Precision Medicine: Lecture 11
Markov Decision Processes

Michael R. Kosorok,
Nikki L. B. Freeman and Owen E. Leete

Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina at Chapel Hill

Fall, 2021

Outline

Markov Decision Processes

DTRs over indefinite time horizons

DTRs on an infinite time horizon

Markov Decision Processes

A **reinforcement learning (RL) task** that satisfies the **Markov property** is a **Markov decision process (MDP)**.

Reinforcement Learning Review

- Reinforcement learning problems: Map states to actions in order to maximize a reward.
- Ingredients
  - An **agent** that interacts with its environment to try achieve its goal. There is uncertainty about the environment and actions can affect the future state of the environment.
  - A **policy** is a map from states to actions.
  - A **reward** is the goal in a RL problem.
  - A **value function** gives us the value of the current state, the total expected reward over future states if you start in the current state.

Reinforcement Learning Review (cont.)

- Precision medicine can be viewed through the lens of RL.
- The tools of RL can be applied to precision medicine problems.
- We have previously discussed some of these methods.
  - Q-learning
  - A-learning
  - BOWL

Markov Property

- The state in a RL problem is the information available to the agent.
- Let $s_t \in \mathcal{S}$ denote the state at time t.
- Let $a_t \in \mathcal{A}$ denote the action at time t.
- Markov property:

$$P(s_{t+1}|s_t, \ldots, s_0, a_t, \ldots, a_0) = P(s_{t+1}|s_t, a_t)$$

- The current state and action provide all the needed information to predict what the next state will be.

Markov Decision Process

A reinforcement learning (RL) task that satisfies the Markov property is a **Markov decision process (MDP)**.

- If the state and action spaces are finite, then it is called a finite MDP.
- Let $r_t \in \mathbb{R}$ denote the reward at time t. The reward at t depends on the state at t.
- The goal is to find a policy $\Pi : \mathcal{S} \rightarrow \mathcal{A}$
  - Finite horizon: maximize $E[\sum_{t=0}^{T} r_t | \Pi, s_0]$
  - Infinite horizon: maximize $E[\sum_{t=0}^{\infty} \gamma^t r_t | \Pi, s_0]$, where $0 < \gamma < 1$ is a discount factor.

MDPs in Precision Medicine

- MDPs are a way to frame precision medicine problems.
- The two methods we will consider in this lecture use MDPs to consider estimating dynamic treatment regimes in more general settings than we have seen so far.
- Ertefaie and Strawderman (2018) consider an indefinite time period setting.
- Luckett et al. (2020) consider the infinite time horizon with a clinical application to mHealth and type 1 diabetes.

Outline

Markov Decision Processes

DTRs over indefinite time horizons

DTRs on an infinite time horizon

The indefinite time period setting

- So far we have considered a wide range of methods for estimating dynamic treatment regimes.
  - Q-learning, OWL, etc.
- Each of these methods rely on backward induction to estimate the optimal DTR.
- These methods cannot extrapolate beyond the time horizon in the observed data.
- For diseases such as cancer, this might be ok. For chronic diseases, this limitation is a departure from reality. (Note that some cancers are now being treated as chronic diseases, too).
- Ertefaie and Strawderman (2018) propose a method for estimating optimal DTRs over an indefinite time horizon.

Notation

- S_t, $t = 0, 1, \ldots$ denotes a summary of a subject's health history through time t.
  - S_t takes values in $\mathcal{S}^\star = \mathcal{S} \cup \{c\}$, where $\mathcal{S} \cap \{c\} = \emptyset$.
  - c represents an absorbing state, such as death.
  - For convenience, assume that $\mathcal{S}$ is finite.
- A_t is the treatment assigned at t after measuring S_t.
  - Let $\mathcal{A}$ be a finite set containing m possible treatment actions.
  - For $t \geq 0$ and $x \in \mathcal{S}$, suppose that A_t takes values in $\mathcal{A}_s$ where $\mathcal{A}_s$ is a subset of $\mathcal{A}$ consisting of $0 < m_s \leq m$ treatments and $\cup_{s \in \mathcal{S}} \mathcal{A}_s = \mathcal{A}$.
  - For any t such that $S_t = c$, let $A_t \in \mathcal{A}_c = \{u\}$, where u denotes undefined.

Notation (cont.)

- Let $\mathcal{H} = (S_0, A_0, \ldots, S_{\widetilde{T}-1}, A_{\widetilde{T}-1}, S_{\widetilde{T}})$ be the data on a subject observed from $t = 0$ until death, $\widetilde{T} = \inf\{t \geq 1 : S_t = c\}$.

- Let $\bar{S}_t$ and $\bar{A}_t$ denote the respective process histories up to and including time t.

Assumption 1: Time-homogeneous Markov behavior

Time-homogeneous Markov behavior assumption: The data-generating process satisfies

$$S_{t+1} \perp \{\bar{S}_{t-1}, \bar{A}_{t-1}\} | S_t, A_t$$

for $t \geq 1$. In addition, for $t \geq 1$, the transition probabilities satisfy

$$\begin{aligned} \text{pr}(S_{t+1} = s' | S_t = s, A_t = a) &= \text{pr}(S_1 = s' | S_0 = s, A_0 = a) \\ &= p(s' | s, a) > 0 \end{aligned}$$

for $s' \in \mathcal{S}^\star$, $s \in S$, $a \in \mathcal{A}_s$; moreover

$$\text{pr}(S_{t+1} = c, A_{t+1} = u | S_t = c, A_t = u) = 1.$$

Assumption 2: Positivity

Positivity assumpmtion: Let $p_{A_t|\bar{S}_t,\bar{A}_{t-1}}(a|\bar{s},\bar{a})$ be the probability of receiving treatment a given $\bar{S}_t = \bar{s}$ and $\bar{A}_{t-1} = \bar{a}$. Then

$$p_{A_t|\bar{S}_t,\bar{A}_{t-1}}(a|\bar{s},\bar{a}) > 0$$

for each action $a \in \mathcal{A}_{S_t}$ and for each possible value of $\bar{s}$ and of $\bar{a}$.

The reward function

- For $t \geq 0$ define the reward value $R_{t+1} = r(S_t, A_t, S_{t+1})$.
  - $r(s, a, s')$ is a known function for $(s, s') \in \mathcal{S}^\star$ and $a \in \mathcal{A}_s$.
  - Larger values are better.
  - For $s \in \mathcal{S}$ and $a \in \mathcal{A}_s$, assume that $|r(s, a, s')| \leq \bar{r}$ for $s' \in \mathcal{S}^\star$ and for some fixed $0 < \bar{r} < \infty$.
  - We assume $r(c, u, c) = 0$.
- Let $\gamma > 0$ be a fixed discount factor. The cumulative discounted reward beyond time t may be written as

$$\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}.$$

Infinite-horizon action-value

- Given a deterministic DTR $\pi(s)$ for $s \in \mathcal{S}$, define the infinite-horizon action-value function

$$Q_t^{\pi}(s,a) = E(\sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k} | S_t = s, A_t = a),$$

  where the conditional expectation assumes that all treatment assignments at times $t+k$, for $k \geq 1$, are assigned according to π.

- By definition of the reward function $Q_t^{\pi}(c,a) = Q_t^{\pi}(c,u) = 0$.

Characterization of optimal DTR

- The optimal DTR is characterized as the regime that, if implemented, would lead to an optimal action-value function for each pair (s, a) when $s \in \mathcal{S}$.
- Under assumption 1, $Q_t^{\pi}(s, a)$ does not depend on t and is finite whenever $\gamma < 1$.
- For $\gamma < 1$, define the optimal action-value function as $Q^*(s, a) = \max_{\pi} Q_t^{\pi}(s, a)$, which can be shown to satisfy

$$Q^*(s, a) = E\{R_{t+1} + \gamma \max_{a' \in \mathcal{A}_{S_{t+1}}} Q^*(S_{t+1}, a') | S_t = s, A_t = a\} \quad (1)$$

  for $s \in \mathcal{S}$ and $a \in \mathcal{A}_s$.

- The optimal treatment regime: $\pi^*(s) = \text{argmax}_{a \in \mathcal{A}_s} Q^*(s, a)$.

Estimating $Q^*(s, a)$

- To estimate the optimal DTR, we need to estimate the optimal action-value function (1).
- A way forward is to model $Q^*(s, a)$ as a linear function, say

$$Q^*(s, a) = Q_{\theta_0}(s, a) = \theta_0^\intercal \varphi(s, a)$$

where

  - $Q_\theta(s, a) = \theta^\intercal \varphi(s, a)$,
  - θ is a p-dimensional parameter, and
  - $\varphi(s, a)$ a p-vector of features summarizing (s, a).

- To ensure that $Q_{\theta_0}(c, a) = 0$, we require $\varphi(c, a) = 0$.

Estimating $Q^*(s, a)$ (cont.)

- Under these assumptions, $\pi^*(s) = \text{argmax}_{a \in \mathcal{A}_s} Q_{\theta_0}(s, a)$ for $s \in \mathcal{S}$.

- From here, estimating equations can be devised for subjects who are followed from $t = 0$ to $\widetilde{T}$ and subjects who are only followed until $\breve{T} = \min(T, \widetilde{T})$ where $0 < T < \infty$ corresponds to the end of follow-up (max number of possible decision points).

Estimating equation

- First, we consider an estimating equation for θ_0 for a subject followed from time $t = 0$ to $\widetilde{T}$.
- Define

$$\delta_{t+1}(\theta) = R_{t+1} - Q_\theta(S_t, A_t) + \gamma \max_{a' \in \mathcal{A}_{S_{t+1}}} Q_\theta(S_{t+1}, a')$$

  as the temporal difference error.
- In view of (1), we have

$$E\{\delta_{t+1}(\theta_0)|S_t = s, A_t = a\} = 0 \quad (t \geq 0).$$

- Hence, under our assumptions

$$E\left\{ \sum_{t=1}^{\widetilde{T}-1} \delta_{t+1}(\theta_0)\varphi(S_t, A_t) \right\} = 0$$

  where $\varphi(s, a) = \nabla_\theta Q_\theta(s, a)$.
- This relationship suggests an estimating equation for θ_0.

Estimating equation (cont.)

- With the intuition of the EE on the previous slide, now we consider an estimating equation for a subject that can only be followed until $\breve{T} = \min(T, \widetilde{T})$.
- Because T is fixed, we have $D(\theta_0) = 0$, where

$$D(\theta) = E\left\{ \sum_{t=0}^{T-1} I(\widetilde{T} > t)\delta_{t+1}(\theta)\varphi(S_t, A_t) \right\}$$
$$= E\left\{ \sum_{t=0}^{T-1} \delta_{t+1}(\theta)\varphi(S_t, A_t) \right\}.$$

where the second equality results from $I(\widetilde{T} > t)\delta_{t+1}(\theta) = \delta_{t+1}(\theta)$ for $t \geq 0$.

Estimating equation (cont.)

- Thus

$$\hat{D}(\theta) = \mathbb{P}_n\left\{ \sum_{t=0}^{\check{T}-1} \delta_{t+1}(\theta)\varphi(S_t, A_t) \right\}$$
$$= \mathbb{P}_n\left\{ \sum_{t=0}^{T-1} \delta_{t+1}(\theta)\varphi(S_t, A_t) \right\} \quad (2)$$

  is an unbiased estimating function for θ_0, where the observed data consists of the trajectories $\mathcal{H}_i = (S_{i,0}, A_{i,0}, \ldots, S_{i,\check{T}_i-1}, A_{i,\check{T}_i-1}, S_{\check{T}_i})$.

- If $\hat{D}(\theta) = 0$ has a unique solution $\hat{\theta}$, then $\hat{Q}^*(s,a) = Q_{\hat{\theta}}(s,a)$ and the optimal DTR can be estimated by $\hat{\pi}^*(s) = \text{argmax}_{a \in \mathcal{A}_s} Q_{\hat{\theta}}(s,a)$.

Estimation

- The estimating function (2) is a continuous, piecewise-linear function in θ that is not differentiable everywhere.
- Under regularity conditions, any solution $\hat{\theta}$ can equivalently be defined as a minimizer of

$$\hat{M}(\theta) = \hat{D}(\theta)^{\intercal} \hat{W}^{-1} \hat{D}(\theta)$$

where

$$\hat{W} = \mathbb{P}_n \left\{ \sum_{t=1}^{\check{T}-1} \varphi(S_t, A_t)^{\otimes 2} \right\} = \mathbb{P}_n \left\{ \sum_{t=1}^{T-1} \varphi(S_t, A_t)^{\otimes 2} \right\}$$

- Minimizing $\hat{M}(\theta)$ is equivalent to solving a nonlinear least-squares problem. A Gauss-Newton algorithm can be used to solve.

Consistency and asymptotic normality of $\hat{\theta}$

Letting $\hat{M}(\theta) = \hat{D}(\theta)^\intercal \hat{W}^{-1} \hat{D}(\theta)$ and $M(\theta) = D(\theta)^\intercal W^{-1} D(\theta)$ and under a few more assumptions, it can be shown that: For any sequence of estimators $\hat{\theta}$ with $\hat{M}(\hat{\theta}) = \min_{\theta \in \Theta} \hat{M}(\theta) + o_p(1)$, we have

$$\hat{\theta} \xrightarrow{p} \theta_0$$

and

$$n^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Gamma(\gamma)\Sigma\Gamma(\gamma)^\intercal),$$

where

$\Gamma(\gamma) = \{\dot{D}_{\theta_0}(\gamma)\}^{-1}$,

$\dot{D}_{\theta_0}(\gamma) = E\left(\sum_{t=1}^{T-1} \varphi(S_t, A_t)[\gamma I_{\mathcal{C}\{\pi^*(S_{t+1})\}=1} \varphi\{S_{t+1}, \pi^*(S_{t+1})\} - \varphi(S_t, A_t)]^\intercal \right)$, and

$\Sigma = E\left[\left\{ \sum_{t=0}^{T-1} \varphi(S_t, A_t)\delta_{t+1}(\theta_0) \right\}^{\otimes 2} \right]$.

Outline

Markov Decision Processes

DTRs over indefinite time horizons

DTRs on an infinite time horizon

Introduction

- Luckett et al. (2020) develop estimation techniques (using data collected with mobile devices) for dynamic treatment regimes (which can be implemented as mHealth interventions)

- Motivating example: type 1 diabetes

  - Understand type 1 diabetes (T1D) and how it is managed

  - Develop tailored mHealth interventions for T1D management

What is T1D?

- Autoimmune disease wherein the pancreas produces insufficient insulin
- Daily regimen of monitoring glucose and replacing insulin
- Hypo- and hyperglycemia result from poor management
- Glucose levels affected by insulin, diet, and physical activity
- Studied in an outpatient setting by Maahs et al. (2012)

A Day in the Life of a T1D Patient

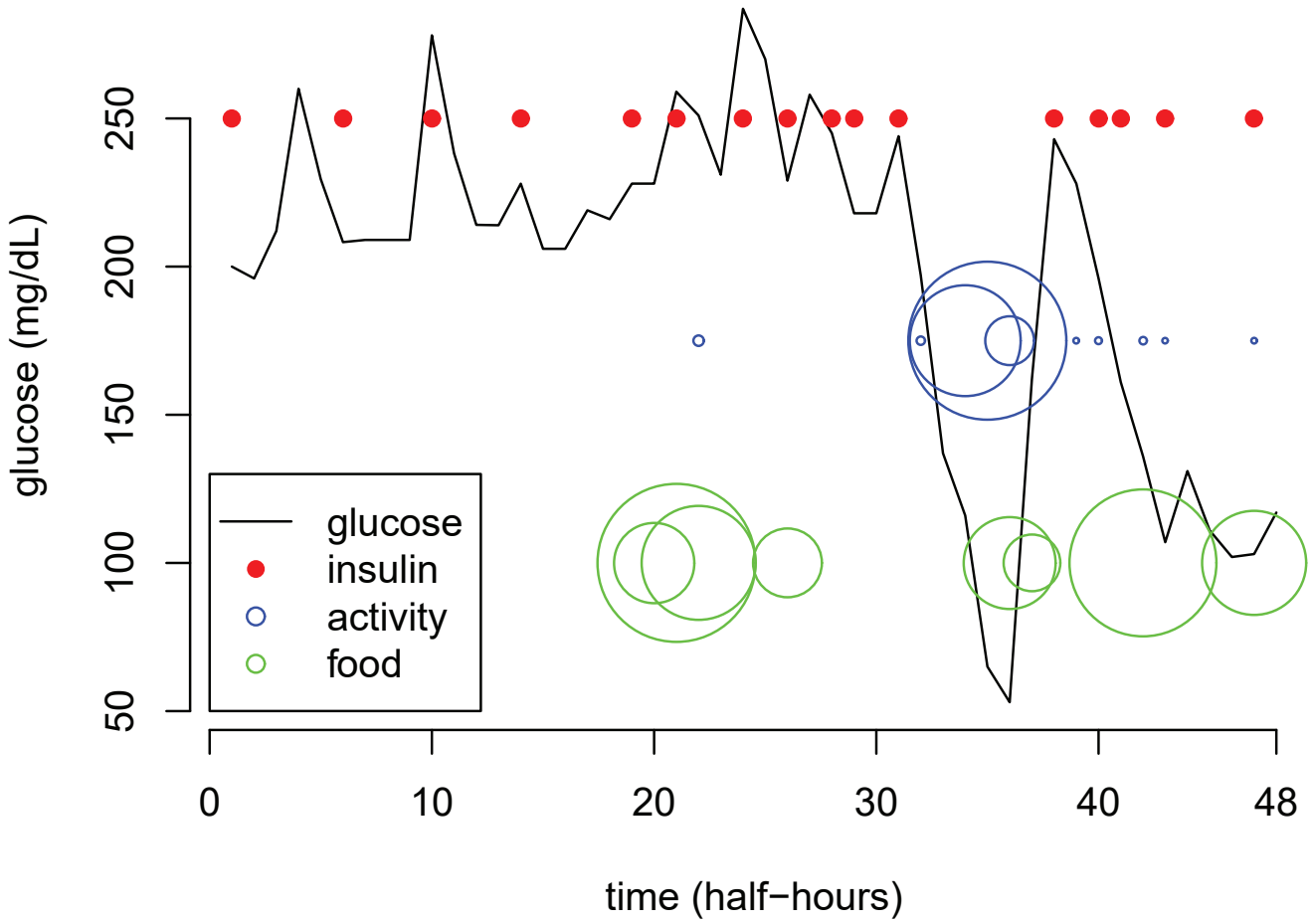

Figure 1: Time course of glucose level and other quantities.

Mobile technology in T1D care

Mobile devices can be used to administer treatment and assist with data collection in an outpatient setting, including

- Continuous glucose monitoring
- Accelerometers to track physical activity
- Insulin pumps to administer and log injections

These technologies can be incorporated using mobile phones.

Goals

Methodological goals:

- Estimate dynamic treatment regimes for use in mobile health
- Infinite time horizon, minimal modeling assumptions
- Observational data with minute-by-minute observations
- Online estimation to facilitate real-time decision making

Clinical goals:

- Provide patients information on the best actions to stabilize glucose
- Recommendations that are dynamic and personalized to the patient

Markov decision processes (MDPs)

Assume the data consist of n i.i.d. trajectories $(\mathbf{S}^1, A^1, \mathbf{S}^2, \ldots, \mathbf{S}^T, A^T, \mathbf{S}^{T+1})$ where $\mathbf{S}^t \in \mathbb{R}^p$, $A^t \in \mathcal{A}$, and there exists a known utility function $U^t = u(\mathbf{S}^{t+1}, A^t, \mathbf{S}^t)$.

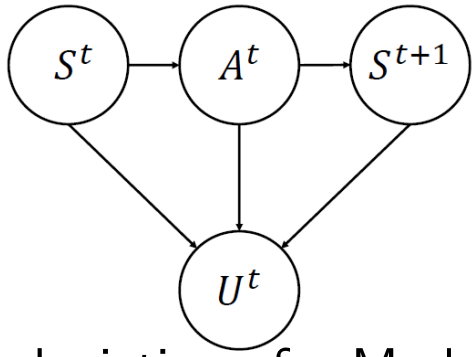

Figure 2: Graphical depiction of a Markov decision process.

Treatment regimes:

- Let $\mathcal{B}(\mathcal{A})$ be the space of distributions on $\mathcal{A}$
- A policy, π, is a function $\pi : \mathrm{dom}\, \mathbf{S}^t \to \mathcal{B}(\mathcal{A})$
- $\pi(a^t; \mathbf{s}^t)$ gives the probability of selecting $a^t \in \mathcal{A}$ when in state $\mathbf{S}^t = \mathbf{s}^t$

The state-value function

- The state-value function is

$$V(\pi, \mathbf{s}^t) = \mathbb{E}\left\{ \sum_{k \geq 0} \gamma^k U^{*(t+k)}(\pi) \middle| \mathbf{S}^t = \mathbf{s}^t \right\}$$

  for a discount factor $\gamma \in (0, 1)$

- For a distribution, $\mathcal{R}$, define the value of π, $V_{\mathcal{R}}(\pi) = \int V(\pi, \mathbf{s}) \mathrm{d}\mathcal{R}(\mathbf{s})$

- For a class of regimes, Π, the optimal regime, $\pi_{\mathcal{R}}^{\mathrm{opt}} \in \Pi$, satisfies

$$V_{\mathcal{R}}(\pi_{\mathcal{R}}^{\mathrm{opt}}) \geq V_{\mathcal{R}}(\pi)$$

  for all $\pi \in \Pi$.

An estimating equation for $V(\pi, \mathbf{s})$

Let $\mu^t(a^t; \mathbf{s}^t) = \Pr(A^t = a^t | \mathbf{S}^t = \mathbf{s}^t)$ for each $t \geq 1$.

Lemma

Assume strong ignorability, consistency, and positivity. Let π denote an arbitrary regime and $\gamma \in (0, 1)$ a discount factor. Then, provided interchange of the sum and integration is justified, the state-value function of π at $\mathbf{s}^t$ is

$$V(\pi, \mathbf{s}^t) = \sum_{k \geq 0} \mathbb{E}\left[\gamma^k U^{t+k} \left\{\prod_{v=0}^{k} \frac{\pi(A^{v+t}; \mathbf{S}^{v+t})}{\mu^{v+t}(A^{v+t}; \mathbf{S}^{v+t})}\right\} \middle| \mathbf{S}^t = \mathbf{s}^t\right].$$

This result will form the basis of an estimating equation for $V(\pi, \mathbf{s})$.

An estimating equation for $V(\pi, \mathbf{s})$ (continued)

It follows that

$$0 = \mathbb{E}\left[\frac{\pi(A^t; \mathbf{S}^t)}{\mu^t(A^t; \mathbf{S}^t)} \left\{U^t + \gamma V(\pi, \mathbf{S}^{t+1}) - V(\pi, \mathbf{S}^t)\right\} \psi(\mathbf{S}^t)\right],$$

for any function ψ (an importance-weighted version of the Bellman equation). An estimating equation for $V(\pi, \mathbf{s})$ is

$$\Lambda_n(\pi, \theta^\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T_i} \frac{\pi(A_i^t; \mathbf{S}_i^t)}{\mu^t(A_i^t; \mathbf{S}_i^t)} \left\{U_i^t + \gamma V(\pi, \mathbf{S}_i^{t+1}; \theta^\pi) - V(\pi, \mathbf{S}_i^t; \theta^\pi)\right\} \nabla_{\theta^\pi} V(\pi, \mathbf{S}_i^t; \theta^\pi),$$

where $V(\pi, \mathbf{S}; \theta^\pi)$ is a parametric model for the state-value function.

V-learning

Given an estimate $\widehat{\theta}_n^{\pi}$, an estimate of the value of π under $\mathcal{R}$ is $\widehat{V}_{n,\mathcal{R}}(\pi) = \int V\left(\pi, \mathbf{s}; \widehat{\theta}_n^{\pi}\right) \mathrm{d}\mathcal{R}(\mathbf{s})$ and an estimate of the optimal policy is $\widehat{\pi}_n = \arg\max_{\pi \in \Pi} \widehat{V}_{n,\mathcal{R}}(\pi)$. Start with an initial policy, π, and repeat until convergence:

1. Estimate $\widehat{\theta}_n^{\pi}$
2. Evaluate $\widehat{V}_{n,\mathcal{R}}(\pi) = \int V\left(\pi, \mathbf{s}; \widehat{\theta}_n^{\pi}\right) \mathrm{d}\mathcal{R}(\mathbf{s})$
3. Take a step to maximize $\widehat{V}_{n,\mathcal{R}}(\pi)$ over a class of policies

Online estimation

Given n i.i.d. copies of accumulating data $\left\{\left(\mathbf{S}^1, A^1, \mathbf{S}^2, \ldots\right)\right\}_{i=1}^{n}$,

- Select actions using estimated policy and estimate a new policy
- Introduce randomness into deterministic policies (e.g., ϵ-greedy)

The optimal policy at time t is based on the estimating equation

$$\begin{aligned}\Lambda_{n,t}(\pi, \theta^\pi) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{v=1}^{t}\frac{\pi(A_i^v; \mathbf{S}_i^v)}{\widehat{\pi}_n^{v-1}(A_i^v; \mathbf{S}_i^v)} \\ &\quad \times \left\{U_i^v + \gamma V(\pi, \mathbf{S}_i^{v+1}; \theta^\pi) - V(\pi, \mathbf{S}_i^v; \theta^\pi)\right\} \nabla_{\theta^\pi} V(\pi, \mathbf{S}_i^v; \theta^\pi).\end{aligned}$$

Note that $\widehat{\pi}_n^{t-1}$ replaces μ^t as the data-generating policy.

Implementation details

- Let $V(\pi, \mathbf{s}_i^t; \theta^\pi) = \Phi(\mathbf{s}_i^t)^\intercal \theta^\pi$ where $\Phi : \mathbb{R}^p \to \mathbb{R}^q$
- Under this model, $\Lambda_n(\pi, \theta^\pi)$ is linear in θ^π
- Let $\widehat{\theta}_n^\pi = \arg\min_{\theta^\pi \in \Theta} \{\Lambda_n(\pi, \theta^\pi)^\intercal \Lambda_n(\pi, \theta^\pi) + \lambda_n(\theta^\pi)^\intercal \theta^\pi\}$
- When there are J possible actions, define the class of policies parametrized by $\beta = (\beta_1^\intercal, \ldots, \beta_{J-1}^\intercal)^\intercal$:

$$\pi_\beta(\mathbf{a}_j; \mathbf{s}) = \frac{\exp(\mathbf{s}^\intercal \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{s}^\intercal \beta_k)},$$

for $j = 1, \ldots, J-1$, and
$\pi_\beta(\mathbf{a}_J; \mathbf{s}) = 1 / \left\{1 + \sum_{k=1}^{J-1} \exp(\mathbf{s}^\intercal \beta_k)\right\}$

- $\widehat{V}_{n,\mathcal{R}}(\pi_\beta) = \mathbb{E}_n \Phi(\mathbf{S})^\intercal \theta^{\pi_\beta}$ is differentiable in β but non-convex

Greedy gradient Q-learning (GGQ)

Introduced by Ertefaie (2014): for a discount factor, γ,

$$Q^{\pi}(\mathbf{s}^t, a^t) = \mathbb{E}\left\{ \sum_{k \geq 0} \gamma^k U^{t+k}(\pi) \Big| \mathbf{S}^t = \mathbf{s}^t, A^t = a^t \right\}.$$

The Bellman optimality equation is

$$Q^{\text{opt}}(\mathbf{s}^t, a^t) = \mathbb{E}\left\{ U^t + \gamma \max_{a \in \mathcal{A}} Q^{\text{opt}}(\mathbf{S}^{t+1}, a) \Big| \mathbf{S}^t = \mathbf{s}^t, A^t = a^t \right\}.$$

which motivates the estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T_i} \left\{ u_i^t + \gamma \max_{a \in \mathcal{A}} Q(\mathbf{s}_i^{t+1}, a; \eta^{\text{opt}}) - Q(\mathbf{s}_i^t, a_i^t; \eta^{\text{opt}}) \right\} \nabla_{\eta^{\text{opt}}} Q(\mathbf{s}_i^t, a_i^t; \eta^{\text{opt}}).$$

After computing the estimate $\widehat{\eta}_n^{\text{opt}}$, the optimal policy in state $\mathbf{s}$ selects action $\widehat{\pi}_n(\mathbf{s}) = \max_{\mathbf{a} \in \mathcal{A}} Q(\mathbf{s}, \mathbf{a}; \widehat{\eta}_n^{\text{opt}})$.

Asymptotic Inference for V-Learning

- We obtain uniform asymptotic normality for key parameters and predictions
- Main technical tools:
  - Donsker theorem for β-mixing stationary processes based on bracketing entropy (Dedecker and Louhichi, 2002)
  - New bracketing entropy preservation results for products of function classes
- Issue: Need Donsker theorems for non-stationary processes for certain types of online V-learning

T1D outpatient data set

- Maahs et al. (2012) followed $N = 31$ patients for a total of 5 days
- Insulin injections were logged via an insulin pump
- Glucose levels were tracked using continuous glucose monitoring
- Dietary intake data were reported in phone interviews
- Utility is the weighted sum of glycemic status
- We want to estimate a treatment policy to control glucose through a personalized insulin regimen or personalized recommendations for insulin, diet, and exercise
- Evaluate with parametric value estimate, $\widehat{V}_{n,\mathcal{R}}(\widehat{\pi}_n) = \mathbb{E}_n \Phi(\mathbf{S})^\intercal \widehat{\theta}_n^{\widehat{\pi}_n}$

An application to T1D data

Action space	Basis	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
Binary	Linear	-6.20	-9.35	-15.99
	Polynomial	-3.91	-9.03	-17.50
	Gaussian	-3.44	-13.09	-25.52
Multiple	Linear	-6.47	-9.92	-0.49
	Polynomial	-2.44	-6.80	-14.48
	Gaussian	-8.45	-3.58	-21.18
Observational policy		-6.77	-11.28	-21.79

Table 1: Parametric value estimates for V-learning applied to type 1 diabetes data.

Example hyperglycemic (229 mg/dL) patient

Action	Probability
No action	< 0.0001
Physical activity	< 0.0001
Food intake	< 0.0001
Food and activity	< 0.0001
Insulin	0.7856
Insulin and activity	0.2143
Insulin and food	0.0002
Insulin, food, and activity	< 0.0001

Table 2: Probabilities for each action as recommended by estimated policy for one example hyperglycemic patient.

Conclusions

- Glucose is known to depend on insulin, diet, and exercise
- Optimal treatment regimes for T1D depend on lifestyle considerations; little work has been done to construct decision rules for T1D management in a statistically rigorous way
- Advantages of V-learning include
  - Flexibility in models for $V(\pi, \mathbf{s}; \theta^{\pi})$ and $\pi_{\beta}(a; \mathbf{s})$
  - Minimal assumptions about the data-generating process
  - Randomized decision rules for online estimation
  - No need to model a non-smooth functional of the data
- A tailored treatment regime delivered through mobile devices may help to reduce hypo- and hyperglycemia in T1D patients