Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

SENSITIVITY ANALYSIS FOR STOPPING CRITERIA WITH APPLICATION TO ORGAN TRANSPLANTATIONS

Xingyu Ren

Department of Electrical and Computer Engineering
Institute for System Research
University of Maryland
College Park, MD 20742, USA

Michael C. Fu

Robert H. Smith School of Business
Institute for System Research
University of Maryland
College Park, MD 20742, USA

Steven I. Marcus

Department of Electrical and Computer Engineering
Institute for System Research
University of Maryland
College Park, MD 20742, USA

ABSTRACT

We consider a stopping problem and its application to the decision-making process regarding the optimal timing of organ transplantation for individual patients. At each decision period, the patient state is inspected and a decision is made whether to transplant. If the organ is transplanted, the process terminates; otherwise, the process continues until a transplant happens or the patient dies. Under suitable conditions, we show that there exists a control limit optimal policy. We propose a smoothed perturbation analysis (SPA) estimator for the gradient of the total expected discounted reward with respect to the control limit. Moreover, we show that the SPA estimator is asymptotically unbiased.

1 INTRODUCTION

This paper is motivated by a kidney transplantation decision-making problem. We consider an end-stage kidney disease (ESKD) patient with a directed living-donor. We assume that the patient is always eligible for transplantation (before they die), and the living-donor organ has a fixed quality and is always available to the patient over the entire decision process. At each decision period, for example, every week or month, the patient health state is inspected and updated, and the decision is whether to transplant depending on the patient health. If the decision is to transplant, the patient receives a *terminal* post-transplantation reward summarizing all the short-term and long-term effect of the transplantation, and the process terminates; otherwise, the patient receives a *intermediate* pre-transplantation reward, and the process continues until a transplantation happens or the patient dies. The goal is to find a policy to maximize the total discounted expected reward. Commonly-used rewards include total discounted expected life years or total discounted quality-adjusted life years (QALYs) (Prieto and Sacristán 2003).

We propose a Markov decision process (MDP) model (Bertsekas 2020) to study this problem, which falls into a special class of MDP models called optimal stopping problems. This type of MDP model has been applied to both liver transplantation (Alagoz et al. 2004; Alagoz et al. 2007b; Alagoz et al. 2007a; Alagoz et al. 2010; Kaufman et al. 2017; Batun et al. 2018) and kidney transplantation (David and Yechiali

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

1985; Bendersky and David 2016; Fan et al. 2020; Ren et al. 2022). Previous work focuses on proving the existence of control limit-type optimal policies. Then, solving MDP problems could be translated into finding an optimal partition of the state space, where each region in the partition is assigned an action; in the scalar-state case, instead of a partition, there is just a single threshold or control limit. Control limit-type policies are important even if they are suboptimal, because they are easy to implement. In general, however, finding an analytic expression for the optimal control limit is difficult. Dynamic programming, one of the most powerful methodologies to solve MDP problems, suffers from the "curse of dimensionality" when the state space or action space is large or even uncountable. In this case, gradient-based optimization methods offer an alternative approach. Moreover, previous work, such as Ren et al. (2022), focuses on the setting of finite state space, while gradient-based methods can be used to solve problems of continuous state spaces. To apply gradient-based optimization methods, one has to compute the gradient of the total expected discounted reward with respect to (w.r.t.) the control limit, where finding an analytic solution is also hard.

In this paper, we focus on estimating the gradient of the total expected discounted reward with respect to the control limit through smoothed perturbation analysis (SPA), a simulation-based method. This is the initial phase of optimization, and fully solving the entire optimization problem will be the focus of future research. The rest of the paper is organized as follows. Section 2 formulates the individual patient organ transplantation problem as a discrete-time, infinite-horizon, *continuous* state space MDP. In Section 3, under suitable conditions, we show the existence of the control limit optimal policy. In Section 4, we propose an SPA estimator (Fu and Hu 1997) for the gradient of the total expected discounted reward w.r.t. the control limit. Moreover, we show that the SPA estimator is asymptotically unbiased. Section 5 reports simulation results illustrating the effectiveness of the SPA estimator. The last section offers conclusion and future research directions.

2 PROBLEM SETTING

In this section, we introduce components of the MDP model. The set of *decision periods* is the natural numbers $\mathbb{N} = \{0, 1, 2, \cdots\}$.

Denote the health state of the patient by $h_n \in S_H := [0, H]$, where a larger value implies worse health. We use an interval $[H_D, H]$, $H_D \in (0, H)$, to represent the death of the patient, i.e., if the patient state h_n is greater than H_D, the patient is deceased. (Representing death as an interval rather than a singleton enables the stochastic kernel for the patient state Markov chain to be expressed using density functions, enhancing clarity in deriving structural results.)

Denote the *post-transplantation state* by P. The MDP will transition into the absorbing state P if the transplantation happens. The *state space* of the MDP is $\mathscr{S} = S_H \bigcup \{P\}$, i.e., at each period n, the state of the MDP s_n is either a scalar patient state h_n, or the post-transplantation state P.

Denote the *action* by a_n. For each $n \in \mathbb{N}$, $a_n \in \mathscr{A} = \{W, T\}$ where $\mathscr{A}$ is the action space including

- W: *wait* for one more period;
- T: accept the kidney for *transplantation*.

The set of state-action pairs $\mathscr{K} := \{(s, a) \mid s \in \mathscr{S},\ a \in \mathscr{A}\}$ consists of four mutually disjoint regions, i.e., $\mathscr{K} = \bigcup_{i=1}^{4} K_i$, where $K_1 = S_H \times \{W\}$, $K_2 = S_H \times \{T\}$, $K_3 = (P, W)$, $K_4 = (P, T)$.

The *Dynamics* of the MDP is defined as follows:

- If action W is taken, the transplant doesn't happen, and the patient will wait until the next decision period. The patient state evolves according to the Markov transition kernel $\mathbb{H}(\cdot|\cdot) : \mathscr{B}(S_H) \times S_H \mapsto [0, 1]$, where $\mathscr{B}(S_H)$ is the collection of Borel subsets of S_H. Specifically, given any current patient state $h_n \in S_H$ and any $B \in \mathscr{B}(S_H)$, at the next period, the patient state h_{n+1} will take a value in B with probability (w.p.) $\int_B \mathbb{H}(dh|h_n)$, where $\mathbb{H}(dh|h_n)$ is a probability measure on measurable space $(S_H, \mathscr{B}(S_H))$.

- The transition kernel $\mathbb{H}$ satisfies the property that once the patient state enters $[H_D, H]$, i.e., the patient dies, they will stay in $[H_D, H]$, and the decision process terminates. In other words, $[H_D, H]$ is an absorbing terminal inverval, i.e., $\int_{H_D}^{H} \mathbb{H}(dh|h_n) = 1,\ \forall h_n \in [H_D, H]$.
- If action T is chosen, the state transitions into the absorbing state P, the decision process terminates.

The general transition kernel $\mathbb{S} : \mathscr{B}(\mathscr{S}) \times \mathscr{K} \mapsto [0,1]$ of the MDP is summarized as follows: for any $n \in \mathbb{N}$, $B \in \mathscr{B}(S_H)$, $h_n \in S_H$,

$$\begin{aligned}
&\mathbb{S}(s_{n+1} = P \mid s_n = P, a_n) = 1,\\
&\mathbb{S}(s_{n+1} \in B \mid s_n = h_n, a_n = W) = \int_B \mathbb{H}(dh|h_n),\\
&\mathbb{S}(s_{n+1} = P \mid s_n = h_n, a_n = T) = 1.
\end{aligned}$$

Reward functions are defined as follows: given patient state $h_n \in S_H$,

- If action W is chosen, an *intermediate pre-transplantation reward* $c(h_n)$ is granted for being alive for one period, where $c(\cdot) : S_H \mapsto \mathbb{R}_+$;
- If action T is chosen, the patient receives a *terminal post-transplantation reward* $r(h_n)$ that evaluates both the short-term and long-term effect of the transplantation, where $r(\cdot) : S_H \mapsto \mathbb{R}_+$.

For $h_n \in [H_D, H]$, i.e., when the patient is deceased, we set $c(h_n) = r(h_n) = 0$. The one-stage reward of the MDP $g(\cdot,\cdot) : S_H \times \mathscr{A} \mapsto \mathbb{R}^+$ is given by

$$g(h,a) = \begin{cases} r(h) & a = T,\\ c(h) & a = W. \end{cases}$$

The *objective* is to find a stationary policy $\pi : S_H \mapsto \mathscr{A}$ maximizing the total discounted expected reward (also known as the value function)

$$V_\pi(h) = \mathbb{E}(\sum_{k=0}^{\infty} \lambda^k g(h_k, \pi(h_k)) | h_0 = h),\ \forall h \in S_H,$$

where $\lambda \in (0,1)$ is a discount factor. We define the maximum of the total discounted expected reward

$$V(h) = \max_{\pi \in \Pi} V_\pi(h),\ \forall h \in S_H,$$

where Π is the set of stationary policies.

3 CONTROL LIMIT POLICY

In this section, we will show the existence of a control limit optimal policy under suitable conditions, which further expand upon the results of Alagoz et al. (2004) to an MDP with a continuous state space. First, we will present several assumptions and preliminary results.

Assumption 1 Both reward functions $c : S_H \mapsto \mathbb{R}_+$ and $r : S_H \mapsto \mathbb{R}_+$ are continuous and nonincreasing.

It follows that for any fixed $a \in \mathscr{A}$, the one-stage reward $g(s,a)$ is nonincreasing and continuous on $\mathscr{S}$. Moreover, g is bounded on $\mathscr{K}$. Because of boundedness of g, for any policy π, its value function $V_\pi(h)$ is bounded on S_H.

Definition 1 (Strong continuity or strong Feller property) Let X, Y by Borel spaces. A stochastic kernel $\mathbb{T} : \mathscr{B}(X) \times Y \mapsto [0,1]$ is said to be strongly continuous if the function $y \mapsto \int v(x)\mathbb{T}(dx|y)$ belongs to $C_b(Y)$, the set of bounded continuous functions on Y, whenever $v \in M_b(X)$, the set of bounded measurable functions on X.

To establish the Bellman optimality condition and value iteration algorithm, we first need to show the strong continuity of the kernel $\mathbb{S}$. We assume the following regularity conditions:

Assumption 2 For any $h \in S_H$, the measure $\mathbb{H}(\cdot|h)$ admits a density $f_{\mathbb{H}}(\cdot|h) : S_H \mapsto \mathbb{R}_+$, satisfying the following conditions:

- For any fixed $h' \in S_H$, $f_{\mathbb{H}}(h'|h)$ is continuous in h.
- $f_{\mathbb{H}}$ is uniformly bounded, i.e., there exists $M > 0$ such that for any $(h', h) \in [0, H]^2$, $f_{\mathbb{H}}(h'|h) < M$.

Lemma 1 The transition kernel $\mathbb{S}$ is strongly continuous, i.e., for every $u \in M_b(\mathscr{S})$, $v(s,a) = \int_S u(s')\mathbb{S}(ds'|s,a)$ is continuous and bounded on $\mathscr{K}$.

Proof. It is enough to show that $\mathbb{S}$ is strongly continuous on $K_1 = [0, H] \times \{W\}$, because $\mathbb{S}(\cdot|s,a)$ is a probability mass on some single absorbing state when $(s,a) \in K_i$, $i = 2,3,4$. Since $a = W$ when $(s,a) \in K_1$, we will drop dependence on a in $v(s,a)$ for simplicity. It suffices to show that $v(h) = \int_S u(s')\mathbb{S}(ds'|s = h, a = W)$ is continuous on $[0, H]$ for any $u \in M_b(\mathscr{S})$. We can write

$$v(h) = \int_0^H u(h')\mathbb{H}(dh'|h). \tag{1}$$

v is bounded, since the measure $\mathbb{H}(dh'|h)$ is finite. Then, by Assumption 2, (1) can be rewritten as

$$\int_0^H u(h')\mathbb{H}(dh'|h) = \int_0^H u(h') f_{\mathbb{H}}(h'|h)dh'.$$

Take any sequence $\{h_n\}$ such that $h_n \to h$ as $n \to \infty$. Since $f_{\mathbb{H}}(h'|h)$ is continuous in h for any $h' \in [0, H]$, $u(h')f_{\mathbb{H}}(h'|h_n) \to u(h')f_{\mathbb{H}}(h'|h)$ as $n \to \infty$. Since both u and $f_{\mathbb{H}}$ are bounded, by the dominated convergence theorem,

$$\int_0^H u(h') f_{\mathbb{H}}(h'|h_n)dh' \to \int_0^H u(h') f_{\mathbb{H}}(h'|h)dh' \text{ as } n \to \infty.$$

Therefore, $v(h)$ is continuous in h. □

Then, by Theorem 4.2.3 and Lemma 4.2.8 in Hernández-Lerma and Lasserre (1996), we can establish the Bellman's optimality condition and value iteration algorithm.

Theorem 1 The optimal value function $V(h)$ is the solution of the optimality equation:

$$V(h) = \min\{r(h), c(h) + \lambda \int_0^H V(h')\mathbb{H}(dh'|h)\}, \forall h. \tag{2}$$

Moreover, the sequence $\{V_k\}$ generated by value iteration

$$V_k(h) = \min\{r(h), c(h) + \lambda \int_0^H V_{k-1}(h')\mathbb{H}(dh'|h)\},$$
$$V_0(h) = 0, \forall h,$$

is a monotonically nondecreasing sequence, i.e., $V_n(h) \le V_{n+1}(h),\ \forall h \in S_H, k \in \mathbb{N}$ and converges pointwise to V, i.e., $V_n \nearrow V^*$.

In reliability theory, the increasing failure rate (IFR) property of a probability distribution is a widely-used concept to depict the deterioration of a system (Ross 1996). For a distribution function F with a density or mass function f, we say that F is IFR if its failure rate function defined by $f(t)/\bar{F}(t)$, where $\bar{F} := 1 - F$ is the complementary (or tail) distribution function, is nondecreasing as a function of t. For our purposes, we define the IFR property for the transition kernel of a Markov chain.

Definition 2 Let $\mathbb{T}$ be a stochastic kernel on $(\mathscr{B}[0,X],[0,X])$. We say that $\mathbb{T}$ has the IFR property if for every $x_0 \in [0,X]$, $b(x) := \int_{x_0}^{X} \mathbb{T}(dx'|x)$ is nondecreasing in x.

Assumption 3 The transition kernel $\mathbb{H}$ has the IFR property.

Assumption 3 has the intuitive explanation that as the patient's health deteriorates, the likelihood of further deterioration increases. The following lemma in Douer and Yechiali (1994) provides a necessary and sufficient condition for Definition 2.

Lemma 2 The stochastic kernel $\mathbb{T}$ on space $(\mathscr{B}[0,X],[0,X])$ is IFR if and only if for any bounded, nonnegative and nondecreasing function $v : [0,X] \mapsto \mathbb{R}^+$, $l(x) = \int_0^X v(x)\mathbb{T}(dx'|x)$ is also nondecreasing.

The monotinicity of the value function V can be easily shown from (2) in Theorem 1 and Lemma 2.

Theorem 2 Under Assumptions 1 through 3, the value function V is nonincreasing on S_H.

Theorem 2 implies that the patient's overall benefit, e.g., the total QALYs, will not increase if the patient health deteriorates.

Theorem 3 provides sufficient conditions for the existence of a control limit optimal policy. Specifically, Theorem 3 shows that there exists an optimal policy π^* that partitions S_H into two intervals:

$$\pi^*(h) = \begin{cases} W & \text{if } h < \theta^*, \\ T & \text{if } h \geq \theta^*, \end{cases} \tag{3}$$

where θ^* is called the optimal *control limit* (or threshold). The optimal action to take depends only on whether the state h is greater than or less than the control limit θ^*, and solving the MDP problem boils down to finding this optimal threshold. To prove Theorem 3, we need the following lemma (Alagoz et al. 2007b) and several additional assumptions.

Lemma 3 Let v be a bounded, nonnegative, and nonincreasing function. If the stochastic kernel $\mathbb{T}$ defined on $(\mathscr{B}[0,X],[0,X])$ is IFR and admits a uniformly bounded density function $f : [0,X]^2 \mapsto \mathbb{R}_+$, then the following results hold: for any $x_1 < x_2$,

- $\int_0^{x_1} v(x)(f(x|x_1) - f(x|x_2))dx \geq v(x_1) \int_0^{x_1} (f(x|x_1) - f(x|x_2))dx$;
- $\int_{x_1}^{X} v(x)(f(x|x_1) - f(x|x_2))dx \geq v(x_1) \int_{x_1}^{X} (f(x|x_1) - f(x|x_2))dx$.

Proof. We only provide a proof for the first part, as the second part can be proved in a similar way. First, we consider the case that v is simple function, i.e., $v(x) = \sum_{i=1}^{n} v_i \mathbf{1}_{A_i}(x)$ for some n, where $\mathbf{1}_{A_i}$ is the indicator function of set A_i and $\{A_i\}_{i=1,\cdots,n}$ is a partition of $[0,x_1]$. v nonincreasing allows us to take each A_i to be an interval. Assume that $v_1 \geq v_2 \geq \cdots \geq v_n$. Then,

$$\begin{aligned}
\int_0^{x_1} v(x)(f(x|x_1) - f(x|x_2))dx &= \sum_{i=1}^{n} v_i \int_{A_i} (f(x|x_1) - f(x|x_2))dx \\
&= v_1 \int_{A_1} (f(x|x_1) - f(x|x_2))dx + \sum_{i=2}^{n} v_i \int_{A_i} (f(x|x_1) - f(x|x_2))dx \\
&\geq v_2 \int_{A_1} (f(x|x_1) - f(x|x_2))dx + \sum_{i=2}^{n} v_i \int_{A_i} (f(x|x_1) - f(x|x_2))dx \\
&\geq v_n \int_0^{x_1} (f(x|x_1) - f(x|x_2))dx \\
&= v(x_1) \int_0^{x_1} (f(x|x_1) - f(x|x_2))dx.
\end{aligned}$$

For any general nonincreasing function v, we can take a monotone sequence of simple functions $\{v_n\}$ such that $v_n \nearrow v$ pointwise. Since f is uniformly bounded, the result follows from the dominated convergence theorem. □

Assumption 4 For any $h_1 < h_2$ and h_0,

$$\int_{h_0}^{H_D} f_{\mathbb{H}}(h|h_1)dh \leq \int_{h_0}^{H_D} f_{\mathbb{H}}(h|h_2)dh.$$

Assumption 4 takes a similar form as the IFR property and can be interpreted similarly, but Assumption 4 is neither a sufficient nor a necessary condition for the IFR property of $\mathbb{H}$.

Assumption 5 For any $h_1 < h_2$,

$$\frac{r(h_1) - r(h_2)}{r(h_2)} \leq \lambda \left(\int_{H_D}^{H} f_{\mathbb{H}}(h|h_2)dh - \int_{H_D}^{H} f_{\mathbb{H}}(h|h_1)dh \right).$$

Assumption 5 has an intuitive explanation that, as the patient health becomes worse, the increment of the probability of death during waiting is greater than the *marginal* reduction in the transplantation reward. Alagoz et al. (2004) presents empirical evidence that Assumptions 4 and 5 are applicable in the context of living-donor liver transplantation.

Theorem 3 Under Assumptions 1 through 5, there exists a control limit optimal policy taking the form of (3).

Proof. By contradiction, suppose that for some $h_1 < h_2$, $T \in a^*(h_1)$, but $a^*(h_2) = W$, where $a^*(h)$ is the set of optimal actions at h. Then, we have

$$\begin{aligned} r(h_1) &\geq c(h_1) + \lambda \int_0^{H_D} V(h') f_{\mathbb{H}}(h'|h_1)dh', \\ r(h_2) &< c(h_2) + \lambda \int_0^{H_D} V(h') f_{\mathbb{H}}(h'|h_2)dh'. \end{aligned}$$

Then,

$$\begin{aligned} r(h_1) - r(h_2) &> c(h_1) - c(h_2) + \lambda \int_0^{H_D} V(h')(f_{\mathbb{H}}(h'|h_1) - f_{\mathbb{H}}(h'|h_2))dh' \\ &\geq \lambda \int_0^{h_1} V(h')(f_{\mathbb{H}}(h'|h_1) - f_{\mathbb{H}}(h'|h_2))dh' + \lambda \int_{h_1}^{H_D} V(h')(f_{\mathbb{H}}(h'|h_1) - f_{\mathbb{H}}(h'|h_2))dh' \\ &\geq \lambda V(h_1) \int_0^{H_D} (f_{\mathbb{H}}(h'|h_1) - f_{\mathbb{H}}(h'|h_2))dh' \\ &= \lambda V(h_1) \left(\int_{H_D}^{H} f_{\mathbb{H}}(h|h_2)dh - \int_{H_D}^{H} f_{\mathbb{H}}(h|h_1)dh \right), \end{aligned}$$

where the second inequality follows from Assumption 1, and the third inequality follows from Lemma 3. By Assumption 5, we have $V(h_1) < r(h_2)$. Since $a^*(h_2) = W$,

$$r(h_2) < V(h_2) \leq V(h_1) \leq V(h_1),$$

which is a contradiction. Therefore, for any $h_1 < h_2$, $T \in a^*(h_1)$ implies that $T \in a^*(h_2)$. □

Theorem 3 has an intuitive explanation that the patient should be transplanted if and only if their health status is worse than some threshold (recall that a larger patient state h_n implies the worse health status).

4 SMOOTHED PERTURBATION ANALYSIS (SPA) ESTIMATOR

In this section, we propose an SPA estimator for the gradient of the value function w.r.t. the control limit. Throughout this section, we suppose that a control limit policy with control limit θ, denoted by π_θ, is implemented. For a fixed initial condition $h_0 \in S_H$, let $V(\theta)$ be the value function associated with π_θ and assume that $V(\theta)$ is differentiable w.r.t. θ. We want to estimate the gradient of $V(\theta)$ w.r.t. θ.

We denote by $h_k(\theta)$, $\forall k$ the patient state at period k, under control limit policy π_θ. For fixed $n \in \mathbb{N}$ and $h_0 \in S_H$, we consider the following sample performance

$$v_n(\theta) = \sum_{k=0}^{n} \lambda^k g(h_k, \pi_\theta(h_k)),$$

i.e., the total discounted reward until period n under policy π_θ. Notice that

$$g(h, \pi_\theta(h)) = \begin{cases} r(h) & h \geq \theta, \\ c(h) & h < \theta. \end{cases}$$

Given a nominal sample path under policy π_θ, suppose that a perturbation of $\Delta\theta$ is introduced to construct a sample path under policy $\pi_{\theta+\Delta\theta}$, called the perturbed sample path. An infinitesimal perturbation analysis (IPA) estimator comes from taking the derivative of each $g(h_k, \pi_\theta(h_k))$ while assuming that the event $\{h_n \neq P,\ h_n < \theta\}$ is unchanged, i.e., the perturbation $\Delta\theta$ results in no change in the transplant decision (thus there is also no change in the sample path). Under this assumption,

$$\frac{dh_k(\theta)}{d\theta} = 0 \text{ w.p. } 1,$$

and

$$\begin{aligned} \frac{dg(h_k, \pi_\theta(h_k))}{d\theta} &= \frac{\partial g(h_k, \pi_\theta(h_k))}{\partial \theta} + \frac{\partial g(h_k, \pi_\theta(h_k))}{\partial h} \frac{dh_k(\theta)}{d\theta} \\ &= \frac{\partial g(h_k, \pi_\theta(h_k))}{\partial \theta} \\ &= 0 \text{ w.p. } 1, \end{aligned}$$

because for fixed h, $g(h, \pi_\theta(h))$ is a function of θ with only one discontinuity of zero measure.

However, the IPA estimator does not capture the discrete changes that occur, for example, when the action in some period changes from "transplant" in the nominal sample path to "wait" in the perturbed sample path. We use SPA to calculate discrete changes caused by the change of the action. Specifically, by conditioning on suitable quantities, we compute the conditional expectation on the change in v_n, and take $\Delta\theta \to 0$. In this MDP model, discrete changes may potentially occur only at $M(n) = \min\{i \leq n : h_i \geq \theta\}$, i.e., the period when a transplantation happens in the nominal sample path. Conditioned on $M(n)$, the event $\{h_{M(n)} \geq \theta\}$ is equivalent to $\{\alpha_n \geq 0\}$, where $\alpha_n := h_{M(n)} - \theta$. The action in the perturbed sample path alters if $\alpha_n < \Delta\theta$. Conditioned on $h_{M(n)-1}$, the SPA estimator is expressed as

$$\begin{aligned} \left(\frac{\partial v_n(\theta)}{\partial \theta}\right)_{SPA} &= \lim_{\Delta\theta \to 0} \mathbb{E}(\Delta v_n(\theta) \mathbf{1}\{\alpha_n \leq \Delta\theta\} | h_{M(n)-1}) / \Delta\theta \\ &= \lim_{\Delta\theta \to 0} \mathbb{E}(\Delta v_n(\theta) | \alpha_n \leq \Delta\theta | h_{M(n)-1}) \mathbb{P}(\alpha_n \leq \Delta\theta | h_{M(n)-1}) / \Delta\theta, \end{aligned}$$

where

$$\mathbb{P}(\alpha_n \leq \Delta\theta | h_{M(n)-1}) = \mathbb{P}(\alpha_n \leq \Delta\theta | \alpha_n \geq 0, h_{M(n)-1})$$

$$= \mathbb{P}(\theta \le h_{M(n)} \le \theta + \Delta\theta | h_{M(n)-1}) / \mathbb{P}(h_{M(n)} \ge \theta | h_{M(n)-1})$$
$$= \frac{\int_\theta^{\theta+\Delta\theta} \mathbb{H}(dh|h_{M(n)-1})}{\int_\theta^H \mathbb{H}(dh|h_{M(n)-1})}.$$

Therefore,

$$\lim_{\Delta\theta\to 0} \mathbb{P}(\alpha_n \le \Delta\theta | h_{M(n)-1})/\Delta\theta = \frac{f_{\mathbb{H}}(\theta|h_{M(n)-1})}{\int_\theta^H \mathbb{H}(dh|h_{M(n)-1})}.$$

The term $\lim_{\Delta\theta\to 0}\mathbb{E}(\Delta v_n(\theta)|\alpha_n \le \Delta\theta, h_{M(n)-1})$ is the extra accumulated reward caused by the change of the action, which is given by

$$\lim_{\Delta\theta\to 0} \mathbb{E}(\Delta v_n(\theta)|\alpha_n \le \Delta\theta, h_{M(n)-1}) = \lambda^{M(n)}(r(\theta)-c(\theta)) - \mathbb{E}\left(\sum_{i=M(n)+1}^{n} \lambda^i g(h_i, \pi_\theta(h_i))|h_{M(n)} = \theta^-\right).$$

The final SPA gradient estimator is given by

$$\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA} = \frac{f_{\mathbb{H}}(\theta|h_{M(n)-1})}{\int_\theta^H \mathbb{H}(dh|h_{M(n)-1})}\left(\lambda^{M(n)}(r(\theta)-c(\theta)) - \mathbb{E}\left(\sum_{i=M(n)+1}^{n} \lambda^i g(h_i,\pi_\theta(h_i))|h_{M(n)}=\theta^-\right)\right).$$

Now we formally show that $\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA}$ is an asymptotically unbiased estimator, i.e.,

$$\mathbb{E}\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA} \to \frac{\partial V(\theta)}{\partial\theta} \text{ as } n\to\infty.$$

First, we show that $\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA}$ is an unbiased estimator of $\frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}$.

Theorem 4 Under Assumptions 1 and 2, $\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA}$ is an unbiased estimator of $\frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}$, i.e.,

$$\mathbb{E}\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA} = \frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}.$$

Proof. We define the following events: for fixed $\Delta\theta$,

$$A_k = \{h_i(\theta) < \theta \text{ or } h_i(\theta) \ge \theta+\Delta\theta,\ i=1,\cdots,k\},\ \forall k,$$
$$B_k = A_k^{\sim},$$

i.e., A_k is the event that a perturbation of size $\Delta\theta$ doesn't cause a change in the transplant decision until time k. Then, we can write

$$\frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta} = \lim_{\Delta\theta\to 0}\left(\frac{\mathbb{E}((v_n(\theta+\Delta\theta)-v_n(\theta))\mathbf{1}\{A_n\})}{\Delta\theta} + \frac{\mathbb{E}((v_n(\theta+\Delta\theta)-v_n(\theta))\mathbf{1}\{B_n\})}{\Delta\theta}\right),$$

where the first term is zero, because the perturbed sample path and the nominal sample path are the same, conditioned on the event A_n. We write the term $\mathbb{E}((v_n(\theta+\Delta\theta)-v_n(\theta))\mathbf{1}\{B_n\})$ as

$$\mathbb{E}(\mathbb{E}((v_n(\theta+\Delta\theta)-v_n(\theta))\mathbf{1}\{B_n\}|h_{M(n-1)})).$$

Note that the event B_n is equivalent to the event $\{M(n)$ is non-empty, $h_{M(n)}(\theta) \geq \theta$, $h_{M(n)}(\theta) < \theta + \Delta\theta\}$. Then,

$$\begin{aligned}
\mathbb{E}(v_n(\theta+\Delta\theta)\mathbf{1}\{B_n\}|h_{M(n)-1}) &= \mathbb{E}(v_n(\theta+\Delta\theta)|h_{M(n)-1}, \mathbf{1}\{B_n\})\mathbb{P}(\mathbf{1}\{B_n\}|h_{M(n)-1}) \\
&= \mathbb{E}(v_n(\theta+\Delta\theta)|h_{M(n)-1}, \mathbf{1}\{B_n\}) \\
&\times \mathbb{P}(\theta \leq h_{M(n)} < \theta+\Delta\theta|h_{M(n)-1}) \\
&\leq \frac{\max_h\{c(h), r(h)\}}{1-\lambda} \int_{\theta}^{\theta+\Delta\theta} f_{\mathbb{H}}(h|h_{M(n-1)})dh \\
&\leq \frac{M\Delta\theta \max_h\{c(h), r(h)\}}{1-\lambda},
\end{aligned}$$

where the first inequality follows from the fact that $v_n(\theta) \leq \sum_{i=0}^{\infty}\lambda^i \max_h\{c(h), r(h)\}$, and the second inequality follows from Assumption 2 that $f_{\mathbb{H}}$ is uniformly bounded by M. Therefore, $\mathbb{E}(v_n(\theta+\Delta\theta)\mathbf{1}\{B_n\}|h_{M(n-1)})/\Delta\theta$ is uniformly bounded for any $\Delta\theta$. By the dominated convergence theorem,

$$\begin{aligned}
\lim_{\Delta\theta\to 0}\frac{\mathbb{E}(v_n(\theta+\Delta\theta)\mathbf{1}\{B_n\})}{\Delta\theta} &= \lim_{\Delta\theta\to 0}\mathbb{E}\left(\frac{\mathbb{E}(v_n(\theta+\Delta\theta)\mathbf{1}\{B_n\}|h_{M(n)-1})}{\Delta\theta}\right) \\
&= \mathbb{E}\left(\lim_{\Delta\theta\to 0}\frac{\mathbb{E}(v_n(\theta+\Delta\theta)\mathbf{1}\{B_n\}|h_{M(n)-1})}{\Delta\theta}\right) \\
&= \mathbb{E}\left(\lim_{\Delta\theta\to 0}\frac{\mathbb{P}(\mathbf{1}\{B_n\}|h_{M(n)-1})}{\Delta\theta} \times \lim_{\Delta\theta\to 0}\mathbb{E}(v_n(\theta+\Delta\theta)|\mathbf{1}\{B_n\}, h_{M(n)-1})\right) \\
&= \mathbb{E}\Bigg(\frac{f_{\mathbb{H}}(\theta|h_{M(n)-1})}{\int_{\theta}^{H}\mathbb{H}(dh|h_{M(n)-1})} \\
&\times \left(\sum_{i=0}^{M(n)-1}\lambda^i c(h_i) + \lambda^{M(n)}c(\theta) + \mathbb{E}\left(\sum_{i=M(n)+1}^{n}\lambda^i g(h_i, \pi_\theta(h_i))|h_{M(n)} = \theta^-\right)\right)\Bigg).
\end{aligned}$$

Similarly, we can derive

$$\lim_{\Delta\theta\to 0}\frac{\mathbb{E}(v_n(\theta)\mathbf{1}\{B_n\})}{\Delta\theta} = \mathbb{E}\left(\frac{f_{\mathbb{H}}(\theta|h_{M(n)-1})}{\int_{\theta}^{H}\mathbb{H}(dh|h_{M(n)-1})} \times \left(\sum_{i=0}^{M(n)-1}\lambda^i c(h_i) + \lambda^{M(n)}r(\theta)\right)\right).$$

It follows that

$$\begin{aligned}
\lim_{\Delta\theta\to 0}\frac{\mathbb{E}((v_n(\theta+\Delta\theta)-v_n(\theta))\mathbf{1}\{B_n\})}{\Delta\theta} &= \mathbb{E}\Bigg(\frac{f_{\mathbb{H}}(\theta|h_{M(n)-1})}{\int_{\theta}^{H}\mathbb{H}(dh|h_{M(n)-1})} \\
&\times (\lambda^{M(n)}(c(\theta)-r(\theta)) + \mathbb{E}\left(\sum_{i=M(n)+1}^{n}\lambda^i g(h_i, \pi_\theta(h_i))|h_{M(n)} = \theta^-\right)\Bigg).
\end{aligned}$$

Therefore, we conclude that $\mathbb{E}\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA} = \frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}$. □

Theorem 5 Under Assumptions 1 and 2, $\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA}$ is an asymptotically unbiased estimator, i.e.,

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA} = \frac{\partial V(\theta)}{\partial\theta}.$$

Proof. Note that $\lim_{n\to\infty}\mathbb{E}(v_n(\theta))=\mathbb{E}(\lim_{n\to\infty}v_n(\theta))=V(\theta)$ because the sequence of random variables $\{v_n\}_{n\in\mathbb{N}}$ is uniformly bounded and converges pointwise. Since we already proved $\mathbb{E}\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA}=\frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}$ in Theorem 4, it remains to show that

$$\lim_{n\to\infty}\frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}=\frac{\partial\lim_{n\to\infty}\mathbb{E}(v_n(\theta))}{\partial\theta}.$$

i.e., passing a derivative through a limit. By Theorem 8.2.3 in Bartle and Sherbert (2010), it suffices to show that $\left(\frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}\right)_{n\in\mathbb{N}}$ is a uniformly convergent sequence on S_H. Note that

$$\left|\frac{\partial\mathbb{E}v_n(\theta)}{\partial\theta}-\frac{\partial\mathbb{E}v_{n-1}(\theta)}{\partial\theta}\right|=\lambda^n\left|\frac{\partial\mathbb{E}g(h_n,\pi_\theta(h_n))}{\partial\theta}\right|.$$

It suffices to show that $\left|\frac{\partial\mathbb{E}g(h_n,\pi_\theta(h_n))}{\partial\theta}\right|$ is bounded, which can be shown similarly as in Theorem 4. We write

$$\frac{\partial\mathbb{E}g(h_n,\pi_\theta(h_n))}{\partial\theta}=\lim_{\Delta\theta\to0}\frac{\mathbb{E}g(h_n(\theta+\Delta\theta),\pi_\theta(h_n(\theta+\Delta\theta)))-\mathbb{E}g(h_n(\theta),\pi_\theta(h_n(\theta)))}{\Delta\theta}.$$

Similar to the proof of Theorem 4, we can write $\mathbb{E}g(h_n(\theta),\pi_\theta(h_n(\theta)))$ as

$$\begin{aligned}\mathbb{E}g(h_n(\theta),\pi_\theta(h_n(\theta)))&=\mathbb{E}(\mathbb{E}(g(h_n(\theta),\pi_\theta(h_n(\theta)))\mathbf{1}\{B_n\}|h_{M(n)-1})),\text{ where}\\ \mathbb{E}(g(h_n(\theta),\pi_\theta(h_n(\theta)))\mathbf{1}\{B_n\}|h_{M(n)-1})&=\mathbb{P}(1\{B_n\}|h_{M(n)-1})\mathbb{E}(g(h_n(\theta),\pi_\theta(h_n(\theta)))|\mathbf{1}\{B_n\},h_{M(n)-1})\\ &\le\max_h\{c(h),r(h)\}\int_\theta^{\theta+\Delta\theta}f_{\mathbb{H}}(h|h_{M(n)-1})dh\\ &\le\Delta\theta M\max_h\{c(h),r(h)\}.\end{aligned}$$

Similarly, $\mathbb{E}g(h_n(\theta+\Delta\theta),\pi_\theta(h_n(\theta+\Delta\theta)))\le\Delta\theta M\max_h\{c(h),r(h)\}$. It follows that

$$\left|\frac{\partial\mathbb{E}g(h_n,\pi_\theta(h_n))}{\partial\theta}\right|\le\lim_{\Delta\theta\to0}\frac{\mathbb{E}g(h_n(\theta+\Delta\theta),\pi_\theta(h_n(\theta+\Delta\theta)))+\mathbb{E}g(h_n(\theta),\pi_\theta(h_n(\theta)))}{\Delta\theta}=2M\max_h\{c(h),r(h)\},$$

which is independent of θ. Therefore, $\left|\frac{\partial\mathbb{E}v_n(\theta)}{\partial\theta}-\frac{\partial\mathbb{E}v_{n-1}(\theta)}{\partial\theta}\right|$ converges uniformly to zero, and $\left(\frac{\partial\mathbb{E}(v_n(\theta))}{\partial\theta}\right)_{n\in\mathbb{N}}$ is a uniformly convergent sequence on S_H. Thus we can pass the derivative through the limit. □

5 SIMULATION EXAMPLE

In this section, we present a simple simulation example to demonstrate the performance of the SPA estimator. We consider an MDP where the decision period is half a year, and the patient state h_n takes values in $S_H=[0,1]$. Suppose that the control limit policy π_θ is implemented for some $\theta\in(0,1)$. For any n, conditioned on $h_n\in[0,\theta]$, h_{n+1} is uniformly distributed over $[h_n,1]$, i.e., $f_{\mathbb{H}}(h'|h)=\mathbf{1}\{1\ge h'\ge h\}/(1-h)$, $\forall h,h'\in[0,1]$. Therefore, the patient health never improves, and it is straightforward to check that $f_{\mathbb{H}}$ satisfies the IFR property. The rewards are defined in terms of expected life years. We define the intermediate pre-transplantation reward $c(h)\equiv0.5$ (years), and the terminal post-transplantation reward function $r(h)=8(1-h)$ (years). Then, the SPA estimator is given by

$$\begin{aligned}\left(\frac{\partial v_n(\theta)}{\partial\theta}\right)_{SPA}&=\frac{f_{\mathbb{H}}(\theta|h_{M(n)-1})}{\int_\theta^H\mathbb{H}(dh|h_{M(n)-1})}\left(\lambda^{M(n)}(r(\theta)-c(\theta))-\mathbb{E}\left(\sum_{i=M(n)+1}^{n}\lambda^i g(h_i,\pi_\theta(h_i))|h_{M(n)}=\theta^-\right)\right)\\ &=\frac{1}{1-\theta}(\lambda^{M(n)}(r(\theta)-c(\theta))-\mathbb{E}(\lambda^{M(n)+1}r(h_{M(n)+1})|h_{M(n)}=\theta^-))\\ &=\frac{1}{1-\theta}(\lambda^{M(n)}(8-8\theta-0.5)-\mathbb{E}(\lambda^{M(n)+1}8(1-h_{M(n)+1})|h_{M(n)}=\theta^-)),\end{aligned}$$

Table 1: Simulation results for sensitivity of value function $V(\theta)$ w.r.t. θ (standard errors in parentheses).

N	θ	SPA	FD($\delta=0.01$)	FD($\delta=0.05$)	FD($\delta=0.1$)
10^2	0.2	$-3.199(0.242)$	$-8.251(5.394)$	$-4.786(1.811)$	$-2.065(0.871)$
	0.5	$-2.668(0.233)$	$-5.523(3.182)$	$-2.560(1.076)$	$-3.512(0.780)$
	0.8	$-1.313(0.225)$	$-3.083(1.411)$	$-1.326(0.552)$	$-0.205(0.242)$
10^4	0.2	$-3.371(0.023)$	$-3.281(0.349)$	$-3.503(0.155)$	$-3.253(0.104)$
	0.5	$-2.997(0.023)$	$-3.120(0.265)$	$-2.991(0.109)$	$-2.920(0.071)$
	0.8	$-1.515(0.022)$	$-1.306(0.114)$	$-1.144(0.043)$	$-0.527(0.028)$
10^6	0.2	$-3.403(0.002)$	$-3.446(0.036)$	$-3.346(0.016)$	$-3.310(0.010)$
	0.5	$-3.019(0.002)$	$-2.948(0.026)$	$-2.945(0.011)$	$-2.867(0.007)$
	0.8	$-1.517(0.002)$	$-1.413(0.011)$	$-1.106(0.004)$	$-0.527(0.003)$

which we compare with the symmetric finite difference (FD) estimator

$$\left(\frac{\partial v_n(\theta)}{\partial \theta}\right)_{FD} = \frac{v_n(\theta+\frac{\delta}{2}) - v_n(\theta-\frac{\delta}{2})}{\delta},$$

where δ is the size of the symmetric difference. We compute both derivative estimators at $\theta = 0.2, 0.5, 0.8$ and test with $\delta = 0.01, 0.05, 0.1$ and number of replications $N = 10^2, 10^4, 10^6$. Simulation results are shown in Table 1 and Figure 1. We have the following observations:

- For a small number of replications, SPA has much smaller bias and standard error (SE) than FD.
- FD at $\delta = 0.1$ has a large bias that can be reduced at the expense of variance.
- FD at $\delta = 0.01$ is almost unbiased but has much larger variance than SPA.
- For a fixed number of replications, the standard error of the SPA estimator is almost identical at different θ, whereas the precision of the FD estimator is proportional to the derivative.

6 SUMMARY AND FUTURE RESEARCH

We proposed a continuous-state MDP model to study the optimal timing of organ transplantation. Under suitable conditions, we proved that there exists a control limit optimal policy. We derived an SPA estimator for the gradient of the value function w.r.t the control limit, which is useful in computing the optimal control limit by gradient-based simulation optimization. Furthermore, we proved that the SPA estimator is asymptotically unbiased and demonstrated its effectiveness using a simulation example. Solving for the optimal control limit through gradient-based optimization methods will be the focus of future research. Another future research direction is to consider the situation where the donor organ's quality and availability may vary over time. Finally, because implementing the SPA estimator requires additional simulation beyond the nominal sample path, comparing it with other unbiased gradient estimators, for example, the generalized likelihood ratio (GLR) method proposed in Peng et al. (2018) and the measure-valued differentiation (MVD) method in Heidergott and Peng (2023), is an important topic warranting further investigation.

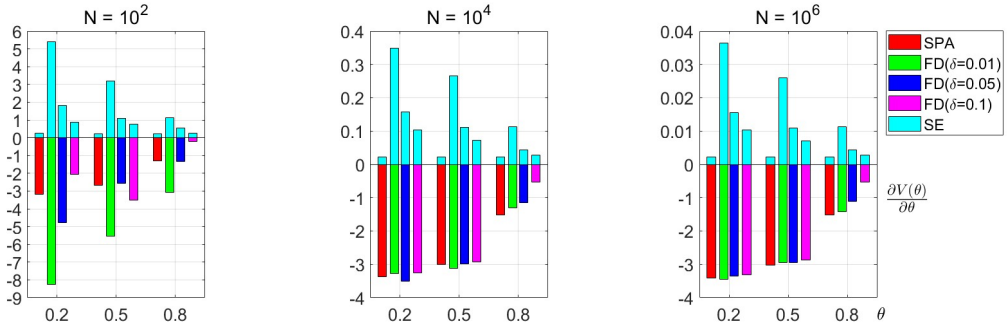

Figure 1: Simulation results for the sensitivity of the value function $V(\theta)$ and their standard errors (SEs).

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant IIS-2123684 and by Air Force Office of Scientific Research under Grant FA95502010211.

REFERENCES

Alagoz, O., H. Hsu, A. J. Schaefer, and M. S. Roberts. 2010. "Markov Decision Processes: A Tool for Sequential Decision Making Under Uncertainty". *Medical Decision Making* 30(4):474–483.

Alagoz, O., L. M. Maillart, A. J. Schaefer, and M. S. Roberts. 2004. "The Optimal Timing of Living-Donor Liver Transplantation". *Management Science* 50(10):1420–1430.

Alagoz, O., L. M. Maillart, A. J. Schaefer, and M. S. Roberts. 2007a. "Choosing Among Living-Donor and Cadaveric Livers". *Management Science* 53(11):1702–1715.

Alagoz, O., L. M. Maillart, A. J. Schaefer, and M. S. Roberts. 2007b. "Determining the Acceptance of Cadaveric Livers Using an Implicit Model of the Waiting List". *Operations Research* 55(1):24–36.

Bartle, R. G., and D. R. Sherbert. 2010. *Introduction to Real Analysis*. 4th ed. New York: John Wiley & Sons.

Batun, S., A. J. Schaefer, A. Bhandari, and M. S. Roberts. 2018. "Optimal Liver Acceptance for Risk-Sensitive Patients". *Service Science* 10(3):320–333.

Bendersky, M., and I. David. 2016. "Deciding Kidney-Offer Admissibility Dependent on Patients' Lifetime Failure Rate". *European Journal of Operational Research* 251(2):686–693.

Bertsekas, D. P. 2020. *Dynamic Programming and Optimal Control*. 4th ed, Volume 1. Belmont, MA: Athena Scientific.

David, I., and U. Yechiali. 1985. "A Time-Dependent Stopping Problem With Application to Live Organ Transplants". *Operations Research* 33(3):491–504.

Douer, N., and U. Yechiali. 1994. "Optimal Repair and Replacement in Markovian Systems". *Stochastic Models* 10(1):253–270.

Fan, W., Y. Zong, and S. Kumar. 2020. "Optimal Treatment of Chronic Kidney Disease With Uncertainty in Obtaining a Transplantable Kidney: An MDP-Based Approach". *Annals of Operations Research* 316(11):269–302.

Fu, M. C., and J.-Q. Hu. 1997. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Boston: Kluwer Academic.

Heidergott, B., and Y. Peng. 2023. "Gradient Estimation for Smooth Stopping Criteria". *Advances in Applied Probability* 55(1):29–55.

Hernández-Lerma, O., and J. B. Lasserre. 1996. *Discrete-time Markov Control Processes: Basic Optimality Criteria*. New York: Springer.

Kaufman, D., A. J. Schaefer, and M. S. Roberts. 2017. "Living-Donor Liver Transplantation Timing Under Ambiguous Health State Transition Probabilities". *Available at SSRN 3003590*.

Peng, Y., M. C. Fu, J.-Q. Hu, and B. Heidergott. 2018. "A New Unbiased Stochastic Derivative Estimator for Discontinuous Sample Performances With Structural Parameters". *Operations Research* 66(2):487–499.

Prieto, L., and J. A. Sacristán. 2003. "Problems and Solutions in Calculating Quality-Adjusted Life Years (QALYs)". *Health and Quality of Life Outcomes* 1:1–8.

Ren, X., M. C. Fu, and S. I. Marcus. 2022. "Optimal Acceptance of Incompatible Kidneys". *arXiv preprint arXiv:2212.01808*.

Ross, S. M. 1996. *Stochastic Processes*. 2nd ed. New York: John Wiley & Sons.

AUTHOR BIOGRAPHIES

XINGYU REN is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of Maryland, College Park. His research interests include stochastic optimization and Markov decision processes. His e-mail address is renxy@umd.edu.

MICHAEL C. FU holds the Smith Chair of Management Science in the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research and an affiliate appointment in the Department of Electrical and Computer Engineering, at the University of Maryland, College Park. His research interests include stochastic gradient estimation, simulation optimization, and applied probability. He served as WSC2011 Program Chair and received the INFORMS Simulation Society's Distinguished Service Award in 2018. He is a Fellow of INFORMS and IEEE. His e-mail address is mfu@umd.edu.

STEVEN I. MARCUS is Professor Emeritus in the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland. He is former Editor-in-Chief of the SIAM Journal on Control and Optimization. His research is focused on stochastic control and estimation, Markov decision processes, and hybrid systems. He is a Fellow of IEEE and SIAM. His email address is marcus@umd.edu.