A Comprehensive View of Fairness through Distributional Stability

Gayane Taturyan
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
gayane.taturyan@telecom-paris.fr

Charlotte Laclau
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
charlotte.laclau@telecom-paris.fr

Stephan Clémençon
LTCI, Télécom Paris
Institut Polytechnique de Paris, France
stephan.clemencon@telecom-paris.fr

Abstract

We view fairness as a property of distributional stability. Rather than assessing a predictor under a fixed data distribution, we study how its predictions change under perturbations that modify the composition of protected groups. A predictor is fair if it remains stable under such shifts. Under this perspective, several classical notions of fairness arise as stability with respect to specific perturbations, with the associated unfairness gap given by a Lipschitz constant of a prediction-rate functional. This formulation also yields guarantees that hold uniformly over a range of demographic compositions at test time, without requiring knowledge of the deployment distribution. It leads to a learning procedure based on convex combinations of reweighted predictors, formulated as a second-order cone program, for which we establish generalization bounds. Experiments on standard benchmarks validate our theoretical findings, demonstrate the competitiveness of our approach in terms of fairness, and highlight its distinctive deployment guarantees relative to existing baselines.

1 Introduction

Fairness in the field of machine learning has been studied primarily from two distinct perspectives, leading to the introduction of two broad categories of definitions. *Group fairness*, on the one hand, requires that predictive outcomes be statistically equal or comparable across groups defined by sensitive characteristics such as gender or ethnic origin, which has given rise to a large number of competing metrics, including demographic parity, equal opportunity, or equalized odds, among others [8, 21]. On the other hand, the concept of *individual fairness* stipulates as a mandatory condition that similar individuals be assigned similar predictive outcomes, which is formalized by a Lipschitz condition with respect to a task-specific metric [16]. These two approaches rest on different foundations and have been studied largely in isolation.

We also note that both of these approaches evaluate fairness relative to a fixed data distribution. In practice, a model trained on one population is often deployed on another, whether in different regions, at different times, or in different institutional contexts. Recent work has begun to investigate how fairness guarantees degrade under such changes [4, 19, 23], but these approaches are either tailored to specific shift models, require knowledge of the deployment distribution at training time, or model the shift as acting on the feature space while leaving the distribution of the sensitive attribute G itself

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

unchanged. Our framework takes a different perspective: we consider perturbations that act directly on the marginal distribution of G, which is the primary driver of deployment bias in practice.

In this article, we propose to approach fairness from the perspective of stability — specifically, the distributional stability of a given prediction function rather than the algorithmic stability derived from learning theory. A classifier is fair if its behavior does not change in the face of perturbations in the data distribution that reflect social considerations, such as changes in the composition of protected groups. This single principle links the concepts of fairness at the individual level and at the group level, which appear as special cases corresponding to different classes of perturbations. This naturally resolves the issue of context shifts during deployment: a classifier trained to remain stable in the face of variations in group distribution inherits fairness guarantees that apply uniformly to all demographic compositions at deployment time, without the need to know the deployment distribution in advance.

Contributions. We define fairness of a predictive rule as a stability property (Section 2). We then establish formal equivalences between individual fairness, demographic parity, equal opportunity, and stability under specific perturbation classes, showing that the unfairness gap is in each case a Lipschitz constant (Section 3). For group fairness, we unify these results into a single optimization problem via a constraint functional measuring deviation between predictions and a target level across deployment contexts; its decomposition ensures guarantees for both fairness and uniform deployment. Restricting to convex mixtures of reweighted classifiers turns the empirical problem into a second-order cone program, for which we establish generalization bounds (Section 4). Numerical experiments confirm the theory's practical applicability (Section 5). Proofs, the Equalized Odds extension, and additional experiments and discussions are deferred to the Supplementary Material.

Related works. Algorithmic fairness, especially for binary classification, has been extensively studied from both theoretical and methodological perspectives. A large part of this literature considers *aware* setting, in which the sensitive attribute is available at the prediction time [8, 10, 12, 18, 20, 34, 35], whereas fewer works address the *unaware* setting [1, 9, 11, 22, 27, 28], in which it is not. A number of works aim to unify fairness-constrained classification within a common framework. A classical example is the reductions approach of Agarwal et al. [1], where the authors reduce fair classification to a sequence of cost-sensitive classification problems and construct a randomized classifier with minimum empirical error under the prescribed constraint. A recent work of Chzhen et al. [11] proposes a unified post-processing framework for randomized classification under general system-level constraints (including group-fairness constraints), based on entropic regularization and stochastic optimization. The stability-based perspective is related to recent work on the robustness of fairness-constrained learning under adversarial perturbations [6], that shows how malicious noise in the training data can significantly affect the validity of fairness constraints. In our work we define fairness itself as a stability under perturbations. Despite an initial resemblance to distributionally robust optimization approaches, our framework is based on specific perturbations induced by a prescribed perturbation rule and admits an importance-weight representation. Unlike DRO, this structural specificity yields non-vacuous deployment guarantees, as we discuss in Appendix A.2.

Notation. For an integer $K > 0$, we define $[K] \stackrel{\text{def}}{=} \{1, \cdots, K\}$ and write Δ_K for the probability simplex in $\mathbb{R}^K$. The notation $\widetilde{\mathcal{O}}$ hides unimportant constants and logarithmic terms.

2 Defining fairness as a property of stability

The concept of stability is ubiquitous in the sciences, particularly in mathematics, as it describes how a "system", such as a machine-learning algorithm, responds to changes in the parameters or data used to define it. The various notions of stability and their applications are far too numerous to list here and we refer to Appendix A.1 for an in-depth description in the machine-learning context. Instead of the main concepts of a predictor's fairness introduced in the literature, we propose a definition based on the concept of stability under changes in the data distribution. As we will see in the following sections, this new definition allows us to subsume the previous ones within a unified formal framework.

2.1 Background and preliminaries

Let us consider the generic problem of predictive learning with fairness concerns regarding sensitive attributes G valued in $\mathcal{G}$, where the random label Y takes values in a space $\mathcal{Y}$ equipped with a metric $d_{\mathcal{Y}} : \mathcal{Y}^2 \to [0, \infty)$, and the predictive information is modelled by a random vector $\boldsymbol{X}$, valued in a

high-dimensional metric space $(\mathcal{X}, d_{\mathcal{X}})$. By $\boldsymbol{z} = (\boldsymbol{x}, g, y)$ is meant a realization in the product space $\mathcal{Z} = \mathcal{X} \times \mathcal{G} \times \mathcal{Y}$ of the random triplet $\boldsymbol{Z} = (\boldsymbol{X}, G, Y)$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with joint distribution $\mathcal{D}$ describing the system under study, by $h : \mathcal{X} \to \mathcal{Y}$ an arbitrary (measurable) predictive rule in *unaware* setting. Equipped with these notations, let us now review the standard definitions that aim to formalize algorithmic fairness. Individual fairness (IF), introduced by Dwork et al. [16], which requires that similar individuals receive similar predictions, is defined as follows.

Definition 1 (Individual Fairness)**.** *A predictor $h : \mathcal{X} \to \mathcal{Y}$ satisfies the IF property when it satisfies the $(d_{\mathcal{Y}}, d_{\mathcal{X}})$-Lipschitz condition for some constant $L < \infty$:*

$$\forall (\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{X}^2, \quad d_{\mathcal{Y}}(h(\boldsymbol{x}), h(\boldsymbol{x}')) \leqslant L \cdot d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}') \,.$$

The concept of group fairness differs in that it requires measures of performance to be identical across groups defined by the sensitive attribute G. Depending on the type of predictive problem considered (e.g., classification, regression, ranking) and the criterion chosen to measure performance, it can take various forms. Among those introduced in the literature, below we highlight the following two major notions for binary classification (i.e., $\mathcal{Y} = \{0, 1\}$), that will serve as main examples in the subsequent analysis. For simplicity, the definitions are given for only two sensitive groups, i.e., $\mathcal{G} = \{0, 1\}$. They can be immediately extended to the case where the sensitive attribute G takes more than two values.

Definition 2 (Demographic Parity)**.** *A predictor $h : \mathcal{X} \to \{0, 1\}$ satisfies* Demographic Parity (DP) *if its predictions are independent of G: $\mathbb{P}(h(\boldsymbol{X}) = 1 \mid G = 0) \;=\; \mathbb{P}(h(\boldsymbol{X}) = 1 \mid G = 1)$.*

The concept of equal opportunity (EO) requires that there be no disparity with respect to a specific aspect of predictive performance.

Definition 3 (Equal Opportunity)**.** *A classifier h satisfies the EO property if it achieves equal true positive rates across groups: $\mathbb{P}(h(\boldsymbol{X}) = 1 \mid Y = 1, G = 0) \;=\; \mathbb{P}(h(\boldsymbol{X}) = 1 \mid Y = 1, G = 1)$.*

The concept of Equalized Odds (EOdd) strengthens the EO property by requiring the equality to hold conditionally on both values of the label; its definition is recalled in Appendix D.

2.2 A new definition of fairness for a predictive rule

We now propose to define the fairness of a predictor as a *stability* property. The concept of stability is already widely used in statistical learning theory, primarily to account for the sensitivity of a learning algorithm to changes in the training examples [7, 24]. An algorithm is said to be stable if replacing a few training data points does not significantly change the learned function. The framework we are developing here is of a different nature. We consider a fixed prediction function h and seek to quantify how its outputs change as the distribution of sensitive attributes G varies. Unlike *domain adaptation* [29], which aims to correct such changes, we characterize predictors whose behavior remains stable in the face of these perturbations, and show that classical notions of fairness naturally follow from the requirement for this specific type of stability. Consider a class $\mathcal{P}$ of (perturbation) operators $\mathbf{P} : D \subset \mathcal{M}_1(\mathcal{Z}) \to \mathcal{M}_1(\mathcal{Z})$ acting on the elements $\mathcal{D}$ of a domain $\mathcal{M}$ included in the set $\mathcal{M}_1(\mathcal{Z})$ of probability measures on $\mathcal{Z}$ to produce a perturbed probability measure $\mathbf{P}(\mathcal{D})$ on $\mathcal{Z}$. Since only the values taken by the variable $h(X)$ are relevant to the definition of individual fairness, we set $\mathcal{Z}$ equal to $\mathcal{X}$ in this case. With this in mind, we propose the following definition, inspired by the formalism of PAC learning.

Definition 4 (Fairness as Distributional Stability)**.** *Consider a class $\mathcal{P}$ of perturbation operators acting on a domain $D \subset \mathcal{M}_1(\mathcal{Z})$ and a prediction function $h : \mathcal{X} \to \mathcal{Y}$. Let $\varepsilon > 0$ and $\Delta(h;\ (\mathcal{D}, \mathbf{P}))$ be a statistical distance between the distributions of the prediction $h(X)$ under $\mathcal{D} \in \mathcal{M}$ and under $\mathbf{P}(\mathcal{D})$. The prediction function h is said to be ε–stable w.r.t. $\mathcal{P}$ and Δ if and only if*

$$\sup_{(\mathcal{D}, \mathbf{P}) \in \mathcal{M} \times \mathcal{P}} \Delta(h;\ (\mathcal{D}, \mathbf{P})) \leqslant \varepsilon \,.$$

The choice of $\mathcal{P}$ and that of Δ are left intentionally general here; specific and interpretable instances are constructed in Section 3. As we shall see, it is natural to consider perturbations of the distribution of the sensitive attribute G in the context of group fairness.

3 From various existing concepts of fairness to a unified approach

In this section, we show that the general formalism introduced accounts for the various concepts of individual and group fairness for specific choices of the class of $\mathcal{P}$ and the statistical distance Δ.

3.1 Individual fairness and local perturbations

The feature space $\mathcal{X}$ being equipped with a task-specific metric $d_{\mathcal{X}}$, as in Dwork et al. [16], consider the class of transformations of $\mathcal{X}$ indexed by $\rho > 0$:

$$\mathcal{T}_\rho = \{T : \mathcal{X} \to \mathcal{X}; \quad T \text{ measurable} \quad \text{and} \quad \forall \boldsymbol{x} \in \mathcal{X},\ d_{\mathcal{X}}(\boldsymbol{x}, T(\boldsymbol{x})) \leq \rho\}\,. \tag{1}$$

Each transformation $T \in \mathcal{T}_\rho$ is naturally associated with a perturbation operator $\mathbf{P}_T$ acting on the distribution of X and assigning to any probability distribution $\mathcal{D}$ on $\mathcal{X}$ its pushforward by T.

Proposition 1 (IF and Stability)**.** *Let $h : \mathcal{X} \to \mathcal{Y}$ be a prediction function and $d_{\mathcal{Y}} : \mathcal{Y}^2 \to [0, \infty)$ a metric on predictions. The following assertions are equivalent.*

1. The predictive function h satisfies the IF property with $(d_{\mathcal{Y}}, d_{\mathcal{X}})$-Lipschitz constant $L < \infty$.

2. For all $\rho > 0$, the predictive function h is $L\rho$–stable w.r.t. to the class $\mathcal{P}_\rho$ of perturbation operators $\mathbf{P}_T$, $T \in \mathcal{T}_\rho$, defined on the domain $\mathcal{M} = \{\delta_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$ of all point masses and taking $\Delta(h;\ (\mathcal{D}, \mathbf{P}))$ as the Wasserstein 1-distance between the pushforward of $h(\boldsymbol{X})$'s distribution under $\mathcal{D}$ and that under $\mathbf{P}(\mathcal{D})$ based on the cost function $d_{\mathcal{Y}}$:

$$\sup_{(\boldsymbol{x}, T) \in \mathcal{X} \times \mathcal{T}_\rho} d_{\mathcal{Y}}(h(\boldsymbol{x}), h(T(\boldsymbol{x}))) \leqslant L\rho\,. \tag{2}$$

Individual fairness is therefore equivalent to distributional stability in the sense of Definition 4, with a stability parameter $L\rho$ proportional to the perturbation parameter ρ.

3.2 Group fairness and resampling perturbations

We now consider the various notions of group fairness. Rather than perturbing individual feature vectors, we consider perturbations that act on the marginal distribution of the sensitive attribute G (assumed binary for simplicity). The magnitude of the perturbation is parametrized here by the proportion $\pi \in (0, 1)$ of group $G = 1$ in the population, and Δ measures how the prediction rates of h change as this proportion varies. We introduce two types of perturbation. The first resamples G globally, while the second resamples G within the positive label class.

Perturbation 1. *The class $\mathcal{P}_\pi^G$ is composed of all operators $\mathbf{P} : \mathcal{M}_1(\mathcal{Z}) \to \mathcal{M}_1(\mathcal{Z})$ such that, for all distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, 1) the conditional distribution of $(\boldsymbol{X}, Y)$ given G is the same under $\mathcal{D}$ and under $\mathbf{P}(\mathcal{D})$, 2) $\mathbb{P}_{\mathbf{P}(\mathcal{D})}(G = 1) = \pi$.*

Such a class is non empty, since the type 1 perturbations of a distribution $\mathcal{D}$ are in 1-to-1 correspondence with the Bernoulli random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with parameter π.

Perturbation 2. *The class $\mathcal{P}_\pi^{G|1}$ is composed of all operators $\mathbf{P} : \mathcal{M}_1(\mathcal{Z}) \to \mathcal{M}_1(\mathcal{Z})$ such that, for all distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, 1) the conditional distribution of $\boldsymbol{X}$ given (Y, G) and the distribution of Y are the same under $\mathcal{D}$ and under $\mathbf{P}(\mathcal{D})$, 2) $\mathbb{P}_{\mathbf{P}(\mathcal{D})}(G = 1 \mid Y = 1) = \pi$.*

Remark 1 (On perturbations)**.** *In Perturbation 2, we fix $\mathbb{P}(\boldsymbol{X} \mid Y, G)$ rather than the stronger $\mathbb{P}(\boldsymbol{X}, Y \mid G)$ used in Perturbation 1. While $\mathbb{P}(\boldsymbol{X}, Y \mid G)$ implies the invariance of $\mathbb{P}(\boldsymbol{X} \mid Y, G)$, it also implies the invariance of $\mathbb{P}(Y \mid G)$. This, combined with the imposed constraint of $\mathbb{P}_{\mathbf{P}(\mathcal{D})}(G = 1 \mid Y = 1) = \pi$, may be infeasible by Bayes' rule. Fixing only $\mathbb{P}(\boldsymbol{X} \mid Y, G)$ ensures feasibility.*

Given a perturbation operator $\mathbf{P} \in \mathcal{P}_\pi^\square$, with $\square \in \{G, G \mid 1\}$ and a base distribution $\mathcal{D}$, we write $\mathcal{D}_\pi \overset{\text{def}}{=} \mathbf{P}(\mathcal{D})$. We show that group fairness constraints can be expressed by a certain descriptive *functional* that depends on the perturbation setting and is affine w.r.t. the perturbation parameter. Throughout the paper, we reserve the label *Claim* for statements that generalize a pattern observed on the running examples (e.g., DP, EO), under conditions tailored to each perturbation rule.

Definition 5 (Prediction rate functional)**.** *For a prediction function $h : \mathcal{X} \to \{0, 1\}$ and a perturbation parameter $\pi \in [0, 1]$, we define the functional $\Phi(h; \pi) \overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}_\pi}[\varphi_h(\boldsymbol{Z})]$, where $\varphi_h : \mathcal{Z} \to [0, 1]$ is some statistic that is linear w.r.t. h.*

Claim 1. *Let $\mathcal{D}_\pi$ be a perturbed distribution obtained from $\mathcal{D}$ according to a given, appropriate perturbation rule and a fixed perturbation parameter $\pi \in [0, 1]$. The prediction rate functional $\Phi(h; \pi)$ admits a form*

$$\Phi(h; \pi) = \alpha(h) + \pi\gamma(h)\,, \tag{3}$$

where $\alpha(h)$ and $\gamma(h)$ are some prediction rate quantities under the initial (unperturbed) distribution $\mathcal{D}$. Moreover, it holds that

$$|\Phi(h;\pi) - \Phi(h;\pi')| = |\gamma(h)||\pi - \pi'|\,, \quad \forall \pi, \pi' \in [0,1]\,, \text{ and} \tag{4}$$

$$\sup_{\pi,\pi' \in [0,1]} |\Phi(h;\pi) - \Phi(h;\pi')| = |\gamma(h)| \stackrel{\text{def}}{=} \mathcal{U}(h)\,, \tag{5}$$

where $\mathcal{U}(h)$ is the explicit fairness gap (unfairness) for each constraint.

The functional $\Phi(h;\pi)$, although at first obscure, clearly captures the intuition behind each constraint, by being defined as a relevant prediction rate under the perturbed distribution. We now show that DP and EO correspond, respectively, to stability under the introduced perturbations. Throughout, $h : \mathcal{X} \to \{0,1\}$ is a possibly randomized prediction function and $G \in \{0,1\}$ is the sensitive attribute.

Demographic Parity and Stability. Let $\mathcal{D}_\pi$ be a perturbed distribution according to Perturbation 1. We define the *global positive* prediction rate under $\mathcal{D}_\pi$ as $P_h(\pi) \stackrel{\text{def}}{=} \mathbb{P}_\pi(h(\boldsymbol{X}) = 1) = \mathbb{E}_{\mathcal{D}_\pi}[\mathbb{I}\{h(\boldsymbol{X})\}]$ (or $\mathbb{E}_{\mathcal{D}_\pi}[h(\boldsymbol{X})]$ when randomized) and denote by $r_g(h) \stackrel{\text{def}}{=} \mathbb{P}(h(\boldsymbol{X}) = 1 \mid G = g)$.

Proposition 2. *The prediction h satisfies DP if and only if $P_h(\pi)$ is constant in π. Moreover,*

$$P_h(\pi) = r_0(h) + \pi(r_1(h) - r_0(h))\,, \text{ and } \sup_{\pi,\pi' \in [0,1]} |P_h(\pi) - P_h(\pi')| = |r_1(h) - r_0(h)|\,.$$

The deviation from stability coincides exactly with the DP gap. This result also connects DP to sample representativity: if the training dataset has group proportions that differ from the population, resampling perturbations model precisely this shift.

Example 1. *Suppose the model predicts a positive outcome for 80% of individuals in group 0 and for 40% of individuals in group 1, so $r_0(h) = 0.8$ and $r_1(h) = 0.4$. The global positive prediction rate under a population with group prevalence π is $P_h(\pi) = 0.8\,(1-\pi) + 0.4\,\pi = 0.8 - 0.4\,\pi$. Thus, as π varies from 0 to 1, the prediction rate decreases linearly from 0.8 to 0.4, and the maximal deviation is $\sup_{\pi,\pi' \in [0,1]} |P_h(\pi) - P_h(\pi')| = |r_0 - r_1| = 0.4$, which is the DP gap.*

The same analysis applies to intra-class resampling.

Equal Opportunity and Stability. Let $\mathcal{D}_\pi$ be a perturbed distribution according to Perturbation 2. We define the *true positive* prediction rate under $\mathcal{D}_\pi$ as $P_{h|Y=1}(\pi) \stackrel{\text{def}}{=} \mathbb{P}_{\mathcal{D}_\pi}(h(\boldsymbol{X}) = 1 \mid Y = 1) = \mathbb{E}_{\mathcal{D}_\pi}\left[\frac{\mathbb{I}\{h(\boldsymbol{X})\}\mathbb{I}\{Y=1\}}{\mathbb{P}(Y=1)}\right]$ (or $\mathbb{E}_{\mathcal{D}_\pi}\left[\frac{h(\boldsymbol{X})\mathbb{I}\{Y=1\}}{\mathbb{P}(Y=1)}\right]$ when randomized) and denote by $q_g(h) \stackrel{\text{def}}{=} \mathbb{P}(h(\boldsymbol{X}) = 1 \mid G = g, Y = 1)$.

Proposition 3. *The prediction h satisfies EO if and only if $P_{h|Y=1}(\pi)$ is constant in π. Moreover,*

$$P_{h|Y=1}(\pi) = q_0(h) + \pi(q_1(h) - q_0(h))\,, \text{ and } \sup_{\pi,\pi' \in [0,1]} |P_{h|Y=1}(\pi) - P_{h|Y=1}(\pi')| = |q_1(h) - q_0(h)|\,.$$

Claim 2. *Due to linearity of φ_h w.r.t. h (Definition 5), there exists a constant $B > 0$ such that for every measurable function $f : \mathcal{X} \to \mathbb{R}$, it holds that $\mathbb{E}[\varphi_f(\boldsymbol{Z})] \leqslant B\mathbb{E}|f(\boldsymbol{X})|$. For example, in the case of Demographic Parity, we have $\varphi_h(\boldsymbol{X}) = h(\boldsymbol{X})$ and $B = 1$.*

Variance control. It is easy to notice that in all of the examples we have $\sup_{\pi,\pi' \in [0,1]} |\Phi(h;\pi) - \Phi(h;\pi')| = |\Phi(h;1) - \Phi(h;0)|$, meaning the supremum is achieved at the endpoints. This rewrites the problem as the usual "minimization under DP/EO". However, our interest lies within the resampling aspect of the problem. Rather than focusing on the worst-case deviation over all perturbations, we now adopt a stochastic viewpoint, in which perturbations are achieved according to a distribution Π. This allows us to quantify fairness in terms of average sensitivity, leading to smoother and more tractable constraints. We propose to reformulate the constraint through variance control.

Definition 6 (Perturbation variance functional)**.** *Let Π be a probability measure on $[0,1]$ and $\pi \sim \Pi$. We define the perturbation variance functional as $\mathbf{V}_\Pi(h) \stackrel{\text{def}}{=} \operatorname{Var}_{\pi\sim\Pi}(\Phi(h;\pi))$.*

From Definition 5, we have that $\Phi(h;\pi) = \alpha(h) + \pi\gamma(h)$. Using the fact that for $a, b \in \mathbb{R}$ and a random variable X, $\operatorname{Var}(a + bX) = b^2 \operatorname{Var}(X)$, we get $\mathbf{V}_\Pi(h) = \gamma(h)^2 \operatorname{Var}_{\pi\sim\Pi}(\pi)$. This equation above leads to the equivalence $\mathcal{U}(h) \leqslant \varepsilon \iff \mathbf{V}_\Pi(h) \leqslant \varepsilon^2 \operatorname{Var}_{\pi\sim\Pi}(\pi)$.

Links between individual and group fairness. As we established, both individual and group fairness can be equivalently expressed by our notion of stability from Definition 4. Let us emphasise one more connection between the two. The functional $\Phi(h;\pi)$ is *Lipschitz* w.r.t. perturbation parameter π (4), and its Lipschitz constant is exactly the unfairness gap. This provides a complementary angle to view fairness notions: while individual fairness requires similar predictions for similar individuals, formalized by Lipschitz property of the prediction function w.r.t. the feature space, our generalized group-fairness formulation requires stability of perturbation-indexed prediction rates under small changes in the perturbation parameter, that is, Lipschitz continuity of the map $\pi \to \Phi(h;\pi)$. However, despite these conceptual analogies and the possibility of extending our variance-based perspective to individual fairness through suitably randomized local perturbations, the methodologies required to achieve fairness in the two settings remain structurally different. We therefore focus from now on on group fairness, leaving explorations of corresponding methods for individual fairness to future work.

4 Methodology: variance and mean control with mixture models

4.1 Problem formulation

Variance and mean control. Under the introduced perturbation-variance functional constraint, the problem reduces to classical fairness constraints, with a scaling governed by the variance of the perturbation-parameter distribution. This holds even for the empirical problem. Rather than controlling only the variability of $\Phi(h;\pi)$ across perturbations, we consider a stricter objective of requiring $\Phi(h;\pi)$ to remain close to a given target level t.

Definition 7 (Perturbation t-variance functional)**.** *Let Π be a probability measure on $(0,1)$ and $\pi \sim \Pi$. For a $t \in [0,1]$, we define the perturbation t-variance functional as*

$$\mathcal{C}_\Pi(h;t) \stackrel{\text{def}}{=} \mathbb{E}_{\pi\sim\Pi}[(\Phi(h;\pi)-t)^2]\,.$$

Adding and subtracting $\mathbb{E}_{\pi\sim\Pi}[\Phi(h;\pi)]$ inside the square yields the following decomposition:

$$\mathcal{C}_\Pi(h;t) = \mathbf{V}_\Pi(h) + (\mathbb{E}_{\pi\sim\Pi}[\Phi(h;\pi)]-t)^2\,. \tag{6}$$

As the decomposition (6) shows, the functional $\mathcal{C}_\Pi(h;t)$ controls not only the variance term equivalent to fairness, but also the mean level of $\Phi(h;\pi)$ itself. Moreover, one can notice that $\mathbf{V}_\Pi(h) = \inf_{t\in[0,1]} \mathcal{C}_\Pi(h;t)\,,$ with $t^\star = \mathbb{E}_{\pi\sim\Pi}[\Phi(h;\pi)]$. Thus, we can state the following obvious lemma.

Lemma 1. *If h satisfies $\mathcal{C}_\Pi(h;t) \leqslant \delta$ for $t \in [0,1]$, then $\mathbf{V}_\Pi(h) \leqslant \delta$ and $(\mathbb{E}_{\pi\sim\Pi}[\Phi(h;\pi)]-t)^2 \leqslant \delta$.*

Remark 2 (On the choice of t)**.** *We can choose t as desired, or, more naturally, we can set $t = \mathbb{E}_{\pi\sim\Pi}[\Phi(h_0,\pi)]$, where h_0 is an unconstrained baseline. The impact of t is studied in Appendix E.4.*

We introduce the following standard regularity assumption on the *loss*.

Assumption 1. *The loss $\ell : [0,1]\times\mathcal{Y} \to [0,1]$ is convex and L-Lipschitz in its first argument.*

We define the corresponding population *risk* of a predictor $h : \mathcal{X} \to [0,1]$ by $\mathcal{R}(h) \stackrel{\text{def}}{=} \mathbb{E}[\ell(h(\boldsymbol{X}),Y)]$. Given a $t \in [0,1]$, our goal is to solve the following constrained problem

$$\min_{h\in\mathcal{H}} \{\mathcal{R}(h) \,:\, \mathcal{C}_\Pi(h;t) \leqslant \delta\}\,. \tag{$\mathcal{P}$}$$

Convex mixtures. In our framework, fairness is characterized through the behaviour of the prediction function across a family of perturbed distributions. This makes it natural to combine predictors trained on different distributions, especially given that model averaging has already been shown to be effective for enforcing fairness in the classical setting [17].

For $\pi_1,\cdots,\pi_K \stackrel{\text{i.i.d.}}{\sim} \Pi$, let $\eta_1,\cdots,\eta_K$ be (unconstrained) *Bayes regression functions* minimizing the risks on respective distributions $\mathcal{D}_{\pi_1},\cdots,\mathcal{D}_{\pi_K}$, that is $\eta_k \in \arg\min_h \mathcal{R}_{\mathcal{D}_k}(h)\,,$ for $k \in [K]$. For $\boldsymbol{\alpha} \in \Delta_K$, we define a *convex mixture* $\boldsymbol{\eta}_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \sum_{k=1}^K \alpha_k\eta_k$.

Rather than optimizing the true risk of the mixture directly, we focus on its linearized convex surrogate, defined as $\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}) \stackrel{\text{def}}{=} \sum_{k=1}^K \alpha_k\mathcal{R}(\eta_k)$. Since the loss is convex in prediction, Jensen's inequality implies that $\mathcal{R}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}) = \mathcal{R}(\sum_{k=1}^K \alpha_k\eta_k) \leqslant \sum_{k=1}^K \alpha_k\mathcal{R}(\eta_k) = \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}})\,,$ showing that this surrogate upper-bounds the actual risk of the convex mixture. Given a $t \in [0,1]$, the problem ($\mathcal{P}$) translates as

$$\boldsymbol{\alpha}^\star \in \arg\min_{\boldsymbol{\alpha}\in\Delta_K} \left\{\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}) \,:\, \mathcal{C}_\Pi(\boldsymbol{\eta}_{\boldsymbol{\alpha}};t) \leqslant \delta\right\}\,. \tag{$\mathcal{P}_{\boldsymbol{\alpha}}$}$$

Empirical problem. Let us define the empirical counterparts of the quantities of our interest and consider the empirical problem of ($\mathcal{P}_{\boldsymbol{\alpha}}$). We define the empirical risk of h on an i.i.d. sample $\mathcal{D}^N = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_N\}$ as $\widehat{\mathcal{R}}_{\mathcal{D}^N}(h) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^N \ell(h(\boldsymbol{X}_i), Y_i)$. Let $\widehat{\eta}_k \in \arg\min_h \widehat{\mathcal{R}}_{\mathcal{D}^N_{\pi_k}}(h)$ estimate η_k for all $k \in [K]$ on the perturbed samples of $\mathcal{D}^N$. For $\boldsymbol{\alpha} \in \Delta_K$, we define $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \sum_{k=1}^K \alpha_k \widehat{\eta}_k$.

Assumption 2. *(Regression rates) It holds that* $\max_{k \leqslant K} \mathbb{E}|\widehat{\eta}_k - \eta_k| \leqslant \rho_N$.

Remark 3. *(On regression rates) Assumption 2 can be satisfied with $\rho_N \leqslant \widetilde{\mathcal{O}}(1/\sqrt{N})$ for parametric cases, and even with faster rates under certain assumptions. For example, under the assumption, that there exist $\beta \in \mathbb{N} \setminus \{0\}$ and $L > 0$, such that $|\eta(\boldsymbol{x}) - \eta_{\boldsymbol{x}}(\boldsymbol{x}')| \leqslant L\|\boldsymbol{x} - \boldsymbol{x}'\|^\beta$, where $\eta_{\boldsymbol{x}}$ is Taylor polynomial of degree β at $\boldsymbol{x}$, and under an additional assumption that the support of $\mathbb{P}_{\boldsymbol{X}}$ is $\mathcal{X} = [-1, 1]^d$ and it admits density uniformly lower and upper bounded on $\mathcal{X}$, then for some constants C_1, C_2, it holds that $\sup_{\mathbb{P}} \mathbb{P}_n(|\eta(\boldsymbol{X}) - \widehat{\eta}(\boldsymbol{X})| \geqslant \epsilon) \leqslant C_1 \exp\left(C_2 N^{-\frac{2\beta}{2\beta+d}} \epsilon^2\right)$ [3, Theorem 3.2]. Here, Assumption 2 is satisfied with $\rho_N \leqslant \widetilde{\mathcal{O}}(N^{-\frac{\beta}{2\beta+d}})$.*

Lemma 2 (Mixture rates)**.** *Under Assumption 2 and Claim 3, $\mathbb{E}|\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}} - \boldsymbol{\eta}_{\boldsymbol{\alpha}}| \leqslant \rho_N$ for all $\boldsymbol{\alpha} \in \Delta_K$.*

Let $\mathcal{D}^n = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n\}$ be i.i.d. samples from $\mathcal{D}$, and $\Pi^m = \{\pi'_1, \cdots, \pi'_m\}$ be i.i.d. samples from Π. For a fixed π', we define $\widehat{\Phi}_n(h; \pi') \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n w_\pi(\boldsymbol{Z}_i)\varphi_h(\boldsymbol{Z}_i)$. Let $\widehat{\mathbf{A}} \in \mathbb{R}^{m \times K}$ be a matrix with elements $\widehat{A}_{j,k} = \widehat{\Phi}_n(\widehat{\eta}_k, \pi_j)$. Notice, that we can rewrite the empirical constraint as $\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}, t) = \frac{1}{m}\|\widehat{\mathbf{A}}\alpha - t\mathbf{1}\|_2^2$. Denoting $\widehat{r}_k \stackrel{\text{def}}{=} \widehat{\mathcal{R}}(\widehat{\eta}_k)$ and $\widehat{\boldsymbol{r}} = (\widehat{r}_1, \cdots, \widehat{r}_K)$, we obtain the following second-order cone program (SOCP)

$$\widehat{\boldsymbol{\alpha}} \in \arg\min_{\boldsymbol{\alpha} \in \Delta_K} \left\{ \widehat{\boldsymbol{r}}^\top \boldsymbol{\alpha} \,:\, \left\| \widehat{\mathbf{A}}\boldsymbol{\alpha} - t\mathbf{1} \right\|_2 \leqslant \sqrt{m\delta} \right\}. \tag{$\mathcal{P}_{\texttt{SOCP}}$}$$

Feasibility and existence. The problem ($\mathcal{P}_{\texttt{SOCP}}$) has a continuous linear objective, and the simplex Δ_K is compact and convex. Therefore, once feasibility is established, existence of an optimizer follows from the Weierstrass extreme value theorem. In classical fairness problems, (strict) feasibility is ensured by constant functions, but in our mixture setting one cannot generally claim that a constant function belongs to the convex hull of $\{\eta_k\}_{k=1}^K$ (or $\{\widehat{\eta}_k\}_{k=1}^K$). However, if we augment both population and empirical dictionaries by constant functions $h_{K+1} \equiv 1$ and $h_{K+2} \equiv 0$, and choose $\boldsymbol{\alpha} = (0, \cdots, 0, t, 1-t)$, yielding a constant predictor $h_{\boldsymbol{\alpha}} \equiv t$, we will have $\mathcal{C}_\Pi(t, t) = 0 < \delta$ (or $\widehat{\mathcal{C}}_{n,m}(t, t) = 0 < \delta$). Note, that these two additional predictors do not affect our general theoretical results, thus we omit them from our discussion, and impose strict feasibility as an assumption.

Assumption 3 (Strict feasibility)**.** *There exists $\boldsymbol{\alpha}^0 \in \Delta_K$ s.t. $\mathcal{C}_\Pi(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t) < \delta$.*

4.2 Constructing perturbations

Let us denote by $p_g \stackrel{\text{def}}{=} \mathbb{P}(G = g)$ and $p_{g,y} \stackrel{\text{def}}{=} \mathbb{P}(G = g \mid Y = y)$ for all $g, y \in \{0, 1\}$.

Assumption 4 (Bounded group probabilities)**.** *Let $p_g, p_{g,y} \in [\underline{p}, 1 - \underline{p}]$, with $\underline{p} \in (0, 1/2)$.*

Assumption 5 (Bounded perturbations)**.** *Let Π be supported on $[\underline{\pi}, 1 - \underline{\pi}]$, with $\underline{\pi} \in (0, 1/2)$.*

The construction of perturbations is twofold. First, we consider *stratified sampling* from Π. Letting $\Pi = U(\underline{\pi}, 1 - \underline{\pi})$, we partition $[\underline{\pi}, 1 - \underline{\pi}]$ uniformly into intervals $I_1, \cdots, I_m$ of length $\frac{(1-2\underline{\pi})}{m}$, and sample $\pi'_j \sim U(I_j)$ independently for $j = 1, \cdots, m$. The stratification yields faster convergence rates, as discussed in the next section. Then, for a perturbation parameter $\pi \in \Pi$, the corresponding distribution shift is represented through *importance-weights*, defined by the Radon-Nikodym derivative as $w_\pi(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{d\mathcal{D}_\pi}{d\mathcal{D}}(\boldsymbol{z}) > 0$. For our examples, Assumptions 4,5 ensure that $\mathcal{D}_\pi \ll \mathcal{D}$, and we can explicitly compute w_π (Appendix C.1). Applying Radon-Nikodym theorem gets us $\Phi(h; \pi) = \mathbb{E}_{\mathcal{D}_\pi}[\varphi_h(\boldsymbol{Z})] = \mathbb{E}[w_\pi(\boldsymbol{Z})\varphi_h(\boldsymbol{Z})]$. This means, that instead of resampling, we can just compute and plug in the respective importance weights onto our expressions.

Claim 3 (Bounded weights)**.** *Under Assumptions 4,5, $\exists W > 0$ s.t. $w_\pi(\boldsymbol{Z}) < W, \forall(\pi, \boldsymbol{Z}) \in \Pi \times \mathcal{Z}$.*

4.3 Convergence rates

We now turn to the convergence analysis of the finite procedure, aiming to understand how the empirical solution $\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}$ is approaching its population counterpart $\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star}$. There are three sources of

Algorithm 1: Pseudo-Algorithm

Require: Samples $\{\boldsymbol{Z}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, perturbation parameter law Π, target t, threshold δ.

1: Sample $\pi_1, \cdots, \pi_K \overset{\text{i.i.d.}}{\sim} \Pi$
2: **for** $k = 1$ **to** K **do**
3: Train $\widehat{\eta}_k$ by weighted ERM on $\{\boldsymbol{Z}_i\}$ with weights $w_{\pi_k}(\boldsymbol{Z}_i)$.
4: **end for**
5: Sample $\pi'_1, \ldots, \pi'_m \overset{\text{i.i.d.}}{\sim} \Pi$.
6: Build $\widehat{\mathbf{A}}$, with $\widehat{A}_{j,k} = \widehat{\Phi}_n(\widehat{\eta}_k, \pi'_j) = \frac{1}{n}\sum_{i=1}^n w_{\pi'_j}(Z_i)\varphi_{\widehat{\eta}_k}(Z_i)$; and $r_k = \widehat{\mathcal{R}}_n(\widehat{\eta}_k)$
7: Solve the SOCP ($\mathcal{P}_{\texttt{SOCP}}$) to obtain $\widehat{\boldsymbol{\alpha}}$.
8: Set $\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}} = \sum_{k=1}^K \widehat{\alpha}_k \widehat{\eta}_k$.

error in our framework, each contributing in its own rate. Under Assumption 2 and Remark 3, *estimating regression functions* η_k from N samples of $\mathcal{D}^N$ yields to an error $\rho_N \asymp N^{-\frac{\beta}{2\beta+d}}$, that under convexity transfers to the mixture $\boldsymbol{\eta}_{\boldsymbol{\alpha}}$ as well. Linearity of φ_h in h, combined with Claims 2,3 leads to overall rate of $|\Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi) - \Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi)| \leqslant \delta_N \asymp \rho_N$. For a fixed π', under Claim 3, and with Bernstein's inequality, the *empirical approximation* of $\Phi(\widehat{\eta}_k; \pi')$ from n samples from $\mathcal{D}^n$ contributes $\delta_n \asymp n^{-\frac{1}{2}}$. The third error comes from *perturbation parameter sampling*, where for any Lipschitz function of π, stratification of $[\underline{\pi}, 1-\underline{\pi}]$ into m intervals replaces the usual $m^{-\frac{1}{2}}$ by $\delta_m \asymp m^{-\frac{3}{2}}$. The *risk analysis* relies on a value continuity argument, that follows from Assumption 3. Under Assumption 1, the approximation of the linearized risk introduces an additional error $\delta_{n(R)} \asymp n^{-\frac{1}{2}}$. The combination of the introduced errors leads to the following (informal) theorem. The formal versions of all the statements from this section, along with their proofs, are given in Appendix C.2.

Theorem 1 (Informal)**.** *Suppose Assumptions 1–5 hold. Then, conditionally on $\mathcal{D}^N$ and with "high probability" over $(\mathcal{D}^n, \Pi^m)$, the following hold.*

1. *Constraint. Let $\widehat{\boldsymbol{\alpha}} \in \Delta_K$ s.t. $\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\alpha}}; t) \leqslant \delta$. Then, $\mathcal{C}_\Pi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; t) \leqslant (\sqrt{\delta} + \delta_n)^2 + \delta_m$.*
2. *Unfairness. It holds that $\mathcal{U}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) \leqslant {}^{(\sqrt{\delta}+\delta_n+\sqrt{\delta_m})}/\sqrt{\mathrm{Var}_\Pi(\pi)}$.*
3. *Uniform Deployment. For all $\pi' \in [\underline{\pi}, 1-\underline{\pi}]$, it simultaneously holds that*

$$|\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi') - t| \leqslant (\sqrt{\delta} + \delta_n + \sqrt{\delta_m}) \cdot (1 + {}^{|\pi' - \mathbb{E}_\Pi[\pi]|}/\sqrt{\mathrm{Var}_\Pi(\pi)}) \ .$$

4. *Risk. There exists a constant $0 < C < \infty$, s.t. it holds that*

$$\overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) - \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star}) \leqslant L \cdot \delta_N + C \cdot (2\sqrt{\delta}(\delta_n + \delta_N) + (\delta_n + \delta_N)^2 + \delta_m) + 2\delta_{n(R)} \ .$$

Theorem 1 shows that the convergence rates for the stability constraint and risk in are within $\widetilde{\mathcal{O}}(n^{-\frac{1}{2}} + m^{-\frac{3}{2}} + N^{-\frac{\beta}{2\beta+d}})$. It also shows that our reformulated problem ($\mathcal{P}_{\boldsymbol{\alpha}}$) doesn't yield any convergence loss for unfairness, whose rate is $\widetilde{\mathcal{O}}(n^{-\frac{1}{2}} + m^{-\frac{3}{4}} + N^{-\frac{\beta}{2\beta+d}})$, and provides additional deployment guarantee. The deployment gap is the largest at the extreme points $\pi' \in \{0, 1\}$, corresponding to single-group populations. Unlike a classifier trained under a classical fairness constraint (e.g., DP,EO) at the training marginal p_1, that only controls the unfairness, and does not prevent $\Phi(h; \pi')$ from drifting linearly as π' changes, the solution $\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}$ of ($\mathcal{P}_{\texttt{SOCP}}$) controls both the *slope* $\gamma(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}})$ and the *absolute level* around t, yielding a strictly stronger and distribution-free deployment guarantee. This also highlights the role of Π as a design parameter. Since the guarantee improves with $\mathrm{Var}_\Pi(\pi)$, choosing broader perturbation law leads to stronger fairness guarantees. Finally, the number of classifiers K appears only in logarithmic factors in the $\widetilde{\mathcal{O}}$ bounds and therefore does not affect the convergence rates.

5 Numerical experiments

We evaluate our approach (referred to as STABLE) on two benchmark datasets, namely Adult Income [5] and COMPAS [2] under both DP and EO constraints, where we consider binary classification with gender and race as sensitive attributes. We also consider a large-scale deployment scenario using

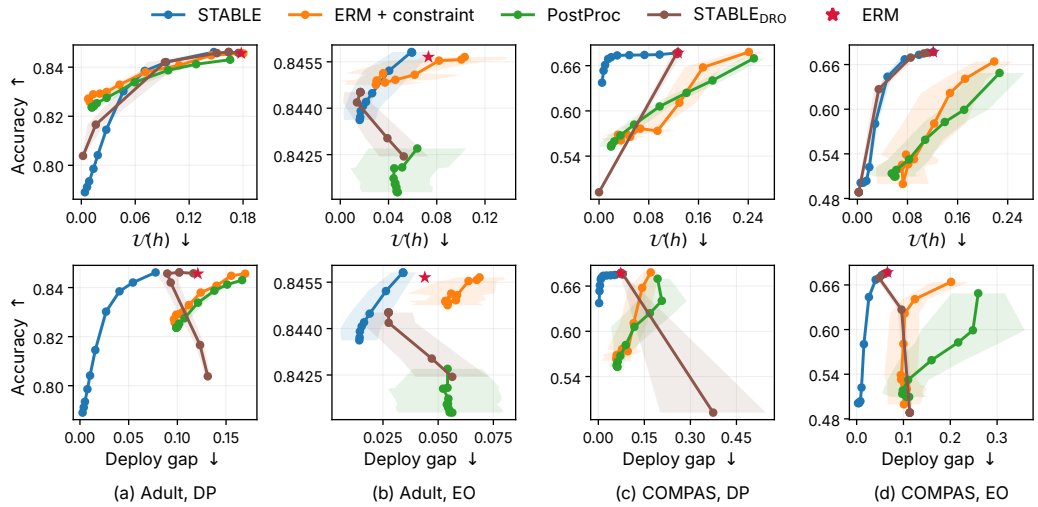


Figure 1: Accuracy–unfairness (top) and accuracy–deployment-gap (bottom) frontiers on Adult and COMPAS. Each point corresponds to one fairness budget of the common grid, averaged over 10 seeds; shaded regions show $\pm$ one standard deviation along the horizontal axis.

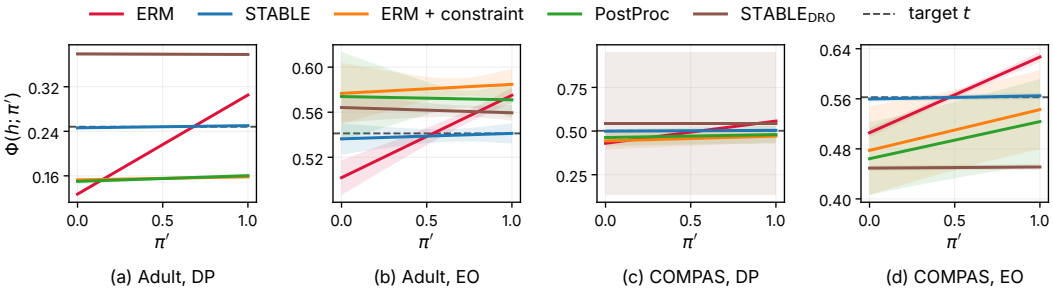


Figure 2: Prediction rate $\Phi(h;\pi')$ as a function of the deployment proportion π' at the tightest fairness budget. Lines show the mean over 10 seeds, shaded regions $\pm$ one standard deviation; the dashed line is the target t

ACS Income [15], a survey of the US adult population covering all 50 states, each with a distinct racial composition. We train on California and evaluate on 11 held-out states.[1].

Baselines STABLE mixes $K = 30$ weighted-ERM logistic regressions, the two constant predictors and, except on COMPAS under DP, $2K_c = 20$ logistic regressions trained on label-flipped data; this choice is made on validation data (Appendix E.1). We compare against ERM under constraints [1], PostProc [11], Shifty [19]. We also report STABLE$_{\text{DRO}}$, an ablation that replaces our objective with an SA-DRO-style one [26] on the same dictionary and perturbations as STABLE (Appendix E). Demographic shift is simulated via importance weighting (see Appendix E.1). The target t is set as the mean prediction of ERM on the training set (DP) or the average group-conditional TPR of ERM (EO). Rather than a single operating point per method, we sweep each method over a common grid of fairness budgets and report the resulting accuracy–unfairness and accuracy–deployment-gap frontiers, averaged over 10 random seeds (5 on ACS Income; protocol in Appendix E.1). For STABLE, a deeper analysis of hyperparameters and the impact of t values on training is conducted in Appendices E.1 and E.4.

Results. Figure 1 reports the Pareto frontiers of all methods (additional plots including Shifty [19] are presented in Appendix E.2). Overall, STABLE is highly competitive with the other models. STABLE outperforms all the models in terms of accuracy–deployment gap trade-off, and achieves the lowest deployment gap among all the baselines across all settings with a large margin. The accuracy–

[1]The code is available at https://github.com/laclauc/stable-fairness.

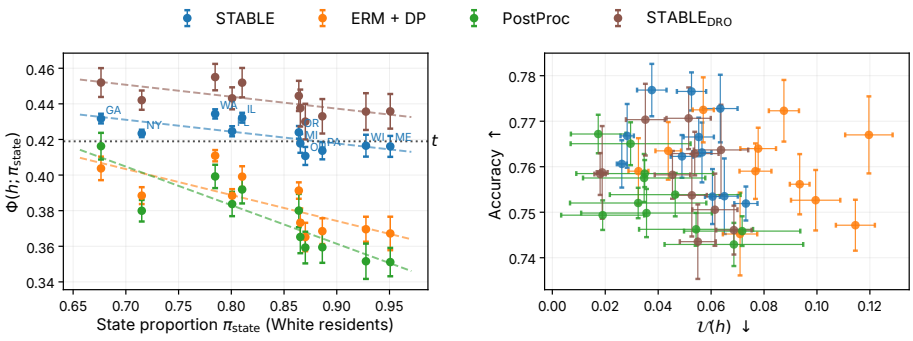


Figure 3: Geographic generalization on ACS Income, at a fixed fairness budget. (left) Prediction rate $\Phi(h;\pi_{\text{state}})$ against the proportion of White residents in each held-out state; dashed lines: linear fits, dotted line: target t. (right) Accuracy against $\mathcal{U}(h)$. Each point is one state, mean $\pm$ std over 5 seeds.

fairness trade-off, on the other hand, varies across settings. On COMPAS, STABLE outperforms ERM+constraint and PostProc both on DP and EO. On the former, for STABLE$_{\text{DRO}}$, the least unfair point is a constant predictor whose level varies across seeds while on EO it gets a comparable accuracy-fairness trade-off but a large deployment gap. On Adult under EO, all methods lie within 0.005 accuracy of each other and their differences in $\mathcal{U}(h)$ are within seed variability. Finally, on Adult under DP, the trade-off is less favourable for STABLE, where at a lower unfairness it has a larger accuracy loss compared to other models. Figure 2 illustrates the prediction rate w.r.t. π'. The prediction rate for classical unconstrained ERM dramatically drifts across population compositions. For PostProc and ERM+constraint the angle of the slope depends on the setting, however, even when the slope is small, the rate still remains far from t. The ablation with SA-DRO highlights the role of the level t and the SOCP formulation: while the prediction rate for STABLE$_{\text{DRO}}$ is nearly flat in π', it still remains away from t.

Overall, everything is consistent with the theory: the SOCP constraint controls both the slope $\gamma(h)$ and the absolute level of $\Phi(h;\pi')$ around t, which is strictly stronger than the classical fairness constraint and shrinks the feasible set accordingly. Moreover, SOCP optimizes a linearized surrogate of the risk, which, despite upper-bounding the true mixture risk, introduces a linearization gap that may occasionally translate into reduced accuracy.

For ACS, Figure 3 compares the methods at a fixed trade-off: the same budget $\delta_0 = 0.01$ for all methods, and STABLE$_{\text{DRO}}$ at its least unfair value of λ. Additional results for both constraints are reported in Appendix E.3. We remark that STABLE remains the closest to t across all held-out states (Figure 3). PostProc and ERM+constraint introduce a systematic geographic bias: their slope $\pi_{\text{state}} \mapsto \Phi(h;\pi_{\text{state}})$ is negative, causing them to drift below t in high-proportion states. The slope of STABLE$_{\text{DRO}}$ is similar the one of STABLE, however still remains far from t. This illustrates that controlling static unfairness does not preclude deployment bias. The not completely flat prediction rate for STABLE is consistent with the fact that the population composition shift doesn't exactly follow Perturbation 1 from theory. On the accuracy–unfairness trade-off, STABLE overall attains accuracies higher or comparable to other models. Results for EO are deferred to the Appendix E.3.

6 Conclusion and perspectives

We introduced a stability-based perspective on fairness and showed that both individual and group fairness arise from it through appropriate perturbation classes. We then introduced a stronger constraint that jointly controls group fairness and the stability of the prediction rate across perturbations, and showed that when considering a linearized risk, the problem admits a tractable SOCP solution. We further established convergence guarantees from empirical to population quantities and illustrated the practical relevance of the problem through numerical experiments and comparisons to various benchmarks. One limitation is that our method addresses a specific tractable surrogate rather than the original problem in full generality. A natural next step is therefore to derive a more direct solution of the original problem and to establish end-to-end guarantees.

Acknowledgments and Disclosure of Funding

This work is supported by Hi! PARIS and ANR/France 2030 program (ANR-23-IACL-0005) and by the PEPR IA FOUNDRY (ANR-23-PEIA-0003). This work was supported by the Caisse des Dépôts as part of a research partnership with Télécom Paris

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] J-Y Audibert and AB Tsybakov. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35(2):608–633, 2007.

[4] Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. Preserving the fairness guarantees of classifiers in changing environments: A survey. *ACM Comput. Surv.*, 57 (6), February 2025. ISSN 0360-0300. doi: 10.1145/3637438. URL https://doi.org/10.1145/3637438.

[5] Barry Becker and Ronny Kohavi. Adult Dataset, 1996. URL https://archive.ics.uci.edu/dataset/2/adult.

[6] Avrim Blum, Princewill Okoroafor, Aadirupa Saha, and Kevin M. Stangl. On the vulnerability of fairness constrained learning to malicious noise. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4096–4104. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/blum24a.html.

[7] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2: 499–526, March 2002. ISSN 1532-4435. doi: 10.1162/153244302760200704. URL https://doi.org/10.1162/153244302760200704.

[8] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, page 13–18, USA, 2009. IEEE Computer Society. ISBN 9780769539027. doi: 10.1109/ICDMW.2009.83. URL https://doi.org/10.1109/ICDMW.2009.83.

[9] E. Celis, L. Huang, V. Keswani, and N. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[10] S. Chiappa, R. Jiang, T. Stepleton, A. Pacchiano, H. Jiang, and J. Aslanides. A general approach to fairness with optimal transport. In *AAAI*, 2020.

[11] Evgenii Chzhen, Mohamed Hebiri, and Gayane Taturyan. Randomized multi-class classification under system constraints: a unified approach via post-processing, 2025. URL https://arxiv.org/abs/2512.14246. arXiv preprint arXiv:2512.14246.

[12] C. Denis, R. Elie, M. Hebiri, and F. Hu. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25(130):1–46, 2024.

[13] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979. doi: 10.1109/TIT.1979.1056087.

[14] L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Inf. Theor.*, 25(2):202–207, September 2006. ISSN 0018-9448. doi: 10.1109/TIT.1979.1056032. URL https://doi.org/10.1109/TIT.1979.1056032.

[15] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490, 2021.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[17] Jean-David Fermanian, Dominique Guégan, and Xuwen Liu. Fair learning by model averaging. *Risk and Decision Analysis*, 11(1-2):20–49, 2025. doi: 10.1177/15697371251321734.

[18] S. Gaucher, N. Schreuder, and E. Chzhen. Fair learning with wasserstein barycenters for non-decomposable performance measures. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2436–2459. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/gaucher23a.html.

[19] Stephen Giguere, Blossom Metevier, Yuriy Brun, Philip S. Thomas, Scott Niekum, and Bruno Castro da Silva. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*, 2022.

[20] Paula Gordaliza, Eustasio del Barrio, Fabrice Gamboa, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *ICML*, pages 2357–2365, 2019.

[21] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 29, 2016.

[22] X. Hou and L. Zhang. Finite-sample and distribution-free fair classification: Optimal trade-off between excess risk and fairness, and the cost of group-blindness. *arXiv preprint arXiv:2410.16477*, 2024.

[23] Zhimeng (Stephen) Jiang, Xiaotian Han, Hongye Jin, Guanchu Wang, Rui Chen, Na Zou, and Xia Hu. Chasing fairness under distribution shift: A model weight perturbation approach. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 63931–63944. Curran Associates, Inc., 2023.

[24] Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Comput.*, 11(6):1427–1453, August 1999. ISSN 0899-7667. doi: 10.1162/089976699300016304. URL https://doi.org/10.1162/089976699300016304.

[25] Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 275–282, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608974.

[26] Haoyu Lei, Amin Gohari, and Farzan Farnia. On the inductive biases of demographic parity-based fair learning algorithms. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

[27] Thibaud Leteno, Michael Perrot, Charlotte Laclau, Antoine Gourru, and Christophe Gravier. Fair text classification via transferable representations. *Journal of Machine Learning Research*, 26(239):1–47, 2025.

[28] L. Oneto, M. Donini, and M. Pontil. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[29] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. 2022.

[30] W. H. Rogers and T. J. Wagner. A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules. *The Annals of Statistics*, 6(3):506 – 514, 1978. doi: 10.1214/aos/1176344196. URL https://doi.org/10.1214/aos/1176344196.

[31] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

[32] Jake A. Soloff, Rina Foygel Barber, and Rebecca Willett. Bagging provides assumption-free stability. *JMLR*, 2024. URL https://arxiv.org/abs/2301.12600.

[33] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

[34] R. Xian, L. Yin, and H. Zhao. Fair and optimal classification via post-processing. In *International conference on machine learning*, pages 37977–38012. PMLR, 2023.

[35] X. Zeng, E. Dobriban, and G. Cheng. Fair bayes-optimal classifiers under predictive parity. *Advances in Neural Information Processing Systems*, 35:27692–27705, 2022.

Organization of the Supplementary Material

The Supplementary Material is organized as follows.

- Appendix A provides additional literature review and further discussion.
- Appendix B contains the main proofs establishing the equivalence between existing fairness notions and our stability-based definition, together with additional discussion.
- Appendix C provides the explicit forms of the importance weights, the proofs of the convergence results for the constraints and the risk, along with some auxiliary lemmas needed in the analysis.
- Appendix D extends the framework to Equalized Odds. We introduce the relevant notions and prove the associated equivalence result with stability.
- Appendix E is dedicated to numerical experiments. We provide additional implementation details and conduct further experiments.

A Discussions on related works

Machine learning models are increasingly used in high-stakes decisions, such as hiring, healthcare, and criminal justice. In such settings, biases already present in the data may be inherited or even amplified by the resulting models, with potentially serious consequences for the individuals affected. Algorithmic fairness seeks to understand, quantify, and mitigate these effects. Since this question also involves philosophical and social considerations, it naturally lies at the intersection of several fields; in this paper, however, we focus on its mathematical aspect and do not aim to resolve the broader question of what fairness ought to be.

A.1 Stability and learning theory

Many notions of stability have been introduced in the Statistics and Machine-Learning literature to define and quantify the *stability property* of a statistical learning algorithm, meaning roughly that a slight change in the data (distribution) has no significant effect on the decision function produced by the algorithm. Depending on the definition of stability chosen, it is possible to use concentration results to deduce error bounds for decision rules from the stability property satisfied by the algorithm producing them, *e.g.*, ERM with a certain cost function and/or with additive penalization, bagging. This idea is not new at all and dates back at least to the late 70's in Statistics, refer to [30], [14], [13] and [24] for instance. The concept of stability has more recently been revisited in new forms by the machine learning community. See in particular the concept of *uniform stability* in [7] yielding exponential bounds and the refinements in [25]. More generally, this line of research has been adopted by many authors, too numerous to be listed in an exhaustive fashion, dealing with various approaches (*e.g.* bagging in [32]) and issues (clustering, latent variable analysis, ranking).

A.2 Comparison with Distributionally Robust Optimization

While our framework bears a resemblance to distributionally robust optimization (DRO), the two approaches differ in both structure and the guarantees they provide, with interesting connections that we now discuss.

DRO reminder. Given a divergence d (here KL) and radius $\varepsilon > 0$, define the uncertainty set $\mathcal{U}_\varepsilon = \{\mathcal{D}' : \mathrm{KL}(\mathcal{D}'\|\mathcal{D}) \leq \varepsilon\}$. DRO solves

$$\min_h \sup_{\mathcal{D}' \in \mathcal{U}_\varepsilon} \mathbb{E}_{\mathcal{D}'}[\ell(h(\boldsymbol{X}), Y)].$$

Our demographic shifts $\mathcal{D}_\pi$ (Perturbation 1) satisfy $\mathcal{D}_\pi \in \mathcal{U}_\varepsilon$ whenever $\varepsilon \geq \mathrm{KL}(\mathcal{D}_\pi\|\mathcal{D})$. Since $\mathcal{D}_\pi$ only resamples G while keeping $P(\boldsymbol{X}, Y \mid G)$ fixed, the KL reduces to the KL between two Bernoulli distributions:

$$\mathrm{KL}(\mathcal{D}_\pi\|\mathcal{D}) = \pi \log \frac{\pi}{p_1} + (1-\pi) \log \frac{1-\pi}{1-p_1}.$$

To cover all deployment proportions $\pi' \in [\bar{\pi}, 1-\bar{\pi}]$, DRO requires a radius at least

$$\varepsilon^*(\bar{\pi}) = \max_{\pi \in [\bar{\pi}, 1-\bar{\pi}]} \left[\pi \log \frac{\pi}{p_1} + (1-\pi) \log \frac{1-\pi}{1-p_1} \right],$$

which is attained at one of the endpoints $\pi \in \{\bar{\pi}, 1-\bar{\pi}\}$ by convexity of the KL divergence. Since $\mathrm{KL}(\mathrm{Bernoulli}(\pi) \| \mathrm{Bernoulli}(p_1))$ is continuous and strictly positive for $\pi \neq p_1$, we have $\varepsilon^*(\bar{\pi}) > 0$ uniformly in $p_1 \in [\bar{p}, 1-\bar{p}]$; in particular, $\varepsilon^* = \Omega(1)$ whenever $|\bar{\pi} - p_1| \geq c$ for some constant $c > 0$. Table 1 illustrates representative values.

Table 1: Minimum DRO radius $\varepsilon^*(\bar{\pi})$ required to cover all demographic shifts $\pi' \in [\bar{\pi}, 1-\bar{\pi}]$, for varying training group proportion p_1 and perturbation lower bound $\bar{\pi}$.

$p_1 \setminus \bar{\pi}$	0.1	0.2	0.3
0.2	1.15	0.83	0.58
0.3	0.79	0.53	0.34
0.5	0.37	0.19	0.08

At radius $\varepsilon = \varepsilon^* = \Omega(1)$, the uncertainty set $\mathcal{U}_\varepsilon$ is sufficiently large to contain distributions that arbitrarily corrupt the feature mechanism $P(\boldsymbol{X} \mid G)$ and the label mechanism $P(Y \mid G)$. Consequently, the worst-case risk over $\mathcal{U}_\varepsilon$ is controlled by corruptions of $P(\boldsymbol{X}|G)$ and $P(Y|G)$ that are unrelated to demographic fairness, hence making the bound uninformative about the performance under such demographic shift.

Why our framework avoids this problem. The uncertainty set $\mathcal{U}_\varepsilon$ contains distributions that alter $P(\boldsymbol{X} \mid G)$, $P(Y \mid G)$, and other aspects of the data generating process entirely irrelevant to demographic fairness. Our framework instead restricts to the *one-dimensional manifold* $\{\mathcal{D}_\pi : \pi \in [\bar{\pi}, 1-\bar{\pi}]\}$, which captures exactly the shifts of interest while preserving $P(\boldsymbol{X}, Y \mid G)$. This structural specificity combined with the affine form $\Phi(h;\pi) = \alpha(h) + \pi\gamma(h)$, enables simultaneous control of slope and absolute level via $\mathcal{C}_\Pi(h;t)$. As a result, Theorem 1 provides guarantees that converge to zero as $n, m \to \infty$ – something no DRO bound with radius $\varepsilon = O(1)$ can achieve.

Connection to Group DRO and SA-DRO. Our framework is more closely related to structured DRO approaches such as Group DRO [31]. In the binary case, Group DRO minimizes $\max_{\lambda \in [0,1]} \lambda \mathbb{E}_{D_1}[\ell(\cdot)] + (1-\lambda)\mathbb{E}_{D_0}[\ell(\cdot)]$, which is exactly the worst-case risk over our Perturbation 1 family $\{\mathcal{D}_\pi : \pi \in [0,1]\}$. The two frameworks thus share the same perturbation structure for demographic parity. The differences are twofold. First, the objectives differ: Group DRO minimizes only the worst-case risk, while STABLE controls the variance of $\Phi(h;\pi)$ around a target level t, which we show to be formally equivalent to fairness constraints. Second, Perturbation 2, which resamples G within the positive class to capture equal opportunity, has no natural Group DRO counterpart, as Group DRO cannot vary $P(G=1 \mid Y=1)$ independently of $P(G=1 \mid Y=0)$.

A closely related work is SA-DRO [26], which applies DRO directly on the marginal distribution of the sensitive attribute and adds an explicit fairness regularization term. The same structural differences with Group DRO apply here. Additionally, SA-DRO relies on soft regularization rather than a hard constraint, and therefore does not provide quantitative guarantees on the unfairness gap $\mathcal{U}(h)$ or on the deployment gap.

B Proofs for Section 3

B.1 Individual fairness

Recall that the Wasserstein 1-distance between the pushforward of $h(\boldsymbol{X})$'s distribution under $\mathcal{D}$ and that under $\mathbf{P}(\mathcal{D})$ based on the cost function $d_{\mathcal{Y}}$ is

$$\inf_{\boldsymbol{X} \sim \mathcal{D},\, \boldsymbol{X}' \sim \mathbf{P}(\mathcal{D})} \mathbb{E}[d_{\mathcal{Y}}(h(\boldsymbol{X}), h(\boldsymbol{X}')],$$

where the infimum is taken over all couplings $(\boldsymbol{X}, \boldsymbol{X}')$ of the pair of distributions $\mathcal{D}$ and $\mathbf{P}(\mathcal{D})$.

Proof of Proposition 1.

Proposition (Proposition 1, IF and Stability)**.** *Let $h : \mathcal{X} \to \mathcal{Y}$ be a prediction function and $d_{\mathcal{Y}} : \mathcal{Y}^2 \to [0, \infty)$ a metric on predictions. The following assertions are equivalent.*

1. *The predictive function h satisfies the IF property with $(d_{\mathcal{Y}}, d_{\mathcal{X}})$-Lipschitz constant $L < \infty$.*
2. *For all $\rho > 0$, the predictive function h is $L\rho$-fair w.r.t. to the class $\mathcal{P}_\rho$ of perturbation operators $\mathbf{P}_T$, $T \in \mathcal{T}_\rho$, defined on the domain $\mathcal{M} = \{\delta_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$ of all point masses and taking $\Delta(h;\ (\mathcal{D}, \mathbf{P}))$ as the Wasserstein 1-distance between the pushforward of $h(\boldsymbol{X})$'s distribution under $\mathcal{D}$ and that under $\mathbf{P}(\mathcal{D})$ based on the cost function $d_{\mathcal{Y}}$:*

$$\sup_{(\boldsymbol{x},T)\in\mathcal{X}\times\mathcal{T}_\rho} d_{\mathcal{Y}}(h(\boldsymbol{x}), h(T(\boldsymbol{x}))) \leqslant L\rho\,.$$

Proof. Let us show that $(1 \Rightarrow 2)$. It is easy to notice that

$$\sup_{(\boldsymbol{x},T)\in\mathcal{X}\times\mathcal{T}_\rho} d_{\mathcal{Y}}(h(\boldsymbol{x}), h(T(\boldsymbol{x}))) \leqslant \sup_{(\boldsymbol{x},T)\in\mathcal{X}\times\mathcal{T}_\rho} L d_{\mathcal{X}}(\boldsymbol{x}, T(\boldsymbol{x})) \leqslant L\rho\,.$$

Now let us show that $(2 \Rightarrow 1)$. First, let us fix any $\boldsymbol{x}, \boldsymbol{x}'$, and denote $\rho' := d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')$. Let us also introduce the following transformation

$$T_{\boldsymbol{x}\to\boldsymbol{x}'}(\boldsymbol{z}) = \begin{cases} \boldsymbol{x}', & if \quad \boldsymbol{z} = \boldsymbol{x} \\ \boldsymbol{z}, & otherwise\,. \end{cases}$$

Choosing $\rho = \rho'$, we consider the set $\mathcal{T}_{\rho'}$. We notice that $T_{\boldsymbol{x}\to\boldsymbol{x}'} \in \mathcal{T}_{\rho'}$ and observe that

$$\begin{aligned} d_{\mathcal{Y}}(h(\boldsymbol{x}), h(\boldsymbol{x}')) &= d_{\mathcal{Y}}(h(\boldsymbol{x}), h(T_{\boldsymbol{x}\to\boldsymbol{x}'}(\boldsymbol{x}))) \\ &\leqslant \sup_{\boldsymbol{x}\in\mathcal{X}} d_{\mathcal{Y}}(h(\boldsymbol{x}), h(T_{\boldsymbol{x}\to\boldsymbol{x}'}(\boldsymbol{x}))) \\ &\leqslant \sup_{T\in\mathcal{T}_{\rho'}} \sup_{\boldsymbol{x}\in\mathcal{X}} d_{\mathcal{Y}}(h(\boldsymbol{x}), h(T(\boldsymbol{x}))) \leqslant L\rho' = L d_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{x}')\,. \end{aligned}$$

The proof is concluded. □

B.2 Group fairness

Proof of Proposition 2.

Proposition (Proposition 2, DP and Stability)**.** *Let $h : \mathcal{X} \to \{0, 1\}$ be a (possibly randomized) prediction function, $G \in \{0, 1\}$ be a sensitive attribute, and denote $r_g(h) \stackrel{\text{def}}{=} \mathbb{P}(h(\boldsymbol{X}) = 1 \mid G = g), \forall g \in \{0, 1\}$. Let $\mathcal{D}_\pi$ be a perturbed distribution according to Perturbation 1. We define the **global positive** prediction rate under $\mathcal{D}_\pi$ as $P_h(\pi) \stackrel{\text{def}}{=} \mathbb{P}_{\mathcal{D}_\pi}(h(\boldsymbol{X}) = 1)$. Then, the following statements are equivalent:*

1. *h satisfies Demographic Parity: $r_0(h) = r_1(h)$.*
2. *$P_h(\pi)$ is constant in π on $[0, 1]$ (i.e., the global positive rate is invariant to any resampling of G).*

Moreover, one has the explicit identity

$$P_h(\pi) = r_0(h) + \pi(r_1(h) - r_0(h))\,,$$

so that

$$\sup_{\pi,\pi'\in[0,1]} |P_h(\pi) - P_h(\pi')| = |r_1(h) - r_0(h)|\,.$$

Proof. By the law of total probability, for any $\pi \in [0, 1]$,

$$P_h(\pi) \;=\; \mathbb{P}_{\mathcal{D}_\pi}(h(\boldsymbol{X}) = 1) \;=\; \pi\, r_1(h) + (1 - \pi)\, r_0(h)\,.$$

We note that if $r_0(h) = r_1(h)$, then $P_h(\pi) = \pi r_1(h) + (1 - \pi) r_1(h) = r_1(h)$, which does not depend on π, hence (1) ⇒ (2).

Secondly, if we suppose that $P_h(\pi)$ is constant in π, then for all $\pi, \pi' \in [0,1]$, we have

$$\pi r_1(h) + (1-\pi) r_0(h) \;=\; \pi' r_1(h) + (1-\pi') r_0(h)\,,$$

which can be rewritten as

$$(\pi - \pi')(r_1(h) - r_0(h)) = 0\,.$$

Since we may choose $\pi \neq \pi'$, it follows that $r_1(h) - r_0(h) = 0$, i.e., $r_0(h) = r_1(h)$, which is exactly Demographic Parity (so (2) $\Rightarrow$ (1)).

This proves the equivalence. □

Proof of Proposition 3.

Proposition (Proposition 3, EO and Stability)**.** *Let $h : \mathcal{X} \to \{0,1\}$ be a (possibly randomized) prediction function, $G \in \{0,1\}$ be a sensitive attribute, and denote $q_{g,y}(h) \stackrel{\text{def}}{=} \mathbb{P}(h(\boldsymbol{X}) = 1 \mid G = g, Y = y), \forall g, y \in \{0,1\}$. Let $\mathcal{D}_\pi$ be a perturbed distribution according to Perturbation 2. We define the **true positive** prediction rate under $\mathcal{D}_\pi$ as $P_{h|Y=1}(\pi) \stackrel{\text{def}}{=} \mathbb{P}_{\mathcal{D}_\pi}(h(\boldsymbol{X}) = 1 \mid Y = 1)$. Then the following statements are equivalent:*

1. *h satisfies Equal Opportunity: $q_{0,1}(h) = q_{1,1}(h)$.*
2. *$P_{h|Y=1}(\pi)$ is constant in π on $[0,1]$ (i.e., the true positive rate is invariant to any resampling of G within positive label class).*

Moreover, one has the explicit identity

$$P_{h|Y=1}(\pi) = q_{0,1}(h) + \pi(q_{1,1}(h) - q_{0,1}(h))\,,$$

so that

$$\sup_{\pi,\pi' \in [0,1]} |P_{h|Y=1}(\pi) - P_{h|Y=1}(\pi')| = |q_{1,1}(h) - q_{0,1}(h)|\,.$$

Proof. Let us show that $(1 \Rightarrow 2)$. Denoting $q := q_{0,1}(h) = q_{1,1}(h)$ and applying the law of total probability, we get

$$P_{h|Y=1}(\pi) = \mathbb{P}_{\mathcal{D}_\pi}(h(X) = 1 \mid Y = 1) = (1-\pi) q_{0,1}(h) + \pi q_{1,1}(h) = q\,,$$

which does not depend on π. Now let us show that $(2 \Rightarrow 1)$. Assuming $P_{h|Y=1}(\pi)$ is constant in π, we can state that

$$(1-\pi) q_{0,1}(h) + \pi q_{1,1}(h) = (1-\pi') q_{0,1}(h) + \pi' q_{1,1}(h)\,.$$

Rewriting the above as

$$(\pi - \pi')(q_{1,1}(h) - q_{0,1}(h)) = 0\,,$$

we deduce that $q_{1,1}(h) = q_{0,1}(h)$.

The proof is concluded. □

C Details and proofs for Section 4

C.1 Importance weights representation

In this section we show that the distribution shifts under our consideration can be described by their corresponding importance-weights, defined by the Radon-Nikodym derivative

$$w_\pi(\boldsymbol{x}, g, y) \stackrel{\text{def}}{=} \frac{d\mathcal{D}_\pi}{d\mathcal{D}}(\boldsymbol{x}, g, y)\,.$$

Recall, that $p_g = \mathbb{P}(G = g)\ \forall g \in \{0,1\}$ and $\mathbb{P}(G = g \mid Y = y)\ \forall g, y \in \{0,1\}$.

For Perturbation 1, considering the factorization $\mathbb{P}(\boldsymbol{X}, G, Y) = \mathbb{P}(G)\mathbb{P}(\boldsymbol{X}, Y \mid G)$, and computing the Radon-Nikodym derivative, we get

$$w_\pi(g) = \begin{cases} \frac{\pi}{p_1}\,, & \text{if } g = 1\,, \\ \frac{1-\pi}{p_0}\,, & \text{if } g = 0\,. \end{cases}$$

For Perturbation 2, considering the factorization $\mathbb{P}(\boldsymbol{X}, G, Y) = \mathbb{P}(Y)\mathbb{P}(G \mid Y)\mathbb{P}(\boldsymbol{X} \mid Y, G)$, and computing the Radon-Nikodym derivative, we get

$$w_\pi(g,y) = \begin{cases} \frac{\pi}{p_{1,1}}\,, & \text{if } (g,y) = (1,1)\,, \\ \frac{1-\pi}{p_{0,1}}\,, & \text{if } (g,y) = (0,1)\,, \\ 1\,, & \text{if } y = 0\,. \end{cases}$$

C.2 Convergence rates

First, let us recall the following two crucial lemmas.

Lemma 3. *[33, Hoeffding inequality] Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$ be independent random variables such that $\boldsymbol{X}_i \in [a_i, b_i]$ for every i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^N (\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i]) \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^N (b_i - a_i)^2}\right).$$

Lemma 4. *[33, Bernstein inequality] Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N$ be independent, mean-zero random variables satisfying $|\boldsymbol{X}_i| \leqslant K$ for every i. Let $\sigma^2 = \sum_{i=1}^N \mathbb{E}\boldsymbol{X}_i^2$ is the variance of the sum. Then, for any $t > 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N \boldsymbol{X}_i\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

We now examine the different sources of error arising from the estimation of population-level quantities.

C.2.1 Regression error rates

Recall the following expressions

$$\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star} = \sum_{k\in[K]} \boldsymbol{\alpha}_k^\star \eta_k\,, \qquad \boldsymbol{\eta}_{\widehat{\boldsymbol{\alpha}}} = \sum_{k\in[K]} \widehat{\boldsymbol{\alpha}}_k \eta_k\,, \qquad \widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}^\star} = \sum_{k\in[K]} \boldsymbol{\alpha}_k^\star \widehat{\eta}_k\,, \qquad \widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}} = \sum_{k\in[K]} \widehat{\boldsymbol{\alpha}}_k \widehat{\eta}_k\,.$$

Lemma 5. *Let Assumption 2 and Claims 2–3 hold. Then, for all for every $\boldsymbol{\alpha} \in \Delta_K$ and all $\pi \in [\underline{\pi}, 1-\underline{\pi}]$, it holds that*

$$\left|\Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi) - \Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi)\right| \leqslant BW\rho_N \stackrel{\text{def}}{=} \delta_N\,.$$

Proof. By the linearity of φ_h in h and the definition of the mixture, for any $\boldsymbol{\alpha} \in \Delta_K$, we have

$$|\Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi) - \Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi)| = \left|\sum_{k=1}^K \alpha_k \left[\Phi(\widehat{\eta}_k; \pi) - \Phi(\eta_k; \pi)\right]\right| \leqslant \max_{1\leqslant k\leqslant K} \left|\Phi(\widehat{\eta}_k; \pi) - \Phi(\eta_k; \pi)\right|.$$

For each k and π, from Hölder's inequality and Claims 2–3, we obtain

$$\begin{aligned} |\Phi(\widehat{\eta}_k; \pi) - \Phi(\eta_k; \pi)| &= |\mathbb{E}[w_\pi(\boldsymbol{Z})\varphi_{\widehat{\eta}_k - \eta_k}(\boldsymbol{Z})]| \leqslant \|w_\pi(\boldsymbol{Z})\|_{L^\infty} \mathbb{E}[\varphi_{\widehat{\eta}_k - \eta_k}(\boldsymbol{Z})] \\ &\leqslant WB\mathbb{E}[|\widehat{\eta}(\boldsymbol{Z}) - \eta(\boldsymbol{Z})|]\,. \end{aligned}$$

Assumption 2 concludes the proof. □

C.2.2 Perturbation error rates

Lemma 6 (Stratified sampling)**.** *Let $\Pi = U(\underline{\pi}, 1-\underline{\pi})$ and $\mathcal{I}_1, \cdots, \mathcal{I}_m$ be a uniform partition of $[\underline{\pi}, 1-\underline{\pi}]$ into m intervals of length $|\mathcal{I}| = \frac{(1-2\underline{\pi})}{m}$. We sample $\pi'_j \sim U(\mathcal{I}_j)$ independently for $j = 1, \cdots, m$. Let $f : [\underline{\pi}, 1-\underline{\pi}] \to \mathbb{R}$ be a $L-$Lipschitz function. For all $a_{\delta_m} \in (0,1)$ with probability greater than $1 - a_{\delta_m}$, it holds that*

$$\left|\frac{1}{m}\sum_{j=1}^m f(\pi'_j) - \mathbb{E}_\Pi[f(\pi)]\right| \leqslant \frac{L(1-2\underline{\pi})}{m^{3/2}} \sqrt{2\log(2/a_{\delta_m})} \stackrel{\text{def}}{=} \delta_{L,m}\,.$$

Proof. Denote the stratum mean by $\mu_j \stackrel{\text{def}}{=} \frac{1}{|\mathcal{I}|}\int_{\mathcal{I}_j} f(\pi)\,\mathrm{d}\,\pi$ and the intra-stratum deviation by $\xi_j \stackrel{\text{def}}{=} f(\pi'_j) - \mu_j$. Notice, that $\frac{1}{m}\sum_{j=1}^m \mu_j = \frac{1}{m}\sum_j \frac{1}{|\mathcal{I}|}\int_{\mathcal{I}_j} f(\pi)\,\mathrm{d}\,\pi = \frac{1}{1-2\underline{\pi}}\int_{\underline{\pi}}^{1-\underline{\pi}} f(\pi)\,\mathrm{d}\,\pi = \mathbb{E}_\Pi[f(\pi)]$, therefore $\frac{1}{m}\sum_j f(\pi'_j) - \mathbb{E}_\Pi[f(\pi)] = \frac{1}{m}\sum_{j=1}^m \xi_j$. Moreover, for $\pi \in \mathcal{I}_j$, $|f(\pi) - \mu_j| \leqslant \sup_{\pi,\pi'\in\mathcal{I}_j}|f(\pi) - f(\pi')| \leqslant L|\mathcal{I}|$, so $|\xi_j| \leqslant L|\mathcal{I}|$. Combining this with the facts that $\mathbb{E}[\xi_j] = 0$ for each j, that ξ_j are independent, and applying Hoeffding's inequality, we obtain

$$\mathbb{P}\Big(\Big|\frac{1}{m}\sum_j \xi_j\Big| > \epsilon\Big) \leqslant 2\exp\Big(-\frac{m\epsilon^2}{2L^2|\mathcal{I}|^2}\Big) = 2\exp\Big(-\frac{m^3\epsilon^2}{2L^2(1-2\underline{\pi})^2}\Big).$$

Setting $a_{\delta_m} = 2\exp\Big(-\frac{m^3\epsilon^2}{2L^2(1-2\underline{\pi})^2}\Big)$ and obtaining $\epsilon = \frac{\sqrt{2}L(1-2\underline{\pi})}{m^{3/2}}\sqrt{\log(2/a_{\delta_m})}$ concludes the proof. □

C.2.3 Sample generalization rates

Lemma 7. *For all $a_{\delta_n} \in (0,1)$ with probability greater than $1 - a_{\delta_n}$ and conditionally on $\mathcal{D}^N$, with probability greater than $1 - a_{\delta_n}$ it holds that*

$$\max_{k\in[K],j\in[m]}\Big|\widehat{\Phi}_n(\widehat{\eta}_k;\pi'_j) - \Phi(\widehat{\eta}_k;\pi'_j)\Big| \leqslant \sqrt{\frac{2W^2\log(2Km/a_{\delta_n})}{n}} + \frac{2W\log(2Km/a_{\delta_n})}{3n} \stackrel{\text{def}}{=} \delta_n\,.$$

Moreover, for all $\boldsymbol{\alpha} \in \Delta_K$ and $j \in [m]$, it holds that

$$\Big|\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}};\pi'_j) - \Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}};\pi'_j)\Big| \leqslant \delta_n\,.$$

Proof. Let us fix $k \in [K]$ and $j \in [m]$. The function $\widehat{\eta}_k$ is fixed conditionally on sample $\mathcal{D}^N$. Let us define $\xi_i \stackrel{\text{def}}{=} w_{\pi'_j}(\boldsymbol{Z}_i)\varphi_{\widehat{\eta}_k}(\boldsymbol{Z}_i)$, for $\boldsymbol{Z}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}^n$. We have $\mathbb{E}[\xi_i] = \Phi(\widehat{\eta}_k;\pi'_j)$, $|\xi_i| \leqslant W$, $|\xi_i - \mathbb{E}[\xi_i]| \leqslant W$ and $\mathrm{Var}(\xi_i) \leqslant \mathbb{E}[\xi_i^2] \leqslant W^2$. Applying Bernstein's inequality, for any $\epsilon > 0$ it holds that

$$\mathbb{P}_{\mathcal{D}^n}\Big(\Big|\frac{1}{n}\sum_i \xi_i\Big| > \epsilon \,|\, \mathcal{D}^N\Big) \leqslant 2\exp\Big(-\frac{n\epsilon^2}{2W^2 + \frac{2}{3}W\epsilon}\Big).$$

Setting $\frac{a_{\delta_n}}{Km} = 2\exp\Big(-\frac{n\epsilon^2}{2W^2+\frac{2}{3}W\epsilon}\Big)$, obtaining $\epsilon = \sqrt{\frac{2W^2\log\left(2Km/a_{\delta_n}\right)}{n}} + \frac{2W\log\left(2Km/a_{\delta_n}\right)}{3n}$, and applying union bound over all pairings$(k,j) \in [K]\times[m]$ concludes the proof of the first inequality.

Moreover, using the triangle inequality and the fact that $\alpha_k \geqslant 0$, $\sum_k \alpha_k = 1$, we obtain

$$\begin{aligned}
\Big|\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}};\pi'_j) - \Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}};\pi'_j)\Big| &= \Big|\sum_{k=1}^K \alpha_k(\widehat{\Phi}_n(\widehat{\eta}_k;\,\pi'_j) - \Phi(\widehat{\eta}_k;\pi'_j))\Big| \\
&\leqslant \sum_{k=1}^K \alpha_k\Big|\widehat{\Phi}_n(\widehat{\eta}_k;\pi'_j) - \Phi(\widehat{\eta}_k;\pi'_j)\Big| \\
&\leqslant \max_k |\widehat{\Phi}_n(\widehat{\eta}_k;\pi'_j) - \Phi(\widehat{\eta}_k;\pi'_j)| \leqslant \delta_n.
\end{aligned}$$

The proof is concluded. □

C.2.4 Final constraint rates

Formal version and the proof of the first statement of Theorem 1.

Theorem 2 (Constraint rates)**.** *Let $\widehat{\boldsymbol{\alpha}}$ be a simplex vector, s.t. $\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}};t) \leqslant \delta$. For all $a_{\delta_m}, a_{\delta_n} \in (0,1)$ and conditionally on $\mathcal{D}^N$, with probability greater than $1 - a_{\delta_m} - a_{\delta_n}$ it holds that*

$$\mathcal{C}_\Pi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}};t) \leqslant \Big(\sqrt{\delta} + \delta_n\Big)^2 + \delta_m\,,$$

where

$$\delta_m \stackrel{\text{def}}{=} \frac{2(2-\underline{\pi})(1-2\underline{\pi})}{m^{3/2}}\sqrt{2\log(4/a_{\delta_m})} \quad \textit{and} \quad \delta_n = \sqrt{\frac{2W^2\log(2Km/a_{\delta_n})}{n}} + \frac{2W\log(2Km/a_{\delta_n})}{3n}\,.$$

Proof. Let us fix $j \in [m]$. By the triangle inequality, and Lemma 7, we have

$$|\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t| \leqslant |\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t| + |\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - \Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j)| \leqslant |\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t| + \delta_n \,.$$

Squaring both sides of the above inequality, we get $(\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t)^2 \leqslant (|\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t| + \delta_n)^2$. Averaging over $j = 1, \cdots, m$ yields

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m}(\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t)^2 &\leqslant \frac{1}{m}\sum_j(|\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t| + \delta_n)^2 \\ &= \widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; t) + 2\delta_n \frac{1}{m}\sum_j |\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t| + \delta_n^2 \,.\end{aligned}$$

By Cauchy–Schwarz inequality, and the fact that $\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; t) \leqslant \delta$, we have

$$\frac{1}{m}\sum_j |\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t| \leqslant \sqrt{\frac{1}{m}\sum_j(\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t)^2} = \sqrt{\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; t)} \leqslant \sqrt{\delta} \,.$$

Combining the inequalities above we get

$$\frac{1}{m}\sum_{j=1}^{m}(\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t)^2 \leqslant (\sqrt{\delta} + \delta_n)^2 \,. \tag{7}$$

For any $\boldsymbol{\alpha} \in \Delta_K$, by linearity of $\alpha(\cdot)$[2] and $\gamma(\cdot)$, that follows from linearity of $\Phi(h; \pi)$ in Claim 1, and denoting $\alpha_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \sum_{k=1}^K \alpha_k \alpha(\widehat{\eta}_k)$ and $\gamma_{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \sum_{k=1}^K \alpha_k \gamma(\widehat{\eta}_k)$, we get

$$(\Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi) - t)^2 = (\alpha_{\boldsymbol{\alpha}} - t)^2 + 2(\alpha_{\boldsymbol{\alpha}} - t)\gamma_{\boldsymbol{\alpha}}\pi + \gamma_{\boldsymbol{\alpha}}^2\pi^2 \,.$$

Averaging over $\pi'_1, \cdots, \pi'_m$ and subtracting from $\mathcal{C}_\Pi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; t)$, we get

$$\begin{aligned}\mathcal{C}_\Pi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; t) - \frac{1}{m}\sum_{j=1}^{m}(\Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi'_j) - t)^2 \\ = 2(\alpha_{\boldsymbol{\alpha}} - t)\gamma_{\boldsymbol{\alpha}}\Big\{\mathbb{E}_\Pi[\pi] - \frac{1}{m}\sum_{j=1}^{m}\pi'_j\Big\} + \gamma_{\boldsymbol{\alpha}}^2\Big\{\mathbb{E}_\Pi[\pi^2] - \frac{1}{m}\sum_{j=1}^{m}(\pi'_j)^2\Big\} \,.\end{aligned}$$

Now, let us notice that the maps $\pi \mapsto \pi$ and $\pi \mapsto \pi^2$ are respectively 1–Lipschitz and $2(1 - \underline{\pi})$–Lipschitz on $[\underline{\pi}, 1 - \underline{\pi}]$. Applying Lemma 6 with a union bound, and recalling that $\alpha_{\boldsymbol{\alpha}} \in [0, 1]$ and $\gamma_{\boldsymbol{\alpha}} \in [-1, 1]$, with probability at least $1 - a_{\delta_m}$ and uniformly in $\boldsymbol{\alpha} \in \Delta_K$, we obtain

$$\mathcal{C}_\Pi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; t) \leqslant \frac{1}{m}\sum_j(\Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi'_j) - t)^2 + \frac{2(2 - \underline{\pi})(1 - 2\underline{\pi})}{m^{3/2}}\sqrt{2\log(4/a_{\delta_m})} \,.$$

Finally, specializing at $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}$ and applying (7), we get

$$\mathcal{C}_\Pi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; t) \leqslant \frac{1}{m}\sum_j(\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi'_j) - t)^2 + \delta_m \leqslant (\sqrt{\delta} + \delta_n)^2 + \delta_m \,.$$

The proof is concluded. □

The second statement of Theorem 1 is stated below as a corollary and follows directly from Theorem 2.

Corollary 1 (Unfairness rates)**.** *Under the conditions of Theorem 2, it holds that*

$$\mathcal{U}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) \leqslant \sqrt{\frac{(\sqrt{\delta} + \delta_n)^2 + \delta_m}{\mathrm{Var}_\Pi(\pi)}} \leqslant \frac{\sqrt{\delta} + \delta_n + \sqrt{\delta_m}}{\sqrt{\mathrm{Var}_\Pi(\pi)}} \,.$$

Thus, if we set $\delta = \delta_0^2 \,\mathrm{Var}_\Pi(\pi)$, then we will have $\mathcal{U}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) - \delta_0 \leqslant \widetilde{\mathcal{O}}\Big(\frac{1}{n^{1/2}} + \frac{1}{m^{3/4}}\Big)$.

[2]Do not confuse the functional $\alpha(\cdot)$ from the decomposition of $\Phi(h; \pi)$ with the simplex vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$.

Formal version and the proof of the third statement of Theorem 1.

Theorem 3 (Uniform Deployment Guarantee)**.** *Let $\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}$ be any feasible solution of ($\mathcal{P}_{\texttt{SOCP}}$) with threshold δ and target $t \in [0,1]$. Under the conditions of Theorem 2, For all $a_{\delta_m}, a_{\delta_n} \in (0,1)$ and conditionally on $\mathcal{D}^N$, with probability greater than $1 - a_{\delta_m} - a_{\delta_n}$ over (D^n, Π^m), the following holds simultaneously for **all** $\pi' \in [\underline{\pi}, 1-\underline{\pi}]$*

$$\left|\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi') - t\right| \leqslant (\sqrt{\delta} + \delta_n + \sqrt{\delta_m}) \cdot \left(1 + \frac{|\pi' - \mu_\Pi|}{\sqrt{\mathrm{Var}_\Pi(\pi)}}\right).$$

Proof. It follows from (6) and Theorem 2 that

$$(\mathbb{E}_{\pi\sim\Pi}[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)] - t)^2 \leqslant (\sqrt{\delta} + \delta_n)^2 + \delta_m\,.$$

Taking the square root, we obtain

$$|\mathbb{E}_{\pi\sim\Pi}[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)] - t| \leqslant \sqrt{\delta} + \delta_n + \sqrt{\delta_m}\,.$$

For a fixed π', we have

$$\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi') - t = \underbrace{\left[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi') - \mathbb{E}_\Pi[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)]\right]}_{\text{fluctuation term}} + \underbrace{\left[\mathbb{E}_\Pi[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)] - t\right]}_{\text{bias term}}.$$

Let us denote by $\mu_\Pi \overset{\text{def}}{=} \mathbb{E}_\Pi[\pi]$. For the fluctuation term, by the linearity of $\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)$ and the identity $\mathbb{E}_\Pi[\alpha(h) + \pi\gamma(h)] = \alpha(h) + \mu_\Pi\gamma(h)$, we have

$$\begin{aligned}\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi') - \mathbb{E}_\Pi[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)] &= \big(\alpha(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) + \pi'\gamma(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}})\big) - \big(\alpha(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) + \mu_\Pi\gamma(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}})\big)\\ &= (\pi' - \mu_\Pi)\gamma(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}})\,.\end{aligned}$$

Therefore,

$$\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi') - t = (\pi' - \mu_\Pi)\gamma(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) + \left[\mathbb{E}_\Pi[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)] - t\right]. \tag{8}$$

Applying the triangle inequality to (8) and substituting bounds from Theorem 2,

$$\begin{aligned}\left|\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi') - t\right| &\leqslant |\pi' - \mu_\Pi| \cdot |\gamma(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}})| + \left|\mathbb{E}_\Pi[\Phi(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}; \pi)] - t\right|\\ &\leqslant |\pi' - \mu_\Pi| \cdot \frac{(\sqrt{\delta} + \delta_n + \sqrt{\delta_m})}{\sqrt{\mathrm{Var}_\Pi(\pi)}} + (\sqrt{\delta} + \delta_n + \sqrt{\delta_m})\\ &= \left(1 + \frac{|\pi' - \mu_\Pi|}{\sqrt{\mathrm{Var}_\Pi(\pi)}}\right) \cdot (\sqrt{\delta} + \delta_n + \sqrt{\delta_m})\,.\end{aligned}$$

Since this bound holds for all $\pi' \in [0,1]$ simultaneously, the proof is concluded. □

Let us also prove the following lemma, that we will use in later analysis.

Lemma 8 (Reverse feasibility)**.** *Under the conditions of Lemma 5, Lemma 6 and Lemma 7, let $\sqrt{\delta} > \delta_n + \delta_N$ and $(\sqrt{\delta} - \delta_n - \delta_N)^2 > \delta_m$. We define the* tightened level *as*

$$\widetilde{\delta} \overset{\text{def}}{=} (\sqrt{\delta} - \delta_n - \delta_N)^2 - \delta_m\,.$$

If $\boldsymbol{\alpha} \in \Delta_K$ satisfies $\mathcal{C}_\Pi(\eta_{\boldsymbol{\alpha}}; t) \leqslant \widetilde{\delta}$, then $\widehat{\mathcal{C}}_{n,m}(\widehat{\eta}_{\boldsymbol{\alpha}}; t) \leqslant \delta$.

Proof. Fix any $j \in [m]$. By the triangle inequality,

$$\begin{aligned}\left|\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi'_j) - t\right| &\leqslant \left|\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi'_j) - \Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi'_j)\right| + \left|\Phi(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi'_j) - \Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi'_j)\right| + \left|\Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi'_j) - t\right|\\ &\leqslant \delta_n + \delta_N + \left|\Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi'_j) - t\right|.\end{aligned}$$

Squaring and averaging over $j = 1, \ldots, m$ gives

$$\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; t) = \frac{1}{m}\sum_{j=1}^m \big(\widehat{\Phi}_n(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; \pi'_j) - t\big)^2 \leqslant \frac{1}{m}\sum_{j=1}^m \Big(\left|\Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi'_j) - t\right| + \delta_n + \delta_N\Big)^2. \tag{9}$$

Let us define $g(\pi) \stackrel{\text{def}}{=} (\Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi) - t)^2$. Since the map $\pi \mapsto \Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi)$is affine, g is a polynomial of degree two in π, and in particular Lipschitz on $[\underline{\pi}, 1 - \underline{\pi}]$ with constant at most $2W^2$, applying Lemma 6, we obtain

$$\frac{1}{m}\sum_{j=1}^{m} g(\pi'_j) \leqslant \mathbb{E}_{\Pi}[g(\pi)] + \delta_m = \mathcal{C}_{\Pi}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; t) + \delta_m \leqslant \widetilde{\delta} + \delta_m \,.$$

Applying the Cauchy–Schwarz inequality, we get

$$\frac{1}{m}\sum_{j=1}^{m} \big|\Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi'_j) - t\big| \leqslant \sqrt{\frac{1}{m}\sum_{j=1}^{m} g(\pi'_j)} \leqslant \sqrt{\widetilde{\delta} + \delta_m} \,.$$

Expanding the right-hand-side of (9), we obtain

$$\begin{aligned}
\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; t) &\leqslant \frac{1}{m}\sum_{j} g(\pi'_j) + 2(\delta_n + \delta_N) \cdot \frac{1}{m}\sum_{j} \big|\Phi(\eta_{\boldsymbol{\alpha}}; \pi'_j) - t\big| + (\delta_n + \delta_N)^2 \\
&\leqslant (\widetilde{\delta} + \delta_m) + 2(\delta_n + \delta_N)\sqrt{\widetilde{\delta} + \delta_m} + (\delta_n + \delta_N)^2 \\
&\leqslant \big(\sqrt{\widetilde{\delta} + \delta_m} + \delta_n + \delta_N\big)^2 \,.
\end{aligned}$$

Substituting $\widetilde{\delta} = (\sqrt{\delta} - \delta_n - \delta_N)^2 - \delta_m$ gives

$$\sqrt{\widetilde{\delta} + \delta_m} = \sqrt{(\sqrt{\delta} - \delta_n - \delta_N)^2} = \sqrt{\delta} - \delta_n - \delta_N \,,$$

and finally we get

$$\widehat{\mathcal{C}}_{n,m}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}; t) \leqslant \big((\sqrt{\delta} - \delta_n - \delta_N) + \delta_n + \delta_N\big)^2 = \delta \,.$$

The proof is concluded. □

C.2.5 Risk rates

For prediction functions $h_1, \cdots, h_K$, let us recall the notation of population and empirical linearized risks

$$\overline{\mathcal{R}}(h_{\boldsymbol{\alpha}}) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k \mathcal{R}(h_k) \,, \qquad \widehat{\overline{\mathcal{R}}}_n(h_{\boldsymbol{\alpha}}) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \alpha_k \widehat{\mathcal{R}}_n(h_k) \,.$$

Let us also define

$$\overline{\nu}(\delta') \stackrel{\text{def}}{=} \min_{\boldsymbol{\alpha} \in \Delta_K} \big\{\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}) \,:\, \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; t) \leqslant \delta'\big\} \,, \text{ for } \delta' > 0 \,.$$

The population-optimal mixture at level δ is $\boldsymbol{\alpha}^\star \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\alpha} \in \Delta_K} \{\overline{R}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}) : C_{\Pi}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; t) \leqslant \delta\}$, so that $\overline{\nu}(\delta) = \overline{R}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star})$.

Lemma 9 (Value continuity)**.** *Under Assumption 3, for every* $0 \leqslant \gamma \leqslant \delta - \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t)$ *, it holds that*

$$0 \leqslant \overline{\nu}(\delta - \gamma) - \overline{\nu}(\delta) \leqslant \gamma C \,, \quad \textit{where} \quad C \stackrel{\text{def}}{=} \frac{\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}) - \overline{\nu}(\delta)}{\delta - \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t)} \,.$$

Proof. Let $\boldsymbol{\alpha}^\star_\delta$ be a minimizer in the definition of $\overline{\nu}(\delta)$. Fix $\gamma \in [0, \delta - \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t)]$ and set

$$\lambda_\gamma \stackrel{\text{def}}{=} \frac{\gamma}{\delta - \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t)} \in [0, 1] \,, \qquad \boldsymbol{\alpha}_\gamma \stackrel{\text{def}}{=} (1 - \lambda_\gamma)\boldsymbol{\alpha}^\star_\delta + \lambda_\gamma \boldsymbol{\alpha}^0 \,.$$

The simplex Δ_K is convex, hence $\boldsymbol{\alpha}_\gamma \in \Delta_K$. Moreover, $\boldsymbol{\alpha} \mapsto \Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi)$ is affine for every fixed π, hence $\boldsymbol{\alpha} \mapsto (\Phi(\boldsymbol{\eta}_{\boldsymbol{\alpha}}; \pi) - t)^2$ is convex, and taking the expectation over $\pi \sim \Pi$, again, preserves the convexity. Therefore, it holds that

$$\mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}_\gamma}; t) \leqslant (1 - \lambda_\gamma)\mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star_\delta}; t) + \lambda_\gamma \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t) \leqslant (1 - \lambda_\gamma)\delta + \lambda_\gamma \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t) = \delta - \gamma \,.$$

Hence, $\boldsymbol{\alpha}_\gamma$ is feasible for level $\delta - \gamma$, and thus

$$\overline{\nu}(\delta - \gamma) \leqslant \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}_\gamma}) \,.$$

Using the linearity of $\overline{\mathcal{R}}$, we obtain

$$\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}_\gamma}) = (1-\lambda_\gamma)\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star_\delta}) + \lambda_\gamma \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}) = \overline{\nu}(\delta) + \lambda_\gamma(\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}) - \overline{\nu}(\delta)) \,.$$

Combining the last two, we obtain

$$\overline{\nu}(\delta - \gamma) - \overline{\nu}(\delta) \leqslant \gamma \frac{\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}) - \overline{\nu}(\delta)}{\delta - \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t)} \,.$$

We conclude the proof by noticing that $\overline{\nu}(\delta) \leqslant \overline{\nu}(\delta - \gamma)$ due to monotonicity of the feasible sets. □

Lemma 10. *Under Assumption 2 and Assumption 1, for every $\alpha \in \Delta_K$,*

$$\left|\overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}) - \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}})\right| \leqslant L\rho_N \,.$$

Proof. For each $k \in [K]$, by the Lipschitz property of ℓ in its first argument,

$$\left|\mathcal{R}(\widehat{\eta}_k) - \mathcal{R}(\eta_k)\right| = \left|\mathbb{E}[\ell(\widehat{\eta}_k(\boldsymbol{X}), Y) - \ell(\eta_k(\boldsymbol{X}), Y)]\right| \leqslant L\mathbb{E}|\widehat{\eta}_k(\boldsymbol{X}) - \eta_k(\boldsymbol{X})| \,.$$

Since $\boldsymbol{\alpha} \in \Delta_K$,

$$\begin{aligned}\left|\overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}) - \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}})\right| &= \Big|\sum_{k=1}^K \alpha_k(\mathcal{R}(\widehat{\eta}_k) - \mathcal{R}(\eta_k))\Big| \leqslant \sum_{k=1}^K \alpha_k\left|\mathcal{R}(\widehat{\eta}_k) - \mathcal{R}(\eta_k)\right| \\ &\leqslant L \max_k \mathbb{E}|\widehat{\eta}_k(\boldsymbol{X}) - \eta_k(\boldsymbol{X})| \leqslant L\rho_N \,.\end{aligned}$$

The proof is concluded. □

Lemma 11 (Uniform linearized risk deviation)**.** *For every $a_{\delta_R} \in (0,1)$, with probability at least $1 - a_{\delta_R}$,*

$$\sup_{\boldsymbol{\alpha}\in\Delta_K} \left|\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}) - \widehat{\overline{\mathcal{R}}}_n(\boldsymbol{\eta}_{\boldsymbol{\alpha}})\right| \leqslant \sqrt{\frac{\log\left(2K/a_{\delta_R}\right)}{2n}} \stackrel{\text{def}}{=} \delta_{n(R)} \,.$$

Moreover, the statement is identical for estimated mixture $\widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}$.

Proof. For each fixed $k \in [K]$, the random variables $\ell(h_k(X_i), Y_i)$ are i.i.d. and take values in $[0,1]$. Hoeffding's inequality yields

$$\mathbb{P}_{\mathcal{D}^n}\left(\left|\mathcal{R}(\eta_k) - \widehat{\mathcal{R}}_n(\eta_k)\right| > \epsilon \,|\, \mathcal{D}^N\right) \leqslant 2\exp\left(-2n\epsilon^2\right) .$$

Taking the union bound over $k \in [K]$ gives

$$\mathbb{P}_{\mathcal{D}^n}\left(\max_{k\in[K]}\left|\mathcal{R}(\eta_k) - \widehat{\mathcal{R}}_n(\eta_k)\right| > \epsilon \,|\, \mathcal{D}^N\right) \leqslant 2K\exp\left(-2n\epsilon^2\right) .$$

Choosing $\epsilon = \sqrt{\frac{\log\left(2K/a_{\delta_R}\right)}{2n}}$ yields, with probability at least $1 - a_{\delta_R}$ and conditionally on $\mathcal{D}^N$,

$$\max_{k\in[K]}\left|\mathcal{R}(\eta_k) - \widehat{\mathcal{R}}_n(\eta_k)\right| \leqslant \delta_{n(R)} \,.$$

Finally, for any $\boldsymbol{\alpha} \in \Delta_K$,

$$\left|\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}}) - \widehat{\overline{\mathcal{R}}}_n(\boldsymbol{\eta}_{\boldsymbol{\alpha}})\right| = \left|\sum_{k=1}^K \alpha_k\big(\mathcal{R}(\eta_k) - \widehat{\mathcal{R}}_n(\eta_k)\big)\right| \leqslant \sum_{k=1}^K \alpha_k\left|\mathcal{R}(\eta_k) - \widehat{\mathcal{R}}_n(\eta_k)\right| \leqslant \delta_{n(R)} \,.$$

The proof is concluded. □

Formal version and the proof of the fourth statement of Theorem 1.

Theorem 4 (Risk rates)**.** *Let $\widehat{\boldsymbol{\alpha}}$ be a solution of ($\mathcal{P}_{\texttt{SOCP}}$). Suppose Assumptions 2–5 hold, and n, m, N are large enough so that $\sqrt{\delta} > \delta_n + \delta_N$ and $(\sqrt{\delta} - \delta_n - \delta_N)^2 > \delta_m$. Then, for all $a_{\delta_n}, a_{\delta_m}, a_{\delta_{n(R)}} \in (0,1)$, conditionally on D^N, with probability at least $1 - a_{\delta_n} - a_{\delta_m} - a_{\delta_{n(R)}}$ over (D^n, Π^m), it holds that*

$$\overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) - \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star}) \leqslant L \cdot \rho_N + C \cdot (2\sqrt{\delta}(\delta_n + \delta_N) + (\delta_n + \delta_N)^2 + \delta_m) + 2\delta_{n(R)}\,,$$

where δ_n and δ_m are from Theorem 2, and

$$\delta_{n(R)} = \sqrt{\frac{\log(2K/a_{\delta_R})}{2n}}\,, \qquad C = \frac{\overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}) - \overline{\nu}(\delta)}{\delta - \mathcal{C}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^0}; t)}\,.$$

Proof. Let $\widetilde{\delta} = (\sqrt{\delta} - \delta_n - \delta_N)^2 - \delta_m > 0$ be the tightened threshold from the reverse feasibility (Lemma 8). When n, m, N are "large enough", then $\delta_n, \delta_m, \delta_N$ are "sufficiently" small, so it holds that $\mathcal{C}_\Pi(\boldsymbol{\eta}_{\boldsymbol{\alpha}_0}; t) < \widetilde{\delta}$, that is the feasible set at the level $\widetilde{\delta}$ is non-empty. Let us denote by $\widetilde{\boldsymbol{\alpha}}$ the minimizer of $\overline{\nu}(\widetilde{\delta})$. We have $\mathcal{C}_\Pi(\boldsymbol{\eta}_{\widetilde{\boldsymbol{\alpha}}}; t)$ and $\overline{\nu}(\widetilde{\delta}) = \overline{\mathcal{R}}(\boldsymbol{\eta}_{\widetilde{\boldsymbol{\alpha}}})$. From Lemma 8, we have that $\mathcal{C}_{n,m}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{\alpha}}}) \leqslant \delta$. Since $\widehat{\boldsymbol{\alpha}}$ is the empirical minimizer of ($\mathcal{P}_{\texttt{SOCP}}$), and $\widetilde{\boldsymbol{\alpha}}$ is empirically feasible, we have $\widehat{\overline{\mathcal{R}}}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) \leqslant \widehat{\overline{\mathcal{R}}}_n(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{\alpha}}})$. Combining this, with Lemma 11, we get

$$\overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) \leqslant \widehat{\overline{\mathcal{R}}}_n(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) + \delta_{n(R)} \leqslant \widehat{\overline{\mathcal{R}}}_n(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{\alpha}}}) + \delta_{n(R)} \leqslant \overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{\alpha}}}) + 2\delta_{n(R)}\,.$$

Next, we apply Lemma 10 and obtain

$$\overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\widetilde{\boldsymbol{\alpha}}}) \leqslant \overline{\mathcal{R}}(\boldsymbol{\eta}_{\widetilde{\boldsymbol{\alpha}}}) + L\rho_N = \overline{\nu}(\widetilde{\delta}) + L\rho_N\,.$$

Finally, we are in position to apply value continuity (Lemma 9). Setting

$$\begin{aligned}\gamma &= \delta - \widetilde{\delta} = \delta - (\sqrt{\delta} - \delta_n - \delta_N)^2 + \delta_m \\ &= 2\sqrt{\delta}(\delta_n + \delta_N) + (\delta_n + \delta_N)^2 + \delta_m\,,\end{aligned}$$

we get

$$\overline{\nu}(\widetilde{\delta}) = \overline{\nu}(\delta - \gamma) \leqslant \overline{\nu}(\delta) + \gamma C\,.$$

Combining everything above, we finally obtain

$$\overline{\mathcal{R}}(\widehat{\boldsymbol{\eta}}_{\widehat{\boldsymbol{\alpha}}}) \leqslant \overline{\mathcal{R}}(\boldsymbol{\eta}_{\boldsymbol{\alpha}^\star}) + \gamma C + L\rho_N + 2\delta_{n(R)}.$$

The proof is concluded. □

D Extension to Equalized Odds

Definition 8 (Equalized Odds)**.** *A classifier h satisfies the EOdds property if it achieves equal true positive rates and equal false positive rates across groups:*

$$\mathbb{P}(h(\boldsymbol{X}) = 1 \mid Y = y, G = 0) \;=\; \mathbb{P}(h(\boldsymbol{X}) = 1 \mid Y = y, G = 1)\,, \qquad \forall y \in \{0,1\}\,.$$

To connect EOdds to stability, we consider the following perturbation.

Perturbation 3. *The class $\mathcal{P}_{\boldsymbol{\pi}}^{G|Y}$, with $\boldsymbol{\pi} = (\pi_0, \pi_1)$, is composed of all operators $\mathbf{P} : \mathcal{M}_1(\mathcal{Z}) \to \mathcal{M}_1(\mathcal{Z})$ such that, for all distribution $\mathcal{D} \in \mathcal{M}_1(\mathcal{Z})$, 1) $\boldsymbol{X}$'s conditional distribution given (Y, G) and Y's distribution are the same under $\mathcal{D}$ and under $\mathbf{P}(\mathcal{D})$, 2) $\mathbb{P}_{\mathbf{P}(\mathcal{D})}(G = 1 \mid Y = y) = \pi_y\,, \forall y \in \{0,1\}$.*

Proposition 4 (EOdds and Stability)**.** *Let $h : \mathcal{X} \to \{0,1\}$ be a (possibly randomized) prediction function, $G \in \{0,1\}$ be a sensitive attribute, and denote $q_{g,y}(h) \stackrel{\text{def}}{=} \mathbb{P}(h(\boldsymbol{X}) = 1 \mid G = g, Y = y), \forall g, y \in \{0,1\}$. Let $\mathcal{D}_{\boldsymbol{\pi}}$ be a perturbed distribution according to Perturbation 3. We define the **true positive** and **false positive** prediction rates under $\mathcal{D}_{\boldsymbol{\pi}}$ respectively as $P_{h|Y=1}(\boldsymbol{\pi}) \stackrel{\text{def}}{=} \mathbb{P}_{\mathcal{D}_{\boldsymbol{\pi}}}(h(\boldsymbol{X}) = 1 \mid Y = 1)$ and $P_{h|Y=0}(\boldsymbol{\pi}) \stackrel{\text{def}}{=} \mathbb{P}_{\mathcal{D}_{\boldsymbol{\pi}}}(h(\boldsymbol{X}) = 1 \mid Y = 0)$. Then, the following statements are equivalent:*

1. *h satisfies Equalized Odds: $q_{0,y}(h) = q_{1,y}(h)\,, \forall y \in \{0,1\}$.*

Table 2: Running time of one fit (training and prediction), in seconds: median [interquartile range] over the fairness budgets and seeds (10×10 on Adult and COMPAS, 6×5 on ACS Income), excluding Shifty runs without a solution; for STABLE$_{\text{DRO}}$, dictionary and one value of λ, over the seeds. Shifty uses Setting 2 (unknown demographic shift, $\alpha = 0.25$).

	ERM + constraint		STABLE (ours)		STABLE$_{\text{DRO}}$		PostProc		Shifty	
	DP	EO	DP	EO	DP	EO	DP	EO	DP	EO
Adult	5.86 [4.80, 7.03]	6.34 [5.44, 7.01]	10.4 [8.9, 12.8]	11.3 [10.4, 11.9]	11.8 [9.9, 13.1]	11.5 [9.7, 12.9]	1.70 [1.61, 1.83]	1.89 [1.83, 1.96]	178 [112, 243]	160 [112, 220]
COMPAS	0.91 [0.84, 1.04]	1.02 [0.93, 1.19]	1.97 [1.59, 2.51]	2.68 [2.34, 2.96]	2.63 [2.13, 3.19]	3.02 [2.84, 3.15]	1.40 [1.37, 1.44]	1.54 [1.49, 1.63]	8.07 [4.84, 13.8]	7.63 [4.16, 10.9]
ACS	116 [106, 128]	101 [88, 123]	248 [233, 263]	218 [209, 234]	254 [224, 263]	220 [214, 255]	8 [7, 9]	17 [16, 19]	39 [26, 57]	24 [16, 35]

2. *$P_{h|Y=y}(\boldsymbol{\pi})$ is constant in π_y on $[0,1]\,, \forall y \in \{0,1\}$ (i.e., the true positive and false negative rates are invariant to any intra-class resampling of G).*

Moreover, one has the explicit identity

$$P_{h|Y=y}(\boldsymbol{\pi}) = q_{0,y}(h) + \pi_y(q_{1,y}(h) - q_{0,y}(h))\,, \quad \forall y \in \{0,1\}\,,$$

so that

$$\sup_{\pi_y, \pi'_y \in [0,1]} |P_{h|Y=y}(\pi_y) - P_{h|Y=y}(\pi'_y)| = |q_{1,y}(h) - q_{0,y}(h)|\,, \quad \forall y \in \{0,1\}\,.$$

Proof. Let us show that $(1 \Rightarrow 2)$. Denoting $q_y := q_{0,y}(h) = q_{1,y}(h), \forall y \in \{0,1\}$, and applying the law of total probability, we get

$$P_{h|Y=y}(\boldsymbol{\pi}) = \mathbb{P}_{\mathcal{D}_{\boldsymbol{\pi}}}(h(\boldsymbol{X}) = 1 \mid Y = y) = (1-\pi_y)q_{0,y}(h) + \pi_y q_{1,y}(h) = q_y\,, \forall y \in \{0,1\}\,,$$

which does not depend on π_y.

Now let us show that $(2 \Rightarrow 1)$. Assuming $P_{h|Y=y}(\boldsymbol{\pi})$ is constant in $\pi_y\,, \forall y \in \{0,1\}$, we can state that for $y \in \{0,1\}$,it holds that

$$(1-\pi_y)q_{0,y}(h) + \pi_y q_{1,y}(h) = (1-\pi'_y)q_{0,y}(h) + \pi'_y q_{1,y}(h) \quad \forall y \in \{0,1\}\,.$$

Rewriting the above as

$$(\pi_y - \pi'_y)(q_{1,y}(h) - q_{0,1}(h)) = 0\,,$$

we deduce that $q_{0,y}(h) = q_{1,y}(h)\,, \forall y \in \{0,1\}$. The proof is concluded. □

Importance-weights representation Under Perturbation 3, considering the factorization $\mathbb{P}(\boldsymbol{X}, G, Y) = \mathbb{P}(Y)\mathbb{P}(G \mid Y)\mathbb{P}(\boldsymbol{X} \mid Y, G)$, and computing the Radon-Nikodym derivative, we get

$$w_{\boldsymbol{\pi}}(g,y) = \begin{cases} \frac{\pi_1}{p_{1,1}}\,, & \text{if } (g,y) = (1,1)\,, \\ \frac{1-\pi_1}{p_{0,1}}\,, & \text{if } (g,y) = (0,1)\,, \\ \frac{\pi_0}{p_{1,0}}\,, & \text{if } (g,y) = (1,0)\,, \\ \frac{1-\pi_0}{p_{0,0}}\,, & \text{if } (g,y) = (0,0)\,. \end{cases}$$

E Experiments: additional details and results

All experiments are run on an Apple M2 with 16 GB RAM. The running time of one fit is reported in Table2. In all experiments, we operate in the *unaware* setting: the sensitive attribute G is not available at prediction time.

E.1 Evaluation protocol

Simulating deployment shift. For the Adult Income and COMPAS datasets we simulate the deployment under demographic shift without retraining.

Given a fixed trained model h and a test set $\{(x_i, g_i, y_i)\}$, the functional $\Phi(h; \pi')$ is estimated via importance weighting: each observation receives weight $w_{\pi'}(g_i) = \frac{\pi'}{p_1}\mathbf{1}\{g_i = 1\} + \frac{1-\pi'}{p_0}\mathbf{1}\{g_i = 0\}$, where $p_g = \widehat{\mathbb{P}}(G = g)$ is estimated on the test set.

This reweights the empirical distribution to match a target group proportion π' without requiring a new dataset, and corresponds exactly to Perturbation 1 of Section 3. The deploy gap is then $\sup_{\pi' \in \{0,1\}} |\Phi(h; \pi') - t|$, evaluated at the two extremes.

Table 3: Hyperparameters of STABLE. $\bar{\pi}$ and the dictionary are selected on validation data; $K = 30$, $K_c = 10$ and $m = 60$ are fixed across all settings. The budget is set through δ_0, with $\delta = \delta_0^2 \operatorname{Var}_\Pi$.

	Adult		**COMPAS**		**ACS Income**	
	DP	EO	DP	EO	DP	EO
$\bar{\pi}$	0.02	0.15	0.02	0.02	0.05	0.15
Label-flipped components	✓	✓	✗	✓	✓	✓
δ_0, frontiers	10 values in $[0.005, 0.15]$				6 values in $[0.005, 0.15]$	
δ_0, fixed budget	0.005				0.01	

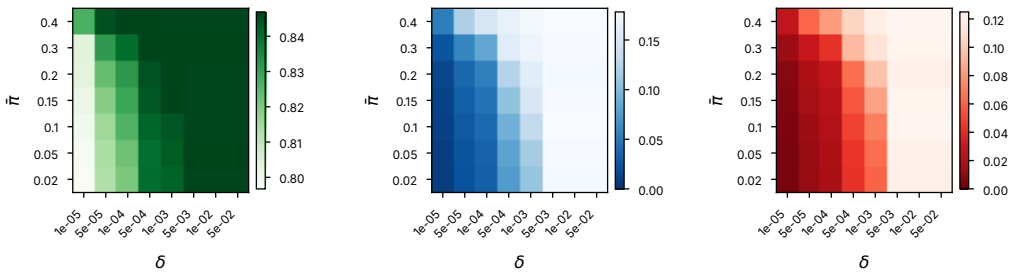
DRO-objective ablation. To isolate the contribution of the STABLE objective, we replace the SOCP by a distributionally robust objective while keeping everything else fixed: the same dictionary $\{\widehat{\eta}_k\}_{k=1}^K$, the same perturbation family and points $\pi'_1, \ldots, \pi'_m$ (Perturbation 1 for DP, Perturbation 2 for EO), and the same values of K, m and $\bar{\pi}$. We denote this variant $\text{STABLE}_{\text{DRO}}$. Following SA-DRO [26], the mixture weights solve

$$\min_{\boldsymbol{\alpha} \in \Delta_K} \max_{j \in [m]} \Big[(1-\lambda) \sum_{k=1}^K \alpha_k \widehat{R}_{j,k} - \rho \, (\pi'_j - \widehat{p}_1)^2 \Big] + \lambda \Big| \sum_{k=1}^K \alpha_k \, \widehat{\gamma}_k \Big|,$$

where $\widehat{R}_{j,k} = \frac{1}{n} \sum_{i=1}^n w_{\pi'_j}(Z_i) \, \ell(\widehat{\eta}_k(X_i), Y_i)$ is the importance-weighted log-loss of the k-th predictor at π'_j, $\widehat{p}_1 = \widehat{\mathbb{P}}(G = 1)$ on the training set, and $\widehat{\gamma}_k$ is the signed group gap of $\widehat{\eta}_k$ (difference of mean predictions between groups for DP, restricted to $Y = 1$ for EO). Since $\widehat{\gamma}$ is linear in $\boldsymbol{\alpha}$, the fairness term is computed in closed form, and the problem is a linear program, solved with CLARABEL via `cvxpy`. We fix $\rho = 0.9$, the default of Lei et al. [26], and sweep the fairness weight $\lambda \in [0, 1)$. Because λ has no correspondence with the fairness budget of the other methods, it is swept on its own grid, chosen per dataset and constraint to cover the transition from the ERM solution to the fair regime: $\{0, 0.3, 0.35, 0.4, 0.45, 0.5, 0.7\}$ on Adult (DP), $\{0, 0.02, 0.05, 0.1, 0.15, 0.2, 0.3\}$ on Adult (EO), $\{0, 0.1, 0.3, 0.5, 0.7, 0.85, 0.95, 0.99\}$ on COMPAS, and $\{0, 0.05, 0.1, 0.2, 0.3, 0.35, 0.4, 0.45, 0.5, 0.6, 0.7, 0.85, 0.95, 0.99\}$ on ACS Income. This variant is an ablation of our method rather than a reproduction of SA-DRO, which trains a neural network, relies on kernel smoothing of the fairness term, and is defined for DP only.

Hyperparameter selection. STABLE learns a convex mixture of K weighted-ERM predictors, each being an ℓ_2-regularised logistic regression (regularisation $C = 1$, `lbfgs` solver), optionally complemented by $2K_c$ label-flipped components ($K_c = 10$), and the two constant predictors (see *Dictionary enrichment* below). The mixture is optimised via a SOCP evaluated on m stratified perturbation points drawn from Π. We fix $K = 30$ (number of mixture components) and $m = 60$ (number of SOCP constraint evaluation points) across all experiments, as preliminary runs showed no sensitivity beyond these values.

The hyperparameters of STABLE are $\bar{\pi}$, the lower bound of the perturbation interval $\Pi = [\bar{\pi}, 1 - \bar{\pi}]$, δ, the SOCP feasibility threshold controlling the trade-off between fairness and accuracy, and the use of the label-flipped components. $\bar{\pi}$ is selected via grid search over $(\bar{\pi}, \delta)$ on a fixed held-out validation split (seed 0 for Adult and COMPAS, three validation states for ACS Income), optimising the deploy gap $\max(|\Phi(h; 0) - t|, |\Phi(h; 1) - t|)$. The search grids are: $\bar{\pi} \in \{0.02, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40\}$, $\delta \in \{10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$. δ is then not fixed but swept: we set $\delta = \delta_0^2 \operatorname{Var}_\Pi$, with $\operatorname{Var}_\Pi = (1 - 2\bar{\pi})^2/12$, and sweep δ_0 over a grid shared with the baselines, $\delta_0 \in \{0.005, 0.0075, 0.01, 0.015, 0.02, 0.03, 0.05, 0.075, 0.1, 0.15\}$ on Adult and COMPAS and $\delta_0 \in \{0.005, 0.01, 0.02, 0.04, 0.08, 0.15\}$ on ACS Income. The two dictionaries, with and without the label-flipped components, reach similar deploy gaps; on the same validation split, we keep the one with the higher mean accuracy over this δ_0 grid, and the dictionary with label-flipped components when the difference is below 0.005. All reported results use the selected $\bar{\pi}$ and dictionary and are averaged over 10 independent random seeds for Adult and COMPAS and 5 for ACS Income (evaluated on independent test states). The selected values are reported in Table 3. We also report the sensitivity analysis for the grid on Adult with DP constraints in Figure 4; the other settings show similar trends. All baselines are swept on the same δ_0 grid. For Shifty, the fairness tolerance is $\varepsilon = \delta_0$, and seeds for which Shifty returns No-Solution-Found are discarded (up to 5 of the 10 seeds at a given budget on



(a) Accuracy↑ (b) $\mathcal{U}(h)\downarrow$ (c) Deployment Gap ↓

Figure 4: Hyperparameter sensitivity on the Adult dataset (DP constraint, seed 0). Each cell reports the value of accuracy (left), $\mathcal{U}(h)$ (center) and deploy gap $\max(|\Phi(h;0)-t|, |\Phi(h;1)-t|)$ (right) as a function of threshold δ and the perturbation lower bound $\bar{\pi}$.

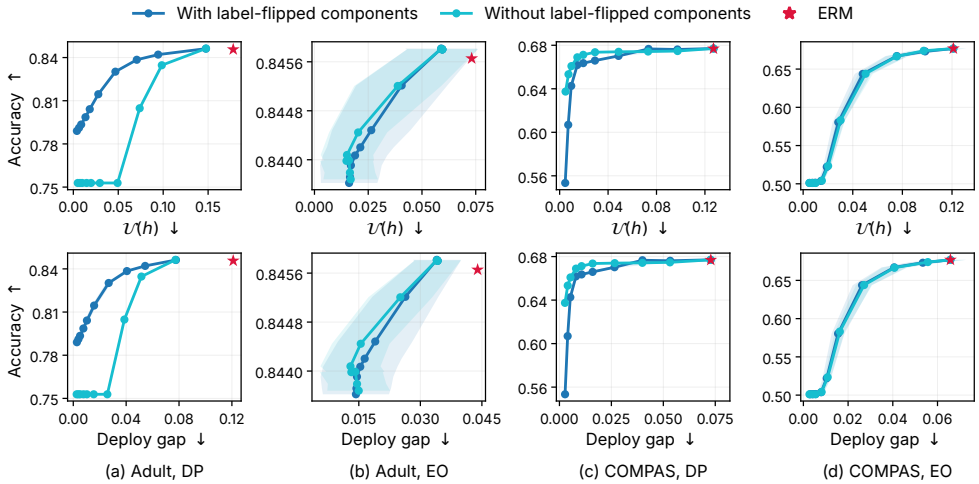


Figure 5: Dictionary ablation: STABLE with and without the label-flipped components, on the same budget grid. Accuracy against $\mathcal{U}(h)$ (top) and against the deployment gap (bottom); mean over 10 seeds, shaded regions show $\pm$ one standard deviation along the horizontal axis.

Adult and COMPAS, and one run out of 60 on ACS Income); all runs used the unknown-shift setting (`setting`$=2$) with $\alpha = 0.25$. ERM and all base classifiers use ℓ_2-regularised logistic regression with $C = 1$ and the `lbfgs` solver. ERM+DP and ERM+EO (Exponentiated Gradient [1], with the demographic-parity and true-positive-rate-parity constraints of `fairlearn`) use $\varepsilon = \delta_0$ as the fairness tolerance. For PostProc, the constraint slack is $\varepsilon = \delta_0$, the inverse temperature of the entropic regularisation is set to $\frac{1}{2}\sqrt{N}\log_2 N$, where N is the size of the unlabeled calibration set (half of the training set), and the number of stochastic optimization steps is $T = 20{,}000$ (25,000 on ACS Income). STABLE$_{\text{DRO}}$ uses the same dictionary as STABLE and is swept over its own grid of λ (see *DRO-objective ablation* above).

Dictionary enrichment. When the dictionary contains only weighted-ERM components, their slopes $\gamma(\widehat{\eta}_k)$ might only span a narrow range and none is close to zero. A small slope can then only be reached by putting weight on the constant predictors added for feasibility, which pulls the mixture toward a constant classifier. Adding $2K_c = 20$ components trained on label-flipped data ($K_c = 10$ flipping rates, applied to each group) provides the mixture with a wider range of slopes. Figure 5 shows that these components remove the accuracy collapse on Adult under DP, have no visible effect under EO, and lower accuracy at the tightest budgets on COMPAS under DP, with unchanged unfairness and deployment gap in all cases. Their use is therefore selected on validation data (Appendix E.1); they are kept in all settings except COMPAS under DP.

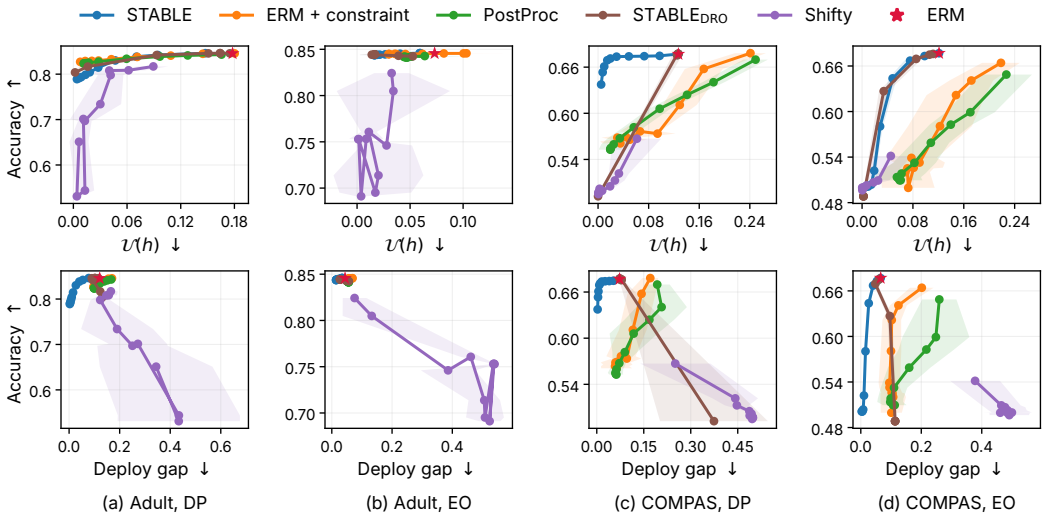


Figure 6: Pareto frontiers with Shifty. One point per fairness budget, mean over 10 seeds (seeds for which Shifty returns no solution are discarded); shaded regions show $\pm$ standard deviation along the horizontal axis.

E.2 Additional comparison to Shifty

Figure 6 adds Shifty to the frontiers of Figure 1. Shifty reaches low values of $\mathcal{U}(h)$ in all settings, but at a large cost in accuracy: on COMPAS its accuracy stays close to 0.5, and on Adult it drops by up to 0.3 with respect to ERM. Its deployment gap, dominated by the offset of its level from t, increases as the budget tightens and is at least as large as that of ERM in every setting. Results on ACS Income, including Shifty, are reported in Table 4 (Appendix E.3).

E.3 ACS geographic experiment.

The ACS Income task [15] is derived from the 2018 American Community Survey and covers over 1.5 million working-age adults across all US states. We train on California ($n = 80{,}000$ subsampled) and select hyperparameters on 3 validation states ($n = 30{,}000$ per state). We use 11 test states, with White resident proportions ranging from 68% to 95%, never seen during hyperparameter selection.

On the ACS geographic experiment under EO (see Figure 7), STABLE remains the closest to t across all held-out states, with the flattest slope $\pi_{\text{state}} \mapsto \Phi_{EO}(h; \pi_{\text{state}})$ among the methods of the figure. ERM+EO and PostProc stay above t in every state, with a steeper slope. The slope of STABLE$_{\text{DRO}}$ is similar to the one of STABLE, however its level remains further from t.

Table 4 reports the results at the same budget under both DP and EO. In both settings, STABLE attains the smallest worst-state deployment gap, with an accuracy higher than or comparable to the other methods. STABLE$_{\text{DRO}}$ reaches an unfairness comparable to STABLE but a larger deployment gap. Shifty attains the smallest $\mathcal{U}(h)$, at the cost of a large accuracy loss and the largest deployment gap. Note that a flat prediction rate alone does not imply a small deployment gap: ERM and Shifty have the smallest slopes, but their level remains further from t than that of STABLE.

E.4 The effect of t

We denote by t_{ERM} the target level estimated from ERM and used throughout our experiments. In addition, we consider a grid of 25 values for t in $[0.05, 0.95]$. We train STABLE under both DP and EO on Adult dataset for all the values of t, using the hyperparameters tuned at t_{ERM}, over 10 random seeds. Figure 8 shows that, for both constraints, the unfairness and the deployment gap remain low over the whole grid, well below those of ERM; under EO, they are slightly larger around t_{ERM}. By contrast, the accuracy depends on t. It is highest for t close to t_{ERM}, where it is close to that of ERM, and decreases as t moves away from it. For extreme values of t, it drops to the accuracy of a constant classifier.

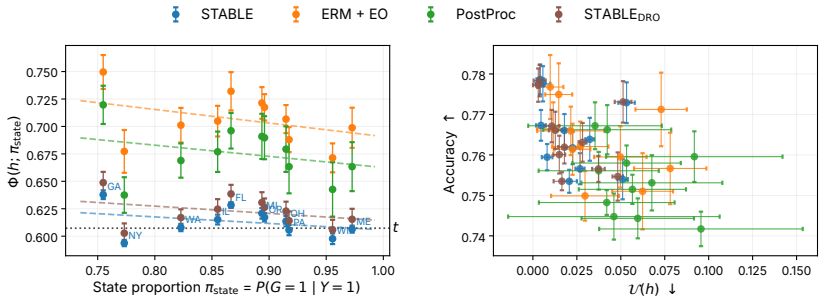


Figure 7: Geographic generalization on ACS Income under EO, at a fixed fairness budget. (left) Prediction rate $\Phi(h; \pi_{\text{state}})$ against $\pi_{\text{state}} = P(G = 1 \mid Y = 1)$ in each held-out state; dashed lines: linear fits, dotted line: target t. (right) Accuracy against $\mathcal{U}(h)$. Each point is one state, mean $\pm$ std over 5 seeds.

Table 4: ACS Income, held-out states, at the fixed budget of Figures 3 and 7 ($\delta_0 = 0.01$; STABLE$_{\text{DRO}}$ at its least unfair λ): mean accuracy and mean $\mathcal{U}(h)$ over the states, worst-state deployment gap, and absolute slope of $\pi_{\text{state}} \mapsto \Phi$. Mean $\pm$ std over 5 seeds.

	DP				EO			
	Acc. ↑	$\mathcal{U}(h)$ ↓	Deploy gap ↓	|slope| ↓	Acc. ↑	$\mathcal{U}(h)$ ↓	Deploy gap ↓	|slope| ↓
ERM	$0.751_{\pm.005}$	$0.122_{\pm.019}$	$0.032_{\pm.007}$	$0.041_{\pm.021}$	$0.751_{\pm.005}$	$0.081_{\pm.022}$	$0.060_{\pm.011}$	$0.025_{\pm.021}$
STABLE	$0.764_{\pm.006}$	$0.052_{\pm.003}$	$0.015_{\pm.002}$	$0.067_{\pm.019}$	$0.765_{\pm.004}$	$0.022_{\pm.002}$	$0.030_{\pm.002}$	$0.059_{\pm.006}$
ERM + constraint	$0.760_{\pm.007}$	$0.079_{\pm.004}$	$0.055_{\pm.007}$	$0.146_{\pm.022}$	$0.763_{\pm.007}$	$0.036_{\pm.008}$	$0.142_{\pm.015}$	$0.124_{\pm.027}$
PostProc	$0.753_{\pm.004}$	$0.040_{\pm.021}$	$0.069_{\pm.008}$	$0.217_{\pm.027}$	$0.754_{\pm.005}$	$0.057_{\pm.023}$	$0.112_{\pm.019}$	$0.105_{\pm.078}$
STABLE$_{\text{DRO}}$	$0.758_{\pm.007}$	$0.048_{\pm.006}$	$0.036_{\pm.007}$	$0.067_{\pm.017}$	$0.765_{\pm.004}$	$0.022_{\pm.001}$	$0.042_{\pm.010}$	$0.066_{\pm.004}$
Shifty	$0.651_{\pm.006}$	$0.006_{\pm.002}$	$0.403_{\pm.010}$	$0.042_{\pm.017}$	$0.640_{\pm.001}$	$0.003_{\pm.001}$	$0.605_{\pm.005}$	$0.011_{\pm.009}$

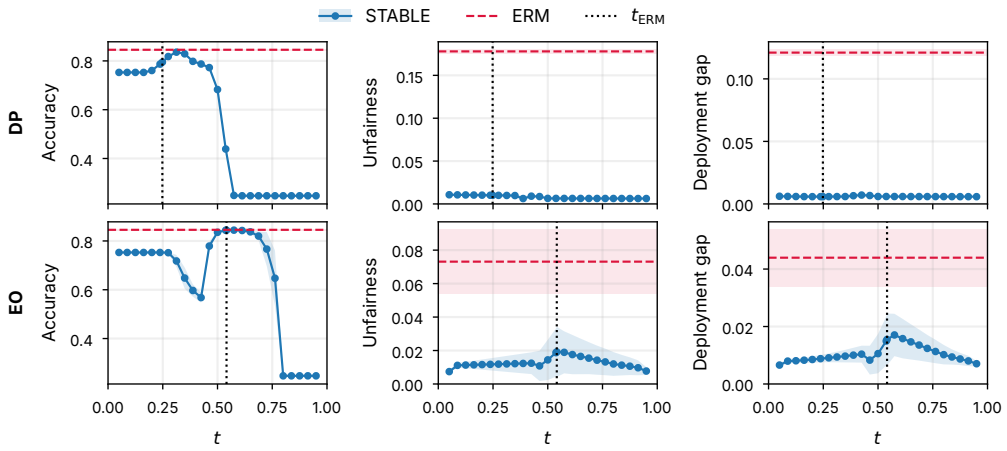


Figure 8: The effect of the target t on STABLE (Adult; top: DP, bottom: EO). Dotted line: t_{ERM}. Shaded regions show $\pm$ std over 10 random seeds.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We introduce a stability-based view of fairness in Definition 4. We show its equivalence to classical notions of fairness in Propositions 1,2,3. We provide generalization bounds in Theorem 1. We illustrate numerical examples in Section 5. Further examples and discussions are included in the Supplementary Material.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the Conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the assumptions are presented and discussed in the main body (Assumptions 1, 2, 3, 4, 5). Due to space limitations, all the proofs are deferred to the Supplementary Material.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the experiments are reproducible. The experiments validate the claims of the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The considered datasets are benchmark datasets. The source code link is provided in Section 5.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the information regarding the training and the evaluation is provided in Section 5, with further discussion and details in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All the reported results are evaluated on 10 independent random seeds for Adult and Compas datasets, and 5 independent random seeds for ACS. The mean results with the respected variances, are presented with all results. Additional discussion is included in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information regarding computer resources is provided in the beginning of Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Since the paper studies algorithmic fairness, its broader positive societal motivation is inherent to the problem setting and is stated in the introduction, as well as in Appendix A.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce high-risk models or datasets of this kind, and the experiments use standard benchmark datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and benchmark methods used in the paper are referenced, and their use follows standard research practices.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a novel definition of fairness based on distributional stability, together with the associated methodology, theoretical results, and numerical results. These are documented in the main text and supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.