Analysis on Noisy Boltzmann Machines and Noisy Restricted Boltzmann Machines

WENHAO LU\textsuperscript{1}, CHI-SING LEUNG\textsuperscript{1}, (Senior Member, IEEE), AND JOHN SUM\textsuperscript{2}, (Senior Member, IEEE)

\textsuperscript{1}Department of Electrical Engineering, City University of Hong Kong, Hong Kong
\textsuperscript{2}National Chung Hsing University, Taichung 402, Taiwan

Corresponding author: Chi-Sing Leung (eeleungc@cityu.edu.hk)

This work was supported in part by the Research Grant from City University of Hong Kong under Grant 9610431.

\section*{ABSTRACT}

The Boltzmann machine (BM) and restricted Boltzmann machine (RBM) models are representative stochastic neural networks, in which neuron states are determined by stochastic activation functions. They are widely used in many applications. However, when analog circuits are used to realize a neural network, noise are not avoidable. The noise could come from external environments, such as power supplies and thermal noise, and they will affect the neurons’ stochastic behaviour in the BM and RBM models. Hence it is important to theoretically study how the noise affect the operation of the BM and RBM models. To best of our knowledge, there are little works related to the analysis on noisy BMs and noisy RBMs. This paper considers that there are additive noise in the inputs of the neurons, and theoretically studies the behaviors of the two models under this imperfect condition. It is found that the effect of additive noise is similar to increasing the temperature factors of the two models. Since the input noise may make the networks to have wrong stochastic behaviour, there is Kullback Leibler (KL) divergence loss in noisy BMs and noisy RBMs. Based on the Gaussian-distributed noise assumption, a noise compensation method is proposed to suppress the effect of additive noise. Experiments show that the proposed noise compensation method can greatly suppress the KL divergence loss. In addition, from the experimental results, our method is also effective for handling non-Gaussian noise.

\section*{INDEX TERMS}

Activation function, Boltzmann machine, restricted Boltzmann machine, state distribution, additive noise.

\section*{I. INTRODUCTION}

Boltzmann machine (BM) is one of stochastic neural network models [11]–[3]. The aim of the BM model is to learn the probability distribution of training data. Many previous works reported its applications, like travelling salesman problems [4], speech processing [5], [6], and pattern recognition [7]–[9]. The restricted Boltzmann machine (RBM) model is a variant of BM, in which connection weights only exist between visible neurons and hidden neurons. The RBM model is very effective for feature extraction and is widely used in dimensionality reduction [10], classification [11], and image processing [12], [13]. In addition, it can act as the building blocks for deep neural networks [14].

Since the original training algorithms of the BM and RBM models are quite computational intensity, some fast training algorithms were developed [15]–[17]. Besides, some GPU-accelerated implementations were reported [18]–[20]. Apart from multi-core CPU or GPU acceleration, some hardware realizations and implementations were reported in [21]–[24]. Also, many neuron-friendly analog elements were developed in the last several years [22], [23], [25]. The aforementioned technology could make the realization of large-scale BMs or RBMs to be possible in the nearing future.

When we realize a neural network based on analog circuits, noise are not avoidable. The noise could come from external environments [26], such as power supplies, and/or thermal noise [27]. In addition, there are random offset drifts in operational amplifiers [28], which are key components in analog neural circuits. In the past, people expected that neural network models have strong resilience to noise or realization imperfection. In fact, for decades ago, many studies reported that without noise-aware procedures, a noisy neural network could have a very poor performance [29]–[35]. Therefore,
many studies on the properties of noisy networks and fault-aware learning algorithms were reported [29]–[35]. However, those results are restricted to feedforward network models only, such as multi-layered network [30], radial basis function network [36], and extreme learning machine [33].

In [37], Sum et al. studied the effect of additive weight and bias noise on BM, but the compensation method for making a BM with certain resilience to noise was not given. Also, the simulation example in [37] is rather simple. To best of our knowledge, there are little works related to the analysis on BM and RBM with input noise.

This paper analyzes the effect of input noise on the BM model, as well as the RBM model. We consider that the noise are time-variant and zero-mean Gaussian-distributed. We first derive the stochastic activation function for BMs and RBMs under Gaussian input noise and find that the effect of input noise is similar to increasing the model’s temperature factors. Hence there are Kullback Leibler (KL) divergence loss in the state distribution of BMs and RBMs. Such loss leads to performance degradation. A compensation scheme is then proposed to suppress the effect of noise. With this scheme, the behaviors of noisy BM and RBM are the same as those of noise-free BMs and RBMs. Simulations are then carried out to verify our theoretical findings.

This paper is organized as follows. Section II introduces the background of the BM and RBM models. Section III shows the noise models for BMs and RBMs. Section IV presents the analytical results and the proposed noise compensation scheme. Section V gives the experimental results. The paper is then concluded in Section VI.

II. BACKGROUND

A. BOLTZMANN MACHINE FOR CLASSIFICATION TASKS

The BM model is a kind of stochastic recurrent networks. The neurons’ states are governed by a stochastic activation function. This paper considers that a BM has \( Q \) neurons and each neuron is connected to all other neurons with a connection matrix \( W \). We denote \( s_q \) as the state of neuron-\( q \) \((q = 1, \ldots, Q)\), where \( s_q \) is equal to either 1 or 0. Let \( s = [s_1, \ldots, s_Q]^T \) be the state vector and \( s_{-q} = [s_1, \ldots, s_{q-1}, s_{q+1}, \ldots, s_Q]^T \) be the state vector where \( s_q \) is not included. The state of neuron-\( q \) is not based on a deterministic way. For neuron-\( q \), given the states of other neurons, the state of neuron-\( q \) is updated based on a stochastic way, given by

\[
\text{Prob}_{BM}(s_q = 1|s_{-q}) = \frac{1}{1 + e^{-u_q/T_o}},
\]

where the input, \( u_q \), is given by

\[
u_q = \sum_{l=1}^Q w_{ql}s_l + b_q.
\]

We can call \( 1 \) as the stochastic activation function. In \( 2 \), \( w_{ql} \) is the weight between the neuron-\( q \) and neuron-\( l \), and \( b_q \) is the bias for neuron-\( q \). The connection matrix \( W \) is symmetric and its main diagonal elements are equal to zero, i.e., \( w_{ql} = w_{ql} (q \neq l) \) and \( w_{qq} = 0 \). The collection of all bias terms are denoted as \( b = [b_1, \ldots, b_Q]^T \). Also, \( T_o \) is called temperature factor.

After the updating process is performed sufficient number of times, from the BM theory [1]–[3] the state distribution \( \text{Prob}_{BM}(s) \) reaches equilibrium. At the equilibrium, the state distribution is given by

\[
\text{Prob}_{BM}(s) = \frac{\exp(-E(s)/T_o)}{\sum_{s'} \exp(-E(s')/T_o)},
\]

where \( E(s) \) is the energy of state \( s \), given by

\[
E(s) = -\frac{1}{2} s^T W s - s^T b.
\]

Here we do not mean that the state of BM converges. The BM theory tells us that the state distribution \( \text{Prob}_{BM}(s) \) reaches equilibrium, at which the probability values are given by \( 3 \).

![Figure 1. The general structure of BM for classification task.](image)

As shown in Fig. 1, the neurons are divided into two groups, visible and hidden. The visible neurons receive the information from the environment or deliver the information to the environment. Let \( v \) be the collection of the states of visible neurons and let \( h \) be the collection of the states of hidden neurons. At the equilibrium, \( 3 \) can be rewritten as

\[
\text{Prob}_{BM}(s) = \frac{\text{Prob}_{BM}(v, h)}{\sum_{v', h'} \exp(-E(v', h')/T_o)}.
\]

In addition, at the equilibrium, the state distribution of the visible neurons are then given by

\[
\text{Prob}_{BM}(v) = \frac{\sum_{h'} \exp(-E(v, h')/T_o)}{\sum_{v', h'} \exp(-E(v', h')/T_o)}.
\]
For the classification task, the visible neurons are divided into input visible neurons and output visible neurons, as shown in Fig. 1. The states of the input visible neurons are denoted as \( v_{in} \), while the states of the output visible neurons are denoted as \( v_{out} \). That means, \( v^T = [v_{in}^T, v_{out}^T] \).

Usually, for the classification task, the states of the input visible neurons are clamped. In this case, (6) becomes

\[
\text{Prob}_{BM}(v_{out}|v_{in}) = \frac{\sum_h \exp(-E(v_{out}, v_{in}, h)/T_0)}{\sum_{v', h'} \exp(-E(v_{out}', v_{in}', h')/T_0)}. \tag{7}
\]

During training, each learning iteration consists of two phases: “clamped phase” and “free-run phase” [1]–[3]. In the clamped phase, the visible neurons are clamped at the training inputs. BM learning algorithms aim at minimizing the KL divergence between the clamped phase state distribution and the free-run phase state distribution.

### B. RESTRICTED BOLTZMANN MACHINE FOR FEATURE EXTRACTION

The RBM model can be considered as a special case of the BM model. It consists of one visible layer and one hidden layer. Each neuron in a layer is connected with all the neurons in another layer, but there is no connection between the neurons in the same layer. Consider that there are \( m \) neurons in the visible layer, whose are denoted as \( v = [v_1, v_2, \ldots, v_m]^T \). Besides, there are \( n \) neurons in the hidden layer, which are denoted as \( h = [h_1, h_2, \ldots, h_n]^T \).

The state of each neuron is binary, i.e., \([0, 1]\). The energy of the network in the state \((v, h)\), given by

\[
E(v, h) = - \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} v_i h_j - \sum_{i=1}^{m} c_i h_i - \sum_{j=1}^{n} d_j v_j, \tag{8}
\]

where \( c_i \)'s and \( d_j \)'s are bias terms of hidden neurons and visible neurons, respectively, and \( w_{ij} \)'s are the interconnection weights between hidden neurons and visible neurons. The state distribution at the equilibrium is given by

\[
\text{Prob}_{RBM}(v, h) = \frac{\exp(-E(v, h)/T_0)}{\sum_{v', h'} \exp(-E(v', h')/T_0)}. \tag{9}
\]

The stochastic activation function for neuron-\( i \) in the hidden layer is given by

\[
\text{Prob}_{RBM}(h_i = 1|v) = \frac{1}{1 + e^{-u_i/T_0}}, \tag{10}
\]

where

\[
u_i = \sum_{j=1}^{m} w_{ij} v_j + c_i. \tag{11}\]

The stochastic activation function for neuron-\( j \) in the visible layer is given by

\[
\text{Prob}_{RBM}(v_j = 1|h) = \frac{1}{1 + e^{-\phi_j/T_0}}, \tag{12}
\]

where

\[
\phi_j = \sum_{i=1}^{n} w_{ij} h_i + d_j. \tag{13}
\]

The RBM model has strong ability to extract useful features from input data for a wide range of classification problems [38]–[40]. There are many possible operations in the RBM. One of them is as follows. As shown in Fig. 2, we first present a data vector in the visible units. Afterwards, we update hidden neurons and visible neurons based on (10) and (12) a number of times. Fig. 3 shows a general structure of using the RBM for classification. In this structure, we present the input and then feed the output of the hidden nodes to a softmax classifier.

During training of an RBM, there are two phases too. In the clamped phase, the visible neurons are clamped in a training input vector. In the free-run phase, the training input vector is first presented in the visible neuron and then all neurons are allowed to be updated. The BM learning algorithms aims at minimizing the KL divergence between the clamped phase state distribution and the free-run phase state distribution.

### III. BM AND RBM WITH NOISE

In implementation, the presence of noise in the circuit may cause a trained BM or RBM works improperly. Especially, the noise can affect the operation of the stochastic activation functions.

For the BM model, the noise component at neuron-\( q \) is denoted as \( \Delta u_q \). In this case, the noisy input is given by

\[
\bar{u}_q = u_q + \Delta u_q. \tag{14}
\]

This paper assumes that \( \Delta u_q \) follows the Gaussian distribution with zero mean and variance equal to \( \sigma^2 \).
Given the noisy component $\Delta u_q$, the stochastic activation function of a noisy BM is a conditional probability (condition on $\Delta u_q$’s), given by

$$\text{Prob}_{\text{NBM}}(s_q = 1|s_{-q}, \Delta u_q) = \frac{1}{1 + e^{-\Delta u_q/T_o}}, \quad (15)$$

It should be noticed that we need to find out the unconditional probability (uncondition on noise) $\text{Prob}_{\text{NBM}}(s_q = 1|s_{-q})$ when the noise present.

For the RBM model, when there are noise in the inputs of the neurons, the effective inputs are modelled as

$$u_i = u_i + \Delta u_i, \quad \phi_j = \phi_j + \Delta \phi_j, \quad (16)$$

where $\Delta u_i$’s and $\Delta \phi_j$’s are the noise components in the hidden and visible neurons, respectively. In this paper, we assume that they follow the Gaussian distribution with zero mean and variance equal to $\sigma^2$.

Given the noise components $\Delta u_i$’s, the stochastic activation functions of a noisy RBM are given by

$$\text{Prob}_{\text{RB}}(h_i = 1|v, \Delta u_i) = \frac{1}{1 + e^{-u_i/T_o}}, \quad \text{Prob}_{\text{RB}}(v_j = 1|h, \Delta \phi_j) = \frac{1}{1 + e^{-\phi_j/T_o}}. \quad (17)$$

Again, we need to find out the unconditional probability $\hat{P}(h_i = 1|v)$ and $\hat{P}(v_j = 1|h)$.

### IV. MAIN RESULT

This section first shows how the noise affect the stochastic activation function and state distribution. Second, a new activation function is proposed to compensate the KL divergence loss caused by the noise. Our result is based on the Haley’s approximation for the Gaussian distribution [41], given by Lemma 1.

**Lemma 1:**

$$\frac{1}{1 + e^{-\rho t}} \approx \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \quad (18)$$

where $\rho = 1.702$.

### A. BOLTZMANN MACHINE

Theorem 1 describes the stochastic activation function of a BM under the effect of noise.

**Theorem 1:** For a noisy BM with an operational temperature $T_o$, the stochastic activation function is given by

$$\text{Prob}_{\text{BM}}(s_q = 1|s_{-q}, \Delta u_q) = \frac{1}{1 + e^{-\Delta u_q/(\eta T_o)}}, \quad (19)$$

where $\eta = \sqrt{1 + \frac{\sigma^2}{\rho^2 T_o^2}}$, and $\sigma^2$ is the variance of the noise.

**Proof:** From Lemma 1, we have

$$\text{Prob}_{\text{BM}}(s_q = 1|s_{-q}, \Delta u_q) = \frac{1}{1 + e^{-\Delta u_q/T_o}}.$$

$$\approx \int_{-\infty}^{\Delta u_q} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz\times \int_{-\infty}^{\Delta u_q} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

$$= \int_{-\infty}^{\Delta u_q} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \quad (20)$$

where

$$\alpha = \rho^2 T_o^2. \quad (21)$$

Since $\Delta u_q$’s follow the Gaussian distribution with zero mean and variance equal to $\sigma^2$, the stochastic activation function becomes,

$$\text{Prob}_{\text{BM}}(s_q = 1|s_{-q}) = \int_{-\infty}^{\infty} \text{Prob}_{\text{BM}}(s_q = 1|s_{-q}, \Delta u_q)$$

$$\times \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(\Delta u_q)^2/2\sigma^2} dx \Delta u_q, \quad (22)$$

Based on (20), we have

$$\text{Prob}_{\text{BM}}(s_q = 1|s_{-q}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi \alpha}} e^{-(\Delta u_q)^2/2\alpha^2} dx \Delta u_q$$

$$\times \frac{1}{\sqrt{2\pi \sigma^2}} e^{-(\Delta u_q)^2/2\sigma^2} dx$$

$$= \frac{1}{\sqrt{2}\pi \sigma^2} \frac{1}{\sqrt{2}\pi \alpha} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2(\alpha + \sigma^2)}} dx$$

$$\times e^{-\frac{(\Delta u_q)^2}{2(\alpha + \sigma^2)}} d\Delta u_q dx$$

$$= \frac{1}{\sqrt{2}\pi (\sigma^2 + \alpha)} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2(\sigma^2 + \alpha)}} dx. \quad (23)$$

From Lemma 1 and (21), we get

$$\text{Prob}_{\text{BM}}(s_q = 1|s_{-q}) \approx \frac{1}{1 + e^{\Delta u_q/(\eta T_o)}}, \quad (24)$$

where $\eta = \sqrt{1 + \frac{\sigma^2}{\rho^2 T_o^2}}$. The proof is completed. ■

Comparing (19) with (1), we find that when noise exist in the operation, the effect of the noise is identical to increasing the temperature from $T_o$ to $\eta T_o$, where $\eta = \sqrt{1 + \frac{\sigma^2}{\rho^2 T_o^2}}$ and $\eta T_o$ is the effective temperature for a noisy BM when noise present. With Theorem 1, the state distribution of a noisy BM is given by Theorem 2.

**Theorem 2:** With the effect of the noise, at the equilibrium, the state distribution of a noisy BM is modified as

$$\text{Prob}_{\text{BM}}(v_{out}|v_{in}) = \frac{\sum_{h'} \exp(-E(v_{out}, v_{in}, h')/(\eta T_o))}{\sum_{v'_{out}} \exp(-E(v'_{out}, v'_{in}, h')/(\eta T_o)).} \quad (25)$$

**Proof:** From Theorem 1, it is easy to obtain that the effect of noise is equivalent to raise the temperature factor.
from $T_o$ to $\eta T_o$. Thus, the corresponding state distribution $\text{Prob}_{\text{NBM}}(v_{\text{out}} \mid v_{\text{in}})$ is modified as

$$\sum_{v_{\text{out}}} \exp \left(-E(v_{\text{out}}, v_{\text{in}}, h')/(\eta T_o)\right).$$

The proof is completed.  

Based on Theorem 2, we can get the state distribution of $v_{\text{out}}$ for noisy BM. Besides, the KL divergence between $\text{Prob}_{\text{BM}}(v_{\text{out}} \mid v_{\text{in}})$ and $\text{Prob}_{\text{NBM}}(v_{\text{out}} \mid v_{\text{in}})$ is

$$D_{\text{BM}} = \sum v_{\text{out}} \text{Prob}(v_{\text{out}}' \mid v_{\text{in}}) \log \frac{\text{Prob}(v_{\text{out}}' \mid v_{\text{in}})}{\text{Prob}(v_{\text{out}} \mid v_{\text{in}})}.$$  

where $\text{Prob}_{\text{BM}}(v_{\text{out}}' \mid v_{\text{in}})$ is the state distribution of the visible neurons of a noise-free BM given by (7). Also, the KL divergence over all possible inputs, is

$$D_{\text{BM}} = \sum v_{\text{in}} \text{Prob}(v_{\text{in}}') D_{\text{BM}}.$$  

Since $\eta T_o > T_o$, we have $\text{Prob}_{\text{BM}}(v_{\text{out}} \mid v_{\text{in}}) \neq \text{Prob}_{\text{NBM}}(v_{\text{out}} \mid v_{\text{in}})$. Therefore, there exists a KL divergence loss for a noisy BM. The following theorem provides a method to compensate the loss.

**Theorem 3:** Given a noise-free BM with training temperature $T_o$, the behaviour of a noisy BM can be the same as that of the noise-free trained BM by setting the operational temperature $T'$ to

$$T' = \sqrt{1 - \frac{\sigma^2}{\rho^2 T_o^2}} T_o,$$  

where $T_o$ is the temperature during training. As a result, the compensated stochastic activation function is

$$\text{Prob}_{\text{NBM}}(s_q = 1 | s_{-q}) = \frac{1}{1 + e^{-u_q)/(\eta T')}} = \frac{1}{1 + e^{-u_q}/T_o}.$$  

**Proof:** Let $T'$ be the operational temperature of the noisy BM. From Theorem 1, the stochastic activation function is given by

$$\text{Prob}_{\text{NBM}}(s_q = 1 | s_{-q}) = \frac{1}{1 + e^{-u_q)/(\eta T')}},$$  

where $\eta = \sqrt{1 + \frac{\sigma^2}{\rho^2 T' T^2}}$. First we have

$$\eta T' = \sqrt{1 + \frac{\sigma^2}{\rho^2 T' T^2}} T'.$$  

When we set $T' = \sqrt{1 - \frac{\sigma^2}{\rho^2 T_o^2}} T_o$, we have $\eta T' = T_o$, and

$$\text{Prob}_{\text{NBM}}(s_q = 1 | s_{-q}) = \frac{1}{1 + e^{-u_q}/(\eta T')} = \frac{1}{1 + e^{-u_q}/T_o} = \text{Prob}_{\text{BM}}(s_q = 1 | s_{-q}).$$  

The proof is completed.  

Suppose a BM is trained under temperature $T_o$. Then the state distribution of a noise-free BM is

$$\text{Prob}_{\text{BM}}(v_{\text{out}} \mid v_{\text{in}}) = \sum v_{\text{out}} \exp(-E(v_{\text{out}}, v_{\text{in}}, h')/(\eta T_o)).$$  

Now consider that there are noise with variance equal to $\sigma^2$ during operation. Let $T'$ be the operational temperature of the noisy BM. From Theorem 2, the state distribution of the noisy BM is given by

$$\text{Prob}_{\text{NBM}}(v_{\text{out}} \mid v_{\text{in}}) = \sum v_{\text{out}} \exp(-E(v_{\text{out}}, v_{\text{in}}, h')/(\eta T')).$$  

From Theorem 3, when we set $T' = \sqrt{1 - \frac{\sigma^2}{\rho^2 T_o^2}} T_o$, $\text{Prob}_{\text{NBM}}(s_q = 1 | s_{-q}) = \text{Prob}_{\text{BM}}(s_q = 1 | s_{-q})$. Thus, the state distribution $\text{Prob}_{\text{NBM}}(v_{\text{out}} \mid v_{\text{in}})$ of a noisy BM becomes the same as the state distribution $\text{Prob}_{\text{BM}}(v_{\text{out}} \mid v_{\text{in}})$ of a standard noisy free BM. In addition, the KL divergence $D_{\text{BM}}$ between $\text{Prob}_{\text{NBM}}(v_{\text{out}} \mid v_{\text{in}})$ and $\text{Prob}_{\text{BM}}(v_{\text{out}} \mid v_{\text{in}})$, given by (26) and (27) becomes zero.

To sum up, when $T_o$ is the temperature during training, we should use $T' = \sqrt{1 - \frac{\sigma^2}{\rho^2 T_o^2}} T_o$ during operation. Nevertheless, from (28), the compensation method has its own limitation, given by

$$1 - \frac{\sigma^2}{\rho^2 T_o^2} > 0.$$  

**B. RESTRICTED BOLTZMANN MACHINE**

The results of BM can be extended to the noisy RBM. Theorem 4 presents the noisy activation function.

**Theorem 4:** For a noisy RBM with an operational temperature $T_o$, the stochastic activation functions are given by

$$\text{Prob}_{\text{NRBM}}(h_j = 1 | v) = \frac{1}{1 + e^{-u_j/(\eta T_o)}},$$  

$$\text{Prob}_{\text{NRBM}}(v_j = 1 | h) = \frac{1}{1 + e^{-\phi_j/(\eta T_o)}},$$  

where $\eta = \sqrt{1 + \frac{\sigma^2}{\rho^2 T_o^2}}$.

**Proof:** The logic of the proof is similar to that of Theorem 1.  

For the noise-free case, the state distribution of the hidden nodes is given by

$$\text{Prob}_{\text{RB}}(h) = \sum h' \exp(-E(h', h')/(\eta T_o)),$$

From Theorem 4, the effect of the noise is to raise the temperature from $T_o$ to $\eta T_o$, where $\eta = \sqrt{1 + \frac{\sigma^2}{\rho^2 T_o^2}}$. The following theorem shows the state distribution of the hidden nodes in a noisy RBM.

**Theorem 5:** For a RBM with noise, during the operation, at the equilibrium, the state distribution of the hidden nodes
is given by

$$\text{Prob}_{\text{NRBM}}(h) = \sum_{\nu} \exp(-E(\nu', h)/(\eta T_o)) / \sum_{\nu'} \exp(-E(\nu', h')/(\eta T_o)).$$  \hspace{1cm} (36)

**Proof:** The proof is similar to that of Theorem 2. The proof is completed. \(\blacksquare\)

For a trained noise-free a RBM with temperature \(T_o\), when noise exist in the operation, the effective temperature becomes \(\eta T_o\), where \(\eta = \sqrt{1 + \frac{\sigma^2}{\rho T_o}}\). There are KL divergence loss in a noisy RBM, given by

$$\mathcal{D}_{\text{RBM}} = \sum_h \text{Prob}_{\text{RBM}}(h) \log \frac{\text{Prob}_{\text{RBM}}(h)}{\text{Prob}_{\text{NRBM}}(h)}.$$ \hspace{1cm} (37)

Theorem 6 shows that the behaviour of a noisy RBM is the same as that of a noise-free RBM.

**Theorem 6:** Given a noise-free RBM with training temperature \(T_o\), the behaviour of a noisy RBM can be the same as that of the noise-free trained RBM by setting the operational temperature \(T'\) to

$$T' = \sqrt{1 - \frac{\sigma^2}{\rho^2 T_o^2}} T_o.$$ \hspace{1cm} (38)

As a result, the compensated stochastic activation functions is

$$\text{Prob}_{\text{NRBM}}(h_i = 1|\nu) = \frac{1}{1 + e^{-u_i/(\eta T')}} = \frac{1}{1 + e^{-u_i/T_o}}$$

$$\text{Prob}_{\text{NRBM}}(h_j = 1|h) = \frac{1}{1 + e^{-\phi_j/(\eta T')}} = \frac{1}{1 + e^{-\phi_j/T_o}}.$$ \hspace{1cm} (39)

**Proof:** The logic of the proof is similar to that of Theorem 3. The proof is completed. \(\blacksquare\)

Suppose \(T_o\) is the temperature during training for a RBM. When noise exist during operation, we should use \(T' = \sqrt{1 - \frac{\sigma^2}{\rho^2 T_o^2}} T_o\) during operation. With this setting, the behaviour of a noisy RBM is the same that of a noise-free RBM. Hence there is no KL divergence loss in the noisy RBM. Again, the compensation method has its own limitation, given by \(1 - \frac{\sigma^2}{\rho^2 T_o^2} > 0\).

**V. SIMULATION**

In the aforementioned sections, we have investigated the effect of noise on the BM and RBM models. In addition, from our findings, we derive a compensation scheme. This section would like to verify our theoretical results. We first train BM and RBM to perform the tasks. Afterwards, we measure the performance loss under the noisy situation. We focus on the performance difference between the standard BM/RBM and noise compensated RM/RBM. Without loss of generality, we set \(T_o = 1\) in our simulations.
standard BM, while another one is our noise compensated BM. For our noise compensated BM, we adjust $T'$ based on Theorem 3. As there is a limitation in Theorem 3, i.e., $\sigma < 1.702$ (for $T_o = 1$). Hence, when the noise level exceeds the limitation, we fix $T'$ at 0.01.

We consider three kinds of noise. The first one is Gaussian noise which matches our theoretical assumption. The other two kinds of noise are zero mean uniform noise and Beta distribution noise. For the Beta distribution noise, we shift the mean such that the noise are zero-mean. The purpose of using uniform and Beta distribution noise to test whether our noise compensated BM can be used to suppress the effect of noise for non-Gaussian cases.

Since the BM model is a kind of stochastic model, we repeat the experiments 1,000 times for each example during testing. We then measure two KL divergence values. One is the KL divergence of the output neuron distributions between the noise-free BM and the noisy standard BM. Another one is the KL divergence of the output neuron’s state distributions between the noise-free BM and the noisy BM with our compensation scheme. In addition, we also measure the classification rates under noisy situations. The results are summarized in Fig. 5.

1) GAUSSIAN NOISE CASE
Our compensation scheme is based on the assumption that the noise follow the Gaussian distribution. The first row of Fig.5 shows the results for the Gaussian noise case. It can be seen that for the standard BM, when noise exist, there is performance degradation. The degradation increases as the noise level increases. When our noise compensated BM is used, up to the noise level of 1.7 to 2, the performance degradation is still small.

| Table 1. Coding for inputs and outputs for the Car Evaluation dataset. |
|-----------------|-----------------|-----------------|-----------------|
| Output          | Quality         | Value and Coding |
|                 | unacc(0.0), acc(0.1), good (1.0), vgood(1) |
| Input           | buying          | low(0.01), med(0.01), high(0.11), vhigh(1.11) |
|                 | marital         | low(0.001), med(0.01), high(0.11), vhigh(1.11) |
|                 | doors           | two(0.001), three(0.01), four(0.11), 5 or more(1.11) |
|                 | persons         | two(0.01), four(0.11), more(1.11) |
|                 | lug_boot        | small(0.01), med(0.01), big (1.11) |
|                 | safety          | low(0.01), med(0.01), high(1.11) |

FIGURE 5. The KL divergence loss and classification rate for standard BM and our noise compensated BM. (a)-(b) Gaussian noise. (c)-(d) zero mean Beta distribution noise. (e)-(f) uniform noise.
From the first row of the figure, when the noise level is equal to 2, the KL divergence loss of the standard BM is $1.713 \times 10^{-4}$. With our noise compensated BM, the KL divergence is $2.496 \times 10^{-6}$ only. This behaviour agrees with the expectation from Theorem 3. In addition, when the noise level is greater than 1.7, our noise compensated BM still has certain ability to suppress the effect of the noise. For example, when the noise level is 3, the KL divergence loss of the standard BM is $5.411 \times 10^{-4}$, but the KL divergence loss of our noise compensated BM is $2.617 \times 10^{-3}$ only.

The classification rate results are shown in the right hand side of the first row in Fig. 5. It can be seen that with our noise compensated BM, up to the noise level of 1.7, the classification rate does not degrade much. For example, when the noise level is equal to 1.8, the classification rate of the standard BM is 0.7094 only. On the other hand, the classification rate of our noise compensated BM can still maintain at 0.7553.

In addition, when the noise level is greater than 1.8, our noise compensated BM still has certain ability to suppress the effect of the noise. For example, when the noise level is 3, the classification rate of the standard BM is 0.6604 only. On the other hand, the classification rate of our noise compensated BM is 0.6828.

2) NON-GAUSSIAN NOISE CASE

Although our compensation scheme is based on the assumption that the noise follow the Gaussian distribution, our compensation scheme still works for non-Gaussian noise.

The second and third rows of Fig. 5 show the results for non-Gaussian noise cases. It can be seen that when our noise compensated BM is used, the performance degradation is very small up to the noise level of 1.8.

For the zero mean Beta distribution noise, when the noise level is 2, the KL divergence loss of the standard BM is $11.2962$.
1.939 × 10^{-4}. On the other hand, the KL divergence of our noise compensated BM is 2.123 × 10^{-5} only.

In addition, when the noise level is greater than 2, our noise compensated BM still has certain ability to suppress the effect of the noise. For example, when the noise level is 3, the KL divergence loss of the standard BM is 6.121 × 10^{-4}. On the other hand, the KL divergence loss of our noise compensated BM is 4.048 × 10^{-4} only.

The classification rate results are shown in the right hand side of the second row in Fig. 5. It can be seen that with our noise compensated BM is used, up to the noise level of 1.8, the classification rate does not decrease much. For example, when the noise level is equal to 1.8, the classification rate of the standard BM is 0.7080. On the other hand, the classification rate of our noise compensated BM can still maintain at 0.7858. In addition, when the noise level is greater than 1.8, our noise compensated BM still has certain ability to suppress the effect of the noise. For example, when the noise level is 3, the classification rate of the standard BM is 0.6543. On the other hand, the classification rate of our noise compensated BM is 0.6733.

For the uniform noise case, the results are similar to those of the other noise cases. To sum up, the results show that our noise compensated BM not only works for the Gaussian noise case, but also works for the non-Gaussian noise cases.

C. RBM FOR HANDWRITTEN DIGIT CLASSIFICATION

For the RBM, it is applied for handwritten digit 0-9 classification [44]. The training and testing set includes 60,000 and 10,000 images, respectively. Both of them are derived from MNIST database [44]. All the digit images are binarized first. Fig. 3 depicts the network configuration of the RBM for feature extraction. There are 784 neurons in the visible layer and 522 neurons in the hidden layer. The RBM is used for extracting the features from the dataset and then the extracted features are fed to a one-layer softmax classifier.

First, an RBM is trained to perform the extract the features from the data set. Afterwards, we train a one-layer softmax classifier to perform the classification task. In the testing stage, zero mean noise with various noise levels are added into the stochastic activation functions. We consider two RBM models in the testing stage. One is the standard RBM, while another one is our noise compensated RBM. For our noise compensated RBM, we adjust $T'$ based on Theorem 6. As there is a limitation in Theorem 6, i.e., $\sigma < 1.702$ (for $T_\sigma = 1$). Hence, when the noise level exceeds the limitation, we fix $T'$ at 0.01. Again, we consider three kinds of noise: Gaussian noise, zero mean uniform noise, and Beta distribution noise.

Since the RBM model is a kind of stochastic models, we repeat the experiments 50 times for each example during testing. We then measure two KL divergence values. Since the number of hidden nodes is very large, it is impossible to measure the hidden node state distribution. The KL divergence measurement used is based on the probability mass functions of the hidden neuron outputs between the noise-free RBM and the noisy standard RBM. Another one is based on the probability mass functions of the hidden neuron outputs between the noise-free RBM and the noisy RBM with our compensation scheme. The results are summarized in Fig. 6.

1) GAUSSIAN NOISE CASE

The first row of Fig. 6 shows the results for the Gaussian noise case. It can be seen that when noise exist, there is performance degradation in the standard RBM. When our noise compensated RBM is used, up to the noise level of 1.7, the KL divergence loss does not increase much. For example, when the noise level is equal to 1.5, the KL divergence loss of the standard RBM is 2.667 × 10^{-3}. On the other hand, the KL divergence of our noise compensated RBM is 2.233 × 10^{-3} only. In addition, when the noise level is greater than 1.7, our compensated RBM still has certain ability to suppress the effect of the noise. For example, when the noise level is 5, the KL divergence loss of the standard RBM is 1.229 × 10^{-2}. On the other hand, the KL divergence loss of our noise compensated RBM is 1.055 × 10^{-2} only.

The classification rate results are shown in the right hand side of the first row in Fig. 6. It can be seen that with our noise compensated RBM, up to the noise level of 1.7, the classification rate does not decrease much. For example, when the noise level is equal to 1.5, the classification rate of our noise compensated RBM can still maintain at 0.9101. Suppose we use around 0.90 as a target rate, the standard RBM can afford the noise level up to 1.5 only, while our compensated RBM affords the noise level up to 2.5.

2) NON-GAUSSIAN NOISE CASE

Although our compensated scheme is based on the assumption that the noise follow the Gaussian distribution, our compensated scheme still works for non-Gaussian noise. The second and third rows of Fig. 6 show the results for non-Gaussian noise cases.

For the zero mean Beta noise case, shown in the second row of Fig. 6, when the noise level is 1.5, the KL divergence loss of the standard RBM is 2.728 × 10^{-3}. On the other hand, the KL divergence of our noise compensated RBM is 2.310 × 10^{-3} only. In addition, when the noise level is greater than 1.7, our noise compensated RBM still has certain ability to suppress the effect of the noise. The classification rate results are shown in the right hand side of the second row in Fig. 6. It can be seen that with our noise compensated RBM is used, up to the noise level around 1.7, the classification rate does not decrease much. Suppose we use around 0.90 as a target rate, the standard RBM affords the noise level up to 1.5, while our compensated RBM affords the noise level up to 2.5.

For the uniform noise case, shown in the third row of Fig. 6, when the noise level is 1.5, the KL divergence loss of the standard RBM is 2.828 × 10^{-3}. On the other hand, the KL
dissimilarity of our noise compensated RBM is $2.318 \times 10^{-3}$ only. In addition, when the noise level is greater than 1.7, it can be seen that our noise compensated RBM still has certain ability to suppress the effect of the noise. The classification rate results are shown in the right hand side of the last row in Fig. 6. It can be seen that with our noise compensated RBM is used, up to the noise level of 1.7, the classification rate does not decrease much. Suppose we use around 0.90 as a target rate, the standard RBM affords the noise level up to 1.5, while our compensated RBM affords the noise level up to 2.5.

To sum up, the results show that our noise compensated RBM not only works for the Gaussian noise case, but also works for the non-Gaussian noise cases.

VI. CONCLUSION
This paper analyzed the behaviors of the BM and RBM models under the noise situation. Our results show that the effect of input noise is equivalent to raise the temperature factor in the stochastic activation function (Theorems 1 and 4) and state distribution (Theorems 2 and 5). To compensate effect of noise, we proposed the revised stochastic activation function (Theorem 3 and Theorem 6). In addition, we discuss the limitation of our compensation scheme. Experiments show that the effect of input noise can be suppressed by applying Theorems 3 and 6. In addition, the experimental results show that when the noise level is greater than the theoretical limit mentioned, our scheme still has certain ability to suppress the effect of noise. Since there is a theoretical limit for our compensation scheme, one of the future works is designed a better protection scheme for high noise levels.

REFERENCES


WENHAO LU is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, City University of Hong Kong. His current research interests include neural networks and machine learning.

CHI-SING LEUNG (Senior Member, IEEE) received the Ph.D. degree in computer science from The Chinese University of Hong Kong, Hong Kong, in 1995. He is currently a Professor with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. He has authored over 120 journal articles in the areas of digital signal processing, neural networks, and computer graphics. His current research interests include neural computing and computer graphics.

Prof. Leung was a member of the Organizing Committee of ICONIP2006. He was a Governing Board Member of the Asian Pacific Neural Network Assembly (APNNA) and the Vice President of APNNA. He received the 2005 IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award for his article titled The Plenoptic Illumination Function, in 2005. He was the Program Chair of ICONIP2009 and ICONIP2012. He is a Guest Editor of several journals, including Neural Computing and Applications, Neurocomputing, and Neural Processing Letters. He is currently an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

JOHN SUM (Senior Member, IEEE) received the B.Eng. degree in electronic engineering from The Hong Kong Polytechnic University, in 1992, and the M.Phil. and Ph.D. degrees in computer science and engineering from The Chinese University of Hong Kong, in 1995 and 1998, respectively. From 1998 to 2004, he taught in several universities in Hong Kong, including Hong Kong Baptist University, the Open University of Hong Kong, and The Hong Kong Polytechnic University. In 2005, he joined the Department of Information Management, Chung Shan Medical University, Taiwan, as an Assistant Professor. Later, he joined the National Chung Hsing University as an Assistant Professor, in 2007, an Associate Professor, in 2009, and a Professor, in 2015. His research interests include neural networks and scheduling.

* * *