J. Korean Math. Soc. **59** (2022), No. 2, pp. 217–233
https://doi.org/10.4134/JKMS.j200406
pISSN: 0304-9914 / eISSN: 2234-3008

MARGIN-BASED GENERALIZATION FOR CLASSIFICATIONS WITH INPUT NOISE

Hi Jun Choe, Hayeong Koh, and Jimin Lee

Abstract. Although machine learning shows state-of-the-art performance in a variety of fields, it is short a theoretical understanding of how machine learning works. Recently, theoretical approaches are actively being studied, and there are results for one of them, margin and its distribution. In this paper, especially we focused on the role of margin in the perturbations of inputs and parameters. We show a generalization bound for two cases, a linear model for binary classification and neural networks for multi-classification, when the inputs have normal distributed random noises. The additional generalization term caused by random noises is related to margin and exponentially inversely proportional to the noise level for binary classification. And in neural networks, the additional generalization term depends on (input dimension) × (norms of input and weights). For these results, we used the PAC-Bayesian framework. This paper is considering random noises and margin together, and it will be helpful to a better understanding of model sensitivity and the construction of robust generalization.

1. Introduction

Machine learning models, particularly deep neural networks, have achieved state-of-the-art performance on various tasks. Machine learning is a task that deals with big data and is exposed to noises or perturbations. Nowadays, it is impossible to use clean data without noise as input when training real-data because it trains a huge amount of data. Therefore, it is meaningful to find a generalization bound considering input noise.

Beside, Szegedy et al. [13] observed a phenomenon that caused a network to output incorrect classifications due to the application of a certain hardly perceptible perturbation, which was found by maximizing the networks prediction

Received July 14, 2020; Revised October 18, 2021; Accepted November 2, 2021.

2020 *Mathematics Subject Classification.* Primary 68Q32, 65Y20.

Key words and phrases. Generalization bound, PAC-Bayesian, margin loss function.

Hi Jun Choe and Hayeong Koh were supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Korea government (No. 2015R1A5A1009350, No. 20181A2A3074566). Jimin Lee was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Korea government (No. 2020R1I1A1A01071731).

error. This means that machine learning models can make incorrect predictions with high confidence by adding adversarial perturbations to inputs. In recent years, the generalization of a robust learning has been investigated to obtain a better understanding of this phenomenon.

It would be great if we could find a generalization bound that can cover all kinds of noises or perturbations at once. But first, the research should be conducted to find a robust generalization bound for each case, such as whether input noise is random or intentional or not. In this paper, we present a generalization bound of classifiers with input noise using margin loss function and related terms to perturbation of input and parameters. For this, we assume that the input noises are random values following Gaussian distribution and apply the PAC-Bayesian framework. The PAC means 'Probably Approximately Correct' [4]. PAC is one of the methods to measure the performance of a model in theory, and although it is difficult to use it in practically measuring the performance of a model, it is conceptually possible to explain why a particular model is better and when the performance of a model is good and when it is bad. Thus, it provides the relationship between the complexity of the learning algorithm and the ability it can achieve. And the ability is explained by the generalization error, the gap between theoretical error and empirical error, and tells a guarantee for learning.

The PAC-Bayes bounds approach proposed by Mcallester [7] contained bounds of error rates in classification. More recent studies combined PAC-Bayesian analysis and margins by Lanford et al. [6] and Mcallester [8]. In addition, under the assumption that changes in network output due to weight perturbations are bounded, Neyshabur et al. [10] applied the PAC-Bayes analysis for feedforward neural networks.

In neural networks, the margin is associated with various tasks. Bartlett et al. [2] suggested that the generalization gap, i.e., the difference between training and test errors, is strongly related to the normalized product of the spectral norms of wight matrices. In addition, they found that normalized margin distribution converge, which led to investigations into margin distribution and margins for hidden layers. Two studies of Jiang et al. [5] and Elsayed et al. [3] focused on reducing the generalization gap. These studies proposed using margin distribution at hidden layers and a method to approximate margins. The papers proposed the coefficient of determination to predict the generalization gap and a flexible loss function that can establish a large margin, respectively. Shen-Huan et al. [12] proposed a new loss function that utilized the margin distribution and found the generalization bound based on *noise stability* proposed by Arora et al. [1] and the PAC-Bayesian framework proposed by Neyshabur et al. [10].

We focus on the bounds that are expressed in the margin. If the output change due to the perturbations of weight parameters can be expressed using the margin loss function, a case could be made for input noise perturbations. To know the sensitivity of classifier to input noise, we want to see how robust classification is to input noise through generalization bounds. The generalization bounds are expressed assuming the worst case, this study is about how much network sensitivity and margins can affect as much as possible. Therefore, it shows a generalization bound according to the fixed number of samples under specific conditions of input noise and margin. In this paper, we find out what the additional generalization term due to input noise has to do with margin.

And, in neural network, we observe that even though the adversarial input perturbation is non-randomly added, and different from our random noise, the additional generalization bounds caused by input noises are similar in that they rely on the product of the input dimension and norms of input and weights as in the adversarial example. It appears that whether or not the perturbations are random is not consequential, because the worst case is considered when we compute the generalization bounds. To find a better robust generalization, other measures are required to handle weight sensitivity or the distribution of input data and these were briefly discussed in the last part of this paper.

1.1. Preliminaries

If hypothesis determines how output is determined for each input, one purpose of machine learning is to approximate the distribution of hypothesis as closely as possible. Let's denote the trained distribution is $\mathcal{S}$ by learning and actual distribution we are looking for is $\mathcal{D}$. The generalization bound tells the maximal difference between these two distributions.

The generalization bounds was first expressed using Kullback-Leibler divergence in the PAC-Bayesian framework by Mcallester [8]. It tells a capacity of variations of parameters stochastically. This is necessary to calculate perturbations of weight parameters and the method of finding KL divergence depends on the type of classification. We use the method by Mcallester for linear classification, and Neyshabur et al. [10] method for multi-classification. The definition of KL divergence is as a follows.

Kullback-Leibler divergence. For two distributions P and Q, the kullback-Leibler divergence is defined as follows:

$$KL(Q\|P) = \sum_x q(x) \log \frac{q(x)}{p(x)},$$

where $p(x)$ and $q(x)$ are the probability density functions of P and Q, respectively. The KL divergence is a measure of the distance between distributions P as a prior and Q as a posterior.

Mcallester [9] and Langford and Shawe-Taylor [6] attempted to use a PAC-Bayesian approach for a SVM or linear binary classifications. In those studies, for networks with weight vectors, Kullback-Leibler divergence is computed using fat shattering arguments. In addition, those studies assume weight vectors are unit-variance isotropic multivariate Gaussian. But recently, the PAC-Bayesian approach is expanded to neural networks with ReLU activation functions. Neyshabur et al. [10] assumed the perturbations of weights are Gaussian random values, and, using this assumption, KL divergence is computed. However, the type of perturbation does not matter if it satisfies the bound on the changes in the output of the networks.

For convenience, we sometimes denote the classifier to $f_{\mathbf{w}}$ where $\mathbf{w}$ is the parameter of classifier. And in some case, the distribution followed by the parameter is written in subscripts. For example, f_Q means a classifier with parameter $\mathbf{w}$ distributed Q, $\mathcal{F}_Q$ means a family of these classifiers. And we indicate that L is the loss function of the classifier and $\hat{L}$ is the experimental version.

The PAC learning is a framework for mathematical analysis of machine learning, in which the learner receives samples and selects the hypothesis from a specific class of possible functions [4]. The introduction of the concept of computational complexity theory to machine learning is an impressive part of the PAC framework. In particular, learners are expected to find efficient functions, and learners themselves must implement efficient procedures. Therefore, the goal of the PAC framework is for a function chosen with high probability (the "probably" part) to have a low generalization error (the "approximately correct" part). The PAC-Bayesian framework was based on the next theorem.

Theorem 1.1 (PAC-Bayesian Theorem)**.** *For any probability distribution* (*measure*) *on a possibly uncountable set $\mathcal{C}$ and any measurable loss function L we have the following where Q ranges over all distributions* (*measures*) *on $\mathcal{C}$. Then, with probability at least $1-\delta$ over the choice of sample set we have*

$$KL(\hat{L}(\mathcal{F}_Q)\|L(\mathcal{F}_Q)) \leq \frac{KL(Q\|P)+\ln\frac{m}{\delta}}{m-1}, \tag{1}$$

where m is the sample size and P is a fixed prior distribution.

In the above theorem, P and Q are the unknown target distribution and the trained distribution, respectively. We can use the Kullback-Leibler divergence to get the distance between the error of the empirical distribution $\hat{L}(\mathcal{F}_Q)$ and the error of the theoretical distribution $L(\mathcal{F}_Q)$. Since, we want the loss of the target distribution $L(\mathcal{F}_Q)$ to be bound using ϵ, which is expressed as follows: With the probability at least $1-\delta$ over the choice of the sample we have that the following holds simultaneously for all $\mathbf{w} \in \mathcal{R}^n$ with $\|w\| = 1$ and $\gamma \geq 0$:

$$L_0(\mathcal{F}_Q) \leq \sup\left\{\epsilon : KL(\hat{L}_\gamma(\mathcal{F}_Q)\|\epsilon) \leq \frac{KL(Q\|P)+\ln\frac{m}{\delta}}{m-1}\right\}.$$

We have the inequality $KL(q\|p) \geq (q-p)^2/(2q)$, and using statement that if $KL(q\|p) \leq x$, then $p \leq q + \sqrt{2qx} + 2x$, the inequality expressed more clearly.

$$L_0(f_{\mathbf{w}}) \leq \hat{L}(f_{\mathbf{w}}) + \sqrt{\frac{2\hat{L}(f_{\mathbf{w}})(KL(\mathbf{w}||P) + \ln \frac{m}{\delta})}{m-1}} + \frac{2(KL(\mathbf{w}||P) + \ln \frac{m}{\delta})}{m-1}. \tag{2}$$

Neyshabur et al. [10] expand the PAC-Bayesian framework to margin-based bounds for neural networks.

Lemma 1.2. *Let $f_{\mathbf{w}}$ be any predictor with parameter $\mathbf{w}$. For all $\gamma > 0$, with probability of at least $1-\delta$ over the training data of size m, we have the following margin bounds*:

$$L_0(f_{\mathbf{w}}) \leq \hat{L}_\gamma(f_{\mathbf{w}}) + 2\sqrt{\frac{2(KL(\mathbf{w}||P) + \ln \frac{2m}{\delta})}{m-1}}. \tag{3}$$

1.2. Outline of the paper

We find margin-based generalization bounds with input noises of the linear model for binary classification and the neural network for multi-classifications in Sections 2 and 3, respectively. Before finding the bounds in each section, we set about the model parameters and input noises. Especially in neural network for multi classification, we compare the bounds with the previous results for adversarial example and describe the need for a new approach.

2. Binary linear classification

In this section, we consider the binary linear classification. Our goal is to express a generalization bound for classification with input noises using the PAC-Bayesian framework. Therefore, the result of this section is the boundary of the $0-1$ loss function of the binary linear classification using the margin loss function and additional terms. This generalization bound provides a case where input noise, a Gaussian random value, is added. Thus we can obtain information about the boundary conditions that are robust to input noise from this result.

We start with a model description.

Assume that the input domain is $\mathcal{X}_{1,n} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$. We consider a linear classifier with parameter $\mathbf{w}$ has unit norm, i.e., $\|\mathbf{w}\|_2 = 1$. Linear classification uses a linear product of input and parameter $\mathbf{w} \cdot \mathbf{x}$, so the output y is naturally in $[-1,1] \in \mathbb{R}$ but the target is $\mathcal{Y} = \{-1,1\}$. We assume that $\mathbf{w} \cdot \mathbf{x} \neq 0$ for all $\mathbf{x}$, and then we predict 1 or -1 depending on whether the result of the dot product of input $\mathbf{x}$ and parameter $\mathbf{w}$ is positive or negative, respectively. Thus for given sample set $\mathcal{S} = (\mathcal{X}, \mathcal{Y})$, if $\mathbf{w} \cdot \mathbf{x}$ and y has same sign, the sample data is classified correctly.

We need information about how accurate the classification is for the sample data, and the loss function tells us. The $0-1$ loss ℓ_0 is defined

$$\ell_0(\mathbf{w},(\mathbf{x},y)) = \begin{cases} 0 & \text{if classification is } \textit{correct}, \\ 1 & \text{if classification is } \textit{incorrect.} \end{cases}$$

The empirical $0-1$ loss function $\hat{L}_0$ is defined as

$$\hat{L}_0(\mathbf{w},\mathcal{S}) = \frac{1}{m}\sum_{i=1}^{m}\ell_0(\mathbf{w},(\mathbf{x}_i,y_i))$$

and $0-1$ loss function L_0 is $L_0(\mathbf{w},\mathcal{S}) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{w}\cdot\mathbf{x}) \leq 0]$.

Since $\mathbf{w}\cdot\mathbf{x} \neq 0$ for all $\mathbf{x}$, there always is a gap between the value $\mathbf{w}\cdot\mathbf{x}$ and the base line $y = 0$. Margin is the minimum value of these intervals for all inputs. Using this margin, margin loss function is defined.

Margin loss function. For a binary-classifier $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x}$, the margin loss function is defined as follows:

$$L_\gamma(\mathbf{w},\mathcal{S}) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}[y(\mathbf{w}\cdot\mathbf{x}) \leq \gamma].$$

The empirical margin loss function is denoted by $\hat{L}_\gamma$. If $\gamma = 0$, it corresponds to the classification loss L_0 and $\hat{L}_0$ represent the expected loss function and the training loss function, respectively.

In next section, we set input noise and then we start the process of finding a generalization bound for binary linear classification using PAC-framework.

2.1. Generalization bound with input noises

We want to express a generalization bound for classification with input noises and this generalization bound includes terms for margin. Initially, we started with the idea that margin can absorb some perturbations from inputs. We assume that the norm of input noise is bounded to the magnitude of margin. And we need to set the relation between the norm of the perturbation and the magnitude of the margin.

First, we assume that input noises are as follows.

The input noise. For noise vector $\mathbf{u}$, we assume that the input with noise is $\tilde{\mathbf{x}} = \frac{\mathbf{x}+\mathbf{u}}{\|\mathbf{x}+\mathbf{u}\|_2}$ and then $\tilde{\mathbf{x}}$ has norm 1. And here, we assume that each element of $\mathbf{u}$ is a random real number independent identically distributed $\mathcal{N}(0,\sigma^2)$ which is a Gaussian distribution with mean 0 and variance σ^2. And for convenience, we denote that the classifier added input noise $\mathbf{u}$ as $\tilde{f}_{\mathbf{w}} = \mathbf{w}\cdot\tilde{\mathbf{x}} = \mathbf{w}\cdot\frac{\mathbf{x}+\mathbf{u}}{\|\mathbf{x}+\mathbf{u}\|_2}$, and let $\tilde{\mathcal{F}}_{\mathbf{w}} = \{\tilde{f}_{\mathbf{w}}|\tilde{f}_{\mathbf{w}} = \mathbf{w}\cdot\tilde{\mathbf{x}}\}$.

To find the bound of the $0-1$ loss for perturbation case with margin loss term, we use fat shattering arguments in [6].

We assume that $\|\mathbf{w}\| = 1$ in this section. We assume that $\mathbf{w}$ has the distribution P whose elements have same value in all directions and the value of elements is a Gaussian random variable. For generalization, we have to get about random weight $\mathbf{w}\prime$ which is not exactly same with $\mathbf{w}$ but quite similar with $\mathbf{w}$. A measurement of the similarity can be considered to the inner product of the two vectors. So, we consider $\mathbf{w}\prime \cdot \mathbf{w} \geq \mu$ for some μ and $\mathbf{w}\prime$ has distribution Q. Then $KL(Q\|P) = \ln \frac{1}{\Phi(\mu)}$ where $\Phi\left(\mu\right)$ is the probability of the Gaussian random variable exceeding μ (see [9]).

Lemma 2.1. *Let an input noise vector* $\mathbf{u}$ *be* $\mathbf{u} \sim N(0, \sigma^2 I)$. *For* $\gamma \geq 0$, *if the magnitude of noise* $\|\mathbf{u}\|_2 < \gamma$ [1], *there exists a constant* c *such that* $\|\mathbf{u}\|_2 = (1-c)\gamma$ *where* $0 < c \leq 1$. *For* $\mathbf{w} \in \mathcal{R}^n$ *with* $\|w\| = 1$, *and for* $\mu \geq 0$, *we have the following.*

$$\hat{L}_0(\tilde{\mathcal{F}}_Q) \leq \hat{L}_\gamma(\mathcal{F}_P) + \Phi\left(c\mu\gamma\right),$$

where $\Phi(z)$ *is the probability that a unit-variance Gaussian random variable exceeds* z.

Proof. For $\mathbf{x} \in \mathcal{R}^n$ with $\|x\| = 1$, we let $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{u}$. Then, separate the input variable $\tilde{\mathbf{x}}$ to two components; $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_{\|} + \tilde{\mathbf{x}}_{\perp}$ where $\tilde{\mathbf{x}}_{\|}$ is the component of $\tilde{\mathbf{x}}$ parallel to $\mathbf{w}$ and $\tilde{\mathbf{x}}_{\perp}$ is $\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_{\|}$, i.e., the component of $\tilde{\mathbf{x}}$ perpendicular to $\mathbf{w}$. For a fixed γ-safe point, consider two orthogonal components $\tilde{\mathbf{x}}_{\|}$, $\tilde{\mathbf{x}}_{\perp}$ and random weight vectors $\mathbf{w}\prime$. If $\mathbf{w}\prime \cdot \mathbf{w} \geq \mu$, we have the following:

$$\begin{aligned}
\Pr_{\mathbf{w}\prime \sim Q}[\ y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}) \leq 0\ ] &= \Pr_{\mathbf{w}\prime \sim Q}[\ -y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}_{\perp}) \geq y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}_{\|})\ ] \\
&= \Pr_{\mathbf{w}\prime \sim Q}[\ -y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}_{\perp}) \geq y(\mathbf{w} \cdot \tilde{\mathbf{x}})(\mathbf{w}\prime \cdot \mathbf{w})\ ] \\
&= \Pr_{\mathbf{w}\prime \sim Q}[\ -y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}_{\perp}) \geq y((\mathbf{w} \cdot \mathbf{x}) + (\mathbf{w} \cdot \mathbf{u}))(\mathbf{w}\prime \cdot \mathbf{w})\ ] \\
&\leq \Pr_{\mathbf{w}\prime \sim Q}[\ -y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}_{\perp}) \geq (\gamma + (\mathbf{w} \cdot \mathbf{u}))\mu\ ] \\
&\leq \Pr_{\mathbf{w}\prime \sim Q}[\ -y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}_{\perp}) \geq (\gamma - \|\mathbf{u}\|_2)\mu\ ] \\
&= \Phi\left(\frac{(\gamma - \|\mathbf{u}\|_2)\mu}{\|\tilde{\mathbf{x}}_{\perp}\|_2}\right) \\
&\leq \Phi\left((\gamma - \|\mathbf{u}\|_2)\mu\right).
\end{aligned}$$

Since the magnitude of noise vector $\mathbf{u}$ is smaller than margin γ, there exists a constant $0 < c \leq 1$ such that $\|\mathbf{u}\|_2 = (1-c)\gamma$. Then, the following is satisfied.

$$\begin{aligned}
\hat{L}_0(\tilde{\mathcal{F}}_Q) &= \mathbb{E}_{(x,y)\sim\mathcal{S}}\left[\Pr_{\mathbf{w}\prime \sim Q}[y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}) \leq 0]\right] \\
&\leq \hat{L}_\gamma(\mathcal{F}_P) + \mathbb{E}_{(x,y)\sim\mathcal{S}}\left[\Pr_{\mathbf{w}\prime \sim Q}[y(\mathbf{w}\prime \cdot \tilde{\mathbf{x}}) \leq 0] \mid y(\mathbf{w} \cdot \mathbf{x}) > \gamma\right] \\
&\leq \hat{L}_\gamma(\mathcal{F}_P) + \Phi\left(\mu(\gamma - \|\mathbf{u}\|_2)\right)
\end{aligned}$$

[1] The $\|\mathbf{u}\|_2$ can be unbounded, but we only consider very small input perturbations. Because we started this study with that an adversarial attack with a very small perturbation can lead to different classification results (in this case, the l_∞ error is limited to ϵ). These are described in the last paragraph of Chapter 3.2.

$$\leq \hat{L}_\gamma(\mathcal{F}_P) + \Phi\left(c\mu\gamma\right). \qquad \square$$

In the above proof, the last inequality holds for $\gamma \leq 1$, but that condition is enough. Since norms of $\mathbf{w}$ and $\mathbf{x}$ are bounded by 1, thus their inner product is bounded by 1, too. So, we consider the case where the margin is less than 1 and this makes sense.

If $c = 1$, then this is the case when there is no noise. And always $\Phi\left(c\mu\gamma\right) \geq \Phi\left(\mu\gamma\right)$ is satisfied. The c is caused by input noise, thus the difference of $\Phi\left(c\mu\gamma\right)$ and $\Phi\left(\mu\gamma\right)$ is the extra term from input noise that we want to find.

Before the next step, we learn more about noise from the following corollary.

Corollary 2.2. *If we compare the case with and without noise, the difference is whether there is the term $y(\mathbf{w} \cdot \mathbf{u})$. The original result is determined by the sign of $y(\mathbf{w} \cdot \mathbf{x})$. Thus the input noise interferes with the result when $(\mathbf{w} \cdot \mathbf{u})$ has the opposite sign with $(\mathbf{w} \cdot \mathbf{x})$. Under the setting $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_{\|} + \tilde{\mathbf{x}}_{\perp}$ in the proof of above lemma, it can be said that only the component in the (opposite) direction parallel to $\mathbf{w}$ in noise affects the result.*

Now, using the theorem proposed by Langford et al. [6] and Mcallester [9], we can find the following margin-based generalization error bounds for the predictor added input noises: Let an input noise vector $\mathbf{u}$ be $\mathbf{u} \sim N(0, \sigma^2 I)$. With the probability at least $1 - \delta$ over the choice of the sample we have that the following holds simultaneously for all $\mathbf{w} \in \mathcal{R}^n$ with $\|w\| = 1$, $\mu \geq 0$ and $\gamma \geq 0$.

$$(4) \qquad L_0(\mathcal{F}_Q) \leq \sup\left\{\epsilon : KL(\hat{L}_\gamma(\mathcal{F}_Q) + \Phi\left(c\mu\gamma\right)\|\epsilon) \leq \frac{KL(Q\|P) + \ln\frac{m}{\delta}}{m-1}\right\}.$$

Theorem 2.3. *Let the binary classifier $\mathcal{F} = \{f_{\mathbf{w}} \,|\, f_{\mathbf{w}} = \mathbf{w} \cdot \mathbf{x}\}$. Then, with probability at least $1 - \delta$, we have*

$$L_0(\tilde{\mathcal{F}}) \leq \hat{L}_\gamma(\mathcal{F}) + \frac{m^{1-c}}{m\gamma^2} + \sqrt{2\left(\hat{L}_\gamma(\mathcal{F}) + \frac{m^{1-c}}{m\gamma^2}\right)\frac{\frac{\ln^+ m\gamma^2}{\gamma^2} + \frac{3}{2}\ln m + \ln\frac{1}{\delta} + 3}{m-1}}$$
$$+ 2\frac{\frac{\ln^+ m\gamma^2}{\gamma^2} + \frac{3}{2}\ln m + \ln\frac{1}{\delta} + 3}{m-1}.$$

Proof. In [9], μ is defined as a function of γ such that $\mu(\gamma) = \frac{\sqrt{2\ln m\gamma^2}}{\gamma}$, and thus $\Phi(\gamma\mu(\gamma)) \leq \frac{1}{m\gamma^2}$. Here, we have $\Phi(c\mu\gamma)$ which is needed to be bounded.

$$\begin{aligned}
\Phi\left(c\mu\gamma\right) &\leq \exp\{-c^2\gamma^2\mu(\gamma)^2/2\} \\
&= \exp\{-c^2\ln\left(m\gamma^2\right)\} \\
&= \frac{1}{(m\gamma^2)^{c^2}} \\
&= \frac{(m\gamma^2)^{1-c^2}}{m\gamma^2}
\end{aligned}$$

$$\leq \frac{(m\gamma^2)^{1-c}}{m\gamma^2}. \qquad \square$$

In the above theorem, the smaller the noise c the closer it is to 1, and then $(m\gamma^2)^{1-c}$ is close to 1, too. Compared with the case of without noises, $(m\gamma^2)^{1-c} > 1$ and it is a loose generalization bound.

The more simple inequality is in the following corollary.

Corollary 2.4. *The above theorem can be expressed simply*:

$$L_0(\tilde{\mathcal{F}}) \leq \alpha + \sqrt{2\alpha\beta} + 2\beta,$$

where

$$\alpha = \hat{L}_\gamma(\mathcal{F}) + \frac{(m\gamma^2)^{1-c}}{m\gamma^2},$$

$$\beta = \frac{\frac{\ln^+ m\gamma^2}{\gamma^2} + \frac{3}{2}\ln m + \ln\frac{1}{\delta} + 3}{m-1}.$$

As in the previous work, we can express the generalization boundary in theorem to more clearly as follows:

$$L_0(\tilde{\mathcal{F}}) \leq \hat{L}_\gamma(\mathcal{F}) + \frac{(m\gamma^2)^{1-c}}{m\gamma^2} + \mathcal{O}\left(\sqrt{\frac{\ln^+ m\gamma^2 + \ln\frac{\delta}{m}}{m\gamma^2}}\right). \tag{5}$$

Without noises, the margin-based generation bound is expressed $L_0(\mathcal{F}) \leq \hat{L}_\gamma(\mathcal{F}) + \mathcal{O}\left(\sqrt{\frac{KL(Q\|P)+\ln\frac{\delta}{m}}{m}}\right)$, and it is from $L_0(\mathcal{F}) \leq \hat{L}_\gamma(\mathcal{F}) + \mathcal{O}\left(\sqrt{\beta}\right)$. However, we use the term $\frac{(m\gamma^2)^{1-c}}{m\gamma^2}$ to indicate the effect of input noise. As a result, the generalization bounds increases as margin and the number of input samples containing noise increases exponentially inversely proportional to the noise level. This means that the more exposure to noise, the looser the boundary.

3. The neural networks for multi-class classification

In this section, we deal with the neural network for multi classification. As in the previous section, our goal is to express a generalization bound for classification with input noises using PAC-Bayesian method. The result of this section will be the boundary of the $0-1$ loss of the neural network for multi-classification using the margin loss function and extra terms. This generalization bound provides a case of input with Gaussian random noise. Thus we can obtain information about a robust boundary conditions to input noise from this result.

Assume that the input domain is $\mathcal{X}_{B,n} = \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}|_2 \leq B\}$. We consider a neural network that classifies input into k-classes. Therefore, its output domain will be $\mathbb{R}^k$. Let $f_{\mathbf{w}}(\mathbf{x}) : \mathcal{X}_B \to \mathbb{R}^k$ be the function for a d-layer feedforward neural network with parameters $\mathbf{w} = \{W_i\}_{i=1}^d$. More specifically, $f_{\mathbf{w}}(\mathbf{x}) =$

$W_d\phi(W_{d-1}\cdots\phi(W_1\mathbf{x}))$. Here, we assume the activation function ϕ is the ReLU function. The ReLU function is piecewise linear and has the Lipschitz property.

Multi-classification has a different output format than binary classification. In binary classification, the output was a scalar, but in multi-classification it is a k-dimensional vector. Thus margin is defined in a different way than binary classification. After each input through the weight parameters, the class of the input is determined by the index having the greatest value. In this process, we call the deference of the biggest and the second value is margin, i.e., the margin γ is $f_{\mathbf{w}}(\mathbf{x})[y]-\max_{i\neq y} f_{\mathbf{w}}(\mathbf{x})[i]$, where $f_{\mathbf{w}}(\mathbf{x})[i]$ is the i-th index of $f_{\mathbf{w}}(\mathbf{x})$ and y is determined index. According to the definition of margin in multi-classification, margin loss function is also defined unlike binary classification.

Margin loss function. For a multi-classifier $f_{\mathbf{w}}$, the margin loss function is defined as follows:

$$L_\gamma(f_{\mathbf{w}}) = \Pr_{(\mathbf{x},y)\sim\mathcal{D}}[f_{\mathbf{w}}(\mathbf{x})[y] - \max_{i\neq y} f_{\mathbf{w}}(\mathbf{x})[i] < \gamma],$$

where $f_{\mathbf{w}}(\mathbf{x})[i]$ is the i-th index of the output when input $\mathbf{x}$ and y is target index. The empirical margin loss function is denoted by $\hat{L}$. If $\gamma = 0$, it corresponds to the classification loss L_0 and $\hat{L}_0$ represent the expected loss function and the training loss function, respectively.

For neural networks for multi-classification, the Kullback-Liebler branch condition was calculated in [10] using the assumption that the weight parameter's perturbations are Gaussian random values. Using this method, we examine the difference of generalization bounds caused by injecting random noises in input. Then the generalization bounds means how flexible the network is to input noises. And we considered the idea that margin can absorb some perturbations from inputs. Therefore, we consider that the norm of input noise is bound to the magnitude of margin. Thus, we need the information about the relation between the norm of the perturbation and the magnitude of the margin. There issues are discussed in the first part of the next section.

3.1. Generalization bounds with input noises

We define input noise as follows.

Input perturbation. For input perturbation vector $\mathbf{u} \in \mathbb{R}^n$, we note that $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{u}$. Here, we assume that each element of $\mathbf{u}$ is a random number independent identically distributed $\mathcal{N}(0, \sigma_u^2)$. For convenience, sometimes we denote that the classifier added input perturbation $\mathbf{u}$ as $\tilde{f}_{\mathbf{w}} = f_{\mathbf{w}}(\mathbf{x} + \mathbf{u})$, i.e., $\tilde{f}_{\mathbf{w}} = W_n\phi(W_{n-1}\cdots\phi(W_1\tilde{\mathbf{x}}))$ where $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{u}$.

In this paper, we examine the effect of margin on training a network with input perturbation. Our goal is to find a generalization error bound for classifier added input perturbation in terms of the margin-based generalization error bound using the method proposed in [10]. The following lemma is the first step.

Lemma 3.1. *Let $f_{\mathbf{w}}(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^k$ be any k-classifier that has parameters $\mathbf{w}$, let P be any distribution on those parameters that is independent of the training data. Let $\tilde{f}_{\mathbf{w}} = f_{\mathbf{w}}(\mathbf{x}+\mathbf{u})$ where $\mathbf{u}$ is an input perturbation. Then, for any γ, $\delta > 0$, with probability of at least $1-\delta$ over the training set of size m, for any $\mathbf{w}$, and any random perturbation $\mathbf{u}$ such that $\Pr_{\mathbf{u}}[\max_{\mathbf{x}\in\mathcal{X}}(|f_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u}) - f_{\mathbf{w}}(\mathbf{x})|_\infty) < \frac{\gamma}{4}] \geq \frac{1}{2}$, we have*:

$$L_0(f_{\mathbf{w}}) \leq \hat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{KL(\mathbf{w}+\mathbf{v}\|P) + \ln\frac{m}{\delta}}{m}}\right). \tag{6}$$

Proof. Let $\mathbf{w}' = \mathbf{w}+\mathbf{v}$ and let $\mathcal{S}_{\mathbf{w}}$ be the set of perturbations that satisfies with the following:

$$\mathcal{S}_{\mathbf{w}} = \{\mathbf{w}' : \max_{\mathbf{x}\in\mathcal{X}} (|f_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u}) - f_{\mathbf{w}}(\mathbf{x})|_\infty) < \frac{\gamma}{4}\}.$$

Let $\tilde{Q}$ be a new distribution over $f_{\tilde{\mathbf{w}}}$ where $\tilde{\mathbf{w}}$ is restricted to $\mathcal{S}_{\mathbf{w}}$. The probability density function of $\tilde{Q}$ is:

$$\tilde{q}(\tilde{\mathbf{w}}) = \frac{1}{Z}\begin{cases} q(\tilde{\mathbf{w}}) & \text{for } \tilde{\mathbf{w}} \in \mathcal{S}_{\mathbf{w}}, \\ 0 & \text{otherwise,} \end{cases}$$

where Z is a normalizing constant.

By trigonometric inequality, we also have $|f_{\mathbf{w}'}(\mathbf{w}+\mathbf{u}) - f_{\mathbf{w}}(\mathbf{x})|_\infty < \frac{\gamma}{4}$. By the definition of $\mathcal{S}_{\mathbf{w}}$, for any $\tilde{\mathbf{w}} \in \mathcal{S}_{\mathbf{w}}$ and $\mathbf{x} \in \mathcal{X}_{B,n}$,

$$\max_{i,j\in[k]} \mid (|f_{\tilde{\mathbf{w}}}(\mathbf{x}+\mathbf{u})[i] - f_{\tilde{\mathbf{w}}}(\mathbf{x}+\mathbf{u})[j]|) - (|f_{\mathbf{w}}(\mathbf{x})[i] - f_{\mathbf{w}}(\mathbf{x})[j]|) \mid \; < \frac{\gamma}{2}.$$

Then, we have

$$L_0(f_{\mathbf{w}}) \leq L_{\frac{\gamma}{2}}(\tilde{f}_{\tilde{\mathbf{w}}}),\; L_{\frac{\gamma}{2}}(\tilde{f}_{\tilde{\mathbf{w}}}) \leq L_\gamma(f_{\mathbf{w}}).$$

Thus,

$$\begin{aligned} L_0(f_{\mathbf{w}}) &\leq \mathbb{E}_{\tilde{\mathbf{w}}}[L_{\frac{\gamma}{2}}(\tilde{f}_{\tilde{\mathbf{w}}})] \\ &\leq \mathbb{E}_{\tilde{\mathbf{w}}}[\hat{L}_{\frac{\gamma}{2}}(\tilde{f}_{\tilde{\mathbf{w}}})] + 2\sqrt{\frac{2(KL(\tilde{\mathbf{w}}\|P) + \ln\frac{2m}{\delta})}{m-1}} \\ &\leq \hat{L}_\gamma(f_{\mathbf{w}}) + 2\sqrt{\frac{2(KL(\tilde{\mathbf{w}}\|P) + \ln\frac{2m}{\delta})}{m-1}} \\ &\leq \hat{L}_\gamma(f_{\mathbf{w}}) + 4\sqrt{\frac{2(KL(\mathbf{w}'\|P) + \ln\frac{6m}{\delta})}{m-1}}. \end{aligned}$$

The last inequality is from the normalized constant Z. □

We began with the idea that some perturbations can be converted into margin. As a natural line of thinking, we must have information about the relationship between the magnitude of perturbations and the margin.

Lemma 3.2. *Let $f_{\mathbf{w}}(\mathbf{x}) : \mathcal{X}_{B,n} \to \mathbb{R}^k$ be a d-layer neural network with ReLU activations for any $d > 0$. Then for any $\mathbf{w}$, and $\mathbf{x} \in \mathcal{X}_{B,n}$, and any perturbations $\mathbf{v} = \{V_i\}_{i=1}^d$ and $\mathbf{u} \in \mathbb{R}^k$ such that $\|V_i\|_2 \leq \frac{1}{d}\|W_i\|_2$, the change in the output of the network can be bounded as follows:*

$$\begin{aligned}&|f_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u}) - f_{\mathbf{w}}(\mathbf{x})|_2 \\ &\leq e\left(\prod_{j=1}^{i+1} \|W_j\|_2\right)\left[\left(\sum_{j=1}^{i+1} \frac{\|V_j\|_2}{\|W_j\|_2}\right)\|\mathbf{x}\|_2 + \|\mathbf{u}\|_2\right],\end{aligned} \tag{7}$$

where e is the natural logarithm.

Proof. Let $f^i_{\mathbf{w}}(\mathbf{x})$ be the output of the i-th layer. We prove the inequality using induction.

Assume that for $i \geq 0$,

$$\begin{aligned}\Delta_i &= |f^i_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u}) - f^i_{\mathbf{w}}(\mathbf{x})|_2 \\ &\leq \left(1+\frac{1}{d}\right)^i \left(\prod_{j=1}^{i} \|W_j\|_2\right)\left[\left(\sum_{j=1}^{i} \frac{\|V_j\|_2}{\|W_j\|_2}\right)\|\mathbf{x}\|_2 + \|\mathbf{u}\|_2\right].\end{aligned}$$

If $i = 0$: $\Delta_0 = \|\mathbf{u}\|_2$ which satisfies the assumed inequality.

For $i \geq 1$:

$$\begin{aligned}\Delta_{i+1} &= |(W_{i+1} + V_{i+1})\ \phi_i(f^i_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u})) - W_{i+1}\ \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))|_2 \\ &\leq |(W_{i+1} + V_{i+1})\ \phi_i(f^i_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u})) - (W_{i+1} + V_{i+1})\ \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))|_2 \\ &\quad +\ |V_{i+1}\ \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))|_2 \\ &\leq (\|W_{i+1}\|_2 + \|V_{i+1}\|_2)\ |\phi_i(f^i_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u})) - \phi_i(f^i_{\mathbf{w}}(\mathbf{x}))|_2 \\ &\quad + \|V_{i+1}\|_2|\phi_i(f^i_{\mathbf{w}}(\mathbf{x}))|_2 \\ &\leq (\|W_{i+1}\|_2 + \|V_{i+1}\|_2)\Delta_i + \|V_{i+1}\|_2|(f^i_{\mathbf{w}}(\mathbf{x}))|_2 \\ &\leq \left(1+\frac{1}{d}\right)\|W_{i+1}\|_2\ \Delta_i + \|V_{i+1}\|_2\left(\prod_{j=1}^{i}\|W_j\|_2\right)\|\mathbf{x}\|_2.\end{aligned}$$

The last inequality is from
$|(f^i_{\mathbf{w}}(\mathbf{x}))|_2 = |W_i(\phi(W_{i-1}\cdots\phi(W_1(\mathbf{x}))))|_2 \leq \left(\prod_{j=1}^i \|W_j\|_2\right)\|\mathbf{x}\|_2$. Substituting Δ_i,

$$\begin{aligned}&\Delta_{i+1} \\ &\leq \left(1+\frac{1}{d}\right)\|W_{i+1}\|_2\ \left(1+\frac{1}{d}\right)^i\left(\prod_{j=1}^{i}\|W_j\|_2\right)\left[\left(\sum_{j=1}^{i}\frac{\|V_j\|_2}{\|W_j\|_2}\right)\|\mathbf{x}\|_2 + \|\mathbf{u}\|_2\right] \\ &\quad + \|V_{i+1}\|_2\left(\prod_{j=1}^{i}\|W_j\|_2\right)\|\mathbf{x}\|_2\end{aligned}$$

$$
\begin{aligned}
&= \left(1+\frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\|W_j\|_2\right)\left[\left(\sum_{j=1}^{i}\frac{\|V_j\|_2}{\|W_j\|_2}\right)\|\mathbf{x}\|_2+\|\mathbf{u}\|_2\right]\\
&\quad+\frac{\|V_{i+1}\|_2}{\|W_{i+1}\|_2}\left(\prod_{j=1}^{i+1}\|W_j\|_2\right)\|\mathbf{x}\|_2\\
&\le \left(1+\frac{1}{d}\right)^{i+1}\left(\prod_{j=1}^{i+1}\|W_j\|_2\right)\left[\left(\sum_{j=1}^{i+1}\frac{\|V_j\|_2}{\|W_j\|_2}\right)\|\mathbf{x}\|_2+\|\mathbf{u}\|_2\right].
\end{aligned}
$$

Here, $\left(1+\frac{1}{d}\right)^{i+1} < e$, it satisfies the assumption, thus the inequality is proved. □

Here, we let $\beta = \left(\prod_{i=1}^{d}\|W_i\|_2\right)^{1/d}$. A network with normalized weights $\widetilde{W}_i = \frac{\beta}{\|W_i\|_2}W_i$ has the same losses as a network with weights $\|W_i\|_2$. Thus we can use β rather than the actual norm $\tilde{\beta}$ as required and it is sufficient if $|\beta-\tilde{\beta}| < \frac{1}{d}\beta$ is satisfied. We will consider a fixed $\tilde{\beta}$ and use the relation $\frac{1}{e}\beta^{d-1} \le \tilde{\beta}^{d-1} \le e\beta^{d-1}$.

By setting, the inequality (7) can be written as:

$$
|f_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u}) - f_{\mathbf{w}}(\mathbf{x})|_2 \le e\beta^d\left[\frac{B}{\beta}\sum_{j=1}^{d}\|V_j\|_2+\|\mathbf{u}\|_2\right]. \tag{8}
$$

Under the assumption that the elements of V_i and u_i are Gaussian distributed random variables and their norms are bounded by $\frac{1}{d}\beta$ and B, respectively, we consider that the variances have the following relation:

$$
\sigma_u \sim \frac{dB}{\beta}\sigma_v. \tag{9}
$$

Using this condition, we can compute the KL-term with respect to the variance of weights.

The bound for $\|V_i\|_2$ we use the following inequality [14]:

$$
\Pr_{V_i\sim N(0,\sigma_v I)}\left[\|V_i\|_2 > t\right] \le 2he^{t^2/2h\sigma_v^2}
$$

and Chebyshev's inequality $\Pr_{\mathbf{u}\sim N(0,\sigma_u^2 I)}\left[|\mathbf{u}|_2 > s\right] \le \frac{n\sigma_u^2}{s^2}$. For the condition with probability of at least $\frac{1}{4}$ for each, s and t are chosen as follows:

$$
\begin{aligned}
|\mathbf{u}|_2 &\le 2\sigma_u\sqrt{n},\\
\|V_i\|_2 &\le \sigma_v\sqrt{2h\ln 8dh},
\end{aligned} \tag{10}
$$

for $1 \le i \le d$. Thus, the inequality (8) is:

$$
|f_{\mathbf{w}+\mathbf{v}}(\mathbf{x}+\mathbf{u}) - f_{\mathbf{w}}(\mathbf{x})|_2 \le e\beta^d\left[\frac{Bd}{\beta}\sigma_v\sqrt{2h\ln 8dh}+2\sigma_u\sqrt{n}\right]
$$

$$
\begin{aligned}
&\leq e\beta^{d-1}\left(dB\sigma_v\sqrt{2h\ln 8dh} + 2dB\sigma_v\sqrt{n}\right)\\
&\leq edB\beta^{d-1}\sigma_v(\sqrt{2h\ln 8dh} + 2\sqrt{n}).
\end{aligned}
$$

The first inequality is from $\beta^{d-1} \leq e\tilde{\beta}^{d-1}$ and the equality is from the relation (9). Now, using the assumption of the perturbation bound $\frac{\gamma}{4}$, we can derive the formula for σ_v:

$$
\sigma_v = \frac{\gamma}{4edB\beta^{d-1}(\sqrt{2h\ln 8dh} + 2\sqrt{n})}. \tag{11}
$$

Finally, we can postulate the following theorem.

Theorem 3.3. *Let $f_{\mathbf{w}}(\mathbf{x}) : \mathcal{X}_{B,n} \to \mathbb{R}^k$ be a k-classifier where the parameters are $\mathbf{w}$, and $\tilde{f}_{\mathbf{w}} = f_{\mathbf{w}}(\mathbf{x}+\mathbf{u})$ where $\mathbf{u} \in \mathbb{R}^n$ is an input noise. For any γ, $\delta > 0$ and $\mathbf{u}$, with probability of at least $1-\delta$ over a training set of size m, we have*:

$$
L_0(\tilde{f}_{\mathbf{w}}) \leq \hat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{B^2d^2(h\ln dh + n)M + \ln\frac{dm}{\delta}}{m\gamma^2}}\right),
$$

where $M = \Pi_{i=1}^d \|W_i\|_2^2 \sum_{i=1}^d \frac{\|W_i\|_F}{\|W_i\|_2^2}$.

Proof. Using the equation (11), the Kullback-Leibler divergence term is bounded by

$$
\begin{aligned}
KL(\mathbf{w}+\mathbf{v}\|P) \leq \frac{\|\mathbf{w}\|_2^2}{2\sigma_{\mathbf{v}}^2} &= \frac{16e^2d^2B^2\tilde{\beta}^{2d-2}(\sqrt{2h\ln 8dh} + 2\sqrt{n})^2}{2\gamma^2}\sum_{i=1}^d \|W_i\|_F^2\\
&\leq \mathcal{O}\left(\frac{B^2d^2\beta^{2d}(h\ln dh + n)}{\gamma^2}\sum_{i=1}^d \frac{\|W_i\|_F}{\beta^2}\right)\\
&\leq \mathcal{O}\left(\frac{B^2d^2(h\ln dh + n)\Pi_{i=1}^d\|W_i\|_2^2}{\gamma^2}\sum_{i=1}^d \frac{\|W_i\|_F}{\|W_i\|_2^2}\right).
\end{aligned}
$$

We applied $(\sqrt{a}+\sqrt{b})^2 \leq \mathcal{O}(a+b)$ to the first inequality.

By Lemma 3.1, for any γ, $\delta > 0$ and $\mathbf{u}$, $\tilde{\beta}$, with probability at least $1-\delta$ over the training set of size m, for any $\mathbf{w}$ such that $|\beta - \tilde{\beta}| < \frac{1}{d}\beta$, we have:

$$
L_0(\tilde{f}_{\mathbf{w}}) \leq \hat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{B^2d^2(h\ln dh + n)M + \ln\frac{m}{\delta}}{m\gamma^2}}\right),
$$

where $\tilde{f}_{\mathbf{w}}(\mathbf{x}) = f_{\mathbf{w}}(\mathbf{x}+\mathbf{u})$ and $M = \Pi_{i=1}^d \|W_i\|_2^2 \sum_{i=1}^d \frac{\|W_i\|_F}{\|W_i\|_2^2}$.

Here, $|\beta - \tilde{\beta}| < \frac{1}{d}\left(\frac{\gamma}{2m}\right)^{1/d}$ is a sufficient condition for $\mathbf{w}$ such that $|\beta - \tilde{\beta}| < \frac{1}{d}\beta$. Thus we can use a cover of size $dm^{\frac{1}{2d}}$, and the left term $\ln\frac{dm}{\delta}$ of the theorem statement is chosen. □

Without input noise, the margin-based generalization bound is as follows:

$$L_0(f_{\mathbf{w}}) \leq \hat{L}_\gamma(f_{\mathbf{w}}) + \mathcal{O}\left(\sqrt{\frac{B^2 d^2 h \ln(dh) M + \ln \frac{dm}{\delta}}{m\gamma^2}}\right).$$

Thus, we can observe that the generalization bound with input noise is relies on (input dimension) × (norms of input and weights); the added term caused by input noise.

3.2. Comparison to adversarial example

We compare our bound with the bounds in [15], those are based on [2] and [10], respectively; they showed the margin-based generalization bounds using different tools, but their results are similar.

Focusing on only effect from the input noises, our result is related to:

$$\frac{n^{1/2}\|W_1\|_F\|W_2\|_F}{\gamma\sqrt{m}}$$

for two-layer neural networks. Yin et al. [15] showed the generalization bounds for ℓ_∞ adversarial attacks using Rademacher complexity. Their results are related to:

- $\frac{\|W\|_{2,\infty}^T k^{3/2} n^{1/2}}{\gamma\sqrt{m}}$ for linear multi-class classifiers where W is from $f_W(\mathbf{x}) = W\mathbf{x}$,
- $\frac{\|W_1\|_1\|W_2^T\|_{2,1}}{\gamma\sqrt{m}}$ for two-layer neural networks for binary-classifications.

The norm of weights is significant. The existence of the terms about the dimensions of input and output depends on the type of norm. Here, we only consider a two-layer classifier; however the number of weights is not significant. In addition, although the PAC-Bayesian tool is a simple approach, the specific classification criteria do not matter.

It appears that whether or not the perturbations are random is not consequential, because the worst case is considered when we compute the generalization bounds. For example, under the ℓ_∞ adversarial attack, the difference ϵ for each coefficient is bounded by $\sqrt{n}\epsilon$ in ℓ_2-norm sense as similar with inequality (7) which contains input noise $\mathbf{u}$ bound, and the norm appears the generalization bounds.

4. Conclusion

We show the generalization bounds increases as margin increases exponentially inversely proportional to the noise level in binary classification. Also, it was confirmed that the more the input is exposed to noise, the greater the generalization bound. In multi-classification, we show the generalization bound with input noise increases in proportion to the product of input dimension and norms of input and weights. This result is similar to the case of the adversarial example considering the norm size of noise, which is considered because the generalization bound considers the worst case.

Schmidt et al. [11] showed that, for the Gaussian and Bernoulli models, the classification error rate differs according to the input data model. The difference is the exponential coefficient of the input dimension. Based on their results, they suggested more prior information is required. Jiang et al. [5] presented a new loss function inspired by the theory of large margin and conducted experiments that showed robustness for both Gaussian noise and adversarial perturbations. These two papers considered margin in hidden layers; we think that tracking margin from the output to the hidden layers would provide a better understanding of how neural networks work and enable more meaningful robust adversarial generalization.

Zhang et al. [16] showed the result from experiments that statistical learning theory has not yet been able to fully explain generalization. Thus various and new attempts to close the gap between theory and experience are needed. This remains our future work.

References

[1] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, *Stronger generalization bounds for deep nets via a compression approach*, arXiv preprint arXiv:1802.05296, 2018.

[2] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, *Spectrally-normalized margin bounds for neural networks*, In Advances in Neural Information Processing Systems 30, pages 6240–6249, 2017.

[3] G. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, *Large margin deep networks for classification*, In Advances in neural information processing systems, pages 842–852, 2018.

[4] D. Haussler, *Probably approximately correct learning*, University of California, Santa Cruz, Computer Research Laboratory, 1990.

[5] Y. Jiang, D. Krishnan, H. Mobahi, and S. Bengio, *Predicting the generalization gap in deep networks with margin distributions*, arXiv preprint arXiv:1810.00113, 2018.

[6] J. Langford and J. Shawe-Taylor, *Pac-bayes & margins*, In Advances in neural information processing systems, pages 439–446, 2003.

[7] D. A. McAllester, *PAC-Bayesian model averaging*, in Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999), 164–170, ACM, New York, 1999. https://doi.org/10.1145/307400.307435

[8] D. A. McAllester, *Pac-bayesian stochastic model selection*, Machine Learning **51** (2003), no. 1, 5–21.

[9] D. McAllester, *Simplified pac-bayesian margin bounds*, In Learning theory and Kernel machines, pages 203–215. Springer, 2003.

[10] B. Neyshabur, S. Bhojanapalli, and N. Srebro, *A pac-bayesian approach to spectrally-normalized margin bounds for neural networks*, International Conference on Learning Representations, 2018.

[11] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, *Adversarially robust generalization requires more data*, In Advances in Neural Information Processing Systems, pages 5014–5026, 2018.

[12] L. Shen-Huan, W. Lu, and Z. Zhi-Hua, *Optimal margin distribution network*, CoRR, abs/1812.10761, 2018.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, *Intriguing properties of neural networks*, arXiv preprint arXiv:1312.6199, 2013.

[14] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math. **12** (2012), no. 4, 389–434. `https://doi.org/10.1007/s10208-011-9099-z`

[15] D. Yin, K. Ramchandran, and P. Bartlett, *Rademacher complexity for adversarially robust generalization*, International Conference on Machine Learning, 2019.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, *Understanding deep learning requires rethinking generalization*, arXiv preprint arXiv:1611.03530, 2016.

Hi Jun Choe
Department of Mathematics
Yonsei University
Seoul 03722, Korea
Email address: `choe@yonsei.ac.kr`

Hayeong Koh
Software Testing & Certification Laboratory
Telecommunication Technology Association
Gyeonggi-do 13591, Korea
Email address: `hykoh@yonsei.ac.kr`

Jimin Lee
Center for Mathematical Analysis & Computation
Yonsei University
Seoul 03722, Korea
Email address: `jmlee524@yonsei.ac.kr`