Applied Reinforcement Learning Seminar: Introduction to Reinforcement Learning

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, 2020

Outline

Decision support in precision medicine

- A central goal in precision medicine is decision support, often in the form of dynamic treatment regimes.
- In this talk, we will explore the multiple decision setting.
- In the multi-stage setting, a dynamic treatment regime is a sequence of decision rules that, at each decision point, selects a next treatment based on a patient's baseline characteristics and accrued information up to that point.
- This mirrors clinical practice where physicians make a series of treatment decisions over the course of a patient's disease based on his/her baseline/evolving characteristics.
- Frequently in biomedical research, the focus is on off-policy (versus on-policy) reinforcement learning.
- In this presentation, we assume there is a finite (possibly random) number of decisions to be made: this the finite horizon setting (versus infinite horizon).

Goals

Our goals in this lecture are to:

- Formalize the multi-decision setting,
- Define the corresponding statistical problem,
- Describe Q-learning and A-learning, and
- Compare Q- and A-learning.

Examples

Two clinical examples:

- Precision medicine for severe burn repair: Hibbard JC, Friedstat JS, Davis SM, Edkins RE, Hultman CS, and Kosorok MR (2018). LIBERTY: A SMART study in plastic surgery. Clinical Trials 15:286-293.
- Personalizing diets for optimizing health in type 1 diabetes: Accelerating Solutions to Optimize Glycemic Control and Weight Management in Young Adults with Type 1 Diabetes: A SMART Study.

Outline

Notation

- K prespecified, ordered decision points indexed by $k = 1, \ldots, K$.
- Final outcome of interest, Y.
  - Assume larger values are better.
- Let Ω be a superpopulation of patients and $\omega \in \Omega$ a patient from this population.
- At the first decision point, a patient ω presents to the physician with baseline covariates denoted by the random variable $S_1(\omega)$ that takes values $s_1 \in \mathcal{S}_1$.

Notation (cont.)

- At each decision point $k = 1, \ldots, K$, assume that there is a finite set of all possible treatment options $\mathcal{A}_k$ with elements a_k.
- Denote by $\bar{a}_k = (a_1, \ldots, a_k)$ a possible treatment history that could be administered through decision k, taking values in $\bar{\mathcal{A}}_k = \mathcal{A}_1 \times \cdots \times \mathcal{A}_k$.
- $\bar{\mathcal{A}}_K$ denotes the set of all possible treatment histories $\bar{a}_K$ through all K decisions.

Potential outcomes

As in the single-decision setting, we will define the optimal multi-stage decision rule in terms of potential outcomes. We have the following potential outcomes

$$W^* = \{S_2^*(a_1), S_3^*(\bar{a}_2), \ldots, S_k^*(\bar{a}_{k-1}), \ldots, S_K^*(\bar{a}_{K-1}), Y^*(\bar{a}_K) \text{ for all } \bar{a}_K \in \bar{\mathcal{A}}_K\}.$$

- $S_k^*(\bar{a}_{k-1})(\omega)$ denotes the value of the covariate information that would arise between decisions $k-1$ and k for a patient $\omega \in \Omega$ in the situation that he or she were to have previously received treatment history $\bar{a}_{k-1}$ and takes values s_k in a set $\mathcal{S}_k, k = 2, \ldots, K$.
- $Y^*(\bar{a}_K)(\omega)$ is the hypothetical outcome that would result for ω were he or she to have been administered the full set of K treatments in $\bar{a}_K$.
- We will write $\bar{S}_k^*(\bar{a}_{k-1})$for $\{S_1, S_2^*(a_1), \ldots, S_k^*(\bar{a}_{k-1})\}$

Dynamic treatment regimes

- A **dynamic treatment regime** (DTR) $d = (d_1, \ldots, d_K)$ is a set of rules that forms an algorithm for treating a patient over time.
- It is "dynamic" because treatment is based on a patient's previous history.
- The feasible treatments for a patient at a particular decision point depend on his or her history.
- $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ is the feasible set of treatments for a patient with history $(\bar{s}_k, \bar{a}_{k-1})$.
  - E.g. At the kth decision point we have the the kth rule $d_k(\bar{s}_k, \bar{a}_{k-1})$ outputs $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1}) \subseteq \mathcal{A}_k$. For $k = 1$, there is no prior treatment and we write $d_1(s_1)$ and $\Psi_1(s_1)$.

Dynamic treatment regimes

- Because $d_k(\bar{s}_k, \bar{a}_{k-1}) \in \Psi_k(\bar{s}_k, \bar{a}_{k-1}) \subseteq \mathcal{A}_k$, d_k need only map a subset of $\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1}$ to $\mathcal{A}_k$. We define these subsets recursively as

$$\begin{aligned}\Gamma_k = \{&(\bar{s}_k, \bar{a}_{k-1}) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1} : \\ &\text{(i) } a_j \in \Psi_j(\bar{s}_j, \bar{a}_{j-1}), j = 1, \ldots, k-1, \text{ and} \\ &\text{(ii) } pr\{\bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k\} > 0\}, k = 1, \ldots, K.\end{aligned}$$

- The Γ_k contain all realizations of covariate and treatment history consistent with having followed such Ψ-specific regimes to decision k.
- Define the class $\mathcal{D}$ of regimes to be the set of all d for which d_k, $k = 1, \ldots, K$, is a mapping from Γ_k into $\mathcal{A}_k$ satisfying $d_k(\bar{s}_k, \bar{a}_{k-1}) \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$ for every $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$.

Outline

Optimal treatment regime

- In the single-stage setting, we defined the optimal treatment regime $Y^*(d^{\text{opt}})$ as satisfying

$$E(Y^*(d)|S_1 = s_1) \leq E(Y^*(d^{\text{opt}})|S_1 = s_1)$$

  for all feasible d and all $s_1 \in \mathcal{S}_1$.
- We will see that the definition in the multi-stage setting is similar.
- Before doing so, we will need to define the potential outcomes associated with d when d is a multi-stage decision rule.

Potential outcomes for a DTR

- For any $d \in \mathcal{D}$, let $\bar{d}_k = (d_1, \ldots, d_k)$ for $k = 1, \ldots, K$ and $\bar{d}_K = d$.
- Define the potential outcomes associated with d as $\{S_2^*(d_1), \ldots, S_k^*(\bar{d}_{k-1}), \ldots, S_K^*(\bar{d}_{K-1}), Y^*(d)\}$ such that for any $\omega \in \Omega$ with $S_1(\omega) = s_1$,

$$
\begin{aligned}
d_1(s_1) &= u_1, \\
S_2^*(d_1)(\omega) &= S_2^*(u_1)(\omega) = s_2, \\
d_2(\bar{s}_2, u_1) &= u_2, \\
&\vdots \\
d_{K-1}(\bar{s}_{K-1}, \bar{u}_{K-2}) &= u_{K-1}, \\
S_K^*(\bar{d}_{K-1})(\omega) &= S_K^*(\bar{u}_{K-1})(\omega) = s_K, \\
d_K(\bar{s}_K, \bar{u}_{K-1}) &= u_K, \\
Y^*(d)(\omega) &= Y^*(\bar{u}_K)(\omega) = y
\end{aligned}
$$

Optimal treatment regime

- With the potential outcomes associated with a DTR d defined, we can now define an optimal treatment regime. An optimal regime, $d^{\text{opt}} \in \mathcal{D}$, satisfies

$$E\{Y^*(d)|S_1 = s_1\} \leq E\{Y^*(d^{\text{opt}})|S_1 = s_1\}$$

  for all $d \in \mathcal{D}$ and all $s_1 \in \mathcal{S}_1$.
- Note that the definition of an optimal regime depends on the particular set Ψ, and hence the class $\mathcal{D}$, of interest.
- This definition of the optimal DTR mirrors the single-stage definition. The subtle difference is the slightly more complicated definition of potential outcome $Y^*(d)$.

Outline

Backward induction

- So far we have set up the multi-stage setting and defined the optimal DTR in terms of potential outcomes.
- Next, we will describe how d^{opt} is determined via backward induction (dynamic programming).
- Later we will consider identification and estimation.

Backward induction (cont.)

At the Kth decision point, for any $\bar{s}_K \in \bar{\mathcal{S}}_K, \bar{a}_{K-1} \in \bar{\mathcal{A}}_{K-1}$ for which $(\bar{s}_K, \bar{a}_{K-1}) \in \Gamma_K$, define

$$d_K^{(1)\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \underset{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})}{\text{argmax}} E\{Y^*(\bar{a}_{K-1}, a_K) | \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\} \tag{1}$$

$$V_K^{(1)}(\bar{s}_K, \bar{a}_{K-1}) = \underset{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})}{\max} E\{Y^*(\bar{a}_{K-1}, a_K) | \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\} \tag{2}$$

Backward induction (cont.)

For $k = K-1, \ldots, 1$, and any $\bar{s}_k \in \bar{\mathcal{S}}_k$, $\bar{a}_{k-1} \in \bar{\mathcal{A}}_{k-1}$ for which $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$, let

$$d_k^{(1)\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = \underset{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})}{\text{argmax}} E[V_{k+1}^{(1)}\{\bar{s}_k, \bar{S}_{k+1}^*(\bar{a}_{k-1}, \bar{a}_k), \bar{a}_{k-1}, a_k\} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k] \quad (3)$$

$$V_k^{(1)}(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} E[V_{k+1}^{(1)}\{\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k]. \quad (4)$$

Backward induction (cont.)

- $d^{(1)\text{opt}} = (d_1^{(1)\text{opt}}, \ldots, d_K^{(1)\text{opt}})$ is a treatment regime.
- The superscript (1) indicates that $d^{(1)\text{opt}}$ provides K rules for a patient presenting prior to decision 1 with baseline information $S_1 = s_1$.
- We have expressed the optimal regime in terms of potential outcomes. Now we will consider identifiability assumptions that will allow us to express the optimal regime in terms of the observed data.
- Under the identifiability conditions, the conditional expectations will be well-defined.

Identification in the multi-stage setting

We now consider when can the DTR be identified.

- Observed data: iid time-ordered random variables $(S_{1i}, A_{1i}, \ldots, S_{Ki}, A_{Ki}, Y_i)$, $i = 1, \ldots, n$, on Ω.
- Data may arise from an observational study or from an intervention study (e.g. SMART)
- To use the observed data from either type of study, several assumptions are required:
  - Consistency: $S_k = S_k^*(\bar{A}_{k-1})$ for $k = 2, \ldots, K$ and $Y = Y^*(\bar{A}_K)$.
  - SUTVA
  - No unmeasured confounders (aka sequential randomization assumption): A strong version is given by $A_k \perp W^*$ after conditioning on $\{\bar{S}_k, \bar{A}_{k-1}\}$, $k = 1, \ldots, K$. (Recall from slide 8 that W^* is the set of potential outcomes.)

Identification (cont.)

- In a SMART, the sequential randomization assumption is satisfied by design.
- Whether or not it is possible to estimate d^{opt} from the available data is predicated on the treatment options in $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$, being represented in the data.
- In a SMART, Ψ defining the class $\mathcal{D}$ of interest would dictate the design.
- In an observational study, all treatment options in $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ at each decision k must have been assigned to some patients.

Introduction to the Q-functions

Under these assumptions, we can express the optimal regimes in terms of the observed data. We now define the following:

$$Q_K(\bar{s}_K, \bar{a}_K) = E(Y|\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K), \tag{5}$$

$$d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \underset{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})}{\text{argmax}} Q_k(\bar{s}_K, \bar{a}_{K-1}, a_K), \tag{6}$$

$$V_K(\bar{s}_K, \bar{a}_K) = \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} Q_K(\bar{s}_K, \bar{a}_{K-1}, a_K), \tag{7}$$

and for $k = K-1, \ldots, 1$,

$$Q_k(\bar{s}_k, \bar{a}_k) = E\{V_{k+1}(\bar{s}_k, S_{k+1}, \bar{a}_k)|\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k\} \tag{8}$$

$$d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = \underset{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})}{\text{argmax}} Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k) \tag{9}$$

$$V_k(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k). \tag{10}$$

(5) and (8) are called **Q-functions**. (7) and (10) are called **value functions**. It follows that $d_k^{(1)\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1})$ and $V_k^{(1)}(\bar{s}_k, \bar{a}_{k-1}) = V_k(\bar{s}_k, \bar{a}_{k-1})$ for $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$, $k = 1, \ldots, K$.

Outline

Q-learning

- Q- and A-learning are two approaches to estimating d^{opt}.
- They are both recursive fitting algorithms.
- The main distinguishing feature is the form of the underlying models.
- With our optimal treatment regime now defined and written in terms of the observed data, we are ready estimate the optimal treatment regime.

Q-learning (cont.)

- An obvious strategy is to directly model and fit the Q-functions.
- Specifically
  1. Posit models $Q_k(\bar{s}_k, \bar{a}_k; \xi_k)$ for $k = K, \ldots, 1$ where ξ_k is a finite dimensional parameter.
  2. Obtain estimator $\hat{\xi}_K$ by solving suitable estimating equations (e.g. OLS, WLS, etc.) for a patient with past history $\bar{S}_K = \bar{s}_K$ and $\bar{A}_{K-1} = \bar{a}_{K-1}$. Plug into $d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}; \xi_K)$ to estimate d_K^{opt}.
  3. Obtain estimator $\hat{\xi}_{K-1}$ for a patient with past history $\bar{S}_{K-1} = \bar{s}_{K-1}$ and $\bar{A}_{K-2} = \bar{a}_{K-2}$ assuming he or she will take the optimal treatment at decision K (i.e. d_K^{opt}). Plug into $d_{K-1}^{\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}; \xi_{K-1})$ to estimate d_{K-1}^{opt}.
  4. Continue in a backward iterative fashion for each $k = K-2, \ldots, 1$.

Q-learning for two stages

To illustrate the Q-learning algorithm, we will look at the two-stage case.

- Let $K = 2$, $\Psi_1(s_1) = \mathcal{A}_1 = \{0, 1\}$ for all s_1 and $\Psi_2(\bar{s}_2, a_1) = \mathcal{A}_2 = \{0, 1\}$ for all $\bar{s}_2$ and $a_1 \in \{0, 1\}$.
- Let $\mathcal{H}_1 = (1, s_1^\intercal)^\intercal$ and $\mathcal{H}_2 = (1, s_1^\intercal, a_1, s_2^\intercal)^\intercal$.
- We can posit linear models for Q-functions

$$Q_1(s_1, a_1; \xi_1) = \mathcal{H}_1^\intercal \beta_1 + a_1(\mathcal{H}_1 \psi_1),$$
$$Q_2(\bar{s}_2, \bar{a}_2; \xi_2) = \mathcal{H}_2^\intercal \beta_2 + a_2(\mathcal{H}_2^\intercal \psi_2)$$

where $\xi_k = (\beta_k^\intercal, \psi_k^\intercal)^\intercal$, $k = 1, 2$.

Q-learning for two stages (cont.)

- With the posited models for the Q-functions, we have

$$\begin{aligned} V_2(\bar{s}_2, a_1; \xi_2) &= \max_{a_2 \in \{0,1\}} Q_2(\bar{s}_2, a_1, a_2; \xi_2) \\ &= \mathcal{H}_2^\intercal \beta_2 + (\mathcal{H}_2^\intercal \psi_2) \times I(\mathcal{H}_2^\intercal \psi_2 > 0), \text{ and} \\ V_1(s_1; \xi_1) &= \max_{a \in \{0,1\}} Q(s_1, a_1; \xi_1) \\ &= \mathcal{H}_1^\intercal \beta_1 + (\mathcal{H}_1^\intercal \psi_1) I(\mathcal{H}_1^\intercal \psi_1 > 0). \end{aligned}$$

- We can see that

$$\begin{aligned} d_1^{\text{opt}}(s_1; \xi_1) &= I(\mathcal{H}_1^\intercal \psi_1 > 0) \\ d_2^{\text{opt}}(\bar{s}_2, a_1; \xi_2) &= I(\mathcal{H}_2^\intercal \psi_2 > 0). \end{aligned}$$

.

- Thus to estimate the optimal DTR, we only need to estimate the regression coefficients ψ_1 and ψ_2 which can be done via OLS, WLS, etc.

Outline

A-Learning

- Advantage learning, or A-learning, is an alternative to Q-learning that does not require the entire Q-function to be specified to estimate the optimal regime.
- To see this, we continue with the two-stage problem. Notice that d_1^{opt} depends only on $\mathcal{H}_1^{\mathsf{T}}\psi_1 = Q_1(s_1, 1; \xi_1) - Q_1(s_1, 0; \xi_1)$.
- Similarly, $d_2^{\text{opt}}(\bar{s}_2, a_1; \xi_2)$ depends only on $\mathcal{H}_2^{\mathsf{T}}\psi_2 = Q_2(\bar{s}_2, a_1, 1; \xi_2) - Q_2(\bar{s}_2, a_1, 0; \xi_2)$.
- Thus for obtaining the optimal treatment regime, we only need to know the contrast function $C_k(\bar{s}_k, \bar{a}_{k-1}) = Q_k(\bar{s}_k, \bar{a}_{k-1}, 1) - Q_k(\bar{s}_k, \bar{a}_{k-1}, 0)$.

A-learning (cont.)

- In the two treatment case, the contrast function is referred to as the optimal-blip-to-zero function (Robbins, 2004).
- Susan Murphy (2003) considers the expression $C_k(\bar{S}_k, \bar{A}_{k-1})[I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k]$. This is referred to as the advantage or regret function.

A-learning algorithm

A-learning proceeds as follows

- Posit models $C_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$, $k = 1, \ldots, K$ for the contrast functions, depending on parameters ψ_k.
- Let $\pi_K(\bar{s}_K, \bar{a}_{K-1}) = \text{pr}(A_K = 1 | \bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1})$ be the propensity of receiving treatment 1 in the observed data as a function of past history.
- Let $\tilde{V}_{(K+1)i} = Y_i$.
- Robbins (2004) showed that all consistent and asymptotically normal estimators for ψ_K are solutions to estimating equations of the form

$$\sum_{i=1}^{n} \lambda_K(\bar{S}_{Ki}, \bar{A}_{(k-1)i}) \{A_{Ki} - \pi_K(\bar{S}_{ki}, \bar{A}_{(K-1)i})\}$$
$$\times \{\tilde{V}_{(K+1)i} - A_{Ki} C_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i}; \psi_K)$$
$$- \theta_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i})\} = 0$$

for arbitrary functions $\lambda_K(\bar{s}_K, \bar{a}_{K-1})$ of the same dimension as ψ_K and arbitrary functions $\theta_K(\bar{s}_K, \bar{a}_{K-1})$.

A-learning algorithm (cont.)

- If $C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K)$ is correct and $\text{var}(Y|\bar{S}_K = s_k, \bar{A}_{K-1} = a_{K-1})$ is constant, the optimal choices of λ_K and θ_K are given by

$$\lambda_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K) = \frac{\partial C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K)}{\partial \psi_K}, \text{ and}$$

$$\theta_K(\bar{s}_{Ki}, \bar{a}_{(K-1)i}) = h_K(\bar{s}_K, \bar{a}_{K-1}) = Q_K(\bar{s}_K, \bar{a}_{K-1}, 0).$$

- To estimate the optimal DTR, we can posit models for λ_K and θ_K and substitute them into the estimating equation to obtain $\hat{\psi}_K$.
- We then have $d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}; \psi_K) = I\{C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K) > 0\}$ which can be estimated by plugging in our estimator for ψ_K.

A-learning algorithm (cont.)

- For $k = K-1, \ldots, 1$, proceed in a backward iterative fashion to yield $\hat{\psi}_k$ by solving analogous estimating equations:

$$\sum_{i=1}^{n} \lambda_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k)\{A_{ki} - \pi_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \phi_k)\}$$
$$\times \{\tilde{V}_{(k+1)i} - A_{ki} C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k)$$
$$- h_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k)\} = 0$$
$$\sum_{i=1}^{n} \frac{\partial h_k(\bar{S}_k), \bar{A}_{k-1}; \beta_k}{\partial \beta_k} \times \{\tilde{V}_{(k+1)i} - A_{ki} C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k)$$
$$- h_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k)\} = 0$$

  where $\tilde{V}_{ki} = \tilde{V}_{(k+1)i} + C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \hat{\psi}_k)[I\{C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \hat{\psi}_k) > 0\} - A_{ki}]$.

- It follows that $\hat{d}_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}; \hat{\psi}_k) = I\{C_k(\bar{s}_k, \bar{a}_{k-1}; \hat{\psi}_k) > 0\}$

Outline

Comparing Q- and A-learning

- Q- and A-learning are both approaches to solving dynamic programming problems using backward induction reasoning. The fundamental difference is what is modeled in order to estimate the optimal policy.
- When the Q-functions are correctly specified, Q-learning may be more efficient than A-learning.
- On the other hand, when the Q-functions are misspecified, A-learning may provide some robustness against such misspecification.
- Q-learning may have practical advantages in the sense that the modeling task involved is familiar to most data analysts.
- A-learning may be preferred in settings where it is expected that the form of the decision rules defining the optimal regime is not overly complex.