Noise-induced degeneration in online learning

Yuzuru Sato [a,b,*], Daiji Tsutsui [c,d], Akio Fujiwara [c]

[a] *RIES/Department of Mathematics, Hokkaido University, Kita 20 Nishi 10, Kita-ku, Sapporo 001-0020, Japan*
[b] *London Mathematical Laboratory, 8 Margravine Gardens, London W68RH, UK*
[c] *Department of Mathematics, Osaka University, Toyonaka, Osaka 560-0043, Japan*
[d] *Technology Management Department, Shanon Incorporated, Minato-ku, Tokyo 108-0073, Japan*

ARTICLE INFO

ABSTRACT

In order to elucidate the plateau phenomena caused by vanishing gradient, we herein analyse stability of stochastic gradient descent near degenerated subspaces in a multi-layer perceptron. In stochastic gradient descent for Fukumizu–Amari model, which is the minimal multi-layer perceptron showing non-trivial plateau phenomena, we show that (1) attracting regions exist in multiply degenerated subspaces, (2) a strong plateau phenomenon emerges as a noise-induced synchronisation, which is not observed in deterministic gradient descent, (3) an optimal fluctuation exists to minimise the escape time from the degenerated subspace. The noise-induced degeneration observed herein is expected to be found in a broad class of machine learning via neural networks.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The least square learning of neural networks is a typical framework in machine learning. The gradient descent is the simplest optimisation algorithm represented by gradient dynamics in a potential. When the input data is finite, gradient descent dynamics fluctuates due to the finite size effects, and is called stochastic gradient descent. In this paper, we study stability of stochastic gradient descent dynamics from the viewpoint of dynamical systems theory.

Learning is characterised as non-autonomous dynamics driven by uncertain input from the external, and as multi-scale dynamics which consists of slow memory dynamics and fast system dynamics. When the uncertain input sequences are modelled by stochastic processes, dynamics of learning is described by a random dynamical system. In contrast to the traditional Fokker–Planck approaches [1,2], the random dynamical system approaches enable the study not only of stationary distributions and global statistics, but also of the pathwise structure of stochastic dynamics. Based on non-autonomous and random dynamical system theory, it is possible to analyse stability and bifurcation in machine learning.

We adopt a *multi-layer perceptron*, a class of feed-forward neural networks, as the parametric model $f(x; \theta)$, where θ is the parameter of neural networks and x is the input data. It is known that multi-layer perceptron with a single hidden layer is a universal function approximator [3], and broadly used for solving generalisation problems.

In recent years, behaviour of *stochastic gradient descent* [4,5] in finite settings, given by the following equation, has become more important:

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

where η is the learning rate, $\{x_i(t)\}$ the input data, and $l(\cdot; \theta)$ a loss function, typically given by the squared norm $\|f(x; \theta) - T(x)\|^2$, where $T(x)$ is the target function. For each time t, a finite set $\{x_i(t)\}_{i=1}^{S}$ of input data, called a *batch* of learning data, is randomly chosen, and Eq. (1) is a stochastic dynamics with finite S.

The *deterministic gradient descent* is given by the gradient dynamics of the averaged potential $E_x[l(x; \theta)]$;

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

where $E_x[\cdot]$ denotes the expectation over x. The averaged potential $E_x[l(x; \theta)]$ is also known as the risk function. The approximation in Eq. (2) corresponds to the assumption that a large set of training data x is given to the system at once, yielding the exact average potential $E_x[l(x; \theta)]$. Eq. (2) is a deterministic gradient dynamics.

In many cases, the gradient descent in multi-layer perceptron exhibits slow convergence to the optimal, known as a *plateau phenomenon* (see Fig. 1), which makes the dynamics of learning slow down [6,7]. The plateau phenomenon is a chain of slow dynamics near attracting regions, during which the loss function

---

* Corresponding author at: RIES/Department of Mathematics, Hokkaido University, Kita 20 Nishi 10, Kita-ku, Sapporo 001-0020, Japan.
E-mail addresses: ysato@math.sci.hokudai.ac.jp (Y. Sato), u274431k@alumni.osaka-u.ac.jp (D. Tsutsui).
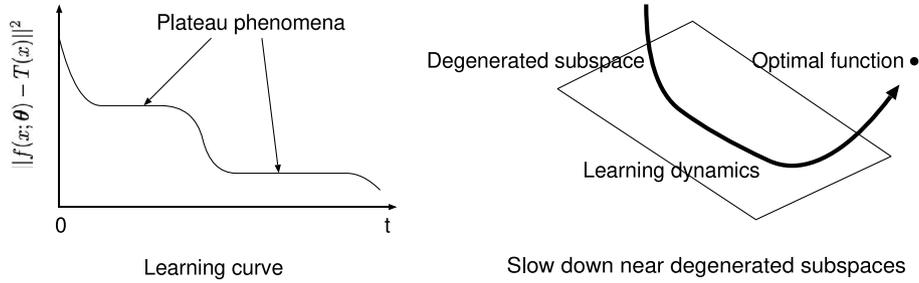
Fig. 1. A schematic view of the plateau phenomena (left) and a stagnant dynamics near the degenerated subspace (right). The dynamics of learning slows down near the attracting region in the degenerated subspace, but eventually escapes to the optimal.

reduces extremely small amount per unit time. Each plateau or trapping is caused by a saddle set or a Milnor attractor, which is an attractor with a zero-measure set of escape holes [8]. In a degenerated subspace, in Eq. (2), these attracting structures cause slow convergence to the optimal [9–11]. Fukumizu and Amari studied the deterministic gradient descent in the minimal three-layer perceptron with one input neuron, two neurons in the hidden layer, and one output neuron, which we call Fukumizu–Amari model, and found non-trivial plateau phenomena based on Milnor attractors [9].

Learning in multi-layer perceptrons with $S = 1$, is called *online learning*. Online learning is modelled by a stochastic gradient descent dynamics with one-dimensional external force, which, however, is not well understood theoretically. Recently, plateau phenomena in online learning has been studied, based on the averaged dynamics $\bar{\theta}(t)$ of stochastic gradient descent with respect to the deterministic potential $E_x[\nabla_{\bar{\theta}} l(x; \bar{\theta}(t))]$ by averaging the stochastic gradient over x [12]. However, such an approximation is not fully valid in online learning, and recently a different approximation with stochastic dynamics near degenerated subspace is studied [13].

In order to elucidate the underlying mechanism of the plateau phenomena in multi-layer perceptrons, we here analyse, as a prototype, stochastic gradient descent of Fukumizu–Amari model from a viewpoint of random dynamical systems without any averaging;

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

where $x(t)$ is a discrete time stochastic process. We show that, in a class of Fukumizu–Amari model, (1) attracting regions exist in multiply degenerated subspaces, (2) a strong plateau phenomenon emerges as a noise-induced synchronisation, which is not observed in deterministic gradient descent, (3) an optimal fluctuation size exists to minimise the escape time from the degenerated subspace.

In the studies of stochastic gradient descent, the importance of the anisotropy of noise has been pointed out [14]. A well-studied field is the Langevin gradient dynamics [15], which is driven by isotropic noise, and is tractable by Fokker–Planck analysis; however, it is essentially different from the dynamics of stochastic gradient descent in the case of small batch size S. Keskar et al. found that stochastic gradient descent with small batch admits the convergence to flat minimisers [16], which results in better generalisation [17]. Some studies reported that small batch methods accelerate learning [18–20].

In contrast, we find that there is an optimal fluctuation size which minimises the escape time. To the best of our knowledge, such a phenomenon has not been reported in studies on either deterministic or stochastic gradient descent. It is thus expected that our random dynamical system approach will shed new light on the analysis of online-learning.

The paper is organised as follows. In Section 2, we introduce the minimal three-layer perceptron, Fukumizu–Amari model, which shows plateau phenomena. In Section 3, we analyse stability of stochastic gradient descent in Fukumizu–Amari model; Global stability in the whole state space (Section 3.1) and local stability in the multiply degenerated subspace (Section 3.2) are discussed. In Section 4, we study phenomenology of "noise-induced degeneration". In Section 5. we give a summary and an overview.

2. Stochastic gradient descent in multi-layer perceptrons

The minimal model of multi-layer perceptrons which exhibits plateau phenomena is given by a three-layer perceptron with gradient descent [9] (see Fig. 2).
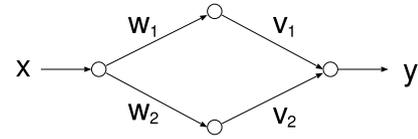
Fig. 2. The three-layer perceptron: The nodes are activation functions given by $\tanh(\cdot)$, and each edge indicates a linear superposition with parameters (w_1, w_2, v_1, v_2). The output y is a function of the input x and the parameters (w_1, w_2, v_1, v_2).

The equation of motion of the stochastic gradient descent (3) is explicitly written as follows;

$$\theta(t + 1) = \theta(t) - \eta \nabla_{\theta} l(x(t); \theta(t)), \quad (t = 1, 2, \ldots) \tag{4}$$

where

$$\theta = (w_1, w_2, v_1, v_2) \in \Theta \tag{5}$$

$$l(x; \theta) = \frac{1}{2}(f(x; \theta) - T(x))^2 \tag{6}$$

$$f(x; \theta) = v_1 \tanh(w_1 x) + v_2 \tanh(w_2 x). \tag{7}$$

Here, Θ is a domain of $\mathbf{R}^4$ called the parameter space, x is an i.i.d. random variable subject to a probability distribution $\rho(x)$, $\eta \in [0, 1]$ is the learning rate, and $T(x)$ is the target function to be learnt. A Fukumizu–Amari model [9] is given by Eq. (4) with a target function

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

and $\rho(x) \sim N(0, \sigma^2)$. The optimal function $f(x; \theta^*)$ with parameters θ^* which minimise $l(x; \theta)$ is given by

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

Given Eqs. (4)–(8), the following subspaces

$$\boldsymbol{\theta}^{\dagger} = (w, w, v_1, 2v - v_1), \ (w, -w, v_1, v_1 - 2v),$$
$$(w_1, w, 0, 2v), \ (w, w_2, 2v, 0) \quad (10)$$

define a class of degenerated functions

$$f(x; \boldsymbol{\theta}^{\dagger}) = 2v \tanh(wx). \quad (11)$$

Typically, the plateau phenomena emerge near the degenerated subspace (10) in the parameter space Θ [12].

In this paper, we focus on the interplay between degenerated subspaces $M_w = \{\boldsymbol{\theta}|w_1 = w_2 = w\}$ and $M_{wv} = \{\boldsymbol{\theta}|w_1 = w_2 = w, v_1 = v_2 = v\}$. In the degenerated subspace M_w, the effective degrees of freedom of Eq. (4) decreases to 3. Although we still have free variable v_1 in M_w, it does not contribute to a better function approximation. Furthermore, when a multiple degeneration to M_{wv} occurs, the effective degrees of freedom decreases to 2, and the dynamics cannot even return to the full parameter space Θ by gradient descent.

In many cases, the dynamics of learning stays near the degenerated subspace for a very long time. This trapping phenomenon is caused by neutral stability with vanishing gradients. In online learning, in addition to this neutrally stable trapping, "stronger trapping" based on multiple degeneration may occur as a noise-induced synchronisation [21–23]. We call this type of degeneration as *noise-induced degeneration* in online learning.

3. Stability analysis of stochastic gradient descent

3.1. Global attraction to the degenerated subspace

We focus on the following random map derived from Fukumizu–Amari model (4)–(7);

$$w_1(t + 1) = w_1(t) - \eta x v_1(t) \frac{h(x; \boldsymbol{\theta}, T)}{\cosh^2(w_1(t)x)}, \quad (12)$$

$$w_2(t + 1) = w_2(t) - \eta x v_2(t) \frac{h(x; \boldsymbol{\theta}, T)}{\cosh^2(w_2(t)x)}, \quad (13)$$

$$v_1(t + 1) = v_1(t) - \eta \tanh(w_1(t)x)h(x; \boldsymbol{\theta}, T), \quad (14)$$

$$v_2(t + 1) = v_2(t) - \eta \tanh(w_2(t)x)h(x; \boldsymbol{\theta}, T), \quad (15)$$

where

$$h(x; \boldsymbol{\theta}, T) = v_1 \tanh(w_1 x) + v_2 \tanh(w_2 x) - T(x). \quad (16)$$

The fluctuation σ^2 of the learning data is a control parameter. Our numerical experiments suggest that, for a broad region of large σ^2 and a positive measure set of initial conditions, and for a finite time, there exists attracting dynamics from the full space Θ to a degenerated subspace

$$M_w = \{\boldsymbol{\theta} \mid w_1 = w_2 = w\}. \quad (17)$$

Furthermore, with large σ^2, we observe attracting dynamics from the degenerated subspace M_w to multiply degenerated subspace

$$M_{wv} = \{\boldsymbol{\theta} \mid w_1 = w_2 = w, \ v_1 = v_2 = v\}. \quad (18)$$

The attraction to M_w is caused by a local minimum of the averaged potential $E_x[l(x; \boldsymbol{\theta})]$. Due to the valley formed by steep gradients of the averaged potential, the gradient dynamics is dominant compared with the stochastic effects and approaches a neighbourhood of M_w quickly (see Appendix A).

To investigate local dynamics near the degenerated subspace M_w, one can introduce the following coordinate system

$$\begin{cases} p = \dfrac{w_1 + w_2}{2}, & q = \dfrac{v_1 + v_2}{2}, \\ r = \dfrac{w_1 - w_2}{2}, & s = \dfrac{v_1 - v_2}{2}, \end{cases} \quad (19)$$

and the corresponding transformed model

$$p(t + 1) - p(t)$$
$$= -\frac{\eta x h}{2} \left[\frac{q(t) + s(t)}{\cosh^2((p(t) + r(t))x)} + \frac{q(t) - s(t)}{\cosh^2((p(t) - r(t))x)} \right], \quad (20)$$

$$q(t + 1) - q(t)$$
$$= -\frac{\eta h}{2} \left[\tanh((p(t) + r(t))x) + \tanh((p(t) - r(t))x) \right], \quad (21)$$

$$r(t + 1) - r(t)$$
$$= -\frac{\eta x h}{2} \left[\frac{q(t) + s(t)}{\cosh^2((p(t) + r(t))x)} - \frac{(q(t) - s(t))}{\cosh^2((p(t) - r(t))x)} \right], \quad (22)$$

$$s(t + 1) - s(t)$$
$$= -\frac{\eta h}{2} \left[\tanh((p(t) + r(t))x) - \tanh((p(t) - r(t))x) \right], \quad (23)$$

where h is given by

$$h = h(x; \boldsymbol{\theta}, T) = (q+s) \tanh((p+r)x) + (q-s) \tanh((p-r)x) - T(x). \quad (24)$$

In what follows, we confine ourselves to the situation where the parameters are in a bounded region and are stagnant away from the origin, which is typical of the synchronising dynamics observed in our numerical experiments.

Theorem 1. *Assuming that*

1. *$\boldsymbol{\theta}$ is near M_w*
2. *$\boldsymbol{\theta}$ is in a bounded region D, where*

$$D := \{\boldsymbol{\theta}|w_1, w_2 \in [\kappa_1, \kappa_1 + \delta], v_1, v_2 \in [-\kappa_2, \kappa_2]\} \ (0 < \kappa_1, \kappa_2, \delta)$$

3. *The target function $T(x)$ is bounded as $|T(x)| \leq L$,*

for a sufficiently small learning rate $\eta > 0$, the dynamics of $s(t)$ is contracting and approaches near 0 in average.

Proof. A Taylor expansion of Eq. (23) yields

$$s(t + 1) = \eta C_1 + [1 - \eta C_2] s(t) + O(r^3) \quad (25)$$

where

$$C_1 = -\frac{rx(2q \tanh(px) - T(x))}{\cosh^2(px)} \quad (26)$$

$$C_2 = \frac{2r^2 x^2}{\cosh^4(px)}. \quad (27)$$

Since $|\frac{x}{\cosh^2 x}| < 1$ and $p \simeq w_1 \simeq w_2$ near M_w, we have

$$|C_1| \leq \frac{|rx|(2|q| + L)}{\cosh^2(px)} \leq \frac{\kappa_2(2\kappa_2 + L)|x|}{\cosh^2(\kappa_1 x)} < \frac{\kappa_2(2\kappa_2 + L)}{\kappa_1} \quad (28)$$

and

$$0 \leq C_2 \leq \frac{2\kappa_2^2 x^2}{\cosh^4(\kappa_1 x)} < \frac{2\kappa_2^2}{\kappa_1^2}. \quad (29)$$

Consequently, for sufficiently small r and η, we have an averaged dynamics

$$s(t + 1) \simeq \epsilon + cs(t), \quad (30)$$

where $\epsilon = \eta \langle C_1 \rangle_x \simeq 0$, and $0 < c = 1 - \eta \langle C_2 \rangle_x < 1$. $\square$

Theorem 1 implies the existence of the second global attraction from the neighbourhood M_w to a neighbourhood of M_{wv}. If $s(t)$ is exactly 0, we observe a total synchronisation with $w_1 = w_2$ and $v_1 = v_2$ on the degenerated subspace M_{wv}. In practice, due to the external inputs x, $s(t)$ approaches 0 and fluctuates around 0 in average.

The multiple degeneration from M_w to M_{wv} is a characteristic behaviour of the stochastic gradient descent dynamics. In the case of the deterministic gradient descent dynamics, the dynamics near M_w in the direction of s is neutral because r converges to 0 monotonically [13]. By contrast, in the stochastic gradient descent dynamics, the dynamics of s can be contracting because r fluctuates around 0 [13]. This phenomenon is comparable with noise-induced synchronisation in random dynamical systems. A typical example of synchronisation is given by uncoupled phase oscillators driven by common noise (see Appendix D). It is known that the Lyapunov exponent of the stochastic phase oscillator is negative while those of the deterministic dynamics is zero [22,24].

In the next section, we show that there exists yet another noise-induced trapping mechanism in the multiply degenerated subspace M_{wv}.

3.2. Local attraction in the degenerated subspace

The equation of motion in M_{wv} is given by the following two-dimensional random map of $w(= w_1 = w_2)$ and $v(= v_1 = v_2)$;

$$w(t+1) = w(t) - \eta x v(t) \frac{2v(t)\tanh(w(t)x) - T(x)}{\cosh^2(w(t)x)}, \quad (31)$$

$$v(t+1) = v(t) - \eta \tanh(w(t)x)[2v(t)\tanh(w(t)x) - T(x)]. \quad (32)$$

or equivalently

$$\begin{pmatrix} w(t+1) - w(t) \\ v(t+1) - v(t) \end{pmatrix} = \eta \cdot g(x; w, v, T(x)), \quad (33)$$

where

$$g(x; w, v, T) = -[2v\tanh(wx) - T(x)] \begin{pmatrix} \dfrac{vx}{\cosh^2(wx)} \\ \tanh(wx) \end{pmatrix}. \quad (34)$$

Theorem 2. *If $T(x) \neq 2v\tanh(wx)$, for a sufficiently small $\eta > 0$, a point (w, v) in the dynamics (31), (32) is either an attracting point or a saddle point.*

Proof. The Jacobian matrix of g is given by

$$J(x; w, v)$$
$$= \begin{bmatrix} -\dfrac{2vx^2[T(x)\tanh(wx) - 3v\tanh^2(wx) + v]}{\cosh^2(wx)} & \dfrac{x[T(x) - vq\tanh(wx)]}{\cosh^2(wx)} \\ \dfrac{x[T(x) - 4v\tanh(wx)]}{\cosh^2(wx)} & -2\tanh^2(wx) \end{bmatrix}, \quad (35)$$

or equivalently

$$J(x; w, v) = -2 \begin{pmatrix} \dfrac{vx}{\cosh^2(wx)} \\ \tanh(wx) \end{pmatrix} \begin{pmatrix} \dfrac{vx}{\cosh^2(wx)} & \tanh(wx) \end{pmatrix}$$
$$- (2v\tanh(wx) - T(x))$$
$$\times \begin{pmatrix} -\dfrac{2vx^2\tanh(wx)}{\cosh^2(wx)} & \dfrac{x}{\cosh^2(wx)} \\ \dfrac{x}{\cosh^2(wx)} & 0 \end{pmatrix}. \quad (36)$$

Let the eigenvalues of J at a point (w, v) be $\mu_-(x; w, v)$ and $\mu_+(x; w, v)$ ($\mu_- \leq \mu_+$) (See Appendix B). Since

$$\det \begin{pmatrix} -\dfrac{2vx^2\tanh(wx)}{\cosh^2(wx)} & \dfrac{x}{\cosh^2(wx)} \\ \dfrac{x}{\cosh^2(wx)} & 0 \end{pmatrix} = -\left(\dfrac{x}{\cosh^2(px)}\right)^2 < 0 \quad (37)$$

for $x \neq 0$, the second term of Eq. (36) has both positive and negative eigenvalues if $2v\tanh(wx) - T(x) \neq 0$. The first term of Eq. (36) is negative semidefinite. Thus, $\mu_-(x; w, v)$ is negative

whenever $2v\tanh(wx) - T(x) \neq 0$. Therefore, for sufficiently small η, the point (w, v) is attracting when $\mu_+(x; w, v)$ is non-positive; otherwise, it is a saddle. □

Theorem 2 implies that depending on x, each point in M_{wv} may be weakly attracting. The dynamics near (w, v) on M_{wv} is characterised by the sign of $\mu_+(x; w, v)$ as long as η is small.

As an example, we investigate the dynamics near $(w, v) = (1/2, 1/2)$ by computing the eigenvalues $\mu_\pm(x; 1/2, 1/2)$ of $J(x; w, v)$ explicitly. We see from Fig. 3 (left) that the dynamics is attracted to the point $(w, v) = (1/2, 1/2)$ when the fluctuation σ^2 is sufficiently large. Put differently, the strong noise may "stabilise" the dynamics near (w, v) in M_{wv}. As a result, the residual time near M_{wv} is extended and a stronger plateau phenomenon is observed.

Fig. 3 (right) shows the numerically computed probability distribution

$$\pi(x; w, v) = \text{Prob}[\mu_+(x; w, v) \leq 0], \quad (38)$$

with $\sigma^2 = 1$, where higher values of $\pi(x; w, v)$ correspond to the darker tones. The red curve C indicates the approximated one-dimensional valley formed by steep gradients of the averaged potential. In this case, C includes a local minimum of the averaged potential. The dark grey region, where $\pi(x; w, v)$ is large, may cause stronger trapping dynamics on M_{wv}. On the contrary, with smaller fluctuation $\sigma^2 = 0.1$, the grey attracting region in Fig. 3 (right) disappears because most points in M_{wv} are saddle points. Thus, there exists another type of noise-induced phenomenon, i.e., the emergence of an attracting region near M_{wv} due to the large fluctuation σ^2, resulting in a substantial extension of the escape time from M_{wv}. In the next section, we discuss the escape time from $M_{w,v}$ phenomenologically.

4. Noise-induced degeneration

Summarising the main results of the previous section, noise-induced degeneration and plateau phenomena emerge through the following three processes;

1. The first global attraction to a neighbourhood of the degenerated subspace M_w. (Appendix A)
2. The second global attraction from the neighbourhood M_w to a neighbourhood of the multiply degenerated subspace M_{wv}. (Theorem 1)
3. The third weak attraction, by the attracting regions in the neighbourhood of M_{wv}. (Theorem 2)

To examine the above three-step scenario, we made numerical experiments to compute the global dynamics and the escape time from the multiply degenerated subspace.

4.1. Global dynamical phenomena

The global dynamics of stochastic gradient descent in Θ is shown in Fig. 4. It depicts converging dynamics to pullback attractors [25] (see Appendix C) in Fukumizu–Amari model with $T(x) = 2\tanh(x) - \tanh(4x)$, $\eta = 0.1$, and $\sigma^2 = 0.1, 1.0$. The red and blue dots represent paths of (w_1, w_2) and (v_1, v_2), respectively, starting from different initial conditions. Both dynamics are plotted together in each panel. The grey points correspond to the optimal attractors θ^*. The degenerated subspaces $w_1 = w_2$ and $v_1 = v_2$ are depicted as a single line. In the case of $\sigma^2 = 1.0$, clear noise-induced degeneration, i.e. $|w_1 - w_2| \to 0$ near M_w at $\tau = 10\,000$, followed by $|v_1 - v_2| \to 0$ near M_{wv} at $\tau = 30\,000$, are observed. Due to this type of noise-induced degeneration, about 20 percentage of the sample paths stay near the attracting region in M_{wv} for an extremely long time, which is shown in a dashed circle at $\tau = 100\,000$.
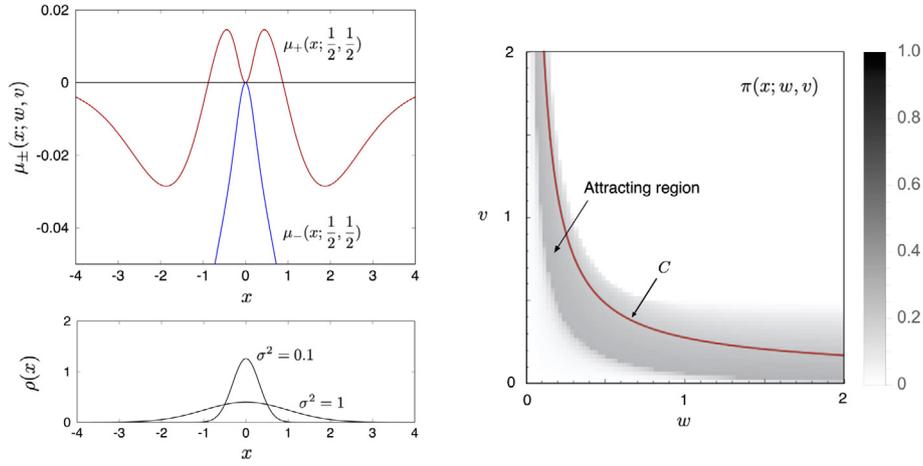
Fig. 3. (left) The eigenvalues $\mu_+(x; 1/2, 1/2)$ (red), and $\mu_-(x; 1/2, 1/2)$ (blue) of $J(x; 1/2, 1/2)$ as well as the distribution $\rho(x)$ are depicted as functions of x. The parameters are set as $T(x) = 2\tanh(x) - \tanh(4x)$, $\eta = 0.1$, and $\sigma^2 = 0.1, 1$. When the fluctuation σ^2 is small, x is frequently sampled near zero, and the point $(w, v) = (1/2, 1/2)$ is a saddle point; otherwise, it is an attracting point. (right) The probability $\pi(x; w, v) = \text{Prob}[\mu_+(x; w, v) \leq 0]$ for $\sigma^2 = 1$ plotted in $[0, 2]^2$ on M_{wv}. The red curve C indicates the valley formed by steep gradients of the averaged potential (see Appendix A). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Escape time from the degenerated subspace

We numerically compute the average escape time from the attracting region on M_{wv} $(w, v > 0)$ for several values of σ^2. The escape time from the region $[-2, 2] \times [-2, 2]$, which includes the multiply degenerated subspace M_{wv} $(w, v > 0)$, is averaged over 10^5 initial conditions and 10^3 noise realisations. In Fig. 5, the average escape time τ^* as a function of the fluctuation size σ^2 is depicted in log–log plot. One can see that larger fluctuation size induces longer escape time because of noise-induced degeneration becoming effective around $\sigma^2 = 0.15$. When noise is weak, around $\sigma^2 = 0.004$, τ^* is large, and when noise is strong, around $\sigma^2 = 0.15$, τ^* is large again. In the intermediate scale, an optimal fluctuation size around $\sigma^2 \simeq 0.07$ exists to minimise the escape time from the degenerated subspace. Thus, the classical Kramer's escape view based on Brownian particle in a potential is insufficient, and the pathwise analysis in terms of random dynamical systems theory is needed.

4.3. Random attractors in stochastic gradient descent

If there exists an attractor A on a subspace M, and A is not an attractor in the full space Θ, it is called a relative attractor in Θ [26]. Thus, if there exists an attractor on M_{wv}, it is a relative attractor. If there is no random compact invariant set in a random dynamical system, all random attractors are random Milnor attractors [27]. Our conjecture is that both relative attractors (or attracting region) in the multiply degenerated subspaces M_{wv} and the optimal attractors in Θ are random Milnor attractors. The dynamics of learning can come and go with each attractor in a long time scale.

In general, non-optimal stable random attractors may exist in stochastic gradient descent learning with small η, because $T(x)$, $f(x; \theta)$, and $\nabla_\theta l(x; \theta)$ are bounded for any $\rho(x)$. In these cases, if an initial point $\theta(0)$ does not belong to the effective basins of the optimal attractors θ^*, the attracting region near $\theta^\dagger$ may become a stable random attractor, which is completely separated from the optimal attractor θ^*, and the orbits can stay there for an arbitrarily long time.

4.4. An extension to larger networks and applications to practical problems

Our results can be extended to a larger three-layer perceptron with N hidden-layer neurons and L dimensional output neurons.

We consider a parameterised function

$$f_i(x; \theta) = \sum_{k=1}^{N} v_k \tanh(w_k x) \quad (i = 1, \ldots, L)$$

where

$$\theta = (w_1, \ldots, w_N, v_1 \ldots, v_N).$$

Near the subspaces defined by

$$w_1 = \cdots = w_j = w_a, \quad w_{j+1} = \cdots = w_M = w_b$$
$$(j = 1, \ldots, M, \ 1 < M < N),$$

$$v_1 = \ldots = v_j = v_a, \quad v_{j+1} = \ldots = v_M = v_b$$
$$(j = 1, \ldots, M, \ 1 < M < N),$$

we have a degenerated function

$$f_i(x; \theta) = j v_a \tanh(w_a x) + (M - j) v_b \tanh(w_b x)$$
$$+ \sum_{k=M+1}^{N} v_k \tanh(w_k x)$$
$$= M g(x; w_a, w_{,b}, v_a, v_b) + \sum_{k=M+1}^{N} v_k \tanh(w_k x), \quad (39)$$

where $i = 1, \ldots, L$, $j = 1, \ldots, M$, and

$$g(x; w_a, w_{,b}, v_a, v_b) = \frac{j}{M} v_a \tanh(w_a x) + \frac{M - j}{M} v_b \tanh(w_b x).$$

Thus, noise-induced degeneration in learning the function $g(x; w_a, w_b, v_a, v_b)$ occurs again in the same way. We have many of such degenerated subspaces hierarchically embedded in the entire parameter space Θ. Indeed, it is known that these singular structures are inevitable, generic properties of neural networks [12].

From the viewpoint of nonlinear physics, the plateau phenomena, noise-induced degeneration, and noise-induced synchronisation between two variables (w_1, v_1) and (w_2, v_2) are all understood as the same dynamical structure. Since synchronisation is a universal phenomenon, they are expected to be observed in a broad class of dynamical systems, and in a broad class of online learning. If the system has a certain local symmetry, synchronisation almost always occurs in that locally symmetric space, and the presented trapping phenomena near the multiply degenerated
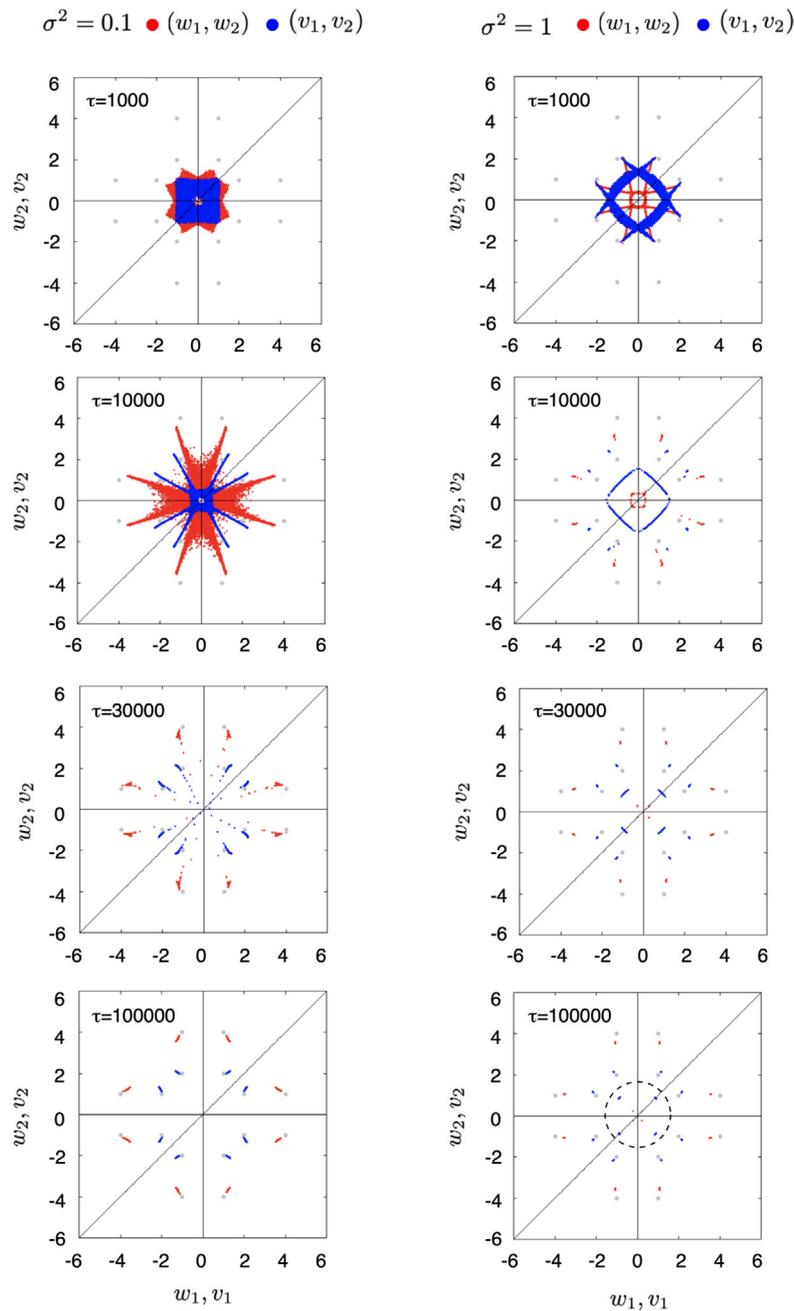
Fig. 4. Finite time pullback attractors (see Appendix C) with the pullback time $\tau = 1000$, $\tau = 10\,000$, $\tau = 30\,000$, and $\tau = 100\,000$ in the full space Θ. Parameters are set as $T(x) = 2\tanh(x) - \tanh(4x)$, $\eta = 0.1$, and $\sigma^2 = 0.1$ (left) and $\sigma^2 = 1.0$ (right). The red and blue dots represent paths of (w_1, w_2) and (v_1, v_2), respectively, starting from different initial conditions. Both dynamics are plotted together in each panel. The grey points correspond to the optimal attractors θ^*. The degenerated subspace $w_1 = w_2$ and $v_1 = v_2$ are depicted as a single line. A typical noise realisation $\{x\}$ is fixed and dynamics is developed with 10^5 different initial conditions $\theta(0) \in [-1, 1]^4$. When $\sigma^2 = 1.0$, the trapping dynamics near M_{wv} is observed in the attracting region indicated by a dashed circle. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

subspace may occur and possibly affect global dynamics. In some cases, the state space degeneration is rather useful, because, if there is no degeneration, we are faced with the problem of over-fitting, which is caused by explosion of the effective degrees of freedom. Thus, it is important to study the trade-off between degeneration and explosion of the effective degree of freedom in terms of nonlinear physics.

In practice, unless the learning requires the full dimension of the parameters, we need degeneration to some extent to avoid over-fitting, although that causes the possible strong plateau phenomena in the learning dynamics. We may choose an "optimal

fluctuation size" to sample the input data x, which is exemplified in Fig. 5, and which, however in general, depends on the complexity of problems and models.

5. Conclusion

In Fukumizu–Amari model, there exist characteristic fluctuation sizes of the training data, with which the dynamics shows strong plateau phenomena. Starting from an initial point in the full space, the dynamics of learning is attracted by a degenerated subspace by the gradient, and then, is attracted by a multiply degenerated subspace by noise-induced synchronisation. We
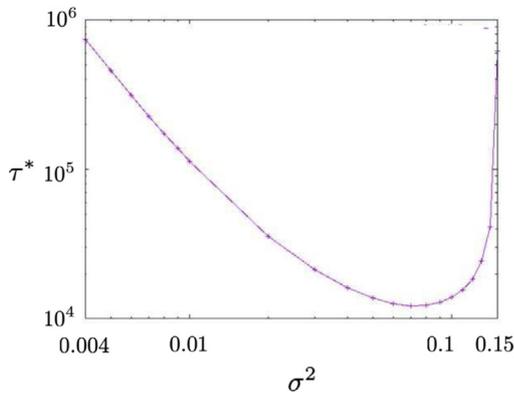
Fig. 5. The averaged escape time from $[-2, 2] \times [-2, 2]$, τ^* as a function of σ^2 in log–log plot. The other parameters are same as in the numerical computation in Fig. 4. Stronger noise induces longer escape time because of noise-induced degeneration. An optimal fluctuation size is $\sigma^2 \simeq 0.07$.

called the second attraction noise-induced degeneration, which is not observed in deterministic gradient descent. Furthermore, when noise is strong, near the multiply degenerated subspace, attracting regions emerge and the residual time near there can be extremely long.

In the attracting region in the multiply degenerated subspace, an optimal fluctuation size exists to minimise the escape time, and stronger noise induces longer escape time. This implies that the classical Kramer's escape view is insufficient, and the path-wise analysis in terms of random dynamical systems theory is needed.

Noise-induced degeneration presented in this paper is expected to be one of the key elements to understand advantages of variety of machine learning to control the appropriate size of the degrees of freedom. As pointed out in [12], the degeneration may be weaken in perceptrons with many hidden layers, which could explain in part the advantage of deep learning. We here add an opposite possibility: the existence of many hidden layer may result in stronger degeneration, which could explain the reason why the over-parameterised neural networks work well.

Further studies on stability of random attractors, geometry of random basins, and bifurcations of stochastic gradient descent of larger neural networks will be reported elsewhere. Our approach would shed new light on various problems in machine learning from a viewpoint of random dynamical systems theory.

CRediT authorship contribution statement

Yuzuru Sato: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualisation. **Daiji Tsutsui:** Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing, Visualisation. **Akio Fujiwara:** Methodology, Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Appendix A. Local minima in the averaged dynamics

In this paper, we have treated $\tanh(\cdot)$ as an activation function. However, in order to analyse the averaged potential, we herein use the function $h(u) := \mathrm{erf}(x/\sqrt{2}) := \sqrt{2/\pi} \int_0^x e^{-t^2/2} dt$ as an approximation of $\tanh(\cdot)$ to understand the qualitative behaviour of the dynamics. According to [28], for a network

$$f(x; \boldsymbol{\theta}) = v_1 h(w_1 x) + v_2 h(w_2 x), \tag{A.1}$$

and a target function which is denoted as

$$T(x) = \nu_1 h(\omega_1 x) + \nu_2 h(\omega_2 x), \tag{A.2}$$

for some values $\nu_1, \nu_2, \omega_1, \omega_2 \in R$, the averaged potential is given by

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= E_x \left[\frac{1}{2} \left(f(x; \boldsymbol{\theta}) - T(x) \right)^2 \right] \\
&= \frac{1}{\pi} \sum_{i,j=1}^{2} v_i v_j \Phi(w_i, w_j) \\
&\quad - \frac{2}{\pi} \sum_{i,a=1}^{2} v_i \nu_a \Phi(w_i, \omega_a) + \text{const},
\end{aligned} \tag{A.3}
$$

$$\Phi(\zeta_1, \zeta_2) := \arcsin \left(\frac{\sigma^2 \zeta_1 \zeta_2}{\sqrt{1 + \sigma^2 \zeta_1^2} \sqrt{1 + \sigma^2 \zeta_2^2}} \right). \tag{A.4}$$

When $w_1 \neq 0$ and $w_2 \neq 0$, the function $L(\boldsymbol{\theta})$ is quadratic in (v_1, v_2), and thus has a minimiser. In particular, when $w_1 = w_2 = w$ is fixed and $v_1 = v_2 = v$,

$$L^*(w, v) := L(\boldsymbol{\theta}) = \frac{4}{\pi} \left(v^2 \Phi(w, w) - v \sum_{a=1}^{2} \nu_a \Phi(w, \omega_a) \right) + \text{const}. \tag{A.5}$$

takes the minimum at

$$v^*(w; \sigma^2) = \frac{\sum_a \nu_a \Phi(w, \omega_a)}{2\Phi(w, w)}. \tag{A.6}$$

Hence, a minimiser of L^* lies in the one-dimensional valley $\{(w, v^*(w; \sigma^2)) | w \in R\}$.

For $(\omega_1, \omega_2, \nu_1, \nu_2) = (1, 4, 2, -1)$ and $\sigma^2 = 1$, in particular, we have the red curve C in Fig. 3 (right) as

$$v^*(w; 1) = \frac{\left(2 \arcsin \left(\frac{w}{\sqrt{2}\sqrt{1+w^2}} \right) - \arcsin \left(\frac{4w}{\sqrt{17}\sqrt{1+w^2}} \right) \right)}{\left(2 \arcsin \left(\frac{w^2}{1+w^2} \right) \right)}. \tag{A.7}$$

Appendix B. Eigenvalues of Jacobian of the dynamics in the multiply degenerated subspace

The Jacobian matrix of g on the multiply degenerated subspace M_{wv} is given by

$$J = \begin{bmatrix} -\dfrac{2vx^2 \left[T(x)\tanh(wx) - 3v\tanh^2(wx) + v \right]}{\cosh^2(wx)} & \dfrac{x[T(x) - vq\tanh(wx)]}{\cosh^2(wx)} \\ \dfrac{x[T(x) - 4v\tanh(wx)]}{\cosh^2(wx)} & -2\tanh^2(wx) \end{bmatrix}. \tag{B.1}$$

The eigenvalues of J are given by

$$\mu_\pm(x; w, v) = A \pm \frac{1}{8 \cosh^4(wx)} \sqrt{B}, \tag{B.2}$$

where

$$A = -1 + \frac{1 + 2v^2 x^2 - vx^2 \tanh(wx)T(x)}{\cosh^2(wx)} - \frac{3v^2 x^2}{\cosh^4(wx)}, \tag{B.3}$$

$$\begin{aligned} B = {}& 32x^2 \cosh^2(wx)C + \\ & \left(-1 + 16v^2 x^2 - 8v^2 x^2 \cosh(2wx) \right. \\ & \left. + \cosh(4wx) + 4vx^2 \sinh(2wx)T(x) \right)^2 \end{aligned} \tag{B.4}$$

$$\begin{aligned} C = {}& 8v^2(2 + \cosh(2wx)) \sinh^2(wx) - v(6 \sinh(2wx) \\ & + \sinh(4wx))T(x) + 2 \cosh^2(wx)T(x)^2. \end{aligned} \tag{B.5}$$

Appendix C. Pullback attractors in random dynamical systems

Let θ act on the probabilistic space of noise realisations Ω, and $\theta_t \omega$ be the path taken at time t by the noise realisation $\omega \in \Omega$. The random dynamical system is represented by the pair (θ, ϕ), where ϕ denotes the dynamics in the state space X, driven by a noise realisation $\theta_t \omega$. The *pullback attractor* $A(t, \omega)$ of a random dynamical system is defined as a random invariant set of X that satisfies

$$\lim_{\tau \to \infty} \text{dist}(\phi(\tau, \theta_{t-\tau}\omega)B, A(t, \omega)) = 0, \tag{C.1}$$

for any bounded set $B \subset X$, where $\text{dist}(C, D)$ denotes the Hausdorff distance between two subsets C and D of X [29].

We call the following τ-pullback image of B as *finite time pullback attractor* or *τ-pullback attractor*, which is given by

$$\widetilde{A}_\tau^B(t, \omega) = \phi(\tau, \theta_{t-\tau}\omega)B, \tag{C.2}$$

where τ is called pullback time. For a given τ, the set $\widetilde{A}_\tau^B(t, \omega)$ represents a finite space–time structure, which may include transient orbits and densities. Each invariant sets in Fig. 4 are finite time pullback attractors with pullback time τ.

Appendix D. Noise-induced synchronisation

A stochastic phase oscillator is given by

$$d\phi = \omega dt + \sin \phi \circ dW_t, \tag{D.1}$$

in Stratonovich form, where ω is a constant, $\phi \in (0, 2\pi]$ is phase on circle, and W_t is the Wiener process with $dW_t \sim N(0, \sigma^2)$. The linearisation along a fixed solution ϕ is given by

$$d\psi = \cos \phi \cdot \psi \circ dW_t, \tag{D.2}$$

Let $r = \log |\psi|$, then we have

$$dr = \cos \phi \circ dW_t, \tag{D.3}$$

or, equivalently in Ito form,

$$dr = -\frac{\sigma^2}{2} \sin^2 \phi dt + \cos \phi dW_t. \tag{D.4}$$

Thus, the Lyapunov exponent λ of (D.1) is given by

$$\lambda = \lim_{T \to \infty} \frac{r(T)}{T} = -\lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{\sigma^2}{2} \sin^2 \phi dt. \tag{D.5}$$

Assuming that the fluctuation σ^2 is small, the dynamics is ergodic, and the invariant density is approximately uniform on circle, we have

$$\lambda \simeq -\frac{1}{2\pi} \int_0^{2\pi} \frac{\sigma^2}{2} \sin^2 \phi d\phi = -\frac{\sigma^2}{4} < 0. \tag{D.6}$$

References

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

[2] Stephan Mandt, Matthew D. Hoffman, David M. Blei, Stochastic gradient descent as approximate bayesian inference, J. Mach. Learn. Res. 18 (2017) 4873–4907.

[3] George Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signals Systems 2 (4) (1989) 303–314.

[4] Herbert Robbins, Sutton Monro, A stochastic approximation method, Ann. Math. Stat. 22 (1951) 400–407.

[5] Shun-ichi Amari, A theory of adaptive pattern classifiers, IEEE Trans. Electron. Comput. 3 (1967) 299–307.

[6] Peter Riegler, Biehl Michael, On-line backpropagation in two-layered neural networks, J. Phys. A: Math. Gen. 28 (1995) L507–L513.

[7] Nam Jin Huh, Jong-Hoon Oh, Kukjin Kang, On-line learning of a mixture-of-experts neural network, J. Phys. A: Math. Gen. 33 (2000) 8663–8672.

[8] John Milnor, On the concept of attractor, in: The Theory of Chaotic Attractors, Springer, 1985, pp. 243–264.

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

[10] Florent Cousseau, Tomoko Ozeki, Shun-ichi Amari, Dynamics of learning in multilayer perceptrons near singularities, IEEE Trans. Neural Netw. 19 (2008) 1313–1328.

[11] Haikun Wei, Jun Zhang, Florent Cousseau, Tomoko Ozeki, Shun-ichi Amari, Dynamics of learning near singularities in layered networks, Neural Comput. 20 (2008) 813–843.

[12] Shun-ichi Amari, Tomoko Ozeki, Ryo Karakida, Yuki Yoshida, Masato Okada, Dynamics of learning in mlp: Natural gradient and singularity revisited, Neural Comput. 30 (1) (2018) 1–33.

[13] Daiji Tsutsui, Center manifold analysis of plateau phenomena caused by degeneration of three-layer perceptron, Neural Comput. 32 (2020) 683–710.

[14] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, Ma Jinwen, The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects, in: ICML, 2019.

[15] Max Welling, Yee Whye Teh, Bayesian learning via stochastic gradient Langevin dynamics, in: ICML, 2011, pp. 681–688.

[16] Sepp Hochreiter, Jürgen Schmidhuber, Flat minima, Neural Comput. 9 (1997) 1–42.

[17] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, Ping Tak Peter Tang, On large-batch training for deep learning: Generalization gap and sharp minima, in: ICLR, 2017.

[18] Elad Hoffer, Itay Hubara, Daniel Soudry, Train longer, generalize better: Closing the generalization gap in large batch training of neural networks, in: NIPS, Vol. 30, 2017, pp. 1731–1741.

[19] Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, Thomas Hofmann, Escaping saddles with stochastic gradients, 2018, arXiv preprint arXiv:1803.05999.

[20] Chen Xing, Devansh Arpit, Christos Tsirigotis, Yoshua Bengio, A Walk with SGD, 2018, arXiv preprint arXiv:1802.08770v4.

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

[22] Jun-nosuke Teramae, Dan Tanaka, Robustness of the noise-induced phase synchronization in a general class of limit cycle oscillators, Phys. Rev. Lett. 93 (20) (2004) 204103.

[23] Yuzuru Sato, Thai Son Doan, Nguyen Van The, Hoang The Tuan, An analytical proof for synchronization of stochastic phase oscillator, 2018, arXiv preprint arXiv:1801.02761.

[24] Yuzuru Sato, Thai Son Doan, Jeroen S.W. Lamb, Martin Rasmussen, Dynamical characterization of stochastic bifurcations in a random logistic map, 2018, arXiv preprint arXiv:1811.03994.

[25] Yuzuru Sato, Mickaël D. Chekroun, Michael Ghil, Convergence rate of snapshot attractors to random strange attractors, 2020, submitted for publication.

[26] Joseph D. Skufca, James A. Yorke, Bruno Eckhardt, Edge of chaos in a parallel shear flow, Phys. Rev. Lett. 96 (17) (2006) 174101.

[27] Peter Ashwin, Minimal attractors and bifurcations of random dynamical systems, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 455 (1987) (1999) 2615–2634.

[28] Michael Biehl, Holm Schwarze, Learning by online gradient descent, J. Phys. A 28 (1995) 643–656.

[29] Mickaël D. Chekroun, Eric Simonnet, Michael Ghil, Stochastic climate dynamics: Random attractors and time-dependent invariant measures, Physica D 240 (21) (2011) 1685–1700.