Counterfactual Decision Support Under Treatment-Conditional Outcome Measurement Error

Growing work in algorithmic decision support proposes methods for combining predictive models with human judgment to improve decision quality. A challenge that arises in this setting is predicting the risk of a decision-relevant target outcome under multiple candidate actions. While counterfactual prediction techniques have been developed for these tasks, current approaches do not account for measurement error in observed labels. This is a key limitation because in many domains, observed labels (e.g., medical diagnoses, defendant re-arrest) serve as a proxy for the target outcome of interest (e.g., biological medical outcomes, recidivism). We develop a method for counterfactual prediction of target outcomes observed under treatment-conditional outcome measurement error (TC-OME). Our method minimizes risk with respect to target potential outcomes given access to observational data and estimates of measurement error parameters. We also develop a method for estimating error parameters in cases where these are unknown in advance. Through a synthetic evaluation, we show that our approach achieves performance parity with an oracle model when measurement error parameters are known and retains performance given moderate bias in error parameter estimates.

1 Introduction

Predictive models are increasingly being introduced to support expert decision-making in real-world tasks. In the medical domain, clinical models have been developed to inform patient treatment decisions by predicting the likelihood of adverse health outcomes (e.g., heart attack, stroke) [Mullainathan and Obermeyer 2022]. In the criminal justice domain, risk assessments have been introduced to inform judicial pre-trial release decisions by predicting the likelihood of a defendant recidivating given release on bail. In these settings, policy makers often wish to estimate the risk of a downstream target outcome under multiple alternative actions [Schulam and Saria 2017].

However, because outcomes were observed under the past decision-making policy, we do not observe the counterfactual outcome that would under a different decision [Pearl 2009]. This makes it challenging to learn a counterfactual model given observational data. Recent work proposes counterfactual modeling and evaluation approaches designed for decision-support settings Coston et al. [2020b,a]. This work builds upon causal inference methods for conditional average treatment effect (CATE) estimation [Abrevaya et al. 2015], which have received growing interest within the machine learning community [Johansson et al. 2018]. We consider counterfactual modeling in two real-world decision support tasks. In selective intervention tasks, a model estimates baseline risk under no intervention (e.g., risk of child neglect given no welfare services [Chouldechova et al. 2018], or poor learning outcomes given no tutoring). In selective opportunity settings, a model estimates risk of a target outcome under a proposed opportunity (e.g., likelihood default given receiving a loan, or performance ratings given a new job). We study these counterfactual prediction settings where the target outcome is subject to measurement error.

Outcome measurement error occurs when policy-relevant target outcomes (e.g., heart attack, defendant recidivism) are imperfectly approximated by proxies (e.g., heart attack diagnosis, defendant re-arrest) in data [Jacobs and Wallach, 2021]. While measurement error and label noise have been studied in previous literature [Menon et al., 2015; Natarajan et al., 2013; Wang et al., 2021; Fogliato et al., 2020; 2021], this challenge has not been addressed in counterfactual prediction settings. Therefore, in this work, we study the novel setting of counterfactual prediction in the presence of treatment-conditional outcome measurement error (TC-OME). We offer the following contributions:

- We formalize the problem of counterfactual prediction under treatment-conditional outcome measurement error.
- We develop a risk minimization approach (FRM-SL; Algorithm 1) for estimating target potential outcomes given observed covariates, past decisions, and proxy outcomes impacted by measurement error. Our method combines covariate adjustment techniques designed for CATE inference [Johansson et al., 2020] with a surrogate loss developed by [Natarajan et al., 2013] for label noise correction.
- We develop a method for estimating treatment-conditional measurement error parameters (CCPE; Algorithm 2). Our approach builds on class probability estimation (CPE) techniques designed for label noise settings [Menon et al., 2015; Scott et al., 2013].
- We evaluate FRM-SL and CCPE via synthetic experiments and show that FRM-SL achieves performance parity with an oracle model given access to ground-truth error parameters. We also show that FRM-SL performance is tolerant to moderate bias in CCPE parameter estimates, and show that bias in estimates decreases as a function of sample size.

2 Problem setup

We consider a counterfactual distribution defined over \( p^*(X, D, Y_0^*, Y_1^*) \), where \( X \in \mathcal{X} \subseteq \mathbb{R}^d \) are covariates, \( D \in \{0, 1\} \) are past decisions, and \( Y_0^*, Y_1^* \in \{0, 1\} \) are target binary potential outcomes of interest to human decision-makers. Under potential outcomes [Rubin, 2005], \( Y_1^* \) is the hypothetical outcome we would see under the baseline condition when \( D := 0 \), while \( Y_1^* \) is the outcome we would observe under the proposed intervention (selective intervention) or opportunity (selective opportunity) when \( D := 1 \). Following the standard setup in causal inference, we only observe \( Y_0^* \) or \( Y_1^* \) for a given instance such that \( Y^* = D \cdot Y_1^* + (1 - D) \cdot Y_0^* \) [Pearl, 2009].

Given the counterfactual joint \( p^* \), we would like to estimate the target quantity:

\[
\eta_d^*(x) := \mathbb{P}(Y^*_d = 1 \mid X = x), \forall x \in X. \tag{1}
\]

where \( d = 0 \) in selective intervention settings, and \( d = 1 \) in selective opportunity settings.

However, rather than sampling directly from \( p^* \), we draw samples i.i.d. from \( p(X, D, Y) \), where \( Y \in \{0, 1\} \) is a binary proxy outcome. Two challenges complicate estimation of \( \eta_d^*(x) \) given samples drawn from \( p \). First, observed proxies \( Y \) arise from potential outcomes such that \( Y = D \cdot Y_1 + (1 - D) \cdot Y_0 \). Therefore, we do not know the outcome that \textit{would have} occurred had the counterfactual decision been made in the past. Second, proxy potential outcomes \( Y_d \) are subject to outcome measurement error. In our \textit{treatment-conditional outcome measurement error} (TC-OME) model, the proxy potential outcome is observed under a false positive rate \( \alpha_d \) and false negative rate \( \beta_d \) given by:

\[
\alpha_d := \mathbb{P}(Y_d = 1 \mid Y_d^* = 0), \beta_d := \mathbb{P}(Y_d = 0 \mid Y_d^* = 1), \forall d \in D \tag{2}
\]

where \( \alpha_d + \beta_d < 1 \). Under this model, the class probability function of the proxy potential outcome \( Y_d \) can be given by:

\[
\mathbb{P}(Y_d = 1 \mid X = x) = \mathbb{P}(Y_d = 1 \mid Y_d^* = 1) \cdot \mathbb{P}(Y_d^* = 1 \mid X = x)
+ \mathbb{P}(Y_d = 1 \mid Y_d^* = 0) \cdot \mathbb{P}(Y_d^* = 0 \mid X = x)
= (1 - \beta_d) \eta_d^*(x) + \alpha_d (1 - \eta_d^*(x)), \forall x \in X, d \in D. \tag{3}
\]
While this assumption follows from the class-conditional model studied in past literature [Menon et al., 2015, Scott et al., 2013], our methods can be readily extended to settings where error parameters not depend on covariates \( X \) or unmeasured confounders \( Z \): \( \mathbb{P}(Y_d \mid Y^*_d) = \mathbb{P}(Y_d \mid Y^*_d, X = x) = \mathbb{P}(Y_d \mid Y^*_d, Z = z) \).

2.1 Identifiability conditions

Figure 1 shows a causal diagram specifying the assumptions we make on the data generating process in a TC-OME setting. Class probability functions \( \eta_d \) and \( \eta_d \) are identifiable if they can be computed uniquely from the observed distribution \( p(X, D, Y) \). Our target causal estimand \( \eta_d(x) \) is not identifiable directly because potential outcomes \( \{Y^*_0, Y^*_1\} \) are unobserved. However, \( \{Y_0, Y_1\} \) are identifiable under a standard set of causal identifiability assumptions:

Assumption 2.1 (Measurement Error Model). Error parameters not depend on covariates \( X \) or unmeasured confounders \( Z \): \( \mathbb{P}(Y_d \mid Y^*_d) = \mathbb{P}(Y_d \mid Y^*_d, X = x) = \mathbb{P}(Y_d \mid Y^*_d, Z = z) \).

Assumption 2.2 (Consistency). An instance receiving decision \( d \in \{0, 1\} \) has outcome \( Y = Y_d \): \( Y = D \cdot Y_1 + (1 - D) \cdot Y_0 \).

Assumption 2.3 (Ignorability). Potential outcomes and decisions are conditionally independent given \( X \): \( \{Y_0, Y_1\} \perp \!\!\!\!\perp D \mid X \).

Assumption 2.4 (Positivity). For any set of covariates \( x \in X \), both decisions have non-zero probability of observation in the data: \( \forall x \in X, d \in \{0, 1\} : p(D = d | X = x) > 0 \).

3 Methodology

First, we develop an estimator for \( \eta_d(x) \) given i.i.d. samples drawn from the observational joint \( p(X, D, Y) \) assuming a priori knowledge of error terms (Section 3.1). In practice, \( \alpha_d \) and \( \beta_d \) are unknown in advance. Therefore, in Section 3.2 we develop an approach for error parameter estimation (Algorithm 2). This approach requires access to observational data from \( p \) and an additional weak separability assumption commonly applied in class-conditional error settings [Menon et al., 2015].

3.1 Risk minimization

Let \( f \in \mathcal{H} \) be a probabilistic decision function belonging to \( \mathcal{H} \subset \{h : \mathcal{X} \to [0, 1]\} \) and let \( \ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}_+ \) be a bounded loss function. The \( \ell \)-risk of \( f \) over \( Y^*_d \) can be given by:

\[
R^*_\ell(f) := \mathbb{E}_{X, Y^*_d}[\ell(f(X), Y^*_d)]
\] (4)

where \( R^*_\ell(f) \) is the marginal risk over the full population. Because \( \eta_d(x) \) is probabilistic, we restrict \( \ell \) to be strictly proper composite such that arg min \( R^*_\ell(f) \) is a monotone transform \( \phi \) of \( \eta_d \) (e.g., the logistic and exponential loss) [Agarwal, 2014, Menon et al., 2015]. This allows for recovering class probabilities from the optimal prediction via the link function \( \phi \).
Johansson et al. [2020] show that the marginal risk over the full population can be decomposed into factual and counterfactual components scaled by $\pi = p(D = 1)$ via:

$$R^*_f(f) := \pi R^*_{d,e}(f) + (1 - \pi) R^*_{1-d,e}(f),$$ (5)

for $d \in \{0, 1\}$. The factual risk, denoted by subscript $d$, can be directly estimated over the sample that received treatment $D = d$. Under ignorability, this factual risk is identifiable as:

$$R^*_{d,e}(f) := \mathbb{E}_{X,Y_d}[\ell(f(X), Y_d^*)] = \mathbb{E}_{X,Y^*|D}[\ell(f(X), Y^*)|D = d]$$ (6)

In theory, the factual risk $R^*_{d,e}(f)$ is a biased estimator for population risk $R^*_f(f)$ in observational settings (i.e., when $X \not \perp D$) [Johansson et al. 2020]. However, previous work has demonstrated empirically that bias correction techniques such as propensity re-weighting offers limited performance improvements in counterfactual risk assessment settings given sufficient sample size and an expressive model class [Coston et al. 2020b]. Therefore, in this work, we develop a minimizer for the factual risk $R^*_{d,e}(f)$ and leave a comparison with re-weighting based approaches for future work.

The factual risk $R^*_{d,e}(f)$ cannot be estimated directly because target potential outcomes $Y^*_d$ are unobserved. We address this challenge by developing a surrogate loss $\hat{\ell}$ such that minimizing factual $\hat{\ell}$-risk w.r.t. proxy outcomes $Y$ is equivalent to minimizing factual $\ell$-risk w.r.t. target outcomes $Y^*$ [Natarajan et al. 2013]. Under ignorability, the factual risk over the proxy potential outcome $Y_d$ is identifiable by conditioning on $D = d$ via:

$$R_{d,e}(f) := \mathbb{E}_{X,Y_d}[\ell(f(X), Y_d)] = \mathbb{E}_{X,Y^*|D}[\ell(f(X), Y^*)|D = d]$$ (7)

Therefore, given knowledge of error parameters, we wish to construct a surrogate loss $\hat{\ell}$ s.t. $R_{d,e}(f)$ gives an unbiased estimator for $R^*_{d,e}(f)$:

$$\mathbb{E}_{Y^*=y^*,D=d}[\hat{\ell}(f(x), Y)] = \ell(f(x), y^*)$$ (8)

By Lemma 1 in [Natarajan et al. 2013], such an $\ell$ can be constructed via:

$$\hat{\ell}(f(x), +1) := \frac{(1 - \alpha_d) \ell(f(x), +1) - \beta_d \ell(f(x), -1)}{1 - \beta_d - \alpha_d}$$
$$\hat{\ell}(f(x), -1) := \frac{(1 - \beta_d) \ell(f(x), -1) - \alpha_d \ell(f(x), +1)}{1 - \beta_d - \alpha_d}$$ (9)

By Lemma 1 in [Natarajan et al. 2013], e.q. [11] converges to the factual risk of the target potential outcome $R^*_{d,e}(f)$ in expectation. Under the condition that $\ell$ is strictly proper composite, $R^*_f(f)$ can be used to recover desired class probabilities $\eta^*_f(x)$. Therefore, given a priori knowledge of $\alpha_d$, $\beta_d$, we can learn an estimator for $\eta^*_f(x)$ given samples from $p(X, D, Y)$. 

By Lemma 1 in [Natarajan et al. 2013], e.q. [11] converges to the factual risk of the target potential outcome $R^*_{d,e}(f)$ in expectation. Under the condition that $\ell$ is strictly proper composite, $R^*_f(f)$ can be used to recover desired class probabilities $\eta^*_f(x)$. Therefore, given a priori knowledge of $\alpha_d$, $\beta_d$, we can learn an estimator for $\eta^*_f(x)$ given samples from $p(X, D, Y)$. 

\[\eta^*_f(x) \]
Learn 3 Construct 2

Compute parameter estimates (0 counterfactual [Xia et al., 2019, Wang et al., 2021]. In our (Weak Separability)
Algorithm 1:
(ii) factual risk minimization with unmodified loss (FRM); and (iii) factual risk minimization with
configurations with (1) no measurement error [OBS performs poorly across all configurations. FRM and oracle methods perform comparably in
predicting outcomes [Pearl, 2009, Coston et al., 2020b] and (2) with symmetric error terms
(0.2, 0.2). In the former case, a surrogate loss is not needed to correct for measurement error. In the

3.2 Error parameter estimation

Directly minimizing \( \hat{R}_{d,\ell}(f) \) is challenging because error parameters are often unknown in advance.
Therefore, we develop a procedure for estimating error terms, then use estimates \( \hat{\alpha}_d, \hat{\beta}_d \) to construct
the surrogate loss. Our parameter estimation approach places a weak separability assumption on \( p^* \):

**Assumption 3.1 (Weak Separability).** Under the proposed decision \( d \), there exist target potential
outcomes \( Y_d^* \) such that \( \inf_{x \in X} \{ \eta_d^*(x) \} = 0 \) and \( \sup_{x \in X} \{ \eta_d^*(x) \} = 1 \).

This assumption has been widely applied in observational label noise settings [Menon et al., 2015
[Xia et al., 2019, Wang et al., 2021]. In our counterfactual setting, this assumption stipulates that there
exists an individual at no risk (\( Y_d^* = 0 \)) and another individual at certain risk (\( Y_d^* = 1 \)) under \( d = 0 \)
in selective intervention settings or \( d = 1 \) in selective opportunity settings. This assumption need
only hold for \( d = 0 \) or \( d = 1 \) depending on the setting.

Given weak separability, error parameters can be estimated by substituting \( \eta_d^*(x) = 0, \eta_d^*(x) = 1 \)
into the TC-OME model (e.g. 3) and solving:

\[
\begin{align*}
\eta_d^*(x) = 0 & \implies \alpha_d = \inf_{x \in X} \eta_d(x) \\
\eta_d^*(x) = 1 & \implies \beta_d = 1 - \sup_{x \in X} \eta_d(x)
\end{align*}
\]

(12)

where we can take the infimum and supremum of \( \eta_d(x) \) because it is a monotone transform of
\( \eta_d^*(x) \) by e.q. 3. Therefore, we estimate \( \alpha_d, \beta_d \) by learning \( \hat{\eta}_d(x) \), then computing the minimum
and maximum over class probabilities predicted on a held-out sample (Algorithm 2). For statistical
purposes, we perform each step on different data folds [Menon et al., 2015]. Estimates \( \hat{\alpha}, \hat{\beta} \) can then
be used to construct \( \hat{\ell} \) and minimize e.q. 11 to learn \( \eta_d^*(x) \) (Algorithm 1).

4 Experiments

**Setup.** TC-OME evaluation on real-world data is challenging due to compounding uncertainty
from (1) unobserved potential outcomes [Pearl, 2009, Coston et al., 2020b] and (2) measurement
[De-Arteaga et al., 2021, Fogliato et al., 2021]. Therefore, we conduct a controlled synthetic
evaluation to validate our approach. Our evaluation emulates a selective intervention setting with
target quantity \( q_0(x) \). We use a unidimensional covariate \( X \sim U \), and sinusoidal functions
\( q_0(x), q_1(x) \) satisfying weak separability. We sample \( Y_0^*, Y_1^* \sim \text{Bern}(q_0^*(x)) \), \forall \in D \) and generate
proxy outcomes by flipping \( Y_0^*, Y_1^* \) with probability defined by \( \alpha_d, \beta_d \). We observe outcomes \( Y \)
by sampling from a propensity function \( \pi(x) = \mathbb{P}(D = 1 \mid X = x) \) that is linear in \( X \). We use an MLP
trained via binary cross-entropy loss in all experiments. Appendix A contains additional details.

**Experiments.** Experiment one (Table 1) compares (i) an observational model targeting \( Y \) (OBS);
(ii) factual risk minimization with unmodified loss (FRM); and (iii) factual risk minimization with
a surrogate loss parameterized by \( \alpha_0, \beta_0 \) (FRM-SL, Algorithm 1). We compare against oracles
predicting \( Y_0^* \) (TARGET*) and \( Y_0^* \) (PROXY*). We report accuracy with respect to \( Y_0^* \) on a held-
out sample in line with a selective intervention setting targeting \( q_0^*(x) \). As shown in Table 1
OBS performs poorly across all configurations. FRM and oracle methods perform comparably in
configurations with (1) no measurement error (\( \alpha_0 = 0, \beta_0 = 0 \)) and (2) with symmetric error terms
(0.2, 0.2). In the former case, a surrogate loss is not needed to correct for measurement error. In the

**Algorithm 1: Factual risk minimization with surrogate loss (FRM-SL)**

**Input:** Data \( W \sim p \)

**Output:** Learned estimator for \( \hat{\eta}_d^*(x) \)

1. Compute parameter estimates \( \hat{\alpha}_d, \hat{\beta}_d \) by e.q. 3. Therefore, we estimate \( \alpha_d, \beta_d \) by learning \( \hat{\eta}_d(x) \), then computing the minimum
2. Construct \( \ell \) parameterized by \( \hat{\alpha}_d, \hat{\beta}_d \)
3. Learn \( \hat{\eta}_d(x) := \arg\min_{f \in H} \hat{R}_{d,\ell}(f) \)

**Algorithm 2: Conditional class probability estimation (CCPE)**

**Input:** Data \( W = \{X_i, D_i, Y_i\}_{i \in N} \sim p \)

**Output:** Parameter estimates \( \hat{\alpha}_d, \hat{\beta}_d \)

1. Partition \( W \) into subsets \( W_1, W_2 \)
2. Learn \( \hat{\eta}_d(x) := \arg\min_{f \in H} \hat{R}_{d,\ell}(f) \) on \( W_1 \)
3. Use \( \hat{\eta}_d(x) \) to estimate error terms on \( W_2 \):
   \[ \hat{\alpha}_d = \min_{x \in X} \{ \hat{\eta}_d(x) \}, \hat{\beta}_d = 1 - \max_{x \in X} \{ \hat{\eta}_d(x) \} \]
We introduce a novel treatment-conditional outcome measurement error model that formalizes key challenges in counterfactual prediction settings. We provide a risk minimization (Algorithm 1) and parameter estimation (Algorithm 2) techniques designed for this setting. Synthetic results demonstrate that FRM+SL provides a strong improvement over FRM alone in asymmetric error settings. Results also show that FRM+SL performance remains robust given bias in \( \hat{\alpha}_d, \hat{\beta}_0 \) in the neighborhood of \( \pm 0.05 \) to \( \pm 1 \) depending magnitude of \( \alpha_0, \beta_0 \).

Experiment two (Figure 2) evaluates CCPE as a function of sample size. We construct \( \pi(x) \) and \( \eta_0(x) \) such that \( \pi(x) \propto \eta_0(x) \). This mirrors the real-world setting in which individuals at high baseline risk are more likely to receive a risk-reducing intervention. Because this results in lower sample density over the high risk region of \( \eta_0(x) \), the learned approximation \( \hat{\eta}_0(x) \) is likely to be worse near the supremum of \( \eta_0(x) \) than at its infimum. As a result, we expect more bias in \( \hat{\beta}_d \) than \( \hat{\alpha}_d \). Figure 2 shows that estimates improve as \( N \) increases, and confirms slower convergence for \( \hat{\beta}_d \) than \( \hat{\alpha}_d \).

## 5 Conclusion

We introduce a novel treatment-conditional outcome measurement error model that formalizes key challenges in counterfactual prediction settings. We provide a risk minimization (Algorithm 1) and parameter estimation (Algorithm 2) techniques designed for this setting. Synthetic results demonstrate that FRM+SL provides a strong improvement over FRM alone in asymmetric error settings. Results also show that FRM+SL performance remains robust given bias in \( \hat{\alpha}_d, \hat{\beta}_d \), and that CCPE recovers reasonable estimates for \( \alpha_d, \beta_d \). Future work will develop additional approaches that also leverage re-weighting, and add additional evaluations on semi-synthetic and real-world datasets.

## 6 Related work

To our knowledge, the treatment-conditional error setting we study is novel in this work. Techniques have been developed for addressing class-conditional [Menon et al. 2015, Scott et al. 2013], group-dependent [Wang et al. 2021], and instance-dependent [Xia et al. 2020] label noise models. [Menon et al. 2015, Scott et al. 2013] and [Northcutt et al. 2021] develop noise rate estimation approaches commonly used in label noise settings. Recently, [De-Arteaga et al. 2021] propose a method leveraging inter-expert consistency to adjust for measurement error in proxy outcomes. Yet this approach is designed for observational, rather than counterfactual, prediction settings. [Coston et al. 2020] develops counterfactual modeling and evaluation approaches for decision-support settings, which target proxy outcomes.
References


Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default [TODO] to [Yes], [No], or [N/A]. You are strongly encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section ??.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We highlight opportunities for future work in [5]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] https://anonymous.4open.science/r/counterfactual-decision-support-under-TCE-F53E/
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Details of experimental setup

Synthetic setup: Conducting evaluations for settings in which measurement error and treatment effects impact observed outcomes is challenging because often there is no way of recovering the true counterfactual outcome or the measurement error parameters. Therefore, we design an initial empirical evaluation via synthetic data to validate our proposed approach. Previous approaches have also used a uni-variate feature and protected attribute Coston et al. [2020b] or 2D synthetic data Natarajan et al. [2013], De-Arteaga et al. [2021]. We plan to extend to these settings and others with semi-synthetic and real-world data in later evaluations.

In all experiments, we use sinusoidal target class probability functions

\[
\eta^*_\alpha(x) = \begin{cases} 
.4 + .4 \cos(9x + 5.5) & x \in [-1, -.61] \\
.5 + .3 \sin(8x + .9) + .15 \sin(10x + .2) + .05 \sin(30x + .2) & x \in (-.61, .921] \\
x^3 & x \in (.921, 1]
\end{cases}
\]

\[
\eta^*_\beta(x) = .5 + .5 \sin(2.9x + .1) \forall x \in [-1, 1].
\]

We design \(\eta^*_\alpha(x)\) to be more challenging to estimate because we use it as our target quantity in this selective intervention setting. In all experiments, we use a linear propensity function \(\pi(x) = .35x + .5, \forall x \in [-1, 1]\).

Model and hyperparameters: All experiments use a MLP implemented via PyTorch with layer sizes \((1, 40, 20, 4, 1)\) and a binary cross-entropy loss satisfying our strictly proper composite criteria for class probability estimation. We run experiments with \(\alpha = .001\).

Evaluation. In all experiments, we split data 70/30 into training and validation folds. We evaluate accuracy with respect to \(Y^*\) on a held-out validation fold. We use accuracy as a performance measure rather than a metric such as AU-ROC because AU-ROC is immune to corruption from our error model (see Menon et al. [2015] for additional details).

Computing Environment Experiments were run on a MacBook Pro with 2.6 GHz 6-Core processor with 32 GB of RAM and Google Colab environment with standard runtime configuration.

A.2 Details of experiment configurations

- **Experiment 1.** Reported in Table 1. We run each setting with \(N = 10000\) and average performance over 10 runs with 40 epochs of training per run.
- **Experiment 2.** Reported in Figure 2. We run each setting with \(N = \{1000, 2000, 3000, 4000, 5000, 10000, 20000\}\) with 150 runs per setting and 30 epochs of training per round. Each round, we sample \(\alpha_0, \beta_0 \sim U(0, .3)\).
- **Experiment 3.** Reported in Figure 3. We run each setting with \(N = 10000\) with 15 runs per setting and 30 epochs of training per round. We vary \(\alpha_d (\beta_d)\) from .1, .2, .3 and hold \(\beta_0 (\alpha_0)\) fixed at 0. We then construct the surrogate loss \(\hat{L}\) with biased parameter estimates \(\hat{\alpha}_0, \hat{\beta}_0\) parameter estimates.
A.3 Baselines

- $\hat{P}_t[Y = 1| X = x]$ (OBS)
- $\hat{P}_t[Y = 1| D = d, X = x]$ (FRM)
- $\hat{P}_t[Y = 1| D = d, X = x]$ (FRM-SL)
- $\hat{P}_t[Y_d = 1| X = x]$ (PROXY$^*$)
- $\hat{P}_t[Y^*_d = 1| X = x]$ (TARGET$^*$)

In causal inference settings, FRM is also referred to as a backdoor covariate adjustment or plug-in estimator [Coston et al. 2020b,a, Pearl 2009].