Chapter 7
Estimation of DTRs for Alternative Outcome Types

Up to this point, our development has focused entirely on the continuous outcome setting. In this chapter, we will turn our attention to the developments that have been made for estimating DTRs for more difficult outcome types including multi-component rewards, time-to-event data, and discrete outcomes. As we shall see, the range of approaches considered in previous chapters have been employed, but additional care and thought must be devoted to appropriately handling additional complexities in these settings.

7.1 Trading Off Multiple Rewards: Multi-dimensional and Compound Outcomes

Most DTR applications involve simple, univariate outcomes or utilities such as symptom scores or even survival times. However, it may be the case that a single dimension of response is insufficient to capture the patient experience under treatment. Recently, for example, Wang et al. (2012) conducted an analysis of a SMART-design cancer treatment study in which the outcome was taken to be a compound score numerically combining information on treatment efficacy, toxicity, and the risk of disease progression. The optimal DTR using the composite endpoint was found to differ from that using simpler endpoints based on a binary or ternary variable indicating treatment success.

Lizotte et al. (2010) considered an approach based on *inverse preference elicitation*. They proposed to find optimal regimes that can vary depending on how a new patient is willing to trade off different outcomes, such as whether he is willing to tolerate some side-effects for a greater reduction in disease symptoms. Specifically, they considered a situation where there were two possible outcomes of interest, R_1 and R_2, whose respective desirability could be described via a weighted sum $Y = \delta R_1 + (1-\delta)R_2$ for $\delta \in [0,1]$. In this situation, Q-functions may be modeled as a function of the two possible outcomes and δ; for example, a linear model for the Q-function might be represented using

B. Chakraborty and E.E.M. Moodie, *Statistical Methods for Dynamic Treatment Regimes*, Statistics for Biology and Health 76, DOI 10.1007/978-1-4614-7428-9_7,

$$Q_j^{opt}(H_j, A_j) = \delta(\beta_{j1}^T H_{j0} + \psi_{j1}^T H_{j1} A_j) + (1-\delta)(\beta_{j2}^T H_{j0} + \psi_{j2}^T H_{j1} A_j).$$

Estimates of $\beta_j = (\beta_{j1}, \beta_{j2})$ and $\psi_j = (\psi_{j1}, \psi_{j2})$ may be obtained by OLS by setting δ to 0 or 1 (Lizotte et al. 2010). This conceptualization of the outcome addresses an important issue for researchers who may wish to propose not a single DTR, but one which may be adapted not only to patient covariates but also to the relative value patients place on different outcomes. For example, Thall et al. (2002) provided an analysis where the response is taken to be a linear combination of the probability of complete remission and the probability of death as judged by a physician with expertise. It would be possible to use the approach of Lizotte et al. (2010) to either leave δ unspecified so that future "users" or "recipients" of the DTR (i.e. patients) could select their preferred weighting on the risks of remission versus death.

As noted by Almirall et al. (2012b), using a linear combination of outcomes as the final response may not in all circumstances be clinically meaningful, but may provide an important form of sensitivity analysis when outcome measures are subjective.

7.2 Estimating DTRs for Time-to-Event Outcomes with Q-learning

While much of the DTR literature has focused on continuous outcomes, research and analyses have been conducted for time-to-event data as well. Here, we briefly review some key developments.

7.2.1 Simple Q-learning for Survival Data: IPW in Sequential AFT Models

Huang and Ning (2012) used linear regression to fit *accelerated failure time* (AFT) models (Cox and Oaks 1984) in a Q-learning framework to estimate the optimal DTR in a time-to-event setting. Consider a two-stage setting, where patients may receive treatment in at least one and possibly two stages of a study. That is, all patients are exposed to some level of the treatment (where we include a control condition as a possible level of treatment) at the first stage. After the first stage of treatment, one of three possibilities may occur to a study participant: (1) the individual is cured by the treatment and does not require further treatment; (2) the individual experiences the outcome event, or (3) the individual requires a second stage of treatment, e.g. because of disease recurrence. Let Y denote the total follow-up time for an individual. If the individual is cured, he is followed until the end of the study and then censored so that Y is the time from the start of treatment to the censoring time; if he experiences the outcome event, Y is the time at which the event occurs. Further, let R denote the time from the initial treatment to the start

of the second stage treatment (assuming this to be the same as the time of disease recurrence), and let S denote the time from the start of the second stage treatment until end of follow-up (due to experiencing the event or the end of the study); then $Y = R + S$. Set $S = 0$ for those individuals who did not experience a second stage of treatment.

First, let us assume that there is no censoring. Then an AFT Q-learning algorithm for time-to-event outcomes proceeds much like that for continuous outcomes:

1. Stage 2 parameter estimation: Using OLS, find estimates $(\hat{\beta}_2, \hat{\psi}_2)$ of the conditional mean model $Q_2^{opt}(H_{2i}, A_{2i}; \beta_2, \psi_2)$ of the log-transformed time of follow-up from the start of the second stage, $\log(S_i)$, for those who experienced a second stage treatment.
2. Stage 2 optimal rule: By substitution, $\hat{d}_2^{opt}(h_2) = \arg\max_{a_2} Q_2^{opt}(h_2, a_2; \hat{\beta}_2, \hat{\psi}_2)$.
3. Stage 1 pseudo-outcome: Set $S_i^* = \max_{a_2} \exp(Q_2^{opt}(H_{2i}, a_2; \hat{\beta}_2, \hat{\psi}_2))$, $i = 1, \ldots, n$, which can be viewed as the time to event that would be expected under optimal second-stage treatment. Then calculate the pseudo-outcome,

$$\hat{Y}_{1i} = \begin{cases} Y_i \text{ if } S_i = 0 \\ R_i + S_i^* \text{ if } S_i > 0 \end{cases} \qquad i = 1, \ldots, n.$$

4. Stage 1 parameter estimation: Using OLS, find estimates

$$(\hat{\beta}_1, \hat{\psi}_1) = \arg\min_{\beta_1, \psi_1} \frac{1}{n} \sum_{i=1}^{n} \left(\log(\hat{Y}_{1i}) - Q_1^{opt}(H_{1i}, A_{1i}; \beta_1, \psi_1) \right)^2.$$

5. Stage 1 optimal rule: By substitution, $\hat{d}_1^{opt}(h_1) = \arg\max_{a_1} Q_1^{opt}(h_1, a_1; \hat{\beta}_1, \hat{\psi}_1)$.

In the presence of censoring, the regressions in steps 1 and 4 above can be performed with *inverse probability weighting* (IPW), where each subject is weighted by the inverse of the probability of not being censored. Because censoring time is a continuous measure, the probability of not being censored can be calculated from the estimated survival curve for censoring, e.g. by fitting a Cox proportional hazards model to estimate the distribution of the censoring times. Huang and Ning (2012) proved consistency and asymptotic normality of the regression parameters under a set of regularity conditions, illustrated good finite-sample performance of the methodology under varying degrees of censoring using a simulation study, and applied the methodology to analyze data from a study on the treatment of soft tissue sarcoma.

7.2.2 Q-learning with Support Vector Regression for Censored Survival Data

In Q-learning, the Q-functions need not always be modeled by linear models. In the RL literature, Q-functions had been modeled via regression trees or more sophisticated variations like *random forests* and *extremely randomized trees* (Ernst et al.

2005; Geurts et al. 2006; Guez et al. 2008) or via kernel-based regression (Ormoneit and Sen 2002). More recently in the DTR literature, Zhao et al. (2011) employed *support vector regression* (SVR) to model the Q-functions in the context of modeling survival time in a cancer clinical trial. These modern methods from the machine learning literature are often appealing due to their robustness and flexibility in estimating the Q-functions. Following Zhao et al. (2011), here we briefly present the SVR method to fit Q-functions.

Stepping outside the RL framework for a moment, consider a regression problem with the vector of predictors $x \in \mathbb{R}^m$ and the outcome $y \in \mathbb{R}$. Given the data $\{x_i, y_i\}_{i=1}^n$, the goal in SVR is to find a (regression) function $f : \mathbb{R}^m \to \mathbb{R}$ that closely matches the target y_i for the corresponding x_i. One of the popular loss functions is the so-called ε-insensitive loss function (Vapnik 1995), defined as: $\mathscr{L}(f(x_i), y_i) = (|f(x_i) - y_i| - \varepsilon)_+$, where $\varepsilon > 0$ and u_+ denotes the positive part of u. The ε-insensitive loss function ignores errors of size less than ε and grows linearly beyond that. Conceptually, this property is similar to that of the robust regression methods (Huber 1964); see Hastie et al. (2009, p. 435) for more details on this similarity, including a graphical representation.

In SVR, typically the regression function $f(\cdot)$ is assumed to take the form $f(x) = \theta_0 + \theta^T \Phi(x)$, where $\Phi(x)$ is a vector of non-linear basis functions (or, features) of the original predictor vector x. Thus, while the regression function employs a linear model involving the transformed features $\Phi(x)$, it can potentially become highly non-linear in the original predictor space, thereby allowing great flexibility and predictive power. It turns out that the problem of solving for unknown f is a convex optimization problem, and can be solved by quadratic programming using Lagrange multipliers (see, for example, Hastie et al. 2009, Chap. 12).

In the context of dynamic treatment regimes, the outcome of interest y (e.g. survival time from cancer) is often censored. The presence of censoring makes matters more complicated and the SVR procedure as outlined above cannot be used without modification. Shivaswamy et al. (2007) considered a version of SVR, without the ε-insensitive property, to take into account censored outcomes. Building on their work, Zhao et al. (2011) developed a procedure called ε-SVR-C (where C denotes censoring) that can account for censored outcomes and has the ε-insensitive property. Below we briefly present their procedure.

In general, we denote interval-censored survival (more generally, time-to-event) data by $\{x_i, l_i, u_i\}_{i=1}^n$, where l and u stand for the lower and upper bound of the interval under consideration. If a patient experiences the death event, then the corresponding observation is denoted by $\{x_i, y_i\}_{i=1}^n$ with $l_i = u_i = y_i$. Also, letting $u_i = +\infty$, one can easily construct a right-censored observation $\{x_i, l_i, +\infty\}$. Given the interval-censored data, consider the following loss function:

$$\mathscr{L}(f(x_i), l_i, u_i) = \max(l_i - \varepsilon - f(x_i), f(x_i) - u_i - \varepsilon)_+.$$

The shape of the loss function for both interval-censored data and right-censored data are displayed in Fig. 7.1.

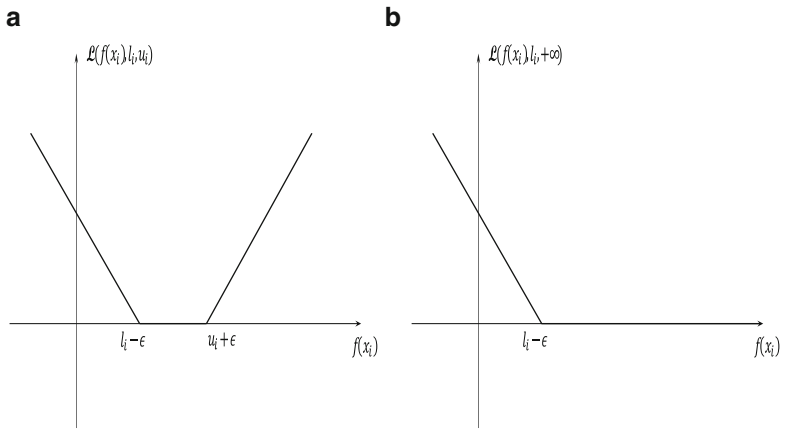

Fig. 7.1 ε-SVR-C loss functions for: (**a**) interval-censored data (*left panel*), and (**b**) right-censored data *(right panel*)

Defining the index sets $L = \{i : l_i > -\infty\}$ and $U = \{i : u_i < +\infty\}$, the ε-SVR-C optimization formulation is:

$$\begin{aligned}\min_{\theta,\theta_0,\xi,\xi'} \frac{1}{2}||\theta||^2 + C_E\Big(\sum_{i\in L}\xi_i + \sum_{i\in U}\xi_i'\Big), \quad \text{subject to}\\ (\theta_0 + \theta^T\Phi(x_i)) - u_i \le \varepsilon + \xi_i,\ i \in U;\\ l_i - (\theta_0 + \theta^T\Phi(x_i)) \le \varepsilon + \xi_i',\ i \in L;\\ \xi_i \ge 0,\ i \in L;\\ \xi_i' \ge 0,\ i \in U.\end{aligned}$$

In the above display, ξ_i and ξ_i' are the so-called *slack variables* and C_E is the cost of error. By minimizing the regularization term $\frac{1}{2}||\theta||^2$ as well as the training error $C_E\Big(\sum_{i\in L}\xi_i + \sum_{i\in U}\xi_i'\Big)$, the ε-SVR-C algorithm can avoid both overfitting and underfitting of the training data.

Interestingly, the solution depends on the basis function Φ only through inner products $\Phi(x_i)^T\Phi(x_j)$, $\forall i, j$. In fact, one need not explicitly specify the basis function Φ; it is enough to specify the *kernel function* $K(x_i,x_j) = \Phi(x_i)^T\Phi(x_j)$. One popular choice of K used by Zhao et al. (2011) is the Gaussian (or radial basis) kernel, given by $K(x_i,x_j) = \exp(-\gamma||x_i - x_j||^2)$. Thus the above optimization problem is equivalent to the following dual problem:

$$\begin{aligned}\min_{\lambda,\lambda'} \frac{1}{2}(\lambda-\lambda')^T K(x_i,x_j)(\lambda-\lambda') - \sum_{i\in L}(l_i-\varepsilon)\lambda_i' + \sum_{i\in U}(u_i+\varepsilon)\lambda_i,\\ \text{subject to}\\ \sum_{i\in L}\lambda_i' - \sum_{i\in U}\lambda_i = 0,\ 0 \le \lambda_i, \lambda_i' \le C_E,\ \ i = 1,\ldots,n.\end{aligned}$$

The tuning parameters γ (in the definition of K) and C_E are obtained by cross-validation to achieve good performance. Once the above formulation is solved to find the optimal values of λ_i and λ_i', say $\hat{\lambda}_i$ and $\hat{\lambda}_i'$, the regression function is given by $\hat{f}(x) = \sum_{i=1}^{n}(\hat{\lambda}_i' - \hat{\lambda}_i)K(x_i,x) + \hat{\theta}_0$. Due to the nature of the constraints in the above optimization problem, typically only a subset of values of $(\hat{\lambda}_i' - \hat{\lambda}_i)$ are non-zero, and the associated data points are called the *support vectors*.

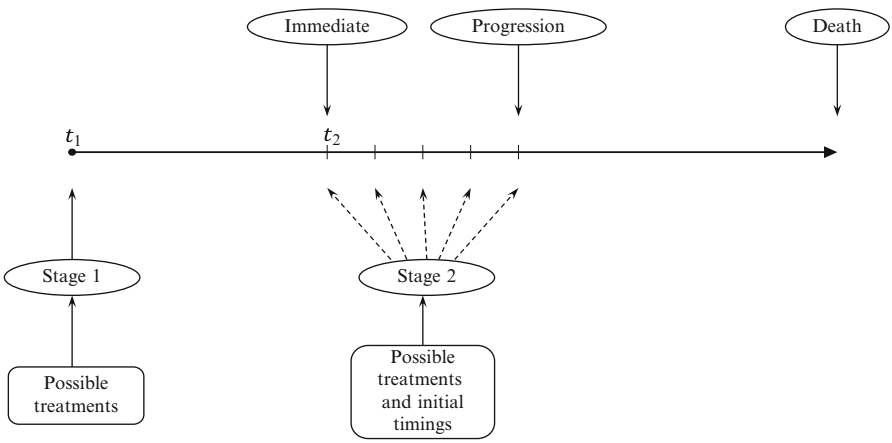

Fig. 7.2 Treatment plan and therapy options for advanced non-small cell lung cancer in a hypothetical SMART design

Zhao et al. (2011) implemented Q-learning in conjunction with the ε-SVR-C method described above in the context of a hypothetical two-stage SMART for treating advanced non-small cell lung cancer; see Fig. 7.2 for a schematic. In addition to the complexity of the problem of selecting optimal stage-1 and stage-2 treatments, another goal was to determine the optimal time to initiate the stage-2 treatment, either immediately or delayed, that would yield the longest overall survival time. Let t_1 and t_2 denote the time-points where the first and second stage treatment decisions are made, respectively. Let the time to disease progression, after initiation of the stage-1 treatment (chemotherapy), be denoted by T_P (for simplicity, it is assumed that $T_P \geq t_2$ with probability 1). Let T_M denote the targeted time after t_2 of initiating the stage-2 treatment. The actual time to initiate the stage-2 treatment is $(t_2 + T_M) \wedge T_P$. At the end of first-stage therapy, i.e. at time t_2, clinicians make a decision about the target start time T_M. Let T_D denote the time of death from the start of therapy (t_1), i.e. the overall survival time. Note that this scenario is more complex than that of the previous section; in the simpler setting of Huang and Ning (2012), $R = (t_2 + T_M) \wedge T_P$ and $S = T_D - R$ or, in the presence of censoring, S will be the time on study following initiation of second treatment (total time minus R).

Acknowledging the possibility of right censoring, denote the patient's censoring time by C and indicator of the event (i.e. of not being censored) by $\delta = \mathbb{I}[T_D \leq C]$. Assume that the censoring is independent of both the death time and the patient

covariates. For convenience, define $T_1 = T_D \wedge t_2$ and $Y_D = \mathbb{I}[T_D \wedge C \geq t_2]$, and also $T_2 = (T_D - t_2)\mathbb{I}[T_D \geq t_2] = (T_D - t_2)\mathbb{I}[T_1 = t_2]$ and $C_2 = (C - t_2)\mathbb{I}[C \geq t_2]$. Note that $T_D = T_1 + T_2$, where T_1 is the time of life lived in $[t_1, t_2]$ and T_2 is the time of life lived after t_2.

As in previous chapters, let H_1 and H_2 denote the histories (e.g. current and past covariates, and also past treatments) available at first and second stage respectively. Also, let A_1 and A_2 denote the treatment choices at the two stages. In this study, the treatment decision at the second stage also involves an initiation time T_M, as discussed above. Thus the stage-2 treatment is two-dimensional, denoted compactly as (A_2, T_M). Define the optimal Q-functions for the two stages as follows:

$$\begin{aligned} Q_2^{opt}\big(H_2, (A_2, T_M)\big) &= E\big[T_2 \big| H_2, (A_2, T_M)\big], \\ Q_1^{opt}(H_1, A_1) &= E\big[T_1 + \mathbb{I}[T_1 = t_2] \max_{(A_2, T_M)} Q_2^{opt}\big(H_2, (A_2, T_M)\big) \big| H_1, A_1\big]. \end{aligned}$$

In case of known Q-functions, the optimal DTR (d_1^{opt}, d_2^{opt}), using a backwards induction argument, would be

$$\begin{aligned} d_2^{opt}(h_2) &= \arg\max_{(a_2, T_M)} Q_2^{opt}(h_2, (a_2, T_M)), \\ d_1^{opt}(h_1) &= \arg\max_{a_1} Q_1^{opt}(h_1, a_1). \end{aligned}$$

When the Q-functions are unknown, they are estimated using suitable models. In the present development, censored outcomes $(T_1 \wedge C, \delta_1 = \mathbb{I}[T_1 \leq C])$ and $(T_2 \wedge C_2, \delta_2 = \mathbb{I}[T_2 \leq C_2])$ are used at both stages. The exact algorithm to perform Q-learning with ε-SVR-C for censored survival data is as follows:

1. For those individuals with $Y_D = 1$ (i.e. those who actually go on to the second stage of treatment), perform right-censored regression using ε-SVR-C of the censored outcome $(T_2 \wedge C_2, \delta_2)$ on the stage-2 variables $(H_2, (A_2, T_M))$ to obtain $\hat{Q}_2^{opt}$.
2. Construct the pseudo-outcome

$$\hat{T}_D = T_1 + \mathbb{I}[T_1 = t_2] \max_{(A_2, T_M)} \hat{Q}_2^{opt}(H_2, A_2, T_M) = T_1 + \mathbb{I}[T_1 = t_2]\hat{T}_2 = T_1 + Y_D\hat{T}_2.$$

3. In fitting Q_1^{opt}, the pseudo-outcome $\hat{T}_D$ is assessed through the censored observation $(\tilde{X}, \tilde{\delta})$, with $\tilde{X} = T_1 \wedge C + Y_D\hat{T}_2 = \hat{T}_D \wedge \tilde{C}$ and $\tilde{\delta} = \mathbb{I}[\hat{T}_D \leq \tilde{C}]$, where $\tilde{C} = C\mathbb{I}[C < t_2] + \infty\mathbb{I}[C_2 \geq t_2]$. Perform ε-SVR-C of $(\tilde{X}, \tilde{\delta})$ on (H_1, A_1) to obtain $\hat{Q}_1^{opt}$.

Once the Q-functions are fitted, the estimated optimal DTR is given by $(\hat{d}_1^{opt}, \hat{d}_2^{opt})$, where the stage-specific optimal rules are given by

$$\begin{aligned} \hat{d}_2^{opt}(h_2) &= \operatorname*{argmax}_{(a_2, T_M)} \hat{Q}_2^{opt}(h_2, (a_2, T_M)), \\ \hat{d}_1^{opt}(h_1) &= \operatorname*{argmax}_{a_1} \hat{Q}_1^{opt}(h_1, a_1). \end{aligned}$$

In the ε-SVR-C steps of the Q-learning algorithm, the tuning parameters C_E and γ are chosen via cross validation over a grid of values. Zhao et al. (2011) reported robustness of the procedure to relatively small values of ε; they set its value at 0.1 in their simulation study.

Zhao et al. (2011) evaluated the above method of estimating the optimal DTR with survival-type outcome in an extensive simulation study. In short, they considered a generative model, the parameters of which could be easily tweaked to reflect four different clinical scenarios resulting in four different optimal regimes. They generated data on 100 virtual patients from each of the 4 clinical scenarios, thus a total of 400 virtual patients. Then the optimal regime was estimated via Q-learning with ε-SVR-C. For evaluation purposes, an independent test sample of size 100 per clinical scenario (hence totaling 400) was also generated. Outcomes (overall survival) for these virtual test patients were evaluated for the estimated optimal regime as well as all possible (12) fixed regimes, using the generative model. Furthermore, they repeated the simulations ten times for the training sample (each of size 400). Then ten different estimated optimal regimes from these ten training samples were applied to the same test sample (of size 400) mentioned earlier. All the results for each of the 13 treatment regimes (12 fixed, plus the estimated optimal) were averaged over the 400 test patients. It was found that the true overall survival was substantially higher for the estimated optimal regime than any of the 12 fixed regimes. They also conducted additional simulations to check the sensitivity of the procedure to the sample size. It was found that for sample sizes $\geq$100, the procedure is very reliable in selecting the optimal regime.

7.3 Q-learning of DTRs for Discrete Outcomes

Moodie et al. (2013) recently tackled the challenging problem of Q-learning for discrete-valued outcomes, and took a less parametric approach to modeling the Q-functions by using generalized additive models (GAMs). Generalized additive models provide a user-friendly means to introducing greater flexibility in modeling the relationship between an outcome and covariates. GAMs are treated as penalized regression splines with different smoothing parameters allowed for each covariate, where the degree of smoothing is selected by generalized cross-validation (Wood 2006, 2011). The automatic parsimony that the approach ensures helps to control the dimensionality of the estimation problem, an important feature in the DTR setting where the covariate space is potentially very large.

Suppose we are in a setting where the outcome at the final stage is discrete, and there are no intermediate rewards. The outcome could represent, for instance, a simple indicator of success such as maintenance of viral load below a given threshold over the course of a study (a binary outcome), or the number of emergency room visits in a given period (a count, possibly Poisson-distributed). When the outcome Y is discrete, the Q-learning procedure must be adapted to respect the constraints on the outcome, for example, Y is bounded in $[0,1]$, or Y is

non-negative. By definition, in a two-stage setting, we have $Q_2^{opt}(H_2,A_2) = E\left[Y\middle|H_2,A_2\right]$ at the final interval. A reasonable modeling choice would be to consider a generalized linear model (GLM). For instance, for a Bernoulli utility, we might choose a logistic model of the form $E\left[Y\middle|H_2,A_2\right] = \text{expit}\left(\beta_j^T H_{j0} + (\psi_j^T H_{j1})A_j\right)$, where $\text{expit}(x) = \exp(x)/(1+\exp(x))$ is the inverse-logit function. Similarly, for a non-negative outcome, we might choose a Poisson family GLM with the canonical link. The key is to choose a link function that is strictly increasing (or decreasing), since this allows maximization of the second-stage Q-function by a maximization of the linear specification in the mean. For example, in the binary outcome setting, since the inverse-logit function is strictly increasing, $\text{expit}\left(\beta_j^T H_{j0} + (\psi_j^T H_{j1})A_j\right)$ can be maximized by maximizing its argument, $\beta_j^T H_{j0} + (\psi_j^T H_{j1})A_j$. Therefore

$$Q_1^{opt}(H_1,A_1;\beta_1,\psi_1) = \max_{a_2} Q_2^{opt}(H_{2i},a_2;\hat{\beta}_2,\hat{\psi}_2) = \text{expit}\left(\hat{\beta}_2^T H_{20,i} + |\hat{\psi}_2^T H_{21,i}|\right),$$

which is bounded by [0,1]. As in the continuous utility setting, the optimal regime at the first interval is defined by

$$\hat{d}_1^{opt}(h_1) = \arg\max_{a_1} Q_1^{opt}(h_1,a_1;\hat{\beta}_1,\hat{\psi}_1).$$

Continuing with the binary outcome example, we have

$$\arg\max_{a_1} Q_1^{opt}(h_1,a_1;\hat{\beta}_1,\hat{\psi}_1) = \arg\max_{a_1} \text{logit}\left(Q_1^{opt}(h_1,a_1;\hat{\beta}_1,\hat{\psi}_1)\right)$$

since the logit function is strictly increasing. We may therefore model the logit of $Q_1^{opt}(H_1,A_1;\beta_1,\psi_1)$ rather than the Q-function itself to determine the optimal DTR.

The Q-learning algorithm for a discrete outcome consists of the following steps:

1. Interval 2 parameter estimation: Using GLM regression with a strictly increasing link function, $f(\cdot)$, find estimates $(\hat{\beta}_2,\hat{\psi}_2)$ of the conditional mean model for the outcome Y, $Q_2^{opt}(H_{2i},A_{2i};\beta_2,\psi_2)$.
2. Interval 2 optimal rule: Set $\hat{d}_2^{opt}(h_2) = \arg\max_{a_2} Q_2^{opt}(h_2,a_2;\hat{\beta}_2,\hat{\psi}_2)$.
3. Interval 1 pseudo-outcome: Set

$$\tilde{Y}_{1i} = \max_{a_2} f(Q_2^{opt}(H_{2i},a_2;\hat{\beta}_2,\hat{\psi}_2)),\ i = 1,\ldots,n.$$

4. Interval 1 parameter estimation: Using ordinary least squares regression, find estimates

$$(\hat{\beta}_1,\hat{\psi}_1) = \arg\min_{\beta_1,\psi_1} \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_{1i} - Q_1^{opt}(H_{1i},A_{1i};\beta_1,\psi_1)\right)^2.$$

5. Interval 1 optimal rule: Set $\hat{d}_1^{opt}(h_1) = \arg\max_{a_1} Q_1^{opt}(h_1,a_1;\hat{\beta}_1,\hat{\psi}_1)$.

The estimated optimal DTR using Q-learning is given by $(\hat{d}_1, \hat{d}_2)$. In a binary outcome scenario, note that unlike in the continuous utility setting, the pseudo-outcome, $\tilde{Y}_{1i}$, does not represent the (expected) value of the second-interval Q-function under the optimal treatment but rather a transformation of that expected outcome.

We briefly consider a simulation study. The data for treatments (A_1, A_2), and covariates (C_1, O_1, C_2, O_2) were generated as in Sect. 3.5. We considered three outcome distributions: normal, Bernoulli, and Poisson, and two forms of the relationship between the outcome and the variables C_1 and C_2. The first setting corresponds to Scenario C of Sect. 3.5 (normal outcome, Q-functions linear in covariates); the second varies only in that a quadratic terms for C_1 and C_2 are included in the mean model. Similarly, settings three and four correspond to a Bernoulli outcome with Q-functions that are, respectively, linear and quadratic in C_1 and C_2, and the final pair of settings correspond to a Poisson outcome with Q-functions that are, respectively, linear and quadratic in the covariates. Results are presented in Table 7.1.

Overall, we observe very good performance of both the linear (correct) specification and the GAM specification of the Q-function when the true confounder-outcome relationship is linear: estimators are unbiased, and the use of the GAM for the Q-function exhibits reasonably variability even for the smaller sample size of 250. In fact the variability of the estimator resulting from a GAM for the Q-function is as low as the linear model-based estimator for the normal and Poisson outcomes, implying there is little cost for the additional flexibility in the cases. When the dependence of the utility on the confounding variables is quadratic, only the decision rule parameters resulting from a GAM for the Q-function exhibits little or no bias and good coverage rates.

Thus, it appears that Moodie et al. (2013) have taken a modest but promising step on the path to a more fully generalized Q-learning algorithm, with the consideration of a flexible, spline-based modeling approach for discrete outcomes. The next step of adapting Q-learning to allow discrete interval-specific outcomes is challenging, and remains an open problem.

7.4 Inverse Probability Weighted Estimation for Censored or Discrete Outcomes and Stochastic Treatment Regimes

Some of the seminal work in developing MSMs for DTR estimation was performed in a survival context, using inverse probability weighting combined with pooled logistic regression to approximate a Cox model for the estimation of the hazard ratio parameters (Hernán et al. 2006; Robins et al. 2008). The methods are gaining popularity in straightforward applications examining, for example, when to initiate dialysis (Sjölander et al. 2011) or antiretroviral therapy (Shepherd et al. 2010). These methods require little adaptation to the algorithm described in Sect. 5.2.2: as with continuous outcomes, data-augmentation is undertaken to create replicates of individuals that are compatible with each regime of interest. The only step that dif-

Table 7.1 Comparison of the performance Q-learning for normal, Bernoulli, and Poisson outcomes when the true Q-function is either linear or quadratic in the covariates: bias, Monte Carlo variance (MC var), Mean Squared Error (MSE) and coverage of 95 % bootstrap confidence intervals (Cover) of the first interval decision rule parameter ψ_{10}. Bias, variance, and MSE are each multiplied by 10.

Adjustment	$n = 250$				$n = 1{,}000$			
method	Bias	MC var	MSE	Cover	Bias	MC var	MSE	Cover
Normal outcome, Q-functions linear in covariates								
None	10.03	0.35	10.41	0.0	10.12	0.09	10.32	0.0
Linear	0.02	0.08	0.08	94.1	0.00	0.02	0.02	93.0
GAM	0.02	0.08	0.08	94.4	0.00	0.02	0.02	93.6
Normal outcome, Q-functions quadratic in covariates								
None	18.18	16.30	4.935	68.1	18.92	4.31	40.11	10.8
Linear	29.64	20.53	108.38	37.9	31.42	4.72	103.46	0.1
GAM	0.21	1.49	1.50	95.2	−0.11	0.40	0.40	92.7
Bernoulli outcome, Q-functions linear in covariates								
None	8.65	1.57	8.97	13.7	8.45	0.19	7.32	0.0
Linear	0.20	1.98	1.98	94.9	0.00	0.28	0.28	95.1
GAM	0.81	4.25	4.25	97.2	0.00	0.28	0.28	95.8
Bernoulli outcome, Q-functions quadratic in covariates								
None	3.77	0.65	2.07	64.8	3.71	0.15	1.53	10.8
Linear	1.54	0.87	1.11	92.5	1.56	0.20	0.44	79.7
GAM	0.06	2.63	2.63	97.2	−0.11	0.32	0.32	97.0
Poisson outcome, Q-functions linear in covariates								
None	8.97	0.70	8.74	5.6	9.49	0.23	9.23	0.0
Linear	0.14	0.11	0.11	93.9	0.14	0.02	0.03	93.8
GAM	0.13	0.11	0.11	95.7	0.14	0.02	0.03	94.5
Poisson outcome, Q-functions quadratic in covariates								
None	4.39	0.19	2.12	15.4	4.32	0.04	1.91	0.0
Linear	−1.01	0.27	0.38	90.1	−1.06	0.07	0.19	72.6
GAM	0.00	0.28	0.28	96.7	0.14	0.64	0.65	94.6

fers is the outcome regression model, which is adapted to the outcome type, using, for example a weighted Cox model or a weighted pooled logistic regression rather than weighted linear regression.

A separate but closely related body of work has focused on survival data primarily in two-phase cancer trials. In the trials which motivated the statistical developments, cancer patients were randomly assigned to one of several initial therapies and, if the initial treatments successfully induced remission, the patient was randomized to one of several maintenance therapies. A wide collection of methods have been developed in this framework, including weighted Kaplan-Meier censoring survivor curves and mean-restricted survival times (Lunceford et al. 2002), an improved estimator for the survival distribution which was shown to be the most efficient among regular, asymptotically linear estimators (Wahed and Tsiatis 2004, 2006). Log-rank tests and sample size calculations have since been developed (Feng and Wahed 2009). While these methods do address estimation of a dynamic regime of the form "what is the best initial treatment? what is the best subsequent treatment if the initial treatment fails?", these methods are typically used to select

from among a small class of initial and maintenance treatment pairs, and have not been developed to select an optimal threshold from among a potentially large list of values.

The general MSM framework for DTR estimation has been further adapted to handle stochastic treatment assignment rules. For example, Cain et al. (2010) considered treatment rules which allowed for a grace period of m months in the timing of treatment initiation, i.e. a rule of the form "initiate treatment within m months of covariate O crossing threshold ψ" rather than "initiate treatment when covariate O crosses threshold ψ".

7.5 Estimating a DTR for a Binary Outcome Using a Likelihood Approach

Thall and colleagues have considered DTRs in several cancer treatment settings, where the typical treatment paradigm is "play the winner, drop the loser" (Thall et al. 2000): a patient given an initial course of a treatment will continue to receive that treatment if it is deemed to be sufficiently successful (e.g. due to partial tumor shrinkage or partial remission), will be switched to a maintenance therapy or follow-up if completely successful, and will be switched to an alternative treatment (sometimes referred to as a salvage therapy) if the initial treatment is unsuccessful. The definition of success on a particular course of treatment may depend on which course it is. For example, in prostate cancer, a success on the first course of treatment requires a decrease of at least 40 % in the cancer biomarker prostate-specific antigen (PSA) from baseline, while success in the second course requires a decrease of at least 80 % in PSA from the baseline value (and, in both cases, no evidence of disease progression).

In a prostate cancer treatment trial, Thall et al. (2000) took a parametric approach to estimating the best sequence of treatments with the goal of maximizing the probability of successful treatment, where success is a binary variable. Four treatment courses were considered. Patients were randomized to one of the four treatments, and if treatment failed, randomized to one of the remaining three options. That is, $\mathscr{A}_1 = \{1,2,3,4\}$ and $\mathscr{A}_2 = \mathscr{A}_1 \setminus a_1$ (where a_1 is the treatment actually given at the first stage). A patient was switched from a treatment after the first failure, or deemed to have had a successful therapy following two successful courses of the same treatment. Thus, the trial can be viewed as a two-stage trial in which patients can have at least one and at most two courses of treatment in the first stage, and at most two courses of treatment in the second stage for a total two to four courses of treatment.

The optimizing criterion for determining the best DTR was the probability of successful therapy. That is, the goal was to maximize $\xi(a,a') = \xi_a + (1-\xi_a)\xi_{a'|a}$, where ξ_a is the probability of a patient success in the first two courses with initial treatment a and $\xi_{a'|a}$ is the probability that the patient has two successful courses with treatment a' following initial (unsuccessful) treatment with a, i.e. under treatment strategy (a,a'). Parametric conditional probability models were posited to

obtain estimates of $\xi(a,a')$ that were allowed to depend on the patient's state and treatment history. For example, letting Y_j take the value 1 if a patient experiences successful treatment on the jth course and 0 otherwise, patient outcomes through the first two courses of therapy can be characterized by the following probabilities:

$$\begin{aligned}\theta_1(a) &= P(Y_1 = 1|A_1 = a)\\ \theta_2(1;(a,a)) &= P(Y_2 = 1|Y_1 = 1, A_1 = A_2 = a)\\ \theta_2(0;(a',a)) &= P(Y_2 = 1|Y_1 = 0, A_1 = a', A_2 = a)\end{aligned}$$

which gives $\xi_a = \theta_1(a)\theta_2(1;(a,a))$. Logistic regression models were proposed for the above probabilities, i.e. logit(θ_j) were modeled as linear functions of treatment and covariate histories for each of the j courses of treatment. These probability models can be extended to depend on state variables such as initial disease severity as well. Once all these models are fitted, one can pick the best DTR, i.e. the best treatment pair (a,a') that maximizes the overall success probability $\xi(a,a')$.

7.6 Discussion

In this chapter, we have considered the estimation of DTRs for a variety of outcome types, including multi-dimensional continuous outcomes, time-to-event outcomes in the presence of censoring, as well as discrete outcomes. Methods used in the literature for such data include Q-learning, marginal structural models, and a fully parametric, likelihood-based approach. In the context of Q-learning, modeling of time-to-event data has been accomplished using accelerated failure time models (with censoring handled by inverse probability weighting) and using the less parametric approach of support vector regression. For discrete outcomes, Q-learning has also been combined with generalized additive models selected by generalized cross-validation, with promising results. The MSM approach has been implemented for discrete failure times only, but can easily be used in a continuous-time setting using a marginal structural Cox model. G-estimation can also be employed assuming an AFT (see Mark and Robins 1993; Hernán et al. 2005) to estimate DTRs, however the approach remains under-utilized, perhaps because of the relative lack of standard software with which it can be implemented.