mHealth Adventures in Sequential Experimentation & Reinforcement Learning!

Susan A Murphy
07.07.18
Outline

- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question
mHealth Science Goals

• Promote behavior change and maintenance of this change
  – Assist user in achieving long term goals
    • Recovery from addictions; avoid heart attacks; maintain independence
  – Manage chronic illness

• Test, evaluate, develop causal behavioral science
mHealth Treatment Actions
Outline

• mHealth
• **Micro-Randomized Trials**
• Challenges to RL
• HeartSteps V1, V2
• A Butchered Bandit
• Open Question
Micro-randomized Trial

- Micro-randomized trial = combination of factorial experimental design with explicitly controlled exploration
- Exploration via use of online forecasting (and RL algorithms)
- Multiple treatment factors occurring at different time scales and which target different rewards
- Probabilistic budgets on # of treatment pushes to manage habituation/burden
- Off-policy, after study is over, causal inference
BariFit MRT to Promote Weight Maintenance Among People Who Received Bariatric Surgery

Prior to the study
Each participant randomized 2x at baseline

Each day of study
- Set Goals
  - Use 60th percentile of daily step counts over 10 prior days as goal

Observations
- step count (sensor)
- whether participant tracks food (sensor)
- interaction with app (sensor)
- weight (self report)
- food intake (self report)

Start Intervention
- Tailored text message to become physically active
- Average 1.5×/day
- No intervention
- Average 3.5×/day

30 minutes after prompt
- Proximal Outcome
  - physical activity (steps taken)
  - weather
  - time of day
  - day of week

Following day
- Proximal Outcome
  - Did participant complete food log?

Daily Goals
- Rest (no goal) 1 of every 7 days on average
- Start Intervention
  - Remind participant to track food
  - Average 1× every other day

Distal Outcome
- Growth in step count

PI: P Klasnja
Location & Funding: Kaiser Permanente
Sense$^2$Stop

Sense$^2$Stop MRT for Stress Management in Newly Abstinent Smokers

Every minute of every day starting with quit date | For two hours after intervention is delivered | Measured via EMA and puffMarker over 10 days
---|---|---

Observations
- stress (via AutoSense sensor suite)
- motion (via accelerometer)
- smoking (via self report)

- Available? NO → No intervention
- Available? YES is stressed? NO → Remainder of times
  → No intervention
  → Average 1.5x/day
  → Prompt use of stress-management exercises

- Available? YES is stressed? YES → Average 1.5x/day
  → Prompt use of stress-management exercises
  → No intervention

PI: S Kumar
Location: Northwestern University, B. Spring, (P.I.)
Funding: NIBIB through funds provided by the trans-NIH Big Data to Knowledge initiative U54EB020404
Outline

• mHealth
• Micro-Randomized Trials
• Challenges to RL
• HeartSteps V1, V2
• A Butchered Bandit
• Open Question
Challenges to RL

• State is large yet partially observed: “unknown-unknowns”
  – Non-stationary reward
• Treatment actions that tend to have positive effects on immediate rewards but negative impact on future rewards via user habituation/burden.
• High noise within/between user
• Clinical populations (e.g., small numbers of users)
• Off-policy causal inference to further develop behavioral science
Outline

• mHealth
• Micro-Randomized Trials
• Challenges to RL
• HeartSteps V1, V2
• A Butchered Bandit
• Open Question
HeartSteps (PI Klasnja)

Goal: Develop an mobile activity coach for individuals who are at high risk of adverse cardiac events.

Three iterative studies:
- V1: 42 day micro-randomized pilot study with 37 sedentary individuals,
- V2: 90 day micro-randomized (partially via a bandit) study,
- V3: 365 day personalized study
Data from wearable devices that sense and provide treatments

- On each individual: $S_1 A_1 R_2, \ldots, S_t, A_t, R_{t+1}, \ldots$

- $t$: Decision point
- $S_t$: Context accrued after $t-1$ and up to/including decision point $t$ (high dimensional)
- $A_t$: Action at $t^{th}$ decision point (treatment)
- $R_{t+1}$: Reward (e.g., utility, cost) accrued after time $t$ and prior to time $t+1$
HeartSteps V1

Heartsteps MRT to Promote Physical Activity Among Sedentary People

Each day of study
Observations are continuous (except self report)
Randomizations to activity prompts occur 5x/day at likely times for increasing physical activity

Observations
- location (via GPS)
- weather (via internet)
- motion (via wrist band)
- usefulness of prompt (via user indication)
- self report of activity (via app in evenings)

Next 30 minutes after intervention is delivered
Measured via accelerometer throughout study

Proximal Outcome
physical activity (steps taken)

Distal Outcome
Overall activity in the 42-day study

Proximal Outcome
physical activity (steps taken)

Following day

Start Intervention
Prompt planning of next day’s activity

Start Intervention
Tailored prompt to become physically active

Intervention tailored on
- weather
- location
- time of day
- day of week

No intervention
Average of 1x every other day

Driving?
YES
NO

Walking?
YES
NO

Average 2x/day

R

Average 3x/day

No intervention

Stop

No intervention

1x/day in evening

5x/day

No intervention

Average of 1x every other day
Actions

• Tailored activity suggestion (provide yes/no)

• The set of actions may depend on the context, $S_t$
Some Results from HeartSteps V1

1) The tailored activity suggestion, as compared to no activity suggestion, indicates an initial increase in step count over succeeding 30 minutes by approximately 271 steps but by day 20 this increase is only approximately 65 steps.

2) Features that appear to predict succeeding 30 minute step count:
   1) Time in study, recent number of messages sent, location, variability in step count in 60 min window over previous 7 days, prior 30 min step count, total steps on prior day, current temperature
Some Results from HeartSteps V1

1) The tailored activity suggestion, as compared to no activity suggestion, indicates an initial increase in step count over succeeding 30 minutes by approximately 271 steps but by day 20 this increase is only approximately 65 steps.

2) Features that appear to interact with treatment on succeeding 30 minute step count:
   1) Time in study, recent number of messages sent, location, variability in step count in 60 min window over previous 7 days, prior 30 min step count, total steps on prior day
Goal

During HeartSteps V2 select binary treatment action--whether to provide tailored activity suggestion--online so as to maximize the sum of rewards for each user over the 90 day study (subject to constraints).

1. 5 decision points per day (according to user’s work schedule)
2. Reward is the 30 minute stepcount following each decision point $t$. 
Outline

• mHealth
• Micro-Randomized Trials
• Challenges to RL
• HeartSteps V1, V2
• A Butchered Bandit
• Open Question
Linear Thompson Sampling Bandit

1) Linear model for expected reward, e.g. \( r(S_t, A_t) = E[R_{t+1}|S_t, A_t] = \eta^T f(S_t, A_t) \)

2) Initialize \( \eta \) parameters in expected reward with a prior distribution.

3) Given \( S_t, A_t, R_{t+1} \) calculate posterior distribution of \( \eta \). Mean, covariance matrix of this posterior distribution is \( \eta_t, \Sigma_t \).

4) Given \( S_{t+1} \), the probability of selecting treatment, \( A_{t+1} = a \), is given by the posterior probability that treatment \( a \) has the highest expected reward.
Challenges

1) Noisy data

Our solution: Bandit algorithm

- Bandit algorithms learn faster than full RL algorithms
- The bandit acts as a regularizer (discount rate is 0): trade speed of learning (reduced variance) with bias
- Learn a low dimensional parameterization of the expected reward: linear model in context and treatment. $E[R_{t+1}|S_t, A_t] = \eta^T f(S_t, A_t)$
- Use a Gaussian prior on $\eta$ with mean, variance based on the data from Heartsteps V1 and a baseline no-treatment week of data from Heartsteps V2
Challenges

2) **Nonstationarity**: Over longer periods of time, the expected reward function will likely change.
   - Due to inability to fully sense, known and unknown, aspects of user’s current context.

Our solution:

- Promote continual exploration: Use a Gaussian process prior in the model for the expected reward, e.g.
  \[ E[R_{t+1}|S_t, A_t] = f(S_t, A_t)^T \eta_t \quad \text{where} \quad \eta_{t+1} = \eta_0 + \gamma(\eta_t - \eta_0) + \epsilon_t \]
  \[ \epsilon_t \sim N(0, 1 - \gamma^2) \]
Challenges

3) The immediate effects are primarily positive; the delayed effects are primarily negative. →
   • Algorithm may falsely learn that “always treat” is best, yet there are better policies.

Our Solution
• Add a low variance proxy for the “value function” to the current reward
Challenges

4) Need to ensure ability to conduct “off-policy learning” and causal inference after bandit study.

Our solution:

• Use explicit randomization to explore: Thompson Sampling Bandit
• Ensure the no-treatment selection probability lies in an interval bounded away from 0 and 1; here [.2, .9]
Challenges

5) Expected reward, $E[R_{t+1}|S_t, A_t]$ is likely a complex function of context, $S_t$

Proposed Solution: Center the treatment indicator by binary treatment selection probability, $\pi_t$
Proposal: For binary $A_t$:

replace $E[R_{t+1}|S_t = s, A_t = a] = \eta^T f(s, a)$

with

$E[R_{t+1}|S_t = s, A_t = a] = b_t(s) + \eta^T f(s)(a - \pi_t)$

where

$b_t(s)$ is an unspecified baseline (maybe nonlinear, non-stationary)

$\eta^T f(s)(a - \pi_t)$ is centered since $\pi_t$ is the probability of selecting treatment $A_t = a = 1$

In the Thompson-Sampling update of expected reward use a working (but likely mis-specified) approximation for $b_t(s)$. 
Median Regret
500 simulated users

Quartiles of Regret
500 simulated users

Context, s, is 3 dimensional
True $b_t(s)$ is nonlinear
Linear working model for $b_t(s)$

No proxy value
Only Gaussian prior
No Gaussian process prior
Outline

- mHealth
- Micro-Randomized Trials
- Challenges to RL
- HeartSteps V1, V2
- A Butchered Bandit
- Open Question
Long Term Goal: Continually learning mHealth App

The learning algorithm is part of the mHealth app

- Incorporate continual learning in the rollout of a mHealth application.
- Learning algorithm makes structural assumptions so as to trade bias and variance in learning
Optimality Criterion?

Multiple goals for learning algorithm

• Track best policy
• Intermittent off-policy inferences
  – causal not just correlational.
  – concern different outcomes than the reward.
  – use different structural assumptions
  – should be valid even if the structural assumptions made by the treatment learning algorithm are false.
Optimality Criterion?

Ideas:

• Minimize finite time $T$ regret subject to bounds on power to detect a particular causal effect at time $T$?
• Minimize finite time $T$ regret subject to bounds on exploration probability?
Collaborators!
Notes

Context \( s_t = [1, s_{t2}, s_{t3}, s_{t4}]^T \)

nonlinear generative model

\[
    r_t = I(a_t = 1) \theta_1^T s_t + I(a_t = 2) \theta_2^T s_t + [1.32, 1.36, .58][1, s_{t3}, s_{t4}]^T + 2I(||s_{t3}|| \leq 0.8)
\]

Analysis model used for action centering.

\[
    r_t = (I(a_t > 0) - \pi_t) [I(\bar{a}_t = 1) \theta_1^T s_t + I(\bar{a}_t = 2) \theta_2^T s_t] + \pi_t [I(\bar{a}_t = 1) \xi_1^T s_t + I(\bar{a}_t = 2) \xi_2^T s_t] + \eta^T s_t
\]

Total params: 20.

Analysis model for std Thompson.

\[
    r_t = I(a_t = 1) \theta_1^T s_t + I(a_t = 2) \theta_2^T s_t + \eta^T s_t
\]

Total params: 12.

\[
\text{Theta1} = [.116, -.275, -.233, .0425]
\]

\[
\text{Theta2} = [.116, .275, -.233, .0425].
\]
**Pis:** M Walton, S Murphy, and M Rabbi Shuvo  
**Location:** University of Michigan  
**Funding:** Michigan Institute for Data Science (PI S. Murphy), University of Michigan Injury Center (PI M. Walton)
Engagement with JOOL

MRT to Promote Engagement with Purpose-driven Well-being App

**Observations**
- activity (via accelerometer)
- surveys (via app)

**Each day of study**

- Select time of day:
  - 1/6 chance for each on weekdays
  - 1/5 chance on weekends
  - 8:30 am (weekday only)
  - 12:30 pm
  - 5:30 pm
  - 6:30 pm
  - 7:30 pm or
  - 8:30 pm

**Decision Rules for Recency**

<table>
<thead>
<tr>
<th>IF</th>
<th>THEN ASK</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 1 days since last engaged app</td>
<td>fewer than 3 days since last notification?</td>
</tr>
<tr>
<td>2-9 days since last engaged app</td>
<td>fewer than 2 days since last notification?</td>
</tr>
<tr>
<td>10-29 days since last engaged app</td>
<td>fewer than 6 days since last notification?</td>
</tr>
<tr>
<td>30+ days since last engaged app</td>
<td>fewer than 15 days since last notification?</td>
</tr>
</tbody>
</table>

**Within 24 hours of push notification**

**Proximal Outcome**
- Engaged app

**Distal Outcome**
- More consistent self-monitoring

**PI:** Victor Strecher, PhD, MPH, CEO of JOOL Health

**Location & Funding:** Ann Arbor, MI

**URL:** [https://www.joolhealth.com](https://www.joolhealth.com)