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 that 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. 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 that groups have similar distributions.

Feature Noise Induces Loss Discrepancy Across Groups

Fereshte Khani \(^1\)  Percy Liang \(^1\)

\(^1\)Department of Computer Science, Stanford University. Correspondence to: Fereshte Khani <fereshte@stanford.edu>.

Proceedings of the 37\textsuperscript{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).
We consider the following regression setup. 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.

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

Let \( h : \mathbb{R}^d \rightarrow \mathbb{R} \) 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 \( \ell \) 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. Figure 1 shows the underlying causal graph.

\[ \text{Figure 1: Schematic view of the causal graph considered in this work. The output } y \text{ is a deterministic function of unobserved random vector } z. \]

Our result are two folds. First, we show that feature noise causes high statistical loss discrepancy which is 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 3). Second, we show that using group information \( o_{+g} \) alleviates the statistical loss discrepancy but causes high counterfactual loss discrepancy (Proposition 3).

To understand the effect of using group information better, we further decompose the incurred loss discrepancy into two terms, where the first term is related to the moments of training distribution, and the second term is related to the moments of the test distribution. We show that the high statistical loss discrepancy of \( o_{-g} \) 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 high loss discrepancy of \( o_{+g} \) is mainly due to differences in the training distribution, and it does not vanish immediately after shifts in the population (Proposition 5).

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 in the Communities and Crime and Students datasets where groups have different means, variances, and sizes, noise leads to high loss discrepancy. On the other hand, in 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 least squares estimator using \( o_{-g} \) vanishes immediately while the loss discrepancy of the estimator using \( o_{+g} \) vanishes slower with the rate studied in Proposition 5.

2. Setup

We consider the following regression setup. 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...
For a predictor $z$ (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 counterfactual loss discrepancy (CLD)

Definition 2. (Counterfactual Loss Discrepancy (CLD)) For a predictor $h$, observation function $o$, and loss function $\ell$, counterfactual loss discrepancy is the expected difference between the loss of an individual and its counterfactual counterpart:

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

where $L_0' = \mathbb{E}[\ell(h(o(z, g', u)), y)|z]$.

There are many concerns regarding CLD when group identity is an immutable characteristic (e.g., race and sex) (Holland 1986; Freedman 2004; Holland 2003), we review and discuss the critics in Section 7. Note that CLD and SLD are not comparable. A model can treat similar individuals differently 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 proof.

It is clear that the observation function can asymmetrically affect groups and causes 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 groups’ distributions that are more susceptible to have high loss discrepancy under feature noise.

3. Feature Noise

We are interested in additive noise with mean zero and independent of other variables ($z$ and $y$), which we call feature noise (also known as classical measurement error (Carroll et al. 2006)). We allow any noise distributions—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 feature noise can have different distributions for different features. In the extreme case, we can have noise with infinite variance on one feature which simulates omitting a feature. Feature noise and omitted features are pervasive in real-world applications; examples include test scores for college admissions or interview scores for hiring.

First of all, we note that adding feature noise causes the estimate of $y$ to become dependent on the distribution of the inputs. Note that without feature noise, the Bayes optimal predictor, $\mathbb{E}[y | z]$, 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, $\mathbb{E}[y | x]$ as the best estimate of $y$, depends on the distribution of inputs ($\mathbb{P}_x$),

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

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 estimation inconsistent, in the simplified setting without considering groups. We study the effect of feature noise on groups in Section 4. Let $\beta, \alpha$ denote the true parameters such that for each individual, $y = \beta^\top z + \alpha$, 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).

$$\hat{\beta} = \Sigma_z^{-1}\Sigma_{yv} = (\Sigma_z + \Sigma_u)^{-1}\Sigma_z\beta \quad (6)$$
$$\hat{\alpha} = (\beta - \hat{\beta})^\top \mathbb{E}[z] + \alpha, \quad (7)$$

where for any two random vectors $v$ and $w$, $\Sigma_{vw} = \mathbb{E}[(v - \mu_v)(w - \mu_w)^\top]$ denote the covariance matrix between $v$ and $w$, and we write $\Sigma_v$ for $\Sigma_{vv}$. To simplify notation, let $\Lambda \overset{\text{def}}{=} (\Sigma_z + \Sigma_u)^{-1}\Sigma_u$ 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

\[1\] 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 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., $\mathbb{E}[y | z + u]$) depends on the distribution of data points. The blue line is the best estimate when $z \sim \mathcal{N}(1, 1)$ and the red line is the best estimate when $z \sim \mathcal{N}(3, 1)$.
We now show how feature noise affects SLD (Definition 1) as shown in the 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 \( y = z \).

The estimated parameters the squared error is:

\[
(\Lambda \beta)^\top \Sigma_z \Lambda \beta + ( (I - \Lambda) \beta)^\top \Sigma_u (I - \Lambda) \beta. 
\]

(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 \( \hat{\beta} = \frac{\Sigma_z \Sigma_x^{-1} \beta}{\Sigma_z + \Sigma_u} < \beta \), where \( \Sigma_z + \Sigma_u \) is the relative size of the true signal and is known as attenuation bias. 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.
  \[
  \ell_{\text{res}}(\hat{y}, y) = y - \hat{y}. 
  \]
  (9)

- Squared error: measures overall performance.
  \[
  \ell_{\text{sq}}(\hat{y}, y) = (y - \hat{y})^2. 
  \]
  (10)

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

4.1. Independent noise without group information

In the entrance exam example in Section 2, let’s assume we observe knowledge of students in each subject through some exam performance where it adds some noise to the result of each student on each subject. How does this noise affect the prediction? Is it possible that 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 to high SLD. Formally, let \( u \) denote the feature noise, we define the following observation function,

\[
o_{-g}(z, g, u) \overset{\text{def}}{=} z + u. 
\]

(11)

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

Let’s first consider a simple one-dimensional case to identify the important factors in SLD. Figure 4 shows two groups, where \( z \sim \mathcal{N}(1.05, 1) \) 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 \( (P[g = 1] = 2P[g = 0]) \). The prediction target here is \( y = z \). Let the noise be Gaussian \( u \sim \mathcal{N}(0, 1) \), and we observe \( x = o_{-g}(z, g, u) = z + u \). Concretely, we are interested in the statistical loss discrepancy between groups for the least squares estimator, which predicts \( y \) using \( x \).

As shown in Section 3.1 having noise in the features causes attenuation bias. In this example, we have \( \text{Var} [z] = E[\text{Var} [z \mid g]] + \text{Var} [E[z \mid g]] = \frac{s}{\beta^2} \). Therefore, \( \hat{\beta} = \frac{\Sigma_z + \Sigma_u}{\Sigma_z} = \frac{s}{\beta^2} \), and \( \hat{\alpha} = \frac{\alpha}{\beta^2} \) (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:

\[
y = \frac{5}{11} x + \frac{6}{11} 
\]

(12)

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 \( y = z \). Predicting \( y \) from \( x = o_{-g}(z, g, u) = z + \mathcal{N}(0, 1) \) is the blue line which underestimates the target values for the orange group and thus has high SLD \( o_{-g}, \ell_{\text{sq}} \); however, since the prediction is independent of the group membership, it has CLD = 0. Predicting \( y \) from \( o_{+g}(z, g, u) = [z + \mathcal{N}(0, 1), g] \) is the red lines, which has high CLD since groups are treated differently according to their group membership but low SLD.
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 \((r_{res})\).

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

Secondly, since the green group is in the majority \((P[g = 1] > P[g = 0])\), 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.

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

Thirdly, as shown in \(B\) 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.

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

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 \(o_{-g}\) \(\text{[11]}\).

Let \(\Lambda \overset{\text{def}}{=} (\Sigma_z + \Sigma_u)^{-1} \Sigma_u\). The loss discrepancies for least squares estimator are as follows:

\[
\text{CLD}(o_{-g}, \ell_{sq}) = \text{CLD}(o_{-g}, \ell_{res}) = \mathbb{E}[\Delta \mu_z^2] = 0
\]

\[
\text{SLD}(o_{-g}, \ell_{sq}) = \langle \Lambda \beta \rangle^\top \text{CLD}(o_{-g}, \ell_{sq}) = 0
\]

\[
\text{SLD}(o_{-g}, \ell_{sq}) = \langle \Lambda \beta \rangle^\top \text{CLD}(o_{-g}, \ell_{sq}) = 0
\]

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 \(\Lambda\)), 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. Independent Noise with Group Information**

Assuming group information is not revealed through observation function is a non-realistic assumption. What happens if group information is revealed? In a simple setting where we have feature noise and group information, does the predictor use the group membership information (i.e., assigns non-zero weight on the group membership feature)? Or, the predictor ignores the group membership since all the necessary information \((z)\) is available, and we are in infinite data setup?

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 results in high CLD). On the other hand, this reliance on group information alleviates SLD.

Formally, we define a new observation function as follows:

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

Let’s first go back to the discussed 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 least squares estimator predicts for \(y\) given a noisy version of \(z\) and the group membership \((x = o_{+g}(z, g, u) = [z + u, g])\). 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 SLD\((o_{+g}, \ell_{res})\) = 0. However, this benefit comes at the expense of treating individuals with the same \(z\) differently.

For the squared error \((\ell_{sq})\), 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 least squares estimator using \(o_{+g}\). See Appendix B for the proof.

**Proposition 2.** Consider the observation function \(o_{+g}\) \(\text{[13]}\).

Let \(\Sigma_{z[g]} = \mathbb{E}[\text{Var}[z \mid g]]\), and \(\Lambda' = (\Sigma_{z[g]} + \Sigma_u)^{-1} \Sigma_u\). The estimated parameters using least squares estimator are:

\[
\hat{\beta} = \left[ (I - \Lambda' \beta) \right]^{-1} \mathbb{E}[z \mid g = 0] + \alpha.
\]

The loss discrepancies are as follows:

\[
\text{CLD}(o_{+g}, \ell_{sq}) = \text{CLD}(o_{+g}, \ell_{res}) = \mathbb{E}[\Delta \mu_z^2]
\]

\[
\text{SLD}(o_{+g}, \ell_{sq}) = \text{SLD}(o_{+g}, \ell_{res}) = \mathbb{E}[\Delta \Sigma_z (\Lambda' \beta)]
\]

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

**Proposition 2** states that the coefficient for group membership is \(\beta_z \overset{\text{def}}{=} (\Lambda' \beta) \Delta \mu_z\) (note that this is similar to SLD\((o_{-g}, \ell_{res})\)). By having this coefficient for \(g\), the estimator has SLD\((o_{+g}, \ell_{res})\) = 0, but it also has CLD\((o_{+g}, \ell_{res})\) =
Feature Noise Induces Loss Discrepancy Across Groups

<table>
<thead>
<tr>
<th>Counterfactual Loss Discrepancy (CLD)</th>
<th>Statistical Loss Discrepancy (SLD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ell_{\text{res}}^{o-g}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\ell_{\text{res}}^{o_g}$</td>
<td>$\left</td>
</tr>
<tr>
<td>$\ell_{\text{sl}}^{o-g}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\ell_{\text{sl}}^{o_g}$</td>
<td>$\left</td>
</tr>
</tbody>
</table>

Table 1: Loss discrepancies between groups, as proved in Proposition 1 and 2. In summary: 1. Feature noise without group information ($o_{-g}$) 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.

$|\hat{\beta}_g|$. Going back to our discussed example in Figure 4. The red lines have better performance for each group, both in terms of residuals and squared error. However, this benefit comes at the expense of having high CLD.

Table 1 presents a summary of the computed 8 metrics in this section. We also study general noise in Appendix E, and infinite noise in Appendix F.

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 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 rewritten in terms of train and test distributions as follows (for the sake of space, we only focus on residual loss function $\ell_{\text{res}}$),

$$
\text{CLD}(o_{+g}, \ell_{\text{res}}) = \left| (\Lambda'_{\text{train}}\beta)^{\top} \Delta \mu_z(\text{train}) \right| \\
\text{SLD}(o_{+g}, \ell_{\text{res}}) = \left| (\Lambda'_{\text{train}}\beta)^{\top} (\Delta \mu_z(\text{train}) - \Delta \mu_z(\text{test})) \right| \\
\text{CLD}(o_{-g}, \ell_{\text{res}}) = 0 \\
\text{SLD}(o_{-g}, \ell_{\text{res}}) = \left| (\Lambda'_{\text{train}}\beta)^{\top} \Delta \mu_z(\text{test}) \right|, 
$$

(16)

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

Equation 16 gives us a better understanding about the effect of using group membership in the observation function. In particular, using $o_{+g}$ leads the loss discrepancies of the linear predictor to be more dependent on the training data. As a result, the loss discrepancies are more persistent even when groups start to have the same means due to a covariate shift. We study this persistence of loss discrepancy in the following setup. We consider two distributions:

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

Here we assume groups have same variances/sizes, but the same analysis works if groups have different variances or sizes. The following proposition studies the persistence of loss discrepancies as we see more data from the shifted distribution.

**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 $c_1 = ((\Sigma + \Sigma_u)^{-1}\Sigma_u\beta)^{\top}(2\mu)$, $c_2 = ((\Sigma + \Sigma_u)^{-1}\mu - (\Sigma + \Sigma_u)^{-1}\Sigma_u\mu_{\text{test}})^{\top}(2\mu)$. For a linear predictor which is trained on the above distribution and tested on the shifted distribution, we have:

$$
t \left( c_1 - c_2 \right) \leq \text{SLD}(o_{+g}, \ell_{\text{res}}) = \text{CLD}(o_{+g}, \ell_{\text{res}}) \leq t \left( c_1 + c_2 \right) \quad (17)$$

$$
\text{SLD}(o_{-g}, \ell_{\text{res}}) = \text{CLD}(o_{-g}, \ell_{\text{res}}) = 0. \quad (18)
$$

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

6. Experiments

**Datasets.** We consider three real-world datasets which are common in the fairness literature. See Table 2 for a summary and Appendix F for more details.

Our assumptions do not hold in these datasets: the features are not ideal (they might have information deficiency for 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 ideal) features impacts groups differently, or whether the loss discrepancy remains the same as its initial value. We
Feature Noise Induces Loss Discrepancy Across Groups

<table>
<thead>
<tr>
<th>name</th>
<th>#records</th>
<th>#features</th>
<th>target</th>
<th>features example</th>
<th>group</th>
<th>$P(g = 1)$</th>
<th>$\Delta \mu_y$</th>
<th>$\Delta \sigma^2_y$</th>
<th>$|\Delta \mu_x|_2$</th>
<th>$|\Delta \Sigma_x|_F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;C</td>
<td>1994</td>
<td>91</td>
<td>crime rate</td>
<td>homeless, average income, . . .</td>
<td>race</td>
<td>0.50</td>
<td>1.10</td>
<td>0.96</td>
<td>5.62</td>
<td>12.75</td>
</tr>
<tr>
<td>law</td>
<td>20798</td>
<td>25</td>
<td>final GPA</td>
<td>undergraduate GPA, LSAT, . . .</td>
<td>race</td>
<td>0.86</td>
<td>0.87</td>
<td>0.01</td>
<td>2.24</td>
<td>2.79</td>
</tr>
<tr>
<td>students</td>
<td>649</td>
<td>33</td>
<td>final grade</td>
<td>study time, #absences, . . .</td>
<td>sex</td>
<td>0.56</td>
<td>0.005</td>
<td>0.04</td>
<td>0.42</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Table 2: Statistics of the used datasets. Size of the first group is denoted by $P(g = 1)$ and $\Delta \mu_y$ and $\Delta \sigma^2_y$ denote the difference of mean and variance of the prediction target between groups, respectively.

Figure 5: Statistical loss discrepancy (SLD) and squared error (SE) when (a) independent normal noise ($u \sim N(0, \sigma_u^2)$), 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 $\beta_g$ as a proxy for CLD$^{\text{o}_g, \ell_{\text{sex}}}$.

(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.

observe that the difference between moments of the groups are still relevant factors governing loss discrepancy; and in the presence of feature noise, datasets with groups of 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 train two least squares estimator on two observation functions: $o_{\text{ng}}$ which only have access to non-group features, and $o_{\text{rg}}$ which have access to all features. We consider two types of noise:

1. Equal noise: for different values of $\sigma_u$ we add independent normal noise ($u \sim N(0, \sigma_u^2)$) 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 equivalent to adding normal noise with a very high variance ($u \sim N(0,10000)$) to them sequentially.

Loss discrepancy based on squared error. As we would expect, increases in the amount of 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 interested to see whether the observed increase in the squared error is similar for both groups, or the feature noise affected the groups differently, thus inducing high loss discrepancy. In C&C, as we increase the amount of feature noise ($\sigma_u$), SLD increases (blue lines in Figure 5a), meaning that one group is incurring higher loss in comparison
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 \((o_g)\), the minority group gets higher loss; this fact reflects in the observation that SLD\((o_g, \ell_{res})\) (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 SLD\((o_g, \ell_{res})\) (see the solid blue line). In law (sex) and students dataset, since groups have similar variance and sizes, we did 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 systematically 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 \(\text{SLD}(o_g, \ell_{res}) = 0\)—see the solid green line in Figure 5. However, if group membership is not used \((o_g)\), 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 \(\hat{\beta}_g\) 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., \(z = z + u\)), then CLD\((o_g, \ell_{res}) = |\hat{\beta}_g|\). As shown in Table 4 there is a close relationship between CLD\((o_g, \ell_{res})\) and SLD\((o_g, \ell_{res})\). Although this strong assumption does not hold in practice, we still observe \(\hat{\beta}_g\) is close to SLD\((o_g, \ell_{res})\) (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 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 F for details. For different values of \(K\), we train 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.

As stated in Proposition 4 the estimator without group membership \((o_g)\) should achieve zero residual loss discrepancy immediately. Meanwhile, the loss discrepancy of the estimator which uses group membership \((o_g)\) should vanish more slowly. The dotted green line in Figure 6 shows the immediate vanishing of loss discrepancy for \(o_g\). On the other hand, the loss discrepancy of \(o_g\) is more persistent and converges to zero with rate of \(O\left(\frac{1}{K}\right)\) (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; Simou 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. 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 because of 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 (sometime in conflict with each other) functions [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 in the lack of information about individuals.

Critics of loss discrepancy notions. 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 a perfect accuracy for one group according to a dataset while for that group only people 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 scholars 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 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 exits” (Greiner and Rubin, 2011).

Note that CLD is asking about intentional discrimination through observation function (therefore, it is conditioned on z). SLD differs from CLD, which focuses on loss discrepancies among groups (without any conditioning on z), which might unintentionally occur. See Greiner and Rubin (2011) for more 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 best estimate of y depends on the distribution of the inputs, which might result 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. 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.

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 each one of these points, especially studying more complicated observation functions, and study their effects on loss discrepancy are interesting future directions.

Reproducibility. All code, data and experiments for this paper are available on the Codalab platform at https://worksheets.codalab.org/worksheets/0x7c3f3fbf981e466c9b11c538e881f37e

Acknowledgements. This work was supported by Open Philanthropy Project Award. We would like to thank anonymous reviewers for useful feedback.
References


