Fairness Trade-Offs and Partial Debiasing

Francois Buet-Golfouse UCAHFBU@UCL.AC.UK
University College London, London, United Kingdom

Islam Utyagulov ISLAM.UTYAGULOV@GMAIL.COM
Independent Researcher, London, United Kingdom

Editors: Emtiyaz Khan and Mehmet Gonen

Abstract

Previous literature has shown that bias mitigating algorithms were sometimes prone to overfitting and had poor out-of-sample generalisation. This paper is first and foremost concerned with establishing a mathematical framework to tackle the specific issue of generalisation. Throughout this work, we consider fairness trade-offs and objectives mixing statistical loss over the whole sample and fairness penalties on categories (which could stem from different values of protected attributes), encompassing partial de-biasing. We do so by adopting two different but complementary viewpoints: first, we consider a PAC-type setup and derive probabilistic upper bounds involving sample-only information; second, we leverage an asymptotic framework to derive a closed-form limiting distribution for the difference between the empirical trade-off and the true trade-off. While these results provide guarantees for learning fairness metrics across categories, they also point out to the key (but asymmetric) role played by class imbalance. To summarise, learning fairness without having access to enough category-level samples is hard, and a simple numerical experiment shows that it can lead to spurious results.

1. Introduction

Reducing bias in an algorithm requires a number of steps; first, specify the fairness definitions (and thus fairness metrics) that apply, second, encode them with a penalty so as to measure the discrepancy between outcomes and the perfect fairness scenario, third, choose the trade-off between the original statistical goal and fairness constraints, and, fourth, pick a method to debias (at least partially) the model. There are a number of obstacles to this programme. Indeed, debiasing may come at a cost Rodolfa et al. (2020), but there are also theoretical reasons behind this claim. Jointly tackling multiple fairness definitions is usually difficult if not impossible without a decrease in model performance due to a series of impossibility theorems Kleinberg et al. (2017); Pleiss et al. (2017); Chouldechova (2017); Kim et al. (2020). Note that these results apply not only to fairness-performance but also to fairness-fairness trade-offs.

Research has recently been undertaken to tackle specifically learnability and generalisation in fairness-related problems. In particular, One (2020); Oneto et al. (2020) derive provable probabilistic upper bounds in a fairness setting where one estimates an overall statistical loss and monitors disparities across categories. In Chen et al. (2018), the authors have applied bias-variance decomposition techniques to disentangle the sources of unfairness in a classifier. While this is not directly related to this present work, the principle of risk decomposability turns out to be fruitful in both setups. Finally, in Agrawal et al. (2020), a limiting distribution is established in the simple case where there is a partial requirement involving a fairness metric with two categories.

Our contributions. The novelty of our work consists of the application of known techniques such as PAC inequalities and the central limit theorem to the problem of learning under fairness constraints:

- First, by leveraging results from Donini et al.; One (2020), we develop a PAC framework for fairness trade-offs, tackling explicitly *partial debiasing* (Kim et al., 2020; Agrawal et al., 2020).
- Second, we show that certain properties of the probabilistic upper bounds lead to the need for sample-efficient bias mitigation techniques. In particular, we show that a new quantity, Z_S, which can be understood as measure of sample concentration, plays a crucial role. We also build on Agrawal et al. (2020) and develop an asymptotic framework with a known limiting distribution; this is useful as PAC bounds may not always be very sharp.
- Third, in both frameworks, we put forward the decomposability of generalisation risk, expressed as an upper bound in the PAC realm and a limiting variance in the asymptotic one.
- Last, we illustrate our results on real-life data and show that generalisation indeed presents specific trade-offs.

2. Fairness Metrics and Loss Functions

2.1. Set-Up and Definitions

In this article, we define $s = 1, \cdots, C$ to be a categorical (protected) attribute, $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ to be a set of non-protected features excluding s. $y \in \{-1, +1\}$ is a binary outcome variable and $\hat{y} \in \{-1, +1\}$ is an estimator for y. Finally, $z = (\mathbf{x}, y, s)$. Note that $\hat{y}$ is derived from a learner $h \in \mathcal{H}$, where $\mathcal{H}$ is a given functional space. Furthermore, we define S to be the in-sample (training) empirical distribution, and D to refer to the true distribution.

Throughout this paper, we consider the case of an overall objective as a statistical performance objective, L^0, plus a fairness loss function ϕ. The trade-off is tuned by a hyper-parameter $\lambda \geq 0$. As usual, the aim is to minimise the overall objective, i.e., minimise the statistical loss and the lack of fairness. This set-up is typical for partial debiasing and can be found in Kim et al. (2020).

Definition 1. The *overall objective* or *fairness trade-off* is defined as

$$\begin{aligned}\mathcal{L}_D(h) &= L^0_D(h) + \lambda\phi\left(L^+_{D,1}(h), L^-_{D,1}(h), \cdots, L^+_{D,C}(h), L^-_{D,C}(h)\right)\\ &= L^0_D(h) + \lambda\phi(\mathbf{L}^\pm_D(h)),\end{aligned}$$

where we have used the standard notations (see Shalev-Shwartz and Ben-David (2014)):

$$\begin{aligned} L^0_D(h) &= \mathbb{E}\left[\ell^0(h,z)\right] \\ L^+_{D,a}(h) &= \mathbb{E}\left[\ell^+(h,z)|s=a, y=1\right] \\ L^-_{D,a}(h) &= \mathbb{E}\left[\ell^-(h,z)|s=a, y=-1\right] \end{aligned}$$

and the vector notation $\mathbf{L}^{\pm}_D(h) = \left[L^+_{D,1}(h), L^-_{D,1}(h), \cdots, L^+_{D,C}(h), L^-_{D,C}(h)\right]^T$, for some functions ℓ^0, ℓ^+ and ℓ^-.

Remark 1. In the whole paper, we assume that there exists a uniform bound $B > 0$ on all functions ℓ, i.e., $|\ell(h,z)| \leq B$ for all h and z. ℓ refers to any such function in what follows.

Let us detail two particular cases. If $\lambda = 0$, then the overall objective boils down to the usual risk minimisation problem. If $\ell^0 = 0$ or $\lambda \to +\infty$, then the trade-off becomes a fairness constraint.

We are now in a position to define the empirical counterpart by considering the empirical distribution rather than the true distribution. With the additional notations $\mathcal{N}^{\pm}_a = \{i \in \{1, \cdots, n\}\,; s_i = a, y_i = \pm 1\}$, $|\mathcal{N}^{\pm}_a| = n^{\pm}_a$, the corresponding *sample* versions can be defined as $L^0_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell^0(h, z_i)$, $L^+_{S,a}(h) = \frac{1}{n^+_a}\sum_{i\in\mathcal{N}^+_a} \ell^+(h, z_i)$ and $L^-_{S,a}(h) = \frac{1}{n^-_a}\sum_{i\in\mathcal{N}^-_a} \ell^-(h, z_i)$, where $n_a = n^+_a + n^-_a$, $n = \sum_{a=1}^{C} n_a$, leading to

$$\mathcal{L}_S(h) = L^0_S(h) + \lambda\phi\left(\mathbf{L}^{\pm}_S(h)\right) \tag{1}$$

For simplicity's sake, we refer to $\phi(\mathbf{L}^{\pm}_T(h))$ as $\phi_T(h)$ for $T \in \{D, S\}$.

2.2. Fairness Definitions and Metrics

One can find multiple technical definitions of fairness in the literature; they have been reviewed in various papers Narayanan (2018); Verma and Rubin (2018); Berk et al. (2018); Kim et al. (2020); Agrawal et al. (2020). We offer an overview of the most frequent metrics in Table 1.

Table 1: Frequent fairness definitions.

Fairness metric	Reference	Equality requirement
Equalized false omission rate	Berk et al. (2018)	$\mathbb{P}(y=1\|\hat{y}=-1, s=a) = \mathbb{P}(y=1\|\hat{y}=-1)$
Predictive parity	Chouldechova (2017)	$\mathbb{P}(y=1\|\hat{y}=1, s=a) = \mathbb{P}(y=1\|\hat{y}=1)$
Demographic parity	Calders and Verwer (2010)	$\mathbb{P}(\hat{y}=1\|s=a) = \mathbb{P}(\hat{y}=1)$
Equalized false negative rate	Chouldechova (2017)	$\mathbb{P}(\hat{y}=-1\|y=1, s=a) = \mathbb{P}(\hat{y}=-1\|y=1)$
Predictive equality	Chouldechova (2017)	$\mathbb{P}(\hat{y}=1\|y=-1, s=a) = \mathbb{P}(\hat{y}=1\|y=-1)$
Equality of opportunity	Hardt et al. (2016)	$\mathbb{P}(\hat{y}=1\|y=1, s=a) = \mathbb{P}(\hat{y}=1\|y=1)$
Equalized odds	Hardt et al. (2016)	$\mathbb{P}(\hat{y}=1\|y=y', s=a) = \mathbb{P}(\hat{y}=1\|y=y')$

We note that all fairness metrics considered here can be expressed as an equality requirement on probabilities, hence on the expectation of indicator variables. These fall within our framework.

2.3. Fairness Loss Functions

In addition to picking one or multiple fairness definitions, we need to specify ϕ to measure the discrepancy from perfect fairness, the ideal case.

A first approach, similar to the Calders-Verwer gap Calders and Verwer (2010), consists of looking at the discrepancy between two categories a and a':

$$\Delta^{L,\pm}_{T,a,a'}(h) = L^{\pm}_{T,a}(h) - L^{\pm}_{T,a'}(h),$$

for $T \in \{D, S\}$. It is worth 0 under perfect fairness, but is asymmetric and thus requires defining an advantaged (or benchmark) category a. To avoid the issue of asymmetry, one can follow One (2020), and define ϕ as the sum of all absolute discrepancies across categories:

$$\phi(\mathbf{L}^{\pm}_{T}(h)) = \sum_{a \neq a'} \left\{ \left| L^{+}_{T,a} - L^{+}_{T,a'} \right| + \left| L^{-}_{T,a} - L^{-}_{T,a'} \right| \right\}, \tag{2}$$

for $T \in \{D, S\}$. Notice that, thanks to the reverse triangle inequality, the function ϕ is Lipschitz continuous. One could naturally weigh the various contributions to this fairness function and use another norm than L^1. Other functions involving ratios or relative differences, are also possible.

Finally, in Kim et al. (2020), the authors consider a convex combination of multiple loss functions $\phi^{(1)}, \cdots, \phi^{(M)}$:

$$\phi(\mathbf{L}^{\pm}_{T}(h)) = \sum_{j=1}^{M} \lambda_j \phi^{(j)} \left(\mathbf{L}^{\pm}_{T}(h) \right).$$

Note that if the $\phi^{(j)}$'s are Lipschitz-continuous (with respect to the loss vector $\mathbf{L}^{\pm}_{T}(h)$), with respective Lipschitz constant $K^{(j)}$, then the overall loss ϕ is itself Lipschitz-continuous with constant $K^{\phi} = \sum_{j=1}^{M} \lambda_j K^{(j)}$.

3. Learning Fairness Trade-Offs

In this Section, we consider the learning problem from different angles. We start by considering the statistical loss L^0 and derive a bound based on the Rademacher complexity observed in each category. To do so, we borrow from the statistical learning toolbox and briefly review Rademacher complexities. We then move on to learning the fairness part per se, namely study the generalisation properties of $\phi_S(h)$. Finally, we bring the two together and derive probabilistic upper bounds on fairness trade-off generalisation.

3.1. Learning the Statistical Loss

Let us first focus on the *statistical* component of the loss, i.e., L^0_T for $T \in \{S, D\}$. Much of this paper is based on theory of bounds derived from Rademacher complexities, introduced in Bartlett and Mendelson (2003) and surveyed in Boucheron et al. (2005). Recent textbooks such as Shalev-Shwartz and Ben-David (2014); Mohri et al. (2012) provide excellent introductions to the topic. We start by recalling the definition of Rademacher complexities, as indicated in Boucheron et al. (2005):

Definition 2. The Rademacher complexity (or average) of a function ℓ is given by

$$R(\ell \circ \mathcal{H} \circ \mathcal{S}) = \mathbf{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \ell\left(h, z_i\right) \right| \right]. \tag{3}$$

Note that definitions can slightly vary (depending for instance on the presence or absence of the absolute value in the definition), but all downstream results are qualitatively similar, up to some multiplicative constants.

To make this more concrete, we can consider a simple (but usual) case (as described in Shalev-Shwartz and Ben-David (2014)). We suppose that $\|\mathbf{x}\|_2 \leq R$ almost surely. In addition, if we let $\mathcal{H} = \{\mathbf{w}; \|\mathbf{w}\|_2 \leq R'\}$ and assume that the loss function ℓ is of the type $\ell(\mathbf{w}, (\mathbf{x}, y)) = \rho\left(\mathbf{w}^T \mathbf{x}, y\right)$, where $|\rho|$ is bounded by B and is L^ρ-Lipschitz, then, almost surely,

$$R(\ell \circ \mathcal{H} \circ \mathcal{S}) \leq \frac{L^\rho R R'}{\sqrt{n}}. \tag{4}$$

Rademacher complexities are key to establishing learnability bounds and lead to fundamental results in statistical learning theory. In particular, we will make use of a standard result (see Boucheron et al. (2005)).

Proposition 1. *With probability at least* $1-\delta$*, it holds*

$$\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \leq 2R(\ell \circ \mathcal{H} \circ \mathcal{S}) + B\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \tag{5}$$

It will also be useful to introduce *conditional* Rademacher complexities that will be used throughout this paper. In particular, since we have a partition of the sample in terms of categories $\mathcal{S} = \bigcup_{a=1}^{C} \mathcal{N}_a$, we can consider the Rademacher complexity of that particular sample:

$$R(\ell \circ \mathcal{H} \circ \mathcal{N}_a) = \mathbf{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n_a} \left| \sum_{i \in \mathcal{N}_a} \sigma_i \ell\left(h, z_i\right) \right| \right]. \tag{6}$$

Lemma 1. *The sample Rademacher complexity can be bounded from above by the weighted sum of conditional Rademacher complexities:*

$$R(\ell \circ \mathcal{H} \circ \mathcal{S}) \leq \sum_{a=1}^{C} \frac{n_a}{n} R(\ell \circ \mathcal{H} \circ \mathcal{N}_a), \tag{7}$$

where $R(\ell \circ \mathcal{H} \circ \mathcal{N}_a) = \mathbf{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n_a} \left|\sum_{i \in \mathcal{N}_a} \sigma_i \ell\left(h, z_i\right)\right|\right]$.

Proof This comes directly from the sub-additivity of the absolute value and the supremum.

One can further interpret the non-negative gap $\sum_{a=1}^{C} \frac{n_a}{n} R(\ell \circ \mathcal{H} \circ \mathcal{N}_a) - R(\ell \circ \mathcal{H} \circ \mathcal{S})$ as a diversification benefit amongst categories. Finally, this leads us to a proposition leveraging our results so far:

Proposition 2. *With probability at least* $1-\delta$*,*

$$\sup_{h \in \mathcal{H}} \left|L_D^0(h) - L_S^0(h)\right| \leq 2\sum_{a=1}^{C} \frac{n_a}{n} R(\ell^0 \circ \mathcal{H} \circ \mathcal{N}_a) + B\sqrt{\frac{2\log\frac{2}{\delta}}{n}}. \tag{8}$$

Proof This follows directly from applying Lemma 1 to Proposition 1.

3.2. Learning Fairness Requirements

Let us now move on to the *fairness* learning part; as before, we would like to find a bound on the *distribution* fairness loss given the *sample* fairness loss, i.e., a probabilistic bound on the difference $\phi_D(h) - \phi_S(h)$.

Proposition 3. *Under the assumption that ϕ is K^ϕ-Lipschitz, it holds, with probability at least $1-\delta$, and where $Z_S := \sum_{a=1}^{C} \sqrt{\frac{n}{n_a^+}} + \sqrt{\frac{n}{n_a^-}}$, that*

$$\begin{aligned}\sup_{h\in\mathcal{H}} |\phi_D(h) - \phi_S(h)| &\leq 2K^\phi \sum_{a=1}^{C} R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-) \\ &\quad + K^\phi B Z_S \sqrt{\frac{2\log\frac{4C}{\delta}}{n}}.\end{aligned}$$

The proof can be found in Section C.1 of supplementary materials and the role played by Z_S is investigated in Section 4.

3.3. Simultaneous Learning of Statistical Performance and Fairness

We have established probabilistic bounds on both the statistical performance criterion and the fairness penalty. However, we wish to study both jointly, and determine the behaviour of the chosen fairness trade-off. As mentioned previously, this trade-off is very flexible and can accommodate multiple situations.

3.3.1. Bounding Loss and Fairness

First, we may wish to determine a probabilistic upper bound on $L_D^0(h) - L_S^0(h)$ and $\phi_D(h) - \phi_S(h)$ jointly. By a simple application of the union bound, we obtain the following result:

Proposition 4. *With probability at least $1-\delta$, it holds* jointly *that*

$$\begin{aligned}\sup_{h\in\mathcal{H}} \left|L_D^0(h) - L_S^0(h)\right| &\leq 2R(\ell^0 \circ \mathcal{H} \circ \mathcal{S}) + B\sqrt{\frac{2\log\frac{4}{\delta}}{n}} \\ \sup_{h\in\mathcal{H}} |\phi_D(h) - \phi_S(h)| &\leq 2K^\phi \sum_{a=1}^{C} R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-) \\ &\quad + K^\phi B Z_S \sqrt{\frac{2\log\frac{8C}{\delta}}{n}}.\end{aligned}$$

3.3.2. Bounding the Trade-Off

As mentioned above, the existence of various impossibility results and the empirical findings showing that statistical performance tends to decrease as fairness requirements increase have highlighted the interest for partial debiasing methods. It is however important to be able to determine the generalisation properties that such objectives can lead to. We can perform a similar analysis on the trade-off objective $\mathcal{L}_D(h)$ by using the same arguments to get an upper bound on this objective.

Proposition 5. *With probability at least* $1-\delta$*, it holds*

$$\begin{aligned}
\sup_{h\in\mathcal{H}} |\mathcal{L}_D(h) - \mathcal{L}_S(h)| &\leq 2R(\ell^0 \circ \mathcal{H} \circ \mathcal{S}) + B\sqrt{\frac{2\log\frac{4}{\delta}}{n}} \\
&\quad +2\lambda K^{\phi}\sum_{a=1}^{C} R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-) \\
&\quad +\lambda K^{\phi} B Z_S \sqrt{\frac{2\log\frac{8C}{\delta}}{n}} \\
&\leq 2\sum_{a=1}^{C}\Bigg\{\frac{n_a^+}{n}R(\ell^0 \circ \mathcal{H} \circ \mathcal{N}_a^+) + \frac{n_a^-}{n}R(\ell^0 \circ \mathcal{H} \circ \mathcal{N}_a^-) \\
&\quad +\lambda K^{\phi} R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + \lambda K^{\phi} R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-)\Bigg\} \\
&\quad +B\left(1+\lambda K^{\phi} Z_S\right)\sqrt{\frac{2\log(8C/\delta)}{n}}.
\end{aligned}$$

The upper bound can be understood as the sum of individual contributions plus the usual $O\left(\sqrt{\log(1/\delta)/n}\right)$ factor. Let us now establish a related inequality, linking the distribution value of the trade-off under the distribution optimum and under the sample optimum. Let us denote by $h_T^* = \arg\inf_{h\in\mathcal{H}} \mathcal{L}_T(h)$ for $T \in \{D, S\}$ an optimal classifier derived on either the underlying distribution or the sample distribution of the objectives $\mathcal{L}_T$. This is an important issue for learnability of fairness as it provides some guarantees on how far the actual optimal trade-off is from the one we would obtain by picking the optimum computed on the sample $\mathcal{S}$. This result provides a fairness pendant to the usual case in statistical learning (see for instance Theorem 26.5 in Shalev-Shwartz and Ben-David (2014)).

Theorem 1. *With probability at least* $1-\delta$*, it holds*

$$0 \leq \mathcal{L}_D(h_S^*) - \mathcal{L}_D(h_D^*) \leq UB_{\mathcal{S}} + B\left(1+\lambda Z_S K^{\phi}\right)\frac{3}{\sqrt{2}}\sqrt{\frac{\log(16C/\delta)}{n}}, \tag{9}$$

where $h_T^* = \arg\inf_{h\in\mathcal{H}} \mathcal{L}_T(h)$ *for* $T \in \{D, S\}$*, and* $UB_{\mathcal{S}} := 2\sum_{a=1}^{C}\left\{\frac{n_a^+}{n}R(\ell^0 \circ \mathcal{H} \circ \mathcal{N}_a^+) + \frac{n_a^-}{n}R(\ell^0 \circ \mathcal{H} \circ \mathcal{N}_a^-) + \lambda K^{\phi} R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + \lambda K^{\phi} R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-)\right\}$.

The proof is given in Section C.2 of supplementary materials. What is particularly interesting about this result is the fact that the upper bound remains the same as in Proposition 5, and only the $O\left(\sqrt{\log(1/\delta)/n}\right)$ factor changes. Let us point out that this result is *practical* in nature in the sense that one would use h_S^* in practice but would still look for out-of-sample generalisation, hence the importance of the term $\mathcal{L}_D(h_S^*)$.

4. Practical Consequences of PAC for Fairness Trade-Offs

4.1. Decomposing Upper Bounds on Generalisation

One is led to decompose the overall loss in terms of each category's contributions to the overall learning upper bound.

Definition 3. The (half-)contribution – denoted by $\overline{R}_{\mathcal{N}_a}$– of each category $a = 1, \cdots, C$, to the fairness trade-off learning upper bound is

$$\begin{aligned}\overline{R}_{\mathcal{N}_a} \quad := \quad & \frac{n_a^+}{n} R(\ell^0 \circ \mathcal{H} \circ \mathcal{N}_a^+) + \frac{n_a^-}{n} R(\ell^0 \circ \mathcal{H} \circ \mathcal{N}_a^-) \\ & + \lambda K^\phi R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + \lambda K^\phi R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-).\end{aligned}$$

Remark 2. This structure is quite interesting as it shows that the first two terms representing contributions to the statistical loss upper bound are weighted by their proportions of the overall sample. Thus, even if the Rademacher complexity of the category is high, it can be counterbalanced by the fact that it will not affect the overall average loss. On the other hand, the contributions coming from the fairness part are unweighted. In this case, a low sample size would usually lead to a higher Rademacher complexity.

Now, $\overline{R}_{\mathcal{N}_a}$ can be decomposed into finer components:

$$\overline{R}_{\mathcal{N}_a} = \overline{R}^{0,+}_{\mathcal{N}_a} + \overline{R}^{0,-}_{\mathcal{N}_a} + \overline{R}^{+}_{\mathcal{N}_a} + \overline{R}^{-}_{\mathcal{N}_a}, \tag{10}$$

with obvious notations. One can then determine the contributions coming from the union of certain categories or of positives or negatives.

4.2. Z_S as Sample Concentration

First, let us remark that the usefulness of probabilistic upper bounds comes from the fact that the ones that have been established only rely on observable quantities coming from the sample under consideration. This is powerful as a practical check, but it also suggests a way of amending the initial minimisation exercise by including the upper bound to the sample component, i.e., optimise $\mathcal{L}_D(h) + \mathrm{UB}_{\mathcal{S}}(h)$. Second, a quantity that is ubiquitous in our analysis (see, for instance, Theorem 1), is $Z_S = \sum_{a=1}^C \sqrt{\frac{n}{n_a^+}} + \sqrt{\frac{n}{n_a^-}}$. As noticed previously, this is a measure of concentration within the sample that only depends on in-sample data. The more unequal the count of categories' sample sizes, the higher Z_S. This can be formalised in the following proposition:

Proposition 6. *The constant Z_S can be expressed in terms of the empirical class sample proportions:*

$$Z_S = \sum_{a=1}^{C} \left[(\widehat{p}_a^+)^{-\frac{1}{2}} + (\widehat{p}_a^-)^{-\frac{1}{2}} \right] = \frac{2C}{\sqrt{M_{-1/2}(\widehat{\mathbf{p}})}}, \tag{11}$$

where $M_\alpha(\widehat{\mathbf{p}})$ is the α-generalised mean [1] *of the vector $\widehat{\mathbf{p}} = (\widehat{p}_a^\pm)_a^\pm$. In addition, under the assumption that $n_a^\pm \geq 1$ for all $a = 1, \cdots, C$, and $n > 2C - 1$, the following inequality holds:*

$$(2C)^{3/2} \leq Z_S \leq (2C-1)\sqrt{n} + \sqrt{\frac{n}{n-(2C-1)}}. \tag{12}$$

Proof To obtain the upper bound, set all cardinals to 1, except for one category with $n - (2C - 1)$. For the lower bound, set $n_a^\pm = \frac{n}{2C}$.

[1] The α-generalised mean of K numbers $x_1, \cdots, x_K$ is given by $\left(\frac{1}{K}\sum_{k=1}^K x_k^\alpha\right)^{1/\alpha}$.

The bounds on Z_S offer two take-aways. First, the lower bound grows super-linearly as a function of the number of categories, which is problematic especially to tackle *intersectionality*. Second, the upper bound on $Z_S/\sqrt{n}$ converges to a *non-zero* constant as n goes to infinity. This is an extreme scenario whereby a category can be made up of one individual only, but this stresses the fact that class imbalance can hinder learning in the context of fairness. Z_S is model-agnostic and thus useful to determine *ex-ante* how much debiasing can be applied on a given dataset.

5. Asymptotic Regime

One of the possible issues of using PAC learning in practice is the fact that it tends to consider worst-case scenarios (e.g., by deriving probabilistic inequalities over a whole space of possible classifiers). To mitigate this possible drawback, we adopt a very different viewpoint and set ourselves in a framework which generalises Proposition 1 in Agrawal et al. (2020). Our objective is to derive a limiting distribution for the rescaled fairness trade-off after having trained *one* particular classifier, which we then consider fixed. In doing so, we can thus assess the different contributions to the limiting variance *given* the choice of classifier. In addition to corresponding to real-life modelling, this also has the benefit of allowing for *closed-form* results (as opposed to upper bounds in PAC learning). Both approaches are – as will be shown – very complementary.

Let us say a few words about the idealised data generating process that hypothetically produces our observations. In particular, we consider that we have an infinite population of independently and identically distributed datapoints, according to a mixture distribution with probability weights $p_a^{\pm}$. Conditional on a and $y = \pm 1$, we can then draw $\mathbf{x}$. All draws are independent. Let us introduce some additional notation: the intra-category variances are given by $\left(\sigma_a^{0,\pm}\right)^2 = \mathbb{V}_{s=a,y=\pm 1}\left[\ell^0(h,z)\right]$, $\left(\sigma_a^{\pm}\right)^2 = \mathbb{V}_{s=a,y=\pm 1}\left[\ell^{\pm}(h,z)\right]$. Furthermore, $K_{D,a}^{\pm} = \partial_{x_a^{\pm}}\phi_D(h)$, i.e., $K_{D,a}^{\pm}$ is the entry corresponding to $a^{\pm}$ in the gradient $\nabla\phi_D(h)$. We can now introduce our main result, which extends Agrawal et al. (2020) to a multi-category and multi-dimensional setting.

Theorem 2. *Let $h \in \mathcal{H}$, under mild assumptions, we have the following convergence in distribution*

$$\sqrt{n}\left(\mathcal{L}_S(h) - \mathcal{L}_D(h)\right) \xrightarrow[n\to+\infty]{d} N\left(0, \mathbb{V}_{\lim}(h)\right), \tag{13}$$

where

$$\mathbb{V}_{\lim}[h] = \sum_{a^{\pm}} p_a^{\pm}\left(\sigma_a^{0,\pm}\right)^2 + \lambda^2 \frac{(K_a^{\pm})^2}{p_a^{\pm}}\left(\sigma_a^{\pm}\right)^2 + \textit{covariance terms} \tag{14}$$

The exact assumptions and the proof can be found in Section C.3 of supplementary materials. In a nutshell, the variance of the limiting trade-off distribution is the addition of

- The sum of intra-category $\ell^{0,\pm}$ losses weighted by their true probabilities $p_a^{\pm}$;
- The sum of intra-category fairness $\ell^{\pm}$ losses weighted by their *inverse* probabilities $p_a^{\pm}$, as in the uniform convergence framework. In addition, the constant $(K_a^{\pm})^2$ represents the sensitivity of the fairness loss function ϕ and is bounded from above by $[K^{\phi}]^2$

in these case where ϕ is K^{ϕ}-Lipschitz continuous, while λ is the hyper-parameter governing the trade-off between statistical and fairness performances.

Qualitatively, the results obtained in the asymptotic and uniform frameworks are similar and point to the conclusion that to learn fairness, one must start by learning category samples. A low probability $p_a^{\pm}$ may not impact the variance of the overall statistical loss but it may lead to a very poor understanding of fairness trade-offs.

6. Experiments

In this Section, we investigate on real datasets the validity of the theoretical results that we have derived. We aim to answer three key research questions:

- **RQ1** Fairness generalisation trade-offs and partial debiasing,
- **RQ2** Out-of-sample disparity and class imbalance,
- **RQ3** Out-of-sample disparity, uncertainty and model complexity.

6.1. Datasets

The three datasets used in the experiments are detailed in Table 2. The protected attribute(s) and the respective outcome are shown in third and fourth columns of Table 2. Typically, only binary protected attributes are considered, whereas we also demonstrate some results for a multivariate protected attribute (e.g., the 'race' attribute in *Adult* dataset from Dua and Graff (2017)). In our experiments we run 5-fold cross-validation (CV). We also run 5-fold CV with different random initialisations to compute uncertainty of associated metrics.

Table 2: Dataset overview

Dataset name	Reference	Protected attribute(s)	Binary outcome
Adult	Dua and Graff (2017)	Sex, Race	Income exceeds 50$K/year
Loan Defaults	I-Cheng Yeh (2009)	Sex	Default payment next month
German Credit	Dua and Graff (2017)	Sex, Sex & Marital Status	Creditability

6.2. Experiment Design

To illustrate the presence of trade-offs, we consider throughout this Section the following partially debiased binary classification problem with C groups:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathbf{x}_i; \boldsymbol{\beta}) + \lambda\phi^{(j)}\left(\frac{1}{n_0}\sum_{j=1}^{n_0}\ell(y_j, \mathbf{x}_j; \boldsymbol{\beta}), \cdots, \frac{1}{n_{C-1}}\sum_{j'=1}^{n_{C-1}}\ell(y_{j'}, \mathbf{x}_{j'}; \boldsymbol{\beta})\right) \tag{15}$$

where ℓ is the logistic loss. We have used two different penalty functions ($j = 1, 2$):

1. *Squared loss*: $\phi^{(1)}(z_0, \cdots, z_{C-1}) = \sum_{c=0}^{C-1}\left(z_c - \left[\frac{1}{C}\sum_{c'=0}^{C-1} z_{c'}\right]\right)^2$
2. *Soft maximum loss*: $\phi^{(2)}(z_0, \cdots, z_{C-1}) = \frac{1}{\alpha}\log\left(\frac{1}{C}\sum_{c=0}^{C-1} e^{\alpha(z_c - [\frac{1}{C}\sum_{c'=0}^{C-1} z_{k'}])}\right)$.

Remark 3. Properties of this soft maximum, called Log-Sum-Exponential, can be found in Nielsen and Sun (2016). When $\alpha \to 0$, the loss goes to 0, whereas when $\alpha \to +\infty$, it attains the maximum disparity.

For the sake of simplicity, the disparity metric considered in our examples in the *mean absolute deviation*, i.e., $\Delta(z_0, \cdots, z_{C-1}) = \sum_{c=0}^{C-1} \left| z_c - \left[\frac{1}{C} \sum_{c'=0}^{C-1} z_{c'} \right] \right|$. Note that similar results are available with different metrics.

6.3. RQ1: Fairness Trade-Offs and Partial Debiasing

In this subsection, we consider Equation 15 with $\phi^{(1)}$ and $C = 2$, for the different values of λ as well as the role of partial debiasing. Figure 1 illustrates a trade off between average logistic loss and group disparity for a range of λ values[2]. Each distinct color represents a different random initialization used in 5-fold CV and the size of each point is monotonically increasing function of the respective λ. Finally, each point's coordinates are the averages of loss and disparity metrics across different folds.

On the *training set* (the top row of Figure 1), there is a clear dependence between the value of λ and the respective average loss and group disparity. An increase in λ tends to decrease disparity, while increasing the average loss, and vice versa. However, this is not always the case on the *test set* (the bottom row of Figure 1), as some intermediate values of λ can achieve both lower average loss and lower disparity, whereas more (or totally) debiased models clearly overfit.

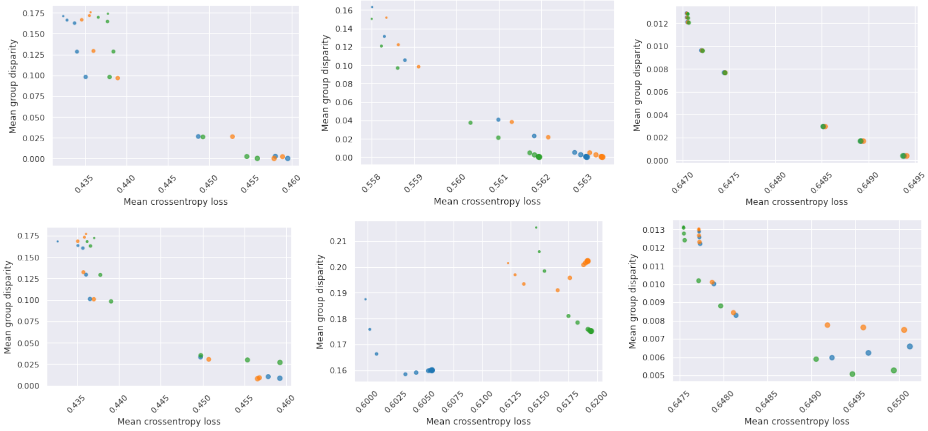

Figure 1: Trade off between mean group disparity and cross-entropy loss on *Adult*, *German Credit*, and *Loan Defaults* data sets. Each colour corresponds to a different initialisation, and the symbol's size corresponds to the λ that controls influence of group disparity on the overall loss. The top row of the Figure depicts results on the *training set*, whereas the bottom row corresponds to the *test set*.

[2] Specifically $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500\}$

6.4. RQ2: Out-of-Sample Disparity and Class Imbalance

We next examine class imbalance to see how the out-of-sample disparity is impacted by the growth of imbalance (e.g. here, we refer to the class imbalance induced by the protected attribute, and not the target) to illustrate Theorem 2.

6.4.1. Down-Sampling in a Bivariate Setup.

First, we down-sample observations from one class, while keeping the original number of observations constant in the other class. As predicted, out-of-sample, the variance of disparity increases with the class imbalance, as shown in Figure 2.

Here, each point corresponds to the standard deviation of overall disparity computed across different folds of data (i.e., the second term in Equation 15) and points of the same color correspond to the same random seed. Different random seeds are needed to obtain several variants of the down-sampled data. The dotted color lines are the plain linear least square estimate through the points of the same color, while the bold black line is least squares' estimate using data points from all seeds; λ is fixed at 50.

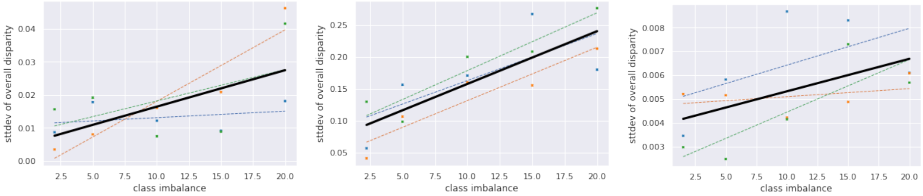

Figure 2: Fairness uncertainty versus class imbalance ratio on the *Adult*, *German Credit*, and *Loan Defaults* datasets. Results are measured on the test set. $\lambda = 50$.

6.4.2. Multivariate Setup.

We now turn our focus to a more realistic setting by considering a multivariate attribute. We fit a neural net with one hidden layer on *Adult* and *German Credit* datasets from Dua and Graff (2017) with 'race' and 'sex & marital status' as protected attributes respectively.[3], with 20 neurons in the hidden layer and using the ReLU activation function, for a range of α values[4] in $\phi^{(2)}$ with $\lambda = 1$, 5 different seeds, and 5 weight decays (equivalent to L2-norm regularisation). For each set of parameters, we then compute the standard deviation of the disparity metric for each group out-of-sample. We verify empirically that the less frequent a protected category is the more uncertain an estimate of group disparity is, see Figure 3. It should be noted that no 'female & single' group is present *German Credit* dataset, which illustrates the limit of debiasing.

[3] The protected attributes defined groups: 'race' has the following categories with the respective counts in brackets: White (41,762), Black (4,685), Asian-Pacific-Islander (1,519), Other (406), Amerindian-Eskimo (470). Similarly, 'sex & marital status' has the following categories and counts: male & divorced (50), female & divorced/married (310), male & single (548), male & married/widowed(92).

[4] $0.001, 0.01, 0.1, 0.5, 1, 10, 100, 1000$

In sum, not only does class imbalance lead to under-representation in non de-biased algorithms, but it also presents important generalisation challenges when estimating and correcting disparity.

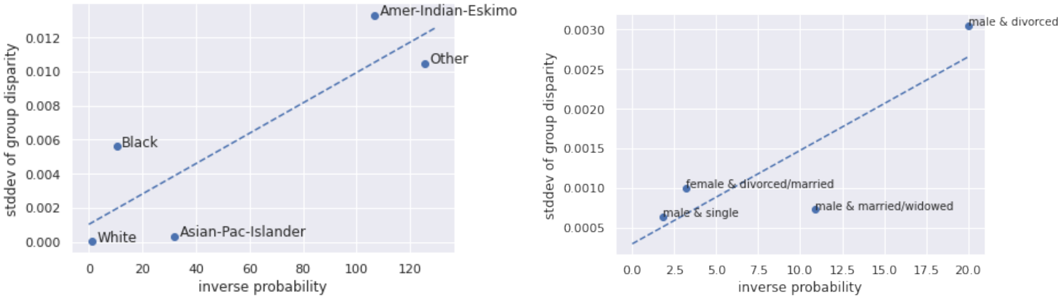

Figure 3: Empirical dependence of individual group disparity standard deviation with respect to class imbalance. *Adult* (left) and *German Credit* (right) datasets with the multicategorical 'race' and 'sex & marital status' protected attributes respectively.

6.5. RQ3: Out-of-Sample Disparity and Model Complexity

Finally, we investigate the impact of model complexity on out-of-sample disparity behaviour. The left plot of Figure 4 depicts a dependence between average overall disparity and its standard deviation. Each point's coordinates are based on results of 5 random initialisations. The size of each point corresponds to the parameter α used in $\phi^{(2)}$, whereas the dotted line is the least squares line through all of the points. Here, the weight decay is 0.1[5] for each of our 5 runs.

The gist from the left plot of Figure 4 is that, by trying to decrease the overall disparity, one also makes it more uncertain out-of-sample, which would suggest that a min-max or fully debiased solution may not be optimal, and partial debiasing may be preferred. Note that this not just an overfitting issue, as the average disparity is also computed on the test set, but an actual trade-off. While the disparity is lower on average on the test set, it also has a higher degree of uncertainty.

In the left plot of Figure 4 we have kept model complexity fixed, but in the right one, we do show how out-of-sample disparity varies with different levels of weight decay. By increasing weight decay levels, we effectively regularise, and as shown in Equation 4 decrease R', that in turn decreases model complexity. We do plot the dependence for several values of α and see the picture, which is consistent with the left plot of Figure 4 - the higher values of α that tend to lead to a smaller average overall disparity is accompanied by a higher standard deviation.

[5] E.g., it corresponds to the parameter γ in the Adam optimizer

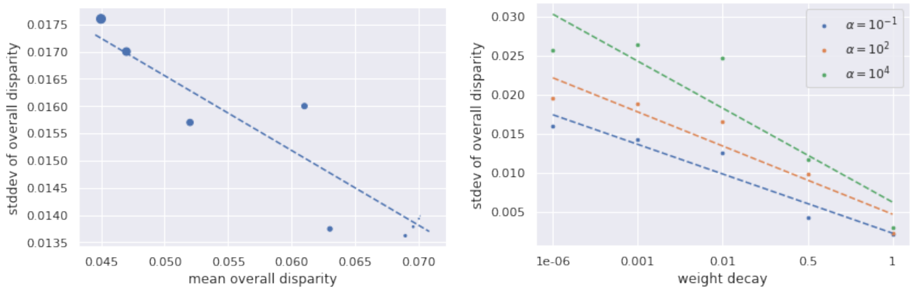

Figure 4: *Left*: Empirical dependence between standard deviation and mean of overall disparity. *Right*: Empirical dependence of standard deviation of overall disparity with respect to the weight decay. *Adult* dataset with 'race' protected attribute.

6.6. Key Observation

These results are important as they signal the necessity to take into account the interplay between class imbalance, model complexity and fairness requirements when designing a fair algorithm. Trade-offs are not limited to accuracy–fairness or fairness–fairness, but also include multiple generalisation "dilemmas". In particular, this implies that attempting to de-bias pre-existing models or simply performing post-processing techniques may encounter such limitations. This is an area for further research.

7. Conclusion

The analysis that we have conducted from two different angles, namely learning with uniform convergence and asymptotics, leads to the same overall qualitative assessment. While we have derived probabilistic upper bounds to prove learning guarantees on the one hand and convergence to a limiting distribution on the other, a striking feature of learning fairness trade-offs is the fundamental difference of regimes between the usual statistical performance criterion measured on the whole sample and the fairness penalty that examines relationships between sub-samples.

Indeed, fairness is not about learning an average distribution, quite the contrary, in the sense that it requires a fine understanding of differences across conditional distributions. Intuitively, in the usual case of an overall statistical loss function, if a category only represents a small portion of the sample, then it also only constitutes a small fraction of the overall loss. On the opposite, from a fairness viewpoint, this implies that it is difficult to make any type of statements regarding the relationship between this particular category and other categories, hence a high risk for a given fairness objective. We have also shown with an empirical investigation that our results translate to practical cases and that small sample sizes or class imbalance can lead to spurious empirical results.

References

Randomized learning and generalization of fair and private classifiers: From pac-bayes to stability and differential privacy. *Neurocomputing*, 416:231–243, 2020. ISSN 0925-2312.

Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Jiahao Chen, Srijan Sood, Sameena Shah, Francois Buet-Golfouse, Bilal A Mateen, and Sebastian Vollmer. Debiasing classifiers: is reality at variance with expectation? 2020.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, March 2003. ISSN 1532-4435.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, August 2018.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: PS*, 9:323–375, 2005. doi: 10.1051/ps:2005018.

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

Irene Y. Chen, Fredrik D. Johansson, and David Sontag. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3543–3554, Red Hook, NY, USA, 2018. Curran Associates Inc.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, June 2017. doi: 10.1089/big.2016.0047.

Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Texts in Statistics. Springer, 1st edition, March 2008. ISBN 0387759700.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3323–3331, December 2016. doi: 10.5555/3157382.3157469. URL https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.

Che-hui Lien I-Cheng Yeh. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, (36), 2009.

Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. Model-agnostic characterization of fairness trade-offs. In *Proceedings of the International Conference on Machine Learning*, volume 37, pages 9339–9349, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 026201825X.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, FAT* 18, New York, USA, 2018.

Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12), 2016.

L. Oneto, M. Donini, M. Pontil, and A. Maurer. Learning fair and transferable representations with theoretical guarantees. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 30–39, 2020.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5680–5689. Curran Associates, Inc., 2017.

Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 142–153, New York, NY, USA, January 2020. Association for Computing Machinery.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Conference on Software Engineering*, pages 1–7, New York, NY, USA, 2018. ACM. doi: 10.1145/3194770.3194776.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness Constraints: Mechanisms for Fair Classification. volume 54 of *Proceedings of Machine Learning Research*, pages 962–970, Fort Lauderdale, FL, USA, 2017. PMLR.

Supplementary Material

In this supplementary material, we provide

1. Applications of fairness to the use cases of disparity and bias,
2. Examples of variance scaling on synthetic data,
3. Proofs of the two main results, Theorem 1 and Theorem 2.

Appendix A. Some Applications of Fair Learning

In this Section, we apply our results to two particular loss functions, namely the Calders-Verwer gap and the Bayes gap.

A.1. Example 1: Learning Disparity

Of particular interest when measuring fairness is the discrepancy that can be observed amongst different groups, for example in terms of a loss L that captures some properties we wish to ascertain. This could be the difference in false negative rates between categories a and a', where a is usually chosen to be a reference group. This concept holds for both the true distribution and the sample distribution.

Definition 4. The (demographic) disparity, or Calders-Verwer gap Calders and Verwer (2010); Chen et al. (2019), between groups a and a' with respect to loss L can be defined as

$$\Delta^L_{T,a,a'}(h) = L_{T,a}(h) - L_{T,a'}(h), \tag{16}$$

for $T \in \{D, S\}$.

Note that it is also quite usual to consider the absolute value of disparity, $\left|\Delta^L_{T,a,a'}(h)\right|$ indicating whether there is equality or not (and magnitude thereof), rather than giving a sign (which relies more on how the baseline category has been chosen). Indeed, trivially, disparity is asymmetric whereas the absolute disparity is symmetric.

It is therefore important to be able to establish the learnability of this widely used metric, which we do in the following proposition.

Proposition 7. *With probability at least $1-\delta$, it holds*

$$\begin{aligned}\sup_{h\in\mathcal{H}}\left|\Delta^L_{D,a,a'}(h) - \Delta^L_{S,a,a'}(h)\right| \quad &\leq \quad 2\left\{R(\ell\circ\mathcal{H}\circ\mathcal{N}_a) + R(\ell\circ\mathcal{H}\circ\mathcal{N}_{a'})\right\} \\ &\qquad + B\sqrt{\frac{2\log\left(\frac{4}{\delta}\right)}{n_a+n_{a'}}}\left[\sqrt{\frac{n_a+n_{a'}}{n_a}} + \sqrt{\frac{n_a+n_{a'}}{n_{a'}}}\right].\end{aligned}$$

Proof We start by observing that by sub-additivity of the sup, we have

$$\sup_{h\in\mathcal{H}}\left|\Delta^L_{D,a,a'}(h) - \Delta^L_{S,a,a'}(h)\right| \leq \sup_{h\in\mathcal{H}}\left|L_{D,a}(h) - L_{S,a}(h)\right| + \sup_{h\in\mathcal{H}}\left|L_{D,a'}(h) - L_{S,a'}(h)\right|.$$

We can then conlude thanks to Proposition 1 and the union bound.

What this upper bound suggests is that in order to learn about disparity, one has to learn the two categories a and a'. If one category has either a large Rademacher complexity or a low count, the upper bound will increase. This result enables us to derive bounds on quantities of interest such as $\left|\Delta^L_{D,a,a'}(h^*_S) - \Delta^L_{S,a,a'}(h^*_S)\right|$ that follows directly from Proposition 7.

A.2. Example 2: Learning Bias

In this Section, we understand bias in a very specific (and non-fairness related) way, as in Chen et al. (2018). In short, one can consider bias –in this context– as the difference between the loss incurred by category a when the classifier is determined by minimising the loss function over the entire sample and the loss in category a when the optimal classifier is derived on a standalone basis (i.e., the classifier explicitly takes into the account the attribute a).

If it is possible to use the characteristics s directly in the model, then one can calibrate an optimal classifier on each category:

$$h^*_D(\mathbf{x}, s) = \sum_{a=1}^{C} \mathbf{1}_{\{s=a\}} h^*_{D,a}(\mathbf{x}). \tag{17}$$

However, we usually do not have access to the characteristic s or cannot include it as a feature (for instance to avoid disparate treatment), so that –in general– $h_D(\mathbf{x}, s) = h^*_D(\mathbf{x})$ for all $s = 1, \cdots, C$. Consequently, we can define the Bayes gap as the difference between the loss obtained between the

Definition 5. The *Bayes gap* for category a is given by

$$\Gamma_{T,a} = L_{T,a}(h^*_S) - L_{T,a}(h^*_{S,a}), \tag{18}$$

for $T \in \{D, S\}$.

In particular, $\Gamma_{S,a} \geq 0$. $\Gamma_{\cdot,a}$ represents the added loss incurred by category a due to the consideration of other categories while selecting the classifier h.

Proposition 8. *The true Bayes gap,* $\Gamma_{D,a} = L_{D,a}(h^*_S) - L_{D,a}(h^*_{S,a})$ *can be learnt as*

$$|\Gamma_{D,a} - \Gamma_{S,a}| \leq 4R(\ell \circ \mathcal{H} \circ \mathcal{N}_a) + 2B\sqrt{\frac{2\log(2/\delta)}{2n_a}}, \tag{19}$$

with probability at least $1 - \delta$.

Proof The proof is fairly straightforward and simply decomposes the Bayes gap in terms of easier building blocks:

$$\begin{aligned} L_{D,a}(h^*_S) - L_{D,a}(h^*_{S,a}) &= L_{D,a}(h^*_S) - L_{S,a}(h^*_S) + L_{S,a}(h^*_S) - L_{S,a}(h^*_{S,a}) \\ &\quad + L_{S,a}(h^*_{S,a}) - L_{D,a}(h^*_{S,a}), \end{aligned}$$

leading to

$$\Gamma_{D,a} - \Gamma_{S,a} = L_{D,a}(h^*_S) - L_{S,a}(h^*_S) + L_{S,a}(h^*_{S,a}) - L_{D,a}(h^*_{S,a}).$$

We thus obtain that

$$\begin{aligned}|\Gamma_{D,a} - \Gamma_{S,a}| &\leq 2 \sup_{h \in \mathcal{H}} |L_{D,a}(h) - L_{S,a}(h)| \\ &\leq 4R(\ell \circ \mathcal{H} \circ \mathcal{N}_a) + 2B\sqrt{\frac{2\log(2/\delta)}{2n_a}},\end{aligned}$$

where the second inequality holds with probability at least $1 - \delta$ by Proposition 1.

Appendix B. Out-of-Sample Variance Scaling

In this Section, we consider a surrogate example to illustrate some real-world consequences of Theorem 2. The behaviour of a number of debiasing methods, both in- and out-of-sample, has been studied on real and surrogate data in the literature Zafar et al. (2017); Agrawal et al. (2020); Oneto et al. (2020) for instance, we thus only provide an illustration of our result on a simple test case, to highlight the role played by sampling variance in dealing with fairness metrics. In keeping with the previous section, we suppose that we have chosen a given classifier $\widehat{y} = h(\mathbf{x}, s)$ and are now considering its statistical properties on a (large) out-of-sample dataset.

B.1. Predictive Parity and Its Limiting Variance

In our simplified experimental setup, we consider $C = 2$, and set $\ell^0 = 0$ with $\lambda = 1$, i.e., we only look at a fairness criterion, and choose a ratio τ_{PP} linked to predictive parity (see Table 1):

$$\phi_D(h) = \tau_{D,\text{PP}} := \frac{\mathbb{P}(\widehat{y} = 1 | y = 1, s = 1)}{\mathbb{P}(\widehat{y} = 1 | y = 1, s = 2)}. \tag{20}$$

The perfectly fair case corresponds to $\tau_{D,\text{PP}} = 1$. We also define the corresponding sample version of this criterion, using the samples $\mathcal{N}_a^+$, $a = 1, 2$:

$$\phi_S(h) = \tau_{S,\text{PP}} = \frac{\frac{1}{n_1^+}\sum_{i \in \mathcal{N}_1^+} \mathbf{1}_{\{\widehat{y}_i = 1\}}}{\frac{1}{n_2^+}\sum_{i \in \mathcal{N}_2^+} \mathbf{1}_{\{\widehat{y}_i = 1\}}}. \tag{21}$$

In this case, one can compute $\mathbb{V}_{\text{lim}}(h)$ directly as

$$\mathbb{V}_{\text{lim}}(h) = \frac{1 - \mathbb{P}(\widehat{y} = 1 | y = 1, s = 1)}{\mathbb{P}(\widehat{y} = 1 | y = 1, s = 1)} \frac{1}{\pi_1} + \frac{1 - \mathbb{P}(\widehat{y} = 1 | y = 1, s = 2)}{\mathbb{P}(\widehat{y} = 1 | y = 1, s = 2)} \frac{1}{\pi_2}, \tag{22}$$

where $\pi_1 = \frac{p_1^+}{p_1^+ + p_2^+}$ and $\pi_2 = 1 - \pi_1 = \frac{p_2^+}{p_1^+ + p_2^+}$.

B.2. Simulation Set-Up and Results

In our simulations, we use the fixed values $\mathbb{P}(\widehat{y} = 1 | y = 1, s = 1) = 0.95$ and $\mathbb{P}(\widehat{y} = 1 | y = 1, s = 2) = 0.99$, leading to $\tau_{D,\text{PP}} = 0.96$. We vary the parameters $m = n_1^+ + n_2^+$, which represents the overall sample size available to us to compute $\tau_{S,\text{PP}}$, and π_1, which is the

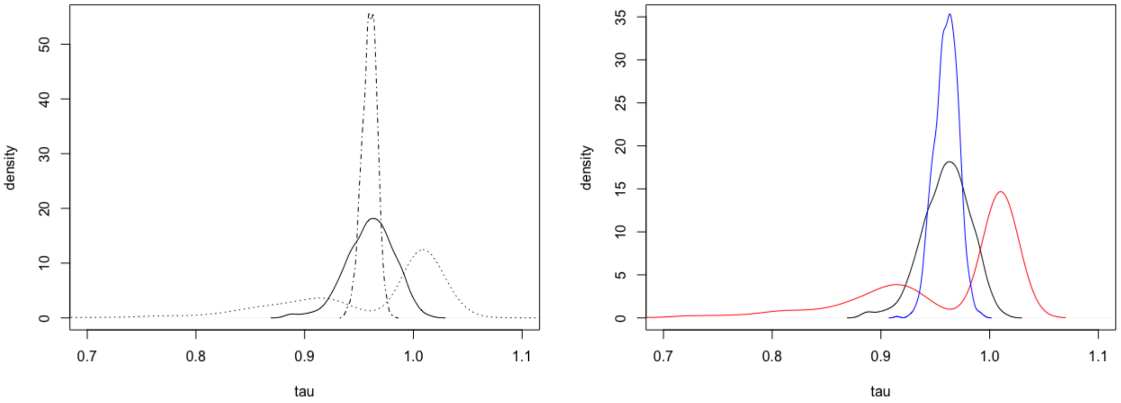

Figure 5: **(a)** Empirical densities of $\tau_{S,\text{PP}}$ for $m = 100$ (*dotted*), 1000 (*black*) and 10000 (*dashed*) at $\pi_1 = 10\%$. **(b)** Empirical densities of $\tau_{S,\text{PP}}$ for $\pi_1 = 1\%$ (*red*), 10% (*black*) and 50% (*blue*) at $m = 1000$.

effective proportion of class 1 versus class 2. Throughout this exercise, we simulate $N = 1000$ times the experiment. Our baseline case is $m = 1,000$ and $\pi_1 = 10\%$.

From Figure 5 (a), one recovers the usual behaviour due to sample size: as m increases, the empirical distribution of $\tau_{S,\text{PP}}$ becomes more peaked. A small sample size (such as $m = 100$) would yield poor accuracy and one can observe that the density is actually bi-modal.

The behaviour with respect to sample size can also be checked by comparing the empirical distribution of $\frac{\tau^{(k)}_{S,\text{PP}} - \tau_{D,\text{PP}}}{\sqrt{\mathbb{V}_{\text{lim}}(h)}}$, where $k = 1, \cdots, N$ is the experiment run, with a standard Gaussian distribution, as predicted by Proposition 2. It can be seen in Figure 6 that there is a good adequacy between both distributions.

The most salient feature, however, is the behaviour of $\tau_{S,\text{PP}}$'s distribution with respect to category probability, as described in Figure 5 (b), where one can see that as the sample becomes balanced, the variance of the estimator $\tau_{S,\text{PP}}$ decreases. When π_1 is very small, say 1%, one recovers the bi-modal distribution that comes with a small sample.

B.3. Key Observation

In a nutshell, low sample sizes and class imbalance can lead to poor generalisation. Indeed, by considering the cases $m = 100, \pi_1 = 10\%$ and $m = 1000, \pi_1 = 1\%$, it is clear that one cannot draw any conclusion regarding the presence of bias nor in terms of which group is the advantaged group. Bi-modality is a particularly interesting characteristic of the empirical distribution: there is a mode strictly below 1 and a mode strictly above 1.

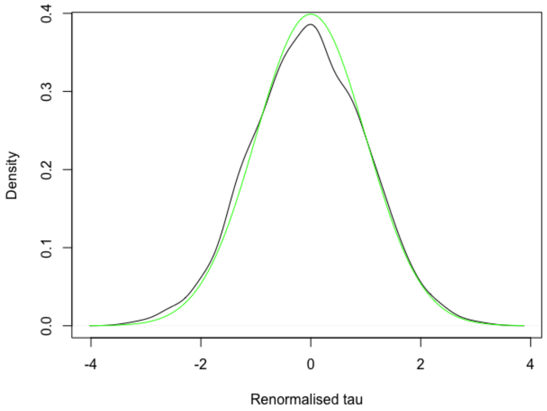

Figure 6: Normalised density of $\tau_{D,\mathrm{PP}}$ (in *black*) and theoretical limiting distribution (in *green*) for $m = 1,000$ and $\pi_1 = 10\%$.

Appendix C. Proofs

C.1. Proof of Proposition 3

Proposition 3 *Under the assumption that ϕ is K^ϕ-Lipschitz, it holds, with probability at least $1-\delta$, and where $Z_S := \sum_{a=1}^C \sqrt{\frac{n}{n_a^+}} + \sqrt{\frac{n}{n_a^-}}$, that*

$$\begin{aligned}\sup_{h\in\mathcal{H}} |\phi_D(h) - \phi_S(h)| \quad &\leq \quad 2K^\phi \sum_{a=1}^C R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-) \\ &\quad + K^\phi B Z_S \sqrt{\frac{2\log\frac{4C}{\delta}}{n}}.\end{aligned}$$

Proof Using the hypothesis that ϕ is K^ϕ-Lipschitz, it comes

$$|\phi_D(h) - \phi_S(h)| \leq K^\phi \sum_{a=1}^C \left| L_{D,a}^+(h) - L_{S,a}^+(h) \right| + \left| L_{D,a}^-(h) - L_{S,a}^-(h) \right|$$

For any $\delta' \in (0,1)$ and any $a = 1, \cdots, C$ it holds

$$\sup_{h\in\mathcal{H}} \left| L_{D,a}^\pm(h) - L_{S,a}^\pm(h) \right| \leq 2R(\ell^\pm \circ \mathcal{H} \circ \mathcal{N}_a^\pm) + B\sqrt{\frac{2\log\frac{2}{\delta'}}{n_a^\pm}},$$

whereby

$$\begin{aligned}\sup_{h\in\mathcal{H}} |\phi_D(h) - \phi_S(h)| \quad &\leq \quad 2K^\phi \sum_{a=1}^C R(\ell^+ \circ \mathcal{H} \circ \mathcal{N}_a^+) + R(\ell^- \circ \mathcal{H} \circ \mathcal{N}_a^-) \\ &\quad + K^\phi B \sum_{a=1}^C \sqrt{\frac{2\log\frac{2}{\delta'}}{n_a^+}} + \sqrt{\frac{2\log\frac{2}{\delta'}}{n_a^-}},\end{aligned}$$

with probability at least $1 - 2C\delta'$, thanks to the union bound. To have $1 - \delta = 1 - 2C\delta'$, one can simply pick $\delta' = \delta/(2C)$. Now, we can write the last term as

$$\begin{aligned}\sum_{a=1}^{C} \sqrt{\frac{2\log\frac{2}{\delta'}}{n_a^+}} + \sqrt{\frac{2\log\frac{2}{\delta'}}{n_a^-}} &= \sqrt{\frac{2\log\frac{2}{\delta'}}{n}} \sum_{a=1}^{C} \sqrt{\frac{n}{n_a^+}} + \sqrt{\frac{n}{n_a^-}} \\ &= Z_S \sqrt{\frac{2\log\frac{2}{\delta'}}{n}},\end{aligned}$$

hence the final result.

C.2. Proof of Theorem 1

Theorem 1 *With probability at least* $1 - \delta$, *it holds*

$$0 \leq \mathcal{L}_D(h_S^*) - \mathcal{L}_D(h_D^*) \leq UB_{\mathcal{S}} + B\left(1 + \lambda Z_S K^{\phi}\right) \frac{3}{\sqrt{2}} \sqrt{\frac{\log(16C/\delta)}{n}}, \tag{23}$$

Proof For the sake of clarity, we divide the proof into different steps.

Step 1. We can start by rewriting

$$\begin{aligned}\mathcal{L}_D(h_S^*) - \mathcal{L}_D(h_D^*) &= \mathcal{L}_D(h_S^*) - \mathcal{L}_S(h_S^*) + \mathcal{L}_S(h_S^*) - \mathcal{L}_S(h_D^*) + \mathcal{L}_S(h_D^*) - \mathcal{L}_D(h_D^*) \\ &\leq \mathcal{L}_D(h_S^*) - \mathcal{L}_S(h_S^*) + \mathcal{L}_S(h_D^*) - \mathcal{L}_D(h_D^*),\end{aligned}$$

since, by definition, $\mathcal{L}_S(h_S^*) - \mathcal{L}_S(h_D^*) \leq 0$.

Step 2. Now, $\mathcal{L}_D(h_S^*) - \mathcal{L}_S(h_S^*) \leq \sup_{h\in\mathcal{H}} |\mathcal{L}_D(h) - \mathcal{L}_S(h)|$ on the one hand, and, on the other hand, we have

$$\begin{aligned}|\mathcal{L}_S(h_D^*) - \mathcal{L}_D(h_D^*)| \leq\ & \left|L_S^0(h_D^*) - L_D^0(h_D^*)\right| \\ & + \lambda K^{\phi} \sum_{a=1}^{C} \left\{ \left|L_{D,a}^+(h_D^*) - L_{S,a}^+(h_D^*)\right| + \left|L_{D,a}^-(h_D^*) - L_{S,a}^-(h_D^*)\right| \right\}.\end{aligned}$$

Step 3. The key insight is to notice that the right-hand side only involves h_D^*, which does *not* depend on the sample $\mathcal{S}$, hence one can apply the Hoeffding inequality directly. Thus, by Hoeffding's inequality, for each a, with probability at least $1 - \delta''$,

$$\left|L_{D,a}^{\pm}(h_D^*) - L_{S,a}^{\pm}(h_D^*)\right| \leq B\sqrt{\frac{\log(2/\delta'')}{2n_a^{\pm}}}.$$

Similarly, with probability at least $1 - \delta''$,

$$\left|L_S^0(h_D^*) - L_D^0(h_D^*)\right| \leq B\sqrt{\frac{\log(2/\delta'')}{2n}}.$$

We can then infer that with probability at least $1-(2C+1)\delta''$

$$\begin{aligned} |\mathcal{L}_S(h_D^*) - \mathcal{L}_D(h_D^*)| &\le B\sqrt{\frac{\log(2/\delta')}{2n}} \\ &\quad +\lambda B K^{\phi} \sum_{a=1}^{C} \sqrt{\frac{\log(2/\delta'')}{2n_a^+}} + \sqrt{\frac{\log(2/\delta'')}{2n_a^-}} \\ &\le B\left(1+\lambda Z_S K^{\phi}\right)\sqrt{\frac{\log(2/\delta'')}{2n}}. \end{aligned}$$

Step 4. From the previous proposition, it comes that with probability at least $1-\delta'$

$$\sup_{h\in\mathcal{H}} |\mathcal{L}_D(h) - \mathcal{L}_S(h)| \le \mathrm{UB}_{\mathcal{S}} + B\left(1+\lambda Z_S K^{\phi}\right)\sqrt{\frac{2\log(8C/\delta')}{n}}.$$

Finally, by choosing $\delta' = \delta/2$ and $\delta'' = \delta/(2(2C+1))$,

$$\begin{aligned} \mathcal{L}_D(h_S^*) - \mathcal{L}_D(h_D^*) &\le \mathrm{UB}_{\mathcal{S}} + B\left(1+\lambda Z_S K^{\phi}\right)\left[\sqrt{\frac{2\log(16C/\delta)}{n}} + \sqrt{\frac{\log(4(2C+1)/\delta)}{2n}}\right] \\ &\le \mathrm{UB}_{\mathcal{S}} + B\left(1+\lambda Z_S K^{\phi}\right)\frac{3}{\sqrt{2}}\sqrt{\frac{\log(16C/\delta)}{n}}, \end{aligned}$$

with probability at least $1-\delta$.

C.3. Proof of Theorem 2

Theorem 2 *Let $h \in \mathcal{H}$, under the assumption that $p_a^{\pm} > 0$ for all a, and that $\nabla\phi_D(h)$ exists and is non-zero, we have the following convergence in distribution*

$$\sqrt{n}\left(\mathcal{L}_S(h) - \mathcal{L}_D(h)\right) \xrightarrow[n\to+\infty]{d} N\left(0, \mathbb{V}_{\lim}(h)\right), \tag{24}$$

where

$$\begin{aligned} \mathbb{V}_{\lim}(h) &= \sum_{a=1}^{C} p_a^+\left(\sigma_a^{0,+}\right)^2 + \lambda^2\frac{(K_a^+)^2}{p_a^+}\left(\sigma_a^+\right)^2 + 2\lambda K_a^+ \, cov_a^+\left(\ell^0(h,z), \ell^+(h,z)\right) \\ &\quad + \sum_{a=1}^{C} p_a^-\left(\sigma_a^{0,-}\right)^2 + \lambda^2\frac{(K_a^-)^2}{p_a^-}\left(\sigma_a^-\right)^2 + 2\lambda K_a^- \, cov_a^-\left(\ell^0(h,z), \ell^-(h,z)\right) \\ &\quad + \sum_{a=1}^{C} p_a^+(1-p_a^+)\left(L_{D,a}^{0,+}\right)^2 + p_a^-(1-p_a^-)\left(L_{D,a}^{0,-}\right)^2 - 2p_a^+p_a^- L_{D,a}^{0,+} L_{D,a}^{0,-} \\ &\quad - \sum_{a\neq a'} \Big(p_a^+ p_{a'}^+ L_{D,a}^{0,+} L_{D,a'}^{0,+} + p_a^- p_{a'}^- L_{D,a}^{0,-} L_{D,a'}^{0,-} \\ &\quad + p_a^+ p_{a'}^- L_{D,a}^{0,+} L_{D,a'}^{0,-} + p_a^- p_{a'}^+ L_{D,a}^{0,-} L_{D,a'}^{0,+} \Big) \end{aligned}$$

Proof For the sake of clarity, we divide this simple (but slightly tedious) proof into multiple steps. In this proof, we make the assumption that ϕ is differentiable with continuous gradient; this is not necessary but simplifies the exposition. The reader is referred to DasGupta (2008) for details regarding the delta method, the central limit theorem and different types of convergence.

Step 1. We start by rewriting the difference between both true and empirical trade-offs $\mathcal{L}_S$ and $\mathcal{L}_D$ for a given $h \in \mathcal{H}$. Thanks to the Taylor formula, there exists $\xi \in [0, 1]$ such that

$$\begin{aligned}\phi_S(h) - \phi_D(h) &= \phi(\mathbf{L}_S^{\pm}(h)) - \phi(\mathbf{L}_D^{\pm}(h)) \\ &= \nabla\phi(\xi\mathbf{L}_D^{\pm}(h) + (1-\xi)\mathbf{L}_S^{\pm}(h))^T \left(\mathbf{L}_S^{\pm}(h) - \mathbf{L}_D^{\pm}(h)\right)\end{aligned}$$

We denote $K_{S,a}^{\pm}$ the partial differential of ϕ with respect to $x_a^{\pm}$, i.e., $K_{S,a}^{\pm} = \partial_{x_a^{\pm}}\phi(\xi\mathbf{L}_D^{\pm}(h) + (1-\xi)\mathbf{L}_S^{\pm}(h))^T \left(\mathbf{L}_S^{\pm}(h) - \mathbf{L}_D^{\pm}(h)\right)$. Some algebra then yields

$$\begin{aligned}&\sqrt{n}\left(\mathcal{L}_S(h) - \mathcal{L}_D(h)\right) \\ &\quad= \sum_{a=1}^{C} \sqrt{\frac{n}{n_a^+}}\sqrt{n_a^+}\left(p_a^+\left[L_{S,a}^{0,+}(h) - L_{D,a}^{0,+}(h)\right] + \lambda K_{S,a}^+\left[L_{S,a}^+(h) - L_{D,a}^+(h)\right]\right) \\ &\quad+ \sum_{a=1}^{C} \sqrt{\frac{n}{n_a^-}}\sqrt{n_a^-}\left(p_a^-\left[L_{S,a}^{0,-}(h) - L_{D,a}^{0,-}(h)\right] + \lambda K_{S,a}^-\left[L_{S,a}^-(h) - L_{D,a}^-(h)\right]\right) \\ &\qquad\qquad+ \sum_{a=1}^{C}\sqrt{n}\left(\frac{n_a^+}{n} - p_a^+\right)L_{S,a}^{0,+}(h) + \sum_{a=1}^{C}\sqrt{n}\left(\frac{n_a^-}{n} - p_a^-\right)L_{S,a}^{0,-}(h).\end{aligned}$$

Step 2. Since $\lim_{n\to+\infty}\frac{n_a^{\pm}}{n} = p_a^{\pm} > 0$ almost surely, this implies trivially that $\lim_{n\to+\infty} n_a^{\pm} = +\infty$ almost surely, for any a. We now recall that, thanks to the continuous mapping theorem, it holds that for any $n_a^{\pm}$, $\lim_{n\to+\infty}\sqrt{n/n_a^{\pm}} = 1/\sqrt{p_a^{\pm}}$ almost surely. Finally, $\lim_{n\to+\infty} K_{S,a}^{\pm} = K_{D,a}^{\pm}$ almost surely, since $\nabla\phi$ is continuous and $\lim_{n\to+\infty}\mathbf{L}_S^{\pm}(h) = \mathbf{L}_D^{\pm}(h)$ almost surely.

Step 3. Let us now consider each term $a^{\pm}$ separately.

$$\begin{aligned}&\sqrt{n_a^{\pm}}\left(p_a^{\pm}\left[L_{S,a}^{0,\pm}(h) - L_{D,a}^{0,\pm}(h)\right] + \lambda K_{S,a}^{\pm}\left[L_{S,a}^{\pm}(h) - L_{D,a}^{\pm}(h)\right]\right) \\ &\quad= \sqrt{n_a^{\pm}}\left(\left[p_a^{\pm}L_{S,a}^{0,\pm}(h) + \lambda K_{D,a}^{\pm}L_{S,a}^{\pm}(h)\right] - \left[p_a^{\pm}L_{D,a}^{0,\pm}(h) + \lambda K_{D,a}^{\pm}L_{D,a}^{\pm}(h)\right]\right) \\ &\qquad\qquad+ \lambda\left(K_{S,a}^{\pm} - K_{D,a}^{\pm}\right)\sqrt{n_a^{\pm}}\left[L_{S,a}^{\pm}(h) - L_{D,a}^{\pm}(h)\right].\end{aligned}$$

By the Central Limit Theorem (CLT), $\sqrt{n_a^{\pm}}\left[L_{S,a}^{\pm}(h) - L_{D,a}^{\pm}(h)\right] \xrightarrow[n\to+\infty]{d} N(0, [\sigma_a^{\pm}]^2)$, but since $\left(K_{S,a}^{\pm} - K_{D,a}^{\pm}\right) \xrightarrow[n\to+\infty]{a.s.} 0$, the whole term $\left(K_{S,a}^{\pm} - K_{D,a}^{\pm}\right)\sqrt{n_a^{\pm}}\left[L_{S,a}^{\pm}(h) - L_{D,a}^{\pm}(h)\right]$ goes to 0 almost surely.

All that remains is to apply the CLT to

$$\sqrt{n_a^{\pm}}\left(\left[p_a^{\pm}L_{S,a}^{0,\pm}(h) + \lambda K_{D,a}^{\pm}L_{S,a}^{\pm}(h)\right] - \left[p_a^{\pm}L_{D,a}^{0,\pm}(h) + \lambda K_{D,a}^{\pm}L_{D,a}^{\pm}(h)\right]\right),$$

and notice that $\mathbb{V}_{s=a,y=\pm1}\left[p_a^{\pm}\ell^{0,\pm}(h,z)+\lambda K_{D,a}^{\pm}\ell^{\pm}(h,z)\right]=[p_a^{\pm}]^2[\sigma_a^{0,\pm}]^2+\lambda^2[K_{D,a}^{\pm}]^2[\sigma_a^{\pm}]^2+2p_a^{\pm}\lambda K_{D,a}^{\pm}\mathrm{cov}_a^{\pm}\left(\ell^0(h,z),\ell^{\pm}(h,z)\right)$.

Note that since each data point in the sample is independent, the CLT can be applied term by term (i.e., all $a^{\pm}$'s have independent limiting distributions).

Step 4. To finish the proof, it simply remains to consider the adjustments due to observing empirical proportions rather than the true class probabilities. We thus need to consider the limiting distribution of the vector $\sqrt{n}\left(\frac{n_a^{\pm}}{n}-p_a^{\pm}\right)_{\pm,a=1,\cdots,C}$. But this is simply the limiting distribution of empirical proportions in repeated trials of a multinoulli distribution, hence the result.

Let us point out that the non-vanishing assumption on the gradient of the fairness loss function ϕ can be relaxed, but leads to the need for a higher-order delta method (see DasGupta (2008), Section 3.4 and results therein).