Blessing from Human-AI Interaction: Super Reinforcement Learning in Confounded Environments

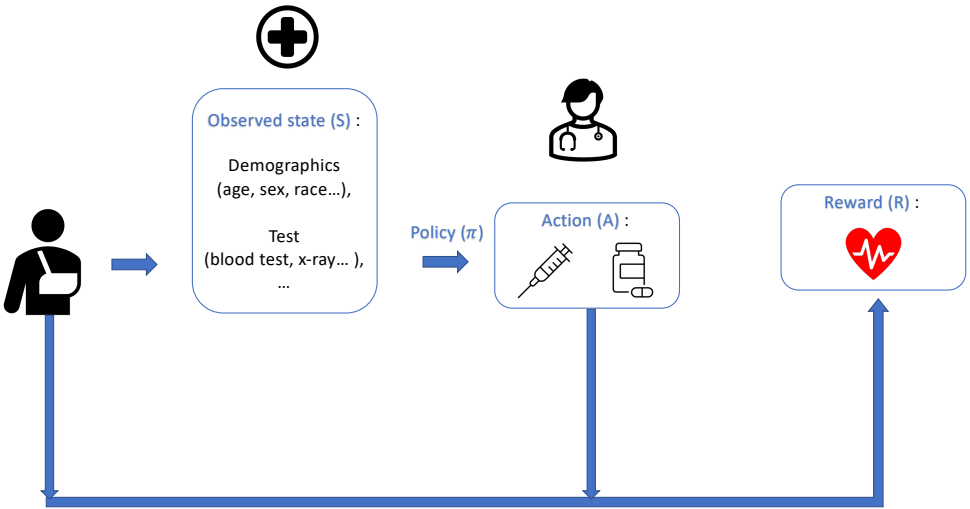
Jiayi Wang

Joint work with Zhengling Qi and Chengchun Shi

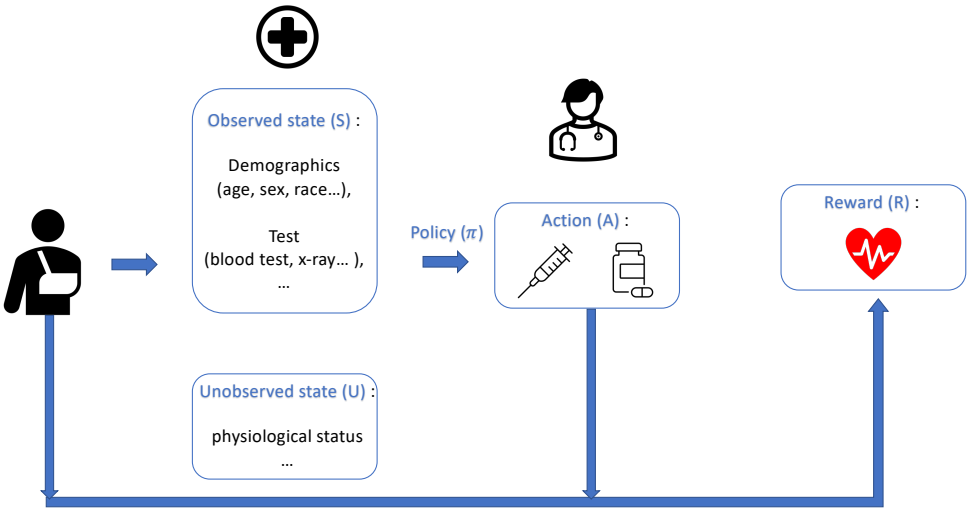
02/03/2023

Confounded Contextual Bandits

- An urgent care example



- An urgent care example



- A random tuple $(S, U, A, \{R(a)\}_{a \in \mathcal{A}})$
  - $\{R(a)\}_{a \in \mathcal{A}}$ denotes a set of the potential/counterfactual rewards under $A = a$.
  - $R = \sum_{a \in \mathcal{A}} R(a) \mathbb{1}(A = a)$.

- A random tuple $(S, U, A, \{R(a)\}_{a \in \mathcal{A}})$
  - $\{R(a)\}_{a \in \mathcal{A}}$ denotes a set of the potential/counterfactual rewards under $A = a$.
  - $R = \sum_{a \in \mathcal{A}} R(a) \mathbb{1}(A = a)$.
- Offline dataset
  - i.i.d copies of (S, A, R): $\{S_i, A_i, R_i\}_{i=1}^n$.
  - A is generated by some behavior policy $\pi^b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(A)$ that depends on both observed and unobserved features.

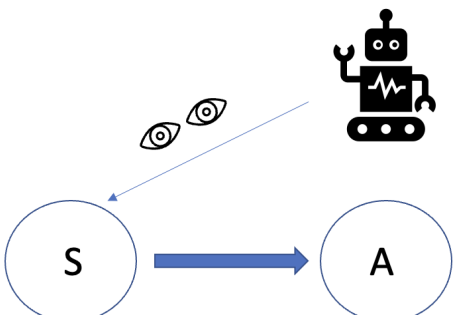
- A random tuple $(S, U, A, \{R(a)\}_{a\in\mathcal{A}})$
  - $\{R(a)\}_{a\in\mathcal{A}}$ denotes a set of the potential/counterfactual rewards under $A = a$.
  - $R = \sum_{a\in\mathcal{A}} R(a)\mathbb{1}(A = a)$.
- Offline dataset
  - i.i.d copies of (S, A, R): $\{S_i, A_i, R_i\}_{i=1}^n$.
  - A is generated by some behavior policy $\pi^b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(A)$ that depends on both observed and unobserved features.
- Target: find the optimal policy π^* such that the value function

$$\mathbb{E}\left\{\sum_{a\in\mathcal{A}} R(a)\pi^*(a \mid S, U)\right\}$$

is maximized, using the offline dataset.

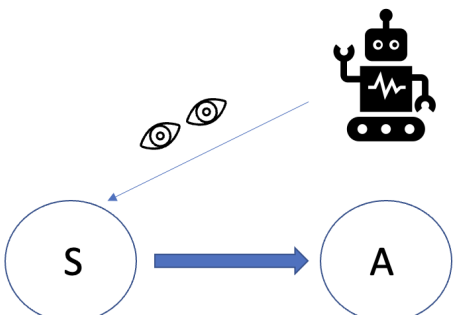
- Since U is unobserved, most existing solutions focus on finding an optimal policy $\pi^* \in \Pi = \{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$ given by

$$\pi^*(a^*|s) = 1 \text{ if } a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[R(a)|S = s], \forall s \in S.$$



- Since U is unobserved, most existing solutions focus on finding an optimal policy $\pi^* \in \Pi = \{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$ given by

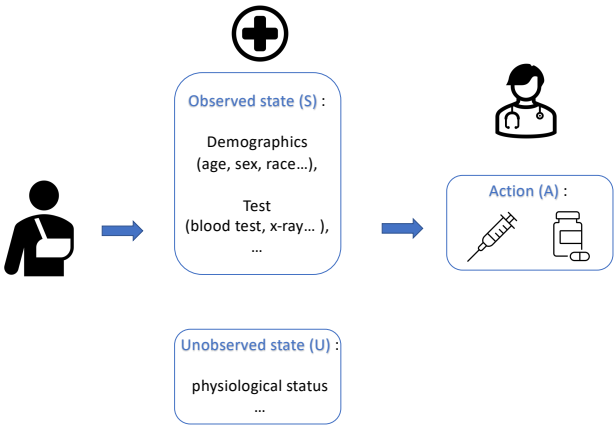
$$\pi^*(a^*|s) = 1 \text{ if } a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[R(a)|S = s], \forall s \in S.$$



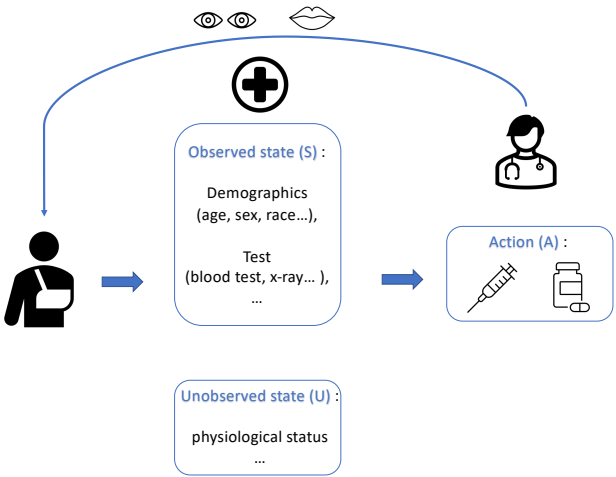
- Can we improve upon this?

- Leverages the input of human expertise, since actions generated by the behavior policy depend on the latent information ($\pi_b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(\mathcal{A})$).

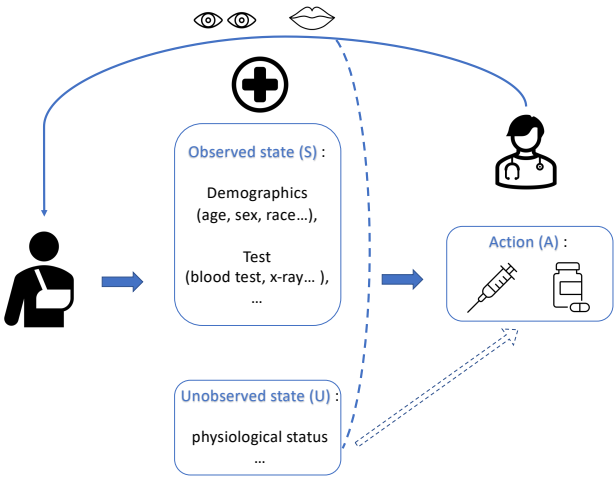
- Leverages the input of human expertise, since actions generated by the behavior policy depend on the latent information ($\pi_b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(\mathcal{A})$).
- The urgent care example



- Leverages the input of human expertise, since actions generated by the behavior policy depend on the latent information ($\pi_b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(\mathcal{A})$).
- The urgent care example

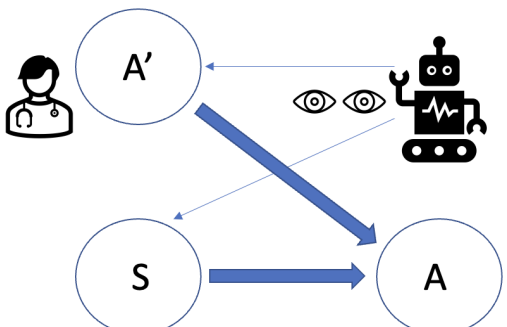


- Leverages the input of human expertise, since actions generated by the behavior policy depend on the latent information ($\pi_b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(\mathcal{A})$).
- The urgent care example



- Aim to find a *super-policy* ν^* in a larger policy class $\Omega = \{\nu : S \times A \to \mathcal{P}(A)\}$ such that

$$\nu^*(a^*|s, a') = 1 \text{ if } a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}\{R(a)|S = s, A = a'\}, \forall (s, a') \in S \times A.$$



- $S \sim Bernoulli(\mathrm{p} = 0.5)$ and $U \sim Bernoulli(\mathrm{p} = 0.5)$ are independent.
- Action is binary and the behavior policy satisfies $P(A = 1|S, U = 1) = P(A = 0|S, U = 0) = 1 - \epsilon$ for some $0 \leq \epsilon \leq 1$.
- $R = 8(A - 0.5)(S - 0.2)(U - 0.3)$.

- $S \sim Bernoulli(\text{p} = 0.5)$ and $U \sim Bernoulli(\text{p} = 0.5)$ are independent.
- Action is binary and the behavior policy satisfies $P(A = 1|S, U = 1) = P(A = 0|S, U = 0) = 1 - \epsilon$ for some $0 \leq \epsilon \leq 1$.
- $R = 8(A - 0.5)(S - 0.2)(U - 0.3)$.

Policy Value	$\mathcal{V}(\pi_b)$	$\mathcal{V}(\pi^*)$	$\mathcal{V}(\nu^*)$
$\epsilon = 0.5$	0.0	**0.4**	**0.4**
$\epsilon = 0$	0.6	0.4	**1.0**
$\epsilon = 1$	-0.6	0.4	**1.0**

- $S \sim Bernoulli(\mathrm{p} = 0.5)$ and $U \sim Bernoulli(\mathrm{p} = 0.5)$ are independent.
- Action is binary and the behavior policy satisfies $P(A = 1|S, U = 1) = P(A = 0|S, U = 0) = 1 - \epsilon$ for some $0 \leq \epsilon \leq 1$.
- $R = 8(A - 0.5)(S - 0.2)(U - 0.3)$.

Policy Value	$\mathcal{V}(\pi_b)$	$\mathcal{V}(\pi^*)$	$\mathcal{V}(\nu^*)$
$\epsilon = 0.5$	0.0	**0.4**	**0.4**
$\epsilon = 0$	0.6	0.4	**1.0**
$\epsilon = 1$	-0.6	0.4	**1.0**

Theorem (Super-Optimality)

$\mathcal{V}(\nu^*) \geq \max\{\mathcal{V}(\pi^*), \mathcal{V}(\pi_b)\}$.

Identification

- Adopt the proximal causal inference framework developed by Tchetgen et al. [2020].

  Specifically, we assume the existence of certain action and reward proxies Z and W in additional to (S, A, R).

- Adopt the proximal causal inference framework developed by Tchetgen et al. [2020].

  Specifically, we assume the existence of certain action and reward proxies Z and W in additional to (S, A, R).

Assumption

(a) $R \perp\!\!\!\perp Z \mid (S, U, A)$;

(b) $W \perp\!\!\!\perp (Z, A) \mid (S, U)$, $W \not\perp\!\!\!\perp U \mid S$;

(c) $R(a) \perp\!\!\!\perp A \mid (S, U)$ *for* $a \in \mathcal{A}$;

(d) *There exists a bridge function* $q : \mathcal{W} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ *such that*

$$\mathbb{E}\left[q(W, a, S) \mid U, S, A = a\right] = \mathbb{E}\left[R \mid U, S, A = a\right]. \tag{1}$$

- Adopt the proximal causal inference framework developed by Tchetgen et al. [2020].

  Specifically, we assume the existence of certain action and reward proxies Z and W in additional to (S, A, R).

Assumption

(a) $R \perp\!\!\!\perp Z \mid (S, U, A)$;

(b) $W \perp\!\!\!\perp (Z, A) \mid (S, U)$, $W \not\perp\!\!\!\perp U \mid S$;

(c) $R(a) \perp\!\!\!\perp A \mid (S, U)$ *for* $a \in \mathcal{A}$;

(d) *There exists a bridge function* $q : \mathcal{W} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ *such that*

$$\mathbb{E}\left[q(W, a, S) \mid U, S, A = a\right] = \mathbb{E}\left[R \mid U, S, A = a\right]. \tag{1}$$

- $\mathcal{V}(\nu) = \mathbb{E}\left[\sum_{a \in \mathcal{A}} q(W, a, S)\nu(a \mid S, A)\right]$ for any $\nu \in \Omega$.

- In practice, one may want to include as many confounders in the policy as possible to achieve the largest super-optimality.

  We further extend the policy class to $\Omega = \{\nu : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \mathcal{P}(\mathcal{A})\}$ and consider the corresponding super-policy ν^*.

- In practice, one may want to include as many confounders in the policy as possible to achieve the largest super-optimality.

  We further extend the policy class to $\Omega = \{\nu : \mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \mathcal{P}(\mathcal{A})\}$ and consider the corresponding super-policy ν^*.

Theorem

Under above Assumptions, if we further suppose some completeness and regularity conditions, solving the following linear integral equation

$$\mathbb{E}\left[q(W, a, S) \mid Z, S, A = a\right] = \mathbb{E}\left[R \mid Z, S, A = a\right], \tag{2}$$

for every $a \in \mathcal{A}$ with respect to q gives a valid bridge function that satisfies (1)*. And the optimal policy ν^* in class Ω is given by*

$$\nu^*(a^* \mid s, z, a') = 1 \quad \text{if} \quad a^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[q(W, a, S) \mid S = s, Z = z, A = a'\right]. \tag{3}$$

Algorithm 1: Learning Algorithm for the contextual bandits under unmeasured confounding

1 **Input:** Data $\mathcal{D} = (S_i, Z_i, A_i, R_i, W_i)_{i=1}^n$.

2 Obtain the estimation of the bridge function $\hat{q}$ by solving the estimation equation (2) using data $\mathcal{D}$

3 Implement any supervised learning method for estimating $\mathbb{E}[\hat{q}(W, S, a) \mid S, Z, A]$.

4 Compute

$$a^* = \arg\max_{a \in \mathcal{A}} \hat{\mathbb{E}}[\hat{q}(W, S, a) \mid S = s, Z = z, A = a'] \quad \forall (s, z, a') \in \mathcal{S} \times \mathcal{Z} \times \mathcal{A}.$$

5 **Output:** $\hat{\nu}^*$ with $\hat{\nu}(a^* \mid s, z, a') = 1$ and $\hat{\nu}(\tilde{a} \mid s, z, a') = 0$ for $\tilde{a} \neq a^*$.

Algorithm 2: Learning Algorithm for the contextual bandits under unmeasured confounding

1 **Input:** Data $\mathcal{D} = (S_i, Z_i, A_i, R_i, W_i)_{i=1}^n$.
2 Obtain the estimation of the bridge function $\hat{q}$ by solving the estimation equation (2) using data $\mathcal{D}$
3 Implement any supervised learning method for estimating $\mathbb{E}[\hat{q}(W, S, a) \mid S, Z, A]$.
4 Compute

$$a^* = \arg\max_{a \in \mathcal{A}} \hat{\mathbb{E}}[\hat{q}(W, S, a) \mid S = s, Z = z, A = a'] \quad \forall (s, z, a') \in \mathcal{S} \times \mathcal{Z} \times \mathcal{A}.$$

5 **Output:** $\hat{\nu}^*$ with $\hat{\nu}(a^* \mid s, z, a') = 1$ and $\hat{\nu}(\tilde{a} \mid s, z, a') = 0$ for $\tilde{a} \neq a^*$.

- Step 2: minimax estmation of conditional moment models.

$$q = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \sup_{g \in \mathcal{G}} \mathbb{E}\left\{ [(q(W, A, S) - R)]\, g(S, Z, A) - \frac{1}{2} g^2(S, Z, A) \right\}$$

Algorithm 3: Learning Algorithm for the contextual bandits under unmeasured confounding

1 **Input:** Data $\mathcal{D} = (S_i, Z_i, A_i, R_i, W_i)_{i=1}^n$.
2 Obtain the estimation of the bridge function $\hat{q}$ by solving the estimation equation (2) using data $\mathcal{D}$
3 Implement any supervised learning method for estimating $\mathbb{E}[\hat{q}(W, S, a) \mid S, Z, A]$.
4 Compute
$a^* = \arg\max_{a \in \mathcal{A}} \hat{\mathbb{E}}[\hat{q}(W, S, a) \mid S = s, Z = z, A = a'] \quad \forall (s, z, a') \in \mathcal{S} \times \mathcal{Z} \times \mathcal{A}.$
5 **Output:** $\hat{\nu}^*$ with $\hat{\nu}(a^* \mid s, z, a') = 1$ and $\hat{\nu}(\tilde{a} \mid s, z, a') = 0$ for $\tilde{a} \neq a^*$.

- Step 3: take $\hat{\mathbb{E}}[\hat{q}(W, S, a) \mid S = \cdot, Z = \cdot, A = \cdot]$ as the solution of

$$\underset{g \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [g(S_i, Z_i, A_i) - \hat{q}(\cdot, \cdot, a)]^2 + \mu \|g\|_{\mathcal{G}}^2,$$

Theorem

Suppose q belongs to certain function class $\mathcal{Q} \subset \mathcal{W} \times \mathcal{S} \times \mathcal{A}$. Define the projection error as $\xi_n := \sup_{q \in \mathcal{Q}, a \in \mathcal{A}} \|g[\cdot, \cdot, \cdot\ ;\ q(\cdot, \cdot, a)] - \hat{g}[\cdot, \cdot, \cdot\ ;\ q(\cdot, \cdot, a)]\|_2$, and the bridge function estimation error as $\zeta_n := \|q - \hat{q}\|_2$. Then we obtain the following regret decomposition

$$\mathcal{V}(\nu^*) - \mathcal{V}(\hat{\nu}^*) \leq 2(\xi_n + p_{\max}\zeta_n),$$

where $p_{\max}$ is some overlap constant.

Theorem

Suppose q belongs to certain function class $\mathcal{Q} \subset \mathcal{W} \times \mathcal{S} \times \mathcal{A}$. Define the projection error as $\xi_n := \sup_{q \in \mathcal{Q}, a \in \mathcal{A}} \|g[\cdot, \cdot, \cdot\ ;\ q(\cdot, \cdot, a)] - \hat{g}[\cdot, \cdot, \cdot\ ;\ q(\cdot, \cdot, a)]\|_2$, and the bridge function estimation error as $\zeta_n := \|q - \hat{q}\|_2$. Then we obtain the following regret decomposition

$$\mathcal{V}(\nu^*) - \mathcal{V}(\hat{\nu}^*) \leq 2(\xi_n + p_{\max}\zeta_n),$$

where $p_{\max}$ is some overlap constant.

Corollary

If the star-shaped spaces $\mathcal{G}$ and $\mathcal{Q}$ are VC-subgraph classes with VC dimensions $\mathbb{V}(\mathcal{G})$, and $\mathbb{V}(\mathcal{Q})$ respectively. Under certain technical assumptions, with probability at least $1 - \delta$,

$$\mathcal{V}(\hat{\nu}^*) - \mathcal{V}(\nu^*) \lesssim n^{-1/2} p_{\max} \sqrt{\log(1/\delta) + \max\{\mathbb{V}(\mathcal{G}), \mathbb{V}(\mathcal{Q})\}}.$$

Application

- A dataset from Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments [SUPPORT Connors et al., 1996].
  SUPPORT examined the effectiveness and safety of direct measurement of cardiac function by Right Heart Catheterization (RHC) for certain critically ill patients in intensive care units (ICU).
- Action: $A = 1 \rightarrow$ measured by RHC in the first 24 hours, otherwise $A = 0$.
  Response: $Y = 1 \rightarrow$ survived or censored at day 30; otherwise $Y = -1$.

- Covariates: demographics, estimated probability of survival, comorbidity, vital signs, physiological status, and functional status.

- Covariates: demographics, estimated probability of survival, comorbidity, vital signs, physiological status, and functional status.
- Ten variables measuring the patient's overall physiological status:
  - subject to substantial measurement error
  - single snapshot of underlying physiological state over time.

- Covariates: demographics, estimated probability of survival, comorbidity, vital signs, physiological status, and functional status.
- Ten variables measuring the patient's overall physiological status:
  - subject to substantial measurement error
  - single snapshot of underlying physiological state over time.
- Following Tchetgen et al. [2020]: $W = (ph1, hema1)$, $Z = (pafi1, paco21)$.

- Three types of policy classes are considered.
  1. SONLY: $\mathcal{S} \to \mathcal{P}(\mathcal{A})$. The policy only depends on the observed state S.
  2. SZONLY: $\mathcal{S} \times \mathcal{Z} \to \mathcal{P}(\mathcal{A})$. The policy depends on on the observed state S and the action proxy Z.
  3. SUPER: $\mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \mathcal{P}(\mathcal{A})$. The super-policy class where the policy depends on the observed state S, the action proxy Z, and observed action A.

- Three types of policy classes are considered.
  1. SONLY: $\mathcal{S} \to \mathcal{P}(\mathcal{A})$. The policy only depends on the observed state S.
  2. SZONLY: $\mathcal{S} \times \mathcal{Z} \to \mathcal{P}(\mathcal{A})$. The policy depends on on the observed state S and the action proxy Z.
  3. SUPER: $\mathcal{S} \times \mathcal{Z} \times \mathcal{A} \to \mathcal{P}(\mathcal{A})$. The super-policy class where the policy depends on the observed state S, the action proxy Z, and observed action A.

Table: Evaluation results of the optimal policies learned from three different policy classes using the RHC data. The averages of evaluation values over 20 random splits are presented. Larger values indicate better performances. Values in the parentheses are standard deviations.

Sonly	**SZonly**	**Super**
0.55 (5.80e-02)	0.55 (5.78e-02)	**0.69** (1.10e-02)

Confounded Sequantial Decision Making

- Confounded POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{U}, \mathcal{A}, T, \mathcal{P}, r)$.
  - $\mathcal{P} = \{\mathcal{P}_t\}_{t=1}^{T}$, $\mathcal{P}_t$ denotes transition kernel from $\mathcal{S} \times \mathcal{U} \times \mathcal{A}$ to $\mathcal{S} \times \mathcal{U}$ at time t

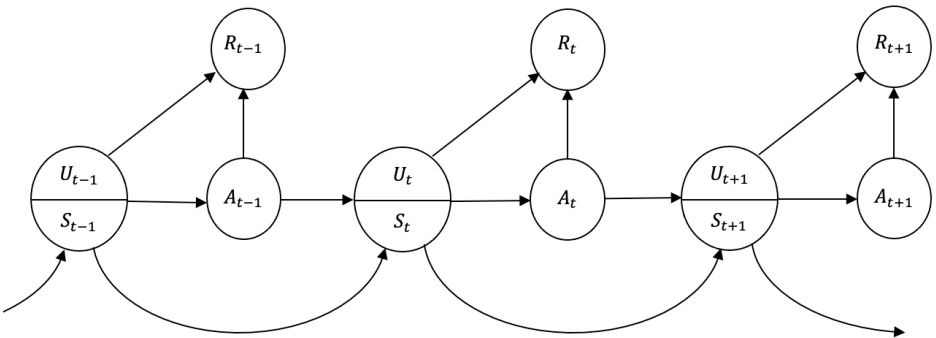
Assumption

(Markovianity) The process $\{S_t, U_t, A_t, R_t\}_{t=1}^{T}$ satisfies the Markov property, i.e., for any t, (R_t, S_{t+1}, U_{t+1}) depends on the past history only through (S_t, U_t, A_t).

- Confounded POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{U}, \mathcal{A}, T, \mathcal{P}, r)$.
  - $\mathcal{P} = \{\mathcal{P}_t\}_{t=1}^{T}$, $\mathcal{P}_t$ denotes transition kernel from $\mathcal{S} \times \mathcal{U} \times \mathcal{A}$ to $\mathcal{S} \times \mathcal{U}$ at time t

Assumption

(Markovianity) The process $\{S_t, U_t, A_t, R_t\}_{t=1}^{T}$ satisfies the Markov property, i.e., for any t, (R_t, S_{t+1}, U_{t+1}) depends on the past history only through (S_t, U_t, A_t).



- Offline dataset:
  - i.i.d. episodes $\{S_{i,t}, A_{i,t}, R_{i,t}\}_{t=1}^{T}$, $i = 1, \ldots, n$.
  - A_t is generated by some behavior policy $\pi_t^b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(A)$.

- Offline dataset:
  - i.i.d. episodes $\{S_{i,t}, A_{i,t}, R_{i,t}\}_{t=1}^{T}$, $i = 1, \ldots, n$.
  - A_t is generated by some behavior policy $\pi_t^b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(A)$.
- Value function given a generic policy $\{\pi_t\}_{t=1}^{T}$:

$$V_t^{\pi}(s, u) = \mathbb{E}^{\pi}[\sum_{t'=t}^{T} R_{t'} \mid S_t = s, U_t = u]$$

- Offline dataset:
  - i.i.d. episodes $\{S_{i,t}, A_{i,t}, R_{i,t}\}_{t=1}^{T}$, $i = 1, \ldots, n$.
  - A_t is generated by some behavior policy $\pi_t^b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(A)$.
- Value function given a generic policy $\{\pi_t\}_{t=1}^{T}$:

$$V_t^{\pi}(s, u) = \mathbb{E}^{\pi}[\sum_{t'=t}^{T} R_{t'} \mid S_t = s, U_t = u]$$

- Target: Estimate an optimal policy that maximizes

$$\mathcal{V}(\pi) = \mathbb{E}[V_1^{\pi}(S_1, U_1)]$$

using the batch data.

- Offline dataset:
  - i.i.d. episodes $\{S_{i,t}, A_{i,t}, R_{i,t}\}_{t=1}^{T}$, $i = 1, \ldots, n$.
  - A_t is generated by some behavior policy $\pi_t^b : \mathcal{S} \times \mathcal{U} \to \mathcal{P}(A)$.
- Value function given a generic policy $\{\pi_t\}_{t=1}^{T}$:

$$V_t^{\pi}(s, u) = \mathbb{E}^{\pi}[\sum_{t'=t}^{T} R_{t'} \mid S_t = s, U_t = u]$$

- Target: Estimate an optimal policy that maximizes

$$\mathcal{V}(\pi) = \mathbb{E}[V_1^{\pi}(S_1, U_1)]$$

  using the batch data.
- Common optimal policy $\pi^* \in \Pi \equiv \{\pi = \{\pi_t\}_{t=1}^{T} \mid \pi_t : \mathcal{S} \times \mathcal{Z}_t \to \mathbb{P}(\mathcal{A})\}$

  v.s.

  Super policy $\nu^* \in \Omega \equiv \{\nu = \{\nu_t\}_{t=1}^{T} \mid \nu_t : \mathcal{S} \times \mathcal{Z}_t \times \mathcal{A} \to \mathcal{P}(\mathcal{A})\}$

Identification

- We assume the existence of certain action and reward proxies $\{Z_t\}_{t=1}^T$ and $\{W_t\}_{t=1}^T$ that can help identify policy values.

- We assume the existence of certain action and reward proxies $\{Z_t\}_{t=1}^T$ and $\{W_t\}_{t=1}^T$ that can help identify policy values.

Assumption

(a) *(Reward proxy)* $W_t \perp\!\!\!\perp (A_t, U_{t-1}, S_{t-1}) \mid (U_t, S_t)$, $W_t \not\perp\!\!\!\perp U_t \mid S_t$, *for* $1 \leq t \leq T$.

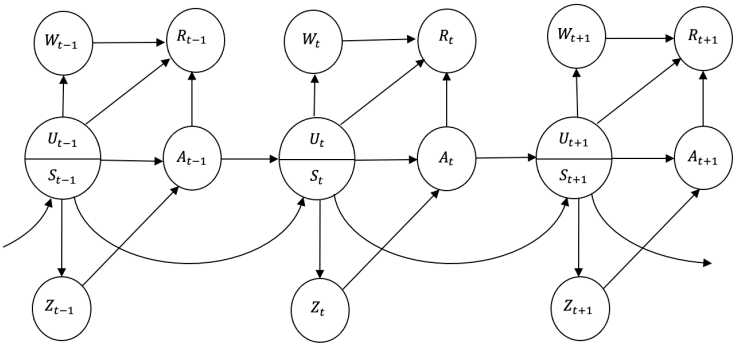
(b) *(Action proxy)* $Z_t \perp\!\!\!\perp (R_t, W_t, S_{t+1}, U_{t+1}, W_{t+1}) \mid (U_t, S_t, A_t)$ *for* $1 \leq t \leq T$.

- We assume the existence of certain action and reward proxies $\{Z_t\}_{t=1}^T$ and $\{W_t\}_{t=1}^T$ that can help identify policy values.

Assumption

(a) *(Reward proxy)* $W_t \perp\!\!\!\perp (A_t, U_{t-1}, S_{t-1}) \mid (U_t, S_t)$, $W_t \not\perp\!\!\!\perp U_t \mid S_t$, *for* $1 \le t \le T$.

(b) *(Action proxy)* $Z_t \perp\!\!\!\perp (R_t, W_t, S_{t+1}, U_{t+1}, W_{t+1}) \mid (U_t, S_t, A_t)$ *for* $1 \le t \le T$.



Theorem

Under appropriate assumptions, there always exist Q-bridge functions $\{q^\nu\}_{t=1}^T$ *satisfying*

$$\mathbb{E}^\nu\left[\sum_{t'=t}^T R_{t'} \mid U_t, S_t, A_t\right] = \mathbb{E}\left[\sum_{a\in\mathcal{A}} q_t^\nu(W_t, S_t, a)\nu_t(a \mid S_t, Z_t, A_t) \mid U_t, S_t, A_t\right], \tag{4}$$

for $t = 1, \ldots, T$. *In particular, set* $q_{T+1}^\nu = 0$, q_t^ν *can be obtained by solving the following linear integral equations for* $t = T, \ldots, 1$,

$$\mathbb{E}\{q_t^\nu(W_t, S_t, A_t) - R_t - V_{t+1}^\nu(W_{t+1}, S_{t+1}, Z_{t+1}, A_{t+1}) \mid Z_t, S_t, A_t\} = 0, \tag{5}$$

where $V_t^\nu(W_t, S_t, Z_t, A_t) = \sum_{a\in\mathcal{A}} q_t^\nu(W_t, S_t, a)\nu_t(a \mid S_t, Z_t, A_t)$.

Theorem

Under appropriate assumptions, there always exist Q-bridge functions $\{q^\nu\}_{t=1}^T$ *satisfying*

$$\mathbb{E}^\nu \left[\sum_{t'=t}^T R_{t'} \mid U_t, S_t, A_t \right] = \mathbb{E} \left[\sum_{a \in \mathcal{A}} q_t^\nu(W_t, S_t, a) \nu_t(a \mid S_t, Z_t, A_t) \mid U_t, S_t, A_t \right], \tag{4}$$

for $t = 1, \ldots, T$. *In particular, set* $q_{T+1}^\nu = 0$, q_t^ν *can be obtained by solving the following linear integral equations for* $t = T, \ldots, 1$,

$$\mathbb{E}\{q_t^\nu(W_t, S_t, A_t) - R_t - V_{t+1}^\nu(W_{t+1}, S_{t+1}, Z_{t+1}, A_{t+1}) \mid Z_t, S_t, A_t\} = 0, \tag{5}$$

where $V_t^\nu(W_t, S_t, Z_t, A_t) = \sum_{a \in \mathcal{A}} q_t^\nu(W_t, S_t, a) \nu_t(a \mid S_t, Z_t, A_t)$.

$$\mathcal{V}(\nu) = \mathbb{E} \left[\sum_{a \in \mathcal{A}} q_1^\nu(W_1, S_1, a) \nu_1(a \mid S_1, Z_1, A_1) \right], \quad \forall \nu \in \Omega.$$

Theorem

Under appropriate assumptions, there always exist Q-bridge functions $\{q^\nu\}_{t=1}^T$ *satisfying*

$$\mathbb{E}^\nu \left[\sum_{t'=t}^T R_{t'} \mid U_t, S_t, A_t \right] = \mathbb{E} \left[\sum_{a \in \mathcal{A}} q_t^\nu(W_t, S_t, a) \nu_t(a \mid S_t, Z_t, A_t) \mid U_t, S_t, A_t \right], \tag{4}$$

for $t = 1, \ldots, T$. *In particular, set* $q_{T+1}^\nu = 0$, q_t^ν *can be obtained by solving the following linear integral equations for* $t = T, \ldots, 1$,

$$\mathbb{E}\{q_t^\nu(W_t, S_t, A_t) - R_t - V_{t+1}^\nu(W_{t+1}, S_{t+1}, Z_{t+1}, A_{t+1}) \mid Z_t, S_t, A_t\} = 0, \tag{5}$$

where $V_t^\nu(W_t, S_t, Z_t, A_t) = \sum_{a \in \mathcal{A}} q_t^\nu(W_t, S_t, a) \nu_t(a \mid S_t, Z_t, A_t)$.

$$\nu^* = \arg\max_{\nu \in \Omega} \mathbb{E} \left[\sum_{a \in \mathcal{A}} q_1^\nu(W_1, S_1, a) \nu_1(a \mid S_1, Z_1, A_1) \right].$$

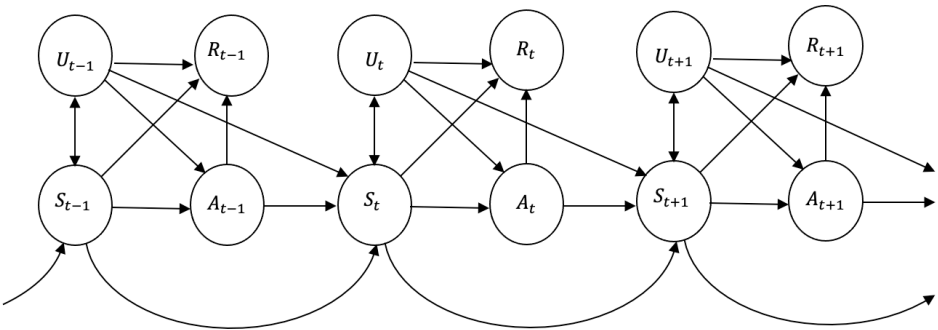
A Practical Algorithm

Assumption (Memoryless Unmeasured Confounding)

For $2 \leq t \leq T$, U_t is independent of past data history (including latent factors in the past) up to time $t-1$ given S_t.

Assumption (Memoryless Unmeasured Confounding)

For $2 \leq t \leq T$, U_t is independent of past data history (including latent factors in the past) up to time $t-1$ given S_t.



Algorithm 4: Super RL for the confounded POMDP

Input: Data $\mathcal{D} = \{\mathcal{D}_t\}_{t=1}^T$ with
$\mathcal{D}_t = \{(S_{i,t}, Z_{i,t}, A_{i,t}, R_{i,t}, W_{i,t}, S_{i,t+1}, Z_{i,t+1}, W_{i,t+1})\}_{i=1}^n$.

Let $\hat{q}_{T+1} = 0$ and $\hat{\nu}_T^*$ be an arbitrary policy.

Repeat for $t = T, \ldots, 1$:

  Obtain an estimator $\hat{q}_t$ for q_t by solving (5) using data $\mathcal{D}_t$ and $\hat{q}_{t+1}$ obtained from the last iteration.

  Compute $\hat{\mathbb{E}}[\hat{q}_t(W_t, S_t, a) \mid S_t = s, Z_t = z, A_t = a']$ for $a \in \mathcal{A}$ using any supervised learning method and obtain the estimated super policy $\hat{\nu}_t^*$ as for every (a', z, s),

$$\hat{\nu}_t^*(a^* \mid s, z, a') = \mathbb{1}\left\{\arg\max_{a \in \mathcal{A}} \hat{\mathbb{E}}[\hat{q}_t(W_t, S_t, a) \mid S_t = s, Z_t = z, A_t = a']\right\}.$$

Output: $\hat{\nu}^* = \{\hat{\nu}_t^*\}_{t=1}^T$.

Theorem

Suppose $q_t \in \mathcal{Q}^{(t)}$ for $1 \leq t \leq T$ and $\hat{\nu}^*$ is computed via Algorithm 4. Then under above Assumptions and appropriate technical conditions, we obtain the following regret decomposition,

$$\mathcal{V}(\nu^*) - \mathcal{V}(\hat{\nu}^*) \lesssim \left(\sum_{t=1}^{T} 2 p_{t,\max} \xi_{t,n} \right) + \sqrt{T \sum_{t=1}^{T} (p^{\omega}_{t,\max})^2 (\zeta_{t,n})^2}.$$

Theorem

Suppose $q_t \in \mathcal{Q}^{(t)}$ for $1 \le t \le T$ and $\hat{\nu}^*$ is computed via Algorithm 4. Then under above Assumptions and appropriate technical conditions, we obtain the following regret decomposition,

$$\mathcal{V}(\nu^*) - \mathcal{V}(\hat{\nu}^*) \lesssim \left(\sum_{t=1}^{T} 2 p_{t,\max} \xi_{t,n} \right) + \sqrt{T \sum_{t=1}^{T} (p^{\omega}_{t,\max})^2 (\zeta_{t,n})^2}.$$

- $p_{t,\max}$, $p^{\omega}_{t,\max}$: overlap constants.
- $\zeta_{t,n}$: Q-bridge function estimation error.
- $\xi_{t,n}$: projection estimation error.

Simulation and Application

- Existing setting from Miao et al. [2022].
- $n = 1000$ and $T = 20$.

Table: Simulation results for the sequential decision making problem. The simulation is performed over 50 simulated datasets. Mean regret values for estimated optimal policies under different policy classes are provided. The smaller regret values indicate better performances. Values in the parentheses are the standard deviations of the regret values.

Sonly	**SZonly**	**Super**
5.4 (1.9e-01)	5.3 (4.7e-01)	**2.2** (4.9e-01)

- Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC-III) dataset (https://physionet.org/content/mimiciii/1.4/) records the longitudinal information (including information of demographics, vitals, labs and scores) of patients who satisfied the sepsis criteria, and the goal is to learn an optimal personalized treatment strategy for sepsis.

- Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC-III) dataset (https://physionet.org/content/mimiciii/1.4/) records the longitudinal information (including information of demographics, vitals, labs and scores) of patients who satisfied the sepsis criteria, and the goal is to learn an optimal personalized treatment strategy for sepsis.
- Despite the richness of data collected at the ICU, the mapping between true patient states and clinical observations is usually ambiguous [Nanayakkara et al., 2022], and therefore makes this dataset fit into the setting of a confounded POMDP.

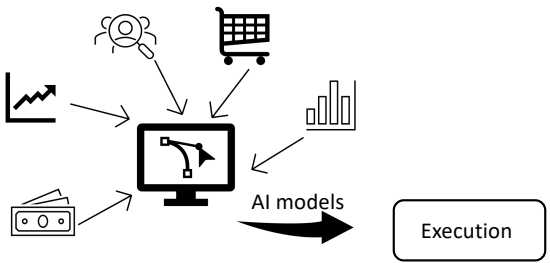
- Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC-III) dataset (https://physionet.org/content/mimiciii/1.4/) records the longitudinal information (including information of demographics, vitals, labs and scores) of patients who satisfied the sepsis criteria, and the goal is to learn an optimal personalized treatment strategy for sepsis.
- Despite the richness of data collected at the ICU, the mapping between true patient states and clinical observations is usually ambiguous [Nanayakkara et al., 2022], and therefore makes this dataset fit into the setting of a confounded POMDP.
- Simplify the action space into 4-dimensional. Fix the horizon at $T = 2$.

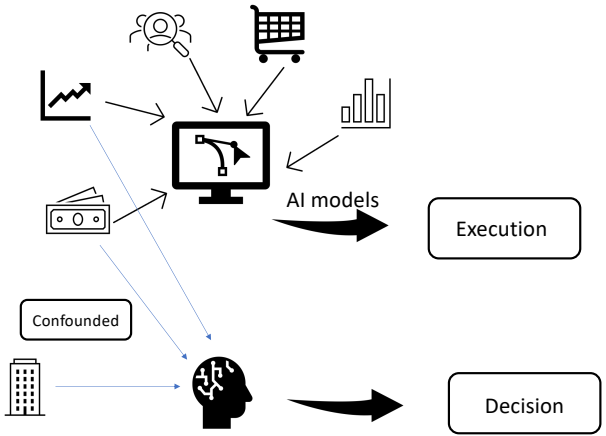
Table: Evaluation results of the optimal policies learned from three different policy classes using the MIMIC-III data. The averages of evaluation values over 20 random splits are presented. Larger values indicate better performances. Values in the parentheses are standard deviations.

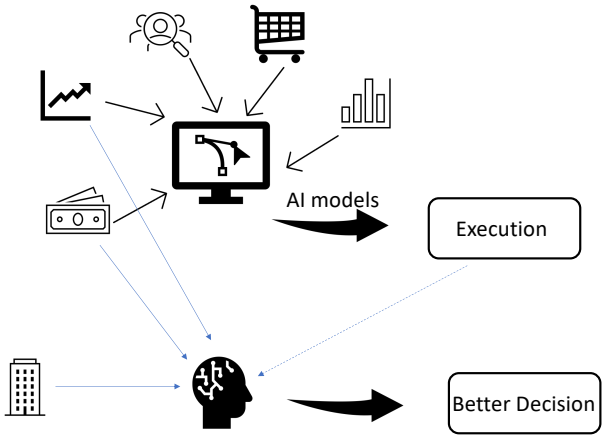
Sonly	**SZonly**	**Super**
-2.83 (5.30e-02)	-2.81 (5.03e-02)	**-1.75** (1.14e-02)

Conclusion

- We introduce super reinforcement learning, which takes the observed action in the offline data as input to enhanced policy learning under endogeneity.
- We establish the identification results for the super-policy in various confounded environments.
- Practical algorithms are provided to perform the super-policy learning with finite-sample regret guarantees.
- Our super policy can be used in "human-in-the-loop" and "machine-in-the-loop".


AI models
Execution


AI models
Execution
Confounded
Decision


AI models
Execution
Better Decision

Thank You!

Alfred F Connors, Theodore Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, et al. The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897, 1996.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

Rui Miao, Zhengling Qi, and Xiaoke Zhang. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models, 2022. URL `https://arxiv.org/abs/2209.10064`.

Thesath Nanayakkara, Gilles Clermont, Christopher James Langmead, and David Swigon. Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment. *PLOS Digital Health*, 1(2):e0000012, 2022.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

Jiayi Wang, Zhengling Qi, and Chengchun Shi. Blessing from experts: Super reinforcement learning in confounded environments. *arXiv preprint arXiv:2209.15448*, 2022.