FRAPPÉ: A Group Fairness Framework for Post-Processing Everything

Alexandru Ţifrea[*1] **Preethi Lahoti**[2] **Ben Packer**[2] **Yoni Halpern**[2] **Ahmad Beirami**[2] **Flavien Prost**[2]

Abstract

Despite achieving promising fairness-error trade-offs, in-processing mitigation techniques for group fairness cannot be employed in numerous practical applications with limited computation resources or no access to the training pipeline of the prediction model. In these situations, post-processing is a viable alternative. However, current methods are tailored to specific problem settings and fairness definitions and hence, are not as broadly applicable as in-processing. In this work, we propose a framework that turns any regularized in-processing method into a post-processing approach. This procedure prescribes a way to obtain post-processing techniques for a much broader range of problem settings than the prior post-processing literature. We show theoretically and through extensive experiments that our framework preserves the good fairness-error trade-offs achieved with in-processing and can improve over the effectiveness of prior post-processing methods. Finally, we demonstrate several advantages of a modular mitigation strategy that disentangles the training of the prediction model from the fairness mitigation, including better performance on tasks with partial group labels.[1]

1. Introduction

As machine learning (ML) algorithms are deployed in applications with a profound social impact, it becomes crucial that the biases they exhibit (Bickel et al., 1975; Dastin, 2018; Mehrabi et al., 2021) are properly mitigated. Of particular importance is being equitable with respect to the different subgroups in the data (i.e. *group fairness*) where groups are determined by sensitive demographic attributes such as race, sex, age, etc (Barocas et al., 2023). To operationalize fairness, one can choose between the several dozen alternative definitions of fairness (Narayanan, 2018; Hort et al., 2023) capturing different notions of equity.

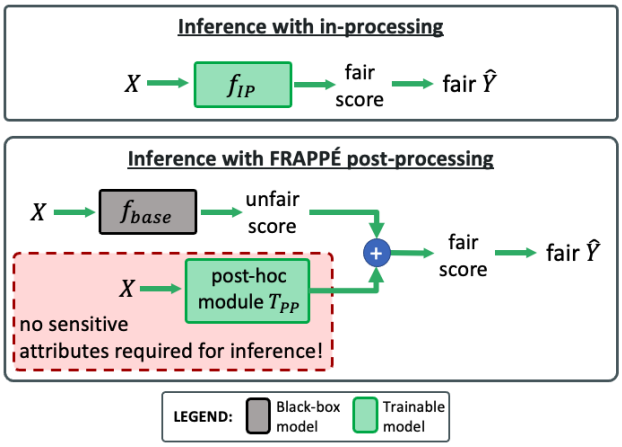


Figure 1: **Inference with FRAPPÉ and in-processing.** FRAPPÉ methods add the output of post-hoc module T_{PP} to the unfair scores output by pre-trained model f_{base}. Unlike prior post-processing methods, FRAPPÉ does not require sensitive attributes for inference. While in-processing trains the entire prediction model f_{IP} to induce fairness, FRAPPÉ only trains the post-hoc module. Note that, for classification, thresholding the predicted scores yields outputs $\hat{Y}$, while for regression $\hat{Y}$ coincides with the score.

One of the most studied mitigation paradigms for group fairness is in-processing (Hort et al., 2023), which changes the training procedure, for instance, by adding a fairness regularizer or constraint to the training loss. In addition to their good performance, in-processing methods are appealing thanks to their broad applicability: to induce a new notion of fairness, one simply needs to quantify fairness violations and use that as a regularizer or constraint to train a new prediction model.

In practice, however, retraining the prediction model to induce fairness is often infeasible. For example, complex prediction models are challenging to retrain when only limited computational resources are available (Cruz & Hardt, 2023). To make matters worse, there is often no access to the param-

[*]Work done during an internship at Google. [1]Department of Computer Science, ETH Zurich [2]Google DeepMind. Correspondence to: Alexandru Ţifrea <alexandru.tifrea@inf.ethz.ch>, Flavien Prost <fprost@google.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]Code is available at https://github.com/google-research/google-research/tree/master/postproc_fairness.

eters of the prediction model, which can only be queried to produce outputs (e.g. logits) for the provided inputs. For instance, when using one of the increasingly popular AutoML platforms (He et al., 2021), one often has little to no control over the training objective, making it impossible to induce the desired fairness notion via in-processing. Additionally, in-processing might not be effective in a multi-component system, as prior research on compositional fairness (Dwork & Ilvento, 2018; Atwood et al., 2023) shows that debiasing each component individually might not lead to a fair outcome.

In these situations, post-processing techniques offer the most compelling way to ensure fair predictions via a post-hoc module that transforms the outputs of a pre-trained base model (Figure 1). However, current post-processing techniques are not nearly as broadly applicable as in-processing. The recent survey of Hort et al. (2023) found that over 200 methods (out of 341 surveyed approaches) used in-processing, covering a broad class of fairness definitions and problem settings. In contrast, the survey identifies only 56 post-processing methods which are tailored to specific problem settings (e.g. problems with binary sensitive attributes (Pleiss et al., 2017; Kim et al., 2020)) and specific fairness definitions (e.g. Hardt et al. (2016) focuses on equal opportunity/odds; Xian et al. (2023) considers statistical parity). The recent method of Alghamdi et al. (2022) significantly extends the applicability of post-processing but is still confined to problems with discrete sensitive attributes and notions of fairness based on a conditional mean score.

Furthermore, existing post-processing methods require that sensitive attributes are known at inference time, despite it being often untenable in practice (Veale & Binns, 2017). This complex landscape leads to the question:

Can we design a post-processing module to induce group fairness which satisfies the following desiderata?

D1. *Works for any pre-trained models that output scores (e.g. logits, continuous labels).*

D2. *Can trade off fairness and prediction error effectively for any quantifiable notion of fairness.*

D3. *Does not require sensitive attributes at inference time.*

To answer this question, in this work we propose a *Fairness Framework for Post-Processing Everything* (**FRAPPÉ**) that turns any regularized in-processing method into a post-processing approach. As highlighted in Figure 2, the resulting method is *modular*, namely it decouples training the (unfair) base model from learning the post-hoc module $T_{\text{PP}}(X)$. Importantly, FRAPPÉ methods are designed to allow training the post-hoc module using *any* arbitrary fairness regularizer. Finally, the post-hoc module of FRAPPÉ methods models a function of covariates X, thus not requiring explicit knowledge of the sensitive attributes at inference time, as shown in Figure 1.

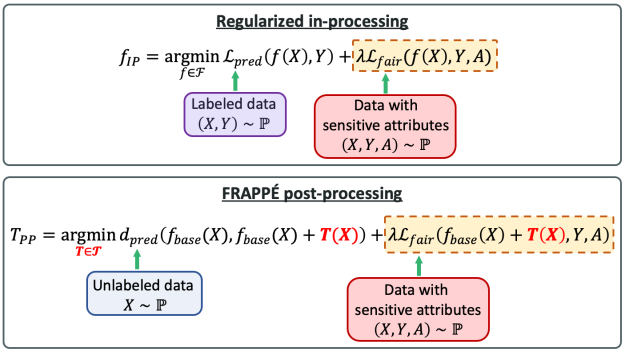


Figure 2: **FRAPPÉ and in-processing training objectives.** Unlike existing post-processing techniques, FRAPPÉ methods can be trained with *any* in-processing fairness regularizer $\mathcal{L}_{fair}$ (orange box). In contrast to in-processing, FRAPPÉ only trains the post-hoc module $T_{\text{PP}}(X)$ instead of the entire prediction model f. Loss terms are computed on data that is labeled, unlabeled or annotated with sensitive attributes, as indicated. d_{pred} measures the difference between the outputs of the base and the fair models (see Section 3).

Our contributions are as follows:

1. Motivated by our theoretical result that establishes a connection between the in- and post-processing training objectives, we propose a novel framework to train a modular post-processing method using an in-processing fairness regularizer (Section 3). The procedure is designed to solve the limitations of prior in- and post-processing approaches captured by desiderata D1 – D3.
2. We complement the theoretical findings with extensive experiments (Section 5) which show that FRAPPÉ methods do not degrade the fairness-error trade-off of their in-processing counterparts for several in-processing regularizers targeting different fairness definitions (MinDiff (Prost et al., 2019) for equal opportunity, Cho et al. (2020) for statistical parity, Mary et al. (2019) for equal odds) and various commonly used datasets (Adult, COMPAS, HSLS, ENEM, Communities & Crime).
3. We demonstrate empirically the advantages of our framework's modular design. Unlike prior post-processing techniques, FRAPPÉ can induce *any* notion of group fairness. Moreover, even when prior approaches are applicable, our experiments reveal that FRAPPÉ methods are on par or better compared to competitive post-processing techniques such as Alghamdi et al. (2022). Finally, we provide evidence that modular FRAPPÉ methods can perform significantly better than their in-processing counterparts on data with partial group labels, when sensitive annotations are scarce (Section 5.3).

2. Problem setting

We consider prediction tasks where the goal is to predict labels $y \in \mathcal{Y}$ (discrete in classification, or continuous for

regression) from covariates $\boldsymbol{x} \in \mathcal{X}$ with low prediction error. A simple learning algorithm, Empirical Risk Minimization or ERM (Vapnik, 1991), that minimizes the prediction loss on an i.i.d. dataset is known to achieve great average-case error. However, this simple strategy does not guarantee good fairness (Menon & Williamson, 2018; Chen et al., 2018; Zhao & Gordon, 2019; Sagawa et al., 2020; Bardenhagen et al., 2021; Sanyal et al., 2022). In applications where fairness is important, it is necessary to adjust the learning algorithm to promote greater equity.

Group fairness. A common fairness consideration is the model impact on sensitive groups. In the framework of group fairness, there exists a sensitive attribute A (discrete or continuous) with respect to which we expect an algorithm to be equitable. Different flavors of group fairness are captured formally by different definitions, e.g. statistical parity (SP) (Calders et al., 2009; Dwork et al., 2012), equal opportunity (EqOpp), equalized odds (EqOdds) (Hardt et al., 2016). We refer to Barocas et al. (2023) and the respective prior works for details on these fairness definitions. Since fairness and predictive performance are often at odds (Menon & Williamson, 2018; Chen et al., 2018; Zhao & Gordon, 2019), the literature focuses on achieving a good trade-off. Two remarkably effective paradigms at reaching a good fairness-error trade-off in practice are in-processing and post-processing (Caton & Haas, 2023; Hort et al., 2023).[2]

In-processing for group fairness. Methods in this category seek to optimize a prediction loss (e.g. cross-entropy, mean squared error) while at the same time encouraging the prediction model to be fair. Regularized in-processing methods (Beutel et al., 2019; Prost et al., 2019; Mary et al., 2019; Cho et al., 2020; Lowy et al., 2022) consider an optimization objective with a fairness violation penalty added to the prediction loss as a regularization term (Figure 2).

Post-processing for group fairness. To induce fairness, post-processing techniques (Hardt et al., 2016; Kamiran et al., 2018; Nandy et al., 2022; Alghamdi et al., 2022; Cruz & Hardt, 2023) adjust the scores output by a pre-trained prediction model f_{base}. The current literature on post-processing methods for classification or regression focuses exclusively on *group-dependent* transformations $f_{\text{fair}}(\boldsymbol{x}) = T_A(f_{\text{base}}(\boldsymbol{x}))$, where f_{base} and f_{fair} are the pre-trained and the fair models, respectively. The *post-hoc transformation* (or *post-hoc module*) T_A is selected based on the value of the sensitive attribute A, from a set containing one learned transformation for each possible value of A.

[2]There also exist pre-processing approaches (Zemel et al., 2013; Madras et al., 2018; Lahoti et al., 2019) that try to debias the data distribution. However, their performance in practice is usually significantly worse compared to in- and post-processing methods (Zehlike et al., 2021; Caton & Haas, 2023; Hort et al., 2023).

3. Proposed framework

In this section we introduce the FRAPPÉ framework that transforms a regularized in-processing method for group fairness into a post-processing one.[3] Unlike prior post-processing approaches, instead of a group-dependent transformation that depends explicitly on the sensitive attribute, FRAPPÉ methods employ an additive term that is a function of all covariates $\boldsymbol{x}$ (the choice of the additive transformation is explained in Section 3.2):

$$f_{\text{fair}}(\boldsymbol{x}) = f_{\text{base}}(\boldsymbol{x}) + T_{\text{PP}}(\boldsymbol{x}). \tag{1}$$

In what follows, we argue that FRAPPÉ methods are specifically designed to overcome the shortcomings of prior post-processing approaches captured in desiderata D1 – D3 (and discussed in more detail in Section 6), while also enjoying the advantages of a modular design (e.g. reduced computation time, no need to access training pipeline and data).

3.1. Theoretical motivation: An equivalence between in- and post-processing for GLMs

We begin by motivating the proposed method by establishing a connection between a regularized in-processing objective and a bi-level optimization problem akin to post-processing. To illustrate this intuition, we consider predictors that are generalized linear models (GLM) (Nelder & Wedderburn, 1972) and take the form $f_\theta(\boldsymbol{x}) = \psi(\boldsymbol{\theta}^\top \boldsymbol{x})$ for parameters $\boldsymbol{\theta} \in \mathbb{R}^D$ and a link function $\psi : \mathbb{R} \to \mathbb{R}$ (e.g. identity or sigmoid, for linear or logistic regression, respectively). Given datasets $\mathcal{D}_{\text{pred}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and $\mathcal{D}_{\text{sensitive}} = \{(\boldsymbol{x}_j, y_j, a_j)\}_{i=1}^m$ drawn i.i.d. from the same distribution[4], we analyze generic regularized optimization problems of the form

$$\begin{aligned}\text{OPT}_{\text{IP}}(\boldsymbol{\theta}; \lambda) = \frac{1}{n} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{pred}}} \mathcal{L}_{pred}(\boldsymbol{x}, y; \boldsymbol{\theta}) \\ + \lambda \mathcal{L}_{fair}(\boldsymbol{\theta}; \mathcal{D}_{\text{sensitive}}),\end{aligned} \tag{2}$$

where $\mathcal{L}_{pred}$ is the prediction loss and $\mathcal{L}_{fair}$ is an arbitrary regularizer capturing a fairness violation penalty. We consider loss functions that can be written as

$$\mathcal{L}_{pred}(\boldsymbol{x}, y; \boldsymbol{\theta}) = F(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top G(\boldsymbol{x}, y), \tag{3}$$

with $F : \mathbb{R}^D \to \mathbb{R}$ strictly convex. The function G is a sufficient statistic with respect to $\boldsymbol{\theta}$ and F can be viewed as a partition function (Wainwright & Jordan, 2008). Standard

[3]While we focus on *regularized* in-processing objectives, our framework can easily be extended to constrained methods as well (Cotter et al., 2019; Chierichetti et al., 2019; Celis et al., 2019), as exemplified in Figure 3c for the method of Agarwal et al. (2018).

[4]Often, in practice, $\mathcal{D}_{\text{pred}}$ and $\mathcal{D}_{\text{sensitive}}$ may even coincide. We leave a discussion of the implications of a potential distribution shift between $\mathcal{D}_{\text{pred}}$ and $\mathcal{D}_{\text{sensitive}}$ as future work.

loss functions for linear regression (e.g. mean squared error) or classification (e.g. logistic loss) follow this pattern.

For every optimization problem that takes the form in Equation (2), consider the following corresponding bi-level optimization problem:

$$\text{OPT}_{\text{PP}}(\boldsymbol{\theta}; \lambda) = D_F(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{base}}) + \lambda \mathcal{L}_{fair}(\boldsymbol{\theta}; \mathcal{D}_{\text{sensitive}}),$$
$$\text{with } \boldsymbol{\theta}_{\text{base}} := \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_{\text{pred}}} \mathcal{L}_{pred}(\boldsymbol{x}, y; \boldsymbol{\theta}), \quad (4)$$

where we denote by $D_F(\boldsymbol{\theta}, \boldsymbol{\phi}) := F(\boldsymbol{\theta}) - F(\boldsymbol{\phi}) - \nabla F(\boldsymbol{\phi})^\top(\boldsymbol{\theta} - \boldsymbol{\phi})$ the Bregman divergence (Bregman, 1967) of the partition function $F(\boldsymbol{\theta})$. Intuitively, the first term encourages that the outputs produced by the GLMs determined by $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{\text{base}}$ are similar.

Example for linear regression. For instance, for linear regression and the mean squared error (MSE), the link function is the identity $\psi(z) = z$, and the loss function can be written as $\ell_{\text{MSE}}(\boldsymbol{x}, y; \boldsymbol{\theta}) = \|\boldsymbol{\theta}^\top \boldsymbol{x} - y\|^2 = \boldsymbol{\theta}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{\theta} - 2y\boldsymbol{\theta}^\top \boldsymbol{x} + c$, for a constant $c \geq 0$. It follows that the Bregman divergence of $F(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{\theta}$ is given by $D_F(\boldsymbol{\theta}, \boldsymbol{\phi}) = (\boldsymbol{\theta} - \boldsymbol{\phi})^\top \boldsymbol{x}\boldsymbol{x}^\top (\boldsymbol{\theta} - \boldsymbol{\phi}) = \|\boldsymbol{\theta}^\top \boldsymbol{x} - \boldsymbol{\phi}^\top \boldsymbol{x}\|^2$, namely the MSE between the outputs of the models parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, for arbitrary $\boldsymbol{\theta}, \boldsymbol{\phi} \in \mathbb{R}^D$.

Equivalence between OPT_{IP} and OPT_{PP}. The following result establishes a connection between the optimization objectives OPT_{IP} (Equation (2)) and OPT_{PP} (Equation (4)) introduced above (the proof is provided in Appendix A).

Proposition 3.1. *Consider the optimization objectives introduced in Equations (2) and (4). There exists a constant $C \in \mathbb{R}$ such that for any $\boldsymbol{\theta} \in \mathbb{R}^D$ and $\lambda \geq 0$ we have*

$$OPT_{PP}(\boldsymbol{\theta}; \lambda) = OPT_{IP}(\boldsymbol{\theta}; \lambda) + C. \quad (5)$$

It follows from Proposition 3.1 that minimizing OPT_{IP} and OPT_{PP} leads to the same solution, for any choice of the regularizer $\mathcal{L}_{fair}$ and the regularization strength λ. In the context of fairness, this result implies that sweeping over the hyperparameter λ gives rise to identical fairness-error Pareto frontiers between any regularized in-processing and the corresponding post-processing method trained with OPT_{PP}. Moreover, since the two optimization problems are identical up to a universal constant, properties established for an in-processing objective (e.g. smoothness (Cho et al., 2020), convergence rate of gradient descent (Lowy et al., 2022), etc) carry over intrinsically to the OPT_{PP} counterpart.

3.2. Proposed post-processing framework

We now describe how to turn the insights revealed by Proposition 3.1 into a practical framework for training a post-processing method for group fairness with an in-processing objective. Moreover, we extend the intuition developed for GLMs to more generic function classes.

First, as mentioned before, numerous in-processing fairness mitigations (Prost et al., 2019; Mary et al., 2019; Cho et al., 2020; Lowy et al., 2022) consider optimization objectives that can be written like OPT_{IP}, where the regularizer $\mathcal{L}_{fair}$ captures a fairness violation penalty. On the other hand, the bi-level problem OPT_{PP} can be used to train a post-processing method, where the inner optimization corresponds to obtaining the pre-trained model parameters $\boldsymbol{\theta}_{\text{base}}$. While Proposition 3.1 only holds for GLMs, both the OPT_{IP} objective and OPT_{PP} can be considered in the context of training more generic model classes (e.g. neural networks).

Furthermore, recall that post-processing methods only modify the outputs of a pre-trained model, instead of training a model from scratch. The GLM scenario introduced above suggests to choose this post-hoc transformation to have the following additive form: $f_{\text{fair}}(\boldsymbol{x}) = \psi((\boldsymbol{\theta}_{\text{base}} + \boldsymbol{\theta}_{\text{T}})^\top \boldsymbol{x})$. In particular, depending on the link function ψ, the post-hoc transformation can be additive in output space (e.g. for linear regression) or logit space (e.g. for logistic regression). More generally, we can use the intuition developed for GLMs to propose the following post-processing transformation of a pre-trained model for regression and classification:

$$f_{\text{fair}}(\boldsymbol{x}) = f_{\text{base}}(\boldsymbol{x}) + T_{\text{PP}}(\boldsymbol{x}), \quad (6)$$

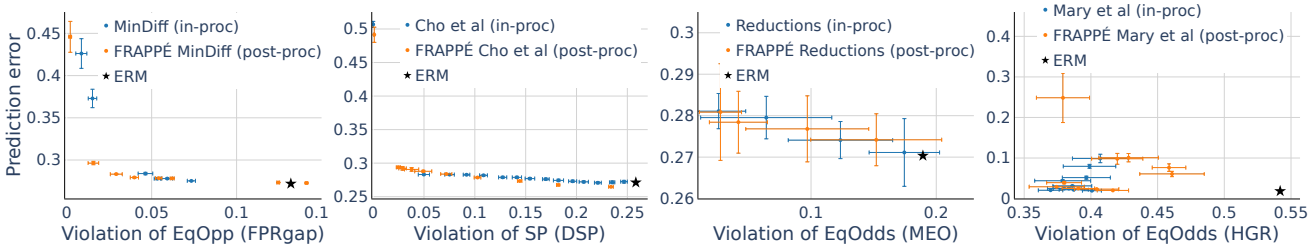
where, for classification, $f_{\text{base}}(\boldsymbol{x})$ and $f_{\text{fair}}(\boldsymbol{x})$ produce vectors of unnormalized logits. Importantly, FRAPPÉ methods only train the fairness correction $T_{\text{PP}}(\boldsymbol{x})$ which is often significantly less complex than the prediction model $f_{\text{base}}(\boldsymbol{x})$, thus decreasing training time compared to in-processing.

In conclusion, for an arbitrary in-processing method that solves a regularized objective like Equation (2), our framework constructs the following optimization problem:

$$T_{\text{PP}} := \arg\min_{T} \frac{1}{|\mathcal{D}_{\text{pp}}|} \sum_{\boldsymbol{x}\in\mathcal{D}_{\text{pp}}} d_{\text{pred}}((f_{\text{base}} + T)(\boldsymbol{x}); f_{\text{base}}(\boldsymbol{x})) + \lambda \mathcal{L}_{fair}(f_{\text{base}} + T; \mathcal{D}_{\text{sensitive}}), \quad (7)$$

where the two terms are computed on datasets $\mathcal{D}_{\text{sensitive}} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^m$ and $\mathcal{D}_{\text{pp}} = \{\boldsymbol{x}_i\}_{i=1}^u$ drawn i.i.d. from the same distribution. Here $\mathcal{L}_{fair}$ is the in-processing fairness regularizer, and d_{pred} is the Bregman divergence in OPT_{PP} or some other notion of discrepancy between the outputs of the f_{fair} and f_{base} models (e.g. KL divergence for classification).

Unlike in-processing, FRAPPÉ is modular and can find a different fairness-error trade-off or mitigate fairness with respect to a different definition (e.g. SP, EqOdds, EqOpp) by just retraining the simple additive term T_{PP}, instead of always retraining a new model f_{fair} from scratch (i.e. only modules in the green boxes in Figure 1 need to be retrained).



(a) $\text{OPT}_{\text{IP}} =$ *MinDiff*. Dataset: HSLS.
(b) $\text{OPT}_{\text{IP}} =$ Cho et al. (2020). Dataset: HSLS.
(c) $\text{OPT}_{\text{IP}} =$ *Reductions*. Dataset: HSLS.
(d) $\text{OPT}_{\text{IP}} =$ Mary et al. (2019). Dataset: Communities & Crime.

Figure 3: Inducing three different definition of fairness (EqOpp, SP, and EqOdds) using in-processing methods and their FRAPPÉ post-processing variant leads to similar Pareto frontiers. Thanks to their modular design, FRAPPÉ methods only need to retrain the post-hoc transformation $T_{\text{PP}}(x)$, instead of the entire prediction model. Appendix D.1 shows similar results on the Adult, COMPAS and ENEM datasets. Notably, FRAPPÉ Mary et al. (2019) is the first post-processing method that can operate on data with continuous sensitive attributes, such as Communities & Crime.

Finally, minimizing the d_{pred} term does not require labeled training data, which makes this procedure suitable when either the labels Y or the sensitive attributes A are difficult to collect (Awasthi et al., 2021; Prost et al., 2021). This observation is particularly important for mitigating the limitations of in-processing on data with partial group labels, when $|\mathcal{D}_{\text{sensitive}}|$ is small, as discussed in Section 5.3.

3.3. Connection to related prior works

The objective in Equation (7) can be seen as a two-step boosting algorithm (Schapire, 1990), where the second step corrects the unfairness of the model f_{base} obtained after the first step, similar to Liu et al. (2021); Bardenhagen et al. (2021). This formulation of post-processing is also related to post-hoc methods for uncertainty calibration (Pleiss et al., 2017; Kumar et al., 2019), as well as techniques for *model reprogramming*, such as Zhang et al. (2023) (see Appendix D.3 for more details). The additive post-hoc transformation that we employ has been considered in the past by works that focus exclusively on disparate performance and use a group-dependent *logit adjustment* to improve worst-group error (Khan et al., 2018; Cao et al., 2019; Menon et al., 2021a). Furthermore, the idea of reusing the in-processing optimization objective to train a post-processing method is reminiscent of recent works on group distributionally robust optimization (Sagawa et al., 2020), which show that last layer retraining is equivalent to training the entire neural network (Menon et al., 2021b; Shi et al., 2023; LaBonte et al., 2023). Finally, similar to our approach, unlabeled data has also been used to improve the trade-off between standard and robust error in the context of adversarial robustness (Carmon et al., 2019; Raghunathan et al., 2020).

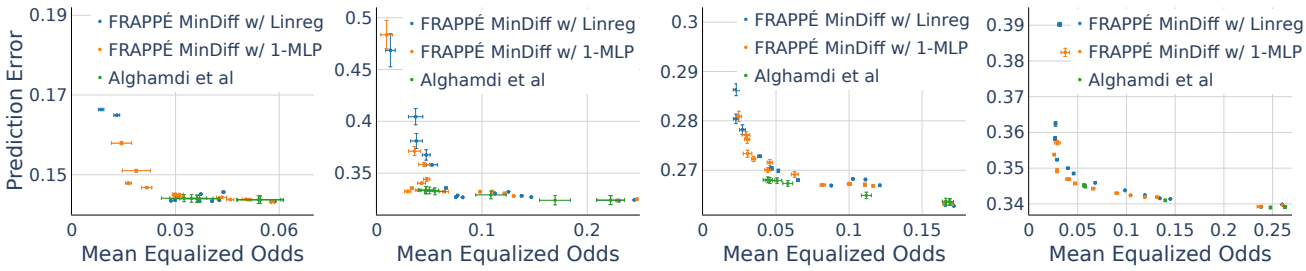
4. Experimental setup

We compare different fairness mitigation techniques by inspecting their fairness-error Pareto frontiers, obtained after varying the λ hyperparameter in Equations (2) and (7). We quantify fairness violations with metrics tailored to each specific fairness definition: the FPR gap for EqOpp, the difference in SP for SP, and the mean equalized odds violation for EqOdds. For EqOdds, we also employ the same evaluation metric as Mary et al. (2019), namely $\text{HGR}_\infty(f(X), A|Y) \in [0, 1]$, which takes small values when predictions $f(X)$ and sensitive attributes A are conditionally independent given true labels Y (see Appendix C.1 for details). For all metrics we report the mean and standard error computed over 10 runs with different random seeds.

In-processing baselines. To induce these notions of fairness, we use the FRAPPÉ framework to obtain a post-processing counterpart for several in-processing methods. We aim to cover a diverse set of regularized in-processing methods that have been demonstrated to perform well in recent thorough experimental studies (Cho et al., 2020; Jung et al., 2022; Lowy et al., 2022; Alghamdi et al., 2022). We identify the following as particularly competitive in-processing methods: i) for EqOpp, we consider MinDiff (Beutel et al., 2019; Prost et al., 2019); ii) for SP, the method of Cho et al. (2020); and iii) for EqOdds, *Reductions* (Agarwal et al., 2018) and the method of Mary et al. (2019). We consider the KL divergence and the MSE as d_{pred} in Equation (7), for classification and regression, respectively.

Datasets. We conduct experiments on standard datasets for assessing fairness mitigation techniques, namely Adult (Becker & Kohavi, 1996) and COMPAS (Angwin et al., 2016), as well as two recently proposed datasets: the high-school longitudinal study (HSLS) dataset (Jeong et al., 2022), and ENEM (Alghamdi et al., 2022). We also evaluate FRAPPÉ on data with continuous sensitive attributes (i.e. the Communities & Crime dataset (Redmond, 2009)), a setting where prior post-processing works cannot be applied. Appendix C.2 provides more details about the datasets.

Prediction models. We consider a broad set of model classes, varying from multi-layer perceptrons (MLPs) or



(a) Adult (b) COMPAS (c) HSLS (d) ENEM

Figure 4: Comparison between FRAPPÉ MinDiff for EqOdds and the best-performing post-processing method (Alghamdi et al., 2022), for random forest pre-trained models. See Figure 10 for a comparison with more post-processing baselines. While in-processing MinDiff cannot be used with non-gradient based models, FRAPPÉ MinDiff performs on-par or better than competitive post-processing approaches such as Alghamdi et al. (2022), even when the post-hoc transformation is as simple as linear regression or a 1-MLP.

gradient-boosted machines (GBMs), to non-gradient based models such as random forests (RFs). Both the in-processing and the pre-trained model are selected from these model classes. For FRAPPÉ, the post-hoc module $T_{\text{PP}}(X)$ is a much less complex model, e.g. linear regression.

Finally, the inherent fairness-error trade-off can pose serious challenges for hyperparameter tuning (Cruz et al., 2021). We adopt the standard practice in the literature, and select essential hyperparameters such as the learning rate so as to minimize prediction error on a holdout validation set, for all the baselines in our experiments. We defer further experimental details to Appendix C.4.

Inducing fairness with partial group labels. Often, in practice, it is challenging to collect data with sensitive attributes, e.g. users of an online service may not be willing to disclose their gender, ethnicity etc (Hashimoto et al., 2018; Coston et al., 2019; Lahoti et al., 2020; Liu et al., 2021; Bardenhagen et al., 2021; Awasthi et al., 2021; Prost et al., 2021). Therefore, the size of $\mathcal{D}_{\text{sensitive}}$ used to train fairness mitigations is significantly reduced, while $\mathcal{D}_{\text{pred}}$ may still be fairly large. In Section 5.3 we present experiments on data with partial group labels, in which $\mathcal{D}_{\text{pred}}$ and $\mathcal{D}_{\text{pp}}$ from Equations (2) and (7) consist of all the available training data, while $\mathcal{D}_{\text{sensitive}}$ contains only a fraction of this data, annotated with sensitive attributes.

5. Experimental results

In this section, we show empirically that FRAPPÉ methods satisfy the desiderata from Section 1. More specifically, we show in extensive experiments that FRAPPÉ methods preserve the competitive fairness-error trade-offs achieved with in-processing techniques, for various notions of fairness (**D2**), while enjoying the advantages of a post-processing method and being entirely agnostic to the prediction model class (**D1**). Moreover, FRAPPÉ methods perform on par or better than existing post-processing approaches, without requiring that sensitive attributes be known at inference time (**D3**). Finally, the FRAPPÉ framework helps to make post-processing fairness mitigations more broadly applicable, providing, for instance, the first post-processing method for data with *continuous* sensitive attributes.

5.1. Can FRAPPÉ perform as well as in-processing?

The data processing inequality (Cover & Thomas, 1991) suggests that it may be challenging for post-processing approaches to match the performance of in-processing methods. In this section, we demonstrate experimentally that FRAPPÉ methods preserve the good fairness-error trade-offs achieved by their in-processing counterparts, for several different notions of fairness and in-processing techniques.

As suggested by the intuition in Section 3.1, Figure 3 confirms that for several fairness definitions it is indeed possible to match the Pareto frontiers of in-processing methods using a modular FRAPPÉ variant. We observe the same equivalence on all datasets we considered (Appendix D.1). While the theoretical result assumes the same function class for the pre-trained model f_{base} and the post-hoc module T_{PP}, these experiments suggest that, in practice, the complexity of T_{PP} (i.e. linear model) can be significantly smaller than f_{base} (i.e. 3-layer MLP). Importantly, the FRAPPÉ variant of Mary et al. (2019) constitutes the first post-processing approach that can be utilized when the sensitive attributes are continuous. We note that the large error bars in Figure 3d are due to the challenges of optimizing the loss of Mary et al. (2019), discussed at length in Lowy et al. (2022).

Computation cost. Since FRAPPÉ methods only train the post-hoc module instead of the entire prediction model, they require only a fraction of the computational resources necessary for in-processing methods. Indeed, it takes 85.4 and 113.4 minutes to obtain the Pareto frontiers of MinDiff and Cho et al. (2020), respectively, on the Adult dataset.

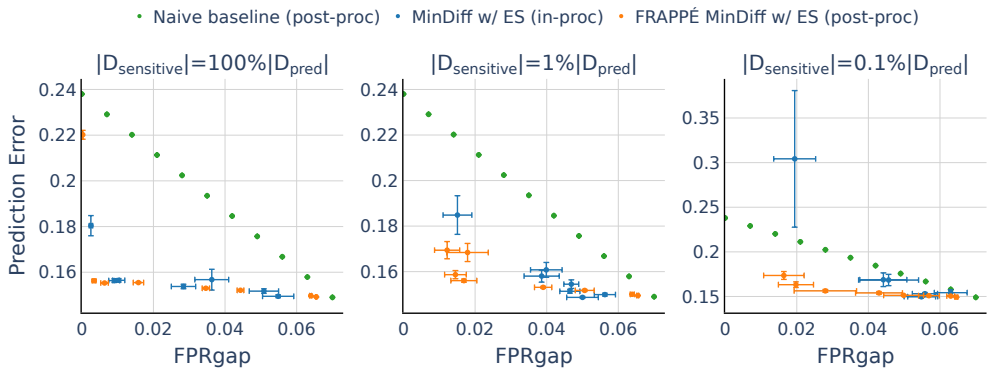


Figure 5: In-processing MinDiff and FRAPPÉ MinDiff with partial group labels on the Adult dataset with optimal early-stopping (ES) regularization. Our post-processing algorithm continues to perform well even in the extreme case where in-processing cannot outperform the trivial baseline described in Section 5.3.

In contrast, their FRAPPÉ variants only require 11.7 and 14.8 minutes, respectively (including the cost of training the base model), which is in line with the computation times we obtain on the same hardware with prior post-processing approaches (e.g. Alghamdi et al. (2022)). Moreover, the modular nature of FRAPPÉ significantly reduces the cost of changing the desired notion of fairness, set of sensitive attributes, or fairness-error trade-off. For instance, a total of two base models (one per dataset) has been used for *all* the FRAPPÉ runs presented in Figure 3, while in-processing methods require retraining the entire prediction model from scratch repeatedly for each point shown in the figures.

5.2. FRAPPÉ compared to other post-processing

Post-processing techniques such as FRAPPÉ methods only train the post-hoc module instead of the entire prediction model, and hence, can mitigate group fairness for *any* class of prediction models. However, unlike FRAPPÉ, which can induce *any* quantifiable notion of fairness, prior post-processing algorithms are only applicable for specific fairness definitions and problem settings. In this section we only focus on settings that are compatible with competitive prior post-processing approaches such as FairProjection (Alghamdi et al., 2022), and show that FRAPPÉ methods perform on-par or better. To illustrate the versatility of FRAPPÉ, we consider non-gradient based models (e.g. random forests (RF)), for which in-processing techniques such as MinDiff cannot be applied directly.

In Figure 4 we compare the Pareto frontiers obtained with FRAPPÉ MinDiff to the recent method of Alghamdi et al. (2022), which significantly outperforms prior post-processing approaches. FRAPPÉ MinDiff can sometimes surpass FairProjection considerably. More specifically, compared to FairProjection, our approach can reduce the MEO by 53%, 37%, 32% and 50% on Adult, COMPAS, HSLS and ENEM, respectively, without sacrificing more than 2% of the prediction error. In Appendix D.2 we compare FRAPPÉ MinDiff with more baselines that perform worse than FairProjection, and present results with several other pre-trained model classes, i.e. logistic regression and GBMs.

5.3. Modular methods on data with partial group labels

In this section we demonstrate how FRAPPÉ methods can alleviate the challenges faced by in-processing, when only training data with partial group labels is available. We argue that the good performance of FRAPPÉ in this regime is due to its modular design and present proof-of-concept experiments on the Adult dataset. Experiment details are deferred to Appendix C.4.

Limitations of in-processing with partial group labels. It has been observed recently that in-processing methods tend to perform poorly when only partial group labels are available for training (Jung et al., 2022; Lokhande et al., 2022; Nam et al., 2022; Sohoni et al., 2022; Zhang et al., 2023). Our experiments corroborate these findings. In particular, we observe that minimizing the objective in Equation (2) can lead to overfitting the fairness regularizer term $\mathcal{L}_{fair}$, as shown in Figure 14 in Appendix D.4. Strong regularization (e.g. early-stopping (Caruana et al., 2000)) can prevent overfitting, but it may induce unnecessary underfitting of the prediction loss $\mathcal{L}_{pred}$ in Equation (2), thus hurting the fairness-error trade-off. Indeed, Figure 5 reveals that the performance of in-processing MinDiff deteriorates significantly on data with partial group labels.

To show how challenging this setting is, we also present, for reference, the performance of a naive post-processing baseline that simply outputs the same prediction as the pre-trained model with probability p, and outputs the more favorable outcome with probability $1 - p$. Varying the probability p interpolates between prioritizing prediction error (for $p = 1$) or fairness (for $p = 0$). Even though this baseline is clearly inferior to in-processing when data

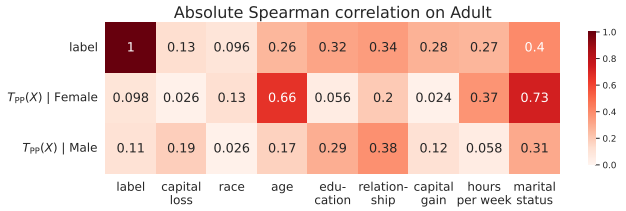


Figure 6: The post-hoc transformation $T_{\text{PP}}(X)$ is highly correlated with features that are predictive of the label (e.g. marital status, relationship), conditioned on gender.

is plentiful (Figure 5 Left), when only partial group labels are available (Figure 5 Right), in-processing struggles to surpass this naive approach.

FRAPPÉ methods with partial group labels. In-processing approaches train a single model to simultaneously minimize both the prediction loss and the fairness regularizer, and hence, finding the right balance between under- and overfitting can be challenging. In contrast, the modular FRAPPÉ methods disentangle the training of the prediction model from the fairness mitigation, and hence, allow for finer-grained control of under- and overfitting.

Indeed, Figure 5 shows a significant gap in performance between MinDiff and its FRAPPÉ variant when sensitive annotations are scarce, despite their similar performance when $\mathcal{D}_{\text{sensitive}}$ consists of all training data. More specifically, both in- and post-processing achieve similar low values of the FPR gap, but only FRAPPÉ can maintain a good prediction error in addition to good fairness. Futhermore, we show in Appendix D.5 that FRAPPÉ does not require early-stopping to outperform (early-stopped) MinDiff, thus eliminating an important hyperparameter.

5.4. Analysis: What does the post-hoc module capture?

We now provide insights about the information captured in the learned post-hoc module $T_{\text{PP}}(X)$. To this end, we use the absolute value of the Spearman's coefficient to measure the statistical correlation between the values of $T_{\text{PP}}(X)$ and each of the input features, conditioned on the sensitive attribute (i.e. gender). For visualization purposes, we focus on datasets with a small number of covariates, i.e. Adult and COMPAS (we defer results on COMPAS to Appendix D.6).

On the one hand, $T_{\text{PP}}(X)$ is expected to be correlated with the sensitive attribute (as illustrated in Appendix D.6), since it is trained specifically to improve fairness with respect to that sensitive attribute. On the other hand, the FRAPPÉ post-hoc module is allowed to depend on all the covariates, not only the sensitive attribute. Therefore, alongside information about gender, $T_{\text{PP}}(X)$ can also embed features that help to achieve a better fairness-error trade-off. Indeed, Figure 6 suggests that $T_{\text{PP}}(X)$ tends to be conditionally correlated with features that are predictive of the class label (e.g. marital status, relationship). In contrast, a group-dependent transformation (like the ones considered by prior post-processing methods) would be conditionally independent of all features. The explicit dependence of $T_{\text{PP}}(X)$ on the entire X (instead of only A), helps FRAPPÉ methods to achieve better fairness-error trade-offs than group-dependent post-processing techniques, as indicated in Section 5.2.

6. Related work

We now highlight the advantages and shortcomings of prior in- and post-processing methods (summarized in Table 1).

In-processing methods can easily induce virtually any quantifiable notion of group fairness into a prediction model (Beutel et al., 2019; Prost et al., 2019; Mary et al., 2019; Cho et al., 2020; Lowy et al., 2022; Baharlouei et al., 2024). Moreover, different mathematical tools can be used to enforce a fairness definition, e.g. EqOdds can be induced using HSIC (Pérez-Suay et al., 2017), Wasserstein distance (Jiang et al., 2020), exponential Rényi mutual information (Mary et al., 2019; Lowy et al., 2022), and Rényi correlation (Baharlouei et al., 2019). In particular, the MinDiff method of Beutel et al. (2019); Prost et al. (2019) uses MMD (Gretton et al., 2012) to great effect, can be easily scaled to multiple groups and tasks (Atwood et al., 2023), and is the standard approach for inducing EqOpp in Tensorflow[5] thanks to its good performance.

On the other hand, in-processing methods require access to the training pipeline and data as they retrain a new prediction model. They can incur large computational costs (Alghamdi et al., 2022; Cruz & Hardt, 2023), and are often tailored to a specific model family (e.g. gradient-based methods (Prost et al., 2019; Lowy et al., 2022), GBMs (Cruz et al., 2023)). Moreover, any change in the fairness definition, the set of sensitive attributes or the desired fairness-error trade-off require retraining the entire prediction model from scratch. These challenges are often cited as important obstacles for the broad adoption of fairness mitigations in practice (Veale & Binns, 2017; Holstein et al., 2019).

Post-processing is often more appealing for real-world applications, since it alleviates the aforementioned shortcomings of in-processing (i.e. it makes no assumptions on the nature of the pre-trained model and is less computationally expensive (Alghamdi et al., 2022; Cruz & Hardt, 2023)). However, current post-processing methods suffer from several limitations that hamper their applicability more broadly in practice. First, existing post-processing approaches are heavily tailored to specific problem settings (e.g. binary labels (Hardt et al., 2016)) and specific fairness definitions

[5] https://www.tensorflow.org/responsible_ai/model_remediation

(e.g. statistical parity (Xian et al., 2023), or fairness definitions based on conditional mean scores (Wei et al., 2020; Alghamdi et al., 2022), which do not include, for instance, fairness notions such as calibration (Pleiss et al., 2017)).

Furthermore, to the best of our knowledge, all prior post-processing mitigations consist in a group-dependent transformation applied to a pre-trained model's outputs. This pattern has two undesired consequences. First, group-dependent transformations require that sensitive attributes are known at inference time. However, in practice, it is often infeasible to collect sensitive attributes at inference time (e.g. asking for the ethnicity of a person before predicting their credit score). Furthermore, attempting to infer the sensitive attribute for test-time samples is also undesirable, due to ethical concerns (Veale & Binns, 2017; Holstein et al., 2019), and harms caused by data biases (Chen et al., 2019; Kallus et al., 2020). Second, prior post-processing approaches only work with discrete (oftentimes even binary) sensitive attributes and cannot be applied to problems with continuous A (e.g. age, income), even though certain in-processing methods are well-suited for this setting (Mary et al., 2019).

7. Conclusion

In this paper we propose a generic framework for training a post-processing method for group fairness using a regularized in-processing objective. We show theoretically and experimentally that FRAPPÉ methods enjoy the advantages of post-processing while not degrading the good fairness-error Pareto frontiers achieved with in-processing. Unlike prior approaches, our method does not require known sensitive attributes at inference time, and can induce any quantifiable notion of fairness on a broad set of problem settings, including when sensitive attributes are continuous (e.g. age, income). Finally, we demonstrate how FRAPPÉ methods can alleviate the drop in performance that affects in-processing when only partial group labels are available.

Impact statement

The framework proposed in this work significantly expands the range of problem settings where post-processing mitigation techniques can be applied. In particular, FRAPPÉ methods can be employed to induce fairness in applications with limited computational resources or with no access to the training pipeline and training data of the prediction model. In addition, FRAPPÉ post-processing methods can help to overcome challenges faced by in-processing methods such as compositional fairness problems (Dwork & Ilvento, 2018; Atwood et al., 2023).

In contrast to FRAPPÉ methods, which are trained on triples $(\hat{Y}, A, Y)$, prior post-processing methods require access to (X, A, Y) for training. However, even for prior methods, access to the features X is still necessary in order to obtain the predictions $\hat{Y}$. Regarding the computation time required to train FRAPPÉ methods, we note that it is similar to other competitive post-processing methods (e.g. Alghamdi et al. (2022)). Both FRAPPÉ methods and prior post-processing techniques perform certain computations to find an appropriate post-hoc transformation. In the case of our approach, it suffices to optimize the parameters of a simple linear regression model to obtain the results shown in our experiments.

When it comes to evaluating algorithmic fairness, popular datasets such as Adult and COMPAS suffer from several limitations which have been pointed out a number of recent works (Bao et al., 2021; Ding et al., 2021; Alghamdi et al., 2022). For this reason, we also report our main experimental results of Sections 5.1 and 5.2 on two recently proposed datasets, HSLS (Jeong et al., 2022) and ENEM (Alghamdi et al., 2022), which specifically address concerns raised about Adult and COMPAS.

Finally, our work does not attempt to provide new arguments in favor of algorithmic fairness. As frequently noted in the ML fairness literature (Corbett-Davies et al., 2017; Kasy & Abebe, 2021; Bao et al., 2021; Barocas et al., 2023), algorithmic interventions to induce fairness are not always aligned with the intended societal impact. Therefore, it remains the object of active research whether notions such as SP, EqOdds and EqOpp are suitable for evaluating the inequity of decision systems (Buyl & De Bie, 2022; Ruggieri et al., 2023; Majumder et al., 2023). Furthermore, we note that our work focuses specifically on mitigating group fairness. Investigating whether our findings also apply to other notions of equity, such as individual fairness (Dwork et al., 2012), remains an important direction for future work.

Acknowledgements

We are grateful to Alexander D'Amour, Ananth Balashankar, Amartya Sanyal, Flavio Calmon, and Jilin Chen for helpful discussions, and thank the anonymous reviewers for feedback on the manuscript. We also thank Flavio Calmon, Hsiang Hsu, Sina Baharlouei and Tom Stepleton for their help with reproducing some of the prior work results.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proceedings of the International Conference on Machine Learning*, 2018.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P. W., Asoodeh, S., and Calmon, F. P. Beyond Adult and COMPAS: Fairness in multi-class prediction via information projection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias. *Propublica*, 2016.

Atwood, J., Tian, T., Packer, B., Deodhar, M., Chen, J., Beutel, A., Prost, F., and Beirami, A. Towards a scalable solution for improving multi-group fairness in compositional classification, 2023.

Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., and Wang, X. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. Rényi fair inference. In *International Conference on Learning Representations*, 2019.

Baharlouei, S., Patel, S., and Razaviyayn, M. f-FERM: A scalable framework for robust fair empirical risk minimization. In *International Conference on Learning Representations*, 2024.

Bao, M., Zhou, A., Zottola, S. A., Brubach, B., Desmarais, S., Horowitz, A. S., Lum, K., and Venkatasubramanian, S. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Bardenhagen, V., Tifrea, A., and Yang, F. Boosting worst-group accuracy without group annotations. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996.

Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., and Chi, E. H. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. Sex bias in graduate admissions: Data from Berkeley. *Science*, 1975.

Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 1967.

Buyl, M. and De Bie, T. Inherent limitations of AI fairness, 2022.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops*, 2009.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019.

Caruana, R., Lawrence, S., and Giles, C. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, 2000.

Caton, S. and Haas, C. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2023.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, 2018.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvtiskii, S. Matroids, matchings, and fairness. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

Cho, J., Hwang, G., and Suh, C. A fair classifier using kernel density estimation. In *Advances in Neural Information Processing Systems*, 2020.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*, 2020.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

Cortes, C., Mohri, M., and Rostamizadeh, A. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 2012.

Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., and Chakraborty, S. Fair transfer learning with missing protected attributes. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Cotter, A., Jiang, H., Wang, S., Narayan, T., You, S., Sridharan, K., and Gupta, M. R. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 2019.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. Wiley-Interscience, 1991.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. On kernel-target alignment. In *Advances in Neural Information Processing Systems*, 2001.

Cruz, A., Belém, C. G., Bravo, J., Saleiro, P., and Bizarro, P. FairGBM: Gradient boosting with fairness constraints. In *The Eleventh International Conference on Learning Representations*, 2023.

Cruz, A. F. and Hardt, M. Unprocessing seven years of algorithmic fairness, 2023.

Cruz, A. F., Saleiro, P., Belém, C., Soares, C., and Bizarro, P. Promoting fairness through hyperparameter optimization. In *IEEE International Conference on Data Mining (ICDM)*, 2021.

Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women, 2018. URL https://www.reuters.com/article/idUSKCN1MK0AG/.

Dheeru, D. and Karra Taniskidou, E. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, 2021.

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

Dwork, C. and Ilvento, C. Fairness under composition. In *Innovations in Theoretical Computer Science Conference*, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference*, 2012.

Gebelein, H. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, 1941.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *Conference on Algorithmic Learning Theory*, 2005.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 2012.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *Proceedings of the International Conference on Machine Learning*, 2018.

He, X., Zhao, K., and Chu, X. AutoML: A survey of the state-of-the-art. *Journal of Knowledge-Based Systems*, 2021.

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the Conference on Human Factors in Computing Systems*, 2019.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 2023.

Jeong, H., Wang, H., and Calmon, F. P. Fairness without imputation: A decision tree approach for fair prediction with missing values. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein fair classification. In *Proceedings of The Uncertainty in Artificial Intelligence Conference*, 2020.

Jung, S., Chun, S., and Moon, T. Learning fair classifiers with partially annotated group labels. In *Conference on Computer Vision and Pattern Recognition*, 2022.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020.

Kamiran, F., Mansha, S., Karim, A., and Zhang, X. Exploiting reject option in classification for social discrimination control. *Journal of Information Sciences*, 2018.

Kasy, M. and Abebe, R. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., and Togneri, R. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

Kim, J. S., Chen, J., and Talwalkar, A. FACT: A diagnostic for group fairness trade-offs. In *Proceedings of the International Conference on Machine Learning*, 2020.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019.

LaBonte, T., Muthukumar, V., and Kumar, A. Towards last-layer retraining for group robustness with fewer annotations. In *Advances in Neural Information Processing Systems*, 2023.

Lahoti, P., Gummadi, K. P., and Weikum, G. Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment*, 2019.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, 2020.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. Just Train Twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Lokhande, V. S., Sohn, K., Yoon, J., Udell, M., Lee, C.-Y., and Pfister, T. Towards group robustness in the presence of partial group labels. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Lowy, A., Baharlouei, S., Pavan, R., Razaviyayn, M., and Beirami, A. A stochastic optimization framework for fair risk minimization. *Transactions on Machine Learning Research*, 2022.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations, 2018.

Majumder, S., Chakraborty, J., Bai, G. R., Stolee, K. T., and Menzies, T. Fair enough: Searching for sufficient measures of fairness. *ACM Trans. Softw. Eng. Methodol.*, 2023.

Mary, J., Calauzènes, C., and Karoui, N. E. Fairness-aware learning for continuous attributes and treatments. In *Proceedings of the International Conference on Machine Learning*, 2019.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 2021.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, 2018.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021a.

Menon, A. K., Rawat, A. S., and Kumar, S. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021b.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2022.

Nandy, P., DiCiccio, C., Venugopalan, D., Logan, H., Basu, K., and El Karoui, N. Achieving fairness via post-processing in web-scale recommender systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Narayanan, A. 21 fairness definitions and their politics. *Tutorial at the Conference on Fairness, Accountability, and Transparency*, 2018.

Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 1972.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. Fair kernel learning. In *Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

Prost, F., Qian, H., Chen, Q., Chi, E. H., Chen, J., and Beutel, A. Toward a better trade-off between performance and fairness with kernel-based distribution matching, 2019.

Prost, F., Awasthi, P., Blumm, N., Kumthekar, A., Potter, T., Wei, L., Wang, X., Chi, E. H., Chen, J., and Beutel, A. Measuring model fairness under noisy covariates: A theoretical perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the International Conference on Machine Learning*, 2020.

Redmond, M. Communities and Crime. UCI Machine Learning Repository, 2009.

Ruggieri, S., Alvarez, J. M., Pugnana, A., State, L., and Turini, F. Can we trust Fair-AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Sanyal, A., Hu, Y., and Yang, F. How unfair is private learning? In *Conference on Uncertainty in Artificial Intelligence*, 2022.

Schapire, R. E. The strength of weak learnability. *Machine Learning*, 1990.

Shi, Y., Daunhawer, I., Vogt, J. E., Torr, P., and Sanyal, A. How robust is unsupervised representation learning to distribution shift? In *International Conference on Learning Representations*, 2023.

Sohoni, N. S., Sanjabi, M., Ballas, N., Grover, A., Nie, S., Firooz, H., and Re, C. BARACK: Partially supervised group robustness with guarantees. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Vapnik, V. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, 1991.

Veale, M. and Binns, R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 2017.

Veldanda, A. K., Brugere, I., Chen, J., Dutta, S., Mishler, A., and Garg, S. Fairness via in-processing in the over-parameterized regime: A cautionary tale with MinDiff loss. *Transactions on Machine Learning Research*, 2023.

Wainwright, M. and Jordan, M. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 2008.

Wei, D., Ramamurthy, K. N., and Calmon, F. Optimized score transformation for fair classification. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2020.

Xian, R., Yin, L., and Zhao, H. Fair and Optimal Classification via Post-Processing. In *Proceedings of the International Conference on Machine Learning*, 2023.

Zehlike, M., Yang, K., and Stoyanovich, J. Fairness in ranking: A survey, 2021.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

Zhang, F., Kuang, K., Chen, L., Liu, Y., Wu, C., and Xiao, J. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *International Conference on Learning Representations*, 2023.

Zhang, G., Zhang, Y., Zhang, Y., Fan, W., Li, Q., Liu, S., and Chang, S. Fairness reprogramming. In *Advances in Neural Information Processing Systems*, 2022.

Zhao, H. and Gordon, G. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems*, 2019.

A. Proof of Proposition 3.1

Proposition A.1. *Consider the optimization objectives introduced in Equations (2) and (4). There exists a constant $C \in \mathbb{R}$ such that for any $\boldsymbol{\theta} \in \mathbb{R}^D$ and $\lambda \geq 0$ we have*

$$OPT_{PP}(\boldsymbol{\theta}; \lambda) = OPT_{IP}(\boldsymbol{\theta}; \lambda) + C. \tag{5}$$

Proof. By the definition of $\boldsymbol{\theta}_{\text{base}}$ and from the first order optimality condition it holds that:

$$\nabla F(\boldsymbol{\theta}_{\text{base}}) = \frac{1}{n} \sum_{i \in [n]} G(\boldsymbol{x}_i, y_i).$$

Plugging this identity into the Bregman divergence of the strongly convex function $F(\boldsymbol{\theta})$ allows us to write it as follows:

$$\begin{aligned} D_F(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{base}}) &= F(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \frac{1}{n} \sum_{i \in [n]} G(\boldsymbol{x}_i, y_i) \\ &\quad + \boldsymbol{\theta}_{\text{base}}^\top \frac{1}{n} \sum_{i \in [n]} G(\boldsymbol{x}_i, y_i) - F(\boldsymbol{\theta}_{\text{base}}) \\ &= \frac{1}{n} \sum_{i \in [n]} \mathcal{L}_{pred}(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}) + C, \end{aligned}$$

where we use the notation $C = \boldsymbol{\theta}_{\text{base}}^\top \frac{1}{n} \sum_{i \in [n]} G(\boldsymbol{x}_i, y_i) - F(\boldsymbol{\theta}_{\text{base}})$ for the terms that are independent of $\boldsymbol{\theta}$.

Rearranging the terms yields that $\text{OPT}_{\text{PP}}(\boldsymbol{\theta}; \lambda) = \text{OPT}_{\text{IP}}(\theta; \lambda) + C$ which concludes the proof. □

B. More related work

First, in order to help position our work in the existing in- and post-processing literature, we present in Table 1 the specific shortcomings that we target with the desiderata D1 – D3 introduced in Section 1.

Method	Require changing prediction model	Require access to training pipeline / data	Computation cost	Sensitive attribute A	Requires A for inference	Fairness definition
In-processing methods e.g. Agarwal et al. (2018); Prost et al. (2019), Mary et al. (2019); Cho et al. (2020)	Yes	Yes	High	Any	No	Any fairness penalty
Existing post-processing methods e.g. Hardt et al. (2016); Kamiran et al. (2018), Alghamdi et al. (2022); Xian et al. (2023)	No	No	Low	Discrete	Yes	Tailored to specific fairness definitions
FRAPPÉ methods	No	No	Low	Any	No	Any fairness penalty

Table 1: In-processing requires retraining the entire prediction model to induce fairness, but can be applied to a broad range of problem settings and to virtually any quantifiable notion of fairness. On the other hand, current post-processing methods are tailored to specific settings and fairness definitions. In contrast, FRAPPÉ methods are as broadly applicable as penalized in-processing methods, while not being confined to applications with access to the training pipeline of the prediction model.

In the remainder of this section we elaborate on some of the limitations of in-processing that have been previously documented in the literature and have not been discussed extensively in Section 6.

In-processing and compositional fairness. It is often the case, in practical applications that multiple prediction models are employed, and their outputs are then all aggregated into a single decision. For instance, a candidate may apply for several jobs, each with their own selection criteria, but the outcome that is of interest to the candidate is whether at least one of the applications is successful. Similarly, complex decision problems may be broken down into finer grained tasks for the purpose of better interpretability. These situations are prone to compositional fairness issues (Dwork & Ilvento, 2018): even if all individual components are fair, it is not guaranteed that the aggregated decision will also be fair. In-processing

techniques are inherently susceptible to limitations due to compositional fairness (Atwood et al., 2023). Indeed, in order to mitigate these issues, in-processing method could train all individual models simultaneously while enforcing that the aggregated decision is fair. However, this procedure raises huge logistical challenges which are often insurmountable in practice. In contrast, post-processing bypasses compositional fairness issues altogether. Applying a post-processing method to the final decision of a complex system with multiple prediction components that are aggregated into a single decision can treat the entire decision system as a black-box and overcome limitations due to compositional fairness. A thorough investigation of this intuition is left as future work.

In-processing when data has partial group labels. In addition to the works mentioned in Section 5.3, there have been a few empirical observations that attempt to study this problem. In particular, Veldanda et al. (2023) investigate the performance of MinDiff (Prost et al., 2019) in the overparameterized regime, where the complexity of the model fit to the training data increases while the sample size stays fixed. The authors show in experiments on image data that explicit regularization (e.g. early stopping) can improve the performance of MinDiff with overparameterized models. However, the impact of overparameterization on the fairness-error Pareto frontiers is not studied in detail.

Furthermore, Lowy et al. (2022) present experiments where the amount of data with sensitive annotations is reduced to 10%. In this case, the authors aim to compare their proposed method to other in-processing strategies, but no catastrophic loss in performance is noticeable. We hypothesize that this is due to the amount of data with sensitive attributes being still large enough to allow for good performance. In particular, our experiments (Figure 5) reveal that when the sensitive data is reduced to 0.1% of the training set size, on Adult in-processing performs no better than the naive strategy of predicting the favorable outcome with probability p and the output of the pre-trained model with probability $1-p$.

C. Experiment details

C.1. Details on fairness definitions

In this section, we provide more details about the notions of group fairness used throughout this paper. We note that this is not an exhaustive list of fairness definitions, and other notions are possible and considered in the literature too (e.g. worst-group error). We refer to surveys such as Caton & Haas (2023) books like Barocas et al. (2023) for more details.

Statistical parity (SP). Also known as demographic parity, SP measures the difference between the frequency of favorable outcomes in the subpopulations determined by the values of the sensitive attribute A (Dwork et al., 2012). To quantify the violation of the SP condition, several works (Donini et al., 2018; Jiang et al., 2020; Cho et al., 2020) consider the difference with respect to statistical parity. Assuming that the favorable outcome is $y=1$, this quantity is defined as follows:

$$\text{DSP}(f) = \sum_a |\mathbb{P}(f(X)=1|A=a) - \mathbb{P}(f(X)=1)|, \tag{8}$$

where the sum is over all the possible values of the sensitive attribute. Note that the sum can also be replaced with a "$\max$" operator in the formulation above.

Equal opportunity (EqOpp). This fairness definition is tailored for settings with discrete labels y and sensitive attributes a. Intuitively, EqOpp asks that a classifier is not more likely to assign the favorable outcome to one of the groups determined by the (discrete) sensitive attribute a. Assuming the negative class $y=0$ is more desirable, one can quantify the fairness of a binary predictor using the following:

$$\text{FPRgap}(f) = |\mathbb{P}(f(X)\neq Y|Y=0,A=0) - \mathbb{P}(f(X)\neq Y|Y=0,A=1)|. \tag{9}$$

This metric can also be generalized to multiclass classification.

Equalized odds (EqOdds). This notion of fairness is satisfied if $A \perp \hat{Y}|Y$. Intuitively, EqOdds penalizes the predictor if it relies on potential spurious correlations between A and Y. One can quantify the violation of this definition of fairness using $\rho(A,\hat{Y}|Y)$, where ρ is a measure of conditional statistical independence (e.g. HSIC (Gretton et al., 2005), CKA (Cristianini et al., 2001; Cortes et al., 2012), HGR (Gebelein, 1941) etc). One can either use one of these quantities to evaluate the fairness of a model (e.g. $\text{HGR}_\infty(f(X),A|Y)$ like in Mary et al. (2019)) or a metric such as mean equalized odds, which,

for binary classification, can be defined as:

$$\text{MEO} = \frac{\text{TPRgap}(f) + \text{FPRgap}(f)}{2}, \tag{10}$$

where $\text{TPRgap}(f)$ is the gap in the true positive rate between groups and is defined similarly to $\text{FPRgap}(f)$.

C.2. Datasets

We briefly describe the datasets used throughout the experiments presented in the paper.

The **Adult** dataset (Becker & Kohavi, 1996) is perhaps the most popular dataset in the algorithmic fairness literature. The task it proposes is to predict whether the income of a person is over the $50,000$\$ threshold, having access to various demographic features. In our experiments, we consider gender as the sensitive attribute. We follow the procedure described in Alghamdi et al. (2022) to pre-process the data.

Alongside Adult, **COMPAS** (Angwin et al., 2016) is also a well-established dataset for evaluating fairness mitigations. It contains information about defendents detained in US prisons. The task is to predict the individual risk of recidivism, while being fair with respect to race. We adopt the pre-processing methodology of Alghamdi et al. (2022) for this dataset.

The **Crimes & Communities** dataset (Redmond, 2009) is also part of the UCI repository (Dheeru & Karra Taniskidou, 2017), like Adult, and contains information about US cities. The task is a regression problem where the goal is to predict the amount of violent crimes and the sensitive attribute is the proportion of an ethnic group in the population. Hence, the sensitive attribute takes continuous values. For this dataset, we use the same pre-processing as Mary et al. (2019).

The **HSLS** dataset (Jeong et al., 2022) contains information about over $23,000$ students from high schools in the USA. The features consist in information about the students' demographic and academic performance, as well as data about the schools. The data is pre-processed using the same procedure as Alghamdi et al. (2022). The task is to predict exam scores while being fair with respect to race.

ENEM (Alghamdi et al., 2022) is a dataset of exam scores collected in Brazilian high schools. The dataset contains demographic and socio-economic information about the students. Once again, we use the same pre-processing methodology as Alghamdi et al. (2022). Similar to HSLS, the goal is to predict the Humanities exam score, while the sensitive attribute is race.

C.3. Baselines

We compare the performance of FRAPPÉ methods obtained with our framework with several competitive in- and post-processing approaches.

In-processing baselines. We consider three different regularized in-processing methods and one constrained in-processing approach for which we construct FRAPPÉ post-processing counterparts. First, MinDiff (Beutel et al., 2019; Prost et al., 2019) is an approach that uses MMD (Gretton et al., 2012) to induce the statistical independence required for various fairness definitions to hold (i.e. EqOpp, EqOdds). The remarkable performance of this method led to it being included in standard fairness toolkits such as `tensorflow-model-remediation`.[6] Furthermore, the method of Cho et al. (2020) employs kernel density estimation to construct a regularizer for certain fairness definition violations. In addition to EqOdds, this method can also be applied to induce SP. Finally, Mary et al. (2019) propose to use the Hirschfeld-Gebelein-Rényi (HGR) Maximum Correlation Coefficient to quantify statistical independence and propose an unfairness regularizer based on this metric. Besides these regularized in-processing methods, we also consider the Reductions approach (Agarwal et al., 2018) in our comparison, which proposes solving a constrained optimization problem. For all the baselines, we use the hyperparameters recommended in the respective papers.

Post-processing baselines. The FairProjection method of Alghamdi et al. (2022) is, to the best of our knowledge, one of the best performing post-processing mitigations. FairProjection adjusts the scores output by a classification method, using a different transformation for each sensitive group in the population. Alternatively, the methods of Hardt et al. (2016); Chzhen et al. (2020) change the decision threshold in a group-dependent manner. These two approaches do not prescribe a way

[6] https://www.tensorflow.org/responsible_ai/model_remediation

to obtain an entire Pareto frontier, but rather a single point on the fairness-error trade-off. Finally, the Rejection-option classification method of Kamiran et al. (2018) exploits uncertainty in the decision of a classifier to decide what labels to output. For all of these methods, we use the results from the public code repository of Alghamdi et al. (2022).

C.4. Experiment details for training FRAPPÉ methods

For the comparison with in-processing methods, we use the pre-trained models recommended in the respective papers (i.e. 3-MLP with 128 hidden units on each layer for MinDiff and Cho et al. (2020), and logistic regression for Agarwal et al. (2018) and Mary et al. (2019). For the FRAPPÉ post-hoc transformation, we use linear regression to model $T_{\text{PP}}(\boldsymbol{x})$. We select the optimal learning rate by minimizing the prediction error on a held-out validation set. To obtain the Pareto frontiers, we vary the λ coefficient that balances the prediction error and the fairness regularizer terms in the loss.

For the comparison with prior post-processing works in Section 5.2 we use FRAPPÉ MinDiff. For these experiments, we employ a variant of MinDiff tailored to EqOdds which encourages not only the FPR gap to be small between sensitive groups, but also the FNR. Once again, we select the optimal learning rate using a validation set, and train linear regression and 1-MLP models with 64 hidden units as the $T_{\text{PP}}(\boldsymbol{x})$ post-hoc transformation. The pre-trained models are obtained following the instructions in Alghamdi et al. (2022).

For the analysis of the post-hoc transformation (Section 5.4), we use a 3-MLP as the pre-trained model and assume $T_{\text{PP}}(X)$ to be a 1-MLP trained using FRAPPÉ MinDiff, like in Section 5.1. We only consider Adult and COMPAS since they have fewer covariates, which makes them suitable for visualization.

For the experiments with partial group labels in Section 5.3 we consider FRAPPÉ MinDiff for EqOpp, with a 3-MLP pre-trained model with 128 hidden units on each layer. We use a 1-MLP with 64 hidden units to model the post-processing transformation. The optimal learning rate and early-stopping epoch are selected so as to minimize prediction error on a held-out validation set.

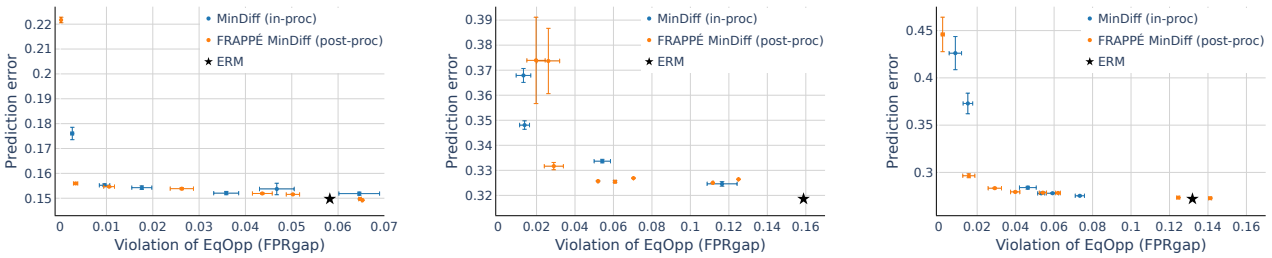
C.5. Measuring computation cost

To compare the computation time of FRAPPÉ methods and compare it to the correspoding in-processing methods we generate Pareto frontiers for each of the settings in Figure 3, where we always select 8 different values for the coefficient that controls the fairness-error trade-off and repeat each experiment 10 times with different random seeds. In total, for each of the three settings in Figure 3 we perform 80 experiments sequentially. For the post-processing methods, we include the time required to train a base model in the reported computation times. The machine we used for these measurements has 32 1.5 GHz CPUs.

D. More experiments

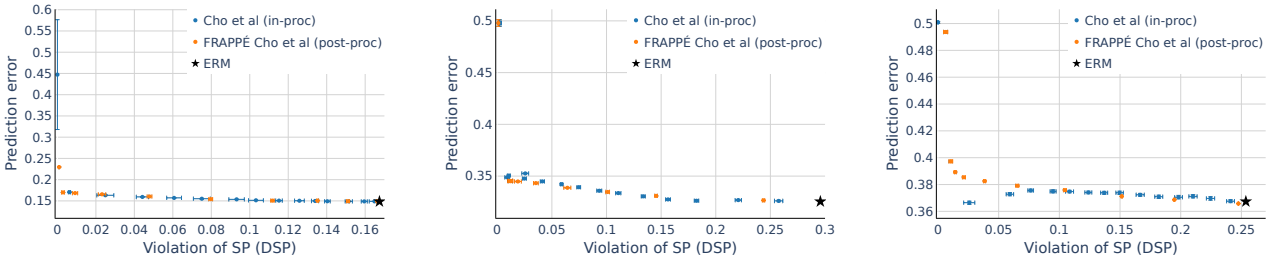
D.1. Equivalence between in- and post-processing on more datasets

Figure 3 demonstrates on the HSLS dataset that FRAPPÉ methods preserve the good fairness-error trade-off achieved by several regularized in-processing approaches, while also enjoying the advantages of post-processing methods. Figures 7 to 9 complement Figure 3 and present similar results on three more datasets: Adult, COMPAS and ENEM.



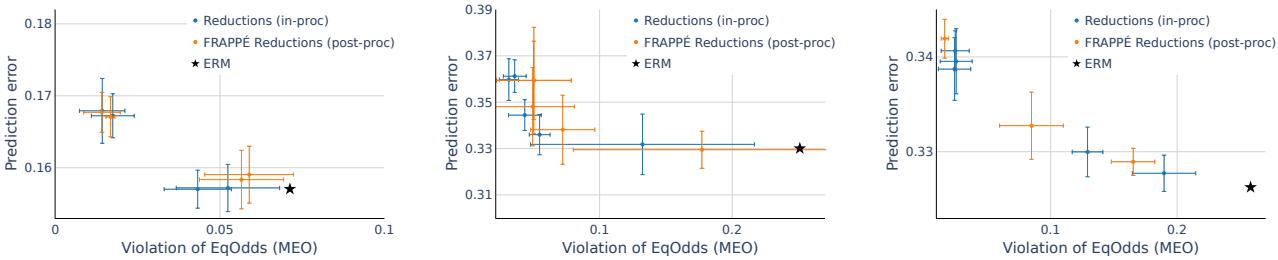
(a) $\text{OPT}_{\text{IP}} = \textit{MinDiff}$ on Adult data. (b) $\text{OPT}_{\text{IP}} = \textit{MinDiff}$ on COMPAS data. (c) $\text{OPT}_{\text{IP}} = \textit{MinDiff}$ on ENEM data.

Figure 7: Inducing EqOpp using in-processing MinDiff and its FRAPPÉ post-processing variant leads to similar Pareto frontiers on several different datasets.



(a) OPT_{IP} = Cho et al. (2020) on Adult. (b) OPT_{IP} = Cho et al. (2020) on COMPAS. (c) OPT_{IP} = Cho et al. (2020) on ENEM.

Figure 8: Inducing SP using the in-processing method of Cho et al. (2020) and its FRAPPÉ post-processing variant leads to similar Pareto frontiers on several different datasets.



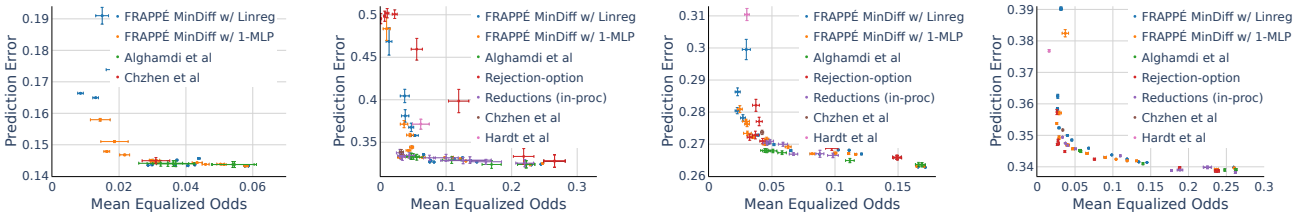
(a) OPT_{IP} = *Reductions* on Adult. (b) OPT_{IP} = *Reductions* on COMPAS. (c) OPT_{IP} = *Reductions* on ENEM.

Figure 9: Inducing EqOdds using in-processing Reductions (Agarwal et al., 2018) and its FRAPPÉ post-processing variant leads to similar Pareto frontiers on several different datasets.

D.2. More comparisons with prior mitigations

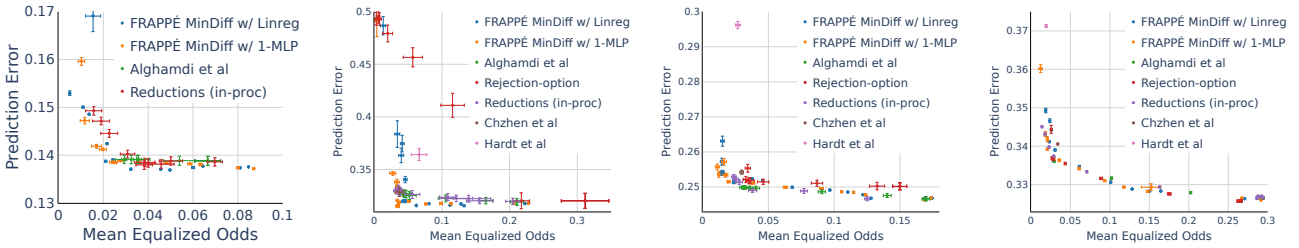
In this section we extend Figure 4 with the results obtained with more in- and post-processing baselines. We consider the same methods as Alghamdi et al. (2022), described in more detail in Appendix C.3. Unless otherwise specified, the techniques presented in Figure 10 are post-processing approaches. Like in Figure 4, we train FRAPPÉ MinDiff for EqOdds, where the post-hoc transformation is modeled by either linear regression or a simple 1-MLP with 64 hidden units. The numbers for all the baselines are collected from the public code of Alghamdi et al. (2022).

In addition to using random forests (RF) as the base model, we also present results for logistic regression and GBMs as the pre-trained model for all three datasets in Figure 12 and Figure 11, respectively. The figures reveal that the same trends observed for RF pre-trained models also occur for other classes of pre-trained models.



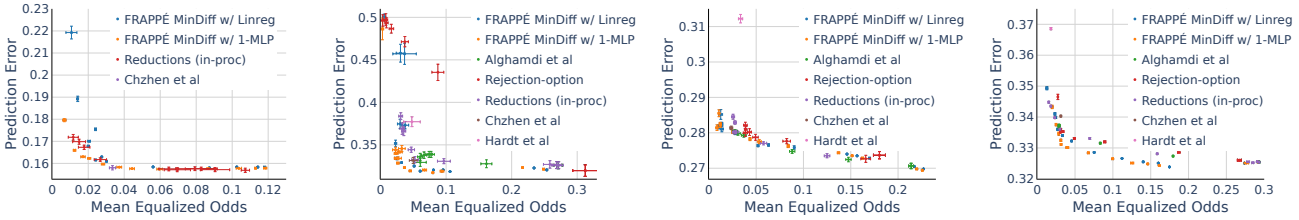
(a) Adult (b) COMPAS (c) HSLS (d) ENEM

Figure 10: Comparison between FRAPPÉ MinDiff for EqOdds and several other in- and post-processing methods for inducing group fairness. The pretrained model is **random forest (RF)**.



(a) Adult (b) COMPAS (c) HSLS (d) ENEM

Figure 11: Comparison between FRAPPÉ MinDiff for EqOdds and several other in- and post-processing methods for inducing group fairness. In contrast to Figure 10, here the pre-trained model is **gradient-boosted machine (GBM)**.



(a) Adult (b) COMPAS (c) HSLS (d) ENEM

Figure 12: Comparison between FRAPPÉ MinDiff for EqOdds and several other in- and post-processing methods for inducing group fairness. In contrast to Figure 10, here the pre-trained model is **logistic regression**.

D.3. Comparison with model reprogramming

Model reprogramming aims to reuse a pretrained model and adjust the inputs in order to elicit outputs with a desired property. In particular, for group fairness, Zhang et al. (2022) consider a somewhat similar optimization objective to FRAPPÉ. However, unlike FRAPPÉ, the method of Zhang et al. (2022) (i.e. Fairness Reprogramming) learns the parameters of a post-hoc transformation of the inputs of a pre-trained prediction model. On the one hand, on image data, choosing this transformation to be a border or a patch (see Figure 1 in Zhang et al. (2022)) leads to remarkable results. To illustrate how FRAPPÉ methods perform on CelebA data, we consider the same experimental setting as in Zhang et al. (2022) and provide in Figure 13 the Pareto frontier obtained with the FRAPPÉ version of FERMI (Lowy et al., 2022).

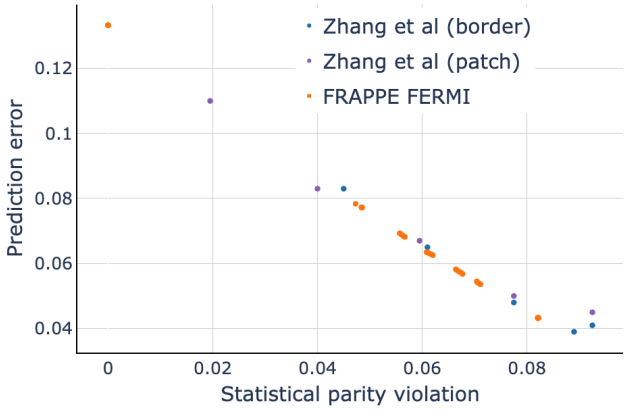


Figure 13: FRAPPÉ FERMI leads to similar Pareto frontiers as Fairness Reprogramming (Zhang et al., 2022).

Our experiments reveal that FRAPPÉ FERMI performs similarly to Fairness Reprogramming on this dataset. We note, however, that unlike Fairness Reprogramming, FRAPPÉ methods do not need to carefully select the family of post-hoc transformations. For Fairness Reprogramming, this choice has great influence on performance, as indicated, for instance, by Figures 3 and 4 in Zhang et al. (2022). Moreover, for Fairness Reprogramming the post-hoc transformation is specific

to a data modality, and different problems may require a human expert to design a set of reasonable candidate post-hoc transformations.

In fact, for tabular data, results in Zhang et al. (2022) suggest that Fairness Reprogramming performs worse than standard techniques such as Zhang et al. (2018), which in turn is outperformed by more recent approaches like Cho et al. (2020). We hypothesize that constructing an appropriate parametric transformation of the inputs (the so-called trigger) for the Fairness Reprogramming method is more challenging for structured tabular data than it is for image or text modalities. In contrast, the FRAPPÉ variant of Cho et al. (2020) matches the Pareto frontiers of the in-processing counterpart on several datasets (including Adult), as indicated in Figures 3b and 8, and thus outperforms both Zhang et al. (2018) and Zhang et al. (2022).

D.4. In-processing MinDiff overfits the fairness regularizer

In this section we provide evidence that suggests that regularized in-processing objectives can overfit the fairness regularizer when trained on data with partial group labels. In particular, we consider MinDiff run on the Adult dataset. We subsample the dataset with sensitive attributes to be only 0.1% of the original training data. As described in Section 5.3, we use the entire training data for the predictive term in the loss.

Figure 14 shows the median EqOpp violation (i.e. the FPR gap) as a function of the number of training epochs of in-processing MinDiff. The learning curves in the figure indicate that when the data with sensitive attributes is sufficiently large, the fairness violation is low on both training and test data. However, for partial group labels, in-processing MinDiff quickly achieves a vanishing fairness regularizer on the training data, while the test FPR gap continues to increase during training.

(a) $|\mathcal{D}_{\text{sensitive}}| = 100\% \cdot |\mathcal{D}_{\text{pred}}|$ (b) $|\mathcal{D}_{\text{sensitive}}| = 0.1\% \cdot |\mathcal{D}_{\text{pred}}|$

Figure 14: In-processing MinDiff achieves a low FPR both on training and test data if data with sensitive attributes is plentiful. When only partial group labels are available, the training FPR gap vanishes, while on the test set the unfairness of the model increases during training. Experiments run on the Adult dataset.

D.5. FRAPPÉ MinDiff without early-stopping for data with partial group labels

Figure 5 shows that with optimal early-stopping regularization, FRAPPÉ MinDiff significantly outperforms its in-processing counterpart when only partial group labels are available for training. In this section, we present evidence that suggests that even without early-stopping, FRAPPÉ post-processing can perform well in this setting. In Figure 15 we compare in-processing MinDiff regularized with early-stopping (like in Figure 5) to FRAPPÉ MinDiff *with no regularization*. The Pareto frontier for FRAPPÉ is once again better than for the in-processing method. Moreover, without the need to regularize the FRAPPÉ method, training does away with careful hyperparameter tuning of the important early stopping iteration.

D.6. More results on what is captured by the learned post-hoc transformation

In this section we complement Figure 6 with additional evidence that the post-hoc transformation learned by FRAPPÉ methods is correlated not only with the sensitive attribute A, but also with features that are predictive of the target class label. Figure 17 shows that the same trends observed on COMPAS, also occur for the Adult dataset.

In addition, Figures 18 and 19 show the (absolute value of the) conditional correlation between $T_{\text{PP}}(X)$ and each of the features, given the sensitive attribute A. For reference, a group-dependent transformation, such as the prior post-processing techniques (Hardt et al., 2016; Alghamdi et al., 2022; Xian et al., 2023) would be constant given A, and hence, statistically

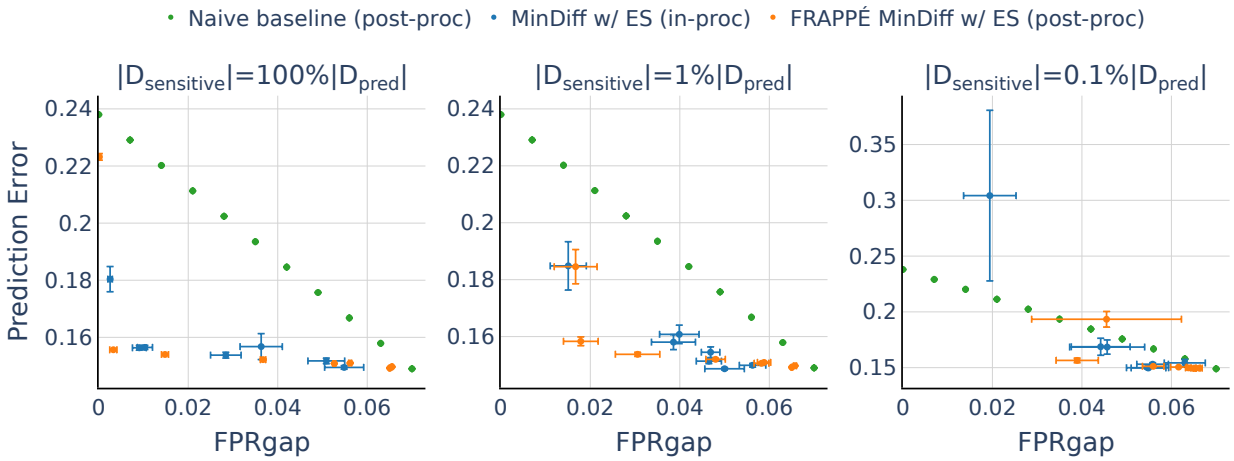


Figure 15: In-processing MinDiff with early-stopping regularization and FRAPPÉ MinDiff *without early-stopping* (ES). On data with partial group labels, the FRAPPÉ method continues to outperform the in-processing variant even without early-stopping regularization. Experiments run on the Adult dataset.

independent of all the other features. In contrast, the post-hoc transformation learned with FRAPPÉ is highly correlated with features that are predictive of the class label (e.g. priors count for COMPAS; age, education or marital status for Adult).

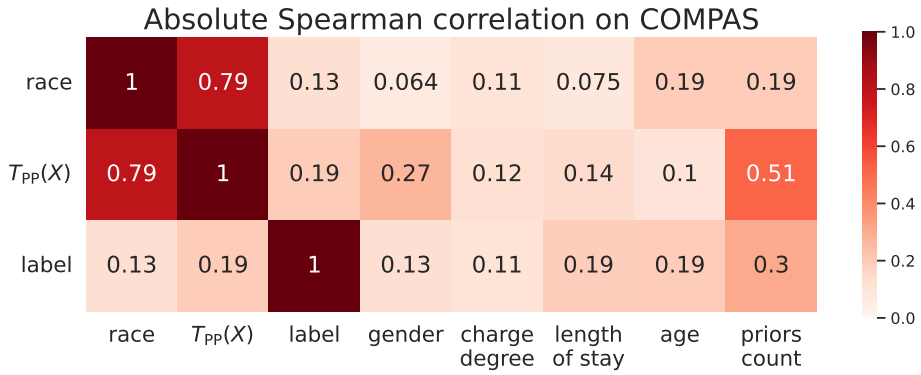


Figure 16: The value of the learned post-hoc transformation $T_{\text{PP}}(X)$ is highly correlated with both the sensitive attribute (i.e. race), as well as with features that are predictive of the class label (e.g. priors count).

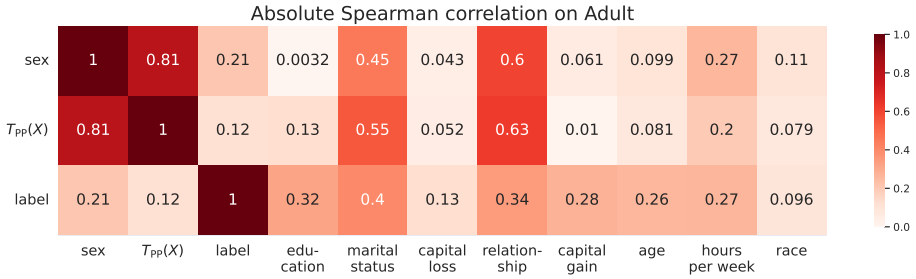


Figure 17: Counterpart of Figure 16, but for the Adult dataset. The value of the learned post-hoc transformation $T_{\text{PP}}(X)$ is highly correlated with the sensitive attribute (i.e. gender), as well as with features that are themselves correlated with A.

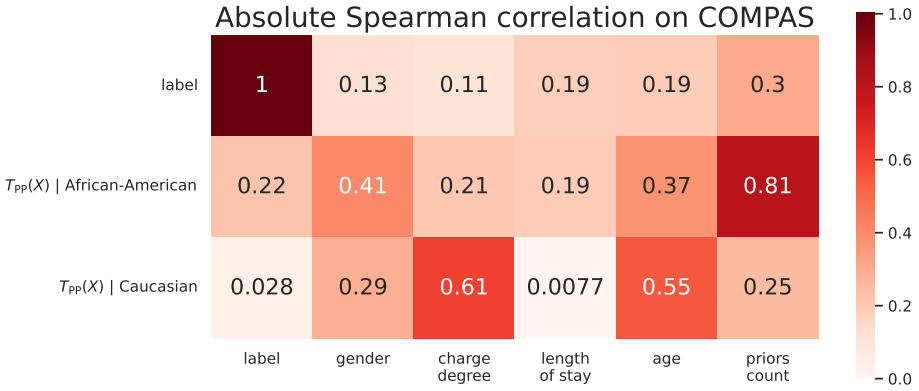


Figure 18: Conditional correlation between $T_{\text{PP}}(X)$ and each of the features, on the COMPAS dataset. The correlation is higher for features that are predictive of the label. In contrast, a group-dependent transformation would be conditionally independent of all features given A.

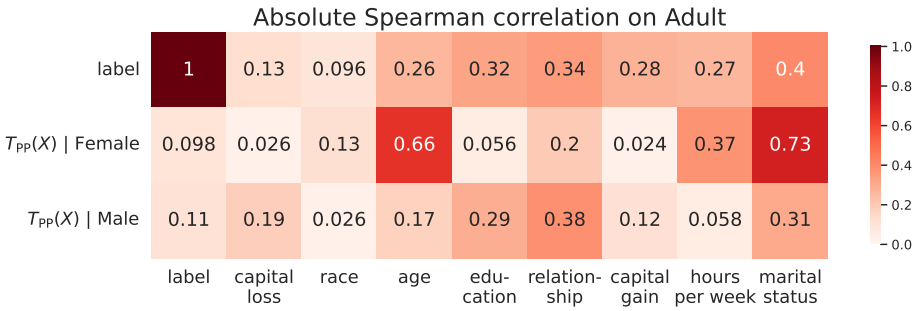


Figure 19: Conditional correlation between $T_{\text{PP}}(X)$ and each of the features, on the Adult dataset. The correlation is higher for features that are predictive of the label. In contrast, a group-dependent transformation would be conditionally independent of all features given A

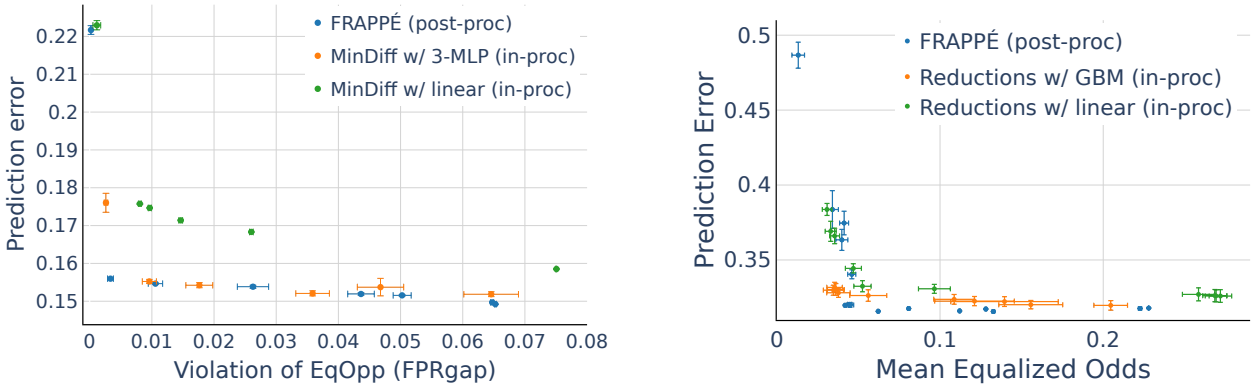
D.7. Varying function complexity for f_{base} and T_{PP}

In this section, we discuss how changing the function class of the prediction model f_{base} or the post-hoc module T_{PP} impacts the performance of FRAPPÉ methods.

Figures 10, 11, and 12 show how FRAPPÉ MinDiff compares to several prior approaches when the base model is a random forest, a gradient-boosted machine or logistic regression, respectively. Notably, the computation cost for training FRAPPÉ MinDiff is roughly unchanged for all three base model classes.

Furthermore, in each of these figures, we consider two different function classes for the post-processing transformation T_{PP}: a linear model and a 1-hidden layer multi-layer perceptron. These experiments confirm the intuition that a more expressive post-hoc transformation for FRAPPÉ can lead to better Pareto frontiers.

Finally, in Figure 20 we highlight how the complexity of the prediction model produced by an in-processing technique affects the Pareto frontier relative to its FRAPPÉ counterpart. We assume the same prediction tasks for Adult and COMPAS as in the rest of the paper. More specifically, the left and right panels use the same pre-processing steps as Figures 4b and 8b, respectively. For this experiment, we consider two different fairness definitions (EqOpp and EqOdds) and two different classes of base models for FRAPPÉ (3-layer multi-layer perceptron, i.e. 3-MLP, and a gradient-boosted machine, i.e. GBM). We compare FRAPPÉ MinDiff to two different in-processing methods, MinDiff (Beutel et al., 2019; Prost et al., 2019) and Reductions (Agarwal et al., 2018). We choose the optimal hyperparameters (e.g. learning rate) using a held-out validation set and the same methodology as in the rest of the experiments. These experiments reveal that FRAPPÉ methods match the performance of in-processing with a complex function class $\mathcal{F}$, which in turn outperforms in-processing with a simpler function class $\mathcal{F} =$ linear models.



(a) Inducing EqOpp on Adult.

(b) Inducing EqOdds on COMPAS.

Figure 20: FRAPPÉ (in blue) matches the performance of in-processing with the same base model complexity (i.e. in orange) and outperforms in-processing that uses less complex linear models (in green). For FRAPPÉ, the post-hoc module is always linear, while the base model is a 3-layer MLP (left) or a GBM (right).