Q-Learning with Censored Data

Yair Goldberg, Michael Kosorok

Department of Biostatistics, University of North Carolina

March 21, 2025

Motivation

Problem Statement

- The problem framework originates from cancer research.
- Finding optimal dynamic treatment regimes that lead to longer survival times.

Statistical challenges

- Incorporate information accrued over time into the decision rule.
- Optimize long-term outcomes rather than short-term responses.
- Data is subject to censoring.*
- The number and timing of treatments are flexible.*

Proposed Method

Develop a methodology for a multistage-decision problem with flexible number of stages in which the rewards are survival times that are subject to censoring.

- Q-learning enables both accrual of information and incorporation of long-term treatment effects.
- Translate to an auxiliary problem with a fixed number of stages.
- Using inverse-probability-of-censoring weighting (IPCW) to adjust for censoring

Review of Q-learning

- Optimal policy π^* with value function $V^*_{t+1}(s_{t+1}, a_t)$:

$$\pi^*_t(s_t, a_{t-1}) = \arg\max_{a_t} E\left[R_t + V^*_{t+1}(S_{t+1}, A_t) \mid S_t = s_t, A_t = a_t\right],$$

$$V^*_{t+1}(s_{t+1}, a_t) = E_{\pi^*}\left[\sum_{i=t+1}^{T} R_i \mid S_{t+1} = s_{t+1}, A_t = a_t\right]$$

- Optimal time-dependent Q-function

$$\begin{aligned} Q^*_t(s_t, a_t) &= E\left[R_t + V^*_{t+1}(S_{t+1}, A_t) \mid S_t = s_t, A_t = a_t\right] \\ &= E\left[R_t + \max_{a_{t+1}} Q^*_{t+1}(S_{t+1}, A_t, a_{t+1}) \mid S_t = s_t, A_t = a_t\right] \end{aligned}$$

Setup

Trajectory sequence: $\{S_1, A_1, \ldots, A_T, S_{T+1}\}$

- State $\mathbf{S}_t = (Z_t, R_{t-1})$
- R_{t-1}: length of the interval between A_{t-1} and A_t
- Z_t: covariates at the beginning of stage, or $Z_t = \emptyset$ if failure happens during stage t

Consider a flexible number of stages and censoring: prevent using Q-learning

- τ: maximal follow-up time
- $\bar{T} \in \{1, \ldots, T\}$: random number of stages truncated by failure/censoring
- $C \in [0, \tau]$: censoring time; $\delta_t \in \{0, 1\}$: censored indicator for stage t

Auxiliary Problem: Applicable for Q-learning

Complete all trajectories to full length

- assume failure at stage t
- for $1 \leq j \leq t$: $S'_j = S_j, A'_j = A_j$
- for $t+1 \leq j \leq T+1$: $S'_j = (\emptyset, 0)$, random draw A'_j

Truncated survival time by τ

- assume survival time exceeds τ from stage t
- $S'_{t+1} = (\emptyset, \tau - \sum_{i=1}^{t-1} R_i)$
- for $1 \leq j \leq t$: $S'_j = S_j, A'_j = A_j$,
- for $t+1 \leq j \leq T+1$ (if $t < T$): $S'_j = (\emptyset, 0)$, random draw A'_j

Relate Auxiliary Problem to Original

The expected truncated-by-τ survival time for a policy π in the original problem:

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

The value function V_π in the auxiliary problem:

$$V_\pi(s_0) \text{ where } V_{\pi,t}(s_t, a_{t-1}) = E_\pi\left[\sum_{i=t}^{T} R_i \mid S_t = s_t, A_t = a_t\right]$$

Decompose Expectation

Decomposing the expectations depends on both the terminal stage and whether the sum of rewards is greater than or equal to τ.

$$E_{0,\pi}\left[\left(\sum_{t=1}^{\bar{T}} R_t\right) \wedge \tau \middle| S_1 = s_o\right] = \sum_{t=1}^{T} \int_{F_t} \left(\sum_{i=1}^{t} r_i\right) f_{t+1,\pi}(s_{t+1}, a_t) d(s_{t+1}, a_t) + \tau \sum_{t=1}^{T} P_{0,\pi}(G_t)$$

$$V_\pi(s_o) = \sum_{t=1}^{T} \int_{F'_t} \left(\sum_{i=1}^{T} r_i\right) f'_{T+1,\pi}(s_{T+1}, a_T) d(s_{T+1}, a_T) + \tau \sum_{t=1}^{T} P_\pi(G'_t)$$

where G_t is the set of trajectories that the sum of rewards is greater than τ at the terminal stage, and G'_t is the set of modified trajectories.

Likelihood for trajectory

- State conditional distribution: $f_t(S_t|\mathbf{S}_{t-1},\mathbf{A}_{t-1})$
- Likelihood under policy π

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

- Conditional distribution f' for modified trajectories

$$f'_t(s'_t|s'_{t-1},a'_{t-1}) = \begin{cases} f_t((z'_t,r'_t)|s'_{t-1},a'_{t-1}) & z'_{t-1}\neq\emptyset, \sum_{i=1}^{t} r'_i < \tau \\ \int_{G'_{z'_t}} f_t((z'_t,r_t)|s'_{t-1},a'_{t-1})dr_t & z'_{t-1}\neq\emptyset, \sum_{i=1}^{t} r'_i = \tau \\ 1_{s'_t=(\emptyset,0)} & z'_{t-1}=\emptyset \end{cases}$$

where $G'_{z'_t} = \{(z'_t,r_t) : \sum_{i=1}^{t} r_i \geq \tau\}$

Likelihood for trajectory

For all π, the following equalities hold true:

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

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

The results obtained for the auxiliary problem can be translated into results regarding the original problem with flexible number and timing of stages.

Algorithm

- Q-learning without censoring:

$$\widehat{Q}_t = \arg\min_{Q_t} \mathbb{E}_n \left[\left(R_t + \max_{a_{t+1}} \widehat{Q}_{t+1}(S_{t+1},(A_t,a_{t+1})) - Q_t(S_t,A_t) \right)^2 \right],$$

- Q-learning with censoring:

$$\widehat{Q}_t = \mathbb{E} \left[\left(R_t + \max_{a_{t+1}} Q^*_{t+1}(S_{t+1},(A_t,a_{t+1})) - Q_t(S_t,A_t) \right)^2 \frac{\delta_t}{S_C\left(\sum_{i=1}^{t'} R_i\right)} \right]$$

- Estimate optimal policy from approximated Q-functions:

$$\hat{\pi}_t(s_t,a_{t-1}) = \arg\max_{a_t} \hat{Q}_t(s_t,(a_{t-1},a_t))$$

Hypothetical Cancer Trial

- Duration of the trial: $u \in [0,3]$
- State: tumor size $(0 \leq T(u) \leq 1)$ and wellness $(0.25 \leq W(u) \leq 1)$
- Decision time-point: u_i such that $T(u_i) = 1$
- Treatment: more aggressive treatment A and less aggressive treatment B

$$W(u_i^+|A) = W(u_i) - 0.5, \quad T(u_i^+|A) = \frac{T(u_i)}{10W(u_i)},$$

$$W(u_i^+|B) = W(u_i) - 0.25, \quad T(u_i^+|B) = \frac{T(u_i)}{4W(u_i)},$$

*$w(u) < 0.25$ leads to immediate failure

- Dynamics at stage i (duration $[u_i, u_{i+1}]$)

$$W(u) = W(u_i^+) + (1 - W(u_i^+))\left(1 - 2^{-(u-u_i)/2}\right), \quad T(u) = T(u_i^+) + \frac{4T(u_i^+)}{3}(u - u_i)$$

End of stage i

- $T(u_{i+1}) = 1$ for some $u_i < u_{i+1} < 3$
- Reaches the end of trail $u = 3$
- Failure event occurs with exponential survival function with mean $3(W(u_i^+) + 2)/20M(u_i^+)$
- Censored by $C \sim \text{Unif}(0, c)$

The number of stages is at least one and not more than three.

Experiments

- Size of trajectory dataset: 40, 80, 120, ..., 400
- Censoring percentage: no censoring, 10%, 20%, 30%
- Evaluate π on a dataset of size 1000
- Repeat simulation 400 times

Experiments Results

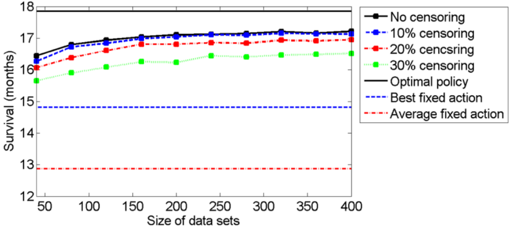

Figure: The expected survival time for different data set sizes with different censoring rates.

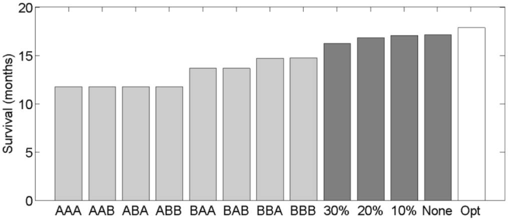

Figure: The expected survival times for fixed treatments and policy estimated from 200 trajectories.

Experiments Results

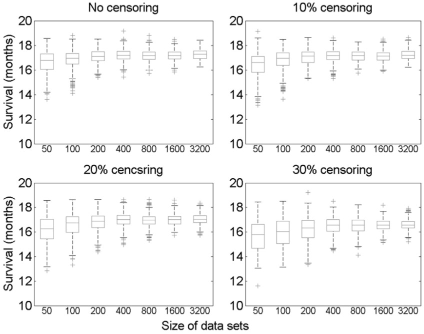

Fig 3.
Distribution of expected survival time (in months) for different data set sizes, with no censoring, 10% censoring, 20% censoring, and 30% censoring. Each box plot is based on 400 repetitions of the simulation for each given data set size and censoring percentage.

Q-function Mismatch

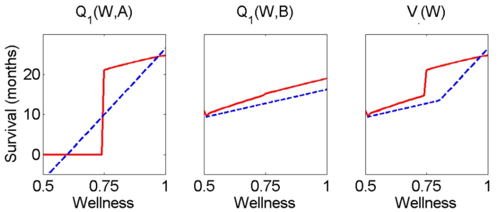

Fig 4.
The Q-functions computed by the proposed algorithm for a size 200 trajectory set. The left panel presents both the optimal Q-function (solid red curve) and the estimated Q-function (dashed blue curve) for different wellness levels and when treatment A is chosen. Similarly, the middle panel shows 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).

Number of Treatments Needed

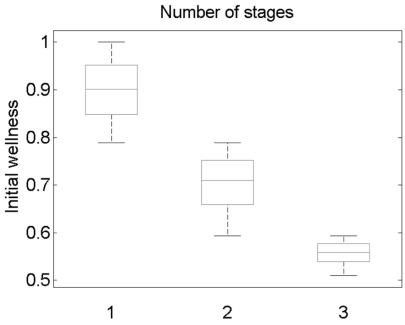

Fig 5.
The number of required treatments for patients that follow the policy $\hat{\pi}$, when no failure event occurs during the trial. The policy $\hat{\pi}$ was estimated from 100 trajectories. The results were computed using a size 100, 000 testing set.

Effect of Ignoring Censoring

Change the uniform censoring to exponential distribution (leaving fewer observations with longer survival times)

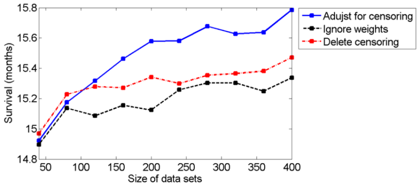

Fig 6.
The solid blue curve, dashed black curve, and dot-dashed red curve correspond to the expected survival times (in months) for different data set sizes, for the proposed algorithm, the algorithm that ignores the weights, and the algorithm that deletes all censored trajectories, respectively. The 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.

Discussion

Contribution

- Proposed a Q-learning algorithm for multistage-decision problems with flexible number of stages in which the rewards are survival times and are subject to censoring.
- Derived the generalization error properties of the algorithm.*
- Demonstrated the algorithm performance using simulations.

Limitation

- Assume censoring is independent of observed trajectories
- The IPCW to correct bias may be inefficient when the percentage of censored trajectories is large.