# ニューラルネットワークの最適化理論

二反田　篤史

確率的勾配降下法はニューラルネットワークの最適化手法として古くより利用されてきた．その有用性を説明するためには非凸最適化問題に対する大域的収束性の証明という困難な問題に踏み込む必要があるが，近年の研究により特定条件下において理解が進みつつある．本稿ではニューラルタンジェントカーネルおよび平均場理論に基づく勾配降下法の収束理論を概説する．

キーワード：ニューラルネットワーク，勾配降下法，ニューラルタンジェントカーネル，平均場理論

## 1. はじめに

深層ニューラルネットワークがさまざまな分野で成功を収めているが，その優れた性能を理論的に裏付けるためには次の問題を解決する必要がある．(I) 非凸最適化問題であるニューラルネットワーク学習に対する最適化手法の大域的収束性（大域的最適解への収束性），(II) 過剰なパラメータ数を備える高次元ニューラルネットワークの汎化誤差保証（未知データへの適合性の保証）．深層学習のパフォーマンスが最適化手法に大いに依存していることから，これらの問題は別々に扱うのではなく最適化の観点から統一的に議論する必要があると考えられている．そのためには非凸最適化問題の大域的収束性の証明という困難な課題に向き合う必要があるが，高次元二層ニューラルネットワークの勾配降下法に対しては特定の条件下で部分的に解決されはじめている．証明の鍵は高次元性のもと二層ニューラルネットワークの学習ダイナミクスをニューラルタンジェントカーネル [1] あるいは平均場理論 [2] に基づき解析することである．本稿ではこれらの理論に関する最近の進展 [3–8] を紹介する．

## 2. 機械学習と最適化

機械学習の目標は入出力空間上の未知のデータ分布に適合する真の入出力関係を獲得することである．この目標は期待損失最小化問題の求解により実行される．一般に探索空間はパラメータを備えた数理モデル $\{g_w : \mathbb{R}^d \to \mathbb{R} \mid w \in \mathbb{R}^p\}$ で表現する．ここで $w \in \mathbb{R}^p$ はパラメータ，ユークリッド空間 $\mathbb{R}^d$ は入力空間に対応する．期待損失最小化問題は次のように定義される．

$$\min_{w\in\mathbb{R}^p} \left\{ \mathcal{L}(w) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)}[\ell(g_w(X), Y)] \right\}. \quad (1)$$

ここで $(X, Y)$ は入力データとその出力を表す確率変数であり未知のデータ分布 $\rho$ に従う．$\ell$ は $g_w(x)$ と $y$ の適合度を示す損失関数（小さいほど適合）である．本稿では損失関数 $\ell(z, y)$ は $z$ について可微分とする．損失関数の代表例として二乗損失 $\ell(g_w(x), y) = \frac{1}{2}(g_w(x) - y)^2$ $(y \in \mathbb{R})$ やロジスティック損失 $\ell(g_w(x), y) = \log(1 + \exp(-y g_w(x)))$ $(y \in \{-1, 1\})$ などがあり，それぞれ実数値を予測する回帰問題，バイナリ値を予測する識別問題で用いられる．この問題の目的関数 $\mathcal{L}$ は期待損失関数，そしてそれを最小化する可測関数 $g_\rho$ はベイズ規則とよばれる．ベイズ規則の獲得が機械学習の目標となるが，数理モデルがこのベイズ規則を含まない場合はその誤差（近似誤差）も機械学習の理論では考慮する必要がある．しかしながら本稿の目的は機械学習の最適化の解説であるため近似誤差の解析には踏み込まないことにする．

一般には $(X, Y)$ のデータ分布 $\rho$ は未知なので実際には独立に得られる有限個のサンプル（訓練データ）$\{(x_i, y_i)\}_{i=1}^n$ を手掛かりに期待損失最小化を試みる，この過程を学習という．代表的なアプローチとしては経験損失最小化問題がある．経験損失最小化問題では期待損失関数をサンプル平均で近似した経験損失関数の最適化を行う．ただし訓練データに対してモデルの表現力が高い場合は未知データに適合しない過学習という問題が起こり得る．このような問題を避けるため経験損失関数に正則化項 $h(w)$ を加えることもある．この場合は特に正則化付き経験損失最小化問題とよばれ次のように定義される．

$$\min_{w\in\mathbb{R}^p} \left\{ \mathcal{L}_n(w) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \ell(g_w(x_i), y_i) + h(w) \right\}. \quad (2)$$

にたんだ　あつし
東京大学大学院情報理工学系研究科
〒 113–8654　東京都文京区本郷 7–3–1
nitanda@mist.i.u-tokyo.ac.jp

正則化の代表例として $L_1$ 正則化 $h(w) = \lambda\|w\|_1$ や $L_2$ 正則化 $h(w) = \frac{\lambda}{2}\|w\|_2^2$ がある．正則化付き経験損失最小化問題の解の期待損失関数値の性質については文献 [9–12] などを参照のこと．

経験損失最小化問題を解くための最も単純な方法は勾配降下法の適用である．すなわち，$w^{(1)} \in \mathbb{R}^p$ を初期点として以下の方法でパラメータを逐次更新する．

$$w^{(t+1)} = w^{(t)} - \eta_t \nabla \mathcal{L}_n(w^{(t)}). \tag{3}$$

ここで $\eta_t > 0$ はステップサイズである．勾配降下法は最適化アルゴリズムによって得られるパラメータの性質を調べるために機械学習では現在においても非常に重要な研究対象であるが，計算コストの観点から大規模機械学習問題において使用されることはあまりない．実際，勾配 $\nabla\mathcal{L}_n(w)$ の評価時に全サンプルについての計算コスト $O(n)$ が発生してしまうという問題がある．そこで，勾配降下法を確率化することでこの問題を解消した確率的勾配降下法あるいはその派生手法が大規模機械学習では有効である．確率的勾配降下法は Robbins and Monro [13] により 1951 年に提案された．勾配降下法では反復点の更新時に勾配の評価を必要とするが，確率的勾配降下法では勾配に観測ノイズが加わる場合を想定する．ここでは次の問題を考える．

$$\min_{w\in\mathbb{R}^p} \Big\{ f(w) \stackrel{\text{def}}{=} \mathbb{E}[g(w,\zeta)] \Big\}. \tag{4}$$

$g$ は $\mathbb{R}^{p+m}$ 上の可微分な実数値関数，$\zeta$ は $\mathbb{R}^m$ に値を取る確率変数，$\mathbb{E}$ は $\zeta$ の従う確率分布による積分である．次に，この最適化問題 (4) に対する確率的勾配降下法を説明する．確率変数 $\zeta$ の分布を期待損失を定義する未知のデータ分布，あるいは経験損失を定義するサンプルによる経験分布にすることで期待損失最小化問題 (1) と経験損失最小化問題 (2) のいずれもこの定式化に含まれており，確率的勾配降下法はどちらの問題にも適用可能であることに注意しておく．ここで $\{\zeta_t\}_{t=1}^{\infty}$ は確率変数 $\zeta$ と同じ分布に従う独立な確率変数の列とする．このとき確率的勾配降下法の更新式は次で定義される．

$$w^{(t+1)} = w^{(t)} - \eta_t \partial_w g(w^{(t)}, \zeta_t).$$

確率変数 $\partial_w g(w^{(t)}, \zeta_t)$ は確率的勾配とよばれ，$w^{(t)}$ において $\mathbb{E}_{\zeta_t}[\partial_w g(w^{(t)}, \zeta_t)] = \nabla f(w^{(t)})$ を満たす．すなわち確率的勾配は勾配の不偏推定量に他ならず平均的な目的関数の減少が期待される．

確率的勾配降下法が収束するためには確率的勾配のノイズの影響を打ち消すためにステップサイズ $\eta_t$ を適切に減少させる必要がある．Robbins and Monro [13] では $\eta_t = O(1/t)$ のもと収束性が証明された．たとえば $g(w,\zeta)$ に $w$ についての強凸性とリプシッツ平滑性を，確率的勾配 $\partial_w g(w,\zeta)$ の分散に有界性を課した場合，$\mathbb{E}[f(w^{(t+1)}) - f_*] = O(1/t)$ という収束性が示される [14]．ここで $f_*$ は目的関数の下限 $f_* = \inf_{w\in\mathbb{R}^p} f(w)$ であり期待値は $\zeta_1, \zeta_2, \ldots, \zeta_t$ について計算される．その後，反復点列の平均をとる Polyak 平均化法を適用すると，より大きなステップサイズで安定的に収束することも示されている [15]．

この設定のもと，経験損失最小化 (2) を例に確率的勾配降下法と勾配降下法の計算量を比較してみよう．ここでは与えられた $\epsilon > 0$ に対し $\mathbb{E}[f(w^{(t+1)}) - f_*] \le \epsilon$ を達成するために必要な微分 $\partial_w g(w,\zeta)$ の計算回数[1]で両手法を評価する．確率的勾配降下法ではパラメータの更新に 1 サンプルしか用いないため，反復ごとの計算コストが $O(1)$ であることに注意すると，$\epsilon$ 誤差の達成に必要な計算量は $O(1/\epsilon)$ となる．一方，勾配降下法は線形収束するものの反復ごとの計算コストは $O(n)$ であるため $\epsilon$ 誤差の達成に必要な計算量は $O(n\log(1/\epsilon))$ となる．このように勾配降下法の確率化によりパラメータの更新回数についての収束性は線形収束から上記で説明したような劣線形収束へと劣化するが，$\epsilon$ 誤差の達成に必要な計算量は勾配降下法の場合と異なりサンプルデータ数 $n$ に非依存になることがわかる．これは大規模機械学習問題に対して確率的勾配降下法がより圧倒的に高速になることを意味し，確率的勾配降下法が機械学習で重宝される理由である．しかしながら理論解析においては勾配降下法もいまだ重要な研究対象であることに注意されたい．

## 3. ニューラルネットワークの学習

ニューラルネットワークとは機械学習モデルの一つであり，畳み込みニューラルネットワークなどに代表される派生モデルは画像認識，音声認識，自然言語処理の分野で非常に高い性能を発揮している．そして対応する経験損失および期待損失最小化問題は非凸最適化問題であるにもかかわらず多くの場合で大域的収束[2]することが経験的に知られている．一般に非凸最適化問題に対する大域的収束性の証明は困難であるが，ニューラルネットワークに対しては特定条件下でその性質が明

[1] 勾配あるいは確率的勾配でパラメータを更新する一次最適化手法の比較においては公平な計算量である．

[2] 数理最適化の文脈とは異なり機械学習では大域的収束性は最適値への収束性を意味することに注意されたい．

らかにされつつある．簡単のため，ここでは次の二層ニューラルネットワークを考える．$a_r \in \mathbb{R}$, $b_r \in \mathbb{R}^d$, $M \in \mathbb{N}$ として，

$$g_w(x) = \sum_{r=1}^{M} a_r \sigma(b_r^\top x)\,.$$

ここで，$\sigma : \mathbb{R} \to \mathbb{R}$ はシグモイド関数 $\sigma(v) = 1/(1+\exp(-v))$, ReLU 関数 $\sigma(v) = \max\{0, v\}$ などの活性化関数で $M$ は中間ノード数，$a_r$, $b_r$ はそれぞれ出力層，入力層のパラメータである．活性化関数によって $g_w$ は一般に非線形関数となる．また中間ノード数 $M$ が増加するにつれて $g_w$ が表せる関数系も増大していく．このように二層ニューラルネットワーク $g_w$ が定める関数系は活性化関数 $\sigma$ の種類と中間ノード数 $M$ に依存するため，機械学習を実行する際には，最適化後の期待損失を推定する手続き（交差検証など）を用いて適当な $\sigma$ と $M$ を選択する．以降，本稿では入力層パラメータ $w = \{b_r\}_{r=1}^M$ の（確率的）勾配降下法による最適化に注目する．活性化関数の非線形性から，この場合でも学習は非凸最適化問題に帰着される．出力層は $a_r = O(1/M)$ あるいは $a_r = O(1/\sqrt{M})$ というスケールで初期化する．この初期化法に応じて勾配降下法の収束解析は平均場理論 [2, 4] とニューラルタンジェントカーネル理論 [1] に分岐する．

### 3.1 ニューラルタンジェントカーネル

二乗損失 $\ell(z, y) = \frac{1}{2}(z-y)^2$ を用いた回帰問題を対象にニューラルタンジェントカーネルの概要を説明する．最適化問題は訓練データ $\{(x_i, y_i)\}_{i=1}^n$ が定める経験損失最小化問題 (2) を考える．正則化項はないもの，すなわち $h \equiv 0$ とする．説明の簡略化のため活性化関数は十分に滑らかとする．このとき，固定ステップサイズ $\eta > 0$ が十分小さければ，目的関数の滑らかさから勾配降下法 $w^{(t+1)} = w^{(t)} - \eta \nabla \mathcal{L}_n(w^{(t)})$ によって目的関数は勾配ノルムの二乗とステップサイズの積の分減少する．

$$\mathcal{L}_n(w^{(t+1)}) \le \mathcal{L}_n(w^{(t)}) - \frac{\eta}{2}\|\nabla \mathcal{L}_n(w^{(t)})\|_2^2\,.$$

この減少量を評価するためにニューラルタンジェントカーネルを導入する．ニューラルタンジェントカーネルは次で定義されるデータ空間上のカーネル関数である．

$$k_w(x, x') = \nabla_w g_w(x)^\top \nabla_w g_w(x')\,. \tag{5}$$

訓練データ $\{x_i\}_{i=1}^n$ 上のグラム行列を $K_w = (k_w(x_i, x_j))_{i,j=1}^n$ とおく．ここで関数 $g_w$ 自身を変数とみたときの関数勾配を

$$\nabla_g \mathcal{L}_n(g_w) = \left(\partial_z \ell(z, y_i)|_{z=g_w(x_i)}\right)_{i=1}^n$$

で定義する．ここでは二乗損失を考えているので関数勾配は $(g_w(x_i) - y_i)_{i=1}^n$ となる．このとき，$K_w$ の最小固有値を $\lambda_w$ とすれば勾配ノルムは次の不等式を満たす．

$$\begin{aligned}\|\nabla \mathcal{L}_n(w^{(t)})\|_2^2 &= n^{-2} \nabla_g \mathcal{L}_n(g_w)^\top K_w \nabla_g \mathcal{L}_n(g_w) \\ &\ge 2\lambda_w n^{-1} \mathcal{L}_n(w).\end{aligned}$$

ゆえに勾配降下法による経験損失の減少は

$$\mathcal{L}_n(w^{(t+1)}) \le \left(1 - \frac{\eta \lambda_{w^{(t)}}}{n}\right) \mathcal{L}_n(w^{(t)})$$

となる．Du et al. [5] は適当な条件下でノード数 $M$ を過剰に大きくとると高確率で $\lambda_{w^{(1)}} > 0$ となることと最適化の過程で $K_{w^{(t)}}$ が初期のグラム行列 $K_{w^{(1)}}$ からあまり変化せず正定値性が保たれつづけ大域的収束することを証明した．またこの理論は $M \to \infty$ のもとで勾配降下法が $k_{w^{(1)}}$ に付随する再生核ヒルベルト空間における勾配降下法に漸近するという事実も示している．次の定理は Du et al. [5] による大域的収束性定理の改良版 [16] である．$H_{1,\infty} = \lim_{M\to\infty} K_{w^{(1)}}$ とおき，$H_{1,\infty}$ の最小固有値を $\lambda_{1,\infty}$ とおく．$\{(x_i, y_i)\}_{i=1}^n$ を $(X, Y)$ の $n$ 個のサンプル，$\|\cdot\|_F$ をフロベニウスノルムとする．

**定理 1.** $\sigma$ は ReLU 関数として，$\|x_i\|_2 = 1, y_i = O(1)$ $(i \in \{1, 2, \ldots, n\})$，$\lambda_{1,\infty} > 0$ とする．このとき，$M = \Omega(n^6/\lambda_{1,\infty}^4)$，$\eta = \Theta(1/\|H_{1,\infty}\|_F)$ となるように設定すると，任意の $\epsilon > 0$ に対し最急降下法によって $O\left(\frac{\|H_{1,\infty}\|_F}{\lambda_{1,\infty}} \log(1/\epsilon)\right)$ 反復以内に高確率で $\mathcal{L}_n(w^{(t)}) \le \epsilon$ が満たされる．

ここでは二層ニューラルネットワークに対する勾配降下法に焦点を当てたが，類似の結果は多層ニューラルネットワークに対する確率的勾配降下法の場合でも成立する．またニューラルタンジェントカーネルの理論とラデマッハー複雑度の解析を組み合わせて Arora et al. [3] は以下の期待損失関数の上界を与えた．訓練データのラベルの列を $y_{1,n} = (y_i)_{i=1}^n$ とおく．このとき，パラメータの初期化と訓練データのサンプリングに関して $1-\delta$ 以上の確率で十分大きな反復数 $T$ に対し次の量は期待損失 $\mathcal{L}(w^{(T)}) = \mathbb{E}_{(X,Y)}[(Y - g_{w^{(T)}}(X))^2]$ の上界となる．

$$\sqrt{\frac{2y_{1,n}^\top H_{1,\infty}^{-1} y_{1,n}}{n}} + O\left(\sqrt{\frac{1}{n}\log\left(\frac{n}{\lambda_{1,\infty}\delta}\right)}\right). \quad (6)$$

ただし，サンプル数 $n$ の増加に伴い $H_{1,\infty}$ の最小固有値は 0 に収束していくためこのバウンドの $n$ についての収束率は一般に $O(1/\sqrt{n})$ よりも遅くなることに注意されたい[3].

### 3.2 平均化確率的勾配降下法による最適収束率

本節では二乗損失の定める期待損失最小化問題に対する平均化確率的勾配降下法の最適性についての研究 [8] を解説する．Arora et al. [3] の上界で具体的な収束率を導出できない理由としては $n \to \infty$ のもとグラム行列 $H_{1,\infty}$ が退化することと固有ベクトルとラベルの関係性が特定されていないことにある．実際，カーネル法を用いた確率的勾配降下法や正則化付き経験損失最小化による推定によって，$O(1/\sqrt{n})$ よりも速い期待損失の収束率 $O(n^{\frac{-2r\beta}{2r\beta+1}})$[17] が真の関数とカーネルが定める積分作用素についての仮定のもとで達成される．ここで，$r \in [1/2, 1]$ はベイズ規則の複雑さであり，$\beta > 1$ は再生核ヒルベルト空間の大きさを表す．このことからニューラルタンジェントカーネルの理論とカーネル法の理論に大きなギャップがあることがわかる．このような状況の中，二層ニューラルネットワークに対する確率的勾配降下法の高速な収束性が適切な設定のもと文献 [8] で示された．以降も入力層パラメータの最適化を考えるが，本節の結果は出力層のパラメータも同時に最適化した場合にも自然に拡張される．

次の確率的勾配降下法を考える．

$$w^{(t+1)} = w^{(t)} - \eta_t \partial_w \ell_t(g_{w^{(t)}}) - \eta\lambda(w^{(t)} - w^{(1)}).$$

ここで $\ell_t(g) = \ell(g(x_t), y_t)$ とおいた．データ $(x_t, y_t)$ は確率的勾配降下法の各反復において真のデータ分布からサンプリングされるものとする．これは以下の初期点周りの正則化付き期待損失に対する確率的勾配降下法に他ならない．

$$\mathcal{L}(g_w) + \frac{\lambda}{2}\|w - w^{(1)}\|_2^2.$$

ここで $\|w - w^{(1)}\|_2^2 = \sum_{r=1}^M \|b_r - b_r^{(1)}\|_2^2$ である．ただし予測は $T$ 反復分の平均 $\overline{w}^{(T+1)} = \frac{1}{T+1}\sum_{t=1}^{T+1} w^{(t)}$ で行い，収束解析も $\overline{w}^{(T+1)}$ を対象とする．パラメータの初期化は $g_{w^{(1)}} \equiv 0$ となるように対称的に行う．すなわち，ノード数 $M$ は偶数とし $a_r = 1/\sqrt{M}$ $(r \in \{1, 2, \ldots, M/2\})$, $a_r = -1/\sqrt{M}$ $(r \in \{M/2+1, M/2+2, \ldots, M\})$ とする．入力層パラメータ $b_r$ $(r \in \{1, 2, \ldots, M/2\})$ は台が単位球に含まれる確率分布 $\mu_0$ に従い初期化し $b_r = b_{r+M/2}$ $(r \in \{1, 2, \ldots, M/2\})$ とする．$M = \infty$ のもとでのニューラルタンジェントカーネルを次のように定義する．

$$k_\infty(x, x') = x^\top x' \mathbb{E}_{b^{(1)}}[\sigma'(b^{(1)\top} x)\sigma'(b^{(1)\top} x')].$$

これは 3.1 節でのニューラルタンジェントカーネル (5) の $M$ についての極限に該当する．次にグラム行列の極限に相当する積分作用素を導入する．確率変数 $(X, Y)$ の確率分布を $\rho$，その $X$ についての周辺分布を $\rho_X$ とする．$K_{\infty,X} \stackrel{\text{def}}{=} k_\infty(X, \cdot)$ とおく．確率測度 $\rho_X$ について二乗可積分関数[4]の成す空間を $L_2(\rho_X)$ とし，$L_2(\rho_X)$ 内の内積 $\langle \cdot, \cdot \rangle_{L_2(\rho_X)}$ を次で定義する．関数 $f, g \in L_2(\rho_X)$ に対し，

$$\langle f, g \rangle_{L_2(\rho_X)} \stackrel{\text{def}}{=} \left(\int f(X)g(X)\mathrm{d}\rho_X\right)^{1/2}.$$

このとき積分作用素 $\Sigma_\infty$ を以下で定義する．関数 $f \in L_2(\rho_X)$ に対して，

$$\Sigma_\infty f \stackrel{\text{def}}{=} \int f(X)K_{\infty,X}\mathrm{d}\rho_X \in L_2(\rho_X).$$

作用素 $\Sigma_\infty$ は自己共役なコンパクト作用素となるのでスペクトル表示することができる．すなわち，$\Sigma_\infty f = \sum_{i=0}^\infty \lambda_i \langle f, \phi_i \rangle_{L_2(\rho_X)} \phi_i$ $(f \in L_2(\rho_X))$ と表される．ここで $\{\lambda_i, \phi_i\}_{i=0}^\infty$ は $\Sigma_\infty$ の $L_2(\rho_X)$ 上の固有値と固有関数であり，固有値は降順に整列しているとする．このとき，積分作用素の冪 $\Sigma_\infty^s$ $(s \in \mathbb{R})$ を $\Sigma_\infty^s f = \sum_{i=0}^\infty \lambda_i^s \langle f, \phi_i \rangle_{L_2(\rho_X)} \phi_i$ で定義する．

以下，平均化確率的勾配降下法のベイズ規則[5] $g_\rho(x) = \mathbb{E}_Y[Y \mid x]$ への収束性を示す定理を紹介する．

**仮定 1.**

- 正数 $C > 0$ が存在し $\|\sigma''\|_\infty \le C$, $\|\sigma'\|_\infty \le 2$, $|\sigma(u)| \le 1 + |u|$ $(\forall u \in \mathbb{R})$ を満たす．
- $\mathrm{supp}(\rho_X) \subset \{x \in \mathbb{R}^d \mid \|x\|_2 \le 1\}$ とし，ラベルは $[-1, 1]$ に値を取るものとする．
- 定数 $r \in [1/2, 1]$ が存在し $\|\Sigma_\infty^{-r} g_\rho\|_{L_2(\rho_X)} < \infty$ を満たす．
- 定数 $\beta > 1$ が存在し $\lambda_i = \Theta(i^{-\beta})$ を満たす．

[3] 固有値の 0 への収束性は $n \to \infty$ のもと $H_{1,\infty}$ が $L_2$ 空間上のトレースが有界な積分作用素に収束することから示される．

[4] 確率測度 $\rho_X$ について測度 0 の集合上でのみ異なる値をとる関数同士は同一視する．

[5] ここでのベイズ規則は二乗損失の期待損失を最小化する可測関数のことであり $g_\rho(x) = \mathbb{E}_Y[Y \mid x]$ と表される．

2 番目の仮定における $\mathrm{supp}(\rho_X)$ は確率測度 $\rho_X$ の台である．$\rho_X$ が $\mathrm{d}x$ について連続な密度関数 $p(x)$ をもつ場合には $\mathrm{supp}(\rho_X)$ は $\{x \in \mathbb{R}^d \mid p(x) > 0\}$ の閉包に他ならない．積分作用素 $\Sigma_\infty$ はカーネル $k_\infty$ による平滑化であるため 3 番目の仮定はベイズ規則 $g_\rho$ に $k_\infty$ による滑らかさを課しているといえる．4 番目の仮定は $k_\infty$ に付随する再生核ヒルベルト空間 $\mathcal{H}_\infty$ の大きさを制御するものである．これらの仮定のもと以下の収束定理が成立する．

**定理 2.** 仮定 1 のもと平均化確率的勾配降下法を実行する．正則化係数を $\lambda = T^{-\beta/(2r\beta+1)}$，固定ステップサイズ $\eta$ は $4(6+\lambda)\eta \leq 1$ を満たすようにとる．このとき，任意の $\epsilon > 0, \delta \in (0,1)$ と $\|\Sigma_\infty\|_{\mathrm{op}} \geq \lambda$ を満たす $T \in \mathbb{Z}_+$ に対して正数 $M_0 \in \mathbb{Z}_+$ が存在し以下が成立する．任意の $M \geq M_0$ に対しパラメータの初期化について $1-\delta$ 以上の確率で

$$\mathbb{E}\left[\|g_{\overline{w}^{(T+1)}} - g_\rho\|^2_{L_2(\rho_X)}\right] \leq \epsilon + \alpha T^{\frac{-2r\beta}{2r\beta+1}}\left(1 + \|\Sigma_\infty^{-r} g_\rho\|^2_{L_2(\rho_X)}\right)$$

を満たす．ここで $\alpha > 0$ はハイパーパラメータに非依存な定数である．

ノード数の下限 $M_0$ を大きくとることで定数 $\epsilon$ はいくらでも小さくできるため，この定理から平均化確率的勾配降下法の収束率は $O(T^{\frac{-2r\beta}{2r\beta+1}})$ であることがわかる．またこれは再生核ヒルベルト空間上の推定問題におけるミニマックス最適 [17] な収束率でありこれ以上改善され得ないものである．確率的勾配降下法の各反復では真の分布からデータを一つサンプリングするので，反復数 $T$ は学習に用いた訓練データ数に他ならない．したがって，収束率 $O(T^{\frac{-2r\beta}{2r\beta+1}})$ における $T$ は 3.1 節における訓練データサイズ $n$ に読み替えることができ，Arora et al. [3] で導出された上界 (6) よりも一般に速いことが確かめられる．

さらに文献 [8] では ReLU を用いた二層ニューラルネットワークのニューラルタンジェントカーネルにより仮定 1 の 3 番目の条件が満たされる場合にも定理を拡張している．具体的には特定のパラメータの初期化分布，入力データ空間 $\mathbb{R}^d$ 上のデータ分布に対し $\beta = 1 + \frac{1}{d-1}$ で 4 番目の条件も成立することを示し，ReLU を平滑化した活性化関数で定まる二層ニューラルネットワークの学習により収束率 $O(T^{\frac{-2rd}{2rd+d-1}})$ が達成されることを証明した．

### 3.3 ニューラルタンジェントカーネルと識別問題

定理 1 でみたように回帰問題においては $n$ に対し過剰なノード数 $\Omega(n^6/\lambda_{1,\infty}^4)$ が大域的収束性に必要であったが，識別問題においては必要なノード数が劇的に減少することが文献 [7] で示された．回帰問題ではニューラルタンジェントカーネルのグラム行列の正定値性が大域的収束性の担保のために重要であったが，識別問題においてはニューラルタンジェントカーネルの陽的表現 $\nabla_w g_w$ を通してデータがマージン付きで識別可能であれば十分である．この後者の条件は前者に比べて大幅に緩く，少ないノード数 $M$ で満たされる．この事実に基づき文献 [7] は少ないノード数のもと，勾配降下法 $w^{(t+1)} = w^{(t)} - \eta \nabla_w \mathcal{L}_n(w^{(t)})$ の大域的収束性の証明と期待識別誤差の上界を与えた．本節ではこの理論を概説する．

二値の識別問題を考えるのでラベル集合を $\{-1,1\}$ とする．損失関数はロジスティック損失 $\ell(z,y) = \log(1+\exp(-yz))$ $(z \in \mathbb{R}, y \in \{-1,1\})$ とする．ここでも二層ニューラルネットワーク $g_w$ の入力層パラメータの最適化を行う．3.2 節同様にパラメータは対称初期化を行うが出力層パラメータについては $\beta \in [0,1)$ に対し $a_r = 1/M^\beta$ $(r \in \{1,2,\ldots,M/2\})$, $a_r = -1/M^\beta$ $(r \in \{M/2+1, M/2+2, \ldots, M\})$ というスケールで初期化する．以下，収束定理のための仮定である．

**仮定 2.**

- $\mathrm{supp}(\rho_X) \subset \{x \in \mathcal{X} \mid \|x\|_2 \leq 1\}$ とする．活性化関数 $\sigma$ は $\mathcal{C}^2$-級で正数 $K_1, K_2 > 0$ が存在し $\|\sigma'\|_\infty \leq K_1$, $\|\sigma''\|_\infty \leq K_2$ を満たす．
- 入力層パラメータの初期化に用いる $\mathbb{R}^d$ 上の確率分布 $\mu_0$ はサブガウシアンとする．すなわち正数 $A, b > 0$ が存在して $\mathbb{P}_{b^{(1)}\sim\mu_0}[\|b^{(1)}\|_2 \geq t] \leq A\exp(-bt^2)$ を満たす．
- 正数 $\gamma > 0$ と可測関数 $v : \mathbb{R}^d \to \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq 1\}$ が存在し次が成立する．任意の $(x,y) \in \mathrm{supp}(\rho) \subset \mathbb{R}^d \times \{-1,1\}$ に対して，

$$y \int \partial_b \sigma(b^{(1)\top} x)^\top v(b^{(1)}) \mathrm{d}\mu_0(b^{(1)}) \geq \gamma. \quad (7)$$

非線形写像 $x \to \partial_b \sigma(b^{(1)\top} x)$ は入力データ空間から無限次元空間への特徴抽出写像であり $M = \infty$ に対応するニューラルタンジェントカーネルの陽表現に他ならない．不等式 (7) はこの特徴抽出写像を通じてデータがマージン $\gamma > 0$ のもと完全識別可能であることを保証するものである．これらの仮定のもと期待識別誤

差 $\mathbb{P}_{(X,Y)\sim\rho}[Yg_{w^{(t)}}(X)\leq 0]$ の収束率が次の定理で示される．ここで勾配降下法はロジスティック損失の経験損失最小化問題に適用されるが，識別問題においてより関心があるのは期待識別誤差の収束性であることに注意されたい．

**定理 3.** 仮定 2 のもと任意の $\epsilon>0$ に対して以下のいずれかの設定[6]で勾配降下法を $T$ 反復実行する．

$$\begin{aligned}&\text{(i) } \beta\in[0,1),\ M=\Omega(\epsilon^{\frac{-1}{1-\beta}}),\ T=\Omega(\epsilon^{-2}),\\ &\qquad \eta=\Theta(\epsilon^{-2}T^{-1}m^{2\beta-1}),\ n=\tilde{\Omega}(\epsilon^{-4}),\\ &\text{(ii) } \beta=0,\ M=\tilde{\Theta}(\epsilon^{-3/2}),\ T=\tilde{\Theta}(\epsilon^{-1}),\\ &\qquad \eta=\Theta(m^{-1}),\ n=\tilde{\Omega}(\epsilon^{-2}).\end{aligned}$$

このとき，勾配降下法により高確率で $T$ 反復以内に $\mathbb{P}_{(X,Y)\sim\rho}[Yg_{w^{(t)}}(X)\leq 0]\leq\epsilon$ が満たされる．

回帰問題では必要ノード数の $n$ についてのオーダーは $\Omega(n^6)$ であったところ，本定理はそれぞれの設定でノード数は $\tilde{\Omega}(n^{1/4})$，$\tilde{\Omega}(n^{3/4})$ で十分であることを示している．したがって，現実的なサイズの二層ニューラルネットワークに対して大域的収束性および汎化保証が与えられたといえる．さらにここで紹介した理論に基づき ReLU 活性化関数の場合では $n$ の対数次数程度まで中間ノード数を減少可能なことが文献 [6] で示された．

### 3.4 平均場理論

本節では二層ニューラルネットワークに対する勾配降下法の平均場理論 [2, 4] について概要のみ述べる．ニューラルタンジェントカーネルの場合と異なり出力層パラメータは $a_r=1/M$ で固定することにする．すなわち $g_w(x)=\frac{1}{M}\sum_{r=1}^{M}\sigma(b_r^\top x)$ とする．このとき，適当な仮定のもと極限 $M\to\infty$ をとるとモデル $g_w$ は $g_{\mu^{(1)}}(x)=\mathbb{E}_{b^{(1)}\sim\mu^{(1)}}[\sigma(b^{(1)\top}x)]$ に概収束する．ここで $\mu^{(1)}$ は入力層パラメータを初期化するための確率分布とする．勾配降下法では初期パラメータ $w^{(1)}=\{b_r^{(1)}\}_{r=1}^M$ を $b_r^{(2)}=b_r^{(1)}-\eta\partial_{b_r}\mathcal{L}_n(w^{(1)})$ に更新するが，この更新を $\mu^{(1)}$ に従う粒子群 $w^{(1)}=\{b_r^{(1)}\}_{r=1}^M$ を $w^{(2)}=\{b_r^{(2)}\}_{r=1}^M$ に変形しているとみなそう．すると粒子群 $w^{(2)}$ は確率分布 $\mu^{(1)}$ からある規則で更新し得られた確率分布 $\mu^{(2)}$ に従っていると解釈できる．したがって，勾配降下法は極限 $M\to\infty$ のもとではパラメータ空間上の確率分布の最適化を行っていると考えられる．以下ではこの観点に基づき確率測度の空間での勾配降下法を導出する．そして実のところそのような確率測度の勾配降下法の粒子を用いた離散化が二層ニューラルネットワークの勾配降下法に他ならないのである．

最適化対象の変数はパラメータ空間 $\mathbb{R}^d$ 上の確率測度 $\mu$ であり，最適化問題は $\min_\mu\mathcal{L}_n(\mu)$ で表される．確率測度 $\mu$ を輸送写像 $\psi:\mathrm{supp}(\mu)\to\mathbb{R}^d$ を用いて $\psi_\sharp\mu$ に更新することを考える．ここで，$\psi_\sharp\mu$ は確率測度の押し出しである．特に $\psi$ は $\mathrm{supp}(\mu)$ 上の滑らかなベクトル場 $\xi:\mathrm{supp}(\mu)\to\mathbb{R}^d$ による摂動 $\psi=id+\xi$ に限定する．このとき，この操作を $t$ 反復し得られる確率測度は $\mu^{(t+1)}=(id+\xi_t)_\sharp\mu^{(t)}=((id+\xi_t)\circ\cdots\circ(id+\xi_1))_\sharp\mu^{(1)}$ という形をとる．最適化手法を構築するにあたり考えるべきは，各反復においての摂動 $\xi_j$ の選び方である．確率測度空間上の勾配降下法の導出を考えると，確率測度 $\mu$ における摂動を $\mathcal{L}_n((id+\xi)_\sharp\mu)$ の $\xi$ についてのフレシェ微分 $\nabla_\xi\mathcal{L}_n((id+\xi)_\sharp\mu)|_{\xi=0}$ とすることは自然である．この場合，適当な仮定のもとで $L_2(\mu)$ 内積による以下の等式が成立する．

$$\begin{aligned}&\mathcal{L}_n((id+\xi)_\sharp\mu)=\mathcal{L}_n(\mu)\\ &+\langle\nabla_\zeta\mathcal{L}_n((id+\zeta)_\sharp\mu)|_{\zeta=0},\xi\rangle_{L_2(\mu)}+O(\|\xi\|^2_{L^2(\mu)}).\end{aligned}$$

これは，摂動についてのテイラーの公式に他ならず，勾配降下法の導出に有用である．実際，$\xi=-\nabla_\zeta\mathcal{L}_n((id+\zeta)_\sharp\mu)|_{\zeta=0}$ が $\mu$ における降下方向であることが直ちに従う．したがって，フレシェ微分あるいは，その推定量を用いた降下法により確率測度についての最適化が実行される．そして実は二層ニューラルネットワークの勾配降下法は $\mu^{(t)}$ に従う粒子群 $w^{(t)}$ を輸送により $\mu^{(t+1)}$ に従う粒子群 $w^{(t+1)}=(id+\xi_t)(w^{(t)})$ へと更新していることに他ならないのである [2].

このようにパラメータについての勾配降下法を確率測度の勾配降下法として捉えると損失関数 $\ell(z,y)$ の $z$ についての凸性を活用できるようになるのである．この観点から極限 $M\to\infty$ における二層ニューラルネットワークの確率測度空間での局所的最適解への収束性が文献 [12] で与えられ，さらに大域的収束性が文献 [4] で与えられた．また，勾配降下法による確率測度の列 $\{\mu^{(t)}\}_{t=1}^\infty$ は確率測度空間におけるワッサースタイン勾配流の離散化に他ならないことも文献 [2] で示されている．

## 4. おわりに

ニューラルネットワークの学習は非凸最適化問題に帰着されるため大域的収束性の証明は困難であったが，特定の条件下ではニューラルタンジェントカーネルお

[6] ランダウ記号 $\tilde{\Omega}$，$\tilde{\Theta}$ は対数項も含んでいる．

よび平均場理論の登場により解決されつつあることを概説した．しかしながらニューラルネットワークを深層にすることの利点の解明はいまだ十分にはなされていない．現代の深層学習の大きな成功を説明するにはさらにこれらの理論を深化させる必要があり今後の発展が期待されるところである．

**謝辞**　本稿で紹介した研究の一部は，JSPS 科研費 JP19K20337 および JST さきがけ JPMJPR1928 の支援を受けたものです．本稿の執筆機会と有益な助言をくださった奥野貴之先生，高野祐一先生に感謝いたします．最後に，共同研究者である鈴木大慈先生に感謝いたします．

## 参考文献

[1] A. Jacot, F. Gabriel and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," In *Advances in Neural Information Processing Systems*, pp. 8571–8580, 2018.

[2] A. Nitanda and T. Suzuki, "Stochastic particle gradient descent for infinite ensembles," *arXiv preprint*, arXiv:1712.05438, 2017.

[3] S. Arora, S. S. Du, W. Hu, Z. Li and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," In *Proceedings of International Conference on Machine Learning*, **36**, pp. 322–332, 2019.

[4] L. Chizat and F. Bach, "On the global convergence of gradient descent for over-parameterized models using optimal transport," In *Advances in Neural Information Processing Systems*, pp. 3040–3050, 2018.

[5] S. S. Du, X. Zhai, B. Poczos and A. Singh, "Gradient descent provably optimizes overparameterized neural networks," *International Conference on Learning Representations*, 2019.

[6] Z. Ji and M. Telgarsky, "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks," *International Conference on Learning Representations*, 2020.

[7] A. Nitanda, G. Chinot and T. Suzuki, "Gradient descent can learn less over-parameterized two-layer neural networks on classification problems," *arXiv preprint*, arXiv:1905.09870, 2019.

[8] A. Nitanda and T. Suzuki, "Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime," *arXiv preprint*, arXiv:2006.12297, 2020.

[9] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, **2**, pp. 499–526, 2002.

[10] S. Mukherjee, R. Rifkin and T. Poggio, "Regression and classification with regularization," *Nonlinear Estimation and Classification*, pp. 111–128, 2003.

[11] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.

[12] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, 2008.

[13] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, **22**, pp. 400–407, 1951.

[14] L. Bottou, F. E. Curtis and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, **60**, pp. 223–311, 2018.

[15] F. Bach and E. Moulines, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

[16] X. Wu, S. S. Du and R. Ward, "Global convergence of adaptive gradient methods for an over-parameterized neural network," *arXiv preprint*, arXiv:1902.07111, 2019.

[17] A. Caponnetto and E. D. Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, **7**, pp. 331–368, 2007.