

A PROOFS

A.1 PROOF OF VALIDITY OF INSTRUMENT

Proof. We check the instrument conditions in order:

1. *Unconfounded Instrument:* $Z \perp\!\!\!\perp U$: The $Z \rightarrow X \leftarrow U$, $V \rightarrow X \leftarrow U$, and $X \rightarrow Y \leftarrow U$ triples are blocked by standard d-separation rules (Pearl et al., 2016). All paths from Z to U must pass through one of these triples so $Z \perp\!\!\!\perp U$.
2. *Exclusion:* $Z \perp\!\!\!\perp Y|X, U$: The $Z \rightarrow X \rightarrow Y$, $X \leftarrow U \rightarrow Y$, and $V \rightarrow X \rightarrow Y$ triples are blocked by standard d-separation rules. All paths from Z to Y must pass through one of these triples so $Z \perp\!\!\!\perp Y|X, U$.
3. *Relevance:* $Z \not\perp\!\!\!\perp X$: There is a $Z \rightarrow X$ edge, which is assumed to be non-degenerate.

Thus, Z is a valid instrument for determining the causal relationship between X and Y . \square

A.2 PROOF OF THEOREM 1

Proof. We simplify notation for clarity in our proof. Consider two vectors of the same dimension, \mathbf{a} and \mathbf{b} . Assume that $\sum_i a_i^2 \leq \epsilon$ and $\sum_i b_i^2 \leq \delta$. This implies that $\|\mathbf{a}\|_2 \leq \sqrt{\epsilon}$ and $\|\mathbf{b}\|_2 \leq \sqrt{\delta}$. Then, by the triangle inequality, $\|\mathbf{a} - \mathbf{b}\|_2 \leq \|\mathbf{a}\|_2 + \|\mathbf{b}\|_2 \leq \sqrt{\epsilon} + \sqrt{\delta}$. Setting $a_i = \sqrt{P(z)}(\mathbb{E}[Y|z] - \mathbb{E}_{\hat{x} \sim g(z)}[\hat{h}(\hat{x})])$ and $b_i = \sqrt{P(z)}(\mathbb{E}_{\hat{x} \sim g(z)}[\hat{h}(\hat{x})] - \mathbb{E}[\hat{h}(x)|z])$ proves that

$$\max_{\hat{h} \in \mathcal{H}} \mathbb{E}_Z[(\mathbb{E}_{x \sim g(z)}[\hat{h}(x)] - \mathbb{E}_{x \sim P(X|z)}[\hat{h}(x)])^2] \leq \delta, \quad (22)$$

$$\mathbb{E}_z[(\mathbb{E}[Y|z] - \mathbb{E}_{\hat{x} \sim g(z)}[\hat{h}(\hat{x})])^2] \leq \epsilon \quad (23)$$

$$\Rightarrow \text{PRMSE}(\hat{h}) = \sqrt{\mathbb{E}_z[(\mathbb{E}[Y|z] - \mathbb{E}_{x \sim P(X|z)}[\hat{h}(x)])^2]} \leq \sqrt{\epsilon} + \sqrt{\delta} \quad (24)$$

\square

A.3 PROOF OF THEOREM 2

Proof. The population version of (12) is

$$\min_{h \in \mathcal{H}} \max_{f \in \mathcal{F}} \mathbb{E}[2(Y - h(X))f(Z) - f^2(Z)] \quad (25)$$

An ϵ -approximate equilibrium is an (\hat{h}, \hat{f}) pair such that:

$$\max_{f \in \mathcal{F}} \mathbb{E}[2(Y - \hat{h}(X))f(Z) - f^2(Z)] - \frac{\epsilon}{2} \quad (26)$$

$$\leq \mathbb{E}[2(Y - \hat{f}(X))\hat{f}(Z) - \hat{f}^2(Z)] \quad (27)$$

$$\leq \min_{h \in \mathcal{H}} \mathbb{E}[2(Y - h(X))\hat{f}(Z) - \hat{f}^2(Z)] + \frac{\epsilon}{2} \quad (28)$$

Taking the derivative w.r.t $f(z)$ of the payoff and setting it equal to 0, we arrive at

$$2P(z)\mathbb{E}[Y - \hat{h}(X)|z] - 2P(z)f(z) = 0 \Rightarrow f(z) = \mathbb{E}[Y - \hat{h}(X)|z]. \quad (29)$$

Plugging this back into (35) gives us the inequality

$$\mathbb{E}_Z[\mathbb{E}[Y - \hat{h}(X)|z]^2] - \frac{\epsilon}{2} \leq \min_{h \in \mathcal{H}} \mathbb{E}[2(Y - h(X))\hat{f}(Z) - \hat{f}^2(Z)] + \frac{\epsilon}{2}. \quad (30)$$

Assuming we are in the realizable setting (e.g. $h(x) = \mathbb{E}[Y|do(x)] \in \mathcal{H}$), $\min_{h \in \mathcal{H}} \mathbb{E}[2(Y - h(X))\hat{f}(Z) - \hat{f}^2(Z)] \leq 0$. Thus, we can write that:

$$\mathbb{E}_Z[\mathbb{E}[Y - \hat{h}(X)|z]^2] - \frac{\epsilon}{2} \leq \frac{\epsilon}{2} \Rightarrow \text{PRMSE}(\hat{h}) \leq \sqrt{\epsilon}. \quad (31)$$

\square

We note that Theorem 2 follows somewhat directly from the main theorems of Dikkala et al. (2020) but that it was not stated in this precise form in their work.

A.4 PROOF OF LEMMA 1

Proof. Notice that

$$\max_{\pi \in \Pi} \mathbb{E}_{s_{t-1}} [(\mathbb{E}_{s_t \sim \hat{T}(s_{t-1}, \pi_1(s_{t-1}))} [\pi(s_t)] - \mathbb{E}_{s_t \sim P(s_t | s_{t-1})} [\pi(s_t)])^2] \leq \delta \quad (32)$$

can be re-written as

$$\max_{\pi \in \Pi} \mathbb{E}_Z [(\mathbb{E}_{x \sim g(z)} [\pi(x)] - \mathbb{E}_{x \sim P(X|z)} [\pi(x)])^2] \leq \delta. \quad (33)$$

Thus, the proof of Theorem 1 holds as written. \square

A.5 PROOF OF LEMMA 2

An ϵ -approximate equilibrium for the policy player is a π such that

$$\max_{f \in \mathcal{F}} \mathbb{E}[2(a_t - \pi(s_t))f(s_{t-1}) - f^2(s_{t-1})] - \frac{\epsilon}{2} \leq \min_{\pi \in \Pi} \mathbb{E}[2(a_t - h(s_t))\hat{f}(s_{t-1}) - \hat{f}^2(s_{t-1})] + \frac{\epsilon}{2}. \quad (34)$$

With a change of notation, we can re-write this as:

$$\max_{f \in \mathcal{F}} \mathbb{E}[2(Y - \pi(X))f(Z) - f^2(Z)] - \frac{\epsilon}{2} \leq \min_{\pi \in \Pi} \mathbb{E}[2(Y - h(X))\hat{f}(Z) - \hat{f}^2(Z)] + \frac{\epsilon}{2}. \quad (35)$$

Thus, the proof of Theorem 2 holds as written.

A.6 PROOF OF THEOREM 3

Proof. By definition,

$$\text{PRMSE}(\pi) = \sqrt{\mathbb{E}_{s \sim d_{\pi_E}} [\mathbb{E}[a' - \pi(s') | s]]^2} = \epsilon. \quad (36)$$

Recall that the measure of ill-posedness of the problem (Dikkala et al., 2020; Chen & Pouzo, 2012) can be defined as

$$\kappa(\Pi) = \sup_{\pi \in \Pi} \frac{\sqrt{\mathbb{E}_{s \sim d_{\pi_E}} [(\pi_E(s) - \pi(s))^2]}}{\sqrt{\mathbb{E}_{s, s', a' \sim d_{\pi_E}} [\mathbb{E}[a' - \pi(s') | s]]^2}} = \sup_{\pi \in \Pi} \frac{\text{RMSE}(\pi)}{\text{PRMSE}(\pi)} \quad (37)$$

Directly,

$$\text{RMSE}(\pi) \leq \epsilon \kappa(\Pi) \quad (38)$$

We repeat the definition of total variation stability of a distribution $P(U)$:

$$\|a - b\|_2 \leq \delta \Rightarrow d_{TV}(a + U, b + U) \leq c\delta. \quad (39)$$

We proceed by noting that TV-stability implies that $\forall s \in \mathcal{S}$,

$$d_{TV}(\pi(s) + U, \pi_E(s) + U) \leq c \|\pi(s) - \pi_E(s)\| \quad (40)$$

$$\Rightarrow d_{TV}(\pi(s) + U, \pi_E(s) + U)^2 \leq c^2 \|\pi(s) - \pi_E(s)\|^2 \quad (41)$$

$$\Rightarrow \mathbb{E}_{s \sim d_{\pi_E}} [d_{TV}(\pi(s) + U, \pi_E(s) + U)^2] \leq c^2 \mathbb{E}_{s \sim d_{\pi_E}} [\|\pi(s) - \pi_E(s)\|^2] = c^2 \text{MSE}(\pi). \quad (42)$$

By Jensen's inequality,

$$\mathbb{E}_{s \sim d_{\pi_E}} [d_{TV}(\pi(s) + U, \pi_E(s) + U)]^2 \leq \mathbb{E}_{s \sim d_{\pi_E}} [d_{TV}(\pi(s) + U, \pi_E(s) + U)^2] \leq c^2 \text{MSE}(\pi). \quad (43)$$

Taking the square root of both sides, we arrive at

$$\mathbb{E}_{s \sim d_{\pi_E}} [d_{TV}(\pi(s) + U, \pi_E(s) + U)] \leq c \text{RMSE}(\pi) \leq c\kappa(\Pi)\epsilon. \quad (44)$$

Lastly, we apply the Performance Difference Lemma of Kakade & Langford (2002) as follows:

$$J(\pi_E) - J(\pi) = T \mathbb{E}_{s, a \sim d_{\pi_E}} [Q^\pi(s, a) - \mathbb{E}_{a' \sim \pi(s)} [Q^\pi(s, a')]] \quad (45)$$

$$= T \mathbb{E}_{s, a \sim d_{\pi_E}} [Q^\pi(s, \pi_E(s) + u + \tilde{u}_1) - \mathbb{E}[Q^\pi(s, \pi(s) + u + \tilde{u}_2)]] \quad (46)$$

$$\leq T^2 \mathbb{E}_{s \sim d_{\pi_E}} [d_{TV}(\pi(s) + U, \pi_E(s) + U)] \quad (47)$$

$$\leq c\kappa(\Pi)\epsilon T^2. \quad (48)$$

We use the fact that the same u would be added to both the learner and the expert's actions and that rewards are in the range $[-1, 1]$ in the third step. \square

A.7 PROOF OF LEMMA 3

Proof.

$$\mathbb{E}[a_t|do(s_t)] = \mathbb{E}[\pi_E(s_t) + u_t + u_{t-1}|do(s_t)] = \pi_E(s_t) + \mathbb{E}[u_t] + \mathbb{E}[u_{t-1}] = \pi_E(s_t) \quad (49)$$

$$\mathbb{E}[a_t|s_t] = \mathbb{E}[\pi_E(s_t) + u_t + u_{t-1}|s_t] = \pi_E(s_t) + \mathbb{E}[u_t] + \mathbb{E}[u_{t-1}|s_t] = \pi_E(s) + \mathbb{E}[u_{t-1}|s_t] \quad (50)$$

$$\pi_{BC}(s) - \pi_E(s) = \mathbb{E}[a_t|s_t] - \mathbb{E}[a_t|do(s_t)] = \mathbb{E}[u_{t-1}|s_t] = \mathbb{E}[u|s] \quad (51)$$

□

B EXPERIMENT DETAILS

B.1 LUNARLANDER EXPERIMENTS

For ease of simulation, we remove the legs from the LunarLander vehicle (the joints connecting them to the main body have a state that is not recorded in the observed state), remove the dispersion noise, and generate trajectories with a fixed ground layout.

For all learned functions, we use two-layer ReLU MLPs with 64 hidden units. We use the Adam optimizer (Kingma & Ba, 2014) for behavioral cloning and DoubIL and use the optimistic variant for ResiduIL. We apply a weight decay of 1e-3 to all. We train all methods for 50k steps.

PARAMETER	VALUE
LEARNING RATE	3E-4
BATCH SIZE	128

Table 2: Parameters for behavioral cloning.

For computational ease, we only learn the mean of $P(a|s)$ for DoubIL and add fresh standard normal noise on-top of it to simulate drawing actions. For more complex noise models, one would need to use a moment matching algorithm (Swamy et al., 2021) in the first stage.

PARAMETER	VALUE
LEARNING RATE	3E-4
BATCH SIZE	128
NUM. SAMPLES FOR \mathbb{E}	8

Table 3: Parameters for DoubIL.

For implementing the “double samples” for the gradient, we compute $\mathbb{E}_1[a' - \pi(s')|s]$ and $\mathbb{E}_2[a' - \pi(s')|s]$ using independent samples. Then, we apply a stop-gradient operator to the former expectation before taking a product between the expectations and averaging over s :

$$L(\pi) = \mathbb{E}_s[\odot(\mathbb{E}_1[a' - \pi(s')|s])\mathbb{E}_2[a' - \pi(s')|s]]. \quad (52)$$

This loss function has the correct gradient as it uses independent samples for computing the two expectations.

PARAMETER	VALUE
LEARNING RATE	5E-5
BATCH SIZE	128
BC REGULARIZER WEIGHT	5E-2
f NORM PENALTY	1E-3
ADAM β s	0, 1E-2

Table 4: Parameters for ResiduIL.

B.2 LQG EXPERIMENTS

We compute the optimal policy for the following canonical linear system via solving a Discrete-Time Algebraic Ricatti Equation via the standard iterative method:

$$x_t = Ax_{t-1} + Bu_{t-1} \quad (53)$$

$$J(K) = \sum_t^T x_t^T Q x_t + (Kx_t)^T R K x_t \quad (54)$$

$$A = \begin{bmatrix} 1 & \Delta T \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0.5(\Delta T)^2 \\ \Delta T \end{bmatrix}, Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, R = [0.1], \Delta T = 0.1$$

This is the dynamics of a “sliding brick on a frozen lake.” We then simulate rollouts of 200 timesteps with u_t being drawn i.i.d. from the standard normal distribution. We confound actions with the sum of confounders going H steps back:

$$a_t = K^* s_t + \sum_{j=t-H}^t u_j. \quad (55)$$

We simulate 1000 such rollouts to compute [\(21\)](#) empirically. We calculate $\mathbb{E}[X|z] = \mathbb{E}[s_t|s_{t-H}] = (A + BK^*)^H s_{t-H}$ analytically instead of via samples due to the small value of the quantity in comparison to the variance of the noise.