深層学習の数理

鈴木大慈
東京大学大学院情報理工学系研究科数理情報学専攻
理研AIP

2023年(令和5年)7月18日

東北大学集中講義

鈴木大慈

所属
- 東京大学大学院情報理工学系研究科数理情報学専攻・准教授
- 東大次世代知能科学研究センター研究部門研究者（研究知能部門）
- 理化学研究所 革新知能統合研究センター 深層学習理論チーム チームリーダー

鈴木大慈
情報理工学系研究科

専門
- 機械学習，数理統計学，統計的学習理論

主な研究内容
- 深層学習を含む様々な学習機構について理論的側面から研究を進めています．学習理論を通じて各種学習手法の汎化性能や学習アルゴリズムの収束性能を解明し複雑な学習過程の本質への理解を深め，理論をもとに新しい機械学習手法の構築や応用への還元を行っています．また，確率的最適化などの方法により大規模かつ複雑な機械学習問題を効率的に解く手法の開発も行っています．

応用

解釈可能性：
説明可能性，データの可視化，メンテナンスの容易化

各種テクニックの解析：
アーキテクチャの解析，損失関数の設計，最適化技法の解析

深層学習の原理解明：
「表現理論」「汎化誤差理論」「最適化理論」

学習の本質解明：
“良い”学習手法の特徴付け，統一理論，深層学習を優越する方法論の提唱

基礎

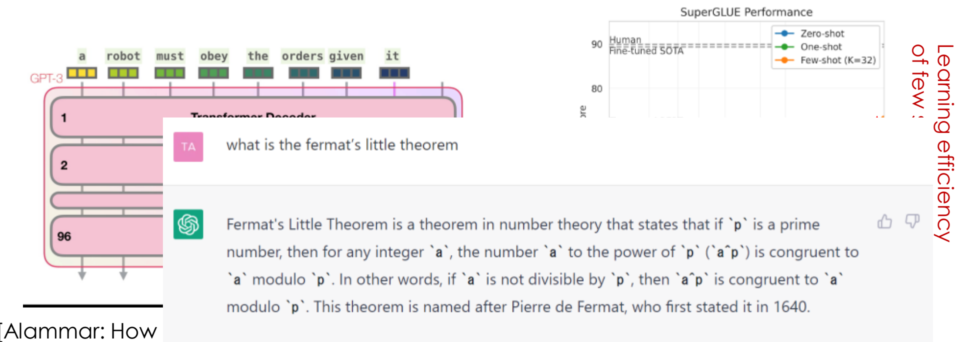
著書/授賞
- 『確率的最適化（機械学習プロフェッショナルシリーズ）』講談社，2015年8月8日．
- 金森敬文，鈴木大慈，竹内一郎，佐藤一誠：『機械学習のための連続最適化(機械学習プロフェッショナルシリーズ)』講談社，2016年12月7日．
- 文部科学大臣表彰・若手科学者賞「深層学習の原理解明に向けた統計的学習理論の研究」．文部科学省，2020年4月7日．
- 第11回日本統計学会研究業績賞 (2017年度)．2017年9月5日．
- Satoshi Hayakawa and Taiji Suzuki:日本神経回路学会論文賞．日本神経回路学会，2021年9月23日．
- 日本応用数理学会，ベストオーサー賞（論文部門）．2019年9月4日．

研究室URLとメール連絡先
- http://ibis.t.u-tokyo.ac.jp/suzuki/
- taiji@mist.i.u-tokyo.ac.jp

深層学習の広がり

様々なタスクで高い精度
なぜ？

Large language model

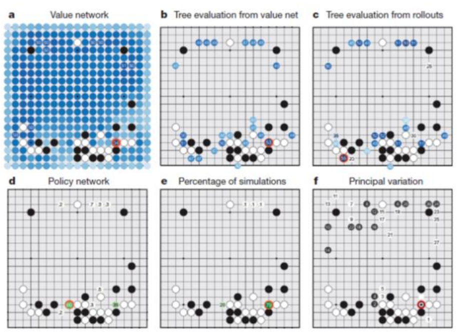


[Alammar: How
Animations,
https://jalammar.github.io/how-gpt3-works-visualizations-animations/]

[ChatGPT. OpenAI2022]

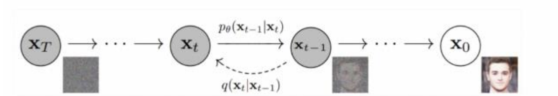
[Brown et al. "Language Models are Few-Shot Learners", NeurIPS2020]

AlphaGo/Zero



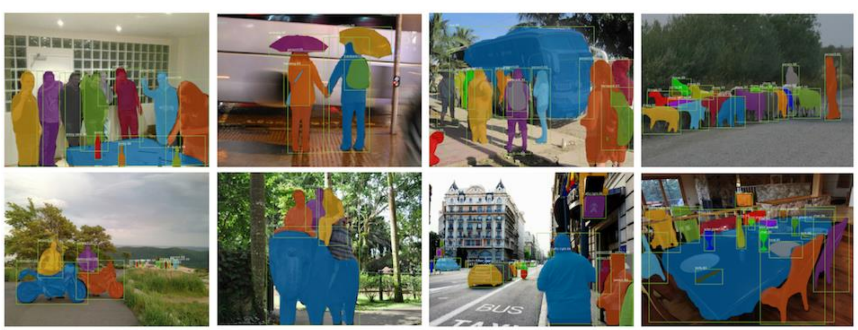
[Silver et al. (Google Deep Mind): Mastering the game of Go with deep neural networks and tree search, Nature, 529, 484—489, 2016]

Generative models (diffusion models)



[Ho, Jain, Abbeel: Denoising Diffusion Probabilistic Models. 2020]

Image recognition



[He, Gkioxari, Dollár, Girshick: Mask R-CNN, ICCV2017]

Stable diffusion, 2022.

Jason Allen "Théâtre D'opéra Spatial" generated by **Midjourney**. Colorado State Fair's fine art competition, 1st prize in digital art category

解決すべき問題点

なぜ深層学習はうまくいくのか？

- 「○○法が良い」という様々な仮説の氾濫.
- <u>世界的課題</u>

- 原理解明
- どうすれば“良い”学習が実現できるか？→新手法の開発

学会の問題意識

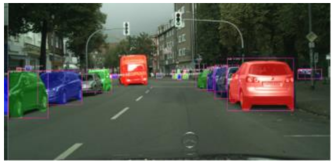
Ali Rahimi's talk at NIPS(NIPS 2017 Test-of-time award presentation)

Ali Rahimi's talk at NIPS(NIPS 2017 Test-of-time award presentation)

Ali Rahimi's talk at NIPS2017 (test of time award).
“Random features for large-scale kernel methods.”
“錬金術”という批判

民間の問題意識

- 中で何が行われているか分からないものは用いたくない.
- 企業の説明責任．深層学習のホワイトボックス化.



理論の必要性

深層学習(AI)の研究

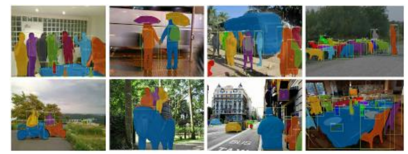
応用

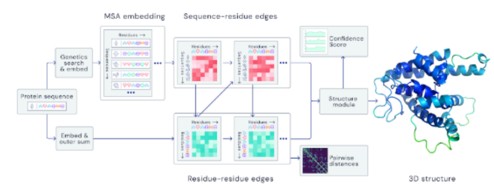
応用：AI手法の各種問題への応用

画像生成，パターン認識，たんぱく質構造予測

物理の対応物

半導体，新材料の開発





方法論：各種機械学習手法の開発

損失関数の設計，正則化法の開発，学習アルゴリズムの開発

各種方法論の<u>定式化</u>，学習アルゴリズムの開発

理論：統計的学習理論，最適化理論

深層学習の理論，収束レート解析，最適化アルゴリズム

各種機械学習手法の<u>原理解明</u>，最適性の<u>理論的保証</u>，
アルゴリズムの数理研究

素粒子論など

我々の得意分野

基礎

深層学習の理論概観

応用

解釈可能性：
説明可能性，データの可視化，メンテナンスの容易化

各種テクニックの解析：
アーキテクチャの解析，損失関数の設計，最適化技法の解析

深層学習の原理解明：
表現理論，汎化誤差理論，最適化の収束理論

学習の本質解明：
“良い”学習手法の特徴付け，統一理論，深層学習を優越する方法論の提唱

基礎

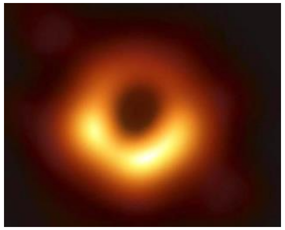
今日の範囲

理論を通して深層学習の不可思議な挙動を理解したい.
- 説明責任
- 可能性と限界の把握
- 学習手法設計の指針

応用から基礎まで広い範囲に“理論”は遍在する.

数学の役割

物理学



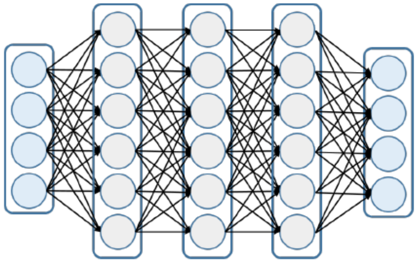
現象

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu}$$

- 相対性理論
- 量子力学
- リーマン幾何
- 関数解析

数学

機械学習（情報学）



深層学習

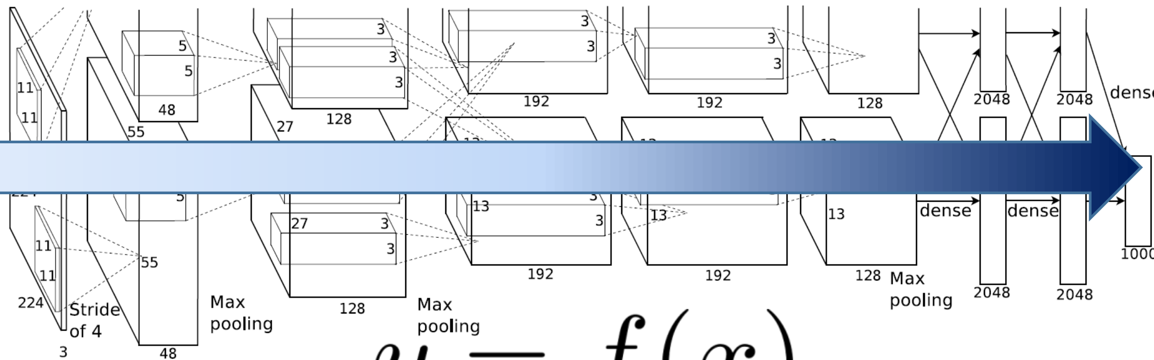
数学者・物理学者も参入

- 確率論
- 関数解析
- Wasserstein幾何
- 熱拡散方程式
- 統計学
- 最適化理論
- 数値計算

数学

NNの構造

画像



-猫　(y=1)
-犬　(y=2)
-人間 (y=3)

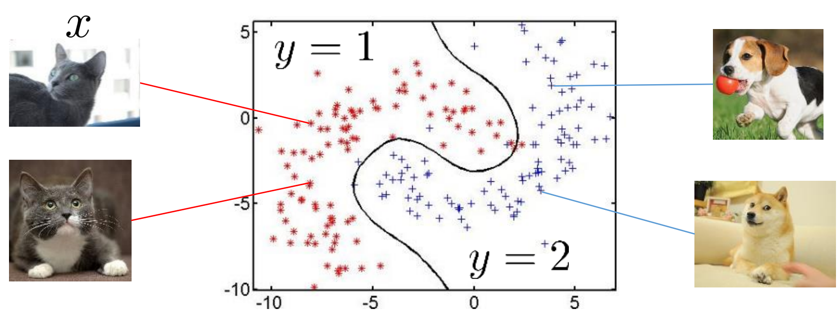
$$y = f(x)$$

$$x$$

ベクトル $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$

$$y$$

$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}$ ベクトル



学習：「関数」をデータに当てはめる
モデル：関数の集合（例：深層NNの表せる関数の集合）

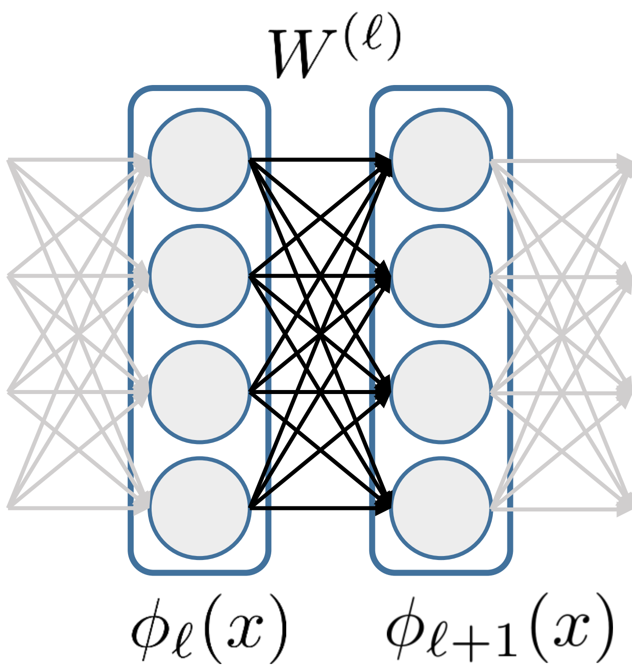
各層の詳細

アフィン変換＋活性化関数

$$\phi_{\ell+1}(x) = \eta(W^{(\ell)}\phi_\ell(x) + b^{(\ell)})$$

$$W^{(\ell)} \in \mathbb{R}^{m_{\ell+1}\times m_\ell}$$

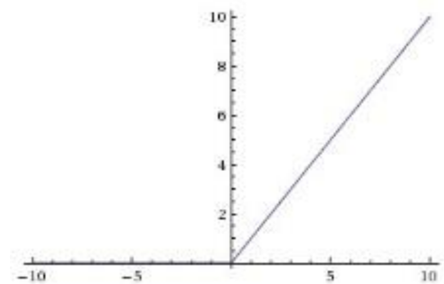
$$b^{(\ell)} \in \mathbb{R}^{m_{\ell+1}}$$



活性化関数

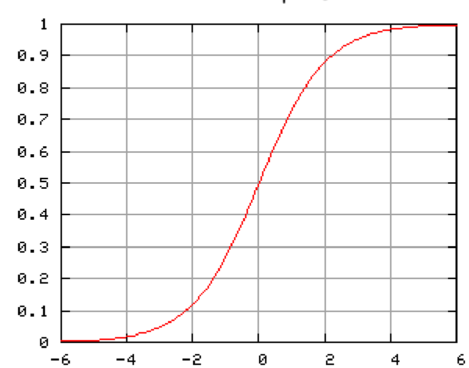
- ReLU (Rectified Linear Unit)

$$\eta(u) = \max\{u, 0\}$$



- シグモイド関数

$$\eta(u) = \frac{1}{1+e^{-u}}$$



訓練誤差と汎化誤差

パラメータ θ ：ネットワークの構造を表す変数
損失関数 $\ell(Y, f(X, \theta))$ ：パラメータ θ がデータをどれだけ説明しているか

予測誤差：損失の期待値

$$\mathrm{E}[\ell(Y, f(X, \theta))]$$

$$L(\theta)$$

本当に最小化したいもの.

訓練誤差：有限個のデータで代用

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i, \theta))$$

$$\widehat{L}(\theta)$$

代わりに最小化するもの.

(訓練データはテストデータと
同じ分布に従っていると仮定)

この二つには大きなギャップがある.
[過学習]

※クラスタリング等，教師なし学習も尤度を使ってこのように書ける.

学習理論の設定

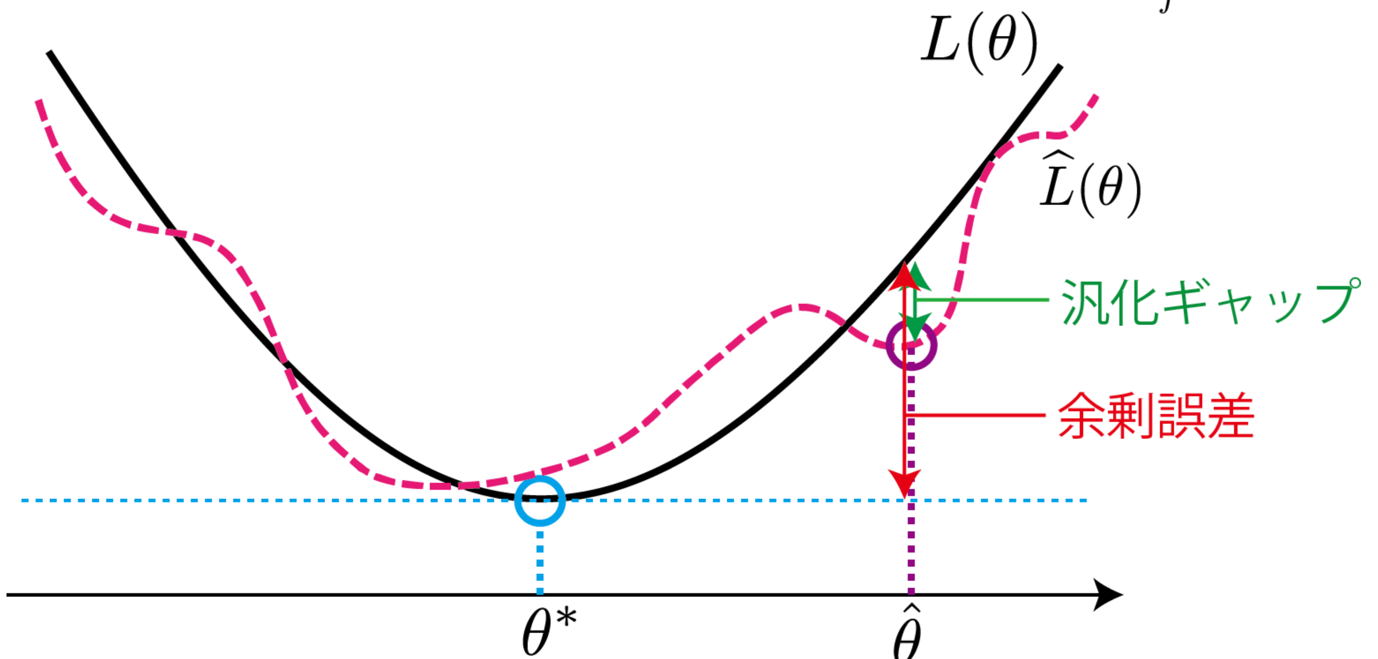
- 汎化ギャップ(汎化誤差)と余剰誤差

Generalization gap　　　Excess risk

汎化誤差: $L(\hat{\theta}) - \widehat{L}(\hat{\theta})$

余剰誤差: $L(\hat{\theta}) - \inf_{\theta} L(\theta)$

もしくは $L(\hat{\theta}) - \inf_{f} \mathbb{E}[\ell(Y, f(X))]$



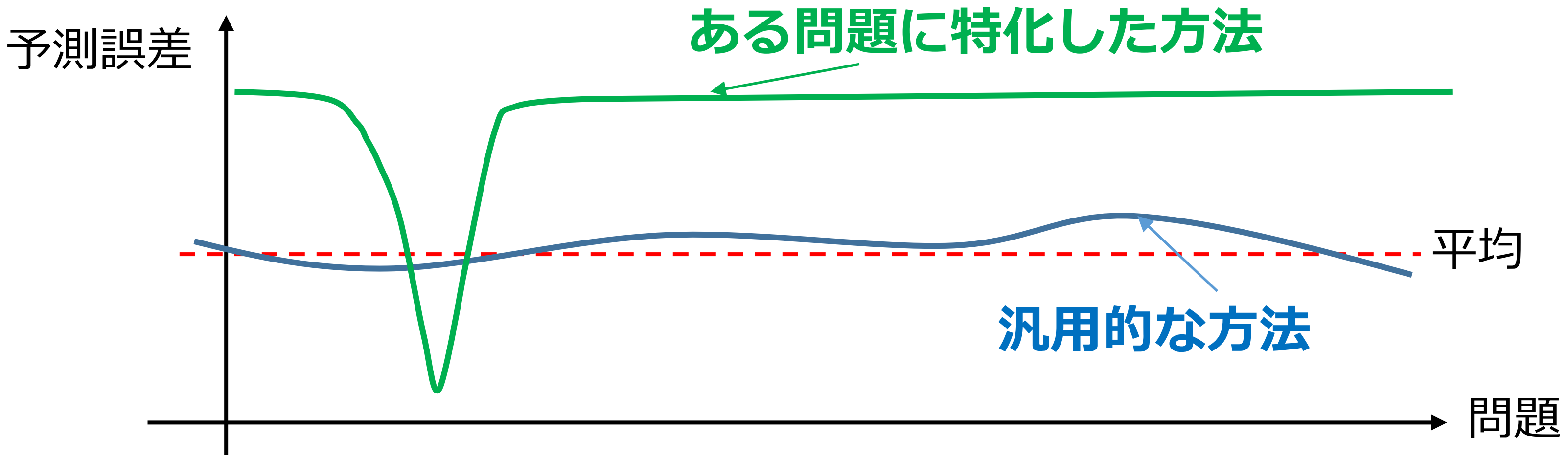
汎化誤差　最適化誤差　汎化誤差2　モデルバイアス

$$L(\hat{\theta}) - L^* = L(\hat{\theta}) - \hat{L}(\hat{\theta}) + \hat{L}(\hat{\theta}) - \hat{L}(\theta^*) + \hat{L}(\theta^*) - L(\theta^*) + L(\theta^*) - L^*$$

学習機の複雑さと学習能力

- No free lunch theorem

「あらゆる問題で最高の性能を出す汎用的学習機は実現不可能であり，ある問題に特殊化された手法に勝てない」



学習手法は「どこかを“贔屓”する必要がある」
→ モデリングの重要性 (オッカムの剃刀)

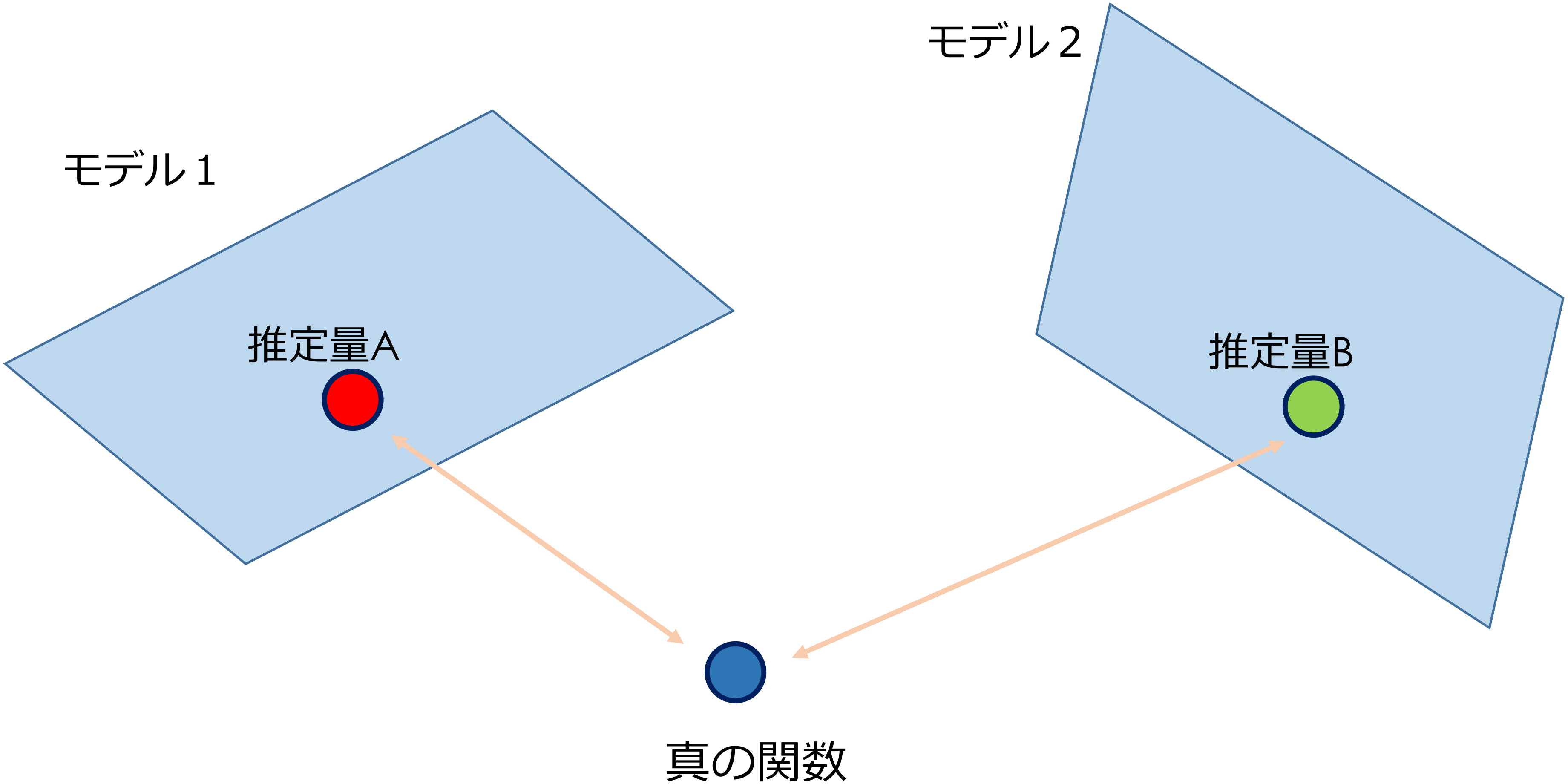
DLも例外ではない: CNN, Transformer, ResNet, …

William of Ockham：1285-1347．スコラ学の神学者，哲学者.
No free lunch theorem: [D.H.Wolpert and W.G. Macready: 1995,1997][Y.C. Ho and D.L. Pepyne: 2002]

「なんでもできる手法は個別問題で負ける」

「必要以上に複雑なモデルを当てはめると失敗する」

（注：すでに学習済みのモデルに関する話ではない．大量のデータで学習した汎用モデルが個別問題でも勝つことはある．あくまでゼロ知識の状態から学習を始めた場合の精度比較である．）



各推定量は対応するモデル内に値を取るし，正則化や事前分布の影響でモデル内でも取りやすい値と取りにくい値がある.
→ そういった場所を贔屓しているということ.

リスクの概念

- **真によらず常に予測誤差を最小化する推定量は存在しない.**
  - 「最適」な推定量をどう特徴づけるか？

「汎化誤差を最小化する」みたいな文言はそれ自体厳密な意味はない.

$\mathcal{P}$ ：真の分布のモデル　　$P^* \in \mathcal{P}$: 真の分布

$D^n = (x_i, y_i)_{i=1}^n$ ：訓練データ　　$\hat{\theta}, \tilde{\theta}$: 推定量

「真の分布のモデル」は本当に実際のデータ生成過程がそのモデルに入っている必要はなく仮想的なもので良い.
あくまで，その推定量がどの設定で最適であるかを特徴づけるための仮想的モデルと考えてよい.

■ ミニマックス最適性

$$\sup_{P^*\in\mathcal{P}} \mathbb{E}_{D^n\sim P^*}[L(\hat{\theta})] = \inf_{\tilde{\theta}:\mathrm{Estimator}} \sup_{P^*\in\mathcal{P}} \mathbb{E}_{D^n\sim P^*}[L(\tilde{\theta})]$$

(最悪誤差が最小)

■ 許容性　　つぎのような$\tilde{\theta}$が存在しない:

$$\mathbb{E}_{D^n\sim P^*}[L(\tilde{\theta})] \leq \mathbb{E}_{D^n\sim P^*}[L(\hat{\theta})] \quad (\forall P^* \in \mathcal{P})$$

$$\mathbb{E}_{D^n\sim P^*}[L(\tilde{\theta})] < \mathbb{E}_{D^n\sim P^*}[L(\hat{\theta})] \quad (\exists P^* \in \mathcal{P})$$

(一様に改善されることはない)

■ ベイズ最適性　　π_0: 事前分布

ベイズリスク $\int \mathbb{E}_{D^n\sim P^*}[L(\hat{\theta})]\mathrm{d}\pi_0(P^*)$ を最小にする推定法$\hat{\theta}$.
(→ ベイズ推定量)

(事前分布で重みづけたリスクを最小化)

深層学習の理論

理論的課題

表現能力

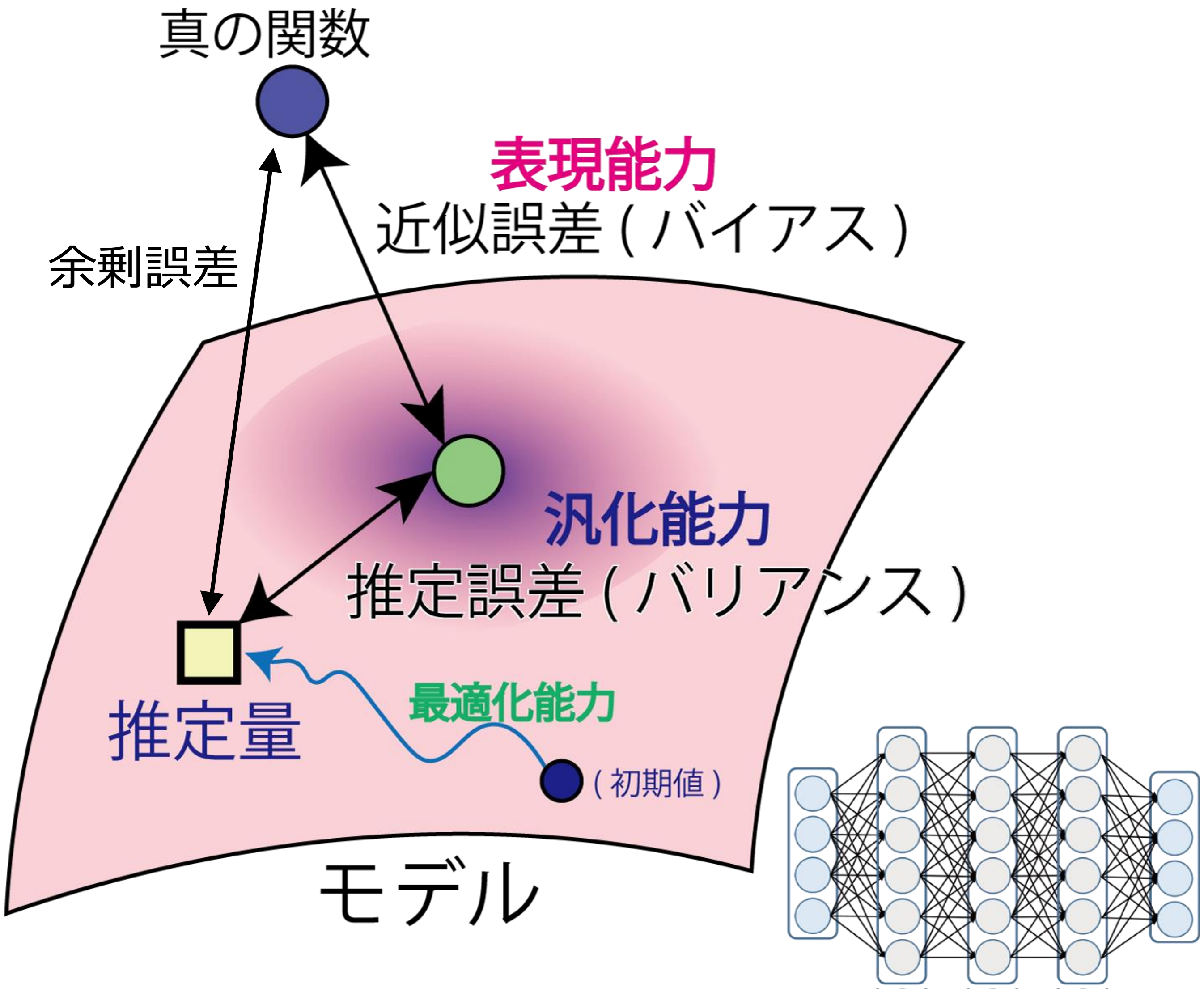
どれだけ難しい問題まで学習できるようになるか？

汎化能力

有限個のデータで学習した時，どれだけ正しく初見のデータを正解できるようになるか？

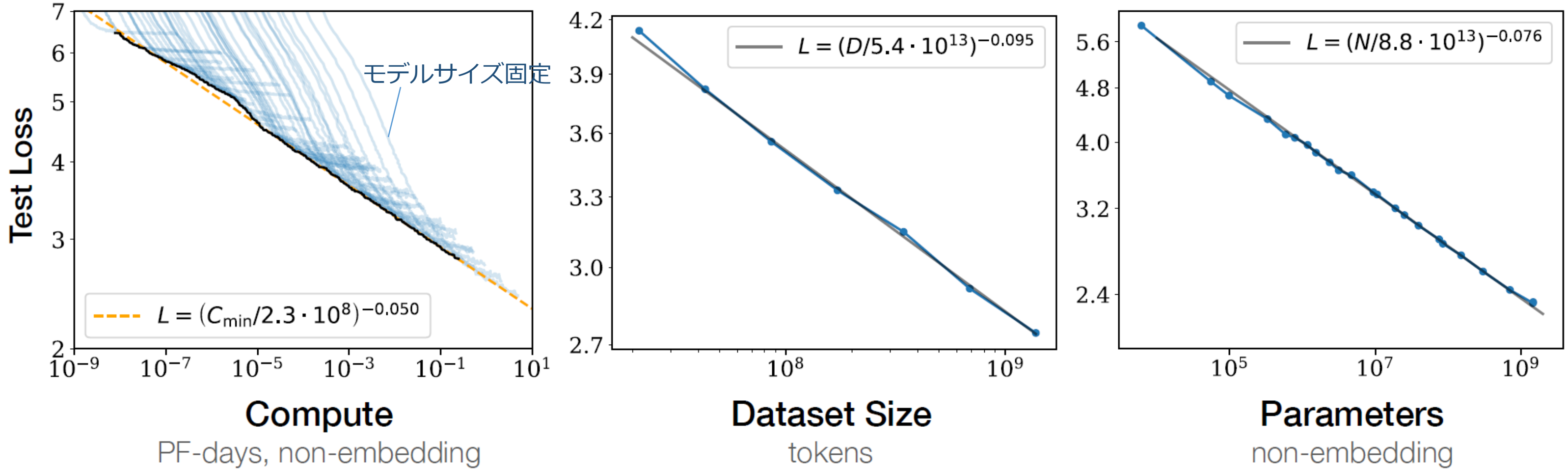
最適化能力

最適な重みを高速に計算機で求めることが可能か？


真の関数
表現能力
近似誤差（バイアス）
余剰誤差
汎化能力
推定誤差（バリアンス）
推定量
最適化能力
（初期値）
モデル

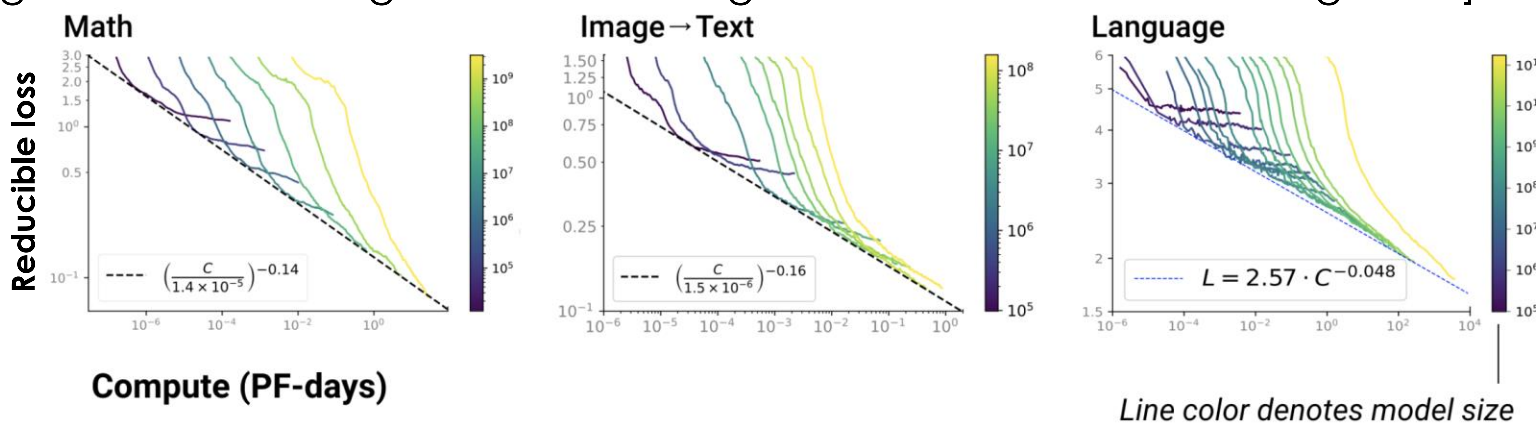
スケーリング則

[Kaplan et al.: Scaling Laws for Neural Language Models, 2020]



（基本的に訓練データサイズと思ってよい）

[Henighan et al.: Scaling Laws for Autoregressive Generative Modeling, 2020]



$$\boxed{\log(\text{予測精度}) = -\alpha \log(n) + \log(C)}$$

[Brown et al.: Language Models are Few-Shot Learners, 2020]　(GPT-3モデルの解析)

基本的考え方

- スケーリング則は古典的な学習理論でも現れる.

真のモデル

(正規直交系 in L2)

$$f^\circ(x) = \sum_{j=1}^{\infty} \alpha_j \varphi_j(x)$$

観測データ： $y_i = f^\circ(x_i) + \epsilon_i$

ただし

$\mu_j \sim j^{-a}$ を用いて

$\sum_{j=1}^{\infty} \alpha_j^2/\mu_j < \infty$

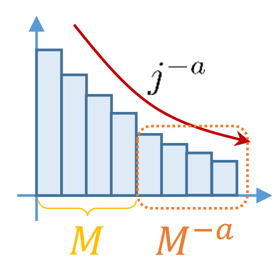
学習モデル

正則化学習法 (カーネル法)

$$\hat{f}(x) = \sum_{j=1}^{M} \hat{\alpha}_j \varphi_j(x) \Leftarrow \min_{(\hat{\alpha}_j)_{j=1}^{M}} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{M}\hat{\alpha}_j\varphi_j(x_i)\right)^2 + \lambda \sum_{j=1}^{M}\hat{\alpha}_j^2/\mu_j$$

予測誤差

$$\|f^\circ - \hat{f}\|_{L_2(P_X)}^2 \le C\left(\overbrace{\frac{M}{n}}^{\text{バリアンス}} + \overbrace{M^{-a}}^{\text{バイアス}}\right)$$



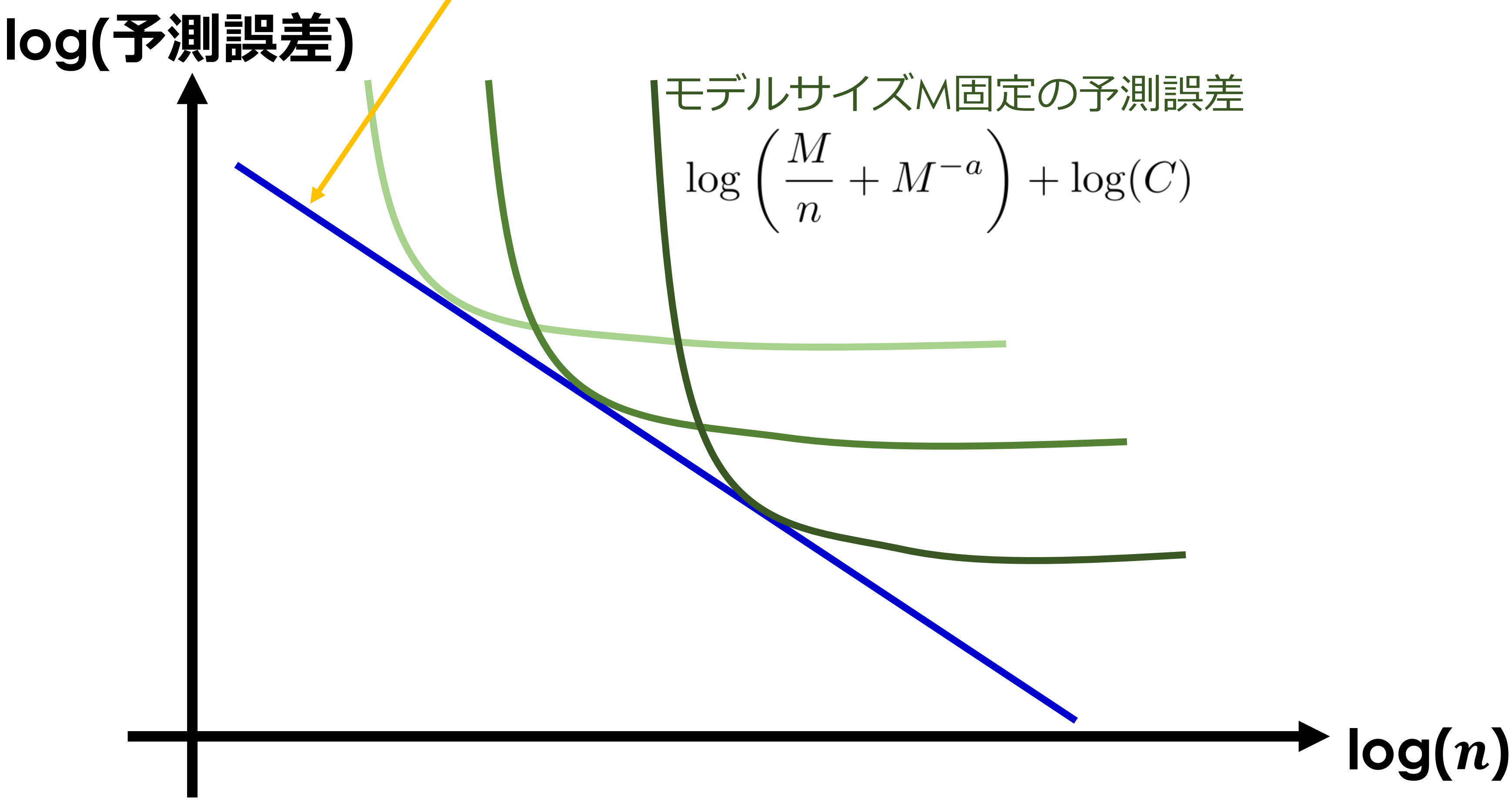
$$\xrightarrow{\text{最適なモデルサイズ}} \quad C n^{-\frac{a}{1+a}}$$

$$\log(\text{予測誤差}) = -\frac{a}{1+a}\log(n) + \log(C)$$

バリアンス=モデルの次元/n
バイアス=切り捨てた係数の二乗和

最適なモデルサイズの予測誤差

$$\log(\text{予測誤差}) = -\frac{a}{1+a}\log(n) + \log(C)$$



カーネル法の学習理論

- Caponnetto and De Vito. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, volume 7, pp.331–368 (2007).
- Steinwart and Christmann. *Support Vector Machines*. 2008.

関連する最近の論文

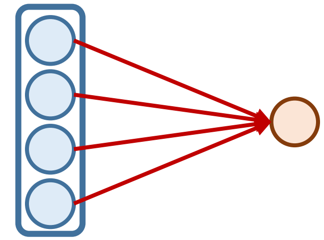
- Mei, Misiakiewicz, Montanari. Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration. arXiv:2101.10588.
- Bordelon, Canatar, Pehlevan. Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks. arXiv:2002.02561.
- Canatar, Bordelon, Pehlevan. Spectral Bias and Task-Model Alignment Explain Generalization in Kernel Regression and Infinitely Wide Neural Networks. arXiv:2006.13198.

深層学習との違い

- 線形モデル

非線形化

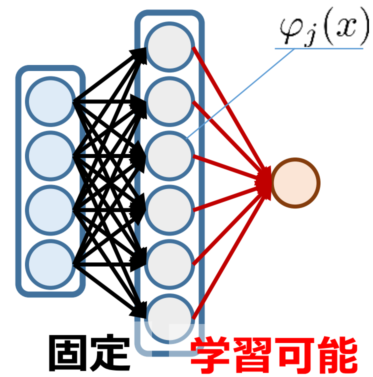
$$f(x) = \sum_{j=1}^{d} \alpha_j x_j$$



- カーネルモデル

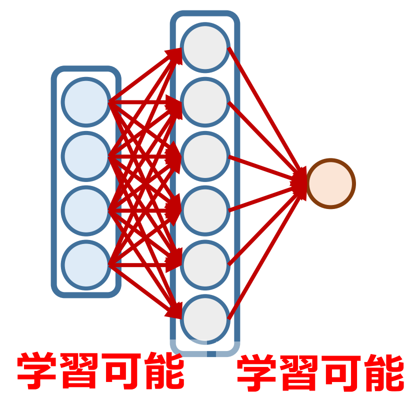
可変基底化

$$f(x) = \sum_{j=1}^{M} \alpha_j \underbrace{\varphi_j(x)}_{\text{固定}}$$



- 深層モデル

$$f(x) = \sum_{j=1}^{M} \alpha_j \underbrace{\varphi_j(x;\theta)}_{\text{学習可能}}$$

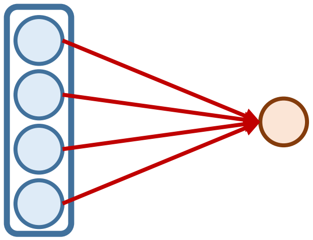


深層学習との違い

- 線形モデル

非線形化

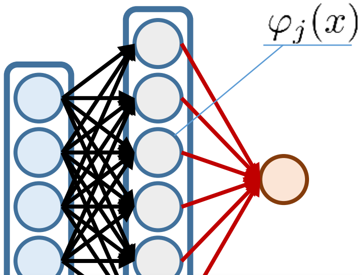
$$f(x) = \sum_{j=1}^{d} \alpha_j x_j$$



- カーネルモデル

可変基底化

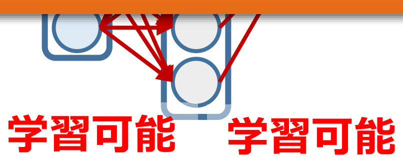
$$f(x) = \sum_{j=1}^{M} \alpha_j \varphi_j(x)$$



問題意識

- 基底を学習可能にすることで何が良くなるか？
- 逆に過学習を起こさないか？
- 最適化可能か？

学習可能



参考資料

Deep learning theory lecture notes

by Matus Telgarsky

- https://mjt.cs.illinois.edu/dlt/

その他，学習理論の教科書

- 一様バウンドを含む経験過程および無限次元統計モデルの教科書:
  Giné and Nickl, *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, 2016.
- 一様バウンドや Fast learning rate が網羅的に収録された教科書:
  Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.
- 経験過程の教科書: van der Vaart and Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.
- 統計的漸近論の教科書: van der Vaart, *Asymptotic statistics*. Cambridge University Press, 2000.
- 学習理論の教科書 (和書): 金森敬文, 統計的学習理論 (MLP シリーズ). 講談社, 2015.
- 学習理論の教科書:
  - Mohri et al., *Foundations of machine learning*. MIT press, 2018.
  - Shalev-Shwartz and Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

第1部
深層学習の表現能力

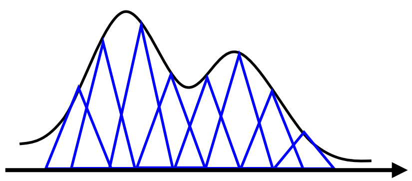
表現能力「万能近似能力」

二層ニューラルネットワークは
どんな関数も任意の精度で近似できる.

理論的にはデータが無限にあり，素子数が無限にあるニューラルネットワークを用いればどんな問題でも学習できる.

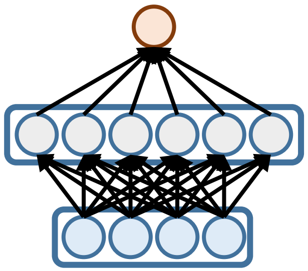
[Hecht-Nielsen,1987][Cybenko,1989]

「関数近似理論」

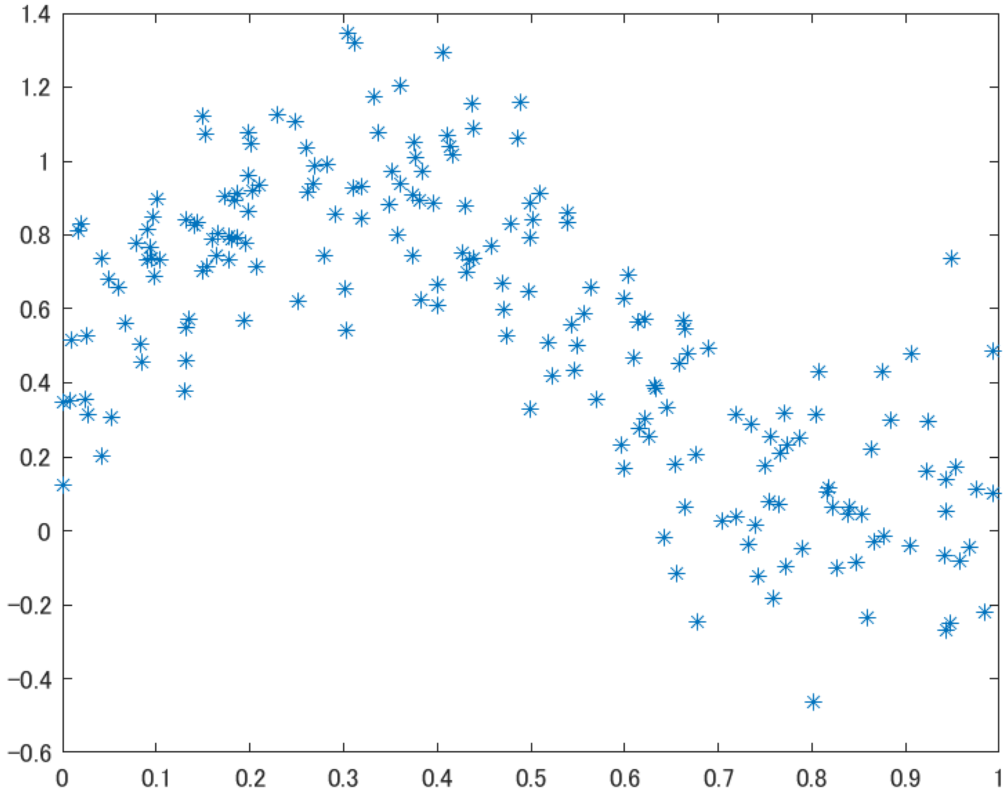


年		基底関数	空間
1987	Hecht-Nielsen	対象毎に構成	$C(R^d)$
1988	Gallant & White	Cos	$L_2(K)$
	Irie & Miyake	integrable	$L_2(R^d)$
1989	Carroll & Dickinson	Continuous sigmoidal	$L_2(K)$
	Cybenko	Continuous sigmoidal	$C(K)$
	Funahashi	Monotone & bounded	$C(K)$
1993	Mhaskar + Micchelli	Polynomial growth	$C(K)$
2015	Sonoda + Murata	**Unbounded**, admissible	$L_1(R^d)/L_2(R^d)$

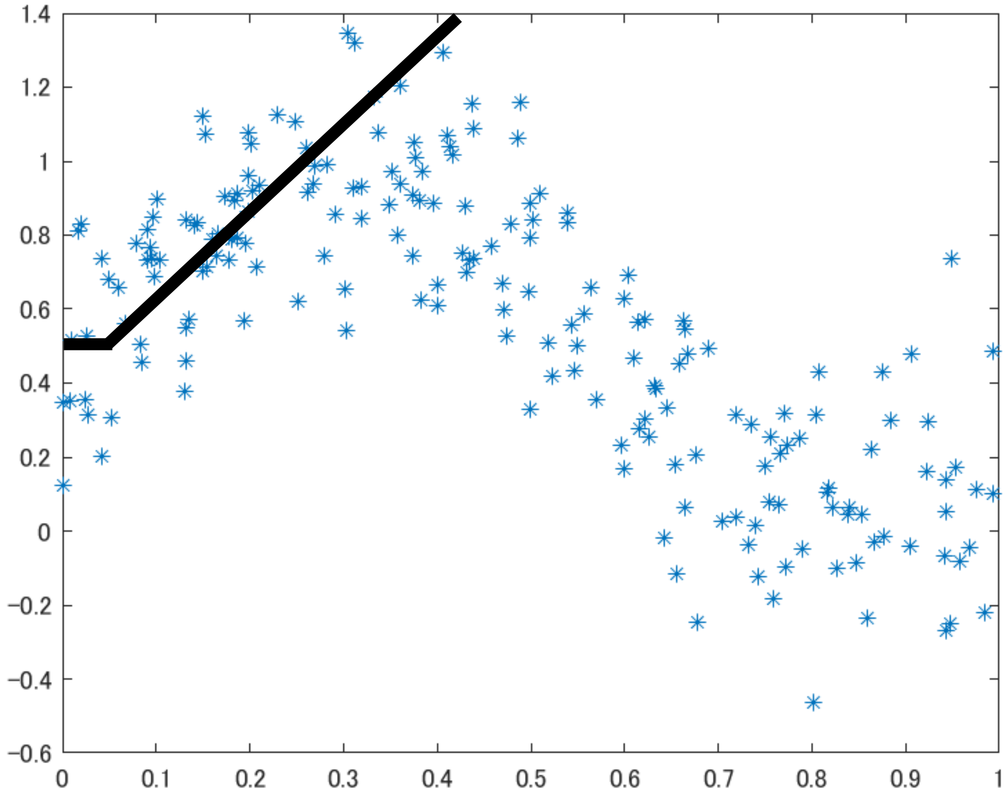
$$\hat{f}(x) = \sum_{j=1}^{m} v_j \eta(w_j^\top x + b_j)$$

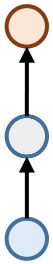


関数近似の様子



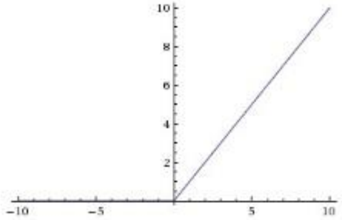
関数近似の様子





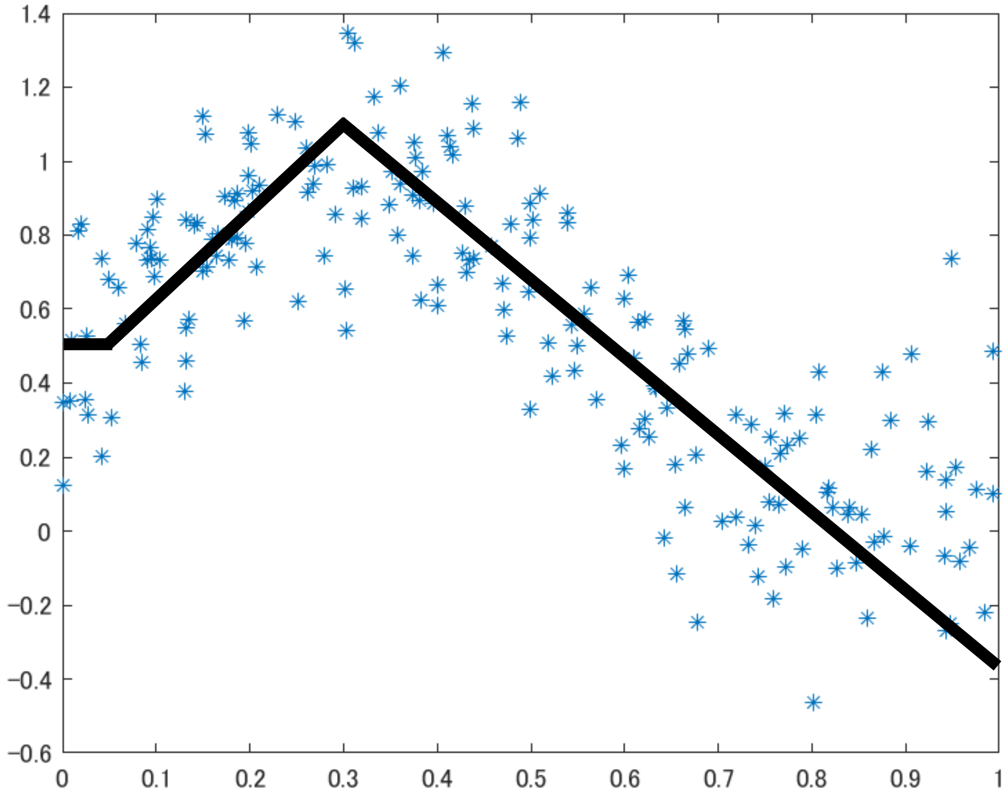
$$a_1\eta(b_1x + c_1)$$

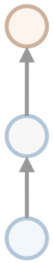
ReLU活性化関数

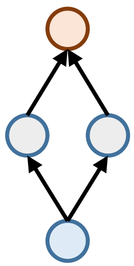


$$\eta(u) = \max\{u, 0\}$$

関数近似の様子



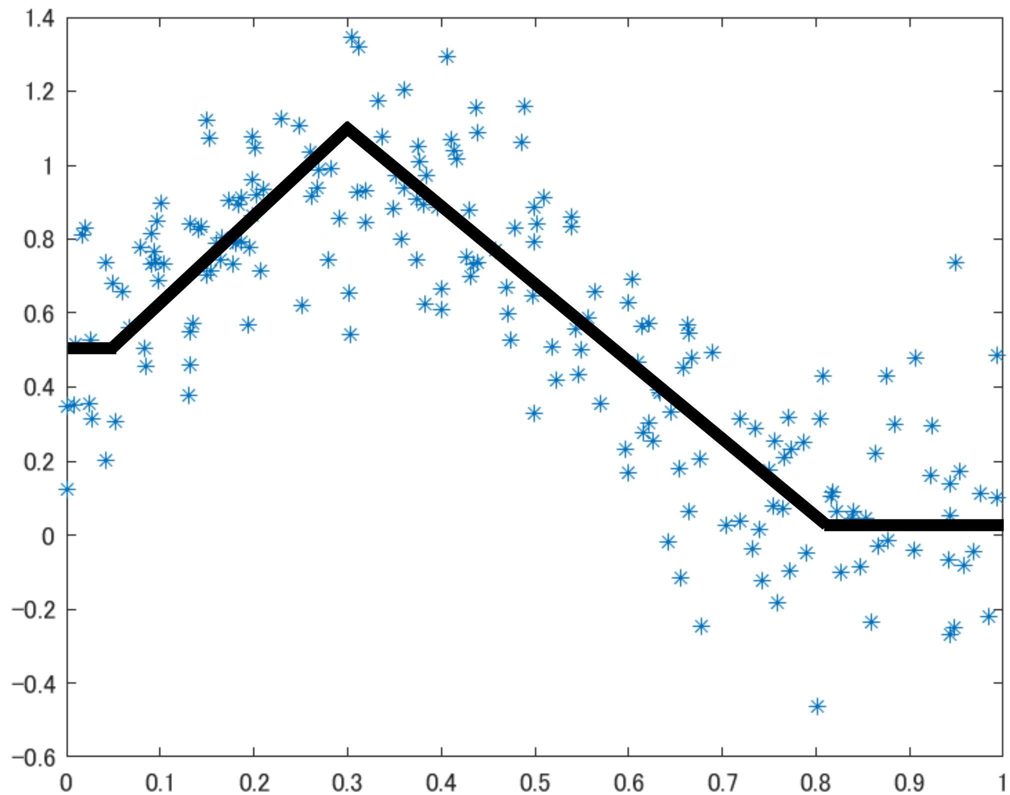


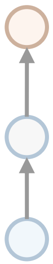


$a_1\eta(b_1x+c_1)$

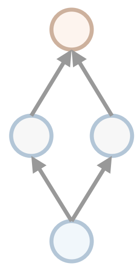
$$a_1\eta(b_1x+c_1)$$
$$+\,a_2\eta(b_2x+c_2)$$

関数近似の様子

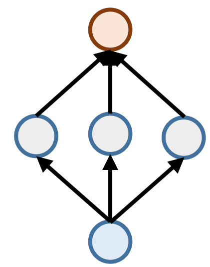




$$a_1\eta(b_1x+c_1)$$

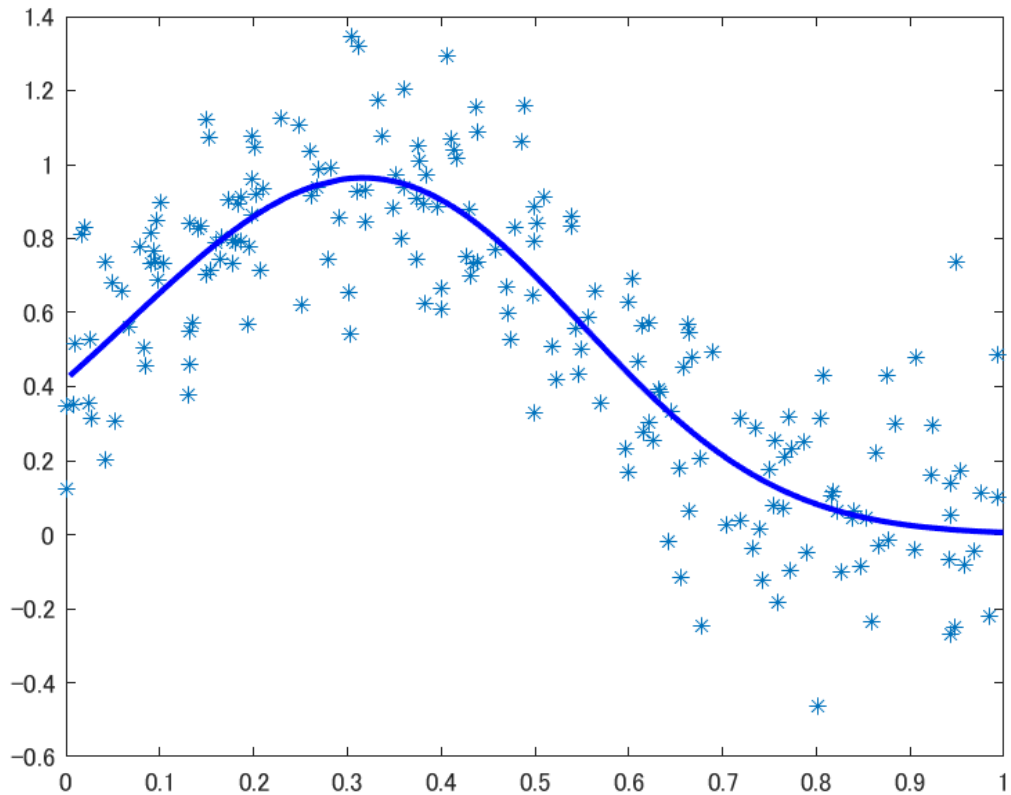


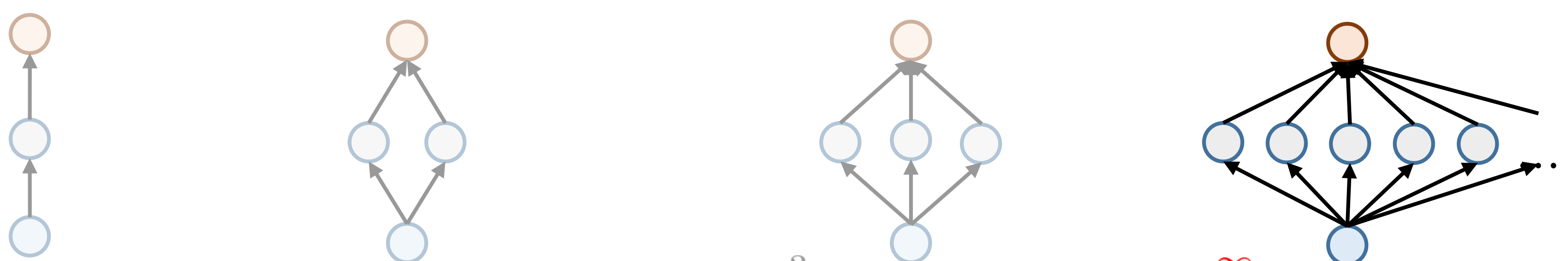
$$a_1\eta(b_1x+c_1)$$
$$+\,a_2\eta(b_2x+c_2)$$



$$\sum_{j=1}^{3} a_j\eta(b_jx+c_j)$$

関数近似の様子





$$a_1\eta(b_1x+c_1)$$

$$a_1\eta(b_1x+c_1) + a_2\eta(b_2x+c_2)$$

$$\sum_{j=1}^{3} a_j\eta(b_jx+c_j)$$

$$\sum_{j=1}^{\infty} a_j\eta(b_jx+c_j)$$

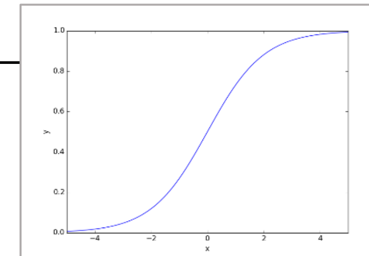
連続関数の近似

- Cybenkoの理論

[Cybenko: Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems,* 2(4): 303-314, 1989]

定義

活性化関数ηがシグモイド的 ⇔ $\eta(x) \to \begin{cases} 1 & (x \to \infty) \\ 0 & (x \to -\infty) \end{cases}$



定理

活性化関数ηが連続なシグモイド的関数なら，任意の$f \in C([0,1]^d)$と，任意の$\epsilon > 0$に対して，ある$g(x) = \sum_{i=1}^{N} \alpha_i \eta(a_i x_i + b_i)$ が存在して，

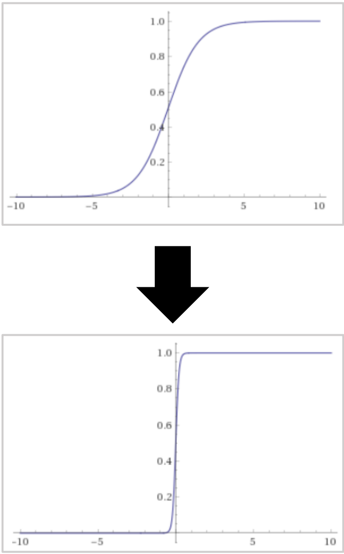
$$\sup_{x \in [0,1]^d} |f(x) - g(x)| \leq \epsilon$$

とできる.

証明の直感的概略

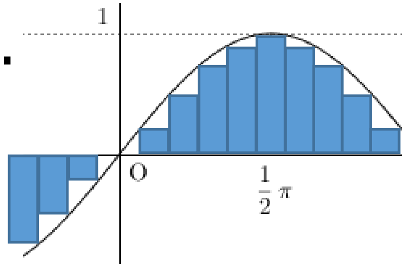
- シグモイド型の関数に対し，

$$h\left(a(\alpha^\top x+\beta)+\theta\right)\xrightarrow{a\to\infty}\begin{cases}1 & (\alpha^\top x+\beta>0)\\ h(\theta) & (\alpha^\top x+\beta=0)\\ 0 & (\alpha^\top x+\beta<0)\end{cases}$$



が成り立つ．つまり，スケールを適切に選べば，
階段関数をいくらでもよく近似できる．

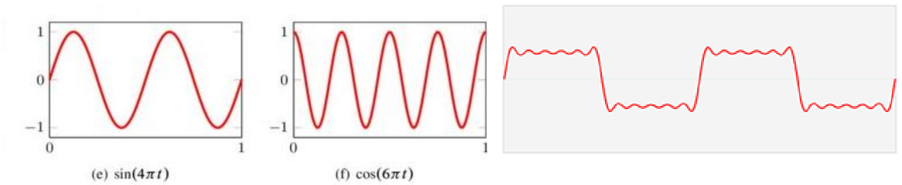
- 階段関数を近似できれば，それらを足し引きすることで，
  $\cos(\alpha^\top x+\beta)$ や $\sin(\alpha^\top x+\beta)$ をいくらでもよく近似できる．
- cos, sinが実現できるならFourier(逆)変換もできる．
- 任意の連続関数が近似できる．



積分表現 (Ridgelet変換)

- Fourier変換

$$f(x) = \int_{\omega \in \mathbb{R}^d} F(\omega) e^{i\omega^\top x} \mathrm{d}\omega$$

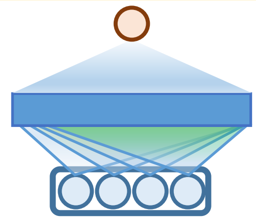


[Wikipedia,フーリエ変換]

- Ridgelet変換

NNはFourier変換におけるsin,cosの代わりに非線形ノードの足し合わせで関数を表現.

$$f(x) = \int T(a,b)\eta(a^\top x - b)\mathrm{d}b\mathrm{d}a$$



ウェーブレット変換 + ラドン変換

ある$\psi: \mathbb{R} \to \mathbb{R}$が以下の「許容条件」を満たすとする：

$$K_{\psi,\eta} := (2\pi)^{d-1} \int_{\mathbb{R}\setminus\{0\}} \frac{\bar{\hat{\psi}}(\xi)\hat{\eta}(\xi)}{|\xi|^d} \mathrm{d}\xi < \infty$$

($\hat{\psi}, \hat{\eta}$はFourie変換)

$$\mathscr{R}_\psi f(a,b) = \int_{\mathbb{R}^d} f(x)\overline{\psi(x^\top a - b)}\|a\|\mathrm{d}x$$

(Ridgelet変換)

$$\mathscr{R}_\eta^\dagger T(x) = \int_{a\in\mathbb{R}^d}\int_{\mathbb{R}} T(a,b)\eta(a^\top x - b)\|a\|^{-1}\mathrm{d}b\mathrm{d}a$$

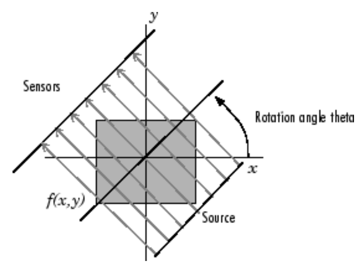
(双対Ridgelet変換)

定理

$$(\mathscr{R}_\eta^\dagger \mathscr{R}_\psi f)(x) = K_{\psi,\eta} f(x)$$

(再構成定理)

CTスキャン



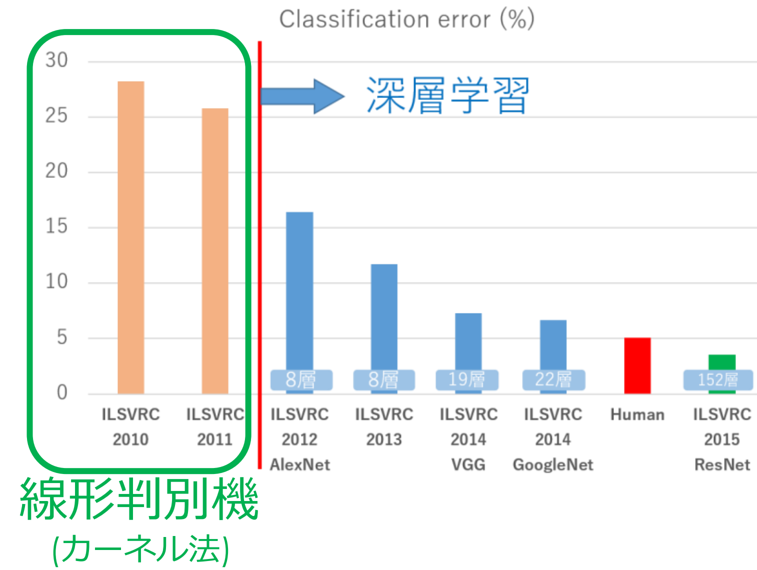
カーネル法

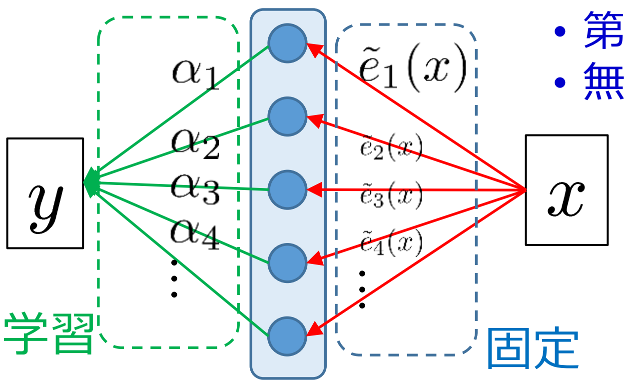
- 浅い手法の代表格.
- これも万能近似能力がある.

第1層目を固定した横幅無限の2層ニューラルネットワーク

似た手法
- スプライン法
- 局所多項式回帰
- シリーズ推定量

(線形推定量と呼ばれるクラス)



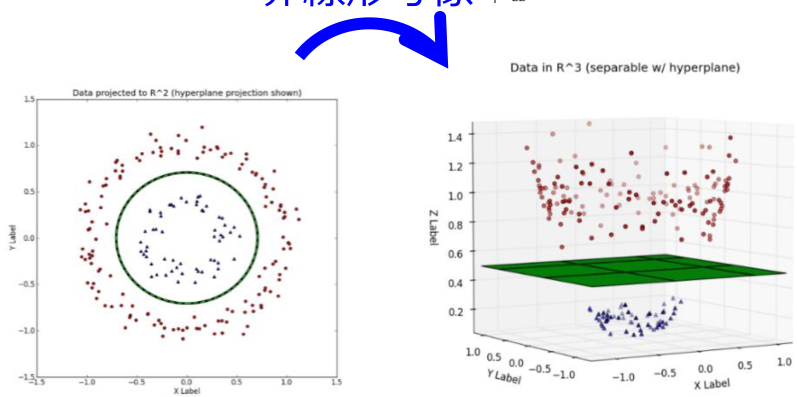


- 第一層目固定
- 無限個の素子

$$k(x, x') = \sum_{m=1}^{\infty} \tilde{e}_m(x)\tilde{e}_m(x') \text{ : カーネル関数}$$

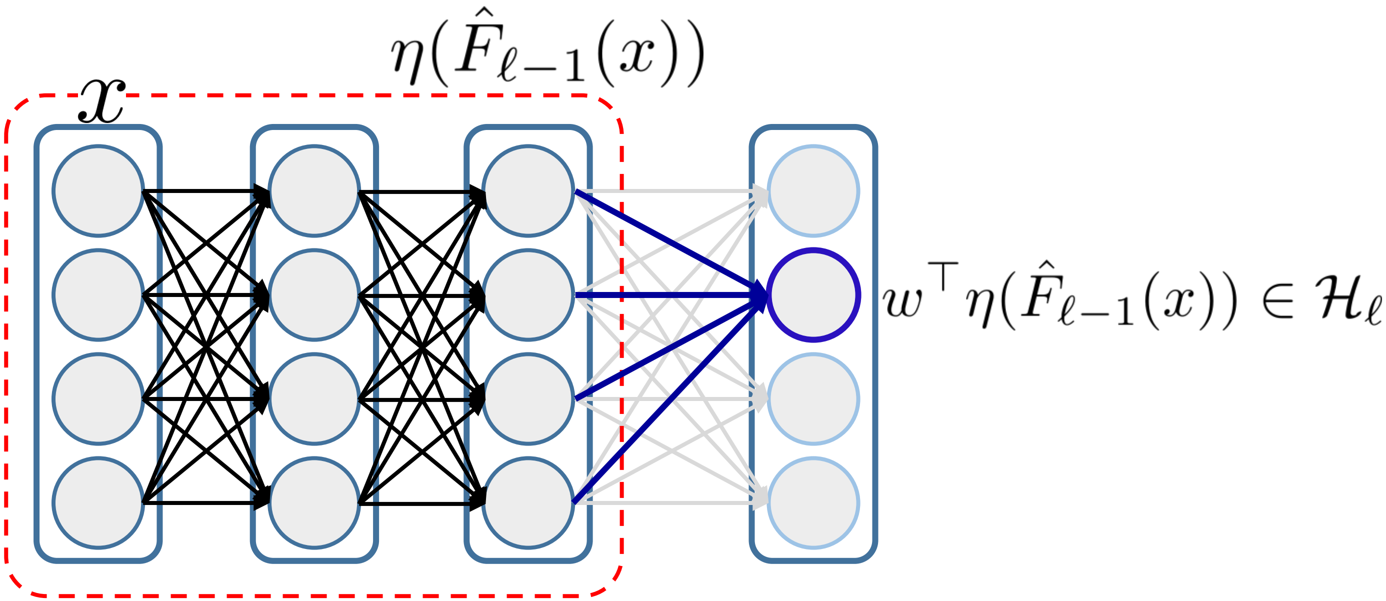
再生核ヒルベルト空間の理論

非線形写像 ϕ_x



深層NNとカーネル関数

深層NNは「カーネル関数をデータに合わせて学習する方法」と言える



$$\hat{k}_\ell(x,x') = \eta(\hat{F}_{\ell-1}(x))^\top \eta(\hat{F}_{\ell-1}(x'))$$

$\hat{k}_\ell$に対応した再生核ヒルベルト空間

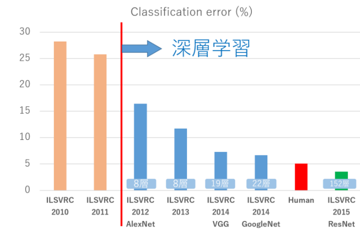
$$\mathcal{H}_\ell = \{f(x) = w^\top \eta(\hat{F}_{\ell-1}(x)) \mid w \in \mathbb{R}^{m_\ell}\},$$
$$\|f\|_{\mathcal{H}_\ell} = \|w\|.$$

(より正確には同じfを与えるwの中でinfを取る)

See Montavon et al. (2011); Bach (2015)

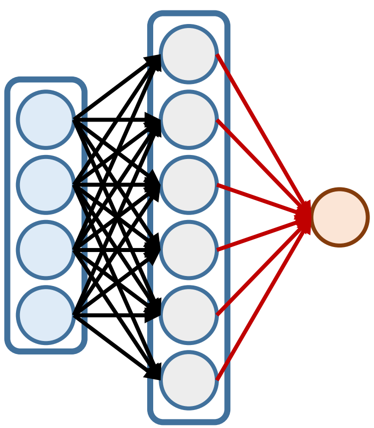
これまでわかったこと

- **[理論]** 万能近似能力という意味では浅層で十分.
- **[実際]** 実際は多層を使うことが多い.

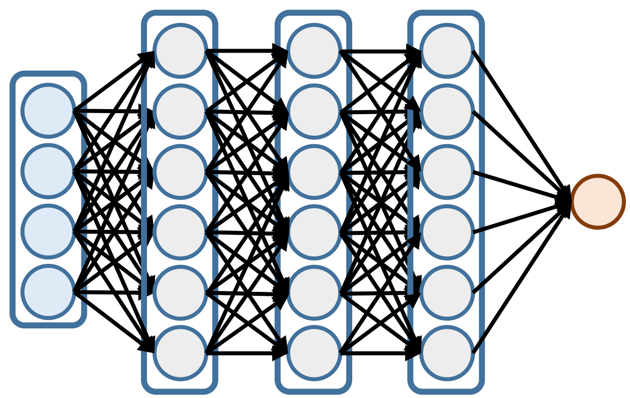


→　この差はどう埋める？

カーネル法
浅層



多層ニューラルネット
深層学習



→「表現力」を比べてみる.

(1)ランダム特徴量の近似能力

ランダム特徴： $\hat{f}(x) = \sum_{j=1}^{M} r_j \sigma(u_j^\top x + b_j)$

これは近似対象に合わせて自由に設定

ランダムに生成(球面上の一様分布) (固定)

疑問： ある対象の関数f^*を近似するのに必要なランダム特徴量の数Mはどれくらい？

実は，一つのニューロンを近似するのにも$M = \exp(\Omega(d))$必要． **(次元の呪い)**

つまり，ランダムに特徴量を生成する方法は非常に効率が悪い．
深層学習のように<u>第一層目のパラメータも学習した方が効率的</u>． (特徴学習の必要性)

※ ランダム特徴量の方法は，カーネルの低ランク近似法とみなせる．

定理 (Yehudai and Shamir 2019)

ある $u^* \in \mathbb{R}^d$, $b^* \in \mathbb{R}$ s.t. $\|u^*\| \leq d^2$ が存在して，

$$\mathbb{E}_X\left[(\hat{f}(X) - \sigma(u^{*\top} X + b^*))^2\right] \leq 0.1$$

であるためには，$M \cdot \max_j |r_j| \geq \exp(\Omega(d))$ が必要．

[Yehudai and Shamir. On the power and limitations of random features for understanding neural networks. NeurIPS2019.]

(2-a) Barron classとランダム特徴量

π: probability measure

- πで決まる再生核ヒルベルト空間：　　　　　(σ: ReLU)

$$\mathcal{H}_\pi = \left\{ \int_{\mathbb{S}^{d-1}} a(w)\sigma(w^\top x)\pi(\mathrm{d}w) \;\middle|\; \int_{\mathbb{S}^{d-1}} a(w)^2 \pi(\mathrm{d}w) < \infty \right\}$$

$$\|f\|_{\mathcal{H}_\pi}^2 := \mathbb{E}_{w\sim\pi}[|a(w)|^2] \quad \text{ただし } f = \int a(w)\sigma(w^\top x)\pi(\mathrm{d}w)$$

（実際，これは再生核を$k(x,x') = \int_{\mathbb{S}^{d-1}} \sigma(w^\top x)\sigma(w^\top x')\pi(\mathrm{d}w)$とするRKHSになっている．）

- M個のランダム特徴量で張られるモデル：

$$\mathcal{H}_{\mathrm{rand}}(M) = \left\{ \hat{f}(x) = \sum_{j=1}^{M} r_j \sigma(u_j^\top x) \;\middle|\; r_j \in \mathbb{R} \right\}$$

ランダムに生成して固定
(近似対象によって変えない)

定理 (Barron classの近似定理; E, Ma, Wu, 2019)

ある確率測度πが存在して，そのπによって決まる再生核ヒルベルト空間の元$f \in \mathcal{H}_\pi$ ($\|f\|_{\mathcal{H}_\pi} \leq 1$) を適切に取ってくると，以下が成り立つ:

$$\inf_{\hat{f}\in\mathcal{H}_{\mathrm{rand}}(M)} \|f - \hat{f}\|_{L_2(P_X)} \gtrsim \frac{1}{dM^{1/d}}$$

→ **次元の呪い**　(誤差ϵを達成するには$M = \epsilon^{-\Omega(d)}$必要)

[E, Ma, Wu: A Comparative Analysis of Optimization and Generalization Properties of Two-layer Neural Network and Random Feature Models Under Gradient Descent Dynamics. arXiv:1904.04326, 2019]

(2-b) NNによる次元の呪いの解消

一方で，近似対象の関数fに対して，適切に一層目のパラメータ$(u_j)_{j=1}^{M}$を設定すると，次元の呪いを回避できる．

これも動かせる

$$\mathcal{H}_{\mathrm{NN}}(M) = \left\{ \hat{f}(x) = \sum_{j=1}^{M} r_j \sigma(u_j^\top x) \;\middle|\; r_j \in \mathbb{R},\ u_j \in \mathbb{S}^{d-1} \right\}$$ ：NNの集合

$$\inf_{\hat{f} \in \mathcal{H}_{\mathrm{NN}}} \|f - \hat{f}\|_{L_2(P_X)}^2 = O\left(\frac{1}{M}\right)$$

NNによって関数近似の効率が改善されている．

$$M^{-1/d} \Rightarrow M^{-1}$$

(一層目固定)　(一層目可変)

直感：ランダム特徴量だと(d-1)次元単位球面全体を覆うのに，$1/\epsilon^d$個のニューロンが必要．

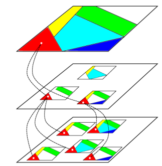
[E, Ma, Wu: A Comparative Analysis of Optimization and Generalization Properties of Two-layer Neural Network and Random Feature Models Under Gradient Descent Dynamics. arXiv:1904.04326, 2019]

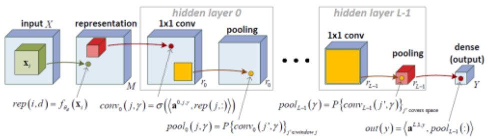
多層化による表現力の増大

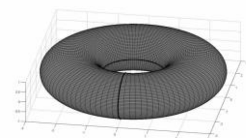
参考

深さに対して指数関数的に“表現力”が上がる．

- **超平面アレンジメント** [Montufar et al., 2014]
  空間の領域分割数
- **多項式展開，テンソル解析** [Cohen et al., 2016; Cohen & Shashua, 2016]
  単項式の次数
- **代数トポロジー** [Bianchini & Scarselli, 2014]
  ベッチ数(Pfaffian)
- **リーマン幾何 + 平均場理論** [Poole et al., 2016]
  埋め込み曲率

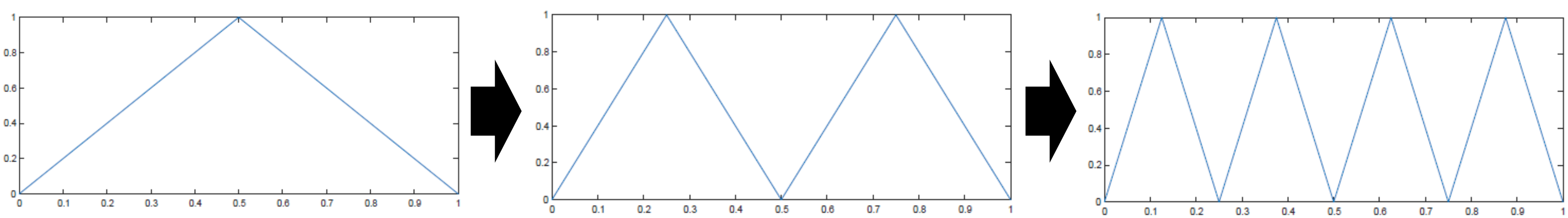






特に，対称性の高い関数の近似は深層NNが有利

$$h(x) = \begin{cases} 2x & (0 \leq x \leq 1/2) \\ 2(1-x) & (1/2 \leq x \leq 1) \\ 0 & (\text{otherwise}). \end{cases}$$



$h(x)$ $\quad$ $h \circ h(x)$ $\quad$ $h \circ h \circ h(x)$

多層が得する例 (1): 領域分割数

参考

NNの“表現力”：領域を何個の多面体に分けられるか？

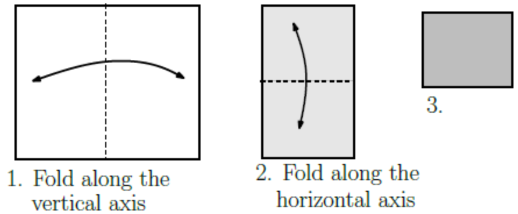
- 層の数に対して表現力は**指数的**に上がる.

$$\left(\frac{n}{n_0}\right)^{L-1}\sum_{j=0}^{n_0}\binom{n}{j}$$

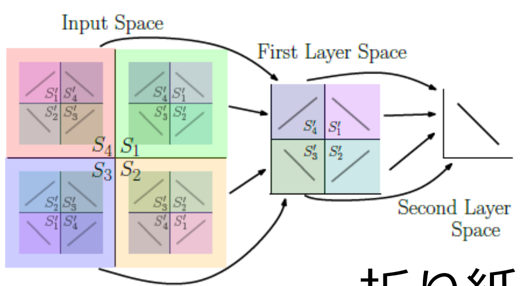
- 中間層のユニット数 (横幅) に対しては**多項式的**.

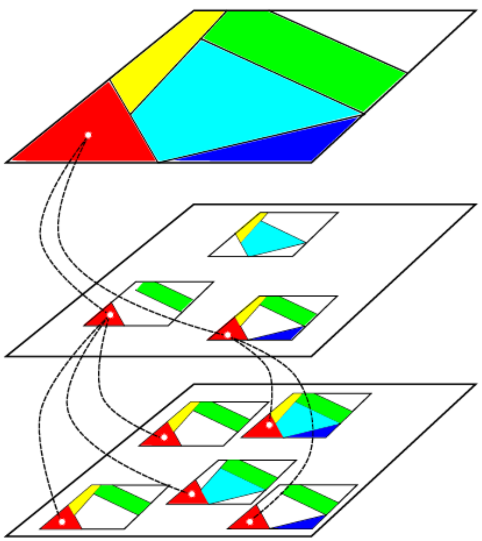
$$\sum_{j=0}^{n_0}\binom{n}{j}$$

L：層の数
n：中間層の横幅
n_0：入力の次元



(a)





折り紙のイメージ

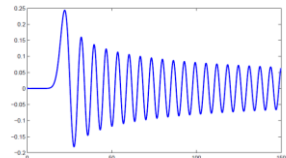
Montufar, Guido F., et al. "On the number of linear regions of deep neural networks." 2014.

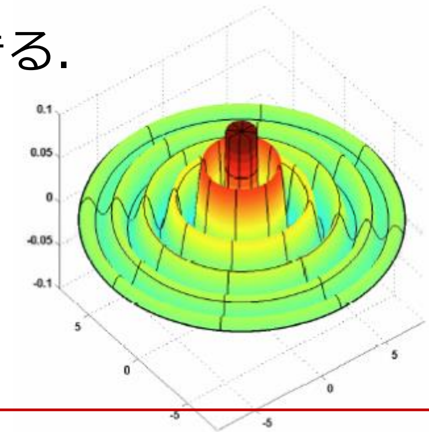
多層が得する例 (2): 対称な関数 参考

深層NNは特徴量を適切に抽出することで次元の呪いを回避できる.

$$g(\|x\|^2) = g(x_1^2 + x_2^2 + \cdots + x_{d_x}^2)$$

gはBessel関数を元に構成





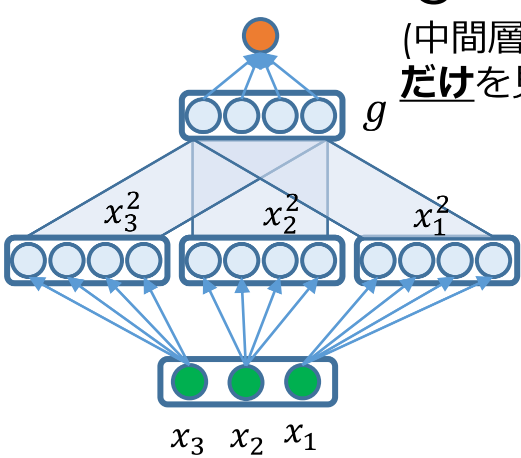
三層 (中間層二層) : $O(\mathrm{poly}(d_x))$ノードで十分
二層 (中間層一層) : $\Omega(\exp(d_x))$ノードが必要 (次元の呪い)

(Eldan&Shamir, 2016)

三層

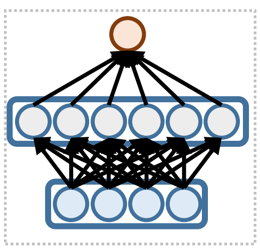
まず二乗和$x_1^2 + \cdots + x_{d_x}^2$を作ってからgを作用.

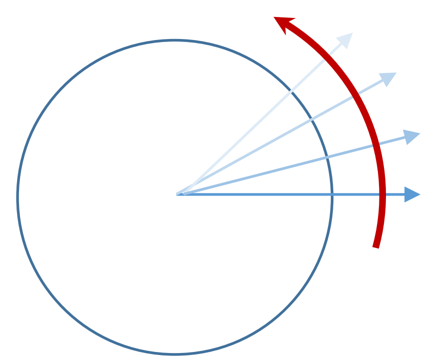
(中間層で**座標軸方向だけ**を見ればよい)



二層

全方向をケアする必要がある
(座標軸方向だけではダメ)



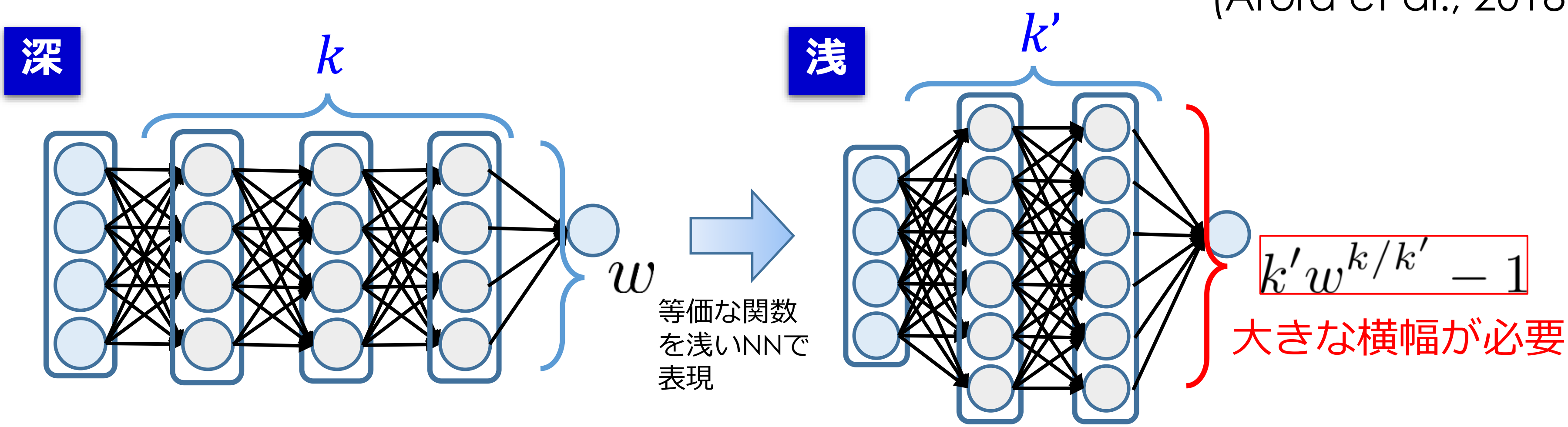


多層が得する例 (3): 区分線形関数の表現

参考

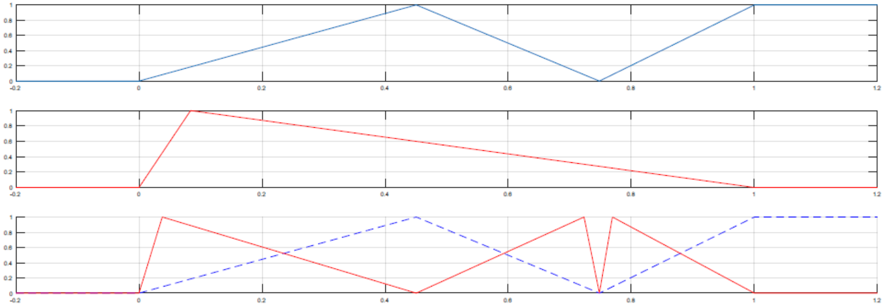
- 任意の区分線形関数($\mathbb{R}^d \to \mathbb{R}$)は深さ$\lceil \log_2(d+1) \rceil$のReLU-DNNで表現可能
- ある横幅w，縦幅kのReLU-DNNが存在して，それを縦幅$k'(< k)$のネットワークで表現するには横幅$k'w^{k/k'}-1$が必要.

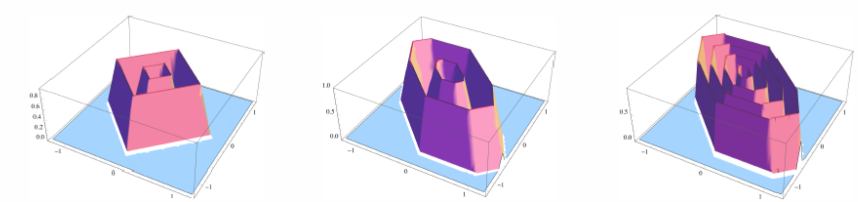
(Arora et al., 2018)



やはり層の深さに対し指数関数的に表現力が増加

上記のネットワークの例





(a) $H_{\frac{1}{2},\frac{1}{2}} \circ N_{\ell_1}$ (b) $H_{\frac{1}{2},\frac{1}{2}} \circ \gamma_{Z(\mathbf{b}^1,\mathbf{b}^2,\mathbf{b}^3,\mathbf{b}^4)}$ (c) $H_{\frac{1}{2},\frac{1}{2},\frac{1}{2}} \circ \gamma_{Z(\mathbf{b}^1,\mathbf{b}^2,\mathbf{b}^3,\mathbf{b}^4)}$

多層が得する例 (4): 有理関数の近似

参考

- 有理関数をReLU-DNNで近似

$$p : [0,1]^d \to [-1,+1],\ q : [0,1]^d \to [2^{-k}, +1]$$

：r次多項式

p/q をReLU-DNNで近似したい

あるReLU-DNN fが存在してノード数と近似誤差が次のように抑えられる：

ノード数

$$\mathrm{O}(\mathrm{poly}(k,r,d)\underline{\mathrm{poly}(\log(1/\epsilon))})$$

近似誤差

$$\sup_{x\in[0,1]^d} \left| f(x) - \frac{p(x)}{q(x)} \right| \le \epsilon$$

- ReLU-DNNを有理関数で近似

k-層で各層のノード数mの任意のReLU-DNN fに対しては，次数と近似誤差が以下で抑えられる有理関数p/qが存在：

次数（分母qと分子pの次数の最大値）

$$\mathrm{O}(\log(k/\epsilon)^k m^k)$$

深さに対して指数的に増大

近似誤差

$$\sup_{x\in[0,1]^d} \left| f(x) - \frac{p(x)}{q(x)} \right| \le \epsilon$$

- ReLU-DNNを多項式で近似： $\Omega(\mathrm{poly}(1/\epsilon))$ の次数が必要

→有理関数に比べて表現力が低い

多層が得する例 (4): 有理関数の近似

参考

- 有理関数をReLU-DNNで近似

$$p : [0,1]^d \to [-1,+1],\ \ q : [0,1]^d \to [2^{-k}, +1]$$ **: r次多項式**

p/q をReLU-DNNで近似したい

あるReLU-DNN fが存在してノード数と近似誤差が次のように抑えられる：

ノード数

$$\mathrm{O}(\mathrm{poly}(k,r,d)\underline{\mathrm{poly}(\log(1/\epsilon))})$$

近似誤差

$$\sup_{x\in[0,1]^d}\left|f(x) - \frac{p(x)}{q(x)}\right| \le \epsilon$$

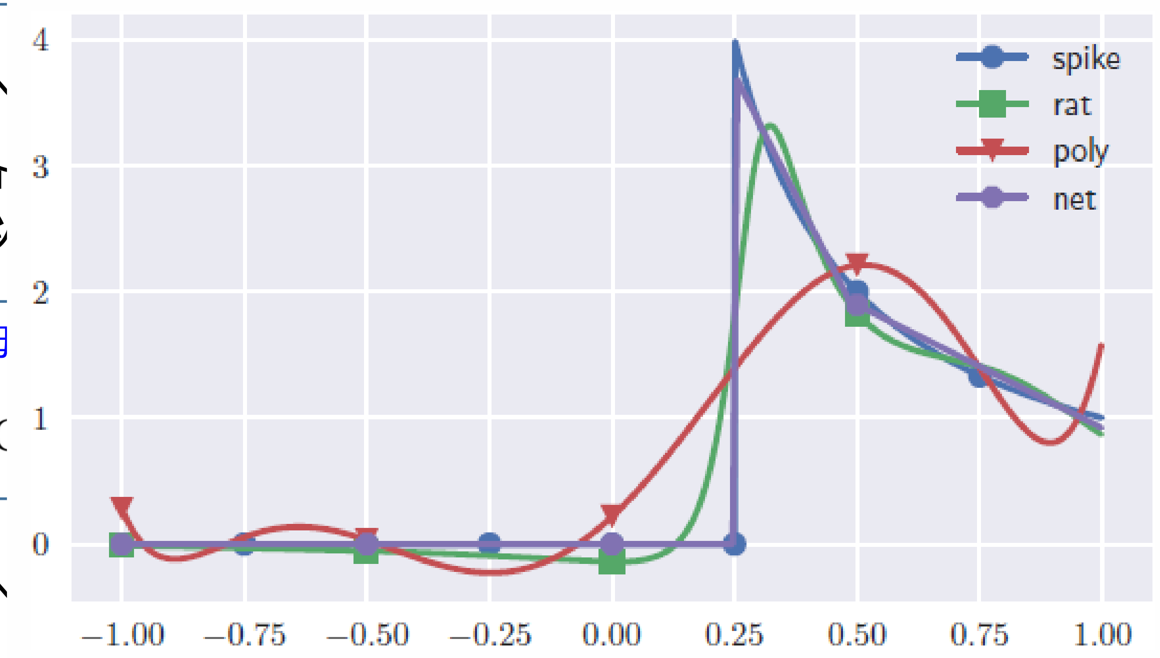
- ReLU-DNN

k-層で各 … 次

数と近似

次数（分母

$$\mathrm{O}(\mathrm{l}$$ … $$\frac{p(x)}{q(x)}\Big| \le \epsilon$$



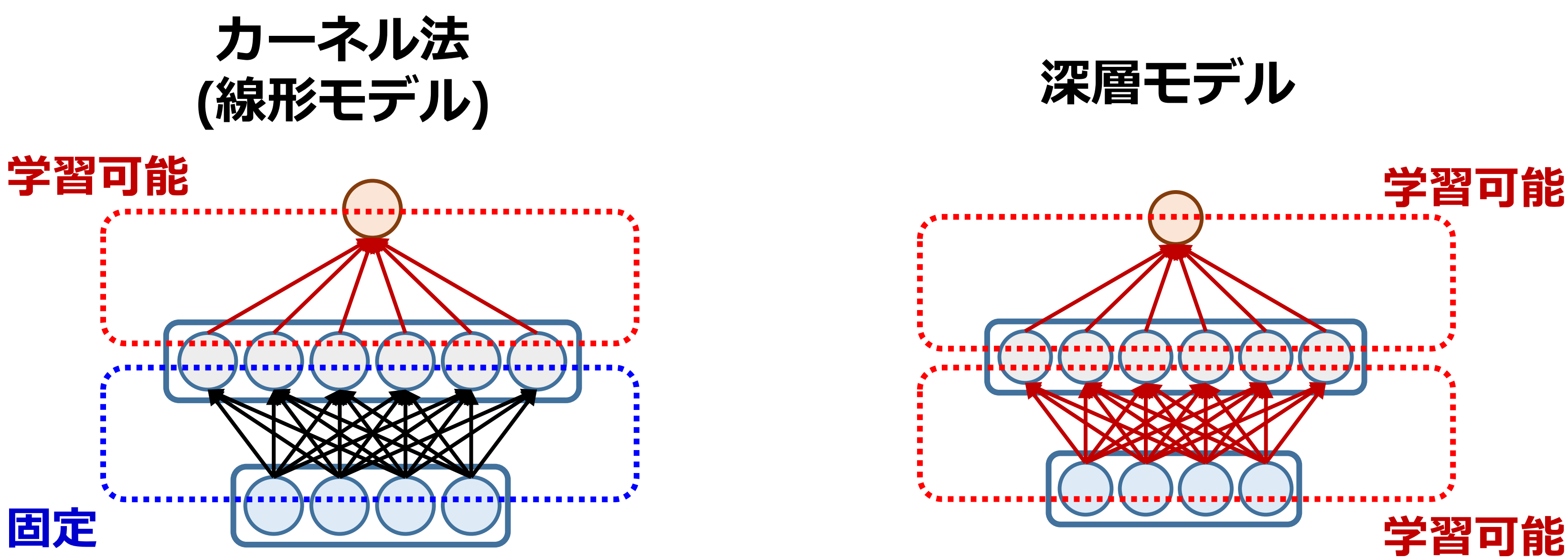
- ReLU-DNN … 要

表現力が低い

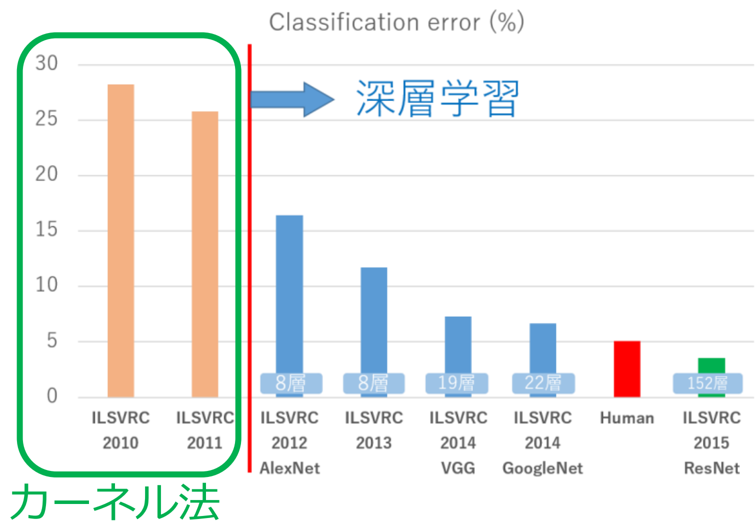
深層学習の適応的推定能力
-ノンパラメトリック回帰理論-

問題意識

ニューラルネットワークは「なぜ良いのか？」



解析対象
- 統計的効率性
- 最適化の効率性
(トレードオフの関係)



- **深層とカーネル法の性能を理論的に比較**
- **深層学習の最適化を理論的に解析**
- ➢ **深層学習の「適応能力」は本質的**

適応能力とは？

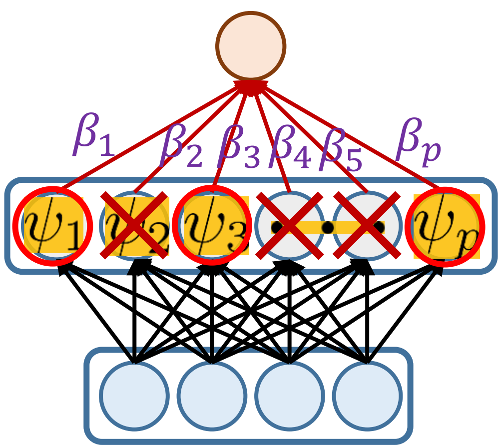
- 代表的な適応的推定法：**スパース推定**

やりたいこと：入出力関係を表す関数$y = f(x)$ を推定したい.

- 事前にたくさんの基底を用意
- それらの線形結合で真の関数を表現

$$\hat{f}(x) = \beta_1\psi_1(x) + \beta_2\psi_2(x) + \cdots + \beta_p\psi_p(x)$$

- 基底は冗長性があるように用意しておき,
  データに合わせて必要な基底のみ選択して使う.
  →スパース推定
- データに合わせて基底を選ぶので適応的と言う.



$\beta_2 = 0$
$\beta_4 = 0$
$\beta_5 = 0$

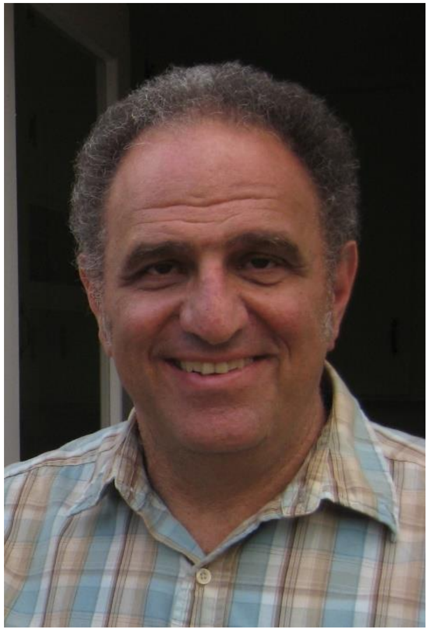
多くの係数が0
⇒ スパース

スパース推定の恩恵
- 推定精度が上がる.
- 解釈性が向上する.

David Donoho

Gauss prize (2018)
Sparse estimation,
wavelet-shrinkage,
compressive sensing，...

Robert Tibshirani



Lasso (1996)
→ L1正則化によるスパース推定

Wavelet shrinkage (Donoho&Johnstone, 1992)
→ Besov空間，スパース推定の適応能力

深層学習との関係

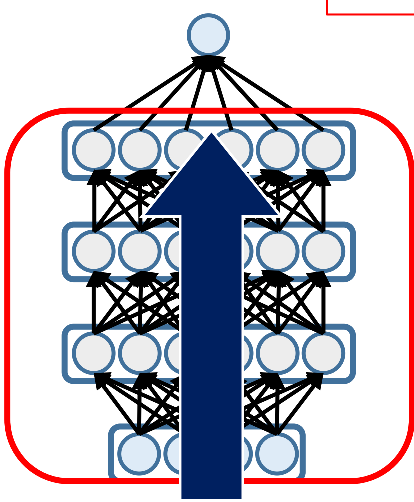
- スパース推定：
  基底を沢山用意してその中から**選択**.
- 深層学習：
  基底を事前に用意せず，データに合わせて**構築**.

$$\hat{f}(x) = \sum_{j=1}^{p} \beta_j \boxed{\psi_j}(x)$$

基底関数

非適応的手法：
- ➢ データによらず全部使う：リッジ回帰

適応的手法：
- ➢ 沢山の候補の中から選ぶ：スパース推定
- ➢ データに合わせて構築：深層学習



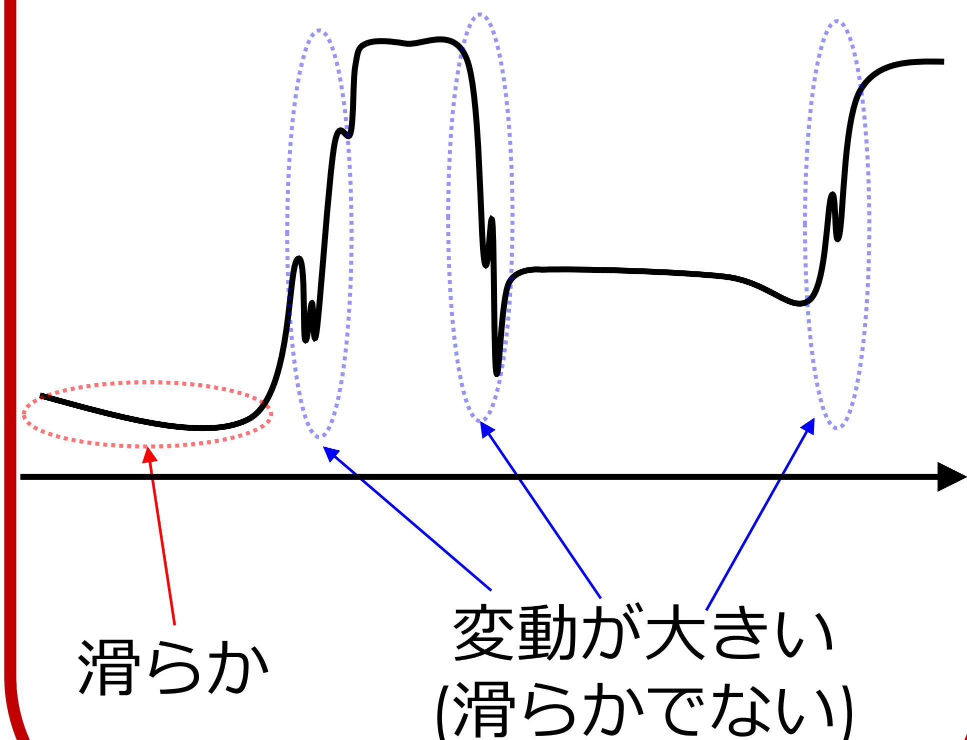
中間層で基底を構成

適応的推定能力

典型的な例

[Imaizumi&Fukumizu, 2019]
[Suzuki, 2019]

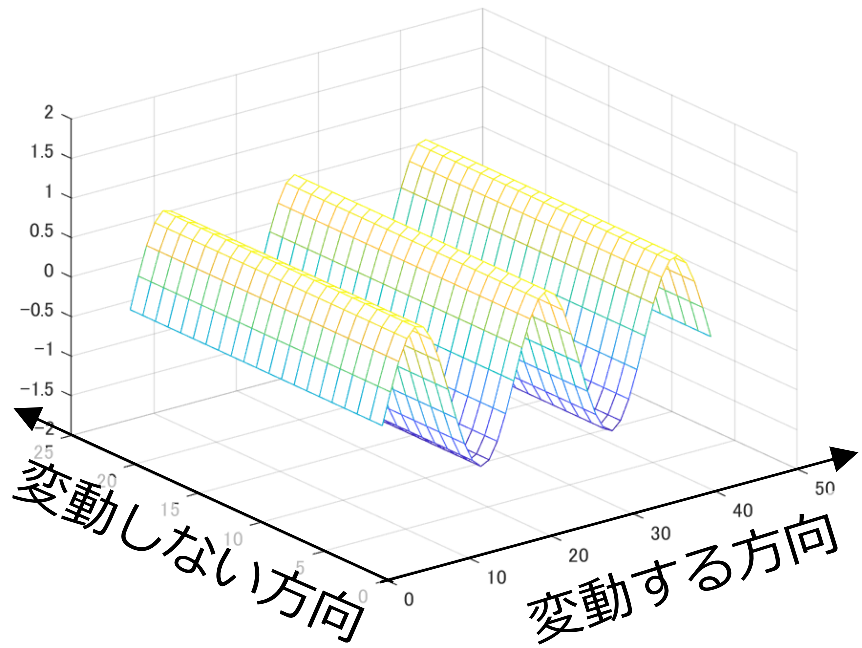
滑らかな部分と
そうでない部分が混在



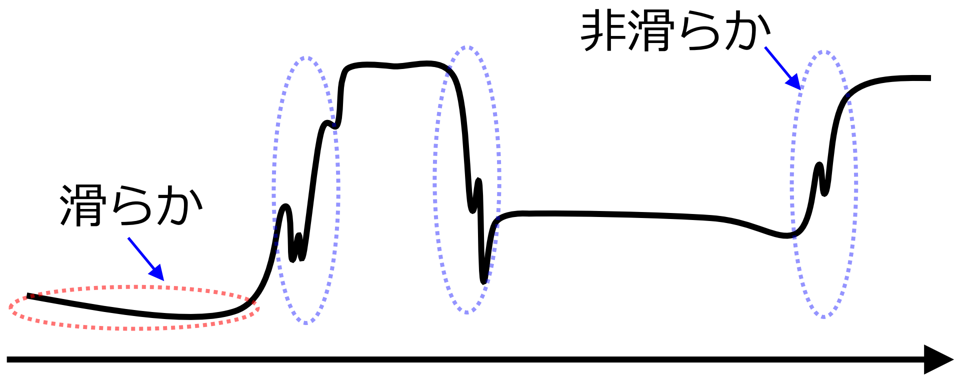
Besov空間

[Schmidt-Hieber, 2019] [Nakada&Imaizumi, 2019][Chen et al., 2019][Suzuki&Nitanda, 2019]

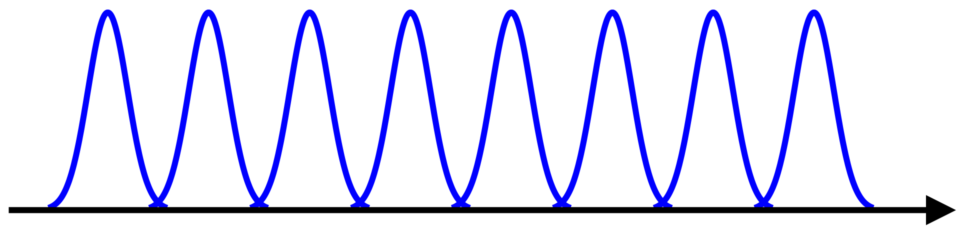
大きく変動する方向と
そうでない方向が混在



直感的説明

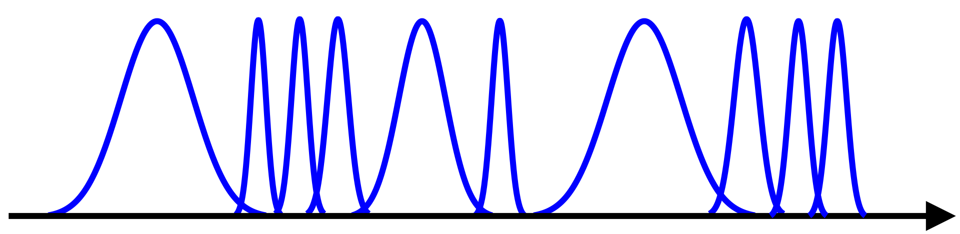


• ガウスカーネルを用いた関数近似 (カーネル法・非適応的)



同じ幅の基底の線形結合 → 効率悪い

• NNによる関数近似



場所によって解像度(幅)の違う基底を生成可能 → 効率良い

統計的推定理論による比較

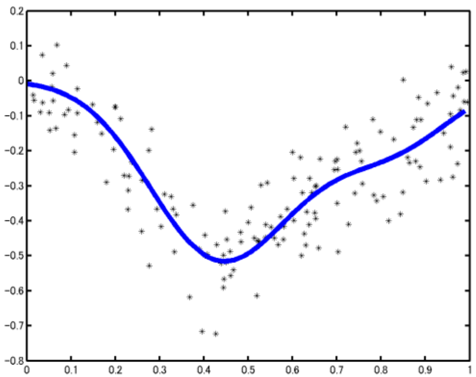
深層 vs 浅層 の統計理論
→「関数近似精度/推定精度」を比べてみる.

「多層」による特徴抽出と推定精度

ノンパラメトリック回帰の設定

$$y_i = f^{\circ}(x_i) + \xi_i \quad (i = 1, \dots, n)$$

$\xi_i \sim N(0, \sigma^2)$は観測誤差



推定誤差 (平均二乗誤差):

$$\mathbb{E}[\|\hat{f} - f^{\circ}\|^2_{L_2(P)}] < ?$$

※実はこれは二乗損失の平均余剰誤差になっている.

$$\mathbb{E}[L(\hat{\theta}) - \inf_f \mathbb{E}[\ell(Y, f(X))]]$$

Hölder, Sobolev, Besov空間

$\Omega = [0,1]^d \subset \mathbb{R}^d$

- **Hölder space** $(\mathcal{C}^\beta(\Omega))$

$$\|f\|_{\mathcal{C}^\beta} = \max_{|\alpha|\le m} \|\partial^\alpha f\|_\infty + \max_{|\alpha|=m} \sup_{x,y\in\Omega} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|^\beta}$$

直観的意味

滑らかさ

$$\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + \|D^s f\|_{L^p(\Omega)}$$

空間的一様性

- **Besov space** $(B^s_{p,q}(\Omega))$ $(0 < p, q \le \infty, 0 < s \le m)$

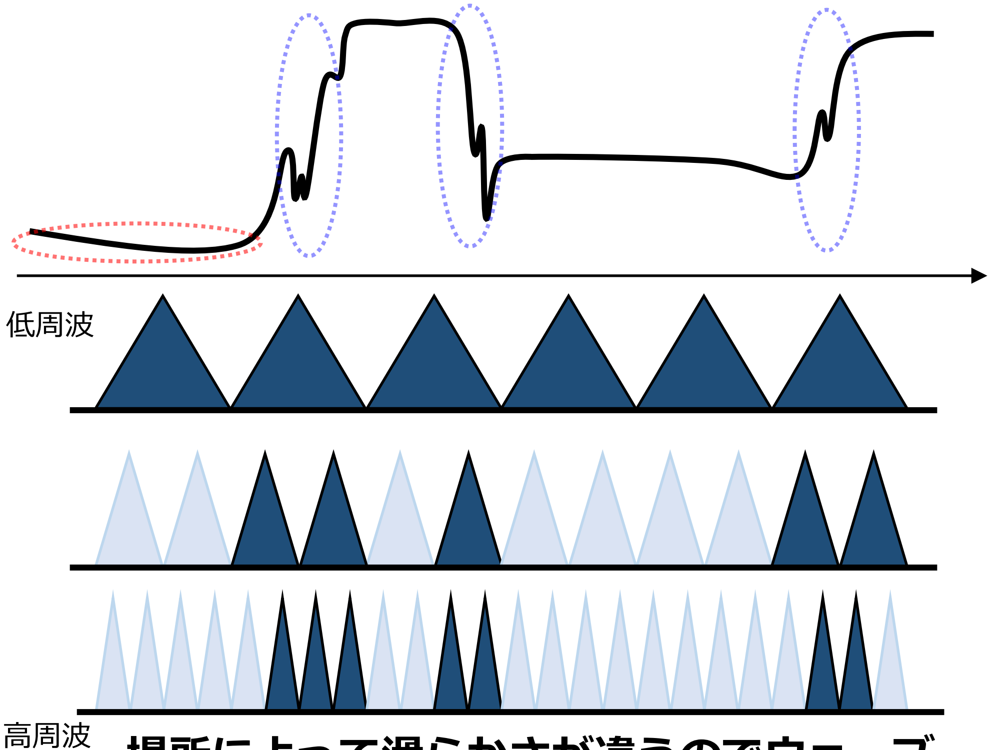
空間的非一様性

$$\omega_m(f,t)_p := \sup_{\|h\|\le t} \left\| \sum_{j=0}^m (-1)^{m-j} \binom{m}{j} f(\cdot + jh) \right\|_{L^p(\Omega)},$$

$$\|f\|_{B^s_{p,q}(\Omega)} = \|f\|_{L^p(\Omega)} + \left(\int_0^\infty [t^{-s} \omega_m(f,t)_p]^q \frac{\mathrm{d}t}{t} \right)^{1/q}.$$

滑らかさの度合い

Besov空間とスパース性との関係

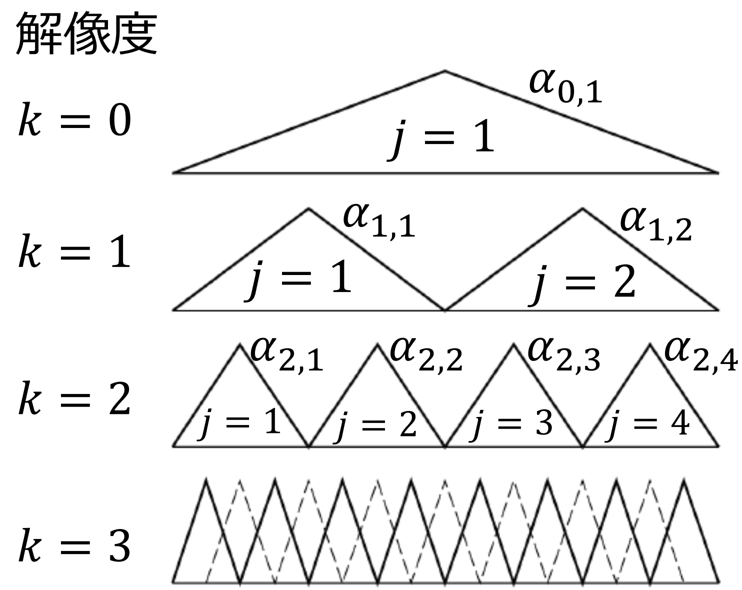


場所によって滑らかさが違うのでウェーブレット基底のスパースな線形結合が有効

$$f = \sum_{k \in \mathbb{N}_+} \alpha_k \phi_k$$

Wavelet基底による展開

Wavelet基底



Multiresolution expansion

$$\|f\|_{B^s_{p,q}} = \left(\sum_{k \in \mathbb{N}_+} |\alpha_k|^p \right)^{1/p} \quad (0 < p)$$

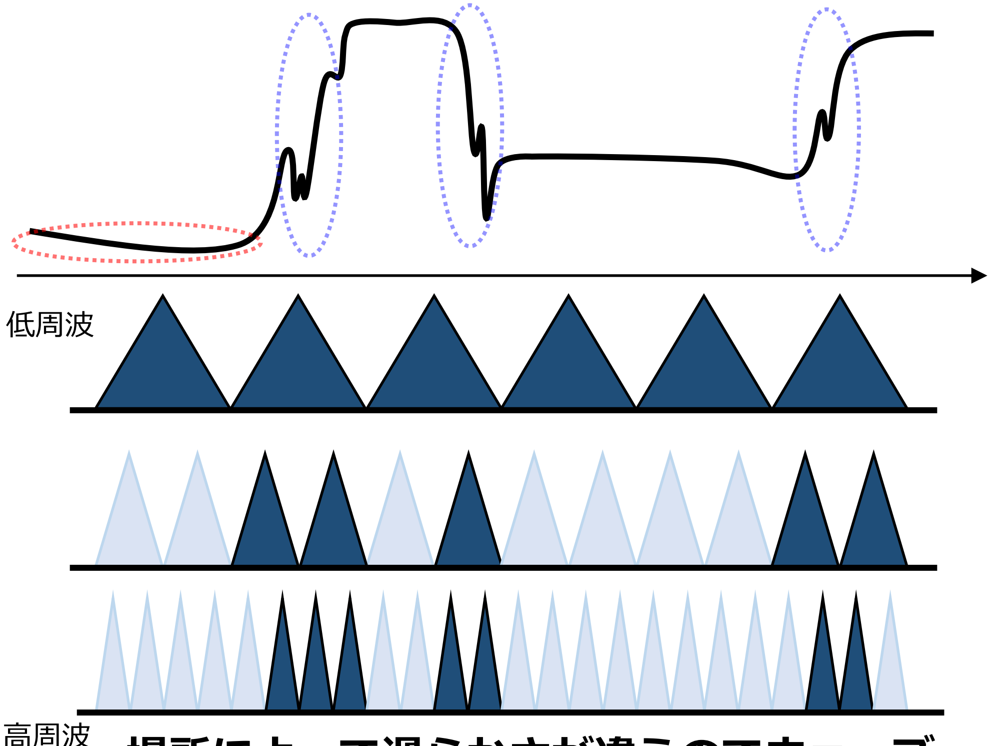
(informal)

$$\|f\|_{B^s_{p,q}} \simeq \left[\sum_{k=0}^{\infty} \{ 2^{sk} (2^{-kd} \sum_{j \in J(k)} |\alpha_{k,j}|^p)^{1/p} \}^q \right]^{1/q}$$

小さなp = スパースな係数

空間的な滑らかさの非一様性

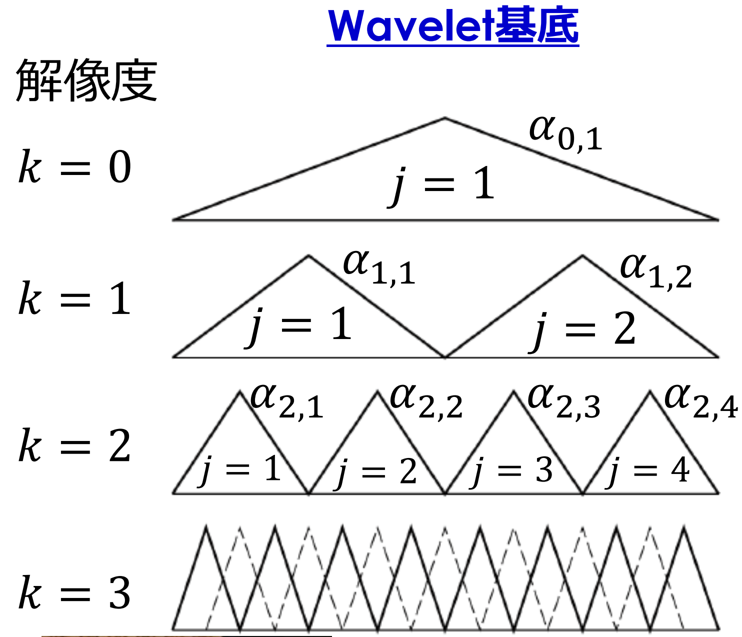
Besov空間とスパース性との関係



場所によって滑らかさが違うのでウェーブレット基底のスパースな線形結合が有効

$$f = \sum_{k \in \mathbb{N}_+} \alpha_k \phi_k$$

Wavelet基底による展開



solution expansion

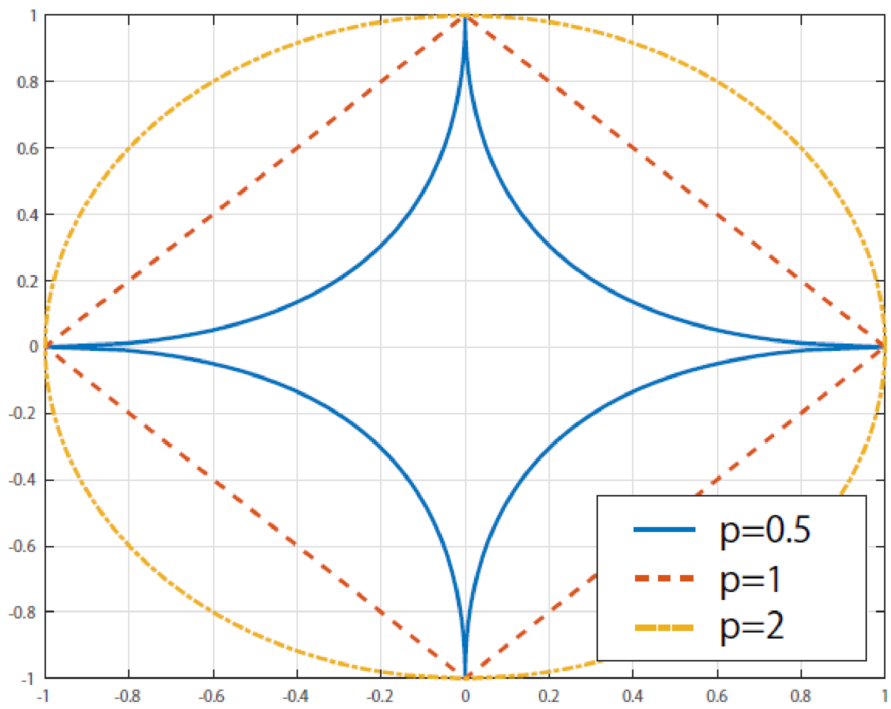
$\|f$ $\left. |\alpha_k|^p \right)^{1/p}$ $\mathbb{N}_+$ $(0 < p)$

$$\|f\|_{B^s_{p,q}} \simeq \left[\sum_{k=0}^{\infty} \{ 2^{sk} (2^{-kd} \sum_{j \in J(k)} |\alpha_{k,j}|^p)^{1/p} \}^q \right]^{1/q}$$

Wavelet-shrinkage

小さなp = スパースな係数

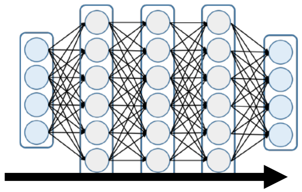
空間的な滑らかさの
非一様性



L^p ノルムのボール

「浅い」学習との比較

- 深層学習は場所によって解像度を変える適応力がある
  →学習効率が良い
- 浅い学習は様々な関数を表現できる基底を
  あらかじめ十分用意して“待ち構える”必要がある.
  →学習効率が悪い



仮定 $f^\circ \in B^s_{p,q}([0,1]^d)$: 真が“Besov空間”に入っている.

[Suzuki, ICLR2019]

線形推定量 (非適応的手法)

カーネルリッジ回帰等：

$$n^{-\frac{2s-2d(1/p-1/2)_+}{2s+d-2d(1/p-1/2)_+}}$$

最適ではない

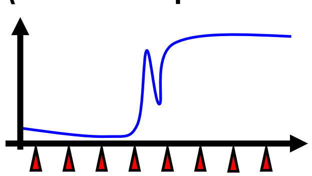
≫

深層学習

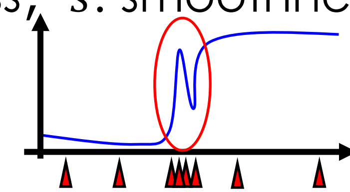
$$n^{-\frac{2s}{2s+d}}$$

最適

(n: sample size, p: uniformity of smoothness, s: smoothness)



一様な解像度



適応的解像度

平均二乗誤差 $\mathrm{E}\left[\|\hat{f}-f^*\|^2\right]$がサンプルサイズが増えるにつれ減少するレート

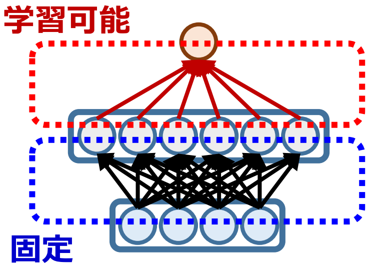
ミニマックス最適性の意味で理論上これ以上改善できない精度を達成できている.

- Wavelet shrinkageより弱い条件
- 基底を用意せず最適化するだけでOK

線形推定量

“浅い” 学習法

Kernel ridge regression: 正則化付き最小二乗推定量



$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^\infty} \sum_{i=1}^{n} (y_i - \beta^\top \psi(x_i))^2 + \lambda \beta^\top \beta$$

$\psi : \mathcal{X} \to \mathbb{R}^\infty$ (特徴マップ) 固定

$K_{X,X} = (\psi(x_i)^\top \psi(x_j))_{i,j=1}^{n,n}$ グラム行列 (カーネル関数)

$$\boxed{\hat{f}(x) = K_{x,X}(K_{X,X} + \lambda \mathrm{I})^{-1} \underline{Y}}$$

(see also [Imaizumi&Fukumizu, 2019])

線形推定量: 観測値$Y = (y_i)_{i=1}^n$に対して線形な推定量.

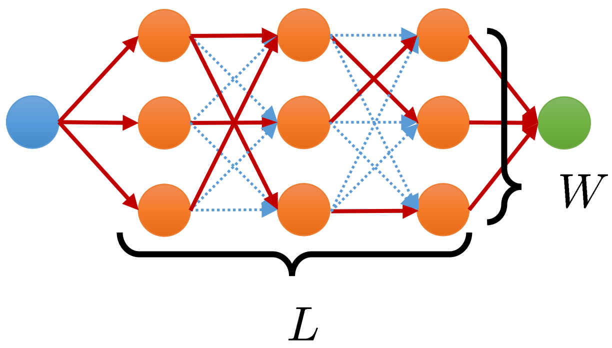
$$X_n = (x_1, \dots, x_n)$$

$$\hat{f}(x) = \sum_{i=1}^{n} \varphi_i(x; X_n) \underbrace{y_i}_{\text{線形}}$$

例

- Kernel ridge estimator
- Sieve estimator
- Nadaraya-Watson estimator
- k-NN estimator

証明: (0) ノーテーション



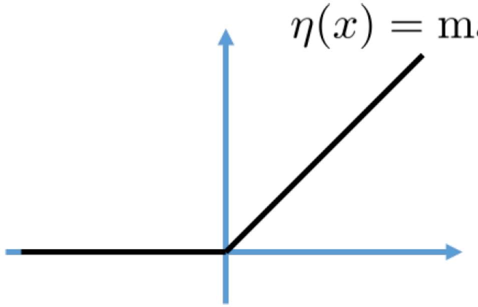
$$f(x) = (W^{(L)}\eta(\cdot) + b^{(L)}) \circ (W^{(L-1)}\eta(\cdot) + b^{(L-1)}) \circ \cdots \circ (W^{(1)}x + b^{(1)})$$

$\mathcal{F}(L, W, S, B)$
- 縦幅：L
- 横幅：W
- 枝の数：S
- 各パラメータの上限：B

の深層NNモデルの集合

- 活性化関数はReLUを仮定

$\eta(x) = \max\{x, 0\}$



証明: (1) 近似誤差の評価

- $0 < p,q,r \le \infty$と$0 < s < \infty$が以下を満たすとする:

$$s > d(1/p - 1/r)_+ \qquad (L^r\text{-可積分性})$$

- mを$s < \min\{m, m-1+1/p\}$を満たす整数とする.

深層ニューラルネットワークの近似誤差

ある自然数Nと用いて深さL, 横幅W, 枝の数S, ノルム上界Bを以下のように定める：

$$L = O(\log(N)), \qquad W = O(N),$$
$$S = O(N\log(N)), \qquad B = O(N^{(d/p-s)_+}),$$

すると，深層NNは以下の誤差でBesov空間の元を近似できる：

大体パラメータ数

$$\sup_{f^\circ \in U(B^s_{p,q}([0,1]^d))} \inf_{\check{f} \in \mathcal{F}(L,W,S,B)} \|f^\circ - \check{f}\|_{L^r([0,1]^d)} \lesssim N^{-s/d}.$$

Pinkus (1999), Mhaskar (1996): $p = r$かつ$1 \le p$, ReLU活性化関数ではない.
Petrushev (1998): $p = r = 2$, ReLU活性化関数ではない ($s \le k + 1 + (d-1)/2$).

証明: (1) 近似誤差の評価の導出

- **Step 1**: Besov空間の基底展開

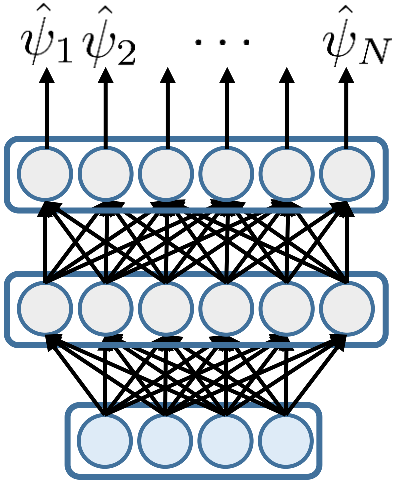
$$f^\circ \in \mathcal{F} \quad \Rightarrow \quad f^\circ(x) = \sum_{i=1}^{\infty} \alpha_i \psi_i(x)$$

$$f^\circ = \sum_{i=1}^{N} \alpha_i \psi_i + \underbrace{\cancel{\sum_{i=N+1}^{\infty} \alpha_i \psi_i}}_{\|\cdot\|_{L^r} \lesssim N^{-s/d}}$$

∵ B-Splineによる適応的近似

[DeVore & Popov, 1988; Dung, 2011]

- **Step 2**: 各基底をDNNで近似.



$\psi_i \simeq \hat{\psi}_i$: DNNによる近似.

$$\Rightarrow \quad \check{f} = \sum_{i=1}^{N} \alpha_i \hat{\psi}_i$$: 線形結合

- **Step 3**: 二つの評価を統合

$$\|f^\circ - \check{f}\|_{L^r} \leq \sum_{i=1}^{N} |\alpha_i| \underbrace{\|\psi_i - \hat{\psi}_i\|_{L^r}}_{\lesssim O(e^{-L})} + \underbrace{\|\textstyle\sum_{i=N+1}^{\infty} \alpha_i \psi_i\|_{L^r}}_{\lesssim N^{-s/d}}$$

証明: (2) バイアス-バリアンス分解

$$\mathrm{E}[\|f^\circ - \hat{f}\|_{L^2(P_X)}^2] \lesssim \underbrace{\frac{S[L\log(BW) + \log(Ln)]}{n}}_{\text{Variance}} + \underbrace{\inf_{f \in \mathcal{F}(L,W,S,B)} \|f - f^\circ\|_{L^2(P_X)}^2}_{\text{Bias}}$$

（局所Rademacher complexityを用いて証明）

古典的なノンパラ回帰の方法でOK．DNNに関する評価は[Schmidt-Hieber, 2019; Hayakawa&Suzuki,2020]

深さ　横幅　スパース性 (非零パラメータ数)　各パラメータの絶対値の上界

$$L = O(\log(N)), W = O(N), S = O(N\log(N)), B = O(N^{(d/p-s)_+})$$

なら

$$\text{Bias} = N^{-s/d} \qquad \text{Variance} = \frac{N\log(N)^3}{n}$$

⇒ バイアスとバリアンスのトレードオフをバランスすればよい.

証明: (3) 推定精度の導出

- 最小二乗解 (訓練誤差最小化)

$$\hat{f} = \mathop{\arg\min}_{\bar{f}: f \in \mathcal{F}(L,W,S,B)} \sum_{i=1}^{n} (y_i - \bar{f}(x_i))^2$$

ただし, $\bar{f} = \min\{\max\{f, -F\}, F\}$ (clipping).

定理 (推定精度)

$\|f^{\mathrm{o}}\|_{B^s_{p,q}} \leq 1$, $\|f^{\mathrm{o}}\|_\infty \leq 1$ かつ $0 < p, q \leq \infty$, $s > d(1/p - 1/2)_+$ のとき,
$N \asymp n^{\frac{d}{2s+d}}$ とすることで,

$$\|f^{\mathrm{o}} - \hat{f}\|^2_{L^2(P_X)} \leq n^{-\frac{2s}{2s+d}} \log(n)^3.$$

$p = q = \infty$のとき, Schmidt-Hieber (2017) に帰着.

スパース推定との繋がり

Wavelet shrinkage

Lasso

圧縮センシング
- Donoho
- Candes&Tao

機械学習の興隆

スパース推定の流行

ビッグデータブーム

ILSVRC Supervision優勝

第三次AIブーム
- **深層学習**
- **産業応用**

深層学習の理論
- 適応的推定理論
- Besov空間を用いた解析

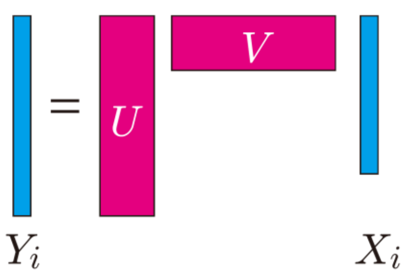
1990 1995 2000 2006 2010 2011 2012 2019

他にも様々な理論が

- 真の関数f°の形状によって深層が有利になる

縮小ランク回帰

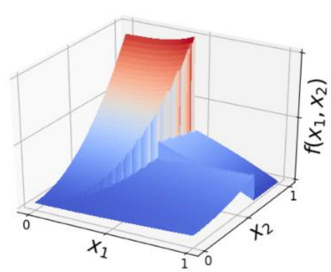
特徴空間の次元が低い状況は深層学習が得意



区分滑らかな関数

[Imaizumi&Fukumizu, 2019]

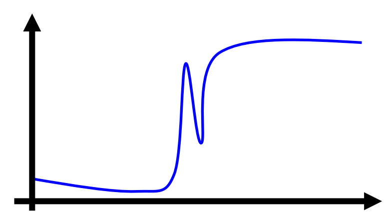
不連続な関数の推定は深層学習が得意



Besov空間

[Suzuki, 2019]

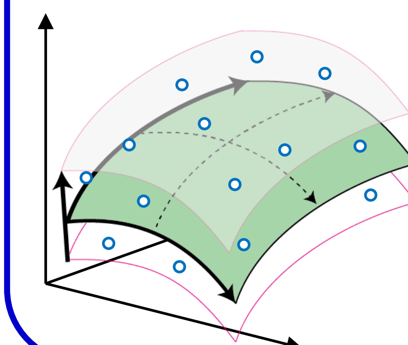
滑らかさが非一様な関数の推定は深層学習が得意



低次元データ

[Schmidt-Hieber, 2019] [Nakada&Imaizumi, 2019] [Chen et al., 2019] [Suzuki&Nitanda, 2019]

データが低次元部分空間上に分布していたら深層学習が有利



推定精度	縮小ランク回帰	区分滑らかな関数	Besov空間	低次元データ
深層	$\frac{r(M+N)}{n}$	$n^{-\frac{2s}{2s+d}} \vee n^{-\frac{\alpha}{\alpha+D-1}}$	$n^{-\frac{2s}{2s+d}}$	$n^{-\frac{2s}{2s+D}}$
カーネル	$\frac{MN}{n}$	$\frac{1}{\sqrt{n}}$	$n^{-\frac{2s-2d(1/p-1/2)_+}{2s+d-2d(1/p-1/2)_+}}$	$n^{-\frac{2(s-D/p+d/2)}{2(s-D/p+d/2)+d}} \vee n^{-\frac{2s}{2s+D}}$

数学的に一般化

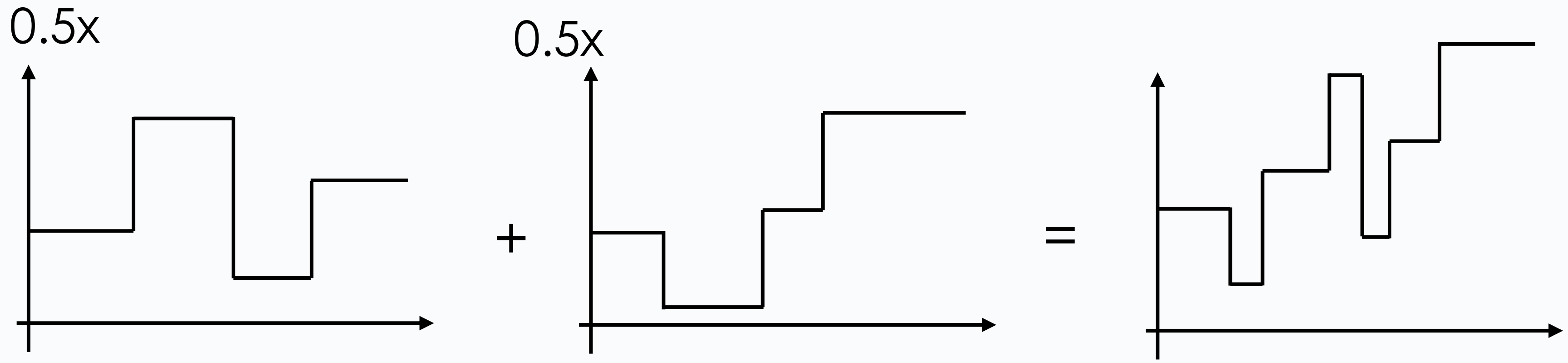
「滑らかさの非一様性」「不連続性」「データの低次元性」
凸結合を取って崩れる性質をもった関数の学習は深層学習が強い

→ 様々な性質を“凸性”で統一的に説明できる

例：ジャンプが３か所の区分定数関数

深層:$1/n$
カーネル: $1/\sqrt{n}$

$\sqrt{n}$倍の違い



ジャンプ**3**か所　　ジャンプ**3**か所　　ジャンプ**6**か所

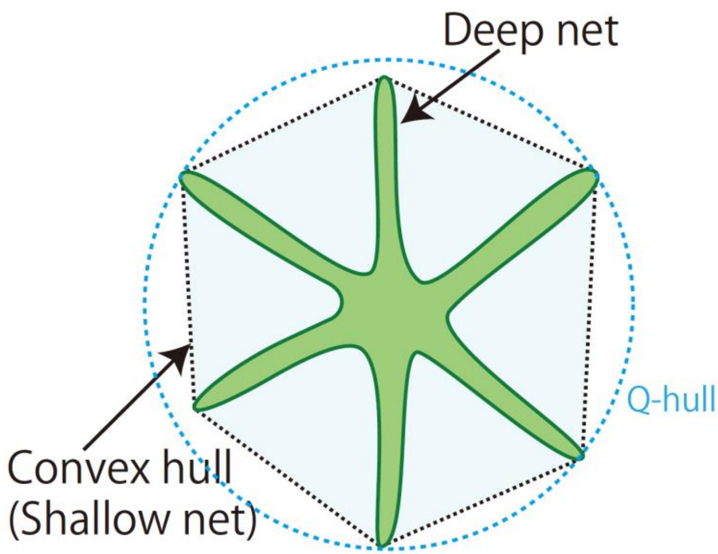
→ **「スパース性」**と**「非凸性」**

[Satoshi Hayakawa and Taiji Suzuki: On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. Neural Networks, Vol.123, pp. 343—361, 2020]

線形推定量の最悪誤差

線形推定量：$\hat{f}(x) = \sum_{i=1}^n y_i \varphi_i(x_1,\dots,x_n;x) + b$ と書ける**任意の推定量**

例: カーネルリッジ回帰 $\hat{f}(x) = K_{x,X}(K_{X,X} + \lambda \mathrm{I})^{-1} Y$ (“浅い”学習法とみなす)



- 凸包を取って大きく膨らむような関数クラスにおいては深層NNを使うメリットがある.
- 特徴量(基底関数)を適応的に作ることで，ターゲットを「狙い撃ち」して近似.
- 一方，線形推定量は広めにモデルを取って「待ち構えて」いる必要がある.

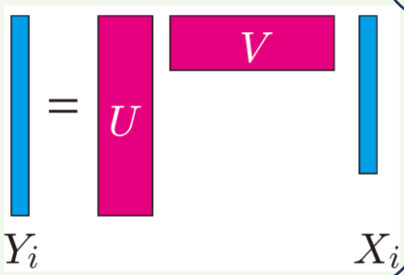
$$\inf_{\hat{f}:\mathrm{Linear}} \sup_{f^\circ \in \mathcal{F}} \mathrm{E}[\|\hat{f} - f^\circ\|_{L_2(P)}^2] = \inf_{\hat{f}:\mathrm{Linear}} \sup_{f^\circ \in \mathrm{conv}(\mathcal{F})} \mathrm{E}[\|\hat{f} - f^\circ\|_{L_2(P)}^2]$$

さらに条件を仮定すれば「Q-hull」まで拡張できる.

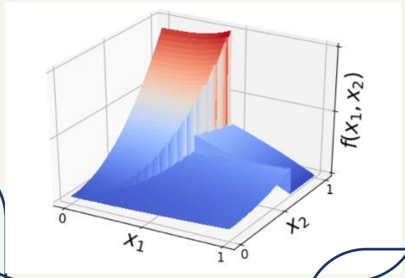
[Hayakawa&Suzuki: 2019][Donoho & Johnstone, 1994]

数学的一般化

縮小ランク回帰

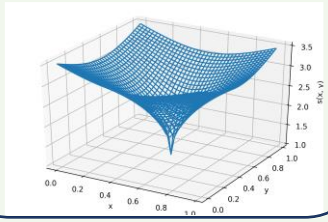


区分滑らかな関数

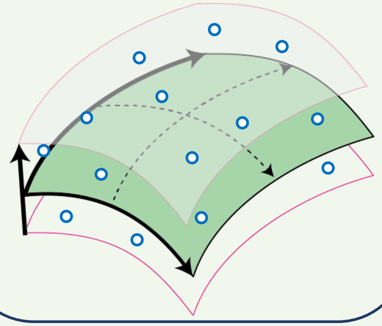


Besov空間

変動指数Besov空間



低次元データ



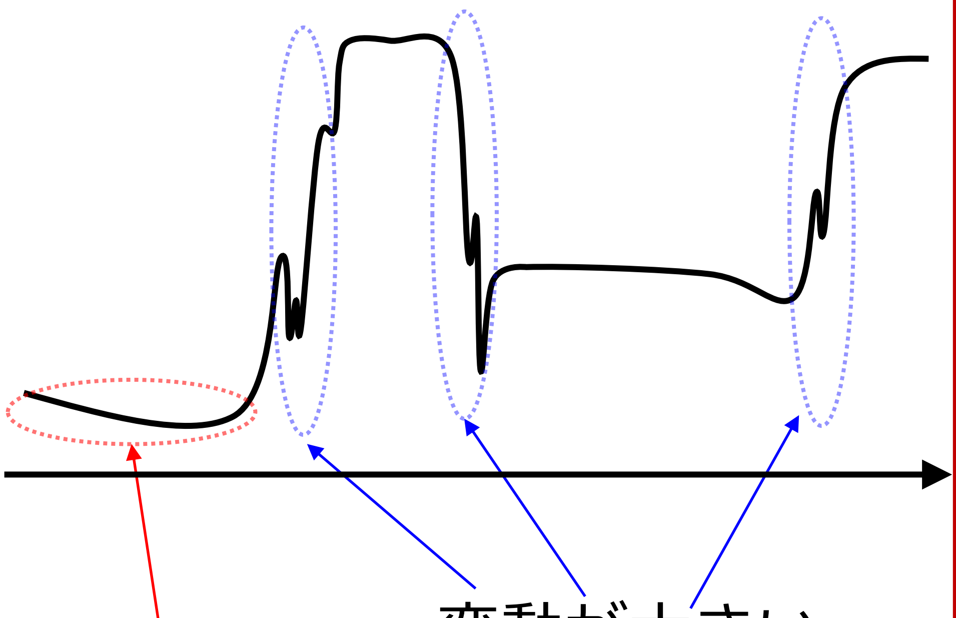
非凸性
スパース性

適応的推定法

- (勾配)ブースティング
- スパース推定
- 深層学習

典型的な例

滑らかな部分と
そうでない部分が混在

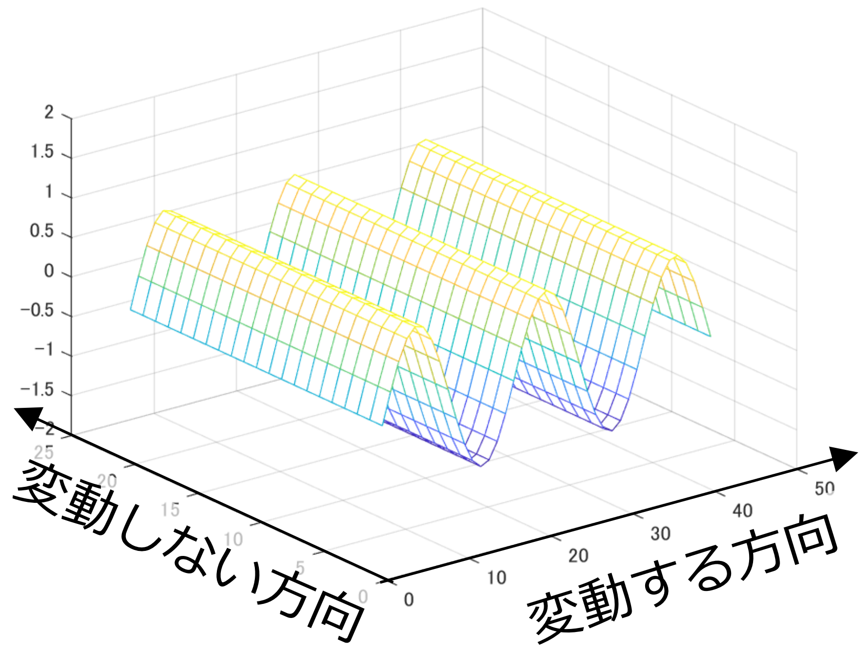


滑らか

変動が大きい
(滑らかでない)

Besov空間

大きく変動する方向と
そうでない方向が混在



高次元データの特徴抽出

深層ニューラルネットは特徴抽出に長けている.

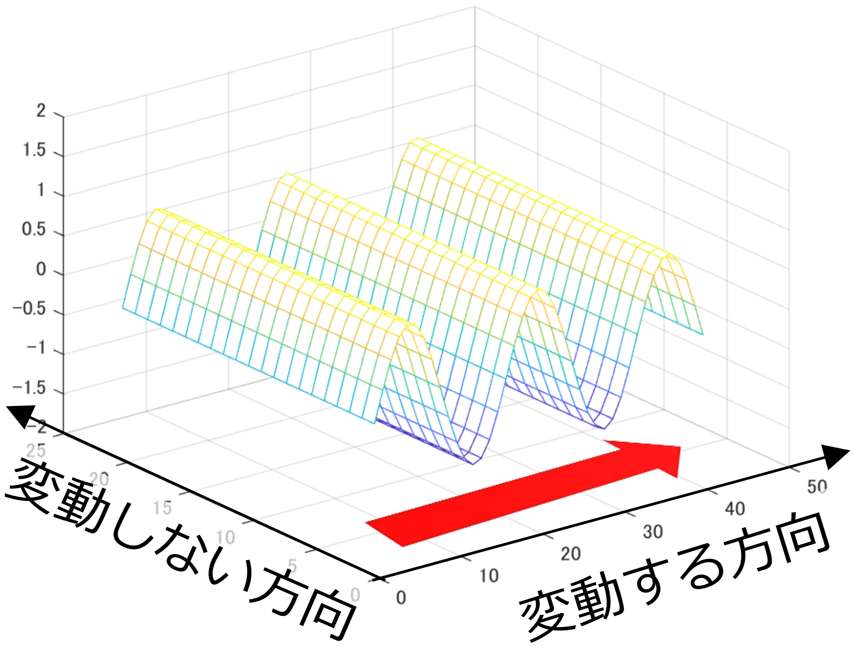
➢ 合成関数を推定できる.

(合成される各関数h_ℓが“単純”であれば層に分割することで得をする)

$$f^\circ(x) = h_H \circ \underbrace{h_{H-1} \circ \cdots \circ h_1}_{\text{特徴抽出}}(x)$$

（無駄な情報を削除）

典型例



- **ガウスカーネルを用いた関数近似**

  変動する方向を特定できない
  →次元の呪い

- **NNによる関数近似**

  変動する方向を表す特徴量を中間層で抽出
  →次元の呪いを回避

次元の呪い

推定誤差のバウンド：

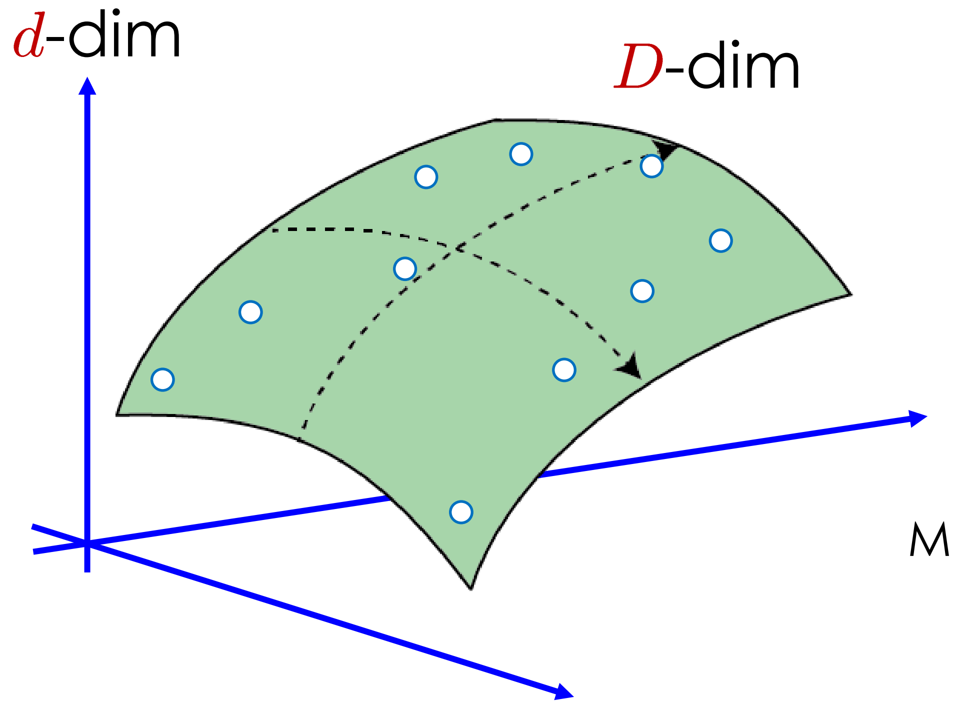
$$n^{-\frac{2s}{2s+d}}$$

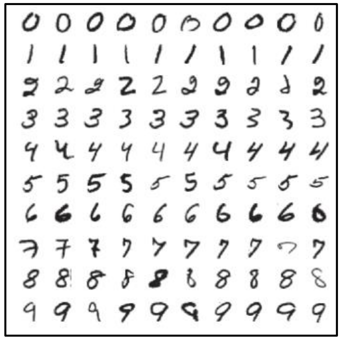
近似誤差のバウンド:

$$N^{-\frac{s}{d}}$$

→ **<u>次元の呪い</u>**

アプローチ (1): 多様体回帰





MNIST: 784 dim/
13.4 intrinsic-dim
[Facco et al. 2017]

- **Classic nonparametric method:** Bickel & Li (2007); Yang & Tokdar (2015); Yang & Dunson (2016).
- **Deep learning:** Nakada & Imaizumi (2019); Schmidt-Hieber (2019); Bauer & Kohler (2019); Chen et al. (2019a,b); Liu et al. (2021).

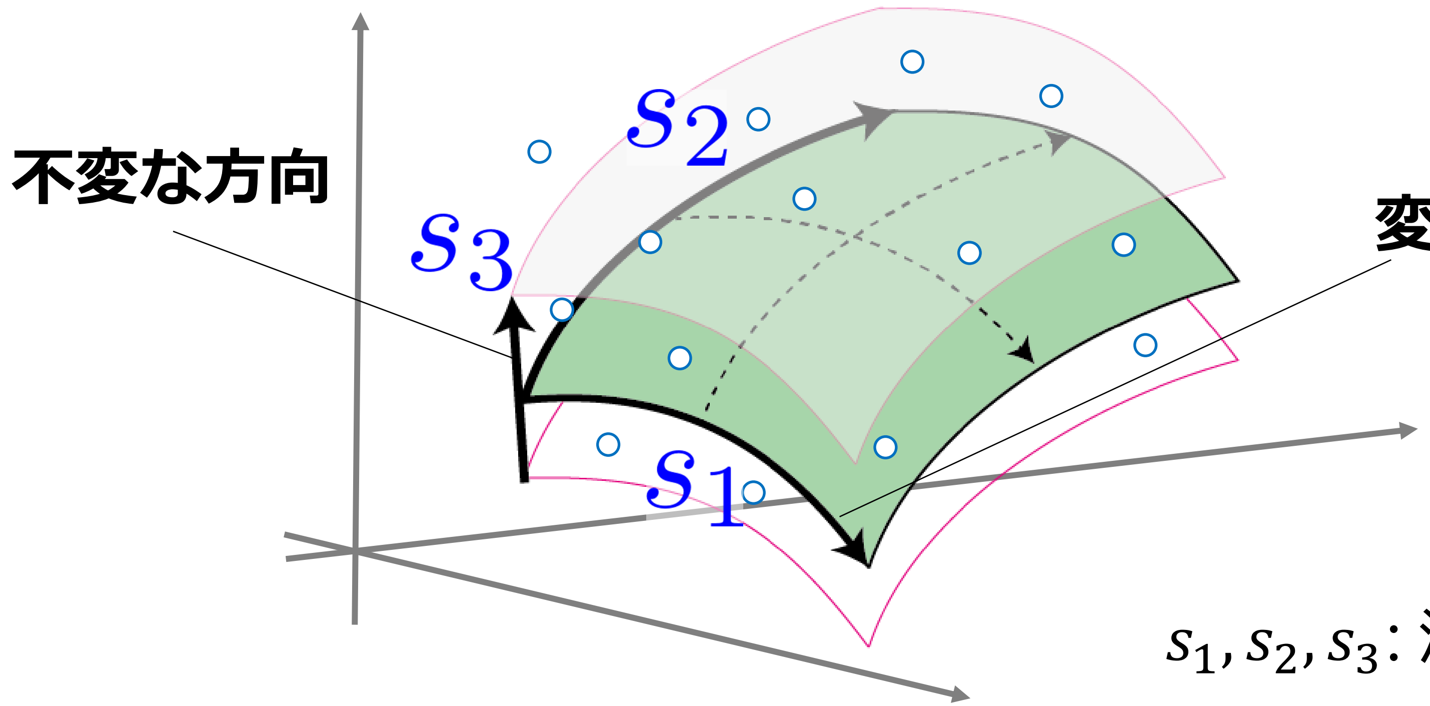
$$n^{-\frac{2s}{2s+D}}$$

データが低次元多様体*に分布していれば次元の呪いを回避できる！

* Nakada&Imaizumi (2019) では非滑らかな低次元構造も許容 (Hausdorff次元が小さい)

アプローチ (2): 関数の平滑性の非等方性

[Suzuki&Nitanda: Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. NeurIPS2021.]



s_1, s_2, s_3: 滑らかさ

(非平滑) $s_1, s_2 \ll s_3$ (平滑)

データが低次元多様体からはみ出る場合：

- 真の関数の滑らかさが方向に依存.
- 多様体に直交する方向にはほぼ定数 (滑らかさ大)

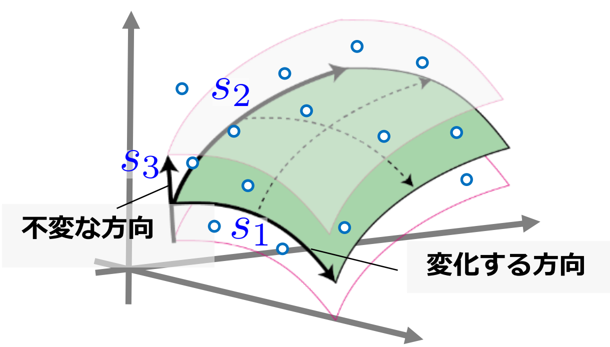
(超)高次元入力NNの学習理論

Suzuki&Nitanda: Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. NeurIPS2021, spotlight.

- **真の関数が方向によって異なる滑らかさを持つ状況ではDNNは重要な方向を見つけ，次元の呪いを回避する．**
- **一方で，浅い学習法は次元の呪いを受ける．**

関連研究：
- 教師生徒設定における大域的最適化と次元の呪いの回避
  Suzuki and Akiyama: ICLR2021, spotlight.
- 深層学習の浅層学習への優位性:
  Hayakawa and Suzuki: Neural Networks 2020, 日本神経回路学会論文賞.

s_1, s_2, s_3: 滑らかさ



真の関数の滑らかさが方向によって大きく異なる状況滑らかでない方向が少なければ次元の呪いを回避できる．
→「**非等方的Besov空間**」を用いた理論．

真の関数のモデル：
非等方的Besov空間の元の合成関数

$$f^\circ(x) = h_H \circ \cdots \circ h_1(x)$$

$h_\ell : \mathbb{R}^{m_\ell} \to \mathbb{R}^{m_{\ell+1}}$: **非等方的Besov空間** $(B^{s^{(\ell)}}_{p,q})$.

➢ 滑らかさが方向によって異なる関数空間
➢ 合成することで様々な形状を実現
(例：多様体上の関数: 一層目で座標を抽出，二層目がその座標上の関数)

Def. (非等方的Besov空間)

$$\Delta_h^r(f)(x) := \Delta_h^{r-1}(f)(x+h) - \Delta_h^{r-1}(f)(x),$$
$$\Delta_h^0(f)(x) := f(x) \qquad (h \in \mathbb{R}^d)$$

$$w_{r,p}(f,t) = \sup_{h \in \mathbb{R}^d : |h_i| \le t_i} \|\Delta_h^r(f)\|_p$$

$$s = (s_1, \dots, s_d) \in \mathbb{R}^d_{++}$$
$$|f|_{B^s_{p,q}} := \begin{cases} \left(\sum_{k=0}^{\infty} [2^k w_{r,p}(f, (2^{-k/s_1}, \dots, 2^{-k/s_d}))]^q\right)^{1/q} & (q < \infty), \\ \sup_{k \ge 0} 2^k w_{r,p}(f, (2^{-k/s_1}, \dots, 2^{-k/s_d})) & (q = \infty). \end{cases}$$
$$\|f\|_{B^s_{p,q}} := \|f\|_p + |f|_{B^s_{p,q}}$$

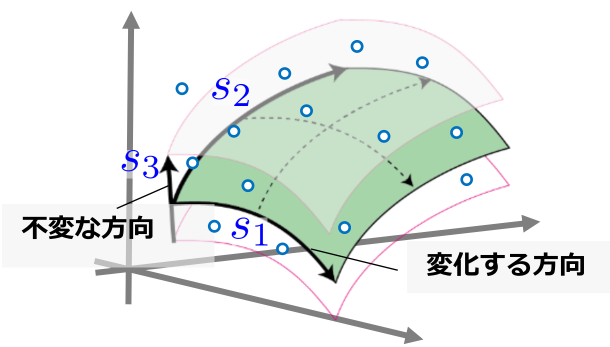
(超)高次元入力NNの学習理論

Suzuki&Nitanda: Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. NeurIPS2021, spotlight.

- **真の関数が方向によって異なる滑らかさを持つ状況ではDNNは重要な方向を見つけ，次元の呪いを回避する．**
- **一方で，浅い学習法は次元の呪いを受ける．**

関連研究：
- 教師生徒設定における大域的最適化と次元の呪いの回避
  Suzuki and Akiyama: ICLR2021, spotlight.
- 深層学習の浅層学習への優位性:
  Hayakawa and Suzuki: Neural Networks 2020, 日本神経回路学会論文賞.

s_1, s_2, s_3: 滑らかさ



真の関数の滑らかさが方向によって大きく異なる状況滑らかでない方向が少なければ次元の呪いを回避できる．
→ 「**非等方的Besov空間**」を用いた理論．

真の関数のモデル：
非等方的Besov空間の元の合成関数

$$f^\circ(x) = h_H \circ \cdots \circ h_1(x)$$

$h_\ell : \mathbb{R}^{m_\ell} \to \mathbb{R}^{m_{\ell+1}}$: **非等方的Besov空間** ($B^{s^{(\ell)}}_{p,q}$)

- ➢ 滑らかさが方向によって異なる関数空間
- ➢ 合成することで様々な形状を実現

(例：多様体上の関数: 一層目で座標を抽出，二層目がその座標上の関数)

Def. (非等方的Besov空間)

直感

$$\|f\|_{B^{s}_{p,q}} = \|f\|_{L^p} + \sum_{i=1}^{d} \left\| \frac{\partial^{s_i} f}{\partial x_i^{s_i}} \right\|_{L^p}$$

$$|f|_{B^s_{p,q}} := \begin{cases} \left(\sum_{k=0}^{\infty} [2^k w_{r,p}(f,(2^{-k/s_1},\dots,2^{-k/s_d}))]^q\right)^{1/q} & (q<\infty), \\ \sup_{k\geq 0} 2^k w_{r,p}(f,(2^{-k/s_1},\dots,2^{-k/s_d})) & (q=\infty). \end{cases}$$

$$\|f\|_{B^s_{p,q}} := \|f\|_p + |f|_{B^s_{p,q}}$$

推定誤差の評価

$$f^\circ(x) = h_H \circ \cdots \circ h_1(x) \qquad h_\ell \in B_{p,q}^{(s_1^{(\ell)},\dots,s_{m_\ell}^{(\ell)})}([0,1]^{m_\ell}) \quad h_\ell : [0,1]^{m_\ell} \to [0,1]^{m_{\ell+1}}$$

$$\hat{f} = \mathop{\arg\min}_{f:\ \text{deep neural-net}} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

(最小二乗推定量)

※今回は最適化手法に関しては議論せず，最適化はできるものと仮定する．

主結果

Let $\tilde{s}^{(\ell)} := \left(\frac{1}{s_1^{(\ell)}} + \cdots + \frac{1}{s_{m_\ell}^{(\ell)}}\right)^{-1}$, $\quad \tilde{s}^{*(\ell)} := \tilde{s}^{(\ell)} \prod_{k=\ell+1}^{H} [(\min_j s_j^{(k)} - 1/p) \wedge 1]$

$$\boxed{\mathrm{E}[\|\hat{f} - f^\circ\|_{L^2(P_X)}^2] \lesssim \max_{\ell \in [H]} n^{-\frac{2\tilde{s}^{*(\ell)}}{2\tilde{s}^{*(\ell)}+1}} \log(n)^3}$$

各方向への滑らかさの調和平均が収束レートを決める．

例: $H = 1$の時

$$n^{-\frac{2\tilde{s}}{2\tilde{s}+1}}$$

$$\tilde{s} = (s_1^{-1} + \cdots + s_d^{-1})^{-1}$$

少ない数の方向においてs_iが小さく(滑らかでない)，その他の方向にはs_iが大きい(滑らか)であるとき，次元の呪いを回避できる．

浅い学習方法との比較 (informal)：

深層

$$n^{-\frac{2\tilde{s}}{2\tilde{s}+1}}$$

$$\tilde{s} = (s_1^{-1} + s_2^{-1} + s_3^{-1})^{-1}$$

浅層

$$n^{-\frac{2s_1}{2s_1+d}}$$

(次元の呪いを受ける)

- **特徴抽出能力の重要性を理論的に正当化**
- **浅い学習方法は一番滑らかでない方向の滑らかさ(s_1)が支配的で，次元の呪いを受ける．**
- **証明には「凸法の議論」を用いる．**

線形推定量との比較 (より正確なステートメント)

$$\mathcal{F} = \{ f^\circ = g(Wx+b) \mid g \in U(B^s_{p,q}([0,1]^D)),\ W \in \mathbb{R}^{D\times d},\ b \in \mathbb{R}^D \}$$

(s.t. $Wx + b \in [0,1]^D$ for any $x \in [0,1]^d$)

f°はD-次元部分空間にのみ依存

If $s > \frac{D}{d-D}\left(\frac{d}{2} - \frac{D}{p} + c\right)$

(非適応的)

深層

$$n^{-\frac{2s}{2s+D}}$$

($n^{-\frac{2s}{2s+1}}$ when $D = 1$)

$\ll$

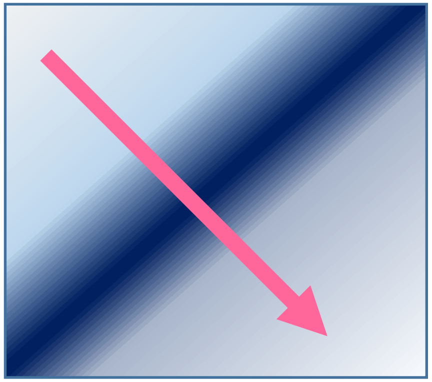
線形推定量

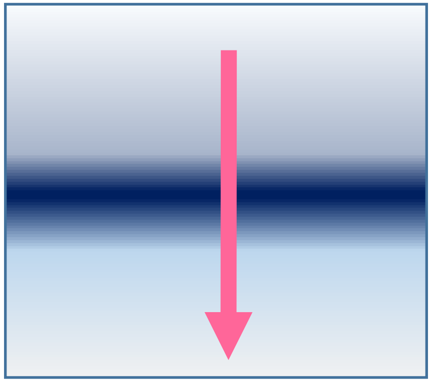
$$n^{-\frac{2(s-D/p+d/2+c)}{2(s-D/p+d/2+c)+d}}$$

$c = 1$ if $D < d/2$, $c = 0$ if $D \geq d/2$.

($n^{-\frac{2s+d}{2s+2d}}$ when $D = 1$ and $p = 1$)

深層にすることで次元の呪いを回避できている.





無限次元入力NN

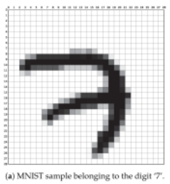
無限次元入力

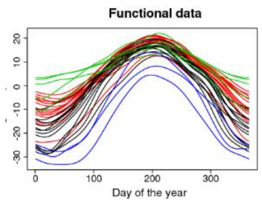
(画像，音声信号, 自然言語,...)

[Okumoto&Suzuki: Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. ICLR2022.]

無限 (高) 次元データ

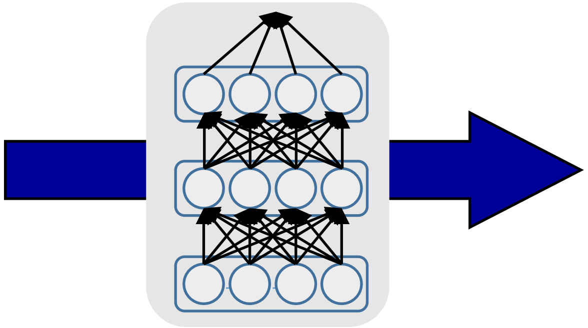
画像データ　関数データ





- 音声
- 文章

⋮



出力
(実数)

[Ramsay, J., Hooker, Giles, & Graves, Spencer. (2009). Functional data analysis with R and MATLAB (Use R!). Dordrecht: Springer.]

さきほどの議論は入力の有限次元性を使っていた.
→ 実は**無限次元**まで拡張できる.

無限次元入力に対する深層学習の統計理論

- 次元に依存しないバウンド (有限次元の拡張)
- 畳み込みNNによる特徴量の抽出

$$\mathbb{E}[\|\hat{f} - f^\circ\|_{L_2(P_X)}^2] \lesssim n^{-\frac{2(\tilde{a}-v)}{2(\tilde{a}-v)+1}} (\log n)^{\frac{2}{q}+2} \max\{(\log n)^{4/q}, (\log n)^4\}$$

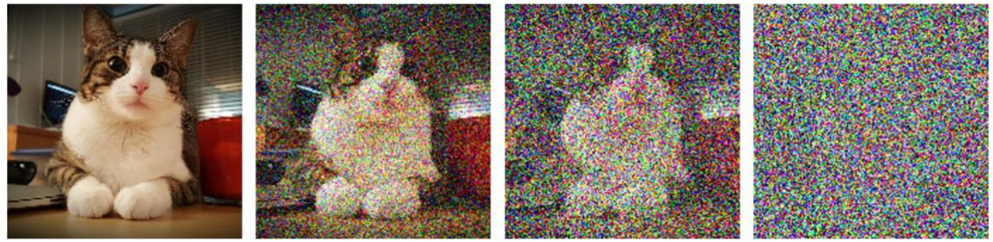
拡散モデルの統計理論

[Kazusato Oko, Shunta Akiyama, Taiji Suzuki: Diffusion Models are Minimax Optimal Distribution Estimators. ICML2023]

Stable diffusion, 2022.

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Forward process



Backward process

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$(Y_t \sim X_{\overline{T}-t})$

経験スコアマッチング推定量:

$$\hat{s} = \mathop{\arg\min}_{s \in \mathrm{DNN}} \frac{1}{n}\sum_{i=1}^{n}\int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t|X_0=x_{0,i}}\left[\|s(X_t,t) - \nabla \log p_t(X_t|x_{0,i})\|^2\right]\mathrm{d}t$$

定理

Let $\hat{Y}$ be the r.v. generated by the backward process w.r.t. $\hat{s}$, then

$$\mathbb{E}_{D_n}\left[\mathrm{TV}(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}}\log^9(n),$$ (s: 密度関数の滑らかさ)

$$\mathbb{E}_{D_n}\left[\mathrm{W}_1(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s+1-\delta}{2s+d'}}$$ (for any $\delta > 0$).

どちらも（ほぼ）ミニマックス最適 [Yang & Barron, 1999; Niles-Weed & Berthet, 2022].

(Estimator for W_1 distance requires some modification)

Transformerの推定理論

[Shokichi Takakura, Taiji Suzuki: Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input. ICML2023]

Transformerの性質

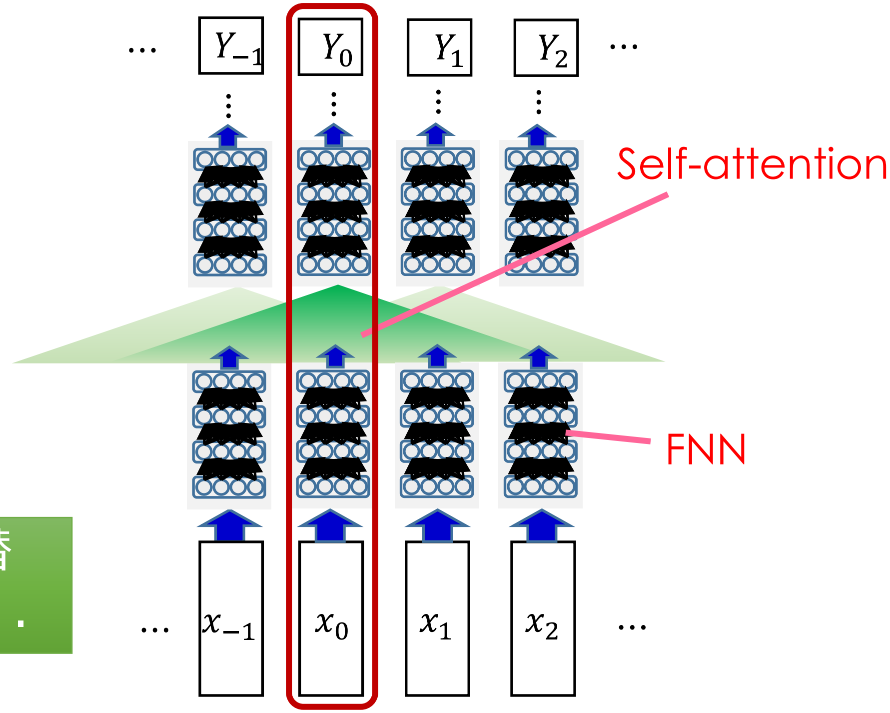
- かなり広いトークン幅から重要なトークンを選べる.

→ 次元の呪い？

- 入力に依存して重要なトークンを選択できる.

→ 次元の呪いを回避！

入力に依存して重要なトークンを切り替えることで，関数を「切り替えている」.



定理 (推定誤差)

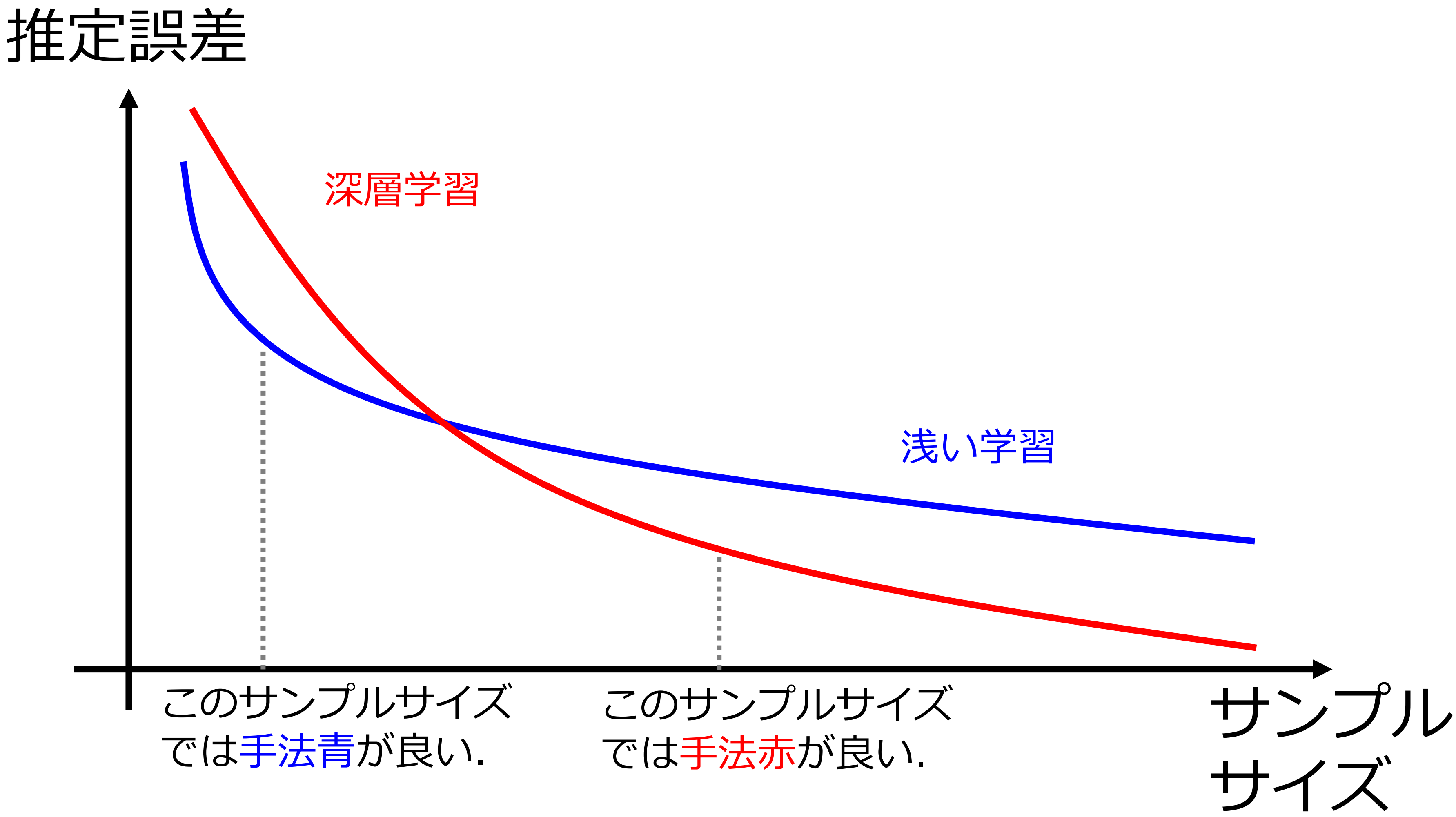
$$\frac{1}{r-l+1}\sum_{j=l}^{r}\mathbb{E}[\|\hat{F}_j - F_j^\circ\|_{L_2(P_X)}^2] \lesssim n^{-\frac{2a^\dagger}{2a^\dagger+1}}(\log n)^{2/\alpha+2+\max\{4/\alpha,4\}}$$

(ほぼミニマックス最適)

➢ 入力が無限次元でも多項式オーダーの収束レート.

収束レートに関する注意

注意：収束レートが速いからといって，その手法が常に良いとは限らない．



第2部
深層学習の汎化誤差
-Overparametrization-

- これまでの議論は，実は問題に合わせて「適切なサイズのネットワーク」を用いた場合の議論であった.

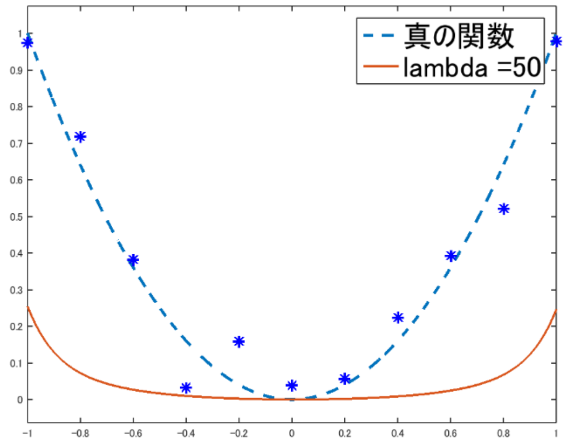
- 実際は，かなりサイズの大きなネットワークを用いる.

→ **Overparameterization**
(過剰パラメータ化)

過学習

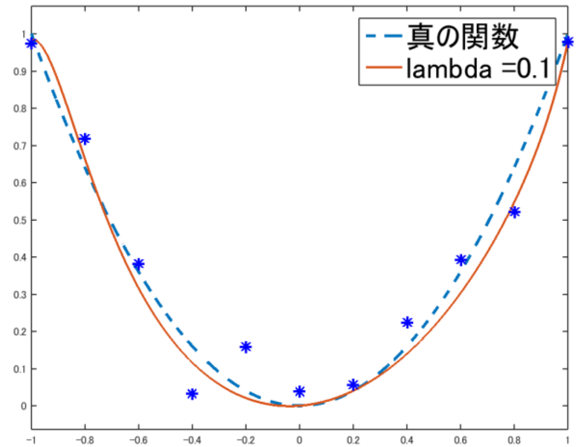
「なんでも表現できる方法」が最適とは限らない
少しのノイズにも鋭敏に反応してしまう

悪い学習結果



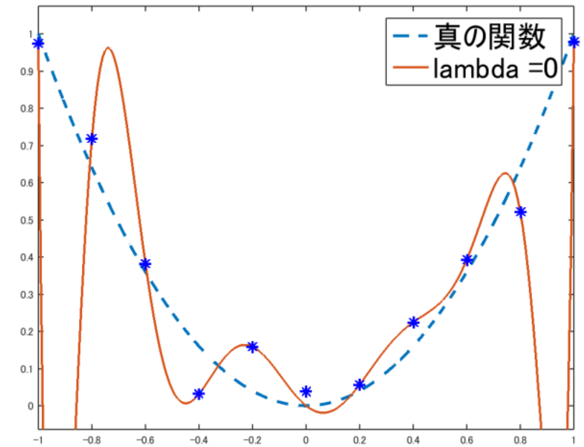
過小学習
説明力が低すぎる

良い学習結果



適切な学習
説明力が適切

悪い学習結果



過学習
説明力が高すぎる
（複雑すぎる）

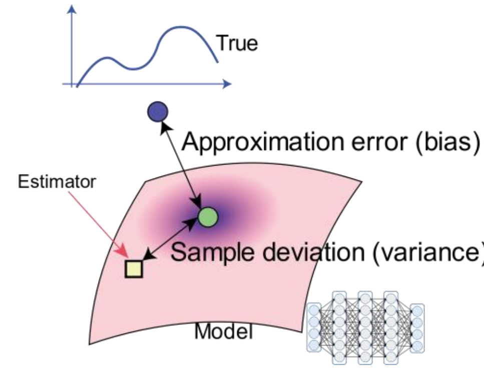
一見当てはまりが良いので危険

学習に用いるデータには誤りも含まれる

通常の学習理論

モデル $\mathcal{F}$: d-次元パラメータ
n: データサイズ

$$\hat{f} \leftarrow \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

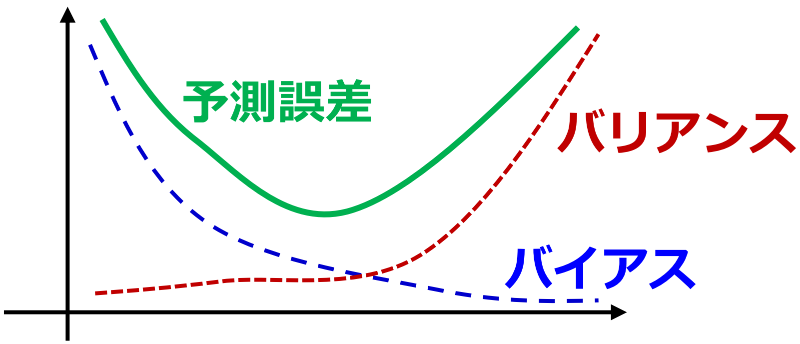


$$\text{予測誤差} = \frac{d}{n} + \underbrace{\inf_{f'\in\mathcal{F}} \|f - f'\|_{L^2(P_X)}^2}_{\text{(近似誤差)}}$$

Variance　　Bias

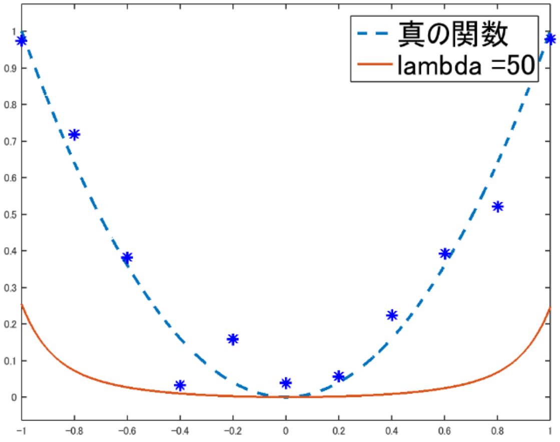
データサイズ (n) ≪ パラメータ数 (d)
の場合，通常の学習理論をナイーブに当てはめると失敗する．

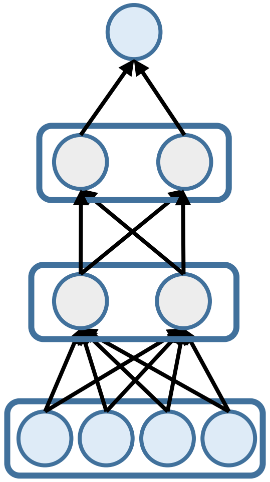
※カーネル法のような無限次元モデルもバイアスバリアンス分解で説明できるので，あくまで「ナイーブに」当てはめた場合の話．



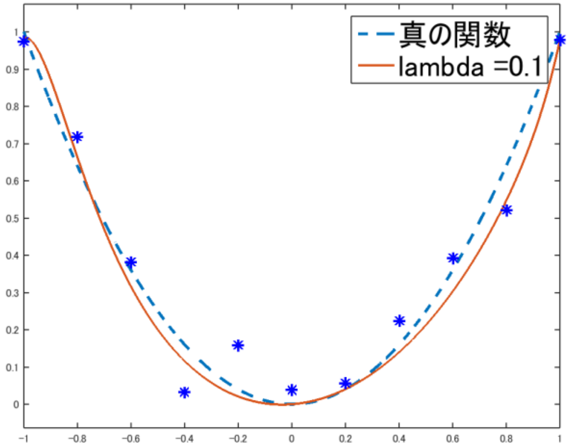
ニューラルネットワークでは？

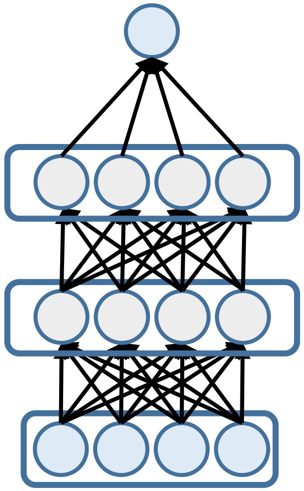
過小学習



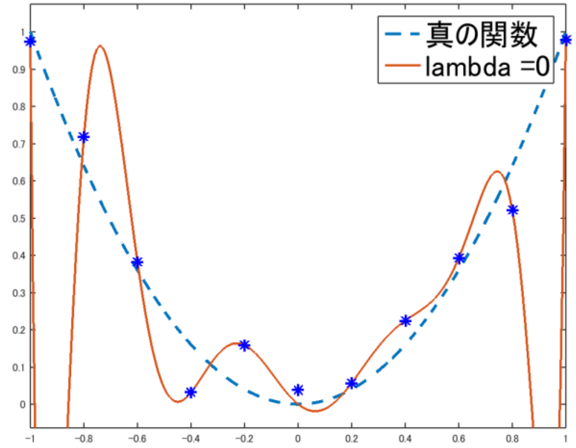


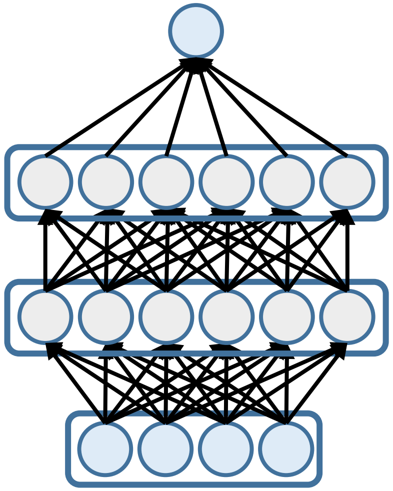
適切な学習





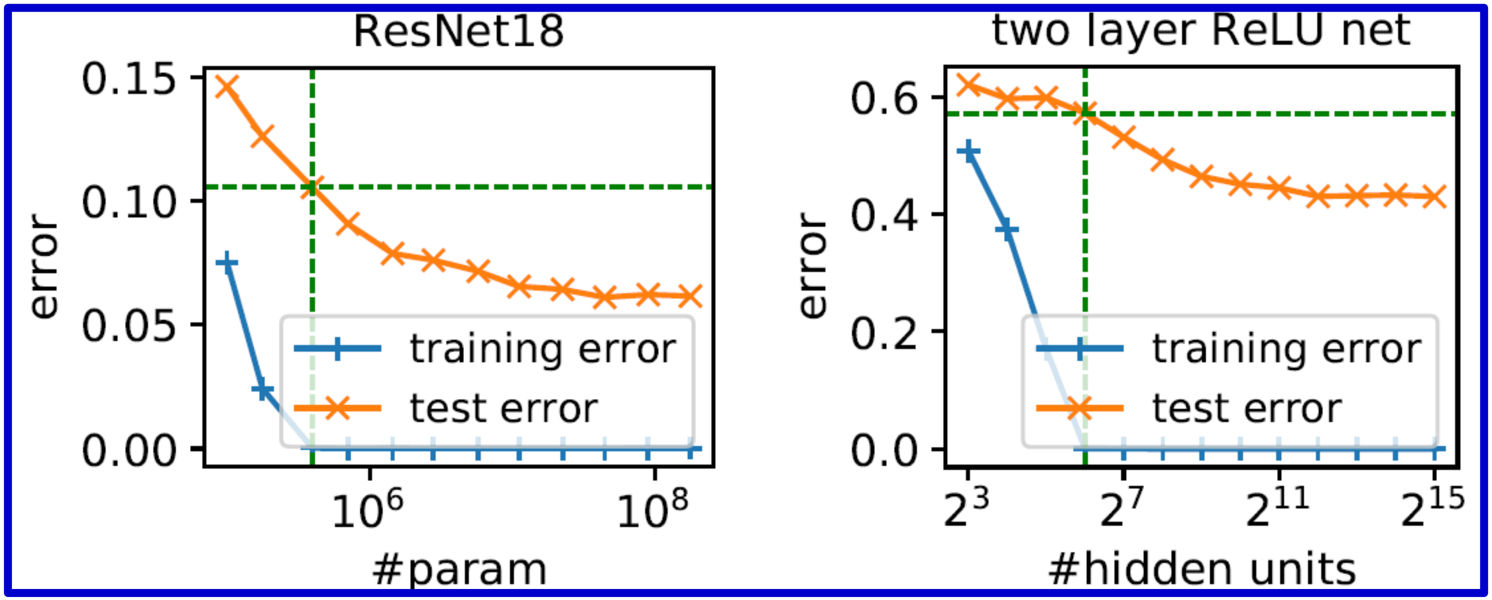
過学習





従来の学習理論

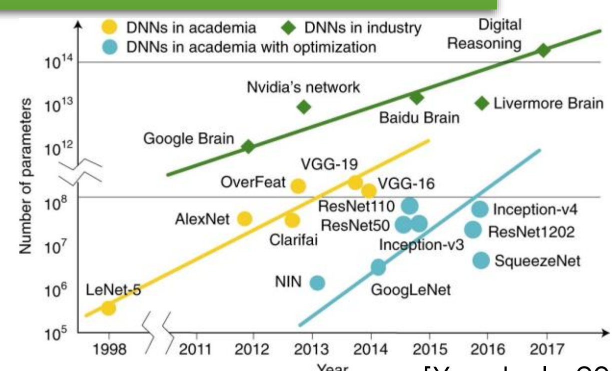
実際は...



[Neyshabur et al., ICLR2019]

ネットワークのサイズを大きくしても過学習しない

データサイズ：120万
モデルパラメータサイズ：10億



[Xu et al., 2018]

深層ニューラルネットの冗長性

「**Overparametrization**」
パラメータサイズがデータサイズを超えている状況
での汎化性能を説明したい.

パラメータ数 ≫ **データサイズ** ≫ **実質的自由度**

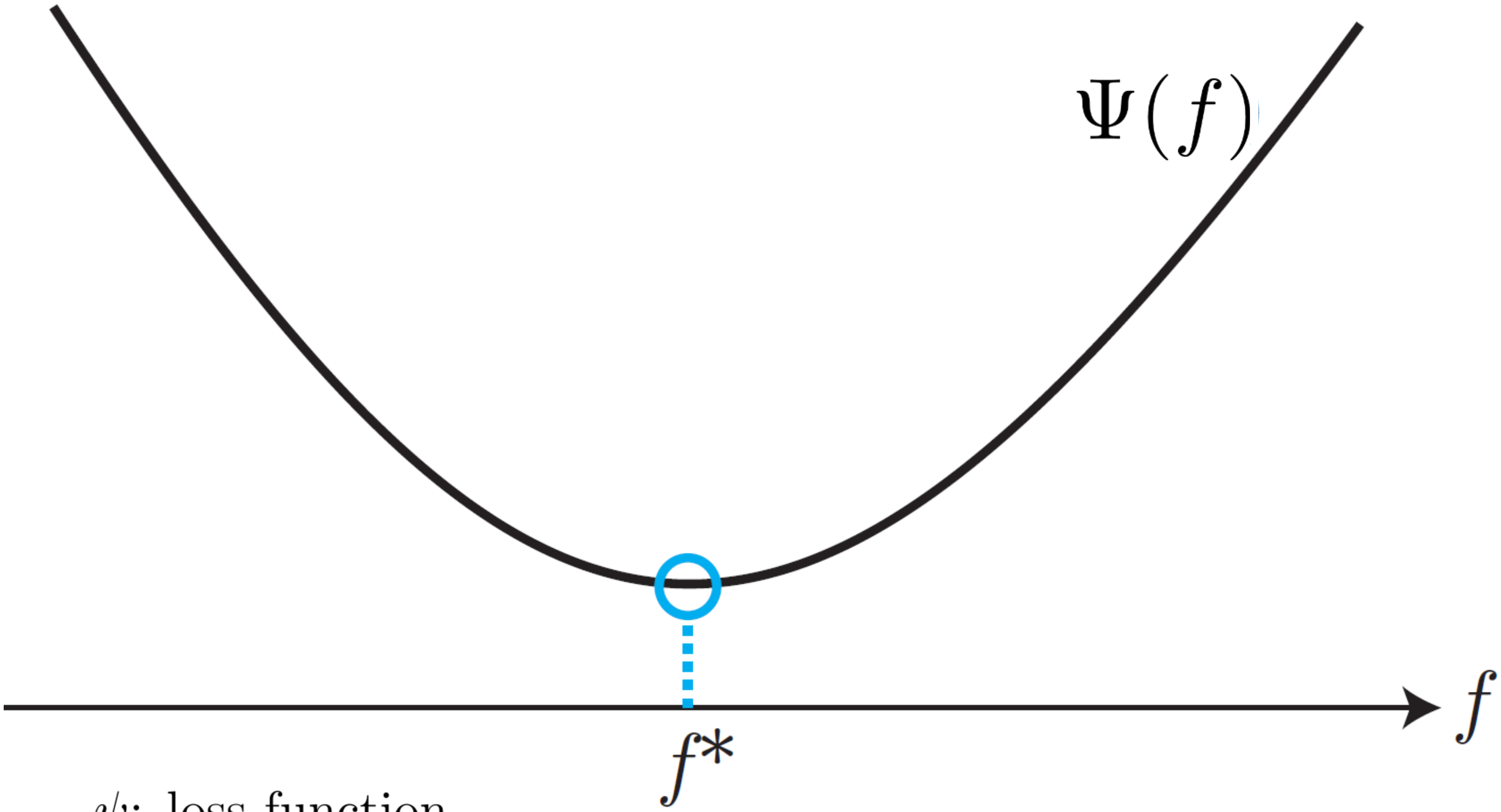
数十億　　数百万　　数十万

[仮説] 見かけの大きさ (パラメータ数) よりも
実質的な大きさ (自由度) はかなり小さいはず.

"実質的自由度"を調べる研究：
- ノルム型バウンド
- 圧縮型バウンド

「実質的自由度」として何が適切かを見つけることが理論上問題になる.

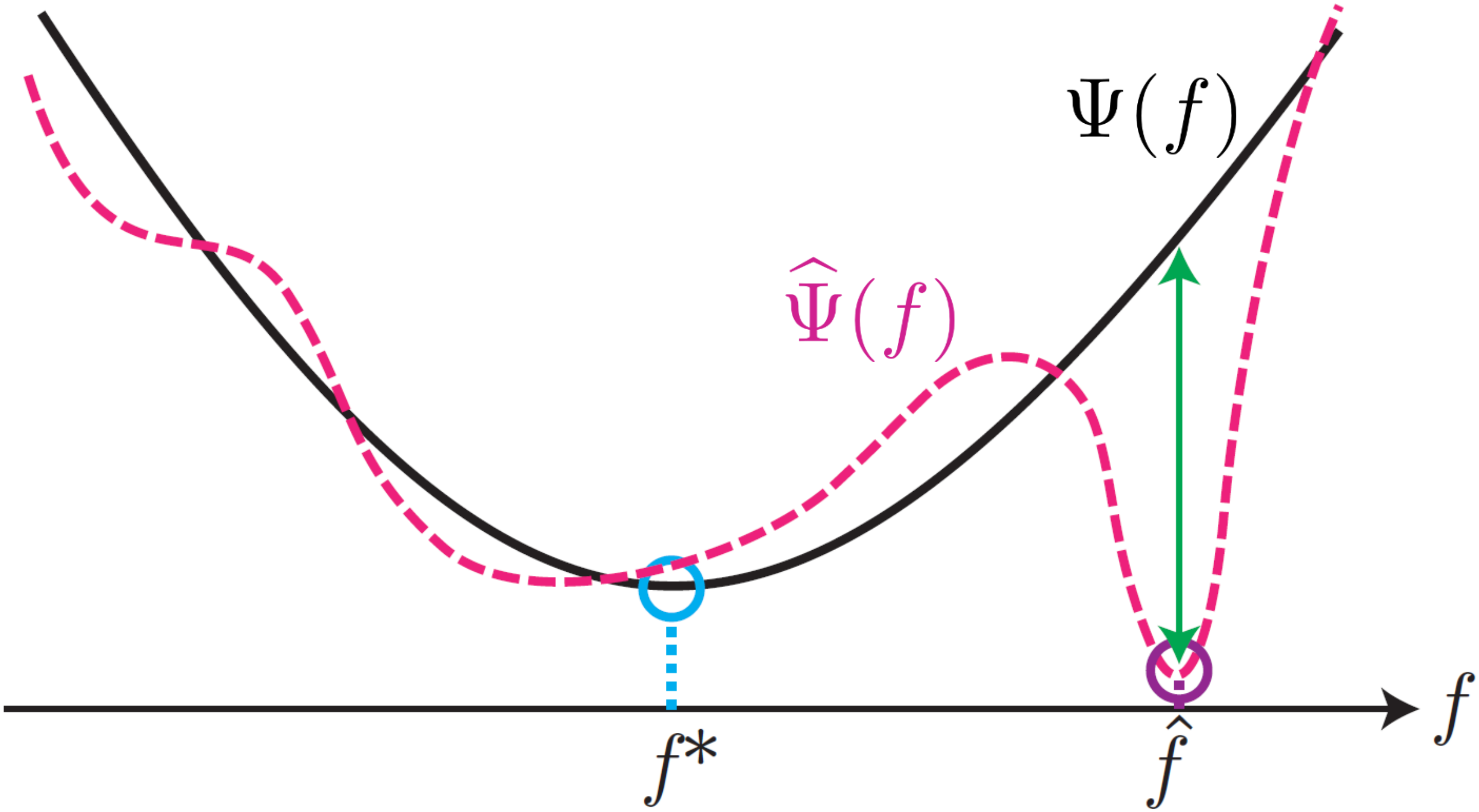
Uniform bound



ψ: loss function

$$\widehat{\Psi}(f) := \frac{1}{n}\sum_{i=1}^{n}\psi(y_i, f(x_i)) \qquad \Psi(f) := \mathbb{E}[\psi(Y, f(X))]$$

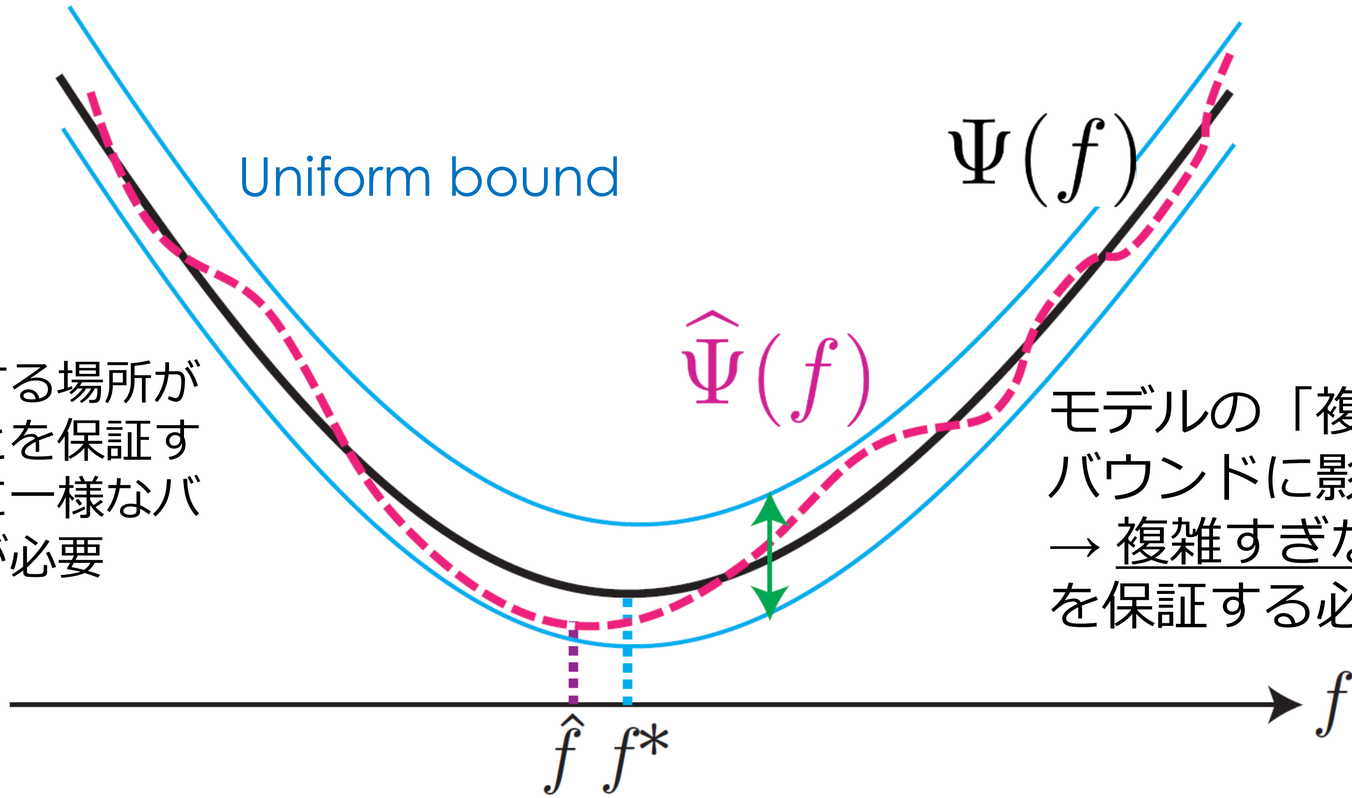
Uniform bound



“運良くデータに強く当てはまる”場所があるかもしれない.
→ 過学習

Uniform bound

過学習する場所がないことを保証するために一様なバウンドが必要



モデルの「複雑さ」がバウンドに影響する
→ 複雑すぎないことを保証する必要がある.

Uniform bound

$$\Psi(\hat{f}) - \widehat{\Psi}(\hat{f}) \leq \sup_{f \in \mathcal{F}} \left\{ \Psi(f) - \widehat{\Psi}(f) \right\}$$

Rademacher complexity

$$\bar{R}_n = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \psi(y_i, f(x_i)) \right]$$

Dudley integral (covering num.)

$$\frac{1}{\sqrt{n}} \int_{1/n}^{\infty} \sqrt{\log(N(\epsilon, \|\cdot\|_\infty, \mathcal{F}))} \mathrm{d}\epsilon$$

深層学習の汎化誤差バウンド（抜粋）

ノルム型バウンド

Author	Rate	Bound type
Neyshabur et al. (2015)	$\frac{2^L \prod_{\ell=1}^L R_{\ell,F}}{\sqrt{n}}$	Norm base
Bartlett et al. (2017)	$\frac{\prod_{\ell=1}^L R_{\ell,2}}{\sqrt{n}} \left(\frac{R_{\ell,2\to 1}^{2/3}}{R_{\ell,2}^{2/3}} \right)^{3/2}$	Norm base
Neyshabur et al. (2017)	$\frac{\prod_{\ell=1}^L R_{\ell,2}}{\sqrt{n}} \sqrt{L^2 W \sum_{\ell=1}^L \frac{R_{\ell,F}^2}{R_{\ell,2}^2}}$	Norm base
Golowich et al. (2018)	$\prod_{\ell=1}^L R_{\ell,F} \min\left\{ \frac{1}{n^{1/4}}, \sqrt{\frac{L}{n}} \right\}$	Norm base
Li et al. (2018) Harvey et al. (2017)	$\frac{\prod_{\ell=1}^L R_{\ell,2} \sqrt{L^2 W^2}}{\sqrt{n}}$	VC-dim **Naïve bound**
Arora et al. (2018)	$\sqrt{\frac{L^2 \max_{1\le i\le n} \lvert\widehat{f}(x_i)\rvert^2 \sum_{\ell=1}^L \frac{1}{\mu_\ell^2 \mu_{\ell\to}^2}}{n}}$	Compression
Baykal et al. (2018)	$\sqrt{\frac{L^2 \max_{1\le i\le n} \lvert\widehat{f}(x_i)\rvert^2 \sum_{\ell=2}^L (\hat{\Delta}^{\ell\to})^2 \sum_{i=1}^W S_i^\ell}{n}}$	Compression
Suzuki et al. (2018)	$\sum_{\ell=2}^L \sqrt{\lambda_\ell} + \sqrt{\frac{\sum_{\ell=1}^L m_{\ell+1}^\sharp m_\ell^\sharp}{n}}$	Compression

圧縮型バウンド

L: depth

$W = \max_\ell m_\ell$: width

R_F: Frobenius norm, R_2: operator norm

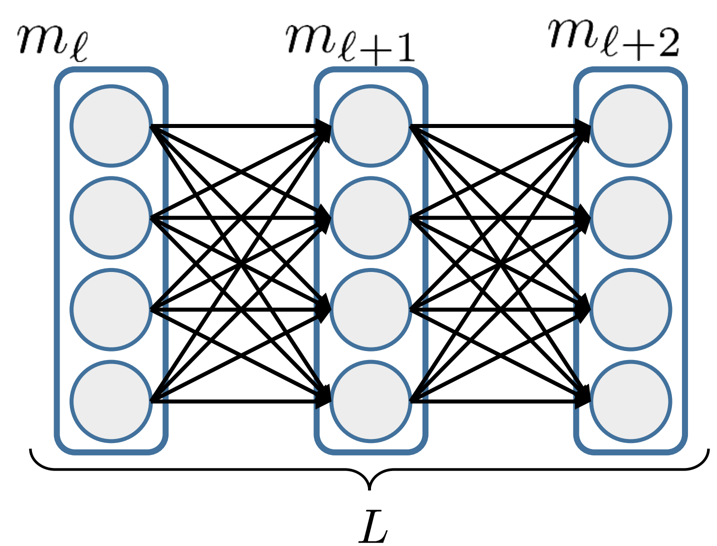
Naïve bound (VC-次元)

$$\Psi(\hat{f}) \leq \hat{\Psi}(\hat{f}) + \boxed{\;?\;}$$

VC-次元 [Harvey et al.2017]

$$\sqrt{L\frac{\sum_{\ell=1}^{L} m_\ell m_{\ell+1}}{n}\log(n)}$$

$$\left(\sqrt{L\frac{\#\text{ of parameters}}{n}}\right)$$



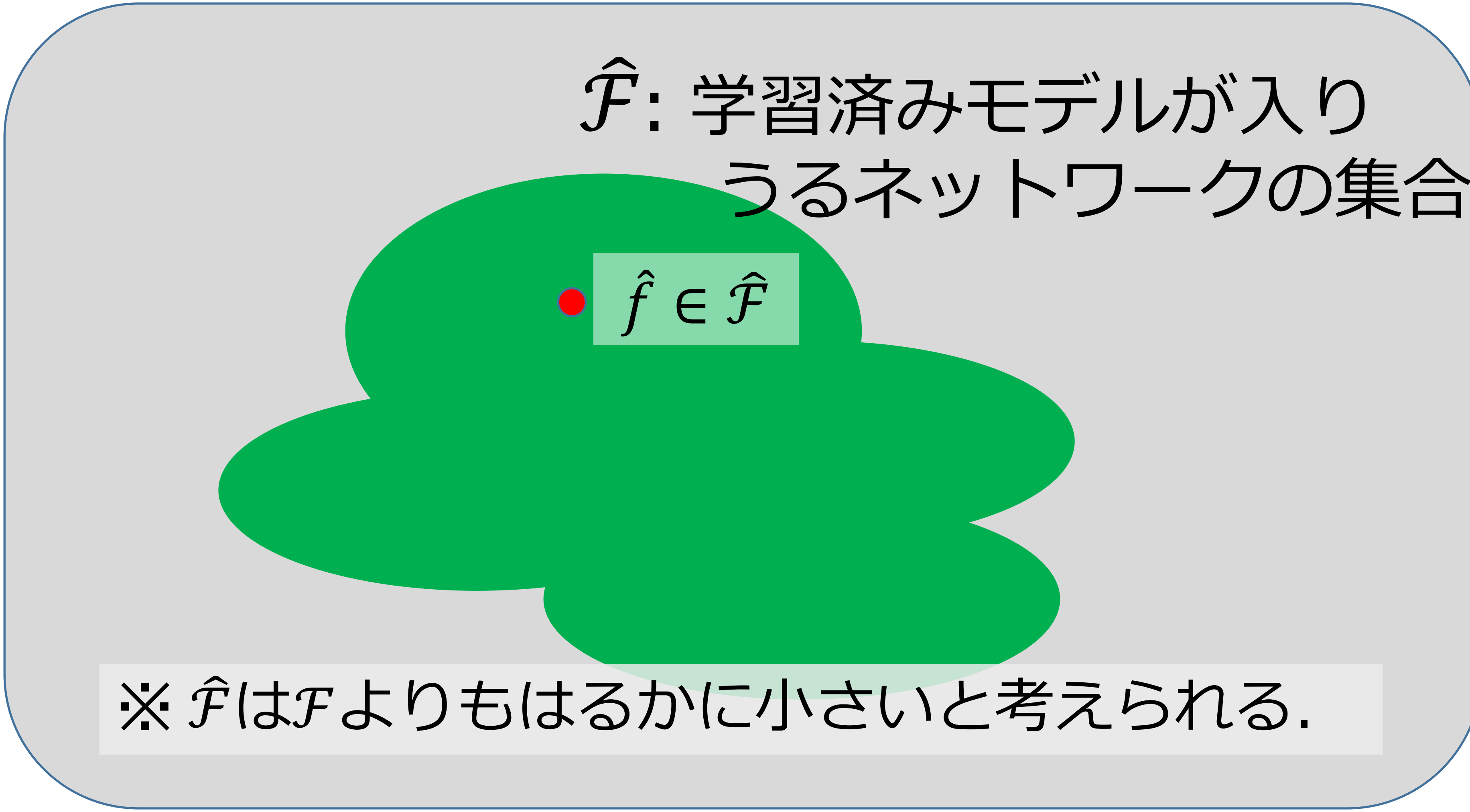
- m_ℓ: width of ℓ-th layer
- L: depth

☹ パラメータ数 $\sum_{\ell=1}^{L} m_\ell m_{\ell+1}$ がそのままバウンドに現れてしまう.

☹ <u>パラメータ数≫データサイズの状況を説明できていない.</u>

どうやって改善するか？

$\mathcal{F}$: ネットワーク全体



“典型的な学習済みネットワーク”の集合を解析する.

ノルム型バウンド

NN model: $f(x) = (W^{(L)}\eta(\cdot)) \circ (W^{(L-1)}\eta(\cdot)) \circ \cdots \circ (W^{(1)}x)$

- Bartlett et al. (2017): Spectrally-normalized margin bound

$$\boxed{\frac{1}{\sqrt{n}} \prod_{\ell=1}^{L} R_{\ell,2} \left(\sum_{\ell=1}^{L} \frac{R_{\ell,2\to1}^{2/3}}{R_{\ell,2}^{2/3}} \right)^{3/2}}$$

$$R_{\ell,2} := \|W^{(\ell)}\| = \sigma_1(W^{(\ell)}) \qquad R_{\ell,2\to1} := \sum_j \|W_{j,:}^{(\ell)}\|$$

(最大特異値)

➢ $\prod_{\ell=1}^{L} R_{\ell,2}$ は中間層から最終層にどう誤差が伝搬するかを表すリプシッツ定数.

➢ $R_{\ell,2\to1}/R_{\ell,2}$ は行列$W^{(\ell)}$のスパース性を表す．もし$W^{(\ell)}$が小さなノルムの行をたくさん含むなら，この値は小さくなる.

☺ 横幅に依存しない → 過剰パラメータ化されたネットワークにも使える!

[P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6241–6250, 2017]

ノルム型バウンド (2)

NN model: $f(x) = (W^{(L)}\eta(\cdot)) \circ (W^{(L-1)}\eta(\cdot)) \circ \cdots \circ (W^{(1)}x)$

- Neyshabur et al. (2017): PAC-Bayes bound

$$\prod_{\ell=1}^{L} R_{\ell,2} \sqrt{\frac{L^2 \max_\ell m_\ell \sum_{\ell=1}^{L} \frac{R_{\ell,\mathrm{F}}}{R_{\ell,2}}}{n}}$$

$R_{\ell,2} := \|W^{(\ell)}\| = \sigma_1(W^{(\ell)})$ (maximum singular value)

$R_{\ell,\mathrm{F}} := \|W^{(\ell)}\|_{\mathrm{F}}$: Frobenius norm (smaller than $R_{\ell,2\to1}$)

- $R_{\ell,2\to1}$の代わりに$R_{\ell,F}$ (which is smaller) 出ている.
- その代わり，横幅m_ℓも出てきている.
- 重み行列がスパースではない状況では，こちらの方がタイト.

[B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. ICLR2018.]

- Golowich et al. (2018)

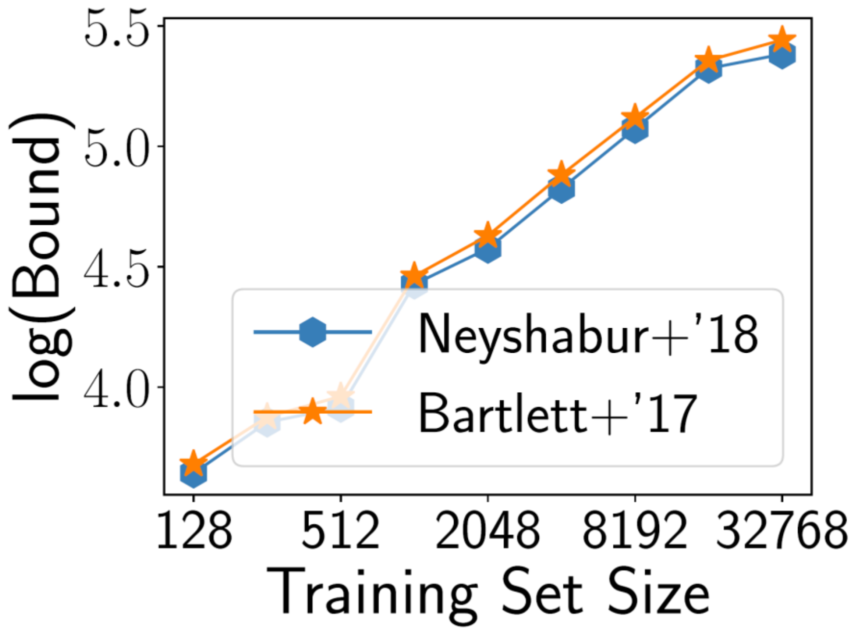
$$\prod_{\ell=1}^{L} R_{\ell,\mathrm{F}} \min\left\{\frac{1}{n^{1/4}}, \sqrt{\frac{L}{n}}\right\}$$

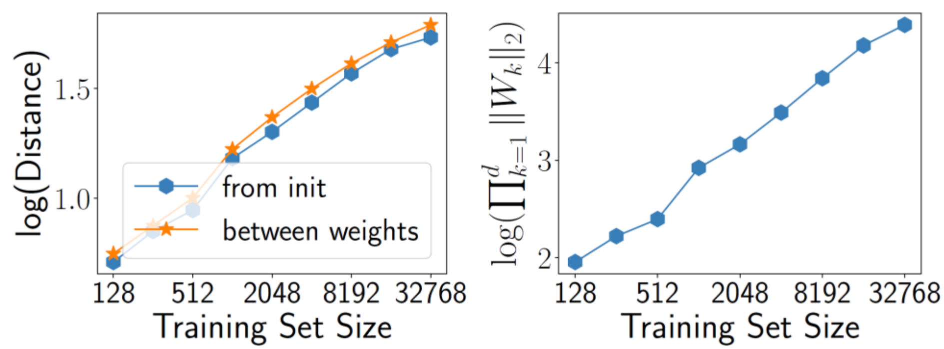
- 完全にFrobeniusノルムだけで特徴づけられている.
- 横幅が出てこない.

[N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. COLT2018, pp. 297–299.]

批判

ノルム型バウンドは大きすぎる.



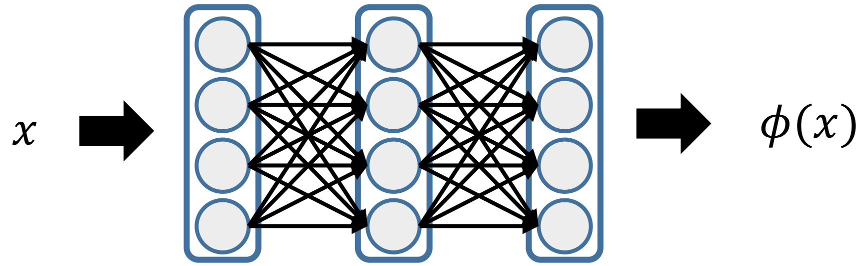


[Nagarajan and Kolter: Uniform convergence may be unable to explain generalization in deep learning. NeurIPS2019]

実験結果：訓練データサイズが増えるほどバウンドが大きくなる.

- これはリプシッツ定数 ($\prod_{\ell=1}^{L}\|W_\ell\|_2 = \prod_{\ell=1}^{L} R_{\ell,2}$) が大きくなることによる.
- この量はよりデータ依存な量に置換できる (Wei&Ma (2019), Arora et al. (2018)). → かなりタイトになる.

データ依存なリプシッツ定数



κ:「データ依存な」リプシッツ連続性:

$$\|\phi(x) - \phi(x')\| \le \kappa \|x - x'\|$$

for any training data point x and its neighborhood point x'.

$$\boxed{\frac{1}{\sqrt{n}} \left(\sum_{\ell=1}^{L} \kappa^{2/3} R_{\ell,2\to1}^{2/3} \right)^{3/2}}$$

(informal)

[Wei and Ma: Data-dependent Sample Complexity of Deep Neural Networks via Lipschitz Augmentation. NeurIPS 2019]

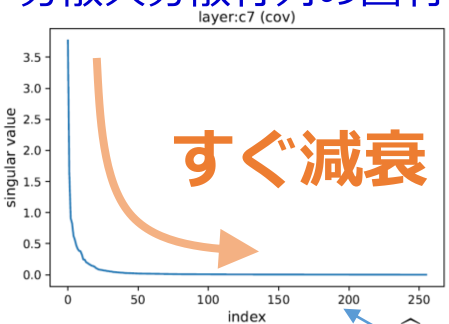
➢ このバウンドは先ほどのバウンドよりかなりタイトになる.

圧縮型バウンド

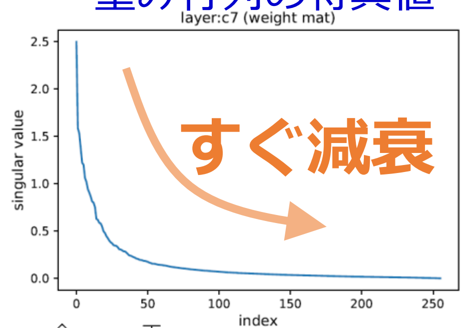
- 中間層の**分散共分散行列の固有値分布**
- 中間層の**重み行列の特異値分布**

が速く減衰するなら圧縮しやすい．

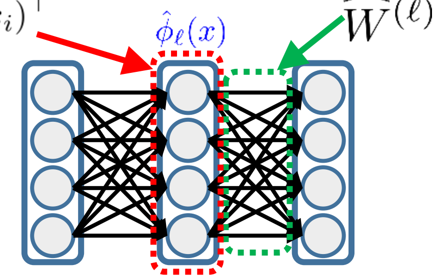
分散共分散行列の固有値



重み行列の特異値



$$\widehat{\Sigma}_\ell = \frac{1}{n}\hat{\phi}_\ell(x_i)\hat{\phi}_\ell(x_i)^\top$$



分散共分散行列も重み行列も
特異値が速く減衰
→**<u>小さい統計的自由度</u>**
(AIC, Mallows' Cp)

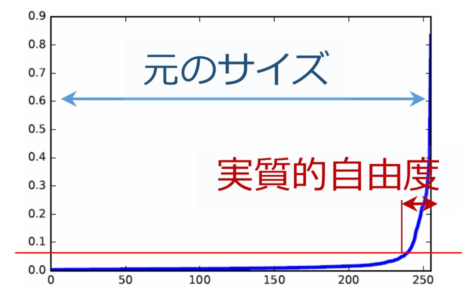
$$\Psi(\widehat{f}) \le \hat{\Psi}(\widehat{f}) + O\left(\left(L\frac{\sum_{\ell=1}^{L} m_\ell}{n}\log(n)\right)^{\frac{2\alpha}{1+2\alpha}} + \sqrt{L^{1+\delta}\frac{(\sum_{\ell=1}^{L} m_\ell)^{\frac{4/\beta}{4/\beta+2(1-1/2\alpha)}}}{n}\log(n)^3}\right)$$

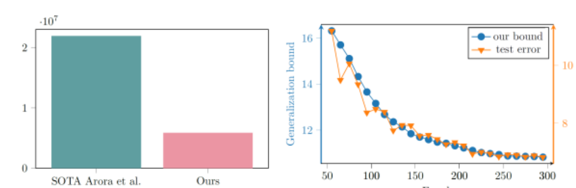
カーネル法の理論
(そもそもカーネルは無限次元モデル)

(次ページに詳細)

[実験的観察] 実際に学習したネットワークは圧縮しやすい.

	元サイズ	圧縮可能サイズ
Layer	Original	Our bound
1	1,728	1,013
4	147,456	84,499
6	589,824	270,216
9	1,179,648	50,768
12	2,359,296	4,583
15	2,359,296	3,886
	大	小



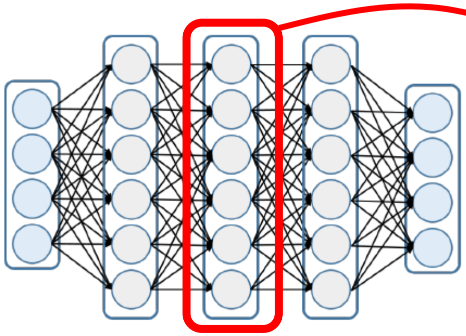


(a) Bound comparison (b) Generalization bound

[Suzuki: Fast generalization error bound of deep learning from a kernel perspective. AISTATS2018]
[Li, Sun, Liu, Suzuki and Huang: Understanding of Generalization in Deep Learning via Tensor Methods. AISTATS2020]
[Suzuki, Abe, Nishimura: Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network, ICLR2020]
[Suzuki et al.: Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. IJCAI-PRICAI 2020]

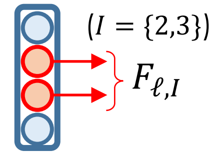
中間層の実質的自由度

参考



$F_\ell(x) \in \mathbb{R}^{m_\ell}$: 中間層の特徴量 (m_ℓ次元)

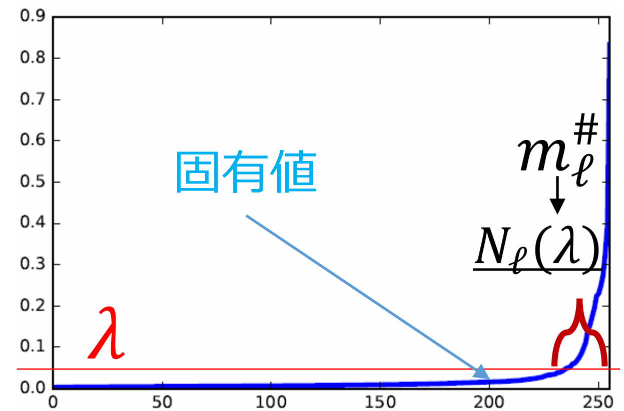
$F_{\ell,I}(x) \in \mathbb{R}^{|I|}$: 部分ベクトル



対応する共分散行列: $\Sigma_\ell := \mathbb{E}[F_\ell F_\ell^\top]$

その固有値: $\mu_j^{(\ell)}$ $(j = 1, \dots, m_\ell)$

統計的自由度: $N_\ell(\lambda) = \sum_{j=1}^{m_\ell} \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)} + \lambda}$



定理 (Bach, 2017)

任意の$\lambda > 0$に対して，$m_\ell^\# = \Omega(N_\ell(\lambda)\log(N_\ell(\lambda)))$個の特徴量の集合$\hat{I}_\ell \subset [m_\ell]$ ($|\hat{I}_\ell| = m_\ell^\#$)が存在して，

$$\sup_{w \in \mathbb{R}^{m_\ell}:\|w\| \le 1} \inf_{\hat{w} \in \mathbb{R}^{m_\ell^\sharp}} \mathbb{E}[(w^\top F_\ell(X) - \hat{w}^\top F_{\ell,\hat{I}_\ell}(X))^2] \lesssim \lambda$$

($\hat{I}_\ell$に対応した部分ベクトル)

つまり，中間層を大体$m_\ell^\#$個の特徴量で代替できる．

分散共分散行列と重み行列の低ランク性

$$f(x) = (W^{(L)}\eta(\cdot)) \circ (W^{(L-1)}\eta(\cdot)) \circ \cdots \circ (W^{(1)}x)$$

- 近似的に低ランクな重み行列と分散共分散行列:
  - $\sigma_j(\widehat{W}^{(\ell)}) \lesssim j^{-\alpha}$
  - $\sigma_j(\widehat{\Sigma}^{(\ell)}) \lesssim j^{-\beta}$

特異値が早く減衰

($\sigma_j(\cdot)$: j-th largest sigular-value)

+ Other boundedness condition.

定理 (Suzuki, Abe, Nishimura, ICLR2020)

$$\Psi(\hat{f}) \leq \hat{\Psi}(\hat{f})$$

$$+ O\left(\left(L\frac{\boxed{\sum_{\ell=1}^L m_\ell}}{n}\log(n)\right)^{\boxed{\frac{2\alpha}{1+2\alpha}}} + \sqrt{L^{1+\delta}\frac{\boxed{\left(\sum_{\ell=1}^L m_\ell\right)^{\frac{4/\beta}{4/\beta+2(1-1/2\alpha)}}}}{n}\log(n)^3}\right)$$

where $\delta = \dfrac{\beta}{\frac{4\alpha}{(2\alpha-1)} + \beta}$.

VC-次元によるバウンドより大きく改善:

横幅の二乗

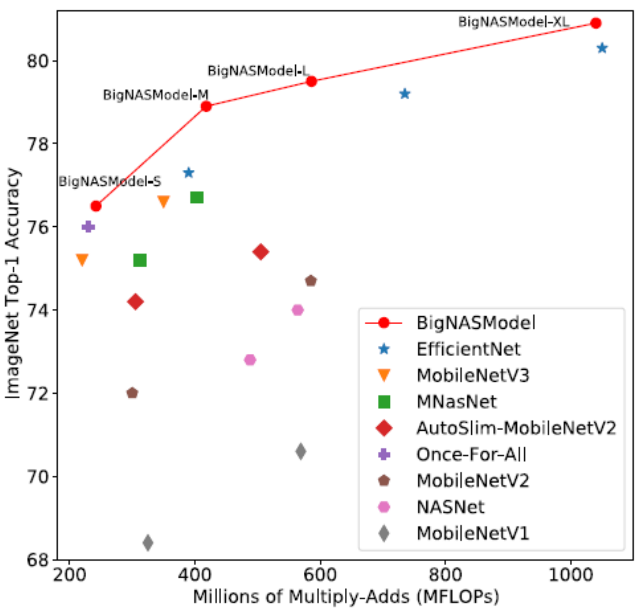
$$\Psi(\hat{f}) \leq \hat{\Psi}(\hat{f}) + \sqrt{L\frac{\sum_{\ell=1}^L \boxed{m_\ell m_{\ell+1}}}{n}\log(n)}$$

BigNAS

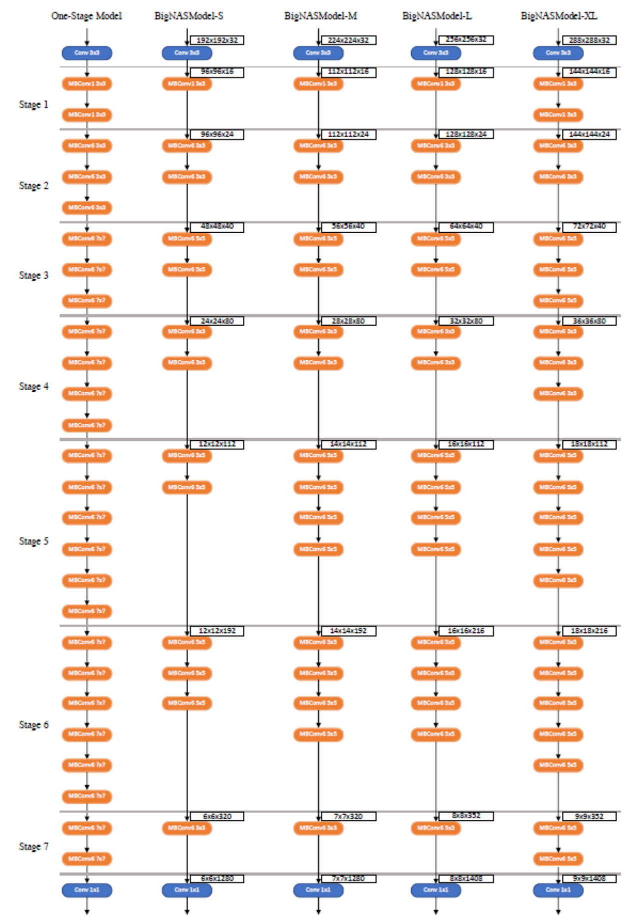
[Yu et al.: BigNAS: Scaling Up Neural Architecture Search with Big Single-Stage Models. ECCV2020]

(理論と関係あるNAS手法)

- 学習後のネットワークが圧縮できるように学習
- 大きなネットワークから小さなネットワークを生成できる
- EfficientNetを上回る効率性を実現



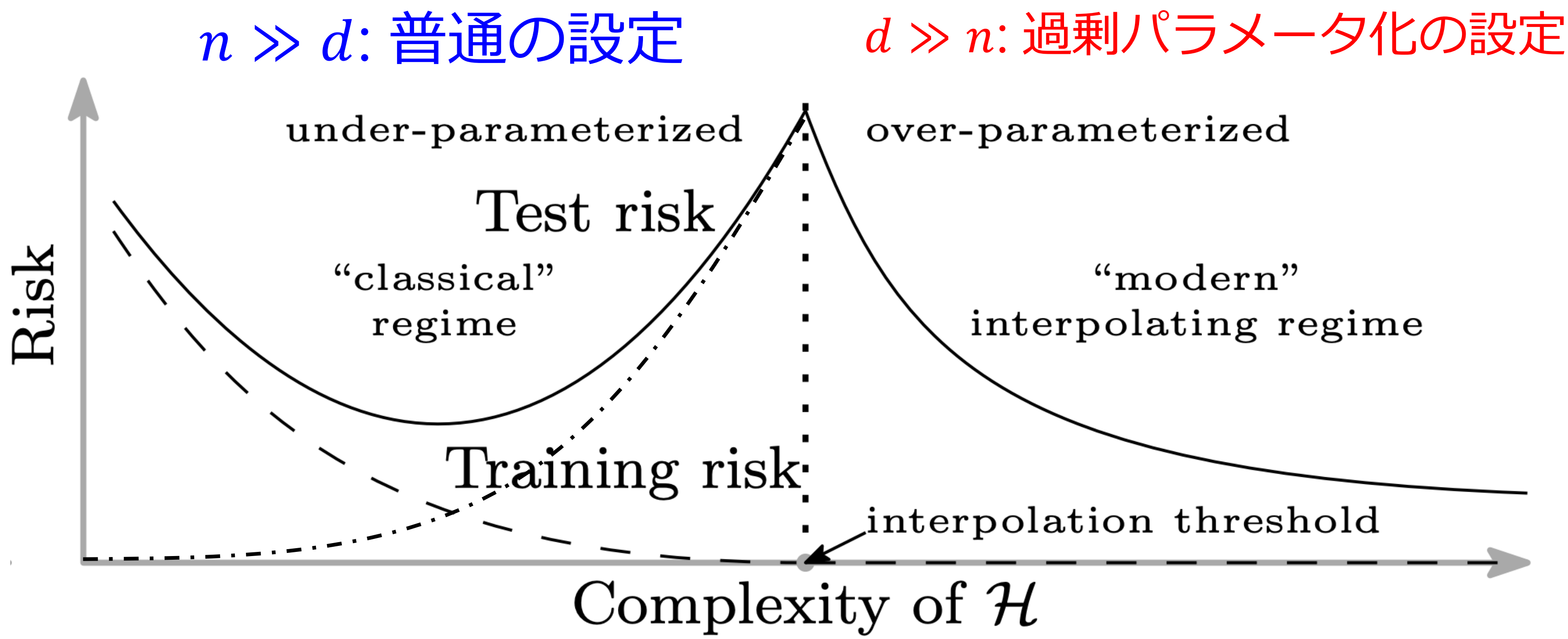
Group	Model Family	Params	FLOPs	Top-1
200M FLOPs	MobileNetV1 0.5×	1.3M	150M	63.3
	MobileNetV2 0.75×	2.6M	209M	69.8
	AutoSlim-MobileNetV2	4.1M	207M	73.0
	MobileNetV3 1.0×	5.4M	219M	75.2
	MNasNet A1	3.9M	315M	75.2
	Once-For-All	4.4M	230M	76.0
	Once-For-All finetuned	4.4M	230M	76.4
	BigNASModel-S	**4.5M**	**242M**	**76.5**
400M FLOPs	NASNet B	5.3M	488M	72.8
	MobileNetV2 1.3×	5.3M	509M	74.4
	MobileNetV3 1.25×	8.1M	350M	76.6
	MNasNet A3	5.2M	403M	76.7
	EfficientNet B0	5.3M	390M	77.3
	BigNASModel-M	**5.5M**	**418M**	**78.9**
600M FLOPs	MobileNetV1 1.0×	4.2M	569M	70.9
	NASNet A	5.3M	564M	64.0
	DARTS	4.9M	595M	73.1
	EfficientNet B1	7.8M	734M	79.2
	BigNASModel-L	**6.4M**	**586M**	**79.5**
1000M FLOPs	EfficientNet B2	9.2M	1050M	80.3
	BigNASModel-XL	**9.5M**	**1040M**	**80.9**



圧縮できるように学習するとスクラッチ学習より性能が向上することもある.

Overparametrizeされた ネットワークの統計学

Double-descent (二重降下)



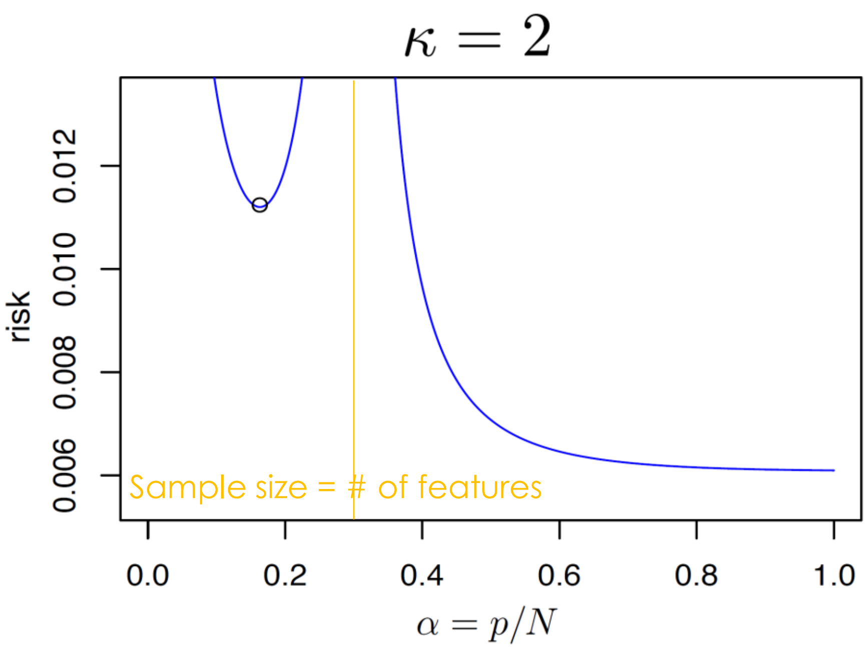
[Belkin et al.: Reconciling modern machine learning practice and the bias-variance trade-off. 2018]

- モデルがある複雑度 (サンプルサイズ) を超えた後，第二の降下が起きる.
- モデルサイズがデータより多いと推定量の<u>バリアンスがむしろ減る</u>.

※設定によるので注意が必要.

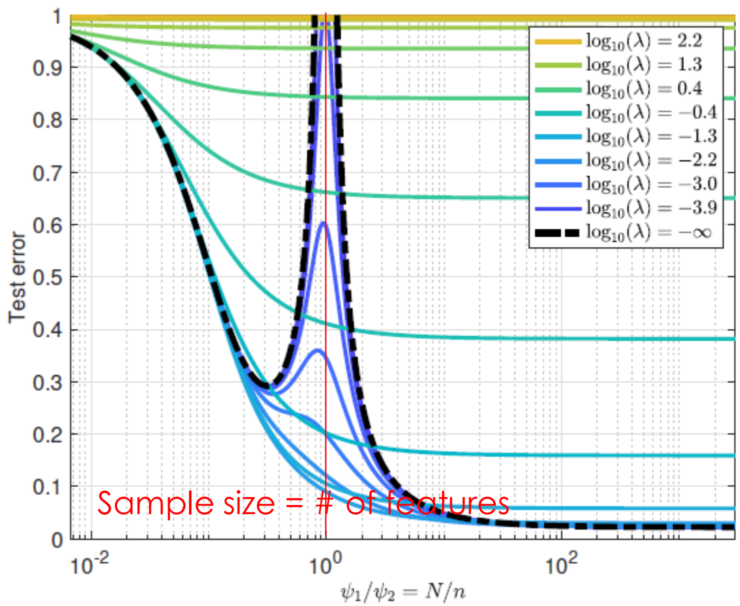
実際の例

Principal component regression



(いくつの主成分を用いたか)
Populationの分散共分散を知っているとして，
その主成分を利用

2-layer neural network



(いくつのニューロンを用いたか)

[Xu and Hsu: On the number of variables to use in principal component regression. NeurIPS2019.]

[Mei and Montanari. "The generalization error of random features regression: Precise asymptotics and double descent curve." *arXiv preprint arXiv:1908.05355* (2019)]

典型的なアプローチ (抜粋)

- **ランダム行列理論**
  - $d/n \to \gamma > 0$という漸近的設定で，厳密なリスクの導出
  - Marchenko–Pastur則， Stieltjes変換

$$\tilde{m}(z) := \frac{1}{d}\mathrm{Tr}\left[\left(X^\top X/n - zI\right)^{-1}\right] \quad \xrightarrow{a.s.} \quad m(z)$$

$$m(z) = \int_0^\infty \frac{1}{\tau(1-\gamma(1+zm(z))) - z}\mathrm{d}H(\tau) \qquad (H(\tau)\text{: Spectral measure of } \Sigma_x)$$

■ Dobriban&Wager: High-dimensional asymptotics of prediction: Ridge regression and classification. The Annals of Statistics, 46(1):247–279, 2018.
■ Hastie et al.: Surprises in High-Dimensional Ridgeless Least Squares Interpolation, arXiv:1903.08560.
■ Song&Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. Communications on Pure and Applied Mathematics. arXiv:1908.05355 (2019).

- **集中不等式による評価**
  - 有限サンプルサイズにおける予測誤差の上界評価 ($n < \infty$)
  - 収束レートが評価できる. ($n \to \infty$を取る前の振る舞いを評価)

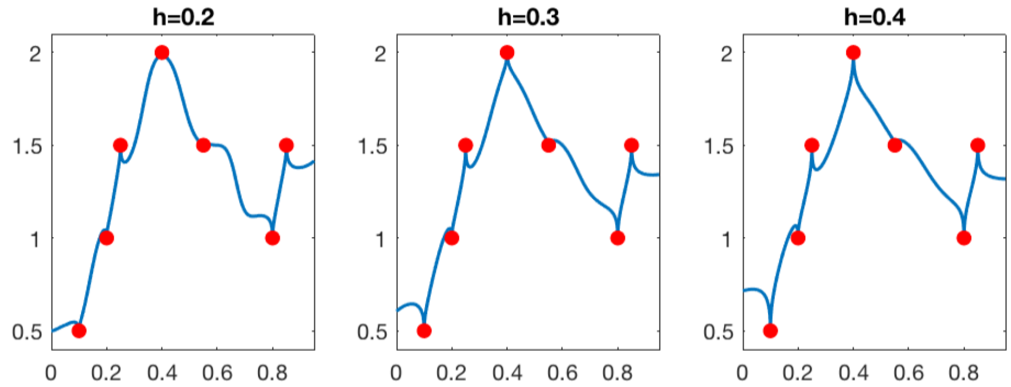
■ Belkin, Rakhlin&Tsybakov: Does data interpolation contradict statistical optimality? AISTATS2019.
■ Bartlett, Long, Lugosi&Tsigler: Benign Overfitting in Linear Regression. PNAS, 117(48):30063-30070, 2020.
■ Liang&Rakhlin: Just interpolate: Kernel "Ridgeless" regression can generalize. The Annals of Statistics, 48(3):1329–1347, 2020.

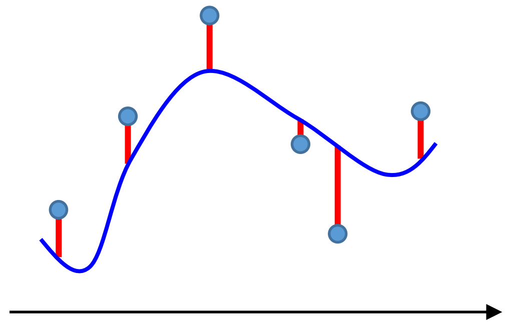
- **CGMT (Convex Gaussian min-max Theorem)**

■ Thrampoulidis, Oymak & Hassibi: Regularized linear regression: A precise analysis of the estimation error. COLT2015.
■ Thrampoulidis, Abbasi & Hassibi: Precise error analysis of regularized m-estimators in high dimensions. IEEE Transactions on Information Theory, vol. 64, no. 8, pp. 5592–5628, 2018.

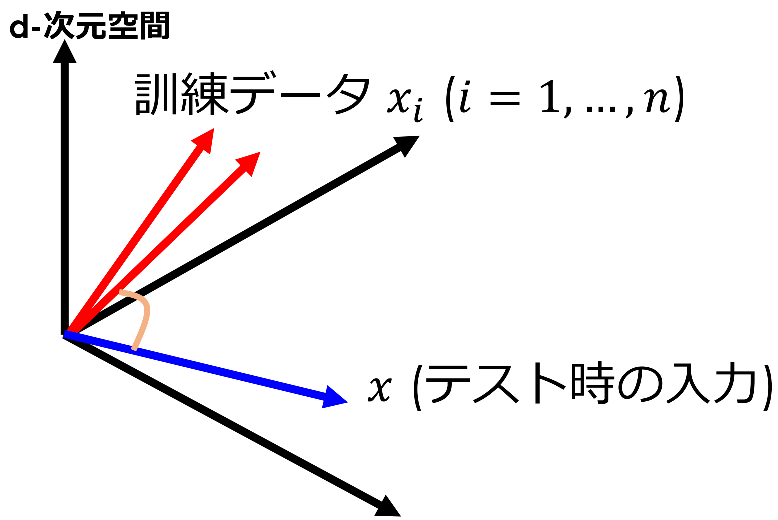
直感

[Belkin, Rakhlin&Tsybakov: Does data interpolation contradict statistical optimality? AISTATS2019]

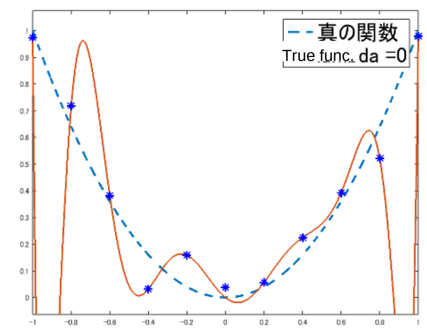




- 過剰パラメータ化されたモデルは “スパイク” 成分を持つ.
- スパイク成分がノイズを説明.
- 大まかな関数形はモデルの主成分が説明.



- 「スパイク成分」とはほぼ直交.
- 直交するには高次元性が必要.
- 高次元空間で2つのランダムベクトルはほぼ直交.



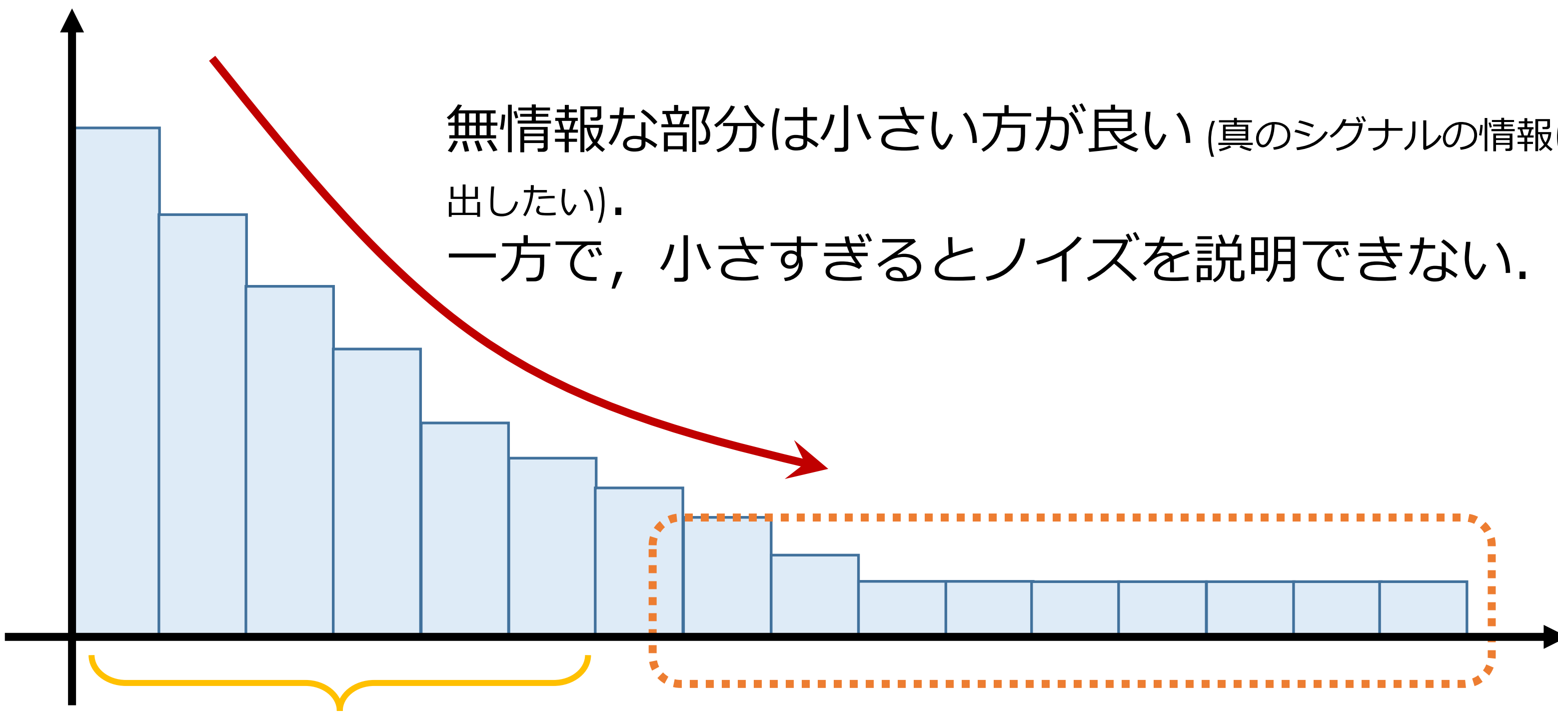
悪性過学習
(モデルの説明力が中途半端なので，無理してノイズも説明)

線形モデル: 分散共分散行列での直感

$$\Sigma_X = \mathbb{E}[xx^\top] \in \mathbb{R}^{d\times d}$$

Eigenvalues of Σ_X (spectrum)

無情報な部分は小さい方が良い (真のシグナルの情報は取り出したい).
一方で，小さすぎるとノイズを説明できない.



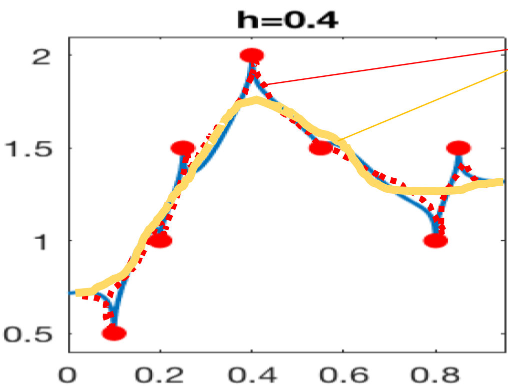
情報のある部分:
真のシグナルを説明

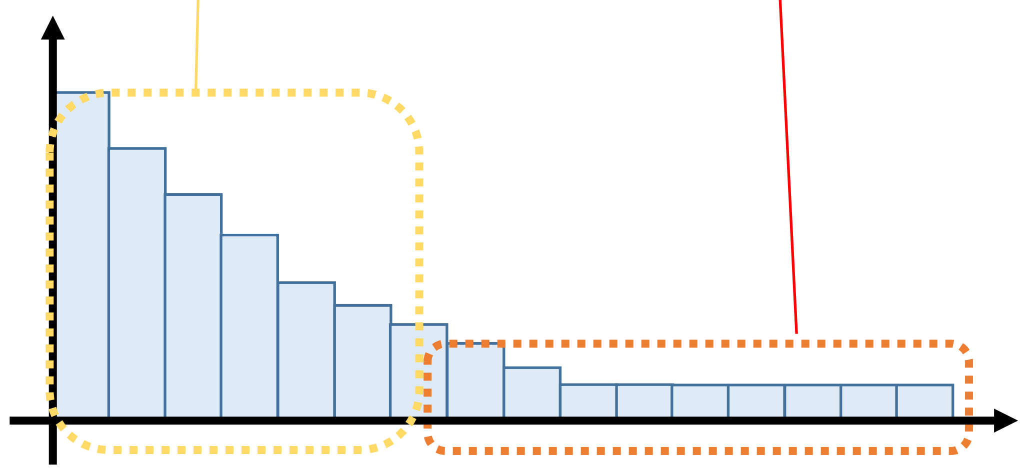
無情報な部分:
<u>ノイズを説明</u>

※サンプルサイズに応じてどこから無情報になるかは変わる．あくまで「そのサンプルサイズの解像度では無情報に見える」ということ.

シグナル/ノイズ分解

$$\hat{f}(x) = \hat{\beta}^\top x = \underbrace{x^\top (X^\top X)^+ X^\top X \beta^*}_{\text{Signal part}} + \underbrace{x^\top (X^\top X)^+ X^\top \epsilon}_{\text{Noisy part}}$$





問題設定: 補間推定量

$$y_i = x_i^\top \beta + \epsilon_i \qquad (d > n)$$

overparameterization

$$\mathbb{E}[\epsilon_i] = 0,\ \ \mathrm{Var}(\epsilon_i) = \sigma^2,\ \ \mathrm{Cov}(x_i) = \Sigma$$
$$(X = [x_1, \ldots, x_n]^\top,\ Y = [y_1, \ldots, y_n]^\top)$$

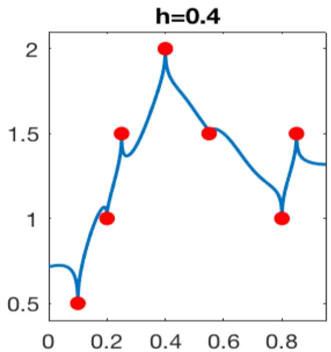
- **<u>最小ノルム補間推定量</u> (Minimum-norm interpolator):**

過学習しているものの中で一番単純な推定量

$$\hat{\beta} = \mathop{\arg\min}_{\beta \in \mathbb{R}^d} \ \|\beta\|$$

$$\text{s.t.} \quad y_i = x_i^\top \beta \ (\forall i = 1, \ldots, n)$$: interpolation

$$\Rightarrow \quad \hat{\beta} = (X^\top X)^+ X^\top Y$$



- **予測誤差 (predictive risk, excess risk):**

$$\mathbb{E}_\epsilon [\|\hat{\beta} - \beta\|_\Sigma^2 | X]$$

$$\hat{\beta}_\lambda := \arg\min_{\beta \in \mathbb{R}^d} \|X\beta - Y\|^2 + \lambda \|\beta\|^2 \quad \Rightarrow \quad \hat{\beta} = \lim_{\lambda \searrow 0} \hat{\beta}_\lambda$$

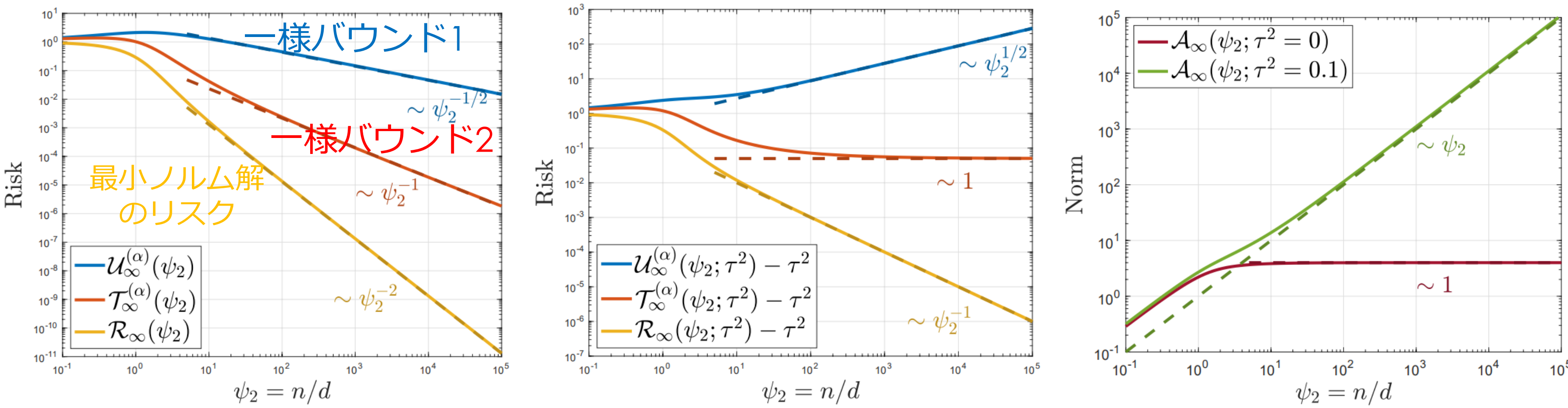
Ridge regression

一様バウンドではダメ

[Yang, Bai, Mei: Exact Gap between Generalization Error and Uniform Convergence in Random Feature Models. ICML2021.]

[Bartlett&Long, JMLR, 22(204):1−15, 2021]も参照

$\alpha = 1.5$



(a) Noiseless response ($\tau^2 = 0$) (b) Noisy response ($\tau^2 = 0.1$) (c) Minimum norm $\mathcal{A}_\infty(\psi_2)$

以下の量の$n, N, d \to \infty$の極限を取る．ただし，$N/d \to \infty, n/d \to \psi_2$.

$$U(A,N,n,d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A} \left(R(\boldsymbol{a}) - \widehat{R}_n(\boldsymbol{a})\right),$$

一様バウンド1 (Aに最小ノルム解のノルムxαを代入)
経験誤差=0とは限らない解の中で一様バウンドを取る.

$$T(A,N,n,d) \equiv \sup_{(N/d)\|\boldsymbol{a}\|_2^2 \le A, \widehat{R}_n(\boldsymbol{a})=0} R(\boldsymbol{a}).$$

一様バウンド2 (Aに最小ノルム解のノルムxαを代入)
経験誤差=0の中で一様バウンドを取る．$\alpha > 1$の時は最小ノルム解以外も含まれる.

ランダム特徴モデル

$$\mathcal{F}_{\mathrm{RF}}(\boldsymbol{\Theta}) \equiv \left\{ f(\boldsymbol{x}) = \sum_{j=1}^N a_j \sigma\left(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle / \sqrt{d}\right) : \boldsymbol{a} \in \mathbb{R}^N \right\}$$

N:特徴数, d:入力の次元, n:サンプルサイズ

ランダム行列のアプローチ

ランダム行列理論による漸近解析

- Hastie et al.: Surprises in High-Dimensional Ridgeless Least Squares Interpolation, Ann. Statist. 50(2): 949--986.

$y_i = x_i^\top \beta^* + \epsilon_i$ $\hat{\beta} = (X^\top X)^+ X^\top Y$: 最小ノルム補間推定量

proportional limit $n, d \to \infty, d/n \to \gamma$における漸近リスク

$$R(\gamma) := \lim_{n\to\infty} \mathbb{E}_\epsilon[\|\hat{\beta} - \beta^*\|_\Sigma^2 | X] =?$$

Theorem

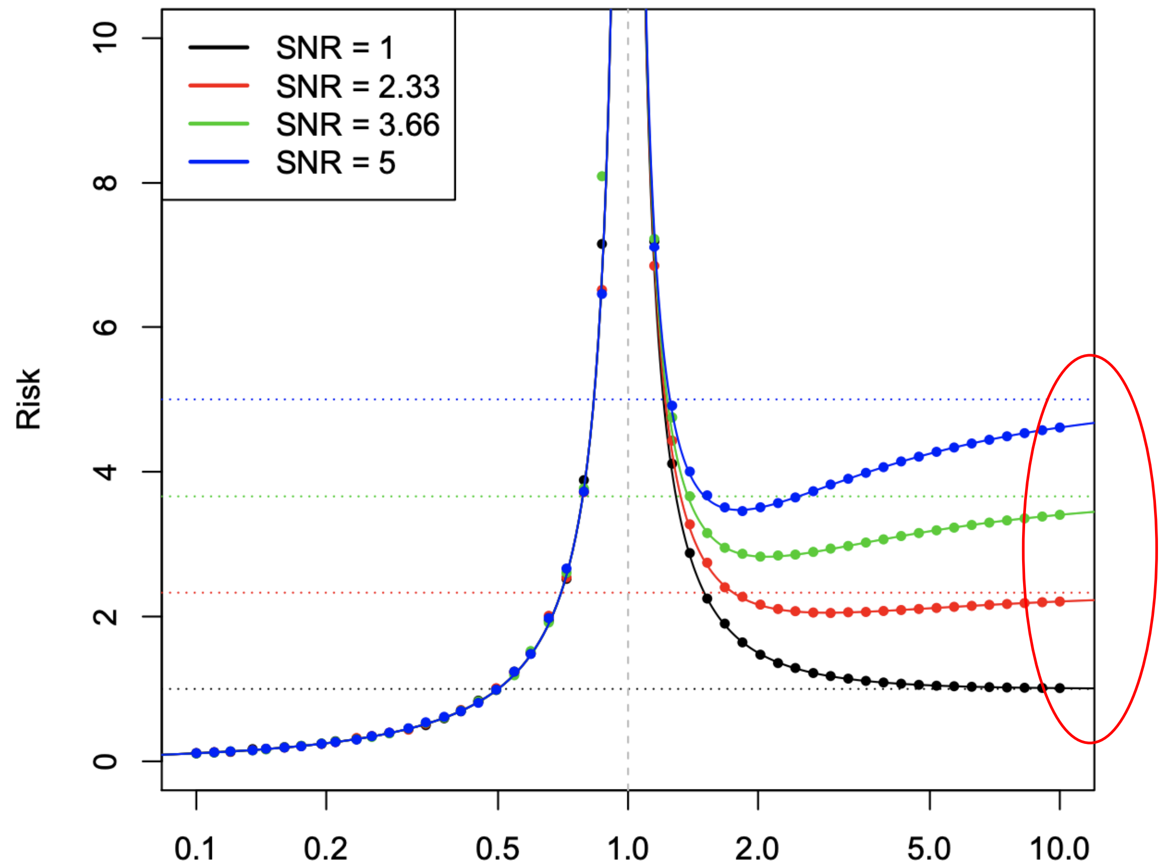
$r^2 := \|\beta^*\|^2$, $\sigma^2 := \mathbb{E}[\epsilon_i^2]$

$n, d \to \infty$かつ$d/n \to \gamma \in (0, \infty)$ (proportional limit)
の極限で期待予測誤差は以下の値に概収束する:

$$R(\gamma) = \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & (\gamma < 1) \\ \underbrace{r^2\left(1 - \frac{1}{\gamma}\right)}_{\text{Bias}} + \underbrace{\sigma^2 \frac{1}{\gamma - 1}}_{\text{Variance}} & (\gamma > 1) \end{cases}$$

$$\mathbb{E}_{x,D_n}[(x^\top\beta^* - x^\top\hat{\beta})^2] = \underbrace{\mathbb{E}_x[(x^\top\beta^* - x^\top\mathbb{E}[\hat{\beta}])^2]}_{B:\ \text{Bias}} + \underbrace{\mathrm{tr}[\mathrm{Cov}(\hat{\beta})\Sigma_x]}_{V:\ \text{Variance}}$$

Numerical Experiment



(d/n)

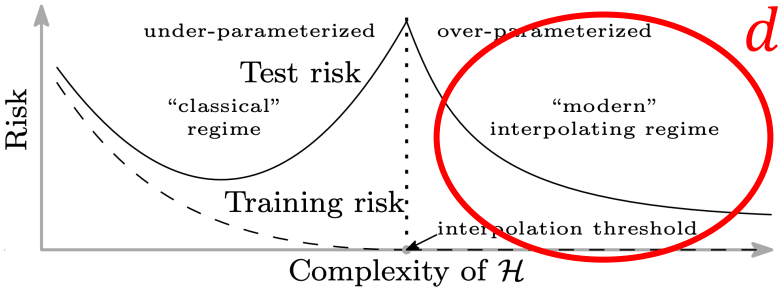
Bias = $||\beta^*||^2$, Variance = 0.
※ここでは$\hat{\beta} = 0$である．これをもって「推定している」とは言いにくいので注意．

$r^2 := \|\beta^*\|^2,\ \sigma^2 := \mathbb{E}[\epsilon_i^2]$

$$R(\gamma) = \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & (\gamma < 1) \\ \underbrace{r^2\left(1 - \frac{1}{\gamma}\right)}_{\text{Bias}} + \underbrace{\sigma^2 \frac{1}{\gamma - 1}}_{\text{Variance}} & (\gamma > 1) \end{cases}$$

前処理付き勾配法

Amari, Ba, Grosse, Li, Nitanda, Suzuki, Wu, Xu: When Does Preconditioning Help or Hurt Generalization? ICLR2021.



$d \gg n$: overparameterized regime

$$y_i = x_i^\top \beta^* + \epsilon_i$$

$$\min_{\beta \in \mathbb{R}^d} \underbrace{\|\beta\|_{P^{-1}}^2}_{=\beta^\top P^{-1} \beta} \quad \text{s.t.} \quad y_i = x_i^\top \beta \quad (\forall i \in [n])$$

(interpolation)

Q: Pによって予測性能がどのように影響受けるか？

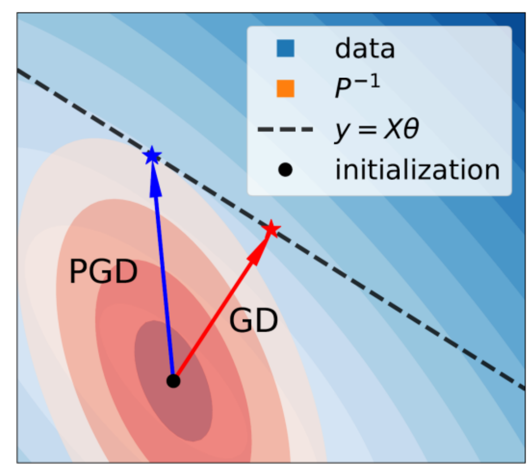
$$\frac{\mathrm{d}\beta(t)}{\mathrm{d}t} = -PX^\top (Y - X\beta(t))/n$$

Preconditioned Gradient Descent

$P = I$: Gradient descent (GD)

$P = \Sigma_x^{-1}$: Natural Gradient descent (NGD)

(真の分布で期待値取ったFisher情報行列)



最適な前処理行列

Bias-variance分解

$$\mathbb{E}_{x,D_n}[(x^\top\beta^* - x^\top\hat{\beta})^2] = \underbrace{\mathbb{E}_x[(x^\top\beta^* - x^\top\mathbb{E}[\hat{\beta}])^2]}_{B:\ \mathrm{Bias}} + \underbrace{\mathrm{tr}[\mathrm{Cov}(\hat{\beta})\Sigma_x]}_{V:\ \mathrm{Variance}}$$

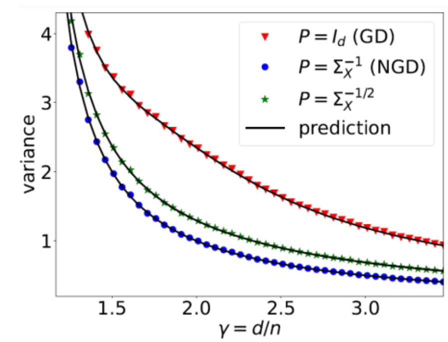
定理 (informal) [Amari, Ba, Grosse, Li, Nitanda, Suzuki, Wu, Xu: When Does Preconditioning Help or Hurt Generalization? ICLR2021]

$d/n \to \gamma > 1$ $(n \to \infty)$の極限における予測誤差の漸近値を厳密に導出

1. バリアンス:

$P = \Sigma_x^{-1}$ (Fisher情報行列) がバリアンスを最小化.

NGDがバリアンスの意味で最適.



(dが増えるほどバリアンスが減る)

2. バイアス:

No free-lunch: 事前に最適なPは決定できない:

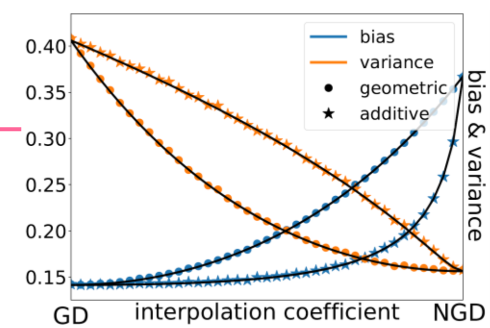
- 真が等方的分布 $\Sigma_{\beta^*} = I$に従っていれば**GD**が良い.
- 真が非等方的分布$\Sigma_{\beta^*} = \Sigma_x^{-1}$に従っていれば**NGD**が良い.

(ベイズの設定: β^*の実現値に関する予測誤差の期待値を比較 $\mathrm{E}[\beta^*\beta^{*\top}] = \Sigma_{\beta^*}$)

GDとNGDの中間が良い

$$\begin{cases} \textbf{Additive: } P = (\alpha\Sigma_x + (1-\alpha)I_d)^{-1} \\ \textbf{Geometric: } P = \Sigma_x^{-\alpha} \end{cases}$$

$\alpha \in [0,1]$



より詳細な結果

(A2) $\Sigma_{XP} := P^{1/2}\Sigma P^{1/2}$のスペクトル分布が$H_{XP}$に弱収束すると仮定.

- **$m(z)$ を自己整合条件を満たす関数とする:** $$\frac{1}{m(z)} = -z + \gamma \int \frac{\tau}{1+\tau m(z)} \mathrm{d}H_{XP}(\tau)$$

$\rightarrow \frac{1}{n}XPX^\top$のスペクトルの漸近分布を表現.

1. バリアンス:

$$V \xrightarrow{\mathrm{p}} \sigma^2 \left(\lim_{\lambda \to +0} m'(-\lambda) m^{-2}(-\lambda) - 1 \right)$$

$V \geq \sigma^2(\gamma-1)^{-1}$ and equality holds by $P = \Sigma^{-1}$

(A3) PとΣが同じ固有ベクトルUを共有.

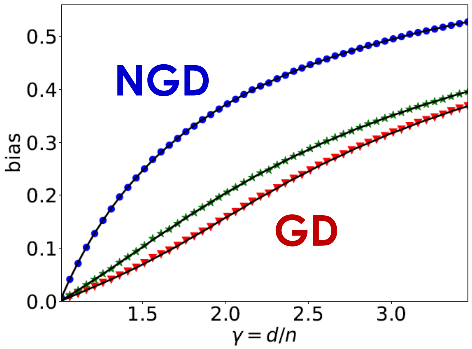
2. バイアス:

$$\mathbb{E}_{\beta^*}[B] \xrightarrow{\mathrm{p}} \lim_{\lambda \to +0} m'(-\lambda) m^{-2}(-\lambda) \mathbb{E}[v_x v_\theta (1 + v_{xp} m(-\lambda))^{-2}]$$

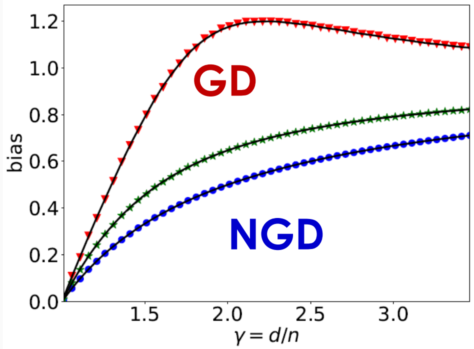
ただし，(e_x, e_θ, e_{xp})は $\Sigma, \Sigma_{XP}, \mathrm{diag}(U^\top \Sigma_{\beta^*} U)$の固有値で$(v_x, v_\theta, v_{xp})$に弱収束するものとする.

Experiments

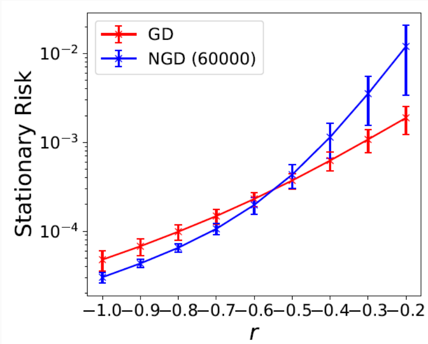
Bias:



$\Sigma_\theta = \boldsymbol{I}_d$ (isotropic).



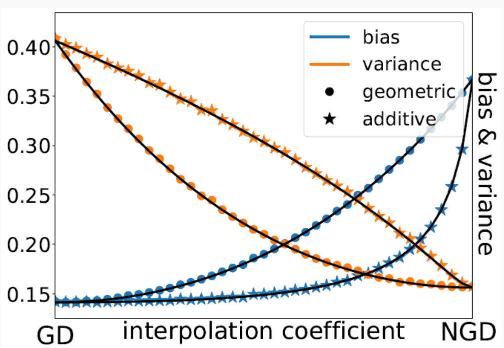
$\Sigma_\theta = \Sigma_X^{-1}$ (misaligned).



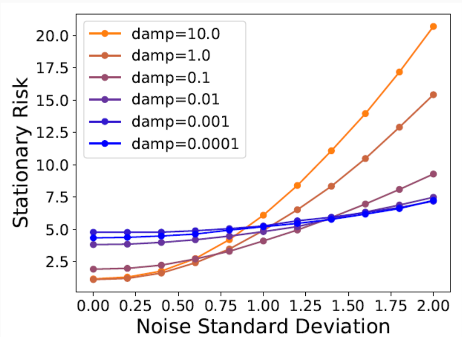
Two-layer MLP (MNIST).

Bias/Variance trade-off:

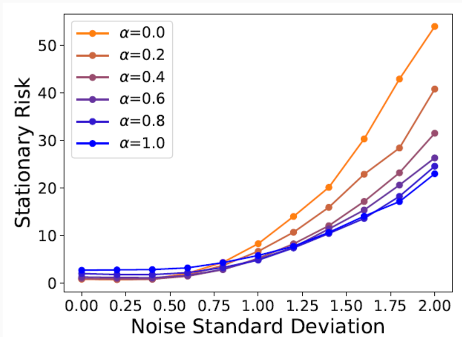
$$\begin{cases} \textbf{Additive: } P = (\alpha\Sigma_x + (1-\alpha)I_d)^{-1} \\ \textbf{Geometric: } P = \Sigma_x^{-\alpha} \end{cases}$$



Monotonicity of bias/variance.



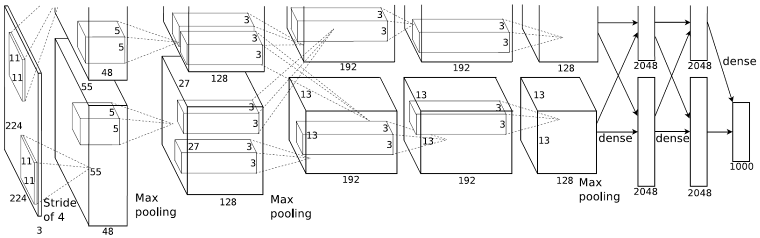
Additive interpolation (MLP).

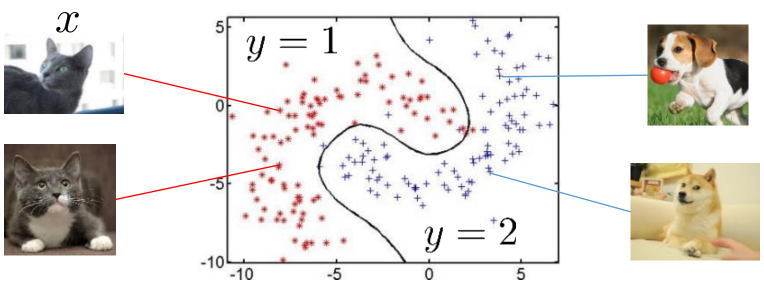


Geometric interpolation (MLP).

第3部
深層学習の最適化
-非凸性の影響-

深層学習の“学習”





深層ニューラルネットワークをデータにフィットさせるとは？

$$L(W) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(W)$$

W: パラメータ

i番目のデータで正解していれば小さく，間違っていれば大きく

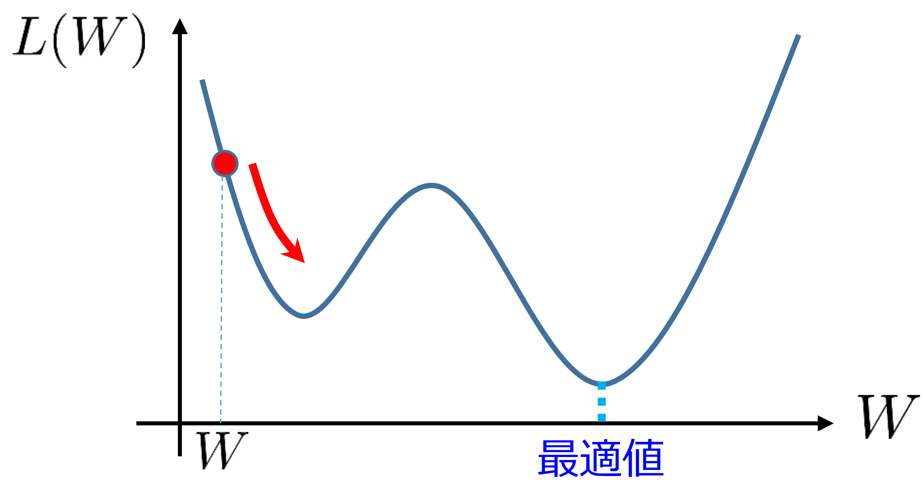
損失関数：データへの当てはまり度合い

損失関数最小化

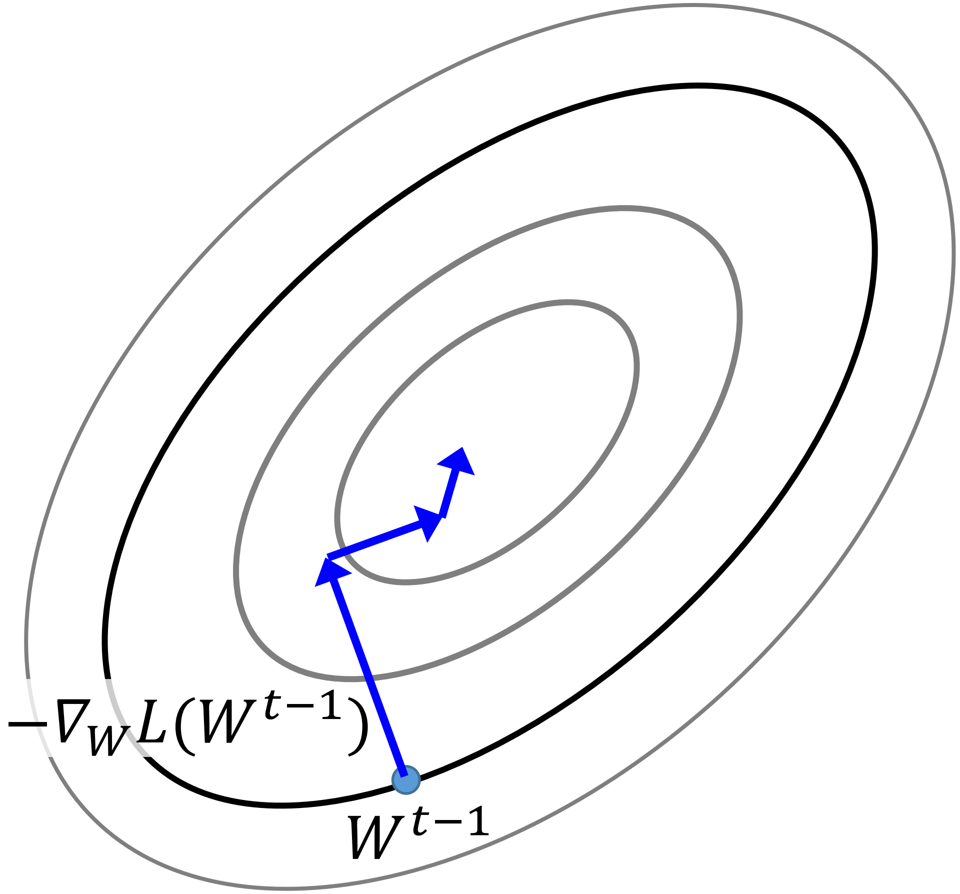
$$\min_{W} L(W)$$

(Wは数十億次元)

通常，**確率的勾配降下法**で最適化



勾配降下法

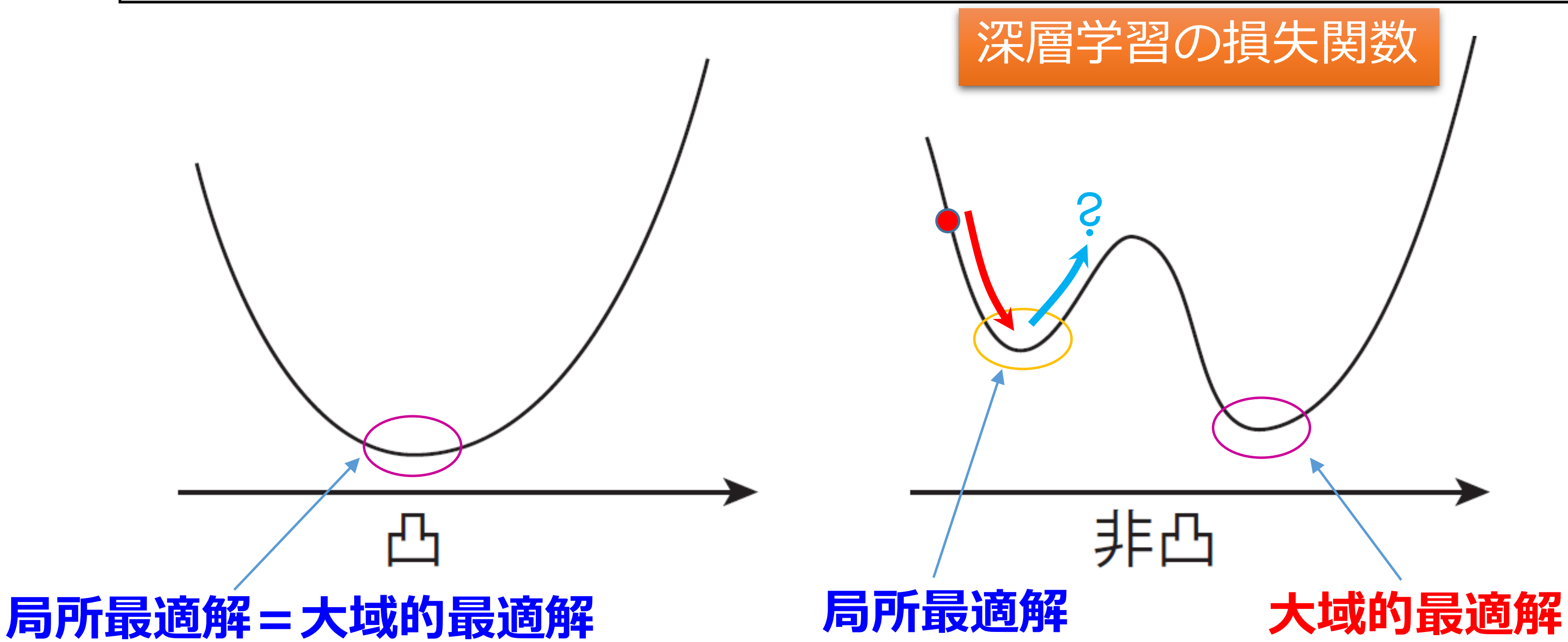


$$W^t = W^{t-1} - \eta \nabla_W L(W^{t-1})$$

問題点

目的関数が非凸関数

凸関数 $\boxed{\theta f(x) + (1-\theta) f(y) \geq f(\theta x + (1-\theta) y) \quad (\forall x, y \in \mathbb{R}^p, \theta \in [0,1])}$



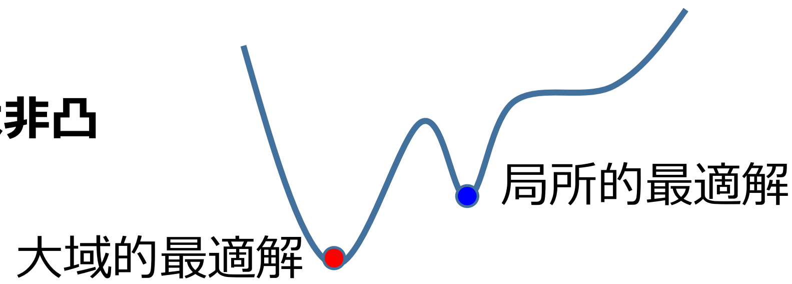
局所最適解や鞍点にはまる可能性あり

“**狭い**”ネットワークの学習はNP-完全:

- Judd (1988), Neural Network Design and the Complexity of Learning.
- Blum&Rivest (1992), Training a 3-node neural network is NP-complete.

局所最適性

深層学習の目的関数は非凸



- <u>線形深層NN</u>の局所的最適解は全て大域的最適解：
  Kawaguchi, 2016; Lu&Kawaguchi, 2017.

  ※ただし対象は<u>線形NN</u>のみ.

  → 臨界点が大域的最適解であることの条件も出されている
  (Yun, Sra&Jadbabaie, 2018)

- 低ランク行列補完の局所的最適解は全て大域的最適解：
  Ge, Lee&Ma, 2016; Bhojanapalli, Neyshabur&Srebro, 2016.

$$\min_{U \in \mathbb{R}^{M \times k}} \sum_{(i,j) \in E} (Y_{i,j} - (UU^\top)_{i,j})^2$$

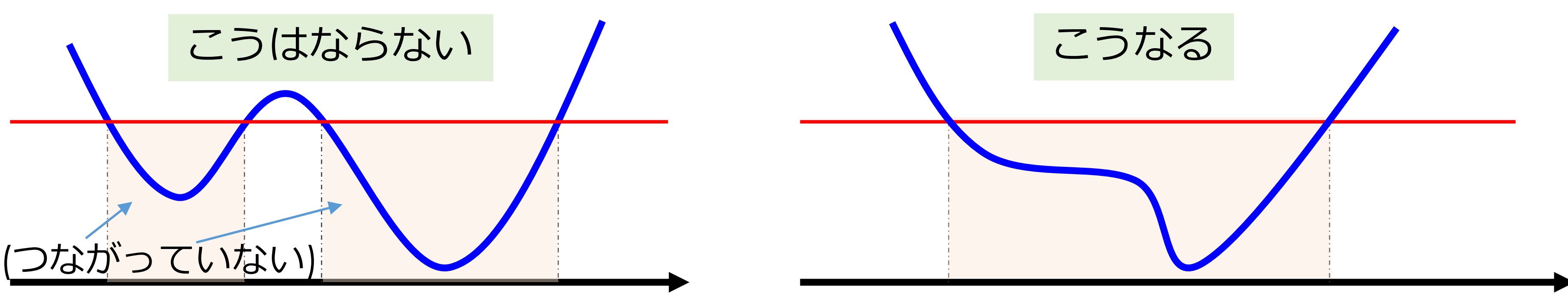
Loss landscape

- 横幅の広いNNの訓練誤差には孤立した局所最適解がない．（局所最適解は大域的最適解とつながっている） ※とはいえ，勾配法で大域的最適解に到達可能かは別問題．

定理

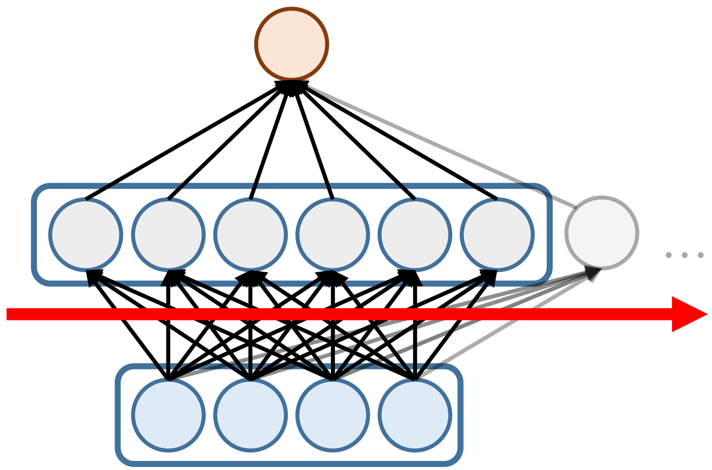
n個の訓練データ$(x_i, y_i)_{i=1}^n$が与えられているとする．損失関数ℓは凸関数とする．
任意の連続な活性化関数について，横幅がデータサイズより広い（$M \geq n$）二層NN$f_{(a,W)}(x) = \sum_{m=1}^M a_m \eta(w_m^\top x)$に対する訓練誤差$\hat{L}(a,W) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_{(a,W)}(x_i))$の任意のレベルセットの弧状連結成分は大域的最適解を含む．言い換えると，任意の局所最適解は大域的最適解である．

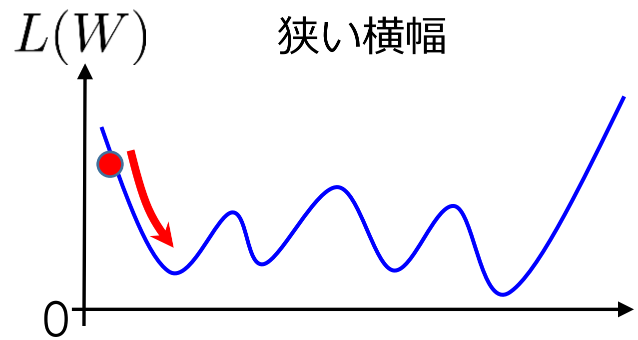
[Venturi, Bandeira, Bruna: Spurious Valleys in One-hidden-layer Neural Network Optimization Landscape JMLR, 20:1-34, 2019.]

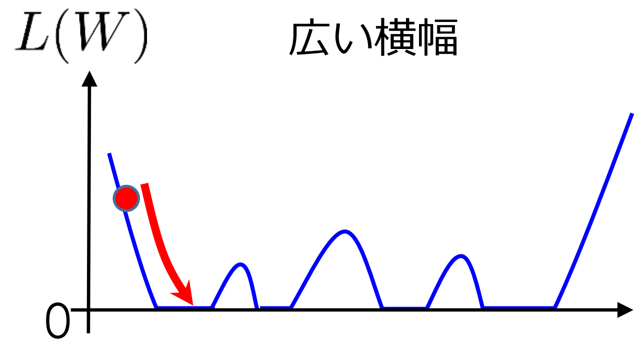


オーバーパラメトライゼーション

横幅が広いと局所最適解が大域的最適解になる.







自由度が高いので，目的関数を減少させる方向が見つけやすい.

- 二種類の解析手法
  - Neural Tangent Kernel (NTK)
  - Mean-field analysis (平均場解析)

二つのスケーリング

$$f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

- Neural Tangent Kernelのregime (lazy learning)
  - ➢ $a_j = O(1/\sqrt{M})$ [Jacot+ 2018][Du+ 2019][Arora+ 2019] (Xavier initialization/He initialization)
- 平均場解析の設定 (mean field)
  - ➢ $a_j = O(1/M)$ [Nitanda & Suzuki (2017), Chizat & Bach (2018), Mei, Montanari, & Nguyen (2018)]

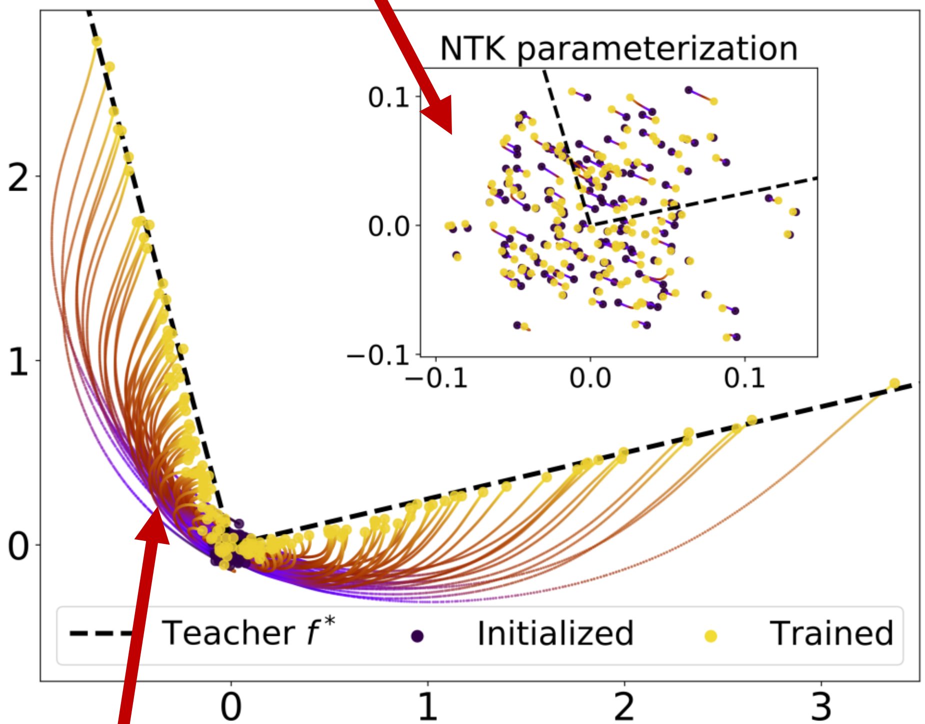
初期化のスケーリングによって，初期値と比べて学習によって動く大きさの割合が変わる.
→ 学習のダイナミクス，汎化性能に影響

（解析の難しさも違う）

変数のスケールと特徴学習

$$f(x) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} r_j \sigma(w_j^\top x)$$

NTK: Large scale initialization → features are (almost) freezed.



二層NNにおける一層目のパラメータの軌跡

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

[Ba et al., 2022]

Mean field: Small scale initialization → features need to move significantly.

$$f(x) = \frac{1}{M} \sum_{j=1}^{M} r_j \sigma(w_j^\top x)$$

(参考) 初期化スケールと学習率の取り方

- ABCパラメトライゼーション [Yang&Hu, 2021]

$$x^l(\xi) = \phi(h^l(\xi)) \in \mathbb{R}^n, \quad h^{l+1}(\xi) = W^{l+1}x^l(\xi) \in \mathbb{R}^n, \quad \text{for } l = 1, \ldots, L-1,$$ n:横幅

(1) パラメータ設定

$$W^l = n^{-a_l} w^l$$

(w^lが学習パラメータ)

A

(2) 初期化法

$$w^l_{\alpha\beta} \sim \mathcal{N}(0, n^{-2b_l})$$

B

(3) 学習率のスケール

$$\eta n^{-c}$$

C

$$h^1 = W^1\xi \in \mathbb{R}^n, x^l = \phi(h^l) \in \mathbb{R}^n, h^{l+1} = W^{l+1}x^l \in \mathbb{R}^n, f(\xi) = W^{L+1}x^L$$

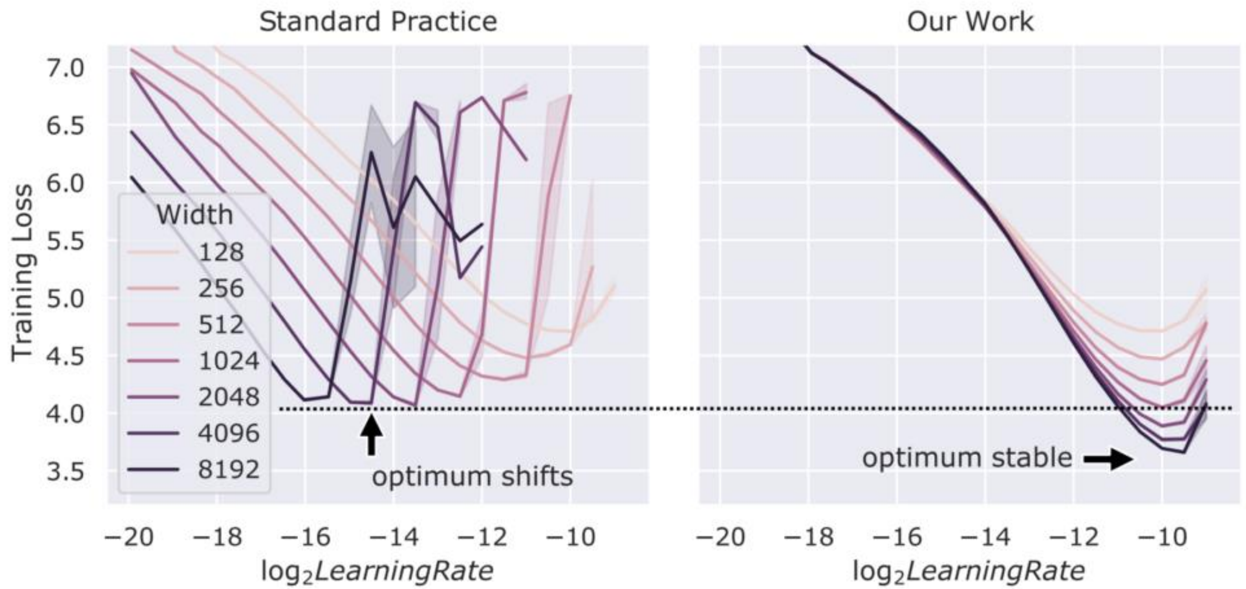
	Definition	SP (w/ LR $\frac{1}{n}$)	NTP	MFP ($L=1$)	μP (ours)
a_l	$W^l = n^{-a_l} w^l$	0	$\begin{cases} 0 & l=1 \\ 1/2 & l \geq 2 \end{cases}$	$\begin{cases} 0 & l=1 \\ 1 & l=2 \end{cases}$	$\begin{cases} -1/2 & l=1 \\ 0 & 2 \leq l \leq L \\ 1/2 & l=L+1 \end{cases}$
b_l	$w^l_{\alpha\beta} \sim \mathcal{N}(0, n^{-2b_l})$	$\begin{cases} 0 & l=1 \\ 1/2 & l \geq 2 \end{cases}$	0	0	$1/2$
c	$LR = \eta n^{-c}$	1	0	-1	0
r	Definition 3.2	$1/2$	$1/2$	0	0
$2a_{L+1} + c$		1	1	1	1
$a_{L+1} + b_{L+1} + r$		1	1	1	1
Nontrivial?		✓	✓	✓	✓
Stable?		✓	✓	✓	✓
Feature Learning?				✓	✓
Kernel Regime?		✓	✓		

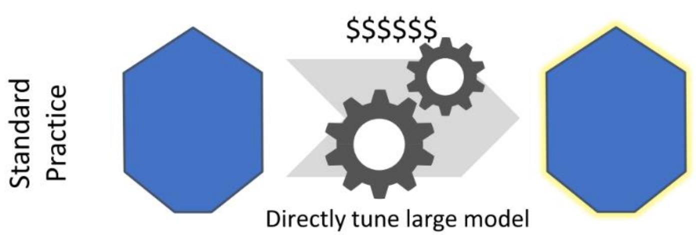
(適切なスケーリング)

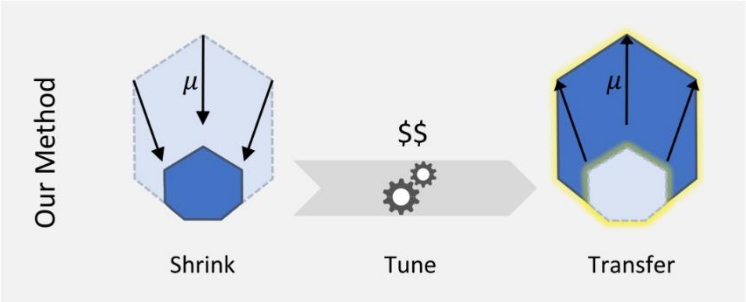
[Yang&Hu: Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks. ICML2021.]

[Yang et al.:Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. arXiv:2203.03466]

小さいモデルのハイパーパラメータを大きなモデルに転用できる．
GPTの学習に利用．数億円の学習コストを抑えられる．







Neural Tangent Kernel

[Jacot, Gabriel, & Hongler (2019)]

$f_W(x) = \sum_{j=1}^M a_j \eta(w_j^\top x)$

重要

テイラー展開　　特徴写像とみなせる

$$f_W(x) \simeq (W - W^{(0)})^\top \nabla_W f_{W^{(0)}}(x)$$ (線形モデル)

初期値のスケールが大きいので，初期値周りの線形近似でデータにフィットできてしまう.

カーネル関数は特徴写像の内積

$$\Rightarrow \quad k_W(x,x') = \langle \nabla_W f_{W^{(0)}}(x), \nabla_W f_{W^{(0)}}(x') \rangle$$

$$= \sum_{j=1}^M \underbrace{a_j^2}_{\frac{1}{M}} (x^\top x')\, \eta'(w_j^\top x)\eta'(w_j^\top x')$$

Neural Tangent Kernel

各ニューロンの微分で得られる
特徴写像の内積を全ニューロンで平均

最適化のダイナミクスや汎化性能などは，
NTKをカーネル関数とするカーネル法としてとらえられる．

Optimization in NTK regime

以下のように初期化する:

- $a_j \sim (\pm 1)\frac{1}{\sqrt{M}}$ (+, − is generated evenly)
- $w_j \sim N(0, I)$

$$f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j^\top x)$$

Theorem [Arora et al., 2019]

$M = \Omega(n^2 \log(n)/\lambda_{\min})$とすれば, 勾配法によって大域的最適解へ**線形収束**し，その汎化誤差は$\sqrt{\boldsymbol{y}^\top (K_{W(0)})^{-1} \boldsymbol{y}/n}$で抑えられる.

$\exp(-ct)$で収束

See also[Du et al., 2018; Allen-Zhu, Li & Song, 2018; Li & Liang, 2018]

- 訓練誤差0の解に線形収束する.
- 汎化誤差も一応抑えられている.

- データに完全にフィットさせてしまうので過学習の可能性あり.
- Early stoppingや正則化を入れれば過学習を防げる. (次ページ)

NTKの収束定理の導出

連続時間ダイナミクスを考える.

Model： $f_W(x) = \sum_{j=1}^{M} a_j \eta(w_j{}^\top x)$

- a_j は固定
- w_j を学習

[Jacot, Gabriel&Hongler, NeurIPS2018]

$$\frac{\mathrm{d}w_j}{\mathrm{d}t} = -\nabla_{w_j} \hat{L}(f_W)$$ (Gradient descent, GD)

$$= -\frac{1}{n}\sum_{i=1}^{n} \ell_i'(f_W(x_i)) a_j \nabla_{w_j}\eta(w_j^\top x_i)$$

$$\nabla_{w_j}\eta(w_j^\top x_i) = x_i \eta'(w_j^\top x_i)$$

(勾配法による更新)

$$\frac{\mathrm{d}f_W(x)}{\mathrm{d}t} = \sum_{j=1}^{M} a_j \nabla_{w_j}^\top \eta(w_j^\top x) \frac{\mathrm{d}w_j}{\mathrm{d}t}$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\sum_{j=1}^{M} a_j^2 \nabla_{w_j}^\top \eta(w_j^\top x)\nabla_{w_j}\eta(w_j^\top x_i)}_{k_W(x,x_i)}\right)\underbrace{\ell_i'(f_W(x_i))}_{\text{residual}}$$

O(1/M)：
特徴写像の内積の平均

(関数勾配)

$k_W(x, x_i)$

Neural Tangent Kernel

目的関数の減少速度

参考

$$\frac{\mathrm{d}\hat{L}(f_W)}{\mathrm{d}t} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{d}f_W(x_i)}{\mathrm{d}t}\ell_i'(f_W(x_i))$$

$$= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ell_i'(f_W(x_i))\underbrace{k_W(x_i,x_j)}_{(K_W)_{i,j}}\ell_j'(f_W(x_j))$$

$$= -\frac{1}{n^2}\|\nabla_f\hat{L}(f_W)\|_{K_W}^2$$

$$\leq -\lambda_{\min}\frac{1}{n^2}\|\nabla_f\hat{L}(f_W)\|^2$$

($\lambda_{\min}$: グラム行列の**最小固有値**)

Fact [Du et al., 2018; Allen-Zhu, Li & Song, 2018]

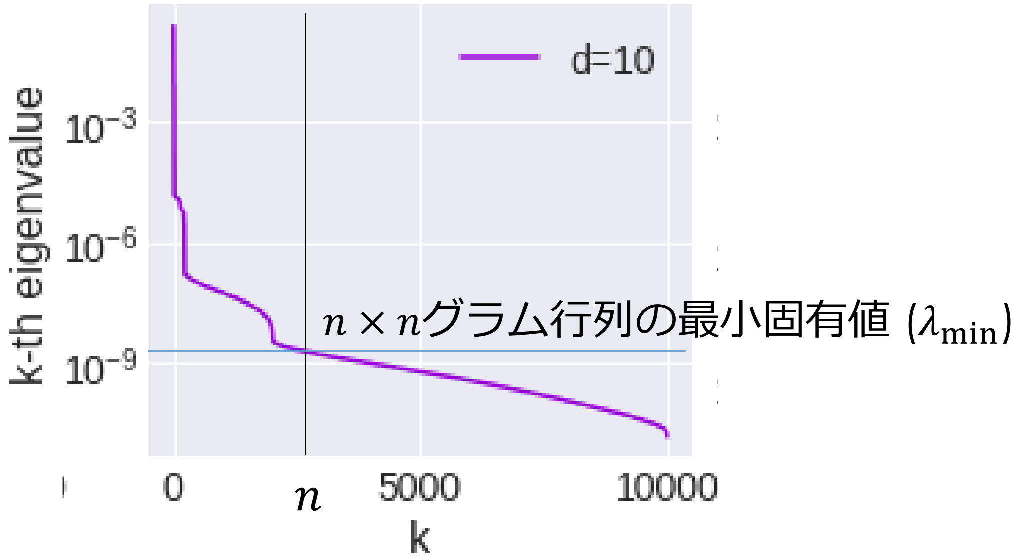
- ランダム初期化しておけば, $K_{W^{(0)}} \succ \epsilon I$ が高確率で成立.
- 最適化の最中に最小固有値は正のまま ($\geq \epsilon/2$).

⇒ **線形収束** ($\exp(-\lambda_{\min} t)$)

Spectral bias

- 最適化の観点からはoverparameterizationは有用に見える.
- 汎化誤差はどうであろうか?

NTKの固有値分布



- グラム行列の最小固有値は小さい ($1/\mathrm{poly}(n)$).
- 固有値の減少レートは多項式オーダー (理論+実験).

→ Spectral bias: 汎化の意味では好ましい.

Kernelによる平滑化という視点

- Frechet 微分 in $L_2(P_n)$: $\nabla_f \hat{L}(f)$

$$\nabla_f \hat{L}(f) = (\ell_i'(f(x_i)))_{i=1}^n$$

$$\hat{L}(f+h) = \hat{L}(f) + \langle \nabla_f \hat{L}(f), h \rangle_{L_2(P_n)} + o(\|h\|_{L_2(P_n)}^2)$$

- **平滑化**積分作用素:

$$T_k f(x) := \int k(x,x') f(x') \mathrm{d}P_n(x')$$

$$\boxed{T_{k_W} \phi_j = \mu_j \phi_j}$$

- NTKにおける勾配は関数勾配を平滑化したもの:

$$\frac{\mathrm{d} f_W}{\mathrm{d}t} = -T_{k_W} \nabla_f \hat{L}(f_W)$$

勾配を平滑化！

$$\left(= -\frac{1}{n}\sum_{i=1}^n k_W(\cdot, x_i) \ell_i'(f_W(x_i))\right)$$

(導出は前のページ参照「NTKの収束定理の導出」)

k_Wが高周波成分に小さな固有値を持てば，T_{k_W}は平滑化作用素として働く
→ 帰納的バイアス (inductive bias).

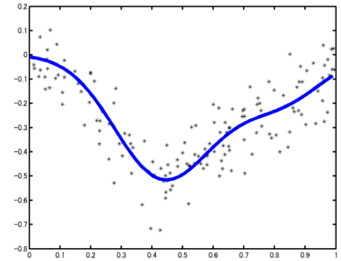
NTK regimeでのSGD

参考

モデル:

$$f_{a,W}(x) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} a_j \eta(w_j{}^\top x)$$

(We train both of first and second layers)



目的関数:

$$L(a,W) = \underbrace{\mathbb{E}[(Y - f_{a,W}(X))^2]}_{\text{期待損失}} + \underbrace{\frac{\lambda}{2}(\|a - a^{(0)}\|^2 + \|W - W^{(0)}\|_{\mathrm{F}}^2)}_{\text{初期値からのずれ}}$$

$Y = f^*(X) + \epsilon$ (ノイズありの観測)

Averaged Stochastic Gradient Descent

for $t = 0$ **to** $T-1$ **do**

Randomly draw a sample $(x_t, y_t) \sim \rho$

Perform SGD update for all $j \in \{1, \dots, M\}$:

$$a_j^{(t+1)} = a_j^{(t)} - \alpha_t[\nabla_a \ell(y_t, f_{a^{(t)},W^{(t)}}(x_t)) + \lambda(a^{(t)} - a^{(0)})]$$

$$W_j^{(t+1)} = W_j^{(t)} - \alpha_t[\nabla_W \ell(y_t, f_{a^{(t)},W^{(t)}}(x_t)) + \lambda(W^{(t)} - W^{(0)})]$$

end for

Return $\bar{a}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} a^{(t)}$, $\bar{W}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1} W^{(t)}$.

NTKにおける余剰誤差の速い収束

[Nitanda&Suzuki: Fast Convergence Rates of Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime, 2020.]

仮定：真の関数がNTKの作るRKHSに入っているとする.

NTK設定で適切な正則化を入れたSGDは“速い学習レート”を達成できる.
→ NTKによるsmoothingのおかげ.

Thm (速い収束レート)

f_T: T回更新後の解

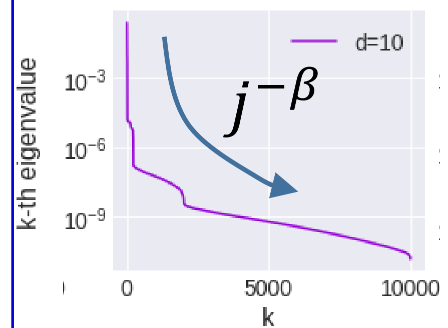
NTKの固有値の減衰レート

$$\mathbb{E}[\|f_T - f^*\|_{L_2}^2] \leq \epsilon_M + O(T^{-\frac{2r\beta}{2r\beta+1}})$$

$M \to \infty$で0に収束する項

速い学習レート
($O(1/\sqrt{T})$より速い)

NTKのスペクトル



→ $T^{-\frac{2r\beta}{2r\beta+1}}$ は**ミニマックス最適レート.**
(各種パラメータの意味は次ページに詳細)

仮定

参考

2層NNのNTK:

$$k_\infty(x,x') = \mathbb{E}_{w^{(0)}}[\eta(w^{(0)\top}x)\eta(w^{(0)\top}x')] + \mathbb{E}_{w^{(0)}}[\eta'(w^{(0)\top}x)\eta'(w^{(0)\top}x')x^\top x]$$

横幅無限における積分作用素:

$$T_{k_\infty}f(x) = \int k_\infty(x,x')f(x')\mathrm{d}P_X$$

population

スペクトル分解: $T_{k_\infty}\phi_j = \mu_j\phi_j$, $k_\infty(x,x') = \sum_{j=1}^\infty \mu_j\phi_j(x)\phi_j(x')$

仮定

- $f^*(x) = \mathrm{E}[Y|X=x]$ が次のように書ける:
  $$T_{k_\infty}^r h = f^*$$
  for $h \in L_2(P_X)$, and $r \in [1/2,1]$.
- 固有値減衰条件:
  $$\mu_j = O(j^{-\beta}).$$

真の関数の平滑性

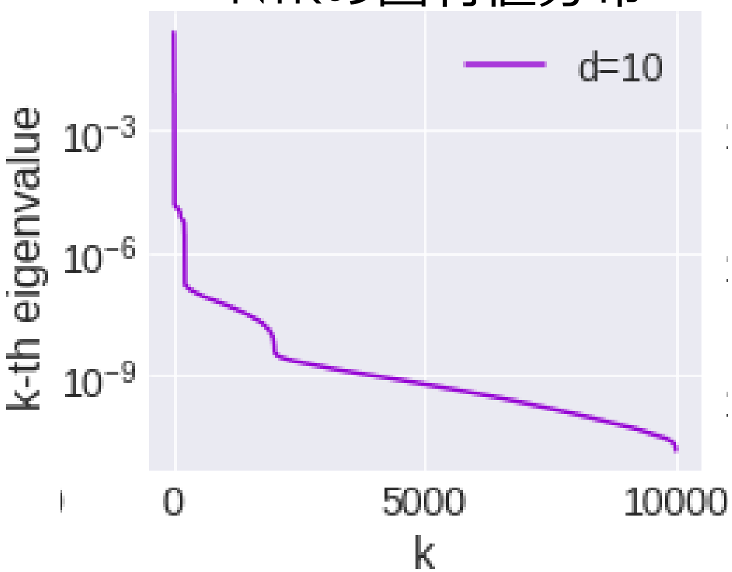
カーネル関数の"複雑さ"

カーネルリッジ回帰の解析における標準的な仮定; see, e.g., Dieuleveut et al. (2016); Caponnetto and De Vito (2007) (rの条件はやや強め).

$$k_\infty(x,x') = \sum_{m=1}^{\infty} \lambda_m \phi_m(x)\phi_m(x')$$

NTKの固有値固有関数分解
$(\phi_m)_{m=1}^{\infty}$: 固有関数. $L_2(P_X)$内の正規直交基底.

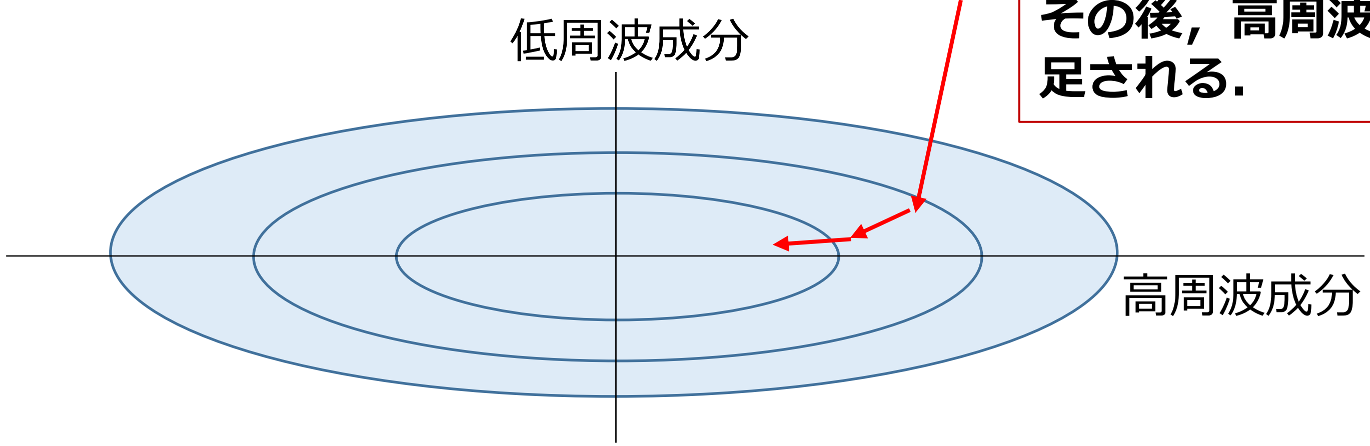
NTKの固有値分布



実際のNTKの固有値は多項式オーダーで減衰する.

[Bietti&Mairal (2019); Cao et al. (2019); Ronen et al. (2019)]

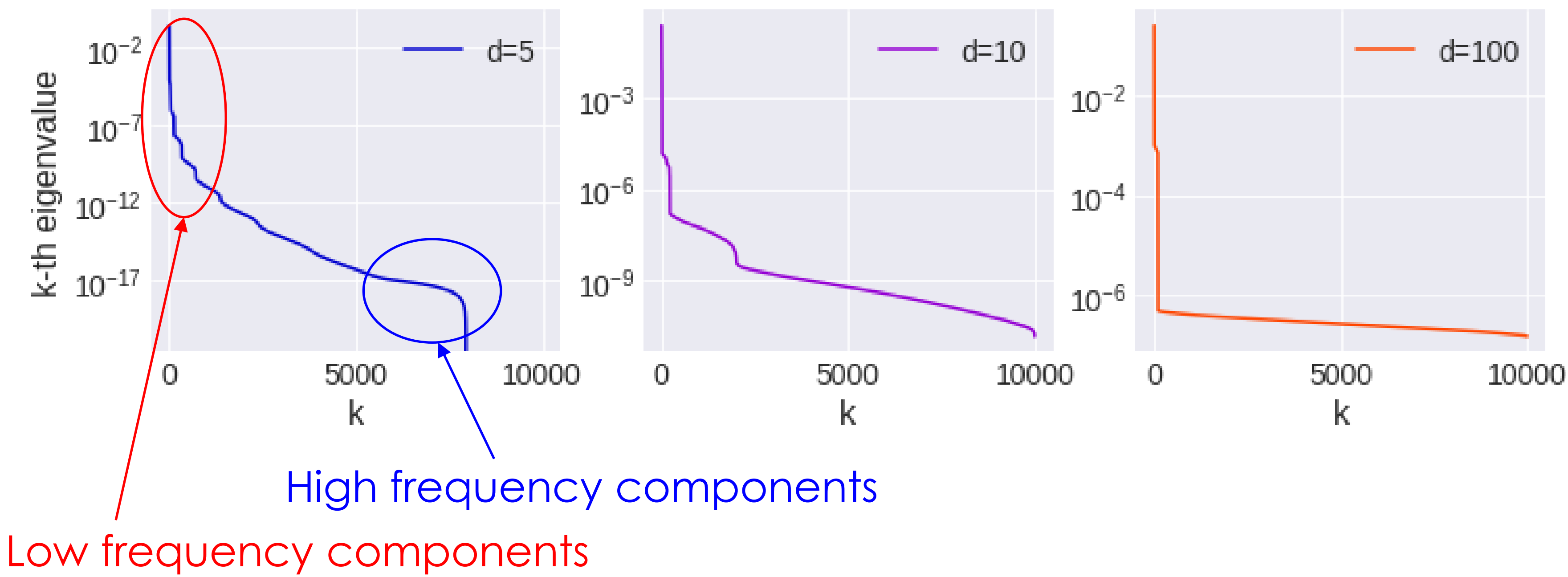
低周波成分が最初に補足される. その後，高周波成分が徐々に補足される.



NTKの陰的正則化

参考

固有値減少の数値実験による検証



理論

ReLU, $(a, w) \sim N(0, I)$:

$$k_{\mathrm{NTK}}(x, y) = \sum_{k=0}^{\infty} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(x) Y_{k,j}(y)$$

- $Y_{k,j}$: spherical harmonics functions with degree k.
- $\mu_k \sim k^{-d}$.
- $N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}$

See Cho&Saul (2009), Xie,Liang&Song (2017), Back (2017), Bietti&Marial (2019) for inductive bias.

二層NNのNTKはmultiple kernel

参考

$$f_{a,W}(x) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} a_j \eta(w_j{}^\top x)$$

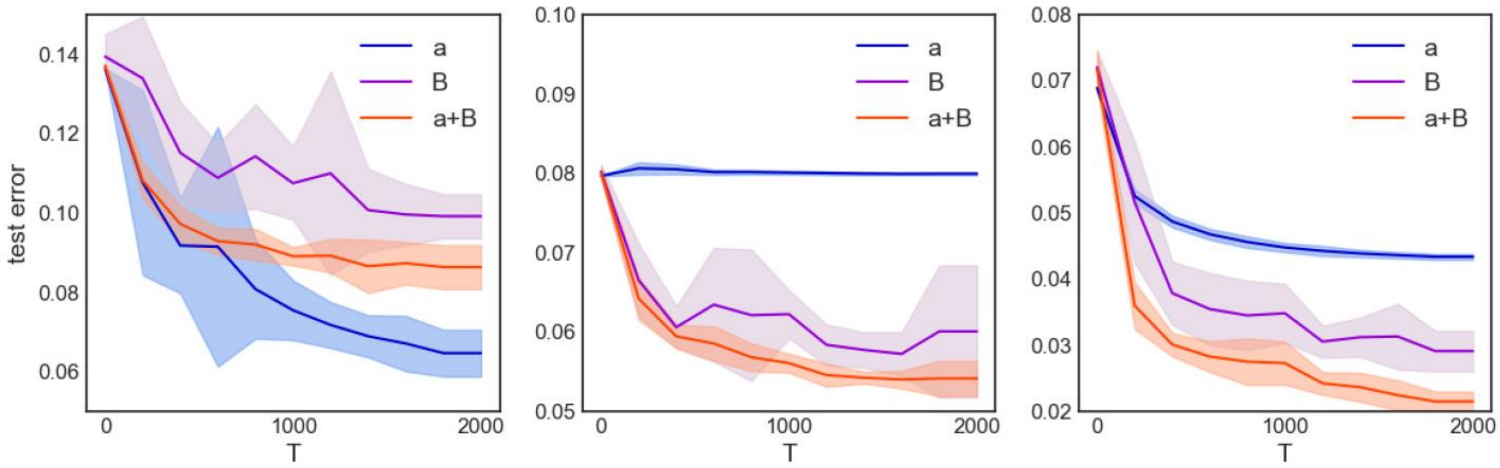
$$k_{a,W}(x,x') = k_a(x,x') + k_W(x,x')$$

二層目のNTK　一層目のNTK

一層目と二層目のカーネルの和：multiple kernel

仮定: f^* is in

RKHS w.r.t. NTK for the 2nd layer.　RKHS w.r.t. NTK for the 1st layer.　RKHS w.r.t. NTK for the both layer.



- 二層NNのNTKによる学習は，multiple kernel learningの効果がある．
- 多層NNを用いることはモデルmisspecificationに対してよりロバストになる．

Beyond kernel

問題点：NTKは解析がしやすいが，結局カーネル法の範疇なので深層学習の“良さ”が現れない．

➢ NTKをはみ出す理論の試みがいくつかなされている．

（今後発展が予想される）

- Allen-Zhu&Li (2019,2020)

  Allen-Zhu&Li: What Can ResNet Learn Efficiently, Going Beyond Kernels? NIPS2019.
  Allen-Zhu&Li: Backward Feature Correction: How Deep Learning Performs Deep Learning. arXiv:2001.04413.

  (ResNet型ネットワークでカーネルを優越する状況)

- Li, Ma&Zhang (2019)

  Li, Ma&Zhang: Learning Over-Parametrized Two-Layer ReLU Neural Networks beyond NTK. arXiv:2007.04596.

  (テンソル分解の理論で深層学習がカーネルを優越することを示した)

- Bai&Lee (2020)

  Bai&Lee: Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks. ICLR2020.

  (二次のテイラー展開まで使う)

2層NNの特徴学習のダイナミクス解析

- (<u>**NTKを超えて**</u>)勾配法によって特徴量がどう学習されるか？

  → 色々な研究がある.
- NTK近似が成り立たない領域では非線形性が強くなり，最適化のダイナミクスの解析が難しくなる.

- 最近では妥協点として，勾配法の最初の段階（1ステップ or 少数ステップ）でどのような特徴量が獲得できるかを解析する研究が複数なされている.
  - 少数ステップで得られる情報は限られているが，それでも予測性能の改善が示せる.
  - 今後はより非線形性の強い特徴量学習のダイナミクスの解析が進むと思われる.

※ 計算量をあまり気にしなければ勾配ランジュバン動力学を用いた学習の解析は完全に非線形な特徴学習をとらえられる.

勾配法とKernel alignment

$$f_{\mathrm{NN}}(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle x, w_i \rangle) = \frac{1}{\sqrt{N}} a^\top \sigma(W^\top x)$$

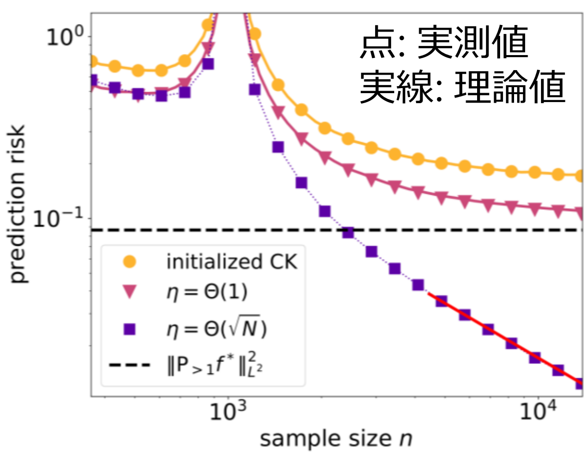
問：勾配法でWを更新することで，データに合った特徴量を獲得できるか？
答：NTK近似が成り立つ領域からはみ出るくらい大きなステップサイズを用いれば，一回の更新で意味のある特徴量の方向を得ることができる．
→ カーネルAlignment，特徴量学習.

$$W_{k+1} = W_k + \eta \sqrt{N} \nabla L(f_{\mathrm{NN}})$$

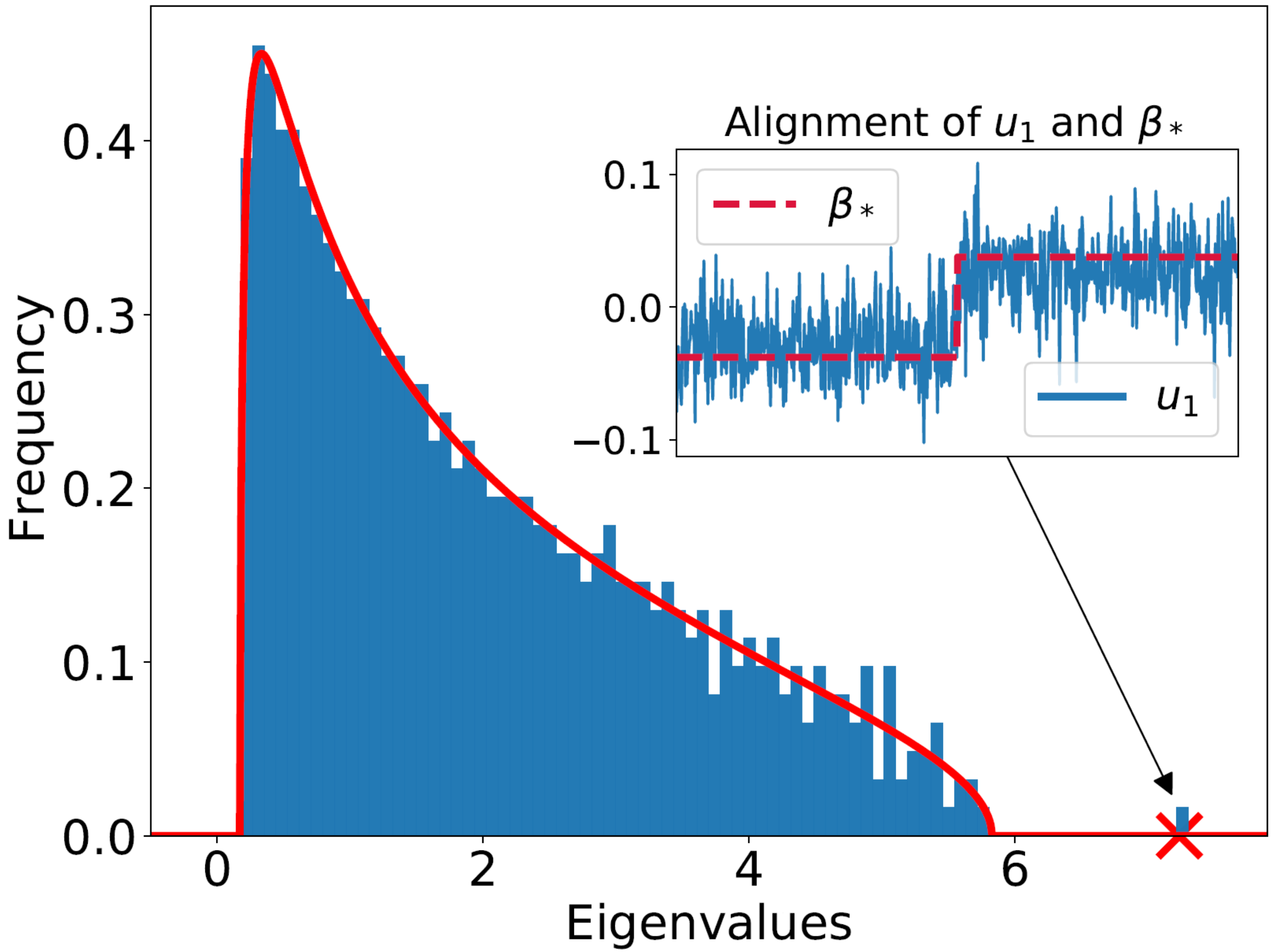
$n, d, N \to \infty$の極限を考え，勾配法１回の更新後の予測誤差を評価してみる．

- $\eta = \sqrt{N}$のように大きな更新を用いると，任意の線形モデルによるリッジ回帰を優越する．
- $\eta = 1$では最適なリッジ回帰を優越しないが初期値Wは優越.
- $\eta = o(1)$では初期値Wと同じ予測誤差 (NTK-regime).

Gaussian equivalence property + Random matrix theory
→ Exact risk evaluation.



勾配法とKernel alignment

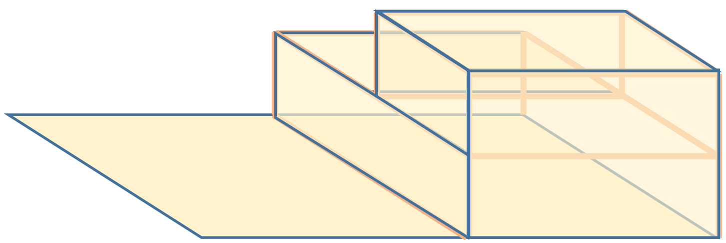


関連研究

The first few step of GD with large learning rate can extract informative features.

- Staircase function [Abbe et al., NeurIPS2021; Abbe et al., arXiv2202.08658]

  Small number of gradient descent can extract nonlinear features to estimate "staircase" function. The trained features for GD can outperform random feature model.



- Benign overfitting with feature learning
  [Cao et al., arXiv:2202.06526; Frei et al., arXiv:2202.05928]

  Gradient descent in two-layer NN can yield benign overfitting and achieves almost the Bayes error in binary classification.

問題設定

観測モデル:

$$y_i = f^*(x_i) + \epsilon_i \quad (i = 1, \dots, n)$$

where $x_i \sim N(0, I)$, $\epsilon_i \sim N(0,1)$, and $x_i \in \mathbf{R}^d$.

➢ 平均場スケールの2-層NNを学習：

$$f_{\mathrm{NN}}(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle x, w_i \rangle) = \overbrace{\frac{1}{\sqrt{N}}}^{\text{平均場レジーム } O(1/N)} a^\top \sigma(W^\top x) \quad (\because a_i = O_p(1/\sqrt{N}))$$

where $a_i \sim N(0, \underbrace{1/N}_{\text{var}})$ and $W_{ij} \sim N(0, \underbrace{1/d}_{\text{var}})$.

経験リスク:

$$\mathcal{L}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

予測リスク:

$$\mathcal{R}(f) = \mathbb{E}[(f^*(X) - f(X))^2]$$

問: 勾配法による特徴学習で予測誤差を改善できるか?

真の関数がシングルインデックスモデルで書ける状況を解析:

$$f^*(x) = \sigma^*(\langle x, \beta^* \rangle)$$

勾配法

- 一層目のパラメータWの更新:

$$W_{t+1} = W_t + \eta\sqrt{N}\underbrace{\left(-\nabla_W \mathcal{L}(f_{\mathrm{NN}}^{(t)})/2\right)}_{G_t}$$

$$G_t = -\frac{1}{n}X^\top\left[\left(\frac{1}{\sqrt{N}}\left(\frac{1}{\sqrt{N}}\sigma(XW_t)a - y\right)a^\top\right)\odot\sigma'(XW_t)\right]$$

✓ 二層目のパラメータaは最適化の途中で固定.
✓ あくまでWの特徴学習のダイナミクスに注目.

1. 特に，勾配法の**“１ステップ更新“** に注目:

$$W_1 = W_0 + \eta\sqrt{N}G_0$$

2. その後，二層目はリッジ回帰で推定: $\phi_{\mathrm{GD}}(x) = \frac{1}{\sqrt{N}}\sigma(W_1^\top x)$

$$\hat{a}_{\mathrm{GD}} = \mathop{\arg\min}_{a\in\mathbb{R}^N}\left\{\frac{1}{n}\sum_{i=1}^n(\tilde{y}_i - \langle a, \phi_{\mathrm{GD}}(\tilde{x}_i)\rangle)^2 + \frac{\lambda}{N}\|a\|^2\right\}$$

$(\tilde{x}_i, \tilde{y}_i)_{i=1}^n$: i.i.d. copy of $(x_i, y_i)_{i=1}^n$

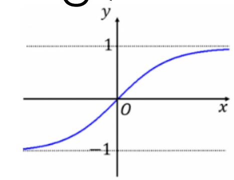
➢ W_1 (1回更新後) は，初期値W_0より真の関数f^*に「相関している」と考えられる．→より良い予測誤差.

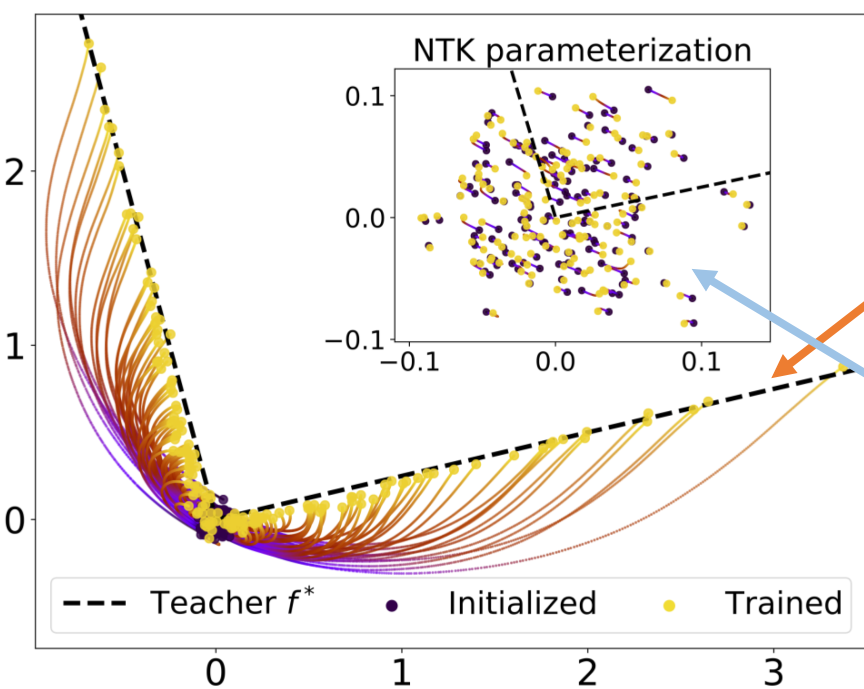
仮定

$$f_{\mathrm{NN}}(x) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} a_i \sigma(\langle x, w_i \rangle) = \frac{1}{\sqrt{N}} a^\top \sigma(W^\top x)$$

1. **Proportional limit:** $n, d, N \to \infty$ and $n/d \to \psi_1,\ N/d \to \psi_2$.
2. **活性化関数の性質:**
   1. σ は有界，$\max\{\|\sigma'\|_\infty, \|\sigma''\|_\infty, \|\sigma'''\|_\infty\} \le C < \infty$.
   2. $\mathrm{E}[\sigma(z)] = 0, \mathrm{E}[z\sigma(z)] \neq 0$ for $z \sim N(0,1)$.　(ちょっと強い条件)
3. **教師の条件:** f^*はリプシッツ連続で$\|f^*\|_{L^2(P_X)} = \Theta_d(1)$.

e.g., tanh





Wの勾配法における軌跡 ($d = 2$).
f^*は二つのReLUニューロンの和.

1. 平均場設定では，各ニューロンは初期値から多きく動いて目標となる真の関数の方向 (黒い破線; 二つのニューロン) を向く.

2. NTKのスケールでは，一層目のパラメータはほとんど動かず特徴学習ができていない.

See also Akiyama&Suzuki (2021), Chizat (2019).

ランダム特徴量

ランダム特徴量 (特徴量の学習無し)

- **Conjugate kernel** at initialization:

$$\phi_{\mathrm{CK}}(x) = \frac{1}{\sqrt{N}}\sigma(W_0^\top x)$$

正確な漸近解析がかなり研究されている (e.g., [Louart, Liao, and Couillet, 2018; Mei and Montanari, 2019])

- **NTK** (Neural tangent kernel):

$$\phi_{\mathrm{NTK}}(x) = \frac{1}{\sqrt{Nd}}\mathrm{Vec}(\sigma'(W_0^\top x)x^\top)$$

$$\hat{a}_{\mathrm{RF}} = \mathop{\arg\min}_{a\in\mathbb{R}^N}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle a, \phi_{\mathrm{RF}}(x_i)\rangle)^2 + \frac{\lambda}{N}\|a\|^2\right\}$$

RF ∈ {CK,NTK}

学習されたカーネル (１ステップGD):

$$\phi_{\mathrm{GD}}(x) = \frac{1}{\sqrt{N}}\sigma(W_1^\top x)$$

よくわかってない

$$\hat{a}_{\mathrm{GD}} = \mathop{\arg\min}_{a\in\mathbb{R}^N}\left\{\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \langle a, \phi_{\mathrm{GD}}(\tilde{x}_i)\rangle)^2 + \frac{\lambda}{N}\|a\|^2\right\}$$

$(\tilde{x}_i, \tilde{y}_i)_{i=1}^n$: i.i.d. copy of $(x_i, y_i)_{i=1}^n$

$$\hat{a}_{\mathrm{RF}} \quad \text{vs.} \quad \hat{a}_{\mathrm{GD}}$$

RFの限界

[El Karoui (2010); Ghorbani et al. (2019); Hu and Lu (2020); ...]

定理 (RFの予測誤差の下限)

$$\inf_{\lambda>0} \min\{\mathcal{R}_{\mathrm{CK}}(\lambda), \mathcal{R}_{\mathrm{NTK}}(\lambda)\} \geq \|P_{>1} f^*\|_{L^2(P_X)}^2 + o_{p,d}(1)$$

$$P_{>1} f^* := (I - P_{\leq 1}) f^*$$

ただし，$P_{\leq 1}$は線形関数がなす空間への$L^2(P_X)$-空間内での射影.

Remark: 同じことが "回転不変カーネル"でもなりたつ [El Karoui (2010)].

- 高次元かつ$d/n = O(1)$の状況では，ランダム特徴量を使っている限り真の関数f^*の線形要素しか取り出せない.
- f^*の非線形性が強ければ，精度が出せない.

($n = d^{k+1-\epsilon}$ならk次多項式まで出てくる [Ghorbani et al. 2021; Mei et al. 2021])

これは，高次元設定では以下のような中心極限定理が成り立つからである:

$$a^\top \phi_{\mathrm{CK}}(x_i) = \frac{1}{\sqrt{N}} a^\top \sigma(W_0^\top x_i) \approx \frac{1}{\sqrt{N}} a^\top (\mu_1 W_0^\top x_i + \mu_2 z)$$

[Gaussian等価性]
後のスライドを参照

最初の勾配ステップはほぼランクが１

- 勾配G_tは，ランク１行列で近似できる．
  ⇒ W_1のスペクトル分布に「スパイク」が現れる！

$$G_t = -\frac{1}{n}X^\top\left[\left(\frac{1}{\sqrt{N}}\left(\frac{1}{\sqrt{N}}\sigma(XW_t)a - y\right)a^\top\right)\odot\sigma'(XW_t)\right]$$

(σ'の非線形性より，一般的には低ランクにならない．しかし高次元だと低ランクになる)

(Gordon-Slepian ineq.; Hanson-Wright inequ)

定理 (勾配のランク１近似)

$$G_0 = \frac{1}{\eta\sqrt{N}}(W_1 - W_0) \qquad (\because W_1 = W_0 + \eta\sqrt{N}G_0)$$ に注意．

$$\mu_1 = \mathbb{E}[z\sigma(z)],\quad \mu_2 = \sqrt{\mathbb{E}[\sigma(z)^2] - \mu_1^2},\quad \text{where } z\sim\mathcal{N}(0,1).$$

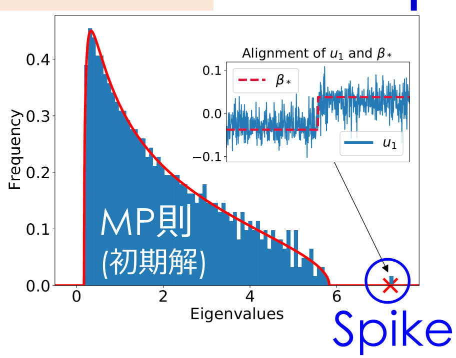
(活性化関数の１次成分と2次成分)

$$A := \frac{\mu_1}{n\sqrt{N}}X^\top y a^\top$$ (ランク1行列)とすれば，以下を得る:

$$\|G_0 - A\| \lesssim \frac{\log^2(n)}{\sqrt{n}}\cdot\|G_0\|$$

with high probability for sufficiently large n, d, N.

$W_1 = W_0 + \eta\times(\text{rank one matrix})$.
⇒ ステップサイズηが大きいと，スパイクが支配的．



Alignment with the Target Function

参考

f^* can be decomposed into

first order higher order

$$f^*(x) = \mu_0^* + \mu_1^* \langle x, \beta_* \rangle + P_{>1} f^*(x)$$

where $\mu_1^* \beta_* = \mathbb{E}_{x \sim N(0,I)}[x f^*(x)]$.

Define $\bar{\mu} \coloneqq \lim_{d\to\infty} \|f^*\|_{L^2(P_X)}$, and

$$\theta_1 := \sqrt{\bar{\mu}^2 \psi_1^{-1} + \mu_1^{*2}} \cdot \mu_1 \eta, \quad \theta_2 := \mu_1 \mu_1^* \eta.$$

$n/d \to \psi_1$, $N/d \to \psi_2$

Theorem (Alignment to the target function)

Suppose that $\eta = \Theta(1)$ and $\sigma_1(W_1)$ is the largest singular value of W_1 and u_1 is the corresponding left singular vector, then

(i) $\theta_1 > \psi_2^{1/4}$ (large step size):

$$|\langle u_1, \beta_* \rangle|^2 \to \frac{\theta_2^2}{\theta_1^2} \left(1 - \frac{\psi_2 + \theta_1^2}{\theta_1^2(\theta_1^2 + 1)} \right), \quad \sigma_1(W_1) \to \sqrt{\frac{(1+\theta_1^2)(\psi_2 + \theta_1^2)}{\theta_1^2}}$$

(ii) $\theta_1 < \psi_2^{1/4}$ (small step size):

$$|\langle u_1, \beta_* \rangle|^2 \to 0, \qquad \sigma_1(W_1) \to 1 + \sqrt{\psi_2}$$

In both cases, we have

$$|\langle u_i, \beta_* \rangle|^2 \to 0 \quad (\forall i \geq 2).$$

Alignment with the Target Function

参考

f^* can be decomposed into

first order higher order

$$f^*(x) = \mu_0^* + \mu_1^* \langle x, \beta_* \rangle + P_{>1} f^*(x)$$

where $\mu_1^* \beta_* = \mathbb{E}_{x \sim N(0,I)}[x f^*(x)]$.

Define $\bar{\mu} \coloneqq \lim_{d\to\infty} \|f^*\|_{L^2(P_X)}$, and

$$\theta_1 := \sqrt{\bar{\mu}^2 \psi_1^{-1} + \mu_1^{*2}} \cdot \mu_1 \eta, \quad \theta_2 := \mu_1 \mu_1^* \eta.$$

$n/d \to \psi_1$, $N/d \to \psi_2$

Theorem (Alignment to the target function)

Suppose that $\eta = \Theta(1)$ and $\sigma_1(W_1)$ is the lar
the corresponding left singular vector, then

Phase transition for the leading singular value: i.e., under sufficiently large step size, the "spike" in the weight matrix aligns with the leading term of the true function.

(i) $\theta_1 > \psi_2^{1/4}$ (large step size):

$$|\langle u_1, \beta_* \rangle|^2 \to \frac{\theta_2^2}{\theta_1^2}\left(1 - \frac{\psi_2 + \theta_1^2}{\theta_1^2(\theta_1^2 + 1)}\right), \quad \sigma_1(W_1) \to \sqrt{\frac{(1+\theta_1^2)(\psi_2 + \theta_1^2)}{\theta_1^2}}$$

The "alignment" becomes more significant as we (i) increase the step size; (ii) use more data (i.e., larger ψ_1)

(ii) $\theta_1 < \psi_2^{1/4}$ (small step size):

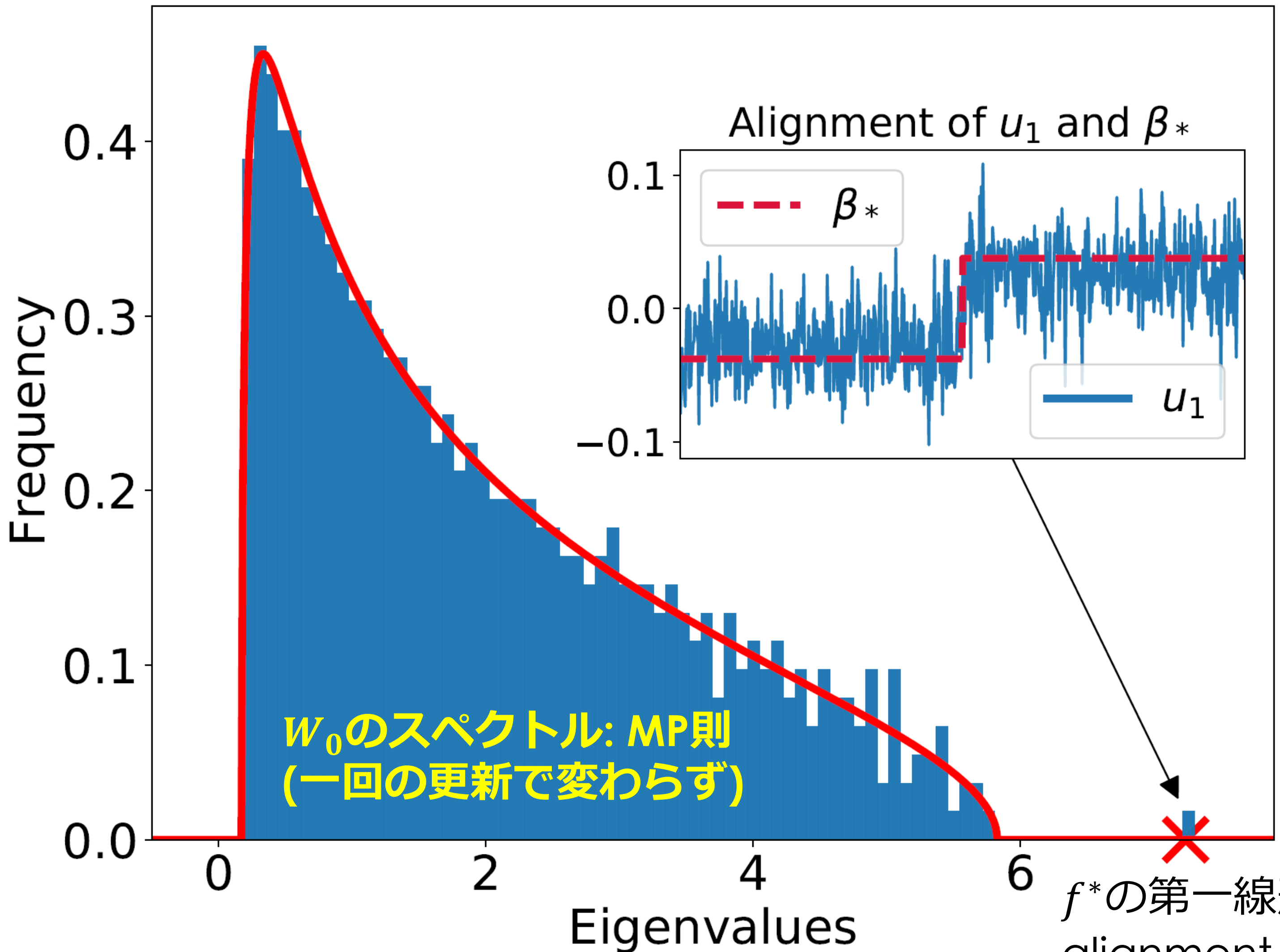
$$|\langle u_1, \beta_* \rangle|^2 \to 0, \quad \sigma_1(W_1) \to 1 + \sqrt{\psi_2}$$

In both cases, we have

$$|\langle u_i, \beta_* \rangle|^2 \to 0 \quad (\forall i \geq 2).$$

Non-leading singular vectors are **near orthogonal to the linear part of the true function** in the high dimensional setting.

W_1のスペクトル分布のスパイク



f^*の第一線形成分への alignment (特徴学習)

Gaussian Equivalence Property

- tステップ目のconjugate kernel :

$$\phi_{\mathrm{CK}}(x) = \frac{1}{\sqrt{N}}\sigma(W_t^\top x)$$

- **Gaussian等価なモデル**:

$$\phi_{\mathrm{GE}}(x) = \frac{1}{\sqrt{N}}(\mu_1 W_t^\top x + \mu_2 z)$$

線形成分+ガウス雑音
(非線形部分はガウス雑音になってしまう)

$\mu_1 = \mathbb{E}[z\sigma(z)], \quad \mu_2 = \sqrt{\mathbb{E}[\sigma(z)^2] - \mu_1^2},$
where $z \sim \mathcal{N}(0,1)$.

$$\hat{a}_{\mathrm{F}} = \arg\min_{a \in \mathbb{R}^N} \left\{ \frac{1}{n}\sum_{i=1}^{n} (\tilde{y}_i - \langle a, \phi_{\mathrm{F}}(\tilde{x}_i)\rangle)^2 + \frac{\lambda}{N}\|a\|^2 \right\}$$

where $\mathrm{F} \in \{\mathrm{CK}, \mathrm{GE}\}$.

$(\tilde{x}_i, \tilde{y}_i)_{i=1}^n$: i.i.d. copy of $(x_i, y_i)_{i=1}^n$

初期解に関する解析は[Hu and Lu, 2020]

定理 (Gaussian等価性)

$\eta = \Theta(1)$として，$f^*(x) = \sigma^*(\langle x, \beta^* \rangle)$とする．すると，有界な$t$に対する最初の$t$ステップでは以下が成り立つ:

$$|\mathcal{R}_{\mathrm{CK}}(\lambda) - \mathcal{R}_{\mathrm{GE}}(\lambda)| = o_{p,d}(1)$$
$$\mathcal{R}_{\mathrm{GE}}(\lambda) \geq \|P_{>1} f^*\|_{L^2(P_X)}^2$$

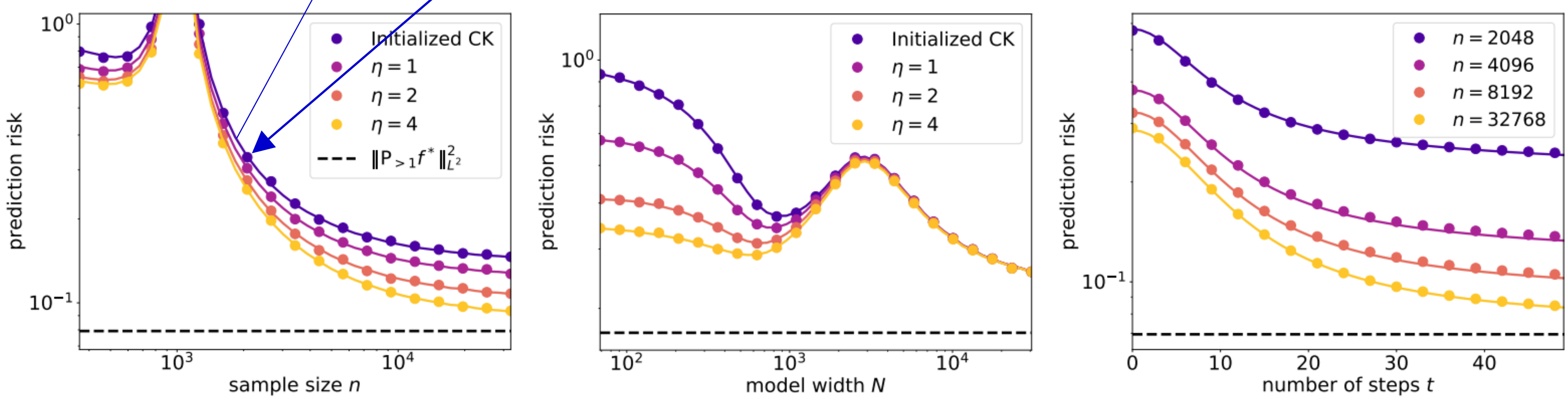
- これは特徴量の空間における“中心極限定理”と言える． (L2正則化によってa_iの全成分が満遍なく小さくなるので)
- 非線形部分はi.i.d.ガウス雑音のようにふるまう (非線形項の情報は消える).
- 小さなステップサイズで定数回勾配法で更新しても限界がある.

数値実験との整合性

1. GETは初期解(ランダム特徴量)だけでなく，**学習したモデル**でも成り立つ.
2. GETを使うことで，ランダム行列の理論から**予測誤差を厳密に計算可能**.
   (なぜならカーネル関数が“線形化”されているので).

Solid line: GEから理論的予測.

Dots: 計算機によるシミュレーション.



$\eta = \Theta(1)$を用いて学習した特徴量を用いたリッジ回帰の予測誤差

より大きなステップサイズでより多くのステップを費やした方がより良い精度
→ 特徴学習が予測誤差を改善している！

初期値のCKからの改善

($R_W(\lambda)$ is the ridge regression estimator using W for the first layer.)

- $\eta = \Theta(1)$ (中間的な大きさのステップサイズ):

$$\boxed{\mathcal{R}_{W_0}(\lambda) - \mathcal{R}_{W_1}(\lambda) \xrightarrow{p} \exists \delta > 0}$$

線形な領域

  - 1ステップのGDは必ず精度を改善させる.
  - しかし，大きく改善させることはない．なぜなら $\mathcal{R}_{W_1}(\lambda) \geq \|P_{>1} f^*\|^2_{L^2(P_X)}$.

- $\eta = \Theta(\sqrt{N})$ (大きな学習率): $f^*(x) = \sigma^*(\langle x, \beta^* \rangle)$ を仮定

Maximal update parameterization (μP) [Yang and Hu, 2020] として知られている.

- $\tau^* = 0$ if $\sigma = \sigma^* = \mathrm{erf}$.
- $\tau^* \ll 1$ if $\sigma = \sigma^* = \tanh$.

$$\tau^* = \inf_{\eta > 0} \mathbb{E}_{\xi_1 \sim N(0,1)} \left[\sigma^*(\xi_1) - \mathbb{E}_{\xi_2 \sim N(0,1)} [\sigma(\eta \xi_1 + \xi_2)] \right]$$

(モデルのズレを表す量)

$n/d \to \psi_1$, $N/d \to \psi_2$

$$\boxed{\mathcal{R}_{W_1}(\lambda) \leq 16\tau^* + C(\sqrt{\tau^*}\psi_1^{-1/2} + \psi_1^{-1}) + o_p(1)}$$

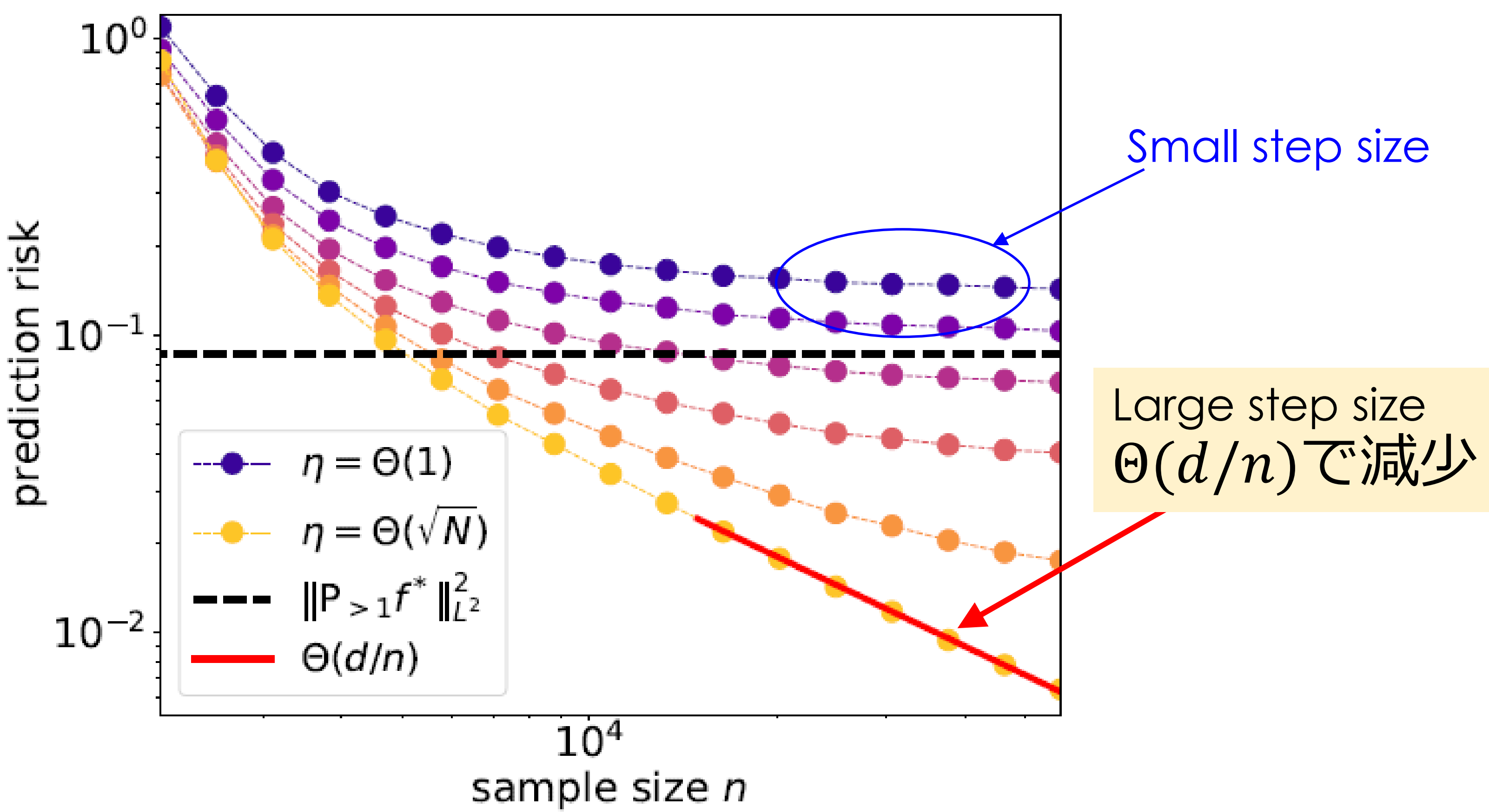
非線形領域

$\|W_1 - W_0\|_{\mathrm{F}} \succeq \|W_0\|_{\mathrm{F}}$

学習率を大きくすることで，精度を大きく改善できる
($\tau^* \ll \|P_{>1} f^*\|^2$の状況で). ※バイアス項を振りなおせば$\tau^* = 0$とできる.

考察

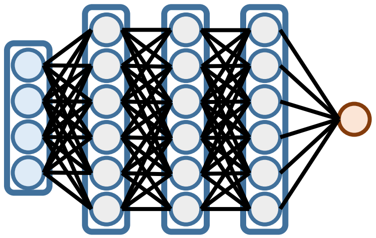
Corollary もし$\sigma = \sigma^* = \mathrm{erf}$なら，$\tau^* = 0$.
特に，$R_{W_1}(\lambda) = \Theta(\psi_1^{-1}) = \Theta(d/n)$ である.

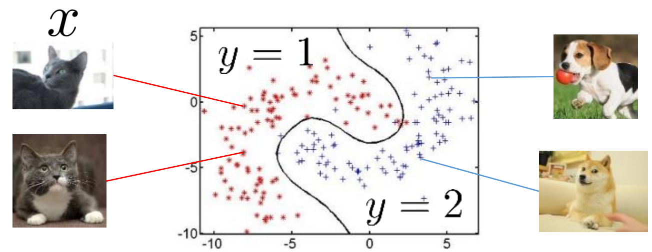


Predictive risk of ridge regression on CK obtained by one step GD (empirical simulation, $d = 1024$): brighter color represents larger step size scaled as $\eta = N^\alpha$ for $\alpha \in [0,1/2]$. We chose $\sigma = \sigma^* = \mathrm{erf}$, $\psi_2 = 2, \lambda = 10^{-3}$, and $\sigma_\epsilon = 0.1$.

機械学習の注意点

Learning = Fitting a model to the training data



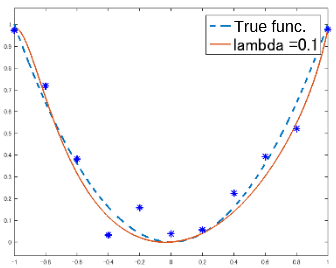


$$f_\theta(x) = W_L \sigma(W_{L-1} \cdots \sigma(W_1 x))$$

Find a model that fits the data well by minimizing a loss function.

$$\min_\theta \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$$

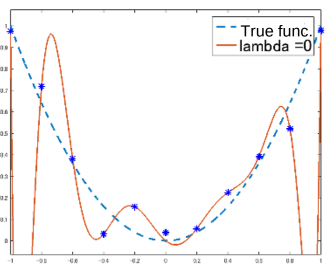
○ Good pred.



ML is intended to obtain good **prediction**.

- Lower classification error.
- Better recommendation.

× Bad pred. (overfit)



Black-box model

They do not find causality but correlation.

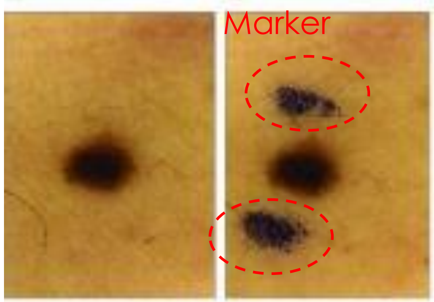
Better prediction ≠ Better interpretability/explainability:

- How is a particular answer produced?
- Is the output produced by a correct reasoning?

Learning a Shortcut

ML models sometimes find a "**shortcut**" that lowers the training error but yields an unreasonable decision rule.

- Medical image recognition:
  Marker marked by doctors produces unintentional bias.



[Winkler et al., Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. JAMA dermatology, 155(10):1135–1141, 2019]

Camera, lighting conditions, skin-colors also produce bias.

- General object recognition:

A specific object is linked to the background color.
➢ ML model recognizes via the BG-color.
➢ It fails if BG-color changes.

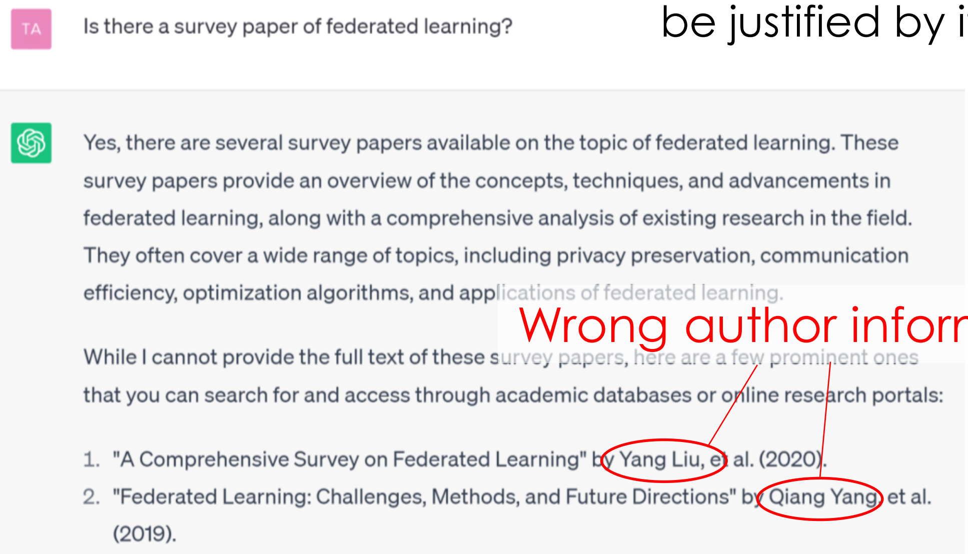
[Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. Distilling localization for self-supervised representation learning. arXiv: 2004.06638, 2020]

Interpretability is necessary for **high-risk stakes problems**.
➢ E.g., cancer diagnosis.

Social issue: Hallucination

- Hallucination

A confident response by an AI that does not seem to be justified by its training data.


TA Is there a survey paper of federated learning?

Yes, there are several survey papers available on the topic of federated learning. These survey papers provide an overview of the concepts, techniques, and advancements in federated learning, along with a comprehensive analysis of existing research in the field. They often cover a wide range of topics, including privacy preservation, communication efficiency, optimization algorithms, and applications of federated learning.

While I cannot provide the full text of these survey papers, here are a few prominent ones that you can search for and access through academic databases or online research portals:

1. "A Comprehensive Survey on Federated Learning" by Yang Liu, et al. (2020).
2. "Federated Learning: Challenges, Methods, and Future Directions" by Qiang Yang, et al. (2019).

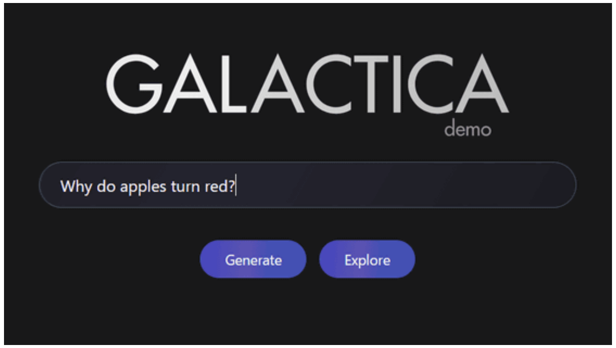
Wrong author information

The situation could get worse because it is estimated that 33–46% of crowd workers use LLMs.
→ AI would be trained by AI generated data.

[Veselovsky et al: Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. arXiv:2306.07899]

- Ethical issue

Galactica: LLM by Meta trained mainly by scientific papers and documents. (published on 15th/Nov/2022)



➢ Produced several hallucination including vicious racist responses backed up by non-existing references.
➢ Was closed to public in 3 days.

[Taylor et al: Galactica: A Large Language Model for Science. 2022]

Ethical issues

- Gender bias in natural language processing

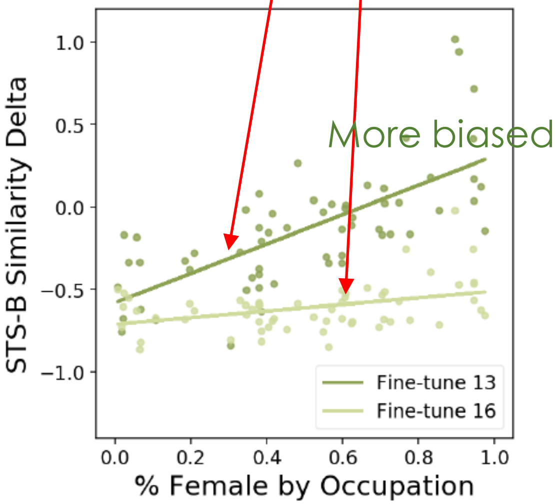
Different results from different random seeds

Semantic textual similarity (STS) task

$$\text{sim}(\text{“a woman is walking”}, \text{“a doctor is walking”}) - \text{sim}(\text{“a man is walking”}, \text{“a doctor is walking”}).$$

Are the results affected by the gender ratio of the occupations?

- ➢ Random seed to fine tune the BERT model largely affected the result.
- ➢ **It is difficult to detect from the predictive accuracy.**



[D'Amour et al. (Google), Underspecification Presents Challenges for Credibility in Modern Machine Learning. JMLR, 23(226):1−61, 2022]

- Copyright issue

Article 30-4 of the Copyright Law in Japan permits the use of a copyrighted work for training ML models, but it does not include any procedure for gaining permission in advance from copyright holders.

["Copyrighted Works Get Flimsy Protection from AI Under Japanese Law," The Yomiuri Shimbun, published on April 29, 2023]

- ➢ Japanese government announced that that necessary measures will be established by summarizing the issues such as cases in which learning of copyrighted works by AI constitutes an infringement. 「知的財産推進計画2023」

G7: Hiroshima AI process

- Agreed to establish **Hiroshima AI process.**
- They are determined to work together and with others to "advance international discussions on inclusive artificial intelligence (AI) governance and interoperability to achieve our common vision and goal of **trustworthy AI**, in line with our shared democratic values."

[HP of Ministry of Foreign Affairs of Japan. https://www.mofa.go.jp/ecm/ec/page1e_000673.html]

[Cabinet Office report: G7 Hiroshima Summit (Session 1 (Working Lunch) "Toward an International Society of Cooperation, Not Division and Conflict/World Economy" Summary)]
[内閣府資料: G7広島サミット（セッション１（ワーキング・ランチ）「分断と対立ではなく協調の国際社会へ／世界経済」概要）]

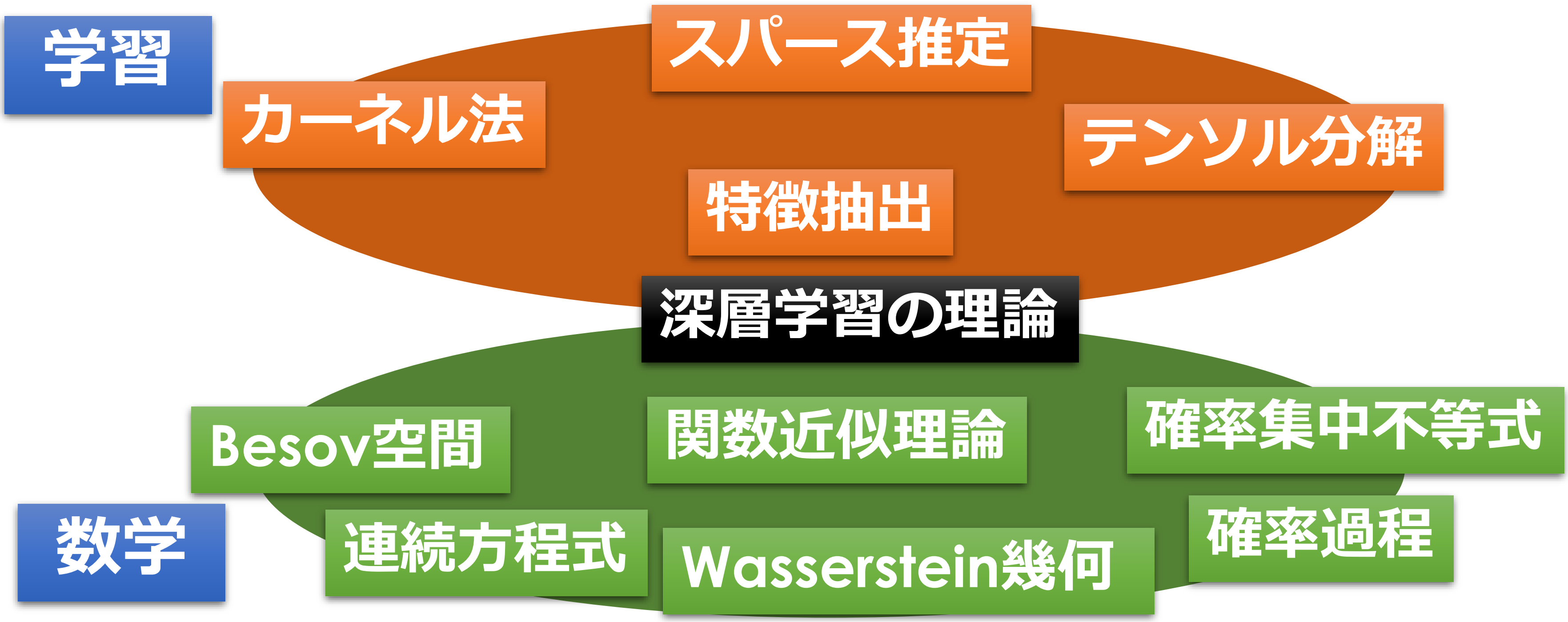
The G7 Digital and Technology Ministers' Meeting was held in Takasaki City, Gunma Prefecture, prior to the G7 Summit, and adopted a joint statement that included the promotion of the development of "human-centered and reliable artificial intelligence (AI).

1. **Facilitation of Cross-Border Data Flows and Data Free Flow with Trust.**
2. **Secure and Resilient Digital Infrastructure:**
3. **Internet Governance.**
4. **Innovating the Economic Society and Enhanci**
5. **Promoting Responsible AI and AI Governance**
6. **Competition Policy on Digital Market.**

1. reinforce democratic values
2. respect for human rights and fundamental freedoms
3. collective efforts to promote interoperability between AI governance frameworks

まとめ

- 深層学習はなぜうまくいくのか？[世界的課題]
- 数学による深層学習の原理究明
  - ➢ 「表現能力」，「汎化能力」，「最適化」

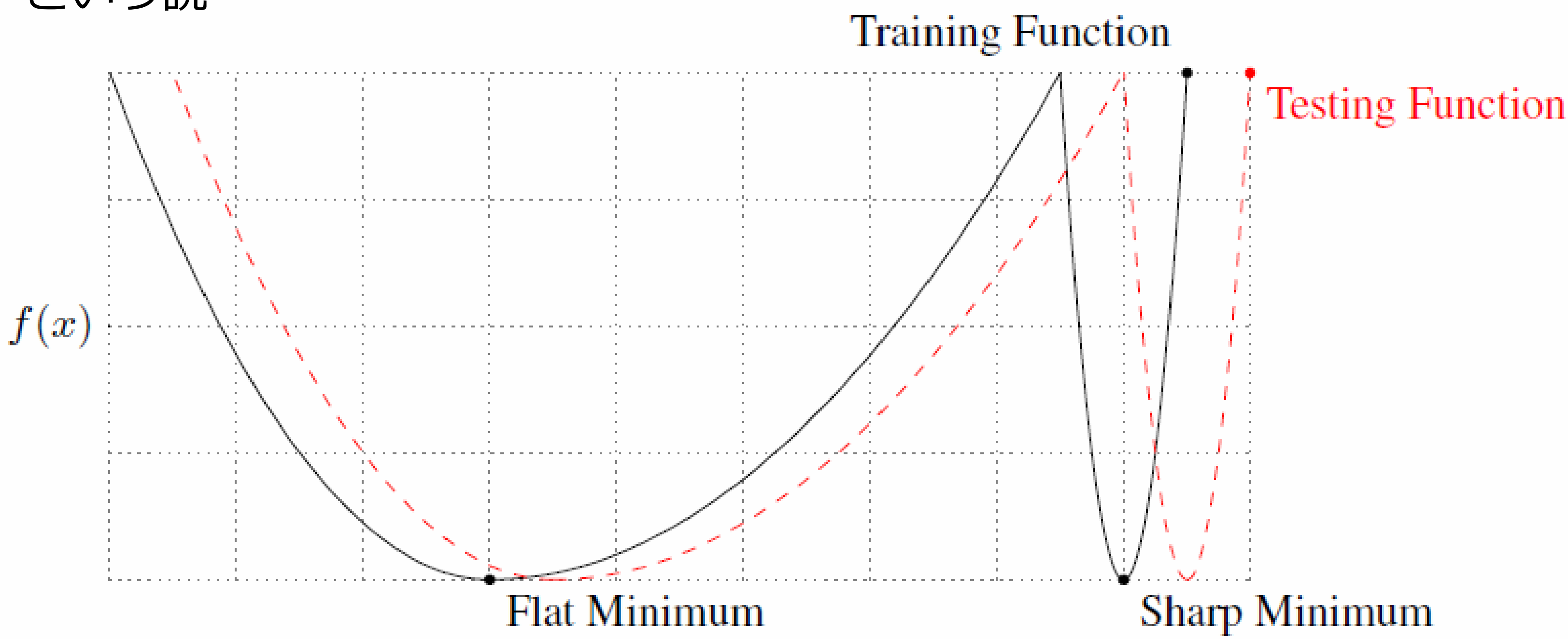


理論により深層学習を“謎の技術”から“制御可能な技術”へ
深層学習を超える方法論の構築へ

ノイズあり勾配法と大域的最適性
(参考資料：最終日に説明予定)

Sharp minima vs flat minima

SGDは「フラットな局所最適解」に落ちやすい→良い汎化性能を示す
という説



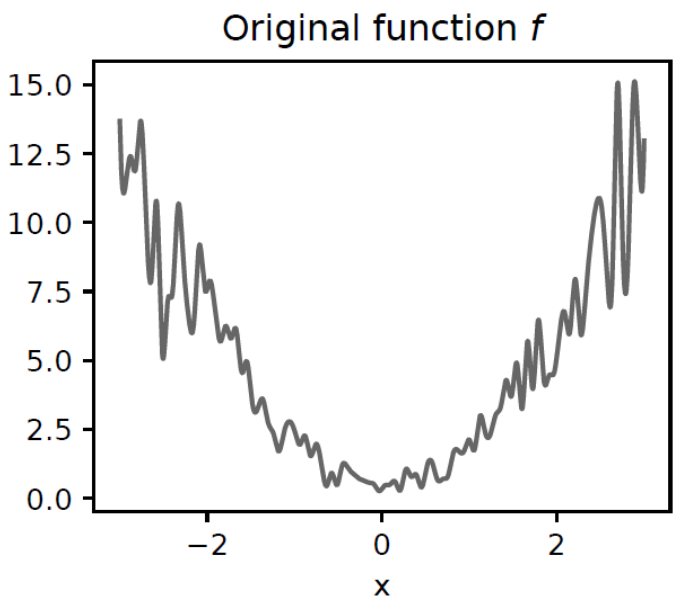
Keskar, Mudigere, Nocedal, Smelyanskiy, Tang (2017):
On large-batch training for deep learning: generalization gap and sharp minima.

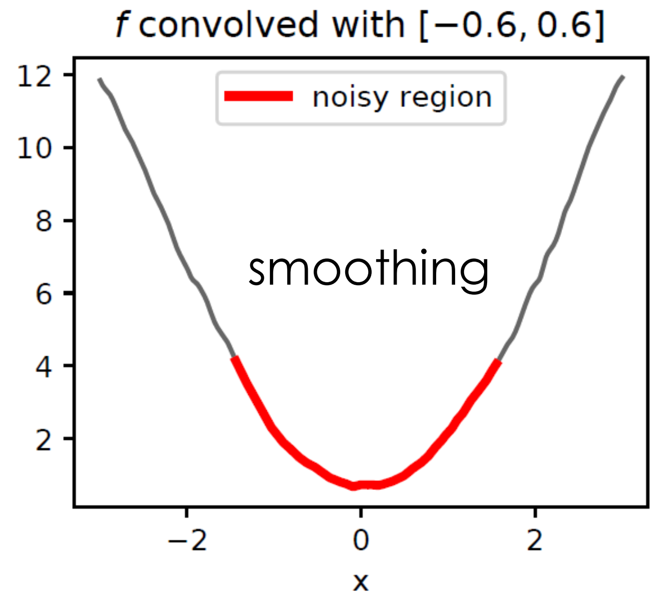
$$\theta_t = \theta_{t-1} - \underbrace{\alpha_b \left(\frac{1}{b} \sum_{j=1}^{b} \nabla_\theta \ell(z_{i_j}; \theta) \right)}_{\cong\text{正規分布}}$$

→ ランダムウォークはフラットな領域にとどまりやすい

- 「フラット」という概念は座標系の取り方によるから意味がないという批判. (Dinh et al., 2017)
- PAC-Bayesによる解析 (Dziugaite, Roy, 2017)

ノイズによる平滑化効果





[Kleinberg, Li, and Yuan, ICML2018]

確率的勾配を用いる ⇒ 解にノイズを乗せている⇒ 目的関数の平滑化

$$x_t = x_{t-1} - \eta(\nabla L(x_{t-1}) + \xi_t) \qquad (y_t = x_t + \eta\xi_t)$$

$$\Rightarrow y_t = y_{t-1} - \eta\xi_{t-1} - \eta\nabla L(y_{t-1} - \eta\xi_{t-1})$$

$$\Rightarrow \mathbb{E}_{\xi_{t-1}}[y_t] = y_{t-1} - \eta\nabla\mathbb{E}_{\xi_{t-1}}[L(y_{t-1} - \eta\xi_{t-1})]$$

ノイズを加えて平滑化した目的関数 $\bar{L}(y_t) = \mathbb{E}_{\xi_t}[L(y_t - \eta\xi_t)]$ **を最適化.**

関連研究: Graduated optimization

- Graduated non-convexity

Blake and Zisserman: *Visual reconstruction*, volume 2. MIT press Cambridge, 1987.

- Gaussian kernelとの畳み込み

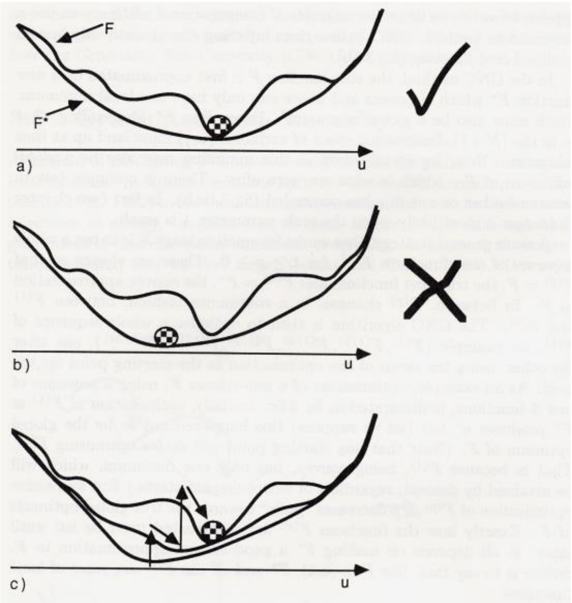
Z. Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. SIAM Journal on Optimization, 6(3):748-768, 1996.

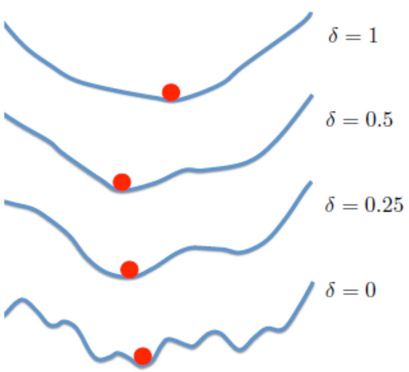
- Graduated optimization

Hazan, Levy, and Shalev-Shwartz: On graduated optimization for stochastic non-convex problems. *International conference on machine learning*, pp. 1833-1841, 2016.

σ-nice性の導入．多項式オーダーでの収束.

$$\hat{L}_\delta(x) = \mathrm{E}_{u\sim U(\mathrm{B}(\mathrm{R}^d))}[L(x+\delta u)]$$





Survey:
Mobahi and Fisher III. On the link between gaussian homotopy continuation and convex envelopes. *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 43-56, 2015.

GLD/SGLD

- 確率的勾配ランジュバン動力学
  Stochastic Gradient Langevin Dynamics (SGLD)

$$\min_{x \in \mathbb{R}^d} L(x) = \min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell_i(x)$$

(非凸)

β: 逆温度パラメータ

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

(勾配ランジュバン動力学)

定常分布： $\pi \propto \exp(-\beta L(X))$

時間離散化

[Gelfand and Mitter (1991); Borkar and Mitter (1999); Welling and Teh (2011)]

GLD:

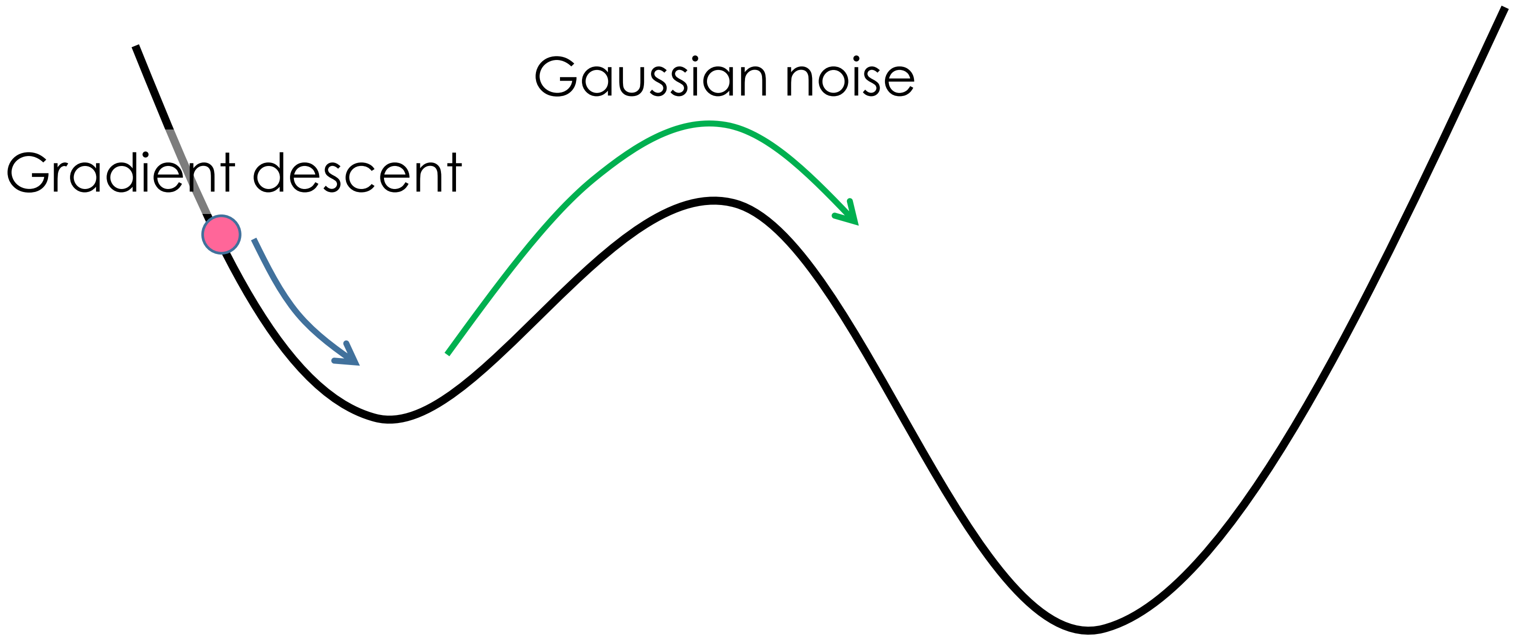
$$X_{t+1} = X_t - \eta \nabla L(X_t) + \sqrt{2\eta\beta^{-1}}\xi_t$$

(Euler-Maruyama scheme)

$\xi_t \sim \mathrm{N}(0, I)$

SGLD:

$$X_{t+1} = X_t - \underbrace{\eta \frac{1}{|I_B|} \sum_{i \in I_B} \nabla \ell_i(X_t)}_{\text{確率的勾配}} + \sqrt{2\eta\beta^{-1}}\xi_t$$

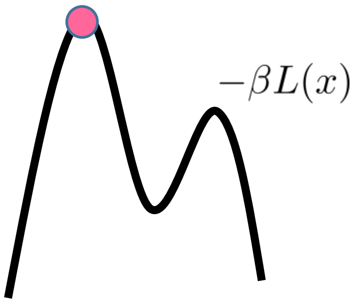

Gaussian noise
Gradient descent

定常分布

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

適当な条件のもとGLDの定常分布は以下で与えられる:

$$\pi_\infty(\mathrm{d}x) \propto \exp(-\beta L(x))\mathrm{d}x$$

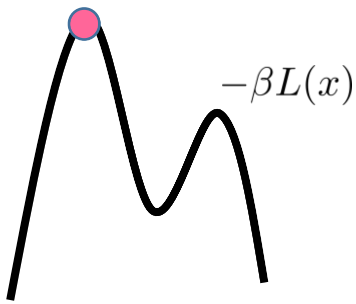


- βが大きければ，定常分布は最適解周りに集中する.
- 定常分布からのサンプリングにも使える (というかこっちが本来の目的).

対数ソボレフ不等式と幾何的エルゴード性

定常分布

$$\pi_\infty(\mathrm{d}x) \propto \exp(-\beta L(x))\mathrm{d}x$$



例：
- 二次関数+有界関数
- Weak Morse型関数

対数ソボレフ不等式 (π_∞の性質):

任意の(π_∞に対して絶対連続な)確率分布 $\mathrm{d}\nu = f\ \mathrm{d}\pi_\infty$ に対し,

$$\int f\log(f)\mathrm{d}\pi_\infty \leq 2c_{\mathrm{LS}}\int \frac{\|\nabla f\|^2}{f}\mathrm{d}\pi_\infty$$

$$(D(\nu\|\pi_\infty) \leq 2c_{\mathrm{LS}} I(\nu\|\pi_\infty))$$

KL-div　　Fisher-div

$$D(\mu\|\nu) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu,\quad I(\mu\|\nu) = \int \left\|\nabla \log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|^2 \mathrm{d}\mu$$

➡ **幾何的エルゴード性**

ρ_t: X_tの周辺分布

$$D(\rho_t\|\pi_\infty) \leq \exp(-2t/c_{\mathrm{LS}})D(\rho_0\|\pi_\infty)$$

定常分布へKL-divergenceの意味で指数オーダーの収束

[Bakry, Gentil, and Ledoux: *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014. Th. 5.2.1]

対数ソボレフ不等式の十分条件

$$L(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(x) + \lambda_1 \|x\|^2$$

Bounded perturbation lemma:

$$|\ell_i(x)| \leq B \ (\forall x) \Rightarrow c_{\mathrm{LS}} \leq \frac{1}{2\lambda_1 \beta} \exp(4\beta B)$$

[R. Holley and D. Stroock. Logarithmic sobolev inequalities and stochastic Ising models. Journal of statistical physics, 46(5-6):1159–1194, 1987.]

強凸な場合: $L(x)$ がμ-強凸 ⇒ $c_{\mathrm{LS}} \leq 1/(\mu\beta)$

[Bakry and Émery, 1985]

- **散逸的 (dissipative):**

$$\exists m > 0,\ b \geq 0,\ \langle x, \nabla L(x) \rangle \geq m\|x\|^2 - b$$

- 平滑性:

$$\exists M,\ \|\nabla L(x) - \nabla L(y)\| \leq M\|x - y\|$$

$$\Rightarrow \quad c_{\mathrm{LS}} \leq \frac{2m^2 + 8M^2}{m^2 M \beta} + \left(\frac{6M(d+\beta)+2}{m}\right) e^{O(\beta+d)}$$

[Raginsky, Rakhlin and Telgarsky, 2017]

離散時間ダイナミクスの収束レート

過程: L はM-平滑: $\exists M,\ \|\nabla L(x) - \nabla L(y)\| \leq M\|x - y\|$

定理

[Vempala and Wibisono, 2019]

ν_k: Marginal distribution of X_k (discrete time dynamics)

$$D(\nu_k \| \pi_\infty) \lesssim \exp(-k\eta/c_{\mathrm{LS}}) D(\nu_0 \| \pi_\infty) + 8 c_{\mathrm{LS}} d M^2 \eta$$

定理 (informal)

散逸性と平滑性の条件のもと (and other technical condition),

$$\mathrm{E}[L(X_k)] - L(X^*) \lesssim \exp\left(-ck\eta/c_{\mathrm{LS}}\right) + c_{c_{\mathrm{LS}},\beta,d}\,\eta + \frac{d \log(\beta+1)}{\beta}$$

幾何的エルゴード性　　時間離散化の誤差　　$E_{\pi_\infty}[L(X)] - L(X^*)$

定常分布が最適解まわりにどれだけ集中しているか

where $c, c_{C_{LS},\beta,d} > 0$ are constants.

[Raginsky, Rakhlin and Telgarsky, 2017; Xu, Chen, Zou, and Gu, 2018; Erdogdu, Mackey and Shamir, 2018]

- 逆温度パラメータβが十分大きければ，目的関数が非凸でも最適解の近くに到達できる．
- ただし，一般には対数ソボレフ不等式はβに指数的に依存することに注意．
  （そうでない場合もある：強凸目的関数，Weak Morse関数）

References

- **Finite dimensional Langevin dynamics:**

➢ Convergence in low (convex case): Dalalyan and Tsybakov, 2012; Dalalyan, 2016; Durmus and Moulines, 2015, ..

➢ Non-convex Optimization: Raginsky et al., 2017; Xu et al., 2018; Erdogdu, Mackey and Shamir, 2018

➢ Log-Sobolev inequality: Vempala and Wibisono, 2019.

- **Infinite dimensional Langevin dynamics:**

➢ Continuous time:

- Existence & Uniqueness of invariant measure: Da Prato and Zabczyk,1992; Maslowski, 1989; Sowers, 1992.
- Geometric ergodicity: Jacquot and Royer, 1995; Shardlow, 1999; Hairer, 2002, Its explicit rate: Goldys and Maslowski, 2006.

➢ Discrete time:

- Weak approximation rate of discretized scheme: Hausenblas, 2003; Debussche, 2011; Bréhier, 2014; Bréhier and Kopec 2016.

Other topics (MCMC in Hilbert space):

- preconditioned Crank–Nicolson (pCN): Hairer et al., 2014; Eberle, 2014; Vollmer, 2015; Rudolf and Sprungk, 2018.
- Metropolis-Adjusted Langevin Algorithm (MALA): Durmus and Moulines, 2015; Beskos et al., 2017.

GLDによるNNの最適化

二層NNの最適化

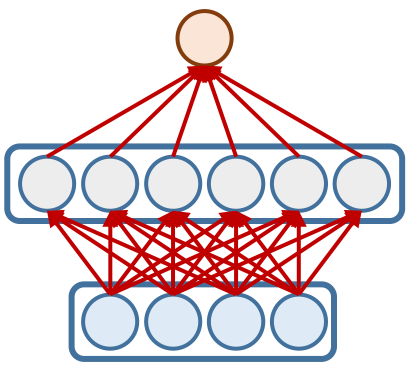
観測モデル:

$$y_i = f^\circ(x_i) + \xi_i \qquad (i = 1, \dots, n)$$

where $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}),\ \xi_i \sim (\text{mean } 0,\ \text{variance } \sigma^2,\ \text{bounded})$

教師生徒設定 with ReLU活性化関数:

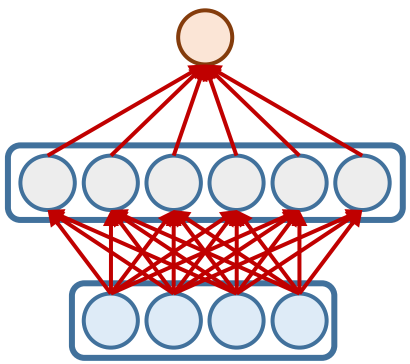
Teacher



$$f^\circ(x) = \sum_{j=1}^{m} a_j^\circ \sigma(\langle w_j^\circ, x \rangle)$$

σ : ReLU

Student (overparameterization)



$$f_\Theta(x) = \sum_{j=1}^{m} a_j \sigma(\langle w_j, x \rangle)$$

- **教師と生徒で同じサイズとする.** (おそらく緩和可能)
- 生徒は教師を多項式時間で推定できるか？

勾配ランジュバン動力学による最適化

L2-正則化あり経験損失関数:

$$\widehat{\mathcal{R}}_\lambda(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{M} a_j \sigma(\langle w_j, x\rangle)\right)^2 + \lambda \underbrace{\sum_{j=1}^{M}(a_j^2 + \|w_j\|^2)}_{\text{L2-正則化}}$$

$$\sum_{j=1}^{M} |a_j|\|w_j\| \leq \frac{1}{2}\sum_{j=1}^{M}(a_j^2 + \|w_j\|^2)$$

ReLUではL2-正則化はスパース正則化 (L1-正則化) でもある.

二段階最適化

(1) 探索フェーズ: GLD

$$\Theta^{(k+1)} = \Theta^{(k)} - \eta^{(1)}\nabla\widehat{\mathcal{R}}_\lambda(\Theta^{(k)}) + \sqrt{\frac{2\eta^{(1)}}{\beta}}\underbrace{\zeta^{(k)}}_{\text{ガウスノイズ}}$$

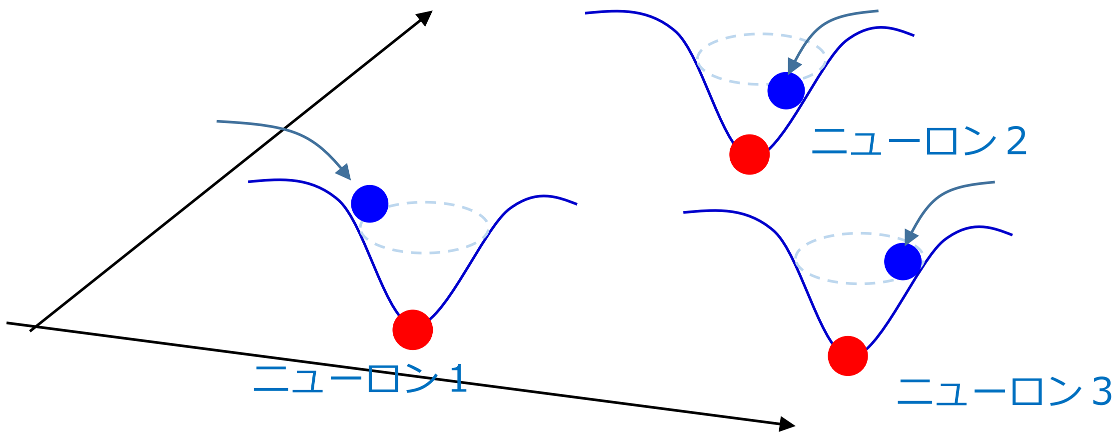
(2) 収束フェーズ: ノイズ無し・正則化なしの勾配法

$$\Theta^{(k+1)} = \Theta^{(k)} - \eta^{(1)}\nabla\widehat{\mathcal{R}}_0(\Theta^{(k)})$$

収束の様子

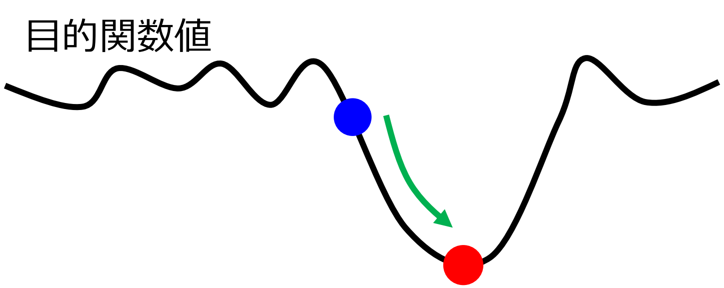
探索フェーズ:

最適解近傍への収束



収束フェーズ:

最適解まわりでは大体強凸
→ 普通の勾配法で線形収束



二段階の収束フェーズ：実験的にも観測されている (Hidden progress)

収束解析

[Akiyama and Suzuki: Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods. arXiv:2205.14818]

- $\widehat{\mathcal{R}}_\lambda(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{M} a_j\sigma(\langle w_j, x\rangle)\right)^2 + \lambda\sum_{j=1}^{M}(a_j^2 + \|w_j\|^2)$
- $\mathcal{R}(\Theta) = \mathbb{E}[(f^\circ(X) - f_\Theta(X))^2]$
- $W^\circ = (w_1^\circ, \dots, w_m^\circ)$のm番目の特異値は$\sigma_{\min}$で下から抑えられる.

定理 (informal)

$(m, d, \sigma_{\min}$固定の元で)適当な定数時間K_1が存在して,

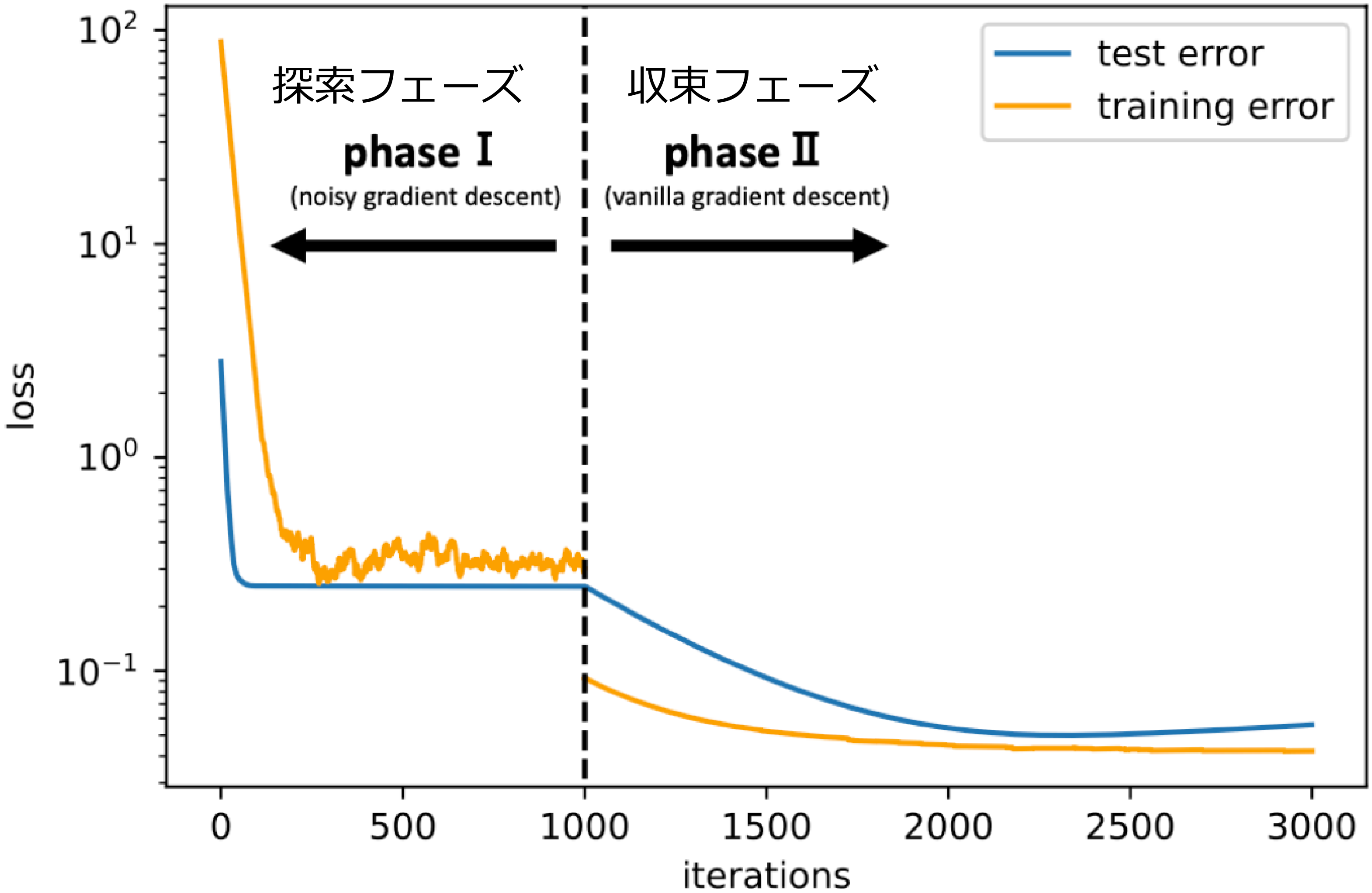
Phase 1 (大域的探索フェーズ): しばらく探索すると期待損失をある閾値以下まで減らせる (真に近くなる)

$$R\left(\Theta^{(K_1)}\right) \le \epsilon_0$$

Phase 2 (収束フェーズ): 大域的最適解へ線形収束

$$\hat{R}_0\left(\Theta^{(k)}\right) - \hat{R}_0(\Theta^*) \le c_1 \exp(-c_2(k - K_1)).$$

数値実験：二段階の学習ダイナミクス



予測誤差の比較

- 前ページのアルゴリズムで最適化して得られた解Θ^k

$$\|f_{\Theta^k} - f^\circ\|_{L^2(P_X)}^2 \lesssim \frac{\sigma_{\min}^{-4} m^5 \log(n)}{n}$$

- 線形推定量の予測誤差の下限:

$$R_{\mathrm{lin}} \gtrsim n^{-\frac{d+2}{2d+2}}$$

For $d = 2$: $n^{-2/3}$
For large d: $n^{-1/2}$

線形推定量は次元の呪いを受ける
→ 特徴量も学習することで改善

Related work

Non-overparameterized setting:

- [Li & Yuan, 2017] showed global convergence under $M = m$ and a special network structure (ResNet like structure).
- [Zhong et al., 2017] showed local convergence under $M = m$, i.e., they showed convergence when the initial solution is close to the true parameter.

Overparameterized setting:

- [Li, Ma & Zhang, 2020] showed global convergence of GD for an overparameterized setting $M > m$.

$$f^\circ(x) = \sum_{j=1}^{m} |\langle w_j, x\rangle|$$

True network

$$\mathcal{L}(\hat{f}) - \mathcal{L}(f^\circ) = O(d^{-(1+Q)})$$

where Q is a small constant.

  - Tensor decomposition technique is used.
  - The true network has a special structure.
  - The convergence is not exactly shown (it converges as $d \to \infty$).

- [Chizat, 2019]: Convergence to sparse solution with sparse reg.
  - BLASSO [De Castro & Gamboa, 2012]

2-homogeneous activation + NDSC condition

ReLU **Guarantee** ← [Akiyama&Suzuki, 2021]

2層NNの学習・粒子勾配法
(参考資料：最終日に説明予定)

- 前のスライドでは横幅が小さいNNの最適化を考えた.
- 実際は過剰パラメータ化した横幅が広いNNを使うことが多い (表現力を上げるため).

- 先のGLDの収束レートには次元が現れるので,そのままでは使えない.

(実際，次元に指数関数的に依存する)

→ 平均場ランジュバン動力学

2層NNのGLDによる最適化

$$f_\Theta(x) = \frac{1}{M}\sum_{j=1}^{M} h_{\theta_j}(x)$$

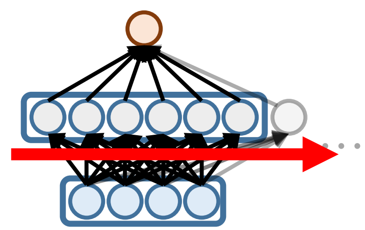
例: $h_\theta(x) = r\sigma(w^\top x)$ for $\theta = (r, w)$

$$\mathrm{d}\theta_{j,t} = -\nabla_{\theta_j} L(f_{\Theta_t})\mathrm{d}t + \sqrt{2/\beta}\,\mathrm{d}B_t$$

ニューロンが沢山あると普通のGLDの理論が適用できない.
しかし，平均場ランジュバン動力学の理論により理論保証ができる.
(逆にニューロン数無限大の極限を考えると理論保証可能になる)

多粒子化（平均場）：

$$f_\Theta(x) = \frac{1}{M}\sum_{j=1}^{M} h_{\theta_j}(x) \xrightarrow{M\to\infty} f_\mu(x) = \int h_\theta(x)\mathrm{d}\mu(\theta)$$



定理 (Hu, Ren, Šiška, and Szpruch, 2021; Mei, Montanari, and Nguye, 2018)

$M\to\infty, t\to\infty$の極限で粒子θ_jの分布μ_tは以下の分布に収束:

$$\mu_\infty = \mathop{\arg\min}_{\mu\in\mathcal{P}} L(f_\mu) + \frac{1}{\beta}\mathrm{Ent}(\mu)$$

エントロピー
$(\mathrm{Ent}(\mu) = \int \log(\mu)\mathrm{d}\mu)$

重要：分布μに対しては凸関数！(if 損失が凸)

MF-LDの収束

目的関数 $$F(\mu)=\frac{1}{n}\sum_{i=1}^{n}\ell_i(f_\mu)+\lambda_1\mathbb{E}_\mu[\|x\|^2]$$ (損失関数+正則化)

$$\frac{\delta F(\mu)}{\delta\mu}(x)=\frac{1}{n}\sum_{i=1}^{n}\ell_i'(f_\mu)h_x(z_i)+\lambda_1\|x\|^2$$

$$f_\mu(x)=\int h_\theta(x)\mathrm{d}\mu(\theta)$$

平均場ランジュバン動力学:

$$\mathrm{d}X_t=-\nabla\frac{\delta F(\mu_t)}{\delta\mu}(X_t)\mathrm{d}t+\sqrt{2\lambda_2}\mathrm{d}B_t \qquad \mu_t=\mathrm{Law}(X_t)$$

(わかりにくいが単純に各ニューロンを勾配法で動かして微小ノイズを加えていることに対応)

近接点更新解: (c.f., Mirror descent, exponentiated gradient)

$$p_\mu(x)\propto\exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta\mu}(x)\right)$$

損失を線形化して得られる解

$$p_\mu=\arg\min_{\nu\in\mathcal{P}}(\nu-\mu)\frac{\delta F(\mu)}{\delta\mu}+\lambda_2\mathrm{Ent}(\nu)$$

定理 (Entropy sandwich) [Nitanda, Wu, Suzuki (AISTATS2022)][Chizat (2022)]

p_{μ_t}が一様に対数ソボレフ不等式 (定数α)を満たすとすると,

$$\mathcal{L}(\mu_t)-\mathcal{L}(\mu^*)\le\exp(-2\alpha\lambda_2 t)(\mathcal{L}(\mu_0)-\mathcal{L}(\mu^*)).$$ (線形収束！)

ただし, $\mathcal{L}(\mu)=F(\mu)+\lambda_2\mathrm{Ent}(\mu)$

応用例

- **平均場二層ニューラルネットワーク**

$$f(x) = \frac{1}{M}\sum_{j=1}^{M} r_j \sigma(w_j^\top x) \xrightarrow{M\to\infty} f_\mu(x) = \int r\sigma(w^\top x)\mathrm{d}\mu(r,w)$$

$$f_\mu(z) = \int h_\theta(z)\mathrm{d}\mu(\theta)$$

Example: $h_\theta(x) = r\sigma(w^\top x)$ for $\theta = (r,w)$

目的関数 $$F(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

μで微分

$$\frac{\delta F(\mu)}{\delta\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i'(f_\mu)h_x(z_i) + \lambda_1\|x\|^2$$

- **MMD最小化によるノンパラメトリック密度推定**

$$F(\mu) = \mathrm{MMD}^2(g*\mu, \hat{\mu}_n) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$$

- $\mathrm{MMD}^2(\nu_1,\nu_2) := \int k(x,x')\mathrm{d}\nu_1(x)\mathrm{d}\nu_1(x') - 2\int k(x,x')\mathrm{d}\nu_1(x)\mathrm{d}\nu_2(x') + \int k(x,x')\mathrm{d}\nu_2(x)\mathrm{d}\nu_2(x')$

  k: 正定値カーネル

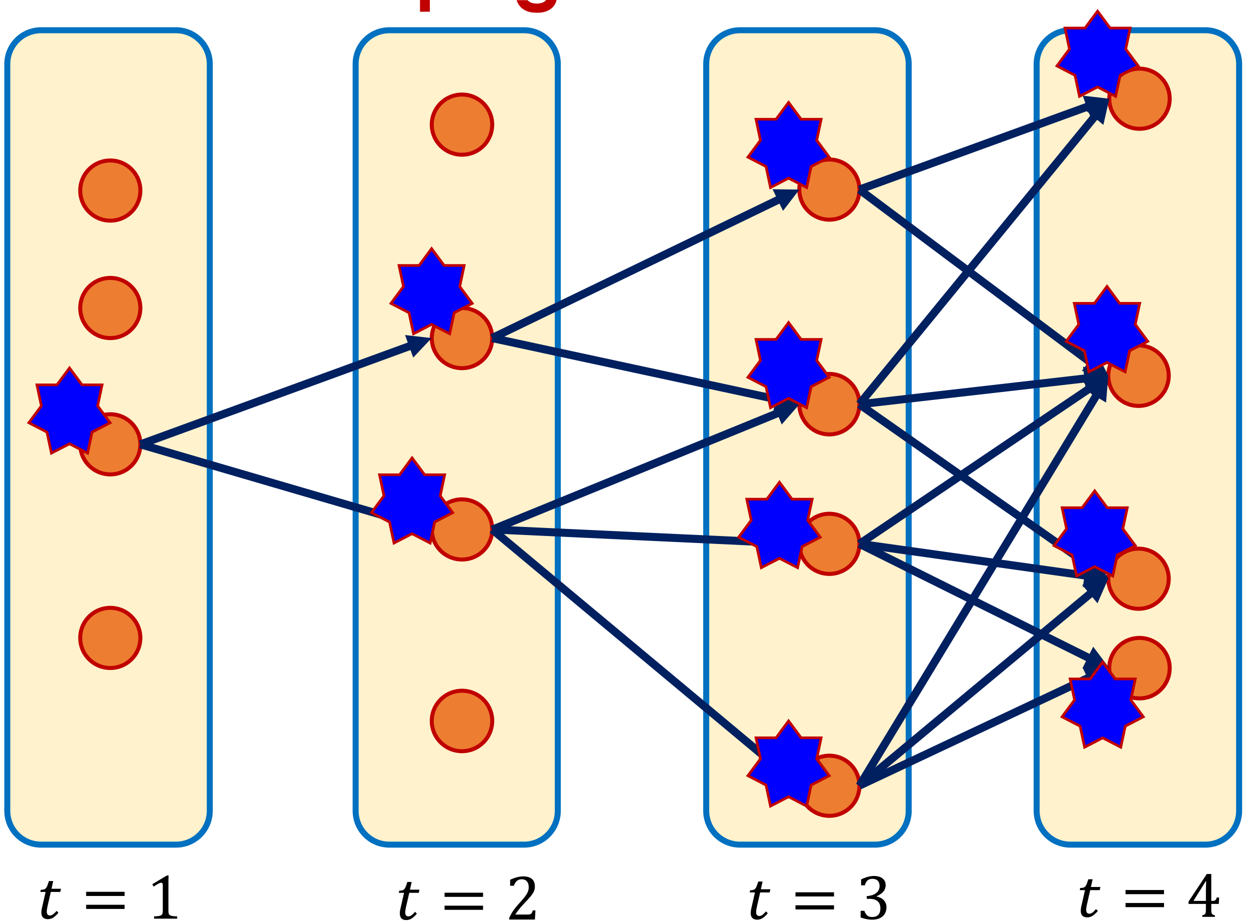
- $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$: 経験分布 (訓練データ)
- $g(x) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}}\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$

- **ベイズ事後分布の変分推論**

難しさ: McKean-Vlasov過程

- 粒子間相互作用のある確率微分方程式はMcKean-Vlasov過程として知られている． (McKean, Kac,…, 60年代)
- 離散時間・有限粒子での収束を示す際にはPropagation of chaosの評価が難しい． (粒子を増やすことでそれぞれがあたかも独立に振る舞う現象)

Propagation of chaos



一つの粒子の微小な変化が他の粒子に伝播して増幅される可能性がある．

研究の流れ

平均場NNの線形収束
連続時間・無限粒子
[Nitanda, Wu, Suzuki (AISTATS2022)]
[Chizat (2022)]

時間・空間離散化：「二重ループの手法」

- PDA法 [Nitanda, Wu, Suzuki: NeurIPS2021]
- P-SDCA法 [Oko, Suzuki, Wu, Nitanda: ICLR2022]
- 無限次元拡張 [Nishikawa, Suzuki, Nitanda: NeurIPS2022]

難しい：**Propagation of chaos**
(McKean, Kac,…, 60年代より)

空間離散化・連続時間:
Uniform-in-time propagation of chaos

- Super対数Sobolev不等式
  [Suzuki, Nitanda, Wu (ICLR2023)]
- Leave-one-out型評価
  [Chen, Ren, Wang (arXiv2022)]

時間・空間離散化・<u>確率的勾配</u>：
「<u>一重ループの手法</u>」

Suzuki, Wu, Nitanda
(arXiv:2306.07221)

離散時間・有限横幅の手法

粒子双対平均化法
(Particle Dual Averaging; PDA)

[Nitanda, Wu, Suzuki: NeurIPS2021]

$$\min_{q:\text{prob.density}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell\left(\mathbb{E}_q[h_\theta(x_i)], y_i\right) + \lambda_1 \mathbb{E}_q[\|\theta\|^2]} + \lambda_2 \mathbb{E}_q[\log(q)]$$

近似

qに関する線形汎関数で近似 (勾配を用いる)

$\mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)]$ (線形近似; $\bar{\boldsymbol{g}}^{(t)}$は基本的に勾配)

$\bar{g}^{(t)}$の決定に<u>双対平均化法</u>のルールを用いる

$$\min_{q:\text{prob.density}} \mathbb{E}_{\theta\sim q}[\bar{g}^{(t)}(\theta)] + \lambda_2 \mathbb{E}_q[\log(q)]$$

解: $q^{(t+1)}(\theta) \propto \exp(-\bar{g}^{(t)}(\theta)/\lambda_2)$ 具体形が得られる.

→ この分布からは以下の勾配ランジュバン動力学を用いてサンプリング可能:

$$\mathrm{d}\theta_t = -\nabla(\bar{g}^{(t)}(\theta)/\lambda_2)\mathrm{d}t + \sqrt{2}\mathrm{d}\xi_t.$$

時間離散化

$$\theta_k = \theta_{k-1} - \eta\nabla\bar{g}^{(t)}(\theta)/\lambda_2 + \sqrt{2\eta}\xi_{k-1}$$

計算量解析：

1. **外側ループ**: $\mathcal{L}(\hat{q}^{(t)}) - \mathcal{L}(q^*) \leq O(1/t)$
2. **内側ループ**: $T_t = \tilde{O}\left(t^2\exp(8/\lambda_2)/(\lambda_1\lambda_2)\right)$ (GLDによる)

⇒合計: $O(\epsilon^{-3})$の勾配アップデートで十分.

➢ **<u>初の多項式オーダー最適化手法</u>**

粒子確率的双対座標上昇法
(Particle Stochastic Dual Coordinate Ascent; P-SDCA)

[Oko, Suzuki, Wu, Nitanda: ICLR2022]

主問題

$$\min_p P(p) = \frac{1}{n}\sum_{i=1}^{n}\ell_i\left(\int p(\theta)h_i(\theta)\right) + \lambda_1\int\|\theta\|^2 p(\theta)\mathrm{d}\theta + \lambda_2\int p(\theta)\log(p(\theta))\mathrm{d}\theta$$

|| by Fenchelの双対定理

双対問題 $\ell_i^*(g) := \sup_{u\in\mathbb{R}}\{ug - \ell_i(u)\}$

$$-\min_{g\in\mathbb{R}^n} D(g) = \frac{1}{n}\sum_{i=1}^{n}\ell_i^*(g_i) + \lambda_2\log\left(\int q[g](\theta)\mathrm{d}\theta\right)$$

ただし $q[g](\theta) := \exp\left\{-\frac{1}{\lambda_2}\left(\frac{1}{n}\sum_{i=1}^{n}h_i(\theta)g_i + \lambda_1\|\theta\|^2\right)\right\}$

- 双対変数の座標をランダムに選択し，その座標に関して最適化.

→確率的双対座標上昇法

計算量解析：

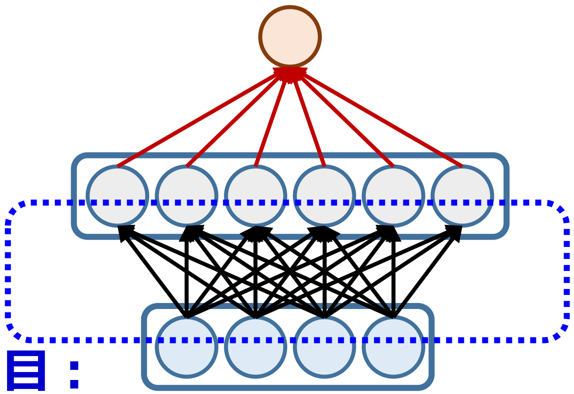
双対ギャップϵ_Pを達成するのに必要な外側ループ数：

$$t_{\mathrm{end}} = 2\left(n + \frac{1}{\lambda_2\gamma}\right)\log\left(\frac{nC}{\epsilon_P}\right)$$

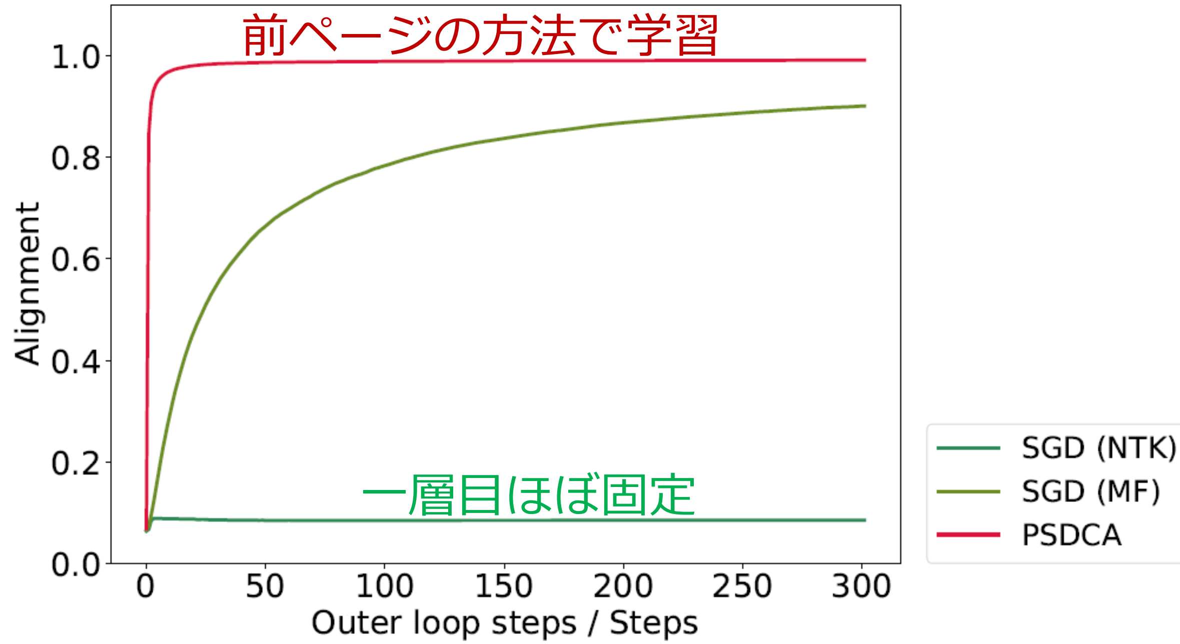
➢ <u>指数オーダーでの収束を達成</u>

➢ <u>サンプルサイズnへの依存を緩和</u>

Kernel alignment



一層目：
特徴抽出



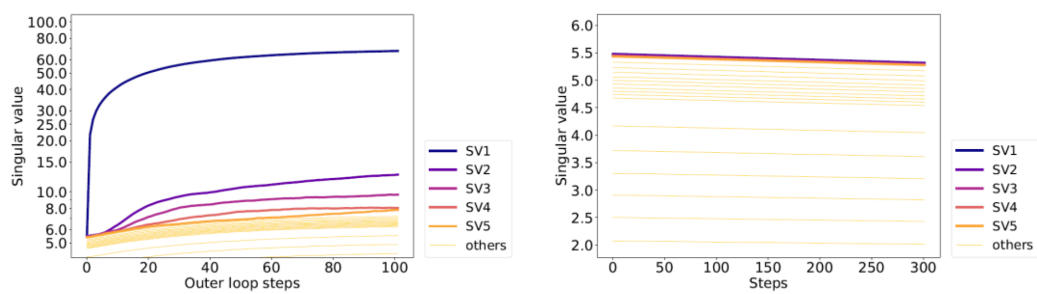
カーネルalignment:

$$A(k_W) := \frac{\langle K_W, yy^\top \rangle_{\mathrm{F}}}{\sqrt{\langle K_W, K_W \rangle_{\mathrm{F}} \langle yy^\top, yy^\top \rangle_{\mathrm{F}}}}$$

一層目で抽出された特徴量が教師信号(y)とどれだけ相関しているか？
→ 高いほど特徴量が真の関数の成分を多く含んでいる．

一層目も学習することで真の関数により適合した特徴量が学習できている．

固有値の分布：



(a) P-SDCA (mean field regime)　(b) SGD (NTK regime)

研究の流れ

平均場NNの線形収束
連続時間・無限粒子
[Nitanda, Wu, Suzuki (AISTATS2022)]
[Chizat (2022)]

時間・空間離散化：「二重ループの手法」

- PDA法 [Nitanda, Wu, Suzuki: NeurIPS2021]
- P-SDCA法 [Oko, Suzuki, Wu, Nitanda: ICLR2022]
- 無限次元拡張 [Nishikawa, Suzuki, Nitanda: NeurIPS2022]

難しい：**Propagation of chaos**
(McKean, Kac,…, 60年代より)

空間離散化・連続時間:
Uniform-in-time propagation of chaos

- Super対数Sobolev不等式
  [Suzuki, Nitanda, Wu (ICLR2023)]
- Leave-one-out型評価
  [Chen, Ren, Wang (arXiv2022)]

時間・空間離散化・<u>確率的勾配</u>：
「<u>一重ループの手法</u>」

Suzuki, Wu, Nitanda
(arXiv:2306.07221)

一重ループの方法

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

(時間離散化)

$$X_{k+1}^{(i)} = X_k^{(i)} - \eta v_k^i + \sqrt{2\eta\lambda_2}\xi_k^{(i)}$$

ただし $\mathbb{E}[v_k^i] = \nabla \frac{\delta F(\hat{\mu}_k)}{\delta \mu}(X_k^i)$ かつ $\hat{\mu}_k = \frac{1}{N}\sum_{i=1}^N \delta_{X_k^{(i)}}$

(確率的勾配)　　(空間離散化)

- 時間離散化: $X_t \to X_k^{(i)}$
- 空間離散化: N粒子で近似 $(\hat{\mu}_k)$ [もっとも難しい]
- 確率的勾配: 勾配計算を軽量化 $\left(\nabla \frac{\delta F(\mu)}{\delta \mu}(x) = \mathbb{E}[v_k(x;\mu)]\right)$

収束解析

$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2}\frac{\delta F(\mu)}{\delta\mu}(x)\right)$$: proximal Gibbs measure

定理 (1ステップ更新の減少)

p_μは対数Sobolev不等式を定数αで満たすとする.
損失関数の凸性と平滑性の仮定のもと,

$$\begin{aligned}
&\mathscr{L}^{(N)}(\hat\mu_{k+1}) - \mathcal{L}(\mu^*) \\
&\le \exp(-\lambda_2\eta_k\alpha)\left(\mathscr{L}^{(N)}(\hat\mu_k) - \mathcal{L}(\mu^*)\right) \\
&\quad + C\left(\underbrace{\eta_k^3 + \lambda_2\eta_k^2}_{\text{時間離散化}} + \underbrace{\frac{\eta_k}{N}}_{\text{空間離散化}} + \underbrace{\eta_k^{\frac{3}{2}}\lambda_2^{\frac{1}{2}}\sigma_k\tilde\sigma_k}_{\text{確率的勾配}}\right)
\end{aligned}$$

$$\begin{cases}
\sigma_k^2 = \max_i \mathbb{E}\left[\|v_k^i - \mathbb{E}[v_k^i]\|^2\right] \\
\tilde\sigma_k^2 = \max_i \mathbb{E}\left[\left\|\nabla v_k^{i\top}(\mathscr{X}) - \nabla\nabla^\top \frac{\delta F(\mu_{\mathscr{X}})}{\delta\mu}(X^i)\right\|_{\mathrm{op}}^2\right]
\end{cases}$$

既存研究では粒子数は時間に対して指数関数的に依存
[Mei et al., 2018; Javanmard et al., 2019; De Bortoli et al., 2020]

Assumption:

1. $F: \mathcal{P} \to \mathbb{R}$ is convex and has a form of $F(\mu) = L(\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2]$.
2. (smoothness) $\left\|\nabla\frac{\delta L(\mu)}{\delta\mu}(x) - \nabla\frac{\delta L(\nu)}{\delta\mu}(y)\right\| \le C(W_2(\mu,\nu) + \|x-y\|)$ and (boundedness) $\left\|\nabla\frac{\delta L(\mu)}{\delta\mu}(x)\right\| \le R$.

確率的勾配の計算量

SGD-MFLD:

$$F(\mu) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mu)$$ (有限和),

$$v_k^i = \frac{1}{B}\sum_{j\in I_k} \frac{\delta f_j(\hat\mu_k)}{\delta\mu}(X_k^i)$$ (確率的勾配)

(Mini-batch size = B)

$$\mathscr{L}^{(N)}(\hat\mu_k) - \mathcal{L}(\mu^*) \lesssim \exp(-\lambda_2\eta k\alpha) + \frac{1}{\alpha\lambda_2}\left(\underbrace{\eta^2+\lambda_2\eta}_{\text{時間離散化}} + \underbrace{\frac{1}{N}}_{\text{空間離散化}} + \underbrace{\frac{(n-B)\sqrt{\eta\lambda_2}}{B(n-1)}}_{\text{確率的勾配}}\right)$$

更新回数のバウンド:

By setting $\eta = O\left(\epsilon\alpha \wedge (\lambda_2\epsilon\alpha)^2 \frac{B^2(n-1)^2}{(n-B)^2\lambda_2} \wedge \sqrt{\lambda_2\epsilon\alpha}\right)$,
the iteration complexity becomes

$$k = O\left(\frac{1}{\epsilon\alpha} + \left(\frac{1}{\lambda_2\epsilon\alpha}\right)^2 \frac{\lambda_2(n-B)^2}{B^2(n-1)^2} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1})$$

to achieve $\epsilon + O(1/(\lambda_2\alpha N))$ accuracy.

➢ $B = n \wedge \sqrt{1/(\lambda_2\alpha\epsilon)}$ is the optimal mini-batch size. $\to k = O(\log(\epsilon^{-1})/\epsilon)$.

分散縮小勾配法

SVRG-MFLD:

$$F(\mu) = \frac{1}{n}\sum_{j=1}^{n} f_j(\mu)$$ (有限和),

$$\boxed{v_k^i = \frac{1}{B}\sum_{j\in I_k}\nabla\frac{\delta f_j(\hat{\mu}_k)}{\delta\mu}(X_k^{(i)}) - \frac{1}{B}\sum_{j\in I_k}\nabla\frac{\delta f_j(\dot{\mu})}{\delta\mu}(\dot{X}^{(i)}) + \nabla\frac{\delta F(\dot{\mu})}{\delta\mu}(\dot{X}^{(i)})}$$

(分散縮小勾配)
($\dot{X}$はm回に一回更新)

$$\begin{aligned}
&\mathscr{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*) \\
&\lesssim \exp(-\lambda_2\eta k\alpha) \\
&\quad + \frac{1}{\lambda_2\alpha}\left(\eta^2 + \lambda_2\eta + \frac{1}{N} + \frac{n-B}{B(n-1)}\lambda_2^{1/2}\eta\sqrt{m(\eta+\lambda_2)}\right)
\end{aligned}$$

時間離散化　空間離散化　確率的勾配の誤差

線形GLDの既存解析 [Kinoshita, Suzuki: NeurIPS2022] の非線形への拡張/改善

更新回数：

$$\eta = \epsilon\alpha \wedge \sqrt{\lambda_2\alpha\epsilon},$$

$$k = \frac{1}{\lambda_2\alpha\eta}\log(1/\epsilon) = O\left(\frac{1}{\epsilon\alpha} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1})$$ ただし$B = \sqrt{m} = n^{1/3}$.

総勾配計算回数：

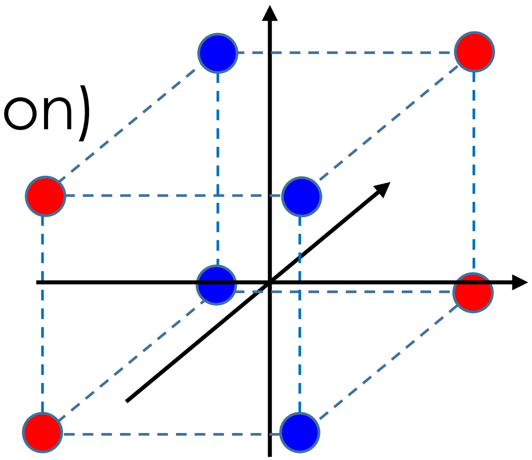
$$Bk + \frac{nk}{m} \lesssim n^{1/3}\left(\frac{1}{\alpha\epsilon} + \sqrt{\frac{1}{\lambda_2\alpha\epsilon}}\right)\frac{1}{\lambda_2\alpha}\log(\epsilon^{-1}).$$

$\sqrt{n}$ in Kinoshita&Suzuki (2022)

統計的性質

- ℓ_i: ロジスティック損失
- $h_z(x) = \bar{R} \cdot [\tanh(\langle x_1, z \rangle + x_2) + x_3]/2$

- k-スパースパリティ問題
  - ➢ $X \sim \mathrm{Unif}(\{-1,1\}^d)$ (up to freedom of rotation)
  - ➢ $Y = X_{i_1} X_{i_2} \dots X_{i_k}$ for $i_j \in [d]$ with $i_j \neq i_l$.



Q: この問題設定でカーネル法を上回る？
A: Yes.

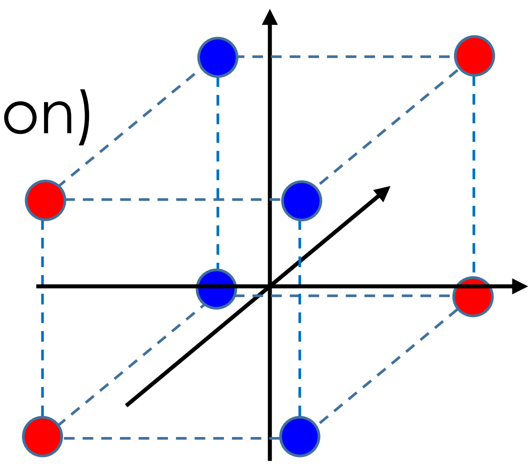
[Suzuki, Wu, Oko, Nitanda: Feature learning via mean-field Langevin dynamics: classifying sparse parities and beyond. 2023]

Authors	regime/method	width	class error	number of iterations
Ji and Telgarsky (2019)	NTK/SGD	d^8	d^2/n	d^2/ϵ
Telgarsky (2023)	NTK/SGD	d^2	d^2/n	d^2/ϵ
Barak et al. (2022)*	Two phase SGD	$O(1)$	$d^{(k+1)/2}/\sqrt{n}$	d/ϵ^2
Telgarsky (2023)	mean-field/GF	d^d	d/n	∞
Wei et al. (2019)	mean-field/WF	∞	d/n	∞
Ours*	mean-field/MFLD	$e^{O(d)}$	$\exp(-O(\sqrt{n}/d))$	$e^{O(d)}$
Ours*	mean-field/MFLD	$e^{O(d)}$	d/n	$e^{O(d)}$

統計的性質

- ℓ_i: ロジスティック損失
- $h_z(x) = \bar{R} \cdot [\tanh(\langle x_1, z\rangle + x_2) + x_3]/2$

- k-スパースパリティ問題
  - ➢ $X \sim \mathrm{Unif}(\{-1,1\}^d)$ (up to freedom of rotation)
  - ➢ $Y = X_{i_1} X_{i_2} \dots X_{i_k}$ for $i_j \in [d]$ with $i_j \neq i_l$.



Q: この問題設定でカーネル法を上回る？

A: Yes.

[Suzuki, Wu, Oko, Nitanda: Feature learning via mean-field Langevin dynamics: classifying sparse parities and beyond. 2023]

特徴学習によって次元への依存性が改善されている.

Authors	regime/method	width	class er[illegible]	[illegible]
Ji and Telgarsky (2019)	NTK/SGD	d^8	$d^2/$[illegible]	[illegible]
Telgarsky (2023)	NTK/SGD	d^2	d^2/n	d^2/ϵ
Barak et al. (2022)*	Two phase SGD	$O(1)$	$d^{(k+1)/2}/\sqrt{n}$	d/ϵ^2
Telgarsky (2023)	mean-field/GF	d^d	d/n	∞
Wei et al. (2019)	mean-field/WF	∞	d/n	∞
Ours*	mean-field/MFLD	$e^{O(d)}$	$\exp(-O(\sqrt{n}/d))$	$e^{O(d)}$
Ours*	mean-field/MFLD	$e^{O(d)}$	d/n	$e^{O(d)}$