FairSVM: A Mixed-Integer Programming Framework for Fairness-Constrained Support Vector Machines

Gabriele Marchesi
Department of Management, Information and Production Engineering
University of Bergamo
Viale G. Marconi 5, Dalmine 24044, Italy
g.marchesi15@studenti.unibg.it

Francesca Maggioni
Department of Management, Information and Production Engineering
University of Bergamo
Viale G. Marconi 5, Dalmine 24044, Italy
francesca.maggioni@unibg.it

Bismark Singh
School of Mathematical Sciences
University of Southampton
Southampton SO17 1BJ, UK
b.singh@southampton.ac.uk

Abstract

Machine learning classifiers are increasingly deployed in high-stakes domains such as credit scoring, hiring, and criminal justice, where concerns about algorithmic bias have become central. Existing approaches to algorithmic fairness often rely either on removing sensitive features ("fairness through unawareness") or on post-hoc corrections, which limit transparency and flexibility. We propose a mixed–integer programming framework, FAIRSVM, that embeds fairness constraints directly into the training of soft–margin Support Vector Machines. Our formulation expresses multiple group fairness notions — including statistical parity, predictive equality, equal opportunity, equalized odds, and conditional statistical parity — as linear constraints within the SVM model. Our approach enables explicit control over the trade-off between predictive accuracy and fairness through adjustable parameters. We evaluate FAIRSVM on three widely studied datasets (German Credit, Adult Income, COMPAS) and compare it against both optimization-based (FAIROCT) and model-agnostic (CR, EXPG, RTO) benchmarks. Results show that FAIRSVM substantially reduces group disparities with only limited loss in accuracy, while offering greater flexibility in navigating fairness–accuracy trade-offs. These findings highlight the potential of optimization-based formulations as a foundation for developing next-generation fairness-aware machine learning models.

1 Introduction

Machine learning (ML) algorithms are increasingly shaping consequential decisions in domains such as criminal justice, public policy, and healthcare. Risk prediction tools in [Angwin et al.,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML×OR Workshop: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2022] , recommendation and classification systems for homeless individuals [Azizi et al., 2018], and diagnostic support systems in medicine [Fatima and Pasha, 2017] provide applications where algorithmic outcomes directly affect the lives of individuals. Alongside their predictive power, these models often raise concerns about *algorithmic fairness*: the possibility that systematic biases in data or model design may lead to discriminatory treatment of individuals or groups defined by sensitive attributes such as age, gender, or race.

Currently, there exists a broad literature to define and enforce fairness in ML. One family of approaches relies on pre-processing, e.g., *fairness through unawareness*, which removes sensitive features prior to model training. While straightforward, such strategies are generally insufficient because proxies for sensitive attributes can remain embedded in other features. In contrast, in-processing methods incorporate fairness constraints directly into the learning algorithm, allowing practitioners to balance predictive accuracy with fairness criteria. Several group-level notions of fairness have been proposed, including statistical parity, predictive equality, equal opportunity, and equalized odds. More advanced metrics, such as conditional statistical parity, further account for relevant context (e.g., prior risk factors) in defining fairness across groups.

Among the classification methods, Support Vector Machines (SVMs) remain a widely used and well-understood technique, valued for their strong predictive performance and relatively transparent geometric interpretation [Vapnik, 2006, Cortes and Vapnik, 1995]. In this work, we adopt the soft-margin SVM formulation [Cortes and Vapnik, 1995, Blanco et al., 2020] as a basis for incorporating fairness constraints. Our approach introduces mixed-integer programming (MIP) formulations that directly embed different group fairness metrics into the SVM training problem. This is what we term a "FAIRSVM" which allows explicit control over the trade-off between accuracy and fairness through adjustable parameters.

The key aim of this work is to propose a general framework for integrating multiple fairness metrics into SVM classifiers, thereby contributing both methodological advances and empirical insights into the study of fairness in machine learning. To this end, we evaluate FAIRSVM on benchmark datasets with known disparities — German Credit, Adult Income, and COMPAS — comparing its performance against state-of-the-art fairness interventions, including the optimization-based FairOCT [Jo et al., 2023] which employs classification trees of depth $d = [1, 2, 3]$ and model-agnostic methods such as Correlation Remover (CR) [Weerts et al., 2023], Exponentiated Gradient (ExpG) [Agarwal et al., 2018] and Randomized Threshold Optimized (RTO) [Hardt et al., 2016]. Our results demonstrate that FAIRSVM substantially reduces group disparity (in some cases by up to 70%) with only a marginal reduction in accuracy, while providing greater flexibility in navigating fairness–accuracy trade-offs than the existing alternative approaches.

The rest of this work is organized as follows. Section 2 formalizes several group fairness metrics as optimization constraints within the SVM framework. Section 3.2 presents our experimental design and results. Section 4 concludes by discussing the implications for fairness-aware optimization and outlining directions for future research.

2 Fairness metrics and FAIRSVM formulation

We incorporate fairness into classification by embedding some of the group fairness notions mentioned in Section 1 as constraints within a SVM. Let the training set be $D = \{(x_i, y_i)\}_{i \in I}$, with $x_i \in \mathbb{R}^J$, $y_i \in \{-1, 1\}$, and decision boundary $w^\top x - \gamma = 0$. We denote by Q the sensitive attribute with levels $p \in P$, and — when using conditional metrics — by L a conditioning attribute with levels $l \in L$. We encode predictions by binary variables $u_i \in \{0, 1\}$, where $u_i = 1$ if instance i is classified positive. Finally, we specify a tolerance parameter $\delta \in [0, 1]$ to denote the maximum allowed disparity.

We reformulate five widely studied group fairness metrics:

(i) Statistical Parity (SP): equal positive prediction rates across groups [Dwork et al., 2012]; i.e. $P[\hat{Y} = 1|Q = p] \approx P[\hat{Y} = 1|Q = p']$.

(ii) Predictive Equality (PE): equal false positive rates [Chouldechova, 2017]; i.e. $P[\hat{Y} = 1|Q = p, Y = -1] \approx P[\hat{Y} = 1|Q = p', Y = -1]$.

(iii) Equal Opportunity (EOpp): equal true positive rates [Hardt et al., 2016], i.e. $P[\hat{Y} = 1|Q = p, Y = 1] \approx P[\hat{Y} = 1|Q = p', Y = 1]$.

(iv) Equalized Odds (EOdds): combines PE and EOpp, requiring parity of both false and true positive rates [Hardt et al., 2016].
(v) Conditional Statistical Parity (CSP): extends SP by conditioning on legitimate factors L [Corbett-Davies et al., 2017], i.e. $P[\hat{Y} = 1|Q = p, L = l] \approx P[\hat{Y} = 1|Q = p', L = l]$.

Next, we embed fairness constraints for each of the five metrics into a soft-margin SVM via a quadratic MIP by bounding the difference between group rates by δ.

$$\min_{w,\gamma,\xi,u} \quad \tfrac{1}{2}\|w\|_2^2 + \nu \sum_{i \in I} \xi_i \tag{1a}$$

$$\text{s.t.} \quad y_i(w^\top x_i - \gamma) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i \in I, \tag{1b}$$

$$\text{(fairness constraint: SP, PE, EOpp, EOdds, or CSP)}, \tag{1c}$$

$$-M(1 - u_i) \leq w^\top x_i - \gamma - \varepsilon, \quad w^\top x_i - \gamma \leq M u_i, \quad u_i \in \{0, 1\}. \tag{1d}$$

Model (1) defines the FAIRSVM and differs only in the fairness constraint selected. In the next section, we evaluate the proposed FAIRSVM against existing fairness-aware classification approaches on benchmark datasets.

3 Experiments

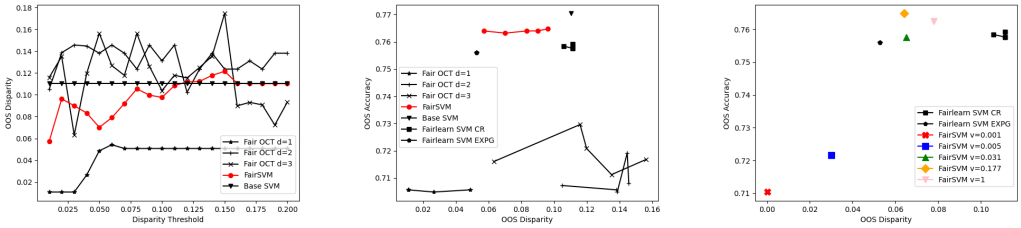
3.1 Data

In this section, we evaluate the proposed FAIRSVM on three real-world datasets: the German Credit, Adult Income, and COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) datasets from the University of California Irvine (UCI) Machine Learning repository [Dua and Graff, 2017] and ProPublica [Angwin et al., 2016]. These datasets differ in size (from 1000 to $6000+$ records), number of features ($6 - 20$) and types of sensitive attributes (age, gender, and race). We test the German and Adult with the SP, PE, EOpp, and EOdds metrics, while COMPAS is tested with CSP (using prior criminal record as conditioning variable). Further details on pre-processing, sample sizes, and feature definitions are reported in Appendix A.1. For each dataset, we split 75%/25% into training and test sets and repeat experiments across five random seeds. We select the regularization parameter ν from five logarithmically spaced values between 0.001 and 1 retaining the best out-of-sample accuracy, and we vary the disparity tolerance $\delta \in [0.01, 0.55]$ in increments of 0.01. All models are solved with a time limit of three hours on a 16-core, 32-thread AMD EPYC CPU machine.

3.2 Results

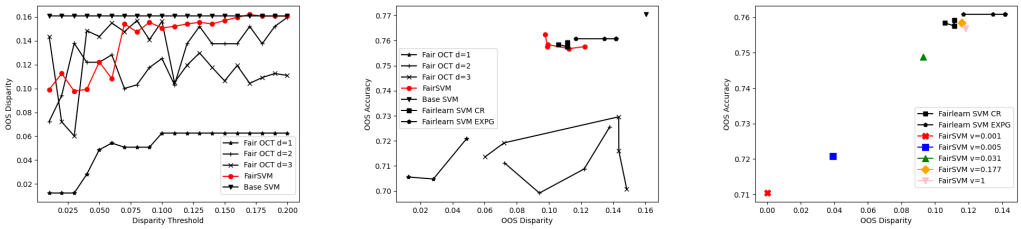
Figure 1 provides representative results for the German Credit dataset under the PE metric; complete results for all datasets and metrics appear in Appendix A.2. Specifically, Figure 1a shows how varying δ in FAIRSVM and FairOCT translates into out-of-sample disparity. FAIRSVM consistently reduces disparity by 30–70% relative to the baseline soft-margin SVM, while maintaining feasible computational times. Figure 1b highlights the classical trade-off between accuracy and fairness. Although the baseline SVM attains high accuracy at the cost of considerable disparity, FAIRSVM achieves reductions in disparity of up to 50% for small values of δ, with only minor losses in accuracy (typically below 5%). When compared with FairOCT, FAIRSVM attains substantially higher accuracy at comparable levels of fairness, particularly relative to FairOCT with $d = 3$. The RTO method is omitted from the figure due to its consistently high disparities, which range from 0.6 to 0.9. Finally, Figure 1c highlights the effect of the regularization parameter ν. Small ν emphasizes fairness, often yielding near-perfect parity at the cost of accuracy, whereas larger ν favors predictive performance while still reducing disparities compared to the unconstrained SVM. This adaptability is a distinctive feature of FAIRSVM, allowing practitioners to select solutions along the fairness–accuracy tradeoff frontier rather than being constrained to a single tradeoff point.

Figure 2 provides results for the German Credit dataset under the EOdds metric. Figure 2a shows that FAIRSVM reduces disparity by nearly 40% relative to the baseline soft-margin SVM, achieving similar performance to FairOCT with $d = 2$. Figure 2b demonstrates that FAIRSVM outperforms model-agnostic methods in terms of both accuracy and fairness. When compared with FairOCT



(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure 1: Results for the German Credit dataset under the *Predictive Equality* (PE) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.



(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure 2: Results for the German Credit dataset under the *Equalized Odds* (EOdds) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.

at different depths, FAIRSVM attains higher accuracy but also higher disparity. Finally, Figure 2c reveals patterns consistent with those observed under the PE metric: smaller values of ν enhance fairness at the expense of predictive accuracy, while larger values of ν prioritize accuracy while still yielding moderate reductions in disparity.

To summarize, our computational experiments demonstrate that FAIRSVM (a) effectively reduces group-level disparities across multiple fairness definitions, (b) incurs only limited accuracy losses, and (c) provides explicit control over the fairness–accuracy balance. These findings highlight the potential of *optimization-based* formulations to serve as a flexible foundation for fairness-aware ML.

4 Conclusions

We introduced FAIRSVM, a MIP framework that integrates multiple group fairness constraints directly into linear SVM training. By formulating *statistical parity*, *predictive equality*, *equal opportunity*, *equalized odds*, and *conditional statistical parity* as linear constraints, FAIRSVM allows a unified treatment of fairness notions within a single optimization model.

Our empirical computational evaluation on three widely used benchmarks (German Credit, Adult Income, COMPAS) shows that FAIRSVM can substantially reduce unfairness while maintaining competitive accuracy compared to existing methods. Promising future work directions include extending the approach to non-linear kernels and so-called "deep" representations, developing scalable solution techniques for larger datasets, and exploring the integration of individual fairness and causal fairness notions into the optimization framework.

References

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69. Proceedings of Machine Learning Research, 2018.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Compas analysis, 2016. URL https://github.com/propublica/compas-analysis.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. *Machine bias*, pages 254–264. Auerbach Publications, Mar. 2022. ISBN 9781003278290. doi: 10.1201/9781003278290-37.

M. J. Azizi, P. Vayanos, B. Wilder, E. Rice, and M. Tambe. Designing fair, efficient, and interpretable policies for prioritizing homeless youth for housing resources. In *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 35–51. Springer, Springer International Publishing, 2018. ISBN 9783319930312. doi: 10.1007/978-3-319-93031-2_3.

V. Blanco, J. Puerto, and A. M. Rodríguez-Chía. On lp-support vector machines and multidimensional kernels. *Journal of Machine Learning Research*, 21(14):1–29, 2020. URL https://dl.acm.org/doi/10.5555/3455716.3455730.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, jun 2017. doi: 10.1007/bf00994018.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sept. 1995. ISSN 1573-0565. doi: 10.1007/bf00994018.

D. Dua and C. Graff. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, 2017. University of California, Irvine, School of Information and Computer Sciences. Accessed: 2025-09-01.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

M. Fatima and M. Pasha. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 09(01):1–16, 2017. ISSN 2150-8410. doi: 10.4236/jilsa.2017.91001.

M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, pages 3315–3323. Curran Associates, Inc., 2016.

N. Jo, S. Aghaei, J. Benson, A. Gomez, and P. Vayanos. Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 181–192, 2023.

V. Vapnik. *Estimation of dependences based on empirical data*. Springer New York, 2006. ISBN 9780387342399. doi: 10.1007/0-387-34239-7.

H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio. Fairlearn: Assessing and improving fairness of AI systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

In this appendix, we provide details on our datasets (Appendix A.1) and additional computational results (Appendix A.2) complementing Section 3.1 and Section 3.2 of the main text, respectively.

A.1 Additional details on datasets

The first dataset is *German Credit*, which classifies individuals as having good or bad credit. The sensitive feature is *age*, converted from continuous to binary (≤ 25 v/s. > 25). The discriminated group is the younger one, which tends to receive worse credit scores. This dataset has 1,000 samples and is the smallest one in our analysis.

The second dataset is *COMPAS* (Correctional Offender Management Profiling for Alternative Sanctions), widely used for predicting risk of recidivism within two years, with known bias against non-White defendants. The sensitive feature *race* was reduced from six classes to four by aggregating the smallest groups. This dataset contains 6,172 data points.

The third dataset is *Adult Income*, where the task is to classify whether income exceeds \$50,000. The sensitive feature is *sex*, with women as the disadvantaged group. Originally containing about 30,000 samples, it was sub-sampled to 2,700 for computational ease.

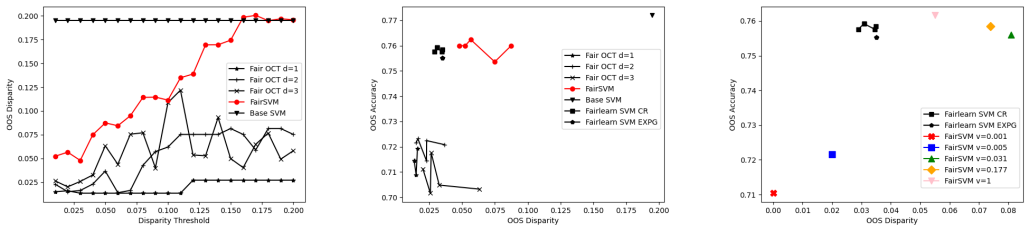
All datasets were pre-processed to remove missing values and to encode categorical variables numerically. Table A.1 provides a summary.

Table A.1: Dataset information

Dataset	Samples	Features	Sensitive feature	Discriminated class
German Credit	1,000	20	Age	≤ 25 years
Adult Income	2,700	14	Sex	Women
COMPAS	6,172	6	Race	Non-white criminals

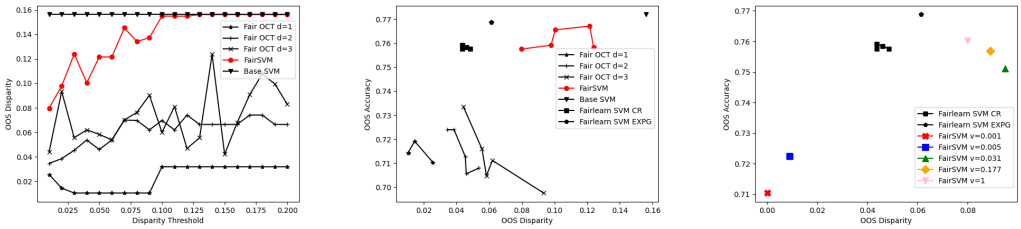
A.2 Additional computational results

The following figures are analogous to Figure 1 and Figure 2 of the main text.



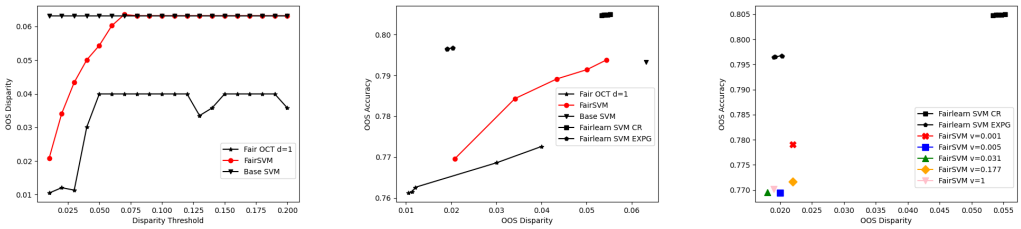
(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure A.3: Results for the German Credit dataset under the *Statistical Parity* (SP) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.



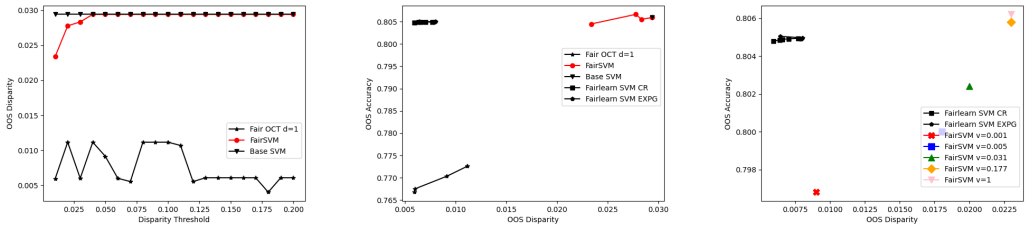
(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure A.4: Results for the German Credit dataset under the *Equal Opportunity* (EOpp) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.



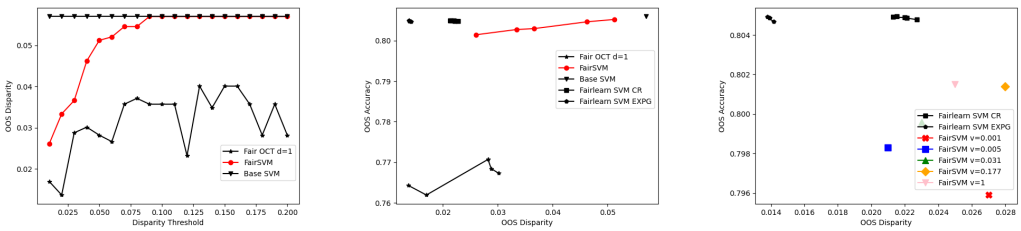
(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure A.5: Results for the Adult Income dataset under the *Statistical Parity* (SP) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.



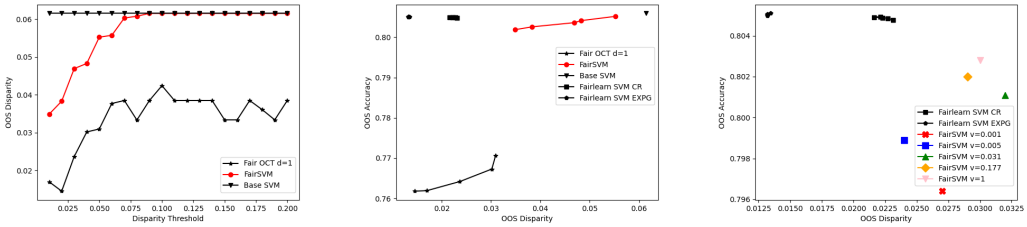
(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure A.6: Results for the Adult Income dataset under the *Predictive Equality* (PE) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.



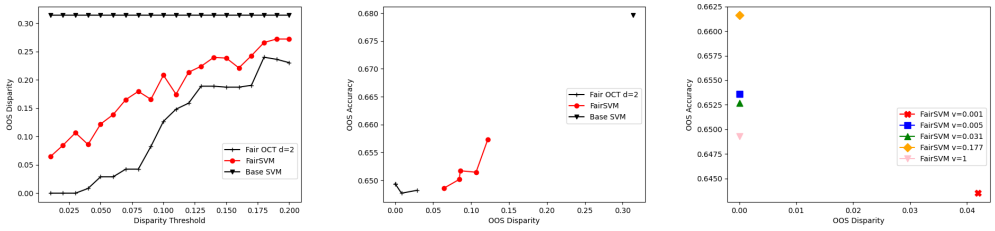
(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure A.7: Results for the Adult Income dataset under the *Equal Opportunity* (EOpp) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.



(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure A.8: Results for the Adult Income dataset under the *Equalized Odds* (EOdds) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.



(a) Disparity v/s δ (b) Accuracy–disparity frontier (c) Effect of ν at low δ

Figure A.9: Results for the COMPAS dataset under the *Conditional Statistical Parity* (CSP) metric. (a) Out-of-sample disparity v/s. δ; (b) accuracy–disparity trade-offs across methods; (c) sensitivity to ν.