# LEARNING REPRESENTATIONS OF INSTRUMENTS FOR PARTIAL IDENTIFICATION OF TREATMENT EFFECTS

**Jonas Schweisthal**[1, 2, 5] **Dennis Frauen**[1, 2] **Maresa Schröder**[1, 2] **Konstantin Hess**[1, 2]

**Niki Kilbertus**[2, 3, 4] **Stefan Feuerriegel**[1, 2]

## ABSTRACT

Reliable estimation of treatment effects from observational data is crucial in fields like medicine, yet challenging when the unconfoundedness assumption is violated. We leverage arbitrary (potentially high-dimensional) instruments to estimate bounds on the conditional average treatment effect (CATE). Our contributions are three-fold: (1) We propose a novel approach for partial identification by mapping instruments into a discrete representation space that yields valid CATE bounds, essential for reliable decision-making. (2) We derive a two-step procedure that learns tight bounds via neural partitioning of the latent instrument space, thereby avoiding instability from numerical approximations or adversarial training and reducing finite-sample variance. (3) We provide theoretical guarantees for valid bounds with reduced variance and demonstrate effectiveness through extensive experiments. Overall, our method offers a new avenue for practitioners to exploit high-dimensional instruments (e.g., in Mendelian randomization).

## 1 INTRODUCTION

Estimating the *conditional average treatment effect (CATE)* from observational data is crucial for personalized medicine (Feuerriegel et al., 2024). For example, assessing the impact of alcohol consumption on cardiovascular diseases (Holmes et al., 2014) often relies on real-world data such as electronic health records. Reliable CATE estimation typically assumes *unconfoundedness* (Rubin, 1974); i.e., no unobserved confounders exist between treatment $A$ and outcome $Y$. When this assumption is violated, **instrumental variables (IVs)** $Z$, which affect $A$ but not $Y$ except through $A$, are employed (as in randomized studies with non-compliance (Imbens & Angrist, 1994)).

**Motivational example: Mendelian randomization.** In Mendelian randomization, genetic instruments $Z$ are used to estimate the effect of exposures (e.g., alcohol consumption) on outcomes (e.g., cardiovascular diseases) (Pierce et al., 2018). However, genetic instruments are high-dimensional and relate non-linearly to treatment, challenging existing IV methods that assume linearity, or other parametric or structural forms (Hartford et al., 2017; Singh et al., 2019; Xu et al., 2021). A promising alternative is **partial identification** of the CATE by estimating upper and lower bounds (Manski, 1990). Early works derived bounds for discrete IV settings (Balke & Pearl, 1997), while methods for continuous instruments typically require unstable optimization such as adversarial training (Kilbertus et al., 2020; Padh et al., 2023).

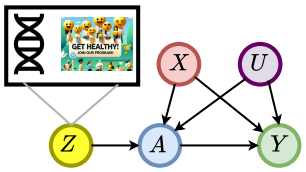


Figure 1: IV setting with complex instruments $Z$, observed confounders $X$, unobserved confounders $U$, binary treatment $A$, and outcome $Y$.

**Related work.** Existing machine learning approaches for CATE estimation with IVs largely focus on point identification. Some extend two-stage least-squares to non-linear settings (Singh et al., 2019; Xu

[1]LMU Munich
[2]Munich Center for Machine Learning
[3]School of Computation, Information and Technology, TU Munich
[4]Helmholtz Munich
[5]Corresponding author (jonas.schweisthal@lmu.de)

et al., 2021) or employ deep conditional density estimation (Hartford et al., 2017), and others develop doubly/multiply robust methods (Kennedy et al., 2019; Ogburn et al., 2015; Semenova & Chernozhukov, 2021; Syrgkanis et al., 2019; Frauen & Feuerriegel, 2023). Recent efforts in Mendelian randomization also target point identification but impose strict assumptions such as linearity or homogeneity (Legault et al., 2024; Malina et al., 2022). In contrast, the partial identification literature seeks to bound causal effects when point identification is unattainable. Early work derived bounds for bounded outcomes (Robins, 1989; Manski, 1990) and later extended these ideas to discrete IVs and treatments (Balke & Pearl, 1994; 1997; Swanson et al., 2018). For continuous instruments, existing methods either impose strong assumptions on treatment responses or require unstable adversarial training (Gunsilius, 2020; Hu et al., 2021; Kilbertus et al., 2020; Padh et al., 2023) and are not tailored for binary treatments.

**Research gap and contributions.** Reliable machine learning methods for partial identification of the CATE with complex, high-dimensional instruments remain underexplored. Our work fills this gap by leveraging high-dimensional instruments, avoiding strict parametric assumptions, and sidestepping unstable optimization procedures. We propose an IV method for partial identification of the CATE with complex instruments. Our approach maps complex instruments to a discrete representation (see Fig. 2) and employs a two-step neural partitioning procedure that reduces estimation variance. We validate our method both theoretically and empirically.

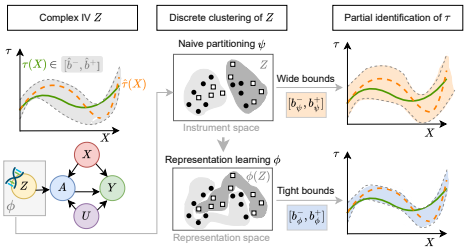


Figure 2: Mapping complex instruments $Z$ to a discrete representation $\phi(Z)$ yields tight bounds on the CATE.

## 2 Problem Setup

**Setting:** We focus on the standard IV setting (Angrist et al., 1996; Wooldridge, 2013) with complex instruments $Z \in \mathcal{Z} \subseteq \mathbb{R}^d$ (e.g., gene data, text, images) that may be continuous and high-dimensional. We assume an i.i.d. observational dataset $\mathcal{D} = \{z_i, x_i, a_i, y_i\}_{i=1}^n$ sampled from $(Z, X, A, Y) \sim \mathbb{P}$, where $X \in \mathcal{X} \subseteq \mathbb{R}^p$, $A \in \mathcal{A} \subseteq \{0, 1\}$, and $Y \in \mathcal{Y} \subseteq [s_1, s_2]$. Unobserved confounders $U$ between $A$ and $Y$ are allowed. We assume the causal structure in Fig. 1: $Z$ affects $A$ but has no direct effect on $Y$, and $Z$ is independent of $X$. An extended discussion is provided in Appendix B.

**Notation:** The *response function* is defined as $\mu^a(x, z) := \mathbb{E}[Y|X = x, A = a, Z = z]$, and the *propensity score* as $\pi(x, z) := \mathbb{P}(A = 1|X = x, Z = z)$.

**CATE:** Using the potential outcomes framework (Rubin, 1974) with $Y(a)$ as the potential outcome under $A = a$, the CATE is defined as $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$.

**Identifiability:** We make the standard assumptions in partial identification with IVs (Angrist et al., 1996): **Assumption 1** (Consistency): $Y(A) = Y$, **Assumption 2** (Exclusion): $Z \perp\!\!\!\perp Y(A) \mid (X, A, U)$, and **Assumption 3** (Independence): $Z \perp\!\!\!\perp (U, X)$. However, these do not suffice for identifying $\tau(x)$ (Gunsilius, 2020) without additional (and often unrealistic) assumptions (e.g., linearity or additive noise). This motivates our focus on partial identification.

**Objective:** Our goal is to estimate valid bounds $(b^-(x), b^+(x))$ for $\tau(x)$ such that $b^-(x) \leq \tau(x) \leq b^+(x),\ \forall\, x \in \mathcal{X}$, while minimizing the expected width $\mathbb{E}_X[b^+(X) - b^-(X)]$. Formally, we solve

$$b_*^-, b_*^+ \in \underset{b^-, b^+}{\arg\min}\, \mathbb{E}_X[b^+(X) - b^-(X)] \quad \text{s.t.} \quad b^-(x) \leq \tau(x) \leq b^+(x) \quad \forall\, x \in \mathcal{X}. \tag{1}$$

## 3 Partial Identification of the CATE with Complex Instruments

### 3.1 Overview

We now describe our method for solving the partial identification problem in Eq. (1). Since $\tau(x)$ is unknown, we proceed as follows:

**Outline:** (1) Learn a discretized representation $\phi(Z)$ of $Z$. (2) Derive closed-form bounds using $\phi$. (3) Express these bounds in terms of estimable quantities.

Existing closed-form bounds for discrete $Z$ (e.g., (Manski, 1990)) are not directly applicable to continuous or high-dimensional $Z$ because (1) they require evaluation over all combinations $l, m \in \mathcal{Z}^2$, and (2) sampling only a subset can incur high variance in low-density regions. Hence, we derive custom bounds for binary treatments and complex instruments. We next present the theory from a population view (Sec. 3.2) and a finite-sample view (Sec. 3.3).

### 3.2 Population view

**Theorem 1** (Bounds for arbitrary instrument discretizations)**.** *Let $\phi : \mathcal{Z} \to \{0, 1, \dots, k\}$ be any mapping from $Z$ to a discrete representation. Define*

$$\mu^a_\phi(x, \ell) = \int_Z \frac{\mu^a(x, z)\mathbb{P}(\phi(Z) = \ell \mid Z = z)}{\mathbb{P}(A = a, \phi(Z) = \ell)} \, \mathbb{P}(A = a \mid Z = z) \, \mathbb{P}(Z = z) \, \mathrm{d}z,$$
$$\pi_\phi(x, \ell) = \int_Z \frac{\pi(x, z)\mathbb{P}(\phi(Z) = \ell \mid Z = z)}{\mathbb{P}(\phi(Z) = \ell)} \, \mathbb{P}(Z = z) \, \mathrm{d}z. \tag{2}$$

*Then, under Assumptions 1, 2, and 3, the CATE $\tau(x)$ is bounded by*

$$b^-_\phi(x) \le \tau(x) \le b^+_\phi(x),$$

*with*

$$b^+_\phi(x) = \min_{l,m} b^+_{\phi;l,m}(x) \quad \textit{and} \quad b^-_\phi(x) = \max_{l,m} b^-_{\phi;l,m}(x), \tag{3}$$

$$b^+_{\phi;l,m}(x) = \pi_\phi(x, l)\mu^1_\phi(x, l) + (1 - \pi_\phi(x, l))s_2 - (1 - \pi_\phi(x, m))\mu^0_\phi(x, m) - \pi_\phi(x, m)s_1, \tag{4}$$
$$b^-_{\phi;l,m}(x) = \pi_\phi(x, l)\mu^1_\phi(x, l) + (1 - \pi_\phi(x, l))s_1 - (1 - \pi_\phi(x, m))\mu^0_\phi(x, m) - \pi_\phi(x, m)s_2.$$

*Proof.* See Appendix A. □

Theorem 1 shows that valid closed-form bounds for $\tau(x)$ are obtained for any $\phi$. We thus adjust Eq. (1) to yield a new objective by optimizing over $\phi$ such that

$$\phi^* \in \arg\min_{\phi \in \Phi} \mathbb{E}_X \Big[ b^+_\phi(X) - b^-_\phi(X) \Big]. \tag{5}$$

No additional validity constraints are needed, as the bounds are ensured by Theorem 1 and depend only on estimable quantities.

### 3.3 Finite-sample view

In practice, we estimate the bounds from Theorem 1 using finite data. Let $\hat{\pi}(x, z)$, $\hat{\mu}^a(x, z)$, and $\hat{\eta}(z)$ be initial estimators for $\pi(x, z)$, $\mu^a(x, z)$, and $\eta(z) = \mathbb{P}(A = 1 \mid Z = z)$, respectively. We then estimate the nuisance functions as

$$\hat{\mu}^a_\phi(x, \ell) = \frac{1}{\sum_{j=1}^n \mathbb{1}\{\phi(z_j) = \ell,\, a_j = a\}} \sum_{j=1}^n \hat{\mu}^a(x, z_j) \, \mathbb{1}\{\phi(z_j) = \ell\} \Big( a\hat{\eta}(z_j) + (1 - a)(1 - \hat{\eta}(z_j)) \Big), \tag{6}$$

$$\hat{\pi}_\phi(x, \ell) = \frac{1}{\sum_{j=1}^n \mathbb{1}\{\phi(z_j) = \ell\}} \sum_{j=1}^n \hat{\pi}(x, z_j) \, \mathbb{1}\{\phi(z_j) = \ell\}. \tag{7}$$

Plugging these into Eq. (3) gives estimates $\hat{b}^-_\phi(x)$ and $\hat{b}^+_\phi(x)$.

A naive approach would use $(\hat{b}^-_\phi(x), \hat{b}^+_\phi(x))$ to optimize Eq. (5), but in finite samples this is infeasible unless the complexity of $\phi$ is controlled. To demonstrate this, we state

**Lemma 1** (Tightness-bias-variance trade-off)**.** *Let $\mathbb{E}_n$ and $\mathrm{Var}_n$ denote expectation and variance over the sample (size $n$). Then,*

$$\mathbb{E}_n \Big[ \Big( b^+_*(x) - \hat{b}^+_\phi(x) \Big)^2 \Big] \le 2 \Big( \underbrace{\Big( b^+_*(x) - b^+_\phi(x) \Big)^2}_{\textit{(i) Population tightness}} + \underbrace{\mathbb{E}_n \Big[ b^+_{\phi^*}(x) - \hat{b}^+_\phi(x) \Big]^2}_{\textit{(ii) Estimation bias}} + \underbrace{\mathrm{Var}_n \Big( \hat{b}^+_\phi(x) \Big)}_{\textit{(iii) Estimation variance}} \Big). \tag{8}$$

*Proof.* See Appendix A. □

Lemma 1 decomposes the mean squared error of $\hat{b}^+_\phi(x)$ into (i) population tightness, (ii) estimation bias, and (iii) estimation variance. Increasing the complexity of $\phi$ (e.g., more partitions) reduces (i) but increases (iii). To further illustrate (iii), we have

**Theorem 2** (Asymptotic distributions of estimators)**.** *It holds that*

$$\sqrt{n}\hat{\mu}^a_\phi(x,\ell) \xrightarrow{d} \mathcal{N}\Big(\mu^a_\phi(x,\ell), \frac{1}{p_{\ell,\phi}}\Big(\frac{\mathrm{Var}(g(Z) \mid \phi(Z)=\ell)}{c} + d\Big)\Big),$$

$$\sqrt{n}\hat{\pi}_\phi(x,\ell) \xrightarrow{d} \mathcal{N}\Big(\pi_\phi(x,\ell), \frac{1}{p_{\ell,\phi}}\,\mathrm{Var}(h(Z) \mid \phi(Z)=\ell)\Big), \tag{9}$$

*with* $c = q^2_{\ell,\phi}$, $d = \frac{\theta^2_\ell(1-p_{\ell,\phi}q_{\ell,\phi})}{q^3_{\ell,\phi}}$, $p_{\ell,\phi} = \mathbb{P}(\phi(Z)=\ell)$, $q_{\ell,\phi} = \mathbb{P}(A=a \mid \phi(Z)=\ell)$, $g(Z) = \hat{\mu}^a(x,Z)\Big(a\hat{\eta}(Z) + (1-a)(1-\hat{\eta}(Z))\Big)$, $h(Z) = \hat{\pi}(x,Z)$, *and* $\theta_{\ell,\phi} = \mathbb{E}\big[g(Z) \mid \phi(Z)=\ell\big]$.

*Proof.* See Appendix A. □

Since the variance increases when $p_{\ell,\phi}$ is small, we aim to restrict $\phi$ to avoid low-density partitions. Overall, Lemma 1 and Theorem 2 reveal an inherent trade-off between bound tightness and estimation variance.

**Learning objective for the representation $\phi$:** To balance between learning tight bounds and controlling variance, we propose to optimize our adjusted objective given by

$$\phi^* \in \arg\min_{\phi\in\Phi} \mathbb{E}_X\Big[\hat{b}^+_\phi(X) - \hat{b}^-_\phi(X)\Big] \quad \text{s.t.} \quad \hat{p}_{\ell,\phi} > \varepsilon, \tag{10}$$

for some $\varepsilon > 0$ and all $\ell \in \{1,\dots,k\}$. We next present a neural method to learn tight bounds using this objective.

## 4 Neural method for CATE bounds with complex instruments

In this section, we propose a neural method to learn tight and valid bounds. Our method consists of two stages (see Algorithm 1): ① learning initial estimators of the three nuisance functions and ② learning an optimal representation $\phi^*$ to minimize the bound width. Our approach is model-agnostic, so arbitrary machine learning models (e.g., pre-trained encoders for gene data) can be used. An overview is shown in Fig. 3 (pseudocode in Appendix H).

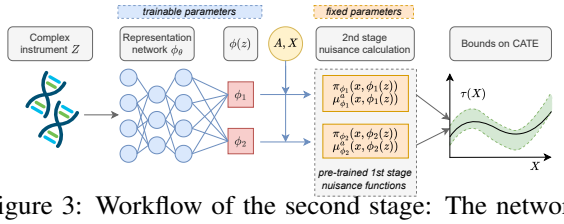


Figure 3: Workflow of the second stage: The network $\phi_\theta$ learns discrete latent representations of the complex $Z$. Using the pre-trained $\hat{\mu}$, $\hat{\pi}$, and $\hat{\eta}$, we compute the nuisance estimates via Eq. (6) and Eq. (7) to yield the bounds.

① **Initial nuisance estimation:** We use any suitable machine learning model (e.g., feed-forward neural network) to learn the first-stage nuisance functions $\hat{\mu}^a(x,z) = \hat{\mathbb{E}}[Y \mid X=x, A=a, Z=z]$, $\hat{\pi}(x,z) = \hat{\mathbb{P}}(A=1 \mid X=x, Z=z)$, $\hat{\eta}(z) = \hat{\mathbb{P}}(A=1 \mid Z=z)$. Since $Z$ and $X$ are potentially high-dimensional, the architectures use separate encoding layers for each variable followed by shared layers. For $\hat{\mu}^a(x,z)$, two outcome heads for $A \in \{0,1\}$ ensure that the treatment effect is preserved (Shalit et al., 2017).

② **Representation learning:** In the second stage, we train a neural network $\phi_\theta$ (with parameters $\theta$) to learn discrete representations of $Z$ that yield tight bounds while controlling estimation variance. On top of the final encoder layer, we apply the Gumbel-softmax trick (Jang et al., 2017) to learn $k$ *discrete* representations, with $k$ chosen as a hyperparameter.

**Custom loss function:** We transform our objective from Eq. (10) into a compositional loss with three terms:

**① Width minimization loss**

$$\mathcal{L}_{\mathrm{b}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{b}^+_{\phi_\theta}(x_i) - \hat{b}^-_{\phi_\theta}(x_i)\Big)$$

(Minimizes average bound width)

**② Regularization loss**

$$\mathcal{L}_{\mathrm{reg}}(\theta) = -\sum_{j=1}^{k}\log\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\phi_\theta(z_i)=j\}\Big)$$

(Controls variance by balancing partition size)

**③ Auxiliary guidance loss**

$$\mathcal{L}_{\mathrm{aux}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{1}\{\phi_\theta(z_i)=j\}\log\Big(p_\zeta(z_i)\Big)$$

(Promotes heterogeneity among partitions)

The final training loss is then

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{b}}(\theta) + \lambda\,\mathcal{L}_{\mathrm{reg}}(\theta) + \gamma\,\mathcal{L}_{\mathrm{aux}}(\theta), \tag{11}$$

| Metric | Dataset 1 | | | Dataset 2 | | |
|---|---|---|---|---|---|---|
| | Naïve | **Ours** | **Rel. Improvement** | Naïve | **Ours** | **Rel. Improvement** |
| **Coverage[↑]** | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | 0.00% | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | 0.00% |
| **Width[↓]** | $1.22 \pm 0.05$ | $\mathbf{1.05 \pm 0.01}$ | 13.9% | $1.31 \pm 0.16$ | $\mathbf{1.14 \pm 0.16}$ | 13.0% |
| **MSD[↓]** | $0.28 \pm 0.06$ | $\mathbf{0.03 \pm 0.03}$ | 89.3% | $0.09 \pm 0.06$ | $\mathbf{0.06 \pm 0.06}$ | 33.3% |

Table 2: **Datasets 1 and 2**: Comparison of NAÏVE vs. Ours regarding coverage, width, and MSD. Relative improvements in green.

with hyperparameters $\lambda$ and $\gamma$. Here, $\lambda$ controls the trade-off between bound tightness and estimation variance. $\gamma$ controls an auxiliary guidance loss that improves training convergence by promoting higher heterogeneity between partitions.

A key advantage of our method is its efficiency and robustness compared to alternating or adversarial training. In stage 2, only the discretization network $\phi_\theta$ is updated while the first-stage nuisance estimators remain fixed. This allows reusing the trained nuisance networks across different second-stage settings (e.g., varying $k$), making training more computationally efficient and robust.

## 5 EXPERIMENTS

**Baselines:** Existing methods focus on (a) point identification with strong assumptions, (b) partial identification with continuous treatments, or (c) discrete instruments. We focus on complex instruments with binary treatments. Hence, a fair comparison is precluded. Instead, we demonstrate the validity and tightness of our bounds. For comparison, we propose a NAÏVE baseline that first discretizes the instruments via $k$-means clustering and then learns the nuisance functions with respect to the discretized instruments to apply the existing discrete bounds from Lemma 2.[6]

| Metric | Naïve | Ours | Rel. Improve |
|---|---|---|---|
| **coverage*[↑]** | $0.96 \pm 0.09$ | $\mathbf{0.99 \pm 0.01}$ | 3.4% |
| **Width*[↓]** | $1.88 \pm 0.04$ | $\mathbf{1.85 \pm 0.04}$ | 1.8% |
| **MSE*[↓]** | $0.12 \pm 0.01$ | $\mathbf{0.11 \pm 0.01}$ | 9.2% |
| **MSD[↓]** | $0.10 \pm 0.10$ | $\mathbf{0.03 \pm 0.02}$ | 70.3% |

Table 1: **Dataset 3**: Comparison regarding coverage with oracle bounds, width, and MSD.

**Data:** We simulate data mimicking Mendelian Randomization, so that the ground-truth CATE is known for evaluation. In Datasets 1 and 2 a one-dimensional continuous instrument (polygenic risk score, (Pierce et al., 2018)) is simulated, with Dataset 1 modeling $\pi(x, z)$ as a simple function and Dataset 2 as a complex function. Dataset 3 uses high-dimensional instruments (SNPs, (Burgess et al., 2020)) to test our method in an even more complex setting. In all datasets, the CATE is heterogeneous in $X$ (see Appendix D).

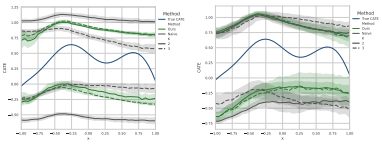

Figure 4: **Datasets 1 and 2:** Estimated bounds on the CATE over 5 runs for different $k$. Left: simple $\pi(x, z)$. Right: complex $\pi(x, z)$.

**Performance metrics:** We report *coverage*: frequency that the true CATE lies within the estimated bounds; *width*: average bound width (lower is better); and *MSD*: mean squared difference of predicted bounds over different $k$, reflecting robustness. For Dataset 3, we can approximate oracle bounds, and thus define *coverage**, *width**, and *MSE** by filtering runs with oracle bound coverage $\geq 95\%$, to compare tightness without overconfident predictions.

**Implementation details:** For our method, we use MLPs for the first-stage nuisance estimation and an MLP with Gumbel-softmax discretization for learning $\phi_\theta$. For the NAÏVE baseline, we use $k$-means clustering to discretize $Z$ and then identical MLP architectures for the nuisance functions. [7]

**Results:** Tables 2 and 1 compare our method with the NAÏVE baseline over multiple runs and different choices of $k$. We observe that: (i) Both methods reach nearly perfect coverage for the true CATE; for Dataset 3 our method achieves better coverage with respect to the oracle bounds. (ii) Our method learns tighter bounds (lower width, width*, and MSE*) compared to NAÏVE. (iii) Our method is robust across different $k$ values, as shown by a low MSD and stable performance in Figs. 4 and 5.

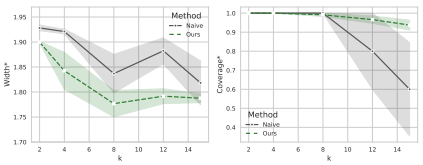


Figure 5: **Dataset 3:** Sensitivity analysis showing width* (left) and coverage* (right) over 5 runs for different $k$.

[6] We provide additional comparisons in Appendix E.

[7] Further details are in Appendix C.

**Sensitivity over $k$:** For Datasets 1 and 2, Fig. 4 shows estimated bounds for varying $k$. For Dataset 3, Fig. 5 plots width* and coverage* versus $k$. Our method shows robust performance (stable width* and near-optimal coverage*) while the NAÏVE baseline varies widely and loses coverage for higher $k$. This demonstrates that our learned representation $\phi$ is the key source of performance gain.

**Conclusion:** We propose a novel method for learning tight bounds on treatment effects using complex instruments (i.e., continuous, high-dimensional instruments with non-trivial relationships to treatment). The experimental results demonstrate the validity, tightness, and robustness of our bounds.

## REFERENCES


Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds, and applications. In *UAI*, 1994.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Stephen Burgess, Christopher N Foley, Elias Allara, James R Staley, and Joanna MM Howson. A robust and efficient method for mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1):376, 2020.

Yash Chandak, Shiv Shankar, Vasilis Syrgkanis, and Emma Brunskill. Adaptive instrument design for indirect experiments. In *ICLR*, 2023.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Dennis Frauen and Stefan Feuerriegel. Estimating individual treatment effects under unobserved confounding using binary instruments. In *ICLR*, 2023.

M Maria Glymour, Eric J Tchetgen Tchetgen, and James M Robins. Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339, 2012.

Florian Gunsilius. A path-sampling method to partially identify causal effects in instrumental variable models. *arXiv preprint*, arXiv:1910.09502, 2020.

Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121, 2024.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *ICML*, 2017.

Michael V Holmes, Caroline E Dale, Luisa Zuccolo, Richard J Silverwood, Yiran Guo, Zheng Ye, David Prieto-Merino, Abbas Dehghan, Stella Trompet, Andrew Wong, et al. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ*, 349:g4164, 2014.

Yaowei Hu, Yongkai Wu, and Xintau Wu. A generative adversarial framework for bounding confounded causal effects. In *AAAI*, 2021.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ICLR*, 2017.

Edward H. Kennedy, Scott A. Lorch, and Dylan S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019.

Niki Kilbertus, Matt J. Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. In *NeurIPS*, 2020.

Alice Kongsted and Anne Molgaard Nielsen. Latent class analysis in health research. *Journal of Physiotherapy*, 63(1):55–58, 2017.

Marc-André Legault, Jason Hartford, Benoît J Arsenault, Archer Y Yang, and Joelle Pineau. A novel and efficient machine learning mendelian randomization estimator applied to predict the safety and efficacy of sclerostin inhibition. *medRxiv*, 2024.

Stephen Malina, Daniel Cizin, and David A Knowles. Deep mendelian randomization: Investigating the causal knowledge of genomic deep learning models. *PLoS Computational Biology*, 18(10): e1009880, 2022.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2):319–323, 1990.

SC Matz, JD Teeny, Sumer S Vaid, H Peters, GM Harari, and M Cerf. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692, 2024.

Katherine L Milkman, Mitesh S Patel, Linnea Gandhi, Heather N Graci, Dena M Gromet, Hung Ho, Joseph S Kay, Timothy W Lee, Modupe Akinola, John Beshears, et al. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20):e2101165118, 2021.

Elizabeth L. Ogburn, Andrea Rotnitzky, and James M. Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B*, 77(2):373–396, 2015.

Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, and Niki Kilbertus. Stochastic causal programming for bounding treatment effects. In *CLeaR*, 2023.

Judea Pearl. Causal inference from indirect experiments. *Artificial Intelligence in Medicine*, 7(6): 561–582, 1995.

Brandon L Pierce, Peter Kraft, and Chenan Zhang. Mendelian randomization studies of cancer risk: a literature review. *Current Epidemiology Reports*, 5:184–196, 2018.

James M Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: A Focus on AIDS*, pp. 113–159, 1989.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Jonas Schweisthal, Dennis Frauen, Mihaela van der Schaar, and Stefan Feuerriegel. Meta-learners for partially-identified treatment effects across multiple environments. In *ICML*, 2024.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 2017.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *NeurIPS*, 2019.

Sonja A Swanson, Miguel A Hernán, Matthew Miller, James M Robins, and Thomas S Richardson. Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018.

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. In *NeurIPS*, 2019.

Jeffrey M. Wooldridge. *Introductory Econometrics: A modern approach*. Routledge, 2013. ISBN 9781136586101.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *ICLR*, 2021.

# A PROOFS

## A.1 PROOF OF THEOREM 1

We begin by stating a result from the literature that obtains valid bounds for discrete instruments.

**Lemma 2** ((Swanson et al., 2018; Schweisthal et al., 2024))**.** *Under Assumptions 1 and 2, the CATE is bounded via*

$$b^-(x) \leq \tau(x) \leq b^+(x), \tag{12}$$

*with*

$$b^+(x) = \min_{l,m} b^+_{l,m}(x) \quad \textit{and} \quad b^-(x) = \max_{l,m} b^-_{l,m}(x) \tag{13}$$

*where*

$$b^+_{l,m}(x) = \pi(x,l)\mu^1(x,l) + (1-\pi(x,l))s_2 - (1-\pi(x,m))\mu^0(x,m) - \pi(x,m)s_1, \tag{14}$$

$$b^-_{l,m}(x) = \pi(x,l)\mu^1(x,l) + (1-\pi(x,l))s_1 - (1-\pi(x,m))\mu^0(x,m) - \pi(x,m)s_2. \tag{15}$$

*Proof of Theorem 1.* First, note that, for a given representation $\phi$, the representation $\phi(Z)$ is still a valid (discrete) instrument that satisfies Assumptions 1 and 2. Hence, we can apply Lemma 2 using $\phi(Z)$ as an instrument and immediately obtain the bounds from Theorem 1, but with *representation-induced nuisance functions* $\mu^a_\phi(x,\ell) = \mathbb{E}[Y|X=x, A=a, \phi(Z)=\ell]$ and $\pi_\phi(x,\ell) = \mathbb{P}(A=1|X=x,\phi(Z)=\ell)$ for $\ell \in \{0,\dots,k\}$.

We can write the representation-induced response function as

$$\begin{aligned}
&\mathbb{E}[Y|X=x,A=a,\phi(Z)=\ell] \overset{Z \perp\!\!\!\perp X}{=} \int_Z \mathbb{E}[Y|X=x,A=a,Z=z]\mathbb{P}(Z=z|A=a,\phi(Z)=\ell)\,\mathrm{d}z \\
&= \int_Z \mathbb{E}[Y|X=x,A=a,Z=z]\frac{\mathbb{P}(\phi(Z)=\ell|A=a,Z=z)\mathbb{P}(A=a|Z=z)\mathbb{P}(Z=z)}{\mathbb{P}(A=a|\phi(Z)=\ell)\mathbb{P}(\phi(Z)=\ell)}\,\mathrm{d}z \\
&= \frac{1}{\mathbb{P}(A=a|\phi(Z)=\ell)\mathbb{P}(\phi(Z)=\ell)} \\
&\qquad \int_Z \mathbb{E}[Y|X=x,A=a,Z=z]\mathbb{P}(\phi(Z)=\ell|A=a,Z=z)\mathbb{P}(A=a|Z=z)\mathbb{P}(Z=z)\,\mathrm{d}z \\
&= \frac{1}{\mathbb{P}(A=a|\phi(Z)=\ell)\mathbb{P}(\phi(Z)=\ell)} \\
&\qquad\quad \int_Z \mathbb{E}[Y|X=x,A=a,Z=z]\mathbb{P}(\phi(Z)=\ell|Z=z)\mathbb{P}(A=a|Z=z)\mathbb{P}(Z=z)\,\mathrm{d}z
\end{aligned} \tag{16}$$

and the representation-induced propensity score as

$$\begin{aligned}
&\mathbb{P}(A=1|X=x,\phi(Z)=\ell) \overset{Z \perp\!\!\!\perp X}{=} \int_Z \mathbb{P}(A=1|X=x,Z=z)\mathbb{P}(Z=z|\phi(Z)=\ell)\,\mathrm{d}z \\
&= \int_Z \mathbb{P}(A=1|X=x,Z=z)\mathbb{P}(\phi(Z)=\ell|Z=z)\frac{\mathbb{P}(Z=z)}{\mathbb{P}(\phi(Z)=\ell)}\,\mathrm{d}z \\
&= \frac{1}{\mathbb{P}(\phi(Z)=\ell)}\int_Z \mathbb{P}(A=1|X=x,Z=z)\mathbb{P}(\phi(Z)=\ell|Z=z)\mathbb{P}(Z=z)\,\mathrm{d}z,
\end{aligned} \tag{17}$$

which completes the proof. □

### A.2 Proof of Lemma 1

*Proof.* The result follows from

$$\mathbb{E}_n\left[\left(b_*^+(x)-\hat{b}_\phi^+(x)\right)^2\right]=\mathbb{E}_n\left[\left(b_*^+(x)-b_{\phi^*}^+(x)+b_{\phi^*}^+(x)-\hat{b}_\phi^+(x)\right)^2\right] \tag{18}$$

$$\leq 2\left(\left(b_*^+(x)-\hat{b}_\phi^+(x)\right)^2+\mathbb{E}_n\left[\left(b_{\phi^*}^+(x)-\hat{b}_\phi^+(x)\right)^2\right]\right) \tag{19}$$

$$\overset{(*)}{(=)}2\left(\left(b_*^+(x)-\hat{b}_\phi^+(x)\right)^2+\mathbb{E}_n\left[b_{\phi^*}^+(x)-\hat{b}_\phi^+(x)\right]^2+\text{Var}_n(\hat{b}_\phi^+(x))\right), \tag{20}$$

where we used the bias-variance decomposition for the MSE for $(*)$. □

### A.3 Proof of Theorem 2

*Proof.* We derive the asymptotic distributions of the estimators $\hat{\mu}_\phi^a(x,\ell)$ from Eq. (6) and $\hat{\pi}_\phi(x,\ell)$ from Eq. (7). We proceed by analyzing the numerator and denominator of each estimator. First, we show that both are asymptotically normal and then we apply the delta method to obtain the asymptotic distribution of the ratios.

**Distribution of $\hat{\mu}_\phi^a(x,\ell)$:** Recall from Equation (6) that we can write $\hat{\mu}_\phi^a(x,\ell)$ as

$$\hat{\mu}_\phi^a(x,\ell)=\frac{S_n}{N_n}, \tag{21}$$

where

$$S_n=\frac{1}{n}\sum_{j=1}^n W_j,\quad \text{with}\quad W_j=\hat{\mu}^a(x,z_j)\mathbb{1}\{\phi(z_j)=\ell\}[a\hat{\eta}(z_j)+(1-a)(1-\hat{\eta}(z_j))], \tag{22}$$

$$N_n=\frac{1}{n}\sum_{j=1}^n D_j,\quad \text{with}\quad D_j=\mathbb{1}\{\phi(z_j)=\ell,a_j=a\}. \tag{23}$$

We define the moments

$$\mu_W=\mathbb{E}[W]=p_\ell\theta_\ell \tag{24}$$

$$\sigma_W^2=\text{Var}(W)=p_\ell(\gamma_\ell-p_\ell\theta_\ell^2) \tag{25}$$

$$\mu_D=\mathbb{E}[D]=p_\ell q_\ell \tag{26}$$

$$\sigma_D^2=\text{Var}(D)=p_\ell q_\ell(1-p_\ell q_\ell) \tag{27}$$

$$\text{c}_{WD}=\text{Cov}(W,D)=p_\ell q_\ell\theta_\ell(1-p_\ell), \tag{28}$$

where $p_\ell=\mathbb{P}(\phi(Z)=\ell)$, $q_\ell=\mathbb{P}(A=a\mid\phi(Z)=\ell)$, $\theta_\ell=\mathbb{E}[g(Z)\mid\phi(Z)=\ell]$, and $\gamma_\ell=\mathbb{E}[g(Z)^2\mid\phi(Z)=\ell]$, with $g(Z)=\hat{\mu}^a(x,Z)(a\hat{\eta}(Z)+(1-a)(1-\hat{\eta}(Z))$. Note that, for better readability, in this proof we avoid the double indexing showing the dependency on $\phi$ which we used in the theorem in the main paper.

By the central limit theorem, we know that

$$\sqrt{n}\begin{pmatrix}S_n\\N_n\end{pmatrix}\xrightarrow{d}\mathcal{N}_2\left(\mu=\begin{pmatrix}\mu_W\\\mu_D\end{pmatrix},\Sigma=\begin{pmatrix}\sigma_W^2&\text{c}_{WD}\\\text{c}_{WD}&\sigma_D^2\end{pmatrix}\right). \tag{29}$$

Let $f(s,n)=\frac{s}{n}$. We are interested in the asymptotic distribution of the ratio $\hat{\mu}_\phi^a(x,\ell)=f(S_n,N_n)$. The delta method states that

$$\sqrt{n}f(S_n,N_n)\xrightarrow{d}\mathcal{N}_2\left(f(\mu_W,\mu_D),\nabla f^\top(\mu_W,\mu_D)\Sigma\nabla f(\mu_W,\mu_D)\right) \tag{30}$$

Using that the gradient is $\nabla f^\top(\mu_W, \mu_D) = \left(\frac{1}{\mu_D}, -\frac{\mu_W}{\mu_D^2}\right)$, we can obtain the asymptotic variance via

$$\nabla f^\top(\mu_W, \mu_D)\Sigma\nabla f(\mu_W, \mu_D) = \frac{\sigma_W^2}{\mu_D^2} - 2\frac{\mu_W \mathrm{c}_{WD}}{\mu_D^3} + \frac{\mu_W^2\sigma_D^2}{\mu_D^4} \tag{31}$$

$$= \frac{1}{p_\ell}\left(\frac{(\gamma_\ell - \theta_\ell^2)}{q_\ell^2} + \frac{\theta_\ell^2(1-p_\ell q_\ell)}{q_\ell^3}\right) \tag{32}$$

$$= \frac{1}{p_\ell}\left(\frac{\mathrm{Var}(g(Z) \mid \phi(Z) = \ell)}{q_\ell^2} + \frac{\theta_\ell^2(1-p_\ell q_\ell)}{q_\ell^3}\right). \tag{33}$$

**Distribution of $\hat{\pi}_\phi(x, \ell)$:** Recall from Equation (7) that we can write $\hat{\pi}_\phi(x, \ell)$ as

$$\hat{\pi}_\phi(x, \ell) = \frac{S_n}{N_n}, \tag{34}$$

where

$$S_n = \frac{1}{n}\sum_{j=1}^n W_j, \quad \text{with} \quad W_j = \hat{\pi}(x, z_j)\mathbb{1}\{\phi(z_j) = l\}, \tag{35}$$

$$N_n = \frac{1}{n}\sum_{j=1}^n D_j, \quad \text{with} \quad D_j = \mathbb{1}\{\phi(z_j) = l\}. \tag{36}$$

We define the moments

$$\mu_W = \mathbb{E}[W] = p_\ell\theta_\ell \tag{37}$$

$$\sigma_W^2 = \mathrm{Var}(W) = p_\ell(\gamma_\ell - p_\ell\theta_\ell^2) \tag{38}$$

$$\mu_D = \mathbb{E}[D] = p_\ell \tag{39}$$

$$\sigma_D^2 = \mathrm{Var}(D) = p_\ell(1-p_\ell) \tag{40}$$

$$\mathrm{c}_{WD} = \mathrm{Cov}(W, D) = p_\ell\theta_\ell(1-p_\ell), \tag{41}$$

where $p_\ell = \mathbb{P}(\phi(Z) = \ell)$, $\theta_\ell = \mathbb{E}[h(Z) \mid \phi(Z) = \ell]$, and $\gamma_\ell = \mathbb{E}[h(Z)^2 \mid \phi(Z) = \ell]$, with $h(Z) = \hat{\pi}(x, Z)$.

By the central limit theorem, we know that

$$\sqrt{n}\begin{pmatrix} S_n \\ N_n \end{pmatrix} \xrightarrow{d} \mathcal{N}_2\left(\mu = \begin{pmatrix} \mu_W \\ \mu_D \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_W^2 & \mathrm{c}_{WD} \\ \mathrm{c}_{WD} & \sigma_D^2 \end{pmatrix}\right). \tag{42}$$

We can then calculate the asymptotic variance using the delta method as above and obtain

$$\nabla f^\top(\mu_W, \mu_D)\Sigma\nabla f(\mu_W, \mu_D) = \frac{\sigma_W^2}{\mu_D^2} - 2\frac{\mu_W \mathrm{c}_{WD}}{\mu_D^3} + \frac{\mu_W^2\sigma_D^2}{\mu_D^4} \tag{43}$$

$$= \frac{1}{p_\ell}(\gamma_\ell - \theta_\ell^2) \tag{44}$$

$$= \frac{1}{p_\ell}\mathrm{Var}(h(Z) \mid \phi(Z) = \ell). \tag{45}$$

□

## B REAL-WORLD RELEVANCE AND VALIDITY OF ASSUMPTIONS

In this section, we elaborate on the real-world relevance of our considered setting and show that our assumptions often hold and are even weaker than the ones of existing approaches. For that, we draw upon two real-world settings.

### B.1 MENDELIAN RANDOMIZATION

Mendelian randomization (MR; the main motivational example from our paper) is a widely used method from biostatistics to estimate the causal effect of some treatment or exposure (such as alcohol consumption) on some outcome (such as cardiovascular diseases). We refer to Pierce et al. (2018) for an introduction to MR, which also shows that MR is widely used in medicine. For that, genetic variants (such as different single nucleotide polymorphisms, SNPs) are used as instruments where it is known that they only influence the exposure but not directly the outcome. Our method for partial identification with complex instruments is perfectly suited for this common real-world application. Depending on the use case, either a predefined genetic risk score (Burgess et al., 2020) as a continuous variable, or up to hundreds of SNPs are used simultaneously as IVs to strengthen the power of the analysis, resulting in high-dimensional instruments (Pierce et al., 2018).

**Validity of assumptions:** The IV assumptions used in our paper such as the exclusion and independence assumptions can be ensured by expert knowledge (e.g., given some observed confounder age ($X$), genetic variations (Z) do not affect age) or, in some cases, they can be even directly tested for (Glymour et al., 2012). In contrast, existing methods for MR rely on additional hard assumptions on top such as the knowledge about the parametric form of the underlying data-generating process. Especially with such high-dimensional IVs, misspecification of these models may result in significantly biased effect estimates. In contrast, our method does not rely on any parametric assumption and also no additional assumptions compared to previous methods, thus enabling more reliable causal inferences in the real-world application of MR by using *strictly weaker* assumptions than existing work.

### B.2 INDIRECT EXPERIMENTS

With indirect experiments (IEs), we show that, in principle, our method is not constrained to medical applications but is also highly useful in various other domains. IEs are widely applied in various areas such as social sciences or public health to estimate causal effects in settings with non-adherence, i.e., where people cannot be forced to take treatments but rather be encouraged by some nudge (Pearl, 1995). For instance, researchers might be interested in estimating the effect of some treatment such as participating in a healthcare program ($T$) on some health outcome $Y$ by randomly assigning nudges $Z$ (IVs) in the form of different text messages on social media promoting participation. Here, common nudges (IVs) are in the form of, for instance, text or even image data and thus high-dimensional, showing the necessity of a method capable of handling complex IVs such as ours.

In principle, our method can be applied to every setting with continuous or multi-dimensional IVs where one wants to avoid making the hard untestable assumptions necessary for point identification such as linearity or additivity (e.g., Hartford et al. (2017)). Specific examples for applications with high-dimensional IVs are text-based nudges for encouraging vaccinations (Milkman et al., 2021), or various kinds of experiments where text nudges are generated by different strategies such as for political microtargeting (Hackenburg & Margetts, 2024) or for personalized persuasion in general (Matz et al., 2024).

Another important application area is online marketing. Concrete use cases involve extended A/B testing for evaluating the benefits of new features, e.g., when one is interested in the effect of a new version of an app on user engagement. Here, users with features such as age, gender, and content preferences ($X$) can be nudged by emails or push notifications ($Z$) to test a new feature such as using a new version of an app ($A$) to estimate its effect on engagement metrics such as screen time ($Y$). Further, our method could also be extended to improve current methods for optimizing instrument designs for indirect experiments that for now assume identifiability is possible (e.g., Chandak et al. (2023)).

**Validity of assumptions:** As a major benefit of IEs, the IV assumptions are *ensured per design* as the IVs are randomly assigned, and, thus they always hold. Hence, our method provides a promising tool for evaluating the effects of IEs.

## C IMPLEMENTATION AND TRAINING DETAILS

**Model architecture:** For all our models, we use MLPs with ReLU activation function. For $\hat{\mu}_\phi^a$, we use 2 layers to encode $X$ and 3 layers to encode $Z$. Then, we concatenate the outputs and add 2 additional shared layers. Finally, we calculate the outputs by a separate treatment head for $A = 0$ and $A = 1$ to ensure the expressiveness of $A$ for predicting $Y$. For $\hat{\pi}$, we use the same architecture. For $\hat{\eta}$, we use 3 layers. For $\phi_\theta$, we also use 3 layers and apply discretization on top of the $K$ outputs (Jang et al., 2017). For the nuisance parameters of the $k$-means baseline, we use the same models as for $\hat{\mu}_\phi^a$ and $\hat{\pi}$ for a fair comparison. We use a neuron size of 10 for all hidden layers.

**Training details:** For training our nuisance functions, we use an MSE loss for the functions learning the continuous outcome $Y$ and a cross-entropy loss for functions learning the binary treatment $A$. For all models, we use the Adam optimizer with a learning rate of $0.03$. We train our models for a maximum of 100 epochs and apply early stopping. For our method, we fixed $\lambda = 1$ and performed random search to tune for $[0, 1]$ for $\gamma$. We use PyTorch Lightning for implementation. Each training run of the experiments could be performed on a CPU with 8 cores in under 15 minutes.

## D DATA DESCRIPTION

**Dataset 1:** We simulate an observed confounder $X \sim \text{Uniform}[-1, 1]$ and an unobserved confounder $U \sim \text{Uniform}[-1, 1]$.

The instrument $Z$ is defined as

$$Z \sim \text{Mixture}\left(\frac{1}{2}\text{Uniform}[-1, 1] + \frac{1}{4}\text{Beta}(2, 2) + \frac{1}{4}(-\text{Beta}(2, 2))\right). \tag{46}$$

We define $\rho$ as

$$\rho = \frac{1}{1 + \exp\left(-\left((2|Z| - \max(Z)) + X + 0.5 \cdot U\right)\right)}. \tag{47}$$

Then, the propensity score is given by

$$\pi = (\rho - 0.5) \cdot 0.9 + 0.5. \tag{48}$$

We then sample our treatment assignments from the propensity scores as

$$A \sim \text{Bernoulli}(\pi). \tag{49}$$

The conditional average treatment effect (CATE) is defined as

$$\tau(X) = -\frac{(2.5X)^4 + 12\sin(6X) + 0.5\cos(X)}{80} + 0.5. \tag{50}$$

The outcome $Y$ is then generated by

$$Y = (X + 0.5U + 0.1 \cdot \text{Laplace}(0, 1)) \cdot 0.25 + \tau(X) \cdot A. \tag{51}$$

**Dataset 2:** We keep the other properties but change the propensity score to be more complex, which results in harder-to-learn optimal representations of $Z$ for tightening the bounds. The propensity score is given by

$$\pi = \sin(2.5Z + X + U) \cdot 0.48 + 0.48 + \frac{0.04}{1 + \exp(-3|Z|)}. \tag{52}$$

**Dataset 3:** We simulate $X$ and $U$ as above. Then, we sample a $d$-dimensional $Z \in \{0, 1\}^d$ with $d = 20$ as

$$Z \sim \text{Binomial}(d, 0.5). \tag{53}$$

Thus, our modeling is here inspired by using multiple SNPs (appearances of genetic variations) as instruments (Burgess et al., 2020), where we simulate potential variations for 20 genes.

Then, we define

$$\rho = \sum_{j=1}^{d} [\mathbb{1}\{j \leq 5\} Z_j] \tag{54}$$

and the propensity score, inspired by the more complex setting of Dataset 2, as

$$\pi = 0.48 \sin(10\rho + X + U) + 0.48 + \frac{0.04}{1 + \exp(-3|5\rho|)}. \tag{55}$$

Then, we define the CATE as

$$\tau(X) = -\frac{-(1.6X + 0.5)^4 + 12\sin(4X + 1.5) + \cos(X)}{80} + 0.5. \tag{56}$$

and the outcome dependent on $\tau$, $X$ and $U$ analogously as for Datasets 1 and 2.

**Dataset 4:** To test our method even in higher-dimensional settings, we consider a 4th dataset with **100-dimensional IVs**. For that, we adapt the DGP from dataset 3 but set $d = 100$. Then we adjust the latent discrete IV score as

$$\rho = \sum_{j=1}^{d} [\mathbb{1}\{j \leq 25\} Z_j]. \tag{57}$$

By Eq. (54) and Eq. (57), we ensure that some of the modeled SNPs are irrelevant for $\pi$ and thus do not affect the treatment or exposure $A$. Thereby, we focus on realistic settings in practice, where the relevance of instruments cannot always be ensured which imposes challenges especially for existing methods for point identification, but not for our approach. Further, we ensure that the latent score $\rho$ can only take 5 discrete levels for dataset 3 and 25 discrete levels for dataset 4. This allows us to approximate oracle bounds using the discrete bounds on top of $\rho$ by leveraging Lemma 2 such that we can evaluate our method and the baseline in comparison to oracle bounds.

To create the simulated data used in Sec. 5, we sample $n = 2000$ from the data-generating process above. We then split the data into train (40%), val (20%), and test (40%) sets such that the bounds and deviation can be calculated on the same amount of data for training and testing.

| Dataset | Method | $k$ | Coverage[↑] | Width[↓] |
|---|---|---|---|---|
| Dataset 1 | Naïve | 2 | $1.00 \pm 0.00$ | $1.62 \pm 0.06$ |
| | | 3 | $1.00 \pm 0.00$ | $0.83 \pm 0.16$ |
| | Ours | 2 | $1.00 \pm 0.00$ | $1.01 \pm 0.05$ |
| | | 3 | $1.00 \pm 0.00$ | $1.09 \pm 0.04$ |
| Dataset 2 | Naïve | 2 | $1.00 \pm 0.00$ | $1.34 \pm 0.19$ |
| | | 3 | $1.00 \pm 0.00$ | $1.28 \pm 0.20$ |
| | Ours | 2 | $1.00 \pm 0.00$ | $1.13 \pm 0.19$ |
| | | 3 | $1.00 \pm 0.00$ | $1.15 \pm 0.31$ |

Table 3: **Datasets 1 and 2:** Sensitivity over $k$.

# E Additional Results

## E.1 Additional results for sensitivity over $k$

## E.2 Additional baselines

As mentioned in the main paper, existing methods are not designed for our considered setting of continuous or high-dimensional IVs with binary treatments. However, to further show the advantages and necessity of our tailored method, we compare with two additional baselines that were not developed for our task but which we adapted for our task, namely, one from uncertainty quantification for point estimates and one from the discrete instruments setting:

(i) *DeepIV with bootstrapped confidence intervals*. DeepIV (Hartford et al., 2017) is a neural method tailored for high-dimensional instruments when point identification can be ensured. This requires the *additional assumption* of additivity of the unobserved confounding, which usually cannot be ensured and is not necessary for our method. For DeepIV, we can approximate confidence intervals using bootstrapping. Here, we approximate confidence intervals with a confidence level of $95\%$, indicating an expected coverage of $95\%$ if assumptions were not violated. However, note that these intervals can *only* adjust for statistical uncertainty, but *not* for identifiability uncertainty due to the violation of causal assumptions. Thus, this baseline acts as an additional motivation for why bound estimators such as our method are important.

(ii) *Discretized IVs*: As a further additional baseline, we proceed by directly discretizing the high-dimensional IVs and then estimating the existing bounds for discrete IVs. Hence, *one loses information* from the IV due to the discretization. Our implementation here is the same as for the naïve baseline, however, the $k$ partitions are not learned by $k$-means clustering but instead defined by a simple grouping rule. To ensure a fair comparison, we average the results of experiments conducted with the same number of partitions $k$ for all methods.

| **Metric** | DeepIV (CI) | Discretized | Naïve | **Ours** | **Rel. Improvement** |
|---|---|---|---|---|---|
| **Coverage[↑]** | $0.52 \pm 0.29$ | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | $0.0\%$ |
| **Coverage*[↑]** | $0.00 \pm 0.00$ | $0.99 \pm 0.01$ | $0.96 \pm 0.09$ | $\mathbf{0.99 \pm 0.01}$ | $0.0\%$ |
| **Width*[↓]** | — | $1.91 \pm 0.04$ | $1.88 \pm 0.04$ | $\mathbf{1.85 \pm 0.04}$ | $1.8\%$ |
| **MSE*[↓]** | — | $0.13 \pm 0.01$ | $0.12 \pm 0.01$ | $\mathbf{0.11 \pm 0.01}$ | $9.2\%$ |
| **MSD[↓]** | — | $0.08 \pm 0.03$ | $0.10 \pm 0.10$ | $\mathbf{0.03 \pm 0.02}$ | $70.3\%$ |

Table 4: **Dataset 3**: Comparison of methods (Naïve vs Ours) on coverage and width metrics with relative performance improvement. Note: "—" means that there are no reliable runs for which the corresponding performance metrics could be calculated.

**Results:** We report our results for Dataset 3 in Table 4. We observe that the DeepIV method, as expected, gives *falsely* overconfident bounds with only about $53\%$ coverage of the true CATE and no coverage of the oracle bounds. Thus, there are no reliable runs for which the other metrics could be calculated (denoted by "—" in the tables). This emphasizes the necessity for using bound estimators. Further, we observe that the discretized baseline gives *more conservative* and *wider* bounds under similar coverage (higher Width* and MSE*) and performs less robustly with regard to $k$ (higher MSD). In sum, the results confirm the strong performance of our method.

## E.3 High-dimensional dataset

To show the validity of our method in even more high-dimensional settings, we added additional experiments with 100-dimensional IVs. For that, we introduced our Dataset 4 (see Appendix D). We report the results for our method and the same baselines as in the previous section. Further, for

| Metric | DeepIV (CI) | Discretized | Naïve | Ours | Rel. Improvement |
|---|---|---|---|---|---|
| **Coverage[↑]** | $0.01 \pm 0.00$ | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | 0.0% |
| **Coverage*[↑]** | $0.00 \pm 0.00$ | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | $1.00 \pm 0.00$ | 0.0% |
| **Width*[↓]** | — | $1.90 \pm 0.06$ | $1.82 \pm 0.13$ | $\mathbf{1.75 \pm 0.08}$ | 3.7% |
| **MSE*[↓]** | — | $0.26 \pm 0.03$ | $0.23 \pm 0.05$ | $\mathbf{0.21 \pm 0.03}$ | 10.9% |
| **MSD[↓]** | — | $0.05 \pm 0.03$ | $0.10 \pm 0.04$ | $\mathbf{0.05 \pm 0.01}$ | 48.2% |

Table 5: **Dataset 4** (100-dimensional IVs): Comparison of methods (Naïve vs Ours) on coverage and width metrics with relative performance improvement. Note: "—" means that there are no reliable runs for which the corresponding performance metrics could be calculated.

the higher-dimensional setting, we varied the hyperparameter $k$ over $[2, 5, 7, 10, 20]$ for all bound estimation methods. We observe similar patterns as for our other dataset. In particular, the DeepIV baseline fails *entirely* to provide reliable bounds. In summary, our method shows robust performance by providing tighter and more reliable bounds than the baseline, even in high-dimensional settings. This emphasizes the applicability of our bounds in even more complex settings.

### E.4 Ablation studys

To further examine the robustness of our method in non-standard settings, we perform two additional ablation studies, one for varying the DGP and one for varying the selected nuisance models.

**Linear DGP:** To analyze if our flexible method also performs robustly in simple settings, we evaluate our method which uses neural networks at every stage on a simple linear DGP. For that we adapt our Dataset 3 and use linear functions for the dependencies between the variables. We report the results in Table 6. As expected, our method performs also robustly in the simpler linear setting and outperforms the baseline by a clear margin again. Summarized, our method shows strong performance which emphasizes its applicability to datasets of various complexity levels.

| Metric | Naïve | Ours | Rel. Improve |
|---|---|---|---|
| **Coverage[↑]** | $1.00 \pm 0.00$ | $\mathbf{1.00 \pm 0.00}$ | 0.0 |
| **Coverage*[↑]** | $0.92 \pm 0.18$ | $\mathbf{1.00 \pm 0.00}$ | 8.6% |
| **Width*[↓]** | $2.07 \pm 0.04$ | $\mathbf{1.99 \pm 0.05}$ | 3.9% |
| **MSE*[↓]** | $0.10 \pm 0.01$ | $\mathbf{0.08 \pm 0.01}$ | 20.0% |
| **MSD[↓]** | $0.08 \pm 0.08$ | $\mathbf{0.04 \pm 0.03}$ | 50.0% |

Table 6: **Linear DGP**: Comparison of methods across key metrics. Relative performance improvements in green.

**Non-linear DGP with linear models:** In our method, we leverage neural networks at all stages to allow for consistent and flexible estimation of all properties. However, since our method is model-agnostic in principle, we analyze the behavior of our method when using non-flexible (mis-specified) models. For that, we implement our method and the baseline by using linear models for the nuisance estimates and evaluate the performance on our non-linear Dataset 3 (i.e., the nuisances and the bounds are misspecified). We report the results in Table 7. As expected, because of the misspecification of the nuisance models, full coverage of the bounds cannot be guaranteed. However, our method still outperforms the naive baseline evidently with respect to coverage and MSD while yielding similar bound tightness. Further, with coverage to the oracle bounds over 90% and low MSD, our method still predicts close to valid bounds robustly over different runs which is unlike the naive baseline. This shows that our method is also robust against misspecification of the nuisance models as when using linear models for non-linear datasets.

| Metric | Naïve | Ours | Rel. Improve |
|---|---|---|---|
| **Coverage[↑]** | $0.96 \pm 0.06$ | $\mathbf{1.00 \pm 0.00}$ | 4.1% |
| **Coverage*[↑]** | $0.59 \pm 0.28$ | $\mathbf{0.91 \pm 0.04}$ | 54.2% |
| **Width*[↓]** | $1.91 \pm 0.02$ | $\mathbf{1.91 \pm 0.03}$ | 0.0% |
| **MSE*[↓]** | $0.14 \pm 0.04$ | $\mathbf{0.14 \pm 0.02}$ | 0.0% |
| **MSD[↓]** | $0.20 \pm 0.11$ | $\mathbf{0.02 \pm 0.01}$ | 90.0% |

Table 7: **Non-linear DGP with linear nuisance models**: Comparison of methods across key metrics. Relative performance improvements in green.

# F ROLE OF NUMBER OF PARTITIONS $k$

## F.1 WHY OUR METHOD IS ROBUST TO DIFFERENT CHOICE OF $k$

One major advantage of our method is that it is clearly less sensitive to the hyperparameter $k$ than, for example, the naïve baseline. Empirically, we demonstrate this in our experiments by lower variance and stable behavior over varying $k$, especially visible in the low values of MSD. This is due to the combination of learning flexible representations tailored to minimize bound width (allowing us to estimate tight bounds already for low $k$) while ensuring reliable estimates of the nuisance functions in the second stage by using our regularization loss in Eq. (**??**) (ensuring robust behavior also for higher $k$).

Note that the robustness of our method is especially beneficial when applying our method to real-world settings in causal inference. In real-world settings from causal inference, hyperparameter tuning and model evaluation are not directly possible because oracle CATE or oracle bounds are not known. Thus, the robustness against suboptimal selection of hyperparameters such as $k$ is crucial. In the following, we provide further high-level theoretical insights into the role of $k$ and propose practical recommendations for selecting $k$ in real-world applications.

**Estimation error for different** $k$: The hyperparameter $\lambda$ controls the regularization loss in Eq. (**??**), i.e., it tries to maximize $\hat{p}_{\ell,\phi} = \hat{\mathbb{P}}(\phi_\theta(Z) = \ell) > \varepsilon$ for all $\ell \in 1, \ldots, k$. Thus, if we choose $\lambda$ high enough, then we enforce that $\hat{p}_{\ell,\phi} = 1/k$ for all $\ell \in 1, \ldots, k$. Plugged into Theorem 12, the asymptotic variances for the nuisance estimators are $k\left(\frac{\mathrm{Var}(g(Z)|\phi(Z)=\ell)}{c} + d\right)$ for $\hat{\mu}^a_\phi(x, \ell)$, and $k\left(\mathrm{Var}(h(Z) \mid \phi(Z) = \ell)\right)$ for $\hat{\pi}_\phi(x, \ell)$, respectively. Thus, for large enough $\lambda$, the variance of the nuisance estimators (and, thus, also likely of the final bounds) will increase for increasing $k$. However, as an interesting side note, for a fixed (not too large) $\lambda$, the penalization term in Eq. (**??**) will also grow with growing $k$ due to the same reason, which yields an automated stabilization for higher $k$. This is also shown in our experiments where higher values of $k$ do *not* necessarily result in a higher variance.

**Bound tightness for different** $k$: On a population level, the bounds get tighter with growing $k$. This follows straightforwardly from Theorem 1, since using more $k$ increases the flexibility of $\phi$. While the exact bound width is highly non-trivial, we can use results from Schweisthal et al. (2024) about bounds for the CATE with discrete instruments to give some intuition. Specifically, in our setting, for some $x$, the bound width is bounded by $b^+_\phi(x) - b^-_\phi(x) \leq \min_{l,m}\{(s_2 - s_1)(2 - \pi_\phi(x, \ell) - (1 - \pi_\phi(x, m)))\}$ with $\ell, m \in \{1, \ldots, k\}$. This has two major implications. First, if for some $x$, $\phi$ is learned such that $\phi(x, \ell)$ is close to 1 for some $l$ and $\pi_\phi(x, m)$ is close to 0 for some $m$, the bound width is close to zero ("point identification"). Second, if the optimal partitioning function $\phi$ is the same for all $x$ (implying $b(x) = b$), then setting $k = 3$ can be sufficient to yield the tightest bounds. This is because, by using a flexible network for $\phi$, the partitions can be learned such that partition 1 yields propensity scores as close as possible to zero (as the data allows), partition 2 yields propensity scores as close as possible to 1, and partition 3 contains all $z$ resulting in propensity scores between those values. Note, however, that this is only valid in population but can result in highly unreliable estimation in finite sample data.

## F.2 PRACTICAL GUIDELINES FOR SELECTING $k$

Although we showed that our method is designed to be robust against different selections of $k$, we provide two potential guidelines for how to choose $k$ in real-world settings where ground-truth CATE or bounds are not available for model selection.

*Approach 1: Expert-informed approach.* In some medical applications, physicians might already know or make an educated guess about a number of underlying clusters of patient characteristics such as genetic variants. For instance, this is a common assumption in subgroup identification or latent class analysis in medicine where patient groups are characterized by having similar responses to treatments or showing similar associations with diseases (Kongsted & Nielsen, 2017). Thus, no data-driven approach is necessary here but one can integrate existing domain knowledge.

*Approach 2: Data-driven for hypothesis confirmation.* Often, physicians are interested in whether some treatment or exposure has a positive or negative effect (i.e., lower bound $> 0$ or upper bound

$< 0$) for at least some observations $x$. Thus, $k$ can be selected by increasing $k$ until such an effect can be observed while holding the variance minimal. Then, the variance can be approximated (e.g., by bootstrapping to test for the reliability of the corresponding bound model and its effect). Thus, this approach can be used when our method is used as a support tool for hypothesis confirmation.

Last, straightforwardly, from an exploratory perspective, all hyperparameters ($k$, $\lambda$, $\gamma$) can be altered together to examine the behavior of bound width and estimation variance to post-hoc find a suitable hyperparameter configuration for a dataset that fulfills the subjective preferences of the practitioner.

## G SENSITIVITY ANALYSIS

We perform a sensitivity analysis over the hyperparameters in our custom loss function. We report the results in Fig. 6 and Fig. 7 for dataset 3 and for $k = 3$. We observe that $\gamma$ does not affect the bound size but can be optimized to reduce estimation variance, as mentioned in the motivation of our auxiliary guidance loss. Thus, $\lambda$ demonstrates the trade-off between tightness and variance and shows the importance of our regularization loss. Here, $\lambda$ can be increased to reduce the variance. In our experiments, the optimal trade-off between reduced variance and bound tightness also results in optimal oracle coverage, showing the practicability of our regularization.

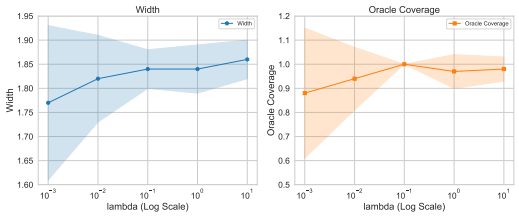


Figure 6: Sensitivity over $\lambda$. Left: Average bound width. Right: Oracle coverage. Averaged over 5 runs $\pm$ sd.

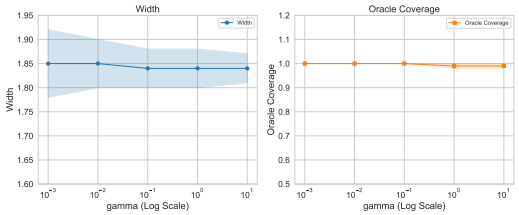


Figure 7: Sensitivity over $\gamma$. Left: Average bound width. Right: Oracle coverage. Averaged over 5 runs $\pm$ sd.

## H TRAINING PROCEDURE

**Algorithm 1:** Two-stage learner for estimating bounds with complex instruments

**Input :** observational data sampled from $(Z, X, A, Y)$, epochs $e$, batch size $n_b$, neural network $\phi_\theta$ with parameters $\theta$, learning rate $\delta$

**Output :** bounds $\hat{b}^-_{\phi_\theta}(x), \hat{b}^+_{\phi_\theta}(x)$

`// First stage (nuisance estimation)`

$\hat{\mu}^a(x,z) \leftarrow \hat{\mathbb{E}}[Y \mid X = x, A = a, Z = z]$

$\hat{\pi}(x,z) \leftarrow \hat{\mathbb{P}}(A = 1 \mid X = x, Z = z)$

$\hat{\eta}(z) \leftarrow \hat{\mathbb{P}}(A = 1 \mid Z = z)$

`// Second-stage (partition learning and bound calculation)`

**for** $\epsilon \in \{1, \ldots, e\}$ *in batches* **do**

  **for** $\ell \in \{1, \ldots, k\}$ **do**

    $\hat{\mu}^a_{\phi_\theta}(x,\ell) = \frac{1}{\sum_j^{n_b} \mathbb{1}\{\phi_\theta(z_j)=\ell, A=a)\}} \sum_j^{n_b} \hat{\mu}^a(x,z_j)\mathbb{1}\{\phi_\theta(z_j)=\ell\}(a\hat{\eta}(z_j) + (1-a)(1-\hat{\eta}(z_j)))$

    $\hat{\pi}_{\phi_\theta}(x,\ell) = \frac{1}{\sum_j^{n_b} \mathbb{1}\{\phi_\theta(z_j)=\ell\}} \sum_j^{n_b} \hat{\pi}(x,z_j)\mathbb{1}\{\phi_\theta(z_j)=\ell\})$

  **end**

  $\hat{b}^+_{\phi_\theta}(x) = \min_{l,m} \hat{b}^+_{\phi_\theta;l,m}(x), \quad \hat{b}^-_{\phi_\theta}(x) = \max_{l,m} \hat{b}^-_{\phi_\theta;l,m}(x)$ for $l, m \in \{1, \ldots, K\}$

  $\mathcal{L}(\theta) \leftarrow \mathcal{L}_{\text{b}}(\theta) + \lambda\mathcal{L}_{\text{reg}}(\theta) + \gamma\mathcal{L}_{\text{aux}}(\theta)$ as per Sec. 4

  $\theta \leftarrow \theta - \delta\nabla_\theta\mathcal{L}(\theta)$

**end**

`// Final bounds`

**return** $\hat{b}^-_{\phi_\theta}(x), \hat{b}^+_{\phi_\theta}(x)$