論文紹介：Gradient Descent Provably Optimizes Over-parameterized Neural Networks

宮崎優

工学系研究科物理工学専攻

November 28, 2022

原論文

スライド

発表の流れ

定理の紹介

活性化関数をReLU $\sigma(z) = z\mathbb{I}\{z \geq 0\}$とする2層のニューラルネットワークを考える。

$$f(\mathbf{W}, \mathbf{a}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma \left(\mathbf{w}_r^\top \mathbf{x} \right)$$

ここでデータセット$\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$についての損失関数

$$L(\mathbf{W}, \mathbf{a}) = \sum_{i=1}^{n} \frac{1}{2} \left(f\left(\mathbf{W}, \mathbf{a}, \mathbf{x}_i\right) - y_i \right)^2 \tag{1}$$

について、第1層パラメータの勾配降下(GD)

$$\mathbf{W}(k+1) = \mathbf{W}(k) - \eta \frac{\partial L(\mathbf{W}(k), \mathbf{a})}{\partial \mathbf{W}(k)} \tag{2}$$

を考えると、mが十分大きく、$\mathbf{W}(0)$をランダム初期化で与えるなどの条件において線形レートで0損失を達成する。

証明の概略

前提：パラメータ$\mathbf{W}$空間のダイナミクスについて調べるのは損失関数の非凸性や非平滑性から難しい⇒予測空間のダイナミクスに着目

$\boxed{\text{グラム行列}\mathbf{H}(k)\text{の最小固有値が解析で重要な役割を果たす}}$

証明の概略

前提：パラメータ$\mathbf{W}$空間のダイナミクスについて調べるのは損失関数の非凸性や非平滑性から難しい⇒予測空間のダイナミクスに着目

$$\boxed{\text{グラム行列}\mathbf{H}(k)\text{の最小固有値が解析で重要な役割を果たす}}$$

1. 初期条件$k=0$においてはグラム行列$\mathbf{H}(0)$の最小固有値が、期待値を取ったグラム行列の最小固有値で下から抑えられる。

証明の概略

前提：パラメータ$\mathbf{W}$空間のダイナミクスについて調べるのは損失関数の非凸性や非平滑性から難しい⇒予測空間のダイナミクスに着目

グラム行列$\mathbf{H}(k)$の最小固有値が解析で重要な役割を果たす

1. 初期条件$k = 0$においてはグラム行列$\mathbf{H}(0)$の最小固有値が、期待値を取ったグラム行列の最小固有値で下から抑えられる。
2. このグラム行列は活性化のパターン($\mathbb{I}\left\{\mathbf{w}_r^\top \mathbf{x}_i \geq 0\right\}$)にのみ依存し、イテレーションを進めても殆どの活性化パターンが変わらない場合、グラム行列(とその最小固有値)は初期値に近いことを示す。

証明の概略

前提：パラメータ$\mathbf{W}$空間のダイナミクスについて調べるのは損失関数の非凸性や非平滑性から難しい⇒予測空間のダイナミクスに着目

グラム行列$\mathbf{H}(k)$の最小固有値が解析で重要な役割を果たす

1. 初期条件$k = 0$においてはグラム行列$\mathbf{H}(0)$の最小固有値が、期待値を取ったグラム行列の最小固有値で下から抑えられる。
2. このグラム行列は活性化のパターン($\mathbb{I}\left\{\mathbf{w}_r^\top \mathbf{x}_i \geq 0\right\}$)にのみ依存し、イテレーションを進めても殆どの活性化パターンが変わらない場合、グラム行列(とその最小固有値)は初期値に近いことを示す。
3. 損失関数のダイナミクスはグラム行列のスペクトルに支配され、グラム行列の最小固有値が下界を持つ限り、線形収束となる。

これらの結果から経験損失最小化に対するReLU活性化NNのグローバルな定量的収束結果を示す。

過剰パラメータ、ランダム初期化の条件が本質的

ノーテーション

- $[n] = \{1, 2, \ldots, n\}$
- 集合Sに対し、その一様分布を$\mathrm{unif}\{S\}$と書く。
- $\mathbb{I}\{A\}$を事象Aの指示関数とする。
- 標準正規分布を$N(\mathbf{0}, \mathbf{I})$とする。
- ベクトルのユークリッドノルムを$\|\cdot\|_2$とし、行列のフロベニウスノルムを$\|\cdot\|_F$とする。
- 行列$\mathbf{A}$が半正定値のとき、$\lambda_{\min}(\mathbf{A})$のその最小固有値とする。
- 2つのベクトルのユークリッド内積を$\langle\cdot,\cdot\rangle$とする。
- Oはランダウの記号で、Ωはその不等式が逆の場合。

連続時間の解析

まずはGDのステップ幅を無限に小さくした連続極限で考えてみる。

GDの連続時間極限

$$\frac{d\mathbf{w}_r(t)}{dt} = -\frac{\partial L(\mathbf{W}(t), \mathbf{a})}{\partial \mathbf{w}_r(t)} \tag{3}$$

ここで入力$\mathbf{x_i}$と時刻tにおける予測$u_i(t) = f(\mathbf{W}(t), \mathbf{a}, \mathbf{x}_i)$とし、
$\mathbf{u}(t) = (u_1(t), \ldots, u_n(t)) \in \mathbb{R}^n$を時刻$t$での予測ベクトルとするときに
次の重要な定理が成り立つ。

連続時間の定理

Theorem 2.1 (Convergence Rate of Gradient Flow)

すべての$i \in [n]$について、

- $\|\mathbf{x}_i\|_2 = 1$
- ある定数Cに対して$|y_i| \leq C$
- $\mathbf{H}_{ij}^{\infty} = \mathbb{E}_{\mathbf{w} \sim N(\mathbf{0},\mathbf{I})}\left[\mathbf{x}_i^\top \mathbf{x}_j \mathbb{I}\left\{\mathbf{w}^\top \mathbf{x}_i \geq 0, \mathbf{w}^\top \mathbf{x}_j \geq 0\right\}\right]$とした際に、$\lambda_{\min}(\mathbf{H}^{\infty}) \triangleq \lambda_0 > 0$[a]

が満たされるとする。
ここで$\mathbf{w}_r \sim N(\mathbf{0}, \mathbf{I}), a_r \sim \mathrm{unif}\left[\{-1, 1\}\right], r \in [m]$のように初期化し、隠れ層のノード数を$m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$とするとランダム初期化に対して少なくとも$1-\delta$の確率で以下を満たす。

$$\|\mathbf{u}(t) - \mathbf{y}\|_2^2 \leq \exp(-\lambda_0 t)\,\|\mathbf{u}(0) - \mathbf{y}\|_2^2 \tag{4}$$

[a] 入力が並行でなければ最小固有値は正[Xie et al., 2017, Du et al., 2019]。

Th. 2.1についての補足

- 仮定の中で最も本質的なのは、行列$\mathbf{H}^{\infty}$が正定値であること。
- 隠れ層のノード数は$m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$であり、サンプル数$n$と最小固有値$\lambda_0$に依存する。この過剰パラメータ条件も大域最小解への到達を保証するのに本質的な役割を果たす。
- この依存性については改善される可能性がある。
- $\|\mathbf{u}(t) - \mathbf{y}\|_2^2$が指数関数的に減衰するため、収束レートは線形であるが、このレートはλ_0に依存するが、隠れ層のノード数mには依存しない。

Th. 2.1の証明

まずはそれぞれの予測$u_i(t)$についてのダイナミクスを計算する。

$$\begin{aligned}\frac{d}{dt}u_i(t) &= \sum_{r=1}^{m}\left\langle \frac{\partial f\left(\mathbf{W}(t),\mathbf{a},\mathbf{x}_i\right)}{\partial \mathbf{w}_r(t)}, \frac{d\mathbf{w}_r(t)}{dt}\right\rangle \\ &= \sum_{j=1}^{n}(y_j-u_j)\left\langle \sum_{r=1}^{m}\frac{\partial f\left(\mathbf{W}(t),\mathbf{a},\mathbf{x}_i\right)}{\partial \mathbf{w}_r(t)}, \frac{\partial f\left(\mathbf{W}(t),\mathbf{a},\mathbf{x}_j\right)}{\partial \mathbf{w}_r(t)}\right\rangle \\ &\triangleq \sum_{j=1}^{n}(y_j-u_j)\,\mathbf{H}_{ij}(t)\end{aligned} \tag{5}$$

ここで$\mathbf{H}(t)$は時間依存する対称行列

$$\mathbf{H}_{ij}(t) = \frac{1}{m}\mathbf{x}_i^\top\mathbf{x}_j\sum_{r=1}^{m}\mathbb{I}\left\{\mathbf{x}_i^\top\mathbf{w}_r(t)\geq 0, \mathbf{x}_j^\top\mathbf{w}_r(t)\geq 0\right\}. \tag{6}$$

である。

Th. 2.1の証明

ここで以下の補題からmが大きいとき、$\mathbf{H}(0)$の最小固有値が高い確率で下から抑えられる。

Lemma 2.2

$m = \Omega\left(\frac{n^2}{\lambda_0^2}\log\left(\frac{n}{\delta}\right)\right)$のとき、少なくとも$1-\delta$の確率で$\|\mathbf{H}(0)-\mathbf{H}^{\infty}\|_2 \le \frac{\lambda_0}{4}$かつ $\lambda_{\min}(\mathbf{H}(0)) \ge \frac{3}{4}\lambda_0$を満たす。

全ての固定された(i,j)ペアについて、$\mathbf{H}_{ij}(0)$を独立ランダム変数の平均とする。Hoeffding不等式を用いると確率$1-\delta'$で

$$\left|\mathbf{H}_{ij}(0)-\mathbf{H}_{ij}^{\infty}\right| \le \frac{2\sqrt{\log(1/\delta')}}{\sqrt{m}}$$

$\delta' = n^2\delta$とし、(i,j)についてunion boundを取ると、全ての(i,j)ペアについて少なくとも確率$1-\delta$で以下を満たす。

$$\left|\mathbf{H}_{ij}(0)-\mathbf{H}_{ij}^{\infty}\right| \le \frac{4\sqrt{\log(n/\delta)}}{\sqrt{m}}$$

故に

$$\begin{aligned}\|\mathbf{H}(0) - \mathbf{H}^{\infty}\|_2^2 &\leq \|\mathbf{H}(0) - \mathbf{H}^{\infty}\|_F^2 \\ &\leq \sum_{i,j} \left|\mathbf{H}_{ij}(0) - \mathbf{H}_{ij}^{\infty}\right|^2 \\ &\leq \frac{16n^2 \log(n/\delta)}{m}\end{aligned}$$

であるから、 $m = \Omega\left(\frac{n^2 \log(n/\delta)}{\lambda_0^2}\right)$ のとき $\|\mathbf{H}(0) - \mathbf{H}^{\infty}\|_2 \leq \frac{\lambda_0}{4}$ であり、$\lambda_{\min}(\mathbf{H}(0)) \geq \lambda_{\min}(\mathbf{H}^{\infty}) - \frac{\lambda_0}{4} = \frac{3\lambda_0}{4}$ より補題は示される。 □

Th. 2.1の証明

更に以下の補題を用いることで$\mathbf{W}(t)$が$\mathbf{W}(0)$に近いとき、$\mathbf{H}(t)$が$\mathbf{H}(0)$に近く、最小固有値の下界を与えることができる。

Lemma 2.3

小さい正の定数cについて与えられた時刻t及び全ての $r \in [m]$ で $\|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2 \leq \frac{c\delta\lambda_0}{n^2} \triangleq R$を仮定したとき、初期化に対して少なくとも$1-\delta$の確率で$|H_{ij}(t) - H_{ij}(0)| < \frac{\lambda_0}{4}$かつ$\lambda_{\min}(\mathbf{H}(t)) > \frac{\lambda_0}{2}$を満たす。

事象をA_{ir}を以下のように定義する。

$$A_{ir} = \left\{ \exists \mathbf{w} : \|\mathbf{w} - \mathbf{w}_r(0)\| \leq R, \mathbb{I}\left\{ \mathbf{x}_i^\top \mathbf{w}_r(0) \geq 0 \right\} \neq \mathbb{I}\left\{ \mathbf{x}_i^\top \mathbf{w} \geq 0 \right\} \right\}$$

これは$\left|\mathbf{w}_r(0)^\top \mathbf{x}_0\right| < R$と同値である。 anti-concentration inequality of Gaussianから$P(A_{ir}) = P_{z \sim N(0,1)}(|z| < R) \leq \frac{2R}{\sqrt{2\pi}}$.であるから、

$$
\begin{aligned}
&\mathbb{E}\left[\left|H_{ij}(t) - H_{ij}(0)\right|\right] \\
=&\mathbb{E}\left[\frac{1}{m}\left|\mathbf{x}_i^\top \mathbf{x}_j \sum_{r=1}^{m}\left(\mathbb{I}\left\{\mathbf{w}_r(0)^\top \mathbf{x}_i \geq 0, \mathbf{w}_r(0)^\top \mathbf{x}_j \geq 0\right\}\right.\right.\right. \\
&\left.\left.\left.\quad -\mathbb{I}\left\{\mathbf{w}_r(t)^\top \mathbf{x}_i \geq 0, \mathbf{w}_r(t)^\top \mathbf{x}_j \geq 0\right\}\right)\right|\right] \\
\leq&\frac{1}{m}\sum_{r=1}^{m}\mathbb{E}\left[\mathbb{I}\left\{A_{ir} \cup A_{jr}\right\}\right] \leq \frac{4R}{\sqrt{2\pi}}
\end{aligned}
$$

(i,j)について足し上げると$\mathbb{E}\left[\sum_{(i,j)=(1,1)}^{(n,n)}\left|H_{ij}(t) - H_{ij}(0)\right|\right] \leq \frac{4n^2R}{\sqrt{2\pi}}$であり、Markov inequalityを用いると少なくとも$1-\delta$の確率で$\sum_{(i,j)=(1,1)}^{(n,n)}\left|H_{ij}(t) - H_{ij}(0)\right| \leq \frac{4n^2R}{\sqrt{2\pi}\delta}$である。行列の摂動論を用いると

$$
\|\mathbf{H}(t) - \mathbf{H}(0)\|_2 \leq \|\mathbf{H}(t) - \mathbf{H}(0)\|_F \leq \sum_{(i,j)=(1,1)}^{(n,n)}\left|H_{ij}(t) - H_{ij}(0)\right| \leq \frac{4n^2R}{\sqrt{2\pi}\delta}
$$

すなわち$\lambda_{\min}(\mathbf{H}(t)) \geq \lambda_{\min}(\mathbf{H}(0)) - \frac{4n^2R}{\sqrt{2\pi}\delta} \geq \frac{\lambda_0}{2}$ □

Th. 2.1の証明

次の補題は$\mathbf{H}$の最小固有値が下から抑えられるとき、線形収束レートで経験損失が0に収束すること、全ての$r \in [m]$について$\mathbf{w}_r(t)$が初期化に近いことを主張する。

Lemma 2.4

$0 \le s \le t$及び$\lambda_{\min}(\mathbf{H}(s)) \ge \frac{\lambda_0}{2}$を仮定する。このとき以下が成り立つ。

- $\|\mathbf{y} - \mathbf{u}(t)\|_2^2 \le \exp(-\lambda_0 t)\|\mathbf{y} - \mathbf{u}(0)\|_2^2$
- 任意の$r \in [m]$について$\|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2 \le \frac{2\sqrt{n}\|\mathbf{y}-\mathbf{u}(0)\|_2}{\sqrt{m}\lambda_0} \triangleq R'$

予測のダイナミクスは$\frac{d}{dt}\mathbf{u}(t) = \mathbf{H}(\mathbf{y} - \mathbf{u}(t))$であるから、損失のダイナミクスは以下のようになる。

$$\begin{aligned}\frac{d}{dt}\|\mathbf{y} - \mathbf{u}(t)\|_2^2 &= -2(\mathbf{y} - \mathbf{u}(t))^\top \mathbf{H}(t)(\mathbf{y} - \mathbf{u}(t)) \\ &\le -\lambda_0\|\mathbf{y} - \mathbf{u}(t)\|_2^2\end{aligned}$$

$\frac{d}{dt}\left(\exp(\lambda_0 t)\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2\right) \leq 0$であることから$\exp(\lambda_0 t)\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2$は減少関数であり、損失関数を以下のように抑える事ができる。

$$\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 \leq \exp(-\lambda_0 t)\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2^2$$

$0 \leq s \leq t$に注意すると

$$\begin{aligned}
\left\|\frac{d}{ds}\mathbf{w}_r(s)\right\|_2 &= \left\|\sum_{i=1}^{n}(y_i-u_i)\frac{1}{\sqrt{m}}a_r\mathbf{x}_i\mathbb{I}\left\{\mathbf{w}_r(s)^\top\mathbf{x}_i \geq 0\right\}\right\|_2 \\
&\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\left|y_i-u_i(s)\right| \leq \frac{\sqrt{n}}{\sqrt{m}}\left\|\mathbf{y}-\mathbf{u}(s)\right\|_2 \\
&\leq \frac{\sqrt{n}}{\sqrt{m}}\exp\left(-\lambda_0 s\right)\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2
\end{aligned}$$

$$\left\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\right\|_2 \leq \int_0^t\left\|\frac{d}{ds}\mathbf{w}_r(s)\right\|_2 ds \leq \frac{\sqrt{n}\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2}{\sqrt{m}\lambda_0} \qquad \square$$

Th. 2.1の証明

次の補題は$R' < R$すなわち$m = \Omega\left(\frac{n^5\|\mathbf{y}-\mathbf{u}(0)\|_2^2}{\lambda_0^4}\right)$のときにLemma2.3及び2.4の仮定が全ての$t \geq 0$で成り立つことを主張する。

Lemma 2.5

$R' < R$のとき、全ての$t \geq 0$で以下が成り立つ。

- $\lambda_{\min}(\mathbf{H}(t)) \geq \frac{1}{2}\lambda_0$
- 全ての$r \in [m]$で$\|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2 \leq R'$
- $\|\mathbf{y} - \mathbf{u}(t)\|_2^2 \leq \exp(-\lambda_0 t)\,\|\mathbf{y} - \mathbf{u}(0)\|_2^2$

補題の結論が時刻tで成り立たないと仮定する。
$\|\mathbf{w}_r(t) - \mathbf{w}_r(0)\| \geq R'$もしくは$\|\mathbf{y} - \mathbf{u}(t)\|_2^2 > \exp(-\lambda_0 t)\,\|\mathbf{y} - \mathbf{u}(0)\|_2^2$となる$r \in [m]$が存在するならば、Lemma2.3より$\lambda_{\min}(\mathbf{H}(s)) < \frac{1}{2}\lambda_0$を満たす$s \leq t$が存在する。

Lemma 2.2から以下で定義されるt_0が存在することがわかる。

$$t_0 = \inf\left\{t > 0 : \max_{r\in[m]} \|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2^2 \geq R\right\}$$

故にt_0で$\|\mathbf{w}_r(t_0) - \mathbf{w}_r(0)\|_2^2 = R$を満たす$r \in [m]$が存在する。
Lemma2.2から$t' < t_0$で$\mathbf{H}(t_0) \geq \frac{1}{2}\lambda_0$である。しかし、Lemma2.3から、$\|\mathbf{w}_r(t_0) - \mathbf{w}_r(0)\|_2 < R' < R$であり、これは矛盾している。また、$\lambda_{\min}(\mathbf{H}(t)) < \frac{1}{2}\lambda_0$を満たす時刻$t$では

$$t_0 = \inf\left\{t \geq 0 : \max_{r\in[m]} \|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2^2 \geq R\right\}$$

が存在し、全く同じ議論から矛盾が示せる。 $\square$

Th. 2.1の証明

$R' < R$すなわち$m = \Omega\left(\frac{n^5\|\mathbf{y}-\mathbf{u}(0)\|_2^2}{\lambda_0^4}\right)$のときに、以下の評価が成り立つ。

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{y}-\mathbf{u}(0)\|_2^2\right] &= \sum_{i=1}^{n}(y_i^2 + y_i\mathbb{E}\left[f(\mathbf{W}(0),\mathbf{a},\mathbf{x}_i)\right] + \mathbb{E}\left[f(\mathbf{W}(0),\mathbf{a},\mathbf{x}_i)^2\right]) \\ &= \sum_{i=1}^{n}(y_i^2+1) = O(n)\end{aligned}$$

Markov不等式から少なくとも$1-\delta$の確率で$\|\mathbf{y}-\mathbf{u}(0)\|_2^2 = O\left(\frac{n}{\delta}\right)$であり、このバウンドに代入すればTh. 2.1は示される。

両方の層を学習させる場合

両方の層のパラメータを学習させる場合のダイナミクスは以下の連立方程式で記述される。

$$\frac{d\mathbf{w}_r(t)}{dt} = -\frac{\partial L(\mathbf{W}(t), \mathbf{a}(t))}{\partial \mathbf{w}_r(t)}, \frac{da_r(t)}{dt} = -\frac{\partial L(\mathbf{W}(t), \mathbf{a}(t))}{\partial a_r(t)}$$

この場合であっても、線形収束レートで損失は下降する。

Theorem 2.6 (Convergence Rate of Gradient Flow for Training Both Layers)

*Th. 2.1*と同様の仮定のもと、$m = \Omega\left(\frac{n^6 \log(m/\delta)}{\lambda_0^4 \delta^3}\right), \mathbf{w}_r(0) \sim N(\mathbf{0}, \mathbf{I}), a_r(0) \sim \mathrm{unif}\left[\{-1, 1\}\right], r \in [m]$のとき、少なくとも$1 - \delta$の確率で$\|\mathbf{u}(t) - \mathbf{y}\|_2^2 \leq \exp\left(-\lambda_0 t\right) \|\mathbf{u}(0) - \mathbf{y}\|_2^2$が成り立つ。

(証明略)

離散時間の場合の収束レート

次に離散時間の勾配降下(2)についての定理について述べる。

Theorem 3.1 (Convergence Rate of Gradient Descent)

*Th. 2.1*と同じ仮定で、隠れ層のノード数を$m = \Omega\left(\frac{n^6}{\lambda_0^4\delta^3}\right)$、初期化を$\mathbf{w}_r \sim N(\mathbf{0}, \mathbf{I}), a_r \sim \mathrm{unif}\left[\{-1, 1\}\right], r \in [m]$、ステップサイズを$\eta = O\left(\frac{\lambda_0}{n^2}\right)$とすると、ランダム初期化に対して少なくとも$1 - \delta$の確率で以下が満たされる。

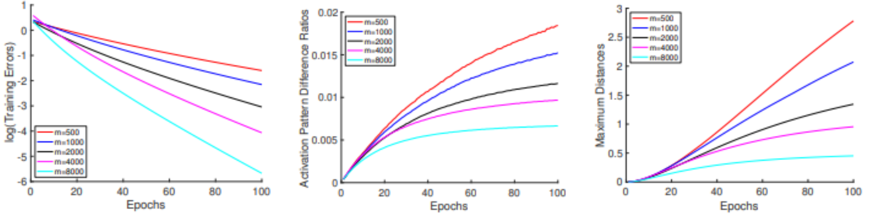
$$\|\mathbf{u}(k) - \mathbf{y}\|_2^2 \leq \left(1 - \frac{\eta\lambda_0}{2}\right)^k \|\mathbf{u}(0) - \mathbf{y}\|_2^2$$

(証明略)

数値実験

- 100エポックのGDをステップサイズ固定
- $n = 1000$のデータを$d = 1000$次元球から一様に生成し、ラベルは1次元の標準正規分布から生成
- 幅mを変化させて、各エポックkに対してa)収束レート、b)パターン変化$\frac{\sum_{i=1}^{m}\sum_{r=1}^{m}\mathbb{I}\left\{\mathrm{sign}\left(\mathbf{w}_r(0)^\top \mathbf{x}_i\right)\neq\mathrm{sign}\left(\mathbf{w}_r(k)^\top \mathbf{x}_i\right)\right\}}{mn}$、c)初期値との距離$\max_{r\in[m]}\left\|\mathbf{w}_r(k)-\mathbf{w}_r(0)\right\|_2$の最大値、を評価

数値実験



(a) Convergence rates. (b) Percentiles of pattern changes. (c) Maximum distances from initialization.

Figure 1: Results on synthetic data.

Fig. 1(a) mが大きくなるほどよい収束性⇒$\mathbf{H}(t)$が安定になり、最小固有値がより大きくなった可能性

Fig. 1(b) mが大きくなるほどパターン変化は少⇒Lemma 2.3の効果

Fig. 1(c) mが大きくなるほど距離の最大値は小⇒Lemma 2.4の効果

結果から得られたこと・考えたこと

- ReLUを活性化関数とするニューラルネットワークの損失関数は非凸かつ非平滑であるにも関わらず、単純な勾配降下によって大域最適解に非常に早いレートで収束

$$\|\mathbf{u}(t) - \mathbf{y}\|_2^2 \leq \exp(-\lambda_0 t) \|\mathbf{u}(0) - \mathbf{y}\|_2^2$$

- ダイナミクスを特徴づけるグラム行列の最小固有値の値が本質的
  - 力学系的解析の集中不等式などの確率的解析という印象
  - 特殊な問題設定に特化したような道具は使っていないように見えるので、拡張はいろいろとありそう
  - 個人的にはNeural Tangent Kernelとの関係が気になる

- 確率的要素は主に初期化で入っており、ダイナミクス自体は決定論的だが、確率的勾配法ではどう変わってくるのか？

References

Simon S. Du, Xiyu Zhai, Barnabas Poczos, Aarti Singh (2019)
Gradient Descent Provably Optimizes Overparameterized Neural Networks
International Conference on Learning Representations.

Bo Xie, Yingyu Liang, Le Song (2017)
Diverse Neural Network Learns True Target Functions
International Conference on Artificial Intelligence and Statistics.

公式集

Theorem 6.1 (Markov Inequality)

任意の正数$t > 0$と非負の確率変数$Z \geq 0$について

$$P[Z \geq t] \leq \frac{\mathbb{E}[Z]}{t}$$

Theorem 6.2 (Hoeffding Inequality)

確率変数$Z_1, Z_2, \ldots, Z_n$を独立な確率変数とし、全ての$i \in [n]$で$Z_i \in [a_i, b_i]$とし、$S_n = Z_1 + \ldots + Z_n$とする。任意の正数$t > 0$に対し

$$P[S_n - \mathbb{E}[S_n] \geq nt] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

公式集

Theorem 6.3 (Union Bound)

事象$A_1, A_2, \ldots$について

$$P\left[\bigcup_i A_i\right] \leq \sum_i \Pr[A_i]$$

Theorem 6.4 (Anti-concentration Inequality of Gaussian)

任意の正数$t > 0$に対して$P_{Z \sim N(0,1)}(Z < t) \leq \frac{t}{\sqrt{2\pi}}$