

# 鈴木 G 研究紹介

東京大学

情報理工学系研究科 数理情報学専攻

鈴木大慈

2024年

## 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

## 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.
- ▶ ICLR2020 outstanding paper award: Atsushi Nitanda, and Taiji Suzuki: Optimal Rates for Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime.
- ▶ 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.



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

## JST関連:

CREST共同研究者: 福水健次「数理知能表現による深層構造学習モデルの革新」, 吉田 朋広「先端的確率統計学が開く大規模従属性モデリング」

## 主な活動場所

- 国内: IBIS, 統計連合大会
- 国外: NeurIPS, ICML, ICLR, ACML, ...

# 研究の大枠

## 機械学習

統計学

数理基盤

数理最適化

### 学習精度の理論

「最適な精度」の方法を構築

#### 高次元スパースモデル

- 低ランクテンソル
  - RKHSのテンソル積
  - 交互最適化
  - ベイズ推定法（ガウス過程回帰）

汎化誤差解析・  
ミニマックス最適性

#### 深層学習

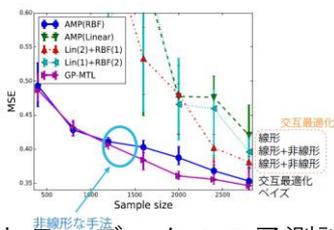
- 汎化誤差解析
    - カーネル法の理論
    - 関数近似理論
  - 基盤モデル解析
  - 深層学習の最適化
- 深層学習の理論的説明  
と構造決定への応用

### 計算手法の理論

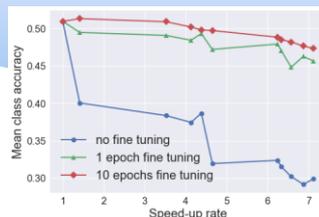
「高速計算」手法の構築

#### 正則化学習の高速最適化

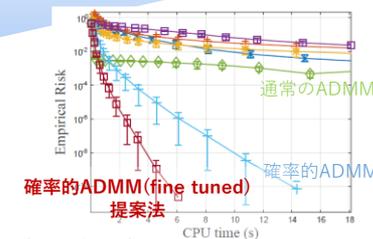
- 大規模高次元データ上での最適化
    - 確率的最適化
    - 交互方向乗数法
    - 連合学習
- 最適収束速度を達成するアルゴリズムの導出



レストランデータでの予測誤差



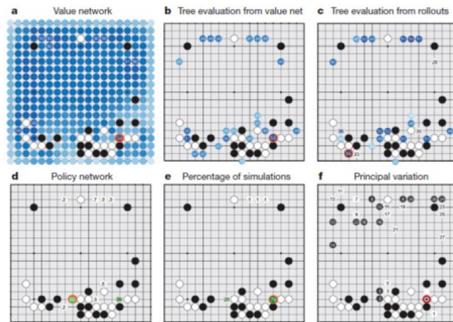
深層ニューラルネットの圧縮



(構造的) 正則化の高速解法

様々なタスクで高い精度なぜ？

## 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]

## Image recognition



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

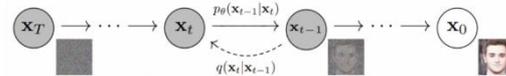
## Large language model

[Alammar: How GPT-3 Animations, <https://jalammar.github.io/how-gpt3-works-visualizations-animations/>]

[ChatGPT. OpenAI2022]

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

## Generative models (diffusion models)



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



Stable diffusion, 2022.



Jason Allen "Théâtre D'opéra Spatial" generated by **Midjourney**. Colorado State Fair's fine art competition, 1<sup>st</sup> 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？

- 「〇〇法が良い」という様々な仮説の氾濫。
- 世界的課題
- 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.”

“鋪金術”という批判

### 民間の問題意識

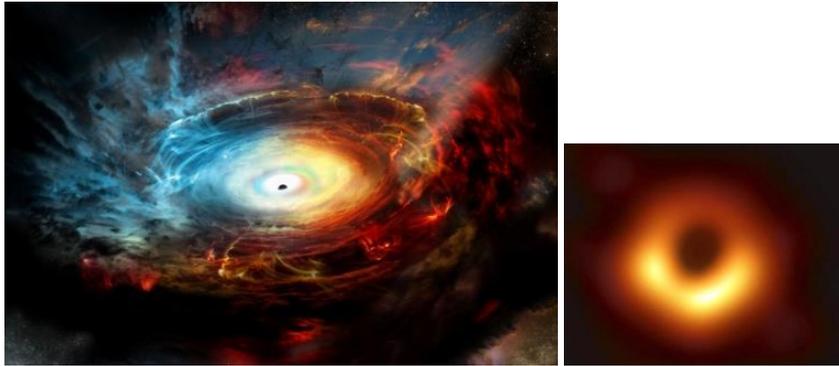
- 中で何が行われているか分からないものは用いたくない。
- 企業の説明責任. 深層学習のホワイトボックス化.



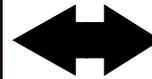
[理論]

誰もが使っている手法を  
誰もが分かっていないという状況を改善

## Physics



physical phenomenon

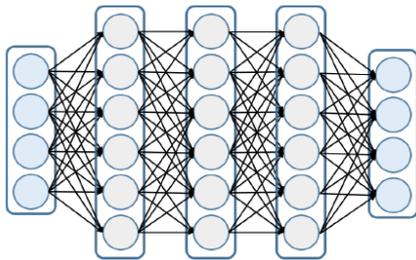


$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = \frac{8\pi G}{c^4}T_{\mu\nu}$$

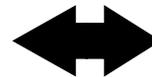
- Theory of Relative
  - Riemannian geometry
- Quantum mechanics
  - Functional analysis

**Mathematics**

## Machine learning



Deep learning



Several mathematicians/physicists join the ML community.

- Prob. theory
- Functional anal.
- Wasserstein geom.
- Diffusion equation
- Statistics
- Optimization
- Numerical analysis

**Mathematics**

# 深層学習の理論概観



## 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

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。

**我々の研究分野**

## 我々の結果

### 関数表現理論

どれだけ難しい問題まで学習できるようにになるか？

Besov空間での近似理論  
積分表現理論  
グラフCNNの表現能力

### 汎化能力

有限個のデータで学習し、どれだけ正しく新しいデータで正解できるようにになるか？

再生核理論による自由度解析  
NTKの収束理論  
モデル圧縮型汎化理論  
スパース理論との融合

### 最適化理論とその実践

最適な重みを高速に計算機で求めることが可能か？

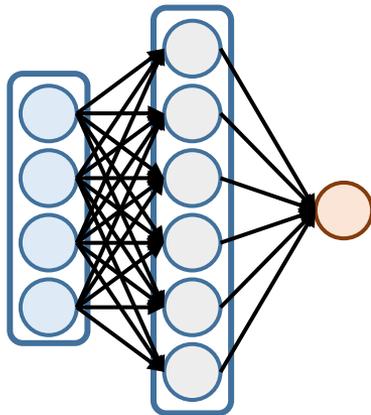
並列計算による高速化  
確率的加速勾配法  
NTKの理論解析  
無限次元勾配Langevin動力学

# 深層学習の表現能力

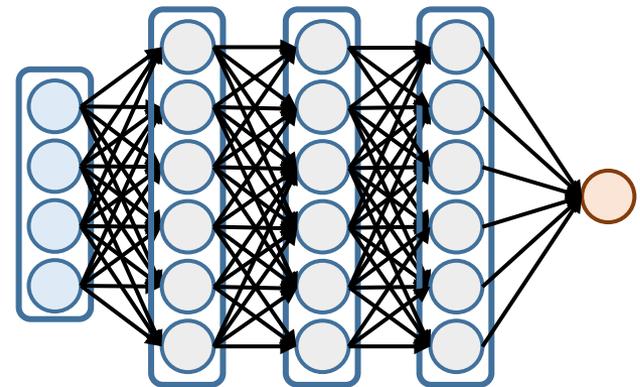
- **[理論]** 万能近似能力という意味では2層で十分.
- **[実際]** 実際は多層を使うことが多い.  
→ この差はどう埋める？



カーネル法  
従来法



多層ニューラルネット  
深層学習



# なぜ深層学習が良いのか？

統計的決定理論におけるミニマックス最適性理論で特徴づけ可能

## 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

$$Y_i = U V X_i$$

深層

$$\frac{r(M+N)}{n}$$

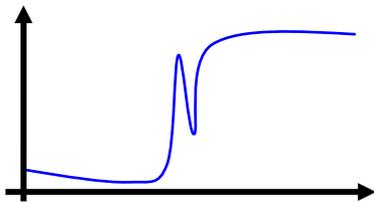
カーネル

$$\frac{MN}{n}$$

## 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



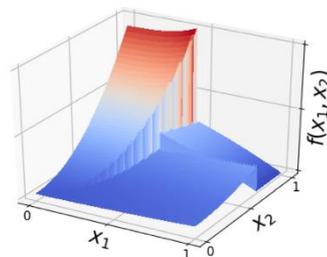
$$n^{-\frac{2s}{2s+d}}$$

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

## 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



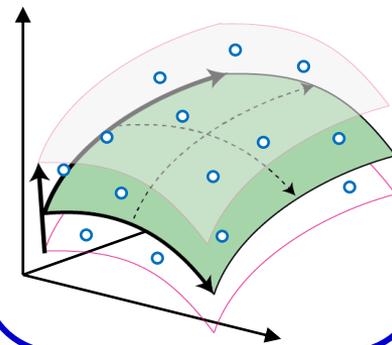
$$n^{-\frac{2s}{2s+d}} \vee n^{-\frac{\alpha}{\alpha+D-1}}$$

$$\frac{1}{\sqrt{n}}$$

## 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



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

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

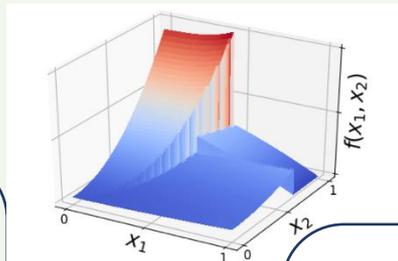
推定精度

# 数学的一般化

縮小ランク回帰

$$Y_i = U V X_i$$

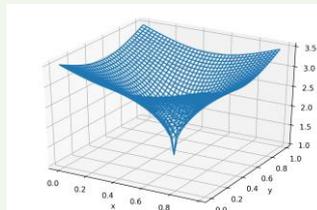
区分滑らかな関数



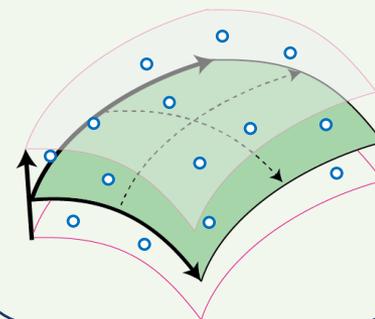
Besov空間



変動指数  
Besov空間



低次元データ



非凸性  
スパース性

# 無限次元入力NN

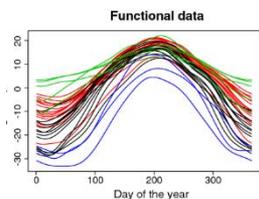
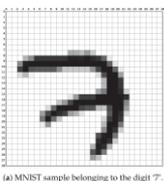
## 無限次元入力

(画像, 音声信号, 自然言語,...)

### 無限 (高) 次元データ

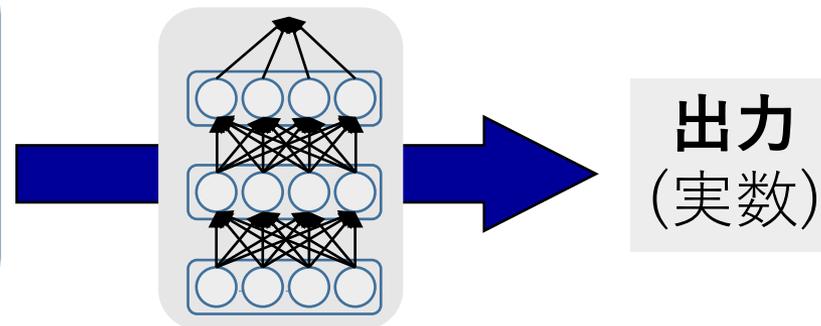
画像データ

関数データ



- 音声
- 文章
- 

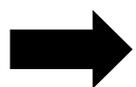
[Okumoto&Suzuki: Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. ICLR2022. Spotlight]



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

典型的なノンパラメトリック回帰のバウンド:  $n^{-\frac{2s}{2s+d}}$

( $s$ : 真の関数の滑らかさ,  $d$ : 入力の次元)



## 次元の呪い

**我々の貢献:** 無限次元入力に対する深層学習の統計理論

異方的平滑性: 真の関数が座標軸方向によって異なる滑らかさを持つ。

- 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

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

# 拡散モデルの統計理論

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

(2% of all submissions)



Stable diffusion, 2022.

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

Forward process



Backward process

$$dY_t = (Y_t + 2\nabla \log(p_{\bar{T}-t}(Y_t))) dt + \sqrt{2} dB_t$$

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

経験スコアマッチング推定量:

$$\hat{s} = \arg \min_{s \in \text{DNN}} \frac{1}{n} \sum_{i=1}^n \int_{t=\underline{T}}^{\bar{T}} \mathbb{E}_{X_t | X_0 = x_{0,i}} [\|s(X_t, t) - \nabla \log p_t(X_t | x_{0,i})\|^2] dt$$

定理

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

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

**学習のボトルネックを究明**

$$\mathbb{E}_{D_n} [\text{TV}(\hat{Y}, X_0)] < n^{-\frac{s}{2s+d}} \log^9(n) \quad (t=0 \text{ 近辺の学習複雑度})$$

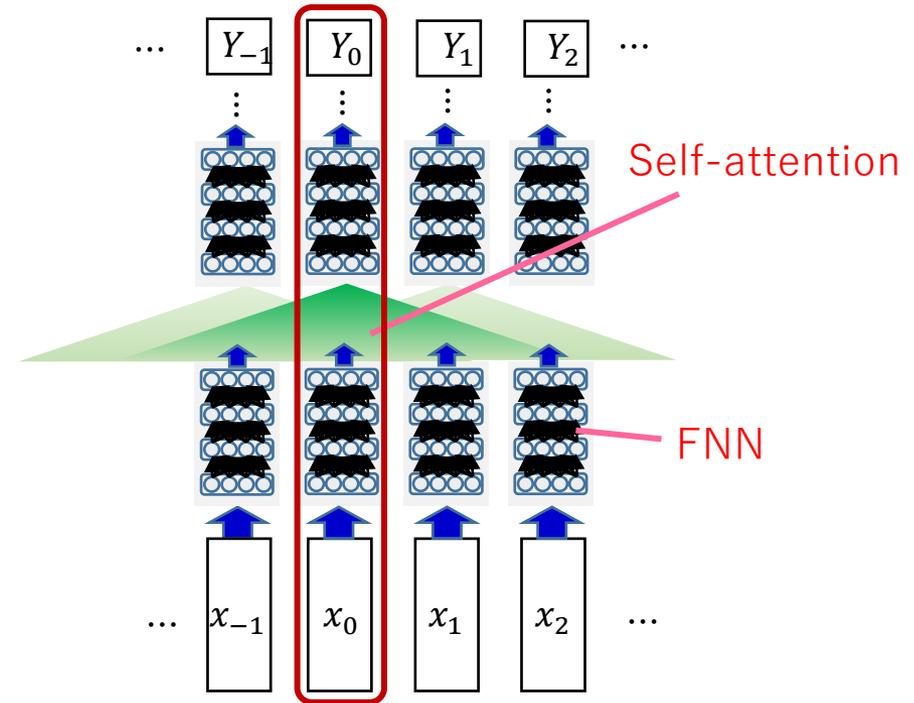
どちらも (ほぼ) **ミニマックス最適** [Yang & Barron, 1999; Niles-Weed & Berthet, 2022].

(Estimator for  $W_1$  distance requires some modification)

[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ぶ。  
→ 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！



## 定理 (推定誤差)

$$\frac{1}{r-l+1} \sum_{j=l}^r$$

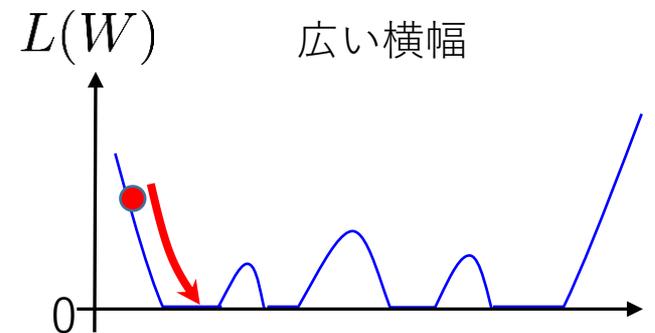
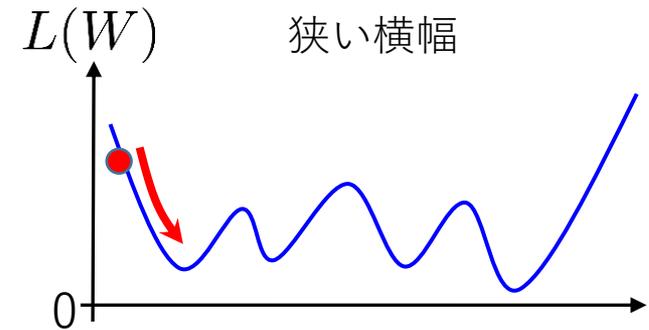
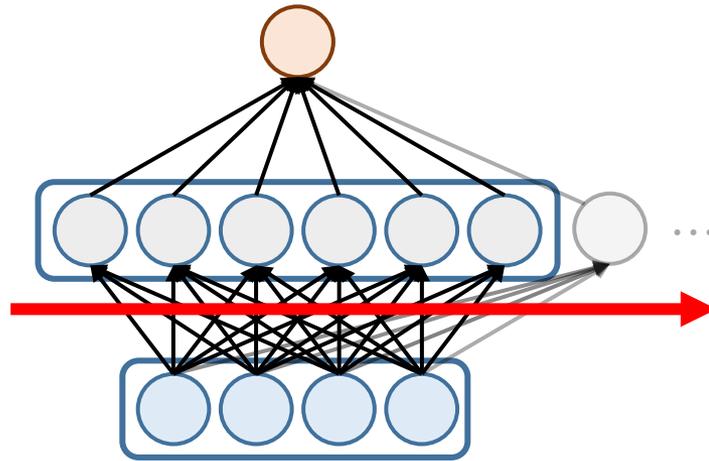
State-Space-Modelといった新しいモデルが提案されているが、少なくともトークンを取捨選択できる性質は必要

$\alpha \in \{4/\alpha, 4\}$

(最適)

- 入力が無限次元でも多項式オーダーの収束レート。

- 横幅が広いと局所最適解が大域的最適解になる。



自由度が上がるため、初期値が最適解(完全フィット)の近くに位置する。

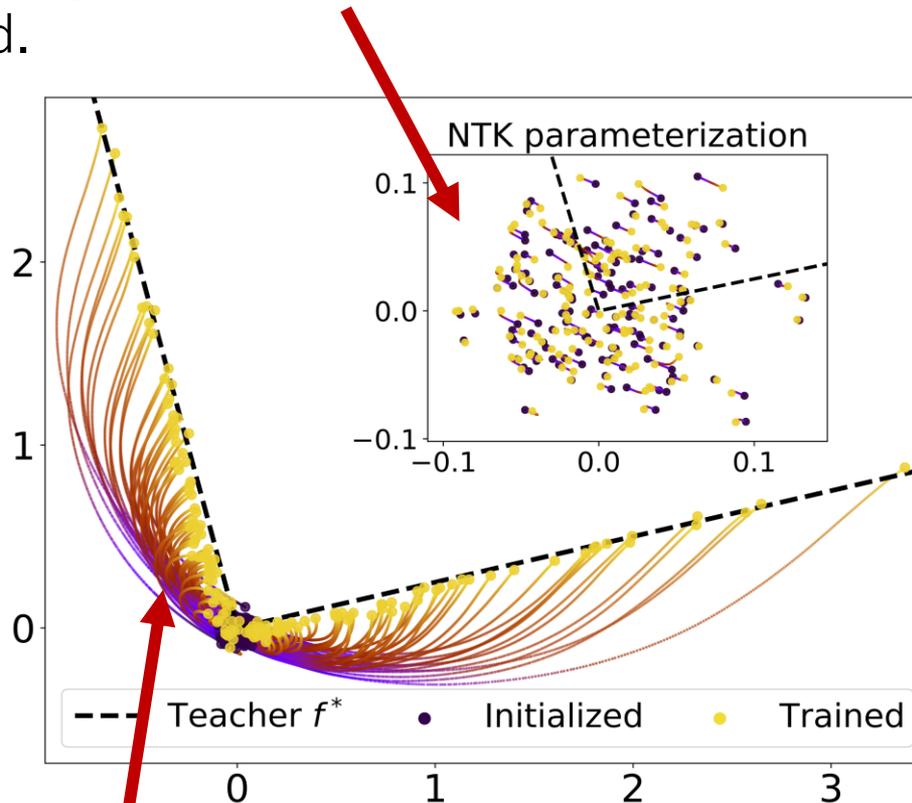
- オーバーパラメトライゼーション
  - Neural Tangent Kernel
  - Mean-field analysis (平均場解析)

[Nitanda & Suzuki, arXiv:1905.09870]

[Nitanda & Suzuki, arXiv:1712.05438.] [Ba, Erdogdu, Suzuki, Wu, Zhang, ICLR2020]

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

**NTK:** Large scale initialization  $\rightarrow$  features are (almost) frozen.



Optimization trajectory of first layer parameters in a 2-layer NN

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

[Ba et al., 2022]

**Mean field:** Small scale initialization  $\rightarrow$  features need to move significantly.

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

# 平均場解析

- ニューラルネットワークの最適化をパラメータの分布最適化としてみなす。

$$f(x) = \frac{1}{M} \sum_{j=1}^M a_j \eta(w_j^\top x) \xrightarrow{M \rightarrow \infty} \int a \eta(w^\top x) \rho(a, w) da dw$$

➡  $(a, w)$ に関する確率密度 $\rho$ による平均とみなせる:

$f$ の最適化  $\Leftrightarrow$   $\rho$ の最適化

$$\frac{d\rho_t}{dt} = -\nabla \cdot (v_t \rho_t)$$

連続方程式

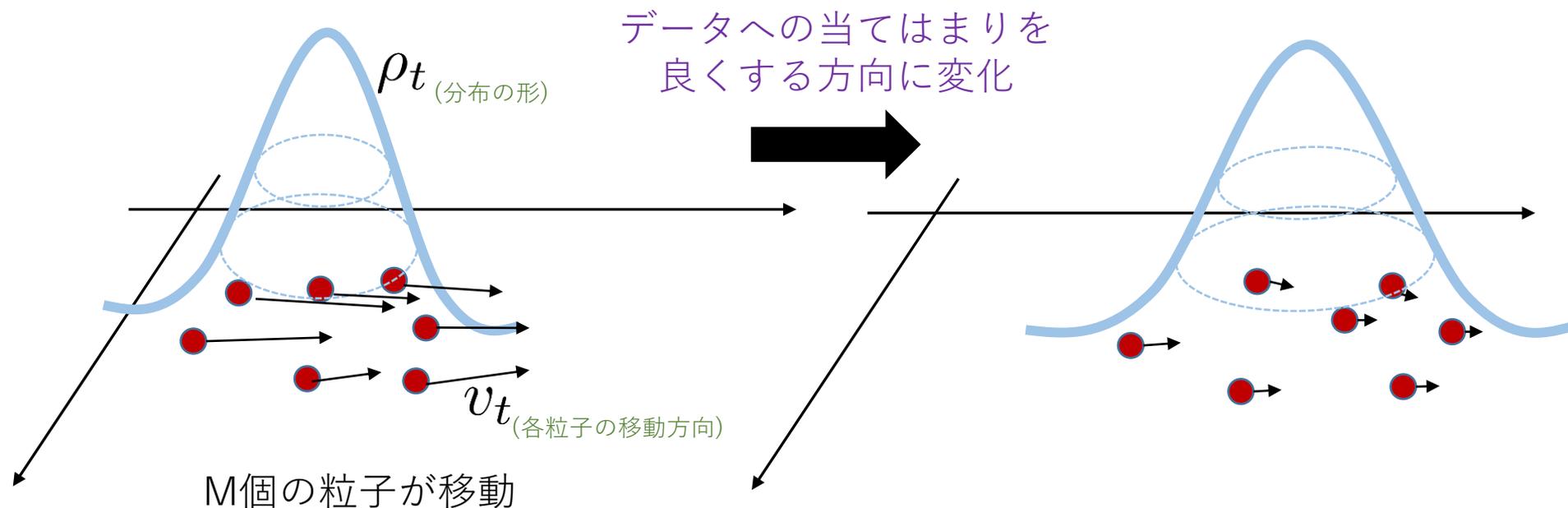
連続方程式, Wasserstein勾配流  
(流体力学, 確率論)

# 粒子勾配降下法

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

1つの粒子

- 各ニューロンのパラメータを一つの粒子とみなす。
- 粒子全体の分布が最適化される。



最適解への収束が示せる。

**連続方程式, Wasserstein勾配流, 無限次元伊藤過程**

(流体力学, 確率論)

$$\mathcal{L}(\mu) := \underline{F(\mu)} + \lambda_2 \text{Ent}(\mu)$$

[非線形汎関数]

(凸は仮定する)

$$\mu^* = \arg \min_{\mu \in \mathcal{P}} \mathcal{L}(\mu)$$

**Remark** GLDの場合

$$F(\mu) = \int L(x) d\mu$$

$$\Rightarrow \frac{\delta F(\mu)}{\delta \mu}(\cdot) = L(\cdot)$$

$$\Rightarrow dX_t = -\nabla L(X_t) dt + \sqrt{2\lambda_2} dB_t$$

**平均場ランジュバン動力学:** (局所線形近似を最小化させる方向に移動)

$$dX_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t) dt + \sqrt{2\lambda_2} dB_t \quad \mu_t = \text{Law}(X_t)$$

Fokker-Planck方程式:

$$\partial_t \mu_t = \nabla \cdot \left[ \mu_t \nabla \frac{\delta F(\mu_t)}{\delta \mu} \right] + \lambda_2 \Delta_x \mu_t$$

$$\frac{d}{dt} \mathcal{L}(\mu_t) = -\lambda_2^2 I(\mu_t || p_{\mu_t})$$

(Fisher divergence)

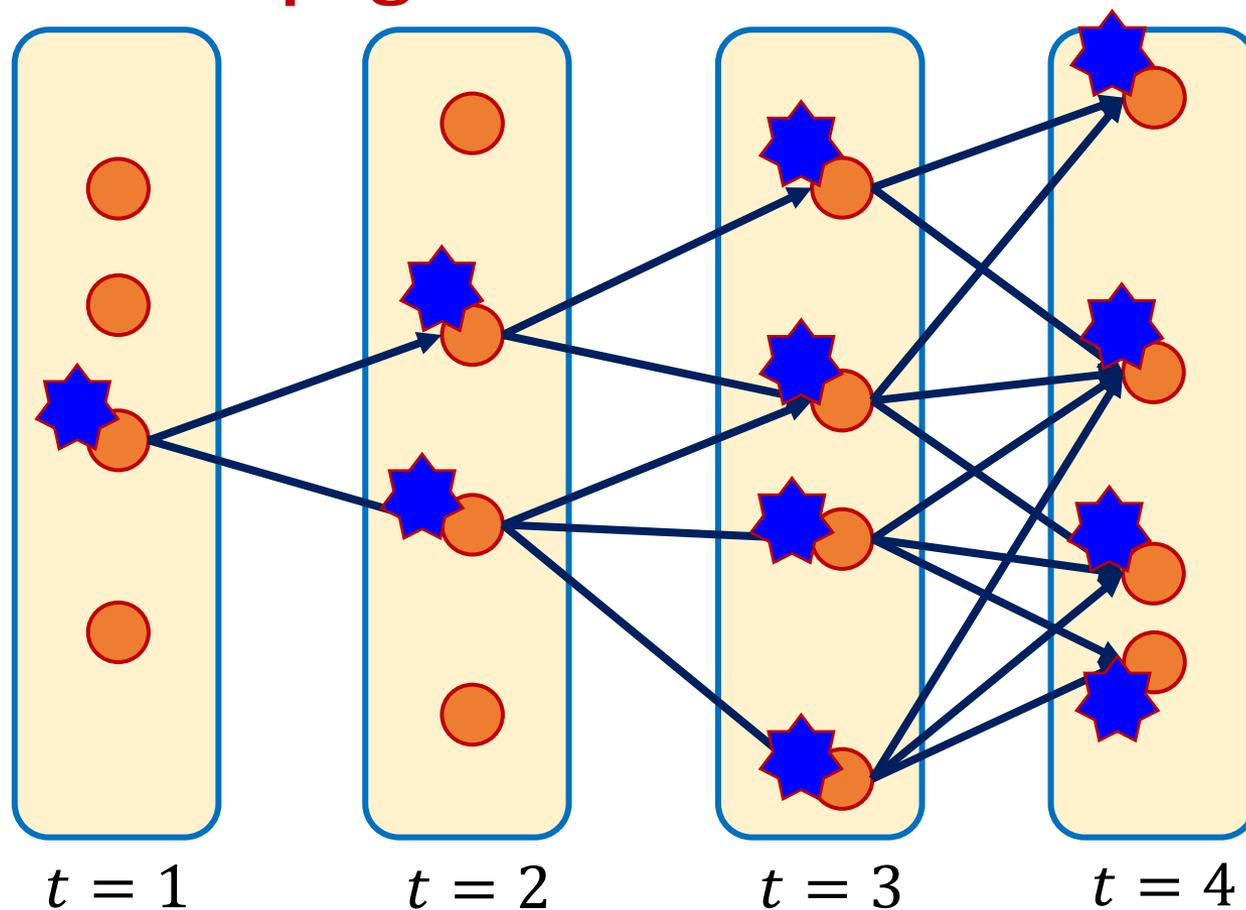
$$p_\mu(x) \propto \exp\left(-\frac{1}{\lambda_2} \frac{\delta F(\mu)}{\delta \mu}(x)\right)$$

これは $\mathcal{L}$ を最小化させるWasserstein勾配流である。  
(適当な滑らかさと凸性の条件より線形収束)

# 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粒子の相互作用の評価が難しい。

## Propagation of chaos (漸近的に独立なように振る舞う)

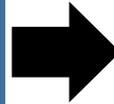


一つの粒子の微小な変化が他の粒子に伝播して増幅される可能性がある。

## Infinite particles / Continuous time

### Linear convergence of mean field Langevin:

[Nitanda, Wu, Suzuki (AISTATS2022)]  
[Chizat (TMLR2022)]



## Finite particle / Discrete time

### Double loop method:

- PDA [Nitanda, Wu, Suzuki: NeurIPS2021]
- P-SDCA [Oko, Suzuki, Wu, Nitanda: ICLR2022]
- Infinite-dim extension [Nishikawa, Suzuki, Nitanda: NeurIPS2022]

Difficult :

**Propagation of chaos (McKean, Kac, ..., 60's)**

## Finite particle / Continuous time

### Uniform-in-time propagation of chaos:

- Super log-Sobolev ineq.  
[Suzuki, Nitanda, Wu (ICLR2023)]
- Leave-one-out type evaluation/Uniform-log-Sobolev  
[Chen, Ren, Wang (arXiv2022)]



## Finite particle / Discrete time

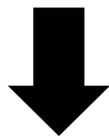
### Single loop method:

Time-space discretization,  
stochastic gradient

- Noisy gradient descent for 2-layer NN

# 離散化誤差

$$dX_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t) dt + \sqrt{2\lambda_2} dB_t$$



(時間離散化)

$$X_{k+1}^{(i)} = X_k^{(i)} - \eta_k v_k^i + \sqrt{2\eta_k \lambda_2} \xi_k^{(i)}$$

where  $\mathbb{E}[v_k^i] = \nabla \frac{\delta F(\hat{\mu}_k)}{\delta \mu}(X_k^i)$  and  $\hat{\mu}_k = \frac{1}{N} \sum_{i=1}^N \delta_{X_k^{(i)}}$

(確率的勾配)

(空間離散化：N粒子)

[Suzuki, Nitanda, Wu (NeurIPS2023)]

## 定理 (離散化誤差バウンド)

$$\mathcal{L}^{(N)}(\hat{\mu}_k) - \mathcal{L}(\mu^*) \lesssim \exp(-\lambda_2 \eta k \alpha) + \frac{1}{\alpha \lambda_2} \left( \underbrace{\eta^2 + \lambda_2 \eta}_{\text{Time discr.}} + \underbrace{\frac{1}{N}}_{\text{Space discr.}} + \underbrace{\frac{(n-B)\sqrt{\eta \lambda_2}}{B(n-1)}}_{\text{Stochastic approx.}} \right)$$

- 一樣对数Sobolev不等式, 超对数Sobolev不等式

## Linear convergence of mean field Langevin dynamics

- Nitanda, Wu, Suzuki (AISTATS2022)

## Stochastic Gradient Langevin dynamics

- Analysis SVRG-GLD: Kinoshita, Suzuki (NeurIPS2022)
- Infinite dimensional SGLD: Muzellec et al. (COLT2022)

## “Double loop” method with convergence guarantee

- PDA: Nitanda, Wu, Suzui (NeurIPS2021)
- P-SDCA: Oko, Suzuki, Wu, Nitanda (ICLR2022)
- Infinite-dim extension: Nishikawa, Suzuki, Nitanda, Wu (NeurIPS2022)

## Superiority of DL with opt. guarantee

- Infinite-dim GLD: Suzuki (NeurIPS2020), Suzuki, Akiyama (ICLR2021)
- Teacher student: Akiyama, Suzuki (ICML2021, ICLR2023)
- Benefit of feature learning (k-parity problem): Suzuki, Wu, Oko, Nitanda (NeurIPS2023)
- Anisotropic regularization: Nitanda, Oko, Suzuki, Wu (ICLR2024)

## Uniform-in-time propagation of chaos:

- Super log-Sobolev inequality: Suzuki, Nitanda, Wu (ICLR2023)

## Analysis of single loop method:

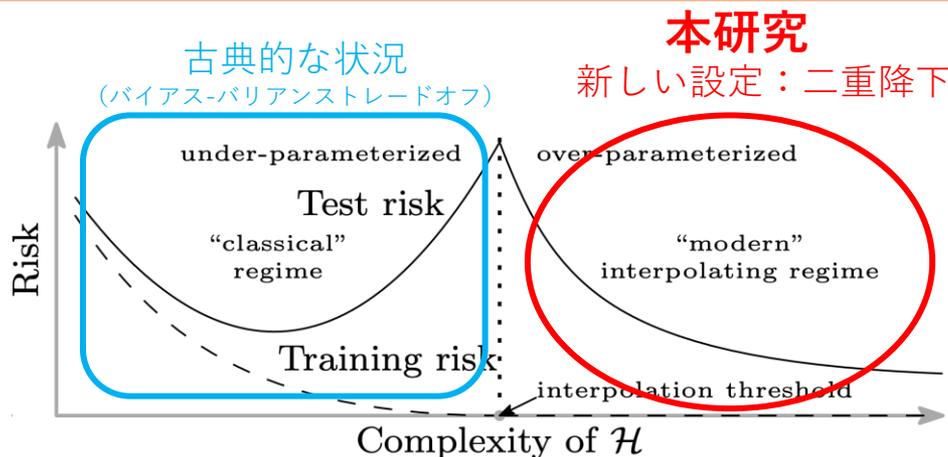
- Convergence analysis with finite particle/discrete time alg. : Suzuki, Nitanda, Wu (NeurIPS2023)

Extension

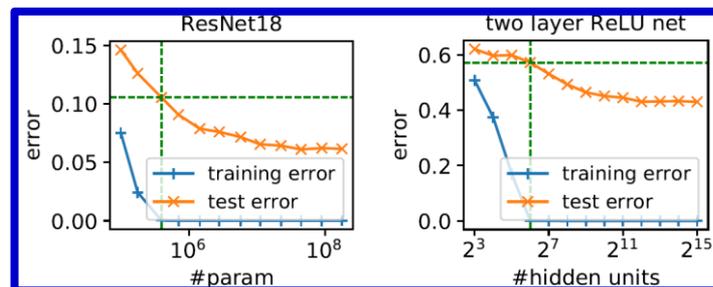
**Extension to RL:** Yamamoto et al. (2023), **Minimax Problem:** Kim et al. (ICLR2024)

# 過剰パラメータNNの予測誤差

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



モデルの次元 > データサイズ  
一見過学習していても汎化する



[Neyshabur et al., ICLR2019]

$$\frac{d\beta(t)}{dt} = -P\nabla L(\beta(t))$$

(前処理付き勾配法)

深層学習の最適化法は色々ある。

- 勾配法 ( $P = I$ )
- 自然勾配法 ( $P = \Sigma_x^{-1}$ )

どれが良い？

真の信号に依存してどの手法が良いかを説明

各学習法によって得られる解の予測誤差の漸近値を解析的に導出

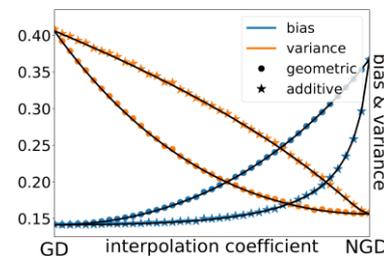
系

## 1. バリエンス

自然勾配法が一番良い。

## 2. バイアス

- 真がモデルに沿っていれば勾配法が良い。
- 真がずれていれば自然勾配法が良い。



# 勾配法と Kernel alignment

[Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, Greg Yang: High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation. NeurIPS2022.]

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

**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？

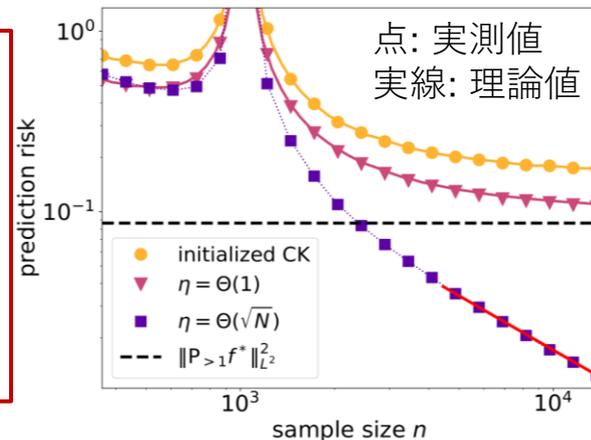
**答**：大きなステップサイズを用いれば，一回の更新で意味のある特徴量の方向を得ることができる。

→ カーネルAlignment, 特徴量学習.

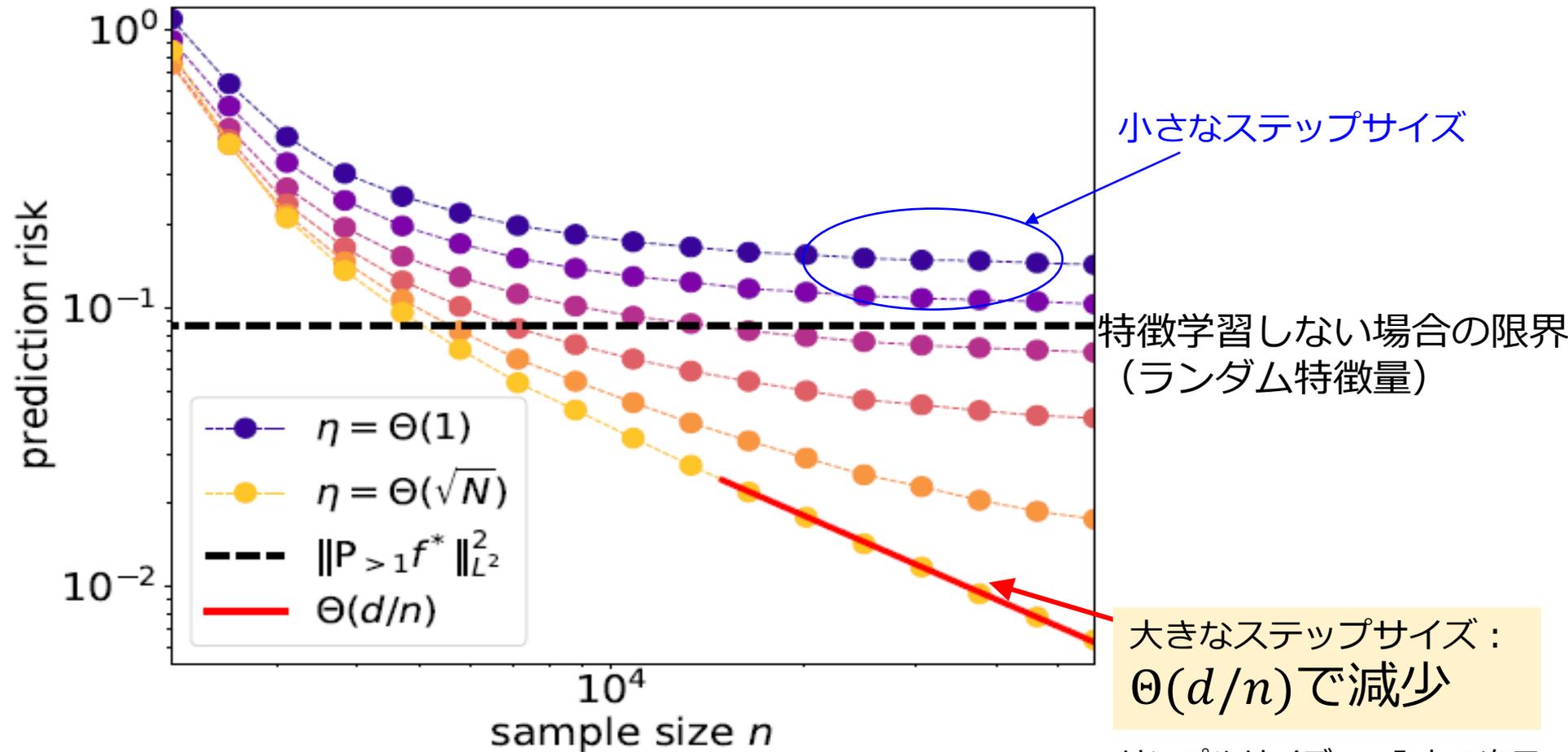
$$W_{k+1} = W_k + \eta \sqrt{N} \nabla L(f_{\text{NN}})$$

$n, d, N \rightarrow \infty$ 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。

- $\eta = \sqrt{N}$ : 大きなステップサイズを用いると，ランダム特徴モデルによるリッジ回帰を優越する。
- $\eta = 1$ : 中間的なステップサイズでは横幅無限大のランダム特徴リッジ回帰を優越しないが初期値 $W$ は優越。
- $\eta = o(1)$ : 小さなステップサイズでは初期値 $W$ と同じ予測誤差 (NTK-regime). 特徴学習の効果なし。



(点線：理論値, 丸印：実験)



( $n$ : サンプルサイズ,  $d$ : 入力の次元)

勾配法一回分による更新後の予測誤差.  
更新に用いるステップサイズごとにプロット

# ノイズあり勾配法の最適性

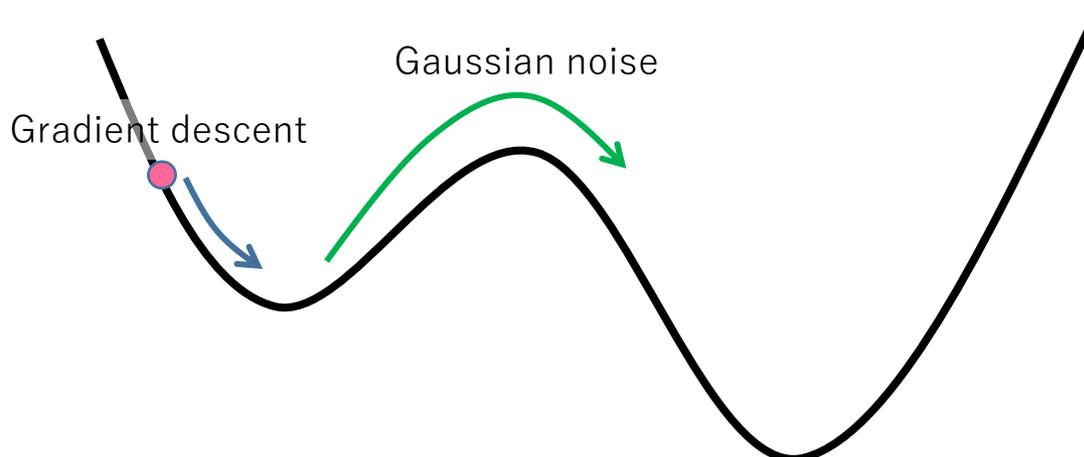
深層学習が浅い学習方法に優れていることを証明

- 大域的最適化が達成可能
- 入力が高次元でも良い予測を達成

Suzuki: Generalization bound of globally optimal non-convex neural network training:  
Transportation map estimation by infinite dimensional Langevin dynamics. NeurIPS2020 (spotlight).

$$X_{n+1} = X_n - \eta \left( \nabla L(X_n) + \frac{\lambda}{2} \nabla \|X_{n+1}\|_{\mathcal{H}_K}^2 \right) + \sqrt{2\frac{\eta}{\beta}} \xi_n$$

$$\int \hat{L}(W_k) d\pi_{(k)}(W_k) - \int \hat{L}(W) d\pi_{\infty}(W) \lesssim \exp(-\Lambda_{\eta}^* k \eta) + \frac{\sqrt{\beta}}{\Lambda_0^*} \eta^{1/2-\kappa}$$



- パラメータが無限個あっても大域的最適解に収束することを証明
- 深層学習は次元の呪いを回避し、カーネル法は次元の呪いを受ける

# 深層学習は次元の呪いを回避する

Suzuki&Akiyama: Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. network training: Transportation map estimation by infinite dimensional Langevin dynamics. ICLR2021 (spotlight).

前ページのノイズあり勾配法で学習した深層NNの予測精度を解析。

**定理**  $n$ : データサイズ

予測誤差  $\mathbb{E}[\|\hat{f} - f^*\|_{L_2(P)}^2]$  の収束レートを導出：

深層

$$n^{-\left(1 + \frac{1}{\gamma}\right)^{-1}}$$

( $\gamma$ : 大)

$$\frac{1}{n}$$

線形 (カーネル法)

$$n^{-\left(1 + \frac{d}{d+11.3}\right)^{-1}}$$

( $d$ : 大)

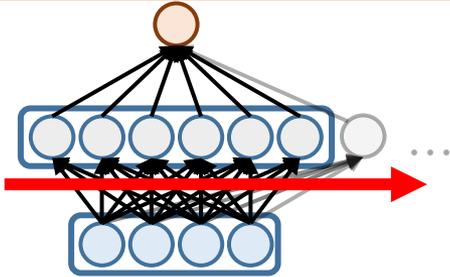
$$\frac{1}{\sqrt{n}}$$

- 深層学習の予測誤差は入力次元に依存しない。
- 浅い学習法はデータの入力次元  $d$  とともに増加する。  
→これは特徴抽出能力の違いによる。

**現実的な最適化の保証ありで性能差を証明**

# Neural Tangent Kernelの理論

Nitanda&Suzuki: Fast Convergence Rates of Averaged Stochastic Gradient Descent under Neural Tangent Kernel Regime, ICLR2021 (oral). **Outstanding paper award.**



ニューラルネットワークの最適化は非凸最適化  
→ 横幅の広いネットワークはより「凸」っぽくなる。

- 確率的最適化により最適な推定レートを達成可能。
- ネットワーク固有の周波数成分のスペクトルが学習効率を決める。

Thm

$f_T$ :  $T$ 回更新後の解

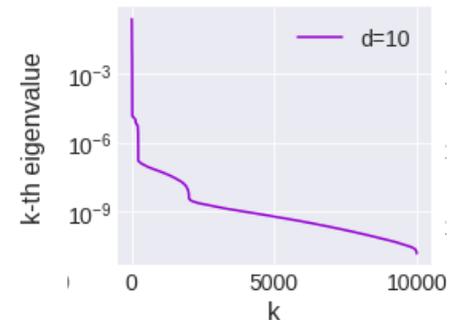
$$\mathbb{E}[\|f_T - f^*\|_{L_2}^2] \leq \epsilon_M + O\left(T^{-\frac{2r\beta}{2r\beta+1}}\right)$$

横幅  $M \rightarrow \infty$  で 0 に収束する項

速い学習レート  
( $O(1/\sqrt{T})$  より速い)

NTK (Neural Tangent Kernel)  
の固有値の減衰レート

NTKのスペクトル



低周波成分

高周波成分

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。

# 分散縮小型確率的勾配Langevin動力学 30

[Kinoshita, Suzuki: Improved Convergence Rate of Stochastic Gradient Langevin Dynamics with Variance Reduction and its Application to Optimization. 2022. arXiv:2203.16217]

**Euler-Maruyama スキーム**

$$\nabla_k = \frac{1}{n} \sum_{i=1}^n \nabla f_i(X_k)$$

$$X_{k+1} = X_k - \eta \nabla_k + \sqrt{2\eta/\gamma} \epsilon_k$$

**分散縮小型SGLD**

$$\tilde{\nabla}_k = \frac{1}{B} \sum_{i_k \in I_k} (\nabla f_{i_k}(X_k) - \nabla f_{i_k}(\tilde{X}^{(s)}) + \nabla F(\tilde{X}^{(s)}))$$

分散縮小

$$X_{k+1} = X_k - \eta \tilde{\nabla}_k + \sqrt{2\eta/\gamma} \epsilon_k$$

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$$

$$\frac{d\nu}{dx}(x) \propto \exp(-\gamma f(x))$$

- 非凸最適化
- サンプリング

- 確率的勾配 + 分散縮小
- 「対数Sobolev不等式の条件」で「KL-divergence」の収束.

Method	Major Assumptions	Criterion*	Criterion $\leq \epsilon$ となるまでの Gradient Complexity**
Dalalyan (2017a)	Smooth, Log-concave (M)	2-Wass.	$\tilde{O}\left(\frac{nd}{\epsilon^2} \cdot \text{poly}(M, L)\right)$
Xu et al. (2018)	Smooth, Dissipative	Weak conv.	$\tilde{O}\left(\frac{nd}{\epsilon}\right) \cdot e^{\tilde{O}(d)}$
Vempala et al. (2019)	Smooth, Log-Sobolev ( $\alpha$ )	KL	$\tilde{O}\left(\frac{n}{\epsilon} \cdot d\gamma^2 L^2 \alpha^{-2}\right)$
Zou et al. (2018)	Smooth, Log-concave (M)	2-Wass.	$\tilde{O}\left(n + \frac{n^{3/4}}{\epsilon^2} + \frac{L^{3/2} n^{1/2} d^{1/2}}{M^{3/2} \epsilon}\right)$
Zou et al. (2019a)	Smooth, Dissipative	2-Wass.	$\tilde{O}\left(n + \frac{n^{3/4}}{\epsilon^2} + \frac{n^{1/2}}{\epsilon^4}\right) \cdot e^{\tilde{O}(\gamma+d)}$
Zou et al. (2021)	Smooth, Dissipative, Warm-start	TV	$\tilde{O}\left(\frac{\gamma^2}{\epsilon^2}\right) \cdot e^{\tilde{O}(d)}$
Zou et al. (2019b)	Smooth, Dissipative	2-Wass.	$\tilde{O}\left(\left(n + \frac{n^{1/2}}{\epsilon^2 \mu^{3/2}}\right) \cdot \sqrt{n} \text{倍速い}\right)$
<b>SVRG-LD/SARAH-LD</b> ( $B = m = \sqrt{n}$ )	Smooth, Log-Sobolev ( $\alpha$ )	KL	$\tilde{O}\left(\left(n + \frac{dn^{1/2}}{\epsilon}\right) \cdot \gamma^2 L^2 \alpha^{-2}\right)$

条件緩和



## 勾配ランジュバン動力学の一般論

- 分散縮小型確率的勾配ランジュバン動力学の収束解析 (NeurIPS2022)
- 無限次元勾配ランジュバン動力学のアルゴリズムと収束理論 (COLT2022)

## 平均場勾配ランジュバン動力学の収束解析とNN最適化への応用

- 無限次元入力ニューラルネットの無限次元GLDによる最適化 (NeurIPS2022)
- 平均場ニューラルネットワークのGLDによる最適化の収束 (AISTATS2022)
- 平均場ニューラルネットワークの新しい最適化手法: 粒子確率的双対座標上昇法 (ICLR2022)
- 有限粒子近似の近似誤差解析 (ICLR2023)

Yuri Kinoshita, Taiji Suzuki: Improved Convergence Rate of Stochastic Gradient Langevin Dynamics with Variance Reduction and its Application to Optimization. NeurIPS2022.

(勾配ランジュバン動力学)

$$X_{k+1} = X_k - \eta \nabla f(X_k) + \sqrt{2\eta/\gamma} \xi_k$$

計算にO(n)かかる (大規模データで困る)  
→ 確率的勾配を用いる  $\tilde{v}_k = \frac{1}{B} \sum_{i \in I_k} f_i(X_k)$

サンプリングや非凸最適化に有用  
(拡散過程生成モデルとも関連)

勾配計算量

- Vempala&Wibisono (2019): 非確率的勾配

$$\tilde{O}\left(\frac{n}{\epsilon} d \gamma^2 L^2 \alpha^{-2}\right)$$

- 我々の結果: 確率的勾配+分散縮小法

$$\tilde{O}\left(\left(n + \frac{\sqrt{nd}}{\epsilon}\right) \gamma^2 L^2 \alpha^{-2}\right)$$

**:  $\sqrt{n}$ 倍高速**

例:

- 二次関数+有界関数
- Weak Morse型関数

Naoki Nishikawa, Taiji Suzuki, Atsushi Nitanda, Denny Wu: Two-layer neural network on infinite dimensional data: global optimization guarantee in the mean-field regime. NeurIPS2022.

### 無限次元入力NNの無限次元GLDによる最適化

有限次元入力

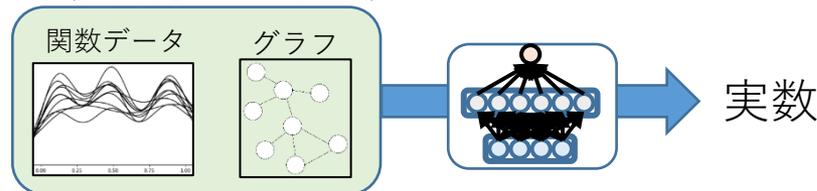
$$f(x) = \frac{1}{M} \sum_{j=1}^M r_j \sigma(w_j^\top x)$$

無限次元入力

$$f(x) = \frac{1}{M} \sum_{j=1}^M r_j \sigma(\langle w_j, x \rangle_{\mathcal{H}})$$

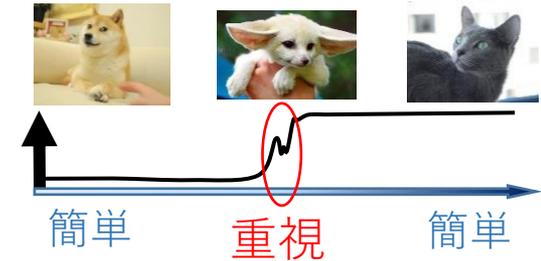
$w_j, x \in \mathcal{H}$  (無限次元ヒルベルト空間)

$x \in \mathcal{H}$  (無限次元ヒルベルト空間)



- 確率的粒子双対平均加法 (PDA), 確率的粒子双対座標降下法 (PSDCA) の両手法に対して, 無限次元入力バージョンを提案.
- 内部ループにて無限次元勾配ランジュバン動力学の理論を援用 (Suzuki et al. COLT2022)

- 深層学習は重要な情報に絞って特徴抽出  
→ 構造的に冗長性が現れる。



**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：**

- 圧縮型バウンド
- ノルム型バウンド
- ...



# 研究成果まとめ

理論

汎化誤差理論を基にした圧縮手法の提案

応用

圧縮可能性による  
汎化誤差解析  
AISTATS2018

実用的圧縮手法の提案  
「Spectral-Pruning」  
IJCAI-PRICAI2020

非圧縮ネットワークの  
圧縮型汎化誤差解析  
ICLR2020

ドメイン適合への応用  
BMVC2020

理論と応用  
の融合

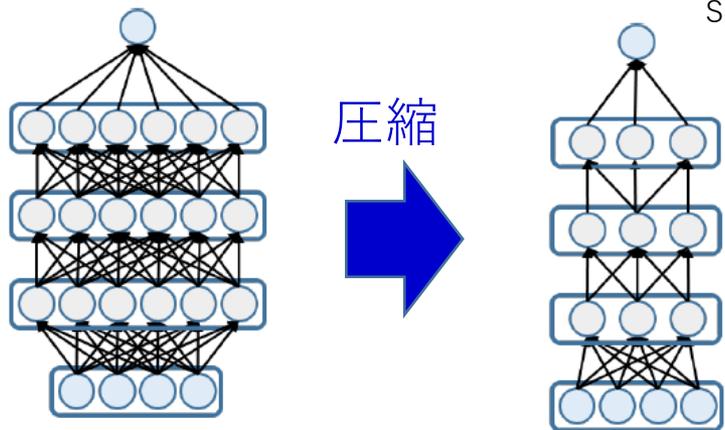
テンソル分解への展開  
AISTATS2020

転移学習への応用  
ICCV2019, Neural  
Architects Workshop

ネットワークは「いかなる時に」「どれくらい」圧縮できるか。  
その圧縮可能性はどう汎化性能に影響するか？

# ニューラルネットワークの圧縮

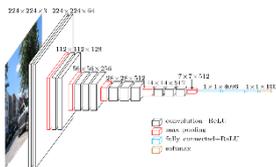
[Suzuki, Abe, Murata, Horiuchi, Ito, Wachi, Hirai, Yukishima, Nishimura: Spectral-Pruning: Compressing deep neural network via spectral analysis, 2018]



- メモリ消費量を減少
- 予測にかかる計算量を減少
- 小型デバイスでの作動に有利 (自動運転など)



## VGG-16ネットワークの圧縮



Model	Top-1	Top-5	# Param.	FLOPs
Original VGG	68.34%	88.44%	138.34M	30.94B
APoZ-2	70.15%	89.69%	51.24M	30.94B
ThiNet-Conv	69.80%	89.53%	131.44M	9.58B
ThiNet-GAP	67.34%	87.92%	8.32M	9.34B
<b>Spec-Conv</b>	<b>70.418%</b>	<b>90.094%</b>	131.44M	9.58B
<b>Spec-GAP</b>	<b>67.540%</b>	<b>88.270%</b>	<b>8.32M</b>	9.34B

提案手法：  
従来手法より良い精度

94%の圧縮  
(精度変わらず)

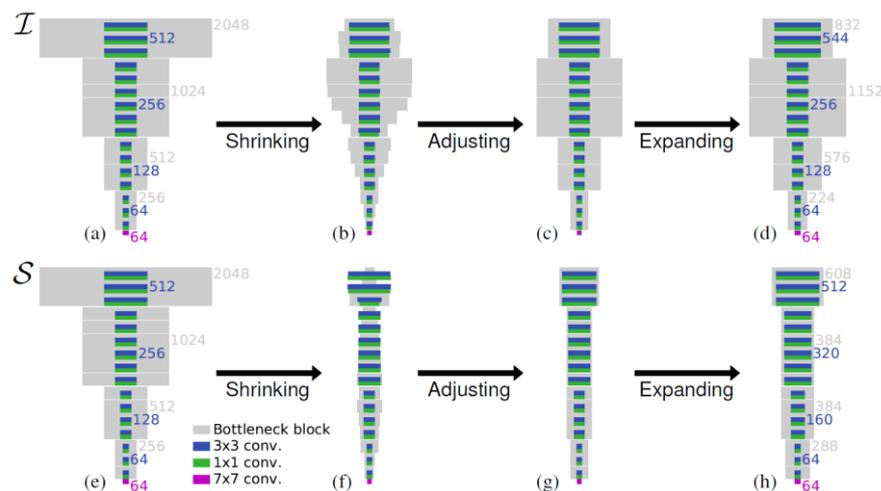
## ResNet-50ネットワークの圧縮

Model	Top-1	Top-5	# Param.	FLOPs
ResNet-50-1	72.89%	91.07%	25.56M	7.75G
ThiNet-70	72.04 %	90.67%	16.94M	4.88G
ThiNet-50	71.01 %	90.02%	12.38M	3.41G
NISP-50-A	72.68%	—	18.63M	5.63G
NISP-50-B	71.99%	—	14.57M	4.32G
<b>Spec-ResA</b>	<b>72.99%</b>	<b>91.56%</b>	12.38M	3.45G
ResNet-50-2	75.21%	92.21%	25.56M	7.75G
Sparse-reg wo/ ft	—	84.2%	19.78M	5.25G
Sparse-reg w/ ft	—	90.8%	19.78M	5.25G
<b>Spec-ResB wo/ ft</b>	<b>66.12%</b>	<b>86.67%</b>	20.69M	5.25G
<b>Spec-ResB w/ ft</b>	<b>74.04%</b>	<b>91.77%</b>	20.69M	5.25G

約半分に圧縮しても精度落ちず

# 転移学習のネットワーク構造決定

- ある閾値以上の固有値をカウント (e.g.,  $10^{-3}$ ).  
→ 縮小したネットワークのサイズとして使う.
- その後, スクラッチから学習 ( $\mathcal{S}$ ) もしくはImageNet事前学習モデルをファインチューニングする ( $\mathcal{I}$ ).



## Network size determination alg.

- 1: Set initial weights.
- 2: Train the whole network to find  $\theta^* = \operatorname{argmin}_{\theta} \mathcal{L}(\theta)$ .
- 3: Calculate eigenspectra  $S_{1:M}$ .
- 4: Calculate intrinsic dimensionalities  $d_{1:M}$  by  $d_{1:M} = \operatorname{len}(S_{1:M} > T)$ .
- 5: Determine new widths  $O'_{1:M}$  by adjusting  $d_{1:M}$ .
- 6: Find the largest  $\omega$  such that  $c(\omega \cdot O'_{1:M}) \leq \zeta$ .
- 7: **return**  $\omega \cdot O'_{1:M}$ .

Backbone	Normalization	Classification		COCO (2× schedule)					
		MACs	#params	AP	AP <sub>50</sub>	AP <sub>75</sub>	AP <sub>S</sub>	AP <sub>M</sub>	AP <sub>L</sub>
ResNet-50 [35]	SyncBN	3.8 G	—	34.5	55.2	37.7	20.4	36.7	44.5
ResNet-50*	GN	4.09 G	25.5 M	35.5	55.6	38.5	21.3	37.5	45.3
ResiaxNet $\mathcal{S}$ 3-50 (MACs)	GN	4.06 G	18.6 M	35.4	55.4	38.6	21.5	37.3	45.2
ResiaxNet $\mathcal{I}$ 1-50 (MACs)	GN	4.05 G	21.7 M	35.5	55.5	38.6	21.4	37.3	<b>46.0</b>
ResiaxNet $\mathcal{I}$ 3-50 (MACs)	GN	4.07 G	22.0 M	35.4	55.6	38.4	21.3	37.8	45.5
ResiaxNet $\mathcal{I}$ 3-50 (params)*	GN	4.92 G	24.7 M	<b>35.8</b>	<b>55.9</b>	<b>38.9</b>	<b>21.8</b>	<b>38.0</b>	45.6
DetNet-59 [35]	SyncBN	4.8+ G	—	<b>36.3</b>	<b>56.5</b>	<b>39.3</b>	22.0	<b>38.4</b>	<b>46.9</b>
DetNet-59 <sup>†</sup>	GN	5.00+ G	18.3+ M	36.2	56.0	<b>39.3</b>	22.1	38.3	46.0
DetiaxNet $\mathcal{I}$ 2-59 (MACs)	GN	4.94+ G	17.4+ M	36.2	56.0	<b>39.3</b>	<b>22.5</b>	38.1	46.0

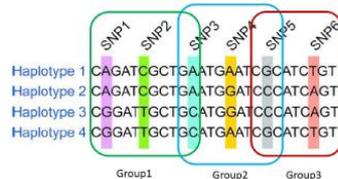


# 確率的最適化 [計算理論]

(構造的)正則化学習

$$\min_{\mathbf{x} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{w}_i^T \mathbf{x}) + \lambda \underbrace{\psi(\mathbf{x})}_{\text{複雑な正則化}}$$

大量サンプル



ゲノムワイド相関解析  
グループ正則化



画像のデノイジング  
Total-Variation正則化

## 従来法

**勾配法**：定番の最適化手法

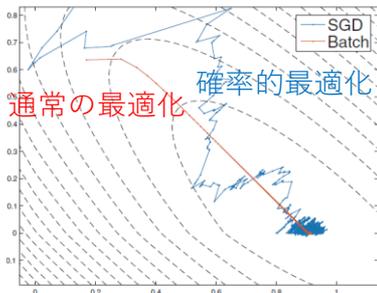
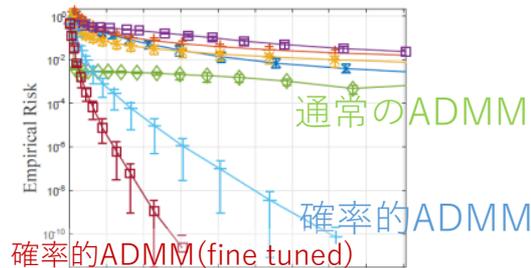
欠点：各更新ごとに全データ点を読むので大きいデータでは非効率。

## 改善法

**確率的最適化法**

各更新で1個のデータ点しか使わない。

→計算効率を大きく改善。



## 計算量比較

誤差  $\epsilon$  を達成するまでの計算量

通常勾配法

$$n\sqrt{\kappa} \log(1/\epsilon)$$

$n$ : サンプルサイズ,  $\kappa$ : 条件数

確率的最適化法

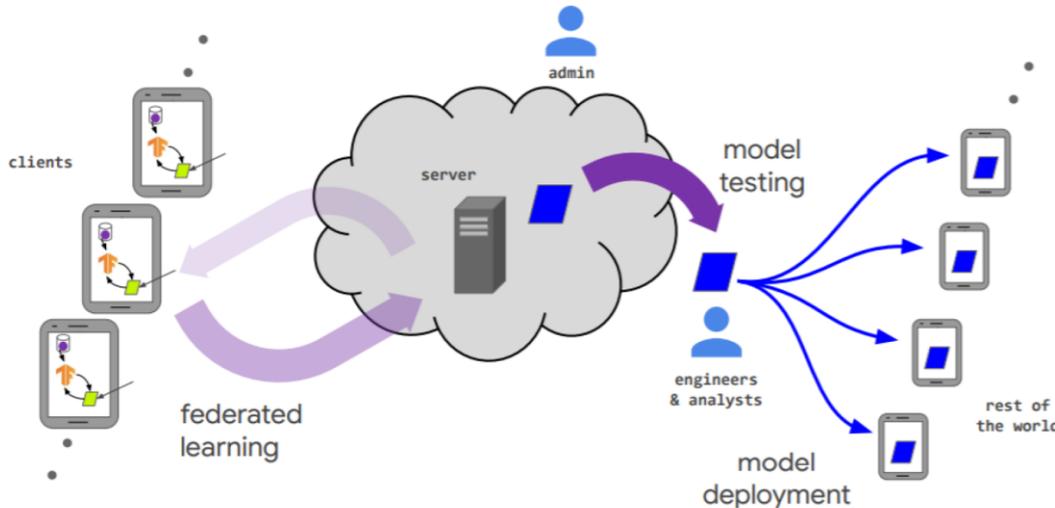
$$(n + \sqrt{n\kappa}) \log(1/\epsilon)$$

$n = 10^6$  なら最大1000倍の高速化

- オンライン型確率的交互方向乗数法 (ICML2013, 2012年度IBISIML 研究会賞)
- 確率的交互方向乗数法による双対座標降下法 (ICML2014)
- グラフ型正則化への応用と加速法 (第28回IBISIML研究会)
- 確率的DC計画 (Nitanda&Suzuki, AISTATS2017)
- 二重確率的分散縮小勾配法 (Murata&Suzuki, NIPS2017; **2016年度IBISIML研究会賞**)



# Federated learningの確率的最適化



[1] Advances and Open Problems in Federated Learning, Kariouzet al., 2019

[Murata, Suzuki: Bias-Variance Reduced Local SGD for Less Heterogeneous Federated Learning. arXiv:2102.03198]

- Communication complexity:

- Local SGD:  $\frac{1}{B\varepsilon} + \frac{1}{BP\varepsilon^2} + \frac{1}{\sqrt{B}\varepsilon^{3/2}} + \frac{\zeta_1}{\varepsilon^{3/2}}$  ← Worse than naïve minibatch SGD!

B: minibatch size, P: number of workers,  $\zeta_i$ : **heterogeneity of workers**

- Our proposal: Bias-Variance Reduced Local SGD method

	Communication Complexity	Communication Complexity ( $B \rightarrow \infty$ )	Assumptions
Minibatch SGD	$\frac{L}{\varepsilon^2} + \frac{\sigma^2}{PB\varepsilon^4}$	$\frac{L}{\varepsilon^2}$	2, 3 and stochastic gradient variance $\leq \sigma^2$
BVR-L-SGD	$\frac{L}{\sqrt{B}\varepsilon^2} + \frac{\sqrt{n}L}{BP\varepsilon^2} + \frac{\zeta_2}{\varepsilon^2} + \frac{n}{BP}$	$\frac{\zeta_2}{\varepsilon^2}$	1, 2, 3

If  $\zeta_2 = o(L)$ , BVR-L-SGD surpasses minibatchSGD!

- 深層学習はなぜ性能が良いのか？
  - 汎化誤差理論 (統計的学習理論)
  - 関数近似理論 (関数解析)
  - 最適化理論 (一次最適化法)
- 確率論を用いた学習アルゴリズムの解析・提案
  - 最適輸送理論, Wasserstein幾何
  - 勾配流
- 高次元統計・確率的最適化
- 企業との共同研究
  - 深層学習・連合学習手法の開発 (東芝)

## 理論を土台にした普遍性の高い基礎研究

### • 統計的学習の理論

#### • 深層学習の理論

- 大規模基盤モデルの理論
- 最適化理論
- 企業との共同研究も複数進行中

#### • 高次元統計

- 過剰パラメータ化, 良性過学習, 特徴学習

#### • カーネル法

### • 機械学習高速計算技法の開発

- 大規模確率的最適化
- 効率的に解ける問題クラスの拡張 (非凸最適化)



理研AIP Youtubeチャンネル  
AIP Open Seminar #20 20210407