Optimization of Sequential Decision Making for Chronic Diseases: From Data to Decisions

2018 INFORMS Annual Meeting
Phoenix, Arizona
November 5, 2018

Brian Denton
Department of Industrial and Operations Engineering
University of Michigan
Fishing is the reason I love decision making under uncertainty!

These slides (and pictures 😊) are on my website:
http://umich.edu/~btdenton
Chronic Diseases

Cancer

Diabetes

Kidney Disease

Heart Disease
PubMed results by methodology in the last 10 years
Talks at this conference
Topics in this tutorial

• Markov Decision Process (MDP) Basics

• Partially Observable Markov Decision Processes (POMDPs)

• Data-Driven Model Parameterization

• Other Models for Medical Decision-Making

• Conclusions
Healthcare problems addressed by MDPs and POMDPs


Dynamic programming (DP) dates back to early work of Richard Bellman in the 1940’s. The 1954 Paper by Bellman describes the foundation for DP. Since its development DP has been applied to fields of mathematics, engineering, biology, chemistry, medicine, and many others. For more history on Richard Bellman see: http://www.gap-system.org/~history/Biographies/Bellman.html
Definitions

A **policy** defines the **action** to take in each possible **state** of the system.

An **optimal policy** defines the **optimal action** to take for each **state** that will achieve some goal such as:

- Maximize rewards gained over time
- Minimize costs paid over time
- Achieve an outcome with high probability
Shortest path problems

• DPs can be used for finding the shortest path that joins two points in a network

• Many problems can be formulated as a shortest path problem
Shortest Path Example

What is the shortest path in this directed graph?
Principle of Optimality

Following is a quote from a 1954 paper by Richard Bellman:

“An optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.”

Dynamic program terminology

Main Elements

- **States**: vertices of the graph
- **Actions**: which vertex to move to
- **Transfer Function**: edges of the graph
- **Rewards**: cost associated with selecting an edge

**Goal**: Starting from vertex S, select the action at each vertex that will minimize total edge distance travelled to reach vertex G
Dynamic program formulation

Let a DP have **states**, $s_t \in S$, **actions** $a_t \in A$, **rewards**, $r_t(s_t, a_t)$, and an **optimal value function**, $v_t(s_t)$, defined for stages $t = 1, 2, ..., T$

**Optimality Equations:**

$$v_t(s_t) = \max_{a_t \in A} \{ r_t(s_t, a_t) + v_{t+1}(s_{t+1}) \}, \quad \forall s_t$$

$$v_T(s_T) = R(s_T), \quad \forall s_T$$

$v_t(s_t)$ is the maximum total reward for all stages $t, t + 1, ..., T$, also called the “optimal value to go”

Transition from $s_t$ to $s_{t+1}$ governed by a **transfer equation**:

$$s_{t+1} = g(s_t, a_t)$$
Assumptions made in this tutorial

- Finite horizon
- The set of decision epochs, $T$, actions, $A$, and states, $S$, are finite
- The decision maker’s goal can be represented by linear additive rewards
What About Uncertainty?

Uncertainty arises in many ways in chronic diseases:

- Future health status
- Treatment effects
- Diagnostic test results
- Procedure outcomes

The first and easiest step to address uncertainty is a Markov decision process (MDP)
Optimality Equations

For all states, $s_t$, and all time periods, $t = 1, \ldots, T - 1$

$$v_t(s_t) = \max_{a_t \in A} \left\{ r_t(s_t, a_t) + \lambda \sum_{s_{t+1} \in S} p(s_{t+1} | s_t, a_t) v_{t+1}(s_{t+1}) \right\}$$

Boundary Condition: $v_T(s_T) = R(s_T), \ \forall s_T$
Fundamental Result

**Theorem:** Suppose $v_t(s_t)$, for all $t$ and $s_t$ is a solution to the optimality equations, then $v_t(s_t) = v_t^*(s_t)$, for all $t$ and $s_t$ the associated actions define the optimal policy $\pi^*$ for the MDP.

**Importance:** This proves solving the optimality equations yields an optimal solution to the MDP.

Reference: These results are an aggregate of results presented in chapter 4 of “Markov Decision Processes: Discrete Stochastic Dynamic Programming,” by Puterman.
Special Structured Policies

Policies with a simple structure are:

• Easier for decision makers to understand
• Easier to implement
• Easier to solve the associated MDPs

General structure of a control limit policy

\[ a_t(s_t) = \begin{cases} 
  a_1, & \text{if } s < s^* \\
  a_2, & \text{if } s \geq s^* 
\end{cases} \]

where \( a_1 \) and \( a_2 \) are alternative actions and \( s^* \) is a control limit.
Monotone Policies

**Definition:** Control limit policies are examples of **monotone** policies. A policy is **monotone** if the *decision rule* at each stage is nonincreasing or nondecreasing with respect to the system state.
Monotonicity: Sufficient Conditions

**Theorem:** Suppose for $t = 1, \ldots, T - 1$

1. $r_t(s, a)$ is nondecreasing in $s$ for all $a \in A$.
2. $q_t(k|s, a)$ is nondecreasing in $s$ for all $k \in S, a \in A$. (IFR Property)
3. $r_t(s, a)$ is **superadditive (subadditive)** on $S \times A$.
4. $q_t(k|s, a)$ is **superadditive (subadditive)** on $S \times A$, $\forall k$
5. $R_T(s)$ is nondecreasing in $s$.

Then there exist optimal decision rules, $d_t^*(s)$, which are nondecreasing (nonincreasing) in $s$ for $t = 1, \ldots, T - 1$.

See Puterman, chapter 4, for discussion of this and related properties.
Some treatment decisions can be viewed as a “stopping time” problem:

- **Statins** lower your risk of heart attack and stroke
- Treatment has **side effects** and cost
- Patients **decision:**
  - initiate statins
  - defer initiation for a year
Model Description

Decision epochs:

- Time horizon: Ages 40-80
- Annual decision epochs

Actions: **Initiate** (Q) or **delay** (C) statin treatment

States:

- Risk factors: Total cholesterol and HDL
- Demographic: Gender, Race, BMI, smoking status, medical history
Stopping Time Problem

Optimality equations:

\[ v_t(s) = \max_{a \in \{Q, C\}} \left\{ R_t(s), r(s, C) + \sum_{j \in S} p(j|s, C)v_{t+1}(j) \right\}, \forall s \in S \]

\[ v_T(s) = R_T(s), \forall s \in S \]

States define patient health status

Action \( C \) represents decision to defer statin initiation, \( Q \) denotes decision to start statins

\( R_t(s) \) is expected survival if statins are initiated
Statin Treatment Markov Chain

Metabolic States before an event has occurred.

L

M

H

V

On Statins

Heart Attack or Stroke

\( R_t(S) \)

\( r(L,W) \)

\( r(M,W) \)

\( r(H,W) \)

\( r(V,W) \)
Rewards

There are various types of reward functions used in health studies like this. The simplest definition for this problem is:

• $r_t(s_t)$ is the time between decision epochs (e.g. 1 year)

• $R_t(s_t)$ is the expected future life years adjusted for quality of life on medication
Computing Transition Probabilities

Transition probabilities between **metabolic states**:

- Longitudinal electronic medical record data for total cholesterol (bad cholesterol) and HDL (good cholesterol) levels for many patients

Transition probabilities from healthy states to **complication state**

- Published cardiovascular **risk models** that estimate the probability of heart attack or stroke in the next year
Computing Transition Probabilities
Estimating Transition Probabilities

Transition probabilities are estimated using longitudinal data for a cohort of patients that includes:

- Laboratory and clinical data (e.g. cholesterol, blood pressure)
- Pharmacy claims data indicating prescriptions

\[
p(s'|s, a) = \frac{n(s, s', a)}{\sum_{s'} n(s, s', a)}, \forall s', s, a
\]
Optimal Treatment Initiation Policy

Bad cholesterol/Good cholesterol

Female

Male
Other Related Examples

The previous example is based on this paper:


Following are extensions:

Other Examples of MDPs for Chronic Disease


MDPs: Where to learn more
Sections

• Markov Decision Process (MDP) Basics
• Partially Observable Markov Decision Processes (POMDPs)
• Data-Driven Model Parameterization
• Other Models for Medical Decision-Making
• Conclusions
Partially Observable MDPs (POMDPs)

Model Elements:

• Decision Epochs: $t = 1, \ldots, T$

• **Core States**: $s_t \in S$

• Actions: $a_t \in A$

• Rewards: $r_t(s_t, a_t)$

• Transition Probability Matrix: $P$

• **Observations**: $o \in O$

• **Observation Probability Matrix**: $Q \in R^{|S| \times |O|}$

Unique to POMDPs
POMDP Sequence of Events

Choose action $a_t(b_t)$ → Transition to new core state, $s_{t+1}$, in the Markov chain → Receive “observation” according to observation probability matrix

unobserved

Updated belief vector, $b_t$ → Bayesian update → Receive reward, $r_t(s_t, a_t)$

Start with $b_0$ → End with $R_T(s_T)$
Sufficient Statistic

The **belief vector** has one element for each state that defines the probability the system is in state $s_t$

$$b_t(s_t) = P(s_t | o_t, a_{t-1}, o_{t-1}, a_{t-2}, ..., o_1, a_0)$$

$h_t$

Complete history of observations (up to $t$) and actions (up to $t - 1$)

The belief vector is a **sufficient statistic** to define the optimal policy for a POMDP.
Bayesian Updating

Belief Update Formula:

\[ b_t(s_t) \equiv P(s_t | h_t) = \frac{P(s_t, o_t, a_{t-1} | h_{t-1})}{P(o_t, a_{t-1} | h_{t-1})} \]

Numerator:

\[ P(s_t, o_t, a_{t-1} | h_{t-1}) = \sum_{s_{t-1} \in S} P(s_t, o_t, a_{t-1}, s_{t-1} | h_{t-1}) \]

\[ = \sum_{s_{t-1} \in S} P(o_t | s_t, a_{t-1}, s_{t-1}, h_{t-1}) P(s_t | a_{t-1}, s_{t-1}, h_{t-1}) P(a_{t-1} | s_{t-1}, h_{t-1}) P(s_{t-1} | h_{t-1}) \]

\[ = P(a_{t-1} | h_{t-1}) P(o_t | s_t) \sum_{s_{t-1} \in S} P(s_t | a_{t-1}, s_{t-1}) b_{t-1}(s_{t-1}) \]
Bayesian Updating

Belief Update Formula:

\[ b_t(s_t) = \frac{P(s_t, o_t, a_{t-1} | h_{t-1})}{P(o_t, a_{t-1} | h_{t-1})} \]

Denominator:

\[ P(o_t, a_{t-1} | h_{t-1}) = \sum_{s_{t'} \in S} \sum_{s_{t-1} \in S} P(s_{t'}, o_t, a_{t-1}, s_{t-1} | h_{t-1}) \]

\[ = \sum_{s_{t'} \in S} \sum_{s_{t-1} \in S} P(o_t | s_{t'}, a_{t-1}, s_{t-1}, h_{t-1}) P(s_{t'} | a_{t-1}, s_{t-1}, h_{t-1}) P(a_{t-1} | s_{t-1}, h_{t-1}) P(s_{t-1} | h_{t-1}) \]

\[ = P(a_{t-1} | h_{t-1}) \sum_{s_{t'} \in S} P(o_t | s_{t'}) \sum_{s_{t-1} \in S} P(s_{t'} | a_{t-1}, s_{t-1}) b_{t-1}(s_{t-1}) \]
Bayesian Updating

Now everything is in terms of transition probabilities, observation probabilities, and the prior belief vector

\[ b_t(s_t) = \frac{P(o_t|s_t) \sum_{s_{t-1} \in S} P(s_t|s_{t-1}, a_{t-1}) b_{t-1}(s_{t-1})}{\sum_{s_{t'} \in S} P(o_t|s_{t'}) \sum_{s_{t-1} \in S} P(s_{t'}|s_{t-1}, a_{t-1}) b_{t-1}(s_{t-1})} \]

Numerator: Probability of observing \( o_t \) and system is in \( s_t \)

Denominator: Probability of observing \( o_t \)
Optimality Equations for POMDPs

Rewards Vector: \( r_t(a_t) = (r_t^1(a_t), \ldots, r_t^S(a_t))' \) denotes the expected rewards under transitions and observations

\[
 r_t^{st}(a_t) = \sum_{o_{t+1} \in \mathcal{O}} \sum_{s_{t+1} \in S} r(s_t, a_t, s_{t+1}, o_{t+1}) p(s_{t+1} | s_t, a_t) p(o_{t+1} | s_{t+1})
\]

Optimality Equations: In POMDPs, the value function is defined on the belief space.

\[
v_t(b_t) = \max_{a_t \in \mathcal{A}} \left\{ b_t \cdot r_t(a_t) + \lambda \sum_{o_{t+1} \in \mathcal{O}} \gamma(o_{t+1} | b_t, a_t) v_{t+1}(T(b_t, a_t, o_{t+1})) \right\}
\]

Boundary Condition: \( v_{T+1}(b_{T+1}) = b_{T+1} \cdot r_{T+1} \)
Solution Methods

• POMDPs are difficult to solve exactly:
  • Time complexity is exponential in the number of actions, observations, and decision epochs
  • Dimensionality in the state space grows with the number of core states

• Complexity class is **P-Space Hard**

• Most approaches rely on approximations: finite grids, supporting hyperplane sampling
POMDP Example: Prostate cancer screening

Age 40

- PSA Test Result
  - Y
  - N

Biopsy Result

Biopsy +

Biopsy -

PSA Test?

Age 41
Biomarker Test: PSA

Graph showing the PSA levels (ng/mL) over age (40-75) for two cohorts: one cancer-free and one with cancer at age 52.
Core States

Markov transitions between prostate cancer states:
- No cancer (NC) → Unobservable
- Cancer present but not detected (C) → Unobservable
- Cancer detected (T) → Treated immediately after detected
- Death (D) → Prostate cancer and other cause mortality
Detailed Model Description

• Decision Epochs, $t = 40, 41, \ldots, 85$

• Health States: Health/cancer status, $s_t$

• Observations: PSA test result, $o_t$

• Observation Matrix: $q_t(o_t|s_t)$

• Rewards: Quality adjusted life years
  • $r_t(\text{NC, No PSA Test}) = 1$
  • $r_t(\text{NC, PSA Test}) = 1 - \delta$
  • $r_t(\text{NC, Biopsy}) = 1 - \mu$
  • $r_t(\text{C, No PSA Test}) = 1$
  • $r_t(\text{C, PSA Test}) = 1 - \delta$
  • $r_t(\text{C, Biopsy}) = 1 - \mu - f\epsilon$

Resource to learn more about QALYs and other public health measures:
Model Data

11,872 patients from Olmsted county, MN with age, PSA, biopsy and cancer information from 1993 through 2006

<table>
<thead>
<tr>
<th>Population size</th>
<th>11,872</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age: Mean(SD)</td>
<td>63.0(12.7)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>96%</td>
</tr>
<tr>
<td>Other</td>
<td>4%</td>
</tr>
<tr>
<td>Outcomes</td>
<td></td>
</tr>
<tr>
<td>Prostate biopsy</td>
<td>908</td>
</tr>
<tr>
<td>Prostate cancer diagnosis</td>
<td>628</td>
</tr>
</tbody>
</table>

Other parameters are drawn from the medical literature
Optimal Policy for Screening

Other examples of POMDPs for chronic disease


POMDPs: Where to learn more

• Tutorial: “POMDPs for Dummies” http://cs.brown.edu/research/ai/pomdp/tutorial/


Sections

- Markov Decision Process (MDP) Basics
- Partially Observable Markov Decision Processes (POMDPs)
- Data-Driven Model Parameterization
- Other Models for Medical Decision-Making
- Conclusions
The Movember Foundation’s GAP3 Cohort

The Movember Foundation launched the Global Action Plan Prostate Cancer Active Surveillance (GAP3) to create a global database:

- includes 15,101 patients from 25 established AS cohorts worldwide

- records longitudinal observations of patients’ clinical and demographic characteristics
Hidden Markov Model (HMM)

- Time periods: annual
- Initial distribution:
  \[ \phi = (\phi_1, 1 - \phi_1) \]
- Transition probabilities:
  \[ A_t = [P(s_{t+1} | s_t)] \]
- Observations:
  \[ O_t = (PSA_t, Biopy_t) \]
Baum-Welch Algorithm for Parameter Estimation

Given the observation sequences

\[ O^{(1)} = \left( O_1^{(1)}, \ldots, O_{T_1}^{(1)} \right), \ldots, O^{(N)} = \left( O_1^{(N)}, \ldots, O_{T_N}^{(N)} \right), \]

**Baum-Welch algorithm**, or equivalently *the EM (expectation-maximization)* estimates the model

\[ \lambda = (\phi, A, B, C, \mu, \sigma) \]

that **locally maximizes the likelihood function**

\[ P(O|\lambda) = \prod_{k=1}^{N} P(O^{(k)}|\lambda) \]

Partially Observable Markov Decision Process

- Objective: to balance the harm of biopsy with the benefit of early detection
- Decision Epochs: every year
- Actions: PSA test only, PSA test and Biopsy

- Hidden States: Low-Risk Cancer, High-Risk Cancer
- Initial Distribution: $\phi$
- Transition Probability Matrix: $A$
- Biopsy Observation Probability Matrix: $B$
- PSA Observation Probability Matrix: $C$

These elements define the decision process and goal

These elements come from the HMM
Results: Optimal Value Function at Age 50

- The optimal policy is a threshold-based policy: if the belief of high-risk state exceeds the threshold, then do biopsy.

Weights are set equally for criteria:
- delay in detection of high risk cancer
- harm from biopsy
Title: A Data-driven Partially Observable Markov Decision Process for Optimizing Individualized Surveillance Strategies for Prostate Cancer

Weiyu Li, Brian Denton

Session: TD76 - Joint Session MIF/HAS: Models and Methods for Improving Patient Outcomes

November 6, 2018, 2:00 PM - 3:30 PM @ West Bldg. 212C
Sections

• Markov Decision Process (MDP) Basics
• Partially Observable Markov Decision Processes (POMDPs)
• Data-Driven Model Parameterization
• Other Models for Medical Decision-Making
• Conclusions
Other Models - Robust MDPs

• All models are subject to uncertainty in model parameter estimates and model assumptions

• Transition probabilities are based on statistical estimates from longitudinal data

• Rewards are based on statistical estimates of mean patient utility, cost, or other performance measures

• Robust MDPs (RMDPs) attempt to account for this uncertainty
RMDP Models

An RMDP assumes TPM is restricted to lie in an uncertainty set, \( U \), leading to the following optimality equations:

\[
\pi^* = \arg\max_{\pi \in \Pi} \min_{P \in U} E^P \left[ \sum_{t=1}^{N-1} r_t(s_t, \pi(s_t)) + r_N(s_N) \right]
\]

**Time Invariant Case** – Adversary selects a single TPM

\[
\pi^* = \arg\max_{\pi \in \Pi} \min_{P \in U} E^P \left[ \sum_{t=1}^{N-1} r_t(s_t, \pi(s_t)) + r_N(s_N) \right]
\]

**Time Varying Case** – Adversary selects a TPM at each epoch

\[
\pi^* = \arg\max_{\pi \in \Pi} \min_{P_t \in U} E^{P_t} \left[ \sum_{t=1}^{N-1} r_t(s_t, \pi(s_t)) + r_N(s_N) \right]
\]
Uncertainty Sets

Many choices of $U$ have been proposed:

- Finite scenario model:
  
  $$U(s_t) = \{ p^1(s_t), p^2(s_t), \ldots, p^K(s_t) \}$$

- Interval model:

  $$U(s_t) = \{ p(s_t) | \underline{p}(s_t) \leq p(s_t) \leq \overline{p}(s_t), p(s_t) \cdot 1 = 1 \}$$

- Ellipsoidal models, relative entry bounds, …
RMDP Case Study: Type 2 Diabetes

Many medications that vary in efficacy, side effects and cost.

Oral Medications:
- Metformin
- Sulfonylurea
- DPP-4 Inhibitors

Injectable Medications:
- Insulin
- GLP-1 Agonists
Treatment Goals

• HbA1C is an important biomarker for blood sugar control

• But disagreement exists about the optimal goals of treatment and which medications to use
Markov Chain for Type 2 Diabetes

Absorbing states: $E^{\text{macro}}$, $E^{\text{micro}}$, $D^0$

HbA1C States

- $\ell(1)$
- $\ell(2)$
- $\ell(3)$
- $\ldots$
- $\ell(k)$

Initiate or delay medication 1

Initiate or delay medication n
Estimating the Uncertainty Set

A combination of laboratory data and pharmacy claims data was to estimate transition probabilities between deciles

\[ p(s'|s, a) = \frac{n(s, s', a)}{\sum_s n(s, s', a)}, \forall s', s, a \]

1 − \( \alpha \) confidence intervals for row \( s \) of the TPM:

\[ [\hat{p}(s'|s, a) - S(\hat{p}(s'|s, a)L, \hat{p}(s'|s, a) + S(\hat{p}(s'|s, a)L] \]

where

\[ S(\hat{p}(s'|s, a)L = \left[ X_{|s|-1, \alpha/2 |s|}^2 \right]^{1/2} \frac{\hat{p}(s'|s, a)(1 - \hat{p}(s'|s, a))}{N(s)} \]
Uncertainty Set with Budget

\[
U(s_t) = \left\{ \begin{array}{l}
p(s_{t+1}|s_t) = \hat{p}(s_{t+1}|s_t) - \delta^L z^L(s_{t+1}) + \delta^U z^U(s_{t+1}), \quad \forall s_{t+1} \\
\sum_{s_{t+1} \in S} p(s_{t+1}|s_t) = 1 \\
\sum_{s_{t+1}} (z^L(s_{t+1}) + z^U(s_{t+1})) \leq \Gamma(s_{t+1}) \\
z^L(s_{t+1}) \cdot z^U(s_{t+1}) = 0, \quad \forall s_{t+1} \\
0 \leq p(s_{t+1}|s_t) \leq 1, \quad \forall s_{t+1}
\end{array} \right.
\]

Properties:
• Can be reformulated as a linear program
• For \( \Gamma = |S| \) can be solved in \( O(|S|) \)
Results

Quality adjusted life years to first health complications for women with type 2 diabetes
Accounting for ambiguity in MDPs

Title: Leveraging decomposition methods to design robust policies for Markov decision processes

Lauren N. Steimle, Brian T. Denton

Session: SD01: Applications of Stochastic Programming
November 4th, 4:30-6:30 PM in North Building 121A

RMDPs: Where to learn more


Model-Free Methods

Two major sources of challenges to solving MDPs are:

1) “curse of dimensionality”
2) “curse of modeling”

“Model-Free” methods are suited to problems of type 2, for which transition probabilities are not known

These methods are known under various names including: reinforcement learning
Model-Free Methods

**Monte Carlo sampling** is a common approach for estimating the expectation of functions of random variables.

Model free approaches use **sample paths** to estimate the value function.

These methods are known under various names including: *reinforcement learning*
Monte-Carlo Sampling

Model free approaches use sample paths to estimate the value function via Monte Carlo sampling

\[ E^\pi \left[ \sum_{t=1}^{N-1} r_t(s_t, \pi(s_t)) + r_N(s_N) \right] \]

\[ \approx \frac{1}{K} \sum_{k=1}^{K} \sum_{t=1}^{N-1} r_t(s_t^k, \pi(s_t^k)) + r_N(s_N^k) \]

Where \( k = 1, \ldots, K \) are random sample paths from the Markov chain.
Monte Carlo Policy Evaluation

A selected policy $\pi$ can be evaluated approximately via Monte Carlo sampling

As $K \to \infty \tilde{v}^\pi(s_0) \to v^\pi(s_0)$

In practice the number of samples, $N$, must be chosen to tradeoff between (a) some desired level of confidence and (b) a computational budget.
Example: Bandit Problem

Consider a game in which your friend holds two coins: 1 coin is fair, the other is biased towards landing heads up.

You know your friend holds two different coins but you don’t know the likelihood of each turning up a head.

Each turn you get to select the coin your friend will flip. If you win you get $1 if you lose you lose $1.

Question: how would you play this game?

Application: medical treatment decisions with multiple treatment options and uncertain rewards
Example: multi-armed bandit

The action is which “arm”, $a$, to try at each decision epoch, and the expected reward for this action is $Q_t(a)$.

Since $Q_t(a)$ is not known exactly it must be estimated as:

$$\tilde{Q}_t(a) = \frac{r_1 + r_2 + \ldots + r_{k_a}}{k_a}$$

Where $k_a$ is the number of times arm $a$ has been sampled.

As $k_a \to \infty$, $\tilde{Q}_t(a) \to Q_t(a)$, thus sampling each arm an infinite number of times will identify the optimal action

$$a^* = \arg\max_{a \in A}\{Q_t(a)\}.$$
Example: multi-armed bandit

Policies obtained from learning attempt to converge to a near optimal policy quickly.

The simplest learning-based policy is the greedy policy:

\[ \hat{a} = \text{argmax}\{\hat{Q}(a)\} \]

Alternatively the \( \epsilon - \text{greedy} \) method explores the action set by randomly selecting actions with probability \( \epsilon \).

As \( k_a \to \infty \), \( Q_t(a) \to Q_t^*(a) \) and the optimal action is selected with probability greater than \( 1 - \epsilon \).
Monte Carlo Policy Iteration

For more complex problems with multiple system states the following algorithm can be used

Algorithm (MC Policy Iteration):

1. For all $s$ initialize $\pi(s)$ and $Q(s, \pi(s))$. Choose a suitably large $N$.

2. **Policy Evaluation:**

   Randomly select a starting pair, $(s, \pi(s))$, and generate a sample path of length $N$

   For all $(s, \pi(s))$ in the sample path compute: $\tilde{Q}^\pi(s, \pi(s)) = \sum_{t=n_s}^{N-1} \lambda^t r_t(s_t, \pi(s_t)) + \lambda^N r_N(s_N)$, where $n_s$ is the index for the first instance state $s$ is encountered.

3. **Policy Improvement:**

   For all $s$: $\pi(s) \in \operatorname{argmax}_{a \in A} \{Q(s, a)\}$

   Return to Step 2;
Other Approaches

- Temporal difference learning
- Q-learning
Example: SMART Trials

Where to Learn More


“Reinforcement Learning: An Introduction”, By Sutton and Barto, MIT Press
Sections

- Markov Decision Process (MDP) Basics
- Partially Observable Markov Decision Processes (POMDPs)
- Data-Driven Model Parameterization
- Other Models for Medical Decision-Making
- Conclusions
Take Away Messages

• **Operations research** has an important role to play in understanding and advancing medical decisions

• **Observational data is an extraordinary resource** but there are important research questions to answer to unlock the value

• There are **extraordinary research opportunities** to bring optimization methods to bear on diseases – you can be the first person to study many diseases
Acknowledgements

Weiyu Li, PhD Student, University of Michigan

Lauren Steimle, PhD Candidate, University of Michigan

Zheng Zhang, Postdoctoral Fellow, University of Michigan

This work was funded in part by grants CMMI-1536444 and CMMI-1462060 from the Operations Engineering program at the National Science Foundation.
Brian Denton
University of Michigan
btdenton@umich.edu

These slides (and pictures 😊) are on my website:
http://umich.edu/~btdenton