Mathematical Neuroscience — An Overview

Shun-ichi Amari
RIKEN Brain Science Institute

Abstract

Mathematical neuroscience studies the fundamental principles of information processing that has realized in the brain through long history of evolution. We use simple and tractable models of neural networks and establish rigorous mathematical theories to study the performances of the models. This reveals the secrets why the brain works well by using a large number of stochastically fluctuating neurons and parallel dynamics of mutual interactions. We give a number of examples of mathematical neuroscience approaches, including statistical neurodynamics, associative memory, neural field theory and dynamics of learning in a hierarchical system including singularity.

1 Introduction

The brain is a very complex system consisting of a huge number of neurons. It forms a system composed of complicated networks, storing information as memories and processing information to make decision. One may say that it is a most sophisticated system that the nature has ever created. There are many approaches of research to the brain.

Brain science searches for the mechanisms and functions of the brain. Researchers study, from the microscopic point of view, the role of genes and proteins to understand the structure and function of neurons by using the techniques of molecular biology. They also use molecular imaging techniques. They search for the molecular mechanisms of synapses and their plastic changes. Systems neuroscientists observe the activities of neurons of behaving animals to study how information is represented in the brain and how it is processed.

Recent developments of measurement technology makes it possible to observe the activities of a human brain by using fMRI, PET, NIRS, EEG, MEG and even multi-electrodes. Cognitive neuroscience makes use of these new technologies. Computational neuroscience builds models of neural networks and studies their behaviors by theory or computer simulations. Researchers try to build realistic models by taking experimental details into account, so that they can explain the behaviors of actual networks by theory. See, for example, Dayan and Abott, 2001.

The brain is an organ which a living animal is equipped with for information processing. A living object is material, but is special in the sense that it has an ability of self-reproduction. It has emerged through the long history of evolution. Information, expressed in terms of DNA molecules, is necessary for this. Hence, a living object keeps information in material. Further, the brain processes information in the environment and decides actions. Higher-order information processing ability as well as structured memory are necessary for this purpose.

The brain has been developed through long years of evolution, and it processes information in neuronal networks. Material has become to be able to keep information and to process it. From this point of view, one may say that information utilizes material to develop itself. The brain sits at the cross point of material and information. It is important to study the material basis of the brain. At the same time, it is necessary to study the principles of information which have realized in material.

Computational neuroscience builds realistic models as faithful to experimental details as possible. This is necessary for explaining actual processes taking place in the brain. The results can then be compared with experimental findings. This is the orthodox way of theory.

However, there would be another approach. The principles of brain information processing are obviously very different from those of computers. Information is distributed over a huge number of neurons and processed by their dynamic interactions. We need to understand the principles which guarantee efficient, reliable and swift information processing in such a distributed system with dynamic interactions. The nature has found such principles through long years of evolution and the brain is a realization of the principles.

We need abstract models of neural networks in order to reveal the principles of parallel and dynamic information processing, which are simple enough to be able to analyze their capabilities deeply. Mathematical neuroscience is an area that studies the possibilities, limitations and performances of information processing by using distributed parallel dynamics (Amari, 1990). It aims at establishing mathematical theories, where models are as simple as possible so long as they do not miss essential features. The nature has realized these principles by using realistic neurons, so that the brain is far more complex than the mathematical models, but they share common principles. If we understand such principles, we can then proceed further to search for realistic models and to find evidences of realization of such principles in the brain.

The present paper shows examples of mathematical neurosciences and its methods briefly.

2 Statistical Neurodynamics (1): Simple Random Networks

Let us consider a very simple model of mathematical neuron, called the McCulloch-Pitts neuron. Let us further consider a network composed of n neurons, and let $x_i(t), \cdots, x_n(t)$ be the outputs of neurons at discrete time t. We assume that their values are 1 or -1, representing firing or non-firing of the neurons. Neuron i receives inputs from other neurons, which are $x_1(t-1), \cdots, x_n(t-1)$. Let w_{ij} be the connection weight from neuron j to neuron i. We neglect the threshold term for simplicity and also external inputs. Then, the state of each neuron changes over time as

$$x_i(t) = \operatorname{sgn}\left\{\sum_j w_{ij} x_j(t-1)\right\}, \tag{1}$$

where sgn is the signature function,

$$\operatorname{sgn}(u) = \begin{cases} 1, & u \geq 0, \\ -1, & u < 0. \end{cases} \tag{2}$$

Let us use abbreviated notations: We use vector

$$\boldsymbol{x}(t) = [x_1(t) \cdots x_n(t)] \tag{3}$$

for the state of the network at time t, matrix $W = (w_{ij})$ for connections, and T_W for state transition operator which is nonlinear, so that (1) is rewritten as

$$\boldsymbol{x}(t) = T_W \boldsymbol{x}(t-1). \tag{4}$$

We study the dynamics (4) described by the state transition operator T_W.

The properties of T_W obviously depends on the matrix W. Statistical neurodynamics (Amari, 1971; Amari, 1974; Amari, Yoshida and Kanatani, 1977) assumes that W is randomly generated from a probability distribution $p(W)$, and searches for the properties that hold for almost all networks having W generated subject to $p(W)$ when n is large. We give two examples.

When w_{ij} are iid (identical and independently distributed) random variables subject to a distribution $p(w)$, it is easy to apply the central limit theorem and the law of large numbers, when n is large. Here, we assume that $p(w)$ is the Gaussian distribution with mean $\bar{w}/\sqrt{n}$ and variance σ_w^2. We can easily treat a general distribution by a similar method. What are common aspects of dynamics of these networks? To answer this question, we introduce macroscopic variables. The first one is the activity level (Amari, 1971),

$$X(\boldsymbol{x}) = \frac{1}{n}\sum x_i. \tag{5}$$

Then, the macroscopic variable at time t is

$$X(t) = \frac{1}{n}\sum x_i(t) \tag{6}$$

and we have a macroscopic state transition law of X, if we could find a function F which satisfies

$$X(t) = F\left\{X(t-1)\right\}. \tag{7}$$

We search for such a function F. To this end, we use the fact that, for a given non-random $x_1, \cdots, x_n$,

$$u_i = \sum_j w_{ij} x_j \tag{8}$$

are independently and identically distributed Gaussian random variables with mean and variance,

$$E\left[u_i\right] = \sqrt{n}\bar{w}X, \quad V\left[u_i\right] = n\sigma^2. \tag{9}$$

Hence, by the law of large numbers, we have

$$F(X) = \operatorname{Prob}\left\{u_i > 0\right\} - \operatorname{Prob}\left\{u_i < 0\right\}. \tag{10}$$

This is calculated as

$$F(X) = \Phi\left(\frac{\bar{w}X}{\sigma}\right), \tag{11}$$

where

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-u}^{u} \exp\left\{-\frac{t^2}{2}\right\} dt. \tag{12}$$

It is easy to prove that, given $\boldsymbol{x}(0)$ at time 0 of which macroscopic variable $X(0)$ is given by

$$X(0) = \frac{1}{n} \sum x_i(0), \tag{13}$$

the next macroscopic variable $X(1)$ is

$$X(1) = \frac{1}{n} \sum x_i(1) = F\left\{X(0)\right\} \tag{14}$$

with a probability close to 1 for almost all randomly generated networks as n tends to infinity. However, there is difficulty for proving that

$$X(2) = F\left\{X(1)\right\}, \tag{15}$$

or more generally

$$X(t) = F\left\{X(t-1)\right\}. \tag{16}$$

This is because, even we show, for a fixed $\boldsymbol{x}$,

$$X' = X\left(T_W \boldsymbol{x}\right) = F\left(X(\boldsymbol{x})\right), \tag{17}$$

this does not automatically guarantees that

$$X'' = X\left(T_W T_W \boldsymbol{x}\right) = F\left\{F(X)\right\}. \tag{18}$$

In order to derive (15) or (17), we used the fact that w_{ij} are independent. However when we calculate (18) by using $T_W T_W \boldsymbol{x}$, the two operators are no more independent. In other words, when we evaluate

$$v_i(t) = \sum w_{ij} x_j(t), \tag{19}$$

$x_j(t)$ are no more constants but depend on W. Therefore, when we deal with a macroscopic dynamics of a random networks, we need to verify if the macroscopic equation of type (18) is valid or not. Computer stimulants show that the marcoscopie dynamics (16) holds. This is discussed in Rozonoer (1969, A, B, C) and Amari et al. Amari (1974), Amari et al (1977). Its theoretical justification is still an unsolved problem.

We can extend the macroscopic dynamics to a network consisting of a number of subnetworks, each of which consists of random networks of different types of neurons. The macroscopic variables are then $\boldsymbol{X} = (X_1, \cdots X_m)$, each X_α, $\alpha = 1, \cdots, m$, denoting the macroscopic activity of subnetwork α. We have the following macroscopic equation

$$X_\alpha(t+1) = \Phi\left(\frac{\sum \bar{w}_{\alpha\beta} X_\beta}{\sigma_\alpha}\right). \tag{20}$$

We can prove that these dynamical equations show monostable, multistable, oscillatory and chaotic behaviors depending on the probability distributions of $w_{ij}^{(\alpha)}$ within a subnetwork and of $w_{ij}^{(\alpha\beta)}$ across subnetworks (Amari, 1971, 1972A).

It is interesting to search for other types of macroscopic variables. Let us study the stability of the state transition T_W of a simple random network. Assume that microscopic state $\boldsymbol{x}$ is mapped to $\boldsymbol{y}$ by T_W

$$\boldsymbol{y}' = T_W \boldsymbol{x}. \tag{21}$$

When $\boldsymbol{x}$ is disturbed and changes to $\boldsymbol{x}'$, how much change is caused in the resultant output $\boldsymbol{y}' = T_W \boldsymbol{x}'$? This is the problem of stability of T_W. We introduce the Hamming distance of two microscopic states

$$D(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{2n} \sum |x_i - x_i'| \tag{22}$$

to show their difference. The problem is to find the relation

$$D' = K(D) \tag{23}$$

where

$$D' = D\left(T_W \boldsymbol{x}, T_W \boldsymbol{x}'\right). \tag{24}$$

We can obtain this relation explicitly when $\bar{w} = 0$ and $\sigma^2 = 1$. The answer given by Amari (1974) is

$$K(D) = \frac{2}{\pi} \sin^{-1} \sqrt{D} \tag{25}$$

The state transition of a random network has a very specific property: When D is small,

$$K(D) \approx \frac{2}{\pi} \sqrt{D}. \tag{26}$$

Since the derivative of $K(D)$ at $D = 0$ is infinity, this means that a small deviation is enlarged extraordinary. So this type of network is convenient for checking small deviations of inputs, since the difference is expanded in the outputs.

If the macroscopic dynamics

$$D_{t+1} = K\left(D_t\right) \tag{27}$$

holds for two serieses $\boldsymbol{x}_t = T_W^t \boldsymbol{x}_0$ and $\boldsymbol{x}_t' = T_W^t \boldsymbol{x}_0'$ and

$$D_t = D\left(\boldsymbol{x}_t, \boldsymbol{x}_t'\right), \tag{28}$$

we can study the state transition diagram of a random network. Since the number of states is 2^n, we have a state-transition graph having 2^n nodes which correspond to 2^n $\boldsymbol{x}$'s. From each node $\boldsymbol{x}$, there emerges one branch and its destination is $T_W \boldsymbol{x}$. Hence, the graph has 2^n directed branches. The state transition graph is highly different from the random state transition graph where the state transition branches are randomly assigned to each node. First, a random neural network has the small-world property, where a small number of nodes monopolize branch outlets. Let p_i be the probability that a node has i ancestors,

$$p_i = \text{Prob}\left\{\left|T_W^{-1}\boldsymbol{x}\right| = i\right\}. \tag{29}$$

Then, the distribution is subject to the power law

$$p_i \propto \frac{1}{i^{3-\alpha}}, \tag{30}$$

for some $0 < \alpha < 1$. See Amari (1974). The small-world network has a number of interesting properties, and one of them is that the transient period of dynamics is very short, and the dynamics of

$$\boldsymbol{x}(t+1) = T_W \boldsymbol{x}(t) \tag{31}$$

converges quickly to its attractor states.

3 Statistical Neurodynamics (2): Associative Memory

The conventional model of associative memory networks can also be analyzed by statistical neurodynamics. See Amari (1972B), Hopfield (1982) and many others. But it has quite different behaviors from the previous random networks, because the connections w_{ij} are random but not independent. Memorized patterns in this network are represented by the states $\boldsymbol{x}$ which satisfy

$$T_W \boldsymbol{x} = \boldsymbol{x}. \tag{32}$$

Let $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_m$ be m patterns to be memorized in a network such that they are attractors of the dynamics

$$T_W \boldsymbol{\xi}_\alpha = \boldsymbol{\xi}_\alpha, \quad \alpha = 1, \cdots, m. \tag{33}$$

Starting at any initial state $\boldsymbol{x}_0$, it is expected that the state reaches one of the memorized pattern $\boldsymbol{\xi}_\alpha$ by the state transition, finding a memorized pattern. When these patterns are generated randomly, we compose the connection matrix in such a way that

$$w_{ij} = \frac{1}{m}\sum \xi_{\alpha i}\xi_{\alpha j} \tag{34}$$

This is a standard way of associative memory. When these patterns are orthogonal, it is easy to see that the patterns are attractors, satisfying

$$T_W \boldsymbol{\xi}_\alpha = \text{sgn}\left(\frac{1}{m}\sum_\beta \boldsymbol{\xi}_\alpha\left(\boldsymbol{\xi}_\alpha \cdot \boldsymbol{\xi}_\beta\right)\right) = \boldsymbol{\xi}_\alpha. \tag{35}$$

When $\boldsymbol{\xi}'_\alpha$'s are randomly generated subject to independent and identical probabilities with $\text{Prob}\left\{\xi_{\alpha i} = 1\right\} = \text{Prob}\left\{\xi_{\alpha i} = -1\right\} = 1/2$, they are asymptotically orthogonal.

The connection weight matrix W is random, but the component w_{ij} are not independent, because they are composed of (34). The law of stability is interesting in this case. The statistical analysis shows that

$$D' = K(D), \quad K(D) = 1 - \frac{1}{2}\Phi\left\{-\frac{(1-2D)n}{m}\right\}. \tag{36}$$

Therefore, when D is small, $D = K(D)$ is nearly equal to 0. Hence the state transition of recalling a memorized pattern is very stable. It is interesting to show that the state transition graph of an associative memory model has a fractal structure (Amari and Maginu, 1988). Here the basin of attractor has very complicated strange boundaries.

4 Pattern Dynamics in Neural Field

A neural field is a continuous version of neural networks, where neurons are located in a 2-dimensional field like cortices. A neural field consists of layers including different types of neurons. We give a general equation describing the dynamics of excitation in the field,

$$\frac{\partial u_i(\boldsymbol{r},t)}{\partial t} = -u_i(\boldsymbol{r},t) + \sum\int w_{ij}\left(\boldsymbol{r},\boldsymbol{r}'\right) f\left\{u_j\left(\boldsymbol{r}',t\right)\right\} d\boldsymbol{r}' + I_i(\boldsymbol{r},t). \tag{37}$$

Here, $\boldsymbol{r} = (r_1, r_2)$ is coordinates of the field, denoting the locations of neurons, $u_i(\boldsymbol{r},t)$ is the average membrane potential of neurons at location $\boldsymbol{r}$ in the i-th layer at time t,

$$z_i(\boldsymbol{r},t) = f\left[u_i(\boldsymbol{r},t)\right] \tag{38}$$

is the output of neurons whose average membrane potential is u_i and f is called the activation function. The functions $w_{ij}\left(\boldsymbol{r},\boldsymbol{r}'\right)$ are connection weights from the neurons at location $\boldsymbol{r}'$ in the j-th layer to the neurons at location $\boldsymbol{r}$ in the i-th layer. $I_i(\boldsymbol{r},t)$ are external inputs to the neurons of i-th layer at position $\boldsymbol{r}$ at time t minus threshold.

This types of dynamics was proposed by Willson and Cown (1973). A simplified version of neural field used by Amari (1977) is a one-dimensional field of one layer, where the activation function is the step function

$$f(u) = \begin{cases} 1, & u \geq 0, \\ 0, & u < 0, \end{cases} \tag{39}$$

and the connection weight function is homogeneous and isotropic, having lateral-inhibition type,

$$w(r, r') = w(|r - r'|), \tag{40}$$

that is, $w(r)$ is positive when r is small and negative otherwise.

Given initial stimuli, we study how an initial excitation pattern $u(r, 0)$ develops in the dynamics. When $u(r) > 0$ holds in a region R of the field, neurons in the region are excited, $z(r) = 1$, $r \in R$. We call such a region an excited region, that is a region

$$R = \{r | u(r) > 0\}. \tag{41}$$

When an excited region is an interval $R = (a_1, a_2)$, it is called a localized excitation pattern or a bump. We first focus on the dynamics of localized excitation patterns.

Let us consider the case where I is constant over the field. In order to analyze how the excited interval develops over time, we remark its boundaries a_1 and a_2. They change over time so that we write $a_1(t)$ and $a_2(t)$ for the boundaries of the excited interval at time t. The boundaries satisfy

$$u\{a_1(t), t\} = u\{a_2(t), t\} = 0 \tag{42}$$

at t. After a short time dt, the boundaries change and they satisfy

$$u\{a_i(t + dt), t + dt\} = 0, \quad i = 1, 2. \tag{43}$$

We have, by Taylor expansion,

$$\frac{\partial u(a_i, t)}{\partial t} + \frac{\partial u(a_i, t)}{\partial r} \frac{da_i}{dt} = 0. \tag{44}$$

The time derivative of $u(a_i, t)$ satisfies (37), and the integration part is written as

$$\int w(r - r') f\{u(r', t)\} dr' = \int_{a_1}^{a_2} w(r - r') dr', \tag{45}$$

since the activation function is 1 in the excited interval (a_1, a_2) and 0 otherwise. We define the integration of the connection function by

$$W(r) = \int_0^r w(s) ds. \tag{46}$$

Then, we have

$$\int_{a_1}^{a_2} w(a_2 - r') dr' = W(a_2 - a_1). \tag{47}$$

We put the derivatives of the waveform of pattern $u(r, t)$ at the boundaries,

$$c_1 = \frac{\partial u(r_1, t)}{\partial r}, \quad c_2 = -\frac{\partial u(r_2, t)}{\partial t}. \tag{48}$$

Taking these into account, we have the equations describing the dynamics of the boundary points,

$$\frac{dr_i(t)}{dt} = -\frac{1}{c_i} W(r_2 - r_1) + I, \quad i = 1, 2. \tag{49}$$

Let

$$a(t) = a_2(t) - a_1(t), \tag{50}$$

which is the length of the excited interval. It develops as

$$\frac{da(t)}{dt} = \frac{1}{c}\{W(a) + I\}. \tag{51}$$

The equilibrium a is the solution of

$$W(a) + I = 0. \tag{52}$$

Its stability is analyzed (Amari, 1977), and we have the following results:

When

$$W(a) + I = 0 \tag{53}$$

has a solution, it is an equilibrium of the length of an excited region. It is stable when

$$w(a) > 0 \tag{54}$$

is satisfied.

We show that, under a certain condition, we can find a stable equilibrium solution. The field can retain a local excitation pattern without further inputs in such a case. Since the field is homogeneous, such a pattern can exist at any position. Hence the field has an infinite number of attractors (stable equilibrium solutions), which is called a line attractor (Seung, 1988). Assume that an initial stimulus arrives at around position b. The stimulus may be very noisy, distributed around b. Initial excitations are arisen at around b and smoothed by the dynamic interactions in the field. Even after the stimulus disappears, a local excitation pattern remains stably at the position b when the field admits a local excitation pattern. Hence, this mechanism can be used as a working memory. Such a mechanism is believed to be used in memorizing input patterns and processing them. One example is a neural field detecting orientations of input bars and keeping the orientations for processing them further. The neural field theory explains the mechanism of the tuning curve in the visual cortex.

More general neural fields give richer behaviors (Ermentrout and Cowan, 1979). See review paper by Coombes and Owen, 2005. When a field consists of an excitatory and inhibitory layers, the field admits a stable travelling region. Its velocity can be controlled by stimuli given outside so that such phenomena will be useful for dynamic information processing.

A two-dimensional neural field has much richer properties. See, for example, Coombes and Owen, 2005. First, it possesses a localized excitation pattern or a bump as well. The stability analysis of the two-dimensional bump was given in a Japanese book in 1978 (Amari), and later proved by some others recently. It also has interesting static patterns consisting of a number of bumps or of a ring shape. Their dynamic behaviors are much more interesting. It has an expanding ring pattern, spiral pattern, breathing pattern and others. It is shown recently that it has also moving bumps. The collision of moving bumps shows very interesting dynamical behaviors.

The neural field model is an important component of computational neuroscience. Moreover, there are lots of applications of dynamic neural field theory to psychological phenomena and robot navigation.

5 Dynamics of Learning in Multilayer Perceptron and Gaussian Mixture

We study the dynamical properties of on-line learning in a hierarchical system such as multilayer perceptrons and Gaussian mixture radial basis systems. It is known that the behavior of a student system trained by using examples given by the teacher system converges to the optimal value, although one cannot avoid a local optimum. However, it is also known that the dynamics of such a learning system suffers from the so-called plateau phenomenon, that the trajectory of learning is attracted to a plateau which is not optimal. It takes long time for the system to get rid of such a plateau. The present section analyzes the dynamical behavior of on-line learning in the neighborhood of a plateau, based on Wei et al. 2008, Wei and Amari, 2008, Cousseau, Ozeki and Amari, 2008.

Plateaus are given rise to by singularities of the parameter space on which learning takes place. This is caused by the symmetric structure of a hierarchical system and we cannot avoid them (Amari, Park and Ozeki, 2006).

Let us consider a function

$$y = k(\boldsymbol{x}), \tag{55}$$

where $\boldsymbol{x}$ is an n-dimensional vector $\boldsymbol{x} = (x_1 \cdots x_n)$ and y is a scalar. We assume that $\boldsymbol{x}$ is an input to a system and y is its output. In order to realize this function, we use a family of functions

$$F = \{f(\boldsymbol{x}, \boldsymbol{\theta})\} \tag{56}$$

parameterized by

$$\boldsymbol{\theta} = (\theta_1 \cdots \theta_n). \tag{57}$$

Our problem is to estimate $\boldsymbol{\theta}$ from a series of input-output examples $D = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)\}$ which we call the training examples. Here, $\boldsymbol{x}_t$ is generated subject to a probability distribution $q(\boldsymbol{x})$ each time independently, and y_t is the output assumed to be generate by

$$y_t = k(\boldsymbol{x}_t) + n_t, \tag{58}$$

that is, y_t is the output from the true system contaminated by noise n_t subject to the standard Gaussian distribution. We assume it is subject to $N(0,1)$ without loss of generality, by rescaling y dividing it by the standard deviation of the noise.

When the true function $k(\boldsymbol{x})$ from which the training examples are generated belongs to F,

$$k(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta}^*), \tag{59}$$

the problem is to estimate the true value $\boldsymbol{\theta}^*$ from the training examples. When $k(\boldsymbol{x})$ does not belong to F, we search for the parameter $\boldsymbol{\theta}^*$ by which $k(\boldsymbol{x})$ is approximated optimally by $f(\boldsymbol{x}, \boldsymbol{\theta}^*)$. Hence, this is a regression problem for the data D.

We use an on-line learning approach such that an estimator $\boldsymbol{\theta}_t$ at time t is modified to $\boldsymbol{\theta}_{t+1}$ by using a new example $\boldsymbol{x}_t, y_t$ at each discrete time $t, t = 1\ 2, \cdots$. This is a method of sequential estimation or regression, called on-line learning. The rule of modifying the parameter is written as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{g}(\boldsymbol{x}_t, y_t, \boldsymbol{\theta}_t) \tag{60}$$

by using an adequate function $\boldsymbol{g}$. Here, η is called a learning constant, which may depend on t. We study the properties of dynamics of the above on-line learning method.

We consider the following two types of the parametric families of functions as F. One is multilayer perceptrons, written as

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{j=1}^{m} v_j \varphi \left(\sum_{i=1}^{n} w_{ji} x_i \right). \tag{61}$$

Here, the system consists of two layers. The first is the hidden layer and the output of the j-th hidden neuron is a nonlinear function of weighted sum of inputs $\boldsymbol{x}$,

$$z_j = \varphi \left(\sum_i w_{ji} x_i \right) = \varphi \left(\boldsymbol{w}_j \cdot \boldsymbol{x} \right), \tag{62}$$

where φ is a sigmoidal function and and we use

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left\{ -\frac{1}{2} s^2 \right\} ds, \tag{63}$$

as an example. The second layer consists of a single neuron, called the output neuron and the output is a linear summation of the outputs of the hidden neurons,

$$z = f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum v_j z_j. \tag{64}$$

The parameters to be estimated are

$$\boldsymbol{\theta} = (v_1, \cdots, v_m \; ; \; w_{11}, \cdots, w_{mn}). \tag{65}$$

The second example uses Gaussian radial basis functions given by

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum v_j \psi \left(\boldsymbol{x} - \boldsymbol{w}_j \right) \tag{66}$$

where

$$\psi(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^n} \exp\left\{ -\frac{|\boldsymbol{x}|^2}{2} \right\} \tag{67}$$

is the Gaussian function. This is similar to the first system of multilayer perceptrons. The parameters to be estimated are

$$\boldsymbol{\theta} = (u_1, \cdots, v_m \; ; \; \boldsymbol{w}_1, \cdots, \boldsymbol{w}_m). \tag{68}$$

We use the stochastic gradient learning method, minimizing a cost function each time. Here, the cost function $l(\boldsymbol{x}, y, \boldsymbol{\theta})$ is given by a half of the square of the difference of the actual output y and the hypothetical output $f(\boldsymbol{x}, \boldsymbol{\theta})$ which the system with parameter $\boldsymbol{\theta}$ emits for given input $\boldsymbol{x}$. This is easily calculated for given parameter $\boldsymbol{\theta}$ when an input-output example $(\boldsymbol{x}, y)$ is given,

$$l(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{2} \left\{ y - f(\boldsymbol{x}, \boldsymbol{\theta}) \right\}^2. \tag{69}$$

The difference is called the error function,

$$e\left(\boldsymbol{x}, y, \boldsymbol{\theta}\right) = y - f(\boldsymbol{x}, \boldsymbol{\theta}). \tag{70}$$

The true system of emitting y can be described by a conditional probability distribution of emitting y from $f(\boldsymbol{x}, \boldsymbol{\theta})$ with additive noise n (58). Then, the conditional probability distribution of y under the condition that input $\boldsymbol{x}$ is given as

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left\{ y - f(\boldsymbol{x}, \boldsymbol{\theta}) \right\}^2 \right] \tag{71}$$

when the parameter is $\boldsymbol{\theta}$. The cost function (69) is interpreted as the negative of log likelihood

$$l(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\log p\left(y|\boldsymbol{x}, \boldsymbol{\theta}\right). \tag{72}$$

Hence, minimizing the cost function is equivalent to maximizing the log likelihood. This gives the maximum likelihood estimator.

The dynamics of the stochastic gradient on-line learning method is written as

$$\boldsymbol{g}(\boldsymbol{x}, y, \boldsymbol{\theta}) = \nabla l\left(\boldsymbol{x}, y, \boldsymbol{\theta}\right) = -e\left(\boldsymbol{x}, y, \boldsymbol{\theta}\right) \nabla f(\boldsymbol{x}, y, \boldsymbol{\theta}) \tag{73}$$

where ∇ is the gradient with respect to $\boldsymbol{\theta}$,

$$\nabla = \left(\frac{\partial}{\partial \theta_1} \cdots \frac{\partial}{\partial \theta_n} \right). \tag{74}$$

In order to analyze the behavior of dynamics of online learning, we use the average learning equation

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \left\langle \nabla l(\boldsymbol{x}, y, \boldsymbol{\theta}) \right\rangle \tag{75}$$

where $< \;\; >$ denotes the average over all possible input $\boldsymbol{x}$ by using the probability distribution (72). We further use the continuous time approximation, and analyze the continuous time average learning equation,

$$\frac{d\boldsymbol{\theta}(t)}{dt} = -\left\langle \nabla l(\boldsymbol{x}, y, \boldsymbol{\theta}) \right\rangle. \tag{76}$$

Let $S = \{\boldsymbol{\theta}\}$ be the parameter space. When

$$f\left(\boldsymbol{x}, \boldsymbol{\theta}_1\right) \neq f\left(\boldsymbol{x}, \boldsymbol{\theta}_2\right) \text{ for } \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \tag{77}$$

holds, the system is identifiable, because only one $\boldsymbol{\theta}$ corresponds to a function $f(\boldsymbol{x}, \boldsymbol{\theta})$. However, when there exist at least two points for which

$$f\left(\boldsymbol{x}, \boldsymbol{\theta}_1\right) = f\left(\boldsymbol{x}, \boldsymbol{\theta}_2\right), \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \tag{78}$$

holds, the system is not identifiable. Let $R(\boldsymbol{\theta})$ be a subset of S in which all the output function is identical,

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = f\left(\boldsymbol{x}, \boldsymbol{\theta}'\right), \quad \boldsymbol{\theta}' \in R(\boldsymbol{\theta}). \tag{79}$$

Then, the all parameter points $\boldsymbol{\theta}'$ in $R(\boldsymbol{\theta})$ are equivalent from the point of view of the input-output behavior. There are many such subsets $R(\boldsymbol{\theta})'$s. We divide the space S by this equivalence relation. Then all the points in a $R(\boldsymbol{\theta})$ reduces to one behavior point. The topological structure of S is destroyed by this division, because every $R(\boldsymbol{\theta})$ reduces to a single point. This gives rise to singularities. We call $R(\boldsymbol{\theta})$ a singular region. We give two typical examples of singularity in a hierarchical system.

1) Eliminating singularity: When $v_i = 0$, whatever $\boldsymbol{w}_i$ is, $v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}) = 0$, so that $R = \{\boldsymbol{\theta} | v_i = 0, \boldsymbol{w}_i \text{ arbitrary}\}$ is a singular region. Since neuron i is eliminated because of $v_i = 0$, it is called an eliminating singularity region.

2) Overlapping singularity:

When $\boldsymbol{w}_i = \boldsymbol{w}_j = \boldsymbol{w}$, so long as $v_i + v_j = v$ holds,

$$v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x}) + v_j \varphi(\boldsymbol{w}_j \cdot \boldsymbol{x}) = v \varphi(\boldsymbol{w} \cdot \boldsymbol{x}). \tag{80}$$

So we cannot identify each of v_i and v_j. The region

$$R = \{\boldsymbol{\theta} \,| \boldsymbol{w}_i = \boldsymbol{w}_j = \boldsymbol{w}, v_i + v_i = v\} \tag{81}$$

is a singular region, depending on $\boldsymbol{w}$ and v. Since the two neurons become identical in this case, we call it an overlapping singularity.

There are other singularities where three neurons overlap and so on, but the above two are typical and other singularities are intersections of the above singularity regions. Therefore, we treat a simple system consisting of two hidden neurons, where $\boldsymbol{\theta} = (v_1, v_2, \boldsymbol{w}_2, \boldsymbol{w}_2)$ and study the dynamical behavior of learning in such a system.

In the critical region $R(v, \boldsymbol{w})$, only one neuron suffices, where $\boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}$ or one of v_1 and v_2 is 0. In this case, its input-output function is written by using a single neuron,

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = v \varphi(\boldsymbol{w} \cdot \boldsymbol{x}). \tag{82}$$

The situation is the same in the Gaussian mixture case.

We introduce a new coordinate system $\boldsymbol{\xi} = (\boldsymbol{v}, w, \boldsymbol{u}, z)$, defined by

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{w}_2 - \boldsymbol{w}_1, \quad \boldsymbol{w} = \frac{1}{2}(\boldsymbol{w}_1 + \boldsymbol{w}_2), \\ z &= \frac{v_1 - v_2}{v_1 + v_2}, \qquad v = v_1 + v_2. \end{aligned} \tag{83}$$

Then, the singular region is represented by $\boldsymbol{u} = 0$ or $z = \pm 1$. Given $(v, \boldsymbol{w})$, singular region $R(v, \boldsymbol{w})$ exists. A singular region $\mathcal{R}$ consists of three parts, one given by $\boldsymbol{u} = 0$, that is, $\boldsymbol{w}_1 = \boldsymbol{w}_2$, which we call the overlapping singular region, and the other is $z = \pm 1$, that is $w_1 = 0$ or $w_2 = 0$, in which one neuron is eliminated. We have $R = R_1 \cup R_2$,

$$R_1(\boldsymbol{v}, w) = \{\boldsymbol{\xi} \,| \boldsymbol{u} = 0, z : \text{arbitrary}; \boldsymbol{v} \text{ and } w \text{ fixed}\}, \tag{84}$$

$$R_2(\boldsymbol{v}, w) = \{\boldsymbol{\xi} \,| \boldsymbol{u} : \text{arbitrary}; z = \pm 1, \boldsymbol{v} \text{ and } w \text{ fixed}\}. \tag{85}$$

In the subspace in which $\boldsymbol{v}$ and w are fixed constants, R_1 is a line in which z changes, and R_2 consists of two n-dimensional surfaces. The singular region R is a composite of one line and two surfaces.

We study the dynamical flow near the singularity $R_1(\boldsymbol{v}, w)$, by using the Taylor expansion with respect to $\boldsymbol{u}$, where $|\boldsymbol{u}|$ is small. The average learning equation is expanded as

$$\dot{\boldsymbol{u}} = \eta \frac{w}{2}(1 - z^2) \langle e \varphi''(\boldsymbol{v} \cdot \boldsymbol{x})(\boldsymbol{u} \cdot \boldsymbol{x})\boldsymbol{x} \rangle + O(|\boldsymbol{u}|^2), \tag{86}$$

$$\dot{z} = -\frac{1 + z^2}{w} \eta \langle \varphi'(\boldsymbol{v} \cdot \boldsymbol{x})(\boldsymbol{u} \cdot \boldsymbol{x}) \rangle + O(|\boldsymbol{u}|^2). \tag{87}$$

We can integrate (86) and (87) in the subspace in which S^* where $\boldsymbol{v} = \boldsymbol{v}^*$ and $w = v^*$ are the optimal values, that is $v^* \varphi(\boldsymbol{w}^* \cdot \boldsymbol{x})$ is the optimal approximation of the target foundation $k(\boldsymbol{x})$ by using only a single neuron. This gives the flows or trajectories of dynamics in the neighborhood of $\boldsymbol{u} = 0$ of S^*. They are given by

$$\frac{1}{2} \boldsymbol{u}^T \cdot \boldsymbol{u} = \frac{2 w^{*2}}{3} \log \left[\frac{(z^2 + 3)^2}{|z|} \right] + c \tag{88}$$

where c is an arbitrary constant.

6 Milnor attractor

The line $R_1(\boldsymbol{v}^*, w^*)$ in S^* is a critical line of the dynamics of learning. Under a certain condition, we show that the line is divided into two parts such that one part is stable (attractive) and the other is unstable (repulsive). In such a case, there are trajectories attracted to the stable part, and the basin of attraction has a finite measure. This is the Milnor type attractor (Milnor, 1985), different from a saddle point. More strongly, the basin of attraction includes a neighborhood of this part, so that all the points near this part are once attracted to it. After being attracted to it, the parameters still move randomly in $R_1(\boldsymbol{v}^*, w^*)$ by stochastic fluctuations until they reach the unstable part (Fukumizu and Amari, 2000). Then, the parameters leave R_1 eventually.

We can prove that R_1^* is unstable (saddle) when $A = (A_{ij})$, $A_{ij} = \langle e\phi_{ij} \rangle$, where

$$\phi_{ij} = \frac{\partial^2}{\partial w_i \partial w_j} \varphi(\boldsymbol{w} \cdot \boldsymbol{x}) \tag{89}$$

includes both positive and negative eigenvalues. When $w^* A$ is positive-semi-definite, the interval $1 < z^2$ is attractive, and the remaining part $1 > z^2$ is repulsive. Hence, the interval $1 < z^2$ behaves like a Milnor attractor. This is the plateau phenomenon, taking long time before getting rid of it. When $w^* A$ is negative-semi-definite, the part $1 > z^2$ is attractive and the other part is repulsive.

7 Conclusions

We have shown examples of mathematical neuroscience approaches. They include 1) the macrodynamics of randomly connected networks and the characteristics of state transition are elucidated. 2) The dynamics of associative memory networks is elucidated from the point of view of macro-dynamics. 3) Pattern formations and pattern interactions in a neural field are investigated by using integro-differential equations. 4) It is studied how the singularity existing in a hierarchical system affects dynamics of learning.

Mathematical approaches are necessary for finding the principles of information processing which the brain utilizes. We need to use simple and tractable models for carrying mathematical analysis. The models are different from the real biological neural networks, but we hope the mathematical theory clarifies the principles from which we can construct realistic models. Mathematical neuroscience has not been fully established and much more efforts should be devoted for constructing such theories.

References

[1] P. Dayan and L.F. Abbott, Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems, MIT Press, 2001.

[2] S. Amari, Mathematical Foundations of Neurocomputing, *Proceedings of the IEEE*, 78, 9, 1443–1463, 1990.

[3] S. Amari, Characteristics of Randomly Connected Threshold-Element Networks and Network Systems, *Proc. IEEE.*, 59, 1, 35–47, 1971.

[4] S. Amari, A Method of Statistical Neurodynamics, *Kybernetik*, 14, 201–215, (Heft 4) 1974.

[5] S. Amari, K. Yoshida and K. Kanatani, A Mathematical Foundation for Statistical Neurodynamics, *SIAM J. Appl. Math.*, 33, 95–126, 1977.

[6] L.I. Rozonoer, Random logical nets; I, *Avtomat. Telemekh.*, 5, 137–147, 1969A.

[7] L.I. Rozonoer, Random logical nets; II, *Avtomat. Telemekh.*, 6, 99–109, 1969B.

[8] L. I. Rozonoer, Random logical nets; III. *Avtomat. Telemekh.*, 7, 127–136, 1969C.

[9] S. Amari, Characteristics of Random Nets of Analog Neuron-Like Elements, *IEEE Trans. Systems, Man and Cybernetics*, SMC-2, 5, 643–657, November 1972A.

[10] S. Amari, Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements, *IEEE Trans. Computers*, C-21, 11, 1197–1206, 1972B.

[11] J.J. Hopfield, Neural network and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences of the United States of America*, 79,8, 2554–8, 1982

[12] S. Amari and K. Maginu, Statistical neurodynamics of associative memory, *Neural Networks*, 1, 1, 63–73, 1988.

[13] H.R. Wilson and J.D. Cowan, A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue, *Kybernetik*, 13, 55–80, 1973.

[14] S. Amari, Dynamics of Pattern Formation in Lateral-Inhibition Type Neural Fields, *Biological Cybernetics*, 27, 77–87, 1977.

[15] H. S. Seung, Continuous attractors and oculomotor control, *Neural Networks*, 11, 1253–58, 1998.

[16] G.B. Ermentrout and J.D. Cowan, A mathematical theory of visual hallucinations patterns, *Biological Cybernetics*, 34, 137–150, 1979.

[17] S. Coombes and M.R. Owen, Bumps, breathers, and waves in a neural network with spike frequency adaptation, *Physical Review Letters*, 94, 148102, 2005.

[18] H. Wei, J. Zhang, F. Cousseau, T. Ozeki and S. Amari, Dynamics of Learning Near Singularities in Layered Networks, *Neural Computation*, 20, 813–843, 2008.

[19] H. Wei and S. Amari, Dynamics of learning near singularities in radial basis function networks, *Neural Networks*, 21, 989–1005, 2008.

[20] F. Cousseau, T. Ozeki and S. Amari, Dynamics of Learning in Multilayer Perceptrons Near Singularities, *IEEE Transactions on Neural Networks*, 19, 8, 1313–1328, 2008.

[21] S. Amari, H. Park and T. Ozeki, Singularities Affect Dynamics of Learning in Neuromanifolds, *Neural Computation*, 18, 1007–1065, 2006.

[22] J. Milnor, On the concept of attractor, *Commum. Math. Phys.*, 99, 177–195, 1985.

[23] K. Fukumizu and S. Amari, Local minima and plateaus in hierarchical structures of multilayer perceptrons, *Neural Networks*, 13, 317–327, 2000.