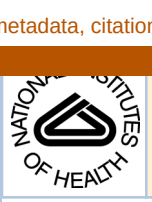

Published in final edited form as:
Ann Stat. 2012 February 1; 40(1): 529–560. doi:10.1214/12-AOS968.

Q-LEARNING WITH CENSORED DATA

Yair Goldberg and **Michael R. Kosorok**
Department of Biostatistics, The University of North Carolina At Chapel Hill, Chapel Hill, NC 27599, U.S.A

Yair Goldberg: ygoldber@bios.unc.edu; Michael R. Kosorok: kosorok@unc.edu

Abstract

We develop methodology for a multistage-decision problem with flexible number of stages in which the rewards are survival times that are subject to censoring. We present a novel Q-learning algorithm that is adjusted for censored data and allows a flexible number of stages. We provide finite sample bounds on the generalization error of the policy learned by the algorithm, and show that when the optimal Q-function belongs to the approximation space, the expected survival time for policies obtained by the algorithm converges to that of the optimal policy. We simulate a multistage clinical trial with flexible number of stages and apply the proposed censored-Q-learning algorithm to find individualized treatment regimens. The methodology presented in this paper has implications in the design of personalized medicine trials in cancer and in other life-threatening diseases.

Keywords and phrases

Q-learning; reinforcement learning; survival analysis; generalization error

1. Introduction

In medical research, *dynamic treatment regimes* are increasingly used to choose effective treatments for individual patients with long-term patient care. A dynamic treatment regime (or similarly, policy) is a set of decision rules for how the treatment should be chosen at each decision time-point, depending on both the patient's medical history up to the current time-point and the previous treatments. Note that although the same set of decision rules are applied to all patients, the choice of treatment at a given time-point may differ, depending on the patient's medical state. Moreover, the patient's treatment plan is not known at the beginning of a dynamic regime, since it may depend on subsequent time-varying variables that may be influenced by earlier treatments and response to treatment. An optimal treatment regime is a set of treatment choices that maximizes the mean response of some clinical outcome at the end of the final time interval (see, for example, Murphy, 2003; Robins, 2004; Moodie et al., 2007).

We consider the problem of finding treatment regimes that lead to longer survival times, where the number of treatments is flexible and where the data is subject to censoring. This type of framework is natural for cancer applications, where the initiation of the next line of therapy depends on the disease progression and thus the number of treatments is flexible. In addition, data are subject to censoring since patients can drop out during the trial. For

SUPPLEMENTARY MATERIAL
Supplement A: Code and data sets
(). Please read the file README.pdf for details on the files in this folder.

example, in advanced non-small cell lung cancer (NSCLC), patients receive one to three treatment lines. The timing of the second and third lines of treatment is determined by the disease progression and by the ability of patients to tolerate therapy (Stinchcombe and Socinski, 2008; Krzakowski et al., 2010). We focus on mean survival time restricted to a specific interval, since in a limited-time study, censoring prevents reliable estimation of the unrestricted mean survival time (see discussion in Karrison, 1997; Zucker, 1998; Chen and Tsiatis, 2001, see also Wahed and Tsiatis, 2006 in the context of sequential decision problems and see Robins et al., 2008 for an alternative approach).

Finding an optimal policy for survival data poses many statistical challenges. We enumerate four. First, one needs to incorporate information accrued over time into the decision rule. Second, one needs to avoid treatments which appear optimal in the short term but may lead to poor final outcome in the long run. Third, the data is subject to censoring since some of the patients may be lost to follow-up and the final outcome of those who reached the end of the study alive is unknown. Fourth, the number of decision points (i.e., treatments) and the timing of these decision points can be different for different patients. This follows since the number of treatments and duration between treatments may depend on the medical condition of the patient. In addition, in the case of patient's death, treatment is stopped. The first two challenges are shared with general multistage decision optimization (Lavori and Dawson, 2004; Moodie et al., 2007). The latter two arise naturally in the context of optimizing survival time, but are applicable to other scenarios as well. Developing valid methodology for estimating dynamic treatment regimes in this flexible timing set-up is crucial for applications in cancer and in other diseases where such structure is the norm and appropriate existing methods are unavailable.

One of the primary tools used in developing dynamic treatment regimes is Q-learning (Murphy et al., 2006; Zhao et al., 2009; Laber et al., 2010; Zhao et al., 2010). Q-learning (Watkins, 1989; Watkins and Dayan, 1992), which is reviewed in Section 2, is a reinforcement learning algorithm. Since we do not assume that the problem is Markovian, we present a version of Q-learning that uses backward recursion. The backward recursion used by Q-learning addresses the first two challenges posed above: It enables both accrual of information and incorporation of long-term treatment effects. However, when the number of stages is flexible, and censoring is introduced, it is not clear how to implement backward recursion. Indeed, finding the optimal treatment at the last stage is not well defined, since the number of stages is patient-dependent. Also, it is not clear how to utilize the information regarding censored patients.

In this paper we present a novel Q-learning algorithm that takes into account the censored nature of the observations using inverse-probability-of-censoring weighting (see, Robins et al., 1994, see also Wahed and Tsiatis, 2006; Robins et al., 2008 in the context of sequential decision problems). We provide finite sample bounds on the generalization error of the policy learned by the algorithm, i.e., bounds on the average difference in expected survival time between the optimal dynamic treatment regime and the dynamic treatment regime obtained by the proposed Q-learning algorithm. We also present a simulation study of a sequential-multiple-assignment randomized trial (SMART) (see Murphy, 2005a, and references therein) with flexible number of stages depending on disease progression and failure event timing. We demonstrate that the censored-Q-learning algorithm proposed here can find treatment strategies tailored to each patient which are better than any fixed-treatment sequence. We also demonstrate the result from ignoring censored observations.

One general contribution of the paper is the development of a methodology for solving backward recursion when the number and timing of stages is flexible. As mentioned previously, this is crucial for applications but has not been addressed previously. In Section

4 we present an auxiliary multistage decision problem that has a fixed number of stages. Since the number of stages is fixed for the auxiliary problem, backward recursion can be used in order to estimate the decision policy. We then show how to translate the original problem to the auxiliary one and obtain the surprising conclusion that results obtained for the auxiliary problem can be translated into results regarding the original problem with flexible number and timing of stages.

An additional contribution of the paper is the universal consistency proof for the algorithm performance. Universal consistency of an algorithm means that for every distribution function on the sample space, the expected loss of the function learned by the algorithm converges in probability to the infimum over all measurable functions of the expected loss (see, for example, Steinwart and Chirstmann, 2008). In Section 6 we prove that when the optimal Q-functions belong to the corresponding approximation spaces considered by the algorithm, the algorithm is universally consistent. The proof presented here is algorithm-specific, but the tools used in the proof are widely applicable for universal consistency proofs when the data are subject to censoring (see, for example, Goldberg et al., 2011). While other learning algorithms were suggested for survival data (see, for example, Biganzoli et al., 1998; Shivaswamy et al., 2007; Shim and Hwang, 2009, see also Zhao et al., 2010 in the context of a multistage decision problem), we are not aware of any other universal consistency proof for survival data.

The paper is organized as follows. In Section 2 we review the Q-learning algorithm and discuss the challenges for adapting the Q-leaning methodology for a framework with flexible number of stages and censored data. We also review existing methods for finding optimal policies. Definitions and notation are presented in Section 3. The auxiliary problem is presented in Section 4. The censored-Q-learning algorithm is presented in Section 5. The main theoretical results are presented in Section 6. In Section 7 we present a multistage-randomized-trial simulation study. Concluding remarks appear in Section 8. Supplementary proofs are provided in Appendix A. A description of and link to the code and data sets used in Section 7 appear in Supplement A

2. Q-learning

2.1. Reinforcement Learning

Reinforcement learning is a methodology for solving multistage decision problems. It involves recording sequences of actions, statistically estimating the relationship between these actions and their consequences and then choosing a policy (i.e., a set of decision rules) that approximates the most desirable consequence based on the statistical estimation. A detailed introduction to reinforcement learning can be found in Sutton and Barto (1998).

In the medical context of long-term patient care, the reinforcement learning setting can be described as follows. For each patient, the stages correspond to clinical decision points in the course of the patient's treatment. At these decision points, actions (e.g., treatments) are chosen, and the state of the patient is recorded. As a consequence of a patient's treatment, the patient receives a (random) numerical reward.

More formally, consider a multistage decision problem with T decision points. Let S_t be the (random) state of the patient at stage $t \in \{1, \ldots, T+1\}$ and let $\mathbf{S}_t = \{S_1, \ldots, S_t\}$ be the vector of all states up to and including stage t. Similarly, let A_t be the action chosen in stage t, and let $\mathbf{A}_t = \{A_1, \ldots, A_t\}$ be the vector of all actions up to and including stage t. We use the corresponding lower case to denote a realization of these random variables and random vectors. Let the random reward be denoted $R_t = r(\mathbf{S}_t, \mathbf{A}_t, S_{t+1})$, where r is a (unknown) time-dependent deterministic function of all states up to stage $t+1$ and all past actions up to stage

t. A trajectory is defined as a realization of $(\boldsymbol{S}_{T+1}, \boldsymbol{A}_T, \boldsymbol{R}_T)$. Note that we do not assume that the problem is Markovian. In the medical context example, a trajectory is a record of all the patient covariates at the different decision points, the treatments that were given, and the medical outcome in numerical terms.

We define a policy, or similarly, a dynamic treatment regime, to be a set of decision rules. More formally, define a policy $\boldsymbol{\pi}$ to be a sequence of deterministic decision rules, $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T\}$, where for every pair $(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, the output of the *t*-th decision rule, $\boldsymbol{\pi}_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, is an action. Our goal is to find a policy that maximizes the expected sum of rewards. The Bellman equation (Bellman, 1957) characterizes the optimal policy $\boldsymbol{\pi}^*$ as one that satisfies the following recursive relation:

$$\pi_t^*(\boldsymbol{s}_t, \boldsymbol{a}_{t-1}) = \underset{a_t}{\operatorname{argmax}}\, E\left[R_t + V_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t) \middle| \boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{A}_t = \boldsymbol{a}_t\right], \tag{1}$$

where the value function

$$V_{t+1}^*(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t) = E_{\pi^*}\left[\sum_{i=t+1}^{T} R_i \middle| \boldsymbol{S}_{t+1} = \boldsymbol{s}_{t+1}, \boldsymbol{A}_t = \boldsymbol{a}_t\right] \tag{2}$$

is the expected cumulative sum of rewards from stage $t+1$ to stage T, where the history up to stage $t+1$ is given by $\{\boldsymbol{s}_{t+1}, \boldsymbol{a}_t\}$, and when using the optimal policy $\boldsymbol{\pi}^*$ thereafter.

Finding a policy that leads to a high expected cumulative reward is the main goal of reinforcement learning. Naively, one could learn the transition distribution functions and the reward function using the observed trajectories, and then solve the Bellman equation recursively. However, this approach is inefficient both computationally and memory-wise. In the following section, we introduce the Q-learning algorithm, which requires less memory and less computation.

2.2. Q-learning

Q-learning (Watkins, 1989) is an algorithm for solving reinforcement learning problems. It is claimed by Sutton and Barto to be one of the most important breakthroughs in reinforcement learning (Sutton and Barto, 1998, Section 6.5). Q-learning uses backward recursion to compute the Bellman equation without the need to know the full dynamics of the process.

More formally, we define the optimal time-dependent Q-function

$$Q_t^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = E\left[R_t + V_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t) | \boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{A}_t = \boldsymbol{a}_t\right].$$

Note that $V_t^*(\boldsymbol{s}_t, \boldsymbol{a}_{t-1}) = \max_{a_t} Q_t^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$, and thus

$$Q_t^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = E\left[R_t + \max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) \middle| \boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{A}_t = \boldsymbol{a}_t\right]. \tag{3}$$

In order to estimate the optimal policy, one first estimates the Q-functions backwards through time $t = T, T-1, \ldots, 1$ and obtains a sequence of estimators $\{\hat{Q}_T \ldots, \hat{Q}_1\}$. The estimated policy is given by

$$\widehat{\pi}_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1}) = \underset{a_t}{\operatorname{argmax}}\, \widehat{Q}_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1}, a_t). \tag{4}$$

In the next section we discuss the difficulties in applying the Q-learning methodology when trajectories are subject to censoring and the number of stages is flexible.

2.3. Challenges With Flexible Number of Stages and Censoring

As discussed in the introduction, our goal is to develop a Q-learning algorithm that can handle a flexible number of stages and that takes into account the censored nature of the observations. We face two main challenges. First, recall that the estimation of the Q-functions in (3) is done recursively, starting from the last stage backward. Thus, when the number of stages is flexible, it is not clear how to perform the base step of the recursion. Second, due to censoring, some of the trajectories may be incomplete. Incorporating the data of a censored trajectory is problematic: even when the number of stages is fixed, the known number of stages for a censored trajectory may be less than the number of stages in the multistage problem. Moreover, the reward is not known for the stage at which censoring occurs.

2.4. Review of Existing Approaches

Finding optimal policies or optimal treatment regimes has been discussed extensively in other work. We discuss shortly some additional work that is related to the approach taken here. However, we are not aware of any other existing approaches that address simultaneously both censoring and flexible number of stages.

The approach closest to our proposal is the censored-Q-learning algorithm of Zhao et al. (2010). Zhao et al. considered a Q-learning algorithm for censored data based on support vector regression adjusted for censoring with fixed number of stages. A simulation study was performed to demonstrate the algorithm's performance; however, the theoretical properties of this algorithm were not evaluated.

A general approach for finding optimal polices that uses backward recursion was studied by Murphy (2003) and Robins (2004) in the semiparametric context, and by Murphy (2005b) in the nonparametric context. These works do not treat flexible number of stages or censoring, and cannot be applied to the framework considered here without some adjustments.

Another approach for finding optimal policies was studied by Orellana et al. (2010), (see also Laan and Petersen, 2007; Robins et al., 2008). Orellana et al. consider dynamic regime marginal structural mean models (Robins, 1999). In this approach, for each regime, one considers all trajectories that comply to the regime up to some point. The trajectories are then censored at the first time-point at which they do not comply to the regime. The contribution of the non-compliant trajectories is redistributed among compliant trajectories that have the same covariate and treatment history, using the inverse-probability-of-censoring weighting. Advantages and disadvantages of this approach compared to the backward recursion approach mentioned above are discussed in Robins et al. (2008, Section 5). We note that it is assumed in their approach that the length of each stage is fixed, an assumption we do not require.

This general issue is also related to the analysis of two-stage randomized trials involving right-censored data studied in a series of papers including Lunceford et al. (2002); Wahed and Tsiatis (2006); Wahed (2009); Miyahara and Wahed (2010). The authors use inverse-probability-of-censoring to correct for censoring. See also Thall et al. (2007) that considers analysis of two-stage randomized trials with interval censoring. However, the main focus of

these works is in finding the best regime from a finite number of optional regimes, as opposed to the individualized-treatment policies addressed in our proposal.

3. Preliminaries

In this section we present definitions and notation which will be used in the paper.

Let T be the maximal number of decision time-points for a given multistage time-dependent decision problem. Note that the number of stages for different observations can be different. For each $t = 1,\ldots,T$, the state S_t is the pair $S_t = (Z_t, R_{t-1})$, where Z_t is either a vector of covariates describing the state of the patient at beginning of stage t or $Z_t = \emptyset$. $Z_t = \emptyset$ indicates that a failure event happened during the t-th stage which has therefore reached a terminal state. R_{t-1} is the length of the interval between decision time-points $t-1$ and t, where we denote $R_0 \equiv 0$. Although in the usual Q-learning context $\sum_{j=1}^{t} R_j$ is the sum of rewards up to and including stage t, in our context it is more useful to think of this sum as the total survival time up to and include stage t. Let A_t be an action chosen at decision time t, where A_t takes its values in a finite discrete space $\mathcal{A}$.

The model assumes that observations are subject to censoring. Let C be a censoring variable and let $S_C(x) = P(C \geq x)$ be its survival function. We assume that censoring is independent of both covariates and failure time. We assume that C takes its values in the segment $[0, \tau]$ where $\tau < \infty$ and that $S_C(\tau) > K_{\min} > 0$. Let δ_t be an indicator with $\delta_t = 1$ if no censoring event happened before the $t+1$-th decision time-point. Note that $\delta_{t-1} = 0 \Rightarrow \delta_t = 0$.

Remark 3.1

Note that for a censoring variable, we define the survival function $S_C(x)$ as P $(C \geq x)$ rather than the usual $P(C > x)$. This is because given a failure time x, we are interested in the probability $P(C \geq x)$. However, to avoid complications that are not of interest to the main results of this paper, we assume that the probability of simultaneous failure and censoring is zero (see, for example, Satten and Datta, 2001).

The inclusion of failure times in the model affects the trajectory structure. Usually, a trajectory is defined as a $(2T+1)$-length sequence $\{S_1, A_1, S_2, \ldots, A_T, S_{T+1}\}$. However, in our context, if a failure event occurs before decision-time-point T, the trajectory will not be of full length. Denote by $\bar{T}$ the (random) number of stages for the individual ($\bar{T} \leq T$). Due to the censoring, the trajectories themselves are not necessarily fully observed. Assume that a censoring event occurred during stage t. Note that this means that $\delta_{t-1} = 1$ while $\delta_t = 0$ and that $C < \sum_{i=1}^{t} R_i$. In this case the observed trajectories have the following structure: $\{S_1, A_1, S_2, \ldots, A_t\}$ and C is also observed.

We now discuss the distribution of the observed trajectories. Assume that n trajectories are sampled at random according to a fixed distribution de-noted by P_0. The distribution P_0 is composed of the unknown distribution of each S_t conditional on $(\boldsymbol{S}_{t-1}, \boldsymbol{A}_{t-1})$ (denoted by $\{f_1, \ldots, f_T\}$) and an exploration policy that generates the actions. Denote the exploration policy by $\boldsymbol{p} = \{p_1, \ldots, p_T\}$ where the probability that action a is taken given history $\{\boldsymbol{S}_t, \boldsymbol{A}_{t-1}\}$ is $p_t(a|\boldsymbol{S}_t, \boldsymbol{A}_{t-1})$. We assume that $p_t(a|\boldsymbol{s}_t, \boldsymbol{a}_{t-1}) \geq L^{-1}$ for every action $a \in \mathcal{A}$ and for each possible value $(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, where $L \geq 1$ is a constant. The likelihood (under P_0) of the trajectory $\{s_1, a_1, s_2, \ldots, a_t, s_{t+1}\}$ is

$$f_1(s_1)p_1(a_1|s_1)\prod_{j=2}^{\bar{t}}(f_j(s_j|\boldsymbol{s}_{j-1},\boldsymbol{a}_{j-1})p_j(a_j|\boldsymbol{s}_j,\boldsymbol{a}_{j-1}))f_{\bar{t}+1}(s_{t+1}|\boldsymbol{s}_{\bar{t}},\boldsymbol{a}_{\bar{t}}).$$

We denote expectations with respect to the distribution P_0 by E_0. The survival time with respect to the distribution P_0 is denoted by $P_0(\sum_{j=1}^{\overline{T}} R_j > x)$. We assume that $G(\tau) > G_{\min} > 0$, i.e., that there is a positive probability that the survival time is greater than τ.

We define policy π to be a sequence of deterministic decision rules, $\{\pi_1, \ldots, \pi_T\}$, where for every non-terminating pair $(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, the output of the t-th decision rule, $\pi_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, is an action. Let the distribution $P_{0,\pi}$ denote the distribution of a trajectory for which the policy π is used to generate the actions. The likelihood (under $P_{0,\pi}$) of the trajectory, $\{s_1, a_1, s_2, \ldots, a_t, s_{\bar{t}+1}\}$ is

$$f_1(s_1)1_{\pi(s_1)=a_1}\prod_{j=2}^{\bar{t}}(f_j(s_j|\boldsymbol{s}_{j-1},\boldsymbol{a}_{j-1})1_{\pi_j(\boldsymbol{s}_j,\boldsymbol{a}_{j-1})=a_j}))f_{\bar{t}+1}(s_{\bar{t}+1}|\boldsymbol{s}_{\bar{t}},\boldsymbol{a}_{\bar{t}}).$$

Our goal is to find a policy that maximizes the expected rewards. Since with probability one $C \leq \tau$, the maximum observed survival time is less than or equal to τ. Thus we try to maximize the truncated-by-τ expected survival time. Formally, we look for a policy $\hat{\pi}$ that approximates the maximum over all deterministic policies of the following expectation:

$$E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right],$$

where $E_{0,\pi}$ is the expectation with respect to $P_{0,\pi}$ and $a \wedge b = \min\{a, b\}$.

4. The Auxiliary Problem

In this section we construct an auxiliary Q-learning model for our original problem. The modified trajectories of the construction are of fixed length T, and the modified sum of rewards is less than or equal to τ. We then show how results obtained for the auxiliary problem can be translated into results regarding the original problem.

For the auxiliary problem, we complete all trajectories to full length in the following way. Assume that a failure time occurred at stage $t < T$. In that case the trajectory up to S_{t+1} is already defined. Write $S'_j = S_j$ for $1 \leq j \leq t+1$ and $A'_j = A_j$ for $1 \leq j \leq t$. For all $t+1 < j \leq T+1$ set $S_j = (\varnothing, 0)$ and for all $t+1 \leq j \leq T$ draw A_j uniformly from $\mathcal{A}$.

We also modify trajectories with overall survival time greater than τ in the following way. Assume that t is the first index for which $\sum_{i=1}^{t} R_i \geq \tau$. For all $j \leq t$, write $S'_j = S_j$ and $A'_j = A_j$. Write $R'_t = \tau - \sum_{i=1}^{t-1} R_i$ and assign $Z'_{t+1} \equiv \varnothing$ and thus the modified state $S'_{t+1} = (\varnothing, R'_t)$. If $t < T$, then for all $t+1 < j \leq T+1$ set $S_j = (\varnothing, 0)$ and for all $t+1 \leq j \leq T$ draw A'_j uniformly from

$\mathcal{A}$. The modified trajectory is given by the sequence $\{S_1', A_1', S_2', \ldots, A_T', S_{T+1}'\}$. Note that trajectories with fewer than $2T+1$ entries and for which $\sum_{i=1}^{t} R_i \geq \tau$ are modified twice.

The n modified trajectories are distributed according to the fixed distribution P which can be obtained from P_0. This distribution is composed of the unknown distribution of each S_t' conditional on $(\boldsymbol{S}_{t-1}', \boldsymbol{A}_{t-1}')$, denoted by $\{f_1', \ldots, f_{T+1}'\}$, and exploration policy $\boldsymbol{p}'$. The conditional distribution f_1' equals f_1, and for $2 \leq t \leq T+1$,

$$f_t'(\boldsymbol{s}_t'|\boldsymbol{s}_{t-1}', \boldsymbol{a}_{t-1}') = \begin{cases} f_t((z_t', r_t')|\boldsymbol{s}_{t-1}', \boldsymbol{a}_{t-1}') & z_{t-1}' \neq \varnothing, \sum_{i=1}^{t} r_i' < \tau \\ \int_{G_{z_t'}} f_t((z_t', r_t)|\boldsymbol{s}_{t-1}', \boldsymbol{a}_{t-1}')\mathrm{d}r_t & z_{t-1}' \neq \varnothing, \sum_{i=1}^{t} r_i' = \tau \\ 1_{\boldsymbol{s}_t'=(\varnothing,0)} & z_{t-1}' = \varnothing \end{cases}, \quad (5)$$

where $G_{z_t'} = \{(z_t', r_t): \sum_{i=1}^{t} r_i \geq \tau\}$ and 1_A is 1 if A is true and is 0 otherwise. The exploration policy $\boldsymbol{p}'$ agrees with $\boldsymbol{p}$ on every pair $(\boldsymbol{S}_t, \boldsymbol{A}_{t-1})$ for which $Z_t \neq \varnothing$ and draws A_t uniformly from $\mathcal{A}$ whenever $Z_t = \varnothing$. The likelihood (under P) of the modified trajectory, $\{s_1', a_1', s_2', \ldots, a_T', s_{T+1}'\}$, is

$$f_1'(s_1')p_1(a_1'|s_1')\prod_{t=2}^{T}(f_t'(s_t'|\boldsymbol{s}_{t-1}', \boldsymbol{a}_{t-1}')p_t(a_t'|\boldsymbol{s}_t', \boldsymbol{a}_{t-1}'))f_{T+1}'(s_{T+1}'|\boldsymbol{s}_T', \boldsymbol{a}_T').$$

Denote expectations with respect to the distribution P by E.

Let $\boldsymbol{\pi}$ be a policy for the original problem. We define a version of the policy $\boldsymbol{\pi}'$ for the auxiliary problem in the following way. For any state $(\boldsymbol{s}_t', \boldsymbol{a}_{t-1}')$ for which $z_t' \neq \varnothing$, the same action is chosen. For any state $(\boldsymbol{s}_t', \boldsymbol{a}_{t-1}')$ for which $z_t' = \varnothing$, a fixed action $a_t \in \mathcal{A}$ is chosen, w.o.l.g., let a_o be chosen. For the auxiliary problem, we say that two policies π_a' and π_b' are equivalent if $\pi_a'(\boldsymbol{s}_t', \boldsymbol{a}_{t-1}') = \pi_b'(\boldsymbol{s}_t', \boldsymbol{a}_{t-1}')$ for every $(\boldsymbol{s}_t', \boldsymbol{a}_{t-1}')$ for which $z_t' \neq \varnothing$. We denote both the original policy and any modified version of it by $\boldsymbol{\pi}$ whenever it is clear from the context which policy is considered. Similarly, we omit the prime from states and actions in the auxiliary problem whenever there is no reason for confusion.

Let $P_{\boldsymbol{\pi}}$ be the distribution in the auxiliary problem where actions are chosen according to $\boldsymbol{\pi}$. The likelihood under $P_{\boldsymbol{\pi}}$ of the trajectory $\{s_1, a_1, s_2, \ldots, a_T, s_{T+1}\}$ is

$$f_1'(s_1)1_{\pi_1(s_1)=a_1}\prod_{t=2}^{T}(f_t'(s_t|\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1})1_{\pi_j(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})=a_t}))f_{T+1}'(s_{T+1}|\boldsymbol{s}_T, \boldsymbol{a}_T).$$

Denote expectations with respect to the distribution $P_{\boldsymbol{\pi}}$ by $E_{\boldsymbol{\pi}}$.

We now define the value functions and the Q-functions for policies in the auxiliary model. For any auxiliary policy $\boldsymbol{\pi}$ define its corresponding value function $V_{\boldsymbol{\pi}}$. Given an initial state s_1, $V_{\boldsymbol{\pi}}(s_1)$ is the expected truncated-by-τ survival time when the initial state is s_1 and the

actions are chosen according to the policy $\boldsymbol{\pi}$. Formally $V_\pi(s_1)=E_\pi[\sum_{i=1}^{T} R_t|S_1=s_1]$ where the truncation takes place since the expectation is taken with respect to the distribution of the modified trajectories. The stage-t value function for the auxiliary policy $\boldsymbol{\pi}$, $V_{\boldsymbol{\pi},t}(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, is the expected (truncated) remaining survival time from the t-th decision time-point, given the trajectory $(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, and when following the policy $\boldsymbol{\pi}$ thereafter. Note, that given $\boldsymbol{s}_t$, the survival time up to the beginning of stage t is known, and thus truncation ensures that the overall survival time is less than or equal to $\boldsymbol{\tau}$. Formally

$$V_{\pi,t}(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})=E_\pi[\sum_{i=t}^{T} R_i|\boldsymbol{S}_t=\boldsymbol{s}_t, \boldsymbol{A}_t=\boldsymbol{a}_t].$$

The stage-t Q-function for the auxiliary policy $\boldsymbol{\pi}$ is the expected remaining (truncated) survival time, given that the state is $(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$, that a_t is chosen at stage t, and that $\boldsymbol{\pi}$ is followed thereafter. Formally,

$$Q_{\pi,t}(\boldsymbol{s}_t, \boldsymbol{a}_t)=E[R_t+V_{\pi,t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t)|\boldsymbol{S}_t=\boldsymbol{s}_t, \boldsymbol{A}_t=\boldsymbol{a}_t].$$

The optimal value function $V_t^*(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})$ and the optimal Q-function $Q_t^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$ are define by (2) and (3), respectively.

The following lemma relates the values of the value function $V_{\boldsymbol{\pi}}$ in the auxiliary problem to the expected truncated-by-$\boldsymbol{\tau}$ survival time for a policy $\boldsymbol{\pi}$ in the original problem.

Lemma 4.1

Let Π be the collection of all policies in the original problem. Then for all $\boldsymbol{\pi} \in \Pi$, the following equalities hold true:

$$V_\pi(s_o)=E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau \middle| S_1=s_o\right], \tag{6}$$

$$V^*(s_o)=\max_{\pi\in\Pi} E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau \middle| S_1=s_o\right]. \tag{7}$$

Proof

We start by decomposing the expectations depending on both the terminal stage and whether the sum of rewards is greater than or equal to $\boldsymbol{\tau}$.

Define

$$F_t=\{\{s_o, a_1, \ldots, s_{t+1}\}: \textstyle\sum_{i=1}^{t} r_i<\tau, z_{t+1}=\varnothing\},$$

$$G_t=\{\{s_o, a_1, \ldots, s_{k+1}\}: t=\min\{j: \textstyle\sum_{i=1}^{j} r_i \geq \tau\}, \text{ and } k=T \text{ or } z_{k+1}=\varnothing\},$$

$$F_t'=\{(\boldsymbol{s}_{T+1}', \boldsymbol{a}_T'): (\boldsymbol{s}_{t+1}', \boldsymbol{a}_t') \in F_t, \{a_{t+1}', \ldots, s_{T+1}\}=\{a_o, (\varnothing, 0), \ldots, (\varnothing, 0)\}\},$$

$$G_t'=\left\{ (\boldsymbol{s}_{T+1}', \boldsymbol{a}_T'): \begin{array}{c} (\boldsymbol{s}_t', \boldsymbol{a}_t') \text{ is a beginning of sequence in } G_t, \\ \{s_{t+1}', a_{t+1}', \ldots, s_{T+1}\}=\{(\varnothing, \tau-\sum_{j=1}^{t-1} r_j), a_o, \ldots, (\varnothing, 0)\} \end{array} \right\}.$$

Denote

$$\boldsymbol{f}_{t,\pi}(\boldsymbol{s}_t, \boldsymbol{a}_{t-1}) = f_1(s_1)\,[1_{\pi(s_1)=a_1}] \prod_{j=2}^{t-1} (f_j(s_j|\boldsymbol{s}_{j-1}, \boldsymbol{a}_{j-1}) 1_{\pi_j(\boldsymbol{s}_j, \boldsymbol{a}_{j-1})=a_j}) f_t(s_t|\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1})$$

and similarly $f'_{t,\pi}$.

Note that

$$E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau \middle| S_1 = s_o\right] = \sum_{t=1}^{T} \int_{F_t} \left(\sum_{i=1}^{t} r_i\right) \boldsymbol{f}_{t+1,\pi}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t) \mathrm{d}\,(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t) + \tau \sum_{t=1}^{T} P_{0,\pi}(G_t), \quad (8)$$

and

$$V_\pi(s_o) = \sum_{t=1}^{T} \int_{F'_t} \left(\sum_{i=1}^{T} r_i\right) \boldsymbol{f}'_{T+1,\pi}(\boldsymbol{s}_{T+1}, \boldsymbol{a}_T) \mathrm{d}\,(\boldsymbol{s}_{T+1}, \boldsymbol{a}_T) + \tau \sum_{t=1}^{T} P_\pi(G'_t). \quad (9)$$

Note that

$$\begin{aligned} \int_{F_t} \left(\sum_{i=1}^{t} r_i\right) \boldsymbol{f}_{t+1,\pi}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t) \mathrm{d}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t) &= \int_{F_t} \left(\sum_{i=1}^{t} r_i\right) \boldsymbol{f}'_{t+1,\pi}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t) \mathrm{d}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t) \\ &= \int_{F'_t} \left(\sum_{i=1}^{T} r_i\right) \boldsymbol{f}'_{T+1,\pi}(\boldsymbol{s}_{T+1}, \boldsymbol{a}_T) \mathrm{d}(\boldsymbol{s}_{T+1}, \boldsymbol{a}_T) \end{aligned} \quad (10)$$

where the first equality follows from (5) and the second follows since there is a one to one correspondence between trajectories in F_t and F'_t, and by construction, for each such trajectory in F'_t we have $\sum_{i=t+1}^{T} r_i = 0$ and

$$[1_{\pi_{t+1}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t)=a_o}] \prod_{j=t+2}^{T} (f'_j(s_j|\boldsymbol{s}_{j-1}, \boldsymbol{a}_{j-1}) 1_{\pi_j(\boldsymbol{s}_j, \boldsymbol{a}_{j-1})=a_o})) f_{T+1}(s_{T+1}|\boldsymbol{s}_T, \boldsymbol{a}_T) = 1.$$

Similarly we show that $P_{0,\pi}(G_t) = P_\pi(G'_t)$. Denote by $\hat{G}_t$ the set of all sequences $(\boldsymbol{s}_t, \boldsymbol{a}_t)$ which are the beginning part of some trajectory in G_t. Note that

$$\begin{aligned} P_{0,\pi}(G_t) &= \int_{\hat{G}_t} \boldsymbol{f}_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})\,[1_{\pi_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})=a_t}] \int_{\{s_{t+1}:\sum_{i=1}^{t} r_i \geq \tau\}} f_{t+1}(s_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \mathrm{d}(\boldsymbol{s}_{t+1}) \mathrm{d}\,(\boldsymbol{s}_t, \boldsymbol{a}_t) \\ &= \int_{\hat{G}_t} \boldsymbol{f}'_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})\,[1_{\pi_t(\boldsymbol{s}_t, \boldsymbol{a}_{t-1})=a_t}] \int_{\{s_{t+1}:\sum_{i=1}^{t} r_i \geq \tau\}} f'_{t+1}(s_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t) \mathrm{d}(\boldsymbol{s}_{t+1}) \mathrm{d}\,(\boldsymbol{s}_t, \boldsymbol{a}_t) \\ &= \int_{G'_t} \boldsymbol{f}'_{T+1}(\boldsymbol{s}_{T+1}, \boldsymbol{a}_t) \mathrm{d}(\boldsymbol{s}_{T+1}, \boldsymbol{a}_t) = P_\pi(G'_t), \end{aligned} \quad (11)$$

where the second equality follows from (5) and the third equality follows from the construction of G'_t.

The first assertion of the lemma, namely, equation (6), follows by substituting the right hand side of the equalities (10) and (11) in (8) for each t and comparing to (9).

The second assertion, (7), is proven by maximizing both sides of (6) over all policies. Note that the maximization is taken over two different sets since each policy in the original problem has an equivalent class of polices in the auxiliary problem. However, since V_{π} is the same for all policies in the same equivalence class, the result follows.

5. The Censored-Q-Learning Algorithm

We now present the proposed censored-Q-learning algorithm. As discussed before, we are looking for a policy $\hat{\pi}$ that approximates the maximum over all deterministic policies of the following expectation:

$$E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right].$$

We find this policy in three steps. First, we map our problem to the corresponding auxiliary problem. Then we approximate the functions $\{Q_1^*, \ldots, Q_T^*\}$ using backward recursion based on (3) and obtain the functions $\{\hat{Q}_1, \ldots, \hat{Q}_T\}$. Finally, we define $\hat{\pi}$ by maximizing $\hat{Q}_t(\boldsymbol{s}_t, (\boldsymbol{a}_{t-1}, a_t))$ over all possible actions a_t.

Let $\{\mathcal{Q}_1, \ldots, \mathcal{Q}_T\}$ be the approximation spaces for the Q-functions. We assume that $Q_t(\boldsymbol{s}_t, \boldsymbol{a}_t) = 0$ whenever $z_t = \varnothing$. In other words, if a failure occurred before the t-th-time-point, Q_t equals zero.

Note that by (3), the optimal t-stage Q-function $Q_t^*(\boldsymbol{s}_t, \boldsymbol{a}_t)$ equals the conditional expectation of $R_t + \max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, (\boldsymbol{A}_t, a_{t+1}))$ given $(\boldsymbol{s}_t, \boldsymbol{a}_t)$. Thus

$$Q_t^* = \underset{Q_t}{\operatorname{argmin}}\, E\left[\left(R_t + \max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, (\boldsymbol{A}_t, a_{t+1})) - Q_t(\boldsymbol{S}_t, \boldsymbol{A}_t)\right)^2\right].$$

Ideally, we could compute the functions $\hat{Q}_t$ using backward recursion in the following way:

$$\widehat{Q}_t = \underset{Q_t}{\operatorname{argmin}}\, \mathbb{E}_n\left[\left(R_t + \max_{a_{t+1}} \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, (\boldsymbol{A}_t, a_{t+1})) - Q_t(\boldsymbol{S}_t, \boldsymbol{A}_t)\right)^2\right],$$

where $\mathbb{E}_n$ is the empirical expectation. The problem is that R_t may be censored and thus unknown.

Note that $E[\delta_t | \sum_{i=1}^{t} R_i] = P(C \geq \sum_{i=1}^{t} R_i) = S_C(\sum_{i=1}^{t} R_i)$ and thus

$$E\left[\frac{\delta_t}{S_C(\sum_{i=1}^{t} R_i)} \middle| \boldsymbol{S}_t, \boldsymbol{A}_t, R_t\right] = 1$$

since $\boldsymbol{S}_t$ includes the information regarding $R_1, \ldots, R_{t-1}$ and C is independent of the covariates and actions.

Thus, for every function $Q_t \in \mathcal{Q}_t$,

$$E\left[\left(R_t+\max_{a_{t+1}}Q^*_{t+1}(\boldsymbol{S_{t+1}},(\boldsymbol{A_t},a_{t+1}))-Q_t(\boldsymbol{S_t},\boldsymbol{A_t})\right)^2\right]=E\left[\left(R_t+\max_{a_{t+1}}Q^*_{t+1}(\boldsymbol{S_{t+1}},(\boldsymbol{A_t},a_{t+1}))-Q_t(\boldsymbol{S_t},\boldsymbol{A_t})\right)^2 E\left[\frac{\delta_t}{S_C(\sum_{i=1}^t R_i)}\Big|\boldsymbol{S}_t,\boldsymbol{A}_t,R_t\right]\right]$$
$$=E\left[E\left[\left(R_t+\max_{a_{t+1}}Q^*_{t+1}(\boldsymbol{S_{t+1}},(\boldsymbol{A_t},a_{t+1}))-Q_t(\boldsymbol{S_t},\boldsymbol{A_t})\right)^2\frac{\delta_t}{S_C(\sum_{i=1}^t R_i)}\Big|\boldsymbol{S}_t,\boldsymbol{A}_t,R_t\right]\right] \quad (12)$$
$$=E\left[\left(R_t+\max_{a_{t+1}}Q^*_{t+1}(\boldsymbol{S_{t+1}},(\boldsymbol{A_t},a_{t+1}))-Q_t(\boldsymbol{S_t},\boldsymbol{A_t})\right)^2\frac{\delta_t}{S_C(\sum_{i=1}^t R_i)}\right].$$

Since Q_t^* is the minimizer of the first expression in the above sequence of equalities, it also minimize the last expression. Thus, we suggest to choose $\hat{Q}_t$ recursively as follows:

$$\operatorname*{argmin}_{Q_t\in\mathcal{Q}_t}\mathbb{E}_n\left[\left(R_t+\max_{a_{t+1}}\widehat{Q}_{t+1}(\boldsymbol{S_{t+1}},(\boldsymbol{A_t},a_{t+1}))-Q_t(\boldsymbol{S_t},\boldsymbol{A_t})\right)^2\frac{\delta_t}{\widehat{S}_C(\sum_{i=1}^t R_i)}\right], \quad (13)$$

where we define $\hat{Q}_{T+1}\equiv 0$, and $\hat{S}_C$ is the Kaplan-Meier estimator of the survival function of the censoring variable S_C. Note that by Remark 3.1, the Kaplan-Meier estimator at x needs to estimate $P(C\geq x)$ rather than $P(C>x)$. This can be done by taking left a continuous version of the Kaplan-Meier estimator that interchanges the roles of failure and censoring events for estimation (see Satten and Datta, 2001).

We define the policies $\hat{\boldsymbol{\pi}}_t$ using the approximated Q-functions $\hat{Q}_t$ as follows:

$$\widehat{\pi}_t(\boldsymbol{s}_t,\boldsymbol{a}_{t-1})=\operatorname*{argmax}_{a_t}\widehat{Q}_t(\boldsymbol{s}_t,(\boldsymbol{a}_{t-1},a_t)).$$

6. Theoretical Results

Let $\{\mathcal{Q}_1,\ldots,\mathcal{Q}_T\}$ be the approximation spaces for the minimization problems (13). Note that we do *not* assume that the problem is Markovian, but, instead, we assume that each Q_t is a function of all the history up to and including stage t. Hence the spaces $\mathcal{Q}_t$ can be different over t.

We assume that the absolute values of the functions in the spaces $\{\mathcal{Q}_t\}_t$ are bounded by some constant M. Moreover, we need to bound the complexity of the spaces $\{\mathcal{Q}_t\}_t$. We choose to use uniform entropy as the complexity measure (see van der Vaart and Wellner, 1996). This enables us to obtain exponential bounds on the difference between the true and empirical expectation of the loss function that involves a random component, namely, the Kaplan-Meier estimator, as in (13) (see Lemma A.6). This is different from Murphy (2005b) who uses the covering number as a measure of complexity (Anthony and Bartlett, 1999, pg 148) for the squared error loss function.

For every $\varepsilon>0$ and measure P, we denote the covering number of $\mathcal{Q}$ by $N(\varepsilon,\mathcal{Q},L_2(P))$, where $N(\varepsilon,\mathcal{Q},L_2(P))$ is the minimal number of closed $L_2(P)$-balls of radius ε required to cover $\mathcal{Q}$. The uniform covering number of $\mathcal{Q}$ is defined as $\sup_P N(\varepsilon M,\mathcal{Q},L_2(P))$ where the supremum is taken over all finitely discrete probability measures P on $\mathcal{Q}$. The log of the uniform covering number is called the uniform entropy (van der Vaart and Wellner, 1996, page 84). We assume the following uniform entropy bound for the spaces $\{\mathcal{Q}_t\}$:

$$\max_{t=\{1,\ldots,T\}}\sup_P\log N(\varepsilon M,\mathcal{Q}_t,L_2(P))<D\left(\frac{1}{\varepsilon}\right)^W, \quad (14)$$

for all $0 < \varepsilon \leq 1$ and some constants $0 < W < 2$ and $D < \infty$, where the supremum is taken over all finitely discrete probability measures, and M is the uniform bound defined above.

In the following, we prove a finite sample bound on the difference between the expected truncated survival times of an optimal policy and the policy $\hat{\pi}$ obtained by the algorithm. As a corollary we obtain that the difference converges to zero under certain conditions.

The proof of the theorem consists of the following steps. First we use Lemma 4.1 to map the original problem to the corresponding auxiliary one. Second, for the auxiliary problem, we adapt arguments given in Murphy (2005b) to bound the difference between the expected value of the learned policy and the expected value of the optimal policy using error terms that involve expectations of both the learned and optimal Q-functions. Third, we bound these error terms by decomposing them to terms that arise due to the difference between the empirical and true expectation, terms that arise due the differences between the estimated and true censoring distribution, and terms that related to the empirical difference between the estimated and optimal Q-function. Fourth, and finally, we obtain a finite sample bound which depends on the complexity of the spaces $\{\mathcal{Q}_t\}$, the deviation of the Kaplan-Meier estimator from the censoring distribution, and the size of the empirical errors in (13).

Theorem 6.1

Let $\{\mathcal{Q}_1, \ldots, \mathcal{Q}_T\}$ be the approximation spaces for the Q-functions. Assume that the uniform entropy bound (14) holds. Assume that n trajectories are sampled according to P_0. Let $\hat{\pi}$ be defined by (4).

Then for any $0 < \eta < 1$, we have with probability at least $1 - \eta$, over the random sample of trajectories,

$$\begin{aligned} &\sup_{\pi \in \Pi} E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right] - E_{0,\hat{\pi}}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right] \\ &\leq 16\varepsilon + \\ &\sum_{t=1}^{T} L^{t/2} \sum_{j=t}^{T} \left(2L^j 4^{j-t} \mathbb{E}_n\left[\frac{\delta_t}{\widehat{S}_C(\sum_{i=1}^{t} R_i)} \left(F(\widehat{Q}_t, \widehat{Q}_{t+1}) - F(Q_t^*, \widehat{Q}_{t+1})\right)\right]_+\right)^{1/2} \end{aligned} \tag{15}$$

for all n that satisfies

$$\max\left\{\frac{5T}{2}\exp\{-nC_1\varepsilon^4 + \sqrt{n}C_2\varepsilon^2\}, TC_3\exp\{-2n\varepsilon^4 + C_4\sqrt{n}\varepsilon^{2(U+\alpha_o)}\}\right\} < \frac{\eta}{2},$$

where

$$\begin{aligned} F(Q_t, Q_{t+1}) &= (R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t)^2, \\ C_1 &= 2(1 - G_{\min})^2 M_1^{-2} K_{\min}^4 (4L)^{-2(T+1)}, \\ C_2 &= C_o(1 - G_{\min}) M_1^{-1} K_{\min}^2 (4L)^{-(T+1)}, \\ C_3 &= C_a \exp\{(4L)^{-(T+1)}\}, \\ C_4 &= C_b (4L)^{(T+1)/2}, \end{aligned}$$

and where $M_1 = (2M + \tau)^2$, C_o is the constant that appears in Bitouzé et al. (1999, Eq. 1), C_a, C_b, and U are the constants that appear in Lemma A.6, and for some α_o small enough such that $U + \alpha_o < 2$.

Before we begin the proof of Theorem 6.1, we note that the bound (15) cannot be used in practice to perform structural risk minimization (see, for example, Vapnik, 1999) for two reasons. First, the bound itself is too loose (see also Murphy, 2005b, Theorem 1, Remark 4). Second, the constants, such as C_a and C_b, are not given, and are model dependent. Interestingly, a bound on C_o was established recently by Wellner (2007). However, this bound is large and simulations suggest that it is not tight. The bound (15) can, however, be used to derive asymptotic rates (Steinwart and Chirstmann, 2008, Chapter 6). Moreover, when the functions Q_t^* are in $\mathcal{Q}_t$, we obtain universal consistency, as stated in the following corollary:

Corollary 6.2

Assume that the conditions of Theorem 6.1 hold. Assume also that for every t, $Q_t^* \in \mathcal{Q}_t$. Then

$$\sup_{\pi \in \Pi} E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right] - E_{0,\widehat{\pi}}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right] \xrightarrow{\text{a.s.}} 0.$$

Proof of Corollary 6.2

Note that for every t, $\hat{Q}_t$ is the minimizer of

$$\mathbb{E}_n\left[\frac{\delta_t}{\widehat{S}_c(\sum_{i=1}^{t} R_i)} F(Q_t, \widehat{Q}_{t+1})\right].$$

Hence, the second expression in the right hand side of (15) equals zero, and the result follows.

Proof of Theorem 6.1

By Lemma 4.1,

$$\sup_{\pi \in \Pi} E_{0,\pi}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right] - E_{0,\widehat{\pi}}\left[\left(\sum_{t=1}^{\overline{T}} R_t\right) \wedge \tau\right] = E[V^*(S_1) - V_{\widehat{\pi}}(S_1)],$$

where the expectation on the right hand side of the equality is with respect to the modified distribution P.

By Lemma 2 of Murphy (2005b) and Remark 2 that follows, for every state $s_o \in \mathcal{S}_1$,

$$V^*(s_o) - V_{\widehat{\pi}}(s_o) \leq \sum_{t=1}^{T} 2L^{t/2} \sqrt{E\left[\left(\widehat{Q}_t(\boldsymbol{S}_t, \boldsymbol{A}_t) - Q_t^*(\boldsymbol{S}_t, \boldsymbol{A}_t)\right)^2 \middle| S_1 = s_o\right]}.$$

Applying Jensen's inequality, we obtain

$$E[V^*(S_1) - V_{\widehat{\pi}}(S_1)] \leq \sum_{t=1}^{T} 2L^{t/2} \sqrt{E\left[\left(\widehat{Q}_t(\boldsymbol{S}_t, \boldsymbol{A}_t) - Q_t^*(\boldsymbol{S}_t, \boldsymbol{A}_t)\right)^2\right]}. \tag{16}$$

We wish to obtain a bound on the expression $E\left[\left(\widehat{Q}_t(\boldsymbol{S}_t, \boldsymbol{A}_t) - Q_t^*(\boldsymbol{S}_t, \boldsymbol{A}_t)\right)^2\right]$ using the expressions $Err_{\widehat{Q}_{t+1}}(\widehat{Q}_t) - Err_{\widehat{Q}_{t+1}}(Q_t^*)$, where

$$Err_{Q_{t+1}}(Q_t) = E\left[\left(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t(\boldsymbol{S}_t, \boldsymbol{A}_t)\right)^2\right],$$

for any pair of function Q_t and Q_{t+1}. To obtain this bound we follow the line of arguments that leads to the bound in Eq. 13 in the proof of Theorem 1 of Murphy (2005b). The bound (19) obtained here is tighter since only the special case of Q_t^* in the second *Err* function is considered. To simplify the following expressions, we write Q_t instead of $Q_t(\boldsymbol{S}_t, \boldsymbol{A}_t)$ whenever no confusion could occur.

For each t,

$$\begin{aligned}
&Err_{\widehat{Q}_{t+1}}(\widehat{Q}_t) - Err_{\widehat{Q}_{t+1}}(Q_t^*) \\
=&E[\widehat{Q}_t^2] - E[(Q_t^*)^2] + 2E[(R_t + \max_{a_{t+1}} \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}))(Q_t^* - \widehat{Q}_t)] \\
=&E[\widehat{Q}_t^2] - E[(Q_t^*)^2] + 2E[(Q_t^* - \widehat{Q}_t)E[(R_t - \max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}))|\boldsymbol{S}_t, \boldsymbol{A}_t]] \\
&+2E[(\max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - \max_{a_{t+1}} \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}))(Q_t^* - \widehat{Q}_t)] \\
=&E[\widehat{Q}_t^2] - E[(Q_t^*)^2] + 2E[(Q_t^*)^2] - 2E[\widehat{Q}_t Q_t^*] \\
&+2E[(\max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - \max_{a_{t+1}} \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}))(Q_t^* - \widehat{Q}_t)] \\
=&E[(\widehat{Q}_t - Q_t^*)^2] \\
&+2E[(\max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - \max_{a_{t+1}} \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}))(Q_t^* - \widehat{Q}_t)],
\end{aligned} \tag{17}$$

where the second to the last equality follows since

$$Q_t^*(\boldsymbol{s}_t, \boldsymbol{a}_t) = E[(R_t - \max_{a_{t+1}} \widehat{Q}_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}))|\boldsymbol{S}_t = \boldsymbol{s}_t, \boldsymbol{A}_t = \boldsymbol{a}_t].$$

Using the Cauchy-Schwarz inequality for the second expression of (17), we obtain

$$\begin{aligned}
&Err_{\widehat{Q}_{t+1}}(\widehat{Q}_t) - Err_{\widehat{Q}_{t+1}}(Q_t^*) \\
&\geq E[(\widehat{Q}_t - Q_t^*)^2] \\
&-2E[(\max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - \max_{a_{t+1}} \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}))^2]^{1/2} E[(Q_t^* - \widehat{Q}_t)^2]^1
\end{aligned}$$

Note that

$$E[\left(\max_{a_{t+1}} Q^*_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - \max_{a_{t+1}} \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1})\right)^2] \le E\left[\max_{a_{t+1}}\left(Q^*_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1})\right)^2\right]$$
$$\le E\left[L\sum_{a\in\mathscr{A}}\left(Q^*_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a) - \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a)\right)^2 p_t(a|\boldsymbol{S}_{t+1}, \boldsymbol{A}_t)\right] \quad (18)$$
$$= LE\left[\left(Q^*_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_{t+1}) - \widehat{Q}_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_{t+1})\right)^2\right],$$

where the first inequality follows since $(\max_a h(a) - \max_a h'(a))^2 \le \max_a (h(a) - h'(a))^2$ and where L is the constant that appears in the definition of the exploration policy $\boldsymbol{p}$ (see Section 3).

Using inequality (18) and the fact that $xy \le \frac{1}{2}(x^2+y^2)$, we obtain

$$Err_{\widehat{Q}_{t+1}}(\widehat{Q}_t) - Err_{\widehat{Q}_{t+1}}(Q^*_t) \ge E[(\widehat{Q}_t - Q^*_t)^2] - E[4L(Q^*_{t+1} - \widehat{Q}_{t+1})^2]^{1/2} E[(Q^*_t - \widehat{Q}_t)^2]^{1/2}$$
$$\ge \tfrac{1}{2}E[(\widehat{Q}_t - Q^*_t)^2] - 2LE[(Q^*_{t+1} - \widehat{Q}_{t+1})^2].$$

Hence

$$E[(\widehat{Q}_t - Q^*_t)^2] \le 2\left(Err_{\widehat{Q}_{t+1}}(\widehat{Q}_t) - Err_{\widehat{Q}_{t+1}}(Q^*_t)\right) + 4LE[(Q^*_{t+1} - \widehat{Q}_{t+1})^2].$$

Using the fact that $\widehat{Q}_{T+1} = Q^*_{T+1} = 0$ we obtain

$$E[(\widehat{Q}_t - Q^*_t)^2] \le 2\sum_{j=t}^{T}(4L)^{j-t}\left(Err_{\widehat{Q}_{j+1}}(\widehat{Q}_j) - Err_{\widehat{Q}_{j+1}}(Q^*_j)\right). \quad (19)$$

We are now ready to bound the expressions $Err_{\widehat{Q}_{j+1}}(\widehat{Q}_j) - Err_{\widehat{Q}_{j+1}}(Q^*_j)$. For any $Q_t \in \mathscr{Q}_t \cup Q^*_t$, $Q_{t+1} \in \mathscr{Q}_{t+1}$, and censoring survival function K: $[0, \tau] \mapsto [K_{\min}, 1]$, where $K_{\min} > 0$, define

$$\mathscr{E}(Q_t, Q_{t+1}, K) = E\left[\frac{\delta_t}{K(\sum_{i=1}^t R_i)}(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t)^2\right],$$
$$\mathscr{E}_n(Q_t, Q_{t+1}, K) = \mathbb{E}_n\left[\frac{\delta_t}{K(\sum_{i=1}^t R_i)}(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t)^2\right]. \quad (20)$$

Note that similarly to (12) we have $Err_{\hat{Q}t+1}(Q_t) = \mathscr{E}(Q_t, \hat{Q}_{t+1}, S_C)$, where S_C is the censoring survival function.

Using this notation we have

$$\begin{aligned} Err_{\widehat{Q}_{t+1}}(\widehat{Q}_t) - Err_{\widehat{Q}_{t+1}}(Q_t^*) &= \mathscr{E}(\widehat{Q}_t, \widehat{Q}_{t+1}, S_C) - \mathscr{E}(Q_t^*, \widehat{Q}_{t+1}, S_C) \\ &\le |\mathscr{E}(\widehat{Q}_t, \widehat{Q}_{t+1}, S_C) - \mathscr{E}(\widehat{Q}_t, \widehat{Q}_{t+1}, \widehat{S}_C)| \\ &+ |\mathscr{E}(\widehat{Q}_t, \widehat{Q}_{t+1}, \widehat{S}_C) - \mathscr{E}_n(\widehat{Q}_t, \widehat{Q}_{t+1}, \widehat{S}_C)| \\ &+ (\mathscr{E}_n(\widehat{Q}_t, \widehat{Q}_{t+1}, \widehat{S}_C) - \mathscr{E}_n(Q_t^*, \widehat{Q}_{t+1}, \widehat{S}_C))+ \\ &|\mathscr{E}_n(Q_t^*, \widehat{Q}_{t+1}, \widehat{S}_C) - \mathscr{E}(Q_t^*, \widehat{Q}_{t+1}, \widehat{S}_C)| \\ &+ |\mathscr{E}(Q_t^*, \widehat{Q}_{t+1}, \widehat{S}_C) - \mathscr{E}(Q_t^*, \widehat{Q}_{t+1}, S_C)|, \end{aligned}$$

where $\hat{S}_C$ is the Kaplan-Meier estimator of S_C, and $(a)_+ = \max\{a, 0\}$. Hence

$$Err_{\widehat{Q}_{t+1}}(\widehat{Q}_t) - Err_{\widehat{Q}_{t+1}}(Q_t^*) \le 2 \sup_{\{Q_t, Q_{t+1}\}} |\mathscr{E}(Q_t, Q_{t+1}, S_C) - \mathscr{E}(Q_t, Q_{t+1}, \widehat{S}_C)| \tag{21}$$

$$+2 \sup_{\{Q_t, Q_{t+1}, K\}} |\mathscr{E}(Q_t, Q_{t+1}, K) - \mathscr{E}_n(Q_t, Q_{t+1}, K)| \tag{22}$$

$$+ (\mathscr{E}_n(\widehat{Q}_t, \widehat{Q}_{t+1}, \widehat{S}_C) - \mathscr{E}_n(Q_t^*, \widehat{Q}_{t+1}, \widehat{S}_C))+. \tag{23}$$

Combining (19) and (21), and substituting in (16), we have

$$\begin{aligned} E[V^*(S_1) - V_{\widehat{\pi}}(S_1)] &\le 2 \sum_{t=1}^{T} L^{t/2} \sum_{j=t}^{T} \sqrt{2(4L)^{j-t}(Err_{Q_{t+1}}(Q_t) - Err_{Q_{t+1}}(Q_t^*))} \\ &\le 8(4L)^{(T+1)/2} \sqrt{\max_t \sup_{\{Q_t, Q_{t+1}\}} |\mathscr{E}(Q_t, Q_{t+1}, S_C) - \mathscr{E}(Q_t, Q_{t+1}, \widehat{S}_C)|} \end{aligned} \tag{24}$$

$$\begin{aligned} &+8(4L)^{(T+1)/2} \sqrt{\max_t \sup_{\{Q_t, Q_{t+1}, K\}} |\mathscr{E}(Q_t, Q_{t+1}, K) - \mathscr{E}_n(Q_t, Q_{t+1}, K)|} \\ &+2 \sum_{t=1}^{T} L^{t/2} \sum_{j=t}^{T} 2^{j-t} L^{j/2} \sqrt{2(\mathscr{E}_n(\widehat{Q}_t, \widehat{Q}_{t+1}, \widehat{S}_C) - \mathscr{E}_n(Q_t^*, \widehat{Q}_{t+1}, \widehat{S}_C))_+}, \end{aligned} \tag{25}$$

where we used the fact that $\sum_{t=1}^{T} L^{t/2} \sum_{j=t}^{T} (4L)^{(j-t)/2} \le 2(4L)^{(T+1)/2}$ for $L \ge 2$ and the fact that $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$.

In the following, we replace the bounds in (24) and (25)) with exponential bounds. We start with (24). Note that $(R_t + \max_{a_{t+1}} Q_{t+1}(\boldsymbol{S}_{t+1}, \boldsymbol{A}_t, a_{t+1}) - Q_t)^2 \le M_1 = (2M + \tau)^2$ for all Q_t, Q_{t+1}. Hence,

$$\sup_{\{Q_t, Q_{t+1}\}} \left| \mathscr{E}(Q_t, Q_{T+1}, S_C) - \mathscr{E}(Q_t, Q_{T+1}, \widehat{S}_C) \right| \le M_1 K_{\min}^{-2} E[|S_C - \widehat{S}_C|],$$

, and thus

$$\begin{aligned} &P\left((4L)^{T/2+2}\sqrt{\max_t \sup_{\{Q_t,Q_{t+1}\}} |\mathscr{E}(Q_t,Q_{T+1},S_C)-\mathscr{E}(Q_t,Q_{T+1},\widehat{S}_C)|}>\varepsilon\right)\\ &\le \sum_{t=1}^{T} P\left((4L)^{T/2+2}\sqrt{\sup_{\{Q_t,Q_{t+1}\}} |\mathscr{E}(Q_t,Q_{T+1},S_C)-\mathscr{E}(Q_t,Q_{T+1},\widehat{S}_C)|}>\varepsilon\right)\\ &\le TP\left((4L)^{T/2+2}\sqrt{M_1K_{\min}^{-2}\left\|S-\widehat{S}_C\right\|_\infty}>\varepsilon\right), \end{aligned} \tag{26}$$

where the first equality follows from the fact that

$$P(\max_{t\in\{1,\ldots,T\}} X_t>c) \le \sum_{t=1}^{T} P(X_t>c). \tag{27}$$

Using a Dvoretzky-Kiefer-Wolfowitz-type inequality for the Kaplan-Meier estimator (Bitouzé et al., 1999, Theorem 2), we have

$$P(\left\|S_C-\widehat{S}_C\right\|_\infty>\varepsilon')<\frac{5}{2}\exp\{-2n(1-G_{\min})^2(\varepsilon')^2+C_o\sqrt{n}(1-G_{\min})\varepsilon'\}, \tag{28}$$

where C_o is some universal constant and $G_{\min}$ is a lower bound on the survival function at τ (see Section 3).

Write $\varepsilon=(4L)^{(T+1)/2}\sqrt{M_1K_{\min}^{-2}\varepsilon'}$, and thus $\varepsilon'=M_1^{-1}K_{\min}^2\varepsilon^2(4L)^{-(T+1)}$. Note that $8(4L)^{T/2+2}\sqrt{M_1K_{\min}^{-2}\left\|S_C-\widehat{S}_C\right\|_\infty}>8\varepsilon$ iff $\left\|S_C-\widehat{S}_C\right\|_\infty>\varepsilon'$. Applying the inequality (28) to the right hand-side of (26) and substituting for ε, we obtain

$$\begin{aligned} &P\left(8(4L)^{(T+1)/2}\sqrt{\sup_t \sup_{\{Q_t,Q_{t+1}\}} |\mathscr{E}(Q_t,Q_{T+1},S_C)-\mathscr{E}(Q_t,Q_{T+1},\widehat{S}_C)|}>8\varepsilon\right)\\ &\le \tfrac{5T}{2}\exp\left\{-2n(1-G_{\min})^2M_1^{-2}K_{\min}^4\varepsilon^4(4L)^{-2(T+1)}\right.\\ &\left.+C_o\sqrt{n}(1-G_{\min})M_1^{-1}K_{\min}^2\varepsilon^2(4L)^{-(T+1)}\right\}\\ &\equiv \tfrac{5T}{2}\exp\{-nC_1\varepsilon^4+\sqrt{n}C_2\varepsilon^2\}, \end{aligned} \tag{29}$$

where $C_1=2(1-G_{\min})^2M_1^{-2}K_{\min}^4(4L)^{-2(T+1)}$ and $C_2=C_o(1-G_{\min})M_1^{-1}K_{\min}^2(4L)^{-(T+1)}$.

We now find an exponential bound for (25). We follow the same line of arguments, replacing the Dvoretzky-Kiefer-Wolfowitz type inequality used in the previous proof with the uniform entropy bound. Recall that by assumption, the uniform entropy bound (14) holds for the spaces $\mathscr{Q}_t$ and thus also for the spaces $\mathscr{Q}_t \cup Q_t^*$. Hence, by Lemma A.6, and (27), for $W'=\max\{W,1\}$ and for all $\alpha>0$, we have

$$\begin{aligned} &P\left(8(4L)^{(T+1)/2}\sqrt{\max_t \sup_{\{Q_t,Q_{t+1},K\}} |\mathscr{E}(Q_t,Q_{t+1},K)-\mathscr{E}_n(Q_t,Q_{t+1},K)|}>8\varepsilon\right)\\ &\le TC_a\exp\{C_b\sqrt{n}(4L)^{-(T+1)/2}\varepsilon^{2(U+\alpha)}-2n(4L)^{-(T+1)}\varepsilon^4\}\\ &\equiv TC_3\exp\{C_4\sqrt{n}\varepsilon^{2(U+\alpha)}-2n\varepsilon^4\}, \end{aligned} \tag{30}$$

where $C_3 = C_a \exp\{(4L)^{-(T+1)}\}$, $C_4 = C_b(4L)^{(T+1)/2}$, and $U = W'(6 - W')/(2 + W')$.

Take n large enough such that the right hand side of (29) and (30) are less than $\eta/2$ and substitute in (24) and (25), respectively, and the result of the theorem follows.

7. Simulation Study

We simulate a randomized clinical trial with flexible number of stages to examine the performance of the proposed censored Q-learning algorithm. We compare the estimated individualized treatment policy to various possible fixed treatments. We also compare the given expected survival times of different censoring levels. Finally, we test the effect of ignoring the censoring.

This section is organized as follows. We first describe the setting of the simulated clinical trial (Section 7.1). We then describe the implementation of the simulation (Section 7.2). The simulation results appear in Section 7.3.

7.1. Simulated Clinical Trial

We consider the following hypothetical cancer trial. The duration of the trial is 3 years. The state of each patient at each time-point $u \in [0, 3]$ includes the tumor size ($0 \le T(u) \le 1$), and the wellness ($0.25 \le W(u) \le 1$). The time-point u_o such that $W(u_o) < 0.25$ is considered the failure time. We define the critical tumor size to be 1. At time u_i such that $T(u_i) = 1$, we begin a treatment. We call the duration $[u_i, u_{i+1}]$ the i-th stage. Note that different patients may have different numbers of stages.

At each time-point u_i, we consider two optional treatments: a more aggressive treatment (A), and a less aggressive treatment (B). The immediate effects of treatment A are

$$\begin{aligned} W(u_i^+|A) &= W(u_i) - 0.5, \\ T(u_i^+|A) &= T(u_i)/(10W(u_i)), \end{aligned} \tag{31}$$

i.e., the wellness at time u_i after treatment A (denoted by $W(u_i^+|A)$) decreases by 0.5 wellness units. The tumor size at time u_i after treatment A (denoted by $T(u_i^+|A)$) decreases by a factor of $1-(10W(u_i))$ which reflects a greater decrease of tumor size for a larger wellness value. Similarly, the immediate effects of the less aggressive treatment B are

$$\begin{aligned} W(u_i^+|B) &= W(u_i) - 0.25, \\ T(u_i^+|B) &= T(u_i)/(4W(u_i)), \end{aligned} \tag{32}$$

which, in comparison to the treatment A, has lower effect on the tumor size but also lower decrease of wellness. The wellness and tumor size at time $u_i < u \le u_{i+1}$ follows the dynamics

$$\begin{aligned} W(u) &= W(u_i^+) + (1 - W(u_i^+))\,(1 - 2^{-(u-u_i)/2}), \\ T(u) &= T(u_i^+) + 4T(u_i^+)\,(u - u_i)/3. \end{aligned} \tag{33}$$

The stage that begins at time-point u_i ends when either $T(u_{i+1}) = 1$ for some $u_i < u_{i+1} < 3$ or when a failure event occurs or at the end of the trial when $u = 3$. During this stage, we model the survival function of the patient as an exponential distribution with mean $3(W(u_i^+)+2)/20M(u_i^+)$.

The trajectories are constructed as follows. We assume that patients are recruited to the trial when their tumor size reaches the critical size, i.e., for all patients $T(0) = 1$, and hence $u_1 = 0$ is the beginning of the first stage. The wellness at the beginning of the first stage, $W(0)$, is uniformly distributed on the segment [0.5, 1]. With equal probability, a treatment $a_1 \in \{A, B\}$ is chosen. If no failure event occurs during the first stage, the first stage ends when either $T(u_2) = 1$ for some $0 = u_1 < u_2 < 3$ or at the end of the trial. If the first stage ends before the end of the trial, then with equal probability another treatment $a_2 \in \{A, B\}$ is chosen. The trial continues in the same way until either a failure time occurs or the trial ends. We note that the actual number of stages for each patient is a random function of the initial state and the treatments chosen during the trial. Due to the choices of model parameters, the number of stages in the above dynamics is at least one and not more than three.

For each trajectory, a censoring variable C is uniformly drawn from the segment $[0, c]$ for some constant $c > 3$, where the choice of the constant c determines the expected percentage of censoring. When an event is censored, the trajectory (i.e., the states and treatments) up to the point of censoring and the censoring time are given.

7.2. Simulation Implementation

The Q-learning algorithm presented in Section 5 was implemented in the Matlab environment. For the implementation we used the Spider library for Matlab[1]. The Matlab code, as well as the data sets, are available online (see Supplement A).

The algorithm is implemented as follows. The input for the algorithm is a set of trajectories obtained according to the dynamics described in Section 7.1. First, the Kaplan-Meier estimator for the survival function of the censoring variable is computed from the given trajectories. Then, we set $\hat{Q}_4 \equiv 0$ and compute $\hat{Q}_i$, $i = 3, 2, 1$ backwardly, as the minimizer of (13) over all the functions $Q_i(s_i, a_i)$ which are linear in the first variable. The policy $\hat{\pi}$ is computed from the functions $\{\hat{Q}_1, \hat{Q}_2, \hat{Q}_3\}$ using (4).

We tested the policy $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3)$ by constructing 1000 new trajectories, in which the choice of treatment at each stage is according to $\hat{\pi}$. 1000 initial wellness values were drawn uniformly from the segment [0.5, 1]. For each wellness value, a treatment was chosen from the set $\{A, B\}$, according to the policy $\hat{\pi}_1$. The immediate effect of the treatment was computed according to (31)–(32). A failure time was drawn from the exponential distribution with mean as described in the previous section; denote this time by f_1. The time that the tumor reached the critical size was computed according to the dynamics (33), and we denote this time by u_2. If both f_1 and u_2 are greater than 3 (the end of the trial) then the trajectory was ended after the first stage and the survival time for this patient was given as 3. Otherwise, if $f_1 \le u_2$, the trajectory is ended after the first stage and the survival time for this patient was given as f_1. If $u_2 < f_1$, then at time u_2, a second treatment is chosen according to the policy $\hat{\pi}_2$. The computation of the reminder of the trajectory is done similarly. The expected value of the policy $\hat{\pi}$ is estimated by the mean of the survival time of all 1000 patients.

We compared the results of the algorithm to all fixed treatment sequences $A_1A_2A_3$, where $A_i \in \{A, B\}$. The expected values of the fixed treatment sequences were computed explicitly. We also compared the results to that of the optimal policy, which was also computed explicitly.

[1]The Spider library for Matlab can be downloaded form http://www.kyb.tuebingen.mpg.de/bs/people/spider/

7.3. Simulation and Results

First, we would like to examine the influence of the sample size and censoring percentage on the algorithm's performance. We simulated data sets of trajectories of sizes 40, 80, 120, …, 400. For each set of trajectories we considered four levels of censoring: no censoring, 10%-censoring, 20%-censoring, and 30%-censoring. Higher levels of (uniform) censoring were not considered since this requires drawing the censoring variable from a segment $[0, c]$ for $c < 3$, which is in contrast to the assumption on the censoring variable (see the beginning of section 3). A policy $\hat{\pi}$ was computed for each combination of data set size and censoring percentage. The policy $\hat{\pi}$ was evaluated on a data set of size 1000, as described in Section 7.2. We repeated the simulation 400 times for each combination of data set size and censoring percentage. The mean values of the estimated mean survival time are presented in Figure 1. A comparison between the different fixed policies, polices obtained by the algorithm for different censoring levels, and the optimal policy appears in Figure 2. As can be seen from both figures, the individualized treatment polices obtained by the algorithm are better than any fixed policy. Moreover, as the number of observed trajectories increases, the expected survival time increases, for all censoring percentages.

We also examined the influence of the sample size and censoring percentage on the distribution of estimated expected survival time. We simulated data sets of sizes 50, 100, 200, …, 3200 and we considered the four levels of censoring as before. As can be seen from Figure 3, the variance decreases when the sample size becomes larger. Also, the variance is smaller for smaller percentage of censoring, although the difference is modest.

Note that the maximum expected survival times obtained by the algorithm are a little bit above 17 months (see both Figures 1 and 2), while the value of the optimal policy is 17.85. The difference follows from the fact that the Q-functions estimated by the algorithm are linear while the optimal Q-function is not (see Figure 4). It is worth mentioning that even in the class of linear functions on which the optimization is done there are Q-functions that yield higher values. This fact is often referred to as the "mismatch" that follows from the fact that optimization of the value function is not performed explicitly, but rather through optimization of the Q-functions (see Tsitsiklis and van Roy, 1996; Murphy, 2005b, for more details).

Figure 5 shows the number of treatments that were needed for patients that followed the policy $\hat{\pi}$ and did not have a failure event during the trial. As can be seen from this figure, patients with high initial wellness need only one treatment. On the other hand, patients with very low initial wellness value need three treatments.

Finally, we checked the effect of ignoring the censoring on the expected survival time. We considered two ways of ignoring the censoring. First, we consider an algorithm that ignores the weights in the minimization problem (13). This is equivalent to deleting the last stage from each trajectory that was censored. We also consider an algorithm that deletes all censored trajectories. In the example presented in Figures 1–5, where uniform censoring takes place, there is a relatively moderate difference between the expected survival time for the proposed algorithm and the other two algorithms that ignore censoring. However, when the censoring variable follows the exponential distribution (leaving fewer observations with longer survival times) the bias from ignoring the censored trajectories is substantial, as can be seen in Figure 6.

8. Summary

We studied a framework for multistage-decision problems with flexible number of stages in which the rewards are survival times and are subject to censoring. We proposed a novel Q-

learning algorithm adjusted for censoring. We derived the generalization error properties of the algorithm and demonstrated the algorithm performance using simulations.

The work as presented is applicable to real-world multi-stage decision problems with censoring. However, two main issues should be noted. First, we assumed that censoring is independent of observed trajectories. It would be useful to relax this assumption and allow censoring to depend on the covariates. Developing an algorithm that works under this relaxed assumption is a challenge. Second, we have used the inverse-probability-of-censoring weighting to correct the bias induced by censoring. When the percentage of censored trajectories is large, the algorithm may be inefficient. Finding a more efficient algorithm is also an open question.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Anthony, M.; Bartlett, PL. Neural Network Learning: Theoretical Foundations. Cambridge University Press; 1999.

Bellman, R. Dynamic Programming. Princeton University Press; 1957.

Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. Statist Med. 1998; 17:1169–1186.

Bitouzé D, Laurent B, Massart P. A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. Ann Inst H Poincaré Probab Statist. 1999; 35:735–763.

Chen P, Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. Biometrics. 2001; 57:1030–1038. [PubMed: 11764241]

Goldberg, Y.; Kosorok, MR.; Lin, DY. Unpublished manuscript. 2011. Support vector regression under right censoring.

Karrison TG. Use of Irwin's restricted mean as an index for comparing survival in different treatment groups–Interpretation and power considerations. Controlled Clinical Trials. 1997; 18:151–167. [PubMed: 9129859]

Kosorok, MR. Introduction to Empirical Processes and Semiparametric Inference. Springer; New York: 2008.

Krzakowski M, Ramlau R, Jassem J, Szczesna A, Zatloukal P, Von Pawel J, Sun X, Bennouna J, Santoro A, Biesma B, Delgado FM, Salhi Y, Vaissiere N, Hansen O, Tan E, Quoix E, Garrido P, Douillard J. Phase III trial comparing Vinflunine with Docetaxel in second-line advanced nonsmall-cell lung cancer previously treated with platinum-containing chemotherapy. Journal of Clinical Oncology. 2010; 28:2167–2173. [PubMed: 20351334]

Laan, MJvd; Petersen, ML. Causal effect models for realistic individualized treatment and intention to treat rules. The International Journal of Biostatistics. 2007; 3:3.

Laber, E.; Qian, M.; Lizotte, DJ.; Murphy, SA. Statistical inference in dynamic treatment regimes. 2010. Avaliable at http://arxiv.org/abs/1006.5831

Lavori PW, Dawson R. Dynamic treatment regimes: Practical design considerations. Clinical Trials. 2004; 1:9–20. [PubMed: 16281458]

Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2002; 58:48–57. [PubMed: 11890326]

Miyahara S, Wahed AS. Weighted Kaplan-Meier estimators for two-stage treatment regimes. Statistics in medicine. 2010

Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007; 63:447–455. [PubMed: 17688497]

Murphy SA. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2003; 65:331–366.

Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in medicine. 2005a; 24:1455–1481. [PubMed: 15586395]

Murphy SA. A generalization error for Q-learning. Journal of Machine Learning Research. 2005b; 6:1073–1097. [PubMed: 16763665]

Murphy SA, Oslin DW, Rush AJ, Zhu J. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. Neuropsychopharmacology. 2006; 32:257–262. [PubMed: 17091129]

Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. The International Journal of Biostatistics. 2010:6.

Robins J, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. Statist Med. 2008; 27:4678–4721.

Robins JM. Association, causation, and marginal structural models. Synthese. 1999; 121:151–179.

Robins, JM. Optimal structural nested models for optimal sequential decisions. In: Lin, D.; Heagerty, PJ., editors. Proceedings of the Second Seattle Symposium in Biostatistics Proceedings of the Second Seattle Symposium in Biostatistics. 2004. p. 189-326.

Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994:89.

Satten GA, Datta S. The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. The American Statistician. 2001; 55:207–210.

Shim J, Hwang C. Support vector censored quantile regression under random censoring. Comput Stat Data Anal. 2009; 53:912–919.

Shivaswamy, PCW.; Jansche, M. A support vector approach to censored targets. Data Mining, 2007 ICDM 2007; Seventh IEEE International Conference on; 2007. p. 655-660.

Steinwart, I.; Chirstmann, A. Support Vector Machines. Springer; 2008.

Stinchcombe TE, Socinski MA. Considerations for second-line therapy of non-small cell lung cancer. Oncologist. 2008; 13:28–36. [PubMed: 18263772]

Sutton, RS.; Barto, AG. Reinforcement Learning: An Introduction. MIT Press; 1998.

Thall PF, Wooten LH, Logothetis CJ, Millikan RE, Tannir NM. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Statist Med. 2007; 26:4687–4702.

TSITSIKLIS JN, van Roy B. Feature-based methods for large scale dynamic programming. Machine Learning. 1996; 22:59–94.

van der VAART, AW.; WELLNER, JA. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer; 1996.

VAPNIK, V. The Nature of Statistical Learning Theory. 2. Springer; 1999.

WAHED AS. Estimation of survival quantiles in two-stage randomization designs. Journal of Statistical Planning and Inference. 2009:139.

WAHED AS, TSIATIS AA. Semiparametric efficient estimation of survival distributions in two-stage randomization designs in clinical trials with censored data. Biometrika. 2006; 93:163–177.

WATKINS, CJCH. PhD thesis. Cambridge University; 1989. Learning from Delayed Rewards.

WATKINS CJCH, DAYAN P. Q-learning. Machine Learning. 1992; 8:279–292.

WELLNER J. On an exponential bound for the KaplanMeier estimator. Lifetime Data Analysis. 2007; 13:481–496. [PubMed: 17805966]

ZHAO, Y.; KOSOROK, MR.; ZD; SMA. Reinforcement learning strategies for clincal trials in non-small cell lung cancer. The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series. 2010. Working Paper 13. Avaliable athttp://biostats.bepress.com/uncbiostat/papers/art13

ZHAO Y, KOSOROK MR, ZENG D. Reinforcement learning design for cancer clinical trials. Statistics in Medicine. 2009; 28:3294–3315. [PubMed: 19750510]

ZUCKER DM. Restricted mean life with covariates: Modification and extension of a useful survival analysis method. Journal of the American Statistical Association. 1998; 93:702–709.

APPENDIX A: SUPPLEMENTARY PROOFS

The main goal of this section is to provide an exponential bound on the difference between the empirical expectation $\mathcal{E}_n(Q_t, Q_{t+1}, K)$ and the true expectation $\mathcal{E}(Q_t, Q_{t+1}, K)$ as a function of the uniform entropy of the class of functions (see (20)). This result appears in Lemma A.6. Similar results for Glivenko–Cantelli classes, Donsker classes, and bounded uniform entropy integral (BUEI) classes can be found in van der Vaart and Wellner (1996) and Kosorok (2008).

Lemma A.1

Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be k sets of functions. Assume that for every $j \in \{1, \ldots, k\}$, $\sup_{f \in \mathcal{F}_j} \|f\|_\infty \leq M_j$. Let $\varphi : \mathbb{R}^k \mapsto \mathbb{R}$ satisfy

$$|\varphi \circ f(x) - \varphi \circ g(x)|^2 \leq c^2 \sum_{j=1}^{k} (f_j(x) - g_j(x))^2 \tag{34}$$

for every $f = (f_1, \ldots, f_k)$, $g = (g_1, \ldots, g_k) \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_k$, where $0 < c < \infty$. Let P be a finitely discrete probability measure. Define $\varphi \circ (\mathcal{F}_1, \ldots, \mathcal{F}_k) = \{\varphi(f_1, \ldots, f_k) : (f_1, \ldots, f_k) \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_k\}$. Then

$$N(\varepsilon c \sum_{j=1}^{k} M_j, \varphi \circ (\mathscr{F}_1, \ldots, \mathscr{F}_k), L_2(P)) \leq \prod_{j=1}^{k} N(\varepsilon M_j, \mathscr{F}_j, L_2(P)). \tag{35}$$

Proof

The proof is similar to the proof of Kosorok (2008, Lemma 9.13). Let $f, g \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_k$ satisfy $\|f_j - g_j\|_{P,2} < \varepsilon M_j$ for $1 \leq j \leq k$. Note that

$$\left\|\varphi \circ f - \varphi \circ g\right\|_{P,2} \leq c \sqrt{\sum_{j=1}^{k} \left\|f_j - g_j\right\|_{P,2}^2} \leq c\varepsilon \sum_{j=1}^{k} M_j$$

which implies (35).

The following two corollaries are direct result of Lemma A.1:

Corollary A.2

Let $\mathcal{K} = \{K : K \text{ is monotone decreasing } K : [0, \tau] \mapsto [K_{\min}, 1]\}$. Define $\mathcal{K}^{-1} = \{1/K : K \in \mathcal{K}\}$. Let P be a finitely discrete probability measure. Then

$$N(\varepsilon K_{\min}^{-1}, \mathscr{K}^{-1}, L_2(P)) \leq N(\varepsilon, \mathscr{K}, L_2(P)).$$

Proof

Note that inequality (34) holds for $k = 1$ and $c = K_{\min}^{-1}$, and the results follow from Lemma A.1.

Corollary A.3

Let $\mathcal{Q} = \{Q(x, a) : x \in \mathbb{R}^p, a \in \{1, \ldots, k\}, \|Q\|_\infty \leq M\}$. Define $\mathcal{Q}^{\max} = \{\max_a Q(x, a) : Q \in \mathcal{Q}\}$. Let P be a finitely discrete probability measure. Then

$$N(\varepsilon k M, \mathcal{Q}^{\max}, L_2(P)) \leq N(\varepsilon M, \mathcal{Q}, L_2(P))^k.$$

Proof

Since $(\max_a h(a) - \max_a h'(a))^2 \leq \max_a (h(a) - h'(a))^2$, inequality (34) holds for $c = 1$. The results now follow from Lemma A.1.

We also need the following lemma and its corollary:

Lemma A.4

Let $\mathcal{F}_1$ and F_2 be two function classes uniformly bounded in absolute value by M_1 and M_2, respectively. Define $\mathcal{F}_1 \cdot \mathcal{F}_2 = \{f_1 \cdot f_2 : f_i \in \mathcal{F}_i\}$. Then

$$N(2\varepsilon M_1 M_2, \mathcal{F}_1 \cdot \mathcal{F}_2, L_2(P)) \leq N(\varepsilon M_1, \mathcal{F}_1, L_2(P)) \cdot N(\varepsilon M_2, \mathcal{F}_2, L_2(P)).$$

Proof

Let $\|f_j - g_j\|_{P,2} \leq \varepsilon M_j$ where $f_j, g_j \in \mathcal{F}_j$, $j = \{1, 2\}$. Note that

$$\begin{aligned} \left\|f_1 \cdot f_2 - g_1 \cdot g_2\right\|_{P,2} &\leq \left\|f_1(f_2 - g_2)\right\|_{P,2} + \left\|g_2(f_1 - g_1)\right\|_{P,2} \\ &\leq M_1 \left\|f_2 - g_2\right\|_{P,2} + M_2 \left\|f_1 - g_1\right\|_{P,2} \leq 2M_1 M_2 \varepsilon. \end{aligned}$$

The result follows.

Corollary A.5

Let $\mathcal{G}$ be a function class uniformly bounded in absolute value by M. Define $\mathcal{G}^2 = \{g^2 : g \in \mathcal{G}\}$. Then

$$N(2\varepsilon M^2, \mathcal{G}^2, L_2(P)) \leq N(\varepsilon M, \mathcal{G}, L_2(P))^2.$$

Proof

Apply Lemma A.4 with $\mathcal{F}_1 = \mathcal{F}_2 = \mathcal{G}$.

We use the previous results to prove the following lemma:

Lemma A.6

Let

$$\mathcal{Q}_t \subset \{Q_t(x,a): x \in \mathbb{R}^{p_t}, a \in \{1,\ldots,k\}, \left\|Q_t\right\|_\infty \leq M\},$$
$$\mathcal{K} = \{K: K \text{ is monotone decreasing } K:[0,\tau] \mapsto [K_{\min}, 1]\},$$
$$\mathcal{R} = \left\{\frac{1}{K(t)}\left(r + \max_a Q_{t+1}(x,a) - Q_t(x,a)\right)^2 : r \in [0,\tau], Q_t \in \mathcal{Q}_t, \mathcal{Q}_{t+1} \in \mathcal{Q}_{t+1}, K \in \mathcal{K}\right\},$$

where $t \in 1, \ldots, T$ and $\mathcal{Q}_{T+1} = \{0\}$. Assume that the uniform entropy bound for each of the spaces $\mathcal{Q}_t$ (14) holds. Then

1 There are constants D' and W' such that $\log N(\varepsilon, \mathcal{R}, L_2(P)) \leq D'\left(\frac{1}{\varepsilon}\right)^{W'}$, where $W' = \max\{W, 1\}$.

2 For every $a > 0$ and $t > 0$,

$$P^*(\sup_{f \in \mathcal{R}} \left|Ef - \mathbb{E}_n f\right| > t) \leq C_a \exp\{C_b \sqrt{n} t^{U+\alpha} - 2nt^2\},$$

where $U = W'(6 - W')/(2 + W')$, the constants C_a and C_b depend only on D', W' and a, and where P^* is outer probability.

Proof

Let $W' = \max\{W, 1\}$. Note that uniform entropy bound (14) for the spaces $\mathcal{Q}_t$ holds also for W'. Note that by Corollary A.3, $\log N(\varepsilon M, \mathcal{Q}_t^{\max}, L_2(P)) \leq Dk^{W'+1}\left(\frac{1}{\varepsilon}\right)^{W'}$. Since $(x+y+z)^2 < 3(x^2+y^2+z^2)$, we can apply Lemma A.1 to the class

$$\mathcal{G} = \{r + \max_a Q_{t+1}(x,a) - Q_t(x,a): r \in [0,\tau], Q_t \in \mathcal{Q}_t, Q_{t+1} \in \mathcal{Q}_{t+1}\},$$

with $c = \sqrt{3}$ and $\varphi(x, y, z) = x + y + z$ to obtain $\log N(\sqrt{3}\varepsilon(2M+\tau), \mathcal{G}, L_2(P)) \leq (\tau + Dk^{W'+1} + D)\varepsilon^{-W'}$, where we used the fact that the segment $[0, \tau]$ can be covered by no more than $\tau/\varepsilon + 1$ balls of radius ε and that $\log(1 + \tau/\varepsilon) \leq \tau/\varepsilon$. By Corollary A.5, we have $\log N(2 \cdot 3\varepsilon(2M+\tau), \mathcal{G}^2, L_2(P)) \leq 2(\tau + Dk^{W'+1} + D)\left(\frac{1}{\varepsilon}\right)^{W'}$ or, equivalently,

$$\log N(\varepsilon M_1, \mathcal{G}^2, L_2(P)) \leq D_1\left(\frac{1}{\varepsilon}\right)^{W'},$$

where $M_1 = (2M + \tau)^2$ is a uniform bound for $\mathcal{G}^2$, and $D_1 = 2(\tau + Dk^{W'+1} + D)6^{-W'}$.

By Kosorok (2008, Lemma 9.11), $\log N(\varepsilon, \mathcal{K}, L_2(P)) \leq D_2\varepsilon^{-1}$ for some universal constant D_2 which is independent of the choice of probability measure P. By Corollary A.2,

$$\log N(\varepsilon K_{\min}^{-1}, \mathcal{K}^{-1}, L_2(P)) \leq D_2\left(\frac{1}{\varepsilon}\right).$$

Applying Lemma A.4 to $\mathcal{R} = \mathcal{K}^{-1} \cdot \mathcal{G}^2$, we obtain

$$\log N(\varepsilon K_{\min}^{-1}, M', \mathcal{R}, L_2(P)) \leq (D_1 + D_2)\left(\frac{1}{\varepsilon}\right)^{W'}.$$

Since this inequality holds for every finitely discrete probability measures P, assertion 1 is proved. The second assertion follows from van der Vaart and Wellner (1996, Theorem 2.14.10).

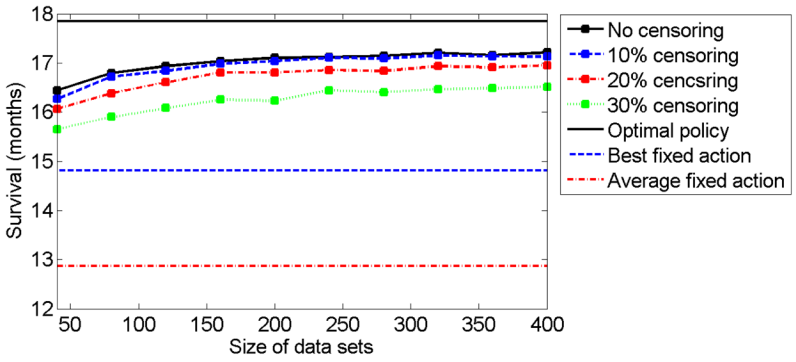

Fig 1.

The solid black curve, dashed blue curve, dot-dashed red curve, and dotted green curve correspond to the expected survival time (in months) for different data set sizes with no censoring, 10% censoring, 20% censoring, and 30% censoring, respectively. The expected survival time was computed as the mean of 400 repetitions of the simulation. The black straight line, blue dashed straight line, and the dot-dashed red straight line correspond to the expected survival times of the optimal policy, the best fixed treatment policy, and the average of the fixed treatment policies, respectively.

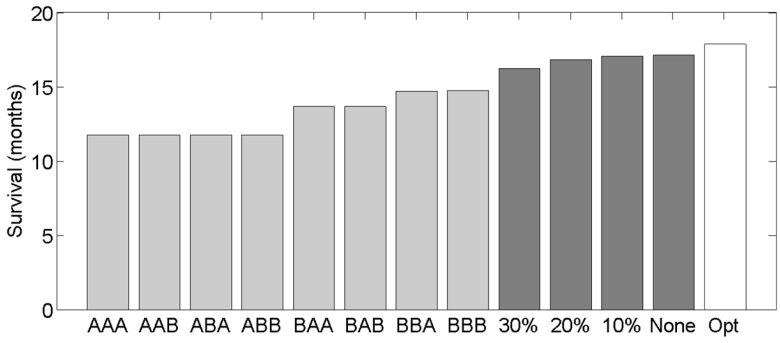

Fig 2.

The eight light gray bars represent the expected survival times for different fixed treatments where $A_1A_2A_3$ indicates the policy that chooses A_i at the i-th stage. The four dark gray bars represent the expected survival times for policy $\hat{\pi}$ obtained by the algorithm with no censoring, 10% censoring, 20% censoring, and 30% censoring. The white bar is the expected value of the optimal policy. The values of the fixed treatments and the optimal policy were computed analytically while the values of $\hat{\pi}$ are the means of 400 repetitions of the simulation on 200 trajectories.

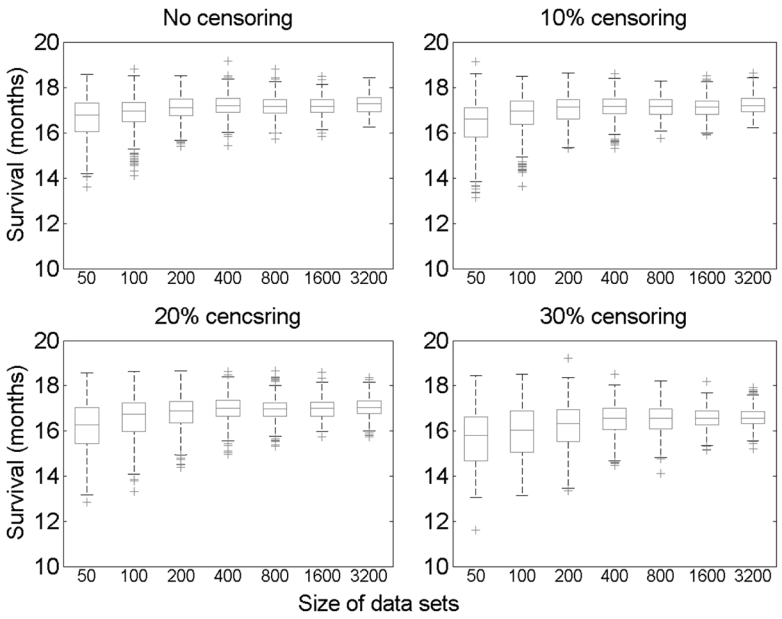

Fig 3.

Distribution of expected survival time (in months) for different data set sizes, with no censoring, 10% censoring, 20% censoring, and 30% censoring. Each box plot is based on 400 repetitions of the simulation for each given data set size and censoring percentage.

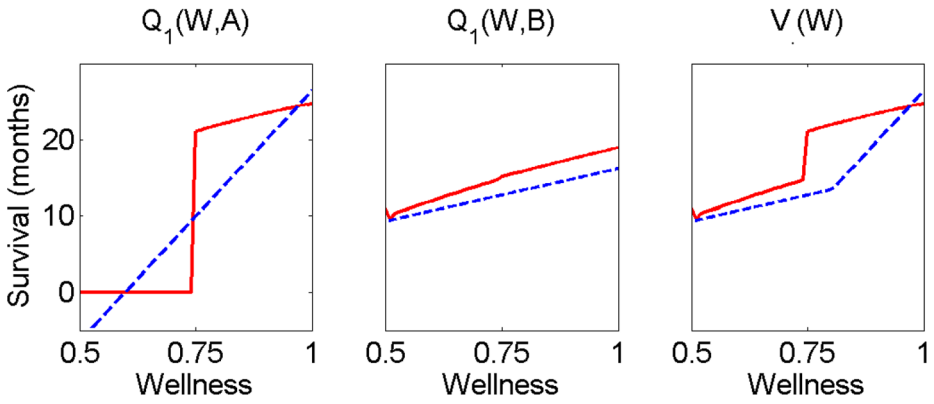

Fig 4.

The Q-functions computed by the proposed algorithm for a size 200 trajectory set. The left panel presents both the optimal Q-function (solid red curve) and the estimated Q-function (dashed blue curve) for different wellness levels and when treatment A is chosen. Similarly, the middle panel shows both Q-functions when treatment B is chosen. The right panel shows the optimal value function (solid red curve) and the estimated value function (dashed blue curve).

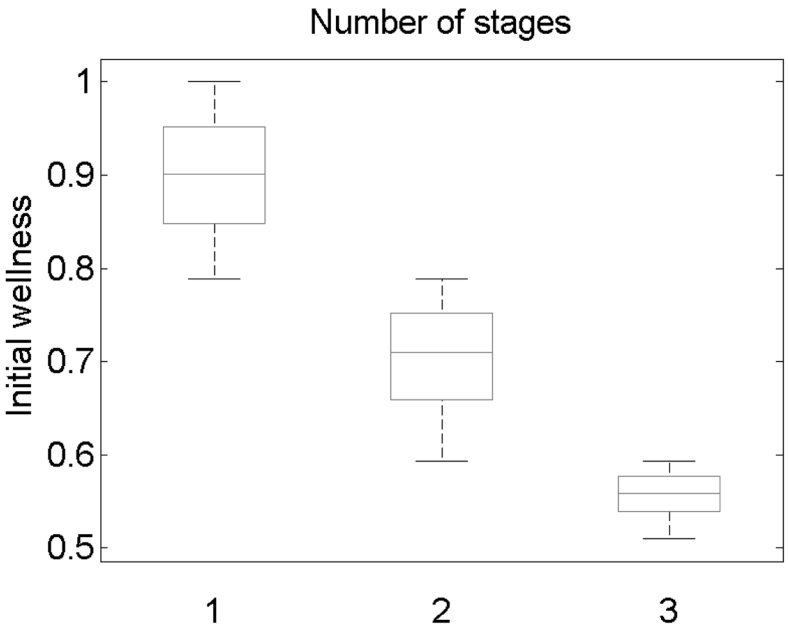

Fig 5.

The number of required treatments for patients that follow the policy $\hat{\pi}$, when no failure event occurs during the trial. The policy $\hat{\pi}$ was estimated from 100 trajectories. The results were computed using a size 100, 000 testing set.

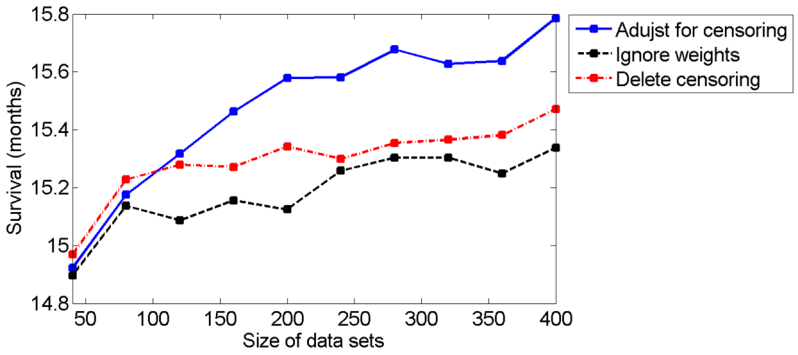

Fig 6.

The solid blue curve, dashed black curve, and dot-dashed red curve correspond to the expected survival times (in months) for different data set sizes, for the proposed algorithm, the algorithm that ignores the weights, and the algorithm that deletes all censored trajectories, respectively. The censoring variable follows the exponential distribution with 50% censoring on average. The expected survival time was computed as the mean of 400 repetitions of the simulation.