Nonparametric Reinforcement Learning for Survival Outcomes

Hunyong Cho, Shannon T. Holloway, and Michael R. Kosorok*

University of North Carolina at Chapel Hill
and North Carolina State University

April 3, 2023
• (Statistical) precision medicine
  • Data-driven decision support for treating patients in the presence of heterogeneity (dynamic treatment regimes or DTRs)
  • Treatment can include drug choice, administrative actions, dosing, timing, potentially modifiable risk factors, and/or other potentially beneficial actions to the patients
  • Must be reproducible and generalizable (empirically and inferentially valid)
Operating Principles

• Observable Constituents:
  • Tailoring variables \((X)\)
  • Choice of treatments and/or potentially modifiable risk factors \((A)\)
  • Vector of outcomes or utilities \((R)\)
  • Could be multiple \((X, A, R)\) triples over time for each patient

• Dynamic Treatment Regime (DTR):
  • Single decision: make a single recommendation for treatment
  • Multiple decision: make a series of interdependent recommendations
  • Continual monitoring: for diabetes, mHealth
• Role of Heterogeneity in the data:
  • Heterogeneity of patients is beneficial (essential) for good precision medicine analysis so that estimated treatment rules are broadly applicable
  • Need heterogeneity of treatment assignment (either naturally or by design) in the data so we can determine best treatment under a variety of situations
 Outline of Overall Pipeline

- **Dynamic Treatment Regime:**
  - $\pi(X)$ gives recommended $A$ to maximize $R$ in future patients
  - Regression: model $R$ as a function of $X$ and $A$ ($Q(X, A) = E[R|X, A]$ is the “value”), with interaction between $X$ and $A$ being most important
  - Policy estimation: directly estimate $\pi(X)$ without necessarily needing $Q(X, A)$ (e.g., outcome weighted learning)
  - Prediction versus prescriptive decision support:
    - Suppose $R = f(X) + Ag(X) + e$, where bigger $R$ is better and $A = \{0 \text{ or } 1\}$
    - We only care about $g(x)$, since rule $\pi(X) = \{1 \text{ if } g(X) > 0, \ 0 \text{ otherwise}\}$ yields optimal decision
    - A focus on prediction may yield information inefficiency through focus on $f(X)$ instead of $g(X)$
The Multi-Decision Setting

- The multi-decision setting:
  - Two or more opportunities for treatment decisions (i.e., cancer treatment involving multiple lines of chemotherapy, other chronic diseases, etc.).
  - Interventions can affect patients in multiple ways
    - Immediate effects (proximal)
    - Delayed effects (distal): sometimes the best treatment is initially harmful but sets the patient up for a better response to certain future treatments
Multi-Decision Setting, cont.

- The basic ingredients:
  - The data: \((X_1, A_1, R_1, \ldots, X_T, A_T, R_T)\), where
    - \(X_1 \in \mathcal{X}_1\) denotes baseline information
    - \(X_t \in \mathcal{X}_t\) denotes interim information collected during treatment stages \(t = 2, \ldots, T\)
    - \(A_t \in \mathcal{A}_t\) denotes treatment and
    - \(R_t\) denotes proximal outcome measured after treatment at stage \(t\), for \(t = 1, \ldots, T\).
  - Define \(H_1 = X_1\) and \(H_t = (H_{t-1}, A_{t-1}, R_{t-1}, X_t)\) so that \(H_t\) is the available patient history at time \(t\) before new action.
  - The data used for analysis is now \((H_1, A_1, R_1, \ldots, H_T, A_T, R_T)\).
The Bellman equation and Q-learning

- \( Q_\pi^k(h, a) = \)
  \[ E[R_k + Q_\pi^{k+1}(H_{k+1}, A_{k+1} = \pi_{k+1}(H_{k+1})) | H_k = h, A_k = a], \]
  \( k = K - 1, K - 2, \ldots, 1, \)

where \( \pi \) is a certain policy that maps
\( H \equiv (H_1, \ldots, H_k) \mapsto A \equiv (A_1, \ldots, A_k). \)
The Bellman equation and Q-learning

- $Q_{k}^{\pi}(h, a) = E[R_{k} + Q_{k+1}^{\pi}(H_{k+1}, A_{k+1} = \pi_{k+1}(H_{k+1})) \mid H_{k} = h, A_{k} = a]$, $k = K - 1, K - 2, \ldots, 1$, where $\pi$ is a certain policy that maps $\mathcal{H} \equiv (\mathcal{H}_{1}, \ldots, \mathcal{H}_{k}) \mapsto \mathcal{A} \equiv (A_{1}, \ldots, A_{k})$.

- Q-learning recursively finds the optimal policy as $\pi_{k}^{*}(h) = \arg \max_{a \in A_{k}} Q_{k}^{\pi_{k}^{*}}(h, a)$, $k = K, K - 1, \ldots, 1$. 
• Regress $R_T$ onto $(H_T, A_T)$ to obtain an estimate of $E[R_T|H_T = h, A_T = a]$, denoted $\hat{Q}_T(h, a)$.

• For each individual, compute $\hat{R}_T = \sup_{a \in A_T} \hat{Q}_T(H_T, a)$.

• Proceeding backwards from $t = T - 1$ to $t = 1$, do the following:
  • Regress $R_t + \hat{R}_{t+1}$ onto $(H_t, A_t)$ to obtain an estimate of $E[R_t + \hat{R}_{t+1}|H_t = h, A_t = a]$, denoted $\hat{Q}_t(h, a)$.
  • For each individual, compute $\hat{R}_t = \sup_{a \in A_t} \hat{Q}_t(H_t, a)$.

The estimated optimal dynamic treatment regime is then $\hat{\pi}_t(h_t) = \arg \max_{a \in A_t} \hat{Q}_t(h_t, a)$, for $t = 1, \ldots, T$. 
Question: Can we find a set of dynamic rules that maximizes the survival outcomes?
Question: Can we find a set of dynamic rules that maximizes the survival outcomes?

A potential solution: Q-learning.
Question: Can we find a set of dynamic rules that maximizes the survival outcomes?

A potential solution: Q-learning.
Dynamic treatment regimes (DTR) for survival outcomes

Question: Can we find a set of dynamic rules that maximizes the survival outcomes?

A potential solution: Q-learning.
Question: Can we find a set of dynamic rules that maximizes the survival outcomes?

A potential solution: Q-learning.

Challenges:
- Number of stages differ (failure or dropout before all planned visits).
- How to do backward recursion for survival data?
How was censoring handled in the literature? Goldberg and Kosorok (2012)

- modified data without loss or addition of information
  - The time increments \( R_k = T_k \) after censoring/failure are left as zero
  - The history after censoring/failure is set as \( H_k = \emptyset \)
  - The actions after censoring/failure are randomly drawn.

- Use Q-learning with the complete 'pseudo' data
- Censoring is handled by inverse probability of censoring weighting (IPCW).
How was censoring handled in the literature?

Goldberg and Kosorok (2012)

- modified data without loss or addition of information
  - The time increments \( R_k = T_k \) after censoring/failure are left as zero
  - The history after censoring/failure is set as \( H_k = \emptyset \)
  - The actions after censoring/failure are randomly drawn.
- Use Q-learning with the complete ‘pseudo’ data
How was censoring handled in the literature?
Goldberg and Kosorok (2012)

• modified data without loss or addition of information
  • The time increments ($R_k = T_k$) after censoring/failure are left as zero
  • The history after censoring/failure is set as $H_k = ∅$
  • The actions after censoring/failure are randomly drawn.

• Use Q-learning with the complete ‘pseudo’ data

• Censoring is handled by inverse probability of censoring weighting (IPCW).
How was censoring handled in the literature?
Goldberg and Kosorok (2012)

- modified data without loss or addition of information
  - The time increments ($R_k = T_k$) after censoring/failure are left as zero
  - The history after censoring/failure is set as $H_k = \emptyset$
  - The actions after censoring/failure are randomly drawn.

- Use Q-learning with the complete ‘pseudo’ data

- Censoring is handled by inverse probability of censoring weighting (IPCW).

However, *independent censoring* was assumed.
Several other relevant methods.

| method                        | $|A_k|$ | failure time | policy class | censoring  | criterion |
|-------------------------------|-------|--------------|--------------|------------|-----------|
| Goldberg et al (2012)         | finite| nonparametric| flexible     | $C \perp T_k$| $E[T]$   |
| Huang et al (2014)            | finite| AFT          | linear       | CI         | $E[T]$   |
| Simoneau et al (2019)         | 2     | AFT          | linear       | CI         | $E[T]$   |
| Jiang et al (2017)            | 2     | PH           | linear       | CI         | $S(t)$   |

- $|A_k|$, the number of treatment arms at stage $k$.
- criterion, the target value being optimized.
- AFT, accelerated failure time; PH, proportional hazards; CI, conditional independence.
- $E[T]$, mean (truncated) survival time; $S(t)$, survival probability at time $t$. 
Several other relevant methods.

| method                      | $|A_k|$ | failure time | policy class | censoring  | criterion      |
|-----------------------------|-------|--------------|--------------|------------|----------------|
| Goldberg et al (2012)      | finite| nonparametric| flexible     | $C \perp T_k$ | $E[T]$         |
| Huang et al (2014)         | finite| AFT          | linear       | CI         | $E[T]$         |
| Simoneau et al (2019)      | 2     | AFT          | linear       | CI         | $E[T]$         |
| Jiang et al (2017)         | 2     | PH           | linear       | CI         | $S(t)$         |
| **new**                    | finite| nonparametric| flexible     | CI         | $E[T], S(t)$   |

- $|A_k|$, the number of treatment arms at stage $k$.
- criterion, the target value being optimized.
- AFT, accelerated failure time; PH, proportional hazards; CI, conditional independence.
- $E[T]$, mean (truncated) survival time; $S(t)$, survival probability at time $t$. 
The proposed method.

- Nonparametric Q-function estimation (random forest).
- Censoring mechanism: covariate-conditionally independent.
- The outcome of interest $= \phi(S')$, some function of the survival probability; $\phi(S')$ can be the (truncated) mean survival time ($E[T \wedge \tau]$) or survival probability at a certain time $t$ ($S(t)$).
The proposed method.

- Nonparametric Q-function estimation (random forest).
- Censoring mechanism: covariate-conditionally independent.
- The outcome of interest $= \phi(S)$, some function of the survival probability;
  $\phi(S)$ can be the (truncated) mean survival time ($E[T \wedge \tau]$) or survival probability at a certain time $t$ ($S(t)$).
- Backward recursion $\Rightarrow$ Slightly more general than Q-learning
• $K$ treatment stages 
  ($A_k \in A_k, \ k = 1, 2, ..., K$).
• $(T_k, U_k)$ are the times to failure and the next treatment at Stage $k$.
• $V_k = T_k \land U_k$.
• $\gamma_k = 1(T_k \leq U_k)$.
• $L_k = \text{“the remaining life” after start of Stage } k$.
• $B_k = \text{time elapsed before } k$. 

Cho, Holloway & Kosorok (UNC-CH BIOS) Nonparametric Reinforcement Learning for Survival Outcomes April 3, 2023 13 / 31
• $K$ treatment stages 
  ($A_k \in A_k$, $k = 1, 2, \ldots, K$).
• $(T_k, U_k)$ are the times to failure and the next treatment at Stage $k$.
• $V_k = T_k \land U_k$.
• $\gamma_k = 1(T_k \leq U_k)$.
• $L_k = \text{“the remaining life” after start of Stage } k$.
• $B_k = \text{time elapsed before } k$.

$L_k$ can be recursively written as,

$$L_k = V_k + (1 - \gamma_k)L_{k+1} \quad \text{for } k < K - 1.$$
- $K$ treatment stages $(A_k \in A_k, \ k = 1, 2, \ldots, K)$.
- $(T_k, U_k)$ are the times to failure and the next treatment at Stage $k$.
- $V_k = T_k \land U_k$.
- $\gamma_k = 1(T_k \leq U_k)$.
- $L_k^* = \text{the remaining life, were the optimal treatments given in later stages} \ (k' > k)$.

\[
L_k = V_k + (1 - \gamma_k) L_{k+1} \quad \text{for} \ k < K - 1.
\]
• $X_k = \min(T_k, U_k, C - B_k) \wedge V_k$ is the observed stage length.
• $\delta_k = 1(V_k \leq X_k)$.
• $X_k = \min(T_k, U_k, C - B_k)$
  $\wedge = V_k$
  is the observed stage length.

• $\delta_k = 1(V_k \leq X_k)$. 
- Backward recursion: Start from stage $K, K - 1, \ldots, 1$.

- For stage $k$,
  - Estimate $S_k$ (the “cumulative” survival curves): $\hat{S}_k(t \mid H_k, A_k = a)$
  - Find $\hat{\pi}_k$ (the stage $k$ decision rule):
    $\hat{\pi}_k(h) = \arg \max_a \phi(\hat{S}_k(\ldots \mid H_k = h, a))$
  - Augmentation: Add the previous stage length to the optimized curve when $\gamma_{k-1} = 0$. $X_{k-1} + L^*_k$ where $L^*_k \sim \hat{S}_{k}\hat{\pi}_k$.

- The final rule: $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_K)^\top$. 
The terminal stage estimator ($k = K$)

- $S_K(t|H_K, A_K)$: the ‘terminal stage’ survivor function of $L_K(= T_K)$.
- Estimated using random survival forest.
- The optimal ITR estimator for stage $K$ is,

$$\hat{\pi}_K(h_K) = \arg\max_{a \in A_K} \phi(\hat{S}_k(t - B_k|H_K = h_K, A_K = a)).$$
DTR for survival outcomes - illustration

\[ S_k(t | H_k = h_{k1}, a) \]

\[ A_k = 0 \]

\[ B_k \]
DTR for survival outcomes - illustration

\[ S_K(t | H_K = h_{K_1}, a) \]

\[ S_K(t | H_K = h_{K_2}, a) \]

\[ A_K = 1 \]

\[ A_K = 0 \]
DTR for survival outcomes - illustration

\[ S_{K-1}(t \mid H_K = h_{K1}, a) = S_K(t - X_{K-1} \mid H_K = h_{K1}, a) \]

\[ S_K(t \mid H_K = h_{K1}, a) \]

\[ X_{K-1} \]

\[ A_K = 0 \]

\[ A_K = 1 \]

\[ B_K \]

\[ S_K(t \mid H_K = h_{K2}, a) \]

\[ A_K = 1 \]

\[ A_K = 0 \]

\[ B_K \]
DTR for survival outcomes - illustration

\[ S_{K-1}(t | H_K = h_{K1}, a) = S_K(t - X_{K-1} | H_K = h_{K1}, a) \]

\[ S_K(t | H_K = h_{K1}, a) \]

\[ A_K = 1 \]

\[ A_K = 0 \]

\[ S_K^*(t | H_K = h_{K1}) \]

\[ S_{K-1}^*(t | H_K = h_{K1}) \]
For earlier stages. Consider stage $k < K$.

- stage length $X_k$ at $k$ is augmented by $\hat{S}_{k+1}^*$. This is done by using $\hat{S}_{k+1}^*(t - X_k | H_k, A_k)$ for each individual. (For those censored during stage $k$, no augmentation is needed.)

- Now the survival distribution of $L_k$ is estimated using the stochastically augmented intervals $\{\hat{S}_{k+1}^*(t - X_{k,i} | H_{k,i}, A_{k,i})\}_i$. 
Generalized random survival forests are used. Modified splitting rules, Modified Kaplan-Meier at terminal nodes.

Properties: uniform consistency under certain regularity conditions.

Simulations validate theory, is effective in example application.
Theorem

Assuming the conditions that follow, the value $\mathcal{V}$ of the estimated optimal dynamic treatment regime, $\hat{\pi}$, is consistent for the truth. I.e.,

$$|\mathcal{V}(\hat{\pi}) - \mathcal{V}(\pi_*)| \rightarrow_P 0,$$

as $n \rightarrow \infty$, where the value ($\mathcal{V}(\pi)$) is either the restricted mean survival time ($E[T^\pi \wedge \tau]$) or the survival probability at a certain time ($S^\pi(t_0)$).
Assumptions for each stage $k$:

1. Stable unit treatment value assumption (SUTVA)
2. $A_k \perp T^a_k \mid H_k, \forall a \in A_k$
3. $\Pr(A_k = a \mid H_k = h) > L_1, \forall a, h, \exists L_1 > 0.$
4. $\Pr(U_k < T_k \land C_k | h) > M, \forall h \in \mathcal{H}_k, \exists M > 0.$
5. $|S_k(t | h_1) - S_k(t | h_2)| \leq L_S\|h_1 - h_2\|,\quad |G_k(t | h_1) - G_k(t | h_2)| \leq L_G\|h_1 - h_2\|, \quad \forall h_1, h_2, \exists 0 < L_S, L_G < \infty.$
6. $1/\zeta \leq f_{H_k}(h) \leq \zeta$
7. $n_{\text{min}} \to \infty$ with $\frac{\log n \log \log n}{n_{\text{min}}} \to \infty$
8. Regular and random-split trees
DTR outline of proof

• We use error bounding methods given in Murphy (2005) and Goldberg and Kosorok (2012) to bound the DTR error by the uniform accuracy of the nonparametric survival estimator at each $1 \leq k \leq K$.

• Specifically, we show that

$$V(\pi^*) - V(\hat{\pi}) \leq \sum_{k=1}^{K} c_k(\phi)$$

$$\times \sqrt{\sup_{h_k, a_k, t \in [0, \tau]} \left| \hat{S}_k(t \mid h_k, a_k) - S_k(t \mid h_k, a_k) \right|},$$

where $c_k(\phi)$ are constants that depend on the reward function $\phi$.

• We then establish the needed uniform consistency and convergence rates.
Suppose the assumptions hold. Let \( \hat{S} = (\hat{S}_1, ..., \hat{S}_2, ..., \hat{S}_K) \) be the sequence of the generalized random survival forest estimators of \( S = (S_1, ..., S_k, ..., S_K) \) such that the \( k \)th stage random survival forest is built based on \( \hat{S}_{k+1} \) for \( k = 1, 2, ..., K - 1 \). Then,

\[
\sup_{t \in [0, \tau], h \in \mathcal{H}_k, k \in \{1, 2, ..., K\}} |\hat{S}_k(t | h) - S_k(t | h)| \rightarrow 0,
\]

in probability as \( n \rightarrow \infty \).
Survival consistency outline of proof

• Results follow from uniform consistency of each $\hat{S}_k$, beginning with $k = K$ and going backwards to $k = 1$.
• We use Z-estimator consistency based on identifiably of the estimating equation (i.e., showing that if the expected Z-function, evaluated at $\theta_n$, goes to zero uniformly over the index, then this forces $\|\theta_n - \theta_0\| \to 0$) combined with uniform consistency of the empirical Z-function (see, e.g., Theorem 2.10 of Kosorok, 2008).
• We use VC-dimension bounded kernel representations of the random forests based on axis-aligned rectangles to obtain consistency of the empirical Z-function.
Suppose the assumptions hold plus a few additional assumptions. Then, for any \( k = 1, 2, \ldots, K \), there exists an \( 1 \leq n_0 < \infty \) such that for all \( n > n_0 \) the following holds with probability \( \geq 1 - \frac{3(K-k+1)}{\sqrt{n}} \):

\[
\sup_{t \leq \tau, h_k} |\hat{S}_k(t; h_k) - S_k(t; h_k)|
\]

\[
\leq \sum_{l=k}^{K} \frac{11}{c_1} \sqrt{\frac{\log(n/n_{\min})\{\log(d_l n_{\min}) + 3 \log \log(n)\}}{n_{\min} \log((1 - \alpha)^{-1})}}
\]

\[
+ \zeta L_S \left\{ \frac{2n_{\min}}{n} \right\} \frac{\log((1 - \alpha)^{-1})}{\log(\alpha^{-1})} \frac{0.991 \varphi}{d_l},
\]

where the constants come from the assumptions.
Theorem

In the context of the previous theorem, $n_{\text{min}}$ and the other tuning parameters can be chosen so that, for some $\eta > 0$,

$$\sup_{t \leq \tau, h_k} |\hat{S}_k(t; h_k) - S_k(t; h_k)| = O_P(n^{-\eta}),$$

and

$$\mathcal{V}(\pi_*) - \mathcal{V}(\hat{\pi}) \leq O_P(n^{-\eta/2}).$$

Thus the convergence rates are polynomial in $n$. 
Simulation results

from Cho, Holloway and Kosorok (2020)
• We applied these methods to an acute myeloid leukemia clinical trial with survival as an outcome (Wahed & Thall, 2013; Xu et al, 2016).

• 210 patients were randomized to frontline treatment (4 possibilities) followed by salvage treatment (2 classes) adaptively chosen by clinicians based on patient status.
from Cho, Holloway and Kosorok (2020)
Clinicians appear to be making treatment selection effectively.

Composite criterion
- Optimize $S(t)$ first and, if tied, use $E(T)$ as the second criterion.

Non-Markov assumption: History matters. However, the disease dynamics need to be stationary within a treatment stage.
• We thank Dr. Donglin Zeng for the discussion of the composite criterion.
• Data credits to Drs. Peter F. Thall, Abdus S. Wahed, and Yanxun Xu (the leukemia data).
• Partial support by grant P01 CA142538 from the National Cancer Institute.
• R package *dtrSurv* on CRAN.