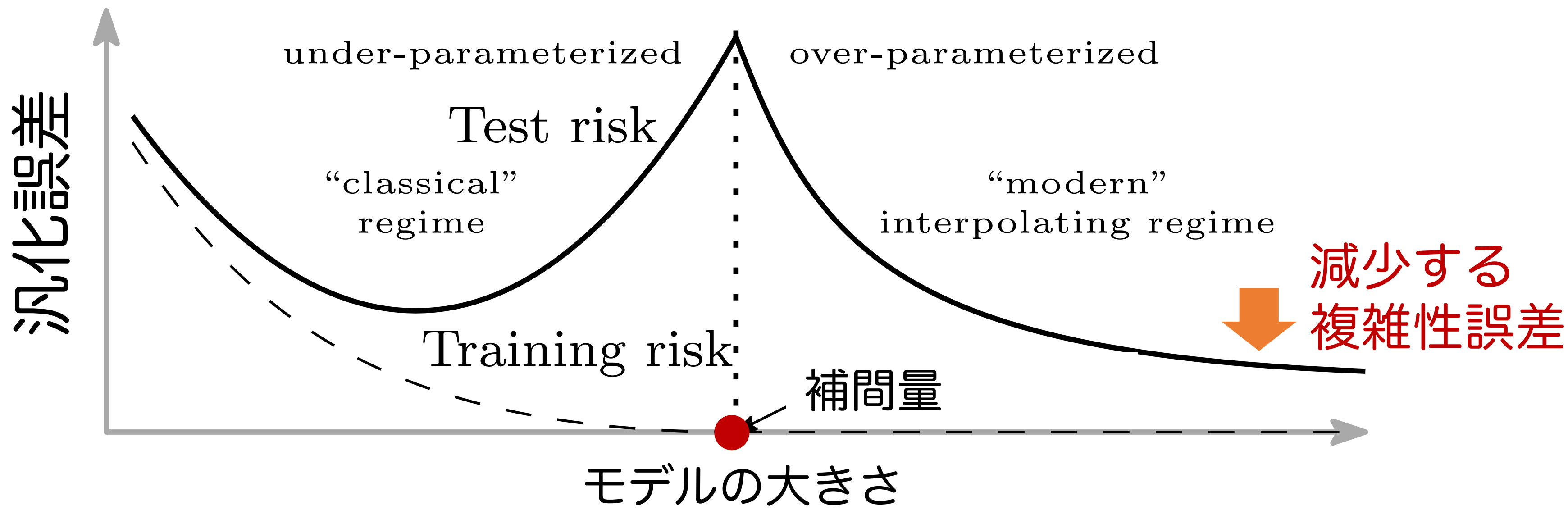
Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS*.

二重降下という考え

二重降下 (double descent)
モデルを過剰に大きくすると、
複雑性誤差（誤差のバリアンス）が減少する現象

二重降下現象

図はBelkin+ 2019より

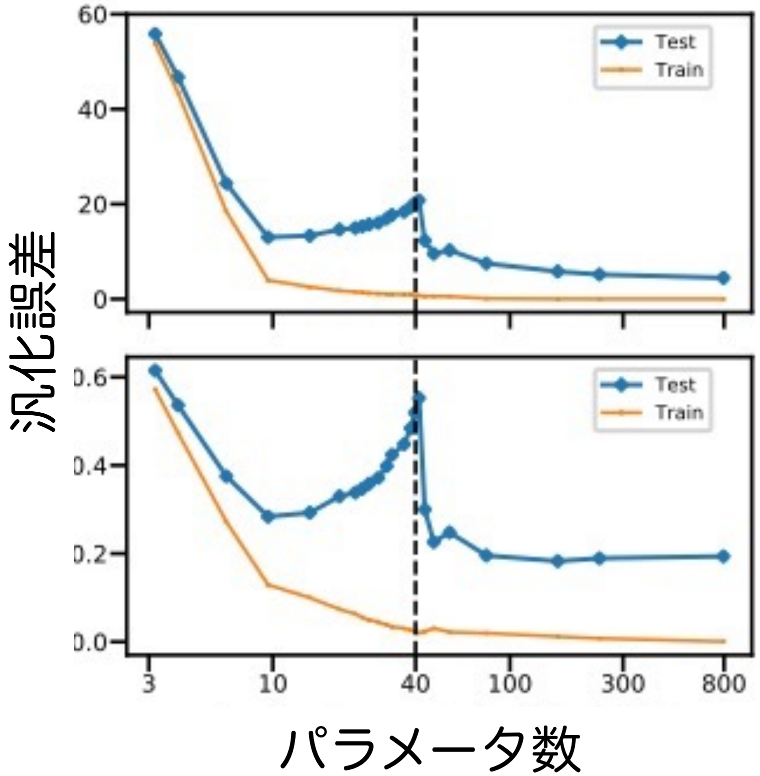


Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS.*

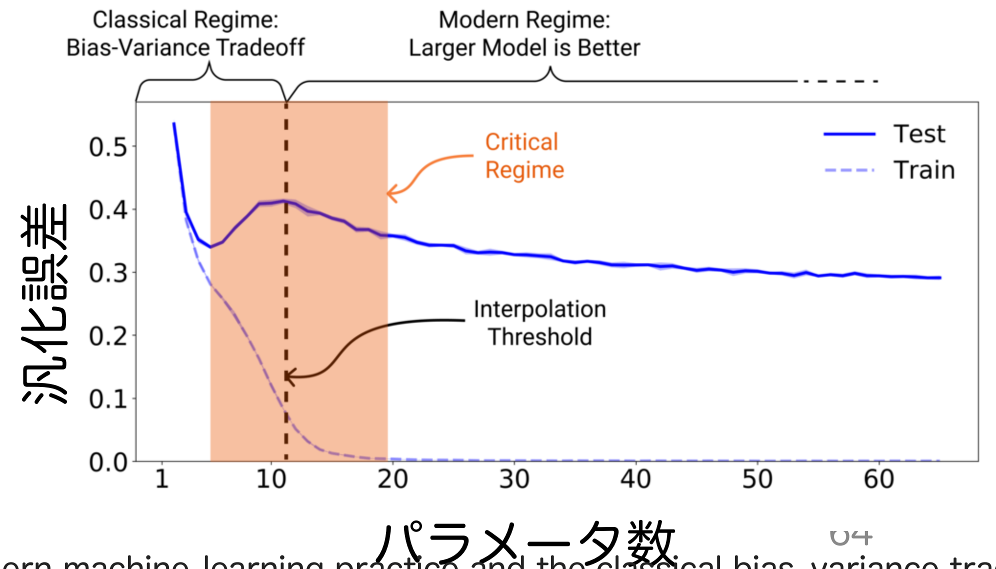
実験による発見

二重降下現象

- シンプルな手法で確認（線形回帰や二層NN）
  - パラメタを増やすと誤差が増加ののち減少 (Belkin+ 2019)



- その後、深層学習でも確認
  - 多層のCNN, ResNetなどで結果が再現 (Nakkiran+ 2020)



Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS*.

これを理論で説明できるか？

<u>線形モデル・2層NNなら可能</u>

線形回帰(単純化)の設定

- 訓練データ $D_n = \{(x_i, y_i)\}_{i=1}^{n}$, x_iはp次元ベクトル
- 線形回帰モデル

$$y_i = {\beta^*}^\top x_i + \varepsilon_i, \quad \beta^* \text{は} p \text{次元パラメータ}$$

補間量 $\hat{\beta} = \operatorname{argmin}\{\|\beta\|_2 : \beta \text{は} \Sigma_{i=1}^{n}(y_i - \beta^\top x_i)^2 \text{を最小化}\}$

線形回帰の汎化誤差

$$\|\beta\|_\Sigma^2 = \beta^\top \Sigma \beta$$

- Σ: X_iの分散共分散行列（$\Sigma = \mathrm{E}[X_i X_i^\top]$）

$$R(\hat{\beta}) = E_\varepsilon\left[\|\hat{\beta} - \beta^*\|_\Sigma^2\right] = \underbrace{\|E_\varepsilon[\hat{\beta}] - \beta^*\|_\Sigma^2}_{= \text{バイアス} B \ (\text{近似誤差})} + \underbrace{\mathrm{tr}[\mathrm{Cov}_\varepsilon(\hat{\beta})\Sigma]}_{= \text{バリアンス} V \ (\approx \text{複雑性誤差})}$$

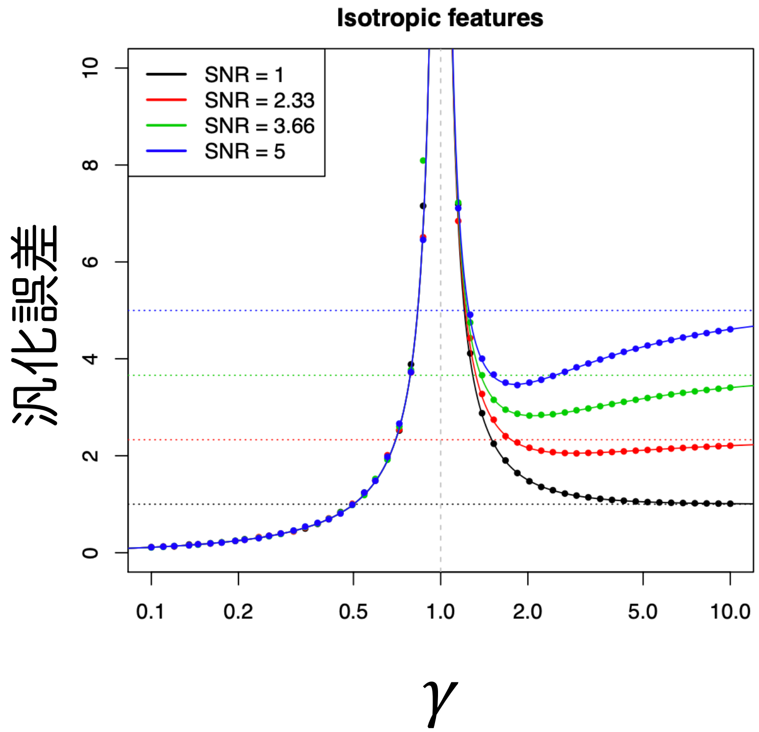
Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *AoS*.

理論による二重効果の再現

Hastie et al. (2022)

$\gamma = \frac{p\ \text{（パラメータ数）}}{n\ \text{（データ数）}}$, σ^2: ノイズ分散

$$\lim_{p,n\to\infty} R(\hat{\beta}) = \begin{cases} \dfrac{\sigma^2\gamma}{1-\gamma}, & (\gamma < 1) \\ \underbrace{\|\beta^*\|_2^2(1-\gamma^{-1})}_{=\text{バイアス}B\ \text{(近似誤差)}} + \underbrace{\dfrac{\sigma^2}{\gamma-1}}_{=\text{バリアンス}V\ \text{(≈複雑性誤差)}}, & (\gamma > 1) \end{cases}$$



モデルの大きさγが増えると
汎化誤差が増加・減少する様子

✔ パラメータ数が増えると($\gamma \to \infty$) 複雑性誤差が減少

× パラメータが多い場合は($\gamma > 1$)近似誤差は残る

Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Stat.*

理論の中身は？

$V(\hat{\beta})$を経験共分散行列の<u>固有値</u>で書き換え

- $X = (x_1, \ldots, x_n)^\top, Z = X\Sigma^{1/2}$ （$n \times p$行列）
- 経験共分散行列 $\hat{\Sigma} = X^\top X/n$ (ランダム行列)
- $\lambda_j(A)$: 行列Aの$j = 1, \ldots, p$番目に大きい固有値

$$V(\hat{\beta}) = \frac{\sigma^2}{n}\mathrm{tr}\left(\hat{\Sigma}^{-1}\Sigma\right) = \frac{\sigma^2}{n}\sum_{j=1}^{p}\frac{1}{\lambda_j(Z^\top Z/n)}$$

$$= \frac{\sigma^2 p}{n}\int \frac{1}{s}\, dF_{Z^\top Z/n}(s)$$

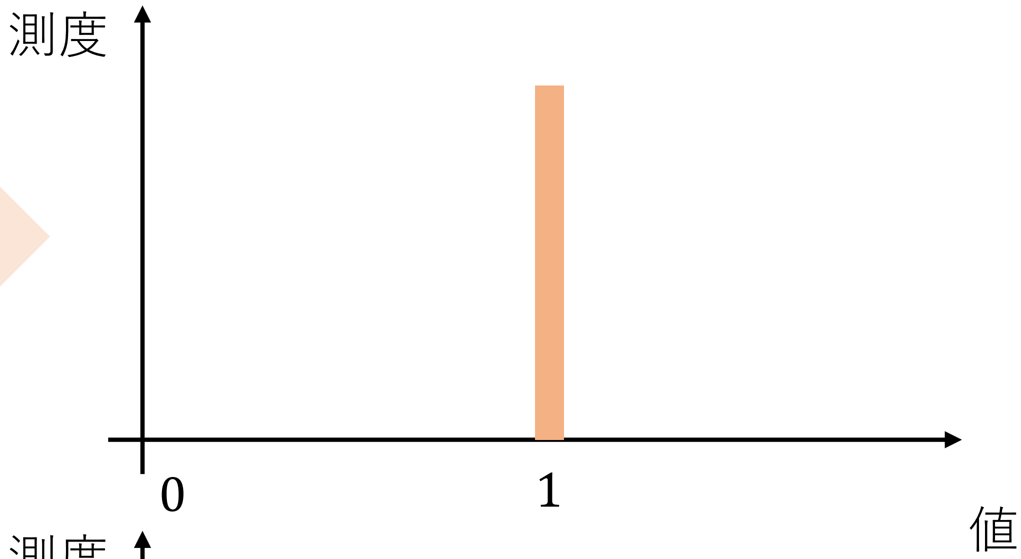
$F_{Z^\top Z/n}(s)$: 行列$Z^\top Z/n$の固有値分布

固有値分布の例

単位行列

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

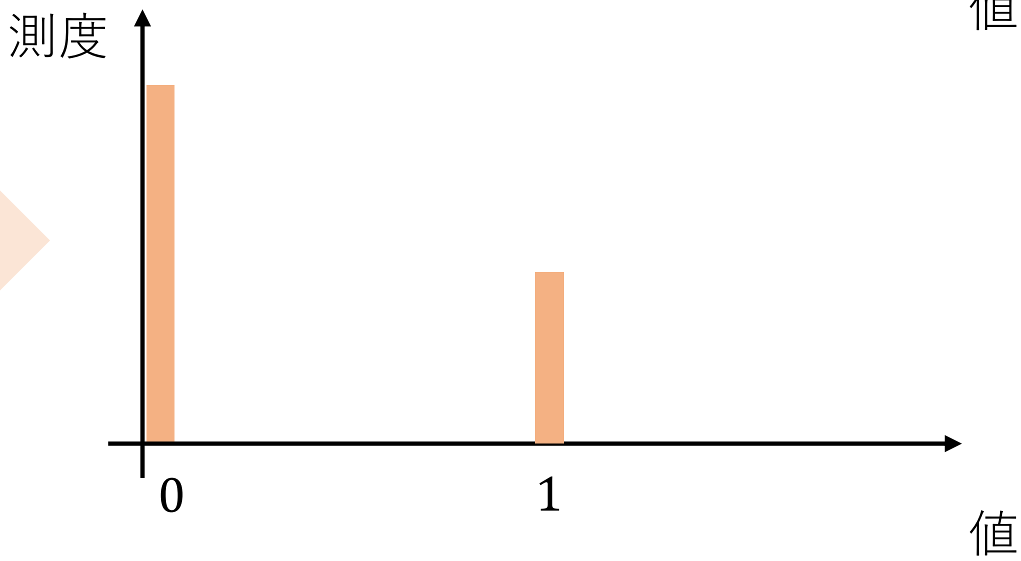
固有値は
1,1,1



特異行列

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

固有値は
1,0,0

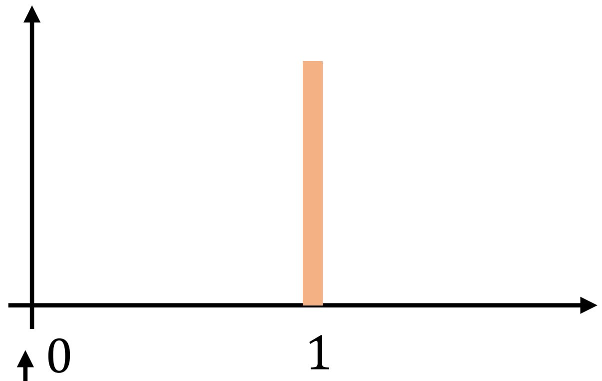


行列が多様な"情報"を持つ時、分布が右に寄る

固有値によるバリアンス評価

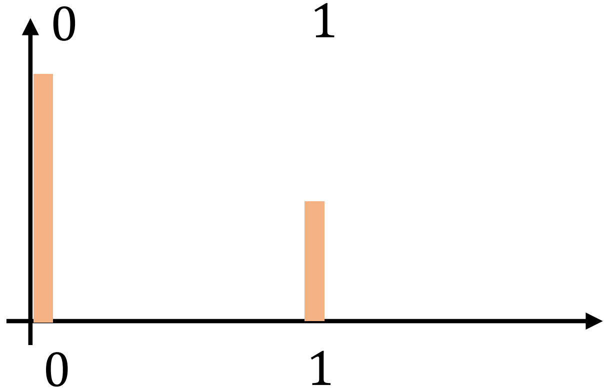
バリアンスは固有値の逆数の和（積分）

$$V(\hat{\beta}) = \frac{\sigma^2 p}{n} \int \frac{1}{s} dF_{Z^\top Z/n}(s)$$



固有値が全て正

→バリアンスは有限



固有値にゼロがある

（例: $p > n$の場合）

→バリアンスが発散

固有値分布が0上にmassを持つかが重要

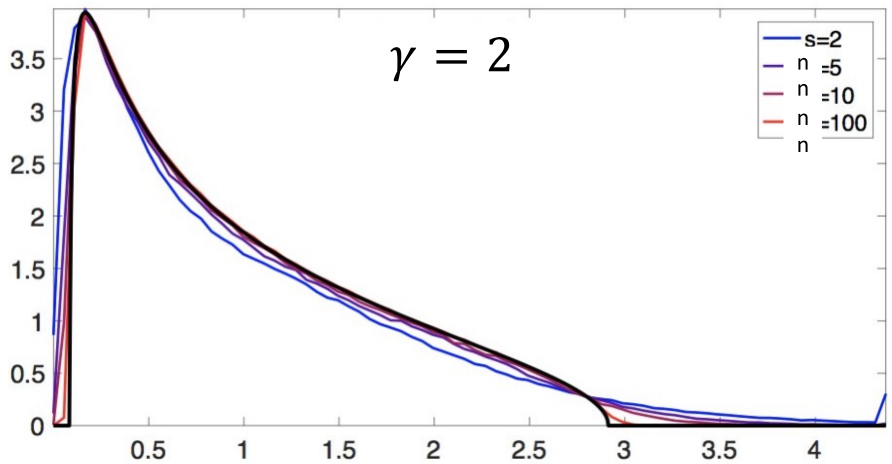
キーとなる固有値分布

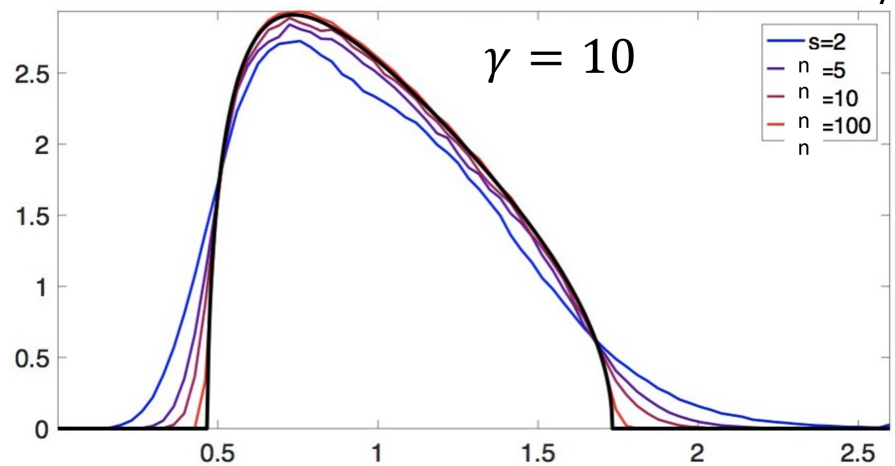
マルチェンコ＝パスツール則 (MP則)

$$\lim_{n,p\to\infty,\ p/n\to\gamma} F_{Z^\top Z/n} = F_\gamma$$

$$dF_\gamma(s) = \frac{\gamma}{2\pi s}\sqrt{(s-s_-)(s_+-s)}1_{[s_-,s_+]},\ s_\pm = \left(1 \pm \sqrt{1/\gamma}\right)^2$$

Peyre(2020)





<u>パラメタ比(γ)が増えると固有値分布がゼロから遠ざかる</u>

($p, n \to \infty$の時、 $p > n$由来のゼロ固有値の影響がなくなる)

Peyre(2020): https://twitter.com/gabrielpeyre/status/928907946108571649

ニューラルネットワークは？

- 限定的な2層ニューラルネットなら理論が通用

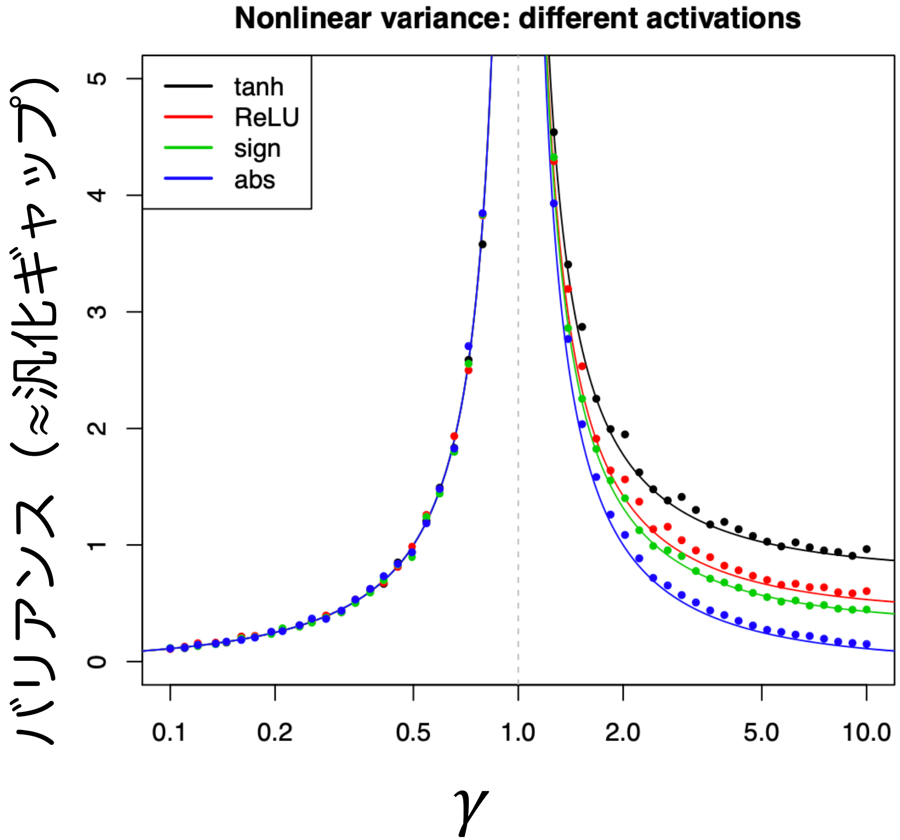
考える二層NN

$$f(x) = \sum_{j=1}^{N} w_j \sigma(a_j^\top x + b)$$

- 一層目：a_j, bは乱数(学習しない)
- 二層目：勾配降下法で学習する

→ **擬似的な線形回帰**

ランダム行列理論を使うため線形回帰に近い設定に持っていくのが大事



モデルの大きさγが増えて、誤差が増加・減少する様子

良性過適合

設定：線形回帰

線形回帰

- 訓練データ $D_n = \{(X_i, Y_i)\}_{i=1}^n$, X_iはd次元ベクトル
- $\phi(X) = (\phi_1(X), \dots, \phi_p(X))$: p次元特徴ベクトル写像
- Σ: $\phi(X_i)$の分散共分散行列（$\Sigma = \mathrm{E}[\phi(X_i)\phi(X_i)^\top]$）

$$Y_i = {\beta^*}^\top \phi(X_i) + \varepsilon_i, \quad \beta^* \text{は} p \text{次元パラメータ}$$

補間量

$$\hat{\beta} = \mathrm{argmin}\left\{\|\beta\|_2 : \beta \text{は} \Sigma_{i=1}^n \left(Y_i - \beta^\top \phi(X_i)\right)^2 \text{を最小化}\right\}$$

汎化誤差（の増加分）

- 予測時の損失の増加分

$$\|\beta\|_\Sigma^2 = \beta^\top \Sigma \beta$$

$$R(\hat{\beta}) = E_{Y,X}\left[\left(Y - \phi(X)^\top \hat{\beta}\right)^2 - (Y - \phi(X)^\top \beta^*)^2\right]$$

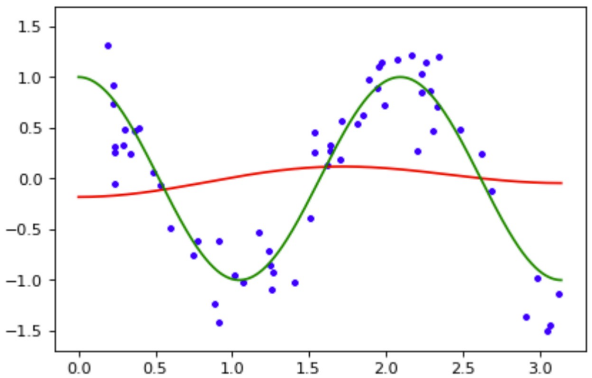
Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *AoS*.

良性過適合という概念

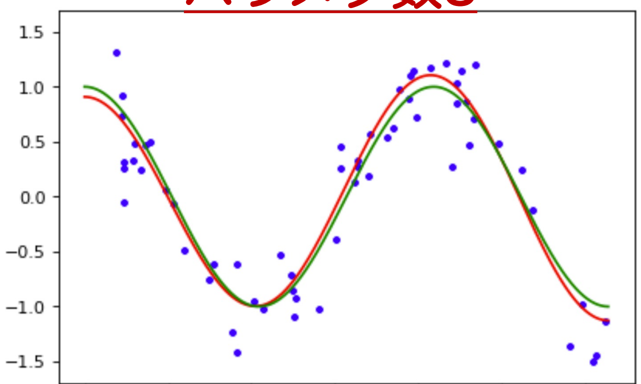
<u>良性過適合 (benign overfitting)</u>
大規模モデルは訓練データへの適合と高い予測性能を両立

緑：真の関数f^*、青： f^* から生成したデータ($n = 60$)、赤：推定した関数

パラメタ数2

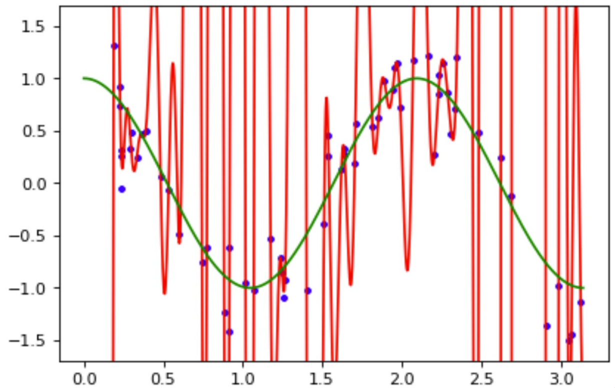


パラメタ数3



適合 ⇒

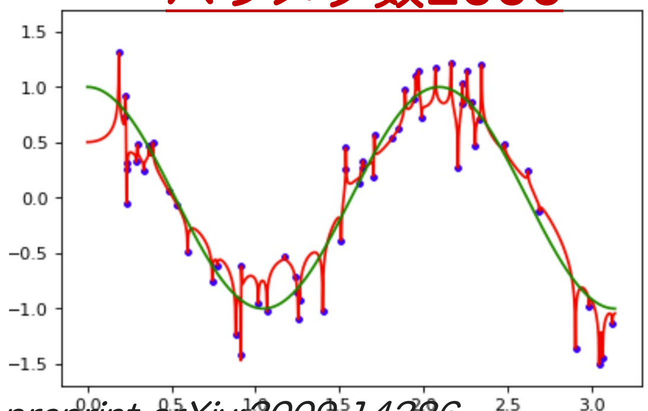
パラメタ数50



⇐ 過適合

⇓ 良性過適合

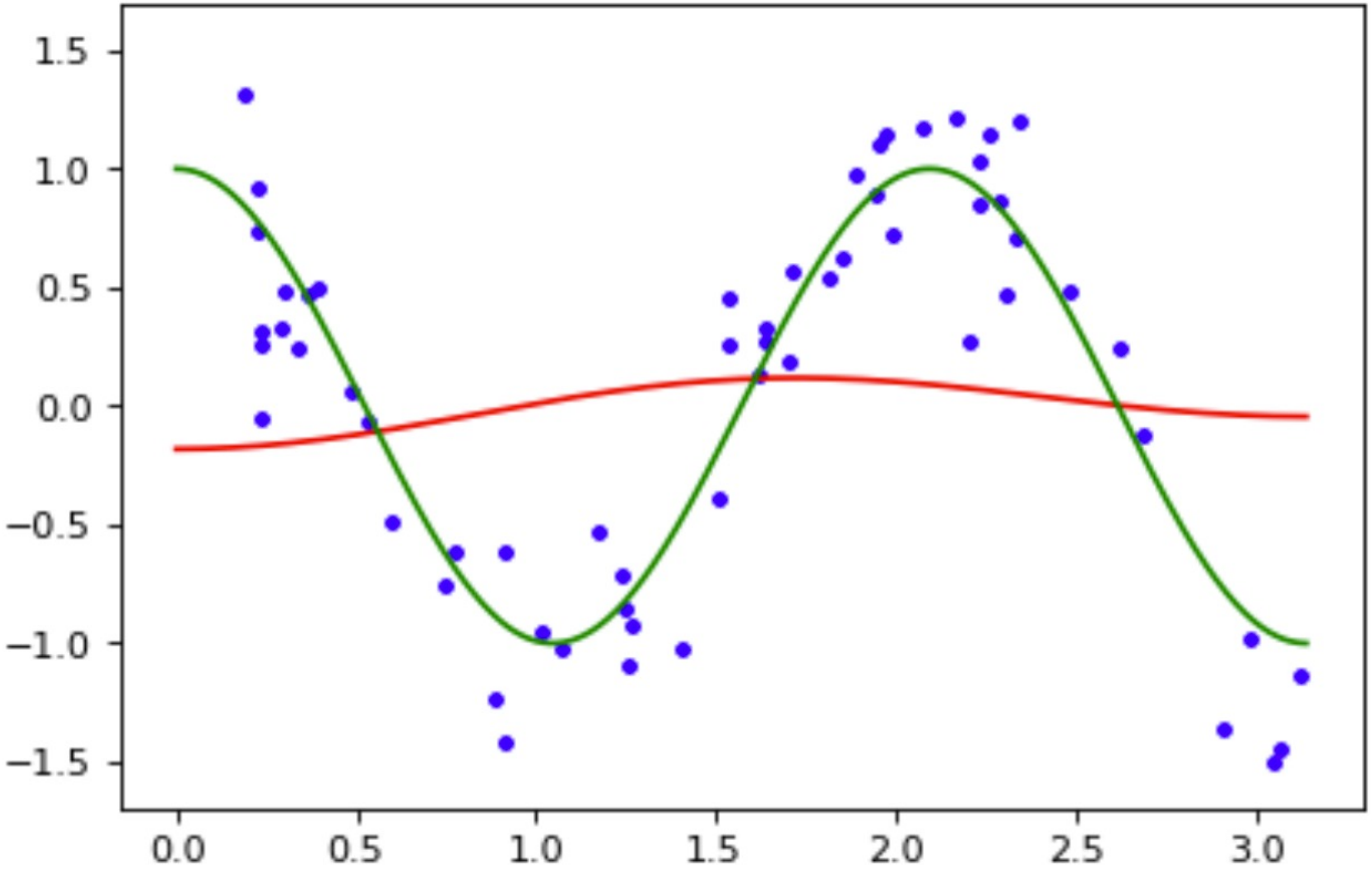
パラメタ数2000



Tsigler, A., & Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

良性過適合の様子

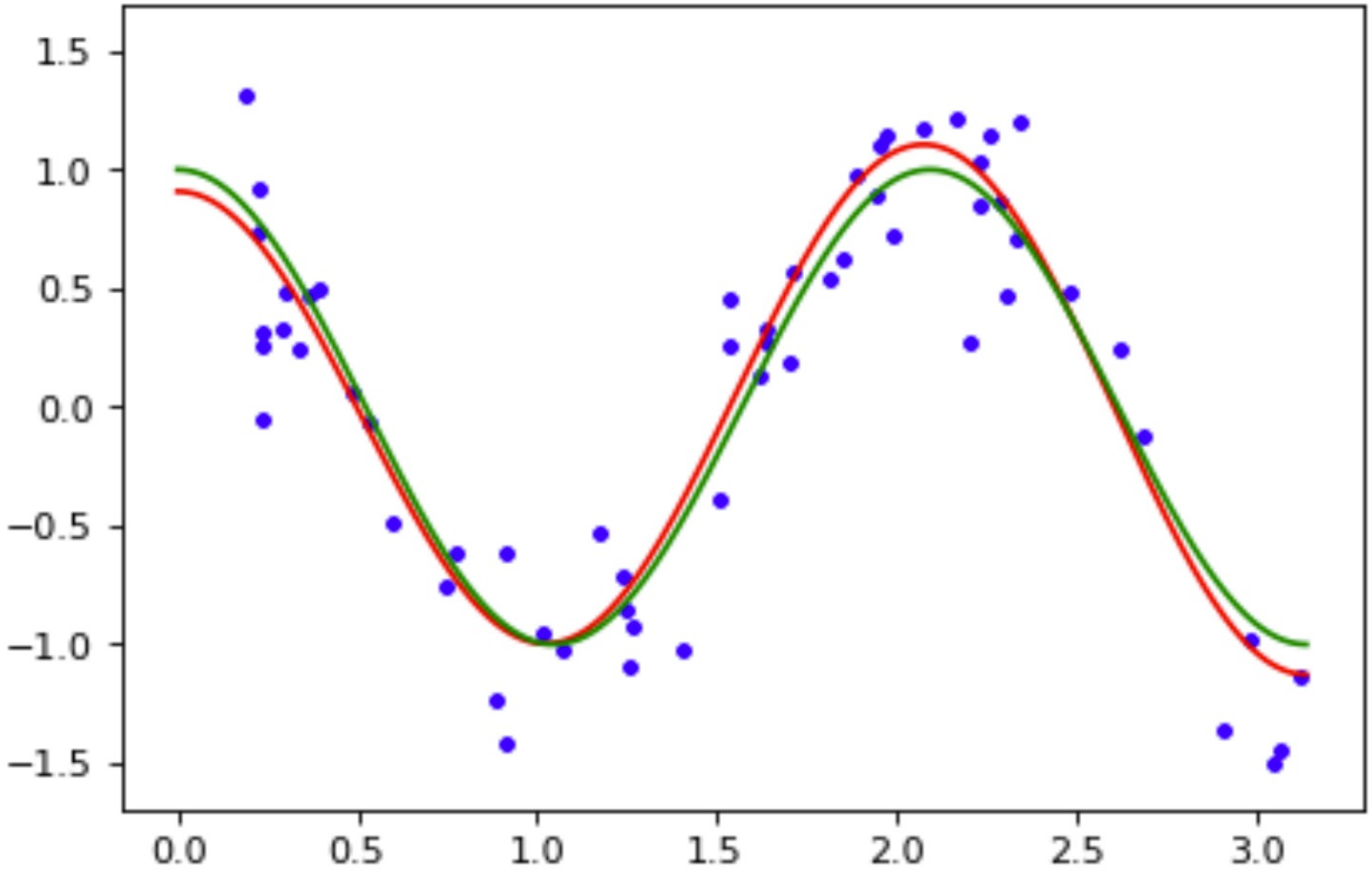
データ数:60 <u>パラメタ数2</u>



緑：真の関数、青：データ、赤：学習した関数

良性過適合の様子

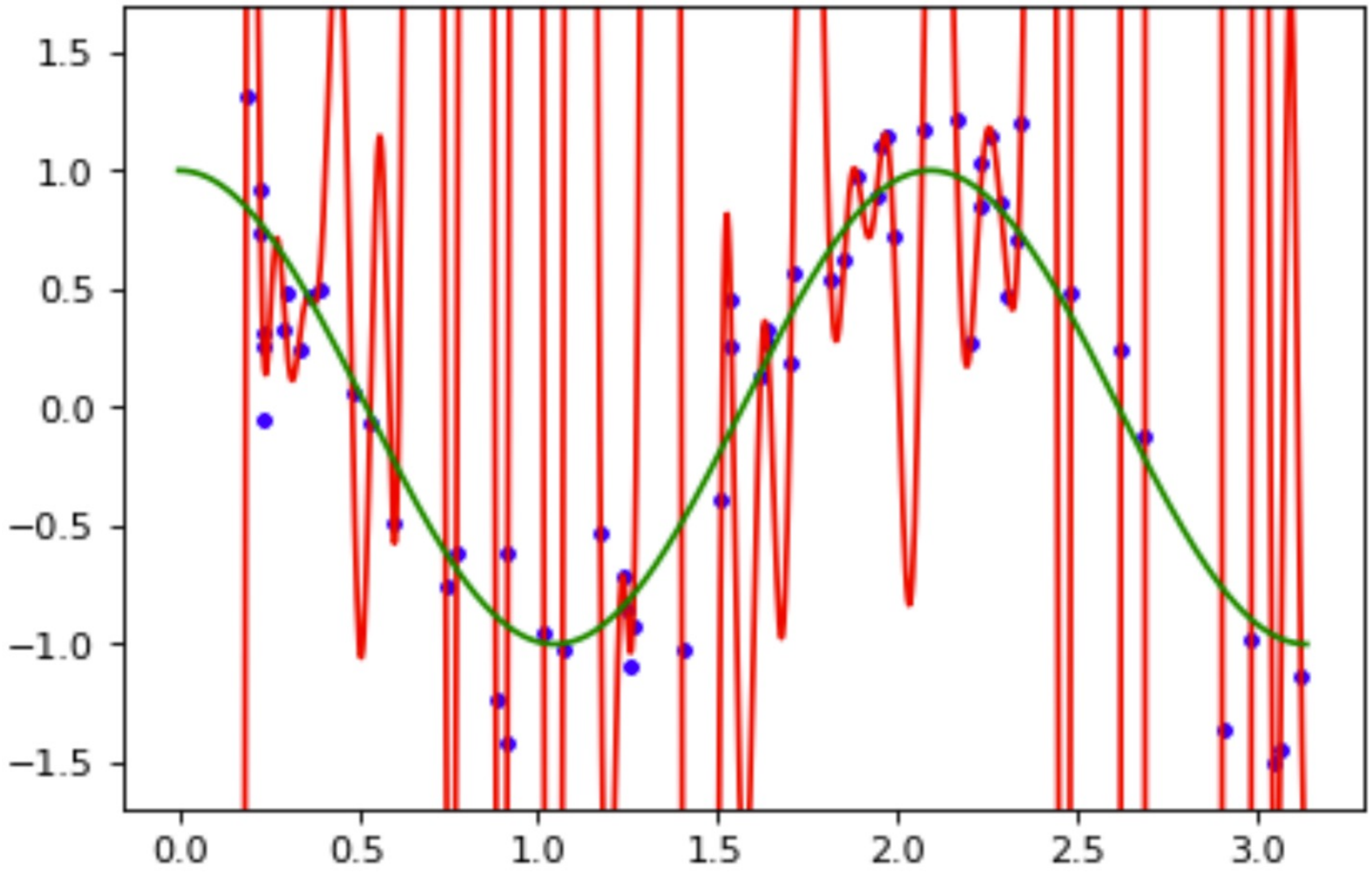
データ数:60 パラメタ数3



緑：真の関数、青：データ、赤：学習した関数

良性過適合の様子

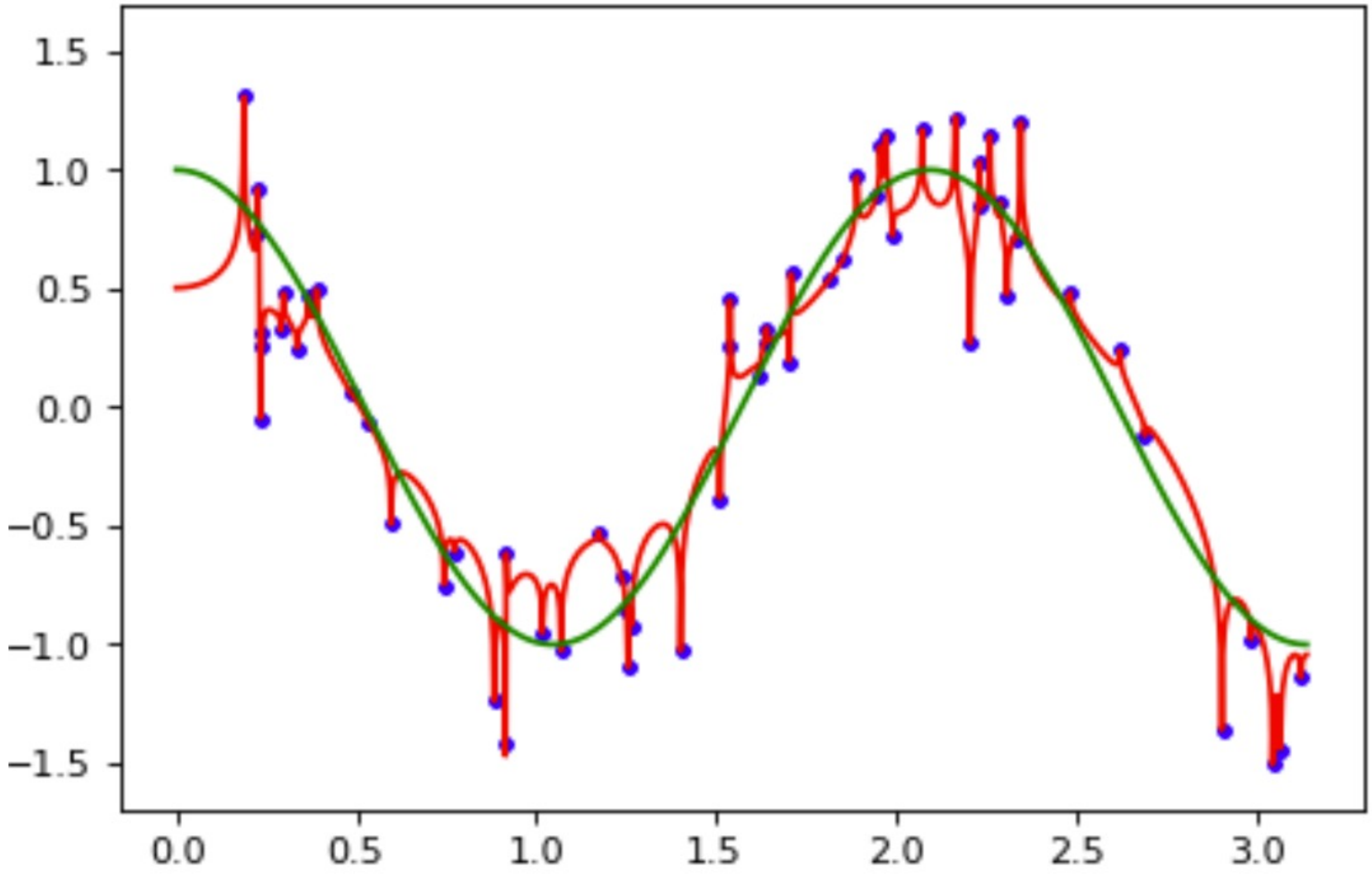
データ数:60 パラメタ数50



緑：真の関数、青：データ、赤：学習した関数

良性過適合の様子

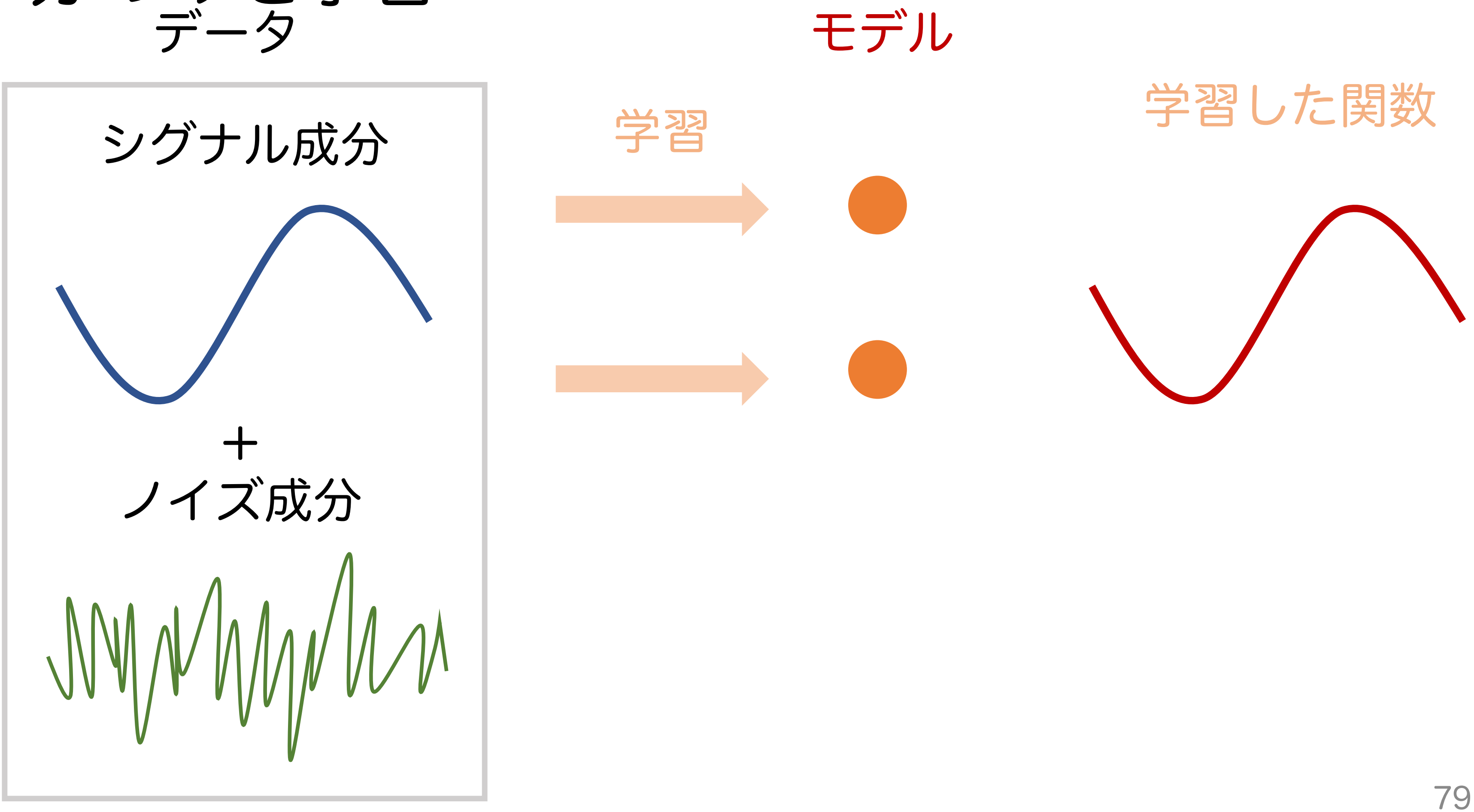
データ数:60 パラメタ数2000



緑：真の関数、青：データ、赤：推定した関数

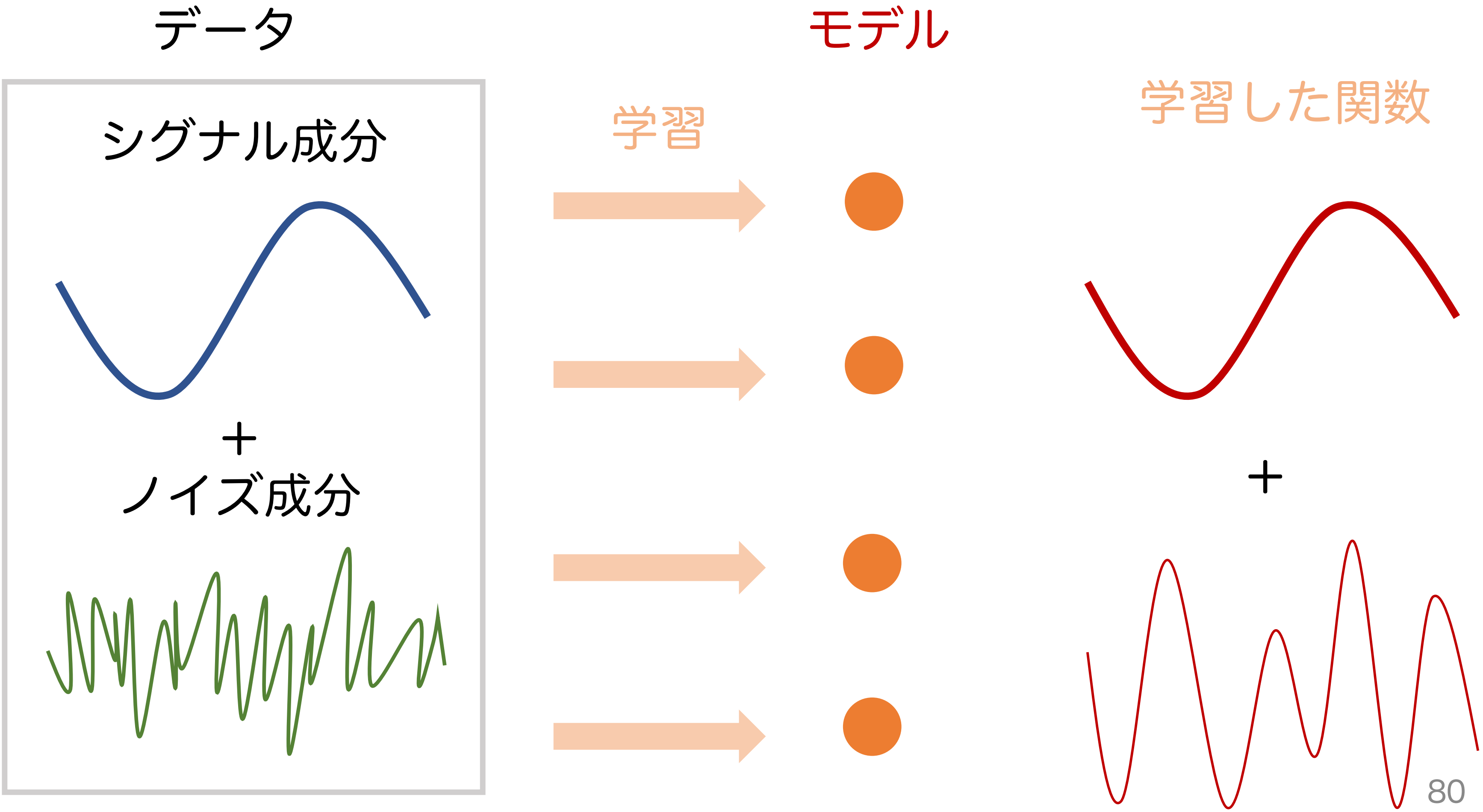
何が起こっている？

- 低次元モデル→少ないパラメータがシグナル成分のみを学習



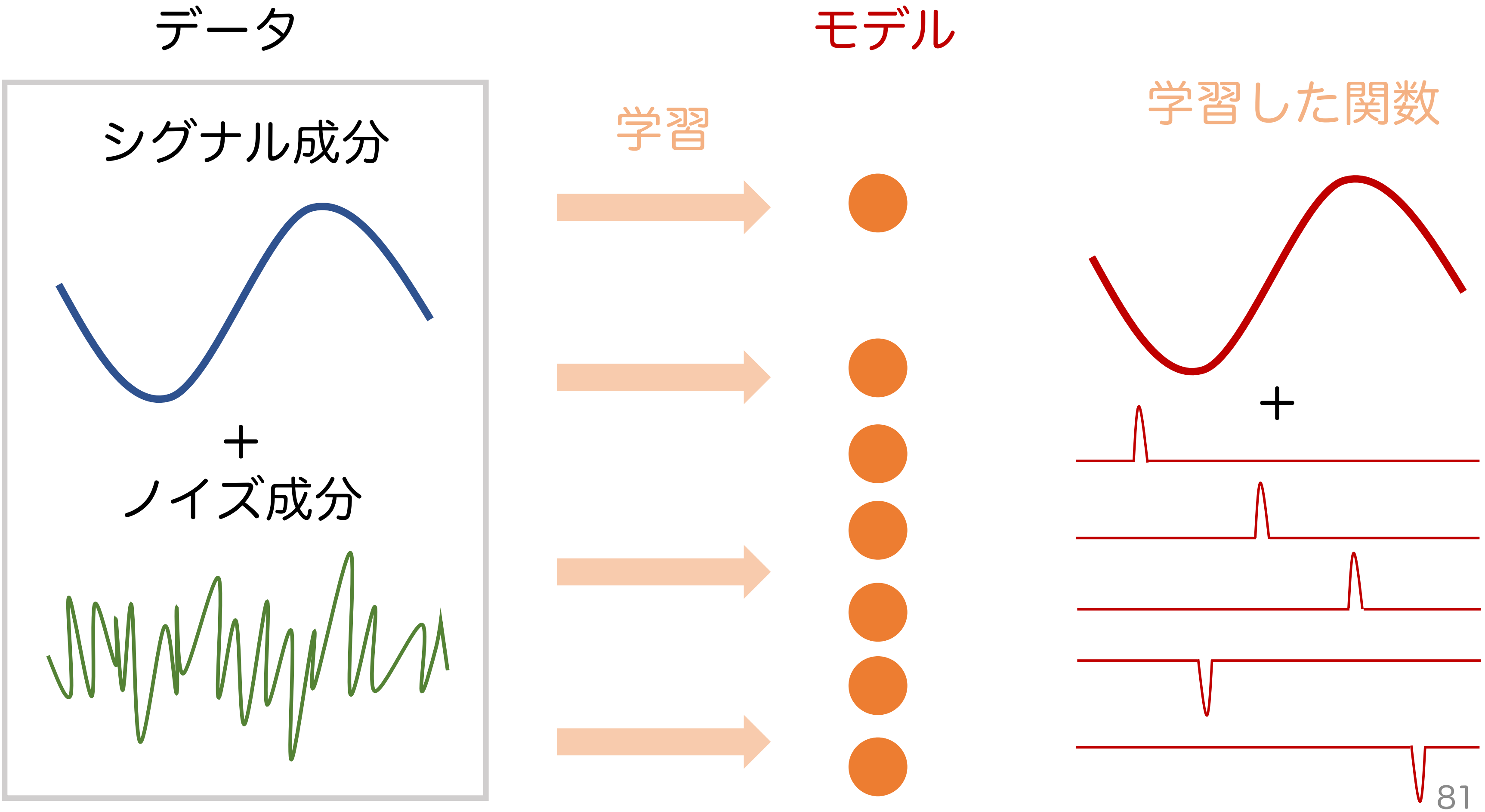
何が起こっている？

- 高次元モデル→余分なパラメータがノイズ成分も学習



何が起こっている？

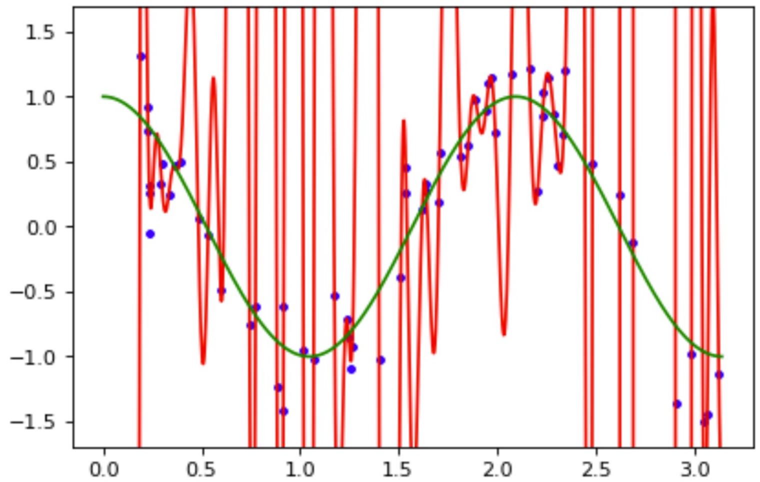
- 過剰パラメータモデル→余分すぎるパラメータ達がノイズを分割



良性過適合という概念

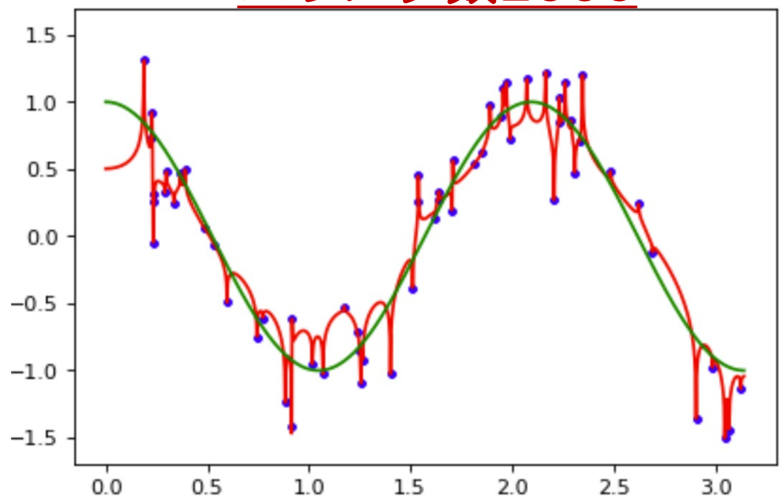
- 過剰パラメータでは、
  モデルは高周波成分でノイズを打ち消す

パラメタ数50



モデルと真の関数の帯域が直交

パラメタ数2000



モデルの高周波帯→データのノイズを補間
モデルの低周波帯→真の関数を学習

高次元における誤差の収束

- $\mathrm{tr}(\Sigma)$: Σのトレース（固有値和）, $R_k(\Sigma)$: Σの実効的なランク

$$\text{線形回帰の汎化誤差} \leq c\left\{\|\theta^*\|_2^2\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \left(\frac{k}{n} + \frac{n}{R_k(\Sigma)}\right)\right\}$$

複雑性誤差のまとめ

- 既存理論との矛盾が大きい
  - 大きなモデルは過学習するという既存理論が否定
  - 深層学習に適した尺度の模索
- 深層モデルのための複雑性誤差の開発
  - 暗黙的正則化、PAC-Bayesなど
  - ただ限界点も多い
- 線形モデルのための過剰パラメータ理論
  - 二重降下・良性過適合などの新現象の発見
  - モデルに制約は大きいが理論が発展

まとめ

深層学習理論

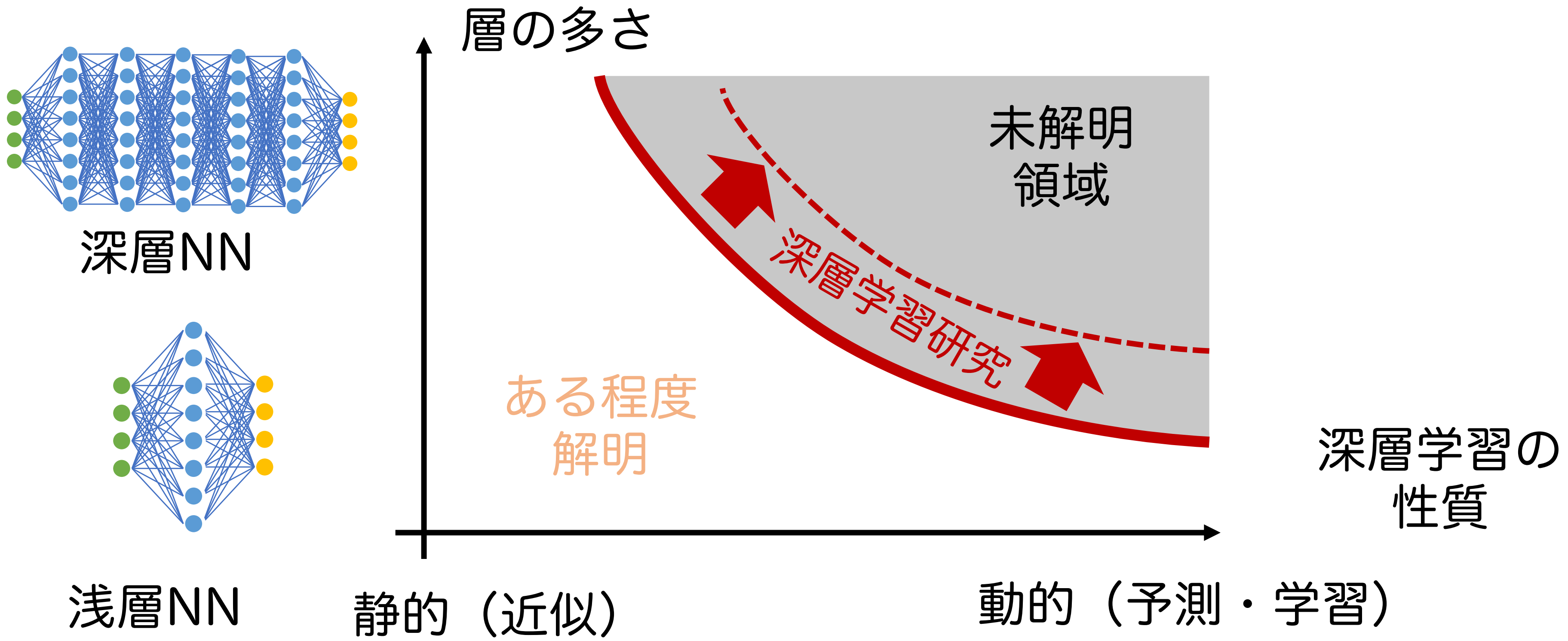
- 近似誤差
  - 多くの洗練された解析→層の役割の解明
- 複雑性誤差
  - 層が多いモデルは未解明点が多い
  - 層が少ない過剰パラメータ理論は発展著しい
    →多層化は今後の課題

人工知能理論

- 近年の発展に応じて解析が必要
- 文脈内学習が一つの突破口

最後に

- 深層モデルの動的な性質ほど解析が難しい



多くの未解決問題・伸び盛りの研究分野
→ 深層学習・人工知能の基礎理論の開発へ