Data-Driven Ranges of Near-Optimal Actions for Finite Markov Decision Processes

Wesley J. Marrero
MGH Institute for Technology Assessment, Harvard Medical School, Boston, MA 02114, wmarrerocolon@mgh.harvard.edu

Mariel S. Lavieri
Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI 48109, lavieri@umich.edu

Ambuj Tewari
Department of Statistics, University of Michigan, Ann Arbor, MI 48109, tewaria@umich.edu

Jeremy B. Sussman, Rodney A. Hayward
Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48105, jeremysu@med.umich.edu, rhayward@med.umich.edu

Markov decision process (MDP) models have been used to obtain non-stationary optimal decision rules in various applications, such as treatment planning in medical decision making. However, in practice, decision makers may prefer other strategies that are not statistically different from the optimal decision rules. To benefit from the decision makers' expertise and provide flexibility in implementing decision strategies, we introduce a new framework for identifying sets of near-optimal actions for finite MDP models. We present a simulation-based dynamic programming algorithm that can be executed using parallel computing and show that it converges to the optimal solutions exponentially fast under fairly mild conditions. The sets of near-optimal actions are modeled as nonparametric simultaneous confidence intervals on the difference between an approximately optimal action and the remaining alternatives. By analyzing the structure of the sets, we characterize their behavior with respect to the modeling data and identify when they can be ordered as a range. Lastly, we show the scalability of our approach by finding ranges of near-optimal antihypertensive treatment choices for 16.72 million adults in the US.

Key words: Markov decision processes, simulation, statistical multiple comparisons, medical decision making, health policy, cardiovascular diseases

1. Introduction

Markov decision process (MDP) models have been used to inform decisions in a wide variety of applications, including medicine, scheduling, transportation, finance, and energy. In many areas of application, such as the management of chronic conditions, the dynamics of the system of interest change over time. This type of problem generally has a finite set of periods when decisions must be made. If all the parameters of a non-stationary MDP are known with certainty, and there are a finite number of states and actions, the backwards induction algorithm can be used to find an optimal decision rule (Puterman 2014, Chang et al. 2013). However, there may be other decision

strategies that we cannot statistically differentiate from the optimal. Therefore, it is important to have a framework to identify sets of decision strategies that lead to similar outcomes.

Identifying sets of near-optimal actions may be useful in a variety of scenarios. When a human being is responsible for controlling a system, a single decision rule may not be enough, as each person has their own decision process (Fard and Pineau 2011). It may be appropriate to assume that some aspect of the decision-making process will be influenced by the decision maker (DM). Moreover, the difference between the performance of an optimal decision rule and other strategies may not be relevant to the DM. The DM could choose between an optimal action and another alternative with similar performance based on their expertise, preference, or other factors. In addition, models are typically estimated from observational data and multiple external sources. This may lead to optimal decision rules that do not perform well in the true system. In this case, the performance of an optimal decision rule may not be statistically different from other strategies. To test for statistical significance before observing the implications of each action in practice, our proposed strategy is to simulate the effect of each action based on an estimated model of the system of interest. We can then provide DMs with a set of actions, which might be optimal, but we do not have enough evidence to differentiate.

In this paper, we are motivated by circumstances in which several actions may have similar performance. We focus on improving the usability and acceptance of MDP models in practice by providing flexibility in implementing decision strategies. Rather than offering a single decision rule, we present DMs with a set of actions from which they may be able to choose. We introduce a new method to obtain sets of near-optimal actions and provide conditions for which the actions in these sets can be ordered as a range.

One area that could benefit from our proposed framework is medical decision making. Within the field of medicine, our methodology may be beneficial for the management of chronic conditions, such as atherosclerotic cardiovascular disease (ASCVD, constituting coronary heart disease (CHD) and stroke). We demonstrate the utility of our methodology by finding flexible treatment plans for the management of ASCVD.

1.1. Medical Decision Making and Atherosclerotic Cardiovascular Disease

MDP models have been widely used to guide medical decision making (Chanchaichujit et al. 2019). In medical decision-making models, the transition dynamics and rewards are often estimated using longitudinal patient data and results from the medical literature. As longitudinal data and medical results are derived with a finite number of observations, these estimates are subject to statistical uncertainty (Steimle et al. 2019).

Translating medical decision-making models into practice is difficult. General medical practitioners may interpret decision rules as cumbersome, confusing, and lacking in credibility (Cabana

et al. 1999). Therefore, it is important to consider practical implications in the design of decision rules. One way such implications can be considered is by providing clinicians with flexibility in implementing protocols while continuing to improve patient outcomes.

In this paper, we apply our proposed framework to manage ASCVD. According to the National Vital Statistics, ASCVD is the leading cause of death in the US (Kochanek et al. 2019). The Heart Disease and Stroke Statistics 2020 Update reports that CHD and stroke account for 42.6% and 17.0% of deaths attributable to cardiovascular diseases in the US, respectively (Virani et al. 2020). One of the leading controllable risk factors of ASCVD is hypertension or high blood pressure (BP). The most recent hypertension guidelines from the American College of Cardiology and American Heart Association (Whelton et al. 2018) have generated considerable controversy among practitioners (Cohen and Townsend 2018). To benefit from physicians' expertise and account for any potentially conflicting perspectives, we design personalized data-driven ranges of treatment options that are within a margin of certainty of the best treatment alternative, based on the estimated transition dynamics and rewards.

1.2. Contributions

We focus on adapting stochastic dynamic programming methods to offer multiple actions per state and time period. To provide easily interpretable decision support, we also obtain insights into how the sets of near-optimal actions behave with respect to the modeling data. Our approach is based on simulation-based dynamic programming (SBDP), nonoverlapping batch means, and statistical multiple comparisons with a control (MCC). Overall, the contributions of this work are as follows.

1. **We develop a new SBDP algorithm, which we will refer to as the simulation-based backwards induction (SBBI).** The idea behind this algorithm is to replace the expectation in the backwards induction algorithm with a sample-average approximation. This concept is widely used in the stochastic programming literature, but mainly for problems with a single time period (Birge and Louveaux 2011). By estimating action-value functions asynchronously at each time period, the SBBI algorithm can be executed using parallel computation.
2. **We provide finite sample, convergence, and asymptotic structural properties of the SBBI algorithm.** Our analysis leverages the Markov property to derive the rate of convergence and the sample complexity of the algorithm. We show that the SBBI algorithm converges almost surely (a.s.) uniformly on the action space under fairly mild conditions. To characterize the asymptotic structural properties of the algorithm, we establish connections between the standard backwards induction and the SBBI algorithm.
3. **We design a new MCC method, which we will refer to as the simulation-based MCC (SBMCC) algorithm.** In contrast to past MCC methods, our algorithm does not

make any distributional assumptions nor assumes equal variances across actions. The reasoning behind the SBMCC algorithm is to divide the output of a simulation model into batches to estimate the distribution of the standardized difference between an (approximately) optimal action and the remaining alternatives. Our approach extends current techniques to obtain confidence intervals in nonoverlapping batch mean methods (Alexopoulos and Goldsman 2004).

4. **We present a new notion of near-optimality by formulating stochastic optimization problems as hypothesis testing problems.** By evaluating the output of the SBBI algorithm as a random sample, we construct simultaneous confidence intervals. Any alternative that is not statistically different from an (approximately) optimal action is part of our sets of near-optimal actions. We highlight that our notion of near-optimality is different from the traditional sense in stochastic programming (Shapiro et al. 2009), as we measure near-optimality in terms of statistical significance.
5. **We offer convergence and asymptotic structural properties of the SBMCC algorithm.** Our convergence analysis includes asymptotic coverage guarantees of the SBMCC algorithm as well as the asymptotic rate of convergence of our algorithm. Taking advantage of the monotonicity of action-value functions and decision rules, we provide insights into how the sets of near-optimal actions will behave asymptotically. We also provide sufficient conditions to guarantee that a set of near-optimal actions will be ordered as a range. These structural results provide managerial insights into how to prioritize the comparison of actions to reduce computational overhead.
6. **We show the scalability of our approach to find ranges of near-optimal actions by applying our method for the management of hypertension.** Using a large sample representative of the adult population in the US with ages between 50 and 54 years old, we provide flexible data-driven decision support that is personalized to each patient's characteristics. We present the implications of flexible hypertension treatment plans at a patient and a population level.

1.3. Organization of the Paper

The remainder of this paper is organized as follows. We begin by providing a review of the relevant literature in Section 2. In Section 3, we provide additional background on MDP models and MCC. We formally define the sets and ranges of near-optimal actions in Section 4. In Section 5, we introduce our algorithms to obtain the sets of near-optimal actions as well as our analysis of the algorithms. We present our case study of flexible hypertension management in Section 6. Finally, conclusions and future research directions are discussed in Section 7.

2. Literature Review

The relevant literature to this research lies in the following fields: (1) simulation-based algorithms for MDP models; (2) methods for the output analysis of simulation models; (3) statistical multiple comparison approaches; (4) decision support models that provide sets of actions; and (5) medical decision-making models. This section highlights some prominent papers in each category and briefly describes how our proposed methodology differs from them.

The description of an MDP as a simulation model can be found in Chang et al. (2013). A large portion of simulation-based algorithms has focused on solving MDP models with large state spaces or in which there is no model of the system dynamics. These methods principally fall under the umbrella of approximate dynamic programming (ADP) or reinforcement learning (RL). Summaries of ADP/RL methods include Powell (2011) and Sutton and Barto (2018). Other simulation-based models have concentrated on solving MDP models with large action spaces (Chang et al. 2013). A key difference between our SBBI algorithm and the model-based methods in the ADP/RL literature is that our algorithm simulates each time period independently. In contrast, they simulate the system dynamics as episodes. Closest to our work, there have been methods that use simulation to estimate the expectation in dynamic programming with small to moderately sized state and action spaces (Powell 2011). In discounted infinite-horizon settings, Haskell et al. (2016) introduced simulation-based value iteration and policy iteration algorithms. Our method differs from the approaches presented by Haskell et al. (2016) in that we focus on finite horizon models and allow for random immediate rewards. Another closely related area to our work is the sample-average approximation in discrete stochastic programming (Kleywegt et al. 2002). Our SBBI algorithm is different from standard multistage discrete stochastic programming models in that actions affect the system dynamics in the future.

There are two notable approaches in steady-state simulation analysis to derive confidence intervals: nonoverlapping batch means and independent replications (Alexopoulos and Goldsman 2004). In our context, both approaches are equivalent as there is no initialization bias, and the observations are independent due to the Markov property. The nonoverlapping batch means and independent replication methods usually rely on asymptotic normality arguments to obtain confidence intervals. Steiger and Wilson (2002) introduced algorithms to attain confidence intervals of a specific precision hinging on normality tests. Our work presents a nonparametric approach to obtain confidence intervals in steady-state simulation analysis without any distributional assumptions.

Our work is also related to the theory of statistical multiple comparisons. Among the types of multiple comparisons, MCC is the most relevant to this research (Dunnett 1955). Similar to many classical statistical methods, MCC assumes normality and equal variances across alternatives. Westfall (2011) proposed a generalization of multiple comparison procedures that allowed for

general distributions using the bootstrap. Another line of work focused on developing MCC methods without the assumption of equal variances (Li and Ning 2012). Even though these alternative formulations allow for general distributions or unequal variances, none of them allow for both of them. Our SBMCC algorithm is a nonparametric approach that does not require equal variances.

There has been limited research in the area of decision support models that provide more than one decision rule. Laber et al. (2014) developed sets of decision rules in the context of dynamic treatment regimes based on clinically significant differences. A crucial distinction between this work and ours is that we focus on statistical significance instead of practical significance. Another difference is that we focus on Markov policies, whereas they center on history-dependent policies. This allows us to consider more than two decision epochs and actions. Ertefaie et al. (2016) also considered the problem of providing a set of suggestions in the context of dynamic treatment regimes. Although our approach has many similarities with this work, a vital distinction is that we identify a control before the statistical inference. This results in fewer comparisons and improved statistical power. Ertefaie et al. (2016) also concentrated on 2-stage history-dependent policies while we focus on Markov policies over a finite planning horizon. The closest work to this article is by Fard and Pineau (2011), where the authors consider the problem of developing sets of near-optimal actions for discounted infinite-horizon MDP models. Compared to this work, we define a new sense of near-optimality in terms of statistical significance, whereas they specify their near-optimality in the same units of the value function. A key contribution of our work to this field is that we characterize the behavior of the sets with respect to the modeling data, including the case where the actions contained in the sets can be ordered as a range.

Within the medical decision-making domain, treatment decision models in the literature include Liu et al. (2017), Negoescu et al. (2017, 2018), Ayer et al. (2019), and Chehrazi et al. (2019). Other studies have focused on finding the optimal time to gather additional information or for a screening procedure, including Sabouri et al. (2017), Hicklin et al. (2018), Suen et al. (2018), Agnihothri et al. (2018), Onen et al. (2018), Lee et al. (2018a), Lin et al. (2018), and Aprahamian et al. (2019). Within the context of cardiovascular diseases, treatment decision models include Lee et al. (2018b), and Zargoush et al. (2018). Researchers have also developed MDP models for the management of cardiovascular diseases (Schell et al. 2016, Steimle et al. 2019, Marrero et al. 2021). While these models recommend a single decision rule, our work provides physicians and their patients with flexibility in the implementation of treatment plans.

3. Overview and Background

In this paper, we focus on finding a sequence of sets of near-optimal actions in the context of discrete-time finite-horizon MDP models with finite state and action spaces. This section intro-

duces our modeling framework, the main notions behind finite MDP models and MCC, and our mathematical notation.

Our modeling framework for a decision period of an MDP is summarized in Figure 1. We begin by representing an MDP as a simulation model. Subsequently, we simulate the immediate rewards and the next state transitions for each state and action at every decision period of the MDP. Rather than using re-sampling techniques, we divide the outputs of the simulation model into batches because the DM can always manage their sample size. Then, we estimate action-value functions using SBDP. We use SBDP because it does not require knowledge of the true underlying probability distribution of the evolution of the system of interest. Instead, SBDP relies on sample realizations of the transition dynamics, which may be obtained through simulation.

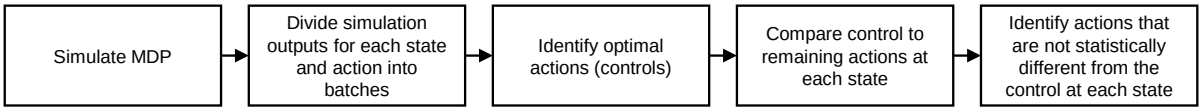

Figure 1 Summary of modeling framework for a decision period of a simulated MDP.

The SBDP framework also allows us to approximate optimal actions with a high degree of accuracy, which serve as controls in the multiple comparison procedures. We use MCC because, once an optimal action (or approximately optimal action) is identified as the control, we are interested in comparing the remaining alternatives with such control. The MCC method requires the least number of comparisons and provides the strongest inference for our purposes. Once the controls are compared to each of the remaining alternatives, we identify the decision strategies that are not statistically different from the optimal actions. Simulation MDP models and MCC are described in more detail in the following subsections.

3.1. Markov Decision Processes

MDP models are used to represent the interactions of a DM with a fully observable system of interest. We use the following notation throughout the paper:

- t: index of discrete time periods; $t \in \mathcal{T}$, where $\mathcal{T} \coloneqq \{1, \ldots, T\}$ is a finite set of time periods. Decisions are made until time $T-1$; the periods $\mathcal{T} \setminus \{T\}$ will be referred to as decision epochs.
- s: state of the system; $s \in \mathcal{S}$, where $\mathcal{S} \coloneqq \{1, \ldots, S\}$ is a finite set of states.
- a: DM's action; $a \in \mathcal{A}$, where $\mathcal{A} \coloneqq \{1, \ldots, A\}$ is a finite set of actions.
- ω: outcome of an exogenous process representing the uncertainty of the system; $\omega \in \Omega$, where Ω is the set of all outcomes.

- $f_{t+1}(s,a,\omega)$: transition function which produces the next state s' given state s, action a, and outcome of the exogenous process ω; $s' = f_{t+1}(s,a,\omega)$, where $f : \mathcal{S} \times \mathcal{A} \times \Omega \mapsto \mathcal{S}$.
- $r_t(s,a,\omega)$: reward associated with state s, action a, and outcome of the exogenous process ω, where the reward function is defined as $r : \mathcal{S} \times \mathcal{A} \times \Omega \mapsto \mathbb{R}_+ \coloneqq \{x \in \mathbb{R} | x \geq 0\}$.
- γ: discount factor of the model; $\gamma \in (0,1]$.

A simulation MDP is formally defined by the tuple $(\mathcal{T}, \mathcal{S}, \mathcal{A}, f, r, \gamma)$. Note that any simulation MDP $(\mathcal{T}, \mathcal{S}, \mathcal{A}, f, r, \gamma)$ with transition function f and rewards r can be transformed into a standard MDP $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \boldsymbol{P}, \rho, \gamma)$ with transition probabilities $\boldsymbol{P} \in [0,1]^{|\mathcal{T} \times \mathcal{S} \times \mathcal{S} \times \mathcal{A}|}$ and rewards $\rho : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}_+$. Simply let $p_t(s'|s,a) \coloneqq \mathbb{E}[\mathbb{1}\{f_{t+1}(s,a,\omega) = s'|s,a\}]$ and $\rho_t(s,a) \coloneqq \mathbb{E}[r_t(s,a,\omega)|s,a]$, where the expectation is taken over ω and $\mathbb{1}\{\cdot\}$ represents an indicator function.

The goal of the DM is to find the policy $\pi \coloneqq (\pi_t(s) : t \in \mathcal{T} \setminus \{T\}, s \in \mathcal{S})$ that maximizes the total expected discounted reward over the planning horizon, i.e.:

$$v_1(s) \coloneqq \max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T-1} \gamma^t r_t(s_t, a_t, \omega_t) + \gamma^T r_T(s_T, \omega_T) \middle| s_1 = s \right],$$

where, s_t, $a_t = \pi_t(s_t)$, and ω_t are the state, action, and random disturbance of the system at time period t, respectively. The expectation is taken with respect to the joint distribution of $\omega_1, \ldots, \omega_T$. Dividing the problem into decision epochs, we get the set of dynamic programming equations:

$$Q_t(s,a) \coloneqq \mathbb{E}\big[r_t(s,a,\omega) + \gamma v_{t+1}\left(f_{t+1}(s,a,\omega)\right) | s, a \big],$$

where $v_t(s) \coloneqq \max_{a \in \mathcal{A}} Q_t(s,a)$ and $Q_T(s,a) = v_T(s) = \mathbb{E}[r_T(s,\omega)|s]$. Starting from the terminal period T and proceeding backwards until decision epoch 1, we can find the optimal set of actions $\mathcal{A}_t^*(s) \coloneqq \arg\max_{a \in \mathcal{A}} Q_t(s,a)$ at each decision epoch t. An optimal decision rule at state s is given by $\pi_t^*(s) \in \mathcal{A}_t^*(s)$, and an optimal policy is defined as $\pi^* = (\pi_t^*(s) : t \in \mathcal{T} \setminus \{T\}, s \in \mathcal{S})$.

3.2. Multiple Comparisons With a Control

When a control or standard action is available, a DM may be interested in comparing the performance of every other action with the performance of such control. In this section, we adapt the ideas from MCC to the context of simulated MDP models. The parameters of interest are $Q_t(s,a^*) - Q_t(s,a)$ for $a^* \in \mathcal{A}_t^*(s)$ and $a \in \mathcal{A}_t^-(s)$, at each decision epoch t and state s. In here, $\mathcal{A}_t^*(s) \subset \mathcal{A}$ denotes the set of optimal actions (potential controls) and $\mathcal{A}_t^-(s) \subset \mathcal{A}$ denotes the set of sub-optimal actions. If a^* is an action such that $Q_t(s,a^*) \geq Q_t(s,a)$ for any $a \in \mathcal{A}$, our objective is to identify as many actions as possible to be inferior than a^*. Assuming equal sample sizes across actions, the $1-\alpha$ simultaneous confidence lower bounds for the difference between a control $Q_t(s,a^*)$ and the remaining action-value functions $\{Q_t(s,a) : a \in \mathcal{A}_t^-(s)\}$ are given by:

$$Q_t(s,a^*) - Q_t(s,a) > \hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) - d_t(s,\alpha)\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*) + \hat{\sigma}_t^2(s,a)\right]}, \tag{1}$$

where $\alpha \in (0,1)$, $\hat{Q}_t(s,a)$ is an estimate of the action-value function $Q_t(s,a) < \infty$, $\hat{\sigma}_t^2(s,a)$ is an estimate of the variance of the action-value function $\sigma_t^2(s,a) < \infty$, and N is the number of observations used to calculate $\hat{Q}_t(s,a)$ and $\hat{\sigma}_t^2(s,a)$. If there are reasons to believe that the action-value functions are normally distributed, the quantile $d_t(s,\alpha)$ can be obtained by solving a double integral whose numerical evaluations are readily available in standard statistical software (Dunnett 1955). However, in most practical situations this assumption may not be true. Westfall (2011) proposed an alternative formulation that allows for general probability distributions. Extending their formulation to unequal variances, we aim to find a quantile $d_t(s,\alpha)$ such that $\mathbb{P}\big(\max_{a\in\mathcal{A}} \psi_t(s,a) \leq d_t(s,\alpha)\big) = 1-\alpha$, where

$$\psi_t(s,a) \coloneqq \frac{\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) - (Q_t(s,a^*) - Q_t(s,a))}{\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*) + \hat{\sigma}_t^2(s,a)\right]}}, \tag{2}$$

is a root statistic corresponding to state s and action a.

4. Ranges of Near-Optimal Actions

From equation (1) we can conclude that an optimal action $a^* \in \mathcal{A}_t^*(s)$ is significantly better than some other action $a \in \mathcal{A}_t^-(s)$ at a significance level α if $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) - d_t(s,\alpha)\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*) + \hat{\sigma}_t^2(s,a)\right]} > 0$. Thus, we cannot conclude that action $a^* \in \mathcal{A}_t^*(s)$ is significantly different from $a' \in \mathcal{A}_t^-(s)$ if $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a') - d_t(s,\alpha)\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*) + \hat{\sigma}_t^2(s,a')\right]} \leq 0$. This leads to our definition of a set of near-optimal actions.

Definition 1. Given N observations, a state $s \in \mathcal{S}$, an optimal action $a^* \in \mathcal{A}_t^*(s)$ such that $Q_t(s,a^*) \geq Q_t(s,a)$ for all $a \in \mathcal{A}$, and a quantile $d_t(s,\alpha)$, a set of actions $\Pi_t(s,\alpha)$ is said to be *α-nonsignificant* with $\alpha \in (0,1)$ if it satisfies:

$$\Pi_t(s,\alpha) \coloneqq \left\{a \in \mathcal{A} : \hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) \leq d_t(s,\alpha)\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*) + \hat{\sigma}_t^2(s,a)\right]}\right\}.$$

A desirable property of $\Pi_t(s,\alpha)$ in practice is to be ordered according to the effect of the actions in the action space $\mathcal{A}_t(s) \subseteq \mathcal{A}$ at state s and decision epoch t, where $\mathcal{A}_t(s) = \mathcal{A}_t^*(s) \cup \mathcal{A}_t^-(s)$. For example, clinicians may find a set of treatment choices more interpretable if the set follows a natural order, such as increasing number of medications. An ordered set of actions in the sense of Definition 2 will be referred as a range of actions.

Definition 2. $\Pi_t(s,\alpha)$ is said to be a *range of α-nonsignificant actions* if it satisfies Definition 1 and if $a, a'' \in \Pi_t(s,\alpha)$ implies that $a' \in \Pi_t(s,\alpha)$ for any actions ordered as $a \leq a' \leq a''$ in $\mathcal{A}_t(s)$.

5. Solution Approach

This section describes our approach to identify sets of α-nonsignificant actions. We first introduce our SBBI algorithm. Subsequently, we study the finite sample, convergence, and asymptotic structural properties of the algorithm. We then present our SBMCC method. Lastly, we examine the asymptotic behavior of our approach and characterize the sets of near-optimal actions.

5.1. Simulation-Based Backwards Induction

We now introduce our algorithm to estimate the action-value functions and identify approximately optimal actions. The algorithm is presented for finding actions $a^* \in \mathcal{A}_t^*(s)$ such that $Q_t(s,a^*) \geq Q_t(s,a)$ for all $a \in \mathcal{A}$ at each $s \in \mathcal{S}$. We aim to estimate $Q_t(s,a)$, $v_t(s)$, and $\mathcal{A}_t^*(s)$ for $t \in \mathcal{T} \setminus \{T\}$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$.

Our SBBI algorithm is included as Algorithm 1. Given a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define a discrete-time stochastic process $(\omega^n : n \in \{1,\ldots,N\})$ as the exogenous information process in the sequential decision problem. Without loss of generality, we let $(\omega^n : n \in \{1,\ldots,N\})$ be a sequence of independent and identically distributed (iid) random variables uniformly distributed on $[0,1]$, denoted by $\mathcal{U}(0,1)$. This is consistent with past work on simulating MDP models (Chang et al. 2013, Haskell et al. 2016).

Algorithm 1: Simulation-based backwards induction (SBBI) algorithm.

Input : $\mathcal{T}$, $\mathcal{S}$, $\mathcal{A}$, N, f, r, $\{v_T(s) : s \in \mathcal{S}\}$.

for $t = T-1, T-2, \ldots, 1$ **do**
  Set $\boldsymbol{Q}_t(s,a) \leftarrow \emptyset$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.
  for $n \leftarrow 1$ **to** N **do in parallel**
    for all $s \in \mathcal{S}$ **do**
      for all $a \in \mathcal{A}$ **do**
        Simulate $\omega^n \sim \mathcal{U}(0,1)$ and determine $s' = f_{t+1}(s,a,\omega^n)$.
        Compute $Q_t^n(s,a) = r_t(s,a,\omega^n) + \gamma \hat{v}_{t+1}(s')$.
        Update $\boldsymbol{Q}_t(s,a) \leftarrow \boldsymbol{Q}_t(s,a) \cup \{Q_t^n(s,a)\}$.
      end for
    end for
  end forpar
  Calculate $\hat{Q}_t(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.
  Set $\hat{\mathcal{A}}_t^*(s) \leftarrow \arg\max_{a \in \mathcal{A}} \hat{Q}_t(s,a)$ for all $s \in \mathcal{S}$.
end for

Output: $\{\boldsymbol{Q}_t(s,a) : t \in \mathcal{T} \setminus \{T\}, s \in \mathcal{S}, a \in \mathcal{A}\}$, $\{\hat{Q}_t(s,a) : t \in \mathcal{T} \setminus \{T\}, s \in \mathcal{S}, a \in \mathcal{A}\}$, $\{\hat{\mathcal{A}}_t^*(s) : t \in \mathcal{T} \setminus \{T\}, s \in \mathcal{S}\}$.

At each decision epoch t, state s, and action a, we simulate a sequence $\boldsymbol{Q}_t(s,a) \coloneqq (Q_t^n(s,a) : n \in \{1,\ldots,N\})$ of $N \in \mathbb{N}_+ \coloneqq \mathbb{N} \setminus \{0\}$ observations. Once the simulation is complete, we approximate the action-value function $Q_t(s,a)$ by its sample mean $\hat{Q}_t(s,a) \coloneqq \frac{1}{N}\sum_{n=1}^N Q_t^n(s,a) = \frac{1}{N}\sum_{n=1}^N r_t(s,a,\omega^n) + \gamma \hat{v}_{t+1}(f_{t+1}(s,a,\omega^n))$. From $\hat{Q}_t(s,a)$, we estimate the value function $v_t(s)$ as $\hat{v}_t(s) \coloneqq \max_{a \in \mathcal{A}} \hat{Q}_t(s,a)$, the set of optimal actions $\mathcal{A}_t^*(s)$ as $\hat{\mathcal{A}}_t^*(s) \coloneqq \arg\max_{a \in \mathcal{A}} \hat{Q}_t(s,a)$, the set

of sub-optimal actions $\mathcal{A}_t^-(s)$ as $\hat{\mathcal{A}}_t^-(s) \coloneqq \mathcal{A} \setminus \hat{\mathcal{A}}_t^*(s)$, and the optimal decision rules as $\hat{\pi}_t^*(s) \in \hat{\mathcal{A}}_t^*(s)$. If the terminal conditions $v_T(s)$ are not known, we also estimate them through their sample mean.

Given ω^n, $f_{t+1}(s,a,\omega^n)$ and $r_t(s,a,\omega^n)$ are deterministic functions of s and a. That is, the distribution of $Q_t^n(s,a)$ is completely determined by ω^n. Because of the Markov property, $Q_t^n(s,a)$ is independent from $Q_t^{n'}(s,a)$ for any $n \neq n'$. Therefore, $\boldsymbol{Q}_t(s,a)$ is a sequence of iid random variables. This allows us to simulate $\boldsymbol{Q}_t(s,a)$ in parallel, which leads to great computational speed gains. We use $\mathbb{F}_t(\cdot,s,a)$ to denote the cumulative distribution function (cdf) of $Q_t^n(s,a)$ and $\mathbb{F}_t(\cdot,s)$ to denote the joint cdf of the set $\{Q_t^n(s,a) : a \in \mathcal{A}\}$. The empirical estimates of the distribution functions are denoted by $\hat{\mathbb{F}}_t(\cdot,s,a)$ and $\hat{\mathbb{F}}_t(\cdot,s)$, respectively.

5.1.1. Finite Sample Properties of the SBBI Algorithm. In this subsection, we provide results on the behavior of the SBBI algorithm with a finite number of observations. The proofs of the claims in this subsection can be found in Appendix A.1. We begin with the following assumption:

ASSUMPTION 1. *The immediate rewards $r_t(s,a,\omega)$ are known constants or non-negative iid random variables from a possibly unknown probability distribution bounded by $R_t(s,a) < \infty$. Further, the terminal rewards $r_T(s,\omega)$ are known constants or non-negative iid random variables from a possibly unknown probability distribution bounded by $R_T(s) < \infty$.*

This assumption implies that the action-value functions $Q_t(s,a)$ are bounded and that their estimates $\hat{Q}_t(s,a)$ are bounded random variables. We now state a result on the convergence rate of the SBBI algorithm.

PROPOSITION 1. *Under Assumption 1, it follows that* $1 - \mathbb{P}(\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)) \leq A \exp\left\{\frac{-N}{2\kappa_t^2}\right\}$, *with* $\kappa_t \coloneqq \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$, *where* $R_t \coloneqq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} R_t(s,a)$.

Proposition 1 implies that the SBBI algorithm converges exponentially fast on N. This finding implies that the SBBI algorithm can efficiently estimate $Q_t(s,a)$, provided that κ_t is not too large compared to N. The result in Proposition 1 leads to the required sample size to guarantee that any action in $\hat{\mathcal{A}}_t^*(s)$ is an optimal action with probability of at least $1-\beta$.

PROPOSITION 2. *Suppose Assumption 1 holds. Then, for any $\beta \in (0,1)$ and a fixed sample size N satisfying $N \geq 2\kappa_t^2 \log(A/\beta)$ it holds that $\mathbb{P}(\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)) \geq 1-\beta$.*

In the context of medical decision-making problems, such as our case study, $\mathcal{A}_t^*(s)$ is usually a singleton. In this case, Proposition 2 gives the number of observations required to ensure that $\mathbb{P}(\hat{\mathcal{A}}_t^*(s) = a^*) \geq 1-\beta$ for $\mathcal{A}_t^*(s) = \{a^*\}$.

5.1.2. Analysis of the SBBI Algorithm. We now present our results on the convergence of the SBBI algorithm. The proofs of the claims in this subsection can be found in Appendix A.2. We begin by showing the uniform almost sure convergence of $\hat{Q}_t(s,a)$ to $Q_t(s,a)$:

THEOREM 1. *Suppose Assumption 1 holds. Then, $\hat{Q}_t(s,a)$ converges to $Q_t(s,a)$ with probability 1 uniformly on $\mathcal{A}$.*

Uniform convergence implies that the required number of observations for convergence is independent of the action. The following corollary is an immediate consequence of Theorem 1:

COROLLARY 1. *Suppose Assumption 1 holds. Then, $\hat{v}_t(s)$ converges to $v_t(s)$ and $\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)$ with probability 1 for N large enough.*

We use similar arguments as in Proposition 2.1 of Kleywegt et al. (2002) to prove Corollary 1 and use their definition for the statement "an event happens with probability 1 for N large enough." An event happens with probability 1 for N large enough if for $\mathbb{P}$-almost every realization of a random sequence $\boldsymbol{\omega} \coloneqq \{\omega^1, \omega^2, \ldots\}$ for $\omega^1, \omega^2, \ldots \in \Omega$ there exists an integer $N(\boldsymbol{\omega})$ such that the event happens for all samples $\{\omega^1, \ldots, \omega^N\} \in \boldsymbol{\omega}$ with $N \geq N(\boldsymbol{\omega})$.

5.1.3. Asymptotic Structural Properties of the SBBI Algorithm. We proceed to present the asymptotic structural properties of the SBBI algorithm. The proofs of the claims in this subsection can be found in Appendix A.3. Structured policies tend to be more intuitive for DMs and are typically easier to implement in practice. Throughout this subsection, we discuss the implications of the results using a medical decision-making example. We begin with the following definition:

DEFINITION 3. Let X be a partially ordered set and $g : X \mapsto \mathbb{R}$. We say that g is an *ϵ-nonincreasing* (*ϵ-nondecreasing*) function if for $x^+ \geq x^-$ in X it holds that $g(x^+) \leq g(x^-) + \epsilon$ $\big(g(x^+) \geq g(x^-) - \epsilon\big)$ for $\epsilon > 0$. Similarly, we say that g is *ϵ-constant* if $|g(x^+) - g(x^-)| \leq \epsilon$ for $\epsilon > 0$.

Definition 3 will be useful to deal with random variables that converge to the same quantity. We use the term ϵ-monotone to describe a function that is either ϵ-nonincreasing or ϵ-nondecreasing. If the sequences of random variables converge to different quantities, the definition of a ϵ-monotone function reduces to the standard definition of a monotone function (using $\epsilon = 0$).

Let $\bar{p}_t(s'|s,a) \coloneqq \sum_{s'' \geq s'} p_t(s''|s,a)$ denote the tail distribution of the transition probabilities, or the probability that the state at decision epoch $t+1$ exceeds s' after choosing action a at state s and decision epoch t. We make the following assumption:

ASSUMPTION 2. *The state space $\mathcal{S}$ can be ordered such that the tail distribution functions $\bar{p}_t(s'|s,a)$ are nondecreasing in s and t, the expected immediate rewards $\mathbb{E}[r_t(s,a,\omega)|s,a]$ are nonincreasing in s and t, and the terminal rewards are nonincreasing in s and $\mathbb{E}[r_{T-1}(s,a,\omega)|s,a] \geq \mathbb{E}[r_T(s,\omega)|s]$.*

The conditions in Assumption 2 with respect to s are the same as in Proposition 4.7.3 in Puterman (2014). This assumption, along with Assumption 1, provide sufficient conditions to ensure that the estimates of the action-value functions and the value functions are ϵ-monotone on s and t for N large enough. We present the ϵ-monotonicity in s in the following proposition:

PROPOSITION 3. *Under Assumptions 1 and 2, $\hat{Q}_t(s,a)$ and $\hat{v}_t(s)$ are ϵ-nonincreasing in s with probability 1 for N large enough.*

In medical decision-making settings, states commonly represent patients' health conditions, and actions represent clinical interventions. The action-value and value functions typically represent a measure of how long and how well patients are expected to live given a clinical intervention, such as life years or quality-adjusted life years. If the health conditions are ordered from the healthiest to the sickest, Proposition 3 implies that sicker patients will have a shorter total expected lifetime than healthier patients. The following proposition presents the ϵ-monotonicity of the estimates of the action-value and value functions in t for N large enough.

PROPOSITION 4. *Suppose Assumptions 1 and 2 hold. Then, $\hat{Q}_t(s,a)$ and $\hat{v}_t(s)$ are ϵ-nonincreasing in t with probability 1 for N large enough.*

In medical decision-making problems, Proposition 4 suggests that patients' total expected lifetime would never increase with their age, a common assumption in the medical literature.

5.2. Simulation-Based Multiple Comparisons with a Control

In this section, we present our method to identify sub-optimal actions that are not statistically different from an (approximately) optimal choice at a significance level α. To derive a set of near-optimal actions, we compare the performance of an (approximately) optimal action with the rest of the alternatives. Similar to Westfall (2011), our formulation aims to find a constant $d_t(s,\alpha)$ such that $\mathbb{P}\big(\max_{a\in\mathcal{A}}\psi_t(s,a) \le d_t(s,\alpha)\big) = 1-\alpha$. Since $Q_t(s,a)$ is unknown, so are $\max_{a\in\mathcal{A}}\psi_t(s,a)$ and its cdf. We denote this cdf as $\mathbb{H}_t$ and use $\mathbb{H}_t(\cdot,\mathbb{F}_t(s))$ when its dependence on the unknown cdf $\mathbb{F}_t(\cdot,s)$ must be emphasized. To address this challenge, we adapt the concept of nonoverlapping batch means to simulated MDP models.

Consider N observations of a simulated MDP. The method of nonoverlapping batch means divides the sequence of N outputs of a simulation into M adjacent nonoverlapping batches, each of size K. Because $\boldsymbol{Q}_t(s,a)$ is a sequence of iid random variables, dividing N outputs of a simulated MDP into M batches is equivalent to executing M independent simulations of the MDP, each with K observations. The m^{th} batch (or simulation replicate) consists of the random variables: $Q_t^{m,1}(s,a), Q_t^{m,2}(s,a), \ldots, Q_t^{m,K}(s,a)$, for $m = 1,\ldots,M$. For each batch m, we estimate the action-value functions and their variance using the sample mean and sample variance over K observations.

After batching, the grand sample mean can be obtained as $\hat{Q}_t(s,a) = \frac{1}{M}\sum_{m=1}^{M}\bar{Q}_t^m(s,a) = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}Q_t^{m,k}(s,a)$ and the variance of the batch sample means as $\hat{\zeta}_t^2(s,a) = \frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{Q}_t^m(s,a) - \hat{Q}_t(s,a)\right)^2$, where $\bar{Q}_t^m(s,a)$ is the sample mean for the m^{th} batch. We obtain an estimate of $\sigma_t^2(s,a)$ by multiplying $\hat{\zeta}_t^2(s,a)$ by the number of observations per batch K. The nonoverlapping batch means method allows us to generate an estimator for $\psi_t(s,a)$ by replacing $\sigma_t^2(s,a)$ for $K\hat{\zeta}_t^2(s,a)$ in equation (2). We use $\hat{\psi}_t(s,a)$ to denote the estimator of $\psi_t(s,a)$.

Using the variability across the M batches, we generate an empirical estimate of $\mathbb{H}_t$, denoted by $\hat{\mathbb{H}}_t(\cdot, \hat{\mathbb{F}}_t(s))$ or simply $\hat{\mathbb{H}}_t$ when is not necessary to highlight its dependence to the empirical cdf $\hat{\mathbb{F}}_t(\cdot, s)$. We now introduce our algorithm to generate $\hat{\mathbb{H}}_t$ and estimate the quantile $d_t(s,\alpha)$ for each state s and decision epoch t. Our SBMCC method is included as Algorithm 2. Let $\mathcal{C}_t(s) \in \mathcal{C}$ denote the set of controls associated with state s and decision epoch t, where $\mathcal{C} := \{\mathcal{C}_t(s) : t \in \mathcal{T} \setminus \{T\}, s \in \mathcal{S}\}$. For each batch m, we generate an estimate of the root statistic as:

$$\bar{\psi}_t^m(s,a) := \frac{\bar{Q}_t^m(s,a^*) - \bar{Q}_t^m(s,a) - \left(\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a)\right)}{\sqrt{K^{-1}\left[\bar{\sigma}_t^2(s,a^*,m) + \bar{\sigma}_t^2(s,a,m)\right]}},$$

where $a^* \in \mathcal{C}_t(s)$ and $\bar{\sigma}_t^2(s,a,m)$ is the sample variance of the m^{th} batch.

Algorithm 2: Simulation-based multiple comparisons with a control (SBMCC) algorithm.

Input : $\mathcal{T}$, $\mathcal{S}$, $\mathcal{A}$, $\mathcal{C}$, M, $\{\bar{Q}_t^m(s,a) : m \in \{1,\ldots,M\}, t \in \mathcal{T}\setminus\{T\}, s \in \mathcal{S}, a \in \mathcal{A}\}$, $\{\hat{Q}_t(s,a) : t \in \mathcal{T}\setminus\{T\}, s \in \mathcal{S}, a \in \mathcal{A}\}$.

for all $t \in \mathcal{T}\setminus\{T\}$ **do**
  Set $\hat{\Psi}_t(s) \leftarrow \emptyset$ for all $s \in \mathcal{S}$.
  for $m \leftarrow 1$ **to** M **do in parallel**
    for all $s \in \mathcal{S}$ **do**
      for all $a \in \mathcal{A}$ and $a^* \in \mathcal{C}_t(s)$ **do**
        Calculate $\bar{\psi}_t^m(s,a)$.
      end for
      Update $\hat{\Psi}_t(s) \leftarrow \hat{\Psi}_t(s) \cup \{\max_{a\in\mathcal{A}} \bar{\psi}_t^m(s,a)\}$.
    end for
  end forpar
  Let $\hat{d}_t(s,\alpha)$ be the empirical $1-\alpha$ quantile of $\hat{\Psi}_t(s)$ for all $s \in \mathcal{S}$.
end for

Output: $\{\hat{d}_t(s,\alpha) : t \in \mathcal{T}\setminus\{T\}, s \in \mathcal{S}\}$.

A key assumption in MCC is that the control is known before observing the data that will be used to evaluate the actions. Thus, we must generate $\mathcal{C}_t(s)$ without knowing $\hat{Q}_t(s,a)$. Several

approaches could be used to identify $\mathcal{C}_t(s)$, such as solving the standard version of the MDP $(\mathcal{T}, \mathcal{S}, \mathcal{A}, \boldsymbol{P}, \rho, \gamma)$ through backwards induction before simulating the MDP and letting $\mathcal{C}_t(s) = \mathcal{A}_t^*(s)$. Alternatively, we could simulate an initial independent replication of the SBBI algorithm and let $\mathcal{C}_t(s) = \hat{\mathcal{A}}_t^*(s)$. Another approach could be to use the estimates of a single batch and let $\mathcal{C}_t(s) = \arg\max_{a \in \mathcal{A}} \bar{Q}_t^m(s,a)$. In the later case, the batch used to obtain $\mathcal{C}_t(s)$ must be excluded from Algorithm 2. When $\mathcal{C}_t(s)$ is obtained from a simulation model, Proposition 2 provides a lower bound on the sample size required such that $\mathcal{C}_t(s) \subseteq \mathcal{A}_t^*(s)$ with high probability.

Once we generate an estimate of the root statistic for each batch, we estimate $d_t(s,\alpha)$ as $\hat{d}_t(s,\alpha) \coloneqq \inf\left\{x \in \mathbb{R} : \hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) \geq 1-\alpha\right\}$. Since the estimates of the root statistics $\{\bar{\psi}_t^m(s,a) : m \in \{1,\ldots,M\}, t \in \mathcal{T} \setminus \{T\}, s \in \mathcal{S}, a \in \mathcal{A}\}$ are mutually independent, Algorithm 2 can be executed in parallel across M, $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{T} \setminus \{T\}$. However, we present the parallel execution of the algorithm across M, $\mathcal{S}$, and $\mathcal{A}$ as it allows for the integration of the SBMCC algorithm to the SBBI algorithm. This integration allows us to investigate the effect of future α-nonsignificant actions in $\Pi_t(s,\alpha)$. Merely replace $\hat{v}_{t+1}(s')$ by $\hat{Q}_{t+1}(s',\tilde{a})$ for $\tilde{a} \in \Pi_{t+1}(s',\alpha)$ in line 7 of Algorithm 1.

5.2.1. Analysis of the SBMCC Algorithm. We now proceed to present our asymptotic results of the SBMCC algorithm. The proofs of the claims in this subsection can be found in Appendix A.4. Let $\Theta \subseteq \mathbb{R}$ denote the set of all possible values of $Q_t(s,a^*) - Q_t(s,a)$. The following proposition shows that the SBMCC algorithm produces the correct overall asymptotic coverage.

PROPOSITION 5. *Suppose Assumption 1 holds. Then, for any $\alpha \in (0,1)$, $a^* \in \mathcal{C}_t(s)$, and N large enough we have that $\mathbb{P}\Big(Q_t(s,a^*) - Q_t(s,a) \in \Theta : \hat{\mathbb{H}}_t\left(\max_{a \in \mathcal{A}}\{\hat{\psi}_t(s,a)\}, \hat{\mathbb{F}}_t(s)\right) \leq 1-\alpha\Big) = 1-\alpha$.*

The result in Proposition 5 implies that the true difference between the performance of a control action and the remaining actions will asymptotically be in a subset of Θ such that $\mathbb{P}\big(\max_{a \in \mathcal{A}} \hat{\psi}_t(s,a) \leq d_t(s,\alpha)\big)$ is exactly the desired confidence level, $1-\alpha$. While all the conditions in Θ involve random variables, all the relevant quantities converge with probability 1 to their true values as $M \to \infty$ and $K \to \infty$ (see Lemmas 5 and 6 in Appendix A.4). Note that our method has similar asymptotic coverage properties to the nonparametric bootstrap method (Tu and Zhou 2000). Proposition 5 allows us to show that a set of actions $\Pi_t(s,\alpha) \subseteq \mathcal{A}_t(s)$ with a quantile $\hat{d}_t(s,\alpha)$ derived from the SBMCC algorithm will asymptotically be a set of α-nonsignificant actions with probability 1. We present this result in the following theorem:

THEOREM 2. *Suppose Assumption 1 holds. Then, for $\hat{d}_t(s,\alpha) = \hat{\mathbb{H}}_t^{-1}(1-\alpha, \hat{\mathbb{F}}_t(s))$ and $a^* \in \mathcal{C}_t(s)$, we have that $\Pi_t(s,\alpha) = \left\{a \in \mathcal{A} : \hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) \leq \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a)\right]}\right\}$ is a set of α-nonsignificant actions with probability 1 for N large enough.*

Theorem 2 generalizes the theoretical basis of MCC as described in Section 3 of Dunnett (1955) to the nonparametric case. This theorem also extends the simultaneous confidence interval methods for MCC without the equal variances assumption described in Section 2 of Li and Ning (2012). It is worth observing that Theorem 2, as well as the results in Dunnett (1955) and Li and Ning (2012), are based on the implicit null hypothesis that all actions are equally good. We now state a result on the rate of convergence of the SBMCC algorithm.

PROPOSITION 6. *Suppose that Assumption 1 is satisfied. Then, it follows that* $\lim_{N\to\infty} N^{1/2} \sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x, \mathbb{F}_t(s))\right| \leq CA^{5/4}\sqrt{2\kappa_t^3}$, *where* C *is the constant appearing in the multivariate Berry-Esseen bound.*

Although the value of the constant C is an open area of research, its current best estimate is $C = 42A^{1/4} + 16$ by Raic (2019). The result in Proposition 6 provides a bound on the convergence rate of Algorithm 2 of order $\mathcal{O}(N^{-1/2})$. We note that this rate of convergence is equivalent to the convergence rate of the central limit theorem (Serfling 1980, Theorem 1.9.5).

5.2.2. Asymptotic Structural Properties of the SBMCC Algorithm. We provide asymptotic structural results of the SBMCC algorithm next. The proofs of the claims in this subsection can be found in Appendix A.5. Similar to Subsection 5.1.3, we discuss the connotations of the assumptions and results in the context of medical decision-making scenarios. We start by stating the relationship between the cardinality of the sets of α-nonsignificant actions and the significance level α.

PROPOSITION 7. *Suppose that Assumption 1 holds. Then,* $|\Pi_t(s, \alpha)|$ *is nonincreasing in* $\alpha \in (0, 1)$. *Moreover, there exist an* α *such that* $\Pi_t(s, \alpha) \subseteq \mathcal{A}_t^*(s)$ *with probability 1 for* N *large enough.*

Under the classical null hypothesis that all actions are equally good, the significance level α indicates the strength of the evidence that a DM, such as a clinician or a patient, requires before concluding that there is sufficient evidence to reject the null hypothesis. The result in Proposition 7 means that if we increase the significance level (e.g., from 0.01 to 0.05), the sets of α-nonsignificant actions will not include more choices. This result allows clinicians to control the cardinality of the sets of α-nonsignificant actions based on their confidence in the rewards and transition functions. If clinicians are not exceptionally confident in the parameterization of their model, smaller values of α, such as 0.01 and 0.05, would be reasonable. On the other hand, if clinicians are highly confident in their model, larger values of α, such as 0.1 and 0.2, may suffice.

We now present sufficient conditions to ensure that the cardinality of the sets of α-nonsignificant actions are ϵ-monotone for N large enough. Consistent with the work of Mannor and Tsitsiklis (2013), we find that there are no recursive algorithms when dealing with the variance of value

functions. As a result, we make assumptions directly on the conditional variance of the value functions. The following assumption gives conditions on the conditional variance of the rewards and value functions:

ASSUMPTION 3. *The conditional variances* $\mathbb{E}[r_t^2(s,a,\omega)|s,a] - \mathbb{E}[r_t(s,a,\omega)|s,a]^2$ *and* $\mathbb{E}[v_t^2(s')|s,a] - \mathbb{E}[v_t(s')|s,a]^2$ *are nonincreasing in* s*,* t*, and* $a \in \mathcal{A}_t(s)$*.*

Recall our medical decision-making setting from Subsection 5.1.3 where the time periods and states represent patients' age and health conditions, respectively, and the actions are clinical interventions at different intensities. Health conditions are ordered from the healthiest to the sickest, and clinical interventions are ordered from the lowest to the highest intensity. The action-value and value functions represent how long and how well patients are expected to live given a clinical intervention. In this setting, Assumption 3 indicates that the effect of clinical interventions will not become more uncertain when patients are sicker or older. It also suggests that more intense clinical interventions will be more certain than less aggressive interventions. Although $\mathbb{E}[v_t^2(s')|s,a] - \mathbb{E}[v_t(s')|s,a]^2$ is not part of the basic model, the conditions can be directly verified after obtaining $v_t(s)$ either via standards backwards induction or after approximating it through $\hat{v}_t(s)$ with the SBBI algorithm.

Besides requirements on the conditional variances, we make an assumption on the nature of the differences of the tail distribution functions and the expected immediate rewards:

ASSUMPTION 4. *The tail distribution functions* $\bar{p}_t(s'|s,a)$ *are subadditive functions on* $\mathcal{S} \times \mathcal{A}$ *and* $\mathcal{T} \times \mathcal{A}$*. Further, the expected immediate rewards* $\mathbb{E}[r_t(s,a,\omega)|s,a]$ *are superadditive functions on* $\mathcal{S} \times \mathcal{A}$ *and* $\mathcal{T} \times \mathcal{A}$*.*

Assumption 4 means that the impact of more intense clinical interventions on the probability that patients get sicker is larger if the patient is sicker or older. It also implies that more aggressive interventions have a larger effect on patients' health when they are sicker or older.

Incorporating Assumptions 1 through 4 provide sufficient conditions to make sure that the sets of α-nonsignificant actions are ϵ-monotone on s for N large enough. We present this result in the following proposition:

PROPOSITION 8. *Suppose Assumptions 1 through 4 hold. Then,* $|\Pi_t(s,\alpha)|$ *is* ϵ*-nonincreasing in* s *with probability 1 for* N *large enough.*

The result in Proposition 8 means that sicker patients will receive less choices than healthier patients. This may be useful from a clinical perspective because sicker patients are more likely to experience adverse events. To ensure that the sets of α-nonsignificant actions are ϵ-monotone on t we also need conditions on the action space $\mathcal{A}_t(s) \subseteq \mathcal{A}$ associated with state s and decision epoch t. We make the following assumption:

ASSUMPTION 5. *The action space $\mathcal{A}_t(s)$ can be ordered such that the tail distribution functions $\bar{p}_t(s'|s,a)$ are nonincreasing in $a \in \mathcal{A}_t(s)$ and the expected immediate rewards $\mathbb{E}[r_t(s,a,\omega)|s,a]$ are nondecreasing in $a \in \mathcal{A}_t(s)$.*

If clinical interventions are ordered from the lowest to the highest intensity, Assumption 5 implies that the more aggressive interventions will result in better immediate and future expected health outcomes. Integrating Assumption 5 to Assumptions 1 through 4 provides sufficient conditions to assure that the sets of α-nonsignificant actions will be ϵ-monotone on t for N large enough.

PROPOSITION 9. *Suppose Assumptions 1 through 5 hold and that $v_t(s) - v_{t+1}(s)$ is nondecreasing in s. Then, $|\Pi_t(s,\alpha)|$ is ϵ-nonincreasing in t with probability 1 for N large enough.*

The additional condition in Proposition 9 indicates that healthier patients will experience smaller differences in terms of life expectancy and quality of life than sicker patients over the planning horizon. Proposition 9 means that patients will receive less choices in their sets of α-nonsignificant actions as they get older. This result may be useful in clinical practice because older patients are typically more likely to experience adverse events. We highlight that the conditions in Propositions 8 and 9 are also sufficient to show ϵ-monotonicity in the approximately optimal decision rules $\hat{\pi}_t^*(s)$ for N large enough in the following remark:

REMARK 1. The conditions in Proposition 8 and Proposition 9 are sufficient to prove that there exists approximately optimal decision rules $\hat{\pi}_t^*(s)$ that are ϵ-monotone on s and t with probability 1 for N large enough, respectively. We prove this remark in Appendix A.5.

Combining Assumptions 1, 2, 3, and 5 we get sufficient conditions to guarantee that the actions contained in the sets of α-nonsignificant actions are arranged as a range (see Definition 2). To account for the possibility of ties among actions, we present our result in terms of ϵ-monotonicity:

THEOREM 3. *Suppose that Assumptions 1, 2, 3, and 5 are satisfied. Then, $\Pi_t(s,\alpha)$ is a ϵ-range of α-nonsignificant actions at state s and decision epoch t with probability 1 for N large enough.*

If there are two clinical interventions of varying intensities contained in a set of α-nonsignificant actions, the result in Theorem 3 implies that any clinical intervention with an intensity between them will also be included in the set. Moreover, once a clinical intervention is proven not to be part of $\Pi_t(s,\alpha)$, it is certain that any more extreme intervention will also not be part of the range. This results in computational gains, especially in the case of large action spaces. A range of near-optimal actions may be more intuitive and interpretable for clinicians than a set without any particular order. As a result, the ranges of α-nonsignificant actions may be easier to implement in medical practice. Note that if $\{Q_t(s,a) : a \in \mathcal{A}\}$ have equal variances (i.e., $\sigma_t^2(s,a) = \sigma_t^2(s,a')$ for all $a \neq a'$),

Assumption 3 is not required to show that $\Pi_t(s,\alpha)$ is a range of actions at state s with probability 1 for N large enough.

A consequence of the conditions in this subsection, in particular Assumption 5, is that clinicians do not need to assume that patients will receive optimal interventions in the next decision epoch. If these conditions are satisfied, clinicians could use any decision rule in the next decision epoch and asymptotically reach at least a subset of the recommendations in the current period. Let $\mathscr{A}_t^*(s,\tilde{a})$ and $\mathit{\Pi}_t(s,\alpha,\tilde{a})$ denote the set of optimal actions and the set of α-nonsignificant actions, respectively, assuming that action $\tilde{a} \in \mathcal{A}_{t+1}(f_{t+1}(s,a,\omega))$ is taken at the next decision epoch. We present our result in the following proposition:

PROPOSITION 10. *Suppose Assumptions 1, 2, 4 and 5 hold. Then, we have that* $\mathscr{A}_t^*(s,\tilde{a}) \subseteq \mathcal{A}_t^*(s)$ *and* $\mathit{\Pi}_t(s,\alpha,\tilde{a}) \subseteq \Pi_t(s,\alpha)$ *for* N *large enough.*

The result in Proposition 10 indicates that DMs do not need to know which decision rule will be followed in the subsequent decision epochs when selecting the set of α-nonsignificant actions in the current period, if Assumptions 1, 2, 4, and 5 are satisfied. This result may be beneficial in clinical practice as future clinical interventions may be unclear due to the uncertainty in patients' health progression. This result provides clinicians and their patients with confidence in the sets of near-optimal actions in the current decision epoch without the burden of potential ambiguity in patients' future health.

6. Case study: Personalized Hypertension Treatment Plans

In this section, we apply our methodology to obtain flexible hypertension treatment plans for the primary prevention of ASCVD. We first provide some background on hypertension treatment and motivate the need for flexible protocols. Next, we describe our MDP, data source, model parameters, and simulation framework. Lastly, we present patients' treatment plans and health outcomes following treatment choices contained in our ranges of near-optimal actions.

6.1. Background on Hypertension Treatment

Among adults with no history of cardiovascular diseases, 40.2% have hypertension (Lamprea-Montealegre et al. 2018). Using the definition from the 2017 Hypertension Clinical Practice Guidelines, stage 1 hypertension is defined as systolic blood pressure (SBP) of 130-139 mm Hg or diastolic blood pressure (DBP) of 80-89 mm Hg (Whelton et al. 2018). Stage 2 hypertension is defined as an SBP of at least 140 mm Hg or a DBP of at least 90 mm Hg. These guidelines provide non-pharmacological and pharmacological recommendations for patients with hypertension and elevated BP, defined as an SBP of 120-129 mm Hg and a DBP smaller than 80 mm Hg. In this

case study, we focus on pharmacological treatment for the primary prevention of ASCVD (i.e., to avoid the first ASCVD event). Our goal is to reduce the prevalence of ASCVD before it develops.

A critical distinction between clinical practice guidelines, such as Whelton et al. (2018), and the optimal decision models in the literature is that they provide clinicians with flexibility in the implementation of hypertension treatment plans. To benefit from clinicians' judgment and account for their patients' preferences, we develop ranges of near-optimal treatment choices for the personalized management of hypertension.

6.2. Markov Decision Process Formulation

The process of sequentially determining antihypertensive medications over a planning horizon is modeled as a finite MDP. We adapt the standard MDP formulation in Schell et al. (2016) to a primary prevention simulation MDP. The objective of the MDP model is to determine the treatment strategy that maximizes the expected discounted life years before an adverse event (i.e., an ASCVD event or death). The elements of our simulation MDP $(\mathcal{T}, \mathcal{S}, \mathcal{A}, f, r, \gamma)$ are as follows:

- $\mathcal{T}$: 10-year planning horizon; $\mathcal{T} = \{1, \ldots, 11\}$, where decisions are made at the beginning of each year $t \in \mathcal{T} \setminus \{11\}$. We use $T = 11$ to represent the effects of treatment on patients' lifetime. This planning horizon is selected based on conversations with our clinical collaborators and the major guidelines for the management of cardiovascular diseases (Whelton et al. 2018).
- $\mathcal{S}$: state space comprising patients' demographic information, clinical observations, and health condition. We separate the state space $\mathcal{S}$ into healthy states $\mathcal{H}$ (before adverse events) and absorbing states $\mathcal{E}$ (after adverse events), based on patients' health conditions (i.e., $\mathcal{S} = \mathcal{H} \cup \mathcal{E}$).
- $\mathcal{A}$: action space composed of 0 to 5 antihypertensive medications at half and standard dosage, for a total of $A = 21$ treatment choices. These medications may include thiazide diuretics, beta-blockers, calcium channel blockers, angiotensin-converting enzyme inhibitors, and angiotensin II receptor blockers. We focus on the number of medications since research suggests that the benefit from antihypertensive treatment is determined by the BP reduction achieved, with little effect attributable to drug-specific factors (Sundström et al. 2014).
- $f_{t+1}(s, a, \omega)$: transition function derived from patients' risk for ASCVD events, the benefit from treatment, fatality likelihoods, and non-ASCVD mortality.
- $r_t(s', a, \omega)$: reward associated with a transition to state s' after action a and outcome ω. We define $r_t(s', a, \omega) = 1 - \Delta(a)$ if $s' \in \mathcal{H}$ and 0 otherwise, where $\Delta(a)$ denotes the treatment-related disutility from medication a.
- $r_T(s', \omega)$: patients' expected lifetime associated with a transition to state s'.
- γ: discount factor of the model. We use $\gamma = 0.97$ as per recommendations in the medical literature (Neumann et al. 2016).

The clinical parameters used throughout our numerical study are listed in Table 1. We provide additional details on the ordering of $\mathcal{S}$ and $\mathcal{A}_t(s)$ to satisfy the conditions established on Theorem 3 in Appendix B.1.

Table 1 Base case parameters

Parameter	Value	Source
BP reduction: standard dosage (half dosage)		
SBP	5.5 (3.7) mm Hg	Sundström et al. (2014), Sussman et al. (2013)
DBP	3.3 (2.2) mm Hg	Sundström et al. (2014), Sussman et al. (2013)
Risk for ASCVD events	Varies by patient	Yadlowsky et al. (2018)
ASCVD risk reduction: standard dosage (half dosage)		
CHD	13% (7%)	Sundström et al. (2014), Sussman et al. (2013)
Stroke	21% (14%)	Sundström et al. (2014), Sussman et al. (2013)
ASCVD risk due to CHD	70%	Virani et al. (2020)
Mortality from ASCVD events		
CHD	Varies by patient	NCHS (2017)
Stroke	Varies by patient	NCHS (2017)
Treatment-related disutility		
Half dosage	0.001	Schell et al. (2016), Sussman et al. (2013)
Full dosage	0.002	Schell et al. (2016), Sussman et al. (2013)
Life expectancy	Varies by patient	Arias and Xu (2019)
Non-ASCVD mortality	Varies by patient	Arias and Xu (2019)

6.3. Data Source

To parameterize our models, we use the National Health and Nutrition Examination Survey (NHANES) dataset from 2009 to 2016. Our primary sample is composed of adult Black or White patients from 50 to 54 years old with no history of heart attack, stroke, or congestive heart failure, for a total population of 16.72 million people. We choose this age group based on conversations with our clinical collaborators. This age group represents a young population with a high prevalence of ASCVD that can benefit significantly from hypertension treatment (Virani et al. 2020, Whelton et al. 2018). Any missing data in the NHANES dataset is imputed using the MissForest package in R. To model how each patient's risk factors may evolve over time, we estimate their progression using linear regression. We regress patients' untreated SBP, total cholesterol (TC), high-density lipoprotein (HDL), and low-density lipoprotein (LDL) on their age, sex, race, smoking status, and diabetes status. The intercept term of each regression model is adjusted by applying the difference between the linear regression fitted value and the observed value in the NHANES data.

6.4. Simulation Framework

We develop a simulation model to evaluate our ranges of α-nonsignificant treatment choices. For comparison purposes, we also evaluate optimal treatment plans as described in Schell et al. (2016) and the 2017 Hypertension Clinical Practice Guidelines (Whelton et al. 2018).

The trajectory of a single patient in our modeling framework is summarized in Figure 2. Before developing treatment plans, we calculate the risk for ASCVD events each year. We then estimate transition probabilities and develop transition functions. Subsequently, we determine the different treatment policies. To derive our ranges of near-optimal treatment choices, we combine the SBBI and SBMCC algorithms. The control treatment choices are identified using the estimates of the first batch of outputs from the SBBI algorithm.

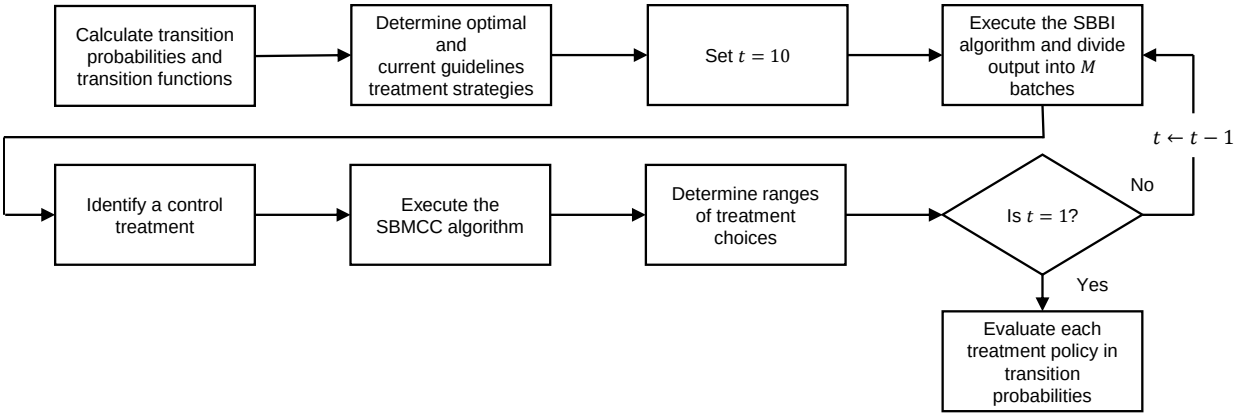

Figure 2 Summary of simulation framework for a single patient. The index t represents the year in the planning horizon (10 years).

We consider three types of treatments contained in the ranges of near-optimal actions: the best performing treatment in the range, the treatment choice with the median number of medications, and the treatment choice with the least amount of medications. These strategies will be referred to as the best in range, median in range, and fewest in range, respectively. We choose the first type of treatment choice to mimic the kind of patient who wants the best possible treatment and completely adheres to prescriptions. Note that the best in range strategy corresponds to the approximately optimal treatment choice obtained with the SBBI algorithm. The other two types of treatment plans (i.e., median in range and fewest in range) aim to represent potential physicians' reactions to patients' nonadherence. Research has shown that prescribing fewer medications may increase patients' adherence to prescriptions (Saini et al. 2009). Once we obtain the different strategies, we evaluate each policy based on the Markov chain embedded in the MDP. The calibration and validation of our simulation model are described in Appendix B.2.

6.5. Analysis

Before evaluating the impact of flexible treatment plans, we derive the number of batches to divide the output of the simulated MDP for each patient in our population (see Appendix B.3). A significance level of $\alpha = 0.05$ is used for all analyses.

To understand the implications of flexible treatment plans at a patient level, we examine the effect of the characteristics of a patient on the width of the ranges of treatment choices. We then study the policy implications of the ranges of near-optimal treatment alternatives by comparing our methodology to the optimal treatment plans and the 2017 Hypertension Clinical Practice Guidelines (current clinical guidelines). First, we inspect the distribution of the number of medications recommended by each treatment strategy. Next, we examine the life-years saved, ASCVD events averted, and expected time to an adverse event (including ASCVD events and non-ASCVD related death) by each treatment strategy, compared to no treatment. Lastly, we explore the proportion of patients in which the ranges of near-optimal actions includes the optimal treatment and the current clinical guidelines (Appendix B.6).

To study the policy implications of each treatment strategy, we divide our population by sex, race, and BP group. We create the BP groups based on the 2017 Hypertension Clinical Practice Guidelines: normal BP, elevated BP, stage 1 hypertension, and stage 2 hypertension. We also perform sensitivity analysis on the treatment strategies by varying the model parameters and assumptions. Our sensitivity analyses are included in Appendix B.7.

6.6. Numerical Results

In this subsection, we evaluate the effect of flexible hypertension treatment plans. We provide insights into the patient and population-level implications of flexible treatment. As we focus on the primary prevention of ASCVD, all the results in this section correspond to patients in the healthy states $\mathcal{H}$. The results of our sensitivity analyses are included in Appendix B.7.1.

6.6.1. Patient-Level Insights from Flexible Treatment. We now evaluate the ranges of near-optimal actions in a series of patient profiles based on patients from the NHANES dataset. For comparison purposes, we also determine the optimal treatment plans and the current clinical guidelines for each patient profile. We first obtain ranges of antihypertensive medications for the following patient profile: a 54-year-old, non-diabetic, non-smoker White male with stage 1 hypertension and low TC, HDL, and LDL. This patient profile will be referred to as the base patient profile. Note that this profile has two major risk factors for ASCVD events, the BP and the HDL levels. We then modify the following characteristics of the patient: sex, race, diabetes status, smoking status, and age.

Figure 3 shows the ranges of near-optimal treatment choices for our selection of patient profiles. We observe in the base patient profile ranges from 4 to 7 treatment choices. They correspond to recommending from 0 to 2 medications at a standard dosage and 1 at half dosage over the planning horizon. We also notice that the best treatment in the ranges matches the optimal treatment every year. This is different from the current clinical guidelines, which do not recommend any pharmacological treatment until year 10, when the patient reaches a 10-year risk for ASCVD events slightly above 10%. The ranges of α-nonsignificant treatment choices identify the years in which the current clinical guidelines are not statistically different from the optimal policy.

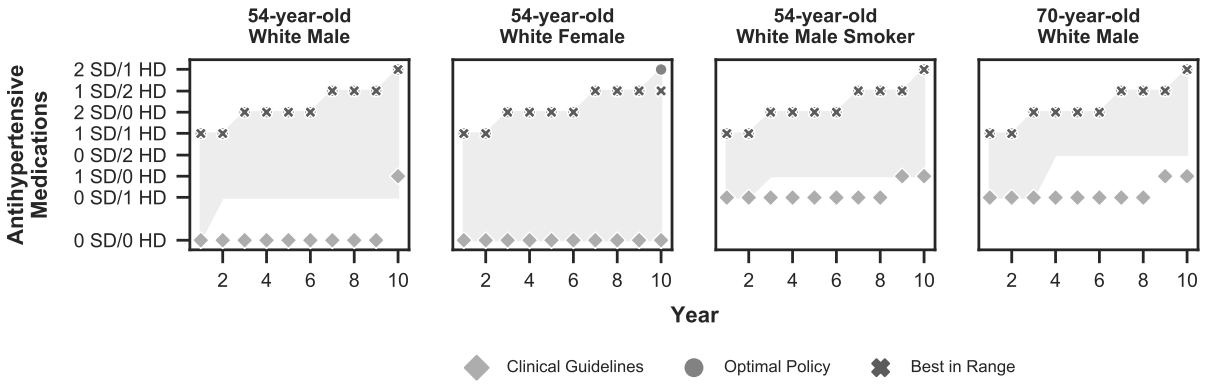

Figure 3 Ranges of near-optimal treatment choices per patient profile. The ranges are highlighted with the gray shaded area in each profile. The labels "SD" and "HD" denote antihypertensive medications at standard dosage and half dosage, respectively.

Changing the sex of the patient to female increases the width of the ranges but decreases the level of aggressiveness of the treatment prescribed. Using this patient profile, we find that the best treatment in range is slightly lower than the optimal treatment at year 10 of our study. However, the optimal treatment is contained in the range, and the optimality gap is relatively small (0.0026 life-years saved). We find a similar behavior by changing the race of the patient to Black. Overall, we notice that a lower risk for ASCVD results in wider ranges and less aggressive treatment. These results are consistent with existing literature (Yadlowsky et al. 2018).

Modifying the diabetes and smoking status of the profile seems to have comparable effects. In both cases, the ranges become narrower, and more aggressive treatment is prescribed. We also note that the current clinical guidelines recommend more aggressive treatment as the risk for ASCVD events increases. Increasing the age of the base patient profile to 70 years has the biggest impact on the behavior of the ranges. We discover ranges from 4 to 5 treatment choices that correspond to recommending from 1 medication at half dosage to 2 medications at a standard dosage and 1 at half dosage over the planning horizon.

6.6.2. Population-Level Insights from Flexible Treatment. We now study the policy implications of providing flexibility in the management of hypertension. The composition of our population is described in Appendix B.4.

Comparison of Treatment Recommendations. By comparing the treatment strategies contained in our ranges of near-optimal actions to the optimal treatment policies and the current clinical guidelines, we can obtain insights into how treatment changes by demographic. The distribution of the treatment recommended by each policy per BP category at years 1 and 10 of our study is shown in Figure 4. Other than more intense treatment over time, we note that the treatment distribution did not change considerably in years 2 through 9. The results segregated by BP category, sex, and race are shown in Figure B.3 in Appendix B.5.

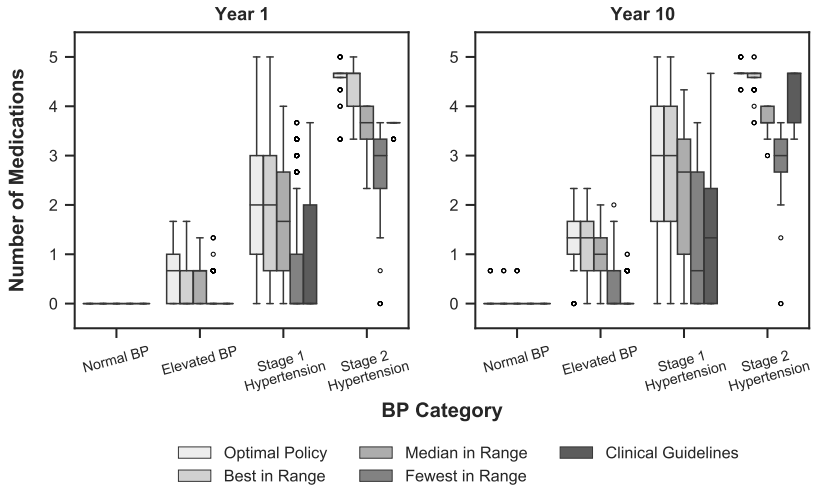

Figure 4 Distribution of treatment at year 1 and year 10 of the study. BP categories are made based on patients' characteristics at year 1.

From the distribution of treatment recommendations, we observe that very few patients receive treatment in the normal BP category at any given year (less than 0.3% of the population). Comparing the treatment strategies contained in our ranges to the optimal treatment plans and the current clinical guidelines, we observe that recommending the best treatment in the range is typically close to optimal. The difference between the optimal policy and the best in range lies in the amount of information each treatment strategy requires. The optimal policy assumes complete certainty of the transition probabilities and immediate rewards, while the best in range only requires samples from their distributions. We find that recommending the best in range is often equivalent to recommending the largest number of medications in the ranges. The best treatment in range is between the optimal treatment and the current guidelines treatment strategy, but considerably closer to optimal treatment in all BP categories. We also note that the current clinical guidelines

are between the fewest and the median in range for patients with normal BP, elevated BP, and stage 1 hypertension. The current clinical guidelines generally recommend between the largest number of medications and the median in range for patients with stage 2 hypertension.

Effect of Treatment Recommendations. Since very few people receive treatment under any of the policies in the normal BP category, we focus on patients with elevated BP, stage 1 hypertension, and stage 2 hypertension. We proceed to evaluate the outcomes of patients under each treatment strategy in terms of the life-years saved, the ASCVD events prevented, and the expected time to an adverse event, compared to no treatment. In total, the best treatment, the median number of medications, and the fewest number of medications contained in the ranges save 2.92, 2.55, and 1.75 million life years over the planning horizon, compared to no treatment. The optimal treatment plans and the clinical guidelines save 3.02 and 1.83 million life years, respectively.

Evaluating our results by BP category, we find that patients with stage 1 hypertension receive the greatest benefit from treatment (Figure 5). We note that patients' health outcomes under the best treatment choice in the ranges are not substantially different from the health outcomes based on optimal treatment. In patients with elevated BP, our treatment strategies outperform the clinical guidelines. The health outcomes of patients with stage 1 hypertension under the clinical guidelines are similar to the life-years saved under the strategy that recommends the fewest number of medications contained in the ranges. All the treatment policies result in similar life-years saved in patients with stage 2 hypertension. Note that the current clinical guidelines are near-optimal in most patients with stage 2 hypertension (see Figure B.5 in Appendix B.6). We include our results separated by sex, race, and BP category in Figure B.4 in Appendix B.5.

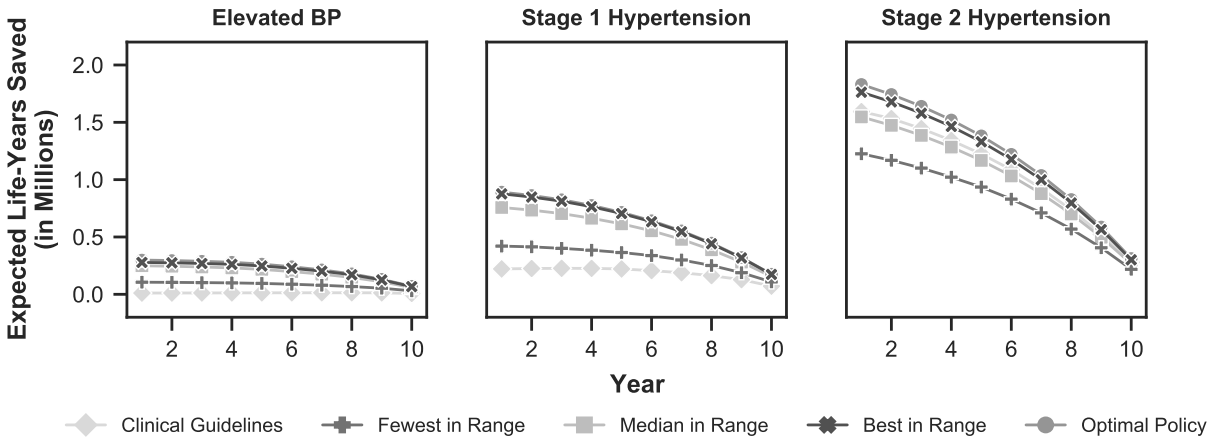

Figure 5 **Life-years saved by each treatment policy compared to no treatment per BP group over the planning horizon.**

We notice a similar pattern when comparing the policies in terms of ASCVD events averted. Over the 10-year planning horizon, the best in range, median in range, and fewest in range prevent 176, 154, and 103 thousand ASCVD events, compared to no treatment. The optimal treatment and current guidelines prevent 180 and 110 thousand ASCVD events, respectively.

Regarding the expected time to an adverse event, we also observe that the best in range is close to the optimal policies. The fewest in range is similar to the clinical guidelines. We notice that events are expected to be delayed in 5.48, 4.61, and 2.93 life years by the best in range, median in range, and fewest in range, respectively. The optimal treatment and the current guidelines delay adverse events in 5.60 and 2.96 life years, respectively.

7. Conclusions

This paper introduced a new method to obtain sets of near-optimal actions in finite MDP models. Additionally, we presented an alternative notion of optimality, which we called α-nonsignificance. We propose two algorithms to obtain the sets of α-nonsignificant actions: the SBBI and the SBMCC algorithm. The SBBI algorithm is based on the standard backwards induction algorithm, and it was created by replacing the expectation with a sample-average approximation. We showed that the estimates attained with the SBBI algorithm converge to their true values with probability 1 exponentially fast. The SBMCC algorithm leverages ideas from the nonoverlapping batch means method to produce simultaneous confidence intervals without any distributional or equal variance assumptions. We proved that the method reaches the correct coverage asymptotically and that the rate of convergence is bounded by a rate of order $\mathcal{O}(N^{-1/2})$. In addition, we provided sufficient conditions to ensure the monotonicity of the sets of α-nonsignificant actions in time and states. Lastly, we gave conditions to guarantee that the sets of near-optimal actions will be ordered as a range. By providing DMs with a set of actions from which they can choose, we improve the usability and acceptance of MDP models in practice.

In our case study, we examined the implications of flexible hypertension treatment plans at a patient and a population level. Several conclusions can be made from this study. First, how much flexibility a physician may receive to treat a patient depends on the patient's characteristics (e.g., age, sex, and race). In general, we find that patients with higher risk for ASCVD events obtain fewer treatment choices in the ranges than patients with lower risk. Second, our treatment strategies outperform the clinical guidelines in patients with elevated BP and stage 1 hypertension. This is an indication that the current clinical guidelines may be under-treating some patients and over-treating other patients. As a result, we observe life-year losses compared to the strategies contained in the ranges of α-nonsignificant treatment choices. The clinical guidelines and our treatment strategies result in similar health outcomes for patients with stage 2 hypertension. Finally, the estimates of

the benefit from treatment have a major effect on the type of treatment patients receive, and the amount of flexibility clinicians may have to treat patients. As new evidence of the effectiveness of BP treatment becomes available, the ranges of near-optimal antihypertensive medications may become more accurate in medical practice.

There are opportunities for future work that build upon our ranges of α-nonsignificant actions. An extension of this work could be to adapt our methodology to partially observable and infinite-horizon MDP models. For the latter type of models, the ideas from Haskell et al. (2016) could be used to obtain empirical estimates of the value and action-value functions. It may also be worthwhile to allow for continuous state and action spaces. Structural results may be necessary to guarantee convergence in these cases. We acknowledge that the SBBI and SBMCC algorithms inherit some of the curses of dimensionality associated with standard backwards induction. Overcoming these difficulties may be an additional area for future work. Additionally, our algorithms are limited by their storage requirements. This issue could be addressed by developing an online method to obtain ranges of α-nonsignificant actions.

Our work could be extended from a clinical perspective by incorporating other conditions, such as high cholesterol or diabetes. Based on communications with our clinical collaborators, we decided to develop ranges of antihypertensive medications as a starting point. Integrating the treatment of multiple conditions will likely result in greater flexibility. Our results provide a lower bound on the amount of flexibility clinicians and their patients could receive in implementing decision strategies. Our work could also be extended by allowing for multiple scenarios of the input data, which may be useful when there is disagreement on how to model patients' health progression. This source of variability could then be used to obtain sets of near-optimal treatment choices. While this modeling framework can be beneficial to provide options when there is debate on how to model the evolution of a system, the methods introduced in this paper provide flexibility when there is a well-calibrated model of the system dynamics such as Yadlowsky et al. (2018).

The ranges of α-nonsignificant actions present a new line of work by handling stochastic optimization problems as hypothesis testing problems. Providing several suggestions at each state and decision epoch in a sequential decision problem presents domain experts with an effective way to integrate their knowledge into mathematical models. A range of near-optimal choices could have many benefits in practice, such as better user experience and flexibility.

References

Agnihothri S, Cui L, Delasay M, Rajan B (2018) The value of mHealth for managing chronic conditions. *Health Care Management Science* ISSN 1386-9620, URL http://dx.doi.org/10.1007/s10729-018-9458-2.

Alexopoulos C, Goldsman D (2004) To batch or not to batch? *ACM Transactions on Modeling and Computer Simulation* 14(1):76–114, ISSN 10493301, URL http://dx.doi.org/10.1145/974734.974738.

Aprahamian H, Bish DR, Bish EK (2019) Optimal Risk-Based Group Testing. *Management Science* 65(9):4365–4384, ISSN 0025-1909, URL http://dx.doi.org/10.1287/mnsc.2018.3138.

Arias E, Xu J (2019) United States Life Tables, 2017. *National Vital Statistics Reports* 68(7).

Ayer T, Zhang C, Bonifonte A, Spaulding AC, Chhatwal J (2019) Prioritizing hepatitis C treatment in U.S. Prisons. *Operations Research* 67(3):853–873, ISSN 15265463, URL http://dx.doi.org/10.1287/opre.2018.1812.

Birge JR, Louveaux F (2011) *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering (New York, NY: Springer New York), ISBN 978-1-4614-0236-7, URL http://dx.doi.org/10.1007/978-1-4614-0237-4.

Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PAC, Rubin HR (1999) Why Don't Physicians Follow Clinical Practice Guidelines? *Journal of the American Medical Association* 282(15):1458, ISSN 0098-7484, URL http://dx.doi.org/10.1001/jama.282.15.1458.

Chanchaichujit J, Tan A, Meng F, Eaimkhong S (2019) *Optimization, Simulation and Predictive Analytics in Healthcare*, 95–121 (Singapore: Springer Singapore), ISBN 978-981-13-8114-0, URL http://dx.doi.org/10.1007/978-981-13-8114-0_5.

Chang HS, Hu J, Fu MC, Marcus SI (2013) *Simulation-Based Algorithms for Markov Decision Processes*. Communications and Control Engineering (Springer London), ISBN 978-1-4471-5021-3, URL http://dx.doi.org/10.1007/978-1-4471-5022-0.

Chehrazi N, Cipriano LE, Enns EA (2019) Dynamics of drug resistance: Optimal control of an infectious disease. *Operations Research* 67(3):619–650, ISSN 15265463, URL http://dx.doi.org/10.1287/opre.2018.1817.

Cohen JB, Townsend RR (2018) The ACC/AHA 2017 Hypertension Guidelines: Both Too Much and Not Enough of a Good Thing? *Annals of Internal Medicine* 168(4):287, ISSN 0003-4819, URL http://dx.doi.org/10.7326/M17-3103.

Dunnett CW (1955) A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association* 50(272):1096–1121, ISSN 1537274X, URL http://dx.doi.org/10.1080/01621459.1955.10501294.

Ertefaie A, Wu T, Lynch KG, Nahum-Shani I (2016) Identifying a set that contains the best dynamic treatment regimes. *Biostatistics* 17(1):135–148, ISSN 14684357, URL http://dx.doi.org/10.1093/biostatistics/kxv025.

Fard MM, Pineau J (2011) Non-deterministic policies in markovian decision processes. *Journal of Artificial Intelligence Research* 40:1–24, ISSN 10769757, URL http://dx.doi.org/10.1613/jair.3175.

Haskell WB, Jain R, Kalathil D (2016) Empirical Dynamic Programming. *Mathematics of Operations Research* 41(2):402–429, ISSN 0364-765X, URL http://dx.doi.org/10.1287/moor.2015.0733.

Hicklin K, Ivy JS, Payton FC, Viswanathan M, Myerse E (2018) Exploring the value of waiting during labor. *Service Science* 10(3):334–353, ISSN 0022-0507, URL http://dx.doi.org/10.1017/S002205070001648X.

Kleywegt AJ, Shapiro A, Homem-De-Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12(2):479–502, ISSN 10526234, URL http://dx.doi.org/10.1137/S1052623499363220.

Kochanek KD, Murphy SL, Xu J, Arias E (2019) Deaths: final data for 2017. *National Vital Statistics Reports* 68(9):1–18, ISSN 15518930.

Laber EB, Lizotte DJ, Ferguson B (2014) Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* 70(1):53–61, ISSN 15410420, URL http://dx.doi.org/10.1111/biom.12132.

Lamprea-Montealegre JA, Zelnick LR, Hall YN, Bansal N, De Boer IH (2018) Prevalence of hypertension and cardiovascular risk according to blood pressure thresholds used for diagnosis. *Hypertension* 72(3):602–609, ISSN 15244563, URL http://dx.doi.org/10.1161/HYPERTENSIONAHA.118.11609.

Lee E, Lavieri MS, Volk M (2018a) Optimal screening for hepatocellular carcinoma: a restless bandit model. *Manufacturing & Service Operations Management* (January 2019):msom.2017.0697, ISSN 1523-4614, URL http://dx.doi.org/10.1287/msom.2017.0697.

Lee EK, Wei X, Baker-Witt F, Wright MD, Quarshie A (2018b) Outcome-Driven Personalized Treatment Design for Managing Diabetes. *Interfaces* 48(5):422–435, ISSN 0092-2102, URL http://dx.doi.org/10.1287/inte.2018.0964.

Li H, Ning W (2012) Multiple comparisons with a control under heteroscedasticity. *Journal of Applied Statistics* 39(10):2275–2283, ISSN 02664763, URL http://dx.doi.org/10.1080/02664763.2012.706269.

Lin Y, Huang S, Simon GE, Liu S (2018) Data-based decision rules to personalize depression follow-up. *Scientific Reports* 8(1):4–11, ISSN 20452322, URL http://dx.doi.org/10.1038/s41598-018-23326-1.

Liu S, Brandeau ML, Goldhaber-Fiebert JD (2017) Optimizing patient treatment decisions in an era of rapid technological advances: the case of hepatitis C treatment. *Health Care Management Science* 20(1):16–32, ISSN 13869620, URL http://dx.doi.org/10.1007/s10729-015-9330-6.

Mannor S, Tsitsiklis JN (2013) Algorithmic aspects of mean-variance optimization in Markov decision processes. *European Journal of Operational Research* 231(3):645–653, ISSN 03772217, URL http://dx.doi.org/10.1016/j.ejor.2013.06.019.

Marrero WJ, Lavieri MS, Sussman JB (2021) Optimal cholesterol treatment plans and genetic testing strategies for cardiovascular diseases. *Health Care Management Science* ISSN 1386-9620, URL http://dx.doi.org/10.1007/s10729-020-09537-x.

NCHS (2017) Health, United States, 2016: with chartbook on long-term trends in health. *Center for Disease Control* 314–317, URL https://www.cdc.gov/nchs/data/hus/hus16.pdf{#}019.

Negoescu DM, Bimpikis K, Brandeau ML, Iancu DA (2017) Dynamic learning of patient response types: an application to treating chronic diseases. *Management Science* (January 2019):mnsc.2017.2793, ISSN 0025-1909, URL http://dx.doi.org/10.1287/mnsc.2017.2793.

Negoescu DM, Bimpikis K, Brandeau ML, Iancu DA (2018) Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science* 64(8):3469–3488, ISSN 15265501, URL http://dx.doi.org/10.1287/mnsc.2017.2793.

Neumann P, Sanders G, Russell L, Siegel J (2016) *Cost-effectiveness in health and medicine* (Oxford University Press).

Onen Z, Sayin S, Gurvit b (2018) Optimal population screening policies for Alzheimer's disease. *IISE Transactions on Healthcare Systems Engineering* 5579:1–36, ISSN 2472-5579, URL http://dx.doi.org/10.1080/24725579.2018.1543738.

Powell WB (2011) *Approximate Dynamic Programming*. Wiley Series in Probability and Statistics (Hoboken, NJ, USA: John Wiley & Sons, Inc.), 2 edition, ISBN 9781118029176, URL http://dx.doi.org/10.1002/9781118029176.

Puterman ML (2014) *Markov decision processes: discrete stochastic dynamic programming* (John Wiley & Sons), ISBN 978-0-471-72782-8.

Raic M (2019) A multivariate Berry – Esseen theorem with explicit constants. *Bernoulli* 25:2824–2853, URL http://dx.doi.org/https://doi.org/10.3150/18-BEJ1072.

Sabouri A, Huh WT, Shechter SM (2017) Screening strategies for patients on the kidney transplant waiting list. *Operations Research* 65(5):1131–1146, ISSN 0030-364X, URL http://dx.doi.org/10.1287/opre.2017.1632.

Saini SD, Schoenfeld P, Kaulback K, Dubinsky MC (2009) Effect of medication dosing frequency on adherence in chronic diseases. *The American journal of managed care* 15(6):e22–33, ISSN 1936-2692, URL http://www.ncbi.nlm.nih.gov/pubmed/19514806.

Schell GJ, Marrero WJ, Lavieri MS, Sussman JB, Hayward RA (2016) Data-driven Markov decision process approximations for personalized hypertension treatment planning. *MDM Policy & Practice* 1(1), ISSN 2381-4683, URL http://dx.doi.org/10.1177/2381468316674214.

Serfling RJ (1980) *Approximation Theorems of Mathematical Statistics*, volume 145 of *Wiley Series in Probability and Statistics* (Hoboken, NJ, USA: John Wiley & Sons, Inc.), ISBN 9780470316481, URL http://dx.doi.org/10.1002/9780470316481.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming*. ISBN 9780898716870, URL http://dx.doi.org/10.1137/1.9780898718751.

Steiger NM, Wilson JR (2002) An improved batch means procedure for simulation output analysis. *Management Science* 48(12):1569–1586, ISSN 00251909, URL http://dx.doi.org/10.1287/mnsc.48.12.1569.438.

Steimle LN, Kaufman DL, Denton BT (2019) Multi-model Markov Decision Processes 1–64.

Suen Sc, Brandeau ML, Goldhaber-Fiebert JD (2018) Optimal timing of drug sensitivity testing for patients on first-line tuberculosis treatment. *Health Care Management Science* 21(4):632–646, ISSN 13869620, URL http://dx.doi.org/10.1007/s10729-017-9416-4.

Sundström J, Arima H, Woodward M, Jackson R, Karmali K, Lloyd-Jones D, Baigent C, et al. (2014) Blood pressure-lowering treatment based on cardiovascular risk: A meta-analysis of individual patient data. *The Lancet* 384(9943):591–598, ISSN 1474547X, URL http://dx.doi.org/10.1016/S0140-6736(14)61212-5.

Sussman J, Vijan S, Hayward R (2013) Using Benefit-Based Tailored Treatment to Improve the Use of Antihypertensive Medications. *Circulation* 128(21):2309–2317, ISSN 00097322, URL http://dx.doi.org/10.1161/CIRCULATIONAHA.113.002290.

Sutton RS, Barto AG (2018) *Reinforcement Learning: An Introduction* (MIT press), 2 edition, ISBN 9780262039246, URL http://dx.doi.org/10.1016/S1364-6613(99)01331-5.

Tu W, Zhou XH (2000) Pairwise comparisons of the means of skewed data. *Journal of Statistical Planning and Inference* 88(1):59–74, ISSN 03783758, URL http://dx.doi.org/10.1016/S0378-3758(99)00206-2.

Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, et al. (2020) *Heart disease and stroke statistics—2020 update: A report from the American Heart Association.* ISBN 0000000000000, URL http://dx.doi.org/10.1161/CIR.0000000000000757.

Westfall PH (2011) On Using the Bootstrap for Multiple Comparisons. *Journal of Biopharmaceutical Statistics* 21(6):1187–1205, ISSN 1054-3406, URL http://dx.doi.org/10.1080/10543406.2011.607751.

Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, DePalma SM, et al. (2018) 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. *Journal of the American College of Cardiology* 71(19):e127–e248, ISSN 07351097, URL http://dx.doi.org/10.1016/j.jacc.2017.11.006.

Yadlowsky S, Hayward RA, Sussman JB, McClelland RL, Min YI, Basu S (2018) Clinical implications of revised pooled Cohort equations for estimating atherosclerotic cardiovascular disease risk. *Annals of Internal Medicine* 169(1):20–29, ISSN 15393704, URL http://dx.doi.org/10.7326/M17-3011.

Zargoush M, Gümüş M, Verter V, Daskalopoulou SS (2018) Designing risk-adjusted therapy for patients with hypertension. *Production and Operations Management* 27(12):2291–2312, ISSN 19375956, URL http://dx.doi.org/10.1111/poms.12872.

Appendix A: Proofs of Analytical Results

This Appendix contains the proof of all our claims. For ease of reading, we have repeated the claims. In addition, we have separated claims with multiple parts with lower case Roman numerals. Since the output of our algorithms can be obtained through independent simulations or by dividing a single simulation into batches, we split the simulation output into M batches (or independent simulations) of K observations, for a total of $N = MK$ samples (see Section 5.2). We present our results in terms of M and K, unless otherwise noted.

A.1. Proofs of Section 5.1.1

PROPOSITION 1. *Under Assumption 1, it follows that* $1 - \mathbb{P}(\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)) \leq A \exp\left\{\frac{-N}{2\kappa_t^2}\right\}$, *with* $\kappa_t \coloneqq \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$, *where* $R_t \coloneqq \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} R_t(s,a)$.

The proof of this proposition depends on the following lemma.

LEMMA 1. *Let* $\hat{\theta}_t(s,a,a') \coloneqq \hat{Q}_t(s,a) - \hat{Q}_t(s,a')$ *for* $a, a' \in \mathcal{A}$. *Under Assumption 1, (i)* $\hat{Q}_t(s,a) \leq \kappa_t$ *and (ii)* $|\hat{\theta}_t(s,a,a')| \leq \kappa_t$.

Proof.

(i) We first show that $0 \leq \hat{Q}_t(s,a) \leq \kappa_t$. Because the rewards are non-negative by Assumption 1 it follows that $\hat{Q}_t(s,a) \geq 0$. The proof proceeds by backwards induction on t with $t = T$ as the base case. Since for every $\omega^{m,k} \in \Omega$ there is always a positive probability of observing an outcome $\tilde{\omega}^{m,k} \in \Omega$ such that $r_T(s,\omega^{m,k}) \leq r_T(s,\tilde{\omega}^{m,k}) \leq R_T(s)$, it holds that $\hat{Q}_T(s,a) \leq \gamma^0 R_T(s) \leq \gamma^0 R_T$. This shows that the claim is true for $t = T$. Suppose $\hat{Q}_{t+1}(s,a) \leq \sum_{\tau=t+1}^{T} \gamma^{\tau-(t+1)} R_\tau$ as the induction hypothesis. By the same argument used to show the claim in the induction base, we get that $\hat{Q}_t(s,a) \leq R_t(s,a) + \frac{\gamma}{MK} \sum_{m=1}^{M} \sum_{k=1}^{K} \hat{v}_{t+1}(s^{m,k})$. Moreover, from the induction hypothesis it holds that $\hat{v}_{t+1}(s^{m,k}) \leq \sum_{\tau=t+1}^{T} \gamma^{\tau-(t+1)} R_\tau$. Therefore, $\hat{Q}_t(s,a) \leq R_t(s,a) + \gamma \sum_{\tau=t+1}^{T} \gamma^{\tau-(t+1)} R_\tau \leq \gamma^0 R_t + \sum_{\tau=t+1}^{T} \gamma^{\tau-t} R_\tau = \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$, completing the inductive step and the proof.

(ii) The claim that $|\hat{Q}_t(s,a) - \hat{Q}_t(s,a')| \leq \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$ follows immediately since $0 \leq \hat{Q}_t(s,a)$ by Assumption 1 and because $\hat{Q}_t(s,a) \leq \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$ by the above analysis. □

Proof of Proposition 1. For every s, we have that:

$$\{\hat{\mathcal{A}}_t^*(s) \not\subseteq \mathcal{A}_t^*(s)\} = \bigcup_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \bigcap_{a'\in\mathcal{A}} \{\hat{Q}_t(s,a) \geq \hat{Q}_t(s,a')\},$$

which implies:

$$\mathbb{P}(\hat{\mathcal{A}}_t^*(s) \not\subseteq \mathcal{A}_t^*(s)) \leq \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \mathbb{P}\left(\bigcap_{a'\in\mathcal{A}} \hat{Q}_t(s,a) \geq \hat{Q}_t(s,a')\right) \leq \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \mathbb{P}\left(\hat{Q}_t(s,a) \geq \hat{Q}_t(s,a')\right),$$

for any $a' \in \mathcal{A}$. Further, consider a mapping $g : \mathcal{A} \setminus \mathcal{A}_t^*(s) \mapsto \mathcal{A}$ such that $\mathbb{E}[\hat{Q}_t(s,g(a))] \geq \mathbb{E}[\hat{Q}_t(s,a)]$. Note that this mapping always exists because if $a \in \arg\max_{\bar{a}\in\mathcal{A}} \mathbb{E}[\hat{Q}_t(s,\bar{a})]$ then we have that $g(a) = a$ so that $\mathbb{E}[\hat{Q}_t(s,g(a))] = \mathbb{E}[\hat{Q}_t(s,a)]$. If $a \notin \arg\max_{\bar{a}\in\mathcal{A}} \mathbb{E}[\hat{Q}_t(s,\bar{a})]$ then we can find a $g(a)$ such that $\mathbb{E}[\hat{Q}_t(s,g(a))] >$

$\mathbb{E}[\hat{Q}_t(s,a)]$. For each $a \in \mathcal{A} \setminus \mathcal{A}_t^*(s)$, define $\vartheta_t^{m,k}(s,a) \coloneqq Q_t^{m,k}(s,a) - Q_t^{m,k}(s,g(a))$, and its corresponding average:

$$\hat{\vartheta}_t(s,a) \coloneqq \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\vartheta_t^{m,k}(s,a) = \hat{Q}_t(s,a) - \hat{Q}_t(s,g(a)).$$

Let $l'(s) \coloneqq \min_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \hat{\vartheta}_t(s,a)$ and $u'(s) \coloneqq \max_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \hat{\vartheta}_t(s,a)$. Note that $\mathbb{E}[\vartheta_t^{m,k}(s,a)] \leq 0$ for all a. Also, notice that $l'(s)$ and $u'(s)$ can be attained because $\mathcal{S}$ and $\mathcal{A}$ are finite and the rewards are bounded for all s, a, and t. Hence, it follows that:

$$\mathbb{P}(\hat{\mathcal{A}}_t^*(s) \not\subseteq \mathcal{A}_t^*(s)) \leq \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \mathbb{P}\left(\hat{\vartheta}_t(s,a) \geq 0\right) = \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \mathbb{P}\left(e^{\tau\sum_m\sum_k \vartheta_t^{m,k}(s,a)} \geq 1\right),$$

for $\tau \geq 0$. By Markov's inequality it holds that:

$$1 - \mathbb{P}(\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)) \leq \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \mathbb{E}\left[e^{\tau\sum_m\sum_k \vartheta_t^{m,k}(s,a)}\right].$$

Since $|u'(s) - l'(s)| \leq \kappa_t$, by Hoeffding's Lemma we get:

$$1 - \mathbb{P}(\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)) \leq \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \exp\left\{\tau\sum_m\sum_k \mathbb{E}[\vartheta_t^{m,k}(s,a)] + \frac{1}{2}MK\tau^2\kappa_t^2\right\}.$$

Minimizing over $\tau \geq 0$ it follows that:

$$\begin{aligned} 1 - \mathbb{P}(\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)) &\leq \min_{\tau\geq 0} \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \exp\left\{\tau\sum_m\sum_k \mathbb{E}[\vartheta_t^{m,k}(s,a)] + \frac{1}{2}MK\tau^2\kappa_t^2\right\} \\ &= \sum_{a\in\mathcal{A}\setminus\mathcal{A}_t^*(s)} \exp\left\{\frac{-MK\left(\frac{1}{MK}\sum_m\sum_k \mathbb{E}[\vartheta_t^{m,k}(s,a)]\right)^2}{2\kappa_t^2}\right\} \leq A\exp\left\{\frac{-N}{2\kappa_t^2}\right\}, \end{aligned}$$

where the second inequality follows because $\left(\frac{1}{MK}\sum_m\sum_k \mathbb{E}[\vartheta_t^{m,k}(s,a)]\right)^2 > 0$ and $N = MK$. □

PROPOSITION 2 *Suppose Assumption 1 holds. Then, for any $\beta \in (0,1)$ and a fixed sample size N satisfying $N \geq 2\kappa_t^2\log(A/\beta)$ it holds that $\mathbb{P}(\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)) \geq 1-\beta$.*

Proof. This results follows by setting the right-hand side of the inequality in Proposition 1 to less or equal than β. Solving for N, we get that $N \geq 2\kappa_t^2\log(A/\beta)$. □

A.2. Proofs of Section 5.1.2

THEOREM 1. *Suppose Assumption 1 holds. Then, $\hat{Q}_t(s,a)$ converges to $Q_t(s,a)$ with probability 1 uniformly on $\mathcal{A}$.*

The proof of this theorem depends on the following lemmas.

LEMMA 2. *Let $(V_n)_{n\in\mathbb{N}}$ and $(W_n)_{n\in\mathbb{N}}$ be sequences of random variables, and let V and W be two other random variables. Suppose that V_n converges a.s. to V and that W_n converges a.s. to W. Then, (i) $V_n + W_n \xrightarrow{a.s.} V + W$ and (ii) $V_nW_n \xrightarrow{a.s.} VW$.*

Proof.

(i) We first show that $V_n + W_n \overset{a.s.}{\to} V + W$. By the triangle inequality, for any $\delta > 0$ we have that:

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty} |V_n + W_n - (V+W)| \geq \delta\right) &\leq \mathbb{P}\left(\left\{\lim_{n\to\infty} |V_n - V| \geq \frac{\delta}{2}\right\} \bigcup \left\{\lim_{n\to\infty} |W_n - W| \geq \frac{\delta}{2}\right\}\right) \\ &\leq \mathbb{P}\left(\lim_{n\to\infty} |V_n - V| \geq \frac{\delta}{2}\right) + \mathbb{P}\left(\lim_{n\to\infty} |W_n - W| \geq \frac{\delta}{2}\right) = 0,\end{aligned}$$

where the second inequality follows from the addition rule of probability and the equality follows because $V_n \overset{a.s.}{\to} V$ and $W_n \overset{a.s.}{\to} W$. This shows that $V_n + W_n \overset{a.s.}{\to} V + W$.

(ii) We now show that $V_n W_n \overset{a.s.}{\to} VW$. Since $V_n W_n - VW = (V_n - V)(W_n - W) + W(V_n - V) + V(W_n - W)$, for any $\delta > 0$ we get:

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty} |V_n W_n - VW| \geq \delta\right) &\leq \mathbb{P}\left(\lim_{n\to\infty} |(V_n - V)(W_n - W)| \geq \frac{\delta}{2}\right) \\ &\quad + \mathbb{P}\left(\lim_{n\to\infty} |W(V_n - V) + V(W_n - W)| \geq \frac{\delta}{2}\right), \qquad (3)\end{aligned}$$

where the inequality follows from the triangle inequality and the addition rule of probability. We first focus on the first term in the right-hand side of equation (3). As $|vw| \leq \frac{1}{2}v^2 + \frac{1}{2}w^2$ for all $v, w \in \mathbb{R}$ it holds that:

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty} |(V_n - V)(W_n - W)| \geq \frac{\delta}{2}\right) &\leq \mathbb{P}\left(\lim_{n\to\infty} \frac{1}{2}(V_n - V)^2 \geq \frac{\delta}{4}\right) + \mathbb{P}\left(\lim_{n\to\infty} \frac{1}{2}(W_n - W)^2 \geq \frac{\delta}{4}\right) \\ &= \mathbb{P}\left(\lim_{n\to\infty} |V_n - V| \geq \sqrt{\frac{\delta}{2}}\right) + \mathbb{P}\left(\lim_{n\to\infty} |W_n - W| \geq \sqrt{\frac{\delta}{2}}\right) = 0,\end{aligned}$$

where the last equality follows because $V_n \overset{a.s.}{\to} V$ and $W_n \overset{a.s.}{\to} W$. We now show that the second term in the right-hand side of equation (3) is also 0. By the triangle inequality, we have that:

$$\mathbb{P}\left(\lim_{n\to\infty} |W(V_n - V) + V(W_n - W)| \geq \frac{\delta}{2}\right) \leq \mathbb{P}\left(\lim_{n\to\infty} |W(V_n - V)| \geq \frac{\delta}{4}\right) + \mathbb{P}\left(\lim_{n\to\infty} |V(W_n - W)| \geq \frac{\delta}{4}\right).$$

For $\mathbb{P}\left(\lim_{n\to\infty} |W(V_n - V)| \geq \frac{\delta}{4}\right)$ it holds that:

$$\begin{aligned}\mathbb{P}\left(\lim_{n\to\infty} |W(V_n - V)| \geq \frac{\delta}{4}\right) &= \mathbb{P}\left(\left\{\lim_{n\to\infty} |W(V_n - V)| \geq \frac{\delta}{4}\right\} \bigcap \{|W| \leq \eta\}\right) \\ &\quad + \mathbb{P}\left(\left\{\lim_{n\to\infty} |W(V_n - V)| \geq \frac{\delta}{4}\right\} \bigcap \{|W| > \eta\}\right) \\ &\leq \mathbb{P}\left(\lim_{n\to\infty} |V_n - V| \geq \frac{\delta}{4\eta}\right) + \mathbb{P}\left(|W| > \eta\right),\end{aligned}$$

for any $\eta \geq 1$. Allowing $\eta \to \infty$ as $n \to \infty$, we get that $\mathbb{P}\left(\lim_{n\to\infty} |W(V_n - V)| \geq \frac{\delta}{4}\right) = 0$ because $V_n \overset{a.s.}{\to} V$ and $\mathbb{P}\left(|W| > \eta\right) = 0$. Repeating the above arguments for $\mathbb{P}\left(\lim_{n\to\infty} |V(W_n - W)| \geq \frac{\delta}{4}\right)$, it holds that:

$$\mathbb{P}\left(\lim_{n\to\infty} |W(V_n - V) + V(W_n - W)| \geq \frac{\delta}{2}\right) = 0.$$

Combining our results, we can conclude that $V_n W_n \overset{a.s.}{\to} VW$. □

LEMMA 3. *Let $\bar{Q}_t(s,a)$ denote the estimate of $Q_t(s,a)$ generated with a single simulation (or batch) of Algorithm 1. Then, $\bar{Q}_t(s,a)$ converges to $Q_t(s,a)$ with probability 1 uniformly on $\mathcal{A}$.*

Proof. Suppose the immediate rewards $r_t(s,a,\omega)$ are bounded iid random variables from an unknown probability distribution at every iteration k and replication m. If the immediate rewards are constant their convergence holds trivially. For simplicity of notation, we use s^k instead of $s^{m,k}$, as we are interested in a single replication of the simulation during this proof ($M=1$).

We prove this result by backwards induction on t, starting with $t=T$ as the base case. Since $Q_T^k(s,a)$ are iid for all $k=1,\dots,K$ with finite mean and variance we have that $\bar{Q}_T(s,a) \xrightarrow{a.s.} Q_T(s,a)$ by the strong law of large numbers (SLLN). Furthermore, since $\mathcal{A}$ is a finite set we have that $\sup_{a\in\mathcal{A}} |\bar{Q}_T(s,a) - Q_T(s,a)| = \max_{a\in\mathcal{A}} |\bar{Q}_T(s,a) - Q_T(s,a)| \leq \sum_{a\in\mathcal{A}} |\bar{Q}_T(s,a) - Q_T(s,a)|$. Hence, $\bar{Q}_T(s,a) \xrightarrow{a.s.} Q_T(s,a)$ implies almost sure convergence uniformly on $\mathcal{A}$. This shows that the claim is true for $t=T$. Suppose that $\bar{Q}_{t+1}(s',a') \xrightarrow{a.s.} Q_{t+1}(s',a')$ uniformly on $\mathcal{A}$ for all s' as the induction hypothesis. For $\bar{Q}_t(s,a)$ we have that:

$$\bar{Q}_t(s,a) = \frac{1}{K}\sum_k Q_t^k(s,a) = \bar{r}_t(s,a) + \frac{\gamma}{K}\sum_k \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s^k,a'),$$

where $\bar{r}_t(s,a) \coloneqq \frac{1}{K}\sum_k r_t(s,a,\omega^k)$. Since $r_t(s,a,\omega^k)$ are bounded iid random variables, we have that $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is well-defined. Consequently, we can deduce that $\bar{r}_t(s,a) \xrightarrow{a.s.} \mathbb{E}[r_t(s,a,\omega)|s,a]$ by the SLLN. Because $\mathcal{S}$ is a finite set and $|r_t(s,a,\omega^k)| < \infty$, we have that $\mathbb{E}[\max_{a'\in\mathcal{A}} Q_{t+1}(s',a')|s,a]$ is also well-defined. Now, note that:

$$\frac{1}{K}\sum_k \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s^k,a') = \frac{1}{K}\sum_k\sum_{s'} \mathbb{1}\{s^k=s'|s,a\} \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s',a').$$

Thus, for a fixed state s' we have that:

$$\frac{1}{K}\sum_k \mathbb{1}\{s^k=s'|s,a\} \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s',a') = \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s',a') \frac{1}{K}\sum_k \mathbb{1}\{s^k=s'|s,a\}.$$

From the induction hypothesis, we can deduce that $\max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s',a') \xrightarrow{a.s.} \max_{a'\in\mathcal{A}} Q_{t+1}(s',a')$ for each s'. Furthermore, by the SLLN we get that $K^{-1}\sum_k \mathbb{1}\{s^k=s'|s,a\} \xrightarrow{a.s.} p_t(s'|s,a)$ for every s'. Hence, by Lemma 2 it follows that $K^{-1}\sum_k \mathbb{1}\{s^k=s'|s,a\} \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s',a') \xrightarrow{a.s.} p_t(s'|s,a) \max_{a'\in\mathcal{A}} Q_{t+1}(s',a')$. Adding over all states, we conclude that:

$$\frac{1}{K}\sum_k\sum_{s'} \mathbb{1}\{s^k=s'|s,a\} \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s',a') \xrightarrow{a.s.} \sum_{s'} p_t(s'|s,a) \max_{a'\in\mathcal{A}} Q_{t+1}(s',a') = \mathbb{E}\Big[\max_{a'\in\mathcal{A}} Q_{t+1}(s',a')\Big|s,a\Big].$$

As $\bar{r}_t(s,a) \xrightarrow{a.s.} \mathbb{E}[r_t(s,a,\omega)|s,a]$ and $\gamma K^{-1}\sum_k \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s^k,a') \xrightarrow{a.s.} \gamma\mathbb{E}[\max_{a'\in\mathcal{A}} Q_{t+1}(s',a')|s,a]$, by Lemma 2 it holds that:

$$\bar{r}_t(s,a) + \gamma K^{-1}\sum_k \max_{a'\in\mathcal{A}} \bar{Q}_{t+1}(s^k,a') \xrightarrow{a.s.} \mathbb{E}[r_t(s,a,\omega) + \gamma \max_{a'\in\mathcal{A}} Q_{t+1}(s',a')|s,a].$$

Therefore, we get that $\bar{Q}_t(s,a) \xrightarrow{a.s.} Q_t(s,a)$ for all a. Because $\max_{a\in\mathcal{A}} |\bar{Q}_t(s,a) - Q_t(s,a)| \leq \sum_{a\in\mathcal{A}} |\bar{Q}_t(s,a) - Q_t(s,a)|$, the a.s. convergence of $\bar{Q}_t(s,a)$ for all a implies that $\bar{Q}_t(s,a) \xrightarrow{a.s.} Q_t(s,a)$ uniformly on $\mathcal{A}$. This completes the inductive step and the proof. □

Proof of Theorem 1. From Lemma 3, we have that $\bar{Q}_t^m(s,a) \xrightarrow{a.s.} Q_t(s,a)$ uniformly on $\mathcal{A}$ for each s and m. Since $\hat{Q}_t(s,a) = \frac{1}{M}\sum_{m=1}^M \bar{Q}_t^m(s,a)$, it holds that $\hat{Q}_t(s,a) \xrightarrow{a.s.} Q_t(s,a)$ by Lemma 2. Since $\max_{a\in\mathcal{A}} |\hat{Q}_t(s,a) - Q_t(s,a)| \leq \sum_{a\in\mathcal{A}} |\hat{Q}_t(s,a) - Q_t(s,a)|$, the a.s. convergence of $\hat{Q}_t(s,a)$ to $Q_t(s,a)$ for all a implies almost sure convergence uniformly on $\mathcal{A}$. □

COROLLARY. 1 *Suppose Assumption 1 holds. Then, (i) $\hat{v}_t(s)$ converges to $v_t(s)$ and (ii) $\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)$ with probability 1 for N large enough.*

Proof.

(i) From Theorem 1, we have that $\max_{a\in\mathcal{A}} |\hat{Q}_t(s,a) - Q_t(s,a)| \overset{a.s.}{\to} 0$ as $N \to \infty$. Thus, it holds that $\hat{v}_t(s) \overset{a.s.}{\to} v_t(s)$ for all s and t.

(ii) The proof is a contrapositive argument. Let $\max_{a\in\mathcal{A}} |\hat{Q}_t(s,a) - Q_t(s,a)| \leq \delta_N$ for $\delta_N > 0$ and $\varrho(s) \coloneqq v_t(s) - \max_{a\in\mathcal{A}_t^-(s)} Q_t(s,a)$. Since for any $a \in \mathcal{A}_t^-(s)$ we have that $v_t(s) > Q_t(s,a)$ and $\mathcal{A}_t^-(s)$ is finite it holds that $\varrho(s) > 0$. Let N be large enough such that $\varrho(s)/2 > \delta_N$. Combining these observations we get that $\hat{v}_t(s) > v_t(s) - \varrho(s)/2 > \hat{Q}_t(s,a)$ and $a \notin \mathcal{A}_t^*(s)$ implies $a \notin \hat{\mathcal{A}}_t^*(s)$. The inclusion $\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)$ follows. □

A.3. Proofs of Section 5.1.3

PROPOSITION 3 *Under Assumptions 1 and 2, (i) $\hat{Q}_t(s,a)$ and (ii) $\hat{v}_t(s)$ are ϵ-nonincreasing in s with probability 1 for N large enough.*

The proof of this proposition relies on the following lemma.

LEMMA 4. *Let X and Y be partially ordered finite sets, $g : X \mapsto \mathbb{R}$ be a monotonic function of X, and $h : X \times Y \mapsto [0,1]$ be a function of Y satisfying $\sum_{x'\geq x} h(x',y) \leq \sum_{x'\geq x} h(x',\bar{y})$, for $y \leq \bar{y}$ with $\sum_{x\in X} h(x,y) = \sum_{x\in X} h(x,\bar{y})$. Then, we have $\sum_{x\in X} h(x,y)g(x) \leq \sum_{x\in X} h(x,\bar{y})g(x)$, when g is nondecreasing in x and $\sum_{x\in X} h(x,y)g(x) \geq \sum_{x\in X} h(x,\bar{y})g(x)$, when g is nonincreasing in x.*

Proof of Lemma 4. We first prove that the claim holds in the nondecreasing case. By definition, we have that $\sum_{x\in X} h(x,y) = \sum_{x\in X} h(x,\bar{y})$. Thus,

$$
\begin{aligned}
0 &= \sum_{x\in X} h(x,y) - \sum_{x\in X} h(x,\bar{y}) \qquad (4)\\
&= \left[\sum_{x\in X} h(x,y) - \sum_{x\in X} h(x,\bar{y})\right] g(x^1)\\
&\geq h(x^1,y)g(x^1) - h(x^1,\bar{y})g(x^1) + \left[\sum_{x\in X\setminus\{x^1\}} h(x,y) - h(x,\bar{y})\right] g(x^2)\\
&\geq \sum_{x'\in\{x^1,x^2\}} h(x',y)g(x') - h(x',\bar{y})g(x') + \left[\sum_{x\in X\setminus\{x^1,x^2\}} h(x,y) - h(x,\bar{y})\right] g(x^3),
\end{aligned}
$$

where $x^1 \leq x^2 \leq x^3 \in X$. The inequalities above follow because $\sum_{x'\geq x} h(x',y) \leq \sum_{x'\geq x} h(x',\bar{y})$ and g is nondecreasing in x. Continuing with this pattern we get that $0 \geq \sum_{x'\in X} h(x',y)g(x') - h(x',\bar{y})g(x')$, which implies that $\sum_{x'\in X} h(x',\bar{y})g(x') \geq \sum_{x'\in X} h(x',y)g(x')$, completing the proof for the nondecreasing case. The proof for the nonincreasing case follows by multiplying (4) by -1 and noticing that the direction of the inequalities change to "$\leq$". □

Proof of Proposition 3.

(i) We first show that $\hat{v}_t(s)$ is ϵ-nonincreasing in s for N large enough. By Proposition 4.7.4 in Puterman (2014), we have that Assumption 2 implies that $v_t(s)$ is nonincreasing in s. Moreover, from Corollary 1 we have that $\hat{v}_t(s) \overset{a.s.}{\to} v_t(s)$ for all s. Thus, for any $\epsilon > 0$ there is an $N^* \in \mathbb{N}$ such that for any $N \geq N^*$ it holds that $\hat{v}_t(\tilde{s}) \leq \hat{v}_t(s) + \epsilon$ for $s \leq \tilde{s}$. It follows that $\hat{v}_t(s)$ is ϵ-nonincreasing in s with probability 1 for N large enough.

(ii) We now show that if $Q_t(s,a)$ is nonincreasing in s, then $\hat{Q}_t(s,a)$ is ϵ-nonincreasing with probability 1 for N large enough. By Assumption 2 we have that $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is nonincreasing in s. Since $v_t(s)$ is nonincreasing in s and $\bar{p}_t(s'|s,a)$ is nondecreasing in s by Assumption 2, applying Lemma 4 with $\tilde{s} \geq s$, $h(x,y) = p_t(s'|s,a)$, $h(x,\bar{y}) = p_t(s'|\tilde{s},a)$, and $g(x) = v_{t+1}(s')$, we get that $\sum_{s' \in \mathcal{S}} p_t(s'|s,a) v_{t+1}(s')$ is nonincreasing in s. Thus, it follows that $Q_t(s,a)$ is nonincreasing in s. Since from Theorem 1 we have that $\hat{Q}_t(s,a) \overset{a.s.}{\to} Q_t(s,a)$, it holds that $\hat{Q}_t(s,a)$ is ϵ-nonincreasing in s with probability 1 for N large enough. □

PROPOSITION 4. *Suppose Assumptions 1 and 2 hold. Then, (i) $\hat{Q}_t(s,a)$ and (ii) $\hat{v}_t(s)$ are ϵ-nonincreasing in t with probability 1 for N large enough.*

Proof. We first show these results when $\mathbb{E}[r_t(s,a,\omega)|s,a]$ and $p_t(s'|s,a)$ are known.

(i) The proof for $v_t(s)$ proceeds by backwards induction on t, starting with $t = T-1$ as the base case. Since $\mathbb{E}[r_{T-1}(s,a,\omega)|s,a] \geq \mathbb{E}[r_T(s,\omega)|s]$ by Assumption 2, we have that $v_{T-1}(s) \geq v_T(s)$. Suppose that $v_{t+1}(s) \geq v_{t+2}(s)$ as the induction hypothesis. Since $v_t(s)$ is nonincreasing in s by Proposition 4.7.4 in Puterman (2014) and $\bar{p}_t(s'|s,a)$ is nondecreasing in t by Assumption 2, applying Lemma 4 with $h(x,y) = p_t(s'|s,a)$, $h(x,\bar{y}) = p_{t+1}(s'|s,a)$, and $g(x) = v_{t+1}(s')$, we get that $\sum_{s' \in \mathcal{S}} p_t(s'|s,a) v_{t+1}(s') \geq \sum_{s' \in \mathcal{S}} p_{t+1}(s'|s,a) v_{t+1}(s')$. As $v_{t+1}(s) \geq v_{t+2}(s)$ from the induction hypothesis, it follows that $\sum_{s' \in \mathcal{S}} p_t(s'|s,a) v_{t+1}(s') \geq \sum_{s' \in \mathcal{S}} p_{t+1}(s'|s,a) v_{t+2}(s')$. Because $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is nonincreasing in t due to Assumption 2, we get that $v_t(s)$ is also nonincreasing in t.

(ii) The proof for $Q_t(s,a)$ follows as a consequence of the nonincreasing behavior of $v_t(s)$ in t. Since $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is nonincreasing t and $\bar{p}_t(s'|s,a)$ is nondecreasing in t by Assumption 2, and $v_t(s)$ is nonincreasing in s and t by (i), it holds that $Q_t(s,a) \geq Q_{t+1}(s,a)$ for any $t \in \mathcal{T} \setminus \{T\}$.

From Theorem 1 we have that $\hat{Q}_t(s,a) \overset{a.s.}{\to} Q_t(s,a)$ and from Corollary 1 we get that $\hat{v}_t(s) \overset{a.s.}{\to} v_t(s)$. This implies that for any $\epsilon > 0$ we can find an $N^* \in \mathbb{N}$ such that $\hat{Q}_t(s,a) \geq \hat{Q}_{t+1}(s,a) - \epsilon$ and $\hat{v}_t(s) \geq \hat{v}_{t+1}(s) - \epsilon$ for any $N \geq N^*$. Hence, if $Q_t(s,a)$ and $v_t(s)$ are nonincreasing in t, then $\hat{Q}_t(s,a)$ and $\hat{v}_t(s)$ are ϵ-nonincreasing with probability 1 for N large enough. □

A.4. Proofs of Section 5.2.1

PROPOSITION 5. *Suppose Assumption 1 holds. Then, for any $\alpha \in (0,1)$, $a^* \in \mathcal{C}_t(s)$, and N large enough we have that $\mathbb{P}\Big(Q_t(s,a^*) - Q_t(s,a) \in \Theta : \hat{\mathbb{H}}_t\left(\max_{a \in \mathcal{A}}\{\hat{\psi}_t(s,a)\}, \hat{\mathbb{F}}_t(s)\right) \leq 1-\alpha\Big) = 1-\alpha$.*

The proof of this proposition depends on the following two lemmas.

LEMMA 5. *Under Assumption 1, (i) $\bar{\sigma}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$, (ii) $\hat{\sigma}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$, and (iii) $K\hat{\zeta}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$.*

Proof.

(i) For simplicity of notation we drop the superscript m in $Q_t^{m,k}(s,a)$ as the batch number m is arbitrary, but fixed. Let $\bar{\nu}_t^2(s,a) \coloneqq \frac{K-1}{K}\bar{\sigma}_t^2(s,a)$. For $\bar{\nu}_t^2(s,a)$ we have:

$$\bar{\nu}_t^2(s,a) = \frac{1}{K}\sum_{k=1}^{K}\left[Q_t^k(s,a) - \bar{Q}_t(s,a)\right]^2 = \frac{1}{K}\sum_{k=1}^{K}\left[Q_t^k(s,a)\right]^2 - \left[\bar{Q}_t(s,a)\right]^2.$$

From Theorem 1 we have that $\bar{Q}_t(s,a) \xrightarrow{a.s.} Q_t(s,a)$ and by Lemma 2 it follows that $\left[\bar{Q}_t(s,a)\right]^2 \xrightarrow{a.s.} Q_t^2(s,a)$. It remains to show that $\frac{1}{K}\sum_{k=1}^{K}\left[Q_t^k(s,a)\right]^2 \xrightarrow{a.s.} \mathbb{E}[(r_t(s,a,\omega)+\gamma v_{t+1}(f_{t+1}(s,a,\omega)))^2\,|s,a]$. For $\frac{1}{K}\sum_{k=1}^{K}\left[Q_t^k(s,a)\right]^2$ it holds that:

$$\frac{1}{K}\sum_{k=1}^{K}\left[Q_t^k(s,a)\right]^2 = \frac{1}{K}\sum_{k=1}^{K} r_t^2(s,a,\omega^k) + \gamma r_t(s,a,\omega^k)\hat{v}_{t+1}(s^k) + \gamma^2\hat{v}_{t+1}^2(s^k),$$

where $s^k = f_{t+1}(s,a,\omega^k)$. Since $r_t(s,a,\omega^k)$ are iid, so are $r_t^2(s,a,\omega^k)$ and we get that $\frac{1}{K}\sum_{k=1}^{K} r_t^2(s,a,\omega^k) \xrightarrow{a.s.} \mathbb{E}[r_t^2(s,a,\omega)|s,a]$ by the SLLN. Further, from the SLLN and Lemma 2 it follows that $K^{-1}\sum_{k=1}^{K} r_t(s,a,\omega^k)\hat{v}_{t+1}(s^k) \xrightarrow{a.s.} \mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a]$ and that $K^{-1}\sum_{k=1}^{K}\sum_{s'}\hat{v}_{t+1}^2(s') \xrightarrow{a.s.} \mathbb{E}[v_{t+1}^2(s')|s,a]$. Therefore,

$$\frac{1}{K}\sum_{k=1}^{K}\left[Q_t^k(s,a)\right]^2 \xrightarrow{a.s.} \mathbb{E}[(r_t(s,a,\omega)+\gamma v_{t+1}(f_{t+1}(s,a,\omega)))^2|s,a].$$

and we have that:

$$\bar{\nu}_t^2(s,a) \xrightarrow{a.s.} \mathbb{E}[(r_t(s,a,\omega)+\gamma v_{t+1}(f_{t+1}(s,a,\omega)))^2|s,a] - \mathbb{E}[r_t(s,a,\omega)+\gamma v_{t+1}(f_{t+1}(s,a,\omega))|s,a]^2.$$

As $\bar{\sigma}_t^2(s,a) = \frac{K}{K-1}\bar{\nu}_t^2(s,a)$ and $\frac{K}{K-1} \to 1$ as $K \to \infty$, we can conclude that $\bar{\sigma}_t^2(s,a) \xrightarrow{a.s.} \sigma_t^2(s,a)$.

(ii) This result is a direct consequence of (i) as $\bar{\sigma}_t^2(s,a)$ is equivalent to $\hat{\sigma}_t^2(s,a)$ if we replace K by N.

(iii) Let $\hat{\nu}_t^2(s,a) \coloneqq \frac{M-1}{M}\hat{\zeta}_t^2(s,a)$. For $\hat{\nu}_t^2(s,a)$ we have:

$$\hat{\nu}_t^2(s,a) = \frac{1}{M}\sum_{m=1}^{M}\left[\bar{Q}_t^m(s,a) - \hat{Q}_t(s,a)\right]^2 = \frac{1}{M}\sum_{m=1}^{M}\left[\bar{Q}_t^m(s,a)\right]^2 - \left[\hat{Q}_t(s,a)\right]^2.$$

From Theorem 1 it holds that $\hat{Q}_t(s,a) \xrightarrow{a.s.} Q_t(s,a)$ and by Lemma 2 it follows that $\hat{Q}_t^2(s,a) \xrightarrow{a.s.} Q_t^2(s,a)$. For $\frac{1}{M}\sum_{m=1}^{M}\left[\bar{Q}_t^m(s,a)\right]^2$ we have:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M}\left[\bar{Q}_t^m(s,a)\right]^2 &= \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{K}\sum_{k=1}^{K} r_t(s,a,\omega^{m,k}) + \gamma\hat{v}_{t+1}(s^{m,k})\right]^2 \\
&= \frac{1}{MK^2}\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l\neq k} r_t(s,a,\omega^{m,k})r_t(s,a,\omega^{m,l}) + 2\gamma r_t(s,a,\omega^{m,k})\hat{v}_{t+1}(s^{m,l}) \\
&\quad + \gamma^2\hat{v}_{t+1}(s^{m,k})\hat{v}_{t+1}(s^{m,l}) \\
&\quad + \frac{1}{MK^2}\sum_{m=1}^{M}\sum_{k=1}^{K} r_t^2(s,a,\omega^{m,k}) + 2\gamma r_t(s,a,\omega^{m,k})\hat{v}_{t+1}(s^{m,k}) + \gamma^2\hat{v}_{t+1}^2(s^{m,k}),
\end{aligned}$$

where $s^{m,k} = f_{t+1}(s,a,\omega^{m,k})$. Moreover, since k is independent of l for every $k \neq l$ it holds that:

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M}\left[\bar{Q}_t^m(s,a)\right]^2 &= \frac{1}{MK^2}\sum_{m=1}^{M} K(K-1)\bar{r}_t^m(s,a)\bar{r}_t^m(s,a) + 2\gamma K(K-1)\bar{r}_t^m(s,a)\frac{1}{K-1}\sum_{l=1}^{K}\hat{v}_{t+1}(s^{m,l}) \\
&\quad + \gamma^2 K(K-1)\frac{1}{K}\sum_{k=1}^{K}\hat{v}_{t+1}(s^{m,k})\frac{1}{K-1}\sum_{l=1}^{K}\hat{v}_{t+1}(s^{m,l}) \\
&\quad + \frac{1}{MK^2}\sum_{m=1}^{M}\sum_{k=1}^{K} r_t^2(s,a,\omega^{m,k}) + 2\gamma r_t(s,a,\omega^{m,k})\hat{v}_{t+1}(s^{m,k}) + \gamma^2\hat{v}_{t+1}^2(s^{m,k})
\end{aligned}$$

where $\bar{r}_t^m(s,a) := \frac{1}{K}\sum_{k=1}^K r_t(s,a,\omega^{m,k})$. By the SLLN and Lemma 2 we can conclude that $M^{-1}\sum_{m=1}^M \bar{r}_t^m(s,a) \overset{a.s.}{\to} \mathbb{E}[r_t(s,a,\omega)|s,a]$, $M^{-1}\sum_{m=1}^M [\bar{r}_t^m(s,a)]^2 \overset{a.s.}{\to} \mathbb{E}[r_t(s,a,\omega)|s,a]^2$, $(MK)^{-1}\sum_{m=1}^M\sum_{k=1}^K r_t^2(s,a,\omega^{m,k}) \overset{a.s.}{\to} \mathbb{E}[r_t^2(s,a,\omega)|s,a]$, and $(MK)^{-1}\sum_{m=1}^M\sum_{k=1}^K \hat{v}_{t+1}(s^{m,k}) \overset{a.s.}{\to} \mathbb{E}[v_{t+1}(f_{t+1}(s,a,\omega))|s,a]$. Thus, it holds that:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^M \left[\bar{Q}_t^m(s,a)\right]^2 \overset{a.s.}{\to}\ & \mathbb{E}[r_t(s,a,\omega)|s,a]^2 - \frac{1}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]^2 \\ & + 2\gamma\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] - \frac{2\gamma}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] \\ & + \gamma^2\mathbb{E}[v_{t+1}(s')|s,a]^2 - \frac{\gamma^2}{K}\mathbb{E}[v_{t+1}(s')|s,a]^2 \\ & + \frac{1}{K}\mathbb{E}[r_t^2(s,a,\omega)|s,a] + \frac{2\gamma}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] + \frac{\gamma^2}{K}\mathbb{E}[v_{t+1}^2(s')|s,a]\end{aligned}$$

and we get:

$$\begin{aligned}\hat{\nu}_t^2(s,a) \overset{a.s.}{\to}\ & \mathbb{E}[r_t(s,a,\omega)|s,a]^2 - \frac{1}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]^2 \\ & + 2\gamma\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] - \frac{2\gamma}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] \\ & + \gamma^2\mathbb{E}[v_{t+1}(s')|s,a]^2 - \frac{\gamma^2}{K}\mathbb{E}[v_{t+1}(s')|s,a]^2 \\ & + \frac{1}{K}\mathbb{E}[r_t^2(s,a,\omega)|s,a] + \frac{2\gamma}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] + \frac{\gamma^2}{K}\mathbb{E}[v_{t+1}^2(s')|s,a] \\ & - \left(\mathbb{E}[r_t(s,a,\omega)|s,a] + \gamma\mathbb{E}[v_{t+1}(s')|s,a]\right)^2 \\ =\ & \frac{1}{K}\mathbb{E}[r_t^2(s,a,\omega)|s,a] + \frac{2\gamma}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] + \frac{\gamma^2}{K}\mathbb{E}[v_{t+1}^2(s')|s,a] \\ & - \frac{1}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]^2 - \frac{2\gamma}{K}\mathbb{E}[r_t(s,a,\omega)|s,a]\mathbb{E}[v_{t+1}(s')|s,a] - \frac{\gamma^2}{K}\mathbb{E}[v_{t+1}(s')|s,a]^2.\end{aligned}$$

Therefore,

$$K\hat{\nu}_t^2(s,a) \overset{a.s.}{\to} \mathbb{E}[(r_t(s,a,\omega)+\gamma v_{t+1}(s'))^2|s,a] - \mathbb{E}[r_t(s,a,\omega)+\gamma v_{t+1}(s')|s,a]^2.$$

Because $\hat{\zeta}_t^2(s,a) = \frac{M}{M-1}\hat{\nu}_t^2(s,a)$ and $\frac{M}{M-1}\to 1$ as $M\to\infty$, it holds that $K\hat{\zeta}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$. □

LEMMA 6. *Under Assumption 1, it holds that* $\sup_{x\in\mathbb{R}}|\hat{\mathbb{H}}_t(x,\hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x,\mathbb{F}_t(s))| \overset{a.s.}{\to} 0$.

Proof. Since $\boldsymbol{Q}_t(s,a)$ is a sequence of independent random variables with common distribution $\mathbb{F}_t(\cdot,s,a)$, it holds that $\sup_{x\in\mathbb{R}}|\hat{\mathbb{F}}_t(x,s,a) - \mathbb{F}_t(x,s,a)| \overset{a.s.}{\to} 0$ by the Glivenko-Cantelli Theorem. Because $Q_t(s,a)$ is independent from $Q_t(s,a')$ for $a\neq a'$, it follows that $\mathbb{F}_t(\cdot,s) = \prod_{a\in\mathcal{A}}\mathbb{F}_t(\cdot,s,a)$ and $\hat{\mathbb{F}}_t(\cdot,s) = \prod_{a\in\mathcal{A}}\hat{\mathbb{F}}_t(\cdot,s,a)$. Since $\hat{\mathbb{F}}_t(\cdot,s,a) \overset{a.s.}{\to} \mathbb{F}_t(\cdot,s,a)$ uniformly on $\mathbb{R}$ we also have that $\hat{\mathbb{F}}_t(\cdot,s) \overset{a.s.}{\to} \mathbb{F}_t(\cdot,s)$ uniformly on $\mathbb{R}^A$. We now show that $\lim_{M\to\infty}\lim_{K\to\infty}\hat{\mathbb{H}}_t(\cdot,\hat{\mathbb{F}}_t(s)) = \lim_{M\to\infty}\lim_{K\to\infty}\mathbb{H}_t(\cdot,\mathbb{F}_t(s))$. First, notice that:

$$\mathbb{H}_t(x,\mathbb{F}_t(s)) = \mathbb{P}\left(\max_{a\in\mathcal{A}}\{\psi_t(s,a)\}\leq x\right) = \prod_{a\in\mathcal{A}}\mathbb{P}\left(\frac{\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) - (Q_t(s,a^*) - Q_t(s,a))}{\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*) + \hat{\sigma}_t^2(s,a)\right]}} \leq x\right),$$

where the second equality holds because $\psi_t(s,a)$ is independent from $\psi_t(s,a')$ for $a\neq a'$. Since $\hat{Q}_t(s,a) \overset{a.s.}{\to} Q_t(s,a)$ from Theorem 1 and $\hat{\sigma}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$ by Lemma 5, by the Central Limit Theorem we have that:

$$\lim_{N\to\infty}\mathbb{P}\left(\frac{\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) - (Q_t(s,a^*) - Q_t(s,a))}{\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*) + \hat{\sigma}_t^2(s,a)\right]}} \leq x\right) = \Phi(x),$$

where $\Phi(\cdot)$ is the cdf of a standard normal random variable. Thus, $\lim_{N\to\infty} \mathbb{H}_t(x, \mathbb{F}_t(s)) = \Phi(x)^A$. Repeating these steps for $\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s))$ and noticing that $K\hat{\zeta}_t^2(s,a) \xrightarrow{a.s.} \sigma_t^2(s,a)$ by Lemma 5, we can deduce that $\lim_{M\to\infty}\lim_{K\to\infty} \hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) = \Phi(x)^A$. Hence, $\lim_{N\to\infty} \hat{\mathbb{H}}_t(\cdot, \hat{\mathbb{F}}_t(s)) = \lim_{N\to\infty} \mathbb{H}_t(\cdot, \mathbb{F}_t(s))$. Moreover, by Pólya's Theorem (Serfling 1980, Theorem 1.5.3) it follows that $\lim_{N\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x, \mathbb{F}_t(s))\right| = 0$ because of the continuity of $\lim_{N\to\infty} \mathbb{H}_t(x, \mathbb{F}_t(s)) = \Phi(x)^A$ for any $x \in \mathbb{R}$. Combining this result with the previous conclusion that $\hat{\mathbb{F}}_t(\cdot, s) \xrightarrow{a.s.} \mathbb{F}_t(\cdot, s)$ uniformly on $\mathbb{R}^A$, we get that $\sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x, \mathbb{F}_t(s))\right| \xrightarrow{a.s.} 0$. □

Proof of Proposition 5. From Lemma 6 we get that $\sup_{\tau\in\mathbb{R}} |\hat{\mathbb{H}}_t(\tau, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(\tau, \mathbb{F}_t(s))| \xrightarrow{a.s.} 0$. Since a.s. convergence implies convergence in distribution, it follows that $\sup_{\tau\in\mathbb{R}} |\hat{\mathbb{H}}_t(\tau, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(\tau, \mathbb{F}_t(s))| \xrightarrow{\mathcal{D}} 0$. Because $\mathbb{H}_t(\cdot, \mathbb{F}_t(s))$ is a true cdf, it is uniformly distributed on $[0,1]$. Therefore, $\hat{\mathbb{H}}_t(\cdot, \hat{\mathbb{F}}_t(s))$ must also follow a $\mathcal{U}(0,1)$ distribution asymptotically and we get that $\mathbb{P}(\hat{\mathbb{H}}_t(\cdot, \hat{\mathbb{F}}_t(s)) \le 1-\alpha) = \mathbb{P}(\mathcal{U}(0,1) \le 1-\alpha) = 1-\alpha$, for N large enough. □

THEOREM 2. *Suppose Assumption 1 holds. Then, for* $\hat{d}_t(s,\alpha) = \hat{\mathbb{H}}_t^{-1}(1-\alpha, \hat{\mathbb{F}}_t(s))$ *and* $a^* \in \mathcal{C}_t(s)$, *we have that* $\Pi_t(s,\alpha) = \left\{a \in \mathcal{A} : \hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) \le \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a)\right]}\right\}$ *is a set of* α*-nonsignificant actions with probability 1 for* N *large enough.*

Proof. From Proposition 5 we have that:

$$
\begin{aligned}
&\mathbb{P}\Big(Q_t(s,a^*) - Q_t(s,a) \in \Theta : \hat{\mathbb{H}}_t\left(\max_{a\in\mathcal{A}}\{\hat{\psi}_t(s,a)\}, \hat{\mathbb{F}}_t(s)\right) \le 1-\alpha\Big)\\
&= \mathbb{P}\Big(\hat{\mathbb{H}}_t\left(\max_{a\in\mathcal{A}}\{\hat{\psi}_t(s,a)\}, \hat{\mathbb{F}}_t(s)\right) \le 1-\alpha\Big)\\
&= \mathbb{P}\Big(\max_{a\in\mathcal{A}}\{\hat{\psi}_t(s,a)\} \le \hat{d}_t(s,\alpha)\Big)\\
&= \mathbb{P}\Big(\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a^1) - \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t(s,a^*) + \hat{\zeta}_t^2(s,a)\right]} \le Q_t(s,a^*) - Q_t(s,a^1),\\
&\ldots, \hat{Q}_t(s,a^*) - \hat{Q}_t(s,a^A) - \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t(s,a^*) + \hat{\zeta}_t^2(s,a)\right]} \le Q_t(s,a^*) - Q_t(s,a^A)\Big) = 1-\alpha,
\end{aligned}
$$

for N large enough, where $a^1, \ldots, a^A \in \mathcal{A}$. Moreover, from Lemma 6 it follows that $\hat{\mathbb{H}}_t(1-\alpha, \hat{\mathbb{F}}_t(s)) \xrightarrow{a.s.} \mathbb{H}_t(1-\alpha, \mathbb{F}_t(s))$ and hence, $\hat{d}_t(s,\alpha) = \hat{\mathbb{H}}_t^{-1}(1-\alpha, \hat{\mathbb{F}}_t(s)) \xrightarrow{a.s.} \mathbb{H}_t^{-1}(1-\alpha, \mathbb{F}_t(s)) = d_t(s,\alpha)$ by Theorem 2.3.1 in Serfling (1980). Consequently, the asymptotic confidence intervals simultaneously contain $Q_t(s,a^*) - Q_t(s,a^1), \ldots, Q_t(s,a^*) - Q_t(s,a^A)$ with probability exactly $1-\alpha$. Furthermore, under the null hypothesis (i.e., all actions have the same performance), any action a such that $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) \le d_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a)\right]}$ is not statistically significant from a^*. □

PROPOSITION 6. *Suppose that Assumption 1 is satisfied. Then, it follows that* $\lim_{N\to\infty} N^{1/2} \sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x, \mathbb{F}_t(s))\right| \le CA^{5/4}\sqrt{2\kappa_t^3}$, *where* C *is the constant appearing in the multivariate Berry-Esseen bound.*

Proof. Let $\Phi_A(x)$ denote the standard normal cdf in $\mathbb{R}^A$. By the triangle inequality we have that:

$$
\begin{aligned}
\lim_{N\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x, \mathbb{F}_t(s))\right| &= \lim_{N\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \Phi_A(x) + \Phi_A(x) - \mathbb{H}_t(x, \mathbb{F}_t(s))\right|\\
&\le \lim_{N\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \Phi_A(x)\right| + \lim_{N\to\infty} \sup_{x\in\mathbb{R}} |\Phi_A(x) - \mathbb{H}_t(x, \mathbb{F}_t(s))|.
\end{aligned} \tag{5}
$$

We show that each component of the right-hand side of equation (5) is bounded by $CA^{5/4}\sqrt{\frac{\kappa_t^3}{2N}}$. By definition, we have that:

$$\begin{aligned}\mathbb{H}_t(x,\mathbb{F}_t(s)) = \mathbb{P}\left(\max_{a\in\mathcal{A}}\{\psi_t(s,a)\}\leq x\right) = \mathbb{P}\Bigg(&\frac{\hat{Q}_t(s,a^*)-\hat{Q}_t(s,a^1)-(Q_t(s,a^*)-Q_t(s,a^1))}{\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*)+\hat{\sigma}_t^2(s,a^1)\right]}}\leq x,\ldots,\\ &\frac{\hat{Q}_t(s,a^*)-\hat{Q}_t(s,a^A)-(Q_t(s,a^*)-Q_t(s,a^A))}{\sqrt{N^{-1}\left[\hat{\sigma}_t^2(s,a^*)+\hat{\sigma}_t^2(s,a^A)\right]}}\leq x\Bigg)\\ = \mathbb{P}\Bigg(&\frac{1}{\sqrt{N}}\sum_n \hat{Z}_n(a^1)\leq x,\ldots,\frac{1}{\sqrt{N}}\sum_n \hat{Z}_n(a^A)\leq x\Bigg),\end{aligned}$$

where $a^1,\ldots,a^A\in\mathcal{A}$ and

$$\hat{Z}_n(a) \coloneqq \frac{Q_t^n(s,a^*)-Q_t^n(s,a)-(Q_t(s,a^*)-Q_t(s,a))}{\sqrt{\hat{\sigma}_t^2(s,a^*)+\hat{\sigma}_t^2(s,a)}}.$$

Since $\hat{\sigma}_t^2(s,a)\overset{a.s.}{\to}\sigma_t^2(s,a)$ from Lemma 5, by Lemma 2 we get that:

$$\hat{Z}_n(a)\overset{a.s.}{\to} Z_n(a) \coloneqq \frac{Q_t^n(s,a^*)-Q_t^n(s,a)-(Q_t(s,a^*)-Q_t(s,a))}{\sqrt{\sigma_t^2(s,a^*)+\sigma_t^2(s,a)}}.$$

Due to the assumption of bounded rewards, we have that $\hat{Q}_t(s,a)$ are also bounded. Since from Theorem 1 it holds that $\hat{Q}_t(s,a)\overset{a.s.}{\to}Q_t(s,a)$, from the Bounded Convergence Theorem it follows that $\lim_{M\to\infty}\lim_{K\to\infty}\mathbb{E}[\hat{Q}_t(s,a)] = Q_t(s,a)$. Moreover, as $Q_t^n(s,a)$ are iid random variables we get that $\lim_{N\to\infty}\mathbb{E}\left[Q_t^n(s,a)\right] = Q_t(s,a)$ and $\lim_{N\to\infty}\mathbb{E}\left[\left(Q_t^n(s,a)\right)^2\right] = \mathbb{E}[(r_t(s,a,\omega)+\gamma v_{t+1}(f_{t+1}(s,a,\omega)))^2|s,a]$. This implies that $\lim_{N\to\infty}\mathrm{Var}\left(Q_t^n(s,a)\right) = \sigma_t^2(s,a)$. Thus, we have that $Z_n(a)$ are iid random variables with $\mathbb{E}\left[Z_n(a)\right]=0$ and $\mathrm{Var}\left(Z_n(a)\right)=1$ for all n. Since the variance of $Z_n(a)/\sqrt{N}$ can be linearly transformed to 1, by the Multivariate Berry-Esseen Theorem we get that:

$$\lim_{N\to\infty}\sup_{x\in\mathbb{R}}|\Phi_A(x)-\mathbb{H}_t(x,\mathbb{F}_t(s))| \leq \frac{CA^{1/4}}{\sqrt{N}}\mathbb{E}\left[\left(Z_n^2(a^1)+Z_n^2(a^2)\ldots+Z_n^2(a^A)\right)^{3/2}\right].$$

We now show that $\mathbb{E}\left[\left(Z_n^2(a^1)+Z_n^2(a^2)\ldots+Z_n^2(a^A)\right)^{3/2}\right]\leq A\sqrt{\frac{1}{2}\kappa_t^3}$. As $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for any $x,y\in\mathbb{R}_+$, it follows that:

$$\begin{aligned}\mathbb{E}\left[\left(Z_n^2(a^1)+Z_n^2(a^2)+\ldots+Z_n^2(a^A)\right)^{3/2}\right] &= \mathbb{E}\left[\left(\sqrt{Z_n^2(a^1)+Z_n^2(a^2)+\ldots+Z_n^2(a^A)}\right)^3\right]\\ &\leq \mathbb{E}\left[\left(\sqrt{Z_n^2(a^1)}+\sqrt{Z_n^2(a^2)}+\ldots+\sqrt{Z_n^2(a^A)}\right)^3\right]\\ &= \mathbb{E}\left[\left(Z_n(a^1)+Z_n(a^2)+\ldots+Z_n(a^A)\right)^3\right].\end{aligned}$$

Expanding $\left(Z_n(a^1)+Z_n(a^2)+\ldots+Z_n(a^A)\right)^3$, we get a summation of terms of the following form: $\mathbb{E}[Z_n^3(a)]$, $\mathbb{E}[Z_n^2(a)Z_n(a')]$, and $\mathbb{E}[Z_n(a)Z_n(a')Z_n(a'')]$ for $a\neq a'\neq a''\in\mathcal{A}_t(s)$. Thus, $\mathbb{E}\left[\left(Z_n(a^1)+Z_n(a^2)+\ldots+Z_n(a^A)\right)^3\right] = \mathbb{E}\left[Z_n^3(a^1)\right]+\mathbb{E}\left[Z_n^2(a^1)Z_n(a^2)\right]+\ldots+\mathbb{E}\left[Z_n^3(a^A)\right]$. Due to the Markov property, we have that $\mathbb{E}[Z_n^2(a)Z_n(a')] = \mathbb{E}[Z_n^2(a)]\mathbb{E}[Z_n(a')] = 0$ and $\mathbb{E}[Z_n(a)Z_n(a')Z_n(a'')] = \mathbb{E}[Z_n(a)]\mathbb{E}[Z_n(a')]\mathbb{E}[Z_n(a'')] = 0$. Hence, $\mathbb{E}\left[\left(Z_n^2(a^1)+Z_n^2(a^2)+\ldots+Z_n^2(a^A)\right)^{3/2}\right]\leq\sum_{a\in\mathcal{A}}\mathbb{E}\left[Z_n^3(a)\right]$.

The proof proceeds by showing that $Z_n(a)\leq\sqrt{\frac{1}{2}\kappa_t^3}$ for all $a\in\mathcal{A}$. Since $Q_t(s,a^*)-Q_t(s,a)\geq 0$, we have that:

$$Z_n(a) = \frac{Q_t^n(s,a^*)-Q_t^n(s,a)-(Q_t(s,a^*)-Q_t(s,a))}{\sqrt{\sigma_t^2(s,a^*)+\sigma_t^2(s,a)}} \leq \frac{Q_t^n(s,a^*)-Q_t^n(s,a)}{\sqrt{\sigma_t^2(s,a^*)+\sigma_t^2(s,a)}}.$$

Further, by Lemma 1 it holds that $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) \leq \kappa_t$, which implies that $Z_n(a) \leq \kappa_t / \sqrt{\sigma_t^2(s,a^*) + \sigma_t^2(s,a)}$. From Theorem 2.2 in **?** we have that $\sigma_t^2(s,a) \geq \kappa_t^{-1}$, for $\sigma_t^2(s,a) > 0$ and it follows that $Z_n(a) \leq \kappa_t / \sqrt{2\kappa_t^{-1}} = \sqrt{\frac{1}{2}\kappa_t^3}$. Since this bound is valid for any $a \in \mathcal{A}$, $\mathbb{E}[Z_n(a)] = 0$, and $\mathrm{Var}(Z_n(a)) = 1$, we have that $\mathbb{E}[Z_n^3(a)] \leq \mathbb{E}\left[Z_n^2(a)\sqrt{\frac{\kappa_t^3}{2}}\right] = \mathbb{E}[Z_n^2(a)]\sqrt{\frac{1}{2}\kappa_t^3} = \sqrt{\frac{1}{2}\kappa_t^3}$. Consequently,

$$\mathbb{E}\left[\left(Z_n^2(a^1) + Z_n^2(a^2) \ldots + Z_n^2(a^A)\right)^{3/2}\right] \leq A\sqrt{\frac{1}{2}\kappa_t^3}, \tag{6}$$

and it follows that $\lim_{N\to\infty} \sup_{x\in\mathbb{R}} |\Phi_A(x) - \mathbb{H}_t(x, \mathbb{F}_t(s))| \leq CA^{5/4}\sqrt{\frac{\kappa_t^3}{2N}}$.

In a similar way,

$$\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) = \mathbb{P}\left(\frac{1}{\sqrt{MK}}\sum_m \sum_k \bar{Z}_{m,k}(a^1) \leq x, \ldots, \frac{1}{\sqrt{MK}}\sum_m \sum_k \bar{Z}_{m,k}(a^A) \leq x\right),$$

where $a^1, \ldots, a^A \in \mathcal{A}_t^-(s)$ and

$$\bar{Z}_{m,k}(a) \coloneqq \frac{Q_t^{m,k}(s,a^*) - Q_t^{m,k}(s,a) - (Q_t(s,a^*) - Q_t(s,a))}{\sqrt{K\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a)\right]}}.$$

Since $K\hat{\zeta}_t^2(s,a) \xrightarrow{a.s.} \sigma_t^2(s,a)$ from Lemma 5, by Lemma 2 we get that:

$$\bar{Z}_{m,k}(a) \xrightarrow{a.s.} Z_{m,k}(a) \coloneqq \frac{Q_t^{m,k}(s,a^*) - Q_t^{m,k}(s,a) - (Q_t(s,a^*) - Q_t(s,a))}{\sqrt{\sigma_t^2(s,a^*) + \sigma_t^2(s,a)}}.$$

As the variance of $Z_{m,k}(a)/\sqrt{MK}$ can be linearly transformed to 1, by the Multivariate Berry-Esseen Theorem we get that:

$$\lim_{M\to\infty}\lim_{K\to\infty}\sup_{x\in\mathbb{R}}\left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \Phi_A(x)\right| \leq \frac{CA^{1/4}}{\sqrt{MK}}\mathbb{E}\left[\left(Z_{m,k}^2(a^1) + Z_{m,k}^2(a^2) \ldots + Z_{m,k}^2(a^A)\right)^{3/2}\right],$$

and by (6) it holds that $\lim_{M\to\infty}\lim_{K\to\infty}\sup_{x\in\mathbb{R}}\left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \Phi_A(x)\right| \leq CA^{5/4}\sqrt{\frac{\kappa_t^3}{2MK}} = CA^{5/4}\sqrt{\frac{\kappa_t^3}{2N}}$. Summing both components of the right-hand side of equation (5) it follows that:

$$\lim_{M\to\infty}\lim_{K\to\infty}\sup_{x\in\mathbb{R}}\left|\hat{\mathbb{H}}_t(x, \hat{\mathbb{F}}_t(s)) - \mathbb{H}_t(x, \mathbb{F}_t(s))\right| \leq 2CA^{5/4}\sqrt{\frac{\kappa_t^3}{2N}} = CA^{5/4}\sqrt{\frac{2\kappa_t^3}{N}}.$$

□

A.5. Proofs of Section 5.2.2

PROPOSITION 7. *Suppose that Assumption 1 holds. Then, (i) $|\Pi_t(s,\alpha)|$ is nonincreasing in $\alpha \in (0,1)$. Moreover, (ii) there exist an α such that $\Pi_t(s,\alpha) \subseteq \mathcal{A}_t^*(s)$ with probability 1 for N large enough.*

Proof.

(i) We first show that $|\Pi_t(s,\alpha)|$ is nonincreasing in $\alpha \in (0,1)$. Note that $\hat{d}_t(s,\alpha)$ is the only quantity in $\Pi_t(s,\alpha)$ that depends on α. Suppose $\alpha_1 < \alpha_2$, by definition we have:

$$\begin{aligned}\hat{d}_t(s,\alpha_1) &= \inf\left\{x \in \mathbb{R} : \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\{\bar{\psi}_t^m(s,a) \leq x\} \geq 1-\alpha_1\right\} \\ &\geq \inf\left\{x \in \mathbb{R} : \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\{\bar{\psi}_t^m(s,a) \leq x\} \geq 1-\alpha_2\right\} = \hat{d}_t(s,\alpha_2).\end{aligned}$$

Thus, $\hat{d}_t(s,\alpha)$ is nonincreasing in α, which implies that $|\Pi_t(s,\alpha)|$ is nonincreasing in α.

(ii) We now show that there exist an α such that $\Pi_t(s,\alpha) \subseteq \mathcal{A}_t^*(s)$ with probability 1 for N large enough. Since if $\mathcal{C}_t(s) = \mathcal{A}_t^*(s)$ the claim follows trivially, suppose that $\mathcal{C}_t(s) = \hat{\mathcal{A}}_t^*(s)$. As $\mathbb{H}_t$ follows a $\mathcal{U}(0,1)$ distribution we have that $d_t(s,\alpha)$ is continuous and strictly increasing by the Continuous Inverse Theorem. The proof proceeds by contradiction. Suppose that there is no $\alpha \in (0,1)$ such that $\Pi_t(s,\alpha) = \hat{\mathcal{A}}_t^*(s)$. Because of the continuity of $\hat{d}_t(s,\alpha)$ there exists a point α such that $\hat{d}_t(s,1) < \hat{d}_t(s,\alpha) < \hat{d}_t(s,0)$ by the Intermediate Value Theorem. Thus, for every K, $\hat{\zeta}_t^2(s,a^*) > 0$, and $\hat{\zeta}_t^2(s,a) > 0$ we can find an α such that $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) > \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a)\right]}$, for any $a^* \in \hat{\mathcal{A}}_t^*(s)$ and all $a \in \hat{\mathcal{A}}_t^-(s)$, contradicting the supposition that there is no $\alpha \in (0,1)$ such that $\Pi_t(s,\alpha) = \hat{\mathcal{A}}_t^*(s)$. As from Corollary 1 it holds that $\hat{\mathcal{A}}_t^*(s) \subseteq \mathcal{A}_t^*(s)$, we get that $\Pi_t(s,\alpha) \subseteq \mathcal{A}_t^*(s)$ with probability 1 for N large enough. □

PROPOSITION 8. *Suppose Assumptions 1 through 4 hold. Then, $|\Pi_t(s,\alpha)|$ is ϵ-nonincreasing in s with probability 1 for N large enough.*

The proof of this proposition depends on the following lemma.

LEMMA 7. *Let X, Y, and Z be partially ordered finite sets, $g : X \mapsto \mathbb{R}$ be a nonincreasing function of X, and $h : X \times Y \times Z \mapsto [0,1]$ be a function satisfying $\sum_{x\in X} h(x,y^+,z^+) + h(x,y^-,z^-) = \sum_{x\in X} h(x,y^+,z^-) + h(x,y^-,z^+)$. Then, we have (i) $\sum_{x\in X}[h(x,y^+,z^+)+h(x,y^-,z^-)]\,g(x) \le \sum_{x\in X}[h(x,y^+,z^-)+h(x,y^-,z^+)]\,g(x)$, if h is a superadditive function and (ii) $\sum_{x\in X}[h(x,y^+,z^+)+h(x,y^-,z^-)]\,g(x) \ge \sum_{x\in X}[h(x,y^+,z^-)+h(x,y^-,z^+)]\,g(x)$, if h is a subadditive function for $y^+ \ge y^- \in Y$, and $z^+ \ge z^- \in Z$.*

Proof.

(i) Since h is a superadditive function, we have that $\sum_{x'\ge x}[h(x',y^+,z^+)+h(x',y^-,z^-)] \ge \sum_{x'\ge x}[h(x',y^+,z^-)+h(x',y^-,z^+)]$ and $\sum_{x\in X} h(x,y^+,z^+) + \sum_{x\in X} h(x,y^-,z^-) - \sum_{x\in X} h(x,y^+,z^-) - \sum_{x\in X} h(x,y^-,z^+) = 0$. Since $g(x)$ is nonincreasing on $x \in X$ it holds that:

$$
\begin{aligned}
0 &= \left[\sum_x h(x,y^+,z^+) + \sum_x h(x,y^-,z^-) - \sum_x h(x,y^+,z^-) - \sum_x h(x,y^-,z^+)\right] g(x^1) \\
&\ge h(x^1,y^+,z^+)g(x^1) + h(x^1,y^-,z^-)g(x^1) - h(x^1,y^+,z^-)g(x^1) - h(x^1,y^-,z^+)g(x^1) \\
&\quad + \left[\sum_{x\in X\setminus\{x^1\}} h(x,y^+,z^+) + \sum_{x\in X\setminus\{x^1\}} h(x,y^-,z^-) - \sum_{x\in X\setminus\{x^1\}} h(x,y^+,z^-) - \sum_{x\in X\setminus\{x^1\}} h(x,y^-,z^+)\right] g(x^2) \\
&\ge \sum_{x'\in\{x^1,x^2\}} h(x',y^+,z^+)g(x') + \sum_{x'\in\{x^1,x^2\}} h(x',y^-,z^-)g(x') - \sum_{x'\in\{x^1,x^2\}} h(x',y^+,z^-)g(x') \\
&\quad - \sum_{x'\in\{x^1,x^2\}} h(x',y^-,z^+)g(x') + \left[\sum_{x\in X\setminus\{x^1,x^2\}} h(x,y^+,z^+) + \sum_{x\in X\setminus\{x^1,x^2\}} h(x,y^-,z^-)\right. \\
&\quad \left. - \sum_{x\in X\setminus\{x^1,x^2\}} h(x,y^+,z^-) - \sum_{x\in X\setminus\{x^1,x^2\}} h(x,y^-,z^+)\right] g(x^3),
\end{aligned}
$$

where $x^1 \le x^2 \le x^3 \in X$, $y^+ \ge y^- \in Y$, and $z^+ \ge z^- \in Z$. This pattern implies that:

$$
0 \ge \sum_{x'\in X} h(x',y^+,z^+)g(x') + \sum_{x'\in X} h(x',y^-,z^-)g(x') - \sum_{x'\in X} h(x',y^+,z^-)g(x') - \sum_{x'\in X} h(x',y^-,z^+)g(x').
$$

Thus, $\sum_{x'\in X}[h(x',y^+,z^+)+h(x',y^-,z^-)]\,g(x') \le \sum_{x'\in X}[h(x',y^+,z^-)+h(x',y^-,z^+)]\,g(x')$, for $y^+ \ge y^- \in Y$ and $z^+ \ge z^- \in Z$.

(ii) For the subadditive case, note that we have $\sum_{x' \geq x} [h(x', y^+, z^+) + h(x', y^-, z^-)] \leq \sum_{x' \geq x} [h(x', y^+, z^-) + h(x', y^-, z^+)]$ by definition. The rest of the proof proceeds in the same way as the superadditive case and we get $\sum_{x' \in X} [h(x', y^+, z^+) + h(x', y^-, z^-)] g(x') \geq \sum_{x' \in X} [h(x', y^+, z^-) + h(x', y^-, z^+)] g(x')$, for $y^+ \geq y^- \in Y$ and $z^+ \geq z^- \in Z$. □

Proof of Proposition 8. To show that $|\Pi_t(s, \alpha)|$ is ϵ-nonincreasing in s, it suffices to demonstrate that $\hat{Q}_t(s, a^*) - \hat{Q}_t(s, a)$ is $\frac{1}{4}\epsilon$-nondecreasing when $\hat{d}_t(s, \alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s, a^*) + \hat{\zeta}_t^2(s, a)\right]}$ is $\frac{3}{4}\epsilon$-nonincreasing. We first prove that $\hat{Q}_t(s, a^*) - \hat{Q}_t(s, a)$ is $\frac{1}{4}\epsilon$-nondecreasing in s with probability 1 for N large enough.

Suppose that $p_t(s'|s, a)$ and $\mathbb{E}[r_t(s, a, \omega)|s, a]$ are known. Then, for any $s \in \mathcal{S}$ we have:

$$\begin{aligned} Q_t(s, a^*) - Q_t(s, a) &= \mathbb{E}[r_t(s, a^*, \omega) + \gamma v_{t+1}(s')|s, a^*] - \mathbb{E}[r_t(s, a, \omega) + \gamma v_{t+1}(s')|s, a] \\ &= \mathbb{E}[r_t(s, a^*, \omega)|s, a^*] - \mathbb{E}[r_t(s, a, \omega)|s, a] + \gamma \left(\sum_{s'} [p_t(s'|s, a^*) - p_t(s'|s, a)] v_{t+1}(s') \right). \end{aligned}$$

By Assumption 4, we have that $\mathbb{E}[r_t(s, a^*, \omega)|s, a^*] - \mathbb{E}[r_t(s, a, \omega)|s, a]$ is nondecreasing in s and it follows that:

$$Q_t(s, a^*) - Q_t(s, a) \leq \mathbb{E}[r_t(\bar{s}, a^*, \omega)|\bar{s}, a^*] - \mathbb{E}[r_t(\bar{s}, a, \omega)|\bar{s}, a] + \gamma \left(\sum_{s'} [p_t(s'|s, a^*) - p_t(s'|s, a)] v_{t+1}(s') \right),$$

for $s \leq \bar{s}$. From Assumption 2, we get that $v_{t+1}(s')$ is nonincreasing in s by Proposition 4.7.3 in Puterman (2014). Since $\bar{p}_t(s'|s, a)$ is subadditive on $\mathcal{S} \times \mathcal{A}$ by Assumption 4, applying Lemma 7 with $h(x, y^+, z^+) = p_t(s'|\bar{s}, a^*)$, $h(x, y^-, z^-) = p_t(s'|s, a)$, $h(x, y^+, z^-) = p_t(s'|\bar{s}, a)$, $h(x, y^-, z^+) = p_t(s'|s, a^*)$, and $g(x) = v_{t+1}(s')$ we get that $\sum_{s'} [p_t(s'|s, a^*) - p_t(s'|s, a)] v_{t+1}(s') \leq \sum_{s'} [p_t(s'|\bar{s}, a^*) - p_t(s'|\bar{s}, a)] v_{t+1}(s')$, for $s \leq \bar{s}$ and $a \leq a^*$. Therefore, it follows that:

$$\begin{aligned} Q_t(s, a^*) - Q_t(s, a) &\leq \mathbb{E}[r_t(\bar{s}, a^*, \omega)|\bar{s}, a^*] - \mathbb{E}[r_t(\bar{s}, a, \omega)|\bar{s}, a] + \gamma \mathbb{E}[v_{t+1}(s')|\bar{s}, a^*] - \gamma \mathbb{E}[v_{t+1}(s')|\bar{s}, a] \\ &= Q_t(\bar{s}, a^*) - Q_t(\bar{s}, a). \end{aligned} \tag{7}$$

Hence, $Q_t(s, a^*) - Q_t(s, a)$ is nondecreasing in s. From Theorem 1 and Lemma 2 we get that $\hat{Q}_t(s, a^*) - \hat{Q}_t(s, a) \xrightarrow{a.s.} Q_t(s, a^*) - Q_t(s, a)$, and it holds that $\hat{Q}_t(s, a^*) - \hat{Q}_t(s, a)$ is $\frac{1}{4}\epsilon$-nondecreasing in s for N large enough.

We now show that $\hat{d}_t(s, \alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s, a^*) + \hat{\zeta}_t^2(s, a)\right]}$ is $\frac{3}{4}\epsilon$-nondecreasing in s with probability 1 for N large enough. Combining Lemma 6 and Theorem 2.3.1 in Serfling (1980) we get that $\hat{d}_t(s, \alpha) \xrightarrow{a.s.} d_t(s, \alpha)$. Since $\hat{\psi}_t(s, a)$ is a pivotal statistic, its distribution does not depend on s. Thus, $d_t(s, \alpha)$ is constant in s and is $\hat{d}_t(s, \alpha)$ $\frac{1}{4}\epsilon$-constant with probability 1 for large enough N. Since $\mathbb{E}[r_t(s, a, \omega)|s, a]$ and $\mathbb{E}[v_{t+1}(s')|s, a]$ are conditionally independent given s and a, we get that $\sigma_t^2(s, a) = \mathbb{E}[r_t^2(s, a, \omega)|s, a] + \gamma^2 \mathbb{E}[v_{t+1}^2(s')|s, a] - \mathbb{E}[r_t(s, a, \omega)|s, a]^2 - \gamma^2 \mathbb{E}[v_{t+1}(s')|s, a]^2$. Further, from Assumption 3 we have that $\mathbb{E}[r_t^2(s, a, \omega)|s, a] - \mathbb{E}[r_t(s, a, \omega)|s, a]^2$ and $\mathbb{E}[v_{t+1}^2(s')|s, a] - \mathbb{E}[v_{t+1}(s')|s, a]^2$ are nonincreasing in s. Since $K\hat{\zeta}_t^2(s, a) \xrightarrow{a.s.} \sigma_t^2(s, a)$ by Lemma 5, it follows that $K\hat{\zeta}_t^2(s, a)$ is $\frac{1}{4}\epsilon$-nonincreasing in s for all a and large enough N. Consequently, $\hat{d}_t(s, \alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s, a^*) + \hat{\zeta}_t^2(s, a)\right]}$ is $\frac{3}{4}\epsilon$-nonincreasing in s with probability 1 for N large enough.

Thus, $\hat{Q}_t(s, a^*) - \hat{Q}_t(s, a) - \frac{1}{4}\epsilon \leq \hat{Q}_t(\bar{s}, a^*) - \hat{Q}_t(\bar{s}, a)$ and $\hat{d}_t(\bar{s}, \alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(\bar{s}, a^*) + \hat{\zeta}_t^2(\bar{s}, a)\right]} \leq \hat{d}_t(s, \alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s, a^*) + \hat{\zeta}_t^2(s, a)\right]} + \frac{3}{4}\epsilon$, for $\epsilon > 0$ and $s \leq \bar{s}$ with probability 1 for N large enough, which completes the proof. □

PROPOSITION 9. *Suppose Assumptions 1 through 5 hold. Furthermore, assume that $v_t(s) - v_{t+1}(s)$ is nondecreasing in s. Then, $|\Pi_t(s,\alpha)|$ is ϵ-nonincreasing in t with probability 1 for N large enough.*

Proof. To show that $|\Pi_t(s,\alpha)|$ is ϵ-nonincreasing in t, it suffices to demonstrate that $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a)$ is $\frac{1}{4}\epsilon$-nondecreasing when $\hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a)\right]}$ is $\frac{3}{4}\epsilon$-nonincreasing. We first prove that $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a)$ is ϵ-nondecreasing in t with probability 1 for N large enough.

Suppose that $p_t(s'|s,a)$ and $\mathbb{E}[r_t(s,a,\omega)|s,a]$ are known. Then, for any $t \in \mathcal{T} \setminus \{T\}$ we have:

$$\begin{aligned} Q_t(s,a^*) - Q_t(s,a) &= \mathbb{E}[r_t(s,a^*,\omega) + \gamma v_{t+1}(s')|s,a^*] - \mathbb{E}[r_t(s,a,\omega) + \gamma v_{t+1}(s')|s,a] \\ &= \mathbb{E}[r_t(s,a^*,\omega)|s,a^*] - \mathbb{E}[r_t(s,a,\omega)|s,a] + \gamma\left(\sum_{s'}\left[p_t(s'|s,a^*) - p_t(s'|s,a)\right]v_{t+1}(s')\right). \end{aligned}$$

By Assumption 4, we have that $\mathbb{E}[r_t(s,a^*,\omega)|s,a^*] - \mathbb{E}[r_t(s,a,\omega)|s,a]$ is nondecreasing in t and it follows that:

$$Q_t(s,a^*) - Q_t(s,a) \le \mathbb{E}[r_{t+1}(s,a^*,\omega)|s,a^*] - \mathbb{E}[r_{t+1}(s,a,\omega)|s,a] + \gamma\left(\sum_{s'}\left[p_t(s'|s,a^*) - p_t(s'|s,a)\right]v_{t+1}(s')\right).$$

Since $v_{t+1}(s')$ is nonincreasing in s from Assumption 2 and Proposition 4.7.3 in Puterman (2014), and $\bar{p}_t(s'|s,a)$ is a subadditive function on $\mathcal{T}\times\mathcal{A}$ from Assumption 4, we get:

$$\sum_{s'}\left[p_t(s'|s,a^*) - p_t(s'|s,a)\right]v_{t+1}(s') \le \sum_{s'}\left[p_{t+1}(s'|s,a^*) - p_{t+1}(s'|s,a)\right]v_{t+1}(s'), \tag{8}$$

by Lemma 7 with $h(x,y^+,z^+) = p_{t+1}(s'|s,a^*)$, $h(x,y^-,z^-) = p_t(s'|s,a)$, $h(x,y^+,z^-) = p_{t+1}(s'|s,a)$, $h(x,y^-,z^+) = p_t(s'|s,a^*)$, and $g(x) = v_{t+1}(s')$. Since $v_{t+1}(s') - v_{t+2}(s')$ is nondecreasing in s' by assumption and $\bar{p}_t(s'|s,a^*) \le \bar{p}_t(s'|s,a)$ from Assumption 5, by Lemma 4 it holds that:

$$\sum_{s'}\left[p_{t+1}(s'|s,a^*) - p_{t+1}(s'|s,a)\right]v_{t+1}(s') \le \sum_{s'}\left[p_{t+1}(s'|s,a^*) - p_{t+1}(s'|s,a)\right]v_{t+2}(s'). \tag{9}$$

Combining equations (8) and (9) we get that $\sum_{s'}\left[p_t(s'|s,a^*) - p_t(s'|s,a)\right]v_{t+1}(s') \le \sum_{s'}\left[p_{t+1}(s'|s,a^*) - p_{t+1}(s'|s,a)\right]v_{t+2}(s')$ and it follows that:

$$\begin{aligned} Q_t(s,a^*) - Q_t(s,a) &\le \mathbb{E}[r_{t+1}(s,a^*,\omega)|s,a^*] - \mathbb{E}[r_{t+1}(s,a,\omega)|s,a] + \gamma\mathbb{E}[v_{t+2}(s')|s,a^*] - \gamma\mathbb{E}[v_{t+2}(s')|s,a] \\ &= Q_{t+1}(s,a^*) - Q_{t+1}(s,a). \end{aligned} \tag{10}$$

Hence, $Q_t(s,a^*) - Q_t(s,a)$ is nondecreasing in t. From Theorem 1 and Lemma 2 we get that $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a) \xrightarrow{a.s.} Q_t(s,a^*) - Q_t(s,a)$, and it holds that $\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a)$ is $\frac{1}{4}\epsilon$-nondecreasing in t for N large enough.

We now show that $\hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a)\right]}$ is $\frac{3}{4}\epsilon$-nonincreasing in t. Combining Lemma 6 and Theorem 2.3.1 in Serfling (1980) we get that $\hat{d}_t(s,\alpha) \xrightarrow{a.s.} d_t(s,\alpha)$. Because $\hat{\psi}_t(s,a)$ is a pivotal statistic, its distribution does not depend on t. Thus, $d_t(s,\alpha)$ is constant in t and $\hat{d}_t(s,\alpha)$ is $\frac{1}{4}\epsilon$-constant in t with probability 1 for large enough N. Since $\mathbb{E}[r_t(s,a,\omega)|s,a]$ and $\mathbb{E}[v_{t+1}(s')|s,a]$ are conditionally independent given s and a, we get that $\sigma_t^2(s,a) = \mathbb{E}[r_t^2(s,a,\omega)|s,a] + \gamma^2\mathbb{E}[v_{t+1}^2(s')|s,a] - \mathbb{E}[r_t(s,a,\omega)|s,a]^2 - \gamma^2\mathbb{E}[v_{t+1}(s')|s,a]^2$. Further, from Assumption 3 we have that $\mathbb{E}[r_t^2(s,a,\omega)|s,a] - \mathbb{E}[r_t(s,a,\omega)|s,a]^2$ and $\mathbb{E}[v_{t+1}^2(s')|s,a] - \mathbb{E}[v_{t+1}(s')|s,a]^2$ are nonincreasing in t. Since $K\hat{\zeta}_t^2(s,a) \xrightarrow{a.s.} \sigma_t^2(s,a)$ by Lemma 5, it follows that $K\hat{\zeta}_t^2(s,a)$ is $\frac{1}{4}\epsilon$-nonincreasing

in t for large enough N. Consequently, $\hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*)+\hat{\zeta}_t^2(s,a)\right]}$ is $\frac{3}{4}\epsilon$-nonincreasing in t with probability 1 for N large enough. Combining this result with equation (10) we get:

$$\begin{aligned}\hat{Q}_t(s,a^*)-\hat{Q}_t(s,a)-\frac{1}{4}\epsilon &\leq \hat{Q}_{t+1}(s,a^*)-\hat{Q}_{t+1}(s,a)\\ &\leq \hat{d}_{t+1}(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_{t+1}^2(s,a^*)+\hat{\zeta}_{t+1}^2(s,a)\right]}\\ &\leq \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*)+\hat{\zeta}_t^2(s,a)\right]}+\frac{3}{4}\epsilon,\end{aligned}$$

with probability 1 for N large enough. □

REMARK 1 The conditions in Proposition 8 and Proposition 9 are sufficient to prove that there exist approximately optimal decision rules $\hat{\pi}_t^*(s)$ that are ϵ-monotone on (i) s and (ii) t with probability 1 for N large enough, respectively.

Proof.

(i) From Assumption 2 and Proposition 4.7.3 in Puterman (2014) we have that $v_{t+1}(s)$ is nonincreasing in s. Since $\bar{p}_t(s'|s,a)$ is subadditive on $\mathcal{S}\times\mathcal{A}$ by Assumption 4, applying Lemma 7 with $h(x,y^+,z^+)=p_t(s'|s^+,a^+)$, $h(x,y^-,z^-)=p_t(s'|s^-,a^-)$, $h(x,y^+,z^-)=p_t(s'|s^+,a^-)$, $h(x,y^-,z^+)=p_t(s'|s^-,a^+)$, and $g(x)=v_{t+1}(s')$ we conclude that $\sum_{s'}p_t(s'|s,a)v_{t+1}(s')$ is superadditive on $\mathcal{S}\times\mathcal{A}$. Because $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is superadditive on $\mathcal{S}\times\mathcal{A}$ by Assumption 4 and the sum of superadditive functions is superadditive, it follows that $Q_t(s,a)$ is superadditive on $\mathcal{S}\times\mathcal{A}$. We get that $\pi_t^*(s)$ are nondecreasing in s from Lemma 4.7.1 in Puterman (2014). The convergence result follows from Theorem 1.

(ii) By Assumption 4 we have that $\bar{p}_t(s'|s,a)$ is subadditive on $\mathcal{T}\times\mathcal{A}$. Further, by Assumption 2 and Proposition 4.7.3 in Puterman (2014) we have that $v_{t+1}(s)$ is nonincreasing in s. By Lemma 7 with $h(x,y^+,z^+)=p_{t+1}(s'|s,a^+)$, $h(x,y^-,z^-)=p_t(s'|s,a^-)$, $h(x,y^+,z^-)=p_{t+1}(s'|s,a^-)$, $h(x,y^-,z^+)=p_t(s'|s,a^+)$, and $g(x)=v_{t+1}(s')$, we get that:

$$\sum_{s'\in\mathcal{S}}\left[p_{t+1}(s'|s,a^+)+p_t(s'|s,a^-)\right]v_{t+1}(s')\geq\sum_{s'\in\mathcal{S}}\left[p_{t+1}(s'|s,a^-)+p_t(s'|s,a^+)\right]v_{t+1}(s'),$$

for $a^+\geq a^-$. Since $v_{t+1}(s')-v_{t+2}(s')$ is nondecreasing in s' by assumption and $\bar{p}_t(s'|s,a^-)\geq\bar{p}_t(s'|s,a^+)$ from Assumption 5, by Lemma 4 it holds that $\sum_{s'}p_t(s'|s,a)v_{t+1}(s')$ is a superadditive function on $\mathcal{T}\times\mathcal{A}$. Because $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is superadditive on $\mathcal{T}\times\mathcal{A}$ by Assumption 4 and the sum of superadditive functions is superadditive, it follows that $Q_t(s,a)$ is superadditive on $\mathcal{T}\times\mathcal{A}$. We get that $\pi_t^*(s)$ are nondecreasing in t from Lemma 4.7.1 in Puterman (2014). The convergence result follows from Theorem 1. □

THEOREM 3. *Suppose that Assumptions 1, 2, 3, and 5 are satisfied. Then, $\Pi_t(s,\alpha)$ is a ϵ-range of α-nonsignificant actions at state s and decision epoch t with probability 1 for N large enough.*

Proof. The proof is by contradiction. Suppose that $a^-,a^+\in\Pi_t(s,\alpha)$ but $a'\notin\Pi_t(s,\alpha)$ with $a^-\leq a'\leq a^+$ for a fixed state s. We first show that $Q_t(s,a)$ is nondecreasing in a. By Assumption 2, we have that $v_t(s)$ is

nonincreasing in s from Proposition 4.7.3 in Puterman (2014). Combining this result with the assumption that $\bar{p}_t(s'|s,a)$ is nonincreasing in a we can deduce that:

$$\sum_{s'} p_t(s'|s,a^-)v_{t+1}(s') \leq \sum_{s'} p_t(s'|s,a')v_{t+1}(s') \leq \sum_{s'} p_t(s'|s,a^+)v_{t+1}(s'),$$

indicating that $\mathbb{E}\big[v_{t+1}(s')|s,a\big]$ is nondecreasing in a. Since $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is nondecreasing a by Assumption 5, we then get that $Q_t(s,a)$ is nondecreasing in a. Moreover, by Theorem 1 we have that $\hat{Q}_t(s,a) \overset{a.s.}{\to} Q_t(s,a)$ uniformly on $\mathcal{A}$. Hence, we can find an N large enough such that $\hat{Q}_t(s,a^-) - \frac{1}{4}\epsilon \leq \hat{Q}_t(s,a') \leq \hat{Q}_t(s,a^+) + \frac{1}{4}\epsilon$ for any $\epsilon > 0$.

We now show that $\sigma_t(s,a^-)^2 \geq \sigma_t(s,a')^2 \geq \sigma_t(s,a^+)^2$. Notice that $\sigma_t^2(s,a) = \mathbb{E}[r_t^2(s,a,\omega)|s,a] + \gamma^2\mathbb{E}[v_{t+1}^2(s')|s,a] - \mathbb{E}[r_t(s,a,\omega)|s,a]^2 - \gamma^2\mathbb{E}[v_{t+1}(s')|s,a]^2$, because $\mathbb{E}[r_t(s,a,\omega)|s,a]$ and $\mathbb{E}[v_{t+1}(s')|s,a]$ are conditionally independent given s and a. From Assumption 3, it follows that $\mathbb{E}[r_t^2(s,a,\omega)|s,a] - \mathbb{E}[r_t(s,a,\omega)|s,a]^2$ and $\mathbb{E}[v_{t+1}^2(s')|s,a] - \mathbb{E}[v_{t+1}(s')|s,a]^2$ are nonincreasing in a. Thus, $\sigma_t^2(s,a)$ is nonincreasing in a and it holds that $\sigma_t^2(s,a^-) \geq \sigma_t^2(s,a') \geq \sigma_t^2(s,a^+)$. Since $K\hat{\zeta}_t^2(s,a) \overset{a.s.}{\to} \sigma_t^2(s,a)$ by Lemma 5, it follows that $K\hat{\zeta}_t^2(s,a)$ is $\frac{1}{4}\epsilon$-nonincreasing in a for large enough N. Hence, we have that $K\hat{\zeta}_t^2(s,a^-) + \frac{1}{4}\epsilon \geq K\hat{\zeta}_t^2(s,a') \geq K\hat{\zeta}_t^2(s,a^+) - \frac{1}{4}\epsilon$ for $\epsilon > 0$ and it follows that:

$$\begin{aligned}\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a^-) - \frac{1}{4}\epsilon \leq \hat{Q}_t(s,a^*) - \hat{Q}_t(s,a') &\leq \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a')\right]} \\ &\leq \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a^-)\right]} + \frac{1}{4}\epsilon,\end{aligned}$$

and

$$\begin{aligned}\hat{Q}_t(s,a^*) - \hat{Q}_t(s,a^+) + \frac{1}{4}\epsilon \geq \hat{Q}_t(s,a^*) - \hat{Q}_t(s,a') &\geq \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a')\right]} \\ &\geq \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*) + \hat{\zeta}_t^2(s,a^+)\right]} - \frac{1}{4}\epsilon,\end{aligned}$$

for any $\epsilon > 0$. Combining these results we find that $a' \in \Pi_t(s,\alpha)$, a contradiction. Consequently, it must hold that $a^-, a', a^+ \in \Pi_t(s,\alpha)$ and $\Pi_t(s,\alpha)$ is an ϵ-range of actions. □

PROPOSITION 10. *Suppose Assumptions 1, 2, 4 and 5 hold. Then, we have that (i) $\mathscr{A}_t^*(s,\tilde{a}) \subseteq \mathcal{A}_t^*(s)$ and (ii) $\mathit{\Pi}_t(s,\alpha,\tilde{a}) \subseteq \Pi_t(s,\alpha)$ for N large enough.*

The proof of this result depends on the following notation. Let $\mathcal{Q}_t(s,a,\tilde{a}) := \mathbb{E}\left[r_t(s,a,\omega) + \gamma Q_{t+1}(f_{t+1}(s,a,\omega),\tilde{a})|s,a,\tilde{a}\right]$ denote the action-value function associated with state s and action a at decision epoch t, assuming that action $\tilde{a} \in \mathcal{A}_{t+1}(f_{t+1}(s,a,\omega))$ is taken at decision epoch $t+1$, and $\hat{\mathcal{Q}}_t(s,a,\tilde{a}) := \frac{1}{MK}\sum_{m=1}^{M} r_t(s,a,\omega^{m,k}) + \gamma\hat{Q}_{t+1}(f_{t+1}(s,a,\omega^{m,k}),\tilde{a})$ denote its empirical estimate. Moreover, let $\mathscr{A}_t^*(s,\tilde{a}) := \arg\max_{a\in\mathcal{A}} \mathcal{Q}_t(s,a,\tilde{a})$ and $\mathscr{A}_t^-(s,\tilde{a}) := \mathcal{A} \setminus \mathscr{A}_t^*(s,\tilde{a})$.

In a similar way, we define the range of α-nonsignificant actions given that action $\tilde{a} \in \mathcal{A}_{t+1}(f_{t+1}(s,a,\omega))$ will be taken at $t+1$ as:

$$\mathit{\Pi}_t(s,\alpha,\tilde{a}) := \left\{a \in \mathcal{A} : \hat{\mathcal{Q}}_t(s,a^*,\tilde{a}) - \hat{\mathcal{Q}}_t(s,a,\tilde{a}) \leq \hat{d}_t(s,\alpha,\tilde{a})\sqrt{M^{-1}\left[\hat{\varsigma}_t^2(s,a^*,\tilde{a}) + \hat{\varsigma}_t^2(s,a,\tilde{a})\right]}\right\},$$

where $\hat{d}_t(s,\alpha,\tilde{a})$ is the $1-\alpha$ empirical quantile of the distribution of the maximum of the root statistics and $\hat{\varsigma}_t^2(s,a,\tilde{a}) := \frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{\mathcal{Q}}_t^m(s,a,\tilde{a}) - \hat{\mathcal{Q}}_t(s,a,\tilde{a})\right)^2$, with $\bar{\mathcal{Q}}_t^m(s,a) := \frac{1}{K}\sum_{k=1}^{K} r_t(s,a,\omega^{m,k}) + \gamma Q_{t+1}(f_{t+1}(s,a,\omega^{m,k}),\tilde{a})$.

Proof of Proposition 10.

(i) We first show that $\mathscr{A}_t^*(s,\tilde{a}) \subseteq \mathcal{A}_t^*(s)$. Let $\bar{a} \in \mathcal{A}_t^-(s)$ and $a^* \in \mathcal{A}_t^*(s)$. We then have that $\mathbb{E}[r_t(s,\bar{a},\omega)+\gamma v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)+\gamma v_{t+1}(s')|s,a^*]$ for $s' = f_{t+1}(s,a,\omega)$. Since $\mathbb{E}[r_t(s,a,\omega)+\gamma v_{t+1}(s')|s,a] = \mathbb{E}[r_t(s,a,\omega)|s,a] + \gamma\mathbb{E}[v_{t+1}(s')|s,a]$, we must have one of these cases: (1) $\mathbb{E}[r_t(s,\bar{a},\omega)|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)|s,a^*]$, (2) $\mathbb{E}[v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[v_{t+1}(s')|s,a^*]$, or (3) $\mathbb{E}[r_t(s,\bar{a},\omega)|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)|s,a^*]$ and $\mathbb{E}[v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[v_{t+1}(s')|s,a^*]$. We want to show that if $\mathbb{E}[r_t(s,\bar{a},\omega)+\gamma v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)+\gamma v_{t+1}(s')|s,a^*]$ then $\mathbb{E}[r_t(s,\bar{a},\omega)+\gamma Q_{t+1}(s',\tilde{a})|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)+\gamma Q_{t+1}(s',\tilde{a})|s,a^*]$, for $\tilde{a} \in \mathcal{A}_{t+1}(s')$. The proof proceeds by showing that $\mathscr{A}_t^*(s,\tilde{a}) \subseteq \mathcal{A}_t^*(s)$ in each case via contrapositive arguments.

Case (1): $\mathbb{E}[r_t(s,\bar{a},\omega)|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)|s,a^*]$. Since $\bar{p}_t(s'|s,a)$ is nonincreasing in a by Assumption 5 and $Q_{t+1}(s',\tilde{a})$ is nonincreasing in s' by Assumption 2 and Proposition 3, we have that $\mathbb{E}[Q_{t+1}(s',\tilde{a})|s,\bar{a}] \leq \mathbb{E}[Q_{t+1}(s',\tilde{a})|s,a^*]$ for any $\tilde{a} \in \mathcal{A}_{t+1}(s')$. Thus, $\mathbb{E}[r_t(s,\bar{a},\omega)+\gamma Q_{t+1}(s',\tilde{a})|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)+\gamma Q_{t+1}(s',\tilde{a})|s,a^*]$ and $\bar{a} \in \mathcal{A}_t^-(s)$ implies that $\bar{a} \in \mathscr{A}_t^-(s,\tilde{a})$. The inclusion $\mathscr{A}_t^*(s,\tilde{a}) \subseteq \mathcal{A}_t^*(s)$ follows.

Case (2): $\mathbb{E}[v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[v_{t+1}(s')|s,a^*]$. Note that $\mathbb{E}[v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[v_{t+1}(s')|s,a^*]$ implies that $\mathbb{E}[v_{t+1}(s')|s,\bar{a}] - \xi < \mathbb{E}[v_{t+1}(s')|s,a^*] - \xi$, for any $\xi \in \mathbb{R}$. Let $\xi_{t+1}(s') \coloneqq v_{t+1}(s') - Q_{t+1}(s',\tilde{a})$. We then have:

$$
\begin{aligned}
&\mathbb{E}[v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[v_{t+1}(s')|s,a^*] \\
&\Rightarrow \mathbb{E}[v_{t+1}(s')|s,\bar{a}] - \mathbb{E}[\xi_{t+1}(s')|s,\bar{a}] < \mathbb{E}[v_{t+1}(s')|s,a^*] - \mathbb{E}[\xi_{t+1}(s')|s,\bar{a}] \\
&\Rightarrow \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})v_{t+1}(s') - \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})\xi_{t+1}(s') < \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)v_{t+1}(s') - \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})\xi_{t+1}(s') \\
&\Rightarrow \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})[v_{t+1}(s') - \xi_{t+1}(s')] < \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)v_{t+1}(s') - \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})\xi_{t+1}(s') \\
&\Rightarrow \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})Q_{t+1}(s',\tilde{a}) < \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)v_{t+1}(s') - \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})\xi_{t+1}(s'). \qquad (11)
\end{aligned}
$$

Since $\mathbb{E}[r_t(s,a^*,\omega)|s,a^*] - \mathbb{E}[r_t(s,a,\omega)|s,a]$ is nondecreasing in s and $\bar{p}_t(s'|s,a)$ is subadditive on $\mathcal{S}\times\mathcal{A}$ by Assumption 4, from Lemma 7 we then get that $\xi_{t+1}(s')$ is nondecreasing in s' (see equation (7) in the proof of Proposition 8). Moreover, as $\bar{p}_t(s'|s,a)$ is nonincreasing in a by Assumption 5, applying Lemma 4 with $h(x,y) = p_t(s'|s,a^*)$, $h(x,\bar{y}) = p_t(s'|s,\bar{a})$, and $g(x) = \xi_{t+1}(s')$ it follows that $\sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)\xi_{t+1}(s') \leq \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})\xi_{t+1}(s')$. Hence, from equation (11) we have:

$$
\begin{aligned}
\sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})Q_{t+1}(s',\tilde{a}) &< \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)v_{t+1}(s') - \sum_{s'\in\mathcal{S}} p_t(s'|s,\bar{a})\xi_{t+1}(s') \\
&\leq \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)v_{t+1}(s') - \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)\xi_{t+1}(s') \\
&= \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)[v_{t+1}(s') - \xi_{t+1}(s')] \\
&= \sum_{s'\in\mathcal{S}} p_t(s'|s,a^*)Q_{t+1}(s',\tilde{a}),
\end{aligned}
$$

which implies that $\mathbb{E}[Q_{t+1}(s',\tilde{a})|s,\bar{a}] < \mathbb{E}[Q_{t+1}(s',\tilde{a})|s,a^*]$. Because $\mathbb{E}[r_t(s,a,\omega)|s,a]$ is nondecreasing in a by Assumption 5, we get that $\mathbb{E}[r_t(s,\bar{a},\omega)+\gamma Q_{t+1}(s',\tilde{a})|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)+\gamma Q_{t+1}(s',\tilde{a})|s,a^*]$ and $\bar{a}\in\mathcal{A}_t^-(s)$ implies that $\bar{a}\in\mathscr{A}_t^-(s,\tilde{a})$. The inclusion $\mathscr{A}_t^*(s,\tilde{a})\subseteq\mathcal{A}_t^*(s)$ holds.

Case (3): $\mathbb{E}[r_t(s,\bar{a},\omega)|s,\bar{a}] < \mathbb{E}[r_t(s,a^*,\omega)|s,a^*]$ and $\mathbb{E}[v_{t+1}(s')|s,\bar{a}] < \mathbb{E}[v_{t+1}(s')|s,a^*]$. This case follows directly from cases (1) and (2).

Since $\mathscr{A}_t^*(s,\tilde{a})\subseteq\mathcal{A}_t^*(s)$ in all 3 cases, we have that $\mathcal{A}_t^*(s)\subseteq\mathscr{A}_t^*(s)$.

(ii) We now show that $\mathit{\Pi}_t(s,\alpha,\tilde{a})\subseteq\Pi_t(s,\alpha)$ via another contrapositive argument. Fix a realization of the sequence of the stochastic process $\boldsymbol{\omega}_N \coloneqq (\omega^n : n\in\{1,\dots,N\})$ and let $\bar{a}\notin\Pi_t(s,\alpha)$. Suppose that $\tilde{a}\in\mathcal{A}_{t+1}^-(f_{t+1}(s,a,\omega))$. If $\tilde{a}\in\mathcal{A}_{t+1}^*(f_{t+1}(s,a,\omega))$ the result is trivially true. We want to prove that:

$$\hat{Q}_t(s,a^*)-\hat{Q}_t(s,\bar{a}) > \hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*)+\hat{\zeta}_t^2(s,\bar{a})\right]}, \tag{12}$$

suggests that $\hat{\mathcal{Q}}_t(s,a^*,\tilde{a}) - \hat{\mathcal{Q}}_t(s,\bar{a},\tilde{a}) > \hat{d}_t(s,\alpha,\tilde{a})\sqrt{M^{-1}[\hat{\varsigma}_t^2(s,a^*,\tilde{a})+\hat{\varsigma}_t^2(s,\bar{a},\tilde{a})]}$ for any $\tilde{a}\in\mathcal{A}_{t+1}^-(f_{t+1}(s,\bar{a},\omega))$. Suppose that $\mathbb{E}[r_t(s,a,\omega)|s,a]$ and $p_t(s'|s,a)$ are known. From the left-hand side of equation (12) we want to show that $\mathcal{Q}_t(s,a^*,\tilde{a}) - \mathcal{Q}_t(s,\bar{a},\tilde{a}) \geq Q_t(s,a^*) - Q_t(s,\bar{a})$. Since $\mathcal{Q}_t(s,a,\tilde{a}) = \mathbb{E}[r_t(s,a,\omega)|s,a] + \gamma\mathbb{E}[Q_{t+1}(s',\tilde{a})|s,a]$ and $Q_t(s,a) = \mathbb{E}[r_t(s,a,\omega)|s,a] + \gamma\mathbb{E}[v_{t+1}(s')|s,a]$ for $s' = f_{t+1}(s,a,\omega)$, $\tilde{a}\in\mathcal{A}_{t+1}(s')$, and $a\in\mathcal{A}_t(s)$, it suffices to show that $\mathbb{E}[Q_{t+1}(s',\tilde{a})|s,a^*] - \mathbb{E}[Q_{t+1}(s',\tilde{a})|s,\bar{a}] \geq \mathbb{E}[v_{t+1}(s')|s,a^*] - \mathbb{E}[v_{t+1}(s')|s,\bar{a}]$. As $\mathbb{E}[r_t(s,a^*,\omega)|s,a^*] - \mathbb{E}[r_t(s,a,\omega)|s,a]$ is nondecreasing in s and $\bar{p}_t(s'|s,a)$ is subadditive on $\mathcal{S}\times\mathcal{A}$ by Assumption 4, from Lemma 7 we get that $v_{t+1}(s') - Q_{t+1}(s',\tilde{a})$ is nondecreasing in s' (see equation (7) in the proof of Proposition 8). Because $\bar{p}_{t+1}(s'|s,a)$ is nonincreasing in a by Assumption 5, from Lemma 4 it holds that:

$$\sum_{s'} p_t(s'|s,\bar{a})\left[v_{t+1}(s') - Q_{t+1}(s',\tilde{a})\right] \geq \sum_{s'} p_t(s'|s,a^*)\left[v_{t+1}(s') - Q_{t+1}(s',\tilde{a})\right],$$

indicating that $\mathbb{E}[Q_{t+1}(s',\tilde{a})|s,a^*] - \mathbb{E}[Q_{t+1}(s',\tilde{a})|s,\bar{a}] \geq \mathbb{E}[v_{t+1}(s')|s,a^*] - \mathbb{E}[v_{t+1}(s')|s,\bar{a}]$. Therefore, $\mathcal{Q}_t(s,a^*,\tilde{a}) - \mathcal{Q}_t(s,\bar{a},\tilde{a}) \geq Q_t(s,a^*) - Q_t(s,\bar{a})$. The convergence result follows from Theorem 1.

For the right-hand side of (12), note that $\hat{\zeta}_t^2(s,a) \geq \hat{\varsigma}_t^2(s,a)$ and $\hat{d}_t(s,\alpha) \geq \hat{d}_t(s,\alpha,a)$ for all a. Both inequalities follow because $\boldsymbol{\omega}_N$ is fixed, which implies that $r_t(s,a,\omega^{m,k}) + \gamma\hat{v}_{t+1}(f_{t+1}(s,a,\omega^{m,k})) \geq r_t(s,a,\omega^{m,k}) + \gamma\hat{Q}_{t+1}(f_{t+1}(s,a,\omega^{m,k}),\tilde{a})$ for every $\omega^{m,k}\in\boldsymbol{\omega}_N$. Thus, we get that $\hat{d}_t(s,\alpha)\sqrt{M^{-1}\left[\hat{\zeta}_t^2(s,a^*)+\hat{\zeta}_t^2(s,\bar{a})\right]} \geq \hat{d}_t(s,\alpha,\tilde{a})\sqrt{M^{-1}[\hat{\varsigma}_t^2(s,a^*)+\hat{\varsigma}_t^2(s,\bar{a})]}$. Note that this inequality holds for any $N\in\mathbb{N}_+$. Combining the results from both sides of equation (12) it follows that $\hat{\mathcal{Q}}_t(s,a^*,\tilde{a}) - \hat{\mathcal{Q}}_t(s,\bar{a},\tilde{a}) > \hat{d}_t(s,\alpha,\tilde{a})\sqrt{M^{-1}[\hat{\varsigma}_t^2(s,a^*,\tilde{a})+\hat{\varsigma}_t^2(s,\bar{a},\tilde{a})]}$. □

Appendix B: Case Study Details

In this section, we provide additional methods and results from our case study.

B.1. Ordering of States and Actions

To obtain a range of near-optimal treatment choices with probability 1, $\mathcal{S}$ and $\mathcal{A}_t(s)$ must be ordered such that $\mathbb{E}[r_t(s',a,\omega)|s,a]$ and $\bar{p}_t(s'|s,a)$ are monotone on $s\in\mathcal{S}$ and $a\in\mathcal{A}_t(s)$, and $\sigma_t(s,a)$ is monotone on $a\in\mathcal{A}_t(s)$ (see Theorem 3).

Given the progression of a patient's risk factors, the state transitions only depend on their health condition. To ensure the monotonicity of $\mathbb{E}[r_t(s',a,\omega)|s,a]$ and $\bar{p}_t(s'|s,a)$ on s, we ordered patients' states at each year in terms of their health condition. As $p_t(s|s,a)=1$ if $s\in\mathcal{E}$ and $p_t(s'|s,a)\in(0,1)$ if $s\in\mathcal{H}$, we order our states so that $\hat{s}<\tilde{s}$ if $\hat{s}\in\mathcal{H}$ and $\tilde{s}\in\mathcal{E}$. This ordering guarantees that $\bar{p}_t(s'|s,a)$ is monotone on s. Since patients only receive a nonzero reward if they transition from $s\in\mathcal{H}$ to $s'\in\mathcal{H}$, $\mathbb{E}[r_t(s',a,\omega)|s,a]$ is monotone on s by construction.

To make sure $\mathbb{E}[r_t(s,a,\omega)|s,a]$, $\bar{p}_t(s'|s,a)$, and $\sigma_t(s,a)$ are monotone on a, we ordered $\mathcal{A}_t(s)$ in terms of number of medications. We note that this ordering achieves the desired result because the reduction in ASCVD risk from treatment is linear in the number of medications.

B.2. Calibration and Validation of Simulation Model

We calibrate the number of events predicted by the ASCVD risk calculator to ensure the number of fatal and non-fatal ASCVD events in our simulation match those of the national statistics. The overall event rates predicted by the risk score are estimated by simulating the first year of the 10-year planning horizon of every patient in our population following the current clinical guidelines. We estimate the likelihood of fatal CHD and stroke events by dividing the fatal event rates from the National Center for Health Statistics (NCHS) by the overall event rates predicted by the risk calculator in our simulation (NCHS 2017). In a baseline simulation in which the study population is untreated, we find that the event rates are calibrated to the national data. A clinical researcher from the University of Michigan Medical School verified the calibration of our model.

Our simulation model is built by discussing the parameters and logic with experts in the field. Practicing clinicians at the University of Michigan Hospital and researchers at the Veterans Affairs Ann Arbor Healthcare System, helped validate our model.

B.3. Convergence Analysis

To select the number of batches for each patient in our population, we first fixed the number of observations per batch to satisfy the conditions in Proposition 2 with $\beta=0.01$. That is, each batch has $K=\left\lceil 2\kappa_t^2\log(21/0.01)\right\rceil$ number of observations, where $\lceil x\rceil:=\min\{y\in\mathbb{Z}|y\geq x\}$. Recall that at each year t and healthy state $s\in\mathcal{H}$ there is a total of $A=21$ treatment choices. Using this approach, the approximately optimal treatment choices identified in each batch are contained in the true sets of optimal actions with a probability of at least 99% for each patient. Note that this approach results in a different number of observations K for each patient in the population. We then increase the number of batches (or simulation replicates) iteratively until the maximum width of the simultaneous confidence intervals across all patients at the first year of our study reaches convergence. Through this method, we find that $M=300$ batches may be enough to obtain a maximum confidence interval width close to the maximum width attained with 1,000 batches (Figure B.1). We also note that using $M=300$ batches, we achieve a maximum confidence interval width of 0.02 life years. This width implies that any treatment choice that results in less than 0.02 life years than the controls will be excluded from the ranges of actions.

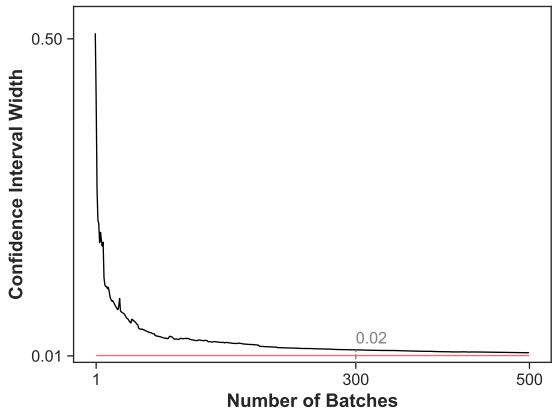

Figure B.1 **Convergence of the maximum confidence interval width over the number of batches. Red line represents the maximum confidence interval width using 1,000 batches (0.01).**

B.4. Study Population

Out of a population of 16.72 million people, 1.33 million (7.96%) are Black females, 7.58 million (45.34%) are White females, 1.08 million (6.44%) are Black males, and 6.73 million (40.26%) are White males. The number of people by sex, race, and BP group (normal BP, elevated BP, stage 1 hypertension, and stage 2 hypertension) at the first year of our study are shown in Figure B.2. We observe that male patients generally have higher BP than female patients. We also notice that 4.71 million people (28.15%) have stage 1 hypertension. This is the second-largest proportion of adults in the US, with ages 50 to 54. Nevertheless, this finding varies by race. While there are more Black adults with stage 1 hypertension than any other BP group (29.17%), the largest proportion of White adults have normal BP (38.59%). These findings are consistent with the most recent age-adjusted hypertension prevalence trends across adults in the US (Virani et al. 2020).

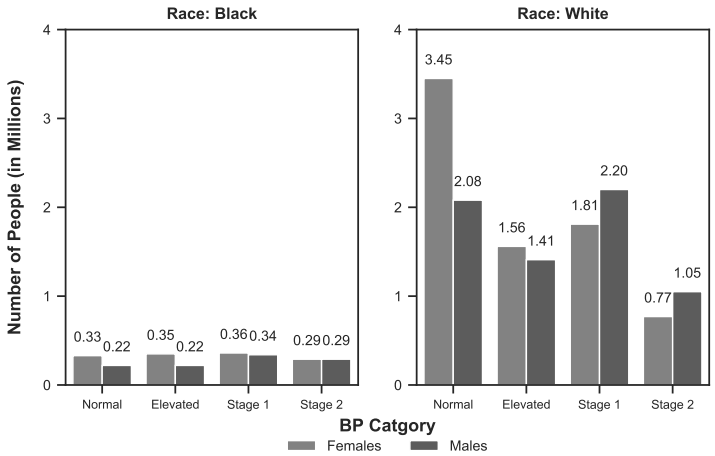

Figure B.2 **Number of people by race, race, and BP category. BP groups are consistent with the BP categories of the 2017 Hypertension Clinical Practice Guidelines. The label "Elevated" denotes elevated BP, "Stage 1" denotes stage 1 hypertension, and the label "Stage 2" denotes stage 2 hypertension.**

B.5. Additional Results

This subsection presents additional results of our case study. All of the results included in this subsection have been described in Section 6 in the main body of the paper.

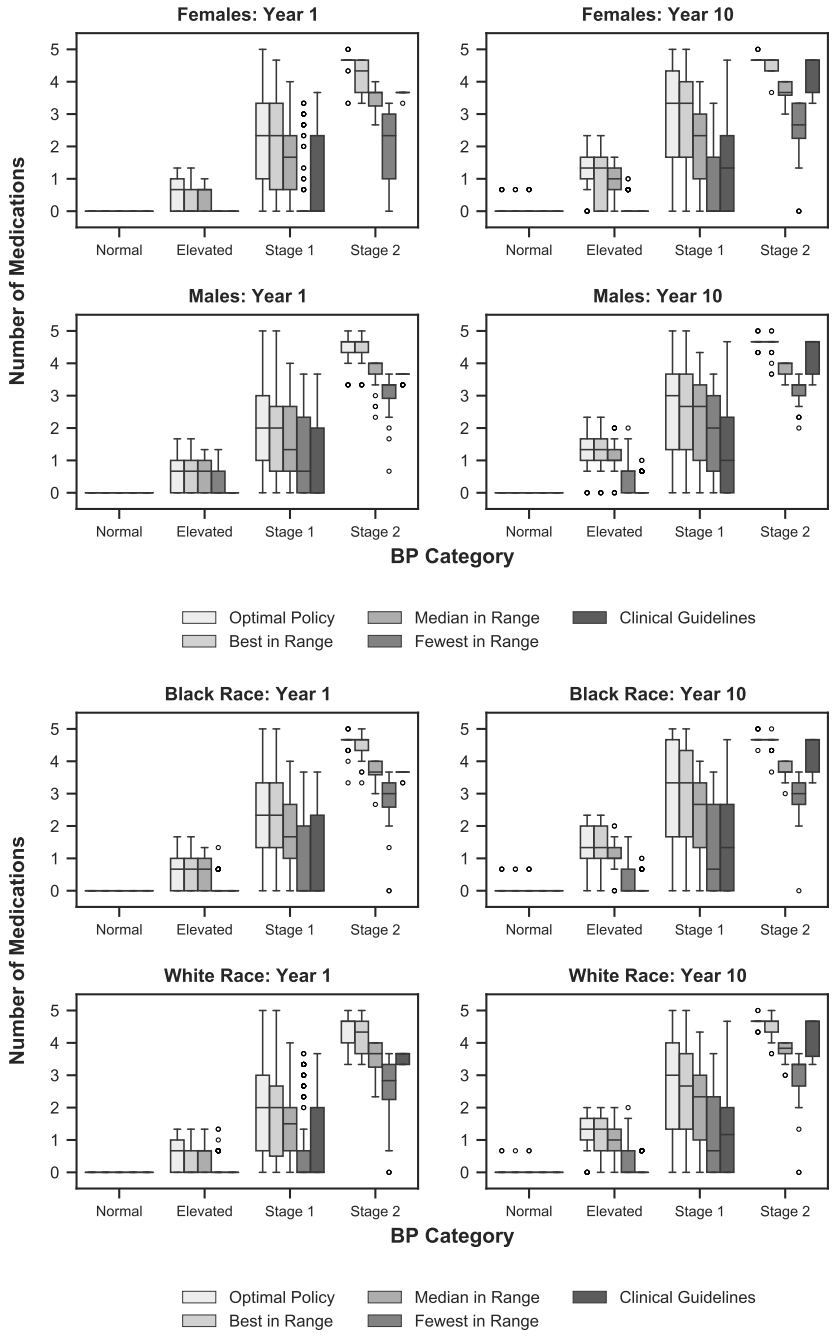

Figure B.3 Distribution of treatment at year 1 and year 10 of the study by sex (top) and race (bottom). BP groups are consistent with the BP categories of the 2017 Hypertension Clinical Practice Guidelines. The label "Elevated" denotes elevated BP, "Stage 1" denotes stage 1 hypertension, and the label "Stage 2" denotes stage 2 hypertension.

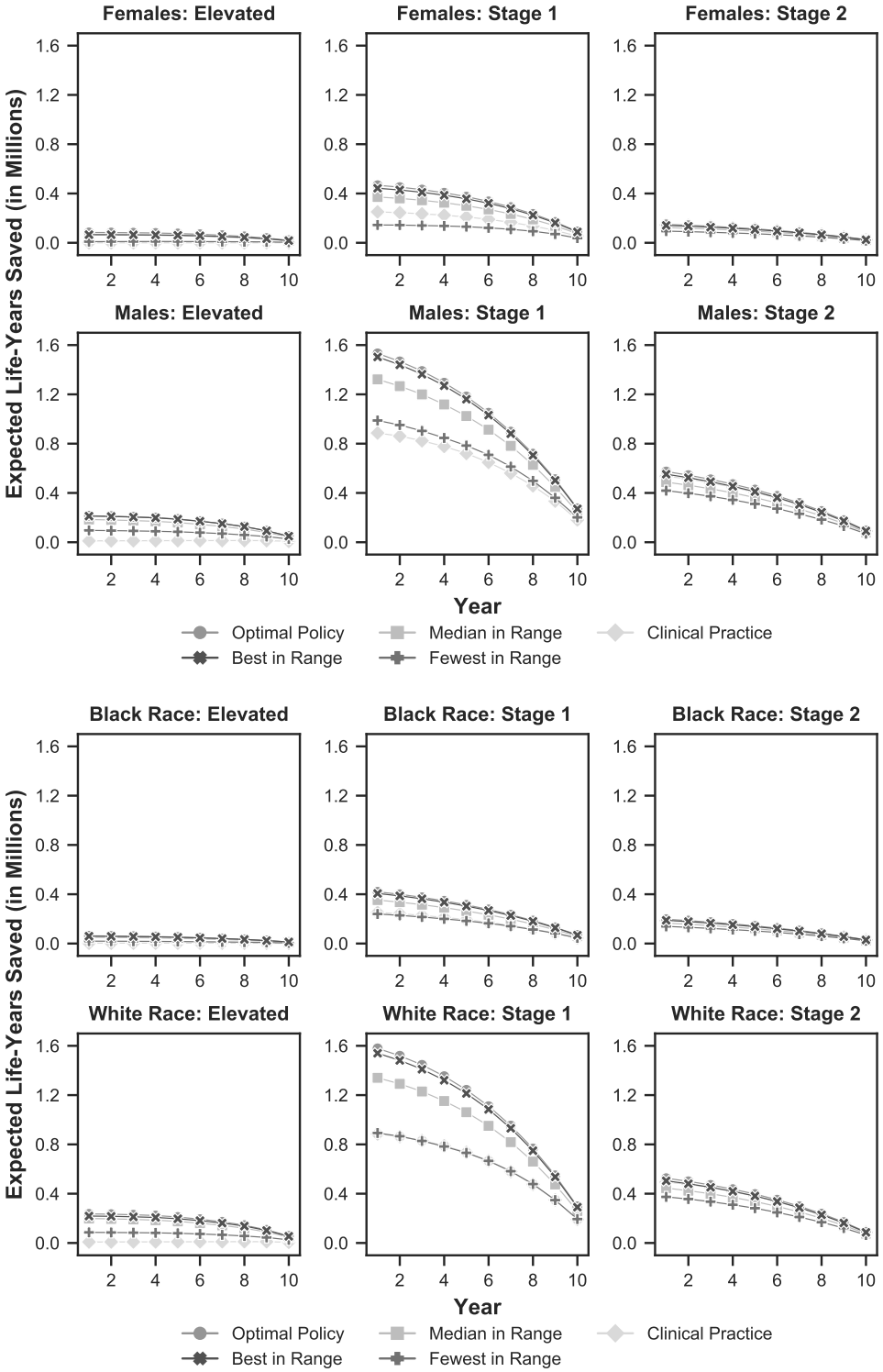

Figure B.4 Life-years saved by each treatment policy compared to no treatment over the planning horizon by sex (top) and race (bottom) per BP group. BP groups are consistent with the BP categories of the 2017 Hypertension Clinical Practice Guidelines. The label "Elevated" denotes elevated BP, "Stage 1" denotes stage 1 hypertension, and the label "Stage 2" denotes stage 2 hypertension.

B.6. Examination of Treatment Choices Contained in the Ranges

The ranges of near-optimal actions always contain the optimal treatment plans in all years, demographics, and BP groups. We also find that the clinical guidelines are contained in the ranges for at least 94.44% of patients with stage 2 hypertension. However, we observe an overall decreasing trend in the proportion of patients treated according to the current clinical guidelines that are included in the ranges in the remaining BP categories. This proportion of patients decrease from 86.48% to 68.44% and from 78.92% to 60.98% over the planning horizon in patients with elevated BP and stage 1 hypertension, respectively. A reason for this may be that the ranges of near-optimal treatment choices are informed by risk, while the current clinical guidelines are mainly driven by BP levels. Also, the current guidelines do not consider the impact of present decisions on patients' future health, while the ranges of α-nonsignificant actions do. The proportion of patients for whom the ranges cover the actions recommended by current clinical guidelines over the 10-year planning horizon stratified by sex, race, and BP category is shown in Figure B.5.

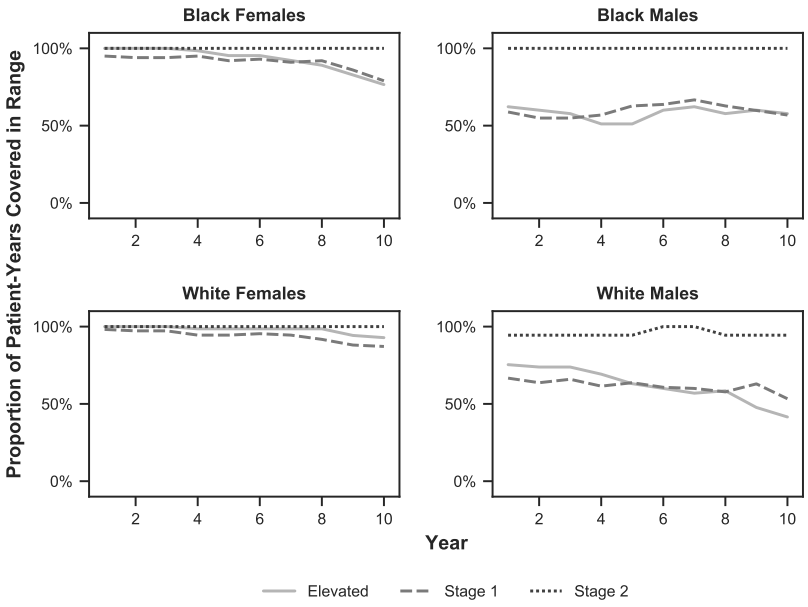

Figure B.5 Proportion of patients of treatment recommendations made by clinical guidelines contained in the ranges of near-optimal actions. The label "Elevated" denotes elevated BP, "Stage 1" denotes stage 1 hypertension, and the label "Stage 2" denotes stage 2 hypertension.

B.7. Sensitivity Analyses

The sensitivity analyses on the model parameters are described in Table B.1. These parameters and their sensitivity analysis values are selected based on communications with our clinical collaborators and the existing literature (Sussman et al. 2013).

We first consider the case that the ASCVD risk reductions obtained from the Blood Pressure Lowering Treatment Trialists' Collaboration (BPLTTC) (Sundström et al. 2014) and the treatment-related disutility are halved or doubled. We also examine a scenario where the treatment-related disutility results in an equal

Table B.1 Sensitivity analysis parameter values.

Parameter	Base case (sensitivity analysis values)
ASCVD risk reductions	BPLTTC (half, double)
Half dosage disutility	0.001 (0.0005, 0.0020, 0.0092)
Standard dosage disutility	0.002 (0.0010, 0.0040, 0.0184)
Action-value function distribution	Empirical (Gaussian)
Future action	Best in range (fewest in range, median in range)
Population	Ages 50-54 (ages 70-74)
Parameter misestimation	None ($\pm$ 50% estimated risk, 50% nonadherence)

number of medications recommended by the optimal treatment strategy and the current clinical guidelines. In this scenario, we use a disutility of 0.0092 for medications at a half dosage and 0.0184 for medications at standard dosage. Second, we perform a sensitivity analysis on the distribution of the action-value functions. Rather than using an empirical estimation of their true distribution, we assume that the action-value functions, including terminal rewards, are normally distributed. In another scenario, we use the treatment choices with the least amount of medications and median number of medications in the next year's range of near-optimal actions, instead of the best treatment choice in the range. We also compare the performance of treatment choices in the ranges of near-optimal actions to the optimal treatment plans and the current clinical guidelines in a secondary population. Each policy is applied in a sample representative of all Black or White adults in the US with ages between 70 and 74 years old (7.55 million people). Finally, we study the case the parameters are misestimated. We contemplate three misestimation scenarios: patients' true risk is half the estimated risk, patients' actual risk is double the estimated risk, and patients' true benefit from treatment is half the estimated benefit.

B.7.1. Results of Sensitivity Analyses. We proceed to study how the treatment strategies are affected by changing the model parameters and assumptions. The results of the sensitivity analysis in the first year of our study are summarized in Table B.2.

If the benefit from treatment is halved, all policies save fewer life years than in the base case. We also note that the ranges contain more treatment choices and that treatment is more aggressive than in the base case. The opposite effect is observed if the benefit from treatment is doubled. In this scenario, fewer medications are necessary to ensure patients' well-being, and fewer treatment choices are within 0.02 life years of the best action, which results in narrower ranges.

We notice that the treatment-related disutility considerably affects the life-years saved by each policy, but the treatment strategies themselves to a lesser extent. If the treatment-related disutility is increased until the optimal treatment policy recommends the same number of medications as the clinical guidelines, we observe a dramatic reduction in the life-years saved by each strategy, the number of medications covered in ranges, and the width of the ranges. In this scenario, the optimal, best in range, median in range, and fewest in range strategies tend to recommend less aggressive treatment, which results in a lower number of life-years saved. Although the current clinical guidelines do not use disutility as a driver for recommending

Table B.2 Summary of sensitivity analyses at the first year of our study.

Sensitivity analysis scenario	Optimal treatment	Life-years saved[a] Best in range	Median in range	Fewest in range	Clinical guidelines	Number of medications[b]	Range width[b]
Base case	3.02	2.92	2.55	1.75	1.83	1.93 (0, 4.33)	3.07 (1, 9)
Treatment benefit							
Halved	2.11	2.02	1.67	1.05	0.99	2.44 (0, 4.67)	5.09 (1, 15)
Doubled	3.44	3.38	3.17	2.47	2.94	1.31 (0, 4.33)	1.81 (1, 4)
Treatment-related disutility							
Halved	3.18	3.07	2.68	1.83	1.89	1.93 (0, 4.33)	3.03 (1, 9)
Doubled	2.73	2.62	2.30	1.58	1.70	1.93 (0, 4.33)	3.21 (1, 10)
Equal treatment	1.29	1.21	0.78	0.71	0.79	1.6 (0, 4)	3.37 (1, 10)

[a] The life-years saved by each policy are presented in millions.

[b] The value outside the parenthesis is the average, the values within the parenthesis are the 5^{th} and 95^{th} quantile across the population of adults in the US with ages between 50 and 54.

treatment, this strategy also results in fewer life-years saved when evaluated in the Markov chain embedded in the MDP.

Assuming that the immediate rewards and terminal rewards are normally distributed does not substantially affect the width of the ranges. Overall, we find that the quantile values $d_t(s, \alpha)$ obtained using the parametric method developed by Dunnett (1955) is reasonably robust to the type of non-normality exhibited in the action-value functions associated with each patient's state and treatment recommendation. This finding is consistent with previous studies on the robustness of Dunnett's method (Westfall 2011). In line with Proposition 10, we also notice that the width of the ranges and the approximately optimal actions do not change with the treatment choice at the next decision epoch if Assumptions 1, 2, 4, and 5 are satisfied. The average range width and number of medications for our base case, assuming normality in the action-value functions, using the action that corresponds to the median number of medications in next year's range, and using the action to corresponds to the fewest number of medications in next year's range are included in Figure B.6.

Applying each treatment strategy to the adult population in the US with ages between 70 and 74 years old, we can draw similar conclusions than with our base case population (adults with ages from 50 to 54). In this population, the treatment strategies contained in the ranges of near-optimal actions save more life years than the current clinical guidelines in every BP category and demographic. This may be because older patients tend to have a higher risk for ASCVD events than younger patients, which translates to more intense treatment by the policies contained in the ranges. We also note that the best treatment in the ranges results in similar health outcomes to the optimal treatment plans. The life-years saved by each policy segregated by sex and BP category as well as by race and BP category are included as Figure B.7.

Figure B.8 shows the proportion of patients whose treatment is covered by the ranges of near-optimal actions despite parameter misestimation. We notice that the ranges of treatment choices are generally robust against event rate misestimation. The largest difference between the proportion of patients whose treatment is covered in the ranges in the base case and the event rate misestimation scenarios is 4.58%. Furthermore, we

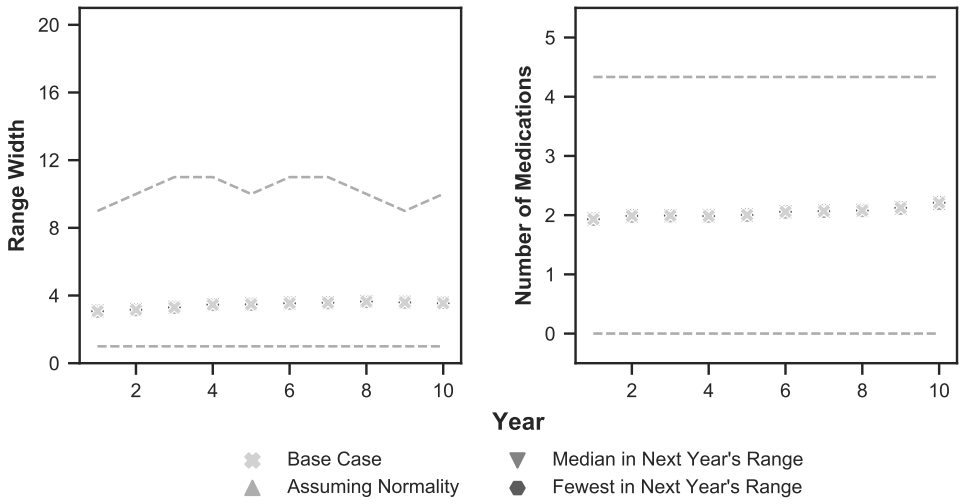

Figure B.6 Average range width and number of medications in the base case, assuming normality in the action-value functions, using the action that corresponds to the median number of medications in next year's range, and using the action to corresponds to the fewest number of medications in next year's range. The label "'Median in Next Year's Range" denotes the median number of medications in next year's range, and the label "Fewest in Next Year's Range" denotes the fewest number of medications in next year's range. Dashed lines represent the 5th and 95th quantile across the population of adults in the US with ages between 50 and 54.

find that the optimal policies are always contained in the ranges of near-optimal actions. While the proportion of patients whose treatment is covered in the ranges remained unchanged in the clinical guidelines treatment strategy in the treatment benefit misestimation scenario, this proportion drops by up to 53.73% in the optimal treatment plans. A potential explanation for this decrease in coverage is that the optimal treatment strategy treats almost twice as aggressively if the true benefit from treatment is half of the misestimated benefit.

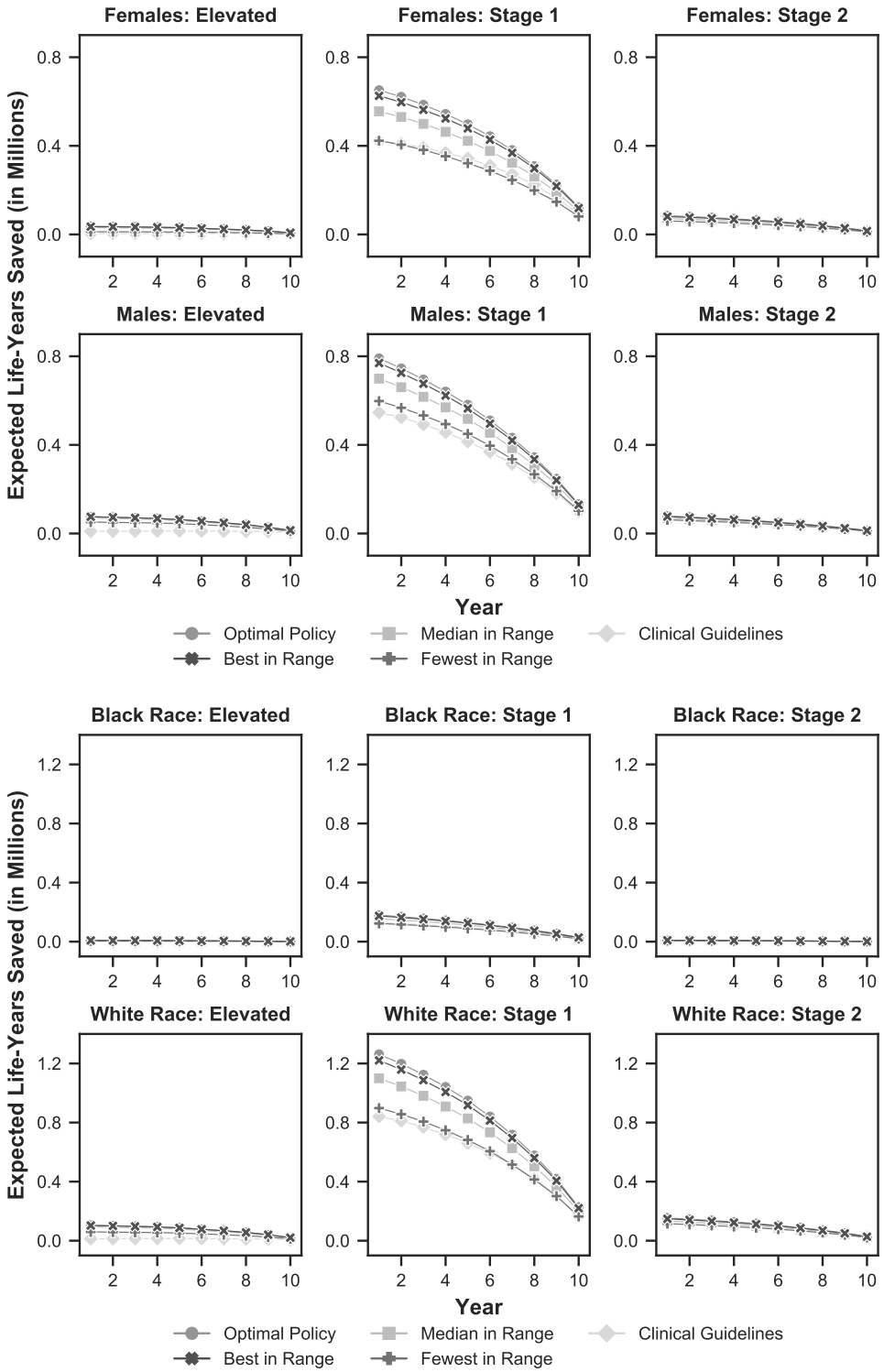

Figure B.7 Life-years saved by each treatment policy compared to no treatment over the planning horizon by sex (top) and race (bottom) per BP group in secondary population of adults with ages between 70 and 74. BP groups are consistent with the BP categories of the 2017 Hypertension Clinical Practice Guidelines. The label "Elevated" denotes elevated BP, "Stage 1" denotes stage 1 hypertension, and the label "Stage 2" denotes stage 2 hypertension.

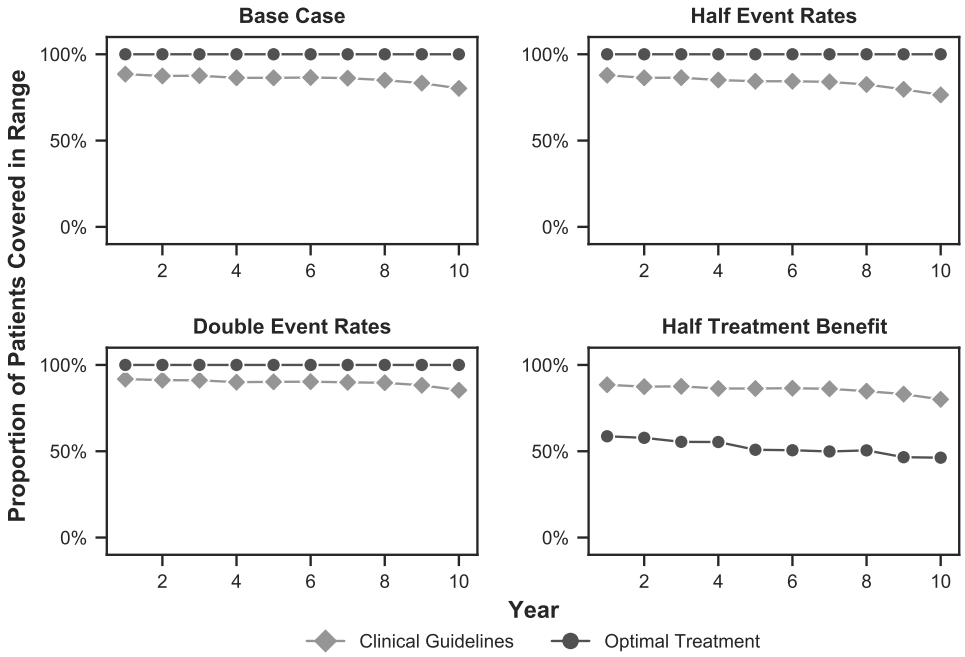

Figure B.8 **Proportion of patients whose treatment recommendations are contained in the ranges of near-optimal actions despite parameter misestimation. Each panel represents a different misestimation scenario: no misestimation (top left), patients' true risk for ASCVD events is half the estimated risk (top right), patients' true risk for ASCVD events is double the estimated risk (bottom left), and patients' true benefit from treatment is half the estimated benefit (bottom right).**