Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity

Peng Liao[1], Kristjan Greenewald[3], Predrag Klasnja[1] and Susan Murphy[2]
University of Michigan[1], Harvard University[2], IBM Research[3]

HEARTSTEPS STUDY

- ***HeartSteps*** is an ongoing, real-life physical activity clinical trial for improving the physical activity of individuals with blood pressure in the stage 1 hypertension range.
- Participants are provided a Fitbit tracker and a mobile phone application on the phone designed to help them improve their physical activity.
- One of the interventions is a **contextually tailored physical activity suggestion** that may be delivered at any of the five user-specified times during each day, corresponding to the user's morning commute, lunch, afternoon, evening, and post-dinner times.
- The content of the suggestion is designed to encourage activity in the current context

REINFORCEMENT LEARNING FRAMEWORK

$$\{S_1, A_1, R_2, S_2, , A_2, R_3, \dots, S_t, A_t, R_{t+1}\}$$

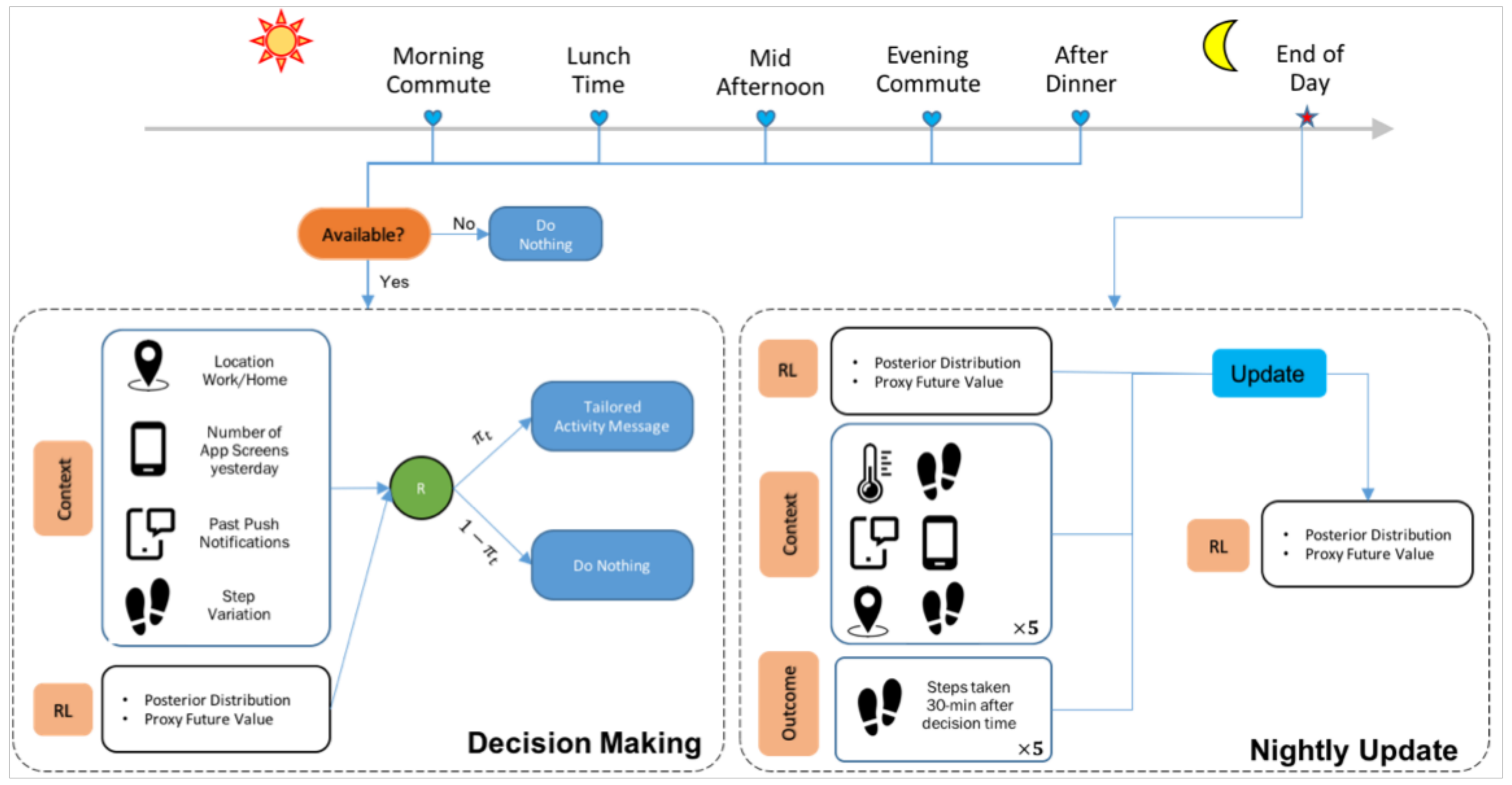
- Decision time t: 5 times/day, 90-day study
- Action A_t : binary - send vs do nothing
- States S_t: location, prior 30-minute step count, daily step count, the temperature, app usage…
- Reward R_{t+1} : number of steps taken in the 30 minutes after decision time
- **Goal**: at each decision time, *determine whether or to send the walking suggestion message* based on user's context, with the goal to maximize total step counts

CHALLENGES TO APPLYING RL IN MOBILE HEALTH

- *The learning algorithm should learn quickly and accommodate noisy data.*
  Most online RL algorithms require the agent to interact many times with the environment prior to performing well. This is impractical in mobile health applications as users can lose interest and disengage quickly.
  Proposal: Use a low-dimensional linear model with an informative prior

- *The learning algorithm should accommodate some model mis-specification and non-stationarity.*
  Due to unobserved aspects of the current context (engagement or burden), observed human behavior is complex to model and often exhibits non-stationarity over longer periods of time
  Proposal: Use action-centering in modeling the reward

- *The learning algorithm must adjust for longer term effects of current actions*
  In mobile health, interventions often tend to have positive effect on the immediate reward, but likely produce negative impact on future rewards due to habituation and/or burden.
  Proposal: Construct a low-variance proxy of future rewards based on a dosage variable using an approximate MDP



PERSONALIZED HEARTSTEPS

- Let $r(s,a)$ be the expected reward given state s and action a. A low-dimensional linear model for immediate treatment effect
$$r(s,1) - r(s,0) = f(s)^\top \beta$$
- Assume a **working** model $r(s,0) \approx g(s)^\top \alpha$ for baseline reward. The posterior distribution of β is found by an <u>action-centered</u> Bayesian linear model:
$$R_{t+1} = g(S_t)^\top \alpha_0 + \boldsymbol{\pi_t} f(S_t)^\top \alpha_1 + (\boldsymbol{A_t} - \boldsymbol{\pi_t}) f(S_t)^\top \beta + \epsilon_t$$
- To capture the **delayed effect of action**, we construct a proxy of future rewards using a dosage variable X_t,
$$X_{t+1} = \lambda X_t + A_t$$
- The proxy value function $\hat{H}(x,a)$ is formed based on a **parsimonious MDP** for the states $S_t = (X_t, Z_t)$, where $\{Z_t\}$ are i.i.d. with some estimated distribution F:
$$\hat{H}(X_t, a) \approx E_{\pi^*}[R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} \dots | X_t, A_t = a]$$
  Note $r(s,a) + \gamma \hat{H}(x,a)$ approximates the standard optimal state-action (Q) value with discount rate γ
- The action A_t is selected stochastically with probability π_t, akin to **Thompson Sampling Bandit**, adjusted by the proxy value: $\tilde{\beta} \sim N(\mu_\beta, \Sigma_\beta)$ and
$$\pi_t = \Pr\left(f(S_t)^\top \tilde{\beta} > \boldsymbol{\gamma \hat{H}(X_t, 0)} - \boldsymbol{\gamma \hat{H}(X_t, 1)}\right)$$

IMPLEMENTATION

- A 42-day pilot study (HS 1.0) is conducted with 37 participants. The feature vectors, as well as the prior distribution of parameters are selected based on GEE analysis of pilot study.
- The tuning parameters are chosen based on a generative model built from HS 1.0. Cross-validation is performed to demonstrate the use of proxy value is able to pick up the delayed effect of treatment – compare with standard Thompson sampling Bandit