arXiv:1905.10607v1 [cs.LG] 25 May 2019

Average Individual Fairness: Algorithms, Generalization and Experiments

Michael Kearns[1], Aaron Roth[1], and Saeed Sharifi-Malvajerdi[1]

[1] *University of Pennsylvania*

May 28, 2019

Abstract

We propose a new family of fairness definitions for classification problems that combine some of the best properties of both statistical and individual notions of fairness. We posit not only a distribution over *individuals*, but also a distribution over (or collection of) *classification tasks*. We then ask that standard statistics (such as error or false positive/negative rates) be (approximately) equalized across *individuals*, where the rate is defined as an expectation over the classification tasks. Because we are no longer averaging over coarse groups (such as race or gender), this is a semantically meaningful individual-level constraint. Given a sample of individuals and classification problems, we design an oracle-efficient algorithm (i.e. one that is given access to any standard, fairness-free learning heuristic) for the fair empirical risk minimization task. We also show that given sufficiently many samples, the ERM solution generalizes in two directions: both to new individuals, and to new classification tasks, drawn from their corresponding distributions. Finally we implement our algorithm and empirically verify its effectiveness.

Contents

1 Introduction **3**
1.1 Our Results . . . 4
1.2 Additional Related Work . . . 5

2 Model and Preliminaries **5**
2.1 AIF: Average Individual Fairness . . . 6
2.2 Notations . . . 7

3 Learning subject to AIF **7**
3.1 The Empirical Fair Learning Problem . . . 8
3.2 A Reductions Approach: Formulation as a Two-player Game . . . 8
3.3 **BEST**: The Learner's Best Response . . . 9
3.4 **AIF-Learn**: Implementation and In-sample Guarantees . . . 11
3.5 Generalization Theorems . . . 14

4 Experimental Evaluation **16**

Acknowledgements **17**

References **17**

Appendix **19**

A Learning subject to False Positive AIF (FPAIF) **19**
A.1 **aBEST**$_{\text{FP}}$: The Learner's approximate Best Response . . . 21
A.2 **FPAIF-Learn**: Implementation and In-sample Guarantees . . . 23
A.3 Generalization Theorems . . . 26

B Proofs of Section 3 **27**
B.1 Preliminary Tools . . . 27
B.2 Algorithm Analysis and In-sample Guarantees . . . 28
B.3 Proof of Theorem 3.4: Generalization over $\mathcal{P}$ . . . 31
B.4 Proof of Theorem 3.5: Generalization over $\mathcal{Q}$ . . . 37
B.5 Proof of Theorem 3.6: Simultaneous Generalization over $\mathcal{P}$ and $\mathcal{Q}$ . . . 39

C Proofs of Section A **42**
C.1 Regret Bounds . . . 42

D Additional Experimental Results **45**

1 Introduction

The community studying fairness in machine learning has yet to settle on definitions. At a high level, existing definitional proposals can be divided into two groups: *statistical* fairness definitions and *individual* fairness definitions. Statistical fairness definitions partition individuals into "protected groups" (often based on race, gender, or some other binary protected attribute) and ask that some statistic of a classifier (error rate, false positive rate, positive classification rate, etc.) be approximately equalized across those groups. In contrast, individual definitions of fairness have no notion of "protected groups", and instead ask for constraints that bind on pairs of individuals. These constraints can have the semantics that "similar individuals should be treated similarly" ([Dwork et al., 2012]), or that "less qualified individuals should not be preferentially favored over more qualified individuals" ([Joseph et al., 2016]). Both families of definitions have serious problems, which we will elaborate on. But in summary, statistical definitions of fairness provide only very weak promises to individuals, and so do not have very strong semantics. Existing proposals for individual fairness guarantees, on the other hand, have very strong semantics, but have major obstacles to deployment, requiring strong assumptions on either the data generating process or on society's ability to instantiate an agreed-upon fairness metric.

Statistical definitions of fairness are the most popular in the literature, in large part because they can be easily checked and enforced on arbitrary data distributions. For example, a popular definition ([Hardt et al., 2016, Kleinberg et al., 2017, Chouldechova, 2017]) asks that a classifier's *false positive rate* should be equalized across the protected groups. This can sound attractive: in settings in which a positive classification leads to a bad outcome (e.g. incarceration), it is the *false positives* that are harmed by the errors of the classifier, and asking that the false positive rate be equalized across groups is asking that the harm caused by the algorithm should be proportionately spread across protected populations. But the meaning of this guarantee to an individual is limited, because the word *rate* refers to an average over the population. To see why this limits the meaning of the guarantee, consider the example given in [Kearns et al., 2018]: imagine a society that is equally split between gender (Male, Female) and race (Blue, Green). Under the constraint that false positive rates be equalized across both race and gender, a classifier may incarcerate 100% of blue men and green women, and 0% of green men and blue women. This equalizes the false positive rate across all protected groups, but is cold comfort to any individual blue man and green woman. This effect isn't merely hypothetical — [Kearns et al., 2018, Kearns et al., 2019] showed similar effects when using off-the-shelf fairness constrained learning techniques on real datasets.

Individual definitions of fairness, on the other hand, can have strong individual level semantics. For example, the constraint imposed by [Joseph et al., 2016, Joseph et al., 2018] in online classification problems implies that the false positive rate must be equalized across all pairs of individuals who (truly) have negative labels. Here the word *rate* has been redefined to refer to an expectation over the randomness of the classifier, and there is no notion of protected groups. This kind of constraint provides a strong individual level promise that one's risk of being harmed by the errors of the classifier are no higher than they are for anyone else. Unfortunately, in order to non-trivially satisfy a constraint like this, it is necessary to make strong realizability assumptions.

1.1 Our Results

We propose an alternative definition of individual fairness that avoids the need to make assumptions on the data generating process, while giving the learning algorithm more flexibility to satisfy it in non-trivial ways. We consider that in many applications each individual will be subject to decisions made by *many classification tasks* over a given period of time, not just one. For example, internet users are shown a large number of targeted ads over the course of their usage of a platform, not just one: the properties of the advertisers operating in the platform over a period of time are not known up front, but have some statistical regularities. Public school admissions in cities like New York are handled by a centralized match: students apply not just to one school, but to many, who can each make their own admissions decisions ([Abdulkadiroğlu et al., 2005]). We model this by imagining that not only is there an unknown distribution $\mathcal{P}$ over individuals, but there is an unknown distribution $\mathcal{Q}$ over classification problems (each of which is represented by an unknown mapping from individual features to target labels). With this model in hand, we can now ask that the error rates (or false positive or negative rates) be equalized across all individuals — where now *rate* is defined as the average over *classification tasks* drawn from $\mathcal{Q}$ of the probability of a particular individual being incorrectly classified.

We then derive a new oracle-efficient algorithm for satisfying this guarantee in-sample, and prove novel generalization guarantees showing that the guarantees of our algorithm hold also out of sample. Oracle efficiency is an attractive framework in which to circumvent the worst-case hardness of even *unconstrained* learning problems, and focus on the *additional computational difficulty* imposed by fairness constraints. It assumes the existence of "oracles" (in practice, implemented with a heuristic) that can solve weighted classification problems absent fairness constraints, and asks for efficient reductions from the fairness constrained learning problems to unconstrained problems. This has become a popular technique in the fair machine learning literature (see e.g. [Agarwal et al., 2018, Kearns et al., 2018]) — and one that often leads to practical algorithms. The generalization guarantees we prove require the development of new techniques because they refer to generalization in *two orthogonal directions* — over both individuals and classification problems. Our algorithm is run on a sample of n individuals sampled from $\mathcal{P}$ and m problems sampled from $\mathcal{Q}$. It is given access to an oracle (in practice, implemented with a heuristic) for solving ordinary cost sensitive classification problems over some hypothesis space $\mathcal{H}$. The algorithm runs in polynomial time (it performs only elementary calculations except for calls to the learning oracle, and makes only a polynomial number of calls to the oracle) and returns a *mapping from problems to hypotheses* that have the following properties, so long as n and m are sufficiently large (polynomial in the VC-dimension of $\mathcal{H}$ and the desired error parameters): For any α, with high probability over the draw of the n individuals from $\mathcal{P}$ and the m problems from $\mathcal{Q}$

1. *Accuracy*: the error rate (computed in expectation over new individuals $x \sim \mathcal{P}$ and new problems $f \sim \mathcal{Q}$) is within $O(\alpha)$ of the *optimal* mapping from problems to classifiers in $\mathcal{H}$, subject to the constraint that for every pair of individuals x, x' in the support of $\mathcal{P}$, the error rates (or false positive or negative rates) (computed in expectation over problems $f \sim Q$) on x and x' differ by at most α.

2. *Fairness*: with probability $1 - \beta$ over the draw of new individuals $x, x' \sim \mathcal{P}$, the error rate (or false positive or negatives rates) of the output mapping (computed in expectation over problems $f \sim Q$) on x will be within $O(\alpha)$ of that of x'.

The mapping from new classification problems to hypotheses that we find is derived from the dual variables of the linear program representing our empirical risk minimization task, and we crucially rely on the structure of this mapping to prove our generalization guarantees for new problems $f \sim Q$.

1.2 Additional Related Work

The literature on fairness in machine learning has become much too large to comprehensively summarize, but see [Mitchell et al., 2018] for a recent survey. Here we focus on the most conceptually related work, which has aimed to bridge the gap between the immediate applicability of statistical definitions of fairness with the strong individual level semantics of individual notions of fairness. One strand of this literature focuses on the "metric fairness" definition first proposed by [Dwork et al., 2012], and aims to ease the assumption that the learning algorithm has access to a task specific fairness metric. [Kim et al., 2018a] imagine access to an oracle which can provide unbiased estimates to the metric distance between any pair of individuals, and show how to use this to satisfy a statistical notion of fairness representing "average metric fairness" over pre-defined groups. [Gillen et al., 2018] study a contextual bandit learning setting in which a human judge points out metric fairness violations whenever they occur, and show that with this kind of feedback (under assumptions about consistency with a family of metrics), it is possible to quickly converge to the optimal fair policy. [Yona and Rothblum, 2018] consider a PAC-based relaxation of metric fair learning, and show that empirical metric-fairness generalizes to out-of-sample metric fairness. Another strand of this literature has focused on mitigating the problems that arise when statistical notions of fairness are imposed over coarsely defined groups, by instead asking for statistical notions of fairness over exponentially many or infinitely many groups with a well defined structure. This line includes [Hébert-Johnson et al., 2018] (focusing on calibration), [Kearns et al., 2018] (focusing on false positive and negative rates), and [Kim et al., 2018b] (focusing on error rates).

2 Model and Preliminaries

We model each individual in our framework by a vector of features $x \in \mathcal{X}$, and we let each learning problem[1] be represented by a binary function $f \in \mathcal{F}$ mapping $\mathcal{X}$ to $\{0, 1\}$. We assume probability measures $\mathcal{P}$ and $\mathcal{Q}$ over the space of individuals $\mathcal{X}$ and the space of problems $\mathcal{F}$, respectively. Without loss of generality, we assume throughout that the support of the distributions $\mathcal{P}$ and $\mathcal{Q}$ are $\mathcal{X}$ and $\mathcal{F}$, respectively. In the training phase there is a *fixed* (across problems) set $X = \{x_i\}_{i=1}^n$ of n individuals sampled independently from $\mathcal{P}$ for which we have available labels of m learning tasks represented by $F = \{f_j\}_{j=1}^m$ drawn independently from $\mathcal{Q}$[2]. Therefore, a training data set of n individuals X for m learning tasks F takes the form: $S = \left\{x_i, (f_j(x_i))_{j=1}^m\right\}_{i=1}^n$. We summarize the notations we use for individuals and problems in Table 1.

In general the function class $\mathcal{F}$ will be unknown. We will aim to solve the (agnostic) learning problem over a hypothesis class $\mathcal{H}$, which need bear no relationship to $\mathcal{F}$. We assume throughout that $\mathcal{H}$ contains the constant classifiers h^0 and h^1 where $h^0(x) = 0$ and $h^1(x) = 1$ for all x. This

[1] we will use the terms: problem, task, and labeling interchangeably.

[2] Throughout we will use subscript i to denote individuals and j to denote learning problems.

Table 1: Summary of notations for individuals vs. problems.

	space	element	distribution	data set	sample size	empirical dist.
individual	$\mathcal{X}$	$x \in \mathcal{X}$	$\mathcal{P}$	$X = \{x_i\}_{i=1}^n$	n	$\widehat{\mathcal{P}} = \mathcal{U}(X)$
problem	$\mathcal{F}$	$f \in \mathcal{F}$	$\mathcal{Q}$	$F = \{f_j\}_{j=1}^m$	m	$\widehat{\mathcal{Q}} = \mathcal{U}(F)$

assumption will allow us to argue about feasibility of the constrained optimization problems that we will solve. We allow for randomized classifiers, which we model as learning over $\Delta(\mathcal{H})$, the probability simplex over $\mathcal{H}$.

Unlike usual learning settings where the primary goal is to learn a single hypothesis $p \in \Delta(\mathcal{H})$, our objective is to learn a *mapping* $\psi \in \Delta(\mathcal{H})^{\mathcal{F}}$ that maps learning tasks $f \in \mathcal{F}$ represented as labelings of the training data to hypotheses $p \in \Delta(\mathcal{H})$. We will therefore have to formally define the error rates incurred by a mapping ψ and use them to formalize a learning task subject to our proposed fairness notion. For a mapping ψ, we write ψ_f to denote the classifier corresponding to $f \in \mathcal{F}$ under the mapping, i.e., $\psi_f = \psi(f) \in \Delta(\mathcal{H})$.

Notice in the training phase, there are only m learning problems $F = \{f_j\}_{j=1}^m$ to be solved, and therefore, the corresponding empirical fair learning problem reduces to learning m randomized classifiers $\boldsymbol{p} = (p_1, p_2, \ldots, p_m) \in \Delta(\mathcal{H})^m$, where p_j is the learned classifier for the jth problem $f_j \in F$. In general, learning m specific classifiers for the training problems might not give any generalizable rule mapping new problems to hypotheses — but the specific algorithm we propose will, in the form of a set of weights (derived from the dual variables of the ERM problem) over the training individuals.

2.1 AIF: Average Individual Fairness

Definition 2.1 (Individual and Overall Error Rates of a Mapping ψ)**.** *For a given individual $x \in \mathcal{X}$, a mapping $\psi \in \Delta(\mathcal{H})^{\mathcal{F}}$, and distributions $\mathcal{P}$ and $\mathcal{Q}$ over $\mathcal{X}$ and $\mathcal{F}$, the individual error rate of x incurred by ψ is defined as follows:*

$$\mathcal{E}(x, \psi; \mathcal{Q}) = \mathop{\mathbb{E}}_{f \sim \mathcal{Q}} \left[\mathop{\mathbb{P}}_{h \sim \psi_f} [h(x) \neq f(x)] \right]$$

The overall error rate of ψ is:

$$err(\psi; \mathcal{P}, \mathcal{Q}) = \mathop{\mathbb{E}}_{x \sim \mathcal{P}} [\mathcal{E}(x, \psi; \mathcal{Q})]$$

In the body of this paper, we will focus on a fairness constraint that asks that the individual error rate should be approximately equalized across all individuals. In Section A of the Appendix, we extend our techniques to equalizing false positive and negative rates across individuals.

Definition 2.2 (Average Individual Fairness (AIF))**.** *In our framework, we say a mapping $\psi \in \Delta(\mathcal{H})^{\mathcal{F}}$ satisfies "(α, β)-AIF" (reads (α, β)-approximate Average Individual Fairness) with respect to the distributions $(\mathcal{P}, \mathcal{Q})$ if there exists $\gamma \geqslant 0$ such that:*

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}} (|\mathcal{E}(x, \psi; \mathcal{Q}) - \gamma| > \alpha) \leqslant \beta$$

2.2 Notations

We briefly fix some notations:

- For an event A, $\mathbb{1}[A]$ represents the indicator function of A. $\mathbb{1}[A] = 1$ if A occurs.
- For a natural number $n \in \mathbb{N}$, $[n] = \{1, 2, \ldots, n\}$.
- $\mathcal{U}(S)$ represents the uniform distribution over the set S.
- For a mapping $\psi : A \to B$ and $A' \subseteq A$, $\psi|_{A'}$ represents ψ restricted to the domain A'.
- For a hypothesis class $\mathcal{H}$, $\Delta(\mathcal{H})$ represents the probability simplex over $\mathcal{H}$.
- $d_{\mathcal{H}}$ denotes the VC dimension of the hypothesis class $\mathcal{H}$.
- $CSC(\mathcal{H})$ denotes a *cost sensitive classification oracle* for $\mathcal{H}$ which is defined below.

Definition 2.3 (Cost Sensitive Classification (CSC) in $\mathcal{H}$)**.** *Let $D = \{x_i, c_i^1, c_i^0\}_{i=1}^n$ denote a data set of n individuals x_i where c_i^1 and c_i^0 are the costs of classifying x_i as positive (1) and negative (0) respectively. Given D, the cost sensitive classification problem defined over $\mathcal{H}$ is the optimization problem:*

$$\arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} \left\{ c_i^1 h(x_i) + c_i^0 \left(1 - h(x_i)\right) \right\}$$

An oracle $CSC(\mathcal{H})$ takes the data set $D = \{x_i, c_i^1, c_i^0\}_{i=1}^n$ as input and outputs the solution to the optimization problem. We use $CSC(\mathcal{H}; D)$ to denote the classifier returned by $CSC(\mathcal{H})$ on an input data set D. We say that an algorithm is oracle efficient if it runs in polynomial time given the ability to make unit-time calls to $CSC(\mathcal{H})$.

3 Learning subject to AIF

In this section we first cast the learning problem subject to the AIF fairness constraints as the constrained optimization problem (1) and then develop an oracle efficient algorithm for solving its corresponding empirical risk minimization (ERM) problem (in the spirit of [Agarwal et al., 2018]). In the coming sections we give a full analysis of the developed algorithm including its *in-sample* accuracy/fairness guarantees, and define the mapping it induces from new problems to hypotheses, and finally establish *out-of-sample* bounds for this trained mapping.

Fair Learning Problem subject to $(\alpha, 0)$-AIF

$$\begin{aligned} \min_{\psi \in \Delta(\mathcal{H})^{\mathcal{F}}, \gamma \in [0,1]} \quad & \operatorname{err}(\psi; \mathcal{P}, \mathcal{Q}) \\ \text{s.t. } \forall x \in \mathcal{X}: \quad & |\mathcal{E}(x, \psi; \mathcal{Q}) - \gamma| \leqslant \alpha \end{aligned} \tag{1}$$

Definition 3.1 (OPT)**.** *Consider the optimization problem (1). Given distributions $\mathcal{P}$ and $\mathcal{Q}$, and fairness approximation parameter α, we denote the optimal solutions of (1) by $\psi^\star(\alpha; \mathcal{P}, \mathcal{Q})$ and $\gamma^\star(\alpha; \mathcal{P}, \mathcal{Q})$, and the value of the objective function at these optimal points by $OPT(\alpha; \mathcal{P}, \mathcal{Q})$. In other words*

$$OPT(\alpha; \mathcal{P}, \mathcal{Q}) = err(\psi^\star; \mathcal{P}, \mathcal{Q})$$

We will use OPT as the benchmark with respect to which we evaluate the accuracy of our trained mapping. It is worth noticing that the optimization problem (1) has a nonempty set of feasible points for any α and any distributions $\mathcal{P}$ and $\mathcal{Q}$ because the following point is always feasible: $\gamma = 0.5$ and $\psi_f = 0.5h^0 + 0.5h^1$ (i.e. random classification) for all $f \in \mathcal{F}$ where h^0 and h^1 are all-zero and all-one constant classifiers.

3.1 The Empirical Fair Learning Problem

We start to develop our algorithm by defining the *empirical* version of (1) for a given training data set of n individuals $X = \{x_i\}_{i=1}^n$ and m learning problems $F = \{f_j\}_{j=1}^m$. We will formulate the empirical problem as finding a *restricted* mapping $\psi|_F$ by which we mean the domain of the mapping is restricted to the training set $F \subseteq \mathcal{F}$. We will later see how the dynamics of our proposed algorithm allows us to extend the restricted mapping to a mapping from the entire space $\mathcal{F}$. We slightly change notation and represent a restricted mapping $\psi|_F \in \Delta(\mathcal{H})^F$ explicitly by a vector $\boldsymbol{p} = (p_1, \ldots, p_m) \in \Delta(\mathcal{H})^m$ of randomized classifiers where $p_j \in \Delta(\mathcal{H})$ corresponds to $f_j \in F$. Notice the empirical versions of the overall error rate and the individual error rates incurred by the mapping $\boldsymbol{p}$ (see Definition 2.1) can be expressed as:

$$\operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{P}}_{h_j \sim p_j}\left[h_j(x_i) \neq f_j(x_i)\right] \tag{2}$$

$$\mathcal{E}\left(x, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) = \frac{1}{m}\sum_{j=1}^{m} \mathop{\mathbb{P}}_{h_j \sim p_j}\left[h_j(x) \neq f_j(x)\right] \tag{3}$$

Using these empirical quantities, we cast the empirical version of the fair learning problem (1) as the constrained optimization problem (4) where there is one constraint for each individual in the training data set. Note that the optimization problem (4) forms a linear program and that we considered a slightly relaxed version of (1) where instead of $(\alpha, 0)$-AIF, we require $(2\alpha, 0)$-AIF (of course now with respect to the empirical distributions) only to make sure the optimal solution $(\psi^\star, \gamma^\star)$ of (1) (in fact $\psi^\star$ restricted to F) is feasible in (4) as long as enough samples are acquired. This will appear later in the generalization analysis of our proposed algorithm.

Empirical Fair Learning Problem

$$\begin{aligned} \min_{\boldsymbol{p} \in \Delta(\mathcal{H})^m,\, \gamma \in [0,1]} \quad & \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \\ \text{s.t. } \forall i \in \{1, \ldots, n\}: \quad & \left|\mathcal{E}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - \gamma\right| \leqslant 2\alpha \end{aligned} \tag{4}$$

3.2 A Reductions Approach: Formulation as a Two-player Game

We use the dual perspective of constrained optimization problems to reduce the fair ERM (4) to a two-player game between a "Learner" (primal player) and an "Auditor" (dual player). Towards

deriving the Lagrangian of (4), we first rewrite its constraints in $\boldsymbol{r}(\boldsymbol{p}, \gamma; \widehat{\mathcal{Q}}) \leqslant 0$ form where

$$\boldsymbol{r}\left(\boldsymbol{p}, \gamma; \widehat{\mathcal{Q}}\right) = \begin{bmatrix} \mathcal{E}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - \gamma - 2\alpha \\ \gamma - \mathcal{E}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - 2\alpha \end{bmatrix}_{i=1}^{n} \in \mathbb{R}^{2n} \tag{5}$$

represents the "fairness violations" of the pair $(\boldsymbol{p}, \gamma)$ in one single vector. Let the corresponding dual variables be represented by $\boldsymbol{\lambda} = \left[\boldsymbol{\lambda}_i^+, \boldsymbol{\lambda}_i^-\right]_{i=1}^n \in \Lambda$, where $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}_+^{2n} \mid ||\boldsymbol{\lambda}||_1 \leqslant B\}$. Note we place an upper bound B on the ℓ_1-norm of $\boldsymbol{\lambda}$ in order to reason about the convergence of our proposed algorithm. B will eventually factor into both the run-time and the approximation guarantees of our solution. Using Equation (5) and the introduced dual variables, we have that the Lagrangian of (4) is

$$\mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}\right) = \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + \boldsymbol{\lambda}^T \boldsymbol{r}\left(\boldsymbol{p}, \gamma; \widehat{\mathcal{Q}}\right) \tag{6}$$

We therefore consider solving the following minmax problem:

$$\min_{\boldsymbol{p} \in \Delta(\mathcal{H})^m, \gamma \in [0,1]} \max_{\lambda \in \Lambda} \mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}\right) = \max_{\lambda \in \Lambda} \min_{\boldsymbol{p} \in \Delta(\mathcal{H})^m, \gamma \in [0,1]} \mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}\right) \tag{7}$$

where strong duality holds because $\mathcal{L}$ is linear in its arguments and the domains of $(\boldsymbol{p}, \gamma)$ and $\boldsymbol{\lambda}$ are convex and compact ([Sion, 1958]). From a game theoretic perspective, the solution to this minmax problem can be seen as an equilibrium of a zero-sum game between two players. The primal player (Learner) has strategy space $\Delta(\mathcal{H})^m \times [0, 1]$ while the dual player (Auditor) has strategy space Λ, and given a pair of chosen strategies $(\boldsymbol{p}, \gamma, \boldsymbol{\lambda})$, the Lagrangian $\mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}\right)$ represents how much the Learner has to pay to the Auditor — i.e. it defines the payoff function of a zero-sum game in which the Learner is the minimization player, and the Auditor is the maximization player.

Using no regret dynamics, an approximate equilibrium of this zero-sum game can be found in an iterative framework. In each iteration, we let the dual player run the *exponentiated gradient descent* algorithm and the primal player *best respond*. The best response problem of the Learner can be decoupled into $(m+1)$ separate minimization problems and that in particular, the optimal classifiers $\boldsymbol{p}$ can be viewed as the solutions to m *weighted classification problems* in $\mathcal{H}$ where all m problems share the same weights $\boldsymbol{w} = [\lambda_i^+ - \lambda_i^-]_i \in \mathbb{R}^n$ over the training individuals. In the following subsection, we derive and implement the best response of the Learner where we use the learning oracle $CSC(\mathcal{H})$ (see Definition 2.3) to solve the weighted classification problems.

3.3 BEST: The Learner's Best Response

We formally describe and analyze the best response problem of the Learner in this subsection and summarize the results in a subroutine called **BEST**. In each iteration of the described iterative framework, the Learner is given some $\boldsymbol{\lambda} \in \Lambda$ picked by the Auditor and she wants to solve the following minimization problem.

$$\operatorname*{arg\,min}_{\boldsymbol{p} \in \Delta(\mathcal{H})^m, \gamma \in [0,1]} \mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}\right)$$

We will use Equations (2) and (3) to expand the Lagrangian (6) and decouple the above minimization problem into $(m + 1)$ minimization problems, each depends only on one variable the Learner has to pick.

$$
\begin{aligned}
&\underset{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]}{\arg\min} \quad \mathcal{L}\left(\boldsymbol{p},\gamma,\boldsymbol{\lambda}\right)\\
\equiv &\underset{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]}{\arg\min} \quad \mathrm{err}\left(\boldsymbol{p};\widehat{\mathcal{P}},\widehat{\mathcal{Q}}\right)+\sum_{i=1}^{n}\left\{\lambda_i^+\left(\mathcal{E}\left(x_i,\boldsymbol{p};\widehat{\mathcal{Q}}\right)-\gamma\right)+\lambda_i^-\left(\gamma-\mathcal{E}\left(x_i,\boldsymbol{p};\widehat{\mathcal{Q}}\right)\right)\right\}\\
\equiv &\underset{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]}{\arg\min} \quad \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}\left(x_i,\boldsymbol{p};\widehat{\mathcal{Q}}\right)+\sum_{i=1}^{n}\left\{\lambda_i^+\left(\mathcal{E}\left(x_i,\boldsymbol{p};\widehat{\mathcal{Q}}\right)-\gamma\right)+\lambda_i^-\left(\gamma-\mathcal{E}\left(x_i;\boldsymbol{p};\widehat{\mathcal{Q}}\right)\right)\right\}\\
\equiv &\underset{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]}{\arg\min} \quad \sum_{i=1}^{n}\left\{\lambda_i^- - \lambda_i^+\right\}\gamma+\sum_{i=1}^{n}\left(\frac{1}{n}+\lambda_i^+-\lambda_i^-\right)\mathcal{E}\left(x_i,\boldsymbol{p};\widehat{\mathcal{Q}}\right)\\
\equiv &\underset{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]}{\arg\min} \quad \sum_{i=1}^{n}\left\{\lambda_i^- - \lambda_i^+\right\}\gamma+\frac{1}{m}\sum_{j=1}^{m}\left\{\sum_{i=1}^{n}\left(\frac{1}{n}+\lambda_i^+-\lambda_i^-\right)\underset{h_j\sim p_j}{\mathbb{P}}\left[h_j(x_i)\neq f_j(x_i)\right]\right\}
\end{aligned}
$$

Therefore, the minimization problem of the Learner gets nicely decoupled into $(m+1)$ minimization problems. Let $w_i = \lambda_i^+ - \lambda_i^-$ for all i, and acccordingly, let $\boldsymbol{w} = [w_1,\ldots,w_n]^\top \in \mathbb{R}^n$. First, the optimal value for γ is chosen according to

$$
\gamma = \mathbb{1}\left[\sum_{i=1}^{n} w_i > 0\right] \tag{8}
$$

And for learning problem j, the following minimization problem must be solved.

$$
\underset{p_j\in\Delta(\mathcal{H})}{\arg\min}\sum_{i=1}^{n}(1/n+w_i)\underset{h_j\sim p_j}{\mathbb{P}}\left[h_j(x_i)\neq f_j(x_i)\right] \equiv \underset{h_j\in\mathcal{H}}{\arg\min}\sum_{i=1}^{n}(1/n+w_i)\,\mathbb{1}\left[h_j(x_i)\neq f_j(x_i)\right]
$$

where the equivalence holds since the Learner can choose to put all the probability mass on a single classifier. This minimization problem represents exactly a weighted classification problem. Since we work with cost sensitive classification oracles in this paper, we further reduce the weighted classification problem to a CSC problem that can be solved by a call to the CSC oracle for $\mathcal{H}$ ($CSC(\mathcal{H})$). For problem $j\in[m]$, let

$$
c_{i,j}^1 = (w_i+1/n)(1-f_j(x_i)) \quad , \quad c_{i,j}^0 = (w_i+1/n)f_j(x_i)
$$

represent the costs associated with individual $i\in[n]$. Observe that the above weighted classification problem can be now casted as the following CSC problem.

$$
h_j = \underset{h\in\mathcal{H}}{\arg\min}\sum_{i=1}^{n} c_{i,j}^1\,h(x_i)+c_{i,j}^0\,(1-h(x_i)) \tag{9}
$$

To sum up, in each iteration of the algorithm the Auditor first uses the exponentiated gradient descent algorithm to update the dual variable $\boldsymbol{\lambda}$ (or correspondingly, the vector of weights $\boldsymbol{w}$ over the training individuals) and then the Learner picks $\gamma = \mathbb{1}\left[\sum_{i=1}^{n} w_i > 0\right]$ and solves m cost sensitive classification problems casted in (9) by calling the oracle $CSC(\mathcal{H})$ for all $1\leqslant j\leqslant m$. We have the best response of the Learner written in Subroutine 1. This subroutine will be called in each iteration of the final AIF learning algorithm.

Subroutine 1: BEST– best response of the Learner in the AIF setting

Input: dual weights $\boldsymbol{w} \in \mathbb{R}^n$, training examples $S = \left\{x_i, (f_j(x_i))_{j=1}^m\right\}_{i=1}^n$

$\gamma \leftarrow \mathbb{1}\left[\sum_{i=1}^n w_i > 0\right]$
for $j = 1, \ldots, m$ **do**
  $c_i^1 \leftarrow (w_i + 1/n)(1 - f_j(x_i))$ for $i \in [n]$.
  $c_i^0 \leftarrow (w_i + 1/n) f_j(x_i)$ for $i \in [n]$.
  $D \leftarrow \{x_i, c_i^1, c_i^0\}_{i=1}^n$
  $h_j \leftarrow CSC\left(\mathcal{H}; D\right)$
end
$\boldsymbol{h} \leftarrow (h_1, h_2, \ldots, h_m)$

Output: $\left(\boldsymbol{h}, \gamma\right)$

3.4 AIF-Learn: Implementation and In-sample Guarantees

In Algorithm 2 (**AIF-Learn**), with a slight deviation from what we described in the previous subsections, we implement the proposed algorithm. The deviation arises when the Auditor updates the dual variables $\boldsymbol{\lambda}$ in each round, and is introduced in the service of arguing for generalization. To counteract the inherent adaptivity of the algorithm (which makes the quantities estimated at each round data dependent), at each round t of the algorithm, we draw a *fresh* batch of m_0 problems to estimate the fairness violation vector $\boldsymbol{r}$ (5). From another viewpoint – which is the way the algorithm is actually implemented – similar to usual batch learning models we assume we have a training set F of m learning problems upfront. However, in our proposed algorithm that runs for T iterations, we partition F into T equally-sized (m_0) subsets $\{F_t\}_{t=1}^T$ uniformly at random and use only the batch of problems F_t at round t to update the dual variables $\boldsymbol{\lambda}$. Without loss of generality and to avoid technical complications, we assume $|F_t| = m_0 = m/T$ is a natural number. This is represented in Algorithm 2 by writing $\widehat{\mathcal{Q}}_t = \mathcal{U}\left(F_t\right)$ for the uniform distribution over the batch of problems F_t , and $\boldsymbol{h}_t|_{F_t}$ for the associated learned classifiers for F_t. We will see this modification will only introduce an extra $\widetilde{O}(\sqrt{1/m_0})$ term to the regret of the Auditor and thus we have to assume m_0 is sufficiently large (Assumption 3.1) so that the Auditor has in fact low regret.

Notice **AIF-Learn** takes as input an approximation parameter $\nu \in [0, 1]$ which will quantify how close the output of the algorithm is to an equilibrium of the introduced game, and it will accordingly propagate to the accuracy bounds. One important aspect of **AIF-Learn** is that the algorithm maintains a vector of weights $\boldsymbol{w}_t \in \mathbb{R}^n$ over the training individuals X and that each $\widehat{p}_j$ learned by our algorithm is in fact an average over T classifiers where classifier t is the solution to a CSC problem on X weighted by $\boldsymbol{w}_t$. As a consequence, we propose to extend the learned restricted mapping $\widehat{\boldsymbol{p}} \in \Delta(\mathcal{H})^m$ to a mapping $\widehat{\psi} = \widehat{\psi}\left(X, \widehat{W}\right) \in \Delta(\mathcal{H})^{\mathcal{F}}$ that takes *any* problem $f \in \mathcal{F}$ as input (represented to $\widehat{\psi}$ by the labels it induces on the training individuals), uses X along with the set of weights $\widehat{W} = \{\boldsymbol{w}_t\}_{t=1}^T$ to solve T CSC problems in a similar fashion, and outputs the average of the learned classifiers denoted by $\widehat{\psi}_f \in \Delta(\mathcal{H})$. This extension is consistent with $\widehat{\boldsymbol{p}}$ in the sense that $\widehat{\psi}$ restricted to F will be exactly the $\widehat{\boldsymbol{p}}$ output by our algorithm. We have the pseudocode for $\widehat{\psi}$ written in detail in Mapping 3 and we let **AIF-Learn** output $\widehat{\psi}$.

Algorithm 2: AIF-Learn – learning subject to AIF

Input: fairness parameter α, approximation parameter ν,
training data set $X = \{x_i\}_{i=1}^n$ and $F = \{f_j\}_{j=1}^m$

$$B \leftarrow \frac{1+2\nu}{\alpha}, \quad T \leftarrow \frac{16B^2(1+2\alpha)^2\log(2n+1)}{\nu^2}, \quad \eta \leftarrow \frac{\nu}{4(1+2\alpha)^2 B}, \quad m_0 \leftarrow \frac{m}{T}, \quad S \leftarrow \left\{x_i, (f_j(x_i))_{j=1}^m\right\}_{i=1}^n$$

Partition F uniformly at random: $F = \{F_t\}_{t=1}^T$ where $|F_t| = m_0$.
$\boldsymbol{\theta}_1 \leftarrow \mathbf{0} \in \mathbb{R}^{2n}$
for $t = 1, \ldots, T$ **do**
  $\lambda_{i,t}^{\bullet} \leftarrow B\frac{\exp(\theta_{i,t}^{\bullet})}{1+\sum_{i',\bullet'}\exp(\theta_{i',t}^{\bullet'})}$ for $i \in [n]$ and $\bullet \in \{+,-\}$
  $\boldsymbol{w}_t \leftarrow [\lambda_{i,t}^{+} - \lambda_{i,t}^{-}]_{i=1}^n \in \mathbb{R}^n$
  $(\boldsymbol{h}_t, \gamma_t) \leftarrow \mathbf{BEST}(\boldsymbol{w}_t; S)$
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \eta \cdot \boldsymbol{r}\left(\boldsymbol{h}_t|_{F_t}, \gamma_t; \widehat{\mathcal{Q}}_t\right)$
end
$\widehat{\boldsymbol{p}} \leftarrow \frac{1}{T}\sum_{t=1}^T \boldsymbol{h}_t$, $\widehat{\gamma} \leftarrow \frac{1}{T}\sum_{t=1}^T \gamma_t$, $\widehat{\boldsymbol{\lambda}} \leftarrow \frac{1}{T}\sum_{t=1}^T \boldsymbol{\lambda}_t$, $\widehat{W} \leftarrow \{\boldsymbol{w}_t\}_{t=1}^T$

Output: average plays $\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right)$, mapping $\widehat{\psi} = \widehat{\psi}\left(X, \widehat{W}\right)$ (see Mapping 3)

We start the analysis of Algorithm 2 by establishing the regret bound of the Auditor over T rounds of the algorithm. The regret bound will help us pick the number of iterations T and the learning rate η so that the Auditor has sufficiently small regret (bounded by ν). Notice the Learner uses her best response in each round of the algorithm which implies that she has zero regret. Since in this subsection we eventually want to state in-sample guarantees (i.e., guarantees with respect to the distributions $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{Q}}$), we work with the restricted mapping $\widehat{\psi}|_F = \widehat{\boldsymbol{p}}$. In the next subsection we will focus on generalizations in our framework and there we will have to state the guarantees for the mapping $\widehat{\psi}$. We defer all the proofs to the Appendix

Lemma 3.1 (Regret of the Auditor)**.** *Let $0 < \delta < 1$. Let $\{\boldsymbol{\lambda}_t\}_{t=1}^T$ be the sequence of exponentiated gradient descent plays (with learning rate η) by the Auditor to the given $\{\boldsymbol{h}_t, \gamma_t\}_{t=1}^T$ of the Learner over T rounds of Algorithm 2. We have that for any set of individuals X, with probability at*

Mapping 3: $\widehat{\boldsymbol{\psi}}\left(X, \widehat{W}\right)$ – pseudocode for the mapping $\widehat{\psi}$ output by Algorithm 2

Input: $f \in \mathcal{F}$ (represented to $\widehat{\psi}$ as $\{f(x_i)\}_{i=1}^n$)

for $t = 1, \ldots, T$ **do**
  $c_i^1 \leftarrow (w_{i,t} + 1/n)(1 - f(x_i))$ for $i \in [n]$
  $c_i^0 \leftarrow (w_{i,t} + 1/n)f(x_i)$ for $i \in [n]$
  $D \leftarrow \{x_i, c_i^1, c_i^0\}_{i=1}^n$
  $h_{f,\boldsymbol{w}_t} \leftarrow CSC\left(\mathcal{H}; D\right)$
end
$\widehat{\psi}_f \leftarrow \frac{1}{T}\sum_{t=1}^T h_{f,\boldsymbol{w}_t}$

Output: $\widehat{\psi}_f \in \Delta(\mathcal{H})$

least $1-\delta$ over the problems F, the (average) regret of the Auditor is bounded as follows.

$$\frac{1}{T}\max_{\boldsymbol{\lambda}\in\Lambda}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t,\gamma_t,\boldsymbol{\lambda}\right)-\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t,\gamma_t,\boldsymbol{\lambda}_t\right)\leqslant B\sqrt{\frac{\log\left(2nT/\delta\right)}{2m_0}}+\frac{B\log\left(2n+1\right)}{\eta T}+\eta B\left(1+2\alpha\right)^2$$

The last two terms appearing in the above regret bound come from the usual regret analysis of the exponentiated gradient descent algorithm. However, the first term originates from a high probability Chernoff-Hoeffding bound because as explained before, the Auditor is using — instead of the whole set of problems F — only m_0 randomly selected problem F_t to estimate the vector of fairness violations $\boldsymbol{r}$ at round t. Hence at each round t, the difference of fairness violation estimates — one with respect to F_t and another with respect to F — will appear in the regret of the Auditor which can be bounded by the Chernoff-Hoeffding's inequality. We will therefore have to assume that m_0 is sufficiently large to make the above regret bound small enough.

Assumption 3.1. *For a given confidence parameter δ, inputs α and ν of Algorithm 2, we suppose throughout this section that the number of fresh problems m_0 used in each round of Algorithm 2 satisfies $m_0 \geqslant O\left(\frac{\log(nT/\delta)}{\alpha^2\nu^2}\right)$, or equivalently $m = m_0 \cdot T \geqslant O\left(\frac{T\log(nT/\delta)}{\alpha^2\nu^2}\right)$.*

Following Lemma 3.1 and Assumption 3.1, we characterize the average plays $(\widehat{\boldsymbol{p}},\,\widehat{\gamma},\,\widehat{\boldsymbol{\lambda}})$ output by Algorithm 2 in the following lemma. Informally speaking, this lemma guarantees that neither player would gain more than ν if they deviated from these proposed strategies output by the algorithm. This is what we call a *ν-approximate equilibrium* of the game. The proof of the lemma follows from the regret analysis of the Auditor and is fully presented in the Appendix.

Lemma 3.2 (Average Play Characterization)**.** *Let $0<\delta<1$. We have that under Assumption 3.1, for any set of individuals X, with probability at least $1-\delta$ over the labelings F, the average plays $(\widehat{\boldsymbol{p}},\widehat{\gamma},\widehat{\boldsymbol{\lambda}})$ output by Algorithm 2 forms a ν-approximate equilibrium of the game, i.e.,*

$$\begin{aligned}\mathcal{L}\left(\widehat{\boldsymbol{p}},\widehat{\gamma},\widehat{\boldsymbol{\lambda}}\right) &\leqslant \mathcal{L}\left(\boldsymbol{p},\gamma,\widehat{\boldsymbol{\lambda}}\right)+\nu \quad \textit{for all} \;\; \boldsymbol{p}\in\Delta(\mathcal{H})^m,\gamma\in[0,1]\\ \mathcal{L}\left(\widehat{\boldsymbol{p}},\widehat{\gamma},\widehat{\boldsymbol{\lambda}}\right) &\geqslant \mathcal{L}\left(\widehat{\boldsymbol{p}},\widehat{\gamma},\boldsymbol{\lambda}\right)-\nu \quad \textit{for all} \;\; \boldsymbol{\lambda}\in\Lambda\end{aligned}$$

We are now ready to present the main theorem of this subsection which takes the guarantees provided in Lemma 3.2 and turns them into accuracy and fairness guarantees of the pair $(\widehat{\boldsymbol{p}},\widehat{\gamma})$ using the specific form of the Lagrangian (6). The theorem will in fact show that the set of randomized classifiers $\widehat{\boldsymbol{p}}$ achieves optimal accuracy up to $O(\nu)$ and that it also satisfies $(O(\alpha),0)$-AIF notion of fairness, all with respect to the empirical distributions $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{Q}}$.

Theorem 3.3 (In-sample Accuracy and Fairness)**.** *Let $0<\delta<1$ and suppose Assumption 3.1 holds. Let $(\widehat{\boldsymbol{p}},\widehat{\gamma})$ be the output of Algorithm 2 and let $(\boldsymbol{p},\gamma)$ be any feasible pair of variables for the empirical fair learning problem (4). We have that for any set of individuals X, with probability at least $1-\delta$ over the labelings F,*

$$err\left(\widehat{\boldsymbol{p}};\widehat{\mathcal{P}},\widehat{\mathcal{Q}}\right)\leqslant err\left(\boldsymbol{p};\widehat{\mathcal{P}},\widehat{\mathcal{Q}}\right)+2\nu$$

and that $\widehat{\boldsymbol{p}}$ satisfies $(3\alpha,0)$-AIF with respect to the empirical distributions $(\widehat{\mathcal{P}},\widehat{\mathcal{Q}})$. In other words, for all $i\in[n]$,

$$\left|\mathcal{E}\left(x_i,\widehat{\boldsymbol{p}};\widehat{\mathcal{Q}}\right)-\widehat{\gamma}\right|\leqslant 3\alpha$$

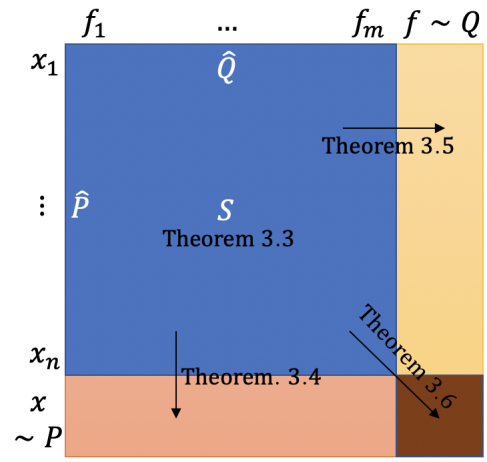


Figure 1: Illustration of generalization directions.

3.5 Generalization Theorems

When it comes to out-of-sample performance in our framework, unlike in usual learning settings, there are two distributions we need to reason about: the individual distribution $\mathcal{P}$ and the problem distribution $\mathcal{Q}$ (see Figure 1 for a visual illustration of generalization directions in our framework). We need to argue that $\widehat{\psi}$ induces a mapping that is accurate with respect to $\mathcal{P}$ and $\mathcal{Q}$, and is fair for almost every individual $x \sim \mathcal{P}$, where fairness is defined with respect to the true problem distribution $\mathcal{Q}$. Given these two directions for generalization, we state our generalization guarantees in three steps visualized by arrows in Figure 1. First, in Theorem 3.4, we fix the empirical distribution of the problems $\widehat{\mathcal{Q}}$ and show that the output $\widehat{\psi}$ of Algorithm 2 is accurate and fair with respect to the underlying individual distribution $\mathcal{P}$ as long as n is sufficiently large. Second, in Theorem 3.5, we fix the empirical distribution of individuals $\widehat{\mathcal{P}}$ and consider generalization along the underlying problem generating distribution $\mathcal{Q}$. It will follow from the dynamics of the algorithm, as well as the structure of $\widehat{\psi}$, that the learned mapping $\widehat{\psi}$ will remain accurate and fair with respect to $\mathcal{Q}$. We will eventually put these pieces together in Theorem 3.6 and argue that $\widehat{\psi}$ is accurate and fair with respect to the underlying distributions $(\mathcal{P}, \mathcal{Q})$ simultaniously, given that both n and m are large enough. We will use OPT (see Definition 3.1) as a benchmark to evaluate the accuracy of the mapping $\widehat{\psi}$.

Theorem 3.4 (Generalization over $\mathcal{P}$)**.** *Let $0 < \delta < 1$. Let $\left(\widehat{\psi}, \widehat{\gamma}\right)$ be the outputs of Algorithm 2, and suppose*

$$n \geqslant \tilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right)$$

where $d_{\mathcal{H}}$ is the VC dimension of $\mathcal{H}$. We have that with probability at least $1 - 5\delta$ over the observed data set (X, F), the mapping $\widehat{\psi}$ satisfies $(5\alpha, \beta)$-AIF with respect to the distributions $\left(\mathcal{P}, \widehat{\mathcal{Q}}\right)$, i.e.,

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left(\left|\mathcal{E}\left(x, \widehat{\psi}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| > 5\alpha\right) \leqslant \beta$$

and that,

$$err\left(\widehat{\psi}; \mathcal{P}, \widehat{\mathcal{Q}}\right) \leqslant OPT\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right) + O\left(\nu\right) + O\left(\alpha\beta\right)$$

The proof of this theorem will use standard VC-type generalization techniques where a Chernoff-Hoeffding bound is followed by a union bound (accompanied with the two-sample trick and Sauer's Lemma) to guarantee a uniform convergence of the empirical estimates to their true expectation. However, compared to the standard VC-based sample complexity bounds in learning theory, there is an extra factor of m because there are m hypotheses to be learned and that the α^2 factor appears in the denominator since in our setting a uniform convergence for all – pure – classifiers will *not* simply lead to a uniform convergence for all – randomized – classifiers without blowing up the sample complexity (specifically when proving generalization for fairness). We will therefore directly prove uniform convergence for randomized classifiers and that our argument will go through by sparsifying the distributions $\widehat{\boldsymbol{p}} = \widehat{\psi}|_F$ (taking samples of size $\widetilde{O}\left(1/\alpha^2\right)$ from $\widehat{\boldsymbol{p}}$) coupled with a uniform convergence for $\widetilde{O}\left(1/\alpha^2\right)$-sparse classifiers (randomized classifiers with support size $\leqslant \widetilde{O}\left(1/\alpha^2\right)$) and this is how α^2 shows up in the sample complexity bound.

Theorem 3.5 (Generalization over $\mathcal{Q}$)**.** *Let $0 < \delta < 1$. Let $\left(\widehat{\psi}, \widehat{\gamma}\right)$ be the outputs of Algorithm 2 and suppose*

$$m \geqslant \widetilde{O}\left(\frac{\log(n)\log(n/\delta)}{\nu^4\alpha^4}\right)$$

We have that for any set of observed individuals X, with probability at least $1 - 6\delta$ over the observed problems F, the learned mapping $\widehat{\psi}$ satisfies $(4\alpha, 0)$-AIF with respect to the distributions $\left(\widehat{\mathcal{P}}, \mathcal{Q}\right)$, i.e.,

$$\mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}}\left(\left|\mathcal{E}\left(x, \widehat{\psi}; \mathcal{Q}\right) - \widehat{\gamma}\right| > 4\alpha\right) = 0$$

and that,

$$err\left(\widehat{\psi}; \widehat{\mathcal{P}}, \mathcal{Q}\right) \leqslant OPT\left(\alpha; \widehat{\mathcal{P}}, \mathcal{Q}\right) + O(\nu)$$

This theorem will follow directly from Chernoff-type concentration inequalities where the fact that in each round the Auditor in our algorithm is using only a fresh batch of randomly selected m_0 problems to estimate the fairness violations will help us to prove concentration without appealing to a uniform convergence. The sample complexity for m stated in this theorem is equivalent to that of Assumption 3.1 because we needed almost the same type of concentration for controlling the regret of the Auditor in the previous subsection. Having proved generalization separately for $\mathcal{P}$ and $\mathcal{Q}$, we are now ready to state the final theorem of this section which provides generalization guarantees simultaneously over both distributions $\mathcal{P}$ and $\mathcal{Q}$.

Theorem 3.6 (Simultaneous Generalization over $\mathcal{P}$ and $\mathcal{Q}$)**.** *Let $0 < \delta < 1$. Let $\left(\widehat{\psi}, \widehat{\gamma}\right)$ be the outputs of Algorithm 2 and suppose*

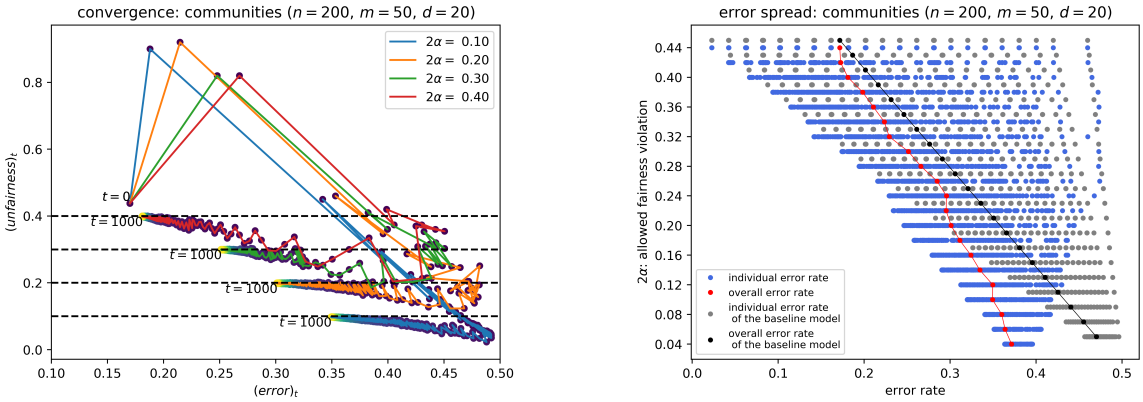
$$n \geqslant \widetilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right) \quad , \quad m \geqslant \widetilde{O}\left(\frac{\log(n)\log(n/\delta)}{\nu^4\alpha^4}\right)$$

where $d_{\mathcal{H}}$ is the VC dimension of $\mathcal{H}$. We have that with probability at least $1 - 12\delta$ over the observed data set (X, F), the learned mapping $\widehat{\psi}$ satisfies $(6\alpha, 2\beta)$-AIF with respect to the distributions $(\mathcal{P}, \mathcal{Q})$, i.e.,

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left(\left|\mathcal{E}\left(x, \widehat{\psi}; \mathcal{Q}\right) - \widehat{\gamma}\right| > 6\alpha\right) \leqslant 2\beta$$

and that,

$$err\left(\widehat{\psi}; \mathcal{P}, \mathcal{Q}\right) \leqslant OPT(\alpha; \mathcal{P}, \mathcal{Q}) + O(\nu) + O(\alpha\beta)$$



(a) convergence plot: communities data set

(b) error spread plot: communities data set

Figure 2: (a) Error-unfairness trajectory plots illustrating the convergence of algorithm **AIF-Learn**. (b) Error-unfairness tradeoffs and individual errors for **AIF-Learn** vs. simple mixtures of the error-optimal model and random classification. Gray dots are shifted upwards slightly to avoid occlusions.

To prove this theorem we basically start with the guarantees we have for the empirical distributions $(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$ and lift them into their corresponding guarantees for $(\mathcal{P}, \widehat{\mathcal{Q}})$ by Theorem 3.4. We will then have to take another "lifting" step from $(\mathcal{P}, \widehat{\mathcal{Q}})$ to $(\mathcal{P}, \mathcal{Q})$ which is not quite similar to what we have shown in Theorem 3.5 and will be proved as a separate lemma in the Appendix. Note that the bounds on n and m in Theorem 3.6 are mutually dependent: n must be *linear* in m, but m need only be *logarithmic* in n, and so both bounds can be simultaneously satisfied with sample complexity that is only polynomial in the parameters of the problem.

4 Experimental Evaluation

We have implemented the **AIF-Learn** algorithm and conclude with a brief experimental demonstration of its practical efficacy using the Communities and Crime dataset[3], which contains U.S. census records with demographic information at the neighborhood level. To obtain a challenging instance of our multi-problem framework, we treated each of the first $n = 200$ neighborhoods as the "individuals" in our sample, and binarized versions of the first $m = 50$ variables as distinct prediction problems. Another $d = 20$ of the variables were used as features for learning. For the base learning oracle assumed by **AIF-Learn**, we used a linear threshold learning heuristic that has worked well in other oracle-efficient reductions ([Kearns et al., 2018]). To elaborate, we build linear regression models for predicting the cost vectors $C_0 \in \mathbb{R}^n$ and $C_1 \in \mathbb{R}^n$ (where C_y is the vector of costs for predicting y) and classify an individual according to the lowest predicted cost by the regression models. We note that all results are reported in-sample, as our primary interest here is not in generalization (which we addressed in the theory), but in the effectiveness of our algorithm for the constrained optimization problem it is solving on the data, and how it compares to a simple alternative.

[3]Described in detail and available for download at http://archive.ics.uci.edu/ml/datasets/communities+and+crime

Despite the absence of worst-case guarantees for the linear threshold heuristic, **AIF-Learn** seems to empirically enjoy the strong convergence properties suggested by the theory. In Figure 2(a) we show trajectory plots of the learned model's error (x axis) versus its fairness violation (variation in cross-problem individual error rates, y axis) over 1000 iterations of the algorithm for varying values of the allowed fairness violation 2α (dashed horizontal lines). In each case we see the trajectory eventually converge to a point which saturates the fairness constraint with the optimal error.

In Figure 2(b) we provide a more detailed view of the behavior and performance of **AIF-Learn**. The x axis measures error rates, while the y axis measures the allowed fairness violation. For each value of the allowed fairness violation 2α (which is the allowed gap between the smallest and largest individual errors on input α), there is a horizontal row of 200 blue dots showing the error rates for each individual, and a single red dot representing the overall average of those individual error rates. As expected, for large α (weak or no fairness constraint), the overall error rate is lowest, but the spread of individual error rates (unfairness) is greatest. As α is decreased, the spread of individual error rates is greatly narrowed, at a cost of greater overall error.

A trivial way of achieving zero variability in individual error rates is to make all predictions randomly. So as a baseline comparison for **AIF-Learn**, the gray dots show the individual error rates achieved by different mixtures of the unconstrained error-optimal model with random classifications, with a black dot representing the overall average of these rates. When the weight on random classification is low (weak or no fairness, top row of gray dots), the overall error is lowest and the individual variation (unfairness) is highest. As we increase the weight on random classification, variation or unfairness decreases and the overall error gets worse. It is clear from the figure that **AIF-Learn** is considerably outperforming this baseline, both in terms of the average errors (red vs. black lines) and the individual errors (blue vs. gray dots). Additional experimental results are provided in the Appendix.

Acknowledgements

AR is supported in part by NSF grants AF-1763307 and CNS-1253345, and an Amazon Research Award.

References

[Abdulkadiroğlu et al., 2005] Abdulkadiroğlu, A., Pathak, P. A., and Roth, A. E. (2005). The new york city high school match. *American Economic Review*, 95(2):364–367.

[Agarwal et al., 2018] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69, Stockholmsmässan, Stockholm Sweden. PMLR.

[Chouldechova, 2017] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

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

[Gillen et al., 2018] Gillen, S., Jung, C., Kearns, M., and Roth, A. (2018). Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609.

[Hardt et al., 2016] Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

[Hébert-Johnson et al., 2018] Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR.

[Joseph et al., 2018] Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2018). Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163. ACM.

[Joseph et al., 2016] Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333.

[Kearns et al., 2018] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577.

[Kearns et al., 2019] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109. ACM.

[Kearns and Vazirani, 1994] Kearns, M. J. and Vazirani, U. V. (1994). *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, USA.

[Kim et al., 2018a] Kim, M., Reingold, O., and Rothblum, G. (2018a). Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852.

[Kim et al., 2018b] Kim, M. P., Ghorbani, A., and Zou, J. Y. (2018b). Multiaccuracy: Black-box post-processing for fairness in classification. *CoRR*, abs/1805.12317.

[Kleinberg et al., 2017] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, page 43. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Mitchell et al., 2018] Mitchell, S., Potash, E., and Barocas, S. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

[Shalev-Shwartz, 2012] Shalev-Shwartz, S. (2012). Online learning and online convex optimiza-

tion. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

[Sion, 1958] Sion, M. (1958). On general minimax theorems. *Pacific J. Math.*, 8(1):171 – 176.

[Yona and Rothblum, 2018] Yona, G. and Rothblum, G. N. (2018). Probably approximately metric-fair learning. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5666–5674. PMLR.

Appendix

A Learning subject to False Positive AIF (FPAIF)

Definition A.1 (Individual False Positive/Negative Error Rates)**.** *For a given individual $x \in \mathcal{X}$, a mapping $\psi \in \Delta(\mathcal{H})^{\mathcal{F}}$, and distribution $\mathcal{Q}$ over the space of problems, the Individual false positive/negative rate incurred by ψ on x are defined as follows:*

$$\mathcal{E}_{FP}(x, \psi; \mathcal{Q}) = \frac{1}{\mathop{\mathbb{P}}\limits_{f \sim \mathcal{Q}}[f(x) = 0]} \cdot \mathop{\mathbb{E}}\limits_{f \sim \mathcal{Q}}\left[\mathop{\mathbb{P}}\limits_{h \sim \psi_f}[h(x) = 1,\ f(x) = 0]\right]$$

$$\mathcal{E}_{FN}(x, \psi; \mathcal{Q}) = \frac{1}{\mathop{\mathbb{P}}\limits_{f \sim \mathcal{Q}}[f(x) = 1]} \cdot \mathop{\mathbb{E}}\limits_{f \sim \mathcal{Q}}\left[\mathop{\mathbb{P}}\limits_{h \sim \psi_f}[h(x) = 0,\ f(x) = 1]\right]$$

Definition A.2 (FPAIF fairness notion)**.** *In our framework, we say a mapping $\psi \in \Delta(\mathcal{H})^{\mathcal{F}}$ satisfies "(α, β)-FPAIF" (reads (α, β)-approximate False Positive Average Individual Fairness) with respect to the distributions $(\mathcal{P}, \mathcal{Q})$ if there exists $\gamma \geqslant 0$ such that*

$$\mathop{\mathbb{P}}\limits_{x \sim \mathcal{P}}(|\mathcal{E}_{FP}(x, \psi; \mathcal{Q}) - \gamma| > \alpha) \leqslant \beta$$

In this section we consider learning subject to the FPAIF notion of fairness. The FPAIF fairness notion basically asks that the individual false positive rates be approximately (corresponds to α) equalized across almost all (corresponds to β) individuals. Learning subject to equalizing false negative rates can be developed similarly. We will be less wordy in this section as the ideas and the approach that we take are mostly similar to those developed in Section 3. We start off by casting the fair learning problem as the constrained optimization problem (10) where a mapping ψ is to be found such that all individual false positive rates incurred by ψ are within α of some γ. As before, we denote the optimal error of the optimization problem (10) by OPT and will consider that as a benchmark to evaluate the accuracy of our algorithm's trained mapping.

Fair Learning Problem subject to $(\alpha, 0)$-FPAIF

$$\begin{aligned} \min_{\psi \in \Delta(\mathcal{H})^{\mathcal{F}},\, \gamma \in [0,1]} \quad & \mathrm{err}(\psi; \mathcal{P}, \mathcal{Q}) \\ \text{s.t. } \forall x \in \mathcal{X}: \quad & |\mathcal{E}_{\mathrm{FP}}(x, \psi; \mathcal{Q}) - \gamma| \leqslant \alpha \end{aligned} \tag{10}$$

Definition A.3. *Consider the optimization problem (10). Given distributions $\mathcal{P}$ and $\mathcal{Q}$ and fairness approximation parameter α, we denote the optimal solutions of (10) by $\psi^\star(\alpha; \mathcal{P}, \mathcal{Q})$ and $\gamma^\star(\alpha; \mathcal{P}, \mathcal{Q})$, and the value of the objective function at these optimal points by $OPT(\alpha; \mathcal{P}, \mathcal{Q})$. In other words*

$$OPT(\alpha; \mathcal{P}, \mathcal{Q}) = err(\psi^\star; \mathcal{P}, \mathcal{Q})$$

It is important to observe that the optimization problem (10) has a nonempty set of feasible points for any α and any distributions $\mathcal{P}$ and $\mathcal{Q}$ because the following is always feasible: $\gamma = 0$ and $\psi_f = h^0$ for all $f \in \mathcal{F}$ where h^0 is the all-zero constant classifier. Since the distributions $\mathcal{P}$ and $\mathcal{Q}$ are generally not known, we instead solve the *empirical* version of (10). Consider a training data set consisting of n individuals $X = \{x_i\}_{i=1}^n$ drawn independently from $\mathcal{P}$ and m problems $F = \{f_j\}_{j=1}^m$ drawn independently from $\mathcal{Q}$. We formulate the empirical fair learning problem in (11) where we find an optimal fair mapping of the training problems F to $\Delta(\mathcal{H})$ given the individuals X.

Empirical Fair Learning Problem

$$\begin{aligned} \min_{\boldsymbol{p} \in \Delta(\mathcal{H})^m,\, \gamma \in [0,1]} \quad & \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \\ \text{s.t. } \forall i \in \{1, \dots, n\}: \quad & \left|\mathcal{E}_{\text{FP}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - \gamma\right| \leqslant 2\alpha \end{aligned} \tag{11}$$

As also discussed in Section 3, solving the empirical problem (11) will only give us a – restricted – mapping $\boldsymbol{p} = \psi|_F$ by which we mean we learn ψ only on the finite domain $F \subseteq \mathcal{F}$. It is not clear for now how we can extend the restricted mapping $\psi|_F \in \Delta(\mathcal{H})^m$ to a mapping $\psi \in \Delta(\mathcal{H})^{\mathcal{F}}$ defined over the entire function space $\mathcal{F}$; however, we will see the specific form of our algorithm (as in the AIF setting: learning a set of weights over the training individuals) will allow us to come up with such an extension.

We use the dual perspective of constrained optimization problems followed by no regret dynamics to reduce (11) to a two-player zero-sum game between a Learner (primal player) and an Auditor (dual player), and design an iterative algorithm to get an approximate equilibrium of the game. To do so, we first rewrite the constraints of (11) in $\boldsymbol{r}_{\text{FP}}\left(\boldsymbol{p}, \gamma; \widehat{\mathcal{Q}}\right) \leqslant 0$ form where

$$\boldsymbol{r}_{\text{FP}}\left(\boldsymbol{p}, \gamma; \widehat{\mathcal{Q}}\right) = \begin{bmatrix} \mathcal{E}_{\text{FP}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - \gamma - 2\alpha \\ \gamma - \mathcal{E}_{\text{FP}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - 2\alpha \end{bmatrix}_{i=1}^n \in \mathbb{R}^{2n} \tag{12}$$

represents the "fairness violations" of the pair $(\boldsymbol{p}, \gamma)$ in one single vector. Let the corresponding dual variables be represented by $\boldsymbol{\lambda} = \left[\boldsymbol{\lambda}_i^+, \boldsymbol{\lambda}_i^-\right]_{i=1}^n \in \Lambda$, where $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}_+^{2n} \mid ||\boldsymbol{\lambda}||_1 \leqslant B\}$. Using Equation (12) and the introduced dual variables, we have that the Lagrangian of (11) is

$$\mathcal{L}(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}) = \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + \boldsymbol{\lambda}^T \boldsymbol{r}_{\text{FP}}\left(\boldsymbol{p}, \gamma; \widehat{\mathcal{Q}}\right) \tag{13}$$

Therefore, we focus on solving the following minmax problem:

$$\min_{\boldsymbol{p} \in \Delta(\mathcal{H})^m,\, \gamma \in [0,1]} \max_{\lambda \in \Lambda} \mathcal{L}(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}) = \max_{\lambda \in \Lambda} \min_{\boldsymbol{p} \in \Delta(\mathcal{H})^m,\, \gamma \in [0,1]} \mathcal{L}(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}) \tag{14}$$

Using no regret dynamics, an approximate equilibrium of this zero-sum game (i.e. a saddle point of $\mathcal{L}$) can be found in an iterative framework. In each iteration, we let the dual player run the *exponentiated gradient descent* algorithm and the primal player run an approximate version of her *best response* using the oracle $CSC(\mathcal{H})$ that solves cost sensitive classification problems in $\mathcal{H}$ (see Definition 2.3). In the following subsection, we will first describe the best response problem of the Learner, and show how the best response depends on the estimates $\widehat{\boldsymbol{\rho}} = [\widehat{\rho}_i]_{i=1}^n \in \mathbb{R}^n$ – where $\widehat{\rho}_i$ is the fraction of problems in F that maps x_i to 0 – in addition to the weights $\boldsymbol{w} = \left[\lambda_i^+ - \lambda_i^-\right] \in \mathbb{R}^n$ maintained by the Auditor. We will then argue that to avoid injecting correlation into the algorithm (so as to argue later on about generalization) we have to "perturb" the best response of the Learner by using some other set of estimates $\widetilde{\boldsymbol{\rho}} = [\widetilde{\rho}_i]_{i=1}^n$ which is given to the algorithm and is independent of F. This is why the Learner is basically using an *approximate* version of her best response.

Definition A.4. *For an individual $x \in \mathcal{X}$, let ρ_x and $\widehat{\rho}_x$ represent the probability that x is labelled 0 by a randomly sampled function $f \sim \mathcal{Q}$ and $f \sim \widehat{\mathcal{Q}}$, respectively. In other words*

$$\rho_x = \mathop{\mathbb{P}}_{f\sim\mathcal{Q}}[f(x) = 0] \quad , \quad \widehat{\rho}_x = \mathop{\mathbb{P}}_{f\sim\widehat{\mathcal{Q}}}[f(x) = 0]$$

we will use $\rho_i \equiv \rho_{x_i}$ and $\widehat{\rho}_i \equiv \widehat{\rho}_{x_i}$ to denote the corresponding probabilities for x_i in the training set of individuals X.

Remark A.1. *Observe that using the introduced notation, for a mapping ψ and $x \in \mathcal{X}$,*

$$\mathcal{E}_{FP}(x, \psi; \mathcal{Q}) = \left(\frac{1}{\rho_x}\right) \mathop{\mathbb{E}}_{f\sim\mathcal{Q}}\left[\mathop{\mathbb{P}}_{h\sim\psi_f}[h(x) = 1,\ f(x) = 0]\right]$$

$$\mathcal{E}_{FN}(x, \psi; \mathcal{Q}) = \left(\frac{1}{1-\rho_x}\right) \mathop{\mathbb{E}}_{f\sim\mathcal{Q}}\left[\mathop{\mathbb{P}}_{h\sim\psi_f}[h(x) = 0,\ f(x) = 1]\right]$$

and that $\mathcal{E}(x, \psi; \mathcal{Q})$ can be written as a linear combination of $\mathcal{E}_{FP}(x, \psi; \mathcal{Q})$ and $\mathcal{E}_{FN}(x, \psi; \mathcal{Q})$:

$$\mathcal{E}(x, \psi; \mathcal{Q}) = \rho_x \cdot \mathcal{E}_{FP}(x, \psi; \mathcal{Q}) + (1 - \rho_x) \cdot \mathcal{E}_{FN}(x, \psi; \mathcal{Q})$$

A.1 aBEST$_{\mathbf{FP}}$: The Learner's approximate Best Response

In each iteration of the algorithm, the Learner is given some $\boldsymbol{\lambda}$ of the dual player and she wants to solve the following minimization problem.

$$\begin{aligned}
&\mathop{\arg\min}_{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]} \mathcal{L}(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}) \\
\equiv &\mathop{\arg\min}_{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]} \mathrm{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + \sum_{i=1}^n \left\{\lambda_i^+\left(\mathcal{E}_{\mathrm{FP}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - \gamma\right) + \lambda_i^-\left(\gamma - \mathcal{E}_{\mathrm{FP}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right)\right)\right\} \\
\equiv &\mathop{\arg\min}_{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\,\gamma\in[0,1]} \frac{1}{n}\sum_{i=1}^n \mathcal{E}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) + \sum_{i=1}^n \left\{\lambda_i^+\left(\mathcal{E}_{\mathrm{FP}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) - \gamma\right) + \lambda_i^-\left(\gamma - \mathcal{E}_{\mathrm{FP}}\left(x_i; \boldsymbol{p}; \widehat{\mathcal{Q}}\right)\right)\right\}
\end{aligned}$$

Let $\widehat{\rho}_i$ be defined as in Definition A.4. We have that by Remark A.1

$$\mathcal{E}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) = \widehat{\rho}_i \cdot \mathcal{E}_{\mathrm{FP}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right) + (1 - \widehat{\rho}_i) \cdot \mathcal{E}_{\mathrm{FN}}\left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}}\right)$$

Now let $w_i = \lambda_i^+ - \lambda_i^-$ for all i (Accordingly let $\boldsymbol{w} = [w_1, \ldots, w_n]^\top \in \mathbb{R}^n$). We have that the above minimization problem is equivalent to

$$\equiv \underset{\boldsymbol{p} \in \Delta(\mathcal{H})^m,\, \gamma \in [0,1]}{\arg\min} \; -\gamma \sum_{i=1}^n w_i + \sum_{i=1}^n \left\{ \left(\frac{\widehat{\rho}_i}{n} + w_i \right) \mathcal{E}_{\mathrm{FP}} \left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}} \right) + \left(\frac{1 - \widehat{\rho}_i}{n} \right) \mathcal{E}_{\mathrm{FN}} \left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}} \right) \right\}$$

Now we can use the fact that

$$\mathcal{E}_{\mathrm{FP}} \left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}} \right) = \frac{1}{m\widehat{\rho}_i} \sum_{j=1}^m \mathop{\mathbb{P}}_{h_j \sim p_j} \left[h_j(x_i) = 1, f_j(x_i) = 0 \right]$$

$$\mathcal{E}_{\mathrm{FN}} \left(x_i, \boldsymbol{p}; \widehat{\mathcal{Q}} \right) = \frac{1}{m(1 - \widehat{\rho}_i)} \sum_{j=1}^m \mathop{\mathbb{P}}_{h_j \sim p_j} \left[h_j(x_i) = 0, f_j(x_i) = 1 \right]$$

to conclude that the best response of the Learner is equivalent to minimizing the following function over the space of $(\boldsymbol{p}, \gamma) \in \Delta(\mathcal{H})^m \times [0, 1]$.

$$\begin{aligned} &- \gamma \sum_{i=1}^n w_i \\ &+ \frac{1}{m} \sum_{j=1}^m \left\{ \sum_{i=1}^n \left(\frac{1}{n} + \frac{w_i}{\widehat{\rho}_i} \right) \mathop{\mathbb{P}}_{h_j \sim p_j} \left[h_j(x_i) = 1, f_j(x_i) = 0 \right] + \left(\frac{1}{n} \right) \mathop{\mathbb{P}}_{h_j \sim p_j} \left[h_j(x_i) = 0, f_j(x_i) = 1 \right] \right\} \end{aligned}$$

Therefore, as in the AIF setting, the minimization problem of the Learner gets nicely decoupled into $(m+1)$ disjoint minimization problems. First, the optimal value for γ is chosen according to

$$\gamma = \mathbb{1} \left[\sum_{i=1}^n w_i > 0 \right] \tag{15}$$

and that for learning problem j, the following cost sensitive classification problem must be solved.

$$h_j = \underset{h \in \mathcal{H}}{\arg\min} \sum_{i=1}^n \left\{ c_{i,j}^1 \, h(x_i) + c_{i,j}^0 \, (1 - h(x_i)) \right\} \tag{16}$$

where the costs are

$$c_{i,j}^1 = \left(\frac{1}{n} + \frac{w_i}{\widehat{\rho}_i} \right) (1 - f_j(x_i)) \quad , \quad c_{i,j}^0 = \left(\frac{1}{n} \right) f_j(x_i)$$

One major distinction between the Learner's best response in the FPAIF setting versus that of the AIF setting is that the empirical quantities $\{\widehat{\rho}_i\}_{i=1}^n$ (which is estimated using the data set F) appear in the costs of the CSC problems for the FPAIF setting. As a major consequence, the generalization (with respect to $\mathcal{Q}$) arguments we had in the AIF setting won't work in this section because now the labels $\{f_j(x_i)\}_{i,j}$ and the estimates $\{\widehat{\rho}_i\}_{i=1}^n$ are correlated. We therefore assume in this section that each individual $x_i \in X$ comes with an estimate $\widetilde{\rho}_i$ of the rate ρ_i that is independent of the data set F. More precisely stated, we assume our algorithm has access to estimates $\{\widetilde{\rho}_i\}_{i=1}^n$ such that for all $i \in [n]$, $|\widetilde{\rho}_i - \rho_i| \leqslant \sqrt{\log(n)/2m_0}$ where m_0 (will be specified exactly in our proposed algorithm) will be essentially the number of fresh problems

that the Auditor is using in each round of the Algorithm. In fact, similar to what we did for the AIF setting, we will randomly partition F into T batch of size m_0: $F = \{F_t\}_{t=1}^T$ and will let the Auditor to use only F_t at round $t \in [T]$ of the algorithm to update the vector of fairness violations $\boldsymbol{r}_{\text{FP}}$. Notice assuming access to the estimates $\{\widetilde{\rho}_i\}_{i=1}^n$ is not farfetched because we can assume there was one more batch of m_0 problems, say F_0, and that the quantities $\{\widetilde{\rho}_i\}_{i=1}^n$ were estimated using the batch F_0 which is independent of $F = \{F_t\}_{t=1}^T$. The upper bound we required for the difference $|\widetilde{\rho}_i - \rho_i|$ will just simply follow from a Chernoff-Hoeffding's bound.

Assumption A.1. *For m_0 specified later on, we assume in this section that our algorithm has access to quantities $\{\widetilde{\rho}_i\}_{i=1}^n$, where we have that for all $i \in [n]$: $|\widetilde{\rho}_i - \rho_i| \leqslant \sqrt{\frac{\log(n)}{2m_0}}$.*

Under Assumption A.1, we now modify the best response of the Learner and let it use the estimates $\{\widetilde{\rho}_i\}_{i=1}^n$ instead of $\{\widehat{\rho}_i\}_{i=1}^n$. This will consequently make the learner to accumulate regret over the course of the algorithm and that this is why we will call it the *approximate* best response of the Learner. We have the approximate best response of the Learner (called $\textbf{aBEST}_{\text{FP}}$) written in Subroutine 4.

Subroutine 4: $\textbf{aBEST}_{\text{FP}}$ – approximate best response of the Learner in the FPAIF setting

Input: dual weights $\boldsymbol{w} \in \mathbb{R}^n$, estimates $\widetilde{\boldsymbol{\rho}} \in \mathbb{R}^n$, training examples $S = \left\{x_i, (f_j(x_i))_{j=1}^m\right\}_{i=1}^n$

$\gamma \leftarrow \mathbb{1}\left[\sum_{i=1}^n w_i > 0\right]$
for $j = 1, \ldots, m$ **do**
  $c_i^1 \leftarrow (w_i/\widetilde{\rho}_i + 1/n)(1 - f_j(x_i))$ for $i \in [n]$
  $c_i^0 \leftarrow (1/n) f_j(x_i)$ for $i \in [n]$
  $D \leftarrow \{x_i, c_i^1, c_i^0\}_{i=1}^n$
  $h_j \leftarrow CSC\left(\mathcal{H}; D\right)$
end
$\boldsymbol{h} \leftarrow (h_1, h_2, \ldots, h_m)$

Output: $\left(\boldsymbol{h}, \gamma\right)$

A.2 FPAIF-Learn: Implementation and In-sample Guarantees

We implement the introduced game theoretic framework in Algorithm 5 and call it **FPAIF-Learn**. The overall style of the algorithm is similar to **AIF-Learn** (Algorithm 2) except that **FPAIF-Learn** takes a set of estimates $\widetilde{\boldsymbol{\rho}} \in \mathbb{R}^n$ as input and that $\widetilde{\boldsymbol{\rho}}$ is used in the approximate best response of the Learner. Once again, in the service of arguing for generalization, we split F into T batches of size m_0 uniformly at random and let the Auditor use only a fresh batch of m_0 problems in each round of the algorithm to update the dual variables $\boldsymbol{\lambda}$, and accordingly the weights $\boldsymbol{w}$ over the individuals. This is reflected in the algorithm by writing $\widehat{\mathcal{Q}}_t = \mathcal{U}(F_t)$ for the uniform distribution over F_t and $\boldsymbol{h}_t|_{F_t}$ for the corresponding learned classifiers of F_t. The algorithm will terminate after $T = O\left(\log(n)/(\nu^2\alpha^2)\right)$ iterations and output the average plays of the Learner and the Auditor, along with a mapping $\widehat{\psi} = \widehat{\psi}\left(X, \widetilde{\boldsymbol{\rho}}, \widehat{W}\right) \in \Delta(\mathcal{H})^{\mathcal{F}}$ which is the object we wanted to learn. In fact, $\widehat{\psi}$ extends the learned restricted mapping $\widehat{\boldsymbol{p}} = \widehat{\psi}|_F$ of the algorithm from the finite domain F to the entire space $\mathcal{F}$. We have the pseudocode for the mapping $\widehat{\psi}$ written in detail in Mapping 6.

Algorithm 5: FPAIF-Learn – learning subject to FPAIF

Input: fairness parameter α, approximation parameter ν,
estimates $\widetilde{\boldsymbol{\rho}} = [\widetilde{\rho}_i]_{i=1}^n$,
training data set $X = \{x_i\}_{i=1}^n$ and $F = \{f_j\}_{j=1}^m$

$$B \leftarrow \tfrac{1+2\nu}{\alpha}, \quad T \leftarrow \tfrac{16B^2(1+2\alpha)^2\log(2n+1)}{\nu^2}, \quad \eta \leftarrow \tfrac{\nu}{4(1+2\alpha)^2 B}, \quad m_0 \leftarrow \tfrac{m}{T}, \quad S \leftarrow \left\{x_i, (f_j(x_i))_{j=1}^m\right\}_{i=1}^n$$

Partition F uniformly at random: $F = \{F_t\}_{t=1}^T$ where $|F_t| = m_0$.
$\boldsymbol{\theta}_1 \leftarrow \mathbf{0} \in \mathbb{R}^{2n}$
for $t = 1, \ldots, T$ **do**
  $\lambda_{i,t}^{\bullet} \leftarrow B\frac{\exp(\theta_{i,t}^{\bullet})}{1+\sum_{i',\bullet'}\exp(\theta_{i',t}^{\bullet'})}$ for $i \in [n]$ and $\bullet \in \{+,-\}$
  $\boldsymbol{w}_t \leftarrow [\lambda_{i,t}^{+} - \lambda_{i,t}^{-}]_{i=1}^n \in \mathbb{R}^n$
  $(\boldsymbol{h}_t, \gamma_t) \leftarrow \mathbf{aBEST}_{\mathrm{FP}}(\boldsymbol{w}_t; \widetilde{\boldsymbol{\rho}}, S)$
  $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \eta \cdot \boldsymbol{r}_{\mathrm{FP}}\left(\boldsymbol{h}_t|_{F_t}, \gamma_t; \widehat{\mathcal{Q}}_t\right)$
end
$\widehat{\boldsymbol{p}} \leftarrow \frac{1}{T}\sum_{t=1}^T \boldsymbol{h}_t$, $\widehat{\gamma} \leftarrow \frac{1}{T}\sum_{t=1}^T \gamma_t$, $\widehat{\boldsymbol{\lambda}} \leftarrow \frac{1}{T}\sum_{t=1}^T \boldsymbol{\lambda}_t$, $\widehat{W} \leftarrow \{\boldsymbol{w}_t\}_{t=1}^T$

Output: average plays $\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right)$, mapping $\widehat{\psi} = \widehat{\psi}\left(X, \widetilde{\boldsymbol{\rho}}, \widehat{W}\right)$ (see Mapping 6)

The analysis of Algorithm 5 follows the same style and uses the same ideas as in the AIF learning section. We will start by establishing the regret bounds for the Learner and the Auditor in Lemma A.1 and Lemma A.2, respectively, and will show just as before how these regret bounds can be eventually turned into in-sample accuracy and fairness guarantees. One major distinction is that when working with false positive rates, the quantities $\rho_x, \widehat{\rho}_x, \widetilde{\rho}_x$ introduced before will show up in the algorithm's analysis. Define for the training individuals $X = \{x_i\}_{i=1}^n$

$$\rho_{\min} = \min_{i\in[n]} \rho_i \quad , \quad \widehat{\rho}_{\min} = \min_{i\in[n]} \widehat{\rho}_i \quad , \quad \widetilde{\rho}_{\min} = \min_{i\in[n]} \widetilde{\rho}_i$$

We will in fact see $\widetilde{\rho}_{\min}$ and $\widehat{\rho}_{\min}$ will appear in the regret bounds of the Learner and that of the Auditor, respectively.

Mapping 6: $\widehat{\boldsymbol{\psi}}\left(X, \widetilde{\boldsymbol{\rho}}, \widehat{W}\right)$ – pseudocode for the mapping $\widehat{\psi}$ output by Algorithm 5

Input: $f \in \mathcal{F}$ (represented to $\widehat{\psi}$ as $\{f(x_i)\}_{i=1}^n$)

for $t = 1, \ldots, T$ **do**
  $c_i^1 \leftarrow (w_{i,t}/\widetilde{\rho}_i + 1/n)(1 - f(x_i))$ for $i \in [n]$
  $c_i^0 \leftarrow (1/n)f(x_i)$ for $i \in [n]$
  $D \leftarrow \{x_i, c_i^1, c_i^0\}_{i=1}^n$
  $h_{f,\boldsymbol{w}_t} \leftarrow CSC\left(\mathcal{H}; D\right)$
end
$\widehat{\psi}_f \leftarrow \frac{1}{T}\sum_{t=1}^T h_{f,\boldsymbol{w}_t}$

Output: $\widehat{\psi}_f \in \Delta(\mathcal{H})$

Lemma A.1 (Regret of the Learner)**.** *Let $0 < \delta < 1$. Let $\{\boldsymbol{h}_t, \gamma_t\}_{t=1}^T$ be the sequence of approximate best response plays by the Learner to the given $\{\boldsymbol{\lambda}_t\}_{t=1}^T$ of the Auditor over T rounds of Algorithm 5. We have that for any set of individuals X, with probability at least $1-\delta/2$ over the problems F, the (average) regret of the Learner is bounded as follows:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) - \frac{1}{T}\min_{\boldsymbol{p}\in\Delta(\mathcal{H})^m,\, \gamma\in[0,1]}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}_t\right) \leqslant \frac{4B}{\widetilde{\rho}_{min}}\sqrt{\frac{\log\left(8nT/\delta\right)}{2m_0}}$$

Lemma A.2 (Regret of the Auditor)**.** *Let $0 < \delta < 1$. Let $\{\boldsymbol{\lambda}_t\}_{t=1}^T$ be the sequence of exponentiated gradient descent plays (with learning rate η) by the Auditor to the given $\{\boldsymbol{h}_t, \gamma_t\}_{t=1}^T$ of the Learner over T rounds of Algorithm 5. We have that for any set of individuals X, with probability at least $1-\delta/2$ over the problems F, the (average) regret of the Auditor is bounded as follows: For any $\boldsymbol{\lambda}\in\Lambda$,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}\right) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) \leqslant \frac{2B}{\widehat{\rho}_{min}}\sqrt{\frac{\log\left(8nT/\delta\right)}{2m_0}} + \frac{B\log\left(2n+1\right)}{\eta T} + \eta B\left(1+2\alpha\right)^2$$

Observe that in order to control the regret of the Learner and the Auditor at level $O(\nu)$ we need to assume that m_0 is large enough such that the regret bound of the Learner and the first term appearing in the regret bound of the Auditor are sufficiently small.

Assumption A.2. *For a given confidence parameter δ, inputs α and ν of Algorithm 5, we assume throughout this section that the number of fresh problems m_0 used in each round of Algorithm 5 satisfies $m_0 \geqslant O\left(\frac{\log(nT/\delta)}{\alpha^2\nu^2\rho_{min}^2}\right)$, or equivalently $m = m_0\cdot T \geqslant O\left(\frac{T\log(nT/\delta)}{\alpha^2\nu^2\rho_{min}^2}\right)$.*

Note that Assumption A.2 immediately implies via a Chernoff bound that $\widehat{\rho}_{\min} \geqslant \rho_{\min}/2$ and that it also implies via Assumption A.1 that $\widetilde{\rho}_{\min} \geqslant \rho_{\min}/2$. In the following lemma we characterize the average play of the Learner and the Auditor. The proof of this lemma follows from the regret bounds developed in Lemma A.1 (Regret of the Learner) and Lemma A.2 (Regret of the Auditor) and uses exactly the same techniques as in the proof of Lemma 3.2.

Lemma A.3 (Average Play Characterization)**.** *Let $0 < \delta < 1$. Let $\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right)$ be the average plays output by Algorithm 5. We have that under Assumption A.2, for any set of observed individuals X, with probability at least $1-\delta$ over the observed labelings F, the average plays $\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right)$ forms a ν-approximate equilibrium of the game, i.e.,*

$$\begin{aligned}
\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) &\leqslant \mathcal{L}\left(\boldsymbol{p}, \gamma, \widehat{\boldsymbol{\lambda}}\right) + \nu \quad \textit{for all } \boldsymbol{p}\in\Delta(\mathcal{H})^m, \gamma\in[0,1] \\
\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) &\geqslant \mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \boldsymbol{\lambda}\right) - \nu \quad \textit{for all } \boldsymbol{\lambda}\in\Lambda
\end{aligned}$$

We conclude this subsection with our main Theorem A.4 that provides in-sample accuracy and fairness guarantees for the learned set of classifiers $\widehat{\boldsymbol{p}}\in\Delta(\mathcal{H})^m$ of Algorithm 5. This theorem follows immediately from Lemma A.3 and the proof is pretty much similar in style to the proof of Theorem 3.3.

Theorem A.4 (In-sample Accuracy and Fairness)**.** *Let $0 < \delta < 1$ and suppose Assumption A.2 holds. Let $(\widehat{\boldsymbol{p}}, \widehat{\gamma})$ be the output of Algorithm 5 and let $(\boldsymbol{p}, \gamma)$ be any feasible pair of variables for the empirical fair learning problem (11). We have that for any set of individuals X, with probability at least $1 - \delta$ over the labelings F,*

$$err\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \leqslant err\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu$$

and that $\widehat{\boldsymbol{p}}$ satisfies $(3\alpha, 0)$-FPAIF with respect to $(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}})$. In other words, for all $i \in [n]$,

$$\left|\mathcal{E}_{FP}\left(x_i, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| \leqslant 3\alpha$$

A.3 Generalization Theorems

We now consider the mapping $\widehat{\psi}$ learned by Algorithm 5 and study the generalization bounds both for accuracy and fairness. As in the AIF learning setting, we state our generalization theorems in three steps. We first consider in Theorem A.5 the empirical distribution of the problems $\widehat{\mathcal{Q}}$ and see how we can lift the guarantees from $\widehat{\mathcal{P}}$ into the true underlying distribution of individuals $\mathcal{P}$. We will then consider generalization only over the problem generating distribution $\mathcal{Q}$ in Theorem A.6. We will eventually in Theorem A.7 provide accuracy and fairness guarantees for the learned mapping $\widehat{\psi}$ with respect to the distributions $(\mathcal{P}, \mathcal{Q})$. We will use OPT (defined formally in Definition A.3) as a benchmark to evaluate the accuracy of the mapping $\widehat{\psi}$. Proofs of these theorems are similar to their counterparts in the AIF section.

Theorem A.5 (Generalization over $\mathcal{P}$)**.** *Let $0 < \delta < 1$. Let $\left(\widehat{\psi}, \widehat{\gamma}\right)$ be the outputs of Algorithm 5, and suppose*

$$n \geqslant \widetilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right)$$

where $d_{\mathcal{H}}$ is the VC dimension of $\mathcal{H}$. We have that with probability at least $1 - 5\delta$ over the observed data set (X, F), the mapping $\widehat{\psi}$ satisfies $(5\alpha, \beta)$-FPAIF with respect to $\left(\mathcal{P}, \widehat{\mathcal{Q}}\right)$, i.e.,

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left(\left|\mathcal{E}_{FP}\left(x, \widehat{\psi}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| > 5\alpha\right) \leqslant \beta$$

and that,

$$err\left(\widehat{\psi}; \mathcal{P}, \widehat{\mathcal{Q}}\right) \leqslant OPT\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right) + O\left(\nu\right) + O\left(\alpha\beta\right)$$

Theorem A.6 (Generalization over $\mathcal{Q}$)**.** *Let $0 < \delta < 1$. Let $\left(\widehat{\psi}, \widehat{\gamma}\right)$ be the outputs of Algorithm 5 and suppose*

$$m \geqslant \widetilde{O}\left(\frac{\log\left(n\right)\log\left(n/\delta\right)}{\rho_{min}^2\nu^4\alpha^4}\right)$$

where $\rho_{min} = \min_{i \in [n]} \rho_i$. We have that for any set of observed individuals X, with probability at least $1 - 6\delta$ over the observed problems F, $\widehat{\psi}$ satisfies $(4\alpha, 0)$-FPAIF with respect to $\left(\widehat{\mathcal{P}}, \mathcal{Q}\right)$, i.e.,

$$\mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}}\left(\left|\mathcal{E}_{FP}\left(x, \widehat{\psi}; \mathcal{Q}\right) - \widehat{\gamma}\right| > 4\alpha\right) = 0$$

and that,

$$err\left(\widehat{\psi};\widehat{\mathcal{P}},\mathcal{Q}\right) \leqslant OPT\left(\alpha;\widehat{\mathcal{P}},\mathcal{Q}\right) + O\left(\nu\right)$$

Theorem A.7 (Simultaneous Generalization over $\mathcal{P}$ and $\mathcal{Q}$)**.** *Let $0 < \delta < 1$. Let $\left(\widehat{\psi},\widehat{\gamma}\right)$ be the outputs of Algorithm 5 and suppose*

$$n \geqslant \widetilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right) \quad , \quad m \geqslant \widetilde{O}\left(\frac{\log\left(n\right)\log\left(n/\delta\right)}{\rho_{inf}^2\nu^4\alpha^4}\right)$$

where $d_{\mathcal{H}}$ is the VC dimension of $\mathcal{H}$ and $\rho_{inf} = \inf_{x\in\mathcal{X}} \rho_x$. We have that with probability at least $1 - 12\delta$ over the observed data set (X, F), the learned mapping $\widehat{\psi}$ satisfies $(6\alpha, 2\beta)$-FPAIF with respect to the distributions $(\mathcal{P}, \mathcal{Q})$, i.e.,

$$\mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}_{FP}\left(x,\widehat{\psi};\mathcal{Q}\right) - \widehat{\gamma}\right| > 6\alpha\right) \leqslant 2\beta$$

and that,

$$err\left(\widehat{\psi};\mathcal{P},\mathcal{Q}\right) \leqslant OPT(\alpha;\mathcal{P},\mathcal{Q}) + O\left(\nu\right) + O\left(\alpha\beta\right)$$

B Proofs of Section 3

B.1 Preliminary Tools

Theorem B.1 (Additive Chernoff-Hoeffding Bound)**.** *Let $X = \{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables with $a \leqslant X_i \leqslant b$ and $\mathbb{E}\left[X_i\right] = \mu$ for all i. We have that for all $s > 0$,*

$$\mathbb{P}_X\left[\left|\frac{\sum_i X_i}{n} - \mu\right| \geqslant s\right] \leqslant 2\exp\left(\frac{-2ns^2}{(b-a)^2}\right)$$

Lemma B.2 (Sauer's Lemma (see e.g. [Kearns and Vazirani, 1994]))**.** *Let $\mathcal{H}$ be a class of binary functions defined on $\mathcal{X}$ where the VC dimension of $\mathcal{H}$, $d_{\mathcal{H}}$, is finite. Let $X = \{x_i\}_{i=1}^n$ be a data set of size n drwan from $\mathcal{X}$ and let $\mathcal{H}(X) = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$ be the set of induced labelings of $\mathcal{H}$ on X. We have that $|\mathcal{H}(X)| \leqslant O\left(n^{d_{\mathcal{H}}}\right)$.*

Theorem B.3 (Exponentiated Gradient Descent Regret (see corollary 2.14 of [Shalev-Shwartz, 2012]))**.** *Let $\Lambda' = \left\{\boldsymbol{\lambda}' \in \mathbb{R}_+^d : ||\boldsymbol{\lambda}'||_1 = B\right\}$. Suppose the exponentiated gradient descent algorithm with learning rate η is run on the sequence of linear functions $\left\{f_t(\boldsymbol{\lambda}') = (\boldsymbol{\lambda}')^\top \boldsymbol{r}_t\right\}_{t=1}^T$ where $\boldsymbol{\lambda}' \in \Lambda'$ and $||\boldsymbol{r}_t||_\infty \leqslant L$ for all t. Let $\boldsymbol{\lambda}'_t$ denote the exponentiated gradient descent play at round t. We have that the regret of the algorithm over T rounds is:*

$$Regret_T(\Lambda') = \max_{\boldsymbol{\lambda}'\in\Lambda'} \sum_{t=1}^T (\boldsymbol{\lambda}')^\top \boldsymbol{r}_t - \sum_{t=1}^T (\boldsymbol{\lambda}'_t)^\top \boldsymbol{r}_t \leqslant \frac{B\log\left(d\right)}{\eta} + \eta B L^2 T$$

and that for $\eta = O\left(1/\sqrt{T}\right)$, we have that $Regret_T(\Lambda') = O\left(\sqrt{T}\right)$.

B.2 Algorithm Analysis and In-sample Guarantees

Proof of Lemma 3.1. Fix the set of observed individuals $X \in \mathcal{X}^n$. Recall that $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}_+^{2n} : ||\boldsymbol{\lambda}||_1 \leqslant B\}$ is the set of strategies for the Auditor. Now let $\Lambda' = \{\boldsymbol{\lambda}' \in \mathbb{R}_+^{2n+1} : ||\boldsymbol{\lambda}'||_1 = B\}$. Any $\boldsymbol{\lambda} \in \Lambda$ is associated with a $\boldsymbol{\lambda}' \in \Lambda'$ which is equal to $\boldsymbol{\lambda}$ on the first $2n$ coordinates and has the remaining mass on the last one. Let $\widehat{\boldsymbol{r}}_t = \boldsymbol{r}\left(\boldsymbol{h}_t|_{F_t}, \gamma_t; \widehat{\mathcal{Q}}_t\right)$ be the vector of fairness violations – estimated over only the m_0 problems F_t of round t – that the Auditor is using in Algorithm 2, and let $\widehat{\boldsymbol{r}}'_t \in \mathbb{R}^{2n+1}$ be equal to $\widehat{\boldsymbol{r}}_t$ on the first $2n$ coordinates and zero in the last one. We have that for any $\boldsymbol{\lambda} \in \Lambda$ and its associated $\boldsymbol{\lambda}' \in \Lambda'$, and in particular for $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}'_t$ of Algorithm 2, and all $t \in [T]$,

$$\boldsymbol{\lambda}^\top \widehat{\boldsymbol{r}}_t = (\boldsymbol{\lambda}')^\top \widehat{\boldsymbol{r}}'_t \quad , \quad \boldsymbol{\lambda}_t^\top \widehat{\boldsymbol{r}}_t = (\boldsymbol{\lambda}'_t)^\top \widehat{\boldsymbol{r}}'_t \tag{17}$$

Now by Theorem B.3, and using the observation that $||\widehat{\boldsymbol{r}}'_t||_\infty = ||\widehat{\boldsymbol{r}}_t||_\infty \leqslant 1 + 2\alpha$, we have that for any $\boldsymbol{\lambda}' \in \Lambda'$,

$$\sum_{t=1}^{T} (\boldsymbol{\lambda}')^\top \widehat{\boldsymbol{r}}'_t \leqslant \sum_{t=1}^{T} (\boldsymbol{\lambda}'_t)^\top \widehat{\boldsymbol{r}}'_t + \frac{B \log(2n+1)}{\eta} + \eta (1+2\alpha)^2 BT$$

Consequently by Equation (17), we have that for any $\boldsymbol{\lambda} \in \Lambda$,

$$\sum_{t=1}^{T} \boldsymbol{\lambda}^\top \widehat{\boldsymbol{r}}_t \leqslant \sum_{t=1}^{T} \boldsymbol{\lambda}_t^\top \widehat{\boldsymbol{r}}_t + \frac{B \log(2n+1)}{\eta} + \eta (1+2\alpha)^2 BT$$

Now let $\boldsymbol{r}_t = \boldsymbol{r}\left(\boldsymbol{h}_t, \gamma_t; \widehat{\mathcal{Q}}\right)$ be the vector of fairness violations estimated over all problems F and notice that the regret bound must be with respect to $\boldsymbol{r}_t$ and not $\widehat{\boldsymbol{r}}_t$. With that goal in mind, we have that for any $\boldsymbol{\lambda} \in \Lambda$,

$$\sum_{t=1}^{T} \boldsymbol{\lambda}^\top \boldsymbol{r}_t \leqslant \sum_{t=1}^{T} \boldsymbol{\lambda}_t^\top \boldsymbol{r}_t + \sum_{t=1}^{T} (\boldsymbol{\lambda}_t - \boldsymbol{\lambda})^\top (\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t) + \frac{B \log(2n+1)}{\eta} + \eta (1+2\alpha)^2 BT$$

We will use Chernoff-Hoeffding's inequality to bound the difference $(\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t)$ in ℓ_∞ norm. Let $\widehat{\psi}_t = \widehat{\psi}(X, \boldsymbol{w}_t)$ (see Mapping 3) where $\boldsymbol{w}_t$ is the vector of weights used in round t of the algorithm and observe that we can rewrite $\widehat{\boldsymbol{r}}_t$ and $\boldsymbol{r}_t$ in terms of $\widehat{\psi}_t$:

$$\widehat{\boldsymbol{r}}_t = \begin{bmatrix} \mathcal{E}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right) - \gamma_t - 2\alpha \\ \gamma_t - \mathcal{E}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right) - 2\alpha \end{bmatrix}_{i=1}^{n} \quad , \quad \boldsymbol{r}_t = \begin{bmatrix} \mathcal{E}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right) - \gamma_t - 2\alpha \\ \gamma_t - \mathcal{E}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right) - 2\alpha \end{bmatrix}_{i=1}^{n}$$

Hence, bounding the difference $(\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t)$ in ℓ_∞ norm involves bounding the terms

$$\left| \mathcal{E}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right) - \mathcal{E}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right) \right|$$

for all i. Notice we can now view the batch of problems F_t as independent draws from the distribution $\widehat{\mathcal{Q}}$. Therefore, it follows from the Chernoff-Hoeffding's Theorem B.1 that with probability at least $1 - \delta$ over the set of problems F, for any $\boldsymbol{\lambda} \in \Lambda$,

$$\sum_{t=1}^{T} (\boldsymbol{\lambda}_t - \boldsymbol{\lambda})^\top (\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t) \leqslant \sum_{t=1}^{T} ||\boldsymbol{\lambda}_t - \boldsymbol{\lambda}||_1 \cdot ||\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t||_\infty \leqslant B \sum_{t=1}^{T} ||\widehat{\boldsymbol{r}}_t - \boldsymbol{r}_t||_\infty \leqslant BT \sqrt{\frac{\log(2nT/\delta)}{2m_0}}$$

which implies with probability at least $1-\delta$ over the problems F, for any $\boldsymbol{\lambda} \in \Lambda$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}\right) - \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) \leqslant B\sqrt{\frac{\log\left(2nT/\delta\right)}{2m_0}} + \frac{B\log\left(2n+1\right)}{\eta T} + \eta\left(1+2\alpha\right)^2 B$$

completing the proof. □

Proof of Lemma 3.2. Let

$$R_{\boldsymbol{\lambda}} := B\sqrt{\frac{\log\left(2nT/\delta\right)}{2m_0}} + \frac{B\log\left(2n+1\right)}{\eta T} + \eta\left(1+2\alpha\right)^2 B$$

be the average regret of the Auditor. We have that for any $\boldsymbol{p} \in \Delta(\mathcal{H})^m$ and $\gamma \in [0,1]$,

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{p}, \gamma, \widehat{\boldsymbol{\lambda}}\right) &= \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}_t\right) \quad \text{(by linearity of } \mathcal{L}\text{)} \\
&\geqslant \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) \quad \text{((}\boldsymbol{h}_t, \gamma_t\text{) is Learner's Best Response)} \\
&\geqslant \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \widehat{\boldsymbol{\lambda}}\right) - R_{\boldsymbol{\lambda}} \quad \text{(w.p. } 1-\delta \text{ over } F \text{ by Lemma 3.1)} \\
&= \mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) - R_{\boldsymbol{\lambda}}
\end{aligned}$$

And that for any $\boldsymbol{\lambda} \in \Lambda$ we have:

$$\begin{aligned}
\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \boldsymbol{\lambda}\right) &= \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}\right) \quad \text{(by linearity of } \mathcal{L}\text{)} \\
&\leqslant \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) + R_{\boldsymbol{\lambda}} \quad \text{(w.p. } 1-\delta \text{ over } F \text{ by Lemma 3.1)} \\
&\leqslant \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \boldsymbol{\lambda}_t\right) + R_{\boldsymbol{\lambda}} \quad \text{((}\boldsymbol{h}_t, \gamma_t\text{) is Learner's Best Response)} \\
&= \mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) + R_{\boldsymbol{\lambda}}
\end{aligned}$$

Now we have to pick T and η of Algorithm 2 such that $R_{\boldsymbol{\lambda}} \leqslant \nu$ where ν is the approximation parameter input to the algorithm. First let $B = \frac{1+2\nu}{\alpha}$ and observe that if $m_0 \geqslant \frac{2B^2\log(2nT/\delta)}{\nu^2} = \frac{2(1+2\nu)^2\log(2nT/\delta)}{\nu^2\alpha^2}$ (see Assumption 3.1 on m_0) we have that

$$R_{\boldsymbol{\lambda}} \leqslant \frac{\nu}{2} + \frac{B\log\left(2n+1\right)}{\eta T} + \eta\left(1+2\alpha\right)^2 B$$

Next, let $T = \frac{16B^2(1+2\alpha)^2\log(2n+1)}{\nu^2}$, and $\eta = \frac{\nu}{4(1+2\alpha)^2 B}$ which makes the last two terms appearing in the RHS of the inequality to sum to $\nu/2$, and accordingly $R_{\boldsymbol{\lambda}} \leqslant \nu$. Therefore, we have shown

that for any X, with probability at least $1-\delta$ over the observed labelings F,

$$\begin{aligned}
\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) &\leqslant \mathcal{L}\left(\boldsymbol{p}, \gamma, \widehat{\boldsymbol{\lambda}}\right) + \nu \quad \text{for all } \boldsymbol{p} \in \Delta(\mathcal{H})^m, \gamma \in [0,1] \\
\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) &\geqslant \mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \boldsymbol{\lambda}\right) - \nu \quad \text{for all } \boldsymbol{\lambda} \in \Lambda
\end{aligned}$$

□

Proof of Theorem 3.3. Let $(\boldsymbol{p}, \gamma)$ be any feasible point of the constrained optimization problem (4) (note as discussed in the body of the paper there is at least one) and define $\boldsymbol{\lambda}^\star \in \Lambda$ to be

$$\boldsymbol{\lambda}^\star := \begin{cases} 0 & \text{if } r_{k^\star}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}; \widehat{\mathcal{Q}}\right) \leqslant 0 \\ Be_{k^\star} & \text{if } r_{k^\star}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}; \widehat{\mathcal{Q}}\right) > 0 \end{cases}$$

where $k^\star = \arg\max_{1 \leqslant k \leqslant 2n} r_k\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}; \widehat{\mathcal{Q}}\right)$. Observe that with probability $1-\delta$ over F,

$$\begin{aligned}
\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) &\leqslant \mathcal{L}\left(\boldsymbol{p}, \gamma, \widehat{\boldsymbol{\lambda}}\right) + \nu \quad \text{(by Lemma 3.2)} \\
&= \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + \widehat{\boldsymbol{\lambda}}^\top \boldsymbol{r}\left(\boldsymbol{p}, \gamma; \widehat{\mathcal{Q}}\right) + \nu \\
&\leqslant \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + \nu
\end{aligned}$$

and that,

$$\begin{aligned}
\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right) &\geqslant \mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \boldsymbol{\lambda}^\star\right) - \nu \quad \text{(by Lemma 3.2)} \\
&= \operatorname{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + (\boldsymbol{\lambda}^\star)^\top \boldsymbol{r}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}; \widehat{\mathcal{Q}}\right) - \nu \\
&\geqslant \operatorname{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) - \nu
\end{aligned}$$

Combining the above upper and lower bounds on $\mathcal{L}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}, \widehat{\boldsymbol{\lambda}}\right)$ implies

$$\operatorname{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \leqslant \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu$$

Now let's prove the bound on fairness violation. Once again using the above upper and lower bounds, we have that

$$(\boldsymbol{\lambda}^\star)^\top \boldsymbol{r}\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}; \widehat{\mathcal{Q}}\right) \leqslant \operatorname{err}\left(\boldsymbol{p}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) - \operatorname{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu \leqslant 1 + 2\nu$$

By definition of $\boldsymbol{\lambda}^\star$,

$$\max_{1 \leqslant k \leqslant 2n} r_k\left(\widehat{\boldsymbol{p}}, \widehat{\gamma}; \widehat{\mathcal{Q}}\right) \leqslant \frac{1+2\nu}{B}$$

which implies for all $1 \leqslant i \leqslant n$,

$$\left|\mathcal{E}\left(x_i, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| \leqslant 2\alpha + \frac{1+2\nu}{B}$$

And the proof is completed by the choice of $B = (1+2\nu)/\alpha$ in Algorithm 2. □

B.3 Proof of Theorem 3.4: Generalization over $\mathcal{P}$

When arguing about generalization over the distribution $\mathcal{P}$, we will have to come up with a *uniform convergence* for all *randomized* classifiers. It is usually the case in learning theory – for example when considering the concentration of classifiers' errors – that a uniform convergence for pure classifiers will immediately imply the uniform convergence for randomized classifiers without blowing up the sample complexity. However, in our setting and in particular for our notion of fairness, we will have to directly argue about the uniform convergence of randomized classifiers. Although there are possibly infinitely many of those randomized classifiers, we actually need to consider only the $\widetilde{O}\left(1/\alpha^2\right)$-sparse classifiers by which we mean the distributions over $\mathcal{H}$ with support of size at most $\widetilde{O}\left(1/\alpha^2\right)$. This $1/\alpha^2$ factor will accordingly show up in our final sample complexity bound for n.

Definition B.1 (Sparse Randomized Classifiers)**.** *We say a randomized classifier $p \in \Delta(\mathcal{H})$ is r-sparse if the support of p is of size at most r. We denote the set of all r-sparse randomized classifiers of the hypothesis class $\mathcal{H}$ by $\Delta_r(\mathcal{H})$, and the elements of $\Delta_r(\mathcal{H})$ by $p^{(r)}$.*

Lemma B.4 (Uniform Convergence of AIF Fairness Notion in $\Delta_r(\mathcal{H})^m$)**.** *Let $0 < \delta < 1$ and $r \in \mathbb{N}$ and let T be the number of iterations in Algorithm 2. Suppose*

$$n \geqslant \widetilde{O}\left(\frac{r\, m\, d_{\mathcal{H}} + \log\left(T/\delta\right)}{\beta^2}\right)$$

We have that for any set of problems F, with probability at least $1-\delta$ over the individuals X: for all $\boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m$ and all $\gamma \in \left\{0, \frac{1}{T}, \frac{2}{T}, \ldots, 1\right\}$,

$$\left| \mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left(\left|\mathcal{E}\left(x, \boldsymbol{p}^{(r)}; \widehat{\mathcal{Q}}\right) - \gamma\right| > \alpha\right) - \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}}\left(\left|\mathcal{E}\left(x, \boldsymbol{p}^{(r)}; \widehat{\mathcal{Q}}\right) - \gamma\right| > \alpha\right)\right| \leqslant \beta$$

Proof of Lemma B.4. The proof of this Lemma will use standard techniques for proving uniform convergence in learning theory such as the "two-sample trick" and Sauer's Lemma B.2. To simplify notation, let's call, for any $\boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m$ and any $\gamma \in S_T := \left\{0, \frac{1}{T}, \frac{2}{T}, \ldots, 1\right\}$,

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left(\left|\mathcal{E}\left(x, \boldsymbol{p}^{(r)}; \widehat{\mathcal{Q}}\right) - \gamma\right| > \alpha\right) := g\left(\boldsymbol{p}^{(r)}, \gamma; \mathcal{P}\right) \tag{18}$$

Define event $A(X)$ as follows:

$$A(X) = \left\{\exists\, \boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m, \gamma \in S_T : \left|g\left(\boldsymbol{p}^{(r)}, \gamma; \mathcal{P}\right) - g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_X\right)\right| > \beta\right\}$$

where $\widehat{\mathcal{P}}_X \equiv \widehat{\mathcal{P}}$ represents the uniform distribution over X. Our ultimate goal is to show that given the sample complexity for n stated in the lemma, $\mathbb{P}_X\left[A(X)\right]$ is small. Suppose besides the original data set X, we also have an "imaginary" data set of individuals $X' = \{x'_i\}_{i=1}^n \in \mathcal{X}^n$ sampled *i.i.d.* from the distribution $\mathcal{P}$. Define event $B(X, X')$ as follows:

$$B\left(X, X'\right) = \left\{\exists\, \boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m, \gamma \in S_T : \left|g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{X'}\right) - g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_X\right)\right| > \frac{\beta}{2}\right\}$$

Claim: $\mathbb{P}_X\left[A(X)\right] \leqslant 2\,\mathbb{P}_{(X,X')}\left[B(X, X')\right]$. Proof: It suffices to show that

$$\mathbb{P}_{(X,X')}\left[B(X, X') \,|\, A(X)\right] \geqslant 1/2$$

because

$$\mathbb{P}_{(X,X')}\left[B\left(X,X'\right)\right] \geqslant \mathbb{P}_{(X,X')}\left[B(X,X') \,|\, A(X)\right] \mathbb{P}_X\left[A\left(X\right)\right]$$

Let the event $A(X)$ hold and suppose the pair $\boldsymbol{p}_\star^{(r)}$ and $\gamma_\star$ satisfy

$$\left|g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \mathcal{P}\right) - g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \widehat{\mathcal{P}}_X\right)\right| > \beta$$

We have that by the triangle inequality and a Chernoff-Hoeffding's bound:

$$\begin{aligned}
&\mathbb{P}_{(X,X')}\left[B(X,X') \,|\, A(X)\right] \\
&\geqslant \mathbb{P}_{(X,X')}\left[\left|g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \widehat{\mathcal{P}}_{X'}\right) - g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \widehat{\mathcal{P}}_X\right)\right| > \frac{\beta}{2}\right] \\
&\geqslant \mathbb{P}_{(X,X')}\left[\left|g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \mathcal{P}\right) - g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \widehat{\mathcal{P}}_X\right)\right| - \left|g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \mathcal{P}\right) - g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \widehat{\mathcal{P}}_{X'}\right)\right| > \frac{\beta}{2}\right] \\
&\geqslant \mathbb{P}_{X'}\left[\left|g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \mathcal{P}\right) - g\left(\boldsymbol{p}_\star^{(r)}, \gamma_\star; \widehat{\mathcal{P}}_{X'}\right)\right| < \frac{\beta}{2}\right] \\
&\geqslant 1 - 2e^{-n\frac{\beta^2}{2}} \\
&\geqslant 1/2
\end{aligned}$$

Following the claim, it now suffices to show that $\mathbb{P}_{(X,X')}\left[B(X,X')\right]$ is small. Consider the following thought experiment: Let T and T' be two empty sets. For each $i \in [n]$ toss a fair coin independently and

- if it lands on Heads, put x_i in T and x_i' in T'.
- if it lands on Tails, put x_i' in T and x_i in T'.

We will later denote the randomness induced by these coin flips by "coin" in our probability statements. It follows immediately by our construction that the distribution of (T,T') is the same as the distribution of (X,X'), and therefore

$$\mathbb{P}_{(X,X')}\left[B(X,X')\right] = \mathbb{P}_{(T,T')}\left[B(T,T')\right] = \mathbb{P}_{(X,X',\text{coin})}\left[B(T,T')\right] \tag{19}$$

where

$$B(T,T') = \left\{\exists\, \boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m, \gamma \in S_T : \left|g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T'}\right) - g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_T\right)\right| > \frac{\beta}{2}\right\}$$

But we have that

$$\mathbb{P}_{(X,X',\text{coin})}\left[B(T,T')\right] = \mathbb{E}_{(X,X')}\left[\mathbb{P}_{\text{coin}}\left[B(T,T')\right]\right] \tag{20}$$

where we use the fact that the coin flips are independent of the random variables (X,X') and thus, conditioning on (X,X') won't change the distribution of "coin". Following Equations (19) and (20), it now suffices to show that for – any (X,X') – $\mathbb{P}_{\text{coin}}\left[B(T,T')\right]$ is small. Fix the data sets (X,X'). Fix $\boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m$ and $\gamma \in S_T$ and let

$$I = \left\{i \in [n] : g\left(\boldsymbol{p}^{(r)}, \gamma; x_i\right) \neq g\left(\boldsymbol{p}^{(r)}, \gamma; x_i'\right)\right\}$$

where recall that by Equation (18):

$$g\left(\boldsymbol{p}^{(r)}, \gamma; x_i\right) = \mathbb{1}\left[\left|\mathcal{E}\left(x_i, \boldsymbol{p}^{(r)}; \widehat{\mathcal{Q}}\right) - \gamma\right| > \alpha\right]$$

Let $|I| = n' \leqslant n$. Observe that

$$\mathbb{P}_{\text{coin}}\left[\left|g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T'}\right) - g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T}\right)\right| > \frac{\beta}{2}\right] = \mathbb{P}_{\text{coin}}\left[|\text{heads}(I) - \text{tails}(I)| > \frac{n\beta}{2}\right]$$

where heads(I) and tails(I) denote the number of heads and tails of the coin on indices I. But a triangle inequality followed by two Chernoff-Hoeffding bounds imply

$$\begin{aligned}
\mathbb{P}_{\text{coin}}\left[|\text{heads}(I) - \text{tails}(I)| > \frac{n\beta}{2}\right] &\leqslant \mathbb{P}_{\text{coin}}\left[\left|\frac{\text{heads}(I)}{n'} - \frac{1}{2}\right| > \frac{n\beta}{4n'}\right] \\
&\quad + \mathbb{P}_{\text{coin}}\left[\left|\frac{1}{2} - \frac{\text{tails}(I)}{n'}\right| > \frac{n\beta}{4n'}\right] \\
&\leqslant 4e^{-2n'(n\beta/4n')^2} \\
&\leqslant 4e^{-n\beta^2/8}
\end{aligned}$$

Therefore, we have proved that for any $\boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m$ and any $\gamma \in S_T$,

$$\mathbb{P}_{\text{coin}}\left[\left|g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T'}\right) - g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T}\right)\right| > \frac{\beta}{2}\right] \leqslant 4e^{-n\beta^2/8}$$

Now it's time to apply Sauer's Lemma B.2 to get a uniform convergence for all $\boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m$ and all $\gamma \in S_T$. Notice once the data sets (X, X') are fixed, there are at most $O\left((2n)^{d_{\mathcal{H}}\, r\, m}\right)$ number of randomized classifiers $\boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m$ induced on the set $\{X, X'\}$. Here $2n$ comes from the fact that X and X' have a combined number of $2n$ points. $d_{\mathcal{H}}$ is the VC dimension, and r and m show up in the bound because each $p_j^{(r)}$ in $\boldsymbol{p}^{(r)}$ is r-sparse and that there are a total of m randomized classifiers $p_j^{(r)}$ in $\boldsymbol{p}^{(r)}$. We also have that $|S_T| = T + 1$. Consequently by a union bound over all induced $\boldsymbol{p}^{(r)}$ and all $\gamma \in S_T$:

$$\begin{aligned}
\mathbb{P}_{\text{coin}}\left[B(T, T')\right] &= \mathbb{P}_{\text{coin}}\left[\exists\, \boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m, \gamma \in S_T : \left|g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T'}\right) - g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T}\right)\right| > \frac{\beta}{2}\right] \\
&\leqslant \sum_{\boldsymbol{p}^{(r)}, \gamma} \mathbb{P}_{\text{coin}}\left[\left|g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T'}\right) - g\left(\boldsymbol{p}^{(r)}, \gamma; \widehat{\mathcal{P}}_{T}\right)\right| > \frac{\beta}{2}\right] \\
&\leqslant O\left((2n)^{d_{\mathcal{H}}\, r\, m}\right) \cdot (T+1) \cdot 4e^{-n\beta^2/8}
\end{aligned}$$

where the second sum is actually over all the induced r-sparse randomized classifiers on the set $\{X, X'\}$ and all $\gamma \in S_T$. This shows so long as

$$n \geqslant \widetilde{O}\left(\frac{r\, m\, d_{\mathcal{H}} + \log\left(T/\delta\right)}{\beta^2}\right)$$

we have that for any (X, X'), $\mathbb{P}_{\text{coin}}\left[B(T, T')\right] \leqslant \delta/2$. Therefore

$$\mathbb{P}_X\left[A(X)\right] \leqslant 2\,\mathbb{P}_{(X,X')}\left[B\left(X, X'\right)\right] = 2\,\mathbb{E}_{(X,X')}\left[\mathbb{P}_{\text{coin}}\left[B(T, T')\right]\right] \leqslant \delta$$

completing the proof. □

Proof of Theorem 3.4. We will use Lemma B.4 in the proof of this theorem. We are interested in the fairness violation and accuracy of the mapping $\widehat{\psi} = \widehat{\psi}(X, F)$ returned by Algorithm 2 with respect to the distributions $\mathcal{P}$ and $\widehat{\mathcal{Q}}$. Fix the set of problems F and observe that when we work with the empirical distribution of the problems $\widehat{\mathcal{Q}}$, we have that

$$\mathcal{E}\left(x, \widehat{\psi}; \widehat{\mathcal{Q}}\right) = \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) \quad , \quad \operatorname{err}\left(\widehat{\psi}; \mathcal{P}, \widehat{\mathcal{Q}}\right) = \operatorname{err}\left(\widehat{\boldsymbol{p}}; \mathcal{P}, \widehat{\mathcal{Q}}\right)$$

where $\widehat{\boldsymbol{p}}$ is the set of m randomized classifiers output by Algorithm 2. Let's first prove the generalization for fairness. Define event $A(X)$:

$$A(X) = \left\{ \mathop{\mathbb{P}}_{x \sim \mathcal{P}} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma} \right| > 5\alpha \right) - \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma} \right| > 3\alpha \right) > \beta \right\}$$

We will eventually show that under the stated sample complexity for n in the theorem, $\mathbb{P}_X[A(X)]$ is small. Consider the distributions $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_m)$ over $\mathcal{H}$. Let

$$r = \frac{\log(12nm/\delta)}{2\alpha^2}$$

For each $j \in [m]$, consider drawing r independent samples from the distribution $\widehat{p}_j$ and define $\widehat{p}_j^{(r)}$ to be the uniform distribution over the drawn samples. We will abuse notation and use $\widehat{p}_j^{(r)}$ to denote both the drawn samples and the uniform distribution over them. Now define

$$\widehat{\boldsymbol{p}}^{(r)} = \left(\widehat{p}_1^{(r)}, \widehat{p}_2^{(r)}, \ldots, \widehat{p}_m^{(r)}\right) \in \Delta_r(\mathcal{H})^m$$

which is the "r-sparsified" version $\widehat{\boldsymbol{p}}$. One important observation that follows from the Chernoff-Hoeffding's theorem is that for – any – $X' = \{x_i'\}_{i=1}^n \in \mathcal{X}^n$, with probability at least $1 - \delta/6$ over the draws of $\widehat{\boldsymbol{p}}^{(r)}$, we have that for all $i \in [n]$:

$$\left| \mathcal{E}\left(x_i', \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x_i', \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) \right| \leqslant \alpha \tag{21}$$

In other words, for any set of individuals $X' \in \mathcal{X}^n$ of size n, we have that:

$$\mathbb{P}_{\widehat{\boldsymbol{p}}^{(r)}} \left[\mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}'} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) \right| > \alpha \right) \neq 0 \right] \leqslant \delta/6 \tag{22}$$

where $\widehat{\mathcal{P}}'$ denotes the uniform distribution over X'. Now define events B, C, D as follows:

$$B\left(X, \widehat{\boldsymbol{p}}^{(r)}\right) = \left\{ \left| \mathop{\mathbb{P}}_{x \sim \mathcal{P}} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma} \right| > 4\alpha \right) - \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma} \right| > 4\alpha \right) \right| > \frac{\beta}{2} \right\}$$

$$C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right) = \left\{ \mathop{\mathbb{P}}_{x \sim \mathcal{P}} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) \right| > \alpha \right) > \frac{\beta}{2} \right\}$$

$$D\left(X, \widehat{\boldsymbol{p}}^{(r)}\right) = \left\{ \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) \right| > \alpha \right) \neq 0 \right\}$$

It follows by the triangle inequality that

$$\begin{aligned} \mathbb{P}_X[A(X)] &= \mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}[A(X)] \\ &\leqslant \underbrace{\mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}\left[B\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right]}_{\text{term 1}} + \underbrace{\mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}\left[C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right]}_{\text{term 2}} + \underbrace{\mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}\left[D\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right]}_{\text{term 3}} \end{aligned}$$

So to prove that $\mathbb{P}_X[A(X)]$ is small, it suffices to show that all the three terms appearing in the RHS of the above inequality are small. We will in fact show each term $\leqslant \delta/3$:

- term 1: Notice to bound this term we need to prove a *uniform convergence* for all r-sparse set of randomized classifiers $\boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m$ and all γ of the form c/T for some nonnegative integer $c \leqslant T$ (because the $\widehat{\gamma}$ output by our algorithm has this form). But we have already proved this uniform convergence in Lemma B.4. In fact if we define the event:

$$B_1(X) = \Big\{\exists\, \boldsymbol{p}^{(r)} \in \Delta_r(\mathcal{H})^m, \gamma \in \{0, \frac{1}{T}, \frac{2}{T}, \ldots, 1\} : \\ \left| \mathop{\mathbb{P}}_{x \sim \mathcal{P}} \left(\left| \mathcal{E}\left(x, \boldsymbol{p}^{(r)}; \widehat{\mathcal{Q}}\right) - \gamma \right| > 4\alpha \right) - \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}} \left(\left| \mathcal{E}\left(x, \boldsymbol{p}^{(r)}; \widehat{\mathcal{Q}}\right) - \gamma \right| > 4\alpha \right) \right| > \frac{\beta}{2} \Big\}$$

by Lemma B.4, so long as,

$$n \geqslant \widetilde{O}\left(\frac{r\, m\, d_{\mathcal{H}} + \log\left(T/\delta\right)}{\beta^2}\right) \equiv \widetilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right) \quad \text{(Recall: } T = O\left(\log\left(n\right)/\nu^2\alpha^2\right))$$

we have that $\mathbb{P}_X\left[B_1(X)\right] \leqslant \delta/3$. And this implies

$$\mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}\left[B\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right] \leqslant \mathbb{P}_X\left[B_1\left(X\right)\right] \leqslant \delta/3$$

- term 2: Observe that

$$\mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}\left[C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right] = \mathbb{E}_X\left[\mathbb{P}_{\widehat{\boldsymbol{p}}^{(r)}}\left[C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right]\right]$$

Notice when we condition on X, the conditional distribution of both $\widehat{\boldsymbol{p}}$ and $\widehat{\boldsymbol{p}}^{(r)}$ will actually change because they both depend on X, but it is the case that for any X, $\widehat{\boldsymbol{p}}^{(r)}$ (given X) will be still independent draws from $\widehat{\boldsymbol{p}}$ (given X) and this is what we actually need. We will show for any X:

$$\mathbb{P}_{\widehat{\boldsymbol{p}}^{(r)}}\left[C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right] \leqslant \delta/3$$

Fix X. Let $X' = \{x_i'\}_{i=1}^n$ be a *new* data set of individuals drawn independently from the distribution $\mathcal{P}$ and let $\widehat{\mathcal{P}}'$ denote the uniform distribution over X'. Define the events:

$$C_1\left(X', \widehat{\boldsymbol{p}}^{(r)}\right) = \\ \left\{ \left| \mathop{\mathbb{P}}_{x \sim \mathcal{P}} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) \right| > \alpha \right) - \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}'} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) \right| > \alpha \right) \right| > \frac{\beta}{2} \right\}$$

$$C_2\left(X', \widehat{\boldsymbol{p}}^{(r)}\right) = \left\{ \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}'} \left(\left| \mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x, \widehat{\boldsymbol{p}}^{(r)}; \widehat{\mathcal{Q}}\right) \right| > \alpha \right) \neq 0 \right\}$$

It follows by a triangle inequality that:

$$\begin{aligned} \mathbb{P}_{\widehat{\boldsymbol{p}}^{(r)}}\left[C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right] &= \mathbb{P}_{X', \widehat{\boldsymbol{p}}^{(r)}}\left[C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right] \\ &\leqslant \mathbb{P}_{X', \widehat{\boldsymbol{p}}^{(r)}}\left[C_1\left(X', \widehat{\boldsymbol{p}}^{(r)}\right)\right] + \mathbb{P}_{X', \widehat{\boldsymbol{p}}^{(r)}}\left[C_2\left(X', \widehat{\boldsymbol{p}}^{(r)}\right)\right] \\ &= \mathbb{E}_{\widehat{\boldsymbol{p}}^{(r)}}\left[\mathbb{P}_{X'}\left[C_1\left(X', \widehat{\boldsymbol{p}}^{(r)}\right)\right]\right] + \mathbb{E}_{X'}\left[\mathbb{P}_{\widehat{\boldsymbol{p}}^{(r)}}\left[C_2\left(X', \widehat{\boldsymbol{p}}^{(r)}\right)\right]\right] \end{aligned}$$

Given $\widehat{\boldsymbol{p}}^{(r)}$, we have by a Chernoff-Hoeffding's inequality that as long as the sample complexity for n is met:

$$\mathbb{P}_{X'}\left[C_1\left(X', \widehat{\boldsymbol{p}}^{(r)}\right)\right] \leqslant \delta/6$$

On the other hand, given X', it follows by Equation (22) that,

$$\mathbb{P}_{\widehat{\boldsymbol{p}}^{(r)}}\left[C_2\left(X', \widehat{\boldsymbol{p}}^{(r)}\right)\right] \leqslant \delta/6$$

Hence:

$$\mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}\left[C\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right] \leqslant \delta/3$$

- term 3: Equation (22) implies:

$$\mathbb{P}_{X, \widehat{\boldsymbol{p}}^{(r)}}\left[D\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right] = \mathbb{E}_X\left[\mathbb{P}_{\widehat{\boldsymbol{p}}^{(r)}}\left[D\left(X, \widehat{\boldsymbol{p}}^{(r)}\right)\right]\right] \leqslant \delta/3$$

We finally proved $\mathbb{P}_X\left[A\left(X\right)\right] \leqslant \delta$. In other words, we have proved that for any F, with probability at least $1-\delta$ over the individuals X,

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left(\left|\mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| > 5\alpha\right) - \mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}}\left(\left|\mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| > 3\alpha\right) \leqslant \beta$$

On the other hand, the in-sample guarantees provided in Theorem 3.3 implies for any X, with probability $1-\delta$ over the observed problems F, the pair $(\widehat{\boldsymbol{p}}, \widehat{\gamma})$ of Algorithm 2 satisfies

$$\mathop{\mathbb{P}}_{x \sim \widehat{\mathcal{P}}}\left(\left|\mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| > 3\alpha\right) = 0$$

Hence, as long as

$$n \geqslant \widetilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right) \tag{23}$$

we have that with probability at least $1-2\delta$ over the observed data set (X, F),

$$\mathop{\mathbb{P}}_{x \sim \mathcal{P}}\left(\left|\mathcal{E}\left(x, \widehat{\boldsymbol{p}}; \widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| > 5\alpha\right) \leqslant \beta$$

which shows $\widehat{\boldsymbol{p}}$, or equivalently the mapping $\widehat{\psi}$, satisfies $(5\alpha, \beta)$-AIF with respect to the distributions $\mathcal{P}$ and $\widehat{\mathcal{Q}}$. Now let's look at the generalization (with respect to $\mathcal{P}$) error of $\widehat{\boldsymbol{p}}$. Let

$$\psi^\star = \psi^\star\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right), \quad \gamma^\star = \gamma^\star\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right), \quad \mathrm{err}\left(\psi^\star; \mathcal{P}, \widehat{\mathcal{Q}}\right) = \mathrm{OPT}\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right)$$

be defined as in Definition 3.1. Let also $\psi^\star|_F = \boldsymbol{p}^\star \in \Delta(\mathcal{H})^m$. Since we are working with the empirical distribution of problems $\widehat{\mathcal{Q}}$,

$$\mathrm{err}\left(\boldsymbol{p}^\star; \mathcal{P}, \widehat{\mathcal{Q}}\right) = \mathrm{err}\left(\psi^\star; \mathcal{P}, \widehat{\mathcal{Q}}\right) = \mathrm{OPT}\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right)$$

We would like to compare the accuracy of $\widehat{\boldsymbol{p}}$ with respect to the distributions $\mathcal{P}$ and $\widehat{\mathcal{Q}}$, i.e. $\mathrm{err}\left(\widehat{\boldsymbol{p}}; \mathcal{P}, \widehat{\mathcal{Q}}\right)$, to $\mathrm{OPT}\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right)$ as a benchmark. We establish this comparison in three steps:

1. a bound on the difference between $\mathrm{err}\left(\widehat{\boldsymbol{p}}; \mathcal{P}, \widehat{\mathcal{Q}}\right)$ and its in-sample version $\mathrm{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)$. Note a uniform convergence can be achieved for all $\boldsymbol{p} \in \Delta(\mathcal{H})^m$ by standard generalization techniques (Sauer's Lemma and the two-sample trick, as discussed in the proof of Lemma B.4), without even needing to consider the uniform convergence for r-sparse randomized

classifiers. Because when investigating the concentration of the overall error rate which is in the form of nested expectations, the linearity of expectation helps us to directly turn a uniform convergence for all sets of pure classifiers $\boldsymbol{h} \in \mathcal{H}^m$ into a uniform convergence for all randomized classifiers $\boldsymbol{p} \in \Delta(\mathcal{H})^m$ without blowing up the sample complexity by a factor of $1/\alpha^2$. We can therefore guarantee that for all F, with probability at least $1-\delta$ over X, as long as the sample complexity (23) is satisfied:

$$\left|\operatorname{err}\left(\widehat{\boldsymbol{p}}; \mathcal{P}, \widehat{\mathcal{Q}}\right) - \operatorname{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right| \leqslant O(\alpha\beta) \tag{24}$$

2. a bound on the difference between $\operatorname{err}\left(\boldsymbol{p}^\star; \mathcal{P}, \widehat{\mathcal{Q}}\right)$ and its in-sample version $\operatorname{err}\left(\boldsymbol{p}^\star; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)$. This is very simple and it follows immediately from the Chernoff-Hoeffding's theorem that for all F, with probability at least $1-\delta$ over X, as long as (23) is satisfied:

$$\left|\operatorname{OPT}\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right) - \operatorname{err}\left(\boldsymbol{p}^\star; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right| \leqslant O(\alpha\beta)$$

3. the pair $(\boldsymbol{p}^\star, \gamma^\star)$ is feasible in the empirical learning problem (4). As a consequence, Theorem 3.3 implies for any X, with probability $1-\delta$ over F:

$$\operatorname{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \leqslant \operatorname{err}\left(\boldsymbol{p}^\star; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu$$

Finally, putting together all three pieces explained above, we have that with probability at least $1-3\delta$ over the observed data set (X, F),

$$\operatorname{err}\left(\widehat{\boldsymbol{p}}; \mathcal{P}, \widehat{\mathcal{Q}}\right) \leqslant \operatorname{OPT}\left(\alpha; \mathcal{P}, \widehat{\mathcal{Q}}\right) + O(\nu) + O(\alpha\beta)$$

completing the proof. □

B.4 Proof of Theorem 3.5: Generalization over $\mathcal{Q}$

Proof of Theorem 3.5. Fix the set of observed individuals $X = \{x_i\}_{i=1}^n$ and recall $\widehat{\mathcal{P}}$ denotes the uniform distribution over X. Let

$$\widehat{\psi} = \widehat{\psi}\left(X, \widehat{W}\right) \quad \text{where} \quad \widehat{W} = \{\boldsymbol{w}_t\}_{t=1}^T$$

be the mapping returned by Algorithm 2. We would like to lift the accuracy and fairness guarantees of $\widehat{\psi}$ from the empirical distribution $\widehat{\mathcal{Q}}$ up to the true underlying distribution $\mathcal{Q}$. First observe that by the definition of $\widehat{\psi}$ (Mapping 3), for any $x \in \mathcal{X}$ and any distribution Q:

$$\mathcal{E}\left(x, \widehat{\psi}; \mathcal{Q}\right) = \frac{1}{T}\sum_{t=1}^T \mathcal{E}\left(x, \widehat{\psi}_t; \mathcal{Q}\right) \tag{25}$$

where $\widehat{\psi}_t = \widehat{\psi}(X, \boldsymbol{w}_t)$ is a mapping defined only by the weights $\boldsymbol{w}_t$ of round t over the individuals X. $\widehat{\psi}_t$ in fact takes a function f, solves one CSC problem on X weighted by $\boldsymbol{w}_t$ and then returns the learned classifier. For any $t \in [T]$, we are interested in bounding the difference

$$\left|\mathcal{E}\left(x, \widehat{\psi}_t; \mathcal{Q}\right) - \mathcal{E}\left(x, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right)\right|$$

Notice following the dynamics of our algorithm, $\boldsymbol{w}_t$ (and accordingly $\widehat{\psi}_t$) has dependence on the batches of problems $\{F_{t'}\}_{t'=1}^{t-1}$ and therefore we cannot directly invoke the Chernoff-Hoeffding's theorem to bound the above difference. We instead use the estimates $\mathcal{E}(x, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t)$ where $\widehat{\mathcal{Q}}_t = \mathcal{U}(F_t)$ as an intermediary step to bound the above difference. Observe that a simple triangle inequality implies

$$\left|\mathcal{E}\left(x, \widehat{\psi}_t; \mathcal{Q}\right) - \mathcal{E}\left(x, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right)\right| \leqslant \left|\mathcal{E}\left(x, \widehat{\psi}_t; \mathcal{Q}\right) - \mathcal{E}\left(x, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right)\right| + \left|\mathcal{E}\left(x, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right) - \mathcal{E}\left(x, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right)\right|$$

Now we can invoke the Chernoff-Hoeffding's Theorem B.1 to bound each term appearing above separately. The batch of problems F_t can be seen as independent draws from both distributions $\mathcal{Q}$ and $\widehat{\mathcal{Q}}$. It therefore follows that with probability $1-\delta$ over the draws of the batch F_t,

$$\left|\mathcal{E}\left(x, \widehat{\psi}_t; \mathcal{Q}\right) - \mathcal{E}\left(x, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right)\right| \leqslant 2\sqrt{\frac{\log\left(4/\delta\right)}{2m_0}}$$

We now use Equation (25) to translate this bound to a bound for $\widehat{\psi}$. We have that for any $x \in \mathcal{X}$, with probability at least $1-\delta$ over the draws of all problems F,

$$\left|\mathcal{E}\left(x, \widehat{\psi}; \mathcal{Q}\right) - \mathcal{E}\left(x, \widehat{\psi}; \widehat{\mathcal{Q}}\right)\right| \leqslant 2\sqrt{\frac{\log\left(4T/\delta\right)}{2m_0}} \tag{26}$$

And therefore, with probability at least $1-\delta$ over all problems F, for all $i \in [n]$,

$$\left|\mathcal{E}\left(x_i, \widehat{\psi}; \mathcal{Q}\right) - \mathcal{E}\left(x_i, \widehat{\psi}; \widehat{\mathcal{Q}}\right)\right| \leqslant 2\sqrt{\frac{\log\left(4nT/\delta\right)}{2m_0}} \leqslant \nu\alpha \tag{27}$$

where the second inequality follows from Assumption 3.1 on m_0. Now by a simple triangle inequality

$$\begin{aligned}
&\mathop{\mathbb{P}}_{x\sim\widehat{\mathcal{P}}}\left(\left|\mathcal{E}\left(x, \widehat{\psi}; \mathcal{Q}\right) - \widehat{\gamma}\right| > 4\alpha\right) \\
&\leqslant \mathop{\mathbb{P}}_{x\sim\widehat{\mathcal{P}}}\left(\left|\mathcal{E}\left(x, \widehat{\psi}; \mathcal{Q}\right) - \mathcal{E}(x, \widehat{\psi}; \widehat{\mathcal{Q}})\right| > \alpha\right) + \mathop{\mathbb{P}}_{x\sim\widehat{\mathcal{P}}}\left(\left|\mathcal{E}(x, \widehat{\psi}; \widehat{\mathcal{Q}}) - \widehat{\gamma}\right| > 3\alpha\right)
\end{aligned}$$

The first term appearing in the RHS of the above inequality is zero with probability $1-\delta$ over F following Inequality (27). The second term is zero with probability $1-\delta$ over F as well following the in-sample guarantees provided in Theorem 3.3. Therefore, we conclude that for any set of observed individuals X, with probability at least $1-2\delta$ over the observed labelings F, as long as Assumption 3.1:

$$m = T \cdot m_0 \geqslant O\left(\frac{T\log\left(nT/\delta\right)}{\nu^2\alpha^2}\right) = \widetilde{O}\left(\frac{\log\left(n\right)\log\left(n/\delta\right)}{\nu^4\alpha^4}\right) \tag{28}$$

holds, the mapping $\widehat{\psi}$ satisfies $(4\alpha, 0)$-AIF with respect to the distributions $\left(\widehat{\mathcal{P}}, \mathcal{Q}\right)$:

$$\mathop{\mathbb{P}}_{x\sim\widehat{\mathcal{P}}}\left(\left|\mathcal{E}\left(x, \widehat{\psi}; \mathcal{Q}\right) - \widehat{\gamma}\right| > 4\alpha\right) = 0$$

Now let's look at the generalization (with respect to $\mathcal{Q}$) error of $\widehat{\psi}$. Let

$$\psi^\star = \psi^\star\left(\alpha; \widehat{\mathcal{P}}, \mathcal{Q}\right), \quad \gamma^\star = \gamma^\star\left(\alpha; \widehat{\mathcal{P}}, \mathcal{Q}\right), \quad \text{err}\left(\psi^\star; \widehat{\mathcal{P}}, \mathcal{Q}\right) = \text{OPT}\left(\alpha; \widehat{\mathcal{P}}, \mathcal{Q}\right)$$

be defined as in Definition 3.1. We would like to compare the accuracy of $\widehat{\psi}$ with respect to the distributions $\widehat{\mathcal{P}}$ and $\mathcal{Q}$, i.e. $\text{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \mathcal{Q}\right)$, to $\text{OPT}\left(\alpha; \widehat{\mathcal{P}}, \mathcal{Q}\right)$ as a benchmark. We follow the same approach we used in the proof of Theorem 3.4 to achieve an upper bound for the difference between $\widehat{\psi}$'s error and the optimal error. It follows directly from Inequality (27) that with probability $1-\delta$ over the problems F, as long as the sample complexity (28) is satisfied,

$$\left|\text{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \mathcal{Q}\right) - \text{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\mathcal{E}\left(x_i, \widehat{\psi}; \mathcal{Q}\right) - \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}\left(x_i, \widehat{\psi}; \widehat{\mathcal{Q}}\right)\right| \leqslant \nu\alpha \tag{29}$$

and that by a simple application of the Chernoff-Hoeffding's inequality, as long as the sample complexity (28) is satisfied, with probability $1-\delta$ over F we have that

$$\left|\text{err}\left(\psi^\star; \widehat{\mathcal{P}}, \mathcal{Q}\right) - \text{err}\left(\psi^\star; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right| \leqslant O\left(\nu\alpha\right) \tag{30}$$

On the other hand, with probability $1-\delta$ over F, the pair $(\psi^\star|_F, \gamma^\star)$ – where $\psi^\star|_F$ is the mapping $\psi^\star$ restricted to the problems F – is feasible in the empirical learning problem (4). Because, with probability $1-\delta$ as long as the sample complexity (28) is satisfied:

$$\left|\mathcal{E}\left(x_i, \psi^\star|_F; \widehat{\mathcal{Q}}\right) - \gamma^\star\right| \leqslant \left|\mathcal{E}\left(x_i, \psi^\star|_F; \widehat{\mathcal{Q}}\right) - \mathcal{E}\left(x_i, \psi^\star; \mathcal{Q}\right)\right| + \left|\mathcal{E}\left(x_i, \psi^\star; \mathcal{Q}\right) - \gamma^\star\right| \leqslant \alpha + \alpha = 2\alpha$$

Therefore, by Theorem 3.3, we have that with probability $1-\delta$ over F,

$$\text{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) = \text{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \leqslant \text{err}\left(\psi^\star|_F; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu = \text{err}\left(\psi^\star; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu \tag{31}$$

Putting together Inequalities (29), (30), and (31), we conclude that for any X, with probability at least $1-4\delta$ over the problems F,

$$\text{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \mathcal{Q}\right) \leqslant \text{OPT}\left(\alpha; \widehat{\mathcal{P}}, \mathcal{Q}\right) + O\left(\nu\right)$$

□

B.5 Proof of Theorem 3.6: Simultaneous Generalization over $\mathcal{P}$ and $\mathcal{Q}$

We have proved so far:

- Theorem 3.4 (Generalization over $\mathcal{P}$): $\left(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \xrightarrow{\text{lifted}} \left(\mathcal{P}, \widehat{\mathcal{Q}}\right)$.
- Theorem 3.5 (Generalization over $\mathcal{Q}$): $\left(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \xrightarrow{\text{lifted}} \left(\widehat{\mathcal{P}}, \mathcal{Q}\right)$

Before we prove Theorem 3.6 which considers simultaneous generalization over both distributions $\mathcal{P}$ and $\mathcal{Q}$, we first prove the following Lemma B.5. Notice when proving the final generalization theorem, we take the following natural two steps in lifting the guarantees:

$$\left(\widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \to \left(\mathcal{P}, \widehat{\mathcal{Q}}\right) \to \left(\mathcal{P}, \mathcal{Q}\right)$$

The first step of the above scheme is exactly what we have proved in Theorem 3.4. However, we cannot directly invoke Theorem 3.5 to prove the second step since we had the empirical distribution of individuals (i.e. $\widehat{\mathcal{P}}$) in that theorem. In the following Lemma, we basically prove the second step where the distribution over individuals is $\mathcal{P}$.

Lemma B.5. *Suppose*

$$n \geqslant \widetilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right) \quad , \quad m \geqslant \widetilde{O}\left(\frac{\log\left(n\right)\log\left(n/\delta\right)}{\nu^4\alpha^4}\right)$$

and let $\widehat{\psi} = \widehat{\psi}\left(X, \widehat{W}\right)$ *be the output of Algorithm 2. We have that for any* X*, with probability at least* $1-2\delta$ *over the problems* F*,*

$$\mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| > \alpha\right) \leqslant \beta$$

and that

$$\left|err\left(\widehat{\psi};\mathcal{P},\mathcal{Q}\right) - err\left(\widehat{\psi};\mathcal{P},\widehat{\mathcal{Q}}\right)\right| \leqslant O\left(\alpha\beta\right) + O\left(\nu\alpha\right)$$

Proof. Let's prove the first part of the Lemma. Let $X' = \{x_i'\}_{i=1}^n \subseteq \mathcal{X}^n$ be – any – set of n individuals. Recall that Inequality (26) developed in the proof of Theorem 3.5 implies with probability $1-\delta$ over the problems F, for all $i \in [n]$:

$$\left|\mathcal{E}\left(x_i',\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x_i',\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| \leqslant 2\sqrt{\frac{\log\left(4nT/\delta\right)}{2m_0}} \leqslant \nu\alpha$$

In other words, if $\widehat{\mathcal{P}}'$ represents the uniform distribution over X', we have that for any $X' \subseteq \mathcal{X}^n$, with probability at least $1-\delta$ over F,

$$\mathop{\mathbb{P}}_{x\sim\widehat{\mathcal{P}}'}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| > \alpha\right) = 0 \tag{32}$$

This doesn't give us what we want because this inequality works for any distribution with support of size at most n while we want a guarantee for any $x \sim \mathcal{P}$ and that $\mathcal{P}$ might have an infinite-sized support. For random variables F, consider the event $A\left(F\right)$:

$$A\left(F\right) = \left\{\mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| > \alpha\right) > \beta\right\}$$

We eventually want to show that $\mathbb{P}_F\left[A\left(F\right)\right]$ is small. Following the guarantee of (32) we consider an auxiliary data set of individuals X' to argue that $\mathbb{P}_F\left[A\left(F\right)\right]$ is in fact small. To formalize our argument, let $X' = \{x_i'\}_{i=1}^n \sim \mathcal{P}^n$ be a *new* set of n individuals drawn independently from the distribution $\mathcal{P}$ and let $\widehat{\mathcal{P}}'$ denote the uniform distribution over X'. Define events $B\left(F,X'\right)$ and $C\left(F,X'\right)$ which depend on both F and X':

$$B\left(F,X'\right) = \left\{\left|\mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| > \alpha\right) - \mathop{\mathbb{P}}_{x\sim\widehat{\mathcal{P}}'}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| > \alpha\right)\right| > \beta\right\}$$

$$C\left(F,X'\right) = \left\{\mathop{\mathbb{P}}_{x\sim\widehat{\mathcal{P}}'}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| > \alpha\right) \neq 0\right\}$$

We have that

$$\begin{aligned}\mathbb{P}_F\left[A\left(F\right)\right] &= \mathbb{P}_{F,X'}\left[A\left(F\right)\right] \\ &\leqslant \mathbb{P}_{F,X'}\left[B\left(F,X'\right)\right] + \mathbb{P}_{F,X'}\left[C\left(F,X'\right)\right] \quad \text{(follows from a triangle inequality)} \\ &= \mathbb{E}_F\left[\mathbb{P}_{X'}\left[B\left(F,X'\right)\right]\right] + \mathbb{E}_{X'}\left[\mathbb{P}_F\left[C\left(F,X'\right)\right]\right]\end{aligned}$$

But Inequality (32) implies for any X', $\mathbb{P}_F\left[C\left(F,X'\right)\right] \leqslant \delta$. And that by a simple application of the Chernoff-Hoeffding's inequality, as long as the sample complexity for n stated in the Lemma is met, we have that for any F, $\mathbb{P}_{X'}\left[B\left(F,X'\right)\right] \leqslant \delta$. We have therefore shown that

$$\mathbb{P}_F\left[A\left(F\right)\right] \leqslant 2\delta$$

proving the first part of the Lemma. The second part of the Lemma can be proved similarly with the same idea of considering an auxiliary data set $X' \sim \mathcal{P}^n$. Skipping the details, we can in fact show that with probability $1-2\delta$ over F,

$$\begin{aligned}&\left|\operatorname{err}\left(\widehat{\psi};\mathcal{P},\mathcal{Q}\right) - \operatorname{err}\left(\widehat{\psi};\mathcal{P},\widehat{\mathcal{Q}}\right)\right| \\ &\leqslant \mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| \\ &\leqslant \left|\mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{P}}'}\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right|\right| \\ &+ \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{P}}'}\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| \\ &\leqslant O\left(\alpha\beta\right) + O\left(\nu\alpha\right)\end{aligned}$$

□

Proof of Theorem 3.6. First, observe that by a traingle inequality

$$\begin{aligned}&\mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \widehat{\gamma}\right| > 6\alpha\right) \\ &\leqslant \mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right)\right| > \alpha\right) + \mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\widehat{\mathcal{Q}}\right) - \widehat{\gamma}\right| > 5\alpha\right)\end{aligned}$$

We will argue that as long as,

$$n \geqslant \widetilde{O}\left(\frac{m\, d_{\mathcal{H}} + \log\left(1/\nu^2\delta\right)}{\alpha^2\beta^2}\right) \quad , \quad m \geqslant \widetilde{O}\left(\frac{\log\left(n\right)\log\left(n/\delta\right)}{\nu^4\alpha^4}\right)$$

we can further bound the above inequality by 2β. Lemma B.5 implies for any X, the first term appearing in the RHS of the above inequality is bounded by β with probability $1-2\delta$ over the problems F. The second term is bounded by β as well by Theorem 3.4 with probability $1-3\delta$ over the observed individuals and problems (X,F) (in the theorem we actually state w.p. $1-6\delta$ because 3δ comes from the accuracy guarantee). Putting these two pieces together, we have that with probability at least $1-5\delta$ over the data set (X,F),

$$\mathop{\mathbb{P}}_{x\sim\mathcal{P}}\left(\left|\mathcal{E}\left(x,\widehat{\psi};\mathcal{Q}\right) - \widehat{\gamma}\right| > 6\alpha\right) \leqslant 2\beta$$

which implies $\widehat{\psi}$ satisfies $(6\alpha, 2\beta)$-AIF with respect to the distributions $(\mathcal{P}, \mathcal{Q})$. Now let's look at the generalization error of $\widehat{\psi}$. Let

$$\psi^\star = \psi^\star(\alpha; \mathcal{P}, \mathcal{Q}), \quad \gamma^\star = \gamma^\star(\alpha; \mathcal{P}, \mathcal{Q}), \quad \mathrm{err}(\psi^\star; \mathcal{P}, \mathcal{Q}) = \mathrm{OPT}(\alpha; \mathcal{P}, \mathcal{Q})$$

be defined as in Definition 3.1. We will again use the triangle inequality to write:

$$\begin{aligned}\left|\mathrm{err}\left(\widehat{\psi}; \mathcal{P}, \mathcal{Q}\right) - \mathrm{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right| \leqslant &\left|\mathrm{err}\left(\widehat{\psi}; \mathcal{P}, \mathcal{Q}\right) - \mathrm{err}\left(\widehat{\psi}; \mathcal{P}, \widehat{\mathcal{Q}}\right)\right| \\ &+ \left|\mathrm{err}\left(\widehat{\psi}; \mathcal{P}, \widehat{\mathcal{Q}}\right) - \mathrm{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right|\end{aligned}$$

It follows from Lemma B.5 that the first term appearing in the RHS is bounded by $O(\alpha\beta) + O(\nu\alpha)$ with probability $1 - 2\delta$ over the problems F, and we also have that the second term is bounded by $O(\alpha\beta)$ as well with probability $1 - \delta$ over the individuals X (see (24)). Therefore, with probability $1 - 3\delta$ over the data set (X, F),

$$\left|\mathrm{err}\left(\widehat{\psi}; \mathcal{P}, \mathcal{Q}\right) - \mathrm{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right| \leqslant O(\alpha\beta) + O(\nu\alpha)$$

It follows similarly that with probability $1 - 2\delta$ over X and F,

$$\left|\mathrm{err}(\psi^\star; \mathcal{P}, \mathcal{Q}) - \mathrm{err}\left(\psi^\star; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right)\right| \leqslant O(\alpha\beta) + O(\nu\alpha)$$

and as in the proof of Theorem 3.5, we can show that with probability $1 - \delta$ over F, the pair $(\psi^\star|_F, \gamma^\star)$ – where $\psi^\star|_F$ is the mapping $\psi^\star$ restricted to the domain F – is feasible in the empirical fair learning problem (4), which again implies by Theorem 3.3 that with probability $1 - \delta$ over F

$$\mathrm{err}\left(\widehat{\psi}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) = \mathrm{err}\left(\widehat{\boldsymbol{p}}; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) \leqslant \mathrm{err}\left(\psi^\star|_F; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu = \mathrm{err}\left(\psi^\star; \widehat{\mathcal{P}}, \widehat{\mathcal{Q}}\right) + 2\nu$$

Putting all inequalities together, we have that with probability $1 - 7\delta$ over (X, F),

$$\mathrm{err}\left(\widehat{\psi}; \mathcal{P}, \mathcal{Q}\right) \leqslant \mathrm{OPT}(\alpha; \mathcal{P}, \mathcal{Q}) + O(\nu) + O(\alpha\beta)$$

□

C Proofs of Section A

We prove only the regret bounds of the players (Lemma A.1 and A.2) in this section. The rest of the proofs are similar to their counterparts in the AIF section.

C.1 Regret Bounds

Proof of Lemma A.1. Recall the Learner is accumulating regret only because they are using the given quantities $\{\widetilde{\rho}_i\}_{i=1}^n$ instead of $\{\widehat{\rho}_i\}_{i=1}^n$ that depend on the observed data F. In other words, at round $t \in [T]$ of the algorithm instead of best responding to

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{h}, \gamma, \boldsymbol{\lambda}_t\right) &= -\gamma \sum_{i=1}^n w_{i,t} \\ &+ \frac{1}{m} \sum_{j=1}^m \left\{ \sum_{i=1}^n \left(\frac{1}{n} + \frac{w_{i,t}}{\widehat{\rho}_i}\right) \mathbb{1}\left[h_j(x_i) = 1, f_j(x_i) = 0\right] + \left(\frac{1}{n}\right) \mathbb{1}\left[h_j(x_i) = 0, f_j(x_i) = 1\right] \right\}\end{aligned}$$

The Learner is in fact minimizing

$$\widetilde{\mathcal{L}}\left(\boldsymbol{h},\gamma,\boldsymbol{\lambda}_t\right) = -\gamma\sum_{i=1}^{n} w_{i,t}$$
$$+\frac{1}{m}\sum_{j=1}^{m}\left\{\sum_{i=1}^{n}\left(\frac{1}{n}+\frac{w_{i,t}}{\widetilde{\rho}_i}\right)\mathbb{1}\left[h_j(x_i)=1, f_j(x_i)=0\right]+\left(\frac{1}{n}\right)\mathbb{1}\left[h_j(x_i)=0, f_j(x_i)=1\right]\right\}$$

Observe that for any pair of $\left(\boldsymbol{h},\gamma\right)$,

$$\begin{aligned}\left|\mathcal{L}\left(\boldsymbol{h},\gamma,\boldsymbol{\lambda}_t\right)-\widetilde{\mathcal{L}}\left(\boldsymbol{h},\gamma,\boldsymbol{\lambda}_t\right)\right| &\leqslant \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\frac{w_{i,t}}{\widetilde{\rho}_i}-\frac{w_{i,t}}{\widehat{\rho}_i}\right|\mathbb{1}\left[h_j(x_i)=1, f_j(x_i)=0\right]\\ &\leqslant \sum_{i=1}^{n}\left|\frac{w_{i,t}}{\widetilde{\rho}_i}-\frac{w_{i,t}}{\widehat{\rho}_i}\right|\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\left[f_j(x_i)=0\right]\\ &= \sum_{i=1}^{n}\widehat{\rho}_i\left|\frac{w_{i,t}}{\widetilde{\rho}_i}-\frac{w_{i,t}}{\widehat{\rho}_i}\right|\\ &\leqslant \sum_{i=1}^{n}\frac{|\widehat{\rho}_i-\widetilde{\rho}_i|\cdot|w_{i,t}|}{\widetilde{\rho}_i}\end{aligned}$$

A simple application of Chernoff-Hoeffding's inequality (Theorem B.1) implies with probability $1-\delta$ over the observed labelings F, uniformly for all i,

$$|\widehat{\rho}_i-\rho_i| \leqslant \sqrt{\frac{\log\left(2n/\delta\right)}{2m_0}}$$

On the other hand by Assumption A.1, we have that for all $i\in[n]$,

$$|\widetilde{\rho}_i-\rho_i| \leqslant \sqrt{\frac{\log\left(n\right)}{2m_0}}$$

Therefore for a given t, with probability $1-\delta$ over F,

$$\begin{aligned}\left|\mathcal{L}\left(\boldsymbol{h},\gamma,\boldsymbol{\lambda}_t\right)-\widetilde{\mathcal{L}}\left(\boldsymbol{h},\gamma,\boldsymbol{\lambda}_t\right)\right| &\leqslant \frac{2}{\widetilde{\rho}_{\min}}\sqrt{\frac{\log\left(2n/\delta\right)}{2m_0}}\sum_{i=1}^{n}|w_{i,t}|\\ &= \frac{2}{\widetilde{\rho}_{\min}}\sqrt{\frac{\log\left(2n/\delta\right)}{2m_0}}\sum_{i=1}^{n}\left|\lambda_{i,t}^{+}-\lambda_{i,t}^{-}\right|\\ &\leqslant \frac{2B}{\widetilde{\rho}_{\min}}\sqrt{\frac{\log\left(2n/\delta\right)}{2m_0}}\end{aligned}$$

Now, let $\left(\boldsymbol{h}_t^{\star},\gamma_t^{\star}\right)=\arg\min_{\boldsymbol{h},\gamma}\mathcal{L}\left(\boldsymbol{h},\gamma,\boldsymbol{\lambda}_t\right)$ be the best response of the Learner to $\mathcal{L}$ and recall the actual play of the Learner is in fact $\left(\boldsymbol{h}_t,\gamma_t\right)=\arg\min_{\boldsymbol{h},\gamma}\widetilde{\mathcal{L}}\left(\boldsymbol{h},\gamma,\boldsymbol{\lambda}_t\right)$. Observe that the

above inequality implies for any t, with probability $1-\delta$ over F,

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) &\leqslant \widetilde{\mathcal{L}}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) + \frac{2B}{\widetilde{\rho}_{\min}}\sqrt{\frac{\log\left(4n/\delta\right)}{2m_0}} \\ &\leqslant \widetilde{\mathcal{L}}\left(\boldsymbol{h}_t^\star, \gamma_t^\star, \boldsymbol{\lambda}_t\right) + \frac{2B}{\widetilde{\rho}_{\min}}\sqrt{\frac{\log\left(4n/\delta\right)}{2m_0}} \\ &\leqslant \mathcal{L}\left(\boldsymbol{h}_t^\star, \gamma_t^\star, \boldsymbol{\lambda}_t\right) + \frac{4B}{\widetilde{\rho}_{\min}}\sqrt{\frac{\log\left(4n/\delta\right)}{2m_0}}\end{aligned}$$

It follows immediately that with probability $1-\delta$ over the observed labelings F, for all $t \in [T]$

$$\mathcal{L}\left(\boldsymbol{h}_t, \gamma_t, \boldsymbol{\lambda}_t\right) - \min_{\boldsymbol{p},\gamma} \mathcal{L}\left(\boldsymbol{p}, \gamma, \boldsymbol{\lambda}_t\right) \leqslant \frac{4B}{\widetilde{\rho}_{\min}}\sqrt{\frac{\log\left(4nT/\delta\right)}{2m_0}}$$

which completes the proof. □

Proof of Lemma A.2. The proof is almost similar to that of Lemma 3.1. The terms $\frac{B\log(2n+1)}{\eta T} + \eta\left(1+2\alpha\right)^2 B$ come from the regret analysis of the exponentiated gradient descent algorithm.The only difference is that now we have to bound the difference between false positive rates, one with respect to the distribution $\widehat{\mathcal{Q}}$ and another with respect to the distribution $\widehat{\mathcal{Q}}_t$, for all i and t:

$$\left|\mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right) - \mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right)\right|$$

Let $\widehat{\rho}_{i,t} = \mathbb{P}_{f\sim\widehat{\mathcal{Q}}_t}\left[f(x_i)=0\right]$ and recall $\widehat{\rho}_i = \mathbb{P}_{f\sim\widehat{\mathcal{Q}}}\left[f(x_i)=0\right]$. We have that

$$\mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right) = \left(\widehat{\rho}_i\right)^{-1} \cdot \mathop{\mathbb{E}}_{f\sim\widehat{\mathcal{Q}}}\left[\mathop{\mathbb{P}}_{h\sim\widehat{\psi}_{t,f}}\left[h(x_i)=1,\ f(x_i)=0\right]\right]$$

$$\mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right) = \left(\widehat{\rho}_{i,t}\right)^{-1} \cdot \mathop{\mathbb{E}}_{f\sim\widehat{\mathcal{Q}}_t}\left[\mathop{\mathbb{P}}_{h\sim\widehat{\psi}_{t,f}}\left[h(x_i)=1,\ f(x_i)=0\right]\right]$$

And therefore

$$\begin{aligned}&\left|\mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right) - \mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right)\right| \\ &\leqslant \frac{1}{\widehat{\rho}_i}\left\{\left|\mathop{\mathbb{E}}_{f\sim\widehat{\mathcal{Q}}}\left[\mathop{\mathbb{P}}_{h\sim\widehat{\psi}_{t,f}}\left[h(x_i)=1,\ f(x_i)=0\right]\right] - \mathop{\mathbb{E}}_{f\sim\widehat{\mathcal{Q}}_t}\left[\mathop{\mathbb{P}}_{h\sim\widehat{\psi}_{t,f}}\left[h(x_i)=1,\ f(x_i)=0\right]\right]\right| + \left|\widehat{\rho}_i - \widehat{\rho}_{i,t}\right|\right\}\end{aligned}$$

This inequality follows from the simple fact that for $a \leqslant b$ and $c \leqslant d$, all in $\mathbb{R}_+$:

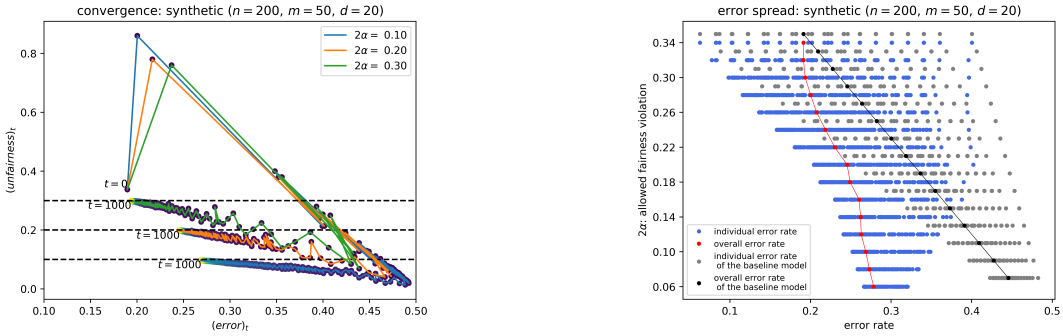
$$\left|\frac{a}{b} - \frac{c}{d}\right| = \left|\frac{a}{b} - \frac{c}{b} + \frac{c}{b} - \frac{c}{d}\right| \leqslant \frac{1}{b}\left|a-c\right| + c\left|\frac{1}{b} - \frac{1}{d}\right| \leqslant \frac{1}{b}\left|a-c\right| + \frac{1}{b}\left|b-d\right|$$

Now by viewing the batch F_t as a random sample from the set of problems F, we can apply the Chernoff-Hoeffding's inequality to bound each difference appearing above separately. We therefore have that with probability $1-\delta$ over the draws of F_t:

$$\left|\mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}\right) - \mathcal{E}_{\text{FP}}\left(x_i, \widehat{\psi}_t; \widehat{\mathcal{Q}}_t\right)\right| \leqslant \frac{2}{\widehat{\rho}_i}\sqrt{\frac{\log\left(4/\delta\right)}{2m_0}}$$

The rest is similar to how the proof for Lemma 3.1 proceeds. □

D Additional Experimental Results



(a) convergence plot: synthetic data

(b) error spread plot: synthetic data

Figure 3: (a) Error-unfairness trajectory plots illustrating the convergence of algorithm **AIF-Learn**. (b) Error-unfairness tradeoffs and individual errors for **AIF-Learn** vs. simple mixtures of the error-optimal model and random classification. Gray dots are shifted upwards slightly to avoid occlusions.

In addition to the experiments on the Communities and Crime data set, we ran additional experiments on synthetic data sets to further verify the performance of our algorithms empirically. Given the parameters n (number of individuals), m (number of problems), d (dimension), and a coin with bias $q \geqslant 0.5$, an instance of the synthetic data is generated as follows:

First n individuals $\{x_i\}_{i=1}^n$ are generated randomly where $x_i \in \{\pm 1\}^d$. Let the first 75% of the individuals be called "group 1" and the rest "group 2". For each learning task $j \in [m]$, we randomly sample two weight vectors $w_{j,\text{maj}}, w_{j,\text{min}} \in \{\pm 1\}^d$. Now consider a matrix $Y \in \{0,1\}^{n \times m}$ that needs to be filled in with labels. We intentionally inject unfairness into the data set (because otherwise fairness would be achieved for free due to the random generation of data) by considering "majority" and "minority" groups for each learning task (majority group will be labelled using $w_{j,\text{maj}}$ and the minority by $w_{j,\text{min}}$). For each learning task j, individuals in group 1 get q chance of falling into the majority group of that task and individuals in group 2 get $1 - q$ chance (this is implemented by flipping a coin with bias q for each individual). Once the majority and minority groups for each learning task are specified, the labels $Y = [y_{ij}]$ are set according to $y_{ij} = \text{sign}(w_{i,j}^T x_i)$ where $w_{i,j} \in \{w_{j,\text{maj}}, w_{j,\text{min}}\}$. (Notice $q = 1$ corresponds to a fixed (across learning tasks) majority group of size $0.75n$, and $q = 0.5$ corresponds to the case where the majority groups are formed completely at random and that they have size $0.5n$ in expectation.)

It is "expected" that solving the unconstrained learning problem on an instance of the above synthetic data generation process would result in discrimination against the minority groups which are formed mostly by the individuals in group 2, and therefore our algorithm should provide a fix for this unfair treatment. We therefore took an instance of the synthetic data set with parameters $n = 200, m = 50, d = 20, q = 0.8$ and ran our algorithm on varying values of allowed fairness violation 2α (notice on an input α to the algorithm, the individual error rates are allowed to be within at most 2α of each other). We present the results (which are reported in-sample) in Figure 3 where similar to the results on the Communities and Crime data set, we get convergence after $T = 1000$ iterations (or 50,000 calls to the learning oracle, see Figure 3(a)). Furthermore, Figure 3(b) suggests that our algorithm on the synthetic data set considerably outperforms the baseline that only mixes the error-optimal model with random classification.