Probabilistic Neural Network: Frequency and Moment Learnings

Kyung Soo Rim\textsuperscript{1} and U Jin Choi\textsuperscript{2}

\textsuperscript{1}Department of Mathematics, Sogang University, Seoul, Korea
E-mail: ksrnim@sogang.ac.kr

\textsuperscript{2}Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Daejeon, Korea
E-mail: uchoi@kaist.ac.kr

April 16, 2020

Abstract—We introduce probabilistic neural networks that describe unsupervised synchronous learning on an atomic Hardy space and space of bounded real analytic functions, respectively. For a stationary ergodic vector process, we prove that the probabilistic neural network yields a unique collection of neurons in global optimization without initialization and back-propagation. During learning, we show that all neurons communicate with each other, in the sense of linear combinations, until the learning is finished. Also, we give convergence results for the stability of neurons, estimation methods, and topological statistics to appreciate unsupervised estimation of a probabilistic neural network. As application, we attach numerical experiments on samples drawn by a standing wave.

Index Terms—probabilistic neural network, synchronous learning, unsupervised learning, frequency learning, moment learning, network probability, governing probability, energy function, partition function, atomic Hardy space, bounded real analytic function, learning rate, full learning, Koopman mode decomposition, dynamic mode decomposition, active path, topological statistic.

1. INTRODUCTION

Nowadays, learning algorithms and architectures are currently being developed dramatically for deep neural networks using the back-propagation of the gradient descent method. In the mathematical theory of artificial neural networks, the universal approximation theorems, which were proved by G. Cybenko in 1989 \cite{Cybenko} and by K. Hornik showed in 1991 \cite{Hornik}, state that a feed-forward network with hidden layers containing a finite number of neurons can approximate continuous functions on compact subsets of $\mathbb{R}^n$, with an activation function. However, it does not concern the algorithmic learnability of those parameters.

The most common form of machine learning including recurrent neural networks, deep or not, is supervised learning. Wearing the back-propagation process in training, it has not provided global optimization, and has been dependent on initial conditions. There are many splendid results for the neural networks as main books such as \cite{Bishop}, \cite{Goodfellow}, \cite{Bengio}, \cite{Courville}, \cite{Haykin}, \cite{Murphy} are well rewritten mathematically, to provide examples. As the next generation of deep neural networks, many scientists mention unsupervised learning \cite{Bengio} thanks to advances in their architecture and ways of training them. Also, they expect unsupervised learning to become far more important in the long term, because there widely exist unsupervised signals, e.g., that are originated from human and animal learning.

The aim of this article is to find a network that gives the globally optimal unsupervised synchronous learning without any initialization and back-propagation. Using the probabilistic method and theory of dynamics, we define two types of probabilistic neural network and derive unique collections of neurons such that their network probabilities are the global solution for the observed samples.

The learning process of the probabilistic neural network is not carried out sequentially by hierarchical layers but is transmitted to all neurons at the same time as input. Once sample data is presented, all neurons in the probabilistic neural network interact simultaneously until the learning is complete. Also, there is no back-propagation. Meanwhile, the better the data (e.g. independent and identical distributed or stationary ergodic data) for learning get, the better the probabilistic neural network predicts. In addition, the probabilistic neural network gives a certain criterion of learning rate, from which we can control the amount of the observed samples. Thus, the probabilistic neural network is closer to the biological human brain \cite{Watson}, \cite{Geirhos}, \cite{Janssen}, \cite{Schutt}, \cite{Rauber}, \cite{Bethge}, \cite{Wichmann}.

This article is organized as follows. In the next section, we define energy and partition functions for a network probability which contains hidden parameters as unknown neurons. In section 3 we consider the Kullback-Leibler divergence between two probability densities and discuss that the energy function of the network probability is expanded in a space of functions as an infinite sum. The Fréchet derivative of a cross entropy is obtained to identify hidden parameters which are to be a solution of Fréchet partial differential equations. In section 4 we discuss the atomic Hardy space and space of bounded real analytic functions, on which we derive the exact forms of hidden parameters which are determined uniquely. The functions of the two spaces play a role of energy functions for frequency and moment learnings, respectively. As corollaries, we have simpler forms of parameters. Although an energy function may be an infinite series, we prove that the network probability equipped with the partial sums of the energy function converges to the limiting distribution in $L^1$-norm. In section 5 we prove that communication emerges between neurons during learning, until the learning is complete. Moreover, the learning rates of probabilistic neural networks are defined and explored in the section of application. In section 6 we derive a dynamical system from a cumulative distribution function for a time series of random vectors. If a sequence of samples is the stationary ergodic processes, then we can generate plenty of samples using the Koopman mode decomposition (KMD) of the induced dynamical system, linear or nonlinear \cite{Koopman}. In section 7 we prove that the empirical distribution function derived from a stationary ergodic process converges to the limiting distribution in $L^1$-norm. This guarantees the stability of convergence for empirical distribution functions. For probabilistic neural networks, we define the active path for a signal, and compute a likelihood of the signal to exist on the network in section 8. In addition, the physical interpretation is introduced by suitable topology on the active path.
path. Furthermore, more elaborate version of examples will be examined in section 9 including topological statistics for estimations.

Throughout this paper we use the following general notations:

For a random vector $X$ with its value $x$, $X_t$ means a random vector at time $t$ with its value $x_t$, and $X_{t:k}$ is the $k$th random vector of $X_t$ with its value $x_{t:k}$. Moreover, $(X)$ and $(x)$ denote sequences of $X$ and its value $x$, respectively. The notation of $E(X)_n$ means the expectation of $X$ with a probability distribution of $p$. For quantities $A$ and $B$, we write $A \leq B$ if there is a constant $C_n$ which depends only on $n$ such that $A \leq C_n B$, where possibly depending on some other variables as well, we append them to $n$. Also, $A \approx B$ means $A \leq B$, $B \leq A$, and write $A \equiv B$ when $A$ is defined as $B$. If an operations appear between multi-indexes, e.g., an inequality, combinatorial notation, partial derivative, etc., it follows the rule of multi-index operations. Especially, $\dot{x}$ is the derivative of $x$ with respect to $t$ when $t$ is regarded as time, and $Z^*$ the set of all non-negative integers. Finally, the notation of $| \cdot |$ denotes the absolute value of a scalar or multi-index, or the euclidean norm of a vector. Sometimes one can meet ‘.’ just like $| \cdot |$ without any concrete variable. To avoid abusing notations, we omit any variable if we do not need to.

2. ENERGY AND PARTITION FUNCTIONS

Let $X = (X_1, \ldots, X_n)$ be a random vector from $P_0$ unknown, namely, a governing probability. Assume that there is a probability $P$ of $X$, namely, a network probability, with $Y$ a collection of parameters such that

$$P_0(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y),$$

where $y = (y_n)$ is a countable collection of complex numbers. For simplicity of expression, we also call $p_0(x_1, \ldots, x_n)$ and $p(x_1, \ldots, x_n \mid y)$ a governing probability and network probability for an observed sample vector $x = (x_1, \ldots, x_n)$ with a collection of parameters $y = (y_n)$, respectively, which play a role of neurons.

Throughout the article, assume that probability distributions defined on a Borel $\sigma$-algebra assign a positive probability to a nonempty open set, the entropy of $p$ is finite, and for each $x$, $p(x \mid y)$ is continuously differentiable as a function of $y$. Since $p_0$ does not contain a neuron, it alone cannot describe a neural network. From information of $x$ fixed, we devote to find $y$ at which $p(x \mid y)$ equals $p_0(x)$. The fact that a sample $x$ is observed, implies that a certain network probability causes $x$, and thus, the network probability does not vanish identically. We rewrite $p(x \mid y)$ as a quotient of two positive functions,

$$p(x \mid y) = \frac{f(x, y)}{g(y)},$$

where $f(x, y)$ is an integrable function for $x$ combined with $y$ and $g(y) = \int f(x, y)dx$ is a partition function of $y$ with respect to $f$. By (1), it follows that

$$p_0(x) \propto f(x, y)$$

for each $y$. This means that the governing probability is proportional to $f$. From $f(x, y) > 0$, (2) is written as

$$p(x \mid y) = \frac{e^{-E(x, y)}}{Z(y)},$$

where $E(x, y) = -\ln f(x, y)$ is called an energy function for the network and rewrite $g(y)$ as $Z(y)$ conventionally. If the components $X_1, \ldots, X_n$ of $X$ are i.i.d., then by

$$E(x_1, \ldots, x_n, y) = E(x_1) + \cdots + E(x_n, y),$$

where $E(x_n; y) = -\ln f(x_n, y)$.

Sometimes, a flow graph is useful to understand a random process. A signal-flow graph is a network of directed links that are interconnected at certain points called nodes $y_n$. A probabilistic neural network is also a signal-flow graph which consists of an observed sample and parameters of a network probability that satisfies (1). A hidden node $y_n$ has associated every input signal $x_n$. The probabilistic neural network is represented by means of Figure 1 which consists of two parts of the sampling and learning processes. The former provides samples from the data-driven method if we need it, while the latter approximates the values of parameters.

3. THE KULLBACK-LEIBLER DIVERGENCE

According to (1) and (2), the goal is to identify $y$ such that

$$p_0(x) = \frac{e^{-E(x, y)}}{Z(y)},$$

We call a component $y_n$ of $y$ in (2) a neuron of the network probability or governing probability. To solve the equation (5), information of $E(x; y)$ is very important. In this study we are devoted to analyzing it by a linearization which is expressed in a suitable infinite dimensional space.

For probability distributions $p_1$ and $p_2$ defined on the same probability space, the Kullback-Leibler divergence between $p_1$ and $p_2$ is defined by

$$D_{KL}(p_1 \parallel p_2) = \mathbb{E} \left( \ln \frac{p_1}{p_2} \right),$$

which is defined only for $x$, where $p_2(x) = 0$ implies $p_1(x) = 0$. Although the Kullback-Leibler divergence is not a distance, it satisfies the following three conditions:

(a) $D_{KL}(p_1 \parallel p_2) \geq 0$ (Gibbs’ inequality).
(b) $D_{KL}(p_1 \parallel p_2) = 0$ if and only if $p_1 = p_2$ a.e. (identity of indiscernibles).
(c) $D_{KL}(p_1 \parallel p_2) \neq D_{KL}(p_2 \parallel p_1)$ (asymmetry).
For non-negative measurable functions $f_1$ and $f_2$, by the equality condition of Jensen’s inequality, $(b)$ is extended to $D_{KL}(f_1 \parallel f_2) = 0$ if and only if

$$f_1 = cf_2 \text{ a.e.}$$

for some constant $c$.

The Kullback-Leibler divergence with $p_0$ and $p$ instead of $p_1$ and $p_2$ yields that

$$D_{KL}(p_0 \parallel p) = -\mathbb{E}(\ln p(\cdot \mid y))_{p_0} - \left( -\mathbb{E}(\ln p_0) \right) \text{ cross entropy of } p_0 \text{ and } y \text{ entropy of } p_0 \tag{6} \equiv H(p_0, p) - H(p_0) \geq 0.$$

So, for all $y$, $H(p_0, p) \geq H(p_0)$ and the Kullback-Leibler divergence can be written as the cross entropy of $p_0$ and $p$, minus the entropy of $p_0$. To see this, by the identity of indiscernibles of the Kullback-Leibler divergence, we have to only find $y$ such that $D_{KL}(p_0 \parallel p) = 0$. Two quantities of $D_{KL}(p_0 \parallel p)$ and $H(p_0, p)$ are identical by the constant $H(p_0)$ difference. The first step toward figuring out the most efficient solution is to determine $y$ such that

$$\arg \min_y D_{KL}(p_0 \parallel p(\cdot \mid y)) = \arg \min_y H(p_0, p(\cdot \mid y)). \tag{7}$$

Unfortunately, it is difficult to clarify $(7)$ directly because of nonlinearity of $H(p_0, p(\cdot \mid y))$. To overcome that issue, we will expand the nonlinear energy function in suitable Banach spaces.

Let $r, r'$ be a pair of conjugate exponents with $1 \leq r \leq \infty$. For $1 \leq r' < \infty$, let $E(x; y) \in \ell'(\mathbb{Z}^n, \mathcal{F}(S)) \otimes \ell'(\mathbb{Z}^n, \mathbb{C}) \equiv \ell'(\mathcal{F}(S)) \otimes \ell'(y)$ be an operator such that $(\phi_\alpha) \in \ell'(\mathcal{F}(S))$ and $(y_\alpha) \in \ell'(y)$, where $\mathcal{F}(S)$ is a proper function space on a compact sample space $S$ and $\otimes$ a tensor product. If $r' = \infty$, then $(y_\alpha)$ is chosen in $c_0(\mathbb{Z}^n, \mathbb{C}) \equiv c_0$ as a subspace of $\ell^\infty(\mathbb{Z}^n, \mathbb{C})$.

Suppose that $E$ has the form of

$$E(x; y) = \sum_\alpha \phi_\alpha(x)y_\alpha \tag{8}$$

such that for $1 \leq r < \infty$,

$$\|\phi_\alpha\|_{\ell'}$$

is uniformly bounded on $S$, and for $r = \infty$,

$$\sup_\alpha |\phi_\alpha|$$

is uniformly bounded on $S$. Note that $E$ converges at every pair of $x$ and $y$ by Hölder’s inequality. For the partition function of $E$, furthermore, the integrability of $e^{-E(x; y)}$ for any $y$, is always assumed.

**Lemma 3.1.** If $E$ satisfies $(8)$, then $D_{KL}(p_0 \parallel p)$ is well defined and its Fréchet derivative is induced by

$$\partial D_{KL}(p_0 \parallel p)(h) = \sum_\alpha \partial_{\phi_\alpha} H(p_0, p)h_\alpha \tag{9}$$

for $h \in \ell'$ if $1 \leq r' < \infty$ and $h \in c_0$ if $r' = \infty$.

**Proof.** From $(9)$, we only prove the lemma with $H(p_0, p)$ instead of $D_{KL}(p_0 \parallel p)$, since $H(p_0)$ is a positive number. We first show the summability of $H(p_0, p)$. By the definition of the cross entropy in $(9)$,

$$H(p_0, p) = \mathbb{E}(E(\cdot \mid y) + \ln Z(y))_{p_0}$$

$$\equiv \mathbb{E}(\sum_\alpha \phi_\alpha(\cdot)y_\alpha) + \ln \mathbb{E}(e^{-\sum_\alpha \phi_\alpha(x)y_\alpha}dx) \tag{10} \equiv I_1 + I_2.$$

By assumption, $I_2$ is readily finite. We show the convergence of $I_1$. Fix $y$. If $r = \infty$, then by the triangle inequality, $I_1$ is bounded by

$$\mathbb{E}\left(\sup_\alpha |\phi_\alpha|\right)_{p_0} \|y\|_{\ell'} \leq \mathbb{E}\sup_\alpha |\phi_\alpha|_{L^\infty(S)} \|y\|_{\ell'} < \infty.$$

For $1 \leq r < \infty$, by the triangle inequality and by Hölder’s inequality again, $I_1$ is less than or equal to

$$\mathbb{E}\left(\sum_\alpha |\phi_\alpha|^{\frac{1}{r'}}\right)_{p_0} \left\|\sum_\alpha |\phi_\alpha|\right\|_{L^r(S)} \|y\|_{\ell'} \left\|\sum_\alpha \phi_\alpha y_\alpha\right\|_{\ell'} < \infty.$$

We will find the Fréchet derivative for $H(p_0, p)$: By interchangeability of integral and limit signs, the ordinary partial derivative of $H(p_0, p(\cdot \mid (y_\alpha), \ldots, y_0))$ with respect to $y_\alpha$ exists. Indeed,

$$\partial_{\phi_\alpha} H(p_0, p(\cdot \mid y))$$

$$= \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0} + \mathbb{E}\left(\frac{\partial_{\phi_\alpha} Z(y)}{Z(y)}\right)_{p_0} \tag{12}$$

$$= \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0} + \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0} - \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0} \tag{12}$$

$$= \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0} - \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0}$$

$$= \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0} - \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0}$$

$$= \mathbb{E}(\partial_{\phi_\alpha} E(\cdot \mid y))_{p_0}.$$
Similarly,
\[
\frac{|H|}{\|h\|_{p'}} \to 0
\]
as \(\|h\|_{p'} \to 0\).

Thus,
\[
\partial D_{KL}(p_0\| p)(h) = \partial H(p_0, p) = \sum_\alpha \partial_{\phi_\alpha} H(p_0, p) h_\alpha
\]
for \(h\) such that \(\|\text{supp} h\| < \infty\). Since clipped sequences are dense in \(\ell^r\) if \(1 \leq r < \infty\) and in \(c_0\) if \(r = \infty\), it is sufficient to show that \(\partial D_{KL}(p_0\| p)\) is a bounded linear functional.

For \(1 \leq r < \infty\), by (12), and by Minkowski inequality,
\[
\left(\sum_\alpha |\partial_{\phi_\alpha} H(p_0, p)|^r\right)^{1/r} \leq \left(\sum_\alpha |E(\phi_\alpha) - \mathbb{E}(\phi_\alpha)|^r\right)^{1/r} \leq 2 \|\phi_\alpha\|_r \|\phi_\alpha\|_{L^\infty(S)} < \infty.
\]

If \(r = \infty\), then by (12) and by the triangle inequality,
\[
|\partial_{\phi_\alpha} H(p_0, p)| \leq 2 \sup_\alpha \|\phi_\alpha\|_{L^\infty(S)} < \infty.
\]

From the form of (13), those give the boundedness of \(D_{KL}(p_0\| p)\).

In Theorem 3.1, we call \(\partial_{\phi_\alpha} H(p_0, p)\) a Fréchet partial derivative of \(H(p_0, p)\) to distinguish it from ordinary derivatives. Now we are ready to solve (14) partially except for uniqueness.

**Theorem 3.2.** If \(E\) satisfies the conditions of (8) and \(D_{KL}(p_0\| p)\) has the minimum at \(y\), then
\[
\mathbb{E}(\phi_\alpha)_{p_0} = \mathbb{E}(\phi_\alpha)_{y}
\]
holds at \(y\). Moreover, a network probability \(p\) for \(p_0\) is given by
\[
p(x \mid y) = \frac{e^{-E(x, y) - s}}{Z} = p_0(x),
\]
where \(s = \mathbb{E}(\ln \frac{1}{p_0 x})\), and \(Z\) is the partition function with respect to the energy function \(E + s\).

**Proof.** Let \(y\) be a minimum point of \(D_{KL}(p_0\| p)\). Combining the boundedness of the Fréchet derivative only with finite sequence \(h\) with Lemma 3.1, we have
\[
\frac{d}{dt} D_{KL}(p_0\| p)(p(x \mid y + th)) \bigg|_{t=0} = \partial D_{KL}(p_0\| p)(p(x \mid y)) = 0.
\]
Equivalently, from (12),
\[
\mathbb{E}(\phi_\alpha)_{p_0(x)} - \mathbb{E}(\phi_\alpha)_{p(x)} = 0
\]
for all \(\alpha\). The obtained energy function \(E\) will only differ from it by some added constant.

Let \(\tilde{p} = e^{-\bar{E}}\) be a non-negative function, where \(\bar{E} = E + s\) with \(s = \mathbb{E}(\ln \frac{1}{p_0 x})\) evaluated by the solution \(y\). Note the shift \(s\) is finite from (10) and the finiteness of \(H(p_0)\). It follows that
\[
D_{KL}(p_0\| \tilde{p}) = E\left(\ln \frac{p_0}{\tilde{p}}\right)_{p_0} = \mathbb{E}(\ln p_0)_{p_0} - \mathbb{E}(\ln \tilde{p})_{p_0} = \mathbb{E}(\ln p_0)_{p_0} + \mathbb{E}(E)_{p_0} + s = 0,
\]
where \(s\) is the smallest number, since \(D_{KL}(p_0\| p)\) has the minimum at \(y\). By identity of indiscernibles, \(\tilde{p} = cp_0\) for a positive constant \(c\). The fact of
\[
c = \int_S cp_0 dx = \int_S \tilde{p} dx
\]
produces that \(c\) must be the partition function \(Z\) of \(\tilde{p}\). Thus, we have the desired network probability,
\[
p(x \mid y) = \frac{e^{-\bar{E}(x, y)}}{Z} = p_0(x)
\]
equipped with \(y\). Therefore, the proof is complete.

As a special case, if \(D_{KL}(p_0\| p)\) has the minimum at \(y = 0\), then \(E = 0\) and \(p = p_0\) is a uniform probability distribution on \(S\). It is not simple to apply Theorem 3.2 if the number of first-order Fréchet partial differential equations is not finite. Nevertheless, (13) is a useful necessary condition to decide extreme points. A suitable solution of (14) is a candidate for the minimum point of \(D_{KL}(p_0\| p)\).

**Example 3.1.** Let \(p_0(x) = e^{-2x} / \sinh(2)\) be a memereyless distribution on \([-1, 1]\). We assume that the governing probability model is given. (In the next section, we introduce sample driven methods without any information on governing probabilities.) We may put \(f(x) = e^{-E(x, y)}\), where \(E(x, y) = y_0 + y_1x\).

If \(y_1 = 0\), then \(Z(y_1 = 0) = 2e^{-y_0}\) and \(p(x \mid y) = 1/2\) which is a uniform distribution on \([-1, 1]\). A random sample from \(p_0\), does not follow the uniform distribution with probability 1. Thus, we assume \(y_1 \neq 0\), the partition function \(Z(y_1 \neq 0) = \frac{2}{y_1} e^{-y_0 \sinh(y_1)}\), and
\[
p(x \mid y_1) = \frac{y_1 e^{-y_0 s}}{2 \sinh(y_1)}.
\]

From \(\partial_\phi E = \phi_0 = 1\) and \(\partial_\phi E = \phi_1 = x\), (14) is calculated at
\[
\mathbb{E}(1)_{p_0} = \mathbb{E}(1)_{y_1}, \text{ and } \mathbb{E}(X)_{p_0} = \mathbb{E}(X)_{y_1}.
\]
We obtain \(y_1 = 2\) and a network probability is derived as
\[
p(x \mid y) = p(x \mid y_1 = 2) = p_0(x).
\]
In fact, the collection of \(y_1 = 2\) and \(Z(y_1 = 2) = \sinh(2)\) is determined uniquely by Theorem 4.3 with
\[
y_1 = \frac{d}{dx} \ln \frac{1}{p_0 x} \bigg|_{x=0} = 2, \text{ and } \ln Z(y_1 = 2) = \ln \frac{1}{p_0 x} \bigg|_{x=0} = \ln \sinh(2),
\]
and \(y_k = \frac{e^{dt}}{dt} \ln \frac{1}{p_0 x} \bigg|_{x=0} = 0 \text{ for } k \geq 2\).

If \(X_1, \ldots, X_n\) are i.i.d., then Example 3.1 can be extended to
\[
p(x_1, \ldots, x_n \mid y) = e^{-2 \sum_{k=1}^n x_k} \sinh^2(2).
\]

4. Probabilistic Neural Networks

In this section, we introduce two function spaces: an atomic Hardy space and space of real analytic functions, in which energy functions will be taken. On the spaces, Theorem 4.2 and 4.3 characterize the neurons using Theorem 3.2. The proofs of two theorems adopt theory of functions to bypass a large amount of calculations of (14) and to give the uniqueness.

We say that a distribution \(f\) belongs to the \(H^1\)-Hardy space if for some Schwartz function \(\phi \in \mathbb{R}^n\) with \(\int_{\mathbb{R}^n} \phi(x) dx \neq 0\), the maximal function
\[
M_\phi f(x) = \sup_{t > 0} |f \ast \phi_t(x)|
\]
is integrable, where \( \phi_t(x) = \phi(x/t)/t^n \). Then the \( H^1 \)-Hardy space is a Banach space with norm \( \|f\|_{H^1} = \int_{\mathbb{R}^n} |M_\phi f| \, dx \). If we define a function \( a \), namely, an \( H^1 \)-atom, such that

(a) \( a \) is supported in a cube \( Q \),
(b) \( |a| \leq |Q|^{-1} \) almost everywhere,
(c) \( \int_{Q} a(x) \, dx = 0 \),

then by the atomic decomposition theorem, \( f \) can be written as an infinite linear combination of atoms \( a_k \) of \( f = \sum_{k=1}^{\infty} y_k a_k \) whose norm is equivalent to \( \sum_{k=1}^{\infty} |y_k| \), where \( \sum_{k=1}^{\infty} |y_k| < \infty \) for complex numbers \( y_k \).

Let \( \mathbb{T}^n = [-\pi, \pi]^n \) be a torus. We define an atomic Hardy space \( H^1_a(\mathbb{T}^n) \equiv H^1_a \) by

\[
H^1_a = \left\{ \sum_{\alpha \in \mathbb{Z}^n \setminus \{0\}} y_\alpha \omega_\alpha(x) \left| \sum_{\alpha \in \mathbb{Z}^n \setminus \{0\}} |y_\alpha| < \infty \right. \right\}
\]

with the norm \( \|f\|_{H^1_a} = \sum_{\alpha \neq 0} |y_\alpha| \), where \( \omega_\alpha(x) = e^{i\alpha \cdot x} \).

Note that \( \omega_\alpha \) satisfies (a), (b), and (c) of \( H^1 \)-atom on \( \mathbb{T}^n \) instead of \( \mathbb{R}^n \).

We denote the subspace of \( L^1(\mathbb{T}^n) \) whose element has zero integral (i.e., direct current (DC) free) by \( L^1_0(\mathbb{T}^n) \equiv L^1_0 \).

**Proposition 4.1.** The space \( H^1_a \) is dense in \( L^1_0 \).

**Proof.** Let \( \sum_{\alpha \in \mathbb{Z}^n \setminus \{0\}} y_\alpha \omega_\alpha(x) \in L^1_0 \). The series converges uniformly and absolutely. So, the limit is continuous and by interchanging sum with integral, its integral vanishes. Thus, \( H^1_a \subset L^1_0 \).

For a bounded function in \( L^1_0 \), taking a circular convolution of the function and a sequence of mollifiers, by Fubini’s theorem we have an \( L^1 \)-convergent sequence of smooth functions. By the reproducing property of mollifiers, the limit of convolutions recovers the bounded function in the \( L^1 \)-norm. Since bounded functions are dense in \( L^1_0 \), the set of smooth functions whose integrals vanish, is also dense in \( L^1_0 \).

Let \( f \) be an \( m \) times continuously differentiable function in \( L^1_0 \). If \( m > 1 + n/2 \), then by the Fourier series representation for smooth functions, \( f = \sum_{\alpha \neq 0} \hat{f}(\alpha) \omega_\alpha \), where \( (\hat{f}(\alpha)) \in \ell^1 \) and

\[
\hat{f}(\alpha) = \frac{1}{2^n} \int_{\mathbb{T}^n} f(x) \omega_\alpha(x) \, dx
\]

is the \( \alpha \)th Fourier coefficient of \( f \). The fact of vanishing integral of \( f \) implies \( f \in H^1_a \). Hence, \( H^1_a \) is dense in \( L^1_0 \) and the proof is complete. □

By normalization, we suppose that a bounded random vector belongs to \( \mathbb{T}^n \). As one of the main results, the next theorem gives the unique solution of \( \hat{f} \) in the concept of frequency-analysis for \( E(x; y) \in H^1_a \), which is embedded in \( \ell^\infty(F(\mathbb{T}^n)) \). However, \( H^1_a \) is a candidate space for energy functions until the integrability of \( e^{-Z} \) is guaranteed. We will prove it in Theorem 4.4.

**Theorem 4.2.** Let \( X = (X_1, \ldots, X_n) \in \mathbb{T}^n \) be a random vector from \( p_0 \). If \( E \in H^1_a \), then \( y \) is the unique solution of \( p_0(x) = \alpha \phi_t(x) \), determined by

\[
y_\alpha = \ln \frac{1}{p_0(\alpha_k)}(\alpha_k) \quad \text{if} \quad \alpha_k = |\alpha|
\]

\[
y_\alpha = \ln \frac{1}{p_0(\alpha_k)}(0) \quad \text{elsewhere},
\]

\[
\ln Z(y) = \sum_{k=1}^{n} \ln \frac{1}{p_0(x_k)}(0),
\]

where \( \tilde{\alpha} \) is a Fourier coefficient on \( \mathbb{T}^n \).

**Proof.** Let \( \alpha \neq 0 \). From the i.i.d. property,

\[
\ln p_0(x) = \sum_{k=1}^{n} \ln p_0(x_k).
\]

By Theorem 4.2,

\[
y_\alpha = \frac{1}{2^n} \sum_{k=1}^{n} \int_{\mathbb{T}^n} \tilde{\omega}_\alpha(x) \ln \frac{1}{p_0(x_k)}(x) \, dx \quad \text{(18)}
\]

\[
\ln Z = \frac{1}{2^n} \sum_{k=1}^{n} \int_{\mathbb{T}^n} \ln \frac{1}{p_0(x_k)}(x) \, dx.
\]

In the integrand of (18), \( 1/p_0(x_k) \) depends only on \( x_k \). Thus, (18) vanishes if \( \alpha \) has a non-trivial component whose index is different from \( k \). □
In the process of estimation it is advantageous to classify and analyze neurons into bundles. We divide these into disjoint collections. For $n$ the number of features, $y_\alpha : \mathbb{Z}^n \rightarrow \mathbb{C}$ is a complex-valued function over $\alpha \mapsto y_\alpha$. For a positive integer $k$, we define the $k$-cell $C_k(k)$ of $y_\alpha$ whose frequency order size is $k$, more precisely,

$$C_k(k) = \{ y_\alpha \mid |\alpha| = |\alpha_1| + \cdots + |\alpha_n| = k \}.$$ 

For $k = 0$, set $C_k(0) = \{ D_c \}$ a singleton of zero index.

Sometimes it is more efficient to restrict neurons within an appropriate finite subset. We denote a finite sub-collection of indexes by $\mathcal{D}$, namely, a dictionary, and its configuration is freely selectable. In order to appreciate the representation of estimation from neurons, it is necessary to seize the point of drawn samples.

Figure 2: Architecture of the frequency learning for $(x_1, x_2, x_3)$. The blue area represents the neurons of the dictionary $\mathcal{D} = \{0, 0, 0, 1, 0, 0, 0, 1, 0\}$, where 0 is the index of $D_c$ and $f \equiv -r$ for a positive integer $r$. Any pair of neurons is connected by a red dashed bidirectional arrow in the sense of Theorem 4.1. This explains that communication of neurons emerges in the process of frequency learning.

We are occasionally interested in geometric quantities of an energy function, e.g., a mean, variation, skewness, or kurtosis, etc. To calculate the moments of the energy, assume that $p_0$ is sufficiently smooth. Let $A_0(\mathbb{T}^n) \equiv A_0$ be the space of all real analytic functions on the interior of $\mathbb{T}^n$, which fix the origin, with the supremum norm and write $f \in A_0$ as

$$f(x) = \sum_{\alpha \in \mathbb{Z}^n \setminus \{0\}} y_\alpha x^\alpha,$$

where $y_\alpha \in \mathbb{R}$. From the definition of $A_0$, $(y_0) \in \ell^1$.

The result below is one of the main theorems that gives the unique solution for $f$ on $A_0$, in the concept of moment-analysis, which is embedded in $\ell^\infty(\mathcal{F}(\mathbb{T}^n) \otimes \ell^1)$. Let us note that the integrability of $e^{-Ex}$ is shown in Theorem 4.4.

**Theorem 4.3.** Let $X = (X_1, \ldots, X_n) \in \mathbb{T}^n$ be a random vector from $p_0$. If $E \in A_0$, then $y$ is the unique solution of $p_0(x) = p(x \mid y)$, determined by

$$y_\alpha = \frac{1}{\alpha!} \frac{\partial^{\alpha} Z}{\partial x^\alpha} \ln \frac{1}{p_0(x)} \bigg|_{x = 0}, \quad \ln Z(y) = \ln \frac{1}{p_0(x)} \bigg|_{x = 0},$$

i.e.,

$$p_0(x) = e^{-\ln Z(y) - \sum_{\alpha \in \mathbb{Z}^n} y_\alpha x^\alpha}.$$  

We denote $\ln Z(y)$ by $D_c$ in Theorem 4.3 and call it the direct current of the network. Neurons $y_\alpha$ are partial derivatives of $\ln 1/p_0$ at 0. By the reason of $y_\alpha$, representing quantitative measures of the shape of $E$, we call the process of Theorem 4.3 a moment learning for $X$. For a positive integer $k$, we collect $y_\alpha$ having a moment order $\alpha$ such that $|\alpha| = k$ and put it by $C_k(k)$, namely, the $k$th cell of $y$. In addition, write $C_k(0) = \{ D_c \}$. In Figure 3 we draw the architecture of Theorem 4.3.

**Proof of Theorem 4.3.** Let $x$ be the realization vector of $X$. According to $x^\alpha$ is uniformly bounded by 1 and $y \in \ell^1$. Lemma 4.1 produces that the Fréchet derivative of $E$ with respect to $y_\alpha$ is

$$\partial_{y_\alpha} E = x^\alpha,$$

and so, for $\alpha \neq 0$,

$$\int_\mathbb{T}^n (p(x \mid y) - p_0(x)) x^\alpha dx = 0.$$

The polynomials are dense in $A_0$, that implies that the integral of $p(x \mid y) - p_0(x)$ is constant, and besides, the constant must be 0, since both are probability distributions. Hence, there is the unique $y$ such that $p(x \mid y) = p_0(x)$. Analyticity of $E$ turns out to be

$$E(x; y) = \sum_{\alpha \neq 0} y_\alpha x^\alpha = -\ln Z(y) - \ln p_0(x).$$

By the uniqueness of the power series of (19), we have

$$y_\alpha = \frac{1}{\alpha!} \frac{\partial^{\alpha} Z}{\partial x^\alpha} \ln \frac{1}{p_0(x)} \bigg|_{x = 0}, \quad \ln Z(y) = \ln \frac{1}{p_0(x)} \bigg|_{x = 0},$$

for $\alpha \neq 0$ and $D_{k1}(p_0 || p) = 0$ uniquely. Therefore, $p(x \mid y) = p_0(x)$ and the proof is complete. $\square$

From 1.D. features, the non-trivial neurons are located only on axes.

**Corollary 4.3.1.** If all components $X_1, \ldots, X_n$ of $X$ are 1.D. from $p_0$, then for $\alpha \neq 0$,

$$y_\alpha = \frac{1}{\alpha!} \frac{\partial^{\alpha} Z}{\partial x^\alpha} \ln \frac{1}{p_0(x_k)} \bigg|_{x_k = 0}, \quad \ln Z(y) = \sum_{k=1}^{n} \ln \frac{1}{p_0(x_k)} \bigg|_{x_k = 0}.$$

**Proof.** By the 1.D. property,

$$\ln p_0(x) = \sum_{k=1}^{n} \ln p_0(x_k).$$

By Theorem 4.3

$$y_\alpha = \frac{1}{\alpha!} \frac{\partial^{\alpha} Z}{\partial x^\alpha} \sum_{k=1}^{n} \ln \frac{1}{p_0(x_k)} \bigg|_{x_k = 0},$$

In the summand of (21), $\ln 1/p_0(x_k)$ depends only on $x_k$. Thus, (21) vanishes if at least two components of $\alpha$ are non-zero for $\alpha \neq 0$, when $\alpha = 0$, the result follows directly. $\square$

The corollary below shows that the lower and upper magnitudes of a governing probability control the size of an energy function from $\ln 1/p_0(x) = E(x; y) + D_c$.

**Corollary 4.3.2.** If $y$ is the the solution in Theorem 4.2 or 4.3 then

$$\text{ess inf} \ln \frac{1}{p_0(x)} - D_c \leq E(x; y) \leq \text{ess sup} \ln \frac{1}{p_0(x)} - D_c$$

for almost every $x$. 


Remark 4.1. The blue area represents the neurons in the dictionary $\mathcal{D} = \{0, 100, 010, 001, 200\}$, where 0 is the index of $D_0$. Any pair of neurons is connected by a red dashed bidirectional arrow in the sense of Theorem 5.2. This explanation that communication of neurons emerges in the process of moment learning.

The next result ensures the $L^1$-norm convergence of approximated network probabilities equipped with a finite sum of $E$.

**Theorem 4.4.** Let $E_N(x; y) = \sum_{0 \neq |\alpha| \leq N} y_\alpha \omega_\alpha(x)$ or $\sum_{0 \neq |\alpha| \leq N} y_\alpha x^\alpha$. Then

$$\lim_{N \to \infty} \int_0 \left| e^{-E(x; y)} - e^{-E_N(x; y)} \right| dx = 0.$$ 

**Proof.** By Taylor’s series expansion, by the triangle inequality, and by the uniform convergence,

$$\int_0 \left| e^{-E(x; y)} - e^{-E_N(x; y)} \right| dx$$

$$\leq \sum_{k=1}^{\infty} \frac{1}{k^k} \int_0 e^{-E(x; y)} \left| E(x; y) - E_N(x; y) \right|^k dx$$

$$\leq \sum_{k=1}^{\infty} \frac{1}{k^k} \left( \sum_{|\alpha| > N} |y_\alpha| \right)^k \int_0 e^{-E(x; y)} dx$$

$$= \sum_{k=1}^{\infty} \frac{1}{k^k} \left( \sum_{|\alpha| > N} |y_\alpha| \right) Z(y)$$

$$= \left( \sum_{|\alpha| > N} |y_\alpha| \right) - 1 Z(y).$$

The last term goes to 0 as $N \to \infty$, since $y \in \ell^1$. Therefore, the proof is complete. \hfill \Box

**Remark 4.1.** (i) Theorems 4.2 and 4.3 provide that the components of a random vector are unidirectionally connected to every neuron simultaneously.

(ii) One may generalize the energy space to a separable space that satisfies summable conditions of $\mathcal{B}$.

(iii) Architectures of Theorems 4.2, 4.3 say that cells of higher order neurons are responsible for higher resolutions of frequencies and moments.

(iv) Theorem 4.4 guarantees the $L^1$-convergence of $e^{-E_N}$. This enables us to approximate likelihoods equipped with $E_N$ instead of $E$.

5. **FULL LEARNING AND COMMUNICATION OF NEURONS**

With finite samples, the induced $p_0$ is an approximation. Moreover, any calculated neuron $\hat{y}_\alpha$ is also an approximation of $y_\alpha$. In this section, we will settle a learning status up to the full learning of $\hat{y}_\alpha = y_\alpha$.

Since $\mathbb{T}^n$ is second countable in the standard topology, it is generated by a countable basis $(B_k)$. If $x \in \mathbb{T}^n$ does not have a dense orbit $(x_k)$, then the orbit is disjoint with some $B_k$. From the assumption in advance that a nonempty open set has a positive probability, we have

$$0 = \frac{1}{m} \sum_{k=0}^{m-1} \mathbbm{1}_{B_k}(x_k) \neq \mathbb{E}(\mathbbm{1}_{B_k}(X))p_0 > 0$$

for any $m$, where $\mathbbm{1}_{B_k}$ is an indicator function of $B_k$. By Birkhoff ergodic theorem, the set of all $x$ whose orbit is not dense in $\mathbb{T}^n$ has measure zero. Hence, we conclude the following remark. The Birkhoff ergodic theorem will be described in section 6 more precisely.

**Remark 5.1.** For almost every $x$, its orbit is dense in $\mathbb{T}^n$.

5.1. **FREQUENCY LEARNING** Let $(x)$ be stationary ergodic samples. Take a partition of $\mathbb{T}^n$ with non-overlapping rectangles such that every rectangle contains a single sample, whose volume is written as $\Delta_x$. On $H_1$, from the approximation,

$$\hat{y}_\alpha \approx \frac{1}{2^n} \sum_{x} \tilde{\omega}_\alpha(x) \ln \frac{1}{p_0(x)} \Delta_x,$$  

we have

$$\hat{y}_\alpha \approx y_\alpha \ln \frac{1}{p_0}(\alpha) = \ln \frac{1}{p_0}(\alpha + \beta)$$

$$= \sum_{\gamma \in \mathbb{Z}^n} \tilde{\omega}_\gamma(\alpha + \beta - \gamma)(x) \ln \frac{1}{p_0}(\gamma)$$

$$\approx \frac{1}{2^n} \sum_{\gamma \in \mathbb{Z}^n} \sum_{x} \tilde{\omega}_{\alpha - \gamma}(x) \Delta_x \hat{y}_\gamma,$$

where $\hat{y}_\alpha$ is an approximation of $\hat{y}_\alpha$. Every $\hat{y}_\alpha$ is an infinite linear combination of $\tilde{y}_\alpha$, approximately, especially, $|\alpha| \to 0$ as $|\gamma| \to 0$ is negligible if $|\gamma|$ is sufficiently large according to Theorem 4.2.

**Theorem 5.1.** If $x_1, \ldots, x_m$ are realization of a stationary ergodic vector process $X_1, \ldots, X_m$, then for any $\alpha$,

$$\hat{y}_\alpha \approx \frac{1}{2^n} \sum_{k=1}^{m} \tilde{\omega}_\alpha(x_k) \ln \frac{1}{p_0(x_k)} \Delta_x$$

$$\to y_\alpha \left( \frac{1}{2^n} \sum_{\gamma \in \mathbb{Z}^n} \sum_{k=1}^{m} \tilde{\omega}_{\alpha - \gamma}(x_k) \Delta_x \hat{y}_\gamma \right) \approx \hat{y}_\alpha$$

as $m \to \infty$, where $\hat{y}_\alpha$ is an approximation of $D_\alpha$.

**Proof.** From the estimate of $|\ln 1/p_0(x_k)| = |E(x_k; y)| \leq \sum_{|\alpha|} |y_\alpha| = ||E||_{H_1}$, Remark 5.1 gives the convergence of

$$\hat{y}_\alpha - \frac{1}{2^n} \int_{\mathbb{T}^n} \tilde{\omega}_\alpha(x) \ln \frac{1}{p_0(x)} dx = y_\alpha$$

as $m \to \infty$. In addition, by Remark 5.1 by (24), and by the boundedness of $|\tilde{\omega}_\alpha| = 1$,

$$\frac{1}{2^n} \sum_{\gamma \in \mathbb{Z}^n} \sum_{k=1}^{m} \tilde{\omega}_{\alpha - \gamma}(x_k) \Delta_x \hat{y}_\gamma \to \sum_{\gamma \in \mathbb{Z}^n} y_\gamma \frac{1}{2^n} \int_{\mathbb{T}^n} \tilde{\omega}_{\alpha - \gamma}(x) dx$$

$$= \sum_{\gamma \in \mathbb{Z}^n} y_\gamma \delta_\gamma(\gamma)$$

$$= y_\alpha$$
as \( m \to \infty \), where \( \delta_\alpha(\gamma) \) is a Dirac delta. Therefore, the proof is complete.

Theorem 5.1 says bidirectional communication between neurons during learning. Red dashed arrows of Figures 2 denote connections between \( y_a \) and \( y_e \), the connection is defined by

\[
\mathbf{\nabla}_\alpha(x) \equiv \frac{1}{2^n} \sum_{k=1}^{m} \tilde{\omega}_{\alpha - \gamma}(x_k) \Delta x_k.
\]

We define a learning rate for frequency by the inverse proportion to the maximum of differences between samples, i.e.,

\[
\frac{1}{\text{max}_k |\mathbf{T}(k)|}.
\]

Remark 5.2. (i) The ergodic property in Theorems 5.1 and 5.2 can be replaced with the i.i.d. property.

(ii) At the ultimate time, the learning process is complete and every neuron will be independent of each other.

(iii) To derive a governing probability, its limiting distribution have to be differentiable. Especially, in Theorem 5.2 \( p_0 \) must be analytic. Otherwise, a suitable analytic approximation could be replaced with \( p_0 \). We present an example in section 9.

Because all samples come from \( p_0 \), the lack of samples may cause inaccuracy to calculate \( y_a \) of Theorems 5.1 and 5.2. Although there are several sampling methods, e.g., Metropolis-Hastings sampling, Gibbs sampling, both depend on prior information and have accumulated error. In the following section, we introduce another sampling method which is very effective, whose convergence is also proven mathematically.

6. Dynamical Systems

In this section we introduce the Koopman mode decomposition (KMD) for a dynamical system as a preprocessing method for samples. We briefly review the results of Rowley et al. [13]. The KMD has two remarkable properties [13] Rowley, Mezić, Bagheri, Schlatter, Henningson]. First, it could remove redundant features and noise. Second, we could extract an amount of samples from snapshots, even if a sample comes from a nonlinear dynamical system.

We derive a dynamical system for time series of random vectors. Let \( X_t = (X_{t,1}, \ldots, X_{t,n}) \) be a random vector with continuous time \( t \) from \( p_0 \) and \( F(x_{t,j}) \) the cumulative distribution function of \( X_{t,j} \). Then

\[
\partial_x F(x_{t,j}) = p_0(x_{t,j}), \quad \frac{d}{dt} F(x_{t,j}) = p_0(x_{t,j}) \tilde{x}_{t,j},
\]

where \( p_0(x_{t,j}) \) is the marginal probability distribution of \( X_{t,j} \).

Theorem 6.1. Suppose that \( F(x_{t,j}) \) is continuously differentiable with respect to \( t \) and \( x_{t,j} \) for \( 1 \leq j \leq n \). Then the dynamical system for continuous time series of \( x_{t,j} \) is obtained by

\[
\tilde{x}_{t,j} = \Psi_j(x_{t,j}),
\]

where \( \Psi_j(x_{t,j}) = \frac{d}{dt} F(x_{t,j}) \).

For a discrete time series time of random vectors, which is our concern, let \( X_r = (X_{r,1}, \ldots, X_{r,n}) \) be a discretization of \( (X_t) \). From (29), the dynamical system for \( x_{r,j} \) \((r \in \mathbb{Z}^+)\) is approximated by

\[
x_{r+1,j} = x_{r,j} + \frac{d}{dt} F(x_{r,j}) \tilde{h}_{r,j}
\]

and put

\[
x_{r+1} = \tilde{\Psi}(x_r),
\]

where \( h_{r,j} \) is time step and \( \Psi_j(x_r) = (\Psi_1(x_r), \ldots, \Psi_n(x_r)) \) is a mapping on \( M \subset \mathbb{R}^n \) of an invariant compact manifold.

The space \( L^2(\mu) \) consists of complex valued \( L^2 \)-functions on \( M \) with \( d\mu \equiv p_0(x) dx \) and the \( L^2 \)-norm. For a subspace \( \mathcal{H} \) of \( L^2(\mu) \), the Koopman operator \( \mathcal{K} : \mathcal{H} \to \mathcal{H} \) is defined by

\[
\mathcal{K} \phi = f \phi \Psi,
\]

We call \( \lambda \in \mathbb{C} \) an eigenvalue of \( \mathcal{K} \) associated with eigenfunction \( \phi \in \mathcal{H} \) if \( \mathcal{K}\phi = \lambda \phi \).
Remark 6.1. (i) In the dynamical system of (30), it is hard to calculate \( \Phi \), because the cumulative distribution function \( F \) does not reveal itself.

(ii) The Koopman operator is a linear, i.e., \( \mathcal{K}(af + bg) = a\mathcal{K}f + b\mathcal{K}g \) \((f, g \in \mathcal{H}, \ a, b \in \mathbb{C})\).

To avoid mathematical ambiguity, we need two fundamental assumptions on \( \mathcal{K} \). First, \( \mathcal{K} \) is bounded on \( \mathcal{H} \) which is dense in \( L^2(\mu) \). Second, the spectrum of \( \mathcal{K} \) consists of only discrete spectrum. If \( f \) is an infinite linear combination of Koopman eigenfunctions, then

\[
\mathcal{K}^r f(x) = \sum_{k=1}^{\infty} \lambda_k^r v_k \phi_k(x) \tag{32}
\]

for \( r \in \mathbb{Z}^+ \), where \( (\lambda_k, \phi_k) \) is a pair of the Koopman eigenvalue and corresponding eigenfunction, \( v_k \) a coordinate sequence of \( r \) relative to Koopman eigenfunctions (13).

Putting \( E(x) = [f_1(x), \ldots, f_n(x)]^\top \), namely, an ensemble of observables in \( \mathcal{H} \), where \( n^\prime \) is the number of \( E \), we have

\[
\mathcal{K}^r E = E(\Psi^r) = \sum_{k=1}^{n^\prime} \lambda_k^r v_k \phi_k \tag{33}
\]

for \( r \in \mathbb{Z}^+ \), where a column vector of \( v_k \in \mathbb{C}^n \) is the \( k \)th coordinates of \( E \) relative to Koopman eigenfunctions, which is called the \( k \)th Koopman mode with respect to \( (\lambda_k, \phi_k) \).

We say that a measure preserving mapping \( \Psi \) is ergodic with respect to \( \mu \) when either \( \mu(E) = 0 \) or \( \mu(E) = 1 \) for any measurable set \( E \subset M \) with \( \Psi^{-1}(E) = E \). Birkhoff ergodic theorem affirms that if \( f \in L^2(\mu) \subset L^1(\mu) \), then

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{r=0}^{N-1} f \circ \Psi^r(x) = \int_M f \ d\mu \quad \text{a.e.} \ x \in M. \tag{34}
\]

Write the measurement of values of \( E \) along a single trajectory of \( \Psi \) starting at an initial vector \( x \in M \) as

\[
Y = \begin{pmatrix}
E(x) & E(\Psi(x)) & \ldots & E(\Psi^{m-1}(x))
\end{pmatrix},
\]

\[
Y' = \begin{pmatrix}
E(\Psi(x)) & E(\Psi^2(x)) & \ldots & E(\Psi^m(x))
\end{pmatrix}
\]

which consist of rows of data snapshots. Put \( A = Y'Y^+ \), where \( Y^+ \) is the Moore-Penrose pseudo-inverse of \( Y \). According to Remark 5.1, we have the following relation between the KMD and dynamic mode decomposition (DMD) (13) Rowley, Mezić, Bagheri, Schlatter, Henningson).

Theorem 6.2. Suppose that (30) is ergodic and \( E \) spans an \( n_0 \)-dimensional invariant subspace. Let \( (\lambda_k, \phi_k) \) be a pair of eigenvalue and corresponding eigenfunction of \( \mathcal{K} \). If \( \phi_k \) belongs to the span of \( f_1, \ldots, f_n' \), i.e.,

\[
\phi_k = w_1 f_1 + \cdots + w_n' f_n' = w \cdot E,
\]

then for almost every \( x, w \in \mathcal{E}(\text{orbit}(x)) \), which is a left eigenvector of \( A \) with eigenvalue \( \lambda_k \).

On a finite orbit \( (\hat{\Psi}^k(x))^{m-1}_{k=0} \), eigenvalues and eigenvectors of \( A \) approximate Koopman eigenvalues, eigenfunctions, and modes. Those can be derived by the dynamic mode decomposition (DMD) that was defined by Schmid and Sesterhenn in [13] Schmidt, Sesterhenn] and [14] Schmidt to extract the spatial flow structures that evolve linearly with time.

The DMD algorithm can be used to compute the augmented modes which approximate the Koopman modes. From (12) Tu, Rowley, Luchtenburg, Brunton, Kutz], by taking the singular value decomposition put \( Y = U \Sigma V^+ \). Then, \( A = Y' \Sigma V^{-1} U^+ \).

For computational efficiency, take projection \( \tilde{A} \) onto the mode of proper orthogonal decomposition (POD), by \( \tilde{A} = U^+ \Sigma U = Y' \Sigma V^{-1} W \), where columns of \( W \) are eigenvectors and \( A \) is a diagonal matrix containing the corresponding eigenvalues \( \lambda_k \). Thus, the eigenvectors of \( A \) (DMD modes) are given by columns of \( \Phi \),

\[
\tilde{\Phi} = Y' \Sigma V^{-1} W.
\]

With the low-rank decomposition in hand, the projected future solution can be constructed for all time. With time step \( h_t \), the solutions at all future times are approximated by

\[
y_t \approx \Phi \exp(\Omega t) B, \tag{35}
\]

where \( \Omega = \text{diag}(\ln(\lambda_k)/h_t) \) is a diagonal matrix, and \( B \) consists of the initial amplitudes of each mode, precisely, \( B = \Phi^+ y \) at time \( t = 0 \).

As a consequence of Theorem 6.2 and (35), the inversion \( E^{-1} \) from observables back to state-space yields \( x_r \).

Corollary 6.2.1. If there exists the inverse of \( E \) such that \( E(x) = y \), then \( x_r \approx E^{-1}(\Phi \exp(\Omega t) B) \).

If \( E \) is identified, e.g., \( f_k(x) = x_k \), then Corollary 6.2.1 generates data which follows (30) and we obtain sufficiently large size of data. For more information for the KMD and DMD, refer to (11) Korda, Mezić, (4) Arbabi, Mezić.

Remark 6.2. (i) The ergodic property in Theorem 6.2 can be replaced with the I.I.D. property.

(ii) The ergodicity allows us to reduce in computational cost in contrast to the I.I.D. property and to raise accuracy up for learning.

7. \( L^1 \)-Stability of Empirical Distribution Functions

In this section, it will be discussed the convergence of empirical distribution functions for sample vectors. Suppose that a stochastic vector process of \( (X_r) \) from a governing probability \( \rho_X \) with values in \( M \). Let \( F \) be the (multivariate) cumulative distribution function of \( X \). Since the set of discontinuities is an \( F_0 \) set, by Cavalieri’s principle, the componentwise monotonicity of \( F \) permits continuity except possibly measure zero set.

The multivariate empirical distribution function of \( (X_r) \) is given by

\[
\hat{F}_m(x) = \frac{1}{m} \sum_{r=1}^{m} \mathbb{I}_{X_r \leq x}. \tag{36}
\]

where \( X_r \leq x \) means \( X_r \leq x \) componentwise. If \( (X_r) \) is a sequence of I.I.D. random vectors, then \( \hat{F}_m(x) \) converges to \( F \) in the supremum norm by Vapnik-Chervonenkis theorem (16) p.823, Example 1 of p.833] a generalization of Glivenko-Cantelli theorem.

For a sequence of locally integrable functions, we say that the sequence is \( L^1 \)-stable if the sequence is a Cauchy sequence in \( L^1(\mathbb{R}^n) \). We prove the \( L^1 \)-stability of multivariate empirical distribution functions of (36) on \( \mathbb{R}^n \), even if those may not be integrable.
Theorem 7.1. If \((X_i)\) is a stationary ergodic process from \(p_0\) on \(\mathbb{R}^n\) such that \(\mathbb{E}(|X_{1,j}|) < \infty\) \((1 \leq j \leq n)\), then \(\hat{F}_m\) converges pointwise to \(F\) almost everywhere and is \(L^1\)-stable.

Proof. For each \(x\) put \(f_x(\cdot) = \mathbb{1}_{\leq x} \in L^1(\mu)\), where \(d\mu(x) = p_0(x)\,dx\). By Birkhoff ergodic theorem, there exists a set of almost all initial samples such that
\[
\hat{F}_m(x) = \frac{1}{m} \sum_{r=1}^m f_x(X_r) \to \mathbb{E}(f_x) = F(x) \quad (37)
\]
at almost every \(x\) as \(m \to \infty\). In fact, the componentwise monotonicity of \(\hat{F}_m\) implies the componentwise monotonicity of the limit of \(\hat{F}_m\). By Caravelli’s principle, the limit is continuous except at most measure zero set.

As \(\hat{F}_m - \hat{F}_{m'}\) is a linear combination of indicator functions, it vanishes if a vector variable is larger than every \(x_k\). It supports compactly and belongs to \(L^1(\mathbb{R}^n)\), accordingly the integral of \(|\hat{F}_m - \hat{F}_{m'}|\) is well defined for \(m\) and \(m'\). Let \(R\) be a positive integer and split \(\mathbb{R}^n\) into two parts of
\[
\mathbb{R}^n = [-R, R]^n \cup \mathbb{R}^n \setminus [-R, R]^n.
\]

Then
\[
\int_{\mathbb{R}^n} |\hat{F}_m - \hat{F}_{m'}|\,dx = \int_{[-R, R]^n} |\hat{F}_m - \hat{F}_{m'}|\,dx + \int_{\mathbb{R}^n \setminus [-R, R]^n} |\hat{F}_m - \hat{F}_{m'}|\,dx \quad (38)
\]

Estimate of \(I\): By the compactness of \([-R, R]^n\), the Lebesgue dominated convergence theorem and Birkhoff ergodic theorem imply
\[
I \to 0 \quad (39)
\]
as \(m, m' \to \infty\).

Estimate of \(II\): The region can be written the union of two part separations, precisely,
\[
\mathbb{R}^n \setminus [-R, R]^n = \bigcup_{k=0}^{n-1} \mathbb{R}^k \times (\mathbb{R} \setminus [-R, R]) \times \mathbb{R}^{n-1-k}.
\]

Let \(b = \{0, 1\}\) be a binary multi-index and \(b'\) the negation of \(b\). The notation of \(x^b\) means the selection of corresponding components \(x_j\) when \(b_j = 1\). By Fubini’s theorem,
\[
II \leq \sum_{|b|=1} \int_{\mathbb{R}^n \setminus [-R, R]} |\hat{F}_m - \hat{F}_{m'}|(dx)^b(\,dx)^{b'} \equiv \sum_{|b|=1} \int_{\mathbb{R}^n \setminus [-R, R]} V_{m,m'}(x^b')(\,dx)^{b'}, \quad (40)
\]

where \((dx)^b = dx^b\) and
\[
V_{m,m'}(x^b') = \int_{\mathbb{R}^n \setminus [-R, R]} |\hat{F}_m - \hat{F}_{m'}|(\,dx)^b.
\]

Estimate of \(V_{m,m'}(x^b')\): By the change of variables,
\[
V_{m,m'}(x^b') = \int_{(\mathbb{R}^n \setminus [-R, R])^b} |\hat{F}_m - \hat{F}_{m'}|(\,dx)^b + \int_{[R, \infty)^b} |\hat{F}_m - \hat{F}_{m'}|(\,dx)^b
\]

where \(-1 \odot x^b\) is the componentwise multiplication only for non-zero components of \(b\). By the triangle inequality, the last sum of integrals is bounded by
\[
\int_{[R, \infty)^b} |\hat{F}_m(\cdot) - \hat{F}_{m'}(\cdot)|(dx)^b + \int_{[R, \infty)^b} |\hat{F}_m - \hat{F}_{m'}|(\,dx)^b = \int_{[R, \infty)^b} \hat{v}_m(x)(\,dx)^b + \int_{[R, \infty)^b} \hat{v}_{m'}(x)(\,dx)^b \equiv A_b + B_b, \quad (41)
\]

Estimate of \(A_b\) and \(B_b\). Since \(\hat{F}_m(x)\) and \(\hat{F}_{m'}(x)\) are sums of products of Heaviside functions, by Birkhoff ergodic theorem,
\[
A_b = \frac{1}{m} \sum_{i=1}^m \int_{[R, \infty)^b} (\mathbb{1}_{X_i^b \geq R} - \mathbb{1}_{X_i^b < R}) \, dx \mathbb{E}((\mathbb{1}_{X_i^b < R})\mathbb{1}_{X_i^b \geq R})_{p_0} \quad (42)
\]
as \(m' \to \infty\). Similarly,
\[
B_b \to \mathbb{E}((\mathbb{1}_{X_i^b < R})\mathbb{1}_{X_i^b \geq R})_{p_0} \quad (43)
\]
as \(m' \to \infty\). Finally, according to \([38] \sim [43]\),

\[
\limsup_{m, m' \to \infty} \int_{\mathbb{R}^n} |\hat{F}_m - \hat{F}_{m'}|\,dx \leq \limsup_{m, m' \to \infty} \sum_{|b|=1} \int_{\mathbb{R}^n \setminus [-R, R]} V_{m,m'}(x^b')(\,dx)^{b'}
\]

\[
\leq \sum_{|b|=1} \limsup_{m, m' \to \infty} \int_{\mathbb{R}^n \setminus [-R, R]} V_{m,m'}(x^b')(\,dx)^{b'} = 2 \sum_{|b|=1} \int_{\mathbb{R}^n \setminus [-R, R]} \mathbb{E}((\mathbb{1}_{X_i^b < R})\mathbb{1}_{X_i^b \geq R})_{p_0} (\,dx)^{b'} = 2 \int_{\mathbb{R}^n \setminus [-R, R]} \mathbb{E}((\mathbb{1}_{X_i^b < R})\mathbb{1}_{X_i^b \geq R})_{p_0} (\,dx)^{b'} \quad (44)
\]

By the Lebesgue dominated convergence theorem, we can interchange the limit of \(R \to \infty\) and two integrals of \([44]\). Then, the integrand of \([44]\) goes to 0. Therefore, \(\hat{F}_m\) is \(L^1\)-stable for almost every initial sample.

8. Estimations

We describe how the designed and learned neurons respond to a signal. For convenience, we define a lexicographical ordering to classify neurons. For \(\alpha, \beta \in \mathbb{Z}^n\), define the ordering by

\[
\alpha < \beta \text{ if either } |\alpha| = \sum_{i=1}^n |\alpha_i| < \sum_{i=1}^n |\beta_i| = |\beta| \text{ or } \sum_{i=1}^n |\alpha_i| = \sum_{i=1}^n |\beta_i| \text{ and } \alpha_i < \beta_i
\]

for the largest \(i\) such that \(\alpha_i \neq \beta_i\). If \(\alpha_i = \beta_i\) for all \(i\), then we define \(\alpha = \beta\). For example, let \(n = 2\). Putting \(\hat{m} = -m\) for \(m > 0\), we have
of A and B. The active path for \( x \) is defined by

\[
\Gamma = \bigcup_{k=0}^{\infty} \{ \Gamma_k \}.
\]

We say that any subset \( \mathcal{D} \) of indexes is a dictionary if an index size is at most \( N \) so that \( \| F_{m\mathcal{D}} - F_{m\mathcal{D}'} \|_1 < \epsilon \) for \( m, m' \geq N \).

Usually, an active path is considered in the dictionary. So, we call \( \Gamma_{\mathcal{D}} \equiv \Gamma \cap \mathcal{D} \) the active path for \( x \) with respect to \( \mathcal{D} \).

**Active path.** We will select the best matching neurons for a signal, namely, the active path. Those are chosen so that the likelihood of \( x \) is maximized in the following way. From \( e^{-E} \equiv e^{-\text{Re}(E)} \), collect indexes \( \alpha \) such that \( \text{Re}(C_w(k) \odot V_w(k))_\alpha < 0 \) \( (\Gamma_w(k))_\alpha \) and put it by

\[
\Gamma_k = \left\{ \alpha \right\},
\]

where \( (A \odot B)_\alpha \) is the \( \alpha \)th component of the Hadamard product

\[
\begin{align*}
\text{Figure 4: Block diagram for learning: Part I is the area to generate more samples. Part II controls the KMD and the learning process is achieved in Part III.}
\end{align*}
\]

\[
\begin{align*}
\text{Figure 5: Block diagram for estimation: Part I consists of a signal and a dictionary of indexes, in which the evaluation vector and likelihood are computed. Part II chooses the indexes at which neurons have a negative projection to maximize a network probability. In Part III, an active path for the signal is estimated.}
\end{align*}
\]

**Probability of active neuron.** For a frequency learning, we define the probability of active neurons (POAN) on \( \Gamma_{\mathcal{D}} \) by

\[
\left\{ \left. p_{\mathcal{D}}(x) \middle| \alpha \right. \in \Gamma_{\mathcal{D}} \right\}.
\]

For a moment learning, in extension to definition, each probability of active neurons must be 1.

**Likelihood.** The likelihood of \( x \) is a non-negative real number induced by \( p(x \mid y) \). The value means the possibility that \( x \) belongs to the network. The value is relatively large if and only if \( x \) may be likely to happen in the network. See Examples 8.1 and 8.2 and Examples of section 9.

With energy functions assumed in advance, in the following examples we will explain how to calculate estimations of the network, i.e., an active path, probability of active neuron, and likelihood. Figures 4 and 5 show the learning and estimation procedures, respectively.

**Example 8.1 (Frequency learning).** Suppose that \( E \) is given by

\[
E(x_1, x_2; y) = -3 \sin \pi x_1 + \cos \pi x_2 + 2 \cos 2\pi x_2
\]

and \((1/12, 1/6)\) is a signal to be estimated by the network.
The energy function has the indexed form of

\[-3 \sin \pi x_1 + \cos \pi x_2 + 2 \cos 2\pi x_2\]

\[= \frac{3}{2}(\omega_{10} - \omega_{10}) + \frac{1}{2}(\omega_{01} + \omega_{01}) + \omega_{02} + \omega_{02},\]

the value of the partition function

\[Z = \int_{\mathbb{R}^2} e^{-E} dx = 47.9883,\]

and the DC value \(D_c = \ln Z = 3.8710.\) Thus, the network probability is given by

\[p(x_1, x_2 \mid y) = e^{-3.8710 + 3 \sin \pi x_1 - \cos \pi x_2 - 2 \cos 2 \pi x_2}\]

(see Figure 6). Non-trivial cells are

\[C_{\Omega}(0) = (3.8710), \quad C_{\Omega}(1) = \begin{pmatrix} \frac{1}{2} \\ -3/2 \\ 1/2 \end{pmatrix}, \quad C_{\Omega}(2) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},\]

and evaluation vectors that we need for \((1/12, 1/6)\) are

\[V_{\omega}(0) = (1), \quad V_{\omega}(1) = \begin{pmatrix} (1 - \sqrt{3})/2 \\ (\sqrt{3} - 1)/2 \\ (\sqrt{3} + 1)/2 \end{pmatrix}, \quad V_{\omega}(2) = \begin{pmatrix} (1 - \sqrt{3})/2 \\ (\sqrt{3} - 1)/2 \\ (\sqrt{3} + 1)/2 \end{pmatrix}.\]

Hence, for \((1/12, 1/6)\), the active path, POAN of \(\Gamma\), and likelihood are as follows.

\[\Gamma = \left\{ \frac{10}{10}, \frac{20}{20} \right\}, \quad \text{POAN}(\Gamma) = \left\{ \frac{1}{2}, \frac{1}{2}, \frac{1}{2} \right\},\]

and

\[p(1/12, 1/6 \mid y) = 0.007,\]

respectively.

Since the likelihood is relatively small (Figure 6), we may guess that \((1/12, 1/6)\) may happen with small possibility in the network.

**Example 8.2 (Moment learning).** Suppose that \(E\) is defined by

\[E(x_1, x_2; y) = -0.5 x_1 - 2 x_2 + 4 x_1 x_2 + 3 x_2^2\]

and \((-4/5, 4/5)\) is a signal to be estimated by the network.

The indexed form of \(E\) is

\[-0.5 x_1 - 2 x_2 + 4 x_1 x_2 + 3 x_2^2\]

\[= -0.5 x_1^{10} - 2 x_1^{11} + 2 x_2^{12} + 3 x_2^{12},\]

its value of the partition function

\[Z = \int_{\mathbb{R}^2} e^{-E} dx = 4.2883\]

and the DC value \(\ln Z = 1.4559 = D_c\). So, the network probability is given by

\[p(x_1, x_2 \mid y) = e^{-1.4559 + 0.5 x_1 + 2 x_1 x_2 - 3 x_2^2}\]

(refer to Figure 7). All non-trivial cells are

\[C_{\Omega}(0) = (1.4559), \quad C_{\Omega}(1) = \begin{pmatrix} 0.5 \\ 2 \end{pmatrix}, \quad C_{\Omega}(2) = \begin{pmatrix} 0 \\ -4 \\ 0 \end{pmatrix},\]

and some evaluation vectors that we need for \((-4/5, 4/5)\) are

\[V_{\omega}(0) = (0), \quad V_{\omega}(1) = \begin{pmatrix} 4/5 \\ 4/5 \end{pmatrix}, \quad V_{\omega}(2) = \begin{pmatrix} 16/25 \\ -16/25 \\ -16/25 \end{pmatrix}.\]

Hence, for \((-4/5, 4/5)\), the active path, likelihood are

\[
\Gamma = \left\{ (10), (20) \right\}, \quad p(-4/5, 4/5 \mid y) = 1.4683,
\]

respectively. The likelihood is relatively large (Figure 6), and so, we may guess that \((-4/5, 4/5)\) may happen with large possibility in the network.

**INTERPRETATION OF ESTIMATIONS.** To simplify what the active path says we give a topological structure on \(\Gamma_D\). Setting \(d(y_\alpha, y_\beta) = |\alpha - \beta|\) on \(\Gamma_D\), we define the number of basic open balls \(N : \mathbb{Z} \rightarrow \bar{\mathbb{N}}\) by

\[N(k) = \frac{1}{2} |k\text{-topology}|,\]

where

\[k\text{-topology} = \left\{ (\alpha, \beta) : d(y_\alpha, y_\beta) = k \right\} \text{ on } \Gamma\]

and \(\bar{\mathbb{N}}\) is the extended natural numbers.

The definition \(45\) could be changed according to learning models. For more methods, refer to [6].
9. APPLICATION TO A STANDING WAVE

A standing wave (or a stationary wave) appears on the surface of a liquid in a vibrating container or on vibrating strings, is oscillation in time but whose peak amplitude outline does not move in space. The locations at which the amplitude is minimum are defined as nodes, and the locations where the amplitude is maximum are defined as antinodes.

For the realization $x$ of the amplitude random variable $X$, the differential equation of the normalized wave is represented by $\ddot{x} + x = 0$, where $-1 \leq x \leq 1$. Putting $x_1 = x$ and $x_2 = \dot{x}$, we have the dynamical system of

$$\frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

which contains the standing wave.

9.1. FREQUENCY LEARNING. Suppose that there are 32 samples with time step $h_t = 0.2$ for $0 \leq t \leq 2\pi$ in Figure 8.

From snapshots of Figure 8, the KMD enables us to generate amplitude samples of size 6,284 ($h_t = 0.001$) as Figure 10. We regard $\hat{F}_{m,284}$ as the limiting distribution in Figure 11 by cubic spline interpolation. With the assumption of differentiability of $p$ with respect to $x$, by taking partial derivatives, we approximate the governing probability $p_0$ as Figure 12.

\[ |\Upsilon_{\alpha}(\gamma)| \text{ for } \alpha = -50, 0, \text{ and } 50. \]
According to (25), every connection function of $\Upsilon_\alpha(\gamma) = \frac{1}{2} \sum_{k=1}^{284} \exp(-\pi i (\gamma - \alpha) x_k) \cdot 0.001$

is a translation of $\Upsilon_0$. Figure 15 shows three kinds of absolute values of $\Upsilon_\alpha$. By (26), the learning rate of frequency with the sample regenerates is as follows,

$$\frac{1}{\max_\alpha \text{Var}(\Upsilon_\alpha)} = \frac{1}{\text{Var}(\Upsilon_0)} = 221.62.$$  

**INTERPRETATION OF ESTIMATIONS.** Let $\mathcal{D}$ be a dictionary of $|\alpha| \leq 100$. Red dots of Figures 14 denote the active neurons for $x = 0.75$, $1$, $1.5$, and $3$, respectively.

Active neurons have specific patterns formed by bundles of neighboring neurons. In pictures of Figure 14 we draw neighborhood groups of red areas of rectangles with rounded corners on positive index set because of symmetry. As $x$ increases to 1, the adjacent groups of active neurons are distributed between frequencies 10 and 30 in (b) of Figure 14. On the other hand, if $x$ is far from 1, the number of groups decreases, whereas the ball size increases. Moreover, if $x = 3$, then there is no such groups between 10 and 30 in (d) of Figure 14.

In addition, we examine topological statistics on $\Gamma_\mathcal{D}$ for $x = 0.75$, $1$, $1.5$, and $3$. In the union of $k$-balls of each active path, the number of balls, mean, and variance of indexes are compared in Figures 15, 16, and 17, respectively.

Finally, from the network probability,

$$\arg \max_{x \in \mathbb{R}} p(x \mid y) = \arg \max_{-1 \leq x \leq 1} p(x \mid y) = \pm 1,$$

i.e., the likelihood has the maximum at the node and antinode. Here, a random variable of amplitude for a standing wave has the maximum likelihood at peaks. With all $y_m$, (47) and $p_0(x)$ are identical. We can calculate likelihoods $p(x \mid \mathcal{D})$ if the dictionary is main focus of interest.

9.2. **MOMENT LEARNING.** We derive the network probability for $(x, \dot{x})$ as a moment learning and analyze estimations of the network through the topological interpretation. Suppose that there are 32 samples of $(x, \dot{x})$ drawn from (46) with time step $h_t = 0.2$ for $0 \leq t \leq 2\pi$ in Figure 18. By (39), the empirical distribution function $F_{32}(x, \dot{x})$ is shown in Figure 19.

Using the KMD with $h_t = 0.001$, we gain 6,284 samples (Figure 20) with which we draw the limiting distribution $F(x, \dot{x})$ and governing probability $p_0(x, \dot{x})$ in Figures 21 and 22, respectively. Here, $p_0$ is calculated by partial derivatives of a finite-difference method at centers of $300 \times 300$ equally spaced bins.

The sample has a uniform time step $h_t = 0.001$. By (28) the learning rate for moment is given by

$$\frac{1}{\max_k \|h_k\|} = \frac{1}{0.001} = 1000.$$

According to topological statistics of Figures 15 $\sim$ 17 mean- and variation-distributions have unstable behaviors on small balls when $x = 1$. Note that $x = 1$ is the magnitude of an antinode. A loss of stem in figures is caused by absence of $k$-ball in $\Gamma_\mathcal{D}$, and so, there does not exist any topological statistic at the position.
On the other hand, the induced $p_0$ is almost cylindrical and every partial derivative of $p_0$ vanishes at the origin. More precisely, $p_0$ is an approximation of singular measure which is supported in the unit circle $S^1$ and

$$\int_{\mathbb{T}^2} p_0(x, \dot{x}) \, dx \, d\dot{x} = 1.$$ 

Thus, $\ln 1/p_0$ does not belong to $A_0$ on $\mathbb{T}^2$. For this reason, the energy function cannot be expanded as a power series near the origin.

To avoid a dead end, we adopt an auxiliary function as a pullback limit of $p_0$. Since $p_0$ could be regarded as the limit of suitable analytic functions, take an approximation of $p_0$, for example,

$$p_a(x, \dot{x}) = \begin{cases} C(1 - x^2 - \dot{x}^2)^{-1/2} & \text{if } x^2 + \dot{x}^2 \leq 1 \\ 0 & \text{otherwise}, \end{cases} \quad (48)$$

where $C$ is chosen such that $\|p_a\|_{L^1} = 1$ (see Figure 23).

Table I presents some examples of likelihoods of $p_a(x, 0 \mid y)$ and $p_0(x, 0)$ with zero velocity. All express that $(1, 0)$ has the maximum likelihood when $\dot{x} = 0$.

| $(x, 0)$ | $p_a(x, 0 \mid y)$ | $p_0(x, 0)$ |
|----------|------------------|------------|
| $(0, 0)$ | 0.1592           | 0          |
| $1/\sqrt{2}$ | 0.2221       | 0          |
| 1        | $\infty$       | 23.9926    |
| $|x| > 1$ | 0               | 0          |

Table I: Likelihoods.

Figure 15: The number of $k$-distance open balls on each active path for $x = 0.75, 1, 1.5, 3, 0$. Distribution of the number of $k$-topology for $x = 3, 0$ are more regular and simpler than others.

Figure 16: The means of $k$-distance neurons on active paths for $x = 0.75, 1, 1.5, 3, 0$. The mean-distributions for small balls for $x = 0.75, 1, 1.5$ are more unstable than others.

Figure 17: The variances of $k$-distance neurons on active paths for $x = 0.75, 1, 1.5, 3, 0$. The most unstable variance-distribution is of $x = 1$.

More precisely,

$$\arg \max_{x \in \mathbb{R}} p_a(x, 0 \mid y) = \arg \max_{x \in \mathbb{R}} p_0(x, 0) = \pm 1,$$

and thus, we conclude that the likelihood of $(\pm 1, 0)$ has the maximum when $\dot{x} = 0$. In general, for $-1 \leq \dot{x} \leq 1$, we obtain $x = \pm \sqrt{1 - \dot{x}^2}$ which have the maximum likelihood of $p_0(x, \dot{x})$. Therefore, $(x, \pm \sqrt{1 - x^2})$ is most likely in the network (refer to Figures 24 and 25). In fact, the relation between $x$ and $\dot{x}$ is equal to $x^2 + \dot{x}^2 = 1$ for $0 \leq x \leq 2\pi$, e.g., $\dot{x} = \cos t$ when $x = \sin t$.

Figure 18: Samples $(x, \dot{x})$ of size 32.

Figure 19: The empirical distribution function $\hat{F}_{32}(x, \dot{x})$. 
INTERPRETATION OF ESTIMATIONS. Let \( \mathcal{D} \) be a dictionary of indexes \( \alpha \) such that \( 0 \leq \alpha_j \leq 20 \) (\( j = 1, 2 \)). The partial derivatives of \( p_\alpha \) at \((0, 0)\) yield non-trivial neurons. For estimation and interpretation, we consider 8 signals of \((x, \dot{x}) = (\pm 1, 0), (0, \pm 1), (\pm 1, \pm 1), \) and \((\pm 1, \mp 1), \) where double signs are in same order.

In Figures 24 and 27, we draw neurons and active paths induced from \( p_\alpha \), in which red dots denote active paths for signals. For simplicity, we adopt \( \| \cdot \|_1 \)-norm to generate the topology, with which we examine topological statistics on \( \Gamma_\mathcal{D} \). For each \( k \), on the collection of all \( k \)-balls, the number of balls, mean, and variance of indexes are compared in Figures 28, 29, and 30, respectively.

Note that all active paths do not contain any \( k \)-ball for \( k \geq 4 \). In Figures 29 and 30 each position has values of two components. The right and left sides toward the distance increasing are the quantities from \( \alpha_1 \) and \( \alpha_2 \), respectively.

There are plenty of mathematical research results for classification ([3], [7], [9], [12]). Those could be useful tools for interpreting unsupervised estimates.

**Remark 9.1.** Without reproducing process of data, we could approximate the neurons of a probabilistic neural network. Indeed, the drawn sample would give the similar distribution of neurons to the probabilistic neural network. In many cases we do not need thousands or more of data to calculate neurons. However, the accuracy of the derived governing probability may be poor and the learning rate is low, because the accuracy depends on the sample size and quality.
Figure 25: Both have the symmetric transposition of axes compared to Figure 24.

Figure 26: Both have non-trivial active neurons in contrast to Figure 25. Active neurons of (a) are negative but not (b). Topological statistics are different (Figures 28 ∼ 30).

Figure 27: Both are similar to each other. Topological statistics are also different (Figures 28 ∼ 30).

Figure 28: The number of $k$-balls on each active path. The numbers of $k$-balls for $(1, 0), (0, 1)$ are smaller.

Figure 29: The mean-distributions of $k$-balls on each active paths. Means for $(1, 0), (-1, 0)$ are similar to each other and so do $(0, 1), (0, -1)$. In addition, Two groups change positions with each other.
Figure 30: The variance-distributions of $k$-balls on each active paths. Variances for $(1, 0)$, $(-1, 0)$ are similar to each other and so do $(0, 1)$, $(0, -1)$. In addition, two groups change positions with each other.

10. Conclusion and Further Study

The probabilistic neural network as an unsupervised learning model provides globally unique learning model as an optimal solution for observed samples in probability. The networks are essentially data driven models that do not depend on the backpropagation because our proposed networks are not trained through iterative computation. Hereby, these network models would be applied in a direction where quality is more important than amount of data in industry, e.g., such as financial service and bioengineering industries. These networks provide the learning rates which depend on the drawn samples, and thus, a probabilistic neural network maximizes the utilization of data. The application of probabilistic neural network technologies, especially for medical image classification, disease prediction, and object recognition, could provide accurate and reliable results.

To appreciate the estimation of artificial intelligence we may need certain physical interpretations and analysis of active neurons. For example, a suitable topology on the collection of all active paths or a dictionary containing a meaningful concept, e.g., the standard hearing range 20 to 20,000 Hz for humans, the $\alpha$-brainwave range 8 to 12.99 Hz, the first-order moments for (conditional) expectations, etc.

REFERENCES

[1] H. Arbabi and I. Mezić, Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator. SIAM J. Appl. Dyn. Syst. 16(4), 2096–2126, 2017.
[2] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. IEEE Trans.Pattern Anal. Machine Intelli. (35), pp. 1798–1828, 2013.
[3] C. M. Bishop, Pattern recognition and machine learning. Springer, 2006.
[4] G. Cybenko. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 2, pp. 303–314, 1989.
[5] S. Furaoa and O. Hasegawa. An incremental network for on-line unsupervised classification and topology learning. Neural Networks 19, 90–106, 2006.
[6] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, F. A. Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. ArXiv abs/1706.06969, 2017.
[7] I. Goodfellow, Y. Bengio and A. Courville. Deep learning. MIT Press, 2016.
[8] L. Grafakos. Classical and modern Fourier Analysis Prentice Hall, 2003.
[9] S. Haykin. Neural networks and learning machines. Pearson, 2016.
[10] K. Hornik. Approximation capabilities of multilayer feedforward networks. Neural Networks 4(2), pp. 251–257, 1991.
[11] M. Korda and I. Mezić. On convergence of extended dynamic mode decomposition to the Koopman operator. Jour. Nonlinear Science 28, pp. 687–710, 2018.
[12] K. P. Murphy. Machine learning: A probabilistic perspective (Adaptive Computation and Machine Learning series). The MIT Press, 2012.
[13] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson. Spectral analysis of nonlinear flows. Jour. Fluid Mech. 641, pp. 115–127, 2009.
[14] P. Schmid. Dynamic mode decomposition of numerical and experimental data. Jour. Fluid Mech. 656, pp. 5–28, 2010.
[15] P. Schmid and J. Sesterhenn. Dynamic mode decomposition of numerical and experimental data. American Physical Society, 61st Annual Meeting of the APS Division of Fluid Dynamics, November 23–25, 2008, abstract id: MR.007.
[16] G. R. Shorack and J. A. Wellner. Empirical processes with applications to statistics. John Wiley & Sons, Inc., 1986.
[17] E. M. Stein. Harmonic Analysis: Real-variable methods, orthogonality, and oscillatory Integrals, Princeton University Press, 1993.
[18] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. Jour. Comp. Dynamics 1(2), pp. 391–421, 2014.
[19] D. Watson. The rhetoric and reality of anthropomorphism in artificial intelligence. Minds and Machines 29, pp. 417–440, 2019.

Kyung Soo Rim received the B.S. degree from Korea University (KU), Korea in 1990, the MS and Ph.D. degree in harmonic analysis of mathematics from Korea Advanced Institute of Science and Technology (KAIST), Korea in 1992 and 1996, respectively. He is a full professor in the department of mathematics at Sogang University in Seoul, Korea and served as the Dean of Academic and Admissions Affairs from 2015-18. His recent research interests focus on the probabilistic machine intelligence related to the Koopman operator of dynamical systems including the continuity of convolution and compositions between function spaces.

U Jin Choi received the B.S. degree from Seoul National University (SNU), Korea in 1974 and Ph.D. degree from Carnegie Mellon University in 1987, U. S. A., both in mathematics. He is currently emeritus professor at Korea Advanced Institute of Science and Technology (KAIST) and a member of Research Advisory Committee at National Institute of Mathematical Sciences (NIMS), Korea and served as a full professor in the department of mathematic sciences at KAIST until 2011. His current research interests are statistics, dynamical systems theory, and machine intelligence in viewpoint of mathematics including stochastic calculus.