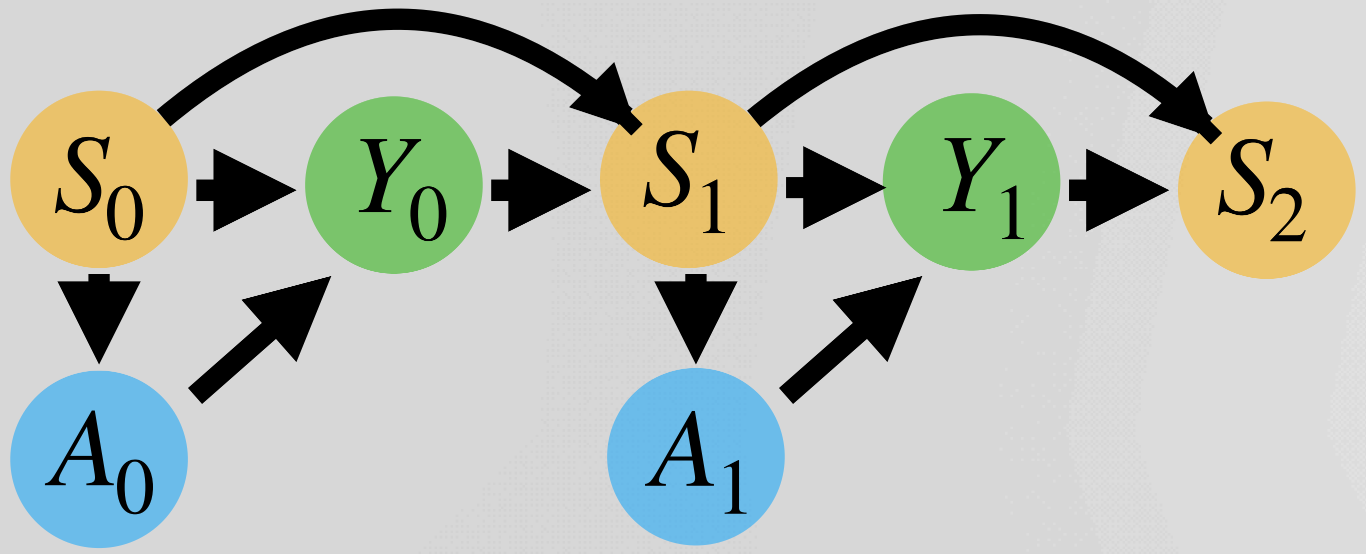
Double Reinforcement Learning in Semiparametric Markov Decision Processes

Lars van der Laan
Joint with Aurelien Bibaut, David Hubbard, Allen Tran, and Nathan Kallus
ACIC 2025

"Automatic double reinforcement learning in semiparametric markov decision processes
with applications to long-term causal inference." arXiv preprint arXiv:2501.06926 (2025).

Motivation

sequential causal inference



- **Many real-world decisions are made sequentially over time**
  - Daily movie recommendations
  - Treatment dosage by visit
- **Questions in sequential causal inference:**
  - What is the optimal treatment or action to take at each time?
  - What is the long-term causal effect of a given policy?
- **Reinforcement Learning: a framework for sequential decision-making**

$$A_t := f_A(S_t, U_{A_t})$$

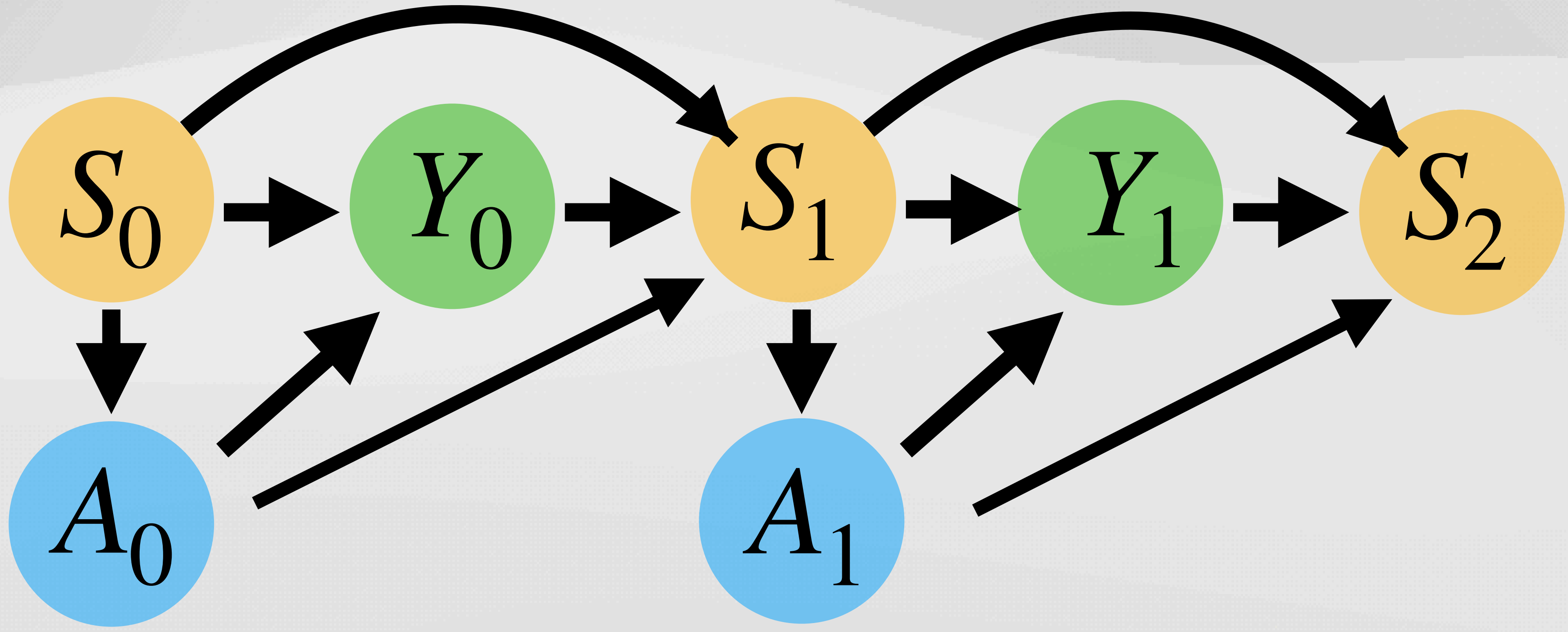
$$Y_t := f_Y(A_t, S_t, U_{Y_t})$$

$$S_{t+1} := f_S(Y_t, A_t, S_t, U_{S_{t+1}})$$

Murphy, Susan A. "Optimal dynamic treatment regimes." Journal of the Royal Statistical Society Series B: Statistical Methodology 65.2 (2003): 331-355.
Chakraborty, Bibhas, and Susan A. Murphy. "Dynamic treatment regimes." *Annual review of statistics and its application* 1.1 (2014): 447-464.
Bibaut, Aurelien, et al. "Sequential causal inference in a single world of connected units." *arXiv preprint arXiv:2101.07380* (2021).
Tran, Allen, Aurélien Bibaut, and Nathan Kallus. "Inferring the long-term causal effects of long-term treatments from short-term experiments." *arXiv preprint arXiv:2311.08527* (2023).

Causal model

- **We assume data follows a <u>Markov decision process</u> (MDP)**
- At each time t, decision-maker is
  - **given <u>state</u>** S_t summarizing current context
  - **takes <u>action</u>** A_t based on S_t
  - **receives <u>outcome</u>** Y_t (cost/reward)
  - **transitions to <u>next state</u>** S_{t+1} based on (S_t, A_t, Y_t)

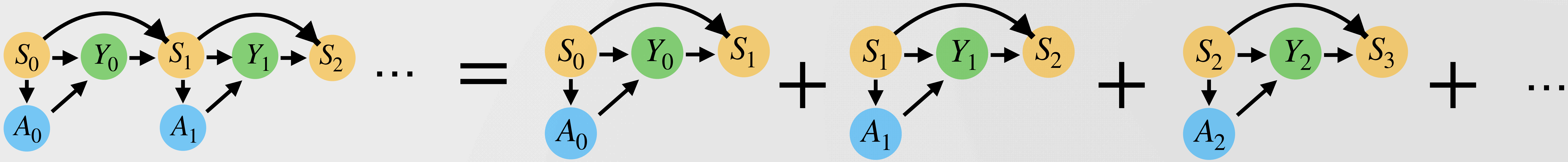


Time-homogeneity of MDP

- **Action taken, outcome received, and state transition don't depend directly on time**

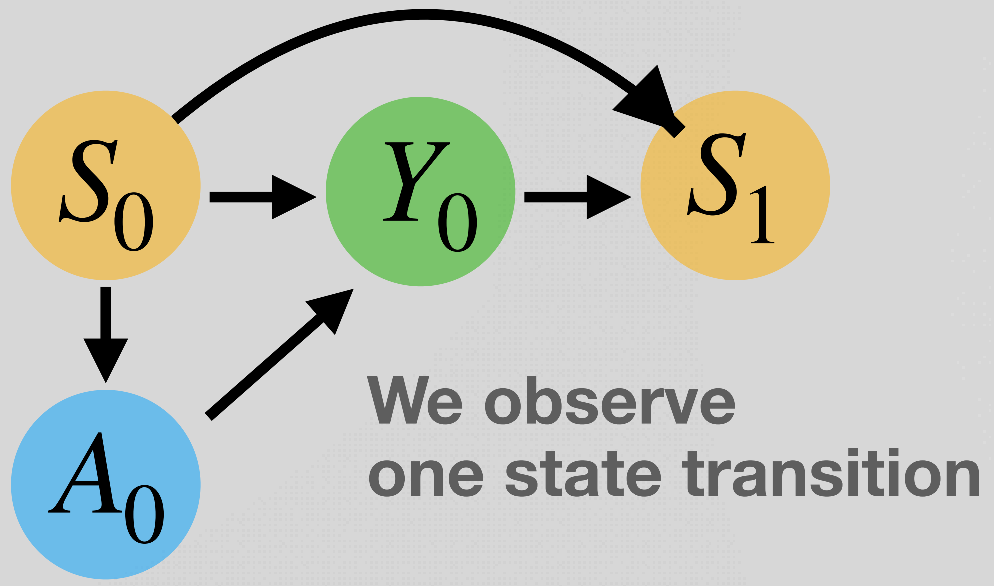
- That is, the following distributions are time-invariant:

Data-generating policy	Reward distribution	State transition distribution
$(A_t \mid S_t = s)$	$(Y_t \mid A_t = a, S_t = s)$	$(S_{t+1} \mid Y_t = y, A_t = a, S_t = s)$



Sequential process obtained by composing single time-step transitions

Objective: Long-term policy evaluation



We observe one state transition

- **Short-term data** : $\{(S_{0,i}, A_{0,i}, Y_{0,i}, S_{1,i}) \sim P_0\}_{i=1}^{n}$
- **Policy π :** probability of taking action a in state s is $\pi(a \mid s)$
- **Our goal** : learn **(long-term) policy value** for discount factor $\gamma \in [0,1]$:

$$\boxed{\psi_0 = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t Y_t(\pi) \right]}$$

Expectation of discounted cumulative reward under counterfactual MDP that follows π

- γ is a “time horizon” that controls how far into the future we look.

Identification via Q-function

- **The <u>Q-function</u> is** $\boxed{q_0(a, s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t Y_t(\pi) \mid A_0 = a, S_0 = s\right]}$

- Policy value ψ_0 equals expectation $E_0[V^\pi(q_0)(S_0)]$, where

$$V^\pi(q_0)(s) = \int q_0(a, s)\pi(a \mid s)da$$

- **Q-function identified by <u>Bellman equation</u>:**

$$\boxed{q_0(a, s) = E_{P_0}\left[Y_0 + \gamma V^\pi(q_0)(S_1) \mid A_0 = a, S_0 = s\right]}$$

In state s,
the value of action a = reward of action a + value from following π starting from S_1 × discount rate
and then following π

Double reinforcement learning

- **DRL** provides efficient nonparametric inference for policy value *(Kallus et al., 2020)*
- Doubly-robust AIPW-style estimator:

Weights

Bellman residual for q_n

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} V^{\pi}(q_n)(S_{0,i})}_{\text{plug-in estimator}} \quad \underbrace{+\frac{1}{n}\sum_{i=1}^{n} d_n(S_{0,i}, A_{0,i})\left\{Y_{0,i} + \gamma V^{\pi}(q_n)(S_{1,i}) - q_n(A_{0,i}, S_{0,i})\right\}}_{\text{augmentation term}}$$

where q_n estimates q_0 and d_n estimates density ratio d_0

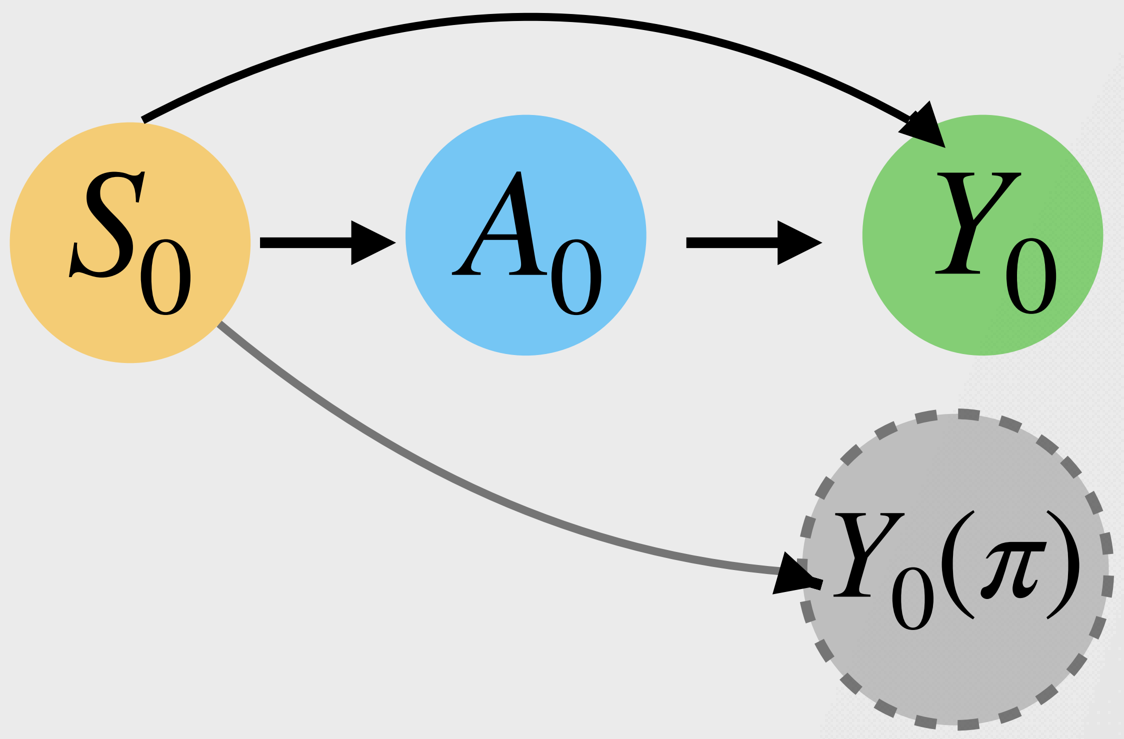
Kallus, Nathan, and Masatoshi Uehara. "Double reinforcement learning for efficient off-policy evaluation in markov decision processes." *Journal of Machine Learning Research* 21.167 (2020): 1-63.
Kallus, Nathan, and Masatoshi Uehara. "Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning." *Operations Research* 70.6 (2022): 3282-3302.
van Der Laan, Mark J., et al. "Online targeted learning for time series." Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies (2018): 317-346.

Overlap challenges in DRL

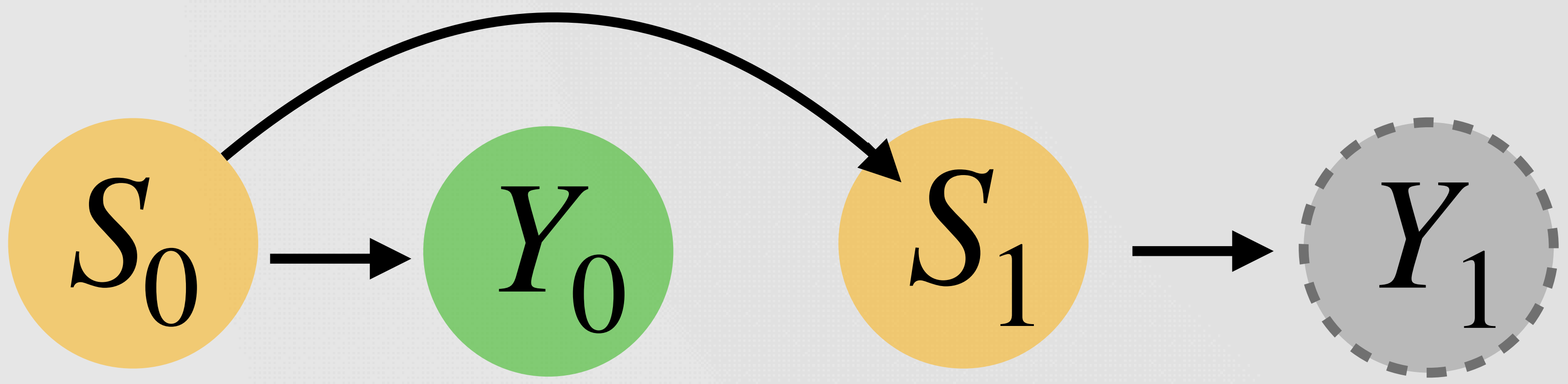
- **Requires existence and <u>finite variance</u> of density ratio:**

$$d_0(a, s) := \underbrace{\frac{\pi(a \mid s)}{P_0(A_0 = a \mid S_0 = s)}}_{\text{overlap between target and behavior policy}} \times \underbrace{\sum_{t=0}^{\infty} \gamma^t \frac{d\mathbb{P}^\pi(S_t = s)}{dP_0(S_0 = s)}}_{\text{overlap between future and initial state distributions}}$$

Need to impute $Y_0(\pi)$ from Y_0

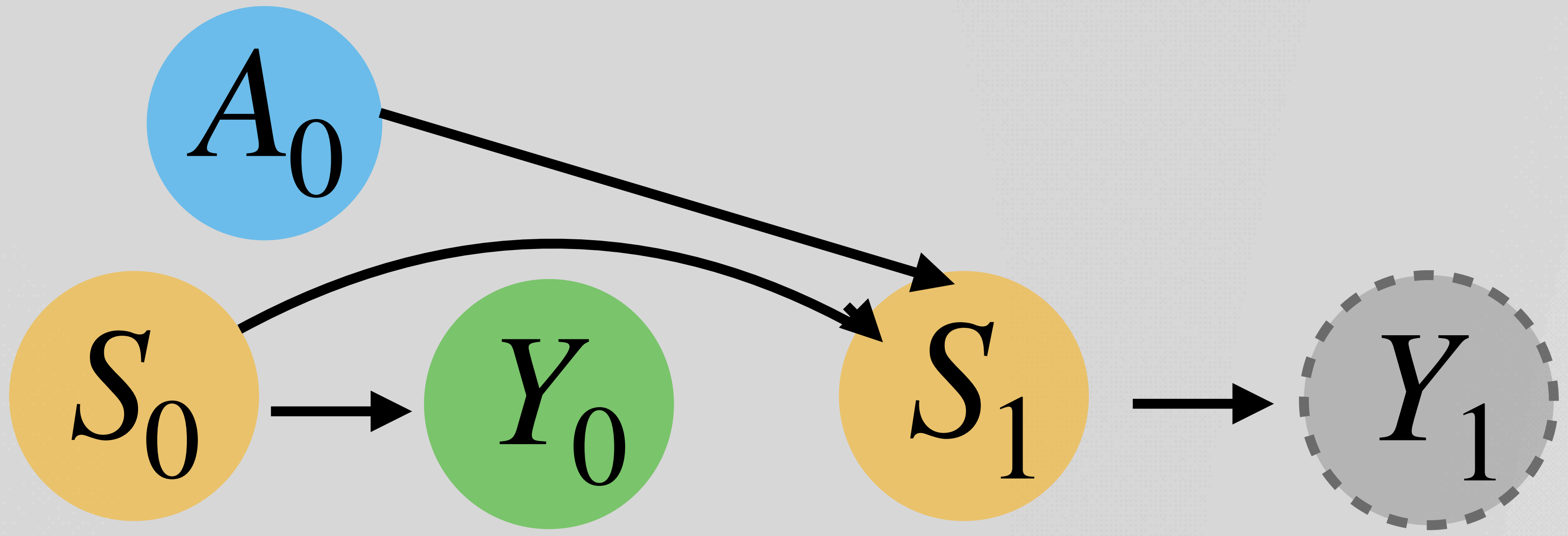


Need to impute $(S_1, \cdot\,)$ from (S_0, Y_0).



Mehrabi, Mohammad, and Stefan Wager. "Off-policy evaluation in markov decision processes under weak distributional overlap." *arXiv preprint arXiv:2402.08201* (2024).

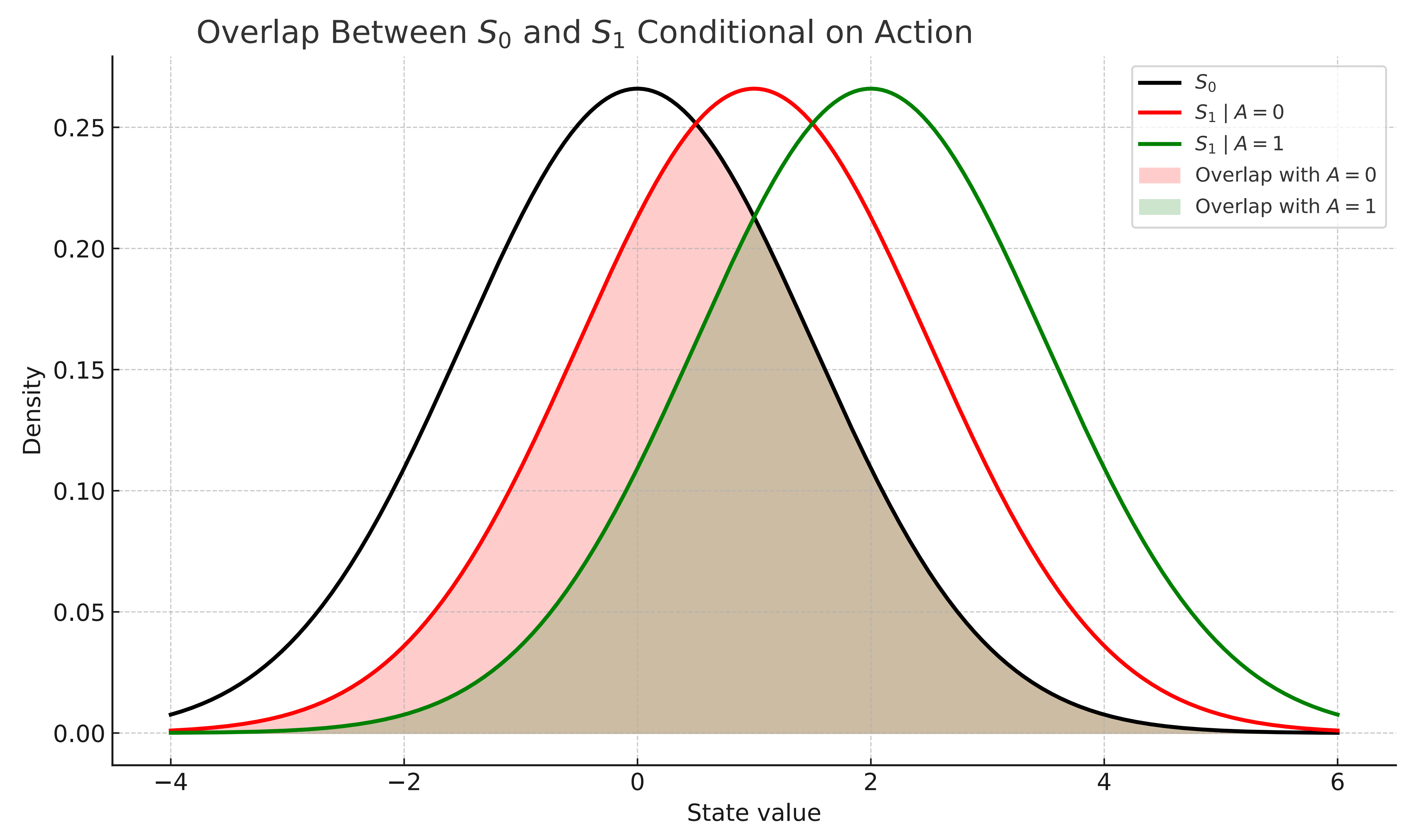
Intertemporal overlap



- **Violated when either time or the policy π induces states that are rare or unseen**

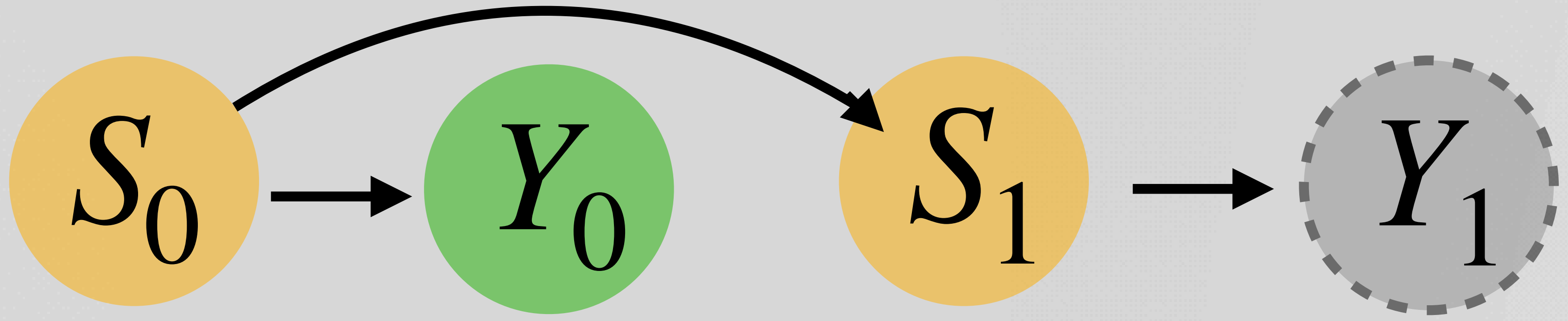
- **Why we care?**
    - Leads to unstable and high variance estimators
    - May cause lack of identification altogether

- **Even a concern in randomized experiments since S_1 is <u>post-treatment</u>**



Mehrabi, Mohammad, and Stefan Wager. "Off-policy evaluation in markov decision processes under weak distributional overlap." *arXiv preprint arXiv:2402.08201* (2024).

How to relax overlap assumptions?



- **Semiparametric restrictions on Q-function reduce overlap requirements.**
  - Allows for extrapolation of outcomes for rare or unseen states
- **Possible semiparametric models:**

Linear model

$$q_0(A_0, S_0) = \varphi(A_0, S_0)^\top \beta$$

Partially linear model

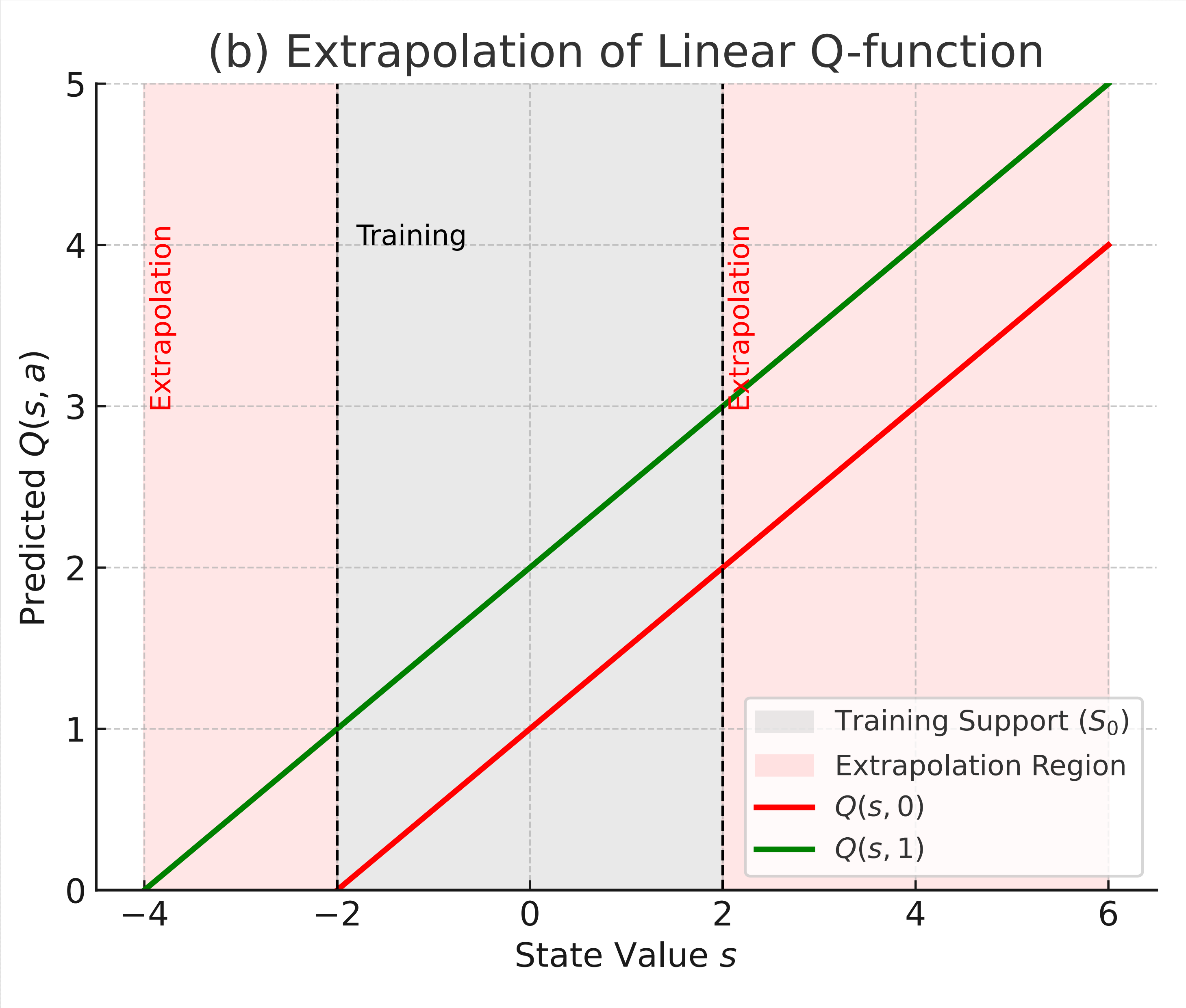
$$q_0(1,S_0) - q_0(0,S_0) = \beta^\top S_0$$

Additive model with $S_0 = (X, Y, Z)$

$$q_0(A_0, S_0) = f_0(A_0, X) + g_0(A_0, Y) + h_0(A_0, Z)$$

Dimension-reduction

$$q_0(A_0, S_0) = \tilde{q}_0(\varphi(A_0, S_0))$$



Our contributions

1. **DRL with semiparametric restrictions on Q-function q_0**
    - Automatic debiasing procedure applies to any linear functional
    - Model-robust inference on best approximation (e.g., BLP)
2. **Model misspecification induces only second-order bias**
    - Valid inference with sieves and data-driven model-selection
    - Reduce variance without sacrificing nonparametric validity
3. **Debiased plug-in estimation via Bellman calibration (remainder of talk)**

Challenge of nuisance estimation in DRL

- **DRL requires estimation of Q-function q_0 and density ratio d_0**
- **Q-function is "easy" to estimate:**
  - Bellman equation says that $q_0(A_0, S_0) = E[Y_0 + \gamma V^\pi(q_0)(S_1) \mid A_0, S_0]$
  - If we knew q_0, we could regress $Y_0 + \gamma V^\pi(q_0)(S_1)$ on (A_0, S_0)

Fitted Q-iteration

```
1. k=0; Initialize $q_n^{(0)} = 0$;
2. Iterate until convergence:
    • update $q_n^{(k+1)}$ by regressing $Y_0 + \gamma V^\pi(q_n^{(k)})(S_1)$ on $(A_0, S_0)$
    • increment: $k = k + 1$
```

Challenge of nuisance estimation in DRL

- Debiasing requires estimation of density ratio d_0
- **Challenging:** need to solve minimax problem

$$\boxed{d_0 = \arg\min_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} L_0(f, g)}$$

- **Issues:**

Computationally expensive

Unstable optimization

Bias due to model misspecification

Our solution

- **Can we avoid estimation of density ratio altogether?**
  - Yes, if we calibrate the Q-function estimator

- **Key result: *Bellman calibration* suffices for debiasing**

If q_n solves bellman equation with $q_n(a, s)$ as 1D dimension reduction:

$$q_n(a, s) \approx E[Y_0 + \gamma V^{\pi}(q_n)(S_1) \mid q_n(A_0, S_0) = q_n(a, s)]$$

Then, the plug-in estimator $\frac{1}{n}\sum_{i=1}^{n} V^{\pi}(q_n)(S_{0,i})$ is asymptotically normal

Isotonic Bellman calibration

- ***We propose <u>isotonic Bellman calibration</u>, extending isotonic calibration to MDPs.***

Machine Learning

Fit Q-function q_0

$$\xrightarrow{q_n}$$

<u>Bellman calibration</u>

Regress

$$Y_{0,i} + \gamma V^{\pi}(q_n)(S_{1,i})$$

on

$$q_n(A_{0,i}, S_{0,i})$$

using isotonic regression

Post-hoc
1D regression (cheap compute)
No tuning
One line of code

$$\xrightarrow{g_n \circ q_n}$$

Plug-in

Estimator

$$\frac{1}{n}\sum_{i=1}^{n} V^{\pi}(g_n \circ q_n)(S_{0,i})$$

Properties of Bellman-calibrated plug-in

Estimator

$$\frac{1}{n}\sum_{i=1}^{n} V^{\pi}(g_n \circ q_n)(S_{0,i})$$

- **Semiparametric efficient under model** with $q_0(A_0, S_0)$ as **1D dimension reduction** of (A_0, S_0)
- **Asymptotically linear** and **superefficient** under **nonparametric model**
- **Relaxes overlap** condition to **finite variance of 1D density ratio**:

$$d_{q_0}(a, s) := \sum_{t=0}^{\infty} \gamma^t \frac{d\mathbb{P}^{\pi}(q_0(A_t, S_t) = q_0(a, s))}{dP_0(q_0(A_0, S_0) = q_0(a, s))}$$

Conclusion

- **DRL faces two key challenges:**
  1. Requires Inter-temporal overlap across states, on top of treatment overlap
  2. Debiasing requires min-max estimation of density ratio nuisance
- **Our solutions:**
  - Semiparametric extension of DRL to relax overlap
  - Bellman calibration of Q-function debiases without nuisance estimation
- **Note: Bellman-calibration tackles both overlap and nuisance estimation challenges.**

"Automatic double reinforcement learning in semiparametric markov decision processes
with applications to long-term causal inference." arXiv preprint arXiv:2501.06926 (2025).