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

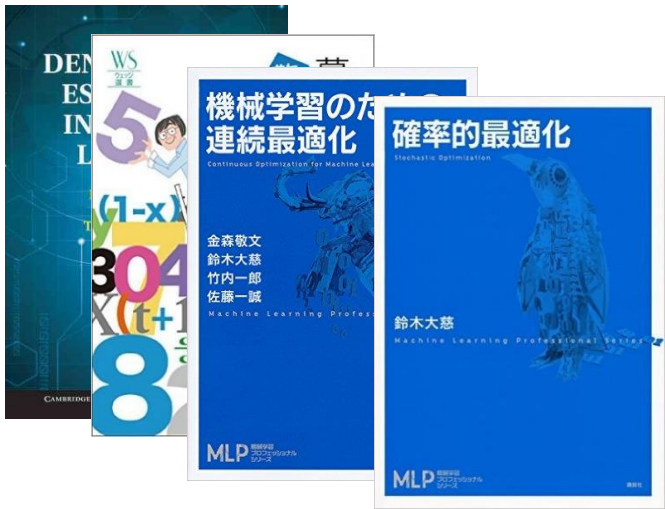
2023年

略歴

- 2009年　東京大学情報理工学系数理情報学専攻　博士課程修了
- 2009年　同専攻助教
- 2013年7月　東京工業大学数理・計算科学専攻准教授
- 2014年10月　さきがけ研究員・兼任（数学協働領域）
- 2017年4月　東京大学数理情報科学専攻准教授
- 2016年10月 **理研AIP「深層学習理論チーム」チームリーダー・兼任**

専門：統計的学習，数理統計学，最適化技法

ノンパラメトリック統計（カーネル法），スパース高次元推定手法，テンソル推定，確率的最適化



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
- モデル圧縮
- 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

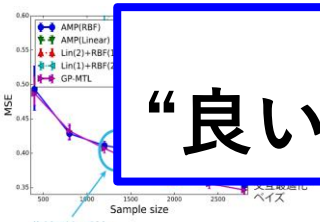
計算手法の理論
「高速計算」手法の構築

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
  - DC計画

最適収束速度を達成するアルゴリズムの導出

「より速くより正確に」

“良い学習”はどうすれば達成できるか？



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

理論的課題

我々の結果

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？

→ Besov空間での近似理論
積分表現理論
グラフCNNの表現能力
(適応的関数近似)

汎化能力

有限個のデータで学習し，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？

→ Besov空間での学習理論
再生核理論による自由度解析
モデル圧縮型汎化理論
スパース理論との融合
最適化との融合
(過剰パラメータ・二重降下)

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？

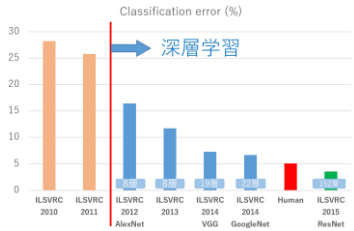
→ ノイズ付勾配法の大域的最適性
粒子勾配降下法
並列計算による高速化
確率的加速勾配法
(NTK, 平均場理論)

実問題への還元 (企業との連携)：モデル圧縮，並列計算

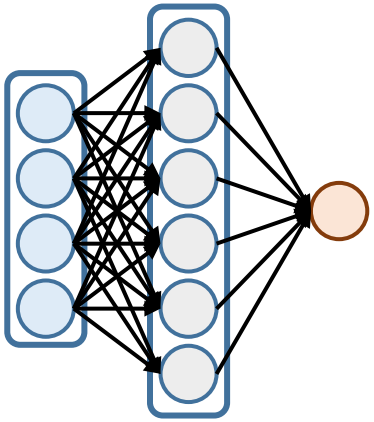
深層学習の表現能力

- **[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.

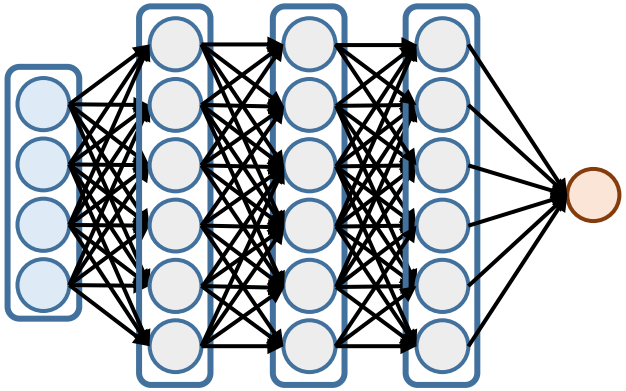
  → この差はどう埋める？



カーネル法
従来法



多層ニューラルネット
深層学習



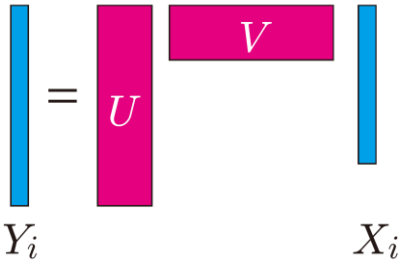
[鈴木大慈: 文部科学大臣表彰・若手科学者賞. 2020年4月7日]

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

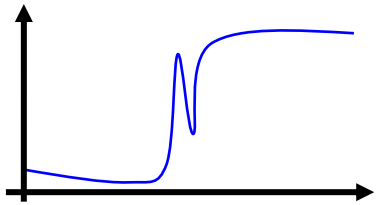
特徴空間の次元が低い状況は深層学習が得意



Besov空間

[Suzuki, 2019]

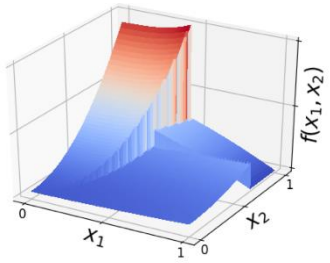
滑らかさが非一様な関数の推定は深層学習が得意



区分滑らかな関数

[Imaizumi&Fukumizu, 2019]

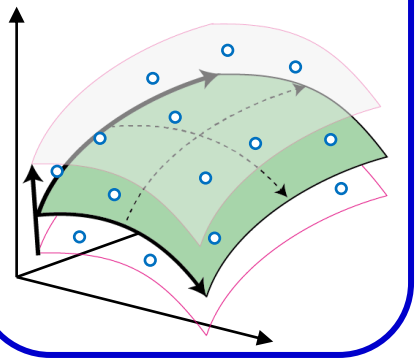
不連続な関数の推定は深層学習が得意



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

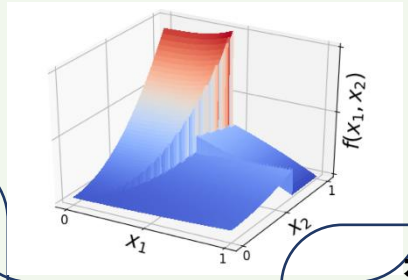


深層	$\frac{r(M+N)}{n}$	$n^{-\frac{2s}{2s+d}}$	$n^{-\frac{2s}{2s+d}} \vee n^{-\frac{\alpha}{\alpha+D-1}}$	$n^{-\frac{2s}{2s+D}}$
カーネル	$\frac{MN}{n}$	$n^{-\frac{2s-2d(1/p-1/2)_+}{2s+d-2d(1/p-1/2)_+}}$	$\frac{1}{\sqrt{n}}$	$n^{-\frac{2(s-D/p+d/2)}{2(s-D/p+d/2)+d}} \vee n^{-\frac{2s}{2s+D}}$

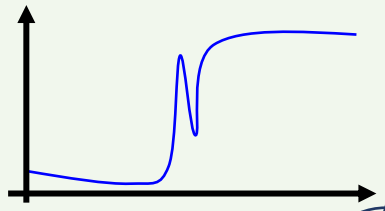
推定精度

数学的一般化

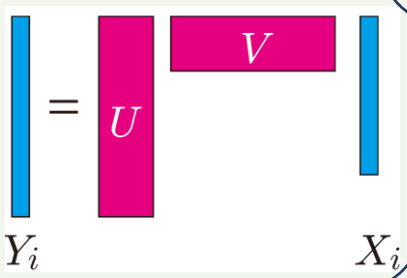
区分滑らかな関数



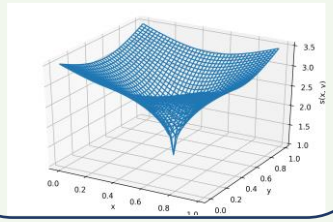
Besov空間



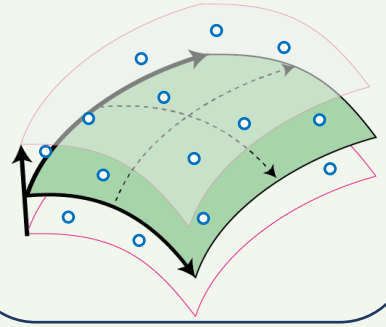
縮小ランク回帰



変動指数
Besov空間



低次元データ



非凸性
スパース性

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

無限次元入力NN

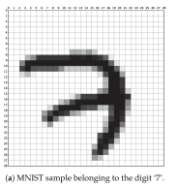
無限次元入力

(画像, 音声信号, 自然言語,...)

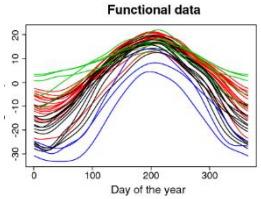
[Okumoto&Suzuki: Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. ICLR2022. Spotlight]

無限 (高) 次元データ

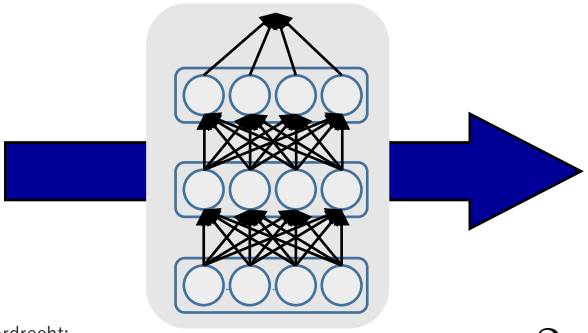
画像データ　　関数データ



(a) MNIST sample belonging to the digit '7'.



- 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.]

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

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

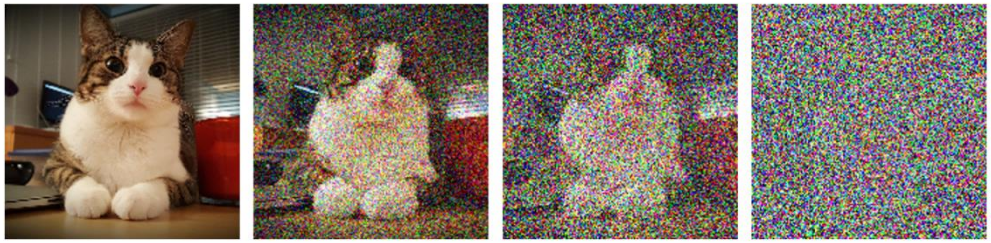
拡散モデルの統計理論

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

Stable diffusion, 2022.

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

Forward process



Backward process

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

経験スコアマッチング推定量:

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

定理

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

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

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

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]

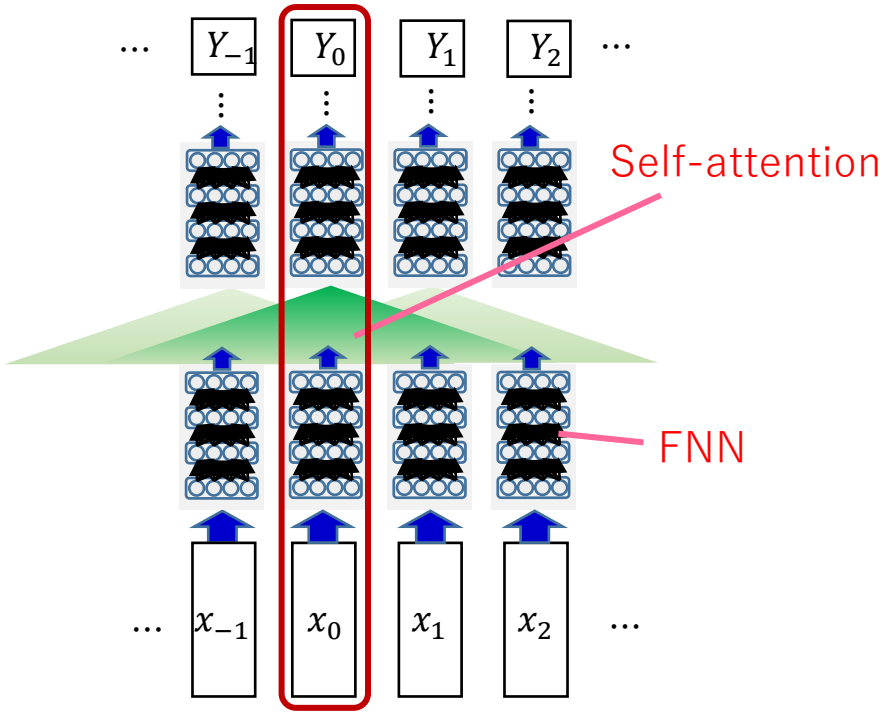
Transformerの性質

- 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！



定理（推定誤差）

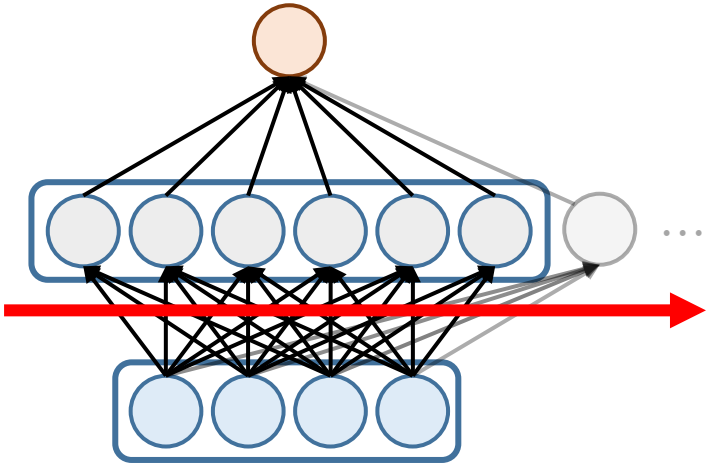
$$\frac{1}{r-l+1}\sum_{j=l}^{r}\mathbb{E}[\|\hat{F}_j - F_j^\circ\|_{L_2(P_X)}^2] \lesssim n^{-\frac{2a^\dagger}{2a^\dagger+1}}(\log n)^{2/\alpha+2+\max\{4/\alpha,4\}}$$

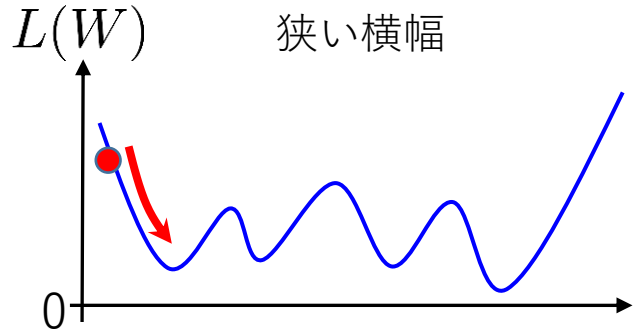
（ほぼミニマックス最適）

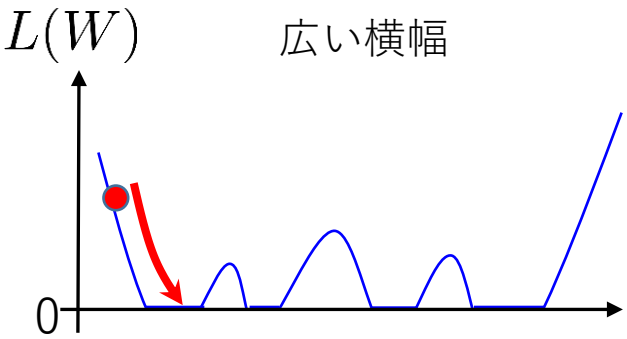
➢ 入力が無限次元でも多項式オーダーの収束レート.

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 [Nitanda &Suzuki, arXiv:1905.09870]
  - Mean-field analysis (平均場解析)

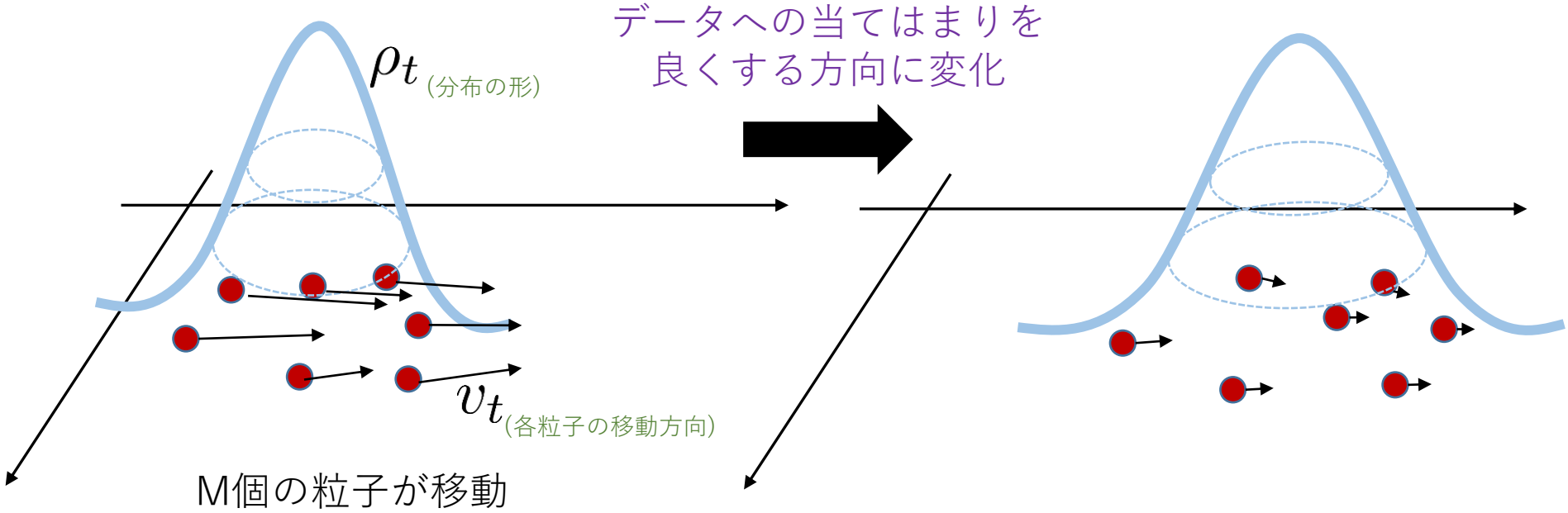
[Nitanda &Suzuki, arXiv:1712.05438.][Ba,Erdogdu,Suzuki, Wu, Zhang, ICLR2020]

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

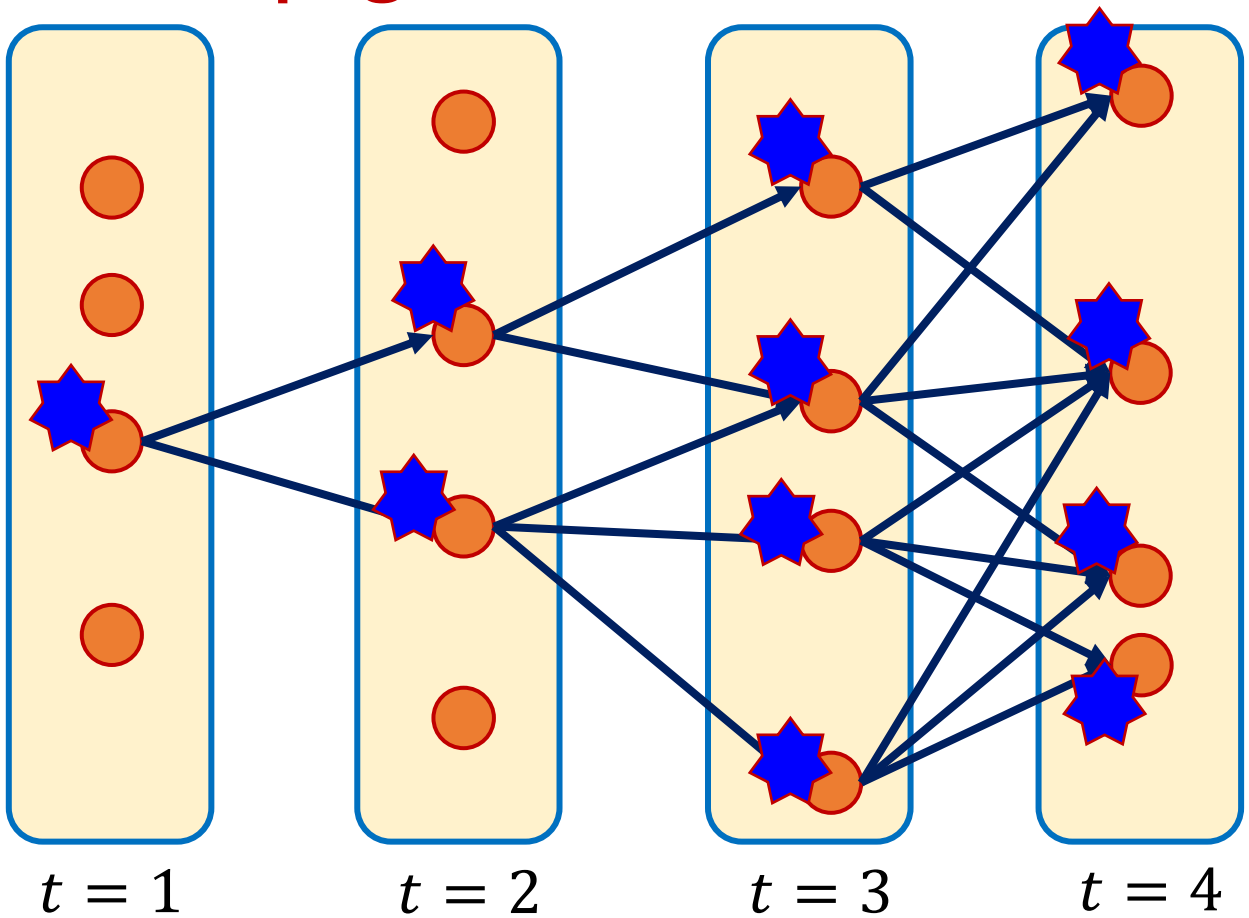
(流体力学，確率論)

[Nitanda&Suzuki, 2017][Chizat&Bach, 2018][Mei, Montanari&Nguyen, 2018][Suzuki, 2020]

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．

正則化学習の再定式化

[Nitanda, Wu, Suzuki: Particle Dual Averaging: Optimization of Mean Field Neural Networks with Global Convergence Rate Analysis. NeurIPS2021]
[Oko, Suzuki, Nitanda, Wu: Particle Stochastic Dual Coordinate Ascent: Exponential convergent algorithm for mean field neural network optimization. ICLR2022]
[Nitanda, Wu, Suzuki: Convex Analysis of the Mean Field Langevin Dynamics. AISTATS2022]

$$\min_{\nu:\mathcal{P}(\Theta)} \frac{1}{n}\sum_{i=1}^{n} \ell\Big(\mathbb{E}_{\theta\sim\nu}[h_\theta(x_i)], y_i\Big) + \lambda\mathbb{E}_\nu[\|\theta\|^2]$$

Prob meas.

L2-正則化

ℓ:滑らかな損失関数,
h_θ: パラメータθのニューロン

$$\text{i.e., } h_\theta(x) = r\sigma(w^\top x) \text{ for } \theta = (r, w)$$

負エントロピー正則化を加える：分布は密度を持ち，最適化しやすくなる.

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

$$\lambda_2 \mathrm{KL}(\nu, N(0, \lambda_2/\lambda_1 I))$$

KL-div from a Gaussian distribution.

密度関数に関して凸関数
→ 通常の凸最適化手法を援用できる.

難しさ

- 基本的に無限次元最適化.
- 期待値の陽な記述は得られない.
- McKean-Vlasov過程を直接扱うと難しい.

→ **解決策:**

- 粒子を用いたモンテカルロ近似.
- 勾配ランジュバン動力学等でサンプリング.
- 各ステップごと内部イテレーションを回して分布を収束させる.

(1) 二重ループ法

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

[Nitanda, Wu, Suzuki: NeurIPS2021]

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

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; $\bar{g}^{(t)}$は基本的に勾配）

$\bar{g}^{(t)}$の決定に双対平均化法のルールを用いる

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

↳ 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** $\mathcal{L}(\hat{q}^{(t)}) - \mathcal{L}(q^*) \le O(1/t)$
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.**

➢ **初の多項式オーダー最適化手法**

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：

$$t_{\text{end}} = 2\left(n + \frac{1}{\lambda_2\gamma}\right)\log\left(\frac{nC}{\epsilon_P}\right)$$

➢ 指数オーダーでの収束を達成

➢ サンプルサイズnへの依存を緩和

(2) 一重ループ法

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

(時間離散化)

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

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)

定理 (離散化誤差バウンド)

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

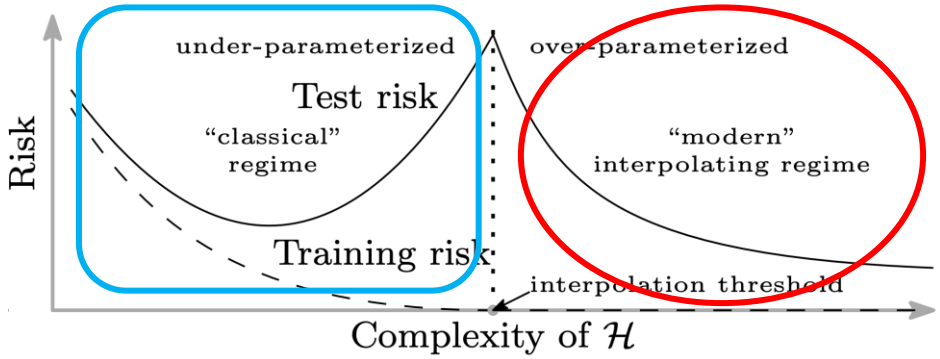
- 一様対数Sobolev不等式，超対数Sobolev不等式

過剰パラメータNNの予測誤差

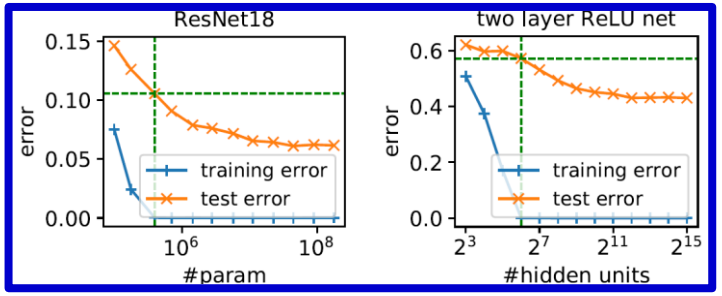
Amari, Ba, Grosse, Li, Nitanda, Suzuki, Wu, Xu: When Does Preconditioning Help or Hurt Generalization? ICLR2021.

古典的な状況
（バイアス-バリアンストレードオフ）

本研究
新しい設定：二重降下



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]

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

（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？

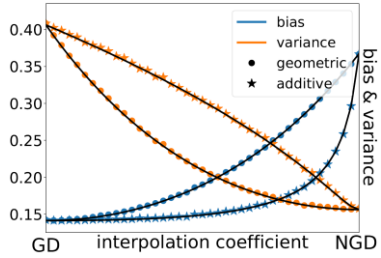
真の信号に依存してどの手法が良いかを解明

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. バリアンス

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.]

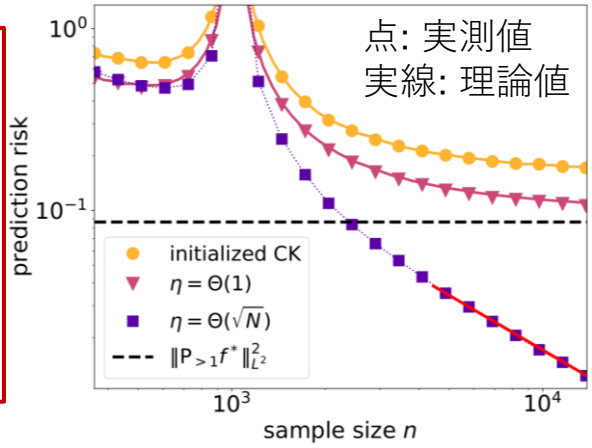
$$f_{\mathrm{NN}}(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle x, w_i \rangle) = \frac{1}{\sqrt{N}} a^\top \sigma(W^\top x)$$

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：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．

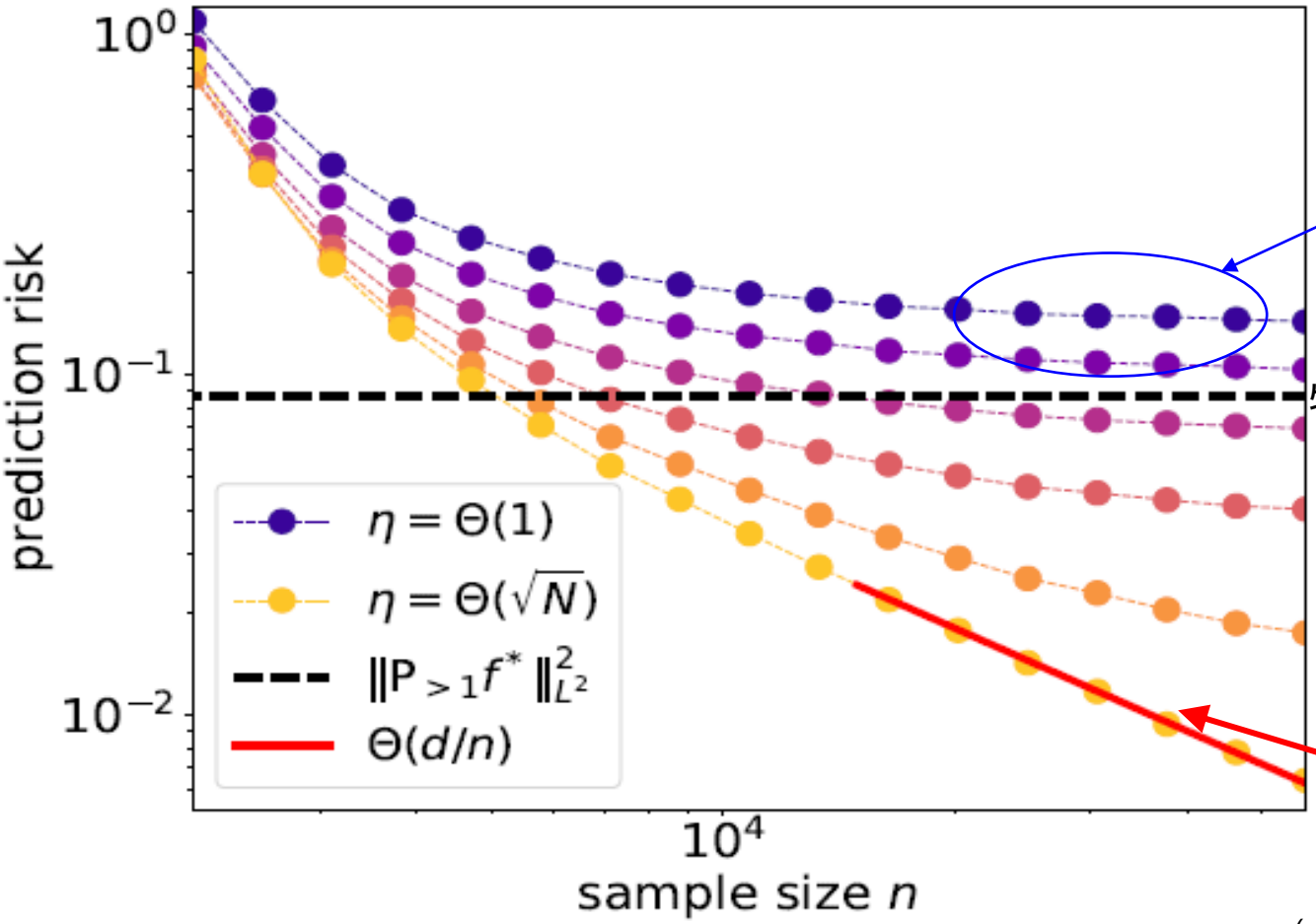
- $\eta = \sqrt{N}$: 大きなステップサイズを用いると，ランダム特徴モデルによるリッジ回帰を優越する．
- $\eta = 1$: 中間的なステップサイズでは横幅無限大のランダム特徴リッジ回帰を優越しないが初期値Wは優越．
- $\eta = o(1)$: 小さなステップサイズでは初期値Wと同じ予測誤差 (NTK-regime)．特徴学習の効果なし．



実験との整合性

参考資料

(点線：理論値，丸印：実験)



小さなステップサイズ

特徴学習しない場合の限界
（ランダム特徴量）

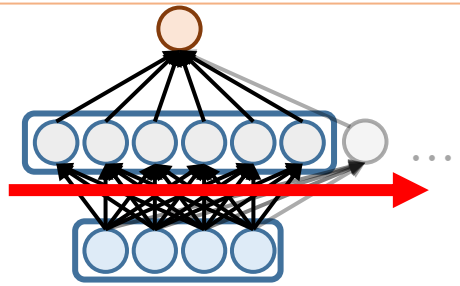
大きなステップサイズ：
$\Theta(d/n)$で減少

(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

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

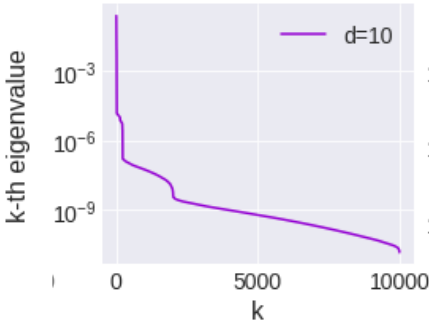
NTK (Neural Tangent Kernel) の固有値の減衰レート

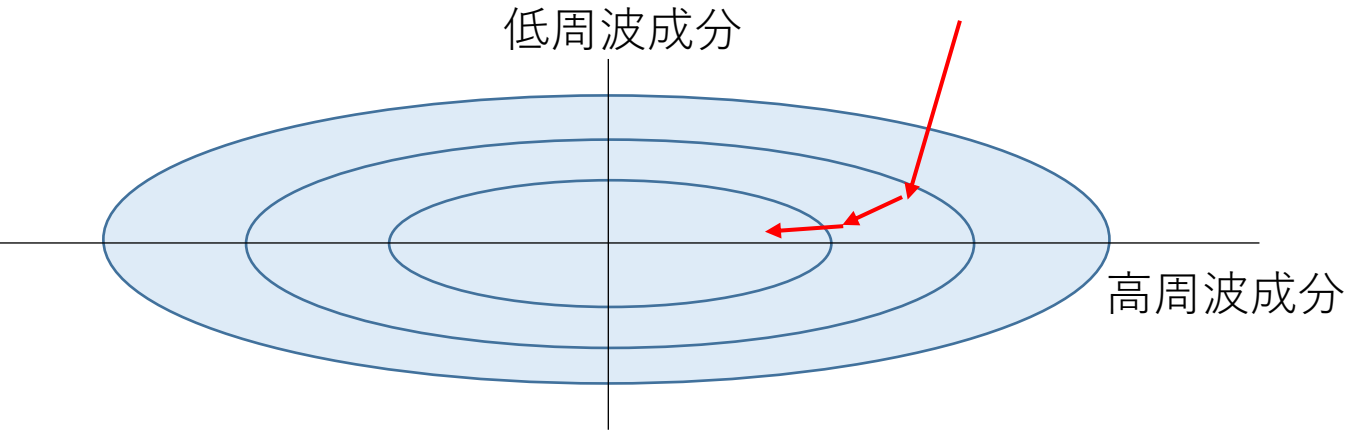
$$\mathbb{E}[\|f_T - f^*\|_{L_2}^2] \leq \epsilon_M + O(T^{-\frac{2r\beta}{2r\beta+1}})$$

横幅$M \to \infty$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





**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.**

ノイズあり勾配法の最適性

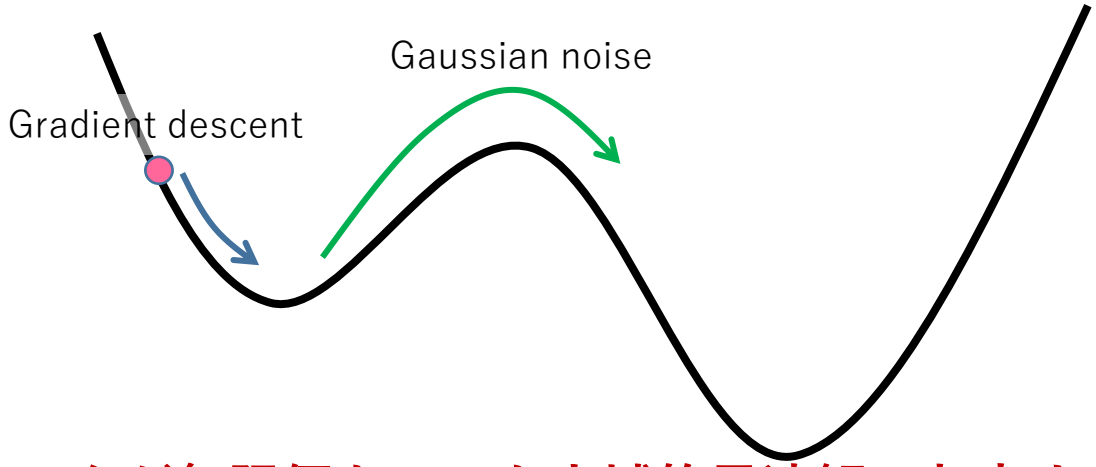
深層学習が浅い学習方法に優れていることを証明

- 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).

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

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



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

深層

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

(γ: 大) → $1/n$

線形（カーネル法）

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

(d: 大) → $1/\sqrt{n}$

- 深層学習の予測誤差は入力の次元に依存しない.
- 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

分散縮小型確率的勾配Langevin動力学

[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

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

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

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

- 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.	$\tilde{O}\left(\frac{nd}{\epsilon^2}\cdot \mathrm{poly}(M,L)\right)$
Xu et al. (2018)	Smooth, Dissipative	Weak conv.	$\tilde{O}\left(\frac{nd}{\epsilon}\right)\cdot \mathrm{e}^{\tilde{O}(d)}$
Vempala et al. (2019)	Smooth, Log-Sobolev (α)	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.	$\tilde{O}\left(n+\frac{n^{3/4}}{\epsilon^2}+\frac{n^{1/2}}{\epsilon^4}\right)\cdot \mathrm{e}^{\tilde{O}(\gamma+d)}$
Zou et al. (2021)	Smooth, Dissipative, Warm-start	TV	$\tilde{O}\left(\frac{\gamma^2}{\epsilon^2}\right)\cdot \mathrm{e}^{\tilde{O}(d)}$
Zou et al. (2019b)	Smooth, Dissipative	2-Wass.	$\tilde{O}\left((n+\frac{n^{1/2}}{\epsilon^2 \mu_*^{3/2}})/\right.$
SVRG-LD/SARAH-LD ($B = m = \sqrt{n}$)	Smooth, Log-Sobolev (α)	KL	$\tilde{O}\left(\left(n+\frac{dn^{1/2}}{\epsilon}\right)\cdot \gamma^2 L^2 \alpha^{-2}\right)$

条件緩和

$\sqrt{n}$倍速い

勾配ランジュバン動力学の研究

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 $\tilde{\nabla}_k = \frac{1}{B}\sum_{i\in I_k} f_i(X_k)$

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

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

：$\sqrt{n}$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

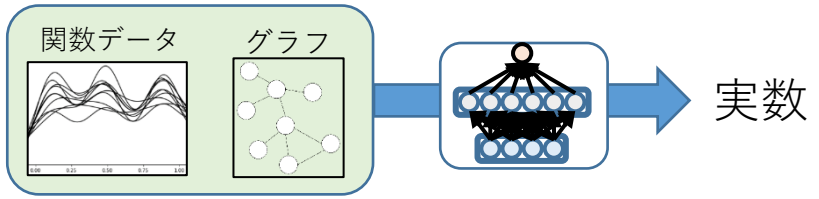
$$f(x) = \frac{1}{M}\sum_{j=1}^{M} r_j \sigma(w_j^\top x)$$

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）

深層ニューラルネットの冗長性

- 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

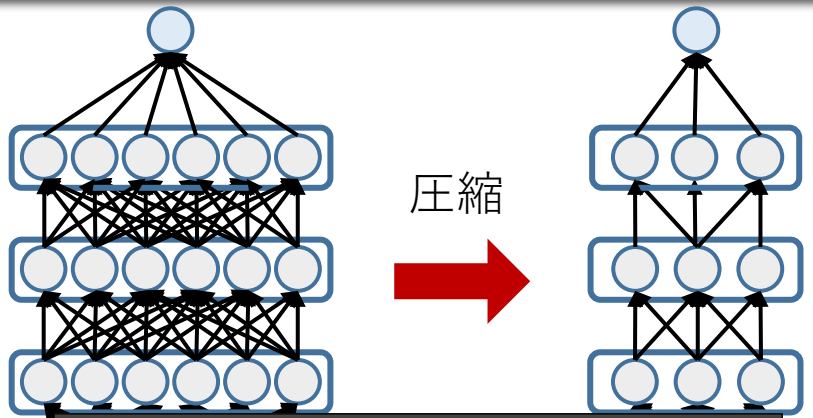
[仮説] 見かけの大きさ (パラメータ数) よりも
実質的な大きさ (自由度) はかなり小さいはず.

“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
- …

圧縮型バウンド

「圧縮」できるネットワークは過学習しない



カーネル法の理論

- 中間層の**分散共分散行列の固有値分布**で圧縮率を評価.
- 「**テンソル分解**」の援用によりCNNの詳細な評価も実現.

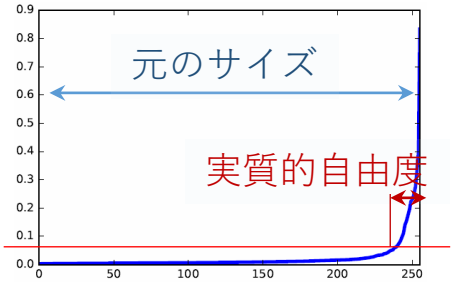
$$\Psi(\widehat{f}) \leq \hat{\Psi}(\widehat{f}) + O\left(\left(L\frac{\sum_{\ell=1}^{L} m_\ell}{n}\log(n)\right)^{\frac{2\alpha}{1+2\alpha}} + \sqrt{L^{1+\delta}\frac{\left(\sum_{\ell=1}^{L} m_\ell\right)^{\frac{4/\beta}{4/\beta+2(1-1/2\alpha)}}}{n}\log(n)^3}\right)$$

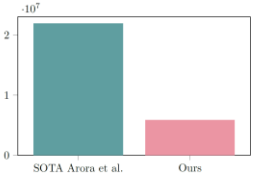
[Suzuki, Abe, Nishimura: Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network, ICLR2020, spotlight]

[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 et al.: Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. IJCAI-PRICAI 2020]

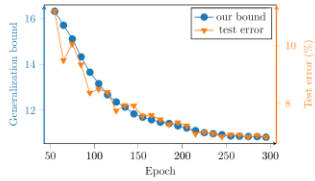
[実験的観察] 実際に学習したネットワークは圧縮しやすい.

	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





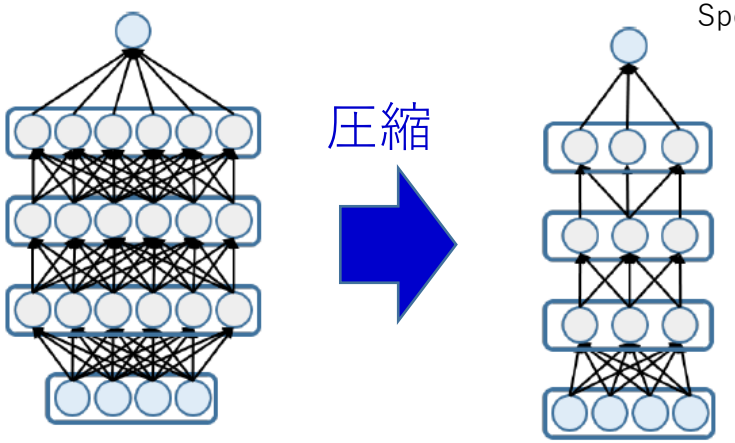
(a) Bound comparison



(b) Generalization bound

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]

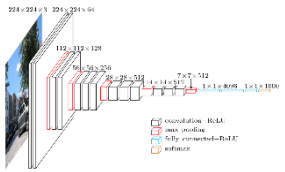


- メモリ消費量を減少
- 予測にかかる計算量を減少

→ 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
Spec-Conv	**70.418%**	**90.094%**	131.44M	9.58B
Spec-GAP	**67.540%**	**88.270%**	8.32M	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）

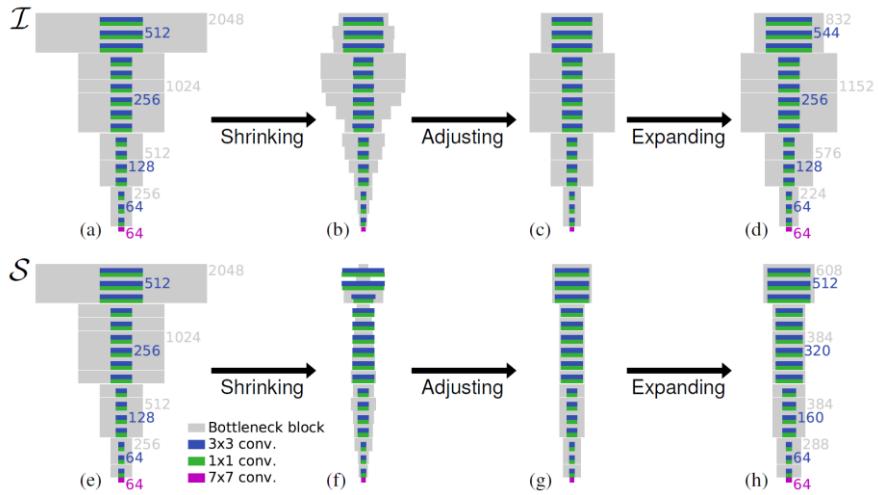
ResNet-50ネットワークの圧縮

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
Spec-ResA	72.99%	91.56%	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
Spec-ResB wo/ ft	66.12%	86.67%	20.69M	5.25G
Spec-ResB w/ ft	74.04%	91.77%	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 $\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta)$.
3: Calculate eigenspectra $S_{1:M}$.
4: Calculate intrinsic dimensionalities $d_{1:M}$ by $d_{1:M} = \text{len}(S_{1:M} > T)$.
5: Determine new widths $O'_{1:M}$ by adjusting $d_{1:M}$.
6: Find the largest ω such that $c(\omega \cdot O'_{1:M}) \leq \zeta$.
7: **return** $\omega \cdot O'_{1:M}$.

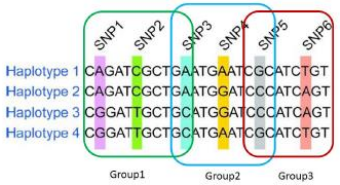
Backbone	Normalization	Classification		COCO (2× schedule)					
		MACs	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50 [35]	SyncBN	3.8 G	—	34.5	55.2	37.7	20.4	36.7	44.5
ResNet-50*	GN	4.09 G	25.5 M	35.5	55.6	38.5	21.3	37.5	45.3
ResiaxNet $\mathcal{S}$3-50 (MACs)	GN	4.06 G	18.6 M	35.4	55.4	38.6	21.5	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	**46.0**
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	**35.8**	**55.9**	**38.9**	**21.8**	**38.0**	45.6
DetNet-59 [35]	SyncBN	4.8+ G	—	**36.3**	**56.5**	**39.3**	22.0	**38.4**	**46.9**
DetNet-59†	GN	5.00+ G	18.3+ M	36.2	56.0	**39.3**	22.1	38.3	46.0
DetiaxNet $\mathcal{I}$2-59 (MACs)	GN	4.94+ G	17.4+ M	36.2	56.0	**39.3**	**22.5**	38.1	46.0


カーネル法
ガウス過程
学習理論
スパース推定
深層学習

確率的最適化 [計算理論]

(構造的)正則化学習

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



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

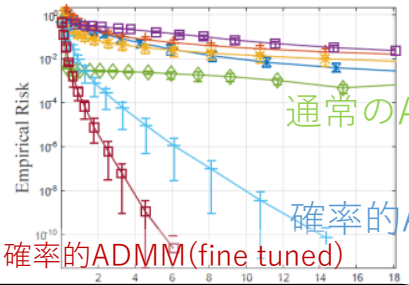
従来法 **勾配法**：定番の最適化手法

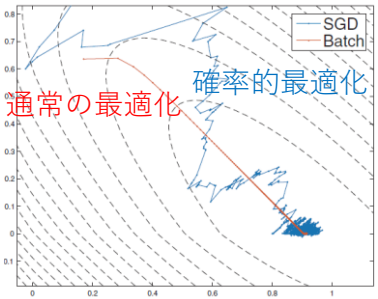
欠点：各更新ごとに全データ点を読むので大きいデータでは非効率.

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 ε を達成するまでの計算量

n:サンプルサイズ, κ:条件数

通常の勾配法 $n\sqrt{\kappa}\log(1/\epsilon)$ ➡ 確率的最適化法 $(n+\sqrt{n\kappa})\log(1/\epsilon)$

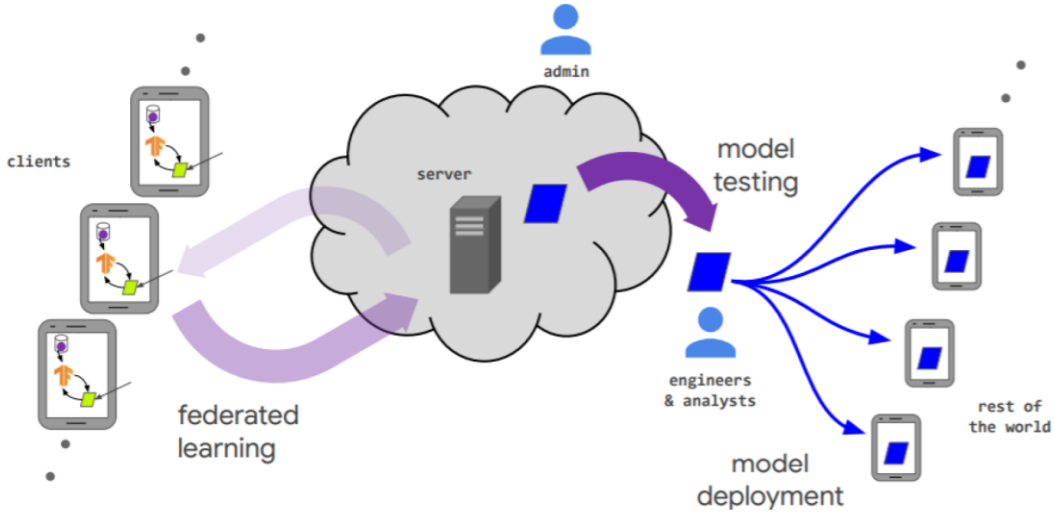
$n = 10^6$なら最大1000倍の高速化

-オンライン型確率的交互方向乗数法(ICML2013, 2012 年度IBISML 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IBISML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IBISML研究会賞**)

Machine Learning Summer School 2015
SIAM-Optimization

書籍：確率的最適化，機械学習のための連続最適化

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, **ζ_i: heterogeneity of workers**
- Our proposal: Bias-Variance Reduced Local SGD method

	Communication Complexity	Communication Complexity ($B\to\infty$)	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!

研究課題

- 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)

まとめ

理論を土台にした普遍性の高い基礎研究

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

- **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)