

On the Relationship between Populated Regions and Adversarial Robustness in Deep Neural Networks

Seongjin Park^{*†}
SAIT
Suwon, South Korea
ssjin.park@samsung.com

Haedong Jeong^{*†}
Sogang University
Seoul, South Korea
haedong@sogang.ac.kr

Tair Djanibekov^{*†}
MBZUAI
Abu Dhabi, United Arab Emirates
tair.djanibekov@mbzuai.ac.ae

Giyoun Jeon[†]
LG AI Research
Seoul, South Korea
giyoung.jeon@lgresearch.ai

Jinseok Seol^{†‡}
Dankook University
Yongin, South Korea
jinseok.seol@dankook.ac.kr

Jaesik Choi[‡]
KAIST & INEEJI
Daejeon & Seongnam, South Korea
jaesik.choi@kaist.ac.kr

Abstract—In general, deep neural networks (DNNs) are evaluated by the generalization performance measured on unseen data excluded from the training phase. Along with the development of DNNs, the generalization performance converges to the state-of-the-art performances and it becomes difficult to evaluate DNNs solely based on this metric. The robustness against adversarial attack has been used as an additional metric to evaluate DNNs by measuring their vulnerability. However, few studies have been performed to analyze the adversarial robustness in terms of the geometry in DNNs. In this work, we perform an empirical study to analyze the internal properties of DNNs that affect model robustness under adversarial attacks. In particular, we propose the novel concept of the populated region set (PRS), where training samples are actually populated, to represent the internal properties of DNNs in a practical setting. From systematic experiments with the proposed concept, we provide empirical evidence to validate that a low PRS ratio has a strong relationship with the adversarial robustness of DNNs. We also devise a PRS regularizer leveraging the characteristics of PRS to improve the adversarial robustness without adversarial training.

Index Terms—Decision Region, Adversarial Robustness, Robust Training

I. INTRODUCTION

With the steep improvement of the performance of Deep Neural Networks (DNNs), their applications are expanding in the real world [1]. For real world application, it may be necessary to choose the best model among the candidates. Traditionally, the generalization performance which measures the objective score on the test dataset excluded in the training phase, is used to evaluate the models [2]. However, it is non-trivial to evaluate DNNs based on this single metric. For example, if two networks with the same structure have the similar test accuracy, it is ambiguous which is better. Robustness against adversarial attacks, measure of the vulnerability, can be

^{*}Equally contributed authors.

[†]Most of this work was done while the authors were at KAIST.

[‡]Both are corresponding authors.

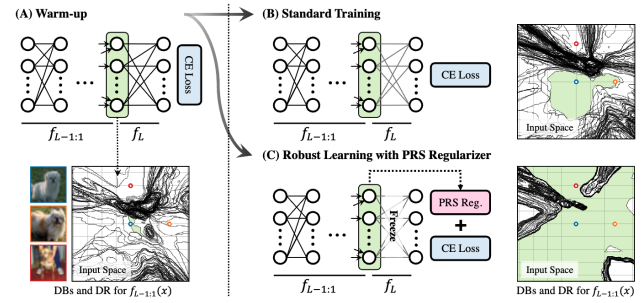


Fig. 1: An illustrative comparison of each training method with CIFAR-10, and visualization for decision boundaries/regions (DBs/DRs) of penultimate layer in the input space ($f^{(L-1:1)}(x)$). For visualization, we randomly select three dog images and depict section of input space. The green area indicates DR which the blue boxed image populates. (A) Warm-up stage for VGG-16 with standard training (cross-entropy loss). (B) Standard training after warm up stage. (C) The robust learning with devised PRS regularizer after warm up stage. We identify that each training scheme induces different configuration of DBs/DRs, which represents different internal properties of DNNs.

an alternative to evaluate DNNs [3]–[9]. Adversarial attacks aim to induce model misprediction by perturbing the input with small magnitude. Most previous works were focused on the way to find adversarial samples by utilizing the model properties such as gradients with respect to the loss function. Given that the adversarial attack seeks to find the perturbation path on the model prediction surface over the input space, robustness can be expressed in terms of the geometry of the model. However, few studies have been performed to interpret the robustness with the concept of the geometric properties of

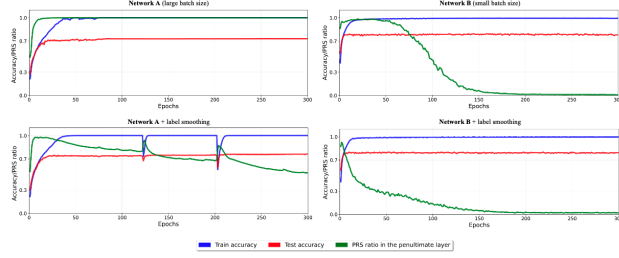


Fig. 2: Training/test accuracy and the PRS ratio on the penultimate layer on CNN-6 with batch size 2048 and 128. We select the networks at the 300th epoch and denote these two CNN-6 by Network A and B, respectively, throughout the paper (PRS ratio of Network A: 0.99, and Network B: 0.007). We also should be scenario where the label smoothing is applied.

DNNs. From a geometric viewpoint, the internal properties of DNNs are represented by the boundaries and the regions. It is shown that the DNNs with piece-wise linear activation layers are composed of many linear regions, and the maximal number of these regions is mathematically related to the expressivity of DNNs [10], [11]. As these approaches only provide the upper bound for the expressivity with the same structured model, it does not explain how much information the model actually expresses.

In this work, we investigate the relationship between the internal properties of DNNs and the adversarial robustness. In particular, our approach analyzes the internal characteristics from the perspective of the decision boundary (DB) and the decision region (DR), which are basic components of DNNs [12]. To avoid insensitivity of the maximal number of linear regions in the same structure assumption, we propose the novel concept of the populated region set (PRS), which is a set of DRs containing at least one sample included in the training dataset. Since the PRS can be considered as the feasible complexity of the model, we hypothesize that the size of PRS is related to the robustness of network. To validate our hypothesis, we perform systematic experiments with various structures of DNNs and datasets. Our observations are summarized as follows:

- 1) The models with the same structure can have different size of PRS, although they have similar generalization performance. We empirically show that the model with a small size of the PRS tends to show higher robustness compared to that with a large size. (in Section IV-B)
- 2) We observe that when the model achieves a low PRS ratio, the linear classifier which maps the penultimate features to the logits has high cosine similarity between parameters corresponding to each class (in Section IV-B).
- 3) We verify that the size of intersection of the PRS from the training/test dataset is related to the robustness of model. The model with a high PRS inclusion ratio of test samples has higher robustness than that with a low

PRS inclusion ratio (in Section IV-C).

- 4) We devise a novel regularizer leveraging the characteristics of PRS to improve the robust accuracy without adversarial training (in Section V).
- 5) We also provide a theoretical analysis of the relationship between the proposed PRS ratio and techniques known to enhance robustness, such as adjusting the batch size and applying label smoothing (in Section VI).

II. RELATED WORK

The adversarial attack which reveals the vulnerability of DNNs, is mainly used to validate the reliability of the trained network. As an early stage for adversarial attacks, the fast gradient sign method (FGSM) [3] based on the gradient with respect to the loss function and the multi-step iterative method [13] are proposed to create adversarial examples to change the model prediction with a small perturbation. Recently, many studies on effective attacks in various settings have been performed to understand the undesirable decision of the networks [14]–[16]. In terms of factors affecting robustness, [17] provide evidence to argue that training with a large batch size can degrade the robustness of the model against the adversarial attack from the perspective of the Hessian spectrum.

With increasing interest in the expressive power of DNNs, there have been several attempts to analyze DNNs from a geometric perspective [18], [19]. In these studies, the characteristics of the decision boundary or regions formulated by the DNNs are mainly discussed. [10] show that the cascade of the linear layer and the nonlinear activation organizes the numerous piece-wise linear regions. They show that the complexity of the decision boundary is related to the maximal number of these linear regions, which is determined by the depth and the width of the model. [11] extend the notion of the linear region to the convolutional layers and show the better geometric efficiency of the convolutional layers. Compared to previous work, our work focuses on the practical decision region which the trained network actually utilizes. It has also been shown that the manifolds learned by DNNs and the distributions over them are highly related to the representation capability of a network [20]. While these studies highlight the benefits of increasing expressivity of DNNs as the number of regions increases, interpreting the vulnerability of DNNs with the geometry is another important topic. [21] show that a model with thick decision boundaries induces robustness. [22] show that a decision boundary with a small curvature acquires the high robustness of the model. These approaches focus on the decision boundaries, while this paper suggests to focus on the decision regions, which are composed by the surrounding decision boundaries.

III. PROBLEM SETTING

This section describes the internal properties of DNNs from the perspective of DBs and DRs. The DBs of the DNN classifier is mainly defined as the borderline between DRs for classification, where the prediction probability of two different

classes are the same [23]. To expand the notion of DBs and DRs to the internal feature-level, we re-define the DBs in the classifier that generalizes the existing definition of DBs. We then propose the novel concept of the Populated Region Set (PRS) that describes the specific DRs used from the network for training samples.

A. Decision Boundary and Region

Let the classifier with L layers be $F(x) = (f^{(L)} \circ \phi \circ f^{(L-1)} \circ \phi \circ \dots \circ \phi \circ f^{(1)})(x) = f^{(L:1)}(x)$, where x is the sample in the input space $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\phi(\cdot)$ denotes the non-linear activation function¹. For the l -th layer, $f^{(l)}(\cdot)$ denotes a linear operation and $f_i^{(l:1)}(\cdot)$ denotes the value of the i -th element of the feature vector $f^{(l:1)}(x) \in \mathbb{R}^{d_l}$. We define the decision boundary for the i -th neuron of the l -th layer.

Definition III.1 (Decision Boundary (DB)). *The i -th decision boundary $DB_i^{(l)}$ at the l -th layer is defined as:*

$$DB_i^{(l)} = \{x \in \mathcal{X} \mid f_i^{(l:1)}(x) = 0\}. \quad (1)$$

We note that the $DB_i^{(l)}$ with $l < L$ divides the input space $\mathcal{X}$ based on the hidden representation of the l -th layer (i.e., existence of feature and the amount of feature activation). There are d_l boundaries and the configuration of the DBs are arranged by the training. As input samples in the same classification region are considered to belong to the same class, the input samples placed on the same side of the internal $DB_i^{(l)}$ share the similar feature representation. In this sense, we define the decision region, which is surrounded by DBs.

Definition III.2 (Decision Region (DR)). *Let $\sigma \in \{-1, +1\}^{d_l}$ be a signature vector. Then the **decision region** $DR_\sigma^{(l)}$, which shares the signature of the feature representation, is defined as:*

$$DR_\sigma^{(l)} = \{x \in \mathcal{X} \mid \text{sign}(f_{1:l}(x)) = \sigma\}. \quad (2)$$

Fig. 1 presents each training scheme for VGG-16 with CIFAR-10 and the internal DBs/DRs of the penultimate layer ($f^{(L-1:1)}(x)$). To visualize the DBs and DRs in the 2D space, we randomly select three training images (red, blue, and orange box) in dog class and make a hyperplane with these images. The standard training ((A) $\rightarrow$ (B)) and the proposed robust learning ((A) $\rightarrow$ (C)) are performed after warm-up stage (a stage without the regularizer), respectively. We identify that the proposed regularizer induces different configuration of DBs/DRs in the input space compared to the standard training ((A) $\rightarrow$ (B)).

B. Populated Region Set

It is well-studied that the number of DRs is related to the representation power of DNNs [10], [11]. In particular, the expressivity of DNNs with partial linear activation function is quantified by the maximal number of the linear regions and this number is related to the width and depth of the structure. We believe that although the maximal number can be one

¹Although there are various activation functions, we only consider ReLU activation for this paper.

measure of expressivity, the trained DNNs with finite training data² cannot handle the entire regions to solve the task. To only consider DRs that the network uses in the training process, we devise the train-related regions where training samples are populated more frequently. We define the populated region set (PRS), which is a set of DRs containing at least one sample included in the training dataset. PRS will be used to analyze the relationship between the geometrical property and the robustness of DNNs in a practical aspect.

Definition III.3 (Populated Region Set (PRS)). *From the set of every DRs of the model f and given the dataset $\mathcal{D}$, the **populated region set** $PRS^{(l)}$ is defined as:*

$$PRS^{(l)}(\mathcal{D}) = \{DR_\sigma^{(l)} \mid DR_\sigma^{(l)} \cap \mathcal{X}_{\mathcal{D}} \neq \emptyset\}, \quad (3)$$

where $\mathcal{X}_{\mathcal{D}}$ denote the inputs of the dataset.

We note that the size of the PRS is bounded to the size of given dataset $\mathcal{D}$. When $|PRS^{(l)}| = \mathcal{D}$, each sample in training dataset is assigned to each distinct DR in the l -th layer. To compare the PRS of networks, we define the PRS ratio as $|PRS^{(l)}|/|\mathcal{D}|$, which measures the ratio between the size of the PRS and the given dataset. Fig. 2 presents a comparison between two equivalent neural networks (A and B) with six convolution blocks (CNN-6) trained on CIFAR-10 varying only the batch size (2048 and 128, respectively). Here, we first observe the following: (1) Training with a large batch size is known to be less robust compared to its small-batch counterpart from various perspectives, and we can see that this phenomenon is expressed through the PRS ratio. Going forward, we will utilize Networks A and B, which differ only in PRS, to make direct comparisons. (2) Conversely, label smoothing is known to enhance robustness, and we confirm that this effect also manifests itself through the PRS ratio. We observe that even with larger batch sizes, the PRS ratio tends to decrease gradually. Inspired by this observation, we incorporate aspects of label smoothing when designing a PRS-based regularizer later. The theoretical analysis of these two observations (batch size and label smoothing in relation to the PRS ratio) will be discussed in Section VI. From the fact that the penultimate layers are widely used as feature extraction, we only consider the PRS ratio on the penultimate layer in the remainder of the paper.

IV. ROBUSTNESS UNDER ADVERSARIAL ATTACKS

In this section, we perform experiments to analyze the relationship between the PRS ratio and the robustness. We evaluate the robustness of the network using the fast gradient sign method (FGSM) [3], basic iterative method (BIM) [13] and projected gradient descent (PGD) [15] method widely used for the adversarial attacks. The untargeted adversarial attacks using training/test dataset are performed for the various perturbations ($\epsilon = 0.0313, 0.05, 0.1$).

²In general, the number of training data is smaller than the maximal number of the linear region.

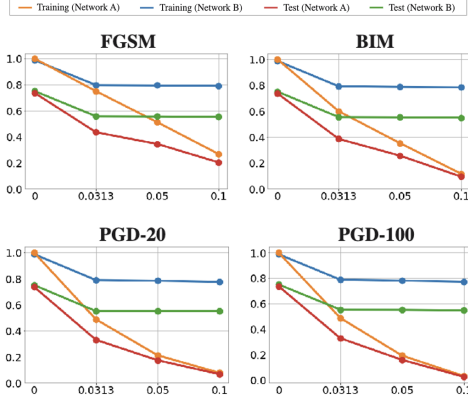


Fig. 3: Robust accuracy under various adversarial attack methods on networks A and B. The x-axis indicates perturbation ϵ and the y-axis indicates the training/test robust accuracy.

A. Experimental Setup

For the systematic experiments, we select three different structures of DNNs to analyze: (1) a convolutional neural network with six convolution blocks (CNN-6), (2) VGG-16 [24], and (3) ResNet-18 [25]. We train³ basic models with fixed five random seeds and four batch sizes (64, 128, 512 and 2048) over three datasets: MNIST, F-MNIST [26], and CIFAR-10 [27]. For the extensive analysis on the correlation between the PRS ratio and properties of network, we extract candidates from each basic model with the various epochs. Then we apply the test accuracy threshold to guarantee the sufficient and similar performance.

B. PRS and Robustness

First, we compare the two models (Network A and B in Fig. 2) with similar test accuracy but different PRS ratio⁴. Fig. 3 presents the results of robust accuracy (accuracy against adversarial attack) under the FGSM, BIM (5-step), PGD-20 (20-step), and PGD-100 (100-step) on L_∞ with $\alpha = 2/255$. We identify that Network B (low PRS ratio) is more robust than Network A (high PRS ratio) under all adversarial attacks.

As the PGD-20 shows the similar robust accuracy compare to a PGD-100, we focus on an analysis under the PGD-20 in the rest of the paper. We measure the PRS ratio and the robust accuracy in all models and datasets to verify the relationship between the PRS ratio and the robustness. For the experiments, we take the magnitude of ϵ as follow: MNIST = 0.3, F-MNIST = 0.1, and CIFAR10 = 0.0313 on L_∞ norm. Fig. 4 presents the experimental results according to the model structure under the PGD attack. To quantify the relation, we calculate the coefficient of the regression line and perform significance test to validate the trend. From Fig. 4, we identify that the PRS

³Cross-entropy loss and Adam optimizer with learning rate 10^{-3} is used.

⁴We note that different PRS ratios are obtained by different batch size of Network A (2048) and B (128).

ratio has an inversely correlated relationship with the robust accuracy in most cases.

From the above observations, we empirically confirm that the PRS ratio is related to the robustness against adversarial attacks. In order to investigate the evidence that the low PRS ratio causes robustness for the gradient-based attack, we perform an additional analysis of failed attack samples. In the gradient-based attack, as the magnitude of the gradient is a crucial component to success, we first count the ratio of the zero gradient samples in the failed attack samples.

Fig. 5a shows the ratio of success samples (light green bar), failure samples with non-zero gradient (blue bar) and zero gradient (red bar) in all samples. We note that the failed attack samples with non-zero gradients maintain the index of the largest logit as the true class after attack. To analyze the reason of failure, we examine the change of the logits under the adversarial attack. This change is shown in Fig. 5b. To clarify the difference of the change of the logits between Network A and B, we select the examples of successful attack on Network A but failed attack on Network B. In Network B, the logits move on almost parallel direction, which causes the predicted label to be maintained as the true class.

To explain the parallel change of the logit vector, we hypothesize that the DBs corresponding to each class node have similar configuration in the input space. However, it is intractable to measure the similarity between DBs in the entire network due to the highly non-linear structure and the high dimensional input space. To simplify our hypothesis, we only measure the cosine similarity between the parameters which map the features on the penultimate layer to logits (*i.e.*, final layer). Fig. 6 presents that the similarity matrices for Networks A and B. When we compare the matrix between the two models at the 300th epoch, we identify that Network B (low PRS ratio) has higher cosine similarity between each parameter in the final layer. We note that the cosine similarity between each parameter in the final layer can be considered as the degree of parallelism for the normal vectors in the linear classifier. We also confirm that the decrease of the PRS ratio is aligned with the increase of the similarity of parameters in Fig. 6b, when we consider the graph in Fig. 2. To verify the relationship between PRS ratio and the cosine similarity, we measure the PRS ratio and the cosine similarity between each parameter in all models. Table I shows the results of the correlation experiment for the relationship between the PRS ratio and the cosine similarity. We identify that the PRS ratio has an inverse correlation for the cosine similarity between each parameter in the final layer.

C. PRS and Test Samples

When we regard the model as a mapping function from the input space to the feature space, handling unseen data in a known feature domain is significant in a perspective of the generalization performance. Hence, if the majority of samples from the test dataset are assigned to the training PR, the model can be considered to be learned the informative and general concept of feature mapping. For example, if the

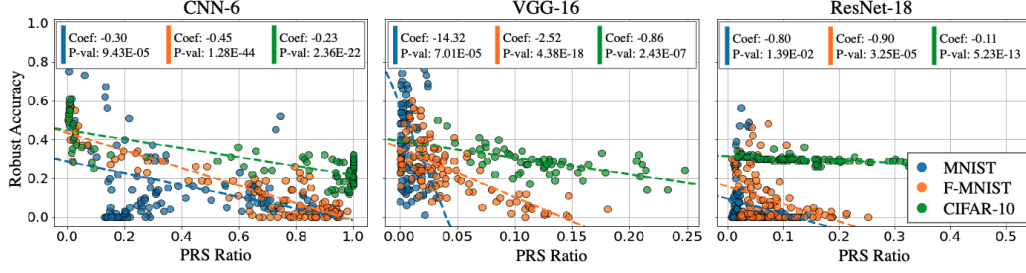


Fig. 4: Relationship between the PRS ratio and robust accuracy attacked by PGD method in various models and datasets. The colored dots are for the independent models. The colored dashed lines indicate the trend for each dataset.

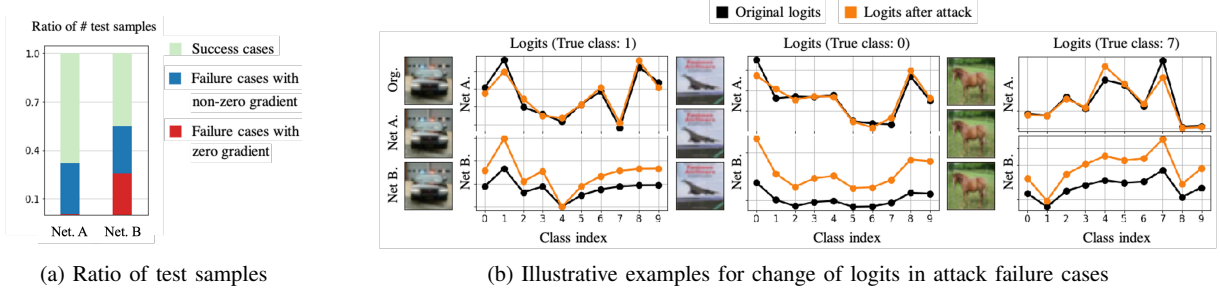


Fig. 5: (a) Comparison of the ratio of the zero gradient in the failure attack for the test samples under the PGD-20 attack on L_∞ with $\epsilon = 0.0313$ (Network A and B). (b) The illustrative examples of attacked samples on Network A and B which is failed on B, and the corresponding logits before/after the attack. After the attack, the logits move on almost parallel direction with the original logits in Network B.

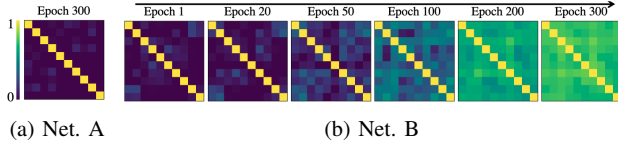


Fig. 6: (a) Cosine similarity (CS) matrix for a final layer on Network A. (b) CS matrix for epochs on Network B. As the epoch increases, the CS for each parameter increases.

arbitrary test sample is mapped to the training PR, we expect that a similar decision will appear. However, it is non-trivial to guess which type of decision will appear when the test sample is mapped to out of the training PR. To investigate the differences between the test samples which are included and excluded in the training PR, we evaluate the test accuracy under adversarial attack for each group. For a comparison, we divide both the inclusion and exclusion groups with $1k$ correctly predicted test samples.

Fig. 7 shows the robust accuracy under the FGSM, BIM with a 5-step, and the PGD-20 and PGD-100 on L_∞ . Although the robust accuracy of each test group decreases as the epsilon becomes larger, we observe that the inclusion group is more robust against all types of attacks compared to the exclusion group. Table I presents the results of the correlation experiment for the relationship between the PRS ratio and the inclusion

ratio of the test samples for the training PR. We compute the inclusion ratio as the ratio of the test samples mapped to the training PR. In Table I, we identify that the PRS ratio and the inclusion ratio have inversely correlated relationship. As we previously verify that the included test samples show high robustness, we empirically confirm that the low PRS ratio is related to the robustness under adversarial attacks.

D. PRS and Training Samples

From previous section, we empirically observe that the vulnerability of individual test samples is related to PRS defined by training samples. In this section, we categorize the PRS to expand this relationship. At first, we define the major DR for each class c which includes the majority of training samples.

Definition IV.1 (Major Region (MR)). *For a training dataset $\mathcal{D}_c$ denoting the instances with class c , the **major region** for l -th layer and class c is defined as:*

$$MR_c^{(l)} = \underset{DR_\sigma^{(l)} \in PRS^{(l)}(\mathcal{D}_c)}{\operatorname{argmax}} |\mathcal{X}_{\mathcal{D}_c} \cap DR_\sigma^{(l)}| \quad (4)$$

We note that since the training samples are finite, large number of samples occupying $MR_c^{(l)}$ directly means the low PRS ratio. We refer the remained DRs (*i.e.*, not MR) as the **extra regions** (ERs).

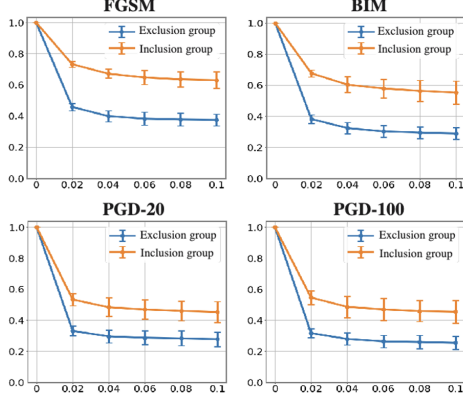


Fig. 7: Test accuracy under adversarial attacks for inclusion/exclusion groups for CNN6 on CIFAR-10 for five fixed random seeds. The x-axis indicates perturbation ϵ and the y-axis indicates the robust accuracy. The blue/orange line indicates the exclusion/inclusion groups, respectively. The exclusion group is shown to be more vulnerable under adversarial attacks.

1) *Comparison of MR and ER*: At first, we observe the distribution of training samples for type of region corresponding each class in VGG-16 trained with CIFAR-10. Fig. 9a depicts the distribution of training samples for MR (sky blue and black dashed line) and ER for each class. Although the training samples are distributed the various regions, in almost case, we identify that there are MR for entire class. To compare the characteristics of samples populated each region, we randomly selected $10k$ training samples from MR and ER. We perform adversarial attack for selected samples and measure the confidence of the prediction (logit value for the target class). Fig. 9b and 9c show the robust accuracy and confidence for MR and ER, respectively. We empirically verify the training samples in MR have higher adversarial robustness and the network predicts these samples with high confidence.

2) *Relationship between MR and Confidence*: From the empirical observations that samples belonging to MR are relatively robust, we hypothesize that samples located closer to center of MR tends to be more robust.

Definition IV.2 (Major Region Mean Vector (MRV)). *For a major region $MR_c^{(l)}$, the major region mean vector is defined as:*

$$MRV_c^{(l)} = \frac{1}{|MR_c^{(l)}|} \sum_{x \in MR_c^{(l)}} f_{l:1}(x). \quad (5)$$

To verify our hypothesis, we measure Euclidean distance between MRV and training samples with success/failure of adversarial attack. From Fig. 8a, we identify that the samples far from MRV tend to be vulnerable for the adversarial attack in the entire class. Furthermore, in Fig. 8b, we identify the inversely correlated relationship between the confidence and the distance to MRV for the failed attack samples.

V. ROBUST LEARNING VIA PRS REGULARIZER

In previous Sections, we empirically verify that PRS is related to the adversarial robustness. In particular, (1) inclusion of MR, and (2) distance to MRV are highly related to vulnerability of individual samples. From these insights, we devise a novel regularizer leveraging the properties of PRS to improve the adversarial robustness.

A. Regularizer via PRS

At first, we design the regularizer to reduce the distance between feature vector and MRV. To guarantee the quality of feature representation which constructs plausible PRS, we utilize the warm-up stage for the classifier. In the warm-up stage, the classifier is trained with cross-entropy loss function $\mathcal{L}_{ce}$ during the T -th epoch. After warm-up stage, we construct the MRV for each class and use it after T -th epoch. The regularizer for MRV is defined as:

$$\mathcal{L}_{MRV} = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \left(MRV_y^{(l)} - f^{(l:1)}(x) \right)^2. \quad (6)$$

We note that because $\mathcal{L}_{MRV}$ reduce the distance to MRV in the feature space, the arbitrary sample can have the opportunity for inclusion of MR. However, it is non-trivial to guarantee for inclusion based on Euclidean distance, because (1) in general, the feature vector is encoded in the high dimensional space, and (2) highly non-linear embedding of DNNs.

As mentioned in Section III-B, we additionally blend label smoothing techniques to ensure that the training samples are included into MR. We adopt label smoothing α into the MR regularizer, whereas we pull a feature into corresponding MRV upto $(1-\alpha)$ (compared to the original MRV regularizer), while also pulling the feature towards MRV of other classes in the ratio of α . We denote this regularizer as $\mathcal{L}_{MRV}^\alpha$.

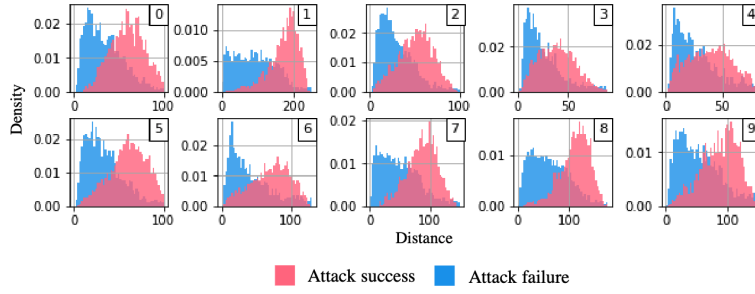
The final objective becomes:

$$\mathcal{L}_{PRS} = \lambda_{ce} \mathcal{L}_{ce} + \lambda_{MRV} \mathcal{L}_{MRV}^\alpha \quad (7)$$

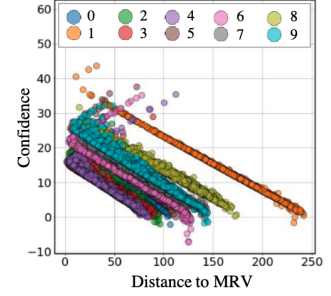
where λ_{ce} and λ_{MRV} are hyperparameters. We perform simple grid search to set hyperparameters and use $\lambda_1 = 0.2$, $\lambda_2 = 0.8$, and $\alpha = 0.1$ for remaining experiments. We denote the loss function with $\alpha = 0$ by $\mathcal{L}_{MR}$.

B. Experimental Results

To verify the effectiveness of our proposed method, we apply the regularizer to CNN-6, VGG-16 and ResNet-18 for the classification task on CIFAR-10. We set standard training ($\mathcal{L}_{ce}$) and Adversarial training (AT) based on PGD-20 attack on L_∞ as the baselines. For warm-up stage, we use parameters of classifier with standard training ($T = 50$), and freeze the final layer parameters to observe the effect of the change of PRS for robust and test accuracy. Fig. 1 depicts the training procedure for each training scheme and trained DBs and DRs which represent the state of PRS. Table II shows the results for each training scheme on various architectures on CIFAR-10. We verify that the proposed method can improve the robust accuracy while maintaining the test accuracy. We

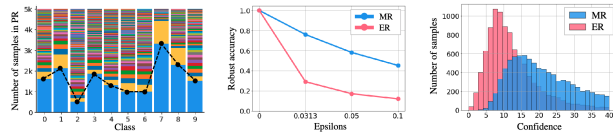


(a) Distribution of the distance between the each feature vector and the MRV



(b) Relationship between the distance to the MRV and the confidence

Fig. 8: Experimental results under VGG-16 on CIFAR-10. (a) Distribution of Euclidean distance to MRV for training samples. The blue histogram indicates the failed attack samples and the red histogram indicates the success attack samples. The white box in the upper right presents each class. (b) Relationship between the distance to MRV and the confidence. The colored dots represent training samples which are vulnerable under adversarial attack per each class.



(a) The number of training samples in DR for each class. The maximum number of samples for each class is 5k and the sky blue bar indicates each MR. (b) Robust accuracy under PGD-20 attacks on L_∞ for the samples in MR and ER. (c) Class confidence for the samples in MR and ER.

Fig. 9: Experimental results under VGG-16 on CIFAR-10. (a) The number of training samples populated in DR for each class. The maximum number of samples for each class is 5k and the sky blue bar indicates each MR. (b) Robust accuracy under PGD-20 attacks on L_∞ for the samples in MR and ER. (c) Class confidence for the samples in MR and ER.

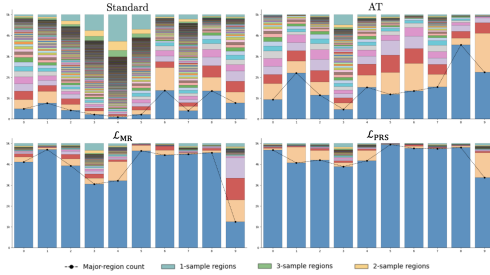


Fig. 10: Experimental results under VGG-16 on CIFAR-10. The distribution of DR for each class by our proposed method. The sky blue bar per class indicates each MR.

note that the proposed PRS regularizer does not use the adversarial examples to improve the adversarial robustness. It means that our method can have strength in the perspective of computation time. We further investigate the change of PRS after training for each method. Fig. 10 shows MR and ER of each class for each training scheme. We identify that the both

TABLE I: The coefficient and p -values of the regression analysis between the PRS ratio and each properties.

Model	Dataset	Cosine Similarity		Inclusion Ratio	
		coef.	p -value	coef.	p -value
CNN-6	MNIST	-0.76	5.17e-29	-0.72	1.09e-126
	F-MNIST	-0.58	2.43e-46	-0.79	1.78e-133
	CIFAR-10	-0.65	4.35e-54	-0.80	5.45e-136
VGG-16	MNIST	-14.70	2.70e-8	-0.53	4.66e-43
	F-MNIST	-3.38	4.11e-19	-0.48	2.22e-24
	CIFAR-10	0.28	2.92e-1	-0.79	2.07e-9
ResNet-18	MNIST	-2.28	1.36e-11	-0.53	1.58e-98
	F-MNIST	-2.24	1.66e-16	-0.65	6.38e-38
	CIFAR-10	-1.29	4.20e-14	-0.71	2.87e-35

TABLE II: Comparison of robust and test accuracy under PGD-20 attacks on L_∞ for CIFAR-10.

Model	Method	Robust Acc.	Test Acc.	PRS Ratio
CNN-6	Standard	38.82 $\pm$ 2.73	77.92 $\pm$ 0.34	.101 $\pm$.010
	AT	53.79 $\pm$ 0.42	70.65 $\pm$ 0.17	.099 $\pm$.001
	$\mathcal{L}_{MR}$ (ours)	53.59 $\pm$ 0.07	80.30 $\pm$ 0.72	.018 $\pm$.003
	$\mathcal{L}_{PRS}$ (ours)	57.10 $\pm$ 0.45	80.47 $\pm$ 0.09	.009 $\pm$.002
VGG-16	Standard	39.94 $\pm$ 1.28	80.28 $\pm$ 0.24	.115 $\pm$.012
	AT	58.18 $\pm$ 0.13	75.22 $\pm$ 0.05	.069 $\pm$.002
	$\mathcal{L}_{MR}$ (ours)	60.42 $\pm$ 0.36	78.61 $\pm$ 0.15	.038 $\pm$.007
	$\mathcal{L}_{PRS}$ (ours)	63.55 $\pm$ 0.72	79.31 $\pm$ 0.22	.018 $\pm$.003
ResNet-18	Standard	33.48 $\pm$ 0.08	76.96 $\pm$ 0.15	.065 $\pm$.001
	AT	50.65 $\pm$ 0.20	73.03 $\pm$ 0.05	.046 $\pm$.004
	$\mathcal{L}_{MR}$ (ours)	49.31 $\pm$ 0.65	76.51 $\pm$ 0.08	.061 $\pm$.003
	$\mathcal{L}_{PRS}$ (ours)	50.59 $\pm$ 0.12	76.81 $\pm$ 0.11	.059 $\pm$.002

proposed regularizers can increase the number of populated training samples in MR compared to the standard training (*i.e.*, reduce of PRS ratio). During the period of reviewing process, we share our source code in a anonymized repository⁵.

⁵https://anonymous.4open.science/r/PRS_ICDM25

VI. THEORETICAL ANALYSIS

In this section, we provide a theoretical analysis about the relationship between the batch size and PRS ratio, and also the relationship between the label smoothing and PRS ratio, whereas those two robustness-related components were heavily affecting the PRS ratio.

A. Preliminaries

For concreteness, we define and redefine some concepts with more rigorous notations.

Definition VI.1 (Class Decision Boundary). *In a softmax-based neural network model for classification task, a **logit** $l_{c;\theta}$ is defined per class c as follows:*

$$l_{c;\theta}(x) = w_{c;\theta}^\top z_{x;\theta} + b_{c;\theta}, \quad (8)$$

where $w_{c;\theta}$, $b_{c;\theta}$ denote the weight and bias parameter in the last layer for the class c , and $z_{x;\theta}$ is features of the penultimate layer for the input x . A **class decision value** $CDV_{c,c'}(x;\theta)$ is defined by the difference between logits of the two classes c and c' for an input x :

$$CDV_{c,c'}(x;\theta) = l_{c;\theta}(x) - l_{c';\theta}(x). \quad (9)$$

A **class decision boundary** $CDB_{c,c'}(\theta)$ is a set of features with zero decision values:

$$CDB_{c,c'}(\theta) = \{z_{x;\theta} \mid CDV_{c,c'}(x;\theta) = 0\}. \quad (10)$$

Definition VI.2 (Class Decision Region). A **class decision region** $CDR_c(\theta)$ for class c is defined as a intersection of regions determined by decision boundaries with other classes c' other than c :

$$CDR_c(\theta) = \bigcap_{c' \neq c} \{z_{x;\theta} \mid CDV_{c,c'}(x;\theta) > 0\}. \quad (11)$$

Definition VI.3 (Class Decision Distance). A **class decision distance** CDD is the minimal Euclidean distance from the feature point $z_{x;\theta}$ to a decision boundary:

$$CDD(z_{x;\theta}, CDB_{c,c'}(\theta)) = \frac{|CDV_{c,c'}(x;\theta)|}{\|w_{c;\theta} - w_{c';\theta}\|}. \quad (12)$$

Definition VI.4 (Margin Distance). A **margin distance** MD is the minimal CDD :

$$MD(z_{x;\theta}) = \min_{c' \neq c} CDD(z_{x;\theta}, CDB_{c,c'}(\theta)). \quad (13)$$

B. Batch Size and PRS Ratio

The proof sketch proceeds as follows: Large batch sizes are known to yield sharp minima due to the Hessian eigenvalue argument [28], and such sharp minima induce higher parameter sensitivity. Highly sensitive parameters lead to sensitive decision boundaries, resulting in smaller decision regions. Ultimately, these smaller decision regions cause a higher PRS ratio.

Definition VI.5 (Sharpness). Let $\theta \in \mathbb{R}^p$ denote parameters for an empirical loss function $\mathcal{L}$, then for a minima θ^* where $\nabla \mathcal{L}(\theta^*) = 0$, we define **sharpness** $S_{\mathcal{L}}(\theta^*)$ by the largest

eigenvalue of the Hessian at the minima, namely $S_{\mathcal{L}}(\theta^*) = \lambda_{\max}(H(\theta^*))$, where $\lambda_{\max}$ denote the largest eigenvalue of $H(\theta^*) = \nabla_{\theta}^2 \mathcal{L}(\theta^*)$, and we say a minima θ^* is sharper than another minima θ^{**} whenever $S_{\mathcal{L}}(\theta^*) > S_{\mathcal{L}}(\theta^{**})$.

For a small perturbation $\Delta\theta \in \mathbb{R}^p$, let us expand $\mathcal{L}(\theta)$ around a minima θ^* using the second-order Taylor expansion:

$$\mathcal{L}(\theta^* + \Delta\theta) \approx \mathcal{L}(\theta^*) + \nabla_{\theta} \mathcal{L}(\theta^*)^\top \Delta\theta + \frac{1}{2} \Delta\theta^\top H(\theta^*) \Delta\theta. \quad (14)$$

Note that since θ^* is a minima, the gradient at θ^* vanishes to zero ($\nabla_{\theta} \mathcal{L}(\theta^*) = \mathbf{0}$) thus we get:

$$\mathcal{L}(\theta^* + \Delta\theta) - \mathcal{L}(\theta^*) \approx \frac{1}{2} \Delta\theta^\top H(\theta^*) \Delta\theta. \quad (15)$$

Definition VI.6 (Parameter sensitivity). For a given perturbation $\Delta\theta$ of with $\|\Delta\theta\| = \epsilon$, we define **parameter sensitivity** $PS_{\mathcal{L},\epsilon}(\theta)$ as the maximal change in the loss function:

$$PS_{\mathcal{L},\epsilon}(\theta) = \max_{\|\Delta\theta\|=\epsilon} |\mathcal{L}(\theta + \Delta\theta) - \mathcal{L}(\theta)|. \quad (16)$$

Lemma VI.1. A sharp minima yields high parameter sensitivity.

Proof. Since most loss functions are C^2 [29], that is, the function, its derivatives, and its second derivatives are all continuous. Thus we assume the Hessian matrix to be a symmetric (interchangeable partial derivatives). Now consider eigen-decomposition of the Hessian $H(\theta^*) = U\Lambda U^\top$ where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ and $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ and U is a orthonormal matrix consisting of eigenbasis u_i s. Let $\mathbf{a} = U^\top \Delta\theta$ be a coordinates of $\Delta\theta$ on eigenbasis, namely $\Delta\theta = U\mathbf{a} = \sum_{i=1}^p a_i u_i$. Then, by substitution we get $\frac{1}{2} \Delta\theta^\top H(\theta^*) \Delta\theta = \frac{1}{2} \mathbf{a}^\top \Lambda \mathbf{a} = \frac{1}{2} \sum_{i=1}^p \lambda_i a_i^2 \leq \frac{1}{2} \lambda_{\max} \|\mathbf{a}\|^2 = \frac{1}{2} \lambda_{\max} \|\Delta\theta\|^2$. Thus we have $PS_{\mathcal{L},\epsilon}(\theta^*) \approx \frac{1}{2} S_{\mathcal{L}}(\theta^*) \epsilon^2$, giving the desired property that the parameter sensitivity increases as the minima becomes sharper. $\square$

By examining the change of decision values with fixed x , we indirectly prove that the shift of decision boundary depends on the parameter perturbation.

Lemma VI.2. High parameter sensitivity leads to high class decision value sensitivity.

Proof. Without loss of generality, let us assume that the perturbed parameter θ' from θ is mostly related to the logit of a class c , namely, $z_{x;\theta'} \approx z_{x;\theta}$ and $l_{c';\theta'}(x) \approx l_{c';\theta}(x)$ (at least one class should be related to the parameter perturbation, and instability of the feature space is out of the current scope). Since the softmax-based classification loss is locally Lipschitz w.r.t. logits [29],

$$|\mathcal{L}(\theta') - \mathcal{L}(\theta)| \leq L_c |l_{c;\theta'}(x) - l_{c;\theta}(x)| \quad (17)$$

holds for a Lipschitz constant L_c . Now consider the change of class decision values:

$$\begin{aligned} & |\text{CDV}_{c,c'}(x; \theta') - \text{CDV}_{c,c'}(x; \theta)| \\ &= |(l_{c;\theta'}(x) - l_{c';\theta'}(x)) - (l_{c;\theta}(x) - l_{c';\theta}(x))| \\ &= |(l_{c;\theta'}(x) - l_{c;\theta}(x)) - (l_{c';\theta'}(x) - l_{c';\theta}(x))| \\ &\approx |l_{c;\theta'}(x) - l_{c;\theta}(x)| \geq \frac{1}{L_c} |\mathcal{L}(\theta') - \mathcal{L}(\theta)|. \end{aligned} \quad (18)$$

Thus, with high parameter sensitivity and large change in the loss leads to significant shift in the decision values. $\square$

Lemma VI.3. *Sensitive class decision boundary implies smaller class decision regions in terms of perturbation robustness.*

Proof. Under class decision boundary shifts of magnitude up to $\epsilon > 0$ while perturbing the parameter to θ' from θ , the perturbation-safe decision region CDR shrinks as:

$$\hat{\text{CDR}}_c(\theta') = \{z_{x;\theta} \in \text{CDR}_c(\theta) \mid \text{MD}(z_{x;\theta}) > \epsilon\}, \quad (19)$$

while again, assuming that $z_{x;\theta'} \approx z_{x;\theta}$, and since $\epsilon > 0$, we get $\hat{\text{CDR}}_c(\theta') \subsetneq \text{CDR}_c(\theta)$. $\square$

Lemma VI.4. *Smaller decision region causes higher PRS ratio.*

Proof. Suppose there exist two feature points $z_{x;\theta}, z_{x';\theta}$ belonging to the same populated region. Reducing the safe margin of class decision regions moves at least one decision boundary closer to these points, causing the decision boundary to intersect the region containing the two points. Geometrically, this means at least one of the feature dimensions approaches zero, flipping its sign and thus changing the feature signature for at least one point. Therefore, the original populated region splits into multiple populated regions, directly increasing the PRS ratio. $\square$

Theorem VI.5. *Large-batch training is less robust to adversarial attacks.*

Proof. As shown in [28], large-batch training yields sharp minima. By Lemma VI.1, VI.2, VI.3, and VI.4, sharp minima implies parameter sensitivity and parameter sensitivity leads to sensitive decision boundary, where sensitive decision boundary creates smaller decision regions in terms of perturbation robustness, causing higher PRS ratio, which is negatively correlated with adversarial robustness. $\square$

C. Label Smoothing and PRS Ratio

Here, we prove explicitly that label smoothing reduces the PRS ratio. The proof sketch is as follows: First, we confirm that label smoothing bounds the decision value from above. Using this fact, we show that the rank of vectors formed by feature differences decreases, thereby demonstrating a reduction in the intra-class variance, yielding reduced PRS ratio. For simplicity, we redefine some concepts [30].

Definition VI.7 (Label smoothing). *For a K -classed classification problem, let a data label y be denote an one-hot label in $\{0, 1\}^K$. In **label smoothing**, we replace y by*

$$y^{(\epsilon)} = (1 - \epsilon)y + \frac{\epsilon}{K}\mathbf{1}, \quad (20)$$

for class c and smoothing factor $0 < \epsilon < 1$.

Lemma VI.6. *Label smoothing bounds decision value differences from above.*

Proof. Note that for class c and c' with $c' \neq c$, we have the followings: $y_c^{(\epsilon)} = 1 - \epsilon + \frac{\epsilon}{K}$, $y_{c'}^{(\epsilon)} = \frac{\epsilon}{K}$. Therefore, the decision value differences at optimality are explicitly bounded:

$$\begin{aligned} \text{CDV}_{c,c'}(x, \theta) &= l_{c;\theta}(x) - l_{c';\theta}(x) \\ &= \log(\exp(l_{c;\theta}(x) - l_{c';\theta}(x))) \\ &= \log\left(\frac{\exp(l_{c;\theta}(x)) / \sum_{c''} \exp(l_{c'';\theta}(x))}{\exp(l_{c';\theta}(x)) / \sum_{c''} \exp(l_{c'';\theta}(x))}\right) \\ &\Rightarrow \log\left(\frac{y_c^{(\epsilon)}}{y_{c'}^{(\epsilon)}}\right) = \log\left(\frac{1 - \epsilon + \frac{\epsilon}{K}}{\frac{\epsilon}{K}}\right) \\ &= \log\frac{K(1 - \epsilon) + \epsilon}{\epsilon} = B < \infty. \end{aligned} \quad (21)$$

$\square$

For simplicity, let us denote $w_{cc'} = w_{c;\theta} - w_{c';\theta}$ and $b_{cc'} = b_{c;\theta} - b_{c';\theta}$ so that $l_{c;\theta}(x) - l_{c';\theta}(x) = w_{cc'}z + b_{cc'}$.

Lemma VI.7. *Bounded decision value difference reduces intra-class variance.*

Proof. For features $z^{(a)}, z^{(b)}$ of two samples $x^{(a)}, x^{(b)}$ from the same class c , since we have $\text{CDV}(c, c')(x, \theta) \rightarrow B$, we get:

$$\begin{aligned} & \text{CDV}(c, c')(x^{(a)}, \theta) - \text{CDV}(c, c')(x^{(b)}, \theta) \\ &= (w_{cc'}z^{(a)} + b_{cc'}) - (w_{cc'}z^{(b)} + b_{cc'}) \\ &= w_{cc'}^\top(z^{(a)} - z^{(b)}) \rightarrow B - B = 0, \end{aligned} \quad (22)$$

for $\forall c, c'$. Let the matrix of class differences W_c be

$$W_c = [w_{c1}, w_{c2}, \dots, w_{cc-1}, w_{cc+1}, \dots, w_{cC}] \in \mathbb{R}^{d \times (C-1)}. \quad (23)$$

Then, the feature differences lie in the null space $\mathcal{V} = \text{Null}(W_c^\top)$. Namely, near optimality, the differences satisfy: $z^{(a)} - z^{(b)} \in \mathcal{V}$. Using the rank-nullity theorem, the null-space dimension satisfies: $\dim(\mathcal{V}) = d - \text{rank}(W_c^\top)$. Typically, we have $\text{rank}(W_c^\top) = K - 1 \geq 1$, and thus: $\dim(\mathcal{V}) = d - (K - 1) < d$. The intra-class covariance matrix Σ_c can be represented as: $\Sigma_c = \mathbb{E}_{z|c}[(z - \mu_c)(z - \mu_c)^\top]$, where $\mu_c = \mathbb{E}_{z|c}[z]$. Note that $W_c^\top(z - \mu_c) = 0$ for all z in class c . Since covariance matrix is always symmetric, we apply eigendecomposition to the covariance matrix as follows: $\Sigma_y = U\Lambda^\top U$ where $U = [u_1, u_2, \dots, u_d]$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$. Note that the range of Σ_c is always linear combination of $(z - \mu_c)(z - \mu_c)^\top v$ for some vector v and expected z s, so by the fact that $W_c^\top(z - \mu_c)(z - \mu_c)^\top v = 0(z - \mu_c)^\top v = 0$, we

have $\text{range}(\Sigma_c) \subset \mathcal{V}$. Therefore, we get $\text{rank}(\Sigma_c) \leq \dim(\mathcal{V}) = d - (K - 1)$, so at least $K - 1$ eigenvalues of Σ_c must be exactly zero: $\lambda_{d-(K-2)} = \lambda_{d-(K-3)} = \dots = \lambda_d = 0$. Therefore, the trace of covariance is strictly bounded by the reduced dimensionality: $\text{Tr}(\Sigma_c) = \sum_{i=1}^d \lambda_i = \sum_{i=1}^{d-(K-1)} \lambda_i$. Without label smoothing however, no eigenvalues would necessarily vanish, potentially yielding larger variance: $\text{Tr}(\Sigma_c^{(\epsilon=0)}) = \sum_{i=1}^d \lambda_i^{(\epsilon=0)} \geq \sum_{i=1}^{d-(K-1)} \lambda_i$, assuming that $\lambda_i^{(\epsilon=0)} \approx \lambda_i$ for $1 \leq i \leq d - (K - 1)$. $\square$

Theorem VI.8. *Label smoothing lowers PRS ratio.*

Proof. From Lemma VI.6, VI.7, we get lower $\text{Var}(z|c)$, inducing higher chance of $P(\text{PR}(z^{(a)}) = \text{PR}(z^{(b)}))$, where PR denotes the occupying populated region, at least giving larger the number of the same signatures, thus lowering the PRS ratio. $\square$

VII. CONCLUSION

We analyze the geometrical properties of DNNs affecting adversarial robustness and introduce the Populated Region Set (PRS) to establish this relationship. Experiments show that PRS correlates with robustness: (1) Networks with low PRS ratios are more robust to gradient-based attacks and exhibit higher parameter parallelism in the final layer. (2) Low-PRS networks include more test samples within training regions; these included samples show higher robustness. (3) A PRS regularizer improves robustness without adversarial examples, and adversarial training further reduces PRS ratio while enhancing robust accuracy. Our work provides a geometrical interpretation of robustness through decision regions, and we expect PRS to contribute to improving DNN robustness.

ACKNOWLEDGMENT

This work was made in collaboration with Korea Advanced Institute of Science & Technology and INEEJI Co., Ltd. Also, this work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No.RS-2022-II220984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation].

REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [2] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
- [3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [4] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.
- [5] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, "Learning with a strong adversary," *arXiv preprint arXiv:1511.03034*, 2015.
- [6] D. Jakubovitz and R. Giryas, "Improving dnn robustness to adversarial attacks using jacobian regularization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 514–529.
- [7] 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.
- [8] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.
- [9] Z. Zhong, Y. Tian, and B. Ray, "Understanding local robustness of deep neural networks under natural variations," in *International Conference on Fundamental Approaches to Software Engineering*. Springer International Publishing Cham, 2021, pp. 313–337.
- [10] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," *Advances in neural information processing systems*, vol. 27, 2014.
- [11] H. Xiong, L. Huang, M. Yu, L. Liu, F. Zhu, and L. Shao, "On the number of linear regions of convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 514–10 523.
- [12] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Classification regions of deep neural networks," *arXiv preprint arXiv:1705.09552*, 2017.
- [13] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.
- [14] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 1277–1294.
- [15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
- [16] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, 2018.
- [17] Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. W. Mahoney, "Hessian-based analysis of large batch training and robustness to adversaries," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [18] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Artificial intelligence and statistics*. PMLR, 2015, pp. 192–204.
- [19] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," *Advances in neural information processing systems*, vol. 27, 2014.
- [20] N. Lei, D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S.-T. Yau, and X. Gu, "A geometric understanding of deep learning," *Engineering*, vol. 6, no. 3, pp. 361–374, 2020.
- [21] Y. Yang, R. Khanna, Y. Yu, A. Gholami, K. Keutzer, J. E. Gonzalez, K. Ramchandran, and M. W. Mahoney, "Boundary thickness and robustness in learning models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6223–6234, 2020.
- [22] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, "Robustness via curvature regularization, and vice versa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9078–9086.
- [23] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Empirical study of the topology and geometry of deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3762–3770.
- [24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [26] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.
- [27] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [28] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.
- [29] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.
- [30] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.