Fairness Transferability Subject to Bounded Distribution Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

Given an algorithmic predictor that is "fair" on some *source* distribution, will it still be fair on an unknown *target* distribution that differs from the source within some bound? In this paper, we study the *transferability of statistical group fairness* for machine learning predictors (*i.e.*, classifiers or regressors) subject to bounded distribution shift, a phenomenon frequently caused by user adaptation to a deployed model or a dynamic environment. Herein, we develop a bound characterizing such transferability, flagging potentially inappropriate deployments of machine learning for socially consequential tasks. We first develop a framework for bounding violations of statistical fairness subject to distribution shift, formulating a generic upper bound for transferred fairness violation as our primary result. We then develop bounds for specific worked examples, adopting two commonly used fairness definitions (*i.e.*, demographic parity and equalized odds) for two classes of distribution shift (*i.e.*, covariate shift and label shift). Finally, we compare our theoretical bounds to deterministic models of distribution shift as well as real-world data.

1 Introduction

Distribution shift is a common, real-world phenomenon that affects machine learning deployments when the *target* distribution of examples (features and labels) encountered by a learned policy diverges from the *source* distribution it was trained for. For socially consequential decisions guided by machine learning, such shifts in the underlying distribution can invalidate fairness guarantees and cause harm by exacerbating social disparities. Unfortunately, distribution shift can be technically difficult or impossible to model at training time. As we shall explain later, these changes may involve a variety of factors that require highly customized treatments. Nonetheless, we still wish to certify the robustness of the fairness of a policy on possible target distributions.

In this paper, we provide a general framework encompassing a wide array of settings for quantifying the robustness of statistical group fairness guarantees. We assume that the target distribution is drawn from a bounded domain, thus reducing the hard problem of modelling distribution shift dynamics to a more tractable, adversarial problem. By bounding fairness violations in this way, we can detect potentially inappropriate policy applications prior to deployment.

This work bridges a gap between recent literature on *domain adaptation*, which has largely focused on the transferability of prediction accuracy rather than fairness, and *algorithmic fairness*, which has typically considered static distributions or prescribed models of distribution shift. Our work is the first to systematically bound quantifiable violations of statistical group fairness while remaining agnostic to (1) the mechanisms responsible for distribution shift, (2) the method of measurement of such change between distributions, and (3) the specific statistical definition of fairness applied.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Our primary result is a bound on a policy's potential "violation of statistical group fairness", defined in terms of the differences in policy outcomes between groups, when applied to a target distribution shifted relative to the source distribution within known constraints. Such bounds are useful when a learned policy will be deployed on a distribution that differs from the training or target distribution with some uncertainty, *e.g.*, whenever a training population represents only a random sample of the target population. Such settings also arise when a target distribution may only be approximately modelled or forecast. For example, upon deducing loan-approval policies, strategic agents with unknown utilities may respond by changing their features or abstaining from classification, thus shifting the distribution of examples [15, 29, 33]. In addition, exogenous pressures, such as economic trends and noise, may also drive distribution shift. In Figure 1, we show how a real-world distribution shift in demographic and income data for US states between 2014 and 2018 may increase fairness violations while decreasing accuracy for a hypothetical classifier trained on the 2014 distribution. In such settings, it is useful to quantify how fairness guarantees transfer across distributions shifted within some bound, thus allowing the deployment of unfair machine learning policies to be avoided.

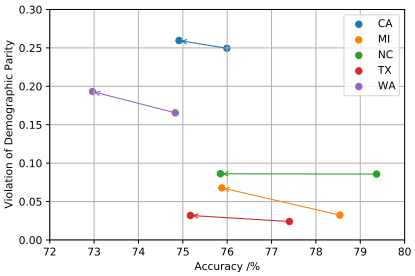


Figure 1: In Section 7, we evaluate our bounds against historical, temporal distribution shifts in demographics and income recorded by the US Census Bureau [10]. The above figure depicts changes to accuracy and violation of demographic parity when a classifier initially trained on US state-specific data for 2014 is reused on 2018 data, thus exemplifying the negative potential effects of distribution shift.

1.1 Related Work

Our work relates to a number of recently explored domains of study in the literature: Specifically, we share a similar setting with recent efforts on *domain adaptation*, which have largely focused on characterizing the effects of distribution shift on prediction performance rather than fairness. Additionally, our work builds on investigations of *algorithmic fairness* as well as the *dynamic modeling* of distributions in response to deployed machine learning policies. We reference specific prior work in these domains in Appendix B, in the supplementary material. The most closely related works to the present study focus on how certain measures of fairness are affected when policies are subject to specific types of distribution shift.

Fairness subject to Distribution Shift: A number of recent studies have considered specific examples of fairness transferability subject to distribution shift [25, 7, 27, 24]. In particular, Schumann et al. [25] examine *equality of opportunity* and *equalized odds* as definitions of group fairness subject to distribution shifts quantified by an $\mathcal{H}$-divergence function; Coston et al. [7] consider *demographic parity* subject to a *covariate shift* assumption while group identification remains unavailable to the classifier; Singh et al. [27] focus on common group fairness definitions for binary classifiers subject to a class of distribution shift that generalizes covariate shift and label shift by preserving some conditional probability between variables; and Rezaei et al. [24] similarly consider common binary classification fairness definitions, and equalized odds in particular, subject to covariate shift. While we address similar settings to these works as special cases of our bound, we propose a unifying formulation for a broader class of statistical group fairness definitions and distribution shifts. In doing so, we recognize that particular settings recommend themselves to more natural measures of distribution shift, providing examples in Section 4.1, Section 4.2, and Section 5).

1.2 Our Contributions

Our primary contribution is formulating a general upper bound for a policy's violation of statistical group fairness subject to group-dependent distribution shifts (i.e., Equation (11)). We characterize the space of statistical group fairness definitions and possible distribution shifts by appeal to premetric functions (Definition 2.1). Our work explores common sets of simplifying assumptions for this bound as special cases, yielding tractable calculations for several familiar combinations of fairness definitions and subcases of distribution shift (Theorem 4.1, Theorem 5.2) with readily interpretable results. We compare our theoretical bounds to prescribed models of distribution shift in Section 6 and to real-world data in Section 7. Bounding violations of fairness subject to distribution shift allows us to recognize and avoid potentially inappropriate deployments of machine learning.

2 Formulation

The supplementary material includes a table of notation (Appendix A) and all proofs (Appendix F).

2.1 Algorithmic Prediction

We consider two distributions, $\mathcal{S}$ (*source*) and $\mathcal{T}$ (*target*), each defined as a probability distribution over the space of "examples", where each example defines values for three random variables:

- X: a **feature**, *e.g.*, x, with arbitrary domain $\mathcal{X}$.
- Y: a **label**, *e.g.*, y, with arbitrary domain $\mathcal{Y}$.
- G: a **group**, *e.g.*, g or h, with finite, countable domain $\mathcal{G}$.

The predictor's policy, π, intended for $\mathcal{S}$ but used on $\mathcal{T}$, defines a fourth variable for each example:

- $\hat{Y}$: a **predicted label**, *e.g.*, $\hat{y}$, with domain $\hat{\mathcal{Y}} = \mathcal{Y}$.

Using $\mathcal{P}(\cdot)$ to denote the space of probability distributions over some domain, we denote the space of distributions over examples as $\mathbb{D} \coloneqq \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{G})$, such that $\mathcal{S}, \mathcal{T} \in \mathbb{D}$. It will also be useful for us to notate the space of distributions over example *outcomes* associated with a given policy as $\mathbb{O} \coloneqq \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{Y}})$ and the space of distributions over of *group-specific examples* as $\mathbb{G} \coloneqq \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Without loss of generality, we allow the prediction policy π to be stochastic, such that, for any combination (x, g), the predictor effectively samples $\hat{Y}$ from a corresponding probability distribution $\pi(x, g)$. Stochastic classifiers arise in various constrained optimization problems and have been shown to be useful for making problems with custom losses or fairness constraints tractable [8, 14, 22].

We denote the space of nondeterministic policies as $\Pi \coloneqq (\mathcal{X} \times \mathcal{G} \to \mathcal{P}(\hat{\mathcal{Y}}))$, *e.g.*, $\pi \in \Pi$, and utilize the natural transformations that relate the spaces of distributions $\mathbb{D}$, policies Π, and outcomes $\mathbb{O}$:

$$\Pr_{\pi,\mathcal{T}}(\hat{Y}{=}\hat{y}, X{=}x, G{=}g) = \Pr_{\hat{Y}\sim\pi(x,g)}(\hat{Y}{=}\hat{y}) \cdot \Pr_{X,G\sim\mathcal{T}}(X{=}x, G{=}g) \tag{1}$$

We will abuse the $\Pr$ notation for both probability *density* and probability *mass* functions where appropriate and *distributions* when the random variables inside the expression are unconstrained.

2.2 Statistical Group-Fairness

We next define a broad class of *disparity* functions $\Delta^\star \colon \Pi \times \mathbb{D} \to \mathbf{R}$ representing how "unfair" a given policy is for a given distribution, *e.g.*, writing $\Delta^\star(\pi, \mathcal{T})$.

Definition 2.1. We define a *premetric* Ψ on the space of distributions p with respect to q by the properties $\Psi(p \parallel q) \geq 0$ and $\Psi(p \parallel p) = 0$ for all p, q, and refer to the value of Ψ as a "shift".

Definition 2.2. We define *statistical group disparity* $\Delta^\star$ in terms of the symmetrized shifts between group-specific outcome distributions, each mapped from a policy π and a distribution $\mathcal{T}$. We measure shifts between outcome distributions with a given premetric $\Psi \colon \mathbb{O}^2 \to \mathbf{R}$.

$$\Delta^\star(\pi, \mathcal{T}) \coloneqq \sum_{g,h\in\mathcal{G}} \Psi\Big(\Pr_{\pi,\mathcal{T}}(X, Y, \hat{Y} \mid G{=}g) \;\Big\|\; \Pr_{\pi,\mathcal{T}}(X, Y, \hat{Y} \mid G{=}h) \Big) \tag{2}$$

In Definition 2.2, Ψ quantifies the statistical differences in outcomes between groups that are "unfair", where a value of 0 implies perfect fairness. In this work, we assume that Ψ is the same for all g, h and that $\Delta^\star$ is insensitive to relative group size $\Pr(G)$.

Examples Familiar applications of Definition 2.2 include *demographic parity* (DP) and *equalized odds* (EO). A policy satisfying DP gives agents chosen at random any group a uniform chance of receiving a given binary classification, *i.e.* $\hat{\mathcal{Y}} = \{0, 1\}$. We may measure the violation of DP as

$$\Delta^\star{}_{\mathsf{DP}}(\pi, \mathcal{T}) \coloneqq \sum_{g,h\in\mathcal{G}} \Big| \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) - \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}h) \Big| \tag{3}$$

The associated premetric Ψ_{DP} on $p, q \in \mathbb{O}$ is

$$\Psi_{\mathsf{DP}}(p \parallel q) = \left| \Pr_{p}(\hat{Y}{=}1) - \Pr_{q}(\hat{Y}{=}1) \right| \tag{4}$$

To satisfy EO, for binary $\mathcal{Y} = \{0, 1\}$, π must maintain group-invariant true positive and false positive classification rates. We may measure the violation of EO as

$$\Delta^\star{}_{\mathsf{EO}}(\pi, \mathcal{T}) \coloneqq \sum_{g,h \in \mathcal{G}} \sum_{y \in \mathcal{Y}} \left| \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g, Y{=}y) - \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}h, Y{=}y) \right| \tag{5}$$

The associated premetric Ψ_{EO} on $p, q \in \mathbb{O}$ is

$$\Psi_{\mathsf{EO}}(p \parallel q) = \sum_{y \in \mathcal{Y}} \left| \Pr_{p}(\hat{Y}{=}1 \mid Y{=}y) - \Pr_{q}(\hat{Y}{=}1 \mid Y{=}y) \right| \tag{6}$$

Note that the restriction of EO to the $(Y = 1)$ case is known as *Equal Opportunity* (EOp).

We remark that Definition 2.2 provides a unifying representation for a wide array of statistical group "unfairness" definitions and may be used with inequality constraints. That is, we may recover many working definitions of fairness that effectively specify a maximum value of some $\Delta^\star$:

Definition 2.3. A policy π is **ϵ-fair** with respect to $\Delta^\star$ on distribution $\mathcal{T}$ iff $\Delta^\star(\pi, \mathcal{T}) \leq \epsilon$.

2.3 Vector-Bounded Distribution Shift

Suppose, after developing policy π for distribution $\mathcal{S}$, we "realize" some new distribution $\mathcal{T}$ on which the policy is actually operating. This realization may be the consequence of sampling errors during the learning process, strategic feedback to our policy, random processes, or the reuse of our policy on a new distribution for which retraining is impractical. Our goal is to bound $\Delta^\star(\pi, \mathcal{T})$ given knowledge of $\Delta^\star(\pi, \mathcal{S})$ and some notion of how much $\mathcal{T}$ differs from S.

Definition 2.4. $K(p \parallel q)$ is a *divergence* if and only if for all p and q, $K(p \parallel q) \geq 0$ and $K(p \parallel q) = 0 \iff q = p$.

Definition 2.5. Define the group-vectorized shift $\mathbf{D}$ as $\mathcal{S}$ mutates into $\mathcal{T}$ as

$$\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \coloneqq \sum_{g} \mathbf{e}_g D_g\big(\Pr_{\mathcal{T}}(X, Y \mid G{=}g) \parallel \Pr_{\mathcal{S}}(X, Y \mid G{=}g)\big) \tag{7}$$

where $\mathbf{e}_g$ represents a unit vector indexed by g, and each $D_g \colon \mathbb{G}^2 \to \mathbf{R}$ is a divergence. Note that each D_g also defines a premetric, but not necessarily a divergence, on $\mathbb{D}$.

Assumption 2.6. Let there exist some vector $\mathbf{B} \succeq 0$ bounding $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$, where $\preceq$ and $\succeq$ denote element-wise inequalities.

In Assumption 2.6, $\mathbf{B}$ limits the possible distribution shift as $\mathcal{S}$ mutates into $\mathcal{T}$, without requiring us to specify a model for how distributions evolve. When modelling distribution shift requires complex dynamics (*e.g.*, when agents learn and respond to classifier policy), we reduce a potentially difficult dynamical problem to a more tractable, adversarial problem to achieve a bound.

Lemma 2.7. *For all π, $\Delta^\star$, and $\mathbf{D}$, when $\mathbf{B} = 0$, $\Delta^\star(\pi, \mathcal{S}) = \Delta^\star(\pi, \mathcal{T})$.*

Lemma 2.7 indicates that, for a fixed policy π, a change in disparity requires a measurable shift in distributions, confirming intuition.

Restricted Distribution Shift Common assumptions that restrict the set of allowed distribution shifts include *covariate shift* and *label shift*. For covariate shift, the distribution of *labels* conditioned on *features* is preserved across distributions for all groups:

$$\textbf{Covariate shift} \text{ implies } \Pr_{\mathcal{T}}(Y \mid X, G) = \Pr_{\mathcal{S}}(Y \mid X, G) \tag{8}$$

For label shift, the distributions of *features* conditioned on *labels* is preserved across distributions for all groups:

$$\textbf{Label shift} \text{ implies } \Pr_{\mathcal{T}}(X \mid Y, G) = \Pr_{\mathcal{S}}(X \mid Y, G) \tag{9}$$

In Section 4, we explore a deterministic model of a population's response to classification, resulting in covariate shift. We do the same in Section 5 for label shift.

3 General Bounds

We formulate a primary bound in Equation (11), which we subsequently consider under simplifying special cases. *We defer all subsequently omitted proofs for this manuscript to the Appendix.*

Given an element-wise bound $\mathbf{B}$ on the vector-valued shift $\mathbf{D}(\mathcal{T} \parallel \mathcal{S})$ (Assumption 2.6) we may bound the disparity $\Delta^\star$ of policy π on $\mathcal{T}$ by its supremum value over all possible $\mathcal{T}$ within the bound.

Definition 3.1. Define the supremum value v for $\Delta^\star$ subject to $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$ as

$$v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) \coloneqq \sup_{\mathbf{D}(\mathcal{T}\parallel\mathcal{S})\preceq\mathbf{B}} \Delta^\star(\pi, \mathcal{T}) \tag{10}$$

$$\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B} \implies \Delta^\star(\pi, \mathcal{T}) \leq v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) \tag{11}$$

In general, our strategy is to exploit the mathematical structure of the setting encoded by $\Delta^\star$ (*i.e.*, Ψ) and $\mathbf{D}$ to obtain an upper bound for v defined in Equation (10). We first explore general cases of simplifying assumptions before presenting worked special examples for frequently encountered settings. Finally, we compare the resulting theoretical bounds to numerical results and simulations.

3.1 Lipshitz Conditions

The value of v defines a scalar field over the domain $\mathcal{T}$ and thus a conservative vector field $\mathbf{F}$:

$$\mathbf{F} = \nabla_{\mathbf{b}} v \tag{12}$$

For any curve in $\mathbb{D}$ from $\mathcal{S}$ to $\mathcal{T}$, bounds of the form $\mathbf{F} \preceq \mathbf{L}$ for some constant $\mathbf{L}$ along the curve imply a Lipshitz bound on $\Delta^\star$. We visualize a bound in Figure 2 for all possible curves in the region $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$.

Theorem 3.2 (Lipshitz Upper Bound)**.** *If there exists an* $\mathbf{L}$ *such that* $\nabla_{\mathbf{b}} v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{b}) \preceq \mathbf{L}$*, everywhere along some curve from* $\mathcal{S}$ *to* $\mathcal{T}$ *then*

$$\Delta^\star(\pi, \mathcal{T}) \leq \Delta^\star(\pi, \mathcal{S}) + \mathbf{L} \cdot \mathbf{B} \tag{13}$$

The utility of Theorem 3.2 arises when a Lipshitz condition $\mathbf{L}$ is known, but direct computation of v is difficult. We provide an example of a Lipshitz bound in Section 5.

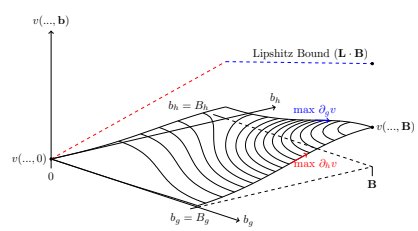


Figure 2: A Lipshitz bound for all curves in the $(0, \mathbf{B})$ $\mathbb{D}$-hyperrectangle.

3.2 Subadditivity Conditions

Definition 3.3. Define w as the maximum *increase* in disparity subject to $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$, *i.e.*, $w(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) \coloneqq v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) - \Delta^\star(\pi, \mathcal{S})$.

Theorem 3.4. *Suppose, in the region* $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$*, that* w *is subadditive in its last argument. That is,* $w(..., \mathbf{a}) + w(..., \mathbf{c}) \geq w(..., \mathbf{a} + \mathbf{c})$ *for* $\mathbf{a}, \mathbf{c} \succeq 0$ *and* $\mathbf{a} + \mathbf{c} \preceq \mathbf{B}$*. If* w *is also locally differentiable, then a first-order approximation of* $w(..., \mathbf{b})$ *evaluated at* 0*,* i.e.*,*

$$\mathbf{L} = \nabla_{\mathbf{b}} w(..., \mathbf{b})\big|_{\mathbf{b}=0} = \nabla_{\mathbf{b}} v(..., \mathbf{b})\big|_{\mathbf{b}=0} \tag{14}$$

provides an upper bound for $v(..., \mathbf{B})$*:*

$$v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) \leq \Delta^\star(\pi, \mathcal{S}) + \mathbf{L} \cdot \mathbf{B} \tag{15}$$

Theorem 3.4 notes that "diminishing returns" in the change of $\Delta^\star$ as the difference of $\mathcal{T}$ with respect to $\mathcal{S}$ is increased implies a bound on $\Delta^\star$ in terms of its local sensitivity to $\mathbf{D}$ at $\mathcal{S}$ (*i.e.*, using a first-order Taylor approximation). Note that, if w is concave in the bounded region, it is also subadditive in the bounded region, but the converse is not true, nor does the converse imply Lipshitzness.

3.3 Geometric Structure

It may happen that $\Psi \colon \mathbb{O}^2 \to \mathbf{R}$ and each $D_g \colon \mathbb{G}^2 \to \mathbf{R}$ share structure that permits a geometric interpretation of distribution shift. While the utility of this observation depends on the specific properties of Ψ and $\mathbf{D}$, we demonstrate a worked example building on Section 4.2 in Appendix D, in which we allow ourselves to select a suitable $\mathbf{D}$ for ease of interpretation.

We proceed to consider worked examples that adopt common assumptions limiting the form of distribution shift and apply common definitions of statistical group fairness.

4 Covariate Shift

4.1 Demographic Parity

The simplest way to work with Equation (10) is to bound the supremum v. We first consider demographic parity (Equation (3)) for $\mathcal{Y} = \{0, 1\}$ and $\mathcal{G} = \{g, h\}$, subject to covariate shift (Equation (8)). We find that the form of $\Delta^\star_{\mathsf{DP}}$ subject to this simplifying assumption recommends itself to a natural choice of $\mathbf{D}$. We first define a *reweighting coefficient* ω_g

$$\omega_g(x) \coloneqq \frac{\Pr_{\mathcal{T}}(X{=}x \mid G{=}g)}{\Pr_{\mathcal{S}}(X{=}x \mid G{=}g)} \tag{16}$$

Theorem 4.1. *For demographic parity between two groups under covariate shift, denoting, for each g, $\beta_g \coloneqq \Pr_{\pi,\mathcal{S}}(\hat{Y}{=}1 \mid G{=}g)$,*

$$\Delta^\star{}_{DP}(\pi, \mathcal{T}) \leq \Delta^\star{}_{DP}(\pi, \mathcal{S}) + \sum_g \left(\beta_g(1-\beta_g) \operatorname*{Var}_{\mathcal{S}}[\omega_g(x)]\right)^{1/2} \tag{17}$$

We notice that $\mathrm{Var}_{\mathcal{S}}[\omega_g(x)]$ recommends itself as a suitable divergence D_g from $\mathcal{S}$ to $\mathcal{T}$. Using basis vectors $\mathbf{e}_g$, we select $\mathbf{D}_{\mathsf{DP}}(\mathcal{T} \parallel \mathcal{S}) = \sum_g \mathbf{e}_g \,\mathrm{Var}_{\mathcal{S}}[\omega_g(x)]$. Thus comparing Equation (17) and the consequent of Equation (13), we can interpret $\Pr_{\pi,\mathcal{S}}(\hat{Y}{=}1)$ in Theorem 4.1, as an upper bound for the *average* value of $\nabla_{\mathbf{b}} v(\Delta^\star, \mathbf{D}, \pi, \mathcal{T}, \mathbf{b})$ along any curve from $\mathcal{S}$ to $\mathcal{T}$. To interpret this result, the closer $\Pr(\hat{Y}{=}1)$ is to 0.5 for any group, the more potentially sensitive the fairness of the policy is to distribution shifts for that group.

Corollary 4.2. *Theorem 4.1 may be generalized to multiple classes $\mathcal{Y} = \{1, 2, ..., m\}$ and multiple groups $\mathcal{G} \in \{1, 2, ..., n\}$:*

$$\Delta^\star_{DP}(\pi, \mathcal{T}) \coloneqq \sum_{y \in \mathcal{Y}} \sum_{g,h \in \mathcal{G}} \left| \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}y \mid G{=}g) - \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}y \mid G{=}h) \right| \tag{18}$$

$$\Delta^\star{}_{DP}(\pi, \mathcal{T}) \leq \Delta^\star{}_{DP}(\pi, \mathcal{S}) + \sum_y \sum_g \left(\beta_{g,y}(1-\beta_{g,y}) \operatorname*{Var}_{\mathcal{S}}[\omega_g(x)]\right)^{1/2} \tag{19}$$

where $\beta_{g,y} = \Pr\Big(\hat{Y}{=}y \mid G{=}g\Big)$.

We remark that in general, binary classification bounds may frequently be generalized to multi-class bounds by redefining fairness violations as a sum of binary-class fairness violations (*i.e.*, same-class *vs.* different-class labels) and summing the bounds on each.

4.2 Equal Opportunity

Consider an example using the $(Y{=}1)$-conditioned case of Equalized Odds—termed *Equal Opportunity* (EOp). Denoting, for each group g, the true positive rate $\beta_g^+ \coloneqq \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid Y{=}1, G{=}g)$ as an implicit function of π and $\mathcal{T}$, we define disparity for EOp as

$$\Delta^\star{}_{\mathsf{EOp}}(\pi, \mathcal{T}) \coloneqq \sum_{g,h \in \mathcal{G}} \left| \beta_g^+ - \beta_h^+ \right| \tag{20}$$

We may bound the realized value of $\Delta^\star{}_{\mathsf{EOp}}(\pi, \mathcal{T})$ by bounding β_g^+ for each group.

Theorem 4.3. *Subject to covariate shift and given $\mathbf{D}, \mathbf{B}$, assume extremal values for β_g^+,* i.e.*,*

$$\forall g, \;\; \big(D_g(\mathcal{T} \parallel \mathcal{S}) < B_g\big) \implies \big(l_g \leq \beta_g^+(\pi, \mathcal{T}) \leq u_g\big) \tag{21}$$

It follows that

$$v(\Delta^\star{}_{EOp}, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) \leq \max_{x_g \in \{l_g, u_g\}} \sum_{g,h} \left| x_g - x_h \right| \tag{22}$$

Corollary 4.4. *$\Delta^\star{}_{EOp}$ cannot exceed $\frac{|\mathcal{G}|^2}{4}$.*

In Appendix D, we bound the extremal values of β_g^+ by geometrically interpreting this quantity as an inner product on an appropriate vector space, utilizing the freedom to select an appropriate $\mathbf{D}$ for our setting.

5 Label Shift

Under label shift ($\Pr_{\mathcal{T}}(X|Y) = \Pr_{\mathcal{S}}(X|Y)$), violations of EO and EOp are invariant, because the independence of $\hat{Y}$ and Y given X implies $\Pr_{\pi,\mathcal{T}}(\hat{Y}|Y) = \Pr_{\pi,\mathcal{S}}(\hat{Y}|Y)$. We therefore focus on the violation of demographic parity (DP) (Equation (3)) subject to the label shift condition, treating a binary classification task over two groups for simplicity.

In this setting, we choose to measure group-specific distribution shifts between $\mathcal{S}$ and $\mathcal{T}$ by the change in proportion of true positive labels, which we refer to as the group *qualification rate* $Q_g(\mathcal{T}) \coloneqq \Pr_{\mathcal{T}}(Y = 1 \mid G = g)$:

$$D_g^{(\mathsf{DP})}(\mathcal{T} \parallel \mathcal{S}) \coloneqq \Big| Q_g(\mathcal{S}) - Q_g(\mathcal{T}) \Big| \leq B_g \tag{23}$$

Theorem 5.1. *A Lipshitz condition bounds* $\nabla_{\mathbf{b}} v(\Delta^\star{}_{DP}, \mathbf{D}^{(DP)}, \pi, \mathcal{S}, \mathbf{b})$ *when*

$$\frac{\partial}{\partial b_g} v(\Delta^\star{}_{DP}, \mathbf{D}^{(DP)}, \pi, \mathbf{S}, \mathbf{b}) \leq (|\mathcal{G}| - 1) \Big| \beta_g^+ - \beta_g^- \Big| \tag{24}$$

for true [false] positive rates $\beta_g^+ \coloneqq \Pr_\pi(\hat{Y}{=}1|Y{=}1, G{=}g)$ *[*$\beta_g^- \coloneqq \Pr_\pi(\hat{Y}{=}1|Y{=}0, G{=}g)$*].*

Because β_g^+ and β_g^- are invariant under label shift given a constant policy π, we elide their explicit dependence on the underlying distribution.

Theorem 5.2. *For demographic parity under label-shift,*

$$\forall g, \quad |Q_g(\mathcal{S}) - Q_g(\mathcal{T})| \leq B_g \implies \Delta^\star_{DP}(\pi, \mathcal{T}) - \Delta^\star_{DP}(\pi, \mathcal{S}) \leq (|\mathcal{G}| - 1) \sum_g B_g \left| \beta_g^+ - \beta_g^- \right| \tag{25}$$

Intuitively, the change in $\Delta^\star{}_{\mathsf{DP}}$ subject to label shift is given by $|\beta_g^+ - \beta_g^-|$, the marginal change in acceptance rates as agents change their qualifications Y, a distribution shift we measure by $|Q_g(\mathcal{S}) - Q_g(\mathcal{T})|$. When β_g^+ is close to β_g^-, the policy looks like a random classifier, and a label shift has limited effect on statistical group disparity. When $|\beta_g^+ - \beta_g^-|$ is large, indicating high classifier accuracy, the effect is larger. Our bound thus exposes a direct trade-off between accuracy and fairness transferability guarantees.

6 Comparisons to Synthetic Distribution Shifts (Demographic Parity)

6.1 Covariate Shift via Strategic Response

Let us consider a specific example of covariate shift (see Equation (8)) caused by a deterministic, group-independent model of *strategic response* in which *agents* are classified by and react to a binary classification policy characterized by group-specific feature thresholds:

$$\hat{Y} \sim \pi(x, g) = \begin{cases} 1 & \text{with probability 1 if } x \geq \tau_g \\ 0 & \text{with probability 1 otherwise} \end{cases} \tag{26}$$

Agents in each group g may modify their feature x to x', after learning τ_g, by incurring a cost $c_g(x, x') \geq 0$. Similar to [15], we define the agent's *utility* to be

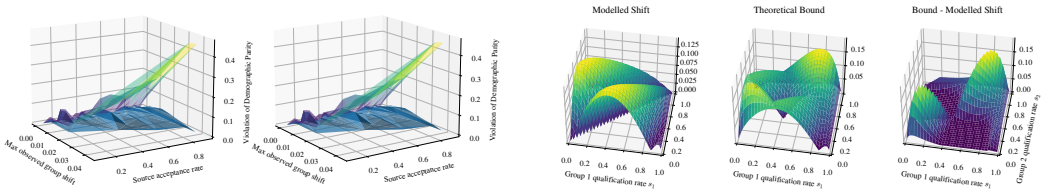
$$u_g(x, x') \coloneqq \beta_g(x') - \beta_g(x) - c_g(x, x'), \quad \beta_g(x) \coloneqq \Pr\Big(\hat{Y}{=}1 \mid X{=}x, G{=}g\Big), \forall g. \tag{27}$$

Contrary to the standard strategic classification setting, we do not assume that feature updates are a misreporting, but that such updates also have the potential of corresponding to updates to the underlying true qualification Y of each agent. This assumption has been made in a recent line of research in incentivizing improvement from human agents subject to such classification [3].

We will, however, assume all agents are rational utility (Equation (27)) maximizers. For a given threshold τ_g and manipulation budget m_g, the best response of an agent with original feature x is:

$$x' = \operatorname*{argmax}_z \; u_g(x, z), \qquad \text{such that } \; c_g(x, z) \leq m_g \tag{28}$$

Under some additional assumptions on the agent's best response behaviors (see Assumption F.3 - Assumption F.5 in Appendix F.1) we can specify the reweighting coefficient $\omega_g(x)$ for our setting (see Figure 5 in the supplementary material) for use with Theorem 4.1.



(a) A stereoscopic (cross-eye view) comparison between the bound of Section 4.1 (gradated) and simulated results for the model of Section 6.1 (blue) in response to a DP-fair classifier. A larger version, Figure 7, is provided in Appendix C in the supplementary material.

(b) A policy satisfying DP is subject to distribution shift prescribed by replicator dynamics. The local change in qualification rates subject to this policy follows the streamlines, while the local change in $\Delta^\star_{\mathsf{DP}}$ is represented with color and compared to the theoretical bound (Theorem 5.2).

Figure 3: Comparison to Synthetic Distribution

Proposition 6.1. *For our assumed setting of strategic response involving DP for two groups $\{g, h\}$, Theorem 4.1 implies*

$$\Delta^\star(\pi, \mathcal{T}) \leq \Delta^\star(\pi, \mathcal{S}) + \tau_g(1-\tau_g)\frac{2}{3}m_g + \tau_h(1-\tau_h)\frac{2}{3}m_h \tag{29}$$

The above result shows that two factors lead to a smaller difference between the source and target fairness violations: a less stochastic classifier (when the threshold τ_g is far away from 0.5) and a smaller manipulation budget m_g (diminishing agents' ability to adapt their feature). These factors both lead to less potential manipulation and result in a tighter upper bound for the fairness violation on $\mathcal{T}$.

6.2 Label Shift via Replicator Dynamics

We now evaluate our theoretical bound, Theorem 5.2, for demographic parity subject to label shift as prescribed by the replicator dynamics model of Raab and Liu [23]. This model assumes $\mathcal{X} = \mathbf{R}$, $\mathcal{Y} = \{0, 1\}$, and a monotonicity condition for $\mathcal{S}$ given by

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{\Pr_{\mathcal{S}}(X{=}x \mid Y{=}1)}{\Pr_{\mathcal{S}}(X{=}x \mid Y{=}0)} > 0 \tag{30}$$

Label shift under the discrete-time (t) replicator dynamics may be expressed in terms of group *qualification rate* $Q_g \coloneqq \Pr_t(Y{=}1 \mid G{=}g)$ and agent utilities, *i.e.*, group- and feature- independent values $U_{y,\hat{y}}$, such that agents adopt new labels with frequencies affine in the current popularity and average utility of the label in the agent's group. Denote the fractions of group-conditioned, feature-independent outcomes with $\rho_g^{y,\hat{y}} \coloneqq \Pr_t(\hat{Y}{=}\hat{y}, Y{=}y \mid G{=}g)$ and abbreviate the fraction-weighted utility as $u_g^{y,\hat{y}}(t) \coloneqq U_{y,\hat{y}}\rho_g^{y,\hat{y}}$. We may then represent the replicator dynamics as

$$Q_g[t+1] = \frac{u_g^{1,1}(t) + u_g^{1,0}(t)}{u_g^{1,1}(t) + u_g^{1,0}(t) + u_g^{0,0}(t) + u_g^{0,1}(t)} \tag{31}$$

To apply Theorem 5.2, we also observe that

$$|\beta_g^+ - \beta_g^-| = \frac{|\rho_g^{1,1} - \rho_g^{0,1}|}{\rho_g^{1,1} + \rho_g^{0,1}} \tag{32}$$

where β_g^+ and β_g^- represent the true positive rate and false positive rate for group g, respectively, and we use the change in qualification rate as our measurement of label shift, *i.e.*, $B_g = |Q_g[t{+}1] - Q_g[t]|$. When demographic parity is perfectly satisfied, we note that the acceptance rate $(\rho_g^{1,1} + \rho_g^{0,1})$ is group-independent.

Theorem 6.2. *For DP subject to label replicator dynamics,*

$$\left|\Delta^\star_{DP}(\pi, \mathcal{T}) - \Delta^\star_{DP}(\pi, \mathcal{S})\right| \leq \sum_g \left|Q_g[t+1] - Q_g[t]\right| \frac{|\rho_g^{1,1} - \rho_g^{0,1}|}{\rho_g^{1,1} + \rho_g^{0,1}} \tag{33}$$

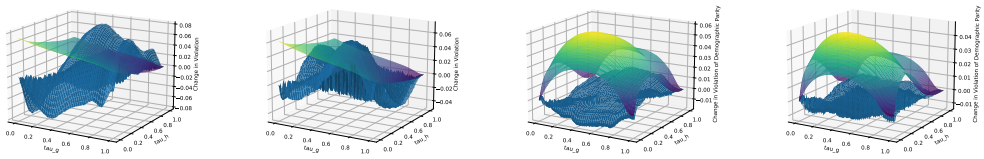
In Section 6.1, we graphically represent all possible states of an initially fair system (thus determining β and ρ as a result of Equation (30)) by the tuple of qualification rates for each group. With the dynamics prescribed by Equation (31), we depict the local evolution of the state subject to a fixed, locally DP-fair policy, as well as the rate of change for the violation of fairness, which we compare to the theoretical bound.

Interpreting our results, we note that the bound lacks information about the relative directions of the change in acceptance rates for each group, and thus over-approximates possible fairness violations when group acceptance rates shift the same direction. When group acceptance rates move in opposing directions, however, the bound gives excellent agreement with the modelled replicator dynamics.

7 Comparisons to Real-World Distribution Shifts (Demographic Parity)

We now compare our special-case theoretical bounds to real-world distribution shifts and hypothetical classifiers. We use American Community Survey (ACS) data provided by the US Census Bureau [13]. We adopt the sampling and pre-processing approaches following the `Folktables` package provided by Ding et al. [10][1] to obtain 1,599,229 data points. The data is partitioned by (1) all fifty US states and (2) years from 2014 to 2018. We use 10 features covering the demographic information used in the UCI Adult dataset [2], including age, occupation, education, *etc.*. as X for our model, select sex as binary protected group, i.e., $G \in \{g = \texttt{female}, h = \texttt{male}\}$, and set the label Y to correspond to whether an individual's annual income is greater than \$50K.

On this data, we train a set of group-dependent, linear threshold classifiers $\Pr_{\pi(x,g)}(\hat{Y}{=}1) = \mathbb{1}[\sigma(w \cdot x) > \tau_g]$, for a range of thresholds τ_g and τ_h for each source distribution. Here, $\sigma(\cdot)$ is the logistic function and w denotes a weight vector. We then consider two types of real-world distribution shift: (1) *geographic*, in which a model trained for one state is evaluated on other US state in the same year, and (2) *temporal*, in which a model trained for 2014 is evaluated on the same state in 2018.



(a) CA $\longrightarrow$ IL (b) CA $\longrightarrow$ NV (c) CA: 2014 $\rightarrow$ 2018 (d) TX: 2014 $\rightarrow$ 2018

Figure 4: Modelled change in DP violation subject to geographic and temporal distribution shift *vs.* direct application of the formulas of Theorem 4.1 and Theorem 5.2, respectively. The x-axis and y-axis of both figures represent the thresholds τ_g and τ_h, respectively used to train the policy.

Finally, we graphically compare the theoretical bounds of Theorem 4.1 and Theorem 5.2 for the increased violation of DP subject to covariate and label shift, respectively, to the simulated violations for our model and data in Figure 4, despite the fact that neither geographic nor temporal distribution shifts are guaranteed to satisfy the assumptions of covariate or label shift. We do this to evaluate the real-world applicability of the covariate and label shift assumptions in conjunction with our theoretical bounds. We provide addtional graphical comparisons in Appendix E.2.

8 Conclusion and Discussion

In this paper, we have developed a unifying framework for bounding the violation of statistical group fairness guarantees when the underlying distribution subject to algorithmic prediction is changed. We hope that this work can generate meaningful discussion regarding the viability of fairness guarantees subject to distribution shift, the bounds of adversarial attacks against algorithmic fairness, and evaluations of robustness with respect to algorithmic fairness. We believe that, just as published measurements are of limited use without reported uncertainties, fairness guarantees must come with bounds on robustness to distribution shift.

[1] This package is available at `https://github.com/zykls/folktables`.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[2] Catherine Blake. Uci repository of machine learning databases. 1998.

[3] Yatong Chen, Jialu Wang, and Yang Liu. Linear classifiers that encourage constructive adaptation. In *Algorithmic Recourse workshop at ICML'21*, 2021.

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

[5] Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

[6] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[7] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 91–98, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314236. URL `https://doi.org/10.1145/3306618.3314236`.

[8] Andrew Cotter, Maya Gupta, and Harikrishna Narasimhan. On making stochastic classifiers deterministic. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[9] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[10] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=bYi_2708mKK`.

[11] 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.

[12] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[13] Sarah Flood, Miriam King, Renae Rodgers Steven Ruggles, and J. Robert Warren. Integrated public use microdata series, current population survey: Version 8.0 [dataset], 2020. URL `https://www.ipums.org/projects/ipums-cps/d030.v8.0`.

[14] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. On fairness, diversity and randomness in algorithmic decision making, 2017.

[15] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, page 111–122, New York, NY, USA, 2016. Association for Computing Machinery.

[16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[17] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

[18] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.

[19] Yang Liu, Yatong Chen, and Jiaheng Wei. Induced domain adaptation, 2021.

[20] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms, 2009.

[21] Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368. ACM, 2019.

[22] Harikrishna Narasimhan. Learning with complex loss functions and constraints. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1646–1654. PMLR, 09–11 Apr 2018.

[23] Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34, 2021.

[24] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. In *AAAI*, 2021.

[25] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. Transfer of machine learning fairness across domains, 2019.

[26] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[27] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, 2021. Association for Computing Machinery.

[28] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[29] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[30] Min Wen, Osbert Bastani, and Ufuk Topcu. Fairness with Dynamics. *arXiv preprint arXiv:1901.08568*, 2019.

[31] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[32] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

[33] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellström, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In *NeurIPS*, 2020.

Checklist

1. For all authors...
    1. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
    2. Did you describe the limitations of your work? [Yes]
    3. Did you discuss any potential negative societal impacts of your work? [Yes]
    4. Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
    1. Did you state the full set of assumptions of all theoretical results? [Yes]
    2. Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
    1. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We submit the code in the supplementary material. We attach a README file for instructions to reproduce the results.
    2. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify the data splits in the experiment details.
    3. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
    4. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] All the experiments are running on CPU.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    1. If your work uses existing assets, did you cite the creators? [Yes]
    2. Did you mention the license of the assets? [N/A]
    3. Did you include any new assets either in the supplemental material or as a URL? [N/A] No new assets curated or released.
    4. Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] No new assets curated or released.
    5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
    1. Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    2. Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    3. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]