SCHULTE, PHILLIP JOEL. Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes. (Under the direction of Anastasios Tsiatis and Marie Davidian.)

In clinical practice, physicians make a series of treatment decisions over the course of a patient’s disease based on his/her baseline and evolving characteristics. A dynamic treatment regime is a set of sequential decision rules that operationalizes this process. Each rule corresponds to a key decision point and dictates the next treatment action among the options available as a function of accrued information on the patient. Using data from a clinical trial or observational study, a key goal is estimating the optimal regime, that, if followed by the patient population, would yield the most favorable outcome on average. Q-learning and advantage (A-)learning are two main approaches for this purpose. We provide a detailed account of Q- and A-learning and study systematically the performance of these methods. The methods are illustrated using data from a study of depression.
Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes

by
Phillip Joel Schulte

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Statistics

Raleigh, North Carolina
2012

APPROVED BY:

.........................................................
Daowen Zhang

.........................................................
Eric Laber

.........................................................
Wenbin Lu

.........................................................
Robin Gardner

.........................................................
Anastasios Tsiatis
Co-chair of Advisory Committee

.........................................................
Marie Davidian
Co-chair of Advisory Committee
DEDICATION

To my parents. Their love and support over the years has been invaluable and made it possible for me to get through my Ph.D. training.
Phillip Joel Schulte, born March 13, 1984, in Chaska, Minnesota, is the son of Charles and Frances. He graduated from Chaska Senior High School in 2002 and attended Saint Olaf College. At Saint Olaf, Phillip pursued mathematics and statistics, including studying mathematics abroad in Budapest, Hungary and engaging in undergraduate statistics research. Phillip graduated from Saint Olaf in 2006 with a B.A. in Mathematics, with concentrations in Statistics and Management Studies. He began his graduate work in the Department of Statistics at North Carolina State University in the fall of 2006. In 2008, Phillip received a Master of Statistics degree and he is set to receive his Ph.D. in Statistics in 2012 under the direction of Drs. Tsiatis and Davidian.
ACKNOWLEDGEMENTS

I would like to thank my advisors, Drs. Marie Davidian and Butch Tsiatis. I learned so much from their expertise and this research could not have been completed without their guidance and knowledge. I’m grateful for the opportunities they both have given me, particularly to Dr. Davidian for her support from a training grant through most of my graduate career and continued funding through research grants. As a result of her support, I’ve been fortunate to travel to conferences, intern at Duke Clinical Research Institute (DCRI), mentor during the summer institute for training in biostatistics (SIBS), and more. Thanks also to Dr. Eric Laber for his contributions and insights over the last year. I’d also like to express my thanks to my committee members and the rest of the NC State Statistics faculty, staff, and students I’ve worked and interacted with. I’d like to acknowledge support from NIH training and research grants T32HL079896 and P01CA142538.

I would like to thank Karen Pieper. Working at DCRI under her supervision has been a truly rewarding experience. I cannot thank Karen enough for her guidance, encouragement, and wealth of knowledge that she has shared with me. I’m also grateful for all the clinical investigators and statisticians I’ve worked and collaborated with at DCRI. It has been a great experience that truly complemented my education.

During my time at Saint Olaf, I was especially inspired and encouraged by Drs. Paul Roback and Cliff Corzatt; a big thanks to each of them. Dr. Corzatt was involved in my study abroad trips to Budapest, Hungary and helped foster interests in advanced mathematics. Taking statistics classes from Dr. Roback encouraged me to pursue a graduate degree in statistics. He was a great mentor for several projects and excellent source of information and advice when deciding to attend a graduate program.

Of course, I would never have gotten to where I am today without the love and support of my family and close friends. I’m grateful for having such amazing and supportive parents, Charles and Frances, who have worked hard to make so much possible throughout my life. My brothers (Brian, Allen, Mike, and Colin) have all been inspirational and I’m thankful for their support. Finally, thank you to my girlfriend Mollie for her support, patience, and encouragement.
TABLE OF CONTENTS

List of Tables ................................................................. vii
List of Figures ................................................................. viii

Chapter 1 Introduction ....................................................... 1
  1.1 Framework and Assumptions ......................................... 3
  1.2 Defining the Optimal Regime ...................................... 8
  1.3 Q- and A-Learning .................................................. 12
    1.3.1 Q-Learning .................................................. 12
    1.3.2 A-Learning ................................................ 14
  1.4 Summary and Discussion .......................................... 17
    1.4.1 Comparison and Practical Considerations ................. 18

Chapter 2 Simulations Under a Single Decision Point, $K = 1$ .... 21
  2.1 Primary Simulations and Results ................................ 23
  2.2 Simulation for the Effects of Sample Size .................... 26
  2.3 Simulations for Alternate Parameterizations ................. 31
    2.3.1 Results for Misspecified Propensity Model ............. 31
    2.3.2 Results for Misspecified Q-function .................... 33
    2.3.3 Results for Misspecification of Both the Propensity Model and Q-function .................. 38
  2.4 Discussion ....................................................... 41

Chapter 3 Simulations Under Two or More Treatment Decisions ... 42
  3.1 $K = 2$ Decisions ................................................ 43
    3.1.1 Primary Simulations and Results ......................... 45
    3.1.2 Simulations for the Effects of Sample Size ............ 50
    3.1.3 Simulations for Alternate Parameterizations .......... 54
  3.2 Moodie and Richardson Simulations ........................... 66
  3.3 Discussion ....................................................... 69

Chapter 4 Data Analysis and Discussion ............................... 71
  4.1 STAR*D Data Analysis .......................................... 71
  4.2 Discussion ....................................................... 75

References ................................................................. 77

Appendices ................................................................. 80
<table>
<thead>
<tr>
<th>Appendix A</th>
<th>Demonstration That Equations (1.5)–(1.8) Define an Optimal Regime</th>
<th>81</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix B</td>
<td>Demonstration of Correspondence in Equations (1.20)–(1.22) Under Assumptions in Chapter 1.1</td>
<td>84</td>
</tr>
<tr>
<td>Appendix C</td>
<td>Justification for $\tilde{V}_{ki}$ in $A$-learning</td>
<td>89</td>
</tr>
<tr>
<td>Appendix D</td>
<td>Demonstration of Equivalence of $Q$- and $A$-learning in a Special Case</td>
<td>90</td>
</tr>
<tr>
<td>Appendix E</td>
<td>Example of Incompatibility of $Q$-function Models</td>
<td>92</td>
</tr>
<tr>
<td>Appendix F</td>
<td>Calculation of $E{H(\hat{d}^{opt})}$ and $R(\hat{d}^{opt})$</td>
<td>93</td>
</tr>
<tr>
<td>Appendix G</td>
<td>Creating “Equivalently Misspecified Pairs” When Both the Propensity Model and $Q$-function are Misspecified</td>
<td>96</td>
</tr>
<tr>
<td>Appendix H</td>
<td>Derivation of $h_1^0(s_1; \beta_1^0)$ and $C_1^0(s_1; \psi_1^0)$ in the Two Decision Point Scenario</td>
<td>97</td>
</tr>
<tr>
<td>Appendix I</td>
<td>Inference for Parameter Estimates</td>
<td>98</td>
</tr>
<tr>
<td>I.1</td>
<td>$Q$-learning</td>
<td>99</td>
</tr>
<tr>
<td>I.2</td>
<td>$A$-learning</td>
<td>101</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3.1  Comparison of $Q$- and $A$-learning for the Moodie and Richardson Scenario, $n = 200$ .................................................. 68
Table 3.2  Comparison of $Q$- and $A$-learning for the Moodie and Richardson Scenario, $n = 1000$ .................................................. 68
Table 4.1  STAR*D data analysis results ............................................. 74
Table 4.2  STAR*D data analysis results after additional exploratory analyses 75
LIST OF FIGURES

Figure 2.1 Comparisons of Q- and A-learning under $K = 1$ and misspecification of the propensity model for the primary simulation .................................................. 24
Figure 2.2 Comparisons of Q- and A-learning under $K = 1$ and misspecification of the Q-function for the primary simulation .................................................. 25
Figure 2.3 Comparisons of Q- and A-learning under $K = 1$ and misspecification of both the propensity model and the Q-function for the primary simulation .................................................. 26
Figure 2.4 Sample size comparison via MSE ratios for estimators of $\psi_1$ under $K = 1$ and misspecification of the Q-function .................................................. 28
Figure 2.5 Sample size comparison via efficiency of $\hat{R}(\hat{d}_Q^{opt})$ and $\hat{R}(\hat{d}_A^{opt})$ under $K = 1$ and misspecification of the Q-function .................................................. 28
Figure 2.6 Sample size comparison via efficiency of $\hat{R}(\hat{d}_Q^{opt})$ and $\hat{R}(\hat{d}_A^{opt})$ under $K = 1$ and misspecification of the Q-function .................................................. 29
Figure 2.7 Sample size comparison via MSE ratios for estimators of $\psi_1$ under $K = 1$ and misspecification of both the propensity model and the Q-function .................................................. 30
Figure 2.8 Sample size comparison via efficiency of $\hat{R}(\hat{d}_Q^{opt})$ and $\hat{R}(\hat{d}_A^{opt})$ under $K = 1$ and misspecification of both the propensity model and the Q-function .................................................. 30
Figure 2.9 Sample size comparison via efficiency of $\hat{R}(\hat{d}_Q^{opt})$ and $\hat{R}(\hat{d}_A^{opt})$ under $K = 1$ and misspecification of both the propensity model and the Q-function .................................................. 31
Figure 2.10 Comparison of Q- and A-learning under $K = 1$ and misspecification of the propensity model, new parameterization .................................................. 32
Figure 2.11 Comparison under $K = 1$ and misspecification of the propensity model, with changes to propensity parameters .................................................. 32
Figure 2.12 Comparison under $K = 1$ and misspecification of the propensity model, with changes to Q-function parameters .................................................. 33
Figure 2.13 Comparison under $K = 1$ and misspecification of the propensity model, with alternate variance of the response .................................................. 34
Figure 2.14 Comparison of Q- and A-learning under $K = 1$ and misspecification of the Q-function, new parameterization .................................................. 34
Figure 2.15 Comparison under $K = 1$ and misspecification of the Q-function, with changes to parameters in the propensity model .................................................. 35
Figure 2.16 Comparison under $K = 1$ and misspecification of the Q-function, with changes to parameters in the propensity model .................................................. 36
Figure 2.17 Comparison under $K = 1$ and misspecification of the Q-function, with changes to the variance of the response .................................................. 37
Figure 2.18 Comparison of under $K = 1$ and misspecification of the $Q$-function, with changes to the variance of the response .......................... 37
Figure 2.19 Comparison of $Q$- and $A$-learning under $K = 1$ and misspecification of both the propensity model and the $Q$-function, new parameterization ................................................................. 38
Figure 2.20 Comparison under $K = 1$ and misspecification of both the propensity model and the $Q$-function, with alternate propensity parameters 39
Figure 2.21 Comparison under $K = 1$ and misspecification of both the propensity model and the $Q$-function, with changes to $Q$-function parameters 40
Figure 2.23 Comparison under $K = 1$ and misspecification of both the propensity model and the $Q$-function, with alternate variance of the response 41

Figure 3.1 Comparison of $Q$- and $A$-learning under $K = 2$ and misspecification of the propensity model, primary results ................................. 47
Figure 3.2 Comparison of $Q$- and $A$-learning under $K = 2$ and misspecification of the $Q$-function, primary results ......................................................... 48
Figure 3.3 Comparison of $Q$- and $A$-learning under $K = 2$ and misspecification of both the propensity models and $Q$-functions, primary results . 49
Figure 3.4 Sample size comparison via MSE ratios for estimators of $\psi_1$ and $\psi_2$ under $K = 2$ and misspecification of the $Q$-function .................. 51
Figure 3.5 Sample size comparison via efficiency of $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ under $K = 2$ and misspecification of the $Q$-function ............... 52
Figure 3.6 Sample size comparison via MSE ratios for estimators of $\psi_1$ and $\psi_2$ under $K = 2$ and misspecification of both the propensity model and the $Q$-function ................................................................. 53
Figure 3.7 Sample size comparison via efficiency of $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ under $K = 2$ and misspecification of the $Q$-function .................. 54
Figure 3.8 Comparison of $Q$- and $A$-learning under $K = 2$ and misspecification of the propensity model, new parameterization .......................... 56
Figure 3.9 Comparison of $Q$- and $A$-learning under $K = 2$ and misspecification of the $Q$-function, new parameterization ............................... 58
Figure 3.10 Comparison under $K = 2$ and misspecification of the $Q$-function, with changes to $Q$-function parameters ................................. 59
Figure 3.11 Comparison under $K = 2$ and misspecification of the $Q$-function, with changes to contrast parameters ............................................ 60
Figure 3.12 Comparison under $K = 2$ and misspecification of the $Q$-function, with changes to contrast parameters ............................................ 61
Figure 3.13 Comparison under $K = 2$ and misspecification of the $Q$-function, with changes to contrast parameters ................................. 62
Figure 3.14 Comparison of $Q$- and $A$-learning under $K = 2$ and misspecification of both the propensity models and $Q$-functions, new parameterization 63

Figure 3.15 Comparison under $K = 2$ and misspecification of both the propensity models and $Q$-functions, with changes to the $Q$-function . . . 64

Figure 3.16 Comparison under $K = 2$ and misspecification of both the propensity models and $Q$-functions, with changes to the contrast parameters 65
Introduction

In the health sciences, an area of considerable current interest is personalized medicine, which involves making treatment decisions for an individual patient using all information available on the patient, including genetic, physiologic, demographic, and other clinical variables, to achieve the “best” outcome for the patient given this information. In treating a patient with an ongoing disease or disorder, a clinician makes a series of decisions based on the patient’s evolving status, so seeking to tailor treatment to the patient. A dynamic treatment regime is a list of sequential decision rules that formalizes this process. Each rule corresponds to a key decision point in the disease or disorder progression and takes as input the information on the patient to that point and outputs the treatment that s/he should receive from among the available options. A key step toward personalized medicine is thus finding the optimal dynamic treatment regime, that which, if followed by the entire patient population, would yield the most favorable outcome on average.

The statistical problem is to estimate the optimal regime based on data from a clinical trial or observational study. $Q$-learning ($Q$ denoting “quality,” Watkins 1989; Watkins and Dayan 1992; Nahum-Shani et al. 2010) and advantage learning ($A$-learning, Murphy 2003; Blatt, Murphy, and Zhu 2004) are two main approaches proposed for this purpose. Both follow from developments on reinforcement learning methods for sequential decision-making in the computer science literature. As described shortly, $Q$-learning is based roughly on posited regression models for the outcome of interest given patient information at each decision point and is implemented through a backwards (in time) recursive fitting procedure that is related to the dynamic programming algorithm (Bather 2000), a standard approach for deducing optimal sequential decisions. $A$-learning involves the
same recursive strategy, but, instead of requiring full regression relationships to be posited, requires only models for the part of the outcome regression involved in representing contrasts among treatments along with models for the probability of observed treatment assignment given patient information at each decision point. As discussed in the sequel, this feature may make $A$-learning more robust to model misspecification than $Q$-learning for consistent estimation of the optimal treatment regime.


Despite increasing interest in estimation of optimal dynamic treatment regimes, there has been little study of the relative merits of $Q$- and $A$-learning, nor of consequences of misspecification of the postulated models involved. Moreover, although descriptions of $Q$- and $A$-learning are available, a self-contained account of both has not been presented. In this dissertation, we provide a detailed description of an appropriate statistical framework in which an optimal regime may be defined formally and introduce $Q$- and $A$-learning in this context. Conditions under which these methods may be expected to yield credible estimators for optimal regimes based on observed data are discussed, and we report on a systematic study of the methods’ performance.

This chapter is organized as follows. Section 1.1 introduces the statistical framework, and Section 1.2 makes precise the form of an optimal regime. We describe and contrast $Q$- and $A$-learning in Section 1.3. Finally, Section 1.4 offers a comparison and discussion of practical considerations for $Q$- and $A$-learning methods.

In the remainder of this dissertation, we conduct a systematic comparison of $Q$- and $A$-learning. Specifically, in Chapter 2, we present extensive simulations evaluating their performance under a single decision point. The multiple decision point scenario is considered, again by extensive simulations, in Chapter 3. Finally, the methods are demonstrated using data from the Sequenced Treatment Alternatives to Relieve Depression
1.1 Framework and Assumptions

We consider the general setting of \( K \) prespecified, ordered decision points, indexed by \( k = 1, \ldots, K \), which may be times or events in the disease or disorder process that necessitate a treatment decision, where, at each point, a set of treatment options is available. Assume that there is a final outcome \( Y \) of interest for which, without loss of generality, large values are preferred. The outcome may be ascertained following the \( K \)th decision, as in the case of CD4 T-cell count at a prespecified follow-up time in a study of HIV infection (Moodie et al. 2007); or may be a function of information accrued over the entire sequence of decisions, as in Henderson et al. (2010), where outcome is the overall proportion of time a measure of blood clotting speed is kept within a target range in a study of dosing of anticoagulant agents.

In order to define an optimal treatment regime and discuss estimation of an optimal regime based on data from an observational study or clinical trial, we first define a suitable conceptual framework. For simplicity, our presentation is heuristic. We imagine that there is a superpopulation of patients, denoted by \( \Omega \), where one may view an element \( \omega \in \Omega \) as a patient from this population. We assume that patients in the population have been treated and otherwise have behaved according to routine clinical practice for the disease or disorder prior to the first treatment decision. Consequently, immediately prior to this first decision, patient \( \omega \) would present to the decision-maker with a set of baseline information (covariates) denoted by the random variable \( S_1 \); we discuss this further below. Thus, \( S_1(\omega) \) is the value of his/her information immediately prior to decision 1 under these conditions, taking values \( s_1 \), say, in a set \( S_1 \). Assume that, at each decision point \( k = 1, \ldots, K \), there is a set of possible treatment options \( A_k \), where we denote elements of \( A_k \) by \( a_k \). We write \( a_k = (a_1, \ldots, a_k) \) to denote a possible treatment history that could be administered through the \( k \)th decision, taking values in the corresponding set \( \bar{A}_k = A_1 \times \cdots \times A_k \). Thus, \( \bar{A}_K \) denotes the set of all possible full treatment histories \( \bar{a}_K \) through all \( K \) decisions.
We then define the potential outcomes (Robins 1986)

\[ W = \{S_2^*(a_1), S_3^*(a_2), \ldots, S_k^*(a_{k-1}), \ldots, S_K^*(a_{K-1}), Y^*(a_K) \text{ for all } a_K \in \mathcal{A}_K\}. \] (1.1)

In (1.1), \( S_k^*(a_{k-1})(\omega) \) denotes the value of covariate information that would arise between the \((k-1)\)th and \(k\)th decision points for a patient \(\omega \in \Omega\) under the hypothetical situation that s/he were to have received previously treatment history \(a_{k-1}\), taking values \(s_k\) in a set \(S_k\), \(k = 2, \ldots, K\). Similarly, \(Y^*(a_K)(\omega)\) is the hypothetical outcome that would result for patient \(\omega\) were s/he to have been administered the full set of \(K\) treatments in \(a_K\). Here and henceforth, this notation implies that, for random variables such as \(S_k^*(a_{k-1})\), \(a_{k-1}\) is an index representing prior treatment history. For convenience, write \(\tilde{S}_k^*(a_{k-1}) = \{S_1, S_2^*(a_1), \ldots, S_k^*(a_{k-1})\}\), \(k = 1, \ldots, K\), where \(\tilde{S}_k^*(a_{k-1})(\omega)\) takes values \(\bar{s}_k\) in \(\tilde{S}_k = S_1 \times \cdots \times S_k\); note that this definition includes the baseline covariate \(S_1\) and is taken equal to \(S_1\) when \(k = 1\). In what follows, for simplicity, we take all random variables to be discrete, but the results we present hold more generally.

Let the random variables \(A_1^{(P)}, \ldots, A_K^{(P)}\) denote the treatments that would be assigned to patients in the population at decisions 1, \ldots, \(K\) under routine clinical practice, so that \(A_k^{(P)}(\omega)\) is the treatment in \(\mathcal{A}_k\) that patient \(\omega\) would receive at decision \(k\), taking values \(a_k \in \mathcal{A}_k\). By routine clinical practice, we mean the conditions under which patients in the population and their providers would make treatment decisions acting as they see fit, emphasized by the superscript \((P)\) (for “population”), to be distinguished from those of a clinical trial, discussed later. Thus, the \(A_k^{(P)}\) characterize the mechanism by which treatments are assigned in the population if patients and clinicians are left to their own devices. Likewise, define the random variables \(S_k^{(P)}\), \(k = 2, \ldots, K\), to be the covariate information that would be observed on patients in the population between decisions \(k-1\) and \(k\) under the treatment assignments \(A_k^{(P)}\), taking values \(s_k \in \mathcal{S}_k\); let \(Y^{(P)}\) be the corresponding observed outcome, taking values \(y\) in a set \(\mathcal{Y}\); and define \(\bar{A}_k^{(P)} = (A_1^{(P)}, \ldots, A_k^{(P)})\), taking values \(\bar{a}_k \in \bar{\mathcal{A}}_k\). Henceforth, as is standard, we make the consistency assumption (e.g., Robins 1994) that the covariates and outcomes that would be observed under these conditions are those that potentially would be seen under the treatments actually received; that is, for patient \(\omega \in \Omega\), \(S_k^{(P)}(\omega) = S_k^*\{A_k^{(P)}(\omega)\}(\omega)\), \(k = 2, \ldots, K\), and \(Y^{(P)}(\omega) = Y^*\{\bar{A}_K^{(P)}(\omega)\}(\omega)\). We also make the stable unit treatment value assumption (Rubin 1978), which ensures that a patient’s covariates and outcome are unaffected by how treatments are allocated to her/him and other patients.
Under this conceptualization, probabilities for events in $\Omega$ are induced by random sampling from this population, as are all probability distributions of the potential data above and observed data that would be obtained from studies carried out in the population. The goal of $Q$- and $A$-learning is to estimate the optimal treatment regime based on data from an observational study or clinical trial carried out in a random sample from this population.

A dynamic treatment regime $d = (d_1, \ldots, d_K)$ is a set of rules that dictates an algorithm for treating a patient over time. At the $k$th decision point, the $k$th rule $d_k(\bar{s}_k, \bar{a}_{k-1})$, say, takes as input the patient’s realized covariate and treatment history prior to the $k$th treatment decision and outputs a value $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1}) \subseteq A_k$; for $k = 1$, there is no prior treatment ($a_0$ is null), and we write $d_1(s_1)$ and $\Psi_1(s_1)$. Here, $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ is the set of feasible treatment options for a patient with realized history $(\bar{s}_k, \bar{a}_{k-1})$, reflecting that some treatment options may be unethical or impossible for patients with certain histories. We discuss considerations for identifying the $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ shortly. Because we consider only regimes where $d_k(\bar{s}_k, \bar{a}_{k-1}) \in \Psi_k(\bar{s}_k, \bar{a}_{k-1}) \subseteq A_k$, $d_k$ need only map a subset of $\mathcal{S}_k \times \mathcal{A}_{k-1}$ to $A_k$. We define these subsets recursively as

$$
\Gamma_k = \left\{ (\bar{s}_k, \bar{a}_{k-1}) \in \mathcal{S}_k \times \mathcal{A}_{k-1} \text{ satisfying } \right. \\
\left. (i) \ a_j \in \Psi_j(\bar{s}_j, \bar{a}_{j-1}), \ j = 1, \ldots, k-1, \text{ and } (ii) \ \text{pr}\{S^*_k(\bar{a}_{k-1}) = \bar{s}_k\} > 0 \right\} \tag{1.2}
$$

for $k = 1, \ldots, K$. Thus, we may define formally the class of all feasible treatment regimes $\mathcal{D}$, say, as the set of all $d = (d_1, \ldots, d_K)$ for which $d_k$, $k = 1, \ldots, K$, is a mapping from $\Gamma_k$ into $A_k$ satisfying $d_k(\bar{s}_k, \bar{a}_{k-1}) \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$ for every $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$.

Intuitively, an optimal regime should represent the “best” way to intervene to treat patients in $\Omega$ who would otherwise behave according to routine clinical practice. We now state with specificity what we mean by this. To this end, for any $d \in \mathcal{D}$, writing $\bar{d}_k = (d_1, \ldots, d_k)$, $k = 1, \ldots, K$, $\bar{d}_K = d$, define the potential outcomes $\{S^*_2(d_1), \ldots, S^*_k(\bar{d}_{k-1}), \ldots, S^*_K(\bar{d}_{K-1}), Y^*(d)\}$ associated with a regime $d \in \mathcal{D}$ such that, for any $\omega \in \Omega$, with $S_1(\omega) = s_1$,

$$
d_1(s_1) = u_1, \ S^*_2(d_1)(\omega) = S^*_2(u_1)(\omega) = s_2, \ d_2(s_2, u_1) = u_2, \ldots, \\
d_{K-1}(\bar{s}_{K-1}, \bar{u}_{K-2}) = u_{K-1}, \ S^*_K(\bar{d}_{K-1})(\omega) = S^*_K(\bar{u}_{K-1})(\omega) = s_K, \\
d_K(\bar{s}_K, \bar{u}_{K-1}) = u_K, \ Y^*(d)(\omega) = Y^*(\bar{u}_K)(\omega) = y. \tag{1.3}
$$
The index \( \bar{d}_{k-1} \) emphasizes that \( \bar{S}_k^*(\bar{d}_{k-1})(\omega) \) represents the covariate information that would arise between decisions \( k-1 \) and \( k \) were patient \( \omega \) to receive the treatments sequentially dictated by the first \( k-1 \) rules in \( d \). Similarly, \( Y^*(d)(\omega) \) is the final outcome that \( \omega \) would experience if s/he were to receive the \( K \) treatments dictated by \( d \).

With these definitions, the expected outcome in the population if all patients with initial state \( S_1 = s_1 \) were to follow regime \( d \) is \( \mathbb{E}\{Y^*(d)|S_1 = s_1\} \). An optimal regime, \( d^{\text{opt}} \in \mathcal{D} \), say, satisfies

\[
\mathbb{E}\{Y^*(d)|S_1 = s_1\} \leq \mathbb{E}\{Y^*(d^{\text{opt}})|S_1 = s_1\} \quad \text{for all } d \in \mathcal{D} \text{ and all } s_1 \in S_1. \tag{1.4}
\]

In Section 1.2, we give the form of \( d^{\text{opt}} \) satisfying (1.4) and demonstrate further optimality properties.

Of course, potential outcomes for a given patient for all \( d \in \mathcal{D} \) are not observed. Thus, the goal is to estimate \( d^{\text{opt}} \) in (1.4) using data from a study carried out on a random sample of \( n \) patients from \( \Omega \) that record baseline and evolving covariate information and the treatments actually received by the participants. We denote the available study data as independent and identically distributed (i.i.d.) time-ordered random variables \( (S_{1i}, A_{1i}, \ldots, S_{Ki}, A_{Ki}, Y_i) \) \( i = 1, \ldots, n \) on \( \Omega \). Here, \( S_1 \) is as before; \( S_k, k = 2, \ldots, K, \) is the covariate information recorded between decisions \( k-1 \) and \( k \), taking values \( s_k \in S_k \); \( A_k, k = 1, \ldots, K, \) is the recorded, observed treatment assignment, taking values \( a_k \in A_k \); and \( Y \) is the observed outcome, taking values \( y \in Y \). As above, we define \( \bar{S}_k = (S_1, \ldots, S_k) \) and \( \bar{A}_k = (A_1, \ldots, A_k), k = 1, \ldots, K, \) taking values \( \bar{s}_k \in \bar{S}_k \) and \( \bar{a}_k \in \bar{A}_k \).

It is important to recognize that the nature of the study generating the available data must be considered carefully. If the data arise from an observational study in which covariate, treatment, and outcome information on \( n \) participants randomly sampled from \( \Omega \) is recorded, with no intervention by investigators, then it is reasonable to assume that the mechanism by which treatments are assigned to the patients in the sample during the study is the same as that for the entire population under routine practice. In this case, for \( k = 1, \ldots, K, A_k = A_k^{(P)} \), so that, under the consistency assumption, for \( k = 2, \ldots, K, S_k(\omega) = S_k^{(P)}(\omega) = S_k^*[\bar{A}_{k-1}^{(P)}(\omega)](\omega) \) and \( Y(\omega) = Y^{(P)}(\omega) = Y^*[\bar{A}_K^{(P)}(\omega)](\omega) \). Here, the form of \( \Psi_k(\bar{s}_k, \bar{a}_{k-1}), k = 1, \ldots, K, \) is determined by treatment choices dictated by clinical practice.

Such a correspondence between the \( S_k, A_k \) and \( S_k^{(P)}, A_k^{(P)} \) is not the case for an intervention study. A clinical trial design that has been advocated for collecting data
suitable for estimating optimal treatment regimes is that of a so-called sequential multiple-assignment randomized trial (SMART, Lavori and Dawson 2000; Murphy 2005). In a SMART involving $K$ pre-specified decision points, each participant is randomized at each decision point to one of a set of feasible treatment options, where, at the $k$th decision, the randomization probabilities may depend on past realized information $\bar{s}_k, \bar{a}_{k-1}$. As we discuss further shortly, as with any clinical trial, an advantage is that the usual issues of confounding associated with an observational study are obviated. However, the treatment assignment mechanism in the study is no longer the same as that in the population under routine practice. More precisely, the sample space is now $\Omega \times \mathcal{A}_K$, where for any element $(\omega \times \bar{a}_K)$, $\omega$ represents the patient randomly sampled from the population, and $\bar{a}_K$ represents the treatments assigned to her/him at all $K$ decisions by the random mechanism dictated by the trial design. Here, then, the observed $A_k(\omega \times \bar{a}_K) = a_k$ and $S_k(\omega \times \bar{a}_k) = S^*_k(\bar{a}_{k-1})(\omega)$. Thus, in contrast to an observational study, $A_k \neq A_k^{(P)}$ and $S_k \neq S_k^{(P)}$ in general. Moreover, the treatment options in $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ are dictated by the trial design so may be different from those in routine practice. In particular, the set of treatment options at each decision might be restricted relative to those available in clinical practice for reasons of logistics, cost, or interest of the trial sponsor in only certain products. We discuss further considerations for using data from a SMART to estimate optimal regimes in Section B of the Appendix.

In order to use the observed data from either type of study to estimate an optimal regime, the critical assumption of no unmeasured confounders, also referred to as the sequential randomization assumption (Robins 1994), must be satisfied. A version of this assumption states that $A_k$ is conditionally independent of $W$ given $\{\bar{S}_k, \bar{A}_{k-1}\}$, $k = 1, \ldots, K$, where $A_0$ is null, written $A_k \perp \perp W|\bar{S}_k, \bar{A}_{k-1}$. In a SMART, this assumption is automatically satisfied by design. In an observational study, this assumption is unverifiable from the observed data. Although in the population patients and their providers may make treatment decisions based on past covariate information available to them, the issue is whether or not all of this information is recorded in the $S_k$; see Section B of the Appendix.
1.2 Defining the Optimal Regime

$Q$- and $A$-learning are two approaches to estimating $d^{\text{opt}}$ satisfying (1.4) under the foregoing framework and assumptions. Both involve similar recursive fitting algorithms; the main distinguishing feature is the form of the respective underlying models. To appreciate the rationale for the methods, one must first understand how $d^{\text{opt}}$ is determined via dynamic programming, also referred to as backward induction. We demonstrate the formulation of $d^{\text{opt}}$ in terms of the potential outcomes and then show how $d^{\text{opt}}$ may be expressed in terms of the observed data under assumptions including those of the last section. In the following, we sometimes highlight dependence on specific elements of quantities such as $\bar{a}_k$, writing, for example, $\bar{a}_k$ as $(\bar{a}_{k-1}, a_k)$.

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

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

(1.5)

$$V_K^{(1)}(\bar{s}_K, \bar{a}_{K-1}) = \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\{Y^*(\bar{a}_{K-1}, a_K)|\bar{S}^*_K(\bar{a}_{K-1}) = \bar{s}_K\}.$$  

(1.6)

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

$$d_k^{(1)\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = \arg \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathbb{E}\{V_{k+1}^{(1)}(\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k)|\bar{S}^*_k(\bar{a}_{k-1}) = \bar{s}_k\},$$

(1.7)

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

(1.8)

so that, for $s_1 \in S_1$, $d_1^{(1)\text{opt}}(s_1) = \arg \max_{a_1 \in \Psi_1(s_1)} \mathbb{E}\{V_2^{(1)}(s_1, S_2^*(a_1), a_1)|S_1 = s_1\}$, and $V_1^{(1)}(s_1) = \max_{a_1 \in \Psi_1(s_1)} \mathbb{E}\{V_2^{(1)}(s_1, S_2^*(a_1), a_1)|S_1 = s_1\}$. Note that the above conditional expectations are well-defined by condition (ii) in (1.2) defining $\Gamma_k$.

It is clear that $d^{(1)\text{opt}} = (d_1^{(1)\text{opt}}, \ldots, d_K^{(1)\text{opt}})$ defined above is a treatment regime, as it comprises a set of rules that uses patient information prior to each decision to assign treatment from among the feasible options. The superscript (1) indicates that $d^{(1)\text{opt}}$ provides a set of $K$ rules for a patient presenting prior to decision point 1 with baseline information $S_1 = s_1$. Note that $d^{(1)\text{opt}}$ is defined in a backward iterative fashion. At the
The $K$th decision, (1.5) gives the treatment among the feasible options at decision $K$ that maximizes the expected potential final outcome given the prior potential information available, and (1.6) is the maximum value achieved. At decisions $k = K - 1, \ldots, 1$, intuitively, (1.7) gives the treatment that maximizes the expected outcome that would be achieved if subsequent optimal rules already defined were followed henceforth.

In Section A of the Appendix, we provide a formal argument demonstrating that $d^{(1)}_{opt}$ defined in (1.5)–(1.8) is an optimal treatment regime in the sense of satisfying (1.4). Note that, because (1.4) is true for any $s_1$, in fact $E\{Y^*(d)\} \leq E\{Y^*(d^{(1)}_{opt})\}$ for any $d \in D$. Thus, from a policy perspective, $d^{(1)}_{opt}$ defines the optimal strategy for treating patients in the population through all $K$ decisions were they to be encountered at the stage of the disease or disorder that precedes decision point 1.

In routine clinical practice, however, patients may be encountered at later stages. Consider a patient $\omega \in \Omega$ for whom the first $\ell - 1$ treatment decisions have been made as seen fit by her/him and her/his provider, $\ell = 2, \ldots, K$. Immediately prior to the $\ell$th decision, the patient would have past history $\bar{s}_{\ell-1}(\omega) = \bar{a}_{\ell-1}$, raising the issue of how best to intervene to treat such a patient henceforth, from the $\ell$th to $K$th decisions. That is, we desire rules $d^{(\ell)}_k(\bar{s}_k, \bar{a}_{k-1})$, $k = \ell, \ell + 1, \ldots, K$, say, that dictate how to treat such patients.

Write $d^{(\ell)} = (d^{(\ell)}_\ell, d^{(\ell)}_{\ell + 1}, \ldots, d^{(\ell)}_K)$ to denote regimes starting at the $\ell$th decision point. Analogous to the above, we define the class $D^{(\ell)}$ of all such feasible regimes to be the set of all $d^{(\ell)}$ for which $d^{(\ell)}_k(\bar{s}_k, \bar{a}_{k-1}) = a_k$ for $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma^{(\ell)}_k$ and $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$ for $k = \ell, \ldots, K$, where

\[
\Gamma^{(\ell)}_k = \{(\bar{s}_k, \bar{a}_{k-1}) \in \mathcal{S}_k \times \mathcal{A}_{k-1} \text{ satisfying} \]

\[(i) a_j \in \Psi_j(\bar{s}_j, \bar{a}_{j-1}), j = \ell, \ldots, k - 1, \text{ and (ii) } \Pr\{\mathcal{V}_{\ell,k}\} > 0 \},
\]

$\mathcal{V}_{\ell,k}$ is the event $\{\bar{s}^{(P)}_{\ell} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}, S_{\ell+1}^*(\bar{a}_\ell) = s_{\ell+1}, \ldots, S_k^*(\bar{a}_{k-1}) = s_k\}$. Then, by analogy to (1.4), we seek $d^{(\ell)}_{opt}$ satisfying

\[
E\{Y^*(\bar{a}_{\ell-1}, d^{(\ell)})|\bar{s}^{(P)}_{\ell} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}\} \leq E\{Y^*(\bar{a}_{\ell-1}, d^{(\ell)}_{opt})|\bar{s}^{(P)}_{\ell} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}\}
\]

(1.9)

for all $d^{(\ell)} \in D^{(\ell)}$ and $\bar{s}_\ell \in \bar{S}_\ell$, $\bar{a}_{\ell-1} \in \bar{A}_{\ell-1}$ for which $\Pr(\bar{s}^{(P)}_{\ell} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}) > 0$. Viewing this as a problem of making $K - \ell + 1$ decisions at decision points $\ell, \ell + 1, \ldots, K$, ...
with initial state \( \bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1} \), by an argument analogous to that in Section A of the Appendix for \( \ell = 1 \) and initial state \( S_1 = s_1 \), it may be shown that \( d^{(1)}_{\ell} \) satisfying (1.9) is given by

\[
\begin{align*}
\bar{d}_{K}^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) &= \arg \max_{\bar{a}_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\{Y^*(\bar{a}_{K-1}, a_K)|\mathcal{V}_{\ell,K}\}, \quad \text{(1.10)} \\
\bar{V}_K^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) &= \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\{Y^*(\bar{a}_{K-1}, a_K)|\mathcal{V}_{\ell,K}\} \quad \text{(1.11)}
\end{align*}
\]

for any \( \bar{s}_K \in \bar{S}_K, \bar{a}_{K-1} \in \bar{A}_{K-1} \) for which \( (\bar{s}_K, \bar{a}_{K-1}) \in \Gamma_K^{(\ell)} \); and, for \( k = K-1, \ldots, \ell \),

\[
\begin{align*}
\bar{d}_{k}^{(\ell)}(\bar{s}_k, \bar{a}_{k-1}) &= \arg \max_{\bar{a}_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathbb{E}\{V_{k+1}^{(\ell)}(\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k)|\mathcal{V}_{\ell,k}\}, \quad \text{(1.12)} \\
\bar{V}_k^{(\ell)}(\bar{s}_k, \bar{a}_{k-1}) &= \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathbb{E}\{V_{k+1}^{(\ell)}(\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k)|\mathcal{V}_{\ell,k}\} \quad \text{(1.13)}
\end{align*}
\]

for any \( \bar{s}_k \in \bar{S}_k, \bar{a}_{k-1} \in \bar{A}_{k-1} \) for which \( (\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k^{(\ell)} \), so that

\[
\bar{d}_{\ell}^{(\ell)}(\bar{s}_\ell, \bar{a}_{\ell-1}) = \arg \max_{a_\ell \in \Psi_\ell(\bar{s}_\ell, \bar{a}_{\ell-1})} \mathbb{E}\{V_{\ell+1}^{(\ell)}(\bar{s}_\ell, S_{\ell+1}^*(\bar{a}_{\ell-1}, a_\ell), \bar{a}_{\ell-1}, a_\ell)|\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}\}.
\]

Comparison of (1.5)–(1.8) to (1.10)–(1.13) shows that the \( \ell \)th to \( K \)th rules of the optimal regime \( d^{(1)}_{\ell} \) that would be followed by a patient presenting at the first decision are not necessarily the same as those of the optimal regime \( d^{(1)}_{\ell} \) that would be followed by a patient presenting at the \( \ell \)th decision. In particular, noting that the conditioning sets in (1.5)–(1.8) are \( \mathcal{V}_{1,K} \) and \( \mathcal{V}_{1,k} \), the rules are \( \ell \)-dependent through dependence of the conditioning sets \( \mathcal{V}_{\ell,k}, \ell = 1, \ldots, K, k = \ell, \ldots, K \), on \( \ell \). However, we demonstrate shortly that these rules coincide under certain conditions.

The foregoing developments define optimal regimes in terms of potential outcomes. To be useful in practice, an optimal regime must be defined in terms of the observed data. To this end, define

\[
\begin{align*}
Q_K(\bar{s}_K, \bar{a}_K) &= \mathbb{E}\{Y|\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K\}, \quad \text{(1.14)} \\
\bar{d}_{K}^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) &= \arg \max_{\bar{a}_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} Q_K(\bar{s}_K, \bar{a}_{K-1}, a_K), \quad \text{(1.15)} \\
\bar{V}_K^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) &= \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} Q_K(\bar{s}_K, \bar{a}_{K-1}, a_K), \quad \text{(1.16)}
\end{align*}
\]

for any \( \bar{s}_K \in \bar{S}_K, \bar{a}_K \in \bar{A}_K \) for which \( \text{pr}(\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}) > 0 \); and for \( k = \)
for any $\bar{s}_k \in \bar{S}_k, \bar{a}_k \in \bar{A}_k$ for which $\text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$. Note that all quantities in (1.14)–(1.19) are expressed entirely in terms of the distribution of the observed data.

In Section B of the Appendix, under the consistency and sequential randomization (no unmeasured confounders) assumptions, along with positivity assumptions on probabilities associated with events involving $\bar{S}_K, \bar{A}_K$ and $\bar{S}_K^{(P)}, \bar{A}_K^{(P)}$ given in Section B of the Appendix, we show that

$$
\Gamma_k \quad \text{for } \ell = 1, \ldots, K \text{ and } k = \ell, \ldots, K.
$$

The equivalence in (1.20)–(1.22) not only demonstrates that an optimal treatment regime can be obtained using the distribution of the observed data but also that the corresponding rules dictating treatment do not depend on $\ell$ under these assumptions. Thus, (1.20)–(1.22) imply that the single set of rules $d^{\text{opt}} = (d_1^{\text{opt}}, \ldots, d_K^{\text{opt}})$ defined in (1.15) and (1.18) is relevant regardless of when a patient presents. That is, treatment at the $\ell$th decision point for a patient who presents at decision 1 and has followed the rules in $d^{\text{opt}}$ to that point would be determined by $d^{\text{opt}}_\ell$ evaluated at his/her history up to that point, as would treatment for a subject presenting for the first time immediately prior to decision $\ell$.

The $Q_k(\bar{s}_k, \bar{a}_k)$ in (1.14) and (1.17) are referred to as the “$Q$-functions,” viewed as measuring the “quality” associated with using treatment $a_k$ at decision $k$ given the history up to that decision and then following the optimal regime thereafter. The “value functions” $V_k(\bar{s}_k, \bar{a}_{k-1})$ in (1.16) and (1.19) reflect the “value” of a patient’s history $\bar{s}_k, \bar{a}_{k-1}$ assuming that optimal decisions are made in the future.

It is worth noting that there may not be a unique $d^{\text{opt}}$. At any decision point $k$, if there is more than one feasible treatment option $a_k$ leading to the maximum value of the
Q-function, then any rule $d_k^{\text{opt}}$ yielding one of these $a_k$ defines an optimal regime.

## 1.3 Q- and A-Learning

### 1.3.1 Q-Learning

From (1.15) and (1.18), the optimal regime $d_k^{\text{opt}}$ is defined in terms of the $Q$-functions (1.14), (1.17). Thus, estimation of $d_k^{\text{opt}}$ based on i.i.d. data $(S_{i1}, A_{i1}, \ldots, S_{Ki}, A_{Ki}, Y_i), i = 1, \ldots, n$, may be accomplished via direct modeling and fitting of the $Q$-functions. This is the approach underlying Q-learning. Specifically, one may posit models $Q_k(\tilde{s}_k, \tilde{a}_k; \xi_k)$, say, for $k = K, K - 1, \ldots, 1$, each depending on a finite-dimensional parameter $\xi_k$. The models may be linear or nonlinear in $\xi_k$ and include main effects and interactions in the elements of $\tilde{s}_k$ and $\tilde{a}_k$.

Estimators $\hat{\xi}_k$ may be obtained in a backward iterative fashion for $k = K, K - 1, \ldots, 1$ by solving suitable estimating equations [e.g., ordinary (OLS) or weighted (WLS) least squares]. Assuming the latter, for $k = K$, letting $\tilde{V}_{(K+1)i} = Y_i$, one would first solve

$$
\sum_{i=1}^{n} \frac{\partial Q_K(\tilde{S}_{Ki}, \tilde{A}_{Ki}; \xi_K)}{\partial \xi_K} \Sigma_K^{-1}(\tilde{S}_{Ki}, \tilde{A}_{Ki})\{\tilde{V}_{(K+1)i} - Q_K(\tilde{S}_{Ki}, \tilde{A}_{Ki}; \xi_K)\} = 0 \quad (1.23)
$$

in $\xi_K$ to obtain $\hat{\xi}_K$, where $\Sigma_K(\tilde{s}_K, \tilde{a}_K)$ is a working variance model. Substituting the model $Q_K(\tilde{s}_K, \tilde{a}_K; \xi_K)$ in (1.15) and accordingly writing $d_K^{\text{opt}}(\tilde{s}_K, \tilde{a}_{K-1}; \xi_K)$, substituting $\hat{\xi}_K$ for $\xi_K$ yields an estimator for the optimal treatment choice at decision $K$ for a patient with past history $\tilde{S}_K = \tilde{s}_K, \tilde{A}_{K-1} = \tilde{a}_{K-1}$. With $\hat{\xi}_K$ in hand, one would form for each $i$, based on (1.16), $\tilde{V}_{Ki} = \max_{a_K \in \Psi_K(\tilde{s}_{Ki}, \tilde{A}_{(K-1)i})} Q_K(\tilde{S}_{Ki}, \tilde{A}_{(K-1)i}, a_K; \hat{\xi}_K)$. To obtain $\hat{\xi}_{K-1}$, setting $k = K - 1$, based on (1.17), one would then solve in $\xi_K$

$$
\sum_{i=1}^{n} \frac{\partial Q_k(\tilde{S}_{ki}, \tilde{A}_{ki}; \xi_k)}{\partial \xi_k} \Sigma_k^{-1}(\tilde{S}_{ki}, \tilde{A}_{ki})\{\tilde{V}_{(k+1)i} - Q_k(\tilde{S}_{ki}, \tilde{A}_{ki}; \xi_k)\} = 0, \quad (1.24)
$$

where $\Sigma_k(\tilde{s}_k, \tilde{a}_k)$ is a working variance model. The corresponding $d_{K-1}^{\text{opt}}(\tilde{s}_{K-1}, \tilde{a}_{K-2}; \hat{\xi}_{K-1})$ then yields an estimator for the optimal treatment choice at decision $K - 1$ for a patient with past history $\tilde{S}_{K-1} = \tilde{s}_{K-1}, \tilde{A}_{K-2} = \tilde{a}_{K-2}$, assuming s/he will take the optimal treatment at decision $K$. One would continue this process in the obvious fashion for
\(k = K - 2, \ldots, 1\), forming \(\tilde{V}_k = \max_{a_k \in \Psi_k(s_{ki}, A_{(k-1)i})} Q_k(\tilde{S}_{ki}, \tilde{A}_{(k-1)i}, a_k; \tilde{\xi}_k)\), and solving equations of form (1.24) to obtain \(\tilde{\xi}_k\) and corresponding \(d_{k}^{\text{opt}}(\tilde{s}_k, \tilde{a}_{k-1}; \tilde{\xi}_k)\).

We may now summarize the estimated optimal regime as \(\hat{d}_Q^{\text{opt}} = (\hat{d}_{Q,1}^{\text{opt}}, \ldots, \hat{d}_{Q,K}^{\text{opt}})\), where

\[
\hat{d}_{Q,k}^{\text{opt}}(s_1) = d_{1,k}^{\text{opt}}(s_1; \hat{\xi}_1), \ldots, \hat{d}_{Q,k}^{\text{opt}}(\tilde{s}_k, \tilde{a}_{k-1}) = d_{k}^{\text{opt}}(\tilde{s}_k, \tilde{a}_{k-1}; \hat{\xi}_k), \quad k = 2, \ldots, K.
\]

(1.25)

It is important to recognize that the estimated regime (1.25) may not be a credible estimator for the true optimal regime unless all the models for the \(Q\)-functions are correctly specified.

We illustrate for the case \(K = 2\), where at each decision there are two feasible treatment options coded as 0 and 1; i.e., \(\Psi_1(s_1) = A_1 = \{0, 1\}\) for all \(s_1\) and \(\Psi_2(\tilde{s}_2, a_1) = A_2 = \{0, 1\}\) for all \(\tilde{s}_2\) and \(a_1 \in \{0, 1\}\). Let \(H_1 = (1, s_1^T)T\) and \(H_2 = (1, s_1^T, a_1, s_2^T)T\). As in many modeling contexts, it is standard to adopt linear models for the \(Q\)-functions; accordingly, consider the models

\[
Q_1(s_1, a_1; \xi_1) = H_1^T \beta_1 + a_1(\hat{H}_1^T \psi_1), \quad Q_2(\tilde{s}_2, \tilde{a}_2; \xi_2) = H_2^T \beta_2 + a_2(\hat{H}_2^T \psi_2),
\]

(1.26)

where \(\xi_k = (\beta_k^T, \psi_k^T)^T\) for \(k = 1, 2\). Note that \(Q_2(\tilde{s}_2, \tilde{a}_2; \xi_2)\) in (1.26) is a model for \(E(Y|\tilde{S}_2 = \tilde{s}_2, A_2 = \tilde{a}_2)\), which is a standard regression problem involving observable data, whereas \(Q_1(s_1, a_1; \xi_1)\) is a model for the conditional expectation of \(V_2(\tilde{s}_2, a_1) = \max_{a_2 \in \{0, 1\}} E(Y|\tilde{S}_2 = s_2, A_1 = a_1, A_2 = a_2)\) given \(S_1 = s_1\) and \(A_1 = a_1\), which is an approximation to a complex relationship involving a maximization. Under (1.26), it is straightforward to deduce that \(V_2(\tilde{s}_2, a_1; \xi_2) = \max_{a_2 \in \{0, 1\}} Q_2(\tilde{s}_2, a_1, a_2; \xi_2) = H_2^T \beta_2 + (\hat{H}_2^T \psi_2)I(\hat{H}_2^T \psi_2 > 0)\) and \(V_1(s_1; \xi_1) = \max_{a_1 \in \{0, 1\}} Q_1(s_1, a_1; \xi_1) = H_1^T \beta_1 + (\hat{H}_1^T \psi_1)I(\hat{H}_1^T \psi_1 > 0)\). Substituting the \(Q\)-functions in (1.26) in (1.15) and (1.18) then yields \(d_{1,1}^{\text{opt}}(s_1; \xi_1) = I(\hat{H}_1^T \psi_1 > 0)\) and \(d_{2,2}^{\text{opt}}(s_2, a_1; \xi_2) = I(\hat{H}_2^T \psi_2 > 0)\).

We have presented (1.23) and (1.24) in the conventional WLS form, with leading term in the summand \(\partial/d \xi_k Q_k(S_{ki}, A_{ki}; \xi_k)\Sigma_k^{-1}(S_{ki}, A_{ki})\); taking \(\Sigma_k\) to be a constant yields OLS. At the \(K\)th decision, with responses \(Y_i\), standard theory implies that this is the optimal leading term when \(\text{var}(Y|S_K = s_K, A_K = a_K) = \Sigma_K(s_K, a_K)\), yielding the efficient (asymptotically) estimator for \(\xi_K\). For \(k < K\), with “responses” \(\tilde{V}_{(k+1)i}\), this theory may no longer apply; however, deriving the optimal leading term involves considerable complication. Accordingly, it is standard to fit the posited models \(Q_k(\tilde{s}_k, \tilde{a}_k; \xi_k)\) via
OLS or WLS; some authors define $Q$-learning as using OLS (Chakraborty, Murphy, and Strecher 2010). The choice may be dictated by apparent relevance of the homoscedasticity assumption on the $\hat{V}_{(k+1)i}$: $k = K, K - 1, \ldots, 1$, and whether or not linear models are sufficient to approximate the relationships may also be evaluated, but see Section 1.4.1.

### 1.3.2 A-Learning

Advantage learning (A-learning, Murphy 2003) is an alternative to $Q$-learning that involves making fewer assumptions on the form of the $Q$-functions. For simplicity, we consider the case of two feasible treatment options coded as 0 and 1 at each decision; i.e., $\Psi_k(\bar{s}_k, \bar{a}_{k-1}) = A_k = \{0, 1\}$, $k = 1, \ldots, K$, though the methodology can be extended to an arbitrary number of treatments at each stage at the expense of complicating the formulation and notation.

To fix ideas, consider (1.26). Note that $d_i^{opt}(s_1; \xi_1)$ implied by (1.26) depends only on $\mathcal{H}_1^T\psi_1 = Q_1(s_1, 1; \xi_1) - Q_1(s_1, 0; \xi_1)$; likewise, $d_i^{opt}(\bar{s}_2, a_1; \xi_2)$ depends on $\mathcal{H}_2^T\psi_2 = Q_2(\bar{s}_2, a_1, 1; \xi_2) - Q_2(\bar{s}_2, a_1, 0; \xi_2)$. This is a special case of the general result that, for purposes of deducing the optimal regime, for each $k = 1, \ldots, K$, it suffices to know the contrast function $C_k(\bar{s}_k, \bar{a}_{k-1}) = Q_k(\bar{s}_k, \bar{a}_{k-1}, 1) - Q_k(\bar{s}_k, \bar{a}_{k-1}, 0)$. This can be appreciated by noting that any arbitrary $Q_k(\bar{s}_k, \bar{a}_k)$ may be written as $h_k(\bar{s}_k, \bar{a}_{k-1}) + a_kC_k(\bar{s}_k, \bar{a}_{k-1})$, where $h_k(\bar{s}_k, \bar{a}_{k-1}) = Q_k(\bar{s}_k, \bar{a}_{k-1}, 1)$, so that $Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k)$ is maximized by taking $a_k = I\{C_k(\bar{s}_k, \bar{a}_{k-1}) > 0\}$; and the maximum itself is the expression $h_k(\bar{s}_k, \bar{a}_{k-1}) + C_k(\bar{s}_k, \bar{a}_{k-1})I\{C_k(\bar{s}_k, \bar{a}_{k-1}) > 0\}$.

The premise of A-learning is thus to model the contrast functions rather than the full $Q$-functions as in $Q$-learning. For $k = K - 1, \ldots, 1$ the latter involve possibly complex relationships, raising concern over the consequences of model misspecification for estimation of the optimal regime. As identifying the optimal regime depends only on correct specification of the contrast functions, A-learning may be less sensitive to mismodeling.

We now describe the A-learning procedure. Assume posited models $C_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$, $k = 1, \ldots, K$, say, for the contrast functions, each depending on a parameter $\psi_k$. Consider the $K$th decision. Given $C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K)$, letting $\pi_K(\bar{s}_K, \bar{a}_{K-1}) = \Pr(A_K = 1|\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1})$ be the propensity of receiving treatment 1 in the observed data as a function of past history and writing $\tilde{V}_{(K+1)i} = Y_i$, Robins (2004) showed that all consistent and asymptotically normal estimators for $\psi_K$ are solutions to estimating equations of the
form

\[
\sum_{i=1}^{n} \lambda_K(\bar{S}_{K,i}, \bar{A}_{(K-1)i}) \{ A_{K_i} - \pi_K(\bar{S}_{K,i}, \bar{A}_{(K-1)i}) \} \\
\times \{ \bar{V}_{(K+1)i} - A_{K_i} C_K(\bar{S}_{K,i}, \bar{A}_{(K-1)i}; \psi_K) - \theta_K(\bar{S}_{K,i}, \bar{A}_{(K-1)i}) \} = 0 \tag{1.27}
\]

for arbitrary functions \( \lambda_K(\bar{s}_K, \bar{a}_{K-1}) \) of the same dimension as \( \psi_K \) and \( \theta_K(\bar{s}_K, \bar{a}_{K-1}) \). Assuming the model \( C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K) \) is correct, if \( \text{var}(Y|\bar{S}_K = s_k, \bar{A}_{K-1} = a_{k-1}) \) is constant, the optimal choices of these functions are \( \lambda_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K) = \partial/\partial \psi_K C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K) \) and \( \theta_K(\bar{s}_K, \bar{a}_{(K-1)i}) = h_K(\bar{s}_K, \bar{a}_{K-1}) \); otherwise, the optimal \( \lambda_K \) is complex (Robins 2004).

To implement estimation of \( \psi_K \) via (1.27), one may adopt parametric models for these functions. Although the appeal of \( A \)-learning is that it obviates the need to specify fully the \( Q \)-functions, one may posit a model for the optimal \( \theta_K \), \( h_K(\bar{s}_K, \bar{a}_{K-1}; \beta_K) \), say. Moreover, unless the data are from a SMART study, in which case the propensities \( \pi_K(\bar{s}_K, \bar{a}_{K-1}) \) would be known, these may also be modeled as \( \pi_K(\bar{s}_K, \bar{a}_{K-1}; \phi_K) \) (e.g., by a logistic regression). These models are only adjuncts to estimating the parameter of interest, \( \psi_K \); interestingly, as long as at least one of these models is correctly specified, (1.27) will yield a consistent estimator for \( \psi_K \), the so-called double robustness property.

Substituting these models in (1.27), one solves (1.27) jointly in \((\psi_K^T, \beta_K^T, \phi_K^T)^T\) with

\[
\sum_{i=1}^{n} \frac{\partial h_K(\bar{S}_{K,i}, \bar{A}_{K-1}; \beta_K)}{\partial \beta_K} \{ \bar{V}_{(K+1)i} - A_{K_i} C_K(\bar{S}_{K,i}, \bar{A}_{(K-1)i}; \psi_K) - h_K(\bar{S}_{K,i}, \bar{A}_{(K-1)i}; \beta_K) \} = 0
\]

and the usual binary regression likelihood score equations in \( \phi_K \). We then have

\[
\hat{d}_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}; \psi_K) = I[C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K) > 0];
\]

as in \( Q \)-learning, substituting \( \hat{\psi}_K \) yields an estimator for the optimal treatment choice at decision \( K \) for a patient with past history \( \bar{S}_K = s_K, \bar{A}_{K-1} = a_{K-1} \).

With \( \hat{\psi}_K \) in hand, as with \( Q \)-learning, the \( A \)-learning algorithm proceeds in a backward iterative fashion to yield \( \hat{\psi}_k, k = K - 1, \ldots, 1 \). At the \( k \)th decision, given models \( h_k(\bar{s}_k, \bar{a}_{k-1}; \beta_k) \) and \( \pi_k(\bar{s}_k, \bar{a}_{k-1}; \phi_k) \), one solves jointly in \((\psi_k^T, \beta_k^T, \phi_k^T)\) a system of estimating equations analogous to those above. As in \( Q \)-learning, the \( k \)th set of equations is based on “responses” \( \hat{V}_{(k+1)i} \), where, for each \( i, \hat{V}_{ki} \) estimates \( V_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}) \). It may be
shown (see Section C of the Appendix) that

\[
E \left( V_{k+1}(S_{k+1}, A_k) + C_k(S_k, A_{k-1}) \mid I\{C_k(S_k, A_{k-1}) > 0\} - A_k \right) = V_k(S_k, A_{k-1}).
\]

The expression \( C_k(S_k, A_{k-1}) \mid I\{C_k(S_k, A_{k-1}) > 0\} - A_k \) is referred to as the advantage or regret function (Murphy 2003), as it represents the “advantage” in response incurred if the optimal treatment at the \( k \)th decision were given relative to that actually received (or, equivalently, the “regret” incurred by not using the optimal treatment). Accordingly, define recursively \( \bar{V}_{ki} = \bar{V}_{(k+1)i} + C_k(S_{ki}, A_{(k-1)i}; \widehat{\psi}_k)[I\{C_k(S_{ki}, A_{(k-1)i}; \widehat{\psi}_k) > 0\} - A_{ki}], k = K, K - 1, \ldots, 1, \bar{V}_{(K+1)i} = Y_i \). The equations at the \( k \)th decision are then

\[
\sum_{i=1}^{n} \lambda_k(S_{ki}, A_{(k-1)i}; \psi_k) \{A_{ki} - \pi_k(S_{ki}, A_{(k-1)i}; \phi_k)\} \times \{\bar{V}_{(k+1)i} - A_{ki}C_k(S_{ki}, A_{(k-1)i}; \psi_k) - h_k(S_{ki}, A_{(k-1)i}; \beta_k)\} = 0,
\]

\[
\sum_{i=1}^{n} \frac{\partial h_k(S_{k}, A_{k-1}; \beta_k)}{\partial \beta_k} \{\bar{V}_{(k+1)i} - A_{ki}C_k(S_{ki}, A_{(k-1)i}; \psi_k) - h_k(S_{ki}, A_{(k-1)i}; \beta_k)\} = 0,
\]

for a given specification \( \lambda_k(s_k, a_{k-1}; \psi_k) \), solved jointly with the maximum likelihood score equations for binary regression in \( \phi_k \). It follows that

\[
d^\text{opt}_k(s_k, a_{k-1}; \widehat{\psi}_k) = I[C_k(s_k, a_{k-1}; \widehat{\psi}_k) > 0].
\]

As above, the optimal \( \lambda_k \) is complex Robins (2004); a reasonable choice is to take \( \lambda_k(s_k, a_{k-1}; \psi_k) = \partial/\partial \psi_k C_k(s_k, a_{k-1}; \psi_k) \) for practical implementation.

Summarizing, the estimated optimal regime \( \hat{d}^\text{opt}_A = (\hat{d}^\text{opt}_{A,1}, \ldots, \hat{d}^\text{opt}_{A,K}) \) is

\[
\hat{d}^\text{opt}_{A,1}(s_1) = d^\text{opt}_{1}(s_1; \widehat{\psi}_1), \quad \hat{d}^\text{opt}_{A,k}(s_k, a_{k-1}) = d^\text{opt}_k(s_k, a_{k-1}; \widehat{\psi}_k), \quad k = 2, \ldots, K.
\]

As with Q-learning, how well \( \hat{d}^\text{opt}_A \) estimates \( d^\text{opt} \) depends on how close the \( C_k(s_k, a_{k-1}; \psi_k) \) are to the true contrast functions.
1.4 Summary and Discussion

In Sections 1.3.1 and 1.3.2, we described estimating parameters in a backwards iterative fashion under $Q$- and $A$-learning, respectively. At each stage, parameters were estimated under Equations 1.24 (1.28) for $Q$-learning ($A$-learning) and a new “response” $\hat{V}_{ki}$ was formed based on the parameter estimates and posited models. Recall that for $Q$-learning, we defined $\hat{V}_{ki} = \max_{a_k \in \psi_k(S_{ki}, A_{(k-1)i})} Q_k(S_{ki}, A_{(k-1)i}; a_k; \hat{\xi}_k), \ k = K, \ldots, 1, \hat{V}_{(K+1)i} = Y_i; \ k = K, K - 1, \ldots, 1, \hat{V}_{(K+1)i} = Y_i$. In this way, one might, starting at the last stage, obtain parameter estimates and form new “responses” at each stage as they work back towards the initial stage. While this approach is convenient for practical implementation of each method, one should recognize that the “responses” at stage $k = K - 1, \ldots, 1$ are functions of parameters that would have been estimated at stages $j = k + 1, \ldots, K$. That is, for $Q$-learning, say, $Q_k(S_k, A_k; \xi_k)$ is a model posited for $\hat{V}_{k+1}$, a function of estimates $\hat{\xi}_{k+1}$ (and, implicitly, also $\hat{\xi}_j, j = k + 2, \ldots, K$). Thus, inference requires simultaneously solving estimating equations for all $k = K, K - 1, \ldots, 1$ jointly. Section I of the Appendix demonstrates the joint estimation process where one can obtain standard errors for parameter estimates that respect this hierarchical constraint using the robust variance estimate. We provide a brief illustrative example in Chapter 3 of the consequences at stages $k = K - 1, \ldots, 1$ of taking a naive approach to estimated standard errors compared to using those outlined in Section I of the Appendix.

Recognizing that $\hat{V}_{ki}$ is a non-smooth function of the data in both $Q$-learning and $A$-learning as described above, Robins (2004) and others have noted that, under certain generative models, methods for estimating optimal dynamic treatment regimes may lead to non-regular asymptotics. Under these generative classes, namely when $\Pr\{C_k(S_k, A_{k-1}) = 0\} > 0$ for $k > 1$, this robust variance estimator in Section I of the Appendix no longer accurately describes the asymptotic behavior of our estimators. Laber et al. (2010) provides the limiting distribution of estimators in the general case case. See Chakraborty et al. (2010), Laber et al. (2010), and Song et al. (2010), among others, for a discussion and for select methods proposed to obtain valid inference under $Q$-learning. We do not consider these methods here or in the sequel, as choosing one or more methods to explore would be arbitrary; instead, our simulations in Chapter 3 do not consider the inferential problem outside the brief illustration.
1.4.1 Comparison and Practical Considerations

When \( K = 1 \), the \( Q \)-function is a model for \( E(Y|S_1 = s_1, A_1 = a_1) \). If in \( Q \)-learning this model and the variance model \( \Sigma_1 \) in (1.23) are correctly specified, then, as noted above, the form of (1.23) is optimal for estimating \( \xi_1 \). Accordingly, even if \( C_1(s_1; \psi_1) \) and \( h_1(s_1; \beta_1) \) are correctly modeled, (1.28) with \( K = 1 \) is generally not of this optimal form for any choice \( \lambda_1(s_1; \psi_1) \), and hence \( A \)-learning will yield relatively inefficient inference on \( \psi_1 \) and hence on the optimal regime. However, if in \( Q \)-learning the \( Q \)-function is mismodeled, but in \( A \)-learning \( C_1(s_1; \psi_1) \) and \( \pi_1(s_1; \phi_1) \) are both correctly specified, then \( A \)-learning will still yield consistent inference on \( \psi_1 \) and hence the optimal regime, whereas inference on \( \xi_1 \) and the optimal regime via \( Q \)-learning may be inconsistent. We assess the trade-off between consistency and efficiency in this case in Chapter 2. For \( K > 1 \), owing to the complications involved in specifying optimal estimating equations for \( Q \) and \( A \)-learning, the relative performance of the methods is not readily apparent; we investigate in Chapter 3.

In certain special cases, \( Q \) and \( A \)-learning lead to identical estimators for the \( Q \)-function (Chakraborty et al. 2010). For example, this holds if the propensities for treatment are constant, as would be the case under pure randomization at each decision point, and certain linear models are used for \( C_1(s_1; \psi_1) \) and \( h_1(s_1; \beta_1) \); see Section D of the Appendix for a demonstration when \( K = 1 \) and \( \text{pr}(A_1 = 1|S_1 = s_1) \) does not depend on \( s_1 \).

As we have emphasized, for \( Q \)-learning, while modeling the \( Q \)-function at decision \( K \) is a standard regression problem with response \( Y \), for decisions \( K - 1, \ldots, 1 \), this involves modeling the estimated value function, which depends on the relationships for subsequent decisions. Ideally, the sequence of posited models \( Q_k(\bar{s}_k, \bar{a}_k; \xi_k) \) should respect this constraint. However, this may be difficult to achieve with standard regression models. To illustrate, consider the models in (1.26), and assume \( S_1, S_2 \) are scalar, where the conditional distribution of \( S_2 \) given \( S_1 = s_1, A_1 = a_1 \) is Normal\( (\mathcal{K}_1^T \gamma, \sigma^2) \), say, \( \mathcal{K}_1 = (1, s_1, a_1)^T \). Recall that \( V_2(\bar{s}_2, a_1; \xi_2) = \mathcal{H}_2^T \beta_2 + (\mathcal{H}_2^T \psi_2)I(\mathcal{H}_2^T \psi_2 > 0) \), where we can write \( \mathcal{H}_2^T \beta_2 = \mathcal{K}_1^T \beta_{21} + s_2 \beta_{22} \) and \( \mathcal{H}_2^T \psi_2 = \mathcal{K}_1^T \psi_{21} + s_2 \psi_{22} \). Then, if the model \( Q_2 \) in (1.26) were correct, from (1.17), ideally, \( Q_1(s_1, a_1) = E[V_2(s_1, S_2, a_1; \xi_2)|S_1 = s_1, A_1 = a_1] \). Letting \( \varphi(\cdot) \) and \( \Phi(\cdot) \) be the standard normal density and cumulative distribution function,
respectively, it may be shown (see Section E of the Appendix) that, under these conditions,

\[
Q_1(s_1, a_1) = \mathbb{E}\{V_2(s_1, S_2, a_1; \xi_2) | S_1 = s_1, A_1 = a_1\} = \mathcal{K}_1^T (\beta_{21} + \gamma \beta_{22}) \\
+ (\mathcal{K}_1^T \psi_{21}) \{1 - \Phi(\eta)\} + \psi_{22} \{\sigma \varphi(\eta) + (\mathcal{K}_1^T \gamma) \{1 - \Phi(\eta)\}\}, \\
\]

where \(\eta = -\mathcal{K}_1^T (\psi_{21}/\psi_{22} + \gamma)/\sigma\), and we have taken \(\psi_{22} > 0\). Contrast the implied true \(Q_1(s_1, a_1)\) in (1.30) to the posited linear model in (1.26); clearly, the true relationship is highly nonlinear in \(s_1, a_1\) and is likely to be poorly approximated by \(Q_1(s_1, a_1; \xi_1)\) in (1.26). Evidently, for larger \(K\), this incompatibility between true and assumed models would propagate from \(K - 1, \ldots, 1\). Thus, while the use of linear models for the \(Q\)-functions is popular in practice, the potential for such mismodeling should be recognized.

An alternative approach that may mitigate the risk of mismodeling is to employ flexible models for the \(Q\)-functions. Zhao, Kosorok, and Zeng (2009) use support vector regression models in place of the linear models described above. Indeed, recent developments in statistical learning suggest a large collection of powerful regression methods that might be used. Many of these methods must be tuned in order to balance bias and variance, a natural approach to which is to minimize the cross-validated mean squared error of the \(Q\)-functions at each decision point. An obvious downside is that the final model may be difficult to interpret, and clinicians may be unwilling to implement “black box” rules. One compromise is to fit a simple, interpretable model, such as a decision tree, to the fitted values of the complex model in order to get a feel for what factors are driving the recommended treatment decisions. One can then check the simple model against scientific theory. If the simple, approximate model appears sensible, then clinicians may be willing to use predictions from the more complex and less interpretable model. For further discussion and references, see Craven and Shavlik (1996).

\(A\)-learning represents a middle ground between \(Q\)-learning and these approaches in that it allows for flexible modeling of the functions \(h_k(\bar{s}_k, \bar{a}_{k-1})\) while maintaining simple parametric models for the contrast functions \(C_k(\bar{s}_k, \bar{a}_{k-1})\). Thus, the resulting decision rule, which depends only on the contrast function, remains interpretable, while the model for the response is allowed to be nonlinear. This is also appealing in that it may be reasonable to expect, based on the underlying science, that the relationship between patient history and outcome is complex while the optimal rule for treatment assignment is dependent, in a simple fashion, on a small number of variables. The flexibility allowed by a semi-parametric model also has its drawbacks. Techniques for formal model building,
critique, and diagnosis are well understood for linear models but much less so for semi-parametric models. Consequently, $Q$-learning based on building a series of linear models may be more appealing to an analyst interested in formal diagnostics.

In the next chapter, we carry out a suite of simulations comparing $Q$- and $A$-learning for a single decision point. Chapter 3 continues with additional simulations for the multiple decision scenario. Finally, Chapter 4 will demonstrate these methods on a data set.
Simulations Under a Single Decision Point, $K = 1$

$Q$- and $A$-learning were introduced in Chapter 1. Specifically, in Section 1.3, each method was detailed and the estimation procedure outlined. In Section 1.4, we discussed some practical implications of each method. Now, we examine the finite sample performance of $Q$- and $A$-learning on a suite of test examples via Monte Carlo simulation. To illustrate the trade-offs between the methods discussed in the preceding chapter, we begin with correctly specified models and then systematically introduce misspecification of the $Q$-function, the propensity model, and both the $Q$-function and propensity model. In all cases, the contrast function is correctly specified, as, when this is not the case, the form of the optimal regime induced by an incorrect contrast function may not include $d^{\text{opt}}$, making interpretation difficult. In all scenarios, 10,000 Monte Carlo replications were used, and, for each generated data set, the estimated optimal regimes $\hat{d}^{\text{opt}}_{Q}$ and $\hat{d}^{\text{opt}}_{A}$ in (1.25) and (1.29) were obtained using the $Q$- and $A$-learning procedures described in Sections 1.3.1 and 1.3.2.

In this chapter, consider one ($K = 1$) decision problems, where, at the decision point, there are two feasible treatment options coded as 0 and 1. In all cases henceforth, we used $Q$-functions of the form $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1C_1(s_1; \psi_1)$ to represent both true and assumed working models. With the contrast functions correctly specified, the parameters $\psi_1$ dictate the optimal regime. Thus, as one measure of performance, we focus on relative efficiency of the estimators of components of $\psi_1$ obtained by $Q$-learning to those obtained by $A$-learning, as reflected by the ratio of their Monte Carlo mean squared
errors (MSEs) (so by MSE of $A$-learning/MSE of $Q$-learning), so that values greater than 1 favor $Q$-learning. Recognizing that $E\{Y^*(d_{opt})\}$ is the benchmark achievable outcome on average, as a second measure, we consider the extent to which the estimated regimes $\hat{d}_Q^{opt}$ and $\hat{d}_A^{opt}$ achieve $E\{Y^*(d_{opt})\}$ if followed by the population. Specifically, for regime $d$ indexed by $\psi_1$, let $H(d) = E\{Y^*(d)\}$, a function of these parameters. Then $H(d_{opt}) = E\{Y^*(d_{opt})\}$ is this function evaluated at the true parameter values, and $H(\hat{d}_{opt})$ is this function evaluated the estimated parameter values for a given data set, where $\hat{d}_{opt}$ represents $\hat{d}_Q^{opt}$ or $\hat{d}_A^{opt}$. Define $R(\hat{d}_{opt}) = E\{H(\hat{d}_{opt})\}/H(d_{opt})$, where the expectation in the numerator is with respect to the distribution of the estimated parameters in $\hat{d}_{opt}$, which may be interpreted as reflecting the efficiency with which $\hat{d}_{opt}$ achieves the performance of the true optimal regime. In Section F of the Appendix, we discuss calculation of $R(\hat{d}_{opt})$.

In this chapter, with $K = 1$, the observed data are $(S_i, A_i, Y_i)$, $i = 1, \ldots, n$. With $\expit(x) = e^x/(1 + e^x)$, to generate the data, we used

$$ S_1 \sim \text{Normal}(0, 1), \quad (2.1) $$

$$ A_1|S_1 = s_1 \sim \text{Bernoulli}\{\expit(\phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2)\}, \quad (2.2) $$

$$ Y|S_1 = s_1, A_1 = a_1 \sim \text{Normal}\{\beta_{10}^0 + \beta_{11}^0 s_1 + \beta_{12}^0 s_1^2 + a_1(\psi_{10}^0 + \psi_{11}^0 s_1), \sigma^2\}. \quad (2.3) $$

Then,

$$ \theta^0 = (\phi_{10}^0, \phi_{11}^0, \phi_{12}^0, \beta_{10}^0, \beta_{11}^0, \beta_{12}^0, \psi_{10}^0, \psi_{11}^0)^T \quad (2.4) $$

indexes our class of generative models. and $d_{opt} = d_1^{opt}$, $d_1^{opt}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0)$. For $A$-learning, we assumed working models $h_1(s_1; \beta_1) = \beta_{10} + \beta_{11} s_1$, $C_1(s_1; \psi_1) = \psi_{10} + \psi_{11} s_1$, and $\pi_1(s_1; \phi_1) = \expit(\phi_{10} + \phi_{11} s_1)$, and for $Q$-learning used $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1)$. Note that these working models involve correctly specified contrast functions and are nested within the true generative models, with $h_1(s_1; \beta_1)$, and hence the $Q$-function, correctly specified when $\beta_{12}^0 = 0$. Similarly, the propensity model $\pi_1(s_1; \phi_1)$ is correctly specified when $\phi_{12}^0 = 0$. To study the effects of misspecification, we systematically varied these two parameters while keeping the others fixed.

The remainder of this chapter is organized as follows. Section 2.1 will provide a scenario where the MSEs of $\hat{\psi}_1$ and $R(\hat{d}_{opt})$ can be compared under correctly specified models, misspecification of the $Q$-function, the propensity model, and both the $Q$-function and propensity models. In Section 2.2 we explore the effects of sample size on the preceding measures, specifically for the cases of misspecification of the $Q$-function.
and misspecification of both the $Q$-function and propensity models. In Section 2.3, we return to a fixed sample size and explore how changes to generative parameters affect the performance measures related to estimation of $\hat{\psi}_1$ and outcomes $R(\hat{d}_{opt})$.

### 2.1 Primary Simulations and Results

In this section, consider sample size $n = 200$ for each Monte Carlo iteration and parameter settings of the form $\theta^0 = (0, -2, \phi_{12}^0, 1, 1, \beta_{12}^0, 1, 0.5)^T$ and $\sigma^2 = 3$. We describe this scenario in the light of an observational study of patients being treated by a physician as follows, assuming $\phi_{12}^0 = 0$ and $\beta_{12}^0 = 0$. Subjects with $S_1 > 0$ were treated in a manner such that they were increasingly more likely to receive treatment 0; whereas patients with $S_1 < 0$ were more inclined to be treated with treatment 1. Subsequently the optimal decision rule suggests treatment 1 for $S_1 < -2$. In this hypothetical case, it would appear that the treating physicians were generally administering treatment to the right people, but many individuals would have received an inferior treatment given their covariates. Thus, attempting to identify the optimal treatment regime through estimation by $Q$- or $A$-learning might guide future clinical choices to yield better outcomes, on average. This is the premise of this section; however, both $Q$- and $A$-learning require models to be posited. Thus, we seek to identify the performance of each under various model misspecifications.

#### Correctly Specified Models

As noted in Section 1.4.1, when all working models are correctly specified, $Q$-learning is more efficient than $A$-learning. Under our class of generative models, this occurs when $\beta_{12}^0 = \phi_{12}^0 = 0$. In this scenario, the relative efficiency of $Q$-learning with respect to $Q$-learning is 1.06 for estimating $\psi_{10}^0$ and 2.74 for estimating $\psi_{11}^0$. Thus, $Q$-learning is a modest 6% more efficient in estimating $\psi_{10}^0$ but a dramatic 174% more efficient in estimating $\psi_{11}^0$. Interestingly, the efficiency of the decision rules produced by $Q$- and $A$-learning is similar, with $R(\hat{d}_{Q}^{opt}) = 0.97$ and $R(\hat{d}_{A}^{opt}) = 0.95$, so that the relative inefficiency in estimation of $\psi_1$ suffered by $A$-learning does not translate in to a regime of poorer quality than that found via $Q$-learning.
Misspecified Propensity Model

An appeal of $A$-learning is the double robustness property noted in Section 1.3.2, which implies that $\psi_1$ should be estimated consistently when the propensity model is misspecified provided that the $Q$-function is correct. Under our class of generative models, this corresponds to $\beta_{12} = 0$ and nonzero $\phi_{12}^0$. In contrast, $Q$-learning does not depend on the propensity model, so its performance is unaffected by this misspecification. Figure 2.1 shows the relative efficiency in estimating $\psi_{10}^0$ and $\psi_{11}^0$ and the efficiency of $\hat{d}_{Q}^{\text{opt}}$ and $\hat{d}_{A}^{\text{opt}}$ as $\phi_{12}^0$ varies from $-1$ to $1$. The leftmost panel shows that there is minimal gain in efficiency by using $Q$-learning instead of $A$-learning in estimation of $\psi_{10}^0$. From the center panel, $Q$-learning yields substantial gains over $A$-learning for estimating $\psi_{11}^0$. Interestingly, the gain in efficiency of $Q$- over $A$-learning is largest when $\phi_{12}^0 = 0$, which corresponds to the propensity model being correctly specified. Letting $\pi_0^0(s_1; \phi_1^0)$ be the true propensity, $\phi_1^0 = (\phi_{10}^0, \phi_{11}^0, \phi_{12}^0)^T$, a possible explanation for this seemingly contradictory result is that, as $|\phi_{12}^0|$ gets larger, $\logit\{\pi_0^0(S_1; \phi_1^0)\} = \phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2$ becomes more profoundly quadratic. Consequently, the estimator for $\phi_{11}$ in the posited model $\pi_1(s_1; \phi_1) = \expit(\phi_{10} + \phi_{11} s_1)$ approaches zero, so that the posited propensity approaches a constant. Because $Q$- and $A$-learning are equivalent under constant propensity, the efficiency gains decrease as $|\phi_{12}^0| \to \infty$. The right panel of Figure 2.1 shows a small gain in efficiency of $\hat{d}_{Q}^{\text{opt}}$ over $\hat{d}_{A}^{\text{opt}}$, with both achieving good performance.

Figure 2.1: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_{A}^{\text{opt}})$ for estimating the true $d^{\text{opt}}$ (right panel) under misspecification of the propensity model. MSE ratios > 1 favor $Q$-learning.
Misspecified Q-function

This scenario examines the second aspect of $A$-learning’s double-robustness and is characterized in our class of true generative models by $\phi_{12} = 0$ and nonzero $\beta_{12}$. Here, $A$-learning leads to consistent estimation while $Q$-learning need not. The left panel of Figure 2.2 shows that the gain in efficiency using $A$-learning is minimal in estimating $\psi_{10}$. The center panel illustrates the bias-variance trade-off associated with choice between $Q$- and $A$-learning. For values of $\beta_{12}$ that are far from zero, the bias in the misspecified $Q$-function dominates the variance, and $A$-learning enjoys smaller MSE while, for small values of $\beta_{12}$, variance dominates bias, and $Q$-learning is more efficient. The right panel shows that large bias in the $Q$-function can lead to meaningful loss (around 10%) in efficiency of $\hat{d}_{Q}^{opt}$ relative to $\hat{d}_{A}^{opt}$.

![Figure 2.2: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{Q}^{opt})$ and $R(\hat{d}_{A}^{opt})$ for estimating the true $d^{opt}$ (right panel) under misspecification of the $Q$-function. MSE ratios > 1 favor $Q$-learning.](image)

Both Propensity Model and $Q$-function Misspecified

In our class of generative models, this corresponds to nonzero values of both $\beta_{12}$ and $\phi_{12}$. Rather than vary both values, (e.g., over a grid), we varied one and chose the other so that it is “equivalently misspecified.” In particular, for a given value of $\phi_{12}$, we selected $\beta_{12} = \beta_{12}(\phi_{12})$ so that the $t$-statistic associated with testing $\phi_{12} = 0$ in the logistic propensity model and the $t$-statistic associated with testing $\beta_{12} = 0$ in the linear $Q$-function would be approximately equal in distribution. Consequently, across data sets,
an analyst would be equally likely to detect either form of misspecification. Details of this construction are given in Section G of the Appendix.

As in the preceding scenario, Figure 2.3 illustrates the bias-variance trade-off associated with $Q$- and $A$-learning. For large misspecification, $A$-learning provides a large enough reduction in bias to yield lower MSE; for small misspecification, $Q$-learning incurs some bias but reduces the variance enough to yield lower MSE. From the right panel of the figure, bias seems to translate into a larger loss in quality of the estimators of $d^{opt}$ than variance.

![Figure 2.3: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}^{opt}_Q)$ and $R(\hat{d}^{opt}_A)$ for estimating the true $d^{opt}$ (right panel) under misspecification of both the propensity model and the $Q$-function. MSE ratios > 1 favor $Q$-learning](image)

2.2 Simulation for the Effects of Sample Size

In the preceding section, we considered a sample size of $n = 200$, which might represent a small observational study or SMART. In many cases, larger sample sizes may be available, say, from observational datasets, registries, or trials involving a common disease or disorder. In this section, we explore the effects of sample size. Specifically, we look at how changes in sample size alter, if at all, the relationships between $Q$-learning and $A$-learning under misspecification of the $Q$-function and misspecification of both the $Q$-function and propensity model. We do not report results under misspecification of only
the propensity model. In this section, we consider sample sizes of \( n = 100 \), \( n = 500 \), and \( n = 5000 \), representing a small, medium, and large study, respectively.

**Effects of Misspecification of the Q-function for Various Sample Sizes**

We first consider the effects of sample size while looking at misspecification of the \( Q \)-function. Consider the data generating class defined in Equations 2.1–2.3 with parameters of the form \( \theta^0 = (0, 1, \phi^0_{12}, 2, 1, \beta^0_{12}, 1, -10)^T \) and \( \sigma^2 = 1 \). Misspecification of the \( Q \)-function requires \( \phi^0_{12} = 0 \) and \( \beta^0_{12} \neq 0 \), such that we expect \( A \)-learning to yield consistent estimates of parameters \( \psi_1 \) while \( Q \)-learning provides biased estimates under misspecification. One might suspect that for smaller sample sizes, variance of the estimates under \( A \)-learning may play more of a role in the comparison. The parameterization above presents a rather idealized case with a strong treatment interaction. That is, in truth, subjects with \( S_1 < 0.1 \) would see benefit from treatment 1 while the others would benefit from treatment 0; and, there are large changes in the response related to this relationship.

The MSEs are compared in Figure 2.4. For \( \hat{\psi}_{10} \), sample size has no effect on the relative efficiency of \( A \)-learning. In this particular parameterization, \( A \)-learning yields a substantial gain in efficiency in estimating \( \psi_{10} \) regardless of sample size. The results for \( \hat{\psi}_{11} \) are shown in the right panel of the same figure. Overall, \( A \)-learning provides substantial gains in efficiency for estimating this parameter under misspecification of the \( Q \)-function. We see that as sample size increases, the relationship indicates increasing favorability towards \( A \)-learning under misspecification.

Furthermore, we look to see how these results relate to the quality of the regime. In Figure 2.5, sample size increases as one moves right among the panels. Indeed, as sample size increases, we would expect both methods to become more idealized. Noting the overall scale of the plots, there is little overall difference between \( Q \)- and \( A \)-learning; but nonetheless, \( A \)-learning yields marginally better outcomes, on average. Figure 2.6 shows a similar result under the data generating parameters \( \theta^0 = (0, -1, \phi^0_{12}, 2, 1, \beta^0_{12}, 1, -10)^T \) and \( \sigma^2 = 1 \) such that the propensity model for treatment more closely follows the true optimal decision rule. That is, subjects with \( S_1 < 0 \) are more likely to receive treatment 1, which, as shown above is the better treatment for them under the optimal decision rule. This represents an observational study where the clinicians assigning treatment more often than not, give the correct treatment to patients. Again, considering the scale, there is little difference between the two options; however, in some cases, \( A \)-learning yields
Effects of Misspecification of Both the Propensity Model and the Q-function for Various sample Sizes

Next, consider the effects of sample size under misspecification of both the Q-function and propensity model. The class of data generating models are defined in Equations
2.1–2.3 and we consider parameters \( \theta^0 = (0, 1, \phi_{12}^0, 2, 1, \beta_{12}^0, 1, -10)^T \) and \( \sigma^2 = 1 \). When both the propensity model and \( Q \)-function are misspecified, \( \beta_{12}^0 \neq 0 \) and \( \phi_{12}^0 \neq 0 \). We use “equivalently misspecified” pairs as previously defined (See Section G).

We compare the MSEs of estimates of \( \psi_1 \) in Figure 2.7. As sample size increases, we notice that the variance of \( \hat{\psi}_{10} \) under \( A \)-learning decreases, and \( A \)-learning becomes increasingly more favorable with respect to this measure. For estimates \( \hat{\psi}_{11} \), there are indications that \( A \)-learning does well under small levels of misspecification, particularly in larger samples. However, for large misspecification, the increased variability and bias from \( A \)-learning yields worse estimates than simply bias under \( Q \)-learning.

Figure 2.8 relates \( R(d_{Q}^{\text{opt}}) \) and \( R(d_{A}^{\text{opt}}) \) under this misspecification scenario. As sample size increases, going left to right among the panels, there is no substantial change to the results except that both methods perform, as expected, closer to the ideal. In the process of creating the misspecified pair \( \{\phi_{12}^0, \beta_{12}^0(\phi_{12}^0)\} \) one could equivalently chose \( \{\phi_{12}^0, -\beta_{12}^0(\phi_{12}^0)\} \) to yield the same results in this scenario. We present the results under this parameterization in Figure 2.9, where all other parameters did not change. Again, with increasing sample size, both methods perform more efficiently. However, sample size does not affect the general relationship among misspecification in this scenario.
Figure 2.7: A comparison of $Q$- and $A$-learning via Monte Carlo MSE ratios for estimators of components of $\psi_1$ under misspecification of both the propensity model and the $Q$-function. MSE ratios $> 1$ favor $Q$-learning. Sample sizes of $n = 100, n = 500,$ and $n = 5000$ are shown.

Figure 2.8: A comparison of $Q$- and $A$-learning via efficiencies $R(\hat{a}_Q^{opt})$ and $R(\hat{a}_A^{opt})$ for estimating the true $a^{opt}$ under misspecification of both the propensity model and the $Q$-function. Sample sizes of $n = 100, n = 500,$ and $n = 5000$ are shown.
Figure 2.9: A comparison of $Q$- and $A$-learning via efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ under misspecification of both the propensity model and the $Q$-function. Sample sizes of $n = 100$, $n = 500$, and $n = 5000$ are shown.

2.3 Simulations for Alternate Parameterizations

Again consider the class of models in Equations 2.1–2.3 with sample size fixed at $n = 1000$. The previous sections have focused on just a few parameterizations of $\theta^0$. Now, we explore the consequences of changing each parameter of the data generating process under the various misspecification scenarios. That is, for $\theta^0 = (\phi^0_{10}, \phi^0_{11}, \phi^0_{12}, \beta^0_{10}, \beta^0_{11}, \beta^0_{12}, \psi^0_{10}, \psi^0_{11})^T$, we seek to determine how changing the parameters $\phi^0_{10}, \phi^0_{11}, \phi^0_{12}, \beta^0_{10}, \beta^0_{11}, \psi^0_{10}, \psi^0_{11}$ might alter results in the context of model misspecification for the propensity, $Q$-function, and both.

2.3.1 Results for Misspecified Propensity Model

As a baseline parameterization to make comparisons to, consider $\sigma^2 = 1$ and $\theta^0 = (\phi^0_{10} = 0, \phi^0_{11} = 0, \phi^0_{12} = 0, \beta^0_{10} = 0, \beta^0_{11} = 1, \beta^0_{12} = 1, \psi^0_{10} = 1, \psi^0_{11} = -10)^T$, where for the misspecified propensity scenario, $\beta^0_{12} = 0$. The results, in Figure 2.10, show results for the efficiency of estimating $\psi_1$ similar to the results of Section 2.1. Under this scenario, again an idealized situation, both $Q$- and $A$-learning are essentially the same with respect to expected potential outcomes $R(\hat{d}^{opt})$. Both perform nearly ideal.

Changes to Parameters in the Propensity Model

If we alter the data generating parameters involved in the propensity, such as $\phi^0_{11} = 0.5$ (previously $\phi^0_{11} = 1$), we want to know how the results change under misspecification of the propensity model. Figure 2.11 yields the results, such that the variance under
Figure 2.10: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ (right panel) under misspecification of the propensity model. MSE ratios $>1$ favor $Q$-learning.

$A$-learning is diminished slightly for estimation of $\psi_{11}$, though $Q$-learning might still be preferred.

Figure 2.11: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ (right panel) under misspecification of the propensity model. MSE ratios $>1$ favor $Q$-learning.

**Changes to Parameters in the Q-function**

The $Q$-function is characterized by two parts, the first $h_1(S_1; \beta_1^0) = \beta_{10}^0 + \beta_{11}^0 S_1 + \beta_{12}^0 S_1^2$ and the second $C_1(S_1; \psi_1^0) = \psi_{10}^0 + \psi_{11}^0 S_1$. In Section F of the Appendix, we show how one would calculate $R(\hat{d}^{opt})$. Thus, for this parameterization, it is easy to show that changing
the parameter $\beta_{10}^0$ would shift the plots of expected outcomes up or down while changes to $\beta_{11}^0$ have no impact at all with respect to efficiency in relation to the ideal expected outcome.

Changes to parameters involved in the contrast, however, may have an impact. Indeed, Figure 2.12 yields the ratio of MSEs and outcome comparisons when $\psi_{11}^0 = 1$ (previously, $\psi_{11}^0 = -10$). While there is no change from the baseline parameterization when comparing the efficiency of estimating $\psi_1$, we see that there is more separation between $Q$- and $A$-learning with respect to $R(\hat{d}_{opt})$; however, this difference, considering the scale, is negligible with $Q$- and $A$-learning both performing nearly ideal to $H(d_{opt})$.

![Figure 2.12: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{Q_{opt}})$ and $R(\hat{d}_{A_{opt}})$ for estimating the true $d_{opt}$ (right panel) under misspecification of the propensity model. MSE ratios $>1$ favor $Q$-learning.](image)

**Changes to $\sigma^2$**

Finally, in the course of looking at misspecification of the propensity, we consider different variance in the response. Figure 2.13 uses the baseline parameterization but with $\sigma^2 = 100$. Even with this substantially larger variance of the response, results are rather unchanged, with both methods performing close to the ideal $H(d_{opt})$.

### 2.3.2 Results for Misspecified Q-function

In the following, we consider the misspecified $Q$-function, such that $A$-learning yields an asymptotically unbiased parameter estimates while the same is not true for $Q$-learning.
Figure 2.13: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true $d_{\text{opt}}$ (right panel) under misspecification of the propensity model. MSE ratios > 1 favor $Q$-learning.

Suppose $\sigma^2 = 1$ and $\theta^0 = (\phi_{10}^0 = 0, \phi_{11}^0 - 1, \phi_{12}^0, \beta_{10}^0 = 0, \beta_{11}^0 = 1, \beta_{12}^0, \psi_{10}^0 = 1, \psi_{11}^0 = -10)^T$ where $\phi_{12}^0 = 0$, are a baseline parameterization. Figure 2.14 shows a comparison of the efficiency with which they estimate parameters $\psi_{10}$ and $\psi_{11}$ as well as their efficiency in achieving $H(d_{\text{opt}})$.

Figure 2.14: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true $d_{\text{opt}}$ (right panel) under misspecification of the $Q$-function. MSE ratios > 1 favor $Q$-learning.
Changes to Parameters in the Propensity Model

Figures 2.15 and 2.16 show how results might change under different parameter settings for $\phi_{10}$ and $\phi_{11}$, respectively. In the former, consider $\phi_{10}^0 = 1$ so that individuals are all, on average, more likely to receive treatment 1. We observe more drastic gains in efficiency under $A$-learning in this scenario relative to the baseline for both estimation of $\psi_{10}$ and outcomes $R(\hat{d}_{opt})$. In the latter (Figure 2.16), we use $\phi_{11}^0 = 0.5$ and the results for $\psi_{10}$ and $R(\hat{d}_{opt})$ are more muted than the baseline parameterization. Indeed, for various alterations to $\phi_{1}^0$, $A$-learning estimates continued to outperform $Q$-learning estimates, but the degree of difference observed was altered by the parameters in the propensity model.

![Figure 2.15: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{opt}^Q)$ and $R(\hat{d}_{opt}^A)$ for estimating the true $d_{opt}$ (right panel) under misspecification of the $Q$-function, with changes to the parameters in the propensity model. MSE ratios > 1 favor $Q$-learning.]

Changes to Parameters in the Q-function

Like Section 2.3.1, we can consider changes to parameters of the $Q$-function. Similar results hold for changes to $\beta_{1}^0$ in that $\beta_{10}^0$ and $\beta_{11}^0$ do not change the results in a substantial way. Furthermore, changes to parameters $\psi_{1}^0$ do not substantially alter the results. With $\psi_{10}^0$ or $\psi_{11}^0$ changed from the baseline model, efficiency of estimation of $\psi_{10}$ and $\psi_{11}$ remains; i.e. there is no change to the two left-most plots. $R(\hat{d}_{opt})$ may be shifted up or down for changes to these parameters.
Changes to $\sigma^2$

Finally, consider smaller or larger variance of the response. In the former, we consider the small variance $\sigma^2 = 10^{-4}$; in the latter, we use $\sigma^2 = 100$. Results of each are in Figures 2.17 and 2.18 for smaller and larger variance, respectively. In the first, we observe perhaps nearly asymptotic results for the ratio of MSEs in the two left-most panels. Variance under $A$-learning (and $Q$-learning) is small. Thus, bias under $Q$-learning quickly outweighs that variance. For $\hat{\psi}_{10}$, the mean square error is approximately half under $A$-learning compared to $Q$-learning. This might suggest a limit to the bias induced under $Q$-learning in these cases.

Under substantially large variance, the two leftmost panels of Figure 2.18 reflect the additional variability of $A$-learning. Interestingly, this change in variance substantially alters the rightmost plot of $R(\hat{d}_{Q}^{\text{opt}})$. In fact, we see that for some levels of misspecification, $A$-learning estimates parameters $\psi_{10}$ and $\psi_{11}$ more efficiently than $Q$-learning, perhaps even substantially so for $\psi_{11}$; however $Q$-learning yields better expected outcomes, on average, with $R(\hat{d}_{Q}^{\text{opt}}) > R(\hat{d}_{A}^{\text{opt}})$. This is certainly not intuitive, but perhaps occurs because, while $Q$-learning estimates may be biased, the combination of $\hat{\psi}_{10}$ and $\hat{\psi}_{11}$ that dictates the value of $S_1$ where optimal treatment switches, is more accurately obtained compared to the more variable $A$-learning estimates.
Figure 2.17: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ (right panel) under misspecification of the $Q$-function, with changes to the variance of the response. MSE ratios > 1 favor $Q$-learning.

Figure 2.18: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ (right panel) under misspecification of the $Q$-function, with changes to the variance of the response. MSE ratios > 1 favor $Q$-learning.
2.3.3 Results for Misspecification of Both the Propensity Model and Q-function

The final comparisons we make in this section are under misspecification of both the propensity model and Q-function using misspecified pairs as noted earlier. We use $\sigma^2 = 1$ and $\theta^0 = (\phi^0_{10} = 0, \phi^0_{11} - 1, \phi^0_{12}, \beta^0_{10} = 0, \beta^0_{11} = 1, \beta^0_{12}, \psi^0_{10} = 1, \psi^0_{11} = -10)^T$ as a baseline parameterization. The results are shown in Figure 2.19 and indicate A-learning yielding better estimation of $\psi_{10}$ and better efficiency with respect to $R(\hat{d}_{opt}^Q)$ than Q-learning.

Figure 2.19: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{opt}^Q)$ and $R(\hat{d}_{opt}^A)$ for estimating the true $d_{opt}$ (right panel) under misspecification of both the propensity model and the Q-function. MSE ratios > 1 favor Q-learning

Changes to Parameters in the Propensity Model

In this exploration of how parameters affect the results under misspecification of both the propensity and Q-function, we first look at changes to the parameters in the propensity model. First, consider a change from the baseline parameterization so that $\phi^0_{12} > 0$ so that there is no difference between A- and Q-learning in either the left or right plots. For estimation of $\psi_{11}$, the middle panel shows that A-learning may yield a more efficient estimator for some levels of misspecification, but Q-learning performs better overall.

We also consider changing $\phi^0_{11} = 0.5$ which yields results similar to the baseline, but with the magnitude of differences between A- and Q-learning diminished (see Figure 2.21).
Figure 2.20: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{Q}^{opt})$ and $R(\hat{d}_{A}^{opt})$ for estimating the true $d^{opt}$ (right panel) under misspecification of both the propensity model and the $Q$-function. MSE ratios $> 1$ favor $Q$-learning.

In each of the two alternative parameterizations considered here, we note that changing parameters of the propensity alters the relationship between $\phi_{12}^0$ and $\beta_{12}^0$ though this does not appear to be the driver behind the results observed.

Figure 2.21: Monte Carlo MSE ratios for estimators of components of $\psi_1$ (left and center panels) and efficiencies $R(\hat{d}_{Q}^{opt})$ and $R(\hat{d}_{A}^{opt})$ for estimating the true $d^{opt}$ (right panel) under misspecification of both the propensity model and the $Q$-function. MSE ratios $> 1$ favor $Q$-learning.
Changes to Parameters in the Q-function

Just as in Sections 2.3.1 and 2.3.2, changes to \( \beta_1^0 \) yield little substantial change in the observed results. The same is not true for changes to parameters of the contrast function. In particular, consider \( \psi_{11}^0 = 1 \) so that treatment 1 is preferred for individuals with \( S_1 > -1 \). In Figure 2.22, we see identical results for the comparison of estimating parameters \( \psi_{10} \) and \( \psi_{11} \). However, results for \( R(\hat{d}_{opt}^Q) \) do not indicate a clearly preferred method.

Figure 2.22: Monte Carlo MSE ratios for estimators of components of \( \psi_1 \) (left and center panels) and efficiencies \( R(\hat{d}_{opt}^Q) \) and \( R(\hat{d}_{opt}^A) \) for estimating the true \( d_{opt} \) (right panel) under misspecification of both the propensity model and the Q-function. MSE ratios > 1 favor Q-learning

Changes to \( \sigma^2 \)

Finally, we look at the scenario where \( \sigma^2 = 100 \), with results shown in Figure 2.23. As we’ve seen in prior results, for some levels of misspecification of both the propensity and Q-function, Q-learning may yield larger \( R(\hat{d}_{opt}^Q) \) despite A-learning yielding more efficient parameter estimates.
2.4 Discussion

In this chapter, we have provided a comprehensive comparison of Q-learning and A-learning in the $K = 1$ decision case. Section 2.1 provides the reader with their first comparisons of the two methods under model misspecification. In the case discussed, Q-learning proved more efficient with respect to both measures when the propensity was misspecified. Results in Section 2.3 agree with this result for many other generating parameterizations. Section 2.1 also described the relationship between the two methods under misspecification of the Q-function, such that for significant misspecification, Q-learning results in biased parameter estimates and worse efficiency with respect to MSEs of parameter estimates and achieving the benchmark $H(d_{opt})$. Sections 2.2 and 2.3 agreed with this result in general. However, select parameterizations, especially those with larger variance of the response, indicate scenarios where Q-learning may yield better efficiency in $R(\hat{d}_{opt})$ despite worse performance in estimation of the key parameters. Finally, the results of Section 2.1, when both the propensity and Q-function are misspecified, indicate no clearly preferred method. Indeed, Sections 2.2 and 2.3 confirm that while A-learning might be preferred for some parameterizations with this misspecification, it is certainly not always the case.
Simulations Under Two or More Treatment Decisions

In the prior chapter, we observed the results of a single decision point study and systematically explored the effects of sample size, each parameter, and misspecification. Now, consider the two decision case, where results can be considered to extend generally to the multiple decision scenario. Again, we examine the finite sample performance of Q- and A-learning on a suite of test examples via Monte Carlo simulation. The intention of this chapter is to illustrate trade-offs discussed in Section 1.4.1 for the more complicated, but applicable, scenario of two decision points. In the two stage \((K = 2)\) framework, we consider two generative classes in which to compare Q- and A-learning. The first, in Section 3.1, mirrors the format of Chapter 2. That is, we look at correctly specified models and will again systematically introduce misspecification of the Q-function, the propensity model, and both the Q-function and propensity model. In the second class, reported in Section 3.2, we consider a scenario related to recent paper by Moodie and Richardson (Moodie et al. 2007). Just as in Chapter 2, throughout this chapter the contrast function is correctly specified so that the form of the optimal regime estimated under Q- or A-learning includes the true \(d^{\text{opt}}\). In all simulations, 10,000 Monte Carlo replications were used, and, for each generated data set, the estimated optimal regimes \(\hat{d}^{\text{opt}}_Q\) and \(\hat{d}^{\text{opt}}_A\) in Equations 1.25 and 1.29 were obtained using the Q- and A-learning procedures described in Sections 1.3.1 and 1.3.2.
3.1 $K = 2$ Decisions

Consider the two stage ($K = 2$) decision problem, where, at each decision point, there are two feasible treatment options coded as 0 and 1. In all cases within this subsection, we used $Q$-functions of the form $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1)$ and $Q_2(\bar{s}_2, \bar{a}_2; \xi_2) = h_2(\bar{s}_2; a_1; \beta_2) + a_2 C_2(\bar{s}_2, a_1; \psi_2)$ to represent both true and assumed working models. With the contrast functions correctly specified, the parameters $\psi_k$, $k = 1, 2$, dictate the optimal regime. Thus, as one measure of performance, we focus on relative efficiency of the estimators of components of $\psi_k$ obtained by $Q$-learning to those obtained by $A$-learning, as reflected by the ratio of their Monte Carlo mean squared errors (MSEs) (so by MSE of $A$-learning/MSE of $Q$-learning), so that values greater than 1 favor $Q$-learning. Recognizing that $E\{Y^*(d_{opt})\}$ is the benchmark achievable outcome on average, as a second measure, we consider the extent to which the estimated regimes $\hat{d}_{opt}^Q$ and $\hat{d}_{opt}^A$ achieve $E\{Y^*(d_{opt})\}$ if followed by the population. Specifically, for regime $d$ indexed by $\psi_1$ and $\psi_2$, let $H(d) = E\{Y^*(d)\}$, a function of these parameters. Then $H(d_{opt}) = E\{Y^*(d_{opt})\}$ is this function evaluated at the true parameter values, and $H(\hat{d}_{opt})$ is this function evaluated the estimated parameter values for a given data set, where $\hat{d}_{opt}$ represents $\hat{d}_{opt}^Q$ or $\hat{d}_{opt}^A$. Define $R(\hat{d}_{opt}) = E\{H(\hat{d}_{opt})\}/H(d_{opt})$, where the expectation in the numerator is with respect to the distribution of the estimated parameters in $\hat{d}_{opt}$, which may be interpreted as reflecting the efficiency with which $\hat{d}_{opt}$ achieves the performance of the true optimal regime. In Section F of the Appendix, we discuss calculation of $R(\hat{d}_{opt})$.

For $K = 2$ decision points, the observed data available for estimation of $d_{opt} = (d_{opt}^1, d_{opt}^2)$ are $(S_{1i}, A_{1i}, S_{2i}, A_{2i}, Y_i)$, $i = 1, \ldots, n$. For these scenarios, we used a class of true generative data models that differs from those of Chakraborty et al. (2010), Song et al. (2010), and Laber et al. (2010) only in that $S_2$ is continuous instead of binary.
generative model is

\[ S_1 \sim \text{Bernoulli}(0.5), \]  
\[ A_1 | S_1 = s_1 \sim \text{Bernoulli}\{\expit(\phi_{10} + \phi_{11}s_1)\}, \]  
\[ S_2 | S_1 = s_1, A_1 = a_1 \sim \text{Normal}(\delta_{10} + \delta_{11}s_1 + \delta_{12}a_1 + \delta_{13}s_1a_1, \sigma_{S_2}^2), \]  
\[ A_2 | S_1 = s_1, S_2 = s_2, A_1 = a_1 \sim \text{Bernoulli}\{\expit(\phi_{20} + \phi_{21}s_1 + \phi_{22}a_1 + \phi_{23}s_2 + \phi_{24}a_1s_2 + \phi_{25}s_2^2)\}, \]

\[ Y | S_1 = s_1, S_2 = s_2, A_1 = a_1, A_2 = a_2 \sim \text{Normal}\{m(s_1, s_2, a_1, a_2), \sigma^2\}, \]

\[ m(s_1, s_2, a_1, a_2) = \beta_{20} + \beta_{21}s_1 + \beta_{22}a_1 + \beta_{23}s_1a_1 + \beta_{24}s_2 + \beta_{25}s_2^2 + a_2(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2). \]

The model is indexed by parameters \( \phi_1^0 = (\phi_{10}, \phi_{11})^T, \delta_1^0 = (\delta_{10}, \delta_{11}, \delta_{12}, \delta_{13})^T, \phi_2^0 = (\phi_{20}, \phi_{21}, \phi_{22}, \phi_{23}, \phi_{24}, \phi_{25})^T, \beta_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25})^T \), and \( \psi_2^0 = (\psi_{20}, \psi_{21}, \psi_{22})^T \), with true \( h_2^0(s_1, s_2, a_1) = \beta_{20} + \beta_{21}s_1 + \beta_{22}a_1 + \beta_{23}s_1a_1 + \beta_{24}s_2 + \beta_{25}s_2^2 \) and contrast function \( C_2^0(s_1, s_2, a_1) = \psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 \), say. Because \( A_1 \) and \( S_1 \) are binary, the true functions \( h_1^0(s_1) = \beta_{10} + \beta_{11}s_1 \) and \( C_1^0(s_1) = \psi_{10} + \psi_{11}s_1 \), are linear in \( s_1; \beta_{10}, \beta_{11}, \psi_{10}, \psi_{11} \), and \( \psi_{11} \) are derived in terms of parameters indexing the generative model in Section H of the Appendix. Thus, the true optimal regime has \( d_{1\text{opt}}^0(s_1) = I(\psi_{10} + \psi_{11}s_1 > 0) \) and \( d_{2\text{opt}}^0(s_1, s_2, a_1) = I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0) \).

We assumed working models for \( A \)-learning of the form \( h_1(s_1; \beta_1) = \beta_{10} + \beta_{11}s_1, \) \( C_1(s_1; \psi_1) = \psi_{10} + \psi_{11}s_1, \) \( \pi_1(s_1; \phi_1) = \expit(\phi_{10} + \phi_{11}s_1), \) \( h_2(s_1, s_2, a_1; \beta_2) = \beta_{20} + \beta_{21}s_1 + \beta_{22}a_1 + \beta_{23}s_1a_1 + \beta_{24}s_2, \) \( C_2(s_1, s_2, a_1; \psi_2) = \psi_{20} + \psi_{21}a_1 + \psi_{22}s_2, \) and \( \pi_2(s_1, s_2, a_1; \phi_2) = \expit(\phi_{20} + \phi_{21}s_1 + \phi_{22}a_1 + \phi_{23}s_2 + \phi_{24}a_1s_2); \) and, similarly, assumed \( Q \)-functions of the form \( Q_1(s_1, s_2; \xi_1) = h_1(s_1; \beta_1) + a_1C_1(s_1; \psi_1) \) and \( Q_2(s_1, s_2, a_1; \xi_2) = h_2(s_1, s_2, a_1; \beta_2) + a_2C_2(s_1, s_2, a_1; \psi_2) \) for \( Q \)-learning, so that the contrast functions are correctly specified in each case. Comparison of the working and generative models shows that the former are correctly specified when \( \phi_{25}^0 \) and \( \beta_{25}^0 \) are both zero and are misspecified otherwise. Thus, we systematically varied these parameters to study the effects of misspecification, leaving all other parameter values fixed for a given simulation.

In the following, we provide a scenario where the MSEs of \( \hat{\psi}_1 \) and \( R(\hat{d}_{\text{opt}}) \) can be compared under correctly specified models, misspecification of the \( Q \)-function, the propensity model, and both the \( Q \)-function and propensity models. We further explore the effects of sample size on the preceding measures, specifically for the cases of misspecification.
of the $Q$-function and misspecification of both the $Q$-function and propensity models. Finally, we return to a fixed sample size and consider how each parameter affects results related to $\hat{\psi}_1$ and $R(\hat{d}^{opt})$.

### 3.1.1 Primary Simulations and Results

In this section, we consider $n = 200$ to represent a dataset of small to moderate size. We take $\phi_0 = (0.3, -0.5)^T$, $\delta_0 = (0, 0.5, -0.75, 0.25)^T$, $\sigma^2 = 2$, $\phi_2 = (0, 0.5, 0.1, -1, -0.1, \phi_{25})^T$, $\beta_2 = (3, 0, 0.1, -0.5, -0.5, \beta_{25})^T$, $\psi_2 = (1, 0.25, 0.5)^T$, and $\sigma^2 = 10$.

#### Correctly Specified Models

Given our working models, this occurs when $\phi_{25} = \beta_{25} = 0$ in the generative models. As discussed previously, $Q$-learning is efficient when the models are correctly specified. Relative efficiencies of $Q$-learning with respect to $A$-learning for estimating $\psi_{10}$, $\psi_{11}$, $\psi_{20}$, $\psi_{21}$, and $\psi_{22}$ are 1.07, 1.03, 1.19, 1.44, and 1.98, respectively. Hence, $Q$-learning is markedly more efficient in estimating the second stage parameters but only modestly so in estimating first stage parameters. More efficient estimators of the underlying parameters do not translate into significantly more efficient estimated regimes, as $R(\hat{d}^{opt}) = 0.96$ and $R(\hat{d}^{opt}) = 0.96$.

In Section 1.4, we noted that, should one intend to draw inference, computation of standard errors of parameter estimates should be done in a way that respects the hierarchical, in time, nature of the estimating equations. Section I of the Appendix describes how one would calculate robust standard errors, if we assume our generating class is not among those leading to non-regular asymptotics. While the generating class above can indeed lead to non-regular asymptotics, we continue under two premises. First, that the following is for illustrative purposes. Second, we recognize that, for this scenario, Cohen’s effect size (Cohen 1988) can be estimated to be 1.47, a substantially large effect; thus we continue with this illustration without concern for non-regularity. In the general case, this inferential problem applies to stages $k = K - 1, ..., 1$. Here, we provide a brief report of estimated standard errors under a naive approach compared to the standard error estimates respecting the hierarchy. Here, consider estimation of $\psi_{10}$ and $\psi_{11}$ in the first stage. Under $Q$-learning, $E[\hat{\psi}_{10,Q}] = 0.39$ with average, across simulations, standard error (aSE) 0.73 using the method outlined in the appendix. A naive approach, that doesn’t account for hierarchy, would estimate aSE= 0.08. Similarly, for $\hat{\psi}_{11}$, $Q$-learning has an
average estimate of $-0.53$ with aSE $0.92$ vs $0.11$ under the naive approach. $A$-learning does not have such a dramatic difference in this simulation. Instead, $E[\hat{\psi}_{10,A}] = 0.41$ with aSE $0.76$ compared to $0.65$ under the naive method; and, $E[\hat{\psi}_{11,A}] = -0.54$ with aSE $0.94$ compared to $0.92$ in the naive case. Underestimation of the standard error of parameter estimates can result in under-coverage of nominal confidence intervals. In the sequel, we do not consider the inferential problem, in part due to complications noted in Section 1.4.

**Misspecified Propensity Model**

The propensity model at the second stage is misspecified when $\phi_{25}^0$ is nonzero. To isolate the effects of such misspecification, we set $\beta_{25}^0 = 0$ and varied $\phi_{25}^0$ between $-1$ and $1$. From Figure 3.1, $Q$-learning is more efficient than $A$-learning for estimation of all parameters in $\psi_1$ and $\psi_2$, and, as in the one decision case, the efficiency gain is largest when the $\phi_{25}^0 = 0$, corresponding to a correctly specified propensity model. From the lower right panel, there appears to be little difference in efficiency of $\hat{d}_{Q_{\text{opt}}}$ and $\hat{d}_{A_{\text{opt}}}$.

**Misspecified $Q$-function**

Under our class of generative models, the $Q$-function is misspecified when $\beta_{25}^0$ is nonzero. We set $\phi_{25}^0 = 0$ to focus on the effects of such misspecification. Figure 3.2 shows that, for the first stage parameters $\psi_{10}^0$ and $\psi_{11}^0$, there is little difference in efficiency between $Q$- and $A$-learning. The upper panels illustrate varying degrees of the bias-variance trade-off between the methods. In particular, in estimating $\psi_{22}^0$, a small amount of misspecification leads to significant bias, and hence $A$-learning produces a much more accurate estimator, while, for $\psi_{20}^0$, the bias-variance trade-off is present but attenuated, and there is little difference between $Q$- and $A$-learning. In estimation of $\psi_{21}^0$, variance appears to dominate bias, and $Q$-learning is preferred for the chosen range of $\beta_{25}^0$ values. From the lower right panel, relative efficiency for estimating $\psi_{22}^0$ weakly tracks the relative efficiencies of the estimated regimes $\hat{d}_{Q_{\text{opt}}}$ and $\hat{d}_{A_{\text{opt}}}$, suggesting that the efficiency gain for $A$-learning in estimating $\psi_{22}^0$ leads to improved estimation of $d_{\text{opt}}$.

**Misspecification of Both the Propensity Model and $Q$-function**

Under our generative model, this scenario corresponds to nonzero values of $\beta_{25}^0$ and $\phi_{25}^0$. Analogous to the one decision case, we chose pairs $(\beta_{25}^0, \phi_{25}^0)$ that are “equivalently
Figure 3.1: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ (lower right panel) under misspecification of the propensity model. MSE ratios $> 1$ favor Q-learning.
Figure 3.2: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ (lower right panel) under misspecification of the $Q$-functions. MSE ratios $>1$ favor $Q$-learning.
misspecified;” see Section G of the Appendix. Figure 3.3 shows the relative efficiency of the two methods. There is no general trend in efficiency of estimation across parameters that might recommend one method over the other. Furthermore, from the lower right panel, there is little difference in efficiency of the estimated regimes. This is as expected, as one should not expect to draw broad conclusions, as neither $Q$- nor $A$-learning need be consistent here. Interestingly, despite misspecification of both models, $\hat{d}_{Q}^{\text{opt}}$ and $\hat{d}_{A}^{\text{opt}}$ still enjoy high efficiency.

![Figure 3.3: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_{A}^{\text{opt}})$ for estimating the true $d^{\text{opt}}$ (lower right panel) under misspecification of both the propensity models and $Q$-functions. MSE ratios $> 1$ favor $Q$-learning.](image-url)
3.1.2 Simulations for the Effects of Sample Size

In the preceding, we considered a sample size of \( n = 200 \) for our simulations, which might represent a small observational study or SMART. In many cases, datasets of larger sample sizes may be available, say, from observational datasets, registries, or trials involving a common disease or disorder. In this section, we explore the effects of different sample sizes. Specifically, we look at how changes in sample size might alter the relationships between \( Q \)-learning and \( A \)-learning under misspecification of the \( Q \)-function and misspecification of both the \( Q \)-function and propensity model. We do not report results under misspecification of only the propensity model. We consider sample sizes of \( n = 100 \), \( n = 500 \), \( n = 1000 \), \( n = 5000 \), and \( n = 10000 \), representing a small, moderate, and several increasingly large datasets.

Effects of Misspecification of the \( Q \)-function for Various Sample Sizes

Consider the data generating class defined in Equations 3.1–3.6 with parameters of the form

\[
\phi_0^0 = (0.3, -0.5)^T, \quad \delta_1^0 = (0, 0.5, -0.75, 0)^T, \quad \sigma_{S_2}^2 = 1, \quad \phi_2^0 = (0, 0.25, 0.3, -0.3, -0.4, \phi_{25}^0)^T, \\
\beta_2^0 = (10, 1, -3, -2, -2.5, \beta_{25}^0)^T, \quad \psi_2^0 = (1.5, 1, -3.5)^T, \quad \sigma^2 = 1.
\]

We first consider the effects of sample size while looking at misspecification of the \( Q \)-function. This requires \( \phi_{25}^0 = 0 \) and \( \beta_{25}^0 \neq 0 \), such that we expect \( A \)-learning to yield consistent estimates of parameters \( \psi_2 \) while \( Q \)-learning provides biased estimates under misspecification. In the prior study (Section 3.1.1), the effects observed at the second stage did not propagate back to the first stage, but that will again be tested. We would expect smaller sample sizes to result in variance of the estimates under \( A \)-learning playing more of a role in the comparison to \( Q \)-learning.

The MSEs of parameters involved in treatment decisions are compared in Figure 3.4. For estimation of \( \psi_{20} \) and \( \psi_{21} \), increasing sample size is associated with favorability towards \( A \)-learning under misspecification of the \( Q \)-function. The same is also true of \( \psi_{22} \) with an even more sharp decline in the curves such that for large sample sizes it takes only a small misspecification for the bias associated with \( Q \)-learning to outweigh any increased variability under \( A \)-learning. Similar to the results observed in the prior section, the differences in estimation of \( \psi_{10} \) and \( \psi_{11} \) are not as substantial; however, there are indications that \( A \)-learning yields better estimates under misspecification with respect to this measure.

We also look to see how these results relate to the quality of the regime through
Figure 3.4: A comparison of $Q$- and $A$-learning via Monte Carlo MSE ratios for estimators of components of $\psi_1$ and $\psi_2$ under misspecification of the $Q$-function. MSE ratios $> 1$ favor $Q$-learning. Sample sizes of $n = 200$, $n = 500$, $n = 1000$, $n = 5000$, and $n = 10000$ are shown.
$R(\hat{d}^{opt})$. In Figure 3.5, sample size increases as one moves right among the top panels, then continues to the second row. We would both methods to become more idealized as sample size increases, which is observed here. $A$-learning yields marginally better outcomes, on average, for simulations of $n = 5000$ and $n = 10000$. Under smaller sample sizes we see that increased efficiency in parameter estimation noted in Figure 3.4 does not necessarily yield efficiency for average expected outcomes.

![Figure 3.5](image)

**Figure 3.5**: A comparison of $Q$- and $A$-learning via efficiencies $R(\hat{d}^{opt}_Q)$ and $R(\hat{d}^{opt}_A)$ for estimating the true $d^{opt}$ under misspecification of the $Q$-function. Sample sizes of $n = 200$, $n = 500$, $n = 1000$, $n = 5000$, and $n = 10000$ are shown.

**Effects of Misspecification of Both the Propensity Model and the Q-function for Various sample Sizes**

Next, consider the effects of sample size under misspecification of both the $Q$-function and propensity model. The class of data generating models are defined in Equations 3.1–3.6 where we use parameters of the form $\phi^0 = (0.3, -0.5)^T$, $\delta^0 = (0, 0.5, -0.75, 0)^T$, $\sigma_{S_1}^2 = 1$, $\sigma_{S_2}^2 = 1$, $\sigma_{Z_1}^2 = 1$, $\sigma_{Z_2}^2 = 1$, $\sigma_{X_1}^2 = 1$, $\sigma_{X_2}^2 = 1$, $\sigma_{Y_1}^2 = 1$, $\sigma_{Y_2}^2 = 1$, $\sigma_{V_1}^2 = 1$, $\sigma_{V_2}^2 = 1$, $\sigma_{W_1}^2 = 1$, $\sigma_{W_2}^2 = 1$, $\sigma_{U_1}^2 = 1$, $\sigma_{U_2}^2 = 1$, $\sigma_{T_1}^2 = 1$, $\sigma_{T_2}^2 = 1$, $\sigma_{Q_1}^2 = 1$, $\sigma_{Q_2}^2 = 1$, $\sigma_{A_1}^2 = 1$, $\sigma_{A_2}^2 = 1$, $\sigma_{R_1}^2 = 1$, $\sigma_{R_2}^2 = 1$, $\sigma_{S_1}^2 = 1$, $\sigma_{S_2}^2 = 1$, $\sigma_{X_1}^2 = 1$, $\sigma_{X_2}^2 = 1$, $\sigma_{Y_1}^2 = 1$, $\sigma_{Y_2}^2 = 1$, $\sigma_{V_1}^2 = 1$, $\sigma_{V_2}^2 = 1$, $\sigma_{W_1}^2 = 1$, $\sigma_{W_2}^2 = 1$, $\sigma_{U_1}^2 = 1$, $\sigma_{U_2}^2 = 1$, $\sigma_{T_1}^2 = 1$, $\sigma_{T_2}^2 = 1$, $\sigma_{Q_1}^2 = 1$, $\sigma_{Q_2}^2 = 1$, $\sigma_{A_1}^2 = 1$, $\sigma_{A_2}^2 = 1$, $\sigma_{R_1}^2 = 1$, $\sigma_{R_2}^2 = 1$, $\sigma_{S_1}^2 = 1$, $\sigma_{S_2}^2 = 1$, $\sigma_{X_1}^2 = 1$, $\sigma_{X_2}^2 = 1$, $\sigma_{Y_1}^2 = 1$, $\sigma_{Y_2}^2 = 1$, $\sigma_{V_1}^2 = 1$, $\sigma_{V_2}^2 = 1$, $\sigma_{W_1}^2 = 1$, $\sigma_{W_2}^2 = 1$, $\sigma_{U_1}^2 = 1$, $\sigma_{U_2}^2 = 1$, $\sigma_{T_1}^2 = 1$, $\sigma_{T_2}^2 = 1$, $\sigma_{Q_1}^2 = 1$, $\sigma_{Q_2}^2 = 1$, $\sigma_{A_1}^2 = 1$, $\sigma_{A_2}^2 = 1$, $\sigma_{R_1}^2 = 1$, $\sigma_{R_2}^2 = 1$, $\sigma_{S_1}^2 = 1$, $\sigma_{S_2}^2 = 1$, $\sigma_{X_1}^2 = 1$, $\sigma_{X_2}^2 = 1$, $\sigma_{Y_1}^2 = 1$, $\sigma_{Y_2}^2 = 1$, $\sigma_{V_1}^2 = 1$, $\sigma_{V_2}^2 = 1$, $\sigma_{W_1}^2 = 1$, $\sigma_{W_2}^2 = 1$, $\sigma_{U_1}^2 = 1$, $\sigma_{U_2}^2 = 1$, $\sigma_{T_1}^2 = 1$, $\sigma_{T_2}^2 = 1$, $\sigma_{Q_1}^2 = 1$, $\sigma_{Q_2}^2 = 1$, $\sigma_{A_1}^2 = 1$, $\sigma_{A_2}^2 = 1$, $\sigma_{R_1}^2 = 1$, $\sigma_{R_2}^2 = 1$, $\sigma_{S_1}^2 = 1$, $\sigma_{S_2}^2 = 1$, $\sigma_{X_1}^2 = 1$, $\sigma_{X_2}^2 = 1$, $\sigma_{Y_1}^2 = 1$, $\sigma_{Y_2}^2 = 1$, $\sigma_{V_1}^2 = 1$, $\sigma_{V_2}^2 = 1$, $\sigma_{W_1}^2 = 1$, $\sigma_{W_2}^2 = 1$, $\sigma_{U_1}^2 = 1$, $\sigma_{U_2}^2 = 1$, $\sigma_{T_1}^2 = 1$, $\sigma_{T_2}^2 = 1$, $\sigma_{Q_1}^2 = 1$, $\sigma_{Q_2}^2 = 1$, $\sigma_{A_1}^2 = 1$, $\sigma_{A_2}^2 = 1$, $\sigma_{R_1}^2 = 1$, $\sigma_{R_2}^2 = 1$, $\sigma_{S_1}^2 = 1$, $\sigma_{S_2}^2 = 1$, $\sigma_{X_1}^2 = 1$, $\sigma_{X_2}^2 = 1$, $\sigma_{Y_1}^2 = 1$, $\sigma_{Y_2}^2 = 1
$\phi_0^* = (0, 0.25, 0.3, -0.3, -0.4, \phi_2^0)^T$, $\beta_0^* = (10, 1, -3, -2, -2.5, \beta_2^0)^T$, $\psi_0^* = (1.5, 1, -3.5)^T$, and $\sigma^2 = 1$. When both the propensity model and $Q$-function are misspecified, $\beta_{12}^0 \neq 0$ and $\phi_{12}^0 \neq 0$. We use “equivalently misspecified” pairs (See Section G of the Appendix) to equate the levels of misspecification.

We compare the MSEs of estimates of key parameters affecting treatment decisions in Figure 3.6. In these results at the second stage, $\hat{\psi}_{21}$ is largely unaffected by sample size. Estimates of $\psi_{20}$ indicate that under these misspecifications, $A$-learning yields an estimate with smaller MSE. Results for $\hat{\psi}_{22}$ indicate that larger samples are associated with less variance under $A$-learning, lending to $A$-learning being preferred for small levels of misspecification. However, for larger misspecification, $Q$-learning might be preferred for all sample sizes.

![Figure 3.6](image)

Figure 3.6: A comparison of $Q$- and $A$-learning via Monte Carlo MSE ratios for estimators of components of $\psi_1$ and $\psi_2$ under misspecification of both the propensity model and the $Q$-function. MSE ratios $> 1$ favor $Q$-learning. Sample sizes of $n = 200, n = 500, n = 1000, n = 5000$, and $n = 10000$ are shown.
Figure 3.7 relates $R(d_{Q}^{\text{opt}})$ and $R(d_{A}^{\text{opt}})$ under this misspecification scenario. Sample size increases from left to right across the top panels then continues to the second row of panels. Throughout, we see no substantial differences between $Q$- and $A$- learning results. As expected, with increasing sample size, both methods perform closer to the benchmark ideal.

![Graph comparing Q- and A-learning](image)

Figure 3.7: A comparison of $Q$- and $A$-learning via efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_{A}^{\text{opt}})$ for estimating the true $d^{\text{opt}}$ under misspecification of both the propensity model and the $Q$-function. Sample sizes of $n = 200$, $n = 500$, $n = 1000$, $n = 5000$, and $n = 10000$ are shown.

### 3.1.3 Simulations for Alternate Parameterizations

Again consider the class of models in Equations 3.1–3.6 with sample size fixed at $n = 1000$. In the prior, we have considered limited parameterizations of $\phi_{1}^{0}, \delta_{1}^{0}, \phi_{2}^{0}, \beta_{2}^{0}, \psi_{2}^{0}$, and $\sigma^{2}$. Now, we explore the consequences of changing each parameter of the data generating
process under the various misspecification scenarios.

In the following, consider a baseline model parameterized as \( \phi_0 = (\phi_{10}^0 = 0.3, \phi_{11}^0 = -0.5)^T, \delta_1^0 = (\delta_{10}^0 = 0, \delta_{11}^0 = 0.5, \delta_{12}^0 = -0.75, \delta_{13}^0 = 0)^T, \sigma_{S_2}^2 = 1, \phi_2^0 = (\phi_{20}^0 = 0, \phi_{21}^0 = 0.25, \phi_{22}^0 = 0.3, \phi_{23}^0 = -0.3, \phi_{24}^0 = -0.4)^T, \beta_2^0 = (\beta_{20}^0 = 10, \beta_{21}^0 = 1, \beta_{22}^0 = -3, \beta_{23}^0 = -2, \beta_{24}^0 = -2.5)^T, \psi_2^0 = (\psi_{20}^0 = 1.5, \psi_{21}^0 = 1, \psi_{22}^0 = -3.5)^T, \) and \( \sigma^2 = 1. \)

Owing to the large number of parameters characterizing this data generating class, we focus only on changes to parameters in the model of the response. That is, for \( \beta_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25})^T, \psi_2 = (\psi_{20}, \psi_{21}, \psi_{22})^T, \) and \( \sigma^2 \) we seek to determine how changing individual parameters from this baseline model might alter results in the context of misspecification for the propensity model, \( Q \)-function, or both.

**Results for Misspecified Propensity**

First, we consider the case of a misspecified propensity model. In these circumstances, we might expect that \( A \)-learning, due to its double-robust properties, will yield consistent estimators of key parameters in the contrast function. \( Q \)-learning should be largely unaffected by misspecification in the propensity model. Figure 3.8 shows results under this scenario. In the top three panels, we compare the MSE for estimates of \( \psi_2 \) under \( Q \)- and \( A \)-learning, such that values above 1 favor \( Q \)-learning. The results coincide with results from Section 3.1.1; there appears to be small gains in efficiency under \( Q \)-learning attributable to the increased variability of \( A \)-learning.

In the lower panels, the two leftmost compare the MSE for estimates of \( \psi_1 \). There is no substantial difference between \( Q \)- and \( A \)-learning with respect to this measure of estimating these parameters. The right panel provides a measure of efficiency in achieving ideal outcomes for a hypothetical population of patients. Under this scenario, \( Q \)- and \( A \)-learning perform equally well with respect to this measure, for all levels of misspecification of the propensity model.

Additional results for other parameterizations of \( \beta_2^0 \) and \( \psi_2^0 \) indicate similar patterns. There does not appear to be a substantial change from the results of Figure 3.8 discussed above under alterations to these parameters.

**Results for Misspecified \( Q \)-function**

Again refer to the baseline parameterization above, but under misspecification of the \( Q \)-function. This scenario provides a test of the double-robust property of \( A \)-learning
Figure 3.8: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}^\text{opt}_Q)$ and $R(\hat{d}^\text{opt}_A)$ for estimating the true $d^\text{opt}$ (lower right panel) under misspecification of the propensity model. MSE ratios > 1 favor Q-learning.
and yields a comparison of the bias-variance tradeoff between $Q$- and $A$-learning. In Figure 3.9, the top panels provide a comparison of $Q$- and $A$-learning with respect to efficiency of estimating parameters $\psi_2$ (values $>1$ favor $Q$-learning). In each case, when misspecification is present, the bias of $Q$-learning quickly outweighs any additional variation under $A$-learning.

The left and center panels on the bottom row explore efficiency of estimating $\psi_1$ - the parameters dictating the treatment regime at the first stage. While there are some indications that $A$-learning yields a result with smaller MSE for extreme misspecification, the difference is small and not substantially different from $Q$-learning. Thus, big gains in efficiency for $A$-learning at the second stage do not appear to propagate back to the first stage estimation. Finally, consider the bottom right panel that plots $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_{A}^{\text{opt}})$. Interestingly, we again see that for some levels of misspecification, $A$-learning may have yielded better estimates of parameters, but does not yield higher efficiency in achieving benchmark outcomes. It appears that this can occur when, despite bias in estimates under $Q$-learning, the combination of parameters affecting treatment decision still yield a treatment regime close to the ideal, on average.

One might consider altering parameters of $h_2(\bar{S}_2, A_1; \beta_2^0)$. In the following, we change the baseline model such that $\beta_2^{03} = -4.5$ representing a situation where outcome is affected by a stronger interaction between $S_1$ and first stage treatment. The top panels of Figure 3.10 resemble the results of the baseline model. The same is true regarding efficiency of estimating $\psi_1$ in the bottom left and center panels. In this scenario, however, the peculiar results observed for efficiency in achieving $H(\hat{d}^{\text{opt}})$ no longer apply. Indeed, for this case, superior estimation of parameters with respect to the MSE due to bias under $Q$-learning does also yield larger $R(\hat{d}^{\text{opt}})$ under $A$-learning.

Another scenario arises under changes to the parameters in the contrast function, parameterized by $\psi_2^0 = (\psi_{20}^0, \psi_{21}^0, \psi_{22}^0)^T$. First, consider a new generative model where $\psi_{20}^0 = 3$, indicating a stronger marginal treatment effect. In Figure 3.11, results for comparing the ratio of MSEs under $A$- and $Q$-learning are unchanged from the baseline model. $R(\hat{d}^{\text{opt}})$ has changed slightly for $Q$- and $A$-learning, again with indications that under some levels of misspecification, $Q$-learning performs better on this measure, despite worse performance in estimating parameters of the contrast functions.

One could alternatively consider the generative model with $\psi_{21}^0 = 0$, such that there is no interaction between treatments at the first and second stage. In figure 3.12, results...
Figure 3.9: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_{Q}^{opt})$ and $R(\hat{d}_{A}^{opt})$ for estimating the true $d_{opt}$ (lower right panel) under misspecification of the $Q$-functions. MSE ratios > 1 favor $Q$-learning.
Figure 3.10: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_{A}^{\text{opt}})$ for estimating the true $d_{C}^{\text{opt}}$ (lower right panel) under misspecification of the $Q$-functions. MSE ratios $> 1$ favor $Q$-learning.
Figure 3.11: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_{A}^{\text{opt}})$ for estimating the true $d_{\text{opt}}$ (lower right panel) under misspecification of the $Q$-functions. MSE ratios $> 1$ favor $Q$-learning.
for estimating parameters of the contrast are unchanged. With respect to efficiency of attaining ideal outcomes, consistent estimators under $A$-learning appear to yield larger $R(\hat{d}^{opt}_A)$ compared to $Q$-learning for most levels of misspecification in the $Q$-function.

Figure 3.12: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}^{opt}_Q)$ and $R(\hat{d}^{opt}_A)$ for estimating the true $d^{opt}$ (lower right panel) under misspecification of the $Q$-functions. MSE ratios > 1 favor $Q$-learning.

Lastly, consider $\psi_{22}^0 = -1.75$ such that the interaction between covariate $S_2$ and second stage treatment is smaller. The top panels of Figure 3.13 compare efficiency of estimating $\psi_2$. For misspecification of the $Q$-function, bias under $Q$-learning leads to better estimation of these parameters under $A$-learning, as evidenced by ratios of MSEs less than 1. There is little difference in the quality of estimates, under this measure, between $Q$- and $A$-learning for estimates of $\psi_1$ in the bottom left and center panels. In this scenario, more dramatic changes can be observed between $Q$- and $A$-learning in their efficiency of achieving $H(d^{opt})$ compared to the baseline model; however, this represents
less than a 1% loss of efficiency.

Figure 3.13: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{opt})$ and $R(\hat{d}_A^{opt})$ for estimating the true $d^{opt}$ (lower right panel) under misspecification of the $Q$-functions. MSE ratios > 1 favor $Q$-learning.

Results for Misspecification of Both the Propensity Model and Q-function

In our final suite of comparing alternate generative model parameterizations, we do so under misspecification of both the propensity model and $Q$-function. The correspondence between misspecification of each model has been previously discussed. We first look at the baseline parameterization, as described at the beginning of this section. Figure 3.14 lets us compare $Q$- and $A$-learning. In the top panels, we see that $A$-learning generally yields better estimation of $\psi_{20}$ under misspecification. The MSE of $\hat{\psi}_{21}$ is approximately the same under $Q$- and iA-learning; and for $\psi_{22}$, estimates favor $A$-learning for small misspecifications of each model, but $Q$-learning for large misspecification.
Figure 3.14: Monte Carlo MSE ratios for estimators of components of \( \psi_2 \) and \( \psi_1 \) (upper row and lower row left and center panels) and efficiencies \( R(\hat{d}_{Q}^{\text{opt}}) \) and \( R(\hat{d}_{A}^{\text{opt}}) \) for estimating the true \( d_{Q}^{\text{opt}} \) (lower right panel) under misspecification of both the propensity models and \( Q \)-functions. MSE ratios > 1 favor \( Q \)-learning.
In the first alternative parameterization, we use $\beta_{23}^0 = -4.5$ indicating a strong interaction between the first stage covariate and treatment with respect to the outcome of interest. This, as shown in Figure 3.15, yields no change to the comparisons of MSEs for estimates of parameters in the contrast functions. Expected outcomes under each regime are improved, such that both $R(d_Q^{\text{opt}})$ and $R(d_A^{\text{opt}})$ nearly achieve the benchmark ideal.

Figure 3.15: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(d_Q^{\text{opt}})$ and $R(d_A^{\text{opt}})$ for estimating the true $d^{\text{opt}}$ (lower right panel) under misspecification of both the propensity models and $Q$-functions. MSE ratios > 1 favor $Q$-learning.

Finally, consider changes to parameters in the contrast function at the second stage, namely so that $\psi_{22}^0 = -1.75$. Little changes are observed from the baseline model when comparing $Q$- and $A$- learning under misspecification of both the propensity model and $Q$-functions. There is no clear pattern indicating favorability of one model or the other under this misspecification scenario with these parameterizations.
Figure 3.16: Monte Carlo MSE ratios for estimators of components of $\psi_2$ and $\psi_1$ (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_{Q}^{\text{opt}})$ and $R(\hat{d}_{A}^{\text{opt}})$ for estimating the true $d_{\text{opt}}$ (lower right panel) under misspecification of both the propensity models and $Q$-functions. MSE ratios $>1$ favor $Q$-learning.
3.2 Moodie and Richardson Simulations

The foregoing simulation scenarios deliberately involve simple models for the \( Q \)-functions in order to allow straightforward interpretation. To investigate the relative performance of the methods in a more challenging setting, we generated data from a scenario similar to that in Moodie et al. (2007) in which the true contrast functions are simple yet the \( Q \)-functions are complex.

The data generating process mimics a study in which HIV-infected patients are randomized to receive antiretroviral therapy (coded as 1) or not (coded as 0) at baseline and again at six months, where the randomization probabilities depend on baseline and six month CD4 counts. Specifically, we generated baseline CD4 count \( S_1 \sim \text{Normal}(450, 150^2) \), and baseline treatment \( A_1 \) was then assigned according to \( A_1|S_1 = s_1 \sim \text{Bernoulli}\{\expit(\phi_{10}^0 + \phi_{11}^0 s_1)\} \). We generated six month CD4 count \( S_2 \), distributed conditional on \( S_1 = s_1, A_1 = a_1 \) as \( \text{Normal}(1.25 s_1, 60^2) \). Treatment \( A_2 \) was then generated according to \( A_2|S_1 = s_1, A_1 = a_1, S_2 = s_2 \sim \text{Bernoulli}\{\expit(\phi_{20}^0 + \phi_{21}^0 s_2)\} \). In contrast to the scenario in Moodie et al. (2007), this allows all possible treatment combinations. The outcome \( Y \) is CD4 count at one year; following Moodie et al. (2007), \( Y \) was generated as \( Y = Y^{\text{opt}} - \mu_1^0(S_1, A_1) - \mu_2^0(S_1, S_2, A_1, A_2) \), where \( Y^{\text{opt}}|S_1 = s_1, A_1 = a_1, S_2 = s_2, A_2 = a_2 \sim \text{Normal}(400 + 1.6 s_1, 60^2) \). Here, \( \mu_1^0(S_1, A_1) \) and \( \mu_2^0(S_1, S_2, A_1, A_2) \) are the true advantage (regret) functions; we took \( C_1^0(s_1) = \psi_{10}^0 + \psi_{11}^0 s_1 \) and \( C_2^0(s_1, s_2, a_1) = \psi_{20}^0 + \psi_{21}^0 s_2 \) to be the true contrast functions, so that, from Section 1.3.2,

\[
\begin{align*}
\mu_1^0(S_1, A_1) &= (\psi_{10}^0 + \psi_{11}^0 S_1)\{I(\psi_{10} + \psi_{11} S_1 > 0) - A_1\}, \tag{3.7} \\
\mu_2^0(S_1, S_2, A_1, A_2) &= (\psi_{20}^0 + \psi_{21}^0 S_2)\{I(\psi_{20} + \psi_{21} S_2 > 0) - A_2\}. \tag{3.8}
\end{align*}
\]

It follows that the optimal treatment regime \( d^{\text{opt}} = (d_1^{\text{opt}}, d_2^{\text{opt}}) \) has \( d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0) \) and \( d_2^{\text{opt}}(s_1, s_2, a_1) = I(\psi_{20}^0 + \psi_{21}^0 s_2 > 0) \). While the true contrast functions are linear in \( \psi_k, k = 1, 2 \), the true implied \( h_1^0(s_1) \) and \( h_2^0(s_1, a_1, s_2) \) are nonsmooth and possibly complex.

Following Moodie et al. (2007), for \( A \)-learning, we assumed working models \( h_1(s_1; \beta_1) = \beta_{10} + \beta_{11} s_1, C_1(s_1; \psi_1) = \psi_{10} + \psi_{11} s_1, h_2(s_1, s_2, a_1; \beta_2) = \beta_{20} + \beta_{21} s_1 + \beta_{22} a_1 + \beta_{23} s_1 a_1 + \beta_{24} s_2, \) and \( C_2(s_1, s_2, a_1; \psi_2) = \psi_{20} + \psi_{21} s_2 \), and assumed propensity models of the form \( \pi_1(s_1; \phi_1) = \phi_{10} + \phi_{11} s_1 \) and \( \pi_2(s_1, s_2, a_1; \phi_2) = \phi_{20} + \phi_{21} s_2 \). For \( Q \)-learning, we analogously assumed \( Q \)-functions \( Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1) \) and \( Q_2(s_1, s_2, a_1, a_2; \xi_2) = \).
\[ h_2(s_1, s_1, a_1; \beta_2) + a_2C_2(s_1, s_2, a_1; \psi_2). \] Note that the contrast functions in each case are correctly specified, as are the propensity models; however, the \( Q \)-functions are misspecified, as the linear models \( h_1(s_1; \beta_1) \) and \( h_2(s_1, s_1, a_1; \beta_2) \) are poor approximations to the complex forms of the true \( h_0^0(s_1) \) and \( h_0^2(s_1, s_2, a_1) \).

We report results for \( n = 200 \) and \( n = 1000 \) with \( \phi_1^0 = (\phi_{10}^0, \phi_{11}^0)^T = (2.0, -0.006)^T, \phi_2^0 = (\phi_{20}^0, \phi_{21}^0)^T = (0.8, -0.004)^T, \psi_1^0 = (\psi_{10}^0, \psi_{11}^0)^T = (250, -1.0)^T, \) and \( \psi_2^0 = (\psi_{20}^0, \psi_{21}^0)^T = (720, -2.0)^T \) in Tables 3.1 and 3.2. Because the \( Q \)-functions are misspecified, not unexpectedly, the \( Q \)-learning estimators for \( \psi_1^0 \) and \( \psi_2^0 \) are biased, while those obtained via \( A \)-learning are consistent owing to the double robustness property. This leads to the dramatic relative inefficiency of \( Q \)-learning reflected by the MSE ratios. Under the assumed models, for the \( n = 200 \) case (Table 3.1), the estimated optimal regime for \( Q \)-learning dictates that, at baseline, antiretroviral therapy be given to patients with baseline CD4 count less than 203.5 (found by solving \( \hat{\psi}_{10} + \hat{\psi}_{11}S_1 > 0 \)), while that estimated using \( A \)-learning gives treatment to those with baseline CD4 count less than 247.4, substantially closer achieving the true optimal CD4 threshold of 250. At the second decision, the estimated optimal regimes obtained by \( Q \)- and \( A \)-learning dictate that therapy be given to patients with six month CD4 count less than 321.5 and 359.4, respectively (each identified by solving \( \hat{\psi}_{20} + \hat{\psi}_{21}S_2 > 0 \)). Again, \( A \)-learning yields an estimated threshold very close to the optimal value of 360.

Similarly, under the case where \( n = 1000 \) (Table 3.2), the estimated optimal regime for \( Q \)-learning dictates that, at baseline, antiretroviral therapy be given to patients with baseline CD4 count less than 199.7, while that estimated using \( A \)-learning gives treatment to those with baseline CD4 count less than 249.1, almost perfectly achieving the true optimal CD4 threshold of 250. Under the data generative process, using the baseline decision rule estimated via \( Q \)-learning may result in as many as 4.4\% of patients who would receive therapy at baseline under the true optimal regime being assigned no treatment. Whereas \( A \)-learning results nearly all patients receiving the ideal treatment. In the second stage, the estimated optimal regimes obtained under \( Q \)- and \( A \)-learning indicate beneficial therapy for patients with six month CD4 counts less than 320.3 and 360.0, respectively. That is, \( A \)-learning almost precisely identifies the true optimal decision rule at the second stage. The rule obtained via \( Q \)-learning is lower, such that 4.3\% of patients who should receive therapy at six months would not if the estimated six month rule from \( Q \)-learning were followed by the population.
Table 3.1: Monte Carlo average (standard deviation) of estimates obtained via Q- and A-learning and ratio of Monte Carlo MSE for the Moodie and Richardson scenario; MSE ratios > 1 favor Q-learning. n = 200.

<table>
<thead>
<tr>
<th>Parameter (true value)</th>
<th>Q-learning</th>
<th>A-learning</th>
<th>MSE ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \psi_{010} = 250 )</td>
<td>160.4 (52.4)</td>
<td>244.5 (43.4)</td>
<td>0.177</td>
</tr>
<tr>
<td>( \psi_{011} = -1.0 )</td>
<td>-0.788 (0.117)</td>
<td>-0.988 (0.095)</td>
<td>0.158</td>
</tr>
<tr>
<td>( \psi_{020} = 720 )</td>
<td>511.2 (110.6)</td>
<td>718.2 (109.7)</td>
<td>0.216</td>
</tr>
<tr>
<td>( \psi_{021} = -2.0 )</td>
<td>-1.590 (0.209)</td>
<td>-1.998 (0.194)</td>
<td>0.178</td>
</tr>
</tbody>
</table>

Table 3.2: Monte Carlo average (standard deviation) of estimates obtained via Q- and A-learning and ratio of Monte Carlo MSE for the Moodie and Richardson scenario; MSE ratios > 1 favor Q-learning. n = 1000.

<table>
<thead>
<tr>
<th>Parameter (true value)</th>
<th>Q-learning</th>
<th>A-learning</th>
<th>MSE ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \psi_{010} = 250 )</td>
<td>154.8 (23.2)</td>
<td>249.1 (18.7)</td>
<td>0.036</td>
</tr>
<tr>
<td>( \psi_{011} = -1.0 )</td>
<td>-0.775 (0.052)</td>
<td>-0.998 (0.041)</td>
<td>0.032</td>
</tr>
<tr>
<td>( \psi_{020} = 720 )</td>
<td>507.3 (49.2)</td>
<td>720.3 (48.4)</td>
<td>0.050</td>
</tr>
<tr>
<td>( \psi_{021} = -2.0 )</td>
<td>-1.584 (0.092)</td>
<td>-2.001 (0.085)</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Using the approach outlined in Section F of the Appendix, we have \( H(d^{\text{opt}}) = 1120 \), whereas, in the case of \( n = 1000 \), \( E\{H(\hat{d}^{\text{opt}}_Q)\} \approx 1117.1 \) (estimated standard error 1.3) and \( E\{H(\hat{d}^{\text{opt}}_A)\} \approx 1119.9 \) (0.3), so that \( R(\hat{d}^{\text{opt}}_Q) \) and \( R(\hat{d}^{\text{opt}}_A) \) are virtually equal to one. Thus, although Q-learning results in poor estimation of parameters in the contrast functions, efficiency loss for estimating the optimal regime is negligible. A similar situation is observed for \( n = 200 \), with \( E\{H(\hat{d}^{\text{opt}}_Q)\} \approx 1116.3 \) (estimated standard error 3.1) and \( E\{H(\hat{d}^{\text{opt}}_A)\} \approx 1119.2 \) (0.9). Again, each nearly achieves the true \( H(d^{\text{opt}}) = 1120 \). A possible explanation is that for the advantage (regret) functions in (3.7) and (3.8), patients near the true treatment decision boundary would have \( C^0_k(\bar{S}_k, \bar{A}_{k-1}) \), \( k = 1, 2 \), close to zero. Thus, even if a regime improperly assigns treatment, patients near this boundary have only a small loss in expected outcome. This, and the aforementioned fact that only a small subset of the population is affected by poor treatment decisions under Q-learning, results in the relatively good expected outcome under the estimated Q-learning regime.
3.3 Discussion

This chapter has compared $Q$- and $A$-learning in two stage simulation studies under a suite of parameterizations and data generating processes. Section 3.1 discussed the $K = 2$ decision points problem where the data generating process allowed us to observe the consequences of misspecified propensity models, misspecified $Q$-functions, and misspecification of both. Section 3.1.1 made a comparison between the two methods under each misspecification scenario. Under misspecification of the propensity model, $Q$-learning more efficiently estimated key parameters for the second stage decision, but did not yield substantial differences from $A$-learning at the first stage or with respect to achieving ideal outcomes. Section 3.1.3 showed a similar trend in results for the misspecified propensity scenario.

We also looked at misspecification of the $Q$-function, where $Q$-learning led to biased parameter estimates at the second stage. In particular, Section 3.1.1 showed preference to $A$-learning for some key parameter estimates, but not all. We also showed that, in most hypothesized scenarios, bias from $Q$-learning meant better outcomes, on average, under $A$-learning. However, in Sections 3.1.2 and 3.1.3, we saw that this was both dependent upon sample size and parameterization. Interestingly, for any parameterization, it appears that there exists a sufficiently large sample size such that the bias from $Q$-learning indeed outweighs any increased variability under $A$-learning, and hence $A$-learning will yield larger $R(\hat{d}^{opt})$.

When both the propensity model and $Q$-function were misspecified there was not a clear indicator for either $A$- or $Q$-learning outperforming the other method. Section 3.1.1 showed some favorability, overall, towards $Q$-learning for that particular scenario, though the advantages were not definitive or substantial. This was further complicated by results from Sections 3.1.2 and 3.1.3 showing that neither method can be assumed to perform better than the other when both models, the propensity and $Q$-function, are misspecified.

Finally, Section 3.2 discussed a generative process similar to the work of Moodie and Richardson (Moodie et al. 2007). This study posited a different approach to creating the outcome; as a result, the $Q$-function is misspecified in a complex way. $A$-learning, due to its double-robust properties, yielded good parameter estimates and outcomes regardless of sample size. $Q$-learning led to severely biased parameter estimates. However, the combination of parameters dictating the treatment regime showed that these large biases did not have a large impact on $Q$-learning’s ability to achieve good outcomes, on average.
In the final chapter of this dissertation, we will implement $Q$- and $A$-learning on a data set from a study of depression.
4.1 STAR*D Data Analysis

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) was a prospective multisite, randomized clinical trial enrolling 4041 patients designed to compare various treatment options for patients with major depressive disorder. The trial involved four levels, where each level consisted of a 12 week period of treatment, with scheduled clinic visits at approximate two week intervals (weeks 0, 2, 4, 6, 9, 12). Severity of depression at any visit was assessed using clinician-rated and self-reported versions of the Quick Inventory of Depressive Symptomatology (QIDS) score (Rush et al. 2003), for which higher values correspond to higher severity. At the end of each level, patients deemed to have an adequate clinical response to that level’s treatment did not move on to future levels, where an adequate response was defined by 12-week clinician-rated QIDS score ≤ 5 (remission) or showing a 50% or greater decrease from the baseline score at the beginning of level 1 (successful reduction). During level 1, all patients were treated with citalopram. Patients continuing to level 2 due to inadequate response were eligible to receive one of up to seven treatment options. We classify these options as either (i) switch: sertraline, bupropion, venlafaxine, or cognitive therapy, or (ii) augment: citalopram plus one of either bupropion, buspirone, or cognitive therapy. Patients assigned to cognitive therapy (alone or augmented with citalopram) were eligible, in the case of inadequate response, to move to a supplementary level 2A and switch to either bupropion or venlafaxine. All patients without adequate response at level 2 (or 2A, if applicable) continued to level 3. Level 3 treatments can again be classified as either (i) switch: mirtazepine or nortriptyline or
(ii) augment with either: lithium or triiodothyronine. Patients without adequate clinical response continued to level 4, requiring a switch to either tranylcypromine or mirtazapine combined with venlafaxine. For a complete description see Rush et al. (2004).

To demonstrate formulation of this problem within the framework of Sections 1.1 and 1.2, we take level 2A to be part of level 2 and consider only levels 2 and 3 of the study, calling them stages (decision points) 1 and 2, respectively \((K = 2)\). Hence, we include in the analysis only the 1260 patients who entered level 2; 330 of these subsequently continued to level 3. Let \(A_k\), \(k = 1, 2\), be the treatment assigned at stage \(k\) (beginning of level \(k + 1\)), taking values 0 (augment) or 1 (switch); both options are feasible for all eligible subjects. Let \(S_{10}\) denote baseline QIDS score and \(S_{11}\) denote the most recent QIDS score at level 1/beginning of level 2, respectively, so that \(S_1 = (S_{10}, S_{11})^T\) is information available immediately prior to the first decision. Similarly, let \(S_2\) be the information available immediately prior to decision 2; here, \(S_2\) is the most recent QIDS score at the end of level 2/beginning of level 3. Finally, let \(T\) be QIDS score at the end of level 3. Because some patients exhibited adequate response at the end of level 2 and did not progress to level 3, we define the outcome of interest to be \(-S_2\) (negative QIDS score at the end of level 2) for patients not moving to level 3 and \(-(S_2 + T)/2\) (average of negative QIDS scores at the end of levels 2 and 3) otherwise. Thus, writing \(L_0 = \max(5, S_{10}/2)\), \(Y = -S_2 I(S_2 \leq L_0) - (S_2 + T)I(S_2 > L_0)/2\), the cumulative average negative QIDS score. Thus, this demonstrates the case where outcome is a function of accrued information over the sequence of decisions.

It is straightforward to deduce from (1.14) that \(Q_2(\bar{s}_2, \bar{a}_2) = E(Y|\bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2) = -s_2\{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + E(-T|\bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2, S_2 > l_0)I(s_2 > l_0)/2\), so that \(V_2(\bar{s}_2, a_1) = -s_2 I(s_2 \leq l_0) + \{s_2 + U_2(\bar{s}_2, a_1)\}I(s_2 > l_0)/2\), where \(U_2(\bar{s}_2, a_1) = \max_{a_2} E(-T|\bar{S}_2 = \bar{s}_2, \bar{A}_1 = \bar{a}_1, A_2 = a_2, S_2 > l_0)\). Thus, from (1.17),

\[
Q_1(s_1, a_1) = E[-S_2 I(S_2 \leq l_0) + \{s_2 + U_2(\bar{s}_2, a_1)\}I(S_2 > l_0)/2|S_1 = s_a, A_1 = a_1].
\]

We describe implementation for Q-learning. At the second decision point, we must posit a model for \(Q_2(\bar{s}_2, \bar{a}_2)\). From the form of \(Q_2(\bar{s}_2, \bar{a}_2)\), we need only specify a model for \(E(-T|\bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2, S_2 > l_0)\); given the form of the conditioning set, this may be carried out using only the data from patients moving to level 3. Let \(s_{k2}\) be the slope of QIDS score over level \(k\) based on \(s_1\) and \(s_2\). We took the model to be of the form \(\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \beta_{23}a_1 + a_2(\psi_{20} + \psi_{21}s_2 + \psi_{22}s_{22} + \psi_{23}a_1)\) so that the posited Q-function
is

\[ Q_2(\bar{s}_2, \bar{a}_2; \xi_2) = -s_2 \{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)\{\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \beta_{23}a_1 \}
+ a_2(\psi_{20} + \psi_{21}s_2 + \psi_{22}s_{22} + \psi_{23}a_1)\}/2, \]

(4.1)

\[ \xi_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \psi_{20}, \psi_{21}, \psi_{22}, \psi_{23})^T. \]

Under (4.1), \( V_2(\bar{s}_2, a_1; \xi_2) = -s_2 \{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)\{\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \beta_{23}a_1 + (\psi_{20} + \psi_{21}s_2 + \psi_{22}s_{22} + \psi_{23}a_1)I(\psi_{20} + \psi_{21}s_2 + \psi_{22}s_{22} + \psi_{23}a_1 > 0)\}/2, \) and the “responses” \( \bar{V}_{2,i} \) for use in (1.24) may then be formed by substituting the estimate for \( \xi_2 \).

We took the posited \( Q \)-function at the first stage to be \( Q_1(s_1, a_1; \xi_1) = \beta_{10} + \beta_{11}s_{11} + \beta_{12}s_{12} + a_1(\psi_{10} + \psi_{11}s_{11} + \psi_{12}s_{12}); \) and \( \xi_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \psi_{10}, \psi_{11}, \psi_{12})^T. \) For \( A \)-learning, we posited models for the functions \( h_k(\bar{s}_k, \bar{a}_{k-1}) \) and \( C_k(\bar{s}_k, \bar{a}_{k-1}), k = 1, 2, \) in the obvious way analogous to the models above, and we took the propensity models to be of the form

\[ \pi_2(\bar{s}_2, a_1; \phi_2) = \expit(\phi_{20} + \phi_{21}s_2 + \phi_{22}s_{22} + \phi_{23}a_1) \text{ and } \pi_1(s_1; \phi_1) = \expit(\phi_{10} + \phi_{11}s_{11} + \phi_{12}s_{12}). \]

The results of this analysis are presented as follows in Table 4.1. In general, the methods yield similar estimates. Neither \( Q \)-nor \( A \)-learning estimates suggest a treatment effect at either stage of the analysis. We notice, however, that many of the parameters in the treatment contrasts tend to have a similar direction for their estimated effect, even though they are not significant. Thus, in the following, we conduct a subsequent analysis using more parsimonious models.

Given the initial results of Table 4.1, we seek to find a more parsimonious model that might yield more insight into these data and offer a more useful model. We use a process similar to backward selection, to determine if a more parsimonious model will adequately describe the relationship between covariates and response. The selection process described is ad hoc in nature and not the focus of this research. We start with \( Q_2(\bar{s}_2, \bar{a}_2; \xi_2) \) and \( Q_1(s_1, a_1; \xi_1) \) parameterized as above. Then, covariates are systematically removed from the posited \( Q \)-functions one at a time, where the covariate with the largest corresponding \( p \)-value of a test for the null hypothesis \( \xi = 0, \xi \in (\beta_{jk}, \psi_{jk}), k = 1, 2 \) and relevant \( j \) previously described, is removed. To retain homogeneity of models between \( Q \)- and \( A \)-learning, the average \( P \)-value between the two methods is used to determine what variable to remove from both algorithms. We also force a hierarchy, such that main effects
Table 4.1: STAR*D data analysis results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Q-learning Estimate</th>
<th>95% CI</th>
<th>A-learning Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>β_{02}</td>
<td>-1.05 (-3.60, 1.50)</td>
<td></td>
<td>-1.18 (-3.75, 1.40)</td>
<td></td>
</tr>
<tr>
<td>β_{12}</td>
<td>-0.77 (-0.94, -0.60)</td>
<td>*</td>
<td>-0.76 (-0.93, -0.58)</td>
<td>*</td>
</tr>
<tr>
<td>β_{22}</td>
<td>1.24 (0.54, 1.93)</td>
<td>*</td>
<td>1.12 (0.42, 1.83)</td>
<td>*</td>
</tr>
<tr>
<td>β_{32}</td>
<td>-0.48 (-1.83, 0.86)</td>
<td></td>
<td>-0.54 (-1.89, 0.82)</td>
<td></td>
</tr>
<tr>
<td>ψ_{02}</td>
<td>0.02 (-0.24, 0.34)</td>
<td></td>
<td>0.96 (-3.51, 5.44)</td>
<td></td>
</tr>
<tr>
<td>ψ_{12}</td>
<td>0.05 (-0.24, 0.34)</td>
<td></td>
<td>-0.01 (-0.31, 0.29)</td>
<td></td>
</tr>
<tr>
<td>ψ_{22}</td>
<td>-0.49 (-2.73, 1.75)</td>
<td></td>
<td>0.69 (-2.04, 3.43)</td>
<td></td>
</tr>
<tr>
<td>ψ_{32}</td>
<td>0.86 (-1.38, 3.10)</td>
<td></td>
<td>0.85 (-1.45, 3.15)</td>
<td></td>
</tr>
</tbody>
</table>

Stage 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Q-learning Estimate</th>
<th>95% CI</th>
<th>A-learning Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>β_{01}</td>
<td>-1.08 (-2.45, 0.30)</td>
<td>*</td>
<td>-0.67 (-2.12, 0.77)</td>
<td></td>
</tr>
<tr>
<td>β_{11}</td>
<td>-0.59 (-0.67, -0.50)</td>
<td>*</td>
<td>-0.61 (-0.70, -0.52)</td>
<td>*</td>
</tr>
<tr>
<td>β_{21}</td>
<td>0.03 (-0.43, 0.49)</td>
<td></td>
<td>0.14 (-0.34, 0.61)</td>
<td></td>
</tr>
<tr>
<td>ψ_{01}</td>
<td>0.82 (-1.31, 2.95)</td>
<td></td>
<td>0.37 (-1.98, 2.73)</td>
<td></td>
</tr>
<tr>
<td>ψ_{11}</td>
<td>0.02 (-0.11, 0.16)</td>
<td></td>
<td>0.04 (-0.11, 0.19)</td>
<td></td>
</tr>
<tr>
<td>ψ_{21}</td>
<td>1.06 (-0.04, 2.17)</td>
<td></td>
<td>0.68 (-0.58, 1.95)</td>
<td></td>
</tr>
</tbody>
</table>

can only be removed from the model if any interaction with that covariate has already been removed. The process concludes when P-value < 0.20 for all remaining main effects or interactions.

Based on this exploratory analysis and selection process, a model was selected such that the posited Q-function is

\[
Q_2(\hat{s}_2, \hat{a}_2; \xi_2) = -s_2\{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)(\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \psi_{20}a_2)/2,
\]

(4.2)

\[\xi_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \psi_{20})^T.\] Under (4.2), \[V_2(\bar{s}_2, a_1; \xi_2) = -s_2\{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)\{\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \psi_{20}I(\psi_{20} > 0)\}/2,\]

and the “responses” \(\tilde{V}_{2,i}\) are formed by substituting the estimate for \(\xi_2\). Based on our exploratory analysis and selection process, we took the posited Q-function at the first stage to be \(Q_1(s_1, a_1; \xi_1) = \beta_{10} + \beta_{11}s_{11} + \beta_{12}s_{12} + a_1(\psi_{10} + \psi_{11}s_{12});\) and \(\xi_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \psi_{10}, \psi_{11})^T.\) For A-learning, we posited models for the functions \(h_k(\bar{s}_k, \bar{a}_{k-1})\) and \(C_k(\bar{s}_k, \bar{a}_{k-1}), k = 1, 2,\) in the obvious way analogous to the Q-functions above. Propensity models were not changed in the course of this exploratory analysis, thus \(\pi_2(\bar{s}_2, a_1; \phi_2) = \expit(\phi_{20} + \phi_{21}s_2 + \phi_{22}s_{22} + \phi_{23}a_1)\)
Table 4.2: STAR*D data analysis results following and exploratory analysis. Asterisks indicate evidence at level of significance 0.05 that the parameter is non-zero.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>95% CI</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Stage 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>-1.46</td>
<td>(-3.47 , 0.55)</td>
<td>-1.47</td>
<td>(-3.49 , 0.54)</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>-0.75</td>
<td>(-0.88 , -0.61) *</td>
<td>-0.75</td>
<td>(-0.88 , -0.61) *</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>1.17</td>
<td>(0.52 , 1.81) *</td>
<td>1.17</td>
<td>(0.52 , 1.81) *</td>
</tr>
<tr>
<td>$\psi_{20}$</td>
<td>1.10</td>
<td>(0.02 , 2.19) *</td>
<td>1.12</td>
<td>(0.03 , 2.22) *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-1.12</td>
<td>(-2.22 , -0.03) *</td>
<td>-0.90</td>
<td>(-2.03 , 0.22)</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>-0.58</td>
<td>(-0.65 , -0.51) *</td>
<td>-0.59</td>
<td>(-0.66 , -0.52) *</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>0.01</td>
<td>(-0.42 , 0.45)</td>
<td>0.11</td>
<td>(-0.34 , 0.57)</td>
</tr>
<tr>
<td>$\psi_{10}$</td>
<td>1.12</td>
<td>(0.43 , 1.80) *</td>
<td>0.90</td>
<td>(0.17 , 1.64) *</td>
</tr>
<tr>
<td>$\psi_{12}$</td>
<td>1.15</td>
<td>(0.20 , 2.10) *</td>
<td>0.84</td>
<td>(-0.24 , 1.92)</td>
</tr>
</tbody>
</table>

and $\pi_1(s_1; \phi_1) = \expit(\phi_{10} + \phi_{11}s_{11} + \phi_{12}s_{12})$.

The results are presented in Table 4.2. At the first stage, $Q$-learning suggests a treatment switch for those with level 1 QIDS slope greater than -0.97 (obtained by solving $1.12 + 1.15s_{12} = 0$); $A$-learning assigns a treatment switch for those with QIDS slope during level 1 greater than -1.07. At the second stage (level 3), the results suggest that all patients should switch treatment and not augment their existing treatments.

### 4.2 Discussion

This dissertation has provided a self-contained and thorough account of $Q$- and $A$-learning methods for estimating optimal dynamic treatment regimes, including a detailed discussion of the underlying statistical framework in which these methods may be formalized and of their relative merits. Our simulation studies confirm that, while $A$-learning may be inefficient relative to $Q$-learning in estimating parameters that define the optimal regime when the $Q$-functions required for the latter are correctly specified, $A$-learning may offer robustness to such misspecification. Nonetheless, $Q$-learning may have practical advantages in that it involves modeling tasks familiar to most data analysts, allowing the use of standard diagnostic tools. On the other hand, $A$-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. However, A-learning increases in complexity with more than two treatment options at each stage, which may limit its appeal. Interestingly, our simulations demonstrate that inefficiency and bias in estimation of parameters defining the optimal regime does not necessarily translate into degradation of performance of the estimated regime for either method.

There remain many unresolved issues in estimation of optimal treatment regimes using these and other methods. Approaches to address the challenges of high-dimensional information and large numbers of decision points are required. Existing methods for model selection focusing on minimization of prediction error may not be best for developing models optimal for decision-making. Formal inference procedures for evaluating the uncertainty associated with estimation of the optimal regime are challenging due to the nonsmooth nature of decision rules, which in turn leads to nonregularity of the parameter estimators; see Chakraborty et al. (2010), Laber et al. (2010), Song et al. (2010), and Laber and Murphy (2011).

This dissertation discussed sequential decision-making in the context of personalized medicine, but many other applications of these methods exist where, at one or more times in an evolving process, an action must be taken from among a set of plausible actions. Indeed, Q-learning was originally proposed in the computer science literature with these more general problems in mind; see Shortreed et al. (2011) for discussion.
REFERENCES


Appendix A

Demonstration That Equations (1.5)–(1.8) Define an Optimal Regime

For \( k = 1, \ldots, K \) and any \( d \in D \), define the random variables \( \alpha_k\{\bar{S}^*_k(\bar{d}_{k-1})\} \) such that

\[
\alpha_k\{\bar{S}^*_k(\bar{d}_{k-1})\}(\omega) = V_k^{(1)}(\bar{s}_k, \bar{u}_{k-1}) \quad \text{(A.1)}
\]

for any \( \omega \in \Omega \), where \((\bar{s}_k, \bar{u}_{k-1})\) are defined by (1.3). We first show that, for any \( d \in D \),

\[
E\{Y^*(d) | S_1 = s_1, S_2^*(d_1), \ldots, S_K^*(d_{K-1})\} \leq E\{Y^*(\bar{d}_{K-1}, d^{(1)opt}_K) | S_1 = s_1, S_2^*(d_1), \ldots, S_K^*(\bar{d}_{K-1})\} = \alpha_K\{s_1, S_2^*(d_1), \ldots, S_K^*(\bar{d}_{K-1})\}. \quad \text{(A.2)}
\]

This follows because, for the set in \( \Omega \) where \( \{S_2^*(d_1) = s_2, \ldots, S_K^*(\bar{d}_{K-1}) = s_K\} \), the left- and right-hand sides of the first line of (A.2) are equal to

\[
E\{Y^*(d) | \bar{S}^*_K(\bar{d}_{K-1}) = \bar{s}_K\} = E\{Y^*(\bar{u}_{K-1}, u_K) | \bar{S}^*_K(\bar{u}_{K-1}) = \bar{s}_K\}, \quad \text{and} \quad \text{(A.3)}
\]

\[
E\{Y^*(d^{(1)opt}_K, d_{K-1}) | \bar{S}^*_K(\bar{d}_{K-1}) = \bar{s}_K\} = E\{Y^*(\bar{u}_{K-1}, d^{(1)opt}_K(\bar{s}_K, \bar{u}_{K-1})) | \bar{S}^*_K(\bar{u}_{K-1}) = \bar{s}_K\}. \quad \text{(A.4)}
\]
respectively. By the definition of $d_{K}^{(1)\text{opt}}$ in (1.5), (A.4) is greater than or equal to (A.3), and, by the definition of $V'_{K}$ in (1.6), (A.4) equals $V'_{K}(\bar{s}_{K}, \bar{u}_{K-1})$. Because these results hold for sets \{\$S_{2}^{\ast}(d_{1}) = s_{2}, \ldots, S_{K}^{\ast}(\bar{d}_{K-1}) = s_{K}\} for any ($s_{2}, \ldots, s_{K}$), and by the definition of $\alpha_{K}$ in (A.1), (A.2) holds. Taking conditional expectations given $S_{1} = s_{1}$ yields

$$E\{Y^{\ast}(d)|S_{1} = s_{1}\} \leq E\{Y^{\ast}(\bar{d}_{K-1}, d_{K}^{(1)\text{opt}})|S_{1} = s_{1}\} \leq E[\alpha_{K}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{K}^{\ast}(\bar{d}_{K-1})\}|S_{1} = s_{1}].$$

(A.5)

The equality in (A.5) holds for any $\bar{d}_{K-1} = (d_{1}, \ldots, d_{K-1})$, hence it must hold for $(d_{1}, \ldots, d_{K-2}, d_{K-1}^{(1)\text{opt}})$. Thus, we also have that

$$E\{Y^{\ast}(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}}, d_{K}^{(1)\text{opt}})|S_{1} = s_{1}\} \leq E[\alpha_{K}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{K-1}^{\ast}(\bar{d}_{K-2}), S_{K}^{\ast}(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}})\}|S_{1} = s_{1}].$$

(A.6)

Similarly, for any $k = K - 1, \ldots, 1$, we can show that

$$E[\alpha_{k+1}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k+1}^{\ast}(\bar{d}_{k})\}|S_{1} = s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k}^{\ast}(\bar{d}_{k-1})] \leq E[\alpha_{k+1}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k+1}^{\ast}(\bar{d}_{k-1}, d_{k}^{(1)\text{opt}})\}|S_{1} = s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k}^{\ast}(\bar{d}_{k-1})] = \alpha_{k}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k}^{\ast}(\bar{d}_{k-1})\},$$

which implies for $k = K - 1, \ldots, 1$,

$$E[\alpha_{k+1}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k+1}^{\ast}(\bar{d}_{k})\}|S_{1} = s_{1}] \leq E[\alpha_{k+1}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k+1}^{\ast}(\bar{d}_{k-1}, d_{k}^{(1)\text{opt}})\}|S_{1} = s_{1}] = E[\alpha_{k}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{k}^{\ast}(\bar{d}_{k-1})\}|S_{1} = s_{1}].$$

(A.7)

Using (A.5) and (A.7) with $k = K - 1$, we thus have

$$E\{Y^{\ast}(d)|S_{1} = s_{1}\} \leq E\{Y^{\ast}(\bar{d}_{K-1}, d_{K}^{(1)\text{opt}})|S_{1} = s_{1}\} \leq E[\alpha_{K}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{K}^{\ast}(\bar{d}_{K-1})\}|S_{1} = s_{1}] \leq E[\alpha_{K}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{K-1}^{\ast}(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}})\}|S_{1} = s_{1}] \leq E[\alpha_{K-1}\{s_{1}, S_{2}^{\ast}(d_{1}), \ldots, S_{K-1}^{\ast}(\bar{d}_{K-2})\}|S_{1} = s_{1}]$$

(A.8)
Because of (A.6), the term in (A.8) is equal to $E\{Y^*(d_{K-2}, d^{(1)opt}_{K-1}, d^{(1)opt}_K)|S_1 = s_1\}$. Hence,

$$E\{Y^*(d)|S_1 = s_1\} \leq E\{Y^*(d_{K-1}, d^{(1)opt}_K)|S_1 = s_1\}$$

$$\leq E\{Y^*(d_{K-2}, d^{(1)opt}_{K-1}, d^{(1)opt}_K)|S_1 = s_1\}$$

$$= E[\alpha_{K-1}\{s_1, S^*_2(d_1), \ldots, S^*_K(d_{K-2})\}|S_1 = s_1]. \quad (A.9)$$

Again, because $d_{K-2}$ is arbitrary, if we replace it by $(d_{K-3}, d^{(1)opt}_{K-2})$, the equality in (A.9) implies

$$E\{Y^*(d_{K-3}, d^{(1)opt}_{K-2})|S_1 = s_1\} = E[\alpha_{K-1}\{s_1, S^*_2(d_1), \ldots, S^*_K(d_{K-3}, d^{(1)opt}_{K-2})\}|S_1 = s_1], \quad (A.10)$$

where, for any $d$, $d_k = (d_k, \ldots, d_K)$. Using (A.7) with $k = K - 2$, (A.9), and (A.10), we obtain

$$E\{Y^*(d_{K-2}, d^{(1)opt}_{K-1})|S_1 = s_1\} = E[\alpha_{K-1}\{s_1, S^*_2(d_1), \ldots, S^*_K(d_{K-2})\}|S_1 = s_1]$$

$$\leq E[\alpha_{K-1}\{s_1, S^*_2(d_1), \ldots, S^*_K(d_{K-2}, d^{(1)opt}_{K-2})\}|S_1 = s_1]$$

$$= E\{Y^*(d_{K-3}, d^{(1)opt}_{K-2})|S_1 = s_1\}$$

$$= E[\alpha_{K-2}\{s_1, S^*_2(d_1), \ldots, S^*_K(d_{K-3})\}|S_1 = s_1].$$

Continuing in this fashion, we may conclude that, for any $d \in \mathcal{D}$,

$$E\{Y^*(d)|S_1 = s_1\} \leq \cdots \leq E\{Y^*(d_{k-1}, d^{(1)opt}_k)|S_1 = s_1\} \leq \cdots \leq E\{Y^*(d^{(1)opt})|S_1 = s_1\},$$

showing that $d^{(1)opt}$ defined in (1.5) and (1.7) is an optimal regime satisfying (1.4).
Appendix B

Demonstration of Correspondence in Equations (1.20)–(1.22) Under Assumptions in Chapter 1.1

We first consider the case $\ell = 1$. We make the positivity assumption that, for any $(\bar{s}_k, \bar{a}_{k-1})$ for which $\Pr(\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$, $\Pr(A_k = a_k | \bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$ if and only if $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$. This ensures that the observed data contain information on the treatments involved in the class of feasible regimes under consideration.

We have $\Gamma_k^{(1)} = \Gamma_k$ by definition, so we need only demonstrate (1.21) and (1.22). We must show that, for any $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$ and $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \ldots, K$,

\[ \Pr(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0, \quad (B.1) \]
\[ \Pr(S_{k+1} = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) = \Pr\{S_{k+1}^* (\bar{a}_k) = s_{k+1} \mid \bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}\}, \quad (B.2) \]
\[ = \Pr\{S_{k+1}^* (\bar{a}_k) = s_{k+1} \mid \bar{S}_j = \bar{s}_j, \bar{A}_{j-1} = \bar{a}_{j-1}, S_{j+1}^* (\bar{a}_j) = s_{j+1}, \ldots, S_k^* (\bar{a}_{k-1}) = s_k\}, \quad (B.3) \]

for $j = 1, \ldots, k$, where we define (B.3) with $j = k$ to be the same as the expression on the right-hand side of (B.2) and take $S_{k+1} = Y$ and $S_{k+1}^* (\bar{a}_k) = Y^* (\bar{a}_k)$.

Assume for the moment that (B.1) is true. We now demonstrate (B.2) and (B.3). For any fixed $k$, by the consistency assumption, the left-hand expression in (B.2) is
equal to \( \Pr\{S'_{k+1}(\tilde{a}_k) = s_{k+1}|\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}, A_k = a_k\} \). It follows by the sequential randomization assumption, which implies \( A_k \perp S'_{k+1}(\tilde{a}_k)|\bar{S}_k, \bar{A}_{k-1} \), that this is equal to the right-hand side of (B.2). The equality in (B.3) follows by induction. Specifically, treating the right-hand side of (B.2) as (B.3) with \( j = k \), the equality follows if we can show that (B.3) being true for a given \( j \) implies that it is also true for \( j - 1 \). For a given \( j = 2, \ldots, k \), by the consistency assumption, (B.3) is equal to \( \Pr\{S'_{k+1}(\tilde{a}_k) = s_{k+1}|\bar{S}_{j-1} = \bar{s}_{j-1}, \bar{A}_{j-2} = \bar{a}_{j-2}, A_{j-1} = a_{j-1}, S'_{j}(\tilde{a}_j) = s_j, \ldots, S'_{k}(\tilde{a}_{k-1}) = s_k\} \). By the sequential randomization assumption, \( A_{j-1} \perp \{S'_j(\tilde{a}_j), \ldots, S'_{k+1}(\tilde{a}_k)\}|\bar{S}_{j-1}, \bar{A}_{j-2} \), so that this expression is equal to \( \Pr\{S'_{k+1}(\tilde{a}_k) = s_{k+1}|\bar{S}_{j-1} = \bar{s}_{j-1}, \bar{A}_{j-2} = \bar{a}_{j-2}, S'_j(\tilde{a}_j) = s_j, \ldots, S'_{k}(\tilde{a}_{k-1}) = s_k\} \), which is (B.3) for \( j - 1 \). Note, then, that this implies that the conditional densities in (B.3), which are \( j \)-dependent, are the same as those on the left-hand side of (B.2), which are not.

We now prove (B.1) by induction. Assume we have shown that \( \Pr(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0 \). Then we must show that \( \Pr(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) > 0 \). If \( \Pr(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0 \), then

\[
\Pr(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k) = \Pr(S_{k+1} = s_{k+1}|\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k)\Pr(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k).
\]

But we have shown above that if (B.1) is true; i.e., \( \Pr(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0 \), then (B.2) and (B.3) are equal for all \( j \) and in particular \( \Pr(S_{k+1} = s_{k+1}|\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) = \Pr(S'_{k+1}(\tilde{a}_k) = s_{k+1}|\bar{S}_k(\tilde{a}_{k-1}) = \bar{s}_k) \). Because \( (\bar{s}_{k+1}, \bar{a}_k) \in \Gamma_{k+1} \), then by condition (ii) of (1.2) defining \( \Gamma_{k+1} \), \( \Pr(S'_{k+1}(\tilde{a}_k) = s_{k+1}|\bar{S}_k(\tilde{a}_{k-1}) = \bar{s}_k) > 0 \) because of (B.4).

Now \( \Pr(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) = \Pr(A_{k+1} = a_{k+1}|\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k)\Pr(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k) \); however, because \( a_{k+1} \in \Psi_k(\bar{s}_{k+1}, \bar{a}_k) \) and by the positivity assumption, \( \Pr(A_{k+1} = a_{k+1}|\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k) > 0 \) and hence \( \Pr(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) > 0 \). The proof is complete by noting that \( \Pr(S_1 = s_1, A_1 = a_1) = \Pr(A_1 = a_1|S_1 = s_1)\Pr(S_1 = s_1) \), where \( \Pr(A_1 = a_1|S_1 = s_1) > 0 \) for \( a_1 \in \Psi(s_1) \) by the positivity assumption.

To demonstrate (1.21) and (1.22) for \( \ell = 1 \), consider first the definitions, given in (1.5) and (1.6), of \( d^{(1)}_K(\bar{s}_K, \bar{a}_{K-1}) \) and \( V^{(1)}_K(\bar{s}_K, \bar{a}_{K-1}) \). These quantities involve the conditional expectation of the potential outcome \( Y^*(\bar{a}_K) \) given \( \bar{S}_K(\tilde{a}_{k-1}) \), which by (B.2)-(B.3) is the same as the conditional expectation of \( Y \) given \( \{\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K\} \). Thus, \( d^{(1)}_K(\bar{s}_K, \bar{a}_{K-1}) \) and \( V^{(1)}_K(\bar{s}_K, \bar{a}_{K-1}) \) are the same as \( d^{opt}_K(\bar{s}_K, \bar{a}_{K-1}) \) and \( V_K(\bar{s}_K, \bar{a}_{K-1}) \).
defined in (1.15) and (1.16). Next, from (1.7) and (1.8), \(d_{K-1}^{1\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2})\) is given by
\[
\arg \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} E[V^{(1)}_{K}\{\bar{s}_{K-1}, S^*_K(\bar{a}_{K-2}, a_{K-1}), \bar{a}_{K-2}, a_{K-1}\}\big| \bar{S}^*_K(\bar{a}_{K-2}) = \bar{s}_{K-1}].
\]
This involves the conditional expectation of \(V^{(1)}_{K}\), a function of \(S^*_K(\bar{a}_{K-1})\), given that \(\bar{S}^*_K(\bar{a}_{K-2}) = \bar{s}_{K-1}\). Again, by (B.2)-(B.3), this is the same as the conditional expectation of the function \(V^{(1)}_{K}\) of \(S_K\) given \(\{\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}\}\). Because we have already shown that \(V^{(1)}_{K}\) is the same as \(V_{K}\), this implies that \(d_{K-1}^{1\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2})\) is given by
\[
\arg \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} E\{V_K(\bar{s}_{K-1}, S_K, \bar{a}_{K-2}, a_{K-1})\big| \bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = (\bar{a}_{K-2}, a_{K-2})\},
\]
which is the same as \(d^{\text{opt}}_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})\) given by (1.18) with \(k = K - 1\). The argument continues in a backward iterative fashion for \(k = K - 2, \ldots, 1\).

Now consider \(\ell > 1\). The sets \(\mathcal{V}_{\ell,K}, \ell = 1, \ldots, K, k = \ell, \ldots, K\), representing events of the form \(\{\bar{s}^{(P)}_{\ell} = \bar{s}_\ell, \bar{A}^{(P)}_{\ell+1} = a_{\ell+1}, S^*_{\ell+1}(\bar{a}_{\ell+1}) = s_{\ell+1}, \ldots, S^*_K(\bar{a}_{K-1}) = s_{K-1}\}\), involved in the definitions of \(\Gamma^{(\ell)}_k\) and (1.10)-(1.13), depend on the random variables \(S^{(P)}_k, \bar{A}^{(P)}_{k-1}\) for \(k = \ell, \ldots, K\), which characterize how treatment assignment and covariate history arise in the population under routine practice. To demonstrate (1.20)-(1.22), in addition to those on the observed random variables given above, we also require sequential randomization and positivity assumptions on the “population” random variables; namely, that \(A^{(P)}_k \perp W|S^{(P)}_k, \bar{A}^{(P)}_{k-1}, k = 1, \ldots, K\); and, for any \((\bar{s}_k, \bar{a}_{k-1})\) for which \(\pr(S^{(P)}_k = \bar{s}_k, \bar{A}^{(P)}_{k-1} = \bar{a}_{k-1}) > 0\), \(\pr(A^{(P)}_k = a_k|S^{(P)}_k = \bar{s}_k, \bar{A}^{(P)}_{k-1} = \bar{a}_{k-1}) > 0\) if and only if \(a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})\), \(k = 1, \ldots, K\). If the observed data are from an observational study where \(S_k, A_k\) are the same as \(S^{(P)}_k, A^{(P)}_k\), these assumptions are equivalent to those on the observed data. For data from a SMART, however, more consideration is required. If the treatment options considered in the trial are restricted relative to those available in practice, then an estimated optimal regime based on the observed data may not be applicable to patients who present at the \(\ell\)th decision with treatment histories involving options not considered in the trial for \(\ell > 1\). The positivity assumption here rules out such patients from consideration. The sequential randomization assumption holds for observed data by design for a SMART. However, whether or not it holds in the population, as we require here, depends on whether or not the covariate information collected in the trial contains the information used by patients and their providers to make treatment decisions in
routine practice. If this is not the case, then the estimated optimal regime based on the trial data is still applicable to patients who present prior to the first decision, \( \ell = 1 \), but may not lead to optimal decision-making for patients presenting at subsequent decision points because the sequential randomization assumption at the population level may no longer hold.

Under these assumptions, it follows by an argument analogous to that above that (B.1)–(B.3) hold with the random variables \( S_k, A_k \) replaced by \( \bar{S}_k^{(P)}, \bar{A}_k^{(P)} \), \( k = 1, \ldots, K \); namely

\[
\text{pr}(\bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k) > 0, \quad \text{(B.5)}
\]

\[
\text{pr}(S_{k-1}^{(P)} = s_{k-1} | \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k)
= \text{pr}\{S_{k-1}^{(P)}(\bar{a}_k) = s_{k-1} | \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k\},
= \text{pr}\{S_{k-1}^{(P)}(\bar{a}_k) = s_{k-1} | \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k\}
= \text{pr}\{S_{k-1}^{(P)}(\bar{a}_k) = s_{k-1} | \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k\}\]
\[
\text{(B.6)}
\]

for \( j = 1, \ldots, k \). We may then show that (1.20) holds as follows. Inspection of \( \Gamma_k \) and \( \Gamma_k^{(\ell)} \) shows both sets involve the same condition (i). Accordingly, we need only demonstrate that, if condition (ii) in \( \Gamma_k \) holds, then so does (ii) in \( \Gamma_k^{(\ell)} \), and vice versa. Condition (ii) in \( \Gamma_k^{(\ell)} \) states that \( \text{pr}(V_{\ell,k}) > 0 \). Because the set \( V_{\ell,k} \subseteq \{ \bar{S}_k^{(P)}(\bar{a}_{k-1}) = \bar{s}_k \} \), condition (ii) in \( \Gamma_k \) follows immediately. In the converse direction, if (ii) of \( \Gamma_k \) holds, then (B.5) holds. Because the set \{ \( \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k \) \} \subseteq V_{\ell,k}, \text{pr}(V_{\ell,k}) > 0 \), which is (iii) of \( \Gamma_k^{(\ell)} \).

Now (1.21) and (1.22) follow by an argument similar to that for \( \ell = 1 \). First, we argue that, for any fixed \( k = 1, \ldots, K \), the probabilities in (B.2) and (B.3) are the same as those in (B.6) and (B.7) for all \( j = 1, \ldots, k \). This follows because (B.3) with \( j = 1 \) is equal to (B.7) with \( j = 1 \). We may now use this to show the result. Consider the definitions of \( a_k^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) \) and \( V_K^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) \) given in (1.10) and (1.11). These quantities involve the conditional expectation of the potential outcome \( Y^*(\bar{a}_K) \) given \{ \( \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}, S_{\ell+1}^{(P)} = s_{\ell+1}, \ldots, S_K^{(P)} = S_K^{(P)}(\bar{a}_{K-1}) = \bar{a}_{K-1} \}\). But, because of the above equivalence of (B.2)–(B.3) and (B.6)–(B.7), this is the same as the conditional expectation of \( Y \) given \{ \( \bar{s}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K \) \}. Thus, \( a_k^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) \) and \( V_K^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) \) are the same as \( a_k^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) \) and \( V_K(\bar{s}_K, \bar{a}_{K-1}) \) defined in (1.15) and (1.16), and this is true for all \( \ell = 2, \ldots, K \). Next, in accordance with (1.21) and (1.22), \( a_k^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) \)
is given by

\[ d_{K-1}^{(\ell)\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \arg \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} E[V_K^{(\ell)}(\bar{s}_{K-1}, S^*_K(\bar{a}_{K-2}, a_{K-1})|\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{S}_{\ell+1}^{(P)} = \bar{a}_{\ell-1}, S^*_K(\bar{a}_{K-2}) = s_{K-1}]]. \]

Note that this involves the conditional expectation of the function \( V_K^{(\ell)} \) of \( S^*_K(\bar{a}_{K-1}) \) given \( S_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}(\bar{s}_\ell) = s_\ell, \ldots, S_{K-1}^*(\bar{a}_{K-2}) = s_{K-1} \). Again, this is the same as the conditional expectation of the function \( V_K^{(\ell)} \) of \( S_K \) given \( \{\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}\} \). Because we have shown that \( V_K^{(\ell)} \) is independent of \( \ell \) and equal to \( V_K \), this implies that

\[ d_{K-1}^{(\ell)\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \arg \max_{\bar{a}_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} E[V_K(\bar{s}_{K-1}, S_K, \bar{a}_{K-2}, a_{K-1})|\bar{S}_K = \bar{s}_K, \bar{A}_{K-1}], \]

which is the same as \( d_{K-1}^{\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}) \) given by (1.18) with \( k = K - 1 \). The argument continues in an backward iterative fashion for \( k = K - 2, \ldots, 1 \).
Appendix C

Justification for $\tilde{V}_{ki}$ in A-learning

We wish to show that

$$E(V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1})[I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k]|\bar{S}_k, \bar{A}_{k-1}) = V_k(\bar{S}_k, \bar{A}_{k-1}). \quad (C.1)$$

Defining $\Gamma(\bar{S}_{k+1}, \bar{A}_k) = V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1})[I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k]$, we may write (C.1) as

$$E[E\{\Gamma(\bar{S}_{k+1}, \bar{A}_k)|\bar{S}_k, \bar{A}_k\}|\bar{S}_k, \bar{A}_{k-1}]. \quad (C.2)$$

The inner expectation in (C.2) may be seen to be equal to

$$E\{V_{k+1}(\bar{S}_{k+1}, \bar{A}_k)|\bar{S}_k, \bar{A}_k\} + C_k(\bar{S}_k, \bar{A}_{k-1})[I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k] = Q_k(\bar{S}_k, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1})[I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k].$$

Substituting $Q_k(\bar{S}_k, \bar{A}_k) = h_k(\bar{S}_k, \bar{A}_{k-1}) + A_kC_k(\bar{S}_k, \bar{A}_{k-1})$, $h_k(\bar{S}_k, \bar{A}_{k-1}) = Q_k(\bar{S}_k, \bar{A}_{k-1}, 0)$, we obtain

$$E[\Gamma(\bar{S}_{k+1}, \bar{A}_k)|\bar{S}_k, \bar{A}_k] = h_k(\bar{S}_k, \bar{A}_{k-1}) + C_k(\bar{S}_k, \bar{A}_{k-1})I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} = V_k(\bar{S}_k, \bar{A}_{k-1}).$$

Substituting this in (C.2) yields the result.
Appendix D

Demonstration of Equivalence of $Q$- and $A$-learning in a Special Case

We take $K = 1$ and let $\text{pr}(A_1 = 1|S_1 = s_1) = \pi$. Consider the $A$-learning estimating equations (1.28) with $k = 1$, and take $\lambda_1(s_1; \psi_1) = \partial / \partial \psi_1 C_1(s_1; \psi_1)$. Then the equations become

$$\sum_{i=1}^{n} \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} (A_{1i} - \pi)\{Y_i - A_{1i}C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0,$$

$$\sum_{i=1}^{n} \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1}\{Y_i - A_{1i}C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0.$$

Likewise, under these conditions, taking $Q_1(s_1, a_1) = a_1C_1(s_1; \psi_1) + h(s_1; \beta_1)$, the $Q$-learning equation is

$$\sum_{i=1}^{n} \frac{\partial Q_1(S_{1i}, A_{1i}; \xi_1)}{\partial \xi_1}\{Y_i - A_{1i}C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0,$$

where, with $\xi_1 = (\psi_1^T, \beta_1^T)^T$,

$$\frac{\partial Q_1(S_{1i}, A_{1i}; \xi_1)}{\partial \xi_1} = \left( A_{1i} \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} \begin{pmatrix} \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1} \end{pmatrix} \right).$$
Thus note that, with \( C_1(s_1; \psi) \) and \( h_1(s_1; \beta) \) linear in functions of \( S_1 \), as long as terms of the form in \( C_1(s_1; \psi) \) are contained in those in \( h_1(s_1, \beta) \), the \( Q \)- and \( A \)-learning estimating equations are identical, as then

\[
\sum_{i=1}^{n} \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} \{ Y_i - A_{1i}C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1) \} = 0.
\]

For example, if \( C_1(s_1; \psi) = \psi_10 + s_1^T \psi_11 \) and \( h_1(s_1; \beta) = \beta_10 + s_1^T \beta_11 \), then note that

\[
\frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} = \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1} = \begin{pmatrix} 1 \\ S_{1i} \end{pmatrix},
\]

and the result is immediate.
Example of Incompatibility of $Q$-function Models

To show (1.30), noting $\mathcal{H}_2 = (1, s_1, a_1, s_2)^T = (K_1^T, s_2)^T$, we have

$$
E\{V_2(s_1S_2, a_1; \xi_2)|S_1 = s_1, A_1 = a_1\} = K_1^T \beta_{21} + \beta_{22} E(S_2|S_1 = s_1, A_1 = a_1)
$$

$$
+ (K_1^T \psi_{21}) E\{I(K_1^T \psi_{21} + S_2 \psi_{22} > 0)|S_1 = s_1, A_1 = a_1\}
$$

$$
+ \psi_{22} E\{S_2 I(K_1^T \psi_{21} + S_2 \psi_{22} > 0)|S_1 = s_1, A_1 = a_1\}.
$$

Taking $\psi_{22} > 0$, we also have $I(K_1^T \psi_{21} + S_2 \psi_{22} > 0) = I(S_2 > -K_1^T \psi_{21}/\psi_{22})$, from which it follows that $E\{I(K_1^T \psi_{21} + S_2 \psi_{22} > 0)|S_1 = s_1, A_1 = a_1\} = 1 - \Phi\{(-K_1^T \psi_{21}/\psi_{22} - K_1^T \gamma)/\sigma\} = 1 - \Phi(\eta)$ for $\eta = -K_1^T (\psi_{21}/\psi_{22} + \gamma)/\sigma$. Similarly, $E\{S_2 I(K_1^T \psi_{21} + S_2 \psi_{22} > 0)|S_1 = s_1, A_1 = a_1\} = E\{S_2 I(S_2 > -K_1^T \psi_{21}/\psi_{22})|S_1 = s_1, A_1 = a_1\}$. It is straightforward to deduce that this is equal to $\int_{\eta}^{\infty} (\sigma t + K_1^T \gamma) \varphi(t) dt = \sigma \varphi(\eta) + (K_1^T \gamma)\{1 - \Phi(\eta)\}$. Using $E(S_2|S_1 = s_1, A_1 = a_1) = K_1^T \gamma$ and combining yields (1.30).
Calculation for $K = 1$. We consider the generative data model in Section 2.1 and treatment regimes of the form $d(s_1) = d_1(s_1) = I(\psi_{10} + \psi_{11}s_1 > 0)$ for arbitrary $\psi_{10}$, $\psi_{11}$. It is possible to derive analytically $H(d) = E\{Y^*(d)\}$ in this case. Under the generative data model, $E[Y^*(d)] = E[E\{Y^*(d)|S_1\}] = E[E\{Y|S_1, A_1 = d_1(S_1)\}] = \beta_0^0 + \beta_1^0 \cdot E(S_1) + \beta_2^0 \cdot E(S_1^2) + E\{I(\psi_{10} + \psi_{11}S_1 > 0)(\psi_{01}^0 + \psi_{11}S_1)\}$, and $S_1 \sim \text{Normal}(0, 1)$. It is straightforward to deduce that $E\{I(\psi_{10} + \psi_{11}S_1 > 0)\} = \text{pr}(S_1 > -\psi_{10}/\psi_{11})$ or $\text{pr}(S_1 < -\psi_{10}/\psi_{11})$ as $\psi_{11} > 0$ or $\psi_{11} < 0$, which is readily obtained from the standard normal cdf. Likewise, $E\{I(\psi_{10} + \psi_{11}S_1 > 0)\} = E(S_1|S_1 > -\psi_{10}/\psi_{11})\text{pr}(S_1 > -\psi_{10}/\psi_{11})$ if $\psi_{11} > 0$ and $E\{S_1 I(\psi_{10} + \psi_{11}S_1 > 0)\} = E(S_1|S_1 < -\psi_{10}/\psi_{11})\text{pr}(S_1 < -\psi_{10}/\psi_{11})$ if $\psi_{11} < 0$, which are again easily calculated in a manner similar to that in Section E. Thus, $E\{Y^*(d_{\text{opt}})\}$ is obtained by substituting $\psi_{10}^0$, $\psi_{11}^0$ in the resulting expression. To approximate $E\{H(\hat{d}_{\text{opt}})\}$ and hence $R(\hat{d}_{\text{opt}})$ for $\hat{d}_{\text{opt}} = d_Q^\text{opt}$ or $d_A^\text{opt}$, we may use Monte Carlo simulation. Specifically, for the $b$th of $B$ Monte Carlo data sets, substitute the estimates $\hat{\psi}_{10,b}$, $\hat{\psi}_{11,b}$, say, defining $\hat{d}_{\text{opt}}$ for that data set in the expression for $E\{Y^*(d)\}$, and call the resulting quantity $U_b$. Then $E\{H(\hat{d}_{\text{opt}})\}$ is approximated by $B^{-1} \sum_{b=1}^{B} U_b$. Combining yields the approximation to $R(\hat{d}_{\text{opt}})$.

Calculation for $K = 2$. The developments are analogous to those above. We consider the generative data model in Section 3.1.1 and treatment regimes of the form $d = (d_1, d_2)$, where $d_1(s_1) = I(\psi_{10} + \psi_{11}s_1 > 0)$ and $d_2(s_1, s_2, a_1) = I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0)$ for arbitrary $\psi_{10}$, $\psi_{11}$, $\psi_{20}$, $\psi_{21}$, $\psi_{22}$. Here, $E\{Y^*(d)\} = E\left(E\{Y^*(d)|S_2(d)|S_1\}|S_1\right) = E\left(E\left(E\{Y|S_2, S_1, A_1 = d_1(S_1), A_2 = d_2(S_2, S_1, d_1(S_1))\}|S_1, A_1 = d_1(S_1)\right)\right)$. Because $S_1$
is binary taking values in \( \{0, 1\} \),

\[
E\{Y^*(d)\} = E\left(E[Y|S_2, S_1, A_1 = d_1(0), A_2 = d_2(S_2, 0, d_1(0)) \mid S_1 = 0, A_1 = d_1(0)\right) \\
\times \Pr(S_1 = 0) \\
+ E\left(E[Y|S_2, S_1, A_1 = d_1(1), A_2 = d_2(S_2, 1, d_1(1)) \mid S_1 = 1, A_1 = d_1(1)\right) \\
\times \Pr(S_1 = 1).
\]

Under the generative model, writing \( a_1 = I(\psi_{10} + \psi_{11}s_1 > 0) \) for brevity, these expectations are of the form

\[
E\left(E[Y|S_2, S_1, A_1 = d_1(s_1), A_2 = d_2(S_2, s_1, d_1(s_1)) \mid S_1 = s_1, A_1 = d_1(s_1)\right) \\
= \beta_{20} + \beta_{21}s_1 + \beta_{22}a_1 + \beta_{23}s_1a_1 + \beta_{24}E\{(S_2|S_1 = s_1, A_1 = d_1(s_1)) \\
+ \beta_{25}E\{S_2|S_1 = s_1, A_1 = d_1(s_1)\} \\
+ (\psi_{20}^0 + \psi_{21}^0a_1)E\{I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\} \\
+ \psi_{22}^0E\{S_2I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\},
\]

for \( s_1 = 0, 1 \). In the generative data model, the conditional distribution of \( S_2 \) given \( S_1, A_1 \) is normal; accordingly, it is straightforward to calculate \( E\{S_2|S_1 = s_1, A_1 = d_1(s_1)\} \), \( E\{S_2^2|S_1 = s_1, A_1 = d_1(s_1)\} \), \( E\{I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\} \), and \( E\{S_2I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\} \) in a manner analogous to those for the case \( K = 1 \). Approximation of \( E\{H(\hat{d}^\text{opt})\} \) and hence \( R(\hat{d}^\text{opt}) \) for \( \hat{d}^\text{opt} = \hat{d}^\text{opt}_Q \) or \( \hat{d}^\text{opt}_A \) may then be carried out as for the case \( K = 1 \).

**Calculation by simulation.** When an analytical expression for \( H(d) = E\{Y^*(d)\} \) for regimes of a certain form \( d \) is not available, \( H(d) \) for a fixed \( d \) may be approximated by simulation using the g-computation algorithm of Robins (1986). We demonstrate for \( K = 2 \), so that \( d = (d_1, d_2) \); the procedure for \( K = 1 \) is then immediate. For total number of simulations \( B \), for each \( b = 1, \ldots, B \), the steps are: (i) Generate \( s_{1b} \) from the true distribution of \( S_1 \); (ii) generate \( s_{2b} \) from the true conditional distribution of \( S_2 \) given \( S_1 = s_{1b} \) and \( A_1 = d_1(s_{1b}) \); (iii) evaluate the true \( E(Y|\bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2) \) at \( \bar{s}_2 = \bar{s}_{2b} = (s_{1b}, s_{2b}) \) and \( \bar{a}_2 = [d_1(s_{1b}), d_2(\bar{s}_{2b}, d_1(s_{1b}))] \), and call the resulting value \( U_b \); and (iv) estimate \( H(d) = E\{Y^*(d)\} \) by \( B^{-1} \sum_{b=1}^B U_b \). When \( d = \hat{d}^\text{opt}_Q \) or \( \hat{d}^\text{opt}_A \), one would follow the above procedure for each Monte Carlo data set. In each of steps (i)–(iii), it
is important to recognize that, while \( \hat{d}^\text{opt}_Q \) and \( \hat{d}^\text{opt}_A \) are determined by the estimated \( \psi \), the distributions from which realizations are generated depend on the true \( \beta \) and \( \psi \). The values of \( E\{H(\hat{d}^\text{opt}_Q)\} \) and \( E\{H(\hat{d}^\text{opt}_A)\} \) may then be approximated by the average of the estimated \( H(\hat{d}^\text{opt}_Q) \) and \( H(\hat{d}^\text{opt}_Q) \) across the Monte Carlo data sets, as before.
Creating “Equivalently Misspecified Pairs”
When Both the Propensity Model and $Q$-function are Misspecified

Consider the $K = 2$ decision point scenario; the developments apply equally to the $K = 1$ setting. To identify pairs $(\beta_{25}^0, \phi_{25}^0)$ that are “equivalently misspecified,” for each of the combinations of $\beta_{25}^0$ and $\phi_{25}^0$ within a pre-specified grid, say $(\beta_{25}^0, \phi_{25}^0) \in [-1, 1] \times [-1, 1]$ with a step size of 0.05, we generate a large data set of size $n = 10,000$ from the generative data model in Section 3.1.1 with all other parameters fixed at their true values. This yields $41 \times 41 = 1681$ combinations and hence such data sets. For each data set, the linear regression model for the response and the logistic model for propensity of treatment assignment are then fitted, and the ratio of standard errors for $\hat{\phi}_{25}$ and $\hat{\beta}_{25}$, $\text{SE}(\hat{\phi}_{25})/\text{SE}(\hat{\beta}_{25})$, say, obtained. We then fit to these values a polynomial model in $\phi_{25}^0$, $f(\phi_{25}^0)$, say, and select the polynomial degree yielding a sufficiently large adjusted $R^2$. Setting $\beta_{25}^0 = \phi_{25}^0/f(\phi_{25}^0)$ then yields the result that the corresponding t-statistics will be approximately equal. These were re-checked in the course of running the simulations so that the t-statistics differed by less than some reasonable value, usually at most a 5 percent difference, as it cannot be guaranteed that they will be precisely the same.
Derivation of $h_1^0(s_1; \beta_1^0)$ and $C_1^0(s_1; \psi_1^0)$ in the Two Decision Point Scenario

We seek to identify the true $h_1^0(s_1)$ and $C_1^0(s_1)$, where $S_1$ and $A_1$ are Bernoulli. With $h_1^0(s_1) = \beta_{10}^0 + \beta_{11}^0 s_1$ and $C_1^0(s_1) = \psi_{10}^0 + \psi_{11}^0 s_1$, it follows that the true $Q$-function at the first decision is $Q_1^0(s_1, a_1) = h_1^0(s_1) + a_1 C_1^0(s_1)$. We thus calculate $Q_1^0(s_1, a_1)$ under the generative model and equate terms to determine the form of $\beta_{10}^0$, $\beta_{11}^0$, $\psi_{10}^0$, and $\psi_{11}^0$. The true value function at the second decision is $V_2^0(S_1, S_2, A_1) = h_2^0(S_1, S_2, A_1) + C_2^0(S_1, S_2, A_1) I\{C_2^0(S_1, S_2, A_1) > 0\}$. Thus, $Q_1^0(s_1, a_1) = E\{V_2^0(S_1, S_2, A_1) | S_1 = s_1, A_1 = a_1\} = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 E\{S_2 | S_1 = s_1, A_1 = a_1\} + \beta_{25}^0 E\{S_2^2 | S_1 = s_1, A_1 = a_1\} + E\{C_2^0(S_1, S_2, A_1) I\{C_2^0(S_1, S_2, A_1) > 0\} | S_1 = s_1, A_1 = a_1\}$. The conditional expectations in this expression may be calculated in a manner analogous to that in Section E to obtain the form of $Q_1^0(s_1, a_1)$. It follows that $Q_1^0(0, 0) = \beta_{10}^0$, $Q_1^0(1, 0) = \beta_{10}^0 + \beta_{11}^0$, $Q_1^0(0, 1) = \beta_{10}^0 + \psi_{10}^0$, and $Q_1^0(1, 1) = \beta_{10}^0 + \beta_{11}^0 + \psi_{10}^0 + \psi_{11}^0$, which may be solved to yield expressions for $\beta_{10}^0$, $\beta_{11}^0$, $\psi_{10}^0$, and $\psi_{11}^0$. 
Appendix I

Inference for Parameter Estimates

We wish to derive the distribution of parameter estimates $\xi_k$ given the sequence of posited models $Q_k(s_k, a_k; \xi_k)$ and the corresponding $h_k(s_k, a_{k-1}; \beta_k)$, $C_k(s_k, a_{k-1}; \psi_k)$, $\xi_k = (\beta_k, \psi_k)$, under $Q$-learning and $A$-learning, respectively. For these purposes, we assume certain data generating classes, avoiding the non-regularity problem recognized by Robins (2004) and others.

In brief, letting $Z_i$, $i = 1, \ldots, n$ be i.i.d. random vectors, we will show that $Q$- and $A$-learning each involve solving a system of estimating equations of the form $\sum_{i=1}^{n} m(Z_i; \theta) = 0$ where $m(\cdot)$ is a function of dimension $q$, the same dimension of the parameter $\theta$, and $E[m(Z; \theta)] = 0$. It follows, under the class of M-estimation, that, if $\theta_0$ denotes the truth, $\hat{\theta}$ is consistent for $\theta_0$ and $n(-1/2)(\hat{\theta} - \theta_0)$ converges in distribution to Normal with mean zero and variance $\Sigma = [-E\{\partial m(Z; \theta)/\partial \theta^T\}]^{-1}E\{m(Z; \theta_0)m^T(Z; \theta_0)\}[-E\{\partial m(Z; \theta_0)/\partial \theta^T\}]^{-1, T}$, which can be estimated with obvious choices detailed in the sequel.

We illustrate the $K = 2$ scenario, though the general concept applies to all $K \geq 2$. Suppose that at each decision there are two feasible treatment options coded as 0 and 1; i.e., $\Psi_1(s_1) = A_1 = \{0, 1\}$ for all $s_1$ and $\Psi_2(\bar{s}_2, a_1) = A_2 = \{0, 1\}$ for all $\bar{s}_2$ and $a_1 \in \{0, 1\}$. We adopt linear models for the $Q$-functions, a standard approach to many modeling situations; accordingly, consider the models

\begin{align}
Q_1(s_1, a_1; \xi_1) &= \mathcal{H}_{1,x}^T \beta_1 + a_1 \mathcal{H}_{1,c}^T \psi_1, \\
Q_2(\bar{s}_2, a_2; \xi_2) &= \mathcal{H}_{2,x}^T \beta_2 + a_2 \mathcal{H}_{2,c}^T \psi_2,
\end{align}

and the corresponding $h_1(s_1; \beta_1) = \mathcal{H}_{1,x}^T \beta_1$, $C_1(s_1; \psi_1) = \mathcal{H}_{1,c}^T \psi_1$, $h_2(\bar{s}_2, a_1; \beta_2) = \mathcal{H}_{2,x}^T \beta_2$. 

98
and \( C_2(s_2, a_1; \psi_2) = \mathcal{H}_{2,c}^T \psi_2 \) for \( A \)-learning. \( \mathcal{H}_{1,x} \) and \( \mathcal{H}_{1,c} \) are vectors that may contain an intercept and \( s_1 \), say \( \mathcal{H}_{1,x} = (1, s_1)^T; \mathcal{H}_{2,x} \) and \( \mathcal{H}_{2,c} \) may contain an intercept, \( s_1, a_1, s_2 \) and appropriate interactions, say \( \mathcal{H}_{2,x} = (1, s_1, a_1, s_2)^T \). In the following, we consider the general formulation. The subscripts \( x \) and \( c \) denote the vector applicable to the non-contrast portion of the \( Q \)-function, \( h_k(s_k, \bar{a}_{k-1}; \beta_k) \) and contrast portion, \( C_k(s_k, \bar{a}_{k-1}; \psi_k), k = 1, 2, \) respectively. Further, call \( \mathcal{B}_k = (\mathcal{H}_{k,x}^T, a_k \mathcal{H}_{k,c}^T)^T, k = 1, 2 \).

In the following, we describe estimation under \( Q \)-learning and \( A \)-learning. For \( Q \)-learning, we suppose a homoscedastic variance at each decision point. For \( A \)-learning, we take the common choices of \( \lambda_k(s_k, \bar{a}_{k-1}; \psi_k) = \partial / \partial \psi_k C_k(s_k, \bar{a}_{k-1}; \psi_k) \) and \( \theta_k(s_k, \bar{a}_{k-1}) = h_k(s_k, \bar{a}_{k-1}) \) for \( k = 1, 2 \).

### I.1 Q-learning

Given Equation 1.23 and 1.24, it is straightforward to set up the appropriate joint estimating equations as follows

\[
\sum_{i=1}^{n} \frac{\partial Q_2(S_{2i}, \bar{A}_{2i}; \xi_2)}{\partial \xi_2} \{ \tilde{V}_{3i} - Q_2(S_{2i}, \bar{A}_{2i}; \xi_2) \} = 0, \quad (I.3)
\]

\[
\sum_{i=1}^{n} \frac{\partial Q_1(S_{1i}, A_{1i}; \xi_1)}{\partial \xi_1} \{ \tilde{V}_{2i} - Q_1(S_{1i}, A_{1i}; \xi_1) \} = 0, \quad (I.4)
\]

where \( \tilde{V}_{3i} = Y_i \) and \( \tilde{V}_{2i} = \max_{a_2 \in \psi_2(S_{2i}, A_{1i})} Q_2(S_{2i}, A_{1i}, a_2; \xi_2) \). If the \( Q \)-functions are correctly specified, we see that this is an unbiased estimating equation in that \( E[\tilde{V}_{(k+1)} - Q_2(S_k, \bar{A}_k; \xi_k)] = 0 \) and the results of consistency and asymptotic normality follow.

Under the \( Q \)-functions proposed in Equations I.1 and I.2, we rewrite Equations I.3 and I.4 as

\[
\sum_{i=1}^{n} \mathcal{B}_{2i} \{ Y_i - \mathcal{H}_{2i,x}^T \beta_2 - A_{2i} (\mathcal{H}_{2i,c}^T \psi_2) \} = 0, \quad (I.5)
\]

\[
\sum_{i=1}^{n} \mathcal{B}_{1i} \{ \tilde{V}_{2i} - \mathcal{H}_{1i,x}^T \beta_1 - A_{1i} (\mathcal{H}_{1i,c}^T \psi_1) \} = 0, \quad (I.6)
\]
Thus, we see that in the submatrix

\[ \hat{\xi}_2 = (\hat{B}_1^T \hat{B}_2)^{-1} \hat{B}_2 Y, \]

\[ \hat{\xi}_1 = (\hat{B}_1^T \hat{B}_1)^{-1} \hat{B}_1^T \bar{V}_2, \]

Let \( H_{k,x} = (H_{k1,x}, H_{k2,x}, \ldots, H_{kn,x})^T \), \( H_{k,c} = (H_{k1,c}, H_{k2,c}, \ldots, H_{kn,c})^T \), and let \( B_k = (B_{k1}, B_{k2}, \ldots, B_{kn})^T, k = 1, 2 \) be a “design matrix”. It then follows from Equations I.5 and I.6 that estimates \( \hat{\xi}_k = (\hat{\beta}_k^T, \hat{\psi}_k^T)^T \) are found by

\[ \hat{\xi}_2 = (\hat{B}_1^T \hat{B}_2)^{-1} \hat{B}_2 Y, \]

\[ \hat{\xi}_1 = (\hat{B}_1^T \hat{B}_1)^{-1} \hat{B}_1^T \bar{V}_2, \]

where \( Y = (Y_1, Y_2, \ldots, Y_n)^T \) and \( \bar{V}_2 = (\bar{V}_{21}, \bar{V}_{22}, \ldots, \bar{V}_{2n})^T \).

Now, we want to obtain a robust variance estimate \( \hat{\Sigma} \) as noted above, where Equations I.5 and I.6 are jointly considered \( \sum_{i=1}^n m_Q(B_{1i}, B_{2i}; \xi_1, \xi_2) = 0 \). First, we consider the obvious estimator for \( -E\{\partial m_Q(B_1, B_2; \xi_1^0, \xi_2^0)/\partial(\xi_1^T, \xi_2^T)\} \) as the sample mean

\[ -n^{-1} \sum_{i=1}^n \frac{\partial m_Q(B_{1i}, B_{2i}; \hat{\xi}_1, \hat{\xi}_2)}{\partial(\xi_1^T, \xi_2^T)} = -n^{-1} \begin{bmatrix} (\hat{B}_2^T \hat{B}_2) & 0 \\ -A_Q & (\hat{B}_1^T \hat{B}_1) \end{bmatrix}, \]

where (I.8)

\[ A_Q = \hat{B}_1^T \begin{bmatrix} H_{2,x} & H_{2,c} \end{bmatrix}, \]

and where \( H_{2,c} = (I\{H_{21,c} \hat{\psi}_2 > 0\} H_{21,c}, I\{H_{22,c} \hat{\psi}_2 > 0\} H_{22,c}, \ldots, I\{H_{2n,c} \hat{\psi}_2 > 0\} H_{2n,c})^T \). Thus, we see that in the submatrix \( A_Q, \hat{B}_1^T H_{2,c} \) is a summation over individuals \( i = 1, \ldots, n \), such that some individuals contribute zero and the others contribute \( \hat{B}_1^T H_{2,c} \) depending on the individual’s \( \bar{V}_{2i} \) in Equation I.7.

Now, let \( B_{ki} = \{\bar{V}_{(k+1)i} - Q_k(\bar{S}_{ki}, \bar{A}_{ki}; \hat{\xi}_k)\} H_{ki} \) and \( B_{ki}^r = (B_{k1}^r, B_{k2}^r, \ldots, B_{kn}^r)^T \). Then, consider the obvious estimator for \( E\{m_Q(B_1, B_2; \xi_1^0, \xi_2^0) m_Q(B_1, B_2; \xi_1^0, \xi_2^0)\} \) as the sample mean

\[ n^{-1} \sum_{i=1}^n m_Q(B_{1i}, B_{2i}; \hat{\xi}_1, \hat{\xi}_2) m_Q(B_{1i}, B_{2i}; \hat{\xi}_1, \hat{\xi}_2) = n^{-1} \begin{bmatrix} (B_2^r)^T B_2^r & (B_2^r)^T B_1^r \\ (B_1^r)^T B_2^r & (B_1^r)^T B_1^r \end{bmatrix}. \]

(1.9)

From here, it is straightforward to obtain \( \hat{\Sigma} \) as the robust variance estimate matrix using
the results of I.8 and I.9,

\[ \hat{\Sigma}_Q = \left[ -n^{-1} \sum_{i=1}^{n} \frac{\partial m_Q(B_{1i}, B_{2i}; \hat{\xi}_1, \hat{\xi}_2)}{\partial (\xi_1^T, \xi_2^T)} \right]^{-1} \]

\[ \times \left\{ -n^{-1} \sum_{i=1}^{n} m_Q(B_{1i}, B_{2i}; \hat{\xi}_1, \hat{\xi}_2) m_Q(B_{1i}, B_{2i}; \hat{\xi}_1, \hat{\xi}_2) \right\} \]

\[ \times \left[ -n^{-1} \sum_{i=1}^{n} \frac{\partial m_Q(B_{1i}, B_{2i}; \hat{\xi}_1, \hat{\xi}_2)}{\partial (\xi_1^T, \xi_2^T)} \right]^{-1.T} \]

### I.2 A-learning

Using the same framework established in the prior section for Q-learning, we now illustrate estimation of parameters and estimation of the variance under A-learning. Let \( \pi_k(s_k, a_{k-1}; \phi_k) = \text{pr}(A_k = 1 | S_k = s_k, A_{k-1} = a_{k-1}; \phi_k) \) be the propensity of receiving treatment 1 in the observed data as a function of past history. Then, Equation 1.28, along with the conditions set forth in the introduction to this appendix, yield the jointly solved estimating equations

\[ \sum_{i=1}^{n} \frac{\partial h_2(S_{2i}, A_{1i}; \beta_2)}{\partial \beta_2} \{ \tilde{V}_{3i} - C_2(S_{2i}, A_{1i}; \psi_2) - h_2(S_{2i}, A_{1i}; \beta_2) \} = 0, \quad (\text{I.10}) \]

\[ \sum_{i=1}^{n} \frac{\partial C_2(S_{2i}, A_{1i}; \psi_2)}{\partial \psi_2} \{ A_{2i} - \pi_2(S_{2i}, A_{1i}; \phi_2) \} \]

\[ \times \{ \tilde{V}_{3i} - C_2(S_{2i}, A_{1i}; \psi_2) - h_2(S_{2i}, A_{1i}; \beta_2) \} = 0, \quad (\text{I.11}) \]

\[ \sum_{i=1}^{n} \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1} \{ \tilde{V}_{2i} - C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1) \} = 0, \quad (\text{I.12}) \]

\[ \sum_{i=1}^{n} \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} \{ A_{1i} - \pi_1(S_{1i}; \phi_1) \} \]

\[ \times \{ \tilde{V}_{2i} - C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1) \} = 0, \quad (\text{I.13}) \]

where \( \tilde{V}_{3i} = Y_i \) and \( \tilde{V}_{2i} = Y_i + C_2(S_{2i}, A_{1i}; \hat{\psi}_2) [I\{C_2(S_{2i}, A_{1i}; \hat{\psi}_2) > 0\} - A_{2i}] \). One would also jointly estimate, if appropriate, parameters from the propensity models using likelihood score equations. For brevity, we do not include this, though it would be straightforward to do; also, in the sequel, we write \( \hat{\pi}_k \) as shorthand for this propensity model.
that may either be known or estimated.

Under the Q-functions, and equivalent contrast and non-contrast functions, proposed in Equations I.1 and I.2, we rewrite Equations I.10 – I.13 as

\[
\sum_{i=1}^{n} \mathcal{H}_{2i,x} \{ Y_i - \mathcal{H}_{2i,x}^T \beta_2 - A_{2i}(\mathcal{H}_{2i,c}^T \psi_2) \} = 0, \tag{I.14}
\]

\[
\sum_{i=1}^{n} \mathcal{H}_{2i,c} \{ A_{2i} - \hat{\pi}_{2i} \} \{ Y_i - \mathcal{H}_{2i,x}^T \beta_2 - A_{2i}(\mathcal{H}_{2i,c}^T \psi_2) \} = 0, \tag{I.15}
\]

\[
\sum_{i=1}^{n} \mathcal{H}_{1i,x} \{ \tilde{V}_{2i} - \mathcal{H}_{1i,x}^T \beta_1 - A_{1i}(\mathcal{H}_{1i,c}^T \psi_1) \} = 0, \tag{I.16}
\]

\[
\sum_{i=1}^{n} \mathcal{H}_{1i,c} \{ A_{1i} - \hat{\pi}_{1i} \} \{ \tilde{V}_{2i} - \mathcal{H}_{1i,x}^T \beta_1 - A_{1i}(\mathcal{H}_{1i,c}^T \psi_1) \} = 0, \tag{I.17}
\]

where

\[
\tilde{V}_{2i} = Y_i + C_2(\bar{S}_{2i}, A_{1i}; \hat{\psi}_2)[I\{ C_2(\bar{S}_{2i}, A_{1i}; \hat{\psi}_2) > 0 \} - A_{2i}]
\]

\[
= Y_i + (\mathcal{H}_{2,c}^T \psi_2)[I\{ \mathcal{H}_{2,c}^T \psi_2 > 0 \} - A_{2i}], \tag{I.18}
\]

Letting \( B_k^A = (\mathcal{H}_{k,x}^T, (a_k - \pi_k)\mathcal{H}_{k,c}^T)^T \), and \( B_k^A = (B_{k1}^A, B_{k2}^A, \ldots, B_{kn}^A)^T \), \( k = 1, 2 \), be an alternative “design matrix” for \( A \)-learning, it then follows from Equations I.14–I.17 that estimates \( \hat{\xi}_k = (\hat{\beta}_k^T, \hat{\psi}_k^T)^T \) are found by

\[
\hat{\xi}_2 = (\hat{\beta}_2^T, \hat{\psi}_2^T)^T = (\mathbf{B}_2^{A,T})^{-1}\mathbf{B}_2^{A,T}Y,
\]

\[
\hat{\xi}_1 = (\hat{\beta}_1^T, \hat{\psi}_1^T)^T = (\mathbf{B}_1^{A,T})^{-1}\mathbf{B}_1^{A,T}\tilde{V}_2,
\]

where \( Y = (Y_1, Y_2, \ldots, Y_n)^T \) and \( \tilde{V}_2 = (\tilde{V}_{21}, \tilde{V}_{22}, \ldots, \tilde{V}_{2n})^T \).

Again, we want to estimate \( \hat{\Sigma} \) under M-estimation as noted above, where Equations I.14 – I.17 are jointly considered \( \sum_{i=1}^{n} m_A(B_{1i}^A, B_{2i}^A; \xi_1, \xi_2) = 0 \). Like in Section I.1, consider the obvious estimator for \( -\mathbb{E}\{ \partial m_A(B_{1i}^A, B_{2i}^A, \xi_1^0, \xi_2^0) / \partial(\xi_1^T, \xi_2^T) \} \) as the sample mean

\[
-n^{-1} \sum_{i=1}^{n} \frac{\partial m_A(B_{1i}^A, B_{2i}^A; \xi_1, \xi_2)}{\partial(\xi_1^T, \xi_2^T)} = -n^{-1} \begin{bmatrix} (-\mathbf{B}_2^{A,T}\mathbf{B}_2) & 0 \\ \mathbf{A}_A & (-\mathbf{B}_1^{A,T}\mathbf{B}_1) \end{bmatrix}, \tag{I.19}
\]

\[
\mathbf{A}_A = \mathbf{B}_1^{A,T}\begin{bmatrix} 0 & \mathbf{H}_{2,c}^{++} \end{bmatrix},
\]

102
and $H_{2,c}^{*} = (\{H_{21,c}, \psi_2 > 0\} - A_{21}[H_{21,c}, I\{H_{22,c}, \psi_2 > 0\} - A_{22}[H_{22,c}, \ldots, I\{H_{2n,c}, \psi_2 > 0\} - A_{2n}[H_{2n,c}]^T]$. Like Q-learning above, we see that in the submatrix $A_A$, $B_1^{A,T}H_{2,c}^{*}$ is a summation over individuals where some individuals contribute zero and the others contribute $B_1^{A,T}H_{2,c}$ depending on whether or not the treatment they received at the second stage was the ideal treatment under $d_{opt}$.

Now, let $B_{ki}^{Ar} = \{\tilde{V}_{(k+1)i} - C_k(S_{ki}, \hat{A}_{(k-1)i}; \hat{\psi}_k) - h_k(S_{ki}, \hat{A}_{(k-1)i}; \hat{\beta}_k)\}B_{ki}^A$ and $B_{k}^{Ar} = \{B_{k1}^{Ar}, B_{k2}^{Ar}, \ldots, B_{kn}^{Ar}\}^T$. It follows that $E\{m_A(B_1^A, B_2^A; \hat{\xi}_1, \hat{\xi}_2)\}m_A^T(B_1^A, B_2^A; \hat{\xi}_1, \hat{\xi}_2) = n^{-1}\begin{bmatrix} (B_2^{Ar,T}B_2^{Ar}) & (B_2^{Ar,T}B_1^{Ar}) \\ (B_1^{Ar,T}B_2^{Ar}) & (B_1^{Ar,T}B_1^{Ar}) \end{bmatrix}$. (I.20)

Finally, once again, it is straightforward to obtain $\hat{\Sigma}$ as the estimated robust variance matrix using the results of I.19 and I.20,

$$
\hat{\Sigma}_A = \left[-n^{-1}\sum_{i=1}^{n} \frac{\partial m_A(B_{1i}^A, B_{2i}^A; \hat{\xi}_1, \hat{\xi}_2)}{\partial (\hat{\xi}_1^T, \hat{\xi}_2^T)}\right]^{-1} \times \left\{n^{-1}\sum_{i=1}^{n} m_A(B_{1i}^A, B_{2i}^A; \hat{\xi}_1, \hat{\xi}_2)m_A^T(B_{1i}^A, B_{2i}^A; \hat{\xi}_1, \hat{\xi}_2)\right\} \times \left[-n^{-1}\sum_{i=1}^{n} \frac{\partial m_A(B_{1i}^A, B_{2i}^A; \hat{\xi}_1, \hat{\xi}_2)}{\partial (\hat{\xi}_1^T, \hat{\xi}_2^T)}\right]^{-1,T}.
$$

103