Dynamic Treatment Regimes under Constraints

Shuping Ruan
Department of Statistics, North Carolina State University

NC STATE

Overview

- Dynamic treatment regimes (DTRs) are sequential decision making problems in precision medicine.
- Most of the current methods for constructing DTRs focus on optimizing a single utility function over a finite number of decision time points (finite horizon).
- However, clinical situations often, in practice, require considering the trade-off among multiple competing outcomes without a priori fixed end of follow-up time point (infinite horizon).
- Hence, we develop a method of estimating constrained optimal dynamic treatment regimes in chronic diseases where patients are monitored and treated throughout life.
- Our method is demonstrated through a simulated randomized clinical trial dataset based on a chemotherapy mathematical model for cancer.

Dataset

- Observed data structure:

$$\mathcal{D} = \left\{ \left(S_0^i, A_0^i, R_0^i, S_1^i, \cdots, S_{T_i-1}^i, A_{T_i-1}^i, R_{T_i-1}^i, S_{T_i}^i \right) \right\}_{i=1}^n$$

  - Assume n i.i.d. trajectories, and the causal inference assumptions for identifiability of the causal effect of a regime
  - $T \in \mathbb{N}$: the total number of follow-up time steps for a patient
  - $S_t \in \mathcal{S}$: the vector of a patient clinical information recorded up to time t, aka *state*. If a patient passed away, $S_t = \emptyset$, aka *absorbing state*
  - $A_t \in \mathcal{A}$: the treatment assignment at time point t, aka *action*
  - $R_t \in \mathbb{R}^J$: the *reward* vector obtained after treatment A_t is assigned

Values of regimes

- A dynamic treatment regime, or *policy*, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, is defined as a function which maps the support of the state variable to the set of the possible treatment assignments.
- The value function $V^\pi(s) \in \mathbb{R}^J$ of a state under a certain policy π is defined as the expected total discounted rewards when the process begins in state s and all decisions are made according to policy π.

$$V^\pi(s) = \mathbb{E}_s^\pi \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}),$$

  where $\mathbb{E}_s^\pi$ is the expectation when the initial state is s and a policy π is followed. γ is the discount factor.
- The state-action value function $Q^\pi(s, a) \in \mathbb{R}^J$ under policy π, is defined similar but the first step takes action a. Equivalently, it can be expressed recursively via the bellman equation,

$$Q^\pi(s, a) = \int_{s' \in \mathcal{S}} p(s' \mid s, a) \left(R(s, a, s') + \gamma V^\pi(s') \right).$$

  However, the transition model $\mathbb{P}$ is unknown in clinical cases, and optimal regimes must be learn from observed dataset.
- Hence, we adopt the least-squared policy evaluation method (LSQ) with Gaussian radial basis functions to estimate the value a regime [1].

Infinite-stage constrained optimal regimes

- Our goal is to use dataset observed over a finite length of time steps to construct a deterministic regime in the setting of infinite horizon constrained Markov decision process.

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

  where ν_{j-1}, for $j = 2, \cdots, J$, are bounds on the constraints.
- To search over a feasible policy space with manageable computation complexity, a quadratic function is used for policy function approximation.
- Interior-point method is used for constrained optimization [2].
- Our method is applied to a dataset simulated using a modified chemotherapy mathematical model, a system of ordinary differential equations (ODE), originally developed by Zhao et al [3].

Results

- Pareto efficient frontier plot

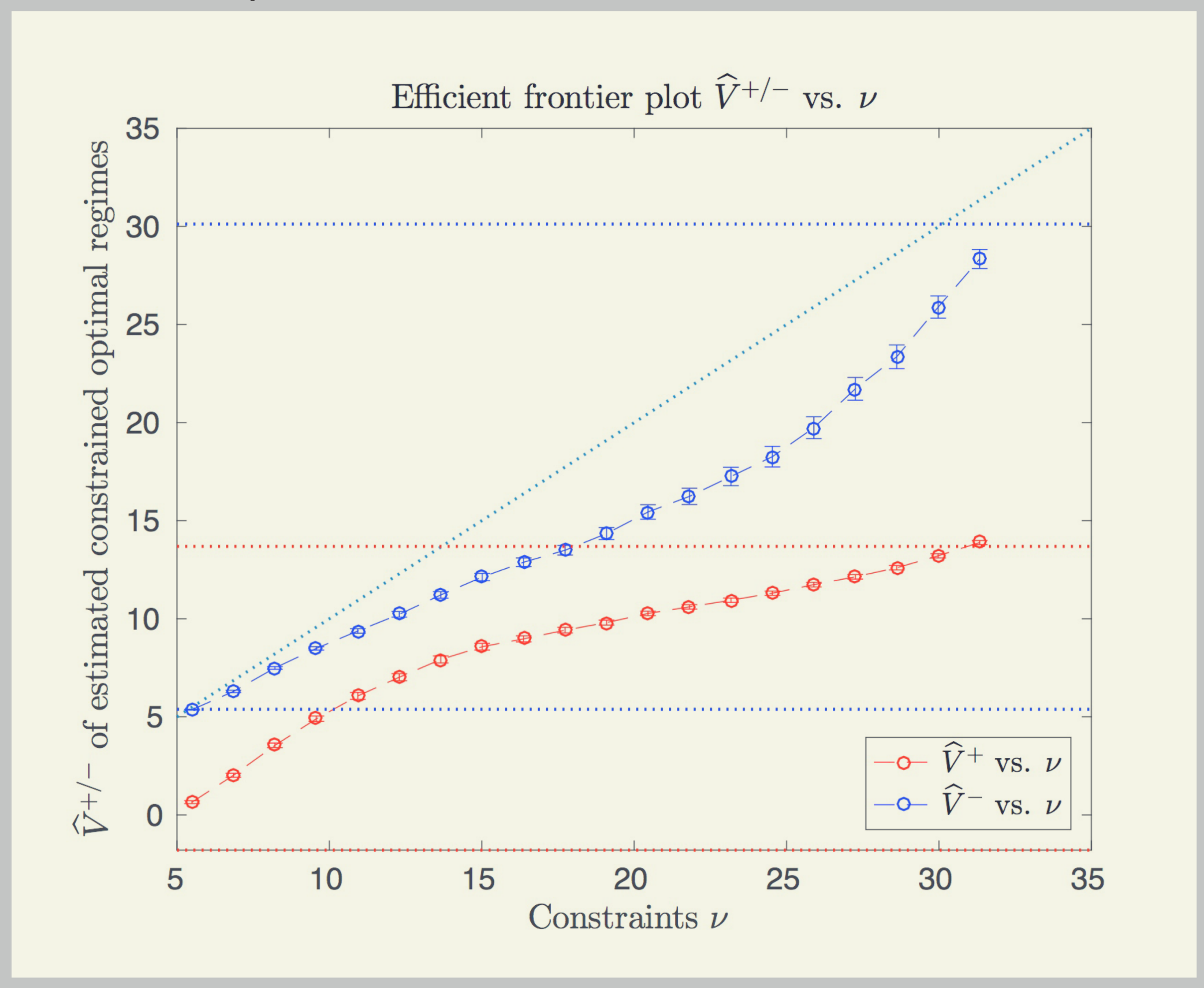

Figure 1: Pareto efficient frontier plot with confidence intervals, 300 Monte Carlo replicates

- Treatment

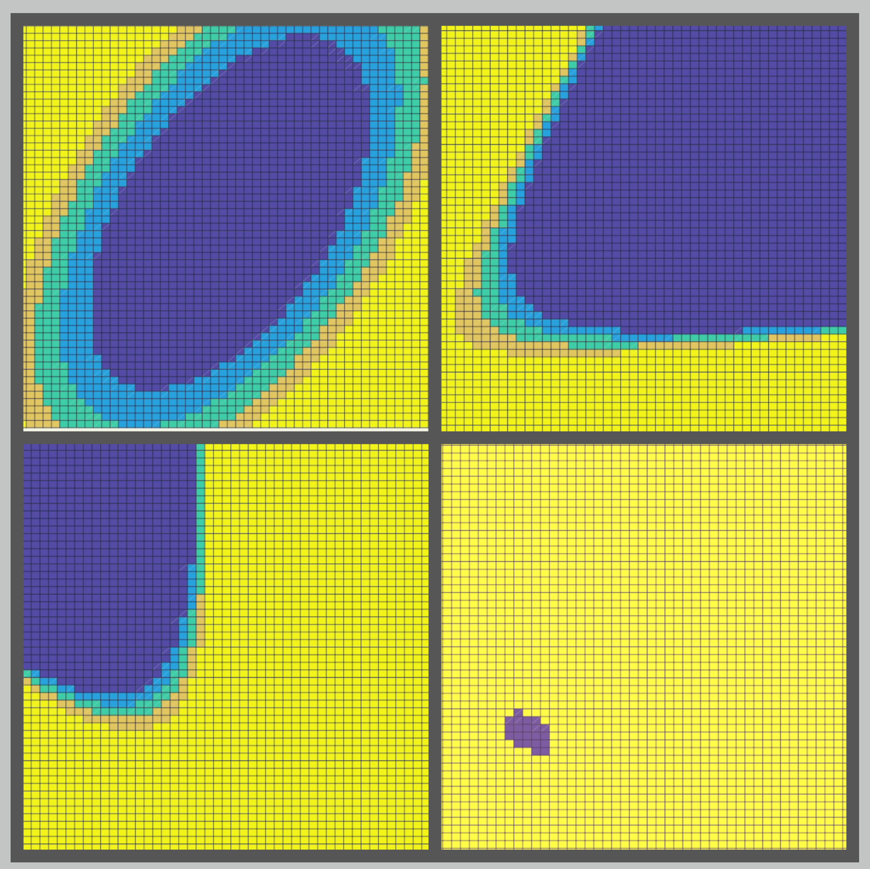

Figure 2: Action for each state under a constraint bound. From left to right, top to bottom, $\nu = 10, 17, 24, 31$. Yellow is a high dose treatment, and blue is low dose treatment.

Conclusion

- We developed a method for constructing infinite-stage constrained optimal treatment regime using LSQ and interior point method.
- Our method is suitable for batch off-line learning, due to the combination of LSQ and policy function approximation.
- Our work is a foray to the practical use of dynamic treatment regimes in the real world applications, where handling multiple objectives in life-long clinical situation is inevitable.

References

[1] Michail G Lagoudakis and Ronald Parr.
Least-squares policy iteration.
The Journal of Machine Learning Research, 4:1107–1149, 2003.

[2] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban.
An interior algorithm for nonlinear optimization that combines line search and trust region steps.
Mathematical Programming, 107(3):391–408, 2006.

[3] Yufan Zhao, Michael R Kosorok, and Donglin Zeng.
Reinforcement learning design for cancer clinical trials.
Statistics in Medicine, 28(26):3294–3315, 2010.

Acknowledgments

- I thank my advisor Dr. Eric Laber for his guidance.

Contact Information

- www.laber-labs.com
- github.com/ShupingR
- sruan@ncsu.edu
- +1 (919) 348-3217