深層学習の汎化誤差のための近似性能と複雑性解析

2019/11/22 IBIS企画セッション

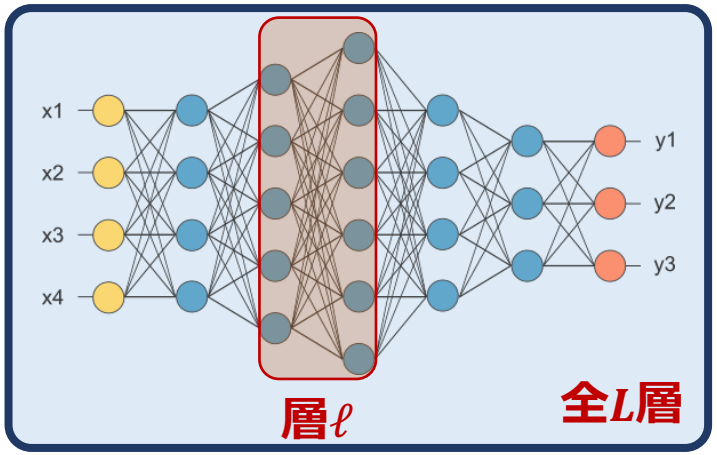
今泉允聡
（統計数理研究所 / 理化学研究所 / JSTさきがけ）

深層学習と問題設定

多層ニューラルネット（DNN）モデル

- 多くの層を持つ（変換を繰り返す）

入力 x



出力 y

ℓ層目の変換

$$h_\ell = \sigma(\Theta_\ell^\top h_{\ell-1})$$

Θ_ℓ: パラメタ (行列)

σ: 活性化関数

sigmoid, ReLUなど

データ上の経験誤差を最小化し、汎化誤差を評価

経験誤差(訓練誤差)

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

f_Θ: DNN, $\Theta = (\Theta_1, \ldots, \Theta_L)$

$\hat{\Theta}$を学習

汎化誤差(精度の尺度)

$$E\left[\mathcal{L}(\hat{\Theta})\right]$$

汎化誤差の分解

誤差を三つの要素で説明

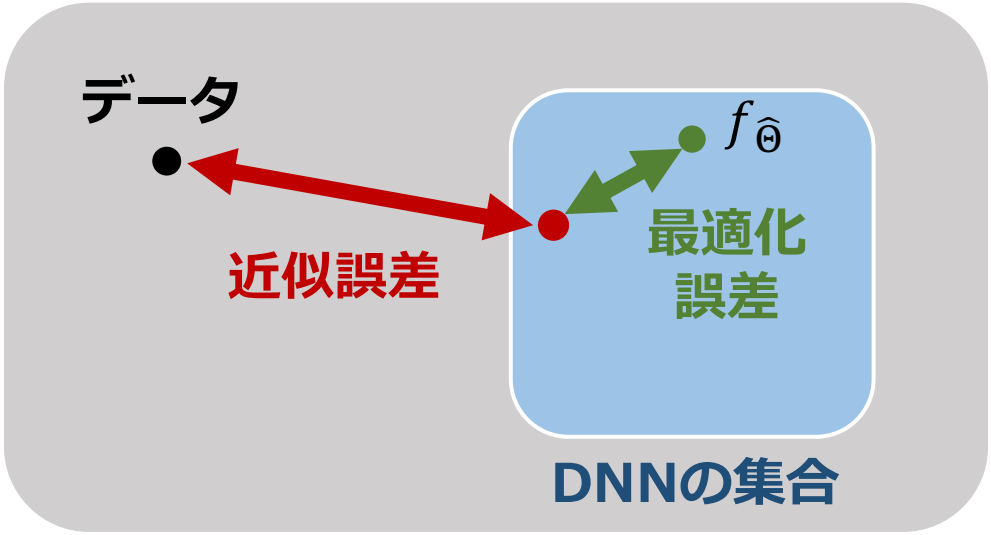
$$\mathrm{E}\left[\mathcal{L}\left(\hat{\Theta}\right)\right] \leq \inf_{\Theta'} \mathcal{L}(\Theta') + \left|E\left[\mathcal{L}\left(\hat{\Theta}\right)\right] - \mathcal{L}\left(\hat{\Theta}\right)\right| + \left|\mathcal{L}\left(\hat{\Theta}\right) - \inf_{\Theta'} \mathcal{L}(\Theta')\right|$$

汎化誤差

近似誤差
DNNの表現力

複雑性誤差
DNNの集合の大きさ

最適化誤差
学習がうまくいくか

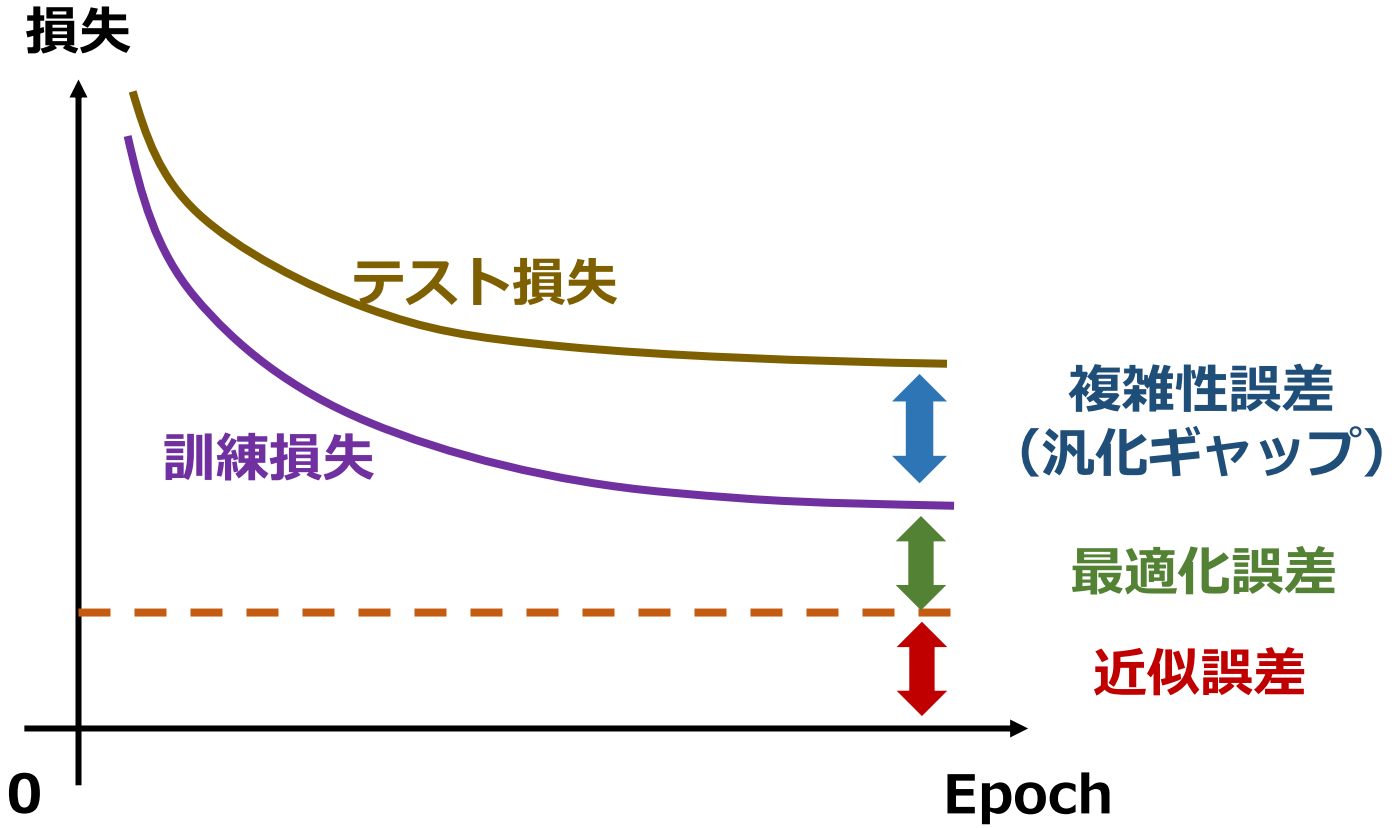


本講演のトピック

- **近似誤差**
- **複雑性誤差**

汎化誤差の分解

- 実際の学習の軌跡との対応



近似誤差

DNNと既存法の差を示す関数近似理論

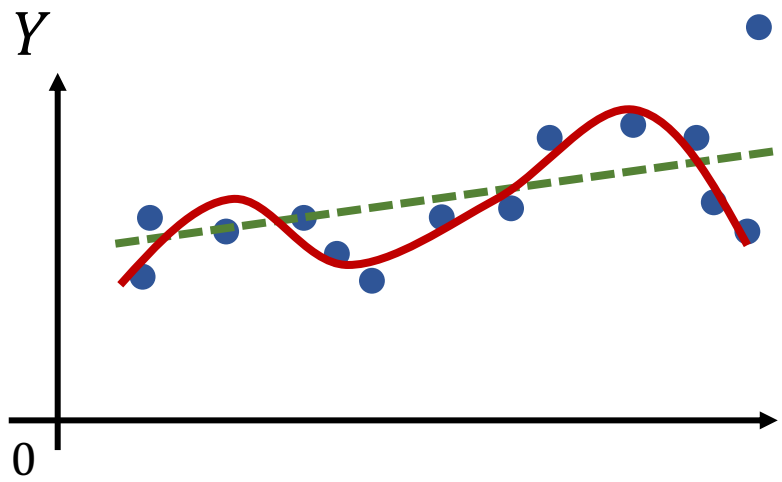
近似誤差とは

- DNN f_Θ が表現できるデータの構造で決まる

回帰の例（損失ℓは二乗損失）
データが $Y = f^*(X) + \varepsilon$ から生成されている時：

$$\inf_{\Theta'} \mathcal{L}(\Theta') \leq \inf_{\Theta} \|f^* - f_\Theta\|_\infty^2 + \text{Noise Terms}$$

Θ: パラメタ
f_Θ：DNN



● **データ**

データの構造を

表現できている

表現できてない

普遍近似定理

よく知られている結果

普遍近似定理 (Cybenko (1989)など)
層が２つのニューラルネットワーク(NN)は、十分な数のパラメタがあれば、連続関数を任意の精度で近似できる。

- 層２つで成立
  → 深層でなくて良い
- 普遍近似は多くの他手法でも成立
  → NNでなくて良い

➡ **なんでDNNを使うの？**

より詳細に近似を調べるには

近似誤差の減衰レート

- パラメタ(エッジ)が増えるときの誤差減少スピード

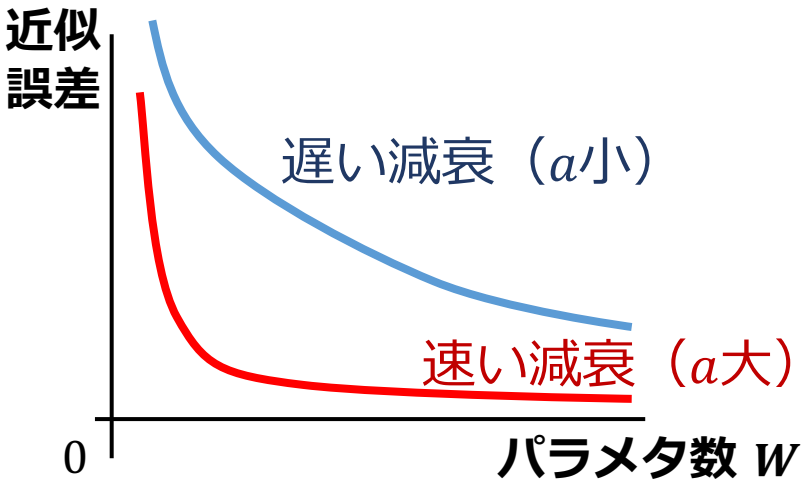
近似誤差の減衰レートa

f^*: 近似対象の関数

f_Θ: W個のパラメタを持つDNN

$$\inf_\Theta \|f^* - f_\Theta\| = O(W^{-a})$$

近似誤差　　パラメタ増で減少



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

滑らかな関数に対する近似レート

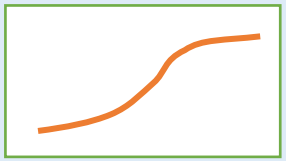
f^*：近似対象（入力d次元、β回微分可能）

f_Θ：DNN（L層, パラメタW個、活性化関数σ）

σがsigmoid等の場合 (Mhaskar (1996)など)

DNNは $L = 2$ のもとで以下を達成：

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



活性化関数がReLUの場合 (Yarotsky (2017)など)

L層のDNNは以下を達成：

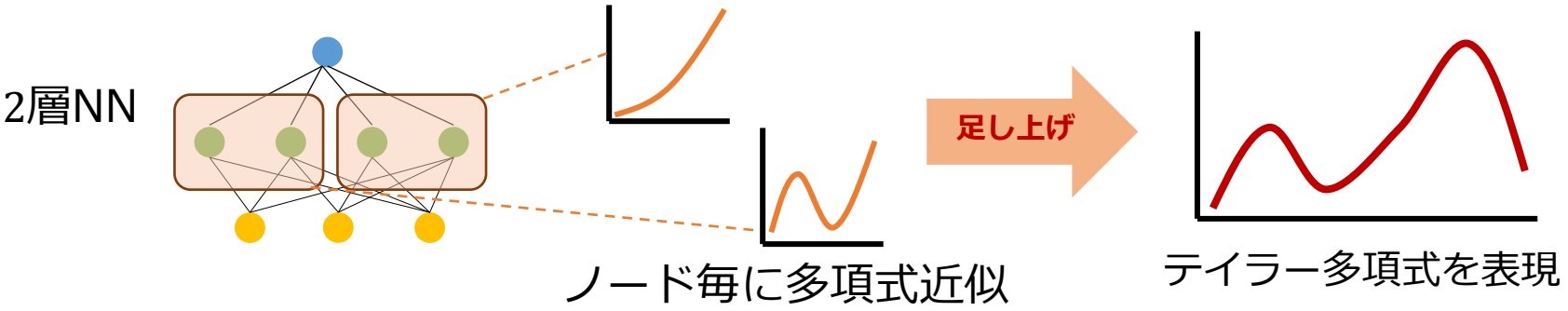
$$\inf_\Theta \|f^* - f_\Theta\| = O(W^{-\beta/d} + 2^{-L})$$

ReLUの尖りから来る影響

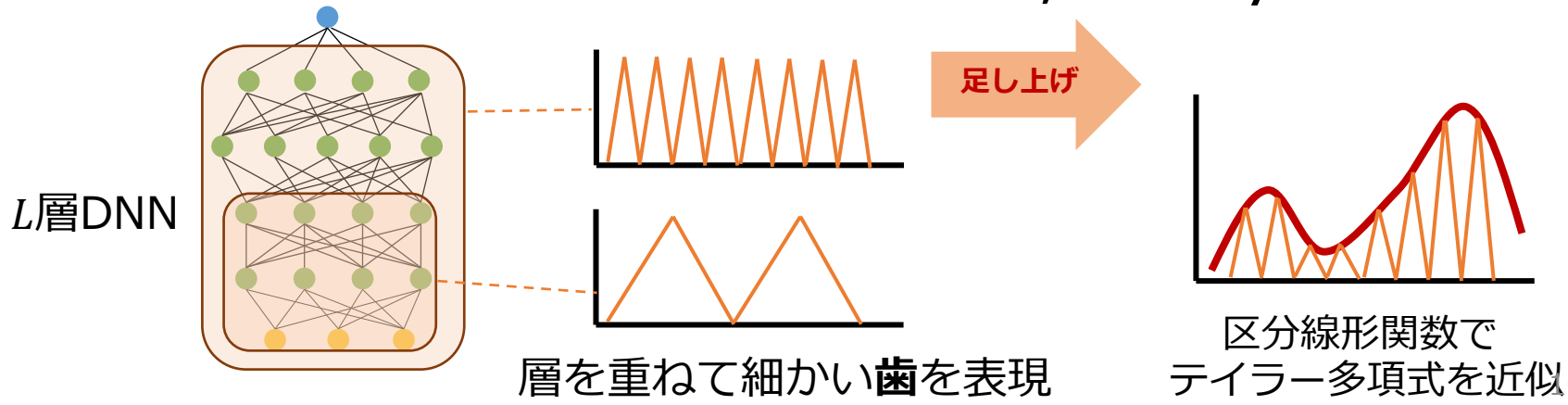
誤差レートβ/dは、 f^*の滑らかさで増加、入力次元で減少。

活性化関数σによる違い

- σが滑らかな場合（sigmoid, softplus）



- σが滑らかでない場合（ReLU, LeakyReLU）



良いレートなの？

- 誤差レートβ/dは理論的に最適

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

すごいぞ！
やはりDNNは
最適なんだ！

- しかし、他手法も同じように最適

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

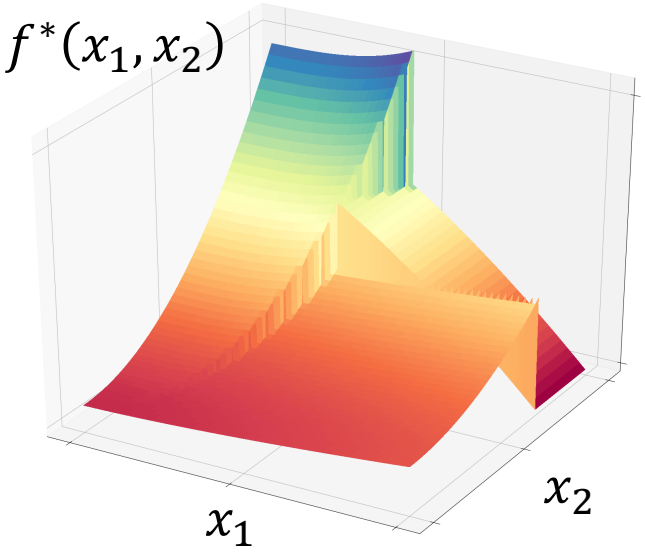
他のも最適だから
結局同じ？

f^*が滑らかなら、DNNと他手法の理論的性能は同等。

DNNが重要になる状況1

汎化性能も優越するが、
ここでは近似性能のみ記述

- f^*が滑らかでない場合、DNNが他に優越



区分上でのみ滑らかな関数

$$f^* = \Sigma_m f_m \otimes 1_{R_m}$$

f_m：滑らかな関数, 1_{R_m}：区分上の指示関数

近似レートの差別化

(Imaizumi & Fukumizu (2019))

DNNのレート：

$O(\max\{W^{-\beta/d}, W^{-\alpha/2(d-1)}\})$

他手法(カーネル等)のレート：

$O(\max\{W^{-\beta/d}, W^{-\alpha/4(d-1)}\})$

（αは区分の境界線の滑らかさ）

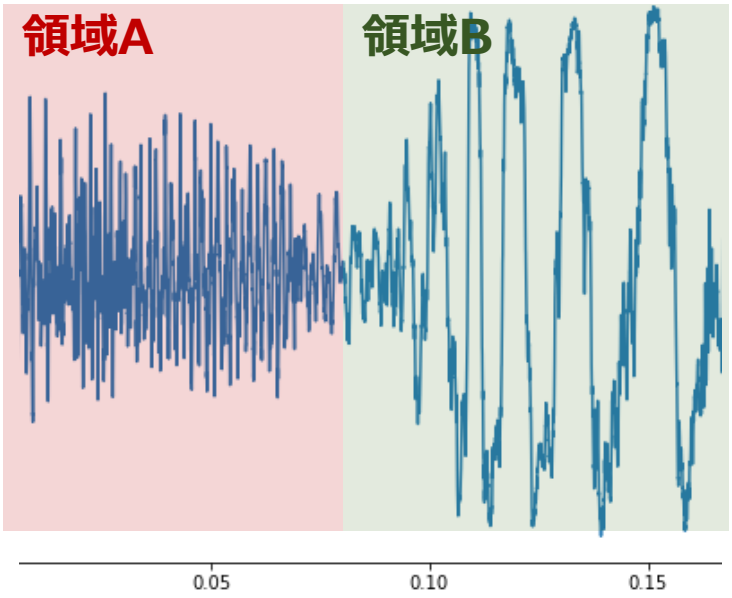
区分の境界が複雑な形

➡ DNNが速いレートを達成

DNNが重要になる状況2

汎化性能も優越するが、
ここでは近似性能のみ記述

- f^*が不均一な滑らかさを持つ場合、DNNが優越



近似レートの差別化

(Suzuki (2019))

DNNのレート：

$$O(W^{-\beta/d})$$

他手法(カーネル等)のレート：

$$O(W^{-(\beta-(1/p-1/2)_+)/d})$$

(pは不均一さの程度)

Besov空間の関数

$$f^* = \Sigma_j c_j \phi_j + \Sigma_{j,k} c_{j,k} \psi_{j,k}$$

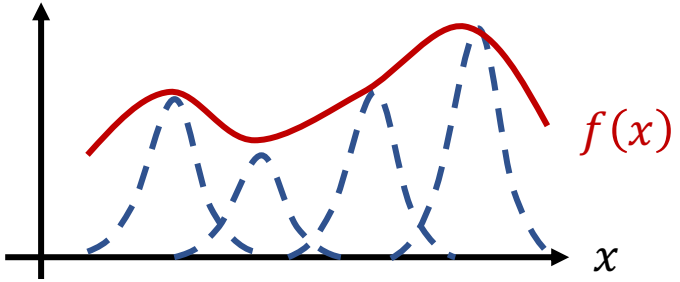
$$\|c_\cdot\|_p + \left(\Sigma_k 2^{qk(\beta+1/2-1/p)} \|c_{\cdot,k}\|_p^q\right)^{1/q} < \infty$$

不均一さがより強い

➡ DNNが速いレートを達成

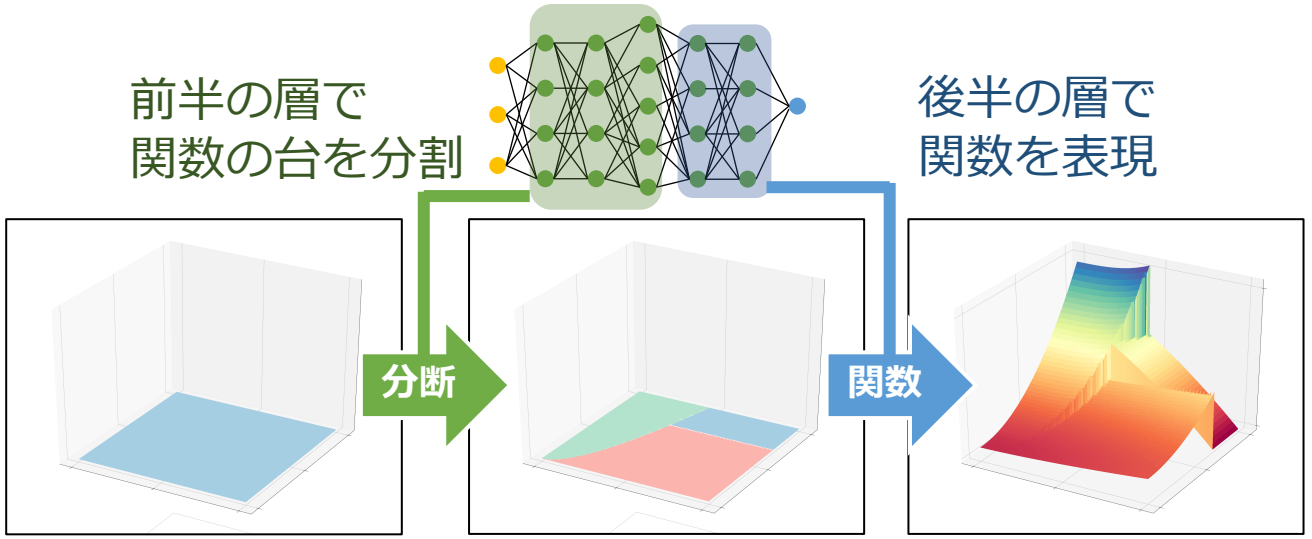
DNNは局所構造を表現できる

- **これまで**：均一な滑らかさ・構造を表現



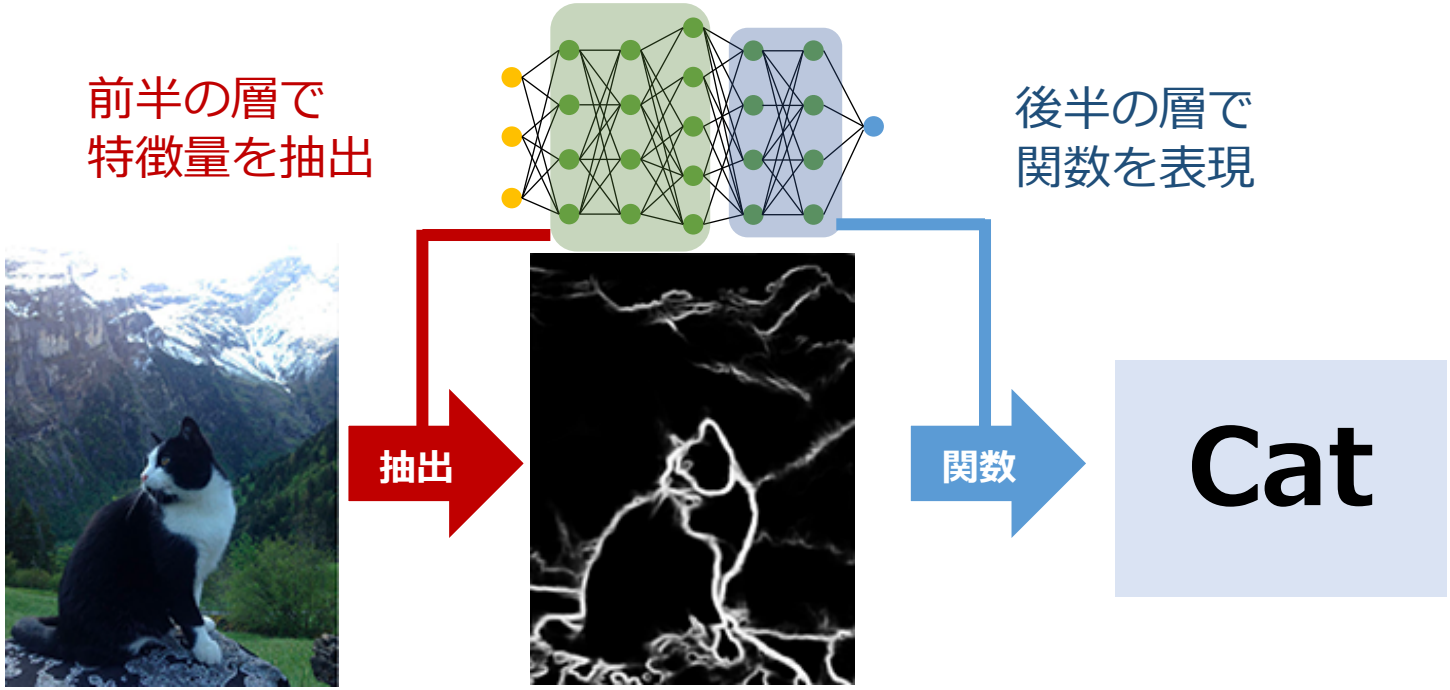
カーネル法
スプライン法
フーリエ法　など

- **DNN**：局所的に滑らかさを変えられる



更なるDNNの役割：特徴量抽出

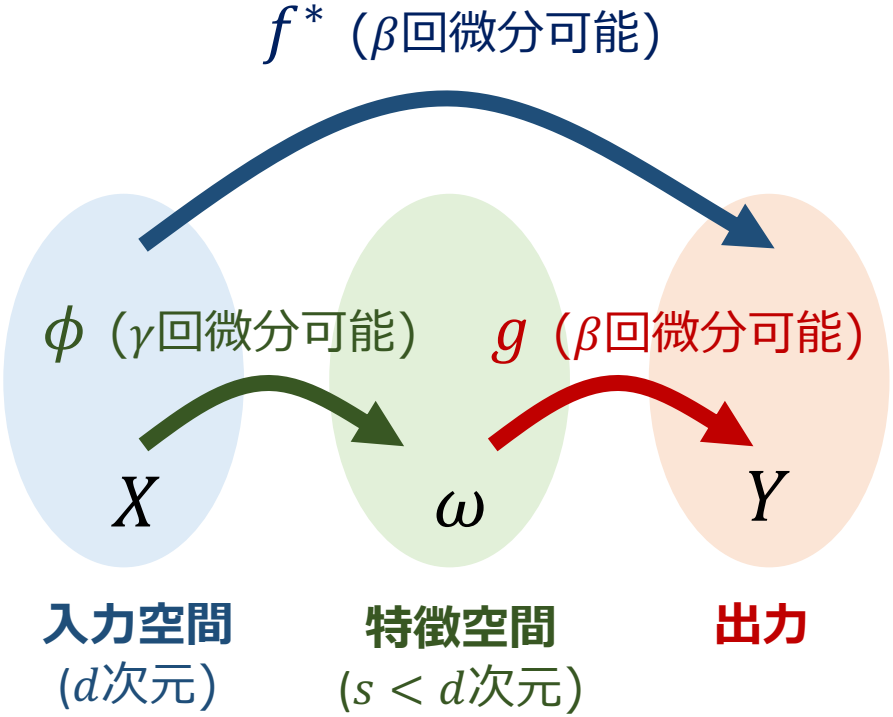
特徴量変換＋関数表現を考える



DNNは複数回の四則演算をする
➡ 既存の変換（フーリエ変換など）が近似的に可能

DNNが重要になる状況3

- 特徴空間への写像がある場合、DNNのレート改善



近似レート

(Peterson et al. (2019))
(Schmidt-Hieber (2019a))

DNNのレート：

$$O(\max\{W^{-\gamma/d}, W^{-\beta/s}\})$$

一般的なレート：

$$O(W^{-\beta/d})$$

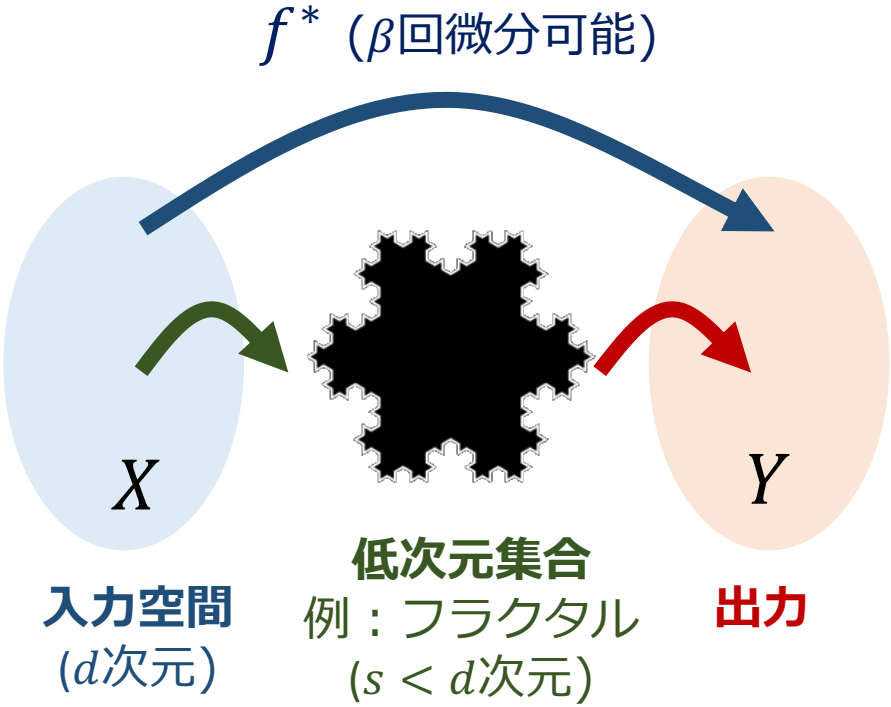
特徴量が低次元・ϕがシンプル
➡ DNNのレート改善

特徴写像ϕ(未知)と関数gの合成

$$f^* = g \circ \phi$$

DNNが重要になる状況3+

- 特徴空間の具体例：低次元集合



低次元特徴量がある時の近似レート

(Nakada & Imaizumi (2019))

DNNのレート：

$$\tilde{O}(W^{-\beta/s})$$

一般的なレート：

$$O(W^{-\beta/d})$$

（s: 低次元集合の次元）

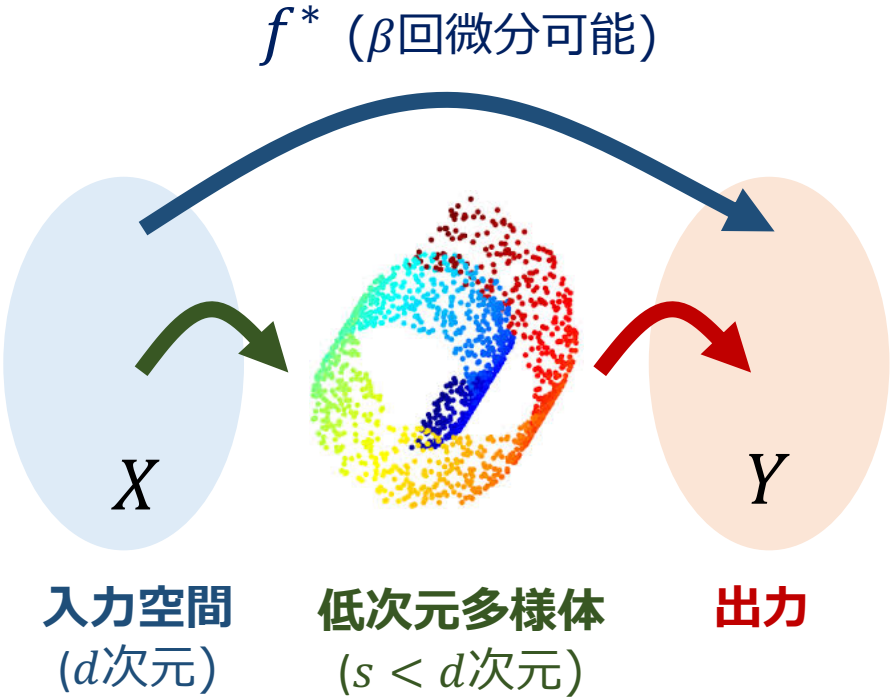
低次元集合Ω上の関数

$$f^*(X), \mathrm{Supp}(X) = \Omega$$

➡ DNNのレート改善

DNNが重要になる状況3+

- 特徴空間の具体例：低次元多様体



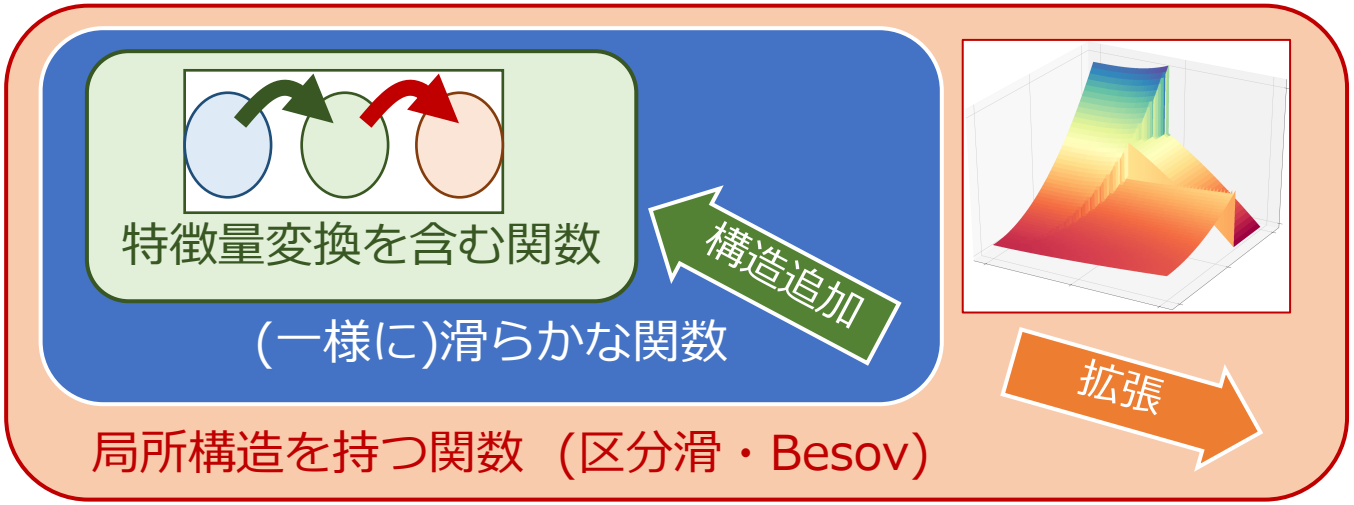
多様体特徴量がある時の近似レート
(Schmidt-Hieber (2019b))
DNNのレート：
$$\tilde{O}(W^{-\beta/s})$$
一般的なレート：
$$O(W^{-\beta/d})$$

多様体上への変換を持つ関数
$$f^* = g \circ \phi$$

➡ DNNのレート改善

近似誤差のまとめ

- DNNが優位・改善する状況の発見



- **知見1：局所構造を持つ関数ではDNNが優位**
  - 既存法(カーネル法など)に対する優位性も示される
- **知見2：特徴量抽出が有効ならDNNは改善**
  - 狭い関数クラスだが、特徴量の構造がDNNに合う

近似誤差の未解決点

- まだ未解決な点も多い

結局、層は
すごく多いのが
いいの？

超深層(100層など)の意義は不明

- 上記の結果は3~5層で成立
- ReLUの場合も$O(\log n)$層くらい

議論：Yarotsky(2018)など

どんな特徴量が
必要なの？

使える特徴量概念は抽象的

- 解析できる特徴は次元などに限定
- 具体的な特徴付けは今後の課題

議論：Alemi+(2016)など

ラフな差別化は出来つつあるが、現象を説明するにはまだ壁

複雑性誤差

なぜ巨大モデルでも汎化する？

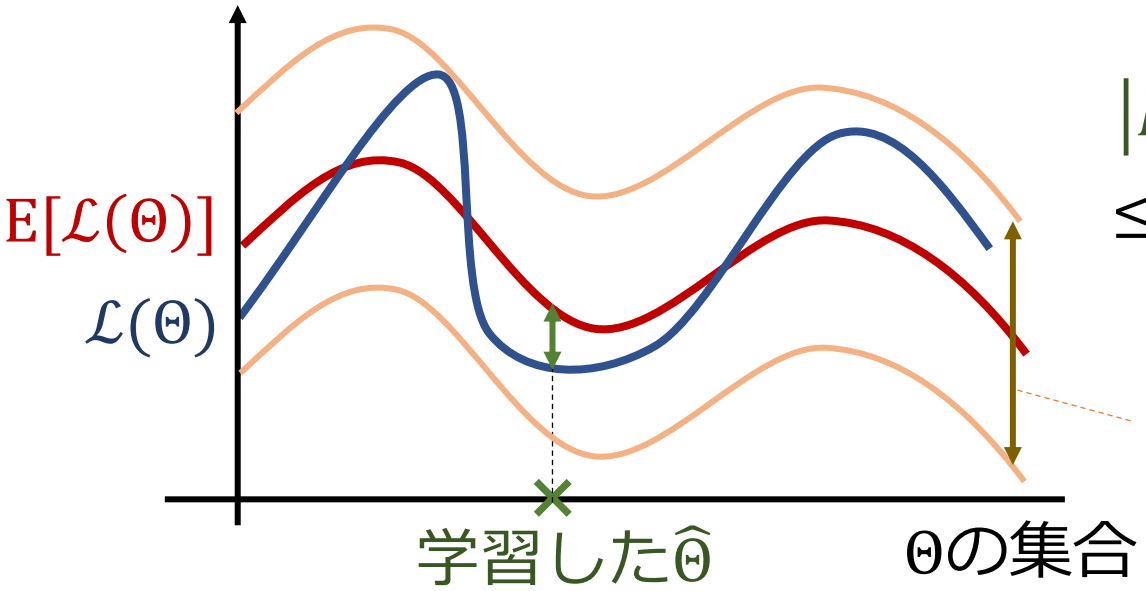
そもそも複雑性誤差とは？

- 汎化誤差(期待値)と訓練誤差(経験平均)との差

$$\left|E\left[\mathcal{L}\left(\hat{\Theta}\right)\right]-\mathcal{L}\left(\hat{\Theta}\right)\right|$$

評価方法：一様収束誤差

- **可能な全てのΘ**上での$\mathrm{E}[\mathcal{L}(\Theta)]$と$\mathcal{L}(\Theta)$の差



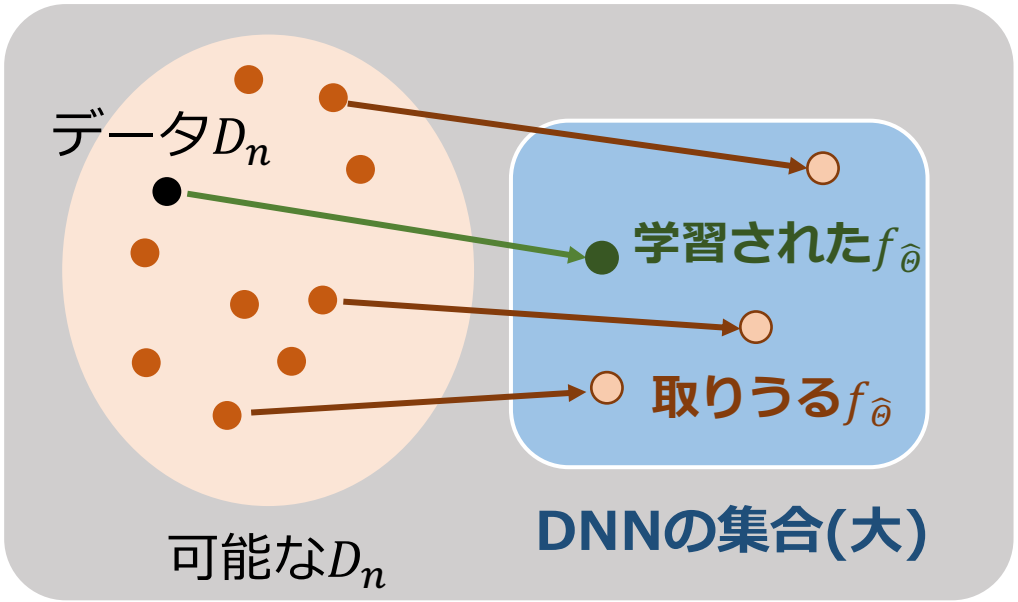
$$\left|E\left[\mathcal{L}\left(\hat{\Theta}\right)\right]-\mathcal{L}\left(\hat{\Theta}\right)\right| \leq \sup_{\Theta'}\left|E\left[\mathcal{L}(\Theta')\right]-\mathcal{L}(\Theta')\right|$$

複雑性誤差の評価

なぜ全てのΘを考える？

アルゴリズムの
ランダムを考慮することもある

- 汎化誤差$\mathrm{E}[\mathcal{L}(\hat{\Theta})]$(期待値)を考えるとは？
  - 可能なD_nすべての場合の平均値を考えること



汎化誤差(期待値)を考える

↓

取りうるD_nをすべて考慮

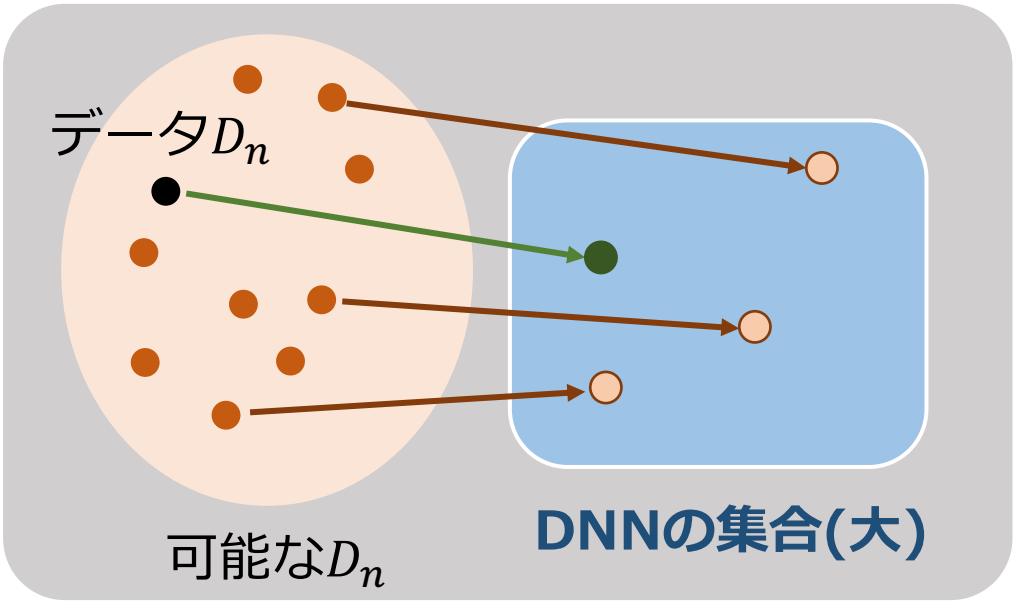
↓

可能なD_nから定まる$f_{\hat{\theta}}$を
すべて考慮

既存理論の考え方

既存理論

- 複雑性誤差＝可能なf_Θの集合  の大きさ



可能なD_nから定まる$f_{\hat{\Theta}}$を
すべて考慮

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

複雑性評価の数学的方法

レートは改善可だが
大きさへの依存は不可

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

n: データ数
W: パラメタ数

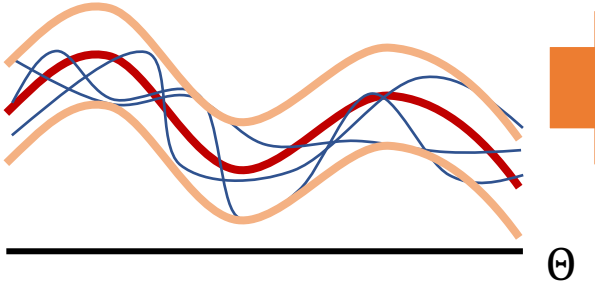
$$\sup_{\Theta}|\mathrm{E}[\mathcal{L}(\Theta)] - \mathcal{L}(\Theta)| = O\left(\frac{1}{\sqrt{n}}\int_0^{\infty}\sqrt{\log N_\delta}\, d\delta\right)$$

可能なf_Θの集合の**大きさ**

導出の流れ

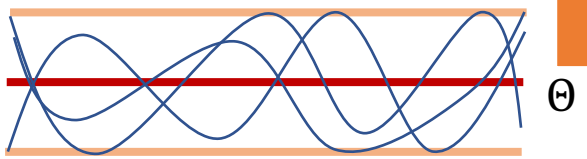
一様誤差

$$\sup_{\Theta}|\mathrm{E}[\mathcal{L}(\Theta) - \mathcal{L}(\Theta)]$$



Rademacher複雑性

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

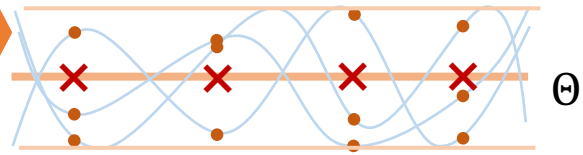


σ_i: Rademacher変数

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）

複雑性はパラメタ数が主

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

$$O\left(\frac{1}{\sqrt{n}}\int_0^{\infty}\sqrt{\log N_\delta}\,d\delta\right) = O\left(\frac{\sqrt{W\log L}}{\sqrt{n}}\right)$$

➡ パラメタ数 W が主な要素

- この理論はDNNの実性能を説明できない

大量のパラメタは
複雑性誤差を上げる

矛盾

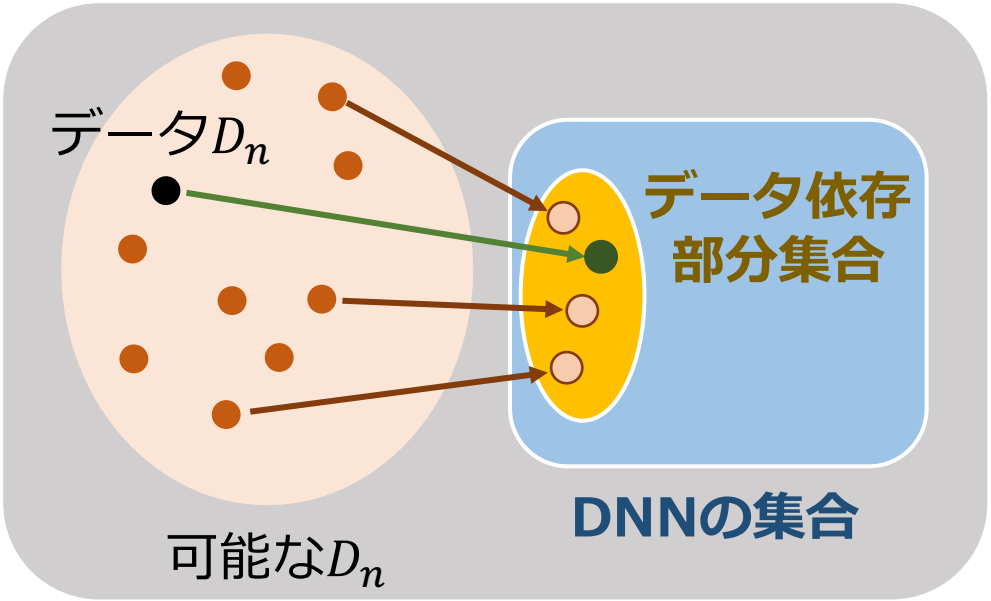
高精度DNNは
膨大なパラメタ数
Alex Net → 6千万
VGG Net → 1億

統計・学習理論の(大)原則

新しい複雑性評価の着想

着想：可能な f_Θ すべてを考える必要は無いのでは？

- データ依存で実現しうるf_Θの部分集合がありそう



これまでの複雑性誤差

可能な全DNN の大きさ

→巨大な複雑性

新しい着想

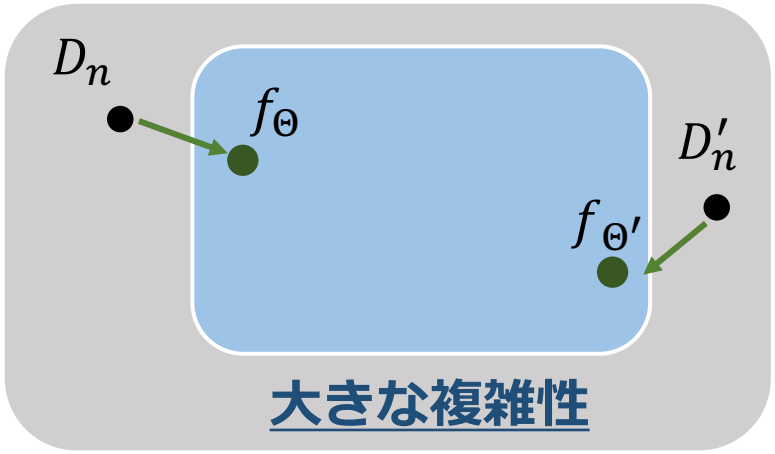
データ依存集合 の大きさ

→小さい複雑性

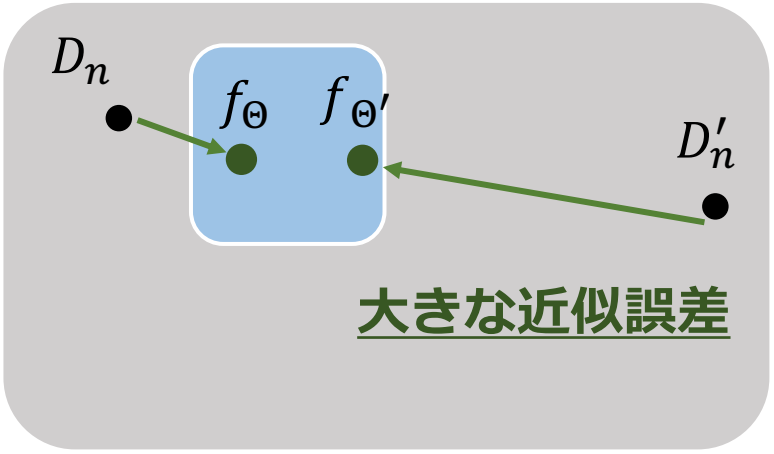
新しい着想を支持する実験

- データ依存集合の重要性を示す**実験**(Zhang+ (2017))
  - 全然違うデータD_n, D_n'でも、
    DNNの近似誤差・複雑性誤差は両方とも**小さい**

既存理論：複雑性誤差＝ の大きさ



or



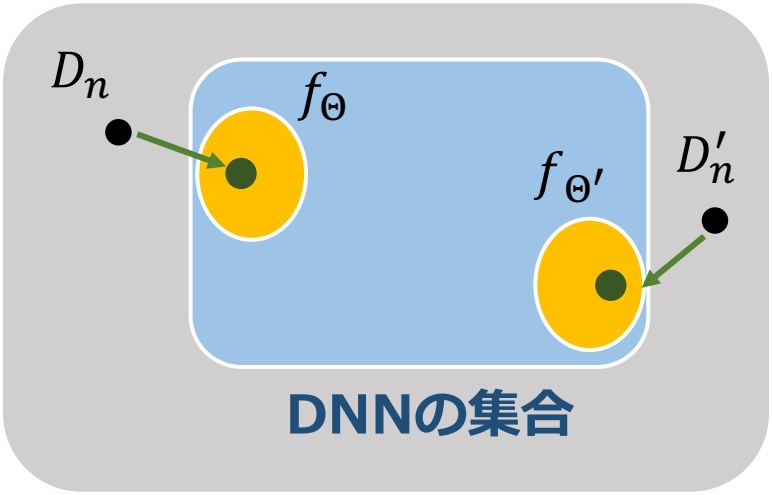
近似誤差：どちらも**小**
複雑性誤差：**大**

近似誤差：どちらかは**大**
複雑性誤差：**小**

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**(Zhang+ (2017))
  - 全然違うデータD_n, D_n'でも、
    DNNの近似・複雑性誤差は両方とも小さい

新しい着想：**複雑性誤差**＝データ依存集合  の大きさ



近似誤差：どちらも**小**
複雑性誤差：常に**小**

実験・実現象と一致！

データ依存集合を考える

- データ依存集合の重要性を示す実験(Zhang+ (2017))
  - 全然違うデータD_n, D_n'でも、
    DNNの近似・複雑性誤差は両方とも小さい

新しい着想：複雑性誤差＝データ依存集合 ● の大きさ

何によって ● は

決まっているの？

試みの一部

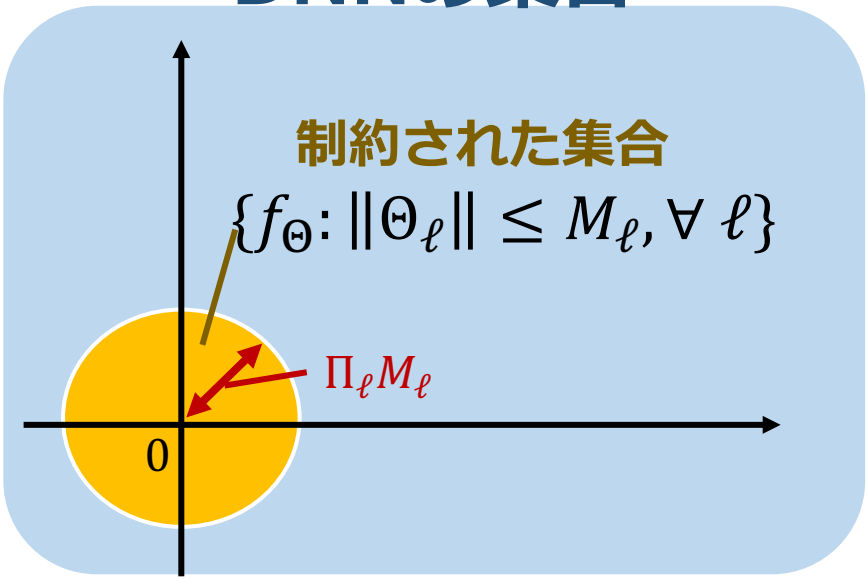
1. ノルム制約(暗黙的正則化)
2. 学習アルゴリズム
3. 各種正則化手法

他多数

試み1. ノルム制約下の複雑性

- パラメタΘが十分に小さい場合
  - 各層$\ell = 1, \ldots, L$の$\|\Theta_\ell\|$がM_ℓで抑えられると仮定

DNNの集合



ノルム制約下での複雑性誤差

(Neyshabur+ (2015a))
(Golowich+ (2019))

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

$B = \max_i \|x_i\|$: データの大きさ

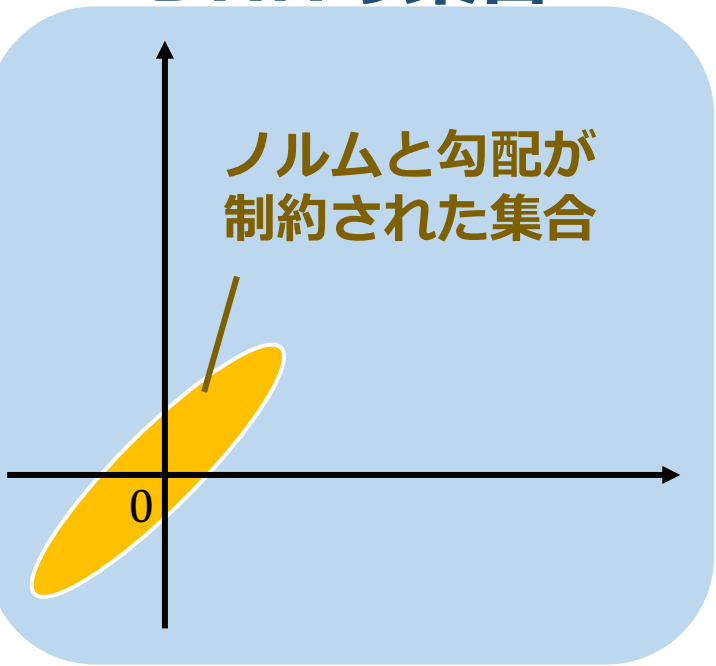
他、Bertlett+ (2017)など

- パラメタ数Wには（陽には）依存しない。
- M_ℓが小さいほど、複雑性誤差は小さくなる。

試み1. ノルム制約下の複雑性

- 他ノルムや、各層の勾配を抑えることも

DNNの集合



ノルム・ヤコビアン制約下の誤差

(Wei & Ma (2019))

$$\tilde{O}\left(\frac{\left(\Sigma_{\ell=1}^{L}(M_\ell' K_\ell)^{2/3} + (M_\ell'' K_\ell')^{2/3}\right)^{3/2}}{\sqrt{n}}\right)$$

M_ℓ', M_ℓ'': Θ_ℓの別ノルムでの上限
K_ℓ, K_ℓ': ℓ層のヤコビアンの上限

試み1. ノルム制約下の複雑性

- ノルム制約の実現性に課題

パラメタのノルムが常に小さければ
複雑性誤差も小さい！

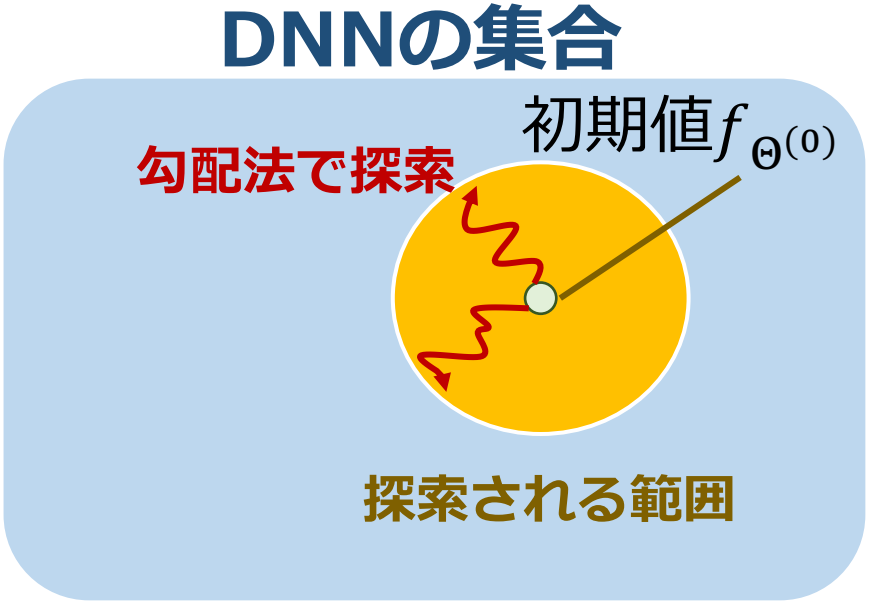
じゃあどういう時にノルムが小さいの？
それはDNN特有なの？

➡ まだ明確な特徴づけは明らかでない

数学的発見による仮説はある
(線形モデルの収束など: Neyshabur+ (2014), Arora+(2019))

試み2. 学習アルゴリズムの影響

- **学習アルゴリズム（勾配法）**の性能を評価
  - 探索される範囲＝実現するf_Θの集合



確率的勾配法(SGD)にも
拡張可能

勾配法による最適化

初期値を設定 $\Theta^{(0)}$

パラメタ更新 $t = 1, \ldots, T$

$$\Theta^{(t+1)} \leftarrow \Theta^{(t)} - \eta_t \nabla_\Theta \mathcal{L}\left(\Theta^{(t)}\right)$$

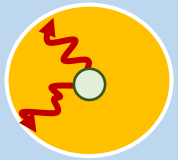
η_t: ステップサイズ

試み2. 学習アルゴリズムの影響

- 探索性能を下げる → 実現するf_Θも減る

DNNの集合

ステップサイズ・更新回数を減



探索範囲は縮小

早く学習をやめれば誤差減

(Hardt+ (2016))
(Kuzborskij+ (2017))

$\eta_t = 0(1/t)$とする

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

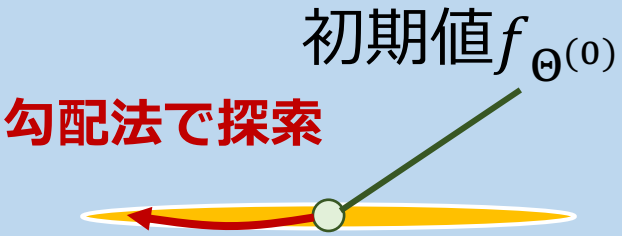
$T \geq 1$: パラメタの更新回数
$q \in (0,1)$: 減衰率

- 探索範囲を狭めれば、複雑性誤差は小さくなる。
- パラメタ数Wの影響は受ける。

試み2. 学習アルゴリズムの影響

- 実際の探索の低次元性を考慮

DNNの集合



低次元集合上の探索で誤差減

(Imaizumi & Sriperumbudur (2019))

$\eta_t = O(1/t)$とする

$$\tilde{O}\left(\frac{d_W L d \log T^{q'}}{n}\right)$$

$d_W < W$: 探索される空間の次元

$q' \in (0,1)$: 減衰率

- 低次元構造で、パラメタ数Wの影響を緩和。
- ただし活性化関数に制約が入る。

試み2. 学習アルゴリズムの影響

- ノルム制約の実現性に課題

学習を控えめにすれば
複雑性誤差は小さい！

それは良いパラメタが
学習できないんじゃないの？

➡ 近似・最適化誤差との関係は明らかでない。

損失関数の形状の研究が必要。
実験的反論もある: Hoffer+ (2018)

複雑性誤差の理論まとめ

問題：DNNは膨大なパラメタを持つが、複雑性誤差が小さい

理論化の方向：複雑性誤差＝データ依存の**モデル部分集合**

- その部分集合の正体は明らかでない

正体の候補

・ノルム制約集合
・学習アルゴリズムの影響
・各種正則化手法（バッチ正規化など）, 他多数

多くの仮説・実験・数学的発見

・暗黙的正則化 / バイアス / 損失関数の形状, etc…

まとめと展望

まとめ・展望

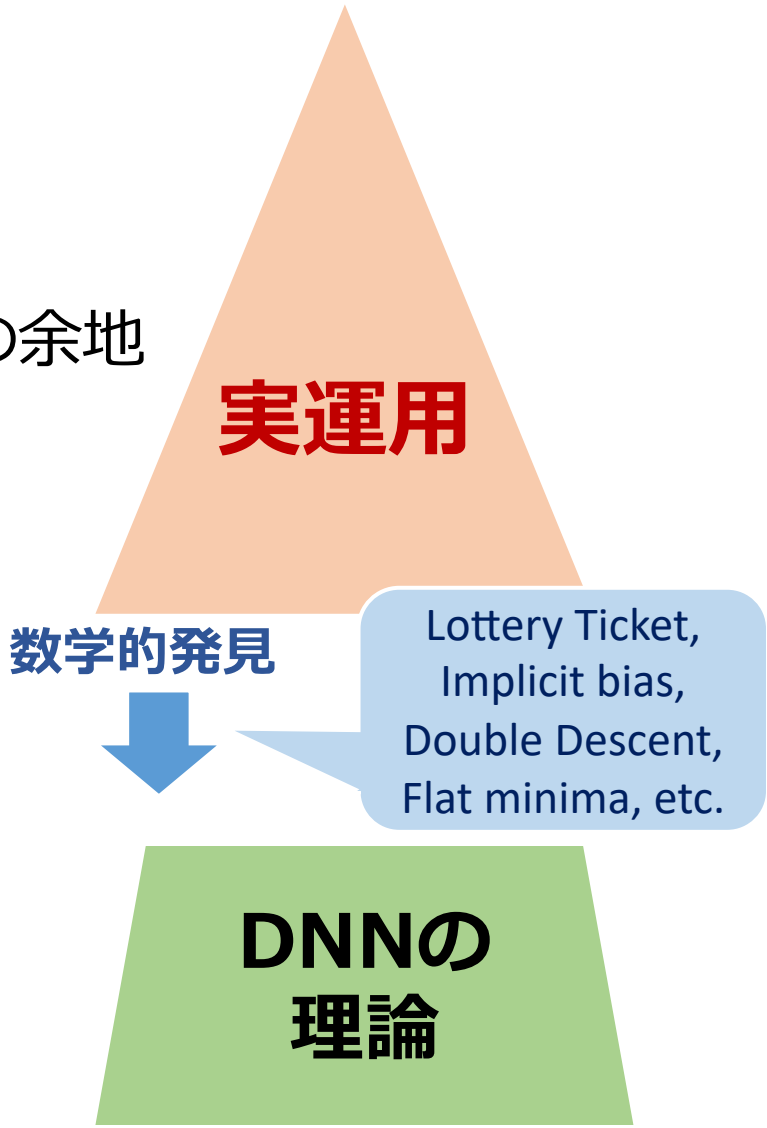
近似誤差

- 進捗しているが更なる発展の余地

複雑性誤差

- 実現象との大きな矛盾
- モデル部分集合のスキーム
  - 未解明な点は多い
  - 他のスキームも提案

実験・数学的発見から
理論の拡張へ



ご静聴ありがとうございました。

Reference

- Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems, 2, 183-192.
- Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. Neural computation, 8(1), 164-177.
- Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. Neural Networks, 94, 103-114.
- DeVore, R. A., Howard, R., & Micchelli, C. (1989). Optimal nonlinear approximation. Manuscripta mathematica, 63(4), 469-478.
- Newman, D. J. and Shapiro, H. S., (1964) Jackson's theorem in higher dimensions, On Approximation Theory (Proceedings of Conference in Oberwolfach, 1963), pp. 208–219.
- Imaizumi, M., & Fukumizu, K. (2018). Deep neural networks learn non-smooth functions effectively. AI & Statistics.
- Suzuki, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality.
- Petersen, P., & Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. Neural Networks, 108, 296-330.
- Schmidt-Hieber, J. (2019a). Nonparametric regression using deep neural networks with ReLU activation function. Annals of Statistics, to appear.

- Nakada, R., & Imaizumi, M. (2019). Adaptive approximation and estimation of deep neural network to intrinsic dimensionality. arXiv preprint arXiv:1907.02177.
- Schmidt-Hieber, J. (2019b). Deep ReLU network approximation of functions on a manifold. arXiv preprint arXiv:1908.00695.
- Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep relu networks. Conference on Learning Theory.
- Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. International Conference on Learning Representations.
- Anthony, M., & Bartlett, P. L. (1999). Neural network learning: Theoretical foundations. cambridge university press.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. International Conference on Learning Representations.
- Neyshabur, B., Tomioka, R., & Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. arXiv preprint arXiv:1412.6614.
- Arora, S., Cohen, N., Hu, W., & Luo, Y. (2019). Implicit Regularization in Deep Matrix Factorization. Neural Information Processing Systems.
- Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In International Conference on Machine Learning.
- Kuzborskij, I., & Lampert, C. H. (2017). Data-dependent stability of stochastic gradient descent. arXiv preprint arXiv:1703.01678.