深層学習・人工知能の理解に向けた数学・物理学的試み

2025/12/10 東京女子大セミナー

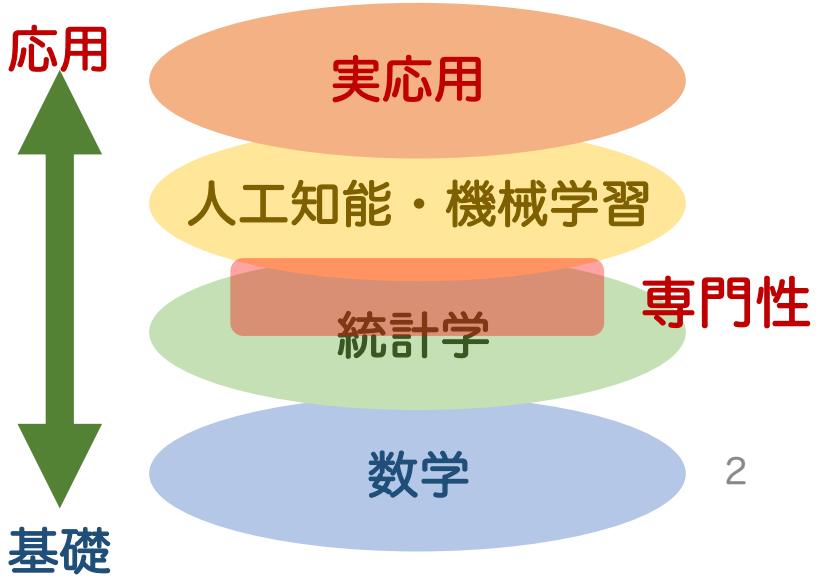
今泉允聡
(東京大学・理研・京大)

今泉允聡
(いまいずみ まさあき)

- ~2017 東京大学 経済院 統計学コース (博士)
- ~2020 統計数理研究所 (PD, 助教)
- 現職：東京大学 相関基礎 物性物理G (准教授)
  (兼) 理化学研究所 革新知能統合研究C (チームディレクター)
  (兼) 京都大学 理学研究科 物理・宇宙物理学専攻 (特定准教授)

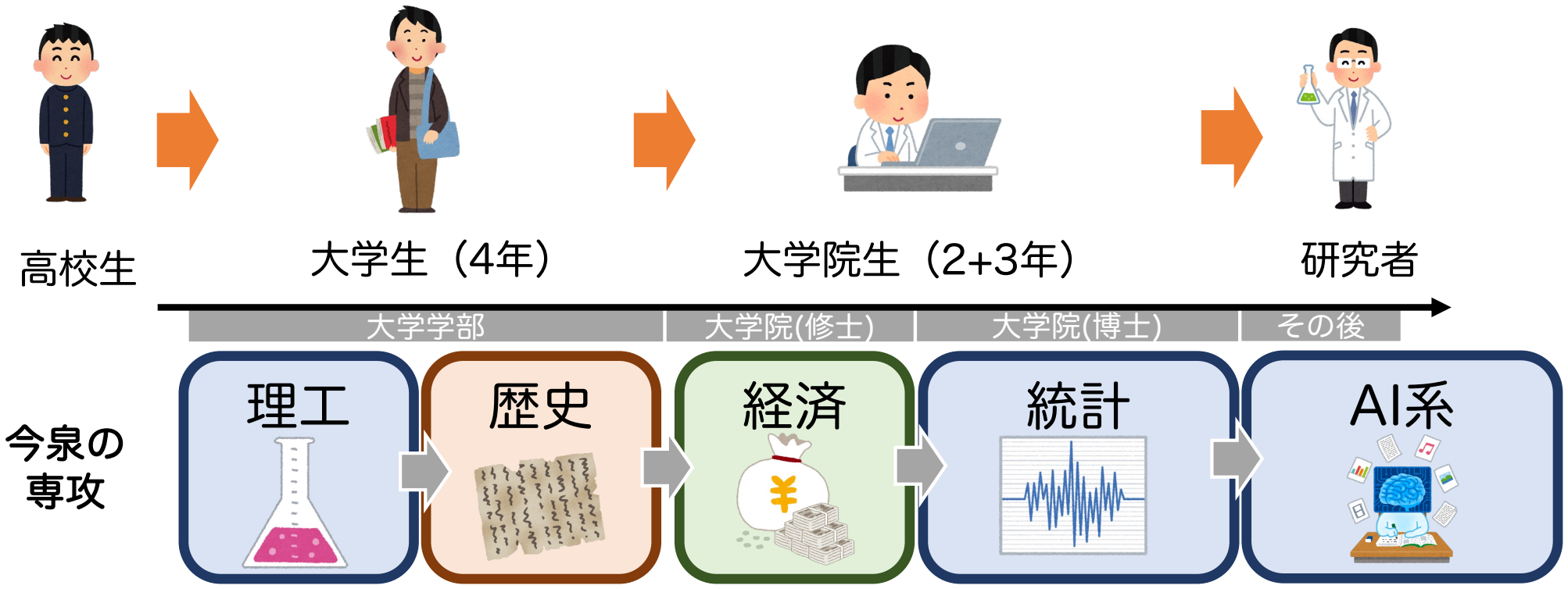
研究関心

- 統計学・機械学習
  - 無限次元統計・深層学習理論
- 物理学と深層学習
  - 統計物理・複雑系アプローチ



研究者になるまで

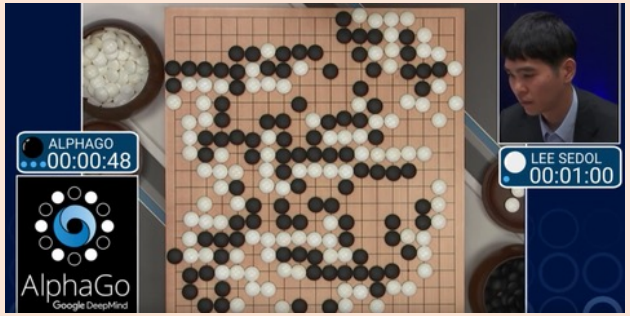
- 研究者になるパターンの典型



AIの登場と深層学習

AI技術

AlphaGo (囲碁AI)



対話できるAI

自動運転

深層学習 (2012年に開発)

AI（人工知能）技術の根幹となる技術

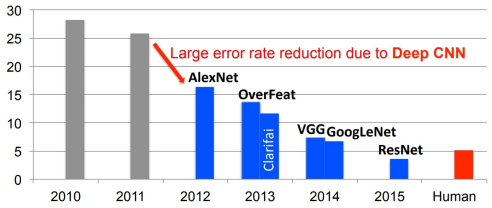
深層学習・AIの発展

基礎研究　深層学習実用化　Transformer登場　ChatGPT公開

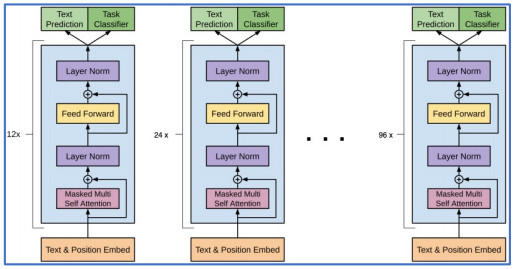
~2000

技術的課題

2012



2017



2022

10Mパラメータ～
画像分類など

200Mパラメータ～
自然言語処理など

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

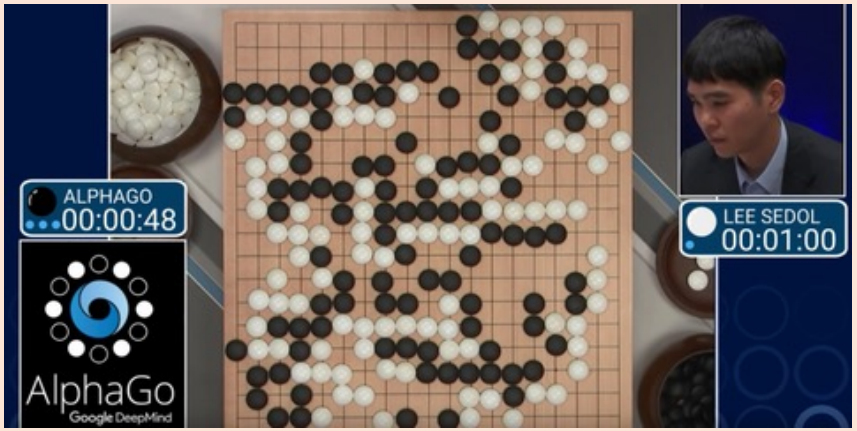
⇒ **原理の解明はまだ発展途上**

内部の解釈や効率的運用に向けて

深層学習・AIの成功例

AlphaGo (DeepMind)

- 囲碁で人間超え
  - 世界トップ棋士に勝利



高精度データ生成

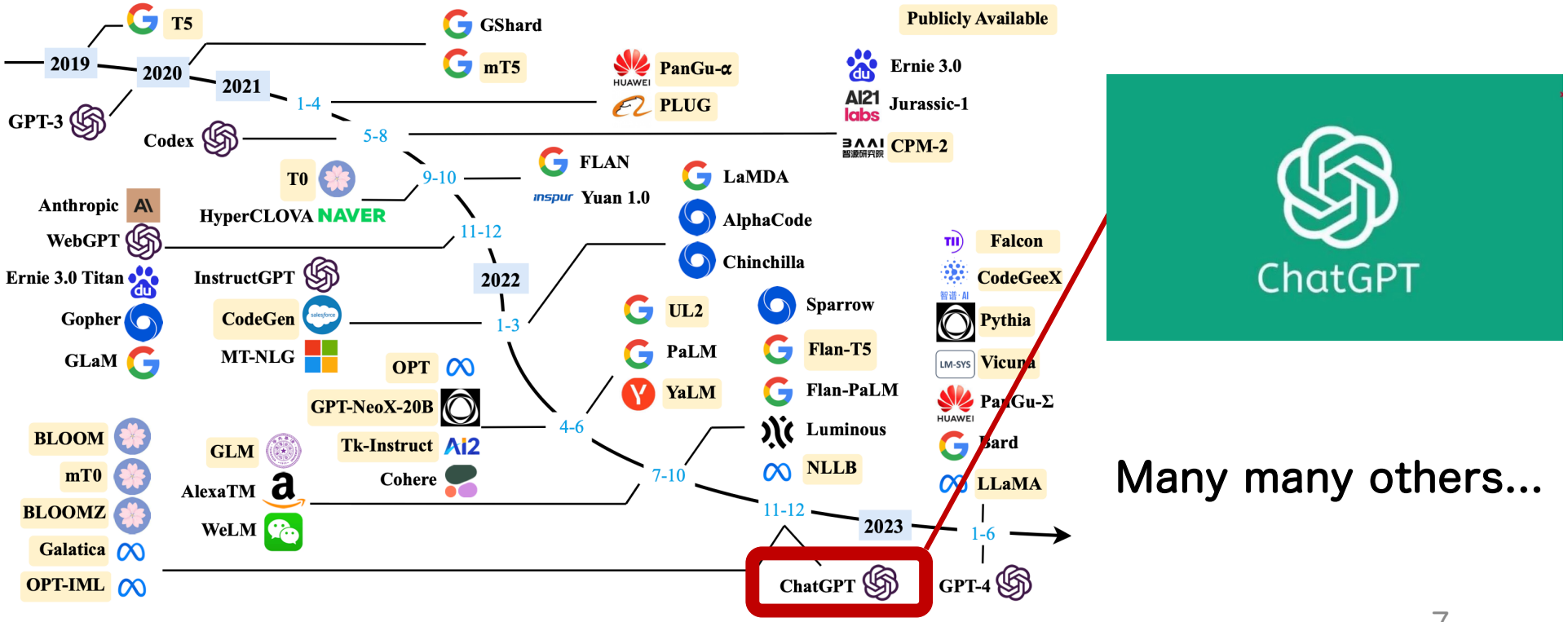
- キーワードから仮想画像の生成

https://stablediffusionweb.com

いくつかのタスクで高い性能を発揮

発展するAI

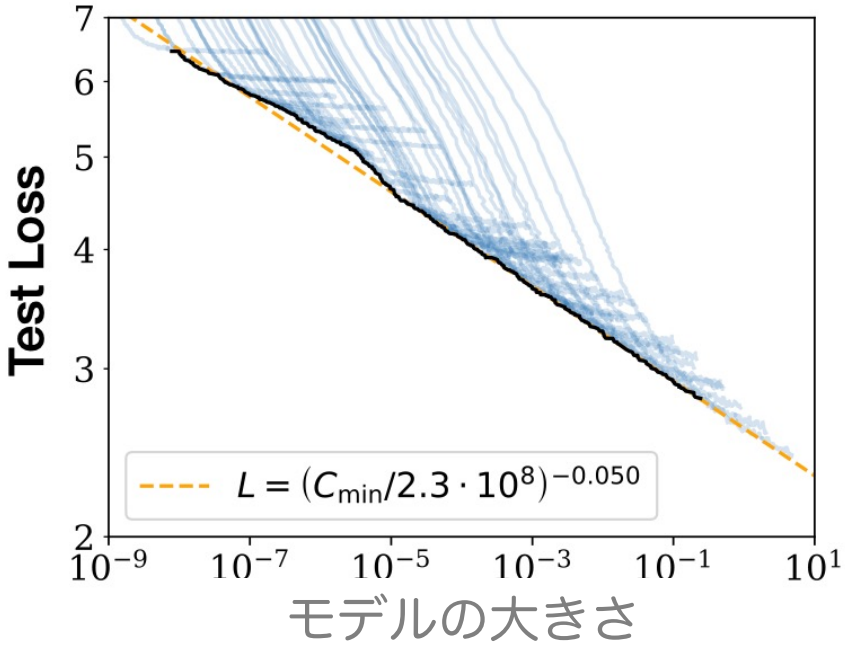
- 汎用的な処理能力を持った学習器
  - 巨大データを用いた事前学習＋FineTune/prompting
  - 例: 大規模言語モデル(Large language model: LLM)



Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

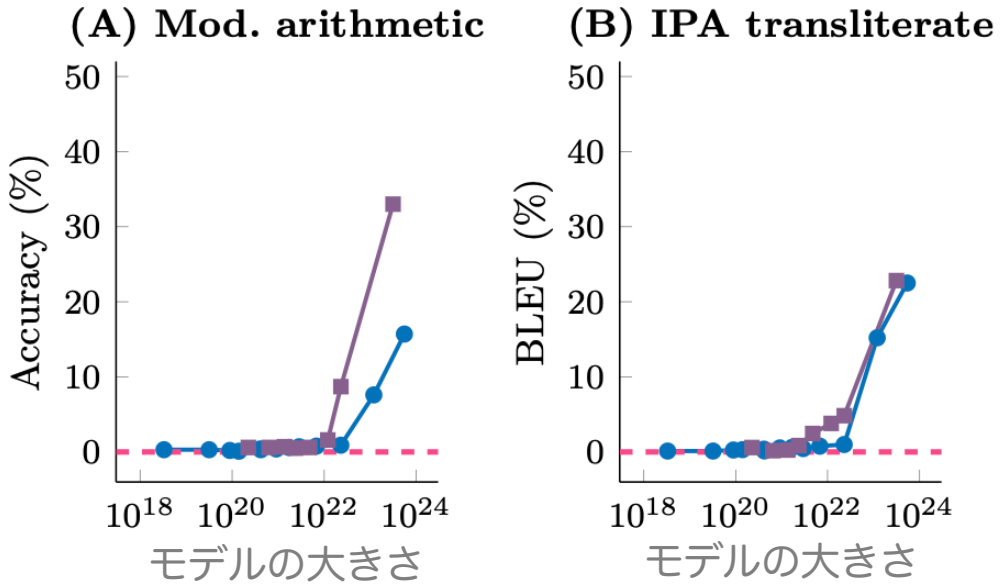
AIの動作で興味を集める新現象

- スケール則



モデルサイズ・データ量が増えると、
損失が単調にベキ則で減少する

- 創発現象



モデルサイズが一定水準を超えると
精度が突然向上する

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. Emergent Abilities of Large Language Models. *TMLR*

理解の不在による壁

深層学習の運用にはまだ問題点が多い

膨大な計算コスト

うまい設定が
分からない
大量に試験しよう

計算がとても大変

ブラックボックスな挙動

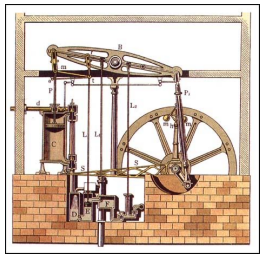
失敗したが
原因は不明！

信頼できる
製品が作れない

実用化の進展には、**原理の理解**が必要

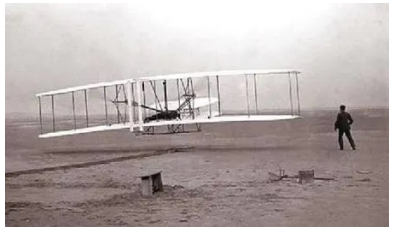
“発見”を理論で記述すること

- 歴史的には共通の現象



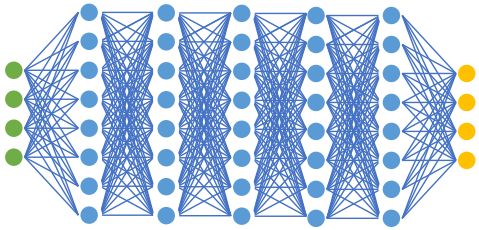
蒸気機関の発明
(1769年)

熱力学の
成立



飛行機の発明
(1903年)

航空力学の
成立

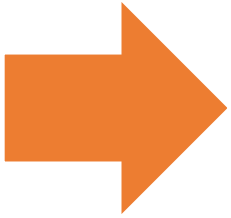


深層学習・
人工知能の発明
(2012年)

？

問：深層学習・AIを理解できる理論は構築できるか？

今日の概要



AIの設計法(深層学習)の謎

数学的な説明と限界

物理学的なアプローチ

深層学習とは

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

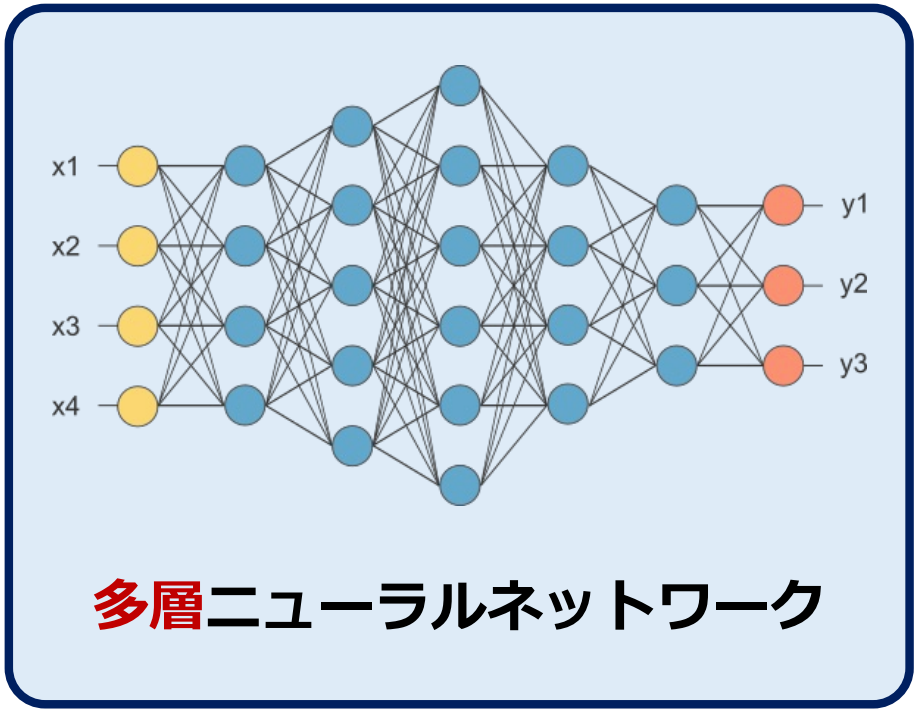
- 入力ベクトルを変換する関数のモデル

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



出力（例：情報）

これは
茶色い猫です

変換

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

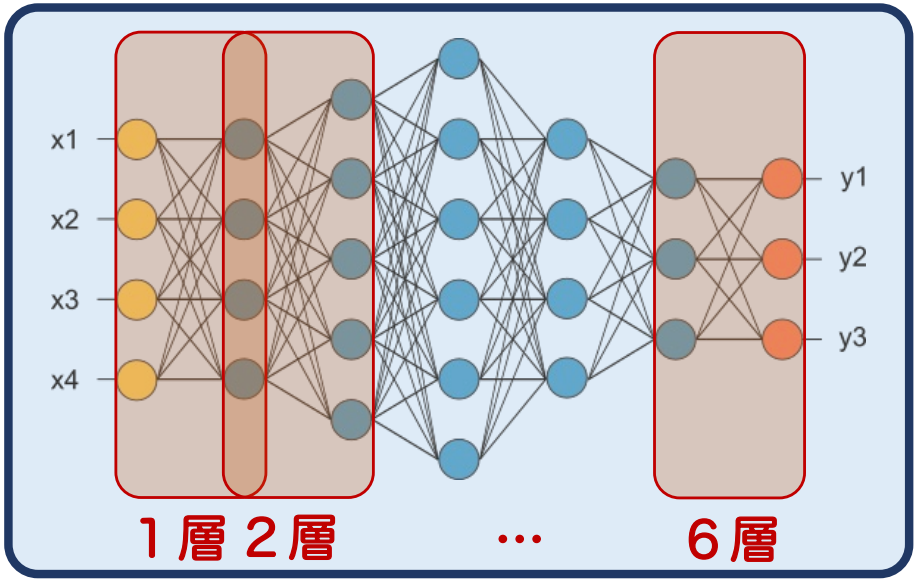
ベクトル y

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

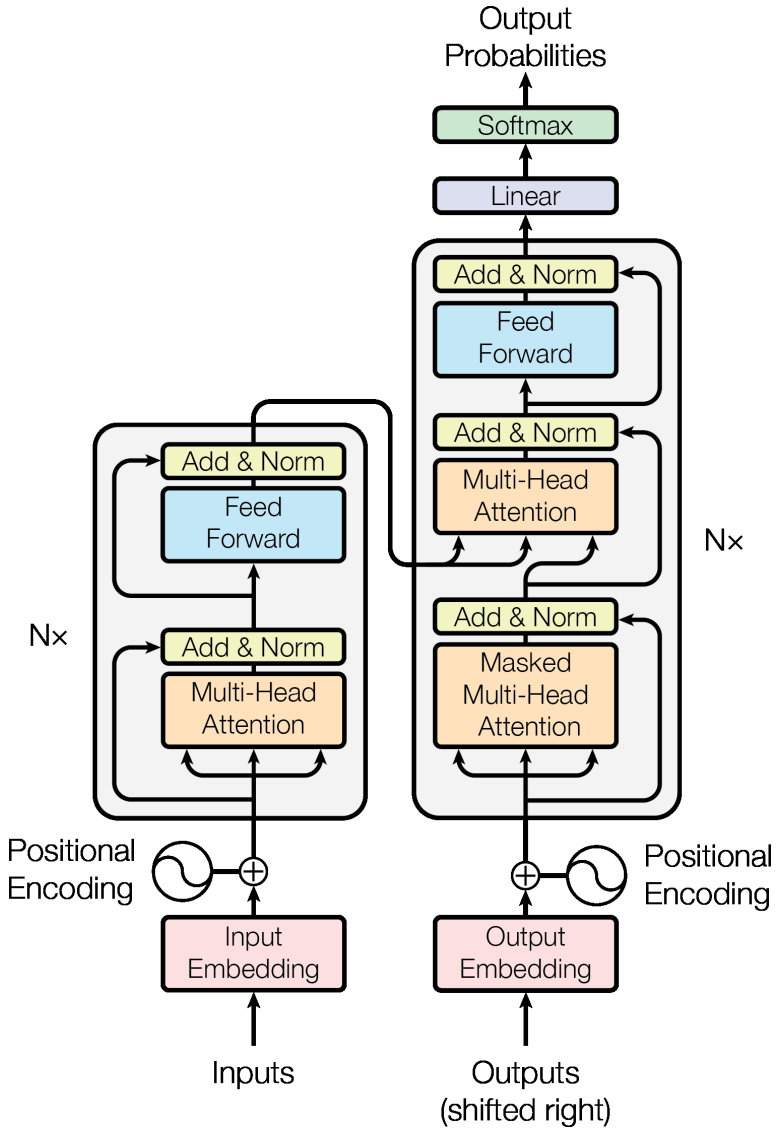
1層目	$z_1 = \sigma(A_1 x + b_1)$
2層目	$z_2 = \sigma(A_2 z_1 + b_2)$
⋮	⋮
6層目	$y = A_6 z_5 + b_6$

A: パラメタ (行列)
b: パラメタ (ベクトル)
σ: 非線型変換

AIへの発展：トランスフォーマー

- アテンション機構を導入したニューラルネットワーク
  - 言語処理や分子構造予測などで非常な高精度を達成

GPT: Generative Pre-trained Transformer



トランスフォーマーのアーキテクチャ

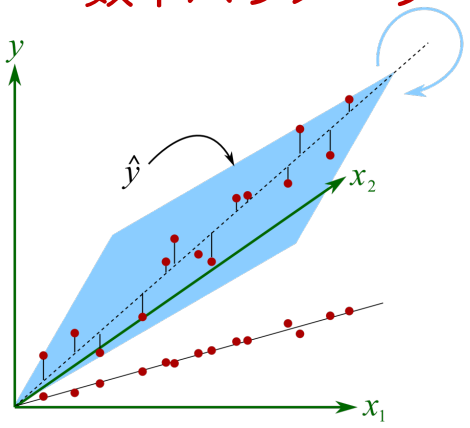
層を増やして巨大化するシステム

- 計算機の発展で数億以上のパラメータを学習可

～2000年

小規模データ解析

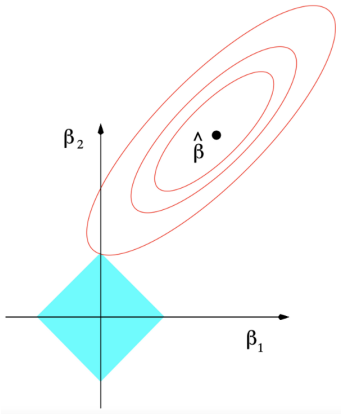
～数十パラメータ



情報量規準・主成分分析など
解釈可能な数次元にデータを縮約する技術

2000年～

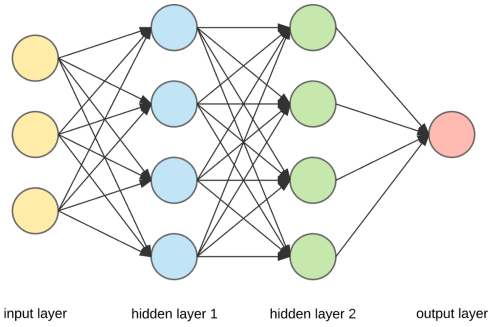
（中程度な）
高次元データ解析
～数千パラメータ



スパース推定・カーネル法など
シンプルな低次元特徴を効率的に発見する技術

2015年～

深層学習
数億以上パラメータ

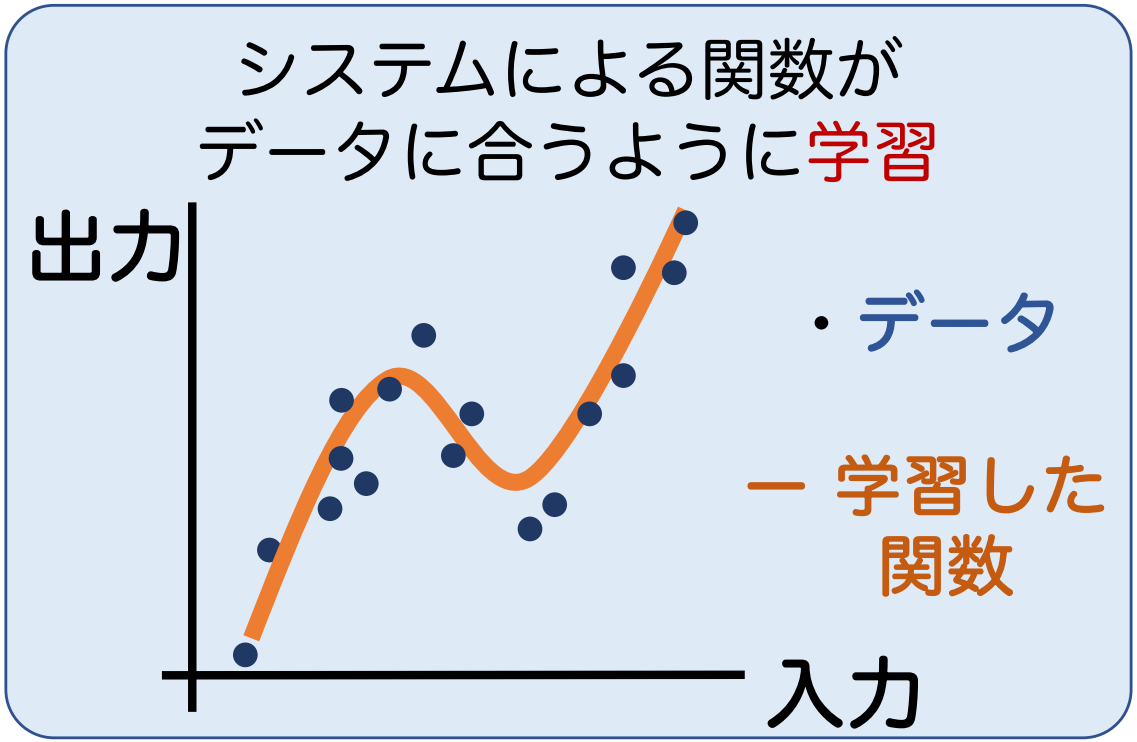


多層ニューラルネットとその拡張
複雑な特徴量を自動的に構成する技術

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

(Y_i, X_i)：データ

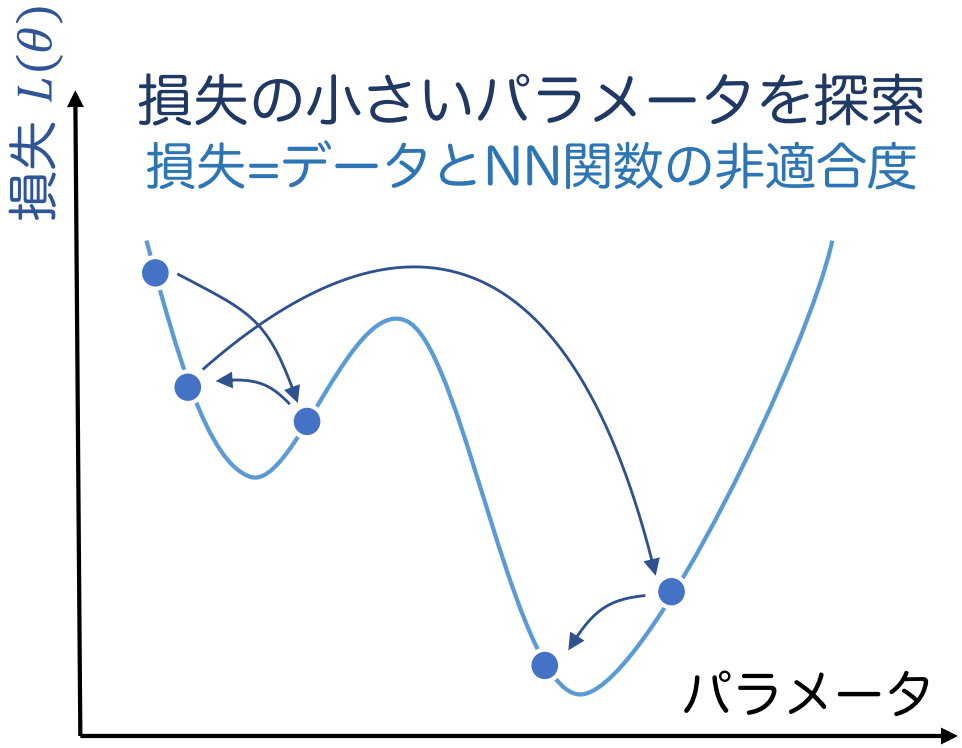
$$\min_{\theta} \sum_i \left(Y_i - f_\theta(X_i)\right)^2$$

損失 $L(\theta)$

＝システムによる関数と
データのズレ

膨大なパラメタはデータから学習

- 学習データに適合するように
  ニューラルネットのパラメータを更新

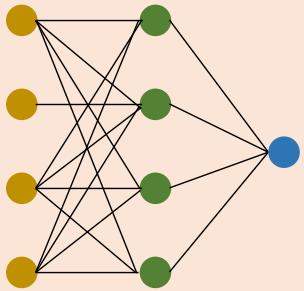
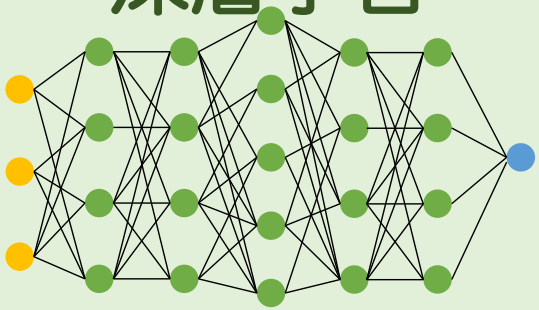


更新アルゴリズム：
勾配降下

$$\theta^t = \theta^{t-1} - \eta \nabla L(\theta^{t-1})$$

深層学習の謎

従来のデータ解析と深層学習の違い

	従来法	深層学習
層の数 (変換の回数)		
	1～2	3～100以上
パラメタの数	数十 ～ 数千	数十万～数億
性能	そこそこ	とても良い

層・パラメータが多いなら、高性能は当たり前？

謎1：多層は不要だと思われていた

理論　層数は少なくて良いと
数学的に証明されているよ

関数推定の最適性定理 (Stone (1980) など)
従来法（１～２層）で理論的に最適。

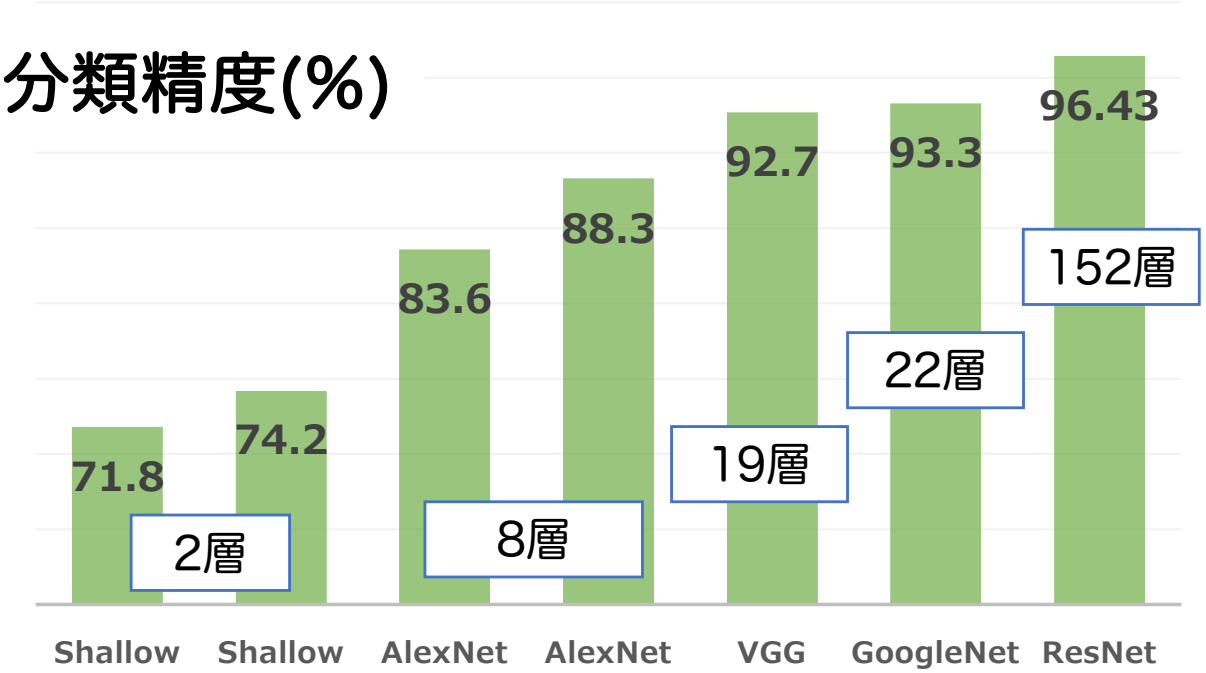
UC Berkley

普遍近似定理 (Cybenko (1989) など)
２層のニューラルネットワークで十分。

Prof. G. Cybenko

謎1：でも多層で性能が上がる

実際
多層にすると性能が向上するよ



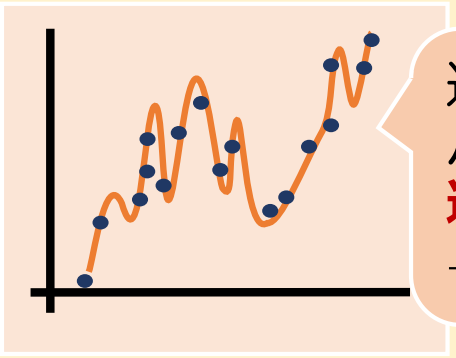
層が増えるほど高い精度を発揮

何層を使えば良い？
仕組みが分からないから
全部試すしかない…

謎2：大規模モデルと過学習

従来理論と深層学習は完全に食い違う

従来のデータ解析理論



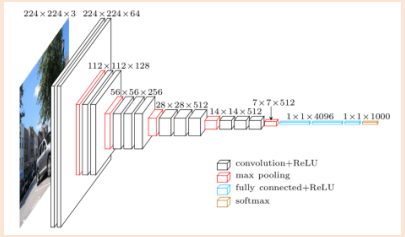
過剰な
パラメタは
過学習する
→性能悪化

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

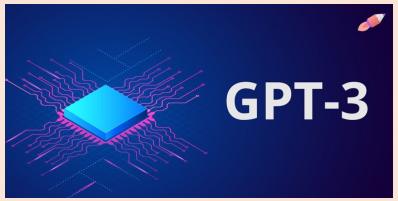
深層学習登場前の**常識**

矛盾

巨大深層学習の成功



VGG19 Net
1億パラメタ



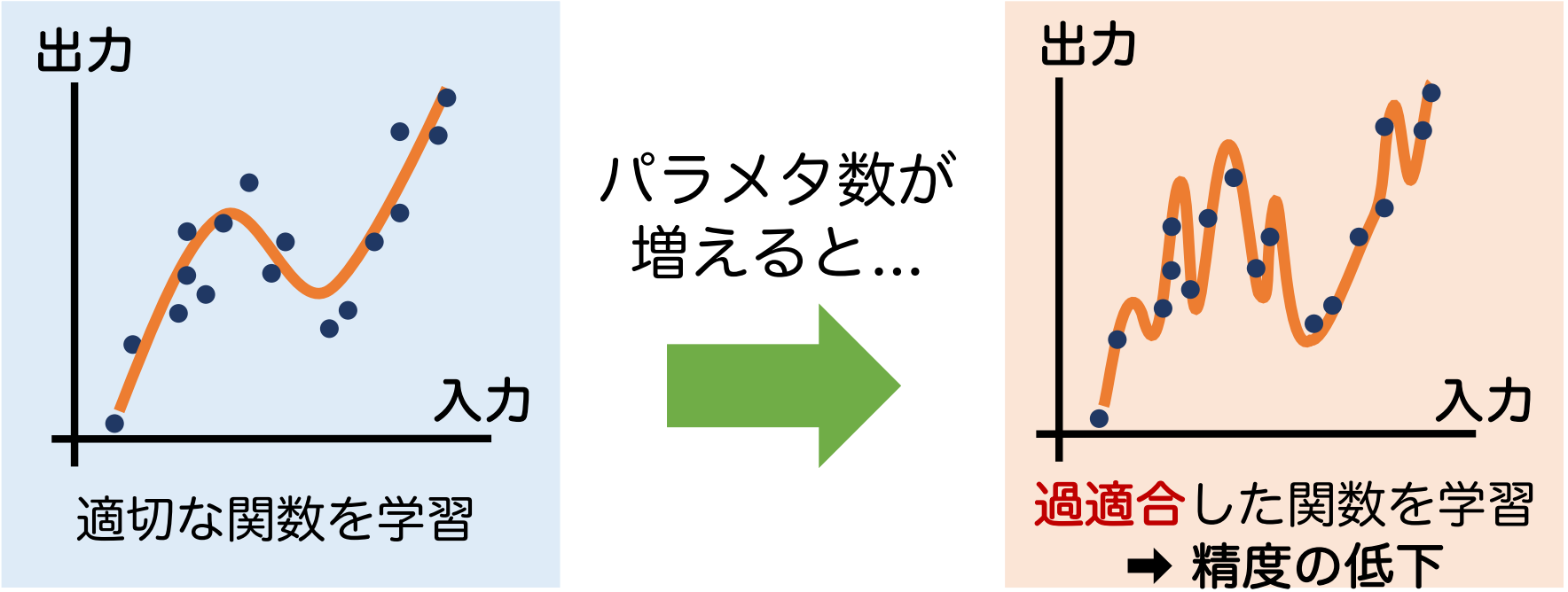
GPT-3
千億パラメタ

パラメタを増やすほど
予測精度が向上

→ 汎化再考：データ解析理論を再構築する必要性

謎2：膨大なパラメータは良くない？

統計数学の(大)原則
大量のパラメタは精度を下げる！

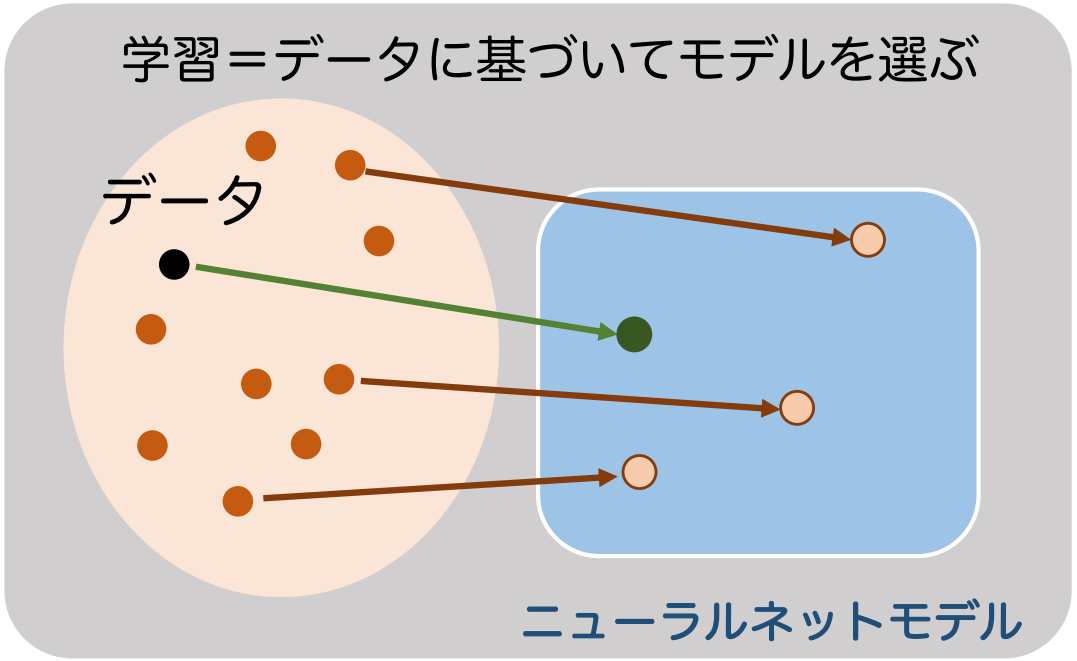


- 既存の統計学はパラメタ数の削減に腐心…
  - 変数選択、スパース推定、正則化、適応化など

謎2：膨大なパラメータは良くない？

<u>従来理論：モデルの大きさが重要</u>

- 過学習＝モデルの大きさが決める



モデルが大きい

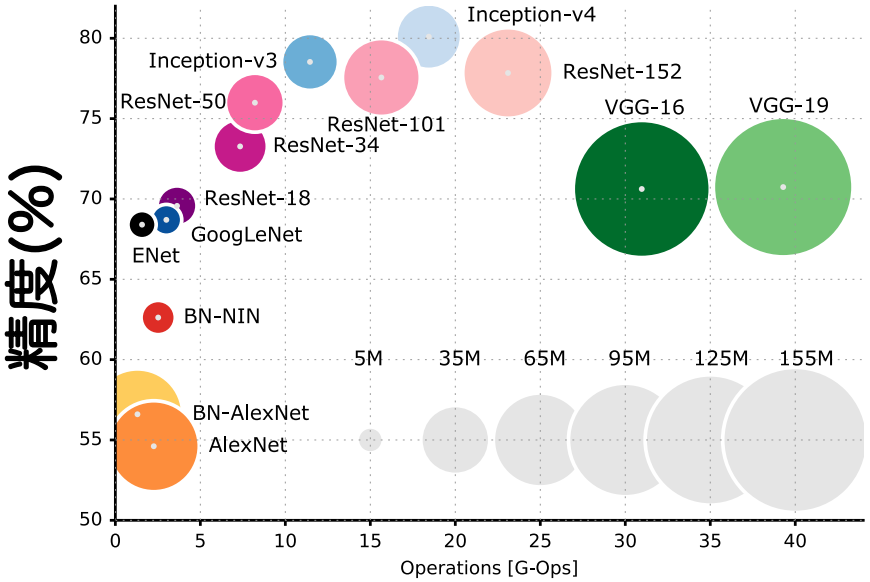
↓

データの摂動に対して
学習モデルも大きく変動

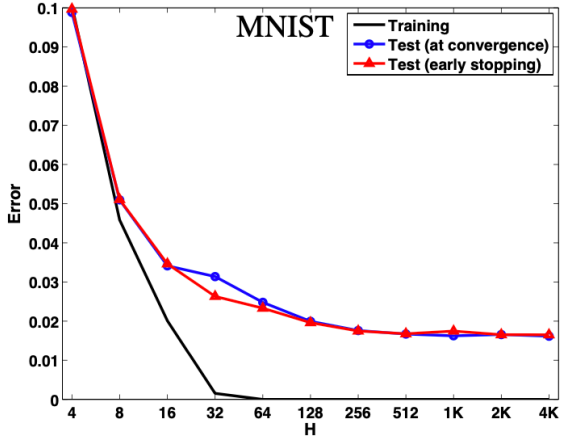
↓

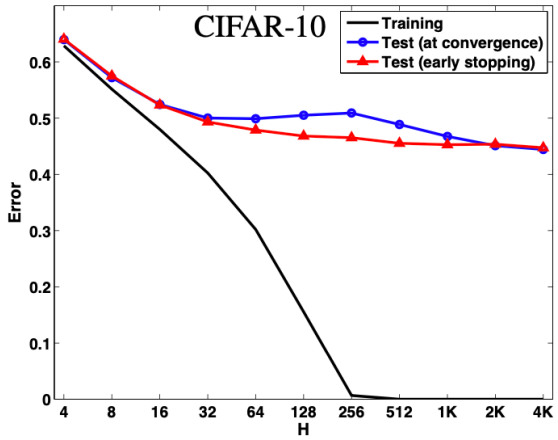
過学習が起こる(はず)

謎2：従来理論は深層学習の実際と乖離



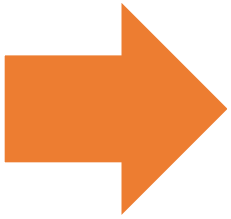
有名ネットワークの
精度とパラメタ数の関係
パラメータ数（丸の大きさ）が増加
することで精度（縦軸）が向上





実データの実験結果
ニューラルネットワークのサイズ
（横軸）の拡大に伴って
汎化誤差（赤線・青線）が減少
(Neyshabur+ 2018)

今日の概要



深層学習とその謎

数学的な説明と限界

物理学的なアプローチ

数学的に説明できたこと

発見1：多層は複雑な関数表現に最適

既存理論よりも複雑な関数を解析

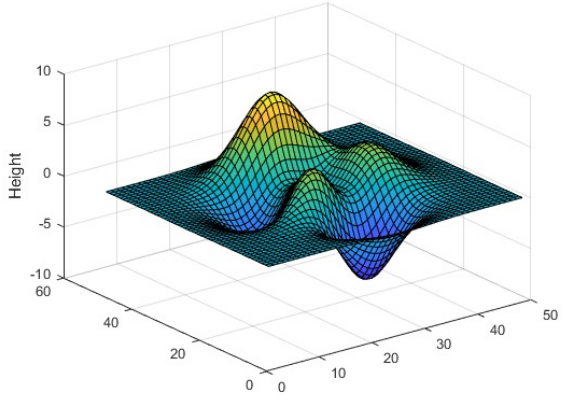
既存の理論が
分析した関数

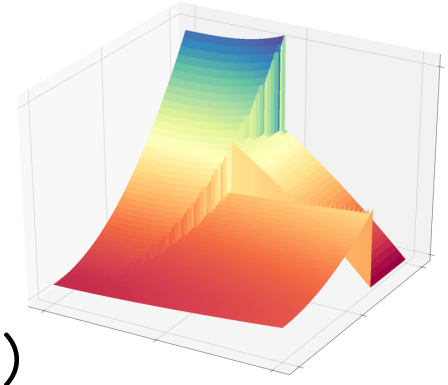
拡張

新しく分析する関数

斉一的な性質を持つ関数

どこでも同じ性質 (例：滑らかさ)



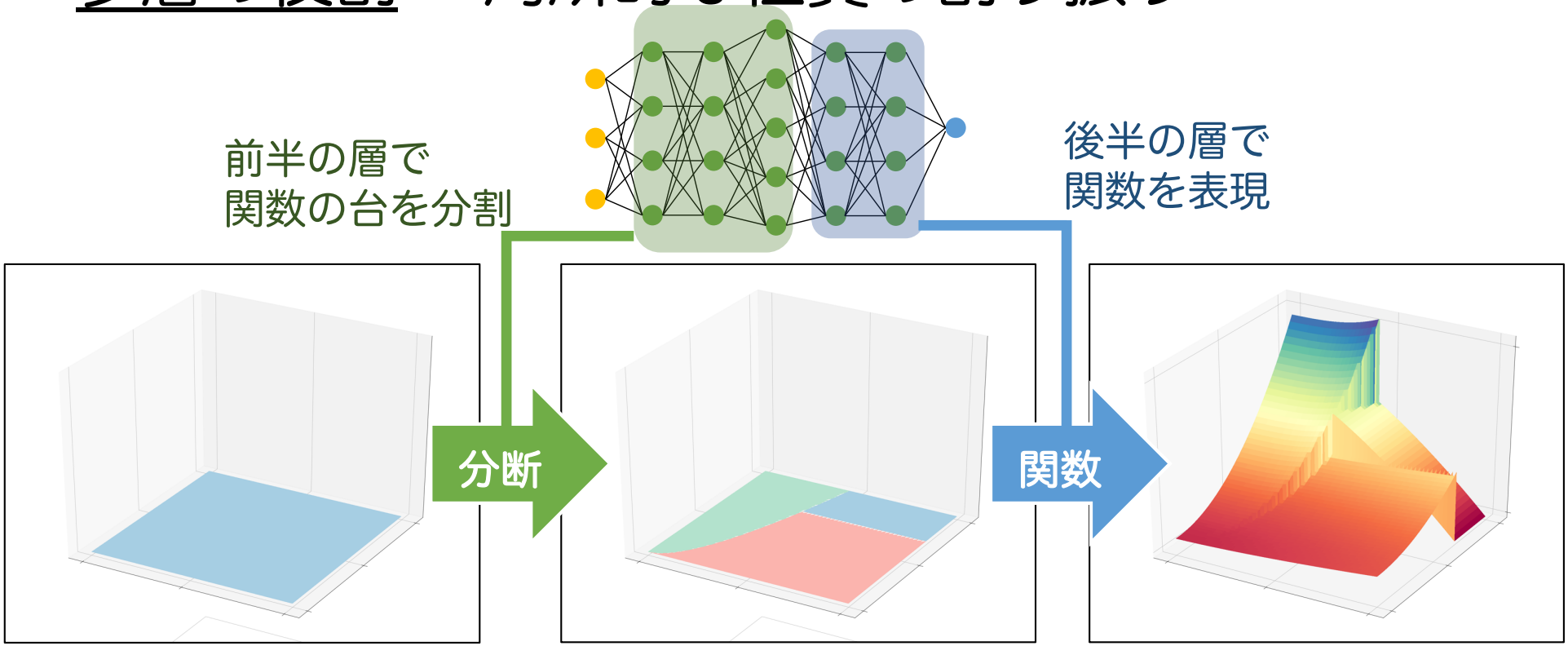


局所構造を持つ関数

場所によって違う性質を持つ（不連続）

発見1：多層は複雑な関数表現に最適

多層の役割： 局所的な性質の割り振り

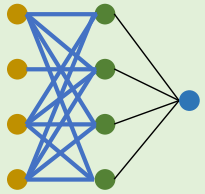


局所構造がある関数には深層学習が適している
（例：物理の相転移現象の表現）

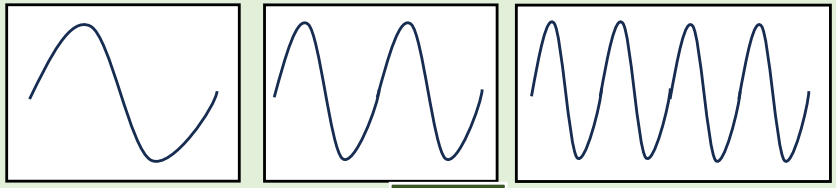
発見1：多層＝データ構造の学習

- 浅い層がデータの基本構造(特徴)を学習
  →データごとにより高精度な予測

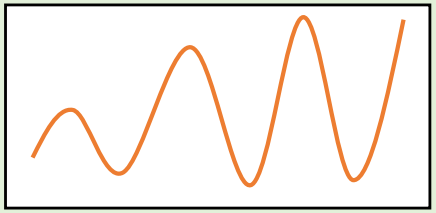
従来法
（層が少ないモデルの学習）



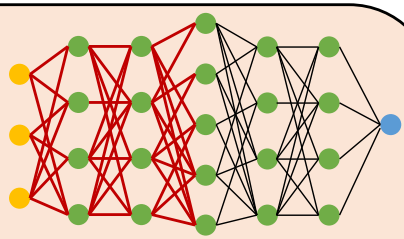
1. 外から特徴を準備 (例:フーリエ基底)



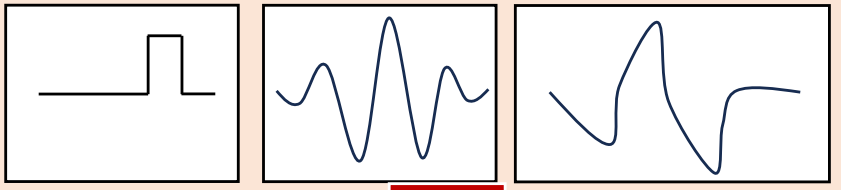
2. 特徴の組み合わせで関数を学習



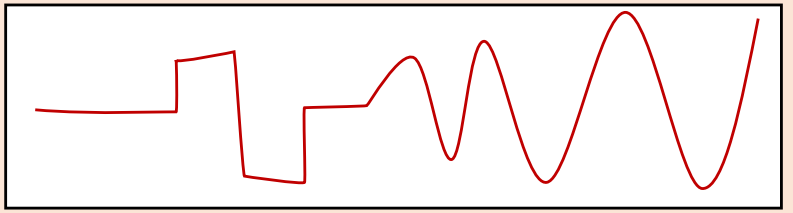
深層学習
（多層モデルの学習）



1. 手前の層がデータの特徴を学習

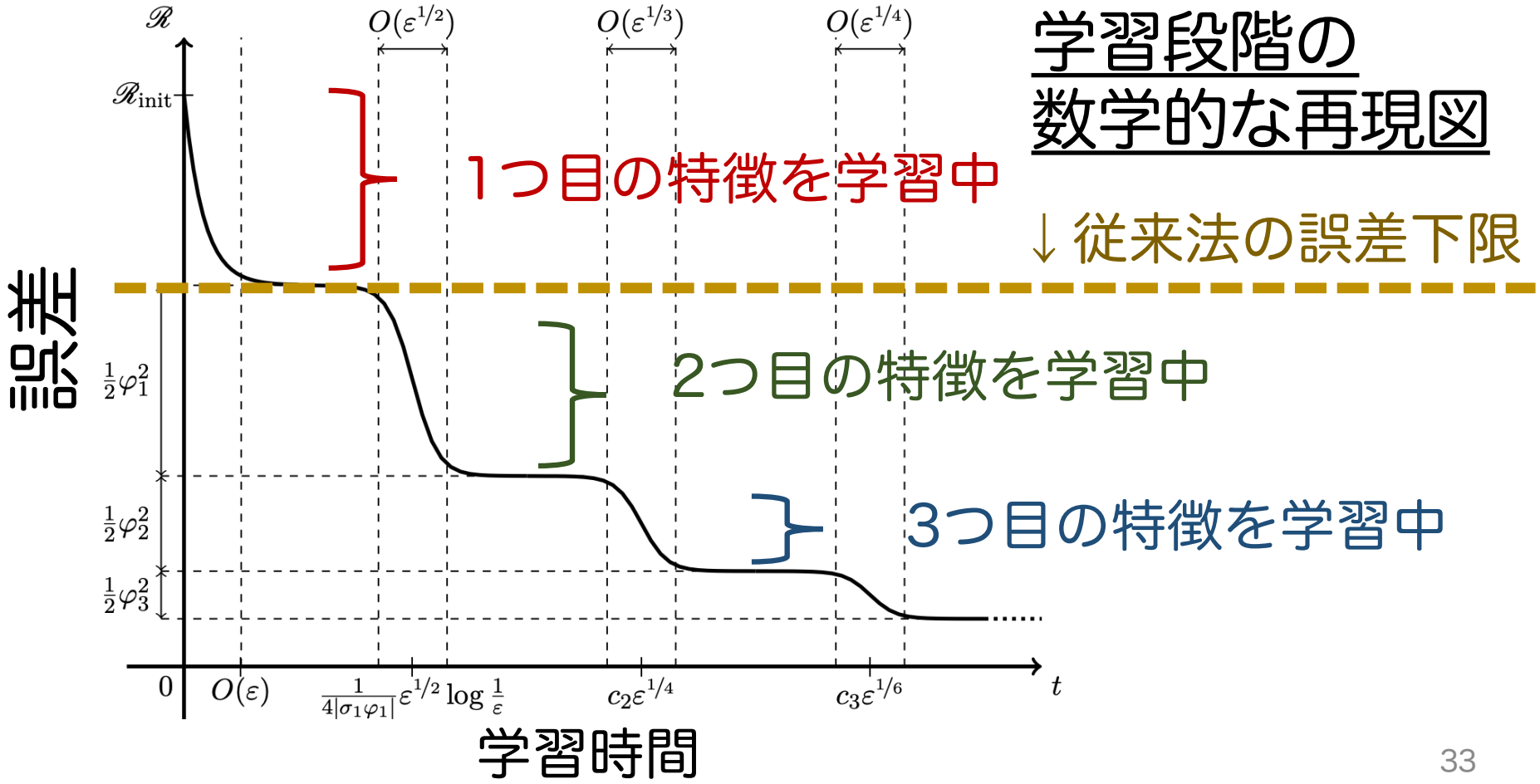


2. 組み合わせで複雑な関数を学習



発見1：特徴学習の段階的進展

- 特徴は1つずつ学習される→段階的発展

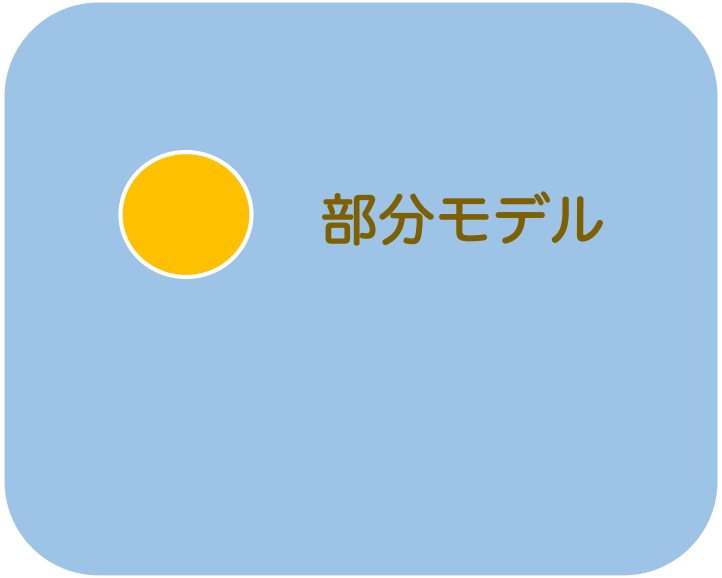


Berthier et al. (2024). Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, 1-84.

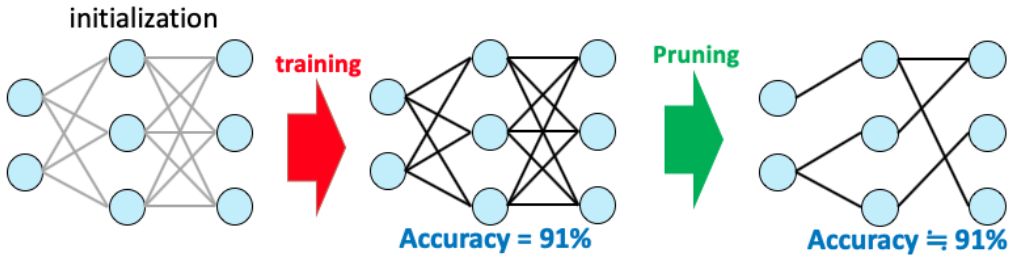
発見2：暗黙的正則化の発想

着想：ニューラルネットモデルすべてを考える必要は無い？

- 実質的に効いている部分もでるがありそう



ニューラルネットモデル
(多パラメタを使う巨大集合)

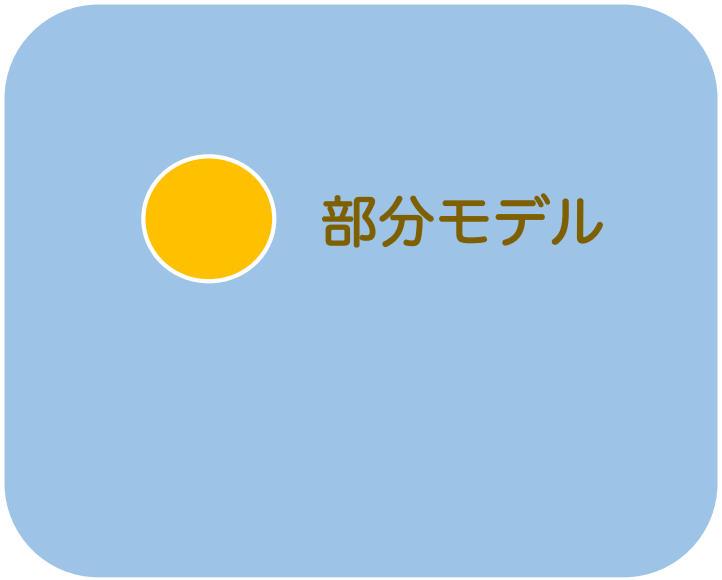


実際、学習後のネットワークは
一部の枝(パラメータ)を削除しても
十分良い予測性能を持つ

発見2：暗黙的正則化

着想：ニューラルネットモデルすべてを考える必要は無い？

- 実質的に効いている部分もでるがありそう



ニューラルネットモデル
(多パラメタを使う巨大集合)

従来理論

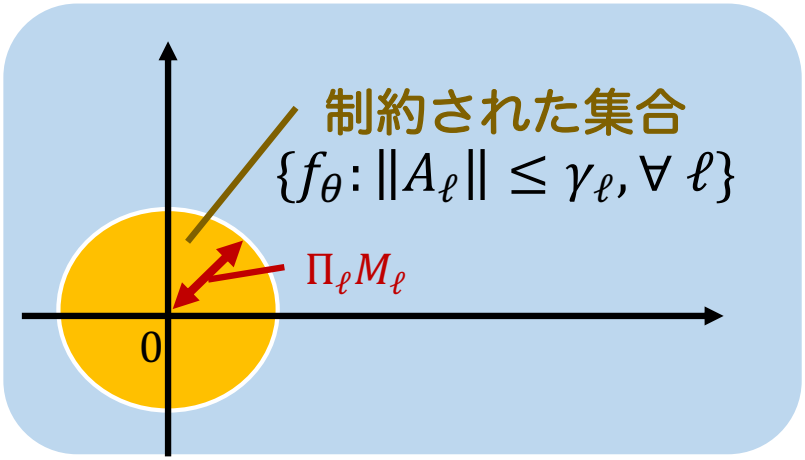
過学習の大きさ
= ニューラルモデル の大きさ

新しい理論

過学習の大きさ
= 部分モデル の大きさ

発見2：部分モデルの過学習理論

仮説1：原点近傍（パラメタ行列 $\|A_\ell\|$ が定数以下）



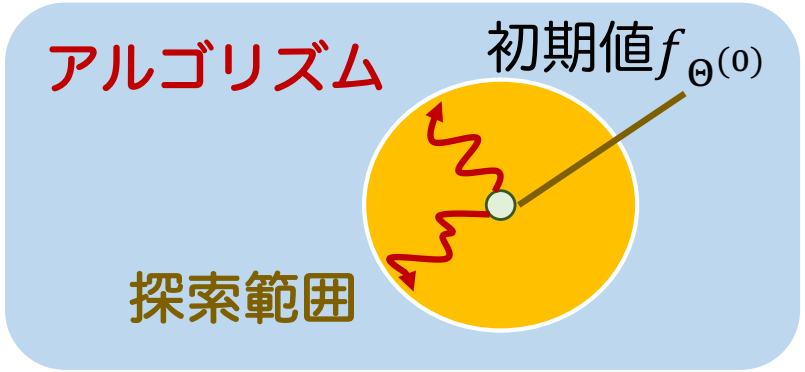
ノルム制約下での過学習誤差

$$O\left(\frac{B\sqrt{L}\prod_{\ell=1}^{L}\gamma_\ell}{\sqrt{n}}\right)$$

B: データの大きさ n: データ数 L:層数

- パラメタ数Wには（陽には）依存しない。

仮説2：探索アルゴリズムの初期値近傍



探索の長さと過学習誤差

$\eta_t \asymp 1/t$とする

$$O\left(\frac{T^q}{n}\right)$$

$T \geq 1$: 更新回数, $q \in (0,1)$: 減衰率

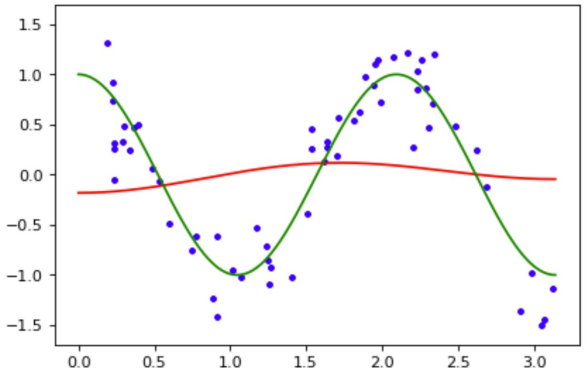
Bartlett, P. L., Foster, D. J., & Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. *NIPS*.
Hoffer, E., Hubara, I., & Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural

発見3：良性過適合

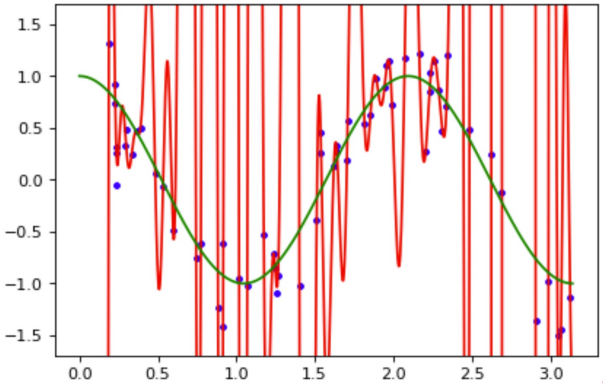
良性過適合 (benign overfitting)
大規模モデルは訓練データへの適合と高い予測性能を両立

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

パラメタ数2



パラメタ数50

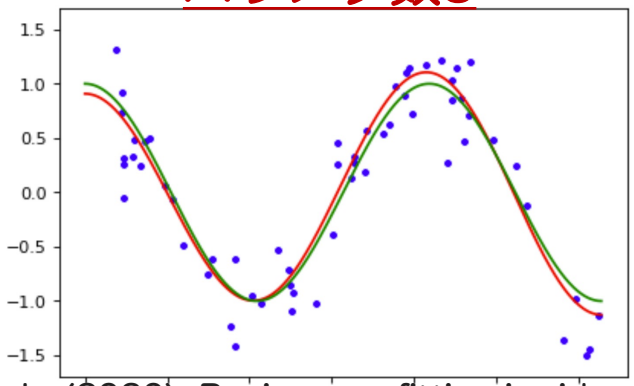


⇐ 過適合

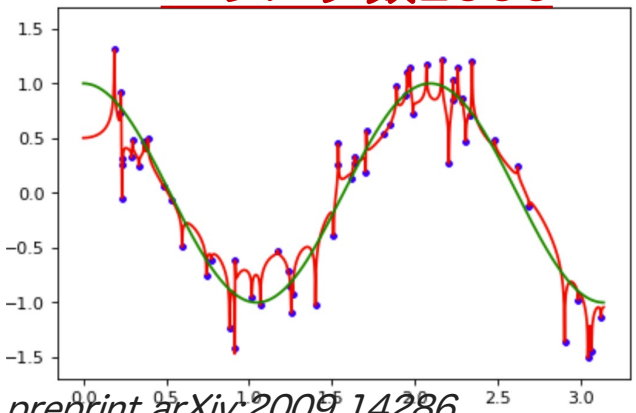
⇓ 良性過適合

適合 ⇒

パラメタ数3



パラメタ数2000

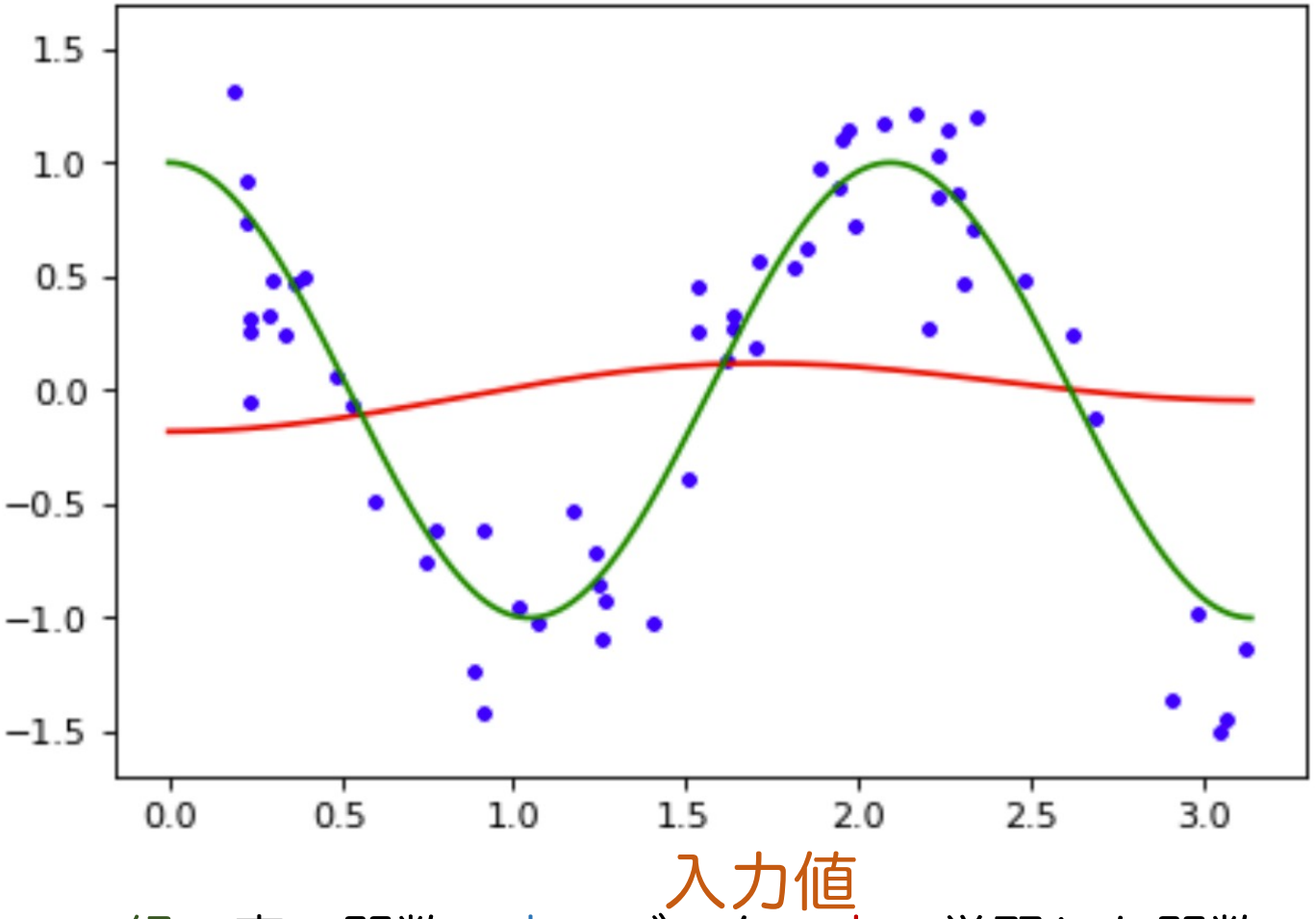


Tsigler, A., & Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

発見3：良性過適合の様子

データ数:60 パラメタ数2

出力値



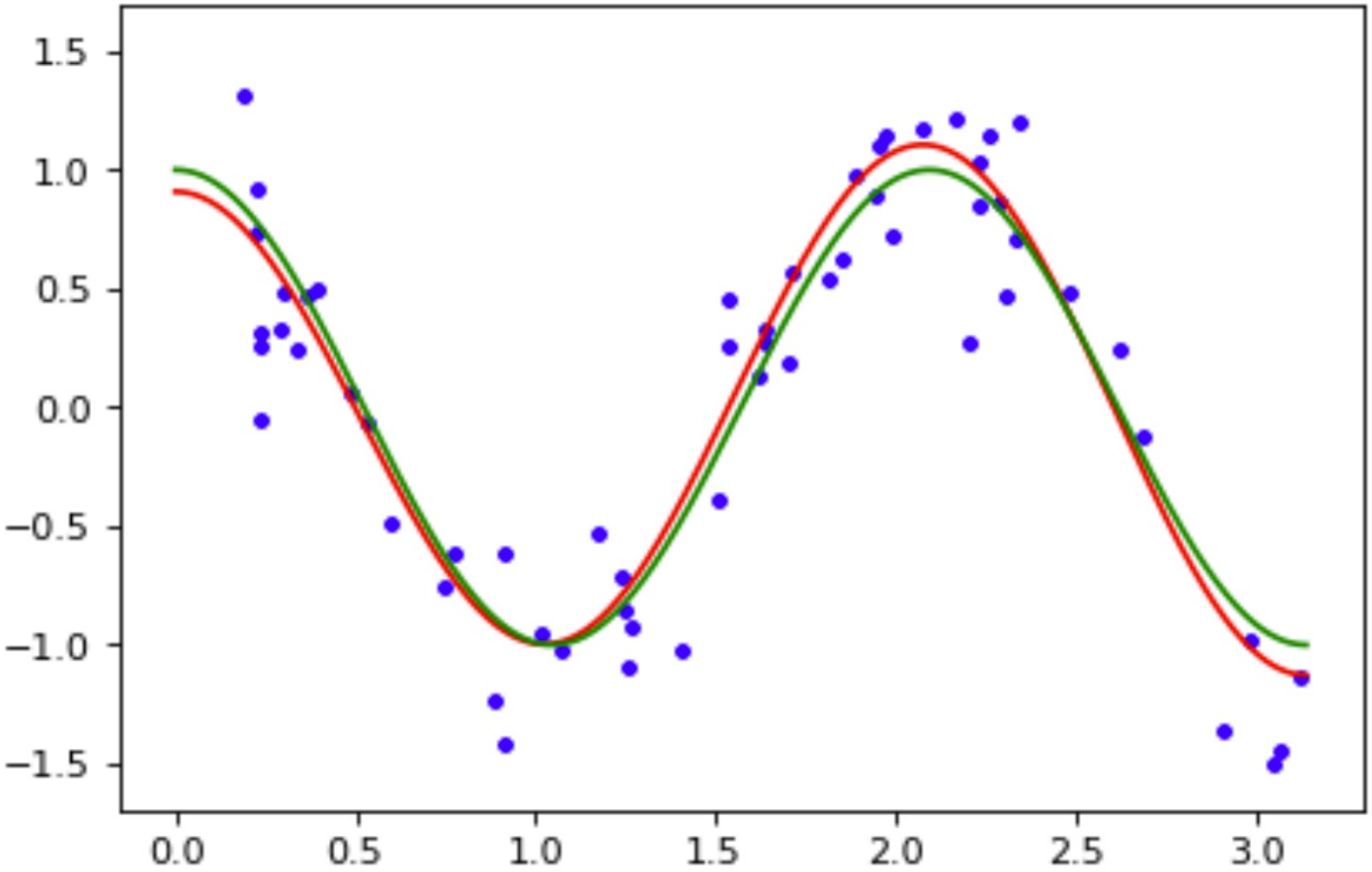
入力値

緑：真の関数、青：データ、赤：学習した関数

発見3：良性過適合の様子

データ数:60 パラメタ数3

出力値

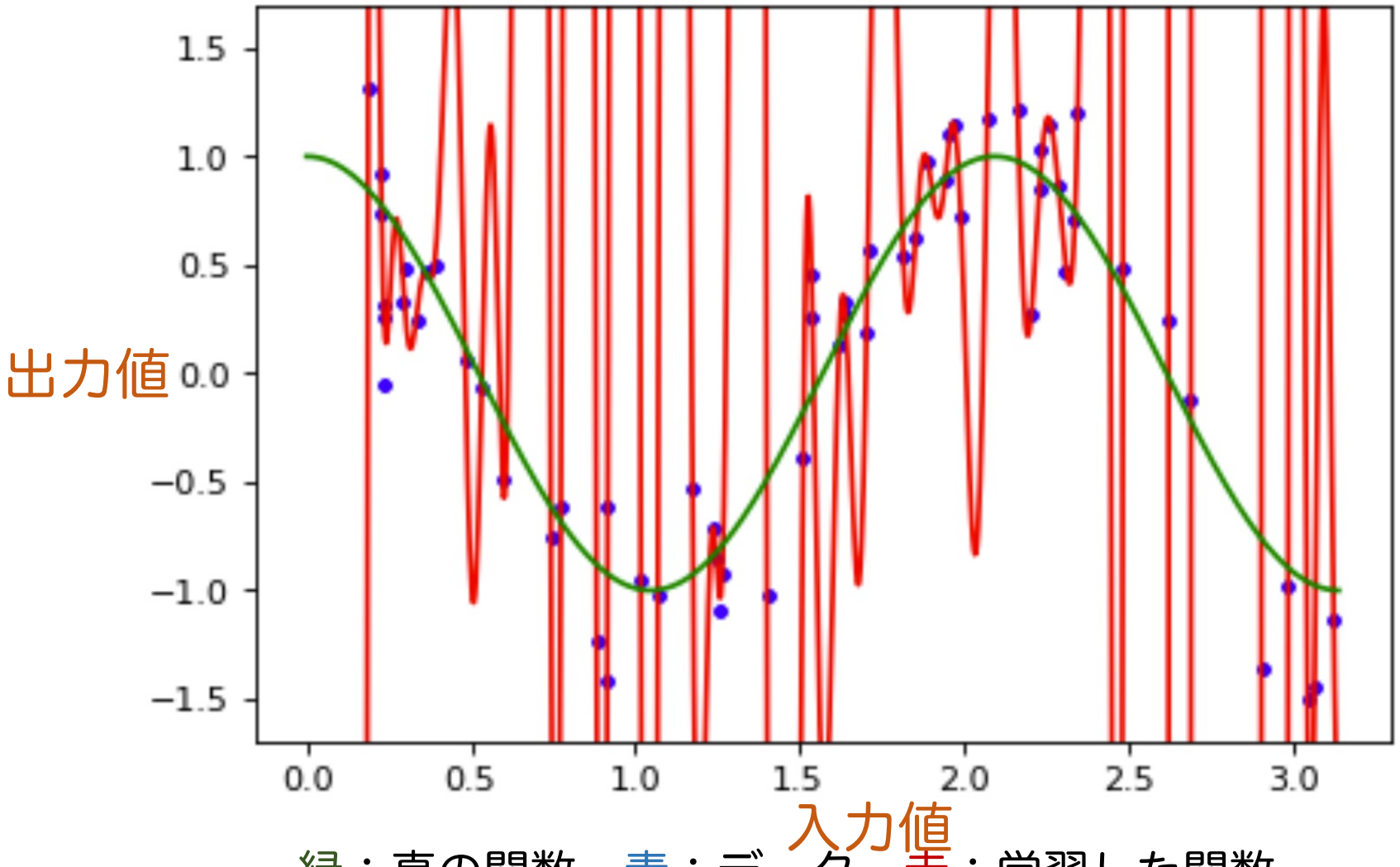


入力値

緑：真の関数、青：データ、赤：学習した関数

発見3：良性過適合の様子

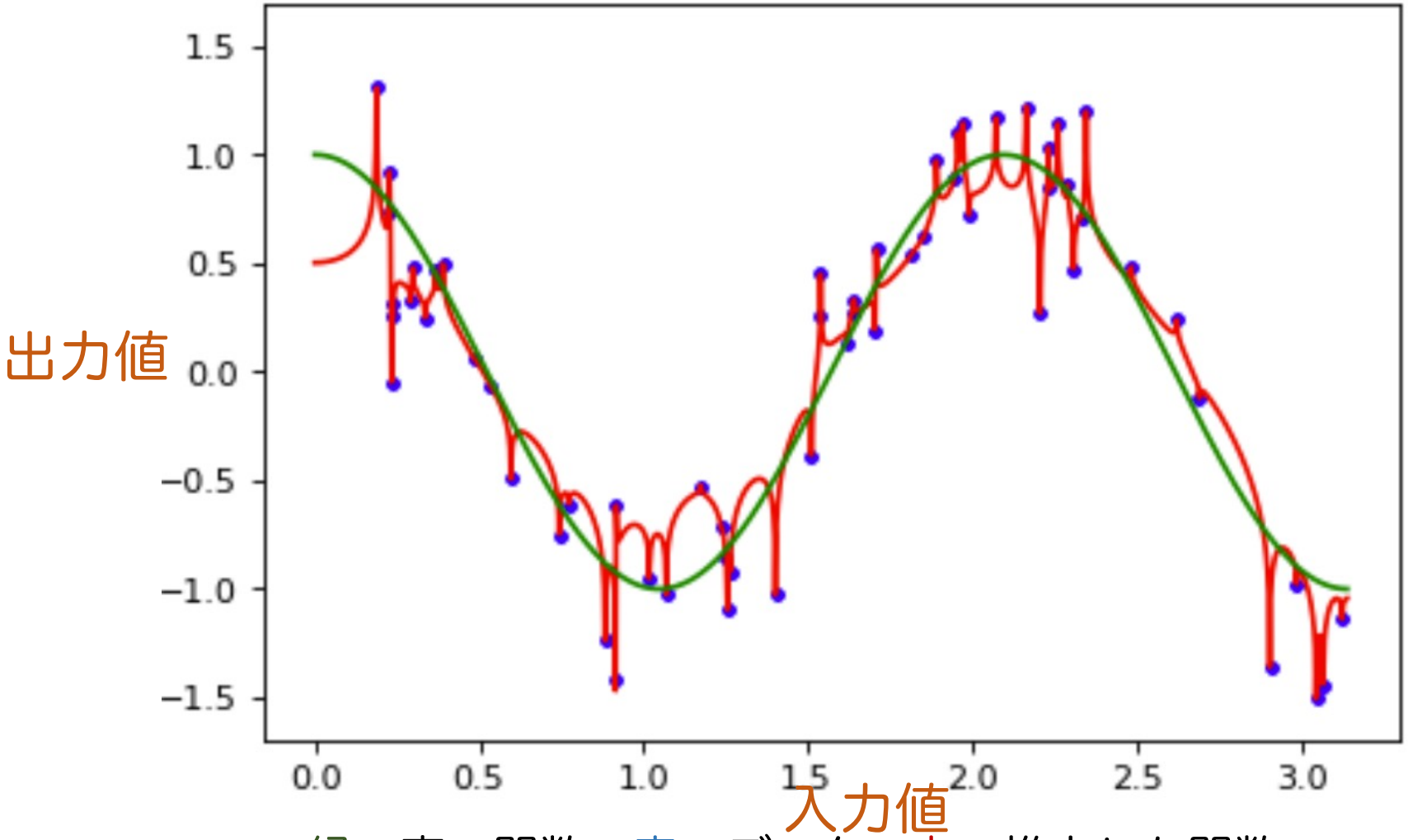
データ数:60 パラメタ数50



緑：真の関数、青：データ、赤：学習した関数

発見3：良性過適合の様子

データ数:60 パラメタ数2000

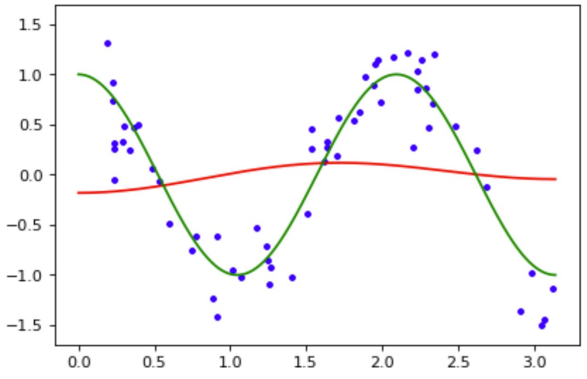


緑：真の関数、青：データ、赤：推定した関数

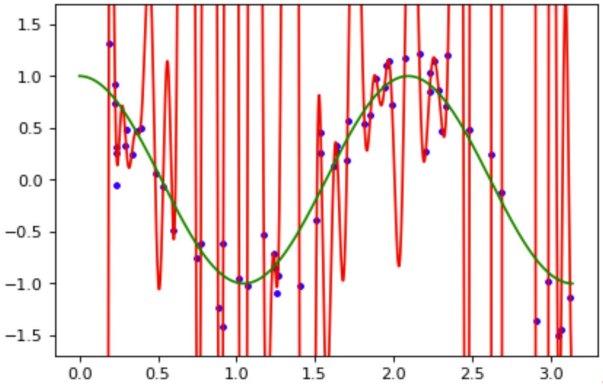
発見3：良性過適合

良性過適合 (benign overfitting)
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
→複数の設定で数学的・理論的な再現も実現

パラメタ数2



パラメタ数50

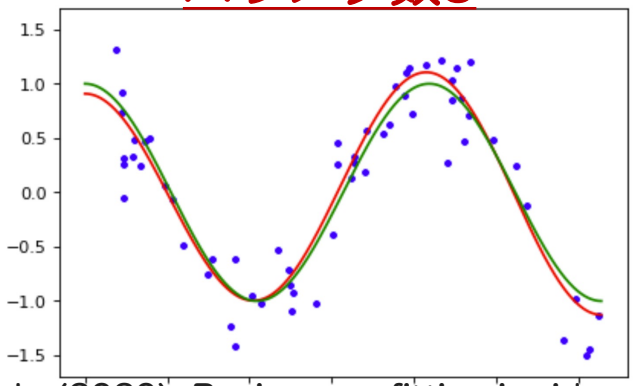


⇐ 過適合

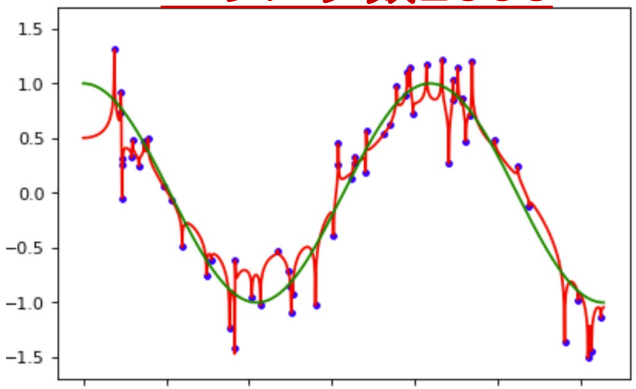
⇓ 良性過適合

適合 ⇒

パラメタ数3



パラメタ数2000



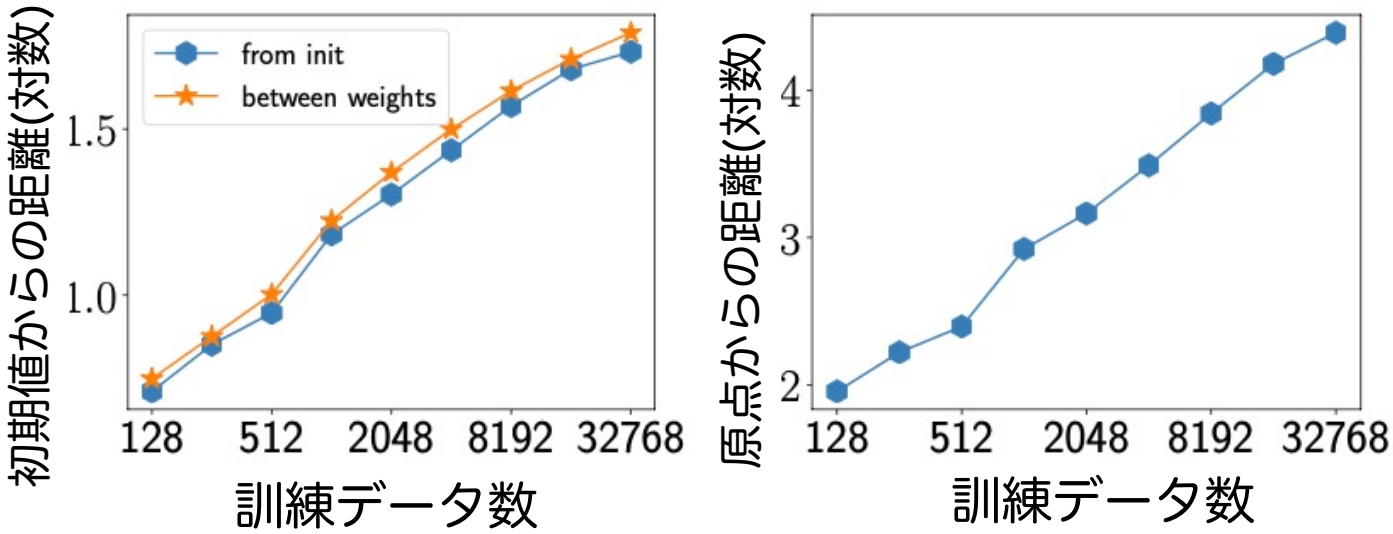
Tsigler, A., & Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

数学的な説明の限界

過学習に関する発見2,3の反論・限界

発見2(暗黙的正則化)への批判：実験的反証

実験：学習すると特定の部分モデルに留まらない
理論：留まることは理論的にも保障されない



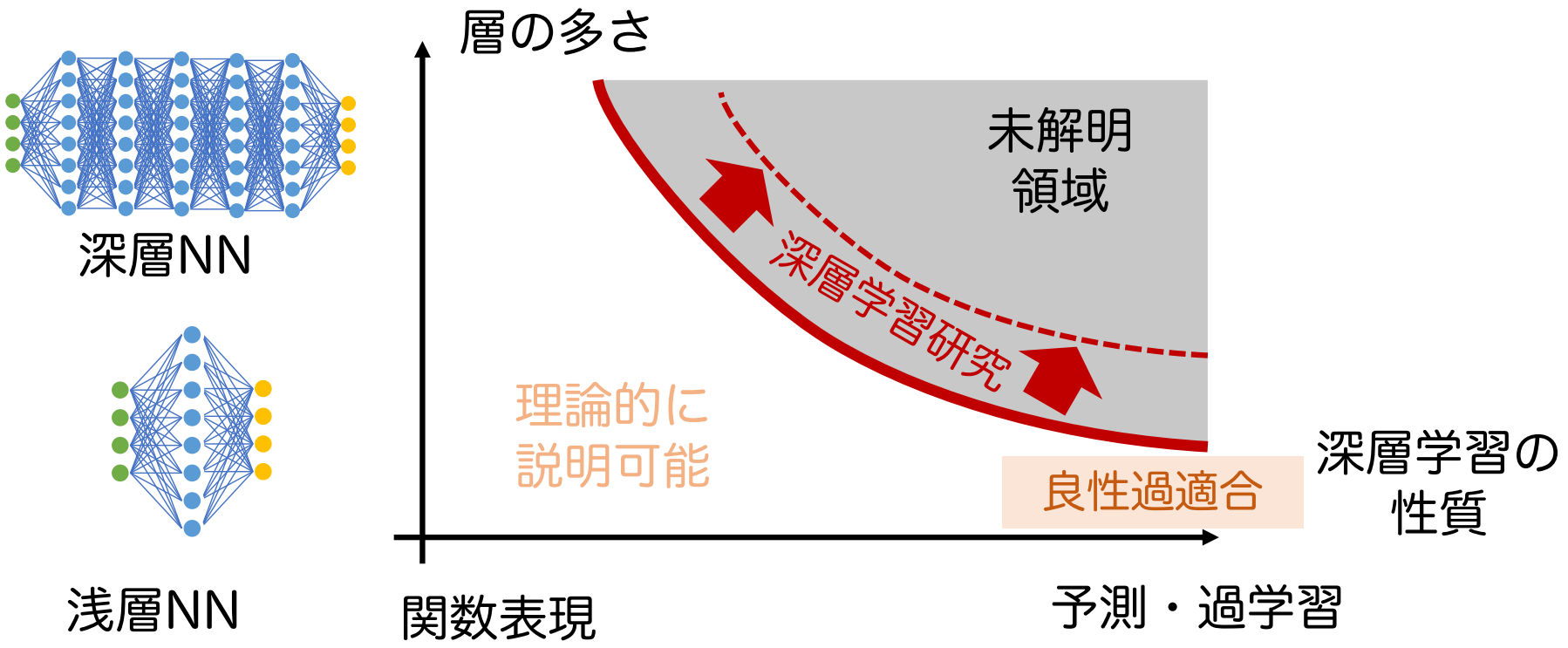
計算機実験で、データ数（横軸）が増えることにパラメタが原点・初期値から遠ざかる（縦軸は距離）様子（Nagarajan+ NeurIPS2019 卓越論文）

→ 暗黙的正則化理論の未解決問題

Nagarajan, V., & Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. *NIPS.*

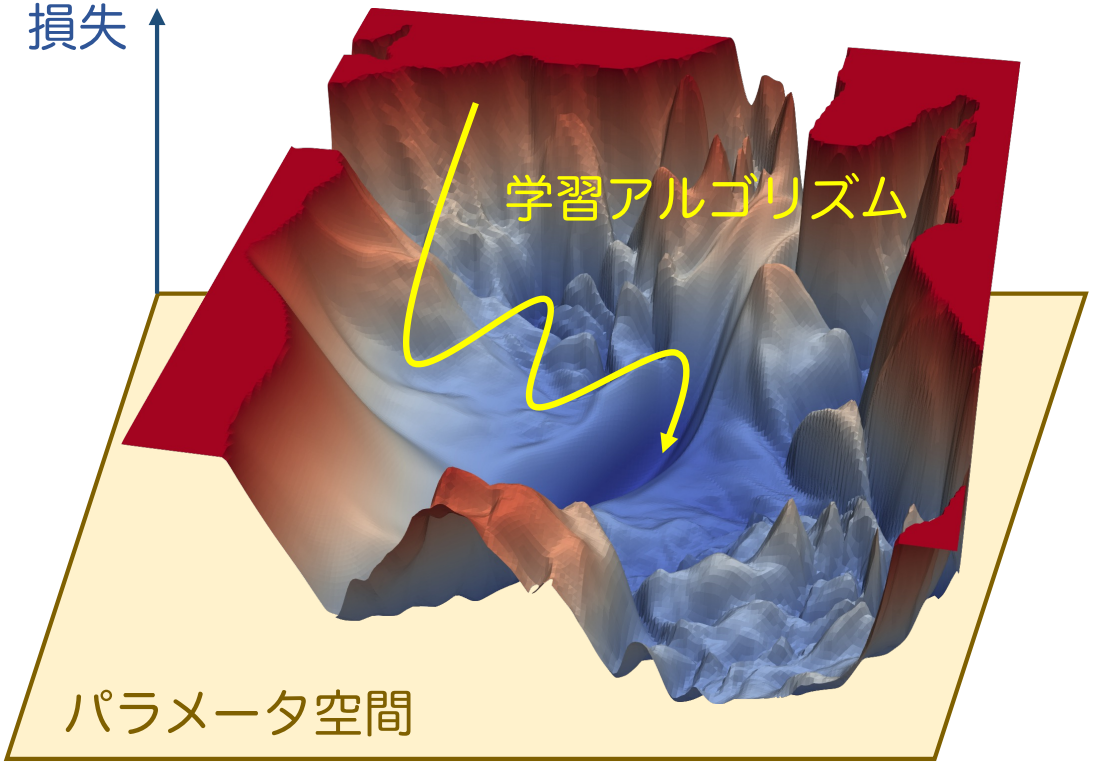
発見3(良性過適合)の限界：多層への非対応

- 良性過適合の理論は層が増えると成立しない



なぜ数学的な説明が難しいか？

- 学習アルゴリズムのカオス的な挙動
  →従来の数学的な方法では近似しきれない



2次元に圧縮した
損失関数の図

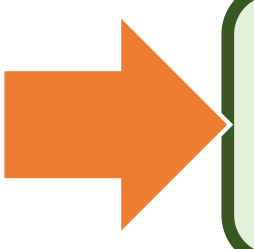
実際は30万次元なので
本物の可視化は困難

層・パラメータ数が増えると数学的理論では解析に限界

今日の概要

深層学習とその謎

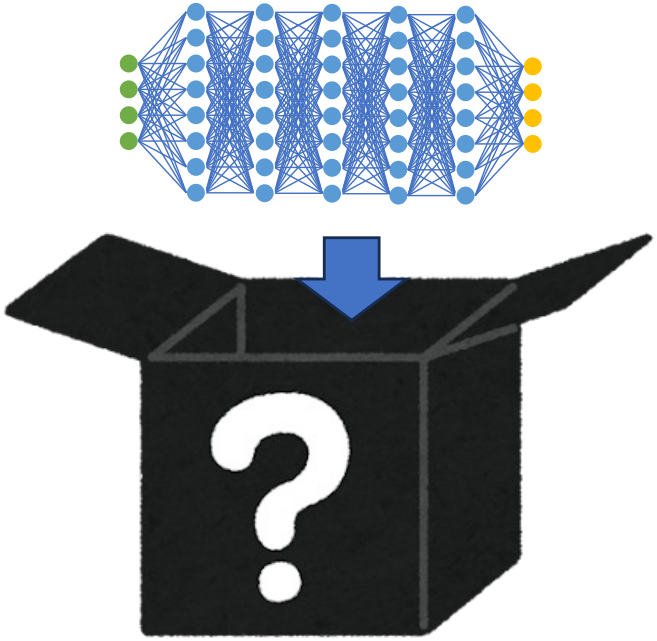
数学的な説明と限界



物理学的なアプローチ

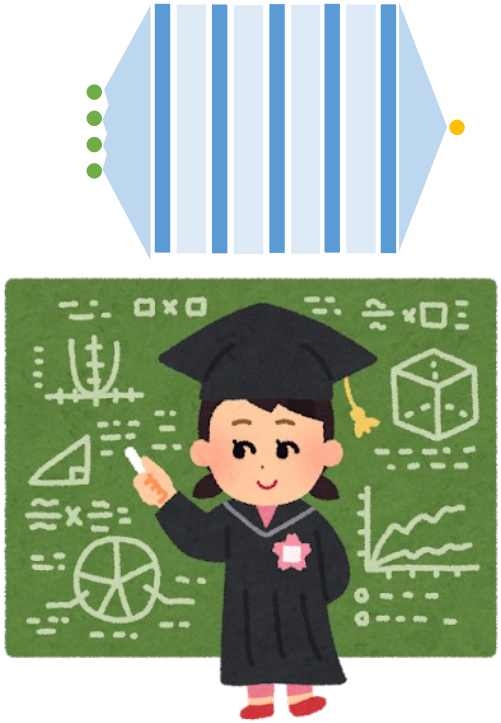
深層学習現象を記述する方程式の導出

- 深層学習（ブラックボックスな現象）を
  物理学的な方程式で記述→完全な再現



計算機上の現象
（ブラックボックス）

物理学的
近似



方程式
（数学的な再現）

利点：精密解析

- 誤差の値を（近似なしで）厳密に求める理論

データの種類と数は～
アーキテクチャは～
アルゴリズムは～

入力

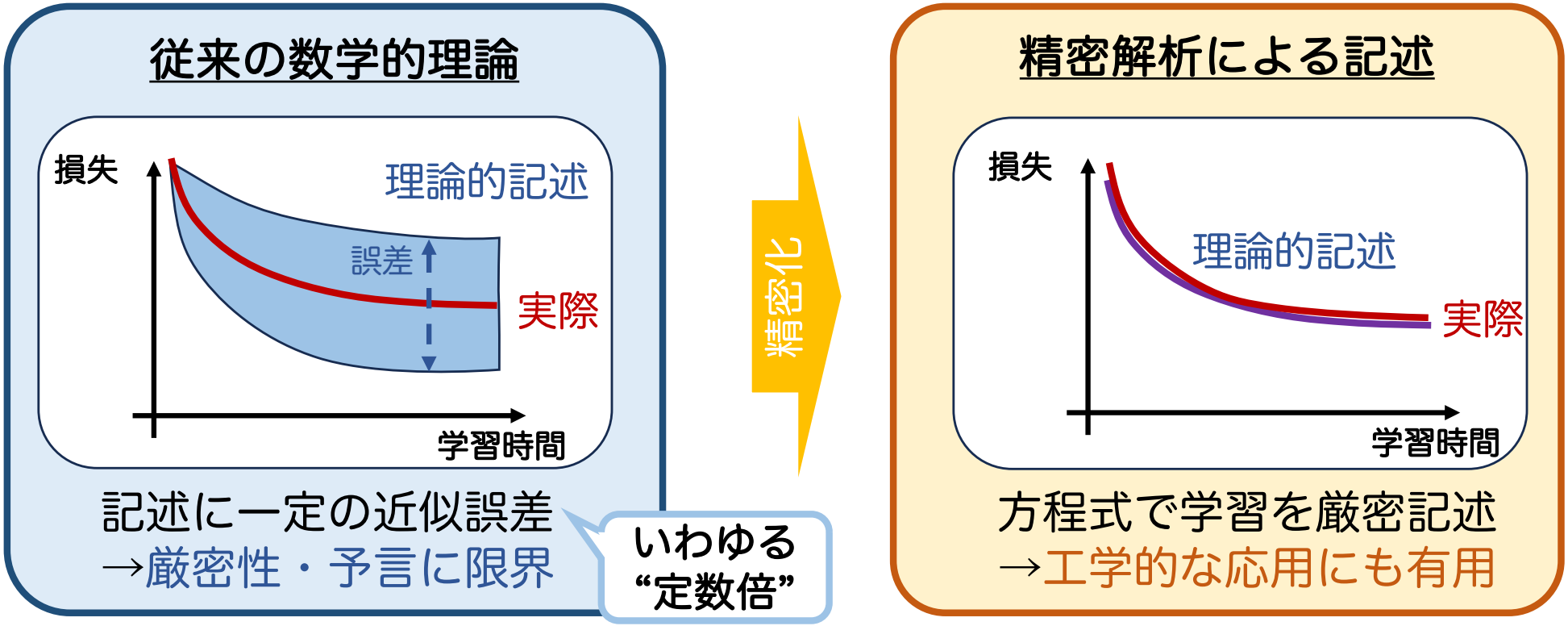
100回更新すると
損失（訓練誤差）は
0.29になるよ！

イメージ例：
弾道予測のための
物理法則

実際の学習をしなくても
学習結果を予言できる理論的な方程式を得る

従来理論の弱点をカバー

- 従来理論は回避できない近似誤差を持つ



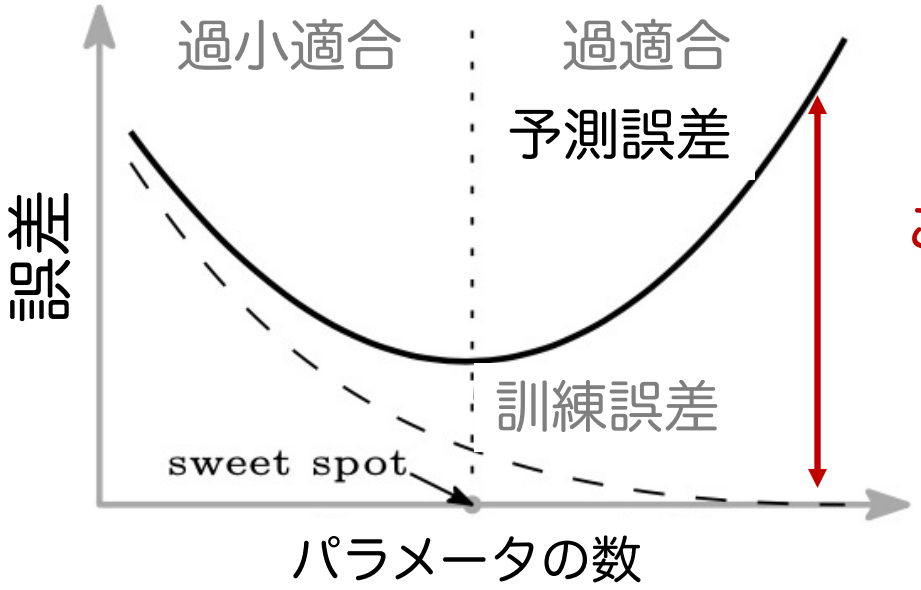
深層学習プロセスは従来理論では誤差が大きく限界
→物理学的な精密解析の有用性

成果例1：二重降下理論

<u>従来理論</u>：バイアス・分散の二律背反

- モデルを必要以上に大きくすると、
  分散が大きくなって過適合が発生する

既存理論の考え



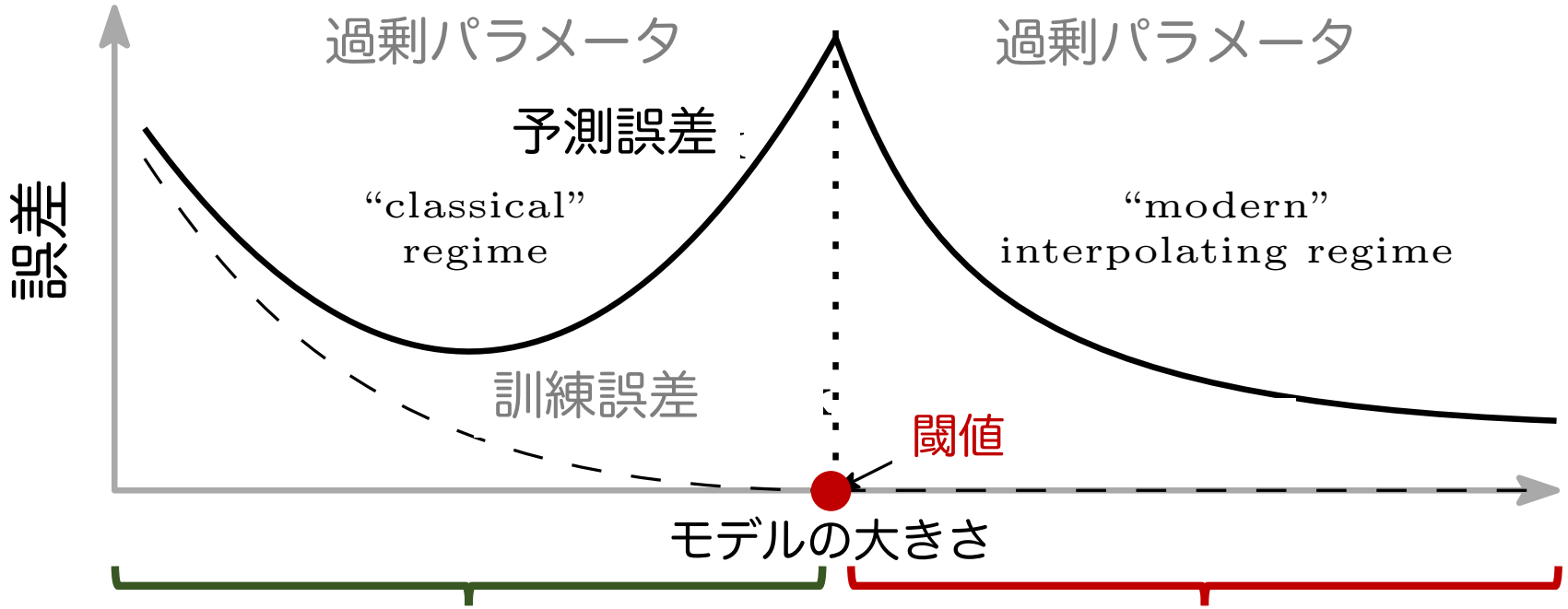
このギャップが分散
（過適合の大きさ）

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.*

成果例1：二重降下理論

二重降下理論
モデルをさらに大きくすると誤差が再び減少



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

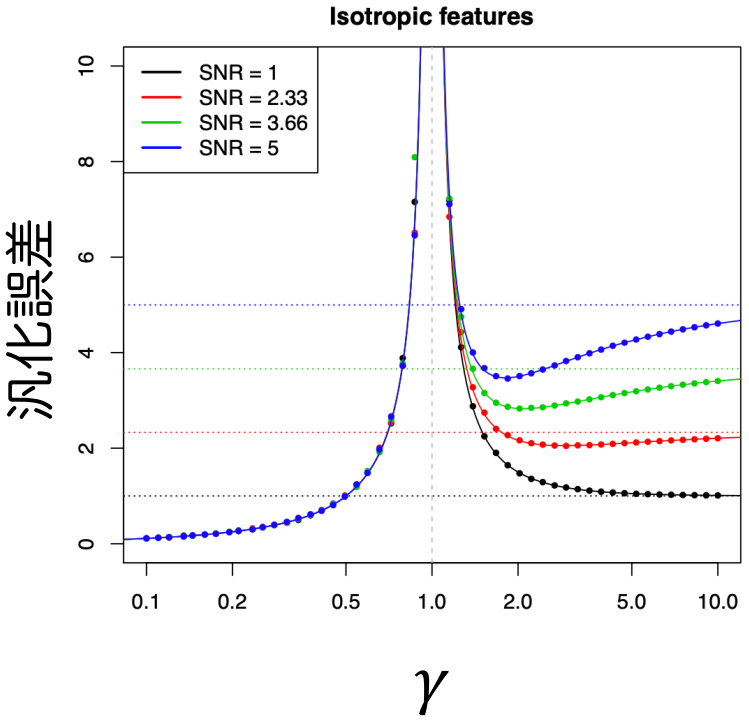
成果例1：二重降下理論

二重降下の理論的再現

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

$$\text{誤差} = \begin{cases} \dfrac{\sigma^2\gamma}{1-\gamma}, & (\gamma < 1) \\ \|\beta^*\|_2^2(1-\gamma^{-1}) + \dfrac{\sigma^2}{\gamma-1}, & (\gamma > 1) \end{cases}$$

$= \text{バイアス}B$（近似誤差）　　$= \text{分散}V$（≈過学習）



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

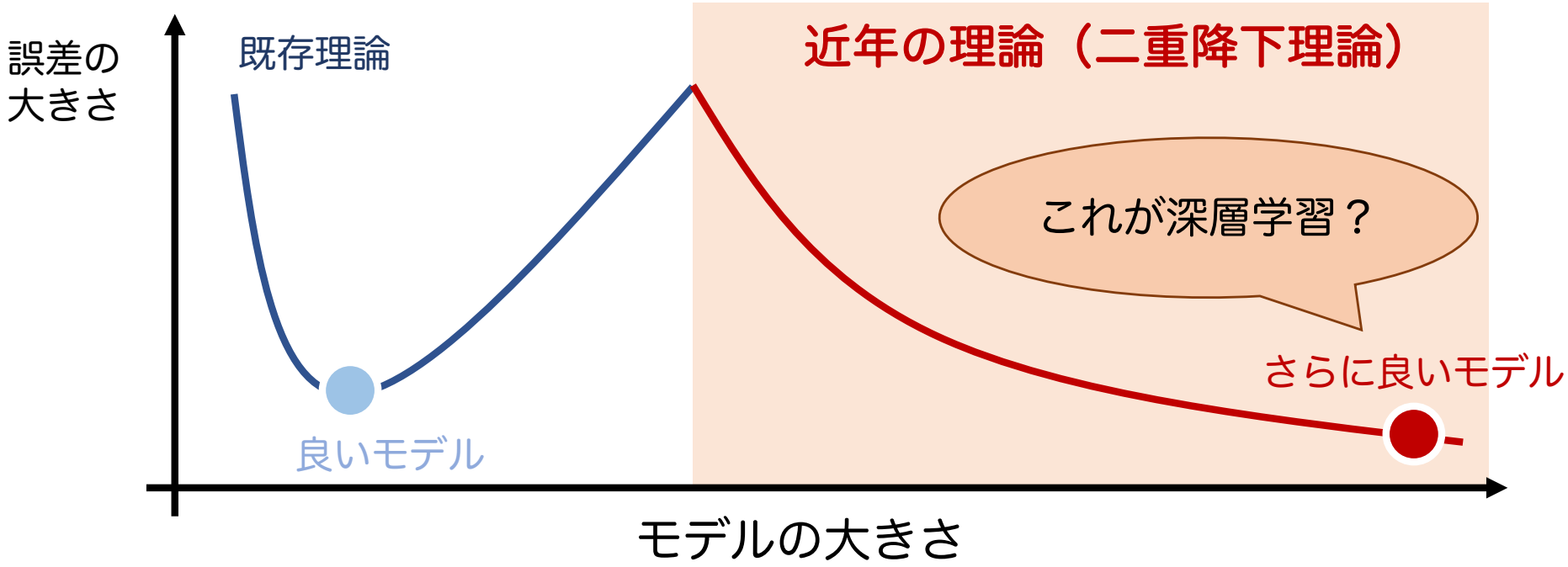
✓ パラメータ数が増えると$(\gamma \to \infty)$ 複雑性誤差が減少

× 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.*

成果例1：二重降下理論

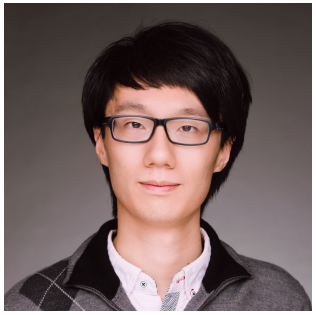
新しいパラダイム：大きいモデルほど安定？



理論的結果：　多様な尺度・理論が提案
汎用的・統一的な理論は今後の課題

成果例2：深層学習ダイナミクスの精密解析

私たちの研究成果

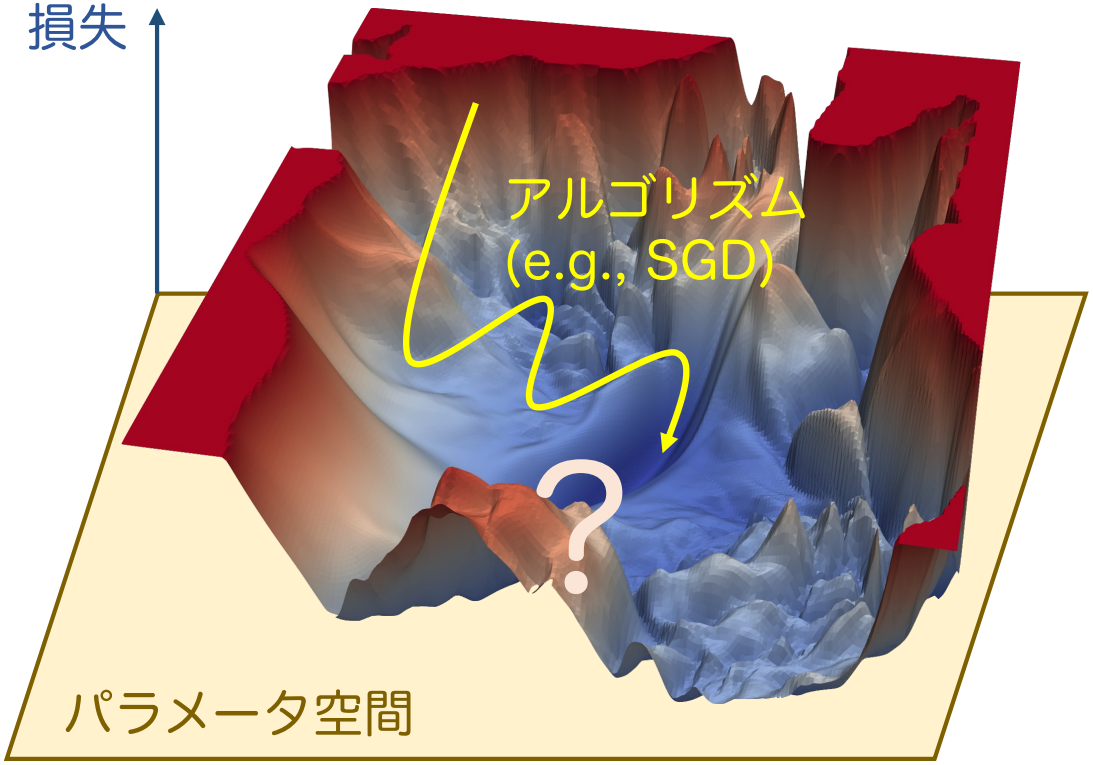


Qiyang Han
(Rutgers U)

Q. Han and M. Imaizumi. Precise gradient descent training dynamics for finite-width multi-layer neural networks. arxiv:2505.04898

深層学習ダイナミクス

- 学習でパラメータが動く過程
  →深層学習理論最大のブラックボックス



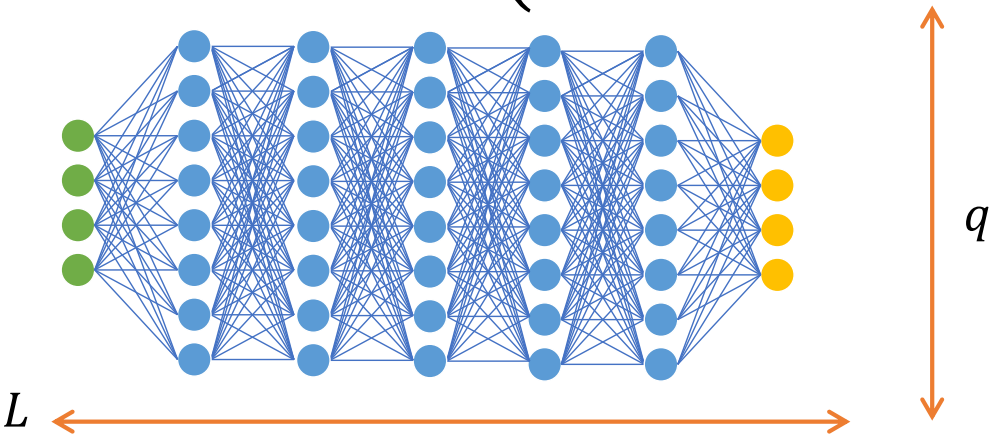
30万次元を
2次元に圧縮して可視化

層が増えるほどより複雑に→解析が難しい

関心：深層ニューラルネット

- 統計モデル：深層ニューラルネットワーク
  - $L \in \mathbb{N}$: 深さ（層の数）, $q \in \mathbb{N}$: ネットワークの幅
  - $m \in \mathbb{N}$: データ点の次元
  - $W_\ell \in \mathbb{R}^{q \times q}$: $\ell = 2, \ldots, L-1$層目の重み行列
  - $\boldsymbol{W} = (W_1, \ldots, W_L)$: 全ての重み行列
  - σ: 活性化関数

$$f_{\boldsymbol{W}}(x) = W_L^{\mathsf{T}} \sigma\left(W_{L-1}^{\mathsf{T}} \cdots \sigma(W_1^{\mathsf{T}} x)\right), x \in \mathbb{R}^m$$



深さ6・幅8の
深層ニューラルネット

学習問題：回帰

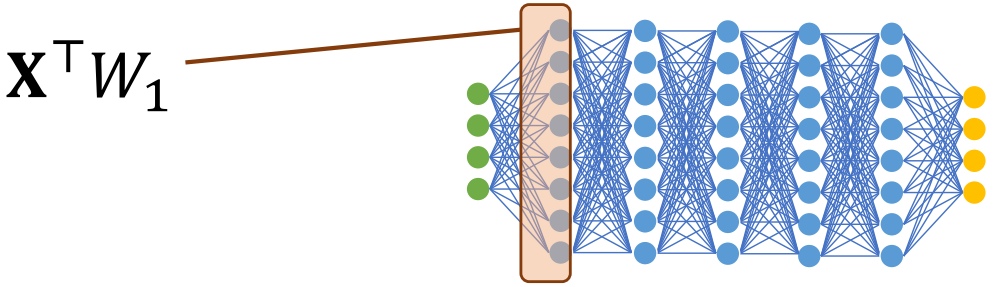
学習の設定

- $(X_i, Y_i) \in \mathbb{R}^m \times \mathbb{R}, i = 1, \ldots, n$: 観測データ
- 二乗損失

$$L(\boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - f_{\boldsymbol{W}}(X_i) \right)^2$$

記法：プレ活性化値

- $\mathbf{X} := (X_1, \ldots, X_n) \in \mathbb{R}^{m \times n}$: データ行列
- $\mathbf{X}^\top W_1 \in \mathbb{R}^{n \times q}$: プレ活性化値 （1層目）



問題関心：汎化誤差

勾配降下アルゴリズムによる学習

- $t = 0,1,2, \ldots$
- $\eta_t > 0$: 学習率（ステップ幅）

$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^{t} - \eta_t \nabla L(\boldsymbol{W}^{t})$$

関心：汎化誤差（テスト誤差）

$$\mathcal{E}(\boldsymbol{W}^{t}) = E_{X,Y}\left[\left(Y - f_{\boldsymbol{W}^{t}}(X)\right)^2\right]$$

- 新しいデータの元での期待損失

→この汎化誤差の動力学(ダイナミクス)を知りたい

ダイナミクスを知るとは？

- 何を知りたい
  - 各時刻 $t = 1,2,\dots,$ で、 $\mathcal{E}(\boldsymbol{W}^t)$ を解析式で表現.

設定: 高次元極限における汎化誤差

- サンプル数 n と次元 m が無限大に
- 比 m/n がある値に収束： $\phi \in (0,\infty)$.

$$\lim_{n,m\to\infty,\ m/n\to\phi\in(0,\infty)} \mathcal{E}(\boldsymbol{W}^t) = ?$$

- 各時刻 $t = 1,2,\dots$ ごとに評価

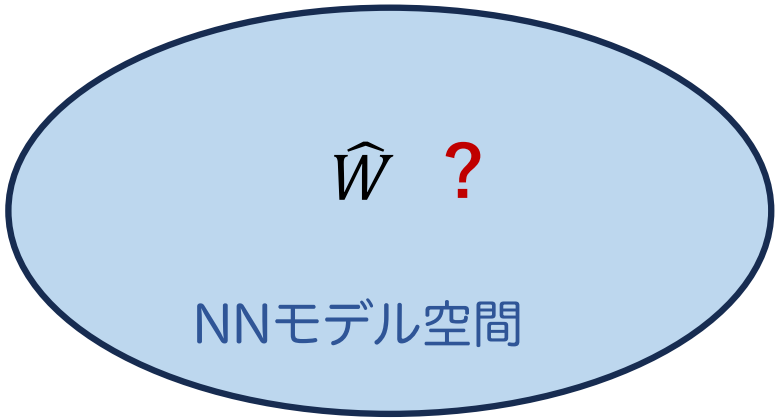
多層NNの精密な誤差を厳密に導出したい.

$\hat{W} \in \operatorname{argmin} L(W)$:
経験誤差最小解

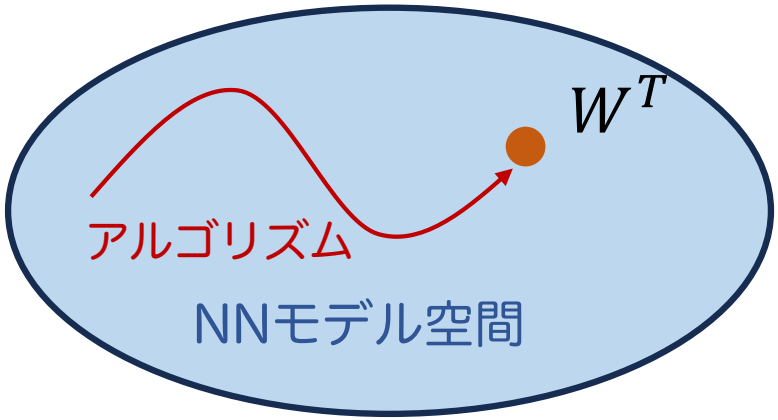
多層の学習ダイナミクスは最大の理論的ボトルネック(のひとつ)

- 巨大モデルの重要な性質は
  動的な性質を使わないと記述できない（はず）

静的な解析



動的な解析



一様上界による誤差評価

$$\mathcal{E}(\hat{W}) \leq L(\hat{W}) + \sup_{W}(\mathcal{E}(W) - L(W))$$

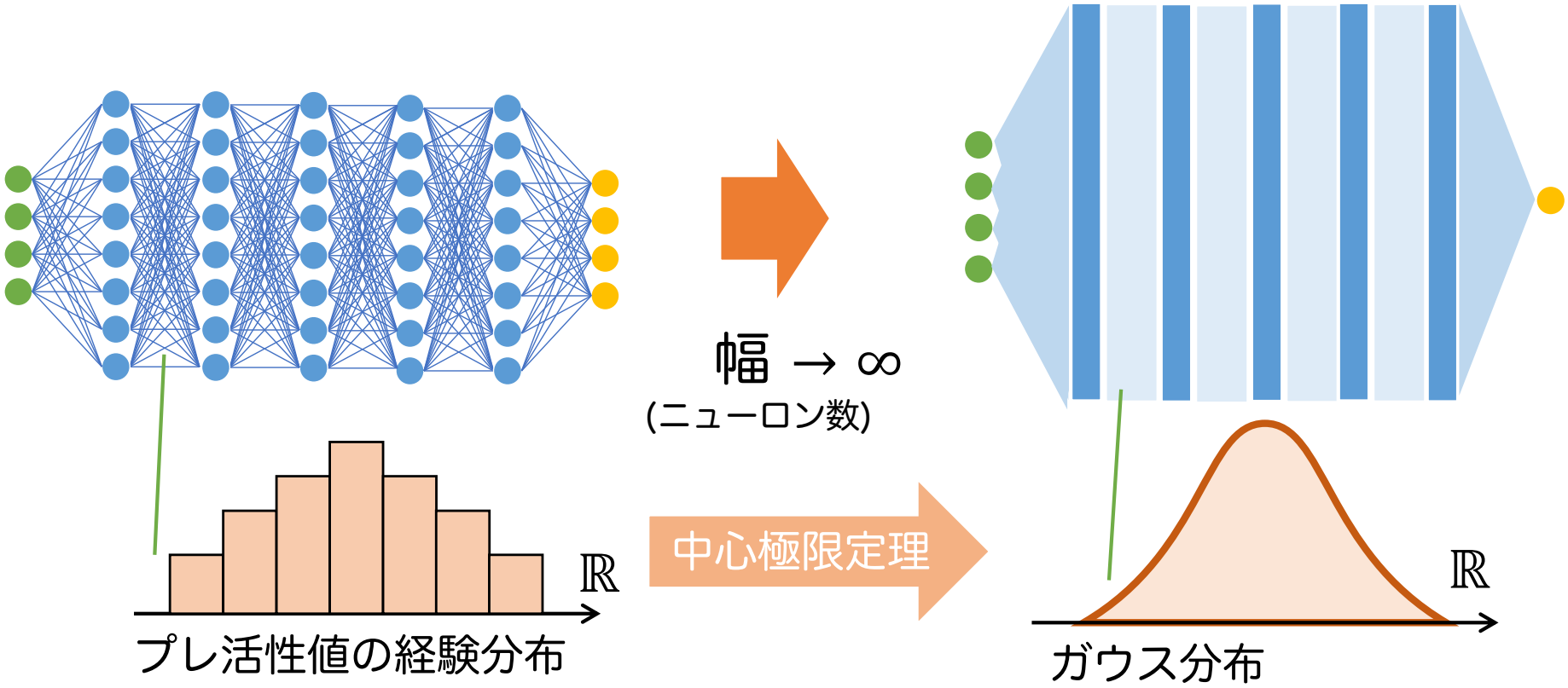
→過学習を説明できない

誤差 $\mathcal{E}(W^T)$ の精密な値を
W^T の構成から解析する
→詳細な解析が可能

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107-115.

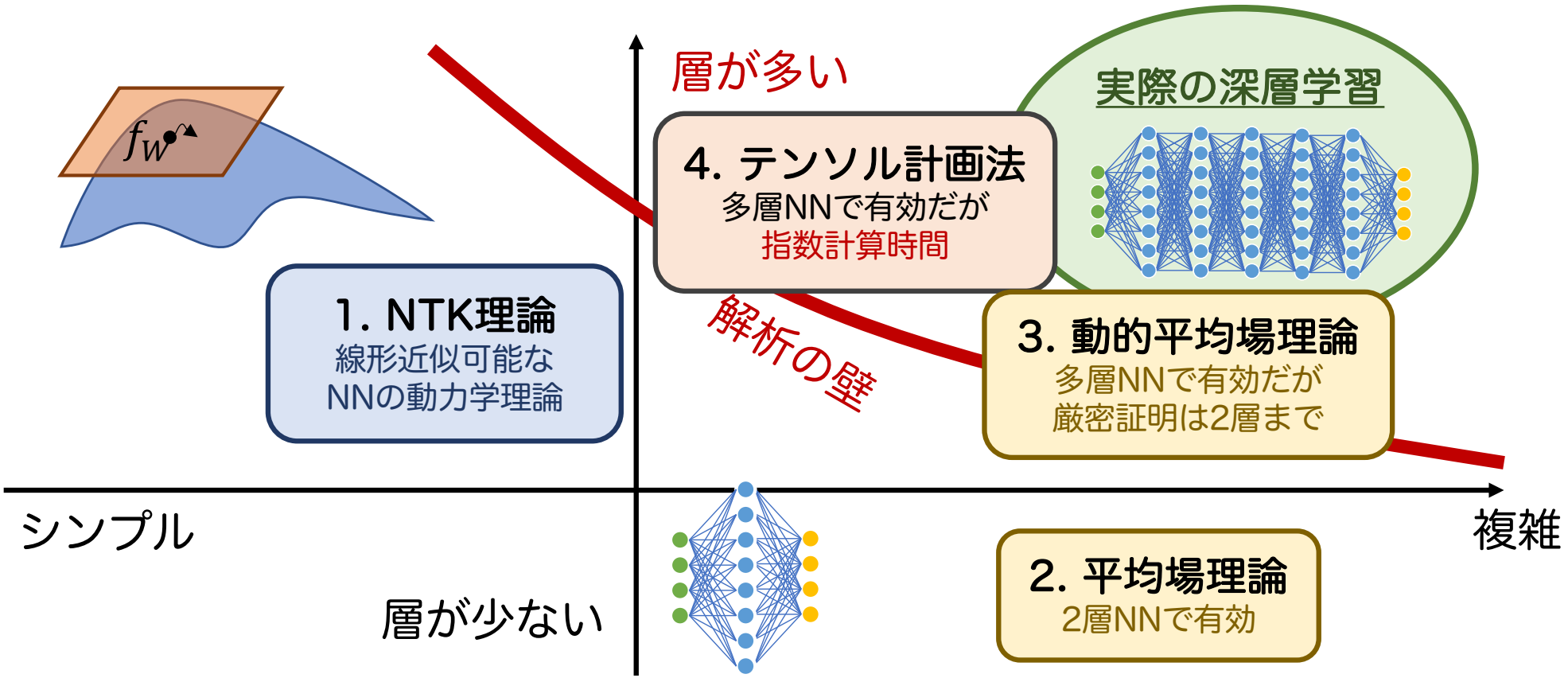
NNダイナミクス精密解析研究

- 各層のプレ活性値の分布を追跡する
  - ニューロン数を無限大にするとガウス分布(正規分布)に収束
    →ニューラルネットのマクロ的な性質を精密追跡



数学的に厳密なNNの精密解析研究

- 多層ニューラルネットのダイナミクスの厳密な解析は長い未解決問題



Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *NeurIPS*.
Song, M., Montanari, A., & Nguyen, P. (2018). A mean field view of the landscape of two-layers neural networks. *PNAS*.
Bordelon, B., & Pehlevan, C. (2022). Self-consistent dynamical field theory of kernel evolution in wide neural networks. *NeurIPS*.
Yang, G. (2020). Tensor programs iii: Neural matrix laws. arXiv.

一部は技術に応用

<u>実用化例</u>：AI学習のハイパーパラメータを選択

- パラメータを物理モデルで選択
  →選んだパラメータを深層学習(例:GPT)に用いる

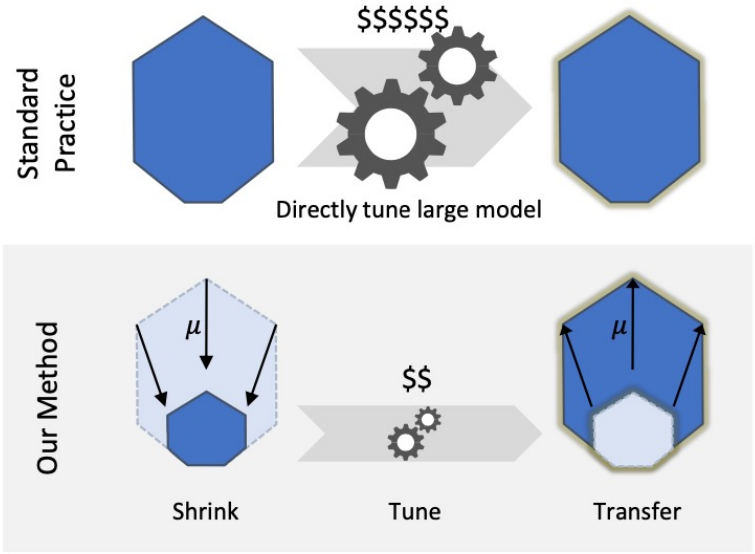


Figure 2: Illustration of μTransfer

ハイパーパラメータ選択は
計算コストが大きい

方程式の活用

物理モデルを用いて
選択してコスト削減

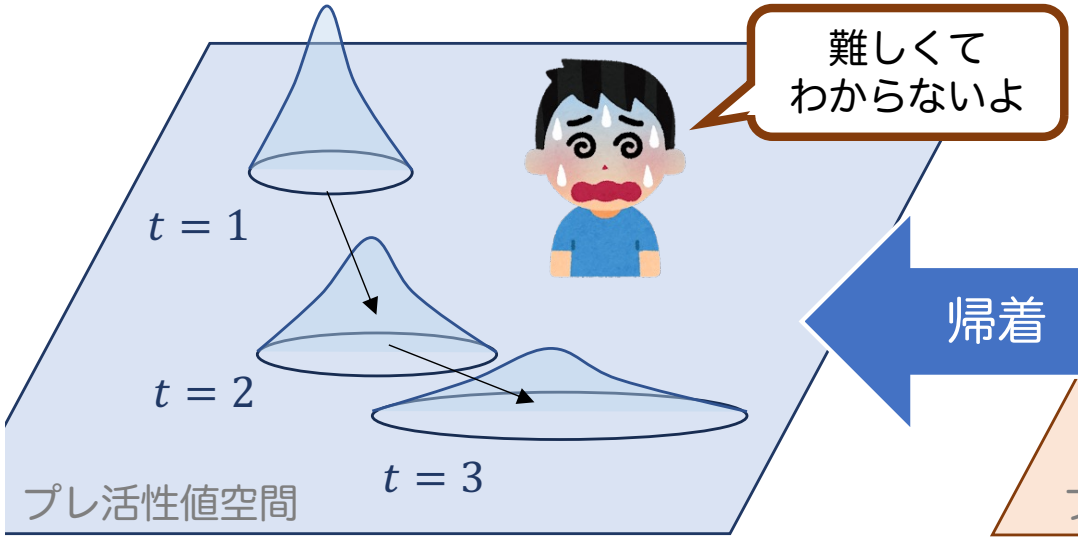
Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., ... & Gao, J. (2022). Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*.

t: 時間

キーテクニック：状態発展

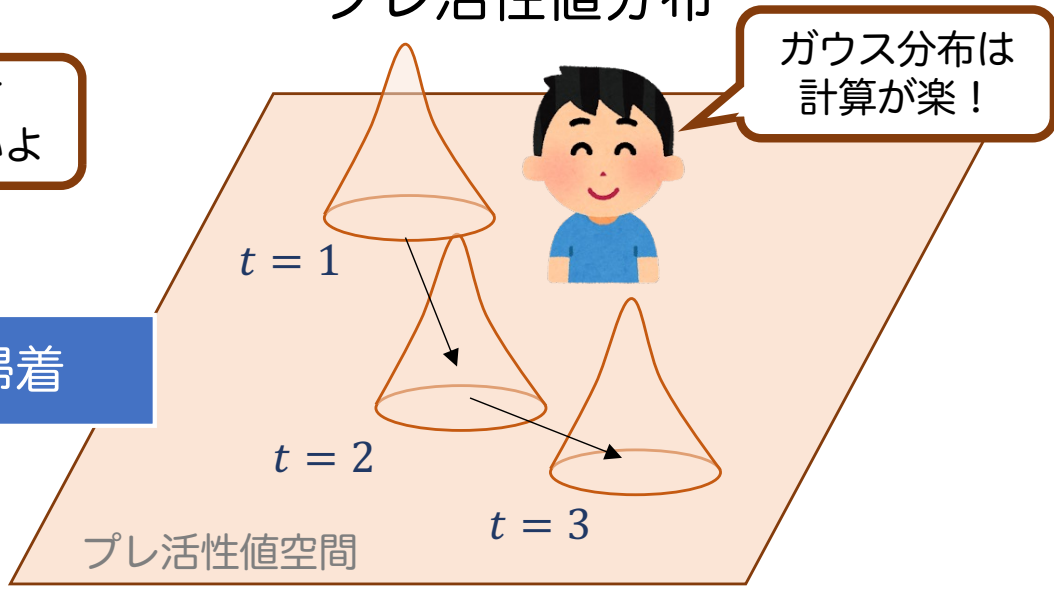
- ダイナミクス中の活性値分布は時間相関で複雑化
  →相関補正によりガウス分布列（状態発展）へ

勾配降下アルゴリズムによる
プレ活性値分布



時間相関によって
複雑(非ガウス)化

補正付きアルゴリズムによる
プレ活性値分布



補正（Onsager補正）により
常にガウス分布

帰着

補正の例：AMPアルゴリズム

- 再帰的なアルゴリズムの一種

設定：線形モデルのための二乗損失最小化

- $\mathbf{Y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times d}$: データベクトル・行列, $R(\cdot)$: 正則化

$$\min_{\theta \in \mathbb{R}^d} \|\mathbf{Y} - \mathbf{X}\theta\|^2 + R(\theta)$$

AMPアルゴリズム (Approximate Message Passing:近似メッセージ伝播法)

$$\theta^{t+1} = \eta(\mathbf{X}^\top \gamma^t + \theta^t)$$

η: 所与の非線形関数(Rで決まる)

$$\gamma^t = \mathbf{Y} - \mathbf{X}\theta^t + (d/n)\gamma^t \langle \eta_t'(\mathbf{X}^\top \gamma^{t-1} + \theta^{t-1}) \rangle$$

$\mathbf{Y} - \mathbf{X}\theta^t$: 残差

$\gamma^t \langle \eta_t'(\mathbf{X}^\top \gamma^{t-1} + \theta^{t-1}) \rangle$: Onsager補正項

Bayati and Montanari. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. IEEE TIT.
Feng et al. (2022). A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*.

GFOM：AMPの拡張

- 長期時間相関を補正
  →NN勾配法の強い
  非線形性を扱える

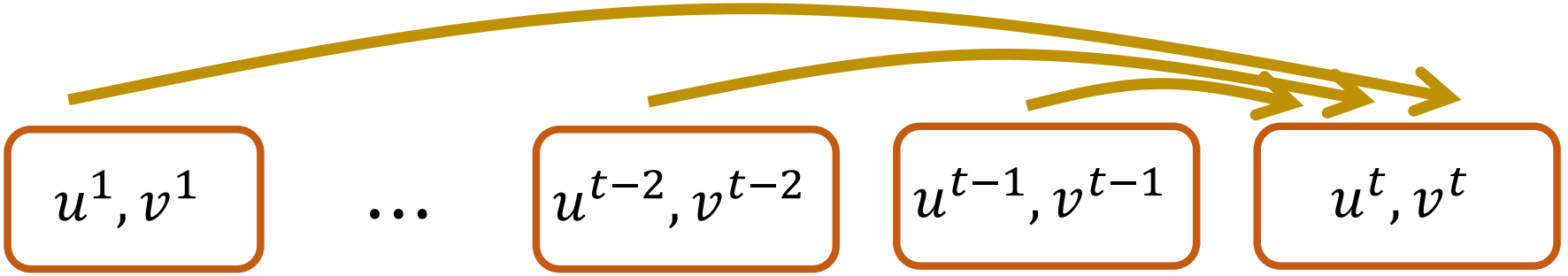
GFOM (General first order method)

$u^t \in \mathbb{R}^m, v^t \in \mathbb{R}^n$: 各時刻のベクトル

$\boldsymbol{A} \in \mathbb{R}^{m \times n}$: ランダム行列

F^1, F^2, G^1, G^2: 非ランダムな関数

$$u^t = \boldsymbol{A} F^1(v^{0:t-1}) + G^1(u^{0:t-1})$$
$$v^t = \boldsymbol{A}^\top G^2(u^{0:t}) + F^2(v^{0:t-1})$$



- G^1とF^2を過去全期間との補正になるよう設計
  → 行列版Onsager補正項

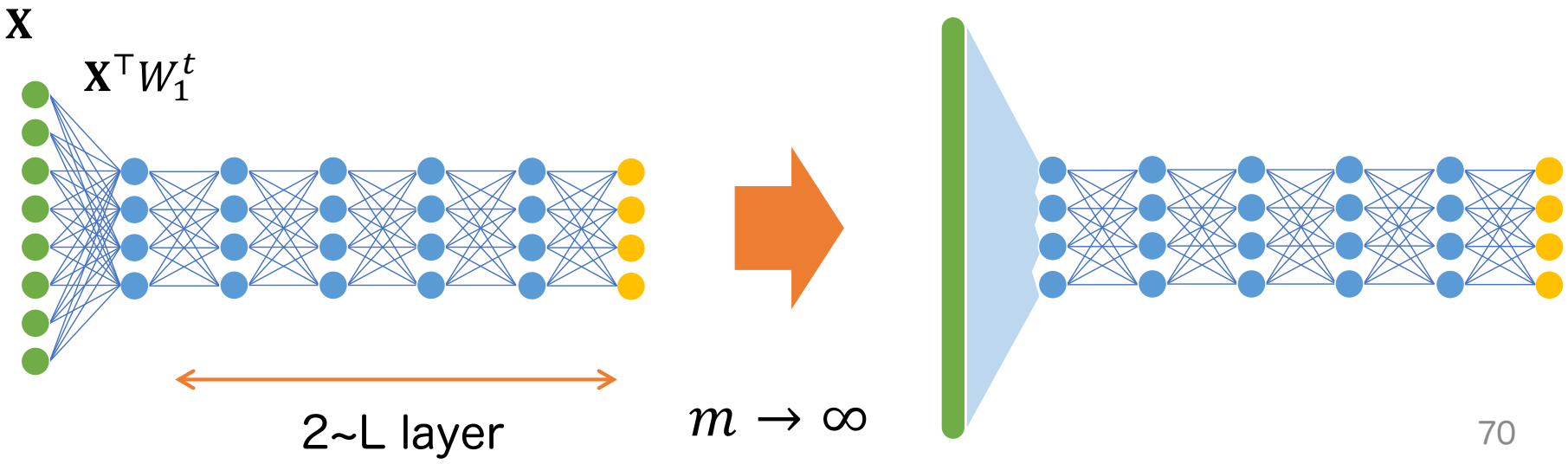
M.Celentano, A.Montanari, and Y.Wei, The Lasso with general Gaussian designs with applications to hypothesis testing, Ann. Statist. 51 (2023).

Q.Han and X.Xu. (2024). Gradient descent inference in empirical risk minimization. *arXiv preprint arXiv:2412.09498*.

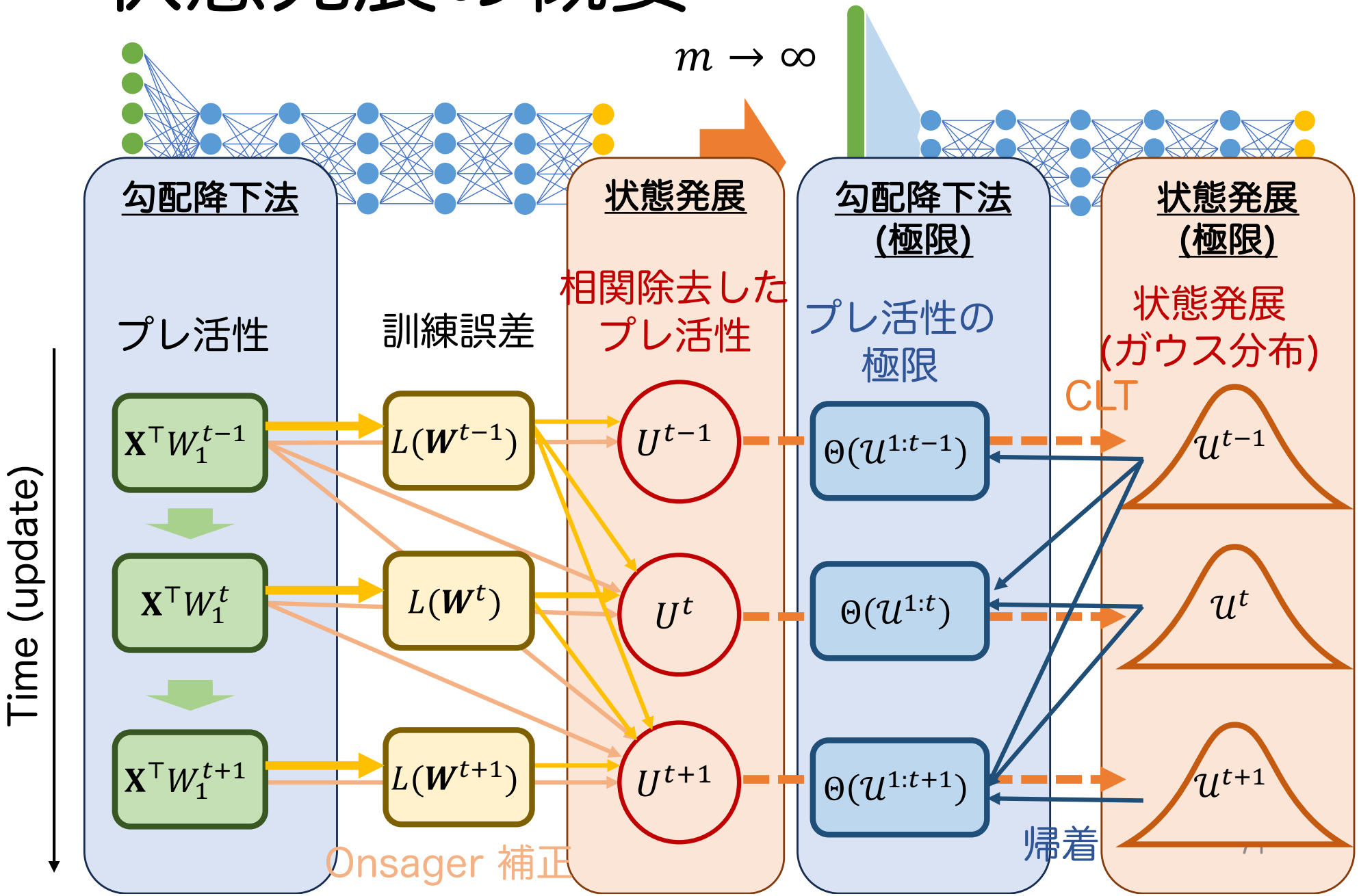
有限幅NNへのGFOMの適応

戦略

- 1. プレ活性値 $\mathbf{x}^\top W_1^t$ の成分分布の状態発展を記述
  - 一層目のみ
- 2. 二層目以降は通常の非線形変換を評価
  - ここでは状態発展は関係ない



状態発展の概要


$m \to \infty$
勾配降下法
状態発展
勾配降下法
(極限)
状態発展
(極限)
プレ活性
訓練誤差
相関除去した
プレ活性
プレ活性の
極限
状態発展
(ガウス分布)
CLT
Time (update)
$\mathbf{X}^\top W_1^{t-1}$
$\mathbf{X}^\top W_1^{t}$
$\mathbf{X}^\top W_1^{t+1}$
$L(\boldsymbol{W}^{t-1})$
$L(\boldsymbol{W}^{t})$
$L(\boldsymbol{W}^{t+1})$
U^{t-1}
U^{t}
U^{t+1}
$\Theta(\mathcal{U}^{1:t-1})$
$\Theta(\mathcal{U}^{1:t})$
$\Theta(\mathcal{U}^{1:t+1})$
$\mathcal{U}^{t-1}$
$\mathcal{U}^{t}$
$\mathcal{U}^{t+1}$
Onsager 補正
帰着

成果例2：学習の動力学

- 得られた方程式（主部分）

細かい部分は
お気になさらず

時間相関の補正行列（Onsager補正）

$$\rho_{t,s} = I_q 1_{s=t} + \sum_{r\in[s+t:t]} (\tau_{t,r} + I_q 1_{r=t})\rho_{r-1,s}$$

補正後の1層目のニューロン分布

$$U^t := \mathbf{X}W_1^t + \phi^{-1}\eta \sum_{s=1}^{t} \left[\left(\nabla_{\mathbf{X}^\top W_1} L(\boldsymbol{W}^{s-1})\right) \cdot \sigma'(\mathbf{X}^\top W_1^{s-1})\right]\rho_{t,s}^\top$$

状態発展 $\mathcal{U}^t$ **（**q**次元ガウス過程）**

$$\mathrm{Cov}(\mathcal{U}^t, \mathcal{U}^s) = \sum_{r\in[1:t], r'\in[1:s]} \rho_{t,r}\Sigma_{r,r'}\rho_{s,r'}^\top$$

状態発展に基づくプレ活性値の極限

$$\Theta(\mathcal{U}^{0:t}) := \mathcal{U}^t - \phi^{-1}\eta \sum_{s=1}^{t-1} \left(\nabla_{\mathbf{X}^\top W_1 = \Theta(\mathcal{U}^{0:s})} L(\boldsymbol{W}^{s-1})\right) \cdot \rho_{t-1,s}^\top$$

仮想ガウス化
アルゴリズム部分

学習アルゴリズム
への帰着部分

Q.Han, M.Imaizumi, "Precise gradient descent training dynamics for finite-width multi-layer neural networks", arXiv:2505.04898

結果：要素分布の極限定理

- プレ活性値分布がガウス分布の変換 $\Theta_\cdot(\mathcal{U}^{1:t})$ に収束

主定理 (Han and Imaizumi 2025)

仮定

- $\mathbf{X}$ の各要素は独立・subGaussian・平均ゼロ・分散$O(1/d)$
- 活性化関数 $\sigma(\cdot) \in C^4$

各 $t \geq 1$ ごとに、全ての $r \geq 1$ と準リプシッツ関数 ψ のもとで

$$E\left[\left\| m^{-1}\sum_{j=1}^{m}\psi\left([\mathbf{X}^\top W_1^{t-1}]_j\right) - E\left[\psi\left(\Theta_\cdot(\mathcal{U}^{1:t})\right)\right]\right\|^r\right] = O(n^{-c}).$$

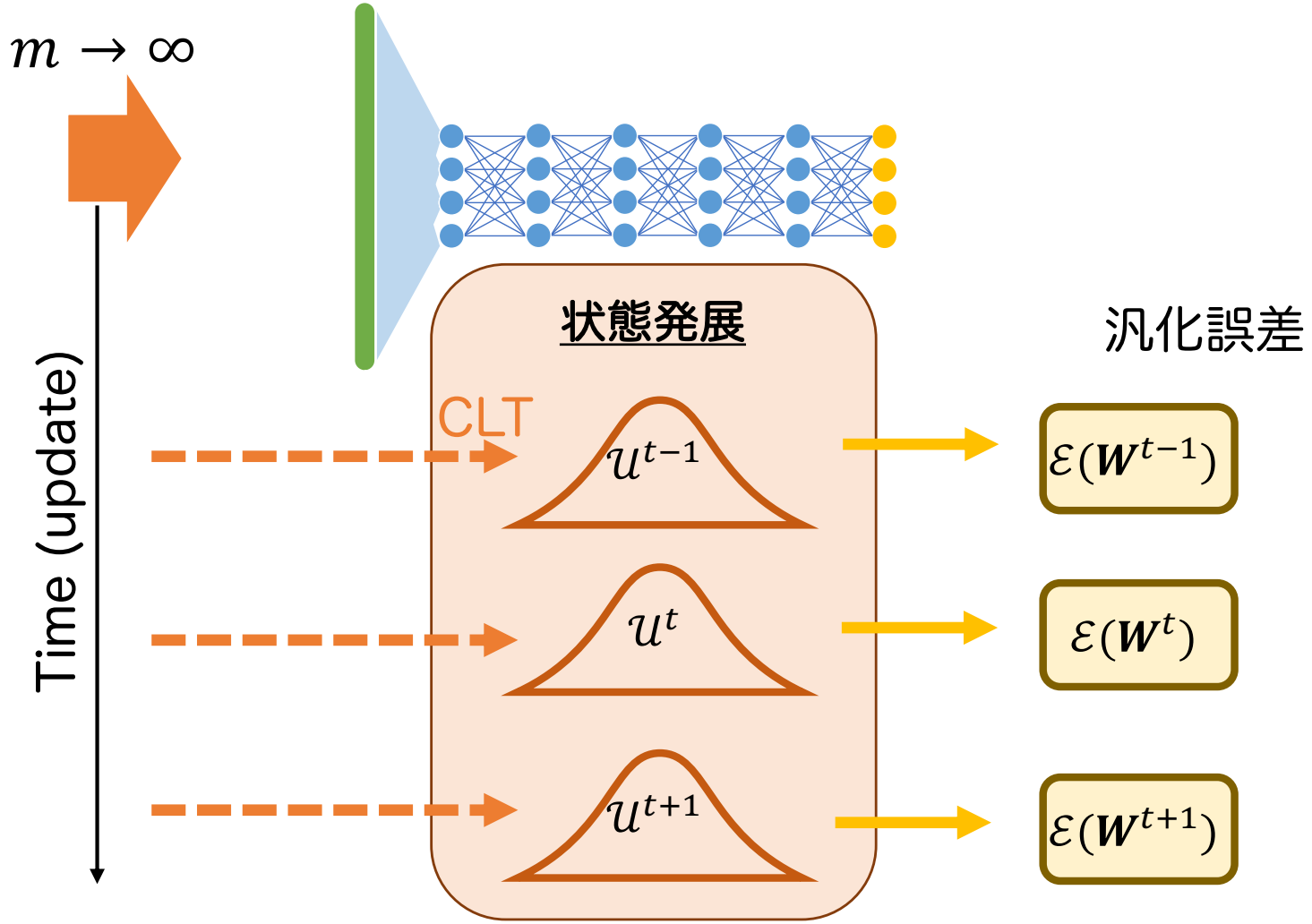
プレ活性化値の成分分布

理論で導出した分布

$c = c(t, q, L, r) > 0$: ある定数

汎化誤差の計算

- $\mathcal{U}^t$ は訓練データに依存しない
  →汎化誤差 $\mathcal{E}(\boldsymbol{W}^t)$ の導出が可能



結果：汎化誤差の精密評価

- $\mathcal{U}^t$ をデータとして用いて順伝播

汎化誤差定理 (Han and Imaizumi 2025)

仮定

- $\mathbf{X}$ の各要素は独立・subGaussian・平均ゼロ・分散$O(1/d)$
- 活性化関数 $\sigma(\cdot) \in C^4$

各 $t = 1,2, \ldots$ ごとに、関数 $\mathcal{R}$ のもとで,

$$|\mathcal{E}(\boldsymbol{W}^t) - E[\mathcal{R}(\mathcal{U}^{t+1}, \mathcal{U}^0)]| = O(n^{-c}).$$

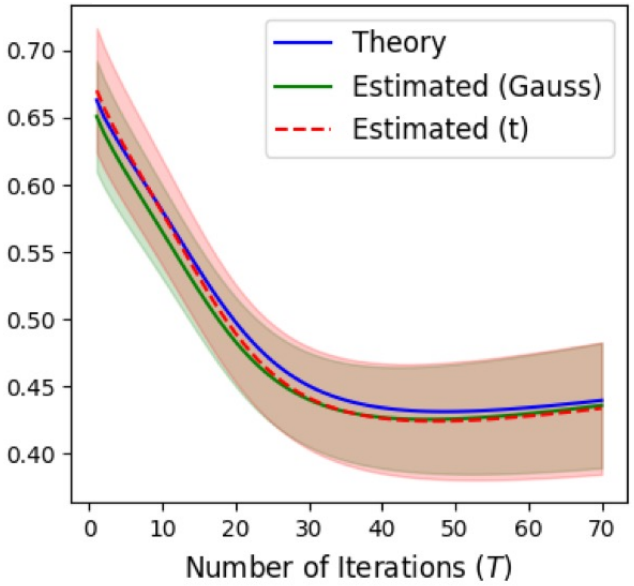
深層NNの汎化誤差　　理論で導出した値

- $\mathcal{R}(u, u') \approx (f_{\mathbf{W}^t}(u) - \varphi(u') - \xi)^2$: 近似した残差
- これより（テストデータを使わないで）汎化誤差を計算可
  - 計算時間は$O(t^2)$

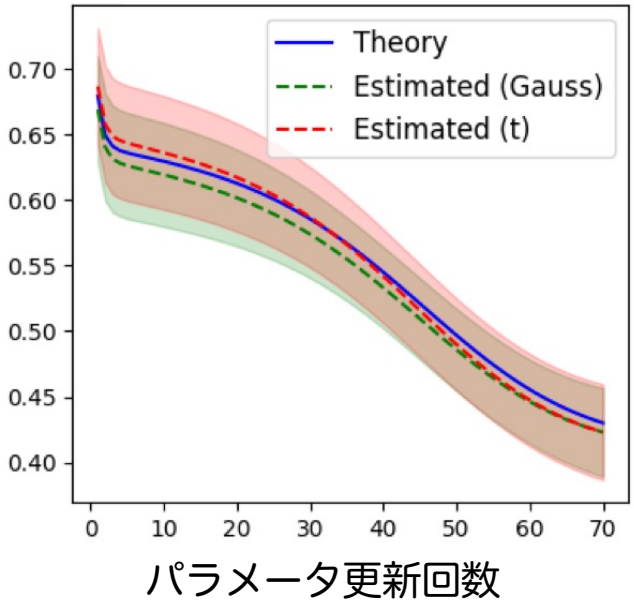
学習ダイナミクスの再現を実証

- 理論と結果の整合性の確認
  - 青：実際のアルゴリズムによる値
  - 緑・赤：理論に基づいた推定量の値

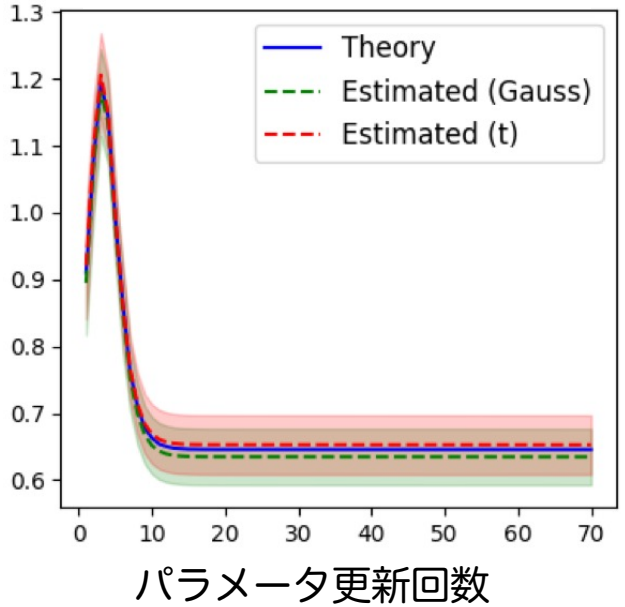
予測誤差(2層)



予測誤差(3層)



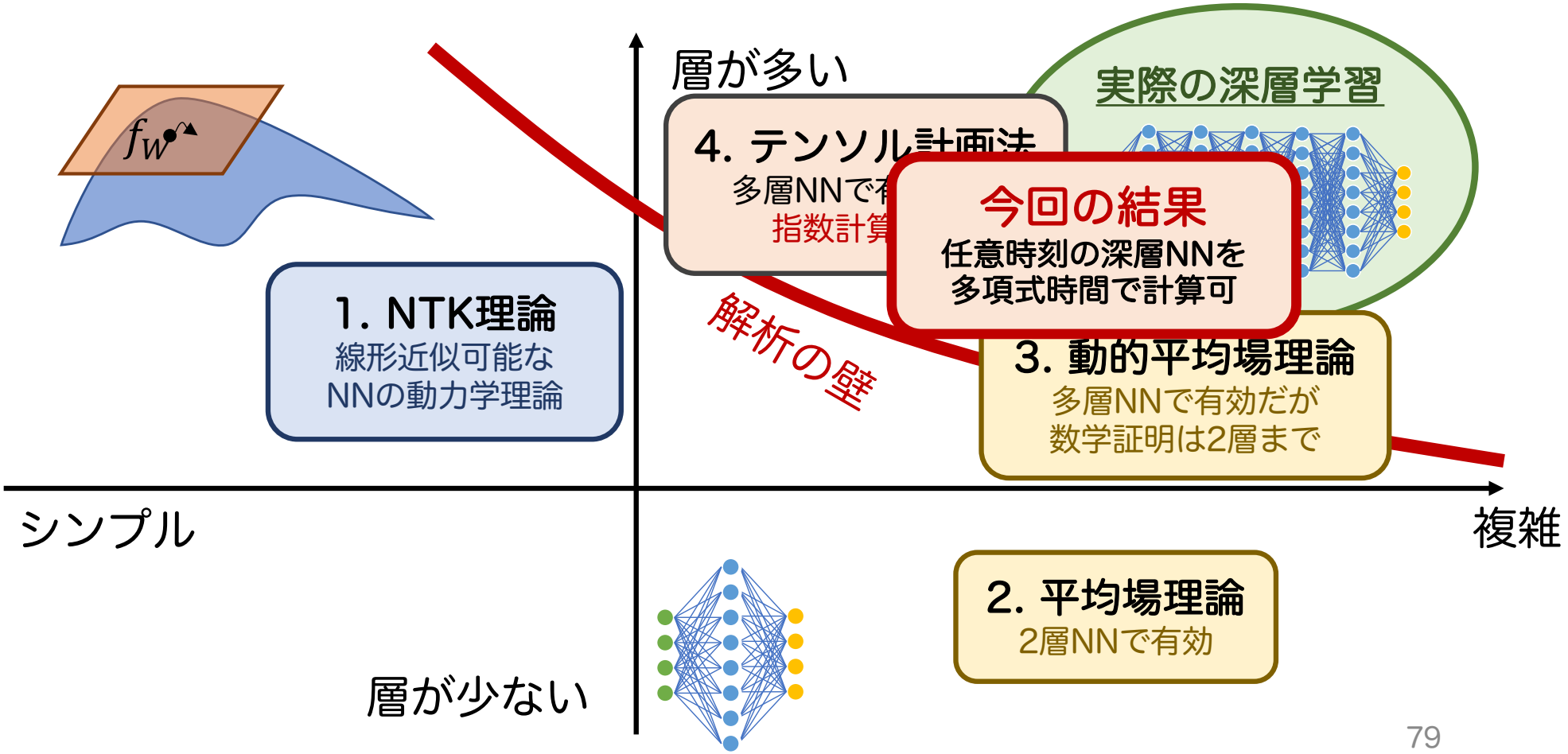
予測誤差(5層)



Q.Han, M.Imaizumi, "Precise gradient descent training dynamics for finite-width multi-layer neural networks", arXiv:2505.04898

研究の立ち位置

- 層の多いNNの複雑な性質を精密記述できた



まとめ

なぜAIは高い性能を発揮できるか？

多層ニューラルネットの役割

→局所関数・データ基本構造の効率的学習

数学的な解析の貢献と限界

→過学習などはアルゴリズム近似などに限界

物理学的なアプローチ

→精密解析による新しい発見と理論化

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

ご静聴ありがとうございました。