Continual Learning in Sequential Decision Making

S.A. Murphy
In Honor of A. Wald, JSM
Father of
Sequential Decision Making

Find the decision making policy, $\pi$ that at each time $n$, inputs present state $S_n$ and outputs one of three actions:

$$\{p_0 & stop, p_1 & stop, continue\}$$

This policy should minimize

$$g\left(w_0 P_{p_0,\pi}[\pi(S_N) = p_1 & stop] + cE_{p_0,\pi}[N]\right)$$

$$+(1 - g)(w_1 P_{p_1,\pi}[\pi(S_N) = p_0 & stop] + cE_{p_1,\pi}[N])$$

where $g$ is a probability,

$w_0, w_1, c$ are positive weights,

$N$ is stopping time.
The Dream!

“Continually Learning Mobile Health Intervention”

• Help maintain healthy behaviors
• Help you achieve your health goals
  – Help you better trade off long term benefit with short term momentary pleasure
• The ideal mHealth intervention
  – will be there when you need it and will not intrude when you don’t need it.
  – will adjust to unanticipated life challenges
Ideas

- **Real-Time Treatment Policies**

- “Personalized Continually Learning Treatment Policy”
  - Warm-Start Policy
  - Online Training Algorithm

- **Clinical Trial**
Treatment Policies are individualized treatments delivered when-ever and where-ever the individual needs help. They are delivered by a wearable device (e.g., smartphone).

\[ O_1, A_1, Y_2, \ldots, O_t, A_t, Y_{t+1}, \ldots \]

\( O_t \): Observations at \( t^{th} \) decision time (high dimensional)

\( A_t \): Action at \( t^{th} \) decision time (treatment)

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

HeartSteps Activity Coach

- Wearable band measures activity and sleep quality; phone sensors measure busyness of calendar, location, weather; utility ratings

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

1) Decision Times (Times at which a treatment can be provided.)
   1) Regular intervals in time (e.g. every 10 minutes)
   2) At user demand

HeartSteps: Approximately every 2-2.5 hours
Examples

2) Observations $O_t$
   1) Passively collected (via sensors)
   2) Actively collected (via self-report)

HeartSteps observations include activity recognition, location, calendar, step count, usefulness ratings, adherence, self-efficacy…….
Examples

3) Actions $A_t$
   1) Treatments that can be provided at decision time
   2) Whether to provide a treatment

HeartSteps: Activity Recommendations
Examples

4) Proximal Response (reward) $Y_{t+1}$

HeartSteps: Activity (step count) over next 60 minutes.
A treatment policy is a mapping from the information on an individual at each decision time $t$, say $S_t$, to a probability distribution on the action space; this will be a distribution for $A_t$.

Example of a parameterized policy for a binary action,

$$P[A_t = 1|S_t = s] = \pi_\theta(1|s) = \frac{e^{\theta^T g(s)}}{1 + e^{\theta^T g(s)}}$$

OR

$$\pi_\theta(1|s) = 1 \text{ if } g(s) > \theta, 0 \text{ otherwise}$$
Desired Treatment Policy

- Individuals recovering from a heart attack
- Develop a coach app on phone. Policy underlying coach app will need to trade off
  - short term improvement/maintenance of physical activity level
  with
  - burden and engagement
to maintain activity over the long term.
Current State of the Art

• Expert-derived treatment policy: Scientists formulate the policy using behavioral/social theories, clinical experience, observational data analyses.

• Concerns that different individuals may do better with different policies (treatment effect heterogeneity)
  • Between individuals
  • Temporal, within individuals
Desired Online Algorithm

- How to best detect and use treatment effect heterogeneity to improve policy?
- Online algorithm should detect that an individual is responding differently from the “average person” to treatment action.
- Online algorithm should detect that an individual is responding differently at later times to a treatment action than at early times.
Summary of Goal

1) Take advantage of existing approach of constructing a policy based on theory, clinical experience—e.g the expert-derived policy

2) Online algorithm should alter treatment policy to improve policy if this will improve average reward.
Ideas

• Real-Time Treatment Policies & Example

• “Personalized Continually Learning Treatment Policy”
  – Warm-Start Policy
  – Online Training Algorithm

• Clinical Trial
Personalized Continually Learning Treatment Policy

1) Two parts:

1) Warm-Start Policy

2) Online Training Algorithm
Warm-Start Policy

- Warm-start policy is stochastic; Warm-start policy is based on expert-derived policy or treatment policy developed via micro-randomized trial.
- Variation in actions can help retard habituation and maintain engagement.
- Parameterized policy can be interpreted/vetted by domain experts.
Example Warm-Start

1) Clinician’s policy: If calendar is not busy and if risk of burden is low, then provide an activity suggestion.

2) Warm-Start policy, $\pi_\theta(a|s) = P[A = a|S = s]$: 

$$
\pi_\theta(1|S_t) = \frac{e^{\theta_0+\theta_1C_t+\theta_2B_t+\theta_3^T g(S_t)}}{1 + e^{\theta_0+\theta_1C_t+\theta_2B_t+\theta_3^T g(S_t)}}
$$

$C_t, B_t \in \{0, 1\}$, $(C_t, B_t) \subset S_t$
Example Warm-Start

Warm-Start policy:

\[
\pi_{\theta}(1|S_t) = \frac{e^{\theta_0 + \theta_1 C_t + \theta_2 B_t + \theta_3^T g(S_t)}}{1 + e^{\theta_0 + \theta_1 C_t + \theta_2 B_t + \theta_3^T g(S_t)}}
\]

How to initialize \(\theta_0, \ldots, \theta_3\)? Set \(\theta_3 = 0\); set remaining \(\theta_0, \theta_1, \theta_2\) so that

| \(C_t\) | \(\pi_{\theta}(1|S_t)\) | 1 | 0 |
|---|---|---|---|
| 1 | .2 | .4 |
| 0 | .4 | \(\sim .84\) |
Personalized Continuously Learning
Treatment Policy

1) Two parts:

1) Warm-Start Policy

2) Online Training Algorithm

Policy is “control law;” online training algorithm is “adaptive control rule” (TL Lai, 2001)
Online Training Algorithm

Training algorithm is used to improve the Warm-Start policy for each person separately; algorithm is a stochastic gradient ascent algorithm.

1) After each decision time and on each individual separately, the training algorithm is used to update the $\theta$ parameters of the current policy, $\pi_\theta$.

2) At the next decision time, treatment, $A_t$, is assigned using the individual’s updated policy.
Background

1) On each individual:

\[ S_1, A_1, Y_2, \ldots, S_T, A_T, Y_{T+1} \]

2) Probabilistic Model: Markov Decision Process

3) Optimality criterion for policy: Average Reward
Markov Decision Process

Markovian Assumptions

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

and

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

Stationarity Assumptions

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

and

\[
E[Y_{t+1}|S_t = s, A_t = a] = r(s, a)
\]
Optimality Criterion for Policy

Average Reward, $\eta_\theta$, for policy $\pi_\theta$:

$$
\eta_\theta = \lim_{T \to \infty} \frac{1}{T} E_\theta \left[ \sum_{t=0}^{T-1} Y_{t+1} \mid S_0 = s \right]
= \sum_s d_\theta(s) \sum_a \pi_\theta(a \mid s) r(s, a)
$$

$E_\theta$ denotes expectation under the stationary distribution, $d_\theta$, associated with $\pi_\theta$. 
1) Each individual follows a possibly different Markovian process. Thus the policy achieving $\max_\theta \eta_\theta$ may differ by individual

2) The Markovian processes are sufficiently similar so that starting from the warm-start policy, a policy achieving $\max_\theta \eta_\theta$ for the individual can be learned quickly.
**Nuisance Parameter: Differential Value**

$V_\theta$ is the Differential Value function

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

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

($\eta_\theta$ is the average reward)
Background: Bellman Equation

Oracle Temporal Difference:

$$\delta_t = Y_{t+1} - \eta_\theta + V_\theta(S_{t+1}) - V_\theta(S_t)$$

Bellman Equation:

$$E_\theta \left[ \delta_t \middle| S_t \right] = 0$$

$S_t, A_t, Y_{t+1}, S_{t+1}$
Bellman’s equation implies that

\[ E_{\theta} \left[ \left( Y_{t+1} - \eta + V(S_{t+1}) - V(S_t) \right) \left( \frac{1}{f(S_t)} \right) \right] \]

will be, for all \( t \), for any vector, \( f(.) \), of appropriately integrable functions, equal to 0 if \( \eta = \eta_\theta, \ V = V_\theta \)
Background for Online Training Algorithm

Policy Gradient Theorem:

\[ \nabla_\theta \eta_\theta = E_\theta [\delta_t \nabla_\theta \log \pi_\theta(A_t|S_t)] \]

where

\[ \delta_t = Y_{t+1} - \eta_\theta + V_\theta(S_{t+1}) - V_\theta(S_t) \]
Stochastic Gradient Algorithm

At the $t^{th}$ decision time, draw action from current policy $\pi_{\theta_t}$ then update estimators of average reward $\hat{\eta}_{\theta_t}$ and differential value, $V_{\theta_t}$. Use these estimators, $\hat{\eta}_t$, $\hat{V}_t$ to form the TD error,

$$\hat{\delta}_t = Y_{t+1} - \hat{\eta}_t + \hat{V}_t(S_{t+1}) - \hat{V}_t(S_t)$$

Move $\hat{\theta}_t$ in direction of empirical gradient,

$$\nabla_\theta \eta(\pi_{\theta}) = \hat{\delta}_t \nabla_\theta \log \pi_{\theta}(A_t|S_t)|_{\theta=\theta_t}$$

to obtain $\hat{\theta}_{t+1}$
Estimating Equations for $\eta_{t}$, $V_{t}$

are based on the fact that $\eta = \eta_{t}$, $V = V_{t}$ is a solution to

$$0 = E_{\theta_{t}} \left[ \left( Y_{t+1} - \eta + V(S_{t+1}) - V(S_{t}) \right) \left( \frac{1}{f(S_{t})} \right) \right]$$

for any appropriately integrable function $z(.)$.

The idea: parameterize $V_{t}$ then solve empirical versions of these equations.
Estimating Equations for $\eta_{\theta_t}$, $V_{\theta_t}$

Parameterize $V_{\theta_t}$ as $v^T f(s)$

Intuition is to solve

$$0 = \sum_{j=0}^{t} \left[ (Y_{j+1} - \eta + v^T f(S_{j+1}) - v^T f(S_j)) \begin{pmatrix} 1 \\ f(S_t) \end{pmatrix} \right]$$

Use first (or second) order incremental solutions.
A Very Little Review!

An exact incremental solution to

\[ 0 = \sum_{j=0}^{t} [Y_{j+1} - \mu] \]

is

\[ \hat{\mu}_{j+1} = \hat{\mu}_j + \sigma_j \left[ Y_{j+1} - \hat{\mu}_j \right] \]

where

\[ \sigma_j = \left( \frac{1}{j+1} \right) \]
Basic Version of Online Training Algorithm

1. **Initialize** Stepsize \( \{\sigma_{\eta,t}, \sigma_{V,t}, \sigma_{\pi,t}\}_{t \geq 0} \) all positive and \( \hat{\theta}_0, \hat{v}_0 \).
   Observe \( S_0 \)

2. **for** \( t \geq 0 \) **do**

3. Draw \( A_t \) from \( \pi_{\hat{\theta}_t}(\cdot | s) \)

4. Observe \( S_{t+1}, Y_{t+1} \).

5. Set \( \hat{\delta}_{t+1} = Y_{t+1} - \hat{\eta}_t + \hat{v}_t^T f(S_{t+1}) - \hat{v}_t^T f(S_t) \)

6. \( \hat{\eta}_{t+1} = \hat{\eta}_t + \sigma_{\eta,t} \hat{\delta}_{t+1} \)

7. \( \hat{v}_{t+1} = \hat{v}_t + \sigma_{V,t} \hat{\delta}_{t+1} f(S_t) \)

8. \( \hat{\theta}_{t+1} = \hat{\theta}_t + \sigma_{\pi,t} \hat{\delta}_{t+1} \nabla \log \pi_{\hat{\theta}_t}(A_t|S_t) \)

9. \( t \leftarrow t + 1 \)

10. **end**
Ideas

• Real-Time Treatment Policies & Example

• “Personalized Continually Learning Treatment Policy”
  – Warm-Start Policy
  – Online Training Algorithm

• Clinical Trial
Proposal

A Two-group Randomized Trial of

“Personalized Continually Learning Treatment Policy” versus

“Standard Care”
Note

Personalized policy must be operationalized, e.g., “locked down” to enhance replicability. Prior to Trial:

1) Warm start policy is pre-specified.

2) Training algorithm must be completely pre-specified
Standard RCT

Standard data analyses can be used to compare Personalized Policy group versus Standard Care group.

1) Survival outcomes, longitudinal outcomes

2) Subpopulation (moderator) analyses. E.g. estimate treatment effect for more severely ill people
Interesting Challenges

• At the end of the trial different people will have different policies, e.g. different $\theta$’s.
  – Testing if the variation between people in their policies at the end of the study is greater than that due to noise.

• Provide error bounds on how much the personalized policy will be an improvement over the warm start—at least on average.
Interesting Challenges

• Methods for using trial data to estimate the average reward for a variety of policies.

• Methods for using trial data to construct a new single “optimal” policy. This new policy could be used in forming a warm-start for a future trial.
Interesting Challenges

• Optimality criterion for online training algorithm: Minimize regret:

\[ \lim_{T \to \infty} E_{online\ alg} \left[ \sum_{t=1}^{T} \left( Y_{t+1} - \eta \theta^* \right) \right] \]

• Or should the regret be an averaged regret?

\[ \lim_{T \to \infty} E \left[ E_{online\ alg} \left[ \sum_{t=1}^{T} \left( Y_{t+1} - \eta \theta^* \right) \right] \right] \]
Interesting Challenges

• How to formulate “Bayesian” heuristics to design the stochastic gradient algorithm and facilitate the balance between using information learned so far with the need to continue exploration?

• Combining information across individuals?
Challenges

• Any method should provide confidence intervals/permit scientists to test hypotheses.

• How to reduce the amount of self-report data (How might you do this?)

• How to accommodate/utilize the vast amount of missing data, some of which will be informative……
Collaborators