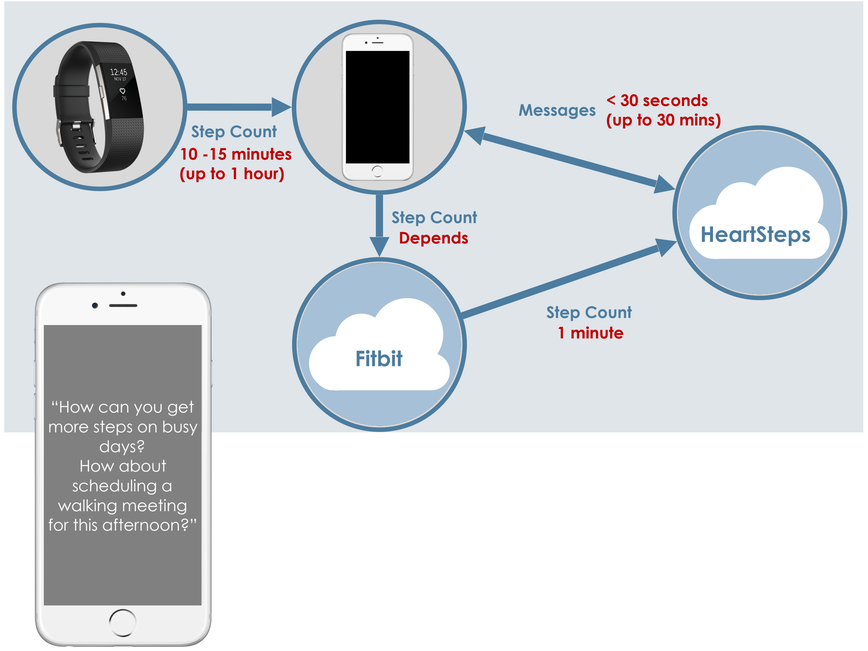
Online RL in HeartSteps: a Robust Contextual Bandit Approach

Kristjan Greenewald, Ambuj Tewari, Predrag Klasnja, Susan Murphy

HeartSteps: Increasing physical activity with mHealth

- Many of the risk factors for heart disease are behavioral - physical inactivity, smoking, and diets high in saturated and trans-fats.
- To successfully adopt a heart-healthy lifestyle, individuals have to make many healthy decisions throughout the day, often when they are not thinking about health but are busy with work, family, and recreational activities.
- HeartSteps aims to support maintenance of physical activity after cardiac patients finish cardiac rehabilitation.
- **Goal**: Provide smartphone-based just in time, adaptive behavioral interventions (JITAI) to increase physical activity in cardiac patients.



RL in HeartSteps

- Intervention is a tailored smartphone notification message. Want to determine optimal times to send the intervention.
- Need for extreme **personalization**. Everyone's routine is different.
- **Reinforcement learning** framework:
  - 5 decision points per day, per user.
  - Observe user **contextual variables** in real time.
  - **Action**: send or don't send an intervention.
  - **Reward**: to be the number of steps taken in the 30 minutes after a decision point.

Requirements

- Context driven
- Do not have the data for pre-study batch trained RL.
- Efficient online learning
  - Need to learn and make decisions starting day one of the study.
- Personalization
  - Each person needs their own mapping from state to action.
  - Study is asynchronous anyway.
- Bottom line: Learn quickly from very few samples.

Contextual variables

- Chosen via rigorous analysis of HeartSteps v1 preliminary study.
- Interaction variables
  - Interaction with good morning message (reject/accept)
  - Number of messages sent in past 7 days
  - Location indicator
- Control variables
  - Log-transformed tracker steps 30 minutes prior to decision point
  - Square root of total steps yesterday
  - Outdoor temperature

Thompson Sampling Contextual Bandit

- [Agrawal et al 2013]: Linear model of the reward for quick learning.
- Bayesian prior and model updates.
- At each decision point, randomly chooses an action with probability determined by the Bayesian posterior of the reward function.
- Guaranteed sublinear regret.

Algorithm 1 Thompson Sampling for Contextual bandits

Set $B = I_d, \hat{\mu} = 0_d,\ f = 0_d$.
for all $t = 1, 2, \ldots,$ **do**
  Sample $\tilde{\mu}(t)$ from distribution $\mathcal{N}(\hat{\mu}, v^2 B^{-1})$.
  Play arm $a(t) := \arg\max_i b_i(t)^T \tilde{\mu}(t)$, and observe reward r_t.
  Update $B = B + b_{a(t)}(t) b_{a(t)}(t)^T,\ f = f + b_{a(t)}(t) r_t,\ \hat{\mu} = B^{-1} f$.
end for

Shortcomings

- Converges to a deterministic policy - inclusion of some randomization improves engagement and offline analyses.
- Sending a message should never have a negative effect, thus bandit will converge to "always send" policy, overtreating users.
- Does not consider future negative impacts to sending a message - "burden", "habituation", etc.
- Assumes the reward function is static, not changing in time.

Robust Contextual Bandit

- Overtreatment: The true RL question is not "**does the action increase reward**", it is "**when are the optimal times to act**".
- Set a soft **budget** on actions/day.
  - Limited data so set to 3, can learn in future.
- Force budget by combining bandit with **feedback control**.

- Bound action probability away from 0 and 1
  - Keeps from overcommitting to deterministic policy.
- Gaussian process prior
  - Allows for continuous adaptation to nonstationarity.

Algorithm 2 Robust Thompson Sampling for HS2

$\theta, \phi \in \mathbb{R}^{d_1}, \mathbb{R}^{d_2}$ respectively.
$\eta = [\theta; \phi] \in \mathbb{R}^{d_1+d_2}$.
Reward noise variance σ^2 (Gaussian).
Given prior diagonal covariance Σ_0, mean μ_0. Set feedback coefficient α, desired average β number of interventions-per-day.
Set $\Sigma = \Sigma_0, \hat{\eta} = \mu_0$.
for 3 week cycles (2 on, 1 off) **do**
  for $days = 1, 2, \ldots, 21$ **do**
    for decision points 1-5 **do**
      Obtain context features s_t and N_t number of actions taken in past 24 hours.
      Feedback mean offset $\delta_\mu = \alpha(N_t - \beta)_+$.
      Compute probability of taking action 1:

$$\pi(t) = \Pr\left(\mathcal{N}\left(s_t^T \hat{\theta} - \delta_\mu, s_t^T \Sigma_{1:d_1, 1:d_1} s_t\right) \geq 0\right)$$

      Make sure $\pi(t) \in [.2, .8]$ by $\pi(t) = \min(.8, \max(\pi(t), .2))$.
      Choose arm $a(t) = 1$ with probability $\pi(t)$.
    end for
    End of day: Observe rewards r_t on the decision points for the day.
    for decision points 1-5 **do**
      Form $\tilde{s}_t = [a(t) s_t; \bar{s}_t]$. Form $R = \Sigma_{t-1}(\tilde{s}_t^T \Sigma_{t-1} \tilde{s}_t + \sigma^2))^{-1} \tilde{s}_t$.
      Use GP prior to propagate from last point, then update

$$\hat{\eta}_{t|t-1} = \mu_0 + \gamma(\hat{\eta}_{t-1} - \mu_0),\ \Sigma_{t|t-1} = \gamma^2 \Sigma_{t-1} + \epsilon^2 \Sigma_0$$

$$\hat{\eta}_t = [\hat{\theta}_t; \hat{\phi}_t] = \hat{\eta}_{t|t-1} + K_t(r_t - \tilde{s}_t^T \hat{\eta}_{t|t-1}),\ \Sigma_t = (I - K_t \tilde{s}_t^T)\Sigma_{t|t-1}$$
$$K_t = \Sigma_{t|t-1} \tilde{s}_t \Gamma_t^{-1},\ \Gamma_t = \tilde{s}_t^T \Sigma_{t|t-1} \tilde{s}_t + \sigma^2$$

    end for
  end for
end for

Current Status

- Finishing incorporation into Google App Engine based system for deployment in HeartSteps v2 study.
- Initial testing and recruitment to begin soon.
- Will be able to evaluate effectiveness of both the intervention and the learning algorithm.
- Working on proving regret bounds.

References

[1] K. Greenewald, A. Tewari, P. Klasnja, S. Murphy, "Action Centered Contextual Bandits," to appear NIPS 2017.