Rotation report: Exploring Accuracy vs Robustness vs Fairness Trade-offs with Multi-Objective Machine Learning

Egor Zverev
ISTA

Abstract

Robustness and fairness have been recognized as important properties of machine learning systems. Previous research demonstrates that in practice optimizing for either of them often decreases accuracy of the model. Recent works moved beyond studying pairwise interactions between these properties and accuracy to investigating trade-offs between all three. However, such works usually develop methods that are applicable only to some task- or field-specific definitions of robustness and fairness. We make a step forward and propose to use a multi-objective framework for studying trade-offs between any number of arbitrary robustness and fairness objectives. We use a classical linear scalarization method to estimate Pareto-fronts for several safety objectives. Furthermore, we utilize recent advancements in hypernetworks to compute Pareto-fronts efficiently. We demonstrate our method on several properties, such as individual and group fairness, adversarial and pairwise robustness. We conclude that hypernetworks are as effective as linear scalarization for computing such trade-offs while offering more computational efficiency.

1. Introduction

Traditionally, machine learning models are trained to minimize a single objective, such as an average error on a training dataset. There are, however, other valuable quantities of interest. In particular, research on robustness (Goodfellow et al., 2015) and fairness (Hardt et al., 2016) has been recognized as beneficial for addressing societal consequences of deploying machine learning models (Hendrycks et al., 2021). Previous works suggest that enforcing either robustness or fairness is in conflict with achieving accuracy as a primary optimization objective (Zhao and Gordon, 2019; Tsipras et al., 2019). While it is important to study pairwise interactions between either robustness and accuracy or fairness and accuracy, it is also desirable to consider all three quantities together. Recently, a number of studies (Xu et al., 2021; Pruksachatkun et al., 2021; Benz et al., 2021) investigated interactions between all three, however, these works develop methods that are applicable only to some particular task- or field-specific notions of fairness and robustness. It is therefore desirable to develop a more general method for studying trade-offs between robustness and fairness.

We propose to use a multi-objective framework (Zitzler and Thiele, 1999) for building such a method. Multi-objective machine learning is concerned with simultaneously minimizing several loss functions. As finding solutions that minimize all given objectives is often impossible, a Pareto-front is estimated using approaches such as linear scalarization (Geoffrion, 1968) or Chebyshev scalarization (Kaisa, 1999). However, these methods require exhaustive computations over the space of all possible choices of scalarization weights. This

flaw, however, can be overcome by using Pareto HyperNetworks (Navon et al., 2021) which are capable of learning the entire Pareto-front simultaneously in a single training procedure.

The main contribution of this work can be summarized as following:

- We propose a framework for finding trade-offs between any number of arbitrary robustness and fairness objectives using multi-objective optimization.
- We apply this method to compute trade-offs between various combinations of fairness and robustness objectives, such as individual and group fairness, adversarial and pairwise robustness.
- We find Pareto-fronts by solving separate optimization problems for multiple linear scalarization (LS) objectives. We also apply hypernetworks and conclude that hypernetworks are as effective as exhaustive LS search for finding trade-offs between fairness and robustness objectives, despite being significantly more efficient.

2. Previous work

2.1 Fairness

Fairness was initially explored by social and philosophical sciences in an attempt to answer questions such as "What is equality of educational opportunities?" (Coleman, 1967). As machine learning became frequently deployed in sensitive decision-making settings, several mathematical notions of fairness arised. Group fairness strives to achieve parity of some statistical measurement across protected groups. Of particular note are concepts of disparate impact (Feldman et al., 2015), demographic parity (Calders and Verwer, 2010), equalized odds and equality of opportunity (Hardt et al., 2016). Contrary to group fairness, individual fairness requires that "similar individuals are treated similarly" (Dwork et al., 2012). All of these definitions remain important for the field, as each of them is useful for some particular setting (Verma and Rubin, 2018).

2.2 Fairness vs Accuracy

While it is not mandatory for accuracy to decrease with increased fairness (Dutta et al., 2020), the existence of trade-off between fairness and accuracy in practice has been noted many times in fairness literature (Calmon et al., 2017; Kusner et al., 2017; Menon and Williamson, 2018). Therefore, researchers strive to build fair machine learning models with sacrificing as little accuracy as possible (Valdivia et al., 2021). Multi-objective perspective is often used to examine Pareto-fronts between fairness and accuracy (Berk et al., 2017; Liu et al., 2022; Ge et al., 2022; Kamani et al., 2021; Liang et al., 2022). We build on top of these works to compute Pareto-fronts between fairness, accuracy and robustness.

2.3 Robustness

In the most general sense, robustness is the capability of machine learning system to endure extreme, unusual or adversarial events (Hendrycks et al., 2021). So far most of the research has been focused on robustness to adversarial attacks (Goodfellow et al., 2015; Dalvi et al., 2004; Biggio and Roli, 2018; Madry et al., 2018; Szegedy et al., 2014). It has been

shown that small synthetic perturbations of a model's output might lead to misclassification (Szegedy et al., 2014). A dominant strategy of overcoming this problem is adversarial training (Bai et al., 2021; Zhao et al., 2022), which is training a model with adversarial loss that accounts for possible perturbations of input data with respect to some l_p-norm. Recent works seem to focus on either improving adversarial training mechanisms (Bai et al., 2021), developing better understanding the phenomena of adversarial attacks (Zhang et al., 2022) or considering unusual types of adversarial attacks besides l_p-pertrubations (Kang et al., 2019).

2.4 Robustness vs accuracy

While optimal robustness and accuracy could be achieved simultaneously under some theoretical conditions (Yang et al., 2020), in practice such conditions are often not met. Thus, many works in robustness agree there is a trade-off between robustness and accuracy (Tsipras et al., 2019; Zhang et al., 2019; Su et al., 2018). Recent papers explore ways of mitigating decrease in accuracy for robust models (Rade and Moosavi-Dezfooli, 2022; Raghunathan et al., 2020).

2.5 Robustness vs Fairness

A number of works have investigated interaction between robustness and fairness. For instance, Pruksachatkun et al. (2021) demonstrated that in Natural Language Processing enforcing word substitution robustness improves fairness, and vice versa. As for adversarial robustness, often class-wise disbalance (mismatch in accuracy between different classes) is used as a definition of fairness (Xu et al., 2021; Benz et al., 2021; Tian et al., 2021). In this case, the research focus is on modifying adversarial training in a way that would help to achieve similar accuracy on all classes. Another notion of fairness in robustness is robustness bias introduced by Nanda et al. (2021) who argues that often data contains groups (sometimes based on gender, race, etc.) that are particularly vulnerable to adversarial attacks.

The most similar work to our own is Benz et al. (2021) that explores trade-offs between accuracy, robustness and fairness. However, the authors use class disbalance as a notion of fairness.

3. Multi-objective optimization for fairness and robustness

In the beginning of this section we will introduce fairness and robustness objectives that we will later use to demonstrate our method. As we strive to build a framework for computing trade-offs between any number of safety objectives, we will use several notions of robustness and fairness. For the rest of this work we will adopt demographic parity (Calders and Verwer, 2010) as a definition of group fairness. We will follow Berk et al. (2017) to enforce individual fairness in the form of regularization for better compatibility with multi-objective optimization instead of employing original constrained-based methods (Dwork et al., 2012).

For dataset $D = S_1 \cup S_2$ of two groups S_1 and S_2 and model f_θ parameterized by θ we define the group fairness loss as:

$$L_G(\theta) = (\frac{1}{n_1}\sum_{x \in S_1} f_\theta(x) - \frac{1}{n_2}\sum_{x \in S_2} f_\theta(x))^2$$

and the individual fairness loss as:

$$L_I(\theta) = \frac{1}{n_1 \cdot n_2} \sum_{(x_i,y_i)\in S_1,(x_j,y_j)\in S_2} I\{y_i = y_j\}(f_\theta(x_i) - f_\theta(x_j))^2$$

Where $n_1 = |S_1|$ and $n_2 = |S_2|$

As for robustness, we will employ traditional adversarial training loss (Madry et al., 2018) as well as alternative pairwise ϵ-loss.

For a dataset $D = (X, Y)$ and loss function L we define the adversarial ϵ-loss as:

$$L_A(\theta) = \frac{1}{n}\sum_{(x,y)\in D} \max_{\delta:||\delta||<\epsilon} L(f_\theta(x+\delta), y)$$

and the pairwise ϵ-loss as:

$$L_P(\theta) = \frac{1}{n}\sum_{x} \mathbf{E}_{\delta:||\delta||<\epsilon} L(f_\theta(x+\delta), f(x))$$

3.1 Multi-objective machine learning

Multi-objective optimization is concerned with the simultaneous minimization of several objective functions $f_1(\theta), ..., f_n(\theta)$. As there is often no optimal point θ that minimizes all given functions simultaneously, the goal is to find Pareto-optimal solutions (Censor, 1977).

For multiple objectives $f_1(\theta), ..., f_n(\theta)$ we say that θ strongly dominates $\hat{\theta}$ if for each objective f_i it holds that $f_i(\theta) \leq f_i(\hat{\theta})$ and there is at least one objective f_j such that $f_j(\theta) < f_j(\hat{\theta})$.

A point θ is called a Pareto-optimal solution if there is no $\hat{\theta}$ that strongly dominates θ. Pareto-optimal solutions form Pareto-front.

One of the simplest ways of finding Pareto-optimal solutions is a search with LS objectives (Geoffrion, 1968). Multiple objectives are replaced with a single one by taking a linear combination of objective functions:

$$\sum_{i=1}^{n} |r_i \cdot f_i(\theta)| \rightarrow min_\theta$$

Point θ obtained by solving this minimization problem for a fixed set of weights $\{r_i\}_i$ is a Pareto-optimal solution (Geoffrion, 1968). For convex optimization problems the full Pareto-front could be obtained by considering all possible combinations of nonnegative weights $\{r_i\}_i$ such that $\sum_i r_i = 1$ (Censor, 1977).

We suggest applying multi-objective optimization for studying trade-offs between accuracy, robustness and fairness. One problem with using a naive LS-based approach is computational cost. Obtaining each Pareto-optimal solution requires solving a separate optimization problem. Since the number of the points that have to be considered grows

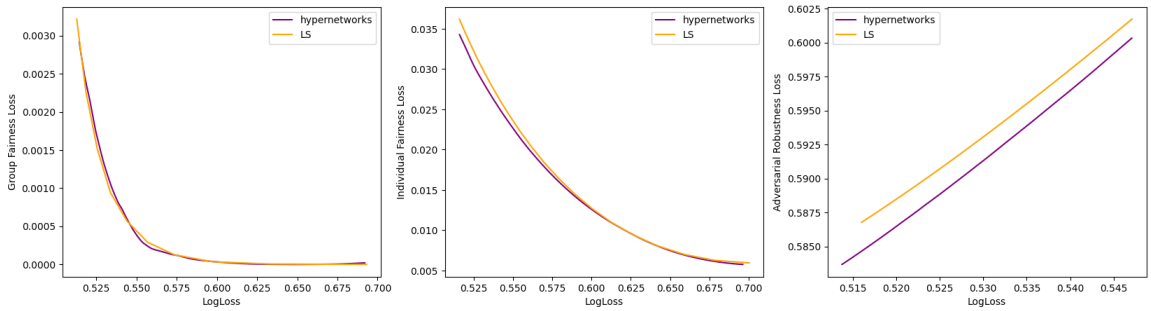

Figure 1: Logloss vs individual fairness, logloss vs group fairness and logloss vs adversarial robustness trade-offs. Comparison of LS and hypernetworks.

exponentially with respect to the number of objectives, estimating a full Pareto-front for a large number of objectives with LS requires impractical amounts of computations.

3.2 Pareto Hypernetworks and fairness vs robustness trade-offs

The problem with exhaustive computations, however, can be overcome by using Pareto HyperNetworks (Navon et al., 2021). Hypernetworks (Ha et al., 2016) are models that predict parameters of another model. They can be used for estimating Pareto-fronts efficiently, as proposed by Navon et al. (2021).

With the LS-based approach, one would fix a combination of weights $\{r_i\}_i$, minimize the objective $\sum_{i=1}^{n} |r_i \cdot f_i(\theta)|$ and then repeat this procedure for many combinations of weights. An alternative approach is to define a hypernetwork $h(r, \phi)$ that takes weights r as an input. The hypernetwork is trained to predict parameters of the original model $\theta = h(r, \phi)$ that should minimize a linear combination of training losses $\{l_i\}_i$ with weights $\{r_i\}_i$. Navon et al. (2021) propose to use a modification of stochastic gradient descent to train hypernetworks. At each iteration weights r are sampled from the Dirichlet distribution with parameter α. Then r is used as a hypernetwork input to predict parameters of the original model θ. The losses with respect to parameters ϕ of the hypernetwork $h(\cdot, \phi)$ are then computed on a batch of training data.

Function `Train Hypernetwork`**:**
  while *not converged* **do**
    $r \sim Dir(\alpha)$
    $\theta(r, \phi) = h(r, \phi)$
    Sample minibatch$(x_1, y_1), ..., (x_B, y_B)$
    $g_\phi = \frac{1}{B} \sum_{i,j} r_i \nabla_\phi l_i(x_j, y_j, \theta(r, \phi))$
    $\phi \leftarrow \phi - \eta g_\phi$
  return ϕ

Navon et al. (2021) demonstrate the effectiveness of their method on several benchmarks. In our work we take one step further and study trade-offs between robustness and fairness by relying on the methodology proposed by Navon et al. (2021).

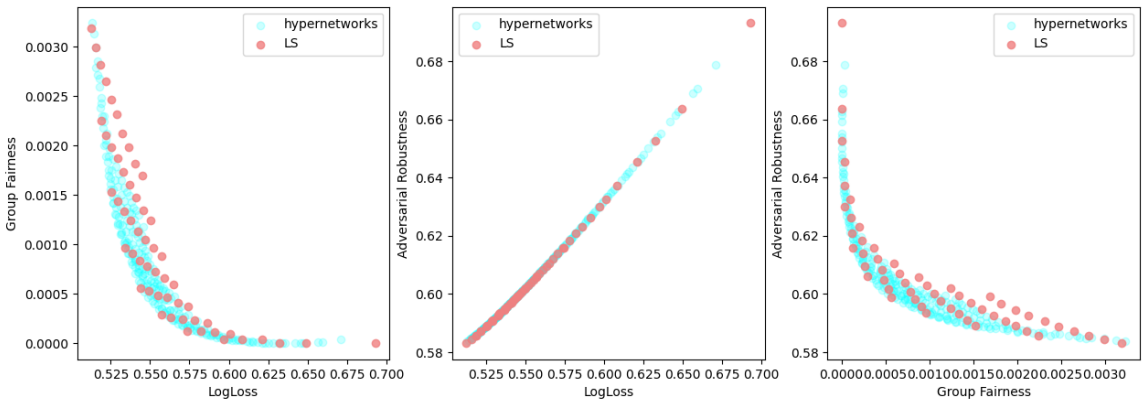

Figure 2: Logloss vs group fairness vs adversarial robustness trade-offs, projected in 2 dimensions. Comparison of LS and hypernetworks.

4. Experiments

In our experimental section we show how to use the proposed methodology to compute trade-offs between various combinations of two, three or four losses. We use a logistic regression model to predict binary salary labels on a portion of the folktables dataset (Ding et al., 2021). "Sex" is used as a protected attribute for fairness. When applying hypernetworks, we use a 3-layer MLP to predict parameters of the main linear model.

4.1 2D Experiments

In this section we demonstrate our method for computing trade-offs between two objectives, one of which is always logloss. The other one is a safety objective, such as group and individual fairness or adversarial robustness. The resulting plots are shown on Figure 1.

Trade-offs between the used notions of fairness and logloss are consistent with existing literature. Indeed, achieving a lower value of fairness would require a sacrifice in logloss, and vice versa. As for adversarial robustness, adversarial loss is positively correlated with logloss for this particular experiment.

We observe that both hypernetworks and LS methods provide almost identical trade-offs, which leads us to a conclusion that hypernetworks are as effective in 2D as the LS-based approach.

4.2 3D Experiments

It is important to move beyond pairwise cases and consider trade-offs between triples of objectives. We demonstrate our framework for estimating trade-offs between logloss, adversarial robustness loss and group fairness loss. For the sake of visualization, we project 3-dimensional Pareto-fronts in two dimensions and compare the resulting projections between hypernetworks and LS approaches.

The graphs for LS are more sparse, as getting each single point requires solving a separate optimization procedure, while for hypernetworks the whole graph can be computed after

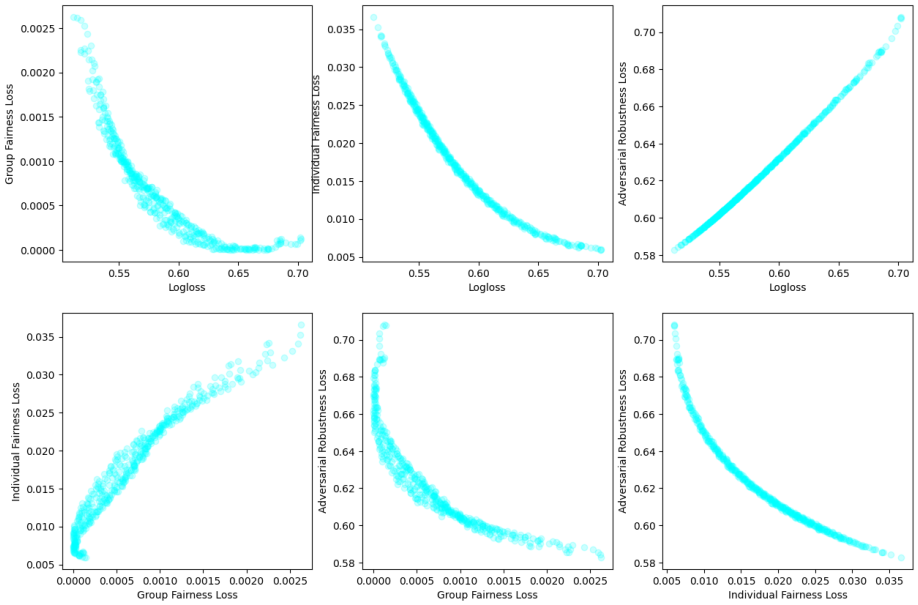

Figure 3: Logloss vs group fairness vs individual fairness vs adversarial robustness trade-offs, projected in 2 dimensions. Trade-offs are computed using hypernetworks only.

training only one model. Despite these differences, clearly both methods produce extremely similar curves. This suggests that both LS-based search and hypernetworks methods are equally effective for finding logloss vs robustness vs fairness trade-offs.

4.3 4D Experiments

In this section we demonstrate that the proposed framework could be used for finding trade-offs between 4 competing objectives, namely individual and group fairness, logloss and adversarial robustness. This time we avoid direct comparison with LS due to computational costs. As before, we project the resulting 4-dimensional Pareto-front in 2-dimensional subspaces (Figure 3). Even in 4 dimensions, trade-offs could still be computed very efficiently with hypernetworks, while doing the same with LS would require a lot of additional resources.

5. Future work

As this study comes with many limitations (extremely simple model, usage of only one dataset and a few selected objectives), we would like to see expansion of this work in the following directions:

- **More datasets:** we explored only one type of problem (binary classification) for one particular set of tabular data (folktables). It is highly desirable to consider other datasets as well.

- **More sophisticated primary model:** It is possible that hypernetworks are as efficient as exhaustive LS search due to the simplicity of the primary model we used (logistic regression). It is thus necessary to perform experiments for more sophisticated primary models.

- **More objectives:** in this study we considered various forms of robustness and fairness. By introducing additional safety objectives (i.e. privacy) we will be able to see the effectiveness of our framework for other safety properties. Furthermore, while we expect this framework to generalize well beyond 4 dimensions, it is not guaranteed and requires verification.

6. Conclusion

Fairness and robustness have been extensively studied, but only recently together. We contribute to this research direction by proposing a framework for computing trade-offs between fairness and robustness. The proposed method is applicable to any number of arbitrary notions of fairness and robustness. We rely on the existing methodology of multi-objective machine learning and hypernetworks to get trade-offs between these quantities. We apply this methodology to a set of safety objectives, such as group and individual fairness and adversarial robustness and confirm its effectiveness. Furthermore, we demonstrate that hypernetworks are as effective as a classical search with linear scalarization objective for computing trade-offs between fairness and robustness, despite being more efficient.

References

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Philipp Benz, Chaoning Zhang, Soomin Ham, Gyusang Karjauv, Adil Cho, and In So Kweon. The triangular trade-off between accuracy, robustness, and fairness. *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV) at CVPR*, 2021.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yair Censor. Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization*, 4(1):41–59, 1977.

James S Coleman. The concept of equality of educational opportunity. *The International Handbook of Educational Research in the Asia-Pacific Region*, 1967.

Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *International Conference on Machine Learning (ICML)*, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2022.

Arthur M Geoffrion. Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications*, 22(3):618–630, 1968.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. *International Conference on Learning Representations (ICLR)*, 2016.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021.

Miettinen Kaisa. *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research and Management Science*. 1999.

Mohammad Mahdi Kamani, Rana Forsati, James Z. Wang, and Mehrdad Mahdavi. Pareto efficient fairness in supervised learning: From extraction to tracing. *arXiv preprint arXiv:2104.01634*, 2021.

Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. *arXiv preprint arXiv:1908.08016*, 2019.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Neural Information Processing Systems (NIPS)*, 2017.

Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy. In *ACM Conference on Economics and Computation (EC)*, 2022.

Suyun Liu, Frank E., Curtis, Fliege Jorg, and Xie Sihong. *Stochastic Multi-Objective Optimization and Its Application to Fairness in Machine Learning*. PhD thesis, 2022.

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 (ICLR)*, 2018.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*, 2018.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

Aviv Navon, Aviv Shamsian, Gal Chechik, and Ethan Fetaya. Learning the pareto front with hypernetworks. In *International Conference on Learning Representations (ICLR)*, 2021.

Yada Pruksachatkun, Satyapriya Krishna, Jwala Dhamala, Rahul Gupta, and Kai-Wei Chang. Does robustness improve fairness? Approaching fairness with word substitution robustness methods for text classification. In *ACL International Joint Conference on Natural Language Processing (IJCNLP)*, 2021.

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations (ICLR)*, 2022.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2020.

Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? - A comprehensive study on the robustness of 18 deep image classification models. In *European Conference on Computer Vision (ECCV)*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of class-wise robustness in adversarial training. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2021.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

Ana Valdivia, Javier Sánchez-Monedero, and Jorge Casillas. How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, 36(4):1619–1643, 2021.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *International Workshop on Software Fairness(FairWare)*, 2018.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning (ICML)*, 2021.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Adversarial robustness through the lens of causality. In *International Conference on Learning Representations (ICLR)*, 2022.

Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Weimin Zhao, Sanaa Alwidian, and Qusay H. Mahmoud. Adversarial training methods for deep learning: A systematic review. *Algorithms*, 15(8), 2022.

E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3 (4):257–271, 1999.