Supplementary material for 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

1. Proof of Sampling Boundaries for Infinite Hypothesis Space Cases

This section delves into the detailed proof of deriving sampling boundaries within the context of an infinite hypothesis space. We focus on the scenario where the confounders are modeled as continuous variables, necessitating a comprehensive measure of the hypothesis space's complexity. To this end, we primarily explore the concept of the Vapnik-Chervonenkis (VC) dimension and its implications for the hypothesis space.

Proof The hypothesis space could be infinite when X is a continuous variable. In this case, we need to measure the complexity of the hypothesis space, and the most common way is to consider the 'VC dimension' of the hypothesis space Mohri et al. (2012). Before introducing the VC dimension, we introduce several related concepts: growth function, dichotomy, and shatter. It is usually calculated in this way: if there exists a sample set of size d that can be shattered by $\mathcal{H}$, but there does not exist any sample sets of size $d+1$ that can be shattered by $\mathcal{H}$, then the VC dimension of $\mathcal{H}$ is d.

Definition A.1. Growth function. The growth function $\Pi_{\mathcal{H}}(n)$ of the hypothetical space, $\mathcal{H}$ is

$$\Pi_{\mathcal{H}}(n) = max|\{\big(h(t_1, x_1), ..., h(t_n, x_n)\big)|h \in \mathcal{H}\}| \tag{1}$$

The growth function $\Pi_{\mathcal{H}}(n)$ represents the maximum number of possible outcomes that the hypothesis space $\mathcal{H}$ can assign to n samples, where the upper boundary is 2^n due to Y being a bivariate.

Definition A.2. Dichotomy. For the hypothetical space $\mathcal{H} = \{h\colon \{T, X\} \rightarrow Y\}, Y \in \{0, 1\}$, a dichotomy is an estimation of Y by h on all samples, i.e., $\{h(t_1, x_1), ..., h(t_n, x_n)\}$.

Definition A.3. Shatter. D' is said to be shattered by $\mathcal{H}$ when the hypothesis space $\mathcal{H}$ can produce all possible dichotomies of D' with size n, i.e., $\Pi_{\mathcal{H}}(n) = 2^n$.

Now we can formally define the VC dimension.

Definition A.4. VC dimension. The VC dimension of the hypothesis space $\mathcal{H}$ is the size of the largest set of examples that can be shattered by $\mathcal{H}$,

$$VC(\mathcal{H}) = max\{n\colon \Pi_{\mathcal{H}}(n) = 2^n\} \tag{2}$$

The VC dimension reflects the learning ability of the hypothesis, and a larger VC dimension denotes a more complex function that can be learned. In general, the VC dimension of the hypothesis space is approximately equal to the number of free variables [20, 28, 29]. It is usually calculated in this way: if there exists a sample set of size d that can be shattered by $\mathcal{H}$, but there does not exist any sample sets of size $d+1$ that can be shattered by $\mathcal{H}$, then the VC dimension of $\mathcal{H}$ is d. The VC dimension of all linear functions is $d = k+1$, k denotes the number of variables.

Theorem A.1. 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 A.1. According to Vapnik and Chervonenkis (2015), for $\forall h \in \mathcal{H}$, we have

$$\begin{aligned} & P\left(\left|E_{D'}[Y(1)] - E_D[Y(1)]\right| \geq \frac{\varepsilon}{2}\right) \\ =& P\left(\left|\frac{1}{n}\sum_{i=1}^{n} h\left(1, x_i\right) - \frac{1}{n}\sum_{i=1}^{n} E\left[h\left(1, x_i\right)\right]\right| \geq \frac{\varepsilon}{2}\right) \\ \leq& 4\Pi_{\mathcal{H}}(2n) \exp\left(-\frac{1}{32} n\varepsilon^2\right). \end{aligned} \tag{3}$$

According to Mohri et al. (2012), the growth function $\Pi_{\mathcal{H}}(n)$ can be bounded as follows:

$$\Pi_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d. \tag{4}$$

Substituting Equation (4) into the inequality, we get:

$$P\left(\left|E_{D'}[Y(1)] - E_D[Y(1)]\right| \geq \frac{\varepsilon}{2}\right) \leq 4\left(\frac{2en}{d}\right)^d \exp\left(-\frac{1}{32} n\varepsilon^2\right). \tag{5}$$

Hence, the probability that the estimation error is within the specified threshold is:

$$P\left(\left|E_{D'}[Y(1)] - E_D[Y(1)]\right| \leq \frac{\varepsilon}{2}\right) \geq 1 - 4\left(\frac{2en}{d}\right)^d \exp\left(-\frac{1}{32} n\varepsilon^2\right). \tag{6}$$

Setting the right side of Equation (6) to be greater than or equal to $1-\delta$, we derive the condition for N:

$$\frac{\varepsilon}{2} = \sqrt{\frac{8d \ln\left(\frac{2eN}{d}\right) + 8 \ln\left(\frac{4}{\delta}\right)}{N}}. \tag{7}$$

■

2. Enhancing Data Augmentation: A Focus on Theorem 4.4

This section serves as an elaboration on Theorem 4.4, specifically addressing the augmentation process's impact on estimation accuracy and confidence. Theorem 4.4 plays a pivotal role in assessing the feasibility of enhancing the accuracy of ATE estimations on an augmented dataset.

The core of our discussion revolves around the confidence in estimations derived from augmented datasets. The key insight from Theorem 4.4 lies in its ability to link the accuracy of augmented dataset estimations directly to the confidence in the estimation of $E[Y|T,X]$. Specifically, if we can assert with confidence $1-\delta$ that the error in estimating $E[Y|T,X]$ on the augmented dataset D' is within a bound ε, i.e., $P(|E_{D'}[y|t,x] - E_D[y|t,x]| \leq \varepsilon) \geq 1-\delta$, then it follows that the accuracy of ATE estimations on D' is similarly bounded, with $P(|\text{SATE}_{\text{D}'} - \text{PATE}_{\text{D}}| \leq 2\varepsilon) \geq 1-\delta$.

This assertion is grounded in the principle of backdoor adjustment, where the ATE is conceptualized as a weighted sum of conditional expectations of Y. Importantly, the confidence in estimating $E[Y|T,X]$ and by extension, ATE relies not on the size of the augmented dataset, but on the dataset size utilized for training the predictive model.

A crucial consideration is the nature of the augmented dataset, which comprises simulated samples rather than samples randomly drawn from the entire dataset. As such, the principles guiding the estimation error and confidence levels outlined in Theorems 4.1 to 4.3 do not directly apply to ATE estimations derived from augmented data. Instead, these metrics are predominantly influenced by the error and confidence associated with estimating the conditional expectation, underscoring the nuanced approach required in leveraging data augmentation to improve causal effect estimations.

3. VC Dimension

This section elaborates on the direct correlation between the VC (Vapnik-Chervonenkis) dimension of the hypothesis space and the number of free variables in a hypothesis, which is foundational for understanding the complexity and capacity of models in machine learning.

To ensure the analytical tractability of this study, our research is predicated on the assumption that the number of free variables is equivalent to the VC dimension, a premise that aligns with common scenarios encountered in empirical research Shalev-Shwartz and Ben-David (2014).

For a given set of indicator functions, if there exist $\mathcal{H}$ samples that can be dichotomized by the functions in the set in every possible manner, amounting to $2^{\mathcal{H}}$ configurations, then we say the function set can 'shatter' $\mathcal{H}$ samples. The VC dimension of the function set is thus defined as the maximum number of samples, $\mathcal{H}$, that it can shatter.

The magnitude of the VC dimension is contingent upon the selected model and the defined hypothesis space. Below, we delineate the VC dimensions of several commonly utilized models:

1. The VC dimension of the hypothesis space constituted by linear hyperplanes in R^d space is $d+1$.

2. For neural networks, the magnitude of their VC dimension is $\mathbf{O}(VD)$, where V represents the number of neurons within the neural network, and D denotes the number of weights, correlating to the number of connections between neurons. This approximation provides a broad estimate; currently, explicit VC bounds for deep neural networks are not well-defined.

Table 1: Probabilities of satisfying the error boundaries at different ε and $1-\delta(\%)$, sampling 10,000 times across four methods.

		BD		PSW		PSS		DML	
C		95.45	99.73	95.45	99.73	95.45	99.73	95.45	99.73
1	0.05	98.92	99.89	98.96	99.89	98.99	99.91	99.25	99.91
	0.025	98.91	99.92	98.96	99.88	99.15	99.90	99.32	99.95
	0.01	98.95	99.91	98.90	99.89	98.83	99.86	99.39	99.96
2	0.05	99.86	100	99.86	100	99.91	99.99	99.90	99.98
	0.025	99.89	100	99.87	100	99.90	99.99	99.85	99.98
	0.01	99.86	99.99	99.78	99.97	99.39	100	99.90	99.99
5	0.05	99.99	99.99	99.93	99.98	99.99	100	99.97	100
	0.025	99.99	100	100	100	99.96	99.99	99.97	99.99
	0.01	100	100	100	100	99.79	99.98	100	100
10	0.05	99.99	99.99	100	100	99.98	100	100	100
	0.025	99.98	99.99	100	100	100	100	100	100
	0.01	99.96	100	99.99	100	100	99.99	99.99	100

4. Additional Experimental Details

Simulation Dataset Construction. We first create a Directed Acyclic Graph (DAG) to represent the relationships among treatment T, outcome Y, and confounders $\{X_i\}$. Each variable T, Y, X_i takes binary values $\{0, 1\}$. We specify $P(x_i)$ and the conditional probabilities $P(t \mid x)$, $P(y \mid t, x)$ to form a joint distribution, then sample data D' via Gibbs sampling (Koller and Friedman, 2009). Given $P(y \mid t)$, the theoretical causal effect is

$$\text{ACE}_D \;=\; P(Y = 1 \mid T = 1) \;-\; P(Y = 1 \mid T = 0),$$

which is set to 0.5 for consistent comparisons.

IHDP Dataset. In addition, we use the Infant Health and Development Program (IHDP) dataset introduced by Hill (2011). It contains 25 covariates (6 continuous and 19 discrete) from a randomized experiment, where treatment corresponds to expert home visits aimed at influencing infants' future cognitive test scores. We employ this dataset for data augmentation (described in Section Experimental Setup) to illustrate how insufficient samples can be mitigated via synthetic sample generation.

Error Thresholds and Confidence Levels. We consider $\varepsilon = \{0.05, 0.025, 0.01\}$. For confidence levels, we evaluate success rates at $1 - \delta \approx 95.45\%$ (2σ) and $1 - \delta \approx 99.73\%$ (3σ) to measure how often the estimation error stays below ε.

Implementation in DoWhy. For each sampling size boundary, we use `DoWhy` (Sharma and Kiciman, 2020) to estimate average treatment effects. We record the proportion of runs (out of 10,000) where the estimation error is within ε, and compare that proportion to the

nominal confidence level $1 - \delta$. This verifies whether the sampling boundaries align with the theoretical guarantees presented in the main text.

5. Extended Results on Different Methods

As seen in Table 1, all four methods consistently surpassed the nominal confidence thresholds across varying numbers of confounders C, error thresholds ε, and confidence levels $(1 - \delta)$.

6. Supplementary Results for Theorem 4.4

This section presents the full detailed results from 100 repeated experiments on the IHDP dataset to verify Theorem 4.4. Each row corresponds to one random seed.

Table 2: Detailed results from 100 repeated experiments on the IHDP dataset for Theorem 4.4. Each row corresponds to one random seed.

Index	**Seed**	**err outcome**	**err ate**	**2 × err outcome**	**check**
1	0	0.2038	0.0261	0.4076	True
2	1	0.1420	0.0977	0.2841	True
3	2	0.2077	0.1454	0.4154	True
4	3	0.1899	0.4201	0.3797	False
5	4	0.1579	0.2224	0.3159	True
6	5	0.2120	0.1215	0.4240	True
7	6	0.2029	0.1523	0.4058	True
8	7	0.2384	0.1242	0.4768	True
9	8	0.1905	0.3935	0.3810	False
10	9	0.2260	0.1794	0.4519	True
11	10	0.2103	0.0066	0.4205	True
12	11	0.1871	0.2335	0.3742	True
13	12	0.1932	0.1834	0.3864	True
14	13	0.1868	0.0706	0.3737	True
15	14	0.1883	0.2743	0.3766	True
16	15	0.2283	0.3520	0.4566	True
17	16	0.1929	0.2449	0.3858	True
18	17	0.2244	0.2918	0.4489	True
19	18	0.1596	0.1854	0.3192	True
20	19	0.1901	0.4597	0.3802	False
21	20	0.2334	0.3125	0.4668	True
22	21	0.2504	0.0023	0.5009	True
23	22	0.1960	0.3287	0.3920	True
24	23	0.1695	0.1754	0.3390	True
25	24	0.1927	0.1523	0.3853	True
26	25	0.1514	0.2635	0.3028	True

(Continued on next page)

(Continued from previous page)

Index	Seed	err outcome	err ate	2 × err outcome	check
27	26	0.2575	0.4741	0.5150	True
28	27	0.1747	0.0339	0.3495	True
29	28	0.1791	0.4770	0.3583	False
30	29	0.2022	0.3152	0.4043	True
31	30	0.1900	0.0648	0.3801	True
32	31	0.1776	0.2747	0.3552	True
33	32	0.1640	0.1006	0.3281	True
34	33	0.2372	0.0331	0.4745	True
35	34	0.2537	0.2513	0.5075	True
36	35	0.2319	0.3121	0.4637	True
37	36	0.2354	0.4259	0.4708	True
38	37	0.1768	0.0800	0.3537	True
39	38	0.1384	0.4427	0.2768	False
40	39	0.2297	0.0226	0.4594	True
41	40	0.1683	0.1661	0.3365	True
42	41	0.1946	0.2713	0.3892	True
43	42	0.2003	0.2598	0.4007	True
44	43	0.2564	0.0042	0.5127	True
45	44	0.1684	0.2066	0.3369	True
46	45	0.1946	0.2203	0.3892	True
47	46	0.1807	0.5667	0.3614	False
48	47	0.1527	0.0993	0.3053	True
49	48	0.1631	0.0713	0.3262	True
50	49	0.1794	0.1135	0.3587	True
51	50	0.2587	0.4802	0.5175	True
52	51	0.1792	0.1318	0.3584	True
53	52	0.1825	0.3886	0.3650	True
54	53	0.1514	0.2937	0.3029	False
55	54	0.2024	0.0283	0.4048	True
56	55	0.2482	0.1029	0.4965	True
57	56	0.1985	0.0502	0.3970	True
58	57	0.1765	0.1386	0.3530	True
59	58	0.1859	0.0205	0.3717	True
60	59	0.1733	0.0599	0.3466	True
61	60	0.1713	0.0217	0.3427	True
62	61	0.2586	0.1327	0.5172	True
63	62	0.1729	0.1523	0.3458	True
64	63	0.1977	0.1140	0.3954	True
65	64	0.2076	0.2564	0.4151	True
66	65	0.2341	0.0679	0.4683	True
67	66	0.2025	0.3255	0.4051	True

(Continued on next page)

(Continued from previous page)

Index	Seed	err outcome	err ate	2 × err outcome	check
68	67	0.2248	0.3303	0.4496	True
69	68	0.1541	0.0708	0.3082	True
70	69	0.2213	0.1033	0.4427	True
71	70	0.2320	0.3010	0.4640	True
72	71	0.1528	0.1884	0.3056	True
73	72	0.1720	0.0620	0.3439	True
74	73	0.1820	0.4251	0.3639	False
75	74	0.2241	0.1921	0.4482	True
76	75	0.1993	0.0536	0.3986	True
77	76	0.1811	0.2620	0.3623	True
78	77	0.2646	0.0094	0.5291	True
79	78	0.2012	0.1166	0.4024	True
80	79	0.1818	0.6128	0.3636	False
81	80	0.2156	0.2333	0.4312	True
82	81	0.1652	0.0373	0.3304	True
83	82	0.2533	0.3093	0.5065	True
84	83	0.1790	0.0978	0.3580	True
85	84	0.1723	0.1921	0.3445	True
86	85	0.1602	0.0192	0.3205	True
87	86	0.2284	0.5094	0.4567	False
88	87	0.1873	0.3103	0.3747	True
89	88	0.2056	0.2640	0.4111	True
90	89	0.1934	0.0139	0.3867	True
91	90	0.2103	0.0088	0.4206	True
92	91	0.1856	0.0474	0.3712	True
93	92	0.2091	0.0982	0.4182	True
94	93	0.1693	0.1417	0.3387	True
95	94	0.2153	0.2177	0.4306	True
96	95	0.1557	0.0289	0.3113	True
97	96	0.1758	0.1513	0.3516	True
98	97	0.1856	0.2599	0.3712	True
99	98	0.1884	0.0645	0.3767	True
100	99	0.1745	0.4314	0.3490	False

The tables above lists all seeds, outcome errors, ATE errors, and checks. Of the 100 trials in total, **89** (marked `check = True`) lie at or below the twofold boundary, $\texttt{err\_ate} \leq 2 \times \texttt{err\_outcome}$. As discussed in the main text (Section 5), the one-sample Wald test rejects a 50% null compliance rate with $z \approx 8.3$ ($p < 10^{-15}$). These data also illustrate how small biases in predicting treatment and control outcomes can inflate final ATE errors in a minority of trials.

7. Implementation Details of Data Augmentation Using Dragonnet

For the data augmentation process, the dataset was divided into training and validation sets with a 70:30 ratio. The network architecture comprised three Dense layers with ELU activation for representing covariates, one Dense layer with Sigmoid activation for the intervention, and three Dense layers with ELU activation for predicting the outcome variable y. L2 regularization was employed between layers to prevent overfitting. Optimization was conducted using Stochastic Gradient Descent (SGD), with a dynamically adjusted learning rate reduced by a factor of 0.5 and a momentum parameter set at 0.9, ensuring robust performance and minimized overfitting, which enabled effective estimation of treatment effects.

References

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

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

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.

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.