MULTI-STATIONARY POINT LOSSES FOR ROBUST MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify that cross-entropy (CE) loss does not guarantee robust boundary for neural networks. The reason is that CE loss has only one asymptotic stationary point. It stops pushing the boundary forward as long as the sample is correctly classified, which left the boundary right next to the samples. A robust boundary should be kept in the middle of samples from different classes, thus maximizing the margins from the boundary to the samples. In this paper, we propose a family of new losses, called multi-stationary point (MS) loss, which introduce additional stationary points beyond the asymptotic stationary point. We prove that robust boundary can be guaranteed by MS loss without losing much accuracy. With MS loss, bigger perturbations are required to generate adversarial examples. We demonstrate that robustness is improved under a variety of adversarial attacks by applying MS loss. Moreover, robust boundary learned by MS loss also performs well on imbalanced datasets. Finally, we modified other losses into two-stationary-point forms, and improved model robustness is observed.

1 INTRODUCTION

Cross-entropy(CE) is the loss function which commonly used in image classification tasks, while generalized CE loss was also used in noisy labels, sound event classifications, and econometric models(Zhang & Sabuncu (2018), Deng et al. (2021), Heckelei et al. (2008), Kazemdehdashti et al. (2018), Kurian et al. (2021)), etc. However, both (Jacobsen et al. (2018)) and (Nar et al. (2019)) draw the same conclusion that CE loss can cause excessive invariance to predict features, i.e., classifiers make decisions relying on only a few highly predictive features. Specifically, CE loss will not continue to optimize the boundary once a data point has been classified correctly based on a few features. We further empirically show that CE loss will only sharpen the inference function to achieve a lower loss rather than moving the decision boundary to a 'better' position once samples are classified correctly. For 'better' boundaries, we mean that the boundaries can lead robust predictions or considering more features into the account, etc.

Neural networks(NNs) are applying to various fields with security requirements, robustness has become an important property (Szegedy et al. (2013)) and (Kurakin et al. (2018b)), however, suggests that the deep neural network is vulnerable to adversarial examples. Adversarial examples are images that added subtle changes artificially on which the neural network makes mistakes but human eyes cannot detect the changes. Specifically, adversarial examples are obtained in the neighborhood of data points in training set, generally L_∞ norm neighborhood. It is clearly that the neighborhood of the training sample crosses the classification boundary and confuses the model. That is to say the reason why neural network is vulnerable to adversarial examples is that decision boundary is too closed to data points, resulting in low robustness. We call the decision boundary which is too closed to data points *near-boundary*. While, models with robust boundary maintains the prediction for an open set in the neighborhood of each training sample.

One of the directions to obtain robust boundary is designing a new loss to enforce a large margin between training instances and decision (Elsayed et al. (2018), Matyasko & Chau (2017)). (Sun et al. (2016)) introduced margin-based penalties to the objective of training DNNs, motivated by theoretical analyses from the perspective of the margin bound.

Given the excessive invariance property of CE loss, we consider *near-boundary* is caused by CE loss. Specially, we think the reason why CE loss dose not optimize to robust boundary is because

CE loss has only one asymptotic stationary point, which result in small optimization range and thus CE loss cannot learn a robust boundary. Thus, we propose a family of new losses with multiple stationary points to overcome the drawback.

Our contributions:

- We propose a family of new losses, called multi-stationary point(MS) loss, which has two stationary points and we provide experimental evidence that MS loss learn a more robust boundary.
- We theoretically show that if a neural network model is trained by minimizing CE loss via gradient decent algorithm, the model dose not always convergence to robust boundary, but MS loss will.
- Several experiments of multi structures on multi datasets with popular white box attack methods has been performed. The result shows that MS loss can greatly enhance the network robustness without losing much precision. At the same time, MS loss also works well in imbalanced datasets.
- We also modified other losses into multi-stationary-point form, and improved model robustness

2 RELATED WORK

Stationary point is an very important nature of function, and it has practical significance in many applications. For example, The optimization of transition-state structures (TSs) is key to the understanding of mechanisms and kinetics of chemical reactions on a computational basis. Transition states are defined as first-order saddle-point structures located on the minimum (reaction) energy path between reactants and products (Bergeler et al. (2015), Van de Vijver & Zádor (2020)). In radar detection, (Park et al. (2019)) proposed SPC technique to make the phase noise of the leakage concentrated on the stationary point of the cosine function, mitigating the Heterodyne FMCW Radar for small drone detection leakage.

Neural network adversarial attacks has become one of the most important test for the robustness of network, notably the white box attacks the most compelling. Considering adversarial attack methods, white box attacks require full access of the model, including the structure and parameters of each layer. Advanced white box attacks are Fast Gradient Sign Method (FGSM) (Goodfellow et al. (2014)), Projected Gradient Descent(PGD)(Madry et al. (2017)), the Carlini and Wagner (C&W)(Carlini & Wagner (2017)), etc. (Schmidt et al. (2018)) thinks that commonly used datasets can provide good accuracy, but not enough to provide a good robustness. Considering the difficulties in practice of training robust classifiers, they further assume that, the difficulties may be the lack of training samples (Krizhevsky (2009)).

In order to improve the robustness of the network, researchers have proposed many defensive methods. A popular method is adversarial training. Adversarial training (Madry et al. (2017), Zhang et al. (2019), Goodfellow et al. (2014), Tramèr et al. (2017)) is a training method, which generates adversarial samples by attacking methods as a way of data augmentation. However, the cost is very expensive, and this method would reduce the performance of neural networks comparing to be trained with the original dataset. There are also many other defense methods, such as attention-based method (Taghanaki et al. (2019), Zoran et al. (2020), Vaishnavi et al. (2020)) and regularization methods (Sokolić et al. (2017), Cisse et al. (2017)), (Kurakin et al. (2018a), Pang et al. (2019b)). The defense methods make changes to the neural network, but does not guarantee universality. Meanwhile, with extra defensive components introduced, the training cost will increase as well.

One of the more effective and universal methods is to design a new loss function to improve the robustness. Several kinds of loss or regularization based robust training method have been put forward. By applying potential characteristics regularization, convolutional neural networks(CNNs) are encouraged to study features between class separability and compactness within the class list(Pang et al. (2019a), Mustafa et al. (2019)). (Pang et al. (2019a)) proposed Max-Mahalanobis center(MMC) loss, which studies identification features and looking for high density feature area. They first calculate each class Max-Mahalanobis center(Pang et al. (2018)), then use the center loss to

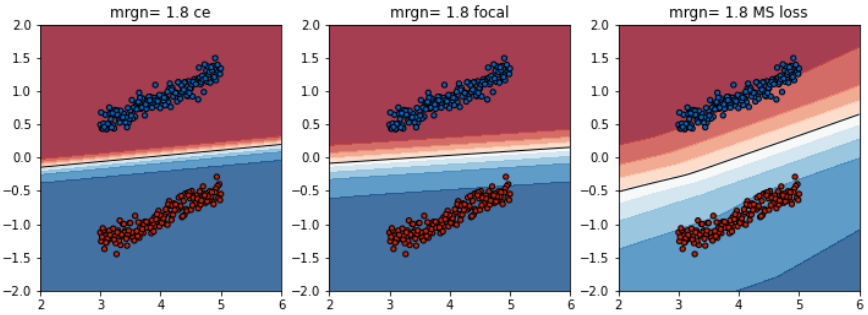

Figure 1: Decision boundary learned by CE loss, focal loss and MS loss, respectively. The classification confidence of the network are divided and visualized by contour lines. The lighter the area, the lower the confidence.

encourage features crowded around the center. But to avoid degradation, the preset center u_y^* for MMC is untrainable. The choice of u_y^* is crucial, which determined features crowded ways, but there is no guidelines for better choice.

3 TWO-STATIONARY-POINT LOSS

3.1 A TOY EXAMPLE OF CLOSED BOUNDARIES

We demonstrate the near-boundaries of CE loss and even focal loss in a toy dataset. We create a two-classes datasets(samples in blue and red) distributed as line segments. The samples are linearly separable with margin equals to 1.8 vertically. A two-layers fully-connected neural network is trained to classify the samples. As shown in the left most subplot of Figure 1, if we use the CE loss, the decision boundary can separate the samples well. However, the decision boundary doesn't maximize the margins between the boundary and samples. Moreover, the area of the lower classification confidence is narrow, since the CE loss keeps pushing the prediction probabilities of the sample to 1 to reduce the loss. From another perspective of the view, narrow lower confidence means the inference function is very sharp at the decision boundary. The toy experiment demonstrates a property of the CE loss, i.e., CE loss stops optimizing the boundary once all the samples are classified correctly, but sharpen the decision boundary to reduce the loss. This indicates that other forces are needed to push CE loss to continue to optimize the boundary.

Compared with CE loss, focal loss(Lin et al. (2017)) is designed for foreground and background imbalance during training process, namely the positive and negative samples imbalance. It reduced the penalty once the sample is correctly classified. As shown in the second subplot in Figure 1, since correctly classified samples has less penalty, the lower confidence area is much larger than the neural network trained under CE loss. However, the decision boundary is still a *near-boundary*.

A more robust decision boundary should maximize the margins from the boundary to samples from both classes. In the third subplot of Figure 1, the decision boundary trained under proposed MS loss is parallel to the line segments which are the samples distributed.

3.2 CE LOSS DOES NOT GUARANTEE TO CONVERGENCE TO A ROBUST BOUNDARY

Why dosen't CE loss convergence to the robust boundary? Different from other works, we analyze the drawbacks of CE loss from optimization. We found that even the model has learned a *near-boundary*, just like the Figure 1, CE loss will not lead the model to find better parameters since there are more than one optimal boundaries for the CE loss.

Notation: Consider the classification problem, let $(X, y) \in \mathcal{D}$ be the training set, $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{N \times d}$ is training sample, N is the size of training set $\mathcal{D}$, $y \in \{1, 2, ..., K\}$ are corresponding targets. Let $f(\theta, x)$ represent a neural network model with θ being the parame-

ters. Denote $o_k(k = 1, ..., K)$ the output of the neural network before $Softmax$, $\hat{y}$ is the predict of the network after $Softmax$. Given a sample x, the output of the network after $Softmax$ is

$$\xi = p(\hat{y} = s|x) = \frac{e^{o_s}}{\sum_{k=1}^{K} e^{o_k}}, \quad s = 1, ..., K \tag{1}$$

Theorem 1 *(CE loss convergence) If a neural network $f(\theta, x)$ is trained to minimize CE loss*

$$L(\theta, x, y) = -ylog(\xi) - (1 - y)log(1 - \xi) \tag{2}$$

and optimized by gradient-based algorithm

$$x^{k+1} = x^k + \alpha \bigtriangledown_\theta L(\theta, x, y) \tag{3}$$

Then for any boundary CE loss has learned, there is always another boundary whose loss value less than the former, where α is learning rate.

In **Theorem 1**, we demonstrate that CE loss makes the network continue to optimize even the network has arrived at the robust boundary. At last, the NNs trained by CE loss deviate the robust boundary, if the training epoches is not appropriate, and that is usually true. That is to say NNs' robust boundary can not be guaranteed by CE loss.

3.3 MS LOSS

As the drawbacks of CE loss portrayed in previous subsection, we think that the reason is CE loss has only one asymptotic stationary point, the optimization range makes it impossible to optimize the network to the best position. CE loss will not continue to optimize boundary location once a data point has classified correctly. When there comes a new data point, CE loss will make a higher score(i.e sharpen the decision boundary), like the Figure 1 dark area, rather than move the decision boundary, if the point is classified correctly. Focal loss is designed for positive and negative feature imbalance, but as the Figure 1 shows, focal loss can also not lead to an optimal boundary. And from Figure 2, focal loss has only one asymptotic stationary point.

Considering one asymptotic stationary point, we propose a family of new losses which has two stationary points, called multi-stationary point(MS) loss. Take the focal loss as an example, the MS loss is

$$FL(o_k) = -\alpha(1 - y)^\gamma log(\xi) + \eta(||\xi||^2 + (||1 - \xi||)^2) \tag{4}$$

o_k is network output before $Softmax$, ξ is the network output after $Softmax$, η is stationary-point coefficient. From Figure 1, we can see that MS loss learned a better boundary than CE loss and focal loss, which is parallel to the datasets in almost everywhere. Intuitively, in line segments datasets paralleled boundary is the boundary human desire, which is similar to the dataset distribution.

The gradient of focal loss about element before $Softmax$ is

$$FL^{'}(o_k) = -\alpha(1 - \xi)^\gamma(\gamma\xi log(\xi) - 1 + \xi) + \eta(2\xi(1 - \xi) + -2(1 - \xi)^2) \tag{5}$$

As Figure2, from the left side image, we can see that once a data point was classified correctly, CE loss will assign a high score to the point, even though the point is closed to the decision boundary. The right figure shows that MS loss has two stationary points, and others have only one asymptotic stationary point. Then we states that MS loss is conducive to NNs' convergence.

Theorem 2 *(MS loss convergence) With same assumption of **Theorem 1** except training loss, MS loss always makes the model $f(\theta, x)$ convergence to robust boundary.*

With above theorem, we prove that MS loss can lead model to convergence to an robust boundary, while CE loss cannot. We show the validity of MS loss theoretically, and from Figure 1, MS loss network has larger margin, which may lead to more stronger robustness to adversarial attacks. Following the advantages of MS are illustrated by experiments.

Proof of **Theorem 1**, **Theorem 2** can be found in appendix.

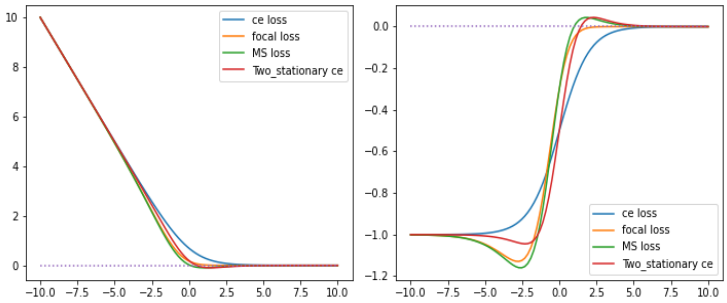

Figure 2: Loss curve of different losses. Left subimage is loss function image, right subimage is corresponding gradient funciont image.

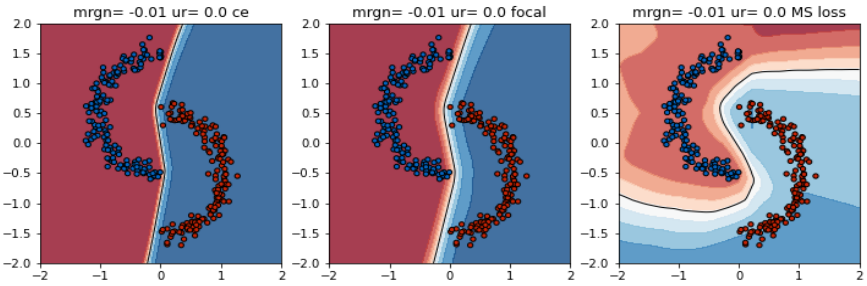

Figure 3: Diagram illustrate the loss performance in linear unseparable classification task. These are the decision boundary learned by CE Loss, focal loss, MS loss, respectively.

4 EXPERIMENT

In this section, we empirically demonstrate several attractive merits of applying MS loss. We firstly validate superior performance of MS loss on linear separable and unseparable datasets. Next, we perform experiments on MNIST, CIFAR10 datasets under white box attacks to state robustness improvement. Finally, we perform a experiment on imbalanced datasets. The result shows that MS loss makes up for the defects of imbalanced datasets to some extent.

4.1 PERFORMANCE ON LINEAR UNSEPARABLE DATA SETS

In Figure 1, we show that MS loss do a great work on linear separable datasets. In this subsection, we perform an classification task in linear unseparable datasets. As Figure 3 shows, we find that decision boundary learned by MS loss is closer to the middle location (robust boundary) than CE loss and focal loss.

With robust boundary, it needs bigger perturbation radius to get adversarial examples. Intuitively, MS loss may be more robust against adversarial attacks. We conduct several adversarial attack experiments in the following.

4.2 ROBUSTNESS AGAINST ATTACKS

(Schmidt et al. (2018)) and others works guess the current datasets in CNNs model can't learn robustness of the model. We think that it is CE loss learned a *near-boundary*, which is closed to training data points, making the model low robustness.

We consider several classic network structures, ResNet-18, ResNet-34, GoogleNet, DenseNet-121 in pytorch on MNIST, CIFAR-10 under the white-attack FGSM, BIM, PGD and UPGD. We apply Adam optimizer with initial learning rate is 0.001 and training for 50 epoches for MNIST and 200 epoches for CIFAR-10. We save the most accurate model. For FGSM, the max perturbation is

Table 1: Classification accuracy(%) for different networks and different losses under several adversarial attack methods.

model	ResNet-18		ResNet-34		GoogleNet		DenseNet-121	
	CE loss	MS loss	CE loss	MS loss	CE loss	MS loss	CE loss	MS loss
FGSM	72.60	**90.84**	21.0	**86.17**	63.50	**72.77**	52.98	**81.19**
BIM	13.70	**69.8**	2.0	**66.34**	11.75	**34.46**	4.57	**61.52**
PGD	0.14	**37.24**	0.02	**40.84**	3.30	3.89	0.81	**27.1**
UPGD	0.15	**41.80**	0.01	**43.88**	3.40	4.4	0.85	**35.3**

Table 2: Adversarial accuracy(%) comparison with other losses.

Perturbation	Attack	MS loss	MMC	SCE	Centerloss	L-DM
eps=8/255	PGD_{10}	**47.8**	36.0	3.7	4.4	19.8
	PGD_{50}	**39.7**	24.8	3.6	4.3	4.9
eps=16/255	PGD_{10}	**46.7**	25.2	2.9	3.1	11.0
	PGD_{50}	**26.7**	17.5	2.6	2.9	2.8

$\epsilon = 0.007$. For BIM $\epsilon = 8/255$. For PGD, $\epsilon = 8/255$, and steps = 40. For UPGD, $\epsilon = 8/255$, $\alpha = 2/255$ and steps = 40. For MNIST dataset, MS loss get an higher accurate than other losses, but the current methods have nearly arrived at perfect accurate. So we omit the MNIST dataset result. The result on CIFAR-10 was shown in Table 1. We can see that MS loss has improved the accurate of all structures and attack methods for at least 20% even more than 65%.

We compared our loss with other baseline MMC(Pang et al. (2019a)), SCE(He et al. (2016)), Center loss(Wen et al. (2016)), L-DM(Wan et al. (2018)) under PGD attack with $\epsilon = 8/255, 16/255$ and steps=10, 50 and untarget. Since the untrainable parameter μ_y^* is preset, and the code is unpublished, the data about baseline is following the original papers. In Table 2, MS loss get the highest adversarial accurate.

4.3 DATA ROBUSTNESS

Imbalanced data is a classic problem in classification tasks, which requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation (He & Garcia (2009)). In short, imbalanced data problem is the

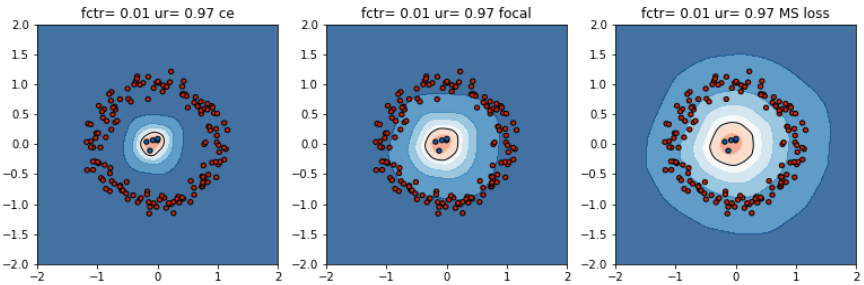

Figure 4: Decision boundary learned by CE loss, focal loss and MS loss, respectively on imbalanced dataset which is lacking 97% data points.

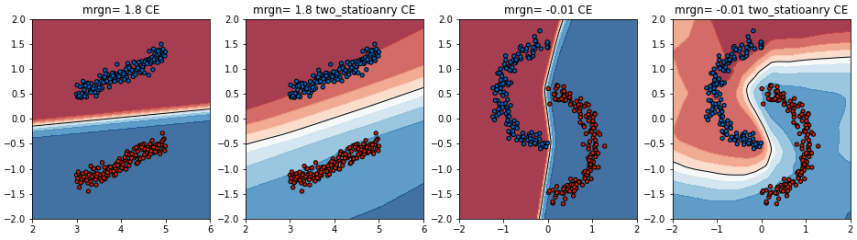

Figure 5: Two stationary point CE performance, stationary-point coefficient is 1.5.

number of samples under each category varies greatly in datasets. Imbalanced data results in high training accuracy but low testing accuracy of NNs model. In this case, it is obvious that the classifier is invalid. So it is very important to improve imbalanced data problem.

Due to epistemic uncertainty, we may never collect a complete data sets, and the work of collecting data could be costly. It would be a cost effective way to improve the robustness problem through the loss function. We conduct a experiment on a two layers of full-connection neural network, to compare the boundary when the network trained from the data in lacking 10%, 30%, 50% and 90% respectively. The experiment results lacking 97% are exhibited in Figure 4, other result can be found in appendix.

From the Figure 4, we can see that decision boundary CE loss learned is far away from the larger volume data, which means the large classification probability of the outer data, and small classification probability the inner data. Compared with CE loss, classification boundary learned by MS loss is closer to the medium location. This reduces the effect of category imbalance on the classification probability.

5 OTHER TWO-STATIONARY-POINT LOSS

In this section, we modify other loss to two-stationary-point form to illustrate the generality of multi-stationary point losses family.

We consider CE loss

$$p_t = \begin{cases} p, & if\ y = 1 \\ 1 - p, & otherwise \end{cases} \tag{6}$$

After $Softmax$, adding stationary-point term to CE loss is

$$L = -log(\xi) + \eta(||\xi||^2 + (||1 - \xi||)^2) \tag{7}$$

Similarly, the gradient of CE loss is

$$CE^{'}(\xi) = \xi - 1 + \eta(2\xi(1 - \xi) + -2(1 - \xi)^2 \tag{8}$$

η is stationary-point coefficient. One can prove that MS loss we proposed is equivalent to (Pezeshki et al. (2021)). But it is not always convergence for any stationary-point coefficient. Following theorem claims the coefficient selection rule.

Theorem 3 *(Stationary-Point Coefficient) Two-stationary-point CE loss learn a robust model, if the stationary-point coefficient* $\eta > 0.5$.

Proof of **theorem 3** can be found in appendix.

From the **theorem 3**, we can know that two-stationary-point CE loss convergence only when stationary-point coefficient $\eta > 0.5$. From Figure 2 we can see that two-stationary-point CE has other stationary point apart from one asymptotic stationary point. As Figure 5 shows, two-stationary-point CE loss learns a robust boundary.

6 CONCLUSION

In this paper, we proposed a family of new losses, called multi-stationary point (MS) loss, which introduce additional stationary points beyond the asymptotic stationary point. Firstly, we states that neural network trained by CE loss cannot guarantee robust decision boundary theoretically. Next, we prove that robust boundary can be guaranteed by MS loss without losing much accuracy. And an experiment of two-layers neural network on toy dataset validated MS loss's effectiveness. Then we conduct several experiments to show that proposed losses improve the model robustness by adversarial attack under several different attack methods and perform well on imbalanced data. Finally, we illustrate that CE loss with effective coefficient can be modified to two stationary points form, at the same time two-stationary-point CE loss can learn a robust boundary. In the future, multi-stationary-point losses's application and its intrinsic nature in high dimension, the training mechanism leading from MS loss deserves further study.

REFERENCES

Maike Bergeler, Carmen Herrmann, and Markus Reiher. Mode-tracking based stationary-point optimization. *Journal of Computational Chemistry*, 36(19):1429–1438, 2015. doi: https://doi.org/10.1002/jcc.23958. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.23958.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017. doi: 10.1109/SP.2017.49.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 854–863. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/cisse17a.html.

Jun Deng, Chunhui Gao, Qian Feng, Xinzhou Xu, and Zhaopeng Chen. Adaptive generalized cross-entropy loss for sound event classification with noisy labels. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 256–260. IEEE, 2021.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/42998cf32d552343bc8e460416382dca-Paper.pdf.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 630–645, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Thomas Heckelei, Ronald C Mittelhammer, and Torbjorn Jansson. A bayesian alternative to generalized cross entropy solutions for underdetermined econometric models. Technical report, 2008.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *Proceedings of the 7th International Conference on Learning Representations (ICLR), 2019*, November 2018.

Abolfazl Kazemdehdashti, Mohammad Mohammadi, and Ali Reaz Seifi. The generalized cross-entropy method in probabilistic optimal power flow. *IEEE Transactions on Power Systems*, 33(5):5738–5748, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition. In Sergio Escalera and Markus Weimer (eds.), *The NIPS '17 Competition: Building Intelligent Systems*, pp. 195–231, Cham, 2018a. Springer International Publishing. ISBN 978-3-319-94042-7.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018b.

Nikhil Cherian Kurian, Pragati Shuddhodhan Meshram, Abhijeet Patil, Sunil Patel, and Amit Sethi. Sample specific generalized cross entropy for robust histology image classification. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1934–1938. IEEE, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

Alexander Matyasko and Lap-Pui Chau. Margin maximization for robust classification using deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 300–307. IEEE, 2017.

Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Kamil Nar, Orhan Ocal, S. Shankar Sastry, and Kannan Ramchandran. Cross-entropy loss and low-rank features have responsibility for adversarial examples. January 2019.

Tianyu Pang, Chao Du, and Jun Zhu. Max-Mahalanobis linear discriminant analysis networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4016–4025. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/pang18a.html.

Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax cross-entropy loss for adversarial robustness, 2019a.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4970–4979. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/pang19a.html.

Junhyeong Park, Seungwoon Park, Do-Hoon Kim, and Seong-Ook Park. Leakage mitigation in heterodyne fmcw radar for small drone detection with stationary point concentration technique. *IEEE Transactions on Microwave Theory and Techniques*, 67(3):1221–1232, 2019. doi: 10.1109/TMTT.2018.2889045.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/0987b8b338d6c90bbedd8631bc499221-Paper.pdf.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/f708f064faaf32a43e4d3c784e6af9ea-Paper.pdf.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017. doi: 10.1109/TSP.2017.2708039.

Shizhao Sun, Wei Chen, Liwei Wang, Xiaoguang Liu, and Tie-Yan Liu. On the depth of deep neural networks: A theoretical view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30 (1), mar 2016. doi: 10.1609/aaai.v30i1.10243.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2013.

Saeid Asgari Taghanaki, Kumar Abhishek, Shekoofeh Azizi, and Ghassan Hamarneh. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses, 2017.

Pratik Vaishnavi, Tianji Cong, Kevin Eykholt, Atul Prakash, and Amir Rahmati. Can attention masks improve adversarial robustness? In Onn Shehory, Eitan Farchi, and Guy Barash (eds.), *Engineering Dependable and Secure Machine Learning Systems*, pp. 14–22, Cham, 2020. Springer International Publishing. ISBN 978-3-030-62144-5.

Ruben Van de Vijver and Judit Zádor. Kinbot: Automated stationary point search on potential energy surfaces. *Computer Physics Communications*, 248:106947, 2020. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2019.106947. URL https://www.sciencedirect.com/science/article/pii/S0010465519302978.

Weitao Wan, Yuanyi Zhong, Tianpeng Li, and Jiansheng Chen. Rethinking feature distribution for loss functions in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 499–515, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19p.html.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

A APPENDIX

A.1 PROOF OF THEOREMS

Theorem 1 *(CE loss convergence) If a neural network $f(\theta, x)$ is trained to minimize CE loss*

$$L(\theta, x, y) = -ylog(\xi) - (1-y)log(1-\xi) \tag{9}$$

and optimized by gradient-based algorithm

$$x^{k+1} = x^k + \alpha \bigtriangledown_\theta L(\theta, x, y) \tag{10}$$

Then for any boundary CE loss has learned, there is always another boundary whose loss value less than the former, where α is learning rate.

Proof: We prove **theorem 1** in 1 dimension, it is same in multi dimension. Assuming we have two data points, -1, 1, then we know that the decision boundary is x=0. Now we assume network trained by CE loss got the decision boundary in x=0 in some weights w_η in the network training process, in which CE loss value is η. CE loss is

$$L(p) = -log(p)$$

the derivation is

$$L^{'}(p) = -\frac{1}{p}$$

and

$$L^{'}(p) < 0$$

is always true. So we assert that there must be another decision boundary got by current state. Assume another decision boundary corresponding weights is w_ξ, the CE loss value is ξ. Then we have

$$w_\xi = w_\eta + \frac{\alpha}{p} \tag{11}$$

in which α is learning rate. Assume $x > 0$ is positive sample, negative sample is also true.

$$\eta = -log\frac{1}{1+e^{1-2(w_\eta x+b)}} \tag{12}$$

$$\xi = -log\frac{1}{1+e^{1-2(w_\eta x+b)-2\frac{\alpha x}{p}}} \tag{13}$$

Due to

$$exp^{1-2(w_\eta x+b)} > e^{1-2(w_\eta x+b)-2\frac{\alpha x}{p}} \tag{14}$$

therefore

$$-log\frac{1}{1+e^{1-2(w_\eta x+b)}} > -log\frac{1}{1+e^{1-2(w_\eta x+b)-2\frac{\alpha x}{p}}} \tag{15}$$

so

$$\eta > \xi \tag{16}$$

Theorem 2 *(MS loss convergence) With same assumption of **Theorem 1** except training loss, MS loss makes the model $f(\theta, x)$ convergence to optimal boundary.*

Proof: Consider one-dimension case, insert 1,1, -1,-1 into MS loss. Let

$$\xi = \frac{1}{1+e^{1-2(w+b)}} \tag{17}$$

$$\gamma = \frac{1}{1 + e^{1-2(b-w)}} \tag{18}$$

η is two-stationary coefficient.

MS loss is

$$-log\xi + \eta\xi^2 + \eta(1-\xi)^2 - log(1-\gamma) + \eta\gamma^2 + \eta(1-\gamma)^2 = L \tag{19}$$

As follow, we separate L to two parts,

$$\begin{aligned} L1 &:= -log\xi + \eta\xi^2 + \eta(1-\xi)^2 \\ L2 &:= -log(1-\gamma) + \eta\gamma^2 + \eta(1-\gamma) * 2 \end{aligned} \tag{20}$$

differentiate them separately,

$$\begin{aligned} L1_w^{'} &= -\frac{1}{\xi}\xi_w^{'} + 2\eta\xi\xi_w^{'} + 2\eta(1-\xi)(-1)\xi_w^{'} \\ &= (-2)(1-\xi)[-1 + 2\eta\xi^2 - 2\eta\xi(1-\xi)] \end{aligned} \tag{21}$$

$$\begin{aligned} L2_w^{'} &= -\frac{1}{1-\gamma}(-1)\gamma_w^{'} + 2\eta\gamma\gamma_w^{'} + 2\eta(1-\gamma) * (-1)\gamma_w^{'} \\ &= 2\gamma(1 + 2\eta\gamma(1-\gamma) - 2\eta(1-\gamma)^2) \end{aligned} \tag{22}$$

$$\begin{aligned} L1_b^{'} &= -\frac{1}{p}\xi_b^{'} + 2\eta\xi\xi_b^{'} + 2\eta(1-\xi)(-1)\xi_b^{'} \\ &= (-2)(1-\xi)[-1 + 2\eta\xi^2 - 2\eta\xi(1-\xi)] \end{aligned} \tag{23}$$

$$\begin{aligned} L2_b^{'} &= -\frac{1}{1-\gamma}(-1)\gamma_b^{'} + 2\eta\gamma\gamma_b^{'} + 2\eta(1-\gamma) * (-1)\gamma_b^{'} \\ &= -2\gamma(1 + 2\eta\gamma(1-\gamma) - 2\eta(1-\gamma)^2) \end{aligned} \tag{24}$$

From formula $(21), (22)$, differentiate w we have

$$(-2)(1-\xi)[-1 + 2\eta\xi^2 - 2\eta\xi(1-\xi)] + 2\gamma(1 + 2\eta\gamma(1-\gamma) - 2\eta(1-\gamma)^2) = 0 \tag{25}$$

Similarly, differentiate b we have

$$(-2)(1-\xi)[-1 + 2\eta\xi^2 - 2\eta\xi(1-\xi)] - 2\gamma(1 + 2\eta\gamma(1-\gamma) - 2\eta(1-\gamma)^2) = 0 \tag{26}$$

incorporate formulas $(equation\ 25), (equation\ 26)$, we get

$$\xi_1 = \frac{\sqrt{\eta(4+\eta)} + \eta}{4\eta} \tag{27}$$

$$\xi_2 = \eta - \frac{\sqrt{\eta(4+\eta)}}{4\eta}, (del) \tag{28}$$

$$\gamma_1 = 3\eta - \frac{\sqrt{\eta(4+\eta)}}{4\eta} \tag{29}$$

$$\gamma_2 = \frac{\sqrt{\eta(4+\eta)} + 3\eta}{4\eta} \tag{30}$$

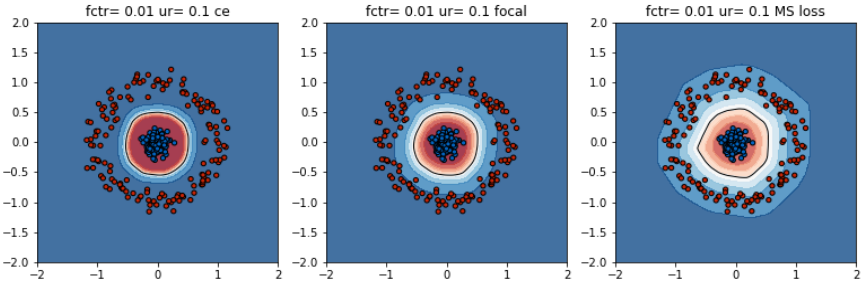

Figure 6: Imbalanced dataset in lacking 10% data points.

then incorporate formula $equation$ 27, $equation$ 28we get $b = 0.5$, so the decision boundary contains point $x = 0$, so is the center of the dataset.

Theorem 3 *(Stationary-Point Coefficient) Improved CE loss learn a robust model, if the stationary-point coefficient* $\eta > 0.5$.

Proof: We consider the regularized CE loss, its formula is

$$L = -log(\xi) + \eta||\xi - 0.5||^2 \tag{31}$$

the derivation of weight is

$$L^{'} = -(1-\xi) * 2x^{\top} + 2\eta * 2x^{\top} * (1-\xi) + 2\eta(1-\xi)^2 * 2x^{\top} \tag{32}$$

it is

$$L^{'} = -(1-\xi) + 2\eta * (1-\xi) - 2\eta(1-\xi)^2 = 0. \tag{33}$$

The we get

$$(1-\xi)[-1 + 2\eta\xi] = 0 \tag{34}$$

because $\xi = \frac{1}{1+e^{1+2(w^{\top}x+b)}} < 1$ is always true, we get

$$\xi = \frac{1}{2\eta} \tag{35}$$

so

$$e^{1+2(w^{\top}x+b)} = 2\eta - 1 > 0 \tag{36}$$

then we get

$$\eta > 0.5$$

A.2 IMBALANCED DATA EXPERIMENT

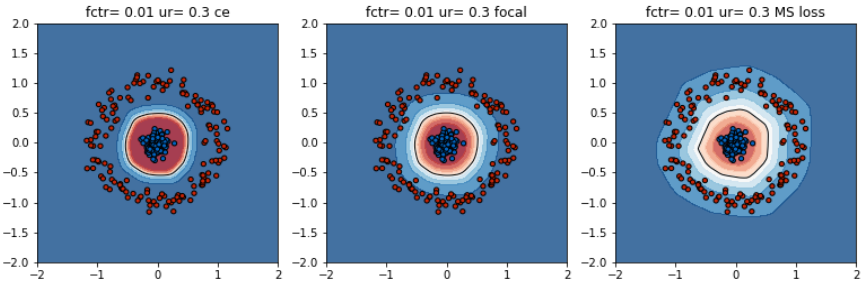

Figure 7: Imbalanced dataset in lacking 30% data points.

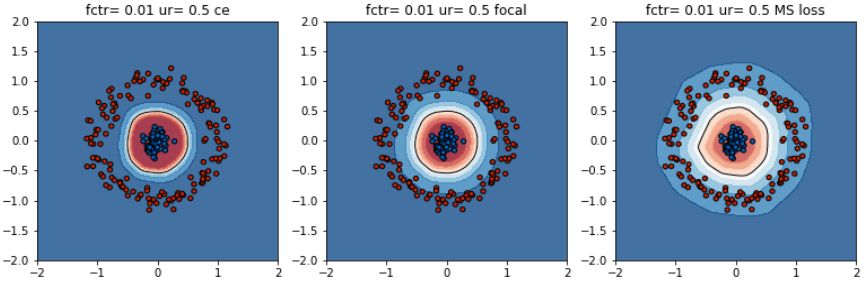

Figure 8: Imbalanced dataset in lacking 50% data points.

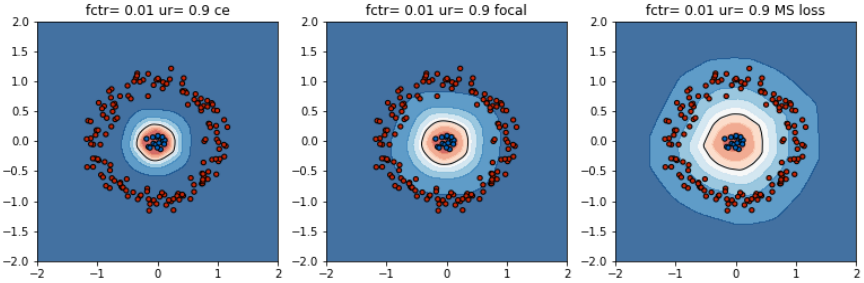

Figure 9: Imbalanced dataset in lacking 90% data points.