Neural Tangent Kernel 概説

作成日 2020.2.24
最終更新 2024.4.18
渡部　海斗[*1]

1　本稿作成の目的

Neural Tangent Kernel (NTK)(Jacot et al., 2018) については既に多くのわかり易い解説 (Rajat, 2019; 甘利俊一, 2019; 鈴木大慈, 2019; 大阪大学医学部 Python 会, 2020) があります．本稿は自分自身の理解を深める，またその過程の中で NTK について様々な方に共有することができればと思い，作成しております．なるべく数学的な厳密性を保った上で作成を進めていくつもりではありますが，何か間違い等ありましたらissuesにあげていただく，あるいは Twitter(@kwignb) までご一報いただけますと非常に助かります．

2　記法とニューラルネットワークの定義式

$\mathcal{D} \subseteq \mathbb{R}^{n_0} \times \mathbb{R}^k$ をデータセットの集合とし，$\mathcal{X} = \{\boldsymbol{x} \mid (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}\}$ と $\mathcal{Y} = \{\boldsymbol{y} \mid (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}\}$ をそれぞれ入力データとラベルとする．中間層が L 層，各層の幅を n_l $(l = 1, \ldots, L)$ とし，出力層の幅 (クラス数) を $n_{L+1} = k$ とする．入力 $\boldsymbol{x} \in \mathbb{R}^{n_0}$ に対し，$h^l(\boldsymbol{x}), x^l(\boldsymbol{x}) \in \mathbb{R}^{n_l}$ を pre-activation function，post-activation function とする．このとき，ニューラルネットワーク (neural network; NN) の再帰関係の定義式を，

$$\begin{cases} h^{l+1} = x^l \boldsymbol{W}^{l+1} + \boldsymbol{b}^{l+1} \\ x^{l+1} = \varphi(h^{l+1}) \end{cases} \quad \text{and} \quad \begin{cases} W_{ij}^l = \dfrac{\sigma_\omega}{\sqrt{n_l}} \omega_{ij}^l \\ b_j^l = \sigma_b \beta_j^l \end{cases} \tag{2.1}$$

とする．ここで，$\varphi : \mathbb{R} \to \mathbb{R}$ はリプシッツ連続 (Lipschitz continuous) かつ 2 回連続微分可能な要素毎の活性化関数 (element-wise activation function) であり，$\boldsymbol{W}^{l+1} \in \mathbb{R}^{n_l \times n_{l+1}}$ と $\boldsymbol{b}^{l+1} \in \mathbb{R}^{n_{l+1}}$ は重み行列とバイアスベクトルを表し，ω_{ij}^l, β_j^l は標準ガウス分布 $\mathcal{N}(0, 1)$ に従って初期化される．$\sigma_\omega^2, \sigma_b^2$ は重みとバイアスの分散である．この定義は標準的な NN の再帰関係の式

$$\begin{cases} h^{l+1} = x^l \boldsymbol{W}^{l+1} + \boldsymbol{b}^{l+1} \\ x^{l+1} = \varphi(h^{l+1}) \\ W_{ij}^l, b_j^l \sim \mathcal{U}\left(-\sqrt{k}, \sqrt{k}\right), \quad k = \dfrac{6}{n_l + n_{l-1}} \end{cases} \tag{2.2}$$

とは異なり，NTK parametrization と呼ばれる (ここでは重みとバイアスの初期化に Glorot の一様分布 (Glorot and Bengio, 2010) を用いている)．NTK parametrization は NN の順伝播時のダイナミクスのみでなく，誤差逆伝播 (backpropagation) 時のダイナミクスも正規化している．

また，各層 l 毎のパラメータベクトル $\boldsymbol{\theta}^l \in \mathbb{R}^{(n_{l-1}+1)n_l}$ と NN の全パラメータベクトル $\boldsymbol{\theta} \in \mathbb{R}^P$ $(P = \sum_{l=0}^{L-1}(n_l + 1)n_{l+1})$ を以下のように定義する．

$$\boldsymbol{\theta}^l \equiv \mathrm{vec}\left(\{\boldsymbol{W}^l, \boldsymbol{b}^l\}\right), \quad \boldsymbol{\theta} = \mathrm{vec}\left(\bigcup_{l=1}^{L} \boldsymbol{\theta}^l\right) \tag{2.3}$$

[*1] watanabe.kaito.xu@alumni.tsukuba.ac.jp

ここで vec(·) は，行列の各列を縦に並べ，1 つの列ベクトルの形にするベクトル化を表す．連続時間 $t \in \mathbb{R}_0^+$ ($\mathbb{R}_0^+$ は 0 を含む正の実数の集合) におけるパラメータの時間依存を $\boldsymbol{\theta}_t$，その初期値を $\boldsymbol{\theta}_0$ とし，NN の出力を $f_t(\boldsymbol{x}) \equiv h_t^{L+1}(\boldsymbol{x}) \in \mathbb{R}^k$ とする．$\hat{\boldsymbol{y}}$ を NN による予測値とすると，損失関数 (loss function) は $\ell(\hat{\boldsymbol{y}}, \boldsymbol{y}) : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ となる．教師あり学習 (supervised learning) では，以下に記述する経験損失 (empirical loss) $\mathcal{L}$ を最小化する $\boldsymbol{\theta}$ の学習を行う．

$$\mathcal{L}_t = \frac{1}{k|\mathcal{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \ell(f_t(\boldsymbol{x}, \boldsymbol{\theta}_t), \boldsymbol{y}) \tag{2.4}$$

3　学習の定義と NTK

学習は損失関数を減らしていくよう，勾配の逆方向にパラメータを変動させていく．バッチ学習を考えるものとして，以下にパラメータ $\boldsymbol{\theta}_t$ の学習を記述する．学習率を η と置く．このとき，

$$\dot{\boldsymbol{\theta}}_t = -\eta \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} \tag{3.1}$$

$$= -\eta \frac{\partial f_t(\mathcal{X})^{\mathrm{T}}}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}}{\partial f_t(\mathcal{X})} \tag{3.2}$$

$$= -\eta \nabla_{\boldsymbol{\theta}} f_t(\mathcal{X})^{\mathrm{T}} \nabla_{f_t(\mathcal{X})} \mathcal{L} \tag{3.3}$$

となる．ここで，$\dot{\boldsymbol{\theta}}$ は $\boldsymbol{\theta}$ の時間微分であり，$\partial/\partial x = \nabla_x$ である．また，$\nabla_{f_t(\mathcal{X})}\mathcal{L}$ は NN の出力 $f_t(\mathcal{X})$ に関する損失の勾配であり，$f_t(\mathcal{X}) = \mathrm{vec}([f_t(\boldsymbol{x})]_{\boldsymbol{x}\in\mathcal{X}}) \in \mathbb{R}^{k|\mathcal{D}|}$ である．

次に，学習によって NN が得る出力を表す関数 $f_t(\mathcal{X})$ がどのように変化していくかを確認する．パラメータ $\boldsymbol{\theta}_t$ の学習と同様，時間微分を考えると，

$$\dot{f}_t(\mathcal{X}) = \frac{\partial f_t(\mathcal{X})}{\partial t} \tag{3.4}$$

$$= \frac{\partial f_t(\mathcal{X})}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial t} \tag{3.5}$$

$$= \nabla_{\boldsymbol{\theta}} f_t(\mathcal{X}) \dot{\boldsymbol{\theta}}_t \tag{3.6}$$

と書くことができる．さらに，式 (3.3) を用いれば，

$$\dot{f}_t(\mathcal{X}) = -\eta \nabla_{\boldsymbol{\theta}} f_t(\mathcal{X}) \nabla_{\boldsymbol{\theta}} f_t(\mathcal{X})^{\mathrm{T}} \nabla_{f_t(\mathcal{X})} \mathcal{L} \tag{3.7}$$

$$= -\eta \hat{\boldsymbol{\Theta}}_t(\mathcal{X}, \mathcal{X}) \nabla_{f_t(\mathcal{X})} \mathcal{L} \tag{3.8}$$

と書くことができる．$\hat{\boldsymbol{\Theta}}_t = \hat{\boldsymbol{\Theta}}_t(\mathcal{X}, \mathcal{X}) \in \mathbb{R}^{k|\mathcal{D}|\times k|\mathcal{D}|}$ は時間 t における Neural Tangent Kernel (NTK) であり，以下のように定義される．

$$\hat{\boldsymbol{\Theta}}_t = \nabla_{\boldsymbol{\theta}} f_t(\mathcal{X}) \nabla_{\boldsymbol{\theta}} f_t(\mathcal{X})^{\mathrm{T}} = \sum_{l=1}^{L+1} \nabla_{\boldsymbol{\theta}^l} f_t(\mathcal{X}) \nabla_{\boldsymbol{\theta}^l} f_t(\mathcal{X})^{\mathrm{T}} \tag{3.9}$$

また，$\mathcal{X}$ 以外の入力 $\boldsymbol{x}' \in \mathbb{R}^{n_0}$ に対する NTK は $\hat{\boldsymbol{\Theta}}_t(\boldsymbol{x}', \mathcal{X})$ と定義できる．この NTK を用いて，学習方程式を関数空間で考える．

4　中間層が無限幅の NN

3 節では NTK の定義を与えたが，一般には関数空間内での学習を考える式 (3.8) の計算は難しい．何故なら，NTK $\hat{\boldsymbol{\Theta}}_t$ は時間毎に変化するためである．しかし，損失関数を平均二乗誤差 (mean square error; MSE)，$\lambda_{\min\,/\,\max}(\boldsymbol{\Theta})$ を NTK $\boldsymbol{\Theta}$ の最小/最大固有値，$\eta_{\mathrm{critical}} \coloneqq 2(\lambda_{\min}(\boldsymbol{\Theta}) + \lambda_{\max}(\boldsymbol{\Theta}))^{-1}$ としたとき，以下の定理が証明されている．

定理 4.1. (**Lee et al., 2019**)

$n_1 = \cdots = n_L = n$ とし，$\lambda_{\min}(\boldsymbol{\Theta}) > 0$ を仮定する．学習率 η が η_{critical} より小さく，任意の入力 $\boldsymbol{x} \in \mathbb{R}^{n_0}$ が $\|\boldsymbol{x}\|_2 \leq 1$ を満たすとき，以下が成立する．

$$\sup_{t\geq 0} \|f_t(\boldsymbol{x}) - f_t^{lin}(\boldsymbol{x})\|_2,\ \sup_{t\geq 0} \frac{\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2}{\sqrt{n}},\ \sup_{t\geq 0} \left\|\hat{\boldsymbol{\Theta}}_t - \hat{\boldsymbol{\Theta}}_0\right\|_F = \mathcal{O}\left(n^{-\frac{1}{2}}\right),\ \text{as} \quad n \to \infty$$

Proof. Appendix A に記載. □

$f_t^{lin}(\boldsymbol{x})$ について，詳細は式 (5.1) で示すが NTK $\hat{\boldsymbol{\Theta}}_0$ を用いて記述されたモデルの出力である．この定理は中間層の幅を無限にしたとき，t を大きく取ったときにも学習の最適なパラメータ $\boldsymbol{\theta}_t$ は $\boldsymbol{\theta}_0$ の近傍にあることを主張しており，さらに $\boldsymbol{\Theta}_t$ を $\boldsymbol{\Theta}_0$ で近似できることも主張している．この定理を用いて，NTK を用いた NN の出力ダイナミクスの記述を行う．

5 NTK を用いた NN の出力ダイナミクス

本節では NN の出力を 1 次のテイラー展開 (Taylor expansion) で置き換えることにより，線形化された NN (linearized neural network; LNN) の学習のダイナミクスを記述する．時間 t 時点における LNN の出力を $f_t^{lin}(\boldsymbol{x})$ とすると，1 次のテイラー展開は，

$$f_t^{lin}(\boldsymbol{x}) \equiv f_0(\boldsymbol{x}) + \nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{x})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \boldsymbol{\omega}_t \tag{5.1}$$

と書ける．ここで，$\boldsymbol{\omega}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_0$ である．f_t^{lin} は第一項がネットワークの初期値であり，第二項が学習中の初期値からの値の変化を示す．$\boldsymbol{\omega}_t$ の時間微分は

$$\dot{\boldsymbol{\omega}}_t = -\eta \nabla_{\boldsymbol{\theta}} f_0(\mathcal{X})^{\mathrm{T}} \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L} \tag{5.2}$$

となる．$\nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{x})$ は学習を通して一定の値を取るので，損失関数に MSE，すなわち $\ell(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{2}|\hat{\boldsymbol{y}} - \boldsymbol{y}|_2^2$ を使用したとき，式 (5.2) について常微分方程式を解けば，

$$\boldsymbol{\omega}_t = -\nabla_{\boldsymbol{\theta}} f_0(\mathcal{X})^{\mathrm{T}} \hat{\Theta}_0^{-1} \left(I - e^{-\eta \hat{\Theta}_0 t/k|\mathcal{D}|}\right) (f_0(\mathcal{X}) - \mathcal{Y}) \tag{5.3}$$

が得られる．つまり，$f_t^{lin}(\mathcal{X})$ は，

$$\begin{aligned}
f_t^{lin}(\mathcal{X}) &= f_0(\mathcal{X}) - \nabla_{\boldsymbol{\theta}} f_0(\mathcal{X}) \nabla_{\boldsymbol{\theta}} f_0(\mathcal{X})^{\mathrm{T}} \hat{\boldsymbol{\Theta}}_0^{-1} \left(I - e^{-\eta \hat{\boldsymbol{\Theta}}_0 t/k|\mathcal{D}|}\right) (f_0(\mathcal{X}) - \mathcal{Y}) && (5.4)\\
&= f_0(\mathcal{X}) - \hat{\boldsymbol{\Theta}}_0 \hat{\boldsymbol{\Theta}}_0^{-1} \left(I - e^{-\eta \hat{\boldsymbol{\Theta}}_0 t/k|\mathcal{D}|}\right) (f_0(\mathcal{X}) - \mathcal{Y}) && (5.5)\\
&= f_0(\mathcal{X}) - \left(I - e^{-\eta \hat{\boldsymbol{\Theta}}_0 t/k|\mathcal{D}|}\right) (f_0(\mathcal{X}) - \mathcal{Y}) && (5.6)\\
&= \left(I - e^{-\eta \hat{\boldsymbol{\Theta}}_0 t/k|\mathcal{D}|}\right) \mathcal{Y} + e^{-\eta \hat{\boldsymbol{\Theta}}_0 t/k|\mathcal{D}|} f_0(\mathcal{X}) && (5.7)
\end{aligned}$$

と書ける．$\mathcal{X}$ 以外の入力 $\boldsymbol{x}'$ に対する出力 $f_t^{lin}(\boldsymbol{x}')$ は，

$$\begin{aligned}
f_t^{lin}(\boldsymbol{x}') &= f_0(\boldsymbol{x}') - \nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{x}') \nabla_{\boldsymbol{\theta}} f_0(\mathcal{X})^{\mathrm{T}} \hat{\boldsymbol{\Theta}}_0^{-1} \left(I - e^{-\eta \hat{\boldsymbol{\Theta}}_0 t/k|\mathcal{D}|}\right) (f_0(\mathcal{X}) - \mathcal{Y}) && (5.8)\\
&= f_0(\boldsymbol{x}') - \hat{\boldsymbol{\Theta}}_0(\boldsymbol{x}', \mathcal{X}) \hat{\boldsymbol{\Theta}}_0^{-1} \left(I - e^{-\eta \hat{\boldsymbol{\Theta}}_0 t/k|\mathcal{D}|}\right) (f_0(\mathcal{X}) - \mathcal{Y}) && (5.9)
\end{aligned}$$

と書ける．したがって，初期化時の NN の出力 $f_0(\mathcal{X})$，$f_0(\boldsymbol{x}')$ と NTK $\hat{\boldsymbol{\Theta}}_0$，$\hat{\boldsymbol{\Theta}}_0(\boldsymbol{x}', \mathcal{X})$ を計算すれば，勾配降下法を実行することなく，LNN の各時間における学習結果を計算できるということである．通常であれば，このように

NN の出力を構成したとしても，各時間における NTK $\hat{\boldsymbol{\Theta}}_t$ を計算する必要があり，何もメリットがない．しかし，4.3 節で述べた定理4.1により，中間層の幅を十分に大きく取れば，任意の学習時間における NTK $\hat{\boldsymbol{\Theta}}_t$ を $\hat{\boldsymbol{\Theta}}_0$ で近似することができるため，関数空間内において NN を線形化することができる．

6 NTK の関連研究

各小節で紹介される論文の記法は元論文に準拠する．

6.1 NTK の関連研究の概要

NTK はJacot et al., 2018によって提案された NN の大域収束性の保証に貢献する理論である．近年，Cho and Saul (2009) により深層 NN (deep neural network; DNN) の学習に対応するカーネル関数が導出され，Neal (2012); Lee et al. (2018) がそのカーネル関数をガウス過程 (gaussian process; GP) の共分散関数として使用することで，中間層の幅を無限に近づけた NN を表現することができることを確認した．NTK は，NN の各層の幅を無限に近づけることで，NTK というグラム行列が決定的 (deterministic) になるということを証明し，特定の条件下においてその行列の正定値性を示すことで，大域的最適解への線形収束を保証した．Lee et al. (2019) はある有限の学習率の設定のもと，モデルの出力のダイナミクスが NTK によって決定されるダイナミクスに従うことを示した．NTK の性質に関する解析も進んでおり，入力の違いに対する安定性と近似能力が優れていることが示されている (Bietti and Mairal, 2019)．Amari (2020) は Jacot et al. (2018) が与えている NTK の数学的に複雑な内容について，任意の目的関数がランダムに生成される目的関数のごく近傍に存在することを幾何的に示した．また，NTK は層の幅を十分に大きくすることを前提としているが，NTK における層の深さの役割に関する研究も進んでいる．Yang and Salman (2019) は識別したい関数の複雑さに応じて最適な深さが存在することを示している．Hanin and Nica (2020) は層の幅と深さの比率を固定して十分に大きくとったときの NTK の平均と分散を研究している．

また，Jacot et al. (2018) が提案した NTK が適用される構造は通常の全結合 NN (fully-connected neural network; FCN) のみであり，畳み込み NN (convolutional neural network; CNN) については Arora et al. (2019) が Convolutional NTK (CNTK) を，自己符号化器 (autoencoder; AE) については Nguyen et al. (2019) が対応する NTK，グラフ NN (graph neural network; GNN) については Du et al. (2019) が Graph NTK (GNTK) を提案している．

6.2 Fast Finite Width Neural Tangent Kernel (ICLR2022)

6.2.1 記法

本論文における記法の定義は以下のとおりとする:

- **N**：NN に入力するデータのバッチサイズ．
  - n：1 から **N** のバッチのインデックス．
- **O**：**N** = 1 に対する NN の出力次元 (クラス数など)．
  - o：1 から **O** のインデックス．
- **W**：FCN の幅，あるいは CNN のチャンネル数．
- **L**：学習可能なパラメータ行列の数．weight sharing がない場合は NN の深さと同義．
  - l：0 から **L** のインデックス．
- **K**：NN の primitive (計算グラフのノード) の数．weight sharing がない場合は NN の深さと同義で，**L** に比例する．

  - k：1 から $\mathbf{K}$ のインデックス．
- $\mathbf{D}$：CNN の入力層と各中間層に含まれるピクセルの総数 (例: 32×32 の画像なら 1024, FCN では 1)．層ごとに空間サイズが変わらないように，SAME または CIRCULAR padding，unit stride，no dilation としている．
- $\mathbf{F}$：CNN の畳み込みフィルタのサイズ (例: 3×3 フィルターなら 9, FCN では 1)．
- $\mathbf{Y}$：primitive y の出力サイズの合計 (例: CNN の場合 $\mathbf{Y} = \mathbf{DW}$，FCN の場合 $\mathbf{Y} = \mathbf{W}$)．状況に応じて単一の primitive の出力サイズ，または全 primitive の出力サイズの合計を表す．
  - y：文脈に応じて中間 primitive 出力，または中間 primitive をパラメータ $y\left(\boldsymbol{\theta}^l\right)$ の関数とする．
- $\mathbf{C}$：primitive Jacobian $\partial y / \partial \boldsymbol{\theta}$ が特定の構造を持つ軸のサイズ ($\mathbf{C}$ は多くの場合，$\mathbf{Y}$ に等しいか，それを占める大きな割合，例えば $\mathbf{W}$ になる)．
  - c：1 から $\mathbf{C}$ の構造軸に沿ったインデックス．

6.2.2 概要

NTK は NN の振る舞いを理解するために重要な役割を果たしている．無限幅の NN を考える際，NTK は解析的に計算できる場合があるが，有限幅の場合には NTK の行列を直接計算する必要があり，これは非常に大きな計算量を要する．

参考文献

S. Amari. Any Target Function Exists in a Neighborhood of Any Sufficiently Wide Random Network: A Geometrical Perspective. *Neural Computation*, 32(8):1431–1447, 2020.

S. Arora, S. S. Du, H. Wei, Z. Li, R. Salakhutdinov, and R. Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A. Bietti and J. Mairal. On the Inductive Bias of Neural Tangent Kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Y. Cho and L. K. Saul. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 342–350, 2009.

S. S. Du, K. Hou, R. R. Salakhutdinov, B. Poczos, R. Wang, and K. Xu. Graph Neural Tangent Kernel: Fusing Graph Neural Networks with Graph Kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5724–5734, 2019.

X. Glorot and Y. Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

B. Hanin and M. Nica. Finite Depth and Width Corrections to the Neural Tangent Kernel. In *International Conference on Learning Representations (ICLR)*, 2020.

A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations (ICLR)*, 2018.

J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

R. M. Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012.

T. V. Nguyen, R. K. Wong, and C. Hegde. Benefits of Jointly Training Autoencoders: An Improved Neural Tangent Kernel Analysis. *arXiv preprint arXiv:1911.11983*, 2019.

V. D. Rajat. Understanding the Neural Tangent Kernel, 2019.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

G. Yang. Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation. *arXiv preprint arXiv:1902.04760*, 2019.

G. Yang and H. Salman. A Fine-Grained Spectral Perspective on Neural Networks. *arXiv preprint arXiv:1907.10599*, 2019.

大阪大学医学部 Python 会. Neural Tangents による無限幅深層ニューラルネットワークの構築とベイズ推論, 2020.

甘利俊一. 新版　情報幾何学の新展開. サイエンス社, 2019.

鈴木大慈. 大阪大学集中講義　深層学習の数理, 2019.

A　定理 4.1 の証明

本節では NTK を用いたバッチ勾配降下における NN の大域的収束性と，勾配降下における NTK の安定性を証明する方法を示す．また，本証明で取り扱うのは NTK parametrization ではなく standard parametrization であるが，若干の変更により NTK parametrization にも適用可能である．

A.1　NTK parametrization と standard parametrization

standard parametization は以下の式で定義される:

$$\begin{cases} h^{l+1} = x^l \boldsymbol{W}^{l+1} + \boldsymbol{b}^{l+1} \\ x^{l+1} = \varphi(h^{l+1}) \end{cases} \quad \text{and} \quad \begin{cases} W_{ij}^l = \omega_{ij}^l \sim \mathcal{N}\left(0, \dfrac{\sigma_\omega^2}{n_l}\right) \\ b_j^l = \beta_j^l \sim \mathcal{N}\left(0, \sigma_b^2\right) \end{cases} \tag{A.1}$$

NTK parametrization(式 (2.1)) と standard parametrization(式 (A.1)) の主な違いは，逆伝播のダイナミクスにおける正規化の有無である．式 (2.1) の場合は逆伝播でも正規化されるが，式 (A.1) の場合は逆伝播は正規化されない．しかし，式 (2.1) は NN の学習にはあまり使用されない．ネットワークが表す関数は式 (2.1) でも式 (A.1) でも同じだが，勾配降下法における訓練のダイナミクスは一般に 2 つの parametrization で異なる．しかし，両 parametrization におけるスケーリングの違いは適切な層ごとの学習率を選ぶことで吸収できることがわかっている．具体的には，まず式 (2.1) については $\boldsymbol{W}^l, \boldsymbol{b}^l$ について独立した層ごとに以下のように学習率を定める．

$$\eta_{\mathrm{NTK},W}^l = \frac{n_l}{n_{\max}\sigma_\omega^2}\eta_0 \quad \text{and} \quad \eta_{\mathrm{NTK},b}^l = \frac{n_l}{n_{\max}\sigma_b^2}\eta_0 \tag{A.2}$$

ここで，η_0 は層によらない学習率であるとする．式 (A.1) の学習率を $\eta_{\mathrm{std}} = \eta_0/n_{\max}$ $(n_{\max} = \max n_l)$ とすると，両 parametrization における学習のダイナミクスは一致する．また，式 (A.2) の両学習率における η_0 の乗法因子を層ごとの Jacobian に組み込むことで，定理 4.1 は式 (A.2) で定義された学習率を持つ NTK network に対しても適用できる．これらの学習率で設定された NTK network は，学習率 η_{std} を持つ standard network と同一の学習のダイナミクスを示すため，十分に広い幅をもつ NTK network の学習のダイナミクスが線形化されるという事実は，standard network に対しても適用される．

A.2　線形化された NN の収束と勾配降下における NTK の安定性

式 (A.1) のように表記された NN について，NTK を用いてバッチ勾配降下における大域的収束性と，NTK の安定性を示す．若干の変更によりこの証明は式 (2.1) にも適用可能である．今，以下の 4 つの仮定を置く．

1. NN のすべての中間層の幅が同一 $(n_1 = \cdots = n_L = n)$ である (本証明は $\min\{n_1, \ldots, n_L\} \to \infty$ のとき，$n_l/n_{l'} \to \alpha_{l,l'} \in (0, \infty)$ の設定に自然に拡張できる)．
2. 解析的 NTK$\boldsymbol{\Theta}$ がフルランク行列で，$0 < \lambda_{\min} < \lambda_{\max} < \infty$.
3. 訓練データ集合 $(\mathcal{X}, \mathcal{Y})$ がコンパクト集合で覆われている．また，任意の $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{X}$ において $\boldsymbol{x} \neq \tilde{\boldsymbol{x}}$ とする．
4. 活性化関数 φ は以下の関係を満たす:

$$|\varphi(0)|, \quad \|\varphi'\|_\infty, \quad \sup_{\boldsymbol{x} \neq \tilde{\boldsymbol{x}}} \frac{|\varphi'(\boldsymbol{x}) - \varphi'(\tilde{\boldsymbol{x}})|}{|\boldsymbol{x} - \tilde{\boldsymbol{x}}|} < \infty \tag{A.3}$$

仮定 2 は $\mathcal{X} \subseteq \{\boldsymbol{x} \in \mathbb{R}^{n_0} : \|\boldsymbol{x}\|_2 = 1\}$ かつ $\varphi(\boldsymbol{x})$ が大きな $\boldsymbol{x}$ に対して多項式的に増加しないときに実際に成立する (Jacot et al., 2018)．本証明において $C > 0$ は $L, |\mathcal{X}|, (\sigma_w^2, \sigma_b^2)$ に依存するが，n には依存しない定数を表す．

$\boldsymbol{\theta}_t$ を時間 t におけるパラメータとする．本証明では以下の略記を使用する．

$$f(\boldsymbol{\theta}_t) = f(\mathcal{X}, \boldsymbol{\theta}_t) \in \mathbb{R}^{|\mathcal{X}| \times k} \tag{A.4}$$

$$g(\boldsymbol{\theta}_t) = f(\mathcal{X}, \boldsymbol{\theta}_t) - \mathcal{Y} \in \mathbb{R}^{|\mathcal{X}| \times k} \tag{A.5}$$

$$J(\boldsymbol{\theta}_t) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_t) \in \mathbb{R}^{k|\mathcal{X}| \times |\boldsymbol{\theta}|} \tag{A.6}$$

standard parametrization における経験的 NTK と解析的 NTK は以下のように定義される．

$$\begin{cases} \hat{\boldsymbol{\Theta}}_t \coloneqq \hat{\boldsymbol{\Theta}}_t(\mathcal{X}, \mathcal{X}) = \dfrac{1}{n} J(\boldsymbol{\theta}_t) J(\boldsymbol{\theta}_t)^\top \\ \boldsymbol{\Theta} \coloneqq \lim\limits_{n \to \infty} \hat{\boldsymbol{\Theta}}_0 \quad \text{in probability} \end{cases} \tag{A.7}$$

解析的 NTK の経験的 NTK への収束は Yang (2019) が厳密に証明している．次に，MSE を以下のとおり与える:

$$\mathcal{L}(t) = \frac{1}{2} \| g(\boldsymbol{\theta})_t \|_2^2 \tag{A.8}$$

$f(\boldsymbol{\theta}_t)$ は平均 0，共分散 $\mathcal{K}$ のガウス分布に収束することから，任意の小さな $\delta_0 > 0$ に対し確率 $(1 - \delta_0)$ 以上で $\| g(\boldsymbol{\theta}_0) \|_2 < R_0$ を満たすような定数 $R_0 > 0$ と n_0 が存在することを示すことができる (ただし，R_0 と n_0 は δ_0，$|\mathcal{X}|$，$\mathcal{K}$ に依存する)．

勾配降下法の更新式は以下の式で与えられる．

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta J(\boldsymbol{\theta}_t)^\top g(\boldsymbol{\theta}_t) \tag{A.9}$$

勾配流 (gradient flow) は以下の式で与えられる．

$$\dot{\boldsymbol{\theta}}_t = -J(\boldsymbol{\theta}_t)^\top g(\boldsymbol{\theta}_t) \tag{A.10}$$

離散的な勾配降下法と勾配流の両方に対し，NN の学習の収束と NTK の安定性を証明する．両証明は Jacobian $J(\boldsymbol{\theta})$ の局所リプシッツ性 (local Lipschitzness) に依存する．

補題 A.1. Jacobian の局所リプシッツ性

ある定数 $K > 0$ が存在して，任意の $C > 0$ に対してランダムな初期化に関して高い確率で以下が成立する．

$$\begin{cases} \dfrac{1}{\sqrt{n}} \| J(\boldsymbol{\theta}) - J(\tilde{\boldsymbol{\theta}}) \|_F & \leq K \| \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} \|_2 \\ \dfrac{1}{\sqrt{n}} \| J(\boldsymbol{\theta}) \| & \leq K \end{cases}, \qquad \forall \boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in B(\boldsymbol{\theta}_0, C n^{-\frac{1}{2}}) \tag{A.11}$$

ここで，$B(\boldsymbol{\theta}_0, R) \coloneqq \{ \boldsymbol{\theta} : \| \boldsymbol{\theta} - \boldsymbol{\theta}_0 \|_2 < R \}$ である．

Proof. 証明はランダムガウス行列の作用素ノルムの上界による．

定理 A.2. (Vershynin, 2010 系 5.35)

$\boldsymbol{A} = \boldsymbol{A}_{N,n}$ を $N \times n$ のランダム行列とし，各成分が独立な標準ガウス分布に従うとする．このとき，任意の $t \geq 0$ に対し，確率 $1 - 2\exp\left(-t^2/2\right)$ 以上で以下が成り立つ．

$$\sqrt{N} - \sqrt{n} - t \leq \lambda_{\min}(\boldsymbol{A}) \leq \lambda_{\max}(\boldsymbol{A}) \leq \sqrt{N} + \sqrt{n} + \sqrt{t} \tag{A.12}$$

$t \geq 1$ に対し，

$$\delta^l(\boldsymbol{\theta}, \boldsymbol{x}) \coloneqq \Delta_{h^l(\boldsymbol{\theta}, \boldsymbol{x})} f^{L+1}(\boldsymbol{\theta}, \boldsymbol{x}) \in \mathbb{R}^{kn} \tag{A.13}$$

$$\delta^l(\boldsymbol{\theta}, \mathcal{X}) \coloneqq \Delta_{h^l(\boldsymbol{\theta}, \mathcal{X})} f^{L+1}(\boldsymbol{\theta}, \mathcal{X}) \in \mathbb{R}^{(k \times |\mathcal{X}|) \times (n \times |\mathcal{X}|)} \tag{A.14}$$

とする．$\boldsymbol{\theta} = \left\{\boldsymbol{W}^l, \boldsymbol{b}^l\right\}$，$\tilde{\boldsymbol{\theta}} = \left\{\tilde{\boldsymbol{W}}^l, \tilde{\boldsymbol{b}}^l\right\}$ を $B(\boldsymbol{\theta}_0, C/\sqrt{n})$ の任意の 2 点とする．定理 A.2 と三角不等式により，$2 \leq l \leq L+1$ のときランダムな初期化に関して高い確率で以下が成立する．

$$\|\boldsymbol{W}^1\|_{\mathrm{op}}, \quad \|\tilde{\boldsymbol{W}}^1\|_{\mathrm{op}} \leq 3\sigma_\omega \frac{\sqrt{n}}{\sqrt{n_0}}, \quad \|\boldsymbol{W}^l\|_{\mathrm{op}}, \quad \|\tilde{\boldsymbol{W}}^l\|_{\mathrm{op}} \leq 3\sigma_\omega \tag{A.15}$$

上記と活性化関数 φ の仮定 (式 (A.3)) を用いると，$\sigma_\omega^2, \sigma_b^2, |\mathcal{X}|, L$ に依存する定数 K_1 が存在して，ランダムな初期化に関して高い確率で以下が成立することを示すことができる．

$$n^{-\frac{1}{2}}\|x^l(\boldsymbol{\theta}, \mathcal{X})\|_2, \quad \|\delta^l(\boldsymbol{\theta}, \mathcal{X})\|_2 \leq K_1 \tag{A.16}$$

$$n^{-\frac{1}{2}}\|x^l(\boldsymbol{\theta}, \mathcal{X}) - x^l(\tilde{\boldsymbol{\theta}}, \mathcal{X})\|_2, \quad \|\delta^l(\boldsymbol{\theta}, \mathcal{X}) - \delta^l(\tilde{\boldsymbol{\theta}}, \mathcal{X})\|_2 \leq K_1\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 \tag{A.17}$$

補題 A.1 はこれら 2 つの不等式から従う．実際，ランダムな初期化に関して高い確率で以下が成り立つ．

$$\|J(\boldsymbol{\theta})\|_F^2 = \sum_l \|J(\boldsymbol{W}^l)\|_F^2 + \|J(\boldsymbol{b}^l)\|_F^2 \tag{A.18}$$

$$= \sum_l \sum_{\boldsymbol{x}\in\mathcal{X}} \|x^{l-1}(\boldsymbol{\theta}, \boldsymbol{x})\delta^l(\boldsymbol{\theta}, \boldsymbol{x})^\top\|_F^2 + \|\delta^l(\boldsymbol{\theta}, \boldsymbol{x})^\top\|_F^2 \tag{A.19}$$

$$\leq \sum_l \sum_{\boldsymbol{x}\in\mathcal{X}} (1 + \|x^{l-1}(\boldsymbol{\theta}, \boldsymbol{x})\|_F^2)\|\delta^l(\boldsymbol{\theta}, \boldsymbol{x})^\top\|_F^2 \tag{A.20}$$

$$\leq \sum_l (1 + K_1^2 n) \sum_{\boldsymbol{x}} \|\delta^l(\boldsymbol{\theta}, \boldsymbol{x})^\top\|_F^2 \tag{A.21}$$

$$\leq \sum_l K_1^2(1 + K_1^2 n) \tag{A.22}$$

$$\leq 2(L+1)K_1^4 n \tag{A.23}$$

同様に以下が成り立つ．

$$\|J(\boldsymbol{\theta}) - J(\tilde{\boldsymbol{\theta}})\|_F^2 = \sum_l \sum_{\boldsymbol{x}\in\mathcal{X}} \|x^{l-1}(\boldsymbol{\theta}, \boldsymbol{x})\delta^l(\boldsymbol{\theta}, \boldsymbol{x})^\top - x^{l-1}(\tilde{\boldsymbol{\theta}}, \boldsymbol{x})\delta^l(\tilde{\boldsymbol{\theta}}, \boldsymbol{x})^\top\|_F^2 + \|\delta^l(\boldsymbol{\theta}, \boldsymbol{x})^\top - \delta^l(\tilde{\boldsymbol{\theta}}, \boldsymbol{x})^\top\|_F^2 \tag{A.24}$$

$$\leq \left(\sum_l (K_1^4 n + K_1^4 n) + K_1 2\right) \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2 \tag{A.25}$$

$$\leq 3(L+1)K_1^4 n\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2 \tag{A.26}$$

□

勾配降下法のケースについて証明する．

定理 A.3. 勾配降下法

先述の 4 つの仮定を満たしているとする．任意の $\delta_0 > 0$, $\eta_0 < \eta_{\mathrm{critical}}$ に対して，ある定数 $R_0 > 0$, $N \in \mathcal{N}$, $K > 1$ が存在して，任意の $n \geq N$ に対して学習率 $\eta = \eta_0/n$ の勾配流を適用したとき，ランダムな初期化に関して確率 $(1 - \delta_0)$ 以上で以下が成り立つ．

$$\begin{cases} \|g(\boldsymbol{\theta}_t)\|_2 \leq \left(1 - \dfrac{\eta_0\lambda_{\min}}{3}\right)^t R_0 \\ \displaystyle\sum_{j=1}^t \|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}\|_2 \leq \frac{\eta_0 K R_0}{\sqrt{n}} \sum_{j=1}^t \left(1 - \frac{\eta_0\lambda_{\min}}{3}\right)^{j-1} \leq \frac{3KR_0}{\lambda_{\min}} n^{-\frac{1}{2}} \end{cases} \tag{A.27}$$

かつ

$$\sup_t \|\hat{\boldsymbol{\Theta}}_0 - \hat{\boldsymbol{\Theta}}_0\|_F \leq \frac{6K^3R_0}{\lambda_{\min}} n^{-\frac{1}{2}} \tag{A.28}$$

Proof. 任意の $n \geq n_0$ に対し，ランダムな初期化に関して確率 $(1-\delta_0/10)$ 以上で以下が成り立つような R_0, n_0 が存在する．

$$\|g(\boldsymbol{\theta}_0)\|_2 < R_0 \tag{A.29}$$

補題 A.1 において $C = \frac{3KR_0}{\lambda_{\min}}$ とする．式 (A.27) を帰納法により証明する．任意の $n \geq n_1$ に対して，ランダムな初期化に関して確率 $(1-\delta_0/5)$ 以上で式 (A.11) と式 (A.29) が成り立つような $n_1 > n_0$ を選ぶ．$t=0$ の場合は明らかであり，t のときに式 (A.27) が成り立つと仮定する．このとき，帰納法の仮定と式 (A.11) の二つ目の不等式より，

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_2 \leq \eta \|J(\boldsymbol{\theta}_t)\|_{\mathrm{op}} \|g(\boldsymbol{\theta}_t)\|_2 \leq \frac{K\eta_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 \tag{A.30}$$

が成り立ち，これは $t+1$ に対する式 (A.27) における一つ目の不等式を与える．また，これは $j = 0, \ldots, t+1$ に対して $\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_0\|_2 \leq \frac{3KR_0}{\lambda_{\min}} n^{-\frac{1}{2}}$ を意味する．二つ目の不等式を証明するため，平均値の定理と $t+1$ ステップ目の勾配降下法の更新式を適用する．

$$\begin{aligned}
\|g(\boldsymbol{\theta}_{t+1})\|_2 &= \|g(\boldsymbol{\theta}_{t+1}) - g(\boldsymbol{\theta}_t) + g(\boldsymbol{\theta}_t)\|_2 && \text{(A.31)}\\
&= \|J(\tilde{\boldsymbol{\theta}}_t)(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + g(\boldsymbol{\theta}_t)\|_2 && \text{(A.32)}\\
&= \| - \eta J(\tilde{\boldsymbol{\theta}}_t) J(\boldsymbol{\theta}_t)^\top g(\boldsymbol{\theta}_t) + g(\boldsymbol{\theta}_t)\|_2 && \text{(A.33)}\\
&\leq \|1 - \eta J(\tilde{\boldsymbol{\theta}}_t) J(\boldsymbol{\theta}_t)^\top\|_{\mathrm{op}} \|g(\boldsymbol{\theta}_t)\|_2 && \text{(A.34)}\\
&\leq \|1 - \eta J(\tilde{\boldsymbol{\theta}}_t) J(\boldsymbol{\theta}_t)^\top\|_{\mathrm{op}} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 && \text{(A.35)}
\end{aligned}$$

ここで，$\tilde{\boldsymbol{\theta}}$ は $\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}$ 間の線形補完である．次に，確率 $(1-\delta_0/2)$ 以上で以下を示す．

$$\|1 - \eta J(\tilde{\boldsymbol{\theta}}_t) J(\boldsymbol{\theta}_t)^\top\|_{\mathrm{op}} \leq 1 - \frac{\eta_0 \lambda_{\min}}{3} \tag{A.36}$$

これは補題 A.1 によって示すことができる．Yang (2019) で示されているように $\hat{\boldsymbol{\Theta}}_0 \to \boldsymbol{\Theta}$ は確率収束するので，任意の $n \geq n_2$ に対して以下の不等式

$$\|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}_0\|_F \leq \frac{\eta_0 \lambda_{\min}}{3} \tag{A.37}$$

が確率 $(1-\delta_0/5)$ 以上で成り立つような n_2 が存在する．仮定 $\eta_0 < 2(\lambda_{\min} + \lambda_{\max})^{-1}$ より，

$$\|1 - \eta_0 \boldsymbol{\Theta}\|_{\mathrm{op}} \leq 1 - \eta_0 \lambda_{\min} \tag{A.38}$$

が成り立つ．したがって，

$$\begin{aligned}
\|1 - \eta J(\tilde{\boldsymbol{\theta}}_t) J(\boldsymbol{\theta}_t)^\top\|_{\mathrm{op}} &\leq \|1 - \eta_0 \boldsymbol{\Theta}\|_{\mathrm{op}} + \eta_0 \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}_0\|_F + \eta \|J(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^\top - J(\tilde{\boldsymbol{\theta}}_t) J(\boldsymbol{\theta}_t)^\top\|_{\mathrm{op}} && \text{(A.39)}\\
&\leq 1 - \eta_0 \lambda_{\min} + \frac{\eta_0 \lambda_{\min}}{3} + \eta_0 K^2 (\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2 + \|\tilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0\|_2) && \text{(A.40)}\\
&\leq 1 - \eta_0 \lambda_{\min} + \frac{\eta_0 \lambda_{\min}}{3} + 2\eta_0 K^2 \frac{3KR_0}{\lambda_{\min}} \frac{1}{\sqrt{n}} \leq 1 - \frac{\eta_0 \lambda_{\min}}{3} && \text{(A.41)}
\end{aligned}$$

が確率 $(1-\delta_0/2)$ 以上で成り立つ．ただし，

$$n \geq \left(\frac{18K^3R_0}{\lambda_{\min}^2}\right)^2 \tag{A.42}$$

である．したがって，

$$N = \max\left\{ n_0, n_1, n_2, \left(\frac{18K^3 R_0}{\lambda_{\min}^2} \right)^2 \right\} \tag{A.43}$$

とすればよい．

式 (A.28) を示すため，以下のことに注意する．

$$\begin{aligned}
\|\hat{\boldsymbol{\Theta}}_0 - \hat{\boldsymbol{\Theta}}_t\|_F &= \frac{1}{n}\|J(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^\top - J(\boldsymbol{\theta}_t)J(\boldsymbol{\theta}_t)^\top\|_F && \text{(A.44)}\\
&\leq \frac{1}{n}\left(\|J(\boldsymbol{\theta}_0)\|_{\mathrm{op}}\|J(\boldsymbol{\theta}_0)^\top - J(\boldsymbol{\theta}_t)^\top\|_F + \|J(\boldsymbol{\theta}_t) - J(\boldsymbol{\theta}_0)\|_{\mathrm{op}}\|J(\boldsymbol{\theta}_t)^\top\|_F\right) && \text{(A.45)}\\
&\leq 2K^2\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_t\|_2 && \text{(A.46)}\\
&\leq \frac{6K^3 R_0}{\lambda_{\min}}\frac{1}{\sqrt{n}} && \text{(A.47)}
\end{aligned}$$

ここでは式 (A.27) の 2 個目の不等式と式 (A.11) を用いた． □