Stochastic Fairness Interventions Are Arbitrary

MATTIA CERRATO*, Johannes Gutenberg University, Germany
MARIUS KÖPPEL*, ETH, Switzerland
KIARA STEMPEL*, Johannes Gutenberg University, Germany
PHILIPP WOLF*, Johannes Gutenberg University, Germany
STEFAN KRAMER, Johannes Gutenberg University, Germany

Bias mitigation techniques offer the opportunity to intervene on statistical models so to reduce the risk that these will discriminate towards certain groups. These techniques rely on learning a mapping from the sensitive data S to some decision variable $\hat{Y}$, usually mediated by the non-sensitive covariates X. Some of the methods available in this space propose to learn a stochastic mapping, which has several theoretical benefits from a computational perspective: namely, randomization makes it possible to compute certain mitigation objectives, and widens the search space for "optimal" models. From the perspective of procedural fairness, however, stochastic mappings may imply arbitrary decisions. In this paper, we study and discuss the distribution of arbitrariness in popular randomized bias mitigation techniques which are currently available in standard fairness toolkits. We find that individuals belonging to different groups may have different risks for arbitrariness; furthermore, we observe different patterns of arbitrariness for different randomized mitigation strategies, and discuss possible causes for this general phenomenon.

Keywords: fairness, bias mitigation, stochasticity, randomized algorithms

Reference Format:
Mattia Cerrato, Marius Köppel, Kiara Stempel, Philipp Wolf, and Stefan Kramer. 2025. Stochastic Fairness Interventions Are Arbitrary. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 7 pages.

1 Introduction and Related Work

Research on bias mitigation focuses on designing methodologies that may perform a given ML task while decreasing the risk of discrimination towards certain relevant social groups. Techniques in this space deal with various classic ML tasks such as classification, regression, and ranking. Today, several bias mitigation software libraries are available [20, 3] and methods will either preprocess [12] the data, postprocess [10] the decisions or include a bias mitigation objective in the cost function of an existing classifier [21, 16], regressor [19] or ranker [23, 22, 4], an idea named in-processing.

*These authors contributed equally to this research. Alphabetical order.

Authors' Contact Information: Mattia Cerrato, mcerrato@uni-mainz.de, Johannes Gutenberg University, Mainz, Germany; Marius Köppel, mkoepp@phys.ethz.ch, ETH, Zurich, Switzerland; Kiara Stempel, Johannes Gutenberg University, Mainz, Germany, kiara.stempel@uni-mainz.de; Philipp Wolf, Johannes Gutenberg University, Mainz, Germany, pwolf01@students.uni-mainz.de; Stefan Kramer, Johannes Gutenberg University, Mainz, Germany, kramer@informatik.uni-mainz.de.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

In this paper, we analyze bias mitigation methods which propose to learn a *stochastic* intervention. These methods map feature values to decision variables while including randomness in some fashion. This idea is not necessarily limited to fairness interventions and bias mitigation. It is straightforwardly possible to obtain a "soft" stochastic binary classifier by sampling from a Bernoulli distribution where the probability of heads, for instance, is set to be equal to a normalized risk score. This is in contrast with a "hard" deterministic classifier, which obtains its decision rule via a hard threshold computed on the risk scores given by the base method. Soft classifiers may be, for instance, helpful in achieving individual fairness, where otherwise a hard decision threshold would mean, by definition, that similar individuals across the decision boundary would obtain very different results, for instance, of access or rejection to some resource.

Stochasticity is an attractive feature in bias mitigation strategies that target group fairness definitions. This is due to the isomorphism between group definitions and mutual information over the relevant random variables. For instance, the independence definition of group fairness in classification requires that $\hat{Y} \perp S$, that is, statistical independence of the two random variables, where $\hat{Y}$ is the decision variable and S a relevant sensitive attribute. Equivalent to statistical independence is that $I(\hat{Y};S) = 0$, the mutual information between the two variables which is 0 if and only if they are independent [6]. This property makes the minimization of mutual information a natural training objective for bias mitigation strategies. However, when two random variables are related by a deterministic function, their mutual information can be ill-defined [9] and in particular infinite when they are continuous. Stochasticity is then an attractive proposition to make it computable and is often pursued in representation learning approaches [24, 15], even if an alternative solution is provided by discretization of one or both variables [5]. Another bias mitigation strategy that makes use of stochasticity is due to Agarwal et al. [1], where bias mitigation is cast as a constrained optimization problem. The problem is then solved by sampling multiple times over some hypothesis class $\mathscr{H}$; critically, the authors propose to learn the optimal distribution Q over $\mathscr{H}$ so that sampling from Q will return, on expectation, the optimal constrained classifier $h \in \mathscr{H}$ for a given classification problem. The authors' rationale is that allowing for randomized classifiers creates a larger search space: the procedure can then be proven to recover the optimal constrained classifier in expectation (see [1], Appendix).

These computational benefits may stand in contrast with basic ideas in process fairness: in particular, Leventhal posed [14] the concept of *consistency* as the idea that "allocative procedures should be consistent across persons and over time". When the mapping $\hat{Y} = f(X)$ is stochastic, the final classification decision may be made partly randomly and thus not consistently, especially if the entropy of $\hat{Y}$ is not distributed equally across individuals.

In this paper, we study how different stochastic fairness interventions affect *decision entropy* and how it is distributed across individuals. Our main finding is that randomized bias mitigation techniques display *disparate arbitrariness* across individuals and groups. We study the distribution of arbitrariness and find that it displays a complex pattern. While sensitive attributes partially explain its variance, they do not offer an exhaustive description of this phenomenon. In particular, different methodologies display different stochastic behaviors. While arbitrariness and uncertainty in bias mitigation have recently been proposed as fairness metrics themselves [13], the present paper is the first one to discuss specifically how stochastic bias mitigation impacts arbitrariness and decision entropy. Our objective with this study is to raise awareness to how stochastic bias mitigation may disparately impact the distribution of randomness and thus, arbitrariness: at this time, we however refrain from stating normatively that randomness should be avoided in bias mitigation.

2 Experimental Methodology

To investigate the distribution of stochasticity in bias mitigation, we set up a "fair classification" task where we employ three separate methodologies. In the following, we will refer to $X,Y,\hat{Y}$ and S as the features, ground truth labels, model predictions, and sensitive attributes respectively. We obtain a binary $\hat{Y}$ via a hard threshold on 0.5 on the (randomized) 0/1 risk score outputted by the underlying model f_θ. Then, $\hat{Y}$ depends on the underlying randomized intervention considered, which we describe in the following.

(1) **Stochastic representation learning**: Feature representations Z are learned stochastically from X, i.e., $Z \sim P(Z \mid X)$, and decisions are made based on these stochastic representations. The process may succinctly be described via the Markov chain $X \to Z \to \hat{Y}$, abstracting away some methodology-specific details. For instance, Zemel et al. and Louizos et al. [24, 15] propose to learn an auto-encoding stochastic mapping $X \to Z \to \hat{X}$ and then a deterministic mapping to $\hat{Y}$. In our experimentation, we employ the randomized model proposed in [5], **BinaryMI**, and the variational autoencoder introduced in [18], **ICVAE**. We employ the implementations given by the respective authors.

(2) **Hypothesis space distributions**: In the constrained optimization framework of Agarwal et al. [1], classifiers are sampled at inference time from a learned distribution $Q(h)$, where $h \in \mathscr{H}$ is an hypothesis in the model class, that is, an individual model. Nonetheless, the actual model employed is sampled from $Q(h)$ at every inference step and thus the mapping $X \to \hat{Y}$ is randomized. We take the implementation of **ExpGrad** available in `fairlearn` [20].

All three methods allow to regulate the strength of the intervention via a hyperparameter. For stochastic representation learning, we take $\gamma \in \{0, 0.1, \dots 1\}$ where lower values imply a weaker intervention on the invariance of the representation Z w.r.t. S. For fair reductions, we take the constraint $\varepsilon = \{0, \dots 0.1\}$ at 10 equispaced intervals and enforce a demographic parity constraint so that $\left|P(\hat{Y}=1 \mid S=0) - P(\hat{Y}=1 \mid S=1)\right| \leq \varepsilon$. To study the distribution of arbitrariness, we then obtain a sample of $\hat{Y}$ by performing inference 1000 times for each data point $x \in X$. Then, we are able to compute the Shannon entropy for the decisions relative to each data point $H(\hat{Y}) = -\sum_{\hat{y} \in \hat{Y}} \hat{y}\, log(\hat{y})$, across all three models and six classification datasets: Adult, COMPAS, German, Banks and Folktables (HI, AK) [7], which we describe more concretely in the Appendix.

3 Results and Discussion

We summarize here some results of our investigation, which we do not give fully due to space constraints.

Disparate arbitrariness. We observe that stochastic, randomized interventions distribute randomness differently across individuals in Figure 1. Decision entropy histograms typically show a peak at 0 entropy; however, we observe consistently across models and datasets that other points in the dataset display non-zero entropy. We report a selection of other decision entropy histograms in the Appendix. From Figure 1, we note that **ExpGrad** exhibits a distinctive pattern where a more relaxed constraint (higher ε) implies lower arbitrariness (lower entropy mass).

Sensitive attribute values or intervention strength do not consistently explain arbitrariness. We show how entropy changes across values of γ and ε in Figure 2, where we plot $H(\hat{Y}|S)$. For **ExpGrad**, we observe minor differences in decision entropy across groups. We do not observe this pattern, however, across all models and datasets. We note that some models display higher decision entropy than others.

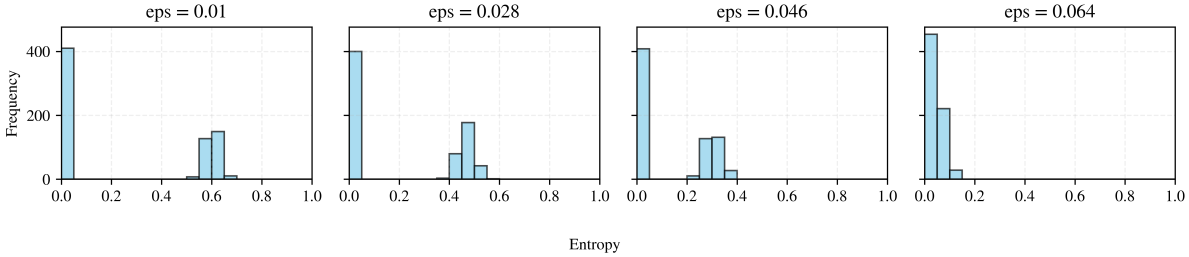

Fig. 1. Mean entropy of predictions $\hat{Y}$ for **ExpGrad** on folktables (AK).

Predicting decision entropy is difficult. We train a linear regression model to predict $H(\hat{Y})$ from X and S. The rationale here is to understand whether it is possible to describe which individuals are more at risk of high decision entropy via their feature or group membership. We report the MSE and R-squared value of the resulting model in the Appendix, Table 1. This approach is not consistently successful across models, datasets and intervention strength, be it representation invariance γ or maximum demographic parity violation ε (Table 2, Appendix).

Our experimental results thus confirm that i) the distribution of arbitrariness in stochastic interventions is disparate across individuals, but that ii) the phenomenon is not easily reducible to any specific issue with a singular intervention strategy or dataset. In future work, we plan to investigate the arbitrariness of these interventions strategy further by analyzing whether, for instance, there are systematic differences between the arbitrariness displayed by fair representation learning [5, 18] and constrained optimization techniques [1]. Another avenue for future applied research is understanding how and when to intervene to reduce stochasticity disparity across individuals.

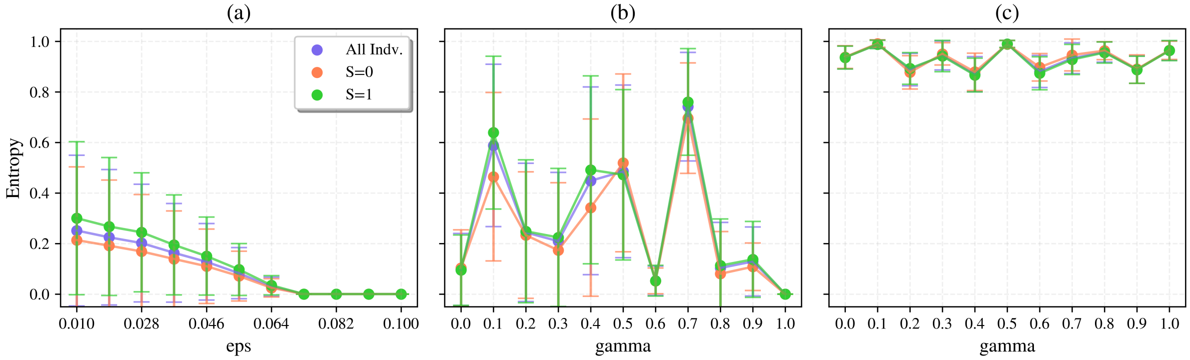

Fig. 2. Entropy of predictions for (a) **BinaryMI** on Adult, (b) **ICVAE** on Adult, (c)**ExpGrad** on Folktables (AK).

References

[1] A. Agarwal et al. “A reductions approach to fair classification”. In: *International conference on machine learning*. PMLR. 2018, pp. 60–69.

[2] J. Angwin et al. *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2023-04-24. 2016.

[3] R. K. Bellamy et al. “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias”. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[4] M. Cerrato et al. “Fair pairwise learning to rank”. In: *DSAA* (2020).

[5] M. Cerrato et al. “Invariant representations with stochastically quantized neural networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 6. 2023, pp. 6962–6970.

[6] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[7] F. Ding et al. “Retiring Adult: New Datasets for Fair Machine Learning”. In: *Advances in Neural Information Processing Systems* 34 (2021).

[8] D. Dua et al. *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml. Accessed: 2023-04-24. 2017.

[9] Z. Goldfeld et al. “The information bottleneck problem and its applications in machine learning”. In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 19–38.

[10] M. Hardt et al. “Equality of opportunity in supervised learning”. In: *Advances in neural information processing systems* 29 (2016).

[11] H. Hofmann. *Statlog (German Credit Data)*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77. 1994.

[12] F. Kamiran et al. “Classifying without discriminating”. In: *ICCC*. 2009.

[13] S. Kuzucu et al. “Uncertainty as a Fairness Measure”. In: *Journal of Artificial Intelligence Research* 81 (2024), pp. 307–335.

[14] Leventhal, Gerald S. “What should be done with equity theory? New approaches to the study of fairness in social relationships”. In: *Social exchange: Advances in theory and research* (1980), pp. 27–55.

[15] C. Louizos et al. “The Variational Fair Autoencoder”. In: *ICLR* (2016).

[16] D. Madras et al. “Learning Adversarially Fair and Transferable Representations”. In: *CoRR* abs/1802.06309 (2018). arXiv: 1802.06309. URL: http://arxiv.org/abs/1802.06309.

[17] S. Moro et al. “A Data-Driven Approach to Predict the Success of Bank Telemarketing”. In: *Decision Support Systems* 62 (June 2014).

[18] D. Moyer et al. “Invariant representations without adversarial training”. In: *NeurIPS* (2018).

[19] H. Narasimhan et al. “Pairwise Fairness for Ranking and Regression”. In: *AAAI* (2020).

[20] H. Weerts et al. “Fairlearn: Assessing and Improving Fairness of AI Systems”. In: *Journal of Machine Learning Research* 24.257 (2023), pp. 1–8. URL: http://jmlr.org/papers/v24/23-0389.html.

[21] Q. Xie et al. “Controllable Invariance through Adversarial Feature Learning”. In: *NeurIPS* (2017).

[22] K. Yang et al. “A nutritional label for rankings”. In: *Proceedings of the 2018 international conference on management of data*. 2018, pp. 1773–1776.

[23] M. Zehlike et al. “Reducing disparate exposure in ranking: A learning to rank approach”. In: *WWW* (2019).

[24] R. Zemel et al. “Learning fair representations”. In: *ICML* (2013).

A Datasets

COMPAS. Introduced by ProPublica [2], COMPAS focuses on evaluating the risk of future crimes among individuals previously arrested, a system commonly used by US judges. The ground truth is whether an individual is arrested within the following two years since the evaluation. The sensitive attribute is ethnicity.

Adult. This dataset, available in the UCI repository [8], pertains to determining whether an individual’s annual salary exceeds $50,000. We take gender to be the sensitive attribute [15, 24].

Bank marketing. Here, the classification task involves predicting whether an individual will subscribe to a term deposit. This dataset (called Banks in the following) exhibits disparate impact and disparate mistreatment concerning age, particularly for individuals under 25 and over 65 years old. [17]

German. The German Credit dataset, contains credit data of individuals with the objective of predicting their credit risk as either high or low risk. The gender of the individuals serves as the sensitive attribute [11].

Folktables. The folktables datasets are a collection of datasets derived from US cencus data, which span multiple years and all states of the USA [7]. Although the dataset supports multiple prediction tasks, we only used the

ACSIncome task, in which the objective is to predict whether an individual´s income exceeds 50.000$. The sensitive attribute is the race of the individual.

B Additional Results

Figure 3 provides three histograms of decision entropies $H(\hat{Y})$ at various intervention intensities. Here, we note that **ExpGrad** shows a pattern of having higher entropy when the intervention is more intense (lower ε). The pattern is not present in **BinaryMI** and **ICVAE**.

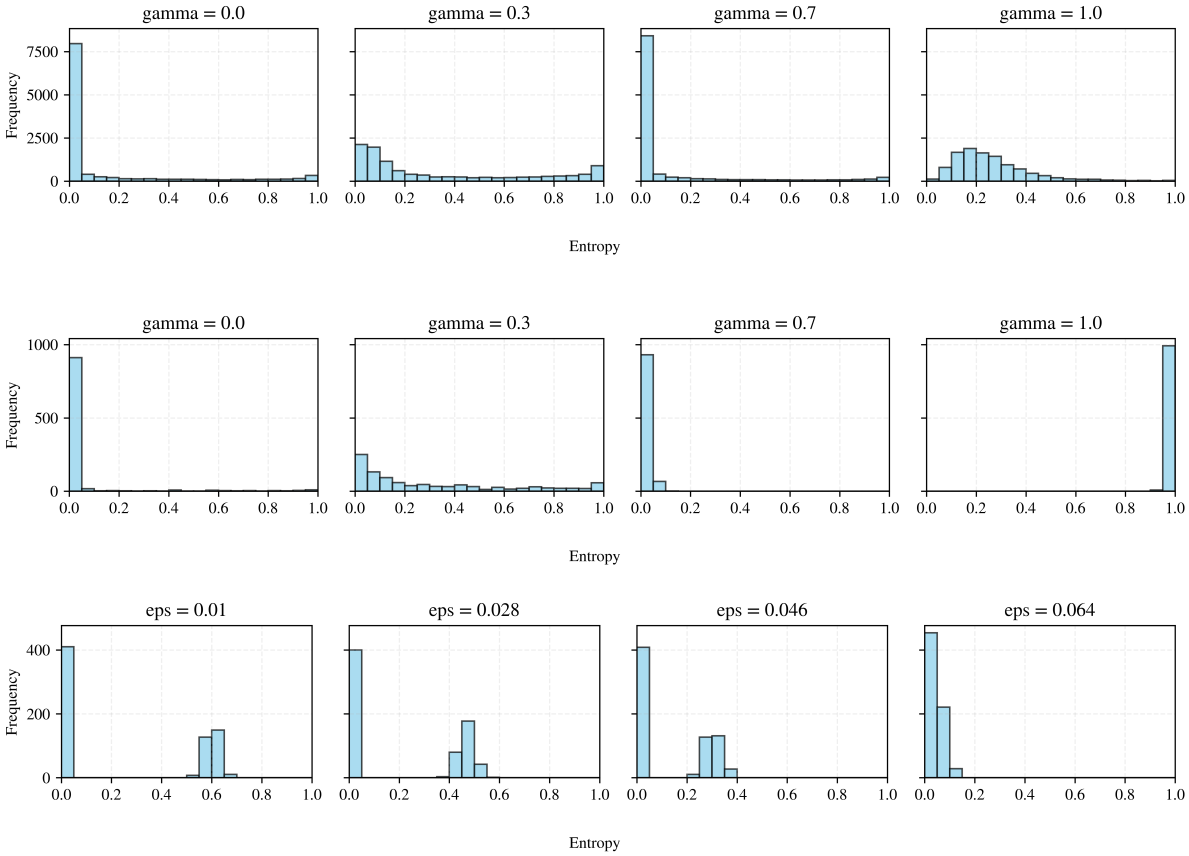

Fig. 3. Mean entropy of predictions $H(\hat{Y})$ for ExponentiatedGradient on folktables_AK, BinaryMI on German, and ICVAE on German.

Tables 1 and 2 show the performance of a linear regression model when predicting the decision entropy $H(\hat{Y})$ from X and S. We observe that MSE and R^2 are very high in most cases, which indicates a failure in being able to predict who is most at risk of arbitrary decisions. Even when the metrics are promising (**ICVAE** on the German dataset), we observe that these metrics fluctuate depending on the strength of the intervention (Table 2).

Table 1. Mean MSE and mean R^2 values for all models and datasets, averaged over all γ or ε values the models were trained on.

Dataset	Model	Mean MSE	Mean R^2
Adult	BinaryMI	>1000	-1.671265972605837e+18
	ExponentiatedGradient	0.07	0.13
	ICVAE	>1000	-1.5250547260659496e+20
Banks	BinaryMI	>1000	-1.2227596374991956e+18
	ExponentiatedGradient	0.06	0.15
	ICVAE	>1000	-2.5127543355692334e+18
Compas	BinaryMI	0.0	0.77
	ExponentiatedGradient	0.03	0.11
	ICVAE	0.02	0.21
Folktables AK	BinaryMI	>1000	-605054755547313.5
	ExponentiatedGradient	>1000	-6.13217413779456e+17
	ICVAE	>1000	-2.0761652335054696e+16
Folktables HI	BinaryMI	>1000	-469404739259224.8
	ExponentiatedGradient	>1000	-51026675754403.53
	ICVAE	>1000	-945121999846983.6
German	BinaryMI	0.02	0.66
	ExponentiatedGradient	0.12	-0.31
	ICVAE	0.01	0.53

Table 2. MSE and R^2 for all γ values when training ICVAE on German.

γ	MSE	R^2
0.0	0.0205	0.6545
0.1	0.0123	0.7028
0.2	0.0082	0.5988
0.3	0.0116	0.6656
0.4	0.0088	0.5068
0.5	0.0128	0.4392
0.6	0.0098	0.6163
0.7	0.0107	0.4084
0.8	0.0039	0.3870
0.9	0.0056	0.4941
1.0	0.0019	0.3447