Harnessing Causality in Reinforcement Learning with Bagged Decision Times

Daiqi Gao, Hsin-Yu Lai, Predrag Klasnja, Susan Murphy
Harvard University, Allen Institute, University of Michigan

Goal: Address Bagged Decision-Time RL

Definition of a bagged decision-time problem:

- A bag contains a sequence of decision times
- All actions in a bag affect one reward, observed at the end of the bag
- It is a stationary MDP across bags but not within.

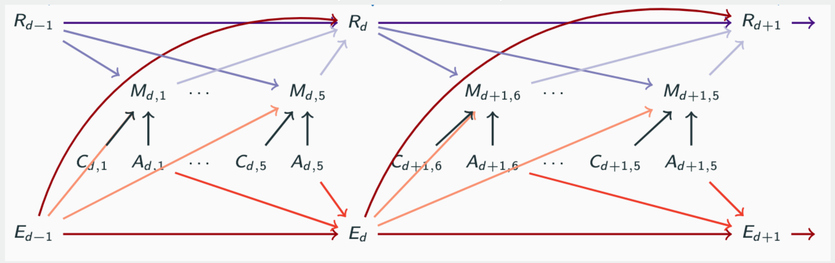
Motivating example — a mobile health intervention, HeartSteps:

- Goal: encourage individuals to increase physical activity (PA)
- Reward: daily survey of "commitment to being active"
- Action: Whether to deliver an activity suggestion
- Decision time: Five times a day
- All five actions within a day affect the daily reward:

Whether a person receives an activity suggestion affects whether they exercise, which affects their sense of commitment to being active.)

Approach: Consider causality

- If some mediators at each decision time are observed, we may utilize these mediators to define an efficient algorithm.
- Continue to use HeartSteps as an example



- R_d : the commitment survey collected at the end of day d
- E_d : the engagement at the end of day d
- $C_{d,k}$: the context measured at each decision time k on day d
- $A_{d,k}$: the action at each decision time k on day d
- $M_{d,k}$: 30-minute step count after each decision time

Main contributions:

- Use the causal DAG to reduce the "problem into a periodic MDP.
- Demonstrate that the Bayesian sufficient statistic of the observed history yields the maximal optimal value function.

Definitions and Theorems

Definition (K-Periodic MDP):

Consider a tuple $(S_{1:K}, A_{1:K}, P_{1:K}, r_{1:K}, \gamma_{1:K}, \nu)$ where ν is the initial state distribution. Assume $P_1 : S_K \times A_K \to \Delta(S_1)$ and $P_k : S_{k-1} \times A_{k-1} \to \Delta(S_k)$ for $1 < k \leq K$.

Theorem (Bellman Optimality Equations can be extended to K-periodic MDP):

For a K-periodic MDP, the functions $Q_{1:K}$ are the optimal Q-functions if and only if these functions satisfy the Bellman optimality equations.

Lemma (Bagged Decision-Time to K-Periodic MDP):

Let $\{U_{d,1:K}\}$ be the states where 1)$U_{d,1:K}$ is bag-invariant, 2)the state transition is Markovian within and across the bag, 3)$R_d \perp U_{t,l} \,|\, U_{d,k}, A_{d,k}$ for any $(t,l) < (d,k)$, 4) $U_{d,k}, A_{d,k}$ blocks the path from $d-1$ to d. Then, by letting $S_{1:K} \equiv \{U_{d,1:K}\}$, $r_{d,1:K-1} \equiv 0$, and $r_K \equiv \{R_d\}$, the bagged decision-time problem is a K-periodic MDP.

Definition (Dynamical Bayesian sufficient statistic D-BaSS)

Given stochastic processes H, R, a process S is a D-BaSS of H w.r.t. R given a process A if 1) $\exists$ a function $F(\cdot)$ s.t. $S_t = F(H_{1:t})$, and 2) $R_{t'} \perp H_{1:t} \,|\, S_t, A_t$ for any $t' \geq t$. Moreover, a process Z is a minimal D-BaSS if there exists a function $f(\cdot)$ s.t. $f(S_t) = Z_t$.

Theorem (Bayesian sufficient statistic yields the maximal optimal value function)

Suppose $\{S_{d,k}\}$ is the minimal D-BaSS of $\{H_{d,k}\}$ w.r.t. $\{R_{d,k}\}$ given $\{A_{d,k}\}$. Then we have $V_k^{U*}(u_{d,k}) \leq V_k^{S*}(s_{d,k})$.

Algorithm: Bagged RLSVI (BRLSVI)

Algorithm 1 Bagged RLSVI (BRLSVI)

Input: Hyperparameters L, λ_d, σ^2.

1: Warm-up: Randomly take actions $A_{d,k} \sim$ Bernoulli(0.5) in bag $d \in \{1:L\}$ for $k \in \{1:K\}$.
2: **for** $d \geq L+1$ **do**
3: **for** $k = K, \ldots, 1$ **do**
4: Construct $\boldsymbol{X}_{1:(d-1),k}$, $Y_{1:(d-1),k}$ with $\widetilde{\boldsymbol{\beta}}_{d-1,1}$ when $k = K$ and with $\widetilde{\boldsymbol{\beta}}_{d,k+1}$ when $k < K$ using (7). Obtain $\boldsymbol{\mu}_{d,k}, \Sigma_{d,k}$ using (8).
5: Draw $\widetilde{\boldsymbol{\beta}}_{d,k} \sim N(\boldsymbol{\mu}_{d,k}, \Sigma_{d,k})$.
6: **end for**
7: **for** $k = 1, \ldots, K$ **do**
8: Observe $H_{d,k}$ and construct $S_{d,k}$.
9: Take $A_{d,k} = \text{argmax}_{a \in \mathcal{A}_k} \phi_k(S_{d,k}, a)^T \widetilde{\boldsymbol{\beta}}_{d,k}$.
10: **end for**
11: Observe $M_{d,K}, E_d, R_d$.
12: **end for**

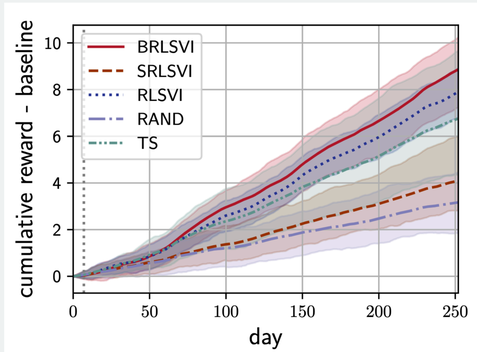
$$\mathcal{Q}_k(s_{d,k}, a_{d,k}) = \phi_k(s_{d,k}, a_{d,k})^T \boldsymbol{\beta}_k, \qquad (7)$$

$$\boldsymbol{X}_{t,k} = \phi_k(S_{t,k}, A_{t,k}),$$

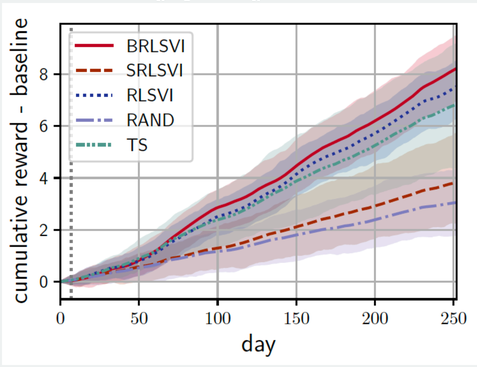
$$Y_{t,k} = \begin{cases} \max_{a \in \mathcal{A}_{k+1}} \phi_{k+1}(S_{t,k+1}, a)^T \widetilde{\boldsymbol{\beta}}_{d,k+1} & \text{if } k < K, \\ R_t + \bar{\gamma} \max_{a \in \mathcal{A}_1} \phi_1(S_{t+1,1}, a)^T \widetilde{\boldsymbol{\beta}}_{d-1,1} & \text{if } k = K, \end{cases} \qquad (8)$$

HeartSteps Simulation Study

- K-Periodic MDP setup:
  $S_{d,k} = [E_{d-1}, R_{d-1}, M_{d,1:(k-1)}, C_{d,k}]$
- Compare with
  - Stationary RLSVI: treat each bag as a decision point and choose K actions at the beginning of each bag.
  - RLSVI with a finite horizon K.
  - Random policy.
  - Thompson sampling where $M_{d,k}$ is optimized.
- Results
  - Vanilla testbed



- Arrow from R_{d-1} to E_d exists



This research was funded by NIH grants P50DA054039, P41EB028242, R01HL125440-06A1, UH3DE028723, and P30AG073107-03 GY3 Pilots.