A Notation

Symbol	Usage
X	A random variable representing an example's *features*.
$\mathcal{X}$	The domain of features X.
Y	A random variable representing an example's *ground truth label*.
$\mathcal{Y}$	The domain of labels Y.
$\hat{Y}$	A random variable representing the *predicted label* for an example.
$\hat{\mathcal{Y}}$	The domain of predicted labels $\hat{Y}$ (distinguished semantically from $\mathcal{Y}$).
G	A random variable representing an example's *group* membership.
$\mathcal{G}$	The domain for group membership G.
π	A learned (non-deterministic) policy for predicting $\hat{Y}$ from X and G.
$\Pr$	A sample probability (density) according to a referenced distribution.
$\mathcal{P}$	The space of probability distributions over a given domain.
$\mathbb{D}$	The space of distributions of *examples* over $\mathcal{X} \times \mathcal{Y} \times \mathcal{G}$.
$\mathbb{O}$	The space of distributions of *outcomes* over $\mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{Y}}$.
$\mathbb{G}$	The space of distributions of *group-conditioned examples* $\mathcal{X} \times \mathcal{Y}$.
$\mathcal{S}$	The *source* distribution in $\mathbb{D}$.
$\mathcal{T}$	The *target* distribution in $\mathbb{D}$ to which π is now applied.
$\mathbf{D}$	A vectorized (by group) premetric for measuring shifts in $\mathbb{D}$.
$\mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{c}$	A vector of element-wise bounds for $\mathbf{D}$.
$\mathbf{e}_g$	A group-specific basis vector.
$\Delta^\star$	A disparity function, measuring "unfairness".
Ψ	A premetric function (see Definition 2.1) for measuring shifts in $\mathbb{O}$.
v	Supremal disparity within bounded distribution shift.
DP	Abbreviation for Demographic Parity.
EO	Abbreviation for Equalized Odds.
EOp	Abbreviation for Equal Opportunity.

Table 1: Primary Notation

B Extended Discussion of Related Work

Domain Adaptation: Prior work has considered the conditions under which a classifier trained on a source distribution will perform well on a given target distribution, for example, by deriving bounds on the number of training examples from the target distribution needed to bound prediction error [2, 27], or in conjunction with the dynamic response of a population to classification [25]. We are interested in a similar setting and concern, but address the transferability of *fairness guarantees*, rather than accuracy. In considering covariate shift and label shift as special cases in this paper, our work may be paired with studies that address the transferability of prediction accuracy under such assumptions [35, 37, 42].

Algorithmic Fairness: Many formulations of fairness have been proposed for the analysis of machine learning policies. When it is appropriate to ignore the specific social and dynamical context of a deployed policy, the statistical regularity of policy outcomes may be considered across individual examples [14] and across groups [41, 15, 8, 19, 6]. In our paper, we focus on such statistical definitions of fairness between groups, and develop bounds for demographic parity [6] and equalized odds [19] as specific examples.

Dynamic Modeling: When the dynamical context of a deployed policy must be accounted for, such as when the policy influences control over the future trajectories of a distribution of features and labels, we benefit from modelling how populations respond to classification. Among this line of work, [23] initiate the discussion of the long-term effect of imposing static fairness constraints on a dynamic social system, highlighting the importance of measurement and temporal modeling in the evaluation of fairness criteria. However, developing such models remains a challenging problem [11, 36, 31, 12, 43, 39, 24, 7, 20, 28, 30]. In particular, [11] discuss causal directed acyclic graphs (DAGs) as a unifying framework on fairness in dynamical systems. In this work, rather than

relying precise models of distribution shift to quantify the transferability of fairness guarantees in dynamical contexts, we assume a bound on the difference between source and target distributions. We thus develop bounds on realized statistical group disparity while remaining agnostic to the specific dynamics of the system.

C Additional Figures

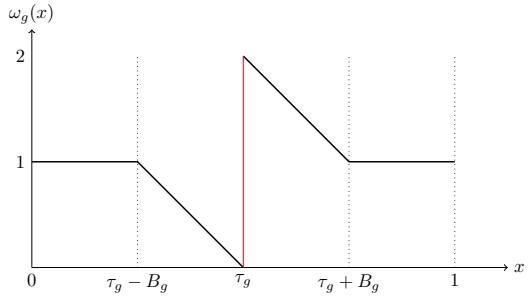


Figure 5: Distribution of the reweighting coefficient $w_g(x)$ for the setting of Covariate shift via Strategic Response.

D A Geometric Interpretation

In this extension of Section 4.2, we fulfill the promise of Section 3.3 and consider a case in which shared structure of between $\Psi\colon \mathbb{O}^2 \to \mathbf{R}$ and each $D_g\colon \mathbb{G}^2 \to \mathbf{R}$ permits a geometric interpretation of distribution shift for Equal Opportunity EOp, building on Theorem 4.3. We continue to defer rigorous proof to Appendix F.

We first recall the definition of the true positive rate of policy π, for each group, on distribution $\mathcal{T}$.

$$\beta_g^+ \coloneqq \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}1 \mid Y{=}1, G{=}g) \tag{29}$$

The true positive rate may be expressed as a ratio of inner products defined over the space of square-integrable L^2 functions on $\mathcal{X}$.[7]

$$\beta_g^+[\mathcal{T}] = \frac{\Pr_{\mathcal{T}}(\hat{Y}{=}1, Y{=}1 \mid G{=}g)}{\Pr_{\mathcal{T}}(Y{=}1 \mid G{=}g)} = \frac{\left\langle \mathsf{r}_g[\mathcal{T}], \mathsf{t}_g \right\rangle_g}{\left\langle \mathsf{r}_g[\mathcal{T}], 1 \right\rangle_g} \tag{30}$$

$$\langle a, b \rangle_g \coloneqq \int_{\mathcal{X}} a(x) b(x) \mathsf{s}_g(x)\, \mathrm{d}x \tag{31}$$

where we use the shorthands

$$\mathsf{r}_g[\mathcal{T}](x) \coloneqq \Pr_{\mathcal{T}}(X{=}x \mid G{=}g) \tag{32}$$

$$\mathsf{s}_g(x) \coloneqq \Pr_{\mathcal{S}}(Y{=}1 \mid X{=}x, G{=}g) \tag{33}$$

$$1(x) \coloneqq 1 \tag{34}$$

$$\mathsf{t}_g(x) \coloneqq \Pr_{\pi}(\hat{Y}{=}1 \mid Y{=}1, X = x, G = g) \tag{35}$$

and assume that $\mathsf{s}_g(x) > 0$ for all x and g.

We observe that the only degree of freedom in β_g^+ as $\mathcal{T}$ varies subject to covariate shift is r_g: by the covariate assumption, s_g is fixed; t meanwhile remains independent of $\mathcal{T}$ for fixed policy π, since π is independent of Y conditioned on X and G.

[7] This precludes distributions with non-zero probability mass concentrated at singular points.

Selection of $\mathbf{D}$ We now select each D_g to be the standard metric for the inner product defined by Equation (31), where, for each group, distributions in $\mathbb{G}$ are mapped to the corresponding vector r_g:

$$\begin{aligned} &D_g\big(\Pr_{\mathcal{T}}(X, Y \mid G{=}g) \parallel \Pr_{\mathcal{S}}(X, Y \mid G{=}g)\big) \\ &\qquad \coloneqq \sqrt{\langle \mathsf{r}_g[\mathcal{S}], \mathsf{r}_g[\mathcal{T}]\rangle_g + \langle \mathsf{r}_g[\mathcal{T}], \mathsf{r}_g[\mathcal{T}]\rangle_g - 2\langle \mathsf{r}_g[\mathcal{S}], \mathsf{r}_g[\mathcal{T}]\rangle_g} \end{aligned} \tag{36}$$

In this geometric picture, $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$ implies that all possible values for $\mathsf{r}_g[\mathcal{T}]$ lie within a ball of radius B_g centered at $\mathsf{r}_g[\mathcal{S}]$. By the normalization condition of a probablity (density) function, denoting $\mathsf{s}_g^{-1}(x) \coloneqq (\mathsf{s}_g(x))^{-1}$, the vector $\mathsf{r}_g[\mathcal{T}]$ must also lie on the hyperplane

$$\int_{\mathcal{X}} \mathsf{r}_g[\mathcal{T}]\,\mathrm{d}x = \langle \mathsf{r}_g[\mathcal{T}], \mathsf{s}_g^{-1}\rangle_g = 1 \tag{37}$$

Recalling Equation (30), the group-specific true positive rate $\beta_g^+[\mathcal{T}]$ for policy π is given by a ratio of the projected distances of r_g along the t_g and $\mathsf{1}$ vectors. Let us therefore denote the projection of $\mathsf{r}_g[\mathcal{T}]$ onto the $(\mathsf{1}, \mathsf{t}_g)$-plane as $\mathsf{r}_g^{\perp}[\mathcal{T}]$. We may then consider the possible values of $\mathsf{r}_g^{\perp}[\mathcal{T}]$ as projections from the intersection of the $\mathsf{r}_g[\mathcal{S}]$-centered hypersphere of radius B_g and the hyperplane of normalized distributions (Equation (37)). Using $\angle(\cdot,\cdot)$ to denote the angle between vectors and denoting $\phi_g' \coloneqq \angle(\mathsf{r}_g, \mathsf{t}_g)$, $\theta_g' \coloneqq \angle(\mathsf{r}_g, \mathsf{1})$, $\phi_g \coloneqq \angle(\mathsf{r}_g^{\perp}, \mathsf{t}_g)$, and $\theta_g \coloneqq \angle(\mathsf{r}_g^{\perp}, \mathsf{1})$, we appeal to the geometric relationship $\langle a, b\rangle = \cos(\angle(a,b))\|a\|\|b\|$ to write

$$\beta_g^+ \frac{\|\mathsf{1}\|}{\|\mathsf{t}_g\|} = \frac{\cos\phi_g'}{\cos\theta_g'} = \frac{\cos\phi_g}{\cos\theta_g} \tag{38}$$

From these observations, we need only bound the ratio between $\cos(\phi_g)$ and $\cos(\theta_g)$ to bound β_g^+. Relating these angles in the $(\mathsf{1}, \mathsf{t}_g)$-plane by $\phi_g = \xi_g - \theta_g$ where $\xi_g \coloneqq \angle(\mathsf{t}_g, \mathsf{1})$, we arrive at the following theorem:

Theorem D.1. *The true positive rate β_g^+ is bounded over the domain of covariate shift $\mathbb{D}_{cov}[\mathbf{B}]$, which we define by the bound $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$, and the invariance of $\Pr(Y{=}1 \mid X{=}x, G{=}g)$ for all x, g, as*

$$\frac{\cos\big(\phi_g^u\big)}{\cos\big(\xi_g - \phi_g^u\big)} \leq \frac{\|\mathsf{1}\|}{\|\mathsf{t}_g\|}\beta_g^+(\pi, \mathcal{T}) \leq \frac{\cos\big(\phi_g^l\big)}{\cos\big(\xi_g - \phi_g^l\big)} \tag{39}$$

with upper (ϕ_g^u) and lower (ϕ_g^l) bounds for ϕ_g represented as

$$\phi_g^l \coloneqq \min_{\mathcal{T} \in \mathbb{D}_{cov}[\mathbf{B}]} \phi_g; \quad \phi_g^u \coloneqq \max_{\mathcal{T} \in \mathbb{D}_{cov}[\mathbf{B}]} \phi_g \tag{40}$$

We obtain a final bound on $\Delta^\star_{\mathsf{EOp}}$ by substituting Equation (39) into Equation (17). We visualize the geometric bound on β_g^+ (Theorem D.1) in Figure 6. In Appendix E.1, we apply this bound to real-world credit score data assuming the model of strategic manipulation given in Section 6.1. Although the result is not an easily interpretted formula, it provides a demonstration of geometric reasoning applied to statistical fairness guarantees.

Finally, we note that, in addition to the constraints considered above, each vector r_g is subject to the positivity condition, $\forall x \in \mathcal{X}, \mathsf{r}_g(x) \geq 0$. The bound developed in this section, however, does not benefit from this additional constraint; we leave this to potential future work.

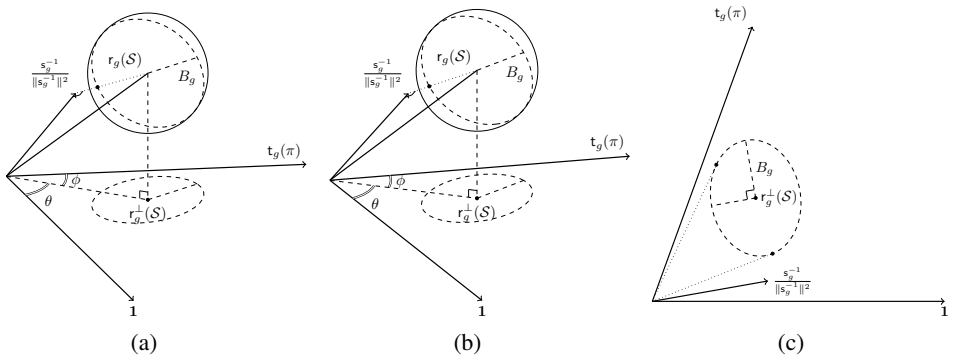


Figure 6: A geometric bound in an infinite-dimensional vector space (*i.e.*, a Hilbert space), represented with a stereoscopic (cross-eye) view in three dimensions (to provide intuition) and an examination of the $(\mathsf{t}_g, \mathbf{1})$-plane. The extreme values of β_g^+ correspond to the extremal angles of ϕ and θ. In this figure, the vector displayed parallel to s_g^{-1} from the origin terminates on the hyperplane of normalized distributions.

E Empirical Evaluations of the Bounds

E.1 Comparisons to Dynamic Models of Distribution Shift

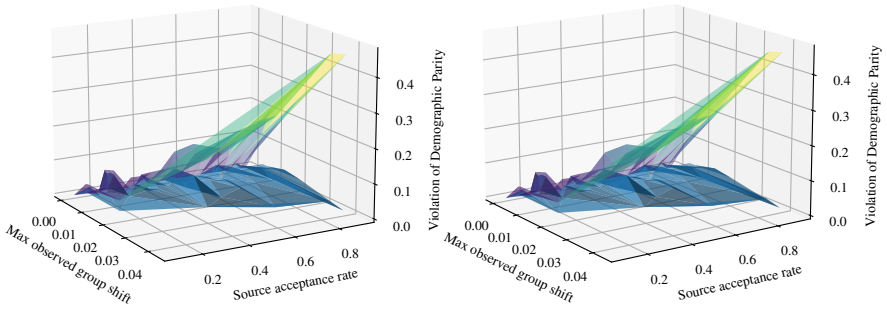


Figure 7: A stereoscopic (cross-eye view) comparison between the bound of Section 4.1 (gradated) and simulated results for the model of Section 6.1 (blue) in response to a DP-fair classifier with different initial group-independent acceptance rates. The x-axis represents the maximum shift D_g over all groups g in response to the classifier.

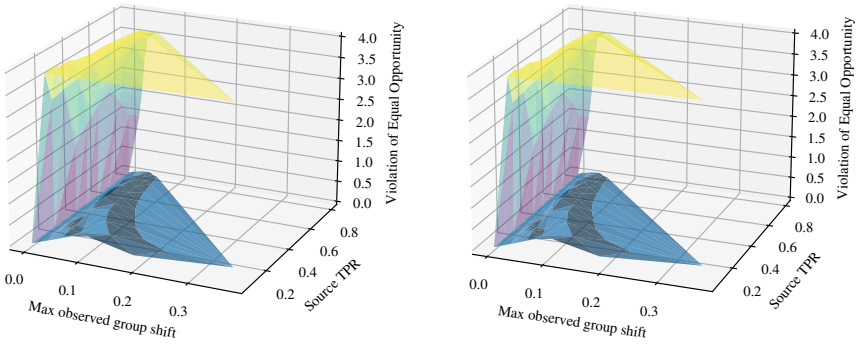


Figure 8: A stereoscopic (cross-eye view) comparison between the theoretical bound of Section 4.2 (gradated) and simulated results for the model of Section 6.1 (blue) in response to a EOp-fair classifier with different initial group-independent true positive rates (TPR). The x-axis represents the maximum shift D_g over all groups g in response to the classifier. As Corollary 4.4 limits the maximum possible value of EO violation, we include this limit as part of the bound.

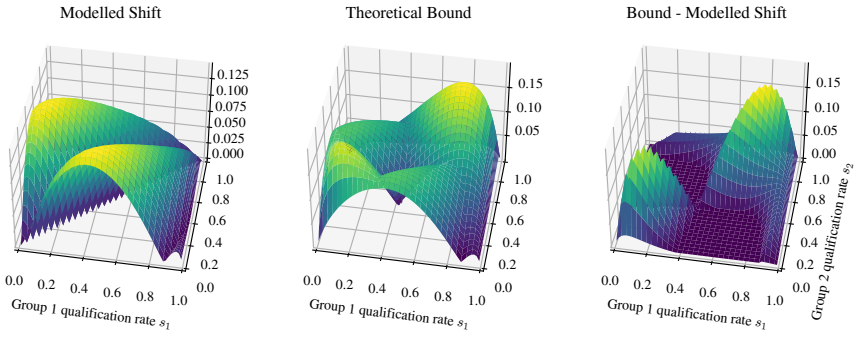
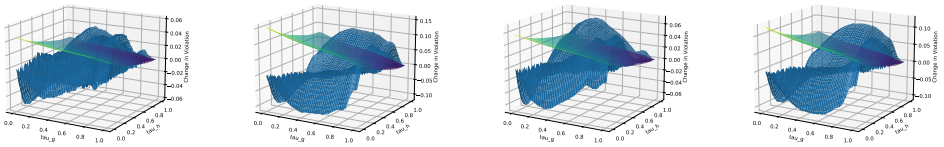


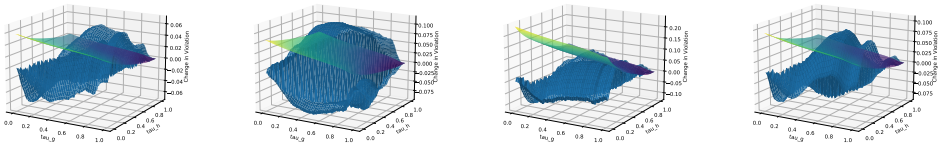
Figure 9: A policy satisfying DP is subject to distribution shift prescribed by replicator dynamics (Section 6.2). Realized disparity increases (blue) are compared to the theoretical bound (Theorem 5.2, gradated), which is tight when group have dissimilar qualification rates.

E.2 Comparisons to Real-World Data

We provide additional graphics comparing bounds on demographic parity or equal opportunity to real-world distribution shifts. Figure 10 compares the covariate shift bound of Theorem 4.1 to the violation of demographic parity for hypothetical policies trained on one US state and deployed in another. Figure 11 compares the label shift bound of Theorem 5.2 to the violation of demographic parity for hypothetical policies trained for a US state in 2014 and deployed in 2018. Figure 12 compares the covariate shift bound of Theorem 4.3 with Theorem D.1 to the violation of equal opportunity for hypothetical policies trained on one US state and deployed in another.

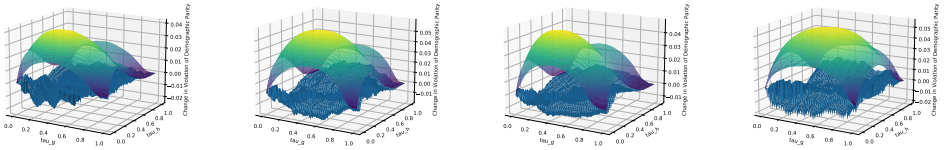


(a) CA $\longrightarrow$ AZ (b) CA $\longrightarrow$ CO (c) CA $\longrightarrow$ FL (d) CA $\longrightarrow$ MD

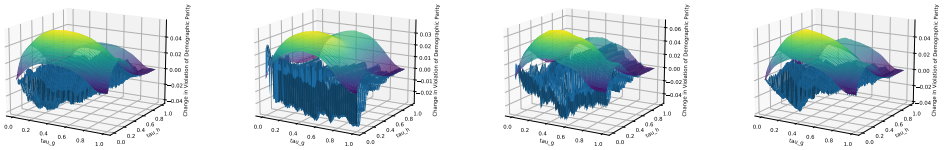


(e) CA $\longrightarrow$ OR (f) CA $\longrightarrow$ UT (g) CA $\longrightarrow$ VT (h) CA $\longrightarrow$ WI

Figure 10: Change in violation of demographic parity for hypothetical policies trained on one US state's data and reused for another state (blue) compared to covariate-shift bounds (Theorem 4.1, gradated). The x-axis and y-axis represent the thresholds τ_g and τ_h, respectively.

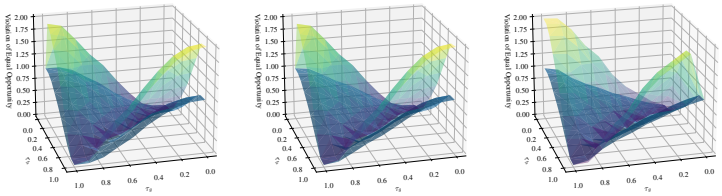


(a) Arizona (AZ) (b) Colorado (CO) (c) Florida (FL) (d) Maryland (MD)



(e) Oregon (OR) (f) Utah (UT) (g) Vermont (VT) (h) Wisconsin (WI)

Figure 11: Change in violation of demographic parity for hypothetical policies trained on 2014 data and reused for 2018 (blue) compared to label-shift bounds (Theorem 5.2, gradated). The x-axis and y-axis represent the thresholds τ_g and τ_h, respectively.



(a) CA $\longrightarrow$ NV (b) CA $\longrightarrow$ IL (c) NV $\longrightarrow$ IL

Figure 12: Change in violation of equal opportunity for hypothetical policies trained on one US state's data and reused for another state (blue) compared to covariate-shift bounds (Theorems 4.3 and D.1, gradated). The x-axis and y-axis represent the thresholds τ_g and τ_h, respectively.

F Omitted Proofs

Proof of Lemma 2.7:

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

Proof. By the definitions of group-vectorized shift (Definition 2.5) and divergence (Definition 2.4) together with the bounded distribution shift assumption (Assumption 2.6), we note

$$\mathbf{B} = 0 \implies \mathbf{D}(\mathcal{T} \parallel \mathcal{S}) = 0 \tag{41}$$

and

$$D_g(\mathcal{T} \parallel \mathcal{S}) = 0 \implies \Pr_{\mathcal{S}}(X, Y \mid G{=}g) = \Pr_{\mathcal{T}}(X, Y \mid G{=}g) \tag{42}$$

Combining these implications and invoking the independence of $\hat{Y} \sim \pi$ and Y conditioned on X and G (Equation (1)), it follows that

$$\mathbf{B} = 0 \implies \forall g, \quad \Pr_{\pi,\mathcal{S}}(X, Y, \hat{Y} \mid G{=}g) = \Pr_{\pi,\mathcal{T}}(X, Y, \hat{Y} \mid G{=}g) \tag{43}$$

Consulting the definition of disparity (Definition 2.2), it follows that $\Delta^\star(\pi, \mathcal{S})$ and $\Delta^\star(\pi, \mathcal{T})$ are equal when $\mathbf{B} = 0$.

Proof of Theorem 3.2:

Statement: 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 0 to $\mathbf{B}$, then

$$\Delta^\star(\pi, \mathcal{T}) \leq \Delta^\star(\pi, \mathcal{S}) + \mathbf{L} \cdot \mathbf{B} \tag{44}$$

Proof. We reiterate that $v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{b})$ defines a scalar field over the non-negative cone $\mathbf{b} \in (\mathbf{R}_+ \cup 0)^{|\mathcal{G}|}$. Treating v as a scalar potential, we may define the conservative vector field $\mathbf{F}$:

$$\mathbf{F} = \nabla_{\mathbf{b}} v \tag{45}$$

This formulation, in terms of a potential, ensures the path-independence of the line integral of $\mathbf{F}$ along any continuous curve C from 0 to $\mathbf{B}$. That is,

$$v(..., \mathbf{B}) - v(..., 0) = \int_C \mathbf{F}(\mathbf{b}) \cdot \mathrm{d}\mathbf{b} \tag{46}$$

Therefore, given a Lipshitz condition for $\mathbf{F}$ along any curve C with endpoints 0 and $\mathbf{B}$, *i.e.* when there exists some finite $\mathbf{L}$ such that

$$\forall \mathbf{b} \in C, \quad \mathbf{F}(\mathbf{b}) \preceq \mathbf{L} \tag{47}$$

and therefore

$$v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) = v(..., 0) + \int_C \mathbf{F}(\mathbf{b}) \cdot \mathrm{d}\mathbf{b} \tag{48}$$

$$\leq \Delta^\star(\pi, \mathcal{S}) + \mathbf{L} \cdot \mathbf{B} \tag{49}$$

By the bounded distribution shift assumption (Assumption 2.6), Lemma 2.7, and the definition of the supremum bound (Definition 3.1), we conclude

$$\Delta^\star(\pi, \mathcal{T}) \leq \Delta^\star(\pi, \mathcal{S}) + \mathbf{L} \cdot \mathbf{B} \tag{50}$$

Proof of Theorem 3.4:

Statement: 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 $\mathbf{a} + \mathbf{c} \preceq \mathbf{B}$. Then, a local, first-order approximation of $w(..., \mathbf{b})$ evaluated at 0, *i.e.*,

$$\mathbf{L} = \nabla_{\mathbf{b}} w(..., \mathbf{b})\big|_{\mathbf{b}=0} = \nabla_{\mathbf{b}} v(..., \mathbf{b})\big|_{\mathbf{b}=0} \tag{51}$$

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

$$v(\Delta^\star, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) \leq \Delta^\star(\pi, \mathcal{S}) + \mathbf{L} \cdot \mathbf{B} \tag{52}$$

Proof. Represent

$$\mathbf{B} = \sum_g \mathbf{e}_g B_g \tag{53}$$

Then, invoking the definition of the derivative as a Weierstrass limit from elementary calculus, as well as Lemma 2.7, and by repeatedly appealing to the assumed subadditivity condition within our domain, we find

$$\begin{aligned}
\mathbf{B} \cdot \mathbf{L} &= \mathbf{B} \cdot \nabla_{\mathbf{b}} v(\pi, \mathcal{S}, \mathbf{b})\big|_{\mathbf{b}=0} && \text{(54a)}\\
&= \sum_g B_g \frac{d}{dx} v(\pi, \mathcal{S}, x\mathbf{e}_g)\big|_{x=0} && \text{(54b)}\\
&= \sum_g B_g \lim_{N\to\infty} N\Big(v(\pi, \mathcal{S}, \frac{1}{N}\mathbf{e}_g) - v(\pi, \mathcal{S}, 0)\Big) && \text{(54c)}\\
&= \sum_g B_g \lim_{N\to\infty} N\Big(w(\pi, \mathcal{S}, \frac{1}{N}\mathbf{e}_g)\Big) && \text{(54d)}\\
&\geq \sum_g B_g \, w(\pi, \mathcal{S}, \mathbf{e}_g) && \text{(54e)}\\
&\geq \sum_g w(\pi, \mathcal{S}, B_g\mathbf{e}_g) && \text{(54f)}\\
&\geq w(\pi, \mathcal{S}, \mathbf{B}) && \text{(54g)}
\end{aligned}$$

Also recall (Definition 3.3)

$$w(\pi, \mathcal{S}, \mathbf{B}) \coloneqq v(\pi, \mathcal{S}, \mathbf{B}) - \Delta^\star(\pi, \mathcal{S}) \tag{55}$$

Therefore, we obtain

$$v(\pi, \mathcal{S}, \mathbf{B}) \leq \Delta^\star(\pi, \mathcal{S}) + \mathbf{B} \cdot \mathbf{L} \tag{56}$$

Lemma F.1. *For each group $g \in \mathcal{G}$, under covariate shift,*

$$\Pr_{\pi,\mathcal{T}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) - \Pr_{\pi,\mathcal{S}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) = \operatorname*{Cov}_{\pi,\mathcal{S}}\Big[\omega_g(\mathcal{T}, \mathcal{S}, X), \Pr_{\pi(X,g)}(\hat{Y}{=}1)\Big] \tag{57}$$

Proof. First, note that $\mathbb{E}_{\mathcal{S}}[\omega_g(\mathcal{T}, \mathcal{S}, x)] = 1$, since

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{S}}[\omega_g(\mathcal{T}, \mathcal{S}, x)] &= \int_{\mathcal{X}} \omega_g(\mathcal{T}, \mathcal{S}, x) \Pr_{\mathcal{S}}(X = x \mid G{=}g)\,\mathrm{d}x\\
&= \int_{\mathcal{X}} \frac{\Pr_{\mathcal{T}}(X{=}x \mid G{=}g)}{\Pr_{\mathcal{S}}(X{=}x \mid G = g)} \Pr_{\mathcal{S}}(X{=}x \mid G{=}g)\,\mathrm{d}x\\
&= \int_{\mathcal{X}} \Pr_{\mathcal{T}}(X{=}x \mid G{=}g)\,\mathrm{d}x = 1
\end{aligned}$$

Then, adopting the shorthand $\omega_g(x) = \omega_g(\mathcal{T}, \mathcal{S}, x)$, we have:

$$\Pr_{\pi,\mathcal{T}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) - \Pr_{\pi,\mathcal{S}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) \tag{58}$$

$$= \int_{\mathcal{X}} \Pr_{\pi(x,g)}(\hat{Y}{=}1) \Pr_{\mathcal{T}}(X{=}x \mid G{=}g)\,\mathrm{d}x - \int_{\mathcal{X}} \Pr_{\pi(x,g)}(\hat{Y}{=}1) \Pr_{\mathcal{S}}(X{=}x \mid G{=}g)\,\mathrm{d}x \tag{59}$$

$$= \int_{\mathcal{X}} \Pr_{\pi(x,g)}(\hat{Y}{=}1)\Big(\omega_g(x) - 1\Big) \Pr_{\mathcal{S}}(X{=}x \mid G{=}g)\,\mathrm{d}x \tag{60}$$

$$= \mathop{\mathbb{E}}_{\mathcal{S}}\Big[\Pr_{\pi(x,g)}(\hat{Y}{=}1)(\omega_g(x) - 1) \mid G{=}g\Big] \tag{61}$$

$$= \mathop{\mathbb{E}}_{\mathcal{S}}\Big[\Pr_{\pi(x,g)}(\hat{Y}{=}1)(\omega_g(x) - \mathop{\mathbb{E}}_{\mathcal{S}}[\omega_g(x)]) \mid G{=}g\Big] \qquad (\text{since } \mathbb{E}_{\mathcal{S}}[\omega_g(x)] = 1)$$

$$= \mathop{\mathbb{E}}_{\mathcal{S}}\Big[(\Pr_{\pi(x,g)}(\hat{Y}{=}1) - \mathop{\mathbb{E}}_{\mathcal{S}}[\Pr_{\pi(x,g)}(\hat{Y}{=}1)](\omega_g(x) - \mathop{\mathbb{E}}_{\mathcal{S}}[\omega_g(x)]) \mid G{=}g\Big] \quad (\mathbb{E}[f(x) - \mathbb{E}[f(x)]] = 0)$$

$$= \mathop{\mathrm{Cov}}_{\pi,\mathcal{S}}\Big[w_g(\mathcal{T}, \mathcal{S}, X), \Pr_{\pi(x,g)}(\hat{Y}{=}1)\Big] \tag{62}$$

Lemma F.2. *If X is a random variable and $X \in [0,1]$, then* $\mathrm{Var}(X) \leq \mathbb{E}[X](1 - \mathbb{E}(X))$.

Proof.

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&\leq \mathbb{E}[X] - (\mathbb{E}[X])^2 && (X \in [0,1]) \\
&= \mathbb{E}[X](1 - \mathbb{E}(X))
\end{aligned}$$

Proof of Theorem 4.1:

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

$$\Delta^\star{}_{\mathsf{DP}}(\pi, \mathcal{T}) \leq \Delta^\star{}_{\mathsf{DP}}(\pi, \mathcal{S}) + \sum_g \Big(\beta_g(1 - \beta_g)B_g\Big)^{1/2} \tag{63}$$

Proof. Again adopting the shorthand $\omega_g(x) = \omega_g(\mathcal{T}, \mathcal{S}, x)$,

$$\Delta^\star(\pi, \mathcal{T}) \tag{64}$$

$$= \Big| \Pr_{\pi,\mathcal{T}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) - \Pr_{\pi,\mathcal{T}}\Big(\hat{Y}{=}1 \mid G{=}h\Big)\Big| \tag{65}$$

$$= \Big| \Pr_{\pi,\mathcal{T}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) - \Pr_{\pi,\mathcal{S}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) \tag{66}$$

$$+ \Pr_{\pi,\mathcal{S}}\Big(\hat{Y}{=}1 \mid G{=}g\Big) - \Pr_{\pi,\mathcal{S}}\Big(\hat{Y}{=}1 \mid G{=}h\Big) \tag{67}$$

$$+ \Pr_{\pi,\mathcal{S}}\Big(\hat{Y}{=}1 \mid G{=}h\Big) - \Pr_{\pi,\mathcal{T}}\Big(\hat{Y}{=}1 \mid G{=}h\Big)\Big| \tag{68}$$

$$\leq \Delta^\star(\pi, \mathcal{S}) + \operatorname*{Cov}_{\mathcal{S}}\Big[\omega_g(x), \Pr_{\pi(x,g)}(\hat{Y}{=}1)\Big] + \operatorname*{Cov}_{\mathcal{S}}\Big[\omega_h(x), \Pr_{\pi(x,h)}(\hat{Y}{=}1)\Big] \qquad \text{(By Lemma F.1)}$$

$$\leq \Delta^\star(\pi, \mathcal{S}) + \sqrt{\operatorname*{Var}_{\mathcal{S}}[\omega_g(x)]} \cdot \sqrt{\operatorname*{Var}_{\mathcal{S}}[\Pr_{\pi(x,g)}(\hat{Y}{=}1)]} + \sqrt{\operatorname*{Var}_{\mathcal{S}}[\omega_h]} \cdot \sqrt{\operatorname*{Var}_{\mathcal{S}}[\Pr_{\pi(x,h)}(\hat{Y}{=}1)]}$$

$$(\big|\operatorname{Cov}[a,b]\big| \leq \sqrt{\operatorname{Var}[a]} \cdot \sqrt{\operatorname{Var}[b]})$$

$$\leq \Delta^\star(\pi, \mathcal{S}) + \sqrt{\operatorname*{Var}_{\mathcal{S}}[\omega_g(x)]} \cdot \sqrt{\mathbb{E}_{\mathcal{S}}[\Pr_{\pi(x,g)}(\hat{Y}{=}1)](1 - \mathbb{E}_{\mathcal{S}}[\Pr_{\pi(x,g)}(\hat{Y}{=}1)])} \tag{69}$$

$$+ \sqrt{\operatorname*{Var}_{\mathcal{S}}(\omega_h(x))} \cdot \sqrt{\mathbb{E}_{\mathcal{S}}[\Pr_{\pi(x,h)}(\hat{Y}{=}1)](1 - \mathbb{E}_{\mathcal{S}}[\Pr_{\pi(x,h)}(\hat{Y}{=}1))]}$$

$$(\hat{Y} \in \{0,1\}, \text{ and Lemma F.2})$$

$$= \Delta^\star(\pi, \mathcal{S}) + \sqrt{\operatorname*{Var}_{\mathcal{S}}[\omega_g(x)]} \cdot \sqrt{\beta_g(1-\beta_g)} + \sqrt{\operatorname*{Var}_{\mathcal{S}}[\omega_h(x)]} \cdot \sqrt{\beta_h(1-\beta_h)}$$

$$(\beta_g = \Pr_{\pi,\mathcal{S}}(\hat{Y}{=}1|G{=}g) = \mathbb{E}_{\mathcal{S}}[\mathbb{1}_{\pi(x,g)}(\hat{Y}{=}1)])$$

$$= \Delta^\star(\pi, \mathcal{S}) + \sum_g \Big(\beta_g(1-\beta_g) \operatorname*{Var}_{\mathcal{S}}[\omega_g(\mathcal{T}, \mathcal{S}, x)]\Big)^{1/2} \tag{70}$$

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

$$\Delta^\star_{\mathsf{DP}}(\pi, \mathcal{T}) \coloneqq \sum_{y \in \mathcal{Y}} \sum_{g,h \in \mathcal{G}} \Big| \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}y \mid G{=}g) - \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}y \mid G{=}h)\Big| \tag{71}$$

$$\Delta^\star{}_{\mathsf{DP}}(\pi, \mathcal{T}) \leq \Delta^\star{}_{\mathsf{DP}}(\pi, \mathcal{S}) + \sum_{y \in \mathcal{Y}} \sum_{g,h \in \mathcal{G}} \big(\beta_{g,y}(1-\beta_{g,y})B_g\big)^{1/2} \tag{72}$$

where $\beta_{g,y} = \Pr\Big(\hat{Y}{=}y \mid G{=}g\Big)$, and assuming $\operatorname{Var}_{\mathcal{S}}[w_g(\mathcal{S}, \mathcal{T}, X)] \leq B_g$.

Proof. We again adopt the shorthand $\omega_g(x) = \omega_g(\mathcal{T}, \mathcal{S}, x)$. We first generalize Lemma F.1. For each group $g \in \mathcal{G}$, under covariate shift, for all $y \in \mathcal{Y}$,

$$\Pr_{\pi,\mathcal{T}}\Big(\hat{Y}{=}y \mid G{=}g\Big) - \Pr_{\pi,\mathcal{S}}\Big(\hat{Y}{=}y \mid G{=}g\Big) = \operatorname*{Cov}_{\pi,\mathcal{S}}\Big[\omega_g(\mathcal{S}, \mathcal{T}, X), \Pr_{\pi(X,g)}(\hat{Y}{=}y)\Big] \tag{73}$$

Retracing the logic of Theorem 4.1, for $\operatorname{Var}_{\mathcal{S}}[\omega_g(\mathcal{T}, \mathcal{S}, x)] \leq B_g$, it follows that

$$\Delta^\star_{\mathsf{DP}}(\pi, \mathcal{T}) \coloneqq \sum_{y \in \mathcal{Y}} \sum_{g,h \in \mathcal{G}} \Big| \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}y \mid G{=}g) - \Pr_{\pi,\mathcal{T}}(\hat{Y}{=}y \mid G{=}h)\Big| \tag{74}$$

$$\leq \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{S}) + \sum_{y \in \mathcal{Y}} \sum_{g,h \in \mathcal{G}} \sqrt{\big(\beta_{g,y}(1-\beta_{g,y}) \operatorname*{Var}_{\mathcal{S}}[\omega_g(x)]\big)} \tag{75}$$

$$\leq \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{S}) + \sum_{y \in \mathcal{Y}} \sum_{g,h \in \mathcal{G}} \sqrt{\big(\beta_{g,y}(1-\beta_{g,y})B_g\big)} \tag{76}$$

$$= \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{S}) + \sum_{y \in \mathcal{Y}} \sum_{g,h \in \mathcal{G}} \big(\beta_{g,y}(1-\beta_{g,y})B_g\big)^{1/2} \tag{77}$$

Proof of Theorem 4.3

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

$$\forall g,\ \ D_g(\mathcal{T} \parallel \mathcal{S}) < B_g \implies l_g \leq \beta_g^+(\pi, \mathcal{T}) < u_g \tag{78}$$

then, for v corresponding to $\Delta^\star{}_{\mathsf{EOp}}$,

$$v(\Delta^\star{}_{\mathsf{EOp}}, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) \leq \max_{x_g \in \{l_g, u_g\}} \sum_{g,h} \Big| x_g - x_h \Big| \tag{79}$$

Proof. Recall that, for this setting,

$$v(\Delta^\star{}_{\mathsf{EOp}}, \mathbf{D}, \pi, \mathcal{S}, \mathbf{B}) = \sup_{\mathbf{D}(\mathcal{T} \| \mathcal{S}) \preceq \mathbf{B}} \Delta^\star{}_{\mathsf{EOp}} \tag{80}$$

and

$$\Delta^\star{}_{\mathsf{EOp}} = \sum_{g,h} |\beta_g^+ - \beta_h^+| \tag{81}$$

This latter expression is convex in each β_g^+. Therefore, $\Delta^\star{}_{\mathsf{EOp}}$ is maximized on the boundary of its domain, *i.e.* $\beta_g^+ \in \{l_g, u_g\}$ for each g, given the assumption of the theorem.

Proof of Corollary 4.4

Statement: The disparity measurement $\Delta^\star{}_{\mathsf{EOp}}$ cannot exceed $\frac{|\mathcal{G}|^2}{4}$.

Proof. We note that each β_g^+ is ultimately confined to the interval $[0, 1]$. Building on our proof for Theorem 4.3, to maximize $\Delta^\star{}_{\mathsf{EOp}}$, we must consider the boundary of this domain, where, for each g, $\beta_g^+ \in \{0, 1\}$. Because the only terms that contribute to $\Delta^\star{}_{\mathsf{EOp}}$ are those in which $\beta_g^+ = 1$ and $\beta_h^+ = 0$ (as opposed to $\beta_g^+ = \beta_h^+$), we seek to maximize the number of such terms. This occurs when as close to half of the groups as possible have one extremal true positive rate (*e.g.*, without loss of generality, $\beta_g^+ = 1$) and the remaining groups have the other (*e.g.*, $\beta_g^+ = 0$). In such cases, $\Delta^\star{}_{\mathsf{EOp}}$ is given by

$$\max \Delta^\star{}_{\mathsf{EOp}} = \lfloor \frac{\mathcal{G}}{2} \rfloor \lceil \frac{\mathcal{G}}{2} \rceil \leq \frac{|\mathcal{G}|^2}{4} \tag{82}$$

Proof of Theorem 5.1:

Statement: A Lipshitz condition bounds $\nabla_{\mathbf{b}} v(\Delta^\star{}_{\mathsf{DP}}, \mathbf{D}, \pi, \mathcal{S}, \mathbf{b})$ when

$$D_g(\mathcal{T} \parallel \mathcal{S}) \coloneqq \Big| Q_g(\mathcal{S}) - Q_g(\mathcal{T}) \Big| \leq B_g \tag{83}$$

Specifically,

$$\frac{\partial}{\partial b_g} v(\Delta^\star{}_{\mathsf{DP}}, \mathbf{D}, \pi, \mathcal{S}, \mathbf{b}) \leq (|\mathcal{G}| - 1) \Big| \beta_g^+ - \beta_g^- \Big| \tag{84}$$

for true positive rates β_g^+ and false positive rates β_g^-:

$$\beta_g^+ \coloneqq \Pr_{\pi}(\hat{Y}{=}1 | Y{=}1, G{=}g) \quad ; \quad \beta_g^- \coloneqq \Pr_{\pi}(\hat{Y}{=}1 | Y{=}0, G{=}g) \tag{85}$$

Proof. We first establish that $D_g^{(\mathsf{DP})}(\mathcal{T} \parallel \mathcal{S}) = |Q_g(\mathcal{S}) - Q_g(\mathcal{T})|$, where

$$Q_g(\mathcal{T}) \coloneqq \Pr_{\mathcal{T}}(Y{=}1 \mid G{=}g) \tag{86}$$

is an appropriate measure of group-conditioned distribution shift (Definition 2.5). That $\mathbf{D}$ satisfies the axioms of a divergence on group-conditioned distributions subject to the label shift assumption ($\Pr_{\mathcal{T}}(X \mid Y, G) = \Pr_{\mathcal{S}}(X \mid Y, G)$) and unchanging group sizes is easily verified:

$$\forall \mathcal{S}, \mathcal{T}, \quad D_g(\mathcal{T} \parallel \mathcal{S}) = |Q_g(\mathcal{S}) - Q_g(\mathcal{T})| \geq 0 \tag{87}$$

$$D_g(\mathcal{T} \parallel \mathcal{S}) = |Q_g(\mathcal{T}) - Q_g(\mathcal{T})| = 0 \tag{88}$$

and

$$\forall g, \quad D_g(\mathcal{T} \parallel \mathcal{S}) = 0 \implies \Pr_{\mathcal{T}}(Y \mid G) = \Pr_{\mathcal{S}}(Y \mid G) \tag{89}$$

$$\implies \Pr_{\mathcal{T}}(Y, X \mid G) = \Pr_{\mathcal{S}}(Y, X \mid G) \tag{90}$$

We next show that $(|\mathcal{G}| - 1)|\beta_g^+ - \beta_g^-|$ is the corresponding Lipshitz bound for the slope of v with respect to B_g, where we recall

$$\forall g, \quad \beta_g^+ \coloneqq \Pr_{\pi, \mathcal{T}}\Big(\hat{Y}{=}1 \mid Y{=}1, G{=}g\Big) \tag{91}$$

$$\forall g, \quad \beta_g^- \coloneqq \Pr_{\pi, \mathcal{T}}\Big(\hat{Y}{=}1 \mid Y{=}{-}1, G{=}g\Big) \tag{92}$$

That is, we wish to show

$$\frac{\partial}{\partial b_g} v(\Delta^\star_{\mathsf{DP}}, \mathbf{D}, \pi, \mathcal{S}, \mathbf{b}) \leq (|\mathcal{G}| - 1)|\beta_g^+ - \beta_g^-| \tag{93}$$

This follows directly from recognition that $\Delta^\star_{\mathsf{DP}}$ is locally always affine in the acceptance rate for each group, with slope bounded by one less than the number of groups.

$$\Delta^\star_{\mathsf{DP}} = \sum_{g,h \in \mathcal{G}} |\beta_g - \beta_h| \implies \frac{\partial}{\partial \beta_g} \Delta^\star_{\mathsf{DP}} \leq |\mathcal{G}| - 1 \tag{94}$$

By the definition of conditional probability,

$$\beta_g \coloneqq \Pr\Big(\hat{Y}{=}1\Big) = \beta_g^+ Q_g + \beta_g^-(1 - Q_g) \tag{95}$$

$$\frac{\partial}{\partial Q_g} \beta_g = \beta_g^+ - \beta_g^- \tag{96}$$

It follows by the chain rule that, for all $\mathcal{T}$ mutated from $\mathcal{S}$ subject to label shift,

$$\frac{\partial}{\partial Q_g(\mathcal{T})} \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{T}) \leq (|\mathcal{G}| - 1)|\beta_g^+ - \beta_g^-| \tag{97}$$

By the linearity of derivatives, for fixed $\mathcal{S}$, this implies that for all $\mathcal{T}$ attainable via label shift,

$$\frac{\partial}{\partial |Q_g(\mathcal{T}) - Q_g(\mathcal{S})|} \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{T}) \leq (|\mathcal{G}| - 1)|\beta_g^+ - \beta_g^-| \tag{98}$$

Since this equation holds for all $\mathcal{T}$, it must also hold when evaluated at v, the supremum of $\Delta^\star$. It follows that

$$\frac{\partial}{\partial B_g} v(\Delta^\star_{\mathsf{DP}}, \mathbf{D}^{(\mathsf{DP})}, \pi, \mathcal{S}, \mathbf{B}) \leq (|\mathcal{G}| - 1)|\beta_g^+ - \beta_g^-| \tag{99}$$

Proof of Theorem 5.2:

Statement: For DP under the bounded label-shift assumption $\forall g, |Q_g(\mathcal{S}) - Q_g(\mathcal{T})| \leq B_g$,

$$\Delta^\star_{\mathsf{DP}}(\pi, \mathcal{T}) \leq \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{S}) + (|\mathcal{G}| - 1) \sum_g B_g \left|\beta_g^+ - \beta_g^-\right| \tag{100}$$

Proof. This follows from the Lipshitz property implied by Theorem 5.1 (Equation (99)) and Theorem 3.2.

F.1 Omitted details for Section 6.1

Lemma F.3. *Recall the covariate shift reweighting coefficient $\omega_g(x)$, defined in Section 4.1.*

$$\omega_g(x) \coloneqq \frac{\Pr_{\mathcal{T}}(X{=}x \mid G{=}g)}{\Pr_{\mathcal{S}}(X{=}x \mid G{=}g)} \tag{101}$$

For our assumed setting,

$$\omega_g(x) = \begin{cases} 1, & x \in [0, \tau_g - m_g) \\ \frac{\tau_g - x}{m_g}, & x \in [\tau_g - m_g, \tau_g) \\ \frac{1}{m_g}(-x + \tau_g + 2m_g), & x \in [\tau_g, \tau_g + m_g) \\ 1, & x \in [\tau_g + m_g, 1] \end{cases} \tag{102}$$

Proof for Lemma F.3:

Proof. We discuss the target distribution by cases:

- For the target distribution between $[0, \tau_g - M_g]$: since we assume the agents are rational, under assumption 6.2, agents with feature that is smaller than $[0, \tau_g - M_g]$ will not perform any kinds of adaptations, and no other agents will adapt their features to this range of features either, so the distribution between $[0, \tau_g - M_g]$ will remain the same as before.
- For target distribution between $[\tau_g - M_g, \tau_g]$, it can be directly calculated from assumption 6.3.
- For distribution between $[\tau_g, \tau_g + M_g]$, consider a particular feature $x^\star \in [\tau_g, \tau_g + M_g]$, under Assumption 6.4, we know its new distribution becomes:

$$\begin{aligned} \Pr_{\mathcal{T}}(x = x^\star) &= 1 + \int_{x^\star - M_g}^{\tau_g} \frac{1 - \frac{\tau_g - z}{M_g}}{M_g - \tau_g + z} dz \\ &= 1 + \int_{x^\star - M_g}^{\tau_g} \frac{1}{M_g} dz \\ &= \frac{1}{M_g}(-x^\star + \tau_g + 2M_g) \end{aligned}$$

- For the target distribution between $[\tau_g + M_g, 1]$: under assumption 6.2 and 6.4, we know that no agents will change their feature to this feature region. So the distribution between $[\tau_g + M_g, 1]$ remains the same as the source distribution.

Thus, the new feature distribution of $x_{\tau_g}^{(M_g)}$ after agents from group g strategic responding becomes:

$$\Pr_{\mathcal{T}}(x) = \Pr\left(x_{\tau_g}^{(M_g)}\right) = \begin{cases} 1, & x \in [0, \tau_g - M_g) \text{ and } x \in [\tau_g + M_g, 1] \\ \frac{\tau_g - x}{M_g}, & x \in [\tau_g - M_g, \tau_g) \\ \frac{1}{M_g}(-x + \tau_g + 2M_g), & x \in [\tau_g, \tau_g + M_g) \\ 0, & \text{otherwise} \end{cases} \tag{103}$$

Proof of Proposition 6.5:

Statement: For our assumed setting of strategic response involving DP for two groups $\{g, h\}$, Theorem 4.1 implies

$$\Delta^\star_{\mathsf{DP}}(\pi, \mathcal{T}) \leq \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{S}) + \tau_g(1 - \tau_g)\frac{2}{3}m_g + \tau_h(1 - \tau_h)\frac{2}{3}m_h \tag{104}$$

Proof. According to Lemma F.3, we can compute the variance of $w_g(x)$: $\mathrm{Var}(w_g(x)) = \mathbb{E}\left[\left(w_g(x) - \mathbb{E}[w_g(x)]\right)^2\right] = \frac{2}{3}M_g$. Then by plugging it to the general bound for Theorem 4.1 gives us the result.

Proof of Theorem 6.6:

Statement: For DP subject to label replicator dynamics,

$$\Delta^\star_{\mathsf{DP}}(\pi, \mathcal{T}) \leq \Delta^\star_{\mathsf{DP}}(\pi, \mathcal{S}) + \sum_g \left| Q_g[t+1] - Q_g[t] \right| \frac{|\rho_g^{1,1} - \rho_g^{0,1}|}{\rho_g^{1,1} + \rho_g^{0,1}} \tag{105}$$

Proof. We may directly substitute

$$|\mathcal{G}| = 2$$
$$B_g = \left|Q_g[t+1] - Q_g[t]\right|$$
$$\left|\beta_g^+ - \beta_g^-\right| = \frac{|\rho_g^{1,1} - \rho_g^{0,1}|}{\rho_g^{1,1} + \rho_g^{0,1}}$$

into Theorem 5.2.

Proof of Theorem D.1:

Statement: The true positive rate β_g^+ is bounded over the domain of covariate shift $\mathbb{D}_{\text{cov}}[\mathbf{B}]$, which we define by the bound $\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B}$, and the invariance of $\Pr(Y{=}1 \mid X{=}x, G{=}g)$ for all x, g, as

$$\frac{\cos\big(\phi_g^u\big)}{\cos\big(\xi_g - \phi_g^u\big)} \leq \beta_g^+(\pi, \mathcal{T}) \leq \frac{\cos\big(\phi_g^l\big)}{\cos\big(\xi_g - \phi_g^l\big)} \tag{106}$$

where

$$\phi_g^l \coloneqq \min_{\mathcal{D} \in \mathbb{D}_{\text{cov}}[\mathbf{B}]} \phi_g[\mathcal{D}]; \quad \phi_g^u \coloneqq \max_{\mathcal{D} \in \mathbb{D}_{\text{cov}}[\mathbf{B}]} \phi_g[\mathcal{D}] \tag{107}$$

Proof. To be rigorous, we may give an explicit expression for $\mathsf{r}_g^\perp$ by implicitly forming a basis in the $(1, \mathsf{t}_g)$-plane via the Gram-Schmidt process.

$$\mathsf{r}_g^\perp \coloneqq \langle \mathsf{r}_g, \mathsf{t}_g \rangle_g \frac{\mathsf{t}_g}{\|\mathsf{t}_g\|^2} + \langle \mathsf{r}_g, \mathsf{u}_g \rangle_g \frac{\mathsf{u}_g}{\|\mathsf{u}_g\|^2} \tag{108}$$

$$\mathsf{u}_g \coloneqq 1 - \langle 1, \mathsf{t}_g \rangle \frac{\mathsf{t}_g}{\|\mathsf{t}_g\|^2} \tag{109}$$

$$\tag{110}$$

From which we may verify that

$$\langle \mathsf{u}_g, \mathsf{t}_g \rangle = 0 \tag{111}$$
$$\langle \mathsf{r}_g^\perp, \mathsf{t}_g \rangle_g = \langle \mathsf{r}_g, \mathsf{t}_g \rangle_g \tag{112}$$
$$\langle \mathsf{r}_g^\perp, \mathsf{u}_g \rangle_g = \langle \mathsf{r}_g, \mathsf{u}_g \rangle_g \tag{113}$$
$$\langle \mathsf{r}_g^\perp, 1 \rangle_g = \langle \mathsf{r}_g, 1 \rangle_g \tag{114}$$

Recalling the relationship between the cosine of an angle between two vectors and inner products:

$$\cos(\angle(a, b)) = \frac{\langle a, b \rangle}{\|a\| \|b\|} \tag{115}$$

It follows from Equation (31) that, defining $\xi_g \coloneqq \angle(\mathsf{t}_g, 1)$,

$$\beta_g^+ \frac{\|1\|}{\|\mathsf{t}_g\|} = \frac{\cos(\angle(\mathsf{r}_g, \mathsf{t}_g))}{\cos(\angle(\mathsf{r}_g, 1))} = \frac{\cos\big(\angle(\mathsf{r}_g^\perp, \mathsf{t}_g)\big)}{\cos\big(\angle(\mathsf{r}_g^\perp, 1)\big)} = \frac{\cos(\phi_g)}{\cos(\xi_g - \phi_g)} \tag{116}$$

By the monotonicity of the final expression above with respect to ϕ_g, for fixed ξ_g:

$$\frac{\mathrm{d}}{\mathrm{d}x} \frac{\cos(x)}{\cos(\xi - x)} = -\frac{\sin(x)\cos(\xi - x) + \cos(x)\sin(\xi - x)}{\cos^2(\xi - x)} = -\frac{\sin(\xi)}{\cos^2(\xi - x)} \tag{117}$$

We note that Equation (117) is strictly negative, thus the expression in Equation (116) must be monotonic for fixed ξ. We may conclude that β_g^+ is extremized with extremal values of ϕ_g, denoted as ϕ_g^u and ϕ_g^l.