ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Mitigating Robustness Bias: Theoretical Results and Empirical Evidences

Danilo Franco, Luca Oneto, and Davide Anguita [*]

University of Genoa - Via Opera Pia 11a, 16145, Genova, Italy

Abstract. Recent research has shown that some learned classifiers can be more easily fooled by an adversary who carefully crafts imperceptible or physically plausible modifications of the input data regarding particular subgroups of the population (e.g., people with particular gender, ethnicity, or skin color). This form of unfairness has been just recently studied, noting the fact that classical fairness metrics, which only observe the model outputs, are not enough but robustness biases need to be measured and mitigated as well. For this reason, in this paper, we will first develop a new metric of fairness which generalizes the current ones and degenerates in the classical ones and then we will develop a theoretical mitigation framework with consistency results able to generate a new empirical mitigation strategy and explain why the current ones actually work.

1 Introduction

Predictive models learned from data are nowadays ubiquitous thanks to massive investments in products able to make them a commodity. In some applications, e.g., games [1], healthcare [2], and text generation [3], these tools have been shown to compare to human capabilities. Nevertheless, these achievements are accompanied by increasing concerns about their impacts on society [4]. In fact, it has been shown how learned classifiers can be easily fooled by an adversary who carefully crafts imperceptible or physically plausible modifications of the input data [5] and that they exhibit the same historical human biases (e.g., threading unfairly subgroups of the population based on gender, ethnicity, or skin color) that are hidden in the data [6]. The problem becomes even worse when these two weaknesses combine together. In fact, many recent works [7–14] actually showed that certain learned classifiers can be more easily fooled by an adversary who carefully crafts imperceptible or physically plausible modifications of the input data regarding particular subgroups of the population producing a robustness bias. For the purpose of measuring this bias, they propose different metrics that, as for the classical ones, are sometimes in contrast with each other [6].

[*] This work is supported in part by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. This work is also partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Then, they propose methods to mitigate these biases with practical approaches partially or not supported by a theoretical background.

In this paper, we will first develop a new generalized metric of fairness (Section 2) which encompasses the current ones for robustness biases and degenerates into the classical ones of fairness. Then we will develop a theoretical mitigation framework (Section 3) with consistency results (i.e., robustness and robustness bias generalization bound) able to generate a new empirical mitigation strategy (Section 4) and explain why the current ones actually work. Section 5 concludes the paper.

2 Robustness Bias

Let $\mathcal{D}=\{(X_1,s_1,y_1),\cdots,(X_n,s_n,y_n)\}$ be a sequence of n samples drawn independently from an unknown probability distribution μ over $\mathcal{X}\times\mathcal{S}\times\mathcal{Y}$, where $\mathcal{Y}=\{\pm1\}$ is the set of binary output labels[1], $\mathcal{S}=\{a,b\}$ represents group membership among two groups[2] (e.g. 'female' or 'male'), and $\mathcal{X}$ is the input space. We note that the input $X\in\mathcal{X}$ may further contain or not the sensitive feature $s\in\mathcal{S}$ in it[3]. Based on a random observation of $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$ one has to estimate $y \in \mathcal{Y} \subseteq \{\pm1\}$ by choosing a suitable hypothesis $h : \mathcal{X} \to \hat{\mathcal{Y}}$ in a set of possible ones $\mathcal{H}$. A learning algorithm selects and hypothesis $h : \mathcal{X}\to\hat{\mathcal{Y}}\subseteq\mathbb{R}$in a set of possible ones $\mathcal{H}$ based on $\mathcal{D}$. The generalization error (i.e., the risk) $\mathsf{L}_\ell(h)=\mathbb{E}_{(X,y)}\{\ell(h(X),y)\}$ together with the empirical one[4] $\hat{\mathsf{L}}_\ell(h)=\hat{\mathbb{E}}_{(X,y)}\{\ell(h(X),y)\}$ associated to an hypothesis h, is defined through a loss function $\ell{:}\hat{\mathcal{Y}}\times\mathcal{Y}\to[0,1]$. The hypothesis h is subject to an adversary which tries to fool the model into mistakes by modifying the observation X according to a set of possible modifications $\mathcal{B}(X)$. We can define then a new loss $\tilde{\ell}(h(X),y)=\sup_{X\in\mathcal{B}(X)}\ell(h(X),y)$ to measure the robustness to this adversary together with the generalization robustness $\mathsf{L}_{\tilde{\ell}}(h)$ and its empirical estimator $\hat{\mathsf{L}}_{\tilde{\ell}}(h)$. Note that when $\mathcal{B}(X)=X$ we have that $\tilde{\ell}=\ell$. Moreover, we request the hypothesis h to be also fair, namely it should not behave differently if applied to subgroups of the population [6]. For this purpose, different metrics have been defined with ϵ-fairness [16] being the most general one encompassing all the most common notions of fairness, e.g., Difference of Demographic Parity (DDP) and Difference of Equal Opportunity.

What we target here is to measure a sort of fusion between robustness and fairness, namely the bias of robustness of the model over the different subgroups in the population.

[1] The extension to multiclass classification is not reported for space constraints.

[2] The extension to multiple subgroups is not reported for space constraints.

[3] The sensitive feature may not be available in the testing phase or it might not be possible to use it as a predictor in the model due to legal requirements [15].

[4] With $\hat{\mathbb{E}}_{(X,y)}$ we indicate the expectation restricted to $\mathcal{D}$.

Definition 1. *Let $\mathtt{G}_{\cdot}(s,y){:}\mathcal{S}{\times}\mathcal{Y}{\to}\mathbb{B}$ be a Boolen-valued function, then h is ϵ-bias-robust if $\mathtt{F}_{\tilde{\ell}}(h){=}|\mathsf{L}_{\tilde{\ell},\mathtt{G}_a,\dot{y}}(h){-}\mathsf{L}_{\tilde{\ell},\mathtt{G}_b,\dot{y}}(h)|{\leq}\epsilon$ where we have defined $\mathsf{L}_{\tilde{\ell},\mathtt{G}_{\cdot},\dot{y}}(h)$ as $\mathsf{L}_{\tilde{\ell},\mathtt{G}_{\cdot},\dot{y}}(h){=}\mathbb{E}_{(X,s,y):\mathtt{G}_{\cdot}(s,y)}\{\tilde{\ell}(h(X),\dot{y})\}$.*

Note that $\hat{\mathtt{F}}_{\tilde{\ell}}(h){=}|\hat{\mathsf{L}}_{\tilde{\ell},\mathtt{G}_a,\dot{y}}(h){-}\hat{\mathsf{L}}_{\tilde{\ell},\mathtt{G}_b,\dot{y}}(h\}|$ where we have defined $\hat{\mathsf{L}}_{\tilde{\ell},\mathtt{G}_{\cdot},\dot{y}}(h)$ as $\hat{\mathsf{L}}_{\tilde{\ell},\mathtt{G}_{\cdot},\dot{y}}(h){=}\hat{\mathbb{E}}_{(X,s,y):\mathtt{G}_{\cdot}(s,y)}\{\tilde{\ell}(h(X),\dot{y})\}$, i.e., the empirical estimator of $\mathtt{F}_{\tilde{\ell}}(h)$. Note also that if $\mathcal{B}(X){=}X$ ϵ-bias-robustness degenerate in the ϵ-fairness, if we use the Hard loss function $\ell(h(X),y){=}[h(X){\neq}y]$, $\mathtt{G}_{\cdot}(s,y){=}[s{=}\cdot]$, and $\dot{y}{=}1$ we get the Difference of Robust Demographic Parity (DRDP), and if $\ell(h(X),y){=}[h(X){\neq}y]$, $\mathtt{G}_{\cdot}(s,y){=}[s{=}\cdot,y{=}+1]$, and $\dot{y}{=}y$ we get the Difference of Robust Equal Opportunity.

3 Theoretical Mitigation Framework

In this paper, we depict a robustness bias mitigation framework deriving the associated consistency results (i.e., robustness and robustness bias generalization bound). Inspired by the risk minimization principle [16, 17], we consider the problem of minimizing the robust risk under robust bias constraint

$$h^* : \arg\min\nolimits_{h\in\mathcal{H}} \ \mathsf{L}_{\tilde{\ell}}(h), \quad \text{s.t. } \mathtt{F}_{\tilde{\ell}}(h) \leq \epsilon, \tag{1}$$

where $\epsilon \in [0,1]$ is the amount of robustness bias that we are willing to bear since $\mathsf{L}_{\tilde{\ell}}(h)$ and $\mathtt{F}_{\tilde{\ell}}(h)$ are obviously in tension with each other [7]. Since the distribution μ is unknown, Problem 1 cannot be solved and we have to replace the deterministic quantities with their empirical counterparts

$$\hat{h} : \arg\min\nolimits_{h\in\mathcal{H}} \ \hat{\mathsf{L}}_{\tilde{\ell}}(h), \quad \text{s.t. } \hat{\mathtt{F}}_{\tilde{\ell}}(h) \leq \hat{\epsilon}, \tag{2}$$

where $\hat{\epsilon} \in [0,1]$, obtaining the counterpart of the empirical risk minimization principle [16, 17] in the case when they want to maximize the robustness minimizing the robustness bias.

In this section, we will show that h^* and $\hat{h}$ are linked one to another. In particular, if the parameter $\hat{\epsilon}$ is chosen appropriately, we will show that, in a certain sense, the estimator $\hat{h}$ is consistent. In order to present our observations, we require that $\sup_{h\in\mathcal{H}} |\mathsf{L}_{\tilde{\ell}}(h){-}\hat{\mathsf{L}}_{\tilde{\ell}}(h)|{\leq}\mathsf{U}_{\tilde{\ell}}(\delta,n,\mathcal{H})$ with probability at least $1{-}\delta$ that where $\mathsf{U}_{\tilde{\ell}}(\delta,n,\mathcal{H})$ goes to zero as n grows to infinity if the class $\mathcal{H}$ is learnable with respect to the loss [17–19].

Note that if a function is learnable with respect to ℓ it is also learnable with respect to $\tilde{\ell}$ as proved and discussed in [18, 19]. Note also that we could use different losses for L and F which we do not discuss here due to space constraints.

At this point, we can present the main result of this section, namely we can prove[5] the following theorem which proves the consistency of $\hat{\epsilon}$ with respect to

[5] Proof is not reported due to space constraints but the idea is the same behind the work of [16] plus some technicalities presented in [19].

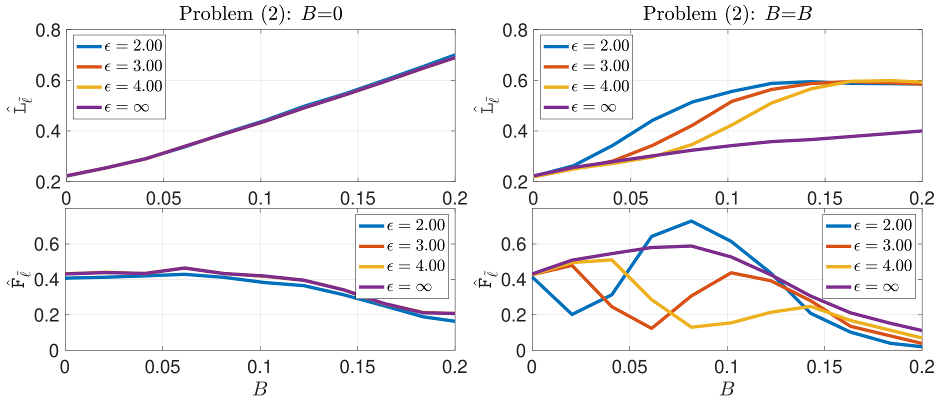

Fig. 1: $\hat{\mathsf{L}}_{\tilde\ell}(h)$ and $\hat{\mathsf{F}}_{\tilde\ell}(h)$ on the test set for the model learned with Problem (2) without (the two graphs on the left) and with (the two graphs on the left) robustness varying ϵ on the Arrhythmia dataset.

$\hat{h}$, namely $\hat{\epsilon}$ tends to $\hat{h}$ as n grows both in terms of robustness and robustness bias.

Theorem 1. *Let $\mathcal{H}$ be a learnable set of functions with respect to the loss function $\tilde\ell$ and let h^* be a solution of Problem* (1) *and let $\hat{h}$ be a solution of Problem* (2) *with $\hat\epsilon = \epsilon + \sum_{s\in\mathcal{S}} \mathsf{U}_{\tilde\ell}(\delta, n_{\mathsf{G}_s}, \mathcal{H})$, where $n_{\mathsf{G}_\cdot} = |\{(X,s,y)\in\mathcal{D} : \mathsf{G}_\cdot(s,y)\}|$. With probability at least $1-6\delta$ it holds simultaneously that*

$$\mathsf{L}_{\tilde\ell}(\hat{h}) - \mathsf{L}_{\tilde\ell}(h^*) \leq 2\mathsf{U}_{\tilde\ell}(\delta, n, \mathcal{H}), \quad \mathsf{F}_{\tilde\ell}(\hat{h}) \leq \epsilon + 2\sum_{s\in\{a,b\}} \mathsf{U}_{\tilde\ell}(\delta, n_{\mathsf{G}_s}, \mathcal{H}). \qquad (3)$$

Note that Theorem 1 clearly explains why some commonly used strategies to mitigate the robustness bias actually work. In fact, most of the current mitigation approaches [7–14], actually rely on the Tikhonov relaxation of Problem (2) where $\hat{h} : \arg\min_{h\in\mathcal{H}} \hat{\mathsf{L}}_{\tilde\ell}(h) + \lambda\hat{\mathsf{F}}_{\tilde\ell}(h)$ with $\lambda \in [0,\infty)$ and where the terms $\hat{\mathsf{L}}_{\tilde\ell}(h)$ and $\hat{\mathsf{F}}_{\tilde\ell}(h)$ are actually relaxed or approximated with different approaches.

4 Empirical Mitigation Framework

In this section, we will instantiate Problem (2) to the case of linear models[6]. In this setting $\mathcal{X}{=}\mathbb{R}^d$, $h(X){=}W{\cdot}X{+}W_0$ for some vector of parameters $W{\in}\mathbb{R}^d$ and the parameter $W_0{\in}\mathbb{R}$. We define $\mathcal{H}$ as an L_2 ball, i.e., $\|W\|_2{=}H{\in}[0,\infty)$. Ideally in Problem (2) we would like to minimize the misclassification error, i.e., $\ell(h(X),y){=}[yh(X){\leq}0]$, with an L_2 ball attach $\mathcal{B}(X){=}\{\tilde{X}{:}\|\tilde{X}{-}X\|_2{\leq}B\}$ with $B{\in}[0,\infty)$ under a particular fairness constraint. For space constraints, we restrict to the DRDP, i.e., $\hat{\mathsf{F}}_{\tilde\ell}(h){=}|\hat{\mathsf{L}}_{\tilde\ell,[s=a],1}(h){-}\hat{\mathsf{L}}_{\tilde\ell,[s=b],1}(h)|$ but the result can be

[6] The extension to Reproducing Kernel Hilbert Space is not reported here for space constraints.

extended to the more general case of the ϵ-bias-robust. In this setting, Problem (2) is NP-hard so intractable. As a consequence we propose a convex relaxation of Problem (2) noting that the empirical robustness can be, in this setting, upper bounded with a convex quantity as follows

$$\hat{\mathsf{L}}_{\tilde{\ell}}(h)\leq\hat{\mathbb{E}}_{(X,y)}\max[0,1-y(W\cdot X+W_0)+HB],$$

where we basically exploited the hinge loss to upper bound the misclassification error. For the DRDP we proceeded analogously

$$\hat{\mathsf{F}}_{\tilde{\ell}}(h)\leq\begin{cases}\hat{\mathbb{E}}_{(X,s=a)}\max[0,1-W\cdot X+W_0+HB]-\hat{\mathbb{E}}_{(X,s=b)}\min[1,-W\cdot X+W_0+HB]\\ \hat{\mathbb{E}}_{(X,s=b)}\max[0,1-W\cdot X+W_0+HB]-\hat{\mathbb{E}}_{(X,s=a)}\min[1,-W\cdot X+W_0+HB]\end{cases}.$$

Note that, substituting the last two results in Problem (2) results in a convex problem.

When $B{=}0$ and $\epsilon{=}\infty$ we get the Linear Support Vector Machine (SVM), when $B{=}0$ and ϵ small enough we get the fair SVM, when $B{\neq}0$ and $\epsilon{=}\infty$ we get the robust SVM, and finally when $B{\neq}0$ and ϵ small enough we get the robust unbiased SVM.

In Figure 1 we report the Empirical Robustness ($\hat{\mathsf{L}}_{\tilde{\ell}}(h)$) and the DRDP ($\hat{\mathsf{F}}_{\tilde{\ell}}(h)$) on the test set (repeating the experiments 30 times) for the model learned with Problem (2) cross-validating W to the optimal value without ($B{=}0$) and with ($B{=}B$ namely the same B has been used both in training and in measuring the performance on the test) robustness varying ϵ on the Arrhythmia dataset [20]. From Figure 1 it is possible to observe that, in order to obtain the best trade-off between $\hat{\mathsf{L}}_{\tilde{\ell}}(h)$ and $\hat{\mathsf{F}}_{\tilde{\ell}}(h)$ in Problem (2), one has to activate both robustness and the constraint, for a specific value of ϵ, supporting the proposed theoretically grounded algorithm.

5 Conclusions

In this work, we investigated the problem that some learned classifiers can be more easily fooled by small perturbations of the input data regarding particular subgroups of the population creating robustness biases that need to be measured and mitigated. For this reason, we first developed a new metric of fairness which generalizes the current ones and degenerates in the classical ones. Then we developed a theoretical mitigation framework with consistency results able to generate a new empirical mitigation strategy and explain why the current ones actually work. Some preliminary results also supported the quality of the proposal. Nevertheless, this work is a first step forward that needs to be more theoretically investigated and supported by stronger empirical evidence also for non-linear and deep models.

References

[1] D. Silver, A. Huang, C. J. Maddison, and Others. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[2] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, and Others. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24:1342–1350, 2018.

[3] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] B. Li, P. Qi, B. Liu, and Others. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[5] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[6] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 55(3):1–44, 2022.

[7] V. Nanda, S. Dooley, and Others. Fairness through robustness: Investigating robustness disparity in deep learning. In *Fairness, Accountability, and Transparency*, 2021.

[8] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, 2021.

[9] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri. On adversarial bias and the robustness of fair machine learning. *arXiv preprint arXiv:2006.08669*, 2020.

[10] A. Fabris, S. Messina, G. Silvello, and Gian A. Susto. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022.

[11] X. Ma, Z. Wang, and W. Liu. On the tradeoff between robustness and fairness. In *Neural Information Processing Systems*, 2022.

[12] C. Tran, K. Zhu, F. Fioretto, and P. Van Henternyck. Fairness increases adversarial vulnerability. *arXiv preprint arXiv:2211.11835*, 2022.

[13] L. E. Richards, E. Raff, and C. Matuszek. Measuring equality in machine learning security defenses. *arXiv preprint arXiv:2302.08973*, 2023.

[14] S. Kamp, A. L. L. Zhao, and S. Kutty. Robustness of fairness: An experimental analysis. In *International Workshops of ECML PKDD*, 2022.

[15] C. Dwork, N. Immorlica, A. T. Kalai, and Others. Decoupled classifiers for group-fair and efficient machine learning. In *Fairness, Accountability and Transparency*, 2018.

[16] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems*, 2018.

[17] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[18] D. Yin, R. Kannan, and P. L. Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, 2019.

[19] L. Oneto, S. Ridella, and D. Anguita. The benefits of adversarial defense in generalization. *Neurocomputing*, 505:125–141, 2022.

[20] H. A. Guvenir, B. Acar, G. Demiroz, and A. Cekin. A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology 1997*, 1997.