MATHEMATICS OF OPERATIONS RESEARCH
Vol. 34, No. 1, February 2009, pp. 210–237
ISSN 0364-765X | EISSN 1526-5471 | 09 | 3401 | 0210

DOI 10.1287/moor.1080.0360

An Optimal Approximate Dynamic Programming Algorithm for the Lagged Asset Acquisition Problem

Juliana M. Nascimento, Warren B. Powell
Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544
{jnascime@alumni.princeton.edu, powell@princeton.edu}

We consider a multistage asset acquisition problem where assets are purchased now, at a price that varies randomly over time, to be used to satisfy a random demand at a particular point in time in the future. We provide a rare proof of convergence for an approximate dynamic programming algorithm using pure exploitation, where the states we visit depend on the decisions produced by solving the approximate problem. The resulting algorithm does not require knowing the probability distribution of prices or demands, nor does it require any assumptions about its functional form. The algorithm and its proof rely on the fact that the true value function is a family of piecewise linear concave functions.

Key words: stochastic learning and adaptive control; stochastic approximation; approximate dynamic programming
MSC2000 subject classification: Primary: 93E35, 62L20; secondary: 90C39
OR/MS subject classification: Primary: optimal control, Markov finite state; secondary: inventory/production, stochastic uncertainty
History: Received August 21, 2006; revised April 20, 2007, January 7, 2008, and June 19, 2008.

1. Introduction. We consider a class of multistage problems called the *lagged asset acquisition problem*. An integer amount x_t of a single asset is purchased at time t, $t = 0, \ldots, T-1$, to be used to satisfy a demand that occurs only at a fixed time T. The price P_t that we pay to acquire assets at time t follows a Markov process. In most practical applications, the price trends upward, but downward fluctuations create buying opportunities. We do not realize the demand $\widehat{D}$ until time T, at which point we receive a random revenue $\hat{r}$ multiplied by the smaller of $\widehat{D}$ and the total we have ordered up to this point. In our problem, x_t is a scalar quantity and can only depend on the prices $P_0, \ldots, P_t$. The goal is to determine $x_0, \ldots, x_{T-1}$ that maximizes

$$\mathbb{E}\left[\sum_{t=0}^{T-1}(-P_t x_t) + \hat{r}\min\left(\widehat{D}, \sum_{t=0}^{T-1} x_t\right)\right]. \tag{1}$$

This problem arises in a number of settings. An energy company may be purchasing futures contracts for oil or gas to lock in a lower price now. Companies purchasing expensive equipment (aircraft, locomotives, power transformers) can often pay less if they place orders for further in the future. Shipping companies purchase space on container ships for a year or more in advance to guarantee space. All of these decisions are made before knowing the true demand, the prices, and the revenues in the future.

Our problem could be solved using classical backward dynamic programming, but two issues might prevent it. First, we may not know the probability distribution of prices, demands, and revenues. There has been an increasing interest in solving stochastic optimization problems using a distribution-free, nonparametric approach. Distribution-free revenue management and multiproduct pricing application can be found in van Ryzin and McGill [27] and Rusmevichientong et al. [18], respectively. A single-period newsvendor problem and its multi-period extension, when the demand distribution is unknown, are considered in Levi et al. [14]. They established bounds on the number of samples required to guarantee that with high probability, the expected cost of the sampling-based policies is arbitrarily close to the optimal policy. Second, even though the state variable only has two dimensions (price and quantity, which we assume are discrete), our state space can still be quite large. In §7, we report on experiments where the state space has as many as 16 million possible values. If we assume the probability distributions are known, exact solutions using classical methods require up to 6.7 hours to compute. Even with one-dimensional state spaces, the cardinality of the state space may still lead to prohibitive computational requirements. In the context of a single-item stochastic lot-sizing problem with known distribution, Halman et al. [11] develops approximation algorithms to deal with it. They also prove that finding an optimal policy is NP-hard.

The goal of this paper is to prove convergence of an algorithm that proceeds by solving problems of the form

$$x_t^n = \underset{x\in\{0,\ldots,M_t\}}{\arg\max} -P_t^n x + \overline{V}_t^{n-1}(P_t^n, R_{t-1}^n + x),$$

where $R_t^n = R_{t-1}^n + x_t^n$ captures cumulative past purchases, $\overline{V}_t^{n-1}(P_t^n, R_t^n)$ is an approximation to the dynamic programming optimal value function, P_t^n is a sample realization of the price we must pay for purchases at time t, and M_t is a known natural number.

Our algorithm and its convergence proof rely on the fact that both the optimal and the approximated value functions $\overline{V}_t^n(P_t, R_t)$ are piecewise linear and concave in the asset dimension with break points on the integers. If we define $\bar{F}_{t,P_t^n,R_{t-1}^n}^{n-1}: \mathbb{R} \to \mathbb{R}$, where

$$\bar{F}_{t,P_t^n,R_{t-1}^n}^{n-1}(x) = -P_t^n x + \overline{V}_t^{n-1}(P_t^n, R_{t-1}^n + x), \tag{2}$$

then the slopes of $\bar{F}_{t,P_t^n,R_{t-1}^n}^{n-1}(x)$ to the left and right of x_t^n (which is an integer break point of $\bar{F}_{t,P_t^n,R_{t-1}^n}^{n-1}(x)$) are used to update $\overline{V}_{t-1}^{n-1}$ obtaining $\overline{V}_{t-1}^n$. As we can see from Equation (2), the slopes used to update $\overline{V}_{t-1}^{n-1}$ both depend on the sample information given by P_t^n and on $\overline{V}_t^{n-1}$, which at iteration n is only an approximation of future profits. As a result, the slopes are biased, causing complications in the convergence proof.

The convergence proof requires that the price process P_t^n has finite support. However, this assumption is not restrictive, as a Markovian discrete process can be obtained from a Markovian continuous process if the original probability distribution is adjusted accordingly to reflect a chosen discretization/truncation scheme. Nevertheless, for theoretical purposes, the discretization increment can be arbitrarily fine. We must note that for an actual implementation of our algorithm we are able to use a continuous price process because we follow sample paths (discretization occurs only in the representation of the value functions).

The dependence on sample information and on the approximation of the value function in the future is common in approximate dynamic programming algorithms (see Bertsekas and Tsitsiklis [4], Sutton and Barto [22]), where an approximation of the future is used to make decisions now, stepping forward in time. The use of separable, piecewise linear approximations has already proven effective on very difficult classes of stochastic resource allocation problems (see Godfrey and Powell [10], Topaloglu and Powell [24]), but as of this writing there are no convergence results for multistage problems.

Our proof technique combines ideas from the field of approximate dynamic programming (notably, Bertsekas and Tsitsiklis [4]) as well as the proof of the SPAR algorithm (successive projective approximation routine) in Powell et al. [16]. Our algorithm is modeled after the SPAR algorithm, which is presented in the context of a two-stage problem. The result is a rare instance of a provably convergent approximate dynamic programming algorithm that uses pure exploitation, which is to say that the decision x_t^n that we make now (based on the value function approximation $\overline{V}_t^{n-1}$) determines the state we visit at $t+1$. Current proofs of convergence for approximate dynamic programming algorithms such as Q-learning (Tsitsiklis [26], Jaakkola et al. [13]) and optimistic policy iteration (Tsitsiklis [25]) require that we visit states (and possibly actions) infinitely often. A convergence proof for a real time dynamic programming (RTDP) (Barto et al. [3]) algorithm that considers a pure exploitation scheme is provided in Bertsekas and Tsitsiklis [4, Proposition 5.3 and 5.4], but it assumes that the initial value function approximations are optimistic in the sense that they are smaller (for a minimization problem) than the optimal ones. It also assumes that the distribution of the random variables are known. We make no such assumptions, but it is important to emphasize that our result depends on the concavity of the objective function.

There are a number of competing approaches to this problem. Because our problem requires integer solutions, we can use any of a vast range of approximate dynamic programming algorithms (Bertsekas and Tsitsiklis [4]), but these lack provable convergence without forcing the algorithm to sample states and actions infinitely often. It should be noted, though, that there is a family of provably convergent algorithms (Singh et al. [21]) that performs decaying exploration schemes and is more likely to achieve better rates of convergence than algorithms where the exploration selection is employed in a more simplistic way. Boltzmann and ϵ-greedy exploration are part of this family of algorithms. From the field of stochastic programming, there are several flavors of Bender's decomposition that can be used (Van Slyke and Wets [28], Higle and Sen [12], Chen and Powell [6]). However, these methods will not handle the random price issue. Another powerful technique is sample average approximation (SAA) (Shapiro [19]), which relies on generating random samples outside of the optimization problems and then solving the corresponding deterministic problems using an appropriate optimization algorithm. Numerical experiments with the SAA approach applied to problems where an integer solution is required can be found in Ahmed and Shapiro [2].

The contributions of the paper are that (a) we propose an approximate dynamic programming algorithm (which avoids the need to enumerate the state space) for the lagged asset acquisition problem; (b) we prove convergence of the algorithm, which is complicated by the fact that it uses pure exploitation and a projection operation that enforces concavity of the approximate value function at each iteration; and (c) we show empirically how the

rate of convergence of our algorithm compares to the rate of convergence of other approximation methods such as RTDP and ϵ-greedy.

This paper is organized as follows: Section 2 defines the problem and the corresponding dynamic programming model. Section 3 describes the algorithmic strategy. Section 4 introduces notation and assumptions for the convergence analysis. Section 5 presents a sketch of the convergence proofs and §6 provides the full proofs. Finally, §7 provides some experimental comparisons against the optimal policy and other approaches and §8 presents the conclusions.

2. Problem formulation and model. In this section, we give a precise description of the problem considered in this paper as well as the assumptions. We also provide the dynamic programming model associated with the problem and identify the structural properties that are exploited in our proof.

The problem is to determine, in each time period $t = 0, \ldots, T-1$, how much should be purchased of a given asset to meet a positive discrete integer random demand $\widehat{D}$ at time T. A strictly positive price P_t is charged for each unit of asset purchased at t and a strictly positive bounded random reward $\hat{r}$ is received for each unit of satisfied demand.

We denote by x_t the amount purchased at each period and we require that $x_t \in \{0, \ldots, M_t\}$, where M_t is a known natural number. Moreover, x_t only relies on information available up until time period t. The price process $P = (P_0, \ldots, P_{T-1})$ is a Markov process independent of the asset level, with finite support $\mathcal{P} = \mathcal{P}_0 \times \cdots \times \mathcal{P}_{T-1}$. The objective is to maximize the expected profit, given by Equation (1).

The decision x_t at each period t depends both on the current unit price of the asset and on the amount of assets purchased up until time $t-1$ (inclusive), which is denoted by R_{t-1}. We note that x_t suffices to depend only on the current price and asset level because the price process is Markov, the immediate rewards do not depend on the past history, and past purchase decisions cannot be modified. We assume that $R_{-1} = 0$. Clearly, $R_t = R_{t-1} + x_t$, for $t = 0, \ldots, T-1$. Note that R_{T-1} denotes the total number of assets acquired over all time periods, which is used to satisfy demand $\widehat{D}$ at T. We have that both the demand $\widehat{D}$ and the reward $\hat{r}$ might be correlated and dependent on the final price P_{T-1}. However, they must be independent of the asset level R_{T-1}. In fact, given the final price P_{T-1}, the demand and the reward are independent of the complete past, implying that they are independent of the intermediate asset levels as well.

The problem can be formulated as a dynamic program. For $t = 0, \ldots, T-2$, the optimality equations $V_t^*\colon \mathcal{P}_t \times [0, B_t] \to \mathbb{R}$, where $B_t = \sum_{i=0}^{t} M_i$, are given by

$$V_t^*(P, R) = \mathbb{E}\Big[\max_{x_{t+1} \in \{0, \ldots, M_{t+1}\}} -P_{t+1}x_{t+1} + V_{t+1}^*(P_{t+1}, R + x_{t+1}) \,\Big|\, P_t = P\Big].$$

For $t = T-1$, $V_{T-1}^*\colon \mathcal{P}_{T-1} \times [0, B_{T-1}] \to \mathbb{R}$ is given by:

$$V_{T-1}^*(P, R) = \mathbb{E}\big[\hat{r} \min(\widehat{D}, R) \mid P_{T-1} = P\big].$$

We point out that the initial decision x_0 is the optimal solution of the optimization problem

$$\max_{x \in \{0, \ldots, M_0\}} -P_0 x + V_0^*(P_0, x),$$

as $R_{-1} = 0$. Moreover, P_0 is known at zero, so no randomness/expectations are involved in the decision-making process.

The state variable at time t is given by $S_t = (P_t, R_t)$. We let $S = (S_0, \ldots, S_{T-1})$ be our state vector and $\mathcal{S} = \mathcal{S}_0 \times \cdots \times \mathcal{S}_{T-1}$ be the state space, where $\mathcal{S}_t = \mathcal{P}_t \times \{0, \ldots, B_t\}$. We are slightly abusing notation here because the time period t should also be included in the definition of the state variable, as the information necessary for the system to move forward is given by S_t combined with t. Therefore, the state vector and the state space are in fact $((S_0, 0), \ldots, (S_{T-1}, T-1))$ and $\bigcup_{t=0}^{T-1} \mathcal{S}_t \times \{t\}$, respectively.

Note that we are using a postdecision state variable, which is the state of the system after the decision x_t is taken (see Powell [15, Chapter 4] for a complete discussion). Postdecision states lead to an inversion of the optimization/expectation order in the value function formula. This inversion allows for more effective computational strategies.

We can show that the optimal value functions are concave and piecewise linear with integer break points in the asset dimension. Therefore, for $t = 0, \ldots, T-1$ and $P \in \mathcal{P}_t$, the value function $V_t^*(P, \cdot)$ can be identified uniquely by its decreasing slopes $(v_t^*(P, 1), \ldots, v_t^*(P, B_t))$, where $v_t^*(P, i) = V_t^*(P, i) - V_t^*(P, i-1)$, $i = 1, \ldots, B_t$. Moreover, if R is an integer, the optimal decision

$$x_t^* = \arg\max_{0 \le x \le M_t} -P_t x + V_t^*(P_t, R + x), \quad t = 0, \ldots, T-1$$

is an integer without having to enforce integrality. We disregard the values at $V_t^*(P,0)$ because the optimal decisions x_{t+1}^* do not change when $V_t^*(P,\cdot)$ is shifted by a constant. In order to simplify notation, let $\bar{\mathscr{S}}_t = \mathscr{P}_t \times \{1,\dots,B_t\}$ and $\bar{\mathscr{S}} = \bar{\mathscr{S}}_0 \times \cdots \times \bar{\mathscr{S}}_{T-1}$. Note that $\bar{\mathscr{S}}_t$ is obtained from $\mathscr{S}_t$ by excluding the pairs $(P,0)$ for all $P \in \mathscr{P}_t$.

We close the section by summarizing the important properties of the optimal value functions and their slopes that are used throughout the paper. The proof is shown in Appendix B.

PROPOSITION 2.1. *The optimal value functions are piecewise linear, with integer break points and concave in the asset dimension. Moreover, for $t = 0,\dots,T-1$ and $(P,R) \in \bar{\mathscr{S}}_t$, the optimal slope $v_t^*(P,R) = V_t^*(P,R) - V_t^*(P,R-1)$ is given by*:

$$v_t^*(P,R) = \mathbb{E}\big[\max(\min(P_{t+1}, v_{t+1}^*(P_{t+1},R)), v_{t+1}^*(P_{t+1},R+M_{t+1})) \mid P_t = P\big]1_{\{t<T-1\}} + \mathbb{E}\big[\hat{r}1_{\{\hat{D}\geq R\}} \mid P_{T-1} = P\big]1_{\{t=T-1\}}. \tag{3}$$

Thus, $v_t^*(P,R)$ is bounded between zero and $\max \hat{r}$, which is the maximum of the support for the reward $\hat{r}$. Furthermore, $(v_t^*(P,1),\dots,v_t^*(P,B_t)) \in \mathscr{C}_t$, where

$$\mathscr{C}_t = \big\{v \in \mathbb{R}^{B_t}\colon v_1 \leq \max \hat{r},\ v_{B_t} \geq 0,\ v_{R+1} \leq v_R \text{ for } R = 1,\dots,B_t - 1\big\}.$$

3. Algorithmic strategy. Our approach to the problem consists of learning the optimal decision given the time period, the amount of assets already available, and the current price. However, the objective is to learn the optimal decision only for asset levels that can be generated by an optimal policy.

Figure 1 describes the ADP-lagged algorithm, a modified version of the SPAR algorithm (Powell et al. [16]). The algorithm starts with initial piecewise linear value function approximations $\bar{V}^0$ represented by their slopes $\bar{v}^0$. As discussed in the previous section, optimal decisions depend only on the slopes of the value functions, thus the algorithm only deals with the slopes instead of the value functions themselves. The initial approximation of the slopes are only required to be decreasing and bounded between zero and $\max \hat{r}$.

At each iteration n and time t, a decision x_t^n is made. This decision is optimal with respect to the sample realization of the price sequence up to time t, the asset level R_{t-1}^n, and the current approximation of the slopes $\bar{v}_t^{n-1}$. It will bring the system to the new asset level $R_t^n = R_{t-1}^n + x_t^n$.

Just after this transition, a sample realization of the slopes of $\bar{V}_t^{n-1}$ to the left and right of (P_t^n, R_t^n) is observed. These samples, denoted by $\hat{v}_{t+1}^n(R_t^n)$ and $\hat{v}_{t+1}^n(R_t^n+1)$, are used to update the slope approximations $\bar{v}_t^{n-1}(P_t^n, R_t^n)$ and $\bar{v}_t^{n-1}(P_t^n, R_t^n+1)$. After that, a projection operation $\Pi_{\mathscr{C},P_t^n,R_t^n}$ is performed in case a violation of the concavity property occurs. For completeness, we assume that $R_{-1}^n = 0$ for all n. We also assume that $\bar{V}_t^n(P,0) = 0$ for all prices $P \in \mathscr{P}_t$.

We denote by $S_t = (P_t, R_t)$ a general state at time t. $S_t^n = (P_t^n, R_t^n)$ represents the actual state visited by the algorithm at iteration n and time t. Moreover, $\{S_t^n\}_{n\geq 0} = \{(P_t^n, R_t^n)\}_{n\geq 0}$ is the sequence of states generated by the algorithm. The same notation holds for the decisions x_t, x_t^n, and $\{x_t^n\}_{n\geq 0}$.

The algorithm also generates the $\{\bar{v}^n\}_{n\geq 0}$ sequences, that is, the sequences of slopes of the value function approximations. It is important to realize that there is one sequence $\{\bar{v}_t^n(P,R)\}_{n\geq 0}$ for each time $t < T$ and $(P,R) \in \bar{\mathscr{S}}_t$. The notation $\{\bar{v}^n\}_{n\geq 0}$ represents the family of all such sequences.

Step 0(i). Pick $\bar{v}_t^0(P,R) \in [0, \max \hat{r}]$ for all t and (P,R) to be monotone decreasing in R.
Step 0(ii). Set $R_{-1}^n = 0$ for all $n \geq 0$.
Step 0(iii). Set $n = 1$.
Step 1. Sample the price sequence $P^n = (P_0^n,\dots,P_{T-1}^n)$, the demand $\hat{D}^n$, and reward $\hat{r}^n$.
Step 2. Do for $t = 0,\dots,T-1$:
  Step 2(a). $x_t^n = \arg\max_{0\leq x\leq M_t} -P_t^n x + \bar{V}_t^{n-1}(P_t^n, R_{t-1}^n + x)$.
  Step 2(b). $R_t^n = R_{t-1}^n + x_t^n$.
  Step 2(c). Observe $\hat{v}_{t+1}^n(R_t^n)$ and $\hat{v}_{t+1}^n(R_t^n+1)$ according to Equation (4).
  Step 2(d). For $(P,R) \in \bar{\mathscr{S}}_t$,

$$z_t^n(P,R) = \begin{cases} (1-\alpha_t^n)\bar{v}_t^{n-1}(P,R) + \alpha_t^n \hat{v}_{t+1}^n(R), & \text{if } P = P_t^n,\ R \in \{R_t^n, R_t^n+1\} \\ \bar{v}_t^{n-1}(P,R), & \text{else} \end{cases}$$

  Step 2(e). $\bar{v}_t^n = \Pi_{\mathscr{C},P_t^n,R_t^n}(z_t^n)$. See Equation (5) for the details.
Step 3. Increase n by one and go to Step 1.

FIGURE 1. ADP-lagged algorithm.

Remember that for $(P, R) \in \bar{\mathcal{S}}_t$, the optimal slope is given by Equation (3). A sample slope is obtained by replacing the expectation in Equation (3) by a sample realization and by replacing v^*_{t+1} by its current approximation. Therefore, the sample slope for $R = 1, \ldots, B_t$ is given by

$$\hat{v}^n_{t+1}(R) = \max\bigl(\min(P^n_{t+1}, \bar{v}^{n-1}_{t+1}(P^n_{t+1}, R)), \bar{v}^{n-1}_{t+1}(P^n_{t+1}, R + M_{t+1})\bigr) 1_{\{t<T-1\}} + \hat{r}^n 1_{\{R \le \hat{D}^n\}} 1_{\{t=T-1\}}, \tag{4}$$

where M_{t+1} is the upper bound on the decision x_{t+1}. We do not define the sample slopes $\hat{v}^n_{t+1}(0)$ and $\hat{v}^n_{t+1}(B_t+1)$, as they are not needed in the update process. Note that for all t, $\hat{v}^n_{t+1}(R) \ge \hat{v}^n_{t+1}(R+1)$ because the projection operation guarantees that $\bar{v}^{n-1}_{t+1}$ is monotone decreasing in the asset dimension. Thus, $\min(P^n_{t+1}, \bar{v}^{n-1}_{t+1}(P^n_{t+1}, R)) \ge \min(P^n_{t+1}, \bar{v}^{n-1}_{t+1}(P^n_{t+1}, R+1))$ and $\bar{v}^{n-1}_{t+1}(P^n_{t+1}, R + M_{t+1}) \ge \bar{v}^{n-1}_{t+1}(P^n_{t+1}, R+1+M_{t+1})$. When $t = T-1$, the sample slope does not depend on a current slope approximation, as is the case for $t < T-1$. This fact is important for the convergence analysis of the algorithm because it implies that $\hat{v}^n_{t+1}(R)$ is an unbiased estimator of $v^*_{t+1}(P, R)$ for $t = T-1$, though it is generally biased for $t < T-1$.

The sample slopes are used to update the approximation slopes $\bar{v}^{n-1}_t$ through a temporary slope vector z^n_t. This step requires the use of a stepsize rule, denoted by α^n_t, where α^n_t is a scalar between zero and one and can depend only on information that became available up until iteration n and time t. We make the standard assumptions that $\sum_{n=1}^{\infty} \alpha^n_t = \infty$ and $\sum_{n=1}^{\infty} (\alpha^n_t)^2 \le B < \infty$ almost surely, where B is a constant.

The projection operator $\Pi_{\mathscr{C}, P^n_t, R^n_t}$ maps the vector z^n_t that may not be monotone decreasing in the asset dimension (the concavity property) into another vector $\bar{v}^n_t$ such that for $P \in \mathscr{P}_t$, $\bar{v}^n_t(P) = (\bar{v}^n_t(P,1), \ldots, \bar{v}^n_t(P, B_t)) \in \mathscr{C}_t$. In this paper, we consider the *level* projection operator (introduced in Topaloglu and Powell [23]). It imposes concavity by simply forcing the violating slopes to be equal to the newly updated ones. For $(P, R) \in \bar{\mathcal{S}}_t$, the operator is given by:

$$\Pi_{\mathscr{C}, P^n_t, R^n_t}(z)(P, R) = \begin{cases} z(P^n_t, R^n_t), & \text{if } P = P^n_t,\ R \le R^n_t,\ z(P,R) \le z(P^n_t, R^n_t) \\ z(P^n_t, R^n_t + 1), & \text{if } P = P^n_t,\ R \ge R^n_t + 1,\ z(P,R) \ge z(P^n_t, R^n_t + 1) \\ z(P, R), & \text{otherwise.} \end{cases} \tag{5}$$

Figure 2 helps us visualize one iteration n of the algorithm at time t. After the algorithm has sampled the price sequence, demand, and reward, and has made the decisions up until time $t-1$, the current price is P^n_t and the total amount of assets purchased so far is R^n_{t-1}. Based on the slope approximation $\bar{v}^{n-1}_t$, the algorithm determines the amount of assets x^n_t to acquire at time t and samples the slopes at $R^n_t = R^n_{t-1} + x^n_t$ and $R^n_t + 1$ as illustrated in Figure 2(a). After the current slope approximations are updated using the sampled slopes, a violation of the concavity property may occur as shown in Figure 2(b). In this case, the projection operation $\Pi_{\mathscr{C}, P^n_t, R^n_t}$ is performed and concavity is restored as in Figure 2(c).

The decision x^n_t maximizes the function $\bar{F}^{n-1}_{t, P^n_t, R^n_{t-1}}(x) = -P^n_t x + \bar{V}^{n-1}_t(P^n_t, R^n_{t-1} + x)$, where for each $(P, R) \in \bar{\mathcal{S}}_t$, we have that $\bar{V}^{n-1}_t(P, R) = \sum_{i=1}^{R} \bar{v}^{n-1}_t(P, i)$, as we have assumed that $\bar{V}^{n-1}_t(P, 0) = 0$. Because $\bar{v}^{n-1}_t$ is monotone decreasing in the asset dimension, $\bar{F}^{n-1}_{t, P^n_t, R^n_{t-1}}(x) \le \bar{F}^{n-1}_{t, P^n_t, R^n_{t-1}}(x+1)$ is equivalent to $\bar{v}^{n-1}_t(P^n_t, R^n_{t-1} + x + 1) \ge P^n_t$. Hence, the solution of the unconstrained optimization problem is characterized by

$$\bar{v}^{n-1}_t(P^n_t, R^n_{t-1} + x) \ge P^n_t, \qquad \bar{v}^{n-1}_t(P^n_t, R^n_{t-1} + x + 1) \le P^n_t. \tag{6}$$

Because of $0 \le x \le M_t$, it may happen that no such solution exists. If $v^{n-1}_t(P^n_t, R^n_{t-1}) \le P^n_t$ then the optimal decision is equal to zero. On the other hand, if $v^{n-1}_t(P^n_t, R^n_{t-1} + M_t + 1) \ge P^n_t$, then the optimal decision is equal to the upper bound M_t.

4. Theoretical conditions and assumptions. We start this section pointing out that the sequence of states $\{S^n_t\}_{n\ge0} = \{(P^n_t, R^n_t)\}_{n\ge0}$ and the sequence of decisions $\{x^n_t\}_{n\ge0}$ generated by the algorithm have at least one accumulation point as the price sequence has finite support and the decisions are integer and bounded, which implies that the resource sequence has finite support as well.

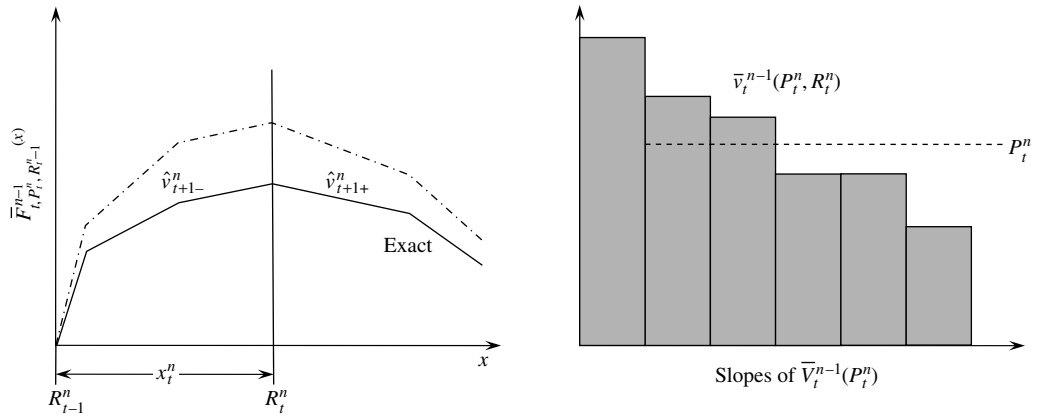
Let $\bar{\mathcal{S}}^*_t$ be the set of all states that are either equal to an accumulation point (P^*_t, R^*_t) of $\{(P^n_t, R^n_t)\}_{n\ge0}$ or are equal to $(P^*_t, R^*_t + 1)$. Moreover, we only consider accumulation points (P^*_t, R^*_t) such that $R^*_t > 0$ and $R^*_t < B_t$.

The slope sequences $\{\bar{v}^n\}_{n\ge0}$ also have an accumulation point, as the set $\mathscr{C}_t$ (defined in Proposition 2.1) is compact and the projection operation guarantees for all iterations n and prices $P \in \mathscr{P}_t$ that $\bar{v}^n_t(P) = (\bar{v}^n_t(P,1), \ldots, \bar{v}^n_t(P, B_t)) \in \mathscr{C}_t$.

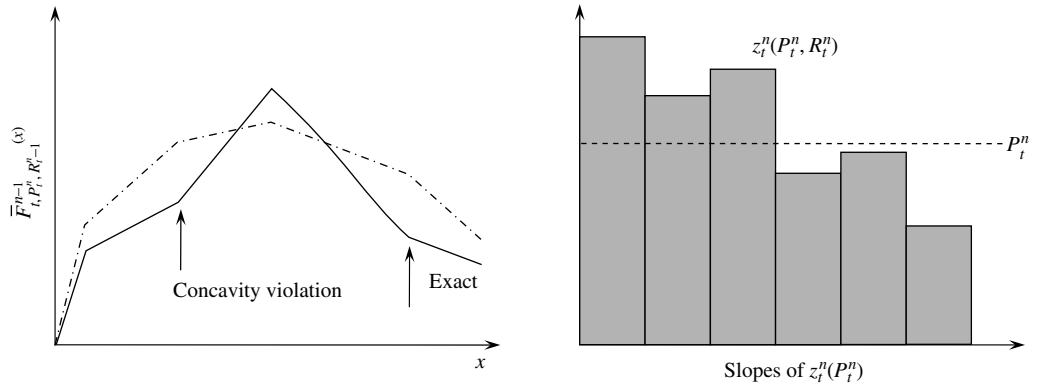
Let $(\Omega, \mathscr{F}, \mathbb{P})$ be our probability space. Define the sigma algebra

$$\mathscr{F} = \sigma\bigl\{(P^n_t, x^n_t, \hat{D}^n, \hat{r}^n),\ n \ge 1,\ t = 0, \ldots, T-1\bigr\}.$$

(a) Current approximate function, optimal decision, and sampled slopes



(b) Temporary approximate function with violation of concavity



(c) Level projection operation: Updated approximate function with concavity restored

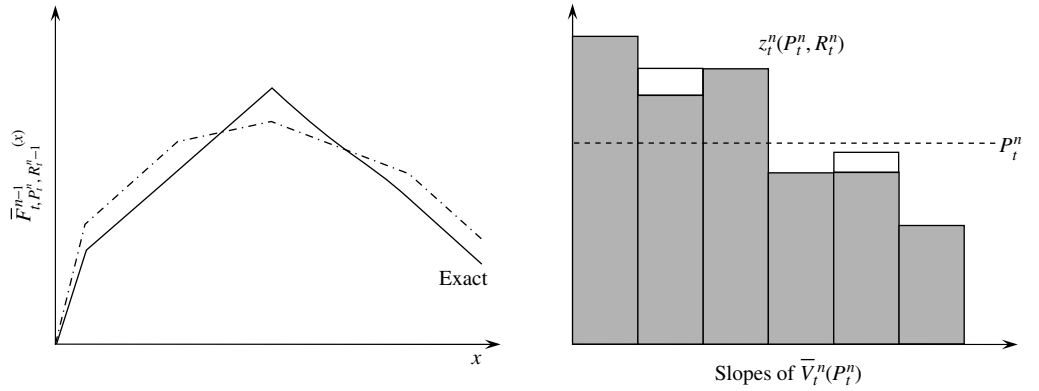


FIGURE 2. Iteration n of the algorithm at time t.

Moreover, let, for $n \geq 1$, $\mathcal{F}^n_T = \sigma\{(P^m_{t'}, x^m_{t'}, \widehat{D}^m, \hat{r}^m),\ m \leq n,\ t' = 0, \ldots, T-1\}$ and, for $t < T$, $\mathcal{F}^n_t = \sigma(\{(P^m_{t'}, x^m_{t'}, \widehat{D}^m, \hat{r}^m),\ m < n,\ t' = 0, \ldots, T-1\} \cup \{(P^n_{t'}, x^n_{t'}),\ t' = 0, \ldots, t\})$. Clearly, $\mathcal{F}^n_t \subset \mathcal{F}^n_{t+1}$ and $\mathcal{F}^n_T \subset \mathcal{F}^{n+1}_0$. Furthermore, given the initial slopes $\bar{v}^0_t(P,R)$, we have that R^n_0 and α^n_0 are in $\mathcal{F}^n_0$ while for $0 < t < T$, $\hat{v}^n_t(R)$, $z^n_{t-1}(P,R)$, $\bar{v}^n_{t-1}(P,R)$, R^n_t, and α^n_t are all in $\mathcal{F}^n_t$. Finally, $\widehat{D}^n$, $\hat{r}^n$, $\hat{v}^n_T(R)$, $z^n_{T-1}(P,R)$, and $\bar{v}^n_{T-1}(P,R)$ are in $\mathcal{F}^n_T$.

We introduce the random index $\bar{N}$, which is used to indicate when an iteration of the algorithm is large enough for convergence analysis purposes. Let $\bar{N}$ be the smallest integer such that for all $t \in \{0, \ldots, T-1\}$, $(\bar{P}, \bar{R}) \in \mathcal{S}_t$, $P \in \mathcal{P}_{t+1}$, and $x \in \{0, \ldots, M_{t+1}\}$, it holds that:

CONDITION 1 (C1). If $\sum_{n=1}^{\infty} 1_{\{(R^n_t, P^n_{t+1}, x^n_{t+1}) = (\bar{R}, P, x)\}} < \infty$ a.s., then $\sum_{n=\bar{N}}^{\infty} 1_{\{(R^n_t, P^n_{t+1}, x^n_{t+1}) = (R, P, x)\}} = 0$ a.s.;

CONDITION 2 (C2). If $\sum_{n=1}^{\infty} 1_{\{\bar{v}^{n-1}_{t+1}(P, \bar{R}+1) < P,\ P^n_t = \bar{P},\ R^n_t = \bar{R}\}} < \infty$ a.s., then $\sum_{n=\bar{N}}^{\infty} 1_{\{\bar{v}^{n-1}_{t+1}(P, \bar{R}+1) < P,\ P^n_t = \bar{P},\ R^n_t = \bar{R}\}} = 0$ a.s.;

CONDITION 3 (C3). If $\sum_{n=1}^{\infty} 1_{\{\bar{v}^{n-1}_{t+1}(P, \bar{R}+1) > P,\ P^n_t = \bar{P},\ R^n_t = \bar{R}\}} < \infty$ a.s., then $\sum_{n=\bar{N}}^{\infty} 1_{\{\bar{v}^{n-1}_{t+1}(P, \bar{R}+1) > P,\ P^n_t = \bar{P},\ R^n_t = \bar{R}\}} = 0$ a.s.

Because $\mathcal{S}_t$ and $\mathcal{P}_{t+1}$ are finite sets, we have that $\bar{N}$ has to satisfy a finite number of constraints, thus it is trivial to see that $\bar{N}$ is finite almost surely. Moreover, from (C1), we see that all states (actions) visited (taken) by the algorithm after iteration $\bar{N}$ are accumulation points of the sequence of states (actions) generated by the algorithm.

Even though the results of the next lemma are only required later in the paper, its proof nicely illustrates the use of the elements just defined: the random set $\bar{\mathcal{S}}_t^*$, the sigma algebra $\mathcal{F}_t^n$, and the random index $\bar{N}$. Hence, we choose to do it now rather than later. The proof relies on an extended version of the Borel-Cantelli lemma as described in Breiman [5, Corollary 5.29] and in Singh et al. [21, Lemma 3].

LEMMA 4.1. *Pick $(\bar{P},\bar{R})\in\bar{\mathcal{S}}_t$. Also pick $P\in\mathcal{P}_{t+1}$ such that $\mathbb{P}\{P_{t+1}=P\mid P_t=\bar{P}\}>0$. On the event that $(P,\bar{R})\notin\bar{\mathcal{S}}_{t+1}^*$, $\sum_{n=\bar{N}}^{\infty}1_{\{\bar{v}_{t+1}^{n-1}(P,\bar{R}+1)<P,\,P_t^n=\bar{P},\,R_t^n=\bar{R}\}}=0$ almost surely. Moreover, let $\bar{R}^*=\sup\{R\colon (\bar{P},\bar{R},P,R)\in\{(P_t^n,R_t^n,P_{t+1}^n,R_{t+1}^n)\}_{n\geq 1}$ and $(P,R)\in\bar{\mathcal{S}}_{t+1}^*\}$. On the event that $(P,\bar{R}+M_{t+1})\notin\bar{\mathcal{S}}_{t+1}^*$, it holds that $\sum_{n=\bar{N}}^{\infty}1_{\{\bar{v}_{t+1}^{n-1}(P,\bar{R}^*+1)<P,\,P_t^n=\bar{P},\,R_t^n=\bar{R}\}}=0$ almost surely.*

PROOF. Fix $(\bar{P},\bar{R})\in\bar{\mathcal{S}}_t$ and $P\in\mathcal{P}_{t+1}$ such that $\mathbb{P}\{P_{t+1}=P\mid P_t=\bar{P}\}>0$. Let $\mathcal{N}=\{n\in\mathbb{N}\colon \bar{v}_{t+1}^{n-1}(P,\bar{R}+1)<P,\,P_t^n=\bar{P},\,R_t^n=\bar{R}\}$ and suppose that $\mathbb{P}\{\omega\colon|\mathcal{N}(\omega)|=\infty\}>0$. Define the event $A_{t+1}^n=\{P_{t+1}^n=P\}$ and the set $\bar{\mathcal{N}}=\mathcal{N}\cap\{n\in\mathbb{N}\colon P_{t+1}^n=P\}$. Clearly, $A_{t+1}^n\in\mathcal{F}_{t+1}^n$. Moreover, $\mathbb{P}\{A_{t+1}^n\mid\mathcal{F}_t^n\}=\mathbb{P}\{A_{t+1}^n\mid P_t^n\}$ and $\mathbb{P}\{A_{t+1}^n\mid P_t^n=\bar{P}\}>0$. Thus, $\{\omega\colon|\mathcal{N}(\omega)|=\infty\}$ is contained in the event $\{\omega\colon\sum_{n=1}^{\infty}\mathbb{P}\{A_{t+1}^n|P_t^n\}(\omega)=\infty\}$. By the extended version of the Borel-Cantelli lemma (Breiman [5, Corollary 5.29]),

$$\left\{\omega\colon\sum_{n=1}^{\infty}\mathbb{P}\{A_{t+1}^n\mid P_t^n\}(\omega)=\infty\right\}=\{\omega\colon\omega\in A_{t+1}^n\text{ for infinitely many }n\text{'s}\},$$

implying that $|\bar{\mathcal{N}}|=\infty$ almost surely whenever $|\mathcal{N}|=\infty$. Pick $\omega\in\Omega$ and $\bar{n}\in\bar{\mathcal{N}}(\omega)$ such that $\bar{n}\geq\bar{N}(\omega)$. Assume that $(P,\bar{R})\notin\bar{\mathcal{S}}_{t+1}^*(\omega)$. Given the decision characterization (6) of the optimization problem solved at each iteration of the algorithm, we have that $x_{t+1}^{\bar{n}}(\omega)=0$ and $R_{t+1}^{\bar{n}}(\omega)=\bar{R}$, leading to the contradiction that $(P,\bar{R})\in\bar{\mathcal{S}}_{t+1}^*(\omega)$, as all states visited by the algorithm at $t+1$ after iteration $\bar{N}(\omega)$ are elements of $\bar{\mathcal{S}}_{t+1}^*(\omega)$. Therefore, on $\{(P,\bar{R})\notin\bar{\mathcal{S}}_{t+1}^*\}$, $|\mathcal{N}|<\infty$ almost surely and from (C2) in the definition of $\bar{N}$, it follows that $\sum_{n=\bar{N}}^{\infty}1_{\{\bar{v}_{t+1}^{n-1}(P,\bar{R}+1)<P,\,P_t^n=\bar{P},\,R_t^n=\bar{R}\}}=0$ almost surely.

Let $\bar{R}^*=\sup\{R\colon(\bar{P},\bar{R},P,R)\in\{(P_t^n,R_t^n,P_{t+1}^n,R_{t+1}^n)\}_{n\geq1}$ and $(P,R)\in\bar{\mathcal{S}}_{t+1}^*\}$ and define $\mathcal{N}^*=\{n\in\mathbb{N}\colon\bar{v}_{t+1}^{n-1}(P,\bar{R}^*+1)>P,\,P_t^n=\bar{P},\,R_t^n=\bar{R}\}$. As before, suppose that $\mathbb{P}\{\omega\colon|\mathcal{N}^*(\omega)|=\infty\}>0$. It holds that $\{\omega\colon|\mathcal{N}^*(\omega)|=\infty\}$ is contained in the event $\{\omega\colon\sum_{m=1}^{\infty}\mathbb{P}\{A_{t+1}^n|P_t^n\}(\omega)=\infty\}$ and the extended version of the Borel-Cantelli lemma tells us that $|\bar{\mathcal{N}}^*|=\infty$ almost surely whenever $|\mathcal{N}^*|=\infty$, where $\bar{\mathcal{N}}^*=\mathcal{N}^*\cap\{n\in\mathbb{N}\colon P_{t+1}^n=P\}$. Pick $\omega\in\Omega$ and $\bar{n}\in\bar{\mathcal{N}}^*(\omega)$ such that $\bar{n}\geq\bar{N}(\omega)$. Assume that $(P,\bar{R}+M_{t+1})\notin\bar{\mathcal{S}}_{t+1}^*(\omega)$. As the decisions are bounded by M_{t+1}, we have that $\bar{R}\leq\bar{R}^*<\bar{R}+M_{t+1}$. Again, given the decision characterization, we have that $x_{t+1}^{\bar{n}}(\omega)>\bar{R}^*-\bar{R}$ and thus $R_{t+1}^{\bar{n}}(\omega)>\bar{R}^*$. Because $R_{t+1}^{\bar{n}}(\omega)$ is an element of the set for which $\bar{R}^*$ is supposed to be the supremum, we get our contradiction. Therefore, on $\{(P,\bar{R}+M_{t+1})\notin\bar{\mathcal{S}}_{t+1}^*\}$, $|\mathcal{N}^*|<\infty$ and from (C3) in the definition of $\bar{N}$, it follows that $\sum_{n=\bar{N}}^{\infty}1_{\{\bar{v}_{t+1}^{n-1}(P,\bar{R}^*+1)>P,\,P_t^n=\bar{P},\,R_t^n=\bar{R}\}}=0$ almost surely. □

For $(P,R)\in\bar{\mathcal{S}}_t$, we present the sets of iterations $\mathcal{N}_t^-(P,R)$ and $\mathcal{N}_t^+(P,R)$. These sets keep track of the effects produced by the projection operation. Let $\mathcal{N}_t^-(P,R)$ ($\mathcal{N}_t^+(P,R)$) be the set of iterations in which the unprojected slope corresponding to state (P,R) was too small (large) and had to be increased (decreased) by the projection operation. Formally,

$$\mathcal{N}_t^-(P,R)=\{n\in\mathbb{N}\colon z_t^n(P,R)<\bar{v}_t^n(P,R)\}$$
$$\mathcal{N}_t^+(P,R)=\{n\in\mathbb{N}\colon z_t^n(P,R)>\bar{v}_t^n(P,R)\}.$$

For example, based on Figure 2(c),

$$n\in\mathcal{N}_t^-(P_t^n,R_t^n-1)\qquad\text{and}\qquad n\in\mathcal{N}_t^+(P_t^n,R_t^n+2).$$

We now introduce the sets of states $\widetilde{\mathcal{S}}_t^-$ and $\widetilde{\mathcal{S}}_t^+$. The states in $\widetilde{\mathcal{S}}_t^-$ ($\widetilde{\mathcal{S}}_t^+$) are the ones for which the projection operation decreased (increased) or kept the same the corresponding unprojected slopes infinitely often, that is, for $(P,R)\in\bar{\mathcal{S}}_t^*$, $\mathcal{N}_t^-(P,R)$ ($\mathcal{N}_t^+(P,R)$) is finite if and only if $(P,R)\in\widetilde{\mathcal{S}}_t^-(\widetilde{\mathcal{S}}_t^+)$. That is,

$$\widetilde{\mathcal{S}}_t^-=\{(P,R)\in\bar{\mathcal{S}}_t^*\colon z_t^n(P,R)\geq\bar{v}_t^n(P,R)\text{ for all }n\geq\bar{N}\}$$
$$\widetilde{\mathcal{S}}_t^+=\{(P,R)\in\bar{\mathcal{S}}_t^*\colon z_t^n(P,R)\leq\bar{v}_t^n(P,R)\text{ for all }n\geq\bar{N}\}.$$

Finally, we impose precisely the conditions that must be satisfied by the stepsizes α_t^n used to update the value function approximations. For $t < T$, the stepsizes satisfy the following conditions:

$$\alpha_t^n \in (0, 1] \qquad \text{and} \qquad \alpha_t^n \in \mathcal{F}_t^n, \tag{7}$$

$$\sum_{n=0}^{\infty} (\alpha_t^n)^2 \leq B < \infty \quad \text{a.s.}, \tag{8}$$

where B is a constant. We also require that

$$\sum_{n=0}^{\infty} \alpha_t^n 1_{\{P_t^n = P^*,\, R_t^n = R^*\}} = \infty \quad \text{a.s.}, \tag{9}$$

where (P^*, R^*) is an accumulation point of the sequence $\{(P_t^n, R_t^n)\}_{n \geq 0}$.

For example, the stepsize rule $\alpha_t^n = 1/(N(P_t^n, R_t^n) + 1)$ satisfies conditions (7)–(9), where $N(P_t^n, R_t^n)$ is the number of visits to state (P_t^n, R_t^n) up until iteration n.

For ease of notation in the next sections, we define a new stepsize sequence $\bar{\alpha}_t^n$ based on the previous one. For $t < T$ and $(P, R) \in \mathcal{S}_t$, let

$$\bar{\alpha}_t^n(P, R) = \alpha_t^n(1_{\{P = P_t^n,\, R = R_t^n\}} + 1_{\{P = P_t^n,\, R = R_t^n + 1\}}).$$

Note that though α_t^n is a scalar, $\bar{\alpha}_t^n$ is a vector with arguments $(P, R) \in \mathcal{S}_t$.

Based on the assumptions (7)–(9), we can trivially prove that $\bar{\alpha}_t^n(P, R) \in [0, 1]$ is $\mathcal{F}_t^n$-measurable and, on $\{(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^*\}$,

$$\sum_{n=0}^{\infty} \bar{\alpha}_t^n(\bar{P}^*, \bar{R}^*)^2 \leq B \quad \text{a.s.} \qquad \text{and} \qquad \sum_{n=0}^{\infty} \bar{\alpha}_t^n(\bar{P}^*, \bar{R}^*) = \infty \quad \text{a.s.} \tag{10}$$

Furthermore, for all positive integers N,

$$\prod_{n=N}^{\infty} (1 - \bar{\alpha}_t^n(\bar{P}^*, \bar{R}^*)) = 0 \quad \text{a.s.} \tag{11}$$

The proof for Equation (11) follows directly from the fact that $\log(1 + x) \leq x$.

As a final remark, we can easily see that $\hat{v}_t^n(R)$, $z_t^n(P, R)$, and $\bar{v}_t^n(P, R)$ are bounded by zero and $\max \hat{r}$ for all iterations n, because the initial approximations are bounded by zero and $\max \hat{r}$ and the stepsizes are between zero and one.

5. Sketch of convergence analysis. We introduce the convergence results we want to prove and sketch the proofs, summarizing the steps that will be used. The full proofs are given in §6.

We are after two main convergence results. The first one is, for each $t < T$ and on $\{(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^*\}$,

$$\bar{v}_t^n(\bar{P}^*, \bar{R}^*) \to v_t^*(\bar{P}^*, \bar{R}^*) \quad \text{a.s.} \tag{12}$$

The second result is, on the event that $(R_{t-1}^*, P_t^*, x_t^*)$ is an accumulation point of the sequence $\{(R_{t-1}^n, P_t^n, x_t^n)\}_{n \geq 0}$,

$$x_t^* = \arg\max_{0 \leq x \leq M_t} -P_t^* x + V_t^*(P_t^*, R_{t-1}^* + x) \quad \text{a.s.}, \tag{13}$$

where V_t^* is the optimal value function.

We use a pointwise argument in all the proofs of almost sure convergence presented in this paper. Thus, we disregard zero-measure events on an as-needed basis.

Equation (13) shows that indeed the algorithm has learned the optimal decision for all states that can be reached by an optimal policy, even if there are two or more recurrent classes of states. It is easy to see this implication. Starting with $t = 0$, we have by assumption that $R_{-1}^* = 0$, as $R_{-1}^n = 0$ for all iterations of the algorithm. Moreover, all prices in $\mathcal{P}_0$ are accumulation points of $\{P_0^n\}_{n \geq 0}$. Thus, Equation (13) tells us that the accumulation points x_0^* of the sequence $\{x_0^n\}$ along the iterations with initial price P_0^* are in fact an optimal policy for period zero when the price is P_0^*. This implies that all accumulation points $R_0^* = x_0^*$ of $\{R_0^n\}_{n \geq 0}$ are asset levels that can be reached by an optimal policy. By the same token, for $t = 1$, every price in $\mathcal{P}_1$ is an accumulation point of $\{P_1^n\}_{n \geq 0}$. Hence, the second result tells us that the accumulation points x_1^* of the sequence $\{x_1^n\}$ along iterations with $(R_0^n, P_1^n) = (R_0^*, P_1^*)$ are indeed an optimal policy for period one when the asset level is R_0^* and

the price is P_1^*. As before, the accumulation points $R_1^* = R_0^* + x_1^*$ of $\{R_1^n\}_{n\geq 0}$ are asset levels that can be reached by an optimal policy. The same reasoning can be applied for $t = 2, \ldots, T-1$.

The main idea to achieve Equation (12) is to define for each $t < T$ and $(P, R) \in \bar{\mathscr{S}}_t$ (introduced in §2) deterministic sequences $\{L_t^k(P,R)\}_{k\geq 0}$ and $\{U_t^k(P,R)\}_{k\geq 0}$ that are provably convergent to $v_t^*(P,R)$ and then prove, for all $k \geq 0$ and n big enough, that

$$L_t^k(\bar{P}^*, \bar{R}^*) \leq v_t^n(\bar{P}^*, \bar{R}^*) \leq U_t^k(\bar{P}^*, \bar{R}^*) \quad \text{a.s. on } \{(\bar{P}^*, \bar{R}^*) \in \bar{\mathscr{S}}_t^*\}. \tag{14}$$

Establishing these inequalities is nontrivial and draws on a proof technique in Bertsekas and Tsitsiklis [4, §4.3.6] (B&T). In our proof, however, we have to handle two significant differences. First, our algorithm uses a pure exploitation strategy whereas B&T assumed that all states are visited infinitely often. Second, we introduce a projection operator to maintain concavity of the approximation. This is not the case in B&T who assume a pure lookup table representation.

To establish Equation (14), we introduce the dynamic programming operator H associated with the asset acquisition problem and the deterministic bounding sequences $\{L^k\}_{k\geq 0}$ and $\{U^k\}_{k\geq 0}$. It is noteworthy that these sequences are completely independent of the algorithm. We also define four stochastic sequences, $\{\bar{s}_-^n\}_{n\geq 0}$, $\{\bar{s}_+^n\}_{n\geq 0}$, $\{\bar{l}^n\}_{n\geq 0}$, and $\{\bar{u}^n\}_{n\geq 0}$, which do depend on the iterations of the algorithm. The first two sequences are called stochastic noise sequences and the last two sequences are called stochastic bounding sequences.

All these elements are combined to prove Equation (14), where the concavity of the value functions plays a major role. Roughly speaking, using properties of the operator H, Lemma 4.1, and concavity, we prove

$$(HL^k)_t(P_t^n, R_t^n) \leq (H\bar{v}^{n-1})_t(P_t^n, R_t^n) \leq (HU^k)_t(P_t^n, R_t^n) \quad \text{a.s.},$$
$$(HL^k)_t(P_t^n, R_t^n + 1) \leq (H\bar{v}^{n-1})_t(P_t^n, R_t^n + 1) \leq (HU^k)_t(P_t^n, R_t^n + 1) \quad \text{a.s.}$$

These inequalities enable us to prove that

$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \leq \bar{u}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) + \bar{s}_{t-}^{n-1}(\tilde{P}^*, \tilde{R}^*) \quad \text{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^-\},$$
$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \geq \bar{l}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) - \bar{s}_{t+}^{n-1}(\tilde{P}^*, \tilde{R}^*) \quad \text{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^+\}.$$

Then, convergence to zero of the noise sequences (a convex combination property of the stochastic bounding sequences and concavity) gives us

$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \leq U_t^k(\tilde{P}^*, \tilde{R}^*) \quad \text{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^-\},$$
$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \geq L_t^k(\tilde{P}^*, \tilde{R}^*) \quad \text{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^+\}.$$

Finally, concavity plays a role again and we obtain Equation (14).

The optimality of the decisions with respect to the optimal value functions represented by Equation (13), is a byproduct of the convergence of the approximate slopes. It is discussed in detail in the next section.

6. Convergence analysis. We present formally the dynamic operator H and the deterministic bounding sequences $\{U^k\}_{k\geq 0}$ and $\{L^k\}_{k\geq 0}$ in §6.1. After that, in §6.2, we state and prove our main theorem, the almost sure convergence of the approximate slopes to the optimal slopes. As part of the proof, we define the stochastic sequences and state technical lemmas as they are needed. To focus on the main ideas of the theorem proof, the proofs of the lemmas will be deferred to Appendix B. Finally, in §6.3, we prove the almost-sure convergence to the optimal decisions.

6.1. The operator H and the bounding sequences. We start by defining the dynamic programming operator H that maps a vector v into a new vector Hv according to the formula

$$(Hv)_t(P,R) = \mathbb{E}[\max(\min(P_{t+1}, v_{t+1}(P_{t+1}, R)), v_{t+1}(P_{t+1}, R + M_{t+1}))1_{\{t<T-1\}} + \hat{r}1_{\{R\leq \hat{D}\}}1_{\{t=T-1\}} \mid P_t = P] \tag{15}$$

for $t = 0, \ldots, T-1$ and $(P,R) \in \bar{\mathscr{S}}_t$.

The following properties can be easily proved.

(i) H has a unique fixed point v^*, where v^* is the vector of slopes of the optimal value functions.

(ii) H is monotone, that is, if $v \leq \tilde{v}$ componentwise, then $Hv \leq H\tilde{v}$.

(iii) $Hv - \eta e \leq H(v - \eta e) \leq H(v + \eta e) \leq Hv + \eta e$, where η is a positive constant and e is a vector with all components equal to one. The inequalities are considered componentwise.

(iv) H is continuous.

We introduce the deterministic bounding sequences $\{U^k\}_{k\geq 0}$ and $\{L^k\}_{k\geq 0}$ and establish three important properties. When we refer to the sequence $\{U^k\}_{k\geq 0}$ without mentioning the time index t and the state $(P,R)\in\bar{\mathcal{S}}_t$, we are referring to the family of sequences $\{U_t^k(P,R)\}_{k\geq 0}$, one for each time $t<T$ and state (P,R). The same is true with the other deterministic sequence $\{L^k\}_{k\geq 0}$.

Let

$$U^0 = v^* + \hat{r}^* e \qquad \text{and} \qquad U^{k+1} = \frac{U^k + HU^k}{2}, \quad k\geq 0, \tag{16}$$

$$L^0 = v^* - \hat{r}^* e \qquad \text{and} \qquad L^{k+1} = \frac{L^k + HL^k}{2}, \quad k\geq 0, \tag{17}$$

where $\hat{r}^* = \max_{\omega\in\Omega}\hat{r}(\omega)$ is well-defined because $\hat{r}$ is a positive bounded random variable that represents the reward.

Note that just like the slopes v^*, for all $k\geq 0$, L^k and U^k are both monotone decreasing in the asset dimension. In the next lemma, the inequality signs applies to all time index t and states (P,R).

LEMMA 6.1. *The sequences $\{U^k\}_{k\geq 0}$ and $\{L^k\}_{k\geq 0}$ satisfy*

$$HU^k \leq U^{k+1} \leq U^k \tag{18}$$

$$HL^k \geq L^{k+1} \geq L^k \tag{19}$$

and both converge to v^*. Furthermore, $U^k > v^*$ and $L^k < v^*$ for all $k\geq 0$.

PROOF. The proof of inequalities (18) and (19) as well as the proof of convergence of the sequences to v^* is given in Bertsekas and Tsitsiklis [4, Lemmas 4.5 and 4.6]. They just require the above-mentioned four properties of operator H.

In order to show that $L^k < v^*$ for all $k\geq 0$, we begin by analyzing L_{T-1}^k. By definition of H, for all $(P,R)\in\bar{\mathcal{S}}_{T-1}$, $(HL^k)_{T-1}(P,R) = v_{T-1}^*(P,R)$ for all $k\geq 0$. We also have that $L_{T-1}^0(P,R) = v_{T-1}^*(P,R) - \hat{r}^* < v_{T-1}^*(P,R)$. Thus, $L_{T-1}^1(P,R) < v_{T-1}^*(P,R)$ and an induction argument on k shows that $L_{T-1}^k(P,R) < v_{T-1}^*(P,R)$ for all $k\geq 0$.

Now, assume that $L_{t+1}^k(P,R) < v_{t+1}^*(P,R)$ for all $k\geq 0$ and $(P,R)\in\bar{\mathcal{S}}_{t+1}$. We prove $(HL^k)_t(P,R) \leq v_t^*(P,R)$ for t when $t=0,\ldots,T-2$. We have, for $(P,R)\in\bar{\mathcal{S}}_t$,

$$\begin{aligned}(HL^k)_t(P,R) &= \mathbb{E}[\max(\min(P_{t+1}, L_{t+1}(P_{t+1},R)), L_{t+1}(P_{t+1}, R+M_{t+1})) \mid P_t = P] \\ &\leq \mathbb{E}[\max(\min(P_{t+1}, v_{t+1}^*(P_{t+1},R)), v_{t+1}^*(P_{t+1}, R+M_{t+1})) \mid P_t = P] = v_t^*(P,R).\end{aligned}$$

Furthermore, $L_t^0(P,R) = v_t^*(P,R) - \hat{r}^* < v_t^*(P,R)$, which implies $L_t^1(P,R) < v_t^*(P,R)$. Again, an induction argument on k shows that $L_t^k(P,R) < v_t^*(P,R)$ for all $k\geq 0$. The proof for U^k follows by a symmetrical argument. □

6.2. Convergence of $\bar{v}_t^n(\bar{P}^*,\bar{R}^*)$. We prove almost-sure convergence of the slopes of the approximate functions to the slopes of the optimal ones on the event $\{(\bar{P}^*,\bar{R}^*)\in\bar{\mathcal{S}}_t^*\}$. In the process, we present the noise and the bounding stochastic sequences. We also introduce three technical lemmas. Their proofs are given in Appendix B. We assume for integers $k\geq 0$, iterations $n\geq 0$ and all possible states (P,R) that $v_T^*(P,R) = U_T^k(P,R) = L_T^k(P,R) = \bar{v}_T^n(P,R) = 0$, as we only need to learn the slopes up until time period $T-1$.

THEOREM 6.1. *Assume the stepsize conditions* (7)–(9). *Fix $t\in\{0,\ldots,T\}$ and $k\geq 0$. Then, there exists an almost surely finite random index $N_t^{*,k}$ such that for all $(\bar{P}^*,\bar{R}^*)\in\bar{\mathcal{S}}_t$, on the event that $\{n\geq N_t^{*,k}, (\bar{P}^*,\bar{R}^*)\in\bar{\mathcal{S}}_t^*\}$, it holds that*

$$L_t^k(\bar{P}^*,\bar{R}^*) \leq \bar{v}_t^{n-1}(\bar{P}^*,\bar{R}^*) \leq U_t^k(\bar{P}^*,\bar{R}^*) \quad a.s. \tag{20}$$

Therefore, on $\{(\bar{P}^*,\bar{R}^*)\in\bar{\mathcal{S}}_t^*\}$,

$$\bar{v}_t^n(\bar{P}^*,\bar{R}^*) \to v_t^*(\bar{P}^*,\bar{R}^*) \quad a.s. \tag{21}$$

PROOF. We show the result for each element in the sample space fixing $\omega\in\Omega$. The proof of the theorem is by backward induction on t. The base case is $t=T$. For any state $(P,R)\in\bar{\mathcal{S}}_t$, integers $k\geq 0$ and $n\geq 0$, it holds that $v_T^*(P,R) = U_T^k(P,R) = L_T^k(P,R) = \bar{v}_T^n(P,R) = 0$. Therefore, the inequalities in Equation (20) are trivial for $t=T$, all $k\geq 0$, and $(P,R)\in\bar{\mathcal{S}}_t$. Thus, for a fixed k, we can pick for example $N_T^{*,k} = \bar{N}$, where $\bar{N}$ as

defined in §4 is an almost surely finite random index that denotes when an iteration of the algorithm is large enough for convergence analysis purposes.

The backward induction proof is completed when we prove, for all $k \geq 0$, that Equation (20) holds for $t = T-1, \ldots, 0$. Given the induction hypothesis for $t+1$, the proof for time period t is divided into two parts. First, for a fixed $k \geq 0$, we prove that there exists an almost surely finite random index N_t^k such that for all states $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t$,

$$\bar{v}_t^{n-1}(\bar{P}^*, \bar{R}^*) \leq U_t^k(\bar{P}^*, \bar{R}^*), \quad \text{a.s. on } \{n \geq N_t^k,\ (\bar{P}^*, \bar{R}^*) \in \widetilde{\mathcal{S}}_t^-\}, \tag{22}$$

$$\bar{v}_t^{n-1}(\bar{P}^*, \bar{R}^*) \geq L_t^k(\bar{P}^*, \bar{R}^*), \quad \text{a.s. on } \{n \geq N_t^k,\ (\bar{P}^*, \bar{R}^*) \in \widetilde{\mathcal{S}}_t^+\}. \tag{23}$$

The proof is by induction on k. Note that it only applies to states in the sets $\widetilde{\mathcal{S}}_t^-$ and $\widetilde{\mathcal{S}}_t^+$.

Then, again for time period t, we show that the upper and the lower bounds also hold for states in $\bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^-$ and $\bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^+$, respectively. We prove for a fixed $k \geq 0$ and state $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t$ the existence of random indices $N_t^{k,u}(\bar{P}^*, \bar{R}^*)$ and $N_t^{k,l}(\bar{P}^*, \bar{R}^*)$ such that

$$\bar{v}_t^{n-1}(\bar{P}^*, \bar{R}^*) \leq U_t^k(\bar{P}^*, \bar{R}^*), \quad \text{a.s. on } \{n \geq N_t^{k,u}(\bar{P}^*, \bar{R}^*),\ (\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^-\}$$

$$\bar{v}_t^{n-1}(\bar{P}^*, \bar{R}^*) \geq L_t^k(\bar{P}^*, \bar{R}^*), \quad \text{a.s. on } \{n \geq N_t^{k,l}(\bar{P}^*, \bar{R}^*),\ (\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^+\}.$$

Therefore, Parts 1 and 2 are put together when we take $N_t^{*,k}$ to be the maximum element of the set

$$\left\{ N_t^k, \max_{(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^-} N_t^{k,u}(\bar{P}^*, \bar{R}^*), \max_{(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^+} N_t^{k,l}(\bar{P}^*, \bar{R}^*) \right\},$$

proving that Equation (20) is true on $\{n \geq N_t^{*,k},\ (\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^*\}$. Figure 3 shows the relationship between the sets of states.

We emphasize that the proof of the theorem consists of two loops. The outside loop is the backward induction on t, where the base case is $t = T$. The inside loop is, for a fixed t, the forward induction on k in the proof of Part 1, where the base case is $k = 0$. Of course, the proof of Part 2 is also contained in the induction on t. We also want to point out that though the indices N_t^k and $N_t^{*,k}$ are independent of the state, the indices $N_t^{k,u}(\bar{P}^*, \bar{R}^*)$ and $N_t^{k,l}(\bar{P}^*, \bar{R}^*)$ as indicated by the notation are state dependent.

We start the backward induction on t. Pick $\omega \in \Omega$. We omit the dependence of the random elements on ω for compactness. Remember that the base case $t = T$ is trivial and we pick $N_T^{*,k} = \bar{N}$. We also pick, for convenience, $N_T^k = \bar{N}$.

Induction Hypothesis. Fix $t \in \{0, \ldots, T-1\}$. For each $k \geq 0$, assume the existence of integers N_{t+1}^k and $N_{t+1}^{*,k}$ such that, for all $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_{t+1}$ and $n \geq N_{t+1}^k$, Equations (22) and (23) are true when $(\tilde{P}^*, \tilde{R}^*)$ is an element of $\widetilde{\mathcal{S}}_{t+1}^-$ and $\widetilde{\mathcal{S}}_{t+1}^+$, respectively. Moreover, when $n \geq N_{t+1}^{*,k}$, the inequalities in Equation (20) hold true for all states $(\bar{P}^*, \bar{R}^*) \in \mathcal{S}_{t+1}^*$.

(i) Part 1 of the induction hypothesis proof.

Assuming the induction hypothesis on $t+1$, we prove for time period t, a fixed $k \geq 0$ and any state $(\tilde{P}^*, \tilde{R}^*) \in \tilde{\mathcal{S}}_t$, the existence of an integer N_t^k such that for $n \geq N_t^k$, inequalities (22) and (23) are true. The proof is by forward induction on k.

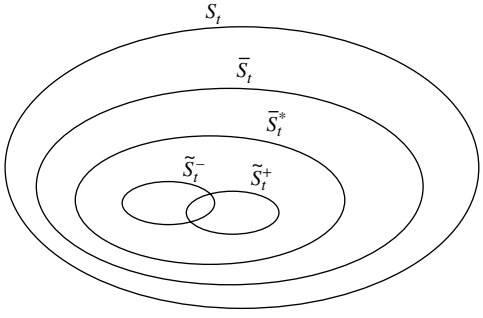


Figure 3. Relationship between the sets of states.

Notes. S_t: Full state space. $\bar{S}_t$: State space minus $(P, 0)$ pairs. $\bar{S}_t^*$: Accumulation point (P^*, R^*) or $(P^*, R^* + 1)$ of $\{(P^n, R^n)\}$. $\tilde{S}_t^-$: Corresponding slope is increased finitely often due to projection operation. $\tilde{S}_t^+$: Corresponding slope is decreased finitely often due to projection operation.

We start with $k = 0$. For every $(\tilde{P}^*, \tilde{R}^*) \in \bar{\mathscr{S}}_t$, $0 \le v_t^*(\tilde{P}^*, \tilde{R}^*) \le \hat{r}^*$ implying, by definition, that $U_t^0(\tilde{P}^*, \tilde{R}^*) \ge \hat{r}^*$ and $L_t^0(\tilde{P}^*, \tilde{R}^*) \le 0$. Therefore, Equations (22) and (23) are satisfied for all $n \ge 1$ because we know that $\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*)$ is bounded by zero and $\hat{r}^*$ for all iterations. Thus, $N_t^0 = \max(1, N_{t+1}^{*,0}) = N_{t+1}^{*,0}$.

The induction hypothesis on k assumes that there exists N_t^k such that for all $n \ge N_t^k$, Equations (22) and (23) are true. Note that we can always make N_t^k larger than $N_{t+1}^{*,k}$, thus we assume that $N_t^k \ge N_{t+1}^{*,k}$. The next step is the proof for $k+1$, i.e., we prove that there exists an integer N_t^{k+1} such that for all $(\tilde{P}^*, \tilde{R}^*) \in \bar{\mathscr{S}}_t$

$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \le U_t^{k+1}(\tilde{P}^*, \tilde{R}^*), \quad \text{if } n \ge N_t^{k+1} \text{ and } (\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^-,$$
$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \ge L_t^{k+1}(\tilde{P}^*, \tilde{R}^*), \quad \text{if } n \ge N_t^{k+1} \text{ and } (\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^+.$$

Before we move on, we depart from our pointwise argument in order to define the stochastic noise sequences and state a lemma describing an important property of these sequences. We start defining $\hat{s}_{t+1-}^n$ and $\hat{s}_{t+1+}^n$ to be the error incurred by observing a sample slope. For $R = 1, \ldots, B_t$,

$$\hat{s}_{t+1-}^n(R) = \hat{v}_{t+1}^n(R) - (H\bar{v}^{n-1})_t(P_t^n, R) \qquad \text{and} \qquad \hat{s}_{t+1+}^n(R) = -\hat{s}_{t+1-}^n(R).$$

Using $\hat{s}_{t+1-}^n$ and $\hat{s}_{t+1+}^n$, we also define the stochastic noise sequences $\{\bar{s}_{t-}^n\}_{n\ge0}$ and $\{\bar{s}_{t+}^n\}_{n\ge0}$. For $(P, R) \in \bar{\mathscr{S}}_t$,

$$\bar{s}_{t-}^n(P, R) = 0 \qquad \text{and} \qquad \bar{s}_{t+}^n(P, R) = 0 \quad \text{on } \{n < N_t^k\},$$

and, on $\{n \ge N_t^k\}$,

$$\bar{s}_{t-}^n(P, R) = \max(0, (1 - \bar{\alpha}_t^n(P, R))\bar{s}_{t-}^{n-1}(P, R) + \bar{\alpha}_t^n(P, R)\hat{s}_{t+1-}^n(R_t^n 1_{\{R \le R_t^n\}} + (R_t^n + 1)1_{\{R > R_t^n\}}))$$
$$\bar{s}_{t+}^n(P, R) = \max(0, (1 - \bar{\alpha}_t^n(P, R))\bar{s}_{t+}^{n-1}(P, R) + \bar{\alpha}_t^n(P, R)\hat{s}_{t+1+}^n(R_t^n 1_{\{R \le R_t^n\}} + (R_t^n + 1)1_{\{R > R_t^n\}})).$$

Remember that $\bar{\alpha}_t^n(P, R) = 0$, except when $(P, R) \in \{(P_t^n, R_t^n), (P_t^n, R_t^n + 1)\}$.

The sample slopes are defined in a way such that

$$\mathbb{E}[\hat{s}_{t+1-}^n(R) \mid \mathscr{F}_t^n] = 0. \tag{24}$$

This conditional expectation expresses the condition that the sample slopes are unbiased. This property, together with the martingale convergence theorem and the boundedness of both the sample slopes and the approximate slopes, is crucial for proving that the noise introduced by the observation of the sample slopes, which replace the observation of true expectations, goes to zero as the number of iterations of the algorithm goes to infinity. This is stated in the next lemma.

LEMMA 6.2. *Pick a state $(\bar{P}^*, \bar{R}^*) \in \bar{\mathscr{S}}_t$. Then, on $\{(\bar{P}^*, \bar{R}^*) \in \bar{\mathscr{S}}_t^*\}$,*

$$\{\bar{s}_{t-}^n(\bar{P}^*, \bar{R}^*)\}_{n\ge0} \to 0 \qquad \textit{and} \qquad \{\bar{s}_{t+}^n(\bar{P}^*, \bar{R}^*)\}_{n\ge0} \to 0 \quad \textit{a.s.} \tag{25}$$

PROOF OF LEMMA 6.2. The proof is given in Appendix B. □

Returning to our pointwise argument where we have fixed $\omega \in \Omega$, we use the convention that the minimum of an empty set is $+\infty$. Let

$$\delta_L^k = \min\left\{\frac{(HL^k)_t(\tilde{P}^*, \tilde{R}^*) - L_t^k(\tilde{P}^*, \tilde{R}^*)}{4} : (\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^+,\ (HL^k)_t(\tilde{P}^*, \tilde{R}^*) > L_t^k(\tilde{P}^*, \tilde{R}^*)\right\}.$$

If $\delta_L^k < +\infty$, we define an integer $N_L \ge N_t^k$ to be such that

$$\prod_{m=N_t^k}^{N_L-1} (1 - \bar{\alpha}_t^m(\tilde{P}^*, \tilde{R}^*)) \le 1/4 \qquad \text{and} \qquad \bar{s}_{t+}^{n-1}(\tilde{P}^*, \tilde{R}^*) \le \delta_L^k \tag{26}$$

for all $n \ge N_L$ and states $(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^+$. Such an N_L exists because both Equation (11) and Equation (25) are true. If $\delta_L^k = +\infty$, then for all states $(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathscr{S}}_t^+$, $(HL^k)_t(\tilde{P}^*, \tilde{R}^*) = L_t^k(\tilde{P}^*, \tilde{R}^*)$ because Equation (19) tells us that $HL^k \ge L^k$. Thus, $L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) = L_t^k(\tilde{P}^*, \tilde{R}^*)$ and we define the integer N_L to be equal to N_t^k.

We can apply a symmetric reasoning to determine δ_U^k and N_U. We just need to consider the deterministic bounding sequence $\{U^k\}_{k\ge0}$, the set $\widetilde{\mathscr{S}}_t^-$, and the noise sequence $\{\bar{s}_{t-}^n\}_{n\ge0}$ instead of $\{L^k\}_{k\ge0}$, $\widetilde{\mathscr{S}}_t^+$, and $\{\bar{s}_{t+}^n\}_{n\ge0}$, respectively.

Finally, let $N_t^{k+1} = \max(N_L, N_U, N_{t+1}^{*,k+1})$. We pick a state $(\tilde{P}^*, \tilde{R}^*) \in \bar{\mathcal{S}}_t$ and assume that $(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathcal{S}}_t^+$. If $L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) = L_t^k(\tilde{P}^*, \tilde{R}^*)$, then inequality $L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) \le \bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*)$ follows from the induction hypothesis. We therefore concentrate on the case where $L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) > L_t^k(\tilde{P}^*, \tilde{R}^*)$.

First, we depart one more time from the pointwise argument to introduce the stochastic bounding sequences $\{\bar{l}_t^n\}_{n\ge 0}$ and $\{\bar{u}_t^n\}_{n\ge 0}$. We also state a lemma combining these sequences with the stochastic noise sequences. For each $(P, R) \in \bar{\mathcal{S}}_t$, we have

$$\bar{l}_t^n(P, R) = L_t^k(P, R) \qquad \text{and} \qquad \bar{u}_t^n(P, R) = U_t^k(P, R) \quad \text{on } \{n < N_t^k\},$$

and on $\{n \ge N_t^k\}$,

$$\bar{l}_t^n(P, R) = (1 - \bar{\alpha}_t^n(P, R))\bar{l}_t^{n-1}(P, R) + \bar{\alpha}_t^n(P, R)(HL^k)_t(P, R)$$
$$\bar{u}_t^n(P, R) = (1 - \bar{\alpha}_t^n(P, R))\bar{u}_t^{n-1}(P, R) + \bar{\alpha}_t^n(P, R)(HU^k)_t(P, R).$$

The next lemma states that the stochastic bounding and noise sequences can be used to provide a bound for the approximate slopes as follows.

LEMMA 6.3. *On* $\{n \ge N_t^k\}$,

$$(HL^k)_t(P_t^n, R_t^n) \le (H\bar{v}^{n-1})_t(P_t^n, R_t^n) \le (HU^k)_t(P_t^n, R_t^n) \quad \textit{a.s. on } \{R_t^n > 0\}$$
$$(HL^k)_t(P_t^n, R_t^n + 1) \le (H\bar{v}^{n-1})_t(P_t^n, R_t^n + 1) \le (HU^k)_t(P_t^n, R_t^n + 1) \quad \textit{a.s. on } \{R_t^n < M_t\}.$$

Moreover, again on $\{n \ge N_t^k\}$,

$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \le \bar{u}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) + \bar{s}_{t-}^{n-1}(\tilde{P}^*, \tilde{R}^*) \quad \textit{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathcal{S}}_t^-\}, \tag{27}$$
$$\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) \ge \bar{l}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) - \bar{s}_{t+}^{n-1}(\tilde{P}^*, \tilde{R}^*) \quad \textit{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathcal{S}}_t^+\}. \tag{28}$$

PROOF OF LEMMA 6.3. The proof is given in Appendix B. □

Back to our fixed ω, a simple inductive argument proves that $\bar{u}_t^n(P, R)$ is a convex combination of $U_t^k(P, R)$ and $(HU^k)_t(P, R)$, and $\bar{l}_t^n(P, R)$ is a convex combination of $L_t^k(P, R)$ and $(HL^k)_t(P, R)$. Therefore, we can write, with $\tilde{b}_t^{n-1} = \prod_{m=N_t^k}^{n-1}(1 - \bar{\alpha}_t^m(\tilde{P}^*, \tilde{R}^*))$,

$$\bar{l}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) = \tilde{b}_t^{n-1}L_t^k(\tilde{P}^*, \tilde{R}^*) + (1 - \tilde{b}_t^{n-1})(HL^k)_t(\tilde{P}^*, \tilde{R}^*).$$

For $n \ge N_t^{k+1} \ge N_L \ge N_t^k$, we have $\tilde{b}_t^{n-1} \le 1/4$. Moreover, $L_t^k(\tilde{P}^*, \tilde{R}^*) \le (HL^k)_t(\tilde{P}^*, \tilde{R}^*)$. Thus, using Equation (17) and the definition of δ_L^k, we obtain

$$\begin{aligned}\bar{l}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) &\ge \tfrac{1}{4}L_t^k(\tilde{P}^*, \tilde{R}^*) + \tfrac{3}{4}(HL^k)_t(\tilde{P}^*, \tilde{R}^*) \\ &= \tfrac{1}{2}L_t^k(\tilde{P}^*, \tilde{R}^*) + \tfrac{1}{2}(HL^k)_t(\tilde{P}^*, \tilde{R}^*) + \tfrac{1}{4}((HL^k)_t(\tilde{P}^*, \tilde{R}^*) - L_t^k(\tilde{P}^*, \tilde{R}^*)) \\ &\ge L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) + \delta_L^k. \end{aligned} \tag{29}$$

We point out that we are concentrating on the case where $L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) > L_t^k(\tilde{P}^*, \tilde{R}^*)$, implying that $\delta_L^k < \infty$ as argued when δ_L^k was defined. Combining Equations (28) and (29), we obtain, for all $n \ge N_t^{k+1} \ge N_L \ge N_t^k$,

$$\begin{aligned}\bar{v}_t^{n-1}(\tilde{P}^*, \tilde{R}^*) &\ge L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) + \delta_L^k - \bar{s}_{t+}^{n-1}(\tilde{P}^*, \tilde{R}^*) \\ &\ge L_t^{k+1}(\tilde{P}^*, \tilde{R}^*) + \delta_L^k - \delta_L^k \\ &= L_t^{k+1}(\tilde{P}^*, \tilde{R}^*),\end{aligned}$$

where the last inequality follows from Equation (26).

To finish the proof of Part 1 of the induction hypothesis, we pick a state $(\tilde{P}^*, \tilde{R}^*) \in \bar{\mathcal{S}}_t$ and assume that $(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathcal{S}}_t^-$. The reasoning for $U_t^{k+1}(\tilde{P}^*, \tilde{R}^*)$ is symmetrical to that for $L_t^{k+1}(\tilde{P}^*, \tilde{R}^*)$, which completes our induction. Thus, we have proved that, for all $k \ge 0$, there exists N_t^k such that (22) and (23) hold for all $n \ge N_t^k$. This concludes the first part of the proof.

(ii) Part 2 of the induction hypothesis proof.

We continue to consider ω picked in the beginning of the proof of the Theorem 6.1. In this part, we take care of the states $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^-$ and $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \backslash \widetilde{\mathcal{S}}_t^+$. In contrast to Part 1, the proof technique here is not by forward induction on k.

A discussion about the projection operation is in order, as this part of the proof is all about states for which the projection operation decreased or increased the corresponding approximate slopes infinitely often.

Remember that at iteration n time period t, we observe the sample slopes $\hat{v}^n_{t+1}(R^n_t)$ and $\hat{v}^n_{t+1}(R^n_t+1)$ and it is always the case that $\hat{v}^n_{t+1}(R^n_t) \geq \hat{v}^n_{t+1}(R^n_t+1)$, implying that the resulting temporary slope $z^n_t(P^n_t, R^n_t)$ is bigger than $z^n_t(P^n_t, R^n_t+1)$. Therefore, according to our projection operator, the updated slopes $\bar{v}^n_t(P^n_t, R^n_t)$ and $\bar{v}^n_t(P^n_t, R^n_t+1)$ are always equal to $z^n_t(P^n_t, R^n_t)$ and $z^n_t(P^n_t, R^n_t+1)$, respectively. Because of our stepsize rule, the slopes corresponding to (P^n_t, R^n_t) and (P^n_t, R^n_t+1) are the only ones updated because of a direct observation of samples slopes at iteration n time period t. All the other slopes are modified only if a violation of the monotone decreasing property occurs. Therefore, the slopes corresponding to states with price $P_t \in \mathcal{P}_t$ different than P^n_t, no matter the asset level $R = 1, \ldots, B_t$, remain the same at iteration n time period t, that is, $\bar{v}^{n-1}_t(P_t, R) = z^n_t(P_t, R) = \bar{v}^n_t(P_t, R)$. On the other hand, it is always the case that the temporary slopes corresponding to states with price P^n_t and asset levels smaller than R^n_t can only be increased by the projection operation. If necessary, they are increased to be equal to $\bar{v}^n_t(P^n_t, R^n_t)$. Similarly, the temporary slopes corresponding to states with price P^n_t and asset levels greater than R^n_t+1 can only be decreased by the projection operation. If necessary, they are decreased to be equal to $\bar{v}^n_t(P^n_t, R^n_t+1)$ (see Figure 2(c)).

Keeping the previous discussion in mind, we can argue that $\widetilde{\mathcal{S}}^+_t$ is a nonempty set. It is easy to see that for each $\bar{P}^* \in \mathcal{P}_t$, if R^{Min} is the minimum asset level such that $(\bar{P}^*, R^{\text{Min}})$ is an accumulation point of $\{(P^n_t, R^n_t)\}_{n\geq 0}$, then the slope corresponding to $(\bar{P}^*, R^{\text{Min}})$ could only be decreased by the projection operation a finite number of iterations, as a decreasing requirement could only be originated from an asset level smaller than R^{Min}. However, no state with price $\bar{P}^*$ and asset level smaller than R^{Min} is visited by the algorithm after iteration $\bar{N}$ because only accumulation points are visited after $\bar{N}$. We thus have that $(\bar{P}^*, R^{\text{Min}})$ is an element of the set $\widetilde{\mathcal{S}}^+_t$. Along the same lines, we can show that $\widetilde{\mathcal{S}}^-_t$ is also a nonempty set.

On that note, if we pick a state $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t$ and assume that $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}^*_t \backslash \widetilde{\mathcal{S}}^+_t$, then there exists another state $(\bar{P}^*, \tilde{R}^*_+)$ where $\tilde{R}^*_+$ is the maximum asset level smaller than $\bar{R}^*$ such that $(\bar{P}^*, \tilde{R}^*_+) \in \widetilde{\mathcal{S}}^+_t$. It could be the case that $(\bar{P}^*, \tilde{R}^*_+) = (\bar{P}^*, R^{\text{Min}})$. Moreover, we show next that for all asset levels R such that $\tilde{R}^*_+ < R \leq \bar{R}^*$, we have that $|\mathcal{N}^+_t(\bar{P}^*, R)| = \infty$. Figure 4 illustrates the situation. A symmetrical property holds if we assume that $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}^*_t \backslash \widetilde{\mathcal{S}}^-_t$.

The argument goes as follows. For any state $(P, R) \in \bar{\mathcal{S}}_t$, remember that $\mathcal{N}^+_t(P, R) = \{n \in \mathbb{N}\colon z^n_t(P, R) > \bar{v}^n_t(P, R)\}$. As discussed in §4, the sets $\widetilde{\mathcal{S}}^+_t$ and $\mathcal{N}^+_t(P, R)$ share the following relationship. Assuming that $(P, R) \in \bar{\mathcal{S}}^*_t$, then $|\mathcal{N}^+_t(P, R)| = \infty$ if and only if the state (P, R) is not an element of $\widetilde{\mathcal{S}}^+_t$. Because we have assumed that $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}^*_t \backslash \widetilde{\mathcal{S}}^+_t$, we have that $|\mathcal{N}^+_t(\bar{P}^*, \bar{R}^*)| = \infty$. If $\tilde{R}^*_+ = \bar{R}^* - 1$, we are done. If $\tilde{R}^*_+ < \bar{R}^* - 1$, we have to consider two cases, namely, $(\bar{P}^*, \bar{R}^* - 1) \in \bar{\mathcal{S}}^*_t$ and $(\bar{P}^*, \bar{R}^* - 1) \notin \bar{\mathcal{S}}^*_t$. For the first case, we have that $|\mathcal{N}^+_t(\bar{P}^*, \bar{R}^* - 1)| = \infty$ from the fact that this state is not an element of $\widetilde{\mathcal{S}}^+_t$. For the second case, because $(\bar{P}^*, \bar{R}^* - 1)$ is not an element of $\bar{\mathcal{S}}^*_t$, its corresponding slope is never updated due to a direct observation of sample slopes for $n \geq \bar{N}$, by the definition of $\bar{N}$. Moreover, every time the slope of $(\bar{P}^*, \bar{R}^*)$ is decreased due to a projection (which is coming from the left), the slope of $(\bar{P}^*, \bar{R}^* - 1)$ has to be decreased as well. Therefore, $\mathcal{N}^+_t(\bar{P}^*, \bar{R}^*) \cap \{n \geq \bar{N}\} \subseteq \mathcal{N}^+_t(\bar{P}^*, \bar{R}^* - 1) \cap \{n \geq \bar{N}\}$, implying that $|\mathcal{N}^+_t(\bar{P}^*, \bar{R}^* - 1)| = \infty$. We then apply the same reasoning for states $(\bar{P}^*, \bar{R}^* - 2), \ldots, (\bar{P}^*, \tilde{R}^*_+ + 1)$, obtaining that the corresponding sets of iterations have an infinite number of elements.

We introduce a lemma that is the key element for the proof of Part 2, once again going away from the pointwise argument.

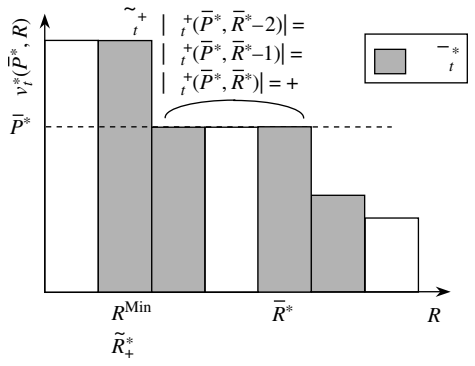


FIGURE 4. Optimal slopes and the corresponding $\mathcal{N}^+_t$ sets.

LEMMA 6.4. *For a given time period t, fix $k \geq 0$ and a state $(P, R) \in \bar{\mathcal{S}}_t$. If there exists a random index $N_t^{k,l}(P, R)$ such that $\bar{v}_t^{n-1}(P, R) \geq L_t^k(P, R)$ on $\{n \geq N_t^{k,l}(P, R), |\mathcal{N}_t^+(P, R+1)| = \infty\}$, then there exists another random index $N_t^{k,l}(P, R+1)$ such that $\bar{v}_t^{n-1}(P, R+1) \geq L_t^k(P, R+1)$ on $\{n \geq N_t^{k,l}(P, R+1)\}$.*

Similarly, if there exists a random index $N_t^{k,u}(P, R)$ such that $\bar{v}_t^{n-1}(P, R) \leq U_t^k(P, R)$ on $\{n \geq N_t^{k,u}(P, R), |\mathcal{N}_t^-(P, R-1)| = \infty\}$, then there exists another random index $N_t^{k,u}(P, R-1)$ such that $\bar{v}_t^{n-1}(P, R-1) \leq U_t^k(P, R-1)$ on $\{n \geq N_t^{k,u}(P, R-1)\}$.

PROOF OF LEMMA 6.4. The proof is given in Appendix B. □

Using the properties of the projection operator, we return to the proof of Part 2 and to our fixed ω. Pick $k \geq 0$ and a state $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t$. We assume that $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \setminus \tilde{\mathcal{S}}_t^+$. Consider the state $(\bar{P}^*, \tilde{R}_+^*)$ where $\tilde{R}_+^*$ is the maximum asset level smaller than $\bar{R}^*$ such that $(\bar{P}^*, \tilde{R}_+^*) \in \tilde{\mathcal{S}}_t^+$. Clearly, this state satisfies the condition of Lemma 6.4 with $N_t^{k,l}(\bar{P}^*, \tilde{R}_+^*) = N_t^k$ (from Part 1 of the proof of Theorem 6.1). Thus, we can apply this lemma in order to obtain an integer $N^{k,l}(\bar{P}^*, \tilde{R}_+^* + 1)$ such that $L_t^k(\bar{P}^*, \tilde{R}_+^* + 1) \leq \bar{v}_t^{n-1}(\bar{P}^*, \tilde{R}_+^* + 1)$, for all $n \geq N_t^{k,l}(\bar{P}^*, \tilde{R}_+^* + 1)$.

After that, we use Lemma 6.4 again, this time considering state $(\bar{P}^*, \tilde{R}_+^* + 1)$. Note that the first application of Lemma 6.4 gave us the integer $N_t^{k,l}(\bar{P}^*, \tilde{R}_+^* + 1)$ necessary to fulfill the conditions of this second usage of the lemma. We repeat the same reasoning, applying Lemma 6.4 successively to the states $(\bar{P}^*, \tilde{R}_+^* + 2), \ldots, (\bar{P}^*, \bar{R}^* - 1)$. In the end, we obtain an integer $N_t^{k,l}(\bar{P}^*, \bar{R}^*)$ such that $L_t^k(\bar{P}^*, \bar{R}^*) \leq \bar{v}_t^{n-1}(\bar{P}^*, \bar{R}^*)$ for all $n \geq N_t^{k,l}(\bar{P}^*, \bar{R}^*)$. Figure 5 illustrates this process.

Similarly, if we assume that $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \setminus \tilde{\mathcal{S}}_t^-$, by successive applications of the second part of Lemma 6.4 we obtain an integer $N_t^{k,u}(\bar{P}^*, \bar{R}^*)$ such that $\bar{v}_t^{n-1}(\bar{P}^*, \bar{R}^*) \leq U_t^k(\bar{P}^*, \bar{R}^*)$ for all $n \geq N_t^{k,u}(\bar{P}^*, \bar{R}^*)$, concluding Part 2 of the proof.

Finally, if we consider:

$$N_t^{*,k} = \max\left\{N_t^k, \max_{(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \setminus \tilde{\mathcal{S}}_t^-} N_t^{k,u}(\bar{P}^*, \bar{R}^*), \max_{(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \setminus \tilde{\mathcal{S}}_t^+} N_t^{k,l}(\bar{P}^*, \bar{R}^*)\right\},$$

then Equation (20) holds for all states $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t^*$ and $n \geq N_t^{*,k}$, concluding the induction on t. □

6.3. Optimality of the decisions. We are ready to prove Equation (13), the second convergence result.

THEOREM 6.2. *Assume the conditions of Theorem 1 are satisfied. For $t = 0, \ldots, T-1$, on the event that $(\bar{v}^*, R_{t-1}^*, P_t^*, x_t^*)$ is an accumulation point of the sequence $\{(\bar{v}^{n-1}, R_{t-1}^n, P_t^n, x_t^n)\}_{n \geq 1}$ generated by the algorithm, x_t^* is almost surely an optimal solution of*

$$\max_{0 \leq x \leq M_t} -P_t^* x + V_t^*(P_t^*, R_{t-1}^* + x). \tag{30}$$

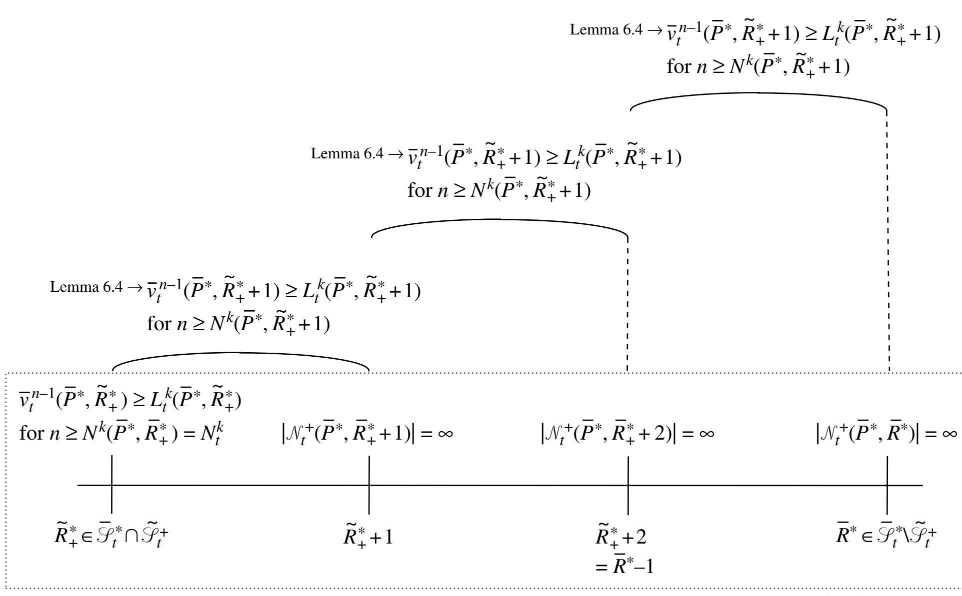


FIGURE 5. Successive applications of Lemma 6.4.

PROOF. Fix $\omega \in \Omega$. As before, the dependence on ω is omitted. At each iteration n and time t of the algorithm, the decision x_t^n is optimal with respect to the price P_t^n, the current asset level R_{t-1}^n, and the value function approximation for price P_t^n, which is piecewise linear with integer break points and is represented by its slopes $\bar{v}_t^n(P_t^n, 1), \ldots, \bar{v}_t^n(P_t^n, B_t)$. Therefore, it follows that either $-P_t^n + \bar{v}_t^n(P_t^n, R_{t-1}^n + x_t^n) \geq 0$ and $-P_t^n + \bar{v}_t^n(P_t^n, R_{t-1}^n + x_t^n + 1) \leq 0$ or

$$-P_t^n + \bar{v}_t^n(P_t^n, R_{t-1}^n) \leq 0, \quad \text{if } x_t^n = 0,$$
$$-P_t^n + \bar{v}_t^n(P_t^n, R_{t-1}^n + M_t + 1) \geq 0, \quad \text{if } x_t^n = M_t.$$

Then, by passing to the limit, we can conclude that each accumulation point $(\bar{v}^*, R_{t-1}^*, P_t^*, x_t^*)$ of the sequence $\{(\bar{v}^{n-1}, R_{t-1}^n, P_t^n, x_t^n)\}_{n\geq 1}$ satisfies either $-P_t^* + \bar{v}_t^*(P_t^*, R_{t-1}^* + x_t^*) \geq 0$ and $-P_t^* + \bar{v}_t^*(P_t^*, R_{t-1}^* + x_t^* + 1) \leq 0$ or

$$-P_t^* + \bar{v}_t^*(P_t^*, R_{t-1}^*) \leq 0, \quad \text{if } x_t^* = 0,$$
$$-P_t^* + \bar{v}_t^*(P_t^*, R_{t-1}^* + M_t + 1) \geq 0, \quad \text{if } x_t^* = M_t.$$

Because states $(P_t^*, R_{t-1}^* + x_t^*)$ and $(P_t^*, R_{t-1}^* + x_t^* + 1)$ are elements of $\bar{\mathcal{S}}_t^*$, it follows from Theorem 6.1 that

$$\bar{v}_t^*(P_t^*, R_{t-1}^* + x_t^*) = v_t^*(P_t^*, R_{t-1}^* + x_t^*) \qquad \text{and} \qquad \bar{v}_t^*(P_t^*, R_{t-1}^* + x_t^* + 1) = v_t^*(P_t^*, R_{t-1}^* + x_t^* + 1).$$

This fact, combined with the given characterization of the accumulation points $(\bar{v}^*, R_{t-1}^*, P_t^*, x_t^*)$, is sufficient to conclude the proof. □

7. Experimental results. The purpose of this section is to analyse and compare empirically the rate of convergence of our approach with the rate of convergence of other convergent Monte Carlo-based algorithms. We would like to emphasize that the main contribution of the paper is the convergence analysis of the ADP-lagged algorithm. Therefore, our intention is not to perform a comprehensive experimental study but rather to provide an illustration of the various aspects that can influence the rate of convergence.

We start by giving a brief description of each approach to which we compare our algorithm. In a batch mode Monte Carlo-based value iteration algorithm (batch), at each iteration n once a sample for the price process, reward, and demand is gathered, we sample slopes at all possible asset levels R and use this information to update the corresponding slopes for the observed sampled prices $P^n = (P_0^n, \ldots, P_{T-1}^n)$. That is, Steps 2(c) and 2(d) of the algorithm described in Figure 1 are replaced by

Step 2(c). Observe $\hat{v}_{t+1}^n(R)$ according to Equation (4) for all R such that $(P_t^n, R) \in \bar{\mathcal{S}}_t$.

Step 2(d). For $(P, R) \in \bar{\mathcal{S}}_t$,

$$z_t^n(P, R) = [(1 - \alpha_t^n)\bar{v}_t^{n-1}(P, R) + \alpha_t^n \hat{v}_{t+1}^n(R)]1_{\{P=P_t^n\}} + \bar{v}_t^{n-1}(P, R)1_{\{P\neq P_t^n\}}.$$

Applying this method, which is synchronous in the sense that all the slopes for the observed prices are updated at once, we want to measure the tradeoff between a synchronous (batch) and an asynchronous approach (our algorithm). Our method is asynchronous in the sense that only two slopes are updated at each iteration n and time t (it can be more if a violation of concavity occurs).

Using a real time dynamic programming (RTDP) approach (Barto et al. [3]), expected values are computed instead of using sample observations. That is, Step 2(c) of the algorithm described in Figure 1 is replaced by:

Step 2(c). Observe $\hat{v}_{t+1}^n(R_t^n)$ and $\hat{v}_{t+1}^n(R_t^n + 1)$ according to:

$$\hat{v}_{t+1}^n(R) = \mathbb{E}\big[\max(\min(P_{t+1}, \bar{v}_{t+1}^{n-1}(P_{t+1}, R)), \bar{v}_{t+1}^{n-1}(P_{t+1}, R + M_{t+1}))1_{\{t<T-1\}} + \hat{r}^n 1_{\{R\leq \hat{D}^n\}}1_{\{t=T-1\}} \mid P_t^n = P\big]. \quad (31)$$

When we compare the computational results of this method to the computational results of our approach, we are observing the tradeoff between more information given by the expectation versus the time spent to do this operation.

A very popular approach in the approximate dynamic literature is Q-learning (Watkins and Dayan [29], Abounadi et al. [1], Rummery and Niranjan [17], Even-Dar and Mansour [9], Cybenko et al. [7], Tsitsiklis [26], Duff [8]), which, like our algorithm is also often used as a model-free algorithmic strategy. However, its state space, namely, $\mathcal{S} \times \mathcal{X}$, where $\mathcal{X}$ is our action space, makes this approach impractical for our problem class. Therefore, instead of implementing a Q-learning approach, we consider an algorithm that only stores the state after the decision is made and samples all possible actions using a decaying exploration scheme called ϵ-greedy. This decaying approach is fully described in Singh et al. [21]. The authors prove its convergence to an optimal

TABLE 1. Instances description – $T = 10$ and $M_t = M = 400$ – discretization = 0.1.

Instances	State space	Initial price	Reward $\hat{r}$	Demand $\widehat{D}$	Price proc.
1	580,000	Constant 20	U(50, 60)	DiscU(180, 250)	RW
2	580,000	Constant 20	U(50, 60)	Poisson(200)	RW
3	4,114,800	$1.7 * \mathrm{U}(1, 12)$	$P_{T-1} * \mathrm{U}(1.03, 1.15)$	Poisson(250)	MR
4	4,114,800	$1.7 * \mathrm{U}(1, 12)$	$P_{T-1} * \mathrm{U}(1.03, 1.15)$	DiscU(180, 220)	MR
5	1,608,000	Constant 40	Constant 25	Poisson(300)	GBM
6	1,608,000	Constant 45	Constant 15	DiscU(225, 375)	GBM

policy and show that each decision is executed infinitely often in every state that is visited infinitely often. Furthermore, as the number of iterations goes to infinity, the decisions are optimal with respect to the current approximation. We want to see how our pure exploitation approach compares to a decaying exploration one.

In our implementation, we define $\epsilon(P, R)$ to be equal to $a/N(P, R)$, where $a \in (0, 1)$ and $N(P, R)$ is the number of visits to state (P, R). We replace Step 2(a) of the algorithm described in Figure 1 by:

Step 2(a). Sample a random variable $\hat{u}$ from a continuous uniform $(0, 1)$ distribution.

$$x_t^n = \begin{cases} \text{Follows a discrete uniform distribution } (0, M_t), & \text{if } \hat{u} < \epsilon(P_t^n, R_{t-1}^n) \\ \underset{0 \le x \le M_t}{\arg\max} -P_t^n x + \overline{V}_t^{n-1}(P_t^n, R_{t-1}^n + x), & \text{otherwise.} \end{cases}$$

We have experimented with values of a ranging from 0.1 to 0.9 with 0.1 increments. Although our convergence proof applies to the case $a = 0$ (no exploration), we found that we obtained the best results with $a = 0.5$ and, therefore, we used this value for all the experimental work in this section. We also observe that epsilon greedy is a fairly simple form of exploration because it ignores the value of visiting a state (see Powell [15, Chapter 10], for a fairly in-depth discussion of these issues). Because the central contribution of this paper is the convergence proof, the results in this section are primarily illustrative and are not intended to represent the best possible implementation of ADP for this problem class.

We note that the projection operation is in effect for all the approaches. Enforcing concavity is not a very time-consuming operation and it both accelerates convergence and guarantees that the optimal solution x_t^n at each iteration is unique and easy to compute using the characterization given by Equation (6).

The instances considered in the experiments are described in Table 1. Problems were randomly generated using different distributions for the rewards $\hat{r}$ and initial prices P_0. Moreover, both discrete uniform (DiscU) and Poisson demand distributions with different parameters were used.

We also created different price processes, namely, random walk (RW), mean reversion (MR), and geometric Brownian motion (GBM), all of which are described below. These processes are unbounded and continuous yet we require, in order to prove convergence of the algorithm, our state space to be bounded and finite. Therefore, the processes are truncated from above and below, and they are discretized. However, except for the RTDP method, all the Monte Carlo-based algorithms (including ours) may use continuous prices, with the discretization occurring only in the value function approximation. Because the RTDP method requires the computation of expected values to determine the slopes, the distribution of the price process has to be adjusted to reflect the truncation levels and the discretization increments under consideration, transforming the original Markovian continuous processes into Markovian discrete processes.

For all instances, the number of time periods considered is 10. That is, the random demand $\widehat{D}$ and the random reward $\hat{r}$ are observed at $T = 10$. Furthermore, the upper bound on the decision quantity x_t, for $t = 0, \ldots, T-1$, is set to $M_t = M = 400$. Table 1 also conveys the size of the state space of each instance when the price process, for all instances, is discretized using a 0.1 increment.

Next, we give the details of the different price processes. The random walk price process is given by $P_t = P_{t-1} + \widehat{P}_t$, where the price increment $\widehat{P}_t$ has a normal distribution with mean $\mu = 0.02$ and standard deviation $\sigma = 1.5$. The mean reversion price process is given by $P_t = P_{t-1} + \widehat{P}_t + 0.5(B_t - P_{t-1})$, where $\widehat{P}_t$ is uniformly distributed between 0.9 and 1.2, and $B_0 = 1.7\bar{U}(1, 12)$ and $B_t = B_{t-1}\bar{U}(0.9, 1.2)$, where $\bar{U}$ is the mean of the corresponding uniform distribution. Finally, the geometric Brownian motion process is given by $P_t = P_{t-1}e^{\widehat{P}_t}$, where $\widehat{P}_t$ is normally distributed with mean $\mu = 0.0125$ and standard deviation $\sigma = 0.087$.

It is easy to see that when the random walk and the geometric Brownian motion are considered, the slopes $v_t^*(P, R)$ given by Equation (3) are monotone increasing in the price dimension. Therefore, for all the different methods and instances 1, 2, 5, and 6, this property is going to be imposed in order to speed up the rate of

convergence. The monotone increasing property is obtained by induction on t. We prove for the random walk process (instances 1 and 2). A similar reasoning can be applied to instances 5 and 6. The base case $t = T-1$ is trivial as, for $(P, R) \in \bar{\mathcal{S}}_{T-1}$, $v^*_{T-1}(P, R) = \mathbb{E}[\hat{r}1_{\{\hat{D}\geq R\}}]$, i.e., the optimal slopes at $T-1$ are independent of the price. The induction hypothesis assumes, given an asset level R, that the optimal slopes at $t+1$ are monotone increasing in the price dimension. Pick $(\bar{P}, R) \in \bar{\mathcal{S}}_t$ and $(\tilde{P}, R) \in \bar{\mathcal{S}}_t$ such that $\bar{P} \geq \tilde{P}$. Also fix $\omega \in \Omega$. Clearly, $\bar{P} + \hat{P}_{t+1}(\omega) \geq \tilde{P} + \hat{P}_{t+1}(\omega)$ and from the induction hypothesis, $v^*_{t+1}(\bar{P} + \hat{P}_{t+1}(\omega), R) \geq v^*_{t+1}(\tilde{P} + \hat{P}_{t+1}(\omega), R)$. Hence,

$$\min(\bar{P} + \hat{P}_{t+1}(\omega), v^*_{t+1}(\bar{P} + \hat{P}_{t+1}(\omega), R)) \geq \min(\tilde{P} + \hat{P}_{t+1}(\omega), v^*_{t+1}(\tilde{P} + \hat{P}_{t+1}(\omega), R)).$$

It also holds that $v^*_{t+1}(\bar{P} + \hat{P}_{t+1}(\omega), R + M_{t+1}) \geq v^*_{t+1}(\tilde{P} + \hat{P}_{t+1}(\omega), R + M_{t+1})$. Therefore,

$$\begin{aligned}&\max(\min(\bar{P} + \hat{P}_{t+1}(\omega), v^*_{t+1}(\bar{P} + \hat{P}_{t+1}(\omega), R)), v^*_{t+1}(\bar{P} + \hat{P}_{t+1}(\omega), R + M_{t+1}))\\ &\quad \geq \max(\min(\tilde{P} + \hat{P}_{t+1}(\omega), v^*_{t+1}(\tilde{P} + \hat{P}_{t+1}(\omega), R)), v^*_{t+1}(\tilde{P} + \hat{P}_{t+1}(\omega), R + M_{t+1})),\end{aligned}$$

proving, by the definition of the optimal slopes, that $v^*_t(\bar{P}, R) \geq v^*_t(\tilde{P}, R)$.

The experiments were run as follows. Using the underlying distributions (given in Table 1), we computed an optimal policy using backward dynamic programming, assuming the prices were discretized to the nearest 0.01. We next randomly generated 50 sets $\tilde{\Omega}^i$, $i = 1, \ldots, 50$, where each set $\tilde{\Omega}^i$ consisted of 2×10^6 sample paths. These sets were used to train and update the value functions of the approximation algorithms.

To evaluate the policies, for each instance we randomly generated a set $\widehat{\Omega}$ of 800 sample paths. For $\omega \in \widehat{\Omega}$, the profit obtained following the optimal policy is given by

$$F^*(\omega) = \sum_{t=0}^{T-1} -P_t(\omega)X^*_t(\omega) + \hat{r}(\omega)\min(\widehat{D}_T(\omega), R_{T-1}(\omega)),$$

where $X^*_t(\omega)$ is the decision determined by the optimal policy (computed exactly using backward dynamic programming) at time t for sample path ω. Hence, the sample mean profit produced by the optimal policy is given by:

$$\bar{F}^* = \frac{1}{800}\sum_{\omega\in\widehat{\Omega}} F^*(\omega).$$

Similarly, for $\omega \in \widehat{\Omega}$, the profit obtained following an approximate policy is given by

$$\widehat{F}^{n,i}(\omega) = \sum_{t=0}^{T-1} -P_t(\omega)\widehat{X}^{n,i}_t(\omega) + \hat{r}(\omega)\min(\widehat{D}(\omega), R_{T-1}(\omega)),$$

where $\widehat{X}^{n,i}_t(\omega)$ is the decision determined by the approximate policy after n iterations of the corresponding approximation algorithm using training set $\tilde{\Omega}^i$. Next, these values are averaged to obtain

$$F^n(\omega) = \frac{1}{50}\sum_{i=1}^{50} \widehat{F}^{n,i}(\omega).$$

Thus, the sample mean profit produced by the approximation algorithm is given by

$$\bar{F}^n = \frac{1}{800}\sum_{\omega\in\widehat{\Omega}} F^n(\omega).$$

Finally, we determine the percentage distance from optimal (in other words, the error incurred by the approximation approach) according to

$$\eta^n = \frac{|\bar{F}^* - \bar{F}^n|}{\bar{F}^*} \times 100. \tag{32}$$

Figure 6 illustrates the rate of convergence of the different approximation methods considered in the paper as a function of the number of iterations (Figure 6(a)) and CPU time (Figure 6(b)). The latter only conveys the results up until the time taken by our algorithm (ADP lagged) to reach 10^5 iterations. Even though Figure 6 refers to instance 1, it reflects the behavior of the methods for the other instances as well.

It is more intuitive to think that a decaying exploration strategy can achieve better rates of convergence than a pure exploitation one. However, Figure 6 shows us that this is not the case for the lagged asset acquisition

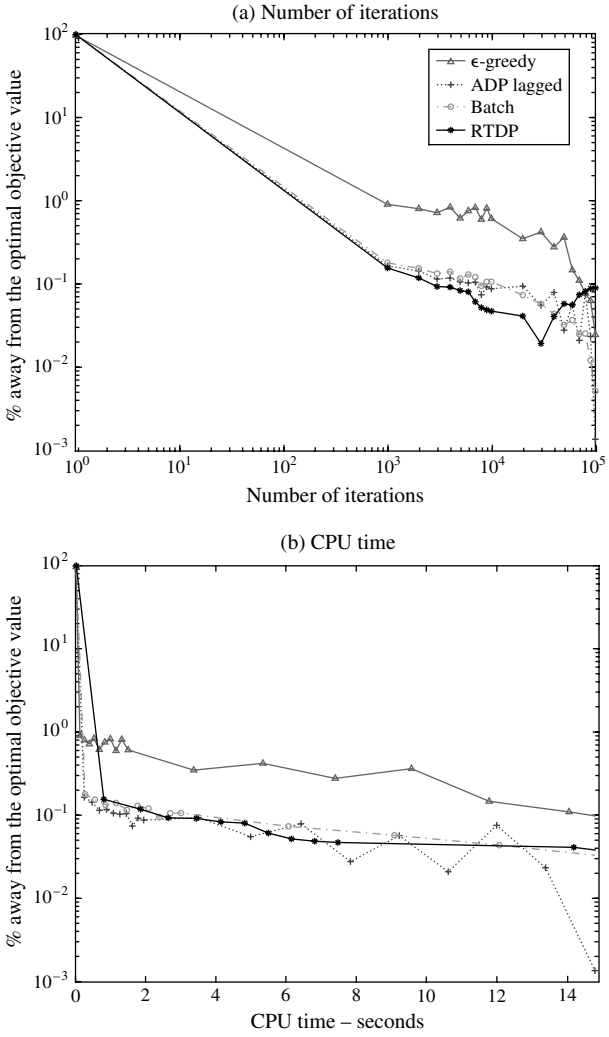


FIGURE 6. Instance 1—% away from optimal objective value.

problem. This might be explained by the fact that in our problem the decision might take 400 different values and the exploration steps, especially in the initial iterations, may lead the algorithm to parts of the state space that add little value to the problem of finding an optimal policy. Thus, instead of accelerating convergence, the exploration is in fact slowing it down. Moreover, even though the ϵ-greedy strategy mimics greedy choices in the limit, with problem instances that have a state space size ranging in the order of 10^5 to 10^7, the algorithm is far from reaching limiting behavior after 10^5 (as in Figure 6) or even 2×10^6 iterations (as in Table 2).

Table 2 shows the time (in seconds) to compute the optimal policy and the time it took each method to be $10\%, 1\%, \ldots, 10^{-3}\%$ away from the optimal policy. All methods were limited to 2 million iterations.

Note that instances 3 and 4 did not reach the $10^{-2}\%$ level. This is due to the fact that these instances use the mean reversion price process, and the monotone increasing property of the slopes in the price dimension does not apply to this process. Hence, this property could not be imposed in order to speed up convergence.

Table 2 also conveys that the computational time for the batch approach is much higher than the computational time of the ADP-lagged approach. It follows that even though the batch method makes better use of the information in each sample realization, the overhead of computing all the slopes at each iteration offsets any benefits. The same is true for the RTDP approach. More information given by the expectation instead of a sample realization does not result in an improvement in the solutions in a competitive amount of time. As pointed out by Figure 6, the common exploitation versus exploration tradeoff has a simple answer for the lagged asset acquisition problem, as the pure exploitation approaches (ours and RTDP) clearly outperformed the ϵ-greedy method, a decaying exploration approach.

We finish this section pointing out that when we compare the computational time spent to obtain the exact solution and the computational time spent by each approximate method, we can infer that despite the fact that we

TABLE 2. Time in seconds to reach the given percentage of the exact solution. All methods were limited to 2×10^6 iterations. Quality of solutions is determined by η^n defined in Equation (32).

		Percent away from optimal objective function				
Instance	Method	10^1	10^0	10^{-1}	10^{-2}	10^{-3}
1	ADP lagged	0.25	0.25	1.62	14.83	
Exact time	Batch	0.27	0.27	2.39	30.02	
6,923.03 secs	RTDP	0.81	0.81	2.65		
	ϵ-greedy	0.13	0.13	16.05		
2	ADP lagged	0.26	0.26	0.26	6.66	
Exact time	Batch	0.21	0.21	0.21		
6,842.21 secs	RTDP	0.81	0.81	0.81	20.36	
	ϵ-greedy	0.14	0.26	16.61		
3	ADP lagged	3.69	29.72	460.79		
Exact time	Batch	5,921.77				
7,658.94 secs	RTDP	21.88	30.7			
	ϵ-greedy	4.29				
4	ADP lagged	3.46	20.62	691.25		
Exact time	Batch	11,816.34				
7,548.53 secs	RTDP	20.96	29.38			
	ϵ-greedy	4.27				
5	ADP lagged	10.53	17.12	27.77	46.83	216.48
Exact time	Batch	129	194.13			
24,149.52 secs	RTDP	406.34	677.7	812.85	812.85	948.04
	ϵ-greedy	4.19	42.45			
6	ADP lagged	0.34	4.9	9.52	206.47	236.63
Exact time	Batch	12.46	31.93			
10,766.61 secs	RTDP	9.18	376.77	376.77	564.38	
	ϵ-greedy	0.26	6.22			

are dealing with a scalar decision, instances of this problem class can easily run into very large state spaces. We can also infer that our ADP-lagged algorithm would be a reasonable method of choice even when the distribution of the random variables are known and we are able to use standard dynamic programming techniques, as the ADP-lagged approach gives very accurate policies quite quickly.

8. Conclusions. We proposed an approximate dynamic programming algorithm to solve the lagged asset acquisition problem. Our algorithm is a sample-based method that uses a pure exploitation scheme at every iteration. The idea is to construct piecewise linear value function approximations, learning their slopes only for certain portions of the state space, which is determined by the algorithm itself. Because the optimal value functions associated with the problem are concave, the algorithm enforces the concavity of the approximations at every iteration through a projection operation.

We prove that the algorithm converges to an optimal policy almost surely. The proof builds on the ideas in Bertsekas and Tsitsiklis [4] as well as in Powell et al. [16], although the former proves convergence only if all states are visited infinitely often and the latter deals only with two-stage problems. The RTDP method is another example of a pure exploitation algorithm that converges to an optimal policy, although it relies on the fact that expected values are computed at each iteration (instead of using just sample realizations) and the initial approximations are optimistic.

The computational experiments illustrate some issues about our problem class and the rate of convergence of the algorithms. First, they show that even though our problem has only two dimensions, obtaining exact solutions (even when the distributions are known) is computationally very expensive. Second, they show that the ADP-lagged algorithm provides very high-quality solutions in a reasonable amount of time, demonstrating that our algorithm can handle continuous prices because only the value functions have to be discretized, not the price process itself. Finally, for the instances considered, we can infer that a pure exploitation scheme accelerates the rate of convergence because both our algorithm and the RTDP approach outperformed the ϵ-greedy approach. We can also infer that using more information either through the computation of expected values (RTDP) or the observation of more sample slopes at a time (batch) does not justify the increased computational time per iteration, as our algorithm still provided better convergence rates. The results also support the idea that the more

structure the faster the convergence, as the rate of convergence for the instances where the monotone increasing property of the slopes in the price dimension did not apply (instances 3 and 4) was slower compared to the other instances.

Appendix A. Summary of elements. Below is a brief summary of notation to assist with reading the proofs. For each random element, we provide its measurability.

- Filtrations
  $\mathcal{F} = \sigma\{(P_t^n, x_t^n, \widehat{D}^n, \hat{r}^n),\ n \geq 1,\ t = 0, \ldots, T-1\}$.
  $\mathcal{F}_T^n = \sigma\{(P_{t'}^m, x_{t'}^m, \widehat{D}^m, \hat{r}^m),\ m \leq n,\ t' = 0, \ldots, T-1\}$.
  $\mathcal{F}_t^n = \sigma(\{(P_{t'}^m, x_{t'}^m, \widehat{D}^m, \hat{r}^m),\ m < n,\ t' = 0, \ldots, T-1\} \cup \{(P_{t'}^n, x_{t'}^n),\ t' = 0, \ldots, t\})$.
  $\mathcal{F}_t^n \subset \cdots \subset \mathcal{F}_T^n \subset \mathcal{F}_0^{n+1} \subset \cdots \subset \mathcal{F}_t^{n+1} \subset \cdots \subset \mathcal{F}$.
- Exogenous information
  $\{P_t^n \in \mathcal{F}_t^n\}_{t=0}^{T-1}$: Markovian price process. Strictly positive, has finite support, and is independent of the asset level.
  $\widehat{D}^n \in \mathcal{F}_T^n$: demand. Positive, discrete, might be dependent on P_{T-1}^n and $\hat{r}^n$.
  $\hat{r}^n \in \mathcal{F}_T^n$: reward. Strictly positive, bounded, might be dependent on P_{T-1}^n and $\widehat{D}^n$.
- Decision and state variables
  $x_t^n \in \mathcal{F}_t^n$: order quantity. Integer valued, $0 \leq x_t^n \leq M_t$ (deterministic bound).
  $(P_t^n, R_t^n) \in \mathcal{F}_t^n$: price and asset quantity (integer, $0 \leq R_t^n \leq B_t = \sum_{i=0}^t M_t$).
  $\mathcal{P}_t$: finite support set of P_t.
- Value function (concave piecewise linear with integer break points)
  $V_t^*(P, R)$: optimal value function at (P, R).
  $\overline{V}_t^n(P, R) = \sum_{i=1}^R \bar{v}_t^n(P, i) \in \mathcal{F}_{t+1}^n$: value function approximation at (P, R). We assume that $\overline{V}_t^n(P, 0) = 0$.
- Slopes (monotone decreasing in R and bounded)
  $v_t^*(P, R)$: slope of the optimal value function at (P, R).
  $z_t^n(P, R) \in \mathcal{F}_{t+1}^n$: unprojected slope of the value function approximation at (P, R).
  $\bar{v}_t^n(P, R) \in \mathcal{F}_{t+1}^n$: slope of the value function approximation at (P, R).
  $\bar{v}_t^*(P, R) \in \mathcal{F}$: accumulation point of $\{\bar{v}_t^n(P, R)\}_{n \geq 0}$.
  $\hat{v}_t^n(R) \in \mathcal{F}_t^n$: sample slope at R.
- Stepsizes (bounded by zero and one, sum is $+\infty$, sum of the squares is $< +\infty$)
  $\alpha_t^n \in \mathcal{F}_t^n$ and $\bar{\alpha}_t^n(P, R) = \alpha_t^n(1_{\{P = P_t^n,\ R = R_t^n\}} + 1_{\{P = P_t^n,\ R = R_t^n + 1\}})$.
- Set of iterations (due to the projection operation)
  $\mathcal{N}_t^-(P, R) \in \mathcal{F}$: iterations in which the unprojected slope at (P, R) was increased.
  $\mathcal{N}_t^+(P, R) \in \mathcal{F}$: iterations in which the unprojected slope at (P, R) was decreased.
- Set of states
  $\mathcal{S}_t$: state space.
  $\bar{\mathcal{S}}_t$: $\mathcal{S}_t$ minus the $(P_t, 0)$ pairs.
  $\mathcal{S}_t^* \in \mathcal{F}$: accumulation points (P_t^*, R_t^*) or $(P_t^*, R_t^* + 1)$ of $\{(P_t^n, R_t^n)\}_{n \geq 0}$.
  $\widetilde{\mathcal{S}}_t^- \in \mathcal{F}$: states in which the projection had not increased the unprojected slopes infinitely often (i.o.).
  $\widetilde{\mathcal{S}}_t^+ \in \mathcal{F}$: states in which the projection had not decreased the unprojected slopes i.o.
- Dynamic programming operator H
- Deterministic bounding sequences $\{L_t^k(P, R)\}_{k \geq 0}$ and $\{U_t^k(P, R)\}_{k \geq 0}$
- Error variables $\hat{s}_{t+1-}^n \in \mathcal{F}_{t+1}^n$ and $\hat{s}_{t+1+}^n \in \mathcal{F}_{t+1}^n$
- Stochastic noise sequences $\{\bar{s}_{t-}^n(P, R)\}_{n \geq 0}$ and $\{\bar{s}_{t+}^n(P, R)\}_{n \geq 0}$
- Stochastic bounding sequences $\{\bar{l}_t^n(P, R)\}_{n \geq 0}$ and $\{\bar{u}_t^n(P, R)\}_{n \geq 0}$

Appendix B. Proofs. We start with the proof of Proposition 2.1, then we present the proofs for the lemmas of the convergence analysis section.

Proof of Proposition 2.1. An induction argument is used. First, we show for time $T-1$ and $(P, R) \in \bar{\mathcal{S}}_{T-1}$ that $v_{T-1}^*(P, R)$ is indeed equal to Equation (3). We also show, for any $y \in (0, 1)$, that $V_{T-1}^*(P, R + y) = V_{T-1}^*(P, R) + y v_{T-1}^*(P, R+1)$. Finally, we argue that $v_{T-1}^*(P, R) \leq v_{T-1}^*(P, R-1)$. Fix $(P, R) \in \bar{\mathcal{S}}_{T-1}$. We have that

$$v_{T-1}^*(P, R) = \mathbb{E}\big[\hat{r} \min(\widehat{D}, R) - \hat{r} \min(\widehat{D}, R-1) \mid P_{T-1} = P\big] = \mathbb{E}\big[\hat{r} 1_{\{\widehat{D} \geq R\}} \mid P_{T-1} = P\big].$$

It is obvious that $\mathbb{E}[\hat{r}1_{\{\widehat{D}\geq R\}} \mid P_{T-1}=P] \leq \mathbb{E}[\hat{r}1_{\{\widehat{D}\geq R-1\}} \mid P_{T-1}=P]$, thus $v^*_{T-1}(P,R) \leq v^*_{T-1}(P,R-1)$. Moreover, for $y \in (0,1)$,

$$\begin{aligned}
V^*_{T-1}(P,R+y) &= \mathbb{E}\big[\hat{r}\min(\widehat{D},R+y) \mid P_{T-1}=P\big] \\
&= \mathbb{E}\big[\hat{r}\widehat{D}1_{\{\widehat{D}\leq R+y\}} + (R+y)\hat{r}1_{\{\widehat{D}>R+y\}} \mid P_{T-1}=P\big] \\
&= \mathbb{E}\big[\hat{r}\widehat{D}1_{\{\widehat{D}\leq R\}} + (R+y)\hat{r}1_{\{\widehat{D}\geq R+1\}} \mid P_{T-1}=P\big] \\
&= \mathbb{E}\big[\hat{r}\min(\widehat{D},R) + y\hat{r}1_{\{\widehat{D}\geq R+1\}} \mid P_{T-1}=P\big] \\
&= V^*_{T-1}(P,R) + yv^*_{T-1}(P,R+1),
\end{aligned}$$

where the transition from the second to the third line follows from the fact that $\widehat{D}$ and R are integer valued.

Now assume for $t+1$ that the optimal value functions are piecewise linear, with integer break points and concave in the asset dimension. Moreover, $v^*_{t+1}(P,R)$ is given by Equation (3). We shall prove that the same claims are true for t, where t is any time period between 0 and $T-2$. Fix $(P,R) \in \bar{\mathcal{S}}_t$. We have that $v^*_t(P,R) = V^*_t(P,R) - V^*_t(P,R-1)$, where

$$V^*_t(P,R) = \mathbb{E}\big[-P_{t+1}x^*_{t+1} + V^*_{t+1}(P_{t+1},R+x^*_{t+1}) \mid P_t=P\big], \tag{B.1}$$

$$V^*_t(P,R-1) = \mathbb{E}\big[-P_{t+1}y^*_{t+1} + V^*_{t+1}(P_{t+1},R-1+y^*_{t+1}) \mid P_t=P\big], \tag{B.2}$$

and x^*_{t+1} and y^*_{t+1} are defined by

$$\begin{aligned}
x^*_{t+1} &= \arg\max_{0\leq x\leq M_{t+1}} -P_{t+1}x + V^*_{t+1}(P_{t+1},R+x), \\
y^*_{t+1} &= \arg\max_{0\leq x\leq M_{t+1}} -P_{t+1}x + V^*_{t+1}(P_{t+1},R-1+x).
\end{aligned}$$

We have that the optimal decision x^*_{t+1} is the maximum element of the set

$$\{0\leq x\leq M_{t+1}\colon\ -P_{t+1}+v^*_{t+1}(P_{t+1},R+x)\geq 0,\ -P_{t+1}+v^*_{t+1}(P_{t+1},R+x+1)\leq 0\}.$$

If the set is empty, then $x^*_{t+1}=0$, if $-P_{t+1}+v^*_{t+1}(P_{t+1},R+1)\leq 0$. Moreover, $x^*_{t+1}=M_{t+1}$, if $-P_{t+1}+v^*_{t+1}(P_{t+1},R+M_{t+1}+1)\geq 0$. The same applies for the optimal decision y^*_{t+1}.

Define the random variable $\hat{v}_{t+1}(R) = \max(\min(P_{t+1}, v^*_{t+1}(P_{t+1},R)), v^*_{t+1}(P_{t+1},R+M_{t+1}))$. We want to show that $v^*_t(P,R) = \mathbb{E}[\hat{v}_{t+1}(R) \mid P_t=P]$. Pick $P_{t+1}\in\mathcal{P}_{t+1}$ such that $\mathbb{P}\{P_{t+1} \mid P_t=P\}>0$. We have to consider three cases.

Case 1. $x^*_{t+1}=y^*_{t+1}=0$. In this case, due to definition of x^*_{t+1} and the induction hypothesis, $P_{t+1}\geq v^*_{t+1}(P_{t+1},R)\geq v^*_{t+1}(P_{t+1},R+M_{t+1})$. Therefore, $\hat{v}_{t+1}(R)=v^*_{t+1}(P_{t+1},R)$.

Case 2. $y^*_{t+1}=x^*_{t+1}+1<M_{t+1}$. In this case, $\hat{v}_{t+1}(R)=P_{t+1}$, because:

$$P_{t+1}\leq v^*_{t+1}(P_{t+1},R) \qquad \text{and} \qquad P_{t+1}\geq v^*_{t+1}(P_{t+1},R+x^*_{t+1}+1)\geq v^*_{t+1}(P_{t+1},R+M_{t+1}+1).$$

Case 3. $x^*_{t+1}=y^*_{t+1}=M_{t+1}$. In this case, $P_{t+1}\leq v^*_{t+1}(P_{t+1},R)$ and $P_{t+1}\leq v^*_{t+1}(P_{t+1},R+M_{t+1})$. Hence, $\hat{v}_{t+1}(R)=v^*_{t+1}(P_{t+1},R+M_{t+1})$. For each case, if we substitute x^*_{t+1} and y^*_{t+1} into $-P_{t+1}(x^*_{t+1}-y^*_{t+1})+V^*_{t+1}(P_{t+1},R+x^*_{t+1})-V^*_{t+1}(P_{t+1},R-1+y^*_{t+1})$, we get that this expression is equal to $\hat{v}_{t+1}(R)$. Thus, we have proved that $v^*_t(P,R)$ is equal to Equation (3). Clearly, as the induction hypothesis tells us that $v^*_{t+1}(P_{t+1},R)\leq v^*_{t+1}(P_{t+1},R-1)$, it follows that $v^*_t(P,R)\leq v^*_t(P,R-1)$.

We finish the proof showing for $y\in(0,1)$ that $V^*_t(P,R+y)=V^*_t(P,R)+yv^*_t(P,R+1)$. Again, we have to consider three cases:

Case 1. $x^*_{t+1}=0$. In this case, $P_{t+1}\geq v^*_{t+1}(P_{t+1},R+1)\geq v^*_{t+1}(P_{t+1},R+1+M_{t+1})$. Thus, because of the induction hypothesis and definition of $\hat{v}_{t+1}(R+1)$,

$$\begin{aligned}
\max_{0\leq x\leq M_{t+1}} -P_{t+1}x + V^*_{t+1}(P_{t+1},R+y+x) &= V^*_{t+1}(P_{t+1},R+y) = V^*_{t+1}(P_{t+1},R) + yv^*_{t+1}(P_{t+1},R+1) \\
&= -P_{t+1}x^*_{t+1} + V^*_{t+1}(P_{t+1},R+x^*_{t+1}) + y\hat{v}_{t+1}(R+1).
\end{aligned}$$

Case 2. $x^*_{t+1}\in\{1,\ldots,M_{t+1}-1\}$. In this case, $\hat{v}_{t+1}(R+1)=P_{t+1}$ because $P_{t+1}\leq v^*_{t+1}(P_{t+1},R+1)$ and $P_{t+1}\geq v^*_{t+1}(P_{t+1},R+M_{t+1}+1)$. Hence,

$$\begin{aligned}
\max_{0\leq x\leq M_{t+1}} -P_{t+1}x + V^*_{t+1}(P_{t+1},R+y+x) &= -P_{t+1}(x^*_{t+1}-y) + V^*_{t+1}(P_{t+1},R+x^*_{t+1}) \\
&= -P_{t+1}x^*_{t+1} + V^*_{t+1}(P_{t+1},R+x^*_{t+1}) + y\hat{v}_{t+1}(R+1).
\end{aligned}$$

Case 3. $x^*_{t+1} = M_{t+1}$. In this case, $P_{t+1} \le v^*_{t+1}(P_{t+1}, R + M_{t+1} + 1)$. Therefore, $\hat{v}_{t+1}(R+1) = v^*_{t+1}(P_{t+1}, R + M_{t+1} + 1)$ and

$$\begin{aligned}\max_{0\le x\le M_{t+1}} -P_{t+1}M_{t+1} + V^*_{t+1}(P_{t+1}, R + y + M_{t+1}) &= -P_{t+1}(x^*_{t+1} - y) + V^*_{t+1}(P_{t+1}, R + x^*_{t+1})\\ &= -P_{t+1}M_{t+1} + V^*_{t+1}(P_{t+1}, R + M_{t+1}) + yv^*_{t+1}(P_{t+1}, R + M_{t+1} + 1)\\ &= -P_{t+1}x^*_{t+1} + V^*_{t+1}(P_{t+1}, R + x^*_{t+1}) + y\hat{v}_{t+1}(R+1).\end{aligned}$$

Because by definition, $V^*_t(P, R+y) = \mathbb{E}[\max_{0\le x\le M_{t+1}} -P_{t+1}x + V^*_{t+1}(P_{t+1}, R + y + x) \mid P_t = P]$, we use Equation (B.1) and $v^*_t(P, R+1) = \mathbb{E}[\hat{v}_{t+1}(R+1) \mid P_t = P]$ to finish the proof. □

Each lemma assumes all the conditions imposed and all the results obtained before its statement in the proof of Theorem 6.1. To improve the comprehension of each proof, all the assumptions are presented beforehand.

Proof of Lemma 6.2.

Assumptions. Assume stepsize conditions (7)–(9).

Fix $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_t$ and $\omega \in \Omega$. Omitting the dependence on ω, we assume that $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}^*_t$. We prove the convergence to zero of the sequence $\{\bar{s}^n_{t-}(\bar{P}^*, \bar{R}^*)\}_{n\ge0}$. The proof for $\{\bar{s}^n_{t+}(\bar{P}^*, \bar{R}^*)\}_{n\ge0}$ is symmetrical. To simplify notation, let $\bar{s}^n_{t-}(\bar{P}^*, \bar{R}^*)$ be denoted by $\bar{s}^{*,n}_t$ and $\bar{\alpha}^n_t(\bar{P}^*, \bar{R}^*)$ be denoted by $\bar{\alpha}^{*,n}_t$. Furthermore, let

$$\hat{\theta}^n_{t+1} = \hat{s}^n_{t+1-}(R^n_t 1_{\{\bar{R}^* \le R^n_t\}} + (R^n_t + 1)1_{\{\bar{R}^* > R^n_t\}}).$$

We have, for $n \ge 1$,

$$(\bar{s}^{*,n}_t)^2 \le \left[(1 - \bar{\alpha}^{*,n}_t)\bar{s}^{*,n-1}_t + \bar{\alpha}^{*,n}_t\hat{\theta}^n_{t+1}\right]^2 = (\bar{s}^{*,n-1}_t)^2 - 2\bar{\alpha}^{*,n}_t(\bar{s}^{*,n-1}_t)^2 + A^n_t, \tag{B.3}$$

where $A^n_t = 2\bar{\alpha}^{*,n}_t\bar{s}^{*,n-1}_t\hat{\theta}^n_{t+1} + (\bar{\alpha}^{*,n}_t)^2(\hat{\theta}^n_{t+1} - \bar{s}^{*,n-1}_t)^2$. We want to show that:

$$\sum_{n=1}^{\infty} A^n_t = 2\sum_{n=1}^{\infty} \bar{\alpha}^{*,n}_t\bar{s}^{*,n-1}_t\hat{\theta}^n_{t+1} + \sum_{n=1}^{\infty}(\bar{\alpha}^{*,n}_t)^2(\hat{\theta}^n_{t+1} - \bar{s}^{*,n-1}_t)^2 < \infty.$$

It is trivial to see that both $\bar{s}^{*,n-1}_t$ and $\hat{\theta}^n_{t+1}$ are bounded. Thus, $(\hat{\theta}^n_{t+1} - \bar{s}^{*,n-1}_t)^2$ is bounded and Equation (8) tells us that

$$\sum_{n=1}^{\infty}(\bar{\alpha}^{*,n}_t)^2(\hat{\theta}^n_{t+1} - \bar{s}^{*,n-1}_t)^2 < \infty. \tag{B.4}$$

Define a new sequence $\{g^n_{t+1}\}_{n\ge0}$, where $g^0_{t+1} = 0$ and $g^n_{t+1} = \sum_{m=1}^n \bar{\alpha}^{*,m}_t\bar{s}^{*,m-1}_t\hat{\theta}^m_{t+1}$. We can easily check that $\{g^n_{t+1}\}_{n\ge0}$ is a $\mathcal{F}^n_T$-martingale bounded in L^2. Measurability is obvious. The martingale equality follows from repeated conditioning and the unbiasedness property. Finally, the L^2-boundedness and consequentially the integrability can be obtained by noticing that $(g^n_{t+1})^2 = (g^{n-1}_{t+1})^2 + 2g^{n-1}_{t+1}\bar{\alpha}^{*,n}_t\bar{s}^{*,n-1}_t\hat{\theta}^n_{t+1} + (\bar{\alpha}^{*,n}_t)^2(\bar{s}^{*,n-1}_t\hat{\theta}^n_{t+1})^2$. From the martingale equality and boundedness of $\bar{s}^{*,n-1}_t$ and $\hat{\theta}^n_{t+1}$, we get

$$\mathbb{E}[(g^n_{t+1})^2 \mid \mathcal{F}^{n-1}_T] \le (g^{n-1}_{t+1})^2 + C\mathbb{E}[(\bar{\alpha}^{*,n}_t)^2 \mid \mathcal{F}^{n-1}_T],$$

where C is a constant. Hence, taking expectations and repeating the process, we obtain from the stepsize assumption (8) and $\mathbb{E}[(g^0_{t+1})^2] = 0$,

$$\mathbb{E}[(g^n_{t+1})^2] \le \mathbb{E}[(g^{n-1}_{t+1})^2] + C\mathbb{E}[(\bar{\alpha}^{*,n}_t)^2] \le \mathbb{E}[(g^0_{t+1})^2] + C\sum_{m=1}^{n}\mathbb{E}[(\bar{\alpha}^{*,m}_t)^2] < \infty.$$

Therefore, the L^2-bounded martingale convergence theorem (Shiryaev [20, p. 510]) tells us that

$$-\infty < \sum_{n=1}^{\infty} \bar{\alpha}^{*,n}_t\bar{s}^{*,n-1}_t\hat{\theta}^n_{t+1} < \infty. \tag{B.5}$$

Inequalities (B.4) and (B.5) show us that $-\infty < \sum_{n=1}^{\infty} A^n_t < \infty$, and so it is valid to write

$$A^n_t = \sum_{m=n}^{\infty} A^m_t - \sum_{m=n+1}^{\infty} A^m_t.$$

Therefore, as $-2\bar{\alpha}_t^{*,n}(\bar{s}_t^{*,n-1})^2 < 0$, inequality (B.3) can be rewritten as

$$(\bar{s}_t^{*,n})^2 + \sum_{m=n+1}^{\infty} A_t^m \le (\bar{s}_t^{*,n-1})^2 + \sum_{m=n}^{\infty} A_t^m. \tag{B.6}$$

Thus, the sequence $\{(\bar{s}_t^{*,n-1})^2 + \sum_{m=n}^{\infty} A_t^m\}_{n\ge 1}$ is nonincreasing and bounded from below as $|\sum_{m=1}^{\infty} A_t^m| < \infty$. Hence, it is convergent. Moreover, as $\sum_{m=n}^{\infty} A_t^m \to 0$ when $n \to \infty$, we can conclude that $\{\bar{s}_t^{*,n}\}_{n\ge 0}$ converges.

Finally, as inequality (B.3) holds for all $n \ge 1$, it yields

$$\begin{aligned}(\bar{s}_t^{*,n})^2 &\le (\bar{s}_t^{*,n-1})^2 - 2\bar{\alpha}_t^{*,n}(\bar{s}_t^{*,n-1})^2 + A_t^n \\ &\le (\bar{s}_t^{*,n-2})^2 - 2\bar{\alpha}_t^{*,n-1}(\bar{s}_t^{*,n-2})^2 + A_t^{n-1} - 2\bar{\alpha}_t^{*,n}(\bar{s}_t^{*,n-1})^2 + A_t^n \\ &\vdots \\ &\le -2\sum_{m=1}^{n} \bar{\alpha}_t^{*,m}(\bar{s}_t^{*,m-1})^2 + \sum_{m=1}^{n} A_t^m.\end{aligned}$$

Passing to the limits, we obtain

$$\limsup_{n\to\infty}(\bar{s}_t^{*,n})^2 + 2\sum_{m=1}^{n} \bar{\alpha}_t^{*,m}(\bar{s}_t^{*,m-1})^2 \le \limsup_{n\to\infty} \sum_{m=1}^{n} A_t^m < \infty.$$

This implies, together with the convergence of $\{\bar{s}_t^{*,n}\}_{n\ge 0}$, that $\sum_{m=1}^{\infty} \bar{\alpha}_t^{*,m}(\bar{s}_t^{*,m-1})^2 < \infty$. On the other hand, stepsize assumption (9) tells us that $\sum_{m=1}^{\infty} \bar{\alpha}_t^{*,m} = \infty$. Hence, there must exist a subsequence of $\{\bar{s}_t^{*,n}\}_{n\ge 0}$ that converges to zero. Therefore, as every subsequence of a convergent sequence converges to its limit, it follows that $\{\bar{s}_t^{*,n}\}_{n\ge 0}$ converges to zero. □

Proof of Lemma 6.3. The proof is divided into two parts.

(i) Part 1.

Proof of inequalities

$$(HL^k)_t(P_t^n, R_t^n) \le (H\bar{v}^{n-1})_t(P_t^n, R_t^n) \le (HU^k)_t(P_t^n, R_t^n), \quad \text{a.s. on } \{R_t^n > 0\} \tag{B.7}$$

$$(HL^k)_t(P_t^n, R_t^n + 1) \le (H\bar{v}^{n-1})_t(P_t^n, R_t^n + 1) \le (HU^k)_t(P_t^n, R_t^n + 1), \quad \text{a.s. on } \{R_t^n < M_t\} \tag{B.8}$$

on the event that $\{n \ge N_t^k\}$.

Assumptions. If $t = T-1$, Equations (B.7) and (B.8) hold trivially with equality without any assumptions, as $(HL^k)_{T-1}(P,R) = (H\bar{v}^{n-1})_{T-1}(P,R) = (HU^k)_{T-1}(P,R) = \mathbb{E}[\hat{r}1_{\{\bar{R}\le\hat{D}\}} \mid P_{T-1} = P]$, for all $(P,R) \in \bar{\mathcal{S}}_{T-1}$. Given $t \in \{0,\dots,T-2\}$ and $k \ge 0$, assume for all states $(\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_{t+1}$ the existence of a random index $N_t^k \ge \bar{N}$ such that

$$L_{t+1}^k(\bar{P}^*, \bar{R}^*) \le \bar{v}_{t+1}^{n-1}(\bar{P}^*, \bar{R}^*) \le U_{t+1}^k(\bar{P}^*, \bar{R}^*) \quad \text{a.s.} \tag{B.9}$$

on $\{n \ge N_t^k,\ (\bar{P}^*, \bar{R}^*) \in \bar{\mathcal{S}}_{t+1}^*\}$. Note that this assumption is possible because the proof of Theorem 1 is done in a backward fashion.

We prove the inequalities in Equation (B.7). The ones in Equation (B.8) are handled in a similar way.

Pick $t \in \{0,\dots,T-2\}$, $k \ge 0$, and $n \ge 0$. Fix $\omega \in \Omega$. Omitting the dependence on ω, assume that $n \ge N_t^k \ge \bar{N}$. Let $(\bar{P}, \bar{R}) = (P_t^n, R_t^n)$ and assume that $\bar{R} > 0$. We pick $P \in \mathcal{P}_{t+1}$ such that $\mathbb{P}\{P_{t+1} = P \mid P_t = \bar{P}\} > 0$. We want to show that

$$\min(P, L_{t+1}^k(P,\bar{R})) \le \min(P, \bar{v}_{t+1}^{n-1}(P,\bar{R})) \le \min(P, U_{t+1}^k(P,\bar{R})).$$

We need to consider two cases.

Case 1. $(P,\bar{R}) \in \bar{\mathcal{S}}_{t+1}^*$. In this case, Equation (B.9) holds for $(P,\bar{R})$. Thus, it is straightforward to see that the inequalities we are trying to prove are true for this case.

Case 2. $(P,\bar{R}) \notin \bar{\mathcal{S}}_{t+1}^*$. The first statement of Lemma 4.1 applies for this case and we have that

$$\sum_{m=\bar{N}}^{\infty} 1_{\{\bar{v}_{t+1}^{m-1}(P,\bar{R}+1) < P,\ P_t^m = \bar{P},\ R_t^m = \bar{R}\}} = 0.$$

As $n \ge N_t^k \ge \bar{N}$, we have that $\bar{v}_{t+1}^{n-1}(P,\bar{R}) \ge \bar{v}_{t+1}^{n-1}(P,\bar{R}+1) \ge P$ and (P, R_{t+1}^n) is an accumulation point of $\{P_{t+1}^m, R_{t+1}^m\}_{m\ge 0}$. Thus, $(P, R_{t+1}^n) \in \bar{\mathcal{S}}_{t+1}^*$ and Equation (B.9) holds for (P, R_{t+1}^n), implying that $P \le U_{t+1}^k(P, R_{t+1}^n) \le U_{t+1}^k(P,\bar{R})$ due to the monotone decreasing property of U_{t+1}^k. We conclude that

$$\min(P, L_{t+1}^k(P,\bar{R})) \le \min(P, \bar{v}_{t+1}^{n-1}(P,\bar{R})) = \min(P, U_{t+1}^k(P,\bar{R})) = P.$$

To simplify the notation, we use $\bar{h}_{t+1}(P, L^k)$, $\bar{h}_{t+1}(P, \bar{v}^{n-1})$ and $\bar{h}_{t+1}(P, U^k)$ to denote $\min(P, L^k_{t+1}(P, \bar{R}))$, $\min(P, \bar{v}^{n-1}_{t+1}(P, \bar{R}))$, and $\min(P, U^k_{t+1}(P, \bar{R}))$, respectively. Note that, for instance, $(HL^k)_t(\bar{P}, \bar{R}) = \mathbb{E}[\max(\bar{h}_{t+1}(P_{t+1}, L^k), L^k_{t+1}(P_{t+1}, \bar{R} + M_{t+1})) \mid P_t = \bar{P}]$. We finalize the proof showing that:

$$\begin{aligned}\max\big(\bar{h}_{t+1}(P, L^k), L^k_{t+1}(P, \bar{R} + M_{t+1})\big) &\le \max\big(\bar{h}_{t+1}(P, \bar{v}^{n-1}), \bar{v}^{n-1}_{t+1}(P, \bar{R} + M_{t+1})\big)\\ &\le \max\big(\bar{h}_{t+1}(P, U^k), U^k_{t+1}(P, \bar{R} + M_{t+1})\big). \qquad \text{(B.10)}\end{aligned}$$

As before, we have to consider two cases. If $(P, \bar{R} + M_{t+1}) \in \bar{\mathcal{S}}^*_{t+1}$, then Equation (B.9) holds for $(P, \bar{R} + M_{t+1})$ and Equation (B.10) is straightforward. On the other hand, if $(P, \bar{R} + M_{t+1}) \notin \bar{\mathcal{S}}^*_{t+1}$, the second part of Lemma 4.1 applies. Hence, $\sum_{m=\bar{N}}^{\infty} 1_{\{\bar{v}^{m-1}_{t+1}(P, \bar{R}^*+1) > P,\, P^m_t = \bar{P},\, R^m_t = \bar{R}\}} = 0$. As $n \ge N^k_t \ge \bar{N}$, we have in particular that $\bar{v}^{n-1}_{t+1}(P, \bar{R}^* + 1) \le P$. Therefore, as $\bar{R} < \bar{R}^* + 1 \le \bar{R} + M_{t+1}$, it follows that $\bar{v}^{n-1}_{t+1}(P, \bar{R} + M_{t+1}) \le \bar{v}^{n-1}_{t+1}(P, \bar{R}^* + 1) \le P$, implying that

$$\max(\bar{h}_{t+1}(P, \bar{v}^{n-1}), \bar{v}^{n-1}_{t+1}(P, \bar{R} + M_{t+1})) = P \le \max(\bar{h}_{t+1}(P, U^k), U^k_{t+1}(P, \bar{R} + M_{t+1}))$$

because we have proved before that $\bar{h}_{t+1}(P, \bar{v}^{n-1}) = \bar{h}_{t+1}(P, U^k) = P$. We also have that Equation (B.9) applies to $(P, \bar{R}^* + 1)$, thus $L^k_{t+1}(P, \bar{R} + M_{t+1}) \le L^k_{t+1}(P, \bar{R}^* + 1) \le \bar{v}^n_{t+1}(P, \bar{R}^* + 1) \le P$, implying that $\max(\bar{h}_{t+1}(P, L^k), L^k_{t+1}(P, \bar{R} + M_{t+1})) \le P$, as $\bar{h}_{t+1}(P, L^k) \le P$ as well. We conclude that Equation (B.10) also holds for this case.

Therefore, we have that

$$\begin{aligned}&\mathbb{E}[\max(\min(P_{t+1}, L^k_{t+1}(P_{t+1}, R^n_t)), L^k_{t+1}(P_{t+1}, R^n_t + M_{t+1})) \mid \mathcal{F}^n_t]\\ &\quad\le \mathbb{E}[\max(\min(P_{t+1}, \bar{v}^{n-1}_{t+1}(P_{t+1}, R^n_t)), \bar{v}^{n-1}_{t+1}(P_{t+1}, R^n_t + M_{t+1})) \mid \mathcal{F}^n_t]\\ &\quad\le \mathbb{E}[\max(\min(P_{t+1}, U^k_{t+1}(P_{t+1}, R^n_t)), U^k_{t+1}(P_{t+1}, R^n_t + M_{t+1})) \mid \mathcal{F}^n_t]\end{aligned}$$

and by definition of the mapping H, we have proved Equation (B.7).

(ii) Part 2.

Proof of inequalities

$$\bar{v}^{n-1}_t(\tilde{P}^*, \tilde{R}^*) \le \bar{u}^{n-1}_t(\tilde{P}^*, \tilde{R}^*) + \bar{s}^{n-1}_{t-}(\tilde{P}^*, \tilde{R}^*) \quad \text{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathcal{S}}^-_t\} \qquad \text{(B.11)}$$

$$\bar{v}^{n-1}_t(\tilde{P}^*, \tilde{R}^*) \ge \bar{l}^{n-1}_t(\tilde{P}^*, \tilde{R}^*) - \bar{s}^{n-1}_{t+}(\tilde{P}^*, \tilde{R}^*) \quad \text{a.s. on } \{(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathcal{S}}^+_t\} \qquad \text{(B.12)}$$

on the event that $\{n \ge N^k_t\}$.

Assumptions. Given $t = 0, \ldots, T-1$, $k \ge 0$ and index N^k_t, assume on $\{n \ge N^k_t\}$ that Equations (22) and (23) hold true. Note that this assumption is feasible because Lemma 6.3 was stated after the induction hypothesis on k in the proof of Part 1 of Theorem 6.1. In fact, the role of this lemma is to assist in proving the existence of N^{k+1}_t such that Equations (22) and (23) are true when $k+1$ is considered.

We prove Equation (B.12). The inequality in Equation (B.11) can be proved using a symmetrical argument. The proof is by induction on n.

Pick $(\tilde{P}^*, \tilde{R}^*) \in \bar{\mathcal{S}}_t$. We fix $\omega \in \Omega$ and omit the dependence on ω. Assume that $(\tilde{P}^*, \tilde{R}^*) \in \widetilde{\mathcal{S}}^+_t$. The proof for the base case $n = N^k_t$ is immediate from the fact that $\bar{s}^{N^k_t - 1}_{t+}(\tilde{P}^*, \tilde{R}^*) = 0$, $\bar{l}^{N^k_t - 1}_t(\tilde{P}^*, \tilde{R}^*) = L^k_t(\tilde{P}^*, \tilde{R}^*)$ and by the assumption that Equation (23) holds true for $n \ge N^k_t$. Now suppose Equation (B.12) is true for a given $n \ge N^k_t$ and we need to prove that $\bar{v}^n_t(\tilde{P}^*, \tilde{R}^*) \ge \bar{l}^n_t(\tilde{P}^*, \tilde{R}^*) - \bar{s}^n_{t+}(\tilde{P}^*, \tilde{R}^*)$.

To simplify the notation, let $\bar{\alpha}^n_t(\tilde{P}^*, \tilde{R}^*)$ be denoted by $\tilde{\alpha}^n_t$ and $\bar{v}^n_t(\tilde{P}^*, \tilde{R}^*)$ be denoted by $\tilde{v}^n_t$. We use the same shorthand notation for $z^n_t(\tilde{P}^*, \tilde{R}^*)$, $\bar{l}^n_t(\tilde{P}^*, \tilde{R}^*)$, and $\bar{s}^n_{t+}(\tilde{P}^*, \tilde{R}^*)$.

Remember that by the construction of $\widetilde{\mathcal{S}}^+_t$, the set of iterations $\mathcal{N}^+_t(\tilde{P}^*, \tilde{R}^*)$ is finite and for all $n \ge N^k_t \ge \bar{N}$, $\bar{v}^n_t(\tilde{P}^*, \tilde{R}^*) \ge z^n_t(\tilde{P}^*, \tilde{R}^*)$. Also (P^n_t, R^n_t) is the state visited by the algorithm at iteration n and time period t. We consider three different cases.

Case 1. $\tilde{P}^* = P^n_t$ and $\tilde{R}^* = R^n_t$.

In this case, $(\tilde{P}^*, \tilde{R}^*)$ is the state being visited by the algorithm at iteration n at time t. Thus,

$$\begin{aligned}\tilde{v}^n_t \ge \tilde{z}^n_t &= (1 - \tilde{\alpha}^n_t)\tilde{v}^{n-1}_t + \tilde{\alpha}^n_t \hat{v}^n_{t+1}(R^n_t)\\ &\ge (1 - \tilde{\alpha}^n_t)(\tilde{l}^{n-1}_t - \tilde{s}^{n-1}_{t+}) + \tilde{\alpha}^n_t \hat{v}^n_{t+1}(R^n_t) - \tilde{\alpha}^n_t (H\bar{v}^{n-1})_t(P^n_t, R^n_t) + \tilde{\alpha}^n_t (H\bar{v}^{n-1})_t(P^n_t, R^n_t) \qquad \text{(B.13)}\\ &\ge (1 - \tilde{\alpha}^n_t)(\tilde{l}^{n-1}_t - \tilde{s}^{n-1}_{t+}) - \tilde{\alpha}^n_t \hat{s}^n_{t+1+}(R^n_t) + \tilde{\alpha}^n_t (HL^k)_t(P^n_t, R^n_t) \qquad \text{(B.14)}\\ &= \tilde{l}^n_t - \big((1 - \tilde{\alpha}^n_t)\tilde{s}^{n-1}_{t+} + \tilde{\alpha}^n_t \hat{s}^n_{t+1+}(R^n_t)\big) \qquad \text{(B.15)}\\ &\ge \tilde{l}^n_t - \big(\max(0, (1 - \tilde{\alpha}^n_t)\tilde{s}^{n-1}_{t+} + \tilde{\alpha}^n_t \hat{s}^n_{t+1+}(R^n_t))\big)\\ &= \tilde{l}^n_t - \tilde{s}^n_{t+}. \qquad \text{(B.16)}\end{aligned}$$

The first inequality is a result of the construction of set $\widetilde{\mathscr{S}}_t^+$ and Equation (B.13) is due to the induction hypothesis. As $n \geq N_t^k$, inequality (B.7) explains Equation (B.14). Finally, Equations (B.15) and (B.16) come from the definition of the stochastic sequences $\tilde{l}_t^n$ and $\tilde{s}_{t+}^n$, respectively.

Case 2. $\tilde{P}^* = P_t^n$ and $\tilde{R}^* = R_t^n + 1$.

This case is analogous to the previous one, except that we use the sample slope $\hat{v}_{t+1}^n(R_t^n + 1)$ instead of $\hat{v}_{t+1}^n(R_t^n)$. We also consider the $(P_t^n, R_t^n + 1)$ component instead of (P_t^n, R_t^n). Moreover, inequality (B.8) instead of Equation (B.7) is used to explain the inequality corresponding to Equation (B.14).

Case 3. Else.

Here, the state $(\tilde{P}^*, \tilde{R}^*)$ is not being updated at iteration n at time t due to a direct observation of sample slopes. Then, $\tilde{\alpha}_t^n = 0$ and, hence,

$$\tilde{l}_t^n = \tilde{l}_t^{n-1} \qquad \text{and} \qquad \tilde{s}_{t+}^n = \tilde{s}_{t+}^{n-1}.$$

Therefore, from the construction of set $\widetilde{\mathscr{S}}_t^+$ and the induction hypothesis,

$$\tilde{v}_t^n \geq \tilde{z}_t^n = \tilde{v}_t^{n-1} \geq \tilde{l}_t^{n-1} - \tilde{s}_{t+}^{n-1} = \tilde{l}_t^n - \tilde{s}_{t+}^n. \quad \square$$

Proof of Lemma 6.4.

Assumptions. Assume stepsize conditions (7)–(9). Moreover, for given $t \in \{0, \ldots, T-1\}$, $k \geq 0$, and index N_t^k, assume for an iteration n that inequalities (B.7) and (B.8) hold true on $\{n \geq N_t^k\}$.

We prove the first statement. The second one is symmetrical.

Fix $\omega \in \Omega$ and omit the dependence of the random elements on w. Given $t \in \{0, \ldots, T-1\}$, $k \geq 0$ and state $(P, R) \in \bar{\mathscr{S}}_t$, if $|\mathscr{N}_t^+(P, R+1)| = \infty$, assume there exists an index $N_t^{k,l}(P, R)$ such that for all iterations $n \geq N_t^{k,l}(P, R)$, it holds that $\bar{v}_t^{n-1}(P, R) \geq L_t^k(P, R)$.

We start by showing that there exists an index $N_t^{k,s}(P, R+1)$ such that $\bar{v}_t^{n-1}(P, R+1) \geq L_t^k(P, R+1) - \bar{s}_{t+}^{n-1}(P, R+1)$ for all $n \geq N_t^{k,s}(P, R+1)$. Then, we show for all $\epsilon > 0$ that there is an integer $N_t^{k,\epsilon}(P, R+1)$ such that $\bar{v}_t^{n-1}(P, R+1) \geq L_t^k(P, R+1) - \epsilon$ for all $n \geq N_t^{k,\epsilon}(P, R+1)$. Finally, using these results, we prove the existence of an integer $N_t^{k,l}(P, R+1)$ such that $\bar{v}_t^{n-1}(P, R+1) \geq L_t^k(P, R+1)$ for all $n \geq N_t^{k,l}(P, R+1)$.

Let $N_t^{k,s}(P, R+1) = \min\{n \in \mathscr{N}_t^+(P, R+1)\colon n \geq N_t^{k,l}(P, R)\} + 1$. Because $|\mathscr{N}_t^+(P, R+1)|$ is infinite, $N_t^{k,s}(P, R+1)$ is well-defined. Note that as $N_t^{k,s}(P, R+1) - 1 \in \mathscr{N}_t^+(P, R+1)$, the slope corresponding to state $(P, R+1)$ was decreased at iteration $N_t^{k,s}(P, R+1) - 1$ and time period t. Hence, $\bar{v}_t^{N_t^{k,s}(P,R+1)-1}(P, R) = \bar{v}_t^{N_t^{k,s}(P,R+1)-1}(P, R+1)$. Redefine the noise sequence $\{\bar{s}_{t+}^m(P, R+1)\}_{m \geq 0}$ introduced in the proof of Theorem 6.1 using $N_t^{k,s}(P, R+1)$ instead of N_t^k.

We prove that $\bar{v}_t^{n-1}(P, R+1) \geq L_t^k(P, R+1) - \bar{s}_{t+}^{n-1}(P, R+1)$ for all $n \geq N_t^{k,s}(P, R+1)$ by induction on n. For the base case $n = N_t^{k,s}(P, R+1)$, from our choice of the index $N_t^{k,s}(P, R+1)$ and the monotone decreasing property of L_t^k, we have that:

$$\bar{v}_t^{n-1}(P, R+1) = \bar{v}_t^{n-1}(P, R) \geq L_t^k(P, R) \geq L_t^k(P, R+1) = L_t^k(P, R+1) - \bar{s}_{t+}^{n-1}(P, R+1).$$

Now, we suppose for a given $n > N_t^{k,s}(P, R+1)$ that $\bar{v}_t^{n-1}(P, R+1) \geq L_t^k(P, R+1) - \bar{s}_{t+}^{n-1}(P, R+1)$. We shall prove that $\bar{v}_t^n(P, R+1) \geq L_t^k(P, R+1) - \bar{s}_{t+}^n(P, R+1)$. We consider two cases.

Case 1. $n \in \mathscr{N}_t^+(P, R+1)$.

In this case, a projection operation took place at iteration n. This fact and the monotone decreasing property of L^k give us

$$\bar{v}_t^n(P, R+1) = \bar{v}_t^n(P, R) \geq L_t^k(P, R) \geq L_t^k(P, R+1) \geq L_t^k(P, R+1) - \bar{s}_{t+}^n(P, R+1).$$

Case 2. $n \notin \mathscr{N}_t^+(P, R+1)$.

The analysis of this case is analogous to the proof of inequality (B.12) of Lemma 6.3. The difference is that we consider $L_t^k(P, R+1)$ instead of $\bar{l}_t^n(P, R+1)$. Therefore, as in Lemma 6.3, we break it down to three possibilities:

Case 2(i). $P = P_t^n$ and $R + 1 = R_t^n$.

In this case, $(P, R+1)$ is the state being visited by the algorithm at iteration n and time t, implying that $\bar{\alpha}_t^n(P, R+1) = \alpha_t^n$. Thus,

$$\begin{aligned}
\bar{v}_t^n(P, R+1) &\geq z_t^n(P, R+1) = (1 - \alpha_t^n)\bar{v}_t^{n-1}(P, R+1) + \alpha_t^n \hat{v}_{t+1}^n(R_t^n) \\
&\geq (1 - \alpha_t^n)(L_t^k(P, R+1) - \bar{s}_{t+}^{n-1}(P, R+1)) + \alpha_t^n \hat{v}_{t+1}^n(R_t^n) \\
&\quad - \alpha_t^n (H\bar{v}^{n-1})_t(P, R_t^n) + \alpha_t^n (H\bar{v}^{n-1})_t(P, R_t^n) \qquad \text{(B.17)}
\end{aligned}$$

$$\geq (1-\alpha_t^n)\big(L_t^k(P,R+1)-\bar{s}_{t+}^{n-1}(P,R+1)\big) - \alpha_t^n \hat{s}_{t+1+}^n(R_t^n) + \alpha_t^n (HL^k)_t(P,R_t^n) \tag{B.18}$$

$$\geq L_t^k(P,R+1) - \big((1-\alpha_t^n)\bar{s}_{t+}^{n-1}(P,R+1) + \alpha_t^n \hat{s}_{t+1+}^n(R_t^n)\big) \tag{B.19}$$

$$\geq L_t^k(P,R+1) - \big(\max(0,(1-\alpha_t^n)\bar{s}_{t+}^{n-1}(P,R+1) + \alpha_t^n \hat{s}_{t+1+}^n(R_t^n))\big)$$

$$= L_t^k(P,R+1) - \bar{s}_{t+}^n(P,R+1). \tag{B.20}$$

The first inequality is a result of the fact that $n \notin \mathcal{N}_t^+(P,R+1)$, meaning that the slope was not decreased due to a projection iteration. The induction hypothesis explains Equation (B.17) while the definition of $\hat{s}_{t+1+}^n(R_t^n)$ and inequality (B.7) explain Equation (B.18). Finally, Equation (B.19) is due to Equation (19), and Equation (B.20) comes from the definition of the stochastic sequence $\{\bar{s}_{t+}^m(P,R+1)\}_{m\geq 0}$.

Case 2(ii). $P = P_t^n$ and $R+1 = R_t^n + 1$.

Similar to the previous case, we have that $\bar{\alpha}_t^n(P,R+1) = \alpha_t^n$. The rest of the analysis is analogous to the previous one, except that we use the sample slope $\hat{v}_{t+1}^n(R_t^n+1)$ instead of $\hat{v}_{t+1}^n(R_t^n)$. We also consider $(H\bar{v}^{n-1})_t(P,R_t^n+1)$ instead of $(H\bar{v}^{n-1})_t(P,R_t^n)$. We use Equation (B.8) instead of Equation (B.7) to justify the inequality equivalent to Equation (B.18).

Case 2(iii). Else.

Here, the state $(P,R+1)$ is not being updated at iteration n and time t due to a direct observation of sample slopes. Then, $z_t^n(P,R+1) = \bar{v}_t^{n-1}(P,R+1)$ and $\bar{s}_{t+}^n(P,R+1) = \bar{s}_{t+}^{n-1}(P,R+1)$. Moreover, from the induction hypothesis and from the fact that $n \notin \mathcal{N}_t^+(P,R+1)$,

$$\bar{v}_t^n(P,R+1) \geq z_t^n(P,R+1) = \bar{v}_t^{n-1}(P,R+1) \geq L_t^k(P,R+1) - \bar{s}_{t+}^{n-1}(P,R+1)$$
$$= L_t^k(P,R+1) - \bar{s}_{t+}^n(P,R+1).$$

Hence, we have proved that for all $k \geq 0$, there exists an integer $N_t^{k,s}(P,R+1)$ such that $\bar{v}_t^{n-1}(P,R+1) \geq L_t^k(P,R+1) - \bar{s}_{t+}^{n-1}(P,R+1)$ for all $n \geq N_t^{k,s}(P,R+1)$.

Pick $\epsilon > 0$. We move on to show the existence of an integer $N_t^{k,\epsilon}(P,R+1)$ such that $\bar{v}_t^{n-1}(P,R+1) \geq L_t^k(P,R+1) - \epsilon$ for all $n \geq N_t^{k,\epsilon}(P,R+1)$. We consider two cases: (i) $(P,R+1) \in \bar{\mathcal{S}}_t^*$ and (ii) $(P,R+1) \notin \bar{\mathcal{S}}_t^*$. For the first case, Lemma 6.2 tells us that $\{\bar{s}_{t+}^n(P,R+1)\}_{n\geq 0}$ goes to zero. Then, there exists $N^\epsilon > 0$ such that $\bar{s}_{t+}^n(P,R+1) < \epsilon$ for all $n \geq N^\epsilon$. Therefore, we just need to choose $N_t^{k,\epsilon}(P,R+1) = \max(N_t^{k,s}(P,R+1), N^\epsilon)$. For the second case, $\bar{\alpha}_t^n(P,R+1) = 0$ for all $n \geq N_t^{k,s}(P,R+1)$ and $\bar{s}_{t+}^{N_t^{k,s}(P,R+1)-1}(P,R+1) = 0$. Thus, $\bar{s}_{t+}^n(P,R+1) = \bar{s}_{t+}^{N_t^{k,s}(P,R+1)-1}(P,R+1) = 0$ for all $n \geq N_t^{k,s}(P,R+1)$ and we just have to choose $N_t^{k,\epsilon}(P,R+1) = N_t^{k,s}(P,R+1)$.

We are ready to conclude the proof. For that matter, we use the result of the previous paragraph. Let $\epsilon = v_t^*(P,R+1) - L_t^k(P,R+1) > 0$. Because $\{L_t^k(P,R+1)\}_{k\geq 0}$ increases to $v_t^*(P,R+1)$, there exists $k' > k$ such that $v_t^*(P,R+1) - L_t^{k'}(P,R+1) < \epsilon/2$. Thus, $L_t^{k'}(P,R+1) - L_t^k(P,R+1) > \epsilon/2$ and the result of the previous paragraph tells us that there exists $N_t^{k',\epsilon/2}(P,R+1)$ such that $\bar{v}_t^{n-1}(P,R+1) \geq L_t^{k'}(P,R+1) - \epsilon/2 > L_t^k(P,R+1) + \epsilon/2 - \epsilon/2 = L_t^k(P,R+1)$ for all $n \geq N_t^{k',\epsilon/2}(P,R+1)$. Therefore, we just need to choose $N_t^{k,l}(P,R+1) = N_t^{k',\epsilon/2}(P,R+1)$ and we have proved that for all $k \geq 0$, there exists $N_t^{k,l}(P,R+1)$ such that $\bar{v}_t^{n-1}(P,R+1) \geq L_t^k(P,R+1)$ for all $n \geq N_t^{k,l}(P,R+1)$. □

Acknowledgments. This research was supported in part by Air Force Office of Scientific Research Grant AFOSR Contract FA9550-08-1-0195. The authors thank the referees for their work, with special thanks to one particularly dedicated reviewer who helped clarify and improve the presentation as well as tighten several critical arguments.

References

[1] Abounadi, J., D. P. Bertsekas, V. Borkar. 2002. Stochastic approximation for non-expansive maps: Q-learning algorithms. *SIAM J. Control Optim.* **41** 1–22.

[2] Ahmed, S., A. Shapiro. 2002. The sample average approximation method for stochastic programs with integer recourse. http://www.optimization-online.org.

[3] Barto, A. G., S. J. Bradtke, S. P. Singh. 1995. Learning to act using real-time dynamic programming, artificial intelligence. *Computat. Res. Interaction Agency* **72**(Special Volume) 81–138.

[4] Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

[5] Breiman, L. 1992. *Probability.* SIAM, Philadelphia.

[6] Chen, Z.-L., W. B. Powell. 1999. A convergent cutting-plane and partial-sampling algorithm for multistage stochastic linear programs with recourse. *J. Optim. Theory Appl.* **102**(3) 497–524.
[7] Cybenko, G., R. Gray, K. Moizumi. 1997. Q-learning: A tutorial and extensions. S. W. Ellacott, J. C. Mason, I. J. Anderson, eds. *Mathematics of Neural Networks: Models, Algorithms, and Applications.* Kluwer Academic Publishers, 24–33.
[8] Duff, M. O. 1995. Q-learning for bandit problems. Technical report, Department of Computer Science, University of Massachusetts, Amherst, MA.
[9] Even-Dar, E., Y. Mansour. 2004. Learning rates for q-learning. *J. Machine Learn. Res.* **5** 1–25.
[10] Godfrey, G. A., W. B. Powell. 2002. An adaptive, dynamic programming algorithm for stochastic resource allocation problems I: Single period travel times. *Transportation Sci.* **36**(1) 21–39.
[11] Halman, N., D. Klabjan, M. Mostagir, J. Orlin. 2006. A fully polynomial time approximation scheme for single-item stochastic lot-sizing problems with discrete demand. Research Paper 4582-06, MIT Sloan School of Managemnt, Cambridge, MA, January, http://ssrn.com/abstract=882101.
[12] Higle, J. L., S. Sen. 1991. Stochastic decomposition: An algorithm for two stage linear programs with recourse. *Math. Oper. Res.* **16**(3) 650–669.
[13] Jaakkola, T., M. I. Jordan, S. P. Singh. 1994. Convergence of stochastic iterative dynamic programming algorithms. J. D. Cowan, G. Tesauro, J. Alspector, eds. *Advances in Neural Information Processing Systems*, Vol. 6. Morgan Kaufmann Publishers, San Francisco, 703–710.
[14] Levi, R., R. Roundy, D. B. Shmoys. 2007. Provably near-optimal sampling-based policies for stochastic inventory control models. *Math. Oper. Res.* **32**(4) 821–839.
[15] Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley and Sons, New York.
[16] Powell, W. B., A. Ruszczyński, H. Topaloglu. 2004. Learning algorithms for separable approximations of stochastic optimization problems. *Math. Oper. Res.* **29**(4) 814–836.
[17] Rummery, G., M. Niranjan. 1994. On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR166, Cambridge University Engineering Department, Cambridge.
[18] Rusmevichientong, P., B. Van Roy, P. W. Glynn. 2006. A non-parametric approach to multi-product pricing. *Oper. Res.* **54**(1) 82–98.
[19] Shapiro, A. 2003. Monte Carlo sampling methods. A. Ruszczyński, A. Shapiro, eds. *Handbooks in Operations Research and Management Science: Stochastic Programming*, Vol. 10. Elsevier, Amsterdam, 353–425.
[20] Shiryaev, A. N. 1996. Probability theory. *Graduate Texts in Mathematics*, Vol. 95. Springer-Verlag, New York.
[21] Singh, S., T. Jaakkola, M. L. Littman, C. Szepesvari. 2000. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learn.* **38**(3) 287–308.
[22] Sutton, R. S., A. G. Barto. 1998. *Reinforcement Learning*. The MIT Press, Cambridge, MA.
[23] Topaloglu, H., W. B. Powell. 2003. An algorithm for approximating piecewise linear concave functions from sample gradients. *Oper. Res. Lett.* **31**(1) 66–76.
[24] Topaloglu, H., W. B. Powell. 2006. Dynamic programming approximations for stochastic, time-staged integer multicommodity flow problems. *INFORMS J. Comput.* **18**(1) 31–42.
[25] Tsitsiklis, J. 2002. On the convergence of optimistic policy iteration. *J. Machine Learn. Res.* **3** 59–72.
[26] Tsitsiklis, J. N. 1994. Asynchronous stochastic approximation and Q-learning. *Machine Learn.* **16** 185–202.
[27] van Ryzin, G., J. McGill. 2000. Revenue management without forecasting or optimization: An adaptive algorithm for determining airline seat protection levels. *Management Sci.* **46**(6) 760–775.
[28] Van Slyke, R., R. Wets. 1969. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM J. Appl. Math.* **17**(4) 638–663.
[29] Watkins, C. J. C. H., P. Dayan. 1992. Q-learning. *Machine Learn.* **8** 279–292.