Dyadic Reinforcement Learning

Shuangning Li

HARVARD UNIVERSITY

Dyadic Reinforcement Learning

Li, Shuangning, Lluis Salvat Niell, Sung Won Choi, Inbal Nahum-Shani, Guy Shani, and Susan Murphy. *arXiv preprint* arXiv:2308.07843 (2023).

Dyadic Reinforcement Learning

Mobile health trial
ADAPTS-HCT

Goal: enhance medication adherence

Adolescent
Cancer Patient

Dyadic Reinforcement Learning

Mobile health trial
ADAPTS-HCT

Goal: enhance medication adherence

Reinforcement learning algorithm:

Decides whether/when to deliver interventions that target the cancer patient 

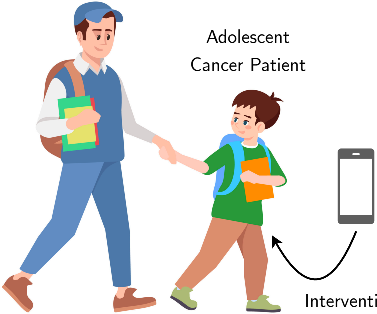
Dyadic Reinforcement Learning

Mobile health trial
ADAPTS-HCT

Goal: enhance medication adherence

Reinforcement learning algorithm:

Decides whether/when to deliver interventions that target the cancer patient 



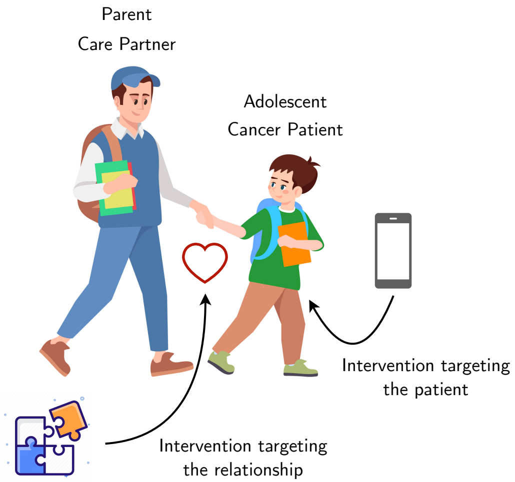
Dyadic Reinforcement Learning

Mobile health trial
ADAPTS-HCT

Goal: enhance medication adherence

Reinforcement learning algorithm:

- Decides whether/when to deliver interventions that target the cancer patient
- Decides whether/when to deliver interventions that target the relationship



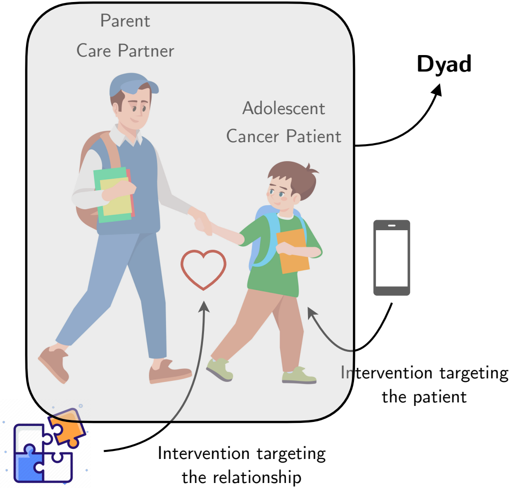
Dyadic Reinforcement Learning

Mobile health trial
ADAPTS-HCT

Goal: enhance medication adherence

Reinforcement learning algorithm:

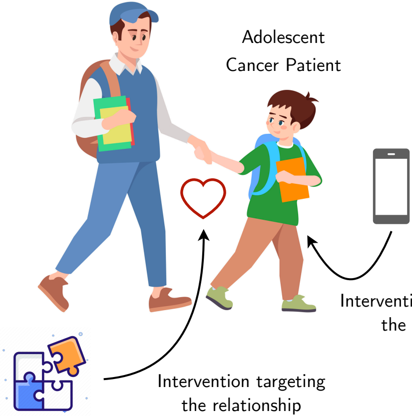
- Decides whether/when to deliver interventions that target the cancer patient
- Decides whether/when to deliver interventions that target the relationship



Dyadic Reinforcement Learning

Intervention targeting the patient:

❖ Daily reminder

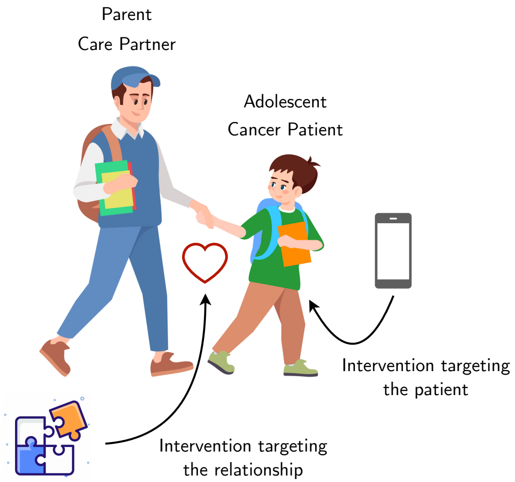


Dyadic Reinforcement Learning

Intervention targeting the patient:
- ❖ Daily reminder

Intervention targeting the relationship:
- ❖ Weekly prompt for the dyad to play a joint game.
- ❖ A puzzle-solving game.
- ❖ The game lasts throughout the week.
- ❖ Once the adolescent takes the medication, it triggers a clue for the parent.
- ❖ If they win the game, a donation is made to their favorite charity.

Smart medication boxes with sensors!



Notation

State

- Weekly state S_w^{weekly}: weekly measurements of the quality of the dyadic relationship.
- Daily state S_t^{daily}: various daily measurements related to adolescents and their parents, such as their step count, sleep duration or mood.

Action

- Weekly action A_w^{weekly}: whether to send the weekly intervention to encourage the adolescent and their parents to participate in the joint game.
- Daily action A_t^{daily}: whether to send the daily reminder to the adolescent.

Reward R_t: whether the adolescent cancer patient takes the medication.

Notation

State

- Weekly state S_w^{weekly}: weekly measurements of the quality of the dyadic relationship.
- Daily state S_t^{daily}: various daily measurements related to adolescents and their parents, such as their step count, sleep duration or mood.

Action

- Weekly action A_w^{weekly}: whether to send the weekly intervention to encourage the adolescent and their parents to participate in the joint game.
- Daily action A_t^{daily}: whether to send the daily reminder to the adolescent.

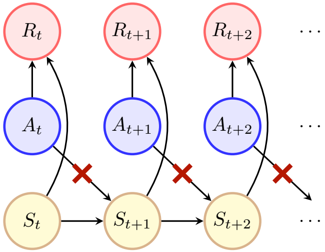
Reward R_t: whether the adolescent cancer patient takes the medication.

Let's focus on the daily state and action temporarily!

Bandit Algorithm?

☺ Contextual bandit algorithms are commonly used in mobile health studies.
They can run reliably and stably in an online environment.

✦ Assumes that there is no delayed effect of actions.

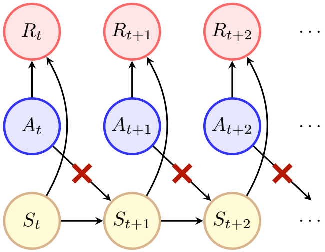


Bandit Algorithm?

Contextual bandit algorithms are commonly used in mobile health studies.
They can run reliably and stably in an online environment.

✦ Assumes that there is no delayed effect of actions.

"Burden effect" in mobile health studies:
Users feel burdened or disengage when receiving too many messages.



Bandit Algorithm?

Contextual bandit algorithms are commonly used in mobile health studies.
They can run reliably and stably in an online environment.

✦ Assumes that there is no delayed effect of actions.

"Burden effect" in mobile health studies.

✦ Assumes stationarity.

Bandit Algorithm?

Contextual bandit algorithms are commonly used in mobile health studies.
They can run reliably and stably in an online environment.

✦ Assumes that there is no delayed effect of actions.

"Burden effect" in mobile health studies.

✦ Assumes stationarity.

Nonstationarity especially within a week.

Bandit Algorithm?

☺ Contextual bandit algorithms are commonly used in mobile health studies.
They can run reliably and stably in an online environment.

✦ Assumes that there is no delayed effect of actions.

☹ "Burden effect" in mobile health studies.

✦ Assumes stationarity.

☹ Nonstationarity especially within a week.

Bandit algorithm:

Variance low ☺ Bias high ☹

RL Algorithm?

☺ Consider RL algorithms that take into account full non-stationarity and delayed effect.
Treat the problem as a finite-horizon reinforcement learning problem.
E.g., one can run the RLSVI (randomized least-squares value iteration) algorithm.

RL Algorithm?

☺ Consider RL algorithms that take into account full non-stationarity and delayed effect.
Treat the problem as a finite-horizon reinforcement learning problem.
E.g., one can run the RLSVI (randomized least-squares value iteration) algorithm.

✦ Need to keep many parameters to capture non-stationarity and delayed effect.

☹
- In ADAPTS-HCT, each dyad will be involved in the trial for 15 weeks $\approx$ 100 days. Need to maintain 100 different Q-functions for each day in the trial.
- Mobile health data is usually very noisy. ➔ Hard to capture delayed effects.

RL Algorithm?

☺ Consider RL algorithms that take into account full non-stationarity and delayed effect.
Treat the problem as a finite-horizon reinforcement learning problem.
E.g., one can run the RLSVI (randomized least-squares value iteration) algorithm.

✦ Need to keep many parameters to capture non-stationarity and delayed effect.

☹
- In ADAPTS-HCT, each dyad will be involved in the trial for 15 weeks $\approx$ 100 days. Need to maintain 100 different Q-functions for each day in the trial.
- Mobile health data is usually very noisy. ➔ Hard to capture delayed effects.

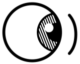
Full RL algorithm:

Variance high ☹ Bias low ☺

Dyadic RL Algorithm

Assumes that the impact of actions does not extend to the following **day**.

Allows for delayed effects.



Bandit algorithm

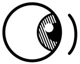
Full RL algorithm

Low Variance
High Bias

High Variance
Low Bias

Dyadic RL Algorithm

Assumes that the impact of actions does not extend to the following **day**.



Bandit algorithm

Low Variance
High Bias

Assumes that the impact of actions does not extend to the following **week**.

Dyadic RL algorithm

Find a balance between variance and bias

Allows for delayed effects.

Full RL algorithm

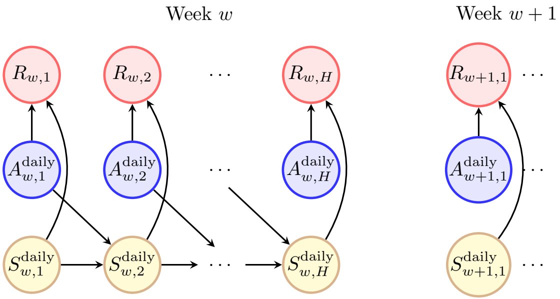
High Variance
Low Bias

Dyadic RL Algorithm

Domain science tells us that:

✓ Weeks exhibit similar structures.

✓ There is a high level of noise in state transitions and rewards.



The algorithm makes the assumption that the impact of actions does not extend to the following **week**.
This assumption is to address the challenge of high noise.

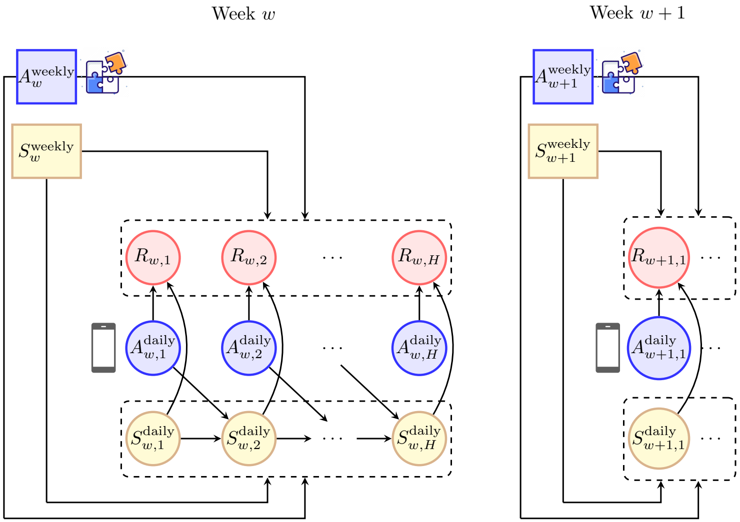
Dyadic RL Algorithm

Two types of interventions:

❖ Daily reminder

❖ Weekly game prompt

Weekly game prompt is expected to impact the dyad throughout the entire week.



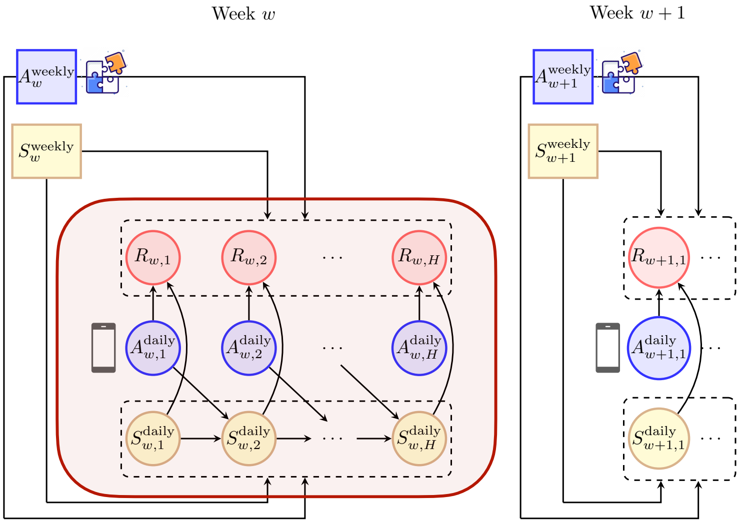
Dyadic RL Algorithm

Daily action:

★ Finite-horizon problem with $H = 7$.

★ We choose to use RLSVI because of its Bayesian nature.

➔ Helps with interpretability.



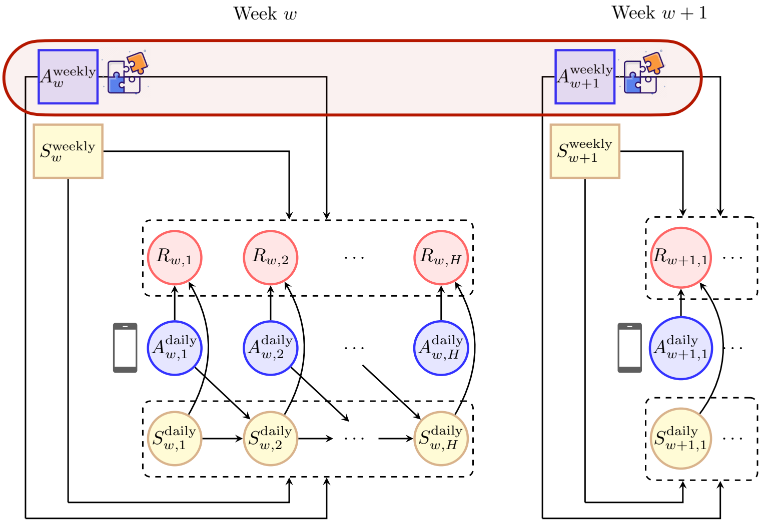
Dyadic RL Algorithm

Weekly action:

★ Contextual bandit problem.

★ We choose to use Thompson Sampling because of its Bayesian nature.

➔ Helps with interpretability.



Dyadic RL Algorithm

Weekly action:

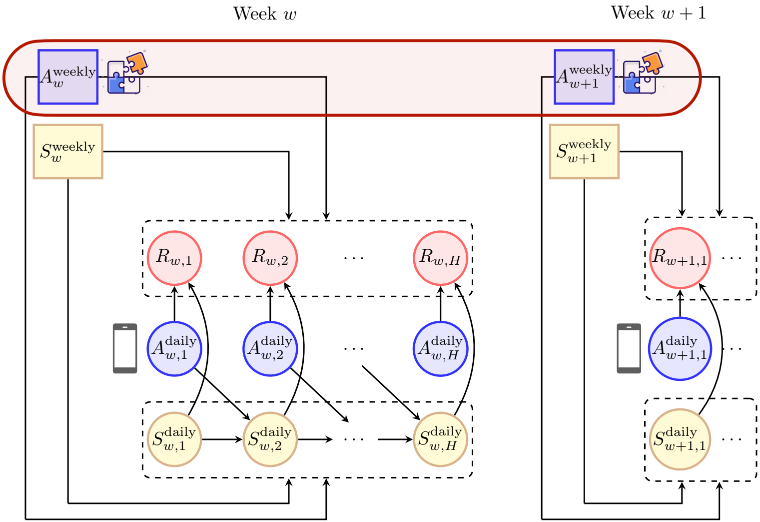
★ Contextual bandit problem.

★ We choose to use Thompson Sampling because of its Bayesian nature.
  ➔ Helps with interpretability.

Reward?

★ Sum of realized rewards: too noisy!

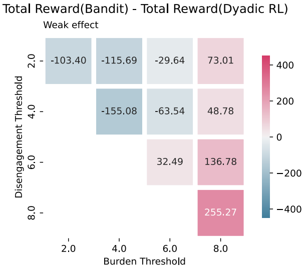
★ Estimate of the Q-function on day 1.

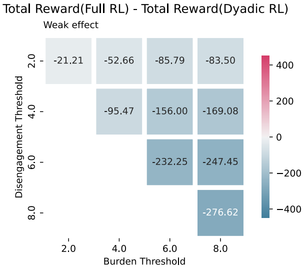


Dyadic RL Algorithm

Theoretically, we establish a regret bound for the dyadic RL algorithm within a tabular setting.

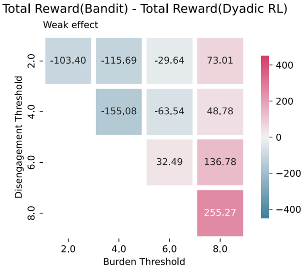
Empirically, we demonstrate the dyadic RL algorithm's performance through simulation studies on both toy scenarios and on a realistic test bed constructed from data collected in a mobile health study.

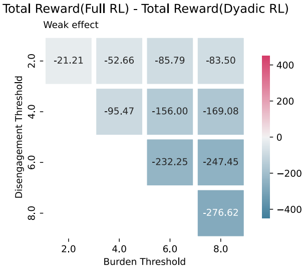


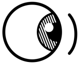


Simulation Test Bed

Total Reward(baseline) - Total Reward(Dyadic RL): green means dyadic RL is performing better







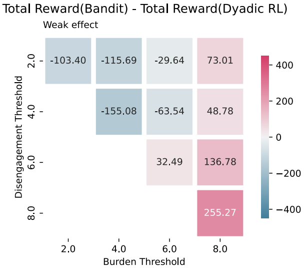
Bandit algorithm

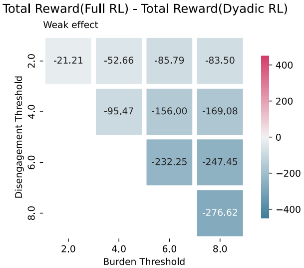
Full RL algorithm

Simulation Test Bed

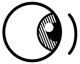
Total Reward(baseline) - Total Reward(Dyadic RL): green means dyadic RL is performing better

More likely to **disengage** when receiving too many messages





More likely to **feel burdened** when receiving too many messages



Bandit algorithm

Full RL algorithm

Thank You!

Shuangning Li

https://lsn235711.github.io/