学習アルゴリズムの大域収束性と帰納的バイアス

二反田篤史
東京大学 / 理研AIP / JSTさきがけ

情報論的学習理論ワークショップ（IBIS2019）
企画セッション「深層学習の理論」
2019年11月22日
@ウインクあいち

発表概要

深層学習に対する（確率的）勾配降下法の理論

近年の重要な進展

- Over-parameterizationの役割
- **帰納的バイアス**

理論的困難

- 非凸最適化問題に対する**大域的収束性**
- 最適化で得られる関数の**汎化性能保証**

1. 深層学習における最適化の研究課題

機械学習における最適化問題

パラメータΘ, 関数$f_\Theta: \mathcal{X} \to \mathbb{R}$,

データとパラメータの適合度（損失関数）$l(y, f_\theta(x))$.

- 二乗損失 $l(y,z) = 0.5(y-z)^2$
- ロジスティック損失 $l(y,z) = \log(1+\exp(-yz))$

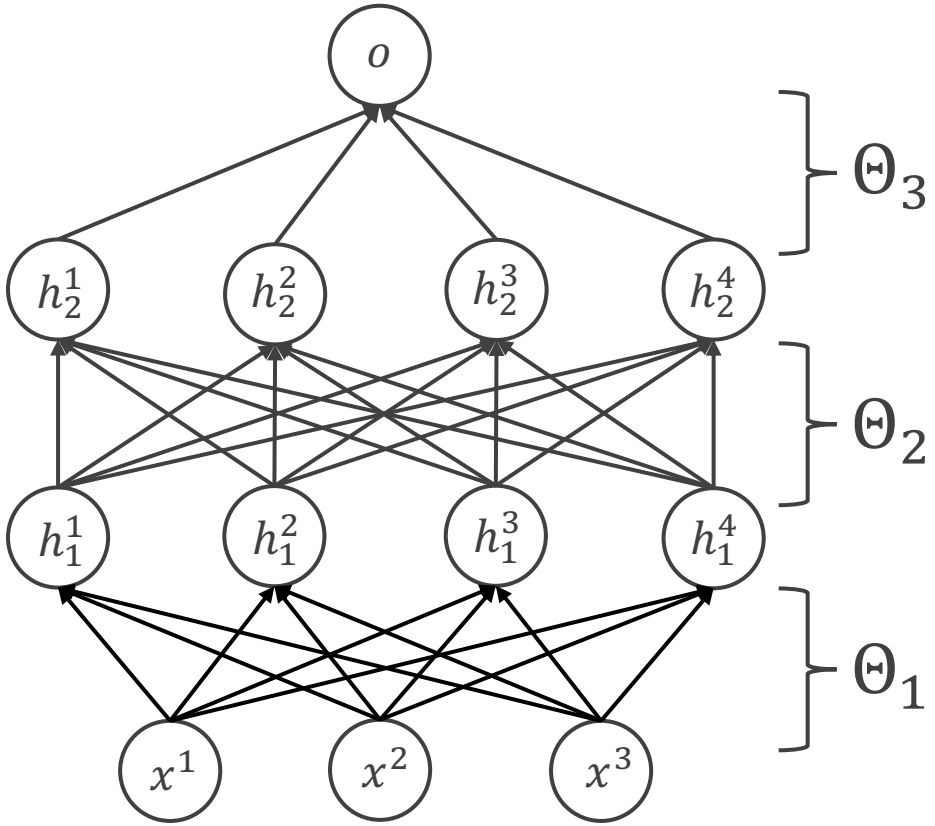
期待損失最小化 真に解きたい問題.

$$\min_{\Theta \in \mathcal{F}} \mathbb{E}[l(Y, f_\Theta(X)]$$

経験損失最小化 訓練データによる近似.

$$\min_{\Theta \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} l\left(y_i, f_\Theta(x_i)\right) + \lambda R(\Theta)$$

深層ニューラルネット



例：三層ニューラルネット

L-層ニューラルネット

パラメータ$\Theta = (\Theta_l)_{l=1}^{L}, \Theta_l \in \mathbb{R}^{n_{l-1} \times n_l}$,

非線形活性化関数

$$\sigma(h) = \max\{0, h\}, \sigma(h) = \frac{1}{1+\exp(-h)}.$$

$$h_0 = x,$$

$$h_l = \sigma\left(\Theta_l^\top h_{l-1}\right) \ (l \in \{0, \dots, L\}),$$

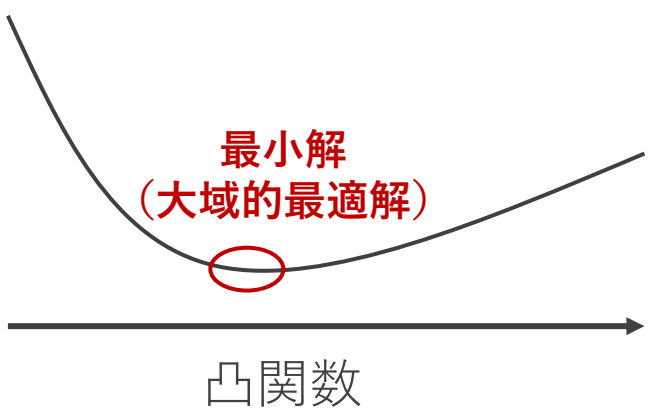
$$o = f_\Theta(x) = h_L.$$

多層ニューラルネットは非線形関数

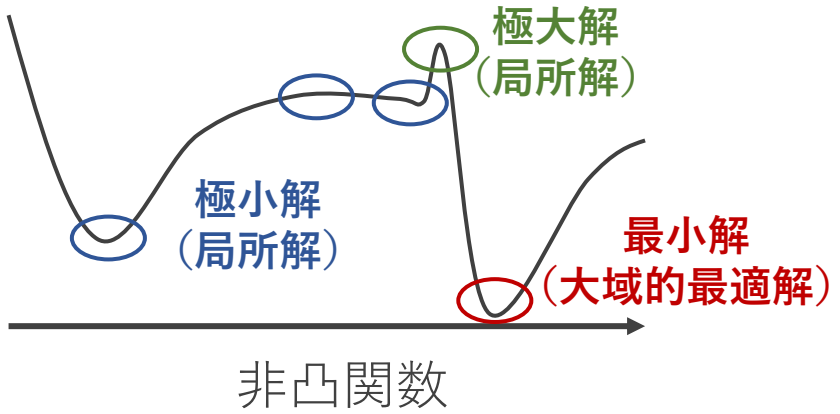
▶ **非凸最適化問題**

非凸最適化問題

一次の勾配法は基本的に**停留点**へ収束：$\nabla_\Theta \mathcal{L}(\Theta^*) = 0.$



凸関数



非凸関数

凸最適化では"停留点＝大域的最適解"

非凸最適化問題に対しての大域収束性保証は困難.

深層学習

高次元ニューラルネットは非凸だが経験的に大域収束する.

Landscape解析

損失関数の性質を解析（真の局所解の存在，大域最適解へのdescent pathの存在）

二乗損失の場合

Negative

- $m = 1$，活性化関数がシグモイド
  ▶ 局所解の数が次元について指数的に増加． [Auer, Herbster, & Warmuth (1996)]
- ReLU，期待損失，教師・生徒ネットワークで同じ中間ノード数$(6 \leq m \leq 20)$
  ▶ 多数の局所解が存在． [Safran & O. Shamir (2018)]
- $n \geq m + 2d - 2$, ReLU
  ▶ 最適解へのdescent pathが存在しないデータが存在（非ゼロ測度） [ICLR submission (2019]

Positive

- Leaky ReLU，$md \geq n$
  ▶ 可微分点かつ局所解は誤差0を達成． [Soudry & Carmon (2016)]
- $m \geq n$，連続な活性化関数
  ▶ 任意の初期点から誤差0 の解へのdescent pathが存在． [Venturi, Bandeira, & Bruna (2018)]

その他，出力層へのskip-connectionが真の局所解を解消する事を示す研究．

平滑化ヒンジ [Liang，Sun，Lee, & Srikant (2018)]，交差エントロピー[Nguyen, Mukkamala, & Hein (2019)].

Landscape解析

Landscape解析に基づく大域的収束性

次の性質を満たせばstrictな鞍点を回避する手法は大域収束.

1. 全ての局所解は最適
2. 全ての鞍点はstrict

該当手法
ノイズ付き勾配法,
負曲率方向の探索法

これらは**二乗活性化関数**に対して示されているが,
（二乗損失の場合 [Soltanolkotabi, A. Javanmard, & J. Lee], 平滑凸損失の場合 [Du & Lee (2018)]）

a. 対象の最適化法が限定的
b. 収束率を出すには更なる仮定も必要
c. 大域解の汎化性能も様々

▶ **勾配法，確率的勾配降下法のより直接的な解析へ**

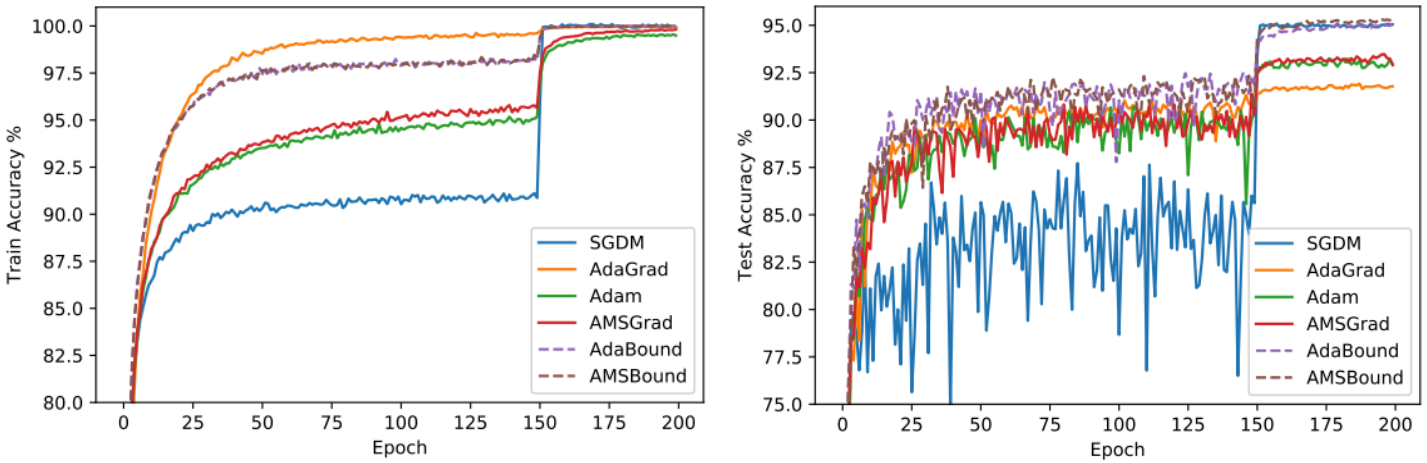
学習アルゴリズムの帰納的バイアス

深層モデルでは**多様な大域解**

大域解の汎化性能も様々

収束先は学習の仕方に依存.

- 層毎のノード数
- パラメータ初期化スケール
- モデルのスケール
- 学習アルゴリズム



左：訓練精度，右：予測精度
[Luo, Xiong, Liu, & Sun (2019)]

学習法が備える帰納的バイアス

学習の条件により本質的に学習ダイナミクスと収束先が特徴付けられる.

▶ 帰納的バイアスで深層モデルを汎化させている説（陰的な正則化）

最適化の重要な研究課題

経験的事実

- 学習アルゴリズムの大域的収束性
- 得られたパラメータの優れた汎化性

これらの現象を説明する理論を構築したい

深層学習における最適化の研究課題

1. 学習アルゴリズムの大域的収束性と収束率
2. 学習法依存の帰納的バイアスの解明

この二つの課題は一体となって解析される場合も多い.

最近のアプローチ：ニューラルタンジェント，Wasserstein勾配流等

2. 高次元ニューラルネットに対する勾配法の大域的収束性と汎化性能解析
（帰納的バイアス：Kernel regime）

勾配法の大域的収束性

2018–2019年，**over-parameterize**された深層ニューラルネットに対し勾配法の大域収束性が示された．[Allen-Zhu，Li, & Song (2019)], [Du, Lee, Li, Wang, Zhai (2019)]

表：大域収束の条件比較（ [Zou & Gu (2019)]より引用）

	Over-para. condition	Iteration complexity	Deep?	ReLU?
Du et al. (2018b)	$\Omega\left(\frac{n^6}{\lambda_0^4}\right)$	$O\left(\frac{n^2\log(1/\epsilon)}{\lambda_0^2}\right)$	no	yes
Wu et al. (2019)	$\Omega\left(\frac{n^6}{\lambda_0^4}\right)$	$O\left(\frac{n\log(1/\epsilon)}{\lambda_0^2}\right)$	no	yes
Oymak and Soltanolkotabi (2019)	$\Omega\left(\frac{n\\|\mathbf{X}\\|_2^6}{\lambda_0^4}\right)$	$O\left(\frac{\\|\mathbf{X}\\|_2^2\log(1/\epsilon)}{\lambda_0}\right)$	no	yes
Du et al. (2018a)	$\Omega\left(\frac{2^{O(L)}\cdot n^4}{\lambda_{\min}^4(\mathbf{K}^{(L)})}\right)$	$O\left(\frac{2^{O(L)}\cdot n^2\log(1/\epsilon)}{\lambda_{\min}^2(\mathbf{K}^{(L)})}\right)$	yes	no
Allen-Zhu et al. (2018b)	$\widetilde{\Omega}\left(\frac{kn^{24}L^{12}}{\phi^8}\right)$	$O\left(\frac{n^6L^2\log(1/\epsilon)}{\phi^2}\right)$	yes	yes
[Zou & Gu (2019)]	$\widetilde{\Omega}\left(\frac{kn^8L^{12}}{\phi^4}\right)$	$O\left(\frac{n^2L^2\log(1/\epsilon)}{\phi}\right)$	yes	yes

DNNの結果

Over-parametrizeでニューラルタンジェントカーネル [Jacot, Gabriel, & Hongler (2019)] のregimeで収束する事が重要．(帰納的バイアス：**NTK regime**)

勾配降下法

経験損失最小化：$\min_{\Theta} \mathcal{L}(\Theta) := \frac{1}{n}\sum_{i=1}^{n} l\left(y_i, f_\Theta(x_i)\right)$.

勾配降下法 $\eta > 0$: 学習率（ステップサイズ），

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta \nabla \mathcal{L}\left(\Theta^{(t)}\right).$$

※ ReLUの場合はsemi-smoothnessを用いる．
[Allen-Zhu, Li, & Song (2019)]

勾配$\nabla \mathcal{L}\left(\Theta^{(t)}\right)$は$\mathcal{L}$を局所的な一次近似：

$$\mathcal{L}(\Theta) = \mathcal{L}\left(\Theta^{(t)}\right) + \nabla \mathcal{L}\left(\Theta^{(t)}\right)^{\mathsf{T}}\left(\Theta - \Theta^{(t)}\right) + O\left(\left\|\Theta - \Theta^{(t)}\right\|_2^2\right).$$

学習率が十分小さい時：

$$\mathcal{L}\left(\Theta^{(t+1)}\right) \leq \mathcal{L}\left(\Theta^{(t)}\right) - \frac{\eta}{2}\left\|\nabla \mathcal{L}\left(\Theta^{(t)}\right)\right\|_2^2 \leq \mathcal{L}\left(\Theta^{(t)}\right).$$

▶ 停留点で無ければ**勾配降下法で目的関数を減少**

ニューラルタンジェントカーネル

[Jacot, Gabriel, & Hongler (2018)]

ニューラルタンジェントカーネル（NTK）は微分$\nabla_\Theta f_\Theta$が定めるカーネル.

NT： $\nabla_\Theta f_\Theta : x \in \mathbb{R}^d \to \nabla_\Theta f_\Theta(x) \in \mathbb{R}^{pd}$,

NTK： $k_\Theta(x, x') = \nabla_\Theta f_\Theta(x)^\top \nabla_\Theta f_\Theta(x')$.

$$k_\Theta(x, x') = \sum_{l=1}^{L} k_{\Theta_l}(x, x') = \sum_{l=1}^{L} \partial_{\Theta_l} f_\Theta(x)^\top \partial_{\Theta_l} f_\Theta(x').$$

訓練データ$(x_i)_{i=1}^n$上のグラム行列：$K_\Theta = \left(k_\Theta(x_i, x_j)\right)_{i,j=1}^n$.

簡単な計算から**関数勾配**： $\nabla_f \mathcal{L}(f_\Theta) = \left(\partial_z l(y_i, f_\Theta(x_i))\right)_{i=1}^n$ に対し,

$$\|\nabla \mathcal{L}(\Theta)\|_2^2 = \frac{1}{n^2} \nabla_f \mathcal{L}(f_\Theta)^\top K_\Theta \nabla_f \mathcal{L}(f_\Theta).$$

ニューラルタンジェントカーネル

[Jacot, Gabriel, & Hongler (2018)]

従って，$\lambda_{min}(K_\Theta) > 0$であれば**大域解でない限り最適化が進む：**

$$\|\nabla_\Theta \mathcal{L}(\Theta)\|_2^2 = \frac{1}{n^2}\nabla_f \mathcal{L}(f_\Theta)^\top K_\Theta \nabla_f \mathcal{L}(f_\Theta) \geq \frac{\lambda_{min}(K_\Theta)}{n^2} \underbrace{\sum_{i=1}^{n} \left|\partial_z l\left(y_i, f_\Theta(x_i)\right)\right|^2}$$

関数勾配ノルム ▶ 大域的最適解でなければ非ゼロ

二乗損失の場合：$\|\nabla_\Theta \mathcal{L}(\Theta)\|_2^2 \geq \frac{2\lambda_{min}(K_\Theta)}{n^2}\mathcal{L}(\Theta)$（パラメータ依存のPL-不等式）

学習中の正定値性を担保するには？

▶ 適切な設定下での**ニューラルネットのOver-parameterization**

NTKによる大域収束性の基本戦略

NTKによる大域収束性を示す基本ステップ

1. 初期点でのNTKの正定値性：$K_{\Theta^{(0)}} \succcurlyeq \lambda I \succ 0$,
2. 最適化中NTKの変化が小さい： $K_{\Theta^{(0)}} \sim K_{\Theta^{(t)}}$.

NTK regime：最適化中，上記性質が保たれる帰納的バイアスにある設定.

※ 二乗損失の場合

▶ NTK regimeでは勾配法は**線形収束（指数収束）**：

$$\mathcal{L}\left(\Theta^{(t+1)}\right) \leq \mathcal{L}\left(\Theta^{(t)}\right) - \frac{\eta}{2}\left\|\nabla_\Theta \mathcal{L}\left(\Theta^{(t)}\right)\right\|_2^2 \leq \underbrace{\left(1 - \frac{1}{2n}\eta\lambda_{min}\left(K_\Theta^{(0)}\right)\right)}_{\text{線形収束性}} \mathcal{L}\left(\Theta^{(t)}\right)$$

NTK regimeの意味

- 正定値カーネル$k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$に付随するRKHSでの勾配降下法：

RKHSでの勾配 $T_k \nabla_f \mathcal{L}(f)(X) = \frac{1}{n} \sum_{i=1}^{n} k(x_i, X) \nabla_f \mathcal{L}(f)(x_i)$により以下の更新

$$f^{(t+1)} = f^{(t)} - \eta T_k \nabla_f \mathcal{L}\left(f^{(t)}\right).$$

- 小さな学習率での勾配法 $\Theta^{(t+1)} = \Theta^{(t)} - \eta \nabla_\Theta \mathcal{L}(\Theta^t)$ は次の関数更新を誘導：

$$f_{\Theta^{(t+1)}} \cong f_{\Theta^{(t)}} - \eta T_{k_{\Theta^{(t)}}} \nabla_f \mathcal{L}\left(f_{\Theta^{(t)}}\right).$$

▶ NTK regimeでは初期カーネル $k_{\Theta^{(0)}}$ に付随するRKHSでの勾配法を近似.
RKHSでの勾配法が帰納的バイアスで早期終了による汎化も期待される.

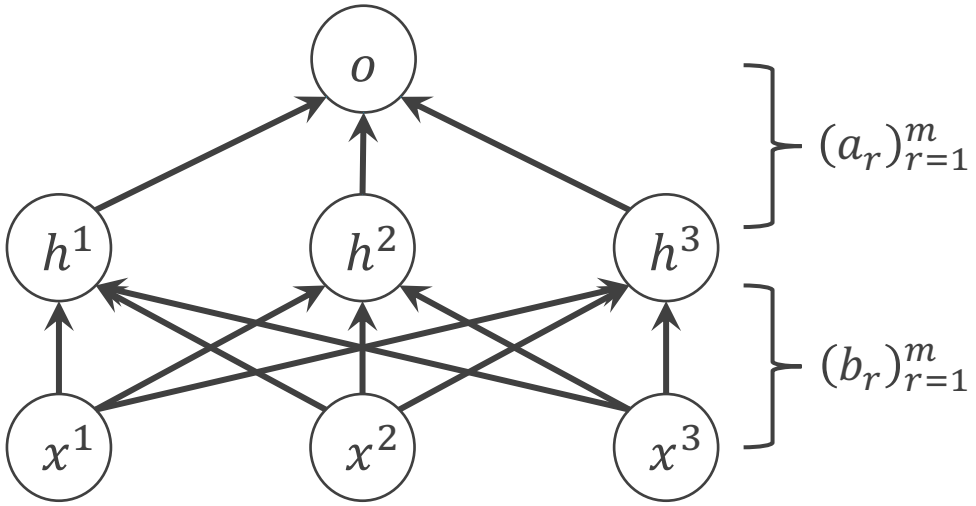
二層ニューラルネット

NTK regimeへの切替は二層の場合が本質的で良く研究されている.

パラメータ: $\Theta_1 = (b_r)_{r=1}^m, \Theta_2 = (a_r)_{r=1}^m$ $\left(b_r \in \mathbb{R}^d, \underbrace{a_r \in \{-\alpha, \alpha\}}_{a_r\text{初期化時の制約}}\right)$,

α: 初期化スケール,

$$f_\Theta(x) = \sum_{r=1}^m a_r \sigma(b_r^\top x)\,.$$



入力層パラメータについて**非凸最適化**で**大域収束性の保証は一般に困難**

適切な設定でのOver-parameterizationでNTK regimeになることを示す.

[Du, Zhai, Poczos, & Singh (2019)]

帰納的バイアス

二層ニューラルネットでは以下の要因で帰納的バイアスが切り替わる：

層の固定の仕方，パラメータの初期化スケール，モデルの出力スケール

- **入力層のみ学習**
  - $\alpha = 1/m$: Mean field regime [Chizat & Bach (2018)], [Mei, Montanari, & Nguyen (2018)]
    $m \to \infty$ での帰納的バイアスは**ワッサースタイン勾配流** [Nitanda & Suzuki 2017]
  - $\alpha = 1/\sqrt{m}$: NTK regime [Du, Zhai, Poczos, & Singh (2019)]
    $m \to \infty$ での帰納的バイアスは$\boldsymbol{k_{\Theta_1}}$**に付随するRKHSでの勾配法**
- **出力層のみ学習**
  ランダム特徴によるNTK regime
  帰納的バイアスは$\boldsymbol{k_{\Theta_2}}$**に付随するRKHSでの勾配法**
- **両層を同時学習** [E, Ma, & Wu (2019)]
  両層が定めるNTK regime，帰納的バイアスは$\boldsymbol{k_{\Theta_1} + k_{\Theta_2}}$**に付随するRKHSでの勾配法**
  十分大きなmに対し，αが小さいと$\boldsymbol{k_{\Theta_2}}$が支配的，$\alpha$が大きいと$\boldsymbol{k_{\Theta_1}}$が支配的
  パラメータの初期化スケールで支配的層が切り替わる

収束解析（入力層の学習）

二層NN: $f_\Theta(x) = \sum_{r=1}^{m} a_r \sigma(b_r^\top x)$, $a_r \sim U\left(\left\{-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\right\}\right)$. （$a_r$は学習中固定）

$$H_1^\infty := \left(\mathbb{E}_b\left[\sigma'(b^T x_i)\sigma'(b^T x_j)x_i^\top x_j\right]\right)_{i,j=1}^{n} = \lim_{m\to\infty} K_{\Theta_1^{(0)}}.$$

※ $\mathbb{E}_b$はbを初期化する分布による期待値.

[Du, Zhai, Poczos, & Singh (2019)]の改良版

定理 [Wu, Du, & Ward (2019)] $\|x\|_2 = 1, y = O(1), \lambda_1 := \lambda_{min}(H_1^\infty) > 0$ とする.

ハイパーパラメータを以下のように設定：

$$m = \Omega\left(\frac{n^6}{\lambda_1^4 \delta^3}\right), \eta = \Theta\left(\frac{1}{\|H^\infty\|_2}\right), T = \underbrace{\tilde{O}\left(\frac{\|H^\infty\|_2}{\lambda_1}\log\left(\frac{1}{\epsilon}\right)\right)}_{\text{線形収束性}}.$$

▶ 勾配法T-反復でϵ-誤差解を達成： $\mathcal{L}\left(\Theta^{(t)}\right) \le \epsilon$.

汎化誤差バウンドも導出可能 [Arora, Du, Hu, & Wang (2019)]：$1-\delta$以上の確率で,

$$\mathbb{E}_{X,Y}\left[l\left(Y, f_{\Theta^T}(X)\right)\right] \le \sqrt{\frac{2y^\top (H^\infty)^{-1} y}{n}} + O\left(\sqrt{\frac{1}{n}\log\left(\frac{n}{\lambda_1 \delta}\right)}\right), \qquad T \ge \Omega\left(\frac{1}{\eta\lambda_1}\log\left(\frac{n}{\delta}\right)\right).$$

※ 但し，ハイパーパラメータは[Wu, Du, & Ward (2019)]と異なる.

収束解析（両層の学習）

二層NN: $f_\Theta(x) = \sum_{r=1}^{m} a_r \sigma(b_r^\top x)$, $a_r \sim U(\{-\alpha, \alpha\})$,

勾配流：$\frac{d\Theta^{(t)}}{dt} = -\nabla \mathcal{L}\left(\Theta^{(t)}\right)$.

$$H^\infty := \alpha^2 H_1^\infty + H_2^\infty, \qquad (\alpha\text{は固定値とする})$$

$$H_1^\infty := \left(\mathbb{E}_b\left[\sigma'(b^T x_i)\sigma'\left(b^T x_j\right)x_i^\top x_j\right]\right)_{i,j=1}^n = \lim_{m\to\infty} \frac{1}{\alpha^2 m} K_{\Theta_1^{(0)}},$$

$$H_2^\infty := \left(\mathbb{E}_b\left[\sigma(b^T x_i)\sigma\left(b^T x_j\right)\right]\right)_{i,j=1}^n = \lim_{m\to\infty} \frac{1}{m} K_{\Theta_2^{(0)}}.$$

定理 [E, Ma, & Wu(2019)] $\|x\|_2 = 1, y = O(1), \lambda_1 := \lambda_{min}(H_1^\infty) > 0, \lambda_2 := \lambda_{min}(H_2^\infty) > 0,$

ノード数mを次のように設定：$m = \Omega\left(\frac{n^6}{\delta(\lambda_1 \wedge \lambda_2)^4} \log\left(\frac{n^2}{\delta}\right)\right)$.

▶ 勾配流で目的関数は指数的収束：

※ 両層の学習によりmがαに非依存.

$$\mathcal{L}\left(\Theta^{(t)}\right) \le \exp\left(-\frac{m}{n}(\alpha^2\lambda_1 + \lambda_2)t\right)\mathcal{L}\left(\Theta^{(0)}\right).$$

αが大きいと入力層，　αが小さいと出力層の学習が支配的となる.

3. 識別問題での二層ニューラルネットの勾配降下法

研究概要

識別問題に対する二層ニューラルネットの勾配降下法の収束解析

- **仮定**
  NTによるデータの識別可能性
  (NTKの正定値性より弱い条件)
- **結果**
  現実的なサイズ$m \ll n$の2層NNに対し勾配法の大域収束性と汎化性を示す.
  Non over-parameterizationに対する収束保証.
  回帰での既存研究では$m \gg n$であった (e.g., $m = \Omega(n^6)$).

下記プレプリントの改良結果を紹介.

A. Nitanda & T. Suzuki. Refined Generalization Analysis of Gradient Descent for Over-parameterized Two-layer Neural Networks with Smooth Activations on Classification Problems. arXiv, 2019a.

2層NNによるロジスティック回帰

- データ: $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} = \{-1,1\}$, $\mathcal{X}\times\mathcal{Y}$上の真の分布$\nu$,
  訓練データ$(x_i, y_i)_{i=1}^n \sim \nu^n$: i.i.d. サンプル.
- モデル: $f_\Theta(x) = \sum_{r=1}^m a_r \sigma(b_r^\top x)$, $a_r \in \{-1,1\}$,　　（入力層のみ学習）
- 損失関数: $l(z,y) = \log(1+\exp(-yz))$.

二値識別の目標は期待識別誤差 $\mathbb{P}_{(X,Y)\sim\nu}[\mathrm{sgn}(f(X)) \neq Y]$ の最小化.
ロジスティック回帰では以下の近似問題で代用.

ロジスティック回帰　ロジスティック損失による経験損失最小化

$$\min_{\Theta_1} \mathcal{L}(\Theta) := \frac{1}{n}\sum_{i=1}^n l\big(y_i, f_\Theta(x_i)\big).$$

回帰問題とNTKの正定値性

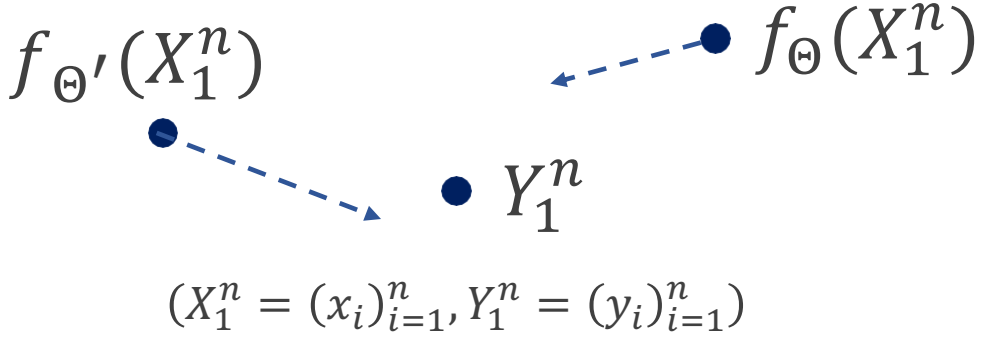
目的関数の減少の条件：

$$\nabla_f \mathcal{L}(f_\Theta)^\top K_\Theta \nabla_f \mathcal{L}(f_\Theta) > 0$$

- NTKのグラム行列の正定値性は関数勾配$\nabla_f \mathcal{L}(f_\Theta)$の方向に必要.
- 二乗損失では$\nabla_f \mathcal{L}(f_\Theta)$ は**任意の方向**を向き得る:

$$\nabla_f \mathcal{L}(f_\Theta) = \left(y_i - f_\Theta(x_i)\right)_{i=1}^n.$$

▶ グラム行列の正定値性を課す理由.



K_Θの正定値性で**関数はラベルに勾配法で接近**
（関数とラベルの位置関係に依らない）

ロジスティック回帰とNTK

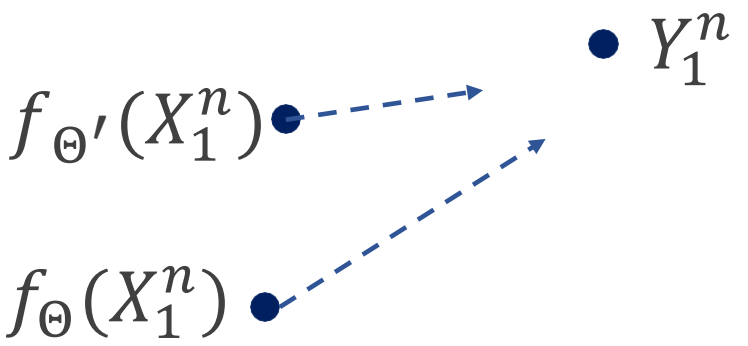
目的関数の減少の条件：

$$\nabla_f \mathcal{L}(f_\Theta)^\top K_\Theta \nabla_f \mathcal{L}(f_\Theta) > 0$$

ロジスティック回帰の関数勾配:

$$\nabla_f \mathcal{L}(f_\Theta) = \left(-y_i \underbrace{\frac{\exp\left(-y_i f_\Theta(x_i)\right)}{1+\exp\left(-y_i f_\Theta(x_i)\right)}}_{(0,1)\text{-値}} \right)_{i=1}^{n}.$$

▶ **ラベルが張る錐上での正定性で十分.**



ロジスティック回帰

$1/(1+\exp(-f_\Theta(X_1^n)))$による$1[Y_1^n = 1]$の回帰.
即ち，ラベルが$\{-\infty, \infty\}$-値の回帰問題.

▶ **関数とラベルの位置関係が不変**
NTKの正定値性は冗長

NTによる識別可能性

仮定 $\exists \rho > 0, \exists v: \mathbb{R}^d \to \{v \in \mathbb{R}^d \mid \|v\|_2 \leq 1\}$ s.t. $\forall (x, y) \in \mathrm{supp}(\nu) \subset \mathcal{X} \times \mathcal{Y}$,

$$\mathbb{E}_b[y \partial_b \sigma(b^\top x)^\top v(b)] \geq \rho.$$

- 無限次元($m = +\infty$) NTによる識別可能性.
- NTKの正定値性より弱い条件.
- ラベルが張る錐上の正定値を誘導：

$$\overline{y}^\top H_1^\infty \overline{y} \geq \rho^2 \|\overline{y}\|_2^2, \qquad (\overline{y} = (\alpha_i y_i)_{i=1}^n, \alpha_i \geq 0).$$

収束解析 [Nitanda & Suzuki (2019a)]

定理 [Nitanda & Suzuki (2019a)] $\mathrm{supp}(\nu^X) \subset \{\|x\|_2 \leq 1\}$, σ はC^2-級で$\|\sigma'\|_\infty, \|\sigma''\|_\infty \leq 1$.
NTの識別可能性を仮定．ハイパーパラメータを以下のいずれかに設定：

(1) $m = \Omega(\epsilon^{-1}), T = \Omega(\epsilon^{-2}), n = \tilde{\Omega}(\epsilon^{-4})$,

(2) $m = \tilde{\Theta}(\epsilon^{-3/2}), T = \tilde{\Theta}(\epsilon^{-1}), n = \tilde{\Omega}(\epsilon^{-2})$.

この時，高確率でT-反復以内ϵ-期待識別誤差を達成：$\exists t \leq T$,

$$\mathbb{P}_{(X,Y)\sim\nu}\left[Y f_{\Theta^{(t)}}(X) \leq 0\right] \leq \epsilon.$$

関連研究と異なり**non over-parameterizationでの大域収束・汎化保証**.

	Activation	Separability	m	n	T
Allen-Zhu, Li, & Liang (2019)	ReLU	Smooth Target	$\tilde{\Omega}(\epsilon^{-10})$	$\Omega(\epsilon^{-4})$	$\tilde{\Theta}(\epsilon^{-2})$
Cao & Gu (2019a)	ReLU	ReLU NN	$\tilde{\Omega}(\epsilon^{-14})$	$\tilde{\Omega}(\epsilon^{-4})$	$\tilde{\Theta}(\epsilon^{-2})$
Cao & Gu (2019b)	ReLU	ReLU NN	$\tilde{\Omega}(\epsilon^{-14})$	$\tilde{\Omega}(\epsilon^{-2})$	$\tilde{\Theta}(\epsilon^{-2})$
Nitanda & Suzuki (2019a)	Smooth	Neural Tangent	$\Omega(\epsilon^{-1})$ $\tilde{\Theta}(\epsilon^{-3/2})$	$\tilde{\Omega}(\epsilon^{-4})$ $\tilde{\Omega}(\epsilon^{-2})$	$\Theta(\epsilon^{-2})$ $\tilde{\Theta}(\epsilon^{-1})$

これらはDNNにも対応.

4. カーネル法・Rich Regimes

RKHSでの確率的勾配降下法（SGD）

[Nitanda & Suzuki (2019b)], [Yashima, Nitanda, & Suzuki (2019)]

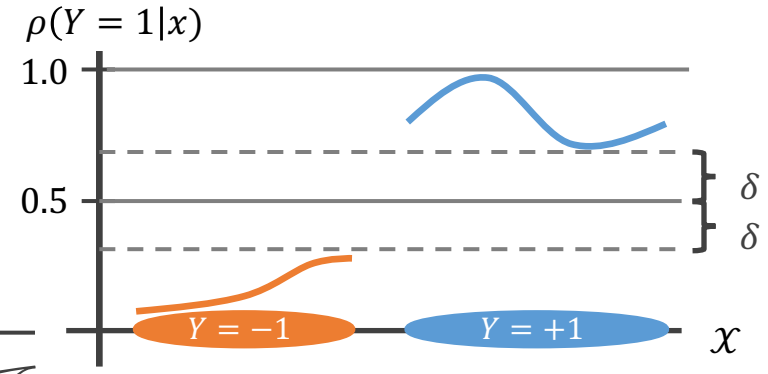
出力層のみの学習はRKHS $\mathcal{H}_k$での学習に対応し，多数の研究がある．

L_2-ロジスティック回帰に対するSGD： $g_t \in \mathcal{H}_k, (x_t, y_t) \sim \rho$,

$$g_{t+1} = (1 - \eta_t \lambda) g_t - \eta_t \partial_z l(g_t(x_t), y_t) k(x_t, \cdot).$$

仮定（強低ノイズ条件）$\exists \delta \in (0, 1/2)$, for X a.e. w.r.t. ρ_X,

$$|\rho(Y = 1|X) - 0.5| > \delta.$$



$$\mathcal{R}(g) := \mathbb{P}_{(X,Y)\sim\rho}[Yg(X) \leq 0], \mathcal{R}_* = \min_g \mathcal{R}(g).$$

定理 [Nitanda & Suzuki (2019b)] 適当な仮定の下，$\exists T_0, \forall T \geq T_0$に対して

$$\mathbb{E}[\mathcal{R}(g_{T+1}) - \mathcal{R}_*] \leq 2 \exp\left(-O(\lambda^2 T) \log^2\left(\frac{1+2\delta}{1-2\delta}\right)\right)$$

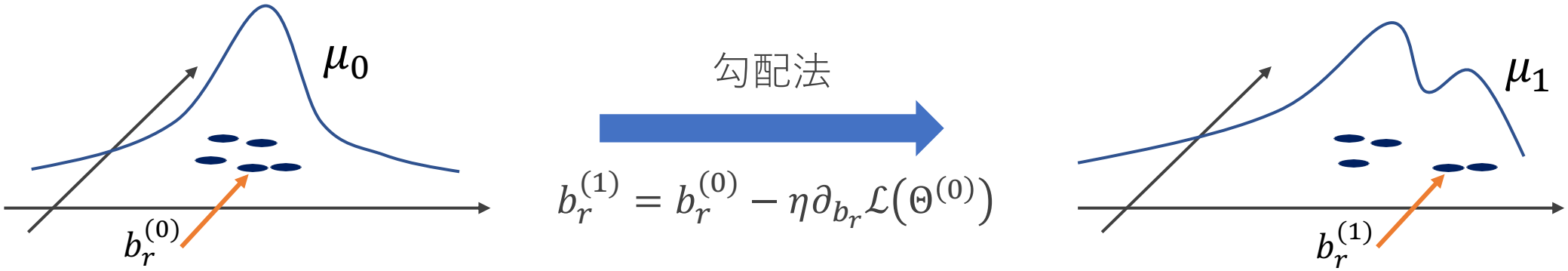
期待識別誤差の線形収束性

- 同様の収束性はRandom Featureモデルでも成立 [Yashima, Nitanda, & Suzuki (2019)]
- 2層NNの両層の学習でも成立（ongoing work）

Particle Gradient Descent [Nitanda & Suzuki (2017)]

二層NN: $f_\Theta(x) = \frac{1}{m}\sum_{r=1}^{m} \sigma(b_r^\top x)$ (mean field regime, MFR), 初期化：$b_r^{(0)} \sim \mu_0$.

$$f_{\Theta^{(0)}}(x) = \frac{1}{m}\sum_{r=1}^{m} \sigma\left(b_r^{(0)\top} x\right) \overset{(m\to\infty)}{\to} f_{\mu_0}(x) = \mathbb{E}_{b^{(0)}\sim\mu_0}\left[\sigma\left(b^{(0)T} x\right)\right].$$



パラメータ$\Theta^{(0)} = \left(b_r^{(0)}\right)_{r=1}^{m}$は分布$\mu_0$のサンプルの集まり.

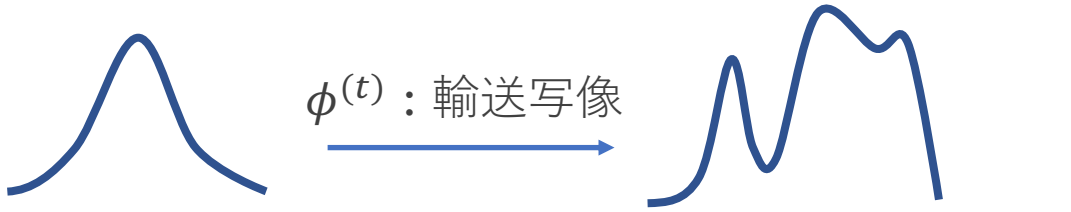
勾配法により$\Theta^{(1)} = \left(b_r^{(1)}\right)_{r=1}^{m}$に更新され，背後にある分布も$\mu_1$に更新される.

▶ **(確率的)勾配法は暗に確率分布を最適化:** $\min_\mu \mathcal{L}(\mu)$. その更新即は？

▶ **(Stochastic) Particle Gradient Descent**

輸送写像による確率測度最適化 [Nitanda & Suzuki (2017)]

$\mathbb{R}^d$上の輸送写像$\phi^{(t)}: \mathbb{R}^d \to \mathbb{R}^d$は分布を変形する.
（即ち，変数変換で確率変数を$\phi^{(t)}$と合成し更新）



$\mu^{(t)}$:確率分布
$b \sim \mu^{(t)}$

$\mu^{(t+1)}$: 更新された分布
$\phi^{(t)}(b) \sim \mu^{(t+1)} = \phi^{(t)}_{\#}\mu^{(t)}$

▶ 確率測度の最適化を輸送写像の最適化$\mathcal{L}(\phi) = \mathcal{L}(\phi_{\#}\mu^{(0)})$に帰着.

▶ $L_2^d(\mu^{(0)})$におけるフレシェ微分（関数勾配）$\nabla_\phi \mathcal{L}(\phi^{(t)})$を用いた**関数勾配法**：

$$\phi^{(t+1)} = \phi^{(t)} - \eta v_t \circ \phi^{(t)}, \qquad (\nabla_\phi \mathcal{L}(\phi^{(t)}) = \exists v_t \circ \phi^{(t)})$$
$$\mu^{(t+1)} = (id - \eta v_t)_{\#}\mu^{(t)}.$$

（付随する確率測度の更新即：SPGD）

これはパラメータの勾配法に対応：$b_r^{(t+1)} \sim (id - \eta v_t)(b_r^{(t)})$. （$m = \infty$では厳密に一致）

▶ **パラメータの勾配法は輸送写像の最適化による確率測度の最適化.**

SPGD法に対し輸送写像空間での局所解への収束$O(\epsilon^{-2})$も保証. [Nitanda & Suzuki (2017)]

ワッサースタイン勾配流 [Nitanda & Suzuki (2017)]

勾配法は無限小の学習率の下では勾配流のダイナミクスに従う.

SPGD法のダイナミクスは？

▶**ワッサースタイン勾配流（W勾配流）**に従う：[Nitanda & Suzuki (2017)]

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t = -\mathrm{div}(-v_t\mu_t), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}(\mu_t) = -\|v_t\|^2_{L_2(\mu_t)}.$$

- W勾配流の**大域収束性**は[Chizat & Bach (2018)]が証明.　（即ち無限次元2-NNの勾配流）
- [Nitanda & Suzuki (2017)] は**収束率付き**の局所収束性をSPGD（即ち**離散ステップの確率的勾配降下法**）に対して証明. 更に改良版SPGDで**有限パーティクルでの収束性**も保証.

その他, ノイズ付きSGDは[Mei, Montanari, & Nguyen (2018)]が解析.

Active Regime

- NTK regimeの要因：微小なパラメータ変化で関数を十分に動かせる(lazy training).
- 出力層固定の2層NNの場合：$\alpha = 1/m$を境界に切り替わる．（MFRはNTK regimeでない）
- モデル$\alpha f\ (\alpha > 0)$：αが大きいとkernel regime. [Chizat & Bach (2019)]

αが小さい場合active (adaptive)なregime.

定理（Homogeneous linear modelの帰納バイアス）$w_+, w_- \in \mathbb{R}^d, \alpha > 0, L \in \mathbb{N}$,
[Woodworth+ (2019)]

$$f_{w_+,w_-}(x) = \alpha^L (w_+^L - w_-^L)^\top x.$$

$w_+(t), w_-(t)$を二乗損失に対するw_+, w_-の勾配流とし $\beta_\alpha(t) = \alpha^L (w_+^L(t) - w_-^L(t))$.

$n \ll d$の時，$\beta_\alpha(\infty)^\top x_i = y_i, (i \in \{1, \dots, n\})$, を満たし以下の特徴付けが成立：

$$\lim_{\alpha \to 0} \beta_\alpha(\infty) = \beta^*_{L_1}, \qquad \lim_{\alpha \to \infty} \beta_\alpha(\infty) = \beta^*_{L_2},$$

ここで$\beta^*_{L_1}, \beta^*_{L_2}$は$\beta^\top x_i = y_i, (i \in \{1, \dots, n\})$, の最小$L_1, L_2$-ノルム解.

▶α が小さい時，L_1-正則化が帰納的バイアス（active regime）
カーネル法はスパース解がある時，その性質を活かせない.

まとめ

- ニューラルネットに対する勾配法の近年の研究を紹介.
- 二層の場合の理論が現状本質的.
- 種々の要因で帰納的バイアスの切り替えが起こる事を紹介.

  NTK regime, Mean field regime, Active regime.

  特にNTK regimeの理論を解説.

より最近の研究の流れ

- NTK regimeの更なる研究（より詳細な解析，NTK regimeの妥当性）.
- NTK以外のregimeの研究.

  Mean field regimeとHomogeneous linear modelはその成功例.

  NTKより高次の帰納的バイアスの研究もされつつある（モデルの二次近似，ResNet）.

参考文献

1章

- P. Auer, M. Herbster, & MK. Warmuth. Exponentially many local minima for single neurons. *NIPS*, 1996.
- D. Soudry & Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv*, 2016.
- D. Soudry & E. Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv*, 2017.
- M. Soltanolkotabi, A. Javanmard, & J. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.
- I. Safran & O. Shamir. Spurious local minima are common in two-layer relu neural networks. *ICML*, 2018.
- S. Du & J. Lee. On the power of over-parametrization in neural networks with quadratic activation. *ICML*, 2018.
- L. Venturi, AS. Bandeira, & J. Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv*, 2018
- S. Liang, R. Sun, J. Lee, & R. Srikant. Adding One Neuron Can Eliminate All Bad Local Minima. *NeurIPS*, 2018.
- Anonymous. Bounds on Over-Parameterization for Guaranteed Existence of Descent Paths in Shallow ReLU Networks. *ICLR submission*, 2019.
- Q. Nguyen, MC. Mukkamala, & M. Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. *ICLR*, 2019.
- L. Luo, Y. Xiong, Y. Liu, & X. Sun. Adaptive Gradient Methods with Dynamic Bound of Learning Rate. *ICLR*, 2019.

2-3章

- A. Jacot, F. Gabriel, & C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *NeurIPS*, 2018.
- S. Du, X. Zhai, B. Poczos, & A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *ICLR*, 2019.
- Z. Allen-Zhu, Y. Li, & Z. Song. A Convergence Theory for Deep Learning via Over-Parameterization. *ICML*, 2019.
- S. Du, J. Lee, H. Li, L. Wang, X. Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. *ICML*, 2019.
- D. Zou & Q. Gu. An Improved Analysis of Training Over-parameterized Deep Neural Networks. *NeurIPS*, 2019.

参考文献

- X. Wu, S. Du, & R. Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network. *arXiv*, 2019.
- S. Arora, S. S. Du, W. Hu, & R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *ICML*, 2019.
- W. E, C. Ma, & L. Wu. A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics. *arXiv*, 2019.
- A. Nitanda & T. Suzuki. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv*, 2019a.
- Z. Allen-Zhu, Y. Li, & Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *NeurIPS*, 2019.
- Y. Cao & Q. Gu. A generalization theory of gradient descent for learning over-parameterized deep relu networks. *arXiv*, 2019a.
- Y. Cao & Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *NeurIPS*, 2019b.

4章

- A. Nitanda & T. Suzuki. Stochastic Gradient Descent with Exponential Convergence Rates of Expected Classification Errors. *AISTATS*, 2019b.
- S. Yashima, A. Nitanda, & T. Suzuki. Exponential Convergence Rates of Classification Errors on Learning with SGD and Random Features. *arXiv*, 2019.
- A. Nitanda & T. Suzuki. Stochastic particle gradient descent for infinite ensembles. *arXiv*, 2017.
- L. Chizat & F. Bach. On the global convergence of gradient descent for over-parameterize models using optimal transport. *NeurIPS*, 2018.
- S. Mei, A. Montanari, & P-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *PNAS*. 2018.
- L. Chizat & F. Bach. On Lazy Training in Differentiable Programming. *NeurIPS*, 2019.
- B. Woodworth, S. Gunasekar, P. Savarese, E. Moroshko, I. Golan, J. Lee, D. Soudry, & N. Srebro. Kernel and Rich Regimes in Overparametrized Models. *arXiv*, 2019.