# Constrained Estimation of Single-stage Optimal Treatment Regimes

Shuping Ruan, Eric Laber
Department of Statistics, North Carolina State University

March 23, 2018

## 1 Introduction

Precision medicine aims to accommodate interventions to individual patient attributes. For chronic illnesses, such as cancer, diabetes and so on, clinicians often need to make sequences of treatment decisions based on the evolving status of the patient's condition. To personalize this multi-stage intervention process, researchers have been developing dynamic treatment regimes (DTRs) to inform clinicians of treatment decisions adaptively. DTRs are a sequential decision making process, of which each decision is made based on the evolving patient status with the goal of maximizing the overall long-term treatment efficacy. It is well-studied in the statistical and biomedical literature [19, 18, 21, 26, 10, 13]. From the standpoint of Markov decision process, reinforcement learning algorithms, such as Q-learning [21], A-learning [3], V-learning [17] etc., are developed to estimate optimal treatment regimes.

Most previous methods for construct optimal dynamic treatment regimes have focused on optimizing a scalar measurement of the long-term efficacy over a fixed time period (finite stage). However, in practice, the clinical situations are more complex. First, they often require consideration of the trade-off among multiple objectives, e.g., effectiveness, side-effect, cost, and so on. The preference of those objectives varies among people and changes over time, thus a single scalar reward or value can not represent the quality of a policy well enough. Thus, considering multiple rewards are necessary. Previous works by Lizotte et al. learn the value function and optimal policy for all preferences, i.e., all the possible convex combination of all the rewards [14, 15]. More recent works adopt multi-objective Markov decision processes (MOMDPs) framework with finite stage. Practical domination is proposed for flexibility based on Pareto domination, and a set of policies that are maximal based on the partial order are treated as indistinguishably optimal [9, 16]. Secondly, patients with chronic diseases are often monitored and treated throughout their life. It often requires taking real-time actions and has no a-priori fixed end point (infinite stage),

and progress is made in infinite stage reinforcement learning for health applications [5, 20, 17].

To deal with the trade-off between multiple objectives, we take a different perspective and adopt the constrained Markov decision processes (CMDPs) framework with infinite horizon. CMDPs are a well-studied framework for reinforcement learning under constraints [2]. The goal is to find the optimal policy, while satisfying constraints on expectations of secondary costs. For many applications of reinforcement learning, the constrained approach is more intuitive and more practical than eliciting a single reward function in order to achieve desirable results. For instance, satisfying safety constraints is necessary for systems that physically interact with humans. Previously, linear programming is used to seek 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 without modeling the underlying dynamics. Recently, Achiam et al [1] proposed constrained policy optimization, a general-purpose policy search algorithm for constrained reinforcement learning guaranteeing near-constraint satisfaction at each iteration. Taking into consideration properties of clinical applications, such as data scarcity and off-policy learning, we develop an algorithm by taking advantage of least-squares policy evaluation and interior-point methods for estimating constrained optimal dynamic treatment regimes. Our method is applied to a simulated cancer trial dataset based on a chemotherapy mathematical model.

# 2 Methodology

## 2.1 Set-up

### Observed Data

We use dataset observed over a finite length of time steps to construct a regime in the setting of infinite horizon Markov decision process. The structure of the available data is $\boldsymbol{D} = \left\{(\boldsymbol{S}_0^i, A_0^i, \boldsymbol{R}_0^i, \boldsymbol{S}_1^i, \cdots, \boldsymbol{S}_{T_i-1}^i, A_{T_i-1}^i, \boldsymbol{R}_{T_i-1}^i, \boldsymbol{S}_{T_i}^i)\right\}_{i=1}^n$, a set of $n$ independent, identically distributed trajectories of
$(\boldsymbol{S}_0, A_0, \boldsymbol{R}_0, \boldsymbol{S}_1, \cdots, \boldsymbol{S}_{T-1}, A_{T-1}, \boldsymbol{R}_{T-1}, \boldsymbol{S}_T)$. Note $T \in \mathbb{N}$ denotes the total number of follow-up time steps for a patient. For each patient, his or her follow up time length $T_i$ may be different. $\boldsymbol{S}_t \in \boldsymbol{\mathcal{S}}$ denotes a vector of patient clinical information recorded up to and including time point $t$, referred as *state* in the reinforcement learning vocabulary. $\boldsymbol{\mathcal{S}} \subseteq \mathbb{R}^m$ denotes the support for state variable. Adopting the time homogeneous Markov decision process model, we assume $\boldsymbol{\mathcal{S}}$ is the same across all time points $t$. $A_t \in \mathcal{A}$ denotes the treatment assignment at time point $t$ after measuring $\boldsymbol{S}_t$, referred as *action* in reinforcement learning. $\mathcal{A}$ denotes the support for treatment assignment, a finite set of all possible treatment options, and is assumed to be the same across all time

points $t$. $\boldsymbol{R}_t \in \mathbb{R}^J$ is the reward obtained after treatment $A_t$ is assigned. We assume the reward, possibly defined based on domain expertise, is a known vector-valued function $\boldsymbol{r}$: $\boldsymbol{\mathcal{S}} \times \mathcal{A} \times \boldsymbol{\mathcal{S}} \to \mathbb{R}^J$, so that $\boldsymbol{R}_t = \boldsymbol{r}(\boldsymbol{S}_t, A_t, \boldsymbol{S}_{t+1})$. The vector-valued reward function is the same across all the time points $t$ as well. Moreover, if a patient dies during the follow-up, say at decision point $t$, we set $\boldsymbol{S}_t = \boldsymbol{\emptyset}$, referred to as the absorbing state in reinforcement learning. Then, the patient's treatment assignment at time $t$ is $A_t = \emptyset$, and his/er length of follow-up $T = t$.

**Potential outcomes**

In reality, a patient can only be assigned to one of the sequences of treatment assignments. Hence, we can only observe the consequence of that treatment sequence, while the others remain unobserved. To identify the average causal effect of a certain regime, we adopt the counter-factual or potential outcomes framework, established by Neyman, Rubin and Robins for assessment of the time-dependent treatment effect from either randomized or observational studies [8, 22, 23]. Let $\bar{\boldsymbol{a}}_t = (a_0, a_1, \cdots, a_t) \in \bar{\boldsymbol{\mathcal{A}}}_t$ be a possible treatment assignment sequence up to time point $t$, $t \geq 0$, where $\bar{\boldsymbol{\mathcal{A}}}_t = \mathcal{A} \times \cdots \times \mathcal{A}$ is the set of all possible the treatment assignment sequences up to time point $t$. Let $\bar{\boldsymbol{s}}_t = (\boldsymbol{s}_0, \boldsymbol{s}_1, \cdots, \boldsymbol{s}_t) \in \bar{\boldsymbol{\mathcal{S}}}_t$ be a possible state sequences up to time point $t$, $t \geq 0$, where $\bar{\boldsymbol{\mathcal{S}}}_t = \boldsymbol{\mathcal{S}} \times \cdots \times \boldsymbol{\mathcal{S}}$ is the set of all possible state sequences up to time point $t$. The set of potential outcomes is $\boldsymbol{W}^* = \left\{\boldsymbol{S}_1^*(a_0), \boldsymbol{S}_2^*(\bar{\boldsymbol{a}}_1), \cdots, \boldsymbol{S}_{t+1}^*(\bar{\boldsymbol{a}}_t), \cdots, \text{ for all } \bar{\boldsymbol{a}}_\infty \in \bar{\boldsymbol{\mathcal{A}}}_\infty\right\}$, 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 $\bar{\boldsymbol{a}}_t$, $t \geq 0$. Moreover, if $\boldsymbol{S}_{t+1}^*(\bar{\boldsymbol{a}}_t) = \boldsymbol{\emptyset}$ happens at time point $t+1$, then $\boldsymbol{S}_{t+2}^*(\bar{\boldsymbol{a}}_{t+1}) = \boldsymbol{S}_{t+3}^*(\bar{\boldsymbol{a}}_{t+2}) = \cdots = \boldsymbol{\emptyset}$, which indicates the patient, if followed treatment assignment sequence $\bar{\boldsymbol{a}}_t$, would have died at time point $t+1$ in the counter-factual world. Rewards with respect to potential states are $\boldsymbol{R}_t^* = \boldsymbol{r}(\boldsymbol{S}_t^*, A_t, \boldsymbol{S}_{t+1}^*)$. The following assumptions are made in the potential outcome framework [8, 24, 23, 22, 7].

- *A1. Consistency*: $\boldsymbol{S}_{t+1} = \boldsymbol{S}_{t+1}^*(\bar{\boldsymbol{A}}_t)$, for all $t \geq 0$.
- *A2. Sequential randomization assumption*: $A_{t+1} \perp\!\!\!\perp \boldsymbol{W}^* \mid \bar{\boldsymbol{S}}_{t+1}, \bar{\boldsymbol{A}}_t$, for all $t \geq 0$.
- *A3. 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 $a_{t+1} \in \mathcal{A}$, $\bar{\boldsymbol{a}}_t \in \bar{\boldsymbol{\mathcal{A}}}_t$ and $\bar{\boldsymbol{s}}_{t+1} \in \bar{\boldsymbol{\mathcal{S}}}_{t+1}$, and all $t \geq 0$.

These assumptions link the potential outcome and the observed data, and are guaranteed in well-designed Sequential, Multiple Assignment, Randomized Trials (SMARTs). Therefore, the observed data from those trials are used to infer the average causal effect of a regime of interest.

**Markov Decision Processes**

To construct a regime in the infinite-horizon setting using a dataset observed over a finite number of time steps, we assume that the underlying dynamics

is a time homogeneous Markov Decision Processes (MDPs). In infinite-horizon setting, MDP is considered as a 5-tuple of $(\boldsymbol{\mathcal{S}}, \mathcal{A}, \mathbb{P}, \boldsymbol{R}, \gamma)$, where $\boldsymbol{\mathcal{S}}$, $\mathcal{A}$ and $\boldsymbol{R}$ is as described above. Additionally, $\mathbb{P}$ is a markovian transition model in which $p(\boldsymbol{s}' \mid \boldsymbol{s}, a)$ denotes the probability density of a transition to state $\boldsymbol{s}'$ when taking action $a$ in state $\boldsymbol{s}$. A discount factor for future reward, $\gamma \in [0, 1)$, is also introduced to form total discounted rewards, which are the value functions to operate constrained optimization on. The following assumptions are made for infinite-stage time homogeneous Markov decision process.

- *A4. 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 $t \geq 1$.
- *A5. 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 $\boldsymbol{s}, \boldsymbol{s}' \in \boldsymbol{\mathcal{S}}$ and $a \in \mathcal{A}$ and $t \geq 0$, where $\boldsymbol{s}$ and $\boldsymbol{s}'$ denote the current state and the next state, respectively.

As the time homogeneity is assumed in infinite-stage(3. setting, time step subscripts maybe dropped for simplicity.

**Values of dynamic treatment regimes**

A dynamic treatment regime is equivalent to a *policy* in reinforcement learning vocabulary, which is mostly to be considered deterministic. Thus, a dynamic treatment regime, $\pi : \boldsymbol{\mathcal{S}} \to \mathcal{A}$, is defined as a function which maps the support of the state variable to the set of the possible treatment assignments. As time homogeneity is assumed, we consider only stationary deterministic regimes, where this mapping function $\pi(\boldsymbol{s})$ does not change over time. Hence, a patient with state $\boldsymbol{S}_t = \boldsymbol{s}$ at time point $t$ will be assigned with treatment $A_t = \pi(\boldsymbol{s})$ for all $t$. The value function $\boldsymbol{V}^{\pi}(\boldsymbol{s})$ of a state under a certain policy $\pi$ is defined as the expected total discounted rewards when the process begins in state $\boldsymbol{s}$ and all decisions are made according to policy $\pi$. Mathematically, $\boldsymbol{V}^{\pi}(\boldsymbol{s}) = \mathbb{E}^{\pi}_{\boldsymbol{s}} \sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}(\boldsymbol{s}_t, a_t, \boldsymbol{s}_{t+1})$, where $\mathbb{E}^{\pi}_{\boldsymbol{s}}$ is the expectation when the initial state is $\boldsymbol{s}$ and a policy $\pi$ is followed. The value function can also be defined recursively via the bellman equation, $\boldsymbol{V}^{\pi}(\boldsymbol{s}) = \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}' \mid \boldsymbol{s}, \pi(\boldsymbol{s})) \left(\boldsymbol{R}(\boldsymbol{s}, \pi(\boldsymbol{s}), \boldsymbol{s}') + \gamma \boldsymbol{V}^{\pi}(\boldsymbol{s}')\right)$. Here, $\boldsymbol{V}^{\pi}(\boldsymbol{s}) \in \mathbb{R}^J$ has the same dimensionality as the reward vector $\boldsymbol{R}$, as we are considering multiple reward functions instead of a scalar reward function. Moreover, the state-action value function under policy $\pi$, $\boldsymbol{Q}^{\pi}(\boldsymbol{s}, a) = \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p\left(\boldsymbol{s}' \mid \boldsymbol{s}, a\right) \left(\boldsymbol{R}\left(\boldsymbol{s}, a, \boldsymbol{s}'\right) + \gamma \boldsymbol{Q}^{\pi}(\boldsymbol{s}', \pi(\boldsymbol{s}'))\right)$, is defined similar but the first step takes action $a$ and $\boldsymbol{Q}^{\pi}(\boldsymbol{s}, a) \in \mathbb{R}^J$. 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. The counter-factual assumptions *(A1-A3)* rule out the confounding phenomena and guarantee the identifiability of the average causal effect of a regime.

**Define infinite-stage constrained optimal dynamic treatment regimes**

Our strategy to cope with multiple competing outcomes is constrained optimization. We optimize the primary outcome of interest, subject to the constraints on the secondary outcomes, over the space of all the possible regimes under consideration, $\Pi$. Here, the average of value functions is referred as competing outcomes, denoted as $\boldsymbol{V}(\pi) = \mathbb{E}\boldsymbol{V}^{\pi}(\boldsymbol{s})$. The space of regimes under consideration may be crafted by experts with domain knowledge via policy function approximation. As We have $\boldsymbol{V}^{\pi}(\boldsymbol{s}) = \Big(V_1(\pi), V_2^{\pi}(\boldsymbol{s}), \cdots, V_q^{\pi}(\boldsymbol{s})\Big)^{\intercal}$, Let $V_1(\pi) = \mathbb{E}V_1(\pi)$ be the primary outcome of interest, and $V_j(\pi) = \mathbb{E}V_j^{\pi}(\boldsymbol{s})$, $j = 2, 3, \cdots, J$ be the secondary outcomes. Mathematically,

$$\begin{aligned} &\max_{\pi \in \Pi} \ V_1(\pi), \\ &\text{subject to } V_j(\pi) \leq \nu_{j-1}, \end{aligned} \tag{1}$$

where $j = 2, \cdots, J$. The constraint upper-bounds $\boldsymbol{\nu} = (\nu_1, \nu_2, \cdots, \nu_{J-1})^{\intercal}$ can be specified based on patient preference and/or expert domain knowledge. Therefore, we define an infinite-stage constrained optimal regime as $\pi^*_{\boldsymbol{\nu}} = \text{argmax}_{\pi \in \Pi} \ V_1(\pi)$, subject to $V_j(\pi) - \nu_{j-1} \leq 0$, where $j = 2, \cdots, J$. Denote the feasible policy space, which is the set of all policy satisfying the constraints, as $\mathcal{F}(\Pi)$. For all $\pi \in \mathcal{F}(\Pi)$ and all $j = 2, \cdots, J$, $V_j(\pi) \leq \nu_{j-1}$. Then, an infinite-stage constrained optimal regime can also be written as $\pi^*_{\boldsymbol{\nu}} = \text{argmax}_{\pi \in \mathcal{F}(\Pi)} V_1(\pi)$. To search over a feasible policy space with manageable computation complexity, we use policy function approximation such that $\pi(\boldsymbol{s}) = \pi(\boldsymbol{s}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^q$ is the indexing parameter for policies. Hence, we use $V_j(\pi)$ and $V_j(\boldsymbol{\theta})$ interchangeably, for all $j$. The search space is reduced from the set of all feasible policies to the feasible space of the indexing parameter $\boldsymbol{\theta}$, denoted as $\mathcal{F}(\boldsymbol{\Theta}) = \{\boldsymbol{\theta} \in \mathbb{R}^q : V_j(\boldsymbol{\theta}) \leq \nu_{j-1}, j = 2, \cdots, J\}$.

To carry out policy search, we need to solve the constrained optimization problem (3.1). This is done using interior-point methods, which are constrained nonlinear optimization methods for finding local optimums, implemented in Matlab fmincon [25, 4]. We also need a method to estimate the value functions using observed dataset. This is done by least-square policy evaluation (LSQ) , a part of the least-squares policy iteration (LSPI) algorithm. [12, 11].

## 2.2 Interior point method for constrained optimization

To solve our constrained optimization problem (3.1) above, interior point algorithm is used. As the optimization softwares often implemented as minimization instead of maximization, we denote $v_1(\boldsymbol{\theta}) = -V_1(\boldsymbol{\theta})$ and $v_j(\boldsymbol{\theta}) = V_j(\boldsymbol{\theta}) - \nu_{j-1}$, for $j = 2, \cdots, J$. Hence, problem (3.1) notation is simplified as

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ v_1(\boldsymbol{\theta}) \\ &\text{subject to } v_j(\boldsymbol{\theta}) \leq 0, \end{aligned} \tag{2}$$

where $j = 2, \cdots, J$. The interior point method solves a following sequence of approximate minimization problem (2), where $\rho$ is always positive and approaches to zero in the limit. For each $\rho > 0$, the approximate problem is

$$\min_{\boldsymbol{\theta}, \boldsymbol{z}} \; f_\rho(\boldsymbol{\theta}, \boldsymbol{z}) = \min \; v_1(\boldsymbol{\theta}) - \rho \sum_{j=2}^{J} \ln(z_j), \text{ subject to } v_j(\boldsymbol{\theta}) + z_j = 0, \qquad (3)$$

where $j = 2, \cdots, J$. There are as many slack variables $z_j$ as there are inequality constraints $\nu_j$. The $z_j$ are restricted to be positive to keep $\ln(z_j)$ bounded. As $\rho$ decreases to zero, the minimum of $f_\rho$ should approach the minimum of $\nu_1$. The added logarithmic term is called a barrier function [25, 4].

## 2.3 Least-squares policy evaluation

Least-squares policy evaluation, is adopted to approximate the state-action value function of a fixed regime/policy. As it is the state-action value function being approximated, instead of the state value function, changing policy is allowed without a model for the underlying dynamics. The exact $\boldsymbol{Q}^\pi$ values for all state-action pairs can be found by solving the linear system of the Bellman equations,

$$\boldsymbol{Q}^\pi(\boldsymbol{s}, a) = \sum_{\boldsymbol{s}' \in \mathcal{S}} p(\boldsymbol{s}'|\boldsymbol{s}, a) \left\{ \boldsymbol{R}(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma \sum_{a' \in \mathcal{A}} \pi(a' \mid \boldsymbol{s}') \, \boldsymbol{Q}^\pi(\boldsymbol{s}', a') \right\},$$

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

Linear approximation is a common way for estimating value functions, so that each component of the vector $\boldsymbol{Q}^\pi(\boldsymbol{s}, a; w)$ are approximated by a linear parametric combination of $K$ basis functions (features). As $\boldsymbol{Q}(\boldsymbol{s}, a) = (Q_1(\boldsymbol{s}, a), Q_2(\boldsymbol{s}, a), \cdots, Q_J(\boldsymbol{s}, a))^\intercal$, the approximation for each component is $Q_j^\pi(\boldsymbol{s}, a; w) = \sum_{k=1}^{K} \phi_{j,k}(\boldsymbol{s}, a) w_{j,k} = \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}, a) \boldsymbol{w}_j$, where $\boldsymbol{w}_j = (w_{j,1}, \cdots, w_{j,K})^\intercal$ are the parameters to estimate. Moreover, 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 $a$ . It is also required that the basis functions $\phi_{j,k}$ are linearly independent to ensure that there are no redundant parameters and that the matrices involved in the computations are full rank.

Substituting each component of the Q function vector with the linear approximator, we get, for $j = 1, \cdots, J$,

$$\boldsymbol{\phi}_j^\intercal(\boldsymbol{s}, a) \boldsymbol{w}_j = \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}, a) R_j(\boldsymbol{s}, a, \boldsymbol{s}') + \gamma \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}, a) \sum_{a' \in \mathcal{A}} \pi(a' \mid \boldsymbol{s}') \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}', a') \boldsymbol{w}_j.$$

This fixed point equation then can be rearranged as

$$\left\{ \boldsymbol{\phi}_j^\intercal(\boldsymbol{s},a) - \gamma \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s},a) \sum_{a' \in \mathcal{A}} \pi(a' \mid \boldsymbol{s}') \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}',a') \right\} \boldsymbol{w}_j = \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s},a) R_j(\boldsymbol{s},a,\boldsymbol{s}'). \tag{4}$$

Given a sample set of 4-tuples $\boldsymbol{\mathcal{D}} = \{\boldsymbol{s}^i, a^i, \boldsymbol{s}^{i\prime}, \boldsymbol{r}^i\}_{i=1}^N$, the equation (3.4) above becomes a over-constrained/overdetermined linear system over the parameter vector $\boldsymbol{w}_j$. The linear system can be written as

$$\boldsymbol{B}_j \boldsymbol{w}_j = \boldsymbol{b}_j,$$

where $\boldsymbol{B}_j = \boldsymbol{\Phi}_j^\intercal(\boldsymbol{\Phi}_j - \gamma \boldsymbol{P}^\pi \boldsymbol{\Phi}_j)$ and $\boldsymbol{b}_j = \boldsymbol{\Phi}_j^\intercal \boldsymbol{R}_j$. Moreover,

$$\boldsymbol{\Phi}_j = \begin{pmatrix} \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^1,a^1) \\ \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^2,a^2) \\ \cdots \\ \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^n,a^n) \end{pmatrix}, \boldsymbol{P}^\pi \boldsymbol{\Phi}_j = \begin{pmatrix} \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}^1,a^1) \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}', \pi(\boldsymbol{s}')) \\ \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}^2,a^2) \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}', \pi(\boldsymbol{s}')) \\ \cdots \\ \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}^n,a^n) \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}', \pi(\boldsymbol{s}')) \end{pmatrix},$$

$$\text{and } \boldsymbol{R}_j = \begin{pmatrix} \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}^1,a^1) R_j(\boldsymbol{s}^1,a^1,\boldsymbol{s}') \\ \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}^2,a^2) R_j(\boldsymbol{s}^2,a^2,\boldsymbol{s}') \\ \cdots \\ \sum_{\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}} p(\boldsymbol{s}'|\boldsymbol{s}^n,a^n) R_j(\boldsymbol{s}^n,a^n,\boldsymbol{s}') \end{pmatrix},$$

where $\boldsymbol{\phi}_j^\intercal(\boldsymbol{s}', \pi(\boldsymbol{s}')) = \sum_{a' \in \mathcal{A}} \pi(a' \mid \boldsymbol{s}') \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}', a')$. Since the transition probability function $p(\boldsymbol{s}' \mid \boldsymbol{s}, a)$ and reward functions $R_j(\boldsymbol{s}, a, \boldsymbol{s}')$ may be unknown, we can construct approximators for $\boldsymbol{B}$ and $\boldsymbol{b}$ using samples. More precisely, we have approximated versions of $\boldsymbol{\Phi}_j$, $\boldsymbol{P}^\pi \boldsymbol{\Phi}_j$ and $\boldsymbol{R}_j$ based on the sample set as follows:

$$\widehat{\boldsymbol{\Phi}}_j = \begin{pmatrix} \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^1,a^1) \\ \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^2,a^2) \\ \cdots \\ \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^n,a^n) \end{pmatrix}, \widehat{\boldsymbol{P}^\pi \boldsymbol{\Phi}}_j = \begin{pmatrix} \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^{1\prime}, \pi(\boldsymbol{s}^{1\prime})) \\ \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^{2\prime}, \pi(\boldsymbol{s}^{2\prime})) \\ \cdots \\ \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^{n\prime}, \pi(\boldsymbol{s}^{n\prime})) \end{pmatrix}, \text{ and } \widehat{\boldsymbol{R}}_j = \begin{pmatrix} r_1 \\ r_2 \\ \cdots \\ r_n \end{pmatrix}.$$

Given $\widehat{\boldsymbol{\Phi}}_j$, $\widehat{\boldsymbol{P}^\pi \boldsymbol{\Phi}}_j$, and $\widehat{\boldsymbol{R}}_j$, $\boldsymbol{B}_j$ and $\boldsymbol{b}_j$ can be approximated as $\widehat{\boldsymbol{B}}_j = n^{-1} \widehat{\boldsymbol{\Phi}}_j^\intercal (\widehat{\boldsymbol{\Phi}}_j - \gamma \widehat{\boldsymbol{P}^\pi \boldsymbol{\Phi}}_j) = n^{-1} \sum_{i=1}^n \boldsymbol{\phi}_j(\boldsymbol{s}^i, a^i) \left( \boldsymbol{\phi}_j(\boldsymbol{s}^i, a^i) - \gamma \boldsymbol{\phi}_j \left( \boldsymbol{s}^{i\prime}, \pi \left( \boldsymbol{s}^{i\prime} \right) \right) \right)$ and $\widehat{\boldsymbol{b}}_j = n^{-1} \widehat{\boldsymbol{\Phi}}_j^\intercal \widehat{\boldsymbol{R}}_j = n^{-1} \sum_{i=1}^n \boldsymbol{\phi}_j(\boldsymbol{s}^i, a^i) r_j^i$. It is shown in the least-squares policy iteration paper that $\lim_{n \to \infty} \widehat{\boldsymbol{B}}_j = \boldsymbol{B}_j$ and $\lim_{n \to \infty} \widehat{\boldsymbol{b}}_j = \boldsymbol{b}_j$, if the samples are uniformly distributed over the state space. Moreover, the Markov property ensures that the solution $\widehat{\boldsymbol{w}}^\pi = \widehat{\boldsymbol{B}}^{-1} \widehat{\boldsymbol{b}}$ will converge to the true solution $\boldsymbol{w}^\pi$ for sufficiently large $n$ whenever $\boldsymbol{w}^\pi$ exists[12, 11].The least-squares policy evaluation algorithm is listed in Algorithm 2 below.

Equipped with least-squares policy evaluation, we can hence calculate the values of any arbitrary regime/policy $\pi$. For $j = 1, \cdots, J$, $V(\pi)$ is estimated by $n^{-1} \sum_{i=1}^n \widehat{Q}_j^\pi(\boldsymbol{s}^i, \pi(\boldsymbol{s}^i)) = n^{-1} \sum_{i=1}^n \boldsymbol{\phi}_j^\intercal(\boldsymbol{s}^i, \pi(\boldsymbol{s}^i)) \widehat{\boldsymbol{w}}_j$.

### 2.3.1 Algorithm

Putting together in the following box, our algorithm uses interior point method for policy search in terms policy indexing parameters $\boldsymbol{\theta}$, and least-squares policy evaluation for policy evaluation.

**Algorithm 1:** Constrained optimal regime with least-squares policy evaluation [12, 11] for policy evaluation and interior-point method [25, 4] for policy search.

**Input :** A sample set $\boldsymbol{D}$ of 4 tuples $(\boldsymbol{s}', a, \boldsymbol{s}, \boldsymbol{r})$
**Output:** Estimated constrained optimal regime $\widehat{\pi}_{\boldsymbol{\nu}}$ indexed by $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\nu}}$

$\pi \leftarrow$ random initialization
until converge
    $\widehat{\boldsymbol{Q}}^{\pi}(\boldsymbol{s}, a) \leftarrow$ least-squares policy evaluation $(\boldsymbol{D}, \pi)$
    $\widehat{\boldsymbol{V}}(\pi) \leftarrow \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{Q}}^{\pi}(\boldsymbol{s}^i, \pi(\boldsymbol{s}^i))$
    $\pi \leftarrow \underset{\pi \in \mathcal{F}(\Pi)}{\operatorname{argmax}} \boldsymbol{V}(\pi)$
end
$\widehat{\pi}_{\boldsymbol{\nu}} \leftarrow \pi$
return $\widehat{\pi}_{\boldsymbol{\nu}}$

**Algorithm 2:** Least-squares policy evaluation (LSQ). [12, 11]

**Input :** A sample set $\boldsymbol{D}$ of 4 tuples $(\boldsymbol{s}', a, \boldsymbol{s}, \boldsymbol{r})$
    $k$: Number of basis functions
    $\phi$: Basis functions
    $\gamma$: Discount factor
    $\pi$: policy whose value function is sought
**Output:** Weights $\widehat{\boldsymbol{w}}^{\pi}$

$\widehat{\boldsymbol{B}} \leftarrow 0$ // $(k \times k)$ matrix
$\widehat{\boldsymbol{b}} \leftarrow 0$ // $(k \times 1)$ vector
for each $(\boldsymbol{s}, a, \boldsymbol{r}, \boldsymbol{s}') \in \boldsymbol{D}$
    $\widehat{\boldsymbol{B}} \leftarrow \widehat{\boldsymbol{B}} + \phi(\boldsymbol{s}, a)\left(\phi(\boldsymbol{s}, a) - \gamma\phi(\boldsymbol{s}', \pi(\boldsymbol{s}'))\right)^{\intercal}$
    $\widehat{\boldsymbol{b}} \leftarrow \widehat{\boldsymbol{b}} + \phi(\boldsymbol{s}, a)\boldsymbol{r}$
$\widehat{\boldsymbol{w}}^{\pi} \leftarrow \widehat{\boldsymbol{B}}^{-1}\widehat{b}$
return $\widehat{\boldsymbol{w}}^{\pi}$

# 3 Simulation

## 3.1 Chemotherapy mathematical model

The chemotherapy mathematical model, a system of ordinary differential equations (ODE), proposed by Zhao et al [27]. is modified and used to generate a hypothetical clinical trial data. Their 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, $\mathcal{A} = \{0.00, 0.25, 0.50, 0.75, 1.00\}$. Two state variables $W_t$ and $M_t$ are considered. $W_t$ denotes the patient negative wellness (toxcity) at time point $t$. $M_t$ denotes the tumor size observed at time point $t$. $A_t$ denotes chemotherapy agent dose at time point $t$. The ODE system is modeled as [27]

$$\dot{W}_t = b_1 \max(M_t, M_0) + c_1(A_t - d_1),$$

$$\dot{M}_t = \big(b_2 \max(W_t, W_0) - c_2(A_t - d_2)\big) \times \mathbb{I}(M_t > 0),$$

where decision points are $t = 0, 1, \cdots, T-1$. Moreover, $\dot{W}_t$ and $\dot{M}_t$ are the transition functions. The indicator function term $\mathbb{I}(M_t > 0)$ means tumor size is absorbed at 0, the patient has been cured and no future tumor recurrence considered. These changing rate yields a piece-wise linear model over time. Constants value are set as $b_1 = 0.1, b_2 = 0.15, c_1 = 1.2, c_2 = 1.2, d_1 = 0.5$ and $d_2 = 0.5$. The initial states are draw as $M_0 \sim \text{Uniform}(0, 2)$ and $W_0 \sim \text{Uniform}(0, 2)$. The initial dose level assignment is draw as $A_0 \sim$ Discrete Uniform$(0.25, 0.50, 0.75, 1.00)$. The state variables for the next time point can be obtained via $W_{t+1} = \max(W_t + \dot{W}_t, 0)$, $M_{t+1} = \max(M_t + \dot{M}_t, 0)$. The dose level assignment is drawn as $A_t \sim$ Discrete Uniform$(0.00, 0.25, 0.50, 0.75, 1.00)$, for $t = 1, \cdots, T-1$.

The survival status of the patient, denoted by $F_t$, is also modeled. We assume everyone is alive at the initial decision point $t = 0$, that is $p_0 = 0$ and $F_0 = 0$. Death events occur during time interval $(t-1, t]$, $t = 1, 2, \cdots, 6$, and are recorded at the end of each interval as variable $F_t$, $t = 1, 2, \cdots, 6$. Assume that survival status depends on both toxicity and tumor size. For each time interval $(t-1, t]$, define the hazard function as $\lambda(t)$, where $\log \lambda(t) = \mu_0 + \mu_1 W_t + \mu_2 M_t$, $\mu_1 = \mu_2 = 1$ and $\mu_0 = -8.5$. This again is a piece-wise linear approximation with $\lambda(t) = \exp(\mu_0 + \mu_1 W_t + \mu_2 M_t)$. Then, the cumulative hazard function during time interval $(t-1, t]$ is $\Delta\Lambda(t) = \sum_{t-1}^{t} \lambda(s)\, ds = \sum_{t-1}^{t} \exp(\mu_0 + \mu_1 W_t + \mu_2 M_t)\, ds = \exp(\mu_0 + \mu_1 W_t + \mu_2 M_t)$. The survival function is $\Delta F(t) = \exp\big(-\Delta\Lambda(t)\big) = \exp\big(-\exp(\mu_0 + \mu_1 W_t + \mu_2 M_t)\big)$. The random event of death during time interval $(t-1, t]$ is drawn as $F_t \sim \text{Bernoulli}(p_t)$, where
$p_t = 1 - \exp\big(-\exp(\mu_0 + \mu_1 W_t + \mu_2 M_t)\big)$. If $F_{t-1} = 1$, then $F_t = 1$. Also, as long as death occurred, all the other state variables at the following decision points are all set to null.

The reward functions here is a bivariate vector, consisting of positive reward and negative reward, denoted as $\boldsymbol{R}_t = (R_t^+, R_t^-)^\intercal$. The positive reward function is used to assess tumor size reduction, while the negative reward to assess the increase of patient negative wellness (toxicity). Specifically, the reward functions

are defined as follow.

$$\begin{aligned} R^+(:,t) = &- 15 \times \mathbb{I}(F_{t+1} = 1) \\ &+ 5 \times \mathbb{I}(F_{t+1} \neq 1, M_{t+1} = 0) \\ &+ 5 \times |M_{t+1} - M_t| \times \mathbb{I}(F_{t+1} \neq 1, M_{t+1} - M_t <= 0) \\ &+ 5 \times |M_{t+1} - M_t| \times \mathbb{I}(F_{t+1} \neq 1, M_{t+1} - M_t \leq -0.5), \end{aligned} \tag{5}$$

$$\begin{aligned} R_t^- = &5 \times |W_{t+1} - W_t| \times \mathbb{I}(W_{t+1} - W_t) \geq -0.5) + \\ &5 \times |W_{t+1} - W_t| \, \mathbb{I}(W_{t+1} - W_t \geq 0.5). \end{aligned} \tag{6}$$

To sum up, the trajectories / training data generated according to the ODE model, where with $N = 1000$ and $T = 6$, are as follow

$$\boldsymbol{S}_0 \xrightarrow{A_0} \underset{\boldsymbol{R}_0}{\curvearrowright} \boldsymbol{S}_1 \xrightarrow{A_1} \underset{\boldsymbol{R}_1}{\curvearrowright} \boldsymbol{S}_2 \xrightarrow{A_2} \underset{\boldsymbol{R}_2}{\curvearrowright} \boldsymbol{S}_3 \xrightarrow{A_3} \underset{\boldsymbol{R}_3}{\curvearrowright} \boldsymbol{S}_4 \xrightarrow{A_4} \underset{\boldsymbol{R}_4}{\curvearrowright} \boldsymbol{S}_5 \xrightarrow{A_5} \underset{\boldsymbol{R}_5}{\curvearrowright} \boldsymbol{S}_6,$$

where $\boldsymbol{S}_t = (M_t, W_t, F_t)$, $t = 0, 1, \cdots, 6$. Moreover, $\boldsymbol{R}_t = (R_t^+, R_t^-), t = 0, 1, \cdots, 5$. There are 7 decision points. The last decision point $T = 6$ has only states $\boldsymbol{S}_6 = (M_6, W_6, F_6)$, without following action nor reward. The trajectories is then broken down into 4-tuples of $(\boldsymbol{s}, a, \boldsymbol{s}', \boldsymbol{r})$ with the time stamps dropped.

## 3.2 Function approximation

### 3.2.1 Q function approximation

To construct linear approximators for Q functions [6], we use $K = 4$ Gaussian radial basis functions and an intercept of one. The Gaussian radial basis function has the form of $\phi(x) = \exp\left(\|x - \mu\|^2 / 2\sigma^2\right)$, where $\mu$ and $\sigma^2$ are the parameters to be specified. Denote the Q function for positive rewards as $Q^+(\boldsymbol{s}, a)$, and the one for negative rewards as $Q^-(\boldsymbol{s}, a)$. As the positive reward function is a function of $M_t$, we only incorporate $M_t$ in the basis functions for estimating $Q^+(\boldsymbol{s}, a)$. Hence, we can rewrite $\widehat{Q}^+(\boldsymbol{s}, a)$ as $\widehat{Q}^+(m, a)$. Specifically, $\widehat{Q}^+(m, a) = \widehat{w}_0^+ + \sum_{k=1}^4 \widehat{w}_k^+ \exp\left(-\|m - \mu_k^+\|^2 / 2(\sigma_k^+)^2\right)$, where $\widehat{w}_k^+$ are the weights to be estimated via least-squares policy evaluation. $\mu_k^+$'s are set as the 20, 40, 60, 80 percentiles of the states, and $\sigma_k^+$'s the average distance of the percentiles to all the sample points. $\widehat{Q}^-(\boldsymbol{s}, a)$ is constructed similarly.

#### Policy function approximation

For policy function approximation [6], we focus on simple decision rule to reduce the search space. Let $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5)^\intercal$ be the index parameters for a policy. The policy function is defined as $\pi(\boldsymbol{s}; \boldsymbol{\theta}) = 0.00 \times \mathbb{I}(f(\boldsymbol{s}, \boldsymbol{\theta}) < 1) + 0.25 \times \mathbb{I}(1 \leq f(\boldsymbol{s}, \boldsymbol{\theta}) < 2) + 0.50 \times \mathbb{I}(2 \leq f(\boldsymbol{s}, \boldsymbol{\theta}) < 3) + 0.75 \times \mathbb{I}(3 \leq f(\boldsymbol{s}, \boldsymbol{\theta}) < 4) + 1 \times \mathbb{I}(f(\boldsymbol{s}, \boldsymbol{\theta}) > 4)$, where $f(\boldsymbol{s}, \boldsymbol{\theta}) = \theta_0 + \theta_1 m + \theta_2 m^2 + \theta_3 w + \theta_4 w^2 + \theta_5 mw$.

### 3.3 Simulation results

The goal here is to maximize $V^+(\pi)$, subject to $V^-(\pi) \leq \nu$, where $\nu$ is the bound on the secondary outcome. We applied our method to the simulated dataset. Table 3.1. shows the values of primary and secondary outcomes of the estimated constrained optimal regimes, along with their standard deviations. Table 3.2. shows the estimated indexing parameters of the estimated regimes, along with their standard deviations. Figure 3.1. shows the values of the primary objective(red) / secondary objective (blue) vs. constraint $\nu$. Figure 3.2-3.6 shows the actions of the estimated regime for each state under different constraint values. As the $\nu$ increases, we start to observe more higher dosages being assigned to patients. Higer dosage leads to better treatment effect (more reduced tumor size), but more toxicity on patient's wellness.

Table 1: Values of estimated optimal regimes under different constraint bounds.

| $\nu$ | $\widehat{V}^+$ | $std^+$ | $\widehat{V}^-$ | $std^-$ |
|---|---|---|---|---|
| 5.49 | 0.39 | 0.36 | 5.42 | 0.16 |
| 6.85 | 1.35 | 0.29 | 6.62 | 0.43 |
| 8.21 | 3.58 | 1.06 | 7.66 | 0.30 |
| 9.57 | 4.07 | 0.89 | 7.77 | 0.35 |
| 10.93 | 4.97 | 0.56 | 9.53 | 1.42 |
| 12.29 | 5.98 | 1.10 | 11.43 | 0.69 |
| 13.65 | 6.13 | 0.99 | 11.90 | 1.46 |
| 15.01 | 7.08 | 0.93 | 14.88 | 0.42 |
| 16.37 | 7.98 | 1.31 | 14.69 | 2.15 |
| 17.73 | 8.89 | 0.61 | 16.84 | 0.57 |
| 19.09 | 9.15 | 0.30 | 16.93 | 0.71 |
| 20.45 | 9.85 | 1.16 | 17.86 | 2.33 |
| 21.81 | 9.76 | 1.01 | 18.18 | 2.88 |
| 23.18 | 9.76 | 1.01 | 18.18 | 2.88 |
| 24.54 | 11.62 | 1.38 | 21.88 | 0.63 |
| 25.89 | 12.02 | 1.17 | 23.45 | 2.47 |
| 27.25 | 12.04 | 1.17 | 23.57 | 2.69 |
| 28.61 | 12.86 | 0.54 | 28.27 | 1.06 |
| 29.98 | 13.69 | 0.57 | 30.25 | 1.04 |
| 31.34 | 14.53 | 0.97 | 31.11 | 1.10 |

The constraint bounds are denoted by $\nu$. $\widehat{V}^+$ denotes the primary outcome values of the estimated regimes. $\widehat{V}^-$ denotes the secondary outcome values of the estimated regimes. Standard deviations of those estimated regime values are reported as well.

Table 2: The estimated indexing parameters of estimated regimes under different constraint bounds.

| $\nu$ | $\widehat{\theta}_{\nu,1}$ | $std_1$ | $\widehat{\theta}_{\nu,2}$ | $std_2$ | $\widehat{\theta}_{\nu,3}$ | $std_3$ | $\widehat{\theta}_{\nu,4}$ | $std_4$ | $\widehat{\theta}_{\nu,5}$ | $std_5$ | $\widehat{\theta}_{\nu,6}$ | $std_6$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5.49 | 0.36 | 0.79 | -0.38 | 0.76 | 0.13 | 0.48 | 0.03 | 0.84 | -0.29 | 0.71 | -0.06 | 0.76 |
| 6.85 | 0.11 | 1.47 | -1.55 | 1.10 | 0.29 | 0.67 | 0.35 | 1.44 | -1.06 | 1.19 | -0.20 | 1.41 |
| 8.21 | -0.00 | 1.66 | -2.02 | 1.09 | 0.58 | 0.80 | 0.27 | 1.48 | -1.26 | 1.40 | -0.41 | 1.53 |
| 9.57 | 0.29 | 1.65 | -2.38 | 1.16 | 0.58 | 0.84 | 0.32 | 1.71 | -1.43 | 1.37 | 0.02 | 1.66 |
| 10.93 | 0.49 | 1.64 | -2.61 | 1.13 | 0.67 | 0.90 | 0.18 | 1.72 | -1.47 | 1.57 | 0.23 | 1.64 |
| 12.29 | 0.46 | 1.70 | -2.90 | 1.16 | 0.71 | 0.89 | 0.08 | 1.90 | -1.27 | 1.58 | 0.26 | 1.59 |
| 13.65 | 0.56 | 1.76 | -3.20 | 1.12 | 0.70 | 0.92 | 0.01 | 1.70 | -1.06 | 1.66 | 0.32 | 1.58 |
| 15.01 | 0.67 | 1.74 | -3.46 | 1.08 | 0.65 | 0.90 | -0.02 | 1.87 | -0.83 | 1.77 | 0.14 | 1.55 |
| 16.37 | 0.89 | 1.81 | -3.49 | 1.10 | 0.48 | 0.95 | -0.05 | 1.95 | -0.43 | 1.81 | 0.16 | 1.61 |
| 17.73 | 0.88 | 1.92 | -3.54 | 1.17 | 0.39 | 0.99 | -0.21 | 2.03 | -0.07 | 1.70 | 0.05 | 1.56 |
| 19.09 | 1.20 | 2.08 | -3.36 | 1.54 | 0.47 | 1.17 | -0.41 | 2.14 | 0.11 | 1.56 | 0.15 | 1.52 |
| 20.45 | 1.69 | 2.18 | -3.06 | 1.79 | 0.72 | 1.44 | -0.84 | 2.28 | 0.13 | 1.52 | 0.16 | 1.61 |
| 21.81 | 1.95 | 2.20 | -2.78 | 2.06 | 1.05 | 1.64 | -0.88 | 2.40 | 0.12 | 1.48 | 0.27 | 1.88 |
| 23.18 | 2.21 | 2.21 | -2.64 | 2.19 | 1.30 | 1.76 | -1.09 | 2.32 | 0.15 | 1.47 | 0.03 | 1.93 |
| 24.54 | 2.46 | 2.27 | -2.46 | 2.23 | 1.46 | 1.86 | -1.32 | 2.31 | 0.36 | 1.43 | -0.14 | 2.13 |
| 25.89 | 2.84 | 2.14 | -2.14 | 2.32 | 1.70 | 2.03 | -1.57 | 2.12 | 0.61 | 1.45 | -0.19 | 2.31 |
| 27.25 | 3.25 | 1.82 | -1.95 | 2.26 | 1.66 | 2.14 | -1.60 | 2.05 | 1.01 | 1.60 | -0.12 | 2.41 |
| 28.61 | 3.41 | 1.73 | -1.86 | 2.26 | 1.58 | 2.08 | -1.10 | 2.06 | 1.46 | 1.65 | -0.05 | 2.53 |
| 29.98 | 3.65 | 1.42 | -1.53 | 2.27 | 2.00 | 2.06 | -0.98 | 2.04 | 1.99 | 1.68 | -0.07 | 2.65 |
| 31.34 | 3.89 | 1.14 | -1.07 | 2.18 | 2.23 | 1.91 | -0.60 | 2.15 | 2.64 | 1.61 | -0.13 | 2.70 |

Here, $\nu$ denotes the constraint bounds. $\widehat{\theta}_{\nu,j}$ denotes the $j$-th component of the estimated parameter vector $\widehat{\boldsymbol{\theta}}_{\nu}$ of the estimated regimes. Standard deviations of those estimated regime values are reported as well.

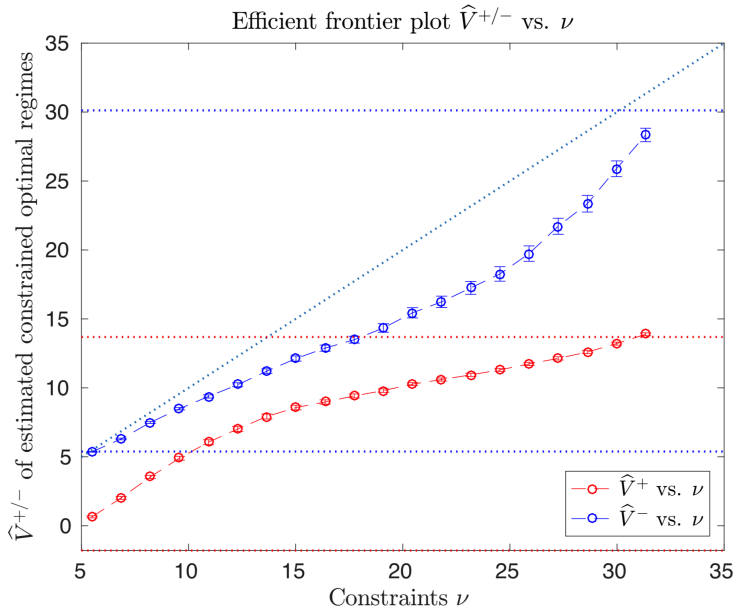


Figure 1: Efficient frontier for estimated constrained optimal regimes (infinite-stage).

The red dashed line is for the primary outcome to maximized. The blue dashed line is for the secondary outcome to be constrained. The red dotted lines are the minima and maxima for unconstrained optimization of the primary objective. The blue dotted lines are the minimal and maximal for unconstrained optimization of the secondary objective.

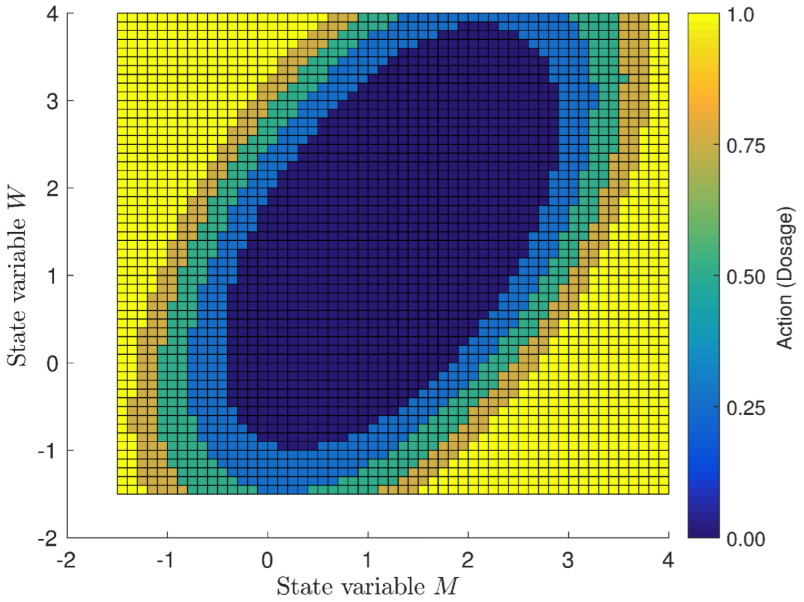


Figure 2: Action for each state under constraint $\nu = 10.93$

Yellow represents high dosage treatment assignment. Blue represents low dosage assignment. As the constraint bound gets loose, more higher dosage treatments are assigned to patients.

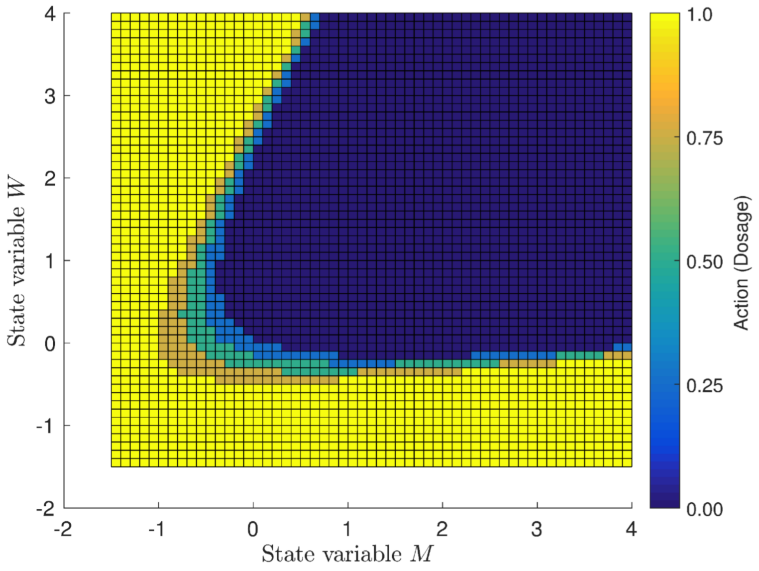


Figure 3: Action for each state under constraint $\nu = 17.73$

Yellow represents high dosage treatment assignment. Blue represents low dosage assignment. As the constraint bound gets loose, more higher dosage treatments are assigned to patients.

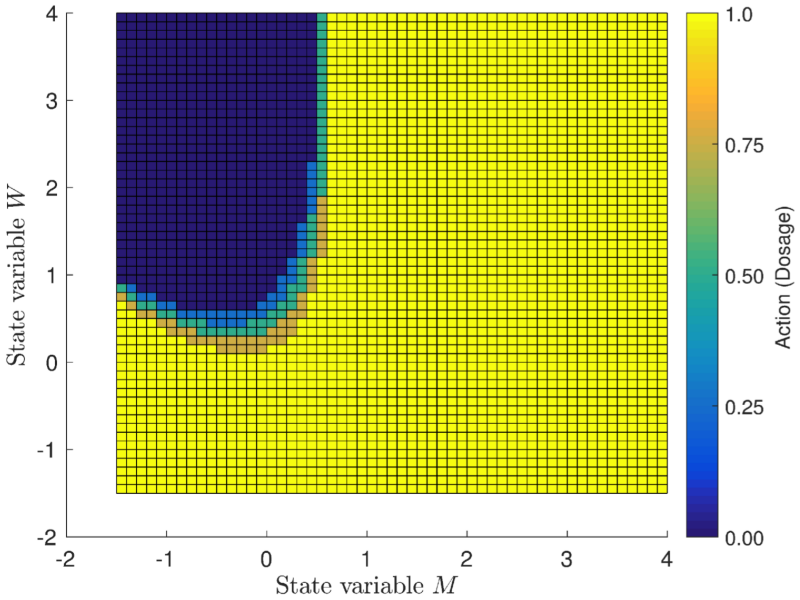


Figure 4: Action for each state under constraint $\nu = 24.54$

Yellow represents high dosage treatment assignment. Blue represents low dosage assignment. As the constraint bound gets loose, more higher dosage treatments are assigned to patients.

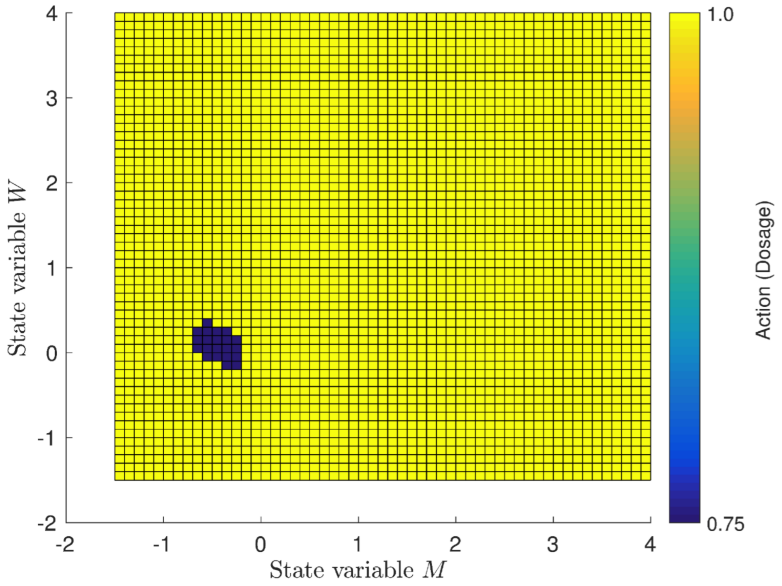


Figure 5: Action for each state under constraint $\nu = 31.34$

Yellow represents high dosage treatment assignment. Blue represents low dosage assignment. As the constraint bound gets loose, more higher dosage treatments are assigned to patients.

# 4 Conclusion and Future

We propose a framework for constrained optimal dynamic treatment regimes to handle the trade-off between the primary objective and all other secondary objectives in infinite stage setting. The simulation results based on the chemotherapy ODE system are presented and visualized. This framework offers an intuitive way for clinicians to exam the trade-off and make treatment decisions based on patient's preference. Different from CPO, our method takes into consideration that clinical data is expensive and scarce. Borrowing strength from least-squares policy evaluation, our method is able to learn from data efficiently. Moreover, least-squares policy evaluation is an iterative method for policy evaluation, and has the advantage of being able to learn both offline and online. Hence, our method can fits in not only the situation where clinical policies needed to be learned after data collect (offline, off-policy batched), but also the situation where online real-time policy learning is needed. Interior-point method is also a well-studies optimization method for constrained estimation. Its theoretical guarantees assure us of good enough optimal solutions. However, it is

obvious that the choice of policy function approximation may have impact on the decision. So does the choice of Q function approximation. Clinical domain expertise may required. Alternatively, automated feature learning techniques for function approximation from the machine learning community can be incorporated. More complex dataset maybe collected, such as text, image, speech and so on, considering the recent technology advancement in mobile devices. How to incorporate those complex information to better describe an individual's state of health is challenging. Rigorous theoretical work for our method is also under investigation.

Besides constraints on the expected value of a policy, we can also consider risk constraints, where the probabilities of adverse events occurring are restricted. Although reinforcement learning is a powerful technique to find optimal treatment regimes for clinical practice, designing appropriate reward functions is crucial for serving the desired clinical purpose, but very difficult. Current approach may not scale well in complex clinical situations or preventive healthcare where multiple subgoals may be involved. How to automatic generate rewards and objectives in complex clinical situations can be an interesting direction for investigation.

Nowadays, many aspects of the clinical practice have been transformed by mobile devices, such as smart phones, tablets, wearable sensors etc. It allows clinicians remotely monitor and intervene patients chronic conditions in real-time. It also allows for adaptive preventive interventions for motivating and maintaining healthy behaviors, such as physical exercise, diets, and so on. To better understand adaptive interventions, interdisciplinary collaborations becomes a necessity among clinicians, medical researchers, behavioral scientists, statisticians, and computer scientists.

# References


[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[2] E. Altman. *Constrained Markov Decision Processes*. Stochastic Modeling Series. Taylor & Francis, 1999.

[3] D Blatt, SA Murphy, and J Zhu. A-learning for approximate planning. *Ann Arbor*, 1001:48109–2122, 2004.

[4] Richard H. Byrd, Mary E. Hribar, and Jorge Nocedal. An Interior Point Algorithm for Large-Scale Nonlinear Programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.

[5] Ashkan Ertefaie. Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings. pages 1–39, 2014.

[6] Alborz Geramifard. A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning. *Foundations and Trends® in Machine Learning*, 6(4):375–451, 2013.

[7] Miguel A Hernan and James M Robins. Estimating causal effects from epidemiological data. *Journal of epidemiology and community health*, (7):578–86, July.

[8] Jerzy Splawa-Neyman, D. M. Dabrowska and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

[9] Eric B. Laber, Daniel J. Lizotte, and Bradley Ferguson. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61, 2014.

[10] Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. Dynamic treatment regimes : technical challenges and applications. 2014.

[11] Michail G. Lagoudakis and Ronald Parr. Model-Free Least-Squares Policy Iteration. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pages 1547–1554, 2001.

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

[13] Kristin A Linn, Eric B Laber, and Leonard A Stefanski. Interactive Q-learning for Probabilities and Quantiles. 2014.

[14] Daniel J Lizotte, Michael H. Bowling, and Susan A. Murphy. Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. *in Proc. of Int. Conf. on Machine Learning*, pages 695–702, 2010.

[15] Daniel J. Lizotte, Michael H. Bowling, and Susan A. Murphy. Linear Fitted-Q Iteration with Multiple Reward Functions. *Journal of Machine Learning Research*, 13:3253–3295, 2012.

[16] Daniel J. Lizotte and Eric B. Laber. Multi-objective markov decision processes for data-driven decision support. *J. Mach. Learn. Res.*, 17(1):7378–7405, January 2016.

[17] Daniel J. Luckett, Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth Mayer-Davis, and Michael R. Kosorok. Estimating Dynamic Treatment Regimes in Mobile Health Using V-learning. 2016.

[18] Erica Moodie. Dynamic treatment regimes. *Clinical trials (London, England)*, 1(5):471, 2004.

[19] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, May 2003.

[20] S A Murphy, Y Deng, E B Laber, H R Maei, R S Sutton, and K. Witkiewitz. A Batch, Off-Policy, Actor-Critic Algorithm for Optimizing the Average Reward. *arXiv*, pages 1–18, 2016.

[21] Susan A Murphy. A Generalization Error for Q-Learning. *Journal of machine learning research : JMLR*, 6:1073–1097, July 2005.

[22] James M. Robins. Causal Inference from Complex Longitudinal Data, 1997.

[23] D. B. Rubin. Discussion of Randomized analysis of experimental data: The Fisher randomization test by D. Basu. *Journal of the American Statistical Association*, (75):591–593, 1980.

[24] Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[25] 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.

[26] Baqun Zhang, Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

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

# Appendices

## A Least-squares policy evaluation (LSQ) algorithm

**Algorithm 3:** LSQ

$A \leftarrow 0$
$b \leftarrow 0$
for each $(s, a, s', r)$
    $A \leftarrow A + \phi(s,a)\left(\phi(s,a) - \gamma\phi(s', \pi(s'))\right)^{\intercal}$
    $b \leftarrow b + \phi(s,a)r$
$w \leftarrow A^{-1}b$
return $w$