オンライン学習における雑音誘起現象

佐藤譲 *
(北海道大学 電子科学研究所 / 理学研究院数学部門,
London Mathematical Laboratory)
Yuzuru Sato
(RIES / Department of Mathematics, Hokkaido University,
London Mathematical Laboratory)

1 学習とランダム力学系

ニューラルネットワークをパラメトリックモデルとみなした最小二乗法は, 典型的な機械学習論のスキームであり, 最適化アルゴリズムとしては, 最も単純な勾配降下法がしばしば用いられる. いま $\boldsymbol{\theta} = (w_1, w_2, \ldots)$ をニューラルネットワークのパラメーター, x を確率変数で表現される入力データ, 多層パーセプトロン[1]を $f(x;\boldsymbol{\theta})$, ターゲット関数を $T(x)$, 損失関数を $l(x;\boldsymbol{\theta}) = ||f(x;\boldsymbol{\theta}) - T(x)||^2$ で与える. 時刻 t で入力される S 個のデータを $\{x_i(t)\}_{i=1,\ldots,S}$ とすると, 学習率 η の勾配降下法 (gradient descent) は以下で与えられる;

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}\,(t) - \eta\frac{1}{S}\sum_{i=1}^{S}\nabla_{\boldsymbol{\theta}} l(x_i(t);\boldsymbol{\theta}(t)). \tag{1}$$

S が有限の場合, (1) は確率的ダイナミクスとなり, 確率的勾配降下法 (stochastic gradient descent) とよばれる.

*ysato@math.sci.hokudai.ac.jp

[1] 入力層, 中間層, 出力層からなるフィードフォワード型ニューラルネットワーク [1].

$S \to \infty$ の極限で平均損失 $E_x[l(x;\boldsymbol{\theta}(t)]$ が存在する場合, (1) は

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}\ (t) - \eta \nabla_{\boldsymbol{\theta}} E_x[l(x;\boldsymbol{\theta}(t))], \tag{2}$$

で与えられる $\boldsymbol{\theta}$ の決定論力学系となる. この $S \to \infty$ の極限で与えられる勾配降下法は決定論的勾配降下法とよばれる. 一方 $S \to 1$ の極限で (1) は入力データを逐次処理するアルゴリズムとなり, オンライン学習のモデルとなる. このとき (1) は

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}\ (t) - \eta \nabla_{\boldsymbol{\theta}} l(x(t);\boldsymbol{\theta}(t)). \tag{3}$$

という, 確率変数 x と決定論変数 $\boldsymbol{\theta}$ からなるランダム力学系となる[2].

多層パーセプトロンは, 中間層の素子数が十分大きい極限で関数近似万能性を持つ [2]. 最もよく研究されている多層パーセプトロンは単一の中間層を持つ 3 層パーセプトロンである[3]. 本論文では力学系理論的視点から 3 層パーセプトロンにおける確率的勾配降下法のダイナミクスを考察する. 確率的ダイナミクスに対するフォッカー・プランク方程式によるアプローチ [4, 5] とは対照的に, ランダム力学系に基づくアプローチによって, 学習ダイナミクスのパスワイズな構造, ランダムアトラクターの安定性, 確率分岐などが議論できる.

2　勾配消失と過学習

多層パーセプトロンの勾配降下法については, 準安定状態である局所最適解へのトラップ, という問題以外に, 典型的かつより非自明な問題が二つあげられる.

1. **勾配消失: 学習過程で勾配が 0 に近くなり, 長時間の停滞が生じる現象**

   一般に多層パーセプトロンのパラメータが縮退した特異領域により勾配消失 (vanishing gradient) が生じ, 学習が停滞する [6]. 損失関数を時間の関数として描いた学習曲線に停滞による平坦な部分 (プラトー) が生じることから, プラトー現象とよばれる[4]. この特異領域はアトラクター上に測度 0 で反発的な

[2] 決定論的勾配降下法 (2) は確率論的勾配降下法 (3) の x に対する平均化ダイナミクスである.

[3] 階層数の多い多層パーセプトロンをディープネットワークとよび, ディープネットワークに基づく学習をディープラーニングとよぶ [3]. 様々な実問題に対してディープラーニングが高性能を示すことが近年認識されているが, そのメカニズムはよくわかっていない. 5 階層程度の機械学習系でもディープラーニングとよばれることがある.

[4] このプラトーはポテンシャル中の平坦領域とは必ずしも対応しないことに注意.

点をもつという意味で, ミルナー型吸引領域である [7, 8]. 典型的には中間層の複数のパラメーターが同期し，学習系の有効自由度が減少して勾配が 0 に近くなる，この停滞状態から脱出し, 再び学習を進行させるためには学習系の有効自由度を上げる必要がある.

2. **過学習**: **入力データに過剰適合した汎用性のない関数が学習される現象**

   一般にニューラルネットワークの関数近似能力を活かすためには, ある程度の数のパラメーターが必要である. しかし逆に多すぎるパラメーターを使って関数を表現しようとすると過学習 (overfitting) が生じる [6]. 例えば極端な場合, データ点をすべて補完する関数, つまり入出力関係の完全な対応表を作ってしまい, 汎化に失敗する. この過学習状態から回復するためには学習系の有効自由度を下げる必要がある.

この二つの問題を力学系的視点から分析するのが本研究の目標である.

3 Fukumizu-Amari モデルにおける雑音誘起縮退現象

中間層に 2 つのニューロンを持つ 3 層パーセプトロン (図 1 参照) がプラトー現象を示すことがわかっている [8]. このミニマルモデルを Fukumizu-Amari モデルとよぶ.

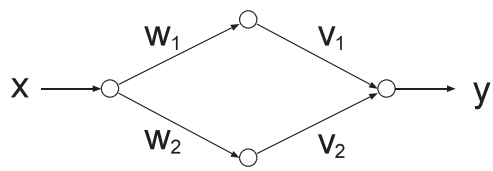

図 1: 中間層に 2 つのニューロンを持つ 3 層パーセプトロン: ノードは活性化関数 $\tanh(\cdot)$. エッジはパラメーター $\boldsymbol{\theta} = (w_1, w_2, v_1, v_2)$ による線形重ね合わせを表す. 出力 y は入力 x とパラメーター $\boldsymbol{\theta}$ の関数 $f(x; \boldsymbol{\theta})$ で与えられる.

同じ学習系でのオンライン学習を考えよう. モデルは以下で与えられる [9].

$$\begin{aligned}
\boldsymbol{\theta}(t+1) &= \boldsymbol{\theta}(t) - \eta \nabla_{\boldsymbol{\theta}} l(x(t); \boldsymbol{\theta}(t)),\ (t = 1, 2, \ldots), && (4)\\
&\boldsymbol{\theta} = (w_1, w_2, v_1, v_2) \in \boldsymbol{R}^4, && (5)\\
&l(x; \boldsymbol{\theta}) = \frac{1}{2}(f(x; \boldsymbol{\theta}) - T(x))^2, && (6)\\
&f(x; \boldsymbol{\theta}) = v_1 \tanh(w_1 x) + v_2 \tanh(w_2 x). && (7)
\end{aligned}$$

x は $N(0, \sigma^2)$ に従うとし, ターゲット関数 $T(x)$ は以下で与える;

$$T(x) = 2\tanh(x) - \tanh(4x). \tag{8}$$

このモデルで損失関数 $l(x; \boldsymbol{\theta})$ を最小化する $f(x; \boldsymbol{\theta}^*)$ は以下の $\boldsymbol{\theta}^*$ で与えられる;

$$\begin{aligned}
\boldsymbol{\theta}^* &= (1, 4, 2, -1),\ (-1, 4, -2, -1),\ (1, -4, 2, 1),\ (-1, -4, -2, 1),\\
&\quad (4, 1, -1, 2),\ (4, -1, -1, -2),\ (-4, 1, 1, 2),\ (-4, -1, 1, -2). \qquad (9)
\end{aligned}$$

図 2 はこのランダム力学系のダイナミクスがプルバック・アトラクターに収束していく過程 (有限時間プルバック・アトラクター [10]) をプルバック時間 $\tau = 1000, 10000, 30000, 100000$ について図示したものである. 入力データ x のゆらぎ σ が大きいとき学習ダイナミクスの停滞が点線円内にみられる. 詳細な安定性解析は [9] を参照. プラトー現象は, $S \to \infty$ の決定論勾配降下法より, $S \to 1$ のオンライン学習系の方がむしろ強化されることがわかった. このことは平均エスケープ時間の計測により観察できる. ゆらぎ σ^2 が小さいと停滞領域からのエスケープ時間は大きいが, ゆらぎが大きくなるにつれて停滞から脱出しやすくなり, エスケープ時間が小さくなる. ところがオンライン学習では $\sigma^2 \simeq 0.07$ 以上にゆらぎを大きくしていくと, エスケープ時間が再び大きくなっていく (図 3 参照). エスケープ時間を最小化する「最適な」ゆらぎサイズは $\sigma^2 \simeq 0.07$ である.

この停滞の強化は, ノイズ同期 [11, 12] に近いメカニズムにより, 特異領域に直行する方向のリアプノフ指数が負になるために生じる. この現象は決定論的勾配降下法, あるいは確率的勾配降下法を決定論勾配系+加法ガウスノイズで近似した系ではとらえることができないエスケープダイナミクスであり, ランダム力学系理論的視点で分析して, はじめてその存在が明らかになったものである. このようなオンライン学習に固有の現象を, 雑音誘起縮退現象 (noise-induced degeneration) と

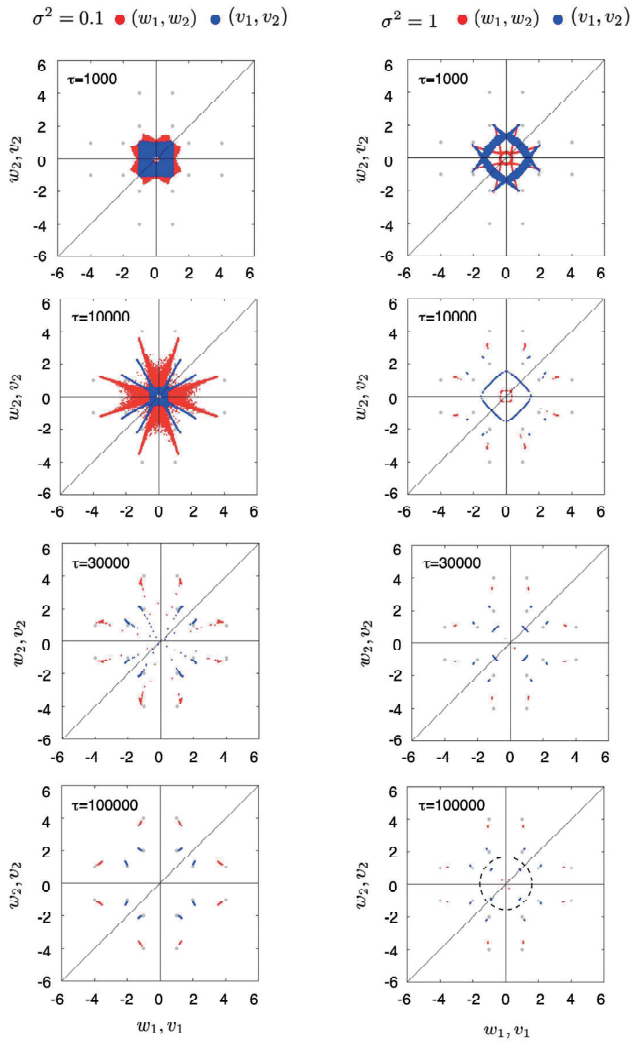

図 2: 有限時間プルバック・アトラクター: 学習率を $\eta = 0.1$ とし, $\sigma^2 = 0.1, 1.0$ の場合について, プルバック時間 $\tau = 1000, 10000, 30000, 100000$ での軌道束を図示した. 赤点が (w_1, w_2), 青点が (v_1, v_2), 灰点が最適解 $\boldsymbol{\theta}^*$ を表す. 縮退部分空間 $w_1 = w_2$, $v_1 = v_2$ は対角線で表示されている. 数値実験では典型的なノイズ系列 $\{x(t)\}$ を固定し, 10^5 個の初期値 $\boldsymbol{\theta}(0) \in [-1, 1]^4$ から出発した軌道をすべて重ね書きして観察する. $\sigma = 1.0$ のとき, 学習ダイナミクスの停滞が点線円内にみられる.

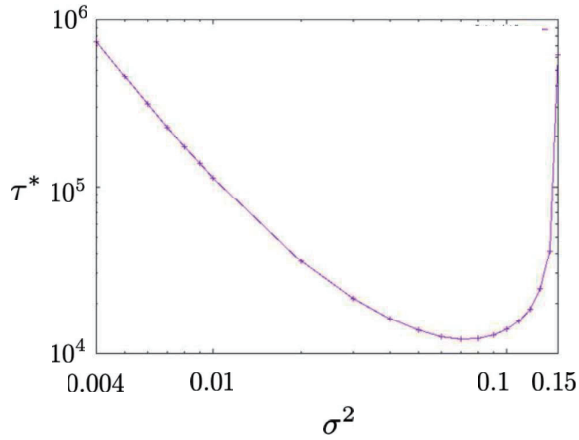

図 3: 平均エスケープ時間とゆらぎ: $\boldsymbol{\theta}(0) \in [-1,1]^4$ から出発した軌道が閉領域 $[-2,2]^4$ から脱出するエスケープ時間 τ^* の平均をゆらぎサイズ σ^2 の関数として log-log プロットで表示した. その他のモデルパラメーターは図 2 と同じである. 入力ゆらぎ σ^2 を大きくするとエスケープ時間 τ^* が大きくなることがみてとれる. エスケープ時間を最小化するゆらぎサイズは $\sigma^2 \simeq 0.07$.

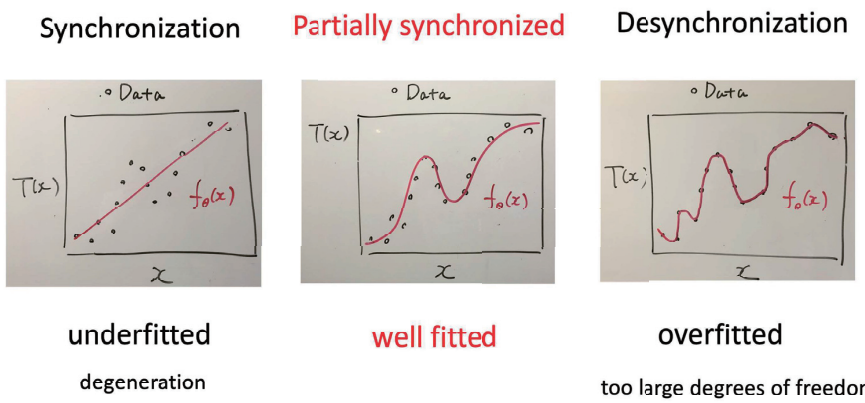

図 4: 勾配消失, 過学習と学習系の有効自由度の関係: 外力により有効自由度の増減が生じ, 有効自由度が低すぎると勾配消失, 高すぎると過学習が生じる. 外力に応じた $\boldsymbol{\theta}$ の部分同期状態が多層パーセプトロンの学習を支えている.

よぶ. 多層パーセプトロンの対称性と $\boldsymbol{\theta}$ の状態空間の階層性より, この現象の普遍性が示唆される [9].

一般の, より大きなネットワークもこの 2 素子 3 階層のパーセプトロンをモジュールとして含むので, 上記の性質を持つ停滞領域が状態空間内に階層的に存在している. これが多層パーセプトロンの学習ダイナミクスの特徴である. このような結合力学系に非自励的外力を加える (データを入力する) と, 状況と文脈に応じて有効自由度の増減が生じ, 有効自由度が低すぎると勾配消失, 高すぎると過学習が生じる (図 4 参照). 汎化にはある程度のネットワークの縮退が必要なので, 外力に応じた $\boldsymbol{\theta}$ の「適度な」部分同期状態が多層パーセプトロンの学習を支えている, といってもよいだろう.

4 結び

近年複数の分野で非自励力学系, ランダム力学系の理論の重要性が認識されている. 本稿で議論した機械学習論の勾配消失問題のような古典的問題も, 非線形確率現象とみなして解析できることがわかってきた. ランダム力学系で生じる様々な複雑現象を理解するには, 既存の力学系理論 · エルゴード理論の概念を外挿するだけではなく, 新しい数学的 · 物理的な概念を構築していくことが必要である.

参考文献

[1] Shun-ichi Amari. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, 3:299–307, 1967.

[2] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[4] Pratic Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *arXiv preprint arXiv:1710.11029*, 2017.

[5] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18:4873–4907, 2017.

[6] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[7] John Milnor. On the concept of attractor. In *The theory of chaotic attractors*, pages 243–264. Springer, 1985.

[8] Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks*, 13(3):317–327, 2000.

[9] Yuzuru Sato, Daiji Tsutsui, and Akio Fujiwara. Noise-induced degeneration in online learning. *arXiv preprint arXiv:2008.10498*, 2020.

[10] Yuzuru Sato, Mickaël D Chekroun, and Michael Ghil. Convergence rate of snapshot attractors to random strange attractors. *submitted*, 2020.

[11] Arkady S. Pikovskii. Synchronization and stochastization of array of self-excited oscillators by external noise. *Radiophysics and Quantum Electronics*, 27(5):390–395, 1984.

[12] Jun-nosuke Teramae and Dan Tanaka. Robustness of the noise-induced phase synchronization in a general class of limit cycle oscillators. *Physical review letters*, 93(20):204103, 2004.