A BATCH, OFF-POLICY ACTOR-CRITIC ALGORITHM FOR OPTIMIZING MOBILE INTERVENTIONS

S.A. Murphy
INFORMS

MD2K
Center of Excellence for
Mobile Sensor
Data-to-Knowledge

The Methodology Center
advancing methods, improving health

INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

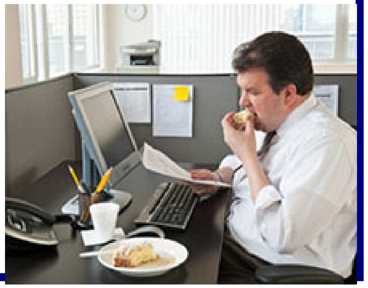
MICHIGAN
STATISTICS

Heart Steps

Smartphone based intervention for improving activity level

- Wearable band measures activity, phone sensors measure busyness of calendar, location, weather, .….

- In which contexts should smartphone ping and deliver activity ideas?



Data from wearable devices that sense and provide treatments

On each of *n* individuals:

$$O_1, A_1, Y_2, \ldots, O_K, A_T, Y_{T+1}$$

O_t: Observations at tth decision time (high dimensional)

A_t : Action at tth decision time (treatment)

Y_{t+1} : Proximal Response (aka: Reward, Utility, Cost)

Setup

1) Actions A_t
   1) Types of treatments that can be provided at decision time
   2) Whether to provide a treatment

2) Observations O_t
   1) Passively collected (location, social context, activity on device)
   2) Actively collected (answers to questions)

Setup

3) Response, Y_{t+1}, (reward or utility or cost)
    1) Proximal measure of clinical outcome
    2) Composite of several outcomes

4) State S_t
    1) Summary of $O_1, A_1, Y_2, \ldots, Y_t, O_t$ that permits the Markovian property; a modeling assumption.

Assumptions

Markovian Assumptions

$$P[S_{r+1} = s' | S_1, A_1, \ldots, S_t, A_t] = P[S_{t+1} = s' | S_t, A_t]$$

and

$$P[Y_{t+1} = y | S_1, A_1, \ldots, S_t, A_t] = P[Y_{t+1} = y | S_t, A_t]$$

Stationarity Assumptions

$$P[S_{t+1} = s' | S_t = s, A_t = a] = p(s' | s, a)$$

and

$$E[Y_{t+1} | S_t = s, A_t = a] = r(s, a)$$

Setup

Unknown transition probabilities

$$P[S_{t+1} = s' | S_t = s, A_t = a] \quad = \quad p(s'|s,a)$$

Unknown reward function

$$E[Y_{t+1} | S_t = s, A_t = a] \quad = \quad r(s,a)$$

Known distribution of actions in data

$$P[A_t = a | S_t = s] = \mu(a|s)$$

Stochastic Treatment Policy

We aim to use the data to construct a parameterized policy, $\pi_\theta(a|s)$ with probabilities bounded away from 0 and 1.

- Variation in actions can help retard habituation and maintain engagement.
- Parameterized $\pi_\theta(a|s)$ can be interpreted/vetted by domain experts

Optimality Criterion (to maximize)

Average Reward, η_θ, for policy π_θ:

$$\begin{aligned} \eta_\theta &= \lim_{K\to\infty} \frac{1}{K} E_\theta \left[\sum_{t=0}^{K-1} Y_{t+1} \middle| S_0 = s \right] \\ &= \sum_s d_\theta(s) \sum_a \pi_\theta(a|s) r(s,a) \end{aligned}$$

E_θ denotes expectation under the stationary distribution, d_θ, associated with π_θ.

Background: Differential Value

V_θ is the Differential Value

$$V_\theta(s) = \lim_{T\to\infty} E_\theta\left[\sum_{t=0}^{T}\Big(Y_{t+1} - \eta_\theta\Big)\Big|S_0 = s\right].$$

$V_\theta(s)$ - $V_\theta(s')$ reflects the difference in sum of centered responses accrued when starting in state *s* as opposed to state *s*′.

(η_θ is the average reward)

Bellman Equation

$$E_\theta \left[Y_{t+1} - \eta_\theta + V_\theta(S_{t+1}) - V_\theta(S_t) \middle| S_t \right] = 0$$

$S_t, A_t, Y_{t+1}, S_{t+1}$

Background

Bellman's equation implies that

$$E\left[\frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)}\Big(Y_{t+1}-\eta+V(S_{t+1})-V(S_t)\Big)\begin{pmatrix}1\\ f(S_t)\end{pmatrix}\right]$$

will be, for all t and for any vector, $f(S_t)$, of appropriately integrable functions, equal to 0 if $\eta=\eta_\theta,\ V=V_\theta$

E denotes averaging over data generating distribution.

Estimating Function

- Construct an approximation for, $V_\theta(s)$:
$f(s)^T v_\theta$ where *f(s)* is a *p* by *1* vector of basis functions evaluated at *s* (p is large)

- Idea is to solve

$$\mathbb{P}_n\left[\sum_{t=1}^{T}\frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)}\Big(Y_{t+1}-\eta+f(S_{t+1})^T v-f(S_t)^T v\Big)\begin{pmatrix}1\\ f(S_t)\end{pmatrix}\right]$$

=0 for $\hat{\eta}_\theta$, $\hat{v}_\theta$

Overview of Algorithm

- The resulting η and v are functions of θ, denote by $\hat{\eta}_\theta, \ \hat{v}_\theta$
  - $\hat{\eta}_\theta, \ \hat{v}_\theta$ are the output of the Critic
- The Actor maximizes $\hat{\eta}_\theta$ over θ to obtain $\hat{\theta}$.
  - this will require repeated calls to the Critic
  - $\hat{\theta}$ is the output of the Actor

CRITIC

Write

$$\mathbb{P}_n\left[\sum_{t=1}^{T}\frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)}\Big(Y_{t+1}-\eta+f(S_{t+1})^T v-f(S_t)^T v\Big)\begin{pmatrix}1\\ f(S_t)\end{pmatrix}\right]$$

$$=\hat{A}_\theta\begin{pmatrix}\eta\\ v\end{pmatrix}-\hat{b}_\theta$$

The critic minimizes

$$||\hat{A}_\theta\begin{pmatrix}\eta\\ v\end{pmatrix}-\hat{b}_\theta||^2+\lambda_c||v||^2$$

to obtain

$$\hat{\eta}_\theta,\ \ \hat{v}_\theta$$

Overview of Actor

- The objective function for the actor is given by

$$\hat{\eta}_\theta = \mathbb{P}_n \left[\sum_{t=1}^{T} \frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)} \Big(Y_{t+1} + f(S_{t+1})^T \hat{v}_\theta - f(S_t)^T \hat{v}_\theta \Big) \right]$$

- Stochastic policy, π_θ

Binary action:

$$\pi_\theta(a|s) \quad = \quad \frac{e^{\theta^T g(s) a}}{1 + e^{\theta^T g(s)}}; \;\; a \in \{0, 1\}$$

Overview of Actor

The policy, π_θ should yield probabilities bounded away from 0, 1.

Chance constraint on θ:

$$\min_a P^* \left[p_0 \leq \pi_\theta(a|S) \leq 1 - p_0\right] \geq 1 - \alpha$$

given α, p_0 and P^*, a reference distribution over states, S.

This constraint is nonconvex; we relax via Markov inequality.

ACTOR

- The actor obtains $\hat{\theta}$ by maximizing

$$\hat{\eta}_\theta = \mathbb{P}_n \left[\sum_{t=1}^{T} \frac{\pi_\theta(A_t|S_t)}{\mu(A_t|S_t)} \Big(Y_{t+1} + f(S_{t+1})^T \hat{v}_\theta - f(S_t)^T \hat{v}_\theta \Big) \right]$$

subject to the constraint, $\theta^T \Sigma_g \theta \leq k_{max}$

$$\Sigma_g = T^{-1} \sum_{t=1}^{T} E^* \left[g(S_t) g(S_t)^T \right]$$

BASICS Mobile

- Smartphone-based intervention to curb heavy drinking and smoking in college students
  - 14 day study
  - Self-report 3x/day (morning, afternoon, evening)
  - Intervention 2x/day (afternoon, evening)
    - Mindfulness-based intervention (A_t=*1*) vs general health information (A_t=*0*)
- <u>Question</u>: Should a mindfulness-based intervention (vs general health info) be provided when there is an increase in need to self-regulate?

BASICS Mobile

- n subjects = 27, T decision points = 28
- Availability: To be available to receive a treatment, the student must complete self-report questions ($I_t = 1$). If the student is available then the student is provided a treatment with probability 2/3.
- Reward, Y_{t+1}, is (-)smoking rate

BASICS Mobile

- S_t is 8 dimensional composed of 5 discrete and 3 continuous valued features.
- Differential value approximated by B-splines and two way products of B-splines constructed from entries in S_t.
- Parameterized policy:

$$\pi_\theta(1|s) \quad = \quad I_t \frac{e^{\theta_0 + \theta_1 g_1 + \theta_2 g_2}}{1 + e^{\theta_0 + \theta_1 g_1 + \theta_2 g_2}}$$

BASICS Mobile

- g_1 is indicator for an increase in self-control demands (1 if yes, 0 if no)
- g_2 is indicator for no burden (1 if yes, 0 if no)
- $\hat{\theta}_0 = .74,\ \hat{\theta}_1 = -.95,\ \hat{\theta}_2 = 2.26 \rightarrow$ An available student with no increase in self-control demands and who is not indicating burden is recommended treatment with probability 0.85

$$\pi_\theta(1|s) \quad = \quad I_t \frac{e^{\theta_0+\theta_1 g_1+\theta_2 g_2}}{1+e^{\theta_0+\theta_1 g_1+\theta_2 g_2}}$$

Challenges

- Average Reward versus Discounted Reward?
  - Burden → disengagement raises the need to pay attention to future.

- This policy will function as a “warm-start” in an online algorithm.

Collaborators