# 非等方的な入力がもたらす<br>非線形ニューラルネットワークの段階的な学習

数理情報学専攻 48-226202　荒田　隼輝
指導教員　定兼 邦彦 教授

## 1　背景

ニューラルネットワークの学習過程は初期値のスケールによって変化する．初期値が大きい場合には，損失が早く減少するが，収束する解はスパースではなく過学習しやすいことが知られている．一方でニューラルネットワークを十分原点に近い値で初期化したときには，スパースな解に収束し汎化性能が高くなることが知られている．さらに近年の研究では，初期値を小さくすることで損失が段階的に減少することが指摘された [1]．このときパラメータ空間では解が鞍点から鞍点へ移動していると予想されており，この学習過程は Saddle-to-Saddle ダイナミクスと呼ばれる．[2, 3] ではいくつかの限定的な問題設定で Saddle-to-saddle ダイナミクスが発生することを証明しているが，解析のために学習率が無限小の勾配流を仮定している．実際には有限ステップ幅の勾配降下法で訓練を行うため，理論とのギャップが生まれてしまう．そのほかにも，初期化の方法，モデルの構造，データの分布に対して強い仮定を必要としており，限定的な理論にとどまっている．

本研究では新しい証明方法を用いることで，これらの仮定がなくても先行研究の結果と整合する Saddle-to-Saddle ダイナミクスが発生することを示した．さらに，提案する証明方法では非線形活性化関数を使用したモデルに対しても拡張が可能であり，活性化関数が線形の場合と同様の学習ダイナミクスを辿ることを示した．モデルが多層の場合には，層ごとの勾配降下法を適用すると，プラトー段階が発生し，重み行列のランクが 1 に近づくことを示した．また Saddle-to-Saddle ダイナミクスではスパースな解が得られる代わりに，学習コストが高いという問題があったが，解析の結果を用いて Saddle-to-Saddle を近似する効率的な最適化アルゴリズムを提案した．

## 2　準備

### 2.1　問題設定

本研究では教師あり学習を扱う．入力データ $X = (x_1, \ldots, x_n) \in \mathbb{R}^{m_0 \times n}$ とこれらに対応するラベル

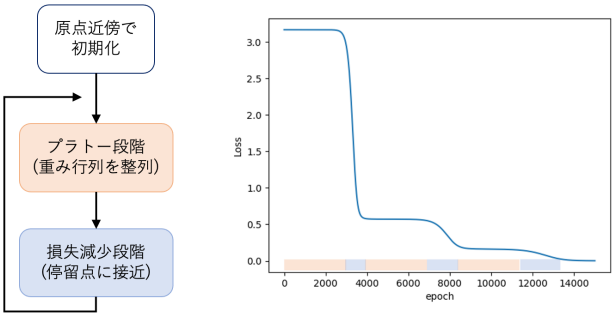


**図 1.** 左: 本研究で発生する Saddle-to-Saddle ダイナミクス，右:Saddle-to-Saddle ダイナミクスと損失の段階的な現象のの対応

$Y = (y_1, \ldots, y_n) \in \mathbb{R}^{m_2 \times n}$ が与えられたとき，損失関数を $l(y, f(x)) : \mathbb{R}^{m_2} \times \mathbb{R}^{m_2} \to \mathbb{R}$ として経験損失 $L = \sum_{i=1}^{n} l(y_i, f(x_i))$ を最小化するようにモデル $f$ のパラメータを最適化する．モデルはニューラルネットワークで，深さが 2，各層の幅が $(m_0, m_1, m_2)$ とする．モデルの出力は $W_l \in \mathbb{R}^{m_l \times m_{l-1}}$ を用いて $f(X) = W_2\sigma(W_1X)$ と表される．$\sigma : \mathbb{R} \to \mathbb{R}$ は活性化関数で要素ごとに作用する．訓練は学習率が $\eta > 0$ の勾配降下法行い，重み行列は次のように更新される．

$$W_l(t+1) = W_l(t) - \eta \nabla L_{W_l}(t).$$

### 2.2　ユニットごとの勾配降下法

**定義 1.** 重み $W_1$ について，$W_1$ の $i$ 行目を $w_{i,1}$，$W_2$ の $i$ 列目を $w_{i,2}$ として $w_i = (w_{i,1}, w_{i,2}) \in \mathbb{R}^{m_0+m_2}$ を $i$ 番目のユニットと呼ぶ．

**補題 1.** 中間層が 1 層の線形ニューラルネットワークの勾配降下法によるパラメータの更新式は，ユニットごとに以下のように表せる．

$$w_i(t+1) = \begin{pmatrix} I & -\eta X D_i(t) \nabla L(t)^T \\ -\eta \left( X D_i(t) \nabla L(t)^T \right)^T & I \end{pmatrix} w_i(t).$$

ただし，

$$\nabla L = \left( \frac{\partial L}{\partial f_L(x_1)}, \ldots, \frac{\partial L}{\partial f_L(x_n)} \right) \in \mathbb{R}^{m_l \times n}.$$

ここで対角行列 $D_i(t) \in \mathbb{R}^{n \times n}$ の $n$ 個の対角要素は

$(D_i(t))_{j,j} = \frac{d\sigma(z)}{dz}|_{z=\langle w_i(t), x_j\rangle}$ とする．特に活性化関数が線形な場合には $(D_i(t))_{j,j} = 1$ である．

## 3　Saddle-to-Saddle ダイナミクス

本研究の設定で発生する Saddle-to-Saddle ダイナミクスは図 1 で表される．原点近傍で初期化すると，初めにプラトー段階が始まる．プラトー段階では損失関数の値がほとんど変化しないという性質から，各ユニットが同じ方向に更新される．次に損失減少段階が始まり，重み行列のランクが 1 のモデルで表現できる鞍点に接近する．その後もプラトー段階と損失減少段階を繰り返し，$k$ 回目の損失減少段階では重み行列のランクが $k$ のモデルで表現できる鞍点に接近する．図 1 左におけるピンク色のプラトー段階と水色の損失減少段階は，図 1 右下部のそれぞれの時刻の色と対応しており，3 回目の損失減少段階で局所最小解に収束していることがわかる．以降では活性化関数が線形の場合の結果を示す．

### 3.1　プラトー段階

**定理 1.** 損失関数がリプシッツ連続であるとする．$(w_i^*)_{i=1}^{m_1}$ を停留点として，$\epsilon = \max_i \|w_i(0) - w_i^*\|$ とする．$P_1$ をある 1 次元空間への直交射影行列とすると，$\delta(\epsilon,\eta) \to 0$ $(\epsilon,\eta \to 0)$ を満たす関数 $\delta$ が存在して，十分大きい $t$ において以下が成り立つ．

$$\frac{\|(I-P_1)w_i(t)\|}{\|P_1 w_i(t)\|} < \delta(\epsilon,\eta)$$

定理 1 より，学習率と初期値を十分小さく取ることで，全てのユニットが同一の方向に更新されることがわかる．

### 3.2　損失減少段階

**補題 2.** あるベクトルへの直交射影行列 $P_1$ が存在して，$t \lesssim \log(\epsilon/\|(I-P_1)w_i(t_1)\|)$ ならば，$\|(I-P_1)w_i(t_1+t)\| \le \epsilon$.

十分小さい $\epsilon$ に対して，損失減少段階でのステップ数を十分大きくとれば鞍点の近傍に収束するため，定理 1 の条件が満たされ再度プラトー段階が始まる．したがって $k$ 回目の損失減少段階では，$k$ 次元線型空間への直交射影行列 $P_k$ を用いて，ある $\epsilon > 0$ について以下の最適化問題を解いていると見なすことができる．

$$\min L(W_1, W_2) \quad s.t. \|(I-P_{v_k})w_i\| \le \epsilon$$

これは近似的なランク制約のもとでの最適化問題となっている．

## 4　効率的な最適化アルゴリズム

Saddle-to-Saddle ダイナミクスは初期値のノルムを十分小さくすることで発生するが，学習に必要なステップ数は $O\left(\log \frac{1}{\|w_i(0)\|}\right)$ となる．定理 1 よりプラトー段階において各ユニットが更新される方向は原点における係数行列の固有ベクトル方向で近似できることがわかるため，プラトー段階を定数ステップで終了するアルゴリズムを考えることができる．図 1 右と図 2 は同一の問題設定で学習アルゴリズムだけを変更している．図 1 では通常の勾配降下法を用い，図 2 では提案する効率的なアルゴリズムを用いた．図 2 では，図 1 左のピンクで表されるプラトー段階がなくなり，その分だけ早く収束していることがわかる．

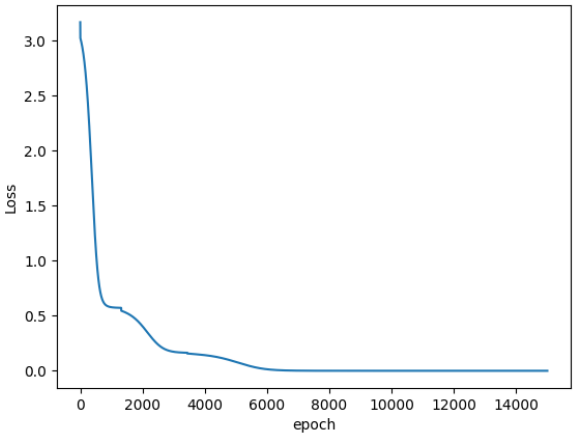


**図 2.** Saddle-to-Saddle ダイナミクスを近似する効率的な最適化アルゴリズムを用いた時の損失の変化

## 参考文献

[1] Gauthier Gidel, Francis R. Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems 32*, pages 3196–3206, 2019.

[2] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2022.

[3] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *the 9th International Conference on Learning Representations*, 2021.