Generalization Bounds for Estimating Causal Effects of Continuous Treatments

Anonymous Author(s)
Affiliation
Address
email

Abstract

We focus on estimating causal effects of continuous treatments (e.g., dosage in medicine), also known as dose-response function. Existing methods in causal inference for continuous treatments using neural networks are effective and to some extent reduce selection bias, which is introduced by non-randomized treatments among individuals and might lead to covariate imbalance and thus unreliable inference. To theoretically support the alleviation of selection bias in the setting of continuous treatments, we exploit the re-weighting schema and the Integral Probability Metric (IPM) distance to derive an upper bound on the counterfactual loss when estimating the average dose-response function (ADRF). The IPM distance builds a bridge between a source (factual) and an infinite number of target (counterfactual) domains. We provide a discretized approximation of the IPM distance with a theoretical guarantee in the practical implementation. Based on the theoretical analysis, we propose a novel algorithm, called **A**verage **D**ose-response esti**M**at**I**on via re-weigh**T**ing schema (ADMIT). ADMIT simultaneously learns a re-weighting network, which aims to alleviate the selection bias, and an inference network, which makes factual and counterfactual estimations. In addition, the effectiveness of ADMIT is empirically demonstrated in both synthetic and semi-synthetic experiments by outperforming the existing benchmarks.

1 Introduction

Causal inference, which estimates causal effects from observational data, is a fundamental problem in identifying causal reactions between actions/treatments and effects. It facilitates the decision-making process in a wide variety of domains, such as economics [1], public policy [2], medicine [3, 4] and advertising [5, 6]. Specifically, causal inference aims to estimate the causal effects of taking different treatments (e.g., whether to take a medication or not) in the binary setting, where neural network-based methods have made a breakthrough [7, 8, 9, 10]. In real-world applications, e.g. a medical experiment, treatments cannot be simplified as binary treatments since the dosage is a critical factor for the effect of the medication. Compared with causal inference for binary treatments, causal inference for continuous treatments, also known as dose-response estimation, deals with an infinite number of treatments, resulting in an infinite number of counterfactuals, far more than one counterfactual of the binary setting.

The success of neural networks on causal inference for binary treatments [11, 12, 13] can be attributed in part to the alleviation of selection bias, which leads to unreliable counterfactual inference since the covariates receiving different treatments are unbalanced due to non-randomized treatments among them. Specially, the two distributions of the treated and control groups are balanced using Integral Probability Metric (IPM) [11] or re-weighting schemes [12, 13]. In the continuous setting, since the control groups can be infinite, these standard methods for alleviating selection bias in binary

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

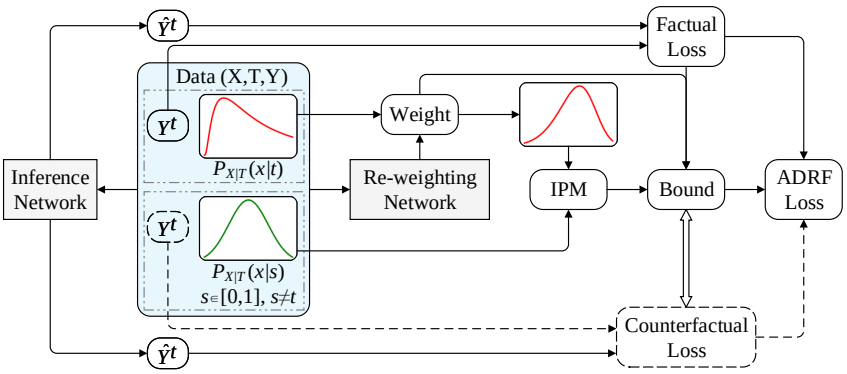

Figure 1: The framework of estimating ADRF using a re-weighting schema and IPM. The counterfactual loss cannot be computed since counterfactual outcomes are not unobserved. In the figure, unobservable variables and incomputable results are indicated by dotted rounded boxes and the infeasible is indicated by dashed arrows. We learn sample weights using a re-weighting network to make the distributions of the subpopulation receiving treatment t and another subpopulation receiving treatment $s \in [0, 1] s \neq t$ be closer in the IPM distance, i.e., more similar, which results in alleviating selection bias in the continuous treatment setting. We also learn an inference network to make factual (counterfactual) estimations. Then we exploit re-weighted factual loss and the IPM distance to bound the counterfactual loss. Finally, we combine the factual loss and the bound of the counterfactual loss as the ADRF loss, which is minimized to simultaneously learn the re-weighting network and the inference network.

treatments cannot be easily extended to mitigate selection bias among a theoretically infinite number of subgroups.

Currently, the Dose Response network (DRNet) [14] is proposed that extends "two heads", which are two outcomes in a neural network for making factual and counterfactual estimations in the binary setting, to "multiple heads" for a continuous dosage by discretizing the dosage into multiple intervals. Several regularization schemes in DRNet, such as distribution matching [7], propensity dropout [15] and matching on balancing scores [16, 17], are exploited to handle selection bias. However, these approaches are heuristic and do not provide theoretical evidence to back up their intuition.

To theoretically support the alleviation of selection bias, this paper derives an upper bound on the counterfactual loss during estimation of the average dose-response function (ADRF). The key points of the derivation lie in the re-weighted factual loss obtained from observational data and an Integral Probability Metric (IPM) that measures the distance between the factual (re-weighted) and counterfactual distributions, as illustrated in Figure 1. The infinite number of counterfactual distributions is the main challenge since the distance can not be calculated with finite samples in practice. In the practical implementation, we provide a discretized approximation of the IPM term and prove that the difference between its approximate and true values converges to zero in probability under the assumption that the covariates' distributions of subpopulations receiving different treatments smoothly shift. On the basis of the derived upper bound and its implementation with a theoretical guarantee, we propose a novel algorithm called **A**verage **D**ose-response esti**M**at**I**on via re-weigh**T**ing schema (ADMIT). ADMIT jointly learns an inference network and a re-weighting neural network to reduce selection bias while making factual and counterfactual estimations. The effectiveness of ADMIT is empirically confirmed by outperforming the existing benchmarks on both synthetic and semi-synthetic experiments in the continuous treatment setting.

Our contributions in this paper are 4-fold: (1) to the best of our knowledge, this is the first study that provides a generalization bound for estimating ADRF, alleviating selection bias; (2) we provide a discretized approximation of the IPM distance with a theoretical guarantee; (3) we propose an algorithm, ADMIT, capable of reducing selection bias while making factual and counterfactual estimations; (4) we conduct both synthetic and semi-synthetic experiments in the continuous treatment setting, empirically demonstrating the effectiveness of ADMIT with superiority over the existing benchmarks.

2 Related Work

Causal inference for continuous treatments is broadly studied through statistical methods [18, 19, 20, 21]. These methods are often based upon the generalized propensity score (GPS) [22], or entropy balancing based method [23] that learns the unit weights by incorporating information about known sample moments to achieve covariate balance, e.g., entropy balancing for continuous treatments (EBCT) [24]. Adaptation of neural networks for causal inference in the continuous treatments setting has attracted the attention of research community recently [14, 25, 26]. Schwab et al. [14] present the Dose Response network (DRNet) to estimate counterfactuals for continuous treatments, an extension of the network architecture proposed in [11]. Specifically, DRNet divides a continuous treatment t into several successive intervals, and trains one separate head for each interval. Bica et al. [26] point out that DRNet might be inflexible since it could not determine the divided intervals dynamically and propose SCIGAN that generates counterfactual outcomes for continuous treatments based on a generative adversarial network (GAN) framework. Meanwhile, for the issue that DRNet produces discontinuous ADRF due to interval discretization of continuous treatments, Varying Coefficient Network (VCNet) [25] lets the weights of the prediction head be continuous functions of the treatment to preserve the continuity of ADRF. Moreover, Nie et al. [25] obtain a doubly robust estimator of ADRF by generalizing the targeted regularization proposed in [10], and provide theoretical guarantees for its asymptotic correctness. Unlike VCNet, our work is derived from the generalization error to reduce the selection bias when estimating ADRF. To balance the covariates among infinite subpopulations, we learn re-sampling weights that reduce the IPM distance between observed and counterfactual groups. As such, we derive an upper bound on the estimated counterfactual error and demonstrate experimentally the proposed algorithm ADMIT based on the derived upper bound outperforms GPS, EBCT, DRNet, SCIGAN and VCNet.

For a discussion on research related to causal inference for binary or continuous treatments, see Appendix A. Appendix A also introduces the relationships between causal inference for continuous treatments and domain adaptation [27, 28, 29].

3 Problem Formulation

Let $\mathcal{X}$ denote the d-dimensional space of covariates and $\mathcal{Y}$ the outcome space. We refer to T as a continuous treatment random variable in an interval $\mathcal{T} = [a, b]$. Without loss of generality, we set $\mathcal{T} = [0, 1]$ in this study. Suppose we have an i.i.d sample of units, indexed by $i = 1, 2, \cdots, n$. Following Rubin's potential outcome framework [30], we postulate the existence of potential outcomes, Y^t, for $t \in \mathcal{T}$. For each unit i we observe the treatment s_i, and the covariate $\boldsymbol{x}_i$. The observed outcome (denoted by Y_i) is defined to be the potential outcome (denoted by $Y_i^{s_i}$) corresponding to the received treatment: $Y_i = Y_i^{s_i}$.

Given an input covariate $\boldsymbol{x}$ and a treatment t, we refer to $\mu(\cdot, \cdot)$ as the individual dose-response function, which is defined as:

$$\mu(t, \boldsymbol{x}) := \mathbb{E}[Y^t | \boldsymbol{X} = \boldsymbol{x}]. \tag{1}$$

Our goal is to derive an estimator of the average dose-response function (ADRF):

$$\mu(t) := \mathbb{E}[Y^t]. \tag{2}$$

By the law of iterated expectation, $\mu(t) = \mathbb{E}[\mu(t, \boldsymbol{X})]$. In this study, the estimator of ADRF $\mu(t)$ is based on estimating the individual dose-response function $\mu(t, \boldsymbol{x})$.

The following assumptions have been made to ensure that ADRF is identifiable from observational data.

Assumption 1. ***(Unconfoundedness)*** *Conditional on the covariates, X, the potential outcomes, Y^t, are independent of the treatment assignment, T,*

$$\{Y^t | t \in [0, 1]\} \perp T | \boldsymbol{X}.$$

This assumption is also known as *no unmeasured confounding* and ensures that the individual dose-response function is identifiable from observational data, i.e., $\mu(t, \boldsymbol{x}) = \mathbb{E}[Y^t | \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y | \boldsymbol{X} = \boldsymbol{x}, T = t]$.

Assumption 2. ***(Overlap)*** *There exists some constant $c > 0$, $\forall x \in \mathcal{X}$ and $\forall t \in \mathcal{T}$,*

$$P_{T|\boldsymbol{X}}(t|\boldsymbol{x}) \geq c.$$

This assumption is the condition that every point in the space of covariates has a non-zero conditional probability density of receiving any value of treatment.

4 Theoretical Analyses and Proposed Algorithm

This section firstly presents a theorem that the error of estimating ADRF is bounded by the marginal loss consisting of the factual and counterfactual losses. The counterfactual loss cannot be calculated from observational data since counterfactual outcomes as the ground truth are unobserved. In the meanwhile, the counterfactual inference might be unreliable due to selection bias. We then derive theoretically optimal weights in alleviating selection bias from the perspective of importance sampling. Although the optimal weights are theoretically guaranteed, they are impractical to learn due to their instability. Under a similar re-weighing schema, a theoretical guarantee is provided to guide the learning of adaptive weights that mitigate the impact of selection bias. We refer the reader to Appendix B for detailed proofs.

4.1 Error in estimating ADRF

Let $\hat{\mu}(t)$ be an estimator for $\mu(t)$, and we are interested in estimators with a small expected mean squared error (EMSE) in estimating ADRF,

$$\mathrm{EMSE}(\mu, \hat{\mu}) = \mathbb{E}[(\hat{\mu}(\mathrm{T}) - \mu(\mathrm{T}))^2]. \tag{3}$$

Since $\mu(t)$ cannot be observed from observational data, this study uses hypotheses $f : \mathcal{T} \times \mathcal{X} \to \mathcal{Y}$ to estimate the individual dose-response function $\mu(t, \boldsymbol{x})$ first. For ease of expression, we write $f(t, \boldsymbol{x})$ as $f_t(\boldsymbol{x})$. After that, we combine these hypotheses to estimate $\mu(t)$,

$$\hat{\mu}(t) = \mathbb{E}[f_t(\boldsymbol{X})]. \tag{4}$$

We bound the error $\mathrm{EMSE}(\mu, \hat{\mu})$ by the loss of hypotheses f_t, for $t \in \mathcal{T}$, with respect to the corresponding potential outcomes. Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function. The expected loss for the treatment and covariate pair $(t, \boldsymbol{x})$ is defined as:

$$l_{f_t}(\boldsymbol{x}) := \mathbb{E}_{Y^t|\boldsymbol{X}}[L(Y^t, f_t(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}]. \tag{5}$$

Theorem 1. *Let L be the squared loss function, i.e., $L(y, y') = (y - y')^2$. For hypotheses f_t of individual dose-response function $\mu(t, \cdot)$ with marginal loss $\epsilon(f_t) = \mathbb{E}[l_{f_t}(\boldsymbol{X})]$, there exists a constant $\sigma_{min} \geq 0$, such that,*

$$\mathrm{EMSE}(\mu, \hat{\mu}) \leq \mathbb{E}_T[\epsilon(f_t)] - \sigma_{min}. \tag{6}$$

Though $\mu(t)$ cannot be observed from observational data, Theorem 1 implies that its expected error can be bounded by the marginal loss $\epsilon(f_t)$, which are related to the prediction loss concerning potential outcomes, i.e., $l_{f_t}(\boldsymbol{x})$. However, the model cannot obtain marginal loss due to the inability to observe the counterfactuals. A re-weighting schema to address this issue is presented next.

4.2 Re-weighting via importance sampling

The fundamental problem of estimating ADRF is that it is impossible to observe all potential outcomes for any unit in data. The potential outcome Y^t could only be observed in the subpopulation that received treatment t, i.e., $T = t$. We refer to $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ as the distribution of $T = t$ subpopulation, and $p_{\boldsymbol{X}}(\boldsymbol{x})$ as the distribution of population. The two distributions, $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ and $p_{\boldsymbol{X}}(\boldsymbol{x})$, are not equal since $\boldsymbol{X}$ and T are not independent. Their difference leads to the *margin loss* $\epsilon(f_t)$ not being equal to the *conditional loss* of hypothesis f_t, $\epsilon(f_t|T = s)$, which is defined as:

$$\epsilon(f_t|T = s) = E_{\boldsymbol{X}|T}[l_{f_t}(\boldsymbol{X})|T = s]. \tag{7}$$

The *conditional loss* $\epsilon(f_t|T = s)$ is the expected loss with respect to the distribution of the subpopulation that received treatment s, i.e., $T = s$, and the *marginal loss* $\epsilon(f_t)$ is the expected loss with respect to the distribution of the population. When the treatments t and s are equal, the *conditional*

loss $\epsilon(f_t|T=s)$ is called *factual loss*, otherwise it is called *counterfactual loss*. The *counterfactual loss* is infeasible since the potential outcome Y^t could only be observed in the subpopulation who receive the treatment t. A minimizer f_t of $\epsilon(f_t)$ is required to minimize the error of estimating ADRF according to Theorem 1, which may be different from a minimizer of $\epsilon(f_t|T=t)$ since $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t) \neq p_{\boldsymbol{X}}(\boldsymbol{x})$. Importance sampling is introduced to estimate $\epsilon(f_t)$ in this work, which aims to estimate $\mathbb{E}_{\boldsymbol{x}\sim g(\boldsymbol{x})}[h(\boldsymbol{x})]$ of a particular distribution $g(\boldsymbol{x})$ while only samples generated from a different distribution $q(\boldsymbol{x})$ are available. The basic idea is to rewrite the expectation as:

$$\mathbb{E}_{\boldsymbol{x}\sim g(\boldsymbol{x})}[h(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}\left[h(\boldsymbol{x}) \cdot \frac{g(\boldsymbol{x})}{q(\boldsymbol{x})}\right], \tag{8}$$

where $\frac{g(\boldsymbol{x})}{q(\boldsymbol{x})}$ is called sampling weight. To estimate $\epsilon(f_t)$ based on the *factual loss* $\epsilon(f_t|T=t)$, a weight $w(x)$ is utilized for each unit to define a *re-weighted factual loss*:

$$\epsilon_w(f_t|T=t) = \mathbb{E}_{X|T}[w(\boldsymbol{X})l_{f_t}(\boldsymbol{X})|T=t]. \tag{9}$$

By Equation (8), the *re-weighted factual loss* is equal to the *marginal loss* when $w(\boldsymbol{x}) = w^*(\boldsymbol{x})$, where

$$w^*(\boldsymbol{x}) = \frac{p_{\boldsymbol{X}}(\boldsymbol{x})}{p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)} = \frac{p_T(t)}{p_{T|\boldsymbol{X}}(t|\boldsymbol{x})}. \tag{10}$$

The term $p_{T|\boldsymbol{X}}(t|\boldsymbol{x})$ is the conditional density of receiving a treatment t given the covariate $\boldsymbol{x}$, which is also known as the *Generalized Propensity Score* (GPS) [22, 31]. In binary treatment cases, the propensity score could balance the distributions of covariates between the treated control groups [32]. Hirano and Imbens [31] stress that the GPS has a similar balancing property in continuous treatment cases. Fong et al. [20] proposed the covariate balancing generalized propensity score methodology by defining a stabilized weight based on the GPS. Interestingly, the weight defined in [20] is exactly the same as $w^*(\boldsymbol{x})$ derived from the sample weight via importance sampling.

The weight $w^*(\boldsymbol{x})$ based on the GPS is theoretically optimal in expectation [22] but may be not practical to obtain. On the one hand, the approaches based on the propensity score may be unstable in the estimations [33, 34]. What's worse is that, this issue may be amplified in continuous treatment cases since the conditional density is difficult to estimate accurately [35]. On the other hand, it could be proved that a large sampling weight $\frac{g(\boldsymbol{x})}{q(\boldsymbol{x})}$ may result in a high variance in the estimations, which could lead to a poor finite sample performance. Therefore, the following subsection focuses on learning adaptive weights to balance observed covariates across subpopulations that received different treatments and giving a generalization bound of the *marginal loss*.

4.3 Generalization bound of the *marginal loss*

The imbalance of covariates across subpopulations that received different treatments caused by selection bias is closely connected with covariate shift in domain adaptation. This study alleviates this imbalance using an IPM distance, which is similar to the distribution distance metrics in domain adaptation, and gives a generalization bound of the *marginal loss* when using adaptive weights in Equation (9).

The *Integral Probability Metric* (IPM) is a class of metrics that measure the distance between two probability distributions, which is defined as:

$$\mathrm{IPM}_{\mathcal{G}}(p,q) = sup_{g\in\mathcal{G}}\left|\int g(\boldsymbol{x})(p(\boldsymbol{x}) - q(\boldsymbol{x}))d\boldsymbol{x}\right|, \tag{11}$$

where $\mathcal{G}$ is a function family of functions $g: \mathcal{X} \to \mathcal{R}$. For simplicity, we abbreviate $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ and $p_{\boldsymbol{X}|T}(\boldsymbol{x}|t)w(\boldsymbol{x})$ as p_t and p_t^w, respectively. We first state a Lemma bounding the difference between the *re-weighted factual loss*, $\epsilon_w(f_t|T=t)$, and the *counterfactual loss*, $\epsilon(f_t|T=s)$, where $s \neq t$.

Lemma 1. *Let $\mathcal{G}$ be a family of functions $l: \mathcal{X} \to \mathcal{R}$. Assume the per-unit expected loss function $l_{f_t} \in \mathcal{G}$. Then for any $s \in [0,1]$ and $s \neq t$, we have:*

$$\epsilon(f_t|T=s) \leq \epsilon_w(f_t|T=t) + \mathrm{IPM}_{\mathcal{G}}(p_s, p_t^w). \tag{12}$$

Based on Lemma 1, we could give a generalization bound of the *marginal loss* through the *re-weighted factual loss* as follows.

Theorem 2. *Let* $\mathrm{IPM}_{max} = max_{s\in[0,1]}\{\mathrm{IPM}_{\mathcal{G}}(p_s, p_t^w)\}$. *The following holds under the conditions of Lemma 1,*

$$\epsilon(f_t) \leq \epsilon_w(f_t|T=t) + \mathrm{IPM}_{max}. \tag{13}$$

Theorem 2 shows that the *marginal loss* is upper bound by the average of the *re-weighted factual loss* and an IPM distance. The first term could be estimated with observed samples. The second term is defined based on the distance between the two probability distribution of the subpopulation receiving treatment t and another subpopulation. Choosing a family of norm-1 reproducing kernel Hilbert space (RKHS) functions leads to IPM being the Maximum Mean Discrepancy (MMD) metric [36]. When involving discrete domains, there may have samples $\{\boldsymbol{x}_1^p, \cdots, \boldsymbol{x}_n^p\} \sim p$ and $\{\boldsymbol{x}_1^q, \cdots, \boldsymbol{x}_m^q\} \sim q$, an estimator of the MMD distance between p and q could be obtained as follows:

$$\widehat{\mathrm{MMD}}_k^2(p,q) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\boldsymbol{x}_{\boldsymbol{i}}^p, \boldsymbol{x}_{\boldsymbol{j}}^p) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} k(\boldsymbol{x}_{\boldsymbol{i}}^p, \boldsymbol{x}_{\boldsymbol{j}}^q) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{i=j}^{m} k(\boldsymbol{x}_{\boldsymbol{i}}^q, \boldsymbol{x}_{\boldsymbol{j}}^q), \tag{14}$$

where k denotes a differentiable kernel, e.g., the Gaussian RBF kernel. A certain number of samples are needed to estimate the MMD as shown in Equation 14. When the treatment is continuous, the samples that received some treatment t may be unavailable since only a finite number of samples are observed while the choice of t is infinite. In other words, the IPM term in Theorem 2 cannot be estimated empirically. To overcome this challenge, we bound the difference between the IPM_{max} and its discretization under the assumption that the probability distributions of subpopulations that received different treatments shift smoothly.

Assumption 3. *Let* p_{s_1} *and* p_{s_2} *denote the conditional probability densities of subpopulations that received treatment* s_1 *and* s_2*, respectively. We assume that there is a constant* α *such that the following inequality holds for all* $s_1, s_2 \in [0,1]$*:*

$$\mathrm{IPM}_{\mathcal{G}}(p_{s_1}, p_{s_2}) \leq \alpha\,|s_1 - s_2|\,. \tag{15}$$

Intuitively, α indicates the rate of the probability distribution shift of subpopulations, and a reasonably small α means the covariates smoothly shift when receiving different treatments.

Based on Assumption 3, the following Lemma states that the IPM_{max} corresponding to continuously shifting domain (for each $s \in [0,1]$) could be approximated as an IPM term based on discrete domains (for each $s \in \{s_1, \cdots, s_n\}$).

Lemma 2. *Suppose we have* n *i.i.d sample of units, and the* i*-th unit received a treatment* $s_i \sim p(t)$*. This study assumes Assumption 3 holds for a constant* α*. Then the following holds,*

$$\mathrm{IPM}_{max} \leq \max_{i\in\{1,\cdots,n\}}\{\mathrm{IPM}_{\mathcal{G}}(p_{s_i}, p_t^w)\} + O_p\left(\frac{\alpha}{\sqrt[3]{n}}\right). \tag{16}$$

Lemma 2 bounds the IPM_{max} by a discrete set $\{\mathrm{IPM}_{\mathcal{G}}(p_{s_i}, p_t^w) | i \in \{1, \cdots, n\}\}$. To empirically estimate $\mathrm{IPM}_{\mathcal{G}}(p_{s_i}, p_t^w)$, the following lemma states that a point s_i could be replaced by its neighbourhood $[s_i, s_i + \delta]$.

Lemma 3. *Let* $p_{\Delta s} = P_{\boldsymbol{X}|T}(\boldsymbol{x}|t \in [s, s+\delta])$ *(*$0 < \delta < 1$*) denotes the conditional density of covariates when* $t \in [s, s+\delta]$*. Then the following holds under Assumption 3,*

$$\mathrm{IPM}_{\mathcal{G}}(p_s, p_t^w) \leq \mathrm{IPM}_{\mathcal{G}}(p_{\Delta s}, p_t^w) + \alpha\delta. \tag{17}$$

Following Lemma 2 and Lemma 3, we bound the *marginal loss* as follows.

Theorem 3. *Suppose we have* n *i.i.d sample of units, and the* i*-th unit received a treatment* s_i*. Let* $\mathrm{IPM}_{\Delta max} = max_{i\in\{1,\cdots,n\}}\{\mathrm{IPM}_{\mathcal{G}}(p_{\Delta s_i}, p_t)\}$*. This study assumes Assumption 3 holds for a constant* α*. Then, for a neighbourhood size* $0 < \delta < 1$ *we have,*

$$\epsilon(f_t) \leq \epsilon_w(f_t|T=t) + \mathrm{IPM}_{\Delta max} + O_p\left(\frac{\alpha}{\sqrt[3]{n}}\right) + \alpha\delta. \tag{18}$$

Contrary to the second term in Theorem 2, $\mathrm{IPM}_{\Delta max}$ could be estimated empirically, since we may get samples that received treatment $s \in [s_i, s_i + \delta]$ by choosing an appropriate neighbourhood size δ.

Algorithm 1 ADMIT: Average dose-response estimation via re-weighting schema

Input: Observed samples $\mathcal{D} = \{(\boldsymbol{x_i}, s_i, Y_i) | i = 1, \cdots, n\}$, loss function $L(\cdot, \cdot)$,function family $\mathcal{G}$ for IPM, batch size n_b, neighbourhood size δ. Initialized parameters: $\theta = [\theta_\phi, \theta_h, \theta_\omega]$

1: **while** θ not converged **do**
2: Sample mini-batch $\{(\boldsymbol{x}_{i_1}, s_{i_1}, y_{i_1}), \cdots, (\boldsymbol{x}_{i_{n_b}}, s_{i_{n_b}}, y_{i_{n_b}})\}$ from $\mathcal{D}$
3: Calculate the weight of each unit: $w_{i_j} = \omega(s_{i_j}, \phi(x_{i_j}))$
4: **for** $l = 1$ to $L = \lceil 1/\delta \rceil$ **do**
5: **for** $k = 1$ to L and $k \neq l$ **do**
6: $\mathrm{IPM}_{\mathcal{G}}(p_{\Delta k}, p^w_{\Delta l}) = \mathrm{IPM}_{\mathcal{G}}(\{w_{i_j} \cdot \boldsymbol{x}_{i_j}\}_{s_{i_j} \in [(l-1)\delta, l\delta]}, \{\boldsymbol{x}_{i_h}\}_{s_{i_h} \in [(k-1)\delta, k\delta]})$
7: **end for**
8: $\mathrm{IPM}^l_{\Delta max} = max_{k \in \{1, \cdots, L\}, k \neq l}\{\mathrm{IPM}_{\mathcal{G}}(p_{\Delta k}, p^w_{\Delta l})\}$
9: **end for**
10: Calculate the gradients of the IPM term:
$g_1 = \Delta_{\theta_\omega} \frac{1}{L} \sum_l \mathrm{IPM}^l_{\Delta max}$
$g_2 = \Delta_{\theta_\phi} \frac{1}{L} \sum_l \mathrm{IPM}^l_{\Delta max}$
11: Calculate the gradients of the *re-weighted factual loss*:
$g_3 = \Delta_{\theta_h} \frac{1}{n_b} \sum_j w_{i_j} \cdot L(h(s_{i_j}, \phi(\boldsymbol{x}_{i_j})), y_{i_j})$
$g_4 = \Delta_{\theta_\phi} \frac{1}{n_b} \sum_j w_{i_j} \cdot L(h(s_{i_j}, \phi(\boldsymbol{x}_{i_j})), y_{i_j})$
12: Update parameters: $[\theta_\phi, \theta_h, \theta_\omega] \leftarrow [\theta_\phi - \eta(g_2 + g4), \theta_h - \eta g_3, \theta_\omega - \eta g_1]$
13: **end while**

Output: Learned representation function $\phi(\cdot)$, outcome function $h(\cdot, \cdot)$, re-weighting function $\omega(\cdot, \cdot)$

To minimize the error ϵ_u in estimating the average dose-response function, Theorem 1 and Theorem 3 suggest that we may minimize the upper bound in Equation (18) with respect to the *re-weighted conditional loss* and the IPM term, which leads us to Algorithm 1 detailed in the following subsection.

4.4 Proposed algorithm

An algorithm, called average dose-response estimation via re-weighting schema (ADMIT) is proposed based on the theoretical results derived in the previous subsections. As illustrated in Figure 2, we use neural networks to build the model, which is mainly composed of three components: a representation network $\phi(\boldsymbol{x})$, a re-weighting network $\omega(t, \phi(\boldsymbol{x}))$, and an inference network $h(t, \phi(\boldsymbol{x}))$.

This work firstly learns a treatment-agnostic representation $z = \phi(\boldsymbol{x})$ of $\boldsymbol{x}$ using all data. The influence of treatment t is different from that of $\boldsymbol{x}$ [11], which might be lost during training if $\phi(\boldsymbol{x})$ is high-dimensional. To address this issue, the varying coefficient neural network following [25] is used to build $h(t, \phi(\boldsymbol{x}))$ and $\omega(t, \phi(\boldsymbol{x}))$. Take the former as an example, $h(t, \phi(\boldsymbol{x}))$ is defined as:

$$h(t, \phi(\boldsymbol{x})) = \mathrm{NN}_{\theta_h(t)}(\phi(\boldsymbol{x})), \tag{19}$$

where $\mathrm{NN}_{\theta_h(t)}$ is a neural network with parameter $\theta_h(t)$. Thus, the treatment t determines the nonlinear function in $\mathrm{NN}_{\theta_h(t)}$.

Based on the theoretical analysis aforementioned, the algorithm aims to seek a re-weighting function ω and outcome function h to minimize the error $\mathrm{EMSE}(\mu, \hat{\mu})$ of estimating the average dose-response function, using the following objective:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \omega(s_i, \phi(\boldsymbol{x_i})) \cdot L(h(s_i, \phi(\boldsymbol{x_i})), Y_i) + \max_{j \in \{1, \cdots, n\}, j \neq i}\{\mathrm{IPM}_{\mathcal{G}}(p_{\Delta s_j}, p^w_{\Delta s_i})\}, \tag{20}$$

where $\theta = [\theta_\phi, \theta_\omega, \theta_h]$. To effectively calculate the IPM, the interval $[0, 1]$ is divided into $\lceil 1/\delta \rceil$ intervals to confirm the neighbourhood of the i-th unit, and details of the calculation are presented in lines 4 to 9 of the Algorithm 1. The model is trained using stochastic gradient descent to minimize the objective. Finally, an estimator of ADRF is built as:

$$\hat{\mu}(t) = \frac{1}{n} \sum_{i=1}^{n} h(t, \phi(\boldsymbol{x_i})). \tag{21}$$

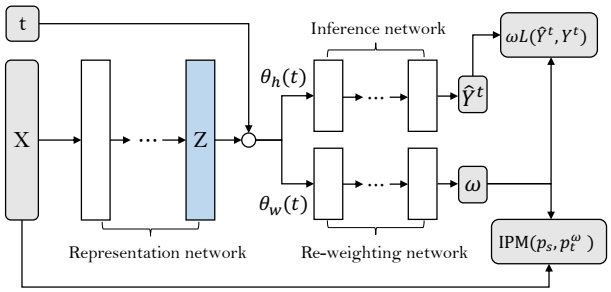

Figure 2: Model architecture of the proposed ADMIT.

5 Experiments

It is difficult to get the ground truth treatment effect on real-world datasets due to the inability to observe the counterfactuals. To deal with this, existing literature [14, 25, 26] often use synthetic/semi-synthetic data for meaningful evaluation. We use one synthetic dataset and two semi-synthetic datasets, News [14, 7] and TCGA [37], to demonstrate the effectiveness of ADMIT. Details on datasets and implementation can be found in Appendix C.

5.1 Experimental setup

Synthetic data generation. We generate covariates $\boldsymbol{x} \sim \text{Unif}(\mathbf{0}, \mathbf{1}) \in \mathbb{R}^6$, and the assigned treatments and their corresponding outcomes are generated as follows:

- $t = (1 + \exp(-t'))^{-1}$, $t' \sim \mathcal{N}(\mu_1, 0.5)$ where
$$\mu_1 = 10\frac{\sin(\max(x_1, x_2, x_3)) + \max(x_3, x_4, x_5)^3}{1 + (x_1 + x_5)^2} + \sin(0.5x_3)(1 + \exp(x_4 - 0.5x_3)) + x_3^2 + 2\sin(x_4) + 2x_5 - 6.5.$$
- $y \sim \mathcal{N}(\mu_2, 0.5)$ where $\mu_2 = \cos(2\pi(t - 0.5))(t^2 + 4\frac{\max(x_1, x_6)^3}{1+2x_3^2}\sin(x_4))$.

Semi-synthetic data generation. We obtain covariates from two real-world datasets, News [14, 7] and TCGA [37]. To generate the assigned treatments and their corresponding outcomes, we generate a set of parameters, $\boldsymbol{v}_i = \boldsymbol{u}_i/||\boldsymbol{u}_i||$ and $i = 1, 2, 3$, where $\boldsymbol{u}_i$ is sampled from a uniform distribution $\text{Unif}(\mathbf{0}, \mathbf{1})$, then

- $t \sim$Beta(α, β) where $\beta = \frac{\alpha-1}{d^*} + 2 - \alpha$ and $d^* = \frac{\boldsymbol{v}_3^T \boldsymbol{x}}{2\boldsymbol{v}_2^T \boldsymbol{x}}$, and
- $y \sim \mathcal{N}(\mu, 0.5)$ where μ=$2(\max(-2, \exp((\frac{\boldsymbol{v}_2^T \boldsymbol{x}}{\boldsymbol{v}_3^T \boldsymbol{x}} - 0.3)))+10\boldsymbol{v}_1^T \boldsymbol{x}) \times 4(t-0.5)^2 \times \sin(\frac{\pi}{2}t)$.

As discussed in [26], a larger α will result in a higher selection bias. Without specific instruction, we set $\alpha = 2$ in our experiments.

Baselines and metrics. For neural network baselines, we compare ADMIT with DRNet [14], VCNet [25] and SCIGAN [26]. With the same settings as in [25], we apply targeted regularization on both DRNet and VCNet, and add a conditional density estimation head for DRNet. For statistical baselines, we compare ADMIT with GPS [31] and entropy balancing for continuous treatments (EBCT) [24]. It is worth mentioning that EBCT learns sample weights from data separately, rather than simultaneously learning sample weights and the inference model as in this paper. When evaluating EBCT, we take VCNet as its inference network to estimate the ADRF. For metrics, we use Expected Mean Square Error (EMSE) on the test dataset, where $\text{EMSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}(t_i) - \mu(t_i))^2$.

5.2 Benchmarks comparison

As shown in Table 1, the proposed ADMIT is significantly more accurate in estimating the average dose-response in both synthetic and semi-synthetic datasets. Considering the dimension of covariates

Table 1: Comparison of ADMIT with (non) neural network-based baselines. Reported performance ($\sqrt{\text{EMSE}}$) of average dose-response estimation on one synthetic and two semi-synthetic datasets. Metrics are reported as Mean $\pm$ Std.

Method	Simulation	TCGA	News
DRNet	0.209 ± 0.0127	0.216 ± 0.0281	0.274 ± 0.0194
SCIGAN	0.638 ± 0.0467	0.301 ± 0.0204	0.707 ± 0.0074
VCNet	0.129 ± 0.0186	0.139 ± 0.0107	0.201 ± 0.0193
VCNet+EBCT	0.162 ± 0.0256	0.126 ± 0.0234	0.196 ± 0.0123
GPS	0.889 ± 0.0021	0.823 ± 0.0019	1.136 ± 0.0024
ADMIT	$\mathbf{0.081 \pm 0.0048}$	$\mathbf{0.071 \pm 0.0074}$	$\mathbf{0.106 \pm 0.0135}$

in TCGA and News is 4000 and around 3000 respectively. This result may indicate that ADMIT is well adapted to the high-dimensional data. Moreover, using sample weights obtained by pre-processing the data with EBCT also enhances the performance on the two semi-synthetic datasets compared with VCNet. This result demonstrates the effectiveness of the re-weighting schema in enhancing the performance of estimating causal effects. Furthermore, the performance boost is smaller compared with ADMIT, which indicates the superiority of the proposed re-weighting strategy in this study.

5.3 Selection bias

As indicated in [26], the variance of the Beta distribution decreases with the increase of α in the semi-synthetic data generation, resulting in higher selection bias. We compare the performance of ADMIT with VCNet, SCIGAN and DRNet on TCGA when varying choices of α from 1 (no bias) to 8 (high bias) to assess the robustness of these methods in different levels of selection bias. On one hand, Figure 3 illustrates that the performance of the four models tends to decrease, i.e. increase $\sqrt{\text{EMSE}}$, as selection bias increases, which is consistent with the theoretical analysis. On the other hand, we found that ADMIT shows consistent performance and outperforms the three baselines across the entire range of evaluated selection bias. This result demonstrates the effectiveness of ADMIT in empirically mitigating the impact of selection bias.

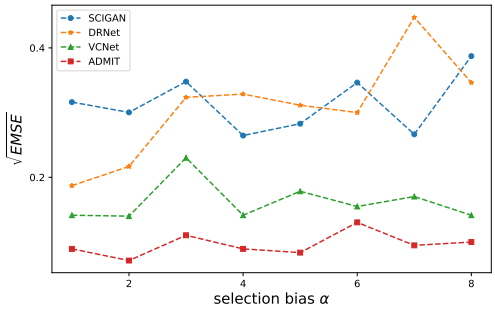

Figure 3: Comparison of four models in terms of their $\sqrt{\text{EMSE}}$ for varying selection bias on TCGA.

6 Conclusion

In this paper, we propose a novel re-weighting schema to mitigate the impact of selection bias in causal inference with continuous treatments. Under this schema, this paper provides and proves a generalization error bound on the estimation of ADRF based on an IPM distance between the observed (re-weighted) and counterfactual distributions, with theoretical evidence on learning sample weights that alleviate the unreliable counterfactual inference problem caused by selection bias. In the practical implementation, we provide a discretized approximation of the IPM distance with a theoretical guarantee, leading to an algorithm ADMIT, that jointly learns re-weighting and inference neural networks. The superiority of ADMIT is validated in both synthetic and semi-synthetic experiments. Experiments show that re-weighting schemas facilitate causal inference.

References

[1] Hal R Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016.

[2] Andrei Svetovidov, Abdul Rahman, and Feras A Batarseh. AI assurance using causal inference: Application to public policy. *arXiv preprint arXiv:2112.00591*, 2021.

[3] Afsaneh Yazdani and Eric Boerwinkle. Causal inference in the age of decision medicine. *Journal of data mining in genomics and proteomics*, 6(1), 2015.

[4] Safoora Gharibzadeh, Kazem Mohammad, Abbas Rahimiforoushani, Atieh Amouzegar, and Mohammad Ali Mansournia. Standardization as a tool for causal inference in medical research. *Archives of Iranian medicine*, 19(9), 2016.

[5] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

[6] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. Causal inference via sparse additive models with application to online advertising. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[7] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International conference on machine learning*, pages 3020–3029. PMLR, 2016.

[8] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Proceedings of the 32nd Advances in Neural Information Processing Systems*, volume 31, 2018.

[9] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[10] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Proceedings of the 33rd Advances in neural information processing systems*, volume 32, 2019.

[11] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[12] Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5880–5887, 2019.

[13] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

[14] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.

[15] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

[16] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[17] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

[18] Yeying Zhu, Donna L Coffman, and Debashis Ghosh. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of causal inference*, 3(1):25–40, 2015.

[19] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

[20] Christian Fong, Chad Hazlett, and Kosuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

[21] Brian G Vegetabile, Beth Ann Griffin, Donna L Coffman, Matthew Cefalu, Michael W Robbins, and Daniel F McCaffrey. Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology*, 21(1):69–110, 2021.

[22] Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

[23] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

[24] Stefan Tübbicke. Entropy balancing for continuous treatments. *Journal of Econometric Methods*, 11(1):71–89, 2022.

[25] Lizhen Nie, Mao Ye, Dan Nicolae, et al. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *Proceedings of the 9th International Conference on Learning Representations*, 2020.

[26] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Proceedings of the 34th Advances in Neural Information Processing Systems*, volume 33, pages 16434–16445, 2020.

[27] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9898–9907. PMLR, 2020.

[28] Hong Liu, Mingsheng Long, Jianmin Wang, and Yu Wang. Learning to adapt to evolving domains. In *Proceedings of the 34th Advances in Neural Information Processing Systems*, volume 33, pages 22338–22348, 2020.

[29] Guillermo Ortiz-Jimenez, Mireille El Gheche, Effrosyni Simou, Hermina Petric Maretic, and Pascal Frossard. CDOT: Continuous domain adaptation using optimal transport. *arXiv preprint arXiv:1909.11448*, 2019.

[30] Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

[31] Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

[32] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[33] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

[34] José R Zubizarreta, Caroline E Reinke, Rachel R Kelz, Jeffrey H Silber, and Paul R Rosenbaum. Matching for several sparse nominal variables in a case-control study of readmission following surgery. *The American Statistician*, 65(4):229–238, 2011.

[35] Miguel A Hernán and James M Robins. Causal inference, 2010.

[36] 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.

[37] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Checklist

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? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (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? [Yes] See Appendix B.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix B.
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] See Appendix C.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix C.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5
   (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] See Appendix C.
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? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix C.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Appendix C.
   (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]