Feature Noise Induces Loss Discrepancy Across Groups

Fereshte Khani [1] **Percy Liang** [1]

Abstract

The performance of standard learning procedures has been observed to differ widely across groups. Recent studies usually attribute this loss discrepancy to an information deficiency for one group (e.g., one group has less data). In this work, we point to a more subtle source of loss discrepancy—feature noise. Our main result is that even when there is no information deficiency specific to one group (e.g., both groups have infinite data), adding the same amount of feature noise to all individuals leads to loss discrepancy. For linear regression, we thoroughly characterize the effect of feature noise on loss discrepancy in terms of the amount of noise, the difference between moments of the two groups, and whether group information is used or not. We then show this loss discrepancy does not vanish immediately if a shift in distribution causes the groups to have similar moments. On three real-world datasets, we show feature noise increases the loss discrepancy if groups have different distributions, while it does not affect the loss discrepancy on datasets where groups have similar distributions.

1. Introduction

Standard learning procedures such as empirical risk minimization have been shown to result in models that perform well on average but whose performance differ widely across groups such as whites and non-whites (Angwin et al., 2016; Barocas and Selbst, 2016). This *loss discrepancy* across groups is especially problematic in critical applications that impact people's lives (Berk, 2012; Chouldechova, 2017). Despite the vast literature on removing loss discrepancy (Hardt et al., 2016; Khani et al., 2019; Agarwal et al., 2018; Zafar et al., 2017), the direct removal of loss discrepancy might introduce other problems such as intra-group loss discrepancy (Lipton et al., 2018) and adverse long-term impacts (Liu et al., 2018). Therefore, it is important to understand the source of loss discrepancy.

Why do such loss discrepancies exist? The literature generally studies sources of loss discrepancy due to an "information deficiency" of one group—that is, one group has, for example, more noise (Corbett-Davies et al., 2017), less training data (Chouldechova and Roth, 2018; Chen et al., 2018), biased prediction targets (Madras et al., 2019), or less-predictive features (Chen et al., 2018). Some work also states that groups have different risk distributions, and thus making hard (binary) decisions on such distributions causes loss discrepancy (Corbett-Davies and Goel, 2018; Canetti et al., 2019). In this work, we show that even under very favorable conditions—i.e., no bias in the prediction targets, *infinite* data, perfect predictive features for both groups, and no hard (binary) decisions (in regression)—adding the *same* amount of feature noise to all individuals still leads to loss discrepancy.

In order to study the effect of feature noise (which includes omitted features as a special case) and use of group information on loss discrepancy, we consider the following regression setup. We assume each individual belongs to a group $g \in \{0, 1\}$ and has latent features $z \in \mathbb{R}^d$ which cause the target $y = \beta^\top z + \alpha$. However, we only observe a noisy version of z through one of the following *observation functions*:

$$o_{-\mathrm{g}}(z, g, u) = [z + u], \tag{1}$$

$$o_{+\mathrm{g}}(z, g, u) = [z + u, g], \tag{2}$$

where $u \in \mathbb{R}^d$ is mean-zero noise independent of the rest of the variables, and the group membership g can be either included or not. We study the discrepancy of both the residual $(y - \hat{y})$, which measures the amount of underestimation and the squared error $((y - \hat{y})^2)$, which measures the overall performance. Abusing terminology, we call both *losses*.

We consider two common flavors of loss discrepancies: (i) *statistical loss discrepancy*, which measures the difference between the expected losses of the two groups (Hardt et al., 2016; Agarwal et al., 2018; Woodworth et al., 2017; Pleiss et al., 2017; Khani et al., 2019); and (ii) *counterfactual loss discrepancy*, which measures the difference between the loss of an individual and a "counterfactual" individual with

[1]Department of Computer Scinece, Stanford University. Correspondence to: Fereshte Khani <fereshte@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

the same characteristics but from another group (Kusner et al., 2017; Chiappa, 2019; Loftus et al., 2018; Nabi and Shpitser, 2018; Kilbertus et al., 2017).

We have two main results. First, we show that without using group information, feature noise causes statistical loss discrepancy determined by four factors: the amount of feature noise and the difference between means, variances, and sizes of the groups. In particular, the loss discrepancy based on residual is proportional to the difference between means, and the loss discrepancy based on squared error is proportional to the difference between variances (Proposition 1). Our second result is that using group information ($o_{+\mathrm{g}}$) alleviates the statistical loss discrepancy but causes high counterfactual loss discrepancy (Proposition 2).

To better understand the effect of using group information, we further decompose the incurred loss discrepancy into two terms, one related to the moments of training distribution, and one related to the moments of the test distribution. We show that the statistical loss discrepancy of $o_{-\mathrm{g}}$ is mainly due to differences in the test distribution, and it vanishes immediately if a shift in distribution causes the groups to have similar distributions. Meanwhile, the loss discrepancy of $o_{+\mathrm{g}}$ is mainly due to differences in the training distribution, and it does not vanish immediately after shifts in the population (Proposition 3).

We validate our results on three real-world datasets: for predicting the final grade of secondary school students , final GPA of law students , and crime rates in the US communities , where the group g is either race or gender. We consider two types of feature noise: (i) adding the same amount of noise to every feature and (ii) omitting features. We show that on the Communities&Crime and Students datasets where groups have different means, variances, and sizes, noise leads to high loss discrepancy. On the other hand, on the Law School dataset, where groups have similar means and variances, noise does not affect the loss discrepancy. Finally, for the datasets with high loss discrepancy, we consider a distribution shift to a re-weighted dataset where groups have similar means and show that the loss discrepancy of the estimator using $o_{-\mathrm{g}}$ vanishes immediately while the loss discrepancy of the estimator using $o_{+\mathrm{g}}$ vanishes more slowly with the rate studied in Proposition 3.

2. Setup

We consider the following regression setup, summarized in Figure 1. We assume each individual belongs to a group $g \in \{0, 1\}$, e.g., whites and non-whites; and has latent (unobservable) features, $z \in \mathbb{R}^d$ which cause the prediction target $y \in \mathbb{R}$. For each individual, we observe $x = o(z, g, u)$ through an observation function o, where $u \in \mathbb{R}^d$ is a random vector representing the source of (feature) noise in observation.

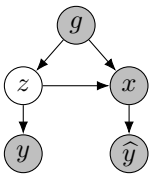


Figure 1: In this work, we consider a prediction problem from x to $\hat{y}$ where the output y is a deterministic function of unobserved random vector z.

As an example, the latent feature (z) is the knowledge of a student in d subjects, and y is her score in an entrance exam, which is a combination of the different subjects. However, we only observe a noisy version of z via exam scores, the school's name, or letter of recommendation, where the latter two might reveal information about group membership (g).

Let $h : \mathbb{R}^d \to \mathbb{R}$ be a predictor, and $\hat{y} = h(o(z, g, u))$ be the predicted value for individual z. We measure the impact of the predictor for an individual through a loss function $\ell(\hat{y}, y)$ (e.g., for squared error, $\ell(\hat{y}, y) = (\hat{y} - y)^2$), abbreviated as ℓ when clear from the context. In the entire paper, we analyze the population setting (infinite data) since we show that the effect of feature noise does not even vanish in this favorable setting.

2.1. Loss Discrepancy Notions

There are two flavors of loss discrepancies: statistical and counterfactual loss discrepancy. Statistical loss discrepancy measures how much two groups are impacted differently. This notion has been studied in economics and machine learning under the names of disparate impact, equal opportunity, classification parity, etc. (Arrow, 1973; Phelps, 1972; Hardt et al., 2016; Corbett-Davies and Goel, 2018). Here, we define statistical loss discrepancy similar to mentioned work, but looking at a general loss function.

Definition 1 (Statistical Loss Discrepancy (SLD))**.** *For a predictor h, observation function o, and loss function ℓ, statistical loss discrepancy is the difference between the expected loss between two groups:*

$$\mathit{SLD}(h, o, \ell) = |\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell \mid g = 0]| \qquad (3)$$

SLD operates at the group level and indicates how much a predictor yields higher loss for one group. However, it does not provide any guarantees at the individual level.

For individuals, counterfactual loss discrepancy measures how much two similar individuals (in our setup, two individuals that have the same z, not necessarily the same x) are treated differently because of their group membership. Here, we define counterfactual loss discrepancy similar to prior work (Kusner et al., 2017; Nabi and Shpitser, 2018), but looking at a general loss function.

Definition 2. *(Counterfactual Loss Discrepancy (CLD)) For a predictor h, observation function o, and loss function*

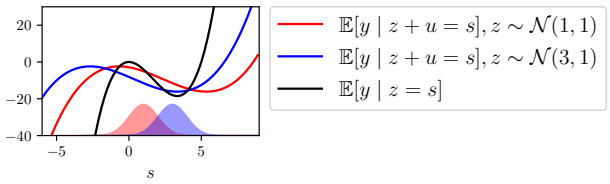


Figure 2: In this example $y = f(z) = z^3 + 5z^2$, and we observe $z + u$ instead of z where $u \sim \mathcal{N}(0, 1)$. As shown in (5), the best estimate of y (i.e., $\mathbb{E}[y \mid z + u]$) depends on the distribution of z. The blue line is the best estimate when $z \sim \mathcal{N}(1, 1)$ and the red line is the best estimate when $z \sim \mathcal{N}(3, 1)$.

ℓ, counterfactual loss discrepancy is the expected difference between the loss of an individual and its counterfactual counterpart:

$$\mathit{CLD}(h, o, \ell) = \mathbb{E}\left[|L_0 - L_1|\right], \tag{4}$$

where $L_{g'} = \mathbb{E}[\ell(h(o(z, g', u)), y)|z]$.

There are many concerns regarding the meaningfulness of CLD when group identity is an immutable characteristic (e.g., race and sex) (Holland, 1986; Freedman, 2004; Holland, 2003), which we discuss further in Section 7.

Note that CLD and SLD are not comparable. A model can have the same loss for similar individuals (CLD $= 0$), but since groups have different distributions over individuals, it can have higher loss for one group (SLD $\neq 0$). Conversely, a model can induce different losses for similar individuals due to their group membership (CLD $\neq 0$), but when averaged over the groups, it can result in similar expected losses for both groups (SLD $= 0$). See Proposition 4 for the construction.

It is clear that the observation function can asymmetrically affect groups and cause high loss discrepancy. For example, the observation function (o) can add noise to the features *only* when $g = 0$ or systematically report a lower value of z for one group. However, in this work, we are interested to see if it is possible that adding the same amount of noise (in a symmetric way) to all individuals affects groups differently (i.e., causes high loss discrepancy). We answer this question in the affirmative and exactly characterize the group distributions that are more susceptible to have high loss discrepancy under feature noise.

3. Feature Noise

We are interested in additive feature noise with mean zero that is independent of other variables (z and y). Feature noise is also known as classical measurement error (Carroll et al., 2006). We allow for any noise distribution—e.g., Laplace, Gaussian, or any discrete distribution—as long as it has mean zero. The independence assumption means u is independent of the value of z_i, but the feature noise can have different distributions for different features. Omitted features can be simulated by having noise with infinite variance. Feature noise and omitted features are pervasive in real-world applications; examples include test scores for college admissions or interview scores for hiring.

Note that without feature noise, i.e., $x = z$, the Bayes optimal predictor, $\mathbb{E}[y \mid x]$, does not depend on the distribution of z, but this no longer holds with feature noise. Formally, let $y = f(z)$ and u denote the additive noise on each feature, i.e., we observe $x = z + u$ instead of z. In this case, the Bayes-optimal predictor $\mathbb{E}[y \mid x]$, depends on the distribution of inputs ($\mathbb{P}_z$):

$$\mathbb{E}[y \mid x] = \frac{\int \mathbb{P}_u(u)\mathbb{P}_z(x-u)f(x-u)du}{\int \mathbb{P}_u(u)\mathbb{P}_z(x-u)du}. \tag{5}$$

Figure 2 shows an example of this dependence.

Feature noise has been extensively studied in linear regression (e.g., Fuller (2009)). In the rest of the paper, we focus on linear regression and show feature noise can cause loss discrepancy across groups.

3.1. Feature Noise in Linear Regression Background

In this section, we give a brief background on how feature noise makes parameter estimates inconsistent, in the simplified setting. We study the effect of feature noise on groups in Section 4. Let β, α denote the true parameters such that for each individual, $y = \beta^\top z + \alpha$, [1] and assume we observe $x = z + u$. When u is feature noise (i.e., mean-zero and independent of other variables), we can analyze the estimated parameters via least squares (Frisch, 1934).

$$\hat{\beta} = \Sigma_x^{-1}\Sigma_{xy} = (\Sigma_z + \Sigma_u)^{-1}\Sigma_z\beta \tag{6}$$

$$\hat{\alpha} = (\beta - \hat{\beta})^\top \mathbb{E}[z] + \alpha, \tag{7}$$

where for any two random vectors v and w, $\Sigma_{vw} = \mathbb{E}[(v - \mu_v)(w - \mu_w)^\top]$ denotes the covariance matrix between v and w, and $\mu_v = \mathbb{E}[v]$ denote the expected value of v. We write Σ_v for Σ_{vv}. To simplify notation, let $\Lambda \stackrel{\text{def}}{=} (\Sigma_z + \Sigma_u)^{-1}\Sigma_u$ denote the noise to signal ratio, then $\hat{\beta} = (I - \Lambda)\beta$ and $\hat{\alpha} = (\Lambda\beta)^\top \mathbb{E}[z] + \alpha$. Finally, for these estimated parameters the squared error is:

$$(\Lambda\beta)^\top \Sigma_z \Lambda\beta + ((I - \Lambda)\beta)^\top \Sigma_u ((I - \Lambda)\beta). \tag{8}$$

Note that the actual estimator only has access to x, but our analysis is in terms of z and u. If all variables are one-dimensional, then $\hat{\beta} = \frac{\Sigma_z}{\Sigma_z + \Sigma_u}\beta < \beta$, where $\frac{\Sigma_z}{\Sigma_z + \Sigma_u}$ is the

[1] Observing a noisy version of y does not change the estimate of parameters in presence of infinite data. For simplicity, we consider noiseless y.

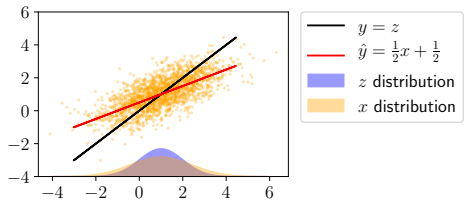


Figure 3: In the presence of feature noise, least squares estimator is not consistent; and the estimated slope (red line) is smaller than the true slope (black line). Here the true feature is $z \sim \mathcal{N}(1, 1)$, the observed features is $x \sim \mathcal{N}(z, 1)$, and the prediction target $y = z$.

relative size of the true signal and is known as attenuation bias (See Wager et al. (2013) for a connection between regularization and feature noise) Figure 3 shows the estimated line which predicts y from x in comparison to the true line which predicts y from z.

4. CLD and SLD for Linear Regression

We now show how feature noise affects SLD (Definition 1) and CLD (Definition 2) for linear regression. We focus on two loss functions when computing CLD and SLD:

- Residual: measures the amount of underestimation.

$$\ell_{\text{res}}(\hat{y}, y) = y - \hat{y}. \tag{9}$$

- Squared error: measures overall performance.

$$\ell_{\text{sq}}(\hat{y}, y) = (y - \hat{y})^2. \tag{10}$$

In this section, we calculate eight metrics according to different notions of loss discrepancy (CLD and SLD), different losses (ℓ_{res} and ℓ_{sq}), and whether group information is used or not ($o_{+\text{g}}$ and $o_{-\text{g}}$). Table 1 demonstrates three points: 1) In the presence of feature noise, SLD is not zero. 2) Using group membership reduces SLD, but increases CLD. 3) Groups are more susceptible to loss discrepancy based on residual when they have different means; they are more susceptible to loss discrepancy based on squared error when they have different variances.

4.1. Effect of Noise Without Group Information

In the entrance exam example in Section 2, suppose we observe each students' exam performance, which is a noisy version of their true knowledge of a subject. How does this noise affect the prediction? Is it possible that this symmetric independent noise over all features and all individuals affect groups differently?

In this section, we show observing a noisy version of z without any information about group membership (g) leads

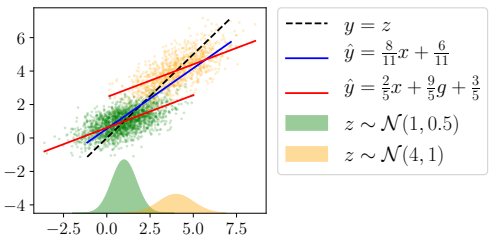


Figure 4: An illustration of feature noise and its effect on CLD and SLD. There are two groups: green and orange. The true function (dashed black line) is $y = z$. The feature noise $u \sim \mathcal{N}(0, 1)$. Predicting y from $x = o_{-\text{g}}(z, g, u) = z + u$ is the blue line which underestimates the target values for the orange group and thus has high SLD($o_{-\text{g}}, \ell_{\text{res}}$); however, since the prediction is independent of the group membership, it has CLD $= 0$. Predicting y from $o_{+\text{g}}(z, g, u) = [z + u, g]$ is the red lines, which has high CLD since groups are treated differently but low SLD.

to high SLD. Formally, let u denote the feature noise, we define the following observation function,

$$o_{-\text{g}}(z, g, u) \stackrel{\text{def}}{=} z + u. \tag{11}$$

In this case, group information is not encoded in the observation function ($o_{-\text{g}}(z, 0, u) = o_{-\text{g}}(z, 1, u)$); therefore, CLD $= 0$. But, as we show, SLD depends on the distribution of z.

Let's first consider a simple one-dimensional case. Figure 4 shows two groups, where $z \sim \mathcal{N}(1, 0.5)$ for the green group ($g = 0$), and $z \sim \mathcal{N}(4, 1)$ for the orange group ($g = 1$), also the $g = 1$ group is twice as likely as the $g = 0$ group ($\mathbb{P}[g = 1] = 2\mathbb{P}[g = 0]$). The prediction target here is $y = z$. Let the noise be Gaussian $u \sim \mathcal{N}(0, 1)$, and we observe $x = o_{-\text{g}}(z, g, u) = z + u$. Concretely, we are interested in the statistical loss discrepancy between groups for the least squares estimator, which predicts y from x.

As shown in Section 3.1, having feature noise results in attenuation bias. In this example, we have $\text{Var}[z] = \mathbb{E}[\text{Var}[z \mid g]] + \text{Var}[\mathbb{E}[z \mid g]] = \frac{8}{3}$. Therefore, $\hat{\beta} = \frac{\Sigma_z}{\Sigma_z + \Sigma_u} = \frac{8}{11}$, and $\hat{\alpha} = \frac{6}{11}$ (see blue line in Figure 4).

Let's see how this attenuation bias affects different groups. As shown in the Figure 4, the prediction target for the orange group is underestimated. Intuitively, if the mean of a group deviates from the mean of the population, then the expected residual for that group is large. Therefore, the difference between means of the groups is a factor in loss discrepancy based on residual (ℓ_{res}).

$$\Delta\mu_z \stackrel{\text{def}}{=} \mathbb{E}[z \mid g = 1] - \mathbb{E}[z \mid g = 0]. \tag{12}$$

Secondly, since the green group is in the majority ($\mathbb{P}[g =$

$1] > \mathbb{P}[g = 0]$), the line has less bias for the green group. Hence, the difference between size of the groups also plays an important role in loss discrepancy.

$$\mathbb{P}[g = 1] - \mathbb{P}[g = 0] \tag{13}$$

Thirdly, as shown in 8, the squared error is related to variance of data points; so intuitively, the difference in variance should also be a main factor in loss discrepancy.

$$\Delta\Sigma_z \stackrel{\text{def}}{=} \text{Var}[z \mid g = 1] - \text{Var}[z \mid g = 0]. \tag{14}$$

Finally, as noise increases, the attenuation bias increases, thus the estimated line deviates more from the true line, leading to a higher loss discrepancy. The following proposition formalizes how SLD depends on the four factors above; see Appendix A for the proof.

Proposition 1. *Consider the observation function o_{-g} (11). Let $\Lambda \stackrel{\text{def}}{=} (\Sigma_z + \Sigma_u)^{-1}\Sigma_u$. The loss discrepancies for least squares estimator are as follows:*

$$\begin{aligned}
&\textit{CLD}(o_{-g}, \ell_{res}) = \textit{CLD}(o_{-g}, \ell_{sq}) = 0 \\
&\textit{SLD}(o_{-g}, \ell_{res}) = \left|(\Lambda\beta)^\top \Delta\mu_z\right| \\
&\textit{SLD}(o_{-g}, \ell_{sq}) = \Big|(\Lambda\beta)^\top \Delta\Sigma_z (\Lambda\beta) \\
&\qquad\qquad - (\mathbb{P}[g = 1] - \mathbb{P}[g = 0])((\Lambda\beta)^\top \Delta\mu_z)^2\Big|.
\end{aligned}$$

where $\Delta\mu_z$ and $\Delta\Sigma_z$ are as defined in (12) and (14).

Proposition 1 states that SLD is not zero in the presence of feature noise. Furthermore, it determines the characteristics of the group distributions which are more prone to incur high SLD. In particular, given fixed variance in z (therefore, fixed Λ), groups with higher difference in means are more susceptible to incur high SLD based on residuals. SLD based on squared error has two terms: the first term is related to the difference between variance of the groups, and the second term is non-zero if groups have different sizes. We observe the effect of the second term in a real-world dataset in Section 6.

4.2. Effect of Noise with Group Information

Now let us consider the setting where the predictor is allowed to use group information. Does the predictor put weight on group membership information (e.g., assigns non-zero weight on the group membership feature), or does the predictor ignore the group membership since all the necessary information (z) is available and we are in the infinite data limit?

In this section, we show that if the observation function reveals the group information, as feature noise increases, the estimator relies more on group information (thus resulting in high CLD). On the other hand, the reliance on group information alleviates SLD.

Formally, we define a new observation function that adds group membership g as a separate feature:

$$o_{+\text{g}}(z, g, u) \stackrel{\text{def}}{=} [z + u, g] \tag{15}$$

Let's first revisit the example in Figure 4. The goal is to predict y (where in this example, we simply have $y = z$). The red lines indicate the estimated line that the least squares estimator predicts for y given a noisy version of z and the group membership ($x = o_{+\text{g}}(z, g, u) = [z + u, g]$). In this case, having g as an additional feature enables the model to have different intercepts for each group. As a result, the average residual for each group is zero; therefore $\text{SLD}(o_{+\text{g}}, \ell_{\text{res}}) = 0$. However, this benefit comes at the expense of treating individuals with the same z differently.

For the squared error (ℓ_{sq}), since each group has its own intercept, the squared error is no longer related to difference in sizes or means. The following proposition characterize CLD and SLD for the least squares estimator using $o_{+\text{g}}$. See Appendix B for the proof.

Proposition 2. *Consider the observation function o_{+g} (15). Let $\Sigma_{z|g} = \mathbb{E}[\textit{Var}[z \mid g]]$, and $\Lambda' = (\Sigma_{z|g} + \Sigma_u)^{-1}\Sigma_u$. The estimated parameters using least squares estimator are:*

$$\hat{\beta} = \begin{bmatrix} (I - \Lambda')\beta \\ (\Lambda'\beta)^\top \Delta\mu_z \end{bmatrix}, \quad \hat{\alpha} = (\Lambda'\beta)^\top \mathbb{E}[z \mid g = 0] + \alpha.$$

The loss discrepancies are as follows:

$$\begin{aligned}
&\textit{CLD}(o_{+g}, \ell_{res}) = \left|(\Lambda'\beta)^\top \Delta\mu_z\right| \\
&\textit{CLD}(o_{+g}, \ell_{sq}) = \left|(\Lambda'\beta)^\top \Delta\mu_z\right| \mathbb{E}\left[\left|(\Lambda'\beta)^\top (2z - \mu_1 - \mu_0)\right|\right] \\
&\textit{SLD}(o_{+g}, \ell_{res}) = 0 \\
&\textit{SLD}(o_{+g}, \ell_{sq}) = \left|(\Lambda'\beta)^\top \Delta\Sigma_z (\Lambda'\beta)\right|,
\end{aligned}$$

where $\Delta\mu_z$ and $\Delta\Sigma_z$ are as defined in (12) and (14), and $\mu_1 \stackrel{\text{def}}{=} \mathbb{E}[z \mid g = 1]$ and $\mu_0 \stackrel{\text{def}}{=} \mathbb{E}[z \mid g = 0]$.

Proposition 2 states that the coefficient for group membership is $\hat{\beta}_{\text{g}} = (\Lambda'\beta)^\top \Delta\mu_z$ (note that this is similar to $\text{SLD}(o_{-\text{g}}, \ell_{\text{res}})$). By having this coefficient for g, the estimator has $\text{SLD}(o_{+\text{g}}, \ell_{\text{res}}) = 0$, but it results in $\text{CLD}(o_{+\text{g}}, \ell_{\text{res}}) = |\hat{\beta}_{\text{g}}|$. Returning to the example in Figure 4, note that the red lines have better performance for each group, both in terms of residuals and squared error.

Table 1 presents a summary of the 8 different types of loss discrepancies. We also study non independent noise in Appendix F, and infinite noise in Appendix E.

5. Persistence of Loss Discrepancy

So far, we assumed the training distribution used to estimate parameters is the same as the test distribution that we are

		Counterfactual Loss Discrepancy (CLD)	Statistical Loss Discrepancy (SLD)
ℓ_{res}	$o_{-\text{g}}$	0	$\lvert(\Lambda\beta)^\top \Delta\mu_z\rvert$
	$o_{+\text{g}}$	$\lvert(\Lambda'\beta)^\top \Delta\mu_z\rvert$	0
ℓ_{sq}	$o_{-\text{g}}$	0	$\lvert(\Lambda\beta)^\top \Delta\Sigma_z(\Lambda\beta) - (\mathbb{P}[g=1] - \mathbb{P}[g=0])((\Lambda\beta)^\top \Delta\mu_z)^2\rvert$
	$o_{+\text{g}}$	$\lvert(\Lambda'\beta)^\top \Delta\mu_z\rvert\mathbb{E}[\lvert(\Lambda'\beta)^\top(2z - \mu_1 - \mu_2)\rvert]$	$\lvert(\Lambda'\beta)^\top \Delta\Sigma_z(\Lambda'\beta)\rvert$

Table 1: Loss discrepancies between groups, as proved in Proposition 1 and 2. In summary: 1. Feature noise without group information ($o_{-\text{g}}$) causes high SLD (first and third row), 2. Using group information reduces SLD but increases CLD (second and forth row), and 3. In loss discrepancies based on residuals the difference between mean is important while for squared error the difference between variances is important.

interested in measuring loss discrepancy with respect to. But what if the train and test distributions are different? Our formulation presented in Proposition 1 and 2 can be refined to be in terms of train and test distributions as follows (for the sake of space, we only focus on residual loss ℓ_{res}),

$$\text{CLD}(o_{-\text{g}}, \ell_{\text{res}}) = 0 \tag{16}$$

$$\text{SLD}(o_{-\text{g}}, \ell_{\text{res}}) = \left|(\Lambda_{(\text{train})}\beta)^\top \Delta\mu_{z(\text{test})}\right| \tag{17}$$

$$\text{CLD}(o_{+\text{g}}, \ell_{\text{res}}) = \left|(\Lambda'_{(\text{train})}\beta)^\top \Delta\mu_{z(\text{train})}\right| \tag{18}$$

$$\text{SLD}(o_{+\text{g}}, \ell_{\text{res}}) = \left|(\Lambda'_{(\text{train})}\beta)^\top (\Delta\mu_{z(\text{train})} - \Delta\mu_{z(\text{test})})\right|, \tag{19}$$

where the subscript denotes whether the statistics are computed on the training or test distribution.

When group information is not used then CLD $= 0$ for any test distribution (16), and the incurred SLD of $o_{-\text{g}}$ is due to the difference in means of the groups in the test distribution and will vanish if groups starts to have same means due to covariate shift (17). On the other hand, when group information is used then CLD $\neq 0$ and it is proportional to the difference in the means of the groups in the *training* distribution (18). Furthermore, in Proposition 2 we showed when group membership is used then $\text{SLD}(o_{+\text{g}}, \ell_{\text{res}}) = 0$; however, this will no longer holds when $\Delta\mu_{z(\text{train})} \neq \Delta\mu_{z(\text{test})}$ due to the covariate shift (19). In summary, using $o_{+\text{g}}$ leads the loss discrepancies of the linear predictor to be more dependent on the training data, thus more persistent even when groups start to have the same means due to a covariate shift.

To study the persistence of loss discrepancy, we instantiate the above expressions in following simple setting. We consider two distributions:

- Initial distribution: The mean of z for group $g = 1$ is $-\mu$, and for group $g = 0$ is μ, the covariance of z for both groups is Σ.
- Shifted distribution: The mean of z for both groups is μ and its covariance is Σ.

Here we assume groups have the same covariances/sizes, but the same analysis applies more generally. The following proposition studies the persistence of loss discrepancies as we see data from the shifted distribution with higher probability.

Proposition 3. *For each $0 \leq t \leq 1$, let the training distribution be a mixture of the initial distribution with probability t and the shifted distribution with probability $1 - t$. Let $c_1 = \left((\Sigma + \Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu)$, $c_2 = \left((\Sigma + \Sigma_u)^{-1}\mu\mu^\top(\Sigma + \Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu)$. For a linear predictor which is trained on the above distribution and tested on the shifted distribution, we have:*

$$t\,(c_1 - |c_2|) \leq \textit{SLD}(o_{+g}, \ell_{res}) = \textit{CLD}(o_{+g}, \ell_{res}) \leq t\,(c_1 + |c_2|) \tag{20}$$

$$\textit{SLD}(o_{-g}, \ell_{res}) = \textit{CLD}(o_{-g}, \ell_{res}) = 0. \tag{21}$$

One way to interpret this proposition is as follows. We start with a batch from initial distribution, at each time step we predict the targets for the shifted distribution and then concatenate the new batch with the correct targets to the training data. At time K, the training data consists of $K + 1$ batches, where one batch is from the initial distribution with $\text{SLD}_{\text{initial}}(o_{+\text{g}}, \ell_{\text{res}})$ and K batches from the shifted distribution. In Proposition 3 terms, we can assume the training data is a mixture of initial distribution with probability $t = \frac{1}{K}$ and the shifted distribution with probability of $1 - t = \frac{K}{K+1}$. The loss discrepancy on the shifted distribution is $\text{SLD}_{\text{new}}(o_{+\text{g}}, \ell_{\text{res}}) \approx \frac{1}{K+1}\text{SLD}_{\text{initial}}(o_{+\text{g}}, \ell_{\text{res}})$, which converges to zero with rate $O(\frac{1}{K})$. For $o_{-\text{g}}$, we have $\text{CLD}(o_{-\text{g}}, \ell_{\text{res}}) = \text{SLD}(o_{-\text{g}}, \ell_{\text{res}}) = 0$ for all K. See Appendix C for the proof.

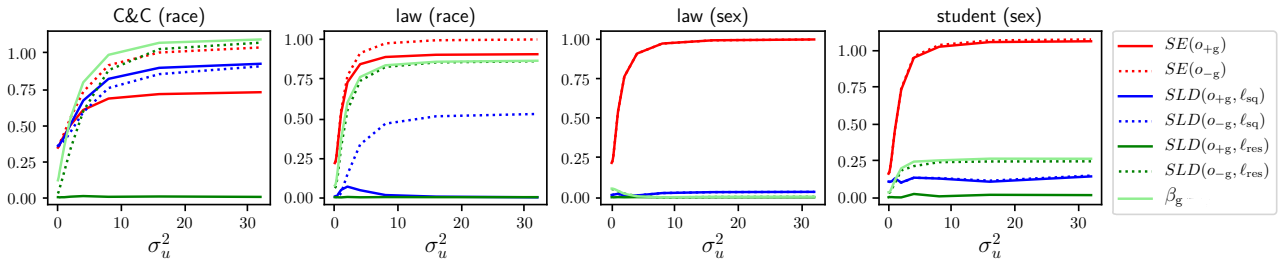
6. Experiments

Datasets. We consider three real-world datasets from the fairness literature. See Table 2 for a summary and Appendix G for more details.

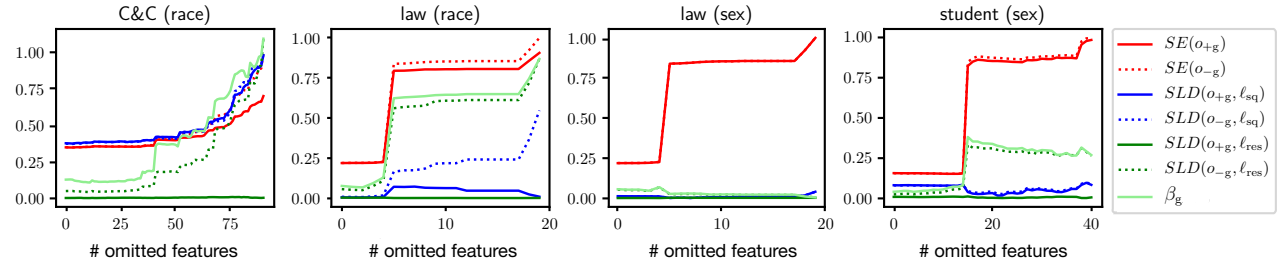
Our assumptions do not hold in these datasets: the features are not ideal (they might have information deficiency specific to one group), the model is misspecified (it is not linear), groups might have different true functions. However, we are still interested to see if adding noise on top of these (not

name	#records	#features	target	features example	group	$\mathbb{P}[g=1]$	$\Delta\mu_y$	$\Delta\sigma_y^2$	$\|\Delta\mu_{\boldsymbol{x}}\|_2$	$\|\Delta\Sigma_{\boldsymbol{x}}\|_F$
C&C	1994	91	crime rate	#homeless, average income, ...	race	0.50	1.10	0.96	5.62	12.75
law	20798	25	final GPA	undergraduate GPA, LSAT, ...	race	0.86	0.87	0.01	2.24	2.79
					sex	0.56	0.005	0.04	0.42	0.51
students	649	33	final grade	study time, #absences, ...	sex	0.59	0.26	0.12	1.40	2.26

Table 2: Statistics of the used datasets. Size of the first group is denoted by $\mathbb{P}[g=1]$ and $\Delta\mu_y$ and $\Delta\sigma_y^2$ denote the difference of mean and variance of the prediction target between groups, respectively.



(a) Adding noise increases squared error (SE) in all datasets; however, noise induces different loss discrepancy across the datasets.



(b) In all datasets except law(sex), omitting features affects groups differently and causes high SLD.

Figure 5: Statistical loss discrepancy (SLD) and squared error (SE) when (a) independent normal noise ($u \in \mathcal{N}(0, \sigma_u^2)$), is added to each feature (except for the group membership) (b) normal noise with high variance is added to the features sequentially (except for the group membership). We report $\hat{\beta}_g$ as a proxy for CLD(o_{+g}, ℓ_{res}).

ideal) features impacts groups differently, or whether the loss discrepancy remains the same as its initial value . We observe that the difference between moments of the groups is still a relevant factor governing loss discrepancy; and in the presence of feature noise, datasets where groups have different means, variances, and sizes are more susceptible to loss discrepancy.

Setup. We standardize all features and the target in all datasets (except the group membership feature) to have mean 0 and variance 1. We run each experiment 100 times, each time randomly performing a 80–20 train-test split of the data, and reporting the average on the test set. We compute the least squares estimator for each of the two observation functions: o_{-g} which only have access to non-group features, and o_{+g} which have access to all features. We consider two types of noise:

1. Equal noise: for different values of σ_u^2 we add independent normal noise ($u \sim \mathcal{N}(0, \sigma_u^2)$) to each feature except the group membership.
2. Omitting features: We start with a random order of the non-group features and omit features, which is nearly equivalent to adding normal noise with a very high variance ($u \sim \mathcal{N}(0, 10000)$) to them sequentially.

Loss discrepancy based on squared error. As expected, increasing the noise results in larger squared errors (SE). We see a smoother increase in Figure 5a, as opposed to the large jumps in Figure 5b, related to the "importance" of a feature.

We are now interested to see whether the observed increase differs across groups, thus inducing high loss discrepancy. In C&C, as we increase the amount of feature noise (σ_u^2), SLD increases (blue lines in Figure 5a), meaning that one group incurs higher loss compared to the other group. In law

(race) dataset, the sizes of the groups are very different (as shown in Table 2, whites represent 86% of the population). Recalling from Proposition 1, when group membership is not used ($o_{-\mathrm{g}}$), the minority group incurs higher loss; this is reflected in the observation that $\mathrm{SLD}(o_{-\mathrm{g}}, \ell_{\mathrm{sq}})$ (dotted blue line) increases as we add more noise. On the other hand, once group membership is used, the group size does not influence the loss discrepancy; therefore, we do not observe an increase in $\mathrm{SLD}(o_{+\mathrm{g}}, \ell_{\mathrm{sq}})$ (see the solid blue line). In law (sex) and students dataset, since groups have similar variance and sizes, we do not observe increase in loss discrepancy.

Loss discrepancy based on residual. When we estimate the parameters of linear regression, the bias (average residual) is always zero. Therefore, as we increase the noise, the average residual remains zero.

Is the average residual also zero for both groups or does adding noise cause a systematical over/underestimate for some groups (i.e., inducing loss discrepancy based on residual)? As discussed in Section 4.2, when group membership is used, then the average residual for each group is always zero ($\mathrm{SLD}(o_{+\mathrm{g}}, \ell_{\mathrm{res}}) = 0$)—see the solid green line in Figure 5. However, if group membership is not used ($o_{-\mathrm{g}}$), as discussed in Section 4.1, feature noise affects groups with different means differently and causes high loss discrepancy based on residuals (see dotted green line in Figure 5). The only dataset in which the loss discrepancy for residuals does not increase is law (sex), in which the considered groups have similar means.

Finally, as shown in Figure 5, weight of the group membership feature ($\hat{\beta}_{\mathrm{g}}$) increases as we increase the feature noise in datasets where groups have different means. Under the strong assumption that the observed features are the same for individuals from different groups (e.g., $x = z + u$), then $\mathrm{CLD}(o_{+\mathrm{g}}, \ell_{\mathrm{res}}) = |\hat{\beta}_{\mathrm{g}}|$. As shown in Table 1, there is a close relationship between $\mathrm{CLD}(o_{+\mathrm{g}}, \ell_{\mathrm{res}})$ and $\mathrm{SLD}(o_{-\mathrm{g}}, \ell_{\mathrm{res}})$. Although this assumption does not hold in practice, we still observe $\hat{\beta}_{\mathrm{g}}$ is close to $\mathrm{SLD}(o_{-\mathrm{g}}, \ell_{\mathrm{res}})$ (see the dotted green line and the light solid green line in Figure 5).

Persistence of loss discrepancy. We now study the persistence of loss discrepancy in the covariate shift setup introduced in Section 5. To simulate the shift, we consider the original distribution (uniform distribution over all data points), and a re-weighted distribution where weights are chosen such that the mean of the features (except for the group membership) and the mean of the prediction target are the same between both groups. We compute such a re-weighting using linear programming; see Appendix H for details. For different values of K, we compute two least squares estimators (with and without group membership) on a batch of size $n = 1000$ from the original distribution and a batch of size Kn from the re-weighted distribution. We then calculate loss discrepancy and average squared error (SE) of both models on the re-weighted distribution.

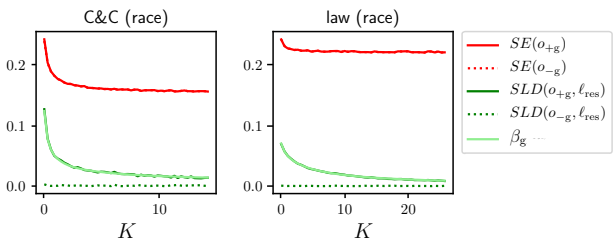


Figure 6: Loss discrepancy for the predictor learned on $o_{-\mathrm{g}}$ (dotted green line) is always zero. Loss discrepancy for the predictor learned on $o_{+\mathrm{g}}$ (solid green line) converges to 0 with rate of $\mathcal{O}(\frac{1}{K})$.

As stated in Proposition 3, the estimator without group membership (o_{-g}) achieves zero residual loss discrepancy immediately (the dotted green line in Figure 6). Meanwhile, the loss discrepancy of the estimator which uses group membership ($o_{+\mathrm{g}}$) vanishes more slowly with rate of $\mathcal{O}(\frac{1}{K})$ (solid green line).

7. Related Work and Discussion

While many papers focus on measuring loss discrepancy (Kusner et al., 2017; Hardt et al., 2016; Pierson et al., 2017; Simoiu et al., 2017; Khani et al., 2019) and mitigating loss discrepancy (Calmon et al., 2017; Hardt et al., 2016; Zafar et al., 2017), there are relatively few that study how loss discrepancy arises in machine learning models arises in the first place.

Chen et al. (2018) decompose the loss discrepancy into three components—bias, variance, and noise. They mainly focus on the bias and variance, and also consider scenarios in which available features are not equally predictive for both groups. There are also lines of work which assume the loss discrepancy of the model is due to biased target values (e.g., Madras et al. (2019)). Some work states that high loss discrepancy is due to lack of data for minority groups (Chouldechova and Roth, 2018). Some assume different groups have different functions (sometime in conflict with each other) (Dwork et al., 2018), and therefore, fitting the same model for both groups is suboptimal. In this work, we showed even when the prediction target is correct (not biased), with infinite data, the same function for both groups, equal noise for both groups, there is *still* loss discrepancy.

Recently, there is some work showing that enforcing fairness constraints without accurately understanding how they change the predictor results in worse outcomes for both

groups. Corbett-Davies and Goel (2018) look at different group fairness notions and show how they can lead to worse results if groups have different risk distributions. Liu et al. (2018) show that enforcing some fairness notions hurts the minorities in the long term. Lipton et al. (2018) show that removing disparate treatment and disparate impact simultaneously causes in-group discrimination. Our result that simple feature noise leads to loss discrepancy even under otherwise favorable conditions points at a more fundamental problem: the lack of information about individuals.

Problems with loss discrepancy notions. In this paper, we study statistical and counterfactual loss discrepancy, which are well established in the literature. However, our results naturally hinge on the meaningfulness of these notions, For completeness, we include a brief discussion about this.

Statistical loss discrepancy (Definition 1) measures the loss discrepancy between groups. SLD is valid if the considered distribution is representative (e.g., i.i.d samples) of the groups' real distribution. For example, consider the loan assignment task in which a model predicts whether a person will default or not. If a model has low SLD in a loan dataset, there is no guarantee for low SLD if we use the model in the real-world. Note that the loan dataset contains examples chosen by an entity to allocate loans (thus not i.i.d samples) and can be biased and not representative of the real-world distribution. Thus, one can claim perfect accuracy for one group according to a dataset, while for that group only people who clearly will not default are in the dataset. See Corbett-Davies and Goel (2018) for more examples.

Regarding counterfactual loss discrepancy (Definition 2), three main concerns need to be considered.

Immutability of group identity: Eminent researchers are opposed counterfactual reasoning with respect to an immutable characteristic (e.g., sex and race). They state that one cannot argue about the causal effect of a variable if its counterfactual cannot even be defined in principle (Holland, 1986; Freedman, 2004). In response, some social scientists study the effect of some mutable variables associated with group membership (mainly associated with the perception of group membership). For example, Bertrand and Mullainathan (2004) studied the effect of "racial soundingness" of a name in a resume for getting an interview. For more discussion on looking at race as a composite variable, see Sen and Wasow (2016).

Post-treatment bias: Characteristics such as race and sex are assigned at-conception before almost all other variables. Thus, considering the effect of these group identities while controlling for other variables that follow birth (z in our setup) introduce post-treatment bias (Rosenbaum, 1984) and can be misguided. Although this is a serious problem, CLD can still answer some valuable questions. For example, according to the disparate treatment law, a person is not liable for discrimination if she behaves in a trait-neutral manner. In particular, an employer can make a decision based on characteristics that are crucial for job performance. Informally, "a decider should avoid its own discrimination not the ones that are already exists" (Greiner and Rubin, 2011). Note that CLD is asking about intentional discrimination through observation function (therefore, it is conditioned on z). CLD differs from SLD, which focuses on loss discrepancies between groups (without any conditioning on z), which might unintentionally occur. See Greiner and Rubin (2011) for further discussion.

Inferring latent variables: in real-world problems, we only observe x and inferring z from data is a hard (if not impossible) task . Inferring z requires many strong assumptions regarding data generation. In this work, we assumed we have access to the latent variable z and focused on the effect of observation function on loss discrepancy. In particular, we showed that even the simple case that x is a noisy version of z leads to loss discrepancy. There is a rich line of work on checking the fairness of a model when true features and observation function need to be inferred from data (Kusner et al., 2017; Nabi and Shpitser, 2018; Chiappa, 2019; Kilbertus et al., 2017; Madras et al., 2019).

8. Conclusion and Future Work

In this work, we first pointed out that in the presence of feature noise, the Bayes-optimal predictor of y depends on the distribution of the inputs, which results in loss discrepancy for groups with different distributions. For linear regression, we showed (i) feature noise causes high SLD, (ii) using group information mitigates SLD but increases CLD, and (iii) using group information also makes the loss discrepancy more persistent under covariate shift. The studied loss discrepancies are not mitigated by collecting more data or designing a group-specific classifier, and designers should think of other methods such as feature replication to estimate the noise and de-noise the predictor (Carroll et al., 2006).

Our results rely on three main points: (i) we assume the true function is linear, (ii) we study the predictor with minimum squared error among linear functions, (iii) we consider two observation functions—feature noise with and without group information. Relaxing these points, especially studying more complex observation functions, would be a productive direction for future work.

Reproducibility. All code, data and experiments for this paper are available on the CodaLab platform at `https://worksheets.codalab.org/worksheets/0x7c3fb3bf981646c9bc11c538e881f37e`.

Acknowledgements. This work was supported by an Open Philanthropy Project Award. We would like to thank Emma Pierson, Pang Wei Koh, Ananya Kumar, and anonymous reviewers for useful feedback.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, pages 60–69.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 23.

Arrow, K. (1973). The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *104 California Law Review*, 3:671–732.

Bechavod, Y. and Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3992–4001.

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., and Smith, A. (2019). From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 309–318.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3539–3550.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 33, pages 7801–7808.

Chouldechova, A. (2017). A study of bias in recidivism prediciton instruments. *Big Data*, pages 153–163.

Chouldechova, A. and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 797–806.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th FUture BUsiness TEChnology Conference*.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133.

Freedman, D. A. (2004). Graphical models for causation, and the identification problem. *Evaluation Review*, 28(4):267–293.

Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*, volume 5. Universitetets Økonomiske Instituut.

Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons.

Greiner, D. J. and Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785.

Hardt, M., Price, E., and Srebo, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Holland, P. W. (2003). Causation and race. *ETS Research Report Series*, 2003(1).

Khani, F., Raghunathan, A., and Liang, P. (2019). Maximum weighted loss discrepancy. *arXiv preprint arXiv:1906.03518*.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in*

Neural Information Processing Systems (NeurIPS), pages 656–666.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4069–4079.

Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8125–8135.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. (2018). Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661.

Pierson, E., Corbett-Davies, S., and Goel, S. (2017). Fast threshold tests for detecting discrimination. *arXiv preprint arXiv:1702.08536*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5684–5693.

Redmond, M. and Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666.

Sen, M. and Wasow, O. (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19(1):499–522.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.

Simoiu, C., Corbett-Davies, S., Goel, S., et al. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216.

Wager, S., Wang, S. I., and Liang, P. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wightman, L. F. and Ramsey, H. (1998). *LSAC national longitudinal bar passage study*. Law School Admission Council.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, pages 1920–1953.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *World Wide Web (WWW)*, pages 1171–1180.

A. Proposition 1

Proposition 1. *Consider the observation function o_{-g} (11). Let $\Lambda \stackrel{\text{def}}{=} (\Sigma_z + \Sigma_u)^{-1}\Sigma_u$. The loss discrepancies for least squares estimator are as follows:*

$$\begin{aligned}
\textit{CLD}(o_{-g}, \ell_{res}) &= \textit{CLD}(o_{-g}, \ell_{sq}) = 0 \\
\textit{SLD}(o_{-g}, \ell_{res}) &= \left|(\Lambda\beta)^\top \Delta\mu_z\right| \\
\textit{SLD}(o_{-g}, \ell_{sq}) &= \Big|(\Lambda\beta)^\top \Delta\Sigma_z (\Lambda\beta) \\
&\qquad - (\mathbb{P}[g=1] - \mathbb{P}[g=0])((\Lambda\beta)^\top \Delta\mu_z)^2\Big|.
\end{aligned}$$

where $\Delta\mu_z$ and $\Delta\Sigma_z$ are as defined in (12) and (14).

Proof. Using least squares estimator, we have:

$$\begin{aligned}
\hat{\beta} =& \Sigma_x^{-1}\Sigma_{xy} \\
=& (\Sigma_z + \Sigma_u + \Sigma_{zu} + \Sigma_{uz})^{-1}(\Sigma_{zy} + \Sigma_{uy}) \qquad (22)
\end{aligned}$$

Due to assumptions (u is independent of other variables), $\Sigma_{zu} = 0$ and $\Sigma_{uy} = 0$, and from (**??**), $\Sigma_{zy} = \Sigma_z\beta$.

$$\begin{aligned}
\hat{\beta} =& (\Sigma_z + \Sigma_u)^{-1}\Sigma_z\beta \\
=& (I - (\Sigma_z + \Sigma_u)^{-1}\Sigma_u))\beta \\
=& (I - \Lambda)\beta \qquad (23)
\end{aligned}$$

The intercept formulation is as follows:

$$\hat{\alpha} = \beta^\top \mathbb{E}[z] + \alpha - \hat{\beta}^\top \mathbb{E}[z] = (\Lambda\beta)^\top \mathbb{E}[z] + \alpha, \qquad (24)$$

Since $o(z, 0, u) = o(z, 1, u)$, we have $\text{CLD}(o_{-\text{g}}, \ell_{\text{res}}) = \text{CLD}(o_{-\text{g}}, \ell_{\text{sq}}) = 0$.

For computing SLD based on residual, we first compute the expected ℓ_{res} for the first group:

$$\begin{aligned}
\mathbb{E}[y - \hat{y} \mid g=0] =& \mathbb{E}[\beta^\top z + \alpha - \hat{\beta}^\top(z+u) - \hat{\alpha} \mid g=0] \qquad (25) \\
=& \mathbb{E}[(\Lambda\beta)^\top z - (\Lambda\beta)^\top \mathbb{E}[z] - \hat{\beta}^\top u \mid g=0] \qquad (26)
\end{aligned}$$

Using $\mathbb{E}[u] = 0$ and $\mathbb{E}[z] = \mathbb{P}[g=0]\mathbb{E}[z \mid g=0] + \mathbb{P}[g=1]\mathbb{E}[z \mid g=1]$ we have:

$$\begin{aligned}
\mathbb{E}[y - \hat{y} \mid g=0] =& (\Lambda\beta)^\top(\mathbb{E}[z \mid g=0] - \mathbb{E}[z]) \qquad (27) \\
=& (\Lambda\beta)^\top(\mathbb{E}[z \mid g=0] - \mathbb{P}[g=0]\mathbb{E}[z \mid g=0] - \mathbb{P}[g=1]\mathbb{E}[z \mid g=1]) \qquad (28) \\
=& (\Lambda\beta)^\top((\mathbb{P}[g=0] - 1)(\mathbb{E}[z \mid g=1] - \mathbb{E}[z \mid g=0])) \qquad (29) \\
=& (\mathbb{P}[g=0] - 1)(\Lambda\beta)^\top \Delta\mu_z. \qquad (30)
\end{aligned}$$

Note that the first group ($g=0$) have lower expected ℓ_r if its size is large or $\Delta\mu_z$ (projected on $\beta^\top\Lambda^\top$) is small. Similarly for the second group ($g=1$) we have:

$$\mathbb{E}[y - \hat{y} \mid g=1] = (1 - \mathbb{P}[g=1])(\Lambda\beta)^\top \Delta\mu_z. \qquad (31)$$

Computing the difference between the expected residuals of the groups we have: $\text{SLD}(o_{-\text{g}}, \ell_{\text{res}}) = |(\Lambda\beta)^\top \Delta\mu|$.

For computing $\text{SLD}(o_{-\text{g}}, \ell_{\text{sq}})$, first note that the squared error can be decomposed to squared of bias and variance,

$$\mathbb{E}[(\hat{y} - y)^2] = \mathbb{E}[(\hat{y} - y)]^2 + \text{Var}[(\hat{y} - y)]. \qquad (32)$$

Using this decomposition, we have:

$$\begin{aligned}
\text{SLD}(o_{-\text{g}}, \ell_{\text{sq}}) =& |\mathbb{E}[\ell_{\text{sq}} \mid g=1] - \mathbb{E}[\ell_{\text{sq}} \mid g=0]| \qquad (33) \\
=& \left|\mathbb{E}[\ell_{\text{res}} \mid g=1]^2 - \mathbb{E}[\ell_{\text{res}} \mid g=0]^2 + \text{Var}[\ell_{\text{res}} \mid g=1] - \text{Var}[\ell_{\text{res}} \mid g=0]\right| \qquad (34)
\end{aligned}$$

Using (30) and (31), we have:

$$\mathbb{E}[\ell_{\text{res}} \mid g=1]^2 - \mathbb{E}[\ell_{\text{res}} \mid g=0]^2 = \mathbb{P}[g=0]^2((\Lambda\beta)^\top \Delta\mu_z)^2 - \mathbb{P}[g=1]^2((\Lambda\beta)^\top \Delta\mu_z)^2 \tag{35}$$

$$= -(\mathbb{P}[g=1] - \mathbb{P}[g=0])((\Lambda\beta)^\top \Delta\mu_z)^2. \tag{36}$$

For the difference between variances, we have:

$$\text{Var}[\ell_{\text{res}} \mid g=1] - \text{Var}[\ell_{\text{res}} \mid g=0] = \text{Var}\left[(\Lambda\beta)^\top z + \hat{\beta}^\top u \mid g=1\right] - \text{Var}\left[(\Lambda\beta)^\top z + \hat{\beta}^\top u \mid g=0\right] \tag{37}$$

$$\tag{38}$$

Since u is independent of z and g the difference is only related to the difference between variances.

$$\text{Var}[\ell_{\text{res}} \mid g=1] - \text{Var}[\ell_{\text{res}} \mid g=0] = \text{Var}[(\Lambda\beta)^\top z \mid g=1] - \text{Var}[(\Lambda\beta)^\top z \mid g=0] \tag{39}$$

$$= (\Lambda\beta)^\top \Delta\Sigma_z (\Lambda\beta) \tag{40}$$

Combining (36) and (40) completes the proof.

□

B. Proposition 2

Proposition 2. *Consider the observation function o_{+g} (15). Let $\Sigma_{z|g} = \mathbb{E}[Var[z \mid g]]$, and $\Lambda' = (\Sigma_{z|g} + \Sigma_u)^{-1}\Sigma_u$. The estimated parameters using least squares estimator are:*

$$\hat{\beta} = \begin{bmatrix} (I - \Lambda')\beta \\ (\Lambda'\beta)^\top \Delta\mu_z \end{bmatrix}, \qquad \hat{\alpha} = (\Lambda'\beta)^\top \mathbb{E}[z \mid g=0] + \alpha.$$

The loss discrepancies are as follows:

$$\begin{aligned}
&\textit{CLD}(o_{+g}, \ell_{res}) = \left|(\Lambda'\beta)^\top \Delta\mu_z\right| \\
&\textit{CLD}(o_{+g}, \ell_{sq}) = \left|(\Lambda'\beta)^\top \Delta\mu_z\right| \mathbb{E}\left[\left|(\Lambda'\beta)^\top (2z - \mu_1 - \mu_0)\right|\right] \\
&\textit{SLD}(o_{+g}, \ell_{res}) = 0 \\
&\textit{SLD}(o_{+g}, \ell_{sq}) = \left|(\Lambda'\beta)^\top \Delta\Sigma_z (\Lambda'\beta)\right|,
\end{aligned}$$

where $\Delta\mu_z$ and $\Delta\Sigma_z$ are as defined in (12) and (14), and $\mu_1 \stackrel{\text{def}}{=} \mathbb{E}[z \mid g=1]$ and $\mu_0 \stackrel{\text{def}}{=} \mathbb{E}[z \mid g=0]$.

Proof. For simplicity, let $z' \stackrel{\text{def}}{=} z + u$. We are interested in finding the best linear estimator for y, given z' and g. According to our assumptions we have:

$$y = \beta^\top z + \alpha \tag{41}$$

$$\mathbb{E}[y \mid z', g] = \beta^\top \mathbb{E}[z \mid z', g] + \alpha \tag{42}$$

First note that, we can represent $\mathbb{E}[z' \mid g]$ according to g linearly as follows:

$$\mathbb{E}[z' \mid g] = \mathbb{E}[z' \mid g=0] + (\mathbb{E}[z' \mid g=1] - \mathbb{E}[z' \mid g=0])g \tag{43}$$

We now write a linear predictor for $\mathbb{E}[z \mid z', g]$ given z' and g with some re-parametrization for simplicity. Define $v \stackrel{\text{def}}{=} z' - \mathbb{E}[z' \mid g]$ and $w \stackrel{\text{def}}{=} g - \mathbb{E}[g]$, we have:

$$\gamma_0 + \gamma_1(\underbrace{z' - \mathbb{E}[z' \mid g]}_{v}) + \gamma_2(\underbrace{g - \mathbb{E}[g]}_{w}) \tag{44}$$

If n denotes the dimension of z then γ_0 is $n \times 1$, γ_1 is $n \times n$ and finally γ_2 is $n \times 1$.

First note that $\mathbb{E}[v] = \mathbb{E}[z'] - \mathbb{E}[\mathbb{E}[z' \mid g]] = 0$. Therefore, we have:

$$\begin{aligned}
\text{Cov}(v, w) =& \mathbb{E}[vw] && (45)\\
=& \mathbb{E}[z'g] - \mathbb{E}[z']\mathbb{E}[g] - \mathbb{E}[\mathbb{E}[z' \mid g]g] + \mathbb{E}[\mathbb{E}[z' \mid g]]\mathbb{E}[g] && (46)\\
=& \mathbb{E}[z'g] - \mathbb{E}[\mathbb{E}[z' \mid g]g] = 0 && (47)\\
& && (48)
\end{aligned}$$

As a result, due to orthogonality, we can compute the parameters as follows:

$$\gamma_1 = \Sigma_v^{-1}\Sigma_{vz} \tag{49}$$

$$\Sigma_v = \mathbb{E}\left[\mathbb{E}\left[(z' - \mathbb{E}[z' \mid g])(z' - \mathbb{E}[z' \mid g])^\top \mid g\right]\right] = \mathbb{E}[\text{Var}[z' \mid g]] \tag{50}$$

For the covariance between v and z we have:

$$\begin{aligned}
\Sigma_{vz} =& \mathbb{E}[vz^\top] - \mathbb{E}[v]\mathbb{E}[z]^\top\\
=& \mathbb{E}\left[(z' - \mathbb{E}[z' \mid g])z^\top\right]\\
=& \mathbb{E}[z'z^\top] - \mathbb{E}[\mathbb{E}[z' \mid g]z^\top]\\
=& \mathbb{E}[zz^\top] + \mathbb{E}[uz^\top] - \mathbb{E}[\mathbb{E}[z + u \mid g]z^\top]\\
=& \mathbb{E}[zz^\top] - \mathbb{E}[\mathbb{E}[z \mid g]z^\top]\\
=& \mathbb{E}\left[\mathbb{E}\left[zz^\top \mid g\right] - \mathbb{E}[z \mid g]\mathbb{E}[z \mid g]^\top\right]\\
=& \mathbb{E}[\text{Var}[z \mid g]] && (51)
\end{aligned}$$

Combining the above equations and presenting $\hat{\beta}$ to two parts ($\hat{\beta}_z$ for coefficient of z' and $\hat{\beta}_g$ for the coefficient of g), we have:

$$\hat{\beta}_z = \gamma_1 = \mathbb{E}[\text{Var}(z' \mid g)]^{-1}\mathbb{E}[\text{Var}[z \mid g]]. \tag{52}$$

Using (43) and noting $\mathbb{E}[z' \mid g] = \mathbb{E}[z + u \mid g] = \mathbb{E}[z \mid g]$, we have $\gamma_2 = \Delta\mu$; therefore, $\hat{\beta}_g = \gamma_2 - \gamma_1\Delta\mu$

Now define Λ' to be:

$$\Lambda' = I - (\mathbb{E}[\text{Var}(z' \mid g)]^{-1}\mathbb{E}[\text{Var}[z \mid g]]) \tag{53}$$

Then the estimated parameters are:

$$\hat{\beta}_z = (I - \Lambda')\beta, \qquad \hat{\beta}_g = (\Lambda'\beta)^\top\Delta\mu, \tag{54}$$

For the intercept we have:

$$\begin{aligned}
\hat{\alpha} - \alpha =& \beta^\top\mathbb{E}[z] - \hat{\beta}_z^\top\mathbb{E}[z] - \hat{\beta}_g\mathbb{E}[g]\\
=& ((I - (I - \Lambda'))\beta)^\top\mathbb{E}[z] - (\Lambda'\beta)^\top\Delta\mu_z\mathbb{E}[g]\\
=& (\Lambda'\beta)^\top(\mathbb{E}[z] - (\mathbb{E}[z \mid g = 1] - \mathbb{E}[z \mid g = 0])\mathbb{E}[g])\\
=& (\Lambda'\beta)^\top\Big(\mathbb{E}[g]\mathbb{E}[z \mid g = 1] + (1 - \mathbb{E}[g])\mathbb{E}[z \mid g = 0] - \mathbb{E}[g]\mathbb{E}[z \mid g = 1] + \mathbb{E}[g]\mathbb{E}[z \mid g = 0]\Big)\\
=& (\Lambda'\beta)^\top\mathbb{E}[z \mid g = 0]. && (55)
\end{aligned}$$

Now that we calculated the estimated parameters, we compute loss discrepancies. SLD based on residuals is zero, since $\mathbb{E}[\ell_{\text{res}} \mid g = 0] = \mathbb{E}[\ell_{\text{res}} \mid g = 1] = 0$. We can calculate SLD based on the squared error using the same techniques as Proposition 1.

$$\text{SLD}(o_{+\text{g}}, \ell_{\text{sq}}) = (\Lambda'\beta)^\top\Delta\Sigma_z(\Lambda'\beta). \tag{56}$$

We now compute $\text{CLD}(o_{+\text{g}}, \ell_{\text{res}})$. Recall $L_{g'} \stackrel{\text{def}}{=} \mathbb{E}[\ell(y, h(o(z, g, u))) \mid z]$.

$$\begin{aligned}
\text{CLD}(o_{+\text{g}}, \ell_{\text{res}}) =& \mathbb{E}\left[|L_1 - L_0|\right] && (57)\\
=& \mathbb{E}\left[\left|\mathbb{E}[y - \hat{\beta}_z^\top z - \hat{\beta}_z^\top u - \hat{\beta}_g - \hat{\alpha} \mid z] - \mathbb{E}[y - \hat{\beta}_z^\top z - \hat{\beta}_z^\top u - \hat{\alpha} \mid z]\right|\right] && (58)\\
=& |\beta_g| && (59)\\
=& \left|(\Lambda'\beta)^\top \Delta\mu_z\right|. && (60)
\end{aligned}$$

For calculating $\text{CLD}(o_{+\text{g}}, \ell_{\text{sq}})$, let $\mu_1 = \mathbb{E}[z \mid g = 1]$ and $\mu_0 = \mathbb{E}[z \mid g = 0]$, then we have:

$$\begin{aligned}
\text{CLD}(o_{+\text{g}}, \ell_{\text{sq}}) =& \mathbb{E}\left[|L_1 - L_0|\right] && (61)\\
=& \mathbb{E}\left[\left|\mathbb{E}\left[\left(y - \hat{\beta}_z^\top (z+u) - \hat{\beta}_g - \hat{\alpha}\right)^2 \mid z\right] - \mathbb{E}\left[\left(y - \hat{\beta}_z^\top (z+u) - \hat{\alpha}\right)^2 \mid z\right]\right|\right] && (62)\\
=& \mathbb{E}\left[\left|((\Lambda'\beta)^\top (z - \mu_1))^2 - ((\Lambda'\beta)^\top (z - \mu_0))^2\right|\right] && (63)\\
=& \mathbb{E}\left[(\Lambda'\beta)^\top \left((z-\mu_1)(z-\mu_1)^\top - (z-\mu_0)(z-\mu_0)^\top\right)(\Lambda'\beta)\right] && (64)\\
=& \mathbb{E}\left[(\Lambda'\beta)^\top \left((\mu_1 - \mu_2)(2z - \mu_1 - \mu_2)^\top\right)(\Lambda'\beta)\right] && (65)\\
=& |((\Lambda'\beta)^\top \Delta\mu_z)| 2\mathbb{E}\left[\left|(\Lambda'\beta)^\top (z - \frac{\mu_1 + \mu_0}{2})\right|\right] && (66)
\end{aligned}$$

□

C. Proposition 3

Proposition 3. *For each $0 \le t \le 1$, let the training distribution be a mixture of the initial distribution with probability t and the shifted distribution with probability $1 - t$. Let $c_1 = \left((\Sigma + \Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu)$, $c_2 = \left((\Sigma + \Sigma_u)^{-1}\mu\mu^\top(\Sigma + \Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu)$. For a linear predictor which is trained on the above distribution and tested on the shifted distribution, we have:*

$$\begin{aligned}
t\left(c_1 - |c_2|\right) \le& \textit{SLD}(o_{+g}, \ell_{res}) = \\
& \textit{CLD}(o_{+g}, \ell_{res}) \le t\left(c_1 + |c_2|\right) && (20)\\
\textit{SLD}(o_{-g}, \ell_{res}) =& \textit{CLD}(o_{-g}, \ell_{res}) = 0. && (21)
\end{aligned}$$

Proof. Due to assumption, $\Delta\mu_{z(\text{test})} = 0$; therefore, $\text{CLD}(o_{-\text{g}}, \ell_{\text{res}}) = \text{SLD}(o_{-\text{g}}, \ell_{\text{res}}) = 0$ and $\text{SLD}(o_{+\text{g}}, \ell_{\text{res}})$ and $\text{CLD}(o_{+\text{g}}, \ell_{\text{res}})$ are equal.

As shown in Proposition 2, $\text{SLD}(o_{+\text{g}}, \ell_{\text{res}}) = (\Lambda'\beta)^\top \Delta\mu_{z(\text{train})}$. The mean difference $\Delta\mu_{z(\text{train})} = 2t\mu$, and converges to zero as t decreases. The challenge is to bound $\Lambda' = (\Sigma + 2t(1-t)\mu\mu^\top + \Sigma_u)^{-1}\Sigma_u$. (Sherman and Morrison, 1950) shows that if A is nonsingular and u, v are column vectors then

$$(A + uv^\top)^{-1} = A^{-1} - \frac{1}{1 + v^\top A^{-1} u} A^{-1} u v^\top A^{-1} \tag{67}$$

We can simplify Λ' using the above equation.

$$((\Sigma + \Sigma_u) + 2t(1-t)\mu\mu^\top)^{-1} = (\Sigma + \Sigma_u)^{-1} - \frac{2t(1-t)}{1 + 2t(1-t)\mu^\top(\Sigma + \Sigma_u)^{-1}\mu}(\Sigma + \Sigma_u)^{-1}\mu\mu^\top(\Sigma + \Sigma_u)^{-1} \tag{68}$$

First note that since we assumed the inverse of covariance matrix $(\Sigma + \Sigma_u)$ exists, therefore, it should be positive definite. As a result $\mu^\top(\Sigma + \Sigma_u)^{-1}\mu \ge 0$. Also note that $0 \le 2t(1-t) \le 1$; therefore, we can have the following bound:

$$0 \leq \frac{2t(1-t)}{1+2t(1-t)\mu^\top A^{-1}\mu} \leq 1 \tag{69}$$

For simplicity, let $r \stackrel{\text{def}}{=} \frac{2t(1-t)}{1+2t(1-t)\mu^\top A^{-1}\mu}$, using the above bound we can bound loss discrepancy when $o_{+\text{g}}$ is used:

$$\begin{aligned}\text{CLD}(o_{+\text{g}}, \ell_{\text{res}}) =& (\Lambda'_{\text{train}}\beta)^\top \Delta\mu_{\text{train}} && (70)\\ =& t\left(\left((\Sigma+\Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu) - r\left((\Sigma+\Sigma_u)^{-1}\mu\mu^\top(\Sigma+\Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu)\right) && (71)\end{aligned}$$

Defining $c_1 \stackrel{\text{def}}{=} \left((\Sigma+\Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu)$ and $c_2 \stackrel{\text{def}}{=} \left((\Sigma+\Sigma_u)^{-1}\mu\mu^\top(\Sigma+\Sigma_u)^{-1}\Sigma_u\beta\right)^\top (2\mu)$ we have:

$$t\left(c_1 - |c_2|\right) \leq \text{SLD}(o_{+\text{g}}, \ell_{\text{res}}) = \text{CLD}(o_{+\text{g}}, \ell_{\text{res}}) \leq t\left(c_1 + |c_2|\right) \tag{72}$$

□

D. CLD and SLD are not comparable

Proposition 4. *There exists a setting where CLD $= 0$ but SLD $\neq 0$, and another where SLD $= 0$ but CLD $\neq 0$.*

Proof. Assume there are only two kinds of individuals: z_1 and z_2. Assume $\frac{1}{4}$ of group one are z_1 and the rest are z_2. Assume $\frac{3}{4}$ of group two are z_1 and the rest are z_2.

If a predictor has loss 1 on z_1 and loss 2 on z_2 (independent of their group membership), then CLD $= 0$ but SLD $= 0.5$. If the loss is the same as above when $g = 0$ but when $g = 1$ the loss is 2 for z_1 and 1 for z_2, then CLD $= 1$, but SLD $= 0$. Let ℓ be squared error, z be a two dimensional vector, and $y = z_1 + z_2$. Let $\mathbb{P}(z_2 = 0 \mid g = 0) = 1$ but $\mathbb{P}(z_2 = 0 \mid g = 1) < 1$. If $o(z, g, u) = z_1$ then CLD $= 0$; however, SLD $\neq 0$ and the loss for the second group is higher. Now if $o(z, g) = [z_1, g]$ then define $h(o(z, g)) = z_1 + \mathbb{E}[y - z_1 \mid g]$ therefore, $SLD = 0$ while CLD $\neq 0$. □

E. Infinite noise

When noise is infinite, the predictor simply predicts $\mu_y = \mathbb{E}[y]$ for all data points when group membership is not available ($o_{-\text{g}}$), and it predicts the average of each group for the members of that group when group membership is available ($o_{+\text{g}}$). In this case, statistical loss discrepancy is only related to the moments of groups on y, see Table 3.

	CLD	SLD	average performance
$\boldsymbol{\ell}_{\text{res}}$	$o_{-\text{g}} : 0$ $o_{+\text{g}} : \Delta\mu_y$	$o_{-\text{g}} : \Delta\mu_y$ $o_{+\text{g}} : 0$	$o_{-\text{g}} : \mathbb{E}[\ell_{\text{res}}] = 0$ $o_{+\text{g}} : \mathbb{E}[\ell_{\text{res}}] = 0$
$\boldsymbol{\ell}_{\text{sq}}$	$o_{-\text{g}} : 0$ $o_{+\text{g}} : 2\Delta\mu_y\mathbb{E}[\lvert y - \mu_y - (\frac{1}{2} - \mathbb{E}[g])\Delta\mu_y\rvert]$	$o_{-\text{g}} : \Delta\sigma_y^2 + (1 - 2\mathbb{E}[g])\Delta\mu_y$ $o_{+\text{g}} : \Delta\sigma_y^2$	$o_{-\text{g}} : \mathbb{E}[\ell_{\text{sq}}] = \sigma_y^2$ $o_{+\text{g}} : \mathbb{E}[\ell_{\text{sq}}] = \sigma_{y\mid g}^2$

Table 3: A summery of metrics in the presence of infinite noise. Here, $\Delta\mu_y \stackrel{\text{def}}{=} \mathbb{E}[y \mid g = 1] - \mathbb{E}[y \mid g = 0]$ and $\Delta\sigma_y^2 \stackrel{\text{def}}{=} \text{Var}[y \mid g = 1] - \text{Var}[y \mid g = 0]$.

F. General noise

The independence assumption on the noise in Section 4.1 and 4.2 enables us to have a closed-form for CLD and SLD. Without any assumptions on the noise, we cannot specify anything about the estimated parameters more than (**??**). However, we can still analyze the form of SLD given the estimated parameters

Proposition 5. *Fix $o(z, g, u)$ as an arbitrary observation function, such that for $x = o(z, g, u)$ the covariance matrix (Σ_x) is invertible. Let $\hat{\beta}$ and $\hat{\alpha}$ be the estimated parameters as computed in (**??**). Equalize dimensions of $\hat{\beta}$, β, z, and x by*

adding extra 0*s at the end; and let* $u = x - z$ *denote the add-on error to the value of the true feature. The statistical loss discrepancies are as follows:*

$$\textit{SLD}(o, \ell_{\textit{res}}) = \left|\hat{\beta}^\top \Delta\mu_x - \Delta\mu_y\right| \tag{73}$$

$$= \left|(\hat{\beta}^\top - \beta^\top)\Delta\mu_z + \hat{\beta}^\top \Delta\mu_u\right|, \tag{74}$$

where for any random variable t, $\Delta\mu_t = \mathbb{E}[t \mid g = 1] - \mathbb{E}[t \mid g = 0]$. *For SLD based on* ℓ_{sq}, *we have:*

$$\textit{SLD}(o, \ell_{sq}) = \left|\Delta\Sigma_y + \hat{\beta}^\top \Delta\Sigma_x \hat{\beta} - 2\hat{\beta}^\top \Delta\Sigma_{xy}\right| \tag{75}$$

$$= \left|(\beta - \hat{\beta})^\top \Delta\Sigma_z (\beta - \hat{\beta}) + \hat{\beta}^\top \Delta\Sigma_u \hat{\beta} - 2(\beta - \hat{\beta})^\top \Delta\Sigma_{zu}\hat{\beta}\right| \tag{76}$$

where for any two random variable s, t *we define* $\Delta\Sigma_{st}$ *to be:*

$$\Sigma_{st} \stackrel{\text{def}}{=} \mathbb{E}[(s - \mu_s)(t - \mu_t)^\top \mid g = 1] - \mathbb{E}[(s - \mu_s)^\top (t - \mu_t) \mid g = 0]. \tag{77}$$

Proof.

$$\begin{aligned}
\text{SLD}(o, \ell_{\text{res}}) &= |\mathbb{E}[\ell_{\text{res}} \mid g = 1] - \mathbb{E}[\ell_{\text{res}} \mid g = 0]| \\
&= \left|\mathbb{E}\left[\beta^\top z + \alpha - \hat{\beta}^\top (z + u) - \hat{\alpha} \mid g = 1\right] - \mathbb{E}\left[\beta^\top z + \alpha - \hat{\beta}^\top (z + u) - \hat{\alpha} \mid g = 0\right]\right| \\
&= \left|(\beta^\top - \hat{\beta}^\top)(\mathbb{E}[z \mid g = 1] - \mathbb{E}[z \mid g = 0]) - \hat{\beta}^\top (\mathbb{E}[u \mid g = 1] - \mathbb{E}[u \mid g = 0])\right| \\
&= \left|(\beta - \hat{\beta})^\top \Delta\mu_z - \hat{\beta}\Delta\mu_u\right|
\end{aligned} \tag{78}$$

$\text{SLD}(o, \ell_{\text{res}})$ can also be formulated as follows:

$$\begin{aligned}
\text{SLD}(o, \ell_{\text{res}}) &= |\mathbb{E}[\ell_{\text{res}} \mid g = 1] - \mathbb{E}[\ell_{\text{res}} \mid g = 0]| \\
&= \left|\mathbb{E}\left[y - \hat{\beta}^\top x - \hat{\alpha} \mid g = 1\right] - \mathbb{E}\left[y - \hat{\beta}^\top x - \hat{\alpha} \mid g = 0\right]\right| \\
&= \left|\Delta\mu_y - \hat{\beta}^\top \Delta\mu_x\right|
\end{aligned} \tag{79}$$

This formulation implies that if $\Delta\mu_x = 0$ and $\Delta\mu_y = 0$ then for any arbitrary linear predictor $\text{SLD}(o, \ell_{\text{res}})$ is always zero.

For computing the statistical loss discrepancy based on squared error, first note that using (**??**), we have: $\hat{\alpha} = \beta\mu_z + \alpha - \hat{\beta}\mu_x \implies \alpha - \hat{\alpha} = -(\beta - \hat{\beta})^\top \mu_z + \hat{\beta}\mu_u$. We can now compute the expected squared error for the first group, as follows:

$$\mathbb{E}[\ell_{\text{sq}} \mid g = 0] = \mathbb{E}\Big[\left(\beta^\top z + \alpha - \hat{\beta}^\top (z + u) - \hat{\alpha}\right)^2 \mid g = 0\Big] \tag{80}$$

$$= \mathbb{E}\Big[\left((\beta - \hat{\beta})^\top (z - \mu_z) - \hat{\beta}^\top (u - \mu_u)\right)^2 \mid g = 0\Big]$$

$$= \mathbb{E}\Big[(\beta - \hat{\beta})^\top (z - \mu_z)(z - \mu_z)^\top (\beta - \hat{\beta}) + \hat{\beta}^\top (u - \mu_u)(u - \mu_u)^\top \hat{\beta} \tag{81}$$

$$- 2(\beta - \hat{\beta})^\top (z - \mu_z)(u - \mu_u)^\top \hat{\beta} \mid g = 0\Big] \tag{82}$$

The expected squared error for $g = 1$ is similar, computing the difference we have:

$$\text{SLD}(o, \ell_{\text{sq}}) = |\mathbb{E}[\ell_{\text{sq}} \mid g = 1] - \mathbb{E}[\ell_{\text{sq}} \mid g = 0]| \tag{83}$$

$$= \left|(\beta - \hat{\beta})^\top \Delta\Sigma_z (\beta - \hat{\beta}) + \hat{\beta}^\top \Delta\Sigma_u \hat{\beta} - 2(\beta - \hat{\beta})^\top \Delta\Sigma_{zu}\hat{\beta}\right| \tag{84}$$

Same as (79), we can compute SLD based on squared error in another form as well:

$$\begin{aligned}\mathbb{E}[\ell_{\text{sq}} \mid g=0] =& \mathbb{E}\left[\left(y-\hat{\beta}^\top x-\hat{\alpha}\right)^2 \mid g=0\right] \\ =& \mathbb{E}\left[\left(y-\hat{\beta}^\top x-\mu_y+\hat{\beta}^\top \mu_x\right)^2 \mid g=0\right] \\ =& \mathbb{E}\left[\left((y-\mu_y)-\hat{\beta}^\top (x-\mu_x)\right)^2 \mid g=0\right]\end{aligned} \tag{85}$$

$$\text{SLD}(o, \ell_{\text{sq}}) = |\mathbb{E}[\ell_{\text{sq}} \mid g=1] - \mathbb{E}[\ell_{\text{sq}} \mid g=0]| = \left|\Delta\Sigma_y + \hat{\beta}^\top \Delta\Sigma_x \hat{\beta} - 2\hat{\beta}^\top \Delta\Sigma_{xy}\right| \tag{86}$$

□

Equation (73) states that if the average over features and the target values are the same between groups then $\text{SLD}(o_{-\text{g}}, \ell_{\text{res}}) = 0$ for any linear predictor. Equation (74) states that if we cannot estimate the true parameters ($\hat{\beta} \neq \beta$) then in order to have $\text{SLD}(o_{-\text{g}}, \ell_{\text{res}}) = 0$ we should enforce different average add-on errors for the groups ($\Delta\mu_u \neq 0$).

G. Datasets

Students[2](Cortez and Silva, 2008) This dataset represents student achievements in secondary education of one Portuguese school in mathematics subject. The data features (x) include student grades, demographic, social, and school-related features and it was collected by using school reports and questionnaires. The target (y) is the final year grade, and we set the groups (g) to be males and females. Students dataset contains the students' 1st and 2nd period grades, which are strongly correlated with the prediction target (the final grade issued at the 3rd period).

Law School Admissions Council's National Longitudinal Bar Passage Study [3](Wightman and Ramsey, 1998) This dataset consists of the records of graduate students in law major. We set the target (y) to be the final Grade Point Average. The features include student grades and school-related features. We consider two versions of this data, one where g is race and the other where g is sex. The Law school dataset contains features such as first-year GPA; which are strongly correlated with the prediction target.

Communities and Crime [4](Redmond and Baveja, 2002) This dataset represents communities within the United States, each data point represents a community, and the goal is to predict the per capital violent crimes (y) given features such as average income in that community. This dataset contains eight continuous features related to the race, specifying the number and percentage of different races within that community. We replaced these features with a single binary feature; which is 1 if a community is in top 50% of communities with a majority of whites, and 0 if otherwise. We set this binary feature to be the indicate the groups.

H. Experimental details for persistence of loss discrepancy

In the experiment section (Section 6), we propose a re-weighted distribution under which groups have similar means. Here we explain how to define a new distribution on data points such that according to the new distribution groups have same mean. Let $X_1 \in \mathbb{R}^{n_1 \times d}$ be the data points for the first group and $X_2 \in \mathbb{R}^{n_2 \times d}$ be the data points for the second group. We compute $p_1 \in \mathbb{R}^{n_1}$ and $p_2 \in \mathbb{R}^{n_2}$, with the following linear programming:

[2] https://archive.ics.uci.edu/ml/datasets/student+performance

[3] Downloaded from https://github.com/jjgold012/lab-project-fairness (Bechavod and Ligett, 2017)

[4] http://archive.ics.uci.edu/ml/datasets/communities+and+crime

$$\max \quad \|p_1\|_1 \tag{87}$$
$$s.t. \quad p_1^\top X_1 = p_2^\top X_2 \tag{88}$$
$$p_1^\top y_1 = p_2^\top y_2 \tag{89}$$
$$0 \leq p_{1i} \leq 1, \qquad i = 1, \ldots, n_1 \tag{90}$$
$$0 \leq p_{2i} \leq 1, \qquad i = 1, \ldots, n_2 \tag{91}$$
$$\|p_1\|_1 = \|p_2\|_1 \tag{92}$$

We then sample points according to $\frac{p_1}{\|p_1\|_1}$ and $\frac{p_2}{\|p_2\|_1}$.