Partial identification of the maximum mean discrepancy with mismeasured data

Ron Nafshi[1] **Maggie Makar**[2]

[1]Fintica AI, Tel Aviv, Israel
[2]University of Michigan, Ann Arbor, MI

Abstract

Nonparametric estimates of the distance between two distributions such as the Maximum Mean Discrepancy (MMD) are often used in machine learning applications. However, the majority of existing literature assumes that error-free samples from the two distributions of interest are available.We relax this assumption and study the estimation of the MMD under ϵ-contamination, where a possibly non-random ϵ proportion of one distribution is erroneously grouped with the other. We show that under ϵ-contamination, the typical estimate of the MMD is unreliable. Instead, we study partial identification of the MMD, and characterize sharp upper and lower bounds that contain the true, unknown MMD. We propose a method to estimate these bounds, and show that it gives estimates that converge to the sharpest possible bounds on the MMD as sample size increases, with a convergence rate that is faster than alternative approaches. Using three datasets, we empirically validate that our approach is superior to the alternatives: it gives tight bounds with a low false coverage rate.

1 INTRODUCTION

Many machine learning methods rely on comparing distances between distributions, with applications ranging from single cell sequencing [Schiebinger et al., 2019] to causal inference [Johansson et al., 2016]. The Maximum Mean Discrepancy (MMD) [Gretton et al., 2012] has emerged as a particularly useful nonparametric notion of distance between distributions. It has been widely used in robust predictive and reinforcement learning [Kumar et al., 2019, Makar et al., 2022, Li et al., 2017, Oneto et al., 2020, Veitch et al., 2021, Goldstein et al., 2022], fairness applications [Prost et al., 2019, Madras et al., 2018, Makar and D'Amour, 2022, Louizos et al., 2015] and distributionally robust optimization [Staib and Jegelka, 2019, Kirschner et al., 2020] among others. Despite its importance and widespread use, the majority of existing work using the MMD assumes that observed samples are measured without error. As we show in this work, if this assumption does not hold, the typical MMD estimate is unreliable.

Here, we study the estimation of the MMD where one of the samples observed is measured with error. Specifically, we consider ϵ-contamination, where a possibly non-random ϵ proportion of one of the two variables is erroneously grouped with the other. This mismeasurement mechanism arises in many important applications. One example of this setting arises from the fairness literature. For example, in settings where we wish to assess if a model gives different predictions across different race groups. Here ϵ-contamination arises if some non-random ϵ proportion of one race group is incorrectly grouped with the other. Beyond fairness, ϵ-contamination arises – for example – when trying to identify if there are biomarkers for Myocardial Infarction (MI). In this setting, we can use the MMD to detect differences in genome sequences between healthy individuals and patients with myocardial MI. Detecting differences between the two groups is complicated due to undiagnosed "silent" MI cases. These silent MI cases represent ϵ-contamination that occurs non-randomly: women's MI cases are more likely to go undiagnosed compared to men [Merz, 2011].

In this paper, we show that the typical MMD estimates are unreliable when the data is collected with ϵ-contamination. Instead, we resort to a partial identification approach, where we estimate upper and lower bounds on the MMD. We characterize upper and lower bounds that are credible, meaning that they contain the true unknown MMD, and sharp, meaning they cannot be made tighter without additional assumptions. Importantly, these bounds are identifiable using the observed contaminated data and an estimate of ϵ. We develop an estimation approach to compute the upper and lower bounds and analyze its behavior in finite samples. Our analysis shows that our approach gives estimates that converge to the sharpest possible upper and lower bounds as the sample size increases at a rate faster than the alternatives.

Our contributions are summarized as follows: **(1)** We show that under ϵ-contamination the typical estimates of the MMD are unreliable, **(2)** We characterize sharp upper and lower bounds on the unknown MMD that are identifiable using only the observed contaminated data, and an estimate of ϵ, **(3)** We propose an estimation approach to compute the upper and lower bounds and analyze its behavior in finite samples showing that its convergence to the true upper and lower bounds depends on the sample size and the value of ϵ, **(4)** We apply our approach to 3 datasets showing that it achieves a superior performance compared to alternative approaches, **(5)** We analyze the sensitivity of our approach to incorrect values of ϵ and give practical guidance on what to do if the true value of ϵ is unknown.

Related work. Most existing work on the MMD focuses on establishing statistically and computationally efficient estimators of the difference between two distributions under the assumption that the observed samples are error-free [Gretton et al., 2012, 2009, Schrab et al., Domingo-Enrich et al., 2023]. However, to our knowledge, the only existing work that tackles the challenge of measurement error is in the context of survival analysis, where the measurement error model arises from the classical right-censoring of the data [Fernández and Rivera, 2021]. By contrast, we study a different measurement error mechanism and suggest methods for partial identification of the MMD.

In the fairness literature, where comparisons between outcomes of different groups is important, Kallus et al. [2022] consider measurement error in the sensitive attribute. They consider a setting where we only have access to an imperfect proxy of the protected class membership and show that typical fairness metrics such as demographic parity and equalized odds are not identifiable. Similar to our work, they develop methods for partial identification of these metrics. A key difference between Kallus et al. [2022] and our work is that the former focuses on comparing a single moment (the mean) of two distributions whereas our work allows a more rigorous comparison of infinitely many moments of two distributions. We also stress that while the methods presented here could be used in a fairness context, they are more widely applicable to any setting where we wish to compare two distributions.

2 PRELIMINARIES

Our goal is to measure the distance between two distributions P_X and P_Y. However, instead of observing $X = \{x_i\}_i^n \sim P_X$, $Y = \{y_i\}_i^n \sim P_Y$, we observe ϵ-contaminated X' and Y', where a possibly non-random ϵ proportion of one of the two variables is incorrectly grouped with the other for $0 < \epsilon < 1$. Without loss of generality, we assume that the two samples have the same size $= n$ and that an ϵ-proportion of X is incorrectly grouped with Y. Specifically, let $C^* = \{c_i^*\}_i^m$, with $m = \lfloor \epsilon n \rfloor$ be the unobserved subset of X that is grouped with Y. We can express the distributions over the observed samples in relation to the true distributions and the unknown contaminated samples as follows:

$$P_{Y'} = (1-\alpha)P_Y + \alpha P_{C^*}$$
$$P_{X'} = (1+\tilde{\alpha})P_X - \tilde{\alpha} P_{C^*},$$

where $\alpha = \epsilon/(1+\epsilon)$ and $\tilde{\alpha} = \epsilon/(1-\epsilon)$. We do not make any additional assumptions about P_{C^*}. Importantly, we do not assume that the contamination is random, meaning *we do not assume* that $P_{C^*} = P_{X'} = P_X$.

We assume that the value of ϵ is known *a priori*, or can be empirically estimated from other data sources. However, in section 5.5, we conduct a sensitivity analysis to examine the performance of our approach and others under violations of this assumption. We use $\mathbb{E}_{P_A}[A]$ to denote the expectation of A according to the distribution $P_A(A)$, $A \cup B$ to denote the union of the set A and B, and $A \setminus B$ to denote the difference between the two sets A and B. We use $\#(A)$ to denote the cardinality of the set A. We use $\mathcal{X}'$ and $\mathcal{Y}'$ to denote the support of X' and Y' respectively.

We focus on the MMD as a measure of distance between distributions [Gretton et al., 2012]:

Definition 1 *For $Z \sim P_Z$, $Z' \sim P_{Z'}$, $\mathcal{F}$ such that $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$, and $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ with k being a positive definite kernel matrix, the MMD is defined as*

$$\mathrm{MMD}(\mathcal{F}, P_Z, P_{Z'}) = \sup_{f\in\mathcal{F}} \left(\mathbb{E}_{P_Z} f(Z) - \mathbb{E}_{P_{Z'}} f(Z')\right),$$

and the witness function f^* is defined as the function attaining the supremum in expression above, with $f^*(t) = \mathbb{E}_{P_Z}[k(Z,t)] - \mathbb{E}_{P_{Z'}}[k(Z',t)]$, up to a normalization constant.

When $\mathcal{F}$ is set to be a general reproducing kernel Hilbert space (RKHS), the MMD defines a metric on probability distributions, and is equal to zero if and only if $P_Z = P_{Z'}$. Throughout, we fix $\mathcal{F}$ to be the RKHS with $\|f\|_{\mathcal{F}} \leq 1$ for all $f \in \mathcal{F}$ and drop $\mathcal{F}$ from the MMD arguments to simplify notation. We use $k(z, z')$ to denote the reproducing kernel of $\mathcal{F}$, and assume that $0 \leq k(x', y') \leq \kappa$ for all $x', y' \in \mathcal{X}', \mathcal{Y}'$.

Gretton et al. [2012], showed that when there is no measurement error, the following empirical estimate of the MMD is unbiased:

$$\widehat{\mathrm{MMD}}(X,Y) = \frac{1}{n(n-1)} \sum_{i,j\neq i} k(x_i, x_j) \quad (1)$$

$$+ \frac{1}{n(n-1)} \sum_{i,j\neq i} k(y_i, y_j) - \frac{2}{n^2} \sum_{i,j} k(x_i, y_i). \quad (2)$$

As we show in the appendix section E, in the ϵ-contamination setting, $\widehat{\mathrm{MMD}}$ is not guaranteed to be an unbiased estimate, meaning $\widehat{\mathrm{MMD}}(X', Y')$ might not converge to $\mathrm{MMD}(P_{X'}, P_{Y'})$. So instead we study partial identifiability of $\mathrm{MMD}(P_X, P_Y)$. Meaning, our goal is to estimate credible and informative lower and upper bounds on the unknown $\mathrm{MMD}(P_X, P_Y)$. For those bounds to be informative, they should be *sharp*, meaning they cannot be made tighter without any additional assumptions.

3 THEORY

Our goal is to estimate upper and lower bounds that reflect our uncertainty in the MMD due to measurement error.

To proceed with our analysis, it is helpful to parameterize the MMD as function of the contaminated samples C. With some abuse of notation, for an arbitrary distribution P_C, we have that:

$$\begin{aligned}\mathrm{MMD}(P_C, P_{X'}, P_{Y'}) &= \sup_{f\in\mathcal{F}} \big[(1-\epsilon)\mathbb{E}_{P_{X'}} f(X') \\ &- (1+\epsilon)\mathbb{E}_{P_{Y'}} f(Y') + 2\epsilon\mathbb{E}_{P_C} f(C)\big], \end{aligned} \tag{3}$$

with $\mathrm{MMD}(P_X, P_Y) = \mathrm{MMD}(P_{C^*}, P_{X'}, P_{Y'})$. Our first result characterizes the sharpest possible bounds that can be attained without additional assumptions.

Proposition 1 *Let $(\mathcal{Y}', \Omega)$ be a measurable space with $Y' \in \mathcal{Y}'$ and let $\mathcal{P}$ be all the probability distributions on $(\mathcal{Y}', \Omega)$. Define $\mathcal{P}(\alpha)$ to be all the possible probability distributions over the unknown C^*, i.e., $\mathcal{P}(\alpha) = \{(P_{Y'}(Y') - (1-\alpha)\varphi)/\alpha : \varphi \in \mathcal{P}\}$, then the following bounds are sharp:*

$$\begin{aligned}&\inf_{P_C\in\mathcal{P}(\alpha)} \mathrm{MMD}(P_C, P_{X'}, P_{Y'}) \le \mathrm{MMD}(P_{C^*}, P_{X'}, P_{Y'}) \\ &\le \sup_{P_C\in\mathcal{P}(\alpha)} \mathrm{MMD}(P_C, P_{X'}, P_{Y'}),\end{aligned}$$

The proofs for proposition 1 and all other statements are presented in the appendix. The intuition for proposition 1 is simple: without any additional assumptions, C^* can take on any values in $\mathcal{Y}'$, and hence its corresponding distribution can be any distribution consistent with the observed data (i.e., any distribution $\in \mathcal{P}(\alpha)$). This means that the sharpest possible upper (lower) bound must be defined with respect to distributions over P_C that maximize (minimize) the MMD.

We use $P_{\overline{C}}$ to denote the distribution that maximizes the third term in proposition 1 and define $P_{\underline{C}}$ similarly. Proposition 1 gives us a recipe for constructing empirical bounds on the true $\mathrm{MMD}(P_{C^*}, P_{X'}, P_{Y'})$. To get an estimate of the upper bound, we need to identify the values of C that render $X' \cup C$ and $Y' \setminus C$ most dissimilar. For a lower bound, we need to identify values of C that render $X' \cup C$ and $Y' \setminus C$ most similar. Unless otherwise noted, we will focus on the analysis of the upper bound of the MMD since the arguments for the lower bound are nearly identical.

We further expand the empirical version of equation 3 to isolate the terms that depend on C, which gives us the empirical objective to optimize. First, we define a weighted version of the empirical witness function,

$$\begin{aligned}\psi(C, X', Y') := \frac{(1-\epsilon)}{n} &\sum_i \sum_j k(x'_i, c_j) \\ - \frac{(1+\epsilon)}{n} &\sum_i \sum_j k(y'_i, c_j) \\ + \frac{\epsilon}{n} &\sum_i \sum_{j\neq i} k(c_i, c_j).\end{aligned}$$

As we show in Lemma A1, in order to estimate $\mathrm{MMD}(P_{\overline{C}}, P_{X'}, P_{Y'})$, we first need to identify $\widehat{C}$:

$$\widehat{C} = \operatorname*{arg\,max}_{C\in Y', \#(C)=m} \psi(C, X', Y'). \tag{4}$$

Note that optimizing ψ under a cardinality constraint in this manner is an NP-hard optimization problem. Instead, we analyze approximation strategies in two regimes: when ϵ can take on any value in [0,1] and when ϵ is sufficiently close to 0. Our analysis relies on analyzing the stability of the estimation algorithms [Bousquet and Elisseeff, 2002].

Approximation strategy for $\epsilon \in [0, 1]$. For any value of ϵ, we can directly maximize equation 4. Noting that: $\max_C \psi(C \in Y', X', Y') \le \max_C \psi(C \in \mathcal{Y}', X', Y')$, we can utilize, for example, iterative optimization algorithms to estimate an approximate $\widehat{C}$. Specifically,

$$\widehat{C}_\circ = \operatorname*{arg\,max}_{C\in\mathcal{Y}', \#(C)=m} \psi(C, X', Y'). \tag{5}$$

The difference between equation 4 and 5 is that 5 can return any value for $\widehat{C}_\circ \in \mathcal{Y}'$, whereas 4 requires that $\widehat{C}_\circ \in Y'$.

While many iterative optimization algorithms can be used to optimize equation 5, we follow Jitkrittum et al. [2016] in using Quasi-Newton methods such as the L-BFGS-B algorithm [Byrd et al., 1995]. For this reason we refer to this iterative optimization approach as the Quasi-Newton optimization **QNO** approach. We stress that our analysis holds for any valid optimization approach.

In proposition 2, we study how fast the estimate based on $\widehat{C}_\circ$ converges to the true upper bound.

Proposition 2 *For* $\mathrm{MMD}(P_{\overline{C}}, P_{X'}, P_{Y'})$ *as defined in proposition 1,* $\widehat{C}_\circ$ *as defined in equation 5, with* $0 \leq k(x', y') \leq \kappa$ *for all* $x', y' \in \mathcal{X}', \mathcal{Y}'$*, we have that:*

$$P_{X',Y'}\Bigg\{|\mathrm{MMD}(P_{\overline{C}}, P_{X'}, P_{Y'}) - \widehat{\mathrm{MMD}}(\widehat{C}_\circ, X', Y')| > b_0 + \varepsilon\Bigg\} \leq 2\exp\left(\frac{-\varepsilon^2 n}{b_1}\right),$$

for $b_0 = 4\sqrt{\kappa}(n^{-1/2} + \epsilon m)$ *and* $b_1 = 2\kappa((1-\epsilon)(1-\epsilon+\epsilon m)^2 + (1+\epsilon)(1+\epsilon+\epsilon m)^2)$.

The proposition shows that the rate of convergence of the empirical MMD defined with respect to $\widehat{C}_\circ$ to the sharp upper bound depends on the sample size, the value of ϵ and the size of the contaminated set m. As ϵ decreases, the estimated $\widehat{\mathrm{MMD}}(\widehat{C}_\circ, X', Y')$ converges faster to its population counterpart $\mathrm{MMD}(P_{\overline{C}}, P_{X'}, P_{Y'})$. At $\epsilon = 0$, we recover the convergence rate of the uncontaminated $\widehat{\mathrm{MMD}}$ (Gretton et al. [2012], theorem 7). As expected, as the sample size increases, the estimate gets closer to its population counterpart. However, the ϵm term in the denominator of the exponent means that the rate of convergence depends unfavorably on the size of the contaminated sample. The next section addresses this.

Approximation strategy for a sufficiently small ϵ. This approach relies on the fact that for a fixed n, and as $\epsilon \to 0$ the third term in equation 4 vanishes.

Specifically for $\epsilon \approx 0$:

$$\psi(C, X', Y') \approx \frac{(1-\epsilon)}{n}\sum_i\sum_j k(x'_i, c_j) - \frac{(1+\epsilon)}{n}\sum_i\sum_j k(y'_i, c_j) = \frac{1}{m}\sum_i \hat{f}'(c_i). \tag{6}$$

where $\hat{f}'$ is a weighted version of the empirical estimate of the witness function definted with respect to the observed contaminated samples.

This means that for ϵ close to 0, maximizing ψ is equivalent to computing the value of the witness function for every sample in Y', and then taking the subset with the highest values to be the estimate of $\widehat{C}$. Consider the following estimate of $\widehat{C}$:

$$\widehat{C}_{\hat{\gamma}} = \{y' : \hat{f}'(y') \geq \hat{\gamma}\} \text{ with } \hat{\gamma} = q(\hat{f}'(Y'), 1-\alpha), \tag{7}$$

where $q(\hat{f}'(Y'), 1-\alpha)$ is defined as the $1-\alpha$ quantile of $\hat{f}'(Y')$. That is, $q(\hat{f}'(Y'), 1-\alpha) = \inf\{\hat{f}'(y') \in \hat{f}'(Y') : (1-\alpha) < \mathrm{CDF}(\hat{f}'(y'))\}$. Equation 7 describes taking the y' samples with weighted witness function values in the top $1-\alpha$ quantile as the candidates for contaminated samples. Next, we show that $\widehat{C}_{\hat{\gamma}}$ is a valid estimate of $\overline{C}$.

Proposition 3 *Let* C_γ *be the solution to equation 7 as* $n \to \infty$*. For a sufficiently small* ϵ*, we have that* $P_{C_\gamma} = P_{\overline{C}}$*, where* $P_{\overline{C}}$ *is defined as the distribution that maximizes the third term in proposition 1.*

While the full proof is stated in the appendix, we find it helpful to highlight the key insight behind proposition 3. The key insight here is that the distribution over C_γ *stochastically dominates* any other distribution over Y' with respect to the transformation $f'(Y')$. Meaning, there exists no other distribution over a subset of Y' with measure α that can give a larger $\mathbb{E}_C[f'(C)]$ than $\mathbb{E}_{C_\gamma}[f'(C_\gamma)]$. We note in passing that this construction extends the classical seminal work by Horowitz and Manski [1995] on estimation of population means using contaminated data to nonparametric estimation of distances between distributions. We refer to this approach as the stochastic dominance (**SD**) approach.

It remains to show that the estimate of the MMD defined with respect to $\widehat{C}_{\hat{\gamma}}$ as estimated using a *finite sample* converges to the true upper bound. We do that in the following proposition.

Proposition 4 *For* $\mathrm{MMD}(P_{\overline{C}}, P_{X'}, P_{Y'})$ *as defined in proposition 1,* $\widehat{C}_{\hat{\gamma}}$ *as defined in equation 7 and* κ *such that* $0 \leq k(x, y) \leq \kappa$ *for all* $x, y \in \mathcal{X}$*. Then as for a sufficiently small* ϵ*:*

$$P_{X',Y'}\Bigg\{|\mathrm{MMD}(P_{\overline{C}}, P_{X'}, P_{Y'}) - \widehat{\mathrm{MMD}}(\widehat{C}_{\hat{\gamma}}, X', Y')| > b_0 + \varepsilon\Bigg\} \leq 2\exp\left(\frac{-\varepsilon^2 n}{b_1}\right)$$

for $b_0 = 4(\kappa/n)^{1/2}(1+\epsilon)$ *and* $b_1 = 2\kappa\big((1-\epsilon)^3 + (1+\epsilon)(1+3\epsilon)^2\big)$.

Proposition 4 shows that unlike QNO, SD avoids the unfavorable dependence on m leading to faster convergence. Similar to proposition 2, at $\epsilon = 0$, we recover the convergence rate of the uncontaminated $\widehat{\mathrm{MMD}}$.

The key advantage of SD over QNO is that it reduces the problem of estimating $\widehat{C}$ to estimating the quantile of the univariate distribution, $P_{f'(Y')}$, which is a single scalar. By contrast, the iterative optimization-based approach needs to identify an $m \times d$ matrix, with d being the dimension of the data. While helpful, the SD approach is limited by the fact that it is a valid approximation only for ϵ sufficiently close to 0. Next, we present our main approach that extends the SD approach making it valid for any value of ϵ.

4 APPROACH

In this section, we describe our main approach to estimating tight and credible upper and lower bounds on the MMD.

Algorithm 1 Our approach (S-SD) for estimating upper bounds

```
Input: $X', Y', \epsilon, S$
$\widehat{C} := \{\}$, $\alpha^{(s)} = \epsilon/(\epsilon + S)$
for $s = 1 \dots S$ do
    $X^{(s)} = X' \cup \widehat{C}, Y^{(s)} = Y' \setminus \widehat{C}$
    Compute $\hat{f}^{(s)}(Y^{(s)})$ as per equation 8
    $\hat{\gamma}_{(1-\epsilon)} = q(\hat{f}^{(s)}(Y^{(s)}), 1 - \alpha^{(s)})$
    $\widehat{C}^s = \{y^{(s)} : \hat{f}^{(s)}(y^{(s)}) \geq \hat{\gamma}_{(1-\epsilon)}\}$
    $\widehat{C} := \widehat{C} \cup \widehat{C}^s$
return $\widehat{\mathrm{MMD}}(\widehat{C}, X', Y')$
```

Algorithm 2 Our approach (S-SD) for estimating lower bounds

```
Input: $X', Y', \epsilon, S$
$\underset{\smile}{C} := \{\}$, $\alpha^{(s)} = \epsilon/(\epsilon + S)$
for $s = 1 \dots S$ do
    $X^{(s)} = X' \cup \underset{\smile}{C}, Y^{(s)} = Y' \setminus \underset{\smile}{C}$
    Compute $\hat{f}^{(s)}(Y^{(s)})$ as per equation 8
    $\hat{\gamma}_{\epsilon} = q(\hat{f}^{(s)}(Y^{(s)}), \alpha^{(s)})$
    $\underset{\smile}{C}^s = \{y^{(s)} : \hat{f}^{(s)}(y^{(s)}) \leq \hat{\gamma}_{\epsilon}\}$

    $\underset{\smile}{C} := \underset{\smile}{C} \cup \underset{\smile}{C}^s$
return $\widehat{\mathrm{MMD}}(\underset{\smile}{C}, X', Y')$
```

Unless otherwise noted, we describe the estimation procedure for constructing the upper bound since the lower bound is nearly identical. Our strategy hinges on identifying $\widehat{C}$, an m-sized subset of Y' which, when removed from Y' and added to X', would render Y' most dissimilar to X', giving us a valid estimate of the the upper bound on the unknown $\widehat{\mathrm{MMD}}(C^*, X', Y')$. Estimating $\widehat{C}$ allows us to estimate $\widehat{\mathrm{MMD}}(\widehat{C}, X', Y')$ in a straightforward manner: we can simply substitute $\widehat{C}$ for C in the empirical version of equation 3.

Our main approach builds upon the SD approach studied in section 3 by addressing its main limitation: that it gives a valid estimate of $\widehat{C}_{\hat{\gamma}}$ only for ϵ sufficiently close to 0. Our approach overcomes this limitation by dividing the task of estimating $\widehat{C}_{\hat{\gamma}}$ into multiple, easier tasks each with an effective $\epsilon^{(s)}$ that is smaller than the true ϵ. Specifically, we divide the estimation process into S steps, in each step we estimate $\widehat{C}^{(s)}_{\hat{\gamma}^{(s)}}$, for $\epsilon^{(s)} = \epsilon/S$. Dividing the estimation into S steps, with each step having ϵ/S-contamination means that each step of the estimation process will have an effective ϵ that is close enough to 0 making equation 7 a valid approximation, and overcoming the main limitation of SD. In the step s of our algorithm, we calculate $\widehat{C}^{(s)}_{\hat{\gamma}^{(s)}} = \{y' \in \widehat{Y}^{(s)} : \hat{f}^{(s)}(\widehat{Y}^{(s)}) \geq \hat{\gamma}^{(s)}\}$, for $\hat{\gamma}^{(s)} = q(\hat{f}^{(s)}(\widehat{Y}^{(s)}), 1 - \alpha^{(s)})$ for $\alpha^{(s)} = \epsilon^{(s)}/(1 + \epsilon^{(s)})$, where

$$\hat{f}^{(s)}(\widehat{Y}^{(s)}) = \Big(1 - \frac{\epsilon}{S}\Big)\frac{1}{n}\sum_i\sum_j k(\hat{x}_i^{(s)}, \hat{y}_j^{(s)}) - \Big(1 + \frac{\epsilon}{S}\Big)\frac{1}{n}\sum_i\sum_j k(\hat{y}_i^{(s)}, \hat{y}_j^{(s)}), \quad (8)$$

with $\widehat{Y}^{(s)} = Y' \backslash \{\widehat{C}^{(1)}_{\hat{\gamma}^{(1)}}, \widehat{C}^{(2)}_{\hat{\gamma}^{(2)}}, \dots \widehat{C}^{(s-1)}_{\hat{\gamma}^{(s-1)}}\}$, and $\widehat{X}^{(s-1)} = X' \cup \{\widehat{C}^{(1)}_{\hat{\gamma}^{(1)}}, \widehat{C}^{(2)}_{\hat{\gamma}^{(2)}}, \dots \widehat{C}^{(s-1)}_{\hat{\gamma}^{(s-1)}}\}$.

We refer to our Stepwise Stochastic Dominance based approach as **S-SD**. We summarize our procedure for estimating the upper and lower bounds in algorithms 1 and 2 respectively. We use $\underset{\smile}{C}$ to denote the counterpart of $\widehat{C}$ defined with respect to the lower bound.

We note that S is a user-specified parameter that takes on values between 0 and m. In section 5.5 we give practical guidance on how to set S. Code for our approach and the experiments in section 5 is available on github.com/mymakar/mmd_uncertainty.

5 EXPERIMENTS

In this section, we (1) analyze the credibility and tightness of our approach and baselines under varying data dimensions, varying sample sizes, and varying values of ϵ. In addition, (2) we examine the computational efficiency of our approach as it compares to baselines. Finally, (3) we examine the sensitivity of our approach to misspecification of ϵ and under varying number of steps S.

To analyze the credibility and the tightness of the bounds estimated using our approach, we compute the False Coverage Rate (FCR) and Mean Interval Width (MIW). For L draws of X', Y' each of size $(1 - \epsilon)n$ and $(1 + \epsilon)n$ respectively, the FCR and the MIW are defined as follows:

$$\begin{aligned} \mathrm{FCR} &= 1 - \frac{1}{L}\sum_i \mathbb{1}\{\widehat{\mathrm{MMD}}(\underset{\smile}{C}, X_i', Y_i') \\ &\quad \leq \widehat{\mathrm{MMD}}(C^*, X_i', Y_i') \leq \widehat{\mathrm{MMD}}(\widehat{C}, X_i', Y_i')\}, \\ \mathrm{MIW} &= \frac{1}{L}\sum_i |\widehat{\mathrm{MMD}}(\underset{\smile}{C}, X', Y') - \widehat{\mathrm{MMD}}(\widehat{C}, X', Y')| \end{aligned}$$

Ablations. We study the following ablations of our approach: **(1) SD**: For $S = 1$, S-SD becomes the same as SD. The performance of SD compared to S-SD highlights the importance of splitting the estimation procedure into S steps. **(2)** Stepwise-QNO (**S-QNO**): Follows the same steps outlined in algorithm 1, however, instead of estimating $\widehat{C}^{(s)}_{\hat{\gamma}}$ and $\underset{\smile}{C}^{(s)}_{\hat{\gamma}}$ as a subroutine, it estimates $\widehat{C}^{(s)}_{\circ}$ and $\underset{\smile}{C}^{(s)}_{\circ}$ following equation 4 using the L-BFGS-B optimization algorithm.

	MIMIC ($N = 100, d = 2$)		FOREST ($N = 100, d = 54$)		BIO ($N = 72, d = 7128$)	
Approach	FCR	MIW	FCR	MIW	FCR	MIW
S-SD (Ours)	$\mathbf{0.0} \pm (\mathbf{0.0})$	$\mathbf{0.137} \pm (\mathbf{0.008})$	$\mathbf{0.0} \pm (\mathbf{0.0})$	$\mathbf{0.088} \pm (\mathbf{0.003})$	$\mathbf{0.1} \pm (\mathbf{0.03})$	$\mathbf{0.075} \pm (\mathbf{0.001})$
S-QNO	$0.08 \pm (0.067)$	$0.119 \pm (0.006)$	$0.02 \pm (0.02)$	$0.084 \pm (0.004)$	$1.0 \pm (0.0)$	$0.059 \pm (0.001)$
QNO	$0.58 \pm (0.069)$	$0.13 \pm (0.006)$	$0.62 \pm (0.069)$	$0.033 \pm (0.006)$	$1.0 \pm (0.0)$	$0.037 \pm (0.001)$
SD	$0.64 \pm (0.068)$	$0.082 \pm (0.01)$	$0.9 \pm (0.042)$	$0.027 \pm (0.005)$	$0.13 \pm (0.034)$	$0.069 \pm (0.001)$
SM	$0.66 \pm (0.067)$	$0.08 \pm (0.01)$	$0.9 \pm (0.042)$	$0.026 \pm (0.004)$	$0.82 \pm (0.038)$	$0.037 \pm (0.001)$
Bootstrap	$0.94 \pm (0.034)$	$0.048 \pm (0.002)$	$0.4 \pm (0.069)$	$0.034 \pm (0.001)$	$0.25 \pm (0.043)$	$0.036 \pm (0.001)$

Table 1: MIW and FCR for all datasets at $\epsilon = 0.2$. Numbers in bold correspond to lowest FCR with smallest MIW. Standard errors (in parentheses) computed by averaging over 100 trials. Results show that our approach performs better than all other approaches when the sample size is small and the dimension is large. In easier settings, our performs comparably to S-QNO.

In each step s, this approach gives an estimate for an m/S subset of candidate contaminated samples. This ablation study highlights the importance of using the SD approach as a subroutine. **(3) QNO**: Similar to S-QNO with $S = 1$.

Baselines. In addition to our main approach and the ablations, we investigate the following baselines: **(1)** Submodular optimization (**SM**): based on the approach suggested in Kim et al. [2016]. It estimates $\widehat{C}$ by converting equation 4 into a submodular function by adding a submodular regularizer. Specifically, it greedily selects samples which maximise the function, $\max_m \hat{f}'(C) + \log\det k(C, C)$, where $\hat{f}'(C)$ is the witness function defined with respect to X' and Y', and $\log\det k(C, C)$ is the log-determinant regularizer. **(2) Bootstrap**: a simple bootstrapping approach, which constructs bounds by resampling both observed groups with replacement and computing the MMD multiple times. The upper and lower bounds are then defined as the $(1 - \alpha)$-th and α quantiles respectively over the distribution of resampled MMD values. The bootstrap estimates are centered around the typical MMD estimate (equation 1), and hence they show how it behaves under ϵ-contamination [1].

For our approach, baselines and ablations, we fix the kernel to be the radial basis kernel (RBF) and use the median heuristic on the contaminated samples to determine bandwidth. Unless otherwise noted, we set the number of steps S for S-SD and S-QNO to be $S = \min(m, 10)$; we take this minimum for when the total number of contaminated samples is less than the total number of steps. We examine the performance of different values of S in section 5.5.

Setup. Since the true value of the contaminated samples C^* is unobserved in real datasets, we resort to semi-simulated data where X, Y represent real data, but the contaminated samples are simulated. We examine the performance of our approach, ablations and baselines in two settings. First, is the nonrandom contamination setting. In this setting, we pick the data points that maximize the difference between the two distributions to be the true contaminated samples. Specifically, we simulate contamination by randomly sampling C^*, a set of size m from the $\min(2m, n)$ samples in X with the largest witness function values, where the witness function here is defined with respect to the uncontaminated X, Y. We then create the observed samples $X' = X \setminus C^*$ and $Y' = Y \cup C^*$. Second, is the random contamination setting, where C^* is sampled at random from X. Since the nonrandom contamination setting is more challenging, we present the results from that setting in the main text. Results from the random contamination setting are presented in the appendix. We define $N = \#(X) + \#(Y)$, the total number of samples, and consider 3 tasks corresponding to 3 datasets:

1. **FOREST**: A publicly available dataset containing measurements of 54 cartographic variables such as elevation and slope [Blackard, 1998]. We consider the task of measuring the distance between the distribution over cartographic properties of two forest types: Lodgepole Pine and Spruce-Fir. We simulate ϵ contamination by flipping an ϵ proportion of Lodgepole Pine ($n = 283, 301$) labels to Spruce-Fir ($n = 211, 840$).

2. **MIMIC**: A publicly available chest radiographs and corresponding clinical data with over 377,000 chest X-ray images and radiology reports [Johnson et al., 2019a,b, Goldberger et al., 2000]. Here, we consider the task of measuring the distance between pneumonia predictions across two race groups – a common task in the fairness literature. In this setting, the sensitive attribute is measured with ϵ-contamination. We use 60% of the data for training the model, 20% for validation, and the remaining 20% for MMD estimation. We use the training and validation data to fine tune a Densenet-121 [Huang et al., 2016] that was pretrained on Imagenet [Deng et al., 2009]. After training the model, we obtain the 2-dimensional logit predictions of the 20% of the data held out for MMD estimation, and simulate ϵ-contamination by changing an ϵ proportion of Black ($n = 3897$) patients to White ($n = 11293$).

3. **BIO**: Unlike the 2-dimensional MIMIC data and 54-dimensional FOREST data, in the third task we examine a more extreme case of high dimensional data with few samples. We use publicly available leukemia gene expression

[1] In the appendix, we explicitly show how the typical estimate of the MMD behaves with varying ϵ

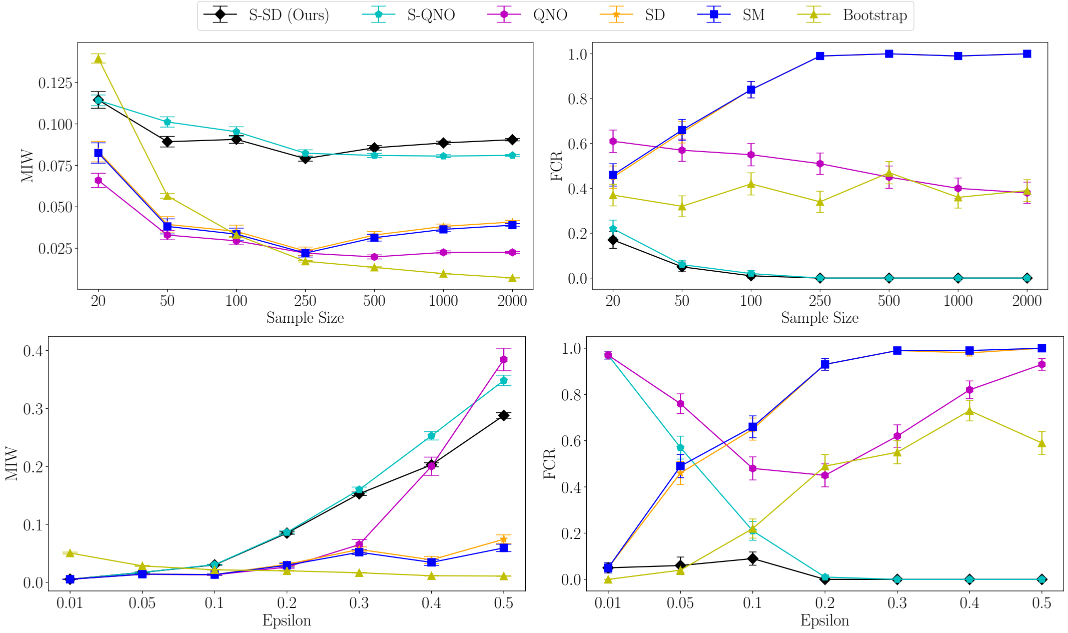

Figure 1: **Top**: Results on FOREST fixing $\epsilon = 0.2$ and increasing sample size from $N = 20$ to $N = 2000$. Bars indicate the SE of the FCR and MIW across all trials. As sample size increases, MIW decreases for all methods, with S-SD providing intervals with the lowest FCR for all sample sizes. **Bottom**: The MIW and FCR for each approach is shown as the intensity of ϵ-contamination varies from $\epsilon = 0.01$ to $\epsilon = 0.5$ in FOREST $(N = 100, d = 54)$. Bars indicate the SE of the FCR and MIW across all trials. As ϵ increases, S-SD reports tight and credible intervals for all values of ϵ.

dataset (BIO) [Golub et al., 1999], which has 7128 measurements of gene expressions from DNA microarrays for 72 samples. The 72 samples are divided into binary groups of leukemia cancer cell types, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), and we conduct the ϵ contamination by flipping ϵ of the ALL $(n = 47)$ to AML $(n = 25)$.

5.1 PERFORMANCE UNDER DIFFERENT DATA DIMENSIONS

In this section, we examine the effect of varying dimension. To do so, we compute the FCR and MIW of bounds estimated on MIMIC $(N = 100, d = 2)$, FOREST $(N = 100, d = 54)$, and BIO $(N = 72, d = 7128)$ in table 1. We focus on the small sample regime as it is much more challenging. To get estimates for the standard error (SE) around the MIW and FCR, we repeat the experiment 100 times on 100 samples picked without replacement for MIMIC and FOREST. For BIO, we create 100 bootstrap samples. We fix $\epsilon = 0.2$, simulate contamination in 100 random samples, and calculate the upper and lower bounds for each approach.

The results in table 1 show that in all settings our approach gives the tightest (smallest MIW) and most credible (lowest FCR) estimates, while SD, QNO and S-QNO return bounds with a higher FCR. In settings where the dimensions are small, S-QNO performs significantly better than QNO. However, both perform poorly when the dimension, d is large. Such a finding makes sense: the stepwise algorithm reduces the dependence on the sample size, however the performance of both QNO and S-QNO appears to have some irreducible dependence on the dimension. This is not surprising, in BIO, for example, S-QNO is solving an optimization problem over an $m/S \times 7128$ parameter space, whereas S-SD is required to estimate the $(1-\alpha)/S$ quantile of a univariate distribution (that is the distribution over the values of the witness function). In this setting where $\epsilon = 0.2$, equation 7 is a poor approximation of equation 4, which explains the poor performance of SD. At $\epsilon = 0.2$ the typical estimate of the MMD (equation 1) is unreliable. Being centered around the typical estimate, Bootstrap is expected to give unreliable bounds. SM also performs poorly since it is designed to find *few* samples that explain the difference between the two corrupted distributions.

Overall, S-SD remains robust even in high dimensions, while other approaches do not. In the appendix, we repeat this experiment with $N = 2000$ for MIMIC and FOREST. The results are largely consistent with the findings presented here. However, as N increases, the estimates for S-QNO in small dimensions become more comparable to S-SD.

For brevity, we present results on the FOREST dataset in the main text but include the similar analyses on MIMIC and BIO in the appendix.

5.2 PERFORMANCE UNDER DIFFERENT SAMPLE SIZES

We study the effect of increasing sample size. Fixing $\epsilon = 0.2$, we vary the sample size from $N = 20$ to $N = 2000$ by sampling from the FOREST dataset. For each sample size, we sample 100 times and compute the mean FCR and MIW and their corresponding standard errors. We plot the results for the MIW in figure 1 (top, left) and the FCR in figure 1 (top, right). The results show that the FCR for our approach, S-QNO and QNO decreases as the sample size increases revealing that these estimates are consistent. However, our approach gives the lowest FCR even in very small samples. In larger samples, S-QNO performs comparably to our approach. SD, SM and the bootstrap method all return overly conservative estimates that do not contain the true MMD.

5.3 PERFORMANCE UNDER DIFFERENT VALUES OF ϵ

We investigate the effect of increasing contamination from $\epsilon = 0.01$ to $\epsilon = 0.9$. Similar to section 5.1, we focus on the small sample regime by fixing N to be 100. We present the results here up to $\epsilon = 0.5$, and the rest in the appendix.

Figure 1 (bottom) shows that for small values of ϵ, QNO and S-QNO perform poorly, giving high FCR. S-QNO resolves some of the issues by dividing the optimization into several steps, but still underperforms compared to our approach. SD gives a biased estimate of the bound for ϵ significantly higher than 0, as expected. Bootstrap gives valid bounds with low FCR only with near negligable values of ϵ, where the typical MMD estimate is approximately valid.

The previous three experiments show that S-SD consistently gives credible and tight estimates of the upper and lower bounds on the value of the true MMD. Next, we examine the sensitivity of S-SD to the number of steps S.

5.4 COMPUTATIONAL EFFICIENCY

Next, we examine the computational efficiency of our approach as compared to baselines. Using the Forest dataset, with $n = 2000$, we vary the value of ϵ and measure the time in seconds that is required for each model to compute the upper and lower bounds. We repeat the experiment 100 times and report mean time and standard errors.

The results, shown in figure 2, indicate that our basic (non-stepwise) approach is the fastest, and particularly it is faster than the non-stepwise QNO approach while our main approach S-SD is faster than its counterpart, S-QNO. Importantly, the plot implies that increasing the value of ϵ has a negligible effect on the computational time of SD and our main approach S-SD.

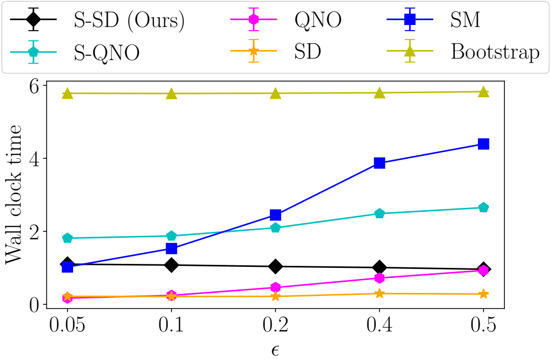

Figure 2: Computational efficiency: x-axis shows the value of ϵ, y-axis shows the wall clock time in seconds. Our basic (non-stepwise) approach

5.5 SENSITIVITY ANALYSES

Sensitivity to incorrect values of ϵ. Next, we examine the sensitivity of S-SD and other methods to incorrect values of ϵ. To do so, we fix the true value of ϵ to be 0.1 but sweep $\tilde{\epsilon}$, the value given to each of the models, from 0.01 to 0.5. This means that the assumption of known/correct level of noise is only satisfied when $\tilde{\epsilon} = \epsilon = 0.1$. Similar to section 5.1, we focus on the small sample regime by fixing N to be 100.

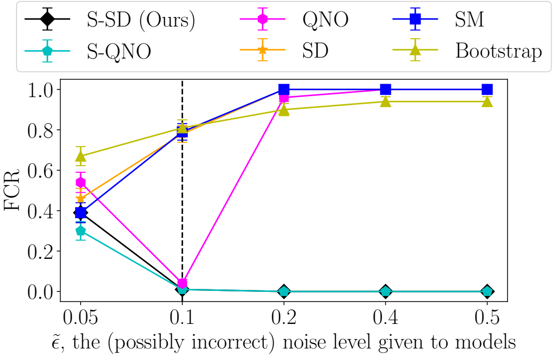

Figure 3: Sensitivity to incorrect values of ϵ. True value of ϵ is 0.1. Models are given $\tilde{\epsilon}$ values shown on the x-axis

Figure 3 shows the value of $\tilde{\epsilon}$ on the x-axis and the corresponding FCR on the y-axis. In the appendix, we present a plot with the corrsponding mean interval widths for each level of $\tilde{\epsilon}$. The results show that only our two approaches (S-SD and S-QNO) achieve an FCR of 0 whenever $\tilde{\epsilon} \geq \epsilon$. This suggests a practical guideline: when in doubt, users should err on the side of a higher ϵ estimate with the trade-off of wider intervals (as reported in the appendix). Other methods do not give such a guarantee: they consistently give overly conservative intervals with poor coverage.

Sensivity to the number of steps. We examine the sensitivity of S-SD to the number of steps S. To do so, we sample

	S-SD (Ours)	
No. of Steps	FCR	MIW
2	$0.21 \pm (0.091)$	$0.082 \pm (0.001)$
3	$0.13 \pm (0.034)$	$0.079 \pm (0.001)$
5	$0.0 \pm (0.0)$	$0.088 \pm (0.001)$
10	$0.0 \pm (0.0)$	$0.08 \pm (0.001)$
20	$0.0 \pm (0.0)$	$0.091 \pm (0.001)$
50	$0.0 \pm (0.0)$	$0.091 \pm (0.001)$

Table 2: Varying number of steps for S-SD in FOREST ($N = 2000, d = 54$) with $\epsilon = 0.2$. Standard errors (in parentheses) over 100 trials. Results imply that setting S to be large gives lower FCR.

$n = 2000$ from FOREST, vary the value of S, and examine the performance of our main approach, S-SD. We repeat the experiment 100 times using 100 different samples from FOREST, each of size 2000 to compute the standard errors around the FCR and MIW.

Table 2 shows the results. The results imply that we can get bound estimates that give a FCR of zero even with a very few number of steps. The MIW increase slightly and starts to plateau as the number of steps increases. This implies that a reasonable choice of S to ensure a low FCR would be the largest possible value which does not lead to a computationally prohibitive number of iterations. Recall that there is a natural upper bound on $S = m$. In the appendix, we repeat this experiment for S-QNO showing similar robustness.

6 CONCLUSION

We studied the problem of comparing two distributions when the data is collected with some measurement error. Specifically, we showed that typical estimates of kernel based distances are unreliable when the data is measured with some ϵ contamination, where an ϵ proportion of one sample is erroneously included with the other. We showed both empirically and theoretically that a straightforward optimization approach to measuring uncertainty has an unfavorable dependence on the size of the contaminated set. Instead, we proposed a stepwise approach to estimate credible and tight upper and lower bounds and showed that it converges faster than alternatives to the true upper and lower bounds. Empirically, we showed that our approach outperforms all baselines. Looking beyond this work, it would be interesting to study other commonly occurring measurement error mechanisms and study their effect on measuring the MMD and other related estimates such as the Hilbert Schmidt independence criterion.

Extensions of this work. While beyond the scope of this work, it might be interesting to understand how our suggested approaches can be used in the context of hypothesis testing, where the goal is to formally test if the two distributions are similar. We note that such a test can be done by combining approaches for hypothesis testing using "interval test statistics" (see Kreinovich et al. [2008] for a summary) with approaches for acquiring empirical estimates of the MMD under the null distribution Gretton et al. [2009].

We also note that extending our approach to settings where both variables are contaminated is likely a trivial extension of our work. Specifically, it might be appropriate to conduct an iterative procedure where we find $\widehat{C}_x$: the samples observed in Y' that are truly sampled from P_X and then find $\widehat{C}_y$ the samples observed in X' that are truly sampled from P_Y iteratively until meeting some convergence criteria.

Acknowledgements

We are thankful for feedback from anonymous reviewers. The work was done when Ron Nafshi was a student at the University of Michigan. This material is based upon work supported by the National Science Foundation under Grant No.2153083 and 2337529. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: `10.24432/C50K5N`.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2: 499–526, 2002.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16 (5):1190–1208, 1995.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Carles Domingo-Enrich, Raaz Dwivedi, and Lester Mackey. Compress then test: Powerful kernel testing in near-linear time. *arXiv preprint arXiv:2301.05974*, 2023.

Tamara Fernández and Nicolás Rivera. A reproducing kernel hilbert space log-rank test for the two-sample problem. *Scandinavian Journal of Statistics*, 48(4):1384–1432, 2021.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Mark Goldstein, Jörn-Henrik Jacobsen, Olina Chau, Adriel Saporta, Aahlad Manas Puli, Rajesh Ranganath, and Andrew Miller. Learning invariant representations with missing data. In *Conference on Causal Learning and Reasoning*, pages 290–301. PMLR, 2022.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999. doi: 10.1126/science.286.5439.531.

Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. *Advances in neural information processing systems*, 22, 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Joel L Horowitz and Charles F Manski. Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302, 1995.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL `http://arxiv.org/abs/1608.06993`.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29, 2016.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr database, Sep 2019a.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, Dec 2019b.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3): 1959–1981, 2022.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally robust bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2174–2184. PMLR, 2020.

Vladik Kreinovich, Hung T Nguyen, and Sa-aat Niwitpong. Statistical hypothesis testing under interval uncertainty: An overview. *International Journal of Intelligent Technologies and Applied Statistics*, 1(1):1–32, 2008.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Maggie Makar and Alexander D'Amour. Fairness and robustness in anti-causal prediction. *arXiv preprint arXiv:2209.09423*, 2022.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

C Noel Bairey Merz. The yentl syndrome is alive and well, 2011.

Luca Oneto, Michele Donini, Giulia Luise, Carlo Ciliberto, Andreas Maurer, and Massimiliano Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Advances in Neural Information Processing Systems*, 33:15360–15370, 2020.

Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176 (4):928–943, 2019.

Antonin Schrab, Ilmun Kim, Benjamin Guedj, and Arthur Gretton. Efficient aggregated kernel tests using incomplete u-statistics. In *Advances in Neural Information Processing Systems*.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Aad W Van Der Vaart, Jon A Wellner, Aad W van der Vaart, and Jon A Wellner. *Weak convergence*. Springer, 1996.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208, 2021.