A Proof of Theorem 1

Here, we only focus on the minimization problem, i.e., $\underline{\hat{\text{ATD}}} \to \underline{\text{ATD}}$. The proof of $\overline{\hat{\text{ATD}}} \to \overline{\text{ATD}}$ will similarly follow.

Assumptions

Before stating our proof, we list the assumptions needed as follows. These are mainly technical assumptions that will simplify the derivations.

1. In this proof we assume all random variables are one-dimensional. [4]
2. Since the true SCM $\mathcal{M}$ is $\mathcal{G}$-constrained (Assumption 1) and linear, we will consider linear $\mathcal{G}$-constrained SCMs as our parameter search space. More concretely, each random variable V_i in an SCM $\mathcal{M}_\mathcal{G}^\theta$ can be written in the following form:
$$V_i = \boldsymbol{\theta}_{V_i}^\top \mathbf{pa}(V_i) + \hat{\boldsymbol{\theta}}_{V_i}^\top \hat{\mathbf{U}}_{\mathbf{C}_i} \tag{14}$$
where $\boldsymbol{\theta}_{V_i} \in \mathbb{R}^{|\mathbf{pa}(V_i)|}$, $\hat{\boldsymbol{\theta}}_{V_i} \in \mathbb{R}^{|\hat{\mathbf{U}}_{\mathbf{C}_i}|}$, and $\theta \in \Theta$ is the concatenation of all $\boldsymbol{\theta}_{V_i}, \hat{\boldsymbol{\theta}}_{V_i}$ in a topological order of V_is. Note that $\theta \in \mathbb{R}^{|E|}$, where E is the set of edges in the causal graph $\mathcal{G}$, containing edges to both observed and unobserved random variables.
3. We assume the set of feasible parameters Θ is a bounded closed subset of $\mathbb{R}^{|E|}$. In practice, even in non-identifiable cases with infinite bounds, we use regularization to ensure the parameters of the network are bounded.
4. The induced probability over observed random variables $P_{\mathcal{M}_\mathcal{G}^\theta}$, which we will write as P_θ, and the true distribution P belong to the Wasserstein space of order $p = 1$, which we refer to as $\mathcal{P}_1$. In other words, $\int_{\mathbb{R}^d} |\mathbf{v}_0 - \mathbf{v}| \, dP_\theta(\mathbf{v}) < +\infty$ for any arbitrary point $\mathbf{v}_0 \in \mathbb{R}^d$. Again, in practice, since the support of all observed random variables is bounded, all probability distributions defined on them belong to the Wasserstein space.

Before the main proof, we will state and prove the following useful lemmas.

Lemma 1. *Let define the set of feasible parameters for n number of samples as $\Theta^n = \{\theta \in \Theta;\ W_1(P_{\mathcal{M}_\mathcal{G}^\theta}, P^n) \le \alpha_n\}$. Then, $P_{\mathcal{M}_\mathcal{G}^{\theta^n}} \to P$ weakly for every sequence of $\theta^n \in \Theta^n$.*

Proof. First, note that the empirical distribution P^n weakly converges to P as $n \to \infty$. Since W_1 metrizes $\mathcal{P}_1$, we have $W_1(P^n, P) \to 0$ [Villani, 2009]. Hence, we can choose the sequence of α_n for the distributional constraint, such that $W_1(P, P^n) \le \alpha_n$ and $\alpha_n \to 0$ as $n \to \infty$. For any parameter $\theta^n \in \Theta^n$, we have $W_1(P_{\mathcal{M}_\mathcal{G}^{\theta^n}}, P^n) \le \alpha_n\}$. Therefore, $W_1(P_{\mathcal{M}_\mathcal{G}^{\theta^n}}, P^n) \to 0$. Using the triangle inequality, we have

$$W_1(P_{\mathcal{M}_\mathcal{G}^{\theta^n}}, P) \le W_1(P_{\mathcal{M}_\mathcal{G}^{\theta^n}}, P^n) + W_1(P^n, P) \tag{15}$$

Therefore, $W_1(P_{\mathcal{M}_\mathcal{G}^{\theta^n}}, P) \to 0$ or equivalently, $P_{\mathcal{M}_\mathcal{G}^{\theta^n}} \to P$ weakly. □

Lemma 2. *For any linear $\mathcal{G}$-constrained SCM $\mathcal{M}_\mathcal{G}^\theta$, we have*

$$\mathbf{V}_\theta = \mathbf{A}(\theta)\hat{\mathbf{U}} \tag{16}$$

where each element in the matrix $\mathbf{A}(\theta) \in \mathbb{R}^{d \times |dim(\hat{\mathbf{U}})|}$ is a continuous function of θ.

Proof. Consider a topological order of observed variables $(V_1, \cdots, V_d)$. We prove the result by induction. For V_1, from eq. 14, we have

$$V_1 = 0 + \hat{\boldsymbol{\theta}}_{V_1}^\top \hat{\mathbf{U}}_{\mathbf{C}_1} = \boldsymbol{\phi}_{V_1}^\top \hat{\mathbf{U}} \tag{17}$$

---

[4] For high-dimensional variables, if the node-level causal graph is given, i.e., the relation between each dimension of variables, we can convert each multi-dimensional variable to multiple one-dimensional ones and follow the same proof. If the node-level causal graph is unknown and there is no inter-dependence between each dimension, one can follow the same proof technique by assuming V_i as a vector in eq. 14. We leave these extensions as future work.

where the elements of $\boldsymbol{\phi}_{V_1}$ matches $\hat{\boldsymbol{\theta}}_{V_1}$ for $\hat{\mathbf{U}}_{\mathbf{C}_1}$ and equal to zero for $\hat{\mathbf{U}} \backslash \hat{\mathbf{U}}_{\mathbf{C}_1}$. Now, assume all variables $V_1, \cdots, V_{d-1}$ can be written as $\boldsymbol{\phi}_{V_i}^\top \hat{\mathbf{U}}$, where $\boldsymbol{\phi}_{V_i}$ is a continuous function of θ. Then, from eq. 14,

$$V_d = \boldsymbol{\theta}_{V_d}^\top \mathbf{pa}(V_d) + \hat{\boldsymbol{\theta}}_{V_d}^\top \hat{\mathbf{U}}_{\mathbf{C}_d} = \boldsymbol{\theta}_{V_d}^\top (\boldsymbol{\phi}_{pa_1(V_d)}^\top \hat{\mathbf{U}}, \cdots, \boldsymbol{\phi}_{pa_r(V_d)}^\top \hat{\mathbf{U}})^\top + \hat{\boldsymbol{\theta}}_{V_d}^\top \hat{\mathbf{U}}_{\mathbf{C}_d} = \boldsymbol{\phi}_{V_d}^\top \hat{\mathbf{U}} \tag{18}$$

where $\boldsymbol{\phi}_{V_d}$ is a linear function of $\boldsymbol{\theta}_{V_d}, \boldsymbol{\phi}_{V_1}, \cdots, \boldsymbol{\phi}_{V_{d-1}}$, and $\hat{\boldsymbol{\theta}}_{V_d}$. Defining matrix $\mathbf{A}(\theta)$ with rows $\boldsymbol{\phi}_{V_i}$ concludes the proof. □

Now, we are ready to prove the main result.

Theorem 2 (Tight Bounds)**.** *Assume the dataset $\{\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(n)}\}$ is generated from a linear SCM. Then, under assumptions 1, and 2, the solution to the constrained optimization problem in eq. 7 converges to the optimal bound over the ATD in infinite samples, i.e.,* $(\hat{\underline{ATD}}, \hat{\overline{ATD}}) \to (\underline{ATD}, \overline{ATD})$.

Proof. The goal is to show the solution to

$$\min_{\theta \in \Theta} \mathrm{ATD}_{\mathcal{M}_{\mathcal{G}}^{\theta}} \text{ s.t. } W_1\left(P_{\mathcal{M}_{\mathcal{G}}^{\theta}}, P^n\right) \leq \alpha_n \tag{19}$$

converges to the solution to

$$\min_{\theta \in \Theta} \mathrm{ATD}_{\mathcal{M}_{\mathcal{G}}^{\theta}}, \quad \text{s.t. } P_{\mathcal{M}_{\mathcal{G}}^{\theta}} = P \tag{20}$$

as $n \to \infty$. We first aim to re-write the value of ATD. Note that, in a linear $\mathcal{G}$-constrained SCM, the partial derivative $\frac{\partial Y_{\mathcal{M}_{\mathcal{G}}^{\theta}(T=t)}(\mathbf{u})}{\partial t}$ is only a function of SCM parameters θ. Let define $g(\theta) = \frac{\partial Y_{\mathcal{M}_{\mathcal{G}}^{\theta}(T=t)}(\mathbf{u})}{\partial t}$. Then,

$$\mathrm{ATD}_{\mathcal{M}_{\mathcal{G}}^{\theta}} = \int_{\Omega} \frac{\partial Y_{\mathcal{M}_{\mathcal{G}}^{\theta}(T=t)}(\mathbf{u})}{\partial t}\Big|_{t=T(\mathbf{u})} dP_{\hat{\mathbf{U}}}(\mathbf{u}) = \int_{\Omega} g(\theta)\, dP_{\hat{\mathbf{U}}}(\mathbf{u}) = g(\theta) \tag{21}$$

Therefore, we need to show the following to conclude the proof:

$$\min_{\theta^n \in \Theta^n} g(\theta^n) \to \min_{\theta^\infty \in \Theta^\infty} g(\theta^\infty) \tag{22}$$

where

$$\begin{aligned} \Theta^n &= \{\theta \in \Theta;\ W_1(P_\theta, P^n) \leq \alpha_n\} \\ \Theta^\infty &= \{\theta \in \Theta;\ P_\theta = P\} = \{\theta \in \Theta;\ W_1(P_\theta, P) = 0\} \end{aligned} \tag{23}$$

Since $\Theta^\infty \subseteq \Theta^n$ for each $n \in \mathbb{N}$, we know that

$$\min_{\theta^n \in \Theta^n} g(\theta^n) \leq \min_{\theta^\infty \in \Theta^\infty} g(\theta^\infty) \tag{24}$$

It is sufficient to show that, for each $\epsilon > 0$, there is n_0 such that for all $n > n_0$ we have

$$\min_{\theta^n \in \Theta^n} g(\theta^n) \geq \min_{\theta^\infty \in \Theta^\infty} g(\theta^\infty) - \epsilon \tag{25}$$

Suppose this is not true, i.e., there exists $\epsilon > 0$ such that for each $n \in \mathbb{N}$ we have

$$\min_{\theta^n \in \Theta^n} g(\theta^n) < g(\theta^\infty) - \epsilon \tag{26}$$

for all $\theta^\infty \in \Theta^\infty$. Let $\theta_\star^n = \arg\min_{\theta \in \Theta^n} g(\theta)$. The sequence $(\theta_\star^n)_{n \in \mathbb{N}}$ is a subset of Θ and therefore is a bounded sequence in $\mathbb{R}^{|E|}$. Thus, by Bolzano-Weierstrass theorem, there exists a convergent sub-sequence $(\theta_\star^{n_i})_{i \in \mathbb{N}}$ that converges to some fixed parameter θ_0. Also, $\theta_0 \in \Theta$ as Θ is closed. Now, since g is continuous, we have

$$g(\theta_0) = \lim_{i \to \infty} g(\theta_\star^{n_i}) \leq g(\theta^\infty) - \epsilon \tag{27}$$

for all $\theta^\infty \in \Theta^\infty$. Hence, $\theta_0 \notin \Theta^\infty$.

On the other hand, using Lemma 2, we have $\mathbf{V}_{\theta_\star^{n_i}} = \mathbf{A}(\theta_\star^{n_i})\hat{\mathbf{U}}$. Since $\mathbf{A}(\theta_\star^{n_i})$ is a continuous function of $\theta_\star^{n_i}$, from the continuous mapping theorem, we have $P_{\theta_\star^{n_i}} \to P_{\theta_0}$ weakly. Also, from Lemma 1, we have $P_{\theta_\star^{n_i}} \to P$. Therefore, $P = P_{\theta_0}$. Since $\theta_0 \in \Theta$, we conclude that $\theta_0 \in \Theta^\infty$, a contradiction. □

Remark. In this proof, we did not separate identifiable and non-identifiable cases. In fact, the notion of identifiability can be seen as a property of the set of feasible parameters Θ^n and Θ^∞. For example, in identifiable cases, we expect the set Θ^∞ to only contain one element while in non-identifiable cases, it consists of multiple possible solutions. Our proof holds as long as Θ^n and Θ^∞ are bounded subsets of $\mathbb{R}^{|E|}$.

B Proof of Corollary 1

Similar to the proof of Theorem 1, define the set of feasible parameters as $\Theta^n = \{\theta \in \Theta \;\; W_1(P_{\mathcal{M}_\mathcal{G}^\theta}, P^n) \leq \alpha_n\}$. Then,

$$
\begin{aligned}
\text{ATE}_{\mathcal{M}_\mathcal{G}^\theta}(d) &= \mathbb{E}_{\mathbf{u}\sim P_{\hat{\mathbf{U}}}}\left[Y_{\mathcal{M}_\mathcal{G}^\theta(T=d)}(\mathbf{u}) - Y_{\mathcal{M}_\mathcal{G}^\theta(T=t_0)}(\mathbf{u})\right] \\
&= \mathbb{E}_{\mathbf{u}\sim P_{\hat{\mathbf{U}}}}\left[\int_{t_0}^{d} \frac{\partial Y_{\mathcal{M}_\mathcal{G}^\theta(T=t)}}{\partial t}\, dt\right] \\
&= (d - t_0)\cdot \mathbb{E}_{\mathbf{u}\sim P_{\hat{\mathbf{U}}}}\left[\int_{t_0}^{d} \frac{\partial Y_{\mathcal{M}_\mathcal{G}^\theta(T=t)}}{\partial t}\,\frac{1}{d-t_0}\, dt\right] \\
&= (d - t_0)\cdot \mathbb{E}_{\mathbf{u}\sim P_{\hat{\mathbf{U}}}}\left[\mathbb{E}_{t\sim \texttt{Unif}[t_0,d]}\left[\frac{\partial Y_{\mathcal{M}_\mathcal{G}^\theta(T=t)}}{\partial t}\right]\right] \\
&= (d - t_0)\cdot \text{UATD}_{\mathcal{M}_\mathcal{G}^\theta}[t_0, d] \qquad (28)
\end{aligned}
$$

Therefore,

$$\theta^* = \arg\min_{\theta\in\Theta^n} \text{UATD}_{\mathcal{M}_\mathcal{G}^\theta}[t_0,d] = \arg\min_{\theta\in\Theta^n} (d-t_0)\cdot \text{UATD}_{\mathcal{M}_\mathcal{G}^\theta}[t_0,d] = \arg\min_{\theta\in\Theta^n} \text{ATE}_{\mathcal{M}_\mathcal{G}^\theta}(d) \tag{29}$$

For the second part, similar to the proof of Theorem 1, we have $\frac{\partial Y_{\mathcal{M}_\mathcal{G}^\theta(T=t)}}{\partial t} = g(\theta)$ for some continuous function g. Then,

$$\text{ATE}_{\mathcal{M}_\mathcal{G}^\theta}(d) \overset{eq.\ 28}{=} (d-t_0)\cdot \mathbb{E}_{\mathbf{u}\sim P_{\hat{\mathbf{U}}}}\left[\mathbb{E}_{t\sim\texttt{Unif}[t_0,d]}\left[\frac{\partial Y_{\mathcal{M}_\mathcal{G}^\theta(T=t)}}{\partial t}\right]\right] = (d-t_0)\cdot g(\theta) \tag{30}$$

The rest of the proof can be directly derived from the proof of Theorem 1.

C ATE and ATD

Here, we provide more intuition on the difference between ATE and ATD. Consider a simple setting with only two random variables T and Y, where the causal graph is $T \to Y$. Here, the average treatment effect is identifiable and can be computed (with infinite samples) using the following formula:

$$\text{ATE}_\mathcal{M}(d) = \mathbb{E}_P[Y|T=d] - \mathbb{E}_P[Y|T=t_0] \tag{31}$$

Now, assume we have a finite dataset $\mathcal{D} = \{(t^{(1)}, y^{(1)}), \cdots, (t^{(n)}, y^{(n)})\}$, where $t^{(i)} \neq t_0$ and $t^{(i)} \neq d$ for all $i \in [n]$. One way to find a (high probability) upper bound on the value of ATE is to solve the following direct optimization:

$$\max_\theta \text{ATE}_{\mathcal{M}_\mathcal{G}^\theta}(d) \quad \text{s.t.} \quad W_1(P_{\mathcal{M}_\mathcal{G}^\theta}, P^n) \leq \alpha_n \tag{32}$$

With no assumption on the regularity of the response functions $Y_{\mathcal{M}_\mathcal{G}^\theta(T=t)}$, it is possible to find a solution to eq. 32 that matches all points in $\mathcal{D}$, i.e., $W_1(P_{\mathcal{M}_\mathcal{G}^\theta}, P^n) = 0$, while getting arbitrarily values of $\text{ATE}_{\mathcal{M}_\mathcal{G}^\theta}(d)$. See Figure 5 for a demonstration. This shows that ATE, in the continuous treatment setting with finite number of samples, is not well-behaved and it is generally impossible to find informative non-parametric bounds on that (see also Gunsilius [2020]). On the other hand, ATD, which is the average partial derivative of response function w.r.t. the observed treatment distribution is defined *globally* over the support of T. Therefore, it is not possible to maximize ATD arbitrarily without violating the distributional constraint in this setting.

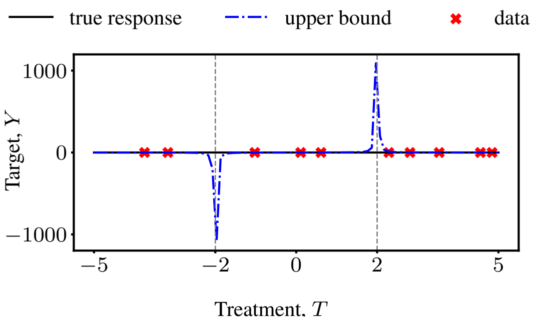

Figure 5: (a) Irregularity of the ATE objective. In finite sample setting, it is possible to find a model that matches the empirical distribution, while creating arbitrarily large values of ATE between $T = 2$ and $T = -2$.

D Algorithm

Algorithm 1 Partial Identification of Average Treatment Derivatives

Input: Dataset $\mathcal{D} = \{X^{(i)}, T^{(i)}, Y^{(i)}\}_{i=1}^{N}$, causal graph $\mathcal{G}$, Sinkhorn approximation parameter ϵ, learning rate γ, Lagrange learning rate γ_L

Output: $\underline{\hat{\text{ATD}}}$: lower bounds on ATD ▷ The algorithm to find $\overline{\hat{\text{ATD}}}$ is similar

Phase 1: Matching Distributions

```
Initialize $\mathcal{M}_\mathcal{G}^{\theta^{(0)}}$ and $i = 0$
while $S_\epsilon(\hat{D}, D)$ not converged do    ▷ Train the model to approximate the observed distribution.
    $\hat{\mathcal{D}} \leftarrow \{\hat{X}^{(j)}, \hat{T}^{(j)}, \hat{Y}^{(j)}\}_{j=1}^{N} \sim \mathcal{M}_\mathcal{G}^{\theta^{(i)}}$    ▷ Generate dataset from the trained SCM
    $\theta^{(i+1)} \leftarrow \theta^{(i)} - \gamma \nabla_\theta S_\epsilon(\hat{\mathcal{D}}, \mathcal{D})$
    $i \leftarrow i + 1$
end while
$\alpha \leftarrow S_\epsilon(\hat{\mathcal{D}}, \mathcal{D})$    ▷ Set the distributional constraint level to the minimum Sinkhorn loss
```

Phase 2: Joint Optimization Phase

```
Initialize Lagrange multiplier $\lambda^{(0)}$ and $j = 0$
while $\text{ATD}_{\mathcal{M}_\mathcal{G}^{\theta^{(i+j)}}}$ not converged do
    $\text{ATD}_{\mathcal{M}_\mathcal{G}^{\theta^{(i+j)}}} \leftarrow$ calculate ATD using eq. 13
    # Alternate optimization
    $\bar{\theta}^{(i+j)} \leftarrow \theta^{(i+j)} - \gamma \nabla_\theta \text{ATD}_{\mathcal{M}_\mathcal{G}^{\theta^{(i+j)}}}$    ▷ Update the parameters to minimize the ATD
    $\hat{\mathcal{D}} \leftarrow \{\hat{X}^{(k)}, \hat{T}^{(k)}, \hat{Y}^{(k)}\}_{k=1}^{N} \sim \mathcal{M}_\mathcal{G}^{\bar{\theta}^{(i+j)}}$
    $\theta^{(i+j+1)} \leftarrow \bar{\theta}^{(i+j)} - \gamma \nabla_\theta S_\epsilon(\hat{\mathcal{D}}, \mathcal{D})$ ▷ Update the parameters to minimize the Sinkhorn loss
    # Lagrange multiplier update
    $\lambda^{(j+1)} \leftarrow \lambda^{(j)} + \gamma_L (S_\epsilon(\hat{\mathcal{D}}, \mathcal{D}) - \alpha)$
    $j \leftarrow j + 1$
end while
return $\text{ATD}_{\mathcal{M}_\mathcal{G}^{\theta^{(i+j)}}}$
```

Extension of Algorithm 1 to ATEs. We can use a similar method to Algorithm 1 for partial identification of average treatment effects (ATEs). The only difference is that, instead of maximizing/minimizing $\text{ATD}_{\mathcal{M}_\mathcal{G}^{\theta}}$, we optimize for the value of $\text{UATD}_{\mathcal{M}_\mathcal{G}^{\theta}}[t_0, d]$. More concretely, we use the following approximation to estimate $\text{UATD}_{\mathcal{M}_\mathcal{G}^{\theta}}[t_0, d]$:

$$\text{UATD}_{\mathcal{M}_\mathcal{G}^{\theta}}[t_0, d] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\epsilon} \left[Y_{\mathcal{M}_\mathcal{G}^{\theta}(T=t^{(i)}+\epsilon)}(\mathbf{u}^{(i)}) - Y_{\mathcal{M}_\mathcal{G}^{\theta}(T=t^{(i)})}(\mathbf{u}^{(i)}) \right] \tag{33}$$

Table 1: The bounds derived by our method over the ATE. The results include the optimal bound.

Causal Graph	Ours	Optimal Bound	True Value
Front-door (Discrete)	(0.4374, 0.5322)	–	0.5085
IV (Discrete)	(-0.5629, -0.0821)	(-0.55, -0.15)	-0.25

where $\{t^{(i)}\}_{i=1}^{n}$ are samples from a uniform distribution within $[t_0, d]$ with a Gaussian tail, and $\{\mathbf{u}^{(i)}\}_{i=1}^{n}$ are the latent variables generated from a uniform distribution. Note that, the only difference between eq. 33 and eq. 13 is the distribution of treatment variables, where in the former we use a uniform distribution with Gaussian tail, while the latter uses the same distribution of treatments in the observed dataset. In our experiments, for the uniform distribution with Gaussian tail, we generate samples from $\mathcal{N}(\mu, \sigma)$, where $\mu = \frac{t_0+d}{2}$ and $\sigma = \frac{d-t_0}{2}$. Then, for each sample within $[t_0, d]$, we generate a new sample from uniform distribution $\texttt{Unif}[t_0, d]$.

E Additional Experiments

E.1 Discrete Setting

To showcase the generality of our framework, we study two datasets with binary treatments. We consider the binary IV dataset described in Duarte et al. [2021], where the true value of ATE is not identifiable, but the optimal bound is known. We also use the Front-door binary dataset in Zhang et al. [2021] where the causal effect is identifiable (See Appendix G). Here, the partial derivatives do not exist, so we directly optimize the ATE. Note that, in the discrete setting, the network cannot generate arbitrary large values in the intervention points without violating the distributional constraint. Table 1 shows our derived bounds and compares them to the optimal bounds. In the identifiable Front-door causal graph, we find a tight bound over the true ATE. In the non-identifiable IV setting, our bound includes the optimal bound with a small gap.

E.2 ACIC Dataset

To demonstrate the applicability of our method on datasets with higher-dimensional covariates, we consider the Atlantic Causal Inference Conference (ACIC) 2019 Data Challenge [Gruber et al., 2019]. The dataset is constructed based on the spam detection data from UCI [Dua and Graff, 2017]. The outcome of interest Y is whether an email is marked as spam or not. The treatment variable T is also a binary variable showing if the number of capital letters in an email exceeds a threshold. There are 22 continuous covariates that correspond to certain word frequencies.

We follow a similar setup as in Guo et al. [2022]. In particular, we consider the problem of partial identification under noisy measurements. Here, the data-generating causal graph is the Back-door setting, and the causal effect of T on Y is identifiable. However, we assume measurement error on the covariates by imposing synthetic noise on them and aim to estimate bounds on ATE under this uncertainty. Since our algorithm can incorporate uncertainty through the distributional constraint, it will lead to a valid and informative bound by choosing an appropriate value for α_n.

We generate a dataset of 2,000 samples using ACIC's data-generating process. Similar to Guo et al. [2022], we synthetically add five different levels of Gaussian noise with mean $\in (0.1, 0.2, 0.3, 0.4, 0.5)$ and standard deviation $\in (0.5, 0.5, 1, 1, 1)$. We then run our algorithm on these five noisy datasets and the original noiseless one, 10 different trials for each. Figure 6 illustrates our derived bounds and the true ATE for each of the noise levels. The results always include the actual value of ATE, while not being too conservative. As the noise level increases, our derived bounds get naturally less informative. [5]

F Implementation Choices

We use similar neural network architectures for each variable in the causal graph. For hyper-parameter tuning, we search over networks with $\{2, 3, 5\}$ hidden layers, $\{16, 64, 256\}$ neurons in each layer,

---

[5] We heuristically choose the value of hyper-parameter α_n by multiplying the minimum Sinkhorn divergence by factors of $(1.2, 1.3, 1.5, 1.6, 1.7)$ for the noise levels, respectively.

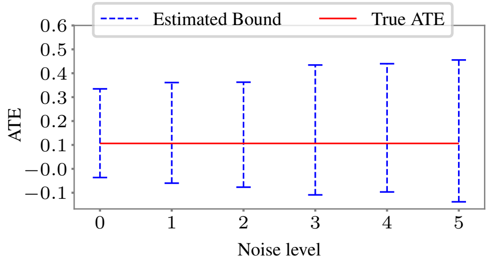

Figure 6: Partial identification of ATE in the ACIC dataset under noisy measurements. Level 0 corresponds to the original data, and 1 to 5 represents the noise levels, from lower to higher measurement error. The result bounds always include the true ATE while not being highly conservative (returning the full support $[-1, 1]$). We take the average over 10 runs with different random seed for each noise level.

learning rate within $\{0.001, 0.005, 0.01\}$, and Lagrange multiplier learning rate $\{0.5, 1\}$. The hyperparameters with the lowest Sinkhorn loss are chosen for each causal graph. We run all experiments for 500 epochs and use all samples ($N = 5000$) in each iteration. To find the level of distribution constraint α_n, we minimize the Sinkhorn loss and check the loss value on a validation set until it does not decrease for 30 epochs. We then choose the minimum value of Sinkhorn loss on the training data as α_n. We use an alternate optimization approach to maximize/minimize the ADE (or ATE). In each step, we first minimize the Sinkhorn loss and then maximize/minimize the value of ADE. We use gradient clipping for the ADE optimizer with the value of 0.2. The learning rate of the ADE optimizer is half the learning rate for the Sinkhorn loss optimizer. To find the upper (lower) bounds on ATD/ATE, we find their maximum (minimum) value within all steps that satisfy the distributional constraint. All implementation is done in PyTorch Lightning using Adam optimizer. To evaluate and calculate gradients of Sinkhorn loss, we use the "geomloss" library [Feydy et al., 2019]. [6]

G Data Generating Processes

Linear Back-door.

$$\begin{aligned} X &\sim \mathcal{N}(2,1) \\ T &\sim 0.1X^2 - X + \mathcal{N}(1,2) \\ Y &\sim 0.5T^2 - TX + \mathcal{N}(0,2) \end{aligned} \tag{34}$$

Nonlinear Back-door.

$$\begin{aligned} &X_1, X_2, X_3 \sim \mathcal{N}(1,1) \\ &T \sim X_1 - X_2 + 2X_3 + 2 + \mathcal{N}(0,3) \\ &Y \sim 3X_1 + X_2 - 0.5X_3 + 3T + \mathcal{N}(0,2) \end{aligned} \tag{35}$$

Front-door.

$$\begin{aligned} U &\sim \mathcal{N}(-1,1) \\ T &\sim U + \mathcal{N}(2,2) \\ X &\sim 2T + \mathcal{N}(1,2) \\ Y &\sim 0.25X^2 - X + U + \mathcal{N}(0,2) \end{aligned} \tag{36}$$

Linear IV (weak instrument, strong confounding).

$$\begin{aligned} Z_1 &\sim \mathcal{N}(-1,1) \\ Z_2 &\sim \mathcal{N}(0,1) \\ U &\sim \mathcal{N}(0,1) \\ T &\sim Z_1 - Z_2 + 0.5U + \mathcal{N}(0,1) \\ Y &\sim 0.5T - 3U + \mathcal{N}(0,1) \end{aligned} \tag{37}$$

---

[6] https://www.kernel-operations.io/geomloss/

Nonlinear IV (strong instrument, weak confounding).

$$\begin{aligned}
&Z_1 \sim \mathcal{N}(-1, 1)\\
&Z_2 \sim \mathcal{N}(0, 1)\\
&U \sim \mathcal{N}(0, 1)\\
&T \sim 3Z_1 + 1.5Z_2 + 0.5U + \mathcal{N}(0, 1)\\
&Y \sim 0.3T^2 - 1.5T + U + \mathcal{N}(0, 1)
\end{aligned} \tag{38}$$

Leaky Mediation.

$$\begin{aligned}
&U_1 \sim \mathcal{N}(1, 1)\\
&U_2 \sim \mathcal{N}(-1, 1)\\
&C \sim \mathcal{N}(0, 1)\\
&T \sim C + \mathcal{N}(0, 1)\\
&X_1 \sim T + U_1 + \mathcal{N}(0, 1)\\
&X_2 \sim 2T + U_2 + \mathcal{N}(0, 1)\\
&Y \sim -1.5X_1 + 2X_2 + U_1 + U_2 + C + \mathcal{N}(0, 1)
\end{aligned} \tag{39}$$

Binary IV [Duarte et al., 2021]. We use the noncompliance IV dataset in Section D.1 from Duarte et al. [2021]. The true value of ATE is -0.25 while the optimal bound is $[-0.55, -0.15]$.

Binary Front-door [Zhang et al., 2021].

$$\begin{aligned}
&U_1 \sim \texttt{Unif}(0, 1)\\
&U_2 \sim \mathcal{N}(0, 1)\\
&T \sim \texttt{Binomial}(1, U_1)\\
&W \sim \texttt{Binomial}(1, \frac{1}{1 + exp(-T - U_2)})\\
&Y \sim \texttt{Binomial}(1, \frac{1}{1 + exp(W - U_1)})
\end{aligned} \tag{40}$$

Linear IV with strong confounding [Padh et al., 2022].

$$\begin{aligned}
&Z \sim \mathcal{N}(0, 1)\\
&U \sim \mathcal{N}(0, 1)\\
&T \sim 0.5Z + 3U + \mathcal{N}(0, 1)\\
&Y \sim T - 6U + \mathcal{N}(0, 1)
\end{aligned} \tag{41}$$

Nonlinear IV with nonlinear interaction between treatment and confounding [Padh et al., 2022].

$$\begin{aligned}
&Z \sim \mathcal{N}(0, 1)\\
&U \sim \mathcal{N}(0, 1)\\
&T \sim 3Z + 0.5U + \mathcal{N}(0, 1)\\
&Y \sim 0.3T^2 - 1.5TU + \mathcal{N}(0, 1)
\end{aligned} \tag{42}$$