深層学習の原理を明らかにする理論の試み

2021/10/21

数学カフェ(https://mathcafe.net)

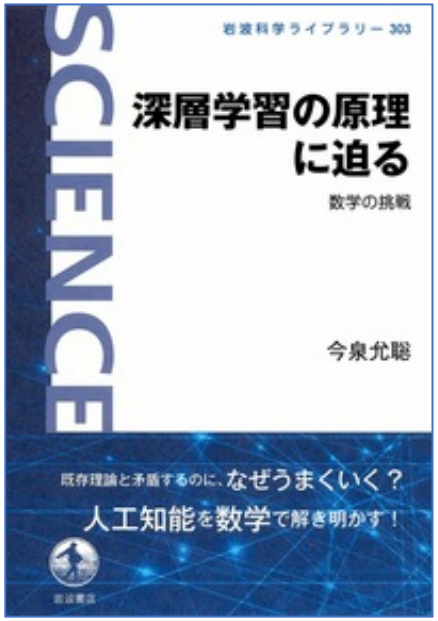
今泉允聡（東京大学）

自己紹介

今泉允聡（いまいずみ まさあき）

経歴

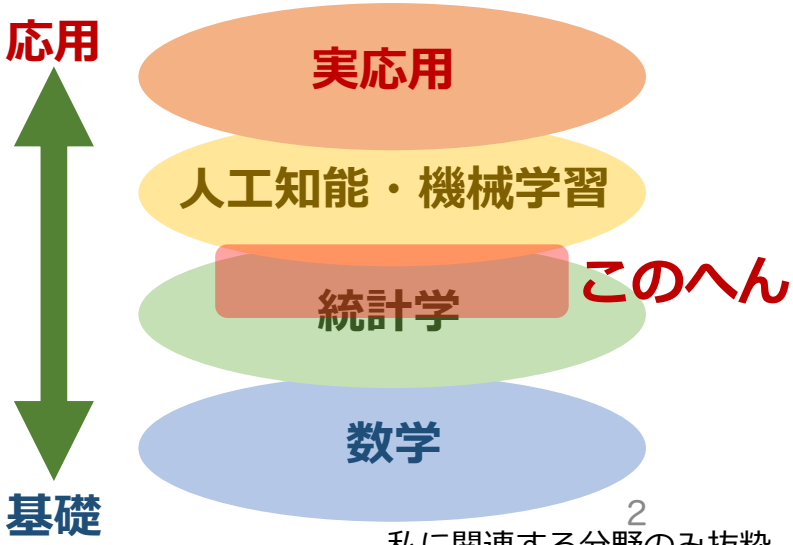
- ~2017 東大 経研 統計専攻（博士）
- ~2020 統計数理研究所 助教
- 2020~ 東大 相関基礎/先進機構 准教授
  （物性理論・統計力学G）
  (兼) 理研AIPセンター、JSTさきがけ



最近出版した紹介書

研究領域

- 統計学
  - 複雑データ、中心極限定理
- 機械学習
  - 深層学習、テンソル



私に関連する分野のみ抜粋

導入：深層学習の”発見”

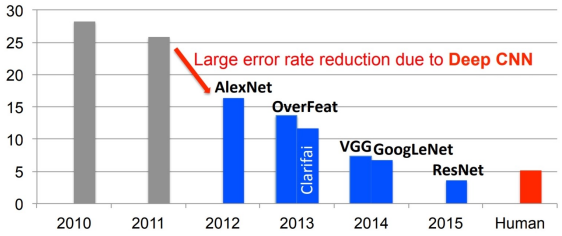
深層学習の”発見”

基礎研究 → **ブレイクスルー** → **実用化の進展**

~2000 | 2012 | 2016~

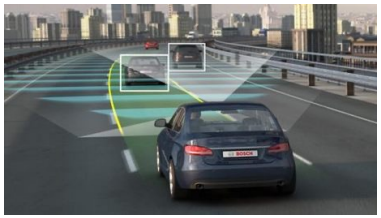
技術的課題

技術発展

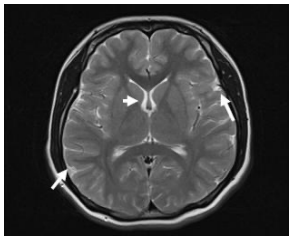


高精度を発揮

画像識別精度: 75% → 96%



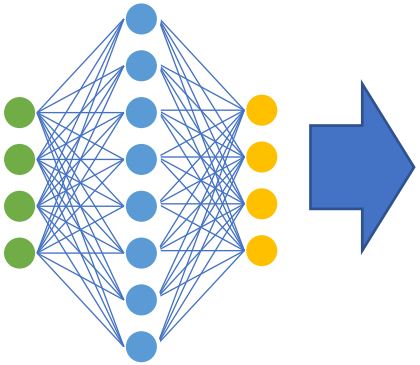
自動運転



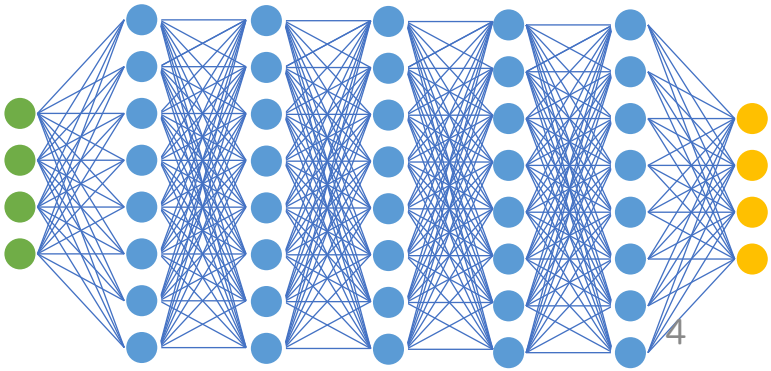
医療診断

発見：統計モデルの層を増やすと精度が大きく向上

（既存法）
浅層モデル



（新手法）
深層学習



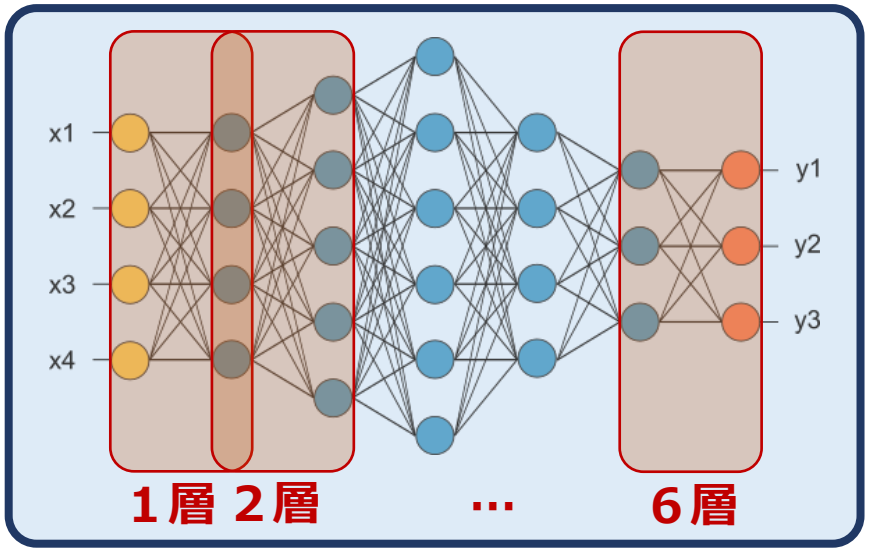
深層学習とは

多層ニューラルネットワークによる統計解析

・変換を層の数だけ繰り返す関数モデル

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

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	$y = A_6 z_5 + b_6$

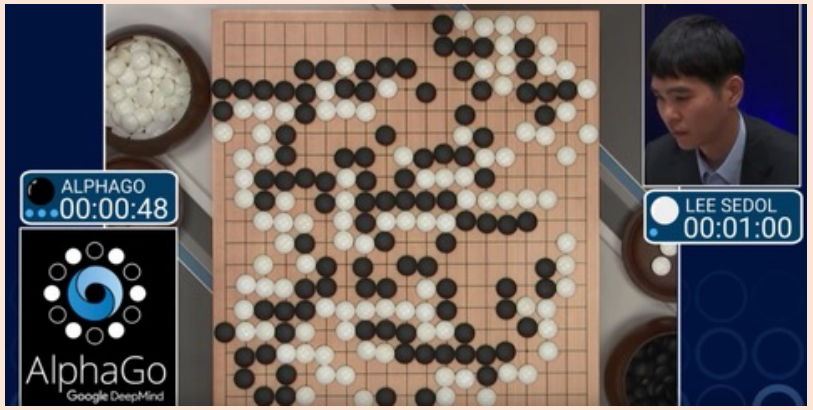
A: パラメタ (行列)
b: パラメタ (ベクトル)
σ: 非線型変換

深層学習の成功例

深層学習の有名な成功例

AlphaGo (DeepMind)

- 囲碁で人間超え
  - 世界トップ棋士に勝利



BERT (Google)

- 言語テストで高得点
  - 人間: 91.2%
  - BERT: 93.2%

文章

In early 2012, NFL Commissioner Roger Goodell stated that the league planned to make the 50th Super Bowl "spectacular" and that it would be "an important game for us as a league".

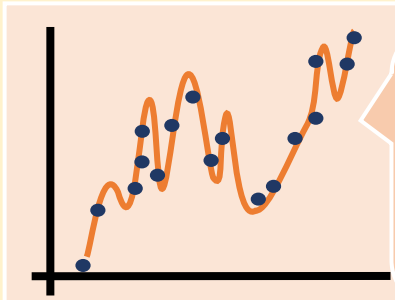
文章に関する問い

Who was the NFL Commissioner in early 2012?
Ground Truth Answers: Roger Goodell | Roger Goodell | Goodell

予測性能をめぐる理論の謎

従来理論と実際の深層学習は完全に食い違う

従来の統計・学習理論

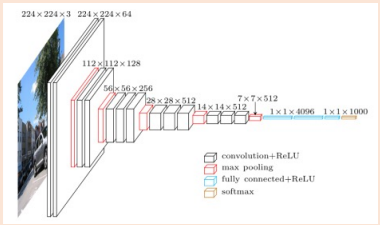


過剰な
パラメタは
過学習する
→性能悪化

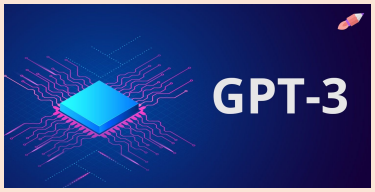
$$\text{誤差} = O\left(\frac{\text{パラメタ数}}{\text{データ数}}\right)$$

矛盾

巨大深層学習の成功



VGG19 Net
1億パラメタ



GPT-3
1兆パラメタ

パラメタを増やすほど
予測精度が向上

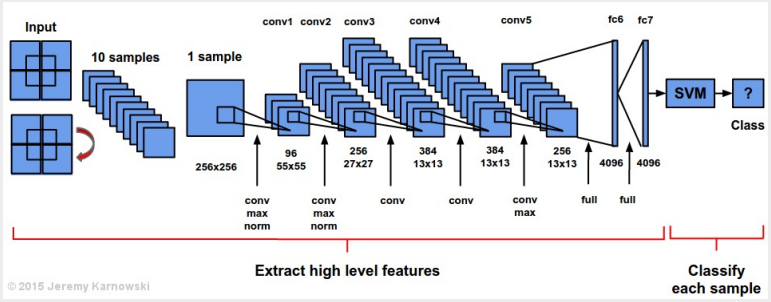
汎化再考：汎化（予測）の理論の再構成が必要

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *ICLR*.

巨大化する深層学習モデル

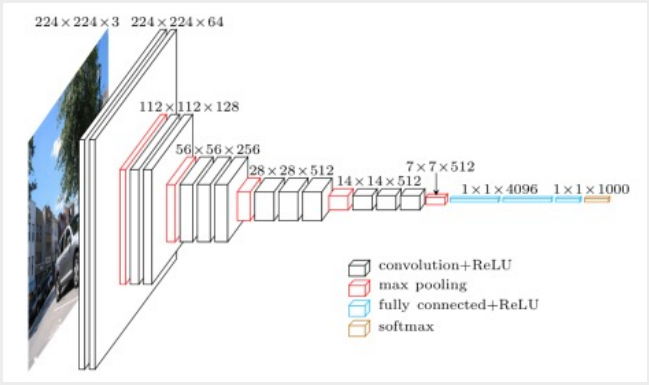
深層化に伴いパラメータ数も膨大化

AlexNet (Toronto U)



層の数：８層
パラメタ数：６千万

VGG19 Net (Oxford U)

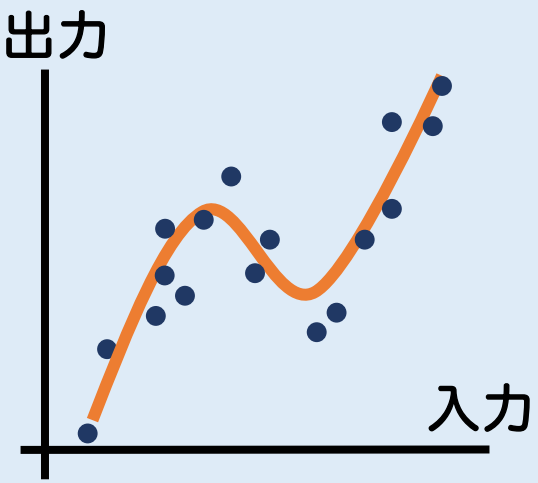


層の数：１９層
パラメタ数：１億

モデルの大きさが、深層学習前の数千倍規模に

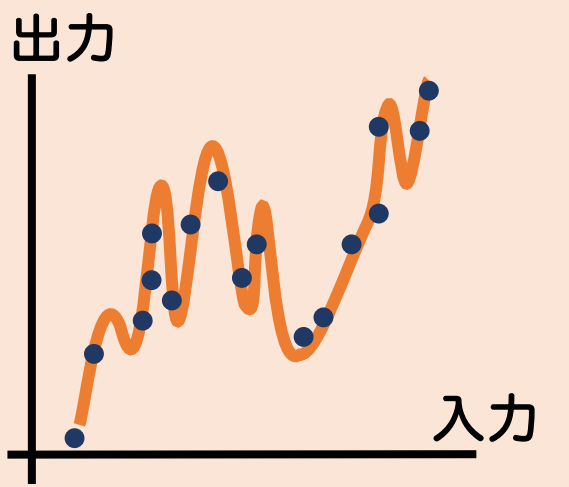
巨大モデルは過適合するはずだった

統計・学習理論の原則
大量のパラメタは精度を下げる！



適切な関数を学習

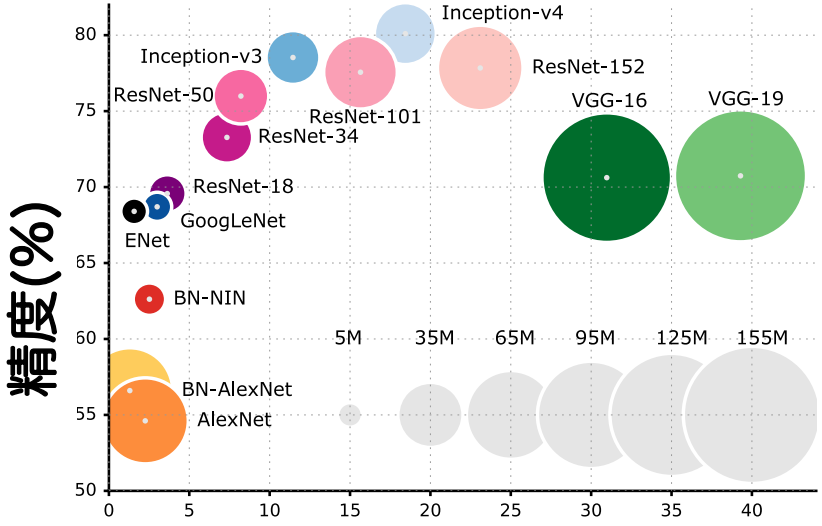
パラメタ数が
増えると…



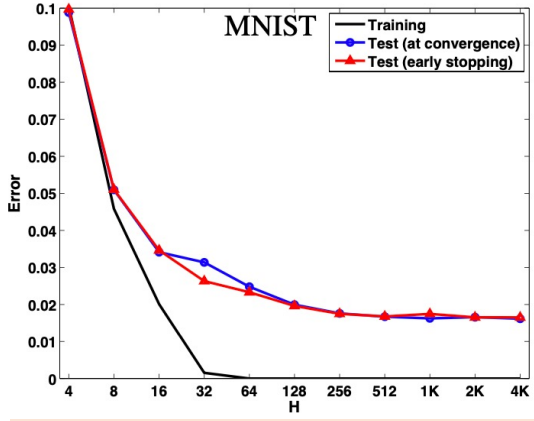
過適合した関数を学習
➡ **精度の低下**

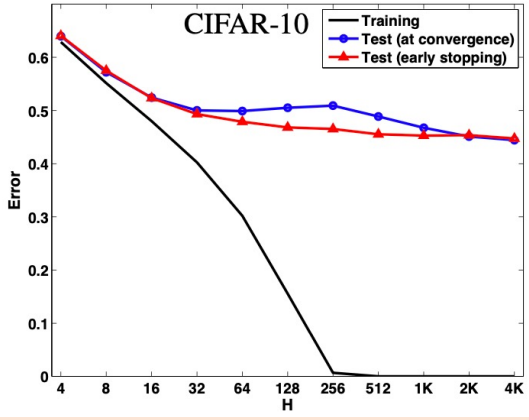
- 既存の理論はパラメタ数の削減に腐心…
  - 変数選択、スパース推定、正則化、適応化など

実際の深層学習は過適合しない



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



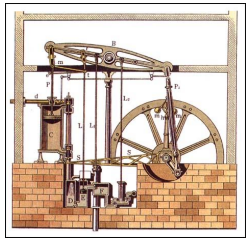


実データの実験結果
ニューラルネットワークのサイズ
（横軸）の拡大に伴って
汎化誤差（赤線・青線）が減少

パラメタ数が増えているのに汎化誤差（期待損失）が減少
➡ 既存理論と完全に矛盾

“発見”を理論で記述すること

- 歴史的には共通の現象

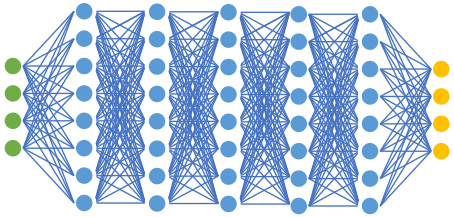


蒸気機関の発明
（1769年）

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？

理論の枠組み

“汎化誤差”で予測性能を測る

ニューラルネットワーク

- 関数 $f(X;\theta)$
  - θは全ての層のパラメータ(A_ℓ, b_ℓ)を並べたもの

学習の設定

- 学習データ（n個の入出力のペア） $(X_1, Y_1), \ldots, (X_n, Y_n)$
- 損失関数：クロスエントロピー・二乗関数など $\ell\big(Y, f(X;\theta)\big)$

- **経験誤差**（学習データ上の損失）

$$L_n(\theta) = n^{-1} \sum_{i=1}^{n} \ell(Y_i, f(X_i;\theta))$$

- **汎化誤差**（損失の期待値）

$$L(\theta) = E[\ell(Y, f(X;\theta))]$$

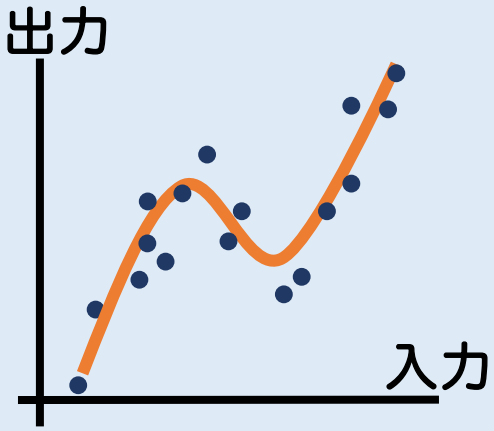
  - 予測誤差の期待値（最終的に小さくしたい値）
    - 学習データには全く依存しない

期待値を考える
＝全ての可能な
データ上の平均

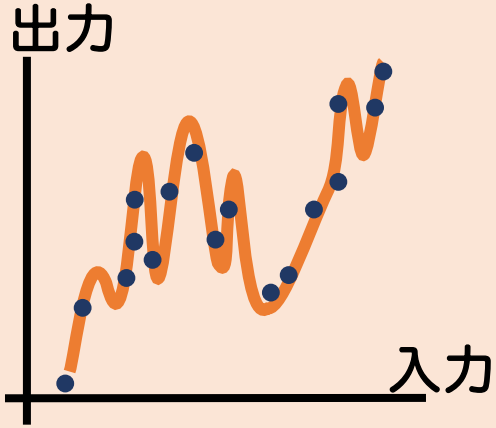
どういう時に予測を間違えるのか？

学習後に予測に失敗＝学習データに**過剰に適合**している

- ある程度の表現力は必要だが、高すぎるとかえって危険



適切な関数を学習



データに過剰に適合

過適合（過学習）：
学習データに過剰に適合し、未知のデータへの予測が弱くなること

え、表現力が高いと
過学習するの？
どうすれば良いんだ…

どんな時に過学習が
起こるかを理解する必要
（＝汎化を理解する）

汎化誤差は
経験誤差＋汎化ギャップ

汎化誤差の要素分解

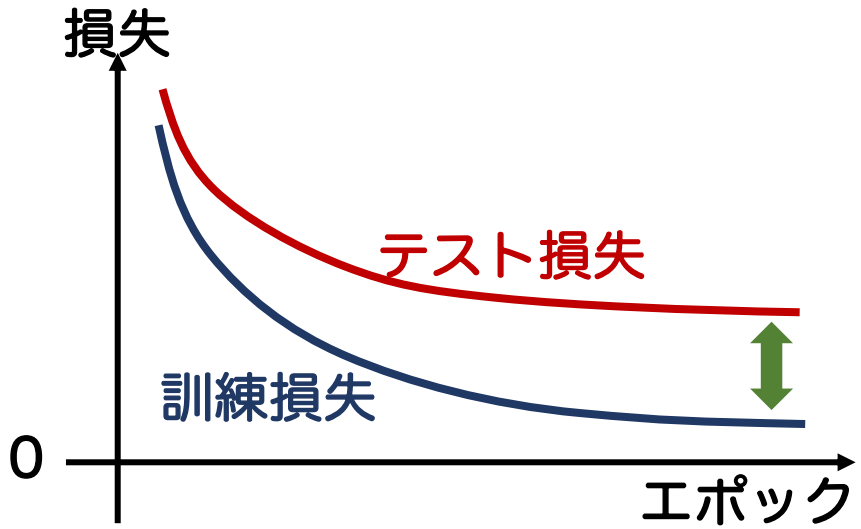
※最適化誤差を加える場合もある

$$L(\theta) = L_n(\theta) + \left(L(\theta) - L_n(\theta)\right)$$

最終的に小さくしたい　アルゴリズムで小さくできる　汎化ギャップ（過適合の程度）

今日の主たる関心

- 実際の計算との対応



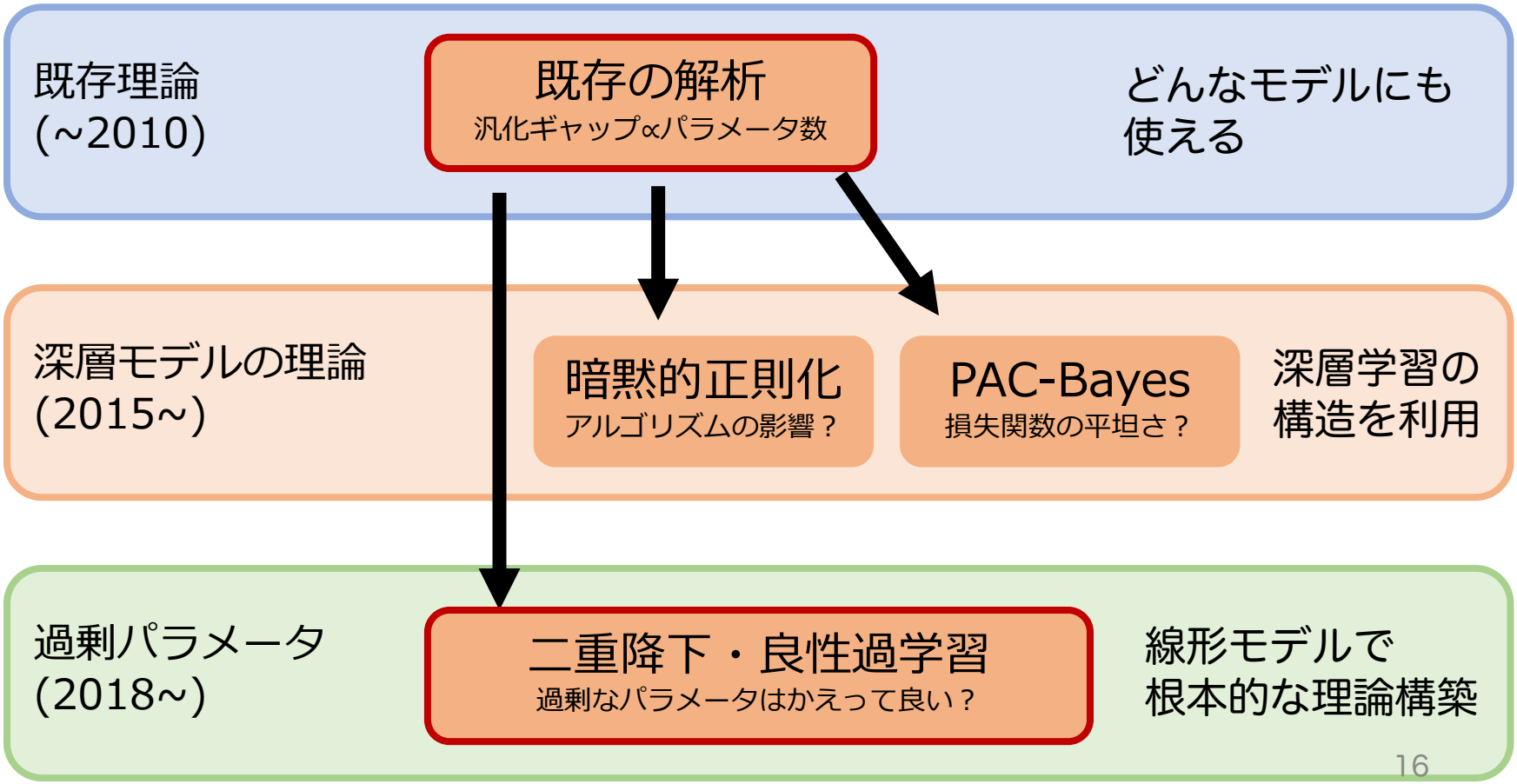
テスト損失(test loss) ≈汎化誤差$L(\theta)$
（学習データ以外のデータ上の平均）

訓練損失(train loss) ＝ 経験誤差$L_n(\theta)$

← 汎化ギャップ

今日のトピック

- 汎化ギャップをめぐる研究の流れ

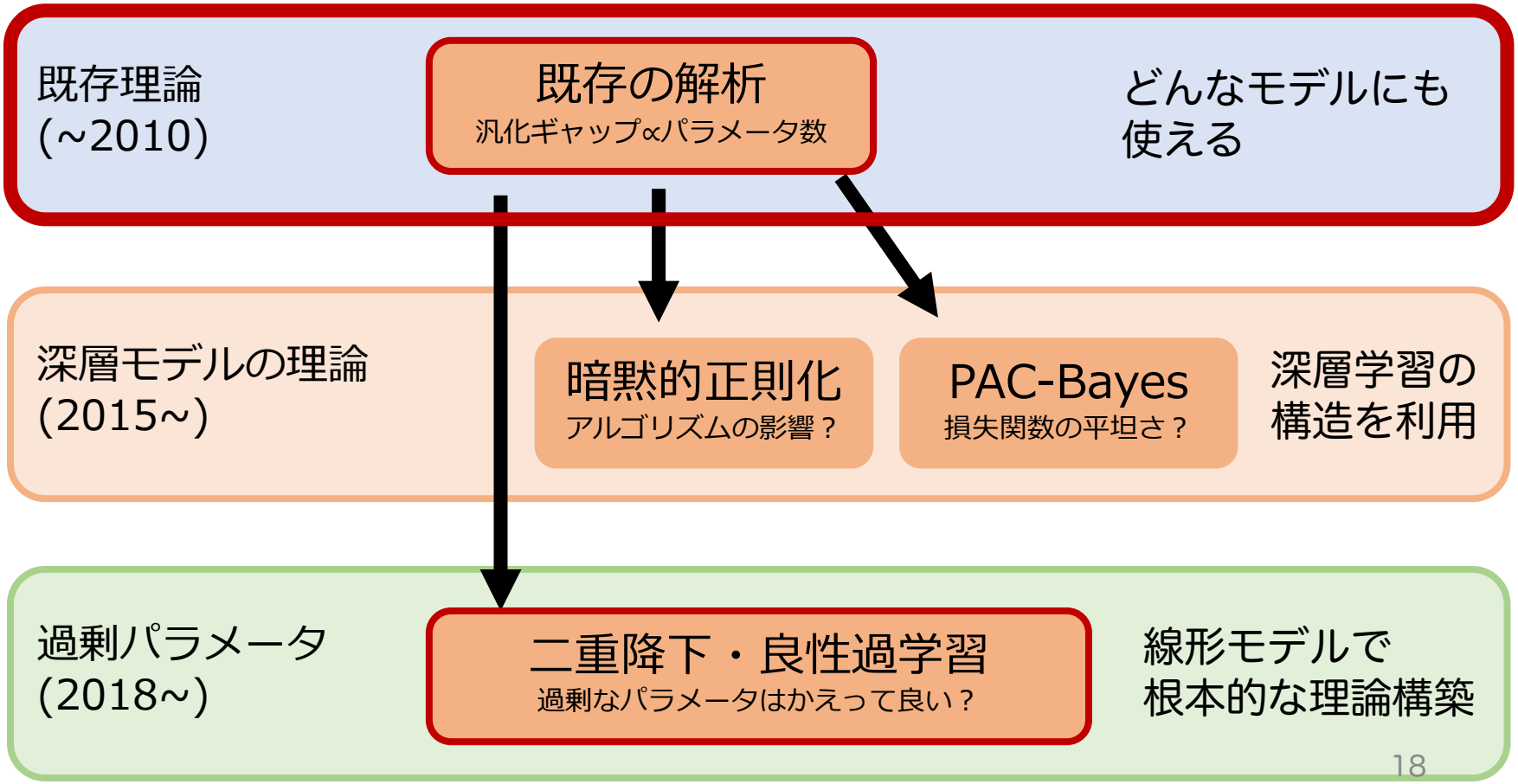


既存理論

汎化ギャップをパラメータ数で説明

今日のトピック

- 汎化ギャップをめぐる研究の流れ



準備：パラメタの学習方法

(確率的)勾配降下法で学習

- 損失関数の勾配を降ることで良いパラメータを探索

アルゴリズムの定式化

勾配降下法

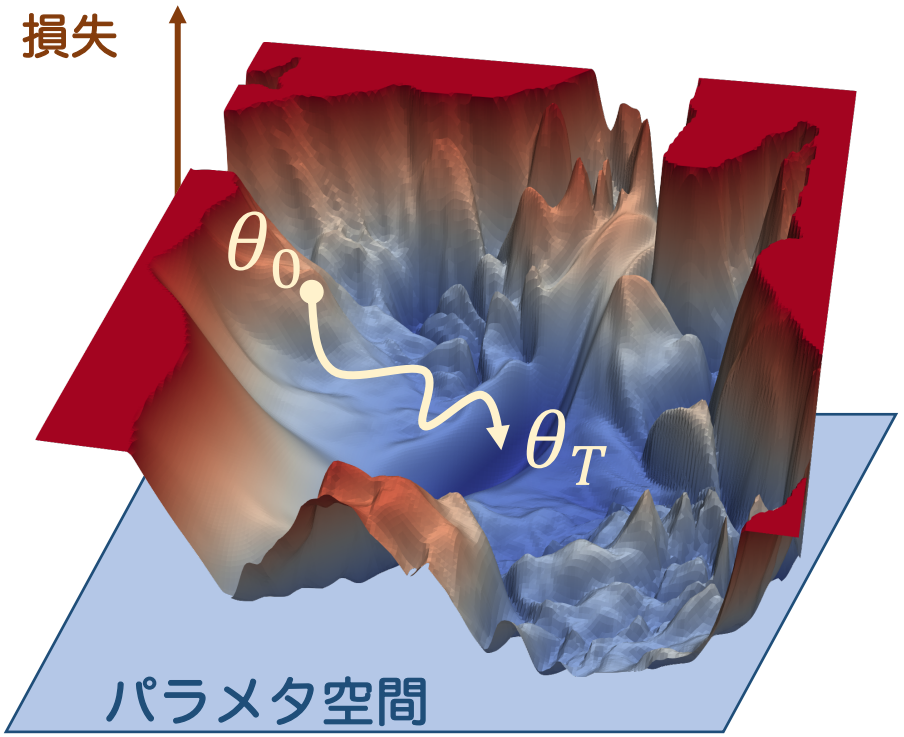
θ_0: 初期値

$$\theta_{t+1} = \theta_t - \eta \nabla L_n(\theta_t),$$

($\eta > 0$: 学習率, $t = 1, \ldots, T$)

確率的勾配降下法

$\nabla L_n(\theta)$を小標本で推定



可視化（次元圧縮）された深層学習の損失関数 (Li et al. 2018)

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. NIPS.

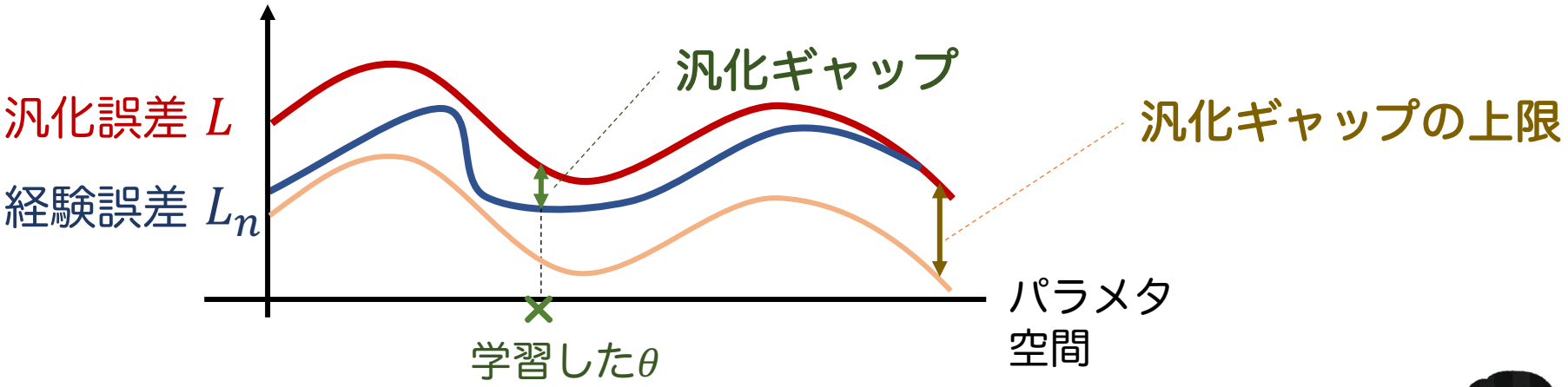
汎化ギャップの調べ方

基本的な考え：汎化ギャップの上限を求める

$$\big(L(\theta) - L_n(\theta)\big) \leq \sup_{\theta'}\big(L(\theta') - L_n(\theta')\big)$$

学習したパラメータθの
汎化ギャップ

可能なパラメータ全て上の
最も大きいギャップ



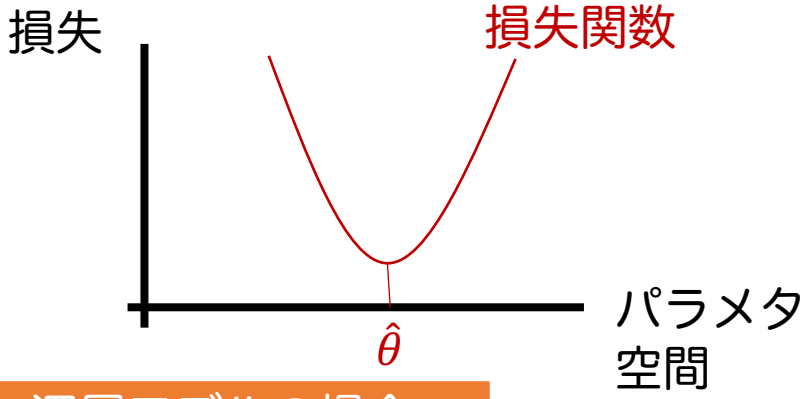
上限がわかればいいの？
知りたい汎化ギャップより
大きくなってない？

それはそうなんだけど、
ギャップ自体を数学的に
評価するのは難しく、
代替案はあまりないんだ

なぜ上限を求めるのか？

- 深層モデルは確率的な変動の記述を難しくする

線形モデルの場合

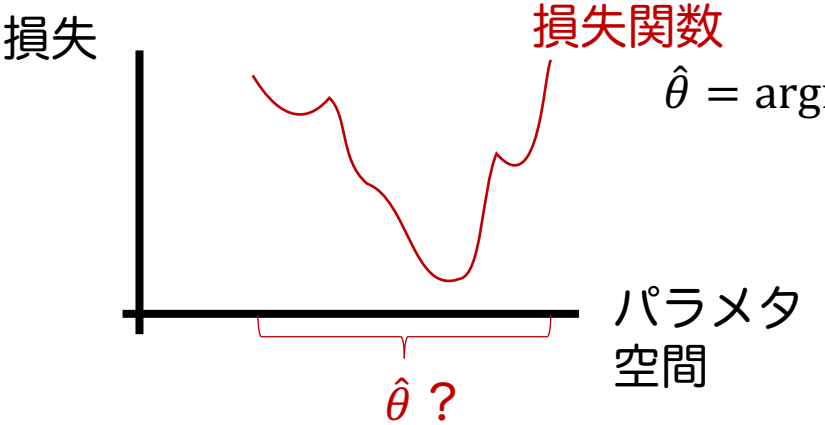


$$\hat{\theta} = \mathrm{argmin}_{\theta} \sum_{i} (Y_i - \theta X_i)^2$$

$$\Rightarrow \hat{\theta} = \left(\Sigma_i X_i^2\right)^{-1} \Sigma_i X_i Y_i$$

簡単にかける！
分かる！

深層モデルの場合



$$\hat{\theta} = \mathrm{argmin}_{\theta=(A_\ell)_{\ell=1}^{L}} \sum_{i} (Y_i - A_L \sigma A_{L-1} \sigma \cdots \sigma A_1 X_i)^2$$

$$\Rightarrow \hat{\theta} = ?????$$

何も分からない
数値解しかない..

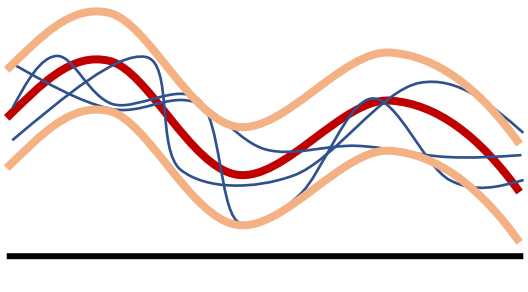
上限の値を知るには

- 二段階のステップで評価する
  - とりうるパラメータθがどれだけ多いかで評価

導出の流れ

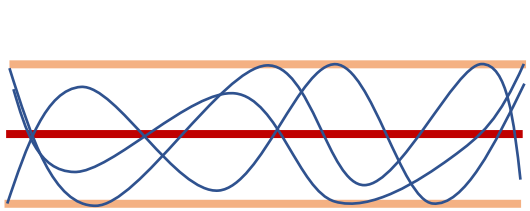
上限

$$\sup_{\theta} |L(\theta) - L_n(\theta)|$$



Rademacher複雑性

$$n^{-1/2}\mathrm{E}\left[\sup_{\Theta} \Sigma_{i=1}^{n} \sigma_i \ell\left(Y_i, f(X_i;\theta)\right)\right]$$

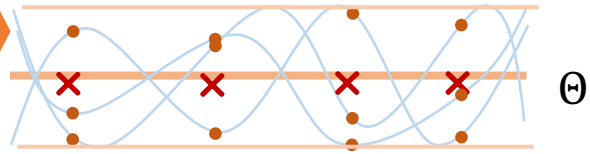


σ_i: Rademacher変数

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）

（ラデマッハ）

キー概念1：Rademacher複雑性

関数集合の大きさを測る概念

- $Z_1, \dots, Z_n$: i.i.d. 確率変数
  - 例： $Z_i = (Y_i, X_i)$
- ある関数の集合 $\mathcal{G} = \{g: \mathcal{Z} \to \mathbb{R}\}$
- Rademacher変数 $\sigma_i, i = 1, \dots, n$
  - 確率1/2で1、確率1/2で-1を取る確率変数

出典：限界数学ゼミガールLINEスタンプ
https://twitter.com/omnisucker/status/1270403374635077635

定義：Rademacher複雑性

$$\mathcal{R}(\mathcal{G}) = E_{\sigma,Z}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(Z_i)\right]$$

関数の集合の大きさを測る概念（の準備）

第一段階の定理

- 以下の確率的上限が成立する。
  - Θ: 可能なパラメータの集合
  - $\mathcal{L}_\Theta := \{\ell(\cdot, f(\cdot;\theta)) : \theta \in \Theta\}$
    ニューラルネットでとりうる損失関数の集合

定理1

$|\ell(y, f(x;\theta))| \leq B$が全ての$x, y, \theta$で成立するとする。このとき、全ての$\delta > 0$について、以下が確率$1-\delta$以上で成立する。

$$\sup_{\theta\in\Theta}\left(L(\theta) - L_n(\theta)\right) \leq 2\mathcal{R}(\mathcal{L}_\Theta) + 2B\sqrt{\frac{\log(2/\delta)}{2n}}$$

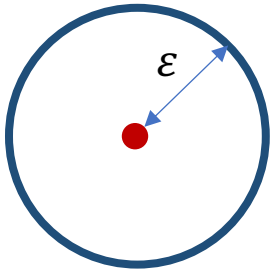
右辺のうち、二項目は比較的小さい値をとる
一項目が重要（まだ詳細には評価できてないけど）

キー概念2：被覆数(カバリング・ナンバー)

- （無限）集合の”大きさ”を測る概念
  - 有限集合なら”個数”が使えるが、無限集合には別尺度が必要
- 距離関数dを備えた集合 Ω を考える。
- $\varepsilon > 0$

定義：被覆数（カバリング・ナンバー）
部分集合$\hat{\Omega} \subset \Omega$が$\Omega$の$\varepsilon$-被覆（カバー）であるとは、全ての$\omega \in \Omega$について、ある$\omega' \in \hat{\Omega}$が$d(\omega, \omega') \leq \varepsilon$を満たすことをいう。
Ωのε-被覆数（カバリング・ナンバー）を以下のように定義：

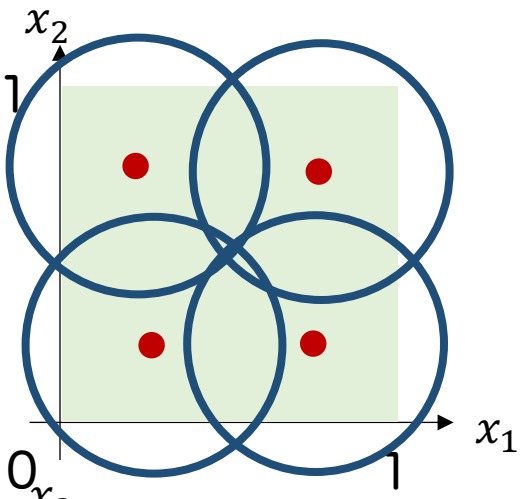
$$\mathcal{N}(\varepsilon, \Omega, d) := \min\left\{\left|\hat{\Omega}\right| : \hat{\Omega}\text{は}\Omega\text{の}\varepsilon\text{–被覆}\right\}$$



- を中心とした半径εの球で
  集合Ωを覆うのに必要な数

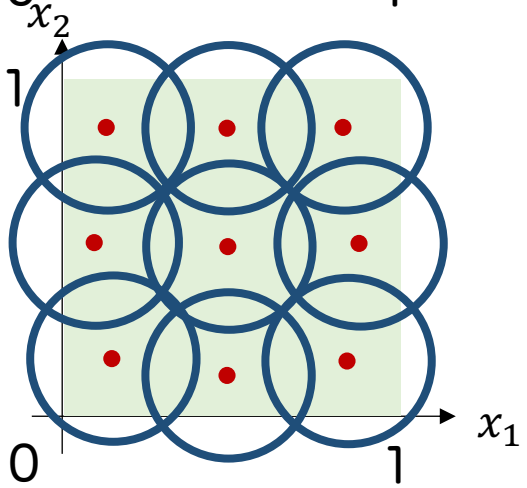
キー概念2：被覆数（カバリング・ナンバー）

- $\Omega = [0,1]^2$ とする。距離は$d_2(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}$.



- ： ε-被覆 $\hat{\Omega}$

$\varepsilon = 2/5$とすると、
$\mathcal{N}(2/5, [0,1]^2, d_2) = 4$



$\varepsilon = 1/5$とすると、
$\mathcal{N}(1/5, [0,1]^2, d_2) = 9$

集合を覆うのに必要な球の数で
集合の大きさを測る

第二段階

- 集合Ωを関数の集合$\mathcal{G} \coloneqq \{g : \mathcal{Z} \to \mathbb{R}\}$とする
  - 距離は$\sup$ノルム$\|\cdot\|_{L^\infty}$で測るとする
    （色々選択肢がある）

定理2

$\|g\|_{L^\infty} \leq B$を満たす関数の集合 $\mathcal{G}$ について、全ての$\delta > 0$について、以下が確率$1 - \delta$以上で成立する：

$$\mathcal{R}(\mathcal{G}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + 12 \int_{\alpha}^{B} \sqrt{\frac{\log \mathcal{N}(\delta', \mathcal{G}, \|\cdot\|_{L^\infty})}{n}} d\delta' \right\} + B\sqrt{\frac{\log(1/\delta)}{n}}$$

関数集合のRademacher複雑性は、
その集合の被覆数の積分値で評価できる

被覆数を評価する

- ニューラルネットによる関数集合の被覆数を計算
- 準備
  - $|\ell(y, y_1) - \ell(y, y_2)| \leq C\|y_1 - y_2\|$が全ての$y, y_1, y_2$について成立
  - Θ：ニューラルネット（L層、パラメータW個）のパラメータθの空間

補題（事実）

・$\mathcal{R}(\mathcal{L}_\Theta) \leq C\mathcal{R}(\{f(\cdot, \theta): \theta \in \Theta\})$ が成立

・$\theta \in \Theta$の各要素が$[-K, K]$に含まれる時、以下が成立：

$$\log \mathcal{N}(\delta, \{f(\cdot, \theta): \theta \in \Theta\}, \|\cdot\|_{L^\infty}) \leq W \log\left(\frac{2LK^L(W+1)^L}{\delta}\right)$$

ニューラルネットの関数集合の大きさは、パラメータ数・層数で決まる。

得られた結果

※ 結果にはいくつかバリエーションがあり、ここではその一つをお伝えしています。

- 追加の仮定
  - $|f(x;\theta)| \leq B'$が全てのx, θについて成立する。

定理3
これまでの仮定のもと、全ての$\delta > 0$について、以下が確率$1 - 2\delta$以上で成立する：

$$\sup_{\theta \in \Theta} \left(L(\theta) - L_n(\theta) \right) \leq C \left(\sqrt{\frac{WL \log(KWL)}{n}} \right) + (2 + 2C) B \sqrt{\frac{\log(1/\delta)}{n}}$$

$C > 0$ 存在する有限の定数

第一項が重要
汎化ギャップの上限は、データ数nで減少、モデルの大きさで増加

要するに何が言えるのか

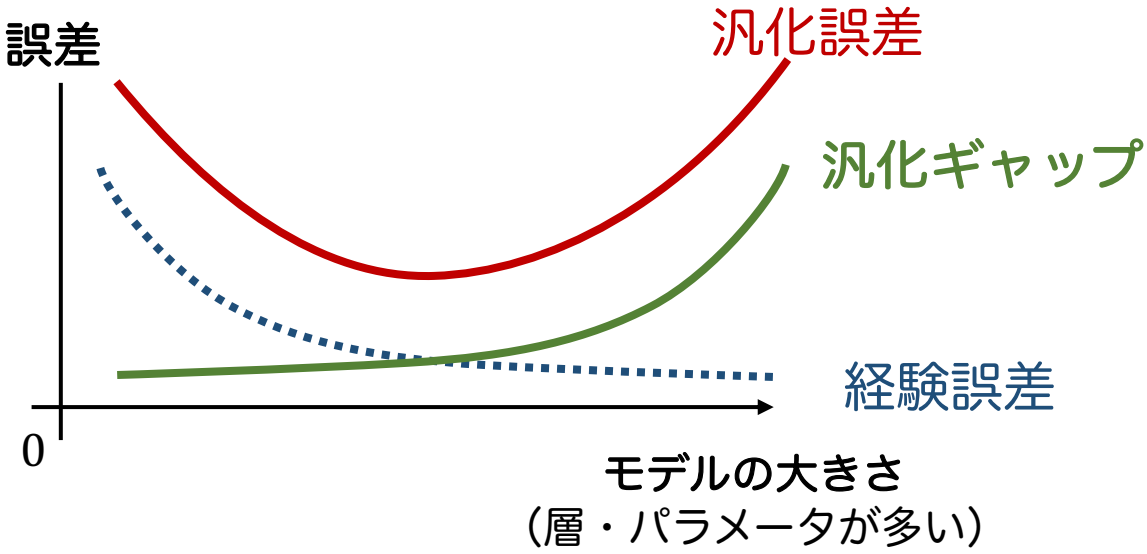
- 汎化ギャップの上限はモデルが大きくなると増加

$$L(\theta) = L_n(\theta) + \left(L(\theta) - L_n(\theta)\right)$$

汎化誤差　経験誤差　汎化ギャップ

$$\leftarrow \quad O\left(\sqrt{\frac{WL\log(KWL)}{n}}\right)$$

層の数が10倍なら
ギャップは約3~4倍



モデルの大きさと各種誤差の関係

この理論は、
深層学習以前の手法の性能を
よく説明していた。

要するに何を言っているのか

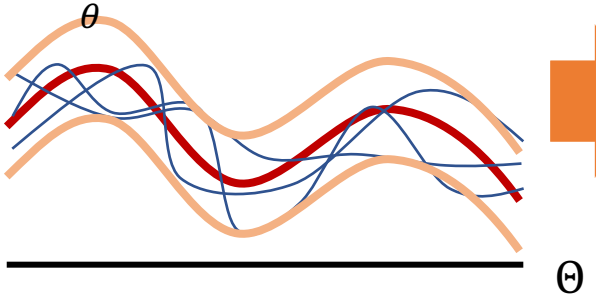
- (関数)集合上に離散点を準備
  →離散点上の確率過程の最大値で評価

$(X_1, \dots, X_p) \sim N(0, I_p)$
$\Rightarrow E\left[\max_{j=1,\dots,p} X_j\right] \leq \sqrt{\log p}$

古くから知られている事実（気持ち）
集合Θ上の確率関数の最大値 $\lesssim_p$ 確率関数の期待値＋log(集合Θの大きさ)$^{1/2}$

関数のズレの上限

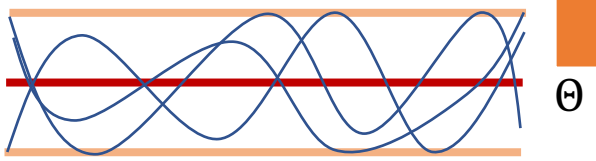
$\sup_{\theta} |L(\theta) - L_n(\theta)|$



汎化誤差を中心に
経験誤差がどうズレ
るかを知りたい

Rademacher複雑性

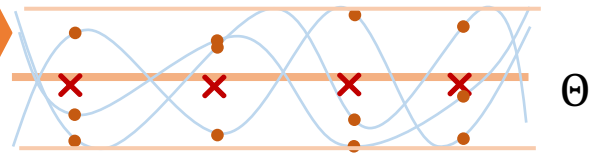
$\mathrm{E}\left[\sup_{\Theta} n^{-1}\Sigma_{i=1}^{n}\sigma_i \ell\left(Y_i, f(X_i;\theta)\right)\right]$



ズレを中心化した
（期待値をゼロに調整）
→Θの大きさを考えればOK

Dudley積分

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



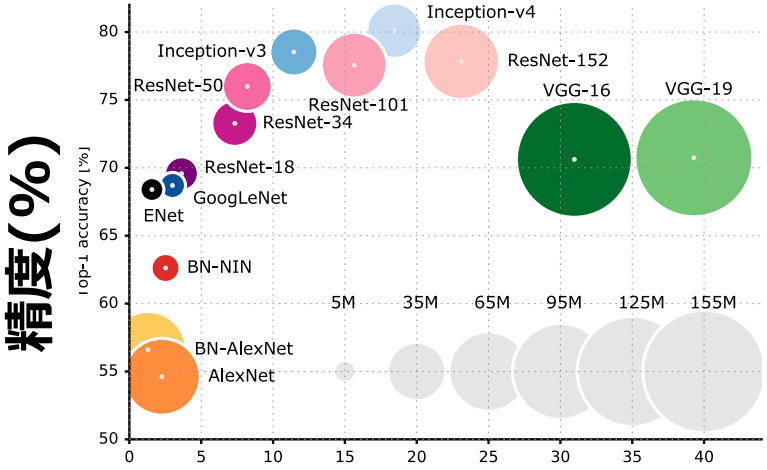
N_δ: δ被覆数 ×: ε-被覆

集合の大きさに
ε-被覆数を採用

既存理論に対する疑義

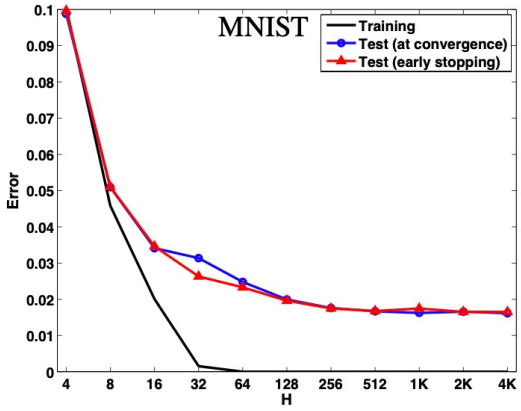
モデル複雑性理論
大量の層・パラメータは汎化ギャップ増加
→精度低下

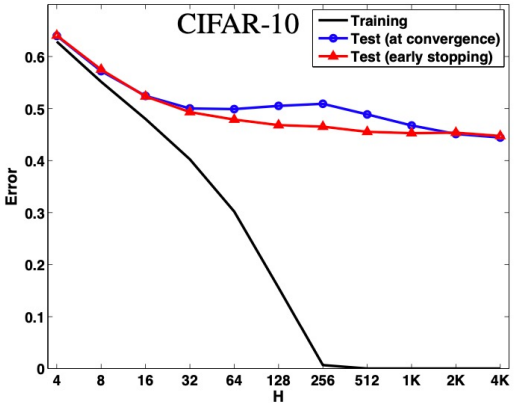
深層学習の実際
実際には精度はかなり上がっているよ



精度とパラメタ数の関係

大きい丸→ パラメタ数が大



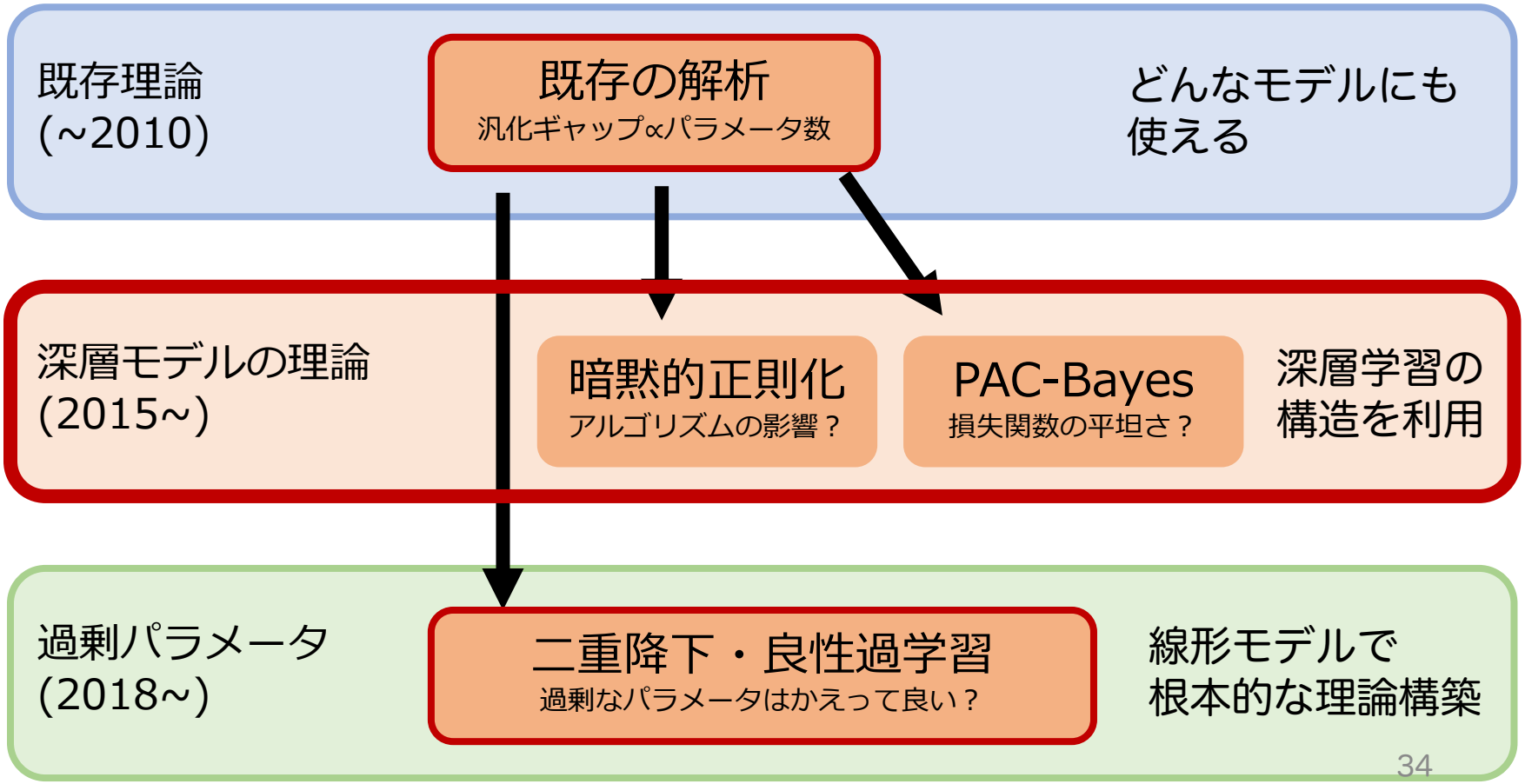


モデルを大きくして汎化誤差が下がる実験結果
横軸（モデルサイズ）、縦軸（汎化誤差）
※ 赤線・青線が汎化誤差

深層学習の試み

今日のトピック

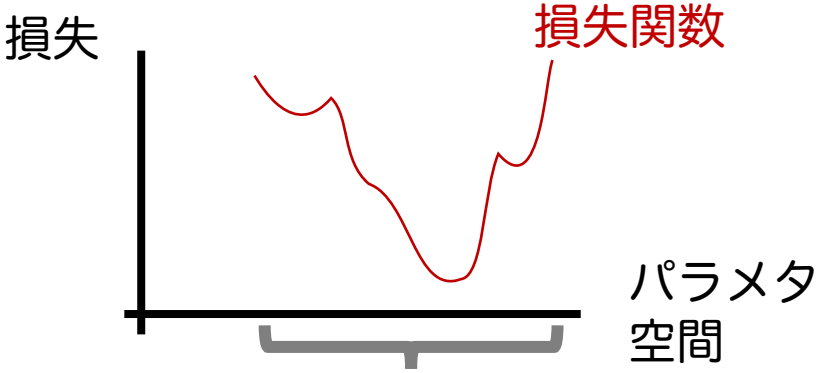
- 汎化ギャップをめぐる研究の流れ



暗黙的正則化の方針

- 学習後のパラメータが存在しうる領域を絞る

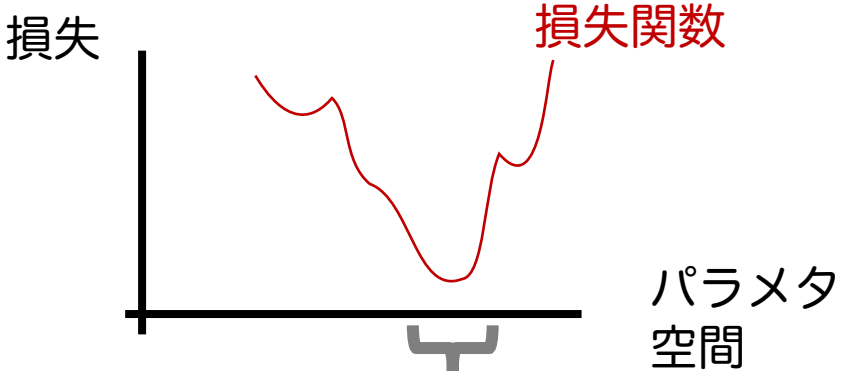
既存理論の考え



パラメタ空間のどこかにはいる

保守的な評価

暗黙的正則化理論



学習されるパラメータは
この辺りには含まれていそう

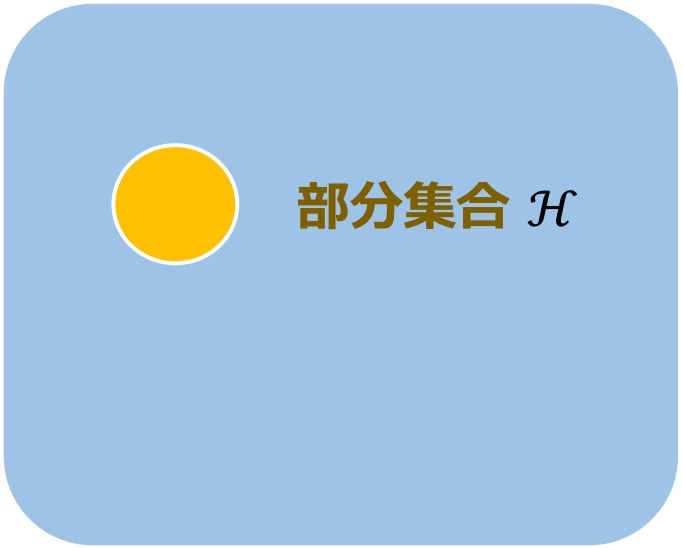
現実的な評価？

暗黙的正則化の着想

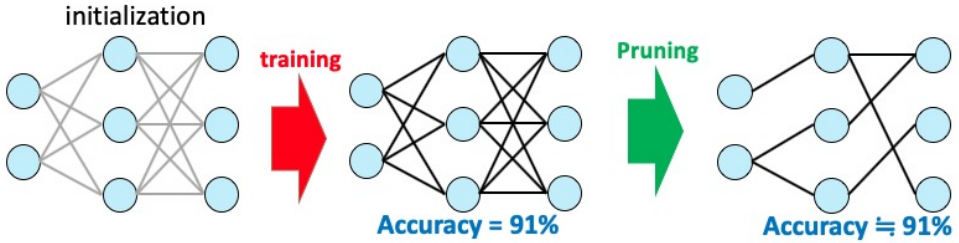
$\mathcal{F}_\Theta := \{f(\cdot,\theta): \theta \in \Theta\}$ ニューラルネットで作る関数の集合

着想：関数集合 $\mathcal{F}_\Theta$ すべてを考える必要は無い？

- 実質的に効いている部分集合$\mathcal{H} \subset \mathcal{F}_\Theta$がありそう



ニューラルネット関数集合 $\mathcal{F}_\Theta$
(多パラメタを使う巨大集合)



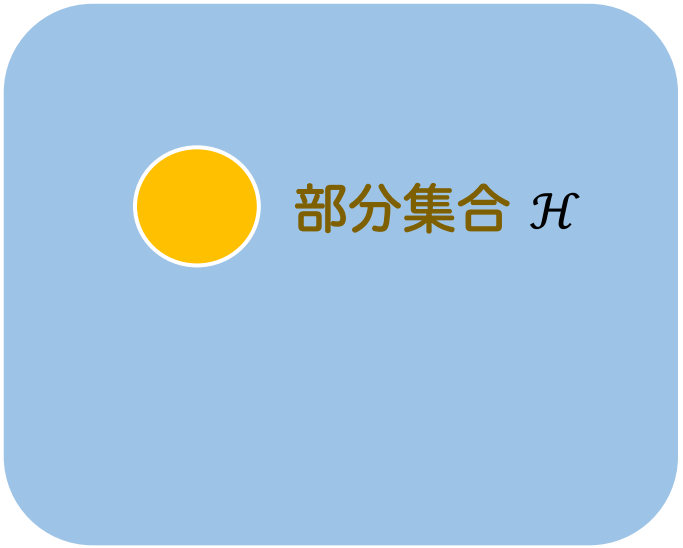
実際、学習後のネットワークは
一部の枝(パラメータ)を削除しても
十分良い予測性能を持つ

暗黙的正則化の着想

$\mathcal{F}_\Theta := \{f(\cdot,\theta): \theta \in \Theta\}$ ニューラルネットで作る関数の集合

着想：関数集合 $\mathcal{F}_\Theta$ すべてを考える必要は無い？

- 実質的に効いている部分集合$\mathcal{H} \subset \mathcal{F}_\Theta$がありそう



ニューラルネット関数集合
$\mathcal{F}_\Theta$
(多パラメタを使う巨大集合)

既存理論
汎化ギャップ ≦ 全関数集合 の大きさ

$$O\left(\int \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}_\Theta, \|\cdot\|_{L^\infty})}{n}}\, d\delta\right)$$

暗黙的正則化
汎化ギャップ ≦ 部分集合 の大きさ

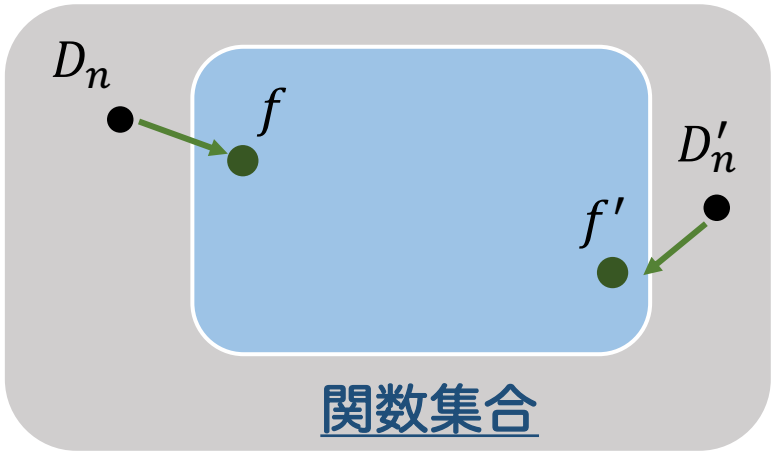
$$O\left(\int \sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{H}, \|\cdot\|_{L^\infty})}{n}}\, d\delta\right)$$

この着想を支持する実験

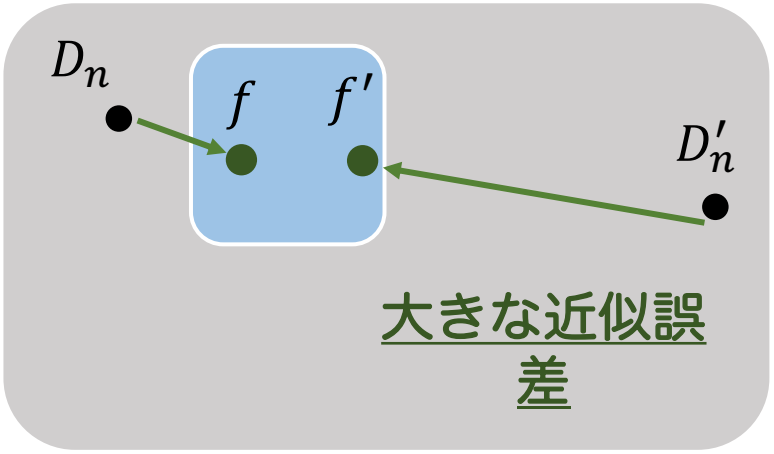
学習データ
$D_n = \{(X_i, Y_i)\}_{i=1}^n$

- データ依存集合の重要性を示す**実験**(2017年)
  - 全然違うデータD_n, D_n'でも、
    DNNの近似誤差・汎化ギャップは**両方とも小さい**

既存理論：汎化ギャップ＝ の大きさ



or



近似誤差：どちらも小
汎化ギャップ：大

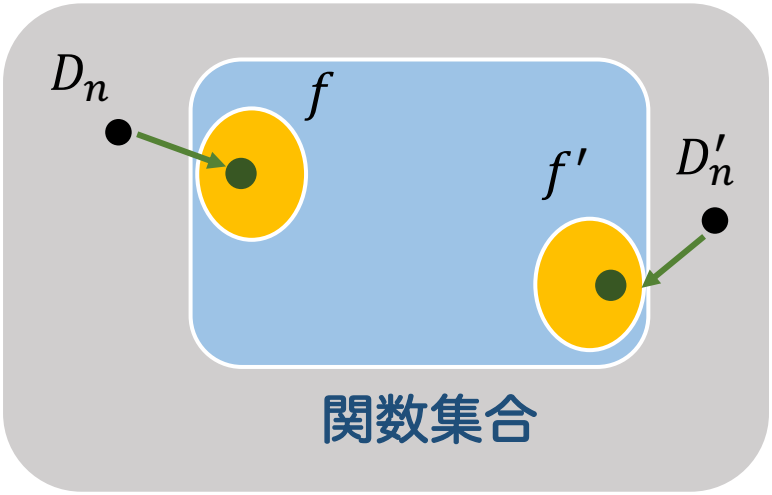
近似誤差：どちらかは大
汎化ギャップ：小

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization (2016). *arXiv preprint arXiv:1611.03530*.

データ依存集合は実験を説明

- データ依存集合の重要性を示す**実験**(2017年)
  - 全然違うデータD_n, D_n'でも、
    DNNの近似誤差・汎化ギャップは**両方とも小さい**

新しい着想：汎化ギャップ≦この部分集合  の大きさ



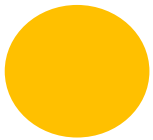
近似誤差：どちらも小
汎化ギャップ：常に小

実験・実現象と一致！

暗黙的正則化：部分集合を考える

- データ依存集合の重要性を示す実験(2017年)
  - 全然違うデータD_n, D_n'でも、
    DNNの近似誤差・汎化ギャップは両方とも小さい

新しい着想：汎化ギャップ≦この部分集合  の大きさ

何によって  は

決まっているの？

いろいろな候補
1. ノルム制約
2. 学習アルゴリズム

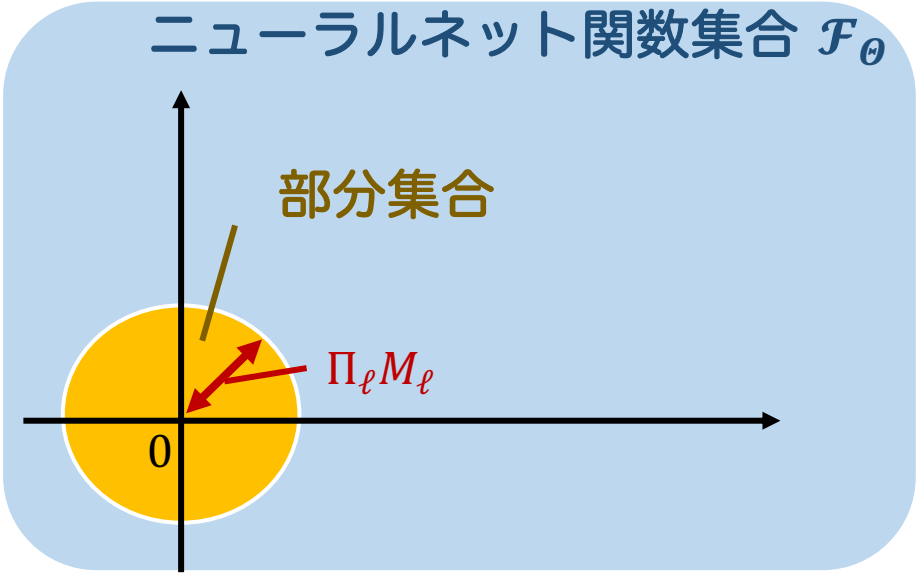
他多数

候補1. パラメータ空間の原点近傍

- **事実**：$\theta = (A_1, b_1, \ldots, A_L, b_L)$の要素は原点近傍に集中
  →パラメタ行列A_ℓのノルム$\|A_\ell\|$がゼロに近いことを想定
  - $\mathcal{H} = \{f(\cdot;\theta): \|A_\ell\| \leq M_\ell, \forall\, \ell\}$

行列ノルムの例：フロベニウスノルム

$$\|A\|_F = \sqrt{\Sigma_i \Sigma_j a_{i,j}^2}$$



原点近傍での汎化ギャップ
(2015~2019年)

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

$B = \max_i \|x_i\|$: データの大きさ

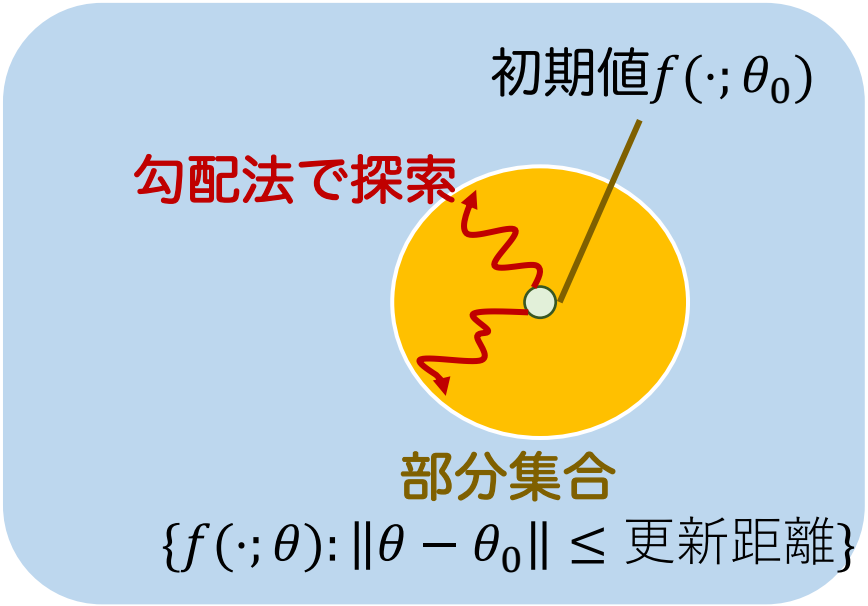
A_ℓがゼロに近ければ近いほど、汎化ギャップが小さくなる
（パラメータ数Wに依存しない）

Bartlett, P., Foster, D. J., & Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. *arXiv preprint*

候補2. 学習初期値の近傍

- **発想**：パラメタθは初期値θ_0の近くに止まっているのでは？
  - (確率的)勾配降下法の近傍だけ考えれば十分？
  - $\mathcal{H} = \{f(\cdot;\theta): \|\theta - \theta_0\| \leq M\}$

ニューラルネット関数集合 $\mathcal{F}_\theta$



パラメータをあまり更新しなければ
汎化ギャップが小さくなる

勾配降下法

初期値を設定 θ_0

パラメタ更新 $t = 1, \ldots, T$

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \nabla L_n(\theta_t)$$

微更新の元での汎化ギャップ
(2016~2017年)

$\eta_t = c/t$とする

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

$T \geq 1$: パラメタの更新回数

$q \in (0,1)$: 減衰率

Hoffer, E., Hubara, I., & Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *arXiv preprint arXiv:1705.08741*.

PAC-Bayes理論

学習後パラメタに摂動（ノイズ）を加える

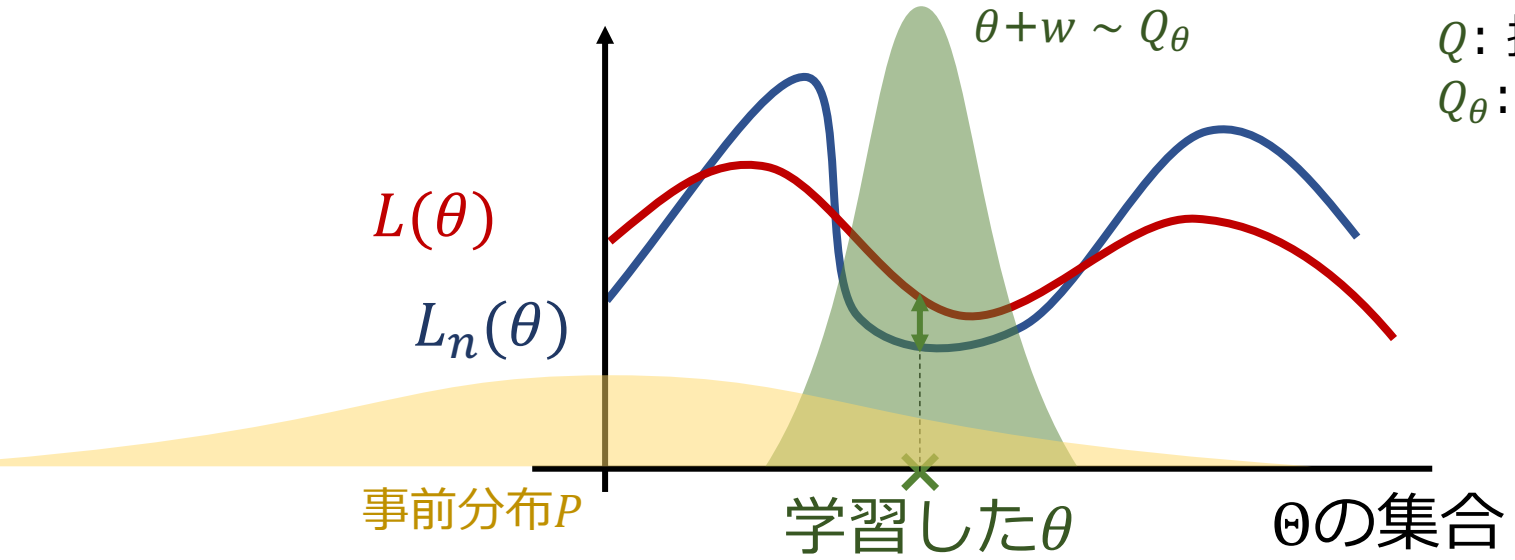
→ この摂動の影響を事前分布を用いて分析

$w \sim Q$

w: 摂動（ノイズ）

Q: 摂動分布

Q_θ: $\theta+w$の分布



摂動分布Q_θと事前分布Pで汎化ギャップを評価

導出の仕組み

カルバック＝ライブラー・ダイバージェンス
（分布P, Qの”違い”を表す尺度）

$$KL(Q||P) = \int \log\left(\frac{q(\theta)}{p(\theta)}\right) q(\theta)\, d\theta$$

p, q: 分布P, Qの確率密度関数

摂動で期待値を取った汎化ギャップを評価

→ パラメータθの近傍を摂動で探索して、 $L(\theta) - L_n(\theta)$ *を集計*

$$\left|E_{w\sim Q}[L(\theta + w) - L_n(\theta + w)]\right| = O\left(\sqrt{\frac{KL(Q_\theta||P)}{n}}\right)$$

摂動付き汎化ギャップの期待値

QとPの違いに依存した上限

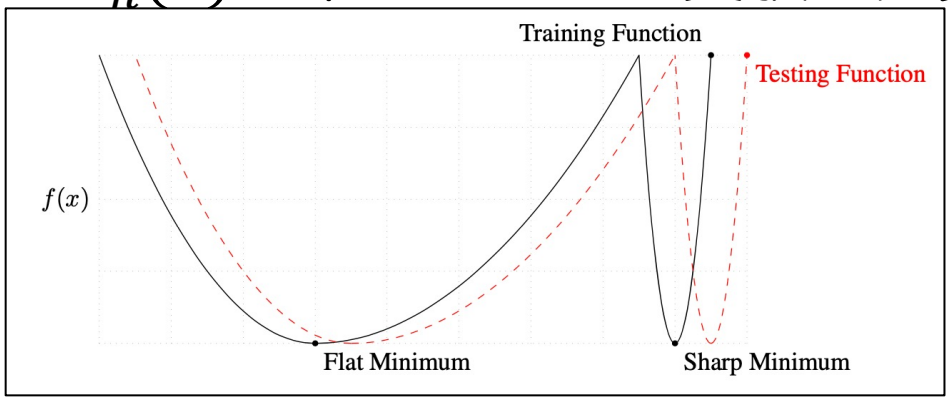
Q, Pは状況に応じて**自由に設定可**
→ モデルの大きさを受けづらい誤差評価

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*.

利点：損失関数の“形状”を解析できる

平坦最小解（flat minima）

- 汎化性能の良いパラメータθの近辺では、誤差関数$L_n(\theta)$が平坦だという実験的発見



平坦最小解のイメージ
(2017年)

PAC-Bayes理論は、**平坦最小解を説明**する

$$\mathrm{E}_{w\sim Q}[L(\theta+w)]-L_n(\theta)\leq\left\{E_{w\sim Q}[L_n(\theta+w)]-L_n(\theta)\right\}+O(\sqrt{KL(Q_\theta||P)/n})$$

汎化ギャップの近似値
(摂動平均との誤差)

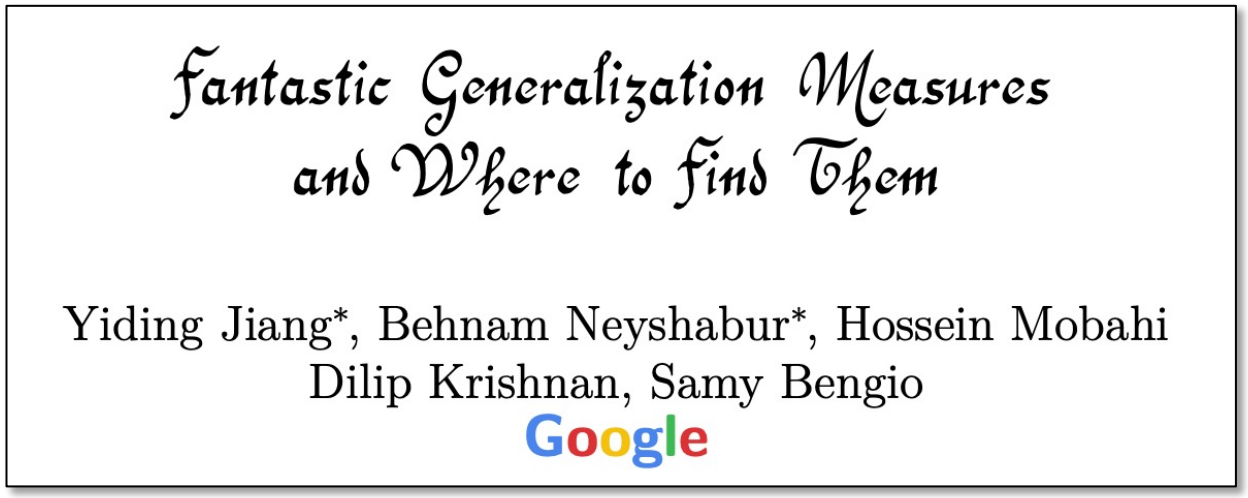
解の平坦さの期待値

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

実験によるPAC-Bayes理論の有用性

- **最も実験と整合的**な理論
  - 実験的に計測した汎化ギャップとの整合性が既存理論や暗黙正則化などよりも高い

Googleによる大規模実験
40種類の理論評価の妥当性を約2000種類のCNNで評価


Fantastic Generalization Measures and Where to Find Them

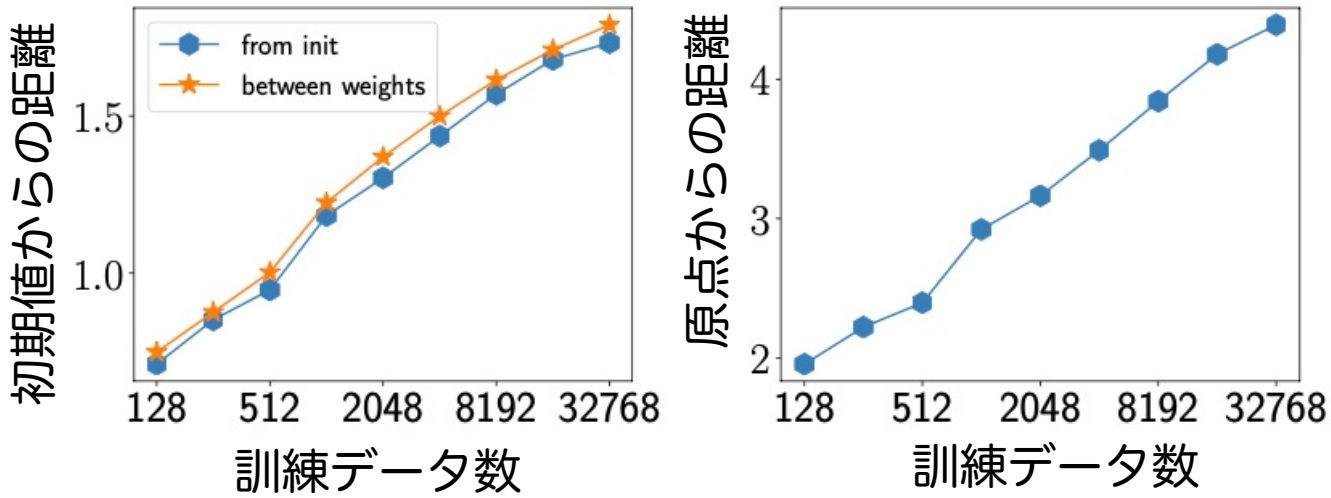
Yiding Jiang*, Behnam Neyshabur*, Hossein Mobahi
Dilip Krishnan, Samy Bengio
Google

当該論文の
タイトル・著者
（2018年）

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., & Bengio, S. (2019). Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.

これらの理論への批判

実験：パラメタは原点・初期値近傍に留まらない
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（2019年）

→ 暗黙的正則化理論の根本的な設定に疑義

Nagarajan, V., & Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. *arXiv preprint arXiv:1902.04742*.

ここまでのまとめ

背景

- 既存理論では、過適合しやすさ∝パラメタ数
- 学習されるパラメタの性質が分からないので保守的

新理論

- パラメタの性質を調べてタイトな(パラメタ数に依存しない)評価
- 学習アルゴリズムの影響など

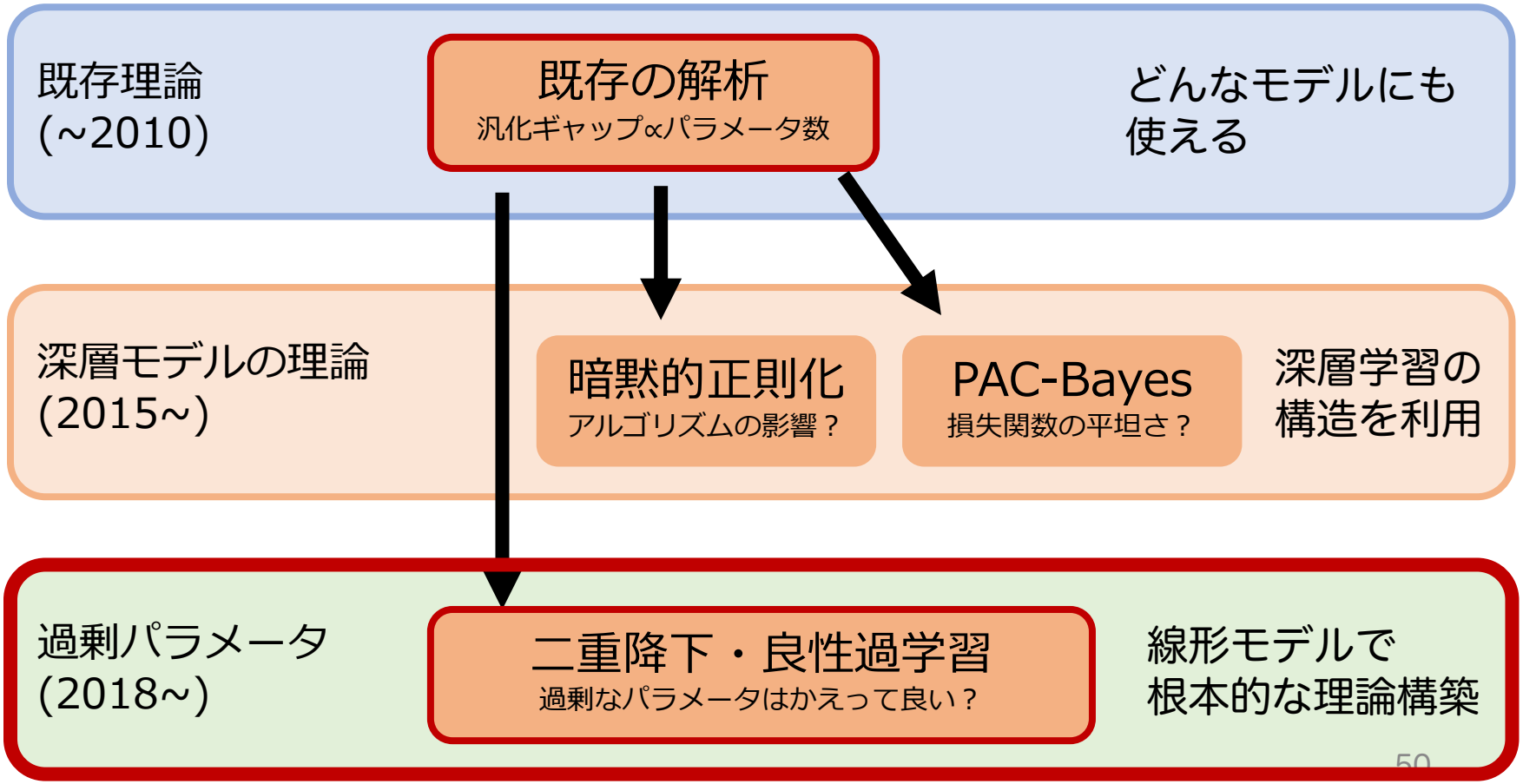
現状

- パラメタの性質を特定しきれていない→批判がある

過剰パラメータの理論

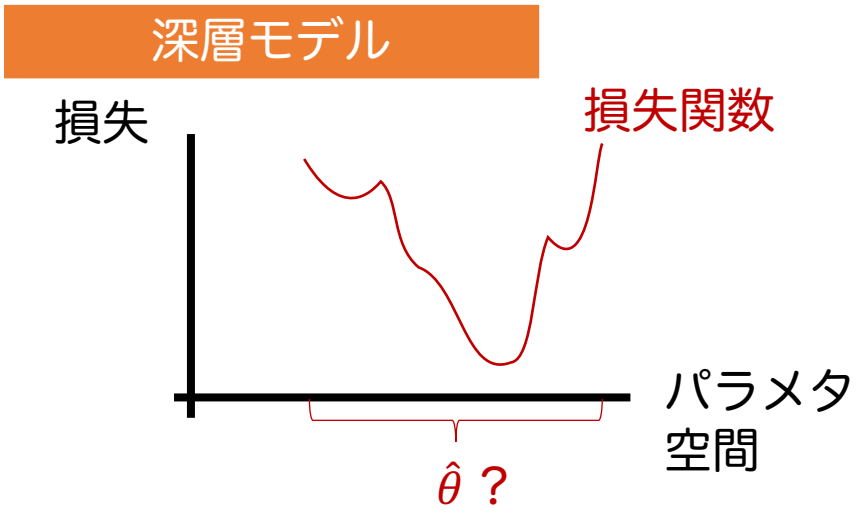
今日のトピック

- 汎化ギャップをめぐる研究の流れ



全体的な方針

- 深層構造(複雑な損失関数)は諦める



層を増やすと
損失関数の形が
難しすぎる…
何も分からない…

とりあえず層のことは忘れて
線形モデルでパラメタの数に
ついて考えよう！

準備：補間量と過剰パラメータ

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関数）

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

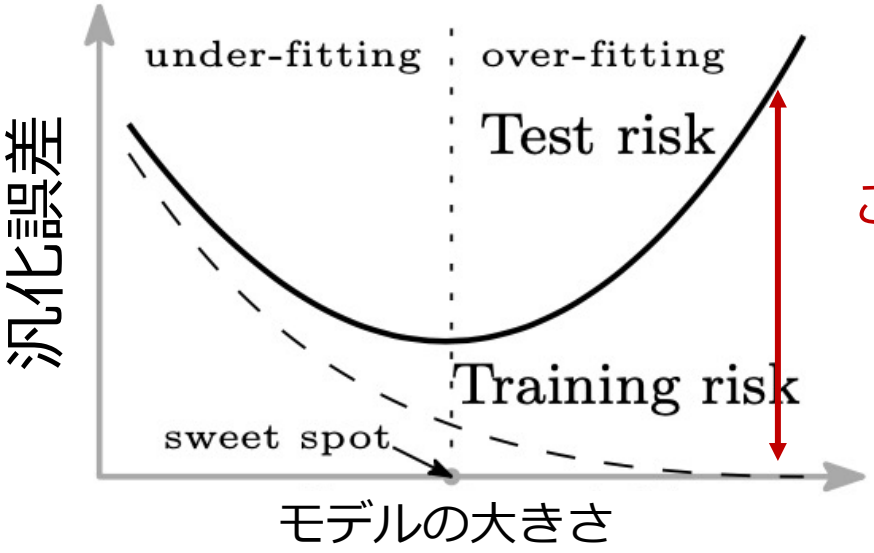
近年の理論的発見

- 補間量の汎化誤差は、過剰パラメタ化のもとで減少
- ただし、理論的には**線形モデル**に限定

典型例1：二重降下理論

<u>従来の理論の考え</u>

パラメータを一定以上増やすと過学習で誤差増加



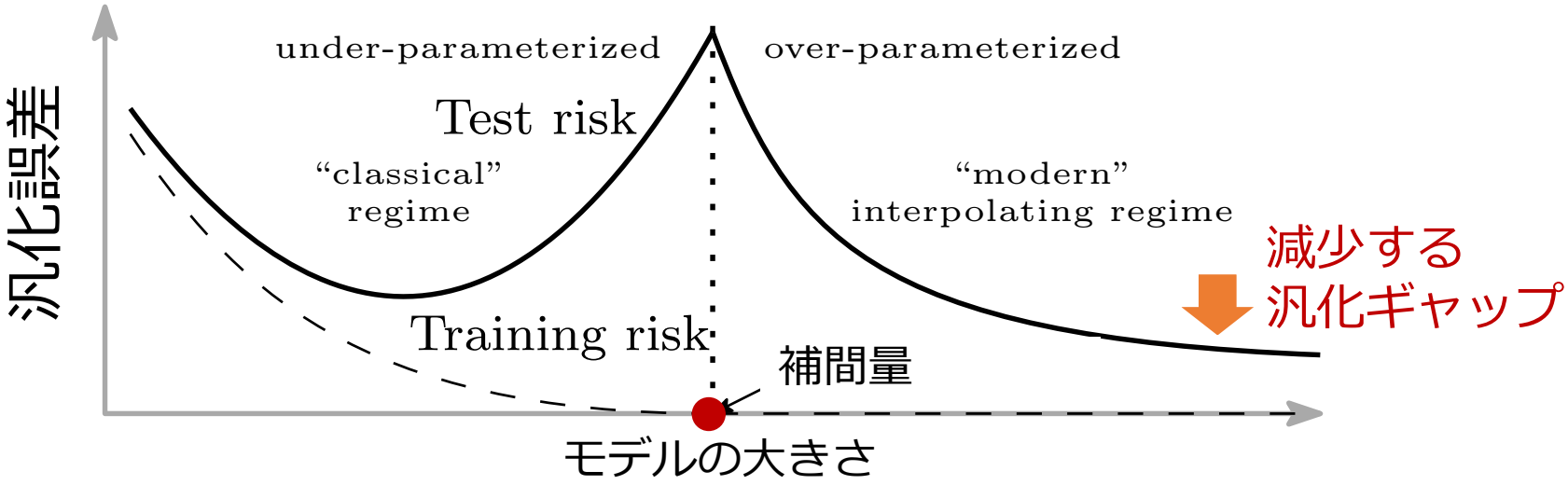
このギャップが汎化ギャップ
（テスト誤差-訓練誤差）

図はBelkin+ 2019より

典型例1：二重降下理論

二重降下 (double descent)
モデルを過剰に大きくすると、
汎化ギャップ（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

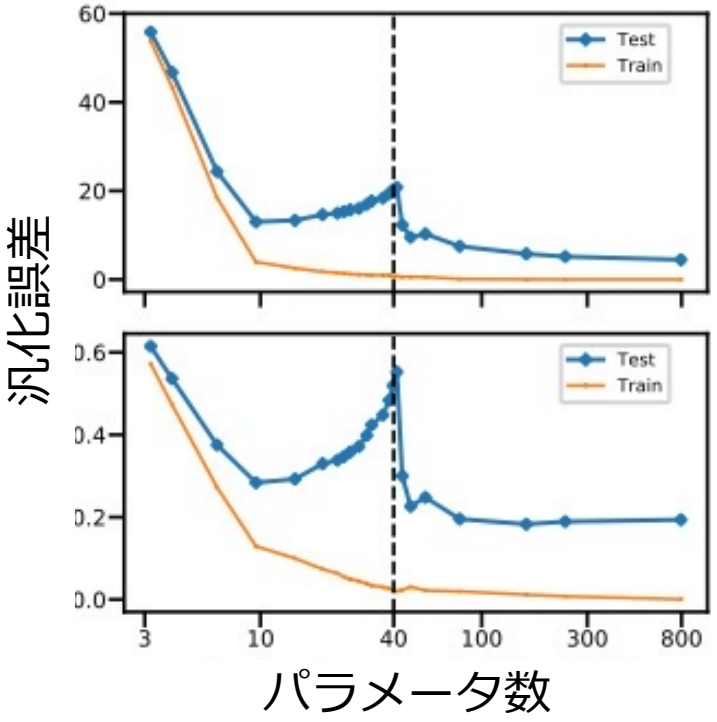


線形モデルでは、これを説明できる

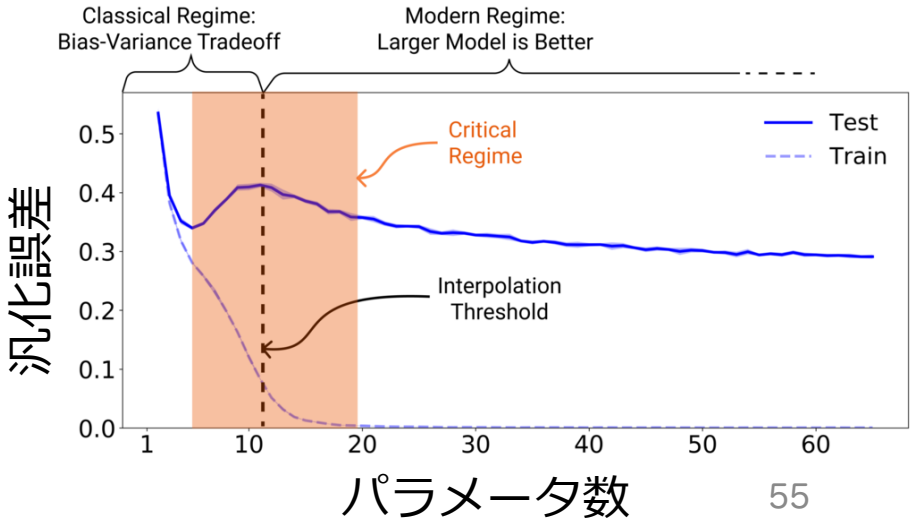
実験による発見

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)



これを理論で説明できるか？

線形回帰の解析

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

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

線形回帰の汎化誤差

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

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

$$L(\hat{\beta}) = E_\varepsilon\left[\|\hat{\beta} - \beta^*\|_\Sigma^2\right] = \|E_\varepsilon[\hat{\beta}] - \beta^*\|_\Sigma^2 + \mathrm{tr}[\mathrm{Cov}_\varepsilon(\hat{\beta})\Sigma]$$

$= \text{バイアス} B(\hat{\beta})$（≈近似誤差）

$= \text{バリアンス} V(\hat{\beta})$（≈汎化ギャップ）

理論の中身は？

$V(\hat{\beta})$を経験共分散行列の固有値で書き換え

- $X = (x_1, \ldots, x_n)^\top, Z = X\Sigma^{1/2}$ （$n \times p$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

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

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

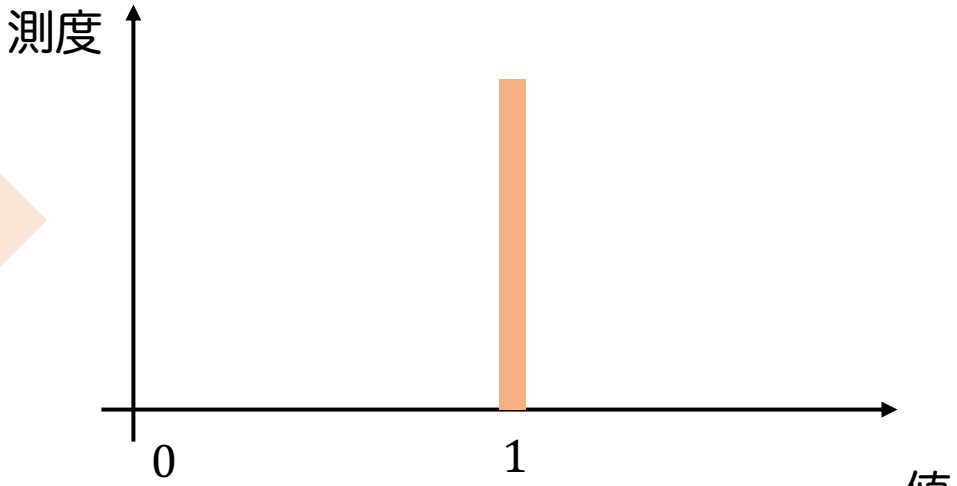
$F_{Z^\top Z/n}(s)$: 行列$Z^\top Z/n$の固有値分布

固有値分布の例

単位行列

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

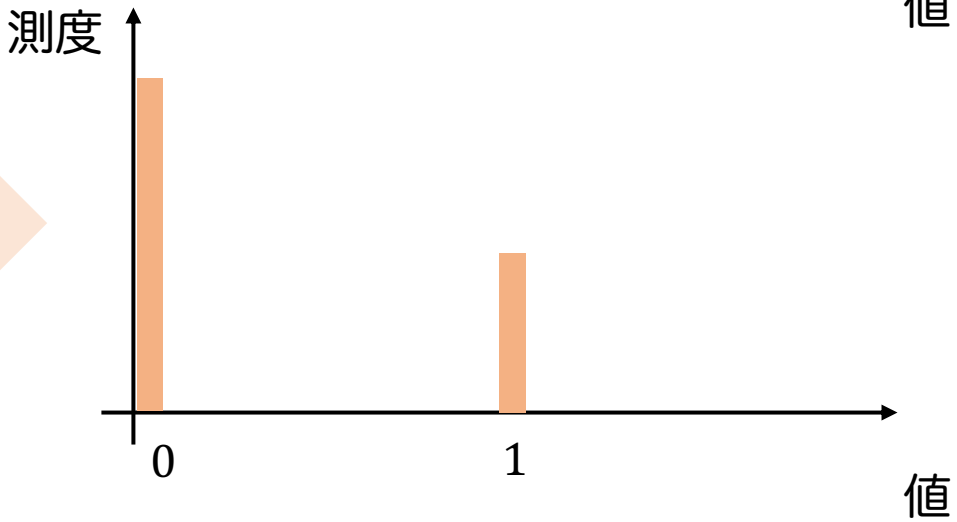
固有値は
1,1,1



特異行列

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

固有値は
1,0,0

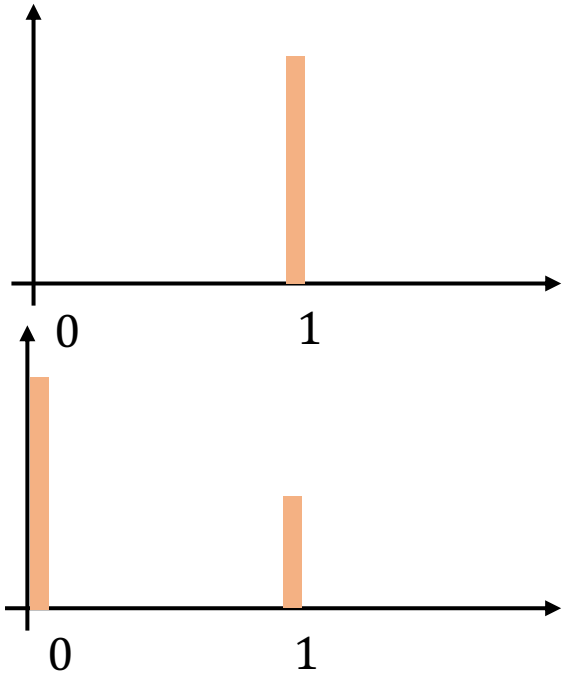


行列が多様な”情報”を持つ時、分布が右に寄る

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

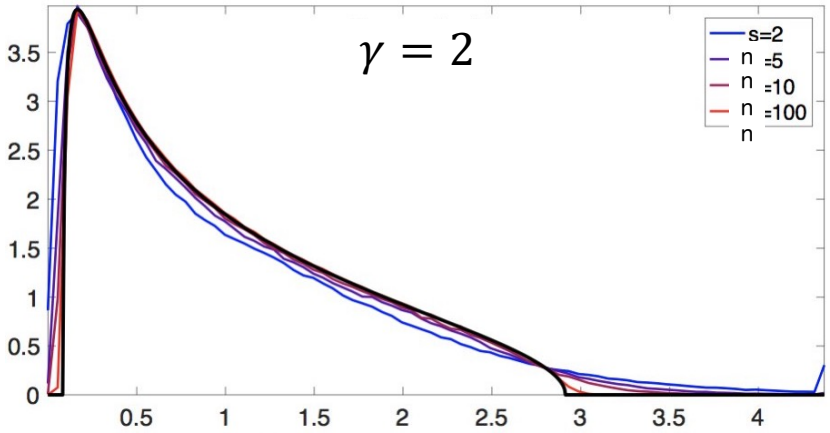
キーとなる固有値分布

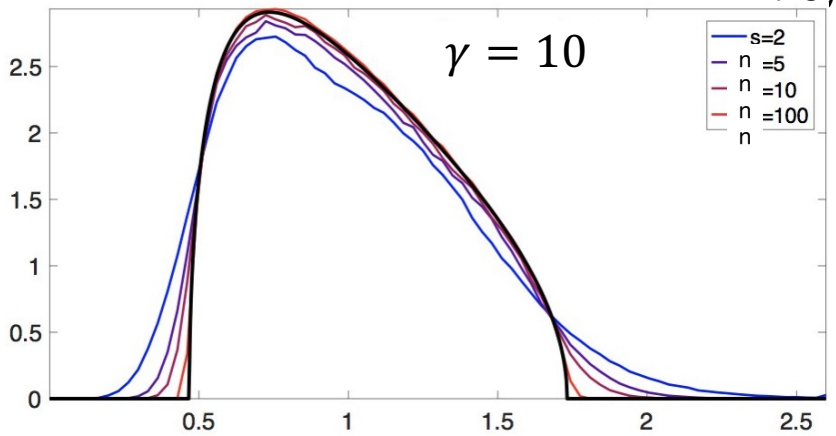
マルチェンコ＝パスツール則 (MP則)

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

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

Peyre(202





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

($p, n \to \infty$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

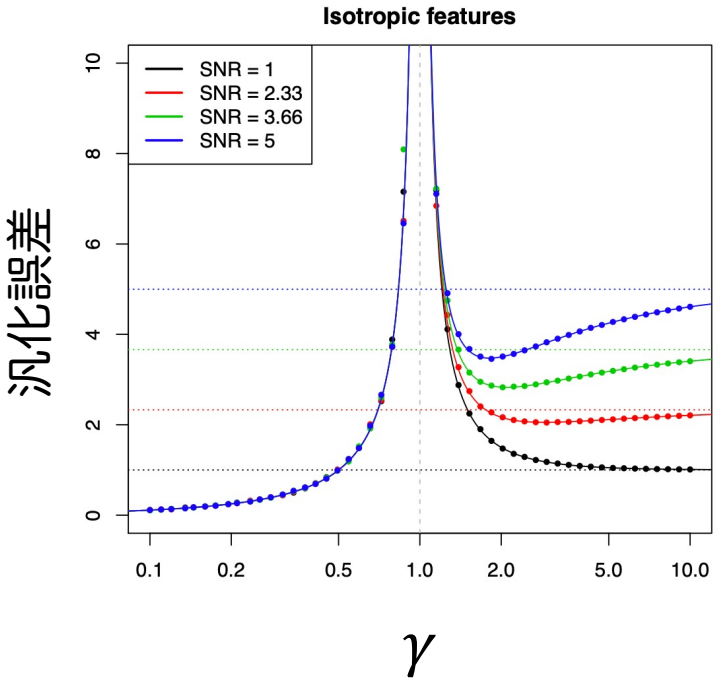
理論による二重効果の再現

線形回帰の汎化誤差(Hastie+ (2019))

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

$$\lim_{p,n\to\infty} L(\hat{\beta}) = \begin{cases} \dfrac{\sigma^2\gamma}{\gamma-1}, & (\gamma < 1) \\ r^2(1-\gamma^{-1}) + \dfrac{\sigma^2}{\gamma-1}, & (\gamma > 1) \end{cases}$$

$= \text{バイアス} B$ （近似誤差）　$= \text{バリアンス} V(\hat{\beta})$ （≈汎化ギャップ）



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、複雑性誤差が減少
- × パラメータが多い場合は、近似誤差は残る

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する

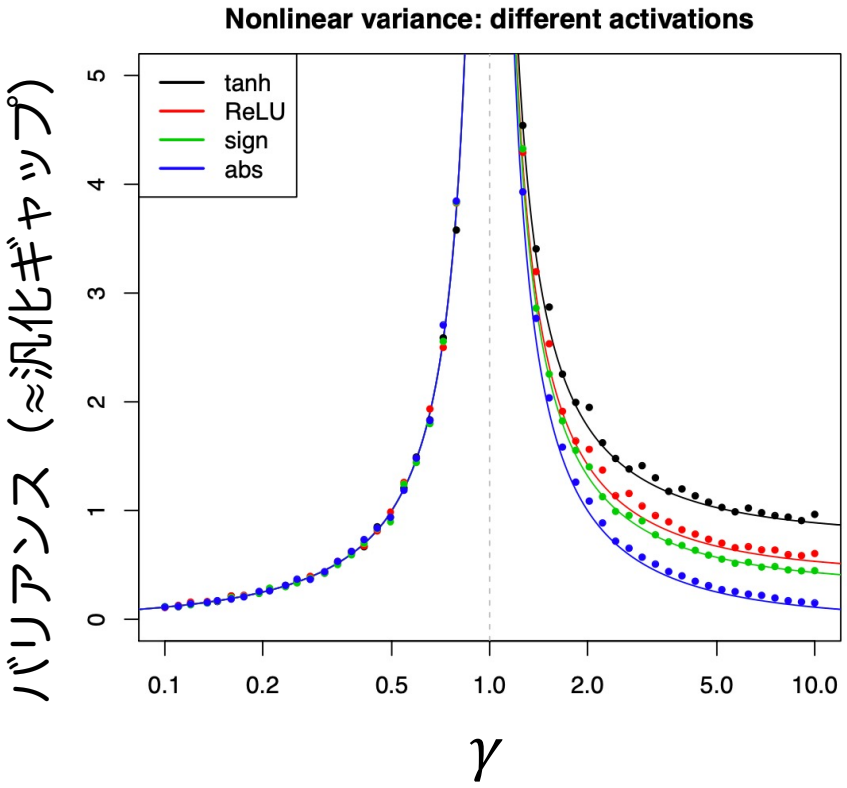
考える二層NN

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

- 一層目：a_j, bは乱数（学習
- 二層目：勾配降下法で学習

→ **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

典型例2：良性過学習

良性過学習 (benign overfitting)
データが特定の条件を満たすとき、過剰パラメタ化した線形回帰の汎化誤差がゼロに収束する

設定は同じ

線形回帰モデル

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

$\Sigma = \mathrm{E}[X_i X_i^\top]$はデータの共分散

補間量

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

線形回帰の良性過学習(Bartlett+ (2020))
共分散 Σ の固有値λ_iが$\lambda_i \asymp i^{-1} \log^{-a}(i), (a > 0)$の時、
n, pが無限の極限で$L(\hat{\beta})$はゼロに収束する。

✓ バイアスとバリアンスの両方がゼロに収束

要は何をやってる？

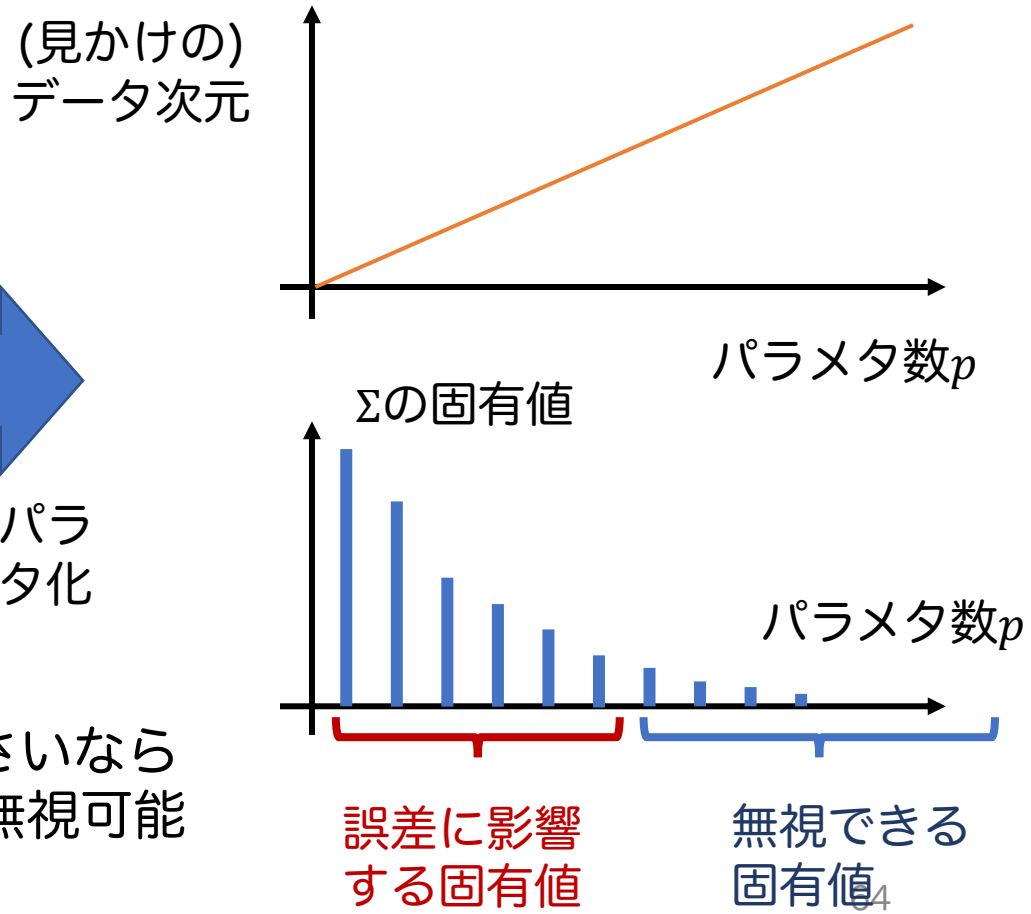
- データのスペクトル(固有値)構造の抽出

線形回帰モデル

$$Y_i = \beta^{*\top} X_i + \varepsilon_i$$

p: パラメタ数

過剰パラメータ化



高い次元の影響力(固有値)が小さいなら
高次元性が誤差に与える影響は無視可能

良性過学習と暗黙的正則化の関係

- 良性過学習の誤差は、
  補間量上の一様収束誤差とみなせる

良性過学習と暗黙的正則化 (Koehler+ (2021))

線形回帰・ガウスデータでは以下が成立：

$$L(\hat{\beta}) \leq \sup_{\beta:\text{補間量}} L(\beta) \leq \text{良性過学習で導かれる誤差}$$

補間量への
暗黙的正則化

一様収束誤差を使った評価

過剰パラメータ化の理論的発見の理解が進んでいる

ここまでのまとめ

設定

- 深層のことは忘れる
- 線形モデルで過剰パラメータを議論

アイディア

- データの固有値を解析
- 見かけの高次元性を回避できる

現状の結果

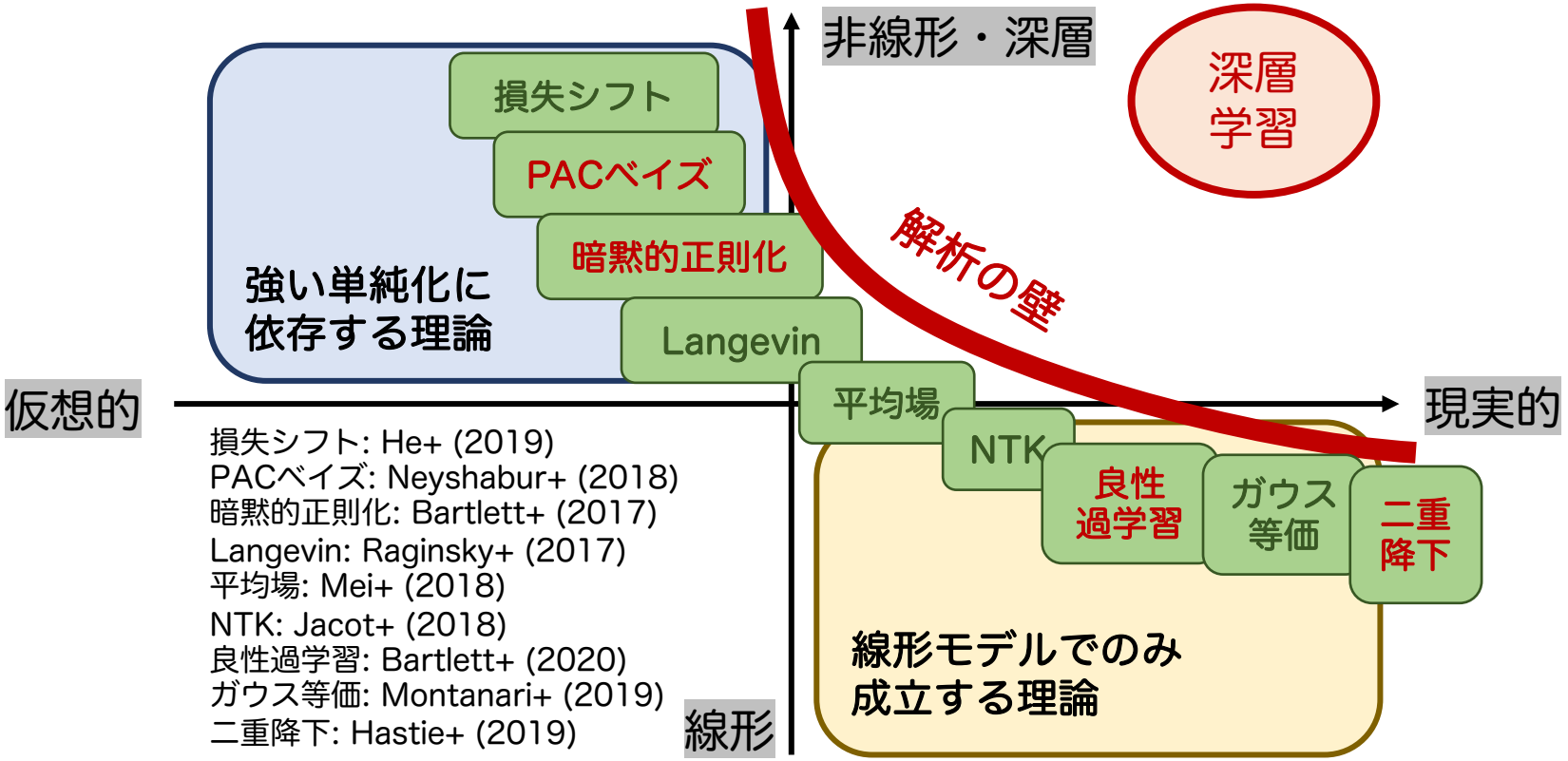
- 新しい高次元統計の結果が出つつある

まとめと今後

新理論の挑戦と限界

近年の深層学習理論

- 革新的な理論が数多く提案、しかし記述できない点は多い



まとめ

背景

- 深層学習の”発見”と、それを記述する理論の不在

目的

- 深層学習という新しい技術をもとに
  深層構造に適応する新しい統計理論を作ること

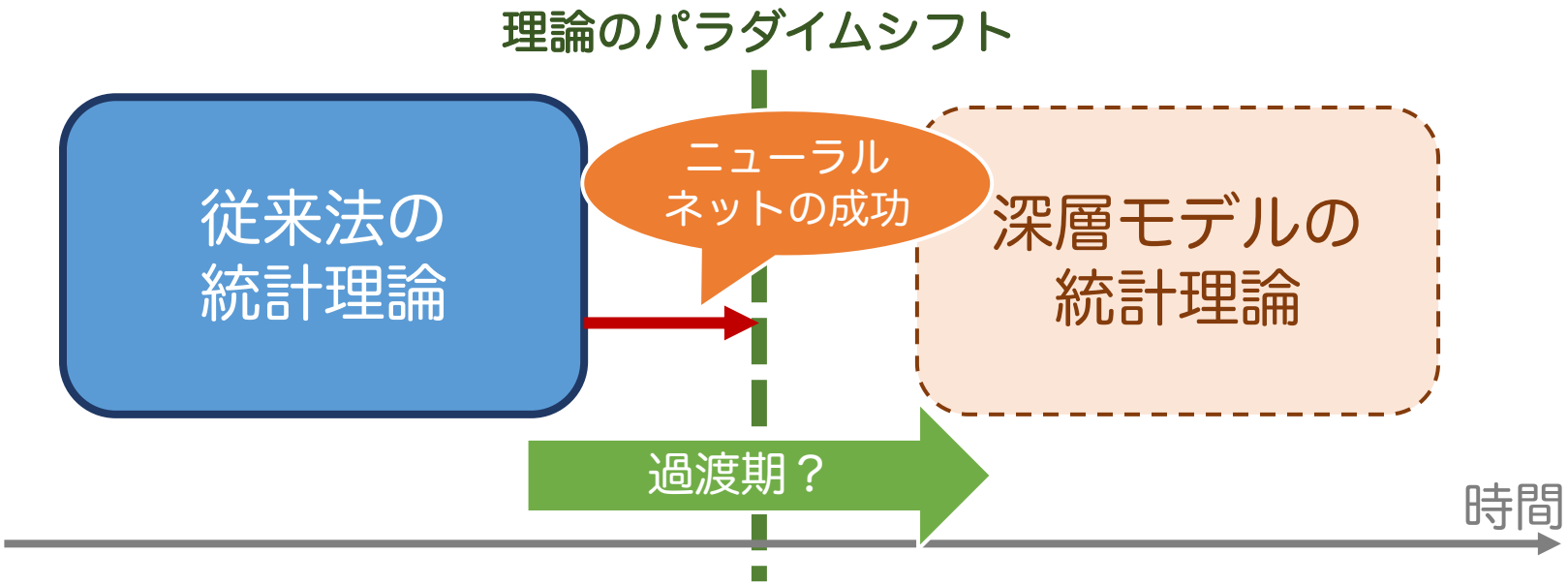
手段

- 近似誤差・複雑性誤差の解析
- 解析設定の拡張や、新しい統計的記述の開発

感想：統計分野に面白い未解決問題がある時代で嬉しい

新理論を創出できる？

- 今後到来する(?)深層統計理論の基盤創出
  - 資金石としての深層ニューラルネットワーク

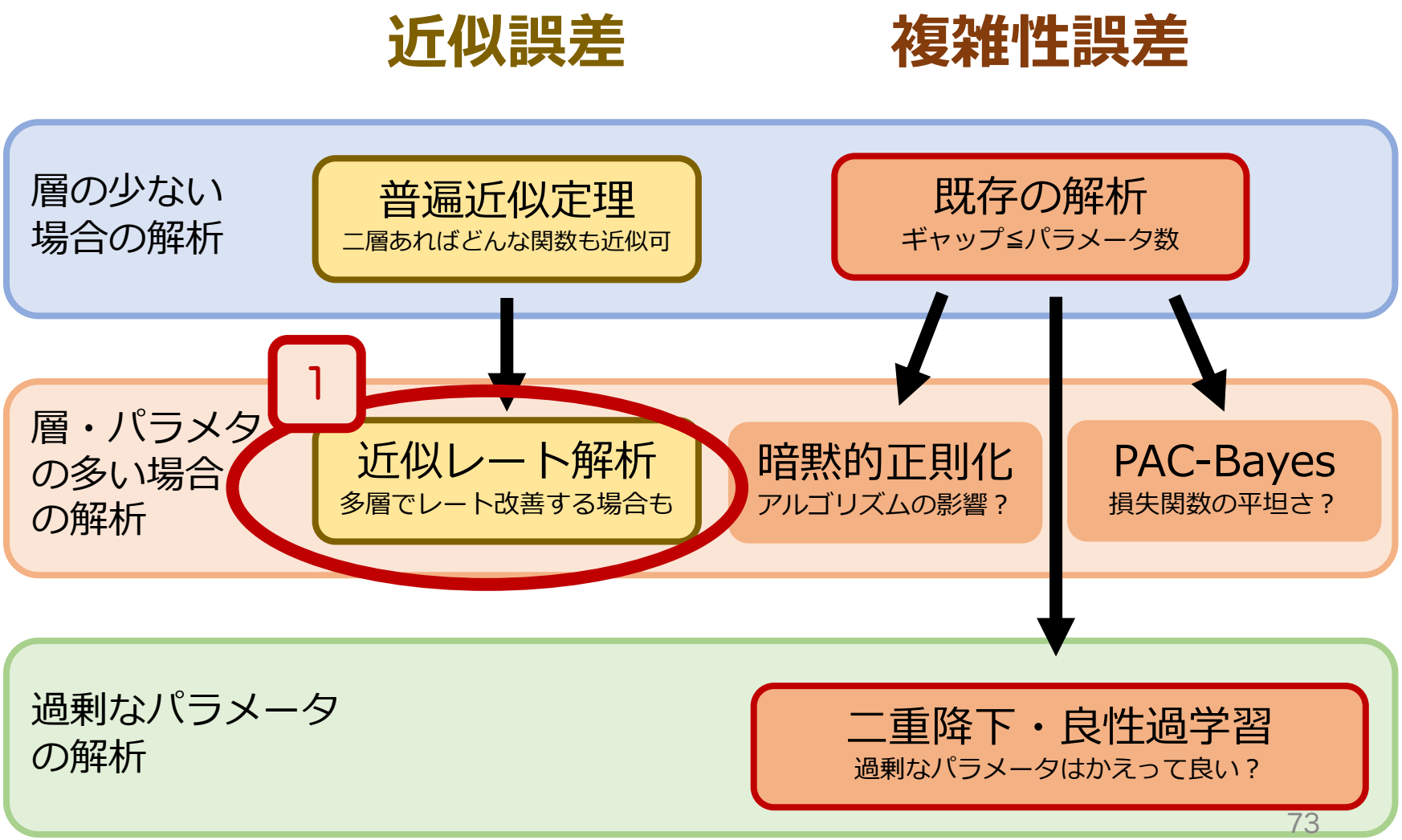


成功するかは不明だが、楽しく研究できる分野

研究1：近似レート解析

ノンパラメトリック回帰による近似誤差の解析

今日の発表の流れ


近似誤差
複雑性誤差
層の少ない
場合の解析
普遍近似定理
二層あればどんな関数も近似可
既存の解析
ギャップ≦パラメータ数
1
層・パラメタ
の多い場合
の解析
近似レート解析
多層でレート改善する場合も
暗黙的正則化
アルゴリズムの影響？
PAC-Bayes
損失関数の平坦さ？
過剰なパラメータ
の解析
二重降下・良性過学習
過剰なパラメータはかえって良い？

統計の枠組みで解析

ノンパラメトリック回帰の設定

n個のi.i.d.観測　$\{(X_i, Y_i) \in [0,1]^D \times \mathbb{R}\}_{i=1}^n$

観測の生成過程　$X \sim P_X$ on $[0,1]^D$

$$\mathrm{Y} = f^*(X) + \epsilon \qquad (\epsilon\text{: ノイズ})$$

真の関数（未知）　$f^*: [0,1]^D \to \mathbb{R}$

- 観測から予測器 $\hat{f}$ を構成

性能の尺度: 汎化誤差（予測誤差）

$$E_{X\sim P_X}\left[\left(\hat{f}(X) - f^*(X)\right)^2\right] =: \left\|\hat{f} - f^*\right\|_{L^2(P_X)}^2$$

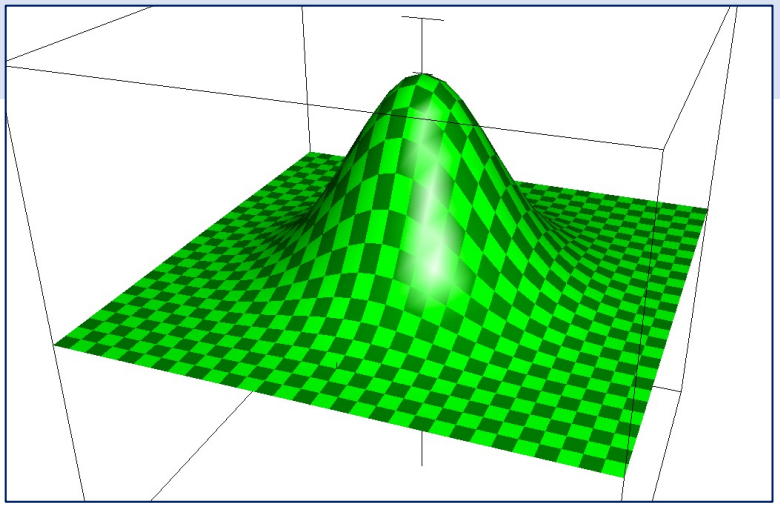
統計理論によるアプローチ

- 統計理論により分かっていること

既存の理論的結果

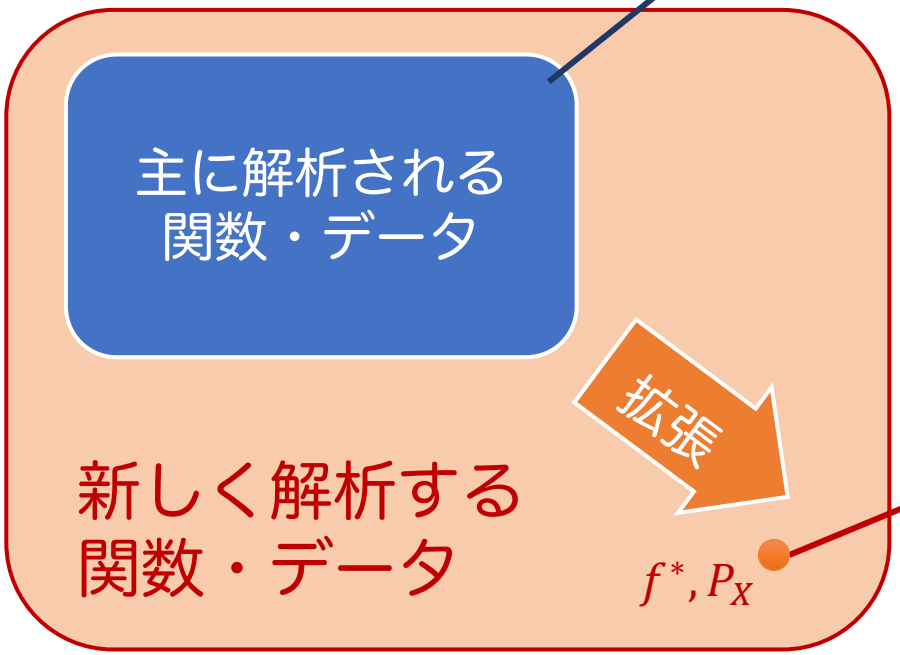
f^* が**滑らか（微分可能）**であるとき, DNN以外にも多くの手法（カーネル法, フーリエ法, Shallow NNなど）が**最適精度**を達成する。

滑らかな関数の例



深層の優位を示す設定の提案

本研究：より広い関数f^*・データ分布P_Xを解析



滑らかな関数・正則データ

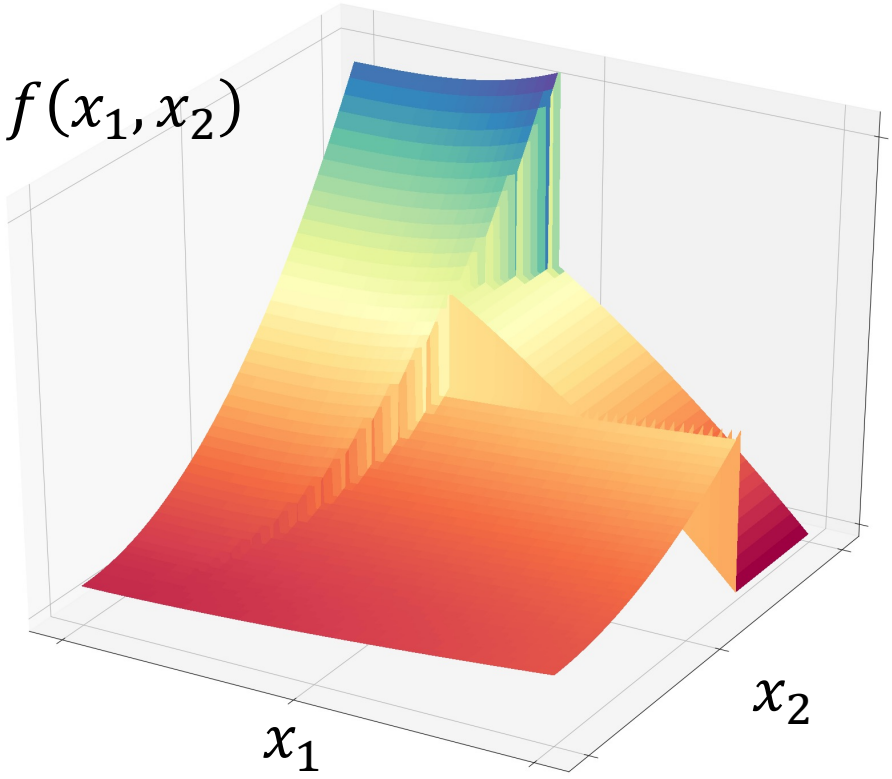
従来手法(非深層法)でも
ある最適性を持つ

非滑らかな関数・
非正則データ

仮説：この設定では深層学習の優越性を証明できる？

研究1：非滑らかな関数f^*

区分上で滑らかな関数 $\mathcal{F}^{PS}$
Piecewise Smooth Function



二次元入力を持つ関数の例
（区分の境界上で非滑らか）

関数の台を区分に分割

入力値の微小な変化で
出力が大きく変化するデータを表現
例：相転移現象

方針1：非滑らかな関数f^*

$\mathcal{F}^{PS}$：区分上で滑らかな関数集合

主定理

データ生成関数f^*が区分上で滑らかな関数$\mathcal{F}^{PS}$の場合：

$$\inf_{\hat{f}^{DL}} \sup_{f^* \in \mathcal{F}^{PS}} E\left[\left\|f^* - \hat{f}^{DL}\right\|^2_{L^2(\mu)}\right] \prec \inf_{\hat{f}^{\mathrm{lin}}} \sup_{f^* \in \mathcal{F}^{PS}} E\left[\left\|f^* - \hat{f}^{\mathrm{lin}}\right\|^2_{L^2(\mu)}\right]$$

深層学習の汎化誤差

$\hat{f}^{DL}$ は深層学習による予測器

従来手法の族の汎化誤差

$\hat{f}^{\mathrm{lin}}$ はフーリエ・カーネル法による予測器

深層学習のミニマックス誤差のオーダーが厳密に優越

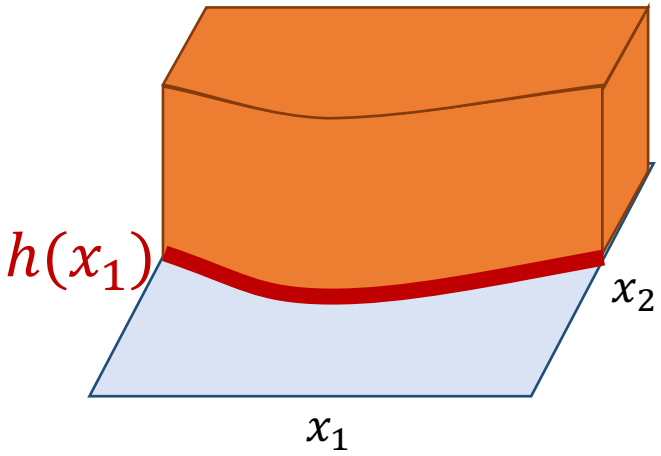
データが**非滑らかな関数**から生成される時

深層学習が従来手法を優越することを証明

DNNによる不連続性解消

不連続な関数の例

- $\{(x_1, x_2): x_1 = h(x_2)\}$上で不連続 / 区分内部で定数



$$1_{\{x_2 \geq h(x_1)\}}(x_1, x_2)$$

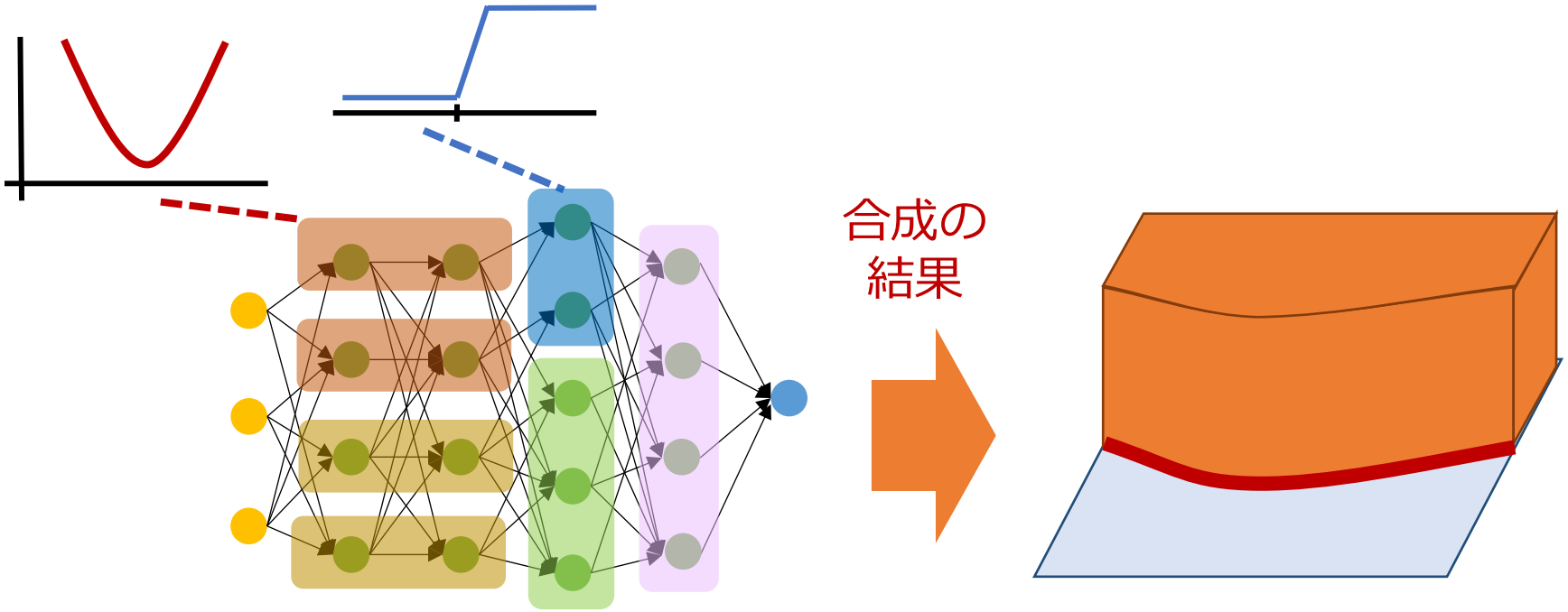
$h(x_1)$: 滑らかな関数

滑らかな関数とステップ関数の合成に分解

- $\boldsymbol{s}(x) := 1_{\{x \geq 0\}}(x)$: ステップ関数
- $1_{\{x_2 \geq h(x_1)\}}(x_1, x_2) = \boldsymbol{s} \circ \underbrace{\left((x_1, x_2) \mapsto \left(x_2 - h(x_1)\right)\right)}_{\text{滑らかな関数}}$

DNNによる近似の方法

- DNNは、非滑らかな関数の各パーツを個別に近似

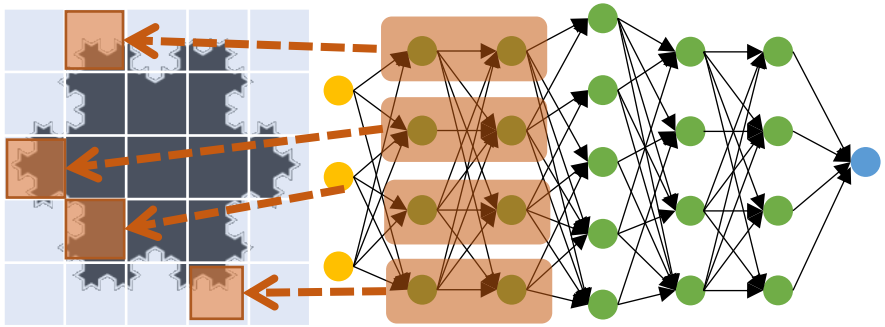


ステップ関数等の近似誤差は無視可能

→滑らかな関数を近似するのと同等の近似誤差

方針2：特異性を持つデータ分布P_X

- 非特異データ(例:フラクタル構造)と深層の関係



フラクタル特徴量

深層構造が
フラクタルの分割に適応

特徴量がS次元フラクタル構造を持つ場合の汎化誤差

β: f^*の滑らかさ、 D: データの次元

$$\tilde{O}\left(n^{-2\beta/(2\beta+S)}\right) < \tilde{O}\left(n^{-2\beta/(2\beta+D)}\right)$$

深層学習の誤差　　通常の誤差

データが**複雑な低次元構造**を持つ時

深層学習がその構造に適合することを証明

研究1のまとめ

目的

- 深層の役割を理解したい

アプローチ

- より複雑な関数の近似誤差を考える

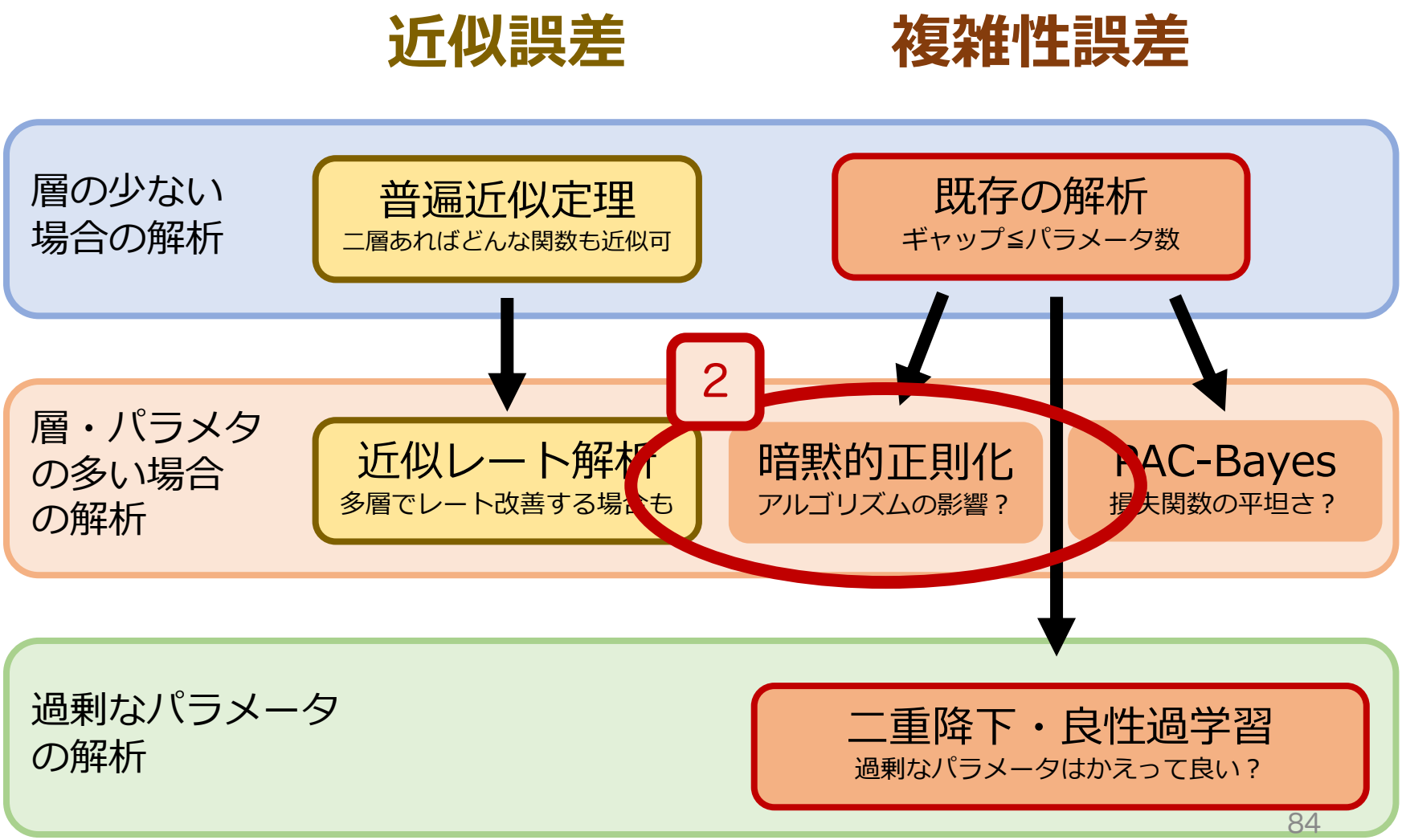
結果

- 深層構造の優位性を証明

研究2：暗黙的正則化

非凸損失局面と過適合の関係

今日の発表の流れ


近似誤差
複雑性誤差
層の少ない
場合の解析
普遍近似定理
二層あればどんな関数も近似可
既存の解析
ギャップ≦パラメータ数
層・パラメタ
の多い場合
の解析
2
近似レート解析
多層でレート改善する場合も
暗黙的正則化
アルゴリズムの影響？
PAC-Bayes
損失関数の平坦さ？
過剰なパラメータ
の解析
二重降下・良性過学習
過剰なパラメータはかえって良い？

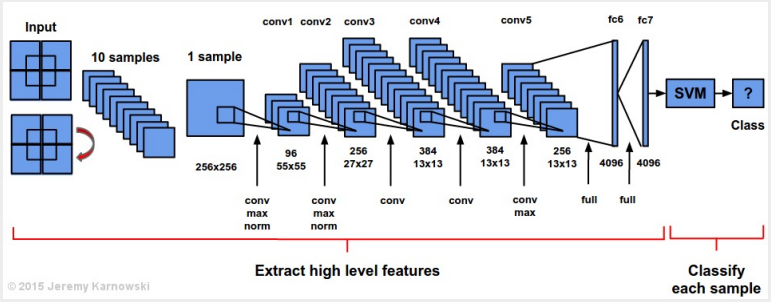
背景

暗黙的正則化理論とそれへの批判

巨大化する深層学習モデル

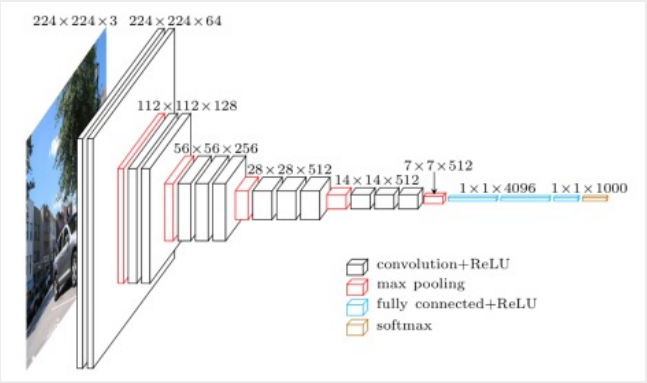
深層化に伴いパラメータ数も膨大化

AlexNet (Toronto U)



層の数：８層
パラメタ数：６千万

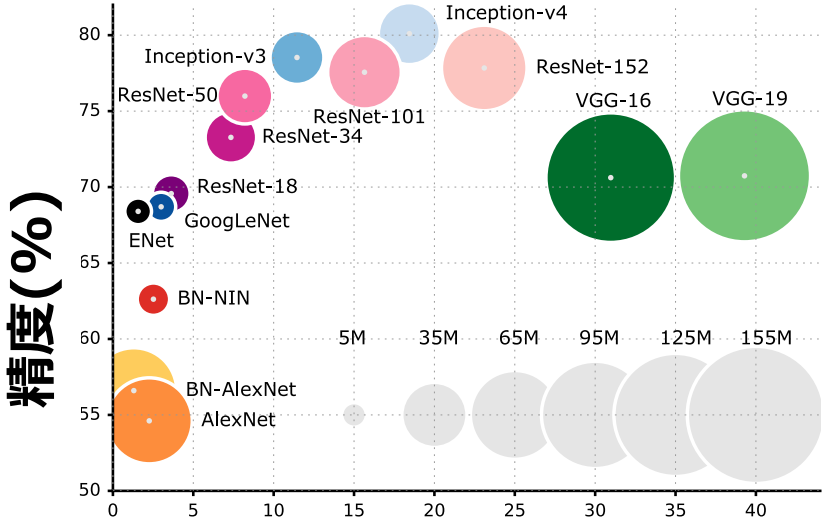
VGG19 Net (Oxford U)



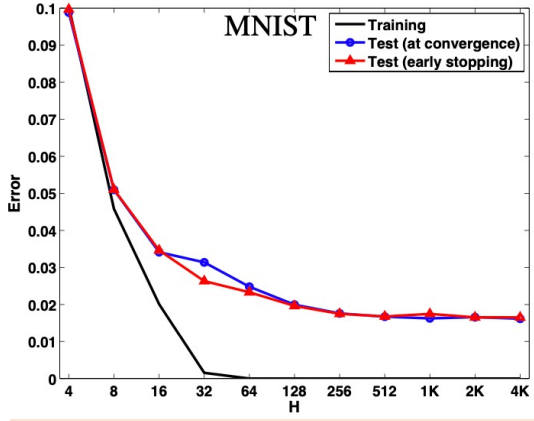
層の数：１９層
パラメタ数：１億

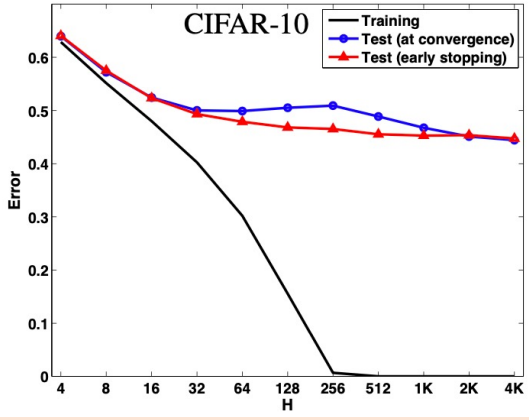
モデルの大きさが、深層学習前の数千倍規模に

実際の深層学習は過適合しない



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

パラメタ数が増えているのに汎化誤差（期待損失）が減少

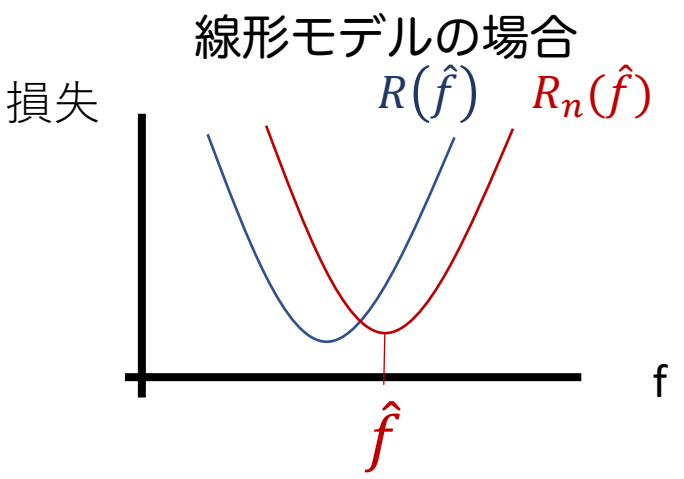
➡ 既存理論と完全に矛盾

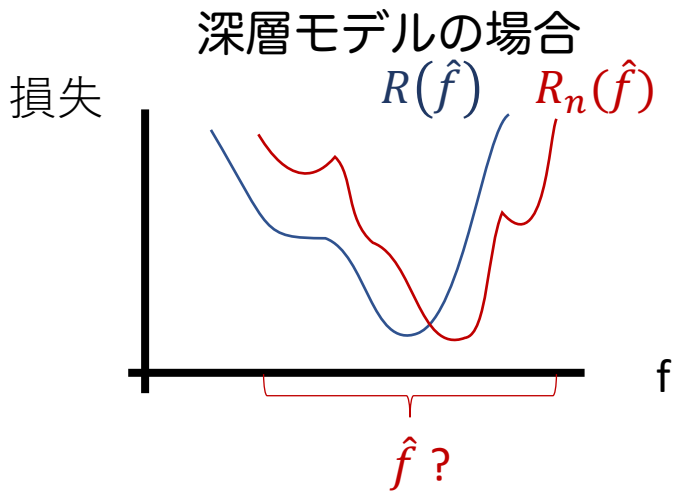
確率挙動と深層構造

- 深層モデルは確率的な変動の記述を難しくする

n個のi.i.d.観測 $Z_1, \ldots, Z_n \in \mathcal{Z}$
学習されたモデル $\hat{f}: \mathcal{Z} \to \mathbb{R}$, 損失 $\ell(Z, f)$
経験誤差: $R_n(f) = n^{-1} \sum_{i=1}^{n} \ell(Z_i, f)$
汎化誤差: $R(f) = E[\ell(Z, f)]$

過学習の尺度： $R(\hat{f}) - R_n(\hat{f})$

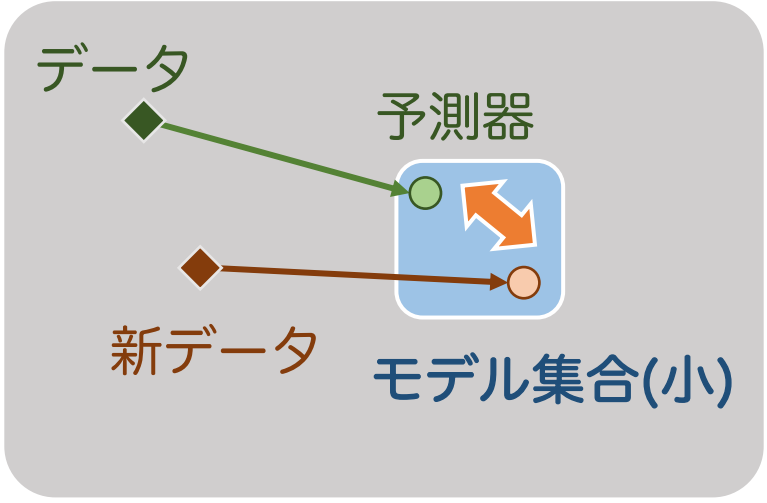




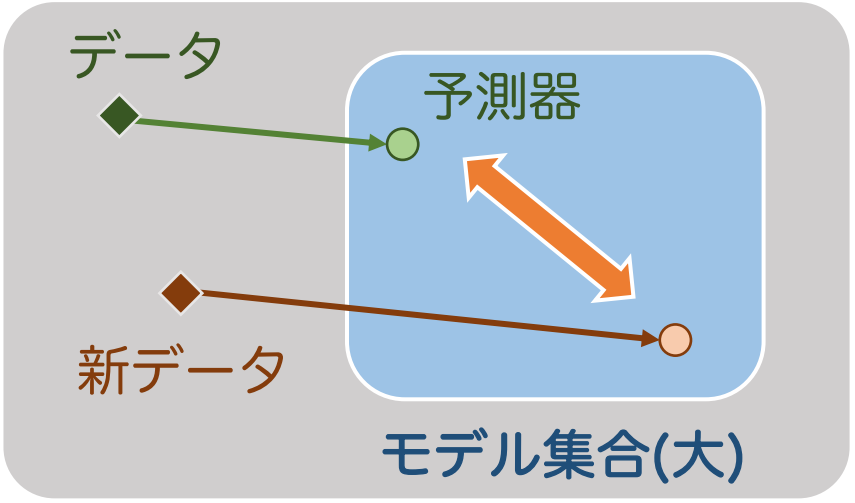
過適合とは？

データの変動に対するモデルの不安定さ

- モデル集合が大きいと、予測器はデータの変動に敏感
- 巨大モデル（パラメータ数大）は不安定



モデルが小さいので**変動**も小さい



モデルが大きいので**変動**も大きい(過適合)

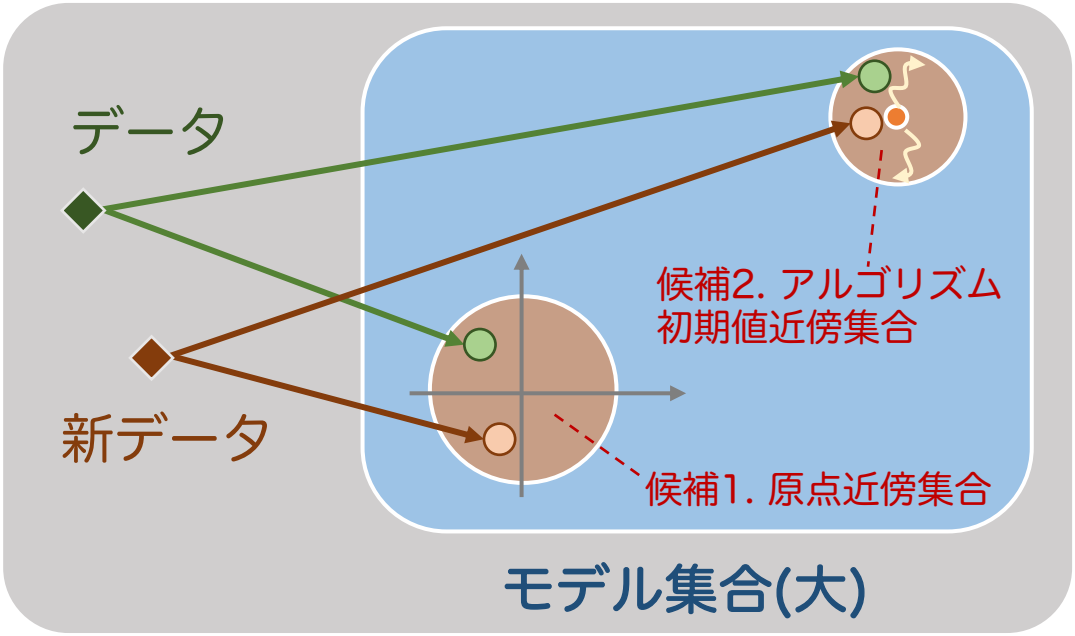
既存理論の主張

過学習による誤差＝モデル集合の大きさ（パラメータの数）

深層学習の過適合の新仮説

暗黙的正則化仮説：実質的なモデル集合はごく一部
- 学習される予測器は特定の部分集合に滞留すると仮定

その部分集合の
正体はなに？



学習された予測器がその部分集合に滞留するなら
深層学習の精度が説明できる（誤差がパラメータ数に依存しない）

Bartlett, P. L., Foster, D. J., & Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. NIPS.

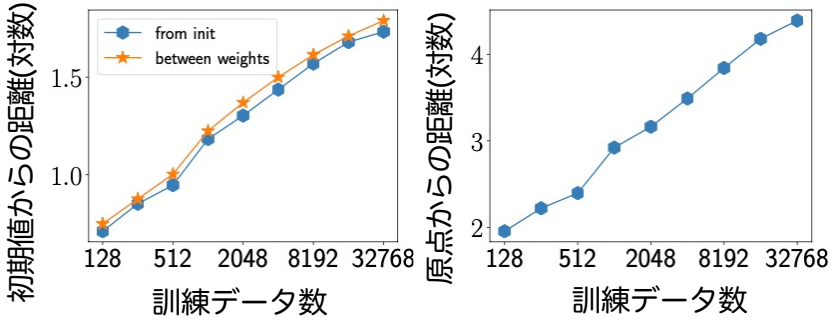
反例の登場：滞留への批判

近年の発見：候補の部分集合(原点・初期値近傍)では
滞留がおこらないことが立証

Nagarajan & Kolter (2019)

実験：パラメタは原点・初期値の近傍に滞留しない

理論：原点・初期値近傍への滞留は数学的に反証



実験上で、データ数（横軸）が増えることに
学習されたモデルが原点・初期値から遠ざかる

理論の前提条件（予測器の滞留）が成立しない
→ 問：どういう部分集合なら滞留するのか？

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

本研究

損失形状を用いた正則化理論の提案

準備：学習アルゴリズムの定式化

モデルのパラメータ学習は損失関数最小化

- 確率勾配降下法(SGD)

問題の定式化

データ $Z_1, \ldots, Z_n$ / 損失関数 ℓ

DNNによる関数 $f \in \mathcal{F}_{NN}$

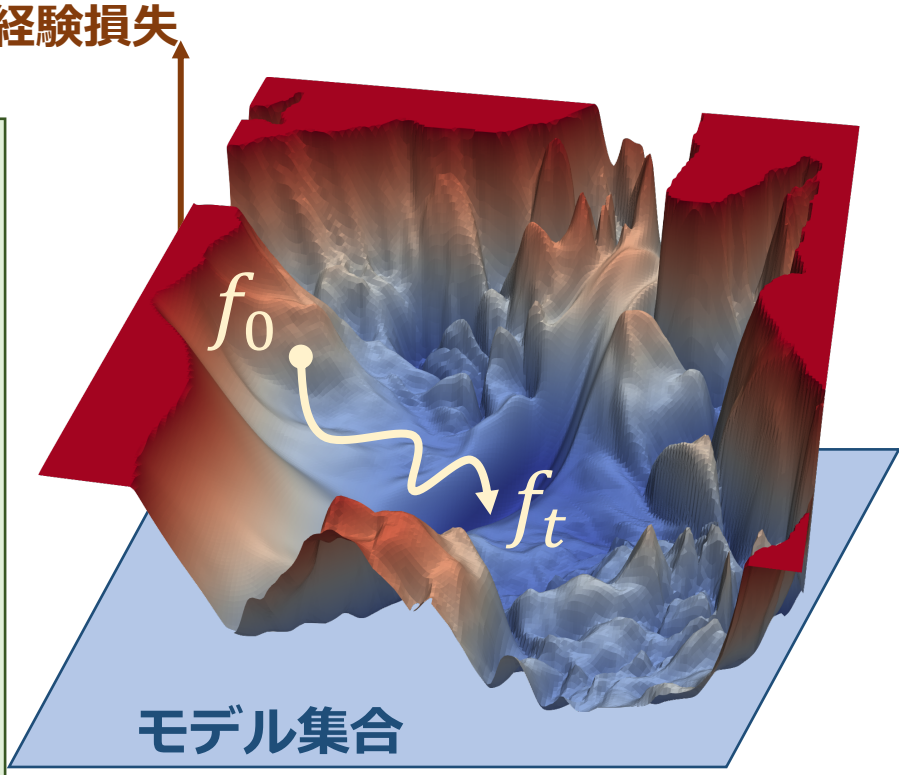
経験損失 $R_n(f) = n^{-1} \sum_{i=1}^{n} \ell(Z_i, f)$

期待損失 $R(f) = E[\ell(Z, f)]$

学習アルゴリズム(SGD)

$$f_{t+1} = f_t - \eta \hat{R}_n(f_t)$$

(η: 学習率, $\hat{R}_n$: ミニバッチ損失)

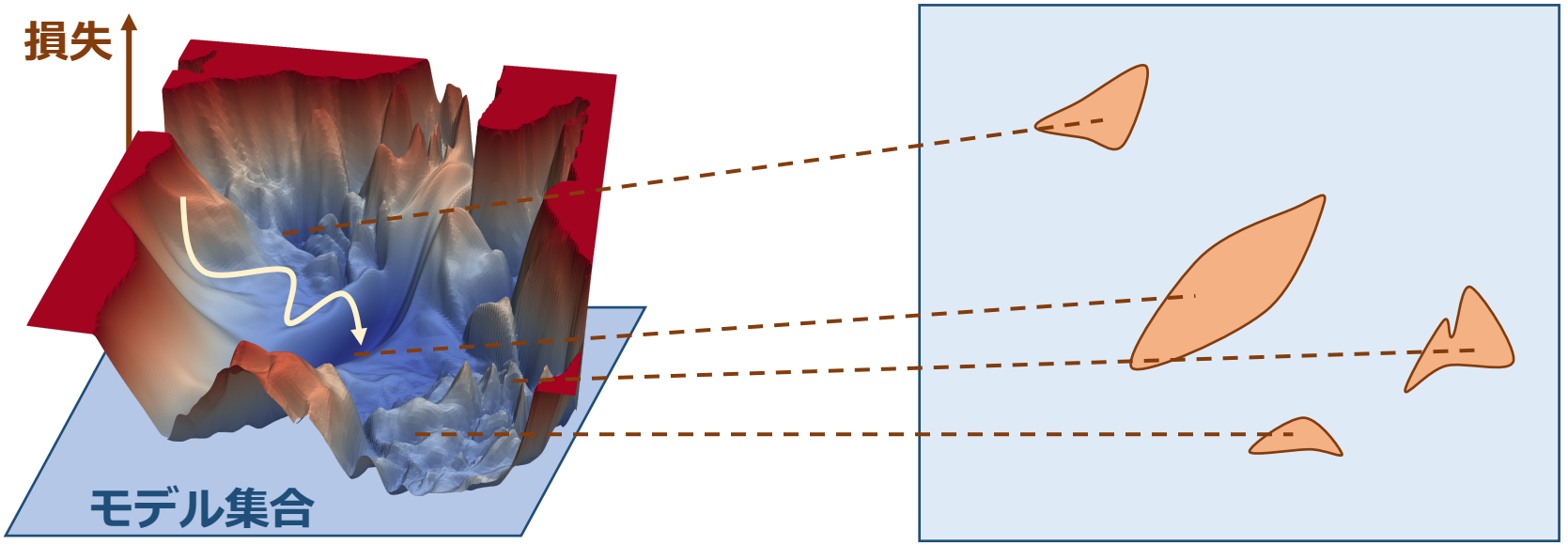


可視化（次元圧縮）された深層学習の損失関数 (Li et al. 2018)

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. NIPS.

本研究の着想

- 損失関数は複雑な形状
  → アルゴリズムは多くの谷で滞留しうる？



可視化（次元圧縮）された損失関数

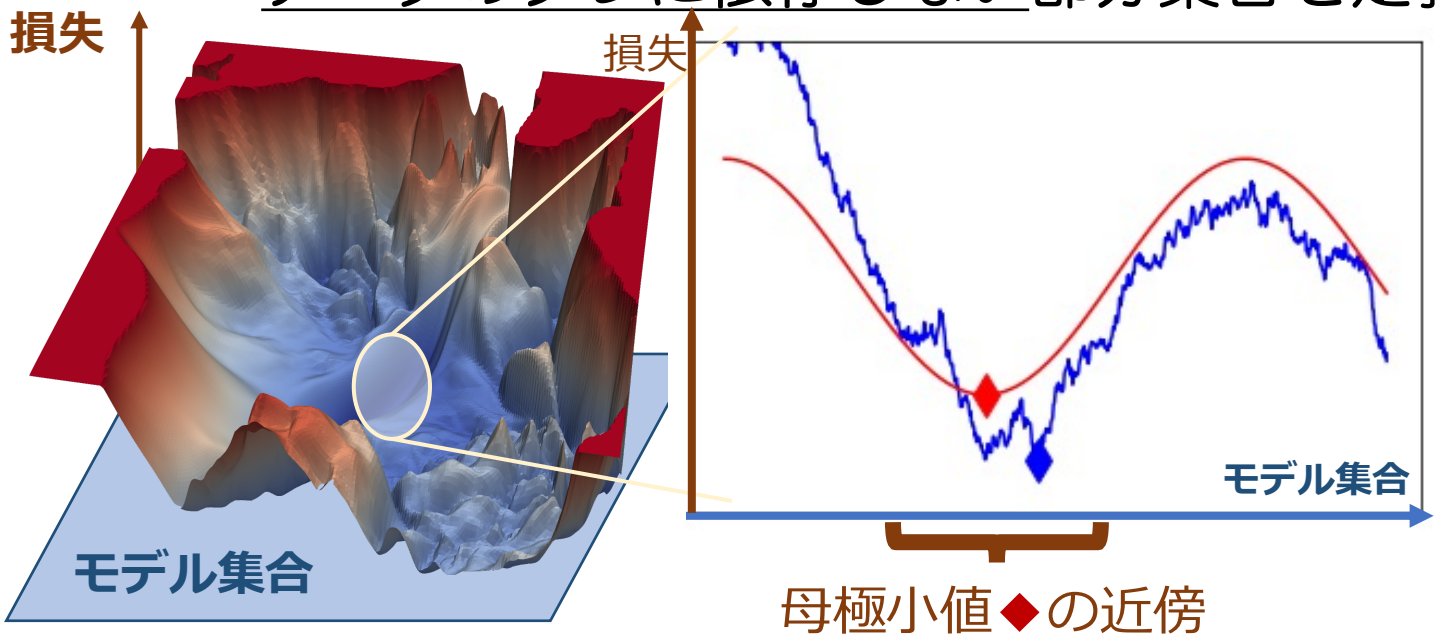
ただし、経験損失は確率変数（確率過程）なので、
理論的な解析にはそのままでは使えない

新しい暗黙的正則化仮説

アイディア: 母極小値(期待損失の極小値)の近傍

- 経験損失$R_n(f)$ の極小値は$R(f)$の極小値の近傍に集中

→データのブレに依存しない部分集合を定義



仮説：パラメタ空間に点在する母極小値が暗黙的正則化を実現する？

滞留保証の誤差理論

δ: 近傍半径
L: DNNの層の数
S: パラメタのスペクトル
K: 局所最小値の数

主定理

確率$C_B\lambda(B_\delta)$以上でアルゴリズムは
母極小値の近傍B_δに滞留し、その時以下が成立：

$$R(\hat{f}^{DL}) \leq \inf_{f \in B_\delta} R(f) + \tilde{O}\left(\frac{K + \delta LS}{\sqrt{n}}\right)$$

深層学習の汎化誤差

母極小値近傍の損失形状
(パラメータ数に不依存)

利点

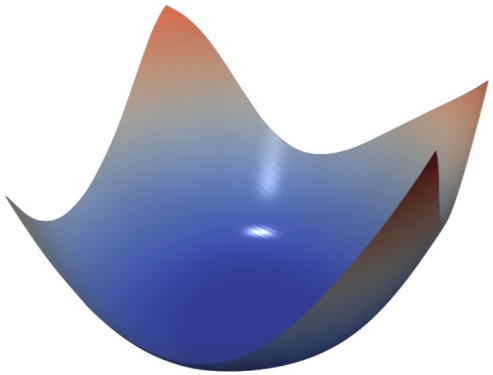
- 母極小値近傍B_δへの滞留を保証
- パラメータ数に依存しない理論誤差評価

Imaizumi, M., & Schmidt-Hieber, J. (2020). Generalization Analysis of Deep Neural Networks with Population Minima.

研究２のまとめ

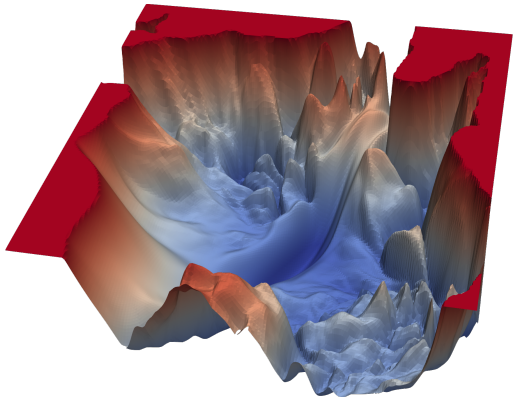
キーワード：暗黙的正則化と滞留

- 本研究の提案する直感
  - 複雑な非凸損失関数が、アルゴリズムを滞留させる
  - 母極小値の近傍集合が予測器の安定性を増す



従来法の損失関数

複雑化



深層学習の損失関数

予想

深層学習のモデル巨大化

↓

損失関数の複雑化

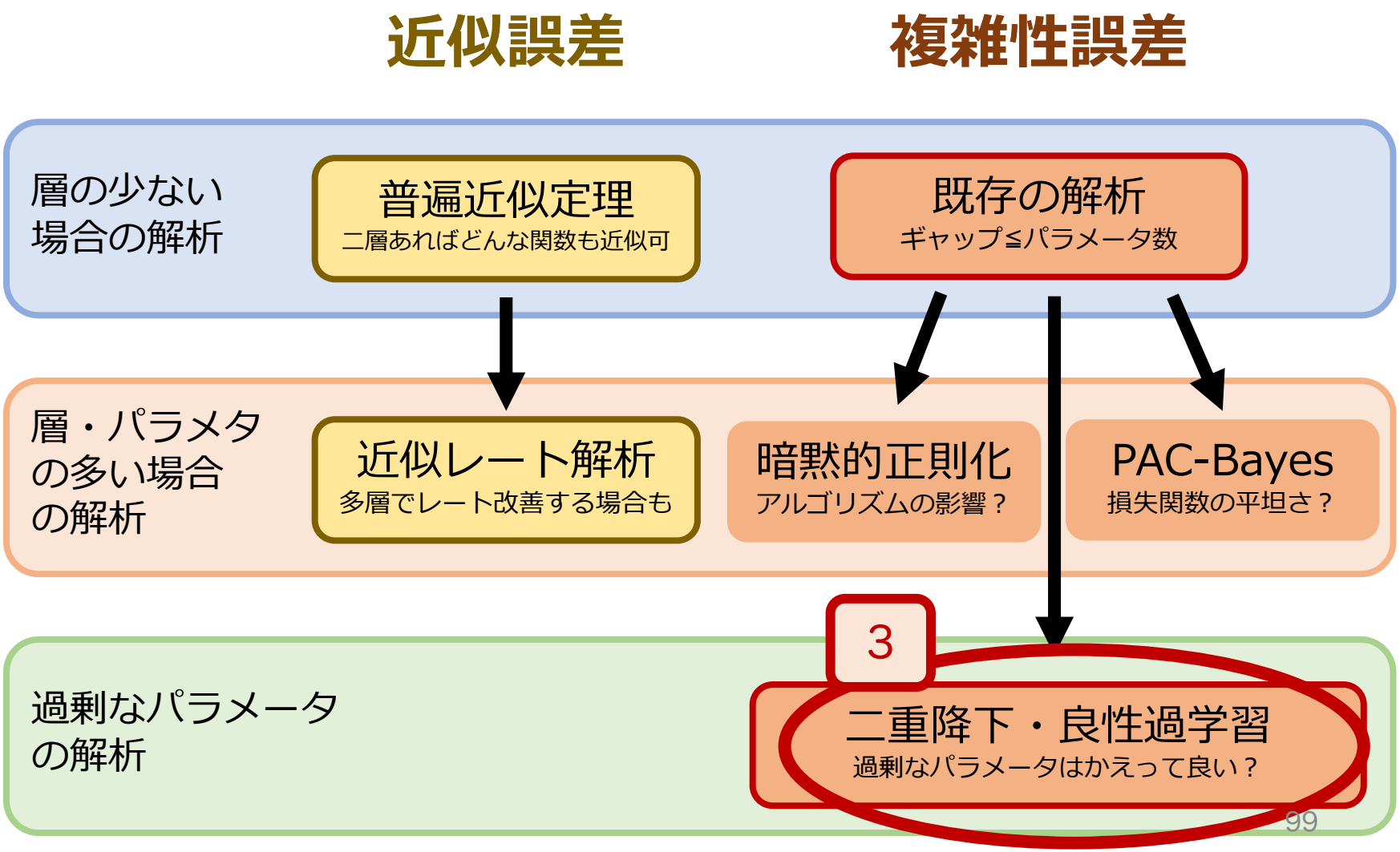
↓

安定的な母極小値の増加

研究3：過剰パラメータ

二重降下とその拡張

今日の発表の流れ

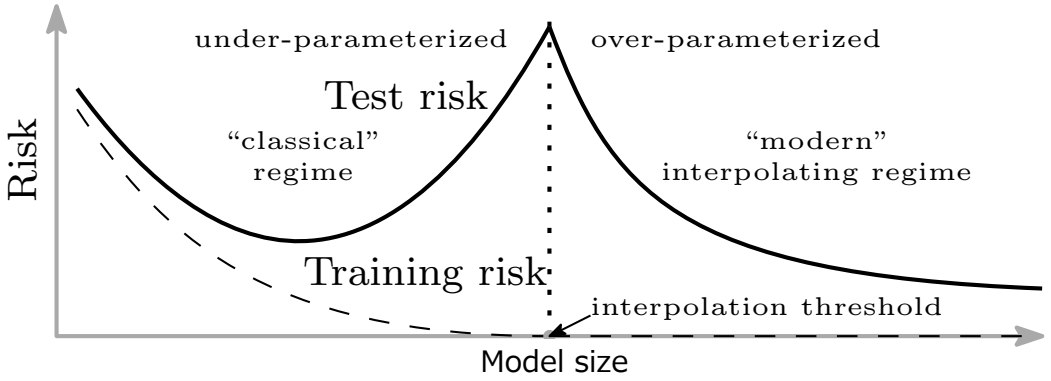

近似誤差
複雑性誤差
層の少ない
場合の解析
普遍近似定理
二層あればどんな関数も近似可
既存の解析
ギャップ≦パラメータ数
層・パラメタ
の多い場合
の解析
近似レート解析
多層でレート改善する場合も
暗黙的正則化
アルゴリズムの影響？
PAC-Bayes
損失関数の平坦さ？
過剰なパラメータ
の解析
3
二重降下・良性過学習
過剰なパラメータはかえって良い？

背景

暗黙的正則化理論とそれへの批判

近年のトピック：二重降下

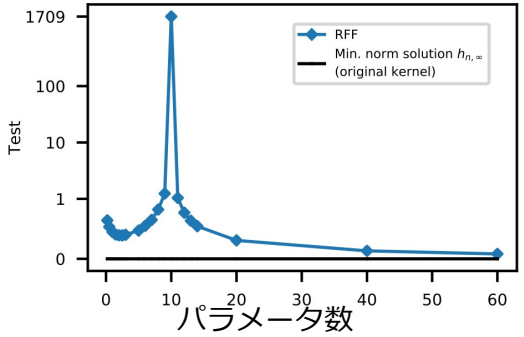
二重降下現象:大きなモデルが漸近的なリスクを減少させる



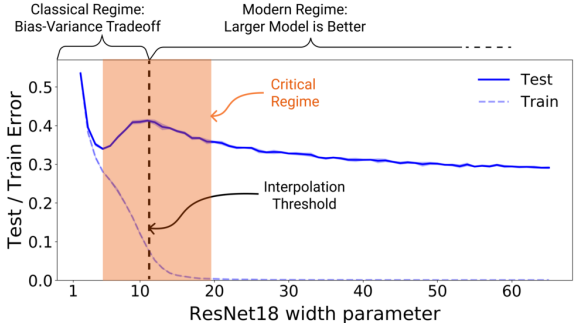
モデルサイズ(γ)が増加すると
リスクが(再び)現象

二重降下のコンセプト図
（Belkin+ (2020)）

- 実験的に実証



ランダム特徴量モデル
Belkin+ (2020)



深層学習
(ResNet / CNN)
Nakkiran+ (2020)

Belkin, Hsu, Ma, and Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the NationalAcademy of Sciences, 116(32):

Nakkiran, Kaplun, Bansal, Yang, Barak, and IlyaSutskever. Deep double descent: Where bigger models and more data hurt, ICLR 2020.

過剰パラメータ理論による解析

β^*: 真のパラメータ
r^2: β^* のL2ノルム
σ^2: ノイズεの分散

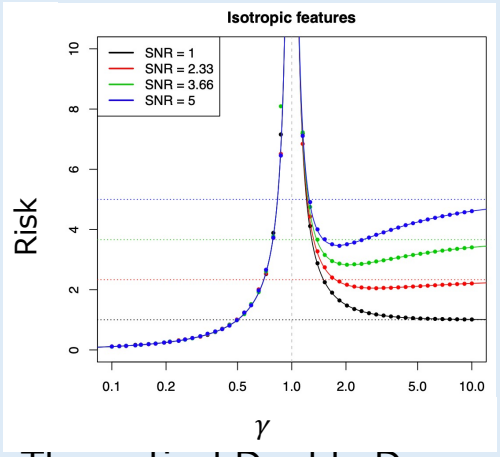
理論的解析

- 線形回帰

$$Y = X^\top \beta + \varepsilon$$

  - $R(\hat{\beta}) \coloneqq E\left[\left(X^\top \hat{\beta} - X^\top \beta^*\right)^2 \middle| X\right]$,
  - $\hat{\beta}$: リッジレス推定量（最小ノルム解）

$$R(\hat{\beta}) = \begin{cases} \sigma^2 \dfrac{\gamma}{1-\gamma} & (\gamma < 1) \\ r^2(1-\gamma^{-1}) + \sigma^2 \dfrac{1}{\gamma - 1} & (\gamma > 1) \end{cases}$$



Theoretical Double Descent
by Hastie+(2019)

- カーネル/特徴量回帰 (Mei+ 2019, Liang+ 2020, etc)
- 線形分類器 (Montanari+ 2019, Lolas 2020, etc)
- (層の少ない)ニューラルネット (Ba+ 2019)

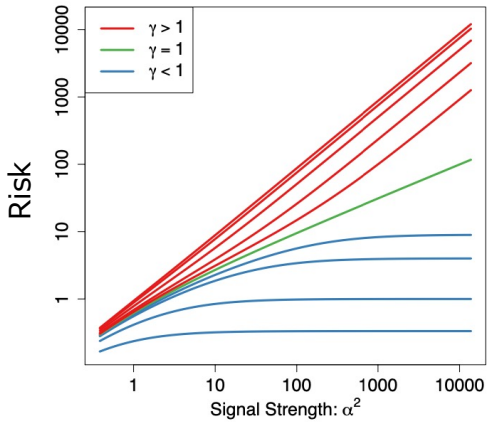
Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises inhigh-dimensional ridgeless least squares interpolation.
Mei and Montanari. The generalization error of random features regression:Precise asymptotics and double descent curve, 2019.
Liang, Rakhlin,. Just interpolate: Kernel “ridgeless” regression cangeneralize.Annals of Statistics, 2020.
Montanari, Ruan, Sohn, and Yan. The generalization error ofmax-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime.
Lolas. Regularization in high-dimensional regression and classification via random matrix theory, 2020
Ba, Erdogdu, Suzuki, Wu, and Zhang. Generalization oftwo-layer neural networks: An asymptotic viewpoint. InInternational Conference on LearningRepresentations, 2019.

異なる過剰パラメータ理論

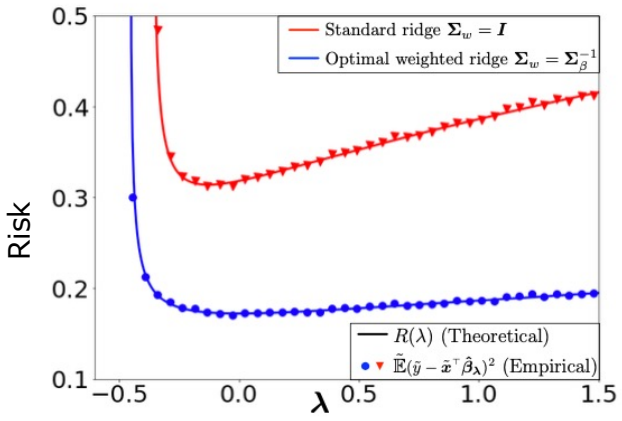
正則化付きモデルの漸近リスク

- 正則化が付くことで、二重降下とは異なる形
- 例：リッジ回帰
  - 正則化推定量 $\hat{\beta}_\tau$

$$\hat{\beta}_\tau = \mathrm{argmin}_\beta n^{-1} \sum^{n} \left(Y_i - X_i^\top \beta\right)^2 + \lambda \|\beta\|^2$$



リスクの相転移
Dobrian+ (2018)



最適な正則化
Wu+ (2020)

Dobriban, Wager High-dimensional asymptotics of prediction: Ridgeregression and classification.The Annals of Statistics, 46(1):247–279, 2018.
Wu and Xu. On the optimal weighted l2 regularization in overparameterized linearregression.Advances in Neural Information Processing Systems, 2020.

二重降下の理論解析

導入: 線形モデルの例 (Hastie+ 2019の例)

線形回帰

- データ $D_n = \{(X_i, Y_i)\}_{i=1}^n$, X_i はp次元(平均ゼロ)
- 線形回帰モデル

$$Y_i = {\beta^*}^\top X_i + \varepsilon_i, \quad \beta^*: p\text{次元パラメタ}$$

リッジレス推定量

$$\hat{\beta} = \mathrm{argmin}\{\|\beta\|_2 : \beta \ minimizes \ \Sigma_{i=1}^n (Y_i - \beta^\top X_i)^2\}$$

汎化誤差の分解

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

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

$$R(\hat{\beta}) = E_\varepsilon\left[\|\hat{\beta} - \beta^*\|_\Sigma^2\right] = \|E_\varepsilon[\hat{\beta}] - \beta^*\|_\Sigma^2 + \mathrm{tr}[\mathrm{Cov}_\varepsilon(\hat{\beta})\Sigma]$$

$= B$ (バイアス)　$= V$ (バリアンス)

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises inhigh-dimensional ridgeless least squares interpolation.

理論のキーは？

V(バリアンス)を経験共分散行列の固有値で表現

- $\boldsymbol{X} = (X_1, \ldots, X_n), \boldsymbol{Z} = \Sigma^{1/2}\mathbf{X}$ （$p \times n$行列）
- 経験共分散行列 $\hat{\Sigma} = \boldsymbol{X}\boldsymbol{X}^\top/n$ (ランダム行列)
- $\lambda_j(A)$: 行列Aのj番目固有値

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

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

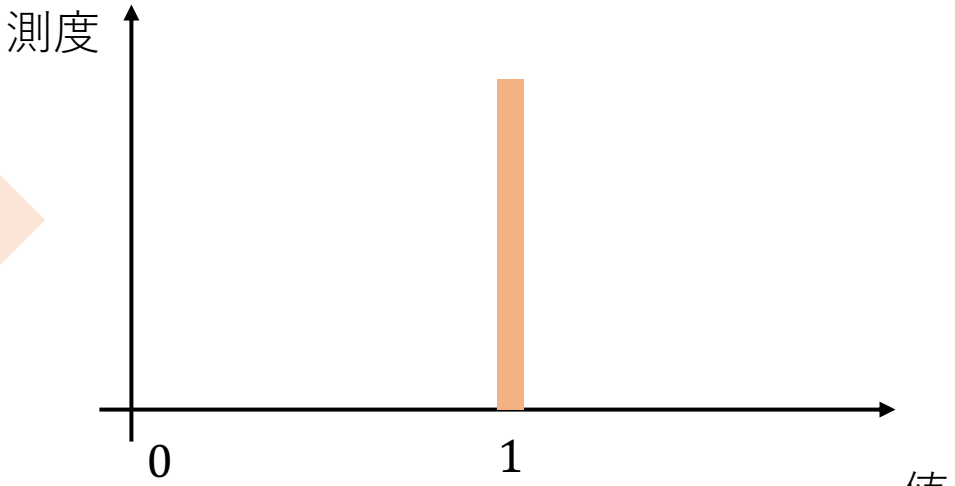
$F_{\boldsymbol{Z}\boldsymbol{Z}^\top/n}(s)$: 行列$\boldsymbol{Z}\boldsymbol{Z}^\top/n$の固有値分布

固有値分布の例

単位行列

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

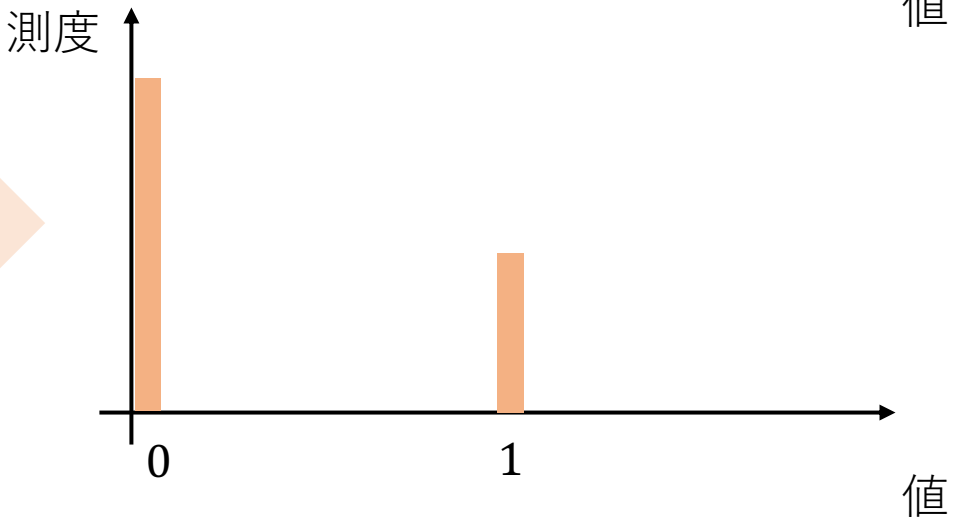
固有値は
1,1,1



特異行列

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

固有値は
1,0,0

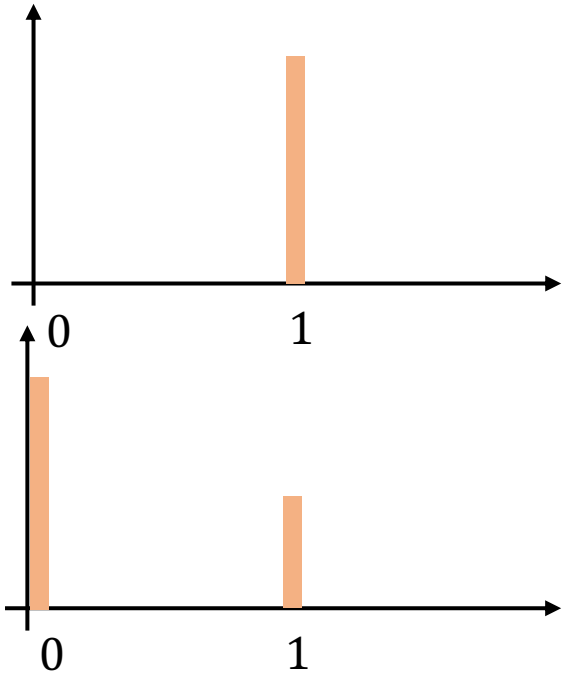


行列が多様な”情報”を持つ時、分布が右に寄る

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）

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



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上に測度を持つかが重要

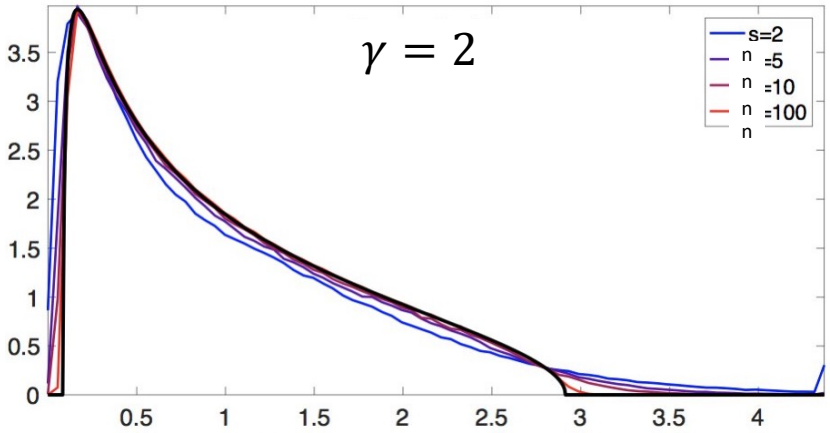
キーとなる固有値分布

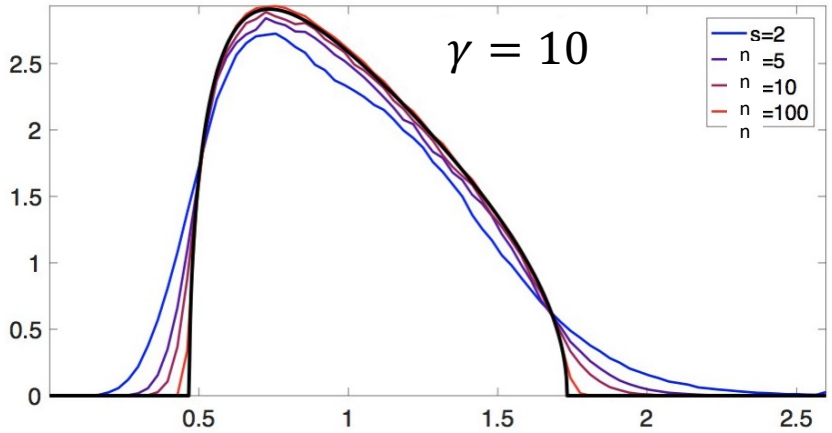
マルチェンコ＝パスツール則 (MP則)

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

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

Peyre(2020)





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

($p, n \to \infty$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

理論的な説明

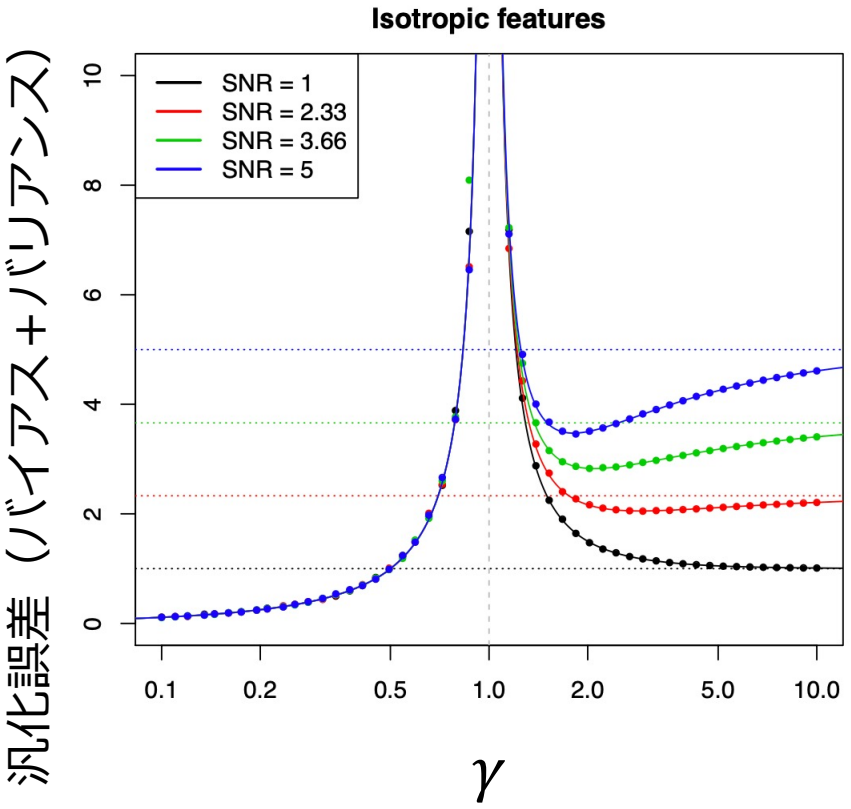
線形モデルの二重降下

線形回帰の汎化誤差のバリアンス項

(Hastie+ (2019))

$$\lim_{p,n\to\infty} V = \begin{cases} \dfrac{\sigma^2\gamma}{\gamma-1}, (\gamma < 1) \\ \dfrac{\sigma^2}{\gamma-1}, (\gamma > 1) \end{cases}$$

σ^2: ノイズε_iの分散



Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises inhigh-dimensional ridgeless least squares interpolation.

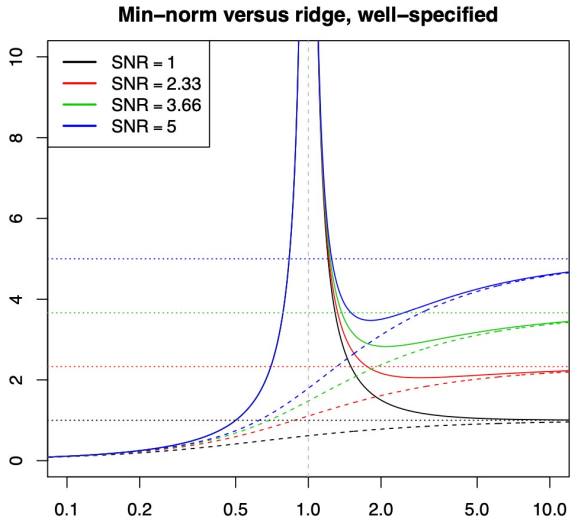
リッジ付き線形回帰の場合

正則化を導入した線形回帰の推定量

$$\hat{\beta} = \mathrm{argmin}_{\beta} \Sigma_{i=1}^{n} (Y_i - \beta^{\top} X_i)^2 / n + \lambda \|\beta\|^2$$

(Hastie+ (2019))

$$V = \frac{\sigma^2 p}{n} \int \frac{1}{s + \lambda} d F_{ZZ^{\top}/n}(s)$$



類似した方法で、漸近的な
リスクが記述できる（点線部分）
二重降下はしないが発散もしない

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises inhigh-dimensional ridgeless least squares interpolation.

本研究

非線形化

問題意識: 複雑なモデルは？

先行研究

- 解析対象: 特徴量写像に線形なモデル

$$\min_{\beta=(\beta_1,..,\beta_p)} \sum_{i=1}^{n} \ell\left(Y_i, g_\beta(X_i)\right) \; s.t. \; g_\beta(x) = \sum_{j=1}^{p} \beta_j \psi_j(x)$$

  - ℓ: 損失, $\psi_j(x)$: 特徴量写像（所与）

例

  - 線形回帰 $\psi_j(x) = x_j$
  - カーネル回帰 $\psi_j(x) = k(x, X_j)$
  - 2層ニューラルネット $\psi_j(x) = \sigma(A_j x), A_j \sim P$

証明テクニックから来る制約

我々の関心

深層ニューラルネットのような
より広いクラスの非線形モデルはどうなる？

我々の設定

正則化つき最尤推定

- $Z_1, \ldots, Z_n$: i.i.d. 観測
- $f_\theta(z)$: 尤度関数 (例：任意のモデル＋損失)
  - θ: p次元パラメタ, $\tau > 0$: 正則化係数

$$\hat{\theta} = \mathrm{argmin}_\theta \; -n^{-1} \sum_{i=1}^{n} \log f_\theta(Z_i) + \frac{\tau}{2} \|\theta\|^2$$

- θ^*: 真のパラメータ ($Z_i \sim f_{\theta^*}$)
- $F_p^* = E[\partial^2 \log f_\theta(Z)]$: Fisher情報行列(FIM)

目的：推定誤差の評価

$$\left\|\hat{\theta} - \theta^*\right\|_{F_p^*}^2 := \left(\hat{\theta} - \theta^*\right)^\top F_p^* \left(\hat{\theta} - \theta^*\right)$$

アイディア

準備: 先行研究の方法

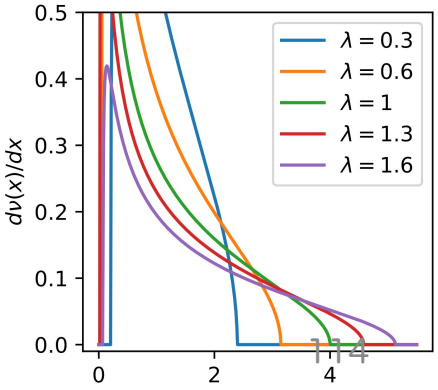
特徴量写像に線形なモデルの場合

- $E\left[\left\|\hat{\beta}-\beta^*\right\|_\Sigma^2\right]=\left\|E\left[\hat{\beta}\right]-\beta^*\right\|_\Sigma^2+\mathrm{tr}\left(\mathrm{Cov}\left(\hat{\beta}\right)\Sigma\right)=:B\left(\hat{\beta}\right)+V\left(\hat{\beta}\right)$

分散項: $V\left(\hat{\beta}\right)\asymp\frac{\sigma^2}{n}\mathrm{tr}\left(\left(\frac{\Psi^\top\Psi}{n}\right)^{-1}\right)\to M_{MP}(\gamma),(n,p\to\infty,p/n\to\gamma)$

- $\Psi=(\psi(X_1),\ldots,\psi(X_n))$: $p\times n$ ランダム特徴量行列
- $M_{MP}(\gamma)$: Marchenko-Pastur (MP) 則のStieltjes変換

- リスクがランダム行列の積の逆行列
  → **Marchenko-Pastur則** (MP則)で極限のリスクを記述できる



Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani.
Surprises inhigh-dimensional ridgeless least squares interpolation.

アイディア

尤度の性質を用いてランダム行列へ帰着

我々の設定：MLE

- $V(\hat{\theta})$: $\hat{\theta} - \theta^*$ のバリアンス項
- $M_n(\theta) \coloneqq -n^{-1}\sum_{i=1}^{n} \log f_\theta(z_i)$

ランダム行列の積！

$$\left\|V(\hat{\theta})\right\|_{F_p^*}^2 \preccurlyeq \frac{1}{n}\mathrm{tr}\left((\partial^2 M_n(\theta^*) + \tau I)^{-1}\right) + O(R) \approx \frac{1}{n}\mathrm{tr}\left(\left(\frac{\hat{J}\hat{J}^\mathsf{T}}{n} + \tau I\right)^{-1}\right) + O(R)$$

$$\to M'_{MP}(\gamma) + o(1)$$

尤度モデルのFIMの近似: $\partial^2 M_n(\theta^*) \approx n^{-1}\hat{J}_{n,p}\hat{J}_{n,p}^\mathsf{T}$

- $\hat{J} = (\partial \log f_\theta(X_1), \ldots, \partial \log f_\theta(X_n))$: $p \times n$ ランダムなJacobi行列
- $M'_{MP}(\gamma)$: 拡張されたMP則のStieltjes変換
  - 列について独立なランダム行列で有効

主結果

Δ：非線形残差

Theorem 2 (漸近的バリアンス)

仮定(後述) のもとで、$\tau \searrow \bar{\tau} \geq 0$ ならば以下が成立：

$$\limsup_{p,n\to\infty,\, p/n\to\gamma} \left\|V(\hat{\theta})\right\|_{F_p^*}^2 \preccurlyeq \lim_{a\to 0} h_{\gamma,\bar{\tau}}(a) + \Delta.$$

定義

- ξ: $\lim_{p\to\infty} F_p^*$ のスペクトル測度
- $h_{\gamma,\bar{\tau}}^{(0)}(a) := \int \frac{1}{\bar{\tau}\lambda/\gamma - a} d\xi(\lambda)$: ξ の重みつきStieltjes変換
- $h_{\gamma,\bar{\tau}}(a) = h_{\gamma,\bar{\tau}}^{(0)}\left(a - \frac{1}{\gamma(1+h_{\gamma,\bar{\tau}}(a))}\right)$: 拡張されたStieltjes変換

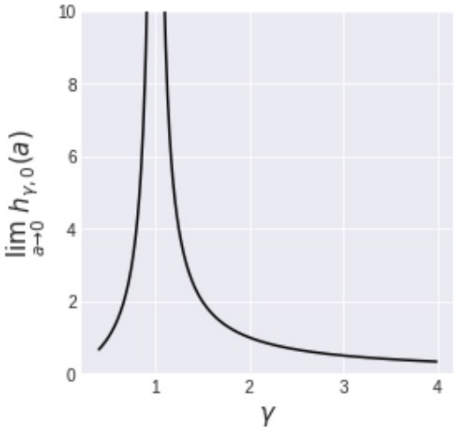
主結果

Theorem 2 (漸近的バリアンス)

<u>仮定(後述)</u> のもとで、$\tau \searrow \bar{\tau} \geq 0$ ならば以下が成立：

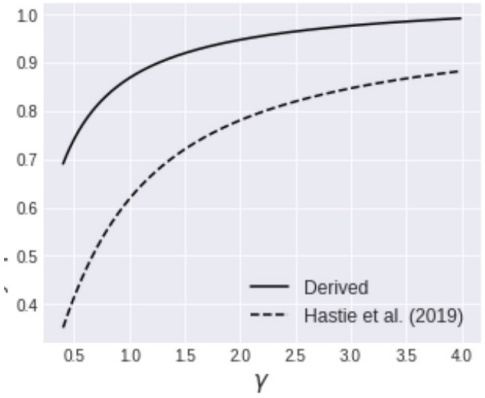
$$\limsup_{p,n\to\infty,\, p/n\to\gamma} \left\| V(\hat{\theta}) \right\|_{F_p^*}^2 \preccurlyeq \lim_{a\to 0} h_{\gamma,\bar{\tau}}(a) + \Delta.$$

$\lim_{a\to 0} h_{\gamma,\bar{\tau}}(a)$ は、漸近的なリスクを複数表現



二重降下
($\bar{\tau} = 0$)

$$\lim_{a\to 0} h_{\gamma,\bar{\tau}}(a) = \begin{cases} \dfrac{\gamma}{1-\gamma} & (\gamma < 1) \\ \dfrac{1}{\gamma - 1} & (\gamma > 1) \end{cases}$$



正則化漸近リスク
($\bar{\tau} > 0, \xi$ is Dirac)

Similar to that of Hastie+ (2019)

Other ξ is applicable

(仮定を満たす)MLEは、二重降下/正則化漸近リスクに従う

主結果

Theorem 3 (漸近的推定誤差)

仮定(後述)が成立かつ$\|\theta^*\|_2^2 \leq r^2$の時、$\tau \searrow \bar{\tau} \geq 0$：

$$\limsup_{p,n\to\infty,\, p/n\to\gamma} \left\|\hat{\theta} - \theta^*\right\|_{F_p^*}^2 \preccurlyeq \lim_{a\to 0} h_{\gamma,\bar{\tau}}(a) + \Delta + r^2.$$

バイアスr^2が残るが、これは漸近的リスク解析では通常

Corollary 1 (汎化誤差)

ノンパラメトリック回帰 $z_i = (y_i, x_i), y_i = g_{\theta^*}(x_i) + \varepsilon_i$ の問題で、

仮定(後述)が成立かつ$\|\theta^*\|_2^2 \leq r^2$の時、$\tau \searrow \bar{\tau} \geq 0$：

$$\limsup_{p,n\to\infty,\, p/n\to\gamma} \left\|g_{\hat{\theta}} - g_{\theta^*}\right\|_{L^2}^2 \preccurlyeq \lim_{a\to 0} h_{\gamma,\bar{\tau}}(a) + \Delta + r^2 + r^4.$$

証明の概要

バイアス・バリアンス分解

- $H_{n,p} = (\partial_\theta^2 M_n(\theta^*) + \tau I)$: 正則化つきHesse行列
- R: $M_n(\theta)$の2次Taylor展開の残差

$$\hat{\theta} - \theta^* = \underbrace{H_{n,p}^{-1}(\partial_\theta M_n(\theta^*) + R)}_{=: V(\hat{\theta})\ (\text{バリアンス})} + \underbrace{H_{n,p}^{-1}\tau\theta^*}_{=: B(\hat{\theta})\ (\text{バイアス})}$$

バリアンス$V(\hat{\theta})$の評価

$$V(\hat{\theta}) \approx \left(\hat{F}_{n,p} + \tau I\right)^{-1}(\partial_\theta M_n(\theta^*) + R)$$

(EIFM近似： $\hat{F}_{n,p} \approx \partial_\theta^2 M_n(\theta^*)$)

$$= \underbrace{\left(\hat{F}_{n,p} + \tau I\right)^{-1}\partial_\theta M_n(\theta^*)}_{\text{MP則を適用}} + o_P(1)$$

(Taylor残差： $R = o_P(1)$)

研究3のまとめ

目的

- 深層モデルでの二重効果を議論したい

アプローチ

- フィッシャー情報行列の分解とMP則

結果

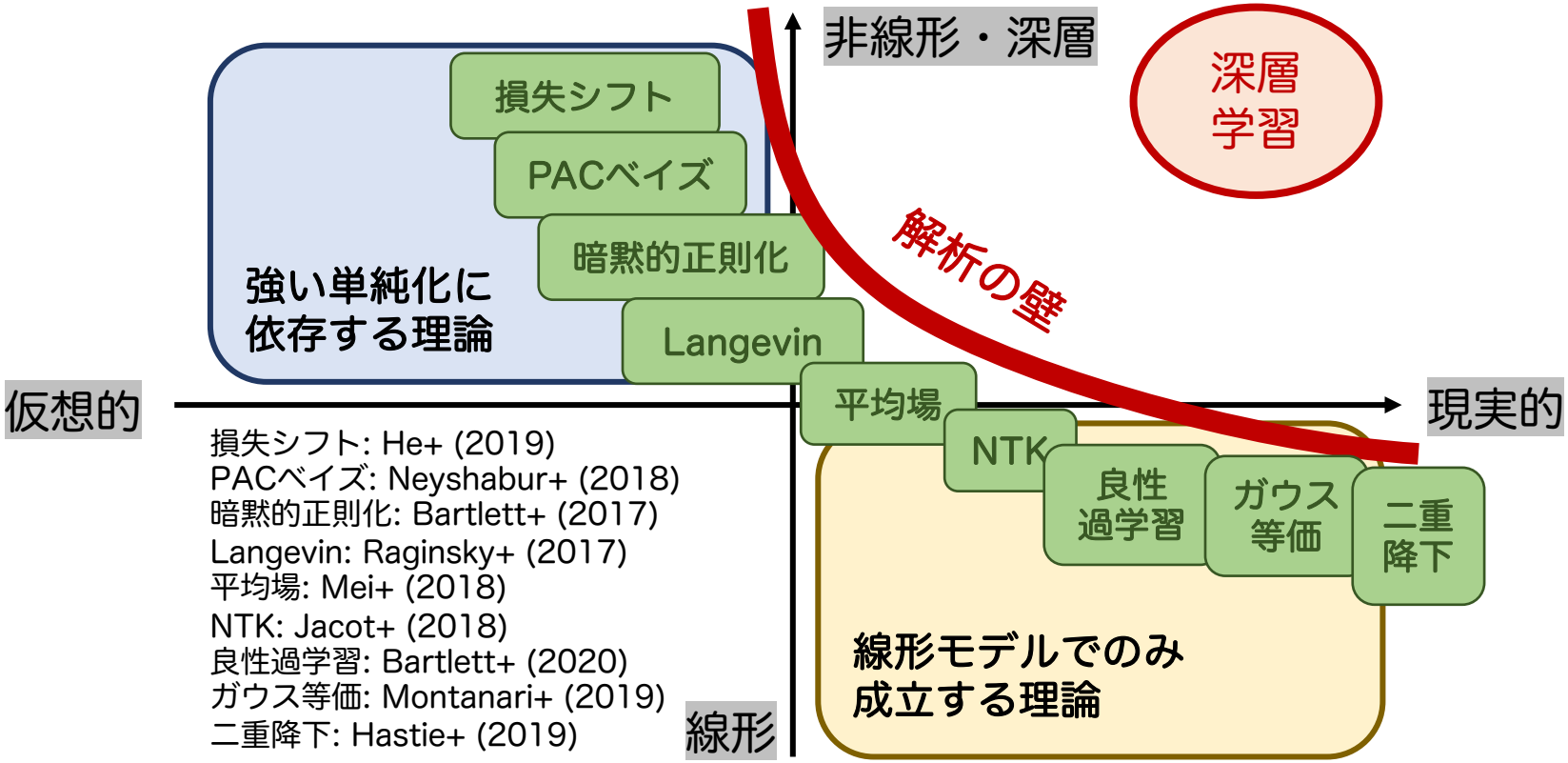
- 二重効果を起こすモデルと条件を特定

まとめと今後

新理論の挑戦と限界

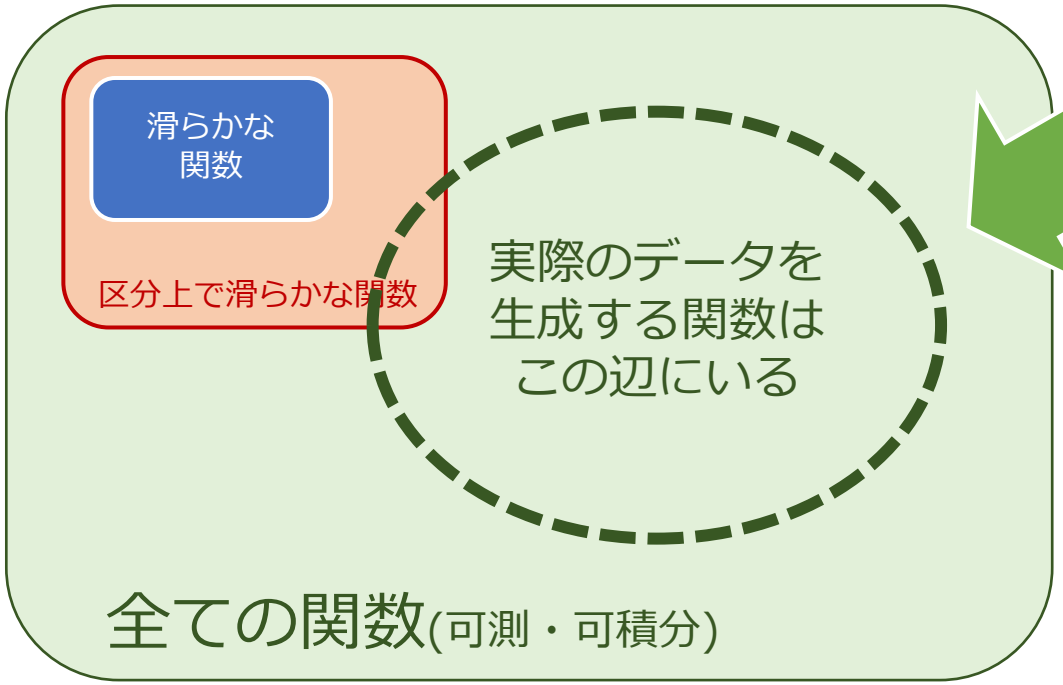
近年の深層学習理論

- 革新的な理論が数多く提案、しかし記述できない点は多い



理論の壁1：線形近似の限界

- 解析できている関数集合(滑らかな関数)は狭い
  - 全関数の集合は大きい(測り方が分からないくらい)



ここを調べれば
いいんじゃないの？

理論の壁1：線形近似の限界

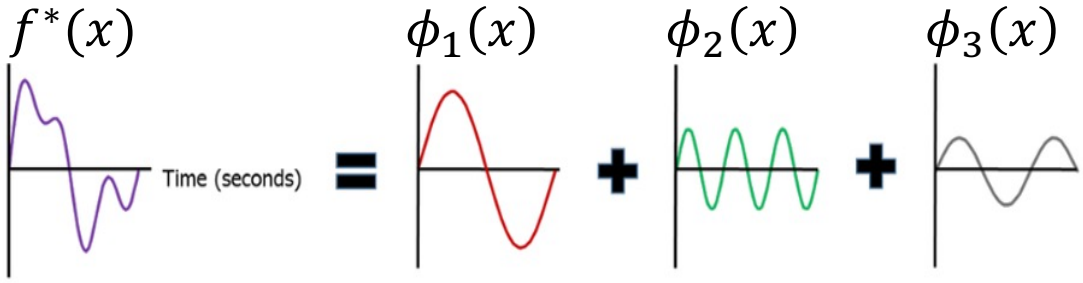
問題：関数の解析には関数の線形表現がほぼ必須

関数の線形表現

f^*が滑らかなら

$$f^*(x) = \sum_j c_j \phi_j(x), c_j \in \mathbb{R}$$

$\phi_j(x)$: 基底関数



例：C^α級+多項式基底、Sobolev空間+三角関数基底、Besov空間+Wavelet基底

- しかし、線形和＝１層で表現できる操作なので、

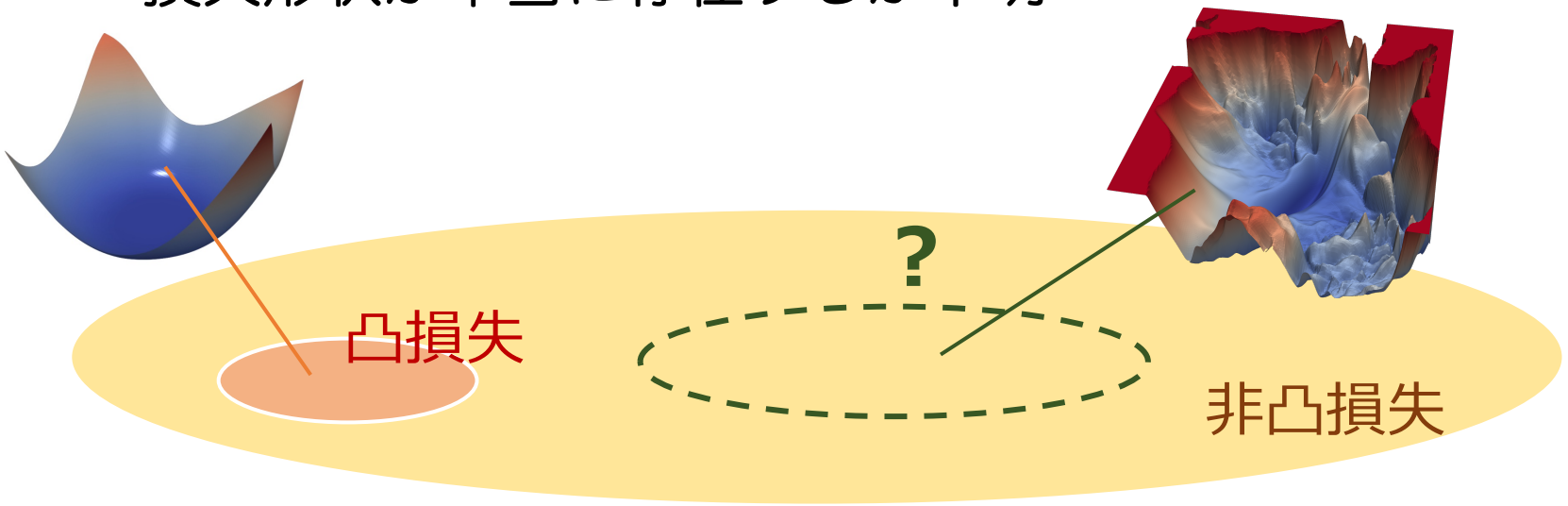
線形表現できる＝非深層モデルでも最適近似できる

- 深層学習を十分に説明する理論には、線形表現に依存しない関数解析が必要
  - 区分上で滑らかな関数はその一具体例

理論の壁2：非凸関数の不透明さ

問題：損失関数の形状には理論的な不明点が多い

- 例：極小値の数 / 近傍構造 / 稠密さ
- 損失形状が本当に存在するか不明



形状の性質が分からないため、過適合の理論も確立されない

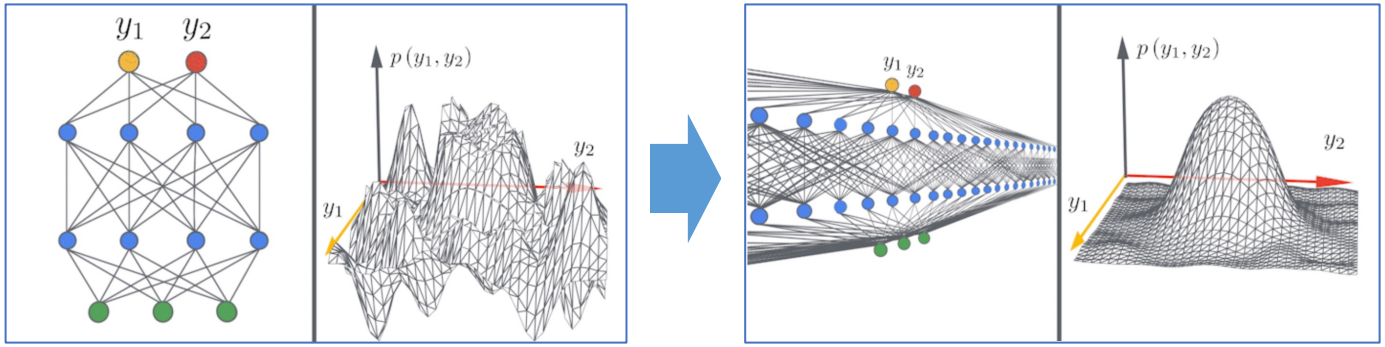
→ 非凸関数を分類・特徴付ける理論が必要

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. NIPS.

理論の壁2 ：非凸関数の不透明さ

現在の深層学習研究の主流

- 凸に近い状況になるように設計を工夫する



例： Neural Tangent Kernel: 層は少なく幅を増やすと損失は凸関数に近づく

- ただし、深層学習固有の良さは失われる

将来的に必要な基礎理論

- 非凸関数を理解し、それに基づく体系を構築

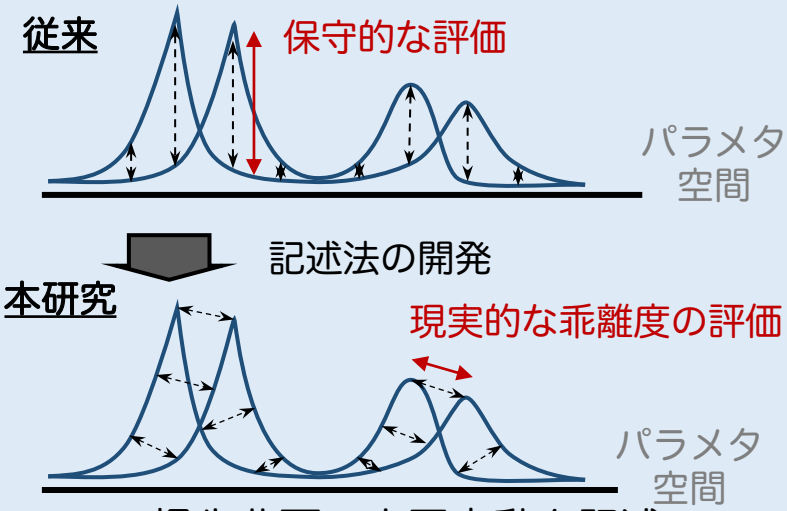
https://ai.googleblog.com/2020/03/fast-and-easy-infinitely-wide-networks.html

今後の個人的な研究方針

深層学習の確率的挙動の精緻化

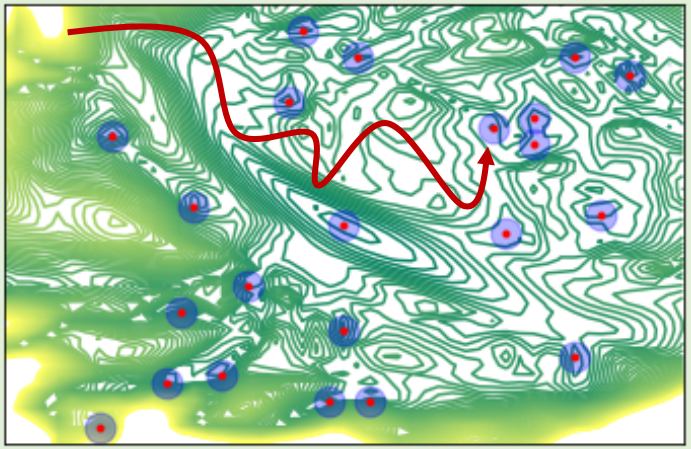
- 出発点：なぜ単純化した設定が必要か？
  →理論の使う道具が実現象を記述できていない

例：損失曲面の変動記述法開発



損失曲面の水平変動を記述
→現実に則した乖離度を評価

研究1：大域探索の暗黙的正則化



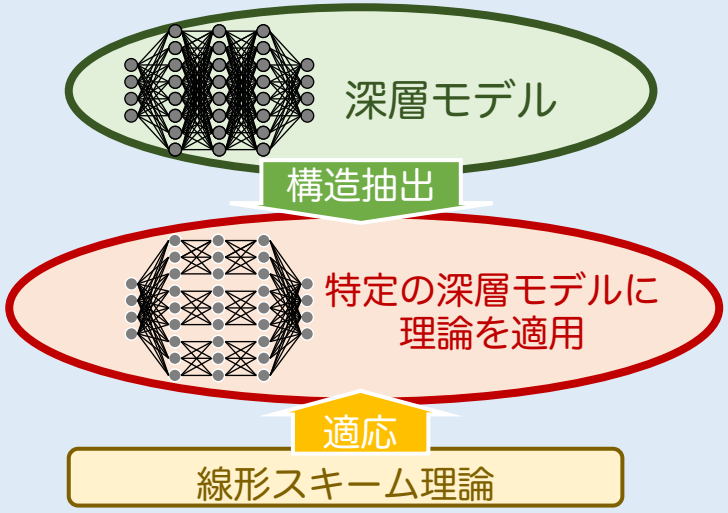
非凸損失上の大域探索が汎化誤差を抑制
技術：非ガウス探索法＋測度変換

今後の個人的な研究方針

線形モデル理論と深層モデルを架橋

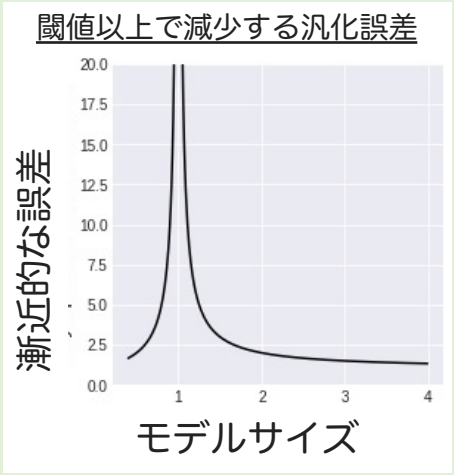
- 出発点：深層モデルを全て扱うのはおそらく無理
  →適応可能なネットワーク構造を特定する

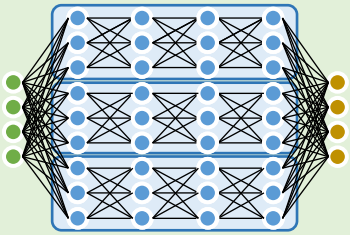
例：適応可能な構造の特定



損失曲面の変動を柔軟に記述
→現実に則した乖離度を評価

研究2：非線形・深層モデルの二重降下





誤差減少を起こす
ネットワーク構造

一定の深層モデルの二重降下を示す
技術：Taylor近似、FIM置換、拡張MP則

まとめ

背景

- 深層学習の”発見”と、それを記述する理論の不在

目的

- 深層学習という新しい技術をもとに
  深層構造に適応する新しい統計理論を作ること

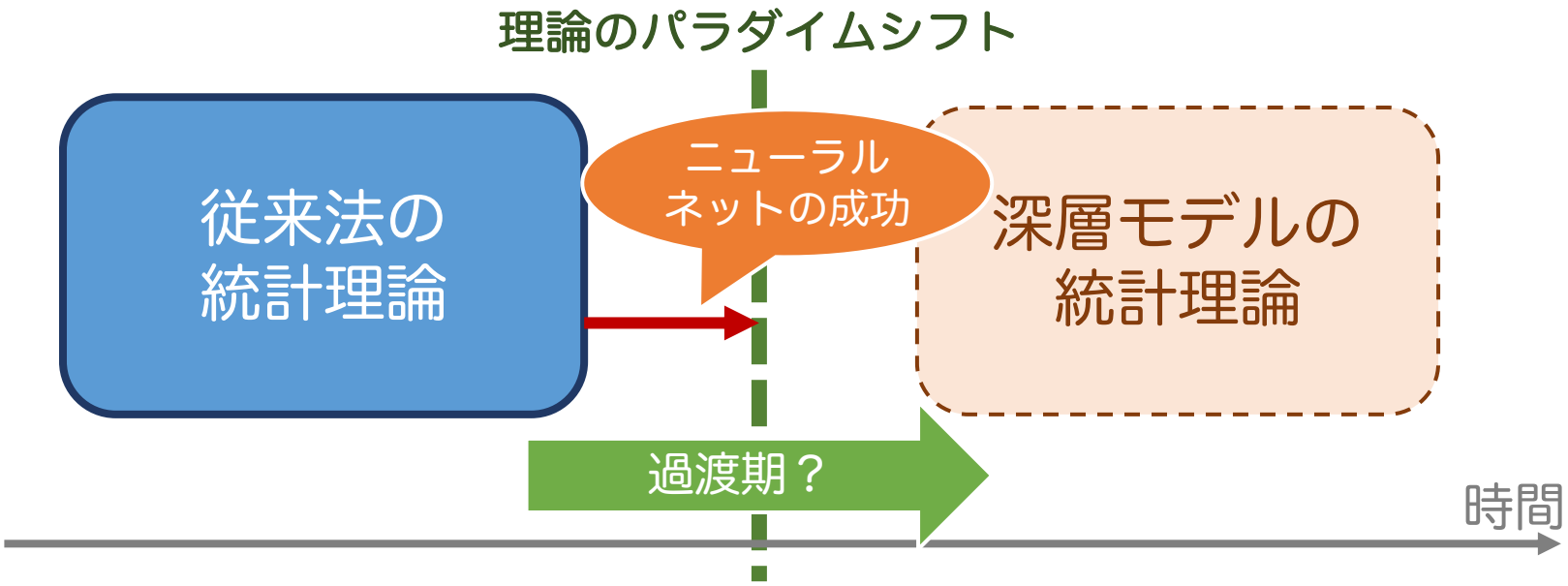
手段

- 近似誤差・複雑性誤差の解析
- 解析設定の拡張や、新しい統計的記述の開発

感想：統計分野に面白い未解決問題がある時代で嬉しい

新理論を創出できる？

- 今後到来する(?)深層統計理論の基盤創出
  - 資金石としての深層ニューラルネットワーク



成功するかは不明だが、楽しく研究できる分野

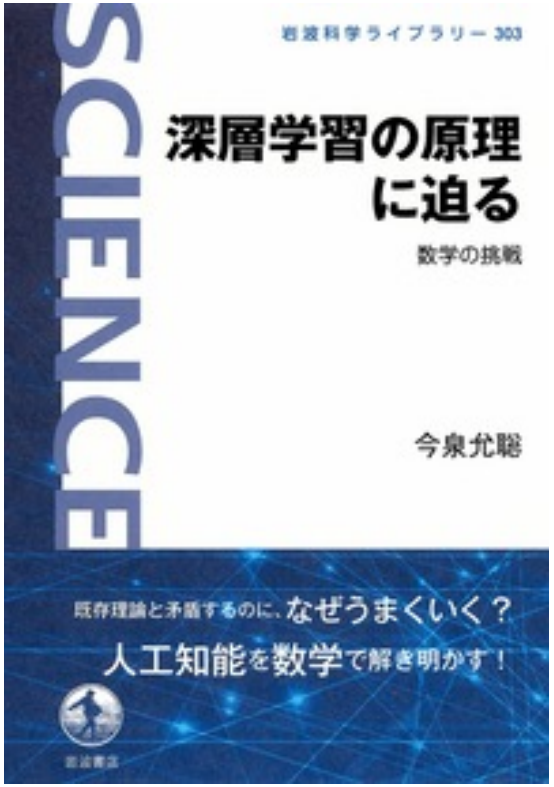
ご清聴ありがとうございました。

宣伝

「深層学習の原理に迫る 数学の挑戦」
岩波科学ライブラリー

- 2019年の統数研オープンハウス講演を書籍化
- 近年の深層学習理論の一部を紹介した新書
  - 一般向けの非技術書(縦書き)
  - さっくり読めます

技術的な書籍はまた後日...



2. 巨大モデルが機能する謎

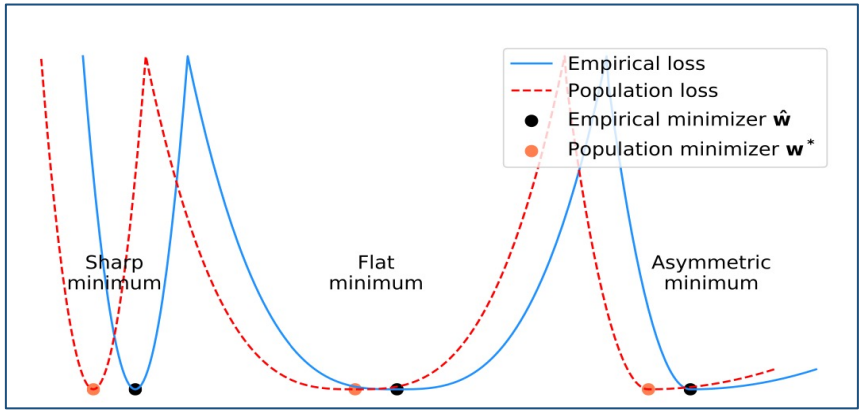
今後への未解決問題

過適合の理論の展望

大きな問：なぜ大規模モデルが過適合しない？

- 損失関数の形状説は実験的に有力視
  - Googleによる大規模実験で立証 (Jiang et al. (2019))

実験に基づく予想
損失関数には良い/悪い極小値が複数あり、
良い極小値が選ばれたら過適合は発生しない



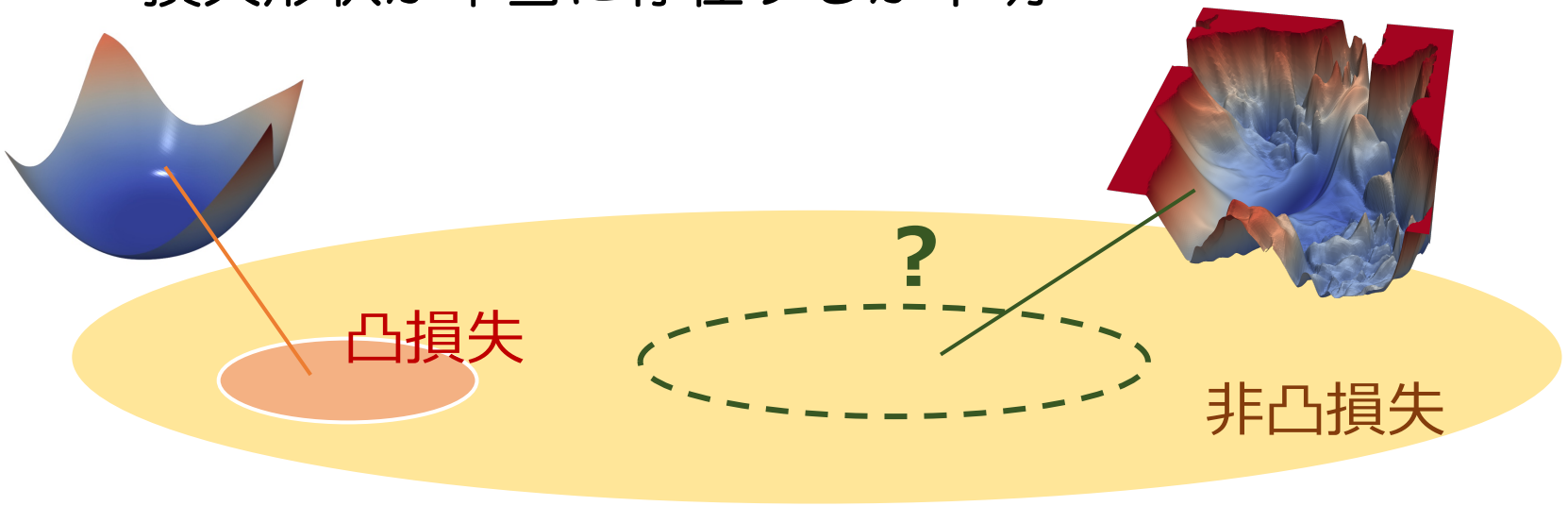
提唱されている
極小値の形状リスト
(He et al. (2019))

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., & Bengio, S. (2019). Fantastic generalization measures and where to find them. *ICLR*.
He, H., Huang, G., & Yuan, Y. (2019). Asymmetric valleys: Beyond sharp and flat local minima. Neurips,.

理論化の壁：非凸関数の不透明さ

問題：損失関数の形状には理論的な不明点が多い

- 例：極小値の数 / 近傍構造 / 稠密さ
- 損失形状が本当に存在するか不明



形状の性質が分からないため、過適合の理論も確立されない

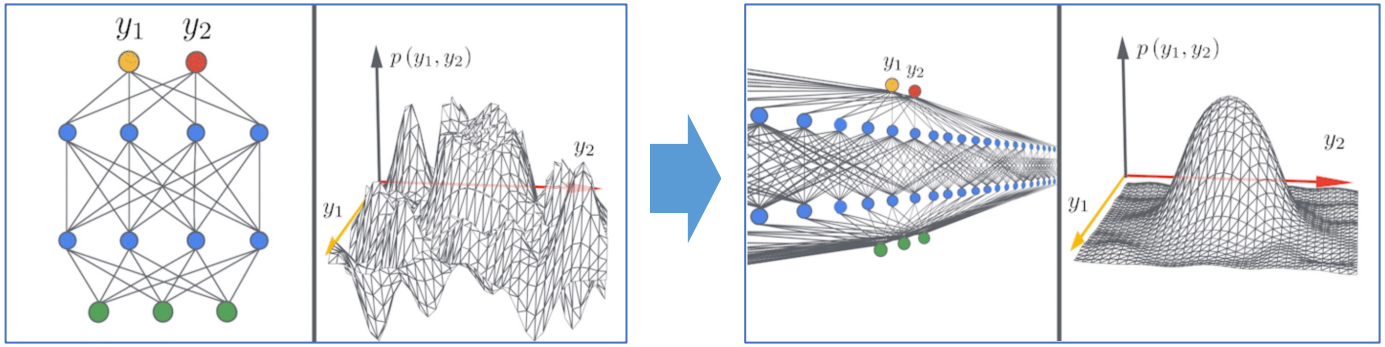
→ 非凸関数を分類・特徴付ける理論が必要

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. NIPS.

理論化の壁：非凸関数の不透明さ

現在の深層学習研究の主流

- 凸に近い状況になるように設計を工夫する



例： Neural Tangent Kernel: 層は少なく幅を増やすと損失は凸関数に近づく

- ただし、深層学習固有の良さは失われる

将来的に必要な基礎理論

- 非凸関数を理解し、それに基づく体系を構築

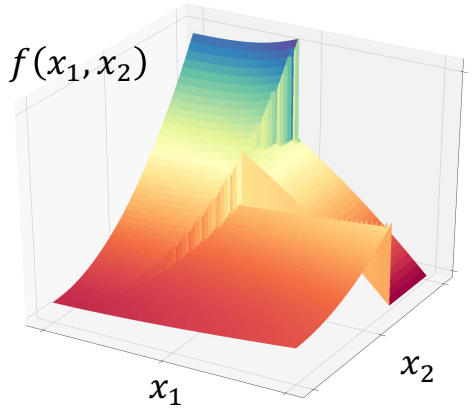
https://ai.googleblog.com/2020/03/fast-and-easy-infinitely-wide-networks.html

まとめ

本研究のまとめ

なぜ理論的理解が必要か？

- 今後の深層学習の”良い”発展に必要



- 1. 層を増やす役割

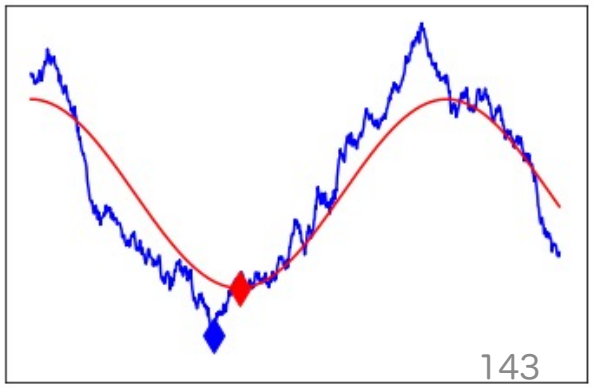
  非滑らかな関数を用いた理論

  - 非深層法との差別化の嚆矢

- 2. 巨大モデルの謎

  母極小値近傍への滞留理論

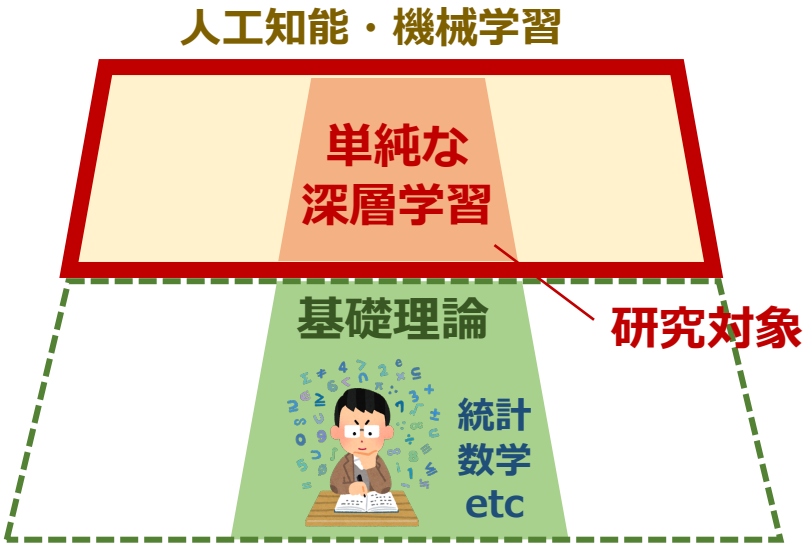
  - 非滞留の批判に耐える
    暗黙的正則化の理論



深層学習研究の方向

昨今の研究の方向

- 現象を単純化する方向
- 既存の基礎理論で説明できる状況をたくさん見つける
  - 例：２層ニューラルネット研究



長期的に必要な方向

- 基礎理論の拡張
- 必要なパラダイム
  - 非線形な関数表現論
  - 非凸関数の特徴付け

ご清聴ありがとうございました。

Note: Minimax最適レート

- Minimax誤差の下限のレート
  - 理論上最良の推定量による最悪の誤差

Minimax最適レート r_n

$\mathcal{F}$：ある関数族

$\bar{f}$：n個の観測に依存する推定量

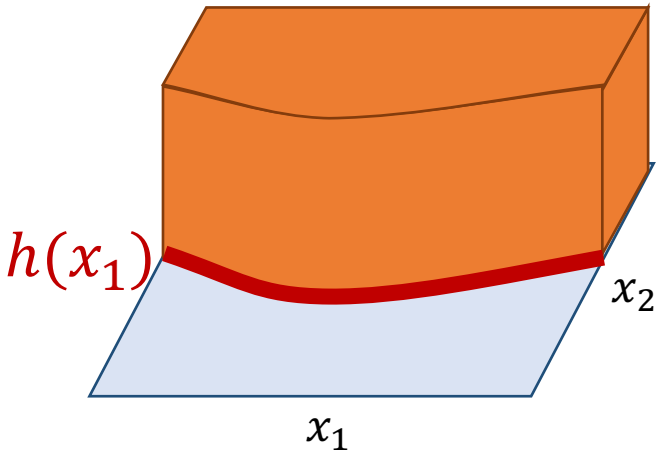
$$\inf_{\bar{f}} \sup_{f^* \in \mathcal{F}} E\left[\left\|f^* - \bar{f}\right\|_{L^2}^2\right] = \Omega(r_n)$$

- 例：$\mathcal{F} = C^\beta$ なら $r_n = n^{-2\beta/(2\beta+D)}$
- 取り扱いが良い最適性の基準

DNNによる不連続性解消

不連続な関数の例

- $\{(x_1, x_2): x_1 = h(x_2)\}$上で不連続 / 区分内部で定数



$$1_{\{x_2 \geq h(x_1)\}}(x_1, x_2)$$

$h(x_1)$: 滑らかな関数

滑らかな関数とステップ関数の合成に分解

- $\boldsymbol{s}(x) := 1_{\{x \geq 0\}}(x)$: ステップ関数
- $1_{\{x_2 \geq h(x_1)\}}(x_1, x_2) = \boldsymbol{s} \circ \underbrace{\left((x_1, x_2) \mapsto \left(x_2 - h(x_1)\right)\right)}_{\text{滑らかな関数}}$

DNNによる近似の方法

各要素の近似にDNNの構成を用いる

- DNN：各層の変換 $z_{\ell+1} = \eta(A_\ell z_\ell + b_\ell)$ の合成
  - ReLU：$\eta(x) = \max\{x, 0\}$ / Sigmoid $\eta(x) = (1 + \exp(-x))^{-1}$

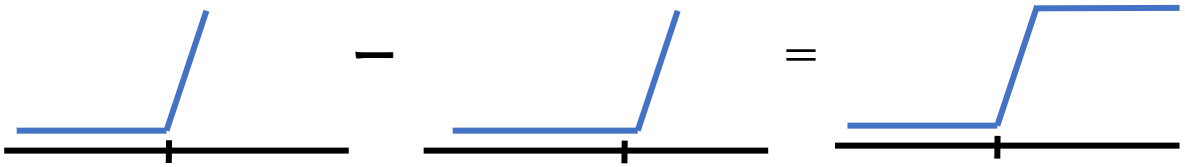
滑らかな関数 → DNNの数層で最適近似可

DNN二層：$\sum_j A_{2,j}\eta(A_1 x + b_1) + b_2 =: \sum_j A_{2,j}\phi_j(x)$

- ηが滑らかなら2層で微分可能関数を最適近似

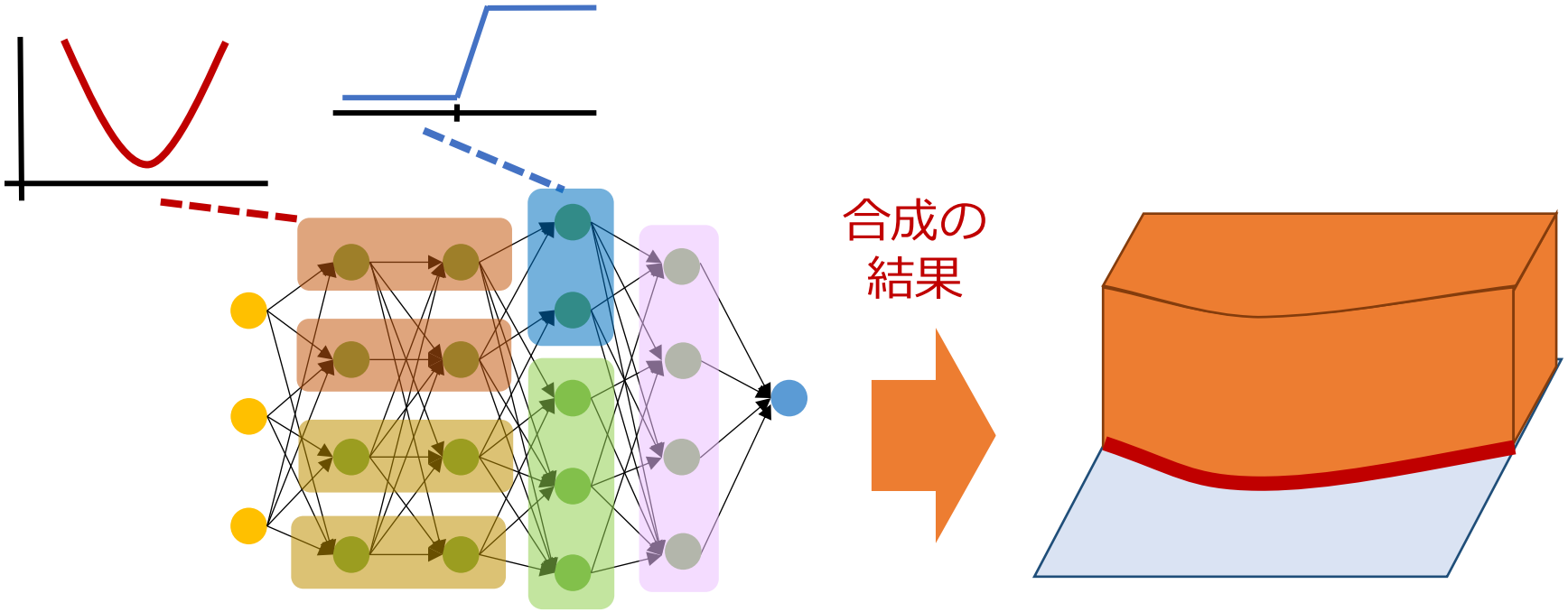
ステップ関数 → 非線型変換 η の差で近似可

- ReLUであれば引き算、Sigmoid関数であればそのまま



DNNによる近似の方法

- DNNは、非滑らかな関数の各パーツを個別に近似



ステップ関数等の近似誤差は無視可能

→滑らかな関数を近似するのと同等の近似誤差

他の方法だとどうなる？

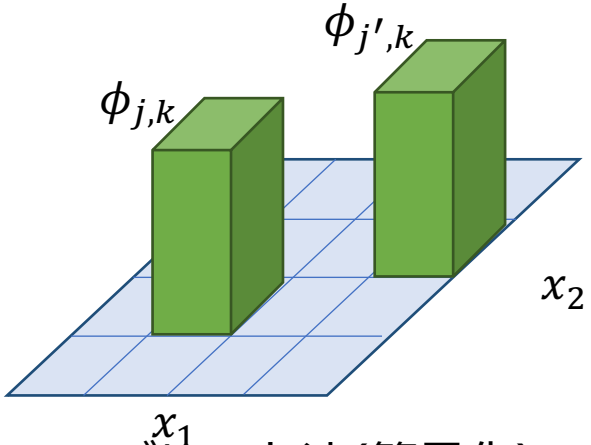
統計で一般的に用いられる方法：基底関数和

$$\hat{f}(x) = \sum_j w_j \phi_j(x)$$

- $\phi_j(x)$：基底関数、 w_j：パラメタ

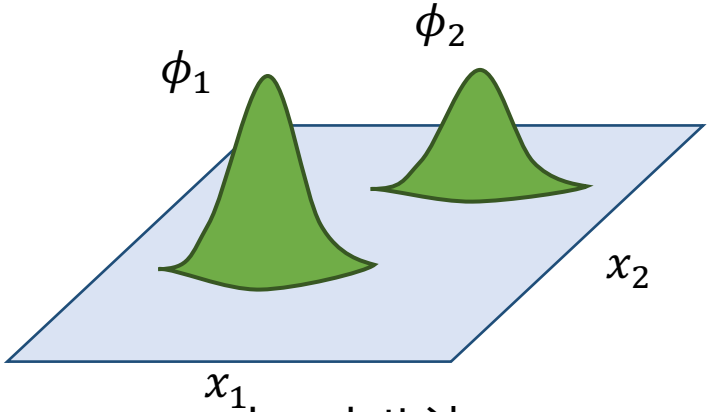
カーネル法
フーリエ法
スプライン法 etc…

- 基底関数の例：（局所的なϕ_jが一般的）



ウェーブレット法(簡易化)

例：$\phi_{j,k}(x) = 2^{-k/2}\Phi(2^{-k}x - j)$



カーネル法

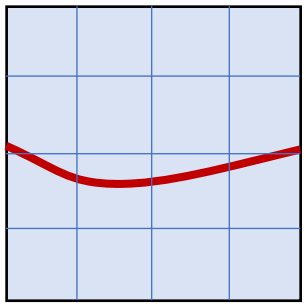
例：$\phi_j(x) = \exp(\|x - x_j\|^2/\sigma^2)$

基底を扱う問題点

基底関数で特異性を記述するのは容易でない

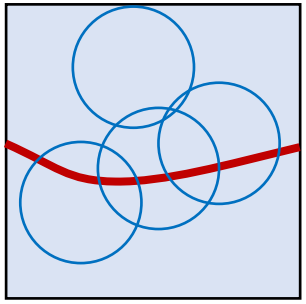
- 普通のϕ_jは、形が特異性にフィットしない

$x_2 = h(x_1)$



ウェーブレット法

$x_2 = h(x_1)$



カーネル法

各基底の
レベルセットの形

- $x_2 = h(x_1)$に重なる基底が多い → 誤差が増加

ϕ_jは$h(x_1)$の構造(滑らかさ)を抽出できない

→ 既存法の近似誤差が増加 → 最適性を喪失

新しい理論の潮流

汎化の原理：深層学習最大の謎（のひとつ）

- 現象を説明する新しい理論を紹介をします。

- 既存の理論

モデルの大きさ

過学習しやすさ
＝モデルの大きさ
（パラメタ数）

- 近年の理論の試み

いやいや違うんだよ
（諸説）

A: 暗黙正則化

B: PAC-Bayes

C: 二重降下