FY2022/2022年度

Deep Learning Theory Team

Taiji Suzuki

深層学習理論チーム　鈴木 大慈

RIKEN Center for Advanced Intelligence Project

平均場解析による深層NNの最適化

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

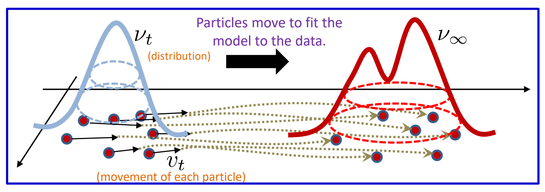
- 無限次元入力ニューラルネットの無限次元GLDによる最適化 (NeurIPS2022)
- 平均場ニューラルネットワークのGLDによる最適化の収束 (AISTATS2022)
- 平均場ニューラルネットワークの新しい最適化手法: 粒子確率的双対座標上昇法 (ICLR2022)
- 有限粒子近似の近似誤差解析 (ICLR2023)

[Atsushi Nitanda, Denny Wu, Taiji Suzuki: Convex Analysis of the Mean Field Langevin Dynamics. AISTATS2022]

2層ニューラルネットワーク:

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

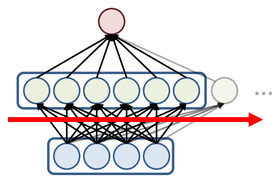
パラメータ $(r_j, w_j)_{j=1}^M$ に関して非線形：**非凸最適化**



過剰パラメータ化（平均場）:

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

パラメータの「**分布**」μ に関する平均：分布に関して線形



目的関数 (正則化学習)

$$F(\mu) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(f_\mu) + \lambda_1 \mathbb{E}_\mu[\|x\|^2] \quad \xrightarrow{\mu\text{で微分}} \quad \frac{\delta F(\mu)}{\delta \mu}(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i'(f_\mu) h_x(z_i) + \lambda_1\|x\|^2$$

➢ **平均場ランジュバン動力学 (ノイズあり勾配降下法):**

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

実は平均場ランジュバン動力学は以下を最適化するWasserstein勾配流：

$$\mu^* = \arg\min_{\mu\in\mathcal{P}} \{F(\mu) + \lambda_2 \mathrm{Ent}(\mu)\} =: \mathcal{L}(\mu)$$

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

ℓ_iが凸であるなら，対数ソボレフ不等式定数αを用いて，

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \le \exp(-2\alpha\lambda_2 t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*)),$$

$$\lambda_2 \mathrm{KL}(\mu||\mu^*) \le \mathcal{L}(\mu) - \mathcal{L}(\mu^*) \le \lambda_2 \mathrm{KL}(\mu||p_\mu).$$

平均場ランジュバン動力学（ノイズ有り勾配降下法）で二層ニューラルネットワークは最適解へ**指数関数的速度で収束**

無限次元入力NNの無限次元GLDによる最適化

[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]

有限次元入力

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

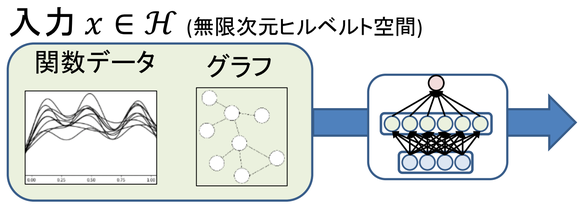
無限次元入力

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

拡張

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

入力 $x \in \mathcal{H}$ (無限次元ヒルベルト空間)



出力（実数）

- 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（Muzellec et al. COLT2022）

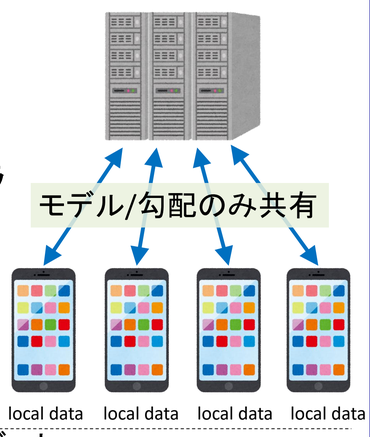
連合学習の最適化理論

連合学習：データを分散させた並列分散学習（1）へ拡張
（1）確率的分散縮小勾配降下法を用いて高速化
（2）既存研究で要求されていた二重ループを一重ループへ

[Murata, Suzuki: Escaping Saddle Points with Bias-Variance Reduced Local Perturbed SGD for Communication Efficient Nonconvex Distributed Learning. NeurIPS2022]
[Oko, Akiyama, Murata, Suzuki: Versatile Single-Loop Method for Gradient Estimator: First and Second Order Optimality, and its Application to Federated Learning: OPT2022, NeurIPS2022 Workshop]

我々の貢献：

- 連合学習の新しい効率的な確率的最適化法の提案
  - ➢ クライアント間分散の縮小
  - ➢ クライアントが似ていれば（Heterogeneous）速い収束
- 一重ループの分散縮小法
  - ➢ 全クライアントを使う必要がなく，ランダムに選んだクライアントを動かせば十分
  - ➢ 通信量を削減できる



収束保証：二次最適性条件も保証

これまでの手法の中で最小の通信回数・計算量

（単一クライアントの最適化手法としても一重ループかつ最適計算量）

- クライアント間勾配同期

I_t:ランダムに選んだクライアント

$$\nabla f_i(x) \approx g_i^t = \begin{cases} \nabla f_i(x^t) & (\text{if } i \in I^t) \\ \frac{1}{p}\sum_{i\in I^t}(\nabla f_i(x^t) - \nabla f(x^{t-1})) + g_i^{t-1} & (\text{otherwise}) \end{cases}$$

提案法「**SLEDGE**」型更新
→クライアント間の分散を縮小

- クライアント内アップデート

Send params $x^{t,0} \leftarrow x^{t-1}$, $g^{t,0} \leftarrow g^{t-1}$ to the client i_t
for $k = 1$ to K **do**
Estimator of $\nabla f(x^{t-1})$

$$x^{t,k} \leftarrow x^{t,k-1} - \eta g^{t,k-1} + \xi^{t,k} \quad (\xi^{t,k} \sim B(0,r))$$

Choose local minibatch $J_{i_t}^{t,k}$ of the client i_t with size b

$$g^{t,k} \leftarrow \frac{1}{b'}\sum_{j\in J_{i_t}^{t,k}} (\nabla f_{i,j}(x^{t,k}) - \nabla f_{i,j}(x^{t,k-1})) + g^{t,k-1}$$

SARAH-type recursive update

$$x^t \leftarrow x^{t,K}$$

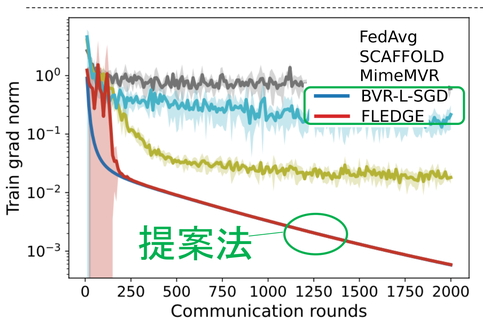


Table 2: Comparison of communication rounds and complexity for a non-convex federated learning problem (2).

Algorithms	Communication rounds	Client sampling (other then x^0)
FedAvg (nonconvex) (Karimireddy et al. 2020)	$\frac{\sigma_c^2}{pe^4} + \frac{\sigma_g}{e^3} + \frac{1}{e^2}$	✓
SCAFFOLD (nonconvex) (Karimireddy et al. 2020)	$\frac{1}{e^2}(\frac{P}{p})^{\frac{2}{3}}$	✓
MimeMVR (nonconvex) (Karimireddy et al. 2021)	$\frac{\zeta'\sigma_c}{\sqrt{p}e^3} + \frac{\sigma_g^2}{pe^2} + \frac{1}{Ke^2} + \frac{\zeta'}{e^2}$	✓
BVR-L-SGD (nonconvex) (Murata and Suzuki 2021)	$\frac{1}{Ke^2} + \frac{\zeta}{e^2}$	×
FLEDGE (nonconvex) (ours)	$\frac{1}{Ke^2} + \frac{\zeta\sqrt{P}}{pe^2} + \frac{\zeta}{e^2}$	✓
BVR-L-PSGD (SOSP) (Murata and Suzuki 2022)	$(\frac{1}{K} + \zeta)(\frac{1}{e^2} + \frac{1}{\delta^2})$	×
FLEDGE (SOSP) (ours)	$(\frac{1}{K} + \zeta)(\frac{1}{e^2} + \frac{1}{\delta^2})$	✓ (requiring $p \gtrsim \sqrt{P} + \frac{\zeta'}{\delta^2}$)
MimeSGD (PL) (Karimireddy et al. 2021)	$\frac{\sigma_c^2}{\mu pe^2} + \frac{L}{\mu}$	✓
FLEDGE (PL) (ours)	$\frac{L}{\mu K} + \frac{\zeta\sqrt{P}}{\mu p} + \frac{\zeta}{\mu} + \frac{P}{p}$	✓

Note: P is the number of clients, μ is the parameter for PL condition, σ_c is the variance between clients, which can be as large as P. ζ is the Hessian-heterogeneity between clients and ζ' in MimeMVR is the Hessian-heterogeneity *between all data* (i.e., $\|\nabla^2 f_{i,j}(x) - \nabla^2 f(x)\| \le \zeta'$). ζ' contains not only the inter-client Hessian-heterogeneity but also the intra-client Hessian-heterogeneity. Thus, $\zeta \le \zeta'$ always holds and moreover it is possible that $\zeta \ll \zeta'$. Importantly, $\zeta \to 0$ does not necessarily means $\zeta' \to 0$.

ランジュバン動力学の理論

勾配ランジュバン動力学の一般論

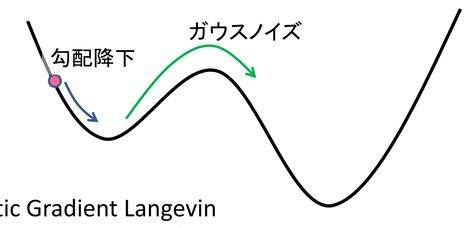
- 分散縮小型確率的勾配ランジュバン動力学の収束解析 (NeurIPS2022)
- 無限次元勾配ランジュバン動力学のアルゴリズムと収束理論 (COLT2022)

連続時間勾配ランジュバン動力学

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

離散時間勾配ランジュバン動力学

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



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

$\nabla F(x)$の計算にO(n)かかる (大規模データで困る):

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

分散縮小型確率的勾配:

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

定常分布

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

- ➢ 定常分布からのサンプリング
- ➢ βが大きければ，定常分布は最適解周りに集中する.

- ➢ 勾配計算量
  - **- Vempala&Wibisono (2019): 非確率的勾配**

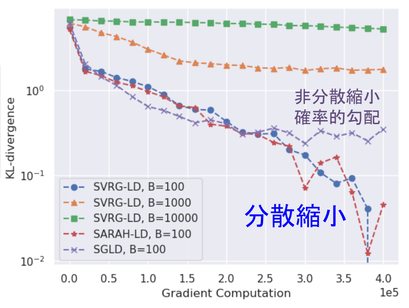
$$\tilde{O}\left(\frac{n}{\epsilon} d\gamma^2 L^2 \alpha^{-2}\right)$$

α: 対数ソボレフ定数
例:
- 二次関数+有界関数
- Weak Morse型関数

  - **- 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



分散縮小型確率的勾配を用いることで$\sqrt{n}$倍少ない計算量でOK.

高次元特徴学習のダイナミクス

- 勾配降下による特徴学習の高次元漸近解析 (NeurIPS2022)
- ノイズあり勾配降下法による特徴学習と汎化誤差解析 (ICLR2023)

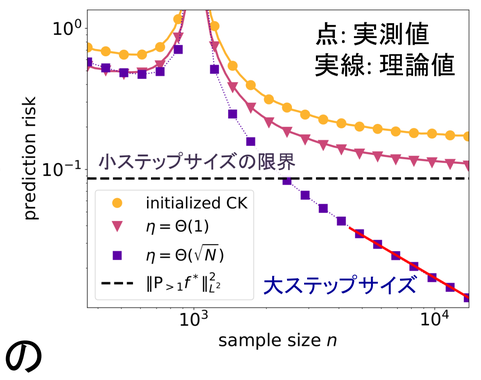
[Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, Greg Yang: High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation. NeurIPS2022]

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

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.

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



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

- ➢ $\eta = \sqrt{N}$: 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.（非線形成分も学習）
- ➢ $\eta = 1$: 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.（線形成分しか学習できない）
- ➢ $\eta = o(1)$: 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）

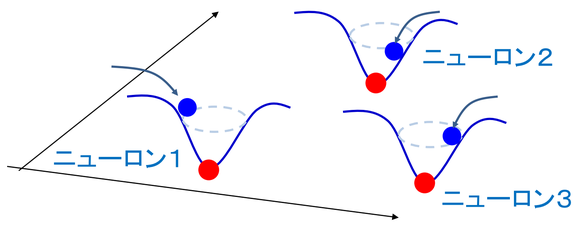
[Akiyama and Suzuki: Excess Risk of Two-Layer ReLU Neural Networks in Teacher-Student Settings and its Superiority to Kernel Methods. ICLR2023]

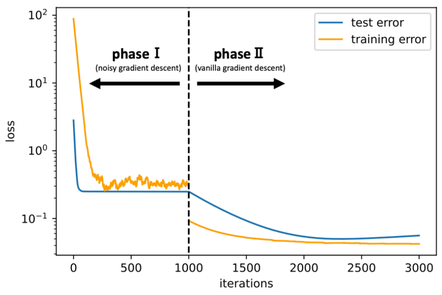
$$\widehat{\mathcal{R}}_\lambda(\Theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{M} a_j\sigma(\langle w_j, x\rangle)\right)^2 + \lambda\sum_{j=1}^{M}(a_j^2 + \|w_j\|^2) \qquad \Theta^{(k+1)} = \Theta^{(k)} - \eta^{(1)}\nabla\widehat{\mathcal{R}}_\lambda(\Theta^{(k)}) + \sqrt{\frac{2\eta^{(1)}}{\beta}}\zeta^{(k)}$$

二層NNの最適化は二段階のフェーズで収束　※全体で**多項式時間でOK**

探索フェーズ:

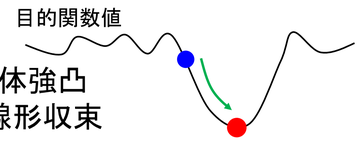
最適解近傍へ到達





収束フェーズ:

目的関数値



最適解まわりでは大体強凸
→ 普通の勾配法で線形収束

ニューラルネットワークの予測誤差

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

（次元の呪いを受けない）

カーネル法の予測誤差

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

（次元の呪いを受ける）