Sampling Boundary for Causal Effect Estimation

Yue Yin YUEYIN@MAIL.HFUT.EDU.CN
Jiaoyun Yang* JIAOYUN@HFUT.EDU.CN
Ning An NING.G.AN@ACM.ORG
Lian Li LLIAN@HFUT.EDU.CN
School of Computer Science and Information Engineering, Hefei University of Technology, 485 Danxia Road, Shushan District, Hefei, Anhui, China

Editors: Hung-yi Lee and Tongliang Liu

Abstract

In causal effect estimation, determining the appropriate sampling size is critical for ensuring reliability and validity in both experimental and observational studies, a challenge closely tied to robust model generalization under limited data conditions in machine learning. This paper tackles these challenges by leveraging the Probably Approximately Correct (PAC) theory to establish a theoretically grounded framework for determining sampling boundaries. We utilize Hoeffding's inequality and Vapnik–Chervonenkis (VC) dimension to set upper boundaries for dataset adequacy in diverse scenarios: no confounders, confounders with a finite hypothesis space, and confounders with an infinite hypothesis space. Our work ensures that if the dataset size exceeds the upper boundary, the error probability for the estimated causal effect stays within a specified threshold at the given confidence level. Additionally, we demonstrate that when the dataset size is inadequate, the error of the estimated average treatment effects is bounded by the estimation of the outcome variable, which forms the theoretical basis for data augmentation strategies to improve the accuracy of causal effect estimation. Extensive experiments on synthetic and semi-synthetic datasets validate the correctness of our presented sampling upper limitations under different error and confidence level constraints. Our findings not only offer a systematic and reliable method for determining sample size in causal effect estimation but also provide actionable guidance for developing causal inference models in data-scarce environments, enhancing their applicability and robustness across fields such as healthcare, social sciences, and policy evaluation.

Keywords: Causal Effect; Sampling Boundary; Probably Approximately Correct.

1. Introduction

Causal effect estimation aims to measure causal relations by quantifying the influence of cause variables on outcomes (Pearl, 2009). It serves as the theoretical foundation for achieving robust modeling and reliable inference, playing an essential role in revealing mechanisms behind complex phenomena and enabling more informed decision-making (Guo et al., 2020). This foundational importance extends across various domains, including data science, social science, and medicine (Imbens and Rubin, 2015). For causal effect estimation, it is critical to first collect data that closely resembles natural sampling.

Data for causal analysis arise from two primary paradigms—randomised controlled trials (RCTs) and observational studies (Guo et al., 2020). RCTs actively assign treatments, block confounders, and are regarded as the gold standard for causal inference, yet they are

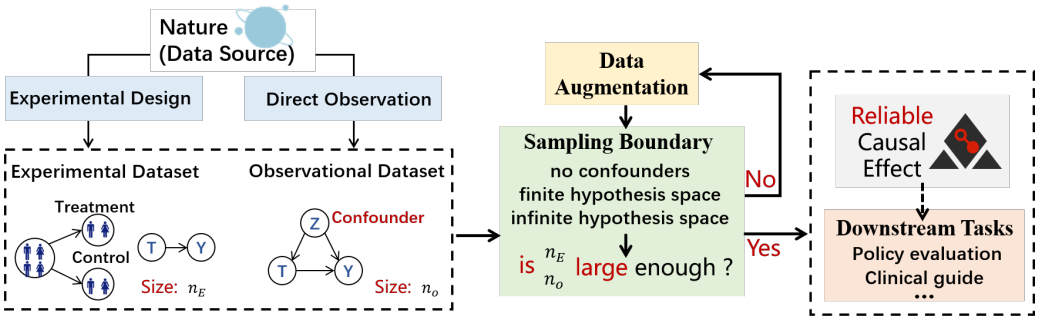

Figure 1: From data sampling to reliable causal estimation. Experimental and observational sampling yield datasets of size n_E and n_O, respectively. The proposed PAC sample–size bound N—instantiated for three confounding regimes—checks whether the current sample size n is sufficient. If $n < N$, we augment the data until $n_{\text{aug}} \geq N$.

costly and can raise ethical concerns (Didero et al., 2021). Observational studies collect data passively, for example through surveys or sensor logging, and are easier to deploy, but confounding is unavoidable (Gentzel et al., 2021). Even when all confounders are measured, their presence inflates the variance of treatment-effect estimates (Freedman, 2008). By the law of large numbers and the central limit theorem, increasing the sample size stabilises these estimates and thereby mitigates residual confounding effects. Consequently, determining an appropriate sample size is pivotal for reliable causal effect estimation (John et al., 2019).

Existing studies on sample-size determination for causal inference fall into two main strands: empirical-equation methods and simulation-based rules of thumb. Empirical equations derive heuristic formulas from practitioner experience. (Sharma et al., 2020) compile standard size-calculation formulas for nursing studies that use both observational and experimental data. (Wolf et al., 2013) propose an empirical expression for structural-equation models, whereas (Taherdoost, 2017) summarises sample-size methods for survey research in social and information-systems domains. The second strand calibrates sample-size rules through simulation studies on benchmark data. (Markoulidakis et al., 2021) recommend collecting 60–80 observations per confounder and per treatment arm; (Kretzschmar and Gignac, 2019) determine the sample size required for stable latent-variable correlations; (Riley et al., 2020) evaluate requirements for clinical prediction models; and (Yang et al., 2021) propose a distribution-preserving heuristic. Although these approaches perform well within their respective assumptions, none provides a universal theoretical guarantee. (Zhang and Bareinboim, 2021) introduce a PAC-learning perspective that moves toward such guarantees, but their analysis does not yet offer an explicit upper bound on the required sample size—a gap that the present study addresses.

We propose a more general approach for deriving sample-size boundaries in causal effect estimation under different confounding scenarios. Specifically, we formulate the estimation of the Population Average Treatment Effect (PATE) from the Sample Average Treatment Effect (SATE) within the Probably Approximately Correct (PAC) learning framework. If the dataset size meets or exceeds our derived boundary, the probability that the estimation error remains below a specified threshold surpasses a predetermined confidence level Mohri

et al. (2012). In that case, we regard the causal effect as PAC-estimable and consider the estimated causal effect to be reliable. Our analysis encompasses three settings—no confounders, confounders with a finite hypothesis space, and confounders with an infinite hypothesis space—and utilizes both the VC dimension and Hoeffding's inequality Shalev-Shwartz and Ben-David (2014) to establish sample-size requirements in each scenario.

When the dataset size does not meet our derived sample-size boundary, we further investigate the use of machine learning models for data augmentation. In particular, we establish that the estimation error for the average treatment effect (ATE) is bounded by the prediction error of the outcome model, implying that improving outcome prediction accuracy can directly enhance causal effect estimation. Therefore, a high-performing machine learning model can be employed to augment the original dataset, alleviating the challenges posed by limited data.

To validate our proposed boundaries and the efficacy of data augmentation, we first conduct simulation studies covering scenarios with 0, 1, 2, 5, and 10 confounders, thereby confirming the accuracy of the derived sample-size thresholds. Subsequently, we apply machine learning-based augmentation to the IHDP dataset (McCormick et al., 1998), demonstrating that augmenting the dataset with model-generated samples significantly improves the precision of causal effect estimates. These findings highlight the practical value of combining theoretically grounded sampling boundaries with data augmentation when the available dataset falls below the recommended threshold.

2. Related Work

Causal Effect Estimation. Propensity Score Matching (PSM) was proposed by Rosenbaum (1987) to reduce bias by matching treatment and control units with similar covariates. Angrist and Pischke (2009) introduced Instrumental Variable (IV) Analysis to tackle endogeneity through external variables affecting treatment allocation but not outcomes. Lee and Lemieux (2010) discussed Regression Discontinuity Design (RDD), exploiting assignment thresholds for more precise local effect estimates, whereas Difference-in-Differences (DiD) uses temporal data from treated and control groups to infer treatment impact (Lechner, 2011). Kline (2023) further showed how Structural Equation Modeling (SEM) can incorporate latent variables and complex interdependencies. Recent machine learning approaches, including Dragonnet (Shi et al., 2019), DRNet (Schwab et al., 2020), SCIGAN (Bica et al., 2020), and EDVAE (Liu et al., 2024), have extended causal inference to continuous treatments and high-dimensional data. Despite these advances, determining an adequate sample size remains a significant challenge, since most methods rely on context-specific assumptions and lack universal theoretical guarantees.

Probably Approximately Correct (PAC) Theory. PAC theory (Mohri et al., 2012) separates "probably" from "approximately correct" via the confidence level $(1-\delta)$ and error tolerance ε. PAC-oriented analyses of treatment-effect estimation have recently examined generalization behavior via estimator-specific discrepancy measures. For example, optimal-transport–based methods (Wang et al., 2023) and proximity-preserving balancing methods (Wang et al., 2025) provide high-probability CATE/PEHE bounds that depend on discrepancy terms and hypothesis complexity, thereby characterizing how error decays with sample size. We adapt PAC concepts to causal effect estimation by introducing "PAC-estimable

Table 1: Summary of notations used throughout this paper.

Notation	Description
T	Treatment variable
Y	Outcome variable
X	Confounders affecting both T and Y
ε	Error threshold for estimated causal effects
$1-\delta$	Confidence level
D	The whole population dataset
D'	Sampled dataset from D
n	Sampled dataset size
N	Sampling boundary
d	VC dimension
C	Number of confounders
$\mathcal{H}$	Hypothesis space
$h_1, \ldots, h_{\lvert\mathcal{H}\rvert}$	Hypotheses in $\mathcal{H}$

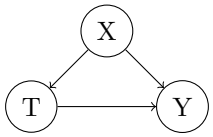

Figure 2: An illustration of a confounder X affecting both treatment T and outcome Y.

causal effects," which establishes a probabilistic framework for accuracy–confidence guarantees. Complementing this line, we derive a PAC sampling boundary $N(\varepsilon, \delta)$ for ATE across three confounding regimes (see §4), providing an explicit criterion for sample-size adequacy at prescribed accuracy and confidence.

Hoeffding's inequality. Hoeffding's inequality (Vapnik, 1995) provides a probability bound for how the average of independent bounded random variables deviates from its expectation. It is widely applied in machine learning for ensuring performance consistency (e.g., boosting (Schapire and Freund, 2013) and high-dimensional analysis (Bühlmann and Van De Geer, 2011)). We leverage Hoeffding's inequality to derive a sample size criterion in causal effect estimation, extending its utility beyond supervised learning contexts.

3. Preliminaries

We next introduce key concepts in causal effect estimation and outline the PAC-based problem formulation. Table 1 lists the main notations used in this paper.

3.1. Related Concepts

Average Treatment Effect (ATE). Let $T = 1$ denote treatment and $T = 0$ denote control. For each unit, $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control, respectively. The Average Treatment Effect (ATE) (Glymour et al., 2016) is

$$ATE = E\big[Y(1)\big] - E\big[Y(0)\big]. \tag{1}$$

We focus on binary treatment and outcome settings, following common empirical scenarios (Curth and van der Schaar, 2021).

Confounders. A confounder X influences both T and Y. As shown in Figure 2, the presence of X can bias causal effect estimation unless properly controlled.

Ignorability Assumption. Given covariate X, treatment assignment T is independent of the potential outcomes $Y(1)$ and $Y(0)$, i.e. $T \perp \{Y(1), Y(0)\} \mid X$ (Yao et al., 2021). Under

this assumption, $E[Y(1) \mid X]$ can be estimated by $E[Y \mid T = 1, X]$ (and similarly for $Y(0)$), enabling causal inference from observational data when X captures all back-door paths.

Back-door Criterion. In a Directed Acyclic Graph (DAG), a set of covariates X satisfies the back-door criterion if (i) no element of X is a descendant of T; and (ii) X blocks every path from T to Y that begins with an arrow pointing into T. Adjusting for such X emulates randomized conditions and provides unbiased estimates of the treatment effect (Pearl, 2009).

Hoeffding's Inequality. To derive sampling boundaries, we use Hoeffding's inequality (Mohri et al., 2012), which bounds the probability that the average of bounded i.i.d. random variables deviates from its expectation. Let $x_1, x_2, \ldots, x_n$ be i.i.d. in $[0, 1]$. For any $\epsilon > 0$,

$$P\Big(\big|\tfrac{1}{n}\sum_{i=1}^{n} x_i - \tfrac{1}{n}\sum_{i=1}^{n} E[x_i]\big| \geq \epsilon\Big) \;\leq\; 2\,\exp\big(-2\,n\,\epsilon^2\big). \tag{2}$$

Even under sampling without replacement, Hoeffding's bound remains valid (although the random variables are not strictly independent, the resulting bound is effectively similar). We apply this inequality to establish probabilistic guarantees on our causal effect estimates.

3.2. Problem Formulation

We first introduce key metrics used in our analysis:

Population Average Treatment Effect (PATE): The true Average Treatment Effect (ATE) across the entire dataset D, which represents the overall effect of the treatment across the population.

Sample Average Treatment Effect (SATE): The ATE estimated from a sampled subset D', reflecting the treatment effect within the sample.

PAC-estimable causal effect. Let ε and $1 - \delta$ denote the given error threshold and confidence level, respectively. We consider the causal effect of T on Y to be PAC-estimable on D' if the probability that the difference between $SATE_{D'}$ and $PATE_D$ falls within ε is at least $1 - \delta$, expressed as:

$$P\left(|SATE_{D'} - PATE_D| \leq \varepsilon\right) \geq 1 - \delta \tag{3}$$

This estimation is inherently random due to its dependence on the particular subset of data sampled. We assume D' is randomly sampled from D, i.e., following the typical independent identically distributed (i.i.d.) sampling approach. Concretely, $D' = \{(T_i, X_i, Y_i)\}_{i=1}^{n}$ is an i.i.d. sample of observable triples, with $Y_i(1), Y_i(0)$ remaining potential outcomes. Let n denote the size of D'. The challenge addressed in this paper is to determine the sampling upper boundary N such that if $n \geq N$, the treatment effect of T on Y is PAC-estimable on D', i.e., the following equation is satisfied:

$$P\left(|SATE_{D'} - PATE_D| \leq \varepsilon \mid n \geq \mathrm{N}\right) \geq 1 - \delta \tag{4}$$

Note that:

$$SATE_{D'} = ATE_{D'} = E_{D'}[Y(1)] - E_{D'}[Y(0)], PATE_D = ATE_D = E_D[Y(1)] - E_D[Y(0)]. \tag{5}$$

We have the following conversion:

$$\begin{aligned} |SATE_{D'} - PATE_D| &= \big|E_{D'}\big[Y(1)\big] - E_{D'}\big[Y(0)\big] - E_D\big[Y(1)\big] + E_D\big[Y(0)\big]\big| \\ &= \big|E_{D'}\big[Y(1)\big] - E_D\big[Y(1)\big] + E_D\big[Y(0)\big] - E_{D'}\big[Y(0)\big]\big| \\ &\leq \big|E_{D'}\big[Y(1)\big] - E_D\big[Y(1)\big]\big| + \big|E_{D'}\big[Y(0)\big] - E_D\big[Y(0)\big]\big|. \end{aligned} \tag{6}$$

Therefore, if

$$|E_{D'}[Y(1)] - E_D[Y(1)]| \leq \frac{\varepsilon}{2} \quad \text{and} \, |E_{D'}[Y(0)] - E_D[Y(0)]| \leq \frac{\varepsilon}{2}, \tag{7}$$

we have

$$|SATE_{D'} - PATE_D| \leq \varepsilon. \tag{8}$$

Because $E_{D'}[Y(1)]$ and $E_{D'}[Y(0)]$ are each sample means of n i.i.d. bounded observations, Hoeffding's inequality yields, for $t \in \{0, 1\}$,

$$P\Big(\big|E_{D'}[Y(t)] - E_D[Y(t)]\big| \geq \varepsilon/2\Big) \leq 2\exp\big(-\tfrac{1}{2}n\varepsilon^2\big). \tag{9}$$

Let $a = E_{D'}[Y(1)] - E_D[Y(1)]$ and $b = E_{D'}[Y(0)] - E_D[Y(0)]$. Allocating probability $\delta/2$ to each tail event gives

$$P\Big(|a| \leq \varepsilon/2 \ \wedge \ |b| \leq \varepsilon/2\Big) \ \geq \ 1 - \delta, \tag{10}$$

which, together with the triangle inequality above, implies $|SATE_{D'} - PATE_D| \leq \varepsilon$ with the same confidence. Consequently it suffices to require:

$$P\left(|E_{D'}[Y(1)] - E_D[Y(1)]| \leq \varepsilon/2 \mid n \geq \mathrm{N}\right) \geq 1 - \delta. \tag{11}$$

4. Sampling Upper Boundary Derivation

This section explores the derivation of sampling upper boundaries under three distinct scenarios, each tailored to different types of confounding variables. Throughout our derivation, we assume that each sample (T_i, X_i, Y_i) is drawn independently and identically (i.i.d.) from the population dataset D, and that the random variables are bounded (e.g., taking values within $[0, 1]$). These assumptions are necessary for applying Hoeffding's inequality and establishing probabilistic bounds on the estimation errors.

4.1. No Confounders Case

In this section, we begin by considering an idealized scenario where there are no confounders between T and Y. This approach not only aids in clearly demonstrating our analytical framework, but also lays the groundwork for understanding more complex situations involving confounders within finite and infinite hypothesis spaces. Assume $Y_i(1)$ is the potential outcome of the sample $i - th$, then $E_{D'}[Y(1)] = \frac{1}{n}\sum_{i=1}^{n} Y_i(1)$. When there are no confounders, the assumption of ignorability is satisfied with the covariate set as null. This is usually done in the experimental setting by directly assigning $T = 1$ to the $i - th$ sample. Then we have:

Theorem 4.1. If $n \geq N = 2\ln(2/\delta)/\varepsilon^2$, Equation (9) is satisfied when there are no confounders.

Proof The expectation of $E[Y_i(1)]$ could be obtained by the average of the potential outcome on the whole population, i.e., $E[Y_i(1)] = E_D[Y(1)]$. Therefore, according to Hoeffding's inequality, we have:

$$\begin{aligned} P\Big(|E_{D'}[Y(1)] - E_D[Y(1)]| \le \frac{\varepsilon}{2}\Big) &= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - E_D[Y(1)]\right| \le \frac{\varepsilon}{2}\right) \\ &= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} E\big[Y_i(1)\big]\right| \le \frac{\varepsilon}{2}\right) \\ &\ge 1 - 2\exp\left(-\frac{n\varepsilon^2}{2}\right). \end{aligned} \tag{12}$$

Let $1 - 2\exp\left(-\frac{1}{2}n\varepsilon^2\right) \ge 1 - \delta$, then $n \ge \frac{2\ln(2/\delta)}{\varepsilon^2}$. ■

According to Theorem 4.1, we need at least $2\ln(2/\delta)/\varepsilon^2$ observations to estimate a single potential–outcome mean within $\varepsilon/2$ at confidence level $1-\delta$. Let each unit be randomly assigned to the treatment group with probability $p \in (0,1)$ and to the control group with probability $1-p$. Because both groups must contain no fewer than N observations, the total sample size must satisfy the following:

$$n \ge \frac{N}{\min\{p, 1-p\}} = \frac{2\,\log(2/\delta)}{\varepsilon^2\,\min\{p, 1-p\}}. \tag{13}$$

In a balanced randomisation scheme ($p = 0.5$), this reduces to $n \ge 4\ln(2/\delta)/\varepsilon^2$

4.2. Finite Hypothesis Space Case

When dealing with confounders, the covariate (or set) X is utilized to determine the potential outcomes $Y_i(1)$ and $Y_i(0)$ for each i-th sample. As introduced earlier, various methods establish a mapping from (T, X) to Y. In many practical scenarios—for instance, when X consists of discrete variables with a limited number of categories—the number of such mappings is inherently finite. We refer to this collection of functions as the finite hypothesis space. Under this assumption, we now formalize the notion of $\mathcal{H}$ and derive the sampling boundary specifically for the finite hypothesis setting. Because $\mathcal{H}$ is of finite size, we can leverage Hoeffding's inequality to establish probabilistic guarantees on the estimation error in causal effect studies with confounders.

Definition 4.1. Hypothesis space (Shalev-Shwartz and Ben-David, 2014). The hypothesis space, denoted as $\mathcal{H}$, includes all possible mappings or functions linking the combined space of treatment variables and covariates (T, X) to the outcome variable Y. In this space, each distinct function, represented as h, acts as a unique estimator for the outcome based on specified covariate and treatment values, and is thus termed a hypothesis.

To elaborate, $\mathcal{H}$ includes a variety of functional forms that can model the relationship between treatment and outcome under different conditions. For example, it encompasses functions like the conditional expectation $h(t, x) = E[Y_i|T_i = t, X_i = x]$. Building on this framework, the potential outcome when treatment is applied, denoted $Y_i(1)$, can be estimated by $h(T_i = 1, x_i)$, succinctly referred to as $h(1, x_i)$. We introduce the following theorem for a finite hypothesis space.

Theorem 4.2. If the sample size n satisfies

$$n \geq N = \frac{2\ln(2|\mathcal{H}|/\delta)}{\varepsilon^2}, \tag{14}$$

the inequality

$$P\left(|E_{D'}[Y(1)] - E_D[Y(1)]| \leq \frac{\varepsilon}{2} \mid n \geq N\right) \geq 1 - \delta \tag{15}$$

holds. Here, $|\mathcal{H}|$ represents the size of a finite hypothesis space. It is the number of combinations of all possible values of X (Mohri et al., 2012).

Before proceeding, it is essential to note that the functions in $\mathcal{H}$ inherently produce outcomes within a bounded range, a necessary condition for the application of Hoeffding's inequality (Hoeffding, 1994). This boundedness assumption is consistent with common empirical scenarios where outcome variables are naturally restricted by either physical limits or by design of the study.

Proof To ensure $|E_{D'}[Y(1)] - E_D[Y(1)]| \leq \frac{\varepsilon}{2}$ for each hypothesis $h \in \mathcal{H}$, we consider:

$$\begin{aligned}
&P\left(\forall h \in \mathcal{H} : |E_{D'}[Y(1)] - E_D[Y(1)]| \leq \frac{\varepsilon}{2}\right) \\
&= P\left(\forall h \in \mathcal{H} : \left|\frac{1}{n}\sum_{i=1}^{n} h\left(1, x_i\right) - \frac{1}{n}\sum_{i=1}^{n} E\left[h\left(1, x_i\right)\right]\right| \leq \frac{\varepsilon}{2}\right) = 1 - P\left(\exists h \in \mathcal{H} : \left|\frac{1}{n}\sum_{i=1}^{n} h\left(1, x_i\right) - \frac{1}{n}\sum_{i=1}^{n} E\left[h\left(1, x_i\right)\right]\right| \geq \frac{\varepsilon}{2}\right) \\
&\geq 1 - P\left(\left(\left|\frac{1}{n}\sum_{i=1}^{n} h_1\left(1, x_i\right) - \frac{1}{n}\sum_{i=1}^{n} E\left[h_1\left(1, x_i\right)\right]\right| \geq \frac{\varepsilon}{2}\right) \vee \left(\left|\frac{1}{n}\sum_{i=1}^{n} h_2\left(1, x_i\right) - \frac{1}{n}\sum_{i=1}^{n} E\left[h_2\left(1, x_i\right)\right]\right| \geq \frac{\varepsilon}{2}\right)\right. \\
&\left.\vee \ldots \vee \left(\left|\frac{1}{n}\sum_{i=1}^{n} h_{|\mathcal{H}|}\left(1, x_i\right) - \frac{1}{n}\sum_{i=1}^{n} E\left[h_{|\mathcal{H}|}\left(1, x_i\right)\right]\right| \geq \frac{\varepsilon}{2}\right)\right) \geq 1 - 2|\mathcal{H}| \exp\left(-\frac{1}{2}n\varepsilon^2\right)
\end{aligned} \tag{16}$$

Setting $1 - 2|\mathcal{H}| \exp\left(-\frac{1}{2}n\varepsilon^2\right) \geq 1 - \delta$, we find $n \geq \frac{2\ln(2|\mathcal{H}|/\delta)}{\varepsilon^2}$. ■

In Theorem 4.2, we demonstrates the method to compute the minimum required sample size to ensure the reliability of causal effect estimates under the condition of a finite hypothesis space, considering the presence of confounders.

4.3. Infinite Hypothesis Space Case

Given that X is a continuous variable, the hypothesis space might potentially expand to infinite proportions. Consequently, gauging the complexity of this space becomes pivotal. A prevalent method in this context is to consider the 'VC dimension' of the hypothesis space, a metric that serves as an indicator of the learning capability of a function set. Notably, a larger VC dimension signals the possibility of learning a more intricate set of models.

Theorem 4.3. If $n \geq N$ and N satisfies $\frac{\varepsilon}{2} = \sqrt{\frac{8d\ln\frac{2eN}{d} + 8\ln\frac{4}{\delta}}{N}}$, equation (9) is satisfied when there are confounders with an infinite hypothesis space $\mathcal{H}$ of VC dimension d.

Proof Central to the proof is the concept of the growth function $m_{\mathcal{H}}(n)$ of the hypothesis space $\mathcal{H}$. This function indicates the maximum number of labels that $\mathcal{H}$ can assign to any given set of n samples, reflecting the complexity of the hypothesis space. Its upper bound is 2^n.

The forthcoming discussion provides a succinct proof of the theorem. For a detailed exposition of the proof, the reader is referred to Section 1 of the Supplementary material.

Drawing from the literature, specifically (Vapnik and Chervonenkis, 2015), for any $h \in \mathcal{H}$, the following inequality holds:

$$P\left(|E_{D'}[Y(1)] - E_D[Y(1)]| \geq \frac{\varepsilon}{2}\right) \leq 4m_{\mathcal{H}}(2n)\exp\left(-\frac{1}{32}n\varepsilon^2\right) \tag{17}$$

Here, the VC dimension, often referred to as the capacity to 'shatter', signifies the largest dataset size that the hypothesis space $\mathcal{H}$ can shatter. Consequently:

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{\mathrm{d}} \binom{n}{i} \leq \left(\frac{e \cdot n}{d}\right)^d \tag{18}$$

According to equations (11) and (12), for $\forall h \in \mathcal{H}$, we have

$$P\left(|E_{D'}[Y(1)] - E_D[Y(1)]| \leq \frac{\varepsilon}{2}\right) \geq 1 - 4\left(\frac{2en}{d}\right)^d exp\left(-\frac{1}{32}n\varepsilon^2\right) \tag{19}$$

Let $1 - 4\left(\frac{2en}{d}\right)^d exp\left(-\frac{1}{32}n\varepsilon^2\right) \geq 1 - \delta$, N should satisfy $\frac{\varepsilon}{2} = \sqrt{\frac{8d\ln\frac{2eN}{d} + 8\ln\frac{4}{\delta}}{N}}$. ■

In Theorem 4.3, we utilize Hoeffding's inequality to ensure the accuracy of the average causal effect, and represent the complexity of the hypothesis space through the growth function and VC dimension, eventually deducing a reliable upper boundary for the sample size. The size of the VC dimension is not influenced by the learning algorithm used, the specific distribution of the dataset, or the objective function being analyzed; it is determined by the model for estimating the causal effect and the defined hypothesis space. Moreover, there is generally a direct correlation between the VC dimension of the hypothesis space and the number of free variables in a hypothesis (Shalev-Shwartz and Ben-David, 2014). We will further delve into this relationship in Section 3 of the Supplementary material.

4.4. Data Augmentation

Although the preceding theoretical analysis holds when the sample size is sufficiently large, practical constraints—such as high costs or limited accessibility—often make it difficult to collect additional real-world data. Consequently, it is not uncommon to encounter datasets that fail to meet the theoretical sampling requirements, thus limiting our ability to make accurate causal effect estimates.

Building on Theorem 4.4, which indicates that the error of estimated average treatment effects is bounded by the accuracy of $E[Y \mid T, X]$, we propose leveraging a well-trained machine-learning model to generate synthetic samples for data augmentation. This approach effectively enlarges the dataset while preserving reliable estimates of $E[Y \mid T, X]$, thereby reaffirming the viability of data augmentation for improving average causal effect estimation when sample sizes are inadequate.

Theorem 4.4. For any arbitrary values of x and z, as long as they satisfy the condition $|E_{D'}[y|t,x] - E_D[y|t,x]| \leq \varepsilon$, it follows that $|SATE_{D'} - PATE_D| \leq 2\varepsilon$ holds true.

In other words, from $\lim_{|D'|\to\infty} E_{D'}[y|t,x] \to E_D[y|t,x]$, we can infer $\lim_{|D'|\to\infty} SATE_{D'} \to PATE_D$. Here, D represents a sufficient large dataset, and the observed dataset D' is a

subset of D. Theorem 4.4. states that if the estimation error of $E[y|t,x]$ is less than or equal to ε, then the resulting estimation error of the causal effect is not greater than 2ε.

Assumption For any t, the mean value calculated by all training sets $E_{D'\subset D}[y|t,x]$ converges unbiasedly to the actual $E_D[y|t,x]$, where E_D denotes the expectation in D.

Proof Given that the data in the dataset is in the form of (T,X,Y). Our goal is to train a machine Q, which given an input (t,x), outputs a probability p that represents the probability of y being labeled as 1.

Q is a machine that learns from train data, and $Q(y|t,x)$ is the distribution of the output of the machine Q in the training set D'. Theoretically, comparing the actual distributions in $Q(y|t,x)$ and D' obtained from the calculation shows the error between the two. In practice, however, it is not possible to directly observe the corresponding distribution of y due to the small amount of data under the same covariate.

According to the do operation calculation formula,

$$E[Y|do(t=1)] = \sum_x E[Y|t=1,x]P(x) \tag{20}$$

x is a backdoor variable, so

$$\begin{aligned} |E_{D'}[Y \mid do(t=1)] - E_D[Y \mid do(t=1)]| &= \left| \sum_x E_{D'}[Y \mid t=1,x]P(x) - \sum_x E_D[Y \mid t=1,x]P(x) \right| \\ &= \left| \sum_x \left[E_{D'}[Y \mid t=1,x] - E_D[Y \mid t=1,x]\right] P(x) \right| \leq \varepsilon, \end{aligned} \tag{21}$$

Next, consider the error between $SATE_{D'}$ and $PATE_D$,

$$\begin{aligned} &|SATE_{D'} - PATE_D| \\ &= |(E_{D'}[Y|\mathrm{do}(x=1)] - E_{D'}[Y|\mathrm{do}(x=0)]) - (E_D[Y|\mathrm{do}(x=1)] - E_D[Y|\mathrm{do}(x=0)])| \\ &\leq |E_{D'}[Y|\mathrm{do}(x=1)] - E_D[Y|\mathrm{do}(x=1)]| + |E_D[Y|\mathrm{do}(x=0)] - E_{D'}[Y|\mathrm{do}(x=0)]|. \end{aligned} \tag{22}$$

Due to the condition $|E_{D'}[Y|do(x=1)] - E_D[Y|do(x=1)]| \leq \varepsilon$, and similarly, $|E_{D'}[Y|do(x=0)] - E_D[Y|do(x=0)]| \leq \varepsilon$, it then follows that $|SATE_{D'} - PATE_D| \leq 2\varepsilon$. ■

Theorem 4.4 furnishes a solid theoretical assurance that given the estimation error of $E[y|t,x]$ is at most ε, the error of the estimated ATE on the augmented dataset will not exceed 2ε, thus validating the rationality of the simulated samples. The elaboration on Theorem 4.4 is in Section 2 of the Supplementary Material.

In the upcoming section 5.3, we utilize the Dragonnet deep learning Shi et al. (2019) model to augment the datasets. Subsequently, the propensity score matching method will be employed to estimate the ATE on both simulated datasets and real IHDP datasets.

5. Experiments

This section validates our sampling boundary approach using both simulated and real-world datasets. We compare the estimated causal effects from sampled data to the theoretical

causal effects and evaluate how well the results align with different error thresholds ε and confidence levels $(1-\delta)$. The detailed steps for constructing the simulation dataset, the specific parameters used in data generation, and further information about the benchmark IHDP dataset are provided in Section 4 of the Supplementary material.

5.1. Experimental Setup

We evaluate estimation accuracy at three tolerance thresholds, $\varepsilon \in \{0.05, 0.025, 0.01\}$, under two confidence levels: $1-\delta \approx 95.45\%\,(2\sigma)$ and $1-\delta \approx 99.73\%\,(3\sigma)$. For each setting, we record the probability that $|SATE_{D'} - PATE_D| \leq \varepsilon$. When this probability exceeds $1-\delta$, the causal-effect estimate is deemed PAC-estimable.

All experiments are run with the Python library `DoWhy` (Sharma and Kiciman, 2020), which identifies causal graphs and computes the Average Treatment Effect (ATE) through a unified estimator. For every theoretical sampling boundary, we draw 10000 bootstrap samples and count how often the estimation error falls below ε; comparing this empirical success rate with the nominal confidence level verifies whether the derived boundary is sufficiently conservative.

5.2. Sampling Boundary Performance

Using the thresholds and confidence levels defined in Section 5.1, we first verify the no-confounder boundary (Section 4.1). For each candidate sample size n, we compute the theoretical bound N from Theorem 4.1, run 10000 Monte-Carlo draws, and record the smallest n_{emp} for which $|SATE_{D'} - PATE_D| \leq \varepsilon$ with probability at least $1-\delta$.

Table 2: Comparison of theoretical N vs. empirical n_{emp} sample sizes (no-confounder scenario). Confidence levels: 2σ (95.45%) and 3σ (99.73%).

ε	2σ		3σ	
	N	n_{emp}	N	n_{emp}
0.05	3027	1858	5286	4193
0.025	12106	7509	21144	17427
0.01	75664	47165	132153	97340

Table 2 contrasts the theoretical N with the empirical n_{emp}. The latter is always smaller, confirming the conservativeness of the bound. Both N and n_{emp} increase as ε tightens or the confidence level rises from 2σ to 3σ, showing that the boundary scales with stricter requirements.

Table 3: Required sample sizes under different numbers of confounders C, for error thresholds ε and confidence levels $(1-\delta)$. As C increases, more samples are needed to achieve the same constraints.

ε	$1-\delta$	$C=1$	$C=2$	$C=5$	$C=10$
0.05	95.45%	3 581	4 136	5 799	8 572
	99.73%	5 841	6 395	8 059	10 831
0.025	95.45%	14 324	16 542	23 197	34 287
	99.73%	23 363	25 581	32 235	43 325
0.01	95.45%	89 527	103 390	144 979	214 293
	99.73%	146 016	159 879	201 468	270 782

Next, we consider confounders under a finite hypothesis space (Section 4.2). As analyzed, increasing the number of confounders C generally enlarges the hypothesis space and demands more samples. Table 3 illustrates how the required sample size grows for $C \in \{1, 2, 5, 10\}$, matching the trend predicted by Theorem 4.2.

To further assess our boundaries under different estimation methods, we varied C, ε, and $(1-\delta)$, and then computed the empirical success rates for four common estimators (Backdoor Criterion, Propensity Score Weighting, Propensity Score Stratification, and Double Machine Learning). The detailed probability results are listed in Section 5 of the Supplementary material. In brief, each method surpassed the nominal confidence levels across all tested conditions, confirming that our theoretical boundaries remain conservative and robust.

Figure 3 depicts the mean error of causal effect estimation for different C, showing that when $\varepsilon = 0.05$, the average error can be as low as 0.016 under the tested simulation settings. This aligns with the conclusions drawn from both the theoretical sample sizes and the empirical success rates, illustrating that our sampling boundaries effectively scale to accommodate growing complexity.

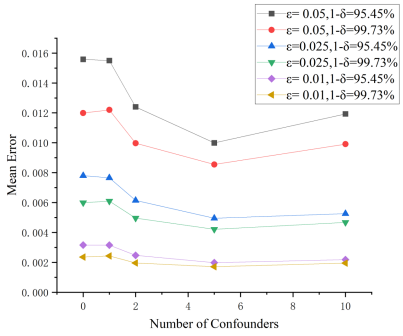

Figure 3: Mean error of causal effect estimation for different numbers of confounders $(0, 1, 2, 5, 10)$. Each point averages 10 000 trials.

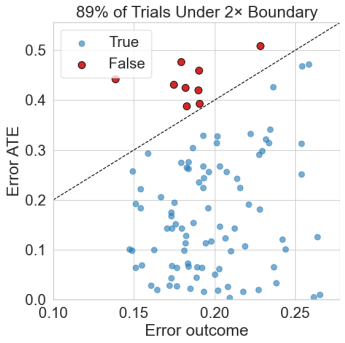

Figure 4: Scatter of outcome-prediction error (x-axis) vs. ATE error (y-axis) over 100 IHDP runs. The dashed line $y = 2x$ marks the two-fold boundary.

5.3. Experiments on Data Augmentation

In practice, observational datasets often fall short of the theoretical sampling boundary from Section 4. We address this issue by employing data augmentation to expand the dataset accordingly. Before implementing augmentation, we validate Theorem 4.4 by comparing the outcome-prediction error and the corresponding ATE error across 100 independent runs on the IHDP dataset.

As shown in Figure 4, an 89% compliance rate (i.e., error ate $\leq 2 \times$ error outcome) was observed. The one-sample Wald test rejects a 50% null compliance rate with $z \approx 8.3$ ($p < 10^{-15}$) indicating that such a success rate can not be coincidental. In these compliant runs, error ate/error outcome averages around 1.35, whereas in trials exceeding the twofold boundary, small offsets in predicting treatment vs. control outcomes can inflate the final ATE error. Data augmentation reduces the mean ATE error from about 0.20 to 0.12,

underscoring how broader covariate coverage mitigates predictive biases and brings more trials under the theoretical boundary.

Next, we create a reduced dataset D (half the theoretical boundary), train a DragonNet Shi et al. (2019) model to generate synthetic covariates, and merge D with these synthetic samples to form D_{aug}. Figures 5 and 6 illustrate the resulting improvements in estimation accuracy under two different settings.

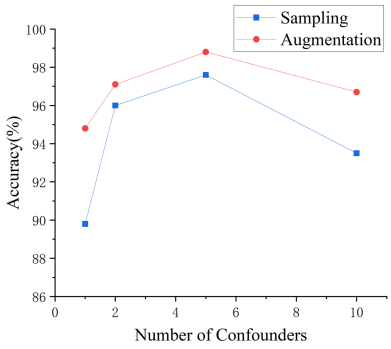

Figure 5: With an error boundary $\mathcal{E} = 0.05$ and confidence level $1 - \delta = 95.45\%$, the accuracy of causal-effect estimation using the Back-door Criterion over 1000 runs.

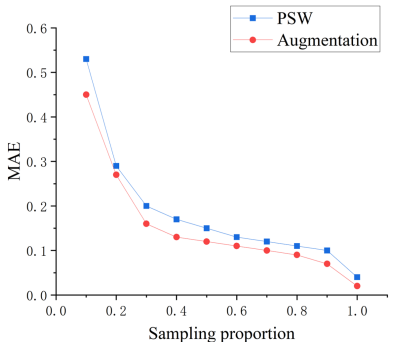

Figure 6: Mean absolute error (MAE) of ATE estimation on the IHDP dataset for different sampling proportions p.

In Figure 5, we focus on the proportion of runs satisfying $|\widehat{\text{ATE}} - \text{ATE}_{\text{true}}| \leq \mathcal{E}$, and observe that adding synthetic samples consistently improves performance. Figure 6 reports the mean absolute error for varying sampling proportions p, reinforcing that augmentation effectively mitigates data scarcity by reducing error and enhancing reliability.

6. Discussion and Conclusion

This introduces a novel perspective on causal effect estimation grounded in the concept of PAC-estimable causal effects. We derive sample-size boundaries for three settings—no-confounder, finite-hypothesis, and infinite-hypothesis spaces—providing reliability guarantees across problem complexities. Furthermore, we established constraints linking expected outcomes and average treatment effects, providing a theoretical basis for leveraging data augmentation when sample sizes are insufficient. Our experiments on both the simulated datasets and the IHDP dataset, under different error thresholds and significance levels, validate the robustness of these sampling boundaries.

Despite these contributions, several limitations remain. First, the derived boundaries are primarily suited to PAC-learnable causal models; for infinite hypothesis spaces, the VC-dimension-based bounds can be loose, making it challenging to attain the required sample size in practice. Second, although our framework can accommodate various causal estimation methods (X-Learner, Dragonnet, CEVAE, etc.), we have not examined the specific error metrics each algorithm may exhibit. Future work will refine these boundaries by considering specific data distributions, and methodological assumptions, thereby enhancing both theoretical rigor and real-world applicability.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (No. 62072153), the Anhui Provincial Key Technologies R&D Program (No. 2022h11020015), and the 111 Center (No. B14025).

References

Joshua D Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ, 2009.

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

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media, Heidelberg, Germany, 2011.

Anna Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021.

Nicole Didero, Marco Costanigro, and Becca BR Jablonski. Promoting farmers market via information nudges and coupons: A randomized control trial. *Agribusiness*, 37(3): 531–549, 2021. doi: 10.1002/agr.21688.

David A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

Amanda M Gentzel, Purva Pruthi, and David Jensen. How and why to use experimental data to evaluate methods for observational causal inference. In *International Conference on Machine Learning*, pages 3660–3671. PMLR, 2021.

Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, Hoboken, NJ, 2016.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4): 1–37, 2020. doi: 10.1145/3397269.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The Collected Works of Wassily Hoeffding*, pages 409–426, 1994. doi: 10.1007/978-1-4612-0865-5_26.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, Cambridge, UK, 2015.

ER John, KR Abrams, CE Brightling, and NA Sheehan. Assessing causal treatment effect estimation when using large observational datasets. *BMC medical research methodology*, 19(1):1–15, 2019. doi: 10.1186/s12874-019-0858-x.

Rex B Kline. *Principles and Practice of Structural Equation Modeling*. Guilford Publications, New York, NY, 2023.

André Kretzschmar and Gilles E Gignac. At what sample size do latent variable correlations stabilize? *Journal of Research in Personality*, 80:17–22, 2019.

Michael Lechner. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224, 2011. doi: 10.1561/0800000014.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010. doi: 10.1257/jel.48.2.281.

Yifan Liu, Jian Wang, and Bo Li. Edvae: Disentangled latent factors models in counterfactual reasoning for individual treatment effects estimation. *Information Sciences*, 652: 119578, 2024.

Andreas Markoulidakis, Peter Holmans, Philip Pallmann, Monica Busse, and Beth Ann Griffin. How balance and sample size impact bias in the estimation of causal treatment effects: a simulation study. *arXiv preprint arXiv:2107.09009*, 2021.

Marie C McCormick, Cecelia McCarton, Jeanne Brooks-Gunn, et al. The infant health and development program: Interim summary. *Journal of Developmental & Behavioral Pediatrics*, 19(5):359–370, 1998.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning.[sl], 2012.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009. doi: 10.1214/09-SS057.

Richard D Riley, Joie Ensor, Kym IE Snell, Frank E Harrell, Glen P Martin, Johannes B Reitsma, Karel GM Moons, Gary Collins, and Maarten Van Smeden. Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368, 2020. doi: 10.1136/bmj.m441.

Paul R Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987. doi: 10.1080/01621459.1987.10478458.

Robert E. Schapire and Yoav Freund. Boosting: Foundations and algorithms. *Kybernetes*, 42(1):164–166, 2013.

Philipp Schwab, Lukas Linhardt, Stefan Bauer, et al. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, Cambridge, 2014.

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

Suresh Kumar Sharma, Shiv Kumar Mudgal, Kalpana Thakur, and Rakhi Gaur. How to calculate sample size for observational and experimental nursing research studies. *National Journal of Physiology, Pharmacy and Pharmacology*, 10(1):1–8, 2020. doi: 10.5455/njppp.2020.10.0930717102019.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Hamed Taherdoost. Determining sample size; how to calculate survey sample size. *International Journal of Economics and Management Systems*, 2, 2017.

Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, USA, 1995.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Measures of complexity: festschrift for Alexey Chervonenkis*, pages 11–30, 2015. doi: 10.1007/978-3-319-21852-6_2.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. In *Advances in Neural Information Processing Systems*, volume 36, pages 5404–5418, 2023.

Hao Wang, Zhichao Chen, Zhaoran Liu, Xu Chen, Haoxuan Li, and Zhouchen Lin. Proximity matters: Local proximity enhanced balancing for treatment effect estimation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, page 2927–2937, 2025.

Erika J Wolf, Kelly M Harrington, Shaunna L Clark, and Mark W Miller. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and psychological measurement*, 73(6):913–934, 2013. doi: 10.1177/0013164413495237.

Jiao-Yun Yang, Jun-Da Wang, Yi-Fang Zhang, Wen-Juan Cheng, and Lian Li. A heuristic sampling method for maintaining the probability distribution. *Journal of Computer Science and Technology*, 36:896–909, 2021. doi: 10.1007/s11390-020-0065-6.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5): 1–46, 2021. doi: 10.1145/3444944.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12207–12215, 2021.