Understanding Adversarially Robust Generalization via Weight-Curvature Index

Yuelin Xu YUELIN.XU@CISPA.DE
CISPA Helmholtz Center for Information Security

Xiao Zhang XIAO.ZHANG@CISPA.DE
CISPA Helmholtz Center for Information Security

Abstract

Despite numerous efforts, achieving adversarial robustness in deep learning remains a critical challenge. Recent studies have discovered that adversarial training, a widely adopted method for improving model robustness against adversarial perturbations, prevalently suffers from robust overfitting. To better characterize the robust generalization of adversarially trained models, we introduce the Weight-Curvature Index (WCI), a novel metric that captures the Frobenius norm of layer-wise weight matrices and the trace of the Hessian matrix with respect to the adversarial loss function. In particular, we establish a theoretical connection between WCI and robust generalization gap under a PAC-Bayesian framework. By analyzing the dynamics of these factors, WCI offers a nuanced understanding of why robust overfitting happens during adversarial training. Experimental results demonstrate a strong correlation between WCI and traditional robustness measures, suggesting the effectiveness of WCI in capturing the learning dynamics of adversarial training.

1. Introduction

Building deep learning models to be resilient to adversarial perturbations remains an enduring challenge. An active line of research [2, 5, 6, 11, 19] has emphasized that adversarial training, where models are optimized to withstand worst-case perturbations, is essential to bolster model robustness. While standard deep learning produces models that can generalize to unseen data well, adversarial training presents a starkly different scenario. In particular, overfitting to the training set severely harms the robust generalization of adversarial training, resulting in models that perform well on adversarial examples in the training set but poorly on those in the test set. This phenomenon, known as robust overfitting [18, 25], underscores a significant gap in our understanding of deep learning generalization under adversarial settings. To mitigate robust overfitting and improve the generalization of adversarial training, various robustness-enhancing techniques have been proposed, such as ℓ_2 weight regularization [27], early stopping [25], data augmentation [38, 39], the use of synthetically generated data [12], adversarial weight perturbation (AWP) [33], sharpness-aware minimization (SAW) [30], to name a few. Nevertheless, there is a lack of principled explanations on why robust overfitting happens and what factors are essential for adversarially robust generalization.

A better theoretical indicator of adversarially robust generalization is crucial for pinpointing the bottleneck of state-of-the-art adversarial training methods and inspiring new insights to further build more robust models. Traditional measures, such as robust loss and error gaps, predominantly assess performance degradation under adversarial perturbations but fail to elucidate the underlying mechanisms driving this vulnerability. These indicators focus on the outcomes of adversarial at-

© Y. Xu & X. Zhang.

tacks without considering the intrinsic characteristics of the model, thereby lacking a theoretical foundation for the continuous improvement of robust generalization. Moreover, a notable limitation of existing metrics is their reliance on both training and testing datasets to assess robustness. Few metrics can characterize robust overfitting and generalization solely from the training dataset [28, 34, 37]. This dependence on additional data impedes a comprehensive understanding of these phenomena and constrains the development of novel algorithms aimed at mitigating robust overfitting and enhancing robust generalization (see Appendix A for detailed discussions of related works).

Contributions. In this paper, we introduce the **W**eight-**C**urvature **I**ndex (WCI), a novel metric to quantify a model's vulnerability to adversarial perturbations, utilizing the Frobenius norm of weight matrices and the trace of the Hessian matrix to comprehensively assess the robust generalization performance of adversarially trained models (Definition 5). Our theoretical foundation, based on PAC-Bayesian theory and second-order loss function approximations, elucidates the interplay between robust generalization, model parameters, and loss landscape curvature (Theorem 4). The WCI enables the evaluation of robust overfitting and generalization using only the training dataset, simplifying robustness assessment and aiding in the development of more effective algorithms. Our empirical results validate the WCI's strong correlation with traditional robustness metrics (Section 2.3). By offering a nuanced understanding of adversarial robustness based on the scale of model parameters and the curvature of the loss landscape, our work provides crucial insights for designing more resilient deep learning models, enhancing their reliability and security.

2. Connecting Robust Generalization with Weight-Curvature Index

In this section, we introduce the definition of the Weight-Curvature Index and provide explanations on why it can serve as a effective measure of a model's robustness against adversarial perturbations.

2.1. Adversarial Risk and PAC-Bayesian Bound

Before introducing the Weight-Curvature Index, we first present a lemma, proven in Appendix B.1, which establishes a PAC-Bayesian adversarial risk bound, inspired by prior work [21, 29].

Lemma 1 (Adversarial Risk with Catoni's bound [1]) *Let $\lambda > 0$, $\epsilon \in (0,1)$, and let $\mathcal{D}$ be any probability distribution. Consider $\mathcal{H}$ as a set of classifiers and $\mathcal{P}$ as a prior distribution supported by $\mathcal{H}$. For each classifier in $\mathcal{H}$, let θ denote the set of parameters that determine the behavior of the network. For a training set $\mathcal{S}$ consisting of n samples (x, y) drawn from $\mathcal{D}$, where x represents the input data and y is the corresponding target label, let g_θ denote the output logits. For any posterior distribution $\mathcal{Q}$ over $\mathcal{H}$, define the expected loss under distribution $\mathcal{D}$ for parameters as $\mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta, y), g_\theta)]$ and the empirical estimate of the loss on the training set $\mathcal{S}$ as $\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta, y), g_\theta)]$. Then, the adversarial risk $\mathcal{R}_{\text{adv}}(f)$ is bounded by:*

$$\begin{aligned}\mathcal{R}_{\text{adv}}(f) \leq & \underbrace{\mathbb{E}_{(x,y)\sim D}[\mathcal{L}((x+\delta, y), g_\theta)] - \mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim D}[\mathcal{L}((x+\delta, y), g_\theta)]}_{\textit{Perturbation Discrepancy}} + \underbrace{\frac{1}{\lambda}\text{KL}[\mathcal{Q}||\mathcal{P}]}_{\textit{KL Divergence}} \\ & + \underbrace{\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta, y), g_\theta)] - \mathcal{L}((x+\delta, y), g_\theta)}_{\textit{Classifier Variability}} + \mathcal{L}((x+\delta, y), g_\theta) + \frac{\lambda C^2}{8n} - \frac{1}{\lambda}\ln\epsilon.\end{aligned} \tag{1}$$

Recall that our objective is to establish a reliable index of robustness generalization capacity, specifically aiming to establish a linkage with adversarial risk $\mathcal{R}_{\text{adv}}(f)$. Thus, we only need to focus on the terms in Eq. 1 that are relevant to adversarial robustness. First, we note that the term $\frac{\lambda C^2}{8n} - \frac{1}{\lambda}\ln\epsilon$ can be detached from robustness metrics. In addition, previous studies has illustrated that the term $\mathcal{L}((x+\delta, y), g_\theta)$ lacks direct correlation with robustness measures [25], and significantly reducing the *Classifier Variability* may also reduce the *Perturbation Discrepancy* component [4, 7, 10, 20]. Consequently, our analysis in the following discussions predominantly explores the influence of the *KL Divergence* and *Classifier Variability* terms on enhancing the model's resistance to adversarial inputs. To further simplify the understanding of the aforementioned two terms, we incorporate hyperprior and second-order loss approximation techniques, which are explained as follows.

Incorporating Hyperprior. We introduce a hyperprior for the standard deviation $\sigma_{\mathcal{P}}$ of the parameters for each weight matrix, following Kim and Hospedales [15]. To maintain prior variance invariant to parameter rescaling, we adopt specialized hyperpriors from Sefidgaran et al. [26]. Specifically, we utilize a uniform hyperprior selected from a finite set of positive real numbers, ensuring precise representation with floating-point arithmetic [32]. This approach guarantees robust Bayesian inference and provides a viable framework for parameter standardization across varying scales.

Assumption 1 *To adhere to standard practice in Bayesian neural networks, we define the prior distribution $\mathcal{P}$ as a Gaussian with zero mean and a diagonal covariance matrix $(\sigma_{\mathcal{P}}^2\mathbf{I})$, ensuring $\sigma_{\mathcal{P}} > 0$. Simultaneously, we model the posterior distribution $\mathcal{Q}$ as a Gaussian characterized by the mean vector θ, representing the network parameters, and a covariance matrix $(\sigma_{\mathcal{Q}}^2\mathbf{I})$, with $\sigma_{\mathcal{Q}} > 0$.*

The following lemma, proven in Appendix B.2, indicates that the *KL divergence* term is directly proportional to the squared Frobenius norm of the parameters when the prior variances are constant.

Lemma 2 (Otto's KL divergence [24]) *To minimize the Kullback-Leibler (KL) divergence term effectively, we equate the prior variance $(\sigma_{\mathcal{P}}^{(l)})^2$ equal to the posterior variance $(\sigma_{\mathcal{Q}}^{(l)})^2$, and denote both variances as $(\sigma^{(l)})^2$. Consequently, the KL divergence is given by:*

$$\mathrm{KL}[\mathcal{Q}||\mathcal{P}] = \sum_l \frac{\|W^{(l)}\|_F^2}{2(\sigma^{(l)})^2} + \text{const}, \tag{2}$$

where $W^{(l)}$ is the weight matrix of the l-th layer and $\|W^{(l)}\|_{\mathrm{F}}^2$ denotes its squared Frobenius norm.

Second-Order Approximation of Loss. To integrate PAC-Bayesian theory with the Hessian matrix of the loss landscape, we employ a second-order approximation to various surrogate loss functions, building on insights from recent research [17, 31, 36]. We model the posterior distribution of the parameters, denoted as $\mathcal{Q}$, as a Gaussian distribution with unit variance.

Assumption 2 *The loss function $\mathcal{L}(\theta, x+\delta, y)$ can be approximated using a second-order approximation. We use the unit-variance Gaussian as the posterior of parameters.*

The following lemma, proven in Appendix B.3, shows that the *Classifier Vulnerability* term under this framework can be quantified by the expectation of the loss over the distribution of parameters.

Lemma 3 (Hessian-based Variability [9]) *By approximating the loss function $\mathcal{L}(\theta, x+\delta, y)$ using a second-order expansion, the relationship between the expected variability in the classifier's performance and the curvature of the loss landscape is given by:*

$$\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta) \leq \sum_l \mathrm{Tr}(H^{(l)}) \cdot (\sigma^{(l)})^2, \tag{3}$$

where $\mathrm{Tr}(H^{(l)})$ denotes the accumulation of the diagonal elements of the Hessian matrix of the training loss with respect to the weight matrix $W^{(l)}$, and $(\sigma^{(l)})^2$ is the variance associated to $W^{(l)}$.

2.2. Defining Weight-Curvature Index

Putting pieces together, the following theorem proves an upper bound on the *KL Divergence* and *Classifier Vulnerability* terms in Eq. 1, which is related to the Frobenius norm of layer-wise model weights and the trace of the corresponding Hessian matrix with respect to adversarial loss.

Theorem 4 (Adversarial Risk Simplification) *Under the same settings as in Lemma 1, we have*

$$\begin{aligned}\frac{1}{\lambda}\mathrm{KL}[\mathcal{Q}||\mathcal{P}] + \mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta)& \\ \leq \sqrt{\frac{2}{\lambda}}\sum_l \sqrt{\|W^{(l)}\|_{\mathrm{F}}^2 \cdot \mathrm{Tr}(H^{(l)})} + \mathrm{const},&\end{aligned} \tag{4}$$

where λ is a normal constant, $\|W^{(l)}\|_{\mathrm{F}}^2$ is the squared Frobenius norm of the weight matrix $W^{(l)}$, and $\mathrm{Tr}(H^{(l)})$ denotes the trace of the Hessian matrix of the loss function with respect to $W^{(l)}$.

The proof of Theorem 4 is provided in Appendix B.4. According to Eq. 4, irrespective of the value of λ that achieves the infimum in the PAC-Bayes bound, a smaller value of the combined metric $\sum_l \sqrt{\|W^{(l)}\|_{\mathrm{F}}^2 \cdot \mathrm{Tr}(H^{(l)})}$ implies tighter robustness bound, which we formally define as follows:

Definition 5 (Weight Curvature Index) *The Weight-Curvature Index is defined as:*

$$\mathrm{WCI}_{\mathrm{adv}} := \sum_l \sqrt{\|W^{(l)}\|_F^2 \cdot \mathrm{Tr}(H^{(l)})}, \tag{5}$$

where $W^{(l)}$ represents the weight matrix of the l-th layer, $\|W^{(l)}\|_F$ denotes the Frobenius norm of the weight matrix $W^{(l)}$, $H^{(l)}$ is the Hessian matrix of the loss function with respect to the parameters of the l-th layer and $\mathrm{Tr}(H^{(l)})$ stands for the trace of the Hessian matrix $H^{(l)}$.

This index quantifies the interaction between the scale of the model's parameters and the curvature of the loss landscape. Note that $\mathrm{WCI}_{\mathrm{adv}}$ is scaled by the magnitude of each weight matrix, ensuring that the metric is invariant to parameter rescaling [22]. A higher $\mathrm{WCI}_{\mathrm{adv}}$ indicates higher vulnerability to adversarial perturbations, while a lower value suggests enhanced robustness.

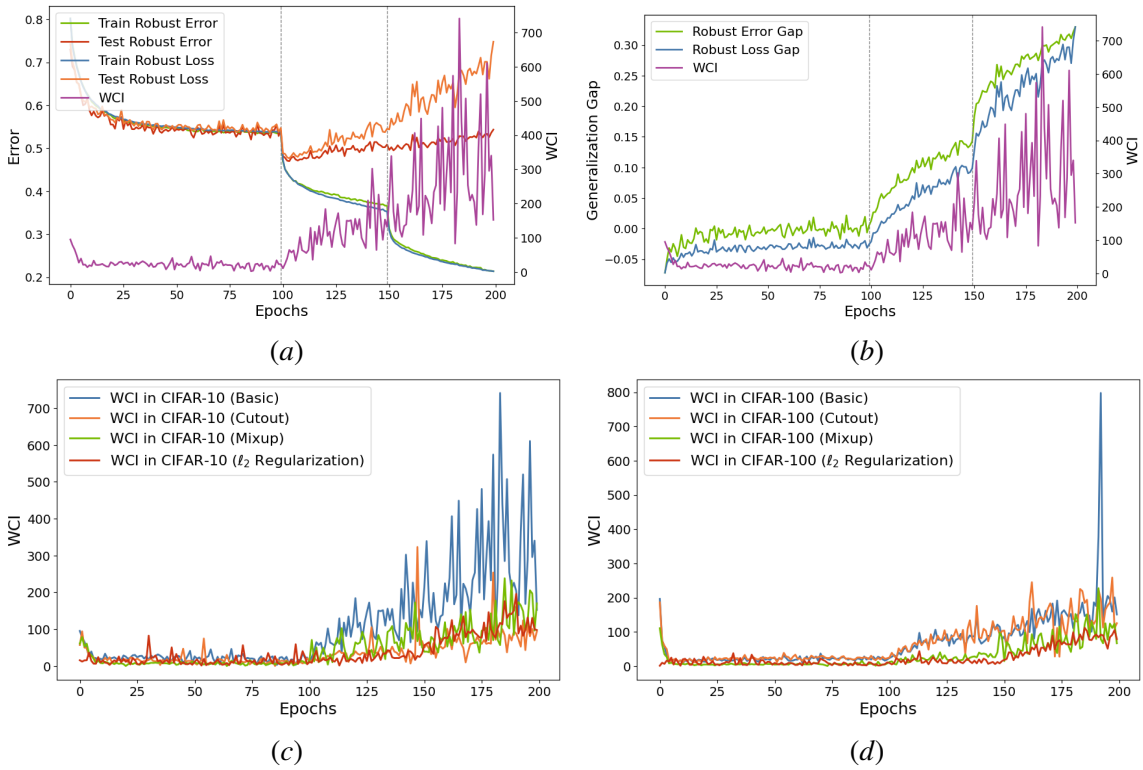

Figure 1: Learning curves of the Weight-Curvature Index of standard adversarial training on CIFAR-10 with respect to (a) Robust error and loss, and (b) robust generalization Gaps. (c) and (d) show WCI curves with various regularization techniques on CIFAR-10 and CIFAR-100 respectively.

2.3. Experiments

In this section, we conduct experiments to examine the effectiveness of the Weight-Curvature Index (WCI) as a metric for assessing the robustness of deep learning models against adversarial attacks. Our experiments are mainly based on the methodology and findings presented in the work by Rice et al. [25]. Figure 1(a) shows the relationship between the WCI, robust loss, and robust error over 200 training epochs, while Figure 1(b) compares the generalization gap (robust loss gap and robust error gap) with WCI. Further analysis in Figures 1(c) and 1(d) compare the WCI curves for models trained on CIFAR-10 and CIFAR-100 using different regularization and data augmentation techniques, including basic adversarial training, cutout, mixup, and ℓ_2 regularization. More details of our experiments are provided in Appendix C. The consistent trend of WCI across datasets and techniques highlights overfitting as a common phenomenon. Our empirical findings demonstrate a strong correlation between WCI and both robust loss and error gaps across training and testing datasets, affirming our theoretical results. Higher WCI values correspond to increased robustness losses and error gaps, especially pronounced during the overfitting periods.

3. Conclusion

We proposed the WCI as a novel metric for quantifying the susceptibility of deep learning models to adversarial perturbations. Our theoretical analysis and empirical evidence demonstrated that the WCI aligns well with traditional measures of robust generalization performance of adversarially trained models. By integrating the scale of model parameters with the curvature of the loss landscape, the WCI provides a more nuanced understanding of adversarial robustness. Our findings indicate that models with higher WCI values exhibit greater susceptibility to adversarial perturbations, as evidenced by increased robust loss and error gaps. This correlation underscores the utility of the WCI in identifying and mitigating overfitting in adversarially robust deep learning models. Moreover, our results suggest that the WCI may serve as a valuable tool for guiding the development of more robust models. Future research should focus on designing algorithms that can better control the model's WCI during training while exploring its applicability across a broader range of architectures and adversarial training methods.

References

[1] Pierre Alquier. User-friendly introduction to pac-bayes bounds. *Found. Trends Mach. Learn.*, 17(2):174–303, jan 2024. ISSN 1935-8237. doi: 10.1561/2200000100. URL https://doi.org/10.1561/2200000100.

[2] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16048–16059. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b8ce47761ed7b3b6f48b583350b7f9e4-Paper.pdf.

[3] David Banks, Víctor Gallego, Roi Naveiro, and David Ríos Insua. Adversarial risk analysis: An overview. *WIREs Computational Statistics*, 14(1):e1530, 2022. doi: https://doi.org/10.1002/wics.1530. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1530.

[4] Arash Behboodi, Gabriele Cesa, and Taco Cohen. A PAC-bayesian generalization bound for equivariant networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=6dfYc2IUj4.

[5] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/32e0bd1497aa43e02a42f47d9d6515ad-Paper.pdf.

[6] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Informa-*

tion Processing Systems Datasets and Benchmarks Track (Round 2), 2021. URL https://openreview.net/forum?id=SSKZPJCt7B.

[7] Darshil Doshi, Tianyu He, and Andrey Gromov. Critical initialization of wide and deep neural networks using partial jacobians: General theory and applications. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=wRJqZRxDEX.

[8] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*, 2017. URL http://auai.org/uai2017/proceedings/papers/173.pdf.

[9] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

[10] Zhijin Ge, Xiaosen Wang, Hongying Liu, Fanhua Shang, and Yuanyuan Liu. Boosting adversarial transferability by achieving flat local maxima. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=AKAMNDe2Sw.

[11] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples, 2021. URL https://arxiv.org/abs/2010.03593.

[12] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A. Mann. Improving robustness using generated data. In *NeurIPS*, pages 4218–4233, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/21ca6d0cf2f25c4dbb35d8dc0b679c3f-Abstract.html.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[14] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

[15] Minyoung Kim and Timothy Hospedales. Bayestune: Bayesian sparse deep model fine-tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=TRuqrVsmZK.

[16] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[17] Bingcong Li and Georgios B. Giannakis. Enhancing sharpness-aware optimization through variance suppression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Sf3t6Bth4P.

[18] Binghui Li and Yuanzhi Li. Towards understanding clean generalization and robust overfitting in adversarial training, 2023. URL https://arxiv.org/abs/2306.01271.

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

[20] Pierre Marion. Generalization bounds for neural ordinary differential equations and deep residual networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=992vogTP1L.

[21] Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. Statistical guarantees for variational autoencoders using PAC-bayesian theory. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=jkPDRHff3s.

[22] Maximilian Mueller, Tiffany Joyce Vlaar, David Rolnick, and Matthias Hein. Normalization layers are all that sharpness-aware minimization needs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lArwl3y9x6.

[23] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

[24] Fabian Otto, Philipp Becker, Vien Anh Ngo, Hanna Carolin Maria Ziesche, and Gerhard Neumann. Differentiable trust region layers for deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qYZD-AO1Vn.

[25] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, pages 8093–8104, 2020. URL http://proceedings.mlr.press/v119/rice20a.html.

[26] Milad Sefidgaran, Abdellatif Zaidi, and Piotr Krasnowski. Minimum description length and generalization guarantees for representation learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Ncb0MvVqRV.

[27] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6969–6980, 2019. doi: 10.1109/CVPR.2019.00714.

[28] David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7787–7797, 2021. doi: 10.1109/ICCV48922.2021.00771.

[29] Paul Viallard, Maxime Haddouche, Umut Simsekli, and Benjamin Guedj. Learning via wasserstein-based high probability generalisation bounds. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3Wrolscjbx.

[30] Zeming Wei, Jingyu Zhu, and Yihao Zhang. Sharpness-aware minimization alone can improve adversarial robustness. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. URL https://openreview.net/forum?id=bxsqPkm2m9.

[31] Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Dkmpa6wCIx.

[32] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf.

[33] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2958–2969. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1ef91c212e30e14bf125e9374262401f-Paper.pdf.

[34] Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15446–15459. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/637de5e2a7a77f741b0b84bd61c83125-Paper-Conference.pdf.

[35] Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. PAC-bayesian spectrally-normalized bounds for adversarially robust generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ydKWoqWZ3t.

[36] Zeke Xie, Qian-Yuan Tang, Mingming Sun, and Ping Li. On the overlooked structure of stochastic gradients. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=H4GsteoL0M.

[37] Yaodong Yu, Zitong Yang, Edgar Dobriban, Jacob Steinhardt, and Yi Ma. Understanding generalization in adversarial training via the bias-variance decomposition, 2021. URL https://arxiv.org/abs/2103.09947.

[38] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. doi: 10.1109/ICCV.2019.00612. URL https://arxiv.org/abs/1905.04899.

[39] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Appendix A. Related Work

The study of overfitting and generalization in deep learning has been a central focus of research for many years. In adversarial learning, overfitting is a prevalent phenomenon, characterized by models performing well on the training set but poorly on the test set. Understanding this phenomenon can significantly enhance the robustness of adversarial learning. Standard deep learning generalization theory frequently employs PAC-Bayesian frameworks to provide probabilistic guarantees on the generalization performance of models. This approach has been instrumental in elucidating neural network behavior and developing methods to enhance their generalization capabilities.

Adversarially Robust Generalization. Robust generalization refers to the capability of models to perform well on both clean data and data subjected to adversarial perturbations. Rice et al. [25] highlight the issue of overfitting in adversarially robust deep learning, showing that models tend to overfit to adversarial examples, thereby reducing their generalization performance on new adversarial attacks. Moreover, the work by Xiao et al. [35] on PAC-Bayesian spectrally-normalized bounds for adversarially robust generalization provides valuable insights into how spectral normalization can improve the robustness and generalization of neural networks. This research underscores the effectiveness of integrating spectral norms into the PAC-Bayesian framework to achieve tighter robustness bounds and enhanced generalization.

Generalization Theory and PAC-Bayesian Frameworks. Dziugaite and Roy [8] applied PAC-Bayesian bounds to deep networks, demonstrating their utility in providing non-vacuous generalization guarantees. Neyshabur et al. [23] further extended these ideas by relating PAC-Bayesian bounds to the stability and complexity of the learning algorithm. These studies demonstrate the versatility of the PAC-Bayesian approach in providing theoretical guarantees for deep learning models.

Sharpness of the Loss Landscape. Recent works have investigated the impact of the loss landscape's sharpness and flatness on model robustness. Jiang et al. [14] proposed a generalization metric based on the sharpness of the loss landscape, showing that flatter minima correlate with better generalization. Foret et al. [9] introduced Sharpness-Aware Minimization (SAM), a method that minimizes sharpness to improve generalization. These studies indicate that flatter loss landscapes are associated with improved generalization and robustness properties.

The Weight-Curvature Index (WCI) proposed in this paper is closely related to the concepts of PAC-Bayesian frameworks and the sharpness of the loss landscape. PAC-Bayesian frameworks utilize the Frobenius norm, while sharpness is primarily analyzed through the trace of the Hessian matrix. Our work builds on these foundations by integrating the PAC-Bayesian framework with second-order loss function approximations to develop the WCI. By integrating these elements, the WCI offers a comprehensive measure of model robustness, capturing the interplay between model parameters and loss landscape curvature.

Appendix B. Detailed Proofs of Main Theoretical Results

B.1. Proof of Lemma 1

To prove Lemma 1, we need to make use of the following definition and lemma. Definition 6 evaluates the robustness of a model against adversarial attacks. Lemma 7 characterizes a fundamental PAC-Bayes bound, known as Catoni's bound [1] but is adapted for adversarially robust learning.

Definition 6 (Adversarial Risk [3]) *Let f refers to the model or classifier being evaluated, $\mathbb{E}$ represents the expected value over the distribution of data points. $(x, y) \sim \mathcal{D}$ indicates that each data point (x, y) is drawn from a distribution $\mathcal{D}$, where x is a feature vector and y is the corresponding label. Let $\mathcal{L}$ denotes the loss function used to measure the discrepancy between the model's prediction on the perturbed input $(x + \delta)$ and the true label y, and g_θ represents the predictive function of the model, parameterized by θ. δ is the perturbation added to the input x, crafted specifically to mislead the model f. The goal of δ in this setting is to find the worst-case perturbation that maximizes the loss, reflecting the model's vulnerability to adversarial examples.*

$$\mathcal{R}_{adv}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x + \delta, y), g_\theta)]$$

Lemma 7 *Let $\lambda > 0$, $\epsilon \in (0, 1)$, and let $\mathcal{D}$ be any distribution. Consider $\mathcal{H}$ as a set of classifiers and $\mathcal{P}$ as a prior distribution supported by $\mathcal{H}$. For each classifier in $\mathcal{H}$, define θ as the set of parameters that determine the behavior of the network. For a training set $\mathcal{S}$ of n samples (x, y) drawn from $\mathcal{D}$, where x is the input data and y is the corresponding target label, let g_θ denote the output logits. For any posterior distribution $\mathcal{Q}$ over $\mathcal{H}$, we define:*

- *The expected loss under distribution $\mathcal{D}$ for parameters:*

$$\mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x + \delta, y), g_\theta)],$$

- *The empirical estimate of the loss on the training set $\mathcal{S}$:*

$$\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x + \delta, y), g_\theta)].$$

The following bound is satisfied with probability at least $1 - \epsilon$:

$$\mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x + \delta, y), g_\theta)] \leq \mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x + \delta, y), g_\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathrm{KL}[\mathcal{Q} \parallel \mathcal{P}] + \ln \frac{1}{\epsilon}}{\lambda}.$$

Proof For the sake of completeness, we present the proof of Lemma 7 first, which is based on the PAC-Bayesian theorem [1]. The bound is derived by applying the PAC-Bayesian theorem to the expected loss under distribution $\mathcal{D}$ and the empirical estimate of the loss on the training set $\mathcal{S}$.

Catoni's bound shows that for any $\lambda > 0$, any $\epsilon \in (0, 1)$,

$$\mathbb{P}_{\mathcal{S}}\left(\forall\rho \in \mathcal{P}(\Theta), \mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathrm{KL}(\rho\|\pi) + \log \frac{1}{\varepsilon}}{\lambda}\right) \geq 1 - \varepsilon.$$

Define the empirical loss $\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x + \delta, y), g_\theta)]$ as the average loss over the training set with perturbations δ, and the expected loss $\mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta, y), g_\theta)]$ as the average loss across the distribution $\mathcal{D}$ considering the same perturbations. Applying Catoni's bound involves a theoretical result that relates the true risk $R(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta, y), g_\theta)]$ of a hypothesis g_θ and its empirical risk $r(\theta) = \mathcal{L}((x + \delta, y), g_\theta)$ on a finite sample set. The bound states that with probability at least $1 - \epsilon$, the following inequality holds for all probability distributions ρ on the hypothesis space Θ induced by $\mathcal{Q}$:

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathrm{KL}(\rho\|\mathcal{P}) + \log \frac{1}{\epsilon}}{\lambda},$$

where C is a bound on the loss function $\mathcal{L}$, and $\mathrm{KL}(\rho\|\mathcal{P})$ represents the Kullback-Leibler divergence from the posterior ρ to the prior $\mathcal{P}$.

For the adversarial setting, this bound becomes particularly useful in assessing how well a model that has been trained with adversarial examples (represented by δ) can generalize from its empirical loss on training data to its expected performance on the overall distribution. The bound provides a trade-off between the empirical loss and the expected loss, with the KL divergence term and the classifier variability term contributing to the generalization error. Since $\mathcal{Q}$ is a distribution over classifiers, integrating the Catoni's bound over $\mathcal{Q}$ yields:

$$\mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta,y),g_\theta)] \leq \mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] + \frac{\lambda C^2}{8n} + \frac{\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) + \log\frac{1}{\epsilon}}{\lambda}.$$

Therefore, we complete the proof of Lemma 7. ■

Proof [Proof of Lemma 1] Using Definition 6 and Lemma 7, we can immediately derive the adversarial risk bound:

$$\begin{aligned}
\mathcal{R}_{adv}(f) :=& \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta,y),g_\theta)] \\
=& \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta,y),g_\theta)] \\
&+ \mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta,y),g_\theta)] \\
\leq& \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathbb{E}_{\theta\sim\mathcal{Q}}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}((x+\delta,y),g_\theta)] \\
&+ \mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta) \\
&+ \lambda^{-1}KL[\mathcal{Q}||\mathcal{P}] + \frac{\lambda C^2}{8n} + \mathcal{L}((x+\delta,y),g_\theta) - \lambda^{-1}\ln\epsilon,
\end{aligned}$$

which completes the proof of Lemma 1. ■

B.2. Proof of Lemma 2

Proof In the context of adversarial machine learning, the Kullback–Leibler (KL) divergence measures how a model's distribution $\mathcal{Q}$, representing the learned classifiers, diverges from a prior distribution $\mathcal{P}$ under adversarial conditions. Specifically, this divergence can be adapted to account for the added complexity introduced by adversarial perturbations to the input data.

The KL divergence term is adapted for adversarial conditions as follows:

$$\mathrm{KL}[\mathcal{Q}||\mathcal{P}] = \sum_l \left[\ln\left(\frac{\sigma_P^{(l)}}{\sigma_Q^{(l)}}\right) + \frac{\|W^{(l)}\|_F^2 + (\sigma_Q^{(l)})^2}{2(\sigma_P^{(l)})^2}\right] + \text{const}.$$

Here, $W^{(l)}$ denotes the weights of the l-th layer of the neural network. The terms $\sigma_P^{(l)}$ and $\sigma_Q^{(l)}$ represent the variances of the prior and posterior distributions of the weights for this layer, respectively. These variances reflect the spread of the weight values and are key to understanding the network's robustness to adversarial attacks; larger variances in the posterior suggest a model that is more sensitive to input perturbations. In the adversarial setting, these variances are particularly crucial as they directly influence the classifier's stability.

The KL divergence is further simplified when the variances of the prior and posterior distributions are equal, which is a common assumption made to facilitate the calculation. In such cases, the KL divergence simplifies to:

$$\mathrm{KL}[\mathcal{Q}||\mathcal{P}] = \sum_l \frac{\|W^{(l)}\|_F^2}{2(\sigma^{(l)})^2} + \text{const.}$$

In this simplified form, the KL divergence is directly proportional to the Frobenius norm of the weight matrices, scaled by the variance of the distributions, and offers a computationally tractable measure for evaluating the divergence in an adversarial machine learning setting. ■

B.3. Proof of Lemma 3

Proof The adversarial loss approximation in the context of adversarial ML involves considering the stability of the training loss in the face of adversarial perturbations. We can express the difference in empirical losses between the perturbed and unperturbed classifier with parameters θ as:

$$\begin{aligned}\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\mathbb{E}_{(x,y)\in S}[\mathcal{L}((x+\delta,y),g_\theta+\epsilon)] \\ - \mathcal{L}((x+\delta,y),g_\theta),\end{aligned} \tag{6}$$

which captures the average adversarial loss over perturbations ϵ drawn from a standard Gaussian distribution. This difference can be approximated using the second-order Taylor series expansion around θ, yielding:

$$\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta) \leq \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\epsilon^\top \nabla_\theta^2 \mathbb{E}_{(x,y)\in S}[\mathcal{L}((x+\delta,y),g_\theta)]\epsilon], \tag{7}$$

and simplifying this further using the trace of the Hessian, we get:

$$\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta) \leq \mathrm{Tr}(\nabla_\theta^2 \mathbb{E}_{(x,y)\in S}[\mathcal{L}((x+\delta,y),g_\theta)]), \tag{8}$$

which quantifies the effect of adversarial perturbations on the loss landscape's curvature and provides a measure for the classifier's adversarial robustness. Note that the Hessian matrix $H^{(l)}$ for the l-th layer of the neural network is defined as:

$$H^{(l)} = \sum_{i,j} \frac{\partial^2 \mathcal{L}((x+\delta,y),g_\theta)}{\partial W^{(l)}[i,j]^2}, \tag{9}$$

which encapsulates the second-order partial derivatives of the loss function with the weights of the layer, indicating how the loss curvature changes in response to perturbations in the weights. Therefore, we can approximate the Classifier Variability Component by the trace of the Hessian. This approximation provides a computationally efficient method to evaluate the classifier's sensitivity to adversarial perturbations, offering insights into the model's robust generalization capabilities:

$$\mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta) \leq \sum_l \mathrm{Tr}(H^{(l)}) \cdot (\sigma^{(l)})^2, \tag{10}$$

which completes the proof. ■

B.4. Proof of Theorem 4

Proof According to Lemma 2 and Lemma 3, we obtain

$$\begin{aligned}
&\lambda^{-1}\mathrm{KL}[\mathcal{Q}||\mathcal{P}] + \mathbb{E}_{\theta\sim\mathcal{Q}}[\mathcal{L}((x+\delta,y),g_\theta)] - \mathcal{L}((x+\delta,y),g_\theta) \\
\leq &\lambda^{-1}\left(\sum_l \frac{\|W^{(l)}\|_F^2}{2(\sigma^{(l)})^2} + \mathrm{const}\right) + \sum_l \mathrm{Tr}(H^{(l)})\cdot(\sigma^{(l)})^2 \\
= &\sum_l \left(\frac{\|W^{(l)}\|_F^2}{2\lambda(\sigma^{(l)})^2} + \mathrm{Tr}(H^{(l)})\cdot(\sigma^{(l)})^2\right) + \mathrm{const} \\
\leq &2\sum_l \sqrt{\frac{\|W^{(l)}\|_F^2}{2\lambda(\sigma^{(l)})^2}\cdot \mathrm{Tr}(H^{(l)})\cdot(\sigma^{(l)})^2} + \mathrm{const} \\
= &\sqrt{\frac{2}{\lambda}}\sum_l \sqrt{\|W^{(l)}\|_{\mathrm{F}}^2\cdot \mathrm{Tr}(H^{(l)})} + \mathrm{const},
\end{aligned}$$

where the last inequality holds because of the Cauchy-Schwarz inequality, and the equality holds if and only if $(\sigma^{(l)})^2 = \sqrt{\frac{\|W^{(l)}\|_{\mathrm{F}}^2}{2\lambda \mathrm{Tr}(H^{(l)})}}$. Therefore, we complete the proof. ■

Appendix C. Experimental Details

C.1. Experimental Setup

- **Dataset**: We utilize the CIFAR-10 and CIFAR-100 datasets [16] for our experiments. Both datasets consist of 32 × 32 color images. CIFAR-10 is composed of 60,000 images divided into 10 classes, with 50,000 training images and 10,000 test images. CIFAR-100, while similar in total number of images and image size, is categorized into 100 classes, presenting a more granular classification challenge due to finer distinctions among the categories.
- **Model Architecture**: The PreActResNet18 model architecture [13] is evaluated in our experiments. PreActResNet18 is a pre-activation variant of the ResNet architecture known for its robustness and performance in various image classification tasks.
- **Adversarial Attack Method**: We employ the Projected Gradient Descent (PGD) method [19] to assess the robustness of the models. PGD is a widely used technique for generating adversarial examples.
- **Training Procedure**:
  - **Initial Learning Rate**: The model is trained with an initial learning rate 0.1.
  - **Learning Rate Decay**: The learning rate is decayed by a factor of 0.1 at the 100th and 150th epochs.

C.2. Evaluation Metrics

The primary metrics used for evaluation are:

- **Weight-Curvature Index (WCI)**: As defined in Definition 5, the WCI measures the interaction between the scale of model parameters and the curvature of the loss landscape, directly calculated on the training set.
- **Robust Loss**: The loss calculated on adversarially perturbed inputs.
- **Robust Error**: The classification error on adversarially perturbed inputs.
- **Robust Loss Gap**: The difference in robust loss between the training and test sets.
- **Robust Error Gap**: The difference in robust error between the training and test sets.