確率的勾配降下法の高次元極限：動的平均場理論による精密解析

西山 颯大 [A, B]　今泉 允聡 [A, B, C]

[A]東京大学　[B]理研AIP　[C]京都大学

背景

◯ 研究対象：確率的勾配降下法 (Stochastic Gradient Descent; SGD)

$\boldsymbol{\theta} \in \mathbb{R}^d$: パラメータ, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$: データ（各行 $\boldsymbol{x}_i \in \mathbb{R}^d$ がサンプル）

n 個のデータ点に対するロスの総和 $L(\boldsymbol{X}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n L(\boldsymbol{x}_i, \boldsymbol{\theta})$ の最小化を考える

SGD: 現代の機械学習の標準的な最適化手法
ランダムに選んだミニバッチ $\mathcal{B}^k \subset \{1, \dots, n\}$ に対して勾配降下（学習率 η, バッチサイズ B）

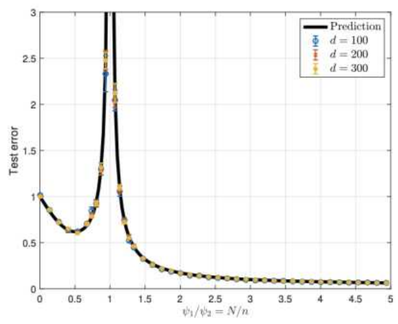
$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k - \eta \nabla_{\boldsymbol{\theta}} L_{\mathcal{B}^k}(\boldsymbol{X}, \boldsymbol{\theta}^k)\,, \quad L_{\mathcal{B}}(\boldsymbol{X}, \boldsymbol{\theta}) := \frac{1}{B}\sum_{i\in\mathcal{B}} L(\boldsymbol{x}_i, \boldsymbol{\theta})\,. \tag{1}$$

勾配ノイズのない勾配降下法とは異なる振る舞い（e.g. より高性能な解に収束）を示す
→ SGD**特有の確率的振る舞いの理解が重要**

◯ 研究方法：高次元漸近論 (High-dimensional asymptotics)

アプローチ: モデルの高次元極限での振る舞いを**精密に**記述する低次元の方程式を導出することにより，
モデルの**典型的・普遍的**な振る舞いを明らかにする

手法: 統計力学，ランダム行列理論など
成果の例: 圧縮センシングの性能限界 [1],
二重効果のメカニズム解明 [2]



2層NNの幅 vs 汎化誤差の精密解析 [2]

◯ 先行研究：SGD の高次元漸近論

1. **オンラインSGDの解析** [3]
   Pros. 2層NNを含む広いクラスのモデルに対して，**オンライン**SGD（パラメータの更新ごとに新しいデータを用いる）の振る舞いを低次元かつ連続時間の方程式で記述
   → 性能の長時間極限の解析や，定性的に異なる振る舞いをする相の分類などが可能
   Cons. データを使い回す設定を扱えず，**過学習**の効果を捉えられない
2. **大きなバッチサイズのSGDの解析** [4]
   Pros. 2層NNを含む広い設定で，大きなバッチサイズ ($B \propto n$) のSGDの高次元極限を低次元の確率過程で記述
   Cons. ▪ 実用では $B \ll n$ が一般的で，設定が非現実的
   ▪ その確率過程は離散時間で，理論的に扱いづらい（定常分布や長時間挙動の解析が難しい）
3. **線形モデルの解析** [5]
   Pros. **線形モデル**のSGD学習の高次元極限を低次元かつ連続時間の方程式で記述
   → さらなる理論的解析が可能で，性能の長時間極限や収束スピードの相転移を解明
   Cons. 線形モデルに設定が限られており，非線形NNの複雑な挙動を捉えられない．

目的

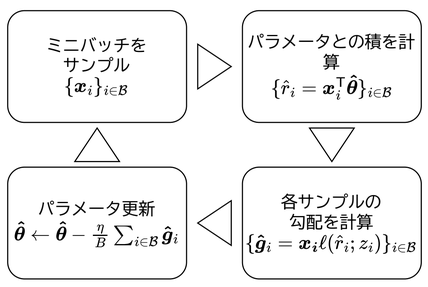
- 2層NNを含む**広いクラスの非線形モデル**で
  （2層NNは，現実のNNの複雑な振る舞いを捉えることのできるモデル）
- **データを使い回す設定**で
  （データを使い回すことは実用的に行われており，過学習の振る舞いを理解するうえで重要）
- SGD の高次元極限を記述する**低次元かつ連続時間**の方程式を導出する
  （連続時間の方程式のほうが理論的に扱いやすく，長時間挙動の解析が可能）

設定

◯ 2層NNのSGD学習

$\hat{\boldsymbol{\theta}} \in \mathbb{R}^{d\times m}$: パラメータ, $z \in \mathbb{R}^n$: ラベルノイズ,
$\ell\colon \mathbb{R}^{m+1} \to \mathbb{R}^m$: ロスの微分

$$\hat{\boldsymbol{\theta}}^{t_{k+1}} = \hat{\boldsymbol{\theta}}^{t_k} - \frac{\eta}{B}\sum_{i\in\mathcal{B}^k} \boldsymbol{x}_i \ell_{t_k}(\hat{r}_i^{t_k}; z_i)^\top\,, \quad \hat{\boldsymbol{r}}^{t_k} = \boldsymbol{X}\hat{\boldsymbol{\theta}}^{t_k}\,. \tag{2}$$



幅 m の2層NNや一般化線形モデル ($m = 1$) を含む

◯ SGDの連続時間近似：確率的勾配流 (Stochastic Gradient Flow; SGF) [6]

SGDを連続時間の**確率微分方程式** (SDE) で近似 → 確率解析の手法を利用可能に

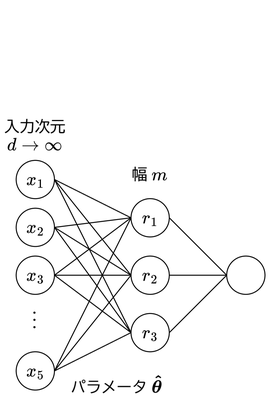
SGDと平均と分散が一致するSDE（確率的勾配流，SGF）を構成

$$\mathrm{d}\boldsymbol{\theta}^t = -\frac{1}{\delta}\boldsymbol{X}^\top\ell_t(\boldsymbol{r}^t; \boldsymbol{z})\,\mathrm{d}t + \sqrt{\frac{\tau}{\delta}}\sum_{i=1}^n \boldsymbol{x}_i\ell_t(r_i^t; z_i)^\top\,\mathrm{d}B_i^t\,, \quad \boldsymbol{r}^t = \boldsymbol{X}\boldsymbol{\theta}^t\,. \tag{3}$$

$\boldsymbol{B}^t \in \mathbb{R}^n$ は Brown 運動，$\delta = n/d$ はデータ・パラメータ比
$\tau = \eta/B$ は**温度パラメータ**（ノイズの大きさを制御）

◯ 仮定

- $\boldsymbol{X} = (x_{ij})_{i\in[n],j\in[d]}$ の成分は独立，平均 0, 分散 $1/d$ の sub-Gaussian 確率変数
- 比例的高次元極限 $n, d \to \infty$, $n/d \to \delta \in (0, \infty)$
- ℓ のLipschitz連続性，$\boldsymbol{\theta}^0, \boldsymbol{z}$ のモーメント条件

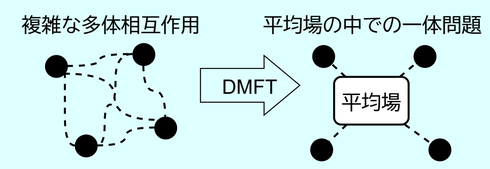


手法: 動的平均場理論 (Dynamical Mean-Field Theory; DMFT)

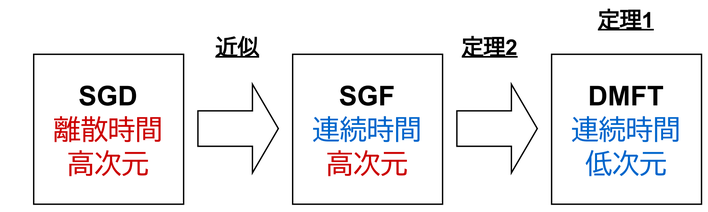
SGFの高次元極限を**動的平均場理論** (DMFT) により低次元の方程式で記述

◯ 動的平均場理論 (DMFT) による解析

- 統計物理学の解析手法
- 高次元のランダムなダイナミクスの典型的な振る舞いを，**低次元**の有効過程で記述
- **アイデア**: パラメータ間の相互作用を，ある1つのパラメータと，他のパラメータが作る平均的な環境（**平均場**）との相互作用に置き換え
  → 高次元の問題を1次元の問題に帰着



結果: DMFT方程式によるSGFの高次元極限の特徴づけ



◯ DMFT方程式: 連続時間，低次元 の確率過程の連立方程式

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\theta^t &= u^t - \Gamma(t)\theta^t - \int_0^t R_\ell(t,s)\theta^s\,\mathrm{d}s\,, \quad u^t \sim \mathrm{GP}(0, C_\ell/\delta)\,,\\ r^t &= w^t - \frac{1}{\delta}\int_0^t R_\theta(t,s)\ell_s(r^s; z)(\mathrm{d}s + \sqrt{\tau\delta}\,\mathrm{d}B^s)\,, \quad w \sim \mathrm{GP}(0, C_\theta)\,,\end{aligned} \tag{4}$$

$$C_\theta(t,t') = \mathbb{E}[\theta^t\theta^{t'\mathsf{T}}]\,, \quad R_\theta(t,t') = \mathbb{E}\left[\frac{\partial\theta^t}{\partial u^{t'}}\right]\,, \quad R_\ell(t,t') = \mathbb{E}\left[\frac{\partial\ell_t(r^t; z)}{\partial w^{t'}}\right]\,, \quad \Gamma(t) = \mathbb{E}[\nabla_r\ell_t(r^t; z)]\,,$$
$$C_\ell(t,t') = \mathbb{E}\left[\ell_t(r^t; z)(1 + \sqrt{\tau\delta}\dot{B}^t)\ell_{t'}(r^{t'}; z)^\mathsf{T}(1 + \sqrt{\tau\delta}\dot{B}^{t'})\right]\,.$$

$B^t \in \mathbb{R}$ は Brown 運動，$\dot{B}^t$ はその微分（ホワイトノイズ），$\partial\theta^t/\partial u^{t'}, \partial\ell_t(r^t; z)/\partial w^{t'}$ は汎関数微分

定理 1: DMFT方程式の解の存在と一意性

ある $T > 0$ に対して，時間 $[0, T]$ で DMFT 方程式 (4) の解が一意的に存在する．

定理 2: DMFT方程式によるSGFの高次元極限の特徴づけ

定理 1 の $T > 0$ と任意の $t \in [0, T]$ に対して，SGF (3) の解 $\boldsymbol{\theta}^t$ の成分の経験分布と，DMFT方程式 (4) の解 θ^t の分布の間の 2-Wasserstein 距離が $n, d \to \infty$ で 0 に確率収束する．

$$\operatorname*{p\text{-}lim}_{n,d\to\infty} W_2\left(\frac{1}{d}\sum_{i=1}^d \delta_{\theta_i^t}, \mathrm{Law}(\theta^t)\right) = 0\,. \tag{5}$$

すなわち，**SGF のパラメータ θ^t の各成分は，DMFT の確率過程 θ^t と同じように振る舞う．**

証明のスケッチ: 時間を離散化 → **近似メッセージ伝搬法** (Approximate Message Passing; AMP) [7] に帰着 → その高次元極限を **状態発展法** (State Evolution) で解析 → 時間を連続化

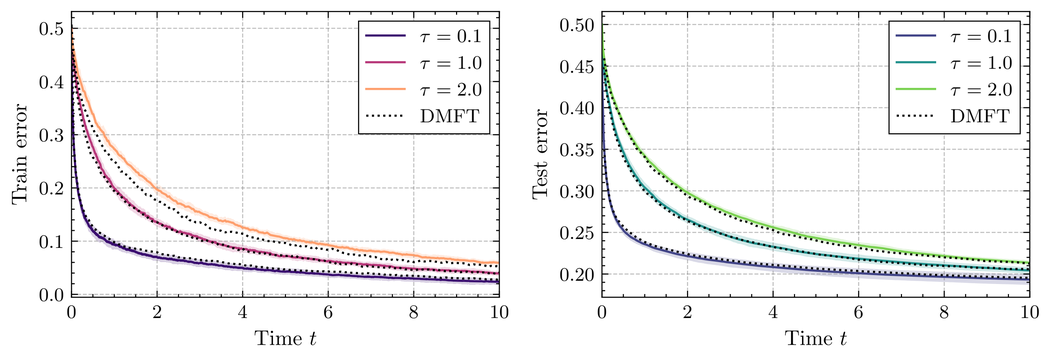


図 1. SGDのシミュレーション ($d = 1024$, $n = 2048$) とDMFT方程式の数値解の比較．高次元SGDの挙動（訓練誤差とテスト誤差）の振る舞いをDMFT方程式が正確に捉えている．

展望

- 具体的なモデルの詳細な解析に応用
- 定常分布解析への応用．DMFT 方程式の定常解を求めることで定常分布が求まる．
  → **SGD学習で見つかる最終的な解の性能の理解へ**

参考文献

[1] Kabashima, Wadayama, and Tanaka. In: *Journal of Statistical Mechanics: Theory and Experiment* (2009).
[2] Mei and Montanari. In: *Communications on Pure and Applied Mathematics* (2022).
[3] Ben Arous, Gheissari, and Jagannath. In: *Advances in Neural Information Processing Systems* (2022).
[4] Mignacco et al. In: *Advances in Neural Information Processing Systems* (2020).
[5] Paquette et al. In: *Mathematical Programming* (2024).
[6] Ali, Dobriban, and Tibshirani. In: *International Conference on Machine Learning*. PMLR. 2020.
[7] Bayati and Montanari. In: *IEEE Transactions on Information Theory* (2011).

snishiyama@g.ecc.u-tokyo.ac.jp