Optimal Treatment Regimes under Constraints

Ph.D. Thesis Presentation

Shuping Ruan

Supervised by Dr. Eric Laber
Department of Statistics
North Carolina State University

December 19, 2017

Precision medicine

- Categorize individuals into subpopulations based on
  - Demographics
  - Genetic information
  - Results of diagnostic test
  - Susceptibility to a certain disease
  - Response to a specific treatment
- Target therapeutic or preventive interventions to individuals
  - Provide more effective treatments
  - Avoid unnecessary side effects and costs

Dynamic treatment regimes

- Definition: a set of decision rules, one function for each decision point, that map patient characteristics to recommended treatments
- Opposite of the acute care model: active but short-term treatment for an urgent medical condition
- The goal is to optimize the long-term cumulative clinical outcome for chronic care
- Analogous to a policy in reinforcement learning, or a controller in control theory
- Apply to time-varying policies in other fields, such as education, marketing, and economics

Optimal treatment regimes

- Most of DTRs focus on a single scalar outcome
- Indirect methods: Q-learning, A-learning, etc.
- Policy search: outcome weighted learning, doubly robust estimators, etc.
- Balancing multiple competing outcomes
  - Treatment effectiveness
  - Side-effect burden
  - Cost and so on
- Current work
  - Composite outcomes
  - Set-valued treatment regimes

Constrained optimal treatment regimes

- Propose a new framework for estimating optimal treatment regimes under constraints
- Optimize the primary outcome of interest, subject to constraints on the secondary outcomes
  - Single-stage setting
  - Multi-stage setting
  - Infinite-stage setting

Single-stage

Dataset

- Dataset
$$\left\{\left(\boldsymbol{X}^i, A^i, \boldsymbol{Y}^i\right)\right\}_{i=1}^{n}$$
- n i.i.d patient trajectories
- $\boldsymbol{X} \in \mathcal{X}$: the patient information collected up to the decision point
- $A \in \mathcal{A}$: the treatment assignment
- $\boldsymbol{Y} \in \mathbb{R}^J$: the vector of outcomes of interest
  - $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_J)^\intercal$
  - Y_1, the primary outcome of interest, the higher the better
  - $Y_2, \cdots, Y_J$, the secondary outcomes of interest, the lower the better

Potential outcome framework

- The counter-factual framework for identifying the causal effect of a certain regime from randomized or observational studies.
- The set of potential outcomes: $\boldsymbol{W}^* = \left\{ \boldsymbol{Y}^* (a), \text{for all } a \in \mathcal{A} \right\}$, where $\boldsymbol{Y}^* (a)$ is the vector-valued outcome that would have been observed if the subject was assigned treatment a.
- Three essential assumptions guarantee that the value for a regime can be estimated using the observed data.

Potential outcome framework

- *Consistency*

$$\boldsymbol{Y} = \boldsymbol{Y}^{*}(A).$$

The actual observed outcome vector $\boldsymbol{Y}$ for an individual who received treatment A is the same as the potential outcome for that individual assigned with the same treatment, regardless of the experimental conditions used to assign treatment. It also implies that there is no interference among individuals.

Potential outcome framework

- *No unmeasured confounders*

$$\boldsymbol{W}^* \perp\!\!\!\perp A \mid \boldsymbol{X}.$$

The set of potential outcomes, $\boldsymbol{W}^* = \left\{ \boldsymbol{Y}^* (a) , \text{for all } a \in \mathcal{A} \right\}$, are conditionally independent of treatment assignment A given patient information $\boldsymbol{X}$. In randomized study, this condition is satisfied by construction. However, it can not be verified in observational studies.

Potential outcome framework

- *Positivity assumption:*

$$\exists\, \epsilon > 0, \text{s.t. } \Pr(A = a \mid \boldsymbol{X}) > \epsilon, \forall a \in \mathcal{A} \text{ w/ prob 1.}$$

This ensures there is a positive probability of receiving every possible treatment assignment for every value of patient covariates in the population. This assumption is satisfied in well-designed randomized studies. It can also be empirically verified in observational studies. Yet, if it is violated, estimating of regimes for certain subsets of patients can be impossible.

A single-stage treatment regime

- A single-stage treatment regime $\pi : \boldsymbol{\mathcal{X}} \to \mathcal{A}$ is a function that maps the support of patient information $\boldsymbol{X}$ to the set of all possible treatments.
- Under a regime π, a patient with $\boldsymbol{X} = \boldsymbol{x}$ is recommended to receive treatment $\pi(\boldsymbol{x})$.
- The vector-valued potential outcome of the regime π is $\boldsymbol{Y}^*(\pi) = \sum_{a \in \mathcal{A}} \boldsymbol{Y}^*(a) \mathbb{I}\{\pi(\boldsymbol{X}) = a\}$.

Values of a regime

- The value functions of a regime π is defined as

$$\boldsymbol{V}(\pi) = \mathbb{E}\boldsymbol{Y}^*(\pi),$$

of which each component is $V_j(\pi) = \mathbb{E}Y_j^*(\pi)$, $j = 1, \cdots, J$. That is the expected outcome if every patient in the population of interest is assigned treatment according to π,

- The Q functions are defined as

$$\boldsymbol{Q}(\boldsymbol{X}, A) = \mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{X}, A),$$

of which each component is $Q_j(\boldsymbol{X}, A) = \mathbb{E}(Y_j \mid \boldsymbol{X}, A)$, $j = 1, \cdots, J$.

- Under the three assumptions, we have

$$\boldsymbol{V}(\pi) = \mathbb{E}(\boldsymbol{Q}(\boldsymbol{X}, \pi(\boldsymbol{X})).$$

Constrained optimal treatment regimes

- A single-stage constrained optimal regime is

$$\max_{\pi \in \Pi} \; V_1(\pi)$$
$$\text{subject to } V_j(\pi) \leq \nu_{j-1},$$

where $j = 2, \cdots, J$, and $\boldsymbol{\nu} = (\nu_1, \nu_2, \cdots, \nu_{J-1})^\intercal$ are the constraints.

Constrained optimal treatment regimes

- Linear decision rules, $\pi(\boldsymbol{x};\boldsymbol{\theta}) = \text{sgn}(\boldsymbol{x}^\intercal\boldsymbol{\theta})$, $\|\boldsymbol{\theta}\|_2^2 = \boldsymbol{\theta}^\intercal\boldsymbol{\theta} = 1$
- Further simplify the notation

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^q} \; v_1(\boldsymbol{\theta})$$
$$\text{subject to } v_j(\boldsymbol{\theta}) \le 0,\; h(\boldsymbol{\theta}) = 0,$$

where $j = 2, \cdots, J$, $v_1(\boldsymbol{\theta}) = -V_1(\boldsymbol{\theta})$, $v_j(\boldsymbol{\theta}) = V_j(\boldsymbol{\theta}) - \nu_j$, and $h(\boldsymbol{\theta}) = \boldsymbol{\theta}^\intercal\boldsymbol{\theta} - 1$.

Interior-point method

- The interior-point method solves a sequence of the following approximate minimization problem,

$$\min_{\boldsymbol{\theta},\boldsymbol{z}} \phi_\mu(\boldsymbol{\theta},\boldsymbol{z}) = v_1(\boldsymbol{\theta}) - \mu \sum_{j=2}^{J} \ln z_j,$$

$$\text{subject to } v_j(\boldsymbol{\theta}) + z_j = 0, \text{ and } h(\boldsymbol{\theta}) = 0,$$

  where μ is always positive and approaches to zero in the limit.
- As μ decreases to zero, the minimums of ϕ_μ form a trajectory path that approaches the feasible optimum of $v_1(\boldsymbol{\theta})$, $\boldsymbol{\theta}^*_{\boldsymbol{\nu}}$, in the limit.

Estimation of the value functions

- $V_j(\boldsymbol{\theta}) = \mathbb{E}\left(Q_j(\boldsymbol{X}, \pi(\boldsymbol{X}, \boldsymbol{\theta}))\right)$, for $j = 1, \cdots, J$.
- A linear model $Q_j(\boldsymbol{X}, A) = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_j + A \cdot \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta}_j$.
- A regime is approximated by $\pi(\boldsymbol{X}) = \text{sgn}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\theta})$.
- Let $(z_1, z_2) = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\theta}, \boldsymbol{x}_1^{\mathsf{T}}\boldsymbol{\beta}_j)$, and $f_{\boldsymbol{\beta}_j}(z_1, z_2; \boldsymbol{\theta})$ be the joint dist'n.
- The estimated values of a regime π are

$$\widehat{V}_j(\boldsymbol{\theta}) = \boldsymbol{x}^{\mathsf{T}}\widehat{\boldsymbol{\alpha}}_j + \iint \text{sgn}(z_1)\, z_2 \widehat{f}_{\widehat{\boldsymbol{\beta}}_j}(z_1, z_2; \boldsymbol{\theta})\, dz_1\, dz_2,$$

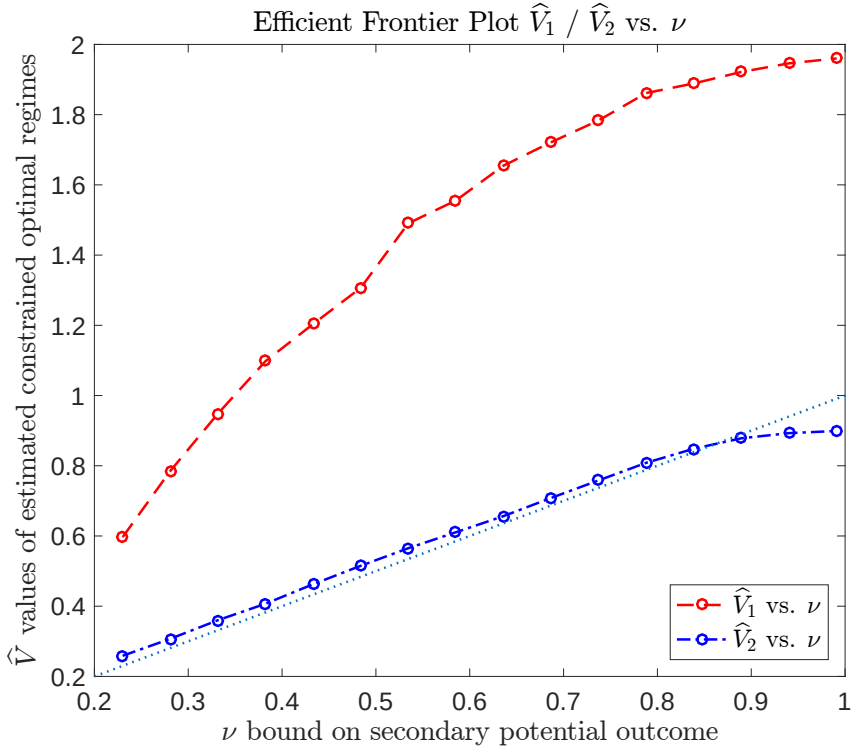
where $\widehat{\boldsymbol{\alpha}}_j$ and $\widehat{\boldsymbol{\beta}}_j$ are the least-squares estimators; $\widehat{f}_{\widehat{\boldsymbol{\beta}}_j}(z_1, z_2; \boldsymbol{\theta})$ is a kernel density estimator (KDE).

Simulation design

- The generative model for simulation is

$$\begin{aligned}
&\boldsymbol{X} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{I}),\\
&A \sim \text{Uniform}\{-1, 1\},\\
&Y_1 = \bar{\boldsymbol{X}}^{\mathsf{T}}\alpha_1 + A \cdot (\bar{\boldsymbol{X}}^{\mathsf{T}}\beta_1) + \epsilon_1, \;\; \epsilon_{Y1} \sim N(0, \sigma_1^2),\\
&Y_2 = \bar{\boldsymbol{X}}^{\mathsf{T}}\alpha_2 + A \cdot (\bar{\boldsymbol{X}}^{\mathsf{T}}\beta_1) + \epsilon_2, \;\; \epsilon_2 \sim N(0, \sigma_2^2),
\end{aligned}$$

  where $\boldsymbol{I}$ is a 2×2 identity matrix and $\bar{\boldsymbol{X}}^{\mathsf{T}} = (1, \boldsymbol{X}^{\mathsf{T}})$. For simplicity, we consider two competing outcomes, i.e., $J = 2$. Also, let $\boldsymbol{X}_0 = \boldsymbol{X}_1 = \boldsymbol{X}$.
- $M_{MC} = 200$, $N_{train} = 1000$, $N_{test} = 10000$.



Multi-stage

Data

- The dataset

$$\{(\boldsymbol{X}_1^i, A_1^i, \boldsymbol{X}_2^i, A_2^i, \cdots, \boldsymbol{X}_T^i, A_T^i, \boldsymbol{Y}^i)\}_{i=1}^n,$$

- n i.i.d. patient trajectories
- $\boldsymbol{X}_t$, the patient covariate collected between the (t-1)-th decision point and the t-th decision point
- A_t, the tth-stage treatment variable
- $\boldsymbol{Y}$, the final outcome vector. Y_1, the primary outcome of interest, the higher the better; $Y_2, \cdots, Y_J$, the secondary outcomes of interest, the lower the better.
- Let $\boldsymbol{H}_t$ denote the patient history information up to the decision point t, i.e., $\boldsymbol{H}_t^\intercal = (\boldsymbol{H}_{t-1}^\intercal, A_{t-1}, \boldsymbol{X}_t^\intercal), \cdots$
- $\bar{A}_t = (A_1, A_2, \cdots, A_t)$ denotes a sequence of treatment history up to time point t

Potential outcome framework

- The set of potential outcomes is $\boldsymbol{W}^* = \{\boldsymbol{X}_2^*(a_1), \boldsymbol{X}_3^*(\bar{a}_2), \cdots, \boldsymbol{X}_T^*(\bar{a}_{T-1}), \boldsymbol{Y}_T^*(\bar{a}_T)$, for all $\bar{a}_t \in \bar{\mathcal{A}}_t, t = 1, 2, \cdots, T\}$
- Assumptions to connect observed data w/ potential outcomes
  - *Consistency:* $\boldsymbol{Y} = \boldsymbol{Y}^*(\bar{A}_T)$, and $\boldsymbol{X}_t = \boldsymbol{X}_t^*(\bar{A}_{t-1})$, $t = 2, \cdots, T$.
  - *Sequential randomization assumption:* $A_t \perp\!\!\!\perp \boldsymbol{W}^* \mid \boldsymbol{H}_t$, for $t = 1, 2, \cdots, T$.
  - *Positivity assumption:* $\exists\, \epsilon_t > 0$, such that $\Pr(A_t = a_t \mid \boldsymbol{H}_t) > \epsilon_t$, w/ prob 1, for all $a_t \in \mathcal{A}_t$, $t = 1, 2, \cdots, T$.

Values of a regime

- The values of a dynamic treatment regime, $\boldsymbol{V}(\pi) = \mathbb{E}\boldsymbol{Y}^*(\pi)$, is defined as the expected final outcome if each patient in the population of interest is treated according to π.
- Values of a regime are estimated using G-computation and the mean-variance modeling techniques by Linn et al.

Constrained optimal treatment regimes

- A multi-stage constrained optimal treatment regime

$$\max_{\pi \in \mathbf{\Pi}} V_1(\pi)$$
$$\text{subject to } V_j(\pi) \leq \nu_{j-1},$$

where $j = 2, 3, \cdots, J$ and $\mathbf{\Pi}$ is the class of dynamic treatment regimes under consideration.

Constrained optimal treatment regimes

- Linear decision rules for each stage, $\pi_t(\boldsymbol{h}_t) = \text{sgn}(\boldsymbol{h}_t^\intercal \theta_t)$, and $\theta_t^\intercal \theta_t = 1$.
- Interior-point method

$$\min_{\boldsymbol{\theta}, \boldsymbol{z}} \phi_\mu(\boldsymbol{\theta}, \boldsymbol{z}) = v_1(\boldsymbol{\theta}) - \mu \sum_{j=2}^{J} \ln z_j,$$

$$\text{subject to } v_j(\boldsymbol{\theta}) + z_j = 0, h_t(\boldsymbol{\theta}_t) = 0,$$

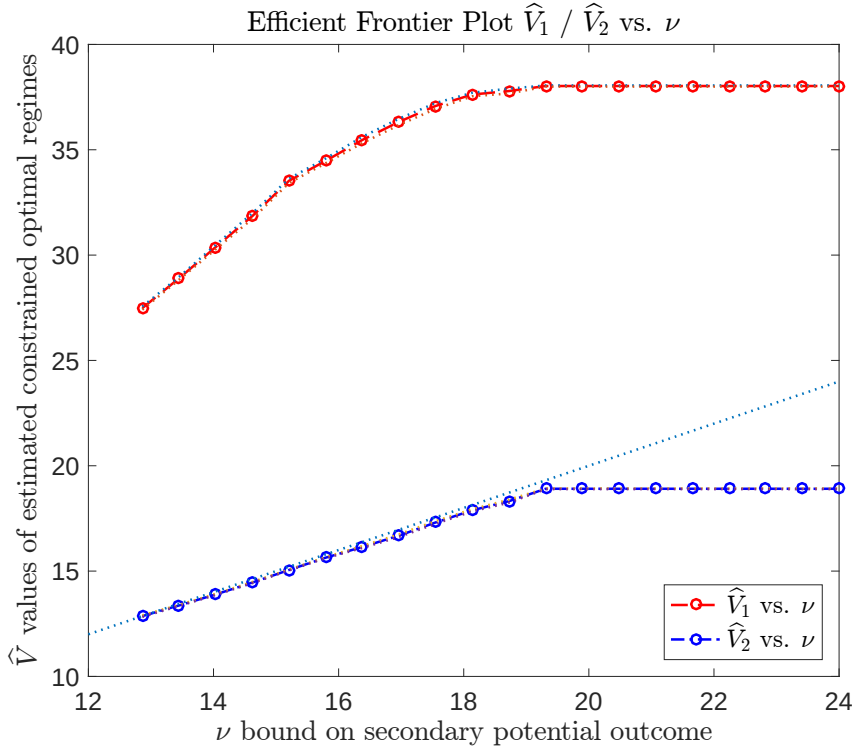
where $v_1(\boldsymbol{\theta}) = -V_1(\boldsymbol{\theta})$, $v_j(\boldsymbol{\theta}) = \boldsymbol{V}_j(\boldsymbol{\theta}) - \nu_j$ for $j = 2, \cdots, J$, and $h_t(\boldsymbol{\theta}_t) = \boldsymbol{\theta}_t^\intercal \boldsymbol{\theta}_t - 1$ for $t = 1, \cdots, T$.

Simulation design

This model is a simple representation of the data from a two-stage sequential randomized trials.

$$
\begin{aligned}
&X_1 \sim \text{Normal}(1,1), \boldsymbol{H}_1 = (1, X_1)^{\mathsf{T}},\\
&A_1 \sim \text{Uniform}\{-1,1\},\\
&X_2 = \boldsymbol{H}_1^{\mathsf{T}}\beta_{1,0} + A_1\boldsymbol{H}_1^{\mathsf{T}}\beta_{1,1} + \epsilon, \epsilon \sim \text{Normal}(0,1),\\
&\boldsymbol{H}_2 = (1, X_2)^{\mathsf{T}},\\
&A_2 \sim \text{Uniform}\{-1,1\},\\
&Y = \boldsymbol{H}_2^{\mathsf{T}}\beta_{2,0,Y} + A_2\boldsymbol{H}_2^{\mathsf{T}}\beta_{2,1,Y} + \epsilon_Y\\
&Z = \boldsymbol{H}_2^{\mathsf{T}}\beta_{2,0,Z} + A_2\boldsymbol{H}_2^{\mathsf{T}}\beta_{2,1,Z} + \epsilon_Z\\
&(\epsilon_Y, \epsilon_Z)^{\mathsf{T}} \sim \text{Normal}(\mathbf{0}_2, \Sigma_{Y,Z})
\end{aligned}
$$

The class of regimes under consideration is $\pi_1 = \text{sgn}(\boldsymbol{h}_1^{\mathsf{T}}\theta_1)$ and $\pi_2 = \text{sgn}(\boldsymbol{h}_2^{\mathsf{T}}\theta_2)$.



Infinite-stage

Constrained reinforcement learning

- Patients with chronic diseases are often monitored and treated throughout their life.
  - Real-time actions
  - No a-priori fixed end point
  - Infinite stage reinforcement learning
- Adopt the constrained Markov Decision Processes (cMDPs) framework for infinite horizon.
- Previously, linear programming is used to sought out constrained optimal policies in the setting of finite cMDPs with known models.
- However, few methods have been proposed for high-dimensional constrained reinforcement learning problems w/o modeling the underlying dynamics.

Data

- The structure of the available data is
$$\boldsymbol{D} = \left\{ (\boldsymbol{S}_0^i, A_0^i, \boldsymbol{R}_0^i, \boldsymbol{S}_1^i, \cdots, \boldsymbol{S}_{T_i-1}^i, A_{T_i-1}^i, \boldsymbol{R}_{T_i-1}^i, \boldsymbol{S}_{T_i}^i) \right\}_{i=1}^{n},$$
- n i.i.d. patient trajectories.
- For each patient, his/er follow up time length T_i may be different.
- $\boldsymbol{S}_t \in \mathcal{S}$ denotes a vector of patient clinical information recorded up to and including time point t
- $A_t \in \mathcal{A}$ denotes the treatment assignment at time point t after measuring $\boldsymbol{S}_t$,
- $\boldsymbol{R}_t \in \mathbb{R}^J$ is the reward obtained after treatment A_t is assigned.

Potential outcome frameworks

- Let $\bar{\boldsymbol{a}}_t = (a_0, a_1, \cdots, a_t) \in \bar{\mathcal{A}}_t$ be a possible treatment assignment sequence up to time point t.
- Let $\bar{\boldsymbol{s}}_t = (\boldsymbol{s}_0, \boldsymbol{s}_1, \cdots, \boldsymbol{s}_t) \in \bar{\mathcal{S}}_t$ be a possible state sequences up to time point t, $t \geq 0$.
- The set of potential outcomes is
  $\boldsymbol{W}^* = \left\{ \boldsymbol{S}_1^*(a_0), \boldsymbol{S}_2^*(\bar{\boldsymbol{a}}_1), \cdots, \boldsymbol{S}_{t+1}^*(\bar{\boldsymbol{a}}_t), \cdots, \text{ for all } \bar{\boldsymbol{a}}_\infty \in \bar{\mathcal{A}}_\infty \right\}$,
  where $\boldsymbol{S}_{t+1}^*(\bar{\boldsymbol{a}}_t)$ is the potential state at $(t+1)$-th time point that would have been observed if the individual had been assigned the treatment sequence $\bar{\boldsymbol{a}}_t$, $t \geq 0$.

Potential outcome frameworks

- *Consistency*: $\boldsymbol{S}_{t+1} = \boldsymbol{S}^*_{t+1}(\bar{\boldsymbol{A}}_t)$, for all $t \geq 0$.
- *Sequential randomization assumption*: $A_{t+1} \perp\!\!\!\perp \boldsymbol{W}^* \mid \bar{\boldsymbol{S}}_{t+1}, \bar{\boldsymbol{A}}_t$, for all $t \geq 0$.
- *Positivity*: there exists $\epsilon_0 > 0$, so that $P\left(A_{t+1} = a_{t+1} \mid \bar{\boldsymbol{S}}_{t+1} = \bar{\boldsymbol{s}}_{t+1}, \bar{\boldsymbol{A}}_t = \bar{\boldsymbol{a}}_t\right) > \epsilon_0$, for all $a_{t+1} \in \mathcal{A}$, $\bar{\boldsymbol{a}}_t \in \bar{\mathcal{A}}_t$ and $\bar{\boldsymbol{s}}_{t+1} \in \bar{\mathcal{S}}_{t+1}$, and all $t \geq 0$.

Markov Decision Process

- *Markov assumption*: $\boldsymbol{S}_{t+1} \perp\!\!\!\perp \left(\bar{\boldsymbol{A}}_{t-1}, \bar{\boldsymbol{S}}_{t-1}\right) \mid (A_t, \boldsymbol{S}_t)$, for all $t \geq 1$.
- *Time homogeneity*: the conditional density $P_t(\boldsymbol{S}_{t+1} = \boldsymbol{s}' \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) = P(\boldsymbol{s}' \mid a, \boldsymbol{s})$ for all $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}$ and $a \in \mathcal{A}$ and $t \geq 0$, where $\boldsymbol{s}$ and $\boldsymbol{s}'$ denote the current state and the next state, respectively.

Value functions of dynamic treatment regimes

- A dynamic treatment regime, $\pi : \boldsymbol{\mathcal{S}} \rightarrow \mathcal{A}$.
- The value function $\boldsymbol{V}^{\pi}(\boldsymbol{s})$ of a state under a certain policy π is defined as the expected total discounted rewards when the process begins in state $\boldsymbol{s}$ and all decisions are made according to policy π.
- $\boldsymbol{V}^{\pi}(\boldsymbol{s}) = \mathbb{E}^{\pi}_{\boldsymbol{s}} \sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1})$.
- Bellman equation,
  $\boldsymbol{V}^{\pi}(\boldsymbol{s}) = \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}' \mid \boldsymbol{s}, \pi(\boldsymbol{s})) \left(\boldsymbol{R}(\boldsymbol{s}, \pi(\boldsymbol{s}), \boldsymbol{s}') + \gamma \boldsymbol{V}^{\pi}(\boldsymbol{s}') \right)$.

Value functions of dynamic treatment regimes

- Moreover, the state-action value function under policy π, $\boldsymbol{Q}^{\pi}(\boldsymbol{s}, a) = \sum_{\boldsymbol{s}' \in \mathcal{S}} p\left(\boldsymbol{s}' \mid \boldsymbol{s}, a\right)\left(\boldsymbol{R}\left(\boldsymbol{s}, a, \boldsymbol{s}'\right) + \gamma\, \boldsymbol{Q}^{\pi}(\boldsymbol{s}', \pi(\boldsymbol{s}'))\right)$, is defined similar but the first step takes action a.
- In clinical cases, the transition model $\mathbb{P}$ is unknown, optimal regimes must be learn from observed dataset.
- In infinite horizon setting, as time steps are dropped, we break n observed trajectories into 4-tuple of $(\boldsymbol{s}, a, \boldsymbol{r}, \boldsymbol{s}')$ for estimating value functions.

Least-squares policy evaluation

- The state-action value function $\boldsymbol{Q}^{\pi}$ is considered the fixed point of the Bellman operator: $\boldsymbol{Q}^{\pi} = T^{\pi}\boldsymbol{Q}^{\pi}$, where the Bellman operator defined as

$$T^{\pi}\boldsymbol{Q}^{\pi} = \sum_{\boldsymbol{s}' \in \mathcal{S}} p(\boldsymbol{s}'|\boldsymbol{s}, a) \left(\boldsymbol{R}(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma \sum_{a' \in \mathcal{A}} \pi(a' \mid \boldsymbol{s}')\, \boldsymbol{Q}^{\pi}(\boldsymbol{s}', a') \right).$$

- The approximation for each component is $Q_j^{\pi}(\boldsymbol{s}, a; \boldsymbol{w}) = \sum_{k=1}^{K} \phi_{j,k}(\boldsymbol{s}, a) w_{j,k} = \boldsymbol{\phi}_j^{\intercal}(\boldsymbol{s}, a)\boldsymbol{w}_j$
  - $\boldsymbol{w}_j = (w_{j,1}, \cdots, w_{j,K})^{\intercal}$ are the parameters to estimate.
  - The basis functions $\boldsymbol{\phi}_j(\boldsymbol{s}, a) = (\phi_{j,1}(\boldsymbol{s}, a), \cdots, \phi_{j,K}(\boldsymbol{s}, a))^{\intercal}$ are arbitrary and fixed, which are often non-linear functions of $\boldsymbol{s}$ and a.

Least-squares policy evaluation

Algorithm 1: Least-squares policy evaluation LSQ

Input : A sample set $\boldsymbol{D}$ of 4 tuples $(\boldsymbol{s}', a, \boldsymbol{s}, \boldsymbol{r})$
k: Number of basis functions
ϕ: Basis functions
γ: Discount factor
π: policy whose value function is sought

Output: Weights $\widehat{\boldsymbol{w}}^{\pi}$

$\widehat{\boldsymbol{A}} \leftarrow 0$ // $(k \times k)$ matrix
$\widehat{\boldsymbol{b}} \leftarrow 0$ // $(k \times 1)$ vector
for each $(\boldsymbol{s}, a, \boldsymbol{r}, \boldsymbol{s}') \in \boldsymbol{D}$

$$\widehat{\boldsymbol{A}} \leftarrow \widehat{\boldsymbol{A}} + \phi(\boldsymbol{s}, a)\left(\phi(\boldsymbol{s}, a) - \gamma\phi(\boldsymbol{s}', \pi(\boldsymbol{s}'))\right)^{\mathsf{T}}$$

$$\widehat{\boldsymbol{b}} \leftarrow \widehat{\boldsymbol{b}} + \phi(\boldsymbol{s}, a)\boldsymbol{r}$$

$\widehat{\boldsymbol{w}}^{\pi} \leftarrow \widehat{A}^{-1}\widehat{b}$
return $\widehat{\boldsymbol{w}}^{\pi}$

Constrained optimal dynamic treatment regimes

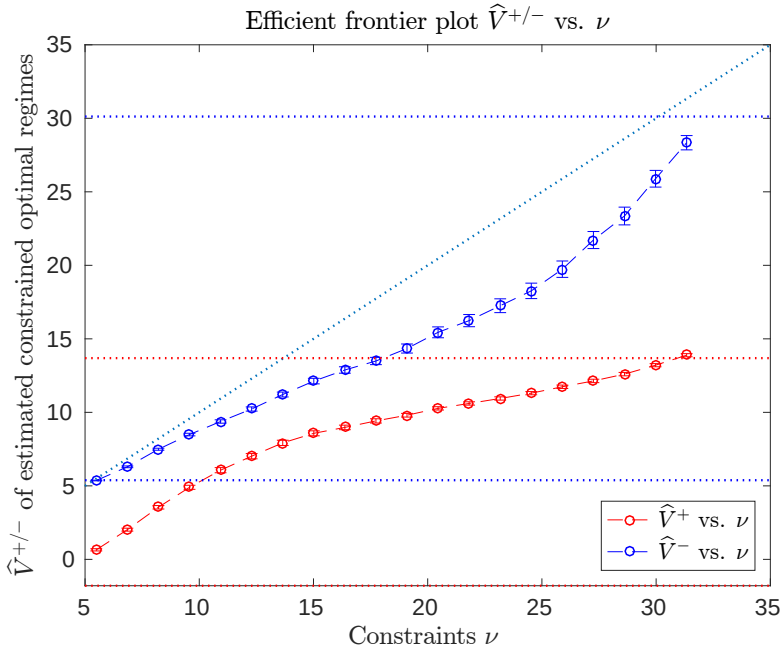
- The average of value functions is referred as competing outcomes, denoted as $\boldsymbol{V}(\pi) = \mathbb{E}\boldsymbol{V}^{\pi}(\boldsymbol{s})$.
- Mathematically,

$$\max_{\pi \in \Pi} \; V_1(\pi),$$
$$\text{subject to } V_j(\pi) \leq \nu_{j-1},$$

where $j = 2, \cdots, J$.

Chemotherapy mathematical model

- The chemotherapy mathematical model, a system of ordinary differential equations (ODE), proposed by Zhao et al, is modified and used to generate a hypothetical clinical trial data.
- The model reflects the capability of the drug to suppress tumor growth, as well as its negative impact on patient wellness due to the toxicity of chemotherapy.
- The dose assignment is discretized to $L = 5$ levels, $\mathcal{A} = \{0.00, 0.25, 0.50, 0.75, 1.00\}$.



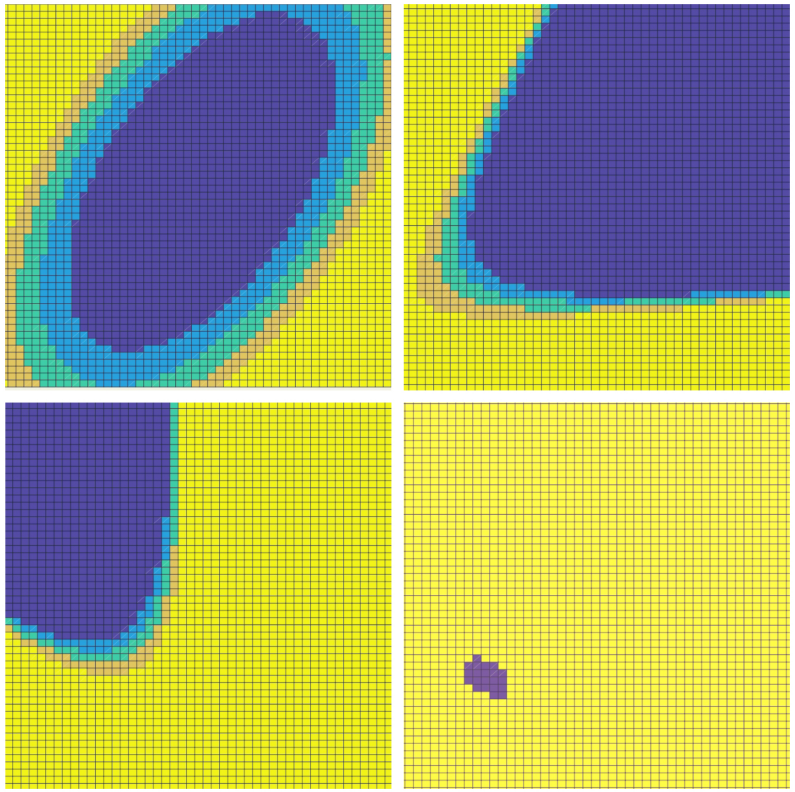


Figure: Treatment assignments under four different constraints

Discussion and Future Work

- Consistency and asymptotic normality were proven for single-stage and finite-stage
- Theoretical work for infinite-stage constrained problem
- Real clinical data application
- Incorporate more complex input data, such as omics data, medical images, etc.
- Learning rewards/objective/constraints from clinical experts

Thank You & Happy Holidays!