ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Q-Learning: Theory and Applications

Jesse Clifton and Eric Laber

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, USA; email: eblaber@ncsu.edu

Annu. Rev. Stat. Appl. 2020. 7:279–301

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041220

Keywords

reinforcement learning, dynamic treatment regimes, model-free, causal inference, policy search

Abstract

Q-learning, originally an incremental algorithm for estimating an optimal decision strategy in an infinite-horizon decision problem, now refers to a general class of reinforcement learning methods widely used in statistics and artificial intelligence. In the context of personalized medicine, finite-horizon Q-learning is the workhorse for estimating optimal treatment strategies, known as treatment regimes. Infinite-horizon Q-learning is also increasingly relevant in the growing field of mobile health. In computer science, Q-learning methods have achieved remarkable performance in domains such as game-playing and robotics. In this article, we (*a*) review the history of Q-learning in computer science and statistics, (*b*) formalize finite-horizon Q-learning within the potential outcomes framework and discuss the inferential difficulties for which it is infamous, and (*c*) review variants of infinite-horizon Q-learning and the exploration-exploitation problem, which arises in decision problems with a long time horizon. We close by discussing issues arising with the use of Q-learning in practice, including arguments for combining Q-learning with direct-search methods; sample size considerations for sequential, multiple assignment randomized trials; and possibilities for combining Q-learning with model-based methods.

1. INTRODUCTION

Optimal sequential decision-making has recently seen a wave of interest across a number of disciplines. In the statistical literature, the problem of making optimal sequences of decisions arises in the context of personalized medicine, and in particular the estimation of dynamic treatment regimes from multistage experimental or observational data (Robins 1986, Murphy 2003), as well as the booming field of mobile health (mHealth), in which high-intensity patient longitudinal data collected through a mobile device are providing new opportunities for designing patient-specific treatment regimes over long time horizons (Tewari & Murphy 2017, Ertefaie & Strawderman 2018, Luckett et al. 2019). Optimal sequential decision-making has long been studied in control theory and computer science (Bellman 1957, Bertsekas & Tsitsiklis 1996, Sutton & Barto 2018) but has seen an explosion of interest with the advent of deep learning and subsequent high-profile successes of reinforcement learning in complex games like chess, go, and poker (Bowling et al. 2015, Mnih et al. 2015, Silver et al. 2017).

The goal of sequential decision-making is to take actions so as to maximize some measure of expected cumulative utility. Utility may correspond to a composite measure of patient well-being in medical contexts, the number of infections in the context of controlling the spread of an infectious disease (Laber et al. 2018), or the points earned in a game (Mnih et al. 2015). One approach to identifying an optimal decision strategy in a sequential decision setting is to estimate the optimal Q-function, which measures the expected cumulative utility of each currently available decision, given that the decision maker will follow the optimal decision strategy in the future. We focus on the class of methods for optimal sequential decision-making that follow this approach. We generically refer to this class of methods as Q-learning.

We begin by formally introducing sequential decision problems and giving a history of Q-learning in Section 2. In Section 3, we discuss the theory of finite-horizon Q-learning, including the issues of bias and nonregularity, the implementation of Q-learning with flexible models, and connections between Q-learning and another class of methods for sequential decision-making called direct search. In Section 4, we discuss approaches to estimating optimal policies in infinite-horizon settings using Q-functions. In Section 5, we explore issues arising in the practical application of Q-learning methods, including a discussion of our preferred approach to the estimation of optimal policies via combining Q-learning and direct search methods. We also present a brief review of recent advances in sample size calculations aimed at sizing sequential, multiple assignment randomized trials (SMARTs) (Murphy 2005b) for the estimation of high-quality treatment regimes via Q-learning, and some remarks on prospects for combining Q-learning with model-based reinforcement learning methods.

2. HISTORY OF Q-LEARNING IN STATISTICS

2.1. Setup and Background

In this section we trace two threads running through the history of reinforcement learning up to contemporary uses of Q-learning in statistics. The first is dynamic programming, which was introduced by Bellman (1957) and later adapted to complex environments with unknown transition dynamics via approximate dynamic programming with function approximators (Bertsekas & Tsitsiklis 1996, Ernst et al. 2005, Powell 2007). The second is the original Q-learning algorithm introduced by Watkins & Dayan (1992), which is an incremental (stochastic approximation) method for estimating the Q-function (defined below) in a Markov decision process (MDP) (Puterman 2009). The work of Murphy (2005a) provides a bridge from these methods, previously confined to the computer science and control theory literatures, to the application of Q-learning to the estimation of treatment regimes in statistics.

Informally, a sequential decision problem is characterized by a (generally stochastic) environment whose context evolves according to transition dynamics that depend on the history of the system as well as the actions (decisions) taken by an autonomous decision maker. Contexts are associated with different utilities. Thus, the decision maker's task is to choose actions that make sequences of high-utility contexts likely; in particular, the decision maker wants to maximize some measure of cumulative expected utility. In medical settings, for instance, the context corresponds to the values of medically relevant patient characteristics, the available decisions are treatments that may be given or recommended to the patient, and the transition dynamics correspond to the biological processes governing how a patient in a given medical state responds to the treatment.

Formally, let $T \in \mathbb{N} \cup \{\infty\}$ be the time horizon (i.e., the number of decision points) in the decision problem, and let $t = 1, \ldots, T$ index these decision points. The domain of the decision problem is defined by:

- $\mathcal{X}_t$: the space of possible decision contexts at time t
- $\mathcal{A}_t$: the space of possible actions (decisions) at time t
- $u_t : \mathcal{X}_t \to \mathbb{R}$: a utility function that measures the immediate value of each context at time t

Define, for each t, the observed utility $Y_t \triangleq u_t(X_t)$ and the history $H_t \triangleq \{X_1, A_1, \ldots, X_t\}$, which belongs to the space of possible histories $\mathcal{H}_t \triangleq \mathcal{X}_1 \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_{t-1} \times \mathcal{X}_t$. Further defining the decision problem are

- $\Psi_t : \mathcal{H}_t \to 2^{\mathcal{A}_t}$: feasible sets that map histories to sets of allowable actions (van der Laan et al. 2005)
- $P_t : \mathcal{X}_{t+1} \times \mathcal{H}_t \times \mathcal{A}_t \to [0, 1]$: transition functions giving the probability distribution over possible states, i.e., $P(\cdot, h_t, a_t)$ is a distribution over next states $X_{t+1} \in \mathcal{X}_{t+1}$ given history $H_t = h_t$ and action $A_t = a_t$
- $\gamma \in (0, 1]$: a discount factor that is required to be less than 1 in infinite-horizon settings to ensure the convergence of the cumulative reward (see Equation 1 below)

For concreteness, consider a two-stage trial ($T = 2$) in which one of two treatments are given at each stage ($\mathcal{A}_1 = \mathcal{A}_2 = \{0, 1\}$). Patient covariates in spaces $\mathcal{X}_1, \mathcal{X}_2$ are measured at baseline and after stage 1, respectively, and an outcome $Y \equiv Y_3$ is measured at the end of the trial. The patients' covariates X_2 and outcomes Y_3 given their treatments and histories at times 1 and 2, respectively, are governed by (probably unknown) transition distributions P_1, P_2.

Define a decision strategy, commonly known as a policy in the computer science literature, to be a sequence of decision rules $\pi \equiv \{\pi_t\}_{t=1}^{T}$, with $\pi_t : \mathcal{H}_t \to \Delta(\mathcal{A}_t)$, where $\Delta(\mathcal{A}_t)$ is the set of probability measures on the action space $\mathcal{A}_t$. Much of the biomedical literature focuses on deterministic decision rules, but stochastic rules are critical in online settings as they both provide a mechanism for exploration (more on this topic in Section 4.2) and allow for the use of gradient-based optimization methods. Define the value of starting in context x under decision strategy π to be

$$V^{\pi}(x) \triangleq \mathbb{E}^{\pi}\left(\sum_{t=1}^{T} \gamma^{t-1} Y_t \mid X_1 = x\right), \qquad 1.$$

where $\mathbb{E}^{\pi}$ denotes the expectation over trajectories in which π is followed at each time step (and we require $\gamma < 1$ if $T = \infty$). We call V^{π} the V-function under π (it is also known as the state-value function; Sutton & Barto 2018, chapter 3). Then, letting $X_1 \sim \nu$ for some initial state distribution ν (which may simply be a point mass $\nu \equiv \delta_{x_1}$ for a known initial state x_1), define the value of decision strategy π to be $V(\pi) \triangleq \int V^{\pi}(x)\mathrm{d}\nu(x)$. We call V the value function. Finally, define the optimal decision strategy in a class Π as

$$\pi^{\text{opt}} \triangleq \arg\max_{\pi\in\Pi} V(\pi); \tag{2.}$$

we assume for simplicity that there is a unique optimal decision strategy in the class Π.

Dynamic programming, in the context of reinforcement learning, refers to a class of algorithms for estimating the optimal policy by taking advantage of the recursive structure of sequential decision problems. In finite-horizon decision problems, the canonical dynamic programming algorithm is backward induction. Backward induction may be defined in terms of the optimal Q-functions, which measure the expected cumulative reward given a particular context-decision pair. For $T < \infty$, define the optimal Q-functions recursively, as follows:

$$\begin{aligned} Q_T^{\text{opt}}(h_T, a_T) &\triangleq \mathbb{E}(Y_T \mid H_T = h_T, A_T = a_T); \\ Q_t^{\text{opt}}(h_t, a_t) &\triangleq \mathbb{E}\left[Y_t + \gamma \max_{a\in\Psi(H_{t+1})} Q_{t+1}(X_{t+1}, a) \mid H_t = h_t, A_t = a_t\right]. \end{aligned} \tag{3.}$$

It follows that for $t = 1, \ldots, T$, $\pi_t^{\text{opt}}(h_t) \triangleq \arg\max_{a\in\Psi(h_t)} Q_t(h_t, a)$. Backward induction is the basis of finite-horizon Q-learning in statistics, which we present below and discuss further in Section 3.

In the infinite-horizon setting, estimation of an optimal strategy requires either solving a finite-horizon approximation by backward induction or imposing additional structure that allows for the computation of the optimal infinite-horizon policy. A common framework for infinite-horizon problems is the (time-homogeneous) MDP, which imposes the following additional assumptions:

- The context spaces, decision spaces, and utility functions are constant, with $\mathcal{X}_t \equiv \mathcal{X}, \mathcal{A}_t \equiv \mathcal{A}$, and $u_t \equiv u$ for $t = 1, \ldots, T$;
- the transition dynamics are Markovian and time-homogenous, i.e., $P_t \equiv P$ for all t, with $P(\cdot \mid H_t, A_t) = P(\cdot \mid X_t, A_t)$; and
- for simplicity, we will also assume that $\Psi_t \equiv \mathcal{A}$ for each t, though this is not essential.

The time-homogenous Markovian structure allows us to define time-homogeneous optimal V- and Q-functions, as opposed to a sequence at functions, one per time step. These are:

$$\begin{aligned} V^{\text{opt}}(x) &\triangleq \mathbb{E}^{\pi^{\text{opt}}}\left(\sum_{t=1}^{\infty} \gamma^{t-1} Y_t \mid X_1 = x\right) \\ Q^{\text{opt}}(x, a) &\triangleq \mathbb{E}^{\pi^{\text{opt}}}\left(\sum_{t=1}^{\infty} \gamma^{t-1} Y_t \mid X_1 = x, A_1 = a\right). \end{aligned} \tag{4.}$$

We now state several well-known properties of the optimal Q-function in the MDP framework, which will later allow us to construct estimators of the optimal policy. We first introduce the Bellman optimality operator B, which acts on functions f as

$$(Bf)(x, a) \triangleq \mathbb{E}\Big[Y_t + \gamma \max_{a'\in\mathcal{A}} f(X_{t+1}, a') \mid X_t = x, A_t = a\Big]. \tag{5.}$$

From Equation 4 we can see that the optimal Q-function satisfies

$$\begin{aligned} Q^{\mathrm{opt}}(x,a) &= \mathbb{E}^{\pi^{\mathrm{opt}}}\left(\sum_{t=1}^{\infty} \gamma^{t-1} Y_t \mid X_1 = x, A_1 = a\right) \\ &= \mathbb{E}\left[Y_0 + \gamma \max_{a'} \mathbb{E}^{\pi^{\mathrm{opt}}}\left(\sum_{t=1} \gamma^{t-1} Y_t \mid X_1, A_1 = a'\right) \mid X_1 = x, A_1 = a\right] \\ &= \mathbb{E}^{\pi^{\mathrm{opt}}}\left[Y_1 + \gamma \max_{a'} Q^{\mathrm{opt}}(X_1, a') \mid X_1 = x_1, A_1 = a_1\right] \\ &= (BQ^{\mathrm{opt}})(x,a). \end{aligned} \qquad 6.$$

The property $Q^{\mathrm{opt}} = BQ^{\mathrm{opt}}$ is called the Bellman optimality equation. A consequence is that Q^{opt} is the unique fixed point of the so-called value iteration algorithm:

```
Algorithm 1 (Value iteration for computing $Q^{\mathrm{opt}}$).
Initialize $Q_0$
for k=0, 1, ... until convergence do
    $Q_{k+1} \leftarrow BQ_k$
```

The Bellman equation and value iteration are used to construct estimators of the optimal infinite-horizon Q-function in Section 4.

As a final remark on the prehistory of Q-learning in computer science, the name Q-learning originates with the incremental online algorithm of Watkins & Dayan (1992) for estimating the Q-function. Assuming finite context and decision spaces, for each newly encountered tuple of observations (X_t, A_t, Y_t, X_{t+1}), the algorithm executes the following updates: $Q_{k+1}(X_t, A_t) \leftarrow Q_k(X_t, A_t) + \alpha\{Y_t + \gamma \max_{a' \in \mathcal{A}} Q_k(X_{t+1}, a') - Q_k(X_t, A_t)\}$ (for a step-size α). Variants of Watkins's Q-learning for the case of large context spaces where Q must be estimated using function approximation have since been proposed and refined (Baird 1995, Precup et al. 2001, Maei et al. 2010).

2.1.1. A partial taxonomy of reinforcement learning methods.

Given that the transition distributions P_t are unknown in most applications, the optimal policy must usually be estimated by either estimating the value function V and optimizing this estimator over a class of decision strategies, or estimating $\{Q_t^{\mathrm{opt}}\}_{t=1}^{T}$ and taking the argmax of the resulting estimators as the estimated optimal policy. These estimators may be obtained in either a model-free or model-based manner, where "model" here refers to a model of the transition dynamics $\{P_t\}_{t=1}^{T}$. Model-free methods directly construct an estimator of V or $\{Q_t^{\mathrm{opt}}\}_{t=1}^{T}$, bypassing the estimation of the transition dynamics, while model-based methods first estimate P_t and subsequently obtain an estimator of the optimal policy (Sutton & Barto 2018, chapter 8). **Table 1** displays examples of algorithms in each of these categories, including those discussed throughout this article. This is far from an exhaustive taxonomy—methods also differ as to whether they are online versus offline, batch versus incremental, on-policy versus off-policy, and so on, and moreover may be hybrids of these categories (see Sections 3.4, 5.1, and 5.3).

Note that Q-learning usually refers to the model-free estimation of the optimal Q-functions, although as model-based dynamic programming may involve the estimation of the optimal Q-functions, one can also speak of model-based Q-learning, as in **Table 1**. But in this article we will focus on Q-learning as it is generally understood, i.e., in the model-free sense.

Table 1 Examples of reinforcement learning algorithms classified according to model-free versus model-based and Q-learning versus policy search, emphasizing methods from the statistical literature

Type	Q-learning	Policy search
Model-free	■ Watkins's Q-learning (Watkins & Dayan 1992; Section 2.1) ■ Finite-horizon Q-learning (Murphy 2005a; Sections 2.2 and 3) ■ Batch infinite-horizon Q-learning (Ertefaie & Strawderman 2018; Section 4.1) ■ Fitted Q-iteration (Ernst et al. 2005; Section 4.1) ■ Deep Q-networks (Mnih et al. 2015; Section 4.4)	■ Robust policy search for DTR (Zhang et al. 2013; Section 3.4) ■ Doubly robust off-policy evaluation (Jiang & Li 2015; Section 4.3) ■ V-learning (Luckett et al. 2019; Section 4.3) ■ Proximal policy optimization (Schulman et al. 2017; Section 4.4)
Model-based	Model-based dynamic programming (Bellman 1957, Bertsekas & Tsitsiklis 1996)	Spatial policy search (Laber et al. 2018)

Abbreviation: DTR, dynamic treatment regime.

2.2. Q-Learning in Statistics

The first methods for the estimation of the value of a policy in a general finite-horizon sequential decision problem were developed in the landmark papers of Robins (1986, 1987, 1989, 1993), Murphy et al. (2001), and Murphy (2003). Consider the problem of estimating the expected potential outcome of a terminal scalar utility Y (measured after T time points) under a policy π, denoted $\mathbb{E}Y^*(\pi)$. Define, for a sequence of variables Z_t, the history $\overline{Z}_t \triangleq \{Z_v\}_{v=1}^t$. Under sequential versions of standard causal inference assumptions (Section 3.1; call these CA) $\mathbb{E}Y^*(\pi)$ may be identified with $V(\pi)$ via the g-formula (assuming for simplicity that covariates at each time point are discrete):

$$\begin{aligned}\mathbb{E}Y^*(\pi) &\overset{\text{CA}}{=} V(\pi)\\ &= \sum_{\overline{x}^t} \left\{\mathbb{E}\{Y \mid \overline{A}_T = [\pi_t(h_t)]_{t=1}^T, \overline{X}^t = (x_t)_{t=1}^T\right\}\\ &\prod_{t=1}^T P\left\{X_t = x_t \mid \overline{A}_{t-1} = [\pi(h_v)]_{v=1}^{t-1}, \overline{X}_{t-1} = \overline{x}_{t-1}\}\right\}.\end{aligned} \qquad 7.$$

The seminal work of Murphy (2003, 2005a) married dynamic programming for optimal policy estimation with the potential outcomes framework (Rubin 1978, Robins 1986) and standard regression methods in what is now known as Q-learning in the statistical literature. Working within this framework, Murphy (2003) showed how the optimal regime may be estimated via the recursive estimation of the regret functions, $\mu_t(h, a) = \max_{a'} Q_t^{\text{opt}}(h_t, a') - Q_t^{\text{opt}}(h_t, a_t)$, thereby bypassing the estimation of the distributions P_t. Murphy (2005a) then introduced what she terms batch Q-learning (from now on, just Q-learning), defined as follows. Suppose we have data from trajectories $i = 1, \ldots, n$ of length $T \in \mathbb{N}$, the ith trajectory consisting of observations $\{X_1^i, A_1^i, Y_1^i, \ldots, X_T^i, A_T^i, Y_T^i\}$. Given a sequence of model classes $\{\mathcal{Q}_t\}_{t=1}^T$ for estimating the optimal Q-functions, Q-learning mimics the backward induction procedure of Equation 3 like so:

$$\begin{aligned}\widehat{Q}_T &= \underset{Q_T \in \mathcal{Q}_T}{\arg\min}\, \mathbb{P}_n[Q_T(H_T, A_T) - Y_T]^2;\\ \widehat{Q}_t &= \underset{Q_t \in \mathcal{Q}_t}{\arg\min}\, \mathbb{P}_n \left\{Q_t(H_t, A_t) - \left[Y_t + \max_{a_{t+1} \in \Psi_{t+1}(H_{t+1})} \widehat{Q}_{t+1}(H_{t+1}, a_{t+1})\right]\right\}^2.\end{aligned} \qquad 8.$$

Robins (2004) introduced a related semiparametric estimator of the optimal policy, but this method has not been used as widely as Q-learning (Vansteelandt & Joffe 2014), perhaps because of its relative complexity. Moodie et al. (2007) provide a discussion of the connections between the methods of Robins (2004) and Murphy (2003).

The max operator in Equation 8 causes difficulties for statistical inference that have been the focus of much research on Q-learning since its introduction to statistics (Chakraborty et al. 2010, 2013; Moodie et al. 2010; Laber et al. 2014b; Song et al. 2015). In addition to problems of statistical inference, developments for finite Q-learning in statistics have included extensions to nonlinear model classes (Zhao et al. 2009, Laber et al. 2014a, Moodie et al. 2014, Zhang et al. 2015, Xu et al. 2016, Zhou & Kosorok 2017), as well as methods for combining Q-learning with policy search methods (Zhang et al. 2012, Zhang et al. 2015). We review each of these in Section 3. Moreover, as statisticians have become increasingly interested in long-horizon problems such as those arising in mHealth and spatial-temporal decision-making, infinite-horizon analogs of the methods reviewed in Section 3 have been developed; we discuss these in Section 4.

3. FINITE-HORIZON Q-LEARNING

For simplicity, we focus on Q-learning in two-stage problems (i.e., in which $T = 2$), but the discussion generalizes to any finite T. The statistical literature on finite-horizon Q-learning is largely concerned with the estimation of optimal decision strategies in a medical context; the data sets used for estimation are often obtained from a multistage clinical trials (see Section 5.2) or observational longitudinal studies. In these settings, the actions are treatments, states are covariates, utilities are outcomes, policies are typically referred to as dynamic treatment regimes, and each trajectory from the decision process corresponds to the complete history of a single patient. Given the prevalence of the usage of finite-horizon Q-learning in medical contexts, we adopt this terminology in this section. **Table 2** serves as a reference for the correspondence between the respective terminologies.

3.1. Backward Induction in the Potential Outcomes Framework

A patient's baseline covariates (measured before stage 1) are denoted $X_1 \in \mathbb{R}^{p_1}$, and $X_2 \in \mathbb{R}^{p_2}$ are time-varying covariates measured after the first treatment and before the second. Treatments are given at times 1 and 2, denoted respectively by $A_1, A_2 \in \{0, 1\}$. The outcome $Y \in \mathbb{R}$ is observed after time point 2 and is the quantity whose expectation we wish to maximize. Finally, we define two history variables, $H_1 = X_1$ and $H_2 = (X_1, A_1, X_2)$, which collect the observations available at times 1 and 2, respectively.

Let $X_2^*(a_1)$ be the potential outcome (Splawa-Neyman et al. 1990; Rubin 1978, 2005; Hernán & Robins 2019) of covariate X_2 if treatment a_1 had been given at time 1, and let $Y^*(a_1, a_2)$ be the

Table 2 Translation between terminology in the dynamic treatment regime and reinforcement learning literatures

Generic	Reinforcement learning	Dynamic treatment regimes	Symbol
Decision	Action	Treatment	A, a
Context	State	Covariate	X, x
Utility	Reward	Outcome	Y, y
Observation unit	Trajectory	Patient	Indexed by i
Time index	Time (or time point)	Stage	Indexed by t
Decision strategy	Policy	(Dynamic) treatment regime	π

potential outcome of the response if treatments a_1 and a_2 had been given at times 1 and 2. The expected response for a regime $\pi = (\pi_1, \pi_2)$ is

$$\mathbb{E}\, Y^*(\pi) \triangleq \mathbb{E}\, Y^*\{\pi_1(H_1), \pi_2(H_2)\}. \tag{9.}$$

In order to be able to estimate the causal effects of candidate regimes from observed data, we need to be able to write Equation 9 in terms of observables only, rather than potential outcomes. This is possible under the following, now standard, causal inference assumptions:

- **(CA 1)** Consistency: $X_2^*(a_1) = X_2$ when a_1 is actually received, and $Y^*(a_1, a_2) = Y$ whenever (a_1, a_2) are actually received.
- **(CA 2)** No unmeasured confounders (a.k.a. sequential ignorability): For any sequence of treatments (a_1, a_2), $A_1 \perp\!\!\!\perp \{X_2^*(a_1), Y^*(a_1, a_2)\} \mid H_1$ and $A_2 \perp\!\!\!\perp Y^*(a_1, a_2) \mid H_2$.
- **(CA 3)** Positivity: $P\{b_1[A_1 = \pi_1(H_1) \mid H_1]b_2[A_2 = \pi_2(H_2) \mid H_2] > 0\} = 1$ for all $\pi \in \Pi$, where P is the joint data-generating distribution, and b_1, b_2 give the conditional probability distributions over actions under the behavior (data-generating) regime given histories at stages 1 and 2, respectively.

Under CA 1–3, Robins's g-computation formula (Robins 1986, 1987) allows us to write the expected potential outcome of a regime in terms of observables:

$$\begin{aligned}\mathbb{E}\, Y^*(\pi) &\overset{\text{CA}}{=} \mathbb{E}\{\mathbb{E}[\mathbb{E}(Y \mid X_1, A_1, X_2, A_2 = \pi_2(H_2)) \mid X_1, A_1 = \pi_1(H_1)]\} \\ &= V(\pi).\end{aligned} \tag{10.}$$

Note that the most troublesome assumption in off-policy, observational settings, that of no unmeasured confounders, is often met automatically in engineering and computer science applications where actions are all taken by the same agent attempting to learn an optimal policy from the data. That is, because the process by which actions are taken is known exactly, there is no question of the confounding of the effects of actions with unmeasured variables. But in medical and epidemiological settings, the focus of most statisticians working in sequential decision-making, the problem of confounding looms large, and it is necessary to make it explicit. [There is only a small literature on the development of methods for the estimation of optimal regimes from potentially confounded observations. Kallus & Zhou (2018) recently presented confounding-robust policy improvement, a direct-search method that parameterizes the degree of possible confounding and attempts to find the policy with the best worst-case value for parameters in a neighborhood of the no-confounding case.]

Writing the g-formula as a repeated expectation (Equation 10) suggests the optimal regime, $\pi^{\text{opt}} = \arg\max_{\pi \in \Pi} \mathbb{E}\, Y^*(\pi)$, can be computed with the backward induction strategy introduced in Section 2. Recall the definition of the optimal Q-functions (Equation 3). In view of the foregoing, assuming $\mathbb{E}(|Y| \mid H_t)$ is bounded almost surely and that the action space is finite, we have that π^{opt} is given by $\pi_t^{\text{opt}}(h_t) = \arg\max_{a_t \in \Psi_t(h_t)} Q^{\text{opt}}(h_t, a_t)$ (Murphy 2003).

3.2. Q-Learning with Linear Models

We first consider the case where each Q-function is modeled linearly. We will also consider the simple case of two treatments at each stage, coded as $\mathcal{A}_1 = \mathcal{A}_2 = \{0, 1\}$. Letting $h_t = (h_{t0}^\intercal, \quad h_{t1}^\intercal)^\intercal$ be vectors summarizing the histories at stages $t = 1, 2$, we model the Q-functions as

$$\begin{aligned}Q_1(h_1, a_1; \beta_1) &= h_{10}^\intercal \beta_{01} + a_1 h_{11}^\intercal \beta_{11}, \\ Q_2(h_2, a_2; \beta_2) &= h_{20}^\intercal \beta_{20} + a_2 h_{21}^\intercal \beta_{21},\end{aligned} \tag{11.}$$

with parameters estimated using least squares

$$\begin{aligned}\widehat{\beta}_2 &\triangleq \arg\min_{\beta_2} \mathbb{P}_n\{Y - Q_2(H_2, A_2; \beta_2)\}^2,\\ \widehat{\beta}_1 &\triangleq \arg\min_{\beta_1} \mathbb{P}_n\{\max_{a_2} Q_2(H_2, a_2; \widehat{\beta}_2) - Q_1(H_1, A_1; \beta_1)\}^2,\end{aligned} \tag{12.}$$

where $\mathbb{P}_n$ denotes the empirical expectation over patient trajectories. Then, the estimated optimal dynamic treatment regime is given by

$$\widehat{\pi}_t^{\text{opt}}(h_t) \triangleq \mathbb{1}(h_{t1}^{\intercal}\beta_{t1} > 0).$$

As touched on in Section 2, the estimator $\widehat{\beta}_1$ is nonregular, meaning that it is sensitive to $1/\sqrt{n}$-perturbations of the data-generating model. This creates difficulties for statistical inference. Define the population-level versions of the estimators in Equation 12:

$$\begin{aligned}\beta_2^* &\triangleq \arg\min_{\beta_2} \mathbb{E}\{Q_2(H_2, A_2; \beta_2) - Y\}^2,\\ \beta_1^* &\triangleq \arg\min_{\beta_1} \mathbb{E}\{Q_1(H_1, A_1; \beta_1) - \max_{a_2} Q_2(H_2, a_2; \beta_2^*)\}^2.\end{aligned} \tag{13.}$$

First, $\widehat{\beta}_1$ is asymptotically biased when $\mathrm{P}(H_{21}^{\intercal}\beta_{21}^* = 0) > 0$. That is, the limiting distribution of $\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ need not have mean zero when there is a positive probability of a null effect (Moodie et al. 2010). Second, as a nonregular estimator, it is impossible to uniformly consistently estimate the sampling distribution of $\widehat{\beta}_1$ (Van der Vaart 1991, Andrews 2000, Leeb & Poetscher 2003, Hirano & Porter 2012). In particular, the limiting distribution $\sqrt{n}(\widehat{\beta}_1 - \beta_1^*)$ is normal if $\mathrm{P}(H_{21}^{\intercal}\beta_{21}^* = 0) > 0$ but nonnormal otherwise.

Because inferential problems arise when $h_{21}^{\intercal}\beta_{21}^*$ is close to 0, Chakraborty et al. (2010) propose shrinking or thresholding values of $h_{21}^{\intercal}\widehat{\beta}_{21}$ near 0. Letting $[z]^+$ denote the positive part of a scalar z, define the pseudooutcome under estimator $\widehat{\beta}_2$ as $\widehat{Y}_{\widehat{\beta}_2} \triangleq \max_{a_2} Q_2(h_2, a_2; \widehat{\beta}_2)$, which in turn satisfies $\max_{a_2} Q_2(h_2, a_2; \widehat{\beta}_2) = h_{20}^{\intercal}\widehat{\beta}_{20} + [h_{21}^{\intercal}\widehat{\beta}_{21}]^+$ [from the definition of $Q_2(\cdot; \widehat{\beta}_2)$ and the 0-1 coding of the actions]. Thus, $\widehat{Y}_{\widehat{\beta}_2}$ is the target of the regression problem which defines the second step of Q-learning. Define a hard-thresholded version of the pseudooutcome:

$$\widehat{Y}_{\widehat{\beta}_2}^{HT} \triangleq h_{20}^{\intercal}\widehat{\beta}_{20} + [h_{21}^{\intercal}\widehat{\beta}_{21}]^+ \mathbb{1}(|h_{21}^{\intercal}\widehat{\beta}_{21}| > \lambda), \tag{14.}$$

for some threshold λ. Similarly, define the soft-thresholded version:

$$\widehat{Y}_{\widehat{\beta}_2}^{ST} \triangleq h_{20}^{\intercal}\widehat{\beta}_{20} + [h_{21}^{\intercal}\widehat{\beta}_{21}]^+ \left[1 - \frac{\lambda}{|h_{21}^{\intercal}\widehat{\beta}_{21}|^2}\right]^+. \tag{15.}$$

Given a modified pseudooutcome $\widehat{Y}_{\widehat{\beta}_2}^{\bullet T} \in \{\widehat{Y}_{\widehat{\beta}_2}^{HT}, \widehat{Y}_{\widehat{\beta}_2}^{ST}\}$, one may estimate the Q-function by using the new pseudooutcome as the target of the second step of Equation 12:

$$\widehat{\beta}_1^{\bullet T} \triangleq \arg\min_{\beta_1} \mathbb{P}_n \left\{Q_1(H_1, A_1; \beta_1) - \widehat{Y}_{\widehat{\beta}_2}^{\bullet T}\right\}^2. \tag{16.}$$

In a similar spirit, Song et al. (2015) propose penalized Q-learning. These authors point out that the previously proposed hard- and soft-thresholding methods (*a*) may suffer from large bias in finite samples (therefore leading to bias in the first stage estimator in regular settings), and (*b*) do not in general possess the oracle property that, as the sample size grows, they will perform as well as the estimator that knows the set of patients with unexceptional laws in advance. They instead propose minimizing a penalized version of the objective in the first step of Q-learning:

$$\widehat{\beta}_2^{Pen} \triangleq \arg\min_{\beta_2} \mathbb{P}_n \left\{ [Q(H_2, A_2; \beta_2) - Y]^2 + p_{\lambda_n}(|H_2^T \beta_2^1|) \right\}, \qquad 17.$$

where p_{λ_n} is a penalty function with tuning parameter λ_n. The benefit of this approach, the authors argue, is that it allows for the identification (through the shrinkage of individual subject effects) of subjects whose effects are 0, thereby allowing us to separate exceptional and nonexceptional laws and use standard inferential procedures for each separately. They demonstrate the consistency and asymptotic normality of their first-stage estimator, but only under the assumption that the stage-two histories take only finitely many values. This assumption assures that treatment effects are either identically zero or well-separated from zero, thereby excluding the case of small but nonzero effects in their asymptotic regime. Laber et al. (2014b) showed that shrinkage methods can be arbitrarily worse than standard (unregularized) Q-learning when such small effects are present.

As for inference, Chakraborty et al. (2010) conducted simulation studies using various bootstrapping procedures for their thresholding estimators, though these were not theoretically motivated. Inference for Song et al.'s (2015) penalized Q-learning follows from the estimation of the asymptotic (normal) distribution of the penalized Q-learning estimator. Chakraborty et al. (2014) propose an m-out-of-n bootstrap procedure (a general method for bootstrapping under nonregularity). Bootstrap consistency for certain nonregular parameters can be obtained by constructing bootstrap confidence intervals using small [i.e., of size $m = o(n)$] subsets of the full data. Chakraborty et al. (2014) present an m-out-of-n bootstrap for inference for the estimated Q-function parameters, including several methods for adaptively choosing m.

The foregoing methods were developed under a fixed parameter framework that fails to capture the finite sample instability of Q-learning estimators (Laber et al. 2014b). We now review two approaches designed to overcome these inferential problems using a moving parameter asymptotic framework: the projection interval of Robins (2004), and the adaptive confidence interval of Laber et al. (2014b).

The method of Robins (2004) can be used to construct a projection region for β_1 as follows. Let $\mathbb{I}_{n,\alpha}(\beta_{21})$ be a valid, $(1-\alpha)\cdot 100\%$ confidence region for β_1 when the true second-stage treatment effect parameter is β_{21}; if the latter were known, we could construct such a confidence region using the asymptotic normality of $\widehat{\beta}_1$. However, β_{21} is unknown, so in the projection interval approach one constructs a conservative confidence region for β_1 by first constructing a confidence region for β_{21} and taking the union of regions $\mathbb{I}_{n,\alpha}(\widetilde{\beta}_{21})$ for each $\widetilde{\beta}_{21}$ in the confidence region for β_{21}. The coverage of this procedure is at least $(1-\alpha-\eta)\cdot 100\%$, where η is the coverage of the region constructed for β_{21}.

The approach of Laber et al. (2014b) is instead to bound (for constants $c \in \mathbb{R}^{\dim(\beta_1)}$) the nonregular $c^\intercal \sqrt{n}(\widehat{\beta}_1 - \beta_1)$ between two regular, uniformly convergent upper and lower bounds. Such a procedure is necessarily conservative; however, since nonregularity occurs only for generative models where $h_1^\intercal \beta_{21}$ is near 0, conservatism can be limited by adapting the bound based on evidence as to whether a subject's second-stage effect is near 0. This is accomplished by using a pretest based on comparing $\widehat{\beta}_{21}$ to a test statistic that is large (i.e., diverges to ∞) if and only if $h_1^\intercal \beta_{21}$ is nonzero.

3.3. Q-Learning with Nonlinear Models

The model in Equation 11 is simple and prone to misspecification. In order to address the problem of misspecification in linear Q-learning, Laber et al. (2014a) present interactive Q-learning (IQ-learning). IQ-learning attempts to replace linear Q-learning with an ordinary mean-variance modeling problem, rendering it amenable to the large toolkit of regression architectures and diagnostics. Define the contrast and main-effect functions for Q_2, $\delta(H_2) = \{Q_2(H_2, 1) - Q_2(H_2, 0)\}/2$, $\mu(H_2) = \{Q_2(H_2, 1) + Q_2(H_2, 0)\}/2$. Then, let g_{h_1,a_1} be the conditional distribution of $\delta(H_2)$ given $H_1 = h_1, A_1 = a_1$. Defining $L(h_1, a_1) \triangleq \mathbb{E}\{\mu(H_2) \mid H_1 = h_1, A_1 = a_1\}$, we can then rewrite $Q_1(h_1, a_1)$ as

$$Q_1(h_1, a_1) = L(h_1, a_1) + \int |z| g_{h_1,a_1}(z) \mathrm{d}z. \tag{18.}$$

IQ-learning, then, estimates Q_1 as

$$\widehat{Q}_1^{IQ}(h_1, a_1) \triangleq \widehat{L}(h_1, a_1) + \int |z| \widehat{g}_{h_1,a_1}(z) \mathrm{d}z. \tag{19.}$$

The estimators $\widehat{L}$ and $\widehat{g}_{h_1,a_1}$ used in this construction are obtained using $\widehat{Q}_2$ from the first step of ordinary Q-learning and the corresponding plugin estimators $\widehat{\delta}, \widehat{\mu}$.

With respect to using more flexible classes of candidate Q-functions, alternatives to IQ-learning include maintaining the linear model framework, but with choices of feature functions h_t, which allow for more flexible function classes (radial basis functions, for instance), or carrying out the sequential regression procedure in Equation 8 with a more flexible class of models, at the cost of foregoing the relative theoretical tractability of the linear framework; Moodie et al. (2014), for instance, use a generalized additive model architecture, and Zhao et al. (2009) use support vector regression. Q-learning with flexible models is also closely related to fitted Q-iteration (a form of approximate value iteration; see Section 4.1) with flexible models (Ernst et al. 2005, Riedmiller 2005, Geurts et al. 2006). Finally, in combining Q-learning with policy search for the estimation of interpretable dynamic treatment regimes, Zhang et al. (2016) (see Section 3.4) use a kernel ridge regression estimator for the Q-function.

3.4. Connections with Policy Search Methods

Finally, Q-learning with policy search (QLPS) refers to a class of methods that use elements of both Q-learning and policy search methods. In the latter approach, one estimates π^{opt} by optimizing an estimator of the policy-value function over a predefined class of treatment regimes (policies). That is, one constructs an estimator $\widehat{V}$ of the policy-value function and takes $\widehat{\pi}^{\text{opt}} \triangleq \arg\max_{\pi \in \Pi} \widehat{V}(\pi)$.

Zhang et al. (2013) provide a general framework for policy search methods, which allows for searching over policies corresponding to candidate Q-functions. Consider T-stage decision problems, and introduce coarsening variables $C_{\pi,i}$ (which measure the agreement of patient i's treatments with policy π):

$$C_{\pi,i} \triangleq \begin{cases} 1, \text{ if } A_1^i \neq \pi_1(X_1^i); \\ \infty, \text{ if } A_t^i = \pi_t(H_t^i) \text{ for } t = 1, \dots, T; \\ \max_{k=2,\dots,T-1} \left\{ k : \prod_{t=1}^{k-1} \mathbb{1}[A_t^i = \pi_t(H_t^i)] = 1 \right\}, \text{ otherwise.} \end{cases} \tag{20.}$$

Define discrete hazard functions $\lambda_{\pi,t}$, which give the probability of the behavior policy ceasing to agree with π at stage t for history H_t, given that the behavior policy has agreed with π until t. Then define the survivor function $K_{\pi,t}$:

$$K_{\pi,t}(\overline{X}_t^i) \triangleq \prod_{k=1}^{t} \{1 - \lambda_{\pi,k}(\overline{X}_k^i)\}. \tag{21.}$$

[$K_{\pi,t}$ play a role analogous to propensity scores (Rosenbaum & Rubin 1983) in the evaluation of single-stage treatment regimes, and indeed depend on a model for propensity scores.] If the coarsening mechanism is correctly specified, all regular, asymptotically linear, consistent estimators for $V(\pi)$ are of the form (Tsiatis 2006)

$$\widehat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{1}(C_{\pi,i}=\infty)}{K_{\pi,T}(\overline{X}_T^i)}Y^i + \sum_{t=1}^{T}\frac{\mathbb{1}(C_{\pi,i}=t)-\lambda_{\pi,t}(\overline{X}_t^i)\mathbb{1}(C_{\pi,i}\ge t)}{K_{\pi,t}(\overline{X}_t^i)}\ell_t(X_t^i)\right\} \qquad 22.$$

for arbitrary functions ℓ_t. Varying the choice of ℓ_t leads to different estimators of the value function. The simplest is the inverse probability weighted estimator, obtained by setting $\ell_t \equiv 0$:

$$\widehat{V}^{\mathrm{IPW}}(\pi) \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(C_{\pi,i}=\infty)}{\widehat{K}_{\pi,T}(\overline{X}_T^i)}Y^i. \qquad 23.$$

$\widehat{V}^{\mathrm{IPW}}$ requires the correct specification of the propensity scores in order to be consistent, and it is relatively inefficient. This motivates the doubly robust estimator:

$$\widehat{V}^{\mathrm{DR}}(\pi) \triangleq \widehat{V}^{\mathrm{IPW}} + \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\mathbb{1}(C_{\pi,i}=t)-\widehat{\lambda}_{\pi,t}(\overline{X}_t^i)\mathbb{1}(C_{\pi,i}\ge t)}{\widehat{K}_{\pi,t}(\overline{X}_t^i)}\widehat{m}_{\pi_t}\{H_t^i, [\pi_v(H_v^i)]_{v=1}^t\}. \qquad 24.$$

In Equation 24, $[\pi_v(H_v^i)]_{v=1}^t$ is the sequence of treatments recommended by regime π for patient i at the corresponding observed histories and $\widehat{m}_{\pi_t}$ is an estimator of the function $m_{\pi_t} \triangleq \mathbb{E}\{Y^*(\pi) \mid H_t^*[(\pi_v)_{v=1}^t] = h_t\}$, where $H_t^*[(\pi_v)_{v=1}^t]$ is the potential history at stage t if π is followed until that stage. $\widehat{V}^{\mathrm{DR}}$ is consistent if either the models for the propensity scores are correctly specified or the models for $\widehat{m}_{\pi_t}$ are, and it is semiparametric efficient if both are.

Finally, there is the augmented inverse probability weighted estimator, which replaces $\widehat{m}_{\pi_t}$ in Equation 24 by estimated optimal Q-functions:

$$\widehat{V}^{\mathrm{AIPW}}(\pi) \triangleq \widehat{V}^{\mathrm{IPW}} + \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\mathbb{1}(C_{\pi,i}=t)-\widehat{\lambda}_{\pi,t}(\overline{X}_t^i)\mathbb{1}(C_{\pi,i}\ge t)}{\widehat{K}_{\pi,t}(\overline{X}_t^i)}\widehat{Q}_t\{X_t^i, [\pi_v(H_v^i)]_{v=1}^t\}. \qquad 25.$$

While $\widehat{m}_{\pi_t}$ must be refitted at each step of optimizing $\widehat{V}^{\mathrm{DR}}$ in π, the $\widehat{Q}_t$s used in $\widehat{V}^{\mathrm{AIPW}}$ are constant, thus improving computational efficiency while hopefully reaping most of the efficiency gains of $\widehat{V}^{\mathrm{DR}}$.

Given an estimator $\widehat{V}$ and class of policies Π, one may estimate the optimal policy as $\widehat{\pi} \triangleq \arg\max_{\pi\in\Pi}\widehat{V}(\pi)$. A variant of QLPS is obtained by choosing $\Pi = \{[\arg\max_{a_1} Q_1(\cdot, a_1), \arg\max_{a_2} Q_2(\cdot, a_2)] : (Q_1, Q_2) \in \mathcal{Q}_1 \times \mathcal{Q}_2\}$ for classes of Q-functions $\mathcal{Q}_1, \mathcal{Q}_2$. While the estimated optimal policy will belong to the same class of regimes as Q-learning with model classes $\mathcal{Q}_1, \mathcal{Q}_2$, if the propensity scores are correctly specified then QLPS will be consistent for the best regime in the class; a similar guarantee does not hold for Q-learning (Qian & Murphy 2011). In Section 5.1 we return to discuss some of the benefits of QLPS methods over using either Q-learning or policy search alone.

Zhang et al. (2016) use the following QLPS procedure, for classes of decision rules Π_1, Π_2:

$$\begin{aligned}
\widehat{Q}_2 &\leftarrow \arg\min_{Q_2 \in \mathcal{Q}_2} \mathbb{P}_n\{Y - Q_2(H_2, A_2)\}^2;\\
\widehat{\pi}_2 &\leftarrow \arg\max_{\pi_2 \in \Pi_2} \mathbb{P}_n \widehat{Q}_2\{H_2, \pi_2(H_2)\};\\
\widehat{Q}_1 &\leftarrow \arg\min_{Q_1 \in \mathcal{Q}_1} \mathbb{P}_n\{\widehat{Q}_2[H_2, \widehat{\pi}_2(H_2)] - Q_1(H_1, A_1)\}^2;\\
\widehat{\pi}_1 &\leftarrow \arg\max_{\pi_1 \in \Pi_1} \mathbb{P}_n \widehat{Q}_1\{H_1, \pi_1(H_1)\}.
\end{aligned} \qquad 26.$$

Zhang et al. (2016) choose Π_1, Π_2 to be classes of decision lists in order to enforce parsimony and interpretability in the estimated policy.

4. INFINITE-HORIZON Q-LEARNING

While Q-learning in statistics originally arose in the context of optimal finite-horizon treatment regimes, there is growing interest among statisticians in decision problems with long horizons. In such problems, data efficiency typically takes precedence over computational efficiency (in contrast to many test-bed problems used in infinite-horizon reinforcement learning, e.g., game-playing and robotics, where data efficiency may be secondary to the ability of an algorithm to make split-second decisions). Applications of this kind include the estimation of optimal treatment strategies to be administered via mHealth technologies (Ertefaie & Strawderman 2018, Luckett et al. 2018) and allocating limited resources in complex spatio-temporal problems such as controlling the spread of an emerging epidemic (Laber et al. 2018). In this section we examine methods for estimating the optimal policy in infinite-horizon settings using Q-functions.

4.1. Estimating the Optimal Infinite-Horizon Q-Function

In this section we work in the setting of MDPs (Section 2.1). Recall that optimal infinite-horizon Q-function for an MDP is characterized by the Bellman equations (Equation 6):

$$\begin{aligned}
Q^{\mathrm{opt}}(x, a) &\triangleq \mathbb{E}^{\pi^{\mathrm{opt}}}\left[\sum_{t=1}^{\infty} \gamma^{t-1} Y_t \mid X_1 = x, A_1 = a\right]\\
&= (BQ^{\mathrm{opt}})(x, a).
\end{aligned}$$

Given a set of data for trajectories indexed by $i = 1, \ldots, n$ and time points indexed by $t = 1, \ldots, T$ and a differentiable parametric model $Q(\cdot; \theta)$ for Q^{opt}, the Bellman equation allows us to derive an estimating equation for θ as follows:

$$\begin{aligned}
&\mathbb{E}[Y_t + \gamma \max_{a'} Q(X_{t+1}, a'; \theta) \mid X_t = x, A_t = a] = Q(x, a; \theta)\\
\Rightarrow\ &\mathbb{E}[Y_t + \gamma \max_{a'} Q(X_{t+1}, a'; \theta) - Q(x, a; \theta) \mid X_t = x, A_t = a] = 0\\
\Rightarrow\ &\mathbb{E}\{[Y_t + \gamma \max_{a'} Q(X_{t+1}, a'; \theta) - Q(x, a; \theta)]\nabla_\theta Q(x, a; \theta) \mid X_t = x, A_t = a\} = 0\\
\Rightarrow\ &\mathbb{E}\left\{\sum_{t=1}^{T-1}[Y_t + \gamma \max_{a'} Q(X_{t+1}, a'; \theta) - Q(x, a; \theta)]\nabla_\theta Q(x, a; \theta) \mid X_t = x, A_t = a\right\} = 0.
\end{aligned} \qquad 27.$$

Then given a set of data for trajectories indexed by $i = 1, \ldots, n$ and time points indexed by $t = 1, \ldots, T$, the empirical estimating equation is

$$\mathbb{P}_n\left\{\sum_{t=1}^{T}[Y_t + \gamma \max_{a} Q(X_{t+1}, a; \theta) - Q(X_t, A_t; \theta)]\nabla_\theta Q(X_t, A_t; \theta)\right\} = 0. \qquad 28.$$

Consider, in particular, linear models of the form $Q(x, a; \theta) = \phi(x, a)^\intercal\theta$. Define $\widehat{D}(\theta) \triangleq \mathbb{P}_n\{\sum_{t=1}^{T}[Y_t + \gamma \max_{a'} \phi(X_{t+1}, a')^\intercal\theta - \phi(X_t, A_t)^\intercal\theta]\phi(X_t, A_t)\}$. Then Equation 28 under the linear model becomes

$$\widehat{D}(\theta) = 0. \tag{29}$$

This is the approach of Ertefaie & Strawderman (2018), who formulate the estimation of θ via the equivalent (under certain conditions) optimization problem:

$$\begin{aligned} \widehat{\theta} &\triangleq \arg\min_{\theta\in\Theta} \widehat{D}(\theta)^\intercal \widehat{W}^{-1}\widehat{D}(\theta), \text{ where} \\ \widehat{W} &\triangleq \mathbb{P}_n\left\{\sum_{t=1}^{T} \phi(X_t, A_t)\phi(X_t, A_t)^\intercal\right\}. \end{aligned} \tag{30}$$

Indeed, the optimization problem in Equation 30 is a batch version of the incremental algorithm of Maei et al. (2010) (Ertefaie & Strawderman 2018), itself a variant of the original incremental algorithm for Q-learning with approximation (Baird 1995). Ertefaie & Strawderman (2018) demonstrate the consistency and asymptotic normality of the solution to Equation 30 as an estimator of Q^{opt} under regularity conditions and a fixed parameter asymptotic framework.

An alternative is to estimate the Q-function via an approximate version of the value iteration algorithm mentioned in Section 2.1. Value iteration with function approximation (Bertsekas & Tsitsiklis 1996), or fitted Q-iteration (FQI) (Ernst et al. 2005), is similar to finite-horizon Q-learning in that it converts the estimation of the Q-function into a sequence of supervised learning problems. Let $\mathcal{Q}_0, \mathcal{Q}_1, \ldots$ be a sequence of candidate Q-functions (i.e., function approximation architectures). Then FQI is given by:

Algorithm 2 (Fitted Q-iteration for estimating Q^{opt}).

```
$\widehat{Q}_0 \leftarrow \arg\min_{Q_0\in\mathcal{Q}_0} \mathbb{P}_n\left\{\sum_{t=1}^{T}[Y_t - Q_0(X_t, A_t)]^2\right\}$
for k=0, 1, … until convergence do
    $Q_{k+1} \leftarrow \arg\min_{Q_{k+1}\in\mathcal{Q}_{k+1}} \mathbb{P}_n\left\{\sum_{t=1}^{T}\left[Y_t + \gamma \max_a \widehat{Q}_k(X_{t+1}, a) - Q_{k+1}(X_t, A_t)\right]^2\right\}$
```

One advantage of FQI over the estimating-equation approach is that it easily allows the analyst to use a wider array of flexible function architectures (such as random forests). And as with finite-horizon Q-learning, FQI allows the analyst to check the adequacy of the estimated Q-function at each iteration.

FQI suffers from bias and nonregularity as in finite-horizon Q-learning (Section 3). Chakraborty et al. (2008) present a variant of the thresholding approach for reducing bias in Q-learning for the case of FQI with linear function approximation.

4.2. Exploration

Methods for estimating the optimal infinite-horizon policy are often used in online sequential decision problems, in which the decision maker observes data over a time horizon T and must make a decision at each of these times based on accumulated data. A central issue in online sequential decision problems is the exploration-exploitation trade-off, which refers to the trade-off between acting on one's current best estimate of the optimal policy (exploiting) and acting so as to gain more information about the environment so as to improve future decisions (exploring). In Q-learning, this refers to the problem that the Q-function is estimated with error at any time step, and therefore the decision maker has a choice between (*a*) following the estimated optimal Q-function or (*b*) acting so as to gain information that would improve the estimated Q-function.

There is an enormous literature on the exploration-exploitation trade-off in reinforcement learning and other sequential decision settings, and a thorough review is beyond the scope of this article; Auer (2002), Villar et al. (2015), Ghavamzadeh et al. (2015), and Russo et al. (2018) provide reviews emphasizing different perspectives on the exploration-exploitation problem.

We present one class of solutions to the exploration-exploitation problem that involve introducing randomness into the parameters of the estimated Q-function in order to induce exploration. An analogous exploration strategy for model-based reinforcement learning (in which one attempts to model the entire transition dynamics of the system rather than just the conditional means of cumulative rewards) is Thompson sampling (Thompson 1933, Russo et al. 2018). In Thompson sampling, a posterior distribution is maintained over the parameters indexing the transition dynamics model. At each decision point this posterior is sampled from, and the decision maker takes the action that is optimal under the sampled model. In the case of Q-learning, we can achieve similar behavior—i.e., sampling from a distribution that concentrates on the solution to the population-level variant of the Q-learning estimating Equation 28 as data accumulate—by perturbing the estimating equation that defines the Q-function estimator as follows:

$$\mathbb{P}_n\left\{\sum_{t=1}^{T} W_t\left[Y_t + \gamma \max_a Q(X_{t+1}, a; \theta) - Q(X_t, A_t; \theta)\right]\nabla_\theta Q(X_t, A_t; \theta)\right\} = 0, \qquad 31.$$

where W_v are multiplier bootstrap weights, e.g., $W_v \overset{\text{iid}}{\sim} \text{Exp}(1)$ (Præstgaard & Wellner 1993, Chatterjee 2005). Similar approaches have been proposed in the machine learning literature (Eckles & Kaptein 2014, Fortunato et al. 2017, Osband et al. 2017, Plappert et al. 2017).

Algorithm 3 (Online Q-learning with bootstrap Thompson sampling).

```
Observe initial data $\{(X_1^i, A_1^i, Y_1^i, X_2)\}_{i=1}^n$
for $T = 2, \ldots, T'$ do
    $\widehat{\theta}_T^{\text{BTS}} \leftarrow \arg\min_{\theta\in\Theta} \Big\| \mathbb{P}_n \Big\{ \sum_{t=1}^{T} W_t \Big[ Y_t + \gamma \max_a Q(X_{t+1}, a; \theta)$
            $- Q(X_t, A_t; \theta) \Big] \nabla_\theta Q(X_t, A_t; \theta) \Big\} \Big\|^2$ (Equation 31)
    $A_T^i \leftarrow \arg\max_{a\in\mathcal{A}} Q(X_T^i, a; \widehat{\theta}_T^{\text{BTS}})$ for $i = 1, \ldots, n$
    Observe $X_{T+1}^i \sim P(\cdot \mid X_T^i, A_T^i)$ and $Y_T^i \equiv u(X_{T+1}^i)$ for $i = 1, \ldots, n$
```

Algorithm 3 displays the online sequential decision algorithm in which the optimal policy is estimated using Ertefaie & Strawderman's (2018) version of Q-learning and exploration is achieved using bootstrap Thompson sampling.

4.3. Q-Learning with Policy Search in Infinite Horizons

As with finite horizons, it is possible to estimate the value of a policy in an infinite-horizon MDP directly, and thereby obtain an estimator of the optimal policy. Policy search methods for MDPs rely on variants of the Bellman equations (Equation 6) for the Q- and V-functions associated with a given policy π:

$$\begin{aligned} Q^\pi(x, a) &= \mathbb{E}\{Y_1 + \gamma Q^\pi[X_2, \pi(X_2)] \mid X_1 = x, A_1 = a\}; \\ V^\pi(x) &= \sum_{a\in\mathcal{A}} \pi(a|x) Q^\pi(x, a), \end{aligned} \qquad 32.$$

where $\pi(a|x)$ is the probability of taking action a under policy π in state x. Given this characterization, we may obtain an estimator $\widehat{Q}^\pi$ using methods analogous to those discussed in Section 4.1 for the estimation of the optimal Q-function. This in turn yields an estimator of the value function for π, $\widehat{V}^{\pi}_{Q}(x) \triangleq \sum_{a\in\mathcal{A}} \pi(a|x)\widehat{Q}^{\pi}(x,a)$. Then, for reference distribution ν, we have the policy-value estimator $\widehat{V}^{\pi} \triangleq \int \widehat{V}^{\pi}_{Q}(x)\mathrm{d}\nu(x)$.

As with finite-horizon policy search, estimating the policy-value function directly means evaluating policies other than the policy that was used to generate the observed data. This can be addressed using inverse probability weighting. Call the behavior policy—i.e., the policy used to generate the data—b, and let $\widehat{b}(a|x)$ be an estimator of the propensity score for action a in state x under the behavior policy.[1] Define the inverse probability weights or importance sampling weights $\rho_t(\pi;\widehat{b}) = \frac{\pi(A_t|X_t)}{\widehat{b}(A_t|X_t)}$ and subsequently define an inverse probability weighting estimator of the policy-value function as $\widehat{V}^{\text{IPW-MDP}}(\pi) \triangleq \mathbb{P}_n\{[\prod_{t=1}^{T} \rho_t(\pi;\widehat{b})][\sum_{v=t}^{T} \gamma^{t-1}Y_v]\}$. This estimator is unbiased and consistent when $\widehat{b}$ is correctly specified, but it suffers from high variance.

To improve upon the above classes of estimators, Jiang & Li (2015) present a doubly robust estimator that extends a doubly robust estimator for contextual bandits (Dudík et al. 2014) and is analogous to $\widehat{V}^{\text{DR}}$ from the finite-horizon setting (Equation 24). The estimator is defined recursively as follows:

$$\begin{aligned}
&\widehat{V}_0^{\text{DR-MDP}}(\pi) \triangleq 0;\\
&\widehat{V}_{T+1-t}^{\text{DR-MDP}}(\pi) \triangleq \mathbb{P}_n\left\{\widehat{V}^{\pi}_{Q}(X_t) + \rho_t(\pi;\widehat{b})\left[Y_t + \gamma V_{T-t}^{\text{DR-MDP}}(\pi) - \widehat{Q}^{\pi}(X_t,A_t)\right]\right\}.
\end{aligned} \qquad 33.$$

The final estimator of the value of π is then $\widehat{V}^{\text{DR-MDP}} \triangleq \widehat{V}_{T}^{\text{DR-MDP}}$. When $\rho(\pi;\widehat{b})$ and $\widehat{Q}^{\pi}$ are independent, we have that $\widehat{V}^{\pi}(x) = \mathbb{E}_{a\sim\widehat{b}(\cdot|x)}\{\rho(\pi;\widehat{b})\widehat{Q}^{\pi}(x,a)\}$; independence may be achieved by estimating $\widehat{Q}^{\pi}$ on a separate data set. In this case, $\widehat{V}^{\text{DR-MDP}}$ has the double-robustness policy that if either the propensity model or the Q-function estimator is correctly specified, then it will consistently estimate the policy-value function at π. The estimator also has lower variance than $\widehat{V}^{Q}$ and $\widehat{V}^{\text{IPW-MDP}}$.

Luckett et al.'s (2018) V-learning is a policy search method that optimizes this estimator of the policy-value function:

$$\begin{aligned}
&\widehat{V}^{\text{VL}}(\pi) \triangleq \int V(x;\widehat{\eta}^{\pi})\mathrm{d}\nu(x), \text{ where } \widehat{\eta}^{\pi} \text{ solves}\\
&\mathbb{P}_n\left\{\sum_{t=1}^{T} \rho_t(\pi;\widehat{b})\Big[Y_t + \gamma V(X_{t+1};\eta^{\pi}) - V(X_t;\eta^{\pi})\Big]\nabla_{\eta^{\pi}} V(X_t;\eta^{\pi})\right\} = 0,
\end{aligned} \qquad 34.$$

where $V(\cdot;\eta^{\pi})$ is a model for the V-function for policy π indexed by parameter η^{π}. Equation 34 is derived from the Bellman equation, which characterizes the V-function of a policy π (Equation 32) analogously to the derivation of the Q-learning estimating equation (Equation 28).

As in the finite-horizon setting, policy search methods may be combined with Q-function methods. We will discuss this again in Section 5.1.

4.4. Deep Reinforcement Learning

Outside of statistics, reinforcement learning has in recent years enjoyed a surge of interest in the computer science community with the advent of deep learning. Among the first major successes of deep reinforcement learning was the deep Q-network (DQN), which achieved performance comparable to that of human experts in a number of Atari video games using only raw pixels as

---

[1] For simplicity, we assume that b depends only on the current context x rather than the history, but this will not generally be true, especially in online settings where actions are taken according to a policy estimated from the entire history. Nevertheless, the assumption may still lead to reasonable performance; see the online simulation experiments of Luckett et al. (2019), for instance. Alternatively, one could have the estimated behavior policy depend on a low-dimensional sufficient statistic for the history.

a state representation (Mnih et al. 2015). DQN is similar to fitted Q-iteration (Algorithm 2), in that it estimates the infinite-horizon Q-function via a sequence of supervised learning problems. The difference is that, in the case of DQN, the target of each supervised learning problem is an estimator of the Q-function trained on a different batch of data. Let $T(i)$ denote the time point of the ith update to the DQN Q-function estimator; let $\mathcal{T}(i)$ be a set of indices sampled uniformly at random from $\{1, \ldots, T(i)\}$, so that $\{(X_t, A_t, Y_t, X_{t+1})\}_{t \in \mathcal{T}(i)}$ are sampled uniformly from the data observed until time $T(i)$; and let $\mathbb{P}_{\mathcal{T}(i)}$ be the empirical expectation taken with respect to this batch of data. Then the Q-function at the ith update of the DQN algorithm is estimated as

$$\widehat{Q}^i \triangleq \arg\min_{Q \in \mathcal{Q}} \mathbb{P}_{\mathcal{T}(i)} \left\{ Y_t + \gamma \max_{a \in \mathcal{A}} \widehat{Q}^{i-1}(X_{t+1}, a) - Q(X_t, A_t) \right\}^2, \qquad 35.$$

where $\mathcal{Q}$ is the class of candidate Q-functions indexed by the parameters of a deep neural network. Rainbow DQN subsequently collected a number of improvements on the original DQN algorithm into a single algorithm, which demonstrated considerably improved performance on the Atari benchmark (Hessel et al. 2018). These include a method for increasing sample efficiency by prioritizing higher-error tuples in the sampling used to construct the DQN objective (prioritized experience replay; Schaul et al. 2015); a method for reducing the finite-sample maximization bias (Smith & Winkler 2006) induced by the max operator (double Q-learning; Hasselt 2010, Hasselt et al. 2016); and the Noisy Net method for inducing and tuning exploration via adding noise to the parameters of the Q-function (Fortunato et al. 2017).

However, policy-based methods represent the state of the art in model-free deep reinforcement learning. One popular approach is the trust region policy optimization (TRPO) algorithm (Schulman et al. 2015) and its refinements under the heading of proximal policy optimization (Schulman et al. 2017). In TRPO, the estimated optimal policy at the ith iteration is obtained by maximizing the objective

$$\widehat{V}^{\mathrm{TRPO}}(\pi) \triangleq \mathbb{P}_{T(i)} \left[\rho_t(\pi; \widehat{\pi}^{i-1}) \widehat{Q}^{\widehat{\pi}^{i-1}}(X_t, A_t) \right] - \iota \left\{ \mathbb{P}_{T(i)} KL \left[\widehat{\pi}^{i-1}(\cdot \mid X_t) \parallel \pi(\cdot \mid X_t) \right] < \lambda \right\}, \qquad 36.$$

where KL is the Kullback-Leibler (KL) divergence, λ is a tuning parameter, and ι is the 0-∞ indicator [yielding a hard constraint on the average KL divergence between $\pi(\cdot|X_t)$ and $\widehat{\pi}^{i-1}(\cdot|X_t)$]. That is, these methods update the previous policy $\widehat{\pi}^{i-1}$ by maximizing an importance-sampling estimator of the policy-value without moving too far away from $\widehat{\pi}^{i-1}$ (in KL divergence). Notice that the policy-value estimator $\mathbb{P}_{T(i)}[\rho_t(\pi; \widehat{\pi}^{i-1}) \widehat{Q}^{\widehat{\pi}^{i-1}}(X_t, A_t)]$ is an estimator of $\mathbb{E}^{\widehat{\pi}^{i-1}}[\rho_t(\pi; \widehat{\pi}^{i-1}) Q^{\pi}(X_t, A_t)] = \mathbb{E}^{\pi}[Q^{\pi}(X_t, A_t)]$, using $\widehat{Q}^{\widehat{\pi}^{i-1}}$ as an estimator of the target-policy Q-function Q^{π}; this is computationally easier than estimating the latter directly. Forcing the policy not to be too far from the previous policy (rather than globally optimizing the policy-value estimator) has two advantages. First, because $\widehat{\pi}^{i-1}$ is used as a stand-in for π in the estimation of the target policy Q-function, the quality of the estimator presumably decays as π moves away from $\widehat{\pi}^{i-1}$. Second, the penalty induces stability in the estimation of the optimal policy, as not only is the policy-value estimator high-variance for a fixed policy, but the class of policies searched over (i.e., one indexed by the parameters of a deep neural network) is extremely large.

5. Q-LEARNING IN APPLICATION

5.1. Q-Learning with Policy Search as the Best of Both Worlds?

Any method for estimating an optimal policy must, at least implicitly, solve two problems: (*a*) It must provide a reasonable criterion for choosing the estimated optimal policy from a given class, and (*b*) it must use a reasonable class of candidate policies. In Q-learning, the class of policies is defined implicitly by the class of candidate Q-functions. Policy search methods, instead, address *a* by attempting to give a good estimator of the value of a policy. But, they do not automatically address *b*, as the policy-value estimator may in principle be optimized over any class of policies.

So, Q-learning naturally enforces the choice of a reasonable class of policies, as an analyst may diagnose the fit of the estimated Q-functions against the actual outcomes and pseudooutcomes at each stage and choose different modeling strategies accordingly—just as with typical regression problems. After conducting such an analysis, an analyst may at least feel confident that they have estimates that do a reasonable job of capturing the relationship between actions and the conditional mean of the outcome. Policy search does not come with any analogous method for ensuring that the policy class is well specified. While this may not be a problem in the context of deep reinforcement learning, where extremely general classes of policies may be searched over to optimize an estimator constructed from huge amounts of data, in statistical contexts, the policy class must allow for data-efficient estimation as well as interpretability by domain experts. Policy search has the advantage of directly optimizing an estimator of the quantity of interest (the policy value), whereas Q-learning estimates the optimal policy only indirectly.

In our view, the best of both worlds can be achieved by using Q-learning as a criterion for policy search and as a method for evaluating a class of candidate policies. When used in constructing a criterion for policy search, Q-function methods can reduce the variance of the objective function (Section 3.4), are flexible and have a low risk of misspecification, and provide a means to estimate the optimal policy nonparametrically. In infinite-horizon settings, one may choose a feature representation ϕ via Q-learning by, for instance, minimizing the cross-validated weighted temporal difference error (Equation 30). Alternatively, one may estimate the transition dynamics of the MDP and attempt to construct a parsimonious representation of the Q-function using Monte Carlo estimates of the same from the estimated transition model. One may then optimize over a differentiable class of stochastic policies indexed by θ that give probability to actions a in state x as an increasing function of $\phi(a, x)^\intercal\theta$; one commonly used such class is $\pi(a|x) \propto \exp\{\phi(a, x)^\intercal\theta\}$.

5.2. Sample Size Considerations

Applying Q-learning to the estimation of an optimal regime from clinical trial data—in particular, from a SMART (Murphy 2005b)—ideally involves sizing the trial so as to construct high-quality estimates of the optimal regime. SMARTs are typically sized for comparisons of two fixed treatment sequences, with estimation of the optimal regime done as a secondary analysis. However, Rose et al. (2019) present sample size calculations for powering tests of the value of the optimal dynamic treatment regime itself (against a standard-of-care null), for both parametric and nonparametric assumptions on the data-generating process. That is, letting $\widehat{\pi}_n$ be an estimator of π^{opt} based on a SMART of sample size n, and letting $B_0 > 0, \gamma, \alpha, \eta, \epsilon, \zeta \in (0, 1)$ be constants, they derive procedures for choosing n such that

$$\begin{aligned} &\text{(POW)} \quad \text{There exists an } \alpha\text{-level test of } H_0 : V(\pi^{\mathrm{opt}}) \leq B_0 \text{ based on } \widehat{\pi}_n \text{ that has power at least} \\ &\qquad\qquad (1-\gamma)\cdot 100 + o(1) \text{ provided } V(\pi^{\mathrm{opt}}) \geq B_0 + \eta, \\ &\text{(OPT)} \quad P\left\{\mathbb{E}\left[Y^*(\widehat{\pi}_n) \mid D_n\right] \geq V(\pi^{\mathrm{opt}}) - \epsilon\right\} \geq 1 - \zeta + o(1). \end{aligned} \tag{37}$$

5.3. Combining Model-Based and Model-Free Q-Learning for Sample Efficiency

In cases where data are scarce, combining model-based estimators (Section 2.1.1) with model-free estimators may improve estimation by reducing variance. A downside of model-based approaches in comparison to model-free ones is the bias introduced by an incorrectly specified transition model. As such, the best strategy may be to adaptively combine model-based and model-free estimators.

An early example of a hybrid model-based and model-free approach is the Dyna-Q architecture, which draws from an estimated dynamics model supplement observed data in order to reduce variance in the estimation of the Q-function (Sutton 1990). However, this algorithm does not attempt to trade off the strengths and weaknesses of its model-based and model-free components. More recent work in the deep reinforcement learning literature has attempted to address this problem. For instance, Kalweit & Boedecker (2017) describe a method in which a model-free approach is augmented using data generated from an estimated dynamics model, with the amount of imagined data increasing with the (bootstrap-estimated) uncertainty in the estimated dynamics model [see also Feinberg et al. (2018) and references therein for more examples of hybrid strategies]. However, these methods still reside in the deep reinforcement learning setting in which large amounts of simulated data are available, rather than being tailored to more data-poor cases. In the direction of optimally combining model-based and model-free approaches in the latter case, Thomas & Brunskill (2016) combine a version of $\widehat{V}^{\text{DR-MDP}}$ (Section 4.3) with model-based estimators of the policy-value function, i.e., estimators derived from the estimated dynamics of the MDP and measuring the value of the target policy in this estimated MDP via simulation. In particular, they present the Model and Guided Importance Sampling Combined (MAGIC) estimator, which takes the estimated MSE-optimal combination of several different blended estimators, each of which uses the model-based value estimator up until a particular time horizon and then $\widehat{V}^{\text{DR-MDP}}$ thereafter. Extending the statistically principled combination of model-free and model-based estimators to other methods is a promising direction for research into improving the data efficiency of reinforcement learning.

One possible application area for a hybrid variant of Q-learning is the control of infectious disease, in particular extending the model-based approach of Laber et al. (2018) for controlling the spread of an emerging epidemic via the optimal allocation of limited resources to incorporate model-free elements, as the former approach depends on a low-dimensional model of disease dynamics that is unlikely to be correctly specified. Among model-free approaches, policy search methods are likely to suffer from prohibitively high variance, given the small number of observations available to decision makers in these contexts paired with the combinatorially large action spaces. Variants of Q-learning may allow for the relatively stable model-free estimation of the optimal policy.

6. DISCUSSION

While the use of Q-learning for the estimation of optimal finite-horizon treatment regimes has matured considerably since its inception with the work of Murphy (2003, 2005a), there are still open research directions, including the extension of the sample size procedures discussed in Section 5.2 to different sets of assumptions on the data-generating model. At the same time, the development of Q-learning in batch infinite-horizon settings, such as mHealth and large-scale spatio-temporal decision-making, is just beginning. Q-learning in these settings shares several of the challenges that Tewari & Murphy (2017) list for contextual bandits as applied to mHealth, including exploring efficiently, finding a good initial policy, balancing policy interpretability and performance, assessing the usefulness of features with respect to the performance of the learned policy, addressing the computational difficulties associated with an online learning and recommendation system implemented on mobile devices, and accounting for missingness in features or rewards. Finally, as touched upon in Section 5.3, the principled combination of model-free Q-learning with model-based methods that reduce variance and take advantage of domain knowledge is a promising direction for improving the performance of Q-learning-based decision support systems.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Andrews DW. 2000. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* 68:399–405

Auer P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* 3:397–422

Baird L. 1995. Residual algorithms: reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, ed. A Prieditis, S Russell, pp. 30–37. San Francisco: Morgan Kaufmann

Bellman R. 1957. *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press

Bertsekas DP, Tsitsiklis J. 1996. *Neuro-Dynamic Programming*. Nashua, NH: Athena Sci.

Bowling M, Burch N, Johanson M, Tammelin O. 2015. Heads-up limit hold'em poker is solved. *Science* 347:145–49

Chakraborty B, Laber EB, Zhao Y. 2013. Inference for optimal dynamic treatment regimes using an adaptive *m*-out-of-*n* bootstrap scheme. *Biometrics* 69:714–23

Chakraborty B, Laber EB, Zhao YQ. 2014. Inference about the expected performance of a data-driven dynamic treatment regime. *Clin. Trials* 11:408–17

Chakraborty B, Murphy S, Strecher V. 2010. Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res.* 19:317–43

Chakraborty B, Strecher V, Murphy S. 2008. Bias correction and confidence intervals for fitted Q-iteration. In *Workshop on Model Uncertainty and Risk in Reinforcement Learning (NIPS 2008)*. **https://cs.uwaterloo.ca/~ppoupart/nips08-workshop/accepted-papers/nips08paper01-final.pdf**

Chatterjee S, Bose A, et al. 2005. Generalized bootstrap for estimating equations. *Ann. Stat.* 33:414–36

Dudík M, Erhan D, Langford J, Li L. 2014. Doubly robust policy evaluation and optimization. *Stat. Sci.* 29:485–511

Eckles D, Kaptein M. 2014. Thompson sampling with the online bootstrap. arXiv:1410.4009 [cs.LG]

Ernst D, Geurts P, Wehenkel L. 2005. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.* 6:503–56

Ertefaie A, Strawderman RL. 2018. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika* 105:963–77

Feinberg V, Wan A, Stoica I, Jordan MI, Gonzalez JE, Levine S. 2018. Model-based value estimation for efficient model-free reinforcement learning. arXiv:1803.00101 [cs.LG]

Fortunato M, Azar MG, Piot B, Menick J, Osband I, et al. 2017. Noisy networks for exploration. arXiv:1706.10295 [cs.LG]

Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. *Mach. Learn.* 63:3–42

Ghavamzadeh M, Mannor S, Pineau J, Tamar A. 2015. Bayesian reinforcement learning: a survey. *Found. Trends Mach. Learn.* 8:359–483

Hasselt HV. 2010. Double Q-learning. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, ed. JD Lafferty, CKI Williams, J Shawe-Taylor, RS Zemel, A Culotta, pp. 2613–21. **https://papers.nips.cc/paper/3964-double-q-learning**

Hasselt HV, Guez A, Silver D. 2016. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2094–100. Menlo Park, CA: AAAI

Hernán M, Robins J. 2019. *Causal Inference*. Boca Raton: Chapman & Hall/CRC

Hessel M, Modayil J, Van Hasselt H, Schaul T, Ostrovski G, et al. 2018. Rainbow: combining improvements in deep reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3215–22. Menlo Park, CA: AAAI

Hirano K, Porter JR. 2012. Impossibility results for nondifferentiable functionals. *Econometrica* 80:1769–90

Jiang N, Li L. 2015. Doubly robust off-policy value evaluation for reinforcement learning. arXiv:1511.03722 [cs.LG]

Kallus N, Zhou A. 2018. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett. **https://papers.nips.cc/paper/8139-confounding-robust-policy-improvement**

Kalweit G, Boedecker J. 2017. Uncertainty-driven imagination for continuous deep reinforcement learning. *PMLR* 78:195–206

Laber EB, Linn K, Stefanski L. 2014a. Interactive model building for Q-learning. *Biometrika* 101:831–47

Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. 2014b. Dynamic treatment regimes: technical challenges and applications. *Electron. J. Stat.* 8:1225

Laber EB, Meyer N, Reich B, Pacifici J, Collazo J, Drake J. 2018. On-line estimation of an optimal treatment allocation strategy for the control of emerging infectious disease. *J. R. Stat. Soc. C* 67:743–70

Leeb H, Poetscher B. 2003. The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econom. Theory* 19:100–42

Luckett D, Laber E, El-Kamary S, Fan C, Jhaveri R, et al. 2018. Receiver operating characteristic curves and confidence bands for support vector machines. arXiv:1807.06711 [stat.ML]

Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR. 2019. Estimating dynamic treatment regimes in mobile health using V-learning. *J. Am. Stat. Assoc.* **https://doi.org/10.1080/01621459.2018.1537919**

Maei HR, Szepesvári C, Bhatnagar S, Sutton RS. 2010. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, ed. J Fürnkranz, T Joachims, pp. 719–26. Madison, WI: Omnipress

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518:529

Moodie EE, Dean N, Sun YR. 2014. Q-learning: flexible learning about useful utilities. *Stat. Biosci.* 6:223–43

Moodie EE, Richardson TS, Stephens DA. 2007. Demystifying optimal dynamic treatment regimes. *Biometrics* 63:447–55

Moodie EE, Richardson TS, Stephens DA. 2010. Estimating optimal dynamic regimes: correcting bias under the null. *Biometrics* 63:447–55

Murphy SA. 2003. Optimal dynamic treatment regimes. *J. R. Stat. Soc. B* 65:331–55

Murphy SA. 2005a. A generalization error for Q-learning. *J. Mach. Learn. Res.* 6:1073–97

Murphy SA. 2005b. An experimental design for the development of adaptive treatment strategies. *Stat. Med.* 24:1455–81

Murphy SA, Van Der Laan M, Robins J. 2001. Marginal mean models for dynamic regimes. *J. Am. Stat. Assoc.* 96:1410–23

Osband I, Russo D, Wen Z, Van Roy B. 2017. Deep exploration via randomized value functions. arXiv:1703.07608 [stat.ML]

Plappert M, Houthooft R, Dhariwal P, Sidor S, Chen RY, et al. 2017. Parameter space noise for exploration. arXiv:1706.01905 [cs.LG]

Powell WB. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. New York: Wiley

Præstgaard J, Wellner JA. 1993. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* 21:2053–86

Precup D, Sutton RS, Dasgupta S. 2001. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning*, ed. CE Brodley, AP Danyluk, pp. 417–24. San Francisco: Morgan Kaufmann

Puterman ML. 2009. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: Wiley

Qian M, Murphy S. 2011. Performance guarantees for individualized treatment rules. *Ann. Stat.* 39:1180–210

Riedmiller M. 2005. Neural fitted Q iteration—first experiences with a data efficient neural reinforcement learning method. In *Machine Learning: ECML 2005*, ed. J Gama, R Camacho, PB Brazdil, A Jorge, L Torgo, pp. 317–28. New York: Springer

Robins J. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* 7:1393–512

Robins JM. 1987. Addendum to a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Comput. Math. Appl.* 14:923–45

Robins JM. 1989. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Serv. Res. Methodol.* 113:159

Robins JM. 1993. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, Vol. 24, pp. 24–33. Washington, DC: Am. Stat. Assoc.

Robins JM. 2004. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, ed. D Lin, PJ Heagerty, pp. 189–326. New York: Springer

Rose EJ, Laber EB, Davidian M, Tsiatis AA, Zhao Y, Kosorok MR. 2019. Sample size calculations for SMARTs. arXiv:1906.06646 [stat.ME]

Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55

Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–58

Rubin DB. 2005. Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Stat. Assoc.* 100:322–31

Russo DJ, Van Roy B, Kazerouni A, Osband I, Wen Z, et al. 2018. A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* 11:1–96

Schaul T, Quan J, Antonoglou I, Silver D. 2015. Prioritized experience replay. arXiv:1511.05952 [cs.LG]

Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. 2015. Trust region policy optimization. *PMLR* 37:1889–97

Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. 2017. Proximal policy optimization algorithms. arXiv:1707.06347 [cs.LG]

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, et al. 2017. Mastering the game of go without human knowledge. *Nature* 550:354

Smith JE, Winkler RL. 2006. The optimizers curse: skepticism and postdecision surprise in decision analysis. *Manag. Sci.* 52:311–22

Song R, Wang W, Zeng D, Kosorok MR. 2015. Penalized Q-learning for dynamic treatment regimens. *Stat. Sin.* 25:901

Splawa-Neyman J, Dabrowska DM, Speed TP. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.* 5:465–72

Sutton RS. 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine Learning Proceedings 1990*, ed. B Porter, R Mooney, pp. 216–24. Amsterdam: Elsevier

Sutton RS, Barto AG. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press. 2nd ed.

Tewari A, Murphy SA. 2017. From ads to interventions: contextual bandits in mobile health. In *Mobile Health: Sensors, Analytic Methods, and Applications*, ed. JM Rehg, SA Murphy, S Kumar, pp. 495–517. New York: Springer

Thomas P, Brunskill E. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, ed. MF Balcan, KQ Weinberger, pp. 2139–48. Brookline, MA: Microtome

Thompson WR. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25:285–94

Tsiatis A. 2006. *Semiparametric Theory and Missing Data*. New York: Springer

van der Laan MJ, Petersen ML, Joffe MM. 2005. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *Int. J. Biostat.* 1:4

Van der Vaart A. 1991. On differentiable functionals. *Ann. Stat.* 19:178–204

Vansteelandt S, Joffe M. 2014. Structural nested models and G-estimation: the partially realized promise. *Stat. Sci.* 29:707–31

Villar SS, Bowden J, Wason J. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Stat. Sci.* 30:199

Watkins C, Dayan P. 1992. Q-learning. *Mach. Learn.* 8:279–92

Xu X, Kypraios T, O'Neill PD. 2016. Bayesian non-parametric inference for stochastic epidemic models using Gaussian processes. *Biostatistics* 17:619–33

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. 2012. Estimating optimal treatment regimes from a classification perspective. *Statistics* 1:103–14

Zhang B, Tsiatis AA, Laber EB, Davidian M. 2013. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* 100:681–94

Zhang Y, Laber EB, Tsiatis A, Davidian M. 2015. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* 71:895–904

Zhang Y, Laber EB, Tsiatis A, Davidian M. 2016. Interpretable dynamic treatment regimes. arXiv:1606.01472 [stat.ME]

Zhao Y, Kosorok M, Zeng D. 2009. Reinforcement learning design for cancer clinical trials. *Stat. Med.* 28:3294–315

Zhou X, Kosorok MR. 2017. Causal nearest neighbor rules for optimal treatment regimes. arXiv:1711.08451 [stat.ML]