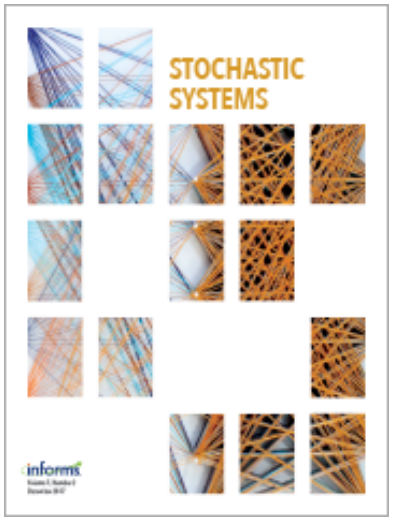

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Polynomial-Time Algorithm for Optimal Stopping with Fixed Accuracy

Yilun Chen, David A. Goldberg

To cite this article:
Yilun Chen, David A. Goldberg (2026) Polynomial-Time Algorithm for Optimal Stopping with Fixed Accuracy. Stochastic Systems

Published online in Articles in Advance 07 Aug 2026

. https://doi.org/10.1287/stsy.2024.0075

https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Articles in Advance, pp. 1–18
ISSN 1946-5238 (online)

Polynomial-Time Algorithm for Optimal Stopping with Fixed Accuracy

Yilun Chen,[a] **David A. Goldberg**[b,*]

[a]School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China; [b]School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14850

*Corresponding author

chenyilun@cuhk.edu.cn (Y. Chen), dag369@cornell.edu (D. A. Goldberg)

https://orcid.org/0000-0002-5761-0286 (D. A. Goldberg)

Received: May 14, 2024
Revised: February 11, 2026; June 16, 2026
Accepted: June 25, 2026
Published Online in Articles in Advance: August 7, 2026

https://doi.org/10.1287/stsy.2024.0075

Abstract. The optimal stopping (OS) problem is important to multiple academic communities and applications. Modern OS tasks often have long horizons and complicated, high-dimensional dynamics, making them especially challenging. Many past approaches have computational cost scaling exponentially in the horizon and/or underlying dimension in the worst case, suffering from the curse of dimensionality. In this work, we develop a novel expansion representation for the OS value. We prove that truncating this expansion yields a simulation-based algorithm that implements an ϵ-optimal stopping policy with computational complexity scaling polynomially in the time horizon and the underlying dimension (for any fixed ϵ). We also explore some connections between our expansion and the martingale duality theory for OS.

Funding: Y. Chen acknowledges support from the National Natural Science Foundation of China (NSFC) [Grants NSFC-72501250 and NSFC-72394361] and the Guangdong Key Lab of Mathematical Foundations for Artificial Intelligence.

Supplemental Material: The online companion is available at https://doi.org/10.1287/stsy.2024.0075.

Keywords: optimal stopping • high-dimensional • path dependent • dynamic programming • computationally efficient algorithm • Monte Carlo simulation • polynomial-time approximation scheme (PTAS)

1. Introduction

We consider the problem of discrete-time optimal stopping (OS): a decision maker observes a stochastically evolving information process $\mathbf{Y} = \{Y_1, \ldots, Y_T\}$, where we assume that (for all t) Y_t is a D-dimensional vector for some $D \geq 1$. Let $\mathcal{F} = \{\mathcal{F}_t, t = 1, \ldots, T\}$ be the natural filtration generated by $\mathbf{Y}$, that is, $\mathcal{F}_t \triangleq \sigma(Y_1, \ldots, Y_t)$. Let $\{g_t : \mathbb{R}^{D \times t} \to \mathbb{R}, t = 1, \ldots, T\}$ be a sequence of measurable reward functions mapping trajectories $Y_{[t]}$ to reals, with $Y_{[t]} = (Y_1, \ldots, Y_t)$. For notational simplicity, let $Z_t \triangleq g_t(Y_{[t]})$, and denote the corresponding stochastic process by $\mathbf{Z}$. Throughout this paper, we assume Z_t is integrable for all t, and the problem of interest is to compute $\text{OPT} = \sup_{\tau \in \mathcal{T}} E[Z_\tau]$, where $\mathcal{T}$ denotes the set of all $\mathcal{F}$-adapted stopping times.

Problems of this kind find an array of applications across areas including operations research and management science, finance, economics, statistics, and computer science. In practice, the problems are often high-dimensional, with long horizons and non-Markovian/path-dependent dynamics, posing significant computational challenges. These challenges have generated substantial research across multiple academic communities and led to a vast literature on approximation algorithms and heuristics. Here, we briefly describe some of the main past approaches to OS, and defer a more detailed discussion to our literature review in Section 1.2.

The canonical approach to OS is to formulate it as a stochastic dynamic program (DP). The associated Bellman equation takes the form $V_t = \max(Z_t, E[V_{t+1} | \mathcal{F}_t]), t = 1, \ldots, T-1$, and $V_T = Z_T$. Here, V_t is the reward-to-go function in period t, and we note that V_t is $\mathcal{F}_t$-measurable for all t. Then $\text{OPT} = E[V_1]$. Naively implementing this recursion, however, takes time exponential in the dimension and/or time horizon, a computational challenge known as the curse of dimensionality. Since the late 1990s, a major line of research has sought to overcome this challenge through approximate dynamic programming (ADP) methods (Tsitsiklis and Van Roy 1999, Longstaff and Schwartz 2001). Here one uses a judicious choice of basis functions and/or random sampling to approximate

the Bellman equation and yield tractable algorithms, although typically having an optimality gap that depends on the quality of these basis functions (and that cannot be made arbitrarily small without incurring an exponential complexity or making extra assumptions).

Another popular approach builds on a dual formulation for OS, pioneered in the seminal work of Davis and Karatzas (1994). They establish a strong duality: $\text{OPT} = \inf_{\mathbf{M}\in\mathcal{M}^0} E[\max_{t=1,\dots,T}(Z_t - M_t)]$, where $\mathcal{M}^0$ denotes the set of mean-zero martingales adapted to $\mathcal{F}$. Moreover, they showed that the Doob–Meyer decomposition of the value functions yields an optimal martingale $\mathbf{M}^*$. Although not offering a resolution to the curse of dimensionality (as computing $\mathbf{M}^*$ requires prior knowledge of the value functions themselves), this duality theory led to substantial algorithmic progress that complements ADP methods.

In this work, we provide new theoretical insights into OS by introducing a novel recursive expansion representation of the OS value, which leads to a polynomial-time algorithm with provable approximation guarantees. Below we summarize our main contributions.

1.1. Main Contributions

1.1.1. Novel Expansion Theory. The standard DP formulation essentially relies on backward induction (in time) to set up the recursive representation of OPT. We develop a novel expansion representation of OPT that builds on fundamentally different logic. More precisely, the first term of our expansion coincides with the expected pathwise maximum $E[\max_{t=1,\dots T} Z_t]$, that is, the hindsight optimal value. The second term is an expectation that can be interpreted as a type of regret, and subsequent terms are defined recursively as the expectation of a maximum that can be interpreted as a higher-order notion of regret. This expansion representation has an intimate connection to the dual martingale formulation and yields an optimal dual martingale different from $\mathbf{M}^*$. Under certain natural regularity conditions on the reward process (e.g., Assumption 1 on bounded rewards, or Assumption A.1 in the appendix regarding the bounded squared coefficient of variation of the maximum reward), we show that truncating our expansion after $O\left(\frac{1}{\epsilon}\right)$ terms yields an ϵ-approximation to OPT, with those terms of the expansion requiring a level of nesting (of conditional expectations) scaling only as $O\left(\frac{1}{\epsilon}\right)$ (independent of the time horizon or dimension). We believe this representation complements the existing literature and offers a novel perspective on the theory of OS.

1.1.2. Polynomial-Time Algorithm. We formalize an algorithm that implements (i.e., approximately computes) an appropriate truncation of our expansion representation via Monte Carlo simulation, given access to an efficient simulator of $\mathbf{Y}$. Using this algorithm, we implement an ϵ-optimal policy in time $T^{O\left(\frac{1}{\epsilon}\right)}$, for any fixed $\epsilon \in (0,1)$. More precisely, the optimal dual martingale associated with our approach has an analogous expansion representation. Our stopping policy truncates this expansion at depth $O\left(\frac{1}{\epsilon}\right)$, uses Monte Carlo simulation to approximate the value of this truncated martingale at each time period, and stops the first time the reward process is sufficiently close to this approximated martingale. As simulating the kth term of the expansion requires implementing depth k nested conditional expectations, this requires a simulation and computational complexity scaling as $T^{O\left(\frac{1}{\epsilon}\right)}$. Because many other approaches (including DP and naive fully nested simulation) have a worst-case complexity growing exponentially with the horizon and/or dimension, our results demonstrate an exponential improvement and serve as a proof of concept that general high-dimensional path-dependent OS problems are polynomially approximable. To be clear, here polynomial time is meant in the fixed-accuracy sense: for each fixed ϵ, the exponent of T is a constant (depending on ϵ). Our algorithms have an exponential (not polynomial) dependence on $\frac{1}{\epsilon}$ itself.

1.2. Related Literature

There is a vast literature on computational OS, and one of the most popular approaches is regression/ADP. Here one fixes a family of basis functions to approximate the DP value functions. Seminal papers in this area include Tsitsiklis and Van Roy (2001) and Longstaff and Schwartz (2001). There was much subsequent work on, for example, nonparametric regression, smoothing spline regression, recursive kernel regression, integer programming, reinforcement learning, multilevel Monte Carlo, and (deep) neural networks (see Lai and Wong 2004; Egloff 2005; Kohler 2008; Belomestny et al. 2015, 2020; Bayer et al. 2021; Becker et al. 2021; Ciocan and Mišić 2022; Sturt 2023; Zhou et al. 2023). These methods mostly focus on achieving empirical success rather than ensuring theoretical guarantees, with their performance relying on how well the choice of basis functions can approximate the true value functions. Those works with theoretical guarantees often require additional continuity assumptions, and/or scale exponentially in the time horizon. Most existing theoretical analyses for these methods focus on their convergence to the best approximation within the fixed set of basis functions (Clément et al. 2002,

Glasserman and Yu 2004, Stentoft 2004, Belomestny 2011, Bouchard and Warin 2012, Bezerra et al. 2020) and do not have guarantees comparable to our own.

Building on the seminal work of Davis and Karatzas (1994), significant progress was made simultaneously by Haugh and Kogan (2004) and Rogers (2002) in their formulation of a dual methodology for OS. Instead of finding an OS time, it formulates and solves a dual optimization problem over the set of adapted martingales. Other dual representations were subsequently discovered (e.g., Jamshidian 2007), and the methodology was extended to more general control problems (Brown et al. 2010). We note that closely related dual formulations and methodologies have also appeared in the study of multistage stochastic optimization (Rockafellar and Wets 1991). Simulation approaches via dual formulation have since led to substantial algorithmic progress (see Andersen and Broadie 2004; Kolodko and Schoenmakers 2004, 2006; Belomestny and Milstein 2006; Chen and Glasserman 2007; Desai et al. 2012; Belomestny et al. 2013, 2019; Christensen 2014; Lelong 2018; Ibáñez and Velasco 2020). To achieve computational efficiency, many of these algorithms adopt ideas similar to those of ADP approaches, for example, fixing a family of basis functions/martingales to approximate the optimal dual martingale. Consequently, the quality of the bounds derived using these approaches typically depends on the expressiveness of the set of basis functions/martingales, or the accuracy of the initial estimation of the value functions. In addition, those works that do provide rigorous approximation guarantees typically require a level of nesting (of conditional expectations) scaling with the time horizon (see Kolodko and Schoenmakers 2006, Chen and Glasserman 2007).

More recently, new machine learning techniques, such as decision trees and deep neural networks, and ideas from robust optimization were introduced to solve the problem of high-dimensional OS and American option pricing (see Kohler et al. 2010, Becker et al. 2019, Becker et al. 2021, Ciocan and Mišić 2022, Sturt 2023). These approaches exhibit excellent performance on a number of benchmarks, but generally do not have strong theoretical guarantees independent of the quality of the basis functions/free from additional continuity assumptions.

1.3. Organization

The rest of this paper is organized as follows. In Section 2, we describe our expansion representation, establish its connection to the martingale duality theory, and provide associated error bounds. In Section 3, we introduce our expansion-inspired algorithm and state and prove our main algorithmic results. We present some closing thoughts and directions for future research in Section 4. We include an appendix, which contains several auxiliary technical lemmas used in our proof, along with a generalization to unbounded rewards. We also include an Online Companion, in which we prove that the well-studied Bermudan max call falls within our theoretical framework and present the results of some numerical experiments (mostly to illustrate the practical limitations of our approach).

2. Novel Expansion Representation for OPT

In this section, we establish our expansion representation of OPT, draw connections to the martingale duality theory, and analyze its convergence rate.

2.1. Novel Expansion Representation for OPT

We begin by giving the simple intuition behind the expansion. We wish to compute $\text{OPT} = \sup_{\tau\in\mathcal{T}} E[Z_\tau]$. Trivially, $\sup_{\tau\in\mathcal{T}} E[Z_\tau] \le E[\max_{t=1,\dots,T} Z_t]$. We next turn this straightforward bound into an equality by compensating with a remainder term, which is characterized by a new OS problem. We introduce a specific martingale $\{E[\max_{i=1,\dots,T} Z_i|\mathcal{F}_t], t = 1,\dots,T\}$, that is, the Doob martingale of $\max_{t=1,\dots,T} Z_t$, adapted to filtration $\mathcal{F}$. By the optional stopping theorem, for any $\tau \in \mathcal{T}$, it holds that $E[E[\max_{t=1,\dots,T} Z_t|\mathcal{F}_\tau]] = E[\max_{t=1,\dots,T} Z_t]$, and thus

$$E[Z_\tau] = E\left[\max_{t=1,\dots,T} Z_t\right] + E\left[Z_\tau - E\left[\max_{t=1,\dots,T} Z_t|\mathcal{F}_\tau\right]\right].$$

Taking supremum on both sides, we conclude that

$$\text{OPT} = E\left[\max_{t=1,\dots,T} Z_t\right] + \sup_{\tau\in\mathcal{T}} E\left[Z_\tau - E\left[\max_{t=1,\dots,T} Z_t|\mathcal{F}_\tau\right]\right]. \tag{1}$$

For $t = 1,\dots,T$, let $Z_t^1 \triangleq Z_t, Z_t^2 \triangleq Z_t^1 - E[\max_{i=1,\dots,T} Z_i^1|\mathcal{F}_t]$ (with probability one (w.p.1) nonpositive). Then (1) becomes

$$\text{OPT} = E\left[\max_{t=1,\dots,T} Z_t^1\right] + \sup_{\tau\in\mathcal{T}} E[Z_\tau^2].$$

There are many bounds for the gap $E[\max_{t=1,\dots,T} Z_t^1] - \text{OPT}$ in the literature on so-called prophet inequalities (see Hill and Kertz 1983, 1992). The novelty of our approach will be to "recurse" this logic on the problem $\sup_{\tau\in\mathcal{T}} E[Z_\tau^2]$ itself, and subsequently a sequence of such problems defined recursively. More formally, let $Z_t^{k+1} \triangleq Z_t^k - E[\max_{i=1,\dots,T} Z_i^k \,|\, \mathcal{F}_t]$ for all $k \geq 1$ and $t \in \{1,\dots,T\}$ (denoting the corresponding stochastic process by $\mathbf{Z}^k$).

It follows from the basic properties of conditional expectation and a straightforward induction (the details of which we omit) that $Z_t^k \leq 0$ for all $t \in \{1,\dots,T\}$ and $k \geq 2$. Then our first main result is as follows.

Theorem 1. $\text{OPT} = \sum_{k=1}^{\infty} E[\max_{t=1,\dots,T} Z_t^k]$.

Before proving Theorem 1, we first state and prove some auxiliary results for $\mathbf{Z}^k$.

Lemma 1. (1) *For all* $k \geq 0$, $\text{OPT} = \sum_{i=1}^{k} E[\max_{t=1,\dots,T} Z_t^i] + \sup_{\tau\in\mathcal{T}} E[Z_\tau^{k+1}]$; (2) *w.p.*1, Z_t^k *is nonpositive and integrable for all* $k \geq 2$ *and* $t = 1,\dots,T$; (3) *for* $t = 1,\dots,T$, $\{Z_t^k, k \geq 2\}$ *is a monotone increasing sequence of random variables (r.v.s.); and* (4) $\mathbf{Z}^k$ *is adapted to* $\mathcal{F}$ *for all* $k \geq 1$.

Proof of Lemma 1. Properties (2), (3), and (4) can be easily verified by straightforward induction arguments, and we omit the details. Let us prove property (1) and proceed by induction. The base case $k = 0$ follows from definitions. Suppose the induction is true for all $j \leq k$ for some $k \geq 0$. Thus, $\text{OPT} = \sum_{i=1}^{k} E[\max_{t=1,\dots,T} Z_t^i] + \sup_{\tau\in\mathcal{T}} E[Z_\tau^{k+1}]$. By the optional stopping theorem, $\forall\, \tau \in \mathcal{T}$,

$$\begin{aligned} E[Z_\tau^{k+1}] &= E\left[Z_\tau^{k+1} - E\left[\max_{t=1,\dots,T} Z_t^{k+1} \,|\, \mathcal{F}_\tau\right]\right] + E\left[\max_{t=1,\dots,T} Z_t^{k+1}\right] \\ &= E[Z_\tau^{k+2}] + E\left[\max_{t=1,\dots,T} Z_t^{k+1}\right]. \end{aligned}$$

Thus, $\sup_{\tau\in\mathcal{T}} E[Z_\tau^{k+1}] = E[\max_{t=1,\dots,T} Z_t^{k+1}] + \sup_{\tau\in\mathcal{T}} E[Z_\tau^{k+2}]$, and by the induction hypothesis,

$$\begin{aligned} \text{OPT} &= \sum_{i=1}^{k} E\left[\max_{t=1,\dots,T} Z_t^i\right] + \sup_{\tau\in\mathcal{T}} E[Z_\tau^{k+1}] \\ &= \sum_{i=1}^{k+1} E\left[\max_{t=1,\dots,T} Z_t^i\right] + \sup_{\tau\in\mathcal{T}} E[Z_\tau^{k+2}]. \quad \square \end{aligned}$$

With Lemma 1 in hand, we would be done with the proof of Theorem 1 if we could show that $\lim_{k\to\infty} \sup_{\tau\in\mathcal{T}} E[Z_\tau^{k+1}] = 0$. We now prove that this is indeed the case.

Proof of Theorem 1. We proceed by constructing a particular stopping time τ_∞ for which $\lim_{k\to\infty} E[Z_{\tau_\infty}^k] = 0$, from which the desired result will follow (as the supremum over stopping times is at least the value achieved by any one stopping time). We will define τ_∞ as the first time that $\lim_{k\to\infty} Z_t^k = 0$. Before formalizing this definition, we will need to argue that there always exists such a time, which we do now. It follows from Lemma 1 and monotone convergence that

$$\{\mathbf{Z}^k, k \geq 1\} \text{ converges almost surely (a.s.), where we define } Z_t^\infty \triangleq \lim_{k\to\infty} Z_t^k. \tag{2}$$

Thus, $\{Z_T^{k+1} - Z_T^k, k \geq 1\}$ converges a.s. to zero. By definition, $Z_T^{k+1} = Z_T^k - \max_{t=1,\dots,T} Z_t^k$, and it follows that $\{\max_{t=1,\dots,T} Z_t^k, k \geq 1\}$ converges a.s. to zero, that is, $\lim_{k\to\infty} \max_{t=1,\dots,T} Z_t^k = 0$ a.s. As the function $\max(\cdot)$ is continuous on $\mathcal{R}^T$, it follows from the continuous mapping theorem for almost sure convergence (see theorem 2.3 of van der Vaart 2000) that we may interchange max and limit, that is, $\lim_{k\to\infty} \max_{t=1,\dots,T} Z_t^k = \max_{t=1,\dots,T} Z_t^\infty$ w.p.1. Thus, because zero and $\max_{t=1,\dots,T} Z_t^\infty$ are both almost sure limits of $\{\max_{t=1,\dots,T} Z_t^k, k \geq 2\}$, it follows from the uniqueness of almost sure limits that

$$\max_{t=1,\dots,T} Z_t^\infty = 0 \text{ w.p.1}, \tag{3}$$

equivalently (because nonpositivity of $\{Z_t^k, k \geq 2\}$ implies nonpositivity of Z_t^∞ for all t),

$$P(\exists\, t \in \{1,\dots,T\} \text{ such that } Z_t^\infty = 0) = 1. \tag{4}$$

We now formally introduce the stopping time τ_∞. Let $\tau_\infty \triangleq \min\{t \in \{1,\dots,T\} : Z_t^\infty = 0\}$, where (4) ensures τ_∞ is well defined, and the adaptedness of $\mathbf{Z}^k$ for all k (along with the measurability of limsup and liminf) ensures τ_∞ is a stopping time.

We next argue that $\lim_{k\to\infty} E[Z^k_{\tau_\infty}] = 0$. First, we observe that the definition of τ_∞ implies that w.p.1 $Z^\infty_{\tau_\infty} = 0$. We now combine with dominated convergence (d.c.) to prove the desired result. That $\{Z^k_{\tau_\infty}, k \geq 2\}$ satisfies the conditions for d.c. follows from the nonpositivity and monotonicity shown in Lemma 1 because $|Z^k_{\tau_\infty}| \leq \sum_{t=1}^T |Z^k_t| \leq \sum_{t=1}^T |Z^2_t|$ for all $k \geq 2$, and Lemma 1 implies that $\sum_{t=1}^T |Z^2_t|$ is integrable. Thus, by d.c., $\lim_{k\to\infty} E[Z^k_{\tau_\infty}] = E[Z^\infty_{\tau_\infty}] = 0$.

Finally, we prove that this fact implies $\lim_{k\to\infty} \sup_{\tau\in\mathcal{T}} E[Z^{k+1}_\tau] = 0$. The nonpositivity of $Z^k_t, t = 1,\ldots,T$ for $k \geq 2$ implies $0 \geq \limsup_{k\to\infty} \sup_{\tau\in\mathcal{T}} E[Z^k_\tau]$. As the supremum over a set of stopping times is at least what one achieves with any given stopping time, we also have that (for all $k \geq 2$) $\sup_{\tau\in\mathcal{T}} E[Z^k_\tau] \geq E[Z^k_{\tau_\infty}]$, and $\liminf_{k\to\infty} \sup_{\tau\in\mathcal{T}} E[Z^k_\tau] \geq \liminf_{k\to\infty} E[Z^k_{\tau_\infty}] = 0$ (as we have already shown $\lim_{k\to\infty} E[Z^k_{\tau_\infty}] = 0$). Combining with the above completes the proof. □

Let us more explicitly give the first few terms. For $k \geq 1$, let $L_k \triangleq E[\max_{t=1,\ldots,T} Z^k_t]$. Then

$$L_1 = E\left[\max_{t=1,\ldots,T} g_t(Y_{[t]})\right] \;;\; L_2 = E\left[\max_{t=1,\ldots,T}\left(g_t(Y_{[t]}) - E\left[\max_{i=1,\ldots,T} g_i(Y_{[i]})|\mathcal{F}_t\right]\right)\right].$$

Note that the first term, L_1, is the only positive term in the expansion, which corresponds to the obvious upper bound. The term L_2 has a "regret" interpretation, as it is the expectation of the maximum of T terms, with term t intuitively the reward for stopping at time t minus your best guess of the all-time maximum given the information up to time t (i.e., the regret of stopping at time t). Later terms are the expectations of similar explicit suprema with analogous interpretations in terms of certain notions of (negative) "higher-order" regret, each of which can be computed by simulation.

2.2. Connection to Martingale Duality

As a byproduct of the expansion representation, we derive a novel optimal dual martingale. Let $S \triangleq \sum_{k=1}^\infty \max_{t=1,\ldots,T} Z^k_t$. Integrability of S is easily seen to follow from Theorem 1 and monotone convergence (as OPT is finite). Let **MAR** denote the Doob martingale such that (s.t.) $MAR_t = E[S|\mathcal{F}_t]$, $t = 1,\ldots,T$. We now prove that $\mathbf{MAR}^0 \triangleq \mathbf{MAR} - E[S]$ is an optimal dual martingale for OS, and begin by proving the following result.

Lemma 2. *$E[MAR_1] =$ OPT, and $P(\max_{t=1,\ldots,T}(Z_t - MAR_t) = 0) = 1$.*

Proof. For all $t = 1,\ldots,T$, $Z_t - MAR_t$ equals

$$Z_t - E\left[\sum_{k=1}^\infty \max_{i=1,\ldots,T} Z^k_i|\mathcal{F}_t\right] = Z^1_t - E\left[\max_{i=1,\ldots,T} Z^1_i|\mathcal{F}_t\right] - E\left[\sum_{k=2}^\infty \max_{i=1,\ldots,T} Z^k_i|\mathcal{F}_t\right]$$
$$= Z^2_t - E\left[\sum_{k=2}^\infty \max_{i=1,\ldots,T} Z^k_i|\mathcal{F}_t\right].$$

By applying the above inductively, we find that for all $j \geq 1$ and $t = 1,\ldots,T$, w.p.1,

$$Z_t - MAR_t = Z^j_t - E\left[\sum_{k=j}^\infty \max_{i=1,\ldots,T} Z^k_i|\mathcal{F}_t\right]. \tag{5}$$

We now take limits in (5) to prove that, w.p.1,

$$Z_t - MAR_t = Z^\infty_t \text{ for all } t = 1,\ldots,T. \tag{6}$$

It follows from Lemma 1 and monotone convergence that $\{E[\sum_{k=j}^\infty \max_{i=1,\ldots,T} Z^k_i|\mathcal{F}_t], j \geq 2\}$ converges a.s. to a limiting nonpositive r.v. with expectation $\lim_{j\to\infty} E[\sum_{k=j}^\infty \max_{i=1,\ldots,T} Z^k_i]$. By Theorem 1 and essentially the same argument we used to prove integrability of S, this expectation must equal zero, and we conclude that this limiting r.v. in fact equals zero w.p.1 for all t. Taking limits (in j) on the right-hand side of (5) then completes the proof of (6). Combining with (3), it follows that $P(\max_{t=1,\ldots,T}(Z_t - MAR_t) = 0) = 1$. As Theorem 1 and the definition of **MAR** imply that $E[MAR_1] =$ OPT, combining the above completes the proof. □

Because $P(\max_{t=1,\ldots,T}(Z_t - MAR_t) = 0) = 1$ implies $E[\max_{t=1,\ldots,T}(Z_t - (MAR_t - \text{OPT}))] =$ OPT, Lemma 2 immediately implies that $\mathbf{MAR}^0$ is indeed an optimal dual martingale in the sense of Davis and Karatzas (1994).

Corollary 1. *$\mathbf{MAR}^0$ is a zero-mean martingale s.t. $E[\max_{t=1,\ldots,T}(Z_t - MAR^0_t)] =$ OPT, and thus $\mathbf{MAR}^0$ is an optimal dual martingale in the sense of Davis and Karatzas (1994).*

We now discuss a more refined notion of optimality to delineate a key difference between $\mathbf{MAR}^0$ and the standard optimal dual martingale $\mathbf{M}^*$ arising from the Doob–Meyer decomposition of the value functions. First, we recall the notion of i-sure optimality, as formalized in Schoenmakers et al. (2013) (our definition also adjusts for the different time indexing used there).

Definition 1. For any $i \in \{0,\dots,T-1\}$, a martingale $\mathbf{M}$ is *i-surely optimal* if $V_{i+1} = \max_{j=i+1,\dots,T}(Z_j - M_j + M_{i+1})$ w.p.1, and is *surely optimal* if it is i-surely optimal for $i = 0,\dots,T-1$.

By Lemma 2, $\max_{j=1,\dots,T}(Z_j - MAR_j^0 + MAR_1^0) = MAR_1^0$. Thus, $\mathbf{MAR}^0$ would be 0-surely optimal if w.p.1 $MAR_1^0 = V_1$ (the first reward-to-go function). As by definition $MAR_1^0 = E[\sum_{k=1}^{\infty} \max_{t=1,\dots,T} Z_t^k | \mathcal{F}_1]$, the desired equality follows from a proof essentially identical to our proof of Theorem 1, but with all expectations replaced by conditional expectations taken with respect to (w.r.t.) $\mathcal{F}_1$, and we omit the details. We conclude the following.

Corollary 2 (0-Sure Optimality). $\mathbf{MAR}^0$ *is 0-surely optimal.*

We note that (as shown in Schoenmakers et al. 2013) 0-sure optimality is a stronger notion than optimality in the sense of Davis and Karatzas (1994), as it requires optimality in a pathwise sense (for the zero-mean martingale $\{M_t - M_1, t = 1,\dots,T\}$), not just in expectation. It is interesting to point out that the standard optimal dual martingale $\mathbf{M}^*$ of Davis and Karatzas (1994) (arising from the Doob–Meyer decomposition of the value functions) is not only 0-surely optimal, but also surely optimal (see Schoenmakers et al. 2013). We now show that our optimal dual martingale $\mathbf{MAR}^0$ does not satisfy this stronger sure optimality property, creating a clear delineation between $\mathbf{MAR}^0$ and $\mathbf{M}^*$, and defer the proof to the appendix, Section A.1.

Lemma 3 (Not Surely Optimal). $\mathbf{MAR}^0$ *is not surely optimal, in contrast to the standard optimal dual martingale* $\mathbf{M}^*$ *of Davis and Karatzas* (1994).

Next, we point out that τ_∞ can be interpreted as a "dual-threshold" stopping time w.r.t. $\mathbf{MAR}$, and is in fact an optimal stopping time. Recall that $\tau_\infty = \min\{t \in \{1,\dots,T\} : Z_t^\infty = 0\}$, and hence, by (6), $\tau_\infty = \min\{t \in \{1,\dots,T\} : Z_t - MAR_t = 0\}$, that is, τ_∞ can indeed be interpreted as a dual-threshold stopping time.

Observation 1. τ_∞ is an optimal solution to the stopping problem $\sup_{\tau \in \mathcal{T}} E[Z_\tau]$, that is, $E[Z_{\tau_\infty}] = \text{OPT}$, because $Z_{\tau_\infty} = MAR_{\tau_\infty}$ (by the dual-threshold interpretation) and $E[MAR_{\tau_\infty}] = \text{OPT}$ by Lemma 2 and the optional stopping theorem.

The polynomial-time stopping policy we define in Section 3 essentially implements the stopping time τ_∞ defined in Observation 1, but with Z^∞ replaced by the approximation Z^k (for $k = O(\frac{1}{\epsilon})$) in the definition $\tau_\infty = \min\{t \in \{1,\dots,T\} : Z_t^\infty = 0\}$ (and "equals zero" replaced by "at least $-\frac{1}{k-1}$"), where Z^k is itself approximated by a depth k nested Monte Carlo simulation.

Let us close this section by comparing our approach to other duality-based approaches appearing in the literature. Essentially all of these previous works either (1) assume that one is given a good approximation to the value function (Belomestny et al. 2009) or a set of basis functions whose span contains a good approximation to the value function (Desai et al. 2012) and use these to construct a good candidate martingale; (2) take a primal-dual approach by starting with a candidate (in general suboptimal) stopping time and generating a (in general suboptimal) martingale by considering the martingale part of the Doob–Meyer decomposition of the value process generated by the stopping time (Andersen and Broadie 2004); (3) construct approximately optimal martingales using neural networks (Ye and Wong 2025); or (4) construct a sequence of recursively defined martingales that (after T iterations) converges to the optimal dual martingale (Kolodko and Schoenmakers 2006). Our approach is conceptually most similar to that of Kolodko and Schoenmakers (2006), directly constructing a sequence of recursively defined martingales without needing to be given approximations to the value functions or an associated basis. The key difference between our work and past such recursive constructions is that our sequence of martingales provably converges to optimality at rate $O(\frac{1}{k})$, whereas previous such constructions (although exhibiting good empirical performance) had no such theoretical guarantees on the optimality gap of the iterates (beyond achieving optimality after the full T iterations, i.e., any theoretical guarantees required a deep level of nesting for long-horizon problems).

2.3. Approximation Guarantees When Truncating the Expansion

The power of Theorem 1 is that it allows for rigorous approximation guarantees when the infinite sum is truncated, under a natural boundedness assumption on $\mathbf{Z}$.

Assumption 1. *Assume Z_t is uniformly bounded with $0 \le Z_t \le U$ w.p.1 for all t, with U a constant independent of both the time horizon T and the dimension D.*

Let $E_k \triangleq \sum_{i=1}^{k} L_i$ denote the finite-sum truncation of our expansion after the first k terms.

Theorem 2. *Suppose Assumption 1 holds. Then, for all* $k \geq 1$, $0 \leq E_k - \text{OPT} \leq \frac{U}{k}$.

Thus, truncating our expansion after k terms yields an absolute error at most $\frac{U}{k}$, independent of the distribution of the underlying stochastic processes $\mathbf{Y}$ and $\mathbf{Z}$ or length of the time horizon, requiring only that the reward functions are bounded in $[0, U]$. Analogous results also hold when $\mathbf{Z}$ is allowed to be negative (and is uniformly bounded both above and below), although we do not formalize that extension here.

To prove Theorem 2, we state and prove a stronger pathwise guarantee, relating $\max_{t=1,\ldots,T} Z_t^k$ to $\max_{t=1,\ldots,T} Z_t^\infty$ (which by Lemma 2 and (6) equals zero). By the logic described at the end of Section 2.2, this will ultimately enable us to construct provably accurate and polynomial-time stopping policies by (intuitively) stopping the first time that $Z_t^k \geq -\frac{U}{k-1}$ (for $k = O(\frac{U}{\epsilon})$), instead of the truly optimal first time that $Z_t^\infty \geq 0$.

Lemma 4. *Suppose that Assumption 1 holds. For* $k \geq 2$, *w.p.*1, $0 \leq -\max_{t=1,\ldots,T} Z_t^k \leq \frac{U}{k-1}$.

Proof. That $0 \leq -\max_{t=1,\ldots,T} Z_t^k$ follows from Lemma 1. It thus suffices to show the upper bound. Consider the last time period T. For $k \geq 1$, $Z_T^{k+1} = Z_T^k - \max_{t=1,\ldots,T} Z_t^k$. Summing over k then yields a telescoping sum, leading to the equation $Z_T^{k+1} = Z_T - \sum_{i=1}^{k} \max_{t=1,\ldots,T} Z_t^i$. By Lemma 1, Z_T^{k+1} is nonpositive because $k \geq 1$, hence $Z_T - \sum_{i=1}^{k} \max_{t=1,\ldots,T} Z_t^i \leq 0$ w.p.1. We may rewrite this inequality as $\sum_{i=2}^{k} \max_{t=1,\ldots,T} Z_t^i \geq Z_T - \max_{t=1,\ldots,T} Z_t$. Again by Lemma 1, we know that Z_t^i is monotone increasing in i for $i \geq 2$ (and all $t \in \{1,\ldots,T\}$). As a result, $\max_{t=1,\ldots,T} Z_t^i$ is also monotone increasing in i for $i \geq 2$. Thus, $\sum_{i=2}^{k} \max_{t=1,\ldots,T} Z_t^i \leq (k-1) \times \max_{t=1,\ldots,T} Z_t^k$. Combining the above arguments, we conclude that $Z_T - \max_{t=1,\ldots,T} Z_t \leq \sum_{i=2}^{k} \max_{t=1,\ldots,T} Z_t^i \leq (k-1) \times \max_{t=1,\ldots,T} Z_t^k$. Because we assume $Z_t \in [0, U]$ for $t = 1,\ldots,T$, it must hold that $Z_T - \max_{t=1,\ldots,T} Z_t \geq -U$. Combining the above, we conclude that $(k-1) \times \max_{t=1,\ldots,T} Z_t^k \geq -U$, and thus $-\max_{t=1,\ldots,T} Z_t^k \leq \frac{U}{k-1}$, completing the proof. □

Proof of Theorem 2. By Lemma 1, $\text{OPT} = \sum_{i=1}^{k} E[\max_{t=1,\ldots,T} Z_t^i] + \sup_{\tau \in \mathcal{T}} E[Z_\tau^{k+1}] = E_k + \sup_{\tau \in \mathcal{T}} E[Z_\tau^{k+1}]$. In light of Lemma 4, $\tau_k \triangleq \min\{t = 1,\ldots,T : Z_t^k \geq -\frac{U}{k-1}\}$ is a well-defined stopping time satisfying $E[Z_{\tau_k}^k] \geq -\frac{U}{k-1}$. Therefore, $E_k - \text{OPT} = -\sup_{\tau \in \mathcal{T}} E[Z_\tau^{k+1}] \leq -E[Z_{\tau_{k+1}}^{k+1}] \leq \frac{U}{k}$. Combining with the fact that $E_k \geq \text{OPT}$ by definitions and the nonpositivity of Z_t^k for $k \geq 2$ proven in Lemma 1 completes the proof. □

We next present a lower bound showing that Theorem 2 is tight up to an absolute multiplicative constant factor (asymptotically in k). Namely, there is a precise sense in which our bound on the rate of convergence of the stated expansion cannot be substantially improved. We defer the proof to the appendix, Section A.2.

Lemma 5 (Lower Bound). *For any given* $k \geq 2$, *there exists a two-period OS problem satisfying Assumption 1 (with* $U = 1$*), with optimal value* opt^k, *such that* $E_k - opt^k = (1 - \frac{1}{k+1})^k \times \frac{1}{k+1}$. *In particular,* $\lim_{k\to\infty} k \times (E_k - opt^k) = e^{-1}$.

Thus, under Assumption 1, truncating our expansion after k terms yields an error that is at most $\frac{1}{k}$ and (for large k) at least (asymptotically) $\frac{e^{-1}}{k}$ (in the worst case).

In this section, we established the convergence rate of our expansion under a boundedness assumption on the reward process (Assumption 1). This condition is not essential for our main results, but rather assumed to enhance readability. Essentially the same set of convergence guarantees can be derived when Assumption 1 is relaxed and replaced by suitable moment bounds on the rewards. We defer a detailed treatment of one such extension (in which a bound is assumed on the squared coefficient of variation of the maximum reward) to the appendix, Section A.5.

3. Polynomial-Time Algorithm and Its Analysis

In this section, we describe and analyze our main algorithm $\mathcal{A}$, which (approximately) implements the stopping policy described informally in Section 2.3, that is, stop the first time $Z_t^k \geq -\frac{U}{k-1}$, with $k = O(\frac{U}{\epsilon})$ and Z_t^k approximated by Monte Carlo simulation (involving depth k nested conditional expectations). Because by (5) and the definition of **MAR**, $Z_t^k = Z_t - E[\sum_{j=1}^{k-1} \max_{i=1,\ldots,T} Z_i^j | \mathcal{F}_t]$, such a stopping time is equivalent to stopping the first time $Z_t - E[\sum_{j=1}^{k-1} \max_{i=1,\ldots,T} Z_i^j | \mathcal{F}_t] \geq -\frac{U}{k-1}$ (with $k = O(\frac{U}{\epsilon})$). Thus, $E[\sum_{j=1}^{k-1} \max_{i=1,\ldots,T} Z_i^j | \mathcal{F}_t]$ serves as an expansion-based approximation to **MAR**, with our stopping time the corresponding approximation to τ_∞.

We note that given such an algorithm, one can easily derive an algorithm for approximating the optimal value (by generating random sample paths, implementing $\mathcal{A}$ on them, and averaging/taking the median), which we later state as a corollary. We begin by formalizing a generative/simulation model for the OS problem.

Assumption 2. *There exists a simulation oracle $\mathcal{B}$ that, given any $t \in \{1,\dots,T\}$ and partial trajectory $\gamma \in \mathbb{R}^{D\times t}$, generates one trajectory of the underlying process $Y_{[T]}$ conditioned on $\{Y_{[t]} = \gamma\}$. Given input $t = 0$, $\mathcal{B}$ generates an unconditioned trajectory of the underlying process $Y_{[T]}$.*

Let us also formalize a computational model for analyzing algorithms. We adopt a model in line with Shapiro and Nemirovski (2005) and Swamy and Shmoys (2012), in which one is given access to an appropriate (simulation) oracle and can perform basic arithmetic operations on real numbers (regardless of the number of bits). Such a model falls within the framework of information complexity (Traub and Werschulz 1998), and we refer the reader to De Klerk (2008) for a more detailed comparison with the Turing-machine/bit complexity model.

Consistent with the assumptions in these works, we suppose that addition, subtraction, multiplication, division, maximum, and minimum of any two numbers can be done in one unit of time, as can the raising of one number to the power of another, regardless of the values of those numbers. We also suppose that the median of n numbers can be computed in n units of computational time, using, for example, the celebrated linear-time method for computing the median (Cormen et al. 2009), and ignoring the absolute constant governing the associated linear run time. We ignore all computational costs associated with reading, writing, and storing numbers in memory, as well as inputting numbers to functions. In general, our model allows for computation over real numbers. We also assume that drawing a sample from a standard Bernoulli, uniform ($\mathcal{U}[0,1]$), or normal ($\mathcal{N}(0,1)$) r.v.s. takes one unit of time. Most importantly, we assume calling $\mathcal{B}$ to generate a (conditioned) trajectory of $\mathbf{Y}$ takes C units of time, and evaluating the reward function $g_t(\cdot)$ takes G units of time. We assume that both G and C scale at most polynomially with T and D.

3.1. Algorithm Description and Guarantees

Before stating algorithm $\mathcal{A}$, we first formalize a family of subroutines $\{\mathcal{B}^k, k \geq 1\}$ that can recursively (approximately) compute the value of Z_t^k conditional on $Y_{[t]} = \gamma$ for any given input $\gamma \in \mathbb{R}^{D\times t}$ (which we denote by $Z_t^k(\gamma)$) to accuracy ϵ with probability at least $1-\delta$ (with ϵ, δ inputs to $\mathcal{B}^k$). When $k = 1$, $\mathcal{B}^1(t,\gamma)$ simply returns the exact reward value $g_t(\gamma)$. When $k \geq 2$, as $Z_t^k(\gamma) = Z_t^{k-1}(\gamma) - E[\max_{i=1,\dots,T} Z_i^{k-1} | Y_{[t]} = \gamma]$, the algorithm $\mathcal{B}^k$ approximates $Z_t^k(\gamma)$ by natural Monte Carlo simulation of the conditional expectation, with repeated recursive calls to $\mathcal{B}^{k-1}$. A detailed description of $\mathcal{B}^k$ can be found in Algorithm 1. Here, we recall that the "median trick" (as we formalize in Lemma A.2 of the appendix, Section A.4) is a direct corollary of Hoeffding's inequality that is commonly used in speeding up estimation algorithms (by returning the median of independent estimations; see, e.g., Wang and Han 2015), and which goes back at least to the work of Nemirovski and Yudin (1983). By carefully adjusting (as a function of k) the parameters ϵ and δ at different depths of recursion, we can ensure that $\mathcal{B}^k$ achieves the desired guarantees. We formalize the performance guarantee of $\mathcal{B}^k$ in the following result, deferring the proof to Section 3.2.

Algorithm 1 ($\mathcal{B}^k$ for Computing Z_t^k)

Input: $t, \gamma, \epsilon, \delta$
Set $n_1 \leftarrow 8\lceil \log(2\delta^{-1}) \rceil, n_2 \leftarrow \lceil \frac{8\log(32)}{\epsilon^2} \rceil$
if $k = 1$ **then**
  Return $g_t(\gamma)$ \\ *Return the reward evaluation*
end if
else
  for $s \leftarrow 1$ *to* n_1 \\ *Generate n_1 independent approximations of $Z_t^k(\gamma)$*
  do
    for $i \leftarrow 1$ *to* n_2 **do**
      $\mathbf{\Gamma}^{s,i} \leftarrow \mathcal{B}(t,\gamma)$ \\ *Generate n_2 independent sample paths of $\mathbf{Y} | Y_{[t]} = \gamma$*
    end
    $\tilde{Z}_t^{k,s}(\gamma) \leftarrow \mathcal{B}^{k-1}(t, \gamma, \frac{\epsilon}{2}, \frac{1}{8}) - \frac{1}{n_2}\sum_{i=1}^{n_2} \max_{j=1,\dots T} \mathcal{B}^{k-1}(j, \mathbf{\Gamma}^{s,i}_{[j]}, \frac{\epsilon}{4}, \frac{1}{16 n_2 T})$
  end
  Return $\text{med}(\tilde{Z}_t^{k,1}(\gamma), \dots, \tilde{Z}_t^{k,n_1}(\gamma))$ \\ *Return the median for the "median trick"*
end

Lemma 6. *Suppose that Assumptions 1 and 2 hold, with $U = 1$ in Assumption 1. Then for all $k \geq 1$, $t \in \{1,\dots,T\}$, $\gamma \in \mathcal{R}^{D\times t}$, $\epsilon, \delta \in (0,1)$, $\mathcal{B}^k(t,\gamma,\epsilon,\delta)$ returns a random number X satisfying $P(|X - Z_t^k(\gamma)| > \epsilon) \leq \delta$ in total computational time at most*

$$(C+G+1) \times \log(2\delta^{-1}) \times 10^{2k^2} \times \epsilon^{-2k} \times (T+2)^k \times \left(1 + \log\left(\frac{1}{\epsilon}\right) + \log(T)\right)^k.$$

We now describe algorithm $\mathcal{A}$, which intuitively stops the first time $Z_t^k \geq -\frac{1}{k-1}$, with $k = O\left(\frac{1}{\epsilon}\right)$ and Z_t^k approximated by a call to $\mathcal{B}^k$. Algorithm $\mathcal{A}$ takes as input accuracy parameter ϵ, and implements an ϵ-optimal (randomized) stopping policy, returning when to stop on the fly (given as input a sequentially revealed trajectory $\mathbf{\Gamma}$ of $\mathbf{Y}$ drawn from the appropriate distribution). The description of $\mathcal{A}$ is provided in Algorithm 2, and its performance guarantee is stated as Theorem 3. We again defer the proof to Section 3.2.

Algorithm 2 ($\mathcal{A}$ for Implementing a Stopping Policy τ_ϵ)

Input: ϵ, $\mathbf{\Gamma}$ (revealed sequentially with decisions made on the fly)
Set $k \leftarrow \lceil 4\epsilon^{-1} \rceil + 1$
for $t \leftarrow 1$ *to* T **do**
  $\tilde{Z}_t^k \leftarrow \mathcal{B}^k(t, \mathbf{\Gamma}_{[t]}, \frac{\epsilon}{4}, \frac{\epsilon}{4T})$ \\ *Compute approximation to* Z_t^k
  if $\tilde{Z}_t^k \geq -\frac{\epsilon}{2}$ **or** $t = T$ **then**
    Return STOP \\ *Stop when approximation to* Z_t^k *is near zero or* T *reached*
  end
end

Theorem 3 (Optimal Stopping Policy). *Suppose that Assumptions 1 and 2 hold, with* $U = 1$ *in Assumption 1. Algorithm* $\mathcal{A}$ *implements a randomized stopping time* τ_ϵ *s.t.* $E[Z_{\tau_\epsilon}] \geq \text{OPT} - \epsilon$. *In addition, at each time step, the decision of whether to stop (if one has not yet stopped) requires total computational time at most* $(C + G + 1) \times \exp(250\epsilon^{-2}) \times T^{12\epsilon^{-1}}$.

Remark 1 (Scaling of the Run Time in the Reward Bound U). We state and prove our main results under the assumption $U = 1$ (as opposed to general $U > 1$) only to enhance readability. The case of general $U > 1$ can be handled by running $\mathcal{A}$ on the corresponding sequence of normalized rewards $\tilde{Z}_t = \frac{Z_t}{U}$, with error parameter $\tilde{\epsilon} = \frac{\epsilon}{U}$. In the language of Theorem 3, this leads to a per-decision complexity of $(C + G + 1) \times \exp(250U^2\epsilon^{-2}) \times T^{12U\epsilon^{-1}} + 1$ to ensure $E[Z_{\tau_\epsilon}] \geq \text{OPT} - \epsilon$ for general $U > 1$. We refer the reader to the appendix, Section A.5, for a generalization of our algorithmic results to the setting in which rewards may be unbounded.

Remark 2 (Randomized Stopping Time). Because $\mathcal{A}$ is simulation based, the stopping strategy τ_ϵ is naturally randomized. Out of considerations of readability, we do not formally review the notion of randomized stopping time here, instead referring the reader to, for example, Solan et al. (2012).

Theorem 3 demonstrates an interesting contrast between our expansion-based approach and standard methods such as DP and naive nested simulation. To achieve a similar performance guarantee, those methods typically require an exponential-in-T run time. In contrast, our approach yields polynomial-in-T algorithms that can achieve any fixed-precision guarantee, in the precise sense that once ϵ is fixed, the exponent of T in the stated complexity bounds is a constant (depending on ϵ). However, let us point out that $\mathcal{A}$ is not a practically efficient algorithm, as is demonstrated in our numerical experiments (see the Online Companion). Our results thus serve as a proof of concept that a polynomial-time algorithm exists for complex OS problems. We leave the development of practical algorithms inspired by these theoretical insights as a direction for future research.

We end this section by noting that algorithm $\mathcal{A}$ also implies an efficient algorithm for approximating the optimal value with high probability (by generating random sample paths, implementing $\mathcal{A}$ on them, and averaging/taking the median). We provide a formal proof in the appendix, Section A.3.

Corollary 3 (Optimal Value Approximation). *Under the same assumptions as Theorem 3, there exists an algorithm* $\mathcal{A}'$ *that takes as input any* $\epsilon, \delta \in (0,1)$ *and achieves the following: in total computational time at most* $\lceil \log(2\delta^{-1}) \rceil \times (C + G + 1) \times \exp(1010\epsilon^{-2}) \times T^{24\epsilon^{-1}}$, *it returns a random number* X *satisfying* $P(|X - \text{OPT}| \leq \epsilon) \geq 1 - \delta$.

3.2. Analysis

We analyze $\mathcal{B}^k$ and $\mathcal{A}$ in this section, thereby proving Lemma 6 and Theorem 3.

Proof of Lemma 6. We proceed by induction. For the base case $k = 1$, we observe that $\mathcal{B}^1$ simply returns the appropriate value $g_t(\gamma) = Z_t^1(\gamma)$, completing the proof in the base case. Next, suppose for induction that the result holds for all $j \leq k-1$ for some $k \geq 2$. We now prove the result holds for k. We first show that $\mathcal{B}^k$ satisfies the desired high-probability error bounds, and begin by proving that in each of the n_1 outer loops, $\mathcal{B}^k$ computes a random number $\tilde{Z}_t^{k,s}(\gamma)$ satisfying $P(|\tilde{Z}_t^{k,s}(\gamma) - Z_t^k(\gamma)| > \epsilon) \leq \frac{1}{4}$. For $s \in \{1, \ldots, n_1\}$ and $i \in \{1, \ldots, n_2\}$, let X_i^s denote $\max_{j=1,\ldots T} \mathcal{B}^{k-1}(j, \mathbf{\Gamma}_{[j]}^{s,i}, \frac{\epsilon}{4}, \frac{1}{16n_2T})$ (as appears in the inner loop of $\mathcal{B}^k$), where we recall that $\mathbf{\Gamma}^{s,i}$ is an independently drawn trajectory distributed as $\mathbf{Y}|Y_t = \gamma$ for each s, i. Let $\bar{X}_i^s \triangleq \max_{j=1,\ldots,T} Z_j^k(\mathbf{\Gamma}_{[j]}^{s,i})$ denote the corresponding exact value of the maximum of $\mathbf{Z}^k$ on each of the trajectories $\Gamma^{s,i}$. Note that $|\bar{X}_i^s| \leq 1$ under Assumption 1 and by the

definition of $\mathbf{Z}^k$, and $E[\bar{X}_i^s] = E[\max_{j=1,\dots T} Z_j^k | Y_{[t]} = \gamma]$. Let us fix some $s \in \{1,\dots,n_1\}$. Then it follows from the induction hypothesis, the Lipschitzness of max, and a union bound that with probability at least $1 - \frac{1}{16}$, $|\bar{X}_i^s - X_i^s| < \frac{\epsilon}{4}$ for all $i = 1,\dots,n_2$. Recall that $n_2 = \lceil \frac{8\log(32)}{\epsilon^2} \rceil$. Applying Hoeffding's inequality (see Lemma A.1) to independent and identically distributed (i.i.d.) r.v.s. $\{\bar{X}_1^s,\dots,\bar{X}_{n_2}^s\}$, we know $|\frac{1}{n_2}\sum_{i=1}^{n_2} \bar{X}_i^s - E[\max_{j=1,\dots,T} Z_j^k | Y_{[t]} = \gamma]| < \frac{\epsilon}{4}$ with probability at least $1 - \frac{1}{16}$. Combining the above with a union bound and the triangle inequality, we conclude that

$$
\begin{aligned}
&P\left(\left|\frac{1}{n_2}\sum_{i=1}^{n_2} X_i^s - E\left[\max_{j=1,\dots,T} Z_j^k | Y_{[t]} = \gamma\right]\right| > \frac{\epsilon}{2}\right) \\
&\le P\left(\left|\frac{1}{n_2}\sum_{i=1}^{n_2} X_i^s - \frac{1}{n_2}\sum_{i=1}^{n_2} \bar{X}_i^s\right| > \frac{\epsilon}{4}\right) + P\left(\left|\frac{1}{n_2}\sum_{i=1}^{n_2} \bar{X}_i^s - E\left[\max_{j=1,\dots,T} Z_j^k | Y_{[t]} = \gamma\right]\right| > \frac{\epsilon}{4}\right), \\
&\le P\left(|X_i^s - \bar{X}_i^s| > \frac{\epsilon}{4} \text{ for some } i \in \{1,\dots,n_2\}\right) + P\left(\left|\frac{1}{n_2}\sum_{i=1}^{n_2} \bar{X}_i^s - E\left[\max_{j=1,\dots,T} Z_j^k | Y_{[t]} = \gamma\right]\right| > \frac{\epsilon}{4}\right),
\end{aligned}
$$

which, by our previous analysis, is at most $\frac{1}{16} + \frac{1}{16} = \frac{1}{8}$. Combining with the fact that $\mathcal{B}^{k-1}(t,\gamma,\frac{\epsilon}{2},\frac{1}{8})$ is an $\frac{\epsilon}{2}$-approximation of $Z_t^{k-1}(\gamma)$ with probability at least $1 - \frac{1}{8}$ by the induction hypothesis, along with another union bound and the definition of $Z_t^k = Z_t^{k-1} - E[\max_{j=1,\dots,T} Z_j^{k-1} | \mathcal{F}_t]$, we conclude that $|\tilde{Z}_t^{k,s}(\gamma) - Z_t^k(\gamma)| \le \epsilon$ with probability at least $1 - \frac{1}{4}$. This guarantee holds for each of the independent terms $\tilde{Z}_t^{k,s}(\gamma), s = 1,\dots,n_1$. Recall that $n_1 = 8\lceil \log(2\delta^{-1}) \rceil$. Then, applying the median trick (as formalized in Lemma A.2 of the appendix, Section A.4), we conclude that $P(|\mathrm{med}(\tilde{Z}_t^{k,1}(\gamma),\dots,\tilde{Z}_t^{k,n_1}(\gamma)) - Z_t^k(\gamma)| \le \epsilon) \ge 1 - \delta$, as desired. Combining the above, we deduce that $\mathcal{B}^k$ satisfies the desired high-probability error bounds.

We next focus on computational cost. To simplify notation, we denote the stated run-time guarantee (modulo the $C + G + 1$ multiplier) by

$$
f_k(\epsilon,\delta) \triangleq \log(2\delta^{-1}) \times 10^{2k^2} \times \epsilon^{-2k} \times (T+2)^k \times \left(1 + \log\left(\frac{1}{\epsilon}\right) + \log(T)\right)^k.
$$

For each fixed $s \in \{1,\dots,n_1\}$, in each of the inner iterations $i = 1,\dots,n_2$, first one call is made to $\mathcal{B}(t,\gamma)$ to generate $\mathbf{\Gamma}^{s,i}$ (at cost C); then T calls are made to $\mathcal{B}^{k-1}(j,\mathbf{\Gamma}_{[j]}^{s,i},\frac{\epsilon}{4},\frac{1}{16n_2T}), j = 1,\dots,T$, each at cost at most $(C+G+1)f_{k-1}(\frac{\epsilon}{4},\frac{1}{16n_2T})$ by the induction hypothesis; then the maximum of a length T array is computed (at computational cost T). One additional call is then made to $\mathcal{B}^{k-1}(t,\gamma,\frac{\epsilon}{2},\frac{1}{8})$, at cost at most $(C+G+1) \times f_{k-1}(\frac{\epsilon}{2},\frac{1}{8})$; the average of n_2 terms is computed and an additional subtraction is performed, at cost $n_2 + 1$. Finally, the median of n_1 numbers is computed at a cost of n_1. Combining with the induction hypothesis, the computational cost of $\mathcal{B}^k(t,\gamma,\epsilon,\delta)$ is thus at most

$$
\begin{aligned}
&n_1\left(n_2 C + n_2 T(C+G+1) f_{k-1}\left(\frac{\epsilon}{4},\frac{1}{16n_2T}\right) + n_2 T + (C+G+1) f_{k-1}\left(\frac{\epsilon}{2},\frac{1}{8}\right) + n_2 + 1 + 1\right) \\
&\le \; 2n_1 \times (C+G+1) \times (n_2+1) \times (T+2) \times f_{k-1}\left(\frac{\epsilon}{4},\frac{1}{16n_2T}\right) \\
&= \; 16\lceil \log(2\delta^{-1}) \rceil \times (C+G+1) \times \left(\left\lceil \frac{8\log(32)}{\epsilon^2} \right\rceil + 1\right) \times (T+2) \times f_{k-1}\left(\frac{\epsilon}{4},\frac{1}{16\lceil \frac{8\log(32)}{\epsilon^2} \rceil T}\right) \qquad (7)
\end{aligned}
$$

With some straightforward yet tedious algebra, it can be shown that

$$
(7) \text{ is at most } (C+G+1) f_k(\epsilon,\delta), \qquad (8)
$$

and we defer the proof of this fact to the appendix, Section A.4 (Lemma A.3). This completes the inductive proof for the run time. Combining the above completes the proof by induction. □

We now turn to Theorem 3. We begin with the following lemma relating the value achieved by a single stopping policy across different stopping problems (defined by $\mathbf{Z}^k$). The logic is essentially the same as that of Lemma 1, with the result following from our definitions, the optional stopping theorem, and a straightforward induction, and we omit the details.

Lemma 7. *For all (randomized) stopping times τ adapted to the (corresponding enlarged[1]) filtration $\mathcal{F}$, and all $k \geq 1$, $E[Z_\tau] = E[Z_\tau^k] + \sum_{i=1}^{k-1} E[\max_{t=1,\ldots,T} Z_t^i]$.*

For $k \geq 1$, let τ_k denote the stopping time that stops the first time that $Z_t^{k+1} \geq -\frac{1}{k}$ (i.e., the same stopping time used in the proof of Theorem 2), where by Lemma 4 such a time exists w.p.1. We now combine several of our past insights (some of which had shown closely related properties) to prove that τ_k is a nearly optimal stopping time.

Corollary 4. *Under the same assumptions as Theorem 3,* $\text{OPT} - E[Z_{\tau_k}] \leq \frac{1}{k} \ \forall\, k \geq 1$.

Proof. It follows from Lemma 1 that

$$\text{OPT} = \sum_{i=1}^{k} E\left[\max_{t=1,\ldots,T} Z_t^i\right] + \sup_{\tau \in \mathcal{T}} E[Z_\tau^{k+1}].$$

As Lemma 7 implies

$$\sum_{i=1}^{k} E\left[\max_{t=1,\ldots,T} Z_t^i\right] = E[Z_{\tau_k}] - E[Z_{\tau_k}^{k+1}],$$

we conclude that

$$\text{OPT} - E[Z_{\tau_k}] = \sup_{\tau \in \mathcal{T}} E[Z_\tau^{k+1}] - E[Z_{\tau_k}^{k+1}].$$

As the nonpositivity of Z_t^k shown in Lemma 1 implies $\sup_{\tau \in \mathcal{T}} E[Z_\tau^{k+1}] \leq 0$, and the definition of τ_k implies $-E[Z_{\tau_k}^{k+1}] \leq \frac{1}{k}$, combining the above completes the proof. □

With Corollary 4 in hand, we now complete the proof of Theorem 3.

Proof of Theorem 3. It is easily verified that for any $\epsilon \in (0,1)$, the stopping time τ_ϵ described in Algorithm 2 is a well-defined, appropriately adapted, randomized stopping time. Recall $k \triangleq \lceil \frac{4}{\epsilon} \rceil + 1$. Let $\mathcal{G}_{1,\epsilon}$ denote the event

$$\left\{ \left| \mathcal{B}^k\left(t, \Gamma_{[t]}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}\right) - Z_t^k(\Gamma_{[t]}) \right| \leq \frac{\epsilon}{4} \ \ \forall t \in \{1,\ldots,T\} \right\},$$

$\mathcal{G}_{2,\epsilon}$ denote the event

$$\left\{ \exists t \in \{1,\ldots,T\} \text{ such that } \mathcal{B}^k\left(t, \Gamma_{[t]}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}\right) \geq -\frac{1}{2}\epsilon \right\},$$

and $\mathcal{G}_{3,\epsilon}$ denote the event $\left\{ Z_{\tau_\epsilon}^k \geq -\frac{3}{4}\epsilon \right\}$. We now prove several straightforward results about the probabilities of these events. First, observe that Lemma 6 and a union bound implies

$$P(\mathcal{G}_{1,\epsilon}) \geq 1 - \frac{\epsilon}{4}. \tag{9}$$

Second, we claim that

$$P(\mathcal{G}_{2,\epsilon} | \mathcal{G}_{1,\epsilon}) = 1. \tag{10}$$

Indeed, as $k = \lceil \frac{4}{\epsilon} \rceil + 1$, and Lemma 4 implies that w.p.1 $\exists t \in \{1,\ldots,T\}$ such that $Z_t^k \geq -\frac{1}{k-1}$, we conclude that w.p.1 $\exists t \in \{1,\ldots,T\}$ such that $Z_t^k \geq -\frac{\epsilon}{4}$. Let t' denote the smallest such time. Conditioning on $\mathcal{G}_{1,\epsilon}$ ensures that $|\mathcal{B}^k(t', \Gamma_{[t']}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}) - Z_{t'}^k(\Gamma_{[t']})| \leq \frac{\epsilon}{4}$. It then follows from the triangle inequality that $\mathcal{B}^k(t', \Gamma_{[t']}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}) \geq -\frac{\epsilon}{2}$, and hence $\exists t \in \{1,\ldots,T\}$ such that $\mathcal{B}^k(t, \Gamma_{[t]}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}) \geq -\frac{1}{2}\epsilon$. Combining the above completes the proof of (10).

Third, we claim that

$$P(\mathcal{G}_{3,\epsilon} | \mathcal{G}_{1,\epsilon} \cap \mathcal{G}_{2,\epsilon}) = 1. \tag{11}$$

Indeed, note that conditional on $\mathcal{G}_{2,\epsilon}$, $\exists t \in \{1,\ldots,T\}$ such that $\mathcal{B}^k(t, \Gamma_{[t]}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}) \geq -\frac{1}{2}\epsilon$. It follows from the definition of τ_ϵ that, conditional on $\mathcal{G}_{2,\epsilon}$, we have $\mathcal{B}^k(\tau_\epsilon, \Gamma_{[\tau_\epsilon]}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}) \geq -\frac{1}{2}\epsilon$. But as we also condition on $\mathcal{G}_{1,\epsilon}$, it follows that

$|\mathcal{B}^k(\tau_\epsilon, \Gamma_{[\tau_\epsilon]}, \frac{\epsilon}{4}, \frac{\epsilon}{4T}) - Z^k_{\tau_\epsilon}(\Gamma_{[\tau_\epsilon]})| \le \frac{\epsilon}{4}$. Combining with an application of the triangle inequality completes the proof of (11).

Combining (9)–(11), it follows that $P(\mathcal{G}^c_{3,\epsilon}) \le \frac{\epsilon}{4}$. Further combining with the fact that our assumptions and Lemma 1 imply $P(Z^k_t \in [-1, 0]) = 1$ for all $t \in \{1, \dots, T\}$, we conclude that

$$\begin{aligned} E[Z^k_{\tau_\epsilon}] &= E[Z^k_{\tau_\epsilon} I(\mathcal{G}_{3,\epsilon})] + E[Z^k_{\tau_\epsilon} I(\mathcal{G}^c_{3,\epsilon})] \\ &\ge -\frac{3}{4}\epsilon - E[I(\mathcal{G}^c_{3,\epsilon})] \ge -\epsilon, \end{aligned}$$

where the inequality $E[Z^k_{\tau_\epsilon} I(\mathcal{G}_{3,\epsilon})] \ge -\frac{3}{4}\epsilon$ follows from the definition of $\mathcal{G}_{3,\epsilon}$. Combining with Lemma 7, we conclude that $E[Z_{\tau_\epsilon}] \ge -\epsilon + \sum_{i=1}^{k-1} E[\max_{t=1,\dots,T} Z^i_t]$. As Theorem 1 and the nonpositivity of Z^k_t shown in Lemma 1 imply $\sum_{i=1}^{k-1} E[\max_{t=1,\dots,T} Z^i_t] \ge \text{OPT}$, combining the above completes the proof of the first part of the theorem. As for the computational cost, the algorithm in each time step makes one call to $\mathcal{B}^{\lceil 4\epsilon^{-1}\rceil+1}$ to compute an $\frac{\epsilon}{4}$−approximation with probability at least $1 - \frac{\epsilon}{4T}$. The desired result then follows from Lemma 6 along with some straightforward algebra, and we omit the details. □

4. Conclusion

In this work, we proved the existence of a polynomial-time (in the horizon T and dimension D) algorithm for complex OS problems with fixed accuracy and access to an efficient simulator. The algorithms come with provable performance guarantees under either boundedness or mild concentration assumptions. Although their complexity is polynomial, it is impractical, as was the case for the first polynomial-time algorithms for other problems in the literature. Our work leaves many interesting directions for future research.

4.1. The Design of Practical Algorithms with Polynomial Run-Time Guarantees

A natural question is whether there exists a practical algorithm with polynomial run-time guarantees. One possible approach is to develop different expansions for the OS value that converge more rapidly, and/or to identify natural assumptions under which such expansions converge more rapidly. Such approaches could also be combined with techniques from reinforcement learning, ADP, simulation (including multilevel Monte Carlo), and parallel computing.

4.2. Generalization to High-Dimensional Online Decision Making Broadly and Other Applications

We believe that our methodology can be extended to a broad family of high-dimensional online decision-making problems in economics, mechanism design/prophet inequalities, finance, statistics, machine learning, operations, and computer science. Preliminary results along these lines for certain problems in multiple stopping, dynamic pricing, and online combinatorial optimization appear in Chen (2021).

4.3. Lower Bounds and Computational Complexity

An interesting set of questions revolve around proving lower bounds on the computational and sample complexity for the problems studied, for example, path-dependent OS. There has been much interesting recent work laying out a theory of computational complexity (with positive and negative results) in the settings of stochastic control and reinforcement learning (Shapiro and Nemirovski 2005; Halman et al. 2014, 2015; Chen and Wang 2017; Sidford et al. 2018; Du et al. 2019; Wang et al. 2021). Better understanding the connection between our approach and those works, as well as the many existing models for online optimization, remains an interesting direction for future research. Furthermore, the question of what quality of approximation can be (efficiently) achieved with a given depth of nesting remains a very interesting open question. Such questions relate to recent studies on the power of adaptivity in optimization (e.g., Balkanski et al. 2017), as well as the literature on parallel and quantum computing, (multistage stochastic) convex optimization, and prophet inequalities.

Acknowledgments

The authors gratefully acknowledge Kaidong Zhang and Di Wu, for their tremendous help with a numerical implementation, and Bobby Kleinberg, for pointing out that the median trick could be applied in our setting. The authors also thank the organizers and participants of the 2018 Symposium on Optimal Stopping in honor of Larry Shepp held at Rice University.

Appendix

A.1. Proof of Lemma 3

Let us consider the following simple example. Suppose the dimension $D = 1$, the horizon $T = 3$, $Y_1 = 1$ w.p.1, $Y_2 = \frac{1}{2}$ w.p.1, Y_3 equals zero w.p. $\frac{1}{2}$ and equals one w.p. $\frac{1}{2}$, and the stopping problem is simply $\sup_{\tau\in\mathcal{T}} E[Y_\tau]$, that is, $Z_t = Y_t$ for $t = 1,2,3$. We note that in this simple example, the state space of the underlying stochastic process $\mathbf{Y}$ consists of two elements: $(1,\frac{1}{2},0)$ and $(1,\frac{1}{2},1)$. As $\max(Z_1,Z_2,Z_3) = 1$ w.p.1, it follows that $Z_1^2 = 0$ w.p.1, $Z_2^2 = -\frac{1}{2}$ w.p.1, $Z_3^2(1,\frac{1}{2},0) = -1$, and $Z_3^2(1,\frac{1}{2},1) = 0$. A straightforward induction further implies that $Z_1^k = 0$ w.p.1, $Z_2^k = -\frac{1}{2}$ w.p.1, $Z_3^k(1,\frac{1}{2},0) = -1$, and $Z_3^k(1,\frac{1}{2},1) = 0$ for all $k \geq 2$. Thus, by (6) and a straightforward calculation, we conclude that $MAR_t^0 = 0$ w.p.1 for all t. Reasoning about the reward-to-go functions, it is easily verified that $V_1 = 1$ w.p.1, $V_2 = \frac{1}{2}$ w.p.1, and $V_3 = Z_3$ w.p.1. As $MAR_t^0 = 0$ w.p.1 for all t, we conclude from Definition 1 that MAR^0 is 0-surely optimal iff, w.p.1, $\max(Z_1,Z_3,Z_3) = 1$, which is indeed the case. However, for MAR^0 to be 1-surely optimal it would have to hold that, w.p.1, $\max(Z_2,Z_3) = V_2 = \frac{1}{2}$, which is clearly not the case, as $P(Z_3 = 1) > 0$. □

A.2. Proof of Lemma 5

We explicitly construct the hard instances, parametrized by $k \geq 1$. We define hard instance k (whose optimal value we denote by opt^k) as follows. In hard instance k, $Z_1 = \frac{k}{k+1}$ w.p.1, and Z_2 is a Bernoulli r.v. with $P(Z_2 = 1) = \frac{k}{k+1}$, $P(Z_2 = 0) = \frac{1}{k+1}$, with the natural filtration generated by the reward process. We now compute the recursively defined $\mathbf{Z}^j$ for this problem, beginning with $j = 2$. Because Z_1 is constant and thus $\mathcal{F}_1$ is the trivial σ-field, we have that $Z_1^2 = Z_1 - E[\max_{i=1,2} Z_i] = \frac{k}{k+1}(\frac{k}{k+1} - 1) + \frac{1}{k+1} \times 0 = -\frac{k}{(k+1)^2}$. A straightforward calculation yields that $Z_2^2 = Z_2 - \max_{i=1,2} Z_i$ is a r.v. taking one of two values: $-\frac{k}{k+1}$ (w.p. $\frac{1}{k+1}$) or zero (w.p. $1 - \frac{1}{k+1}$). Noting that $\mathbf{Z}$ is itself a martingale, and the difference of martingales is a martingale, we find (from the recursive definition of $\mathbf{Z}^j$) that $\mathbf{Z}^j$ is a martingale for all j. It then follows from our recursive definition of $\mathbf{Z}^j$ and a straightforward induction (the details of which we omit) that for all $j \geq 2$, $Z_1^{j+1} = (1 - \frac{1}{k+1}) Z_1^j$, and Z_2^{j+1} is a r.v. taking one of two values: $(k+1) \times Z_1^{j+1}$ (w.p. $\frac{1}{k+1}$) or zero (w.p. $1 - \frac{1}{k+1}$). A straightforward calculation then yields that $Z_1^j = -(1 - \frac{1}{k+1})^{j-2} \times \frac{k}{(k+1)^2}$ for all $j \geq 2$. By Lemma 1, $E_k - opt^k = -\sup_{\tau\in\mathcal{T}}[Z_\tau^{k+1}]$. Because $\mathbf{Z}^{k+1}$ is a martingale (as $\mathbf{Z}^j$ is a martingale for all j), it follows that $-\sup_{\tau\in\mathcal{T}}[Z_\tau^{k+1}] = -E[Z_1^{k+1}] = (1 - \frac{1}{k+1})^{k-1} \times \frac{k}{(k+1)^2} = (1 - \frac{1}{k+1})^k \times \frac{1}{k+1}$. Taking limits completes the proof. □

A.3. Proof of Corollary 3

In this section, we prove Corollary 3.

Proof of Corollary 3. Consider the following algorithm $\mathcal{A}'$, with positive integer parameters n_1, n_2 (which we will later select as appropriate functions of ϵ, δ). For each $i \in \{1,\ldots,n_1\}$, algorithm $\mathcal{A}'$ generates n_2 independent unconditioned trajectories by calling $\mathcal{B}(0)$ n_2 times, and runs algorithm $\mathcal{A}$ (with parameter $\frac{\epsilon}{2}$) on each of these trajectories. This yields $n_1 \times n_2$ values $Z^{i,j}_{\tau_{\frac{1}{2}\epsilon}}$ (for $i = 1,\ldots,n_1$, $j = 1,\ldots,n_2$), representing the values returned by the calls to $\mathcal{A}$. We note that these values are i.i.d. Algorithm $\mathcal{A}'$ then returns $\text{med}(\frac{1}{n_2}\sum_{j=1}^{n_2} Z^{1,j}_{\tau_{\frac{1}{2}\epsilon}}, \ldots, \frac{1}{n_2}\sum_{j=1}^{n_2} Z^{n_1,j}_{\tau_{\frac{1}{2}\epsilon}})$. It follows from Hoeffding's inequality (see Lemma A.1 in Section A.4) that for each $i \in \{1,\ldots,n_1\}$, $P(|\frac{1}{n_2}\sum_{j=1}^{n_2} Z^{i,j}_{\tau_{\frac{1}{2}\epsilon}} - E[Z_{\tau_{\frac{1}{2}\epsilon}}]| \geq \frac{1}{2}\epsilon) \leq \exp(-\frac{1}{2}n_2\epsilon^2)$. Thus, so long as $n_2 \geq 2\log(4)\epsilon^{-2}$, it will hold that (for each $i \in \{1,\ldots,n_1\}$)

$$P\left(\left|\frac{1}{n_2}\sum_{j=1}^{n_2} Z^{i,j}_{\tau_{\frac{1}{2}\epsilon}} - E\left[Z_{\tau_{\frac{1}{2}\epsilon}}\right]\right| \leq \frac{1}{2}\epsilon\right) \geq \frac{3}{4}.$$

We may thus apply the median trick (see Lemma A.2 in Section A.4) to conclude that for $n_1 = 8\lceil\log(2\delta^{-1})\rceil$,

$$P\left(\left|\text{med}\left(\frac{1}{n_2}\sum_{j=1}^{n_2} Z^{1,j}_{\tau_{\frac{1}{2}\epsilon}}, \ldots, \frac{1}{n_2}\sum_{j=1}^{n_2} Z^{n_1,j}_{\tau_{\frac{1}{2}\epsilon}}\right) - E\left[Z_{\tau_{\frac{1}{2}\epsilon}}\right]\right| \leq \frac{1}{2}\epsilon\right) \geq 1 - \delta.$$

Combining with Theorem 3 (which ensures $E[Z_{\tau_{\frac{1}{2}\epsilon}}] \geq \text{OPT} - \frac{1}{2}\epsilon$) and the triangle inequality implies the desired performance guarantee. The complexity analysis follows directly from the complexity guarantee of Theorem 3 after accounting for the $n_1 \times n_2$ iterations (and additional time for the averaging and taking the median). Combining the above with some straightforward algebra completes the proof. □

A.4. Auxiliary Technical Lemmas for Earlier Proofs

In this section, we review several auxiliary technical lemmas used in earlier proofs, including Hoeffding's inequality, the median trick, and a property of the functions f_k.

We first recall Hoeffding's inequality, a standard result from probability theory used often to prove concentration for estimators.

Lemma A.1 (Hoeffding's Inequality). *Suppose that for some $n \geq 1$ and $U > 0$, $\{X_i, i = 1, \ldots, n\}$ are i.i.d., and either $P(X_1 \in [0, U]) = 1$ or $P(X_1 \in [-U, 0]) = 1$. Then, for all $\eta > 0$, $P(|n^{-1}\sum_{i=1}^{n} X_i - E[X_1]| \geq \eta) \leq 2\exp(-\frac{2\eta^2 n}{U^2})$.*

We next recall the median trick—a direct corollary of Hoeffding's inequality that is commonly used in speeding up algorithms for estimating a number (see, e.g., Wang and Han 2015) and which goes back at least to the work of Nemirovski and Yudin (1983). For completeness, and as many variants of the core idea appear throughout the literature, below we state and prove a specific result along these lines.

Lemma A.2 (Median Trick). *Suppose there is a given deterministic real number X, and a randomized* algorithm A *with the following property: the (random) output $\hat{X}$ of A satisfies $P(|X - \hat{X}| \leq \epsilon) \geq \frac{3}{4}$. Consider a new algorithm A′ constructed as follows: Repeat the* algorithm A *for $m = 8\lceil \log(2\delta^{-1}) \rceil$ independent trials (with any random numbers used by A independent across trials), yielding outputs $\{\hat{X}_i, i = 1, \ldots, m\}$, and output the median X' of $\{\hat{X}_i, i = 1, \ldots, m\}$.*[2] *Then $P(|X - X'| \leq \epsilon) \geq 1 - \delta$.*

Proof of Lemma A.2. Suppose the m outputs from the repetitions of A (when running A') are $q_1, \ldots, q_m$ (possibly nondistinct) and that, without loss of generality, these are nondecreasing (i.e., $q_1 \leq \cdots \leq q_m$). For $i \in \{1, \ldots, m\}$, let $Z_i \triangleq I(|q_i - X| \leq \epsilon)$. It may be easily verified from our definition of median that for any $j \in \{1, \ldots, \lfloor \frac{m+1}{2} \rfloor\}$, the interval $[q_j, q_{j+\lfloor \frac{m}{2} \rfloor}]$ must contain the median of the sequence. Also, the event $\{\sum_{i=1}^{m} Z_i > \frac{m}{2}\}$ implies that there exists at least $\lceil \frac{m+1}{2} \rceil$ elements in $\{q_1, .., q_m\}$ that satisfy $|q_i - X| \leq \epsilon$. Conditional on this event, let u denote the index of the smallest of these (in the order $q_1 \leq \cdots \leq q_m$) and v the index of the largest. It may be easily verified that (conditional on that event) $v - u \geq \lfloor \frac{m}{2} \rfloor$, and $u \leq \lfloor \frac{m+1}{2} \rfloor$. Combining with our previous observation about the median, it follows that in this case, $\text{median}(q_1, \ldots, q_m) \in [q_u, q_v]$, and therefore, $|\text{median}(q_1, \ldots, q_m) - X| \leq \epsilon$. Taking the contraposition of the above reasoning, we conclude that $\{|\text{median}(q_1, \ldots, q_m) - X| > \epsilon\}$ implies $\{\sum_{i=1}^{m} Z_i \leq \frac{m}{2}\}$, and thus

$$
\begin{aligned}
P(|X' - X| > \epsilon) &\leq P\left(\frac{1}{m}\sum_{i=1}^{m} Z_i \leq \frac{1}{2}\right) \\
&= P\left(\frac{1}{m}\sum_{i=1}^{m} Z_i - E[Z_1] \leq \frac{1}{2} - E[Z_1]\right) \\
&\leq P\left(\frac{1}{m}\sum_{i=1}^{m} Z_i - E[Z_1] \leq -\frac{1}{4}\right) \text{ because } E[Z_1] \geq \frac{3}{4} \\
&\leq P\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - E[Z_1]\right| \geq \frac{1}{4}\right).
\end{aligned}
$$

Note that $\{Z_i, i = 1, \ldots, m\}$ viewed as an unordered set has the same joint distribution as m i.i.d. r.v.s., each distributed as the indicator of whether a single output of A is within ϵ of X in absolute value. Thus, we may apply Hoeffding's inequality with $U = 1, n = m = 8\lceil \log(2\delta^{-1}) \rceil$, and $\eta = \frac{1}{4}$ to conclude that

$$
P(|X' - X| > \epsilon) \leq 2\exp\left(-\frac{m}{8}\right) \leq \delta,
$$

thus completing the proof. □

Recall our definition of function f_k in Section 3.2:

$$
f_k(\epsilon, \delta) \triangleq \log(2\delta^{-1}) \times 10^{2k^2} \times \epsilon^{-2k} \times (T + 2)^k \times \left(1 + \log\left(\frac{1}{\epsilon}\right) + \log(T)\right)^k.
$$

We next prove that the functions $\{f_k, k \geq 1\}$ satisfy a certain recursive relationship, which we used in the proof of Lemma 6 (specifically in Inequality (8)).

Lemma A.3. *For all $\epsilon, \delta \in (0, 1)$ and $k \geq 1$,*

$$
f_{k+1}(\epsilon, \delta) \geq 16\lceil \log(2\delta^{-1}) \rceil \times \left(\left\lceil \frac{8\log(32)}{\epsilon^2} \right\rceil + 1\right) \times (T + 2) \times f_k\left(\frac{\epsilon}{4}, \frac{1}{16\lceil \frac{8\log(32)}{\epsilon^2} \rceil T}\right),
$$

from which Inequality (8) follows after multiplying both sides by $C + G + 1$.

Proof of Lemma A.3. Combining definitions, the monotonicity of $f_k(\epsilon,\delta)$ in ϵ,δ,η on $(0,1)$, and some straightforward algebra (including the facts that $\lceil x\rceil \le 2x$ for $x \ge \log(2)$, $x+1 \le 2x$ for $x \ge 1$, and $8x+2 \le 10x$ for $x \ge 1$), we have that

$$
\begin{aligned}
&16\lceil\log(2\delta^{-1})\rceil \times \left(\left\lceil \frac{8\log(32)}{\epsilon^2}\right\rceil + 1\right) \times (T+2) \times f_k\left(\frac{\epsilon}{4}, \frac{1}{16\lceil \frac{8\log(32)}{\epsilon^2}\rceil T}\right)\\
&\le 16\lceil\log(2\delta^{-1})\rceil \times \left(8\log(32)\epsilon^{-2}+2\right) \times (T+2) \times f_k\left(\frac{\epsilon}{4}, \frac{1}{16\lceil \frac{8\log(32)}{\epsilon^2}\rceil T}\right)\\
&\le (32\log(2\delta^{-1})) \times (10\log(32)\epsilon^{-2}) \times (T+2) \times f_k\left(\frac{\epsilon}{4}, \frac{\epsilon^2}{256\log(32)T}\right)\\
&= (32\log(2\delta^{-1})) \times (10\log(32)\epsilon^{-2}) \times (T+2)\\
&\quad \times \log\left(512\log(32)T\epsilon^{-2}\right) \times 10^{2k^2} \times 4^{2k} \times \epsilon^{-2k} \times (T+2)^k \times \left(1+\log\left(\frac{4}{\epsilon}\right)+\log(T)\right)^k\\
&\le \log\left(2\delta^{-1}\right) \times (T+2)^{k+1} \times \epsilon^{-2(k+1)} \times \left(1+\log\left(\frac{1}{\epsilon}\right)+\log(T)\right)^{k+1} \times 10^{2k^2} \times 4^{2k} \times 10^4 \times 4^k\\
&\quad \text{because } \log\left(512\log(32)T\epsilon^{-2}\right) \le \log\left(512\log(32)\right)\left(1+\log\left(\frac{1}{\epsilon}\right)+\log(T)\right), 1+\log\left(\frac{4}{\epsilon}\right)+\log(T) \le 4\left(1+\log\left(\frac{1}{\epsilon}\right)+\log(T)\right), \text{and}\\
&\quad 32\times 10\log(32)\times\log(512\log(32)) \le 10^4\\
&\le \log\left(2\delta^{-1}\right) \times 10^{2(k+1)^2} \times (T+2)^{k+1} \times \epsilon^{-2(k+1)} \times \left(1+\log\left(\frac{1}{\epsilon}\right)+\log(T)\right)^{k+1},
\end{aligned}
$$

the final inequality following from the fact that

$$
\begin{aligned}
&10^{2k^2} \times 4^{2k} \times 10^4 \times 4^k\\
&= 10^{2k^2+4} \times 4^{3k} \le 10^{2k^2+4} \times 10^{2k} = 10^{2(k+1)^2-2k+2}\\
&\le 10^{2(k+1)^2} \quad \text{for} \quad k \ge 1. \ \square
\end{aligned}
$$

A.5. Extension: Unbounded Rewards

In this section, we generalize our results to the setting in which rewards may be unbounded, instead imposing the following assumption.

Assumption A.1. *Let $M \triangleq \max_{t=1,\dots,T} Z_t$. Assume that $Z_t \ge 0$ w.p.1 for all t, and that the squared coefficient of variation of M, $\frac{E[M^2]}{(E[M])^2} - 1$, is upper bounded by a constant independent of both the time horizon T and the dimension D. We let $\gamma_0 \triangleq \frac{E[M^2]}{(E[M])^2}$.*

To handle this unbounded setting, we truncate the reward process, prove that this truncation still yields a good approximation of the original problem, and then apply our previous approach and algorithms. Our main result for the unbounded case is the following.

Theorem A.1 (Optimal Stopping Policy for Unbounded Rewards). *Suppose that Assumptions 2 and A.1 hold. Then algorithm $\mathcal{A}$ (with properly chosen parameters and an additional truncation step) implements a randomized stopping time τ_ϵ s.t. $\frac{E[Z_{\tau_\epsilon}]}{\text{OPT}} \ge 1-\epsilon$. In addition, at each time step, the decision of whether to stop (if one has not yet stopped) requires total computational time at most $(C+G+4) \times \exp(1.6\times 10^9 \times \gamma_0^8 \times \epsilon^{-6}) \times T^{3\times 10^4 \times \gamma_0^4 \times \epsilon^{-3}}$.*

As in the bounded case, Theorem A.1 also implies an efficient algorithm for approximating the optimal value with high probability (by generating random sample paths, implementing $\mathcal{A}$ on the truncated problem, and averaging/taking the median). The statement and proof are completely analogous to those of Corollary 3, and we omit the statement and details.

A.5.1. Proof of Theorem A.1

Let us begin by introducing some additional notation. For any $U>0$, define $Z^1_{U,t} \triangleq U^{-1} \times \min(U, Z^1_t)$. For $k \ge 1$, let $Z^{k+1}_{U,t} \triangleq Z^k_{U,t} - E[\max_{i=1,\dots,T} Z^k_{U,i} | \mathcal{F}_t]$, $L_{U,k} \triangleq E[\max_{i=1,\dots,T} Z^k_{U,i}]$, and $E_{U,k} \triangleq \sum_{i=1}^k L_{U,i}$. Let $M_1 = E[\max_{t=1,\dots,T} Z^1_t]$ and $M_2 = E[(\max_{t=1,\dots,T} Z^1_t)^2]$. Then γ_0, that is, the constant specified in Assumption A.1, can be represented as $\gamma_0 = \frac{M_2}{(M_1)^2}$. Before diving into the proof of Theorem A.1, we first state and prove two lemmas. The next result bounds the error induced (in the optimal stopping value) by truncating the rewards.

Lemma A.4. *Suppose $P(Z_t \ge 0) = 1$ for all t. Then, for all $U>0$,*

$$
0 \le \text{OPT} - U \times \sup_{\tau \in \mathcal{T}} E[Z^1_{U,\tau}] \le (M_2)^{\frac{1}{2}} \times \left(\frac{M_1}{U}\right)^{\frac{1}{2}}.
$$

Proof. The first inequality follows from the fact that, w.p.1, $Z^1_t \geq U \times Z^1_{U,t}$ for all $t \in \{1,\dots,T\}$. To prove the second inequality, let τ_* denote an optimal stopping time for the problem $\sup_{\tau\in\mathcal{T}} E[Z^1_\tau]$, where existence follows from Chow et al. (1971). Then, by a straightforward coupling and rescaling,

$$\begin{aligned}
E[Z^1_{\tau_*}] - U \times E[Z^1_{U,\tau_*}] &\leq E[Z^1_{\tau_*} I(Z^1_{\tau_*} > U)] \\
&\leq E\left[Z^1_{\tau_*} I\left(\max_{t=1,\dots,T} Z^1_t > U\right)\right] \\
&\leq E\left[\left(\max_{t=1,\dots,T} Z^1_t\right) \times I\left(\max_{t=1,\dots,T} Z^1_t > U\right)\right] \\
&\leq (M_2)^{\frac{1}{2}} \times \left(P\left(\max_{t=1,\dots,T} Z^1_t > U\right)\right)^{\frac{1}{2}} \text{ by Cauchy–Schwarz} \\
&\leq (M_2)^{\frac{1}{2}} \times \left(\frac{M_1}{U}\right)^{\frac{1}{2}} \text{ by Markov's inequality.}
\end{aligned}$$

Noting that $\sup_{\tau\in\mathcal{T}} E[Z^1_{U,\tau}] \geq E[Z^1_{U,\tau_*}]$ completes the proof. □

The next result shows that OPT may be bounded (from below) as a function of M_1 and γ_0.

Lemma A.5. $\text{OPT} \geq \frac{4}{27} \times \gamma_0^{-1} \times M_1$.

Proof. Recall the celebrated Paley–Zygmund inequality, that is, the fact that for any $\delta \in (0,1)$ and nonnegative r.v. X,

$$P(X > \delta E[X]) \geq (1-\delta)^2 \times \frac{(E[X])^2}{E[X^2]}. \tag{A.1}$$

Now, for $\delta \in (0,1)$, consider the stopping time τ_δ, which stops the first time that $Z^1_t \geq \delta \times M_1$, and stops at time T if no such time exists in $\{1,\dots,T\}$. Then, by nonnegativity and (A.1),

$$E[Z^1_{\tau_\delta}] \geq \delta \times (1-\delta)^2 \times \frac{(M_1)^3}{M_2}.$$

Optimizing over δ (a straightforward exercise in calculus) then completes the proof. □

We next complete the proof of Theorem A.1.

Proof of Theorem A.1. For any $U > 0$ and $\epsilon \in (0,1)$, we may use algorithm $\mathcal{A}$ (as in Theorem 3) to implement a randomized stopping time $\tau_{U,\epsilon}$ s.t. $E[Z^1_{U,\tau_{U,\epsilon}}] \geq \sup_{\tau\in\mathcal{T}} E[Z^1_{U,\tau}] - \epsilon$. Because $\tau_{U,\epsilon} \in \mathcal{T}$, it also yields a stopping time for the original problem $\sup_{\tau\in\mathcal{T}} E[Z^1_\tau]$. By construction, for all $t \geq 1$, w.p.1 $Z^1_t \geq U \times Z^1_{U,t}$. Combining the above, it follows that for any $U > 0$ and $\epsilon \in (0,1)$,

$$\begin{aligned}
E[Z^1_{\tau_{U,\epsilon}}] &\geq U \times E[Z^1_{U,\tau_{U,\epsilon}}] \\
&\geq U \times \sup_{\tau\in\mathcal{T}} E[Z^1_{U,\tau}] - U \times \epsilon.
\end{aligned}$$

Combining with Lemma A.4, the definition of γ_0, and the triangle inequality, we conclude that for all $U > 0$ and $\epsilon' \in (0,1)$,

$$\text{OPT} - E[Z^1_{\tau_{U,\epsilon'}}] \leq \left(\frac{\gamma_0 M_1^3}{U}\right)^{\frac{1}{2}} + U \times \epsilon'. \tag{A.2}$$

For $\epsilon > 0$, let $U^\epsilon \triangleq \left(\frac{27}{2}\right)^2 \times \gamma_0^3 \times \epsilon^{-2} \times M_1$ and $\epsilon' = \epsilon^3 \times \frac{8}{27^3} \times \gamma_0^{-4}$. It then follows from (A.2), Lemma A.5, and some straightforward algebra that for any $\epsilon \in (0,1)$,

$$\text{OPT} - E[Z^1_{\tau_{U^\epsilon,\epsilon'}}] \leq \epsilon \times \text{OPT}.$$

Combining with Theorem 3, and accounting for the extra time needed to truncate and normalize the rewards, completes the proof. □

Endnotes

[1] See Solan et al. (2012) for a formal discussion of randomized stopping times and corresponding enlarged filtrations.

[2] Here, by default, we define the median of an ordered sequence $x_1 \leq \cdots \leq x_n$ (with n even or odd) as $\frac{1}{2}(x_{\lfloor\frac{n+1}{2}\rfloor} + x_{\lfloor\frac{n}{2}\rfloor+1})$.

References

Andersen L, Broadie M (2004) Primal-dual simulation algorithm for pricing multidimensional American options. *Management Sci.* 50(9):1222–1234.

Balkanski E, Rubinstein A, Singer Y (2017) The limitations of optimization from samples. *Proc. 49th Annual ACM SIGACT Sympos. Theory Comput.* (Association for Computing Machinery, New York), 1016–1027.

Bayer C, Redmann M, Schoenmakers J (2021) Dynamic programming for optimal stopping via pseudo-regression. *Finance* 21(1):29–44.
Becker S, Cheridito P, Jentzen A (2019) Deep optimal stopping. *J. Machine Learn. Res.* 20(74):1–25.
Becker S, Cheridito P, Jentzen A, Welti T (2021) Solving high-dimensional optimal stopping problems using deep learning. *Eur. J. Appl. Math.* 32(3):470–514.
Belomestny D (2011) On the rates of convergence of simulation-based optimization algorithms for optimal stopping problems. *Ann. Appl. Probab.* 21(1):215–239.
Belomestny D, Milstein GN (2006) Monte Carlo evaluation of American options using consumption processes. *Internat. J. Theoret. Appl. Finance* 9(4):455–481.
Belomestny D, Bender C, Schoenmakers J (2009) True upper bounds for Bermudan products via non-nested Monte Carlo. *Math. Finance* 19(1):53–71.
Belomestny D, Hildebrand R, Schoenmakers J (2019) Optimal stopping via pathwise dual empirical maximisation. *Appl. Math. Optim.* 79(3):715–741.
Belomestny D, Ladkau M, Schoenmakers J (2015) Multilevel simulation based policy iteration for optimal stopping—Convergence and complexity. *SIAM/ASA J. Uncertainty Quantification* 3(1):460–483.
Belomestny D, Schoenmakers J, Dickmann F (2013) Multilevel dual approach for pricing American style derivatives. *Finance Stochastics* 17(4):717–742.
Belomestny D, Schoenmakers J, Spokoiny V, Zharkynbay B (2020) Optimal stopping via deeply boosted backward regression. *Comm. Math. Sci.* 18(1):109–121.
Bezerra SC, Ohashi A, Russo F, de Souza F (2020) Discrete-type approximations for non-Markovian optimal stopping problems: Part II. *Methodology Comput. Appl. Probab.* 22(3):1221–1255.
Bouchard B, Warin X (2012) Monte-Carlo valuation of American options: Facts and new algorithms to improve existing methods. Carmona R, Del Moral P, Hu P, Oudjane N, eds. *Numerical Methods in Finance* (Springer, Berlin), 215–255.
Brown DB, Smith JE, Sun P (2010) Information relaxations and duality in stochastic dynamic programs. *Oper. Res.* 58(4-part-1):785–801.
Chen Y (2021) Efficient algorithms for high-dimensional data-driven sequential decision-making. Ph.D. thesis, Cornell University, Ithaca, NY.
Chen N, Glasserman P (2007) Additive and multiplicative duals for American option pricing. *Finance Stochastics* 11(2):153–179.
Chen Y, Wang M (2017) Lower bound on the computational complexity of discounted Markov decision problems. Preprint, submitted May 20, https://arxiv.org/abs/1705.07312.
Chow Y, Robbins H, Siegmund D (1971) *Great Expectations: The Theory of Optimal Stopping* (Houghton Mifflin, Boston).
Christensen S (2014) A method for pricing American options using semi-infinite linear programming. *Math. Finance* 24(1):156–172.
Ciocan DF, Mišić VV (2022) Interpretable optimal stopping. *Management Sci.* 68(3):1616–1638.
Clément E, Lamberton D, Protter P (2002) An analysis of a least squares regression method for American option pricing. *Finance Stochastics* 6(4):449–471.
Cormen TH, Leiserson CE, Rivest RL, Stein C (2009) *Introduction to Algorithms* (MIT Press, Cambridge, MA).
Davis MH, Karatzas I (1994) A deterministic approach to optimal stopping. Kelly FP, ed. *Probability, Statistics and Optimisation* (John Wiley & Sons Ltd, New York), 455–466.
De Klerk E (2008) The complexity of optimizing over a simplex, hypercube or sphere: A short survey. *Central Eur. J. Oper. Res.* 16(2):111–125.
Desai VV, Farias VF, Moallemi CC (2012) Pathwise optimization for optimal stopping problems. *Management Sci.* 58(12):2292–2308.
Du SS, Kakade SM, Wang R, Yang LF (2019) Is a good representation sufficient for sample efficient reinforcement learning? *Internat. Conf. Learn. Representations* (OpenReview.net).
Egloff D (2005) Monte Carlo algorithms for optimal stopping and statistical learning. *Ann. Appl. Probab.* 15(2):1396–1432.
Glasserman P, Yu B (2004) Number of paths versus number of basis functions in American option pricing. *Ann. Appl. Probab.* 14(4):2090–2119.
Halman N, Nannicini G, Orlin J (2015) A computationally efficient FPTAS for convex stochastic dynamic programs. *SIAM J. Optim.* 25(1):317–350.
Halman N, Klabjan D, Li CL, Orlin J, Simchi-Levi D (2014) Fully polynomial time approximation schemes for stochastic dynamic programs. *SIAM J. Discrete Math.* 28(4):1725–1796.
Haugh MB, Kogan L (2004) Pricing American options: A duality approach. *Oper. Res.* 52(2):258–270.
Hill TP, Kertz RP (1983) Stop rule inequalities for uniformly bounded sequences of random variables. *Trans. Amer. Math. Soc.* 278(1):197–207.
Hill TP, Kertz RP (1992) A survey of prophet inequalities in optimal stopping theory. *Contemporary Math.* 125(1):191–207.
Ibáñez A, Velasco C (2020) Recursive lower and dual upper bounds for Bermudan-style options. *Eur. J. Oper. Res.* 280(2):730–740.
Jamshidian F (2007) The duality of optimal exercise and domineering claims: A Doob–Meyer decomposition approach to the Snell envelope. *Stochastics* 79(1–2):27–60.
Kohler M (2008) A regression-based smoothing spline Monte Carlo algorithm for pricing American options in discrete time. *AStA Adv. Statist. Anal.* 92(2):153–178.
Kohler M, Krzyżak A, Todorovic N (2010) Pricing of high-dimensional American options by neural networks. *Math. Finance* 20(3):383–410.
Kolodko A, Schoenmakers J (2004) Upper bounds for Bermudan style derivatives. *Monte Carlo Methods Appl.* 10(3–4):331–343.
Kolodko A, Schoenmakers J (2006) Iterative construction of the optimal Bermudan stopping time. *Finance Stochastics* 10(1):27–49.
Lai TL, Wong SS (2004) Valuation of American options via basis functions. *IEEE Trans. Automatic Control* 49(3):374–385.
Lelong J (2018) Dual pricing of American options by wiener chaos expansion. *SIAM J. Financial Math.* 9(2):493–519.
Longstaff FA, Schwartz ES (2001) Valuing American options by simulation: A simple least-squares approach. *Rev. Financial Stud.* 14(1):113–147.
Nemirovsky AS, Yudin DB (1983) *Problem Complexity and Method Efficiency in Optimization*, Wiley-Interscience Series in Discrete Mathematics (John Wiley and Sons, New York).
Rockafellar RT, Wets RJB (1991) Scenarios and policy aggregation in optimization under uncertainty. *Math. Oper. Res.* 16(1):119–147.
Rogers LC (2002) Monte Carlo valuation of American options. *Math. Finance* 12(3):271–286.
Schoenmakers J, Zhang J, Huang J (2013) Optimal dual martingales, their analysis, and application to new algorithms for Bermudan products. *SIAM J. Financial Math.* 4(1):86–116.

Shapiro A, Nemirovski A (2005) On complexity of stochastic programming problems. Jeyakumar V, Rubinov A, eds. *Continuous Optimization* (Springer, Boston), 111–146.

Sidford A, Wang M, Wu X, Yang L, Ye Y (2018) Near-optimal time and sample complexities for solving Markov decision processes with a generative model. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Red Hook, NY), 5186–5196.

Solan E, Tsirelson B, Vieille N (2012) Random stopping times in stopping problems and stopping games. Preprint, submitted November 25, https://arxiv.org/abs/1211.5802.

Stentoft L (2004) Convergence of the least squares Monte Carlo approach to American option valuation. *Management Sci.* 50(9):1193–1203.

Sturt B (2023) A nonparametric algorithm for optimal stopping based on robust optimization. *Oper. Res.* 71(5):1530–1557.

Swamy C, Shmoys DB (2012) Sampling-based approximation algorithms for multistage stochastic optimization. *SIAM J. Comput.* 41(4):975–1004.

Traub JF, Werschulz AG (1998) *Complexity and Information*. Lezioni Lincee, vol. 26862 (Cambridge University Press, Cambridge, UK).

Tsitsiklis JN, Van Roy B (1999) Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Trans. Automatic Control* 44(10):1840–1851.

Tsitsiklis JN, Van Roy B (2001) Regression methods for pricing complex American-style options. *IEEE Trans. Neural Networks* 12(4):694–703.

van der Vaart AW (2000) *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3 (Cambridge University Press, Cambridge, UK).

Wang D, Han Z (2015) Basics for sublinear algorithms. *Sublinear Algorithms for Big Data Applications*, SpringerBriefs in Computer Science (Springer, Cham, Switzerland), 9–21.

Wang Y, Wang R, Kakade SM (2021) An exponential lower bound for linearly-realizable MDPs with constant suboptimality gap. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Vaughan JW, eds. *Advances in Neural Information Processing Systems*, vol. 34 (Curran Associates, Inc., Red Hook, NY).

Ye J, Wong HY (2025) DeepMartingale: Duality of the optimal stopping problem with expressivity. Preprint, submitted October 13, https://arxiv.org/abs/2510.13868.

Zhou Z, Wang G, Blanchet J, Glynn PW (2023) Unbiased optimal stopping via the MUSE. *Stochastic Processes their Appl.* 166(2023):104088.